Statistical Challenges in the Detection of Mutation and Variation using High Throughput Sequencing

A Thesis submitted for the Degree of Doctor of Philosophy

Susanne Pfeifer
Merton College, University of Oxford

Trinity 2012

Statistical Challenges in the Detection of Mutation and Variation using High Throughput Sequencing

Susanne Pfeifer, Merton College, University of Oxford
A Thesis submitted for the Degree of Doctor of Philosophy
Trinity 2012

The aim of this thesis is to obtain a better understanding of mutation rates within as well as between the genomes of humans and chimpanzees using data generated by high throughput sequencers. I will start with a review of the field and an overview of the technologies and protocols used to generate and analyse high throughput sequencing data. I apply some of the discussed techniques to show that there is evidence of a selective advantage of pathogenic *de novo* mutations in the Fibroblast Growth Factor Receptor 3 gene in the male germ line of humans. Furthermore, I use some of the methods to generate a map of genome-wide sequence variation in Western chimpanzees. Ever since Darwin [Darwin, 1871] and Huxley [Huxley, 1863] postulated more than a century ago that African great apes are our closest living evolutionary relatives, the study of chimpanzee individuals is of great scientific interest from an evolutionary point of view, as comparisons between the genomes of human and chimpanzee offer the potential to help to understand the molecular basis for similarities and differences between the two species. I use the generated data to explore the breadth of the nucleotide diversity in the chimpanzee genome in order to shed light on whether or not the local variation in mutation rate has been conserved since the divergence of the two species and to place human nucleotide diversity into perspective with an evolutionary closely related species. I explore the relationship of nucleotide diversity in chimpanzees with specific large-scale genome features to reveal a number of highly significant correlations which explain over 40% of the observed variation. I use data from the 1000 Genomes Project to examine the occurrence of ancestral polymorphisms shared between human and chimpanzee on a genome-wide scale. These ancestral polymorphisms do not only influence fine-scale divergence rates across the genome in very closely related species, they are also good candidates for regions under balancing selection and thus, they are a useful tool to study long-time population demographics and speciation. Using these variants, I postulate that long-term balancing selection may be more common than previously believed. I conclude with a discussion of the results contained in the body of the thesis and suggest a number of areas for future research.

Acknowledgements

The first thing I would like to mention is that I have an attention to detail which often comes along with the problem to cut a long story short, thus I would like to thank my examiners, Dmitry Filatov and Ines Hellmann, for taking the time to consider my thesis. Up until a few years ago, I was strongly convinced that my Master thesis would be the most substantial amount of work that I would ever have to write (and more importantly that I would ever be willing to write), and although I managed quite well to avoid writing this thesis for a good four years, my supervisor, Gil McVean, eventually pushed me softly to get it 'out of my way'. I would like to thank him not only for his support and motivation, but also for giving me the opportunity to work on some really exciting projects. He created an inspiring research environment and even if a growing group size over the last years meant that finding time with him was a bit more difficult, it also resulted in a larger circle of colleges and friends which I really value.

My thesis was supported by a research studentship from the Engineering and Physical Sciences Research Council (EPSRC). Part of the work in this thesis was presented at the 59th Annual Meeting American Society of Human Genetics, Honolulu, USA, October 2009; at the Society for Molecular Biology and Evolution 2010 Annual Meeting, Lyon, France, July 2010; at the EMBO Bioinformatics and Comparative and Genome Analyses Course, Paris, France, July 2011; at the 17th Annual European Meeting of PhD Students in Evolutionary Biology, Seia, Portugal, August 2011; and at the 12th ICHG/61th Annual Meeting American Society of Human Genetics, Montréal, Canada. I would like to thank the EPSRC, EMPSEB, the Department of Statistics, and Merton College, Oxford for financial support in attending these conferences.

I would further like to thank all members of the Mathematical Genetics and Bioinformatics Group of the Department of Statistics for creating a friendly, welcoming, and cookie-filled atmosphere. In particular, I would like to thank Adam Auton without whom I probably would not have survived the 'dark sides' of my DPhil (at least not as well) and who had an incredible patience to keep up with my questions over the last years. I am also thankful to Oliver Venn, Ellen Leffler, Zam Iqbal, Cord Melton, Julian Mallers, Laure Ségurel, Simon Myers, Peter Donnelly, and Molly

Przeworski, with whom it was a pleasure to collaborate in the PanMap Project. Many thanks go to my office mates and colleagues over the years, including Alexander Dilthey, Denise Xifara, Loukas Moutsianas, and David Reshef for their cheerfulness and the occasional nice glass of wine or after-work dinner.

I would further like to thank all the incredible friends I made here, which helped me to keep a good life-work balance: Franck Silva, who, after my arrival in Oxford, turned my world upside-down, but with whom, despite the difficulties, I share some nice memories, Sara-Jane Dunn without whose constant chattering my skills of the English language would still be at a primary school level, Guido Klingbeil for being a good friend and rescuing me out of my many miseries, Samuel Hugueny for all the gossip he shared with me as well as his (I-know-everything-better) advise which I should have listen to more often, Bartu 'Maverick' Ahiska, and Jochen 'Ice' Klingelhöfer for what can only be described as my most eventful carnival, Camille Guillemin de Monplanet, Karl Anderson, and Kambez Benam for some incredible terms in college and a spectacular time-turning ceremony, James Anderson, Jem Pearson, Luke Heaton, Phil Locascio, Aaron Smith, Gabriel Rosser, Matt Gibb, Anna Lewis, Courtney Bishop, Kit Yates, and Tom Doel for plenty of nice formal halls, pub crawls, bops, and parties, Julia Shanks, Simon Fairclough, Katie Moore, and Ricardo Engel to take me into their home as a (longer-time) 'temporary' room mate, Gianna Hessel for calming me down and keeping my sugar level high with plenty of chocolate, and all the other interesting short- and long-term acquaintances that I might have forgotten. You all made Oxford a truly unforgettable place for me! Equally important though, although physically slightly further away, are my two best friends, Ingrid and Martina Höring, who I greatly missed in Oxford, but whose support I always cherished. Thank you so much for all the letters, postcards, pictures, and presents you sent me over the last years!

I am very thankful to Graeme Johnstone, who has been a great source of joy, and who has brought some sunshine into my everyday life. Your companionship has been so important to me in these last months!

Last but not least, I would like to thank some of the most important people in my life. Marko H. Jung who provided me with constant support and strength over the years and who never lost faith in me. You are truly important to me! I am most thankful to my family, especially my mum, who is the foundation for all my love of life and happiness, and without whose support I would not be where I am today. Thank you for everything from all my heart!

Contents

1 Introduction **1**
- 1.1 Causes of Mutation . . . 3
- 1.2 Types of Mutation . . . 4
- 1.3 Mutation Rate Variation . . . 8
- 1.4 Measurements of Polymorphism . . . 15
- 1.5 Linkage Disequilibrium . . . 17
- 1.6 Outline . . . 20

2 Generation and Analysis of High Throughput Sequencing Data **21**
- 2.1 High Throughput Sequencing . . . 22
- 2.2 Alignment of High Throughput Sequencing Data . . . 27
- 2.3 Variant Calling from High Throughput Sequencing Data . . . 40
- 2.4 Validation . . . 45

3 Estimating *De Novo* Mutation Rates from High Throughput Sequencing Data **49**
- 3.1 Fibroblast Growth Factor Receptor 3 . . . 50
- 3.2 Aim of the Study . . . 51
- 3.3 Study Design . . . 52
- 3.4 Knowledge-based Model to Estimate Mutation Levels . . . 55
- 3.5 Black-box Model to Estimate Mutation Levels . . . 59
- 3.6 Analysis of Mutation Levels . . . 65
- 3.7 Discussion . . . 67

4 PanMap - A Map of Genome-wide Sequence Variation in Western Chimpanzees **71**
- 4.1 Aim of the Study . . . 74
- 4.2 Study Design . . . 74
- 4.3 Identification of Genomic Variation . . . 75
- 4.4 Discussion . . . 107

5 PanMap - Analysis of Genetic Diversity in Western Chimpanzees **111**
- 5.1 Aim of the Study . . . 113
- 5.2 Accessible Genome . . . 114
- 5.3 Genome-wide Patterns of Nucleotide Diversity . . . 117
- 5.4 Discussion . . . 131

6 PanMap - Analysis of Variation Shared between Humans and Chimpanzees **135**
- 6.1 Aim of the Study . . . 137
- 6.2 Data . . . 138
- 6.3 Sequence-context of Shared Polymorphisms . . . 145
- 6.4 Simulation of Recurrent Mutations . . . 150
- 6.5 Annotation of Shared Polymorphisms . . . 152
- 6.6 Linkage Disequilibrium Patterns of Shared Polymorphisms . . . 156
- 6.7 Discussion . . . 168

7 Discussion **173**

8 Outlook **179**

A Sequencing Technologies **183**
A.1 First-generation Sequencers 183
A.2 Second-generation Sequencers 186
A.3 Third-generation Sequencers 198

B FASTQ Format **207**

C Alignment with MAQ **209**

D SAM and BAM Format **211**
D.1 SAM Format 211
D.2 BAM Format 213

E FASTA Format **215**

F VCF Format **217**

G MHC Region **221**

H Sequence-context of Coding and Non-coding Shared Polymorphisms **223**

I Concept of Replication Slippage **225**

J Polymorphisms involved in the observed Haplotype Structure **227**

Bibliography **231**

List of Figures

1.1 Types of Mutation: Missense Mutation, Insertion/Deletion, and Frameshift 5
1.2 Types of Mutation: Duplication, Inversion, and Translocation 7
1.3 Concept of Linkage 19

2.1 Illumina Sequencing: Protocol 25
2.2 Comparison between Hash- and BWT-based Alignment Algorithms 28
2.3 Hash-based Alignment: Construction of a Hash Table 30
2.4 MAQ: Indexing and Short Indel Detection 31
2.5 BWT: Construction of a Burrows-Wheeler-transformed String 37
2.6 BWT: Prefix Trie 37

3.1 K650 Wild-type Codon and the Effect of Substitutions 51
3.2 Strategy to Quantify Mutation Levels at the K650 Codon 53
3.3 Characteristics of the Sequenced DNA 55
3.4 Read Error Structure and Quality Score Calibration 57
3.5 Distribution of Read Quality Scores 58
3.6 MCMC Convergence Diagnostics 64
3.7 Mutation Levels estimated by the Titration Series 66
3.8 Mutation Levels of the K650E Mutation 66
3.9 Cumulative Mutation Levels at the K650 Codon 67

4.1 Distribution of the Common Chimpanzee 72
4.2 Alignment Work Flow (Stampy) 78
4.3 Variant Calling Work Flow 80
4.4 Depth of Read Coverage of Chromosome 6 Region 31Mb-34Mb 82
4.5 Recalibration: Residual Error 85
4.6 Recalibration: Reported versus Empirical Quality Scores 86
4.7 Variant Annotations and Variant Filter Criteria 92
4.8 Distribution of Variants at different Allele Frequencies 97
4.9 Distribution of Variants per Chromosome and along Chromosome 1 99
4.10 Read Depth per Individual at Polymorphic Sites 101
4.11 Assessment of SNP Discovery Power 107

5.1 Nucleotide Diversity between Chromosomes 119
5.2 Comparison of the Nucleotide Diversity Estimates π and Θ 121
5.3 Nucleotide Diversity within Chromosomes 123
5.4 Regional Variation of Nucleotide Diversity 124
5.5 Multiple Linear Regression: Diagnostic Plots 126
5.6 Multiple Linear Regression: Predictors to Model Nucleotide Diversity 127
5.7 Multiple Linear Regression: Pairwise Correlation between Predictors 129
5.8 Multiple Linear Regression: Relative Importance of Predictors 129

6.1 Concept of Ancestral Polymorphisms 136
6.2 Distribution of Shared Polymorphisms between Chromosomes 142
6.3 Characteristics of Shared Polymorphisms: Read Coverage and Mapping Quality . . 143
6.4 Characteristics of Shared Polymorphisms: Site Frequency Spectrum 145

6.5 Sequence-context of Shared Polymorphisms 147
6.6 Comparison of Estimates for Obs/Exp Rate of Sharing 148
6.7 10bp Local Sequence-context of Shared and Non-shared Polymorphisms 149
6.8 SNP Density around Non-synonymous Polymorphisms 155
6.9 LD Structure of Shared Polymorphisms 155
6.10 Polymorphisms in the *TRIM31* Gene 167
6.11 Polymorphisms in the *ABO* Gene 170
6.12 Polymorphisms in the *FMO2* Gene 171
6.13 Polymorphisms in the *PRAM1* Gene 172

A.1 Sanger Sequencing: Protocol 185
A.2 454 Sequencing: Protocol 190
A.3 SOLiD Sequencing: Protocol 194
A.4 SOLiD Sequencing: 2 Base Encoding 195
A.5 Ion Torrent PGM Sequencing: Protocol 197
A.6 HeliScope Sequencing: Protocol 200
A.7 PacBio RS Sequencing: Protocol 203
A.8 Nanopore Sequencing: Protocol 204

C.1 Alignment Work Flow (MAQ) 210

H.1 Sequence-context of Shared Non-coding Polymorphisms 223
H.2 Sequence-context of Shared Coding Polymorphisms 224

I.1 Concept of Replication Slippage 225

List of Tables

2.1 Smith-Waterman Algorithm 41

3.1 Nucleotide Substitution Rates 56
3.2 Posterior Means for the Background Parameter W 65

4.1 Sequencing Data 79
4.2 Initial SNP Calls 89
4.3 Filtered SNP Calls from panTro2.1-based Mappings 98
4.4 Filtered SNP Calls from hg18-based Mappings and from the Intersection Set . . . 100
4.5 Validation Experiment 105
4.6 Duplicate Sample Genotype Concordance 106

5.1 Accessible Sites in the panTro2-based and in the hg18-based Data Set 116
5.2 Nucleotide Diversity between Chromosomes 120
5.3 Linear Regression Models for Nucleotide Diversity 130

6.1 Polymorphism Data Sets, Coincident and Shared Polymorphisms 141
6.2 Rate of Shared Polymorphism Pairs due to Recurrent Mutations 151
6.3 Annotation of Polymorphisms 153
6.4 Annotated Shared Polymorphisms 158

A.1 Characteristics of Next-generation Sequencing Platforms 188

D.1 SAM Format: Alignment Information 212

F.1 VCF Format: Header Information 218
F.2 VCF Format: Variant Information 219

G.1 MHC Region 221

J.1 Characteristica of Polymorphisms involved in the observed Haplotype Structure . . 227

Acronyms

A Adenine

ATP Adenosine TriPhosphate

BAC Bacterial Artificial Chromosome

BAM Binary SAM format

BAQ Base Alignment Quality

bp base pair

BWA Burrows-Wheeler Alignment

BWT Burrows-Wheeler Transform

C Cytosine

CCD Charge-Coupled Device

cDNA complementary DNA

CI Confidence Interval

ddNTPs 2',3'-dideoxyNucleotide TriPhosphates

DNA DeoxyriboNucleic Acid

dNTPs 2'-deoxyNucleotide TriPhosphates

EBV Epstein-Barr Virus

emPCR emulsion PCR

EST Expressed Sequence Tag

FDR False Discovery Rate

FGFR2 Fibroblast Growth Factor Receptor 2

FGFR3 Fibroblast Growth Factor Receptor 3

FRET Fluorescent Energy Resonance Transfer

G Guanine

GAII Genome Analyzer II (an Illumina sequencing platform)

GATK Genome Analysis ToolKit

Gb Giga base pairs (= 1,000,000,000bp)

HLA Human Leukocyte Antigen

HCH HypoCHondroplasia

HWE Hardy-Weinberg Equilibrium

indel insertion/deletion

kb kilo base pairs (= 1,000bp)

LD Linkage Disequilibrium

MAF Minor Allele Frequency

MAQ Mapping and Assembly with Qualities

Mb Mega base pairs (= 1,000,000bp)

MCMC Markov Chain Monte Carlo

MHC Major Histocompatibility Complex

Mya Million years ago

NGS Next-Generation Sequencing

PAR1 PseudoAutosomal Region 1

PAR2 PseudoAutosomal Region 2

PCR Polymerase Chain Reaction

PPi Diphosphate Group

PTP PicoTiter Plate

RMSE Root Mean Square Error

RNA Ribonucleic Acid

SADDAN Severe Achondroplasia with Developmental Delay and Acanthosis Nigricans

SAM Sequence Alignment/Map format

SBS Sequencing-By-Synthesis

SD Standard Deviation

SNP Single Nucleotide Polymorphism

STR Single Tandem Repeat

T Thiamine

TDI Thanatophoric Dysplasia Type I

TDII Thanatophoric Dysplasia Type II

Ts Transition

Tv Transversion

VAT Variant Annotation Tool

VCF Variant Calling Format

VNTR Variable Number Tandem Repeat

YRI Yoruba population

ZMW Zero-Mode Waveguide

Chapter 1

Introduction

Molecular evolution is driven by an interplay of different forces: mutation, recombination, selection, and genetic drift. Out of these four, mutation, a random event that leads to an alteration in the DNA sequence of an organism, is one of the most fundamental processes in biology. Often regarded as the raw material of evolution, mutations continuously give rise to new variations in the genome, making them the primary source of genetic diversity within populations as well as the divergence between species. The evolutionary fate of a new mutation depends not only on the frequency with which it arises, but also on its selective effect (deleterious, neutral, or beneficial) as well as its dominance level (the degree to which it affects an individual's fitness in the heterozygous condition). However, as newly introduced mutations are rare, their fate is mostly determined by genetic drift (a random process that causes an allelic variant to raise or fall in frequency, eventually leading to either a fixation or a loss of the variant) rather than natural selection. Only if a mutation reaches an appreciable frequency, natural selection will be sufficiently effective to either favour beneficial mutations that typically increase an individual's fitness (e.g. by improving the reproduction or the chances of survival) or to decrease the frequency of deleterious mutations. Recombination can separate and randomise alleles at different loci onto different genetic backgrounds, introducing new combinations which can be (dis-)favoured by selection.

In the absence of selection, the rate at which a mutation reaches fixation is given by the mutation rate [Kimura, 1983] (the number of new mutations per target region per time; typical target regions are bases, base pairs, genes, gametes, or genomes; typical time units are replication or sexual generations [Johnston, 2001]). The rate of mutation is mainly controlled by two opposing forces: the benefits of reducing deleterious mutations imposing a fitness cost to the carrying individual and the costs of increasing fidelity and accuracy by enabling adaptations to changing environments (e.g. the time and energy required for DNA proof-reading) [Kimura, 1967; Kondrashov, 1988; Leigh, 1970; McVean and Hurst, 1997]. Mutation rates are known not to be uniformly distributed across

the genome but rather to vary between different sites. Unfortunately the evolutionary causes underlying this variation are still poorly understood.

As a result, the study of mutation rates as well as the patterns underlying mutational processes across the genome and the forces driving it are important subjects of research. An understanding of these factors is of obvious importance to gain a better knowledge about many biological phenomena (such as the evolution of sex) and evolutionary processes that shaped genome sequences (such as recombination, replication, repair, and transcription [Duret, 2009]). In fact, a better understanding of the processes that generate and maintain variation within and between species is one of the main goals of evolutionary and population genetics as many key questions rely on correct information about genome-wide mutation rates, such as (i) the estimation of evolutionary distance between two species, (ii) the inference of the ancestral state of an allele, (iii) the construction of phylogenies, (iv) the identification of adaptive evolution, (v) the estimation of ancestral population size, and (vi) the detection of functional elements in the genome [Hodgkinson and Eyre-Walker, 2011]. In comparative and functional genomics, knowledge about mutation rates is important when trying to infer selection from patterns of divergence (as realistic null hypothesis of neutral variation are required) and in medical genetics, the understanding of mutation rates is essential as mutations can have severe impacts on human health by causing genetic diseases [Ellegren et al., 2003]. In addition, information about mutation rates can help to shed light on questions regarding the basis of germ line mutations (e.g. elucidating the relative importance of replication errors as source of mutations) [Ellegren et al., 2003].

Most information about mutation rates and their variation has been obtained through indirect approaches which, due to the fact that genomic mutation rates are generally extremely low (although some human genes exhibit mutation rates as high as 10^{-5} per generation, e.g. mutations in the *FGFR3* gene which can cause achondroplasia as well as several other disorders that affect skeletal or skin development [Orioli et al., 1995; Waller et al., 2008], most regions stay evolutionary relatively conserved, showing much lower mutation rates in the order of $\sim 2.5 \times 10^{-8}$ per nucleotide per generation to avoid harmful substitutions [Ellegren et al., 2003; McVean and Hurst, 1997; Nachman and Crowell, 2000]), are both laboratory complex and imprecise and thus, the present knowledge is, despite ongoing investigations, still limited. However, the advent of next-generation sequencing (NGS) technologies in recent years, which enable the generation of large amounts of data within a short amount of time and at much lower costs than traditional sequencing methods, has permitted the large-scale study of the genome of different individuals of a species, making it possible to directly estimate genomic mutation rates at the DNA level, thereby avoiding some of the problems associated with the usage of indirect methods.

The aim of this thesis is to obtain a better understanding of mutation rates within as well as between humans and our closest living evolutionary relatives, chimpanzees, using data generated by NGS. The study of chimpanzees is of great scientific interest from an evolutionary point of view as comparisons between the genomes of human and chimpanzee might help to understand the molecular basis for similarities and differences between the two species. A comparison between mutation rates of the two closely related species not only allows to address several questions dealing with the variation of mutation rates within as well as between the two lineages, but it also gives an indication of how quickly mutation rates evolve. The exploration of the breadth of chimpanzee genome diversity might shed light on whether or not the local variation in mutation rate has been conserved since the divergence of the two species and it might help to place human nucleotide diversity into perspective with an evolutionary closely related relative.

To give an introduction into this area of research, the remainder of this chapter reviews some of the work which was performed in earlier studies. A general introduction of several factors that cause genetic mutations is presented, together with a summary of the types of mutations found at the DNA sequence level in different organisms and an overview of the current knowledge of mutation rate variation in genomes. The chapter discusses methods which are widely used to measure levels of DNA polymorphism as a surrogate for mutation rate and to study the association between different polymorphisms. The last section gives a synopsis of the aims of this thesis and outlines the content of the further chapters.

1.1 Causes of Mutation

Mutations are caused by several factors, amongst them are (i) the intrinsic stability of the genome and its exposure to endogenous (e.g. oxygen-free radicals produced during aerobic respiration) or exogenous (e.g. UV radiation [Pfeifer et al., 2005] or chemical mutagens such as tobacco [Pfeifer et al., 2002]) mutagenic agents mainly increasing mutation rates in somatic tissue [Friedberg et al., 2004], (ii) the extent of DNA replication errors (e.g. certain genomic regions such as repeats are more vulnerable to mispairing during the replication and recombination processes making them prone to higher error rates [Jeffreys et al., 1988]), (iii) the efficiency of DNA repair processes (e.g. the direct reversal of nucleotide damage by different enzymes or the repair of DNA double-strand breaks by recombination repair and non-homologous end-joining [Eisen and Hanawalt, 1999]), as well as (iv) pathways that modulate the impact of these mutations on the fitness of the individual [Baer et al., 2007]. These factors frequently bias the direction of mutation in certain ways (e.g. repair enzymes might bias the direction of mutation dependent on different local sequence-contexts and different mutation types [Brown and Jiricny, 1988]).

1.2 Types of Mutation

Different types of mutation can be distinguished:

Point Mutations: The most frequent type of mutation is a substitution of a single nucleotide for another (known as a 'point mutation'). If this variant is common in a population (usually if its frequency is >1%), it is referred to as a 'single nucleotide polymorphism' (SNP). There are thought to be about 9-10 million common SNPs with a minor allele frequency (MAF) ≥ 0.05 in the human genome [The International HapMap Consortium, 2007]. SNPs can be distinguished into transitions (Ts) and transversions (Tv). The first comprises all changes that occur either between two different purines (adenine, A, and guanine, G) or between two different pyrimidines (cytosine, C, and thiamine, T), while the latter decode changes between a purine and a pyrimidine. Transitions account for about two-thirds of all point mutations in the human genome [Zhao and Boerwinkle, 2002], an excess over transversions that is most likely caused by the fact that they only need one (instead of two) rare tautomeric form(s) of a nucleotide to generate a stable purine-pyrimidine (purine-purine) mismatch which can escape DNA repair mechanisms [Topal and Fresco, 1976]. If the mutation occurs in a protein-coding region, it can be further classified according to its effect on the protein: mutations that, due to a redundancy in the genetic code (there are three codons specifying a stop codon and 61 codons specifying one of 20 amino acids), do not alter the amino acid sequence are termed 'synonymous', while amino acid altering mutations are referred to as 'non-synonymous'. The latter can be divided into 'missense mutations' which change the previously encoded amino acid into a different one (see Figure 1.1(a)) and 'nonsense mutations' which change a sense codon into a premature stop codon that ends the translation leading to the production of a truncated protein. Point mutations are known to cause a number of human diseases (e.g. sickle cell disease [Thein, 2011] or achondroplasia, the most common type of short-limbed dwarfism [Orioli et al., 1995; Waller et al., 2008]).

Insertions/Deletions: Insertions/deletions (indels), the second most abundant type of mutation in the human genome [Mullaney et al., 2010], add/remove extra nucleotides to/from DNA sequences (see Figures 1.1(b) and 1.1(c)). They range in size from single nucleotides to segments containing thousands of nucleotides [Bhangale et al., 2005; Mills et al., 2006; Weber et al., 2002]. To date, several million indels have been identified in human populations causing a substantial amount of genetic variation. If an indel occurs within a protein-coding region of the genome, it can alter the pattern of DNA sequence translation (e.g. it can cause 'frameshift mutations', if the change in sequence length is not a multiple of three, as depicted

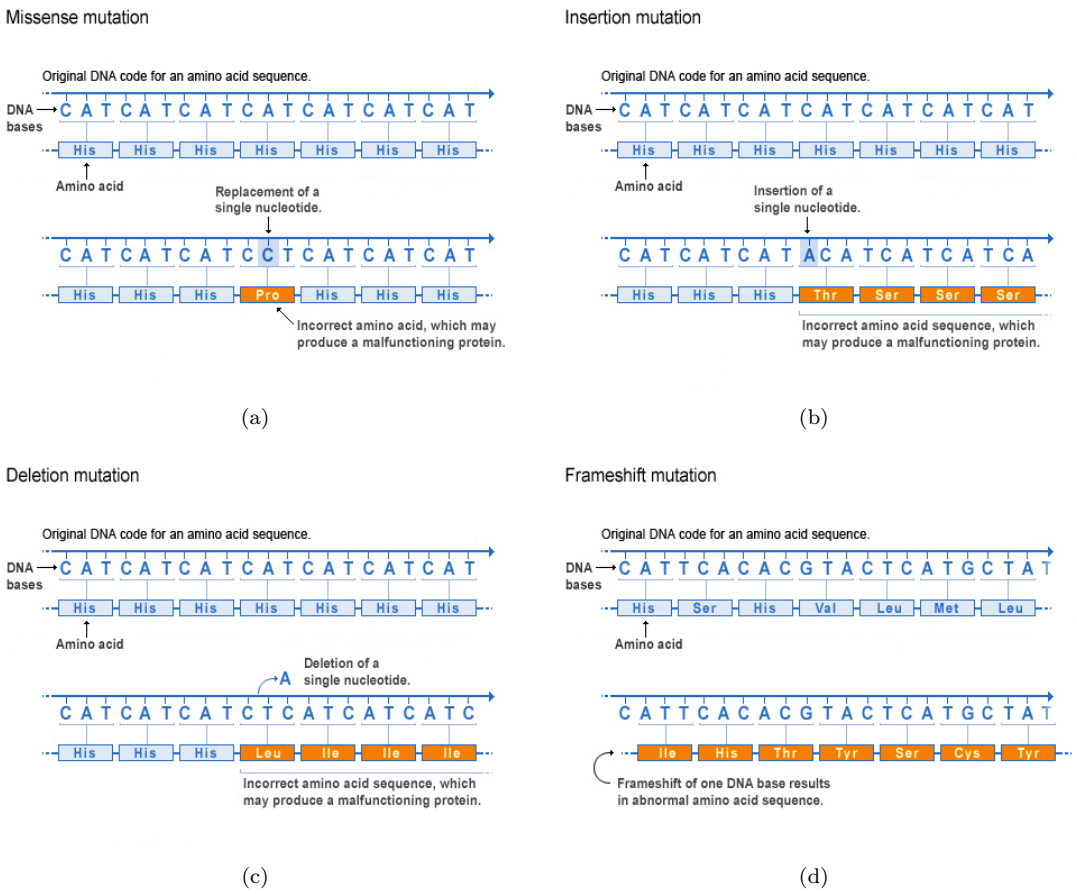


Figure 1.1: (a) A missense mutation is a point mutation which changes the previously encoded amino acid into a different one (in contrast to a synonymous mutation that does not alter the amino acid sequence due to a redundancy in the genetic code) (b) Insertions add extra nucleotides to the DNA sequence. (c) Deletions remove extra nucleotides from the DNA sequence. (d) An insertion/deletion can cause a frameshift mutation if the change in sequence length is not a multiple of three. These figures were taken from U.S. National Library of Medicine.

in Figure 1.1(d), or it can cause translations to end prematurely by introducing a new stop codon). A number of indels are known to map to functionally important sites within human genes, causing a variety of inherited human disorders (e.g. cystic fibrosis which is frequently caused by an indel polymorphism within the *CFTR* gene [Collins et al., 1987]).

Duplications: Duplications occur when a part of a chromosome is copied too many times leading to extra copies of genetic material from the duplicated segment (see Figure 1.2). Duplications of a DNA segment longer than 1kb (also referred to as 'segmental duplications') account for about 3.5-5.2% of the entire human genome sequence [Bailey et al., 2002; Cheung et al., 2003]. (The largest duplicated interval in human spans 1.5 Mb on chromosome Y [Sainz et al., 2006].) They can not only lead to a number of diseases in humans (e.g. anxiety disorders or Parkinson's disease [Emanuel and Shaikh, 2001; Gratacos et al., 2001; Mazzarella and Schlessinger, 1998; Singleton et al., 2003; Stankiewicz and Lupski, 2002]), but they have also played an important role in shaping the evolution of the human genome [Bailey et al.,

2002; Cheung et al., 2003; Hattori et al., 2000; International Human Genome Sequencing Consortium, 2001; Samonte and Eichler, 2002; Zhang et al., 2005]. In addition, the human genome consists to 5-10% of 1-5bp short sequences that are repeated several times (so called 'microsatellites' or 'short tandem repeats', STRs) which are duplicated one or more times [Wright, 2008]. The highly variable number of repeated elements is usually caused by slippage mutations during DNA replication. Therefore, recurrent mutations are much more common for STRs than for other types of mutation (such as SNPs). Larger 5-64bp repeated sequences (referred to as 'minisatellites' or 'variable number tandem repeats', VNTRs), spanning sequence lengths of 10-400kb, can also be found in the genome of most mammals. They can cause genomic instability due to unequal pairing of homologous chromosomes during meiosis [Wright, 2008]. If STR-duplications occur in functionally important regions of the genome, they can give rise to diseases, usually by generating large copy numbers of the repeat, which interfere with the function, or expression of the gene (e.g. fragile X syndrome, an inherited dominant mental retardation disorder [Richards and Sutherland, 1992; Webb, 1989], which is characterised by an expansion of a trinucleotide (CGG)n repeat located in the *FMR1* gene [Verkerk et al., 1991]).

Inversions: An inversion is a rearrangement that resulted from a loop-structure of a chromosomal region which was broken and repaired in the reverse orientation (see Figure 1.2). In this process, genetic material might be lost. If an inversion includes the centromeric region, it is termed a 'pericentric inversion', otherwise it is called a 'paracentric inversion'. In contrast to large inversions, which are rare in the genome, small inversions are more common. Some inversions are known to give rise to human disease (e.g. an inversion in the *F8* gene causes haemophilia A [Wright, 2008]).

Translocation: A transfer of a DNA segment from one chromosome to a non-homologous chromosome is called a 'translocation'. The transfer often happens in a reciprocal fashion (the two non-homologues chromosomes swap segments) which is termed 'balanced translocation', in contrast to an 'unbalanced translocation' in which genetic material is either gained or lost (see Figure 1.2). Translocations can (i) generate hybrid genes or destroy gene functions (e.g. if the breaks occur within genes) and (ii) they can alter gene expression (by bringing the gene under the influence of different promotors and enhancers), which can lead to human disease (e.g. Burkitt's lymphoma, a rare, aggressive B-cell lymphoma, results from chromosomal translocations that involve the *MYC* gene, and either the lambda or the kappa light chain immunoglobulin genes [Boerma et al., 2009; Kirsch et al., 1982; Lenoir et al., 1982]).

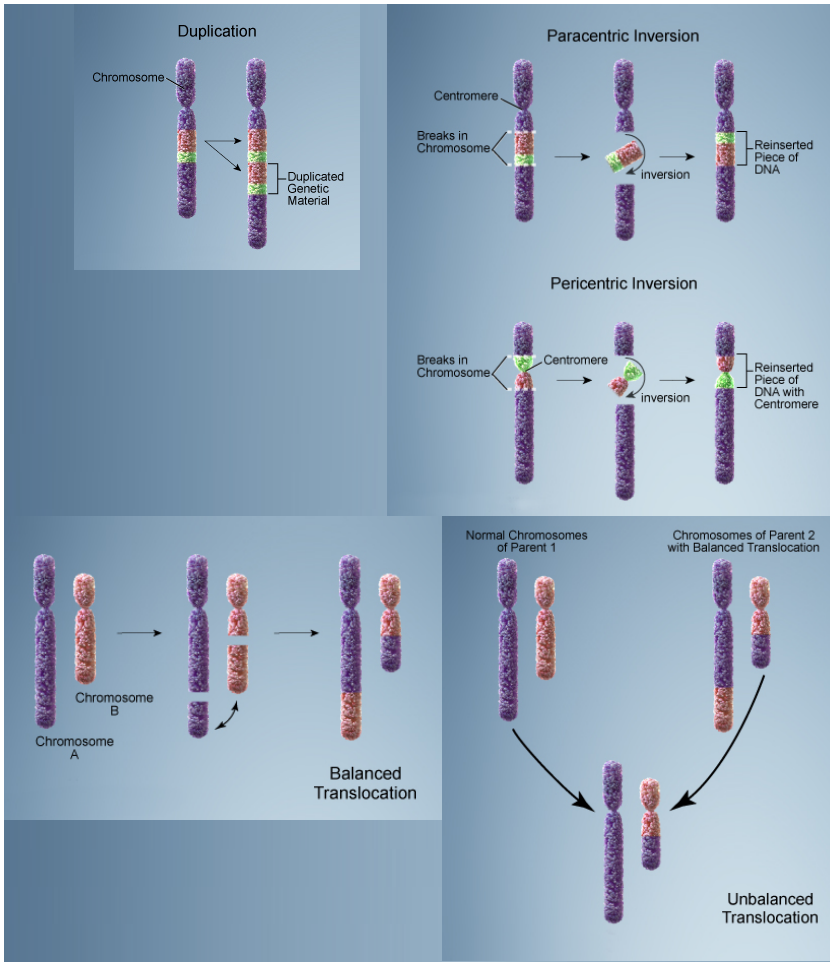


Figure 1.2: Duplications occur when part of a chromosome is copied too many times leading to extra copies of genetic material from the duplicated segment. An inversion is a rearrangement that resulted from a loop-structure of a chromosomal region which was broken and repaired in the reverse orientation. If an inversion includes the centromeric region, it is termed a pericentric inversion, otherwise it is called a paracentric inversion. A transfer of a DNA segment from one chromosome to a non-homologous chromosome is called a translocation. The transfer often happens in a reciprocal fashion (the two non-homologues chromosomes swap segments) which is termed balanced translocation, in contrast to an unbalanced translocation in which genetic material is either gained or lost. These figures were taken from U.S. National Library of Medicine.

The above mentioned mutations can occur in all cell types. Although somatic mutations, arising in non-germ line cells, are very interesting in their own right as they can influence the fitness of an individual (e.g. some of them might cause cancer), this thesis focuses on germ line mutations, as only germ line mutations are heritable in mammals and thus, contribute to the genetic variation of future generations. (However, it should be noted that in many other organisms, such as plants and fungi, reproductive structures are continuously produced from somatic cells, for example during flower development, and thus, somatic mutations can be transmitted to the gametes in these organisms [Johnston, 2001]). In particular, the thesis concentrates on point mutations as they are the most abundant genetic variants in the human genome and their mutational process is probably the best understood [Ellegren et al., 2003]. Point mutations are a very interesting subject of study, as their occurrence is not entirely random: in primates, mutation rates are known to vary

with gender, with genomic location (high rates have been detected for sequences containing for instance CpG-dinucleotides, repeats, or a high purine content [Gojobori et al., 1982; Krawczak et al., 1998; Li et al., 1984; Siepel and Haussler, 2004]), as well as with other genomic features such as recombination rate [Hellmann et al., 2003], and their direction is usually biased towards the following order C/G>T/A, T/A>C/G, C/G>A/T, G/C>C/G, A/T>C/G, and T/A>A/T [Blake et al., 1992; Gojobori et al., 1982; Hess et al., 1994; Hwang and Green, 2004; Li et al., 1984]. Unfortunately, the factors that influence this mutation rate variation across the genome are not very well understood.

1.3 Mutation Rate Variation

Variation in mutation rates has not only been observed between species, but also between different individuals of the same species [Drake, 1992; Duret, 2009; The 1000 Genomes Project Consortium, 2011]. Mutation rates are known to be subject to a multitude of evolutionary pressures causing them to vary between different regions of a single genome [Wolfe et al., 1989]. This variation occurs on various physical scales: on the large scale, mutation rates vary between chromosomes [Chen et al., 2001; Ebersberger et al., 2002; Lercher et al., 2001; Li et al., 2002; The Mouse Genome Sequencing Consortium, 2002] as well as regionally within chromosomes (at both kilobase scales and megabase scales) [Lercher et al., 2001; Matassi et al., 1999; Silva and Kondrashov, 2002; Smith et al., 2002; Williams and Hurst, 2000], while the largest variation can be detected on the smallest scale, at individual sites of the genome [Blake et al., 1992; Gojobori et al., 1982; Hwang and Green, 2004; Zhao and Boerwinkle, 2002]. Although different internal genetic factors have been identified that influence spontaneous mutation rates at different genomic scales (such as transcription, chromatin packaging, biased gene conversion, recombination, as well as sequence characteristics which decrease or increase local mutation rates in the genome [Hess et al., 1994; Krawczak et al., 1998; Siepel and Haussler, 2004; Zhang et al., 2007; Zhao and Boerwinkle, 2002; Zhao and Zhang, 2006]), more complex processes are likely involved, leading to the observed variation in mutation rates across mammalian genomes.

1.3.1 Mutation Rate Variation at the Single Nucleotide Level

Systematic differences of mutation rates at the single nucleotide level are frequently influenced by neighbouring nucleotides (a phenomenon referred to as 'sequence-context effect') [Blake et al., 1992; Gojobori et al., 1982; Hwang and Green, 2004; Zhao and Boerwinkle, 2002]. The most well known and strongest sequence-context effect is the about one order of magnitude higher

frequency of C to T (C>T) transitions at hypermutable CpG-dinucleotides (the 'p' denotes the phosphodiester bond between the adjacent C and G nucleotides on the same DNA strand - thus, this DNA sequence-context effect is also known as 'CpG-effect') accounting for about a quarter of all SNPs in the human genome [Chimpanzee Sequencing and Analysis Consortium, 2005; Fryxell and Moon, 2005; Hwang and Green, 2004; Miller and Kwok, 2001; Rideout et al., 1990; Sved and Bird, 1990; Zavolan and Kepler, 2001; Zhao and Boerwinkle, 2002]. The reason for the elevated mutability of CpG-sites is that many cytosine residues in mammals are methylated at their 5' position of the pyrimidine cycle. Methylcytosine is not very stable and thus, it exhibits a strong tendency to spontaneously deaminate to either TpG or CpA, both of which are less efficiently or incorrectly repaired by the DNA repair machinery [Cooper and Youssoufian, 1988; Coulondre et al., 1978; Ehrlich and Wang, 1981; Holliday and Grigg, 1993]. The CpG-effect, together with a bias in transition-transversion rates, gives rise to particularly high rates of C>T and G>A mutations in the genome [Krawczak et al., 1998; Nachman and Crowell, 2000]. Elevated mutation rates of transversions at CpG-dinucleotides have also been observed, most likely caused by oxidative damage of CpG-dinucleotides (as guanosines are more susceptible to transversions when exposed to oxygen radicals [Agnez-Lima et al., 2001]) and an error-prone repair of altered guanines after methylation and deamination to cytosines [Blake et al., 1992; Cooper and Krawczak, 1990; Hwang and Green, 2004; Nachman and Crowell, 2000].

High mutation rates at CpG-dinucleotides are not uniformly distributed across the genome [Mugal and Ellegren, 2011]. In fact, an analysis of introns in human genes could show that the genomic substitution rate at CpG-sites varies not only in correlation with the level of DNA methylation, but also with (i) non-CpG divergence, indicating that regions which exhibit a higher general mutation rate (e.g. due to a higher rate of replication errors in mitotic cell divisions of the germ line [Kim et al., 2006]) are also more likely to experience higher mutation rates at CpG-sites, a fact that suggests that CpG-mutation rates also depend on genomic parameters common for the general mutation rate variation in the genome, (ii) intronic GC-content as a high GC-content stabilises double-stranded DNA (G:C bonds are stronger than A:T bonds, thus, they increase the melting temperature of the DNA), thereby reducing the CpG mutation rate as methylcytosine deamination occurs >140 times more often on single-stranded than on double-stranded DNA, a context-dependence that exponentially decays over ∼1,500bp to either side of the CpG [Duret and Arndt, 2008; Elango et al., 2008; Frederico et al., 1990, 1993; Fryxell and Moon, 2005; Fryxell and Zuckerkandl, 2000; Tyekucheva et al., 2008], and (iii) germ line transcription levels which might reflect a higher accessibility of DNA repair enzymes during active transcription [Gilbert et al., 2004; Mugal and Ellegren, 2011; Prendergast et al., 2007].

While CpG-dinucleotides exhibit the strongest effect on mutation rates (transition mutation rates are elevated by ∼15-fold relative to the average mutation rate in mammals and ∼30-fold in great apes [Hwang and Green, 2004; Keightley et al., 2011; Siepel and Haussler, 2004]), other adjacent nucleotides have also been shown to elevate mutation rates by two- to threefold but their patterns are usually more complex and less well understood [Blake et al., 1992; Hess et al., 1994; Hwang and Green, 2004; Siepel and Haussler, 2004; Zhao and Boerwinkle, 2002]. This heterogeneity in nucleotide composition can extent as many as 200bp away, possibly reflecting an overall nucleotide bias of the region, but the mechanisms underlying this long-distance bias are still largely unknown [Elango et al., 2008; Krawczak et al., 1998; Silva and Kondrashov, 2002; Smith et al., 2003; Zhao and Boerwinkle, 2002]). However, the strongest effects have been detected for short ranges of up to two or three nucleotides each direction as nucleotides further away seem to have statistically significant but increasingly small effects [Nevarez et al., 2010; Zavolan and Kepler, 2001; Zhao and Boerwinkle, 2002].

1.3.2 Mutation Rate Variation within Chromosomes

Regional variation of genomic mutation rates occurs at both small scales of several kilobases as well as large, megabase, scales [Chimpanzee Sequencing and Analysis Consortium, 2005; Elango et al., 2008; Gaffney and Keightley, 2005; Li et al., 2002; McVean, 2000; Silva and Kondrashov, 2002], with the greatest difference in mutation rate being observed between 2kb blocks in hominoids [Spencer et al., 2006]. Although many factors are thought to correlate with mutation rate variation between different genomic regions, the exact mechanisms as well as their interplay are currently not fully understood.

1.3.2.1 Small-scale Variation

CpG-islands exhibit a higher GC-content and a higher rate of observed versus expected CpG-content compared to the rest of the genome, both of which are known to influence the underlying mutation rate [Arndt et al., 2005; Bird, 1986; Chen et al., 2010; Chimpanzee Sequencing and Analysis Consortium, 2005; Cohen et al., 2011; Elango et al., 2008; Fryxell and Moon, 2005; Hardison et al., 2003; Hellmann et al., 2005; Hwang and Green, 2004; Matassi et al., 1999; Miller and Kwok, 2001; Polak and Arndt, 2008; Rideout et al., 1990; Sved and Bird, 1990; Tyekucheva et al., 2008; Wolfe et al., 1989; Zavolan and Kepler, 2001; Zhao and Boerwinkle, 2002]. Mutation rates at CpG-sites are usually lower within CpG-islands than in the rest of the genome [Cohen et al., 2011; Polak and Arndt, 2008]. At least two possible facts could account for this observation: (i) CpGs within CpG-islands are usually unmethylated and thus, less prone to substitution and

(ii) CpG-islands can play a role in gene regulation, leading to more stable CpG-sites when they are under selection [Hodgkinson and Eyre-Walker, 2011]. Most studies conducted so far promote the first hypothesis as it could be shown (i) that the frequency of other substitutions is not influenced by their location within a CpG-island, indicating that CpG-dinucleotides are subject to lower mutation rates rather than selective pressures [Cohen et al., 2011; Polak and Arndt, 2008], and (ii) that mutation rates of methylated CpGs are decreased by two- to threefold in regions of high GC-content compared to other regions in the genome, possibly due to a lower melting of the DNA duplex as methylcytosine deamination preferentially occurs on single-stranded DNA [Elango et al., 2008; Fryxell and Moon, 2005].

Mutations might cause other mutations at nearby sites (e.g. by affecting the surrounding DNA structure or by affecting a gene involved in DNA replication or repair, leading to mutations in other loci) [Hodgkinson and Eyre-Walker, 2011]. Increased point mutation rates have been observed in the ∼150bp regions flanking indels in humans and other eukaryotes (with the largest effects up to ∼50bp from the indel) [McDonald et al., 2011; Tian et al., 2008]. This effect is stronger the longer the indel is, but even short 1-5bp indels show elevated rates of point mutation in their neighbourhood [Hodgkinson and Eyre-Walker, 2011]. Currently, it is unclear what causes this association but possible explanations are (i) that point mutations and indels occur simultaneously through a single mutation process, (ii) that point mutations and indels occur together within parts of the genome which are generally highly mutable (such as repetitive DNA which exhibits elevated rates of indel mutations and which shows increased rates of error-prone replication due to its tendency to interrupt DNA polymerase [McDonald et al., 2011]), or (iii) that indels cause problems with chromosome pairing during meiosis in individuals heterozygous for the indel, leading to point mutations that segregate with the indel in a population [Tian et al., 2008].

Transcription is thought to affect mutation rates because it requires DNA to be unwound from the nucleosomes, thereby exposing single-stranded DNA to mutagens [Boulikas, 1992; Datta and Jinks-Robertson, 1995; Lippert et al., 1998; Mugal et al., 2009]. Although previous studies observed some effects of transcription on the patterns of mutation (such as an excess of A>G relative to T>C transitions as well as an excess of C>T relative to G>A transitions on the coding, non-template strand; most likely a consequence of transcription-coupled repair [Green et al., 2003]), they could also show that the influence on the overall rate of germ line mutations is very small [Green et al., 2003; Polak and Arndt, 2008; Webster et al., 2004].

Chromatin packaging also appears to affect mutation rates at the regional scale [Prendergast et al., 2007; Sasaki et al., 2009; Teytelman et al., 2008]: open chromatin has been shown to have lower mutation rates than other regions in the genome (e.g. DNase I hypersensitive sites have

lower substitution rates than regions flanking the hypersensitive site [Ying et al., 2010]), an effect that is thought to be caused by different efficiencies of the repair enzymes in open and closed chromatin structures [Filipski, 1988; Gaffney and Keightley, 2005; Matassi et al., 1999]. However, the observed effect might be caused by selection as hypersensitive sites might contain regulatory elements [Boyle et al., 2011].

The two DNA strands are not equal with respect to mutation rates. The two main reasons for strand-asymmetric bias are (i) transcription (e.g. the coding strand frequently shows an excess of G and T over A and C in genic regions as A>G transitions are more common than T>C transitions in transcribed parts of the human genome while there is no difference of these rates for non-transcribed regions [Green et al., 2003]) and (ii) replication (i.e. asymmetric mutation bias with respect to the origin of replication as the leading strand is continuously synthesised while the lagging strand is synthesised in pieces [Lobry, 1996]).

Although both the strength and the direction of the relationship between mutation rates and genomic GC-content is currently unclear (some studies have found a positive correlation between the two [Matassi et al., 1999; Smith et al., 2003] while others have found a negative [Filipski, 1988; Hellmann et al., 2005] or even more complex one [Eory et al., 2010; Wolfe and Sharp, 1993; Wolfe et al., 1989]), GC-content is thought to cause little variation in mutation rate (current estimates are around 10% [Smith et al., 2002]).

1.3.2.2 Large-scale Variation

While mutation rates at CpG-dinucleotides mainly depend on the methylation density [Chen et al., 2010] and the regional GC-content [Elango et al., 2008; Fryxell and Moon, 2005; Tyekucheva et al., 2008], recent work by Chen et al. suggested that the most influential factors on mutational mechanisms at non-CpG-sites are (i) the replication time [Chen et al., 2010; Pink and Hurst, 2010; Stamatoyannopoulos et al., 2009] (late replicating DNA has been shown to have ∼20-30% higher mutation rates in hominoids [Chen et al., 2010; Stamatoyannopoulos et al., 2009], most likely caused by either a greater exposure of DNA to mutagens as it stays single-stranded for longer during the S-phase [Stamatoyannopoulos et al., 2009] or by a decreased mismatch repair activity during the S-phase [Chen et al., 2010; Holmquist and Filipski, 1994]) and (ii) the distance from the telomere [Chen et al., 2010; Chimpanzee Sequencing and Analysis Consortium, 2005; Hellmann et al., 2005; Tyekucheva et al., 2008], followed by (iii) the male recombination rate [Duret and Arndt, 2008; Hellmann et al., 2003, 2005; Lercher and Hurst, 2002; Tyekucheva et al., 2008]. During recombination, biased gene conversion can replace adenine or thiamine nucleotides with guanine or cytosine mutations during the repair of the heteroduplex DNA, leading to a higher

GC-content in regions of high recombination rates [Duret and Arndt, 2008]. A higher local GC-content might reduce regional mutation rates by stabilising double-stranded DNA [Arndt et al., 2005; Elango et al., 2008; Fryxell and Moon, 2005; Hardison et al., 2003; Hellmann et al., 2005; Matassi et al., 1999; Wolfe et al., 1989]. Other, possibly less important, factors affecting mutation rates at non-CpG-sites are gene density [Arndt et al., 2005; Duret and Arndt, 2008; Hardison et al., 2003; Lercher and Hurst, 2002], nucleosome position [Washietl et al., 2008], nuclear lamina binding sites [Hardison et al., 2003], simple repeat content (such as microsatellite or minisatellite repeats which are prone to DNA replication errors) [Hellmann et al., 2005; Jeffreys et al., 1988], and different repair efficiencies [Hawk et al., 2005].

In humans, merely ∼50% of the mutation rate variance observed at the 1Mb level can be explained by the above mentioned factors, as each individual one contributes only a little to the overall variance, and most of them are strongly interrelated with each other [Chen et al., 2010; Hellmann et al., 2005; Tyekucheva et al., 2008]. Thus, other, currently undiscovered, highly complex patterns and processes might also influence regional large-scale variation in mutation rates.

1.3.3 Mutation Rate Variation between Chromosomes

In many species, males exhibit higher rates of point mutations than females [Ellegren and Fridolfsson, 1997; Goetting-Minesky and Makova, 2006; Haldane, 1947; Wilson Sayres and Makova, 2011], a fact that has often been attributed to the male's requirement of a greater number of germ line cell divisions before meiosis per generation during spermatogenesis (compared to oogenesis), in which more mutations are caused by DNA replication errors [Haldane, 1947]. However, several studies suggest that the origin of sex-specific differences in mutation rate is most likely more complex [Chang and Li, 1995; Li et al., 1996; McVean and Hurst, 1997; Miyata et al., 1987] and, as a result, other hypothesis have also been proposed to explain a male-to-female bias in mutation rates such as differences in DNA repair associated with methylation [Haldane, 1947; McVean, 2000; Taylor et al., 2006] as regions on chromosome X have been shown to have a male-biased mutation rate due to differences in germ line methylation patterns [Ketterling et al., 1993] or the lack of particular repair enzymes in mature sperm [Boulikas, 1992]. On the other hand, mutation rates seem to be lower on hemizygous sex chromosomes (about two-fold lower on the mammalian chromosome X) due to strong evolutionary pressures such as negative selection which acts to reduce the occurrence of deleterious recessive mutations as they are effectively dominant in the heterogametic sex and usually effectively hemizygous in the homogametic sex due to dosage compensation [McVean and Hurst, 1997].

As a result of this male-driven evolution, chromosomal mutation rates depend on the way chromosomes are transmitted through the male and female germ lines: mutation rates are highest on chromosome Y as it is exclusively transmitted through the male germ line and lowest on chromosome X which only spends a third of its time in the male sex [Chimpanzee Sequencing and Analysis Consortium, 2005; Ebersberger et al., 2002; Gaffney and Keightley, 2005; Goetting-Minesky and Makova, 2006; Lercher and Hurst, 2002; Malcom et al., 2003; McVean and Hurst, 1997; Wolfe and Sharp, 1993]. Autosomal mutation rates usually lie in between the two as an autosome spends half its time in males and half its time in females. In primates, the male-to-female mutation rate bias is estimated to be ∼2-7 [Bohossian et al., 2000; Chimpanzee Sequencing and Analysis Consortium, 2005; Goetting-Minesky and Makova, 2006; Makova and Li, 2002; Taylor et al., 2006], making sex chromosomes of special interest for the study of mutation rate variation between chromosomes [Ellegren et al., 2003]. The estimates vary greatly depending on the data set used for the analysis as mutation rates show a strong regional variation (e.g. mutation rates have been shown to vary more than 16-fold using different genes [Goetting-Minesky and Makova, 2006]) and thus, sampling bias can be introduced if the data set is limited to a small genomic region. If the rate is predicted from a comparison of human and chimpanzee sequences, ancestral polymorphisms can also pose a problem as chromosome Y has a low diversity [Li et al., 2002]. As a result, predictions of the male-to-female mutation rate are more reliable if either ancestral polymorphisms are accounted for [Ebersberger et al., 2002] or if data between more distantly related primate species is used as ancestral polymorphisms will have a smaller influence [Makova and Li, 2002]. Nevertheless, the predictions of the male-to-female mutation bias are roughly consistent with the ratio of the number of male to female germ line cell divisions in humans [Crow, 2000; Li et al., 2002; Nachman and Crowell, 2000; Shimmin et al., 1993].

Consistent with the fact that most non-CpG mutations are created during DNA replication, mutation rate variation between autosomes and sex chromosomes has been found to be stronger for non-CpG mutations than for CpG-mutations [Taylor et al., 2006]: at non-CpG-sites, the mutation rate on chromosome Y is at least 50% higher than the one on the autosomes, which in turn have about 30% higher mutation rates than chromosome X [Chimpanzee Sequencing and Analysis Consortium, 2005; Ebersberger et al., 2002; Ellegren, 2007; Li et al., 2002]. In contrast, the male-to-female bias in the rate of CpG-mutations has been found to be weak (∼2-3 in humans [Chimpanzee Sequencing and Analysis Consortium, 2005; Taylor et al., 2006]). This observation is consistent with the assumption that the majority of mutations at CpG-sites are caused by methylation-dependent deamination rather than DNA replication errors [Chimpanzee Sequencing and Analysis Consortium, 2005; Duret, 2009; Hwang and Green, 2004; Taylor et al., 2006; Xing

et al., 2004]. Although there is also the possibility that they have been introduced by replication through a mismatch at the CpG-dinucleotide or by a depletion of CpG-sites on chromosome X due to a higher degree of methylation causing a lower rate of transitions at CpG-sites [Krawczak et al., 1998; McVean, 2000].

Mutation rates can not only vary between autosomes and sex chromosomes, but also among autosomes [Chimpanzee Sequencing and Analysis Consortium, 2005; Ebersberger et al., 2002; Gaffney and Keightley, 2005; Lercher and Hurst, 2002; Malcom et al., 2003]. The variance of mutation rates between autosomes is ∼10-fold smaller than between 1Mb regions of the same chromosome [Wolfe and Sharp, 1993]. But while the processes causing variation between the sex chromosomes and autosomes are fairly well characterised, the ones resulting in mutation rate variation among autosomes are still not well understood [Gaffney and Keightley, 2005; Lercher et al., 2001]. It is unclear whether differences on an autosomal level are entirely driven by regional variation (as described above) or whether different, currently unknown, forces play an additional role. Previous studies in rodents could show that different replication times on the autosomes can account for ∼4.5% of the difference in mutation rate [Pink and Hurst, 2010] while chromosomal rearrangements can account for as much as 56% of the variation in mutation rate at a chromosomal level [Pink et al., 2009]. The latter process might be driven by recombination as chromosomes with higher recombination rates might also have higher rates of mutation and thus, also contain more rearrangement events [Pink et al., 2009].

1.4 Measurements of Polymorphism

The majority of this thesis focuses on mutation rate variation in humans and chimpanzees. As the rate of spontaneous germ line mutations is very low in most organisms (e.g. in many multicellular eukaryotes, it is in the order of $\sim 10^{-8}$ per nucleotide per generation [Crow, 1993; Drake et al., 1998; Haldane, 1935; Kondrashov, 2003; Kondrashov and Crow, 1993; Lynch, 2010; Nachman and Crowell, 2000; Roach et al., 2010]), their direct observation is very difficult and requires careful and laborious methods [Kondrashov and Kondrashov, 2010; Roach et al., 2010; Xue et al., 2009]. These techniques are complicated by the fact that mutation rates are influenced by both the genetic background (such as methylation state or sequence-context effects) [Holliday and Grigg, 1993; Rogozin and Pavlov, 2003] and many mutagenic environmental and physiological factors (such as chemicals, radiation, age, and sex) [Kondrashov and Kondrashov, 2010], as well as by the fact that mutation rates can be subject to selection and evolution [Avila et al., 2006; Loewe et al., 2003; Lynch, 2008; Sniegowski et al., 1997]. Nevertheless, most of the current knowledge was

obtained from these indirect, complex, and imprecise methodological approaches. (An overview of the different methods is given in Kondrashov and Kondrashov [2010]). However, recent advances in NGS, permitting to sequence whole genomes at lower costs, make it possible to study patterns of mutation rate variation across genomes directly by analysing DNA polymorphism data based on sequence variation within a species or by comparing orthologous sequences from different species, thereby avoiding some of the issues associated with the use of indirect methods.

Genetic variability in a population can be measured using heterozygosity (the frequency of heterozygotes h at a particular locus in a population) [Nei, 1987]. For a random mating population, the heterozygosity of a n-allelic locus with allele frequencies p_i is expected to be $h = 1 - \sum_{i=1}^{n} p_i^2$, as the genotype frequencies are expected to follow Hardy-Weinberg's law (e.g. at a two-allelic locus with allele frequencies p_1 and p_2, the homozygote frequency is expected to be $p_1^2 + p_2^2$ and the heterozygote frequency is expected to be $2\ p_1\ p_2 = 1 - p_1^2 + p_2^2$). The average heterozygosity H can be calculated as the heterozygosity tends to vary between different loci.

Although the heterozygosity is commonly used to measure polymorphisms in proteins, it is not very well suited to study DNA polymorphisms. The reason is that, due to the high variability of DNA, two randomly chosen sequences are often different from each other (which causes the heterozygosity to be close to 1) and thus, it is difficult to differentiate between different loci. As a result, the statistic π, defined as the number of nucleotide differences per site between two randomly chosen sequences, is used to estimate nucleotide diversity [Nei and Li, 1979]. It is calculated as

$$\pi = \left(\frac{n}{n-1}\right) \sum_{s} H_s,$$

with

$$H_s = \frac{2a_s}{n}\left(1 - \frac{a_s}{n}\right),$$

where a_s is the number of segregating alleles and n is the total number of alleles at a site s.

Another measure of genetic variability is Watterson's estimate Θ which is based on the number of variable positions, corrected by the number of individuals in a sample [Watterson, 1975]:

$$\Theta = \frac{s_n}{a_n},$$

with

$$a_n = \sum_{i=1}^{n-1} \frac{1}{i},$$

where s_n is the number of segregating sites.

In a model in which sites evolve neutrally, π estimates the population genetics parameter $\Theta = 4N_e\mu$, with a mutation rate μ and a constant-sized population of effective size N_e [Sachidanandam et al., 2001]. As a consequence, estimates of nucleotide diversity can be used to estimate the mutation rate at the sequence level [Deng et al., 2006; Deng and Lynch, 1996; Kondrashov, 1998; Li and Deng, 2005; Messer, 2009].

1.5 Linkage Disequilibrium

The inference of mutation rates from nucleotide diversity estimates (as described in the previous section) suffers from several limitations caused by the fact that the method relies on neutrally evolving sequences in the genome (as their sequence divergence is proportional to their mutation rate [Kimura, 1983]). To measure the base-line mutation rate of genomic loci, nucleotide variations at supposedly silent sites of the genome (such as non-coding regions or synonymous substitutions) are studied. However, some of these sites might be under evolutionary constraints (e.g. non-coding regions can be part of regulatory sequences or unidentified protein-coding genes while synonymous changes can create new splice sites or destroy existing ones by turning an exon into an intron and vice versa; in addition to which many organisms also show some sort of codon bias [Dermitzakis et al., 2002]). Hence, the observed differences could be caused by different acting evolutionary forces rather than mutational processes, a scenario which will be reflected in the estimations of mutation rates from DNA polymorphism data.

One example of an acting evolutionary force that can influence the estimation of mutation rates from DNA sequencing data is homologous meiotic recombination, known to vary between individuals, populations, and species [Smukowski and Noor, 2011]. Previous studies could show that different recombination rates across the genome can explain as much as 50% of the genomic variation observed in humans [Brooks and Marks, 1986; Coop and Przeworski, 2007; Nachman, 2002; Smukowski and Noor, 2011] and thus, patterns of linkage disequilibrium (LD) are studied to estimate recombination rates.

In a panmictic population, different SNPs are expected to be associated with each other at random according to their allele frequencies assuming that those SNPs are either unlinked (e.g. because they are positioned on different chromosomes or because they are far away from each other on the same chromosome) or linked but sufficient time has elapsed for recombination to separate and randomise them onto different genetic backgrounds. These SNPs are in a so called 'linkage equilibrium'. Consequently, the frequency of four haplotypes AB, Ab, aB, and ab at any pair of segregating loci, of which one exhibits the alleles A and a at the frequencies p_A and p_a and the

other one exhibits the alleles B and b at the frequencies p_B and p_b, is given by the product of the corresponding allele frequencies ($p_A p_B$, $p_A p_b$, $p_a p_B$, and $p_a p_b$, respectively).

In general, a non-random association between alleles at adjacent loci (referred to as 'linkage disequilibrium' [Lewontin and Kojima, 1960], see Figure 1.3) can frequently be observed as (i) recombination might not have had sufficient time to randomise the haplotypes in a population [Ardlie et al., 2002; Jeffreys et al., 2001], (ii) haplotypes in a population might not have been sampled proportionally due to genetic drift [Ardlie et al., 2002], (iii) the history of a population (such as inbreeding, admixture, migration, founder effects, population growth, and bottlenecks) might have influenced the patterns of LD [Pritchard and Rosenberg, 1999; Service et al., 2001; Slatkin, 2008; Wilson and Goldstein, 2000], and/or (iv) the LD structure might have been maintained by natural selection (e.g. balancing selection or epistatic effects where the fitness of one allele on one locus depends on another locus) [Abecasis et al., 2001; Cannon, 1963; Huttley et al., 1999; Peterson et al., 1995]. Over several generations, the ancestral haplotypes will be broken down by recombination. Thereby, the extent of the LD decay is proportional to the number of generations since the event that created the LD in the first place [Reich et al., 2001] and as a result, LD can be used to infer the history of common ancestry at a particular locus [Hartl and Clark, 1997; Tishkoff et al., 1996]. Nowadays, the availability of large polymorphism data sets permits the quantification of the extent of LD in a genome enabling fine-scale analysis of forces that influence genomic variation [Slatkin, 2008].

When loci are in LD, some haplotypes descended from single ancestral chromosomes occur more or less frequently than expected, leading to block structures of SNP haplotypes (combinations of SNPs on a single chromosome, in humans typically in the range of 10-100kb) [Abecasis et al., 2001; Collins et al., 1999; Dunning et al., 2000; Taillon-Miller et al., 2000]. LD can be measured using the coefficient D which quantifies the disequilibrium as the difference between the observed frequency of a two-locus haplotype, $p_{AB}p_{ab}$, and the frequency with which it would have been expected given randomly segregating alleles, $p_{Ab}p_{aB}$ [Lewontin and Kojima, 1960]. However, normalised coefficients (such as r^2 which divides D^2 by the product of all four allele frequencies, $r^2 = D^2/(p_A p_a p_B p_b)$ [Hill and Robertson, 1968]) are more commonly used. Two alleles which exhibit the same frequencies and which have not yet been separated by recombination are said to be in perfect LD ($r^2 = 1$), in which case, only two of the four possible two-locus haplotypes can be observed (see Figure 1.3(1)). In most other cases, it is difficult to assess the biological meaning of different values. Another shortcoming of r^2 is that it commonly measures LD between pairs of biallelic markers and it is unclear how to combine estimates over several loci as the obtained values are not independent [Pritchard and Przeworski, 2001].

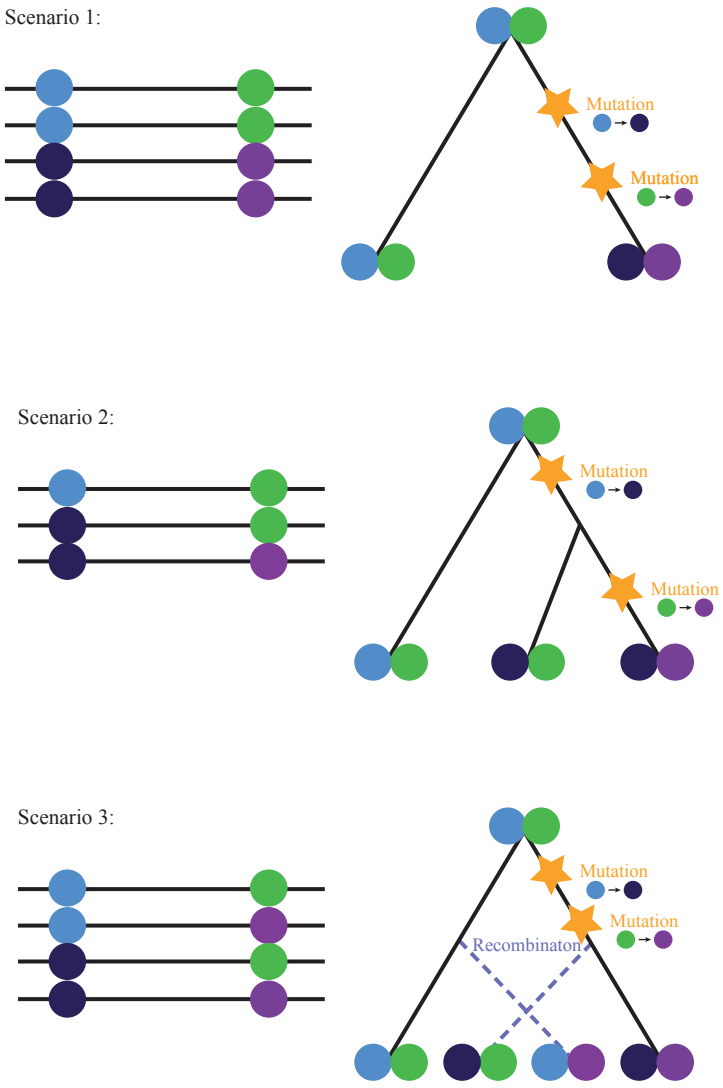


Figure 1.3: Concept of linkage. Linkage equilibrium and disequilibrium might be caused by different mutational and recombinational events as shown in the trees on the right hand side. Three different scenarios are depicted together with the allelic state of the two loci (illustrated on the left-hand side): (1) The two loci are in absolute LD as they share the same mutational history and no recombination occurred between them. (2) The two loci exhibit linkage disequilibrium. Mutations occurred on different lineages without recombination between the loci. (3) A recombination event caused the two loci to be in a linkage equilibrium.

1.6 Outline

The aim of this thesis is to obtain a better understanding of mutation rates within as well as between humans and chimpanzees using data generated by next-generation, high throughput sequencers.

Chapter 2 gives an overview of currently available technologies used to generate high throughput sequencing data as well as of the computational methods required to perform reliable genome-wide calling of polymorphisms from the data generated by them.

Some of the discussed techniques were used in Chapter 3 to investigate whether there is any evidence of a selective advantage of pathogenic *de novo* mutations in the Fibroblast Growth Factor Receptor 3 (*FGFR3*) gene in the male germ line of humans.

In addition, the methods have been used in Chapter 4 to generate a map of genome-wide sequence variation in our closest evolutionary relative, the chimpanzee. The study of chimpanzee individuals is of great scientific interest from an evolutionary point of view, as comparisons between the genomes of human and chimpanzee will help to understand the molecular basis for similarities and differences between the two species.

The generated data set was used in Chapter 5 to explore the breadth of chimpanzee genome diversity in order to shed light on whether or not the local variation in mutation rate has been conserved since the divergence of the two species, to investigate the extent to which site-specific mutation rates vary between the two hominoid lineages, and to place human nucleotide diversity into perspective with an evolutionary closely related species. This is important because nucleotide diversity does not only depend on the underlying mutation rates and acting selective pressures, but also on the demographic history of a population.

Chapter 6 focuses on the occurrence of ancestral polymorphisms shared between human and chimpanzee on a genome-wide scale. These ancestral polymorphisms do not only influence fine-scale divergence rates across the genome in very closely related species, they are also good candidates for regions under balancing selection and thus, they are a useful tool to study long-time population demographics and speciation.

Chapter 7 discusses some of the results contained in the body of the thesis and Chapter 8 suggests a number of areas for future research.

Chapter 2

Generation and Analysis of High Throughput Sequencing Data

In 1977 a British biochemist, named Frederick Sanger, provided, with his development of the chain-termination method, the fundamentals for modern DNA sequencing methods [Sanger et al., 1977] - a remarkable milestone in the history of biological research. Since then, further extensive research as well as achievements in biology, bioinformatics, and computer science, allowed the development of improved sequencing techniques that revolutionised molecular biology and biotechnology by enabling the rapid sequencing of whole organisms.

Nowadays, sequencing has entered the scientific zeitgeist. More than 500 archaea, bacteria, and eukaryote genomes have already been completely sequenced (amongst them reference genomes for human, chimpanzee, and many major model organisms), shifting the focus of most scientific studies towards resequencing analyses e.g. to identify genetic variation in different individuals of a population. Altogether, the demand for DNA sequences has never been greater than right now in the 21st century with application areas such diverse as comparative genomics, forensic, medical, evolutionary, and metagenomic studies, using, amongst others, whole-genome or targeted genome (re-)sequencing methods [Albert et al., 2007; Bashiardes et al., 2005; Dahl et al., 2007; Fredriksson et al., 2007; Hodges et al., 2007; Okou et al., 2007; Porreca et al., 2007; Wheeler et al., 2008b].

Sanger sequencing, the technology used by the Human Genome Project in the late nineties to obtain the finished-grade human genome sequence, is too labour intensive, time consuming, and costly (the initial draft cost ∼US$300 million while the final draft cost ∼US$3 billion [International Human Genome Sequencing Consortium, 2001; Service, 2006; Venter et al., 2001]) to satisfy the high demand of large-scale DNA sequencing of the current projects. As a result, the years following the finished Human Genome Project have experienced a rapid development of massively parallel, high throughput sequencing technologies (see Section 2.1) which are hundreds of times faster and nearly a thousand times cheaper than traditional Sanger sequencing, permitting the analysis of the human genome and its variation [Coombs, 2008; Rothberg and Leamon, 2008].

Data analysis is computationally challenging as the huge amount of data generated by NGS imposes several challenges on statistical analysis and bioinformatic tools, making them some of the most difficult aspects of whole genome resequencing using high throughput technologies. Reliable, comprehensive data analysis is not only complicated by the fact that every platform has specific characteristics and limitations that need to be addressed, but it is also problematic due to the size and complexity of the genome, the large amounts of sequencing data which are rapidly generated, the shorter read length, the higher read error rates, read error profiles that differ between platforms, and the computational demands required to store data and to detect variants on a genome-wide scale with the highest possible accuracy.

The identification of variation in genomes strongly relies on the precise alignment of sequenced reads to a reference genome in order to identify positions or regions at which at least one of the samples differs from the reference sequence (see Section 2.2) as well as reliable and accurate SNP calling to avoid errors by calling positional variants which are produced by misaligned reads or sequencing errors (see Section 2.3). Despite carefully chosen analytical methods, there is often a considerable amount of uncertainty associated with the results (especially for low coverage sequencing) which should be accounted for as it influences the downstream population-genomic analysis. Thus, it is important to validate a subset of the called variants using independent technologies in order to estimate the false discovery rates (FDRs) (see Section 2.4).

This chapter is a review of the current technologies and analytical methods which were used in the body of this thesis to generate and analyse high throughput sequencing data. A full discussion of the experience using these different approaches is given in Chapter 4.

2.1 High Throughput Sequencing

From 1977 to 2005, Sanger sequencing (see Appendix A.1) has been the dominant sequencing technology. Capillary-based, semi-automated Sanger sequencing was one of the most influential innovations in biological research within the last two decades [Shendure and Ji, 2008], leading to some monumental achievements such as the availability of the finished-grade human genome sequence, an initiative led by the International Human Genome Sequencing Consortium and Celera Genomics [International Human Genome Sequencing Consortium, 2001; International Human Genome Sequencing Consortium, 2004; Venter et al., 2001]. However, Sanger sequencing is too expensive and too time consuming to routinely perform high throughput sequencing of large genomes on the scale required to reach the scientific research goals of many modern projects, such as the 1000 Genomes Project. In the last few years, it became apparent that protocol optimisations may have approached exhaustion and a further reduction of costs is unlikely without a change of technol-

ogy. For these reasons, several new sequencing technologies, so called 'second-generation' or NGS technologies have been established to satisfy the demand for faster, more affordable DNA sequencing (see Appendix A.2). They are strengthened by the general progress in diverse scientific fields including nucleotide biochemistry, surface chemistry, polymerase engineering, microscopy, computation, and data storage, as well as the availability of whole genome assemblies for *Homo sapiens* and most other major model organisms, which can be used as a reference in short read mapping. It is likely that these methods will dominate high throughput sequencing for the upcoming years, but it should be noted that other, 'third-generation' (or 'next-next generation'), single-molecule sequencing methods are under ongoing development (see Appendix A.3). These technologies have the potential of even higher throughput of longer sequencing reads in a shorter time with lower overall costs and thus, they might eventually become a standard within the foreseeable future.

Worldwide, mainly four different commercial high throughput sequencing platforms are currently used in research laboratories to parallelise individual reactions and to achieve the sequencing of millions of distinct, short DNA sequences per run: the Genome Sequencer FLX System from 454 Life Science/Roche, the Genome Analyzer from Illumina, SOLiD from Applied Biosystems/Life Technologies, and HeliScope from Helios Bioscience, while Life Technologies' Ion Torrent PGM just entered the market. The following section focuses on Illumina sequencing which was used, in combination with a subset of the tools and techniques described below, (i) in a study investigating whether there is any evidence of a selective advantage of pathogenic *de novo* mutations in the Fibroblast Growth Factor Receptor 3 gene in the male germ line of humans (Chapter 3) and (ii) in a study identifying genome-wide sequence variation in Western chimpanzees (Chapter 4). The workflow as well as the main characteristics of all other above mentioned high throughput sequencing technologies are discussed in Appendix A. However, it should be kept in mind that the field of DNA sequencing is changing rapidly and thus, the information provided here only reflects a snapshot at the time of this writing.

2.1.1 Illumina: Genome Analyzer

In 2006, the second commercially sold and one of today's most widely used NGS platforms, the Genome Analyzer (GA), was released by Illumina [1]. Its concept is based on a sequencing-by-synthesis principle, similar to the one used in Sanger sequencing.

Work Flow Illumina's method starts by preparing a sequencing library by randomly fragmenting genomic DNA into segments with a length of several hundred base pairs (see Figure 2.1(a)).

[1] `http://www.illumina.com`

These fragments are end-repaired to generate 5'-phosphorylated blunt ends (Figure 2.1(b)). Both ends, the 5'-end and the 3'-end, are ligated to two different adapters (Figures 2.1(c) and (d)). Single-stranded templates, generated by melting the double-stranded library using sodium hydroxide [Kircher and Kelso, 2010], are added to the surface of the channels of a glass flow cell. This surface has two oligomers covalently attached on it which are complementary to the two specific adapters that were ligated to the sequence library and thus, enable the immobilisation of the added fragments through hybridisation (Figure 2.1(e)). Unlabelled nucleotides and enzymes (e.g. *Bst* polymerase) are added in order to amplify the single-stranded library molecules, starting from their hybridised double-stranded parts. As both forward and reverse primers are immobilised on the surface of the flow cell, strands, newly created by reverse strand synthesis, can covalently attach themselves to the surface by bending over and hybridising their adaptor sequence to the complementary oligomer, a process referred to as 'bridge polymerase chain reaction' (bridge PCR) [Fedurco et al., 2006] (Figure 2.1(f)). These bridges are denaturated using formamide, resulting in single-stranded templates anchored to the substrate. This process is repeated several times in order to generate thousands of identical copies of each molecule [Bentley et al., 2008; Fedurco et al., 2006]. As all amplicons that arise from a single template during the amplification remain immobilised and clustered to a single physical location on the array, these copies of the original sequence build clusters within a small area on the surface with a diameter of one micron or less. In this way, up to ten million single-molecule clusters can be generated per square centimetre on the flow cell surface (Figure 2.1(g)).

Before the sequence of each cluster can be determined, either the forward or the reverse strands need to be removed by selectively cleaving at base modifications of the oligomer at the flow cell surface as the presence of both strand directions might interfere with the sequencing process (e.g. through complementary base pairing), leaving only single-stranded, identically orientated sequences in each cluster [Kircher and Kelso, 2010]. A sequence primer is annealed on to the adapter sequences which acts as a starting point for the sequencing reaction (Figure 2.1(h)). Four proprietary fluorescently-labelled modified nucleotides are added to sequence the clusters. They are especially designed to possess a reversible termination property allowing each cycle of the sequencing reaction to occur simultaneously in the presence of all four nucleotides. Cluster fluorescence is measured to identify which base has been incorporated: a green laser identifies the incorporation of G and T bases, whereas a red one identifies A and C bases. Two different filters are then used to distinguish between G and T, and A and C, respectively. After signal detection, the fluorophore and the termination

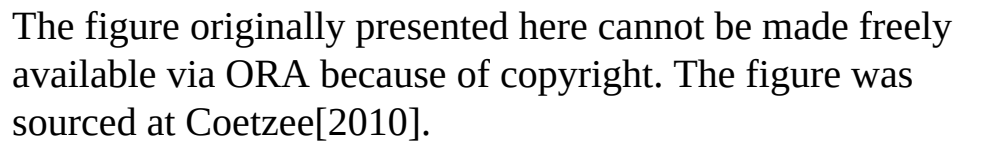


Figure 2.1: Illumina work flow. (a-d) The genomic DNA is randomly fragmented and adapters are ligated at both ends of the fragments. (e) The single-stranded fragments are bound to the surface of the flow cell. (f) Unlabelled nucleotides and enzymes are added. The latter incorporate nucleotides to build double-stranded bridges. (g) A denaturation step breaks the double-stranded templates into single-stranded ones. Amplification generates several million of clusters of double-stranded DNA. (h) A sequencing primer is annealed to each strand. (i-j) All four labelled reversible terminators, primers, and DNA polymerase are added to initiate the first sequencing cycle. (k) After laser excitation, the image of the emitted fluorescence from each cluster on the flow cell is captured and the identity of the first base for each cluster is recorded. The cycles of sequencing are repeated to determine the sequence of bases in a given fragment, a single base at each time. This figure was taken from Coetzee [2010].

modification are removed (Figure 2.1(i)). This process is repeated several times until all nucleotides have been sequenced (Figure 2.1(j)).

Recent protocol updates allow for paired end sequencing (sequencing of both strand directions) by chemically melting and washing away the synthesised sequence and performing additional bridge PCR steps before the starting strand is removed and a second primer is hybridised to sequence the second read [Kircher and Kelso, 2010].

Characteristics Illumina's main advantages are the simplicity of producing DNA libraries and amplifying the DNA fragments, the low per-base costs (∼US$0.50 per Mb), and the high amount of sequencing data that can be generated in a single run (each flow cell is composed of eight separately loadable lanes, each with a capacity of ∼35-60 million clusters allowing independent samples to be sequenced and analysed simultaneously [Kircher and Kelso, 2010]).

However, its major disadvantage is that it produces shorter read lengths than the 454 and Sanger sequencing methods due to its requirement to use chemically modified dye-labelled nucleotides as substrates, combined with a reengineered DNA polymerase, which compromises efficiency and fidelity of nucleotide incorporations [Shendure and Ji, 2008]. Read lengths are currently restricted to up to 150 bases as signal decay (due to dimming effects of the fluorophores [Kircher and Kelso, 2010]) and dephasing (such as the incomplete cleavage of fluorescent labels or termination moieties) increase the background noise and cause the error rate to increase towards the read end [Erlich et al., 2008; Illumina, 2010; Kircher et al., 2009; Lister et al., 2009; Rougemont et al., 2008; Shendure and Ji, 2008]. As a result, the dominant type of error are substitutions. These errors are frequently biased towards A/C, and G/T substitution errors as these nucleotide pairs exhibit similar emission spectra when excited using the same laser, making their distinction using optical filters difficult [Dohm et al., 2008; Kircher et al., 2009]. Further sources of error are the amplification involved in the library preparation and steps involved in the flow cell preparation as it is not possible to control the density of the amplified templates on the slide to avoid mixture or background noise disturbing the imaging process (in contrast to spatially controlled bead locations in the 454 sequencing method) [Kircher and Kelso, 2010]. Average raw error rates usually lie in the order of 1-1.5% but can be as high as 10% in the final cycles [Kircher and Kelso, 2010]. However, higher accuracy bases with error rates of about 0.1% can be identified through quality metrics associated with each base call or consensus sequences [Pettersson et al., 2009; Shendure and Ji, 2008]. Another disadvantage is the time required for a sequencing run. It is beyond that of the 454 technology because multiple data acquisition events are required

for every nucleotide incorporation, as the flow cell size exceeds the scanning microscope field [Rothberg and Leamon, 2008].

Despite the longer run times, the shorter read lengths, and the higher error rates, the fact that the Illumina sequencing approach generates a very large number of sequence reads for a considerably lower price than 454 or Sanger sequencing makes it most applicable in variant discovery studies by whole genome resequencing as well as other studies that can make use of a previously sequenced reference genome [Morrissy et al., 2010; Quail et al., 2009]. However, deeper sampling is required to ensure a high confidence in the determination of genetic differences (i.e. to distinguish true variants from sequence errors). Another aspect widening its application is the possibility of multiplexing: samples are indexed by hybridisation of additional sequencing primers in the same strand direction as part of the ligated adapter [Mamanova et al., 2010; Meyer and Kircher, 2010], allowing multiple samples to be sequenced simultaneously in one lane and to be computationally separated by their barcode.

2.2 Alignment of High Throughput Sequencing Data

Next-generation sequencing is a quickly evolving field. To tap its full potential, a firm understanding of the produced data as well as the tools available for its analysis is required. The alignment of short reads produced by NGS platforms to a reference genome is the first and most essential step of this analysis as it allows to rapidly determine whether the biological and sequencing experiments have succeeded. The amount of output generated by these high throughput sequencers imposes serious challenges to the algorithmic design of alignment tools. First, the reads are much shorter and less accurate than those generated by earlier methods. As a consequence, mutations or sequencing errors have a greater influence on the mapping position which might result in wrong alignments. To obtain maximal information from the data, the different error profiles of NGS platforms should be taken into account and quality scores should be used to assess the mapping quality of the read. An additional problem with the shorter length is that some reads might align equally well to multiple locations as most genomes contain repetitive regions on the length scale of the reads. Secondly, the amount of data is orders of magnitude greater than those produced by capillary-based sequencing. Therefore, algorithms must be optimised for storage space and speed. Traditional alignment algorithms (such as BLAST [Altschul et al., 1990]) can be run on expensive computer grids to map reads from high throughput sequencers. However, new generation alignment tools were especially designed to align short read sequences to a reference genome assembly. They consequently can significantly reduce both analysis time and computing costs.

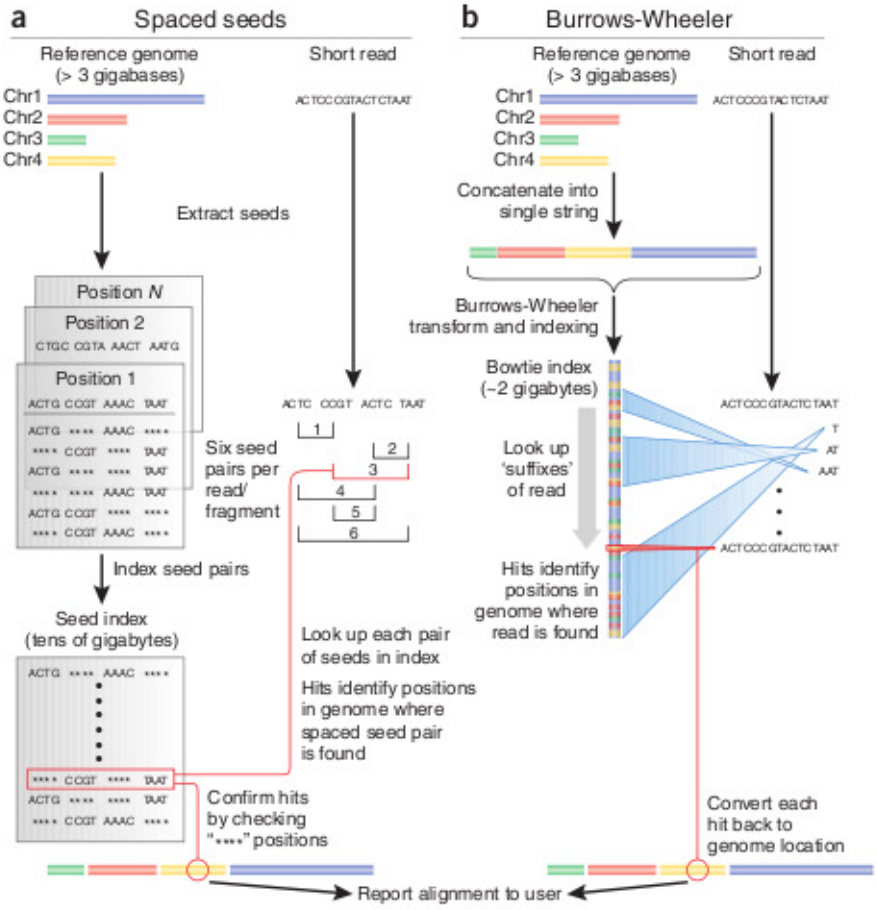


Figure 2.2: Two different algorithmic approaches to align short read sequencing data: (a) A hash-based algorithm based on spaced-seed indexing. Both the reference and the reads are indexed by cutting them into equally sized seeds which are then paired and stored in a hash table. Each pair of seeds can be used to look up the matching positions in the reference sequence. (b) An alignment algorithm based on the Burrows-Wheeler Transform. Starting from the right end, reads are mapped character by character against the transformed reference sequence. With each new character, the algorithm updates an interval (indicated by the blue beams) in the transformed string. Reprinted by permission from Macmillan Publishers Ltd: Trapnell and Salzberg [2009], Nature Biotechnology (`http://www.nature.com/nbt`).

Most of today's commonly used tools for aligning NGS data have a multi-step procedure in common (a schematic overview is given in Figure 2.2). In a first step, a heuristic is used to determine a small set of locations in the reference genome from which the read most likely originated. To ensure that the particular algorithm can be run on a single desktop computer, indices are used for this task as they speed up computations through their sublinear look up time. The ones most frequently implemented in alignment tools are seed hash tables (see Section 2.2.1) and the Burrows-Wheeler Transform (BWT) (see Section 2.2.2). In a second step, more accurate (and thus, slower) alignment algorithms (such as the well established optimal Smith-Waterman algorithm based on dynamic programming, see Section 2.2.3) are run on this subset. Due to the fact that the computational complexity of the algorithms used in the second step depends on the sequence length, it would be computationally infeasible to use these algorithms to perform a full search.

2.2.1 Hash-based Alignment

Hash tables are used in many alignment tools specifically designed for short reads generated by next-generation sequencers (amongst them BFAST [Homer et al., 2009], CloudBurst [Schatz, 2009], ELAND [Cox, 2007], GNUMAP [Clement et al., 2010], MAQ [Li et al., 2008a], MOM [Eaves and Gao, 2009], MOSAIK, NovoAlign, PASS [Campagna et al., 2009], PerM [Chen et al., 2009], ProbeMatch, RazerS [Weese et al., 2009], RMAP [Smith et al., 2009, 2008], SeqMap [Jiang and Wong, 2008], SHRiMP [Rumble et al., 2009], SOAP [Li et al., 2008b], and ZOOM [Lin et al., 2008]). They utilise a hash function to efficiently map certain identifiers to associated values. They can either be build on the reference genome or on the set of reads and permit rapid alignments by scanning the corresponding data with the same hash function. There are some advantages and disadvantages to both procedures: hash tables based on the reference genome are straightforward and have a constant memory requirement. Unfortunately, those requirements can be quite large depending on the size and complexity of the reference sequence. Hash tables built on input reads have variable memory requirements which are dependent on the read number and diversity. However, hashing read sequences needs more processing time as the entire reference genome must be scanned (which can be an overhead especially if only a few reads are mapped) and multi-threading is much harder than for hash tables built on the reference sequence. Hash tables can be generated using various hash functions and seeds. The latter enable genome sequence indexing in such a way that shorter sequences embedded within can quickly be found (similar to an index at the end of a book). The most commonly used types of seeds are

Continuous seeds: Continuous seeds use a sliding window of a particular length. The seeds of length 4 for the example sequence *ATGACGATGA* in Figure 2.3 are: *ATGA*, *TGAC*, *GACG*, *ACGA*, *CGAT*, *GATG*.

Non-contiguous seeds: Non-continuous seeds use a non-overlapping sliding window [Buhler, 2001; Ma et al., 2002]. The seeds of length 4 for the example sequence are: *ATGA*, *CGAT*.

Spaced seeds: Spaced seeds are sequence regions that need to have a particular pattern of matches and mismatches. In an example seed *1101*, *1* represents a sequence position that is required to match. The number of *1*s in the pattern corresponds to the seed weight. The seeds of length 4, weight 3, and the path *1101* for the above mentioned example sequence are: *AT*A*, *TG*C*, *GA*G*, *AC*A*, *CG*T*.

Periodic spaced seeds: The example seed with a path $n * (1101)$ and $n = 2$ is *AT*ACG*T*, *TG*CGA*T*, *GA*GAT*A*.

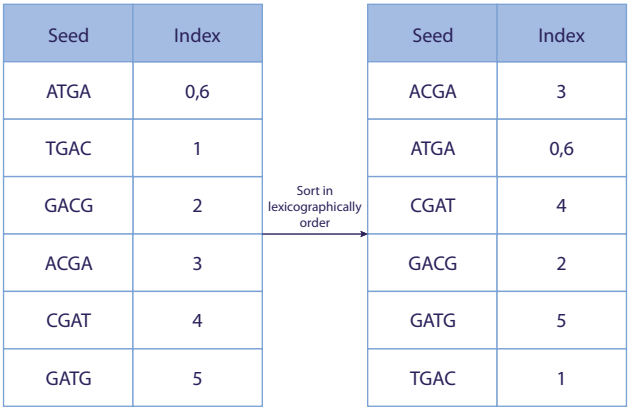


Figure 2.3: A hash table for the sequence *ATGACGATGA* is constructed using continuous seeds of length 4. Step 1: A sliding window of length 4 notes all start positions of a particular seed (note that seed *ATGA* appears twice in the sequence, at position 0 and 6). Step 2: The hash table is sorted in lexicographically order.

It should be noted that all algorithms are quite sensitive to the chosen seed weight. On the one hand, if the seed weight is too small, many false positives will be detected, slowing down the alignment process. On the other hand, if the seed weight is high, more seeds will be required to achieve a sensitive search. Most of the current alignment algorithms use either non-contiguous seeds or spaced seeds to implement hash tables. As the study in Chapter 4 makes use of MAQ (Section 2.2.1.1) for an initial analysis and Stampy (Section 2.2.1.2) with BWA (Section 2.2.2.1) as a premapper for the whole read alignment, these techniques are discussed in more detail. The reason for the choice of Stampy was two-fold: it has been shown that Stampy is one of the most accurate short read aligners currently available while facilitating short running times [Lunter and Goodson, 2011; Nielsen et al., 2011].

2.2.1.1 Mapping and Assembly with Qualities (MAQ)

Mapping and Assembly with Qualities [2] [Li et al., 2008a] is free software that performs alignments from single end as well as paired end short reads generated by Illumina sequencing machines (preliminary functions are available for SOLiD). For the first case, the initial 28bp of each read sequence are indexed using six hash tables corresponding to six non-contiguous spaced seeds (see Figure 2.4(a)) of length 8 and weight 4 (*11110000*, *00001111*, *11000011*, *00111100*, *11001100*, and *00110011*) by taking the nucleotides at the *1* positions of each template and hashing them into a 24-bit integer. The 24-bit integer serves as a hash key and is stored together with its corresponding region in a hash table. If the whole read perfectly aligns the reference genome, all its seeds also perfectly map it. In the case of a mismatch (either due to a SNP or a sequencing error), it must fall within one of the seeds. By applying various seeds, the list of possible candidate alignment

[2] `http://maq.sourceforge.net`

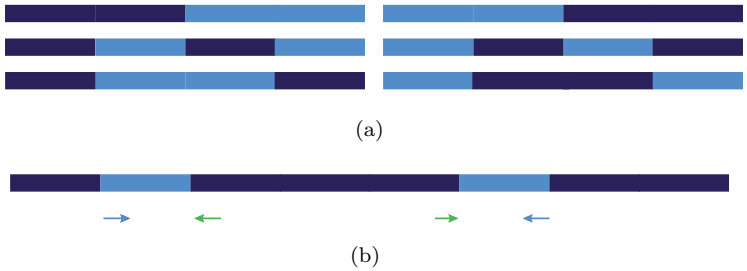


Figure 2.4: (a) MAQ indexing strategy: six hash tables are created by indexing the first 28bp of each read sequence using six non-contiguous spaced seeds. MAQ's algorithm is guaranteed to find all alignments with at most two mismatches within the first 28bp as a mismatch must fall within one of those seeds. (b) Paired end read information can be used in case only one end can be confidently mapped (confidently aligned read is indicated by a green arrow). A more accurate Smith-Waterman alignment is carried out in the region of the mapped read (determined by the inferred insert size; indicated in light blue) to identify the exact read placement and to discover whether indels interrupt the mapping of the unmapped end.

positions can be efficiently winnowed and all alignments with at most two mismatches within the first 28bp of the reads are guaranteed to be found. After the read indexing, the reference genome is scanned three times against the six hash tables on both forward and reverse strands by hashing each 28bp subsequence through two templates at a time (a template and its complementary) to identify potential read alignment positions (referred to as 'hits') within the hash tables. If a hit is found, the 28bp seed is extended without gaps. Its mismatch score q, defined as the sum of the quality values of the mismatching bases, is computed. It then hashes the coordinate of the hit together with the read identifier into another 24-bit integer h. The hit is scored as $q\ 2^{24} + h$. In this step, the algorithm searches for the ungapped match with at most two mismatches with the lowest mismatch score. If a read can be aligned equally well to more than one location (different hits with the same q), the hit with the smallest h is chosen. To reduce the memory footprint, only the position and score of the two best scored hits as well as the number of 0-, 1-, and 2-mismatch hits in the seed region are stored.

In the case of paired end read mapping, six hash tables are build for each end (12 in total) to jointly align the two reads in a mate-pair (to make use of the entire mate-pair information). As in single end read mapping, different rounds of indexing are executed. In each round, four hash tables are put in memory in order to index the first and the second end with two complementary templates each. This strategy guarantees that any hit with at most one mismatch can be found in each round. In a next step, both strands of the reference sequence are scanned for hits. If a hit is found on the forward strand, its position is appended to a queue which keeps the last two hits of this read on the forward strand. If a hit is found on the reverse strand, the queue of its mate is checked to search for hits of the mate on the forward strand within a maximum allowed distance in order to mark the two ends as a pair. In the case that only one end can be confidently aligned, a more accurate Smith-Waterman alignment is carried out in the region of the mapped read to

determine the exact read placement, and to discover whether indels interrupt the mapping of the unmapped end (see Figure 2.4(b)). Thereby, the size and coordinate of the region are estimated from the distribution of all mapped reads by taking the mean separation of read pairs plus or minus twice the standard deviation (SD) [Li et al., 2008a].

In order to handle ambiguities in read alignments, a mapping quality Q_s is assigned to each alignment. The mapping quality is a Phred-scaled quality score [Ewing and Green, 1998] which expresses the probability that the read did not come from the location the alignment algorithm has mapped it to

$$Q_s = -10 \ \log_{10} \ p\,(\text{read is wrongly mapped})\,.$$

Based on the assumption that sequencing errors are independent at different sites of the L-long read z, the probability $p(z|x,u)$ of z coming from position u in the L-long reference sequence x is equal to the product of the error probabilities of the mismatched bases at the aligned position. Assuming a uniform prior distribution $p(u|x)$, the posterior probability $p_s(u|x,z)$ can be calculated as

$$p_s(u|x,z) = \frac{p(z|x,u)}{\sum_{v=1}^{L-l+1} p(z|x,v)}.$$

Thus, the mapping quality of the alignment is

$$Q_s(u|x,z) = -10 \ \log_{10} \ (1 - p_s(u|x,z))\,.$$

As it is impractical to sum over all positions of the reference in large genomes, Q_s is in practice approximated as

$$Q_s = \min \begin{cases} q_2 - q_1 - 4.343 \ \log \ n_2 \\ 4 + (3-k')(\bar{q} - 14) - 4.343 \ \log \ p_1(3-k', 28), \end{cases}$$

where q_2 is the sum of quality values of the mismatching bases of the second best hit, q_1 is the corresponding sum for the best hit, 4.343 is $10/\log 10$, n_2 is the number of hits having the same number of mismatches as the second best hit, k' is the minimum number of mismatches in the seed, $\bar{q}$ is the average base quality in the seed, and $p_1(k, 28)$ is the probability of a perfect hit and a k-mismatch one coexisting given a 28bp sequence. This formula is useful as minimising the sum of quality values of the mismatching bases in the alignment is effectively maximising the posterior probability $p_s(u|x,z)$. If the read maps equally well at multiple positions, the mapping quality zero is assigned to the randomly chosen alignment. This procedure has the advantage that it gives

information about the fraction of reads that can be aligned to the genome as well as the copy number of repetitive regions, and that the coverage is not reduced in an uneven fashion avoiding holes due to repeats. Assuming independent errors, paired end mapping qualities Q_p for uniquely mapping pairs (for example alignments that align in the right orientation at a proper distance) can be derived from the two single end ones, Q_{s1} and Q_{s2}, by computing

$$Q_p = Q_{s1} + Q_{s2}.$$

In case there are multiple consistent hit pairs, their single end mapping qualities are taken as the mapping quality. The time complexity of the algorithm is $O(c_1 N \log N + c_2 L + c_3 N L \ 2^{-k})$ if N reads are aligned on a L-long reference sequence using k-bits in indexing. Thereby, the first term corresponds to the time spend on sorting the indexes, the second on scanning the entire reference genome, and the third one on processing the alignment in case there was a hit [Li et al., 2008a].

2.2.1.2 Stampy

Stampy [3] [Lunter and Goodson, 2011] performs single end as well as paired end hash-based alignments of Illumina reads of up to 4500bp allowing small gaps of up to 20bp. Stampy uses an open-addressing hashing algorithm to construct 15bp long hash keys from every fifth position of the reference genome. Thereby, the 15 nucleotides are encoded in a 30-bit word from which the reverse-complement symmetry is divided out by a transformation which maps words related by reverse-complementing to the same 29-bit word. Due to the fact that the confidential alignment of reads in repetitive regions is a difficult and time consuming task, the reference is first scanned to identify 15bp regions that occur more than 200 times to avoid long hash chains related to repetitive regions. These locations are not included in the hash table but their presence is denoted using an unused word. This flag later enables the skipping of reads which fall into repetitive regions in the genome.

After the hash table creation, all overlapping 15bp regions in the reads are scanned to detect possible mapping positions. If the considered 15mer contains a missing base (i.e. 'N'), all four nucleotide possibilities for this position are considered. If it comprises more than one N, it is neglected. Additionally, all 1-base mismatches are considered for reads shorter than 34bp. For reads up to 49bp, half of the 1-base mismatches and for longer reads, a third of the 1-base mismatches are taken into account. As all positions in the hash table matching the 15mer (or its reverse complement) are obtained, the number of potential candidate locations is reduced by applying a sequence

[3] `http://www.well.ox.ac.uk/project-stampy`

similarity filter. The filter utilises a fingerprint comprising the counts of A, C, and G nucleotides of three 4bp long words in close proximity of the mapped 15mer. The absolute nucleotide differences between the three 4bp long words in the reference and the corresponding positions at the putative genomic region is calculated and a threshold is used to limit the candidate mapping positions. These positions are then considered for full-length alignment using a Needleman-Wunsch global alignment. The single instruction, multiple data banded affine-gap alignment implementation takes individual nucleotide quality scores as well as penalties for both gap opening and extension into account to compute Phred-scaled mapping quality scores. To avoid differences between read errors, polymorphisms and substitutions, a term corresponding to the expected divergence to the reference (default 10^{-3}) is added to this score. At this stage, a 30bp diameter band is used in order to fully consider indels with length of up to 15bp. The best-matching candidate alignment position is determined (if more than one position is equally likely, a deterministic pseudo-random choice is performed) which is realigned with a 60bp diameter band allowing up to 30bp indels. To enable the consideration of large indels, the dynamic programming band is further increased in case the mapping comes within 10bp of the edge of the previous band.

In paired end read mapping, each read is individually aligned. Paired end locations are considered if both reads (i) are close together on the genome (within four SDs of the mean implied insert size), (ii) have a posterior probability of less than 1% of being mapped incorrectly, and (iii) exhibit a sufficiently low estimated probability (less than 1%) of not having found the right candidate. In any other case, a short list of mapping location pairs is generated from which regions are extracted that together constitute 99.9% of the single read posterior mapping quality. For each location, the mate is mapped against the reference around a region determined by the library insert size distribution (plus or minus four SDs from the mean). The best scoring single end mapping locations are added to this list. If the best scoring single end hits are selected as pair, the mapping quality is computed as the product of the single end mapping qualities. Otherwise, it is calculated as the single end posterior of the anchoring read. If no potential candidates could be found for the two reads, the paired end is flagged as unmapped.

For each read or read pair, a mapping quality, expressing the probability that the alignment algorithm has mapped it to the wrong location, is estimated using an approximate Bayesian model. This model takes three different cases into account: (i) the correct location for the read was not considered by the algorithm due to an abundance of errors or variants, (ii) the correct location for the read was considered but not chosen either because an exact repeat was chosen instead or because errors or variants in the read resulted in a better match of a different location, or (iii) the original sequence is not present in the reference sequence. The likelihood that a read pair (r_1, r_2)

is aligned to loci (x^0, y^0) is modelled as

$$L(r_1, r_2, x^0, y^0) = P_r(r_1|x^0)\ P_r(r_2|y^0)\ P_d(y^0 - x^0)\ P_u(x^0),$$

where P_r is the likelihood of an alignment, P_d models the insert size distribution, and P_u is the uniform distribution over the genome. Priors on read errors and variants (both SNPs and indel polymorphisms) are included in P_r while P_d includes a prior on the occurrence of anomalous inferred insert sizes caused by large indels and other structural variants. The posterior probability that the chosen locus is incorrect is given as

$$1 - P(x^0, y^0|r_1, r_2) = 1 - \frac{L(r_1, r_2, x^0, y^0)}{\sum_{(x,y)\in C} L(r_1, r_2, x, y)} \times \frac{\sum_{(x,y)\in C} L(r_1, r_2, x, y)}{\sum_{(x,y)\in \Omega} L(r_1, r_2, x, y)},$$

where C is the set of all candidate positions and Ω indicates all pairs of genomic coordinates. As the last factor can not be efficiently computed, it is replaced by the probability that the correct position of the read was not included in C. This probability is estimated from the nucleotide quality scores. A likelihood ratio test is performed to assess whether or not the inferred sequence similarity is unlikely to have appeared by chance using a random reference sequence. The resulting p-value is used both to cap the posterior probability and as a threshold to judge whether the alignment should be reported [Lunter and Goodson, 2011].

Stampy is about as fast as the previously described MAQ but much slower than BWA which uses a Burrows-Wheeler Transformation (see Chapter 2.2.2). For this reason, Stampy can be run in a hybrid mode in which BWA is used as a prealigner. Thereby, BWA aligns the majority of reads and Stampy is only used to realign reads with two or more mismatches. This procedure enables Stampy to run at 5-10 times its native speed without any reduction in sensitivity [Lunter and Goodson, 2011].

2.2.2 Burrows-Wheeler Transformation-based Alignment

Many aligners developed over the last years (such as BOWTIE [Langmead et al., 2009], SOAP2 [Li et al., 2009c], and BWA [Li and Durbin, 2009] which is further discussed in Section 2.2.2.1) use the Burrows-Wheeler Transformation [Burrows and Wheeler, 1994] to rapidly map short reads to a known reference genome. The Burrows-Wheeler Transformation, also known as 'block-sorting compression', is an algorithm which is predominantly used in data compression methods. BWT implementations of short read aligners are, at the same level of sensitivity, usually much faster than their hash-based counterparts [Flicek and Birney, 2009]. They use a compressed suffix array

(also referred to as 'FM index data structure' after Ferragina and Manzini who introduced the concept [Ferragina et al., 2007]) which allows to count the number of exact hits of a string W of length $|W|$ in $O(|W|)$ time. This index uses a suffix array created from the BWT sequence rather than the original one (which, for mammalian genomes, is much smaller than the genome input size [Graf et al., 2007]). The underlying data structure is generated in two steps: first, BWT is used to reversibly permute the order of the characters in the reference sequences in such a way that sequences existing multiple times occur together in the data structure. Then, the final index, which is used for the read alignment, is created. To perform an alignment, a backward search [Ferragina and Manzini, 2000] with BWT is executed which mimics a top down traversal on a genomic prefix trie. Hence, if inexact matches should be found, distinct substrings which are less than k edit distance away from the query read can be sampled from the implicit prefix trie. Similar to hash-based methods, more sensitive algorithms can be used as soon as a region is identified where the read is most likely to align.

Burrows-Wheeler Transform

Let Σ be an alphabet and let $X = a_0 a_1 \dots a_{n-1}$ be a string with length $|X| = n$. A string always ends with the character \$ which is not part of the alphabet Σ. \$ is lexicographically smaller than all symbols in Σ and it only appears at the end. $X[i] = a_i$ denotes the i-th character of X with $i = 0, 1, \dots, n-1$. A substring of X is indicated by $X[i, j] = a_i \dots a_j$ and a suffix of X by $X_i = X[i, n-1]$. A suffix array S of X is a permutation of the integers $0, 1, \dots, n-1$ in a way that $S(i)$ is the start position of the i-th smallest suffix [Li and Durbin, 2009]. The BWT of X is defined as $B[i] = \$$ if $S(i) = 0$ and $B[i] = X[S(i) - 1]$ otherwise. An example that shows the BWT and the constructed suffix array is given in Figure 2.5.

Prefix Trie

Let X be a string with characters from the alphabet Σ. A prefix trie for X is a tree in which each edge is labeled with one character and the string concatenation of all edge characters on the path from a leaf of the tree to the root gives a unique prefix of X (see Figure 2.6 for an example). Each string concatenation of the edge labels from a node to the root of the prefix trie gives a unique substring of X. Thus, W is an exact substring of X if and only if a node can be found whose concatenation of edge characters to the root represents W. This search can be performed in $O(|W|)$ time (by matching each character in W to an edge beginning from the root as all exact repeats collapse on one path) and it can be optimised by using the prefix information of W.

Position	String
0	AGGAGC$
1	GGAGC$A
2	GAGC$AG
3	AGC$AGG
4	GC$AGGA
5	C$AGGAG
6	$AGGAGC

Sort in lexicographically order

i	S(i)	String	B[i]
0	6	$AGGAG	C
1	3	AGC$AG	G
2	0	AGGAGC	$
3	5	C$AGGA	G
4	2	GAGC$A	G
5	4	GC$AGG	A
6	1	GGAGC$	A

Figure 2.5: A Burrows-Wheeler transformed string is generated in three steps given the original string $X = AGGAGC$. Step 1: X is circulated to create seven strings which are then lexicographically sorted. Step 2: The suffix array $S = (6, 3, 0, 5, 2, 4, 1)$ is formed by taking the positions of the first symbols. Step 3: The BWT string $CG\$GGAA$ is obtained by concatenating the last symbols of the circulated strings.

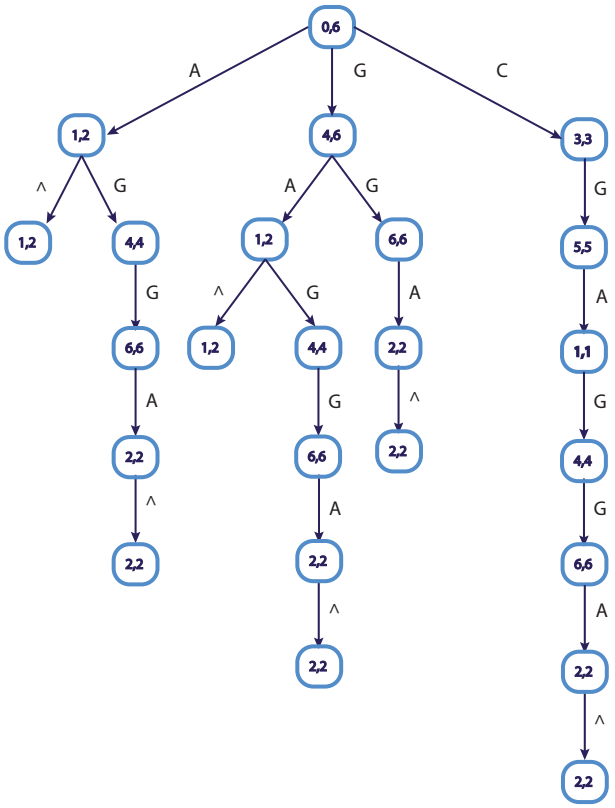


Figure 2.6: The prefix trie of the string $AGGAGC$ in which ^ marks the start of the string. The suffix array interval is indicated by the two numbers in a node.

Due to the fact that the prefix trie of X is equal to the suffix trie of the reverse of X, suffix trie theory can be applied to the prefix trie. The suffixes that have W as a prefix are sorted together as the position of each occurrence of W in X will be in an interval in the suffix array. This suffix array interval of W is defined as $[\underline{R}(W), \overline{R}(W)]$ in which $\underline{R}(W)$ and $\overline{R}(W)$ are defined as follows

$$\underline{R}(W) = \min\left(k : W \text{ is prefix of } X_{S(k)}\right)$$

$$\overline{R}(W) = \max\left(k : W \text{ is prefix of } X_{S(k)}\right)$$

with $\underline{R}(W) = 1$ and $\overline{R}(W) = n - 1$ if W is an empty string. The positions of all occurrences of W in X are given by $S(k) : \underline{R}(W) \leq k \leq \overline{R}(W)$ (see example in Figure 2.6). As the positions of W can be obtained through the information of the suffix array interval, a short read sequence alignment is equivalent to the search for suffix array intervals of substrings of X that match the read.

Backward Search

In 2000, Ferragina and Manzini proved that if W is a substring of X, $\underline{R}(aW) \leq \overline{R}(aW)$ with

$$\underline{R}(aW) = C(a) + O(a, \underline{R}(W) - 1) + 1$$

$$\overline{R}(aW) = C(a) + O(a, \overline{R}(W)),$$

where $C(a)$ is the number of characters in $X[0, n-2]$ that are lexicographically smaller than $a \in \Sigma$ and $O(a, i)$ is the number of occurrences of a in $B[0, i]$ if and only if aW is a substring of X [Ferragina and Manzini, 2000]. Therefore, a backward search which iteratively computes $\underline{R}$ and $\overline{R}$ from the end of W can be performed to test whether W is a substring of X (and in case it is, to count its occurrences). A backward search is equivalent to the exact string matching on a prefix trie.

2.2.2.1 Burrows-Wheeler Alignment Tool (BWA)

BWA [4] [Li and Durbin, 2009] is a free software published under the GPLv3 licence that rapidly aligns short base or colour space reads with a reference genome. It implements two algorithms based on the backward search with BWT which recursively samples distinct substrings from a genome in order to search for suffix array intervals of substrings that match the query. One algorithm is designed for short queries up to 200bp with low error rate (<3%) which does gapped global

[4] `http://bio-bwa.sourceforge.net/`

alignment and supports paired end reads. The second, BWA-SW, is designed for longer reads with more errors. It performs heuristic Smith-Waterman-like alignments to find high scoring local hits.

In contrast to the exact matching discussed above, the two algorithms are slightly altered to enable matches with no more than a specified number of variations (including both mismatches and gaps). In order to search for suffix array intervals of substrings of X which match the query W with at most k mismatches, W's bases are searched backwards for matches and the number of mismatches are counted. (BWA treats all non-A/C/G/T bases within a read as mismatches while the ones in the reference sequence are converted into random nucleotides, a procedure which can lead to false hits.) The search aborts as soon as the mismatch count exceeds k. k is chosen in such a way that <4% of reads with a 2% uniform base error rate may contain more than k differences (for 15-37bp reads: $k = 2$, for 38-63bp reads: $k = 3$, for 64-92bp reads: $k = 4, \ldots$). Furthermore, the number of allowed differences in the seed sequence (the first few tens of base pairs on a read) can be limited due to the fact that they are usually more accurate than the 3'-end. To reduce search space, the backward search is bounded by the array $D(.)$, where $D(i)$ is the lower bound of the number of differences in $W[0, i]$.

For each alignment, a mapping quality score is generated using an algorithm similar to MAQ with the exception that it is assumed that a true mapping can always be found. This assumption counteracts the fact that the formula used in MAQ's algorithm overestimates the probability of missing the true hit and thus, underestimates the mapping quality score. As with MAQ, non-unique reads are placed randomly with mapping quality 0. BWA can also handle paired end reads. For this purpose, all positions of good hits are found and ordered according to their chromosomal coordinates before the algorithm linearly scans through all potential hits to determine a pair (suboptimal hits can be considered). A Smith-Waterman alignment is performed for unmapped reads in case their mate can be reliably mapped.

To reduce the memory footprint of BWA, an algorithm developed by Hon et al. [Hon et al., 2007] was adapted from an earlier implementation in BWT-SW [Lam et al., 2008]. This algorithm allows the utilisation of only n bits of working space for the construction of the suffix array (most algorithms use at least $n\lceil \log_2 n \rceil$) as well as less than 1GB memory for the construction of the BWT. In addition, the BWA transform of the human genome is small enough to fit into less than 2GB of memory [Trapnell and Salzberg, 2009]. Several other modifications were made to the algorithm's design to realistically model biological data (e.g. different penalties for mismatches, gap openings, and gap extensions have been incorporated). More details about BWA can be found in Li and Durbin [2009].

2.2.3 Smith-Waterman Algorithm-based Alignment

In 1981, a variation of the dynamic programming Needleman-Wunsch algorithm [Needleman and Wunsch, 1970], the Smith-Waterman algorithm, was proposed by Temple Smith and Michael Waterman [Smith and Waterman, 1981]. It is a well known algorithm for performing local sequence alignment to determine a pair of segments, one from each of the sequences, such that there is no other pair of segments with greater similarity. Instead of looking at the total sequence, the Smith-Waterman algorithm compares segments of all possible lengths and optimises the similarity measure. The algorithm can be divided into two parts: the construction of an alignment matrix and the backtracking.

The $n \times m$ dynamic programming matrix between two sequences $A = a_1...a_n$ and $B = b_1...b_m$ is constructed using a similarity scoring scheme $s(a, b)$ and the weights w_k for deletions of length k. The cells $sw[k, 0]$ and $sw[0, l]$ for each $0 \leq k \leq n$ and $0 \leq l \leq m$ of the matrix with row i and column j are initialised with zero. Then, for every $0 \leq i \leq n$ and $0 \leq j \leq m$, each cell $sw[i, j]$ containing the maximal value of similarity of two segments ending in a_i and b_j is filled, using the following mathematical formula:

$$sw[i,j] = \max \begin{cases} 0 & \text{(no similarity up to } a_i \text{ and } b_j\text{)} \\ \max_{l \geq 1}(sw[i, j-l] - w_l) & (b_j \text{ is at the end of a deletion of length } l\text{)} \\ \max_{k \geq 1}(sw[i-k, j] - w_k) & (a_i \text{ is at the end of a deletion of length } k\text{)} \\ sw[i-1, j-1] + s(a_i, b_j) & \text{(match or mismatch).} \end{cases}$$

After the matrix is constructed, the pair of segments with maximum similarity is found by initially locating the maximum element of sw. The other matrix elements leading to this maximum are then sequentially determined using a backtracking procedure ending with an element of sw equals to zero. This technique does not only detect the segments but it also produces the corresponding alignment. An example is given in Table 2.1.

2.3 Variant Calling from High Throughput Sequencing Data

The reliable and accurate detection of genetic variants is one of the most important tasks when resequencing a genome to avoid errors by calling positional variants which are produced by misaligned reads or sequencing errors.

	-	C	A	G	C	C	T	C	G
-	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
A	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
A	0.0	0.0	1.0	0.7	0.0	0.0	0.0	0.0	0.0
T	0.0	0.0	0.0	0.7	0.3	0.0	1.0	0.0	0.0
G	0.0	0.0	0.0	1.0	0.3	0.0	0.0	0.7	1.0
C	0.0	1.0	0.0	0.0	2.0	1.3	0.3	1.0	0.3
C	0.0	1.0	0.7	0.0	1.0	3.0	1.7	1.3	1.0
A	0.0	0.0	2.0	0.7	0.3	1.7	2.7	1.3	1.0
T	0.0	0.0	0.7	1.7	0.3	1.3	2.7	2.3	1.0
T	0.0	0.0	0.3	0.3	1.3	1.0	2.3	2.3	2.0
G	0.0	0.0	0.0	1.3	0.0	1.0	1.0	2.0	3.3
A	0.0	0.0	1.0	0.0	1.0	0.3	0.7	0.7	2.0
C	0.0	1.0	0.0	0.7	1.0	2.0	0.7	1.7	1.7
G	0.0	0.0	0.7	1.0	0.3	0.7	1.7	0.3	2.7
G	0.0	0.0	0.0	1.7	0.7	0.3	0.3	1.3	1.3

Table 2.1: The similarity scoring scheme as well as the deletion weights in this example are chosen on an *a priori* statistical basis. Thereby, a match, $a_i = b_j$, produces a similarity value $s(a_i, b_j)$ of unity while a mismatch produce a minus one-third. These values have an average of zero for long, random sequences over an equally probable four letter set $\sum = A, C, G, T$. The deletion weight is $w_k = 1.0 + 1/3k$. The maximal element found (3.3) is highlighted in blue whereas the values of the track back path are highlighted in green. Thus, the obtained alignment contains the two subsequences 'GCCATTG' and 'GCC-TCG'.

2.3.1 Data Preprocessing

A number of steps is necessary to convert the aligned read data into a variant call set, each of which contributes to the accuracy of the final call set. In a first step, lane-level duplicates resulting from the amplification step in the sample preparation as well as library-level duplicates (and, if reads were produced by paired end sequencing, reads that do not satisfy the proper pair conditions) need to be removed to improve the overall quality of the variant calls as they manifest themselves as high read coverage support. In general, all but one copy of the duplicates in the data is removed (based on the outer coordinates of individual reads). Several computer programs have been developed that comprise command-line utilities to perform this task by manipulating alignments before the variant calling (e.g. SAMtools [Li et al., 2009a] or Picard which were used in the study in Chapter 4).

In a second step, base quality scores should be recalibrated as reported quality scores only inaccurately display the actual probability of mismatching the reference genome due to quality variations caused, amongst others, by machine cycle and by sequence-context. This is an important step, as recalibration helps to identify high quality bases and to adjust reported quality scores, leading to an improved accuracy of subsequent SNP calls which rely on these quality scores. Several algorithms have been published that recalibrate the per-base quality scores of sites with no known SNPs by comparing the sequenced reads to the reference genome (e.g. SOAP2 [Li et al., 2009c] and the Genome Analysis Toolkit [5] (GATK) [DePristo et al., 2011; McKenna et al., 2010]). If no SNPs have been previously reported (e.g. in the public catalogue of genetic variant sites, dbSNP [Sherry et al., 2001]), candidate positions for variant sites can be identified first and only the remaining sites

[5] `http://www.broadinstitute.org/gsa/wiki/index.php/The_Genome_Analysis_Toolkit`

can be used in the recalibration step (before another round of SNP calling is performed). The study in Chapter 4 applied the alignment-based recalibration algorithm implemented as part of GATK which was also adopted in the 1000 Genomes Project [The 1000 Genomes Project Consortium, 2010]. For each non-polymorphic site, the algorithm assigns different categories (classified by different features such as raw quality score, position of the nucleotide in the read, dinucleotide-context, and read group) to each nucleotide that aligns to the site. In the next step, it estimates the empirical quality score for each of these categories using the number of mismatches of the nucleotides to the reference sequence. The residual differences between the empirical qualities and the mismatch rates implied by the raw quality scores are added to the raw quality scores to obtain recalibrated quality scores [Nielsen et al., 2011].

The necessity of the third (and most often last) step of the data preparation results from the fact that alignment algorithms map reads individually to the reference position from which they most likely originated. As SNPs occur much more frequently than indels in the genomes of most species, most algorithms rather annotate SNPs than indels at positions mismatching the reference genome. At positions of true indels, alignment artefacts result in false positive SNP calls [Li and Homer, 2010]. One possibility to identify those positions and to improve SNP calls is multiple sequence alignment which locally realigns reads such that the number of mismatching bases is minimised across all reads (e.g. by using GATK's 'IndelRealigner'). However, current implementations are computationally intense and filtering SNPs around predicted indels is hindered by the challenge of correctly identifying new indels in the first place. Another way to avoid those artefacts is calculating a Base Alignment Quality [Li, 2011] on a per-base level which down weights base qualities in regions with high ambiguity in the local alignment, thereby avoiding an excess of false positive SNP calls in low complexity regions of the genome. The newly assigned base qualities can then be used in the initial SNP call.

2.3.2 Variant Discovery

A multitude of bioinformatic tools have been developed to facilitate SNP discovery from short NGS read data. Earlier methods called variants by counting high quality alleles at each site and applying simple cutoff rules (e.g. a heterozygous genotype would be called if the proportion of the non-reference alleles is between 20-80% while a homozygous genotype would be called otherwise [Harismendy et al., 2009; Wang et al., 2008]). While this procedure works well for high coverage (>20X) sequencing data (the probability that a heterozygous individual falls outside the 20-80% range is low), it typically leads to problems in sequencing studies with low or medium coverage (e.g. heterozygous genotypes are usually under-called due to the fixed cutoffs) [Nielsen et al., 2011].

In contrast, newer approaches are based on probabilistic frameworks which not only work better for low coverage data, but which also have the advantage of incorporating statistical uncertainty (and other additional information) in the SNP calling process [Li et al., 2008a, 2009b,c]. They calculate the likelihood that a given locus is homozygous or heterozygous for a variant allele, taking, amongst other factors, error rates and error profiles of different NGS platforms, the read coverage of the locus, as well as the probability that the variant candidate has been produced by a 'bad' mapping into account, leading to SNP calls with associated quality scores (to give information about the uncertainty of the call) [Magi et al., 2010; Nielsen et al., 2011]. In addition, they can be coupled with prior information (such as allele frequencies). In brief, a Bayesian framework is applied, computing the conditional likelihood of the nucleotides at each position by using the Bayes rule:

$$P(G|D) = \frac{P(D|G)\ P(G)}{P(D)}.$$

Using this formula, the posterior probability $P(G|D)$ of a genotype G given the data D can be calculated by multiplying the prior probability of the genotype $P(G)$ with the likelihood that the data was observed given the genotype $P(D|G)$ and dividing it by the normalising constant $P(D)$. The likelihood is calculated using the pileup of bases and the associated quality scores at a given locus. The prior probability for each genotype can either assign equal probability to each genotype or it can be based on other information such as a set of reference data (e.g. information obtained from SNPs from a variant databases such as dbSNP [Sherry et al., 2001]) usually with a strong weight against heterozygous to avoid false positive SNP calls due to sequencing artefacts [Nielsen et al., 2011]. Prior probabilities on the genotype can be improved by simultaneously calling different individuals (e.g. by incorporating either allele frequency information from which genotype probabilities can be computed using the Hardy-Weinberg equilibrium (HWE) or genotype frequency information estimated from larger data sets [Li et al., 2009b; Martin et al., 2010]). The joint calling procedure generally leads to a substantial increase of genotype-calling accuracy compared to the calls obtained independently from each individual [Nielsen et al., 2011]. $P(G|D)$ is then computed for all 10 genotypes (A/A, A/C, A/G, A/T, C/C, C/G, C/T, G/G, G/T, and T/T), and, in general, the genotype with the highest posterior probability is chosen and its probability used to measure the confidence in the call [Nielsen et al., 2011]. In the easiest scenario, a SNP can be called when an individual is either heterozygous or homozygous for the non-reference allele at a locus. However, if multiple individuals are called simultaneously, the joint posterior probability is a better estimate to ascertain the probability that all individuals are homozygous for the reference allele [Nielsen et al., 2011].

Besides true variation, the initial SNP call set contains false positives due to systematic machine artefacts, mismapped reads, and misaligned indels caused by the reference genome which is both incomplete and does not represent all genetic variation. Such false positive calls frequently (i) exhibit an unusual behaviour according to excessive depth of coverage (as a read depth much lower or much higher than expected is suggestive of copy number variation which might lead to false positive SNP calls in the mapping of paralogous sequences), (ii) show a skewed allelic imbalance, (iii) occur preferentially on a single strand, (iv) appear in regions of poor read alignment (indicated by a poor mapping quality), and (v) arise in close proximity to other SNPs. Thus, the majority of such calls can be detected and rejected using different filter criteria based on the above observations (such as deviations from HWE, low quality scores, extreme read depths, and strand bias; appropriate filter criteria will depend on the exact sequencing protocol) which will greatly enhance the accuracy of the SNP call.

Algorithms to compute genomic differences include the software packages Bambino [Edmonson et al., 2011], GATK [DePristo et al., 2011; McKenna et al., 2010], MAQ [Li et al., 2008a], FreeBayes[6], SOAP [Li et al., 2008b] and its improvements SOAP2 [Li et al., 2009c] and SOAP3 [Liu et al., 2012], and Slider [Malhis et al., 2009]. In the study described in Chapter 4, GATK was used for to perform the initial SNP call and the subsequent SNP filtering as it has a good accuracy in multi-sample variant calling [Nielsen et al., 2011].

For studies that use medium coverage sequence data, it is also advisable to refine the initially obtained genotype information before estimating haplotypes. The genotype likelihood information obtained during the SNP call can be used in a genotype-calling algorithm which determines the genotype for each individual at each site. Instead of calling each site individually, most approaches take advantage of the pattern of LD at nearby sites to estimate genotypes which makes them computationally more expensive, but also leads to a significant improvement in the genotype-calling accuracy when multiple samples have been sequenced [Nielsen et al., 2011; The 1000 Genomes Project Consortium, 2010]. In addition, it allows to fill in missing data in the original genotype calls [Browning and Browning, 2007; Dai et al., 2006; Howie et al., 2009; Marchini et al., 2007; Minichiello and Durbin, 2006; Scheet and Stephens, 2006]. For this task, several methods have been published, amongst them IMPUTE [Howie et al., 2009; Marchini et al., 2007], MACH [Li et al., 2010], fastPHASE [Scheet and Stephens, 2006], PHASE [Stephens and Donnelly, 2003; Stephens and Scheet, 2005; Stephens et al., 2001], and BEAGLE [Browning and Yu, 2009; Browning, 2006; Browning and Browning, 2007], the last two of which have been used in the study in Chapter 4. An overview of the different techniques can be found in Marchini and Howie [2010].

[6] `http://bioinformatics.bc.edu/marthlab/FreeBayes`

2.4 Validation

SNP calls do not only contain true variation but also artefacts usually caused by misaligned reads or systematic sequencing errors. In order to estimate the FDR of the SNP calls, a validation experiment should be conducted to determine the presence or absence of SNPs called in short-read sequencing using an independent technology (such as customised SNP genotyping platforms). One example of a widely used platform is Sequenom [7] which was used to validate a subset of the SNPs called in the study described in Chapter 4. It uses a MassARRAY system for SNP genotyping which combines a reproducible primer extension chemistry (a single termination mix and universal reaction conditions are used for all SNPs) with MALDI-TOF mass spectrometry (differences in mass are used to differentiate between SNP alleles). However, it should be noted that different technologies can lead to different (sometimes even disagreeing) results, and dependent on the technique used, real polymorphic sites might be missed. Thus, it is advisable and desirable to validate a data set using multiple independent technologies.

[7] http://www.sequenom.com/home/

Preface

Work from Chapter 3 has been published as:

> Goriely, A., Hansen, R.M., Taylor, I.B., Olesen, I.A., Jacobsen, G.K., McGowan, S.J., Pfeifer, S.P., McVean, G.A., Rajpert-De Meyts, E., and Wilkie, A.O., *Activating Mutations in FGFR3 and HRAS Reveal a Shared Genetic Origin for Congenital Disorders and Testicular Tumors*, Nature Genetics 2009, Vol. 41, No. 11, pp. 1247-52.

The study was designed by Anne Goriely and Andrew Wilkie (as described in Section 3.3). Biological samples were obtained from the Oxford Fertility Clinic. DNA was extracted by Anne Goriely, Aimee Fenwick, and Indira Taylor. All samples for *FGFR3* quantification were processed by Anne Goriely. High throughput sequencing was performed on a Illumina GAIIx by High Throughput Genomics at the Wellcome Trust Centre for Human Genetics. The data was processed by Simon McGowan and Anne Goriely. Statistical analysis and validation were performed by Gil McVean and myself.

Chapter 3

Statistical Methods for Estimating the Rate of De Novo Mutation at FGFR3 using High Throughput Sequencing

The majority of human germ line mutations show a paternal origin [Nachman and Crowell, 2000]. Some of these mutations cause rare inherited congenital disorders of which many do not only exhibit an origin from spontaneous mutations in the germ line of healthy fathers, but also an association with an increased age of the father at the time of conception (about 2-5 years above the population average [Rannan-Eliya et al., 2004]). These mutations are referred to as 'paternal age effect' mutations. The best known examples occur in the genes *HRAS* (causing Costello syndrome) [Sol-Church et al., 2006], *PTPN11* (Noonan syndrome) [Tartaglia et al., 2004], *NF1* (Sporadic neurofibromatosis 1) [Bunin et al., 1997], *RET* (multiple endocrine neoplasia types 2A and 2B) [Carlson et al., 1994], *FGFR2* (Apert, Crouzon, and Pfeiffer syndromes) [Moloney et al., 1996], and *FGFR3* (achondroplasia and Muenke syndrome) [Wilkin et al., 1998]. Paternal age effect mutations have been attributed to an accumulation of replication errors or inefficient repair processes resulting in higher mutation levels in the sperm with increasing age due to the requirement of repeated cycles of replication of spermatogonial stem cells during spermatogenesis, a process known as 'copy error hypotheses' [Goriely et al., 2005]. However, even if there is some evidence supporting this hypothesis, recent work has shown that the prevalence of some of these specific mutations reflects a protein-driven, positive selection of mutant cells according to the functional consequences of the encoded amino acid substitution [Goriely et al., 2009, 2003]. One study affirming this idea is the one of Goriely et al. investigating Fibroblast Growth Factor Receptor 2 (*FGFR2*) gene mutations. The authors proposed in their work that the selection of at least one mutated nucleotide, the one with the highest inferred mutation rate [Kan et al., 2002], nucleotide

755, is protein-driven. This indicates that specific *FGFR2* mutations, which arise at low frequency in diploid spermatogonial stem cells, confer a proliferative advantage to the mutated cell relative to the wild-type ones, leading to a clonal expansion within the testis over time. Goriely et al. further suggested that this concept could be extended to other genes showing related biological features, such as *FGFR3*, which is part of the same signalling pathway than *FGFR2* [Goriely et al., 2005].

3.1 Fibroblast Growth Factor Receptor 3

FGFR3 is one out of four members of the receptor tyrosine kinase family (*FGFR* 1-4) which all share a common organisation structure consisting of three extracellular immunoglobulin-like domains, one hydrophobic transmembrane domain and two cytoplasmic tyrosine kinase subdomains responsible for its catalytic activity [Baujat et al., 2008; Schlessinger, 2000]. Binding of fibroblast growth factor ligands together with heparin to one of the immunoglobulin-like domains induces dimerisation of two *FGFR3*s. The cytoplasmic tyrosine kinase subdomain of one receptor phosphorylates the one of the other receptors [Pantoliano et al., 1994; Shi et al., 1993]. The phosphorylated residues serve as docking sites for adaptor proteins as well as effectors that propagate intracellular signals [Baujat et al., 2008]. These signals regulate many cellular processes which control growth, development, and maintenance of bone and brain tissue. Point mutations in the *FGFR3* gene cause permanent receptor phosphorylation, resulting in the protein being overly active. This interferes with skeletal development and leads to disturbances in bone growth [Sahni et al., 1999; Weksler et al., 1999], causing a series of skeletal dysplasias including hypochondroplasia (HCH), thanatophoric dysplasia type I and II (TDI and TDII), and severe achondroplasia with developmental delay and acanthosis nigricans (SADDAN) [Rannan-Eliya et al., 2004; Wilkin et al., 1998]. Each of these mutations, causing *FGFR3*-related skeletal dysplasias, is a gain-of-function mutation stimulating excessive intracellular signalling [Kannan and Givol, 2000; Vajo et al., 2000]. Their associated phenotypes are similar - dominated by short limbs, long trunk, large head with frontal bossing, and mid-facial hypoplasia [Horton and Lunstrum, 2002]. Nevertheless, their severity is quite different, ranging from the low activated, less severe HCH mutant (mild dwarfism) to the highly activated, much more severe TD one with perinatal lethality [Lievens et al., 2004]. Thereby, the severity of the phenotype associated with different mutations seems to be correlated with the extent of the gain [Naski et al., 1996; Ornitz et al., 1996], whereby the function that is gained depends on the cell in which the *FGFR3* receptor is expressed [Horton and Lunstrum, 2002]. Interestingly, all described dysplasias are caused by related mutations of the same codon, K650, located at cDNA position 1948-1950 (see Figure 3.1).

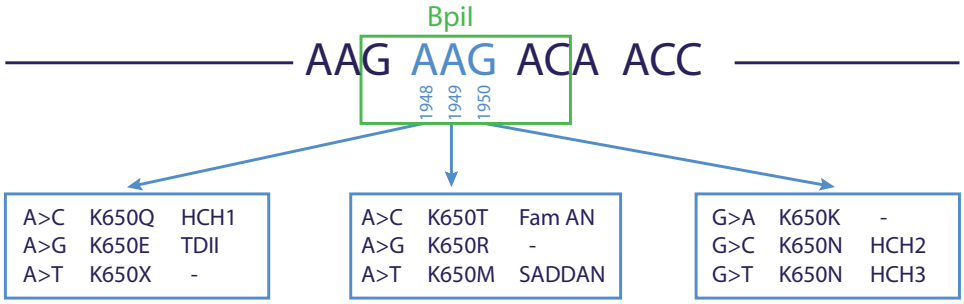


Figure 3.1: DNA region around the wild-type K650 codon (AAG) located at cDNA position 1948-1950 (depicted in light blue). The nine possible single nucleotide substitutions of the codon are shown with their associated germ line defects: hypochondroplasia (HCH), thanatophoric dysplasia type II (TDII), familial acanthosis nigricans (Fam AN), severe achondroplasia with developmental delay and acanthosis nigricans (SADDAN), - (not reported as germ line mutation). The green box indicates the recognition sequence of the restriction enzyme *BpiI* which can be used to differentiate between the wild-type and mutated K650 codons.

3.2 Aim of the Study

Mutations occurring at the K650 codon in the *FGFR3* gene are the cause of several dysplasias such as HCH, TD, and SADDAN. These paternal age effect mutations have generally been attributed to higher mutation levels in the sperm of men of increasing age due to an accumulation of replication errors or inefficient repair processes during spermatogenesis. However, recent work on *FGFR2*, a gene which is part of the same signalling pathway and which comprises related biological features, has shown that the prevalence of some specific mutations is protein-driven and positive selection acts on mutant cells according to the functional consequences of the encoded amino acid change [Goriely et al., 2009, 2003]. For this reason, the aim of this study is to investigate whether there is any evidence for a selective advantage of pathogenic *FGFR3* mutations at the K650 codon in the male germ line of humans, similar to the one previously found in *FGFR2*.

High throughput sequencers, enabling the generation of large amounts of data within a short amount of time and at much lower costs than traditional sequencing methods [Coombs, 2008; Rothberg and Leamon, 2008], have permitted large-scale studies of mutation rates in genomes. Although their produced data offers the potential to a better understanding of genetic diseases, their analysis has proven to be computationally challenging as the huge amount of data generated imposes several challenges on statistical analysis. In contrast to the previous study of the *FGFR2* gene which was based on intensity data obtained from pyrosequencing, the data for this work has been generated by Illumina sequencing (see Section 2.1.1). As every platform has specific characteristics (such as different error rates and error structures) as well as different limitations that need to be addressed, novel statistical methods need to be developed in order to use the output from the Illumina high throughput sequencer to quantify all possible point mutations at the K650 codon in the *FGFR3* gene.

Particularly problematic in NGS studies are the higher error rates which are caused by a multitude of different factors and which hinder the straightforward interpretation of the data. These factors include biases that have been introduced by (i) incomplete enzymatic digestion (performed to enrich for particular mutations), (ii) PCR amplification steps (which are required in the library preparation), or (iii) steps involved in the flow cell preparation (as it is not possible to control the density of the amplified templates on the slide to avoid mixture or background noise disturbing the imaging process [Kircher and Kelso, 2010]). In addition, errors can occur during the sequencing or base-calling process, or during the alignment step. Missing data as well as low coverage further complicate this non-trivial issue. In Illumina sequencing, read errors are known to increase towards the end of the read [Erlich et al., 2008; Illumina, 2010; Kircher et al., 2009; Lister et al., 2009; Rougemont et al., 2008; Shendure and Ji, 2008] as signal decay (due to dimming effects of the fluorophores [Kircher and Kelso, 2010]) and dephasing (caused by the incomplete cleavage of fluorescent labels or termination moieties) increase the background noise. Average raw error rates for reads generated by Illumina sequencing lie in the order of 1-1.5% (but can be as high as 10% in the final cycles [Kircher and Kelso, 2010]), with substitutions as dominant type of error (in contrast to pyrosequencing, where the dominant type of error are indels).

The above mentioned biases, errors, and uncertainties can strongly influence downstream analyses, making it crucial to quantify and account for them. Advanced statistical models are well suited for the analysis of NGS data as they allow to capture subtle but complex features of the machine- and run-dependent error structures. Two different methods are explored in this study to estimate the rate of *de novo* mutations in human sperm from high throughput sequencing data: (i) a model based on the knowledge of the error structure of Illumina reads together with their inbuilt quality scores and (ii) a black-box model assuming that there was no *a priori* information available. As the first model requires a good understanding of the underlying error structure of the Illumina reads as well as information about how well the inbuilt quality scores are calibrated, it is important to have a detailed prior knowledge about the study design.

3.3 Study Design

Ejaculates from 78 healthy men (aged 22.1-73.9 years) and blood samples from eight donors (36.6-73 years) were obtained for quantification of K650 *FGFR3* mutations. DNA samples were divided into three aliquots, two of which were spiked with different amounts of diluted genomic DNA heterozygous for the *FGFR3* 1948A>C (A to C) substitution. Mutation levels were estimated relative to the diluted spike DNA and validated using serial dilutions of *FGFR3*.

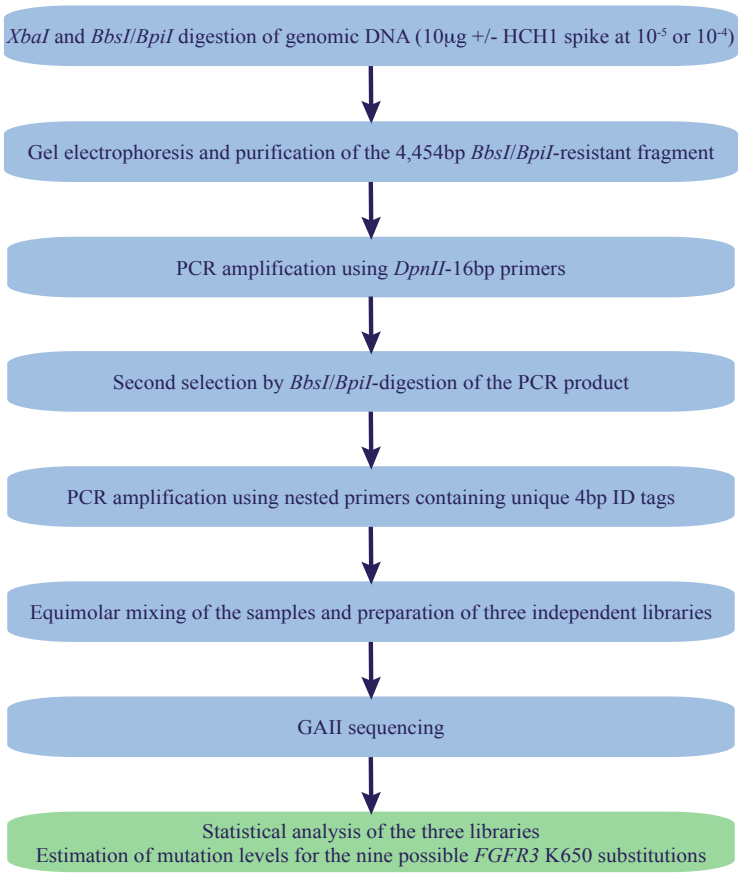


Figure 3.2: Strategy to quantify K650 *FGFR3* mutation levels in human sperm and blood using tagged-oligonucleotide pooled PCR and massively parallel sequencing.

3.3.1 Protocol

The samples were prepared using the following protocol (as illustrated in Figure 3.2): DNA (triplicate 10μg samples corresponding to approximately 3×10^6 copies of the haploid genome) was cut with the enzyme *XbaI* at the positions 11,365 and 15,819 resulting in a 4,454bp fragment containing the K650 codon. Taking into account that five other point mutations of the K650 codon have been previously identified in germ line congenital disorders [Bellus et al., 2000; Castro-Feijo et al., 2008; Zankl et al., 2008], mutant DNA was preselected by digestion with the restriction enzyme *BbsI*/*BpiI*. The recognition sequence of the restriction enzyme is GAAGAC. As a result, it cut wild-type DNA into two fragments (with a length of 2,448bp and 2,006bp) whilst the K650 mutant DNA stayed uncut in one single fragment (with a length of 4,454bp; see Figure 3.1). The next steps consisted of a gel electrophoresis, followed by a gel purification of the *BbsI*-resistant DNA to equally enrich for all *FGFR3* K650 codon substitutions. A PCR was run for 25 cycles using *DpnII*-16bp primers, 5' AACCTCGACTACTACA 3' and 3' GGGGTCATGCCAGTAG 5' for the forward direction and the backward direction, respectively. In order to get rid of wild-type

DNA present through incomplete digestion and to enrich for DNA with mutations at the positions 1948, 1949, and 1950, the entire process was repeated a second time. Instead of the previously utilised primers, the second round of PCR made use of nested primers containing unique 4bp ID tags to differentiate between patients. The primers of the resulting DNA sequences were linked to adapters to enable sequencing using an Illumina Genome Analyzer II.

Following this protocol, three individual sequencing libraries were created, each containing both sperm and blood samples. Two out of the three libraries were additionally spiked with a diluted heterozygous HCH3 DNA sample, showing a 1948 A>C substitution, to provide an internal standard enabling determination of the absolute mutation levels.

3.3.2 Titration Series

Genomic samples heterozygous for TD, *FGFR3* 1948A>G (TD3 and TD6), and SADDAN, *FGFR3* 1949A>T (SAD4), mutations were progressively diluted with normal blood DNA in order to be used in reconstruction experiments: 10μg of blood genomic DNA was mixed with a diluted series of mutant DNA corresponding to final mutation concentrations of 0 (no added mutant DNA), 3×10^{-6} (0.06 ng), 10^{-5} (0.2 ng), 3×10^{-5} (0.6 ng), 10^{-4} (2 ng), and 3×10^{-4} (6 ng), and taken through the same protocol as the sperm and blood samples (i.e. each dilution sample was mixed with three dilutions of the HCH3-spiked DNA). The dilution samples were analysed together with the sperm and blood samples.

3.3.3 Control DNA

Heterozygous mutant genomic samples from HCH3, SAD4, TD3, and TD6 patients as well as DNA of four wild-type individuals were included as controls in the analysis. The control DNA was amplified separately with the same protocol with the exception of the omission of the *BpiI* enzyme from all the incubation steps.

3.3.4 Sequencing

The two spiked DNA libraries (S1 and S2 spiked at a concentration of 10^{-5} and 10^{-4}, respectively) were sequenced at the same time on the same machine but on different lanes whereas the unspiked one (S8) was run on a different machine. The sequence runs generated 6,736,513, 8,820,153, and 2,623,294 reads for the three lanes S1, S2, and S8, respectively, with read lengths of 36bp each. Each read consisted of a 4bp long ID tag (individual for each patient), a 16bp shared *FGFR3* sequence specific primer (5' AACCTCGACTACTACA 3'), the K650 codon region as well as its

Figure 3.3: DNA sequence analysed on the Illumina GAII sequencer using the Gex-DpnII sequence primer (dark blue arrow). The unique tag sequence (XXXX, light blue) was obtained at sequencing cycles 1-4, the 16bp invariant primer sequence at cycles 5-20 (green), and the *FGFR3* K650 codon sequence (pink) at cycles 23-25.

neighbouring bases (see Figure 3.3). As the 16bp primer sequence was shared between all reads, it could be used as a reference to assess the read error structure in respect to error positions, error rates, and sequence-context preferences near erroneous base calls.

3.4 Knowledge-based Model to Estimate Mutation Levels

The read error structure was analysed using the knowledge of the 16bp sequence of the shared primer site in order to assess whether the gained information could be used together with Illumina's inbuilt quality scores to model the rate of *de novo* mutations at the K650 codon in the *FGFR3* gene. All available reads, apart from the ones that contained either an invalid patient tag or more than two mismatches compared to the reference sequence, were used, discarding a total of 3.7%, 1.9%, and 4.1% of all reads within the three libraries S1, S2, and S8, respectively.

3.4.1 Read Error Structure

In a first step, the per-base error rate was calculated. In concordance with previously published studies [Erlich et al., 2008; Illumina, 2010; Kircher et al., 2009; Lister et al., 2009; Rougemont et al., 2008; Shendure and Ji, 2008], error rates have been found to generally increase with read length (with higher per-base error rates towards the 3'-end, as depicted in Figure 3.4(a)). This observation is most likely caused by issues during the sequencing process: in Illumina sequencing, clusters of molecules are sequenced at once, requiring each cluster to give the same signal at the same time. As a consequence of this protocol, both negative frameshifts (due to an incorporation of nucleotides in each cycle which is not 100% complete) as well as positive frameshifts (due to an incomplete removal of the fluorophore which hinders the incorporation of new nucleotides in the next cycle) can result in more than one kind of fluorescent base to be present in one cycle which interferes with the signal interpretation. These shifts accumulate with a higher number of cycles, leading to an increased error rate at later stages of the sequencing process. Nevertheless, frameshifts fail to explain why particular positions in the middle of the sequenced reads were more prone to erroneous (or undefined) base calls than others (such as the second position in all three libraries or the ninth position in the S2 library).

From		Into				
		A	**C**	**G**	**T**	**N**
S1						
	A	-	0.80	0.13	0.06	0.01
	C	0.52	-	0.09	0.39	0.00
	T	0.09	0.55	0.36	-	0.00
S2						
	A	-	0.63	0.25	0.11	0.01
	C	0.39	-	0.13	0.41	0.07
	T	0.11	0.51	0.38	-	0.00
S8						
	A	-	0.64	0.26	0.09	0.01
	C	0.46	-	0.10	0.44	0.00
	T	0.14	0.48	0.37	-	0.01

Table 3.1: Nucleotide substitution rates for the three libraries, S1, S2, and S8. (Note that the substitution rate from guanine into any other base is not provided as the reference sequence only contained one guanine.)

To identify whether inaccurate reads shared common features, the sequence composition close to a mismatch was analysed (by using the bases flanking the error positions and, in the case of the first position of the 16bp sequence primer, the last base of the ID tag as preceding base). Although, there did not seem to be a very strong sequence-context preference, a slight bias towards a guanine or thiamine preceding an error was observed (see Figure 3.4(b)), the cause of which is unclear.

A closer look at the erroneous base calls revealed that base substitution errors were not equally frequent (e.g. the most frequent base to be changed into was a cytosine, preferentially substituting an adenine; see Table 3.1). One of the reasons for the observed bias in substitution errors might be the way signals are detected in the Illumina sequencing process: a green laser is used to detect the incorporation of guanine and thiamine while a red laser detects adenine and cytosine incorporation events. Different filters are required to enhance the brightness of one base type (e.g. guanine over thiamine) in order to distinguish the two different incorporation events. The transversions A>C, C>A, and T>G were among the most frequent ones, pointing to an insufficient discrimination of the respective base emission spectra. However, the signal detection process fails to explain the frequent transition of both C>T and T>C.

Furthermore, the distance between different errors was examined to assess whether a clustering of errors could be observed. Mismatches occurred more frequently in close proximity to each other, indicating that distance played a role in the read error structure. Taking all reads that contained exactly two errors into account, 68.46%, 42.97%, and 48.93% of the erroneous bases in the libraries S1, S2, and S8, respectively, were either at adjacent positions or separated by only one nucleotide (see Figure 3.4(c)). This observation is not very surprising considering that most errors were concentrated at the 3'-end of the read, thus, exhibiting a preference for small distances between each other.

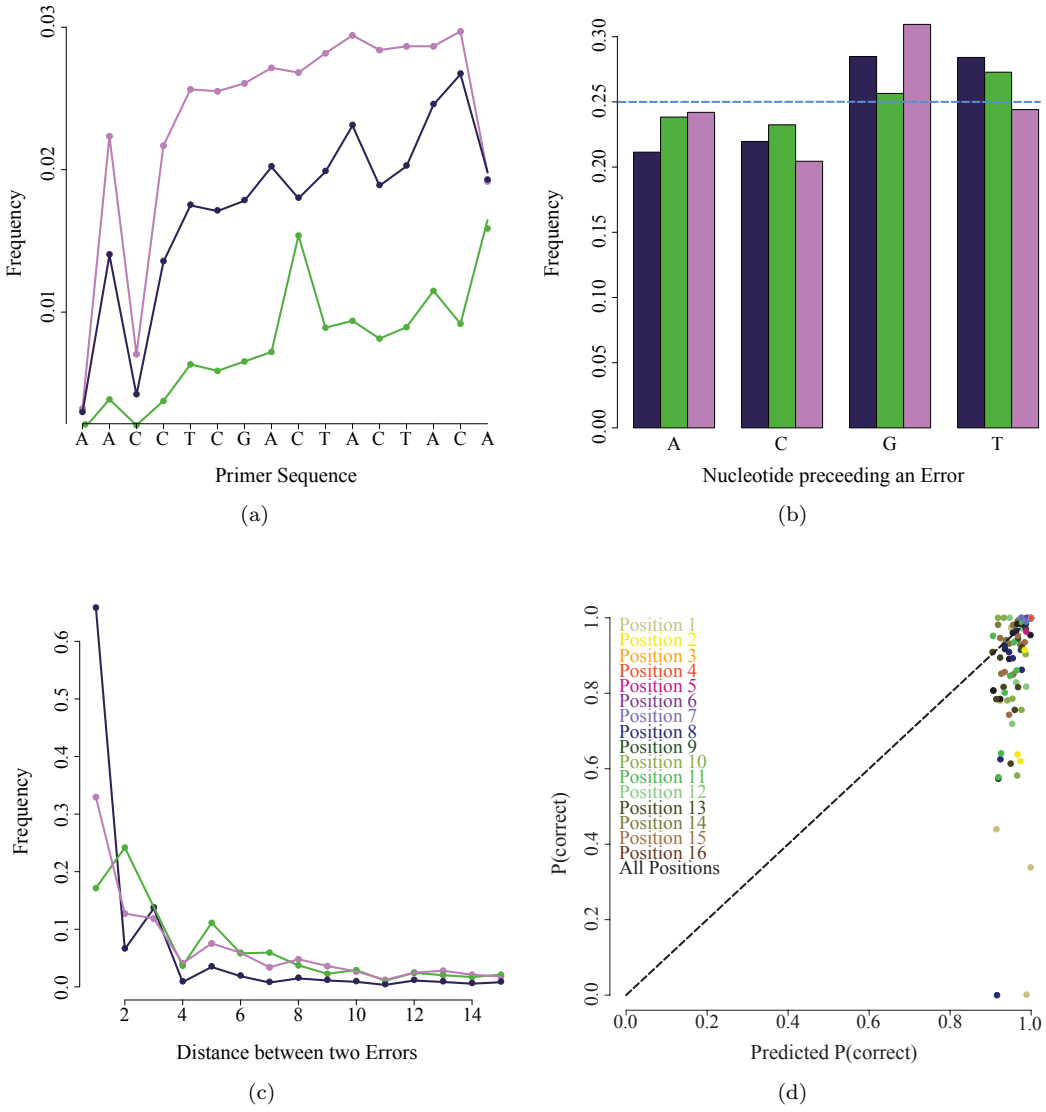


Figure 3.4: (a-c) Error structure of reads in the three libraries (S1: blue, S2: green, S8: pink). (a) Frequency of erroneous base calls along the 16bp shared sequence primer (5' AACCTCGACTACTACA 3'). (b) Frequency of nucleotides preceding an error in a read. The dotted blue line indicates the bar height which would have been expected under the assumption that all nucleotides are equally likely to precede an error. (c) Distance between two errors in a read (0 indicates that the erroneous base calls are next to each other). (d) Calibration of Illumina's inbuilt quality scores for each read position in the reference sequence of the 16bp shared sequence primer.

3.4.2 Quality Score Calibration

For each base call within a read the Illumina base caller, Bustard, reports the call quality by estimating a quality score in a way similar to the Phred score. Phred defines a quality value Q assigned to a base call to be

$$Q = -10 \log_{10}(P),$$

where P is the estimated error probability for that base call [Ewing et al., 1998]. Based on the image output (and without considering a reference sequence), Bustard estimates quality scores which are asymptotically identical to the Phred quality scores

$$Q = -10 \log_{10}\left(\frac{P}{1-P}\right).$$

The initial idea to use these inbuilt quality scores, together with the information gained about the read error structure, to estimate the rate of *de novo* mutations at the K650 codon in the *FGFR3* gene was neglected. The main reason was that the quality scores reported by the Illumina sequencer could not account for the underlying error structure. In particular, high quality values strongly underestimated the probability of an error which led to badly calibrated quality scores (as illustrated in Figure 3.4(d)). This resulted in a substantial amount of high quality values for wrong base calls in all three libraries (see Figure 3.5) which made it impossible to distinguish between a correct and an erroneous base call based on the inbuilt quality scores.

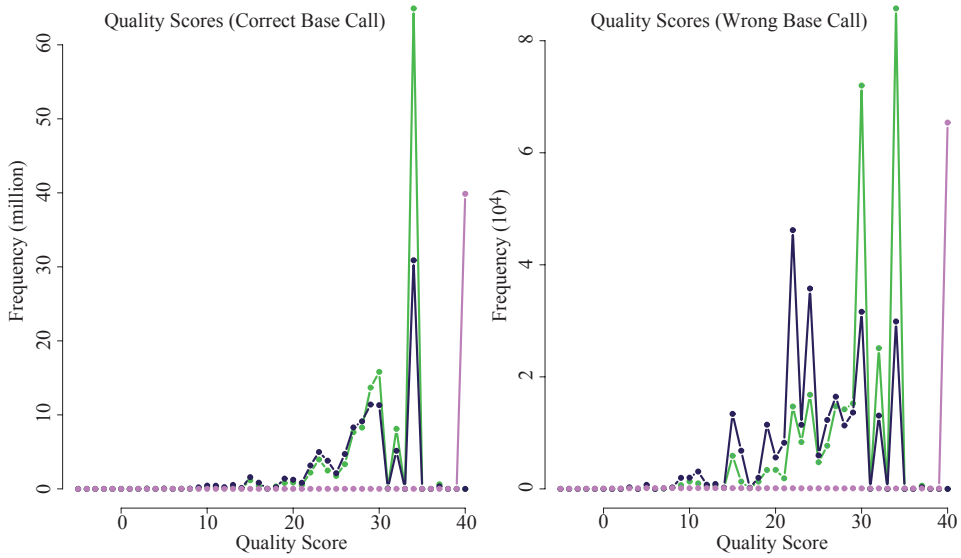


Figure 3.5: Illumina's base quality scores for correct and wrong calls in the three libraries (S1: blue, S2: green, S8: pink)

3.5 Black-box Model to Estimate Mutation Levels

As Illumina's inbuilt quality scores for the reads used in this study were too badly calibrated to fit a model based on the knowledge of the read error structure, a Bayesian approach was used instead to fit a model to the observed counts of each codon in the sequencing data, assuming that there was no *a priori* information available. Thereby, errors and biases derived from PCR amplification and enzymatic digestion during the sample preparation as well as from the sequencing process were modelled directly.

In the model, the normalised frequencies of the non-wild-type mutations after the addition of the spike at concentration s_j in experiment j ($s_0 = 0$, $s_1 = 10^{-5}$, $s_2 = 10^{-4}$) were assumed to be given by

$$f_{ij} = \frac{10^{z_i} + I_{i=\text{spike}}\, s_j}{s_j + \sum_i 10^{z_i}},$$

where z_i is the $\log_{10}$ frequency of mutation i in the sample and I is the indicator function that takes the value 0 or 1 depending on whether the mutation in question is the same as the spike. To account for the fact that the most common form of *FGFR3* mutation occurs at rates of up to 10^{-5} [Baujat et al., 2008; Waller et al., 2008], a normal(-8,2) prior for z_i was chosen for all non-wild-type species. As different lanes and machines on which the sequencing was carried out can affect the outcome of the experiment, each of the three experiments was allowed to have individual signal to noise parameters and biases. Specifically, it was assumed that mutation counts in the sequencing data were multinomial with frequency vector

$$x_j = \text{Dirichlet}(B_j f_{1j} + w_{1j}, B_j f_{2j} + w_{2j}, \dots),$$

where B_j is the signal to noise ratio for experiment j with a uniform(0,1000) prior and w_{ij} is the background for mutation i in experiment j with a uniform(0,100) prior.

The usage of a multinomial-Dirichlet model enabled the easy integration over x to obtain the contribution to the total likelihood from experiment j in a given sample

$$L_j = \frac{\Gamma(B_j + W_j)}{\Pi_i \Gamma(B_{ij} f_{ij} + w_{ij})} \times \frac{\Pi_i \Gamma(B_{ij} f_{ij} + w_{ij} + C_{ij})}{\Gamma(B_j + W_j + C_j)},$$

where

$$W_j = \sum_i w_{ij},$$

C_{ij} is the number of reads observed with mutation i in experiment j, and

$$C_j = \sum_i C_{ij}.$$

The likelihoods were combined across the samples and experiments to give a total likelihood. A Markov Chain Monte Carlo (MCMC) was used to sample from the posterior as dealing with multiple parameters made it difficult to calculate the normalising constant of the posterior analytically.

Markov Chain Monte Carlo Method

While an analytical expression for the posterior can be derived in simple cases, in many situations (e.g. when dealing with multiple parameters, multidimensional parameters, complex model structures, or complex likelihood functions), it is difficult to calculate the normalising constant of the posterior analytically. For this reason, Monte Carlo methods are used to sample from the posterior. One of the most widely used techniques to do this is the MCMC method.

By using the current sample values to randomly generate the next ones, MCMC constructs a Markov chain whose stationary distribution is the required posterior. A Markov chain is a sequence of random variables ($X_1, X_2, \ldots X_t$ where X_t denotes the value of the random variable at time t) generated by a random variable X whose transition probabilities between different values in the state space (the range of possible values for X) depend only on the variable's current state (known as a 'Markov process'):

$$P(X_t|X_1, X_2, \ldots X_{t-1}) = P(X_t|X_{t-1}).$$

The MCMC method has its roots in the Metropolis algorithm [Metropolis et al., 1953; Metropolis and Ulam, 1949] which was generalised by Hastings in 1970 [Hastings, 1970] and brought to prominence by Gelfand and Smith in 1990 [Gelfand and Smith, 1990].

Metropolis-Hastings Markov Chain Monte Carlo Method

A Metropolis-Hastings MCMC with sequential update of each parameter can be used to obtain the posterior density of the parameters X given the data D. Initially, a value of X_0 is chosen, typically from the prior $P(X)$. Then, the Metropolis-Hastings algorithm, consisting of the following steps, is repeated multiple times:

Step 1: A new set of parameters X' is proposed from a proposal distribution (also known as 'transition kernel') which is usually dependent on the current state in step i, X_i: $q(X_i|X')$.

Step 2: The proposal is excepted with probability

$$\alpha(X_i, X') = min\left(1, \frac{P(X')}{P(X_i)} \frac{P(D|X')}{P(D|X_i)} \frac{q(X'|X_i)}{q(X_i|X')}\right),$$

in which case $X_{i+1} = X'$, otherwise $X_{i+1} = X_i$.

Step 3: Increment i by 1.

The chosen proposal distribution greatly influences the mixing of the chain. A chain is said to be 'well mixing' if it explores the space well whilst it is said to be 'poorly mixing' if it stays in small regions of the parameter space for a long time. The mixing of a chain can be adjusted by tuning the SD of the proposal distribution. To increase the acceptance probability, the proposal SD is decreased. However, if the proposal SD is too small, moves are small and have a high acceptance rate but the chain is poorly mixing and values are highly autocorrelated. In contrast, too large values for the proposal SD lead to high autocorrelation and poorly mixing chains for which moves are large, but not often accepted. In both cases, much longer chains are required to reach stationary.

A well-constructed Markov chain, which is (i) irreducible (every state must be eventually reachable from every other state), (ii) aperiodic (to stop the chain oscillating between different states), and (iii) positive recurrent (the expected waiting time to return to a state is finite; once a chain reached its stationary distribution, it will stay in it), is guaranteed to have the posterior $P(X|D)$ as its stationary distribution [Brooks, 1998]. But how long the chain needs to run to reach stationary is not only dependent on (i) the proposal distribution (which may lead to low acceptance rates), but also on (ii) the initial starting point, and (iii) the likelihood surface (as local maxima might trap the chain).

Convergence Diagnostics

MCMC simulations have two shortcomings: (i) the distribution of the samples only converges with i to the posterior and (ii) the samples are typically highly correlated and thus, each sample is not an independent draw from the posterior. Early samples are strongly influenced by the initial distribution of X_i, which are poor representatives from the stationary distribution due to the fact that the chain usually starts far away from the target distribution. Therefore, the variance in the density estimated from a finite number of iterations of the chain can be reduced by removing iterations from the beginning of the chain, known as the 'burn-in' period.

To assess convergence, multiple runs from different initial starting points as well as graphical checks are used in combination with formal convergence tests (e.g. Geweke diagnostic [Geweke,

1992], Gelman and Rubin diagnostic [Gelman and Rubin, 1992], Yu and Mykland diagnostic [Yu and Mykland, 1998], Raftery and Lewis diagnostic [Raftery and Lewis, 1992], Heidelberger and Welch diagnostic [Heidelberger and Welch, 1981, 1983], Riemann sums [Philippe, 2001], multivariate bounds [Brooks and Roberts, 1999], and subsampling [Giakoumatos et al., 1999]; for an overview see Brooks and Roberts [1999]; Cowles and Carlin [1996]; El Adlouni et al. [2006]).

3.5.1 Parameter Estimation

A Metropolis-Hastings MCMC with sequential update of each parameter was used to estimate the parameters of interest. The work flow of the algorithm is shown below using B_j, the signal to noise ratio for experiment j, as an example:

$B_j = 10$	$\leftarrow$	*Step 1:* Start with an initial value.
$B_{j,new} = B_j + r_1$	$\leftarrow$	*Step 2:* Use the current value of B_j to sample a candidate point from a candidate-generating distribution (also referred to as 'proposal' or 'jumping' distribution) which is the probability of returning a value of $B_{j,new}$ given the value B_j. A proposal distribution based on a random walk chain is used, in which the new value $B_{j,new}$ equals the current value B_j plus some normally distributed random variable $r_1 \sim$ normal(0,1).
$\alpha = p(B_{j,new})/p(B_j)$	$\leftarrow$	*Step 3:* Compute the ratio of density ($p()$ denotes the probability density function) at the candidate point $B_{j,new}$ and current point B_j. Note that the unknown normalising constant cancels out as the ratio of $p()$ is considered under two different values.
if $(r_2 < \alpha) : B_j = B_{j,new}$	$\leftarrow$	*Step 4:* If the jump increases the density more than a variable r_2 drawn from a uniform distribution ($r_2 \sim$ uniform(0,1)), accept the candidate point; otherwise, reject it. Return to *Step 2*.

Multiple runs from different starting points were conducted to check for convergence and visual inspection was used to compare observed to expected values as a means of checking model adequacy. Time series traces (plots of the generated random variables versus the number of iterations) were investigated to check whether the chain was poorly mixing (also useful to assess the minimal required burn-in period for a starting value).

As an example, Figure 3.6(a) shows a time series trace of the first 28,000 iterations for the signal to noise ratio B for the experiment without spike addition. Visual inspection of both the trace plot and the density plot indicated that the beginning of the chain was badly mixing and the stationary distribution (which is also the target distribution) was only reached after a burn-in of ∼15,000 iterations. (In comparison, Figure 3.6(b) displays the well mixed chain after removal of the burn-in samples.)

Another way to assess convergence is to explore the autocorrelations between the draws of the Markov chain. Thereby, the lag k autocorrelation ρ_k is the correlation between every draw of the Markov chain x and its k-th lag:

$$\rho_k = \frac{\sum_{i=1}^{n-k}(x_i - \bar{x})(x_{i+k} - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

The k-th lag autocorrelation is expected to decrease as k increases. High autocorrelation for larger values of k indicate that a chain is slowly mixing. Figure 3.6(c) confirms that the chain was poorly mixing at the beginning while, after removal of the burn-in samples (see Figure 3.6(d)), convergence was reached.

A formal convergence diagnostic that can be used to check whether a Markov chain has reached stationary is Geweke's convergence test [Geweke, 1992]. After removing the burn-in period, it splits the sample in two parts (usually the first 10% and the last 50%). If the samples are drawn from the stationary distribution of the chain, the two means are equal and Geweke's statistic has an asymptotically standard normal distribution. A modified Z-test can be used to compare the two samples. It calculates the difference between the two sample means and divides it by its estimated standard error (which is estimated from the spectral density at zero such that it takes any autocorrelation into account). A value larger than two indicates that the mean is still drifting and a longer burn-in period will be needed. If samples from iteration 15,000-28,000 are considered from the example above, the Z-score (computed using the 'geweke.diag' algorithm from Gnu R's 'coda' package) is equal to 1.24, indicating that the burn-in period was sufficient to estimate the parameter of interest.

Another formal convergence diagnostic is the Heidelberg and Welch diagnostic [Heidelberger and Welch, 1981, 1983] which, based on the Cramer-von Mises test, calculates a run length control statistic to accept or reject the null hypothesis that the Markov chain is from a stationary distribution, based on a criterion of relative accuracy for the estimate of the mean. It is calculated in two parts: in the first part, the diagnostic is successively applied to the whole chain, then to the chain in which the first 10%, 20%, 30%, and 40% of the chain are discarded, until either the

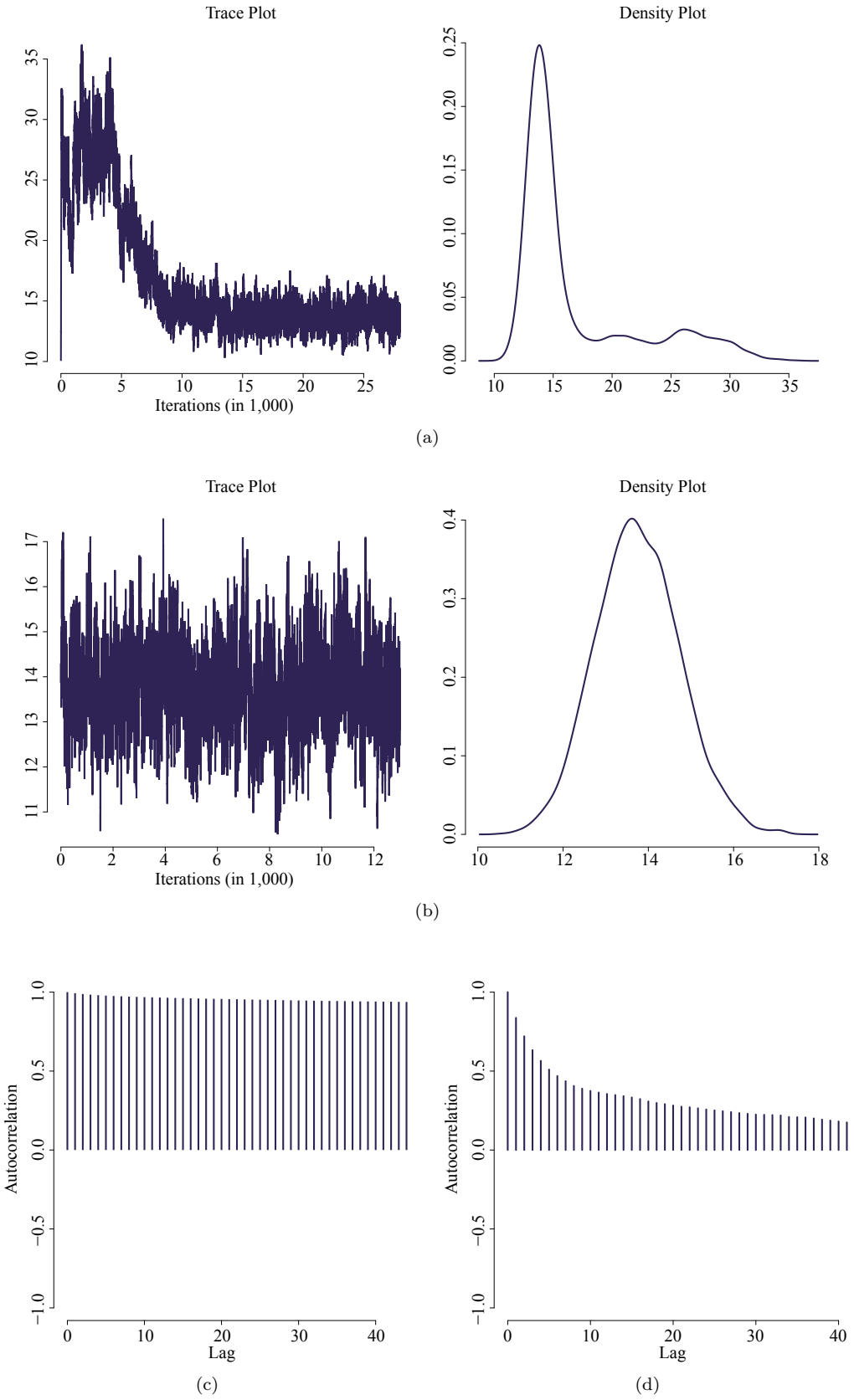


Figure 3.6: Time series trace plot and density plot for the signal to noise ratio B for the experiment without spike addition of a) the first 28,000 iterations (including the burn-in period) and b) after removal of the burn-in samples. Autocorrelations between the draws of the Markov chain c) run for 28,000 iterations (including the burn-in period) and d) after removal of the burn-in samples.

Library	Codon								
	CAG	**GAG**	**TAG**	**ACG**	**AGG**	**ATG**	**AAA**	**AAC**	**AAT**
S1	0.46	10.71	0.35	1.24	0.69	0.99	0.77	0.65	0.67
S2	5.44	3.22	0.44	0.88	0.72	0.73	0.77	0.50	0.67
S8	0.49	4.47	0.32	0.76	0.99	0.55	0.65	0.41	0.49

Table 3.2: Final posterior means for the background parameter W.

null hypothesis is accepted, or 50% of the chain has been discarded. If the test still rejects null hypothesis after 50% of the chain has been discarded, the chain fails the test and will need to be run longer. Otherwise, a half-width test calculates a 95% confidence interval (CI) for the mean, using the portion of the chain which passed the stationarity test. Half the width of this interval is compared with the estimate of the mean. If the ratio between the half-width and the mean is lower than some ϵ, the half-width test is passed. If the test fails, the length of the sample is deemed not long enough to estimate the mean with sufficient accuracy. The Heidelberg and Welch diagnostic was run using Gnu R's 'heidel.diag' algorithm (also from the 'coda' package) and both parts of the test statistic were passed.

The final posterior means for the signal to noise parameter B for the experiments S1, S2, and S8 were 34.75, 122.64, and 18.78, respectively, showing that this parameter varied greatly between the different libraries. The final posterior means for the background parameter W are provided in Table 3.2.

3.6 Analysis of Mutation Levels

Analysis of the reads obtained by high throughput sequencing can reveal whether there is any evidence that mutations in the K650 codon of the *FGFR3* gene might be under selective pressure. If the occurrence of the mutations is protein-driven (as in the *FGFR2* gene), background mutation rates should not be altered and only the mutations at the positions 1948-1950 should show disproportionally elevated levels in sperm (excluding the possibility that the whole region of the target *BbsI* site of *FGFR3* is highly mutable in general). At the same time, mutation levels within the blood samples should not show any selection for a mutation.

In the study, all reads with either apparent frameshift errors, unrecognised ID tags, or quality scores of less than 10 for any of the four ID tag bases or the three bases of the K650 codon were excluded, leaving 71.73%, 85.76%, and 86.86% of the original reads for the three libraries, S1, S2, and S8, respectively.

The three titration series (TD3, TD6, and SAD4), which were progressively diluted with normal blood DNA, were analysed together with biological samples to assess the accuracy and repro-

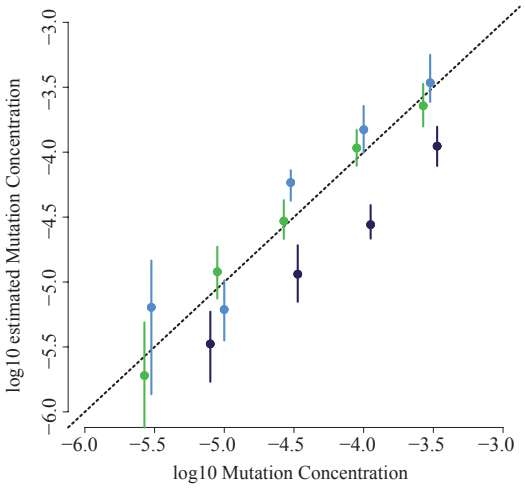


Figure 3.7: The mutation levels estimated by titration of three heterozygous control samples TD3 (light blue), TD6 (green), and SAD4 (dark blue) diluted in normal blood DNA.

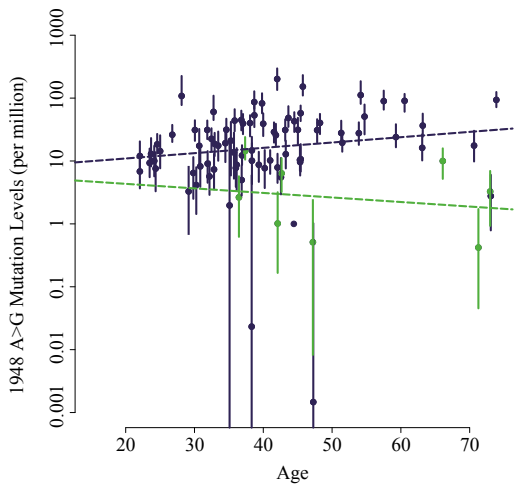


Figure 3.8: Estimated levels of the 1948A>G (K650E, TDII) mutation in blood (green, n = 8) and sperm (blue, n = 78) in relation to the age of the sample donor. Vertical bars indicate the extent of the 95% equal tail probability interval. Dotted lines indicate the regression lines.

ducibility of the proposed method. As illustrated in Figure 3.7, for the two TD samples, a good correspondence between the amount of input DNA and mutation levels was found at least down to a level of 10^{-5}. The measured levels of the SAD series were slightly underestimated (about 3.3-fold) but also strongly correlated with the amount of input DNA.

As expected, all blood samples showed relatively low levels of all mutations. In sperm samples, 73% of the total mutations quantified were caused by a 1948A>G (K650E, TDII) mutation. This mutation was estimated to be present at average levels of about 10^{-5} in sperm based on the overall birth prevalence of thanatophoric dysplasia of about 2.4×10^{-5} [Orioli et al., 1995; Waller et al., 2008]. However, Figure 3.8 shows clearly that it reached much higher mutation levels in sperm samples (with a maximum of 2.1×10^{-4}). The 1948A>G mutation rates have been found to be significantly correlated with the age of the sperm donor (Spearman rank r=0.34, P=0.002). Even if it was the most prevalent mutation, several other substitutions attained levels $> 10^{-5}$ in a minority of sperm samples (see Figure 3.9). The substitution 1949A>C accounted for 17% of all mutations in sperm. It was previously described in a small number of individuals with acanthosis nigricans and mild short stature [Berk et al., 2007; Castro-Feijo et al., 2008]. The two other substitutions that have been observed in constitutional disorders, 1950G>C and 1949A>T, accounted for total mutation levels >1%. In contrast, the three mutations encoding silent, conservative or stop changes, were the least abundant collectively accounting for <0.1% of sequences at the K650 codon. This prevalence of specific mutations may reflect a protein-driven, positive selection of mutant cells according to the functional consequences of the encoded amino acid substitution.

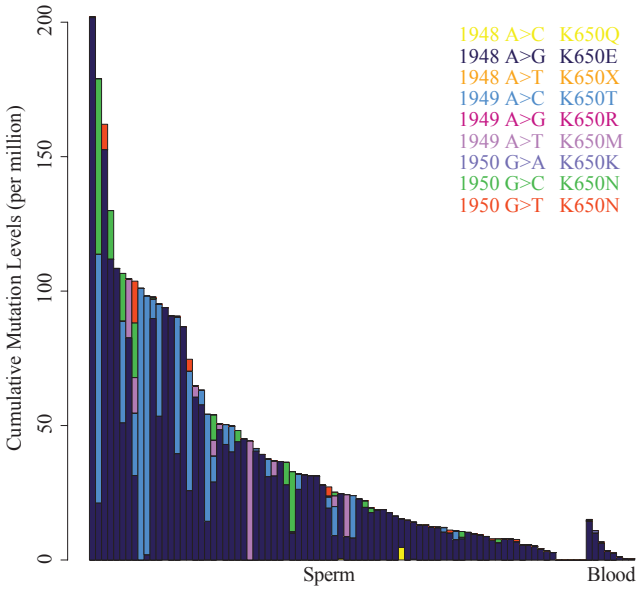


Figure 3.9: Cumulative levels of all nine different nucleotide substitutions at the K650 codon in sperm (left) and blood (right) samples.

3.7 Discussion

Estimating mutation rates from high throughput sequencing data is not an easy task as many different types of errors, biases, or uncertainties can occur in an experiment like the one described above which will need to be taken into account as they can strongly influence downstream analyses. They can be caused by a multitude of different factors including, but not restricted to (i) incomplete enzymatic digestion, (ii) PCR amplification which is not a 100% accurate, as well as (iii) the sequencing, (iv) base-calling, or (v) alignment process. In addition, estimations are further complicated by missing or low coverage data.

In Illumina sequencing, the dominant type of error are substitutions which are frequently clustered at the read end (see Figure 3.4(a)) as constant signal decay and dephasing give rise to increased levels of background noise. These substitutions are biased towards A/C and G/T errors (see Table 3.1) as these nucleotide pairs exhibit similar emission spectra when excited using the same laser, making their distinction using optical filters difficult. In concordance with previous work [Kircher and Kelso, 2010], this study identified error rates in the range of up to 3% per nucleotide (as illustrated in Figure 3.4(a)).

As problems that arise in NGS data are subtle but complex, advanced statistical models are an important tool for the analysis of the data as they allow to capture features and uncertainties of library-, machine-, and run-dependent error structures. However, as this study showed, building statistical models on the knowledge of the error structure of high throughput sequencing data is difficult due to the fact the inbuilt quality scores are not well calibrated. Black-box models, assum-

ing that there is no *a priori* information available, can help to overcome this issue, but extensive validation studies for different libraries, NGS machines, and sequenced codons (as described in Section 3.6) are needed to confirm obtained results.

To enhance the current knowledge of mutation and disease, more extensive molecular data on the human germ line is required. The performance of this work could show that advanced statistical methods can be used to estimate mutation rates in human sperm from high throughput sequencing data down to a level of 10^{-5} whilst capturing subtle features of the library-, machine-, and run-dependent error structure. However, it should be noted that very few mutations in humans have rates as high as 10^{-5} (such as point mutations in the *FGFR2* or *FGFR3* gene which represent some of the most extreme examples of male mutation bias in the human germ line [Crow, 2000]). The majority of spontaneous mutations occurs at much lower rates (previous analyses of DNA sequence data estimated the average substitution frequency per nucleotide site to be in the order of $10^{-7} - 10^{-9}$ [Arnheim and Calabrese, 2009; Kondrashov, 2003; Nachman and Crowell, 2000]) and thus, the discussed model will unfortunately not be suited for the analysis of most other human germ line mutations.

Preface

Work from Chapter 4 has been published as:

> Auton*, A., Fledel-Alon*, A., Pfeifer*, S., Venn*, O., Ségurel, L., Street, T., Leffler, E.M., Bowden, R., Aneas, I., Broxholme, J., Humburg, P., Iqbal, Z., Lunter, G., Maller, J., Hernandez, R.D., Melton, C., Venkat, A., Nobrega, M.A., Bontrop, R., Myers, S., Donnelly†, P., Przeworski†, M., and McVean†, G., *A Fine-Scale Chimpanzee Genetic Map from Population Sequencing*, Science 2012, Vol. 336, No. 6078, pp. 193-198.
>
> * These authors contributed equally to the project.
>
> † These authors jointly supervised the project.

The study was designed by Peter Donnelly, Molly Przeworski, and Gil McVean. Biological samples were obtained from and prepared by Ronald Bontrop and Rory Bowden. High throughput sequencing was performed by High Throughput Genomics at the Wellcome Trust Centre for Human Genetics (as documented in Section 4.2) and data was processed by John Broxholme and Gerton Lunter. Genotyping and phasing was conducted by Julian Maller (Section 4.3.4). The validation experiment (Section 4.3.5) was designed and performed by High Throughput Genomics. The validation data have been analysed by Adam Auton, Peter Humburg, and myself. The SNP discovery power was assessed by Adam Auton (as described in Section 4.3.6), and I conducted all other analyses with input from Adam Auton and Gil McVean.

Chapter 4

PanMap - Creation of a Map of Genome-wide Sequence Variation in Western Chimpanzees

Anthropogeny, the investigation of the origin of humans (Oxford Engl. Dict. 2006), and the philosophical question of what makes humans human and differentiates us from great apes, are not only big scientific challenges, but also a subject of interest for most of us. Chimpanzees are our closest living evolutionary relatives [Goodman, 1999] with whom we share a last common ancestor ∼5-7 million years ago (Mya) according to both molecular and fossil data [Brunet et al., 2002; Chen and Li, 2001; Varki and Altheide, 2005]. Based on morphological and geographical criteria, at least five distinct chimpanzee subspecies can be distinguished [Becquet et al., 2007; Gagneux et al., 1999; Keele et al., 2006; Stone et al., 2002]. They comprise bonobos (*Pan paniscus*) and four extant common chimpanzee subspecies (*Pan troglodytes*) from which bonobos separated ∼1 Mya [Becquet and Przeworski, 2007; Won and Hey, 2005]: Western (*Pan troglodytes verus*), Nigerian-Cameroon (*Pan troglodytes vellerosus*), Central (*Pan troglodytes troglodytes*), and Eastern (*Pan troglodytes schweinfurthii*) populations [Caswell et al., 2008] (see Figure 4.1). To understand human and chimpanzee evolution and to analyse what distinguishes us evolutionary and biomedically from chimpanzees, extensive comparative genetics is required to catalogue genetic changes that have accumulated since the two populations diverged from their most recent common ancestor.

First informations about the genetic code of chimpanzees were obtained by large-scale efforts including the sequencing of more than 100,000 reads from a chimpanzee BAC library [Fujiyama et al., 2002] as well as the sequencing of more than 200,000 protein-coding exons [Clark et al., 2003] followed by more specific studies sequencing the major histocompatibility complex (MHC) class I region [Anzai et al., 2003], the chimpanzee orthologue of human chromosome 21 [The International Chimpanzee Chromosome 22 Consortium, 2004], and parts of the Y chromosome [Rozen et al.,

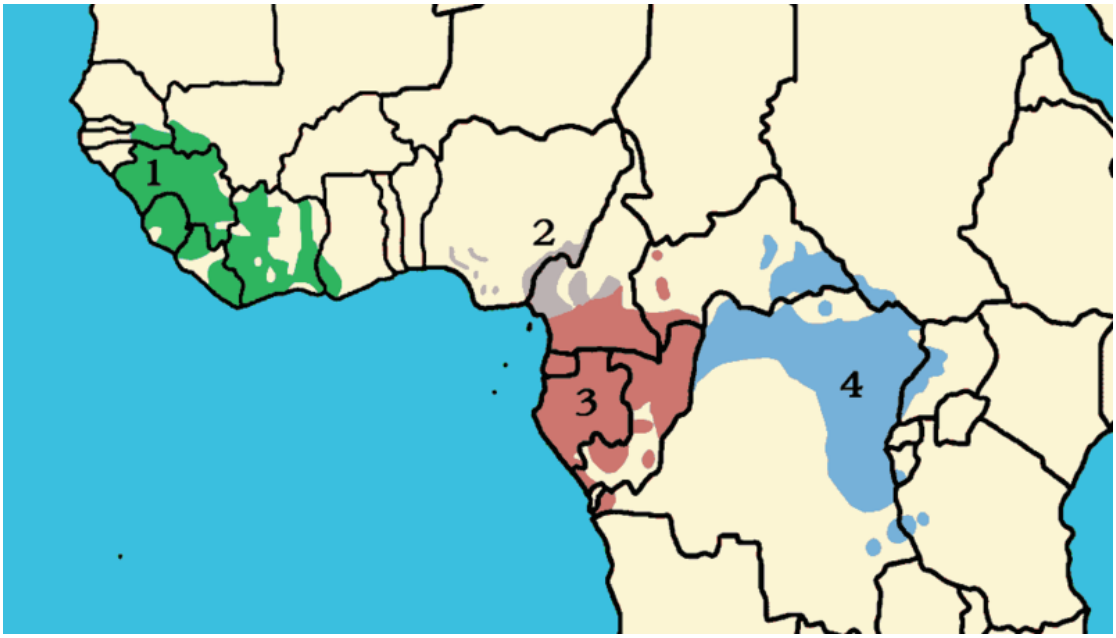


Figure 4.1: Distribution of the common chimpanzee: (1) *Pan troglodytes verus* domiciled in Guinea, Mali, Sierra Leone, Liberia, Cote d'Ivoire, Ghana, and Nigeria, (2) *Pan troglodytes vellerosus* domiciled in Nigeria and Cameroon, (3) *Pan troglodytes troglodytes* domiciled in the Central African Republic, Cameroon, Equatorial Guinea, Gabon, the Democratic Republic of the Congo, and the Republic of the Congo, and (4) *Pan troglodytes schweinfurthii* domiciled in the Central African Republic, Sudan, the Democratic Republic of the Congo, Uganda, Rwanda, Burundi, Tanzania, and Zambia. This figure was created by Luis Fernández Garcìa under the Creative Commons Attribution-Share Alike 3.0 Unported license (based on information by Keele et al. [2006]).

2003]. But it was not until 2005 that the first rough draft sequence of the entire chimpanzee genome was released by the Chimpanzee Sequencing and Analysis Consortium [Chimpanzee Sequencing and Analysis Consortium, 2005]. The analysed sequences mainly originated from a single captive-born male Western chimpanzee, called Clint, from the Yerkes National Primate Research Center [8]. The draft was produced by a whole-genome shotgun sequencing to a BAC library followed by an assembly of 361,782 contigs with a median length of 15,700 bases using the ARACHNE algorithm [Batzoglou et al., 2002; Varki and Altheide, 2005]. To facilitate contig linking, the assembly made use of the finished human genome sequence (NCBI build 34) [International Human Genome Sequencing Consortium, 2001; International Human Genome Sequencing Consortium, 2004]. The draft sequence of the chimpanzee genome has a ∼3.6-fold redundancy in sequencing reads from autosomes (sex chromosomes have half that redundancy in males). It covers ∼2.7 billion nucleotides (corresponding to ∼94% of the entire chimpanzee genome) of which 98% are high quality bases with a quality score of at least 40 (corresponding to an estimated error rate of $\leq 10^{-4}$). However, by largely assembling the chimpanzee genome through alignment to the human genome, gaps as well as biases might have been introduced and there is a risk that segments of the genome have been misassembled. The latter is a particular problem in rapidly evolving segments of the genome such as regions that have either been duplicated in one of the two lineages or that have been selectively

[8] `http://www.yerkes.emory.edu/`

deleted in humans [Olson and Varki, 2003]. The initial ∼3.6X coverage was increased to ∼6X due to the sequencing efforts from the Washington University Genome Sequencing Center [9] which produced an additional ∼2X whole genome coverage by using a combination of whole genome plasmid reads, fosmid and BAC end sequences.

As a draft sequence, the chimpanzee genome still contains gaps and regions of uncertainty. Due to this fragmentary nature of the genome, only ∼2.4Gb of the first ∼3.6X coverage high quality sequence (including 89Mb from chromosome X and 7.5Mb from chromosome Y) could be mapped to the human reference genome (corresponding to ∼80% of the human genome) [Chimpanzee Sequencing and Analysis Consortium, 2005]. Based on this alignment, a study of the Chimpanzee Sequencing and Analysis Consortium confirmed previous estimates that the chimpanzee genome differs from its human counterpart in ∼4% [Chimpanzee Sequencing and Analysis Consortium, 2005; The International Chimpanzee Chromosome 22 Consortium, 2004]. Circa 1.23% result from ∼35 million SNPs. The other ∼3% yield from ∼5 million events that caused differences in the length of two homologous sequences: in total ∼90 million nucleotides of insertions and deletions of which 40-45 million occurred in each lineage. Estimations that correct for the coalescence times in the two species suggest that the fixed divergence rate in SNPs is ∼1.06% or less (as polymorphisms account for 14-22% of the observed divergence rate) whilst the remainder represents species-specific variations [Chimpanzee Sequencing and Analysis Consortium, 2005].

The reliable detection of intra-specific genetic variations requires the availability of genomic sequences from different individuals in a population. This problem is not acute for humans as millions of SNPs that vary from one individual to another have already been detected through both the International HapMap Project [10] and the 1000 Genomes Project [11]. The HapMap Project mapped more than 8 million SNPs of all 9-10 million common SNPs with a MAF ≥ 0.05 in the assembled human genome, catalogued the correlation patterns between nearby variants, and added this information to the public dbSNP database [Mardis, 2011; The International HapMap Consortium, 2005, 2007]. The 1000 Genomes Project on the other hand aims to characterise over 95% of variants with a MAF ≥ 0.01 in genomic regions of ∼1,000 individuals worldwide (ensuring representation of African, American, Asian, and European populations) [The 1000 Genomes Project Consortium, 2010]. However, the problem remains in the population of our closest evolutionary relative as, in spite of its importance, much less polymorphic variation data is available. For this reason, a logical next step is the sequencing of genomes of different chimpanzee individuals to study their genetic diversity on a genome-wide scale.

[9] http://genome.wustl.edu
[10] http://www.hapmap.org/
[11] http://www.1000genomes.org/

4.1 Aim of the Study

Over the last decades, sequencing has led to a better understanding of the human genome and its variation. In contrast to the vast amounts of data on intra-specific genetic diversity of humans, there is only limited polymorphism data and no genetic map available for the chimpanzee, despite its evolutionary significance. Hence, the aim of this study is to create a publicly available map of genome-wide sequence variation in Western chimpanzees. This resource will enable (i) the detection and analysis of polymorphism patterns (including copy number variation) in chimpanzees, (ii) the detection of recombination hotspots and coldspots, (iii) the construction of fine-scale genetic maps, (iv) the examination of evolutionary forces that drove recent specification through the comparison of evolutionary patterns in particular genes of interest, (v) the genome-wide comparative study of human and chimpanzee diversity patterns which might shed light on uniquely human pathogenic mechanisms of disease related to (mal)adaptation during our recent evolutionary past, (vi) the inference of selection in the chimpanzee, and (vii) the improvement of the chimpanzee reference sequence. Low quality nucleotide positions yielding from the variable quality of the draft genome sequence and its assembly process might have resulted in sites that are falsely divergent between humans and chimpanzees. The data generated by sequencing multiple chimpanzee individuals might help to identify and eliminate those sites.

4.2 Study Design

DNA obtained from Epstein-Barr virus (EBV) immortalised cell lines of 10 unrelated Western chimpanzees (nine females and one male) housed at the Biomedical Primate Research Center [12] in the Netherlands was sequenced to 5-11X coverage using paired end sequencing with 50bp reads. Their names/accessions were

- ANNACLARA (PtAC), ♀
- FRITS (PtFR), ♂
- GINA (PtGI), ♀
- LADY (PtLA), ♀
- LIESBETH (PtLI), ♀
- PEARL (PtPE), ♀
- REGINA (PtRG), ♀
- RENEE (PtRN), ♀
- SUSIE (PtSU), ♀
- YVONNE (PtYO), ♀

The rational behind the choice of the study design was a tradeoff between coverage and sample size: 5-11X coverage should give rise to accurate genotype calls and a sample size of 10 individuals should enable obtaining useful LD information to learn about evolutionary processes. As sequencing platform, Illumina's Genome Analyzer GAII was used. One reason for this choice was that

[12] http://www.bprc.nl/

capillary-based sequencing is not only more expensive, but also detection of variation in samples is difficult due to the fact that multiple templates originate from the same PCR product or clone. Thus, a significant representation of each allele in the fragment population is necessary to reliably distinct multiple base calls at a single position. In contrast, NGS technologies enable a deep single nucleotide and indel variant analysis of each fragment individually. However, as error rates for individual next-generation reads are higher than those for traditional Sanger sequenced ones, accuracy can only be achieved by sequencing the same region multiple times, resulting in a higher coverage, and calling a consensus sequence. The Illumina platform produces millions of high-quality reads per run making it well suited for the detection of variation in genomes through resequencing. An additional problem is the short length of NGS reads as their mapping to a reference genome is more difficult (reads might align equally well to multiple locations as most genomes contain repetitive regions on the length scale of the reads). To aid the alignment of short read sequences to a reference genome, sequence information from both ends of a linear fragment (paired end sequencing; paired end reads are separated by ∼300-500bp) can be obtained by Illumina's technology which can be used to learn about the relative position and orientation of a pair of reads in the genome.

In order to use the Illumina chemistry, random paired end shotgun libraries were prepared using adaptive focused acoustics (Covaris) for DNA fragmentation. Samples were fragmented to 500bp giving an expected 450bp insert size. The libraries were sequenced individually on the Illumina GAII platform using standard protocols (a combination of v2 paired end cluster generation kits with v3 cycle sequencing kit and v6 GA recipe and v4 paired end cluster generation kits with v4 cycle sequencing kit and v7 GA recipe). Initial data processing was conducted using the default Illumina software. Sequence reads were given in FASTQ format (see Appendix B).

4.3 Identification of Genomic Variation

The identification of variation in genomes strongly relies on the precise alignment of sequenced reads to a reference genome as well as on accurate SNP calling to avoid errors by calling positional variants which are produced by misaligned reads or sequencing errors. For this reason, the analysis to detect and genotype genomic variants can be divided into four steps: (i) sequence reads need to be aligned to a reference genome in order to identify positions or regions at which at least one of the samples differs from the reference sequence (discovery), (ii) quality metrics are required to remove candidate locations which are likely to be false positives (filtering), (iii) individual alleles at all variant sites are estimated (genotyping and phasing), and (iv) a subset of the called variants is assessed using an independent technology in order to estimate the FDR (validation).

4.3.1 Discovery - Step 1: Alignment

The alignment of reads to a reference genome is the first and most essential step in most sequence data analyses. It not only allows to rapidly determine whether the biological and sequencing experiments have succeeded, but also correct alignments are the prerequirement to detect true genomic polymorphic sites.

Alignments are mainly influenced by the repeat structure of the reference sequence (as alignment algorithms have difficulties to find the correct mapping of the read), the read base quality (as low quality reads might lead to false hits), and the aligner sensitivity (as true hits can be missed if the tool is not sensitive enough). Paired end sequences were generated by the Illumina GAII as they can help to increase the reliability of an alignment. Their major advantage is their ability to compensate for problems arising through short read lengths. For example, it is impossible to accurately place a single read in a repetitive region that is longer than the read itself, but as long as one of the paired ends can be mapped with confidence, the inferred insert size can be used to map its mate. In addition to increasing the mapability of reads and the reliability of alignments, read pairs can also help to correct wrong alignments and to identify small insertions or deletions.

Strategy

At first, two different alignment strategies were used to align short reads to the chimpanzee reference genome: MAQ (Version 0.7.1) [Li et al., 2008a] (the alignment protocol is given in Appendix C) and Stampy (Version 0.95) [Lunter and Goodson, 2011] with BWA (Version 1.5.6) [Li and Durbin, 2009] as a premapper (the alignment protocol is described below). A comparison of the mappings for an entire run (seven lanes, more than 120 million paired end reads) showed that >99.9% of the reads were aligned to exactly the same location by the two algorithms. Due to the fact that Stampy combined with BWA as a premapper was, with a similar sensitivity, ∼10-times faster than MAQ, the entire read data set was aligned only utilising the combination of Stampy and BWA to speed up the mapping process.

In addition, all reads were aligned to a human reference sequence in order to identify SNPs mapping in places with reliable synteny between the two species as preliminary results of the study suggested that the systematic incompleteness of the chimpanzee reference genome in repetitive regions might have a significant effect on the alignment process.

Work Flow Each lane was assigned a unique read-group tag (see SAM format specification in Appendix D.1) before it was individually aligned to the reference sequence given in FASTA format (see Appendix E) using BWA to speed up the alignment process. Only reads that

either aligned with two or more mismatches or that did not align at all were later remapped using Stampy. BWA needed the reference genome to be BWT-indexed

```
bwa index -a bwtsw reference.fasta,
```

where -a defines the algorithm used for the construction of the index (bwtsw is an algorithm which generally works well with whole genomes) whilst Stampy required the generation of a genome file (extension .stidx)

```
stampy.py -G chimp reference.fasta,
```

as well as a hash table (extension .sthash)

```
stampy.py -g chimp -H chimp.
```

After the genome file and the hash table were created and the reference was indexed, a paired end alignment was performed

```
stampy.py -g chimp -h chimp \
          -M sequence1.txt,sequence2.txt \
          --bwaoptions=reference \
          --solexa \
          -o alignment.sam,
```

where -g specifies the genome file, -h indicates the hash table, -M represents the read data that should be mapped (sequence1.txt and sequence2.txt contain an end of the paired end reads each with reads sorted in the same order), - -bwaoptions specifies the location of the BWT index, and - -solexa indicates that the data was obtained from an Illumina platform. As per default, all reads were represented in the output which was given in the standard alignment SAM format. An overview of Stampy's work flow is given in Figure 4.2.

Reference Genomes Reads were aligned to both the genome reference panTro2 (Build 2 Version 1) chimpanzee 6X draft assembly (produced by the Chimpanzee Sequencing and Analysis Consortium and downloaded from the UCSC [13]) and to the human genome assembly hg18 (Build 36.1, assembled by the International Human Genome Project sequencing centres and downloaded from the UCSC [14]) including all chromosomes and unlocalised contigs. Both reference sequences were given in FASTA format. The pseudoautosomal region 1 (PAR1), comprising the 2.6Mb beginning tip of the short arm of the two sex chromosomes in humans

[13] http://hgdownload.cse.ucsc.edu/downloads.html#chimp
[14] http://hgdownload.cse.ucsc.edu/goldenPath/hg18/bigZips/

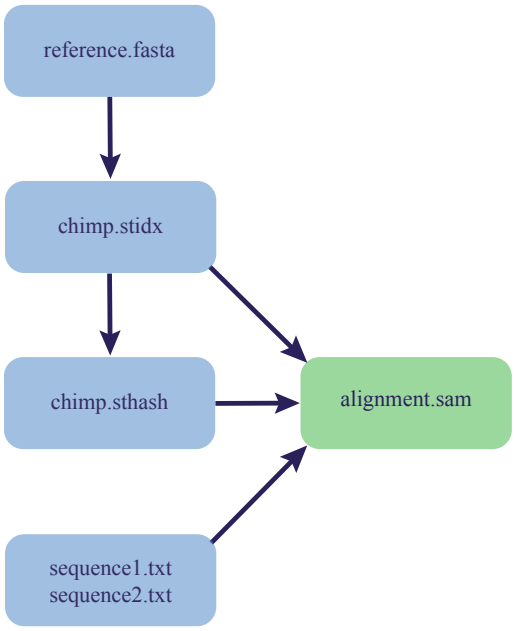


Figure 4.2: Stampy work flow. Step 1: Generation of a genome file (extension .stidx) from the reference sequence (given in FASTA format). Step 2: The genome file is used to build a hash table (extension .sthash). Step 3: The genome file and the hash table are used together with the reads in Illumina FASTQ format (sequence1.txt and sequence2.txt) to perform an alignment, the results of which are given in SAM format.

and chimpanzees [Graves et al., 1998; Rappold, 1993; Raudsepp and Chowdhary, 2008], was masked out on the Y chromosome. In addition, the human-specific pseudoautosomal region 2 (PAR2), comprising the beginning 320kb of the long arm telomere [Freije et al., 1992], was masked out on the human Y chromosome. The two sex chromosomes recombine with each other in the PAR making it a region of true sequence homology between the mammalian X and Y chromosomes. Therefore, if the PAR is not masked in one of the two sex chromosomes, mapping algorithms have problems to align reads with confidence as it appears that there are two possible locations to map the read in the genome. Furthermore, the genome of the EBV, a human herpes virus (human herpes virus 4), was included in the reference. It was used to generate the cell lines from Chimp B lymphocytes (see Section 4.2) resulting in the possibility of 1-100 copies being present in each chimp genome. The EBV standard genome reference is 170kb in size and was downloaded from NCBI [15].

Project Status For the 10 chimpanzee individuals, ∼7.3 billion paired end 50bp reads from 130 lanes (between 12 and 15 lanes per individual) were sequenced using the Illumina GAII and mapped to both the chimpanzee reference genome, panTro2, and the human reference genome, hg18, using Stampy with BWA as a premapper. Table 4.1 shows a summary of the read data.

[15] http://www.ncbi.nlm.nih.gov/sites/entrez?Db=genome&Cmd=ShowDetailView&TermToSearch=19000

Individual	Statistics					
	# Lanes	# Reads^{+}	# Reads*	# Reads*,**	# Proper Pairs*,**	Coverage^{*,**,PP}
PtAC	13	755,214,042	739,574,521	680,743,052	602,246,143	9.96
PtFR	12	739,493,468	721,224,973	629,498,454	597,076,426	9.87
PtGI	14	800,564,548	786,619,880	713,686,396	624,637,156	10.33
PtLA	12	657,217,022	636,303,660	570,370,084	531,965,056	8.80
PtLI	12	683,332,122	667,104,520	630,693,891	588,054,076	9.72
PtPE	13	634,285,664	623,484,379	596,129,401	556,799,292	9.20
PtRG	15	953,373,582	927,922,201	736,815,238	682,799,953	11.29
PtRN	13	733,738,636	657,515,687	384,416,097	337,073,364	5.57
PtSU	14	689,717,116	681,007,137	661,606,036	612,401,674	10.12
PtYO	12	647,744,260	637,837,527	616,132,517	582,447,980	9.63

$^{+}$ = As Stampy maps all reads to the reference genome, this column also represents the raw read data.
* = without lane-level duplicates, ** = without library-level duplicates, ^{PP} = proper pairs

Table 4.1: For the 10 chimpanzee individuals, 130 lanes (12-15 lanes per individual) were sequenced. They were mapped to both the chimpanzee reference genome, panTro2, and the human reference genome, hg18, using Stampy with BWA as a premapper. On average ∼22% of all reads per individual were removed as they were either lane-level duplicates (∼3%) or library-level duplicates (∼12%) or they did not qualify as proper pairs (∼8%) resulting in a mean coverage of 9.45X. For the calculation of the coverage, the genome size was taken to be 3.06Gb.

4.3.2 Discovery - Step 2: SNP Calling

Multiple steps are necessary to perform reliable and accurate SNP calling and to avoid errors by calling positional variants which are produced by misaligned reads or sequencing errors.

Duplicate Removal Before a SNP call, the mapping data obtained from different Illumina lanes should be grouped by individual to enable the removal of duplicates in the alignments as they manifest themselves as high read coverage support (see Section 4.3.2.1). Mapping duplicates are read pairs with identical outer coordinates. They can either be lane-level duplicates resulting from the amplification step in the sample preparation or library-level duplicates.

Base Quality Score Recalibration Reported quality scores only inaccurately display the actual probability of mismatching the reference genome. For this reason, a base quality score recalibration should be carried out to adjust reported quality scores (Section 4.3.2.2).

Base Alignment Quality (BAQ) Alignment algorithms map individual reads to the reference position from which they most likely originated. As SNPs occur much more frequently than indels in the genomes of most species (e.g. ∼35 million SNPs compared to ∼5 million insertion and deletion events occurred in the chimpanzee genome [Chimpanzee Sequencing and Analysis Consortium, 2005]), most algorithms rather annotate SNPs than indels at positions mismatching the reference genome. At positions of true insertions or deletions, alignment artefacts can result in false positive SNP calls [Li and Homer, 2010]. One possibility to identify those positions and to improve SNP calls is multiple sequence alignment which locally realigns reads such that the number of mismatching bases is minimised across all the reads. However, current implementations are computationally very intense and filtering

SNPs around predicted indels is hindered by the challenge of correctly identifying new indels in the first place. Another way to avoid those artefacts is implemented in SAMtools' 'calmd' algorithm [Li et al., 2009a] which calculates a Base Alignment Quality [Li, 2011] on a per-base level (Section 4.3.2.3). It down weights base qualities in regions with high ambiguity in the local alignment, thereby avoiding an excess of false positive SNP calls in low-complexity regions of the genome. The newly assigned base qualities can then be used in the initial SNP call.

Initial SNP Call An initial SNP call is performed (Section 4.3.2.4) and its output evaluated and filtered to allow a distinction between true variations and false positives (Section 4.3.3).

An overview of the SNP calling work flow is given in Figure 4.3.

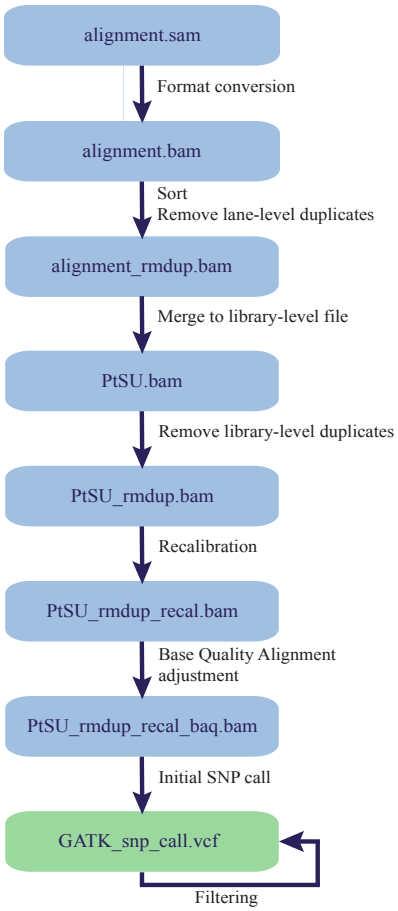


Figure 4.3: SNP calling work flow. Preliminary Step: The alignment file is converted from SAM format to BAM format. Step 1: The alignment is sorted by the leftmost coordinate to enable the removal of lane-level duplicates. Step 2: Duplicate free lane-level files for single chimpanzee individuals (e.g. for Susie, PtSU) are merged to a single library-level file. Step 3: Library-level duplicates are removed. Step 4: Base quality scores are recalibrated. Step 5: A Base Alignment Quality calculation is performed to avoid an excess of false positive SNP calls in low-complexity regions of the genome. Step 6: Raw variants containing non-reference bases as well as genotypes are obtained after an initial SNP call. Step 7: The raw variants are filtered to separate true segregating variation from machine artefacts.

4.3.2.1 Step 2.1: Duplicate Removal

Prior to the initial SNP call, both lane-level duplicates, resulting from the amplification step in the sample, and library-level duplicates were removed to improve the overall quality of the SNP calls. For this purpose, the data was processed using two tools, SAMtools [16] (Version 0.1.12a) [Li et al., 2009a] and Picard [17] (Version 1.14) which comprise command-line utilities that manipulate alignments in SAM or BAM format (see Appendix D.1 and D.2).

Work Flow The first step in the duplicate removal work flow was the indexing of the reference sequence (resulting in a reference.fasta.fai file) to allow its faster access

```
samtools faidx reference.fasta.
```

Alignments stored in SAM format were then converted to BAM format

```
samtools import reference.fasta.fai alignment.sam alignment.bam
```

to enable sorting them with Picard by the leftmost coordinate

```
java -jar SortSam.jar \
     INPUT=alignment.bam \
     OUTPUT=alignment_srt.bam \
     SORT_ORDER=coordinate,
```

before removing lane-level duplicates (only retaining the pair which exhibits the highest mapping quality)

```
samtools rmdup alignment_srt.bam alignment_rmdup.bam.
```

Duplicate free lane-level BAM files for single chimpanzee individuals (e.g. for Susie, PtSU) were merged to a single library-level file using Picard

```
java -jar MergeSamFiles.jar \
     INPUT=alignment1_rmdup.bam \
     INPUT=alignment2_rmdup.bam ... \
     OUTPUT=PtSU.bam.
```

To improve the overall accuracy of SNP calls, library-level duplicates were also removed

```
samtools rmdup PtSU.bam PtSU_rmdup.bam.
```

[16] http://samtools.sourceforge.net/
[17] http://picard.sourceforge.net

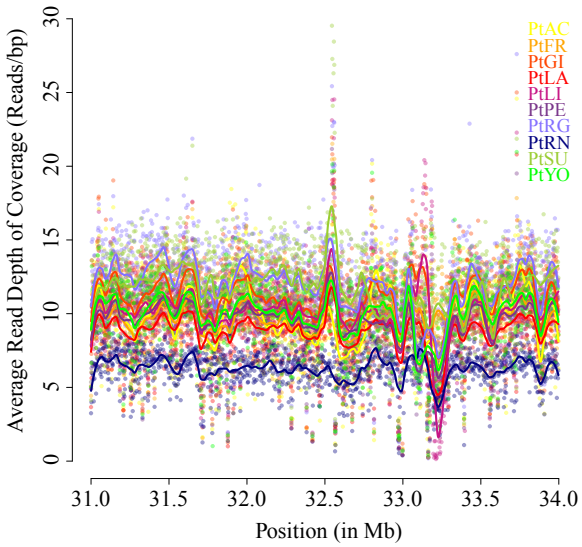


Figure 4.4: Depth of coverage of the 31Mb-34Mb region of chromosome 6 in the 10 chimpanzee individuals after removal of all read duplicates. Coverage was calculated using GATK's 'Depth of Coverage' tool [DePristo et al., 2011; McKenna et al., 2010] (averaged over 2.5kb regions).

Project Status For each individual between 10-54% of all reads were excluded from further analysis due to the following reasons: on average ∼15% of all reads were removed because they were duplicates (∼3% lane-level duplicates and ∼12% library-level duplicates). A further ∼8% of all reads in an individual were excluded because they did not satisfy the proper pair conditions (e.g. the mate did not map to the same chromosome in the opposite strand direction). Regarding only proper pairs without any duplicates, the mean coverage across individuals was 9.45X. A summary of the data is given in Table 4.1.

As an example, Figure 4.4 depicts the depth of coverage in the 10 chimpanzee individuals after removal of all read duplicates of the 31Mb-34Mb region of chromosome 6 corresponding to the MHC which encodes cell-surface molecules that are key components in the control of the adaptive immunological activity. The genes in the MHC region are some of the most polymorphic genes in vertebrate genomes known to date, making it a very difficult target for alignment algorithms. Consequently, a high variability in the read coverage can be observed.

4.3.2.2 Step 2.2: Base Quality Score Recalibration

Reported quality scores often only inaccurately display the actual probability of mismatching the reference genome due to quality variations caused by machine cycle (as mismatches tend to occur more frequently at the 3'-end of Illumina reads) and by sequence-context (e.g. AC and CG dinucleotides often have much lower quality than GT or TG). As the residual errors by machine cycle and by dinucleotide indicated that the reported quality systematically underestimated the

true quality (Figures 4.5(a) and 4.5(c)), a base quality score recalibration was carried out to identify high quality bases, to adjust reported quality scores, and to improve the accuracy of SNP calls.

Work Flow The recalibration was performed using GATK (Version 1.0.3471) [DePristo et al., 2011; McKenna et al., 2010], a toolkit to analyse next-generation read data obtained from Illumina, SOLiD, and 454 sequencers. GATK requires as input a sorted, indexed BAM file containing aligned reads and a FASTA-formatted reference with associated dictionary (.dict) and index (.fasta.fai) files. The index file was already generated during the data preprocessing (see Section 4.3.2.1). The dictionary file was created using Picard (Version 1.14)

```
java -jar CreateSequenceDictionary.jar \
     REFERENCE=reference.fasta \
     OUTPUT=reference.dict.
```

GATK's inbuilt quality score recalibrator improves the accuracy of a reported Phred-scaled base quality score Q_{raw} (QualityScoreCovariate) by providing empirically accurate base quality scores for each base in each read in every lane (ReadGroupCovariate) while also correcting for error covariates such as the read position in a machine cycle (CycleCovariate) and the preceding and current nucleotide (sequencing chemistry effect) observed by the sequencing machine (DinucCovariate). In a first step, the algorithm 'CountCovariates' tabulated bases mismatching with the reference genome at all loci not known to vary in the population categorised by the above mentioned covariates. The obtained information was stored in a recalibration table.

```
java -jar GenomeAnalysisTK.jar \
     -T CountCovariates \
     -I PtSU_rmdup.bam \
     -R reference.fasta \
     -cov ReadGroupCovariate \
     -cov QualityScoreCovariate \
     -cov CycleCovariate \
     -cov DinucCovariate \
     -recalFile PtSU_rmdup_recal.csv,
```

where -I specifies the input file that should be recalibrated, -R indicates the reference sequence, -cov represents the covariates used in the recalibration, and -recalFile specifies the file in which the recalibration table should be stored.

The generated data can be used to calculate an empirical probability of an error given the particular covariates seen at a site, where

$$P(\text{error}) = \frac{\#\text{ mismatches}}{\#\text{ observations}},$$

from which raw empirical quality scores can be computed by Phred-scaling the observed mismatch rate. Thus, in the next step, GATK's 'TableRecalibration' algorithm applied the covariates through a piecewise tabular correction to recalibrate the quality scores of all reads in a BAM file

```
java -jar GenomeAnalysisTK.jar \
     -T TableRecalibration \
     -I PtSU_rmdup.bam \
     -R reference.fasta  \
     -recalFile PtSU_rmdup_recal.csv \
     -outputBam PtSU_rmdup_recal.bam,
```

where -I represents the input file that should be recalibrated, -R specifies the reference sequence, -recalFile indicates the file containing the recalibration table which was generated in the previous step, and -outputBam specifies the file in which the recalibrated data should be stored.

Low quality score bases (per default: bases with a quality score less than 5) were not modified during the recalibration to avoid possible inappropriate elevation. The algorithm 'TableRecalibration' applied a Yates' correction for low occupancy bins to avoid over-fitting: the empirical quality score was inferred from

$$Q_{\text{emp}} = \frac{\#\text{ mismatches} + 1}{\#\text{ bases} + 1}.$$

The recalibrated quality score Q_{recal} was then estimated by breaking the initial quality scores down into linear components

$$Q_{\text{recal}} = Q_{\text{raw}} + \Delta Q_{\text{read group}} + \Delta Q_{Q_{\text{raw}}} + \sum_{\text{covariates}} \Delta\Delta Q_{\text{covariate},Q_{\text{raw}}},$$

where ΔQ and $\Delta\Delta Q$ were the residual differences between the reported quality scores and the empirical mismatch rates for all observations conditioning only on Q_{raw} or on both Q_{raw} and the covariates [DePristo et al., 2011].

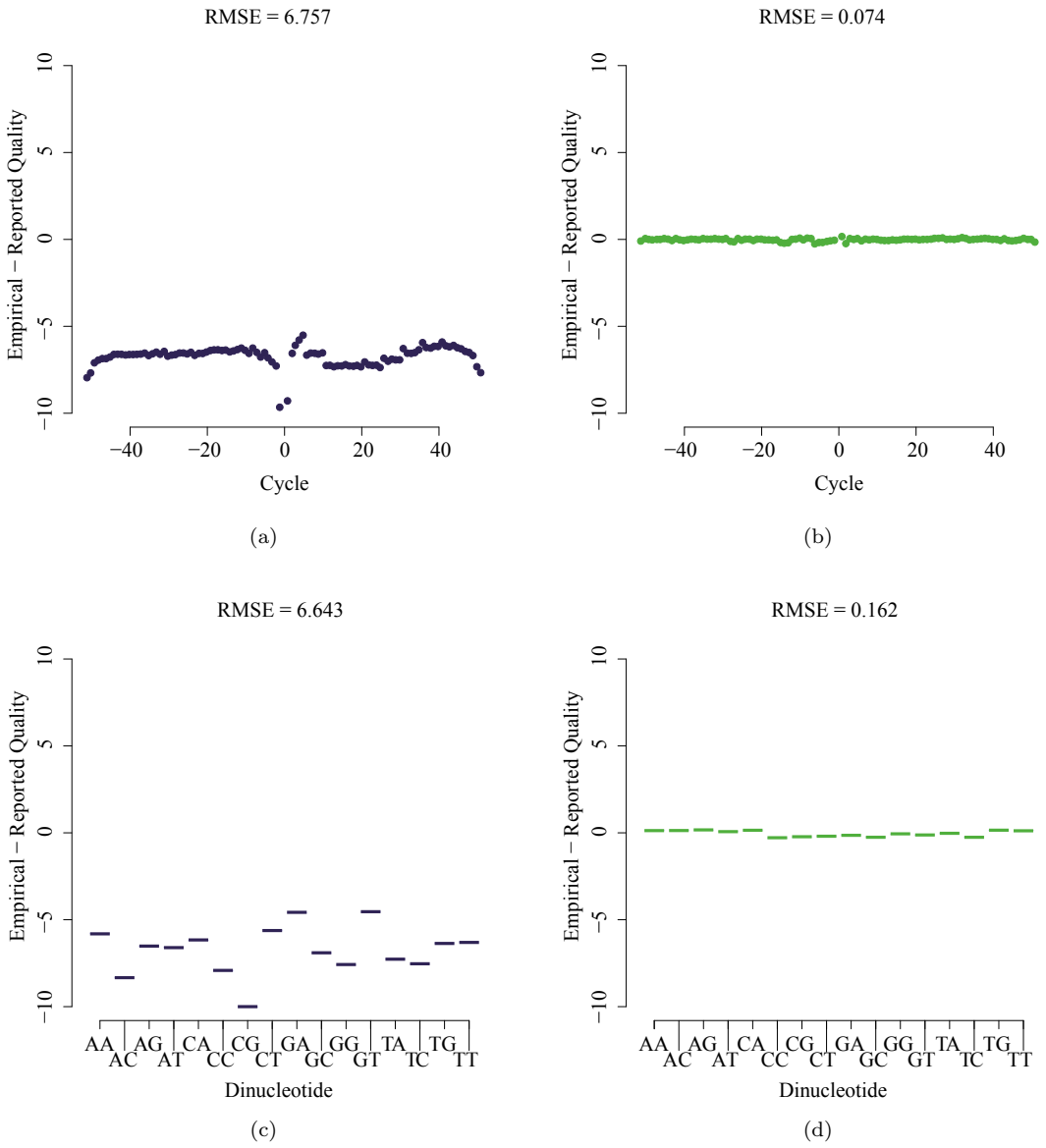


Figure 4.5: The plots indicate the residual error by machine cycle (with positive and negative cycle values given for the first and second read in a pair) (a) before and (b) after recalibration and the residual error for each of the 16 genomic dinucleotide contexts (e.g. the AC context occurs at all sites in a read where C is the current nucleotide preceded by A) (c) before and (d) after recalibration. Empirical quality scores were calculated by Phred-scaling the observed rate of mismatches with the reference genome. Root mean square errors are given for the pre- and post-recalibration curves.

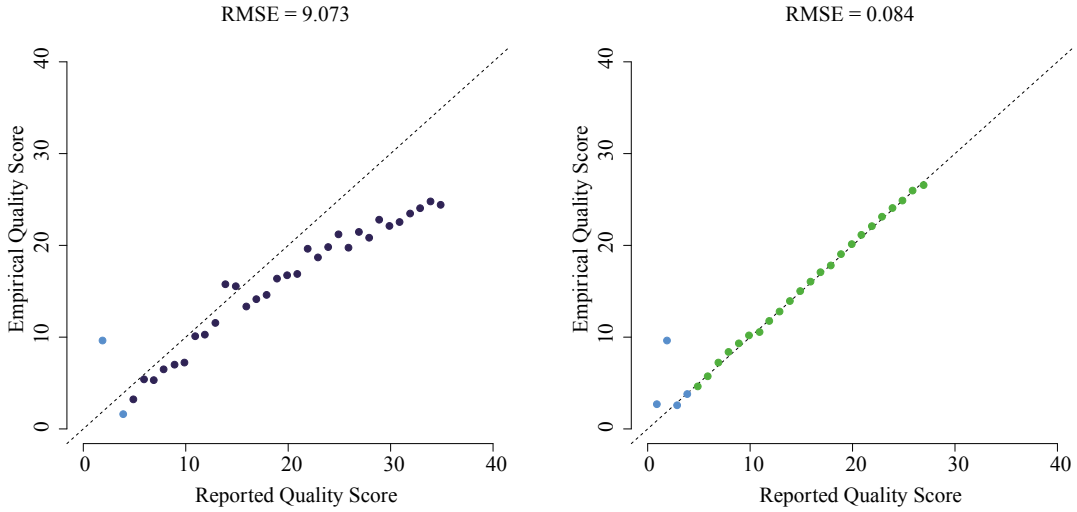


Figure 4.6: Reported versus empirical quality scores (a) before and (b) after recalibration. Empirical quality scores were calculated by Phred-scaling the observed rate of mismatches with the reference genome. Bases with a quality value smaller than 5 (indicated in lightblue) were ignored during the recalibration.

Project Status Recalibration was conducted for the duplicate-free reads mapped against the chimpanzee reference genome (but not for the human-based mappings as they were only used to validate SNP calls). After recalibration, the root mean square error (RMSE) between the empirical and reported quality scores was decreased from 6.757 to 0.074 by machine cycle and from 6.643 to 0.162 by dinucleotide context (see Figure 4.5). In general, taking the read group, the machine cycle, and the dinucleotide covariates into account, the underestimated reported quality scores could be adjusted to be in better concordance with the empirical ones (see Figure 4.6; RMSE before recalibration: 9.073; RMSE after recalibration: 0.084).

4.3.2.3 Step 2.3: Base Alignment Quality

Mapping artefacts caused by gapless alignments, due to the annotation of SNPs rather than indels, can be introduced in single alignments resulting in false positive SNP calls at positions of true indels. Preliminary data analysis suggested an excess of SNPs in low-complexity regions of the genome, a pattern indicative of local misalignments leading to a higher FDR for SNPs in these regions. To reduce the number of these artefacts, a Base Alignment Quality adjustment [Li, 2011] was applied prior to SNP calling which replaced the original base quality score of each nucleotide with the minimum between the original score and a computed BAQ score (a Phred-scaled probability evaluating the probability of a nucleotide being misaligned). This procedure down weighted base qualities in regions with high ambiguity in the local alignment thereby avoiding an excess of false positive SNP calls in low-complexity regions of the genome.

Work Flow BAQ scores were computed at a per-base level using SAMtools' 'calmd' algorithm (Version 0.1.12a) [Li et al., 2009a] which implements a Base Alignment Quality [Li, 2011] adjustment modelling the read genesis to each alignment independently using a profile Hidden Markov Model

```
samtools calmd -Abr PtSU_rmdup_recal.bam \
               reference.fasta > PtSU_rmdup_recal_baq.bam.
```

4.3.2.4 Step 2.4: Initial SNP Call

Initial SNP calling was performed using GATK's 'UnifiedGenotyper' [DePristo et al., 2011; McKenna et al., 2010], a multiple-sample, technology and quality score aware SNP caller. It uses a Bayesian genotype likelihood model to jointly estimate simultaneously the most likely genotypes and allele frequency in a population. The 'UnifiedGenotyper' emits a posterior probability of the occurrence of a segregating variant allele at each locus as well as for the genotype of each sample [DePristo et al., 2011; McKenna et al., 2010]. The single sample genotype likelihood is computed as

$$L(G|D) = P(G)\ P(D|G) = \Pi_{b\ \epsilon\ \text{good bases}}\ P(b|G)$$

where $L(G|D)$ is the likelihood for the genotype G given the data D (computed for all 10 genotypes: A/A, A/C, A/G, A/T, C/C, C/G, C/T, G/G, G/T, and T/T), $P(G)$ is the prior for the genotype (during multi-sample calculation: $P(G) = 1$), $P(D|G)$ is the likelihood of the data given the genotype (computed using the pileup of bases and the associated quality scores at a given locus), and $\Pi_{b\ \epsilon\ \text{good bases}}\ P(b|G)$ is an independent base model which uses a platform-specific confusion matrix. More information can be found in McKenna et al. [2010] and DePristo et al. [2011].

Work Flow GATK's 'UnifiedGenotyper' (Version 1.0.4705) was run using the following command

```
java -jar GenomeAnalysisTK.jar \
     -T UnifiedGenotyper \
     -I PtAC_rmdup_recal_baq.bam \
     -I PtFR_rmdup_recal_baq.bam \
     -I PtGI_rmdup_recal_baq.bam ... \
     -R reference.fasta \
     -o GATK_snp_call.vcf,
```

where -I specifies the input file(s) (one input file per chimpanzee individual; all 10 individuals were used to make joint SNP calls from the autosomes and the PAR whilst only the nine

females were used to call SNPs from the non-PAR of chromosome X), -R indicates the reference sequence, and -o represents the output file in which the SNP calls should be stored (here in VCF format, see Appendix F). GATK's algorithm only includes bases satisfying a minimum base quality, read mapping quality, and pair mapping quality in this calculation. The following parameters were used as set per default: (i) the minimum base quality and the minimum read mapping quality a base or a read must possess for calling were 10, (ii) the maximum fraction of reads with deletions spanning a locus tolerated by the genotyper was 0.05, and (iii) the maximum number of mismatches within a 20bp window on both sides of the target position was 3. To distinguish low from high confidence calls, only genotypes above a certain Phred-scaled quality score were emitted. Per default, this threshold was set to 50.

Project Status Autosomal SNPs and SNPs in the PAR were called from the 10 chimpanzee individuals using GATK's 'UnifiedGenotyper'. Due to the hemizygosity of the one male in the study, SNPs in the non-PAR of chromosome X were called separately, in the nine female chimpanzee individuals only. An overview of the initial SNP calls on the single chromosomes is given in Table 4.2.

The initial call set from the chimpanzee-based mappings contained 6,869,589 autosomal SNPs with a transition-transversion (Ts/Tv) ratio of 1.91. On chromosome X 306,091 SNPs were called, 8,013 SNPs in the PAR and 298,078 SNPs in the non-PAR of chromosome X with a Ts/Tv ratio of 1.65 and 1.42, respectively. The initial call set from the human-based mappings contained 34,627,784 autosomal SNPs with a Ts/Tv ratio of 2.22 and 1,431,038 SNPs on chromosome X, 54,382 SNPs in the PAR and 1,376,656 SNPs in the non-PAR of chromosome X with a Ts/Tv ratio of 1.83 and 2.07, respectively. (Note that the call set from the human-based mappings contained both variation within chimpanzees as well as their divergence from humans.)

In both data sets, SNPs on chromosome X exhibited a lower Ts/Tv ratio than autosomal SNPs, an observation that is expected from the results of previous analyses in humans: in the Pilot 1 study of the 1000 Genomes Project, alignments of high quality read sequences to the human reference genome yielded a Ts/Tv ratio of about 1.85 for SNPs on chromosome X while the Ts/Tv ratio for autosomal SNPs was closer to 2 (personal communication). This disparity is most likely at least partially caused by the difference in diversity rates between the sex chromosomes and the autosomes.

Chromosome	Statistics			
	# $\mathbf{SNPs_C}$	$\mathbf{Ts/Tv_C}$	# $\mathbf{SNPs_H}$	$\mathbf{Ts/Tv_H}$
1	542,633	1.96	2,718,634	2.29
2a	273,741	1.86	N/A	N/A
2b	313,414	1.90	N/A	N/A
2	N/A	N/A	3,009,067	2.17
3	499,497	1.86	2,481,003	2.14
4	492,609	1.83	2,478,800	2.10
5	442,837	1.87	2,244,255	2.14
6	438,836	1.94	2,139,878	2.20
7	420,571	1.92	2,035,799	2.19
8	374,424	1.84	1,964,586	2.05
9	294,165	1.83	1,471,512	2.15
10	327,370	1.96	1,693,672	2.22
11	312,137	1.88	1,688,149	2.15
12	346,637	1.90	1,654,016	2.24
13	236,223	1.89	1,255,897	2.18
14	218,232	1.90	1,102,244	2.22
15	198,642	1.91	1,022,027	2.22
16	221,689	1.89	1,144,411	2.11
17	193,456	2.01	972,450	2.48
18	186,094	1.91	984,151	2.21
19	166,480	1.86	794,383	2.41
20	165,965	1.96	805,099	2.35
21	101,570	2.02	492,401	2.17
22	102,367	2.02	475,350	2.46
X	298,078	1.42	1,376,656	2.07
PAR1	8,013	1.65	54,382	1.83

$_C$ = Chimpanzee-based mappings
$_H$ = Human-based mappings

Table 4.2: Initial SNP calls from the chimpanzee-based mappings contained 6,869,589 autosomal SNPs and 306,091 SNPs from chromosome X, with an average Ts/Tv ratio of 1.91 and 1.42, respectively. The call set from the human-based mappings contained 34,627,784 autosomal SNPs and 1,431,038 SNPs from chromosome X, with an average Ts/Tv ratio of 2.22 and 2.06, respectively. (Note that the call set from the human-based mappings contained both variation within chimpanzees as well as their divergence from humans.)

4.3.3 Filtering

Besides true variation, the initial SNP call set contains false positives due to systematic machine artefacts, mismapped reads, and misaligned indels caused by the reference genome which is both incomplete and does not represent all genetic variation. Such false positive calls often (i) exhibit an unusual behaviour according to excessive depth of coverage (as a read depth much lower or much higher than expected is suggestive of copy number variation which might lead to false positive SNP calls in the mapping of paralogous sequences), (ii) show a skewed allelic imbalance, (iii) occur preferentially on a single strand, (iv) appear in regions of poor read alignment (frequently indicated by a poor mapping quality score), and (v) arise in close proximity to other SNPs. Thus, the majority of such calls can be detected and rejected using different filter criteria based on the above observations.

Work Flow The relationship between each aggregated SNP covariate and an error must be determined before the gained information can be used to filter raw SNPs. For this purpose, the

variant calls were annotated using GATK's 'VariantAnnotator' (Version 1.0.4705) [DePristo et al., 2011; McKenna et al., 2010] with the following parameters:

```
java -jar GenomeAnalysisTK.jar \
     -T VariantAnnotator \
     -I PtAC_rmdup_recal_baq.bam \
     -I PtFR_rmdup_recal_baq.bam \
     -I PtGI_rmdup_recal_baq.bam ... \
     -R reference.fasta  \
     -all \
     -B:variant,VCF GATK_snp_call.vcf \
     -o GATK_snp_call_annotated.vcf,
```

where -I represents the input file(s) (one input file per chimpanzee individual; all 10 individuals were used to annotate SNP calls from the autosomes and the PAR whilst only the nine females were used to annotate SNPs from the non-PAR of chromosome X), -R indicates the reference sequence, -all specifies that all possible annotations should be made, -B is the file containing the SNP calls (here in VCF format), and -o represents the output file in which the annotated SNP calls should be stored.

The following annotations were used:

AC	Number of chromosomes that carry the alternative allele,
AF	Frequency of the allele,
AN	Number of chromosomes,
DP	Depth of coverage at the given position,
Dels	The percentage of reads with deletions spanning this position,
HRun	Length of the longest continuous homopolymer run of the variant allele,
HaplotypeScore	Estimate of the probability that the reads at this locus are coming from no more than two very local haplotypes,
MQ0	Number of reads with a mapping quality of 0 at a locus,
MQ	Root mean square mapping quality of all reads,
QD	Confidence value of the VCF record divided by the sum of depths of all samples with non-reference genotypes, and
SB	Strand bias.

As recommended by best-practise, SNPs were filtered in order to control the Ts/Tv ratio (a commonly used assessor to evaluate the quality of a SNP call as transitions are known to be twice as common as transversions in the human genome [Zhao and Boerwinkle, 2002]), while maintaining as many SNP calls as possible. Each covariate was binned and plotted against the Ts/Tv ratio to manually select filtering thresholds for the variant annotations (see Figure 4.7).

SNPs were removed if

clusterWindowSize=10	Three or more SNPs were found within 10bp,
AF=1	All individuals were fixed for the non-reference allele,
AN<16	The number of chromosomes in the sample with a genotype call was <16,
DP<30 or DP>180	The depth of coverage (summed across individuals) at a given position was <30 or >180,
Dels>0	The percentage of reads with deletions spanning the position was >0,
HRun>2	The length of the longest continuous homopolymer run of the variant allele was >2,
HaplotypeScore>40	The estimated probability that the read bases within 10bp of the locus are consistent with no more than two haplotypes was >40,
MQ0>40 or MQ0/DP>0.4	The number of reads with a mapping quality of 0 at a locus was >40 or MQ0/DP > 0.4,
MQ<25	The root mean square mapping quality of all reads was <25,
QD>27.0 or QD<1.0	The SNP quality score divided by the sum of depth across all samples with non-reference genotypes was >27.0 or <1.0, or
SB>0	There was evidence of a strand bias.

Due to the hemizygosity of the one male in the study, chromosome X was split into the pseudoautosomal and the non-pseudoautosomal regions. SNP calling in the PAR proceeded as for the autosomes while SNPs in the non-PAR of chromosome X were called from the nine females only. As a result, GATK's SNP filter criteria needed to be adjusted to account for the fewer number of chromosomes in the sample.

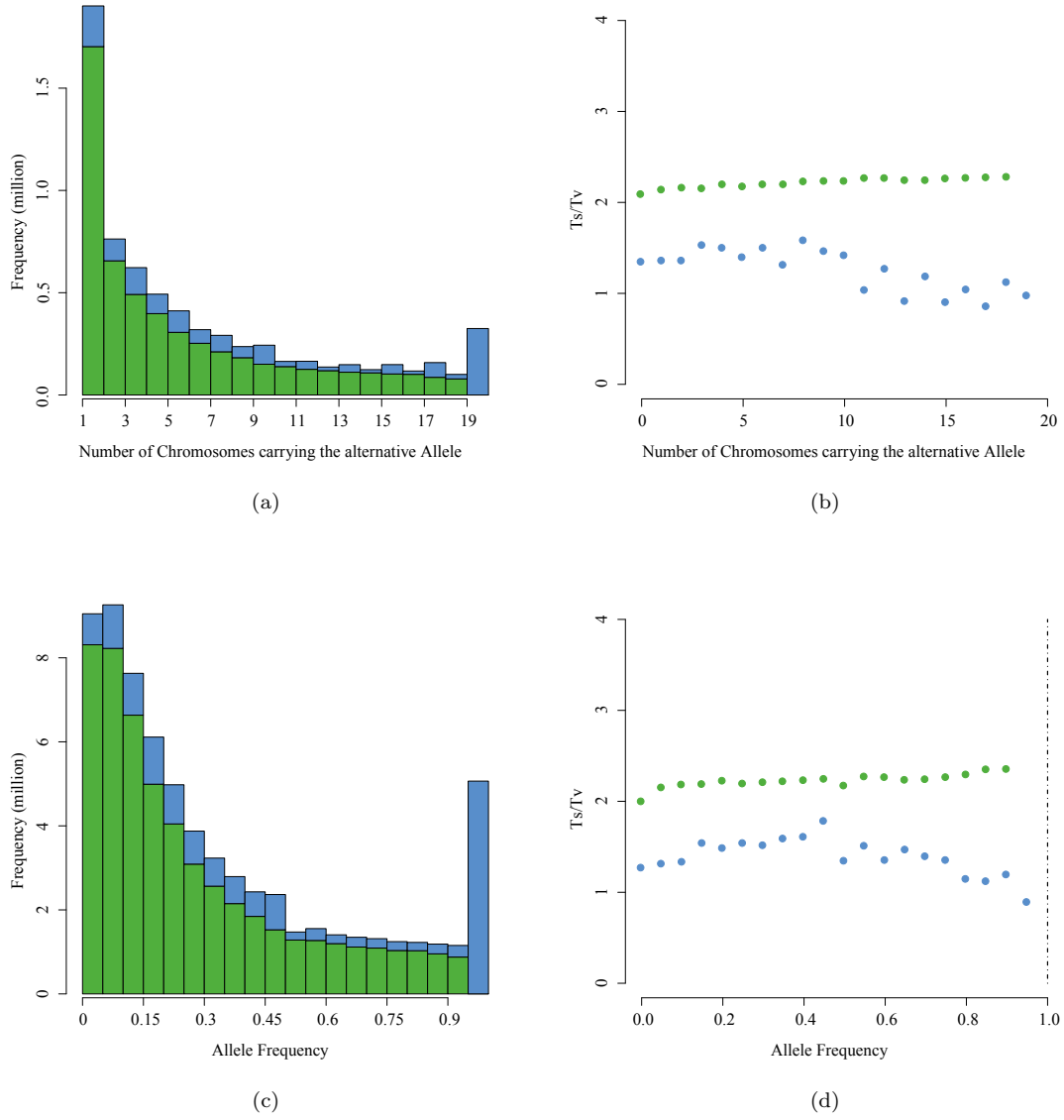
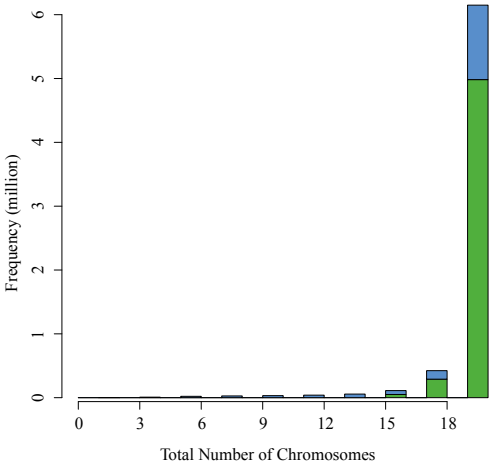
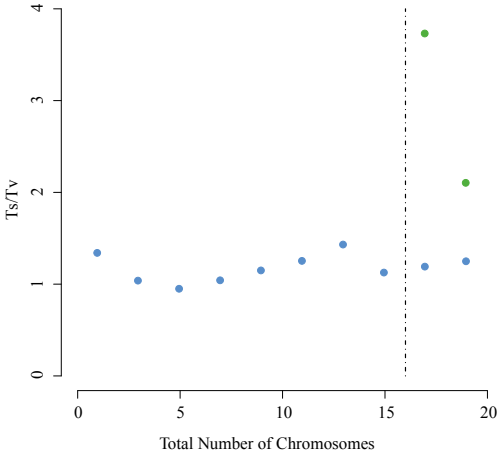


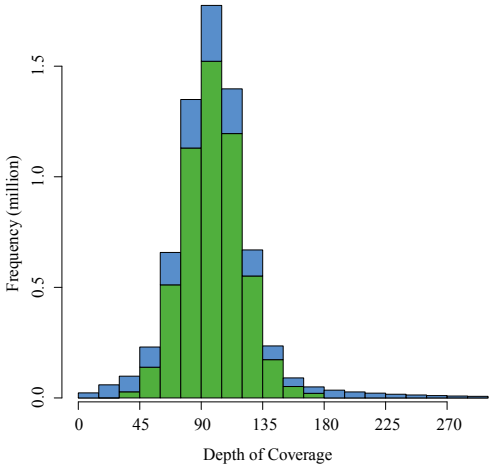
Figure 4.7: Histograms and Ts/Tv ratio of SNP covariates that passed the filter criteria (green) and those that failed (blue). The set thresholds are indicated by a black dashed line. (a,b) Number of chromosomes carrying the alternative allele. (c,d) Allele frequency. (e,f) Total number of chromosomes. (g,h) Depth of coverage. (i,j) Percentage of reads with deletions spanning this position. (k,l) Length of the longest continuous homopolymer. (m,n) Estimate of the probability that the reads at this locus are coming from no more than two very local haplotypes. (o,p) Number of reads with mapping quality 0 at locus. (q,r) Root mean square mapping quality of all reads. (s,t) Quality score over depth. (u,v) Strand bias.



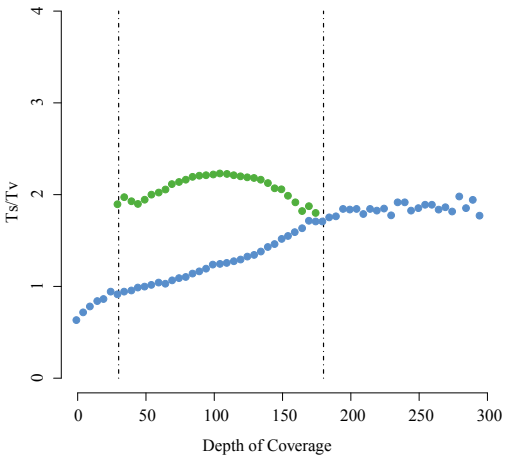
(e)



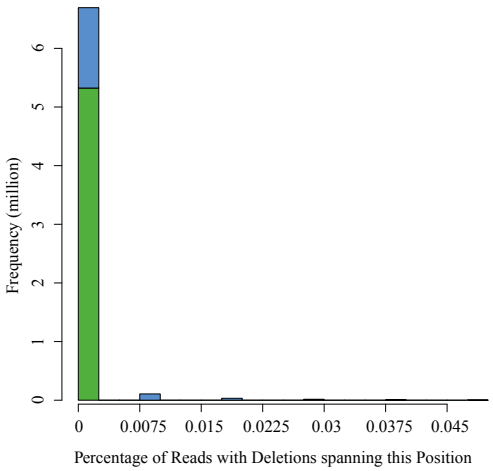
(f)



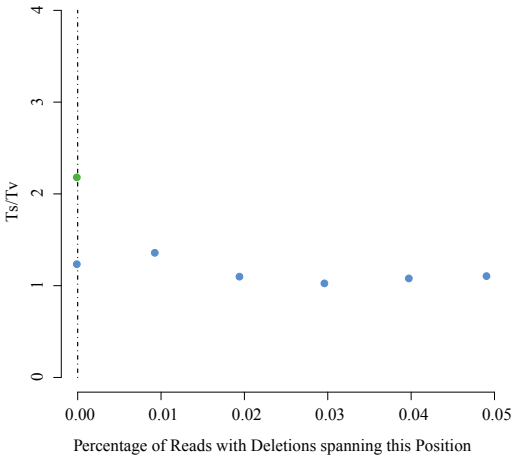
(g)



(h)

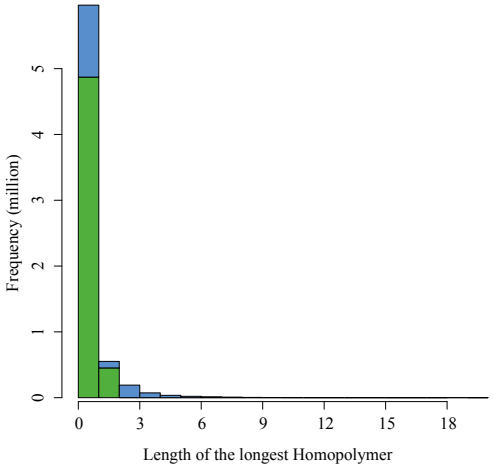


(i)

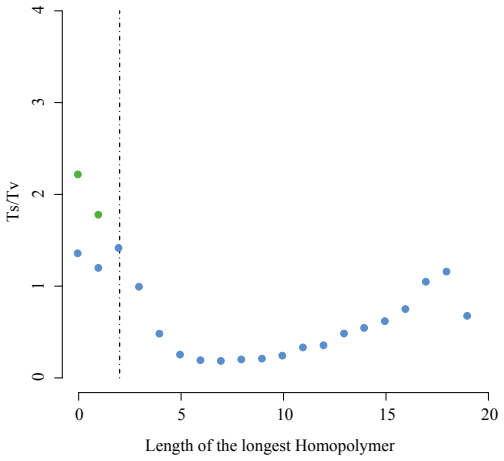


(j)

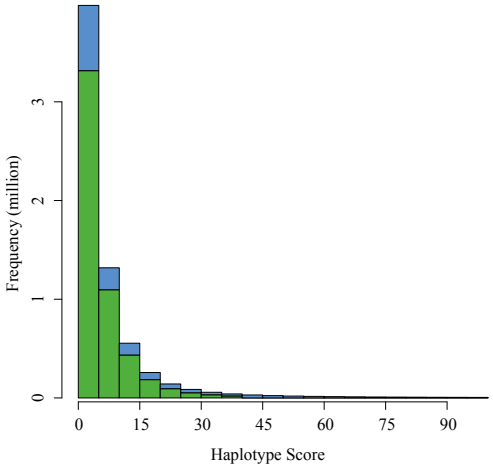
Figure 4.7



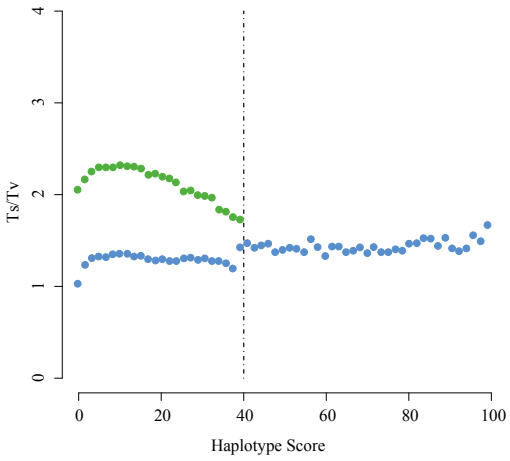
(k)



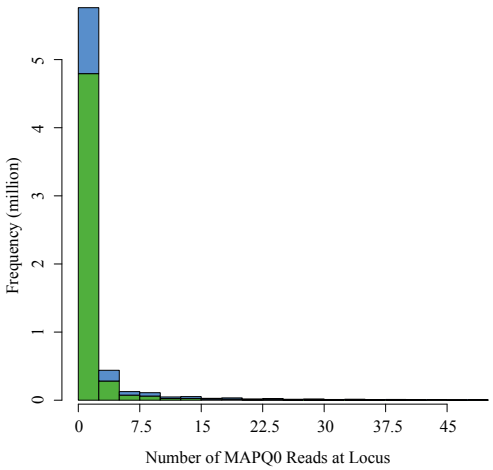
(l)



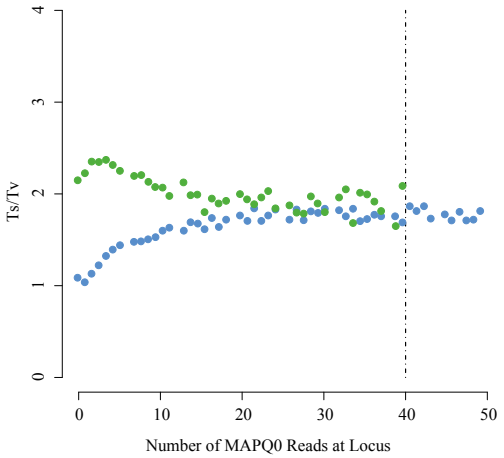
(m)



(n)

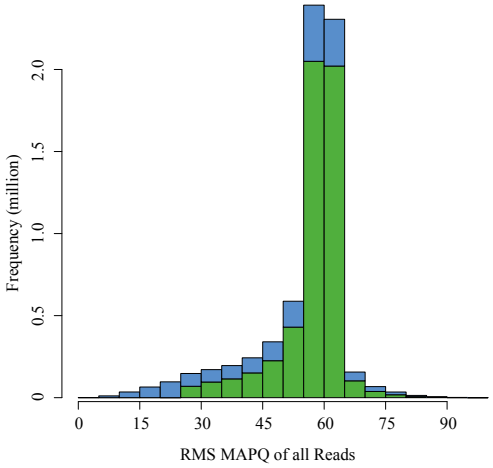


(o)

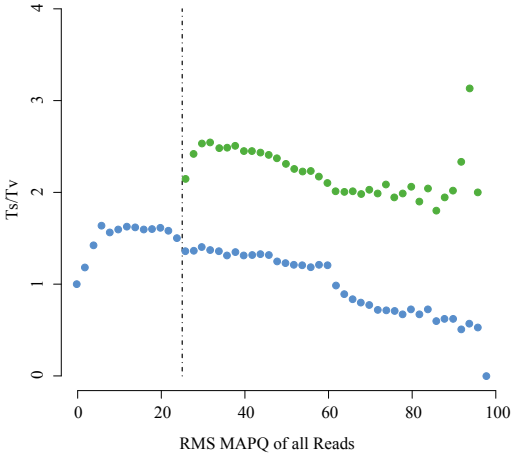


(p)

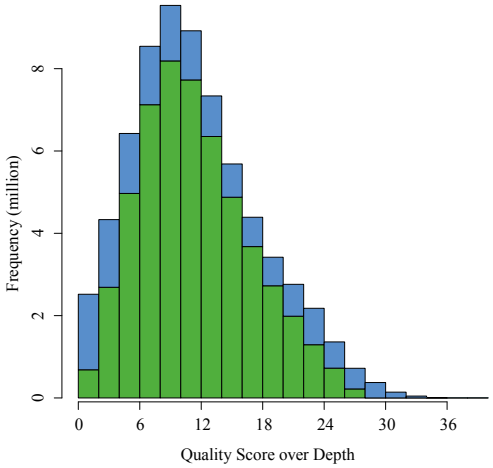
Figure 4.7



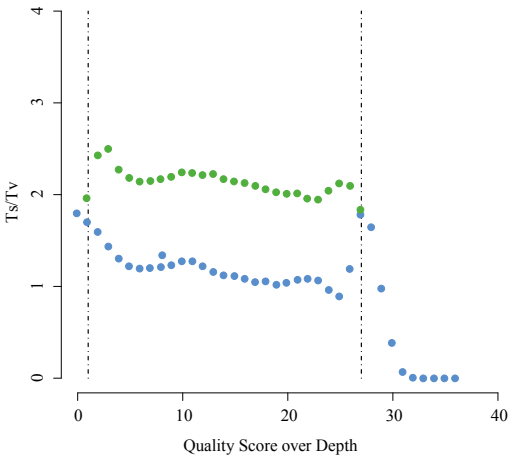
(q)



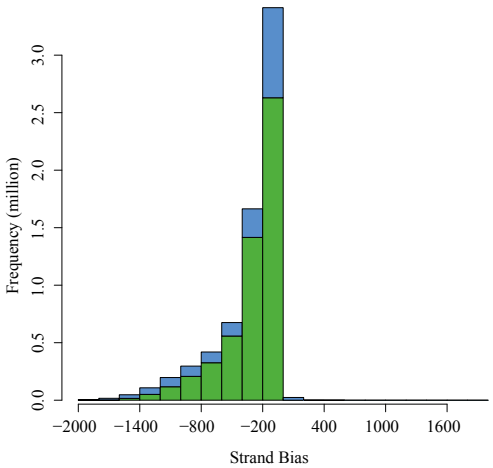
(r)



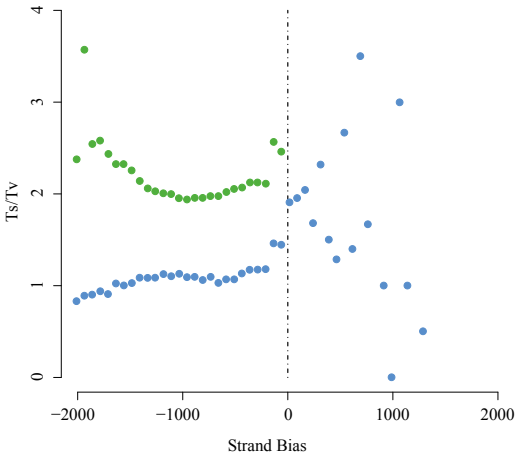
(s)



(t)



(u)



(v)

Figure 4.7

Specifically, the following changes have been made to the previously specified thresholds:

AN<14	The number of chromosomes in the sample with a genotype call was <14,
DP<27 or DP>162	The depth of coverage (summed across individuals) at a given position was <27 or >162, or
QD>24.3 or QD<0.9	The SNP quality score divided by the sum of depth across all samples with non-reference genotypes was >24.3 or <0.9.

The initial SNP calls were filtered using GATK's 'VariantFiltration' (Version 1.0.4705)

```
java -jar GenomeAnalysisTK.jar \
     -T VariantFiltration \
     -I PtAC_rmdup_recal_baq.bam \
     -I PtFR_rmdup_recal_baq.bam \
     -I PtGI_rmdup_recal_baq.bam ... \
     -R reference.fasta  \
     -B:variant,VCF GATK_snp_call_annotated.vcf \
     -o GATK_snp_call_filtered.vcf \
     --clusterWindowSize 10 \
     --filterExpression 'AF>=1.0 || AN<16 || DP<30 || DP>180 || \
                         Dels>0.0 || HaplotypeScore>40.0 || \
                         HRun>2 || MQ<25.0 || MQ0>40 || \
                         (MQ0 \ DP)>0.4 || QD>27.0 || QD<1.0 || SB>0.0',
```

where -I represents the input file(s) (one input file per chimpanzee individual; all 10 individuals were used to filter SNP calls from the autosomes and the PAR whilst only the nine females were used to filter SNPs from the non-PAR of chromosome X), -R specifies the reference sequence, -B indicates the file containing the SNP calls, -o specifies the output file, and - -clusterWindowSize as well as - -filterExpression indicate the previously described filter criteria (in this example for the autosomes/PAR).

In a last step, SNPs were filtered on the basis of the Hardy Weinberg Equilibrium. A p-value for HWE was calculated for each SNP using VCFtools[18] (Version 0.1.3.2) [Danecek et al., 2011] and SNPs with $p < 0.01$ were removed.

[18] http://vcftools.sourceforge.net/

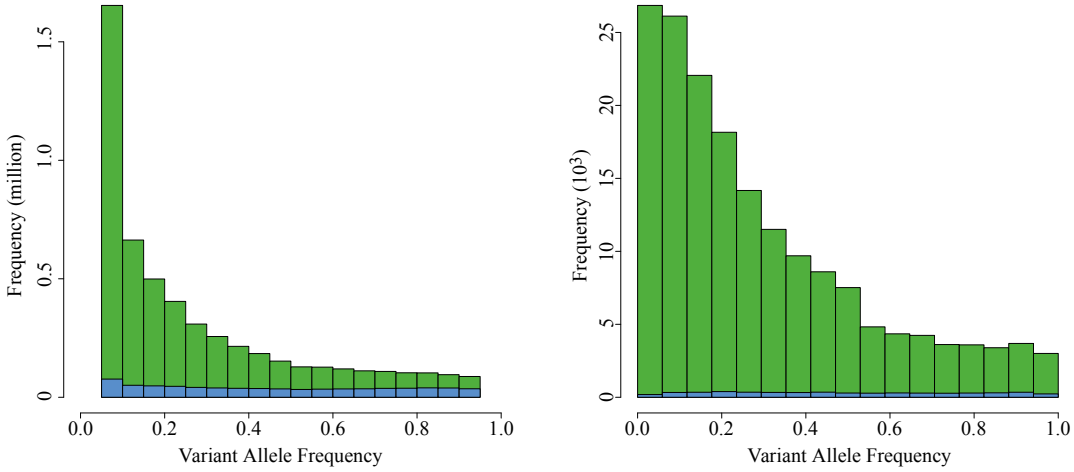


Figure 4.8: (a) The number of autosomal variants at different allele frequencies. (b) The number of variants on chromosome X at different allele frequencies. SNPs that have been previously reported in the public catalogue of variant sites dbSNP (release 125) [Sherry et al., 2001] are indicated in blue while novel SNPs are shown in green.

Project Status The initial SNP call sets from both the chimpanzee-based mappings and the human-based mappings were filtered using GATK's 'VariantFiltration' and VCFtools.

This procedure caused 1,546,288 autosomal SNPs (Ts/Tv: 1.23), 5,251 SNPs in the PAR (Ts/Tv: 1.49), and 125,460 SNPs in the non-PAR of chromosome X (Ts/Tv: 0.93) from the chimpanzee-based mapping to be filtered out as they failed one or more of the above stated criteria. The resulting call set from the chimpanzee-based mappings contained 5,323,301 autosomal SNPs with a Ts/Tv ratio of 2.18 which compares well with the Ts/Tv ratio of 2.10 estimated from high quality substitutions on the chimpanzee lineage (chimpanzee reference base quality $\geq$ 90). Of those SNPs, 14% have been previously ascertained in Western chimpanzees and have been reported in the public catalogue of genetic variant sites dbSNP (release 125) [Sherry et al., 2001] for which chimpanzee SNPs have largely been submitted by the Chimpanzee Sequencing and Analysis Consortium (but also from a few other laboratories). These SNPs show a uniformly distributed allele frequency spectrum for the non-reference allele (see Figure 4.8(a)). This is exactly what is expected if the SNPs were ascertained in only two chromosomes (or in other words one chimpanzee individual). This is not surprising given that the majority of the data generated by the Chimpanzee Sequencing and Analysis Consortium originated from a single male chimpanzee individual [Chimpanzee Sequencing and Analysis Consortium, 2005].

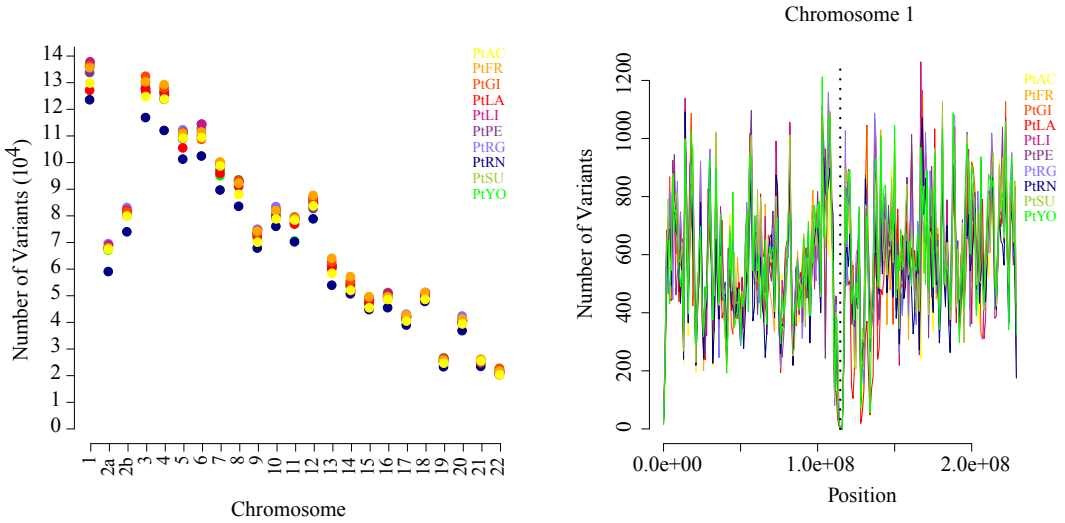
On the other hand, the PanMap study could only recover 51.1% of the dbSNP SNPs (see Table 4.3). Assuming that all SNPs recorded in dbSNP came from Clint, the single captive-

Chromosome	Statistics						
	# SNPs	% dbSNP[1]	% $\text{dbSNP}^{1}_{\text{recov}}$	$\text{Ts/Tv}_{\text{known}}$	$\text{Ts/Tv}_{\text{novel}}$	% Acc[2]	SNP/kb
1	422,154	13.94	50.88	2.29	2.26	85.65	1.84
2a	215,624	14.17	51.38	2.14	2.11	83.62	1.88
2b*	248,364	14.30	51.84	2.20	2.17	86.87	1.85
3	395,984	14.79	52.81	2.12	2.10	87.33	1.94
4	391,850	14.14	53.11	2.06	2.05	87.05	2.01
5	351,334	14.64	52.58	2.15	2.11	86.35	1.91
6	345,909	14.39	52.43	2.17	2.18	85.39	1.99
7	318,515	13.54	50.84	2.18	2.18	84.07	1.99
8	291,031	14.29	51.85	2.15	2.08	85.93	2.01
9	227,515	13.99	52.35	2.06	2.08	70.67	1.64
10	253,910	14.33	52.22	2.28	2.24	84.24	1.88
11	241,990	14.49	52.39	2.21	2.15	83.72	1.80
12	267,916	13.90	53.61	2.21	2.20	87.12	1.98
13	189,311	14.11	53.19	2.12	2.12	69.19	1.63
14	170,913	14.87	51.99	2.18	2.18	72.82	1.59
15	154,803	14.55	53.41	2.25	2.19	69.68	1.55
16	166,507	13.10	48.39	2.13	2.18	71.96	1.84
17	140,515	13.27	48.24	2.45	2.46	78.26	1.69
18	148,689	14.66	52.39	2.16	2.15	87.59	1.92
19	107,275	11.20	49.07	2.37	2.40	69.14	1.66
20	125,534	14.38	53.68	2.33	2.36	84.99	2.02
21	75,581	17.56	26.60	2.22	2.24	65.10	1.63
22	72,077	12.42	51.47	2.52	2.45	56.27	1.44
X	172,618	3.01	35.08	2.06	1.96	5.40	1.23
PAR	2,762	0.00	0.00	N/A	2.00	22.96	1.06

* = Excluding the first 114Mb as the sequence is not available,
[1] = SNPs ascertained in Western chimpanzees in dbSNP version 125,
[2] = Accessibility: site could be called (either passes all filter criteria or it is called as homozygous reference with a 'LowQual' filter status - emitted by GATK's algorithm when there is insufficient evidence for the site to be called as a SNP - in all individuals),
recov = recovered

Table 4.3: Filtered SNP calls from the chimpanzee-based mappings contained 5,323,301 autosomal SNPs with an average Ts/Tv ratio of 2.18. On chromosome X, 2,762 SNPs were called in the PAR and 172,618 SNPs were called in the non-PAR of chromosome X with a Ts/Tv ratio of 2.00 and 1.97, respectively.

born male Western chimpanzee from which the majority of the sequences analysed for the discovery of dbSNP SNPs originated, and that the ancestral alleles x are distributed with frequency $1/x$, the probability of discovering a dbSNP SNP in 10 Western chimpanzee individuals, assuming an ideal scenario of 100% SNP discovery power in the study, is ∼90%. This probability decreases slightly both with a smaller SNP discovery power (the actual power of the PanMap study to detect segregating sites in the chimpanzee population is in the range of ∼85%, see Section 4.3.6) and in the case that dbSNP SNPs were obtained from a wider set of chimpanzee individuals. Nevertheless, a recovery of only 51.1% of the dbSNP SNPs in the PanMap study is much less than expected. Possible explanations for this observation could be (i) a low quality of the PanMap data set (an explanation that can be ruled out given an autosomal FDR of only 2.7%, see Section 4.3.5), (ii) an exclusion of sites in the PanMap data set for which dbSNP SNPs were called (see Section 5.2), (iii) an occurrence of a large amount of false positives in dbSNP (large error rate of dbSNP SNPs have been reported before, e.g. in humans, ∼15-17% false positives have been detected which were most likely caused by errors in sequencing and SNP calling, particularly at heterozygous sites [Mitchell et al., 2004] as



(a) (b)

Figure 4.9: (a) The number of variants identified per chimpanzee individual per chromosome. The variants were distributed appropriately across chromosomes and the numbers of variants identified per sample were highly consistent, although one sample (PtRN) showed 5-10% fewer variants than other samples. (b) The number of variants identified along chromosome 1 in 1Mb windows for each chimpanzee individual.

well as ∼8% false positives caused by incorrect calls of substitutions in paralogous sequences [Musumeci et al., 2010]), or (iv) an underlying (but currently unknown) demographic effect in Western chimpanzees.

The frequency spectrum depicted in Figure 4.8(a) also clearly shows that the addition of more samples facilitates the discovery of a greater number of rare variants in the study. In fact, 86% of the SNPs reported by the PanMap study were novel compared to the dbSNP SNPs, increasing the number of known SNPs in Western chimpanzees by a factor of three.

The variants were distributed appropriately across chromosomes and the numbers of variants identified per sample were highly consistent, although one sample (PtRN) showed a 40% lower average coverage and 5-10% fewer variants than other samples (see Figure 4.9(a)). Variation in SNP density was present (see Figure 4.9(b)) but no strong peaks of SNP density indicative of a high local FDR were observed.

On chromosome X, 2,762 SNPs with a Ts/Tv ratio of 2.00 were called in the PAR whilst 172,618 SNPs with a Ts/Tv ratio of 1.97 were called outside the PAR on chromosome X. The number of variants on chromosome X at different allele frequencies is given in Figure 4.8(b). Despite these encouraging statistics, an analysis of preliminary data showed that the systematic incompleteness of the chimpanzee reference sequence caused issues in the variant calling process. As a result, SNP calls on chromosome X were of substantially lower quality (see Section 4.3.5). In contrast, the 220,040 chromosome X SNPs (9,623 SNPs in the PAR with a

Chromosome	Statistics			
	# $SNPs_H$	Ts/Tv_H	# $SNPs_I$	Ts/Tv_I
1	478,452	2.46	320,829	2.34
2a	N/A	N/A	163,624	2.17
2b	N/A	N/A	191,814	2.24
2	513,304	2.31	N/A	N/A
3	433,061	2.27	302,879	2.17
4	423,443	2.21	294,758	2.12
5	383,123	2.27	264,809	2.18
6	387,742	2.32	261,240	2.24
7	359,318	2.35	234,777	2.24
8	325,814	2.24	218,879	2.16
9	252,177	2.26	173,142	2.13
10	288,812	2.43	191,980	2.33
11	286,943	2.31	182,413	2.24
12	301,486	2.41	202,452	2.30
13	224,155	2.29	145,357	2.18
14	191,828	2.36	129,970	2.26
15	176,646	2.38	114,941	2.28
16	190,904	2.36	121,379	2.24
17	172,260	2.70	102,975	2.56
18	163,659	2.32	116,344	2.22
19	143,432	2.69	63,006	2.63
20	144,865	2.53	97,319	2.42
21	80,577	2.39	58,323	2.29
22	87,982	2.66	52,019	2.57
X	210,417	2.09	107,482	2.11
PAR1	9,623	2.00	1,477	1.95

$_H$ = Human-based mappings
$_I$ = Intersection set of chimpanzee-based and human-based mappings

Table 4.4: Filtered SNP calls from the human-based mappings contained 6,009,982 autosomal SNPs, 9,623 SNPs in the PAR, and 210,417 SNPs in the non-PAR of chromosome X, with an average Ts/Tv ratio of 2.39, 2.00, and 2.09, respectively. The intersection set of SNPs called from chimpanzee-based and human-based mappings contained 4,005,229 autosomal SNPs, 1,477 SNPs in the PAR, and 107,482 SNPs in the non-PAR of chromosome X, with an average Ts/Tv ratio of 2.25, 1.95, and 2.11, respectively.

Ts/Tv ratio of 2.00 and 210,417 SNPs in the non-PAR of chromosome X with a Ts/Tv ratio of 2.09, see Table 4.4) called from the human-based mappings were of better quality. Therefore, an intersection set of chimpanzee-based and human-based SNP calls was generated in order to identify SNPs mapping in places with reliable synteny between the two species.

In order to generate an intersection set between the chimpanzee-based and human-based SNP calls, all 5,323,301 autosomal SNPs as well as 175,380 SNPs on chromosome X from the chimpanzee-based mapping were lifted over from the chimpanzee reference sequence, panTro2, to the human reference genome, hg18. Similarly, the 6,009,982 autosomal SNPs (Ts/Tv: 2.39) and the 220,040 chromosome X SNPs (Ts/Tv: 2.09) from the filtered human-based call (28,617,802 autosomal SNPs with a Ts/Tv ratio of 2.16, 44,759 SNPs in the PAR with a Ts/Tv ratio of 1.79, and 1,166,239 SNPs in the non-PAR of chromosome X with a Ts/Tv ratio of 2.07 from the initial call failed the filter criteria) needed to be lifted over from the human reference genome to the chimpanzee reference genome. For this task, the 'liftOver' tool downloaded from the UCSC [19] web page was used.

[19] `http://hgdownload.cse.ucsc.edu/goldenPath/hg18/liftOver/`

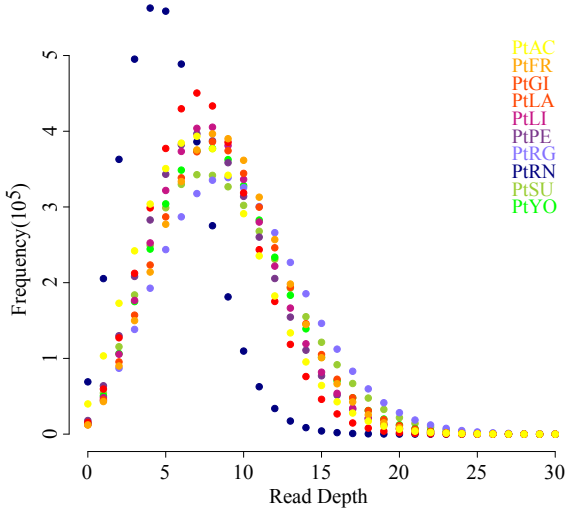


Figure 4.10: Read depth per individual at the polymorphic sites of the intersection data set.

The generated intersection set contained 4,005,229 autosomal SNPs with a Ts/Tv ratio of 2.25. On chromosome X, 108,959 SNPs were identified of which 1,477 SNPs were in the PAR (Ts/Tv: 1.95) and 107,482 SNPs (Ts/Tv: 2.11) were in the non-PAR of chromosome X (see Table 4.4). The read depth per individual at the polymorphic sites of the intersection set is illustrated in Figure 4.10.

As the SNPs found in the intersection of the chimpanzee-based and human-based calls are of higher quality than from the chimpanzee-based call alone (especially on the X chromosome, see Section 4.3.5), it is recommended to use them for chromosome X analyses.

4.3.4 Genotyping and Phasing

As the study used medium coverage sequence data, an imputation-based method was used to refine GATK's genotype likelihoods in a way similar to that adopted by the 1000 Genomes Project before haplotypes were estimated [The 1000 Genomes Project Consortium, 2010].

Work Flow GATK's genotype likelihoods were refined into genotypes using BEAGLE [20] [Browning and Yu, 2009; Browning, 2006; Browning and Browning, 2007], a program for imputing genotypes and inferring haplotype phase. Methods like BEAGLE that are based on imputation can not only improve genotype calls with a high degree of uncertainty through the usage of LD information but they can also be used to fill in missing data in the original genotype calls.

[20] `http://faculty.washington.edu/browning/beagle/beagle.html`

BEAGLE's initial estimates of the haplotype phase were subsequently used as a starting point for PHASE[21] [Stephens and Donnelly, 2003; Stephens and Scheet, 2005; Stephens et al., 2001], a Bayesian statistical method for reconstructing haplotypes from population genotype data [Browning and Browning, 2007]. For this purpose, the data was split into windows consisting of 401 SNPs, with a 100 SNP overlap between adjacent windows. The program was run for 200 iterations, with a burn-in period of 300 iterations using the following command line options:

```
PHASE -MR -l 10 -F 0.05 -X 5 -x 5 input output,
```

where -MR specifies the use of the recombination model, one out of two main models for haplotype reconstruction. PHASE's speed and accuracy of haplotype estimation result from the application of a divide-and-conquer strategy, 'Partition Ligation' [Niu et al., 2002; Stephens and Donnelly, 2003], which divides data into segments of consecutive loci for which a list of plausible haplotypes are computed. The maximum length of each segment is controlled using the -l flag. In order to obtain haplotypes across larger regions, segments are iteratively combined. -F specifies a threshold a haplotype's frequency estimate must surpass in order for that haplotype to be included in the list when merging sections. The -X option makes the final run for the entire data set X times longer than the other runs (X times the burn-in and X times the number of main iterations specified). -x specifies the number of times the algorithm is run from different starting points. The output results from the run with the best average value for the 'goodness of fit'.

The phased haplotypes in each window were subsequently spliced back together by choosing the phase across the splice point with the minimum Hamming distance between the haplotypes in the overlapping region.

Project Status A total of 390,797 genotypes (0.73%) from the chimpanzee-based mappings were fully imputed by BEAGLE. An additional 594,629 genotypes (1.12%) were replaced with a different call by the imputation process. All resulting BEAGLE haplotypes were subsequently rephased using PHASE.

4.3.5 Validation

SNP calls do not only contain true variation, but also artefacts usually caused by misaligned reads or systematic sequencing errors. In order to estimate the FDR of the SNP calls, a validation experiment was conducted to determine the presence or absence of SNPs called in short read sequencing

[21] http://stephenslab.uchicago.edu/software.html#phase

using a preliminary subset of the called variants and Sequenom, an independent technology. Sequenom uses a MassARRAY system for SNP genotyping which combines a reproducible primer extension chemistry (a single termination mix and universal reaction conditions are used for all SNPs) with MALDI-TOF mass spectrometry (differences in mass are used to differentiate between SNP alleles).

Work Flow 300 autosomal SNPs were randomly selected according to the SNP density and a further 115 SNPs were chosen from chromosome X. The minimal distance between two selected SNPs was set to be 1Mb on the autosomes and 100kb on chromosome X (the median distance between two SNPs in the validation set was 6.3Mb and 1.1Mb, respectively). The site frequency spectrum and the Ts/Tv ratio were verified to be similar to those obtained from the entire SNP call set.

SNPs were excluded if (i) the 200bp surrounding sequence contained either more than five SNPs according to dbSNP125 or the study's call set or missing bases in the reference (i.e. 'N'), (ii) the Sequenom software failed to find appropriate primers for the PCR (e.g. because there were too many repeats present in the sequence or it has had a high dimer or hairpin potential), or (iii) the identified primers matched more than twice in the panTro2.1 reference genome. The first criteria caused the exclusion of 4% of the autosomal SNPs and 16% of the SNPs on chromosome X, the second one eliminated 3% and 4% of the SNPs on the autosomes and chromosome X, respectively, whilst the third criteria excluded as much as 35% and 40% of the SNPs.

After the described filtering process, 236 SNPs, 180 on the autosomes and 56 on chromosome X, remained. Multiplexes with a maximum number of 30 SNPs per multiplex were designed using Sequenom's software leading to eight multiplexes of 28-30 SNPs with a total of 232 SNPs, 178 autosomal ones and 54 on chromosome X. The selected SNPs were genotyped in a panel of 93 chimpanzees (including the 10 subjects of the PanMap study). In order to allow for concordance calculations, three individuals (PtFR, PtPE, and PtRG) were genotyped twice.

Project Status The FDR of the preliminary autosomal data set from the chimpanzee-based mapping was estimated to be 5.9% (95% CI = 2.4-9.4%, via normal approximation of binomial distribution) based on 10 sites called as monomorphic and eight out of the 178 autosomal SNPs which were excluded from the calculations as they resulted in bad assays. In contrast, the FDR of the preliminary chromosome X SNPs called from the chimpanzee-based mapping was estimated to be 17.0% (95% CI = 6.9-27.1%) based on eight sites called as monomorphic

and excluding one out of the 54 SNPs which resulted in a bad assay. The considerably higher rate is most likely caused by a lower quality of the chimpanzee reference assembly of chromosome X compared to the autosomes. The reason for this is that the chimpanzee reference mainly originated from a single male Western chimpanzee resulting in less coverage of chromosome X used in the assembly (e.g. there are 28,367 gaps in the chromosome X reference sequence, compared to 9,519, 9,975, and 8,146 in the similarly sized chromosomes 6, 7, and 8, respectively). An overview of the initial Sequenom validation data is presented in Table 4.5.

The results of the validation experiment were used to revise and modify the initial calling and filtering strategy leading to the final filter criteria which are described in Section 4.3.3. 156 of the 178 autosomal SNPs and 39 out of 54 SNPs on chromosome X from the chimpanzee-based mapping were called in the final call set. The validation experiment has not been repeated on the final call set. Nevertheless, it is expected to be more conservative than the preliminary data set from which the SNPs for the validation were chosen. The implied FDR for the final autosomal call set was estimated to be 2.7% (95% CI = 0.08-5.2%) based on four sites called as monomorphic and excluding six SNPs which failed Sequenom assays. The implied FDR for the final chromosome X call set was estimated to be 7.9% (95% CI = 0-16.5%) based on three sites called as monomorphic and one out of the 39 SNPs which was excluded from the calculations as it resulted in bad assays (see Table 4.5).

As the FDR of the chromosome X SNP calls was relatively high, SNP calls based on sites that were called from both the chimpanzee-based and the human-based mappings should be used for further analyses. An intersection set was build from the chimpanzee-based and the human-based calls as described in Section 4.3.3 which included 126 of the 178 autosomal SNPs as well as 27 of the 54 chromosome X SNPs genotyped on Sequenom. The implied FDR for this set was estimated to be 2.5% (95 CI = 0-5.3%) on the autosomes and 7.7% (95% CI = 0-17.9%) on chromosome X based on three and two sites called as monomorphic and excluding six and one SNPs which failed Sequenom assays, respectively (see Table 4.5). Even if the FDR for the chromosome X calls of this set was similar to the one reported for the chimpanzee-based call, it is expected to be more conservative due to the higher quality and completeness of the human reference sequence. Therefore, it is recommended to use the intersection set for chromosome X analyses.

		Statistics		
		Autosomes	**Chromosome X**	**Total**
Good Clusters	**G**	101	30	131
Two Clusters	**T**	33	5	38
Monomorphic	**M**	10	8	18
Caution*	**C**	13	7	20
Skewed Clusters	**S**	3	1	4
Four Clusters	**F**	2	0	2
Bad Assay	**X**	8	1	9
	C/S	1	0	1
	G/C	2	0	2
	C/T	4	0	4
	C/T/S	1	0	1
	M/C	0	1	1
	T/S	0	1	1
	Total	178	54	232

		Statistics		
		Autosomes	**Chromosome X**	**Total**
Good Clusters	**G**	96	27	123
Two Clusters	**T**	28	3	31
Monomorphic	**M**	4	3	7
Caution*	**C**	11	4	15
Skewed Clusters	**S**	2	0	2
Four Clusters	**F**	2	0	2
Bad Assay	**X**	6	1	7
	C/S	1	0	1
	G/C	1	0	1
	C/T	4	0	4
	C/T/S	1	0	1
	M/C	0	0	0
	T/S	0	1	1
	Total	156	39	195

		Statistics		
		Autosomes	**Chromosome X**	**Total**
Good Clusters	**G**	77	21	98
Two Clusters	**T**	23	1	24
Monomorphic	**M**	3	2	5
Caution*	**C**	9	1	10
Skewed Clusters	**S**	2	0	2
Four Clusters	**F**	1	0	1
Bad Assay	**X**	6	1	7
	C/S	1	0	1
	G/C	1	0	1
	C/T	3	0	3
	C/T/S	0	0	0
	M/C	0	0	0
	T/S	0	1	1
	Total	126	27	153

Table 4.5: Sequenom validation experiment: Results of the preliminary SNP call (top), final SNP call (middle), and the intersection set of the chimpanzee-based and the human-based calls (bottom).
* = Some merging of clusters resulted in an increase of no-calls.

Individual	Genotype				
		0/0	**0/1**	**1/1**	**./.**
PtFR					
	0/0	133	1	0	8
	0/1	1	42	1	6
	1/1	0	0	21	1
	./.	1	2	1	11
PtPE					
	0/0	135	1	0	2
	0/1	1	51	2	0
	1/1	0	0	26	1
	./.	1	1	0	8
PtRG					
	0/0	136	1	0	2
	0/1	0	59	0	1
	1/1	0	0	20	0
	./.	1	0	0	9

Table 4.6: Genotype concordance on all sites of the three individuals Frits (PtFR), Pearl (PtPE), and Regina (PtRG) which have been genotyped twice in the performed Sequenom experiment.

The performed Sequenom experiment also enabled the assessment of the genotype accuracies of the three individuals which have been genotyped twice. The genotype concordances of PtFR, PtPE, and PtRG at non-failed assay sites were 98.5%, 98.2% and 99.5%, respectively, in the chimpanzee-based mapping (see Table 4.6). Overall, the genotype concordance for all non-failed Sequenom assays was 98.13% at autosomal SNPs while it rose to 98.80% for sites called as 'good' in the Sequenom assays. In comparison, the genotype concordance between Sequenom and the intersection set sequencing calls of chromosome X was 97.8% for all passed assays and 99.0% for all assays identified as 'good'.

4.3.6 Assessment of SNP Discovery Power

The power to detect sites segregating in the chimpanzee population is dependent on two factors: (i) whether the alternative allele is present amongst the subjects of the study and (ii) how many high quality reads cover the variant site in those individuals carrying the non-reference allele. In order to assess the ability of the PanMap study to discover SNPs of different frequencies, the data set was compared to two external data sources containing the same 10 chimpanzee individuals.

Work Flow First, the call set was compared to 1,081 ENCODE genotypes obtained from Winckler et al.'s study of 38 Western chimpanzees [Winckler et al., 2005]. As the genotypes were given in the hg15/hg16 coordinate system, the original coordinates were converted using UCSC's 'liftOver' tool resulting in a loss of 12 SNPs. The second data set used for comparison was the one of Myers et al. which consisted of 694 sites genotyped in 36 Western, 20 Central, and 17 Nigerian-Cameroon chimpanzee individuals [Myers et al., 2010]. 692 out of the 694 variant sites could be extracted in panTro2.1 coordinates from dbSNP. In both data sets, SNPs that

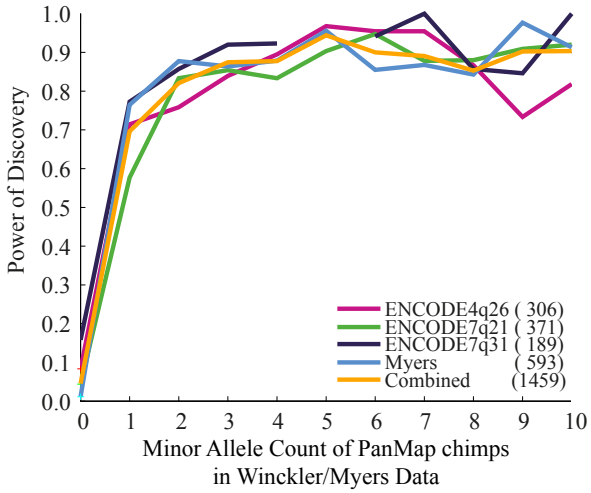


Figure 4.11: Power to detect SNPs present in the sample at different allele frequencies by comparison to previous SNP genotype data [Myers et al., 2010; Winckler et al., 2005]. This figure was kindly provided by Adam Auton.

were found to be non-polymorphic as well as SNPs with more than 50% missing data were excluded (203 and 99 for Winckler et al.'s and Myers et al.'s data set, respectively).

Project Status The discovery power for variants present at least once in the study samples was 86.83% (between 85% and 89.4% in the Winckler et al. study and 87.9% in the study of Myers et al.; see Figure 4.11). Above zero power to detect variants with a MAF of zero implies either a small degree of genotype error in the genotyping studies or false positives in the PanMap study which coincide with SNPs detected in previous studies.

4.4 Discussion

A map of genome-wide sequence variation in Western chimpanzees, *Pan troglodytes verus*, has been created by collecting and analysing genetic variation data obtained from high throughput Illumina sequencing of 10 individuals (nine females and one male) at an average coverage of 9.45X after mapping to the panTro2.1 reference genome. Calling variants from NGS data is far away from an 'off-the-shelf' technology. As described in the previous sections, it is a multi-step procedure which requires the application of different tools and algorithms. Its results are most strongly influenced by both the quality of the reference sequence as well as the choice of appropriate parameter values in the different steps involved in the process.

Despite the efforts of previous projects to increase the coverage of the initial chimpanzee reference genome published in 2005 by the Chimpanzee Sequencing and Analysis Consortium [Chimpanzee Sequencing and Analysis Consortium, 2005] from 3.6X to 6X, the entire genome remained at a rough draft state with many gaps and regions of uncertainty, a situation that has complicated

the variant calling immensely as many of the available tools for variant calling from NGS data were designed to work with the finished grade human reference genome. This issue became most obvious when trying to call SNPs on chromosome X whose reference sequence exhibited even lower coverage (as the sequences on which the reference was build mainly originated from a single male individual): across the autosomes, 5.3 million SNPs (Ts/Tv: 2.18) were discovered with a FDR of less than 3% (as determined through an independent technology) while the 175,380 SNPs (Ts/Tv: 1.97) that were identified on chromosome X exhibited a much higher FDR of ∼8%. This result suggested that systematic incompleteness of the chimpanzee reference sequence in repetitive regions had a significant effect on the variant calling process. As a consequence, all reads have been additionally aligned to a human reference sequence in order to identify SNPs mapping in places with reliable synteny between the two species and to ensure robustness of the results. Although this strategy helped to overcome some of the problems (e.g. the 108,959 SNPs on chromosome X with a Ts/Tv ratio of 2.11 are thought to be of substantially higher quality in the intersection set), in retrospect, a *de novo* assembly could have been a valuable alternative to the reference-guided approach, not only because the study generated large amounts of data (>90X coverage compared to the 6X coverage of the current reference genome), but also because it might have helped to improve the reference genome by removing some of the biases which might have been previously introduced by assembling the chimpanzee reference through alignment to the human genome, a strategy that took the risk that certain segments of the genome might have been misassembled (e.g. rapidly evolving segments such as regions that have either been duplicated in one of the two lineages or that have been selectively deleted in humans).

Many steps in the variant calling process required the setting of a whole range of parameter values (e.g. cutoffs for SNP filtering) which were critical to achieve reliable results. However, the choice of these values (which might vary between different projects depending on their goals) were usually only based on either expert knowledge or on visual inspection of the data. As a consequence, it is very difficult to follow a strict work flow protocol, making variant calling from high throughput sequencing data more of an art than a solid science.

Data Availability

The sequence data (in BAM file format) as well as the variant calls (in VCF file format) are available for download over anonymous FTP from birch.well.ox.ac.uk.

Preface

The mask files have been generated by Adam Auton and myself (as described in Section 5.2). The genetic map used to study the influence of recombination rates on variation in nucleotide diversity in Western chimpanzees (Section 5.3.2.1) was created by Adam Auton and Oliver Venn. I conducted all analyses with input from Gil McVean.

Chapter 5

PanMap - Analysis of Genetic Diversity in Western Chimpanzees

Mutations continuously give rise to new variations in the genome sequence of an organism - a diversity which can be regarded as the raw material for evolution. Evolutionary forces such as recombination, natural selection, and genetic drift can act on this variation in order to influence the rate of evolution and thus, the study of genetic diversity can give insights into the processes that created and shaped this variation. However, the overall extent of genetic variability is not only influenced by the underlying evolutionary forces, but also by the demographic history of a population: events such as bottlenecks, splits, migrations, expansions, or contractions can change the effective size of a population and, as a consequence, can cause mutations to accumulate in characteristic ways. As an example, the expansion of the human population in the past 100,000 years from a few thousand individuals to several billion generated a vast amount of genetic variation, most of which only occurs in a few individuals as the elapsed time was not sufficient to spread those rare variants within the population in the absence of selection [Wright, 2008]. These imprints on local and/or global patterns of sequence variation allow, at least in theory, to trace back the history of a species as a whole by surveying variation in genomes of a sufficient number of individuals from different populations.

In contrast to the vast amounts of data on intra-specific genetic diversity in humans, available through projects such as the International HapMap Project and the 1000 Genomes Project, there was, at least until recently, only limited polymorphism data publicly available for great apes, and in particular for our closest evolutionary relative, the chimpanzee, despite its evolutionary significance. For this reason, only a few studies on the genetic variation in non-human primates have been published. Although most of this previous work has been restricted to distinct genomic locations (such as particular protein variants [King and Wilson, 1975; Sarich and Wilson, 1967] or mitochondrial DNA [Ferris et al., 1981; Morin et al., 1994; Wise et al., 1997]), they enabled

a first glimpse of whether or not the extent of genetic diversity seen in the human genome is representative for large primates [Deinard and Kidd, 1999; Fischer et al., 2004; Gagneux et al., 1999; Gilad et al., 2003; Kaessmann and Paabo, 2002; Kitano et al., 2003; Stone et al., 2002; Yu et al., 2003]. The results of the performed analyses have been in disagreement. Most of the published work indicated that humans exhibit lower levels of intra-specific nuclear sequence variation than most other mammals, including a two- to four-fold lower nucleotide diversity than the chimpanzee [Chimpanzee Sequencing and Analysis Consortium, 2005; Deinard and Kidd, 1999, 1998; Jensen-Seaman et al., 2001; Kaessmann et al., 1999b; Satta, 2001; Stone et al., 2002]: (i) an analysis of a 1,000bp segment of the intergenic non-coding *HOXB6* locus detected a nucleotide diversity of 0.19% in chimpanzees and 0.06% in humans based on eight and four polymorphic sites, respectively [Deinard and Kidd, 1999], (ii) two analyses of a 10,000bp long segment on chromosome X (Xq13.3) observed 33 substitutions in 70 humans and 84 substitutions in 30 chimpanzees leading to nucleotide diversities in the range of 0.033% and 0.13%, respectively [Kaessmann et al., 1999a,b], (iii) a greater nucleotide diversity in chimpanzee has also been found on chromosome Y [Stone et al., 2002], as well as (iv) in mitochondrial DNA [Gagneux et al., 1999; Morin et al., 1994; von Haeseler et al., 1996; Wise et al., 1997], and (v) although a few genes of the major histocompatibility complex showed a greater diversity in humans than in chimpanzee [Bontrop et al., 1995], the majority of the class I and class II genes support the greater diversity in chimpanzees [Adams et al., 2000; Bontrop et al., 1995]. Despite the fact that both variations in mutation rates and selection as well as population subdivision could have resulted in a higher nucleotide diversity in chimpanzees compared to humans, previous work indicated that these factors were unlikely the cause of the observed differences [Deinard and Kidd, 1999; Kaessmann et al., 2001]. In contrast, a larger effective population size in chimpanzees over time might reflect a greater time depth for intra-species variation [Deinard and Kidd, 1999; Kaessmann et al., 2001]. Inconsistent with these findings, other studies have detected less variability in great apes than in humans in certain regions of the genome (e.g. in STRs [Cooper et al., 1998; Wise et al., 1997] as well as in some genes of the MHC [Bontrop et al., 1995]), an observation that might, amongst other factors, result from the restricted geographical habitants of non-human primates.

Most of the previous analyses were only based on a set of specific genes or genomic locations. Thus, their results should be handled with care as certain ascertainment biases are likely to have influenced the outcomes (e.g. most studied STR-markers have originally been selected for their high variability in humans). In contrast, the genetic data generated by the PanMap project (described in Chapter 4) has a much wider genomic representation, making it well suited (and most likely also more reliable) for the study of diversity in chimpanzees at a genome-wide scale.

5.1 Aim of the Study

The study of nucleotide diversity at a genome-wide scale is complicated by the fact that the levels of diversity are dependent on the mutation rate per site per generation as well as the effective population size, both of which are known to vary between different genomic regions [Drake, 1992; Duret, 2009; The 1000 Genomes Project Consortium, 2011; Wolfe et al., 1989]. The effective population size varies across the genome due to evolutionary forces (such as the interplay of acting selective pressures and homologous meiotic recombination) as well as due to the different demographic histories of different loci [Thornton et al., 2007] - an effect that is greatest between the sex chromosomes and the autosomes [Hedrick, 2007; Schaffner, 2004; Vicoso and Charlesworth, 2006]. Similarly, mutation rates are subject to a multitude of evolutionary pressures causing them to vary with gender, with genomic location (high rates have been detected for sequences containing for instance CpG-dinucleotides, repeats, or a high purine content [Gojobori et al., 1982; Krawczak et al., 1998; Li et al., 1984; Siepel and Haussler, 2004]), as well as with other genomic features such as recombination rate [Hellmann et al., 2003]. This variation can not only occur between different chromosomes (between autosomes and sex chromosomes as well as among autosomes) [Chen et al., 2001; Chimpanzee Sequencing and Analysis Consortium, 2005; Ebersberger et al., 2002; Gaffney and Keightley, 2005; Lercher and Hurst, 2002; Lercher et al., 2001; Li et al., 2002; Malcom et al., 2003; The Mouse Genome Sequencing Consortium, 2002], but also regionally (at both kilobase and megabase scales) within individual chromosomes [Lercher et al., 2001; Matassi et al., 1999; Silva and Kondrashov, 2002; Smith et al., 2002; Williams and Hurst, 2000] with the largest variation occurring at the smallest scale, at individual sites of the genome, where mutation rates are frequently influenced by their nucleotide sequence-context [Blake et al., 1992; Gojobori et al., 1982; Hwang and Green, 2004; Zhao and Boerwinkle, 2002].

As a result of both differences in mutation rate and effective population size, levels of nucleotide diversity within a single genome are expected not to be uniform but to vary greatly both between different chromosomes as well as regionally within chromosomes [Ellegren et al., 2003]. The aim of this study is to use the PanMap data set (described in Chapter 4) to explore the breadth of the nucleotide diversity of Western chimpanzees at a genome-wide scale. This will allow to place human nucleotide diversity into perspective with an evolutionary closely related relative. It will enable to shed light on whether or not the local variation in mutation rate has been conserved since the divergence of the two species. In addition, it might also offer new insides into diversity patterns which are unique in the chimpanzee or the human lineage which might give information about chimpanzee and/or human history.

5.2 Accessible Genome

Candidate SNPs from the initial PanMap SNP call were subject to several filter criteria in order to avoid false positives and to control the Ts/Tv ratio, a commonly used assessor to evaluate the quality of a SNP call, while maintaining as many SNPs as possible (see Section 4.3.3 for details of the filtering process). As the applied filter metrics can lead to the exclusion of a substantial fraction of sites in the genome (especially those on coverage and the fraction of reads with low mapping quality scores), mask files, defining which nucleotides were accessible to the variant discovery in the study, were generated for both the panTro2-based mapping and the hg18-based mapping to enable population genetic analysis.

Work Flow Mask files were generated using GATK's 'UnifiedGenotyper' (Version 1.0.4705) [DePristo et al., 2011; McKenna et al., 2010] with the same parameters that were used to make the initial PanMap SNP calls (as described in Section 4.3.2.4), with the exception of the '- all_bases' option which forces GATK's algorithm to make a call at every site for which there was data available:

```
java -jar GenomeAnalysisTK.jar \
     -T UnifiedGenotyper \
     -I PtAC_rmdup_recal_baq.bam \
     -I PtFR_rmdup_recal_baq.bam \
     -I PtGI_rmdup_recal_baq.bam ... \
     -R reference.fasta  \
     -all_bases \
     -o GATK_snp_call_all_bases.vcf,
```

where -I represents the input file(s) (one input file per chimpanzee individual; all 10 individuals were used to make joint calls from the autosomes and the PAR whilst only the nine females were used to call from the non-PAR of chromosome X), -R indicates the reference sequence, -all_bases specifies that a call will be made at every site for which there was data available, and -o represents the output file in which the calls should be stored.

Similar to the SNP calling procedure, the initial calls were filtered using GATK's 'VariantFiltration' (Version 1.0.4705). The same filter criteria were used as for the initial PanMap SNP call (see Section 4.3.3), although it should be noted that certain metrics can not be easily interpreted in the absence of a SNP call. The only exception was the SNP cluster

filter (- -clusterWindowSize) which can not be applied in the case of a call at every site in the genome:

```
java -jar GenomeAnalysisTK.jar \
     -T VariantFiltration \
     -I PtAC_rmdup_recal_baq.bam \
     -I PtFR_rmdup_recal_baq.bam \
     -I PtGI_rmdup_recal_baq.bam ... \
     -R reference.fasta  \
     -B:variant,VCF GATK_snp_call_all_bases.vcf \
     -o GATK_snp_call_all_bases_filtered.vcf \
     --filterExpression 'AF>=1.0 || AN<16 || DP<30 || DP>180 || \
                         Dels>0.0 || HaplotypeScore>40.0 || \
                         HRun>2 || MQ<25.0 || MQ0>40 || \
                         (MQ0 \ DP)>0.4 || QD>27.0 || QD<1.0 || SB>0.0',
```

where -I represents the input file(s), -R specifies the reference sequence, -B indicates the file containing the initial calls, -o specifies the output file, and - -filterExpression indicates the previously described filter criteria (in this example for the autosomes/PAR).

Project Status Mask files were generated for both the panTro2-based data set and the hg18-based data set. For the former, the accessible genome contained on average 77.5% of the reference sequence on the autosomes, 26.1% of the reference sequence in the PAR, and 81.8% of the reference sequence in the non-PAR of chromosome X (see Table 5.1(a); note that for chromosome X, sites were considered accessible when the filter status was either 'PASS' or low quality, 'LowQual', where every individual that has been called, has been called as homozygous reference, but some individuals were missing). On chromosome X, 92.6% of the inaccessible sites in the PAR and 82.2% of the inaccessible sites in the non-PAR of chromosome X did not exhibit any data.

For the hg18-based mapping approach, the accessible genome contained on average 79.8% of the reference sequence on the autosomes, 65.0% of the reference sequence in the PAR, and 91.0% of the reference sequence in the non-PAR of chromosome X (see Table 5.1(b); accessibility on chromosome X was defined as for the panTro2-based data set). On chromosome X, 45.4% of the inaccessible sites in the PAR and 29.7% of the inaccessible sites in the non-PAR of chromosome X did not exhibit any data.

(a)

Chromosome	Statistics				
	Callable[1]	**LowQual**$^{2}_{\text{missing}}$	**LowQual**$^{3}_{\text{not\_homref}}$	**Failed Filters**[4]	**No data/Other**
1	85.65%	2.71%	0.03%	5.94%	5.68%
2a	83.62%	2.66%	0.03%	6.08%	7.62%
2b	47.00%	1.41%	0.01%	2.95%	48.63%
3	87.33%	2.55%	0.03%	5.57%	4.53%
4	87.05%	2.80%	0.03%	5.93%	4.19%
5	86.35%	2.65%	0.03%	6.10%	4.88%
6	85.39%	2.65%	0.03%	6.52%	5.41%
7	84.07%	2.80%	0.03%	7.24%	5.86%
8	85.93%	2.69%	0.03%	6.46%	4.90%
9	70.67%	2.41%	0.02%	5.71%	21.19%
10	84.24%	2.64%	0.03%	6.08%	7.01%
11	83.72%	2.71%	0.03%	5.53%	8.02%
12	87.12%	2.75%	0.03%	5.92%	4.19%
13	69.19%	2.34%	0.02%	4.12%	24.32%
14	72.82%	2.49%	0.02%	4.92%	19.75%
15	69.68%	2.15%	0.02%	4.98%	23.18%
16	71.96%	2.84%	0.03%	7.21%	17.96%
17	78.26%	3.18%	0.03%	6.46%	12.07%
18	87.59%	2.65%	0.03%	5.64%	4.10%
19	69.14%	4.19%	0.03%	7.13%	19.51%
20	84.99%	2.81%	0.03%	5.31%	6.86%
21	65.10%	1.87%	0.02%	3.26%	29.75%
22	56.27%	2.76%	0.02%	5.31%	35.64%
X	5.41%	76.39%	0.02%	3.22%	14.97%
PAR	22.96%	3.17%	0.02%	5.43%	68.42%

(b)

Chromosome	Statistics				
	Callable[1]	**LowQual**$^{2}_{\text{missing}}$	**LowQual**$^{3}_{\text{not\_homref}}$	**Failed Filters**[4]	**No data/Other**
1	79.19%	5.91%	0.03%	5.47%	9.40%
2	86.54%	5.51%	0.03%	5.33%	2.59%
3	88.26%	4.44%	0.03%	4.44%	2.83%
4	87.12%	5.21%	0.03%	5.11%	2.53%
5	86.65%	6.13%	0.03%	5.02%	2.17%
6	86.94%	5.00%	0.03%	5.47%	2.56%
7	83.58%	6.49%	0.03%	7.00%	2.91%
8	86.36%	5.34%	0.03%	5.32%	2.95%
9	69.22%	10.35%	0.03%	5.68%	14.72%
10	84.57%	6.65%	0.03%	5.53%	3.23%
11	85.81%	5.64%	0.03%	5.62%	2.90%
12	88.64%	4.47%	0.03%	4.85%	2.01%
13	75.62%	4.02%	0.03%	3.67%	16.66%
14	73.25%	5.01%	0.03%	4.35%	17.37%
15	69.15%	6.57%	0.02%	4.95%	19.31%
16	72.91%	7.41%	0.03%	7.98%	11.67%
17	84.02%	7.20%	0.03%	7.00%	1.75%
18	88.50%	4.40%	0.03%	4.71%	2.36%
19	72.39%	7.65%	0.04%	6.78%	13.14%
20	85.73%	4.43%	0.03%	4.65%	5.16%
21	62.87%	4.89%	0.02%	4.59%	27.62%
22	57.53%	6.87%	0.03%	5.32%	30.26%
X	5.30%	85.67%	0.04%	6.32%	2.68%
PAR1	49.12%	15.89%	0.08%	19.01%	15.89%

Table 5.1: Accessible sites per chromosome in the (a) panTro2-based and (b) hg18-based data set.
[1] Filter status is either 'PASS' or 'LowQual' (low quality) with all individuals called as homozygous reference.
[2] Filter status is 'LowQual' and every individual that has been called, has been called as homozygous reference, but some individuals were missing.
[3] Filter status is 'LowQual', but some of the individuals were not called as homozygous reference.
[4] Failed one or more of the filter criteria (neither 'PASS' nor 'LowQual').

5.3 Genome-wide Patterns of Nucleotide Diversity

Levels of nucleotide diversity depend on the mutation rate per site per generation which is known to vary between different regions of the genome as a result of a multitude of evolutionary pressures [Mugal and Ellegren, 2011; Wolfe et al., 1989]. Mutation rates vary between chromosomes [Chen et al., 2001; Ebersberger et al., 2002; Lercher et al., 2001; Li et al., 2002; The Mouse Genome Sequencing Consortium, 2002] as well as regionally (at both kb- and Mb-scales) within an individual chromosome [Lercher et al., 2001; Matassi et al., 1999; Silva and Kondrashov, 2002; Smith et al., 2002; Williams and Hurst, 2000], while the largest variation can be detected at individual sites of the genome, where mutation rates are often influenced by their nucleotide sequence-context [Blake et al., 1992; Gojobori et al., 1982; Hwang and Green, 2004; Zhao and Boerwinkle, 2002]. The strongest sequence-context effect has been observed for CpG-dinucleotides, for which the majority of mutations are caused by spontaneous methylation-dependent deamination [Chimpanzee Sequencing and Analysis Consortium, 2005; Cooper and Youssoufian, 1988; Coulondre et al., 1978; Duret, 2009; Ehrlich and Wang, 1981; Holliday and Grigg, 1993; Hwang and Green, 2004; Xing et al., 2004]. In contrast, most non-CpG mutations are generated during DNA replication [Kim et al., 2006]. As a consequence of the different underlying mechanisms, mutation rates at CpG- and non-CpG-sites are thought to be influenced by different internal genomic factors (such as recombination rates as well as sequence characteristics which decrease or increase local mutation rates in the genome [Hess et al., 1994; Krawczak et al., 1998; Siepel and Haussler, 2004; Zhang et al., 2007; Zhao and Boerwinkle, 2002; Zhao and Zhang, 2006]) and thus, these two types should be studied separately.

To analyse patterns of CpG- and non-CpG nucleotide diversity both between different chromosomes as well as within individual chromosomes, each chromosome was divided into continuous windows of different sizes. For each window, π, defined as the number of nucleotide differences per site between two randomly chosen sequences [Nei and Li, 1979], as well as Watterson's estimate of the parameter Θ, a measure of genetic variability based on the number of variable positions corrected by the number of individuals in a sample [Watterson, 1975], were calculated to evaluate the extent of polymorphism within chimpanzee individuals. In a model in which sites evolve neutrally, π should estimate the population genetics parameter $\Theta = 4N_e\mu$. The relationship of regional nucleotide diversity levels in chimpanzees with specific genome features (such as divergence, recombination rates, GC-content, exon content, DNA repeat content, CpG-islands content, distance to the centromere and telomeres, as well as chromosome size) was studied to reveal whether they were any correlations between nucleotide diversity at large scales and these sequence features.

5.3.1 Nucleotide Diversity between different Chromosomes

To look at patterns of nucleotide diversity between chromosomes, each chromosome was divided into continuous windows of 1Mb size (based on chromosome coordinates), for each of which both π and Θ were estimated. Nucleotide diversity levels were averaged over all windows of a chromosome which exhibited at least 80% accessibility after filtering (see Section 5.2).

5.3.1.1 Nucleotide Diversity between Autosomes and Chromosome X

In Western chimpanzees, nucleotide diversity levels on chromosome X (observed from the PanMap data) were lower than on autosomes ($\pi = 4.6 \times 10^{-4}$, corresponding to 67% diversity on the autosomes; see Table 5.2 and Figure 5.1), a finding which was in concordance with previously reported differences of nucleotide diversity levels between chromosome X and autosomes in humans [Sachidanandam et al., 2001].

The underlying population genetic cause might be linked to lower rates of point mutations in females than males in the two species due to the male's requirement of a greater number of germ line cell divisions before meiosis per generation during spermatogenesis (compared to oogenesis) in which more mutations are generated by DNA replication errors [Haldane, 1947]. As a result of this male-driven evolution, mutation rates depend on the way the chromosomes are transmitted through the male and female germ lines: mutation rates are usually highest on chromosome Y as it is exclusively transmitted through the male germ line, intermediate on autosomes which spend half their time in males and half their time in females, and lowest on chromosome X which only spends a third of its time in the male sex [Chimpanzee Sequencing and Analysis Consortium, 2005; Ebersberger et al., 2002; Gaffney and Keightley, 2005; Goetting-Minesky and Makova, 2006; Lercher and Hurst, 2002; Malcom et al., 2003; McVean and Hurst, 1997; Wolfe and Sharp, 1993]. Given the different modes of inheritance, the relative contribution of males and females to evolutionarily accumulated mutations, the male-to-female mutation rate α, can be studied using substitution rates at supposedly neutrally evolving genomic regions (e.g. non-coding regions of the genome [Hardison et al., 2003; Taylor et al., 2006]) between any two chromosome types as they serve as surrogate for mutation rates [Miyata et al., 1987]. In chimpanzees, estimates for the male-to-female mutation rate from genome-wide sequence data range from 3.6 [Sayres et al., 2011] to 6.2 [Presgraves and Yi, 2009], depending on which chromosome types were compared as well as which mutation types were included in the analysis, and whether or not ancestral polymorphisms were corrected for (as they can inflate or deflate the estimates of α in very closely related species) [Makova and Li, 2002; Pink et al., 2009; Taylor et al., 2006; Wilson Sayres and Makova, 2011].

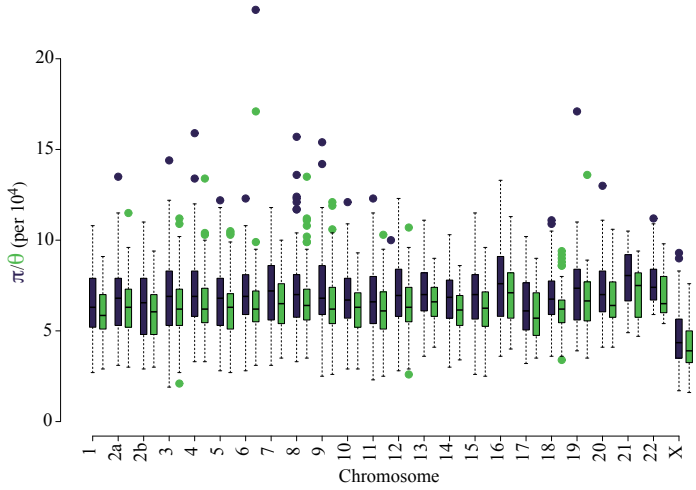


Figure 5.1: Distribution of π (blue) and Θ (green) estimates by chromosomes. Calculated using 1Mb windows which had at least 80% accessibility after filtering (corresponding to 81.7% of the genome).

The different transmission of the chromosomes through the germ lines does not only influences the role of mutation, but also of other evolutionary factors such as selection and genetic drift, as well as demographic factors. As chromosome X is hemizygous in males, all mutations on chromosome X are directly exposed in the male (they are usually also effectively hemizygous in the homogametic sex due to dosage compensation) [McVean and Hurst, 1997]. This results in a stronger influence of both background selection against recessive deleterious mutations as well as positive selection for recessive beneficial mutations [Ellegren, 2009; McVean and Hurst, 1997], the latter of which can reduce diversity levels by hitchhiking [Betancourt et al., 2004; Charlesworth et al., 1987]. In addition, genetic drift can work faster on chromosome X than on autosomes (chromosome X has a reduced effective population size N_e of 3/4 as it undergoes a male meiosis only one third of the time) [Schaffner, 2004].

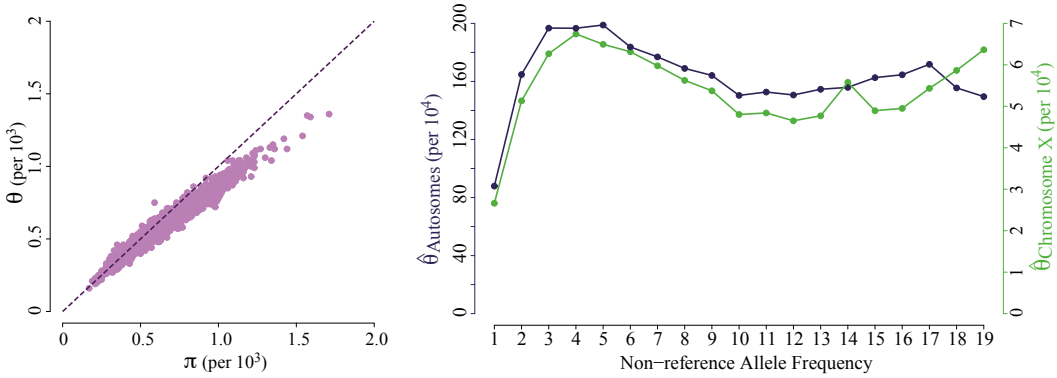
5.3.1.2 Nucleotide Diversity between different Autosomes

In contrast to chromosome X, most autosomes exhibited similar levels of nucleotide diversity (as illustrated in Figure 5.1). Although smaller chromosomes often show differences in several sequence features (such as GC-content and recombination rate), they did not tend to have different levels of nucleotide diversity. In fact, the nucleotide diversity of all autosomes was within the range of one SD of the genome-wide average ($\pi = \Theta = 6.9 \times 10^{-4}$ with a SD of 1.9×10^{-4}). Only three autosomes (namely chromosome 16, 21, and 22) exhibited slightly higher values (outside the 10% genome-wide average for autosomes) of nucleotide diversity ($\pi = 7.7 \times 10^{-4}$, $\pi = 7.9 \times 10^{-4}$, and $\pi = 7.8 \times 10^{-4}$, respectively). Notably, chromosome 21 is also an outlier in humans, showing, in contrast to chimpanzee, less diversity than the 10% genome-wide autosomal average [Sachidanandam et al., 2001]. However, the reasons for the observed small variation of autosomal diversity are not entirely

Chromosome	**Statistics**						
	$\Theta_{\mathbf{Chimp}}$ $(\times 10^{-4})$	$\Theta_{\mathbf{Chimp,CpG}}$ $(\times 10^{-5})$	$\Theta_{\mathbf{Chimp,non\text{-}CpG}}$ $(\times 10^{-4})$	$\pi_{\mathbf{Chimp}}$ $(\times 10^{-4})$	$\pi_{\mathbf{Chimp,CpG}}$ $(\times 10^{-5})$	$\pi_{\mathbf{Chimp,non\text{-}CpG}}$ $(\times 10^{-4})$	$\pi^{*}_{\mathbf{Human}}$ $(\times 10^{-4})$
1	6.0	9.9	5.0	6.5	10.8	5.5	7.7
2a	6.3	9.5	5.3	6.5	10.2	5.8	NA
2b	6.0	9.5	5.1	6.5	9.8	5.5	NA
2	NA	NA	NA	NA	NA	NA	7.4
3	6.3	8.8	5.4	6.8	9.4	5.9	7.5
4	6.5	8.7	5.6	7.2	9.6	6.2	8.1
5	6.2	8.7	5.3	6.7	9.4	5.7	7.2
6	6.4	9.6	5.4	7.0	10.5	6.0	7.4
7	6.6	9.7	5.6	7.2	10.6	6.2	7.6
8	6.5	9.5	5.6	7.2	10.4	6.1	7.7
9	6.5	9.2	5.6	7.2	10.0	6.2	8.1
10	6.2	10.4	5.2	6.8	11.3	5.7	8.3
11	6.0	9.2	5.1	6.6	10.0	5.6	8.4
12	6.4	9.9	5.4	7.1	10.9	6.0	7.6
13	6.5	9.4	5.6	7.1	10.2	6.1	8.0
14	6.1	9.2	5.2	6.7	10.1	5.7	7.4
15	6.2	9.9	5.2	6.8	10.7	5.7	8.8
16	7.1	11.8	5.9	7.7	12.7	6.4	8.3
17	5.9	11.7	4.8	6.4	12.5	5.1	7.8
18	6.2	9.5	5.3	6.9	10.4	5.8	8.1
19	6.7	10.4	5.6	7.4	11.4	6.2	7.6
20	6.6	11.7	5.4	7.2	12.6	5.9	7.2
21	7.1	12.0	5.9	7.9	13.3	6.6	5.2
22	7.0	15.4	5.5	7.8	17.0	6.0	8.6
X	4.1	4.7	3.6	4.6	5.2	4.1	4.7

* Data taken from Sachidanandam et al. [2001].

Table 5.2: Nucleotide diversity in Western chimpanzees between chromosomes averaged over 1Mb windows which had at least 80% accessibility after filtering (corresponding to 81.7% of the genome). Nucleotide diversity levels were measured separately at CpG-sites and non-CpG-sites. For comparison, nucleotide diversity levels in humans are also shown (data was taken from Sachidanandam et al. [2001]).



(a) (b)

Figure 5.2: (a) Comparison of estimates for π and Θ. (b) Watterson's estimate for $\hat{\Theta}$ at different non-reference allele frequencies.

clear: while they might be caused by statistical fluctuations or methodological problems, there is also a chance that they might be meaningful from a biological point of view, in which case differences on an autosomal level might either be entirely driven by regional variation or by other, currently unknown, forces which might play an additional role.

Although π should estimate the population genetic parameter $\Theta = 4N_e\mu$ at selectively neutral sites, a discrepancy between the two estimates of nucleotide diversity was detected: estimates for π were nearly consistently higher than for Watterson's estimate of Θ (see Figure 5.2(a)). This observation is most likely caused by a combination of stochastical fluctuations (for example, different phylogenetic tree shapes can lead to differences in the estimates for π and Θ, e.g. estimates for π should be higher for trees with variants occurring at high frequencies while estimates for Θ should be higher for star-shaped trees) and an underestimation of $\hat{\Theta}$ due to an ascertainment bias in the PanMap data set (see Figure 5.2(b)).

5.3.2 Nucleotide Diversity within Chromosomes

The study of nucleotide diversity levels within individual chromosomes is of great interest due to the fact that some of the observed variation might reflect acting selective pressures [Charlesworth et al., 1995; Hudson and Kaplan, 1995; Kaplan et al., 1989; Nordborg et al., 1996; Simonsen et al., 1995]. Nevertheless, other parameters, such as mutation rates, are known to influence regional diversity levels and thus, they will need to be taken into account when trying to interpret patterns of diversity. Mutation rates, which are equal to divergence rates in neutrally evolving regions of the genome [Li, 1997], are known to vary within chromosomes [Ebersberger et al., 2002; Lercher and Hurst, 2002; Malcom et al., 2003; Wolfe et al., 1989]. Thereby, the strongest variation is usually observed at individual sites [Hwang and Green, 2004] where different mechanisms cause

mutations at CpG- and non-CpG-sites: at CpG-sites, the majority of mutations result from spontaneous methylation-dependent deamination [Chimpanzee Sequencing and Analysis Consortium, 2005; Cooper and Youssoufian, 1988; Coulondre et al., 1978; Duret, 2009; Ehrlich and Wang, 1981; Holliday and Grigg, 1993; Hwang and Green, 2004; Xing et al., 2004] while most non-CpG mutations are generated by errors in the DNA replication process [Kim et al., 2006]. As a consequence, CpG- and non-CpG mutations should be studied separately as different internal genomic factors might affect these two processes. Genomic features which are either directly or indirectly linked with DNA damage or DNA repair mechanisms (e.g. simple repeats or gene density) [Hellmann et al., 2005; Holmquist, 1992; Surralles et al., 2002] might also influence mutation rates at larger scales. One factor of particular importance is the underlying recombination rate as the exact effect of selection on nucleotide diversity levels depends strongly on it: in contrast to the levels of divergence, which, at neutral sites, do not depend on natural selection at linked sites, both positive selection [Hudson, 1994; Kaplan et al., 1989; Przeworski, 2002; Smith and Haigh, 1974] as well as negative selection [Andolfatto, 2001; Aquadro et al., 2001; Begun and Aquadro, 1994; Charlesworth et al., 1993; Hudson and Kaplan, 1995; Kim and Stephan, 2000; Nordborg et al., 1996] reduce levels of diversity at linked neutral sites [Hellmann et al., 2005] and thus, variation-reducing selection will have a greater impact in regions with low recombination rate (as linkage will break down less quickly) [Nachman, 1997; Nachman et al., 1998; Przeworski et al., 2000].

To assess the effect of some of the above mentioned factors on large-scale patterns of nucleotide diversity, the genome was divided into continuous windows of 5Mb, 2Mb, and 1Mb size (based on chromosome coordinates) and both π and Θ were estimated for all windows which exhibited at least 80% accessibility after filtering (see Section 5.2). On the 1Mb-scale, nucleotide diversity levels across the chimpanzee genome ranged from 1.9×10^{-4} to 2.3×10^{-3} with an average value of 6.9×10^{-4} (SD $= 1.9 \times 10^{-4}$). As discussed in Section 5.3.1, chromosome X exhibited the smallest average levels of nucleotide diversity. In general, the pattern of nucleotide diversity across the genome was quite variable (see Figure 5.3): 70.3% of the windows exhibited a value of π within the range of one SD of the genome-wide average. Regions of low genetic diversity frequently corresponded to telomeric and centromeric regions (e.g. a region of low nucleotide diversity is located at the beginning of chromosome 22 which includes the centromere). Highly diverse regions of the genome might have been subject to balancing selection (e.g. the MHC region located at the 31Mb-34Mb region of chromosome 6 which contains some of the most polymorphic genes in vertebrate genomes known to date [Clark, 1997; Fan et al., 1989; Gyllensten and Erlich, 1989; Hughes and Nei, 1988, 1989; Klein, 1987; Lawlor et al., 1988; Mayer et al., 1988]), but other explanations, such as regions of high mutation rate and/or recombination or unrecognised duplications in the chimpanzee

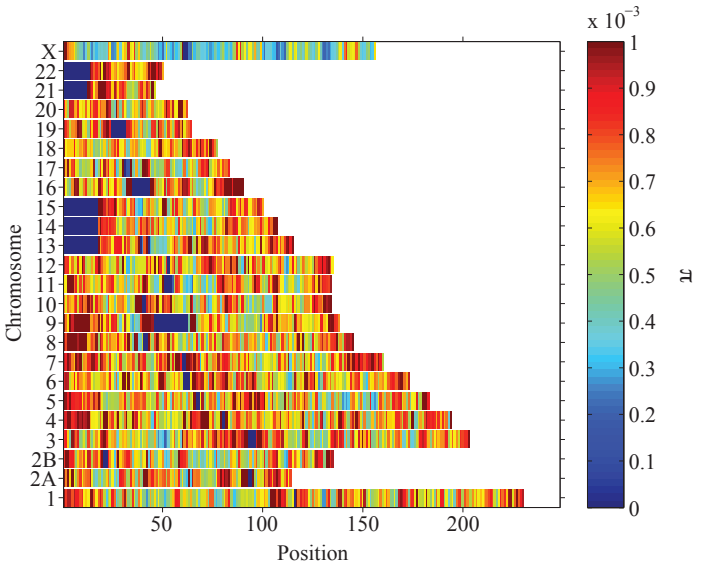


Figure 5.3: Nucleotide diversity levels in Western chimpanzees across chromosomes (estimated in 1Mb windows which had at least 80% accessibility after filtering, corresponding to 81.7% of the genome).

genome, can not be excluded. In fact, previous studies in humans suggested that π can vary 10- to 100-fold between individual genes or subchromosomal regions of the genome due to, amongst others, differences in chromosomal architecture, biological or evolutionary forces (such as different rates of mutation and recombination or natural selection), or population demographic effects (such as bottlenecks, expansions, or admixture) [Brookes, 2008; Kimmel et al., 1998; Reich and Goldstein, 1998]. In Western chimpanzees, nucleotide diversity levels at the 1Mb-scale along single chromosomes have been found to vary strongly as a function of divergence rate, recombination rate, and GC-content as illustrated in Figure 5.4.

5.3.2.1 Predictors of Nucleotide Diversity within Chromosomes

Variations in nucleotide diversity levels can be caused by (i) genetic drift (resulting in variations in the mean coalescence time), (ii) selection, as well as (iii) variations in the underlying mutation rate. The contribution of different genomic factors influencing one or more of these three aspects to the observed 1Mb-scale variation in diversity levels across the genome of Western chimpanzees was assessed using multiple linear regression. In particular, GC-content, divergence rates, recombination rates, exon content, CpG-island content, simple repeat content, distance to telomeres and centromeres, chromosome size, as well as larger-scale diversity (at the 5Mb- and 2Mb-level, measured at the two regions flanking a 1Mb-window) were considered as potential predictors of diversity. From these sequence features used to predict diversity, recombination rates are informative about the effective population size N_e while all other predictors are informative about both N_e and the mutation rate μ.

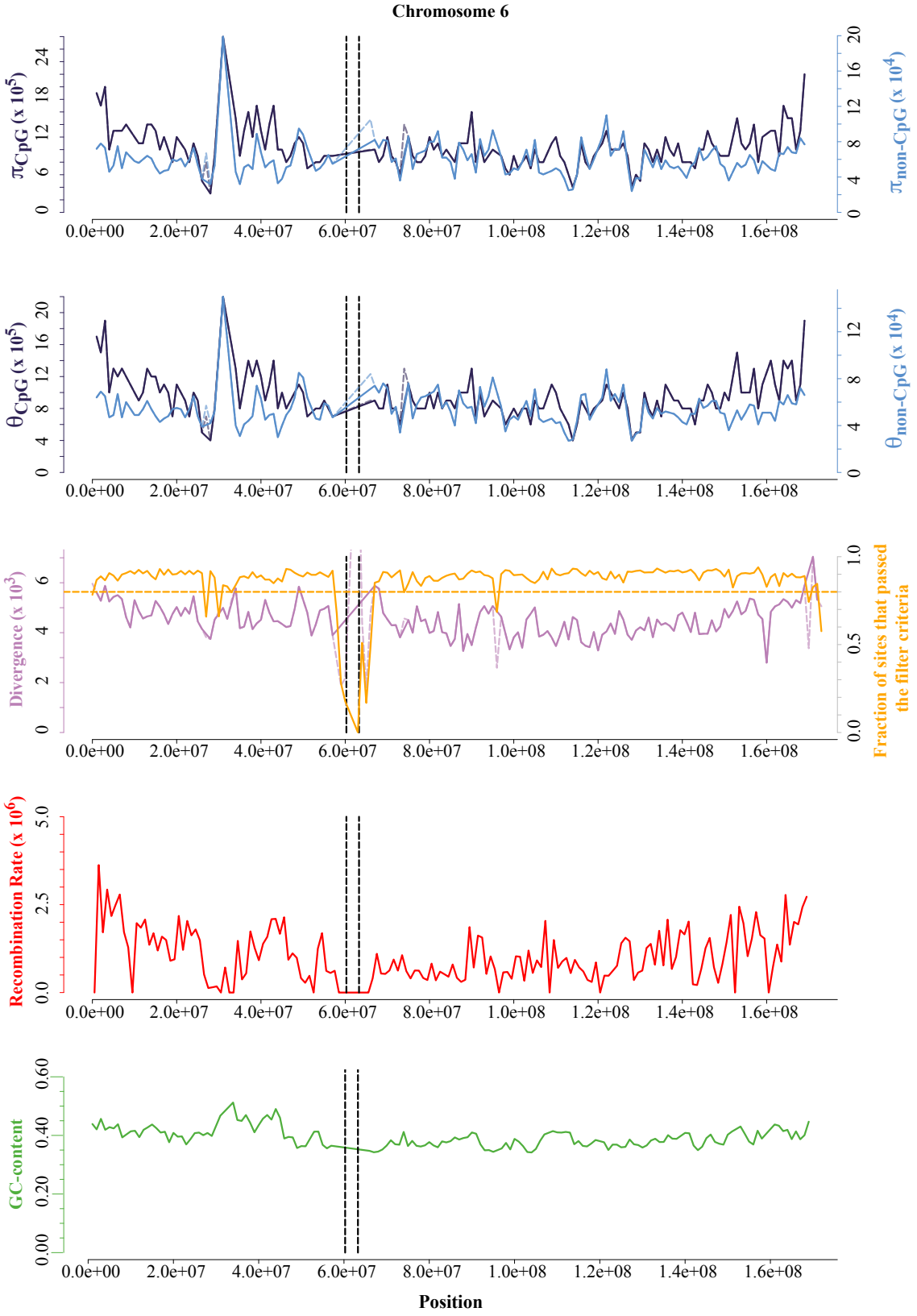


Figure 5.4: Distribution of nucleotide diversity in Western chimpanzees across chromosome 6. Calculated using 1Mb windows which had at least 80% accessibility after filtering. On the 1Mb-scale, the level of nucleotide diversity varies along a single chromosome as a function of divergence (purple), recombination rate (red), and GC-content (green). Dashed lines indicate an accessibility of less than 80%. Accessibility of sites is illustrated as an orange line (a dashed orange line denotes an accessibility of 80% of all sites in a window). Dashed black lines indicate the centromeric region.

The GC-content was measured using information obtained from the panTro2-reference genome. Human-chimpanzee divergence rates were estimated from the alignment of the PanMap high throughput sequencing reads to the human reference genome (version hg18) using the same criteria as for the chimpanzee-based alignment (see Chapter 4). Recombination rates were calculated as the slope of a regression of genetic and physical distances of markers within a 1Mb-window. The genetic map was published in Auton et al. [Auton et al., 2012]. The physical distances were obtained from the mapping of the markers to the human genome (version hg18). The exon content was estimated as the percentage of sequence within exons using the annotation of the chimpanzee genome build panTro2 (extracted from the 'Ensembl Genes' data set of the UCSC Genome table browser [Karolchik et al., 2004]). Similarly, the CpG-island content and the simple repeat content were estimated as the percentage of sequence within CpG-islands and simple repeats, respectively, as obtained from the 'CpG Island' and 'Simple Repeats' track, respectively, provided by the UCSC Genome Table Browser for the chimpanzee genome build panTro2. In addition, the distance from the middle of a given window to the closest centromere, which typically consists of large arrays of repetitive DNA, as well as to the closest telomere, which often contains a high GC-content as well as high rates of recombination, in chimpanzees was measured to obtain information about the influence of large-scale chromosomal structures on nucleotide diversity rates. Locations of centromeres and telomeres were obtained from the UCSC 'Gap' data base table. The length of each chromosome was determined using the information for the panTro2.1 reference genome provided by the UCSC Genome Table Browser.

Estimates for diversity levels were log-transformed to be roughly normally distributed. All parameters were scaled to have variance 1 and mean 0 in order to enable an easier comparison of the different parameters (after the transformation, the slopes directly measure the strength of the relationship between the explanatory and the response variable). All factors were jointly analysed using multiple linear regression. Gnu R's 'leaps' algorithm was applied to perform an exhaustive search for the best subsets of the variables for predicting the nucleotide diversity in linear regression. Standard regression diagnostics were used to evaluate the validity of the model and the influence of outliers (as illustrated in Figure 5.5). As part of the diagnostics, plots of each predictor against the nucleotide diversity in Western chimpanzees were examined in order to see whether the assumption of linearity was violated (see Figure 5.6). Five outliers were excluded from the models: the 1Mb-large regions at (i) 48-49Mb of chromosome 2a, (ii) 10-11Mb of chromosome 3, (iii) 157-158Mb of chromosome 5, (iv) 92-93Mb of chromosome 7, and (v) 53-54Mb of chromosome 20. (After lifting the outliers over to the human reference genome, hg18, using UCSC's 'liftOver' tool, all five of them were found to correspond to regions of segmental duplications).

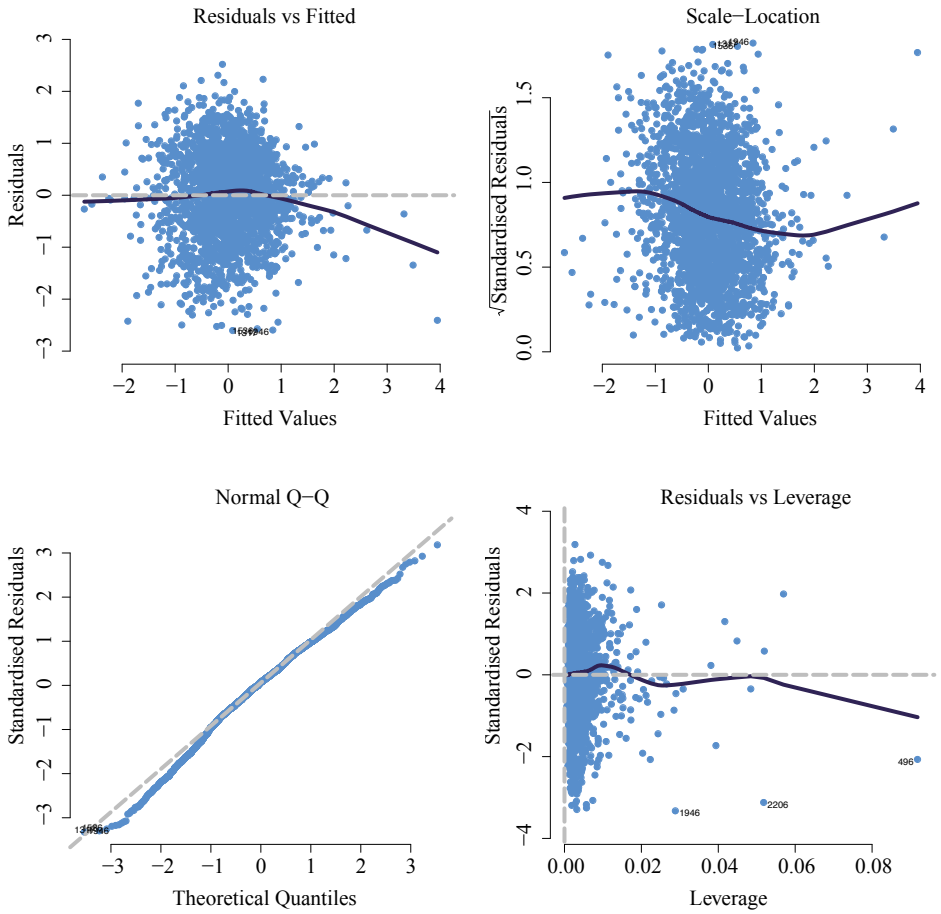


Figure 5.5: Diagnostic plots to check for heteroscedasticity, normality, and influential observations in the multiple linear regression model.

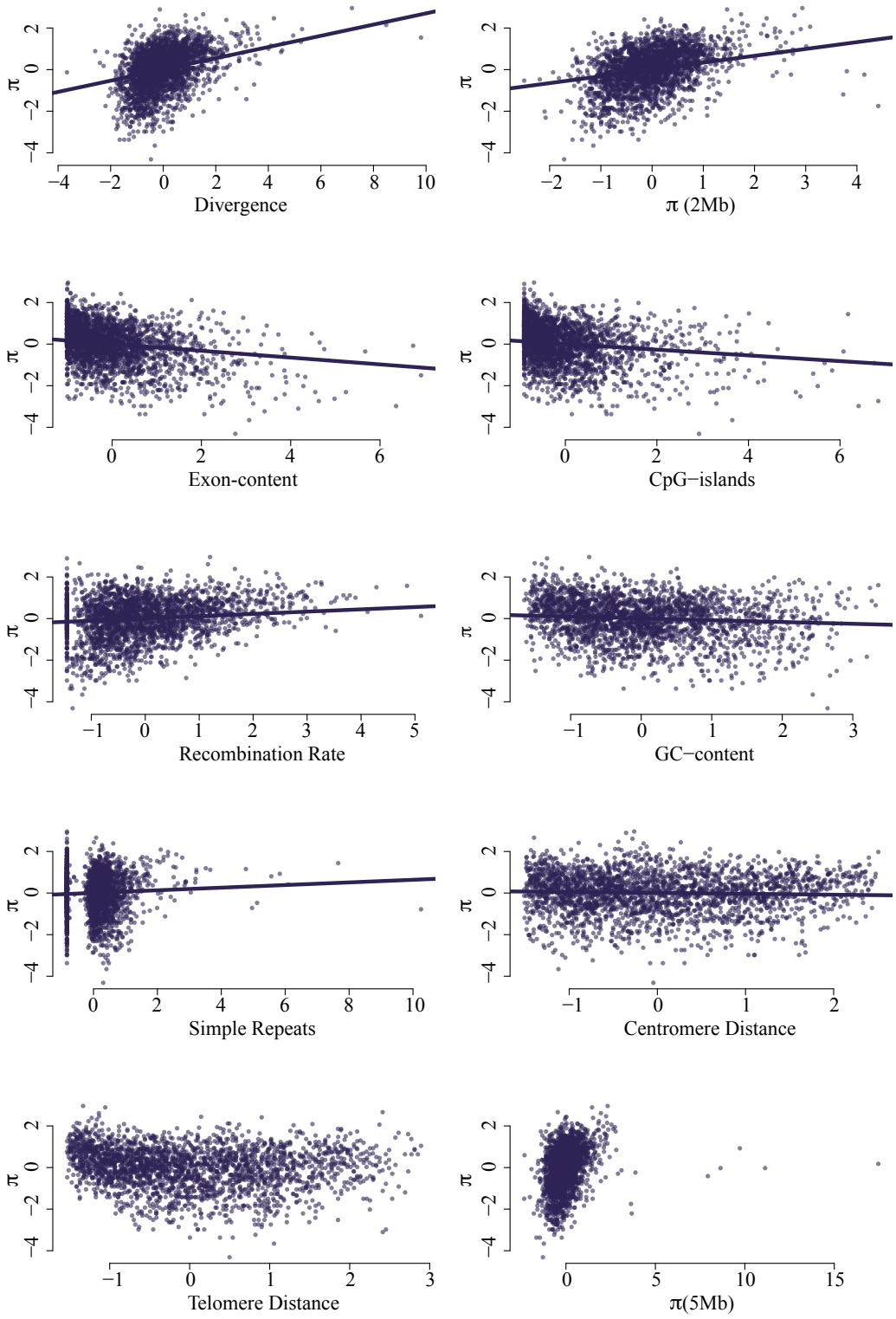


Figure 5.6: Scatterplots of each predictor against the nucleotide diversity estimate π (including regression lines from the full multiple linear regression).

The best model for the overall nucleotide diversity in Western chimpanzees included eight predictors, while best model for the nucleotide diversity at CpG-sites and nonCpG-sites included seven predictors each from the 11 DNA sequence features that might influence mutation rates and that have been considered in the multiple linear regression. These predictors explained 36%, 46%, and 42% of the variance observed in Western chimpanzees, respectively (see Table 5.3(a)). Despite the fact that most predictors were correlated with each other (see Figure 5.7), many of them were included in the best model, suggesting that each of those sequence features explained a unique aspect of the chimpanzee nucleotide diversity. Divergence, recombination rates, GC-content, exon content, CpG-island content, as well as patterns of large 2Mb-scale diversity were significant predictors of chimpanzee diversity at both CpG- and non-CpG-sites. However, their relative importance to the models were quite different: while nucleotide diversity patterns at non-CpG-sites were strongly influenced by divergence rates, diversity rates at the 2Mb-level as well as exon content (see Figure 5.8(a)), the single best predictor at CpG-sites was GC-content (R^2 = 0.282), followed by recombination rates and divergence rates (see Figure 5.8(b)). Although recombination rates and divergence were strongly correlated with each other (see Figure 5.7), recombination rates predicted nucleotide diversity in chimpanzee beyond what was expected from its association with divergence. Similarly, GC-content, exon content and CpG-island content, all three of which were highly correlated with each other, were each included as predictors in the best models. In contrast to the six predictors the models had in common, the distance to the centromere only significantly improved the model of diversity at non-CpG-sites, while the simple repeat content was only a significant predictor of nucleotide diversity at CpG-dinucleotides. Neither the distance to the nearest telomere, the chromosome size, nor the diversity patterns at the 5Mb-level explained any additional part of the variance in nucleotide diversity.

As both nucleotide diversity and divergence, which was the single best predictor for diversity at non-CpG-sites and the third best predictor at CpG-sites, are products of the underlying mutation rate, most of the observed variance in nucleotide diversity arising from mutation rates might be explained by different patterns of divergence. For this reason, a second linear regression was carried out, only using the other 10 DNA sequence features that might influence regional levels of nucleotide diversity to separately investigate their effects on the three models. Interestingly, even without the inclusion of divergence rates in the linear regression analysis, the remaining predictors included in the best models for the overall nucleotide diversity in Western chimpanzees, for the nucleotide diversity at CpG-sites as well as at nonCpG-sites could still explain 32%, 46%, and 38% of the observed variance, respectively (see Table 5.3(b)).

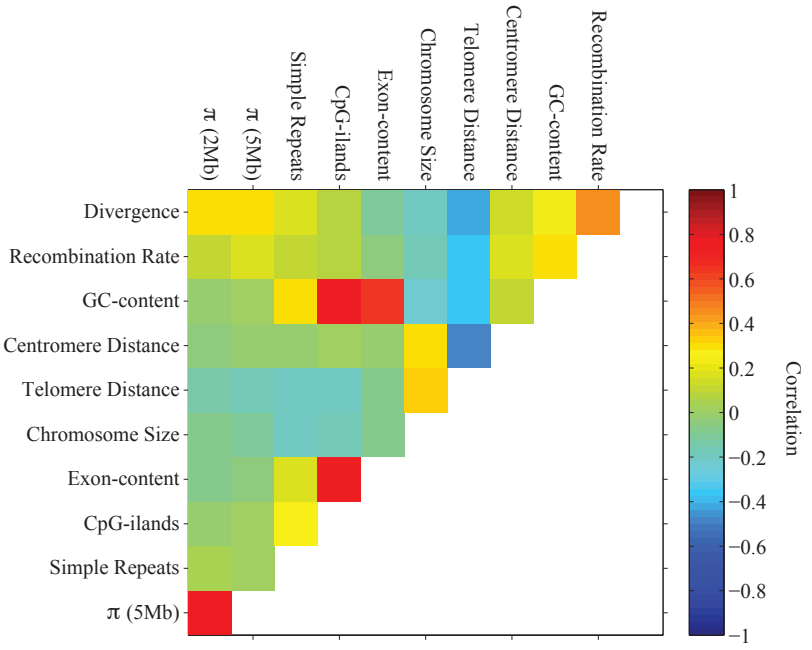


Figure 5.7: Pairwise correlation between different predictors used in the regression model.

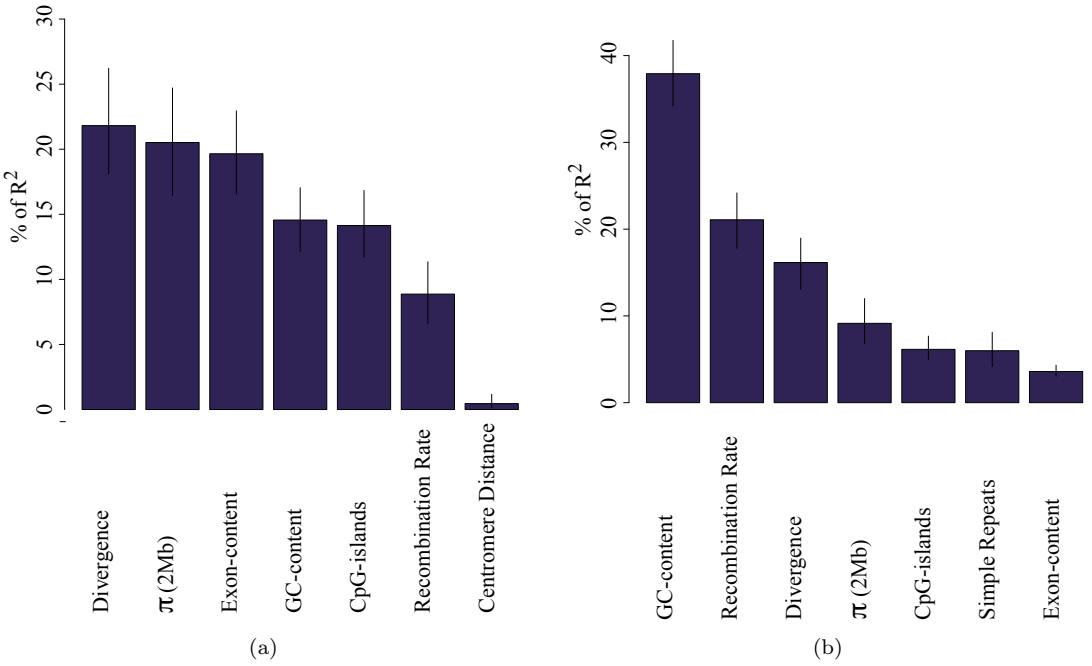


Figure 5.8: Relative importance of each regressor in the multiple linear model for the nucleotide diversity estimate π at (a) non-CpG-sites ($R^2 = 41.96\%$) and (b) CpG-sites ($R^2 = 46.42\%$). The LMG metric [Lindeman et al., 1980] was used to calculate the relative importance of each predictor by partitioned R^2 by averaging over orders (computed using Gnu R's 'rela.impo' package). Metrics were normalised to sum up to 100%.

(a)

	Chimpanzee Diversity					
	$\mathbf{R^2_{all}}$	$\mathbf{Coeff_{all}}$	$\mathbf{R^2_{CpG}}$	$\mathbf{Coeff_{CpG}}$	$\mathbf{R^2_{non\text{-}CpG}}$	$\mathbf{Coeff_{non\text{-}CpG}}$
Divergence		0.27***		0.13***		0.27***
Recombination rate		0.11***		0.14***		0.17***
GC-content		-0.09**		0.60***		-0.26***
Centromere distance		-0.04**		N.S.		-0.05**
Telomere distance		N.S.		N.S.		N.S.
Chromosome size		N.S		N.S.		N.S.
Exon content		-0.16***		-0.14***		-0.15***
CpG-Islands		-0.14***		-0.10***		-0.11***
Simple repeats		0.06**		0.05**		N.S.
π **(5Mb)**		N.S.		N.S.		N.S.
π **(2Mb)**		0.33***		0.21***		0.23***
Full model	0.364		0.464		0.420	

(b)

	Chimpanzee Diversity					
	$\mathbf{R^2_{all}}$	$\mathbf{Coeff_{all}}$	$\mathbf{R^2_{CpG}}$	$\mathbf{Coeff_{CpG}}$	$\mathbf{R^2_{non\text{-}CpG}}$	$\mathbf{Coeff_{non\text{-}CpG}}$
Recombination rate		0.17***		0.18***		0.23***
GC-content		-0.06*		0.62***		-0.25***
Centromere distance		-0.07***		N.S.		-0.06**
Telomere distance		-0.09***		N.S.		-0.06**
Chromosome size		N.S		N.S.		N.S.
Exon content		-0.23***		-0.17***		-0.20***
CpG-Islands		-0.11***		-0.09***		-0.09**
Simple repeats		0.10***		0.07***		0.07***
π **(5Mb)**		N.S.		N.S.		N.S.
π **(2Mb)**		0.44***		0.27***		0.33***
Full model	0.317		0.455		0.384	

Table 5.3: Linear regression models for nucleotide diversity in Western chimpanzees (based on genome-wide data): (a) including divergence rates and (b) excluding divergence rates. The R^2 values represent estimates for the full model (including all significant predictors). The estimates for the coefficients (Coeff) were standardised to simplify the comparison. Significance codes: 0 *** 0.001 ** 0.05 * N.S. = not significant.

5.4 Discussion

Nucleotide diversity levels in Western chimpanzees are not constant across the genome, a variation that might be driven by either evolutionary forces (such as genetic drift or natural selection) or by regional variation in mutation rates. The latter are thought to be controlled by several sequence features, 11 of which were examined to study their relationship with nucleotide diversity in Western chimpanzees using multiple linear regression. Thereby, CpG- and non-CpG-sites were studied separately as their mutation rates as well as the mechanisms responsible for the majority of mutations at these sites are known to be quite different (at CpG-dinucleotides, most mutations are caused by spontaneous methylation-dependent deamination [Chimpanzee Sequencing and Analysis Consortium, 2005; Cooper and Youssoufian, 1988; Coulondre et al., 1978; Duret, 2009; Ehrlich and Wang, 1981; Holliday and Grigg, 1993; Hwang and Green, 2004; Xing et al., 2004] while errors occurring during DNA replication are the main source of mutations at non-CpG-sites [Kim et al., 2006]), both of which might influence nucleotide diversity levels at these sites. Although six factors were significant predictors of chimpanzee diversity at both CpG- and non-CpG-sites (namely divergence, recombination rates, GC-content, exon content, CpG-island content, as well as patterns of large 2Mb-scale diversity), their relative importance to the two models varied: while the model of nucleotide diversity at non-CpG-sites was mainly based on three different sequence features of similar importance, divergence, patterns of large 2Mb-scale diversity, as well as exon content, by far the single best predictor of nucleotide diversity at CpG-sites was regional GC-content ($R^2 = 0.282$). The second best predictor at CpG-sites was recombination rate, which was only the second least influential predictor at non-CpG-sites. This finding is in contrast to previously published work in humans where recombination rates were found to be the single best predictor of nucleotide diversity [Hellmann et al., 2005].

The results of this work clearly demonstrate the significance of studying putative CpG-mutations separately from mutations at non-CpG-sites due to the fact that the relative importance of the above mentioned factors to each mutation-generating mechanism varies between the different sites. In total, ∼46% and ∼42% of the nucleotide diversity in Western chimpanzees observed at the 1Mb-level at CpG- and non-CpG-sites, respectively, can be explained by the studied sequence features, suggesting that other highly complex patterns and processes might also influence regional large-scale variation in mutation rates (such as chromatin structure or replication time). However, it should be kept in mind that both the response as well as the explanatory variables most likely exhibit certain measurement errors which can bias the coefficient estimates resulting from the multiple linear regression and that coefficient estimates might be inflated due to the fact that

most of the regressors are strongly correlated with each other. In addition, the possibility that some of the observed variance might be due to selective constraints on functionally important sites can not be excluded as nucleotide diversity levels were measured on a genome-wide scale rather than for neutrally evolving sites only. Thus, further work is required to assess whether any parts of the observed nucleotide diversity in chimpanzees could be explained by selective pressures (e.g. by examining different selective models such as models of positive selection, negative selection, or a combination of both).

Preface

Work from Chapter 6 has been published as:

> Leffler*, E. M., Gao*, Z., Pfeifer*, S. P., Ségurel*, L., Auton, A., Venn, O., Bowden, R., Bontrop, R., Wall, J. D., Sella, G., Donnelly, P., McVean†, G., and Przeworski†, M., *Multiple Instances of Ancient Balancing Selection Shared between Humans and Chimpanzees*, Science 2013, Vol. 339, No. 6127, pp. 1578-1582.
>
> * These authors contributed equally to the project.
>
> † These authors jointly supervised the project.

Variants have been annotated with the ancestral allele by Adam Auton using the ancestral alignments provided by Javier Herrero of the EBI. Simulations of recurrent mutations in humans and chimpanzees were carried out by Ziyue Gao (see Section 6.4). I conducted all other analyses in this chapter with input from Adam Auton and Gil McVean.

Chapter 6

PanMap - Analysis of Genome-wide Sequence Variation Shared between Humans and Western Chimpanzees

The genome of each human individual consists of about 3.2 billion nucleotides, the majority of which are identical by descent as they have been passed down generation by generation from our ancestors. Nevertheless, individual genomes contain slight variations caused by distinct mutational events in the ancestral germ lines. These germ line mutations, which accumulate in characteristic ways influenced by events during the history of the population (such as bottlenecks, splits, migrations, expansions, or contractions), play a key role in shaping genome evolution as they are the primary source of genetic diversity within populations as well as the divergence between species, both of which natural selection can act on.

Despite the fact that humans and their evolutionary closest living relatives, chimpanzees, largely exhibit a distinct evolutionary history across their genomes, genetic polymorphisms that have already been present in their common ancestor can be detected. These shared variants (also referred to as 'ancestral' or 'trans-species' polymorphisms) are the result of mutations that occurred before the speciation of the two hominoid lineages ∼5-7 Mya [Brunet et al., 2002; Chen and Li, 2001; Varki and Altheide, 2005] and that survived within both species throughout their divergence (see Figure 6.1). However, due to the evolutionary distance between humans and chimpanzees, the occurrence of ancestral polymorphisms is extremely rare. Coalescence theory predicts that the probability of observing a shared polymorphism by chance for selectively neutral alleles in a randomly mating population is very small and even polymorphisms under moderate selection in one of the two lineages will eventually be eliminated by random genetic drift due to the long time separating the two species [Golding, 1992]. Thus, strong balancing selection is required to maintain variants for these extraordinary lengths of time. Several scenarios of balancing selection which work

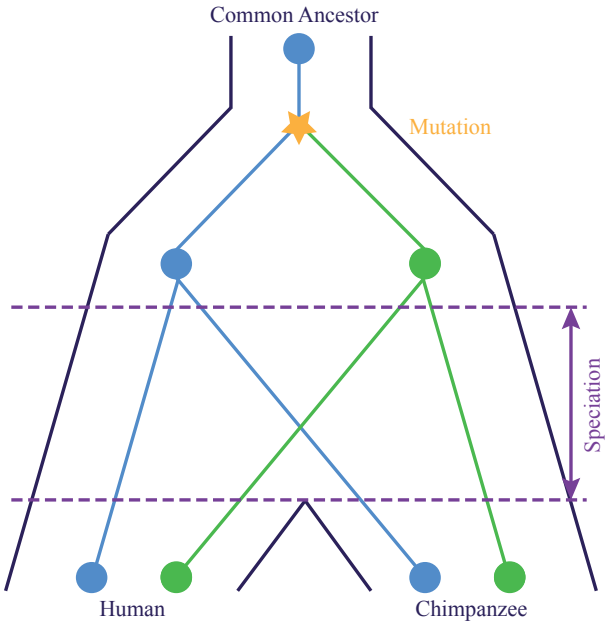


Figure 6.1: Concept of ancestral polymorphisms. Two alleles at a single locus (depicted in blue and green) have been passed on through speciation to human and chimpanzee individuals from their most recent common ancestor.

over the required time-scales have been proposed to explain the maintenance of ancestral polymorphisms. The two main (and potentially non-exclusive) models are (i) over-dominant selection, also known as 'heterozygous advantage' as heterozygous individuals experience a fitness advantage over both homozygous genotypes due to co-dominant gene expression [Doherty and Zinkernagel, 1975; Hughes and Nei, 1988; Takahata and Nei, 1990], and (ii) negative frequency-dependent selection in which a rare allele confers a superior fitness in a population (thus also referred to as 'rare' or 'minority allele advantage') [Clarke and Kirby, 1966; Takahata and Nei, 1990].

Allelic variants that have been inherited before the time of speciation and that have survived in both descendent lineages have been detected in primates. The best known examples of ancestral polymorphisms occur in genes whose function suggest a mechanism whereby strong natural selection acts to maintain a highly diverse set of alleles, such as genes in the immune system gene cluster (also known as MHC) class I and class II antigenes involved in the adaptive immune response [Clark, 1997; Fan et al., 1989; Gyllensten and Erlich, 1989; Hughes and Nei, 1988, 1989; Klein, 1987; Lawlor et al., 1988; Mayer et al., 1988]. Although strong evolutionary forces (such as expansions due to insertions of either HERV or Alu sequences, gene duplication, contractions caused by gene loss, gene conversions, and selection for diversity [Bergstrom et al., 1999; Gagneux and Varki, 2001; Watkins, 1995]) as well as high levels of recombination caused great genetic diversification in this region, it has been shown in several studies that not only have many of these genes been conserved between all hominoids [McAdam et al., 1994; The MHC sequencing consortium, 1999],

but also that extensive sharing of polymorphisms in the antigen-binding regions already exists for millions of years due to long-term balancing selection [Ayala et al., 1994; Bjorkman et al., 1987; Brown et al., 1988; Hedrick, 1998; Hughes and Nei, 1988, 1989; Klein, 1987] (see Section 6.6.1). Few ancestral polymorphisms between humans and chimpanzees have been detected outside the MHC region (e.g. trans-species polymorphisms in the *TRIM5* gene [Cagliani et al., 2010]).

A comparison of the extensive genome-wide sequence variation data for Western chimpanzees, now available through the PanMap Project (see Chapter 4), with intra-specific genetic diversity data of humans might allow the identification of further ancestral polymorphisms shared between the two lineages.

6.1 Aim of the Study

As part of the PanMap Project aimed to document genome-wide sequence variation in our closest evolutionary relative, over four million high confidence autosomal variants and over 100,000 high confidence variants on chromosome X have been identified in regions of orthology to humans by sequencing 10 Western chimpanzees (nine females and one male) at an average coverage of 9.45X (as described in Chapter 4). The availability of high quality polymorphism data for chimpanzees enables the identification of polymorphic sites which are also variable among humans (in particular this study focused on the 59 individuals from the Yoruba, YRI, population from Ibadan, Nigeria, sequenced as part of the 1000 Genomes Project [The 1000 Genomes Project Consortium, 2010]), leading to a map of genome-wide shared polymorphism. This map will allow the systematic search for ancestral polymorphisms shared between the two lineages on a genome-wide scale.

The majority of the observed shared variants will most likely be the result of recurrent mutations at the same locus after speciation, a scenario which is especially important for highly mutable sites of the genome (e.g. CpG-dinucleotides). Other polymorphisms which are shared between the two lineages might be caused by systematic errors in the variant calling process, such as shared mismapping artefacts due to problems with the read alignment (e.g. in repetitive regions of the genome), repeated sequencing errors, ascertainment bias, or shared false positive SNP calls (e.g. due to an incorrect genome assembly of duplications which occurred before the divergence of the species onto the same locus). Although balancing selection which is active over long periods of time is predicted to be extremely rare [Hey, 1999; Przeworski et al., 2000], a subset of the shared variants might be due to the maintenance of polymorphisms that occurred in the common ancestral species due to long-term balancing selection. Their frequency will depend on several factors such as the effective population size and the genetic diversity of the last common ancestor as well as demographic patterns [Hacia et al., 1999].

Therefore, apart from the generation of a map of genome-wide polymorphisms shared between humans and chimpanzees, another goal of this study is to develop methods to distinguish polymorphisms which are shared through ancestry from those that have been introduced either through convergent evolution of unrelated alleles after speciation or due to systematic artefacts. As the occurrence of ancestral polymorphisms by chance is extremely rare, ancestral polymorphisms are good candidates for genes and regions under strong balancing selection and as a result, they are a useful tool to study speciation.

6.2 Data

An overview of the data used in this study is given in Table 6.1.

6.2.1 Polymorphism Data Sets

Human: 10,903,690 polymorphic sites (10,556,156 on the autosomes with a Ts/Tv ratio of 2.06, 345,181 in the non-PAR of chromosome X with a Ts/Tv of 1.90, and 2,353 in the PAR with a Ts/Tv ratio of 1.87) discovered in low-coverage (average 3.4X) sequencing of 59 unrelated individuals (36 females and 23 males) from the Yoruba population as part of the Pilot phase of the 1000 Genomes Project [The 1000 Genomes Project Consortium, 2010] were used. Data from the Yoruba population has been chosen for this analysis as Africans carry considerably more genetic variation than the non-African populations from which samples were obtained in the 1000 Genomes Project. The polymorphism data set was downloaded from the project's FTP site [22].

SNPs were defined as 'shared' between human and chimpanzee if orthologous sites in the human and chimpanzee genome were variable for the exact same nucleotides in both polymorphism data sets. To this end, orthologous positions in the chimpanzee genome needed to be identified for each human SNP (and vice versa). For this purpose, UCSC's 'liftOver' tool was used to lift the SNPs from the human reference genome (hg18) over to the chimpanzee reference genome (panTro2). To ensure that the selected SNPs were part of well-characterised regions in the genome, at least 95% of the 100bp flanking neighbourhood of a SNP (in both directions) was also required to lift over. The lift over of each human SNP as well as its 100bp flanking neighbourhood to the chimpanzee reference assembly resulted in the exclusion of 969,115 SNPs from further analysis.

[22] `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot_data/paper_data_sets/a_map_of_human_variation/low_coverage/snps/`

In the downloaded files [23], the ancestral alleles of the remaining 9,669,391 autosomal SNPs (Ts/Tv: 2.08) were already annotated by the 1000 Genomes Project using the 4-way primate EPO alignments [24]. For these alignments, the human, chimpanzee, orang-utan, and rhesus macaque genomes have been aligned using a combination of Enredo, Pecan, and Ortheus [Paten et al., 2008a,b]: Enredo was used to built an initial graph, using a set of conserved sequences to identify genomic point anchors in all four genomes, which was then modified through a series of graph transformations and simplifications in a way such that the graph's edges represent sets of co-linear segments. These segments were aligned with Pecan which uses either a standard species tree (for segments not representing duplications) or a sequence tree (for any other case) to perform an alignment. The ancestral states of the alleles were derived based on these alignments using Ortheus, a software that uses a Tamura-Nei evolutionary model [Tamura and Nei, 1993] to handle substitutions, and a Hidden Markov Model to infer indels. All calls have been classified using the ancestral, the sister and the ancestral of the ancestral sequences. An ancestral allele could be assigned to 9,484,832 autosomal SNPs (Ts/Tv: 2.08) while 184,559 SNPs could not been assigned an ancestral allele due to either

- a lineage-specific insertion (i.e. 'AA= −', 22,570 SNPs),
- a disagreement of the ancestral allele classification based on information obtained from the ancestral, the sister and the ancestral of the ancestral sequences (i.e. 'AA=N', 60,988 SNPs), or
- because there was no data available (i.e. 'AA=.', 101,001 SNPs).

SNPs that showed a lineage-specific insertion or a disagreement of the ancestral allele classification were excluded. The 265,184 SNPs on chromosome X were all kept in the data set despite them not having an ancestral allele annotation.

A further 55,726 autosomal SNPs were excluded as they were the result of a double mutation (they experienced mutations to more than one allele).

The final data set contained 9,429,106 autosomal SNPs with a Ts/Tv ratio of 2.09, 263,405 SNPs in the non-PAR of chromosome X with a Ts/Tv ratio of 1.94, and 1,779 SNPs in the PAR with a Ts/Tv ratio of 1.93. 14.16% of the autosomal SNPs, 13.04% of the non-PAR chromosome X SNPs, and 19.62% of the PAR SNPs occurred at CpG-sites (a CpG-site was defined as one where the reference allele creates a CpG-site).

[23] `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot_data/technical/reference/ancestral_alignments/`
[24] `ftp://ftp.ensembl.org/pub/release-54/emf/ensembl-compara/epo_4_catarrhini/`

Chimpanzee: 4,005,229 high confidence autosomal SNPs (Ts/Tv: 2.25) and 1,477 SNPs (Ts/Tv: 1.95) in the PAR variable among the 10 individuals sequenced in the PanMap Project were used in this study. Furthermore, 107,482 high confidence SNPs (Ts/Tv: 2.11) from the non-PAR of chromosome X were used. They were called separately, in the nine female chimpanzee individuals only, due to the hemizygosity of the one male in the PanMap study (see details about the intersection polymorphism data set in Section 4.3.3).

Following a similar procedure than the one described for the human polymorphism data set, a lift over of each chimpanzee SNP from the chimpanzee reference genome (panTro2) to the human reference genome (hg18) resulted in the exclusion of 146,761 autosomal SNPs, 5,584 SNPs in the non-PAR of chromosome X, and 165 SNPs in the PAR from further studies.

The ancestral alleles of the remaining 3,858,468 autosomal SNPs (Ts/Tv: 2.24), 101,898 SNPs in the non-PAR of chromosome X (Ts/Tv: 2.12), and 1,312 SNPs in the PAR (Ts/Tv: 1.98) were subsequently annotated using the 6-way primate EPO alignments [25]. An ancestral allele could be assigned to 3,836,570 out of the 3,858,468 autosomal SNPs, 93,841 out of the 101,898 SNPs in the non-PAR of chromosome X, and 380 out of the 1,312 SNPs in the PAR. A total of 23,200 SNPs could not been assigned an ancestral allele as those SNPs either

- showed a lineage-specific insertion (9,011 autosomal SNPs, 606 SNPs in the non-PAR of chromosome X, and 50 SNPs in the PAR),
- showed a disagreement of the ancestral allele classification (12,887 autosomal SNPs, 643 SNPs in the non-PAR of chromosome X, and 3 SNPs in the PAR), or
- there was no data available (6,808 SNPs in the non-PAR of chromosome X, and 879 SNPs in the PAR).

SNPs that showed a lineage-specific insertion or a disagreement of the ancestral allele classification were excluded, resulting in a data set of 3,836,570 autosomal SNPs (Ts/Tv: 2.24), 100,649 SNPs in the non-PAR of chromosome X (Ts/Tv: 2.21), and 1,259 SNPs in the PAR (Ts/Tv: 1.95) for further studies.

In a last step, 10,349 SNPs that were the result of a double mutation were excluded.

The final data set contained 3,826,415 autosomal SNPs with a Ts/Tv ratio of 2.25, 100,459 SNPs in the non-PAR of chromosome X with a Ts/Tv ratio of 2.12, and 1,255 SNPs in the PAR with a Ts/Tv ratio of 1.96. 15.24% of the autosomal SNPs, 11.94% of the non-PAR chromosome X SNPs, and 6.37% of the PAR SNPs occurred at CpG-sites.

[25] `ftp://ftp.ensembl.org/pub/release-59/emf/ensembl-compara/epo_6_primate/`

Chromosome		Statistics			
		# SNPs	Ts/Tv	Fraction in Repeats	Fraction in CpGs
Autosomes	Chimpanzee	3,826,415	2.25	46.31%	15.24%
	Human	9,429,106	2.09	42.17%	14.16%
	Coincident	43,037	4.96	43.18%	52.74%
	Shared	34,140	12.56	43.55%	61.61%
Chromosome X non-PAR	Chimpanzee	100,459	2.12	54.59%	11.94%
	Human	263,405	1.94	50.72%	13.04%
	Coincident	690	4.19	56.38%	46.23%
	Shared	528	9.76	57.95%	55.11%
Chromosome X PAR	Chimpanzee	1,255	1.96	55.46%	6.37%
	Human	1,779	1.93	43.62%	19.62%
	Coincident	8	undef.	62.50%	37.50%
	Shared	8	undef.	62.50%	37.50%

Table 6.1: Characteristics of the polymorphism data sets, coincident and shared polymorphisms.

6.2.2 Coincident Polymorphisms

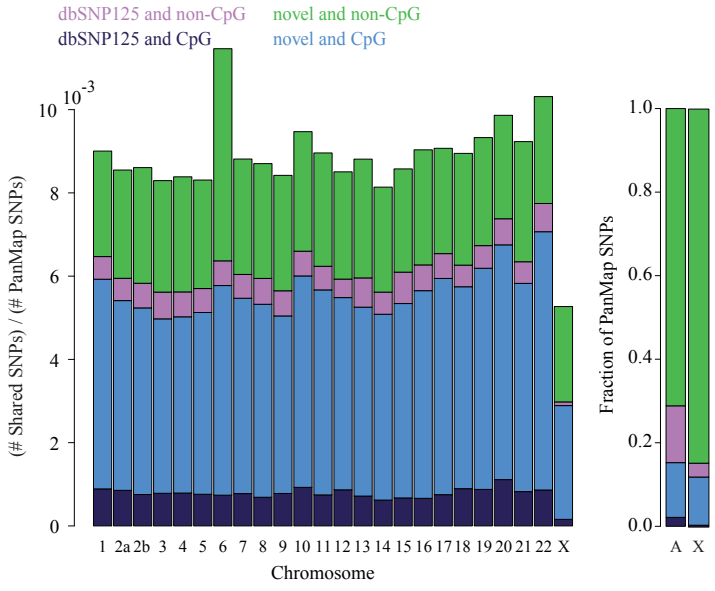
43,735 SNPs were identified that occurred at orthologous sites in the genomes of human and chimpanzee (referred to as 'coincident polymorphisms'): 43,037 SNPs on the autosomes (Ts/Tv: 4.96), 690 SNPs in the non-PAR of chromosome X (Ts/Tv: 4.19), and eight SNPs in the PAR (all eight SNPs encoded transitions). 52.74% of the autosomal SNPs, 46.23% of the non-PAR chromosome X SNPs, and 37.5% of the PAR SNPs occurred at CpG-sites.

6.2.3 Shared Polymorphisms

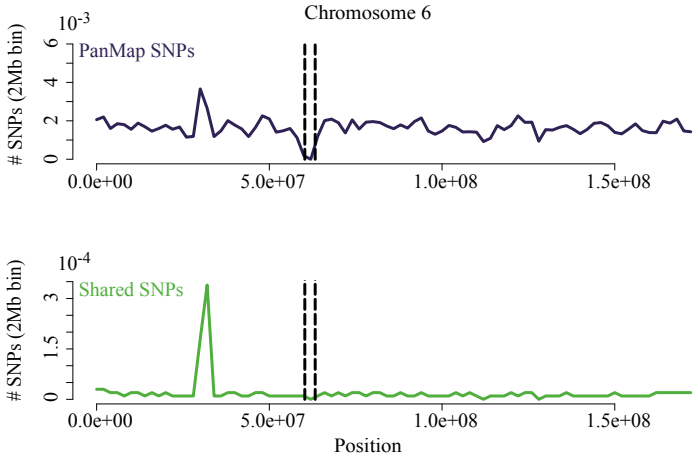
Out of the 43,735 coincident SNPs, 34,676 were variable for the exact same two nucleotides in the polymorphism data sets of both species (referred to as 'shared polymorphisms'): 34,140 SNPs on the autosomes (Ts/Tv: 12.56), 528 SNPs in the non-PAR of chromosome X (Ts/Tv: 9.76), and eight SNPs in the PAR (all eight SNPs encoded transitions). The high Ts/Tv ratios observed for the shared polymorphisms resulted from the fact that the majority of these SNPs occurred at CpG-sites of the genome (∼62% on the autosomes and ∼55% on chromosome X).

6.2.3.1 Distribution of Shared Polymorphisms

The fraction of shared variants across all autosomes is depicted in Figure 6.2(a). The data set was divided into four categories: variants that have already been previously reported in the public catalogue of variant sites dbSNP (release 125) [Sherry et al., 2001] and variants that were novel to the PanMap Project, as well as variants which were putative CpG-mutations and variants at non-CpG-sites. Notably, the fraction of shared variants was not constant across chromosomes. Chromosome 6 exhibited a higher degree of sharing between humans and chimpanzees which was largely caused by a high degree of sharing in the MHC region (see Figure 6.2(b)). Out of the 249,227



(a)



(b)

Figure 6.2: (a) Fraction of shared variants between chromosomes. The data set was divided into SNPs that have already been previously reported in the public catalogue of variant sites dbSNP (release 125) [Sherry et al., 2001] and novel SNPs as well as SNPs which were putative CpG-mutations and SNPs at non-CpG-sites. For comparison, the fraction of SNPs in the entire PanMap set for each of these four categories is given on the right-hand side for autosomes (A) and chromosome X (X). (b) SNP density along chromosome 6 (top) in the PanMap polymorphism data set and in the shared polymorphism data set (bottom). SNP density was measured in 2Mb-sized bins. Dashed black lines indicate the centromeric region.

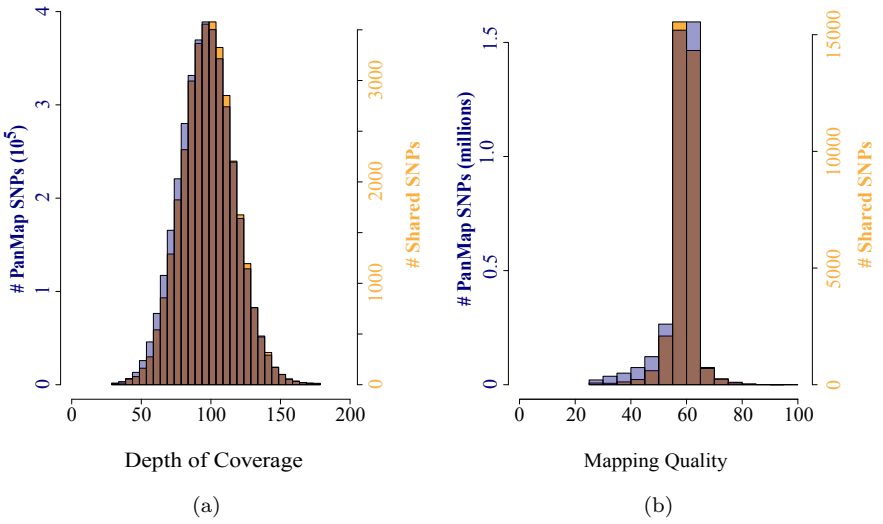


Figure 6.3: (a) Read coverage and (b) mapping quality for all PanMap SNPs (blue) and shared variants (orange).

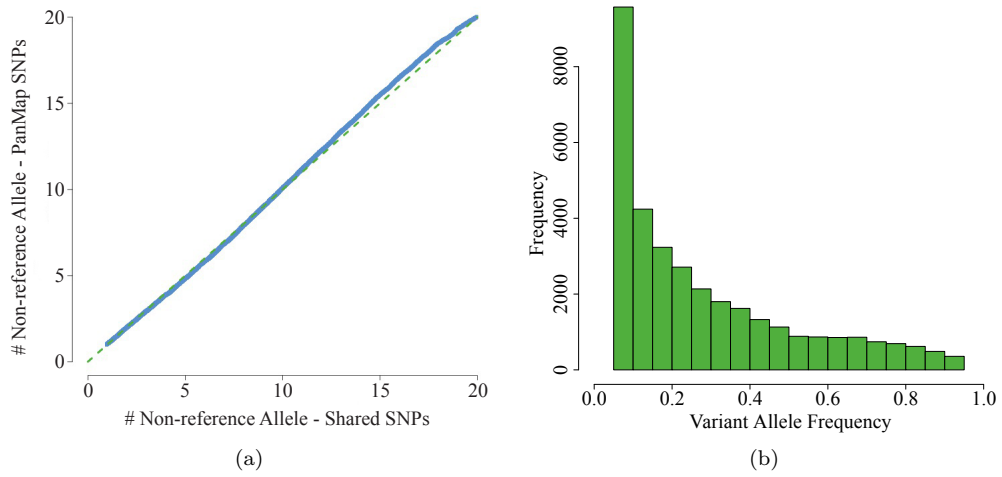
PanMap SNPs detected on chromosome 6, 1.19% occurred in the regions of MHC genes (43 SNPs in the MHC class I and 2,928 SNPs in the MHC class II as defined by the 'Known Genes' data set [Hsu et al., 2006] of the UCSC table browser [Karolchik et al., 2004]; see Table G.1). Of these variants identified in Western chimpanzees, 16.8% were identified as shared with humans: one SNP in the MHC class I gene *HLA-B* and 497 SNPs in MHC class II genes. The 498 shared variants in the MHC region make up a total of 17.4% of all 2,857 variants shared on chromosome 6. Although the genes of the MHC are some of the best-known examples for long-term balancing selection [Ayala et al., 1994; Bjorkman et al., 1987; Brown et al., 1988; Hedrick, 1998; Hughes and Nei, 1988, 1989; Klein, 1987], it should be noted that some of the shared polymorphisms observed in this region are likely due to recurrent mutation rather than shared ancestry. This is due to the fact that there will be more opportunity for mutations in the history of the sample where balancing selection leads to a deep coalescence tree and thus, there will be a higher probability of sharing due to recurrent mutation.

6.2.3.2 Characteristics of Shared Polymorphisms

The shared polymorphisms showed a good coverage (mean autosomal coverage of 100.28 with SD = 19.24; mean coverage on chromosome X of 87.36 with SD = 17.51, see Figure 6.3(a)) of reads with high mapping qualities (mapping quality score of 59.21 on average on the autosomes with SD = 4.24; an average mapping quality score of 59.63 on chromosome X with SD = 4.37; see Figure 6.3(b)) in the studied individuals, comparable with those obtained from all 3,928,129 PanMap SNPs.

The pattern of these shared variants was inconsistent with substantial problems and systematic errors in the SNP calling process as false positive calls are more likely to exhibit an unusual behaviour according to excessive depth of coverage (as a read depth much lower or much higher than expected is suggestive of copy number variation which might lead to false positive SNP calls in the mapping of paralogous sequences) and they appear more often in regions of poor read alignment indicated by low mapping quality scores. Furthermore, the shared SNPs are similar to non-shared SNPs in terms of their proportion in repeats (though with a slightly higher proportion in repeats on chromosome X): while the fraction of the 34,140 autosomal shared SNPs in repeats (44%), as well as the faction of the 528 shared chromosome X non-PAR SNPs in repeats (58%) was in a very similar range as the original data set for chimpanzee (46%, and 55%, respectively), the fraction of shared variants in repeats in the PAR was higher compared to the original PanMap data (63%, and 55%, respectively) but this deviation is most likely caused by the very small sample size of observed variants that were shared in this region (see Table 6.1). Artificial SNP calls due to a misassembly or mismapping of reads (e.g. in repetitive regions of the genome or collapsed duplications) also lead to a skewed allelic imbalance, in particular an excess of alleles with higher frequencies can be observed. In contrast, the site frequency spectrum of the identified shared polymorphisms was very similar to that of the entire PanMap SNP set (see Figure 6.4(a)) suggesting that these shared polymorphisms were unlikely to be artefacts of shared mismapping, incorrect assemblies, or other systematic errors. However, the similarity in the allele frequency spectrum at shared SNPs versus all SNPs also indicates that there was no evidence of substantial ancestry-based sharing (a shift towards more common alleles would be expected if a substantial fraction of shared SNPs were identical by descent, but was not observed; see Figure 6.4(b)).

One major difference between shared and non-shared SNPs was their proportion at CpG-sites: 15% of the autosomal SNPs, 12% of the SNPs in the non-PAR of chromosome X, and 6% of the SNPs in the PAR occur at CpG-dinucleotides genome-wide in chimpanzees (using the chimpanzee reference sequence) while 62%, 55%, and 38%, respectively, of the shared SNPs do. CpG-dinucleotides are known to be hot spots for point mutations in mammalian genomes [Chimpanzee Sequencing and Analysis Consortium, 2005; Hwang and Green, 2004; Keightley et al., 2011; Miller and Kwok, 2001; Siepel and Haussler, 2004], an observation strengthening the hypothesis that the majority of the identified shared polymorphisms were due to recurrent mutations of the same locus rather than being shared through ancestry. The reduced rate of allele sharing on chromosome X might then be explained by a combination of a more rapid coalescence in the ancestor of humans and chimpanzees, the lower mutation rate in females where the X spends 2/3 of its time, and a lower GC-content in the non-PAR of chromosome X [Galtier et al., 2001; Li et al., 2002].



(a) (b)

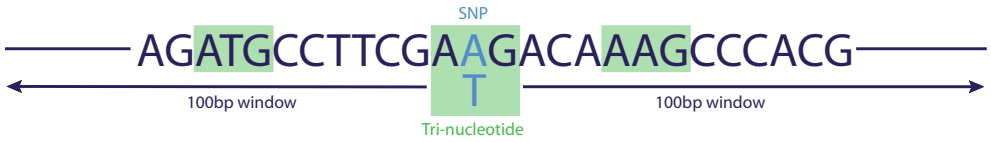
Figure 6.4: (a) QQ-plot of the non-reference allele site frequency spectrum of the shared polymorphisms and the one generated using all PanMap polymorphisms. (b) The number of shared variants at different allele frequencies.

6.3 Sequence-context of Shared Polymorphisms

Mutation rates vary between different regions of the genome, an effect that is strongest at the single nucleotide level where mutation rates are influenced by their sequence-context [Blake et al., 1992; Gojobori et al., 1982; Hwang and Green, 2004; Zhao and Boerwinkle, 2002]. The most well-known and strongest DNA sequence-context effect is the about one order of magnitude higher frequency of C>T transitions at hyper-mutable CpG-dinucleotides which accounts for ∼25% of all SNPs in the human genome [Fryxell and Moon, 2005; Hwang and Green, 2004; Miller and Kwok, 2001; Rideout et al., 1990; Sved and Bird, 1990; Zavolan and Kepler, 2001; Zhao and Boerwinkle, 2002]. The reason for the elevated mutability of CpG-sites is that many cytosine residues are methylated at their 5' position and when deaminated, they exhibit a strong tendency to deaminate to either TpG or CpA both of which are less efficiently or incorrectly repaired by the DNA repair machinery [Cooper and Youssoufian, 1988; Ehrlich and Wang, 1981; Holliday and Grigg, 1993]. While CpGs exhibit the strongest effect on mutation rates (transition mutation rates are elevated by ∼15-fold relative to the average mutation rate in mammals and ∼30-fold in great apes [Hwang and Green, 2004; Keightley et al., 2011; Siepel and Haussler, 2004]), other adjacent nucleotides have also been shown to elevate mutation rates by two- to threefold [Blake et al., 1992; Hwang and Green, 2004; Siepel and Haussler, 2004; Zhao and Boerwinkle, 2002]. Even if previous studies have shown that these effects can extent as many as <10 nucleotides away [Silva and Kondrashov, 2002; Smith et al., 2003], the strongest effects have been detected for short ranges of up to two nucleotides each direction as nucleotides further away seem to have statistically significant but increasingly small effects [Nevarez et al., 2010; Zavolan and Kepler, 2001; Zhao and Boerwinkle, 2002].

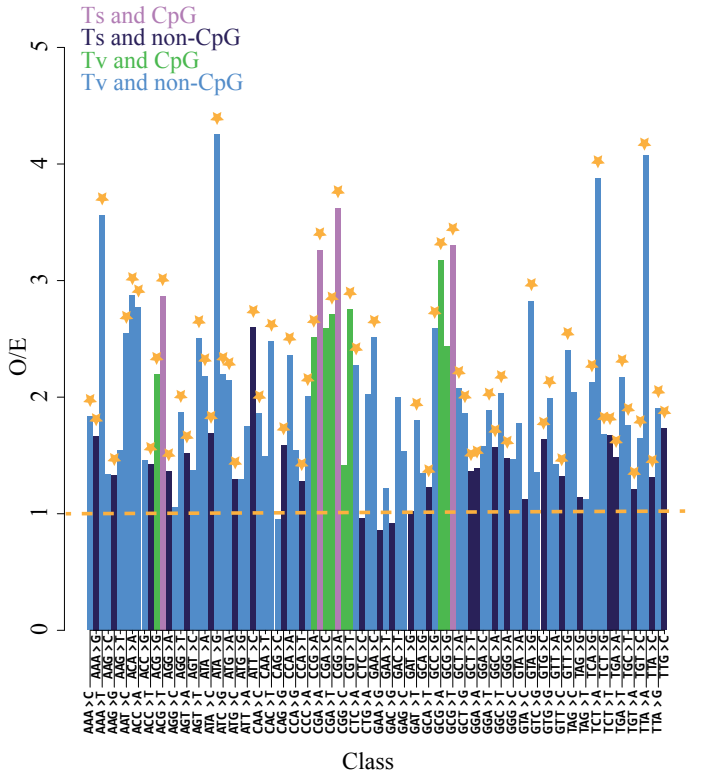
As some regions of the genome are highly mutable, a significant proportion of the shared polymorphisms detected at those sites will most likely be due to recurrent mutations at coincident sites. For this reason, it is important to estimate how many of the detected shared variants can be explained by different sequence-context effects (such as CpG-effects). As directly adjacent sites seem to have the most significant effect on nucleotide substitution rates, a local permutation strategy was developed to control for different trinucleotide sequence composition.

To assess if certain sequence-contexts are more likely to exhibit shared polymorphisms, all possible trinucleotides in the chimpanzee genome that had a polymorphism at their central base were grouped by mutation type (e.g. AAG>T denotes an AAG trinucleotide in which a mutation from the ancestral allele A to the alternate allele T occurred leading to an ATG trinucleotide). For each variant in the PanMap data set, all other locations of both trinucleotides (in this example, AAG and ATG) in the 100bp neighbourhood of the polymorphism were recorded (to partially account for possible variation of mutation rates at a larger scale), together with the information on whether the motife occurrence coincided with a SNP site in humans that had the same alleles.

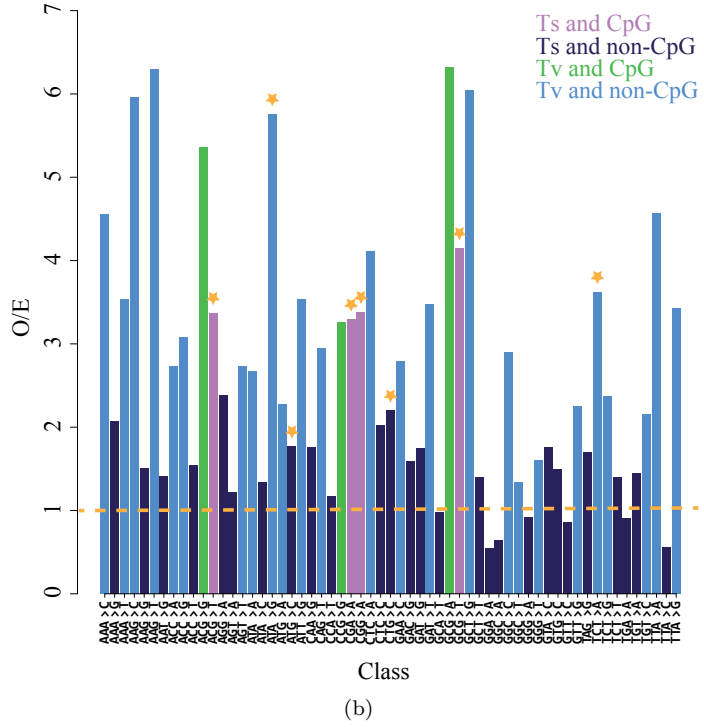


Out of the initial 3,928,129 PanMap SNPs that passed the lift over and ancestral allele requirements described above, 10 were excluded from this analysis as there was a missing reference base (i.e. 'N') at one of the adjacent positions precluding trinucleotide classification. Based on the information of the remaining 3,928,119 variant sites in chimpanzee, 3,826,405 autosomal ones (of which 27,792 SNPs were coding; see Section 6.5) and 101,714 variants on chromosome X, the ratio of the observed rate of sharing between the two species to the expected one was calculated for each class of mutation.

For most trinucleotides and mutation types, there was significant excess sharing for SNPs located on the autosomes (both considering all SNPs and only non-coding SNPs), at both CpG- and non-CpG-sites (see Figures 6.5(a) and H.1). The three mutation classes with the greatest fold-excess sharing were non-CpG transversions (ATA>G, TTA>A, and TCT>A). There was also excess sharing for most classes of mutation on chromosome X as well as for protein-coding SNPs on the autosomes (see Figures 6.5(b) and H.2), but due to the limited amount of data, most of which was not significant. The higher ratios of the observed rate of sharing to the expected one for most mutation classes on chromosome X compared to the autosomes (see Figure 6.6) were most likely caused by a lower nucleotide diversity on chromosome X.



(a)



(b)

Figure 6.5: Ratio of the observed (O) rate of sharing to the expected (E) one per mutation class on (a) the autosomes and (b) chromosome X. To simplify the plots, strands have been collapsed (e.g. ACG>C and CGT>G have been collapsed to one class as CGT is the reverse complement of ACG). Bars have been coloured by whether or not the SNP was a putative CpG-mutation, transition (Ts), or transversion (Tv). * in the plots indicates a p-value<0.01 using a likelihood-ratio test LRT $= -2\,(\mathrm{O}-\mathrm{E}) + 2\,\mathrm{O}\,(\log \mathrm{O} - \log \mathrm{E})$ with the null hypothesis that the ratio is 1 (indicated by an orange line).

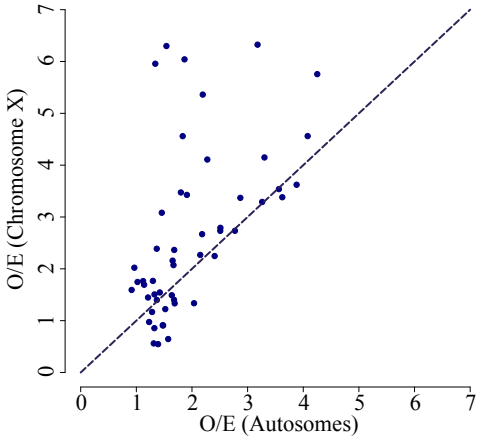


Figure 6.6: Comparison of the estimates for the ratio of the observed (O) rate of sharing to the expected (E) one per mutation class on the autosomes and chromosome X.

In total, 12,667 autosomal variants would have been expected to be shared by chance taking trinucleotide sequence-context effects into account while 34,140 autosomal shared variants were observed (see Section 6.2.3), representing a 2.7-fold excess. Similarly, a 2.9-fold excess of sharing has been observed on chromosome X: 536 SNPs have been detected that are shared between humans and chimpanzees while only 187 SNPs would have been expected by chance under the model. The similar ratio on chromosome X can not be readily explained by ancestral polymorphisms (as less are expected) or the same rate of recurrent mutation than on the autosomes (as the mutation rate is lower on chromosome X). A possible explanation is that there is more heterogeneity in mutation rate on the X chromosome, but this will require further investigation. In any case, more than 60% of the observed shared polymorphisms can not be explained by simple trinucleotide sequence-context effects. Since there was significant excess sharing in non-coding regions alone, variation in selective constraint between codons can not explain the finding, though variation in selective constraint at functional non-coding elements might contribute.

To investigate whether particular sequence features might be responsible for these hotspots of shared variants, the 10bp local sequence composition of shared variants was analysed and compared to the 10bp local sequence-context of all variants in the data set. For each variant, the identities of all nucleotides in its 10bp flanking neighbourhoods (either side of the variant) were determined using the reference sequence panTro2. All other locations of the trinucleotide, composed of the variant itself and its 3' and 5' neighbouring nucleotides, were recorded in a 100bp window around the position of the variant, together with the identities of each nucleotide within the 10bp sequences flanking the trinucleotide. The latter information of local base composition was used to normalise the 10bp sequence pattern around the variant. Similar to the previously described analysis, variants were grouped by their mutation type.

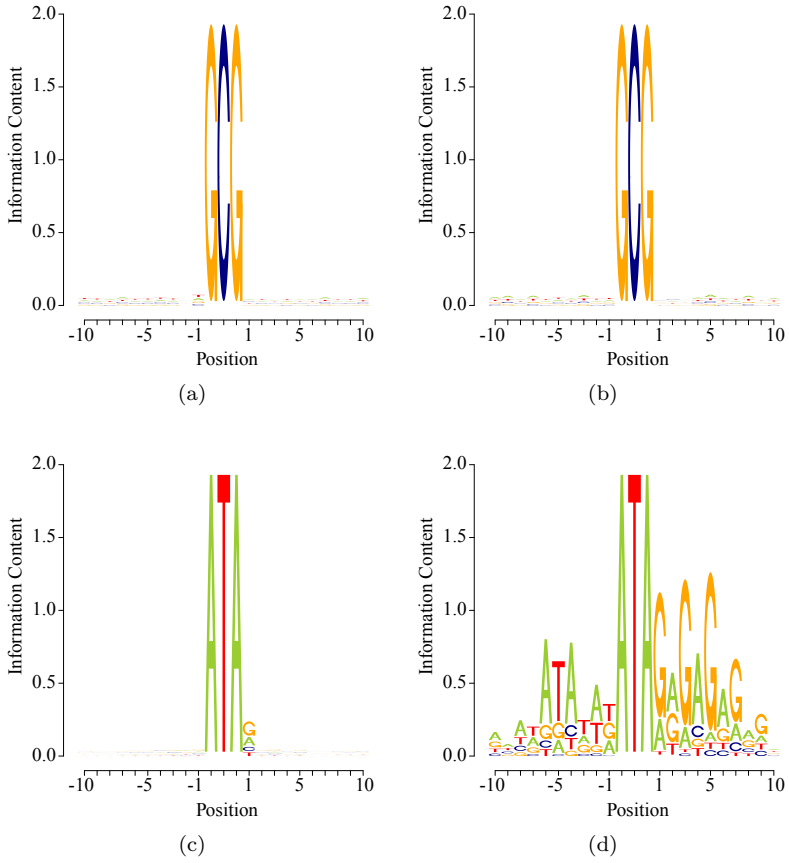


Figure 6.7: 10bp local sequence-context of (a) all GCG>T SNPs in the PanMap data set, (b) GCG>T SNPs shared between humans and chimpanzees, (c) all ATA>G SNPs, and (d) shared ATA>G SNPs.

For the different types of mutations, two distinct patterns have been observed: (i) at CpG-dinucleotide sites, no sequence motife conservation was detected around polymorphisms (neither around all variants in the PanMap data set nor around the ones shared between humans and chimpanzees, see Figures 6.7(a) and 6.7(b), respectively). (ii) While there was also no conservation of local sequence-context identified around all PanMap variants at non-CpG-sites (see Figure 6.7(c)), there was a strong signal for conservation around shared polymorphisms (see Figure 6.7(d)). The characteristics of these patterns were similar for most mutation types: they consisted of two consecutive short repeat sequences which were in agreement with the direction of the observed mutation (in this example an AT repeat followed by an AG repeat at the site of a shared ATA>G mutation), suggesting that these mutations might have been introduced through DNA polymerase slippage (a form of strand pairing that can cause small indels during replication especially in regions of repetitive DNA sequence e.g. in STRs; see Figure I.1) in low complexity regions of the genome. Thus, the observed heterogenerity in mutation rate seems to be mainly the consequence of either mutations at methylated CpG-sites or of strand-slippage mispairing at sites of low sequence complexity.

6.4 Simulation of Recurrent Mutations

To estimate how often shared variants can be expected due to recurrent mutations, data was simulated using the coalescent software ms [Hudson, 2002], matching the actual sample sizes (i.e. 20 chimpanzee haplotypes and 118 human ones). In order to model recurrent mutations although the program assumes the infinite sites model, the simulated sequence was divided into bins and each bin was regarded as a single nucleotide. As a consequence, recurrent mutations occurred when multiple mutations took place in the same bin. Only recurrent mutations that led to shared SNPs in the samples were considered (i.e. that were not fixed in the haplotypes sampled from either species). In these simulations, the phase was assumed to be known without error.

Given that heterogeneity in the mutation rate can greatly increase the rate of recurrent mutations and that CpG-sites in primates have a higher mutation rate than non-CpGs [Hodgkinson and Eyre-Walker, 2011], fragments consisting of both CpGs and non-CpGs were simulated by changing the bin sizes to model two different per nucleotide mutation rates (with the simplified assumption that all CpG and all non-CpG sites have the same mutation rates). Whereas shared SNPs were required to have the exact same two alleles in both species in the screen, recurrent mutations do not necessarily give rise to identical alleles in two species. To take this difference into consideration, the mutation rate in the simulation was modified accordingly. At each specific site, transitions and transversions were assumed to take place at different rates (Ts and Tv) and it was further assumed that two types of transversions were equally likely. Therefore, the probability that a mutation occurred in each species and generated the same derived allele was $Ts^2 + 0.5 \times Tv^2$, and the 'effective mutation rate' μ_e used in the simulation was the square root of this probability.

Given that 1% of dinucleotides in human genome are observed to be CpGs and both sites are hypermutable (since the G corresponds to the C in a CpG dinucleotide on the other strand), 2% of the genome consists of hypermutable CpG sites. If the overall mutation rate in the human genome is taken as 1.2×10^{-8} per generation per base pair [The 1000 Genomes Project Consortium, 2010], with an overall mutation rate at non-CpGs represented as μ, and a A-fold higher mutation rate at CpG-dinucleotides, then:

$$\mu \times 98\% + A\mu \times 2\% = 1.2 \times 10^{-8}.$$

Using the observation that 14.2% of the SNPs in the YRI sample occur at CpG sites (considering whether the ancestral allele creates a CpG site) as an estimate of the overall proportion of the mutations that occur at CpG sites, then:

$$\frac{A\mu \times 2\%}{\mu \times 98\% + A\mu \times 2\%} = 14.2\%.$$

	All	CpG-CpG	CpG-nonCpG	nonCpG-nonCpG
Frequency of simulations with shared SNP pairs	4.3×10^{-3}	2.0×10^{-3}	1.9×10^{-3}	4.3×10^{-4}
Frequency of simulations with shared SNP pairs that pass pairwise LD[1] and phase[2] criteria	1.6×10^{-4}	7.6×10^{-5}	6.0×10^{-5}	2.6×10^{-5}

[1] = r^2 >0.065 in humans (corresponding to p=0.005); Fisher Exact Test p<0.1 in chimpanzee.
[2] = The same alleles are linked in both species.

Table 6.2: The rate at which pairs of shared SNPs are observed within 4kb in coalescent simulations due to neutral, recurrent mutations.

By solving the two equations above, mutation rates of 8.55×10^{-8} and 1.05×10^{-8} at CpGs and non-CpGs, respectively, are obtained, a difference in rates consistent with previous estimates [Kong et al., 2012]. Furthermore, using the observation that 90.0% of the SNPs at CpG sites in YRI are transitions and 63.8% of the SNPs at non-CpG sites are transitions (similar to observations from *de novo* mutations [Kong et al., 2012], transition rates are estimated to be 7.7×10^{-8} at CpGs and 6.7×10^{-9} at non-CpGs, and transversion rates are estimated to be 8.5×10^{-9} at CpGs and 3.8×10^{-9} at non-CpGs. Finally, using these rates and the earlier equation, the effective mutation rates are 7.7×10^{-8} at CpGs and 7.2×10^{-9} at non-CpGs.

500,000 four kb segments were simulated consisting of 80 CpG sites (40 CpG dinucleotides) and 3,920 non-CpG sites, with a recombination rate of 1.2cM/Mb (i.e. the genome sex-averaged rate). The CpG dinucleotides were placed by choosing a position at random along the sequence, without replacement. A simple demographic history was assumed where human and chimpanzee populations maintained constant effective population sizes (N_{YRI}=27,000 and N_{chimp}=17,000, derived from a mutation rate of 1.2×10^{-8} and the diversity levels π of Yoruba and Western chimpanzee [Fischer et al., 2006; Hernandez et al., 2011]) until they merged 250,000 generations ago [Scally and Durbin, 2012]) to form an ancestral population with effective populations size of N_a=50,000 [Wall, 2003].

The simulation results suggested that shared SNP pairs were extremely rare (they occurred at a frequency of 0.43% - 2,142 out of 500,000 replicates - but note that, because there are more non-CpG sites in the genome, the frequency of pairs at two CpG sites was only ∼4 fold higher than pairs at two non-CpG sites; see Table 6.2). However, this estimate must be considered with care as the simulations assumed a fixed recombination rate, a uniform distribution of CpG sites and ignored the relationship of diversity/GC content and recombination, as well as other potentially salient details.

6.5 Annotation of Shared Polymorphisms

Polymorphisms shared through ancestry can only be maintained over the required time-scales by strong evolutionary forces such as balancing selection. Although previous studies have shown that large fractions of the non-coding genome are under evolutionary constraints conserving them across different mammal species [Dermitzakis et al., 2002], the majority of variants that lie in intergenic and intronic regions are thought to be either selectively neutral or the acting selection is too weak to conserve them over the required time-scales [Asthana et al., 2007; Dermitzakis et al., 2005; The Mouse Genome Sequencing Consortium, 2002]. The regions that exhibit the strongest selective pressures are usually positions of functional importance, such as coding sequences [Halligan and Keightley, 2006], and thus, shared polymorphisms which are due to shared ancestry rather than due to recurrent mutations are more likely to occur in genes.

In order to identify substitutions in genes that are potentially ancient in the lineages of both human and chimpanzee, all polymorphisms in the two species were annotated using the Variant Annotation Tool[26] (VAT) which identifies coding SNPs as (non-)synonymous, as introducing a premature stop, removing a stop, or overlapping a splice site, given a gene annotation set. In this study, a preprocessed annotation set derived from the GENCODE[27] Project (annotation release v3b) [Harrow et al., 2006] was used. As only a small fraction of the genome consists of genes (about 1-2% in humans [International Human Genome Sequencing Consortium, 2001]), the proportion of variants expected to lie in coding regions is quite small. An overview of all annotated variants is given in Table 6.3.

Human: In total, 75,137 of the 9,429,106 autosomal YRI SNPs (0.8%) were found in genes: 38,595 encoded non-synonymous changes, 35,744 encoded synonymous changes, 645 encoded changes towards a premature stop codon, and 365 created an overlapping splice site (some substitutions showed multiple effects). On chromosome X, 1,319 out of the 265,184 variants could be annotated (1,308 in the non-PAR of chromosome X and 11 in the PAR, representing 0.5% and 0.6% of all SNPs in those regions, respectively). From the annotated non-PAR chromosome X SNPs, 687 encoded non-synonymous changes, 608 encoded synonymous changes, while 12 substitutions encoded a premature stop codon, and six changes towards an alternative site for splicing have been detected. Out of the 11 annotated SNPs in the PAR, seven encoded non-synonymous substitutions, and four encoded synonymous substitutions. Neither changes towards a premature stop codon nor substitutions generating an alternative site for splicing have been identified.

[26] http://archive.gersteinlab.org/proj/VAT/
[27] http://www.gencodegenes.org/

	Statistics				
	# $SNPs_{(H)}$	# $SNPs_{(C,CpG)}$	# $SNPs_{(C,non\text{-}CpG)}$	# $SNPs_{(S,CpG)}$	# $SNPs_{(S,non\text{-}CpG)}$
Autosomes					
Original data	9,429,106	583,175	3,243,240	21,034	13,106
Annotated*	75,137	9,221	18,571	282	92
Synonymous	35,744	5,058	8,560	166	50
Non-synonymous	38,595	4,089	9,811	116	42
Premature Stop	645	71	165	1	0
Splice Overlap	365	3	90	0	0
Chromosome X non-PAR					
Original data	263,405	11,996	88,463	291	237
Annotated*	1,308	114	330	1	2
Synonymous	608	57	143	1	1
Non-synonymous	687	54	184	0	1
Premature Stop	12	5	4	0	0
Splice Overlap	6	0	1	0	0
Chromosome X PAR					
Original data	1,779	80	1,175	3	5
Annotated*	11	2	5	0	0
Synonymous	4	2	3	0	0
Non-synonymous	7	0	2	0	0
Premature Stop	0	0	0	0	0
Splice Overlap	0	0	0	0	0

* = Some substitutions have had more than one annotation, $_{C}$ = Chimpanzee, $_{H}$ = Human, $_{S}$ = Shared

Table 6.3: Variant annotation.

Chimpanzee: In chimpanzees, 27,792 of the 3,826,415 autosomal SNPs (0.7%) corresponded to substitutions within genes of which 13,900 encoded non-synonymous changes, 13,618 encoded synonymous substitutions, 277 encoded changes towards a premature stop codon, and 97 created an overlapping splice site (as in humans some substitutions showed multiple effects). On chromosome X, 444 out of the 100,459 variants in the non-PAR of chromosome X (0.4%), and seven out of the 1,255 SNPs in the PAR (0.6%) could be annotated. From the seven annotated SNPs in the PAR, two encoded non-synonymous changes while the other five encoded synonymous substitutions. In the non-PAR of chromosome X, 238 out of the 444 annotated SNPs encoded non-synonymous changes, 200 synonymous substitutions, nine changes resulted in the occurrence of a premature stop codon, and one SNP generated an alternative site for splicing (some changes showed multiple effects).

Shared: Out of the 34,140 observed shared polymorphisms on the autosomes, 374 were annotated in genes. 158 of those SNPs encoded non-synonymous substitutions while 216 encoded synonymous ones. One variant encoded a premature stop codon but no substitutions generating an alternative site for splicing have been identified. From the 528 shared SNPs in the non-PAR of chromosome X, three variants could be annotated, one of which encoded a non-synonymous substitution while the other two encoded synonymous changes. In contrast, none of the eight SNPs in the PAR could be annotated.

In theory, synonymous and non-synonymous sites are expected to evolve at the same rate in the absence of selection. This rate reflects the underlying mutation rate. However, as purifying selection acts on most genes to eliminate (often harmful) non-synonymous mutations, substitution rates at synonymous sites usually exceed substitution rates at non-synonymous sites [Hughes and Yeager, 1998]. On the other hand, elevated rates of non-synonymous compared to synonymous nucleotide substitutions can be observed when selective pressures act on the region to enhance codon substitutions [Lopez de Castro et al., 1982].

A deficit of non-synonymous substitutions in the autosomal polymorphism data set shared between humans and chimpanzees was identified at both CpG-dinucleotides and non-CpG-sites (41.1% and 45.7%, respectively) in comparison to the original PanMap data set (44.3% and 52.9%, respectively). This finding indicates that there was no evidence of substantial ancestry-based sharing and it suggests that most of the observed shared polymorphisms were generated in the two lineages by recurrent mutations of the same locus after speciation. This hypothesis was strengthened by the fact that the majority of the autosomal annotated shared variants (more than 75%) were located at highly mutable CpG-dinucleotides.

If non-synonymous shared polymorphisms were under balancing selection, one would expect to find other closely linked variants that hitch-hike with them. (The degree of the hitch-hiking will depend on the extent of recombination between the two loci [Hughes and Yeager, 1998].) In contrast to a locus closely linked to a locus under directional selection, which depicts a reduced rate of polymorphism compared to a neutral one due to the recent fixation of the favourable allele (a pattern referred to as 'selective sweep' as polymorphisms linked to the favoured allele are swept out of the population [Hughes and Yeager, 1998]), a locus tightly linked to an ancient polymorphism under balancing selection shows a higher rate of polymorphisms compared to a neutral locus [Nei and Li, 1980; Strobeck, 1983].

Interestingly, at non-CpG-sites, such a pattern of elevated SNP density was observed in close proximity of non-synonymous polymorphisms which were shared between humans and chimpanzees when compared to all non-synonymous polymorphisms in the PanMap data set (see Figure 6.8). At CpG-dinucleotides, the SNP density exhibited a similar range around both non-synonymous shared as well as non-synonymous non-shared polymorphisms. While this might be a sign of acting balancing selection, the possibility that this observed increase in the rate of variants around non-synonymous shared polymorphisms indicates an underlying local mutability, can not be excluded.

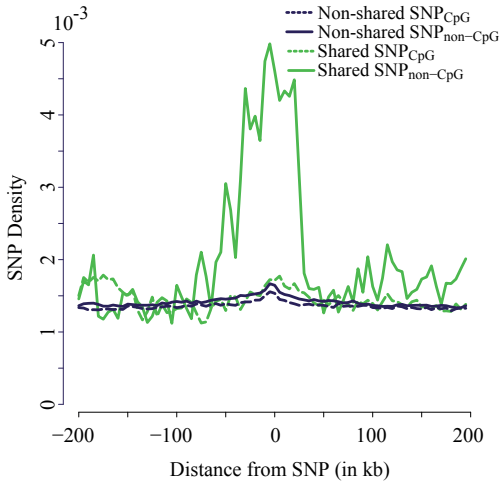


Figure 6.8: SNP density around non-synonymous shared (green) and non-shared (dark-blue) SNPs (measured in 5kb-sized bins). Putative CpG-mutations are indicated by dashed lines.

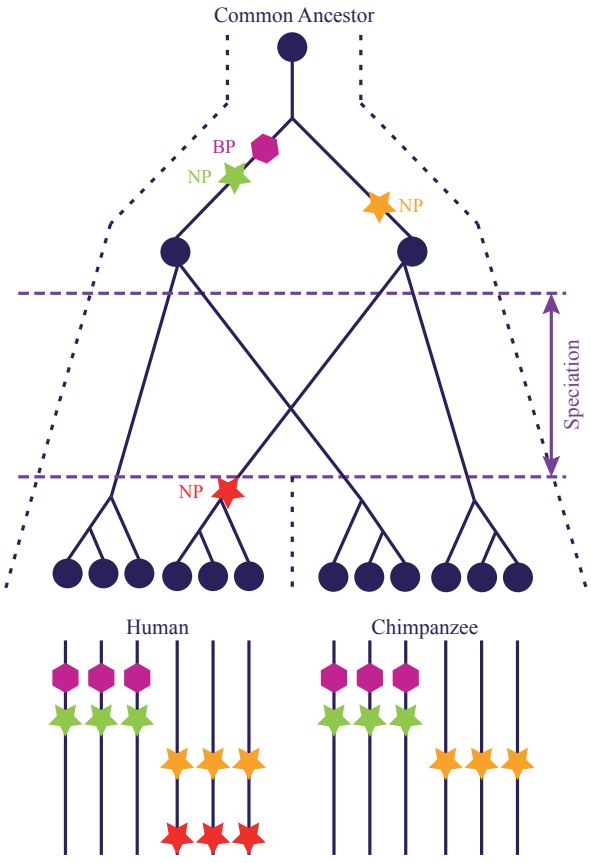


Figure 6.9: A polymorphism under balancing selection (BP) occurred before the human and chimpanzee species split from their most recent common ancestor. On the same branch of the tree, a neutral polymorphism (NP) occurred before the time of speciation. If this neutral variant is closely linked to the one under balancing selection, it can hitch-hike with this shared polymorphism and thus, after the speciation event, it can be detected on the same haplotype in the two descendent lineages if recombination has not yet been broken up the haplotype structure. As a result of this process, the presence of short haplotypes where the shared variant is in very strong linkage disequilibrium with another close by variant can be interpreted as a sign of shared polymorphism occurring through common ancestry.

6.6 Linkage Disequilibrium Patterns of Shared Polymorphisms

Variation along a sequence on a chromosome appears in so-called 'haplotypes', a linked set of alleles inherited *en bloc* from each parent. The analysis of the structure of these haplotypes can reveal relationships of different genetic polymorphisms which might reflect their shared ancestry due to the fact that significant non-random association between alleles of different loci separated by a large genetic distance can be observed as a result of natural selection [Abecasis et al., 2001; Peterson et al., 1995]. In particular, a signature of a shared polymorphism occurring through common ancestry would be the presence of short haplotypes where the shared variant is in very strong linkage disequilibrium with other close-by variants, indicating that the underlying haplotypes have not yet been broken up by recombination (as illustrated in Figure 6.9). It is very unlikely that these closely located polymorphisms have occurred by chance, suggesting that balancing selection acts on them.

For this reason, all shared variants were tested for a signal of strong LD (r^2 >0.8 in chimpanzee or human whereby r^2 > ranges from 0 - random association between two SNPs across the available haplotypes - to 1 - maximal positive or negative association considering the allele frequencies of the two SNPs) with another nearby shared variant (within a 15kb window) using VCFtools (Version 0.1.3.2) [Danecek et al., 2011] with the following parameters:

```
vcftools --vcf Chimp_genotypes_shared.vcf (or --vcf Human_genotypes_shared.vcf)
         --out shared_in_high_LD_15kb_0.8r2
         --min-alleles 2
         --max-alleles 2
         --hap-r2
         --phased
         --ld-window-bp 15000
         --min-r2 0.8,
```

where - -vcf specifies the input file, - -out specifies the file the output is written to, - -min-alleles and - -max-alleles indicate that only sites with a number of alleles within the specified range are included, while the - -hap-r2 option is used to report LD statistics. It informs VCFtools to output a file reporting the r2, D, and D' statistics using phased haplotypes (the haplotype version implies the option - -phased). The - -ld-window-bp option is used to define the maximum physical separation (in base pairs) of SNPs included in the LD calculation. Finally, the - -min-r2 sets a minimum value for r2 below which the LD statistic is not reported.

This analysis was restricted to autosomal shared variants due to the limited amount of data on chromosome X. Of the previously annotated autosomal shared polymorphisms, 31 SNPs showed a strong haplotype structure with another shared variant within either the 59 human individuals from the Yoruba population (seven SNPs), the 10 chimpanzee individuals from the PanMap study (21 SNPs), or both populations (three SNPs: two SNPs in *HLA-DPB1* and one SNP in *DNHD1*). Out of these 31 polymorphisms, 10 were non-synonymous, out of which only six resided outside highly mutable CpG-sites. Out of these six SNPs, four SNPs, two in the *HLA-DQA1* gene and two in the *HLA-DPB1* gene, were part of the MHC class II region while the other were part of the coding region of the genes *FER1L6* and *FMO2*. Table 6.4 lists all 374 shared polymorphisms that were observed in coding regions of genes (the ones showing a strong haplotype structure are indicated in green). The allele frequencies of all shared polymorphisms involved in the haplotype structure is presented in Table J.1 together with their distance to the shared coding polymorphism.

Few polymorphisms shared between humans and chimpanzees have been identified to date, most of which - apart from shared polymorphisms in the *TRIM5* gene [Cagliani et al., 2010] - are part of the MHC class I and class II antigenes [Ayala et al., 1994; Bjorkman et al., 1987; Brown et al., 1988; Hedrick, 1998; Hughes and Nei, 1988, 1989; Klein, 1987]. Therefore, the following two subchapters are dedicated to these known examples to assess whether the study was able to detect shared polymorphisms in those previously described regions. In addition, a third subchapter summarises potentially ancestral polymorphisms outside the MHC and *TRIM* genes which exhibited a strong linkage disequilibrium with nearby shared variants in chimpanzees, humans or both species.

6.6.1 Shared Polymorphisms in the Major Histocompatibility Complex

Some of the most polymorphic genes in vertebrate genomes known to date are part of the major histocompatibility complex, a multigenic region located on the short arm of chromosome 6 in humans (where it is also called the 'human leukocyte antigen' complex, HLA) and chimpanzees encompassing about 3.6Mb. The region is traditionally divided into three distinct gene segments, designated class I, class II, and class III (also referred to as 'inflammatory region' or 'central' MHC as it is separating the other two classes) which are derived from a common ancestor but differ not only structurally but also functionally [Kupfermann et al., 1992; The MHC sequencing consortium, 1999].

The MHC encodes cell-surface molecules that are key components in the control of the adaptive immunological activity. These molecules facilitate the detection and destruction of mutated or pathogen infected cells by specifically binding intracellulary derived antigens (through MHC class I molecules expressed on the surface of nearly all nucleated somatic cells in the body) or antigens

SNP		Statistics			
Chr	**Position**	**Gene**	**(N)S**[1]	**Putative CpG-mutation?**	**Strong haplotype structure?**[2,3]
chr1	1230647	CPSF3L	S	Yes	-
chr1	3232825	PRDM16	S	No	-
chr1	6343810	ACOT7	S	Yes	-
chr1	7714117	CAMTA1	S	Yes	-
chr1	12475070	VPS13D	NS	Yes	-
chr1	19172229	UBR4	S	Yes	-
chr1	20787672	KIF17	NS	Yes	-
chr1	20801210	KIF17	S	Yes	-
chr1	20804117	KIF17	S	Yes	-
chr1	21755955	RAP1GAP	S	Yes	-
chr1	24763295	RCAN3	S	Yes	-
chr1	26280831	TRIM63	S	No	-
chr1	32093618	COL16A1	S	Yes	-
chr1	43524736	ERMAP	NS	Yes	-
chr1	54003877	SLC1A7	S	No	-
chr1	90492913	GBP3	NS	Yes	-
chr1	110203904	C1orf59	S	Yes	-
chr1	110725840	KIAA1324	S	Yes	-
chr1	126649923	DRAM2	NS	Yes	-
chr1	122966810	AMPD1	NS	Yes	-
chr1	120493910	TTF2	S	No	-
chr1	119483516	SPAG17	NS	Yes	-
chr1	135587399	RHBG	NS	Yes	-
chr1	140103913	LY9	NS	No	-
chr1	144514343	LMX1A	S	No	-
chr1	148076920	DPT	S	Yes	-
chr1	150637486	FMO2	NS	No	C
chr1	150637526	FMO2	S	Yes	C
chr1	154688206	TNN	S	Yes	-
chr1	156858971	FAM5B	S	Yes	-
chr1	158656296	TOR3A	S	No	-
chr1	164527577	FAM129A	NS	No	-
chr1	180144553	NR5A2	S	Yes	-
chr1	181501341	CSRP1	S	Yes	-
chr1	181926651	LMOD1	NS	Yes	-
chr1	182158376	GPR37L1	S	No	-
chr1	186883750	IKBKE	NS	Yes	-
chr1	188634414	PLXNA2	S	No	-
chr1	190229759	IRF6	S	No	-
chr1	196338731	USH2A	S	Yes	-
chr1	203443941	FAM177B	NS	No	-
chr1	204293076	CAPN8	NS	No	-
chr1	204369991	CAPN8	S	Yes	-
chr1	204382799	CAPN8	NS	Yes	-
chr1	207817913	CDC42BPA	NS	Yes	-
chr1	214012399	MAP3K19	S	Yes	-
chr1	215076826	TARBP1	S	Yes	-
chr1	218465030	RYR2	S	No	-
chr1	218741381	ZP4	NS	Yes	-
chr2a	1467309	TPO	NS	Yes	-
chr2a	3302921	AC019172.1	S	Yes	-
chr2a	10151518	GRHL1	S	Yes	-
chr2a	38102205	PRKD3	NS	No	-
chr2a	70986694	AAK1	NS	No	-
chr2a	76481959	HK2	NS	No	-
chr2a	87269737	ELMOD3	NS	Yes	-
chr2b	148272627	AC016910.1	S	Yes	-
chr2b	174028576	LRP2	NS	Yes	-
chr2b	183512503	FKBP7	S	Yes	-
chr2b	188096866	NCKAP1	S	No	-

[1] = (Non-)Synonymous
[2] = $r^2 \geq 0.8$ within 15kb
[3] = in chimpanzees (C), in YRI (Y), or in both (B)

Table 6.4: Characteristics of all annotated shared variants. Shared polymorphisms that showed a strong haplotype structure within either the 59 human individuals from the Yoruba population, the 10 chimpanzee individuals from the PanMap study, or both populations are indicated in green.

SNP		Statistics			
Chr	Position	Gene	(N)S[1]	Putative CpG-mutation?	Strong haplotype structure?[2,3]
chr2b	224660888	TTLL4	S	Yes	-
chr2b	233792301	SLC19A3	NS	Yes	-
chr3	20366688	EFHB	S	Yes	-
chr3	39770817	SCN11A	NS	Yes	-
chr3	56229696	LRTM1	NS	Yes	-
chr3	58439538	IL17RD	NS	Yes	-
chr3	65439459	ATXN7	S	No	-
chr3	65594896	PRICKLE2	S	No	C
chr3	68126352	LRIG1	NS	Yes	-
chr3	74635785	SHQ1	NS	No	-
chr3	91824181	EPHA3	NS	Yes	-
chr3	112309162	IFT57	S	Yes	-
chr3	116260538	TMPRSS7	NS	Yes	-
chr3	189156557	KLHL24	S	Yes	Y
chr3	192654915	ST6GAL1	S	No	-
chr3	200047110	CPN2	S	Yes	-
chr3	200066698	LRRC15	S	No	Y
chr3	200067394	LRRC15	S	Yes	Y
chr4	361264	ZNF141	NS	No	-
chr4	6471983	WFS1	S	Yes	-
chr4	8448544	SH3TC1	S	Yes	-
chr4	60274381	C4orf35	NS	No	-
chr4	59907896	AC009570.2	NS	Yes	-
chr4	51959097	FRAS1	S	Yes	-
chr4	126145843	AC053545.2	NS	Yes	-
chr4	180700805	ASB5	S	No	-
chr4	187541940	DCTD	S	No	-
chr4	190961552	KLKB1	NS	Yes	-
chr4	194201924	HSP90AA4P	S	No	Y
chr5	336611	CCDC127	S	Yes	-
chr5	565916	PDCD6	S	Yes	-
chr5	13950688	DNAH5	S	No	-
chr5	74751104	CARD6	NS	No	-
chr5	48475463	MAST4	NS	Yes	-
chr5	40483601	COL4A3BP	S	No	-
chr5	21124820	C5orf36	NS	No	-
chr5	123622566	ZNF474	NS	Yes	-
chr5	129009568	MEGF10	S	Yes	-
chr5	139405481	KLHL3	S	No	-
chr5	152881190	ZNF300	S	Yes	-
chr5	153513818	FAT2	S	Yes	-
chr5	153538510	FAT2	NS	Yes	-
chr5	159669281	SOX30	NS	Yes	-
chr5	162341391	CCNJL	NS	No	-
chr5	162770336	ATP10B	S	Yes	-
chr5	172336779	DOCK2	S	Yes	-
chr5	181436690	GRM6	S	No	-
chr5	183544686	BTNL9	S	Yes	-
chr6	1725834	GMDS	S	No	-
chr6	2635524	C6orf195	NS	No	-
chr6	11377274	NEDD9	NS	Yes	-
chr6	25316492	FAM65B	NS	Yes	-
chr6	26216299	HIST1H2BA	S	No	-
chr6	30756814	TRIM31	NS	Yes	C
chr6	31611973	C6orf205	NS	Yes	-
chr6	31612682	C6orf205	S	Yes	-
chr6	31741577	C6orf15	S	Yes	C
chr6	31741664	C6orf15	S	Yes	C
chr6	32005470	HLA-B	S	No	C
chr6	32178919	AIF1	NS	Yes	-

[1] = (Non-)Synonymous
[2] = $r^2 \geq 0.8$ within 15kb
[3] = in chimpanzees (C), in YRI (Y), or in both (B)

Table 6.4

SNP		Statistics			
Chr	Position	Gene	(N)S[1]	Putative CpG-mutation?	Strong haplotype structure?[2,3]
chr6	33325789	HLA-DQA1	NS	No	C
chr6	33329605	HLA-DQA1	NS	No	-
chr6	33329665	HLA-DQA1	NS	No	-
chr6	33330263	HLA-DQA1	NS	No	C
chr6	33348451	HLA-DQB1	S	No	-
chr6	33348514	HLA-DQB1	S	No	-
chr6	33350080	HLA-DQB1	S	No	C
chr6	33493142	HLA-DOB	NS	Yes	-
chr6	33778430	HLA-DPB1	NS	No	-
chr6	33782690	HLA-DPB1	NS	No	B
chr6	33783414	HLA-DPB1	NS	No	B
chr6	47239885	ENPP5	S	Yes	-
chr6	50748831	RHAG	S	Yes	-
chr6	53023703	PKHD1	NS	Yes	-
chr6	55239819	C6orf142	NS	Yes	-
chr6	76053388	COL12A1	S	Yes	-
chr6	76827145	IMPG1	NS	Yes	-
chr6	90915762	MDN1	NS	Yes	-
chr6	109544860	SCML4	S	Yes	-
chr6	172741783	WDR27	NS	Yes	-
chr7	8143601	ICA1	NS	Yes	-
chr7	14823987	DGKB	S	Yes	-
chr7	20920197	ABCB5	NS	Yes	-
chr7	34125300	RP11-89N17.1	NS	Yes	-
chr7	40126500	AC004837.1	NS	No	-
chr7	48628649	PKD1L1	NS	Yes	-
chr7	75786381	SRCRB4D	NS	Yes	-
chr7	75825795	ZP3	S	Yes	-
chr7	94056116	COL1A2	S	No	-
chr7	95539648	DYNC1I1	S	No	-
chr7	99186425	PTCD1	S	Yes	C
chr7	99201896	PTCD1	NS	Yes	-
chr7	99405131	ZNF498	S	Yes	-
chr7	100831814	MUC3B	NS	Yes	-
chr7	100831817	MUC3B	NS	No	-
chr7	100869198	MUC12	S	No	-
chr7	103364659	RELN	S	No	-
chr7	105968464	CTA-351J1.1	NS	Yes	-
chr7	130643942	C7orf45	NS	Yes	-
chr7	132598134	PLXNA4	S	Yes	-
chr7	132962782	PLXNA4	S	Yes	C
chr7	139225785	SVOPL	S	Yes	-
chr7	151326180	TMEM176A	NS	Yes	-
chr7	151568165	ABCB8	S	Yes	-
chr7	157515960	RNF32	NS	Yes	-
chr8	3335533	CSMD1	S	No	-
chr8	24280171	ESCO2	NS	No	-
chr8	45762952	PRKDC	NS	No	-
chr8	103387757	TM7SF4	NS	No	-
chr8	103481385	DPYS	NS	Yes	-
chr8	114954430	TRPS1	S	No	-
chr8	123071764	FBXO32	S	Yes	-
chr8	123538356	FER1L6	NS	No	Y
chr8	124553919	ZNF572	S	Yes	-
chr8	133143752	ST3GAL1	S	Yes	C
chr8	138579796	COL22A1	NS	Yes	-
chr8	140100372	TRAPPC9	S	Yes	-
chr8	144814256	ZNF517	NS	Yes	-
chr9	15561595	TTC39B	NS	No	-
chr9	33415964	TMEM215	S	Yes	-

[1] = (Non-)Synonymous
[2] = $r^2 \geq 0.8$ within 15kb
[3] = in chimpanzees (C), in YRI (Y), or in both (B)

Table 6.4

SNP		Statistics			
Chr	**Position**	**Gene**	**(N)S**[1]	**Putative CpG-mutation?**	**Strong haplotype structure?**[2,3]
chr9	36579082	OR13J1	S	Yes	-
chr9	70469130	TMEM2	NS	Yes	-
chr9	76309486	GNA14	S	Yes	-
chr9	83268542	SLC28A3	S	Yes	-
chr9	96876074	NCBP1	S	Yes	C
chr9	111568714	SUSD1	NS	Yes	-
chr9	113520990	ZNF618	S	Yes	-
chr9	123943011	AL162724.4	NS	Yes	-
chr9	124158778	GPR144	NS	Yes	-
chr9	127513414	C9orf117	S	Yes	-
chr9	127571419	SH2D3C	NS	Yes	-
chr9	129836859	FNBP1	NS	No	-
chr9	130386217	HMCN2	NS	Yes	-
chr9	130883738	ABL1	S	Yes	-
chr9	131316865	PPAPDC3	NS	Yes	-
chr9	133198102	RALGDS	S	Yes	-
chr9	133338093	ABO	S	No	C
chr9	133793778	SARDH	S	Yes	-
chr9	136752811	NOTCH1	S	Yes	-
chr9	136985403	LCN8	NS	Yes	-
chr9	137498435	FAM166A	S	Yes	-
chr10	20031793	C10orf112	NS	Yes	-
chr10	70928358	CDH23	S	Yes	-
chr10	71350436	ASCC1	NS	No	-
chr10	80624609	SH2D4B	NS	Yes	-
chr10	103858346	PDCD11	S	Yes	-
chr10	120208300	GRK5	NS	Yes	-
chr10	123181469	BTBD16	S	Yes	-
chr10	123328791	HTRA1	S	Yes	-
chr11	936802	TSPAN4	S	Yes	-
chr11	1173115	MUC2	S	Yes	-
chr11	4883890	OR52E2	S	Yes	-
chr11	5377938	OR52V1P	S	Yes	-
chr11	5574754	OR56B1	NS	Yes	-
chr11	5744426	OR52E5	S	Yes	-
chr11	6395029	DNHD1	S	No	B
chr11	6413257	DNHD1	NS	No	-
chr11	6704770	OR10A5	NS	Yes	-
chr11	14184773	SPON1	S	No	-
chr11	17569020	OTOG	NS	No	-
chr11	59173770	PRPF19	S	Yes	-
chr11	59202707	TMEM132A	S	Yes	-
chr11	65797424	SSH3	S	Yes	-
chr11	73796459	KLHL35	NS	Yes	-
chr11	84341079	SYTL2	S	No	-
chr11	84595991	PICALM	S	Yes	-
chr11	90877111	FAT3	S	Yes	-
chr11	112195445	ANKK1	S	Yes	-
chr11	117933726	HYOU1	S	Yes	-
chr11	125256634	SRPR	S	Yes	-
chr12	6781813	NOP2	NS	Yes	-
chr12	65492869	SOX5	S	Yes	-
chr12	49365791	MUC19	NS	Yes	-
chr12	40395367	TROAP	S	Yes	-
chr12	38819991	TMPRSS12	NS	Yes	-
chr12	33802143	MMP19	NS	Yes	-
chr12	29764560	SLC16A7	S	Yes	-
chr12	70889656	PTPRB	S	Yes	-
chr12	73032350	TRHDE	S	Yes	-
chr12	80896298	C12orf64	S	No	-

[1] = (Non-)Synonymous
[2] = $r^2 \geq 0.8$ within 15kb
[3] = in chimpanzees (C), in YRI (Y), or in both (B)

Table 6.4

SNP		Statistics			
Chr	Position	Gene	(N)S[1]	Putative CpG-mutation?	Strong haplotype structure?[2,3]
chr12	81277774	PTPRQ	S	Yes	C
chr12	92051030	DCN	S	Yes	-
chr12	97013155	LTA4H	NS	No	-
chr12	102769269	MYBPC1	S	Yes	-
chr12	105075003	HSP90B1	S	Yes	-
chr12	109443406	WSCD2	S	Yes	-
chr12	120696937	HSPB8	S	Yes	-
chr12	124299863	ZCCHC8	S	Yes	-
chr12	125741730	DNAH10	S	Yes	-
chr12	130983395	GLT1D1	S	No	-
chr13	22263643	AL354828.3	NS	Yes	-
chr13	22921227	SACS	S	Yes	-
chr13	25155973	ATP8A2	S	Yes	-
chr13	27322449	POLR1D	S	Yes	C
chr14	22081010	SLC7A8	S	Yes	-
chr14	77075204	POMT2	S	No	-
chr14	78790724	NRXN3	S	Yes	-
chr14	91001125	AL159191.1	NS	Yes	-
chr14	96572994	AL355102.1	S	No	-
chr15	20515613	CYFIP1	S	Yes	-
chr15	23144949	ATP10A	S	Yes	C
chr15	25663989	OCA2	S	Yes	-
chr15	42870320	SQRDL	S	No	-
chr15	56160355	HSP90AB4P	NS	No	-
chr15	59497438	VPS13C	S	Yes	-
chr15	60707187	LACTB	S	Yes	-
chr15	67146745	PAQR5	S	Yes	-
chr15	71123088	HCN4	S	Yes	-
chr15	78910363	KIAA1199	S	No	-
chr15	82275972	WDR73	S	Yes	-
chr15	82494670	ALPK3	NS	Yes	-
chr15	83224173	AKAP13	S	Yes	-
chr16	370556	AC005202.1	NS	Yes	-
chr16	674849	RHOT2	S	Yes	-
chr16	729449	CCDC78	NS	No	-
chr16	1058529	AC009041.5	NS	No	-
chr16	2073586	NOXO1	S	No	-
chr16	2421100	ABCA3	S	Yes	-
chr16	3331214	OR1F1	NS	Yes	-
chr16	3674535	NLRC3	S	Yes	-
chr16	3718730	BTBD12	S	Yes	-
chr16	7336906	AC022206.2	NS	No	-
chr16	23447400	SCNN1G	S	Yes	-
chr16	23660629	COG7	NS	Yes	-
chr16	47464838	ABCC11	NS	Yes	-
chr16	50009746	NOD2	NS	Yes	-
chr16	57099706	GPR56	NS	Yes	C
chr16	58180349	SLC38A7	NS	Yes	-
chr16	75493081	CHST5	NS	Yes	-
chr16	81365664	BCMO1	NS	Yes	-
chr16	81811986	AC099524.1	S	Yes	-
chr16	82074099	PLCG2	S	Yes	-
chr16	84702153	ATP2C2	S	Yes	-
chr16	86840927	FOXF1	S	Yes	-
chr16	87669777	FBXO31	S	Yes	-
chr16	90466180	GAS8	S	Yes	-
chr17	1761917	SMYD4	NS	Yes	-
chr17	4112516	ZZEF1	S	Yes	-
chr17	5031929	GP1BA	NS	Yes	-
chr17	5247948	USP6	NS	Yes	-

[1] = (Non-)Synonymous
[2] = $r^2 \geq 0.8$ within 15kb
[3] = in chimpanzees (C), in YRI (Y), or in both (B)

Table 6.4

SNP		Statistics			
Chr	Position	Gene	(N)S[1]	Putative CpG-mutation?	Strong haplotype structure?[2,3]
chr17	6722631	KIAA0753	NS	Yes	-
chr17	44716356	DNAH9	S	Yes	-
chr17	29579969	NOS2	NS	Yes	-
chr17	22004414	UNC45B	S	Yes	-
chr17	21239943	RDM1	S	Yes	C
chr17	15808895	HAP1	NS	No	-
chr17	49117853	PPP1R9B	S	Yes	-
chr17	75429060	UNK	S	Yes	-
chr17	78166434	DNAH17	S	Yes	-
chr17	78249300	DNAH17	NS	Yes	-
chr17	78667905	LGALS3BP	S	Yes	-
chr17	80356478	CARD14	NS	Yes	-
chr17	80513672	RNF213	S	Yes	-
chr17	82306192	FASN	S	Yes	-
chr18	11253843	EPB41L3	NS	Yes	-
chr18	7325376	ANKRD12	S	Yes	-
chr18	3409721	CEP192	NS	Yes	-
chr18	49747967	DCC	S	Yes	-
chr18	55052177	ALPK2	S	Yes	-
chr18	55710778	GRP	S	No	-
chr18	71400616	CNDP1	NS	Yes	-
chr18	77113199	ADNP2	S	Yes	-
chr19	2264295	JSRP1	NS	Yes	-
chr19	3033730	TLE6	S	Yes	-
chr19	4931823	M6PRBP1	NS	Yes	-
chr19	7719980	C19orf45	NS	Yes	-
chr19	8306757	FBN3	S	Yes	-
chr19	8742743	PRAM1	NS	Yes	Y
chr19	8835066	ADAMTS10	S	Yes	-
chr19	10624440	TYK2	NS	Yes	Y
chr19	11708382	RGL3	S	No	-
chr19	12618175	ZNF44	NS	Yes	-
chr19	14814233	CD97	S	Yes	-
chr19	14820533	CD97	S	Yes	C
chr19	17181964	NWD1	NS	Yes	-
chr19	19408929	HOMER3	S	Yes	-
chr19	39555286	RPS4P21	S	No	-
chr19	43428199	WDR87	S	Yes	-
chr19	44515912	FBXO17	S	Yes	-
chr19	49172348	AC006276.1	NS	Yes	-
chr19	49984480	ZNF285B	S	Yes	-
chr19	55270788	RCN3	NS	Yes	-
chr19	57619677	ZNF615	NS	Yes	-
chr19	57769946	ZNF616	S	Yes	-
chr19	59919115	LILRB3	S	No	-
chr20	6013586	FERMT1	S	Yes	-
chr20	7855496	HAO1	S	Yes	-
chr20	17673464	BFSP1	S	Yes	-
chr20	17676088	BFSP1	NS	Yes	-
chr20	30395014	CDK5RAP1	S	Yes	-
chr20	54889737	BMP7	NS	Yes	-
chr20	57497417	PHACTR3	S	Yes	-
chr20	60185366	CABLES2	S	Yes	-
chr21	30265188	KRTAP19-4	NS	Yes	-
chr21	39379417	B3GALT5	NS	Yes	-
chr21	42076815	TMPRSS3	NS	Yes	-
chr21	42089295	TMPRSS3	S	Yes	-
chr21	43060521	SIK1	S	Yes	-
chr21	43810386	C21orf32	NS	Yes	-
chr21	45827849	FTCD	NS	Yes	-

[1] = (Non-)Synonymous
[2] = $r^2 \geq 0.8$ within 15kb
[3] = in chimpanzees (C), in YRI (Y), or in both (B)

Table 6.4

SNP		Statistics			
Chr	**Position**	**Gene**	**(N)S**[1]	**Putative CpG-mutation?**	**Strong haplotype structure?**[2,3]
chr22	15712142	AC007064.4	NS	Yes	-
chr22	16658658	BID	NS	No	-
chr22	19107480	SCARF2	S	No	-
chr22	19717581	AIFM3	NS	No	-
chr22	23490492	SGSM1	S	Yes	-
chr22	30011746	INPP5J	NS	Yes	-
chr22	31320787	BPIL2	NS	No	-
chr22	36119048	RAC2	S	Yes	-
chr22	38547045	CACNA1I	NS	Yes	-
chr22	39683951	MCHR1	S	Yes	-
chr22	42217485	BIK	S	Yes	-
chr22	43262460	PARVB	S	Yes	-
chr22	44744093	FBLN1	S	Yes	-
chr22	45498617	GTSE1	NS	Yes	-

[1] = (Non-)Synonymous
[2] = $r^2 \geq 0.8$ within 15kb
[3] = in chimpanzees (C), in YRI (Y), or in both (B)

Table 6.4

from extracellular pathogens (through MHC class II molecules primarily expressed on cells of the white blood cell lineage) to glycoproteins on the cell surface and presenting these peptides to cytotoxic T-cells. The T-cells recognise these cells as different from non-infected ones and release cytokines that initiate an antigen-specific response of the immune system [Trowsdale, 1993].

MHC class I molecules are heterodimers consisting of a heavy chain (also called 'α-chain') encoded by a class of highly polymorphic loci of which three classical ones occur in humans (HLA-A, HLA-B, and HLA-C), and a single-domained β_2-microglobulin encoded by a non-polymorphic loci outside the MHC region [Hughes and Yeager, 1998]. In contrast, MHC class II molecules are heterodimers consisting of an α-chain and a β-chain, both of which are encoded in the MHC region. In mammals, the MHC class II region is subdivided into three isotype regions (HLA-DR, HLA-DP, and HLA-DQ), each comprising an α-chain gene and one or more genes encoding β-chains [de Groot and Bontrop, 1999; Hughes and Yeager, 1998; Slierendregt et al., 1993]. Evolution in the HLA-DQ region mainly occurs through an accumulation of point mutations [Bontrop, 2006; Marsh et al., 2010]. In contrast, HLA-DP evolves mainly due to recombination [Bontrop, 2006; Marsh et al., 2010] while the evolution the HLA-DR region has been shown to be driven by both point mutations and recombination processes [Bontrop, 2006].

The two MHC subfamilies show an excessive variation which is not only expressed at the allele level through a great sequence divergence amongst alleles, allowing diverse immunological and behavioural functions, but also at the population level as those allelic polymorphisms frequently cluster in distinct allelic lineages [Arden and Klein, 1982; de Bakker et al., 2006; Horton et al., 1998; Robinson et al., 2009]. Several ancient polymorphisms in both the MHC class I and class II genes have been shown to have survived over long evolutionary periods predating the human-chimpanzee

speciation (some of them have been shown to even predate the divergence of hominoids and Old World monkeys) [Bontrop et al., 1999; Fan et al., 1989; Gyllensten et al., 1991a; Gyllensten and Erlich, 1989; Gyllensten et al., 1991b; Klein, 1987; Lawlor et al., 1988; Mayer et al., 1988; Otting et al., 1992]. They have been maintained in the two different lineages through strongly acting long-term balancing selection (most likely through over-dominant selection as co-dominantely expressed alleles enable heterozygous individuals to bind a broader spectrum of peptides resulting in a broader immune surveillance than homozygous genotypes).

In concordance with previously published work, this study identified 12 polymorphisms that were shared between humans and chimpanzees in the MHC class I and class II regions: one polymorphism in *HLA-B*, four in *HLA-DQA1*, three in *HLA-DQB1*, one in *HLA-DOB*, and three in *HLA-DPB1* (see Table 6.4). Six of those 12 shared variants showed a strong haplotype structure (four in the 10 chimpanzee individuals and two in both the studied chimpanzee and human individuals) indicating shared ancestry due to a strong long-term balancing selection. Four out of the 12 polymorphisms, one in *HLA-B* and three in *HLA-DQB1*, encoded a synonymous change. These could be selectively neutral, but these loci might hitch-hike with other closely linked ancestral polymorphisms under balancing selection. Out of the eight shared polymorphisms encoding non-synonymous substitutions, only one, the one in the *HLA-DOB* gene, was a putative CpG-mutation which also did not display a strong linkage disequilibrium with nearby shared variants. Thus, it is more likely that it resulted from convergent evolution rather than shared ancestry.

6.6.2 Shared Polymorphisms in the TRIpartite Motif Family

The human *TRIM5* gene belongs to the TRIpartite Motif protein family counting more than 70 members in the human genome [Cagliani et al., 2010]. Despite the family's common domain architecture consisting of an amino-terminal RING finger domain, one or two B-box domains, a long coiled-coil domain, and, additionally in most genes, a C-terminal domain [Reymond et al., 2001], TRIM proteins are implicated in a variety of cellular functions. A large number is involved in processes associated with innate immune response such as *TRIM31* that specifically inhibits the HIV entry in humans [Uchil et al., 2008] and *TRIM5* which encodes an anti-retroviral restriction factor associated with a reduced risk of HIV-infection [Cagliani et al., 2010].

A previous analysis has shown that the *TRIM5* gene is one of the few examples outside the MHC in primates which has been the target of exceptionally strong long-term balancing selection: two polymorphisms, one SNP and one indel, that predate the human-chimpanzee speciation event have been identified that have been maintained by balancing selection for millions of years, possibly due to their effects on transcription-factor binding sites [Cagliani et al., 2010]. In contrast to the

study of Cagliani et al., this work could not detect the previously identified shared polymorphisms in the *TRIM5* gene as no polymorphisms in the *TRIM5* gene have been identified in the 10 Western chimpanzee individuals of the PanMap study. (Cagliani et al.'s paper did not indicate to which subspecies the three chimpanzee individuals belonged on which their study was based on and thus, it is unclear whether these shared variants were not detected in the PanMap study or whether they are simply not present in Western chimpanzees.) However, this study identified shared variants in two other *TRIM* genes, *TRIM31* and *TRIM63*. The *TRIM63* polymorphism encoded a synonymous substitution while the shared polymorphism in the *TRIM31* gene encoded a non-synonymous change. The latter shared polymorphism showed a strong linkage disequilibrium with nearby shared variants in chimpanzees suggesting that balancing selection might act on the gene. However, as the shared variant is a putative CpG-mutation, the possibility of convergent evolution of unrelated alleles after speciation can not be excluded.

Figure 6.10 shows the location of the shared non-synonymous polymorphism in *TRIM31* together with the positions of all annotated coding polymorphisms in human and chimpanzee in this gene. Above and below the central figure, the average read depth per sample is given for the two species.

6.6.3 Shared Polymorphisms outside the MHC and TRIM Genes

Outside the MHC region and genes of the *TRIM* family in which shared polymorphisms between humans and chimpanzees have already been documented in previous studies [Ayala et al., 1994; Bjorkman et al., 1987; Brown et al., 1988; Cagliani et al., 2010; Hedrick, 1998; Hughes and Nei, 1988, 1989; Klein, 1987], 24 shared polymorphisms in 21 genes have been identified in this study which were present on short haplotypes where the shared variants were in very strong linkage disequilibrium with each other in either chimpanzees (16 SNPs), humans (seven SNPs), or both species (one SNP), an indication of shared ancestry through a strongly acting balancing selection. The majority (>79%) of these identified shared polymorphisms encoded synonymous changes. Although synonymous substitutions are frequently selectively neutral, they might hitch-hike with a closely linked locus under selection. Alternatively, they might be under selection if they alter an open-reading frame, transcription factor binding-sites, or if they are epistatic.

Interesting genes with shared polymorphisms with a strong haplotype structure include:

ABO *ABO*[28] encodes a protein, histo-blood group ABO system transferase, that is the basis of the ABO blood group system used to date. The change at a non-CpG-site in this gene, which

[28] http://www.genecards.org/cgi-bin/carddisp.pl?gene=ABO&search=ABO

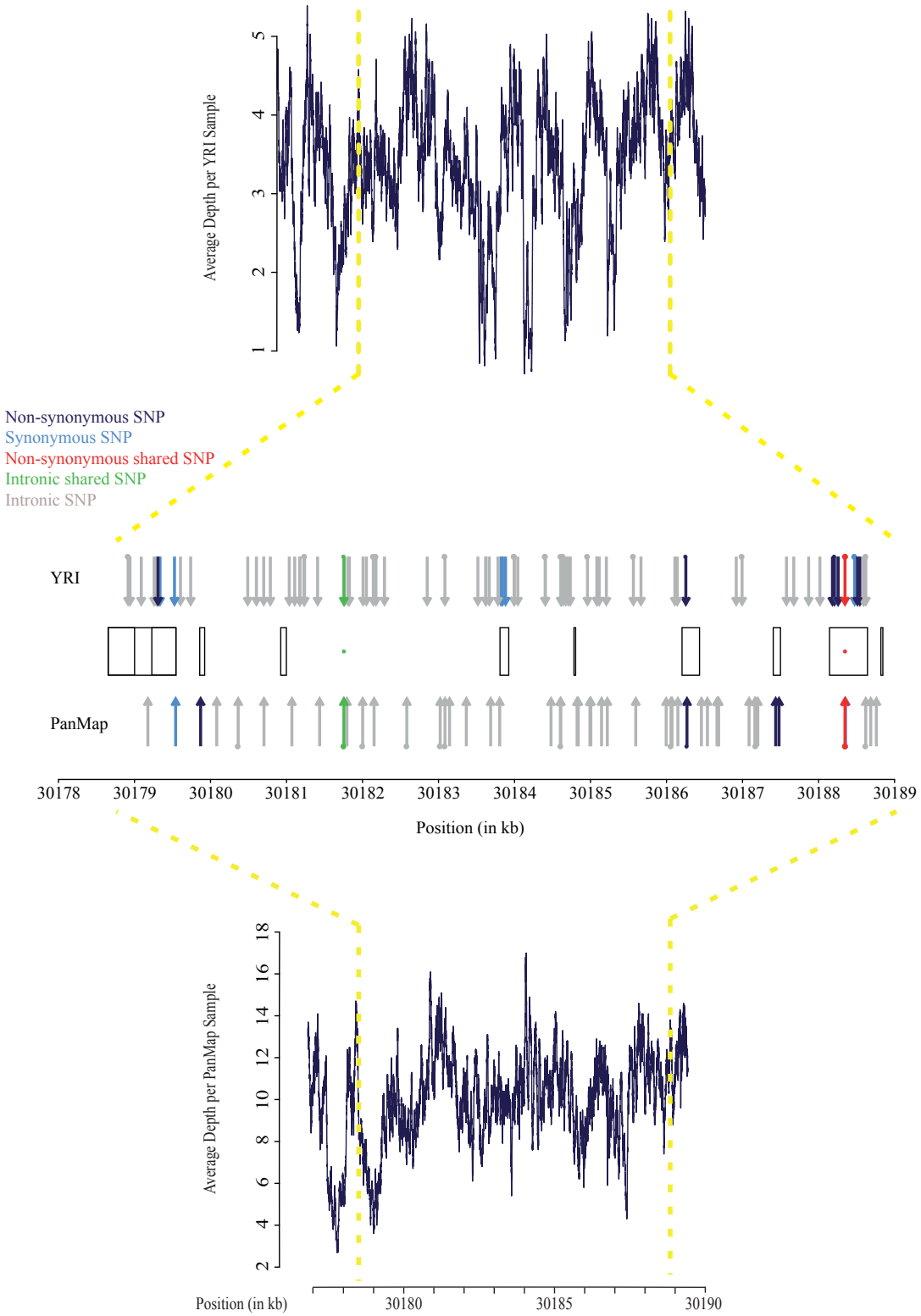


Figure 6.10: Location of the two shared polymorphisms within the *TRIM31* gene: one non-synonymous shared variant which showed a strong linkage disequilibrium with nearby shared variants in chimpanzee (depicted as a red arrow) and one non-coding shared variant (green arrow). The positions of all other synonymous (light-blue arrows) and non-synonymous (dark-blue arrows) coding polymorphisms as well as intronic polymorphisms (grey arrows) in this gene in the human (YRI) and chimpanzee (PanMap) sample are also illustrated. A dot above an arrow marks a putative CpG-mutation. Above and below the central figure, the average read depth per sample is given for the two species; the region of the *TRIM31* gene is marked by yellow dashed lines. The positions are given in human (hg18) coordinates and the gene position was established based on human genic regions extracted from the 'Known Genes' data set [Hsu et al., 2006] of the UCSC table browser [Karolchik et al., 2004]. Exons are illustrated as boxes.

showed a strong haplotype structure in chimpanzees, encoded a synonymous substitution with an alternate allele frequency of 0.75 (see Figure 6.11).

FMO2 FMO2 [29], a flavin-containing monooxygenase, is a NADPH-dependent enzyme that catalyses the oxidation of many drugs and xenobiotics in humans. Two shared polymorphisms which exhibited a strong haplotype structure in chimpanzees have been identified in this gene (one synonymous change at a CpG-site and one non-synonymous change at a non-CpG-site, both with an alternate allele frequency of 0.9; see Figure 6.12) suggesting that the *FMO2* gene might be under balancing selection.

PRAM1 The *PRAM1* [30] gene, which is expressed and regulated during normal myelopoiesis, encodes the PML-RARA-regulated adapter molecule 1. The encoded protein is similar to an adaptor protein, FYN binding protein, involved in T-cell receptor mediated signalling [Clemens et al., 2004; Moog-Lutz et al., 2001]. The non-synonymous shared polymorphism with an alternate allele frequency of 0.78 was in strong linkage disequilibrium with other nearby shared variants in humans (see Figure 6.13).

PTPRQ *PTPRQ* [31] encodes a type III receptor-like protein-tyrosine phosphatase that plays an important role in adipogenesis of mesenchymal stem cells. The protein is able to dephosphorylate a broad range of phosphatidylinositol phosphates which are involved in regulation of survival, proliferation, and sub-cellular architecture [Seifert et al., 2003]. One synonymous shared variant, with an alternate allele frequency 0.95, which showed a strong haplotype structure in chimpanzees, has been detected in this genic region. (No figure could be provided as the genic regions extracted from the 'Known Genes' data set [Hsu et al., 2006] of the UCSC table browser [Karolchik et al., 2004] did not match the regions annotated in the GENCODE data set).

6.7 Discussion

A map of genome-wide sequence variation shared between Western chimpanzees, *Pan troglodytes verus*, and the 59 human individuals from the Yoruba population from Ibadan, Nigeria, sequenced as part of the 1000 Genomes Project, has been created containing 34,140 autosomal polymorphisms with a Ts/Tv ratio of 12.56 and 536 variants on chromosome X with a Ts/Tv ratio of 9.94. Although these shared polymorphisms could have been introduced through systematic se-

[29] http://www.ncbi.nlm.nih.gov/sites/entrez?Db=gene&Cmd=ShowDetailView&TermToSearch=2327
[30] http://www.genecards.org/cgi-bin/carddisp.pl?gene=PRAM1&search=PRAM1
[31] http://www.genecards.org/cgi-bin/carddisp.pl?gene=PTPRQ&search=PTPRQ

quencing errors or false positive SNP calls (e.g. through collapsed paralogues), the pattern of these polymorphisms, in particular their read coverage, their mapping qualities as well as their site frequency spectrum, was inconsistent with substantial problems in the variant calling process. The probability of observing such shared polymorphisms by chance for selectively neutral alleles in a randomly mating population is very small. Consequently, it is more likely that one out of two biological processes generated these shared variants in the genomes of human and chimpanzee: (i) convergent evolution or (ii) shared ancestry due to the maintenance of polymorphisms over millions of years since they first occurred in the common ancestral species through strongly acting long-term balancing selection, a scenario which is thought to be extremely rare.

The study found that variants were shared in excess of what would have been expected by chance. Despite the fact that the majority of these variations were located at CpG-sites, which are known hotspots for point mutations in mammalian genomes, suggesting that they were most likely the result of recurrent mutation at the same locus after speciation rather than being shared through ancestry, it was very difficult to distinguish events of convergent evolution from acting balancing selection for single variant sites. However, short ancestral segments on which a balancing polymorphism is located might contain additional shared ancestral polymorphisms, a scenario that is unlikely to occur either by chance (as recombination should have had sufficient time to break up the underlying haplotypes) or through convergent evolution. As a consequence, this study focused on cases of two or more shared variants in close proximity to each other, which exhibited the same pattern of linkage disequilibrium in the two species. Besides the MHC, for which ancestral shared polymorphisms have been previously detected [Bontrop et al., 1999; Fan et al., 1989; Gyllensten et al., 1991a; Gyllensten and Erlich, 1989; Gyllensten et al., 1991b; Klein, 1987; Lawlor et al., 1988; Mayer et al., 1988; Otting et al., 1992], this work identified over 20 such loci which are good candidates for regions under balancing selection.

The results of this study suggest that long-term balancing selection might be more common than previously believed. Nevertheless, further investigations will be required to rule out events of deep coalescence that occurred by chance (e.g. by the analysis of patterns of variation in additional closely related primate species such as gorilla or bonobo).

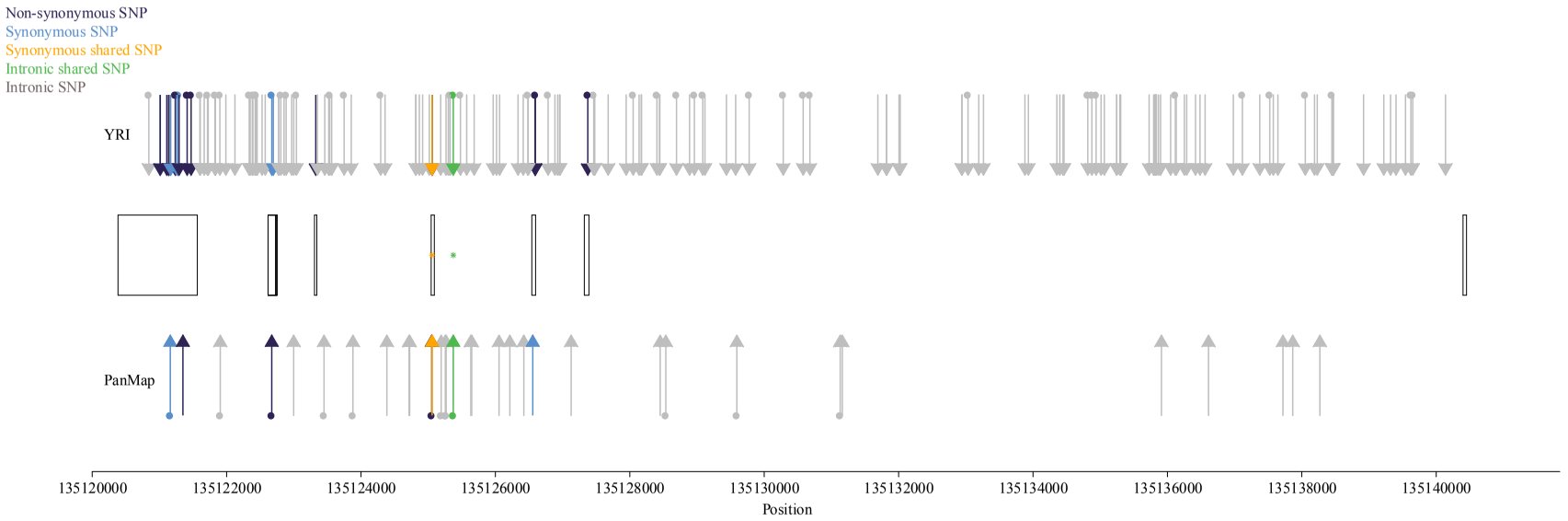


Figure 6.11: Location of the two shared polymorphisms within the *ABO* gene: one synonymous shared variant which showed a strong linkage disequilibrium with nearby shared variants in chimpanzee (depicted as an orange arrow), and one non-coding shared variant (green arrow). The positions of all other synonymous (light-blue arrows) and non-synonymous (dark-blue arrows) coding polymorphisms, as well as intronic polymorphisms (grey arrows) in this gene in the human (YRI) and chimpanzee (PanMap) sample are also illustrated. A dot above an arrow marks a putative CpG-mutation. The positions are given in human (hg18) coordinates and the gene position was established based on human genic regions extracted from the 'Known Genes' data set [Hsu et al., 2006] of the UCSC table browser [Karolchik et al., 2004]. Exons are illustrated as boxes.

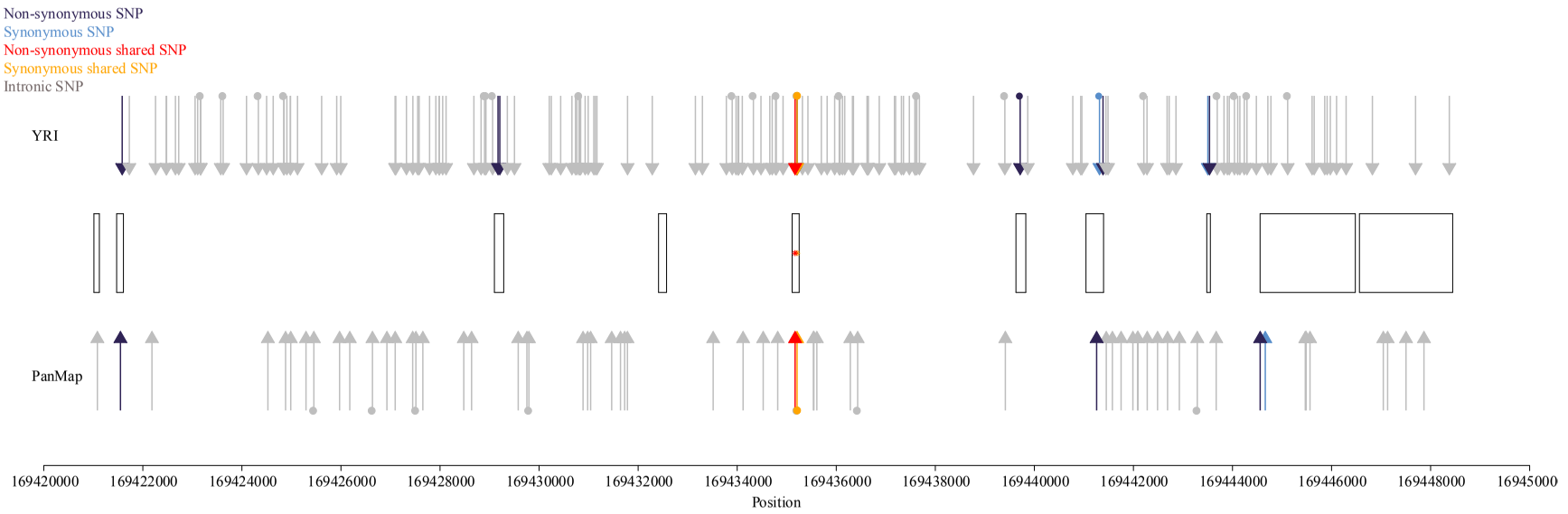


Figure 6.12: Location of the non-synonymous shared polymorphism which showed a strong linkage disequilibrium with nearby shared variants in chimpanzee (depicted as an red arrow) within the *FMO2* gene. The positions of all other synonymous (light-blue arrows) and non-synonymous (dark-blue arrows) coding polymorphisms, as well as intronic polymorphisms (grey arrows) in this gene in the human (YRI) and chimpanzee (PanMap) sample are also illustrated. A dot above an arrow marks a putative CpG-mutation. The positions are given in human (hg18) coordinates and the gene position was established based on human genic regions extracted from the 'Known Genes' data set [Hsu et al., 2006] of the UCSC table browser [Karolchik et al., 2004]. Exons are illustrated as boxes.

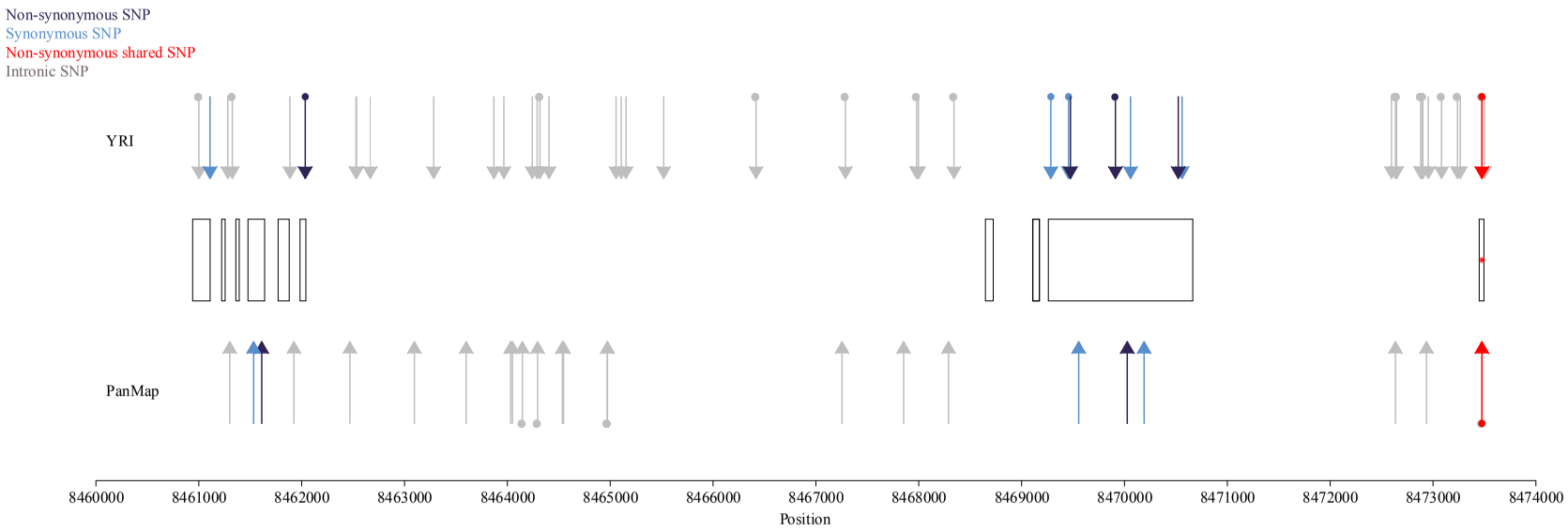


Figure 6.13: Location of the non-synonymous shared polymorphism which showed a strong linkage disequilibrium with nearby shared variants in chimpanzee (depicted as an red arrow) within the *PRAM1* gene. The positions of all other synonymous (light-blue arrows) and non-synonymous (dark-blue arrows) coding polymorphisms, as well as intronic polymorphisms (grey arrows) in this gene in the human (YRI) and chimpanzee (PanMap) sample are also illustrated. A dot above an arrow marks a putative CpG-mutation. The positions are given in human (hg18) coordinates and the gene position was established based on human genic regions extracted from the 'Known Genes' data set [Hsu et al., 2006] of the UCSC table browser [Karolchik et al., 2004]. Exons are illustrated as boxes.

Chapter 7

Discussion

Although germ line mutation rates are generally low throughout most genomes ($\sim 2.5 \times 10^{-8}$ per nucleotide per generation in humans [Ellegren et al., 2003; McVean and Hurst, 1997; Nachman and Crowell, 2000])), they are an interesting subject of study as they vary extensively between different genomic sites. A better knowledge of this variation, as well as the patterns underlying mutational processes across the genome and the forces driving it, are essential in medical genetics as mutations can have severe impacts on human health by causing genetic diseases [Ellegren et al., 2003]. In addition, the understanding of mutation rates and their variation is important to shed light on many biological phenomena (such as the evolution of sex chromosomes or the question of how and why different species evolved) and evolutionary processes that shaped genome sequences (such as recombination, replication, repair, and transcription [Duret, 2009]). In fact, the study of mutation rate variation within and between species is one of the main goals of evolutionary and population genetics as, together with an understanding of nucleotide diversity levels, it can aid gaining a better knowledge of our own history (e.g. mutation rates can be used as molecular clocks). The aim of this thesis was to obtain a better knowledge of mutation rates within as well as between humans and chimpanzees using data generated by high throughput sequencing. In this chapter, I will discuss some of the results obtained in the previous chapters as well as the limitations of the studies.

The results of the study undertaken in Chapter 3 demonstrate that the estimation of mutation rates from high throughput sequencing data is not an easy task as read error structures are influenced by a multitude of factors, making them subtle but complex. These different types of errors, biases, or uncertainties need to be taken into account as they can strongly influence the downstream analysis of the data. Unfortunately, the sequence readout produced by next-generation sequencers is frequently not well enough calibrated to build accurate statistical models on it. As a consequence, 'black-box' models, assuming that there is no *a priori* information available, are

a valuable alternative. However, extensive validation studies (e.g. for different libraries, NGS machines, and runs) are required to validate and confirm obtained results.

The performance of our model could show that advanced statistical methods can be used to estimate mutation rates in human sperm from high throughput sequencing data down to a level of 10^{-5}, whilst capturing subtle features of the library-, machine-, and run-dependent error structure. However, it should be noted that very few germ line mutations have rates as high as 10^{-5} in humans (the majority of spontaneous mutations occurs at a much lower rates of $10^{-7} - 10^{-9}$ [Arnheim and Calabrese, 2009; Kondrashov, 2003; Nachman and Crowell, 2000]) and thus, the discussed model will unfortunately not be suited for the analysis of most other human germ line mutations.

In order to gain better estimates of mutation rates from high throughput sequencing data, lower error rates as well as better calibration would be needed from NGS platforms. Due to the fact that many of the observed errors are generated during either (i) laboratory-intensive library preparation steps (e.g. during an incomplete enzymatic digestion, required to enrich the sample for specific mutations, or during *in vitro* PCR amplifications, which often use a reengineered DNA polymerase, thereby compromising both efficiency and fidelity of nucleotide incorporations) or (ii) steps involved in the actual sequencing process (e.g. in Illumina sequencing, it is not possible to control the density of the amplified templates on the slide, which can result in high background noise if reads are strongly clustered), third-generation single-molecule sequencing technologies might be a good alternative for future studies. They do not only present the advantage of not relying on *in vitro* amplification, but also dephasing and signal decay, which introduce strong background noise in NGS sequencing, do not pose a problem in single-molecule sequencing, which might lead to lower substitution error rates. On the other hand, as next-next generation sequencing platforms just entered the market, a better understanding of the error structures of the different machines will first need to be established before they can be routinely used. Another possibility would be to enhance the usability of current second-generation sequencers by producing enough oversampling, which would allow researches to deal better with errors and biases in the data. However, as sequencing costs are still at a high price, this scenario might be a too expensive solution for many projects.

In Chapter 4, a map of genome-wide sequence variation in Western chimpanzees, *Pan troglodytes verus*, has been created by collecting and analysing genetic variation data obtained from high throughput Illumina sequencing of 10 individuals (nine females and one male) at an average coverage of 9.45X after mapping to the panTro2.1 reference genome. Calling variants from NGS data is a multi-step procedure, which is strongly influenced by both the quality of the reference sequence as well as the choice of 'appropriate' parameter values and cutoffs in the different steps involved in

the process. The choice of these values is very critical to achieve good results, but, unfortunately, it is usually only based on either visual inspection of the data, or expert knowledge. This makes it not only difficult (if not impossible) to follow strict work flow protocols, it also implies that the results will always (at least partly) depend on the person performing the analyses.

The systematic incompleteness of the rough draft chimpanzee reference genome, with its many gaps and regions of uncertainty, complicated the variant calling process. As a consequence of the imperfect state of the reference genome, all reads were additionally aligned to the finished grade human reference genome in order to ensure robustness of the results by identifying SNPs mapping in places with reliable synteny between the two species. Although this procedure worked reasonably well in the case of Western chimpanzees (at least on the autosomes were it led to a FDR of less than 3%), the application of this strategy to other primate species might be less suitable as a greater evolutionary distance to humans causes the genomes to differ to a higher extent. The limitations of this approach were already visible when trying to use this method to improve the quality of the variant calls on chromosome X: even after the construction of an intersection set between the chimpanzee-based and the human-based mapping, the FDR of chromosome X stayed much higher than the autosomal one (in the order of $\sim$8%) as the chimpanzee reference sequence exhibited even lower coverage on the X (the sequences on which the reference was build mainly originated from a single male individual), restricting its use in population genetic analysis. Unfortunately, the low quality of the polymorphism data set on the X chromosome makes it very difficult to address scientific questions about this sex chromosome (e.g. its evolution from autosomes by studying the PAR, the only region in which sex chromosomes recombine in the heterogametic sex and a very interesting subject of study as the strength of selection to reduce genetic diversity depends on the underlying recombination rate which is known to be about 20-fold higher in the PAR than the genome-wide average in humans [May et al., 2002]). In retrospect, a *de novo* assembly might have been a valuable alternative to the reference-guided approach, not only because it could have overcome some of the above mentioned issues but also because it might have aided several follow-up data analyses. It is also most likely the better strategy in order to call variants from other primates which are less related to humans than chimpanzees.

The data set generated in Chapter 4 was used in Chapter 5 to explore the breadth of the nucleotide diversity in the chimpanzee genome, a study which might help to shed light on whether or not the local variation in mutation rate has been conserved since the divergence of the two species and to place human nucleotide diversity into perspective. Similar to many other species, nucleotide diversity levels in Western chimpanzees were found not to be constant across the genome. This variation might be caused by either evolutionary forces (such as genetic drift or natural

selection) or by regional variation in mutation rates which are influenced by several sequence features. Although all point mutations were controlled by the same factors, which explained >40% of the nucleotide diversity observed in chimpanzees, the performed analysis demonstrated that the relative importance of each of these factors is very different for various types of point mutations. This is not an unexpected finding, given that mutation rates vary strongly between CpG- and non-CpG sites as a result of the different mechanisms that cause mutations at these sites (at CpG-dinucleotides, most mutations are caused by spontaneous methylation-dependent deamination [Chimpanzee Sequencing and Analysis Consortium, 2005; Cooper and Youssoufian, 1988; Coulondre et al., 1978; Duret, 2009; Ehrlich and Wang, 1981; Holliday and Grigg, 1993; Hwang and Green, 2004; Xing et al., 2004], while errors during DNA replication are the main source of mutations at non-CpG-sites [Kim et al., 2006]). As a consequence, it is necessary to consider putative CpG-mutations separately from mutations at non-CpG-sites in population genetic analyses.

At non-CpG-sites, the single best predictor for nucleotide diversity was divergence. Fine-scale divergence rates across the genome of very closely related species are known to be influenced by the occurrence of ancestral polymorphisms. Data from the 1000 Genomes Project was used in combination with the polymorphism data set obtained by the PanMap Project in Chapter 6 to examine the appearance of such ancestral polymorphisms between humans and chimpanzees. Consistent with previously published work [Hodgkinson and Eyre-Walker, 2010; Hodgkinson et al., 2009; Johnson and Hellmann, 2011], by considering SNPs in humans that had a SNP in the orthologous position in chimpanzees, this study observed an excess of shared polymorphisms which could not be easily explained by fine-scale sequence-context captured by neighbouring nucleotides. This excess could be the result of an underlying interesting biology or it could arise from less interesting technical artefacts caused by systematic errors in the variant calling process such as shared mismapping artefacts due to problems with the read alignment (e.g. in repetitive regions of the genome), repeated sequencing errors, ascertainment bias, or shared false positive SNP calls (e.g. due to an incorrect genome assembly of duplications which occurred before the divergence of the species onto the same locus). Although, the characteristics of the shared variants were inconsistent with substantial problems in the variant calling process, the possibility of a contribution of shared variants from substitutions in paralogous sequences that occurred before the human-chimpanzee split was not excluded and thus, it will require further investigation. An analysis of the local 10bp sequence composition of shared variants suggested that the observed heterogenerity in mutation rate seems to mainly be the consequence of either mutations at methylated CpG-sites or of strand-slippage mispairing at sites of low sequence complexity. In general, it was difficult to distinguish polymorphisms which were shared through ancestry from those generated by systematic artefacts or

recurrent mutations at individual sites of the genome, thus, this work focused on short haplotypes in which several shared ancestral polymorphisms could be identified, and which exhibited the same pattern of linkage disequilibrium in both species, a scenario that is unlikely to have occurred by chance suggesting that selection acts on them. More than 20 different regions have been identified that fulfilled this criteria. These regions (including the MHC but also other regions of the genome) are good candidates for regions under balancing selection. This result suggests that long-term balancing selection may be more common than previously believed. Nevertheless, to rule out events of deep coalescence, these regions will need to be validated, for instance by studying the patterns of variation in other closely related primates (such as gorilla or bonobo).

In conclusion, the generated data set of genome-wide sequence variation in Western chimpanzees has provided some first insights into patterns of mutation rates and nucleotide diversity of our closest evolutionary relative, and it offers the great potential to address many more. This thesis has demonstrated that, although high throughput sequencing data is a very valuable resource for this research, it needs to be handled with care as its analysis is not straightforward. However, as further data of humans and other primates becomes available, we are presented with the amazing opportunity to gain an even better understanding of the nature of mutation and variation in genomes.

Chapter 8

Outlook

After discussing the results of this thesis in the previous chapter, in this chapter, I take the opportunity to suggest some future directions of the work.

Patterns of nucleotide variation show a high heterogeneity across the genome in many different organisms, including the chimpanzee, which is in part caused by genealogical history as well as regional differences in mutation rates, recombination rates, and other genome features. Nevertheless, some of this variation might reflect patterns of natural selection. To address fundamental questions about the extent and nature of adaptive evolution in the chimpanzee and human lineage, not only broad-scale patterns of nucleotide diversity should be analysed, but the study should be extended to specific genes of interest (e.g. genes involved in the malarial resistance in chimpanzees), which might show evidence for different forms of natural selection. In order to assess the role of selection on different variation patterns, the effect of three models on the relationship between diversity and recombination should be examined: (i) background selection, (ii) recurrent selective sweeps, and (iii) a combination of (i) and (ii). Different types of tests are available to detect selection using genetic data including tests based (i) on species divergence data, (ii) on population data, as well as (iii) tests based on both population and divergence data. While the first group can identify genes (and regions within genes) that have been under natural selection since species divergence, statistical tests based on population data can examine the nature of genetic variation within a single species. The latter are an important tool to gain an understanding of evolutionary adaptation at the molecular level as they can provide information about genomic regions which are or have been influenced by positive selection. With the availability of a map of genome-wide sequence variation in Western chimpanzees from the PanMap Project (see Chapter 4), tests based on population data can now be performed to systematically search for regions in the chimpanzee genome that are under recent selection. One possibility to detect regions under selection in genomes is the usage of statistical tests that identify regions in which frequencies are skewed compared to the neutral model as

they are usually suggestive of non-neutral evolution. These tests include (i) composite likelihood tests [Kim and Stephan, 2002; Nielsen et al., 2005; Williamson et al., 2007] which compare the likelihood of a neutral model to a selective sweep model for a given genomic region whilst taking the background pattern of variation into account, (ii) Tajima's D which detects regions where the mutation frequency spectrum is biased to either rare or common variations [Tajima, 1989], and (iii) statistical tests which are related to Tajima's D (e.g. Fu and Li's $D^\star$ or $F^\star$ [Fu and Li, 1993], or Fay and Wu's H [Fay and Wu, 2000]). Although these tests allow to scan for selected genes throughout the entire genome, they are not as sensitive as other methods for recent or very ancient selection events, and they are dependent on the underlying demographic structure. Thus, other statistics which are more sensitive for recent episodes of selection, such as haplotype-based tests (e.g. the extended haplotype homozygosity [Sabeti et al., 2002] or the integrated haplotype score test [Voight et al., 2006]) based on the idea that non-selected haplotypes vary in length whereas incomplete selective sweeps lead to haplotypes at high frequency with high homozygosity extending over large regions (further than it would be expected under the neutral model) because insufficient time has elapsed for recombination to reduce their length, should also be executed. The availability of population genetic data for the chimpanzee also enables the application of tests which are based on both population and divergence data (such as HKA test [Hudson et al., 1987], or the MK test [McDonald and Kreitman, 1991]). Their main advantage is that they can be applied to a very wide range of time scales.

However, although data of genome-wide sequence variation in Western chimpanzees enables the search for regions under selection in the chimpanzee genome, it should be kept in mind that (i) the sample size of the study was small (10 individuals), and that (ii) only Western chimpanzees were studied, both of which might limit its use. As a consequence, larger sample sizes including chimpanzee individuals from different subspecies will be required for more reliable analyses.

An additional interesting study would be to address the effect that large-scale chromosomal rearrangements might have had on the speciation of the two lineages [Rieseberg, 2001]. Although karyotypes have stayed relatively constant during the primate evolution, 10 large-scale euchromatic rearrangements have been detected between human and chimpanzee karyotypes in previous cytogenetic studies: nine pericentric inversions, and the telomere fusion of two ancestral chromosomes in human that remained separate in the chimpanzee lineage (two chromosomes 2a and 2b in chimpanzee, one chromosome 2 in human) [Fan et al., 2002a,b; Yunis and Prakash, 1982; Yunis et al., 1980]. A model by Navarro and Barton proposed that the evolution of alleles that were compatible within but incompatible between species has been accelerated by the blocked gene flow in rearranged regions of the chromosomes (such as large-scale chromosomal inversions)

[Navarro and Barton, 2003a]. Support of this model came from an analysis of a small set of ∼100 human and chimpanzee genes which showed elevated ratios of non-synonymous to synonymous amino acid changes in rearranged chromosomes compared to the ratios observed in non-rearranged chromosomes, and thus hinting towards a stronger influence of positive selection on rearranged chromosomes [Navarro and Barton, 2003b]. However, the set of genes used for this analysis was small and most likely biased, and contradictory results have been found in a subsequent analysis of chimpanzee ESTs [Zhang et al., 2004]. Thus, the role of euchromatic rearrangements on the primate speciation remains to be unraveled, and the PanMap data set could be used to readdress this issue.

Further extension of this work could focus on the examination of patterns of diversity in regions that are of particular evolutionary interest such as the region of chromosome 2 in humans which has evolved by fusion of the two short arms of its predecessor chromosomes 2a and 2b. Both chromosomes 2a and 2b can still be found in the chimpanzee lineage, making it possible to explore changes in the underlying patterns of diversity as well as recombination rates between the two species that might have been caused by the fusion event.

Appendix A

Sequencing Technologies

A.1 First-generation Sequencers

Fifteen years after the discovery of the double helix in 1953, first methods to sequence DNA became available but modern DNA sequencing only began in 1977 when Sanger, Nicklen and Coulson, as well as Maxam and Gilbert independently published new approaches to sequence DNA. Even if both techniques were very similar, using gel electrophoresis to separate DNA fragments with single base pair resolution [Maxam and Gilbert, 1977; Sanger et al., 1977], the Sanger approach became the method of choice over Maxim and Gilbert's chemical sequencing technique for the next several decades [Hunkapiller et al., 1991]. The main reasons were the development of automated sequencing platforms in the mid-1980s as well as technologies that reduced the costs while improving the sequencing throughput at the same time: efficient library construction and template preparation, dideoxynucleotides with fluorescent moieties [Prober et al., 1987], and thermostable polymerase engineered to accept them [Tabor and Richardson, 1995], as well as implementations of efficient DNA sequence production work flows [Shendure and Ji, 2008]. Since Sanger et al. described for the first time the process of dideoxynucleotide DNA sequencing [Sanger et al., 1977], the method has undergone many changes that made it suitable for large-scale productions. Nowadays, specialised infrastructures including robotics, large computer databases, and instrumentation, as well as bioinformatics to cope with the huge amount of data, are required for high throughput DNA sequencing. Therefore, only a handful of large genome centres worldwide (amongst them the Wellcome Trust Sanger Institute, the Broad Institute, and the Washington University Genome Sequencing Center have both the resources and the expertise to carry out experiments on the scale of whole mammalian-sized genome sequencing, large-scale sequencing of multiple organisms, or resequencing of a large number of genes [Rogers and Venter, 2005].

Work Flow Sanger's sequencing method starts by amplifying the sequence fragment of interest either by cloning it into a known high-copy-number plasmid (using restriction and ligation processes), which is used to transform a bacterial strain (such as *Escherichia coli*) which amplifies the sequence fragment of interest *in vivo* (for shotgun *de novo* sequencing), or by using it in a polymerase chain reaction carried out with primers flanking the target (for targeted resequencing). After sufficient quantities of the template are generated, 2'-deoxynucleotide triphosphates (dNTPs, the standard building blocks of DNA) as well as chain-terminating 2',3'-dideoxynucleotide triphosphates (ddNTPs, chemically modified nucleotides that miss a hydroxyl group at the third carbon atom of the sugar molecule) are added. In any given template molecule, the strand elongation begins at the 3'-end of the primer, which is complementary to a known sequence directly flanking the region of interest, and ends as soon as a ddNTP is incorporated. The relative probability of incorporating a ddNTP during the primer extension is determined by the specific ratio of dNTPs and ddNTPs. In older protocols, four separate primer extension reactions are carried out, each containing template, polymerase, a radioactively labelled primer, dNTPs, and one out of the four ddNTP types (ddATP, ddGTP, ddCTP, ddTTP). As a result, many terminated fragments with only a subset of different lengths (corresponding to the positions of this particular nucleotide in the template sequence) are generated in each reaction. In a next step, the four reactions are electrophoresed in different lanes using a polyacrylamide gel resulting in molecules that are sorted by their weight. As a consequence, this procedure ensures a separation with a single nucleotide resolution allowing for the interpretation of the primary sequence template from the observed bands as each band consists of terminated fragments of a single length. In contrast, current protocols perform a thermal 'cycle sequencing' reaction in which multiple cycles of denaturation, primer annealing, and primer extension are executed to increase the number of terminating strands whilst minimising the amount of template DNA [Carothers et al., 1989; Craxton, 1993; Krishnan et al., 1991; Murray, 1989; Sears et al., 1992; Slatko, 1996]. Only a single primer extension reaction, including all four ddNTPs at the same time, is performed, which requires a specifically designed, thermostable polymerase, such as 'ThermoSequenase', which allows for an efficient incorporation of ddNTPs [Tabor and Richardson, 1995]. ddNTPs are fluorescently labelled using dyes (resonant energy transfer, or ET dyes [Ju et al., 1995]) with the same wavelength but distinct strong emission spectra allowing the distinction of the incorporated ddNTPs by fluorescent energy resonance transfer (FRET) [Ju et al., 1996; Lee et al., 1997]. Reaction results are analysed using an electrophoresis, which is carried out in a long capillary filled with a denaturating polymer within an automated

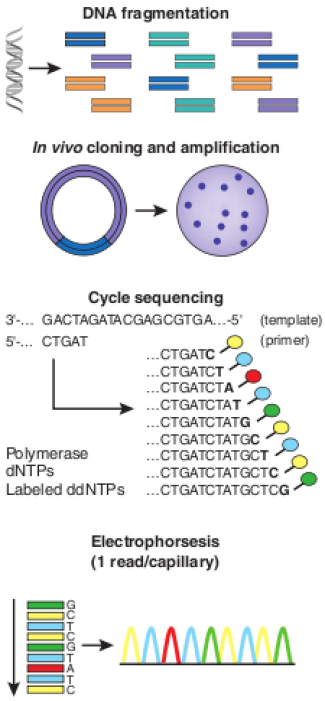


Figure A.1: Work flow of conventional Sanger sequencing. The first step in the high throughput shotgun Sanger sequencing is the fragmentation of the DNA of interest. The fragments are cloned to a plasmid vector which is used to transform *E. coli*. In the following cycle sequencing reaction, fluorescently labelled ddNTPs are used to identify incorporated nucleotides and to terminate the elongation of the DNA strand. As these fragments of discrete size pass a detector, the four-channel emission spectrum is used to generate a trace of the sequence. Reprinted by permission from Macmillan Publishers Ltd: Shendure and Ji [2008], Nature Biotechnology (`http://www.nature.com/nbt`).

sequencing instrument, which typically analyses 96 to 384 independent sequencing reactions simultaneously via an array of capillaries [Emrich et al., 2002; George et al., 1997; Shendure and Ji, 2008; Shibata et al., 2000; Swerdlow and Gesteland, 1990]. FRET is used to identify the terminal base of a fragment as it passes through a transparent component near the end of the capillary where a single wavelength of light excites the fluorophore linked to the ddNTP. Measurements of the emission spectra produce a four-colour sequencing trace whose peak heights are analysed by specifically designed algorithms, so-called 'base callers'. They reveal the DNA sequence, and define the accuracy with which individual base calls have been made (often using the well-established 'Phred' score [Ewing and Green, 1998; Ewing et al., 1998]). An overview of the procedure is depicted in Figure A.1.

Characteristics Current instruments, such as the Applied Biosystems 3xxx series or the GE Healthcare MegaBACE instrument, capable of producing reads with an average length of 600-1,000bp [Hert et al., 2008; Shendure et al., 2008] have a throughput of about 6Mb per day [Kircher and Kelso, 2010]. Per-base raw accuracies are as high as 99.999% [Shendure and Ji, 2008] due to DNA proof-reading and DNA repair mechanisms in the bacterial host organism during the sequence amplification. Apart from the amplification step, sequence

errors can arise from contaminated samples and from polymerase slippage at low complexity regions of the genome such as homopolymer runs (sequence fragments of the same nucleotide) or simple repeat sequences. In general, errors increase towards the end of the read as lower intensities, missing termination variants, and deletions accumulate, and the separation of the electrophorese products becomes less accurate [Kircher and Kelso, 2010]. The refinements and automations of the Sanger technology enabled a massive drop in costs from US$10 to read a single base in 1985 to 10,000 bases 20 years later for the same amount of money to currently about US$500 per 1Mb of DNA sequence (prices per sequencing read strongly vary between commercial facilities and in-house high throughput sequencing centres) [Kircher and Kelso, 2010; Pettersson et al., 2009; Shendure and Ji, 2008; Shendure et al., 2005]. However, compared to NGS methods, Sanger sequencing is still expensive. The high price is mainly caused by the time consuming and labour-intensive sample preparation process involving shotgun cloning of specific DNA fragments into bacteria (a step that can be particularly challenging for certain base compositions and lengths of the sequence fragment, and that is further complicated by possible interactions with the host organism [Kircher and Kelso, 2010]), their distribution onto tens of thousands of plates, the colony picking and their transfer to well plates followed by the amplification of individual clones and the cleanup of the templates [Rothberg and Leamon, 2008]. Nevertheless, Sanger sequencing has the potential to become cheaper by using microfluidic platforms which automate DNA extraction from a smaller amount of DNA [Blazej et al., 2007] as well as *in vitro* amplification and which integrate the individual steps of the sequencing process using a lower reagent volume at the same time [Blazej et al., 2006; Emrich et al., 2002; Hert et al., 2008; Mariella, 2008; Roper et al., 2007]. In addition, analysis times can be shortened by parallelising sequencing separations [Fredlake et al., 2008; Paegel et al., 2002].

A.2 Second-generation Sequencers

Current NGS technologies make it possible to sequence large amounts of data within a few days, shortening the required time by a factor of 100-1,000 based on a daily throughput compared to first-generation sequencers, to substantially lower costs (0.1-4% the costs of traditional Sanger sequencing [Kircher and Kelso, 2010]). These methods have many features in common which distinguish them from the Sanger sequencing method: (i) they overcome cloning biases (present in the conventional vector-based cloning and *Escherichia coli*-based amplification stages used in Sanger sequencing) by making use of DNA libraries consisting of short sequence fragments to which

platform-specific adapters are ligated [Mardis, 2008b]. These adapters are used to immobilise the fragments to a solid surface (usually either a bead or a glass slide) where they are amplified by a polymerase chain reaction. This step results in the presence of clusters (depending on the used technique either in an ordered fashion or randomly dispersed) consisting of multiple copies of a unknown DNA sequence of interest flanked by universal adapter sequences. (ii) A sequencing-by-synthesis (SBS) principle allows millions of sequence reads to be processed in parallel using a single reagent volume (rather than 96 at a time as in conventional capillary-based sequencing), resulting in low overall sequencing costs. For this purpose, all spatially separated clusters are simultaneously decoded by applying a single reagent volume to the array and a cyclic enzymatic process interrogates the identity of one base position at the time for all clusters in parallel. This process is coupled either to the production of light or the incorporation of a fluorescent group which can be detected using a charge-coupled device (CCD). Thus, after multiple cycles, the continuous sequence from each cluster can be obtained from analysing the full series of imaging data. Depending on the platform, sequence information can be obtained from one end of the fragment, or from both ends of either a linear fragment (paired end sequencing; paired end reads are usually separated by ∼300-500bp) or a previously circular fragment (mate pair sequencing; mate pairs are usually separated by ∼1.5-20kb) [Mardis, 2011]. Paired end sequencing libraries require only a small amount of genomic sequence, and are easy to prepare, making them well suited for the sequencing of human-sized genomes. In contrast, mate pair sequencing libraries require larger amounts of input DNA and are more difficult to prepare due to the low yield of circular large DNA molecules, but they provide better information about large-scale structural events [Mardis, 2011]. However, both approaches aid the assembly or the alignment of short read sequences to a reference genome using the provided information about the relative position and orientation of a pair of reads in the genome [Chaisson et al., 2009]. (iii) The *in vitro* amplification step is essential to ensure that the sequencing reaction produces a sufficiently strong signal that can be detected by the platform's optical system, but it is also a source of sequencing error as the used polymerases are never 100% accurate (both errors in the template sequence and amplification biases can be introduced) [Mardis, 2011]. Another source of error is the loss of synchronisation (also referred to as 'dephasing') during the sequencing process as the addition of nucleotides is usually not perfect. This problem becomes more apparent the longer the read extends, which limits the read length produced with NGS platforms to much smaller sizes than those achieved in traditional capillary-based sequencing.

Although the underlying process is similar in all platforms, the concrete protocols differ remarkably, resulting in specific biases and limitations, and the details of a given application deter-

Criteria	Technology				
	454 GS FLX+[1]	Illumina GAII[2]	SOLiD 5500xl[3]	Polonator[4]	HeliScope[5]
Amplification	emPCR (28 μm beads)	Bridge PCR	emPCR (1 μm beads)	emPCR	None (single molecule)
Sequencing chemistry	Polymerase	Polymerase	Ligase	Ligase	Polymerase
Read length (bp)	up to 1,000	up to 150	up to 75	up to 26	25-55
Dominant error	Indels	Substitutions	Substitutions	Substitutions	Deletions
Throughput/day (Gb)	$\sim$ 0.75	up to 2.7	10-15	2-2.5	2.5-4
Time/run (days)	$\sim$ 1	up to 14	up to 7	$\sim$ 4	$\sim$ 8

[1] Roche454 [2011], [2] Illumina [2010], [3] Applied Biosystems [2011], [4] `http://polonator.org/`, [5] `http://helicosbio.com/`

Table A.1: Overview of the distinctive features of the commercially available next-generation sequencers.

mine which platform is best suited for a particular scientific research question. For example, *de novo* sequencing and metagenomics prefer longer read lengths, while for applications that rely on 'tag counting' (such as gene expression quantification, chromatin immunoprecipitation sequencing, methylation analysis, or promotor binding site studies), the number of reads is much more important than their actual length [Kahvejian et al., 2008; Rothberg and Leamon, 2008]. Apart from the throughput and the produced read lengths, the choice of an appropriate sequencing strategy often depends on the raw/consensus accuracies, the possibility to generate mate pair or paired end reads, as well as the cost per raw base, per consensus base, and per read.

A brief description of the methods underlying the three most widely used NGS platforms, the Genome Sequencer from 454 Life Science/Roche, the Genome Analyzer from Illumina, and SOLiD from Applied Biosystems/Life Technologies, as well as Ion Torrent's PGM which has just entered the market, is given in the Sections A.2.1-A.2.3 and in Section 2.1, together with an overview of their main (dis-)advantages (for an overview of their main characteristics see Table A.1). However, it should be kept in mind that all NGS methods are continuously improved. Developments of robust protocols to generate sequencing libraries, fundamental innovations in the experimental design, such as nucleotide chemistry, polymerase activity, fluorescence and its detection, surface chemistry, and image analysis, as well as in the base-calling process, are likely to further improve them in terms of read length, sequence quality, read mapping, total yield, and faster data collection as it has been the case for the Sanger technology.

A.2.1 454 Life Science/Roche: Genome Sequencer FLX

In 2005, the first commercial next-generation sequencer was introduced by 454 Life Science [32], a member of the Roche Diagnostics group [Bentley, 2006; Droege and Hill, 2008; Margulies et al., 2005]. It was the first non-Sanger sequencing technology (relying on a SBS principle using a pyrosequencing protocol) used to sequence an individual human genome [Wheeler et al., 2008a].

[32] `http://www.454.com`

Work Flow The 454 method starts with the construction of DNA libraries using nebulisation, a mechanical process to sheer DNA in order to produce short sequence fragments. These fragments are end-repaired and their 5'-end as well as their 3'-end are ligated to two different specialised common oligonucleotide adapters [Droege and Hill, 2008]. The libraries are diluted to a single molecule concentration, denatured, and hybridised to the surface of hundreds of thousands of 28μm agarose beads, which have millions of oligomers complementary to one of the two adaptor sequences of the fragments attached to them. The DNA strands are amplified on these beads using an emulsion PCR (emPCR) based on an oil and aqueous mixture [Dressman et al., 2003]. This mixture is used to capture individual beads into separate aqueous emulsion droplets containing the PCR reactants which are separated by oil into millions of stable reaction chambers of uniform size [Pettersson et al., 2009; Tawfik and Griffiths, 1998]. These chambers simultaneously amplify a library consisting of unknown DNA fragments flanked by universal adapters with a single pair of primers. As the beads carry one of the two PCR primers on their surface (5'-attached), they are able to capture the PCR amplicons that were generated within an individual chamber resulting in clonally amplified templates [Dressman et al., 2003]. Thereby, each chamber only contains multiple PCR amplicons obtained from a single fragment within the library. In this way, more than a million beads, each comprising up to a million copies of the original annealed DNA, can be produced within a few hours [Margulies et al., 2005].

To enable the detection of incorporated nucleotides, several hundred thousands of these beads are added to the surface of a 454 picotiter plate (PTP) consisting of approximately 1.6 million single wells (with dimensions such that exactly one bead fits per well) in the tips of fused fibre optic strands [Leamon et al., 2003]. In a next step, smaller, magnetic and latex beads of 1μm diameter are added to the PTP. These beads are attached to the active enzymes ATP sulfurylase and luciferase, both of which are required for the pyrosequencing reaction. After the PTP is placed in the sequencer, one side of the semi-ordered array acts as a flow cell to sequentially introduce (and remove) a single dNTP species as well as all required reagent solutions at each cycle. A polymerase extends the primed DNA strand if the introduced nucleotide is complementary to the template strand; a process that only terminates if all complementary nucleotides have been synthesised. The incorporation of a new nucleotide by a polymerase results in the release of a diphosphate group (PPi), which is converted to ATP by ATP sulfurylase. Together with D-luciferin and oxygen, the newly formed ATP is used by luciferase to emit light [Pettersson et al., 2009], which can be detected by a high resolution CCD camera on the other side of the array [Shendure and Ji, 2008]. Up to the level of detector

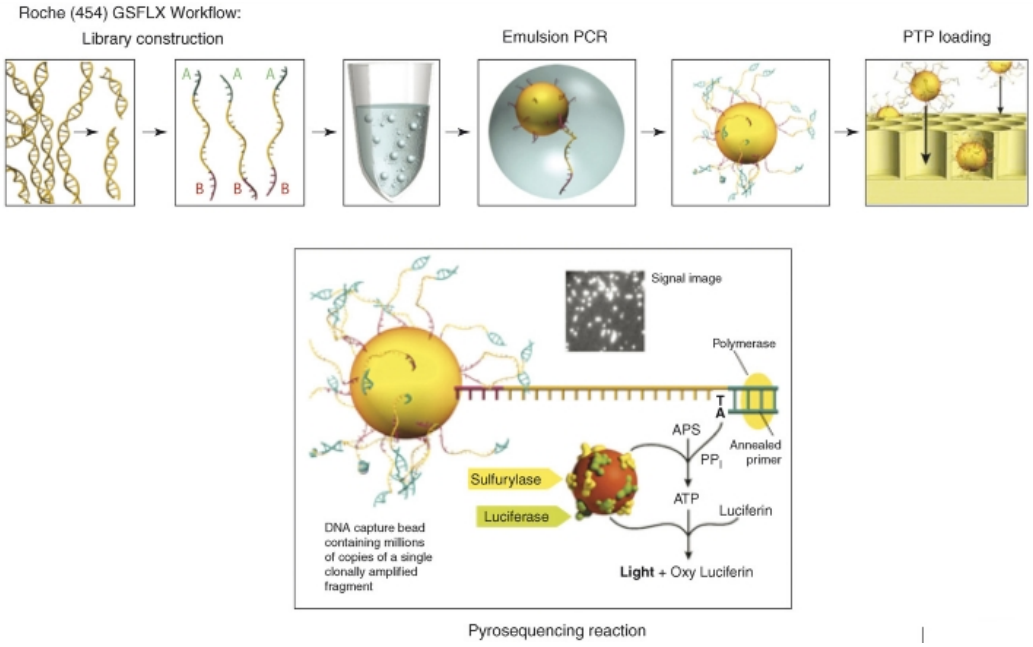


Figure A.2: 454 work flow. Top: The library is constructed by fragmentation of the sample (such as genomic DNA) into small, 300-800bp fragments and ligating adapters (specific for both the 3' and 5' ends) to them which are used for purification, amplification, and sequencing steps. The fragments are immobilised onto beads in a way such that each bead carries a unique single stranded DNA library fragment. Emulsion PCR is used to amplify them in parallel within individual microreactors containing just one bead and thus, excluding competing or contaminating sequences. This process results in several million identical DNA fragments per bead. In a next step, the emulsion PCR is broken while the amplified fragments remain bound to their specific beads which are loaded into a PTP for sequencing. Bottom: The pyrosequencing reaction occurring on nucleotide incorporation to report sequencing by synthesis: after addition of sequencing enzymes, individual nucleotides flow in a fixed order across the hundreds of thousands of wells of the PTP containing one bead each. Addition of one (or more) nucleotide(s) complementary to the template strand results in a chemiluminescent signal recorded by the CCD camera of the platform. Reprinted from Mardis [2008b], Copyright (2008), with permission from Elsevier.

saturation, the emitted light is directly proportional to the amount of pyrophosphate released and thus, it can be used to gain information about the incorporated nucleotide. After each cycle, apyrase is used to degrade unincorporated dNTPs and to stop the reaction by degrading ATP. To ensure that there are no nucleotides remaining in a well, the plate is washed before the next dNTP species is introduced [Ronaghi et al., 1998]. Standard dNTPs are used, with the exception of dATP as it is a substrate of luciferase which could be directly used in the light-emitting reaction. Instead, deoxy-adenosine-5'-(α-thio)-triphosphate is utilised.

The software that detects the level of emitted light per nucleotide incorporation is usually calibrated using a defined single nucleotide pattern in the adaptor sequence which matches the sequence of the first four nucleotide flows (TCGA). A metric, similar to the well-established Phred score used by Sanger sequencers [Ewing et al., 1998], is used to estimate the *ab initio* quality of each base of a read [Margulies et al., 2005]. Quality filters are implemented to remove poor-quality sequences, sequences with more than one initial DNA fragment per bead, and sequences without the initial TCGA sequence [Mardis, 2008a]. 454's work flow is depicted in Figure A.2.

Characteristics The 454 technology is able to produce more than a million reads with lengths of (i) 250bp in the initial GS FLX system, (ii) 450bp in the GS FLX Titanium XLR70 platform (which reduces noise between adjacent wells through a titanium-coated PTP design), and (iii) 700bp in the GS FLX Titanium XL+ sequencer (by far the longest of any NGS platform) with a throughput of 750Mb per day [Pettersson et al., 2009; Roche454, 2011; Voelkerding et al., 2009]. Current read lengths are limited by a decreased efficiency of the enzymes polymerase, luciferase, and apyrase, (or in some cases, the complete loss of enzymes for certain molecules) after a large number of cycles. This can lead to beads that are out of phase: molecules lose synchronism by either falling behind (a negative frameshift, occurring at a frequency of ∼0.1-0.3%, caused by a not 100% complete nucleotide incorporation in a cycle), or by getting ahead of all other templates in the sequence (a positive frameshift, occurring at a frequency of ∼1-2%, caused by nucleotides that are not entirely degraded by apyrase and thus, they can be incorporated after the next nucleotide), which causes enhanced error rates towards the end of the reads as the level of noise increases [Pettersson et al., 2009]. Whereas the sequential flow of nucleotides makes the occurrence of substitution errors rare (in the range of 10^{-3}-10^{-4} [Margulies et al., 2005]), small indel errors can occur if response linearity exceeds the sensitivity of the detector. The latter scenario is especially problematic for homopolymers as it is difficult to interpret the precise number of consecutive incorporations in a single cycle from the signal intensity - a problem that becomes more severe, the longer the homopolymer run - usually resulting in higher error rates for homopolymers [Huse et al., 2007; Margulies et al., 2005]. Further errors can be introduced by the PCR during the *in vitro* amplification (a process that can also add a bias in template representation), by the bead preparation (e.g. the occurrence of 'mixed' beads carrying copies of multiple different sequences which do not reflect real molecules), and by 'ghost' wells (empty wells for which a signal is recorded because of a very strong signal from the neighbouring wells) [Kircher and Kelso, 2010]. Most (but not all) of these issues can be resolved through enough oversampling (although homopolymer runs longer than 10 nucleotides are often miscalled regardless of the coverage [Green et al., 2008; Quinlan et al., 2008; Wicker et al., 2006]), and some problematic reads can be automatically filtered out during the data post-processing [Green et al., 2006; Pettersson et al., 2009].

The major drawbacks of 454 sequencing relative to other NGS technologies are the cumbersome emPCR, the lower throughput, and the higher cost mainly caused by the requirement of rigorous sample washing between nucleotide additions, as well as the requirement of the addition of new enzymes in each cycle [Hert et al., 2008]. Nevertheless, 454 seems to be the

method of choice for scientific projects where long reads are critical, such as in *de novo* whole-genome sequencing, metagenomic approaches, or mapping in repetitive regions [Shendure and Ji, 2008].

A.2.2 Applied Biosystems/Life Technologies: SOLiD

The third NGS platform to enter the market was SOLiD [33] (Sequencing by Oligo Ligation and Detection) introduced by Life Technologies/Applied Biosystems (ABI) in 2007. Its approach is based on work which was carried out at the Harvard Medical School and the Howard Hughes Medical Institute, and which was published in 2005 [Shendure et al., 2005].

Work Flow Different standard approaches can be used to generate single stranded, oligo-adaptor-linked DNA fragments, the starting points of the SOLiD sequencing method (see Sections A.2.1, and 2.1.1). Alternatively, protocols have been developed allowing the construction of mate paired tag libraries with controllable, highly flexible distance distributions between the fragments [Ng et al., 2005; Shendure and Ji, 2008; Shendure et al., 2005]. Generated fragments are attached to 1μm paramagnetic beads containing complementary oligomers on their surface [Dressman et al., 2003] and amplified using emPCR. After amplification, the emulsion is broken and the beads bearing amplification products are randomly dispersed to a solid surface of a specially prepared glass slip as a highly densed, disordered array. Beads are immobilised on the surface either with an acrylamide gel or by covalent attachment. The glass slide is placed into a sequencing machine which can process two of these slides at the same time: one slide receives the sequencing reactants while the other one is imaged.

The sequencing process (Figure A.3) starts with the hybridisation of a universal sequencing primer to the adapters of the library fragments. DNA ligase is added together with a limited set of semi-degenerate, fluorescently labelled octamer oligonucleotides which compete for ligation to the primer at its 3'-end. If an octamer hybridises with a matching fragment adjacent to the universal primer, the phosphate backbone is sealed by the ligase. In a next step, the sequence information is collected for the first two bases on the 3'-end of the octamer. For this purpose, a fixed base of the octamer (the base at either the second or the fifth position depending on the cycle number) is correlated with the fluorophore making its identification possible through a four-colour fluorescent readout using a modified epifluorescence microscope. In this way, data can be efficiently collected for the same base position across all template-bearing beads. However, as there are 16 possible dinucleotides and SOLiD uses

[33] `http://solid.appliedbiosystems.com`

only four colours, 2 base encoding (described below) is necessary to identify each base. To enable a further round of ligation, the sixth, seventh, and eight base of the ligated octamer, as well as the fluorescence group are removed in a cleavage step, leaving a free 5'-phosphate on the by five nucleotides extended primer, which can be used in another sequencing-by-ligation cycle. As a result, each fragment sequence is identified in five nucleotide intervals (for instance bases 5, 10, 15, 20). After a certain number of cycles, synthesised fragments are melted and washed away to reset the system. The next sequencing round targets a different set of positions and thus, it is initiated by a hybridisation of a primer which is set back one base from the adaptor-insert junction (position n-1) or by using mixtures of octamer oligonucleotides where another position is labelled. In order to facilitate the readout of each nucleotide in the sequence, the whole process is repeated several times for different primers.

2 Base Encoding 2 base encoding relies on the fact that for each primer multiple successive ligation cycles in blocks of five nucleotides are carried out in which the first two bases of the primer are interrogated. As there are 16 possible dinucleotides and only four colours, one dye represents four different dinucleotides as shown in the encoding scheme in Figure A.4. The encoding scheme has a number of unique properties: every pair of dinucleotides is represented in the same colour (for instance AC and CA are both represented by a green colour) as well as the complement of these dinucleotides (GT and TG). In contrast to single base encoded sequencing methods (such as Sanger sequencing), the bead colour does not directly refer to the incorporated base. For this reason, colours are recorded over consecutive cycles for each bead allowing each position to get probed twice using the four-dye encoding scheme. The incorporated nucleotide can be determined by analysing the colour that results from the two successive ligation reactions. As the bases are read out of order, polymerase-introduced errors are eliminated. In addition, 2 base encoding enables an additional quality check of the read accuracy: sequencing errors can be distinguished from true polymorphisms during the data analysis step as the first would only be detected in one particular ligation reaction whilst the latter would be detected in both.

Characteristics One of the main advantages of SOLiD is the availability of a high density array of very small beads. Although more than one billion sequencing features could potentially fit on the surface of a standard microscope slide [Shendure and Ji, 2008], current systems are limited to about 300 million beads that can be sequenced in parallel as the bead's signal clarity suffers from other signals overbleeding from beads in close proximity [Kircher and Kelso, 2010]. Similar to the sequencing technologies discussed in the previous sections, *in*

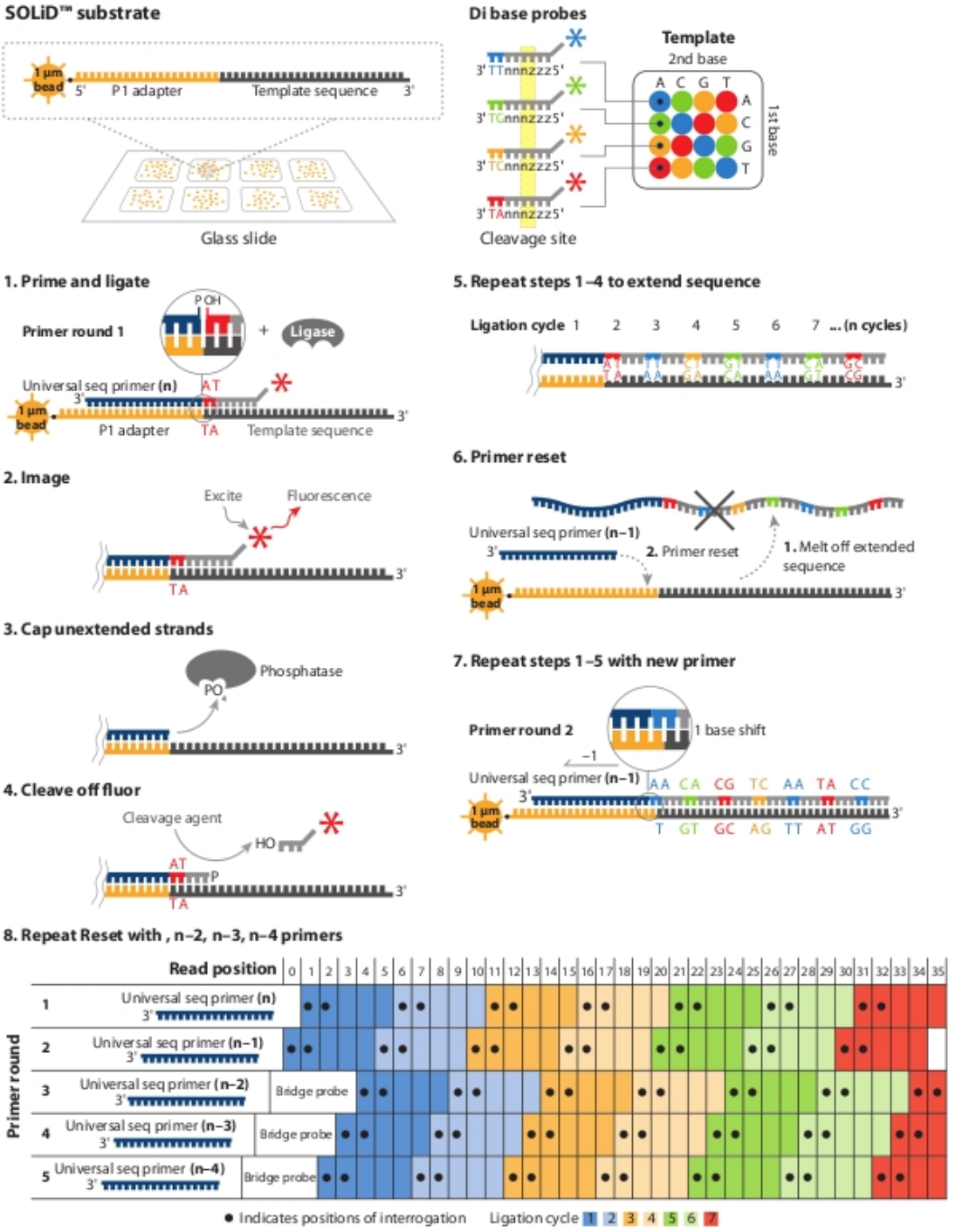


Figure A.3: SOLiD work flow. DNA fragments are amplified on the surface of beads before they are deposited onto a flow cell slide. A primer is annealed to the shared adaptor sequences on each amplified fragment. The DNA is then sequenced using DNA ligase along with octamers specifically fluorescently labelled on the fourth and fifth position. A fluorescent detection follows each ligation step, after which bases as well as the fluorescent group of the octamer are removed to enable another round of ligation. Republished with permission of Annual Review of Genomics and Human Genetics, from Mardis [2008a]; permission conveyed through Copyright Clearance Center, Inc.

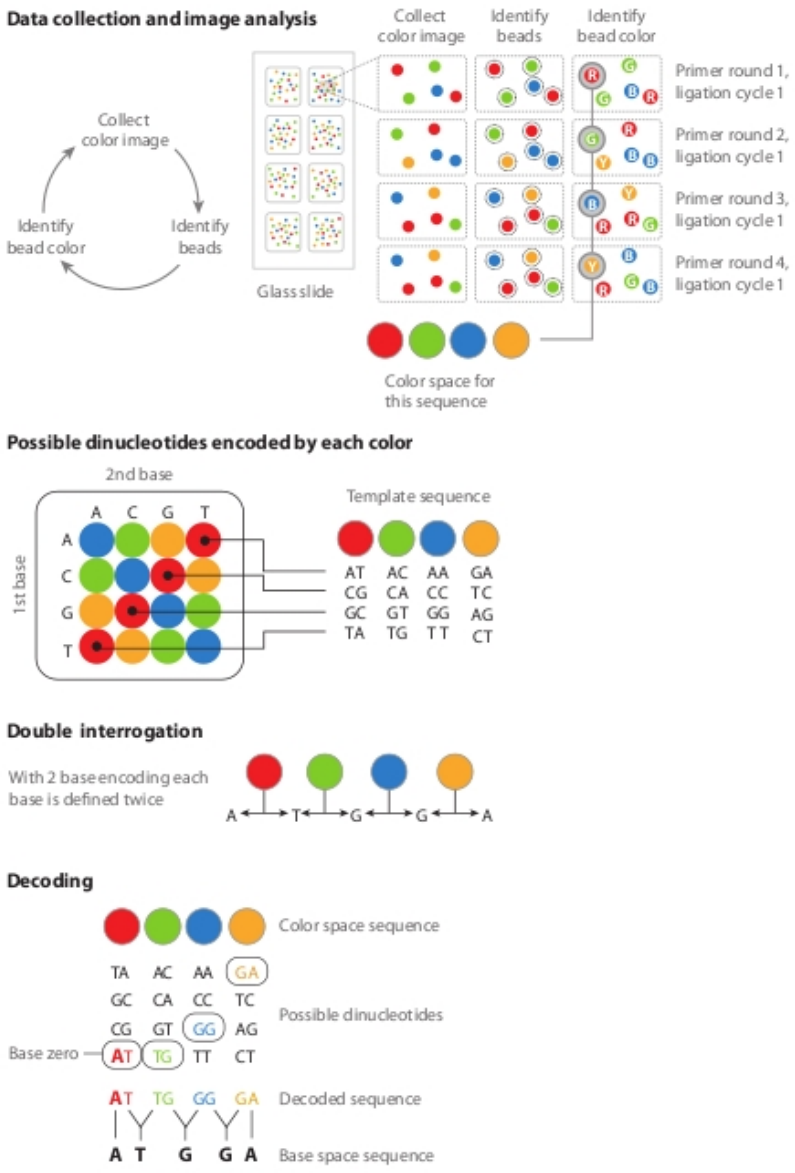


Figure A.4: Principle of the 2 base encoding. True polymorphisms can be distinguished from sequencing errors due to the fact that each fluorescent group identifies a two-base combination and sequence reads can be aligned to a known high quality reference sequence. Republished with permission of Annual Review of Genomics and Human Genetics, from Mardis [2008a]; permission conveyed through Copyright Clearance Center, Inc.

vitro amplification is a potential source of error, as are 'mixed' beads which result in false, artificial readouts. Other potential causes of error are a phasing-induced background noise due to an incomplete cleaving of the fluorophore which might be removed at the next cycle allowing further extension and a decay of the signal due to the random nature of hybridisation which causes a smaller number of molecules to participate in each subsequent ligation cycle [Kircher and Kelso, 2010]. Nevertheless, average error rates are generally low (in the range of 10^{-3}-10^{-4}) [Kircher and Kelso, 2010].

Beside the cumbersome and technically difficult emPCR procedure and the long run times (a run takes up to 7 days), the very short read length of 35-75bp per clonally amplified bead is the major disadvantage of the SOLiD technology, circumventing its usage in *de novo* sequencing [Applied Biosystems, 2011]. However, the throughput is high (about 10-15Gb a day) and run costs are fairly low (about US$0.50 per Mb) [Applied Biosystems, 2011; Kircher and Kelso, 2010], making the platform well-suited for resequencing studies for which high consensus accuracies are important.

Side note: Polonator

Polonator [34], a SOLiD-related system, was constructed by Dover System in collaboration with the Church Laboratory of the Harvard Medical School. It uses fragments generated by paired end shotgun library construction techniques that are subsequently amplified by emPCR and sequenced by ligation on a flow cell [Shendure et al., 2011]. Polonator G.007 was introduced as an open platform with freely downloadable, open-source software and protocols, making it cheaper than other NGS machines. Between eight and 10 million reads with an accuracy of $\sim 98\%$ are generated in a four day run but the major drawback is the short read length of currently 26bp in $13 + 13$ base paired end reads. However, as this system is programmable, it might enable user innovations such as alternative biochemistries [Shendure and Ji, 2008].

A.2.3 Life Technologies: Ion Torrent PGM

Life Technologies' Ion Torrent Personal Genome Machine (PGM) [35] is the first commercially available cyclic-array sequencing platform that uses a voltage change rather than fluorescence for detection.

Work Flow Ion Torrent takes advantage of existing semiconductor technology to create a high-density array of micro-machined wells, each of which holds a single DNA fragment, previously

[34] http://polonator.org/
[35] http://www.iontorrent.com/

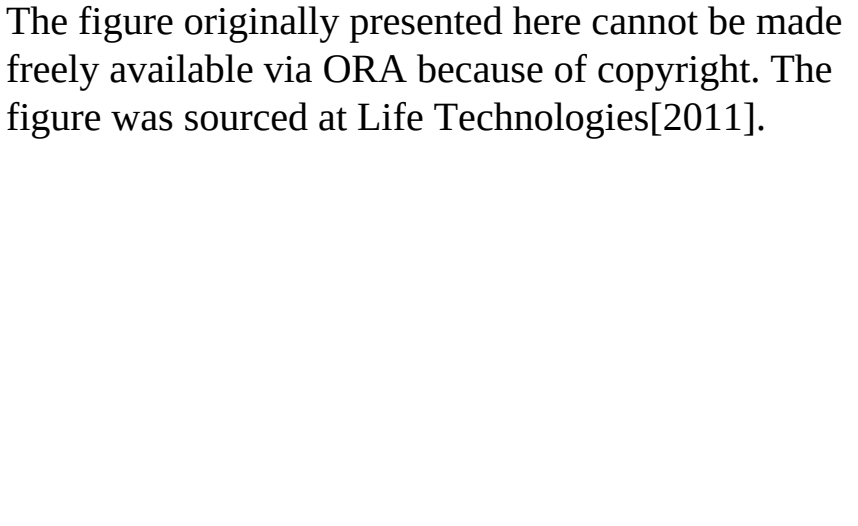


Figure A.5: Ion Torrent work flow. The platform uses a high-density array of micro-machined wells, each of which contains a different unknown DNA template. Nucleotide incorporation into a DNA strand by polymerase releases a hydrogen ion as a by-product. The charge of this ion causes the pH of the solution to change. This change can be detected by the ion sensor beneath the well and the nucleotide that has been incorporated can be inferred. This figure was taken from Life Technologies [2011].

generated from the input DNA following the same protocol than 454 and SOLiD (i.e. em-PCR). In each sequencing cycle, a single dNTP species as well as all other required reagent solutions are added (just as in the 454 method). A polymerase extends the DNA strand if the introduced nucleotide is complementary to the template strand, thereby releasing a hydrogen ion. This release can be sensed by the ion sensor beneath the well as the charge from the ion changes the pH of the solution (see Figure A.5). If there are multiple identical bases on the DNA strand, the chip records multiple identical bases as the voltage is also multiplied. The sequencing is paused after the incorporation of each nucleotide species to wash away unused reagents, allowing a new cycle to begin.

Characteristics Ion Torrent's underlying sequencing principle is very similar to those of all other second-generation technologies, using a typical 'wash-and-scan' scheme. As a consequence, comparable disadvantages are present such as the requirement of a high amount of reagents, sequence biases and errors introduced by the amplification step and a high rate of insertion and deletion errors (e.g. in homopolymeric regions). As a result, the throughput is currently limited to about 100,000 reads with read length in the range of other NGS technologies (∼120bp) [Shendure et al., 2011]. However, the platform also has several unique advantages: neither light, scanning, nor cameras are required, which does not only greatly simplifies and quickens the sequencing procedure, but which also leaves the DNA undamaged and avoids saving large amounts of image data (saving both computational storage space and costs) at the same time.

In 2012, Ion Torrent announced to launch PGM's successor, the Ion Proton Sequencer aiming to sequence an entire human genome in one day for just US$1,000 [36]. The system will rely on a protocol similar to the PGM one, using either an Ion Proton I chip with 165 million sensors (to sequence exoms, planned availability: mid of 2012) or an Ion Proton II Chip with 660 million sensors (to sequence whole human genomes, planned availability: end of 2012).

A.3 Third-generation Sequencers

Next-generation sequencing platforms already facilitate large-scale genome sequencing projects, such as the 1000 Genomes Project and the International Cancer Genome Consortium, but third-generation technologies, relying on a diverse set of ideas, might take genomics even further by using single molecules. Single molecule methods decrease the overall sequencing costs by circumventing laboratory-intense amplification steps and by requiring smaller amounts of reagents due to an often continuous sequencing procedure which avoids several flushing, scanning, and washing steps. These steps are not only time consuming, but they also often introduce systematic biases and sequencing errors (e.g. through dephasing). Another major advantage of most third-generation approaches is that they are able to quickly produce reads with lengths similar to those obtained from capillary-based Sanger instruments, making the analysis of the sequencing data easier and making them better suited for *de novo* assembly.

However, most of the next-next generation sequencing platforms are either in their early stages of access or still in their test phase and thus, the information provided here only reflects the current status of these technologies. Given the immense diversity, only some of the most prominent examples of third-generation sequencers, which lately captured both media and scientific attention, are discussed here, including Helicos' HeliScope, Pacific Biosciences' PacBio RS platform, and Oxford Nanopore Technologies' sequencing strategy (described in Sections A.3.1-A.3.3), but other technologies are on the horizon which are discussed elsewhere (e.g. Shendure et al. [2011]).

A.3.1 Helicos Biosciences: HeliScope

HeliScope, the sequencer of Helicos BioSciences [37], was the first commercially available single molecule sequencing platform released in 2008. Although it does not require an amplification of the genomic fragments, its protocol still relies on the cyclic interrogation of a dense array of sequencing features, making it very similar to the NGS methods discussed in Section A.2.

[36] http://www.lifetechnologies.com/global/en/home/about-us/news-gallery/press-releases/2012/life-techologies-itroduces-the-bechtop-io-proto.html

[37] http://helicosbio.com/

Work Flow Template libraries are generated by random fragmentation of the double-stranded input DNA into strands of 100-200 nucleotides, which are subsequently melted into single strands. A poly-A universal primer sequence is synthesised onto the 3'-end of each fragment using a polyadenylate polymerase. This poly-A-tail is used to immobilise the fragments by hybridisation to poly-T oligomers covalently attached to the surface of a glass slip to obtain a disordered array of primed single molecule sequencing templates [Harris et al., 2008; Zerbino and Birney, 2008]. In the last step of the polyadenylation, each strand is labelled by addition of a fluorescent nucleotide. A laser illuminating the surface of the flow cell allows to determine the position of each fluorescently labelled fragment which is recorded by a CCD camera. After the templates have been imaged, their 3' fluorescent labels are removed and washed away, before the sequencing reaction starts.

At each sequencing cycle, DNA polymerase as well as a single nucleotide species fluorescently labelled with a non-radioactive cyanine dye, Cy5, are added to extend the reverse strand of the sequences, starting from the poly-T oligomers. Thereby, the fluorescent label slows the polymerase down such that no other nucleotide can be incorporated in the same cycle before the washing step occurs, a process referred to as 'virtual termination' [Bowers et al., 2009; Zhu and Waggoner, 1997]. Both the polymerase and all free nucleotides are washed away, before a fluorescence detection system is used to identify array-wide template-dependent nucleotide incorporations. A subsequent cycle using another nucleotide species can be carried out after the fluorescent dye is chemically cleaved from the template and released. Four cycles of single base extension, one for each nucleotide, constitute a 'quad' and several hundred quads are performed for each sequencing run. An overview of the sequencing protocol is given in Figure A.6.

Characteristics One of the main advantages of HeliScope is its high throughput of about 2.5-4Gb per day (with a run time of about 8 days [Shendure et al., 2011]), through the high number of single molecules that can be sequenced in parallel, at an average cost of US$0.33 per Mb [Helicos BioSciences, 2011; Pushkarev et al., 2009]. Due to the asynchronous nature of the sequencing process, reads produced by a single run vary in lengths between 25 and 55 nucleotides, with an average read length of 35bp [Helicos BioSciences, 2011; Pettersson et al., 2009; Pushkarev et al., 2009; Shendure and Ji, 2008]. An average read length of 35bp is much shorter than the read lengths produced by most current second-generation sequencing platforms and might limit the usage of the platform to analyses where short reads can be mapped to an already available reference genome. However, in contrast to NGS technologies,

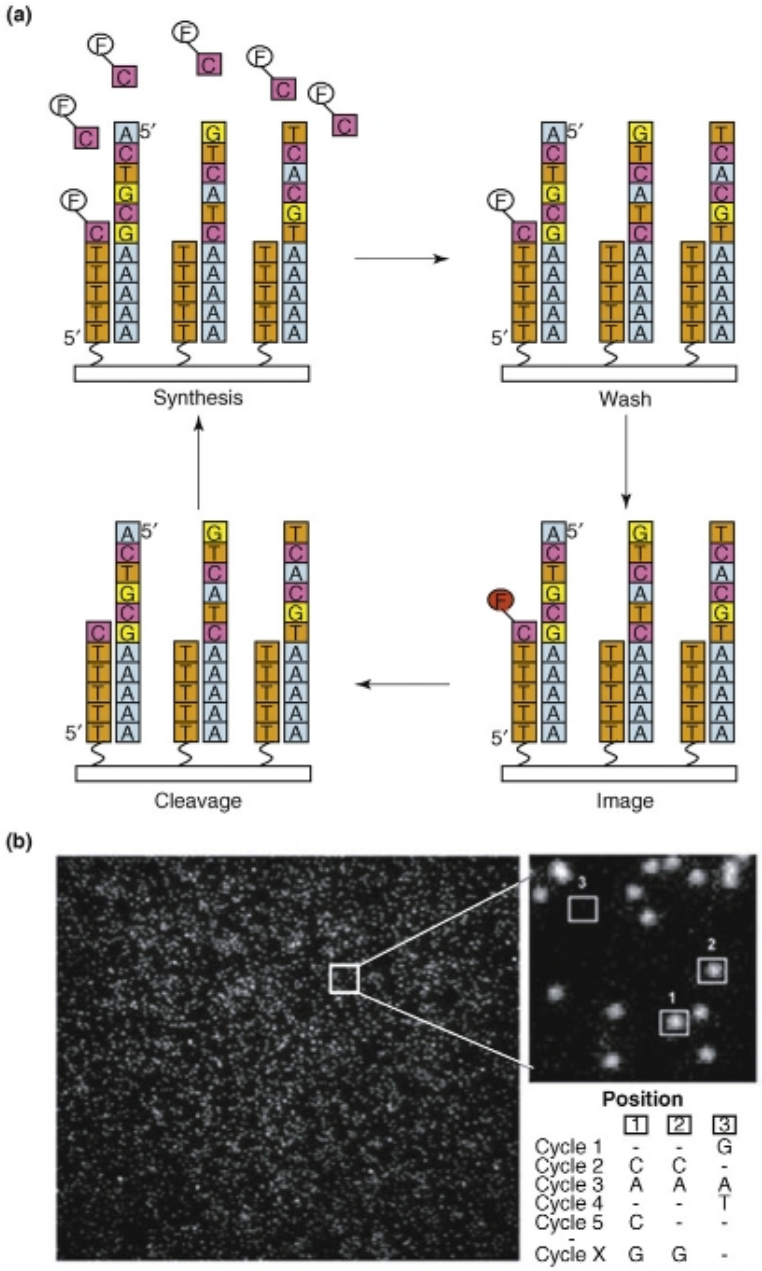


Figure A.6: HeliScope work flow. (a) Top left: The flow cell is incubated with a fluorescently labelled nucleotide which is incorporated in the growing strand in a template-dependent way. Top right: All unincorporated nucleotides are washed away. Bottom right: A fluorescence detection system recognises incorporated nucleotides through the excitation of the attached Cy5 label. Bottom left: The fluorescent label is cleaved off to allow a further round of sequencing. (b) An example for an image and raw data obtained by using HeliScope. The left picture illustrates the poly-T templates after incubation with a fluorescently labelled nucleotide. The right picture shows a close-up view of individual molecules. The position of three templates is marked (1-3) after the second cycle of incubation with labelled cytosine which has been incorporated at positions 1 and 2 but not at 3. Reprinted from Gupta [2008], Copyright (2008), with permission from Elsevier.

HeliScope (as well as other single molecule methods) present the advantage of not relying on *in vitro* amplification, which is not only expensive but which can also introduce copying errors and biases [Gupta, 2008; Service, 2006] leading to a low substitution error rate in the range of 0.01-1% [Shendure and Ji, 2008].

Similar to the 454 method, templates might fall ahead or behind others, due to their sequence or by chance by being unable to incorporate a nucleotide on a given cycle despite having the appropriate base at the next position, but dephasing is not an issue as single molecules rather than clusters are sequenced. Instead, a misincorporation event would manifest as a 'dead' template which would not extend any further whilst a non-incorporation event would appear as a 'pause' in the sequence [Shendure et al., 2004]. On the other hand, using single molecules rather than clusters results in weaker signals and no possibility to correct for misincorporation events. In addition, homopolymers can cause problems as no terminating moieties are used on the nucleotides, an issue that is usually attenuated by limiting the rate of incorporation events [Shendure and Ji, 2008]. A further source of error is the incorporation of contaminated, unlabelled, or non-emitting nucleotides. As a consequence, deletions are the dominant type of error with error rates in the range of 2-7% [Shendure and Ji, 2008].

A two-pass strategy can be implemented to improve the raw sequencing accuracy. A template array with adaptors at both ends is sequenced as described above and then copied entirely. After new strands are synthesised, the original template can be removed by denaturation, as strands are tethered to the surface. Another sequencing event leads to a second sequence for the same template (in the opposite orientation). The two-pass strategy enables to lower both substitution error rates as well as deletion error rates to 0.001% and 0.2-1%, respectively [Shendure and Ji, 2008].

Apart from DNA sequencing, HeliScope can also be used to directly sequence RNA by replacing the DNA polymerase with a reverse transcriptase in the protocol [Ozsolak et al., 2009].

Side Note: VisiGen

Another single molecule sequencer, which can be seen as an improvement over the HeliScope system, is currently developed by VisiGen Biotechnologies (recently acquired by Life Technologies) and might become available soon. DNA sequence can be obtained in real time by monitoring a specially engineered polymerase containing a donor fluorophore as it incorporates bases into a DNA strand. Thereby, each nucleotide carries one of four differently labelled acceptor fluorophores.

Proximity of donor and acceptor fluorophores during the incorporation event results in a FRET signal specific for each type of nucleotide due to an energy transfer from donor to acceptor. To allow for a next incorporation, the fluorophore is released which quenches the signal.

A.3.2 Pacific Biosciences: PacBio RS

In April 2011, Pacific Biosciences[38] launched their third-generation sequencing platform, PacBio RS, commercially [Korlach et al., 2008].

Work Flow PacBio RS uses a silicon dioxide chip with a 100nm thick metal film containing thousands of holes with a diameter in the range of 10-50nm, so-called 'zero-mode waveguides' (ZMW), each with a single DNA polymerase molecule anchored to their bottom, to sequence single molecules in real time (Single Molecule Real Time, SMRT) [Eid et al., 2009]. As the ZMW and its volume is very small (20 zeptolitres = 20×10^{-21} litres), visible laser light, which has a wavelength of ∼600nm, can not pass through the ZMW entirely [Schadt et al., 2010]. Instead, it decays exponentially as it enters the ZMW, only illuminating the bottom 30nm of the ZMW where the DNA polymerase is attached. Nucleotides, fluorescently labelled at their phosphate group, are flooded above the ZMWs. Within microseconds, they diffuse inside the ZMWs and complementary nucleotides are incorporated into the growing DNA strand by the DNA polymerase. Although all nucleotides within the bottom 30nm of the ZMW fluoresce, the signal of the incorporated nucleotide can be distinguished from the free ones as the strand synthesis takes about three orders of magnitude longer than simple diffusion (the incorporation of a new nucleotide by the polymerase requires the formation of a phosphodiester, unlike diffusion which happens in the range of microseconds, this process takes up milliseconds facilitating the detection), resulting in a higher signal intensity that can be used to obtain the sequence information using a CCD camera. After the nucleotide incorporation, the fluorophore attached to the phosphate group is released and the phosphate-dye complex quickly diffuses out of the ZMW [Levene et al., 2003]. In this way, 10 bases can be incorporated per second resulting in a chain of thousands of nucleotides after a few minutes [Korlach et al., 2008]. PacBio RS's workflow is depicted in Figure A.7.

Characteristics Similar to all other single molecule technologies, the PacBio RS does not require any amplification of the input DNA, avoiding systematic biases and sequencing errors. The protocol also does not include any time- and resource-intensive scanning and washing steps, enabling the use of less consumables as well as quicker results than any NGS platform (typical

[38] `http://www.pacificbiosciences.com`

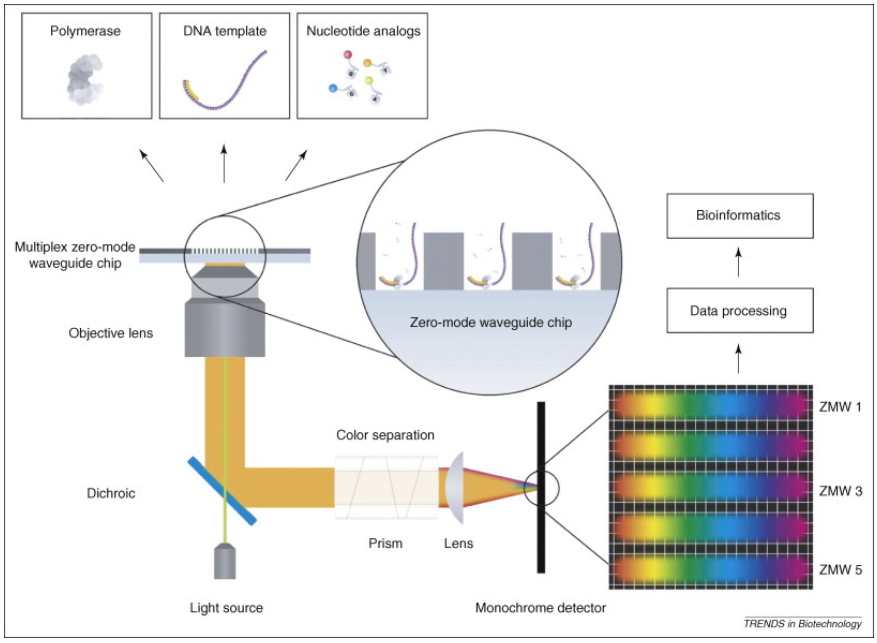
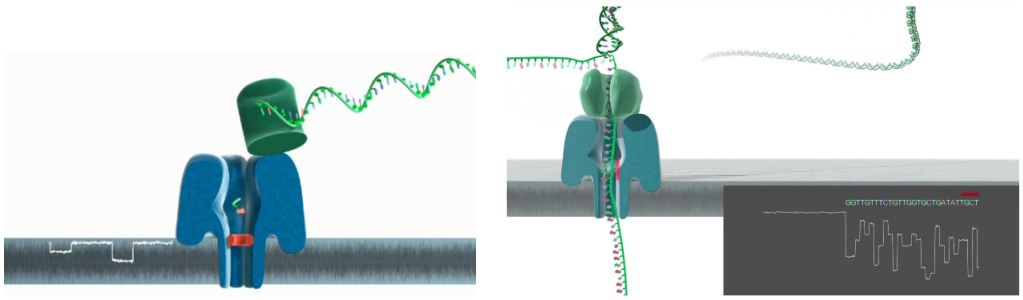


Figure A.7: Pacific Biosciences work flow. Silicon dioxide chips containing thousands of ZMWs are used to sequence single molecules in real time. The incorporation of a fluorescently labelled nucleotide during DNA synthesis can be detected through a laser-beam-mediated illumination within the ZMW. The light flash is separated into a spatial array from which the incorporated nucleotide species can be determined. Reprinted from Gupta [2008], Copyright (2008), with permission from Elsevier.

sequence run time is ∼15 min with a throughput of 100Gb per hour [Mardis, 2011]) [Schadt et al., 2010]. In addition, it allows the direct detection of chemical modifications to the nucleotides (such as methylation) [Schadt et al., 2010]. Read lengths are in the order of thousand nucleotides and are mainly limited by the denaturation of the polymerase through the laser [Kircher and Kelso, 2010; Mardis, 2011]. The main disadvantage of the platform is its high raw read error rate, mainly caused by insertions and deletions (e.g. through an incorporation of an unlabelled nucleotide, through a failure to detect an incorporation event because the emitted photons could not be distinguished from the background noise, or through a failure of the polymerase to incorporate a nucleotide at the first try, making multiple tries look like different incorporation events [Qin et al., 2010; Schadt et al., 2010]), which can be in excess of 5%, hindering the reliable assembly and the alignment of the genomic sequences [Schadt et al., 2010; Shendure et al., 2011].

A.3.3 Oxford Nanopore Technologies

Nanopore sequencing is an ultra-fast DNA sequencing method. It uses a nanopore, either a chemically modified biological membrane protein such as the heptameric α-hemolysin pore [Clarke et al., 2009] or a synthetic pore [Deamer et al., 2000], usually inserted into a membrane created by a synthetic lipid bilayer which has very high electronic resistance. A potential is applied across the bi-



(a) (b)

Figure A.8: (a) Exonuclease sequencing. (b) Strand sequencing. These figures were taken from Oxford Nanopore Technologies' web page.

layer (for example by changing the concentration of salt) resulting in a current flowing through the nanopore. Negatively charged nucleic acids driven towards a positive charge through the nanopore cause characteristic disruptions in the current which can be measured with a cyclodextrin sensor to identify the DNA sequence [Astier et al., 2006; Clarke et al., 2009].

Work Flow Two different sequencing approaches have been implemented by Oxford Nanopore Technologies [39], both using a chemically modified α-hemolysin nanopore, engineered to allow single base resolution within a DNA strand: exonuclease sequencing (see Figure A.8(a)) and strand sequencing (see Figure A.8(b)). In the first, a nanopore is coupled with an exonuclease enzyme (depicted in green) on the extracellular face of the pore which cleaves individual bases from the end of a DNA strand. As they pass through the nanopore, they are transiently bond with an adapter molecule, cyclodextrin (indicated in red), which is covalently attached to the inside surface of the nanopore [Rusk, 2009]. The interaction between the nucleotide and the cyclodextrin blocks the channel and generates a characteristic charge in the membrane electrical conductance. Each kind of nucleotide produces a slightly different conductance change which can be measured in an electric circuit in order to identify the nucleotide in sequence. In the second method, nucleotides are not cleaved from the DNA strand. Instead intact single stranded DNA is sequenced as it passes through the nanopore using an enzyme (shown in green). The enzyme is bound to the end of the double-stranded DNA, which it unwinds and passes as a single strand through the nanopore, one base at a time. Similar to exonuclease sequencing, a disruption in current can be used to determine the nucleotides as they pass through the nanopore. If the DNA is prepared in a way such that it has a hairpin structure at its end, both strands can be read.

Unique to the nanopore technology is the ability to not only read the four different nucleotides but also to directly read methylcytosine without prior bisulfide conversion.

[39] `www.nanoporetech.com/`

Characteristics Oxford Nanopore Technologies' major strength is its ability to quickly sequence reads of much longer lengths than those produced by most NGS and its potential to directly detect individual nucleotide identifications (such as 5'-methylcytosine) during the sequencing process [Clarke et al., 2009]. Similar to Ion Torrent's PGM, nanopore sequencing does neither require fluorescent reagents to label nucleotides nor expensive CCD cameras to record from optical chips, which saves both time and money and which leaves the DNA undamaged. Like all single molecule methods, no amplification step is needed which could introduce sequencing biases and errors. Furthermore, homopolymer runs usually do not pose a problem as the pore records each nucleotide independently. Nevertheless, some obstacles in nanopore-based sequencing must be addressed before commercialisation of a platform such as the reliable and reproducible pore fabrication, the integration of sensor and electronics, motion control, detection sensitivity, DNA orientation as it traverses through the pore and its speed (for instance, if the DNA is driven electrophoretically, it passes through the nanopore too fast to enable accurate resolution of individual bases), the method to distinguish different nucleotides, and parallelisation. However, a better understanding of its full characteristics and drawbacks is only possible after the technology becomes commercially available.

In February 2012, Oxford Nanopore Technologies announced the launch of its new machine, the GridION, for the second half of this year after sequencing the entire 5.4kb genome of the Phi X phage in one continuous read [40]. Also planned is a launch of the world's first miniaturised, disposable USB drive sequencer, called the MinION. The initial system will contain 2,000 nanopores, each of which able to read DNA at a rate of hundreds of kilobases per second with error rates of $\sim$4%.

[40] http://www.nature.com/news/nanopore-genome-sequencer-makes-its-debut-1.10051

Appendix B

FASTQ Format

Different closely related FASTQ formats exist, which compactly store FASTA sequences (see Appendix E) and their base qualities in the form of four lines per sequence

```
@WTCHG_413:1:1:1:17#0
TTTTAAATATTCCTTTTGATATCTCCCATTGCATATCAGCTAATATCATTN
+
976:;;4;934877<;;;54;;8;79;;:236:879:::89;899439;//
```

where the first line contains an `@` followed by the sequence name (a free format field with no length limit), the second line represents the raw sequence in FASTA format, the third line contains a `+` to signal the end of the sequence line which can be either followed by a description or left empty, and the fourth line represents the quality scores in ASCII code (which must have the same length as the sequence line). The major difference between the various FASTQ formats lies in the encoding scheme which is used to indicate the quality scores as the calculation of quality varies between platforms:

Sanger FASTQ Quality scores q_{Phred} are computed using the Phred formula [Ewing and Green, 1998]

$$q_{\text{Phred}} = -10 \ \log_{10} \ p_{\text{Phred}},$$

where p_{Phred} is the probability that the base call is wrong. Hence, probabilities of a miscall of 10%, 1%, and 0.1% lead to Phred scores of 10, 20, and 30, respectively. Sanger FASTQ files use ASCII 33-126 to encode Phred qualities from 0 to 93.

Illumina 1.0 FASTQ Quality scores q_{Illumina} are calculated as

$$q_{\text{Illumina}} = -10 \ \log_{10} \ \left(\frac{p_{\text{Illumina}}(x)}{1 - p_{\text{Illumina}}(x)} \right),$$

where $p_{\text{Illumina}}(x)$ is the error probability. Illumina's quality scores can be easily converted into Phred ones (and vice versa)

$$q_{\text{Phred}} = 10 \ \log_{10} \ \left(10^{\frac{q_{\text{Illumina}}}{10}} + 1 \right)$$

$$q_{\text{Illumina}} = 10 \ \log_{10} \ \left(10^{\frac{q_{\text{Phred}}}{10}} - 1 \right),$$

resulting in the fact that for high scores, the two quality measures are asymptotically identical. Illumina's quality scores range from -5 to 62 using ASCII 33 to 126.

Illumina 1.3+ FASTQ From Genome Analyzer version 1.3 onwards, quality scores are computed using the Phred formula. However, to guarantee interchangeability with their earlier FASTQ format, they introduced yet another format instead of adopting the original Sanger one: Illumina's 1.3+ FASTQ format encodes Phred quality scores in the range from 0 to 62 using ASCII 64 to 126 although normally only the range 0 to 40 is observed.

Appendix C

Alignment with MAQ

Before an alignment of Illumina reads can be performed with MAQ, two preliminary steps are necessary. First, to ensure reliable SNP calling with MAQ, the Illumina raw read format must be converted as it scales base qualities differently than 'mapass' FASTQ format (see Appendix B)

```
maq sol2sanger reads.txt reads.fastq,
```

where reads.txt is the Illumina read sequence file. Second, to reduce disc space and to accelerate file I/O, both reads and the reference sequence need to be converted into a binary format. The read sequences are converted into MAQ's binary FASTQ (BFQ) format by

```
maq fastq2bfq reads.fastq reads.bfq.
```

The reference sequence is converted to MAQ's binary FASTA (BFA) format using

```
maq fasta2bfa reference.fasta reference.bfa.
```

After these initial file transformations, a paired end alignment can be performed

```
maq map alignment.map reference.bfa reads1.bfq reads2.bfq,
```

where reads1.bfq and reads2.bfq contain an end of the paired end reads each with reads sorted in the same order. The output alignment.map is a compressed binary file storing sequences, qualities, positions of the mapped reads, and related information such as the number of mismatches and the mapping quality. MAQ was run with the default options allowing hits with up to two mismatches to be found and discarding alignments with a sum of mismatching base qualities greater than 70. As MAQ usually performs best when the number of reads is around two million (either too few or too many reads yield bad efficiency as MAQ's memory is linear to the number of reads in the alignment), the paired end input files were split in read junks of one million reads each.

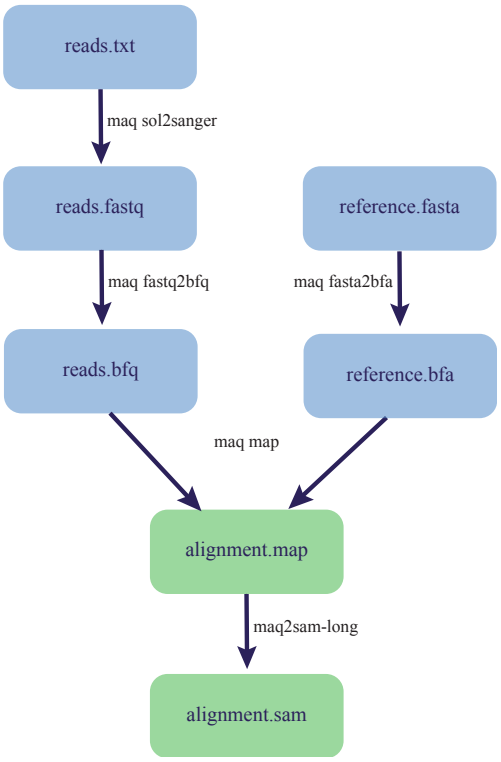


Figure C.1: MAQ work flow. Step 1: Read pre-processing (a) Convert reads in Illumina FASTQ format (.txt) to Sanger FASTQ format (.fastq). (b) Convert reads into MAQ binary FASTQ (.bfq) format. Step 2: Reference pre-processing. Convert reference sequence (.fasta) to MAQ binary FASTA (bfa). Step 3: The binary formats of the reads and the reference are used to perform an alignment. Step 4: MAQ initial alignment format (.map) is converted into standard SAM format (.sam).

Due to the split input files, individual results needed to be merged with the following command:

```
maq mapmerge alignments.map alignment1.map alignment2.map alignment3.map ... .
```

Last, MAQ's map format is converted into standard SAM format (see Appendix D.1)

```
maq2sam-long alignments.map > alignments.sam.
```

An overview of MAQ's work flow is given in Figure C.1.

Appendix D

SAM and BAM Format

D.1 SAM Format

The Sequence Alignment/Map (SAM) format [Li et al., 2009a] is a format for storage of platform-independent NGS data. Reads stored in a TAB-delimited SAM file can be single or paired end with a length of up to 128Mb produced by different platforms. The format consists of two sections, one for the header information in which each line starts with an `@` and one for the alignment information. A brief outline of the SAM format is provided here. For more detail visit SAMtools' [41] web page.

Header Information

The `@HD` header line gives information about the used file format version (`VN`) and the sort order of the alignment (`SO`). It has a further optional tag to indicate the group order (`GO`). The order can either be `unsorted`, `queryname`, or `coordinate`, whereby most operations on the alignments only work on a file which is sorted by the leftmost reference coordinate. The `@RG` line contains the read group information: the unique read group identifier (`ID`), the sample (`SM`) and the used library (`LB`), the name of the sequencing centre (`CN`), platform used to generate the read (`PL`) and its unit (`PU`), and the predicted mean insert size (`PI`). In addition, it can contain information about the date of the run (`DT`) and a description (`DS`). The line starting with `@PG` gives an overview of the program which was used to create the alignment such as its name (`ID`) and version (`VN`) as well as the command line for the run (`CL`). Lines beginning with `@CO` are one-line comments whilst the ones starting with `@SQ` give details about the used sequence dictionary (such as the order of the reference sequences). The only two required fields in the latter are the name of the sequence (`SN`) and its length (`LN`).

[41] `http://samtools.sourceforge.net`

An example for the header section is given below

```
@HD  VN:1.0  SO:coordinate
@RG  ID:WTCHG_1020  SM:PtPE  PL:ILLUMINA  PU:091110_GAII04_0003_1  LB:186/09 \
     CN:WTCHG  PI:530bp
@PG  ID:stampy  VN:v0.85_(November_2009) CL:--comment=@Lane_1_comments.txt \
     --solexa -v0 -g /tmp/Chimp2 -h /tmp/Chimp2 \
     -M s_1_1_sequence.txt,s_1_2_sequence.txt \
     --bwaoptions=/tmp/Chimp2 -o s_1_stampy.sam
@CO  SN:PtPE ID:186/09  GE:Chimp2  SR:gDNA PE  CT:false  PR:Chimpanzee genomes \
     SM:186/09
@SQ  SN:chr1   LN:229974691
@SQ  SN:chr2a  LN:114460064
@SQ  SN:chr2b  LN:248603653
@SQ  SN:chr3   LN:203962478
```

Alignment Information

Each alignment has 11 mandatory fields which describe key information about the alignment. In case the corresponding information is unavailable, the value of the field is set to `*` or `0` (depending on the field). It can further contain a variable number of optional fields which store extra information from the read mapper or sequencing platform and which are presented as key-value pairs in the form `TAG:TYPE:VALUE`. The mandatory fields in order of their appearance encode

Field	Description
QNAME	The name of the query read (or the read pair).
FLAG	A bit-wise flag which gives information about whether the read is paired, its own strand as well as the one of its mate, whether it failed any platform/vendor quality checks, and whether it is a PCR or optical duplicate.
RNAME	The name of the reference sequence.
POS	The 1-based leftmost position of the clipped alignment.
MAPQ	The Phred-scaled estimated probability that the read was aligned to the wrong location assuming that the read was generated from the reference and that the quality scores are correct.
CIGAR	The original CIGAR string describes the pairwise alignment using three characters: `M` for match/mismatch, `I` for insertion, and `D` for deletion. To support splicing, clipping, multi-part and padded alignments, an extended version also encodes `N` for skipped bases on the reference, `S` for soft clipping, `H` for hard clipping, and `P` for padding.
MRNM	The name of the mate's reference (`=` if the same as RNAME).
MPOS	The 1-based leftmost position of the mate alignment.
ISIZE	The inferred insert size.
SEQ	The query sequence on the same strand as the reference sequence.
QUAL	The quality of the queried read.

Table D.1: Fields in the alignment information section.

The SAM format specification [42] gives more information as well as a more detailed overview of each field (including the optional ones). An example for an alignment line is given below

```
GAII03:1:99:1562:362#0  99  chr1  77738  0  51M  *  0  407
TGATTATCCATTGTCACCCTTCCCTCCAAACAGCCACCTCTCCCCTGGAGA
B@9BB=A@@AB@;4;AAA;=@@=431?BAB7<46=@6==?=?=42636:7A
```

D.2 BAM Format

The Binary Alignment/Map (BAM) format is the binary representation of a SAM file. It is compressed by the BGZF library, a generic library developed by the authors to enable rapid access in a zlib-compatible compressed file. More information about both the BGZF compression and the BAM format specification can be found in the SAM format specification.

[42] http://samtools.sourceforge.net/SAM1.pdf

Appendix E

FASTA Format

The FASTA format is a text-based format which can either represent single sequences or many sequences in a single file. It was originally invented by Bill Pearson as an input format for his FASTA suite of tools [Pearson and Lipman, 1988]. It uses standard IUB/IUPAC single letter codes to represent base pairs (with the exceptions that lower-case letters are accepted and are mapped into upper-case ones and a dash can be used to represent a gap):

A Adenosine

C Cytosine

G Guanine

T Thymidine

U Uracil

R G A (puRine)

Y T C (pYrimidine)

K G T (Ketone)

M A C (aMino group)

S G C (Strong interaction)

W A T (Weak interaction)

B G T C (not A) (B comes after A)

D G A T (not C) (D comes after C)

H A C T (not G) (H comes after G)

V G C A (not T, not U) (V comes after U)

N A G C T (aNy)

X masked

- gap of indeterminate length

A sequence in FASTA format is represented as a set of lines. The first line containing a comment or a unique description of the sequence starts with a >. The following lines comprise the actual sequence which ends as soon as another line starts with > (indicating the start of another sequence). An example is given below

```
>Chimpanzee gene for bone gla protein (BGP)
GGCAGATTCCCCCTAGACCCGCCCGCACCATGGTCAGGCATGCCCCTCCTCATCGCTGGGCACAGCCCAGAGGGT
ATAAACAGTGCTGGAGGCTGGCGGGGCAGGCCAGCTGAGTCCTGAGCAGCAGCCCAGCGCAGCCACCGAGACACC
ATGAGAGCCCTCACACTCCTCGCCCTATTGGCCCTGGCCGCACTTTGCATCGCTGGCCAGGCAGGTGAGTGCCCC
CACCTCCCCTCAGGCCGCATTGCAGTGGGGGCTGAGAGGAGGAAGCACCATGGCCCACCTCTTCTCACCCCTTTG
GCTGGCAGTCCCTTTGCAGTCTAACCACCTTGTTGCAGGCTCAATCCATTTGCCCCAGCTCTGCCCTTGCAGAGG
GAGAGGAGGGAAGAGCAAGCTGCCCGAGACGCAGGGGAAGGAGGATGAGGGCCCTGGGGATGAGCTGGGGTGAAC
CAGGCTCCCTTTCCTTTGCAGGTGCGAAGCCCAGCGGTGCAGAGTCCAGCAAAGGTGCAGGTATGAGGATGGACC
TGATGGGTTCCTGGACCCTCCCCTCTCACCCTGGTCCCTCAGTCTCATTCCCCCACTCCTGCCACCTCCTGTCTG
GCCATCAGGAAGGCCAGCCTGCTCCCCACCTGATCCTCCCAAACCCAGAGCCACCTGATGCCTGCCCCTCTGCTC
CACAGCCTTTGTGTCCAAGCAGGAGGGCAGCGAGGTAGTGAAGAGACCCAGGCGCTACCTGTATCAATGGCTGGG
GTGAGAGAAAAGGCAGAGCTGGGCCAAGGCCCTGCCTCTCCGGGATGGTCTGTGGGGGAGCTGCAGCAGGGAGTG
GCCTCTCTGGGTTGTGGTGGGGGTACAGGCAGCCTGCCCTGGTGGGCACCCTGGAGCCCCATGTGTAGGGAGAGG
AGGGATGGGCATTTTGCACGGGGGCTGATGCCACCACGTCGGGTGTCTCAGAGCCCCAGTCCCCTACCCGGATCC
CCTGGAGCCCAGGAGGGAGGTGTGTGAGCTCAATCCGGACTGTGACGAGTTGGCTGACCACATCGGCTTTCAGGA
GGCCTATCGGCGCTTCTACGGCCCGGTCTAGGGTGTCGCTCTGCTGGCCTGGCCGGCAACCCCAGTTCTGCTCCT
CTCCAGGCACCCTTCTTTCCTCTTCCCCTTGCCCTTGCCCTGACCTCCCAGCCCTATGGATGTGGGGTCCCCATC
```

Appendix F

VCF Format

The Variant Calling Format (VCF) [Danecek et al., 2011] is a format to store information about genetic variants. It consists of three sections, one for the meta-information in which each line starts with a `##`, one for the header information (one line starting with a `#`) and one for the variant information. A brief outline of the VCF is provided here. For more details visit the 1000 Genomes web page [43].

Meta-Information

The only required meta-information is the `fileformat` field indicating which VCF format version was used. `INFO` fields containing entries used in the variant information section of the VCF file, `FILTER` fields describing filters which have been applied to the data, and `FORMAT` fields specifying genotypes, are optional. If they are present they must follow the predefined format specifications

```
##INFO=$<$ID=ID,Number=number,Type=type,Description=description$>$
##FILTER=$<$ID=ID,Description=description$>$
##FORMAT=$<$ID=ID,Number=number,Type=type,Description=description$>$
```

where the `Number` entry is an Integer describing the number of values that can be included within the field. Those values can have the following `Types`: Integer, Float, Character, and String. `INFO` fields have an additional Flag type which indicates that the `INFO` field does not contain a value entry.

[43] `http://www.1000genomes.org`

An example for the meta-information section is given below

```
##fileformat=VCFv4.0
##INFO=$<$ID=AN,Number=1,Type=Integer,Description="Total number of alleles in \
called genotypes"$>$
##FILTER=$<$ID=AN,Description="AN$<$16"$>$
##FORMAT=$<$ID=GT,Number=1,Type=String,Description="Genotype"$>$
##FORMAT=$<$ID=AD,Number=.,Type=Integer,Description="Allelic depths for the REF \
and ALT alleles in the order listed"$>$
##FORMAT=$<$ID=DP,Number=1,Type=Integer,Description="Read Depth (only filtered \
reads used for calling)"$>$
##FORMAT=$<$ID=GL,Number=3,Type=Float,Description="Log-scaled likelihoods for \
AA,AB,BB genotypes where A=REF and B=ALT; not applicable if site is not \
biallelic"$>$
##FORMAT=$<$ID=GQ,Number=1,Type=Float,Description="Genotype Quality"$>$
```

Header Information

The header information is presented in a single tab-delimited line. This line contains eight mandatory columns: Missing values in fixed fields are specified with a dot. The mandatory columns are

Field	Description
CHROM	Chromosome identifier from the reference genome. (Alphanumeric String, Required)
POS	Position in the reference genome (1-based). (Integer, Required)
ID	Semicolon separated list of unique identifiers (where available). (Alphanumeric String)
REF	Reference base(s) (multiple bases in the form A,C,G,T,N are permitted). (String, Required)
ALT	Comma separated list of alternate non-reference alleles called on at least one of the samples. (Alphanumeric String; no whitespaces, commas, or angle-brackets are permitted in the ID String itself)
QUAL	Phred-scaled quality score for the assertion made in ALT. (Numeric)
FILTER	`PASS` if this position has passed all filters; otherwise: a semicolon-separated list of codes for filters that failed. (Alphanumeric String)
INFO	Additional information encoded as a semicolon-separated series of short keys with optional values in the format `<key>=<data>[,data]` although some subfields are reserved (see online documentation for details; albeit optional). The exact format of each `INFO` subfield should be specified in the meta-information (as described above).

Table F.1: Fields in the header information section.

followed by a `FORMAT` column header and an arbitrary number of sample IDs if genotype data is present in the file. An example for the header information section is given below

```
#CHROM  POS  ID  REF  ALT  QUAL  FILTER  INFO  FORMAT  \
Individual1  Individual2  Individual3  ...
```

Variant Information

The meta-information section and the header information section are followed by information about genetic variation identified in the data. An example is given below

```
chr1  132703  .  T  A  775.32    AN  AN=14  GT:AD:DP:GL:GQ  ./.  \
1/1:0,5:5:-15.69,-1.51,-0.01:15.04  1/0:0,1:1:-3.20,-0.30,-0.00:3.01  ...
chr1  133337  .  C  G   80.29  PASS  AN=16  GT:AD:DP:GL:GQ  ./.  \
0/1:4,1:4:-3.45,-1.21,-11.38:22.42  0/0:17,2:7:-0.02,-2.12,-23.86:21.06  ...
```

As indicated in the header section, the first and second entries specify the chromosome and the position the variation was found on (e.g. `chr1`, `132703`). This is followed by an identifier (e.g a rs number for a dbSNP variant or a missing value '`.`' if there is no identifier available), the reference base (e.g `T`), any alternate non-reference alleles which were called in at least one sample (e.g `A`), and a Phred-scaled quality score for the assertion made (e.g. `775.32`). The next entry indicates whether or not the variant has passed all `FILTER` criteria set out in the meta-information section (`PASS` or a list of codes for filters that failed e.g. `AN` notifies that the total number of alleles in called genotypes does not fulfil the `FILTER` criteria - in the example above `AN<16`). The last mandatory column gives further information about the called variant (e.g. the total number of alleles in called genotypes `AN` is `14`). If genotype information is present, the required entries are followed by a `FORMAT` field which indicates the data types and order (in the example above `GT:AD:DP:GL:GQ`) and which is followed by one field per sample, with the colon-separated data in this field corresponding to the types specified in the `FORMAT`. In general, the keywords for the `FORMAT` subfields can be defined in the meta-information section (e.g. `AD`) but several keywords are reserved :

Field	Description
FLAG	A bit-wise flag which gives information about whether the read is paired.
GT	Genotype information encoded as alleles values (0 for the reference allele; 1 for the first allele listed in ALT; 2 for the second allele list in ALT; ...; '.' for a missing allele if no call could be made) and separated by either '`/`' (if the genotype call was unphased) or by '`|`' (if it was phased). The genotype information is always the first subfield of the `FORMAT` entry.
DP	Sample read depth at this position. (Integer)
GL	Three floating point log10-scaled likelihoods for AA,AB,BB genotypes where A=REF and B=ALT (only applicable if site is biallelic). (Numeric)
GQ	Genotype quality encoded as a Phred quality. (Numeric)
HQ	Two comma separated Phred-scaled haplotype qualities. (Numeric)
FT	Sample genotype filter gives information about the genotype call (`PASS` indicates that all filters have been passed; a semicolon separated list of codes indicates filters that failed; '.' indicates that no filters have been applied). (Alphanumeric String)

Table F.2: Fields in the variant information section.

Appendix G

MHC Region

	Statistics		
	Gene	**Start Position**	**End Position**
MHC Class I	HLA-A	30,572,528	30,575,871
	HLA-B	32,003,858	32,007,152
	HLA-C	31,918,678	31,921,781
MHC Class II	HLA-DPA1	33,762,124	33,778,345
	HLA-DPB1	33,773,495	33,787,193
	HLA-DPB2	33,810,004	33,826,612
	HLA-DQA1	33,325,715	33,430,756
	HLA-DQA2	33,424,735	33,430,266
	HLA-DQB1	33,342,988	33,350,215
	HLA-DQB2	33,439,486	33,446,913
	HLA-DRA	33,017,744	33,022,942
	HLA-DRB1	33,179,664	33,277,102
	HLA-DRB3	33,179,666	33,192,748
	HLA-DRB4	33,100,900	33,277,078
	HLA-DRB5	33,100,819	33,114,525
	HLA-DRB6	33,137,343	33,143,125

Table G.1: Regions of MHC genes as defined by the 'Known Genes' dataset [Hsu et al., 2006] in the UCSC table browser [Karolchik et al., 2004]

Appendix H

Sequence-context of Coding and Non-coding Shared Polymorphisms

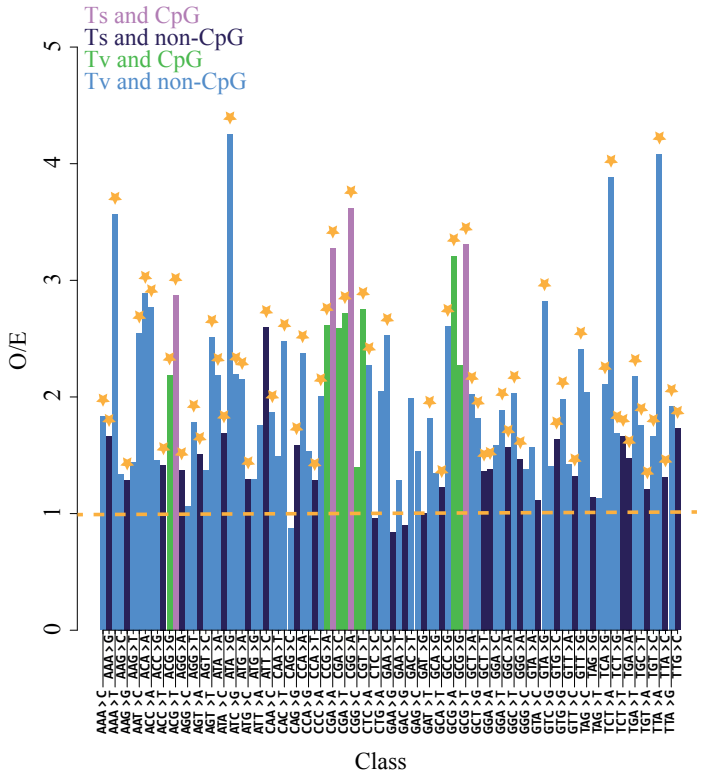


Figure H.1: Ratio of the observed (O) rate of sharing to the expected (E) one per mutation class on the autosomes for non-coding SNPs. To simplify the plots, strands have been collapsed (e.g. ACG>C and CGT>G have been collapsed to one class as CGT is the reverse complement of ACG). Bars have been coloured by whether or not the SNP was a putative CpG-mutation, transition (Ts), or transversion (Tv). * in the plots indicates a p-value<0.01 using a likelihood-ratio test LRT $= -2\ (\mathrm{O} - \mathrm{E}) + 2\ \mathrm{O}\ (\log \mathrm{O} - \log \mathrm{E})$ with the null hypothesis that the ratio is 1 (indicated by an orange line).

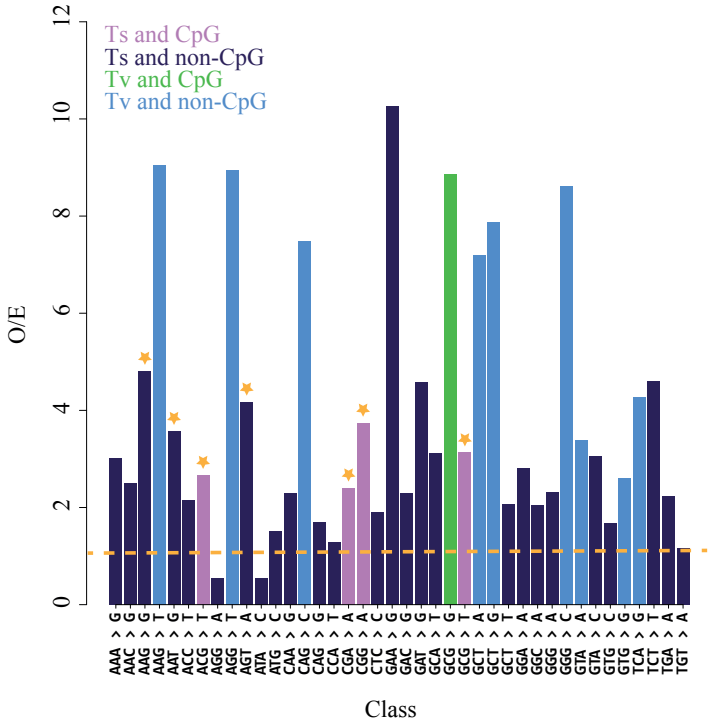


Figure H.2: Ratio of the observed (O) rate of sharing to the expected (E) one per mutation class on the autosomes for protein-coding SNPs. To simplify the plots, strands have been collapsed (e.g. ACG>C and CGT>G have been collapsed to one class as CGT is the reverse complement of ACG). Bars have been coloured by whether or not the SNP was a putative CpG-mutation, transition (Ts), or transversion (Tv). * in the plots indicates a p-value<0.01 using a likelihood-ratio test LRT $= -2\ (\mathrm{O} - \mathrm{E}) + 2\ \mathrm{O}\ (\log \mathrm{O} - \log \mathrm{E})$ with the null hypothesis that the ratio is 1 (indicated by an orange line).

Appendix I

Concept of Replication Slippage

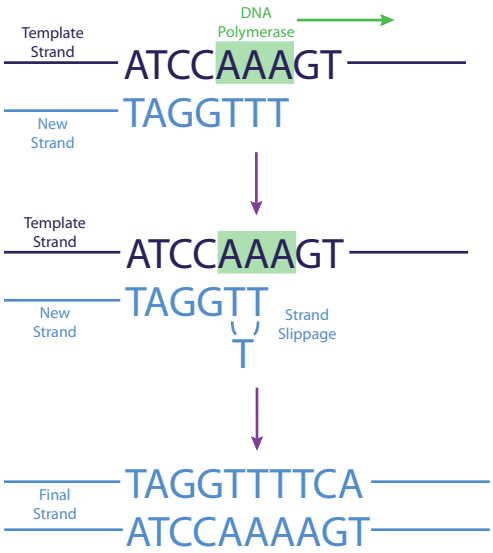


Figure I.1: Replication slippage (a commonly observed replication error) frequently occurs at repetitive sequences when the new strand mispairs with the template strand. Forward slippage (depicted here) on the newly synthesised (3' to 5') strand leads to a deletion while backward slippage on the 5' to 3' strand causes an insertion mutation.

Appendix J

Polymorphisms involved in the observed Haplotype Structure

Shared Coding SNP			Statistics			
Chr	**Pos**	**Gene**	**SNP_{Chr}**	**SNP_{Pos}**	**Distance (in bp)**	**AF**
chr1	150637486	FMO2	chr1	150637526	40	0.90
chr1	150637526	FMO2	chr1	150637486	40	0.90
chr3	65594896	PRICKLE2	chr3	65585938	8958	0.65
chr3	189156557	KLHL24	chr3	189157884	1327	0.90
chr3	200066698	LRRC15	chr3	200067394	696	0.65
chr3	200067394	LRRC15	chr3	200066698	696	0.20
chr4	194201924	HSP90AA4P	chr4	194204093	2169	0.80
chr6	30756814	TRIM31	chr6	30766807	9993	0.70
chr6	31741577	C6orf15	chr6	31728025	13552	0.45
			chr6	31729766	11811	0.45
			chr6	31732715	8862	0.45
			chr6	31732842	8735	0.45
			chr6	31741664	87	0.45
chr6	31741664	C6orf15	chr6	31728025	13639	0.45
			chr6	31729766	11898	0.45
			chr6	31732715	8949	0.45
			chr6	31732842	8822	0.45
			chr6	31741577	87	0.45
chr6	32005470	HLA-B	chr6	32001179	4291	0.75
			chr6	32002459	3011	0.75
			chr6	32002598	2872	0.75
			chr6	32003472	1998	0.75
chr6	33325789	HLA-DQA1	chr6	33320449	5340	0.35
			chr6	33320577	5212	0.45
			chr6	33320449	5340	0.35
			chr6	33320577	5212	0.45
			chr6	33320848	4941	0.45
			chr6	33320979	4810	0.40
			chr6	33321574	4215	0.40
			chr6	33321743	4046	0.40
			chr6	33321753	4036	0.40
			chr6	33321774	4015	0.40
			chr6	33321791	3998	0.40
			chr6	33321992	3797	0.40
			chr6	33322151	3638	0.40
			chr6	33322166	3623	0.40
			chr6	33322664	3125	0.40

Table J.1: The distance of shared coding polymorphisms to each other shared polymorphism involved in the haplotype structure is given together with their alternate allele frequencies (AF).

Shared Coding SNP			Statistics			
Chr	Pos	Gene	SNP_{Chr}	SNP_{Pos}	Distance (in bp)	AF
chr6	33325789	HLA-DQA1	chr6	33323323	2466	0.45
			chr6	33323432	2357	0.40
			chr6	33323775	2014	0.40
			chr6	33324132	1657	0.35
			chr6	33325081	708	0.40
			chr6	33325932	143	0.45
			chr6	33326110	321	0.40
			chr6	33326228	439	0.40
			chr6	33327537	1748	0.40
			chr6	33327793	2004	0.40
			chr6	33328577	2788	0.40
			chr6	33328621	2832	0.40
			chr6	33328884	3095	0.45
			chr6	33328934	3145	0.40
			chr6	33329086	3297	0.40
			chr6	33329500	3711	0.40
chr6	33330263	HLA-DQA1	chr6	33331125	862	0.80
			chr6	33331357	1094	0.80
chr6	33350080	HLA-DQB1	chr6	33338392	11688	0.65
			chr6	33338471	11609	0.60
			chr6	33338508	11572	0.60
			chr6	33338736	11344	0.60
			chr6	33339418	10662	0.65
			chr6	33339490	10590	0.65
			chr6	33340917	9163	0.65
			chr6	33342982	7098	0.65
			chr6	33343330	6750	0.60
			chr6	33343697	6383	0.60
			chr6	33344499	5581	0.60
			chr6	33344507	5573	0.65
			chr6	33344998	5082	0.65
			chr6	33345322	4758	0.60
			chr6	33346326	3754	0.60
			chr6	33346623	3457	0.65
			chr6	33346637	3443	0.60
			chr6	33346681	3399	0.60
			chr6	33349538	542	0.60
			chr6	33349582	498	0.60
			chr6	33349763	317	0.60
			chr6	33350136	56	0.60
			chr6	33350154	74	0.60
			chr6	33350382	302	0.60
			chr6	33350403	323	0.65
			chr6	33350485	405	0.60
			chr6	33350586	506	0.60
			chr6	33350821	741	0.60
			chr6	33351171	1091	0.60
			chr6	33351218	1138	0.60
			chr6	33351419	1339	0.65
			chr6	33351508	1428	0.60
			chr6	33351944	1864	0.60
			chr6	33352231	2151	0.60
			chr6	33352754	2674	0.65
			chr6	33352941	2861	0.65
			chr6	33353002	2922	0.60
			chr6	33353065	2985	0.60
			chr6	33353324	3244	0.60
			chr6	33353500	3420	0.60

Table J.1

Shared Coding SNP			Statistics			
Chr	**Pos**	**Gene**	**SNP_{Chr}**	**SNP_{Pos}**	**Distance (in bp)**	**AF**
chr6	33782690	HLA-DPB1	chr6	33781555	1135	0.95
			chr6	33781760	930	0.95
			chr6	33782160	530	0.95
			chr6	33782980	290	0.95
			chr6	33783187	497	0.95
			chr6	33783414	724	0.95
			chr6	33784407	1717	0.95
			chr6	33784622	1932	0.95
			chr6	33785086	2396	0.95
chr6	33783414	HLA-DPB1	chr6	33781555	1859	0.95
			chr6	33781760	1654	0.95
			chr6	33782160	1254	0.95
			chr6	33782690	724	0.95
			chr6	33782980	434	0.95
			chr6	33783187	227	0.95
			chr6	33784407	993	0.95
			chr6	33784622	1208	0.95
			chr6	33785086	1672	0.95
chr7	99186425	PTCD1	chr7	99185969	456	0.45
			chr7	99189477	3052	0.35
chr7	132962782	PLXNA4	chr7	132960131	2651	0.95
chr8	123538356	FER1L6	chr8	123542996	4640	0.90
chr8	133143752	ST3GAL1	chr8	133144055	303	0.75
			chr8	133144651	899	0.75
chr9	96876074	NCBP1	chr9	96871801	4273	0.95
chr9	133338093	ABO	chr9	133338408	315	0.75
chr11	6395029	DNHD1	chr11	6392589	2440	0.85
chr12	81277774	PTPRQ	chr12	81278844	1070	0.95
chr13	27322449	POLR1D	chr13	27336275	13826	0.10
chr15	23144949	ATP10A	chr15	23141956	2993	0.60
chr16	57099706	GPR56	chr16	57101433	1727	0.60
			chr16	57107827	8121	0.40
chr17	21239943	RDM1	chr17	21252207	12264	0.20
chr19	8742743	PRAM1	chr19	8744260	1517	0.95
chr19	10624440	TYK2	chr19	10623478	962	0.75
chr19	14820533	CD97	chr19	14823497	2964	0.60

Table J.1

Bibliography

Abecasis, G. R., Noguchi, E., Heinzmann, A., Traherne, J. A., Bhattacharyya, S., Leaves, N. I., Anderson, G. G., Zhang, Y., Lench, N. J., Carey, A., Cardon, L. R., Moffatt, M. F., and Cookson, W. O. (2001). Extent and distribution of linkage disequilibrium in three genomic regions. *Am. J. Hum. Genet.*, 68:191–197. (Cited on pages 18 and 156.)

Adams, E. J., Cooper, S., Thomson, G., and Parham, P. (2000). Common chimpanzees have greater diversity than humans at two of the three highly polymorphic MHC class I genes. *Immunogenetics*, 51:410–424. (Cited on page 112.)

Agnez-Lima, L. F., Napolitano, R. L., Fuchs, R. P., Mascio, P. D., Muotri, A. R., and Menck, C. F. (2001). DNA repair and sequence context affect $^1O_2$-induced mutagenesis in bacteria. *Nucleic Acids Res.*, 29:2899–2903. (Cited on page 9.)

Albert, T. J., Molla, M. N., Muzny, D. M., Nazareth, L., Wheeler, D., Song, X., Richmond, T. A., Middle, C. M., Rodesch, M. J., Packard, C. J., Weinstock, G. M., and Gibbs, R. A. (2007). Direct selection of human genomic loci by microarray hybridization. *Nat. Methods*, 4:903–905. (Cited on page 21.)

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410. (Cited on page 27.)

Andolfatto, P. (2001). Adaptive hitchhiking effects on genome variability. *Curr. Opin. Genet. Dev.*, 11:635–641. (Cited on page 122.)

Anzai, T., Shiina, T., Kimura, N., Yanagiya, K., Kohara, S., Shigenari, A., Yamagata, T., Kulski, J. K., Naruse, T. K., Fujimori, Y., Fukuzumi, Y., Yamazaki, M., Tashiro, H., Iwamoto, C., Umehara, Y., Imanishi, T., Meyer, A., Ikeo, K., Gojobori, T., Bahram, S., and Inoko, H. (2003). Comparative sequencing of human and chimpanzee MHC class I regions unveils insertions/deletions as the major path to genomic divergence. *Proc. Natl. Acad. Sci. U.S.A.*, 100:7708–7713. (Cited on page 71.)

Applied Biosystems (2011). 5500 Series Genetic Analysis System specification sheet. Technical report, Applied Biosystems. Available from: `http://media.invitrogen.com.edgesuite.net/solid/pdf/CO18235-5500-Series-Spec-Sheet-F2.pdf`. (Cited on pages 188 and 196.)

Aquadro, C. F., Bauer DuMont, V., and Reed, F. A. (2001). Genome-wide variation in the human and fruitfly: a comparison. *Curr. Opin. Genet. Dev.*, 11:627–634. (Cited on page 122.)

Arden, B. and Klein, J. (1982). Biochemical comparison of major histocompatibility complex molecules from different subspecies of Mus musculus: evidence for trans-specific evolution of alleles. *Proc. Natl. Acad. Sci. U.S.A.*, 79:2342–2346. (Cited on page 164.)

Ardlie, K. G., Kruglyak, L., and Seielstad, M. (2002). Patterns of linkage disequilibrium in the human genome. *Nat. Rev. Genet.*, 3:299–309. (Cited on page 18.)

Arndt, P. F., Hwa, T., and Petrov, D. A. (2005). Substantial regional variation in substitution rates in the human genome: importance of GC content, gene density, and telomere-specific effects. *J. Mol. Evol.*, 60:748–763. (Cited on pages 10 and 13.)

Arnheim, N. and Calabrese, P. (2009). Understanding what determines the frequency and pattern of human germline mutations. *Nat. Rev. Genet.*, 10:478–488. (Cited on pages 68 and 174.)

Asthana, S., Noble, W. S., Kryukov, G., Grant, C. E., Sunyaev, S., and Stamatoyannopoulos, J. A. (2007). Widely distributed noncoding purifying selection in the human genome. *Proc. Natl. Acad. Sci. U.S.A.*, 104:12410–12415. (Cited on page 152.)

Astier, Y., Braha, O., and Bayley, H. (2006). Toward single molecule DNA sequencing: direct identification of ribonucleoside and deoxyribonucleoside 5'-monophosphates by using an engineered protein nanopore equipped with a molecular adapter. *J. Am. Chem. Soc.*, 128:1705–1710. (Cited on page 204.)

Auton, A., Fledel-Alon, A., Pfeifer, S., Venn, O., Segurel, L., Street, T., Leffler, E. M., Bowden, R., Aneas, I., Broxholme, J., Humburg, P., Iqbal, Z., Lunter, G., Maller, J., Hernandez, R. D., Melton, C., Venkat, A., Nobrega, M. A., Bontrop, R., Myers, S., Donnelly, P., Przeworski, M., and McVean, G. (2012). A fine-scale chimpanzee genetic map from population sequencing. *Science*, 336:193–198. (Cited on page 125.)

Avila, V., Chavarrias, D., Sanchez, E., Manrique, A., Lopez-Fanjul, C., and Garcia-Dorado, A. (2006). Increase of the spontaneous mutation rate in a long-term experiment with Drosophila melanogaster. *Genetics*, 173:267–277. (Cited on page 15.)

Ayala, F. J., Escalante, A., O'Huigin, C., and Klein, J. (1994). Molecular genetics of speciation and human origins. *Proc. Natl. Acad. Sci. U.S.A.*, 91:6787–6794. (Cited on pages 137, 143, 157, and 166.)

Baer, C. F., Miyamoto, M. M., and Denver, D. R. (2007). Mutation rate variation in multicellular eukaryotes: causes and consequences. *Nat. Rev. Genet.*, 8:619–631. (Cited on page 3.)

Bailey, J. A., Gu, Z., Clark, R. A., Reinert, K., Samonte, R. V., Schwartz, S., Adams, M. D., Myers, E. W., Li, P. W., and Eichler, E. E. (2002). Recent segmental duplications in the human genome. *Science*, 297:1003–1007. (Cited on page 5.)

Bashiardes, S., Veile, R., Helms, C., Mardis, E. R., Bowcock, A. M., and Lovett, M. (2005). Direct genomic selection. *Nat. Methods*, 2:63–69. (Cited on page 21.)

Batzoglou, S., Jaffe, D. B., Stanley, K., Butler, J., Gnerre, S., Mauceli, E., Berger, B., Mesirov, J. P., and Lander, E. S. (2002). ARACHNE: a whole-genome shotgun assembler. *Genome research*, 12:177–189. (Cited on page 72.)

Baujat, G., Legeai-Mallet, L., Finidori, G., Cormier-Daire, V., and Le Merrer, M. (2008). Achondroplasia. *Best Pract Res Clin Rheumatol*, 22:3–18. (Cited on pages 50 and 59.)

Becquet, C., Patterson, N., Stone, A. C., Przeworski, M., and Reich, D. (2007). Genetic structure of chimpanzee populations. *PLoS Genet.*, 3:e66. (Cited on page 71.)

Becquet, C. and Przeworski, M. (2007). A new approach to estimate parameters of speciation models with application to apes. *Genome Res.*, 17:1505–1519. (Cited on page 71.)

Begun, D. J. and Aquadro, C. F. (1994). Evolutionary inferences from DNA variation at the 6-phosphogluconate dehydrogenase locus in natural populations of drosophila: selection and geographic differentiation. *Genetics*, 136:155–171. (Cited on page 122.)

Bellus, G. A., Spector, E. B., Speiser, P. W., Weaver, C. A., Garber, A. T., Bryke, C. R., Israel, J., Rosengren, S. S., Webster, M. K., Donoghue, D. J., and Francomano, C. A. (2000). Distinct missense mutations of the FGFR3 lys650 codon modulate receptor kinase activation and the severity of the skeletal dysplasia phenotype. *Am. J. Hum. Genet.*, 67:1411–1421. (Cited on page 53.)

Bentley, D. R. (2006). Whole-genome re-sequencing. *Curr. Opin. Genet. Dev.*, 16:545–552. (Cited on page 188.)

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., Boutell, J. M., Bryant, J., Carter, R. J., Keira Cheetham, R., Cox, A. J., Ellis, D. J., Flatbush, M. R., Gormley, N. A., Humphray, S. J., Irving, L. J., Karbelashvili, M. S., Kirk, S. M., Li, H., Liu, X., Maisinger, K. S., Murray, L. J., Obradovic, B., Ost, T., Parkinson, M. L., Pratt, M. R., Rasolonjatovo, I. M., Reed, M. T., Rigatti, R., Rodighiero, C., Ross, M. T., Sabot, A., Sankar, S. V., Scally, A., Schroth, G. P., Smith, M. E., Smith, V. P., Spiridou, A., Torrance, P. E., Tzonev, S. S., Vermaas, E. H., Walter, K., Wu, X., Zhang, L., Alam, M. D., Anastasi, C., Aniebo, I. C., Bailey, D. M., Bancarz, I. R., Banerjee, S., Barbour, S. G., Baybayan, P. A., Benoit, V. A., Benson, K. F., Bevis, C., Black, P. J., Boodhun, A., Brennan, J. S., Bridgham, J. A., Brown, R. C., Brown, A. A., Buermann, D. H., Bundu, A. A., Burrows, J. C., Carter, N. P., Castillo, N., Chiara E Catenazzi, M., Chang, S., Neil Cooley, R., Crake, N. R., Dada, O. O., Diakoumakos, K. D., Dominguez-Fernandez, B., Earnshaw, D. J., Egbujor, U. C., Elmore, D. W., Etchin, S. S., Ewan, M. R., Fedurco, M., Fraser, L. J., Fuentes Fajardo, K. V., Scott Furey, W., George, D., Gietzen, K. J., Goddard, C. P., Golda, G. S., Granieri, P. A., Green, D. E., Gustafson, D. L., Hansen, N. F., Harnish, K., Haudenschild, C. D., Heyer, N. I., Hims, M. M., Ho, J. T., Horgan, A. M., Hoschler, K., Hurwitz, S., Ivanov, D. V., Johnson, M. Q., James, T., Huw Jones, T. A., Kang, G. D., Kerelska, T. H., Kersey, A. D., Khrebtukova, I., Kindwall, A. P., Kingsbury, Z., Kokko-Gonzales, P. I., Kumar, A., Laurent, M. A., Lawley, C. T., Lee, S. E., Lee, X., Liao, A. K., Loch, J. A., Lok, M., Luo, S., Mammen, R. M., Martin, J. W., McCauley, P. G., McNitt, P., Mehta, P., Moon, K. W., Mullens, J. W., Newington, T., Ning, Z., Ling Ng, B., Novo, S. M., O'Neill, M. J., Osborne, M. A., Osnowski, A., Ostadan, O., Paraschos, L. L., Pickering, L., Pike, A. C., Pike, A. C., Chris Pinkard, D., Pliskin, D. P., Podhasky, J., Quijano, V. J., Raczy, C., Rae, V. H., Rawlings, S. R., Chiva Rodriguez, A., Roe, P. M., Rogers, J., Rogert Bacigalupo, M. C., Romanov, N., Romieu, A., Roth, R. K., Rourke, N. J., Ruediger, S. T., Rusman, E., Sanches-Kuiper, R. M., Schenker, M. R., Seoane, J. M., Shaw, R. J., Shiver, M. K., Short, S. W., Sizto, N. L., Sluis, J. P., Smith, M. A., Ernest Sohna Sohna, J., Spence, E. J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C. L., Turcatti, G., Vandevondele, S., Verhovsky, Y., Virk, S. M., Wakelin, S., Walcott, G. C., Wang, J., Worsley, G. J., Yan, J., Yau, L., Zuerlein, M., Rogers, J., Mullikin, J. C., Hurles, M. E., McCooke, N. J., West, J. S., Oaks, F. L., Lundberg, P. L., Klenerman, D., Durbin, R., and Smith, A. J. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456:53–59. (Cited on page 24.)

Bergstrom, T. F., Erlandsson, R., Engkvist, H., Josefsson, A., Erlich, H. A., and Gyllensten, U.

(1999). Phylogenetic history of hominoid DRB loci and alleles inferred from intron sequences. *Immunol. Rev.*, 167:351–365. (Cited on page 136.)

Berk, D. R., Spector, E. B., and Bayliss, S. J. (2007). Familial acanthosis nigricans due to K650T FGFR3 mutation. *Arch Dermatol*, 143:1153–1156. (Cited on page 66.)

Betancourt, A. J., Kim, Y., and Orr, H. A. (2004). A pseudohitchhiking model of X vs. autosomal diversity. *Genetics*, 168:2261–2269. (Cited on page 119.)

Bhangale, T. R., Rieder, M. J., Livingston, R. J., and Nickerson, D. A. (2005). Comprehensive identification and characterization of diallelic insertion-deletion polymorphisms in 330 human candidate genes. *Hum. Mol. Genet.*, 14:59–69. (Cited on page 4.)

Bird, A. P. (1986). CpG-rich islands and the function of DNA methylation. *Nature*, 321:209–213. (Cited on page 10.)

Bjorkman, P. J., Saper, M. A., Samraoui, B., Bennett, W. S., Strominger, J. L., and Wiley, D. C. (1987). The foreign antigen binding site and T cell recognition regions of class I histocompatibility antigens. *Nature*, 329:512–518. (Cited on pages 137, 143, 157, and 166.)

Blake, R. D., Hess, S. T., and Nicholson-Tuell, J. (1992). The influence of nearest neighbors on the rate and pattern of spontaneous point mutations. *J. Mol. Evol.*, 34:189–200. (Cited on pages 8, 9, 10, 113, 117, and 145.)

Blazej, R. G., Kumaresan, P., Cronier, S. A., and Mathies, R. A. (2007). Inline injection microdevice for attomole-scale sanger DNA sequencing. *Anal. Chem.*, 79:4499–4506. (Cited on page 186.)

Blazej, R. G., Kumaresan, P., and Mathies, R. A. (2006). Microfabricated bioprocessor for integrated nanoliter-scale Sanger DNA sequencing. *Proc. Natl. Acad. Sci. U.S.A.*, 103:7240–7245. (Cited on page 186.)

Boerma, E. G., Siebert, R., Kluin, P. M., and Baudis, M. (2009). Translocations involving 8q24 in Burkitt lymphoma and other malignant lymphomas: a historical review of cytogenetics in the light of todays knowledge. *Leukemia*, 23:225–234. (Cited on page 6.)

Bohossian, H. B., Skaletsky, H., and Page, D. C. (2000). Unexpectedly similar rates of nucleotide substitution found in male and female hominids. *Nature*, 406:622–625. (Cited on page 14.)

Bontrop, R. E. (2006). Comparative genetics of MHC polymorphisms in different primate species: duplications and deletions. *Hum. Immunol.*, 67:388–397. (Cited on page 164.)

Bontrop, R. E., Otting, N., de Groot, N. G., and Doxiadis, G. G. (1999). Major histocompatibility complex class II polymorphisms in primates. *Immunol. Rev.*, 167:339–350. (Cited on pages 165 and 169.)

Bontrop, R. E., Otting, N., Slierendregt, B. L., and Lanchbury, J. S. (1995). Evolution of major histocompatibility complex polymorphisms and T-cell receptor diversity in primates. *Immunol. Rev.*, 143:33–62. (Cited on page 112.)

Boulikas, T. (1992). Evolutionary consequences of nonrandom damage and repair of chromatin domains. *J. Mol. Evol.*, 35:156–180. (Cited on pages 11 and 13.)

Bowers, J., Mitchell, J., Beer, E., Buzby, P. R., Causey, M., Efcavitch, J. W., Jarosz, M., Krzymanska-Olejnik, E., Kung, L., Lipson, D., Lowman, G. M., Marappan, S., McInerney, P., Platt, A., Roy, A., Siddiqi, S. M., Steinmann, K., and Thompson, J. F. (2009). Virtual terminator nucleotides for next-generation DNA sequencing. *Nat. Methods*, 6:593–595. (Cited on page 199.)

Boyle, A. P., Song, L., Lee, B. K., London, D., Keefe, D., Birney, E., Iyer, V. R., Crawford, G. E., and Furey, T. S. (2011). High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Res.*, 21:456–464. (Cited on page 12.)

Brookes, A. J. (2008). *Handbook of Human Molecular Evolution*, chapter Single Nucleotide Polymorphism (SNP), pages 588–591. Wiley. (Cited on page 123.)

Brooks, L. D. and Marks, R. W. (1986). The organization of genetic variation for recombination in Drosophila melanogaster. *Genetics*, 114:525–547. (Cited on page 17.)

Brooks, S. and Roberts, G. (1999). On quantile estimation and MCMC convergence. *Biometrika*, 86:710–717. (Cited on page 62.)

Brooks, S. P. (1998). Markov chain monte carlo method and its application. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 47:69–100. (Cited on page 61.)

Brown, J. H., Jardetzky, T., Saper, M. A., Samraoui, B., Bjorkman, P. J., and Wiley, D. C. (1988). A hypothetical model of the foreign antigen binding site of class II histocompatibility molecules. *Nature*, 332:845–850. (Cited on pages 137, 143, 157, and 166.)

Brown, T. C. and Jiricny, J. (1988). Different base/base mispairs are corrected with different efficiencies and specificities in monkey kidney cells. *Cell*, 54:705–711. (Cited on page 3.)

Browning, B. and Yu, Z. (2009). Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. *The American Journal of Human Genetics*, 85:847–861. (Cited on pages 44 and 101.)

Browning, S. R. (2006). Multilocus association mapping using variable-length Markov chains. *Am. J. Hum. Genet.*, 78:903–913. (Cited on pages 44 and 101.)

Browning, S. R. and Browning, B. L. (2007). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.*, 81:1084–1097. (Cited on pages 44, 101, and 102.)

Brunet, M., Guy, F., Pilbeam, D., Mackaye, H. T., Likius, A., Ahounta, D., Beauvilain, A., Blondel, C., Bocherens, H., Boisserie, J. R., De Bonis, L., Coppens, Y., Dejax, J., Denys, C., Duringer, P., Eisenmann, V., Fanone, G., Fronty, P., Geraads, D., Lehmann, T., Lihoreau, F., Louchart, A., Mahamat, A., Merceron, G., Mouchelin, G., Otero, O., Pelaez Campomanes, P., Ponce De Leon, M., Rage, J. C., Sapanet, M., Schuster, M., Sudre, J., Tassy, P., Valentin, X., Vignaud, P., Viriot, L., Zazzo, A., and Zollikofer, C. (2002). A new hominid from the Upper Miocene of Chad, Central Africa. *Nature*, 418:145–151. (Cited on pages 71 and 135.)

Buhler, J. (2001). Efficient large-scale sequence comparison by locality-sensitive hashing. *Bioinformatics*, 17:419–428. (Cited on page 29.)

Bunin, G. R., Needle, M., and Riccardi, V. M. (1997). Paternal age and sporadic neurofibromatosis 1: a case-control study and consideration of the methodologic issues. *Genet. Epidemiol.*, 14:507–516. (Cited on page 49.)

Burrows, M. and Wheeler, D. J. (1994). A block-sorting lossless data compression algorithm. Technical report, Digital Equipment Corporation. (Cited on page 35.)

Cagliani, R., Fumagalli, M., Biasin, M., Piacentini, L., Riva, S., Pozzoli, U., Bonaglia, M. C., Bresolin, N., Clerici, M., and Sironi, M. (2010). Long-term balancing selection maintains trans-specific polymorphisms in the human TRIM5 gene. *Hum. Genet.*, 128:577–588. (Cited on pages 137, 157, 165, and 166.)

Campagna, D., Albiero, A., Bilardi, A., Caniato, E., Forcato, C., Manavski, S., Vitulo, N., and Valle, G. (2009). PASS: a program to align short sequences. *Bioinformatics*, 25:967–968. (Cited on page 29.)

Cannon, G. B. (1963). The effects of natural selection on linkage disequilibrium and relative fitness in experimental populations of Drosophila melanogaster. *Genetics*, 48:1201–1216. (Cited on page 18.)

Carlson, K. M., Bracamontes, J., Jackson, C. E., Clark, R., Lacroix, A., Wells, S. A., and Goodfellow, P. J. (1994). Parent-of-origin effects in multiple endocrine neoplasia type 2B. *Am. J. Hum. Genet.*, 55:1076–1082. (Cited on page 49.)

Carothers, A. M., Urlaub, G., Mucha, J., Grunberger, D., and Chasin, L. A. (1989). Point mutation analysis in a mammalian gene: rapid preparation of total RNA, PCR amplification of cDNA, and Taq sequencing by a novel method. *BioTechniques*, 7:494–496. (Cited on page 184.)

Castro-Feijo, L., Loidi, L., Vidal, A., Parajes, S., Rosn, E., Alvarez, A., Cabanas, P., Barreiro, J., Alonso, A., Domnguez, F., and Pombo, M. (2008). Hypochondroplasia and Acanthosis nigricans: a new syndrome due to the p.Lys650Thr mutation in the fibroblast growth factor receptor 3 gene? *Eur. J. Endocrinol.*, 159:243–249. (Cited on pages 53 and 66.)

Caswell, J. L., Mallick, S., Richter, D. J., Neubauer, J., Schirmer, C., Gnerre, S., and Reich, D. (2008). Analysis of chimpanzee history based on genome sequence alignments. *PLoS Genet.*, 4:e1000057. (Cited on page 71.)

Chaisson, M. J., Brinza, D., and Pevzner, P. A. (2009). De novo fragment assembly with short mate-paired reads: Does the read length matter? *Genome Res.*, 19:336–346. (Cited on page 187.)

Chang, B. H. and Li, W. H. (1995). Estimating the intensity of male-driven evolution in rodents by using X-linked and Y-linked Ube 1 genes and pseudogenes. *J. Mol. Evol.*, 40:70–77. (Cited on page 13.)

Charlesworth, B., Coyne, J. A., and Barton, N. H. (1987). The relative rates of evolution sex-chromosomes and autosomes. *Am. Nat.*, 130:113–146. (Cited on page 119.)

Charlesworth, B., Morgan, M. T., and Charlesworth, D. (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics*, 134:1289–1303. (Cited on page 122.)

Charlesworth, D., Charlesworth, B., and Morgan, M. T. (1995). The pattern of neutral molecular variation under the background selection model. *Genetics*, 141:1619–1632. (Cited on page 121.)

Chen, C. L., Rappailles, A., Duquenne, L., Huvet, M., Guilbaud, G., Farinelli, L., Audit, B., d'Aubenton Carafa, Y., Arneodo, A., Hyrien, O., and Thermes, C. (2010). Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. *Genome Res.*, 20:447–457. (Cited on pages 10, 12, and 13.)

Chen, F. C. and Li, W. H. (2001). Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am. J. Hum. Genet.*, 68:444–456. (Cited on pages 71 and 135.)

Chen, F. C., Vallender, E. J., Wang, H., Tzeng, C. S., and Li, W. H. (2001). Genomic divergence between human and chimpanzee estimated from large-scale alignments of genomic sequences. *J. Hered.*, 92:481–489. (Cited on pages 8, 113, and 117.)

Chen, Y., Souaiaia, T., and Chen, T. (2009). PerM: efficient mapping of short sequencing reads with periodic full sensitive spaced seeds. *Bioinformatics*, 25:2514–2521. (Cited on page 29.)

Cheung, J., Estivill, X., Khaja, R., MacDonald, J. R., Lau, K., Tsui, L. C., and Scherer, S. W. (2003). Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. *Genome Biol.*, 4:R25. (Cited on pages 5 and 6.)

Chimpanzee Sequencing and Analysis Consortium (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437:69–87. (Cited on pages 9, 10, 12, 14, 15, 72, 73, 77, 79, 97, 107, 112, 113, 117, 118, 122, 131, 144, and 176.)

Clark, A. G. (1997). Neutral behavior of shared polymorphism. *Proc. Natl. Acad. Sci. U.S.A.*, 94:7730–7734. (Cited on pages 122 and 136.)

Clark, A. G., Glanowski, S., Nielsen, R., Thomas, P. D., Kejariwal, A., Todd, M. A., Tanenbaum, D. M., Civello, D., Lu, F., Murphy, B., Ferriera, S., Wang, G., Zheng, X., White, T. J., Sninsky, J. J., Adams, M. D., and Cargill, M. (2003). Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science*, 302:1960–1963. (Cited on page 71.)

Clarke, B. and Kirby, D. R. (1966). Maintenance of histocompatibility polymorphisms. *Nature*, 211:999–1000. (Cited on page 136.)

Clarke, J., Wu, H., Jayasinghe, L., Patel, A., Reid, S., and Bayley, H. (2009). Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnology*, 4:265–270. (Cited on pages 203, 204, and 205.)

Clemens, R. A., Newbrough, S. A., Chung, E. Y., Gheith, S., Singer, A. L., Koretzky, G. A., and Peterson, E. J. (2004). PRAM-1 is required for optimal integrin-dependent neutrophil function. *Mol. Cell. Biol.*, 24:10923–10932. (Cited on page 168.)

Clement, N. L., Snell, Q., Clement, M. J., Hollenhorst, P. C., Purwar, J., Graves, B. J., Cairns, B. R., and Johnson, W. E. (2010). The GNUMAP algorithm: unbiased probabilistic mapping of oligonucleotides from next-generation sequencing. *Bioinformatics*, 26:38–45. (Cited on page 29.)

Coetzee, B. (2010). A metagenomic approach using next-generation sequencing for viral profiling of a vineyard and genetic characterization of Grapevine virus Evolution. Master's thesis, Stellenbosch University, South Africa. (Cited on page 25.)

Cohen, N. M., Kenigsberg, E., and Tanay, A. (2011). Primate CpG islands are maintained by heterogeneous evolutionary regimes involving minimal selection. *Cell*, 145:773–786. (Cited on pages 10 and 11.)

Collins, A., Lonjou, C., and Morton, N. E. (1999). Genetic epidemiology of single-nucleotide polymorphisms. *Proc. Natl. Acad. Sci. U.S.A.*, 96:15173–15177. (Cited on page 18.)

Collins, F. S., Drumm, M. L., Cole, J. L., Lockwood, W. K., Vande Woude, G. F., and Iannuzzi, M. C. (1987). Construction of a general human chromosome jumping library, with application to cystic fibrosis. *Science*, 235:1046–1049. (Cited on page 5.)

Coombs, A. (2008). The sequencing shakeup. *Nat. Biotechnol.*, 26:1109–1112. (Cited on pages 21 and 51.)

Coop, G. and Przeworski, M. (2007). An evolutionary view of human recombination. *Nat. Rev. Genet.*, 8:23–34. (Cited on page 17.)

Cooper, D. N. and Krawczak, M. (1990). The mutational spectrum of single base-pair substitutions causing human genetic disease: patterns and predictions. *Hum. Genet.*, 85:55–74. (Cited on page 9.)

Cooper, D. N. and Youssoufian, H. (1988). The CpG dinucleotide and human genetic disease. *Hum. Genet.*, 78:151–155. (Cited on pages 9, 117, 122, 131, 145, and 176.)

Cooper, G., Rubinsztein, D. C., and Amos, W. (1998). Ascertainment bias cannot entirely account for human microsatellites being longer than their chimpanzee homologues. *Hum. Mol. Genet.*, 7:1425–1429. (Cited on page 112.)

Coulondre, C., Miller, J. H., Farabaugh, P. J., and Gilbert, W. (1978). Molecular basis of base substitution hotspots in Escherichia coli. *Nature*, 274:775–780. (Cited on pages 9, 117, 122, 131, and 176.)

Cowles, M. and Carlin, B. (1996). Markov chain monte carlo convergence diagnostics: a comparative review. *J. Amer. Statist. Assoc.*, 91:883–904. (Cited on page 62.)

Cox, A. J. (2007). ELAND: Efficient large-scale alignment of nucleotide databases. Illumina, San Diego. (Cited on page 29.)

Craxton, M. (1993). Cosmid sequencing. *Methods Mol. Biol.*, 23:149–167. (Cited on page 184.)

Crow, J. (2000). The origins, patterns and implications of human spontaneous mutation. *Nat. Rev. Genet.*, 1:40–47. (Cited on pages 14 and 68.)

Crow, J. F. (1993). How much do we know about spontaneous human mutation rates? *Environ. Mol. Mutagen.*, 21:122–129. (Cited on page 15.)

Dahl, F., Stenberg, J., Fredriksson, S., Welch, K., Zhang, M., Nilsson, M., Bicknell, D., Bodmer, W. F., Davis, R. W., and Ji, H. (2007). Multigene amplification and massively parallel sequencing for cancer mutation discovery. *Proc. Natl. Acad. Sci. U.S.A.*, 104:9387–9392. (Cited on page 21.)

Dai, J. Y., Ruczinski, I., LeBlanc, M., and Kooperberg, C. (2006). Imputation methods to improve inference in SNP association studies. *Genet. Epidemiol.*, 30:690–702. (Cited on page 44.)

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., Depristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., and 1000 Genomes Project Analysis Group (2011). The variant call format and VCFtools. *Bioinformatics*, 27:2156–2158. (Cited on pages 96, 156, and 217.)

Darwin, C. (1871). *The Descent of Man and Selection in Relation to Sex*. D. Appleton and Company, New York. (Cited on page III.)

Datta, A. and Jinks-Robertson, S. (1995). Association of increased spontaneous mutation rates with high levels of transcription in yeast. *Science*, 268:1616–1619. (Cited on page 11.)

de Bakker, P. I., McVean, G., Sabeti, P. C., Miretti, M. M., Green, T., Marchini, J., Ke, X., Monsuur, A. J., Whittaker, P., Delgado, M., Morrison, J., Richardson, A., Walsh, E. C., Gao, X., Galver, L., Hart, J., Hafler, D. A., Pericak-Vance, M., Todd, J. A., Daly, M. J., Trowsdale, J., Wijmenga, C., Vyse, T. J., Beck, S., Murray, S. S., Carrington, M., Gregory, S., Deloukas, P., and Rioux, J. D. (2006). A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nat. Genet.*, 38:1166–1172. (Cited on page 164.)

de Groot, N. G. and Bontrop, R. E. (1999). The major histocompatibility complex class II region of the chimpanzee: towards a molecular map. *Immunogenetics*, 50:160–167. (Cited on page 164.)

Deamer, D. W., Akeson, M., and Deamer, D. W. (2000). Nanopores and nucleic acids: prospects for ultrarapid sequencing. *Trends Biotechnol.*, 18:147–151. (Cited on page 203.)

Deinard, A. and Kidd, K. (1999). Evolution of a HOXB6 intergenic region within the great apes and humans. *J. Hum. Evol.*, 36:687–703. (Cited on page 112.)

Deinard, A. S. and Kidd, K. K. (1998). Evolution of a D2 dopamine receptor intron within the great apes and humans. *DNA Seq.*, 8:289–301. (Cited on page 112.)

Deng, H. W., Li, J., Pfrender, M. E., Li, J. L., and Deng, H. (2006). Upper limit of the rate and per generation effects of deleterious genomic mutations. *Genet. Res.*, 88:57–65. (Cited on page 17.)

Deng, H. W. and Lynch, M. (1996). Estimation of deleterious-mutation parameters in natural populations. *Genetics*, 144:349–360. (Cited on page 17.)

DePristo, M., Banks, E., Poplin, R., Garimella, K., Maguire, J., Hartl, C., Philippakis, A., del Angel, G., Rivas, M., Hanna, M., McKenna, A., Fennell, T., Kernytsky, A., Sivachenko, A., Cibulskis, K., Gabriel, S., Altshuler, D., and Daly, M. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43:491–498. (Cited on pages 41, 44, 82, 83, 84, 87, 90, and 114.)

Dermitzakis, E. T., Reymond, A., and Antonarakis, S. E. (2005). Conserved non-genic sequences - an unexpected feature of mammalian genomes. *Nat. Rev. Genet.*, 6:151–157. (Cited on page 152.)

Dermitzakis, E. T., Reymond, A., Lyle, R., Scamuffa, N., Ucla, C., Deutsch, S., Stevenson, B. J., Flegel, V., Bucher, P., Jongeneel, C. V., and Antonarakis, S. E. (2002). Numerous potentially functional but non-genic conserved sequences on human chromosome 21. *Nature*, 420:578–582. (Cited on pages 17 and 152.)

Doherty, P. C. and Zinkernagel, R. M. (1975). Enhanced immunological surveillance in mice heterozygous at the H-2 gene complex. *Nature*, 256:50–52. (Cited on page 136.)

Dohm, J. C., Lottaz, C., Borodina, T., and Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.*, 36:e105. (Cited on page 26.)

Drake, J. W. (1992). Mutation rates. *Bioessays*, 14:137–140. (Cited on pages 8 and 113.)

Drake, J. W., Charlesworth, B., Charlesworth, D., and Crow, J. F. (1998). Rates of spontaneous mutation. *Genetics*, 148:1667–1686. (Cited on page 15.)

Dressman, D., Yan, H., Traverso, G., Kinzler, K. W., and Vogelstein, B. (2003). Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc. Natl. Acad. Sci. U.S.A.*, 100:8817–8822. (Cited on pages 189 and 192.)

Droege, M. and Hill, B. (2008). The Genome Sequencer FLX System–longer reads, more applications, straight forward bioinformatics and more complete data sets. *J. Biotechnol.*, 136:3–10. (Cited on pages 188 and 189.)

Dunning, A. M., Durocher, F., Healey, C. S., Teare, M. D., McBride, S. E., Carlomagno, F., Xu, C. F., Dawson, E., Rhodes, S., Ueda, S., Lai, E., Luben, R. N., Van Rensburg, E. J., Mannermaa, A., Kataja, V., Rennart, G., Dunham, I., Purvis, I., Easton, D., and Ponder, B. A. (2000). The extent of linkage disequilibrium in four populations with distinct demographic histories. *Am. J. Hum. Genet.*, 67:1544–1554. (Cited on page 18.)

Duret, L. (2009). Mutation patterns in the human genome: more variable than expected. *PLoS Biol.*, 7:e1000028. (Cited on pages 2, 8, 14, 113, 117, 122, 131, 173, and 176.)

Duret, L. and Arndt, P. F. (2008). The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet.*, 4:e1000071. (Cited on pages 9, 12, and 13.)

Eaves, H. L. and Gao, Y. (2009). MOM: maximum oligonucleotide mapping. *Bioinformatics*, 25:969–970. (Cited on page 29.)

Ebersberger, I., Metzler, D., Schwarz, C., and Paabo, S. (2002). Genomewide comparison of DNA sequences between humans and chimpanzees. *Am. J. Hum. Genet.*, 70:1490–1497. (Cited on pages 8, 14, 15, 113, 117, 118, and 121.)

Edmonson, M. N., Zhang, J., Yan, C., Finney, R. P., Meerzaman, D. M., and Buetow, K. H. (2011). Bambino: a variant detector and alignment viewer for next-generation sequencing data in the SAM/BAM format. *Bioinformatics*, 27:865–866. (Cited on page 44.)

Ehrlich, M. and Wang, R. Y. (1981). 5-Methylcytosine in eukaryotic DNA. *Science*, 212:1350–1357. (Cited on pages 9, 117, 122, 131, 145, and 176.)

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., Dewinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J., and Turner, S. (2009). Real-time DNA sequencing from single polymerase molecules. *Science*, 323:133–138. (Cited on page 202.)

Eisen, J. A. and Hanawalt, P. C. (1999). A phylogenomic study of DNA repair genes, proteins, and processes. *Mutat. Res.*, 435:171–213. (Cited on page 3.)

El Adlouni, S., Favre, A., and Bobee, B. (2006). Comparison of methodologies to assess the convergence of Markov chain Monte Carlo methods. *Computational Statistics & Data Analysis*, 50:2685–2701. (Cited on page 62.)

Elango, N., Kim, S. H., Vigoda, E., and Yi, S. V. (2008). Mutations of different molecular origins exhibit contrasting patterns of regional substitution rate variation. *PLoS Comput. Biol.*, 4:e1000015. (Cited on pages 9, 10, 11, 12, and 13.)

Ellegren, H. (2007). Characteristics, causes and evolutionary consequences of male-biased mutation. *Proc. Biol. Sci.*, 274:1–10. (Cited on page 14.)

Ellegren, H. (2009). The different levels of genetic diversity in sex chromosomes and autosomes. *Trends Genet.*, 25:278–284. (Cited on page 119.)

Ellegren, H. and Fridolfsson, A. K. (1997). Male-driven evolution of DNA sequences in birds. *Nat. Genet.*, 17:182–184. (Cited on page 13.)

Ellegren, H., Smith, N. G., and Webster, M. T. (2003). Mutation rate variation in the mammalian genome. *Curr. Opin. Genet. Dev.*, 13:562–568. (Cited on pages 2, 7, 14, 113, and 173.)

Emanuel, B. S. and Shaikh, T. H. (2001). Segmental duplications: an 'expanding' role in genomic instability and disease. *Nat. Rev. Genet.*, 2:791–800. (Cited on page 5.)

Emrich, C. A., Tian, H., Medintz, I. L., and Mathies, R. A. (2002). Microfabricated 384-lane capillary array electrophoresis bioanalyzer for ultrahigh-throughput genetic analysis. *Anal. Chem.*, 74:5076–5083. (Cited on pages 185 and 186.)

Eory, L., Halligan, D. L., and Keightley, P. D. (2010). Distributions of selectively constrained sites and deleterious mutation rates in the hominid and murid genomes. *Mol. Biol. Evol.*, 27:177–192. (Cited on page 12.)

Erlich, Y., Mitra, P. P., delaBastide, M., McCombie, W. R., and Hannon, G. J. (2008). Alta-Cyclic: a self-optimizing base caller for next-generation sequencing. *Nat. Methods*, 5:679–682. (Cited on pages 26, 52, and 55.)

Ewing, B. and Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.*, 8:186–194. (Cited on pages 32, 185, and 207.)

Ewing, B., Hillier, L., Wendl, M. C., and Green, P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.*, 8:175–185. (Cited on pages 58, 185, and 190.)

Fan, W. M., Kasahara, M., Gutknecht, J., Klein, D., Mayer, W. E., Jonker, M., and Klein, J. (1989). Shared class II MHC polymorphisms between humans and chimpanzees. *Hum. Immunol.*, 26:107–121. (Cited on pages 122, 136, 165, and 169.)

Fan, Y., Linardopoulou, E., Friedman, C., Williams, E., and Trask, B. J. (2002a). Genomic structure and evolution of the ancestral chromosome fusion site in 2q13-2q14.1 and paralogous regions on other human chromosomes. *Genome Res.*, 12:1651–1662. (Cited on page 180.)

Fan, Y., Newman, T., Linardopoulou, E., and Trask, B. J. (2002b). Gene content and function of the ancestral chromosome fusion site in human chromosome 2q13-2q14.1 and paralogous regions. *Genome Res.*, 12:1663–1672. (Cited on page 180.)

Fay, J. C. and Wu, C. I. (2000). Hitchhiking under positive Darwinian selection. *Genetics*, 155:1405–1413. (Cited on page 180.)

Fedurco, M., Romieu, A., Williams, S., Lawrence, I., and Turcatti, G. (2006). BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Res.*, 34:e22. (Cited on page 24.)

Ferragina, P., Giancarlo, R., Greco, V., Manzini, G., and Valiente, G. (2007). Compression-based classification of biological sequences and structures via the Universal Similarity Metric: experimental assessment. *BMC Bioinformatics*, 8:252. (Cited on page 36.)

Ferragina, P. and Manzini, G. (2000). Opportunistic data structures with applications. In *Proceedings of the 41st Symposium on Foundations of Computer Science (FOCS 2000)*, pages 390–398. IEEE Computer Society. (Cited on pages 36 and 38.)

Ferris, S. D., Brown, W. M., Davidson, W. S., and Wilson, A. C. (1981). Extensive polymorphism in the mitochondrial DNA of apes. *Proc. Natl. Acad. Sci. U.S.A.*, 78:6319–6323. (Cited on page 111.)

Filipski, J. (1988). Why the rate of silent codon substitutions is variable within a vertebrate's genome. *J. Theor. Biol.*, 134:159–164. (Cited on page 12.)

Fischer, A., Pollack, J., Thalmann, O., Nickel, B., and Paabo, S. (2006). Demographic history and genetic differentiation in apes. *Curr. Biol.*, 16:1133–1138. (Cited on page 151.)

Fischer, A., Wiebe, V., Paabo, S., and Przeworski, M. (2004). Evidence for a complex demographic history of chimpanzees. *Mol. Biol. Evol.*, 21:799–808. (Cited on page 112.)

Flicek, P. and Birney, E. (2009). Sense from sequence reads: methods for alignment and assembly. *Nat. Methods*, 6:S6–S12. (Cited on page 35.)

Frederico, L. A., Kunkel, T. A., and Shaw, B. R. (1990). A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. *Biochemistry*, 29:2532–2537. (Cited on page 9.)

Frederico, L. A., Kunkel, T. A., and Shaw, B. R. (1993). Cytosine deamination in mismatched base pairs. *Biochemistry*, 32:6523–6530. (Cited on page 9.)

Fredlake, C. P., Hert, D. G., Kan, C. W., Chiesl, T. N., Root, B. E., Forster, R. E., and Barron, A. E. (2008). Ultrafast DNA sequencing on a microchip by a hybrid separation mechanism that gives 600 bases in 6.5 minutes. *Proc. Natl. Acad. Sci. U.S.A.*, 105:476–481. (Cited on page 186.)

Fredriksson, S., Banr, J., Dahl, F., Chu, A., Ji, H., Welch, K., and Davis, R. W. (2007). Multiplex amplification of all coding sequences within 10 cancer genes by Gene-Collector. *Nucleic Acids Res.*, 35:e47. (Cited on page 21.)

Freije, D., Helms, C., Watson, M. S., and Donis-Keller, H. (1992). Identification of a second pseudoautosomal region near the Xq and Yq telomeres. *Science*, 258:1784–1787. (Cited on page 78.)

Friedberg, E. C., McDaniel, L. D., and Schultz, R. A. (2004). The role of endogenous and exogenous DNA damage and mutagenesis. *Curr. Opin. Genet. Dev.*, 14:5–10. (Cited on page 3.)

Fryxell, K. J. and Moon, W. J. (2005). CpG mutation rates in the human genome are highly dependent on local GC content. *Mol. Biol. Evol.*, 22:650–658. (Cited on pages 9, 10, 11, 12, 13, and 145.)

Fryxell, K. J. and Zuckerkandl, E. (2000). Cytosine deamination plays a primary role in the evolution of mammalian isochores. *Mol. Biol. Evol.*, 17:1371–1383. (Cited on page 9.)

Fu, Y. X. and Li, W. H. (1993). Statistical tests of neutrality of mutations. *Genetics*, 133:693–709. (Cited on page 180.)

Fujiyama, A., Watanabe, H., Toyoda, A., Taylor, T. D., Itoh, T., Tsai, S. F., Park, H. S., Yaspo, M. L., Lehrach, H., Chen, Z., Fu, G., Saitou, N., Osoegawa, K., de Jong, P. J., Suto, Y., Hattori,

M., and Sakaki, Y. (2002). Construction and analysis of a human-chimpanzee comparative clone map. *Science*, 295:131–134. (Cited on page 71.)

Gaffney, D. J. and Keightley, P. D. (2005). The scale of mutational variation in the murid genome. *Genome Res.*, 15:1086–1094. (Cited on pages 10, 12, 14, 15, 113, and 118.)

Gagneux, P. and Varki, A. (2001). Genetic differences between humans and great apes. *Mol. Phylogenet. Evol.*, 18:2–13. (Cited on page 136.)

Gagneux, P., Wills, C., Gerloff, U., Tautz, D., Morin, P. A., Boesch, C., Fruth, B., Hohmann, G., Ryder, O. A., and Woodruff, D. S. (1999). Mitochondrial sequences show diverse evolutionary histories of African hominoids. *Proc. Natl. Acad. Sci. U.S.A.*, 96:5077–5082. (Cited on pages 71 and 112.)

Galtier, N., Piganeau, G., Mouchiroud, D., and Duret, L. (2001). GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. *Genetics*, 159:907–911. (Cited on page 144.)

Gelfand, A. and Smith, A. (1990). Sampling-based approaches to calculating marginal densities. *J. American Statistical Association*, 85:398–409. (Cited on page 60.)

Gelman, A. and Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.*, 7:457–511. (Cited on page 62.)

George, K. S., Zhao, X., Gallahan, D., Shirkey, A., Zareh, A., and Esmaeli-Azad, B. (1997). Capillary electrophoresis methodology for identification of cancer related gene expression patterns of fluorescent differential display polymerase chain reaction. *J. Chromatogr. B Biomed. Sci. Appl.*, 695:93–102. (Cited on page 185.)

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. *Bayesian Statistics*, 4:169–193. (Cited on pages 61 and 63.)

Giakoumatos, S., Vrontos, I., Dellaportas, P., and Politis, D. (1999). A MCMC convergence diagnostic using subsampling. *J. Comput. Graph. Statist.*, 8:431–451. (Cited on page 62.)

Gilad, Y., Bustamante, C. D., Lancet, D., and Paabo, S. (2003). Natural selection on the olfactory receptor gene family in humans and chimpanzees. *Am. J. Hum. Genet.*, 73:489–501. (Cited on page 112.)

Gilbert, N., Boyle, S., Fiegler, H., Woodfine, K., Carter, N. P., and Bickmore, W. A. (2004). Chromatin architecture of the human genome: gene-rich domains are enriched in open chromatin fibers. *Cell*, 118:555–566. (Cited on page 9.)

Goetting-Minesky, M. P. and Makova, K. D. (2006). Mammalian male mutation bias: impacts of generation time and regional variation in substitution rates. *J. Mol. Evol.*, 63:537–544. (Cited on pages 13, 14, and 118.)

Gojobori, T., Li, W. H., and Graur, D. (1982). Patterns of nucleotide substitution in pseudogenes and functional genes. *J. Mol. Evol.*, 18:360–369. (Cited on pages 8, 113, 117, and 145.)

Golding, B. (1992). The prospects for polymorphisms shared between species. *Heredity (Edinb)*, 68:263–276. (Cited on page 135.)

Goodman, M. (1999). The genomic record of Humankind's evolutionary roots. *Am. J. Hum. Genet.*, 64:31–39. (Cited on page 71.)

Goriely, A., Hansen, R. M. S., Taylor, I. B., Olesen, I. A., Krag Jacobsen, G., McGowan, S. J., Pfeifer, S. P., McVean, G. A. T., Rajpert-De Meyts, E., and Wilkie, A. O. (2009). A shared genetic origin for congenital disorders and testicular tumors revealed by activating mutations in FGFR3 and HRAS. *Nat Genet.*, 41:1247–1252. (Cited on pages 49 and 51.)

Goriely, A., McVean, G., Rjmyr, M., Ingemarsson, B., and Wilkie, A. (2003). Evidence for selective advantage of pathogenic FGFR2 mutations in the male germ line. *Science*, 301:643–646. (Cited on pages 49 and 51.)

Goriely, A., McVean, G. A., van Pelt, A. M., O'Rourke, A. W., Wall, S. A., de Rooij, D. G., and Wilkie, A. O. (2005). Gain-of-function amino acid substitutions drive positive selection of FGFR2 mutations in human spermatogonia. *Proc. Natl. Acad. Sci. U.S.A.*, 102:6051–6056. (Cited on pages 49 and 50.)

Graf, S., Nielsen, F. G., Kurtz, S., Huynen, M. A., Birney, E., Stunnenberg, H., and Flicek, P. (2007). Optimized design and assessment of whole genome tiling arrays. *Bioinformatics*, 23:195–204. (Cited on page 36.)

Gratacos, M., Nadal, M., Martin-Santos, R., Pujana, M. A., Gago, J., Peral, B., Armengol, L., Ponsa, I., Miro, R., Bulbena, A., and Estivill, X. (2001). A polymorphic genomic duplication on human chromosome 15 is a susceptibility factor for panic and phobic disorders. *Cell*, 106:367–379. (Cited on page 5.)

Graves, J. A., Wakefield, M. J., and Toder, R. (1998). The origin and evolution of the pseudoautosomal regions of human sex chromosomes. *Hum. Mol. Genet.*, 7:1991–1996. (Cited on page 78.)

Green, P., Ewing, B., Miller, W., Thomas, P. J., and Green, E. D. (2003). Transcription-associated mutational asymmetry in mammalian evolution. *Nat. Genet.*, 33:514–517. (Cited on pages 11 and 12.)

Green, R. E., Krause, J., Ptak, S. E., Briggs, A. W., Ronan, M. T., Simons, J. F., Du, L., Egholm, M., Rothberg, J. M., Paunovic, M., and Paabo, S. (2006). Analysis of one million base pairs of Neanderthal DNA. *Nature*, 444:330–336. (Cited on page 191.)

Green, R. E., Malaspinas, A. S., Krause, J., Briggs, A. W., Johnson, P. L., Uhler, C., Meyer, M., Good, J. M., Maricic, T., Stenzel, U., Prufer, K., Siebauer, M., Burbano, H. A., Ronan, M., Rothberg, J. M., Egholm, M., Rudan, P., Brajkovic, D., Kucan, Z., Gusic, I., Wikstrom, M., Laakkonen, L., Kelso, J., Slatkin, M., and Paabo, S. (2008). A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell*, 134:416–426. (Cited on page 191.)

Gupta, P. K. (2008). Single-molecule DNA sequencing technologies for future genomics research. *Trends Biotechnol.*, 26:602–611. (Cited on pages 200, 201, and 203.)

Gyllensten, U., Sundvall, M., Ezcurra, I., and Erlich, H. A. (1991a). Genetic diversity at class II DRB loci of the primate MHC. *J. Immunol.*, 146:4368–4376. (Cited on pages 165 and 169.)

Gyllensten, U. B. and Erlich, H. A. (1989). Ancient roots for polymorphism at the HLA-DQ alpha locus in primates. *Proc. Natl. Acad. Sci. U.S.A.*, 86:9986–9990. (Cited on pages 122, 136, 165, and 169.)

Gyllensten, U. B., Sundvall, M., and Erlich, H. A. (1991b). Allelic diversity is generated by intraexon sequence exchange at the DRB1 locus of primates. *Proc. Natl. Acad. Sci. U.S.A.*, 88:3686–3690. (Cited on pages 165 and 169.)

Hacia, J. G., Fan, J. B., Ryder, O., Jin, L., Edgemon, K., Ghandour, G., Mayer, R. A., Sun, B., Hsie, L., Robbins, C. M., Brody, L. C., Wang, D., Lander, E. S., Lipshutz, R., Fodor, S. P., and Collins, F. S. (1999). Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. *Nat. Genet.*, 22:164–167. (Cited on page 137.)

Haldane, J. B. (1935). The rate of spontaneous mutation of a human gene. *Journal of Genetics*, 31:317–326. (Cited on page 15.)

Haldane, J. B. (1947). The mutation rate of the gene for haemophilia, and its segregation ratios in males and females. *Ann Eugen*, 13:262–271. (Cited on pages 13 and 118.)

Halligan, D. L. and Keightley, P. D. (2006). Ubiquitous selective constraints in the Drosophila genome revealed by a genome-wide interspecies comparison. *Genome Res.*, 16:875–884. (Cited on page 152.)

Hardison, R. C., Roskin, K. M., Yang, S., Diekhans, M., Kent, W. J., Weber, R., Elnitski, L., Li, J., O'Connor, M., Kolbe, D., Schwartz, S., Furey, T. S., Whelan, S., Goldman, N., Smit, A., Miller, W., Chiaromonte, F., and Haussler, D. (2003). Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Res.*, 13:13–26. (Cited on pages 10, 13, and 118.)

Harismendy, O., Ng, P. C., Strausberg, R. L., Wang, X., Stockwell, T. B., Beeson, K. Y., Schork, N. J., Murray, S. S., Topol, E. J., Levy, S., and Frazer, K. A. (2009). Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol.*, 10:R32. (Cited on page 42.)

Harris, T. D., Buzby, P. R., Babcock, H., Beer, E., Bowers, J., Braslavsky, I., Causey, M., Colonell, J., Dimeo, J., Efcavitch, J. W., Giladi, E., Gill, J., Healy, J., Jarosz, M., Lapen, D., Moulton, K., Quake, S. R., Steinmann, K., Thayer, E., Tyurina, A., Ward, R., Weiss, H., and Xie, Z. (2008). Single-molecule DNA sequencing of a viral genome. *Science*, 320:106–109. (Cited on page 199.)

Harrow, J., Denoeud, F., Frankish, A., Reymond, A., Chen, C. K., Chrast, J., Lagarde, J., Gilbert, J. G., Storey, R., Swarbreck, D., Rossier, C., Ucla, C., Hubbard, T., Antonarakis, S. E., and Guigo, R. (2006). GENCODE: producing a reference annotation for ENCODE. *Genome Biol.*, 7 (Suppl 1):1–9. (Cited on page 152.)

Hartl, D. and Clark, A. G. (1997). *Principals of Population Genetics*. Sinauer, Massachusetts. (Cited on page 18.)

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov Chains and their applications. *Biometrica*, 57:97–109. (Cited on page 60.)

Hattori, M., Fujiyama, A., Taylor, T. D., Watanabe, H., Yada, T., Park, H. S., Toyoda, A., Ishii, K., Totoki, Y., Choi, D. K., Groner, Y., Soeda, E., Ohki, M., Takagi, T., Sakaki, Y., Taudien,

S., Blechschmidt, K., Polley, A., Menzel, U., Delabar, J., Kumpf, K., Lehmann, R., Patterson, D., Reichwald, K., Rump, A., Schillhabel, M., Schudy, A., Zimmermann, W., Rosenthal, A., Kudoh, J., Schibuya, K., Kawasaki, K., Asakawa, S., Shintani, A., Sasaki, T., Nagamine, K., Mitsuyama, S., Antonarakis, S. E., Minoshima, S., Shimizu, N., Nordsiek, G., Hornischer, K., Brant, P., Scharfe, M., Schon, O., Desario, A., Reichelt, J., Kauer, G., Blocker, H., Ramser, J., Beck, A., Klages, S., Hennig, S., Riesselmann, L., Dagand, E., Haaf, T., Wehrmeyer, S., Borzym, K., Gardiner, K., Nizetic, D., Francis, F., Lehrach, H., Reinhardt, R., and Yaspo, M. L. (2000). The DNA sequence of human chromosome 21. *Nature*, 405:311–319. (Cited on page 6.)

Hawk, J. D., Stefanovic, L., Boyer, J. C., Petes, T. D., and Farber, R. A. (2005). Variation in efficiency of DNA mismatch repair at different sites in the yeast genome. *Proc. Natl. Acad. Sci. U.S.A.*, 102:8639–8643. (Cited on page 13.)

Hedrick, P. W. (1998). Balancing selection and MHC. *Genetica*, 104:207–214. (Cited on pages 137, 143, 157, and 166.)

Hedrick, P. W. (2007). Sex: differences in mutation, recombination, selection, gene flow, and genetic drift. *Evolution*, 61:2750–2771. (Cited on page 113.)

Heidelberger, P. and Welch, P. (1981). A spectral method for condence interval generation and run length control in simulations. *Comm. ACM.*, 24:233–245. (Cited on pages 62 and 63.)

Heidelberger, P. and Welch, P. (1983). Simulation run length control in the presence of an initial transient. *Opns Res.*, 31:1109–44. (Cited on pages 62 and 63.)

Helicos BioSciences (2011). True single molecule sequencing performance. Technical report, Helicos BioSciences. Available from: `http://www.helicosbio.com/Technology/TrueSingleMoleculeSequencing/`. (Cited on page 199.)

Hellmann, I., Ebersberger, I., Ptak, S. E., Paabo, S., and Przeworski, M. (2003). A neutral explanation for the correlation of diversity with recombination rates in humans. *Am. J. Hum. Genet.*, 72:1527–1535. (Cited on pages 8, 12, and 113.)

Hellmann, I., Prufer, K., Ji, H., Zody, M. C., Paabo, S., and Ptak, S. E. (2005). Why do human diversity levels vary at a megabase scale? *Genome Res.*, 15:1222–1231. (Cited on pages 10, 12, 13, 122, and 131.)

Hernandez, R. D., Kelley, J. L., Elyashiv, E., Melton, S. C., Auton, A., McVean, G., Sella, G., and Przeworski, M. (2011). Classic selective sweeps were rare in recent human evolution. *Science*, 331:920–924. (Cited on page 151.)

Hert, D. G., Fredlake, C. P., and Barron, A. E. (2008). Advantages and limitations of next-generation sequencing technologies: a comparison of electrophoresis and non-electrophoresis methods. *Electrophoresis*, 29:4618–4626. (Cited on pages 185, 186, and 191.)

Hess, S. T., Blake, J. D., and Blake, R. D. (1994). Wide variations in neighbor-dependent substitution rates. *J. Mol. Biol.*, 236:1022–1033. (Cited on pages 8, 10, and 117.)

Hey, J. (1999). The neutralist, the fly and the selectionist. *Trends Ecol. Evol. (Amst.)*, 14:35–38. (Cited on page 137.)

Hill, W. and Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics*, 38:226–231. (Cited on page 18.)

Hodges, E., Xuan, Z., Balija, V., Kramer, M., Molla, M. N., Smith, S. W., Middle, C. M., Rodesch, M. J., Albert, T. J., Hannon, G. J., and McCombie, W. R. (2007). Genome-wide in situ exon capture for selective resequencing. *Nat. Genet.*, 39:1522–1527. (Cited on page 21.)

Hodgkinson, A. and Eyre-Walker, A. (2010). The genomic distribution and local context of coincident SNPs in human and chimpanzee. *Genome Biol Evol*, 2:547–557. (Cited on page 176.)

Hodgkinson, A. and Eyre-Walker, A. (2011). Variation in the mutation rate across mammalian genomes. *Nat. Rev. Genet.*, 12:756–766. (Cited on pages 2, 11, and 150.)

Hodgkinson, A., Ladoukakis, E., and Eyre-Walker, A. (2009). Cryptic variation in the human mutation rate. *PLoS Biol.*, 7:e1000027. (Cited on page 176.)

Holliday, R. and Grigg, G. W. (1993). DNA methylation and mutation. *Mutat. Res.*, 285:61–67. (Cited on pages 9, 15, 117, 122, 131, 145, and 176.)

Holmquist, G. P. (1992). Chromosome bands, their chromatin flavors, and their functional features. *Am. J. Hum. Genet.*, 51:17–37. (Cited on page 122.)

Holmquist, G. P. and Filipski, J. (1994). Organization of mutations along the genome: a prime determinant of genome evolution. *Trends Ecol. Evol. (Amst.)*, 9:65–69. (Cited on page 12.)

Homer, N., Merriman, B., and Nelson, S. F. (2009). BFAST: an alignment tool for large scale genome resequencing. *PLoS ONE*, 4:e7767. (Cited on page 29.)

Hon, W.-K., Lam, T.-W., Sadakane, K., Sung, W, K., and Yiu, S.-M. (2007). *Algorithmica*, volume 48, chapter A Space and Time Efficient Algorithm for Constructing Compressed Suffix Arrays, pages 23–36. Springer-Verlag New York, Inc. (Cited on page 39.)

Horton, R., Niblett, D., Milne, S., Palmer, S., Tubby, B., Trowsdale, J., and Beck, S. (1998). Large-scale sequence comparisons reveal unusually high levels of variation in the HLA-DQB1 locus in the class II region of the human MHC. *J. Mol. Biol.*, 282:71–97. (Cited on page 164.)

Horton, W. A. and Lunstrum, G. P. (2002). Fibroblast growth factor receptor 3 mutations in achondroplasia and related forms of dwarfism. *Rev Endocr Metab Disord*, 3:381–385. (Cited on page 50.)

Howie, B. N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.*, 5:e1000529. (Cited on page 44.)

Hsu, F., Kent, W. J., Clawson, H., Kuhn, R. M., Diekhans, M., and Haussler, D. (2006). The UCSC Known Genes. *Bioinformatics*, 22:1036–1046. (Cited on pages 143, 167, 168, 170, 171, 172, and 221.)

Hudson, R. R. (1994). How can the low levels of DNA sequence variation in regions of the drosophila genome with low recombination rates be explained? *Proc. Natl. Acad. Sci. U.S.A.*, 91:6815–6818. (Cited on page 122.)

Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18:337–338. (Cited on page 150.)

Hudson, R. R. and Kaplan, N. L. (1995). Deleterious background selection with recombination. *Genetics*, 141:1605–1617. (Cited on pages 121 and 122.)

Hudson, R. R., Kreitman, M., and Aguade, M. (1987). A test of neutral molecular evolution based on nucleotide data. *Genetics*, 116:153–159. (Cited on page 180.)

Hughes, A. L. and Nei, M. (1988). Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature*, 335:167–170. (Cited on pages 122, 136, 137, 143, 157, and 166.)

Hughes, A. L. and Nei, M. (1989). Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. *Proc. Natl. Acad. Sci. U.S.A.*, 86:958–962. (Cited on pages 122, 136, 137, 143, 157, and 166.)

Hughes, A. L. and Yeager, M. (1998). Natural selection at major histocompatibility complex loci of vertebrates. *Annu. Rev. Genet.*, 32:415–435. (Cited on pages 154 and 164.)

Hunkapiller, T., Kaiser, R. J., Koop, B. F., and Hood, L. (1991). Large-scale and automated DNA sequence determination. *Science*, 254:59–67. (Cited on page 183.)

Huse, S. M., Huber, J. A., Morrison, H. G., Sogin, M. L., and Welch, D. M. (2007). Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol.*, 8:R143. (Cited on page 191.)

Huttley, G. A., Smith, M. W., Carrington, M., and O'Brien, S. J. (1999). A scan for linkage disequilibrium across the human genome. *Genetics*, 152:1711–1722. (Cited on page 18.)

Huxley, T. H. (1863). *Evidence as to Man's Place in Nature*. Williams and Norgate, London. (Cited on page III.)

Hwang, D. G. and Green, P. (2004). Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc. Natl. Acad. Sci. U.S.A.*, 101:13994–14001. (Cited on pages 8, 9, 10, 14, 113, 117, 121, 122, 131, 144, 145, and 176.)

Illumina (2010). Genome Analyzer II System. Technical report, Illumina. Available from: `http://www.illumina.com`. (Cited on pages 26, 52, 55, and 188.)

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, 409:860–921. (Cited on pages 6, 21, 22, 72, and 152.)

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. *Nature*, 431:931–945. (Cited on pages 22 and 72.)

Jeffreys, A. J., Kauppi, L., and Neumann, R. (2001). Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat. Genet.*, 29:217–222. (Cited on page 18.)

Jeffreys, A. J., Royle, N. J., Wilson, V., and Wong, Z. (1988). Spontaneous mutation rates to new length alleles at tandem-repetitive hypervariable loci in human DNA. *Nature*, 332:278–281. (Cited on pages 3 and 13.)

Jensen-Seaman, M. I., Deinard, A. S., and Kidd, K. K. (2001). Modern African ape populations as genetic and demographic models of the last common ancestor of humans, chimpanzees, and gorillas. *J. Hered.*, 92:475–480. (Cited on page 112.)

Jiang, H. and Wong, W. H. (2008). SeqMap: mapping massive amount of oligonucleotides to the genome. *Bioinformatics*, 24:2395–2396. (Cited on page 29.)

Johnson, P. L. and Hellmann, I. (2011). Mutation rate distribution inferred from coincident SNPs and coincident substitutions. *Genome Biol Evol*, 3:842–850. (Cited on page 176.)

Johnston, M. O. (2001). *Encyclopedia of Life Sciences*, chapter Mutations and New Variation: Overview, pages 1–10. Nature Publishing Group. (Cited on pages 1 and 7.)

Ju, J., Glazer, A. N., and Mathies, R. A. (1996). Energy transfer primers: a new fluorescence labeling paradigm for DNA sequencing and analysis. *Nat. Med.*, 2:246–249. (Cited on page 184.)

Ju, J., Ruan, C., Fuller, C. W., Glazer, A. N., and Mathies, R. A. (1995). Fluorescence energy transfer dye-labeled primers for DNA sequencing and analysis. *Proc. Natl. Acad. Sci. U.S.A.*, 92:4347–4351. (Cited on page 184.)

Kaessmann, H., Heissig, F., von Haeseler, A., and Paabo, S. (1999a). DNA sequence variation in a non-coding region of low recombination on the human X chromosome. *Nat. Genet.*, 22:78–81. (Cited on page 112.)

Kaessmann, H. and Paabo, S. (2002). The genetical history of humans and the great apes. *J. Intern. Med.*, 251:1–18. (Cited on page 112.)

Kaessmann, H., Wiebe, V., and Paabo, S. (1999b). Extensive nuclear DNA sequence diversity among chimpanzees. *Science*, 286:1159–1162. (Cited on page 112.)

Kaessmann, H., Wiebe, V., Weiss, G., and Paabo, S. (2001). Great ape DNA sequences reveal a reduced diversity and an expansion in humans. *Nat. Genet.*, 27:155–156. (Cited on page 112.)

Kahvejian, A., Quackenbush, J., and Thompson, J. F. (2008). What would you do if you could sequence everything? *Nature Biotechnology*, 26:1125–1133. (Cited on page 188.)

Kan, S., Elanko, N., Johnson, D., Cornejo-Roldan, L., Cook, J., Reich, E., Tomkins, S., Verloes, A., Twigg, S., Rannan-Eliya, S., McDonald-McGinn, D., Zackai, E., Wall, S., Muenke, M., and Wilkie, A. (2002). Genomic screening of fibroblast growth-factor receptor 2 reveals a wide spectrum of mutations in patients with syndromic craniosynostosis. *Am. J. Hum. Genet.*, 70:472–486. (Cited on page 49.)

Kannan, K. and Givol, D. (2000). FGF receptor mutations: dimerization syndromes, cell growth suppression, and animal models. *IUBMB Life*, 49:197–205. (Cited on page 50.)

Kaplan, N. L., Hudson, R. R., and Langley, C. H. (1989). The "hitchhiking effect" revisited. *Genetics*, 123:887–899. (Cited on pages 121 and 122.)

Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D., and Kent, W. J. (2004). The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.*, 32:D493–D496. (Cited on pages 125, 143, 167, 168, 170, 171, 172, and 221.)

Keele, B. F., Van Heuverswyn, F., Li, Y., Bailes, E., Takehisa, J., Santiago, M. L., Bibollet-Ruche, F., Chen, Y., Wain, L. V., Liegeois, F., Loul, S., Ngole, E. M., Bienvenue, Y., Delaporte, E., Brookfield, J. F. Y., Sharp, P. M., Shaw, G. M., Peeters, M., and Hahn, B. H. (2006). Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. *Science*, 313:523–526. (Cited on pages 71 and 72.)

Keightley, P. D., Eory, L., Halligan, D. L., and Kirkpatrick, M. (2011). Inference of mutation parameters and selective constraint in mammalian coding sequences by approximate Bayesian computation. *Genetics*, 187:1153–1161. (Cited on pages 10, 144, and 145.)

Ketterling, R. P., Vielhaber, E., Bottema, C. D., Schaid, D. J., Cohen, M. P., Sexauer, C. L., and Sommer, S. S. (1993). Germ-line origins of mutation in families with hemophilia B: the sex ratio varies with the type of mutation. *Am. J. Hum. Genet.*, 52:152–166. (Cited on page 13.)

Kim, S. H., Elango, N., Warden, C., Vigoda, E., and Yi, S. V. (2006). Heterogeneous genomic molecular clocks in primates. *PLoS Genet.*, 2:e163. (Cited on pages 9, 117, 122, 131, and 176.)

Kim, Y. and Stephan, W. (2000). Joint effects of genetic hitchhiking and background selection on neutral variation. *Genetics*, 155:1415–1427. (Cited on page 122.)

Kim, Y. and Stephan, W. (2002). Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics*, 160:765–777. (Cited on page 180.)

Kimmel, M., Chakraborty, R., King, J. P., Bamshad, M., Watkins, W. S., and Jorde, L. B. (1998). Signatures of population expansion in microsatellite repeat data. *Genetics*, 148:1921–1930. (Cited on page 123.)

Kimura, M. (1967). On the evolutionary adjustment of mutation rates. *Genetics Research*, 9:23–34. (Cited on page 1.)

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge University Press. (Cited on pages 1 and 17.)

King, M. C. and Wilson, A. C. (1975). Evolution at two levels in humans and chimpanzees. *Science*, 188:107–116. (Cited on page 111.)

Kircher, M. and Kelso, J. (2010). High-throughput DNA sequencing–concepts and limitations. *Bioessays*, 32:524–536. (Cited on pages 24, 26, 52, 67, 185, 186, 191, 193, 196, and 203.)

Kircher, M., Stenzel, U., and Kelso, J. (2009). Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome Biol.*, 10:R83. (Cited on pages 26, 52, and 55.)

Kirsch, I. R., Morton, C. C., Nakahara, K., and Leder, P. (1982). Human immunoglobulin heavy chain genes map to a region of translocations in malignant B lymphocytes. *Science*, 216:301–303. (Cited on page 6.)

Kitano, T., Schwarz, C., Nickel, B., and Paabo, S. (2003). Gene diversity patterns at 10 X-chromosomal loci in humans and chimpanzees. *Mol. Biol. Evol.*, 20:1281–1289. (Cited on page 112.)

Klein, J. (1987). Origin of major histocompatibility complex polymorphism: the trans-species hypothesis. *Hum. Immunol.*, 19:155–162. (Cited on pages 122, 136, 137, 143, 157, 165, 166, and 169.)

Kondrashov, A. S. (1988). Deleterious mutations and the evolution of sexual reproduction. *Nature*, 336:435–440. (Cited on page 1.)

Kondrashov, A. S. (1998). Measuring spontaneous deleterious mutation process. *Genetica*, 102-103:183–197. (Cited on page 17.)

Kondrashov, A. S. (2003). Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. *Hum. Mutat.*, 21:12–27. (Cited on pages 15, 68, and 174.)

Kondrashov, A. S. and Crow, J. F. (1993). A molecular approach to estimating the human deleterious mutation rate. *Hum. Mutat.*, 2:229–234. (Cited on page 15.)

Kondrashov, F. A. and Kondrashov, A. S. (2010). Measurements of spontaneous rates of mutations in the recent past and the near future. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, 365:1169–1176. (Cited on pages 15 and 16.)

Kong, A., Frigge, M. L., Masson, G., Besenbacher, S., Sulem, P., Magnusson, G., Gudjonsson, S. A., Sigurdsson, A., Jonasdottir, A., Jonasdottir, A., Wong, W. S., Sigurdsson, G., Walters, G. B., Steinberg, S., Helgason, H., Thorleifsson, G., Gudbjartsson, D. F., Helgason, A., Magnusson, O. T., Thorsteinsdottir, U., and Stefansson, K. (2012). Rate of de novo mutations and the importance of father's age to disease risk. *Nature*, 488:471–475. (Cited on page 151.)

Korlach, J., Marks, P. J., Cicero, R. L., Gray, J. J., Murphy, D. L., Roitman, D. B., Pham, T. T., Otto, G. A., Foquet, M., and Turner, S. W. (2008). Selective aluminum passivation for targeted

immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures. *Proc. Natl. Acad. Sci. U.S.A.*, 105:1176–1181. (Cited on page 202.)

Krawczak, M., Ball, E. V., and Cooper, D. N. (1998). Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. *Am. J. Hum. Genet.*, 63:474–488. (Cited on pages 8, 9, 10, 15, 113, and 117.)

Krishnan, B. R., Blakesley, R. W., and Berg, D. E. (1991). Linear amplification DNA sequencing directly from single phage plaques and bacterial colonies. *Nucleic Acids Res.*, 19:1153. (Cited on page 184.)

Kupfermann, H., Mayer, W. E., O'hUigin, C., Klein, D., and Klein, J. (1992). Shared polymorphism between gorilla and human major histocompatibility complex DRB loci. *Hum. Immunol.*, 34:267–278. (Cited on page 157.)

Lam, T. W., Sung, W. K., Tam, S. L., Wong, C. K., and Yiu, S. M. (2008). Compressed indexing and local alignment of DNA. *Bioinformatics*, 24:791–797. (Cited on page 39.)

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, 10:R25. (Cited on page 35.)

Lawlor, D. A., Ward, F. E., Ennis, P. D., Jackson, A. P., and Parham, P. (1988). HLA-A and B polymorphisms predate the divergence of humans and chimpanzees. *Nature*, 335:268–271. (Cited on pages 122, 136, 165, and 169.)

Leamon, J. H., Lee, W. L., Tartaro, K. R., Lanza, J. R., Sarkis, G. J., deWinter, A. D., Berka, J., Weiner, M., Rothberg, J. M., and Lohman, K. L. (2003). A massively parallel PicoTiterPlate based platform for discrete picoliter-scale polymerase chain reactions. *Electrophoresis*, 24:3769–3777. (Cited on page 189.)

Lee, L. G., Spurgeon, S. L., Heiner, C. R., Benson, S. C., Rosenblum, B. B., Menchen, S. M., Graham, R. J., Constantinescu, A., Upadhya, K. G., and Cassel, J. M. (1997). New energy transfer dyes for DNA sequencing. *Nucleic Acids Res.*, 25:2816–2822. (Cited on page 184.)

Leigh, E. G. (1970). Natural selection and mutability. *American Naturalist*, 104:301–305. (Cited on page 1.)

Lenoir, G. M., Preud'homme, J. L., Bernheim, A., and Berger, R. (1982). Correlation between immunoglobulin light chain expression and variant translocation in Burkitt's lymphoma. *Nature*, 298:474–476. (Cited on page 6.)

Lercher, M. J. and Hurst, L. D. (2002). Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet.*, 18:337–340. (Cited on pages 12, 13, 14, 15, 113, 118, and 121.)

Lercher, M. J., Williams, E. J., and Hurst, L. D. (2001). Local similarity in evolutionary rates extends over whole chromosomes in human-rodent and mouse-rat comparisons: implications for understanding the mechanistic basis of the male mutation bias. *Mol. Biol. Evol.*, 18:2032–2039. (Cited on pages 8, 15, 113, and 117.)

Levene, M. J., Korlach, J., Turner, S. W., Foquet, M., Craighead, H. G., and Webb, W. W. (2003). Zero-mode waveguides for single-molecule analysis at high concentrations. *Science*, 299:682–686. (Cited on page 202.)

Lewontin, R. and Kojima, K. (1960). The evolutionary dynamics of complex polymorphisms. *Evolution*, 14:458–472. (Cited on page 18.)

Li, H. (2011). Improving SNP discovery by base alignment quality. *Bioinformatics*, 27:1157–1158. (Cited on pages 42, 80, 86, and 87.)

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25:1754–1760. (Cited on pages 35, 36, 38, 39, and 76.)

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009a). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25:2078–2079. (Cited on pages 41, 80, 81, 87, and 211.)

Li, H. and Homer, N. (2010). A survey of sequence alignment algorithms for next-generation sequencing. *Brief Bioinformatics*, 11:473–483. (Cited on pages 42 and 79.)

Li, H., Ruan, J., and Durbin, R. (2008a). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, 18:1851–1858. (Cited on pages 29, 30, 32, 33, 43, 44, and 76.)

Li, J. and Deng, H. W. (2005). Estimation of the rate and effects of deleterious genomic mutations in finite populations with linkage disequilibrium. *Heredity*, 95:59–68. (Cited on page 17.)

Li, R., Li, Y., Fang, X., Yang, H., Wang, J., Kristiansen, K., and Wang, J. (2009b). SNP detection for massively parallel whole-genome resequencing. *Genome Res.*, 19:1124–1132. (Cited on page 43.)

Li, R., Li, Y., Kristiansen, K., and Wang, J. (2008b). SOAP: short oligonucleotide alignment program. *Bioinformatics*, 24:713–714. (Cited on pages 29 and 44.)

Li, R., Yu, C., Li, Y., Lam, T. W., Yiu, S. M., Kristiansen, K., and Wang, J. (2009c). SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, 25:1966–1967. (Cited on pages 35, 41, 43, and 44.)

Li, W. (1997). *Molecular Evolution*. Sinauer Associates. (Cited on page 121.)

Li, W., Yi, S., and Makova, K. (2002). Male-driven evolution. *Curr. Opin. Genet. Dev.*, 12:650–656. (Cited on pages 8, 10, 14, 113, 117, and 144.)

Li, W. H., Ellsworth, D. L., Krushkal, J., Chang, B. H., and Hewett-Emmett, D. (1996). Rates of nucleotide substitution in primates and rodents and the generation-time effect hypothesis. *Mol. Phylogenet. Evol.*, 5:182–187. (Cited on page 13.)

Li, W. H., Wu, C. I., and Luo, C. C. (1984). Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. *J. Mol. Evol.*, 21:58–71. (Cited on pages 8 and 113.)

Li, Y., Willer, C. J., Ding, J., Scheet, P., and Abecasis, G. R. (2010). MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.*, 34:816–834. (Cited on page 44.)

Lievens, P. M., Mutinelli, C., Baynes, D., and Liboi, E. (2004). The kinase activity of fibroblast growth factor receptor 3 with activation loop mutations affects receptor trafficking and signaling. *J. Biol. Chem.*, 279:43254–43260. (Cited on page 50.)

Life Technologies (2011). Ion Torrent Technology: How does it work? Technical report, Life Technologies. Available from: `http://www.iontorrent.com/technology-how-does-it-work-more/`. (Cited on page 197.)

Lin, H., Zhang, Z., Zhang, M. Q., Ma, B., and Li, M. (2008). ZOOM! Zillions of oligos mapped. *Bioinformatics*, 24:2431–2437. (Cited on page 29.)

Lindeman, R., Merenda, P., and Gold, R. (1980). *Introduction to Bivariate and Multivariate Analysis*. Scott, Foresman, Glenview, IL. (Cited on page 129.)

Lippert, M. J., Chen, Q., and Liber, H. L. (1998). Increased transcription decreases the spontaneous mutation rate at the thymidine kinase locus in human cells. *Mutat. Res.*, 401:1–10. (Cited on page 11.)

Lister, R., Gregory, B. D., and Ecker, J. R. (2009). Next is now: new technologies for sequencing of genomes, transcriptomes, and beyond. *Curr. Opin. Plant Biol.*, 12:107–118. (Cited on pages 26, 52, and 55.)

Liu, C. M., Wong, T., Wu, E., Luo, R., Yiu, S. M., Li, Y., Wang, B., Yu, C., Chu, X., Zhao, K., Li, R., and Lam, T. W. (2012). SOAP3: ultra-fast GPU-based parallel alignment tool for short reads. *Bioinformatics*, 28:878–879. (Cited on page 44.)

Lobry, J. R. (1996). Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol. Biol. Evol.*, 13:660–665. (Cited on page 12.)

Loewe, L., Textor, V., and Scherer, S. (2003). High deleterious genomic mutation rate in stationary phase of Escherichia coli. *Science*, 302:1558–1560. (Cited on page 15.)

Lopez de Castro, J. A., Strominger, J. L., Strong, D. M., and Orr, H. T. (1982). Structure of crossreactive human histocompatibility antigens HLA-A28 and HLA-A2: possible implications for the generation of HLA polymorphism. *Proc. Natl. Acad. Sci. U.S.A.*, 79:3813–3817. (Cited on page 154.)

Lunter, G. and Goodson, M. (2011). Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res.*, 21:936–939. (Cited on pages 30, 33, 35, and 76.)

Lynch, M. (2008). The cellular, developmental and population-genetic determinants of mutation-rate evolution. *Genetics*, 180:933–943. (Cited on page 15.)

Lynch, M. (2010). Rate, molecular spectrum, and consequences of human mutation. *Proc. Natl. Acad. Sci. U.S.A.*, 107:961–968. (Cited on page 15.)

Ma, B., Tromp, J., and Li, M. (2002). PatternHunter: faster and more sensitive homology search. *Bioinformatics*, 18:440–445. (Cited on page 29.)

Magi, A., Benelli, M., Gozzini, A., Girolami, F., Torricelli, F., and Brandi, M. L. (2010). Bioinformatics for next generation sequencing data. *Genes*, 1:294–307. (Cited on page 43.)

Makova, K. D. and Li, W. H. (2002). Strong male-driven evolution of DNA sequences in humans and apes. *Nature*, 416:624–626. (Cited on pages 14 and 118.)

Malcom, C. M., Wyckoff, G. J., and Lahn, B. T. (2003). Genic mutation rates in mammals: local similarity, chromosomal heterogeneity, and X-versus-autosome disparity. *Mol. Biol. Evol.*, 20:1633–1641. (Cited on pages 14, 15, 113, 118, and 121.)

Malhis, N., Butterfield, Y. S., Ester, M., and Jones, S. J. (2009). Slider–maximum use of probability information for alignment of short sequence reads and SNP detection. *Bioinformatics*, 25:6–13. (Cited on page 44.)

Mamanova, L., Coffey, A. J., Scott, C. E., Kozarewa, I., Turner, E. H., Kumar, A., Howard, E., Shendure, J., and Turner, D. J. (2010). Target-enrichment strategies for next-generation sequencing. *Nat. Methods*, 7:111–118. (Cited on page 27.)

Marchini, J. and Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.*, 11:499–511. (Cited on page 44.)

Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.*, 39:906–913. (Cited on page 44.)

Mardis, E. R. (2008a). Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet*, 9:387–402. (Cited on pages 190, 194, and 195.)

Mardis, E. R. (2008b). The impact of next-generation sequencing technology on genetics. *Trends Genet.*, 24:133–141. (Cited on pages 187 and 190.)

Mardis, E. R. (2011). A decade's perspective on DNA sequencing technology. *Nature*, 470:198–203. (Cited on pages 73, 187, and 203.)

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L., Jarvie, T. P., Jirage, K. B., Kim, J. B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F., and Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437:376–380. (Cited on pages 188, 189, 190, and 191.)

Mariella, R. (2008). Sample preparation: the weak link in microfluidics-based biodetection. *Biomed Microdevices*, 10:777–784. (Cited on page 186.)

Marsh, S. G., Albert, E. D., Bodmer, W. F., Bontrop, R. E., Dupont, B., Erlich, H. A., Fernandez-Vina, M., Geraghty, D. E., Holdsworth, R., Hurley, C. K., Lau, M., Lee, K. W., Mach, B., Maiers,

M., Mayr, W. R., Muller, C. R., Parham, P., Petersdorf, E. W., Sasazuki, T., Strominger, J. L., Svejgaard, A., Terasaki, P. I., Tiercy, J. M., and Trowsdale, J. (2010). Nomenclature for factors of the HLA system, 2010. *Tissue Antigens*, 75:291–455. (Cited on page 164.)

Martin, E. R., Kinnamon, D. D., Schmidt, M. A., Powell, E. H., Zuchner, S., and Morris, R. W. (2010). SeqEM: an adaptive genotype-calling approach for next-generation sequencing studies. *Bioinformatics*, 26:2803–2810. (Cited on page 43.)

Matassi, G., Sharp, P. M., and Gautier, C. (1999). Chromosomal location effects on gene sequence evolution in mammals. *Curr. Biol.*, 9:786–791. (Cited on pages 8, 10, 12, 13, 113, and 117.)

Maxam, A. M. and Gilbert, W. (1977). A new method for sequencing DNA. *Proc. Natl. Acad. Sci. U.S.A.*, 74:560–564. (Cited on page 183.)

May, C. A., Shone, A. C., Kalaydjieva, L., Sajantila, A., and Jeffreys, A. J. (2002). Crossover clustering and rapid decay of linkage disequilibrium in the Xp/Yp pseudoautosomal gene SHOX. *Nat. Genet.*, 31:272–275. (Cited on page 175.)

Mayer, W. E., Jonker, M., Klein, D., Ivanyi, P., van Seventer, G., and Klein, J. (1988). Nucleotide sequences of chimpanzee MHC class I alleles: evidence for trans-species mode of evolution. *EMBO J.*, 7:2765–2774. (Cited on pages 122, 136, 165, and 169.)

Mazzarella, R. and Schlessinger, D. (1998). Pathological consequences of sequence duplications in the human genome. *Genome Res.*, 8:1007–1021. (Cited on page 5.)

McAdam, S. N., Boyson, J. E., Liu, X., Garber, T. L., Hughes, A. L., Bontrop, R. E., and Watkins, D. I. (1994). A uniquely high level of recombination at the HLA-B locus. *Proc. Natl. Acad. Sci. U.S.A.*, 91:5893–5897. (Cited on page 136.)

McDonald, J. H. and Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, 351:652–654. (Cited on page 180.)

McDonald, M. J., Wang, W. C., Huang, H. D., and Leu, J. Y. (2011). Clusters of nucleotide substitutions and insertion/deletion mutations are associated with repeat sequences. *PLoS Biol.*, 9:e1000622. (Cited on page 11.)

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, 20:1297–1303. (Cited on pages 41, 44, 82, 83, 87, 90, and 114.)

McVean, G. (2000). Evolutionary genetics: what is driving male mutation? *Curr. Biol.*, 10:R834–R835. (Cited on pages 10, 13, and 15.)

McVean, G. T. and Hurst, L. D. (1997). Evidence for a selectively favourable reduction in the mutation rate of the X chromosome. *Nature*, 386:388–392. (Cited on pages 1, 2, 13, 14, 118, 119, and 173.)

Messer, P. W. (2009). Measuring the rates of spontaneous mutation from deep and large-scale polymorphism data. *Genetics*, 182:1219–1232. (Cited on page 17.)

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A., and Teller, H. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1091. (Cited on page 60.)

Metropolis, N. and Ulam, S. (1949). The Monte Carlo method. *J. Amer. Statist. Assoc.*, 44:335–341. (Cited on page 60.)

Meyer, M. and Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb Protoc*, 2010:pdb.prot5448. (Cited on page 27.)

Miller, R. D. and Kwok, P. Y. (2001). The birth and death of human single-nucleotide polymorphisms: new experimental evidence and implications for human history and medicine. *Hum. Mol. Genet.*, 10:2195–2198. (Cited on pages 9, 10, 144, and 145.)

Mills, R. E., Luttig, C. T., Larkins, C. E., Beauchamp, A., Tsui, C., Pittard, W. S., and Devine, S. E. (2006). An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res.*, 16:1182–1190. (Cited on page 4.)

Minichiello, M. J. and Durbin, R. (2006). Mapping trait loci by use of inferred ancestral recombination graphs. *Am. J. Hum. Genet.*, 79:910–922. (Cited on page 44.)

Mitchell, A. A., Zwick, M. E., Chakravarti, A., and Cutler, D. J. (2004). Discrepancies in dbSNP confirmation rates and allele frequency distributions from varying genotyping error rates and patterns. *Bioinformatics*, 20:1022–1032. (Cited on page 98.)

Miyata, T., Hayashida, H., Kuma, K., Mitsuyasu, K., and Yasunaga, T. (1987). Male-driven molecular evolution: a model and nucleotide sequence analysis. *Cold Spring Harb. Symp. Quant. Biol.*, 52:863–867. (Cited on pages 13 and 118.)

Moloney, D., Slaney, S., Oldridge, M., Wall, S., Sahlin, P., Stenman, G., and Wilkie, A. (1996). Exclusive paternal origin of new mutations in Apert syndrome. *Nat. Genet.*, 13:48–53. (Cited on page 49.)

Moog-Lutz, C., Peterson, E. J., Lutz, P. G., Eliason, S., Cave-Riant, F., Singer, A., Di Gioia, Y., Dmowski, S., Kamens, J., Cayre, Y. E., and Koretzky, G. (2001). PRAM-1 is a novel adaptor protein regulated by retinoic acid (RA) and promyelocytic leukemia (PML)-RA receptor alpha in acute promyelocytic leukemia cells. *J. Biol. Chem.*, 276:22375–22381. (Cited on page 168.)

Morin, P. A., Moore, J. J., Chakraborty, R., Jin, L., Goodall, J., and Woodruff, D. S. (1994). Kin selection, social structure, gene flow, and the evolution of chimpanzees. *Science*, 265:1193–1201. (Cited on pages 111 and 112.)

Morrissy, S., Zhao, Y., Delaney, A., Asano, J., Dhalla, N., Li, I., McDonald, H., Pandoh, P., Prabhu, A. L., Tam, A., Hirst, M., and Marra, M. (2010). Digital gene expression by tag sequencing on the illumina genome analyzer. *Curr Protoc Hum Genet*, Chapter 11:1–36. (Cited on page 27.)

Mugal, C. F. and Ellegren, H. (2011). Substitution rate variation at human CpG sites correlates with non-CpG divergence, methylation level and GC content. *Genome Biol.*, 12:R58. (Cited on pages 9 and 117.)

Mugal, C. F., von Grunberg, H. H., and Peifer, M. (2009). Transcription-induced mutational strand bias and its effect on substitution rates in human genes. *Mol. Biol. Evol.*, 26:131–142. (Cited on page 11.)

Mullaney, J. M., Mills, R. E., Pittard, W. S., and Devine, S. E. (2010). Small insertions and deletions (INDELs) in human genomes. *Hum. Mol. Genet.*, 19:R131–R136. (Cited on page 4.)

Murray, V. (1989). Improved double-stranded DNA sequencing using the linear polymerase chain reaction. *Nucleic Acids Res.*, 17:8889. (Cited on page 184.)

Musumeci, L., Arthur, J. W., Cheung, F. S., Hoque, A., Lippman, S., and Reichardt, J. K. (2010). Single nucleotide differences (SNDs) in the dbSNP database may lead to errors in genotyping and haplotyping studies. *Hum. Mutat.*, 31:67–73. (Cited on page 99.)

Myers, S., Bowden, R., Tumian, A., Bontrop, R. E., Freeman, C., MacFie, T. S., McVean, G., and Donnelly, P. (2010). Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. *Science*, 327:876–879. (Cited on pages 106 and 107.)

Nachman, M. W. (1997). Patterns of DNA variability at X-linked loci in Mus domesticus. *Genetics*, 147:1303–1316. (Cited on page 122.)

Nachman, M. W. (2002). Variation in recombination rate across the genome: evidence and implications. *Curr. Opin. Genet. Dev.*, 12:657–663. (Cited on page 17.)

Nachman, M. W., Bauer, V. L., Crowell, S. L., and Aquadro, C. F. (1998). DNA variability and recombination rates at X-linked loci in humans. *Genetics*, 150:1133–1141. (Cited on page 122.)

Nachman, M. W. and Crowell, S. L. (2000). Estimate of the mutation rate per nucleotide in humans. *Genetics*, 156:297–304. (Cited on pages 2, 9, 14, 15, 49, 68, 173, and 174.)

Naski, M. C., Wang, Q., Xu, J., and Ornitz, D. M. (1996). Graded activation of fibroblast growth factor receptor 3 by mutations causing achondroplasia and thanatophoric dysplasia. *Nat. Genet.*, 13:233–237. (Cited on page 50.)

Navarro, A. and Barton, N. H. (2003a). Accumulating postzygotic isolation genes in parapatry: a new twist on chromosomal speciation. *Evolution*, 57:447–459. (Cited on pages 180 and 181.)

Navarro, A. and Barton, N. H. (2003b). Chromosomal speciation and molecular divergence–accelerated evolution in rearranged chromosomes. *Science*, 300:321–324. (Cited on page 181.)

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48:443–453. (Cited on page 40.)

Nei, M. (1987). *Molecular Evolutionary Genetics*. Columbia University Press, New York. (Cited on page 16.)

Nei, M. and Li, W. H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci. U.S.A.*, 76:5269–5273. (Cited on pages 16 and 117.)

Nei, M. and Li, W. H. (1980). Non-random association between electromorphs and inversion chromosomes in finite populations. *Genet. Res.*, 35:65–83. (Cited on page 154.)

Nevarez, P. A., DeBoever, C. M., Freeland, B. J., Quitt, M. A., and Bush, E. C. (2010). Context dependent substitution biases vary within the human genome. *BMC Bioinformatics*, 11:462. (Cited on pages 10 and 145.)

Ng, P., Wei, C. L., Sung, W. K., Chiu, K. P., Lipovich, L., Ang, C. C., Gupta, S., Shahab, A., Ridwan, A., Wong, C. H., Liu, E. T., and Ruan, Y. (2005). Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. *Nat. Methods*, 2:105–111. (Cited on page 192.)

Nielsen, R., Paul, J. S., Albrechtsen, A., and Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.*, 12:443–451. (Cited on pages 30, 42, 43, and 44.)

Nielsen, R., Williamson, S., Kim, Y., Hubisz, M. J., Clark, A. G., and Bustamante, C. (2005). Genomic scans for selective sweeps using SNP data. *Genome Res.*, 15:1566–1575. (Cited on page 180.)

Niu, T., Qin, Z. S., Xu, X., and Liu, J. S. (2002). Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *Am J Hum Genet*, 70:157–169. (Cited on page 102.)

Nordborg, M., Charlesworth, B., and Charlesworth, D. (1996). The effect of recombination on background selection. *Genet. Res.*, 67:159–174. (Cited on pages 121 and 122.)

Okou, D. T., Steinberg, K. M., Middle, C., Cutler, D. J., Albert, T. J., and Zwick, M. E. (2007). Microarray-based genomic selection for high-throughput resequencing. *Nat. Methods*, 4:907–909. (Cited on page 21.)

Olson, M. V. and Varki, A. (2003). Sequencing the chimpanzee genome: insights into human evolution and disease. *Nat. Rev. Genet.*, 4:20–28. (Cited on page 73.)

Orioli, I. M., Castilla, E. E., Scarano, G., and Mastroiacovo, P. (1995). Effect of paternal age in achondroplasia, thanatophoric dysplasia, and osteogenesis imperfecta. *Am. J. Med. Genet.*, 59:209–217. (Cited on pages 2, 4, and 66.)

Ornitz, D. M., Xu, J., Colvin, J. S., McEwen, D. G., MacArthur, C. A., Coulier, F., Gao, G., and Goldfarb, M. (1996). Receptor specificity of the fibroblast growth factor family. *J. Biol. Chem.*, 271:15292–15297. (Cited on page 50.)

Otting, N., Kenter, M., van Weeren, P., Jonker, M., and Bontrop, R. E. (1992). MHC-DQB repertoire variation in hominoid and Old World primate species. *J. Immunol.*, 149:461–470. (Cited on pages 165 and 169.)

Ozsolak, F., Platt, A. R., Jones, D. R., Reifenberger, J. G., Sass, L. E., McInerney, P., Thompson, J. F., Bowers, J., Jarosz, M., and Milos, P. M. (2009). Direct RNA sequencing. *Nature*, 461:814–818. (Cited on page 201.)

Paegel, B. M., Emrich, C. A., Wedemayer, G. J., Scherer, J. R., and Mathies, R. A. (2002). High throughput DNA sequencing with a microfabricated 96-lane capillary array electrophoresis bioprocessor. *Proc. Natl. Acad. Sci. U.S.A.*, 99:574–579. (Cited on page 186.)

Pantoliano, M. W., Horlick, R. A., Springer, B. A., Van Dyk, D. E., Tobery, T., Wetmore, D. R., Lear, J. D., Nahapetian, A. T., Bradley, J. D., and Sisk, W. P. (1994). Multivalent ligand-receptor binding interactions in the fibroblast growth factor system produce a cooperative growth factor and heparin mechanism for receptor dimerization. *Biochemistry*, 33:10229–10248. (Cited on page 50.)

Paten, B., Herrero, J., Beal, K., Fitzgerald, S., and Birney, E. (2008a). Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res.*, 18:1814–1828. (Cited on page 139.)

Paten, B., Herrero, J., Fitzgerald, S., Beal, K., Flicek, P., Holmes, I., and Birney, E. (2008b). Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome Res.*, 18:1829–1843. (Cited on page 139.)

Pearson, W. R. and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U.S.A.*, 85:2444–2448. (Cited on page 215.)

Peterson, A. C., Di Rienzo, A., Lehesjoki, A. E., de la Chapelle, A., Slatkin, M., and Freimer, N. B. (1995). The distribution of linkage disequilibrium over anonymous genome regions. *Hum. Mol. Genet.*, 4:887–894. (Cited on pages 18 and 156.)

Pettersson, E., Lundeberg, J., and Ahmadian, A. (2009). Generations of sequencing technologies. *Genomics*, 93:105–111. (Cited on pages 26, 186, 189, 191, and 199.)

Pfeifer, G. P., Denissenko, M. F., Olivier, M., Tretyakova, N., Hecht, S. S., and Hainaut, P. (2002). Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. *Oncogene*, 21:7435–7451. (Cited on page 3.)

Pfeifer, G. P., You, Y. H., and Besaratinia, A. (2005). Mutations induced by ultraviolet light. *Mutat. Res.*, 571:19–31. (Cited on page 3.)

Philippe, A.and Robert, C. (2001). Riemann sums for MCMC estimation and convergence monitoring. *Statist. Comput.*, 11:103–115. (Cited on page 62.)

Pink, C. J. and Hurst, L. D. (2010). Timing of replication is a determinant of neutral substitution rates but does not explain slow Y chromosome evolution in rodents. *Mol. Biol. Evol.*, 27:1077–1086. (Cited on pages 12 and 15.)

Pink, C. J., Swaminathan, S. K., Dunham, I., Rogers, J., Ward, A., and Hurst, L. D. (2009). Evidence that replication-associated mutation alone does not explain between-chromosome differences in substitution rates. *Genome Biol Evol*, 1:13–22. (Cited on pages 15 and 118.)

Polak, P. and Arndt, P. F. (2008). Transcription induces strand-specific mutations at the 5' end of human genes. *Genome Res.*, 18:1216–1223. (Cited on pages 10 and 11.)

Porreca, G. J., Zhang, K., Li, J. B., Xie, B., Austin, D., Vassallo, S. L., LeProust, E. M., Peck, B. J., Emig, C. J., Dahl, F., Gao, Y., Church, G. M., and Shendure, J. (2007). Multiplex amplification of large sets of human exons. *Nat. Methods*, 4:931–936. (Cited on page 21.)

Prendergast, J. G., Campbell, H., Gilbert, N., Dunlop, M. G., Bickmore, W. A., and Semple, C. A. (2007). Chromatin structure and evolution in the human genome. *BMC Evol. Biol.*, 7:72. (Cited on pages 9 and 11.)

Presgraves, D. C. and Yi, S. V. (2009). Doubts about complex speciation between humans and chimpanzees. *Trends Ecol. Evol. (Amst.)*, 24:533–540. (Cited on page 118.)

Pritchard, J. K. and Przeworski, M. (2001). Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.*, 69:1–14. (Cited on page 18.)

Pritchard, J. K. and Rosenberg, N. A. (1999). Use of unlinked genetic markers to detect population stratification in association studies. *Am. J. Hum. Genet.*, 65:220–228. (Cited on page 18.)

Prober, J. M., Trainor, G. L., Dam, R. J., Hobbs, F. W., Robertson, C. W., Zagursky, R. J., Cocuzza, A. J., Jensen, M. A., and Baumeister, K. (1987). A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science*, 238:336–341. (Cited on page 183.)

Przeworski, M. (2002). The signature of positive selection at randomly chosen loci. *Genetics*, 160:1179–1189. (Cited on page 122.)

Przeworski, M., Hudson, R. R., and Di Rienzo, A. (2000). Adjusting the focus on human variation. *Trends Genet.*, 16:296–302. (Cited on pages 122 and 137.)

Pushkarev, D., Neff, N. F., and Quake, S. R. (2009). Single-molecule sequencing of an individual human genome. *Nat. Biotechnol.*, 27:847–850. (Cited on page 199.)

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D. R., Li, J., Xu, J., Li, S., Li, D., Cao, J., Wang, B., Liang, H., Zheng, H., Xie, Y., Tap, J., Lepage, P., Bertalan, M., Batto, J. M., Hansen, T., Le Paslier,

D., Linneberg, A., Nielsen, H. B., Pelletier, E., Renault, P., Sicheritz-Ponten, T., Turner, K., Zhu, H., Yu, C., Li, S., Jian, M., Zhou, Y., Li, Y., Zhang, X., Li, S., Qin, N., Yang, H., Wang, J., Brunak, S., Dore, J., Guarner, F., Kristiansen, K., Pedersen, O., Parkhill, J., Weissenbach, J., Bork, P., Ehrlich, S. D., Wang, J., Antolin, M., Artiguenave, F., Blottiere, H., Borruel, N., Bruls, T., Casellas, F., Chervaux, C., Cultrone, A., Delorme, C., Denariaz, G., Dervyn, R., Forte, M., Friss, C., van de Guchte, M., Guedon, E., Haimet, F., Jamet, A., Juste, C., Kaci, G., Kleerebezem, M., Knol, J., Kristensen, M., Layec, S., Le Roux, K., Leclerc, M., Maguin, E., Minardi, R. M., Oozeer, R., Rescigno, M., Sanchez, N., Tims, S., Torrejon, T., Varela, E., de Vos, W., Winogradsky, Y., and Zoetendal, E. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464:59–65. (Cited on page 203.)

Quail, M. A., Swerdlow, H., and Turner, D. J. (2009). Improved protocols for the illumina genome analyzer sequencing system. *Curr Protoc Hum Genet*, Chapter 18:Unit 18.2. (Cited on page 27.)

Quinlan, A. R., Stewart, D. A., Strmberg, M. P., and Marth, G. T. (2008). Pyrobayes: an improved base caller for SNP discovery in pyrosequences. *Nat. Methods*, 5:179–181. (Cited on page 191.)

Raftery, A. E. and Lewis, S. (1992). How many iterations in the gibbs sampler? In *In Bayesian Statistics 4*, pages 763–773. Oxford University Press. (Cited on page 62.)

Rannan-Eliya, S. V., Taylor, I. B., De Heer, I. M., Van Den Ouweland, A. M., Wall, S. A., and Wilkie, A. O. (2004). Paternal origin of FGFR3 mutations in Muenke-type craniosynostosis. *Hum. Genet.*, 115:200–207. (Cited on pages 49 and 50.)

Rappold, G. A. (1993). The pseudoautosomal regions of the human sex chromosomes. *Hum. Genet.*, 92:315–324. (Cited on page 78.)

Raudsepp, T. and Chowdhary, B. P. (2008). The horse pseudoautosomal region (PAR): characterization and comparison with the human, chimp and mouse PARs. *Cytogenet. Genome Res.*, 121:102–109. (Cited on page 78.)

Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., Lavery, T., Kouyoumjian, R., Farhadian, S. F., Ward, R., and Lander, E. S. (2001). Linkage disequilibrium in the human genome. *Nature*, 411:199–204. (Cited on page 18.)

Reich, D. E. and Goldstein, D. B. (1998). Genetic evidence for a Paleolithic human population expansion in Africa. *Proc. Natl. Acad. Sci. U.S.A.*, 95:8119–8123. (Cited on page 123.)

Reymond, A., Meroni, G., Fantozzi, A., Merla, G., Cairo, S., Luzi, L., Riganelli, D., Zanaria, E., Messali, S., Cainarca, S., Guffanti, A., Minucci, S., Pelicci, P. G., and Ballabio, A. (2001).

The tripartite motif family identifies cell compartments. *EMBO J.*, 20:2140–2151. (Cited on page 165.)

Richards, R. I. and Sutherland, G. R. (1992). Fragile x syndrome: The molecular picture comes into focus. *Trends in Genetics*, 8:249–255. (Cited on page 6.)

Rideout, W. M., Coetzee, G. A., Olumi, A. F., and Jones, P. A. (1990). 5-Methylcytosine as an endogenous mutagen in the human LDL receptor and p53 genes. *Science*, 249:1288–1290. (Cited on pages 9, 10, and 145.)

Rieseberg, L. H. (2001). Chromosomal rearrangements and speciation. *Trends Ecol. Evol. (Amst.)*, 16:351–358. (Cited on page 180.)

Roach, J. C., Glusman, G., Smit, A. F., Huff, C. D., Hubley, R., Shannon, P. T., Rowen, L., Pant, K. P., Goodman, N., Bamshad, M., Shendure, J., Drmanac, R., Jorde, L. B., Hood, L., and Galas, D. J. (2010). Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science*, 328:636–639. (Cited on page 15.)

Robinson, J., Waller, M. J., Fail, S. C., McWilliam, H., Lopez, R., Parham, P., and Marsh, S. G. (2009). The IMGT/HLA database. *Nucleic Acids Res.*, 37:D1013–D1017. (Cited on page 164.)

Roche454 (2011). GS FLX+ System Instrument Specifications. Technical report, Roche454. Available from: `http://454.com/downloads/GSFLX+_InstumentSpecSheet.pdf`. (Cited on pages 188 and 191.)

Rogers, Y. H. and Venter, J. C. (2005). Genomics: massively parallel sequencing. *Nature*, 437:326–327. (Cited on page 183.)

Rogozin, I. B. and Pavlov, Y. I. (2003). Theoretical analysis of mutation hotspots and their DNA sequence context specificity. *Mutat. Res.*, 544:65–85. (Cited on page 15.)

Ronaghi, M., Uhlen, M., and Nyren, P. (1998). A sequencing method based on real-time pyrophosphate. *Science*, 281:363–365. (Cited on page 190.)

Roper, M. G., Easley, C. J., Legendre, L. A., Humphrey, J. A., and Landers, J. P. (2007). Infrared temperature control system for a completely noncontact polymerase chain reaction in microfluidic chips. *Anal. Chem.*, 79:1294–1300. (Cited on page 186.)

Rothberg, J. M. and Leamon, J. H. (2008). The development and impact of 454 sequencing. *Nat. Biotechnol.*, 26:1117–1124. (Cited on pages 21, 27, 51, 186, and 188.)

Rougemont, J., Amzallag, A., Iseli, C., Farinelli, L., Xenarios, I., and Naef, F. (2008). Probabilistic base calling of Solexa sequencing data. *BMC Bioinformatics*, 9:431. (Cited on pages 26, 52, and 55.)

Rozen, S., Skaletsky, H., Marszalek, J. D., Minx, P. J., Cordum, H. S., Waterston, R. H., Wilson, R. K., and Page, D. C. (2003). Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature*, 423:873–876. (Cited on page 71.)

Rumble, S. M., Lacroute, P., Dalca, A. V., Fiume, M., Sidow, A., and Brudno, M. (2009). SHRiMP: accurate mapping of short color-space reads. *PLoS Comput. Biol.*, 5:e1000386. (Cited on page 29.)

Rusk, N. (2009). Cheap third-generation sequencing. *Nature Methods*, 6:244–245. (Cited on page 204.)

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., Ackerman, H. C., Campbell, S. J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R., and Lander, E. S. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419:832–837. (Cited on page 180.)

Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G., Sherry, S., Mullikin, J. C., Mortimore, B. J., Willey, D. L., Hunt, S. E., Cole, C. G., Coggill, P. C., Rice, C. M., Ning, Z., Rogers, J., Bentley, D. R., Kwok, P. Y., Mardis, E. R., Yeh, R. T., Schultz, B., Cook, L., Davenport, R., Dante, M., Fulton, L., Hillier, L., Waterston, R. H., McPherson, J. D., Gilman, B., Schaffner, S., Van Etten, W. J., Reich, D., Higgins, J., Daly, M. J., Blumenstiel, B., Baldwin, J., Stange-Thomann, N., Zody, M. C., Linton, L., Lander, E. S., and Altshuler, D. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, 409:928–933. (Cited on pages 17, 118, 119, and 120.)

Sahni, M., Ambrosetti, D. C., Mansukhani, A., Gertner, R., Levy, D., and Basilico, C. (1999). FGF signaling inhibits chondrocyte proliferation and regulates bone development through the STAT-1 pathway. *Genes Dev.*, 13:1361–1366. (Cited on page 50.)

Sainz, J., Rovensky, P., Gudjonsson, S. A., Thorleifsson, G., Stefansson, K., and Gulcher, J. R. (2006). Segmental duplication density decrease with distance to human-mouse breaks of synteny. *Eur. J. Hum. Genet.*, 14:216–221. (Cited on page 5.)

Samonte, R. V. and Eichler, E. E. (2002). Segmental duplications and the evolution of the primate genome. *Nat. Rev. Genet.*, 3:65–72. (Cited on page 6.)

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.*, 74:5463–5467. (Cited on pages 21 and 183.)

Sarich, V. M. and Wilson, A. C. (1967). Rates of albumin evolution in primates. *Proc. Natl. Acad. Sci. U.S.A.*, 58:142–148. (Cited on page 111.)

Sasaki, S., Mello, C. C., Shimada, A., Nakatani, Y., Hashimoto, S., Ogawa, M., Matsushima, K., Gu, S. G., Kasahara, M., Ahsan, B., Sasaki, A., Saito, T., Suzuki, Y., Sugano, S., Kohara, Y., Takeda, H., Fire, A., and Morishita, S. (2009). Chromatin-associated periodicity in genetic variation downstream of transcriptional start sites. *Science*, 323:401–404. (Cited on page 11.)

Satta, Y. (2001). Comparison of DNA and protein polymorphisms between humans and chimpanzees. *Genes Genet. Syst.*, 76:159–168. (Cited on page 112.)

Sayres, M. A., Venditti, C., Pagel, M., and Makova, K. D. (2011). Do variations in substitution rates and male mutation bias correlate with life-history traits? A study of 32 mammalian genomes. *Evolution*, 65:2800–2815. (Cited on page 118.)

Scally, A. and Durbin, R. (2012). Revising the human mutation rate: implications for understanding human evolution. *Nat. Rev. Genet.*, 13:745–753. (Cited on page 151.)

Schadt, E. E., Turner, S., and Kasarskis, A. (2010). A window into third-generation sequencing. *Hum. Mol. Genet.*, 19:R227–R240. (Cited on pages 202 and 203.)

Schaffner, S. F. (2004). The X chromosome in population genetics. *Nat. Rev. Genet.*, 5:43–51. (Cited on pages 113 and 119.)

Schatz, M. C. (2009). CloudBurst: highly sensitive read mapping with MapReduce. *Bioinformatics*, 25:1363–1369. (Cited on page 29.)

Scheet, P. and Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.*, 78:629–644. (Cited on page 44.)

Schlessinger, J. (2000). Cell signaling by receptor tyrosine kinases. *Cell*, 103:211–225. (Cited on page 50.)

Sears, L. E., Moran, L. S., Kissinger, C., Creasey, T., Perry-O'Keefe, H., Roskey, M., Sutherland, E., and Slatko, B. E. (1992). CircumVent thermal cycle sequencing and alternative manual and automated DNA sequencing protocols using the highly thermostable VentR (exo-) DNA polymerase. *BioTechniques*, 13:626–633. (Cited on page 184.)

Seifert, R. A., Coats, S. A., Oganesian, A., Wright, M. B., Dishmon, M., Booth, C. J., Johnson, R. J., Alpers, C. E., and Bowen-Pope, D. F. (2003). PTPRQ is a novel phosphatidylinositol phosphatase that can be expressed as a cytoplasmic protein or as a subcellularly localized receptor-like protein. *Exp. Cell Res.*, 287:374–386. (Cited on page 168.)

Service, R. F. (2006). Gene sequencing. The race for the $1000 genome. *Science*, 311:1544–1546. (Cited on pages 21 and 201.)

Service, S. K., Ophoff, R. A., and Freimer, N. B. (2001). The genome-wide distribution of background linkage disequilibrium in a population isolate. *Hum. Mol. Genet.*, 10:545–551. (Cited on page 18.)

Shendure, J. and Ji, H. (2008). Next-generation DNA sequencing. *Nat. Biotechnol.*, 26:1135–1145. (Cited on pages 22, 26, 52, 55, 183, 185, 186, 189, 192, 193, 196, 199, and 201.)

Shendure, J., Mitra, R. D., Varma, C., and Church, G. M. (2004). Advanced sequencing technologies: methods and goals. *Nat. Rev. Genet.*, 5:335–344. (Cited on page 201.)

Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M., Wang, M. D., Zhang, K., Mitra, R. D., and Church, G. M. (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, 309:1728–1732. (Cited on pages 186 and 192.)

Shendure, J. A., Porreca, G. J., and Church, G. M. (2008). Overview of DNA sequencing strategies. *Curr Protoc Mol Biol*, Chapter 7:Unit 7.1. (Cited on page 185.)

Shendure, J. A., Porreca, G. J., Church, G. M., Gardner, A. F., Hendrickson, C. L., Kieleczawa, J., and Slatko, B. E. (2011). Overview of DNA sequencing strategies. *Curr Protoc Mol Biol*, Chapter 7:Unit7.1. (Cited on pages 196, 197, 198, 199, and 203.)

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic acids research*, 29:308–311. (Cited on pages 41, 43, 97, 141, and 142.)

Shi, E., Kan, M., Xu, J., Wang, F., Hou, J., and McKeehan, W. L. (1993). Control of fibroblast growth factor receptor kinase signal transduction by heterodimerization of combinatorial splice variants. *Mol. Cell. Biol.*, 13:3907–3918. (Cited on page 50.)

Shibata, K., Itoh, M., Aizawa, K., Nagaoka, S., Sasaki, N., Carninci, P., Konno, H., Akiyama, J., Nishi, K., Kitsunai, T., Tashiro, H., Itoh, M., Sumi, N., Ishii, Y., Nakamura, S., Hazama, M., Nishine, T., Harada, A., Yamamoto, R., Matsumoto, H., Sakaguchi, S., Ikegami, T., Kashiwagi, K., Fujiwake, S., Inoue, K., and Togawa, Y. (2000). RIKEN integrated sequence analysis (RISA) system–384-format sequencing pipeline with 384 multicapillary sequencer. *Genome Res.*, 10:1757–1771. (Cited on page 185.)

Shimmin, L. C., Chang, B. H., and Li, W. H. (1993). Male-driven evolution of DNA sequences. *Nature*, 362:745–747. (Cited on page 14.)

Siepel, A. and Haussler, D. (2004). Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol. Biol. Evol.*, 21:468–488. (Cited on pages 8, 10, 113, 117, 144, and 145.)

Silva, J. C. and Kondrashov, A. S. (2002). Patterns in spontaneous mutation revealed by human-baboon sequence comparison. *Trends Genet.*, 18:544–547. (Cited on pages 8, 10, 113, 117, and 145.)

Simonsen, K. L., Churchill, G. A., and Aquadro, C. F. (1995). Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics*, 141:413–429. (Cited on page 121.)

Singleton, A. B., Farrer, M., Johnson, J., Singleton, A., Hague, S., Kachergus, J., Hulihan, M., Peuralinna, T., Dutra, A., Nussbaum, R., Lincoln, S., Crawley, A., Hanson, M., Maraganore, D., Adler, C., Cookson, M. R., Muenter, M., Baptista, M., Miller, D., Blancato, J., Hardy, J., and Gwinn-Hardy, K. (2003). alpha-Synuclein locus triplication causes Parkinson's disease. *Science*, 302:841. (Cited on page 5.)

Slatkin, M. (2008). Linkage disequilibrium - understanding the evolutionary past and mapping the medical future. *Nat Rev Genet*, 9:477–485. (Cited on page 18.)

Slatko, B. E. (1996). Thermal cycle dideoxy DNA sequencing. *Mol. Biotechnol.*, 6:311–322. (Cited on page 184.)

Slierendregt, B. L., Kenter, M., Otting, N., Anholts, J., Jonker, M., and Bontrop, R. E. (1993). Major histocompatibility complex class II haplotypes in a breeding colony of chimpanzees (Pan troglodytes). *Tissue Antigens*, 42:55–61. (Cited on page 164.)

Smith, A. D., Chung, W. Y., Hodges, E., Kendall, J., Hannon, G., Hicks, J., Xuan, Z., and Zhang, M. Q. (2009). Updates to the RMAP short-read mapping software. *Bioinformatics*, 25:2841–2842. (Cited on page 29.)

Smith, A. D., Xuan, Z., and Zhang, M. Q. (2008). Using quality scores and longer reads improves accuracy of Solexa read mapping. *BMC Bioinformatics*, 9:128. (Cited on page 29.)

Smith, J. M. and Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genet. Res.*, 23:23–35. (Cited on page 122.)

Smith, N. G., Webster, M. T., and Ellegren, H. (2002). Deterministic mutation rate variation in the human genome. *Genome Res.*, 12:1350–1356. (Cited on pages 8, 12, 113, and 117.)

Smith, N. G., Webster, M. T., and Ellegren, H. (2003). A low rate of simultaneous double-nucleotide mutations in primates. *Mol. Biol. Evol.*, 20:47–53. (Cited on pages 10, 12, and 145.)

Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, 147:195–197. (Cited on page 40.)

Smukowski, C. S. and Noor, M. A. (2011). Recombination rate variation in closely related species. *Heredity (Edinb)*, 107:496–508. (Cited on page 17.)

Sniegowski, P. D., Gerrish, P. J., and Lenski, R. E. (1997). Evolution of high mutation rates in experimental populations of E. coli. *Nature*, 387:703–705. (Cited on page 15.)

Sol-Church, K., Stabley, D. L., Nicholson, L., Gonzalez, I. L., and Gripp, K. W. (2006). Paternal bias in parental origin of HRAS mutations in Costello syndrome. *Hum. Mutat.*, 27:736–741. (Cited on page 49.)

Spencer, C. C., Deloukas, P., Hunt, S., Mullikin, J., Myers, S., Silverman, B., Donnelly, P., Bentley, D., and McVean, G. (2006). The influence of recombination on human genetic diversity. *PLoS Genet.*, 2:e148. (Cited on page 10.)

Stamatoyannopoulos, J. A., Adzhubei, I., Thurman, R. E., Kryukov, G. V., Mirkin, S. M., and Sunyaev, S. R. (2009). Human mutation rate associated with DNA replication timing. *Nat. Genet.*, 41:393–395. (Cited on page 12.)

Stankiewicz, P. and Lupski, J. R. (2002). Genome architecture, rearrangements and genomic disorders. *Trends Genet.*, 18:74–82. (Cited on page 5.)

Stephens, M. and Donnelly, P. (2003). A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am. J. Hum. Genet.*, 73:1162–1169. (Cited on pages 44 and 102.)

Stephens, M. and Scheet, P. (2005). Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am. J. Hum. Genet.*, 76:449–462. (Cited on pages 44 and 102.)

Stephens, M., Smith, N. J., and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *American journal of human genetics*, 68:978–989. (Cited on pages 44 and 102.)

Stone, A. C., Griffiths, R. C., Zegura, S. L., and Hammer, M. F. (2002). High levels of Y-chromosome nucleotide diversity in the genus Pan. *Proc. Natl. Acad. Sci. U.S.A.*, 99:43–48. (Cited on pages 71 and 112.)

Strobeck, C. (1983). Expected linkage disequilibrium for a neutral locus linked to a chromosomal arrangement. *Genetics*, 103:545–555. (Cited on page 154.)

Surralles, J., Ramirez, M. J., Marcos, R., Natarajan, A. T., and Mullenders, L. H. (2002). Clusters of transcription-coupled repair in the human genome. *Proc. Natl. Acad. Sci. U.S.A.*, 99:10571–10574. (Cited on page 122.)

Sved, J. and Bird, A. (1990). The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. *Proc. Natl. Acad. Sci. U.S.A.*, 87:4692–4696. (Cited on pages 9, 10, and 145.)

Swerdlow, H. and Gesteland, R. (1990). Capillary gel electrophoresis for rapid, high resolution DNA sequencing. *Nucleic Acids Res.*, 18:1415–1419. (Cited on page 185.)

Tabor, S. and Richardson, C. C. (1995). A single residue in DNA polymerases of the Escherichia coli DNA polymerase I family is critical for distinguishing between deoxy- and dideoxyribonucleotides. *Proc. Natl. Acad. Sci. U.S.A.*, 92:6339–6343. (Cited on pages 183 and 184.)

Taillon-Miller, P., Bauer-Sardina, I., Saccone, N. L., Putzel, J., Laitinen, T., Cao, A., Kere, J., Pilia, G., Rice, J. P., and Kwok, P. Y. (2000). Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. *Nat. Genet.*, 25:324–328. (Cited on page 18.)

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123:585–595. (Cited on page 180.)

Takahata, N. and Nei, M. (1990). Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. *Genetics*, 124:967–978. (Cited on page 136.)

Tamura, K. and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.*, 10:512–526. (Cited on page 139.)

Tartaglia, M., Cordeddu, V., Chang, H., Shaw, A., Kalidas, K., Crosby, A., Patton, M. A., Sorcini, M., van der Burgt, I., Jeffery, S., and Gelb, B. D. (2004). Paternal germline origin and sex-ratio distortion in transmission of PTPN11 mutations in Noonan syndrome. *Am. J. Hum. Genet.*, 75:492–497. (Cited on page 49.)

Tawfik, D. S. and Griffiths, A. D. (1998). Man-made cell-like compartments for molecular evolution. *Nat. Biotechnol.*, 16:652–656. (Cited on page 189.)

Taylor, J., Tyekucheva, S., Zody, M., Chiaromonte, F., and Makova, K. D. (2006). Strong and weak male mutation bias at different sites in the primate genomes: insights from the human-chimpanzee comparison. *Mol. Biol. Evol.*, 23:565–573. (Cited on pages 13, 14, and 118.)

Teytelman, L., Eisen, M. B., and Rine, J. (2008). Silent but not static: accelerated base-pair substitution in silenced chromatin of budding yeasts. *PLoS Genet.*, 4:e1000247. (Cited on page 11.)

The 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467:1061–1073. (Cited on pages 42, 44, 73, 101, 137, 138, and 150.)

The 1000 Genomes Project Consortium (2011). Variation in genome-wide mutation rates within and between human families. *Nat. Genet.*, 43:712–714. (Cited on pages 8 and 113.)

The International Chimpanzee Chromosome 22 Consortium (2004). DNA sequence and comparative analysis of chimpanzee chromosome 22. *Nature*, 429:382–388. (Cited on pages 71 and 73.)

The International HapMap Consortium (2005). A haplotype map of the human genome. *Nature*, 437:1299–1320. (Cited on page 73.)

The International HapMap Consortium (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449:851–861. (Cited on pages 4 and 73.)

The MHC sequencing consortium (1999). Complete sequence and gene map of a human major histocompatibility complex. *Nature*, 401:921–923. (Cited on pages 136 and 157.)

The Mouse Genome Sequencing Consortium (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420:520–562. (Cited on pages 8, 113, 117, and 152.)

Thein, S. L. (2011). Genetic modifiers of sickle cell disease. *Hemoglobin*, 35:589–606. (Cited on page 4.)

Thornton, K. R., Jensen, J. D., Becquet, C., and Andolfatto, P. (2007). Progress and prospects in mapping recent selection in the genome. *Heredity (Edinb)*, 98:340–348. (Cited on page 113.)

Tian, D., Wang, Q., Zhang, P., Araki, H., Yang, S., Kreitman, M., Nagylaki, T., Hudson, R., Bergelson, J., and Chen, J. Q. (2008). Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. *Nature*, 455:105–108. (Cited on page 11.)

Tishkoff, S. A., Dietzsch, E., Speed, W., Pakstis, A. J., Kidd, J. R., Cheung, K., Bonne-Tamir, B., Santachiara-Benerecetti, A. S., Moral, P., and Krings, M. (1996). Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science*, 271:1380–1387. (Cited on page 18.)

Topal, M. D. and Fresco, J. R. (1976). Complementary base pairing and the origin of substitution mutations. *Nature*, 263:285–289. (Cited on page 4.)

Trapnell, C. and Salzberg, S. L. (2009). How to map billions of short reads onto genomes. *Nat. Biotechnol.*, 27:455–457. (Cited on pages 28 and 39.)

Trowsdale, J. (1993). Genomic structure and function in the MHC. *Trends Genet.*, 9:117–122. (Cited on page 164.)

Tyekucheva, S., Makova, K. D., Karro, J. E., Hardison, R. C., Miller, W., and Chiaromonte, F. (2008). Human-macaque comparisons illuminate variation in neutral substitution rates. *Genome Biol.*, 9:R76. (Cited on pages 9, 10, 12, and 13.)

Uchil, P. D., Quinlan, B. D., Chan, W. T., Luna, J. M., and Mothes, W. (2008). TRIM E3 ligases interfere with early and late stages of the retroviral life cycle. *PLoS Pathog.*, 4:e16. (Cited on page 165.)

Vajo, Z., Francomano, C. A., and Wilkin, D. J. (2000). The molecular and genetic basis of fibroblast growth factor receptor 3 disorders: the achondroplasia family of skeletal dysplasias, Muenke craniosynostosis, and Crouzon syndrome with acanthosis nigricans. *Endocr. Rev.*, 21:23–39. (Cited on page 50.)

Varki, A. and Altheide, T. K. (2005). Comparing the human and chimpanzee genomes: searching for needles in a haystack. *Genome Res.*, 15:1746–1758. (Cited on pages 71, 72, and 135.)

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guig, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M.,

Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., and Zhu, X. (2001). The sequence of the human genome. *Science*, 291:1304–1351. (Cited on pages 21 and 22.)

Verkerk, A. J., Pieretti, M., Sutcliffe, J. S., Fu, Y. H., Kuhl, D. P., Pizzuti, A., Reiner, O., Richards, S., Victoria, M. F., and Zhang, F. P. (1991). Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell*, 65:905–914. (Cited on page 6.)

Vicoso, B. and Charlesworth, B. (2006). Evolution on the X chromosome: unusual patterns and processes. *Nat. Rev. Genet.*, 7:645–653. (Cited on page 113.)

Voelkerding, K. V., Dames, S. A., and Durtschi, J. D. (2009). Next-generation sequencing: from basic research to diagnostics. *Clin. Chem.*, 55:641–658. (Cited on page 191.)

Voight, B. F., Kudaravalli, S., Wen, X., and Pritchard, J. K. (2006). A map of recent positive selection in the human genome. *PLoS Biol.*, 4:e72. (Cited on page 180.)

von Haeseler, A., Sajantila, A., and Paabo, S. (1996). The genetical archaeology of the human genome. *Nat. Genet.*, 14:135–140. (Cited on page 112.)

Wall, J. D. (2003). Estimating ancestral population sizes and divergence times. *Genetics*, 163:395–404. (Cited on page 151.)

Waller, D. K., Correa, A., Vo, T. M., Wang, Y., Hobbs, C., Langlois, P. H., Pearson, K., Romitti, P. A., Shaw, G. M., and Hecht, J. T. (2008). The population-based prevalence of achondroplasia and thanatophoric dysplasia in selected regions of the US. *Am. J. Med. Genet. A*, 146A:2385–2389. (Cited on pages 2, 4, 59, and 66.)

Wang, J., Wang, W., Li, R., Li, Y., Tian, G., Goodman, L., Fan, W., Zhang, J., Li, J., Zhang, J., Guo, Y., Feng, B., Li, H., Lu, Y., Fang, X., Liang, H., Du, Z., Li, D., Zhao, Y., Hu, Y., Yang, Z., Zheng, H., Hellmann, I., Inouye, M., Pool, J., Yi, X., Zhao, J., Duan, J., Zhou, Y., Qin, J., Ma, L., Li, G., Yang, Z., Zhang, G., Yang, B., Yu, C., Liang, F., Li, W., Li, S., Li, D., Ni, P., Ruan, J., Li, Q., Zhu, H., Liu, D., Lu, Z., Li, N., Guo, G., Zhang, J., Ye, J., Fang, L., Hao, Q., Chen, Q., Liang, Y., Su, Y., San, A., Ping, C., Yang, S., Chen, F., Li, L., Zhou, K., Zheng, H., Ren, Y., Yang, L., Gao, Y., Yang, G., Li, Z., Feng, X., Kristiansen, K., Wong, G. K., Nielsen, R., Durbin, R., Bolund, L., Zhang, X., Li, S., Yang, H., and Wang, J. (2008). The diploid genome sequence of an Asian individual. *Nature*, 456:60–65. (Cited on page 42.)

Washietl, S., Machne, R., and Goldman, N. (2008). Evolutionary footprints of nucleosome positions in yeast. *Trends Genet.*, 24:583–587. (Cited on page 13.)

Watkins, D. I. (1995). The evolution of major histocompatibility class I genes in primates. *Crit. Rev. Immunol.*, 15:1–29. (Cited on page 136.)

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theor Popul Biol*, 7:256–276. (Cited on pages 16 and 117.)

Webb, T. (1989). *The Fragile X Syndrome*. Oxford University Press. (Cited on page 6.)

Weber, J. L., David, D., Heil, J., Fan, Y., Zhao, C., and Marth, G. (2002). Human diallelic insertion/deletion polymorphisms. *Am. J. Hum. Genet.*, 71:854–862. (Cited on page 4.)

Webster, M. T., Smith, N. G., Lercher, M. J., and Ellegren, H. (2004). Gene expression, synteny, and local similarity in human noncoding mutation rates. *Mol. Biol. Evol.*, 21:1820–1830. (Cited on page 11.)

Weese, D., Emde, A. K., Rausch, T., Doring, A., and Reinert, K. (2009). RazerS–fast read mapping with sensitivity control. *Genome Res.*, 19:1646–1654. (Cited on page 29.)

Weksler, N. B., Lunstrum, G. P., Reid, E. S., and Horton, W. A. (1999). Differential effects of fibroblast growth factor (FGF) 9 and FGF2 on proliferation, differentiation and terminal differentiation of chondrocytic cells in vitro. *Biochem. J.*, 342:677–682. (Cited on page 50.)

Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y. J., Makhijani, V., Roth, G. T., Gomes, X., Tartaro, K., Niazi, F., Turcotte, C. L., Irzyk, G. P., Lupski, J. R., Chinault, C., Song, X. Z., Liu, Y., Yuan, Y., Nazareth, L., Qin, X., Muzny, D. M., Margulies, M., Weinstock, G. M., Gibbs, R. A., and Rothberg, J. M. (2008a). The complete genome of an individual by massively parallel DNA sequencing. *Nature*, 452:872–876. (Cited on page 188.)

Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., Dicuccio, M., Edgar, R., Federhen, S., Feolo, M., Geer, L. Y., Helmberg, W., Kapustin, Y., Khovayko, O., Landsman, D., Lipman, D. J., Madden, T. L., Maglott, D. R., Miller, V., Ostell, J., Pruitt, K. D., Schuler, G. D., Shumway, M., Sequeira, E., Sherry, S. T., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusov, R. L., Tatusova, T. A., Wagner, L., and Yaschenko, E. (2008b). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, 36:13–21. (Cited on page 21.)

Wicker, T., Schlagenhauf, E., Graner, A., Close, T. J., Keller, B., and Stein, N. (2006). 454 sequencing put to the test using the complex genome of barley. *BMC Genomics*, 7:275. (Cited on page 191.)

Wilkin, D. J., Szabo, J. K., Cameron, R., Henderson, S., Bellus, G. A., Mack, M. L., Kaitila, I., Loughlin, J., Munnich, A., Sykes, B., Bonaventure, J., and Francomano, C. A. (1998). Mutations in fibroblast growth-factor receptor 3 in sporadic cases of achondroplasia occur exclusively on the paternally derived chromosome. *Am. J. Hum. Genet.*, 63:711–716. (Cited on pages 49 and 50.)

Williams, E. J. and Hurst, L. D. (2000). The proteins of linked genes evolve at similar rates. *Nature*, 407:900–903. (Cited on pages 8, 113, and 117.)

Williamson, S. H., Hubisz, M. J., Clark, A. G., Payseur, B. A., Bustamante, C. D., and Nielsen, R. (2007). Localizing recent adaptive evolution in the human genome. *PLoS Genet.*, 3:e90. (Cited on page 180.)

Wilson, J. F. and Goldstein, D. B. (2000). Consistent long-range linkage disequilibrium generated by admixture in a Bantu-Semitic hybrid population. *Am. J. Hum. Genet.*, 67:926–935. (Cited on page 18.)

Wilson Sayres, M. A. and Makova, K. D. (2011). Genome analysis substantiate male mutation bias in many species. *Bioessays*, 33:938–945. (Cited on pages 13 and 118.)

Winckler, W., Myers, S. R., Richter, D. J., Onofrio, R. C., McDonald, G. J., Bontrop, R. E., McVean, G. A., Gabriel, S. B., Reich, D., Donnelly, P., and Altshuler, D. (2005). Comparison of fine-scale recombination rates in humans and chimpanzees. *Science*, 308:107–111. (Cited on pages 106 and 107.)

Wise, C. A., Sraml, M., Rubinsztein, D. C., and Easteal, S. (1997). Comparative nuclear and mitochondrial genome diversity in humans and chimpanzees. *Mol. Biol. Evol.*, 14:707–716. (Cited on pages 111 and 112.)

Wolfe, K. H. and Sharp, P. M. (1993). Mammalian gene evolution: nucleotide sequence divergence between mouse and rat. *J. Mol. Evol.*, 37:441–456. (Cited on pages 12, 14, 15, and 118.)

Wolfe, K. H., Sharp, P. M., and Li, W. H. (1989). Mutation rates differ among regions of the mammalian genome. *Nature*, 337:283–285. (Cited on pages 8, 10, 12, 13, 113, 117, and 121.)

Won, Y. J. and Hey, J. (2005). Divergence population genetics of chimpanzees. *Mol. Biol. Evol.*, 22:297–307. (Cited on page 71.)

Wright, A. F. (2008). *Handbook of Human Molecular Evolution*, chapter Genetic Variation: Polymorphisms and Mutations, pages 37–46. Wiley. (Cited on pages 6 and 111.)

Xing, J., Hedges, D. J., Han, K., Wang, H., Cordaux, R., and Batzer, M. A. (2004). Alu element mutation spectra: molecular clocks and the effect of DNA methylation. *J. Mol. Biol.*, 344:675–682. (Cited on pages 14, 117, 122, 131, and 176.)

Xue, Y., Wang, Q., Long, Q., Ng, B. L., Swerdlow, H., Burton, J., Skuce, C., Taylor, R., Abdellah, Z., Zhao, Y., MacArthur, D. G., Quail, M. A., Carter, N. P., Yang, H., and Tyler-Smith, C. (2009). Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. *Curr. Biol.*, 19:1453–1457. (Cited on page 15.)

Ying, H., Epps, J., Williams, R., and Huttley, G. (2010). Evidence that localized variation in primate sequence divergence arises from an influence of nucleosome placement on DNA repair. *Mol. Biol. Evol.*, 27:637–649. (Cited on page 12.)

Yu, B. and Mykland, P. (1998). Looking at Markov samplers through CUSUM path plots: a simple diagnostic idea. *Statist. Comput.*, 8:275–286. (Cited on page 62.)

Yu, N., Jensen-Seaman, M. I., Chemnick, L., Kidd, J. R., Deinard, A. S., Ryder, O., Kidd, K. K., and Li, W. H. (2003). Low nucleotide diversity in chimpanzees and bonobos. *Genetics*, 164:1511–1518. (Cited on page 112.)

Yunis, J. J. and Prakash, O. (1982). The origin of man: a chromosomal pictorial legacy. *Science*, 215:1525–1530. (Cited on page 180.)

Yunis, J. J., Sawyer, J. R., and Dunham, K. (1980). The striking resemblance of high-resolution G-banded chromosomes of man and chimpanzee. *Science*, 208:1145–1148. (Cited on page 180.)

Zankl, A., Elakis, G., Susman, R. D., Inglis, G., Gardener, G., Buckley, M. F., and Roscioli, T. (2008). Prenatal and postnatal presentation of severe achondroplasia with developmental delay and acanthosis nigricans (SADDAN) due to the FGFR3 Lys650Met mutation. *Am. J. Med. Genet. A*, 146A:212–218. (Cited on page 53.)

Zavolan, M. and Kepler, T. B. (2001). Statistical inference of sequence-dependent mutation rates. *Curr. Opin. Genet. Dev.*, 11:612–615. (Cited on pages 9, 10, and 145.)

Zerbino, D. R. and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.*, 18:821–829. (Cited on page 199.)

Zhang, J., Wang, X., and Podlaha, O. (2004). Testing the chromosomal speciation hypothesis for humans and chimpanzees. *Genome Res.*, 14:845–851. (Cited on page 181.)

Zhang, L., Lu, H. H., Chung, W. Y., Yang, J., and Li, W. H. (2005). Patterns of segmental duplication in the human genome. *Mol. Biol. Evol.*, 22:135–141. (Cited on page 6.)

Zhang, W., Bouffard, G. G., Wallace, S. S., and Bond, J. P. (2007). Estimation of DNA sequence context-dependent mutation rates using primate genomic sequences. *J. Mol. Evol.*, 65:207–214. (Cited on pages 8 and 117.)

Zhao, Z. and Boerwinkle, E. (2002). Neighboring-nucleotide effects on single nucleotide polymorphisms: a study of 2.6 million polymorphisms across the human genome. *Genome Res.*, 12:1679–1686. (Cited on pages 4, 8, 9, 10, 91, 113, 117, and 145.)

Zhao, Z. and Zhang, F. (2006). Sequence context analysis of 8.2 million single nucleotide polymorphisms in the human genome. *Gene*, 366:316–324. (Cited on pages 8 and 117.)

Zhu, Z. and Waggoner, A. S. (1997). Molecular mechanism controlling the incorporation of fluorescent nucleotides into DNA by PCR. *Cytometry*, 28:206–211. (Cited on page 199.)