The Genetics of Autoimmune and Proteinuric Disease

Katherine R Bull

St Cross College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2014

This thesis is dedicated to

Ben, Matilda and Ida

Acknowledgements

I would like to thank my supervisor, Professor Richard Cornall, for supporting and challenging me, but also giving me the freedom to pursue interesting questions.

I have been privileged to work with a great team in the Cornall lab over the last three years, particular thanks goes to Tanya Crockford for patiently teaching a total novice some lab skills, to Delphine Baup for help with lentiviral work, Sophia Bennett for getting me started with tissue culture, to Owen Siggs for many stimulating conversations and to the rest of the lab - Consuelo Anzilotti, Aron Chakera, Greg Crawford, Daian Chang, Mukta Deobagkar-Lele, Lucy Garner, Matthew Edmans and Xijin Xu for making the Cornall lab a friendly and fun place to work.

I am hugely grateful to Andy Rimmer, who not only collaborated on the IBD project but supported my attempts to learn Python, and gave much of his time throughout my studentship to assist me with bioinformatic problems and discuss my work. In all but name he has been my second supervisor.

I would also like to thank Gerton Lunter who patiently helped Andy and I with the harder maths, and spotted numerous bugs in our code.

Alistair Pagnamenta in Samantha Knight's group assisted with the SNP array.

Professor Chris Goodnow and his group at Australia National University,

including Anselm Enders and Belinda Whittle provided ENU mutants and DNA.

Steve Brown, Paul Potter, Michelle Simon and Ann-Marie Mallon at MRC Harwell provided sequences for union file analysis.

Bruce Beutler and Owen Siggs generated the strain used for validation of the quality of variant calls using the IBD method.

Professor Timothy Vyse (King's College London) and Professor Earl Silverman (University of Toronto) collaborated and provided SLE patient samples.

Professor Moin Saleem and Dr Hugh McCarthy (University of Bristol) collaborated and provided SRNS patient samples.

During my DPhil I was mentored by Professor Tony Hope, who met me with me every term to discuss the highs and lows and shared his years of experience and wisdom, continuing to meet with me even after his official retirement. Tony provided some much need perspective and moral support - often just at the right moment.

I would like to thank my parents, Alice and John Bull, for being endlessly patient, and never telling me I can't.

Finally I must thank my family, my two fabulous daughters Tilly and Ida, and my wonderful partner Ben, who thought he would see more of me if I became a student, and was sadly disappointed. Thank you Ben for the practical and emotional support and love that keeps me going.

Abstract

The genetics of complex common diseases are not fully understood, but rare variants with large phenotypic effects contribute to heritability. The objective of this thesis is to identify rare variants of relevance to autoimmune and renal disease, by developing ways of analysing whole genome sequencing (WGS) data and exploring the variants identified. Forward genetic experimental approaches are used, both in mutagenised mice, and in humans with extreme trait forms of the steroid resistant nephrotic syndrome (SRNS) and systemic lupus erythematosus (SLE). This work demonstrates that *N*–ethyl–*N*–nitrosourea (ENU) mutations can be distinguished within WGS data, including a hypomorphic mutation in *Lamb2* in a strain with the nephrotic syndrome, a murine model for the milder spectrum of human Pierson syndrome. In a B–cell deficient ENU strain the causative mutation in *Lyn* was isolated by sequencing multiple affected mice and applying an implementation of the Lander–Green algorithm to search for identity by descent. This method for the first time overcomes a rate-limiting step in ENU programmes and offers the potential to accelerate gene discovery, eliminating the need for out–crossing and conventional linkage analysis. Knowledge of the ENU genomic intervals allowed calculation of a mutation frequency, 1.5 mutations per mega base, and modelling of an efficient ENU strategy. Short–lists of candidate variants from 14 unrelated patients with steroid resistant nephrotic syndrome or systemic lupus erythematosus provide a substrate for future experiments, for example a candidate in the Wiskott–Aldrich syndrome gene was excluded using DNA from family members. In conclusion, WGS coupled with identity by descent analysis offers a powerful tool to improve the efficiency of ENU programmes. Rare variant discovery in humans without obvious Mendelian in-

heritance is more challenging and will require strategies to prioritise variants that combine bioinformatic filters and experimental verification in a high throughput way.

Contents

List of Figures

List of Tables

Acronyms

ABCA13 ATP-binding cassette subfamily A member 13.

ACTN4 actinin alpha 4.

ADH1C alcohol dehydrogenase 1C.

AGL amylo-alpha-1 6-glucosidase 4-alpha-glucanotransferase.

ARHGDIA Rho GDP Dissociation Inhibitor (GDI) Alpha.

BANK1 B cell scaffold protein with ankyrin repeats.

BLK B lymphoid tyrosine kinase.

C9ORF117 Chromosome 9 open reading frame 117.

CARD8 Caspase recruitment family member 8.

CBLC Cbl proto-oncogene C E3 ubiquitin protein ligase.

CCDC28A coiled coil domain containing 28A.

CCDC88C Coiled-coil domain containing 88c.

CD2AP CD2-associated protein.

CDHR2 Cadherin related family member 2.

CDON cell adhesion associated, oncogene regulated.

CHPF2 Chondroitin polymerising factor.

CMYA5 Cardiomyopathy 5.

COL4A4 Collagen type IV alpha 4.

DMBT1 Deleted in malignant brain tumors 1.

DNAH11 Dynein axomal heavy chain 11.

DNAH5 Dynein axonal heavy chain 5.

DNASE1l3 deoxyribonuclease I-Like 3 (human).

DNASE1 deoxyribonuclease I (human).

DOCK6 dedicator of cytokinesis 6.

DOCK8 dedicator of cytokinesis 6.

DST dystonin.

Dock2 dedicator of cytokinesis 2.

EFCAB4B EF-hand calcium binding domain 4B.

EHMT2 euchromatic histone- lysine N-methyltransferase 2.

ETS1 V-ETS avian erythroblastosis virus E26 oncogene homolog 1.

FAM76A family with sequence similarity 76, member A.

FOXG1 forkhead box G1.

GDF7 growth differentiation factor 7.

HOXD12 homeobox protein Hox-D12.

HSH2D Haematopoietic SH2 domain containing protein.

IFIH1 interferon induced with helicase C domain 1.

IKZF1 Ikaros family zinc finger 1.

IL12RB1 Interleukin 12 receptor beta 1.

INF1 inverted formin, FH2 and WH2 domain containing.

IQCE IQ motif containing E.

IRF5 interferon regulatory factor 5.

ITIH5 Inter-Alpha-Trypsin Inhibitor Heavy Chain Family, Member 5.

Ighm Immunoglobulin heavy constant mu (murine).

KRT27 keratin 27.

LZTS1 Leucine zipper, putative tumour suppressor 1.

Lamb2 / LAMB2 laminin b2 (murine/ human).

Lyn V-Yes-1 Yamaguchi Sarcoma Viral Related Oncogene Homolog (murine/ human).

MMP17 matrix metalloproteinase-17.

MUC16 Mucin 16.

MYO7B Myosin VIIB.

NCF2 neutrophil cytosolic factor 2.

NEB Nebulin.

NPHS1 nephrin.

OBSCN Obscurin cytoskeletal calmodulin and titin-interacting RhoGEF.

Olfr402 olfactory receptor 402 (murine).

Olfr979 olfactory receptor 979 (murine).

PABPC1P2 poly(A) binding protein 2.

PAGE1 P antigen family member 1.

PARP15 poly ADP ribose polymerase 15.

PLCE1 phospholipase C epsilon 1.

PLCL2 phospholipase C-like 2.

PNPLA7 Patatin-like phospholipase domain containing 7.

PPP6R1 Protein phosphatase 6, regulatory subunit 1.

PRKCA protein kinase C alpha (protein).

PSTPIP1 proline-serine-threonine phosphatase interacting protein 1.

PTPN22 protein tyrosine phosphatase non-receptor type 22.

PTPRG Protein tyrosine phosphatase receptor type G.

Prkca protein kinase C alpha (murine).

PtdIns(4,5)P2 Phosphatidylinositol-4,5-bisphosphonate.

RAET1E retinoic acid early transcript 1E.

RASAL3 Ras protein activator like 3.

RASGRP3 RAS guanyl releasing protein 3 (calcium and DAG-regulated).

RNF17 Ring finger protein 17.

SCN4A Sodium channel, voltage-gated, type IV.

SCRIB Scribbled Planar Cell Polarity Protein.

SENP6 Sentrin specific peptidase 6.

SEPT1 Septin 1.

SGK233 sugen kinase 233.

STARD9 StAR-related lipid transfer domain containing 9.

STAT4 signal transducer and activator of transcription 4.

STK11IP serine / threonine kinase 11-interacting protein.

Sppl2a signal peptide peptidase 2a (murine).

TATDN2 TatD DNase domain containing 2.

THSD7B thrombospondin, type I, domain containing 7B.

TNFSF4 tumor necrosis factor (ligand) superfamily member 4.

TNIP1 TNFAIP3-interacting protein.

TNK2 Tyrosine kinase non-receptor 2.

TTC37 tetratricopeptide repeat domain 37.

TTN Titin.

TYK2 tyrosine kinase 2.

Tlr4 Toll-like receptor 4 (murine).

Tlr7 Toll-like receptor 7 (murine).

Tnfsf13b tumor necrosis factor (ligand) superfamily, member 13b (murine) also known as BAFF.

Tnfsf1a tumor necrosis factor (ligand) superfamily, member 1a (murine).

UBE2L3 ubiquitin-conjugating enzyme E2L.

VWA7 Von Willebrand factor A domain containing 7.

WAS Wiskott Aldrich syndrome.

WT1 Wilms tumour 1.

XIAP X-linked inhibitor of apoptosis.

ZNF703 zinc finger protein 703.

AD autosomal dominant.

AGS Aicardi-Goutieres syndrome.

ALP alkaline phophatase.

ALT alanine transaminase.

ANA anti nuclear antibody.

ANKRD45 ankryin repeat domain containing 45.

ANU Australia National University.

ASCII American standard code for information interchange.

AST aspartate transaminase.

B6 C57BL/6J.

BAF B allele frequency.

BAFF B cell activating factor.

BIP binding immunoglobulin protein.

bp base pair.

CNS central nervous system.

CNV copy number variation.

Da dalton.

dA deoxyadenosine.

DHPLC denaturing high performance liquid chromatography.

DNA deoxyribonucleic acid.

DOG 1,2-Dioctanoyl-sn-glycerol.

dsDNA double stranded deoxyribonucleic acid.

dT deoxythymidine.

EM electron microscopy.

ENU N-ethyl-N-nitrosourea.

ER endoplasmic reticulum.

ESP Exome Sequencing Project (National Heart Lung and Blood Institute).

ESRD end stage renal disease.

FSGS focal segmental glomerulosclerosis.

G_1 first generation ENU offspring.

GBD GTPase binding domain.

GBM glomerular basement membrane.

GDP guanosine diphosphate.

GFP green fluorescent protein.

GO gene ontology.

GWAS genome wide association study.

HDL high density lipoprotein.

HGMD Human Gene Mutation Database.

HLA human lymphocyte antigen.

HMM hidden Markov model.

HUMARA human androgen receptor.

IBD identity by descent.

IBS identity by state.

IFN–α Interferon alpha.

IFN–γ Interferon gamma.

Ig immunoglobulin.

IgD immunoglobulin D.

IgG immunoglobulin G.

IGV integrated genomics viewer.

Kb kilobase.

LDL low density lipoprotein.

LOD logarithm of odds.

MAF minor allele frequency.

Mb mega base.

mRNA messenger ribonucleic acid.

MTB Mybacterium Tuberculosis.

NGS next generation sequencing.

NS non-synonymous.

NZB / NZW New Zealand black / New Zealand white.

OR odds ratio.

PBMC peripheral blood mononucleated cell.

PCR polymerase chain reaction.

PEI polyethylenimine.

RFLP restriction fragment length polymorphism.

RNA ribonucleic acid.

ROH runs of homozygosity.

SD standard deviation.

SLE systemic lupus erythematosis.

SLT specific locus test.

SNP single nucleotide polymorphism.

SRNS steroid resistant nephrotic syndrome.

SSLP simple sequence length polymorphism.

TBMD thin basement membrane disease.

TLR toll like receptor.

TNF tumor necrosis factor.

TNFAIP3 Tumor necrosis factor- alpha-induced protein.

TNFR1 Tumor Necrosis Factor Receptor 1.

Tnfrsf6 tumor necrosis factor receptor superfamily member 6 (Fas).

UPD uniparental disomy.

VCA Verprolin homology-cofilin homology acidic region.

VCF variant call format.

WES whole exome sequencing.

WGS whole genome sequencing.

WH1 WASP homology 1 domain.

WT wild type.

Chapter 1

Introduction

1.1 Overview

Autoimmune disorders affects 8% of the population and include numerous chronic and debilitating conditions (Cooper, Bynum, and Somers 2009). Proteinuria is a significant risk factor for renal disease and cardiovascular mortality (Kannel et al. 1984). Both autoimmunity and proteinuria are frequently associated with kidney disease, and can be part of monogenic rare disease or complex disease with a genetic component, in both cases our understanding of the genetics is incomplete. Current therapeutic options are often limited to broad spectrum immunosuppressive agents with significant side effects.

This thesis seeks to establish efficient methodological techniques to isolate genes carrying rare variants and assess their likely contribution to disease, in order to identify genetic variants with large phenotypic effects that may inform our understanding of rare Mendelian disease and more complex common diseases. Whilst predominantly methodological, the work also explores some functional biology in candidate genetic variants.

Knowledge of disease genetics will illuminate the biological pathways involved,

enriching our understanding of disease mechanisms. It may lead to better diagnostic testing and prognostic information for patients, and to the development of more rational treatments. In nephrology for example, identification of deficiency in alpha-galactoside A in Fabry disease due to mutations in *GLA* (galactosidase, alpha) led to effective treatment with enzyme replacement (Waldek and Feriozzi 2014). Further, understanding of the role of a gene in rare disease may aid the development of treatments for more common diseases. For example patients with rare mutations in the tumour suppressor Von Hippel-Lindau (*VHL*) develop multiple benign and malignant tumours, including renal cell carcinoma (Latif et al. 1993). 80% of sporadic cases of clear cell renal cell carcinoma, which accounts for 3% of all cancer deaths, have somatic mutations in *VHL* (Barry and Krek 2004). Drugs which target vascular endothelial growth factor, downstream of VHL, are effective in treating sporadic renal cell cancer (Clark and Cookson 2008).

Forward genetic approaches have provided an important tool for the discovery of such genes. Next generation sequencing (NGS) techniques such as whole genome sequencing (WGS) now allow the detection of rare variants not previously identified by methods such as microarrays.

Challenges remain however in human phenotyping, data collection, and sequencing analysis. Some of these challenges can be overcome by using mouse forward genetics to identify defects that cause disease.

This thesis explores forward genetic methods both in mouse models with point genetic mutations and humans with suspected single gene defects, developing a more rapid and efficient method to isolate causative mutations in mice subjected to ENU mutagenesis and exploring ways to filter and shortlist candidate rare variants in humans. Candidate variants are explored using genetic and molecular biology techniques.

Analysis of variant data from both ENU mutagenesis and humans suspected of

rare monogenic forms of complex disease, permits an informed comparison of the strengths and limitations of the two approaches.

1.2 Rare variants in rare and common disease

Single gene defects are a tractable approach to elucidate genes of importance both for rare single gene disorders and for understanding more complex disease. There are many monogenic heritable human diseases for which the genetic basis remains unknown, and the large majority of the 20,000 or so genes in the human genome have not been linked with any Mendelian disease. The discovery of the genes underlying rare familial disease is of importance for diagnosis and the development of effective therapies for the 6,000 to 7,000 diseases that individually affect less than 1 in 2,000 people (Montserrat Moliner and Waligora 2013), but collectively affect millions of people worldwide (Haffner, Whitley, and Moses 2002).

An understanding of these rare monogenic diseases can also provide insights into biological mechanisms of importance for more common complex disease. Large scale GWAS over the last decade have revealed many genetic loci associated with common disease. However for almost all diseases studied much less than half of the genetic contribution is known (Manolio et al. 2009). This failure to explain disease heritability has brought into question the 'common disease, common variant' model of disease genetics, in which multiple common variants, each conferring only a small relative risk, are hypothesised to combinatorially contribute to an overall disease risk. The explanation for the missing heritability is likely to be multifactorial in origin, with epistatic interactions between genes, epigenetic effects, complex interactions between genes and environmental factors and large variants such as insertions, deletions and duplications contributing to the 'missing heritability'. One explanation is the contribution of very rare variants, but with large individual effects (Cirulli and Goldstein

2010).

Rare variants are usually defined as being present at minor allele frequencies (MAF) of less than one percent in the population (Panoutsopoulou, Tachmazidou, and Zeggini 2013) but could be unique or private to families or individuals.

Variants with low MAF are less likely to reach significance in a GWAS study due to low frequency in cases. A further limitation of the GWAS approach, when based on array data, is that the variants identified are in most cases not causative variants, but in linkage with some unseen causative variant. This variant, or even the gene involved is often not identified (Goldstein 2009), limiting the value of GWAS variants for understanding disease mechanisms or gene function. These limitations can be overcome in some cases by imputation to infer genotypes at untyped rare variant loci using reference panels (Day-Williams et al. 2011). Furthermore the GWAS study design can be applied to WGS data (Panoutsopoulou, Tachmazidou, and Zeggini 2013). GWAS studies are most effective when the disease is genetically homogeneous, at least in the population studied, however a single disease phenotype, as clinically defined, can be due to a large number of different underlying genetic defects, often in multiple biological pathways.

The human de novo mutation rate is estimated to be 100 mutations per generation (Kondrashov 2002; Kong et al. 2012; Sun et al. 2012a). Due to relatively recent population expansion humans have accumulated large numbers of rare variants under only weak purifying selection. Recent estimates suggest that on average an individual carries 13,500 exonic variants, 86% of which have a MAF of less than 0.5% (Tennessen et al. 2012), a higher estimate predicts as many as 1 variant every 17 bases with MAF of less than 0.5% (Nelson et al. 2012). The rarer the variant the more likely it is to be coding, non–synonymous versus synonymous and predicted deleterious (Li et al. 2010; Marth et al. 2011), so rare variants are more likely than common variants to have functional consequences and cause disease. The collective abundance of rare variants

with functional effects points to a 'common disease , rare variant' explanation for the unknown genetic component in complex diseases.

In addition to rare monogenic disorders, many chronic diseases have familial forms due to rare variants of large effect. For example mutations in classical components of complement cause familial forms of SLE (Pickering and Walport 2000).

Rare variants may be causal for disease or modify disease severity by epistatic interactions with other genes. This dual role has been demonstrated for rare variants in the gene encoding the retrograde intraflagellar transport protein *IFT139*, some variants cause an autosomal dominant renal ciliopathy (nephronopthisis) while others are enriched in ciliopathy cases suggesting a contribution to pathogenic load (Davis et al. 2011). Likewise acquired deficiency of complement component C1q and C1q autoantibodies are observed frequently in SLE patients (Botto and Walport 2002).

It is clear that rare variants with large phenotypic effect sizes contribute to disease risk both in monogenic disease and in more complex, common disease, and that identification of such variants is a powerful method to explore gene function. However the extent to which such rare variants explain the genetic variance in complex disease remains unclear.

The total contribution of other factors to genetic variance, such as genotype by environment interactions, and cumulative effects of large numbers of variants with infinitesimal effects below the detection threshold of GWAS, is uncertain. However identifying rare large effect variants is likely to be the most amenable target for immediate study, particularly as NGS allows us to directly identify increasingly rare variants. The challenge in identifying these variants is in distinguishing those that are causative for a specific disease amongst a background of human genetic heterogeneity (McClellan and King 2010) (Cooper and Shendure 2011).

1.3 Forward genetics as a tool to uncover variants with phenotypic consequences

Forward genetics describes an approach to gene discovery that begins with a phenotype and asks what the underlying genetic defect is. The classical genetic study in human families with a Mendelian disease is a forward genetic experiment, and in the two decades since the discovery of the Huntingdon's gene locus (Gusella et al. 1983) many rare monogenic hereditary disease genes have been discovered, using linkage analysis in family pedigrees to home in on a shared causal genetic variant. Progress is limited by the rarity of these Mendelian families, the need to validate candidates and in some cases the practicalities of collecting accurate phenotypic data.

Forward genetic studies in animals include looking for genetic differences underlying strains arising from spontaneous or induced mutations, in either case proceeding from phenotype to genotype. These studies allow screening of many mutations, generating animal models of human phenotypes and revealing novel gene functions.

In any forward genetic strategy the underlying defect is not known at the outset and genetic linkage is used to converge on the causative variant. Linkage methods are based on the concept that the likelihood of inheriting two genetic loci from a single parental allele, without a recombination, is dependant on the physical distance between the two loci. Genes that are physically near to each other on the same chromosome are more likely to be inherited together. Genetic markers that segregate with a disease in families or pedigrees are used to pinpoint a genomic region and then a gene and a causative variant.

This gene discovery approach is complementary to reverse genetics in animal models, in which a gene or genetic locus is targeted based on a hypothesis about the possible biological consequences of a disruption; this hypothesis dictates the systems interrogated to detect a predicted phenotypic effect. The mutated model is then exam-

ined for phenotypic consequences. Classically reverse genetics has been based on gene knock–out, conditional knock–out or knock–in methods to create transgenic animals. Recent advances in gene targeting such as zinc finger endonucleases, transcription activator like effector nucleases (TALENs) and clustered regularly interspaced short palindromic repeats (CRISPR) systems have improved the precision and speed of gene targeting (Carlson, Fahrenkrug, and Hackett 2012; Wang et al. 2013). These tools allow functional investigation of a gene of interest in vivo. However for gene discovery, such methods remain costly and time consuming. Projects such as the International Mouse Knock Out Consortium and the associated International Mouse Phenotyping Consortium (Brown and Moore 2012) are generating and phenotyping large numbers of knock–out mice from embryonic stem cells. This important resource is limited to exploring the effects of null variants, and 30% will be embryonic lethals (Adams et al. 2013). Gene discovery in mice can be via QTL mapping of spontaneous mutations but this is difficult and often unsuccessful, for example most QTLs identified in mouse strains susceptible to SLE have not been narrowed down to a single causative gene (Morel 2010).

An alternative forward genetic approach is via induction of random mutations across the genome. Mutagenesis programmes screen large numbers of candidate genetic variants for a phenotype or disease model and allow investigators to focus only on genetic variants with phenotypic consequences. This method can be used to generate animals as models of human disease and identify single variants causing a phenotype in the manner of monogenic human disease or rare human variants of large effect. Unlike in the human Mendelian families, once a causative mutation has been identified, the mouse model is immediately available for further functional studies to confirm causality and explore the disease mechanism.

Various techniques have been used to induce mutations, including radiation, chemical mutagenesis and transposon mutagenesis. The alkylating agent *N*–ethyl–*N*–

nitrosourea (ENU) has been used for chemical mutagenesis for over 30 years due to its tolerability and powerful mutagen action (Russell et al. 1979). ENU induces point mutations, mimicking human variation, including viable hypomorphic or gain–of–function alleles, and illuminating gene function. ENU is administered to male mice, inducing mutations in spermatogonial stem cells. These mice are bred to generate first generation, (G_1, mice heterozygous for ENU mutations and screened for dominant phenotypes or third generation, (G_3), for recessive phenotypes (Figure 1.1). Simple primary screening assays for physiological, developmental, neurological cellular or gene regulatory processes can be followed up with more detailed and specific secondary experiments to define a phenotype. Immunisation challenge screens measuring cellular phenotypes by flow cytometry have revealed novel immuno-regulatory genes (Vinuesa et al. 2005; Vinuesa and Goodnow 2004). However the process of identifying underlying causative mutations has limited the efficiency of ENU. Conventionally this has required out crossing to another inbred laboratory strain of mouse and time consuming multiple breeding generations. Until now this has been necessary in order to map a linkage region using single nucleotide polymorphisms (SNPs) or simple sequence length polymorphisms (SSLP) that differ between the two strains. Tracking of the phenotype, strain specific modifiers, and the need for non–lethal phenotypic screens in order to propagate the mice, have all limited the efficacy of ENU programmes. Furthermore, until now estimates of the density of mutations induced by ENU have been varied and subject to biases. Confirmation of putative causative mutations may be possible based on published mouse variants for well characterised genes, but for novel genotype phenotype correlations will require complementation with existing mouse mutants or transgenic phenotypic rescue The identification of causative mutations is further discussed and addressed experimentally in Chapters 3, 4 and 5. The ENU mutation density is calculated in Chapter 5.

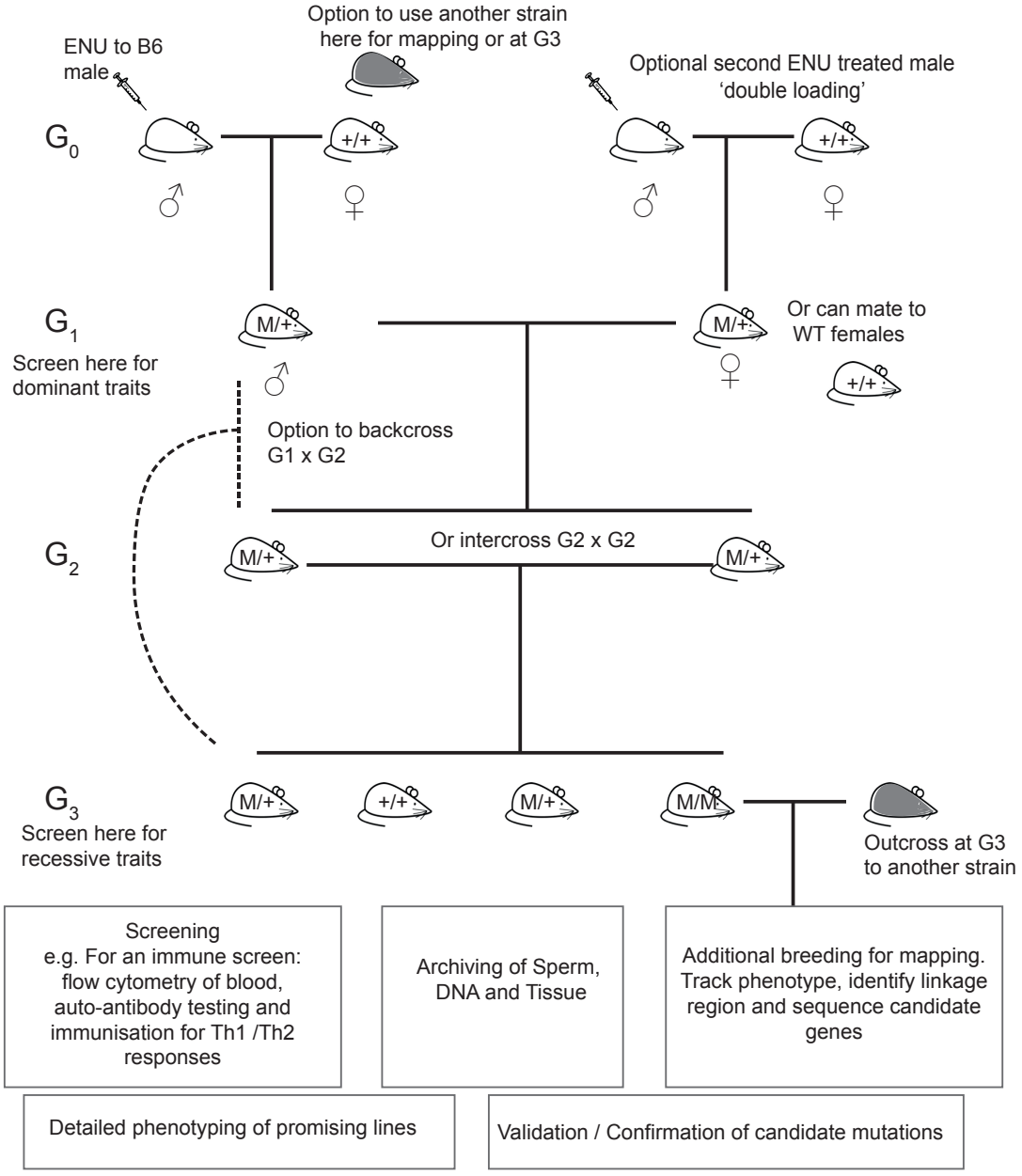


Figure 1.1: ENU treated male B57Bl/6J (B6) mice are bred to generate mice carrying heterozygous and homozygous mutations. M indicates mutant, + indicates wild type (WT). Generation 1 (G_1) or generation 3 (G_3) mice are screened for phenotypes. Selected mutant lines are conventionally mapped using markers for another inbred laboratory strain to which the mutants have been crossed. This outcross can take place at (G_0) or later at (G_3). The use of a second ENU treated male at G_0 or a backcross will increase the mutational burden, but may reduce fertility or litter size.

1.3.1 Mouse versus human forward genetics to discover disease causing genes

ENU mutagenesis has been applied to several model species, most notably the zebra fish and the mouse. The inbred laboratory mouse provides an attractive tool for forward genetic studies due to its homogeneous genetic background. Causative single mutations are not confounded due to interactions with different genetic modifiers in the background, and sequenced ENU mutations can be more easily distinguished due to lack of other genetic variation (Ermann and Glimcher 2012). The C57BL/6J (B6) mouse provided the deoxyribonucleic acid (DNA) for the primary mouse reference genome (Waterston et al. 2002) and thus mapping and variant calling of NGS data from this strain is relatively straightforward. Using a mouse model permits rapid breeding with full knowledge of pedigree structures, deep, quantitative, cellular phenotyping, and immediate access to mutant tissue, cells and whole animals for further functional studies. The mouse can be bred to manipulate the genetics, for example by generating complementary crosses and examining redundancy in pathways.

Although 99% of mouse genes have a homologue in the human genome, only 75% of mouse genes, or 80% of human genes have a 1:1 orthologue in the other species (Waterston et al. 2002; Church et al. 2009) making meaningful comparisons of mouse and human mutations possible for around 75% of genes. Despite this, interspecies differences can limit comparisons even when a mouse orthologue exists. Human phenotypes may be difficult to assay in the mouse (e.g. behavioural or psychological disease) and may differ from observed mouse phenotypes. For example DOCK8 deficient patients are characterised by hyper immunoglobulin E levels, a finding not replicated in the DOCK8 deficient ENU mouse (Lambe et al. 2011; Su, Jing, and Zhang 2011).

(Durbin et al. 2010) Despite the practical advantages of the inbred homogeneous mouse genetic background, this fails to model the complexities of background effects

on human disease. Conversely many mouse phenotypes are lost, modified or exacerbated when a causative locus is crossed onto a different background strain. For example mice with the Y-linked autoimmune accelerator (Yaa) mutation, a translocation of toll like receptor 7 (*Tlr7*) and adjacent genes from the X chromosome to the Y, exhibit double normal expression and increased constitutive and induced signalling by *Tlr7*. (Pisitkun et al. 2006; Subramanian et al. 2006). Yaa male mice on autoimmune prone backgrounds, such as BXSB, develop lupus-like disease and glomerulonephritis (Fossati et al. 1995), but B6.yaa mice are resistant (Izui et al. 1988).

Forward genetic studies in humans avoid these limitations by working directly with patients with the disease or phenotype, but identifying and recruiting families with a rare disease for such studies can be limiting. Increasingly physicians and researchers are collecting large databases of phenotypic information for rare diseases, such as the national renal rare disease database (www.renalradar.org).

Phenotyping in humans can be subjective, and influenced by unknown environmental variables. It is less straightforward to demonstrate causation for a variant on a background of human genetic heterogeneity than in the mouse, but once identified, such genes will be directly relevant for human disease. Chapters 6 explores the limitations and challenges in applying WGS sequencing approaches to humans, searching for rare variants with large effect in complex disease, and examines ways to select and prioritise candidate variants in such individuals.

Ultimately, identifying all genetic variants with large effect sizes will require an integrated approach based on human and model systems, including reverse genetic, gene targeting approaches to validate and explore the function of human genetic variants and forward genetics in mice to accelerate gene discovery, which will be of most value when it can be related back to humans.

1.4 Whole genome sequencing for rare variants

1.4.1 Whole genome sequencing to identify rare variants in humans

WGS targets an individual's complete DNA sequence. The most widely used current techniques rely on generating short reads based on fragments of the sequence being investigated, and then reconstructing these, either by mapping reads onto a pre-determined species specific genome or by de novo assembly of overlapping reads to create a complete genomic sequence (Metzker 2010). The development of WGS techniques has made it possible to determine the sequence of many different species, and many individual humans, revealing inter and intra species variation on an unprecedented scale and advancing medical, population and evolutionary genetics.

The advent of whole genome and whole exome sequencing (WES) techniques has accelerated the pace of disease gene discovery. The human genome was completed in 2001 using Sanger chain-termination sequencing (Lander et al. 2001), since then massively parallel sequencing techniques including pyrosequencing and sequencing by synthesis, and more recently single molecule and ultra long read sequencing techniques, have led to rapid advances in sequencing capacity and speed. Since 2010, when the work in this thesis was initially conceived, targeted exome capture, whole exome sequencing and whole genome sequencing have moved from research methods to also become increasingly widespread clinical tools (Ng et al. 2010c; Bamshad et al. 2011; Katsanis and Katsanis 2013). The parallel advances in next generation sequencing (NGS) technology, and scientific discoveries with clinical utility based on these techniques, are illustrated in Figure 1.2.

Ascribing causality for rare variants is relatively straightforward for diseases inherited in an autosomal recessive fashion. In such cases, the aetiology is clearly monogenic, and the causative variant can be narrowed down to homozygous vari-

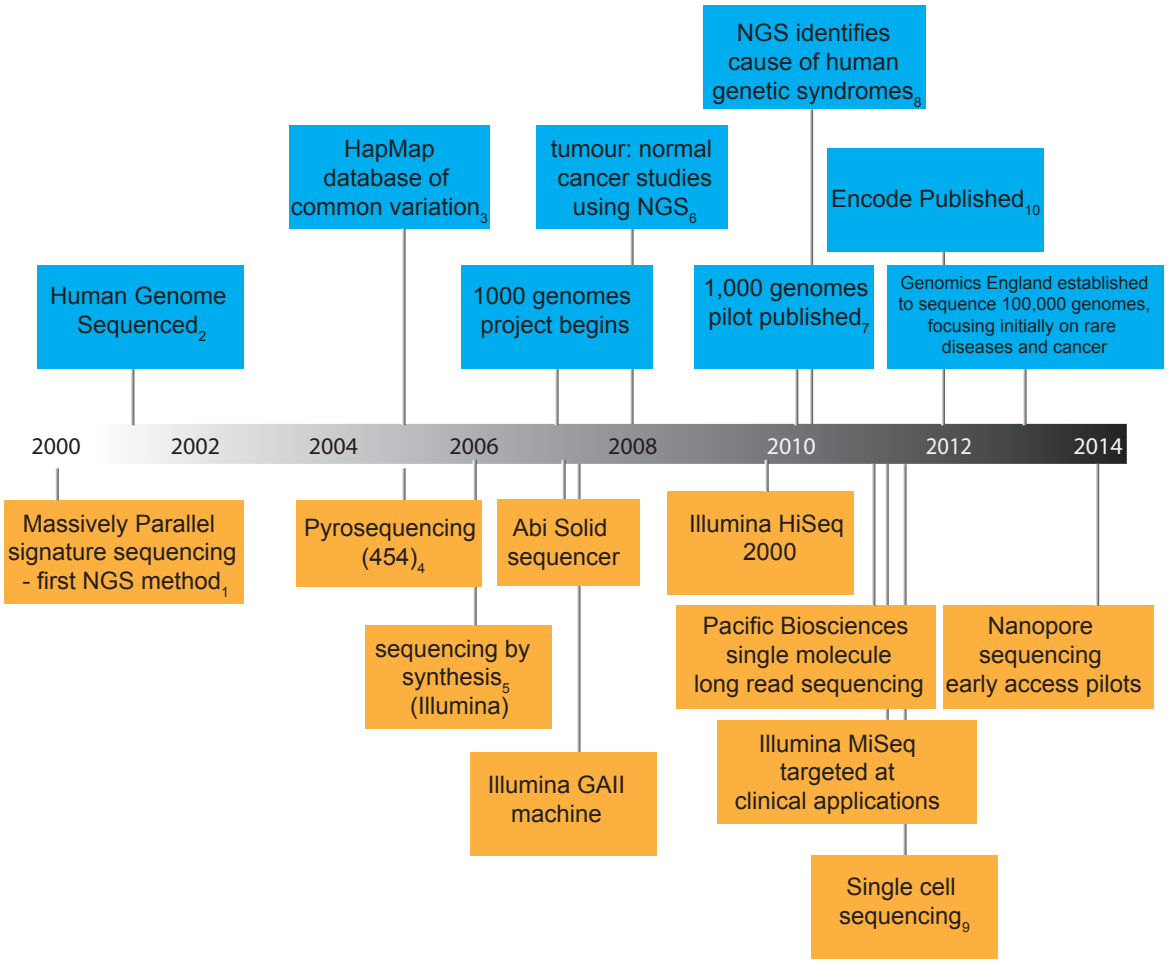


Figure 1.2: Next generation sequencing (NGS) timeline. Relevant technological developments in yellow and examples of key publications and projects in blue. Key references: (1) (Brenner et al. 2000) , (2) (Lander et al. 2001), 3)(International HapMap Consortium 2005), (4) (Margulies et al. 2005), (5)(Bentley et al. 2008), (6)(Chiang et al. 2008), (7)(Durbin et al. 2010), (8) (Lupski et al. 2010; Roach et al. 2010; Ng et al. 2010b), (9) (Navin et al. 2011), (10) (Consortium 2012).

ants observed in only in affected individuals. This method is most effective when more than one affected individual from a family with the inherited disease can be sequenced, and unaffected parents can be checked for heterozygosity at the locus. For example, WES of sisters with steroid resistant nephrotic syndrome (SRNS) and their consanguineous parents revealed a shared loss of function mutation in Rho GDP Dissociation Inhibitor (GDI) Alpha (*ARHGDIA*), which encodes a Rho guanosine diphosphate (GDP) dissociation inhibitor expressed in podocytes (Gupta et al. 2013). Two thirds of mutations in rare disease genes identified using WES to date cause autosomal recessive disease (Boycott et al. 2013). De novo dominant mutations arising as sporadic cases of early onset disease are also relatively easy to identify, by sequencing a trio of affected child and both parents (Gibson et al. 2012). The number of candidate variants will be much larger for familial autosomal dominant diseases; however if a linkage region is known these variants can also be isolated (Palles et al. 2012). Alternatively, for sporadic cases, sequencing of multiple unrelated individuals can be effective, as shown for *MLL2* and Kabuki syndrome based on 53 cases (Ng et al. 2010a).

WGS or WES can also be used to directly identify rare variants causative for rare monogenic forms of more complex disease. This can be achieved by studying a family with an inherited form of the complex disease, such as was demonstrated by the identification of a family with autosomal recessive SLE due to a null mutation in the deoxyribonuclease I-Like 3 (*DNASE1L3*) gene (Al-Mayouf et al. 2011). Mutations in the related gene deoxyribonuclease I-Like 3 (*DNASE1*) have been linked to monogenic SLE, and DNASE1 activity is reduced in SLE patients with multi-genetic aetiology (Martínez Valle et al. 2008), highlighting the relevance of monogenic forms in aiding understanding of the complex disease.

WES studies have also been applied to complex disease cases without evidence of Mendelian inheritance. In autism, highly penetrant copy number variants have

been shown to influence disease risk, suggesting that other variants of large effect may also play a role. Studies of sporadic autism cases (including trio analysis) have identified de novo variants as contributing to risk in at least 10% of cases, being enriched in neurodevelopmental genes in cases and highlighting candidate genes for further study. However most are incompletely penetrant, such that individual de novo variants are not sufficient to fully explain the disease in a patient, and the limited number of genes with variants in multiple patients illustrates the genetic heterogeneity of autism (Sanders et al. 2012a; Neale et al. 2012; O'Roak et al. 2012).

An alternative to sequencing affected individuals in families is the extreme trait sequencing method (Cirulli and Goldstein 2010). This offers a solution to the scarcity of familial cases. Unrelated individuals with severe or early onset forms of the disease are sequenced, based on the assumption that single rare variants of large effect will be enriched in a population at the extreme end of the disease spectrum. The power of such a method will be limited by the genetic heterogeneity of the disease being studied, and by how effectively the extreme phenotype chosen selects for patients with a monogenic aetiology. In all studies in which isolated and unrelated cases are examined, both clinical and genetic heterogeneity will limit comparisons or associations across patients. However a disease-causing variant in only a single patient may be detected if it can be distinguished as likely to be functional and other candidates eliminated. This may be achieved using knowledge of the likely mode of inheritance and comparison with other unaffected family members, combined with validation in larger cohorts. In a WGS study of 962 individuals across the spectrum of a continuous trait: high density lipoprotein cholesterol (HDL–C) levels, the majority of genetic variability (61.8%) was attributable to multiple common variants of small effect, however a few individuals in the tails of the phenotype distribution had single rare variants predicted to have large phenotypic effect, some of which could be immediately validated as known Mendelian variants influencing HDL–C, these variants

were estimated to explain 7.8% of the population variance (Morrison et al. 2013).

1.4.2 Whole genome sequencing or whole exome sequencing for rare variant studies

WES has been the NGS tool of choice for the majority of human rare disease genetic studies and ENU research published to date. WES has been favoured over WGS for economic reasons, and because the majority of known disease variants are in coding regions. WES techniques capture the 1–2% of the genome that consists of coding exons and splice sites using target enrichment methods. Although the emphasis in publications on coding variation reflects an ascertainment bias, our ability to predict the likely functional consequences of non-coding variation is currently limited. The role of non–coding variants in human disease in increasingly recognised (Khurana et al. 2013) and non protein–coding ENU variants have been reported (Lewis et al. 2009; Masuya et al. 2007).

Despite the preference for WES in published studies, there are advantages in moving towards a WGS approach. Both WES and WGS methods are subject to non uniform coverage depth due to a stochastic distribution (Lander and Waterman 1988), GC content (Dohm et al. 2008) and difficulty mapping to repetitive regions. Coverage is more uneven with WES than WGS due to biases in targeted capture methods, hence higher mean coverage depths are required to ensure sensitivity to coding variants, and some coding regions will be consistently difficult to capture (Sims et al. 2014). As the per base price of sequencing falls, and targeted capture and library preparation become a larger proportion of overall cost, WGS will become increasingly affordable. Furthermore WES fails to detect both non–coding variation and the majority of structural variation, both of which are more accessible through WGS. WGS offers a duel benefit for forward genetics, permitting linkage analysis to reduce the search space for a causative variant, and identification of both coding and

non–coding variation within such a region. Chapter 5 demonstrates that WGS rather than WES is necessary for identity by descent based fine mapping in ENU treated mice due to the low density of ENU SNPs.

1.4.3 Inferring causality and functional significance from whole genome sequencing data

Within this thesis 'pathogenic' means leading to disease and the words 'causal' and 'causative' are used to describe a variant that directly gives rise to a phenotype or disease. The term 'candidate' is used to refer to variants that have been prioritised as possible disease causing variants based on parameters such as rarity and effect on protein sequence, but which have not yet been demonstrated to be causal (Table 1.1).

For the ENU mutations described in this work, causality was inferred based on previous publication of similar variants in the same genes associated with consistent mouse phenotypes (Verhagen et al. 2009; Chen, Kikkawa, and Miner 2011), or based on segregation of the mutation with the phenotype and rescue experiments by collaborators (Bergmann et al. 2013).

The strength of available evidence for causality in humans varies, but to ascribe causality to a novel genetic variant in a clinical setting requires several lines of supportive evidence, including segregation with disease in families or case control populations, in silico, in vitro and in vivo evidence that the variant alters protein expression or function, this ideally indicating a mechanism for the disease, identification of the variant or other variants in the same gene in additional cohorts of patients with the same disease and if possible, replication of the genotype and phenotype in an animal model. For clinical interpretation, great caution must be exercised. A genetic variant is not a clinical diagnosis. Even variants previously reported to cause disease may not be pathogenic, 15% of reported cardiomyopathy variants are found, at 1000 times the prevalence of cardiomyopathy, in the general population (Andreasen et al. 2013).

There are two key challenges in identifying causative variants from WGS data. The first is distinguishing true variants from errors generated in the sequencing, alignment and variant calling pipeline. The sources of error in WGS and methods to filter out spurious variant calls, applicable to mouse and human data are described in chapter 3.

The second challenge is to prioritise variants likely to have functional consequences and influence the disease process. In humans the number of candidate variants is typically large due to genetic heterogeneity and abundance of rare variants. This can be addressed by statistical methods and family data, and by computational, in vitro and in vivo methods to test functional effects.

Statistical methods include prioritising variants by comparison with datasets of known variation with known population frequency such as the 1000 Genomes Project (Durbin et al. 2010) and Exome Sequencing Project (Tennessen et al. 2012) to establish rarity. Some ethnicities are at present underrepresented in control datasets, reducing the power of this method to isolate rare variants. Caution must be exercised when interpreting the absence of a variant in population controls as an indication of pathogenicity. Modelling suggests that there is a greater than one percent chance of a benign variant in a single patient not being observed in 10,000 controls, due to the excess of rare alleles in humans, the impact of natural selection and human migrations resulting in population specific variants (Sunyaev 2012). Thus most variants seen in only one sample will nevertheless be benign. The converse situation, exclusion of a variant that is present in other datasets, can also be erroneous as some datasets include patients with known disease and many provide limited or no phenotype data. Excluding such variants assumes complete penetrance of the phenotype irrespective of genetic background.

The statistical methods can be combined with knowledge of the segregation of variants within a family to reduce the number of candidate variants, particularly if

the mode of inheritance is known. Finally confirmation of the variant as causal may be possible by identifying allelic variants in additional unrelated individuals with the disease from a cohort.

In most cases further filtering of candidates based on predicted or observed functional consequences of variation is required. This will usually require a combination of computational functional prediction, and experimental testing of variants. Computational prediction methods allow rapid prioritisation of large numbers of candidates, based on evidence of local phylogenetic conservation, typically using a combination of multiple sequence alignment and prediction of the effect of variation on protein structure. Various algorithms exist to perform these predictions but there is considerable discordance between tools (Li et al. 2013; Hicks et al. 2011). The accuracy of these computational methods is limited by the relevance of the training datasets to the disease model and variant type being studied, and by current knowledge of protein structures. It is important to note that a deleterious variant, under purifying selection, or a damaging variant that alters protein function, may not necessarily be pathogenic, that is lead to a disease phenotype (Table 1.1). Experimental testing of variant function is more accurate but more time consuming and is therefore usually reserved for a short–list or single candidate following the statistical and computational filtering described above. This can involve in vitro analysis of protein expression and function, or in vivo studies using model organisms, such as mouse knock–outs of the gene of interest followed by experimental rescue by expression of wild type or variant protein.

Term	Definition
Deleterious	Under purifying selection
Damaging	Alters protein function
Candidate	Putative disease causing variant - unconfirmed
Pathogenic / Causal	Directly leads to disease

Table 1.1: Definition of terms used to describe variants

1.5 Summary

This thesis explores forward genetic methods to identify gene variants that are causative for disease. Two complementary approaches are applied and compared, ENU mutagenesis in mice and methods to identify rare variants of large effect in humans.

Chapter 3 demonstrates that WGS can be used to identify induced mutations in ENU mice, by filtering variant calls. Causative mutations are isolated in individuals from ENU pedigrees using knowledge of a linkage region and low coverage sequencing.

Chapter 4 applies murine mutagenesis to identify a novel variant causing steroid resistant nephrotic syndrome (SRNS), a predominantly monogenic disease with a missing heritable component. A mutation in Laminin $\beta 2$ (*Lamb2*) is shown to cause nephrotic syndrome in the ENU mouse *nephertiti*. Immunohistochemistry confirms a hypomorphic phenotype. *Nephertiti* provides a model for the milder spectrum of human Pierson syndrome. This work shows that ENU mutagenesis coupled with high throughput sequencing can identify novel variants that model human SRNS.

Chapter 5 develops an implementation of the Lander–Green algorithm to simultaneously identify IBD linkage regions and isolate underlying causative mutations, permitting isolation of damaging ENU mutations without the need for outcrossing. This work also leads to an estimate of the ENU mutation density that avoids biases inherent in previous estimates and permits modelling of an efficient forward genetic ENU programme.

Chapter 6 examines the efficacy of an extreme trait WGS method to look for rare variants that cause SRNS or systemic lupus erythematosus (SLE) a complex and heterogeneous disease. Short–lists of candidate variants provide a substrate for future functional experiments. The chapter demonstrates the effect of applying filters to the candidate variants and the limitations and challenges of WGS in isolated human disease cases.

Chapter 2

Methods

2.1 Reagents

Recipes for reagents used in the experiments described in the methods are given in table 2.1.

Reagent	Recipe
5X Annealing Buffer	400 mM Tris-HCL (pH 9.0), 10 mM MgCl2
TAE Buffer	10mM Tris-acetate, 1mM EDTA from pH 8.0 stock
TE Buffer	10 mM Tris-HCL (pH 7.5), 1mM EDTA from pH 8.0 stock
Ampicillin	50mg/ml stock stored at −20°C; powder from Sigma
L-broth (LB)	1% (w/v) Bactotryptone, 0.5% (w/v) yeast extract, 1% (w/v) NaCl
LB-plates	LB plus 1.5% (w/v) agar
Foetal calf serum (FCS)	European origin. Heat inactivated at 55°C for 40 minutes, stored at −20°C
D10 medium	500mls Dulbecco's modified Eagles medium (DMEM) (Sigma) plus 20mM (5mls) 4–(2–hydroxyethyl)–1–piperazineethanesulfonic acid (HEPES) (Invitrogen), 0.05 mM 2–mercaptoethanol (Sigma), 2mM L-glutamine (Sigma) plus or minus 105 U/L Penicillin
EMEM/ HAMs medium	EMEM (EBSS):Ham's F12 (1:1) + 2mM Glutamine + 1% Non Essential Amino Acids (NEAA) + 15% Foetal Bovine Serum (FBS) (Heat Inactivated)
TBS	50mM Tris-HCL pH 7.6, 150 mM NaCl
Periodic acid Schiff and Methenamine Silver reagents	Richard Allen Scientific
Scots tap water	Magnesium sulphate 10.0g, Sodium Bicarbonate 0.67g, tap water 1 litre
Flow cytometry staining media	Hanks balanced salt solution, 2% FCS, 0.1% NaN3 (sodium azide)

Table 2.1: Reagents, including buffers, bacterial culture reagents, cell culture reagents, histology and flow cytometry reagents

2.2 Generation of ENU mice

ENU treated B6 mice *ENU16CH17a*, *222*, *NIH69b*, *NIH85a*, *NIH19a*, *11_007*, *12_007*, *nephertiti*, *APFNS1015_17* and *ENU22* were generated at the Australian Phenomics Facility, The Australia National University (ANU), Canberra (Nelms and Goodnow 2001).

Male founder mice for each pedigree, 8–15 weeks old, were treated three times 1 week apart with 90–100mg/kg *N*–ethyl–*N*–nitrosourea (Sigma) prepared in 10% ethanol, citrate buffer (pH 5.0). After 8 weeks, treated mice were mated with B6

females. Individual first generation ENU offspring (G_1) progeny were inter–crossed to generate G_2 pairs.

Bruce Beutler's group (Scripps Institute) generated the mouse used to validate the ENU mutation density. This strain was generated using a similar ENU protocol, but the pedigree was generated from a G_1 mouse crossed with an ENU founder. This increases the burden of mutation by increasing the proportion of the genome carrying ENU mutations but will not affect the density of mutations in these regions.

Two strains were treated with ENU using the same dosing, but only one ENU treated founder, at the MRC Harwell facility. The Harwell ENU pipeline currently uses B6 mice crossed with a second strain for mapping purposes in their ENU programme. Both Harwell mice used in the union file were B6 x C3H.

2.3 Mouse phenotyping

2.3.1 Phenotyping of ENU mice in Canberra

Phenotypic screening of G_3 ENU16CH17a mice included flow cytometry of peripheral lymphoid cells (Figure 5.2 and Appendix E). All Phenotypic screens for *ENU16CH17a*, *222*, *NIH69b*, *NIH85a*, *NIH19a*, *APFNS1015_17* and *ENU22* were carried out by Professor Chris Goodnow's group at ANU Canberra.

Professor Warwick Britton's group at the University of Sydney screened *007* for mycobacterium tuberculosis.

2.3.2 Phenotyping of *nephertiti*

Proteinuria was detected by urine dipstick (Multistix, Bayer HealthCare) in third generation (G_3) offspring in the *nephertiti* pedigree.

All further *nephertiti* phenotyping was performed in the Cornall lab (section 4.2.1).

2.3.3 Collection of blood samples

Mice aged 17–25 weeks were terminally anaesthetised and blood samples were collected, by cardiac puncture, into lithium heparin paediatric tubes and centrifuged to separate out the plasma.

2.3.4 Routine clinical chemistry of plasma samples

Clinical chemistry was performed on a Beckman Coulter AU400 semi-automated clinical chemistry analyser by the Mary Lyon Centre's clinical pathology service laboratory at MRC Harwell. All assays were carried out using the manufacturer's instructions, parameter settings and reagents. Samples were analysed for total protein, albumin, total cholesterol, triglycerides urea and creatinine. Electrolytes (sodium, potassium and chloride), total calcium, inorganic phosphate, alanine aminotransferase, aspartate aminotransferase, alkaline phosphatase, HDL cholesterol, LDL cholesterol and glucose were also measured.

The author performed histology and immunohistochemistry and the methods are described below.

2.4 Conventional mapping of ENU mice

B6 *nephertiti* mice were out–crossed to the CBA/J strain for mapping and bred to bring the causative mutation to homozygosity, using dipstick urine testing to track the phenotype. Linkage mapping was performed using simple sequence length polymorphisms (SSLP) and single nucleotide polymorphisms (SNPs). The author calculated maximal logarithm of odds (LOD) scores (Morton 1955) using a range of recombination fractions less than 0.5 at each of 26 polymorphic loci, given the observed alleles in a mean of 27 affected and unaffected mice from the *nephertiti* pedigree (Figure 4.7).

Mapping locations for *222*, *NIH85a*, *NIH19a* and *007* were provided by Professor

Chris Goodnow's group ANU.

2.5 Extraction of mouse DNA

DNA was extracted from tail tissue using a DNAeasy kit (Qiagen).

2.6 Assessment of DNA quality for sequencing

For quantification and quality assessment prior to whole genome sequencing, DNA was analysed by a fluorometric method, in which DNA specific dyes emit on binding to their target molecule. This was either with Qubit (Invitrogen), or PicoGreen assay (Life technologies). For the Qubit assay, 1μl sample DNA was diluted in 199μl of Qubit working solution and processed alongside two standards. For the PicoGreen assay, serial dilutions of DNA in TE were added sequential wells of a micro plate with 200μg PicoGreen reagent, alongside dilutions of a standard with known concentration. Samples were incubated for 5 minutes and examined with a microplate reader, using excitation at 480 and emission reading at 520. The R statistical environment (http://www.r-project.org) was used to generate a linear regression analysis for the standard, with a coefficient of determination of 0.992. By fitting the sample data to the standard curve the sample concentrations can be calculated. Finally, to assess DNA quality, 1μl DNA for each sample was run on an agarose gel alongside a ladder.

2.7 Whole Genome Sequencing

222 was sequenced on 7 lanes of a GAII analyser (Illumina), all other mice were sequenced on a HiSeq 2000 (Illumina). The low coverage mice *NIH69b*, *NIH85a*, *NIH19a*, *11_007*, *12_007* (28 Gb per individual) and *nephertiti* (18 Gb) on one lane per mouse, the 3 low coverage mice to validate ENU mutation density (40Gb

in total), *APFNS1015_17* and *ENU22* using one lane per 3 barcoded individuals. *APFNS1015_17* and *ENU22* were sequenced at ANU and the raw output mapped, called, filtered and processed with the Lander–Green based algorithm by the author. For the *ENU16CH17a* pedigree 3 affected G3 siblings were sequenced using two lanes in an Illumina HiSeq machine (mean 78Gb) per mouse. 100 bp paired-end reads were generated for all mouse sequencing.

2.7.1 Mapping to the reference genome

Reads were mapped to the mouse reference MGSCv37 (mm9) using Stampy (Lunter and Goodson 2011) with BWA settings. *APFNS1015_17* and *ENU22* were mapped to the MGSCv38 / mm10 genome. For *ENU16CH17a* 94.5% of genome was covered at least once, the mean coverage across the genome was 24–fold per mouse.

Human WGS was performed according to WGS500 project protocols to a target depth of 25–fold (Figure 6.1), using 150bp paired end reads and mapping to the human genome GRCh37.

2.7.2 Coverage calculation for whole genome sequencing

Coverage was calculated across the genome from the BAM format mapped read files using BEDTools software (Quinlan and Hall 2010).

Distribution of coverage depth for individual sequenced datasets is shown in figure 3.1, figure 4.8, figure 5.6 and figure 6.1. Table 2.2 gives summary coverage data over the genome and the exome for the low coverage whole genome samples, and gives examples of genomic coverage for the full coverage human datasets. For the low coverage genomes, more of the exome than the genome had at least 1–fold coverage and in some cases 3–fold coverage was also higher, this may be due to the excess of repetitive sequence in non–coding regions, mean exonic coverage was similar to genomic coverage.

Sample	% Genome >= 1	% Genome >= 3	Mean genomic coverage	% Exome >= 1	% Exome >= 3	Mean exonic coverage
222	92.99	83.32	5.31	98.82	86.83	5.32
NIH85a	94.04	79.64	4.91	97.00	79.05	4.55
NIH19a	89.15	56.37	3.11	92.05	55.68	3.06
NIH19a	89.15	56.37	3.11	92.05	55.68	3.06
NIH69b	94.18	80.10	4.9	97.34	81.51	4.9
007_11	89.28	57.37	3.18	92.14	56.75	3.13
007_12	89.29	61.40	3.38	93.64	61.38	3.34
Nephertiti	92.75	84.36	5.79	98.78	91.99	6.51
ENU16-CH17a 1	93.87	93.32	23.08	99.36	99.03	26.68
SRNS 0001	90.31	90.24	33.07			
SLE 17709	91.76	91.65	52.82			

Table 2.2: Coverage summary, as percentage of genome or exome covered by at least 1 read, at least 3 reads (the minimum to call variants in single individuals) and mean coverage depth. The exome coverage was calculated based on the UCSC gene set for mm9.

2.7.3 Variant calling and annotation

For all whole genome sequencing in this thesis, an in-house variant caller Platypus was used (Rimmer et al. 2014). Variants were annotated using Annovar (Wang, Li, and Hakonarson 2010) with Ensembl (release 64) gene annotation.

2.7.4 Filtering WGS ENU variant calls

All scripting to parse variant calls was written by the author in python (www.python.org).

2.7.5 Generation of a union file

Variants previously observed in other ENU pedigrees were removed using a variant union file. To create this file of shared, and thus non–ENU, variation, Platypus was used to call variants from mice from 9 different ENU pedigrees simultaneously. Thus at each variant locus a genotype and genotype likelihood was assigned for all mice.

All data in the union file was generated using the sequencing, mapping and variant calling pipeline described in the Methods and Chapter 3.

2.7.6 Filters

Filters as described in Chapter 3 were used for all mouse variant calls.

In order to identify the number of variants removed by each filter individually a python script was implemented that ran all the filters individually or together. The script counts the number of variants that 'fail' each unique combination of filters to examine the overlap between filters and writes out a variant call format (VCF) file for each mouse consisting of variants not excluded by any filter.

In the three pedigrees analysed using the Lander–Green based algorithm: *ENU16Ch17a*, *APFNS1015_17* and *ENU22*, variants with a high local density of bad reads were removed using a flag generated by the variant caller. After all other filters, remaining calls that were clustered closely together with a threshold of less than 1,000 bp were removed, this later was because multiple variant calls in close proximity often represent regions of error and cause the Lander–Green based algorithm to overestimate the likelihood of a recombination event within a genomic window.

PCR duplicates Duplicates were removed from the data using Picard (http://picard.sourceforge.net/), which incorporates data from the 5' ends of both reads in mapped paired data.

Annotation The ENU variants which passed filtering were annotated using annovar (Wang, Li, and Hakonarson 2010) with Ensembl (release 64) gene annotation. Functional predictions were made with PolyPhen–2 (Adzhubei et al. 2010) using probabilistic classifications based on a model trained with the HumVar dataset, tailored for detection of Mendelian disease caused by mutations with large effects. Although the training dataset consists of human disease causing mutations, the modelling is based on sequence and structural features applicable across species (Adzhubei et al. 2010) and higher PolyPhen–2 scoring has been shown to correlate with damaging murine

ENU mutations (Andrews et al. 2012). The Lyn^{T410A} mutation in *ENU16CH17a* was confirmed independently using exome sequencing (Andrews et al. 2012). Throughout this thesis, both for murine and human variants, a 'benign' PolyPhen–2 score is defined as a probabilistic score of 0.15 or below, 'possibly damaging' as above 0.15 but below 0.85 and 'probably damaging' 0.85 or above. This is the qualitative classification used in PolyPhen–2 (Adzhubei et al. 2010) and must be interpreted in the context of other information about an individual variant or gene.

2.8 The effects of laboratory and strain on shared variation

Platypus was used to call variants from mice from 9 different ENU pedigrees simultaneously. Thus at each variant locus a genotype and genotype likelihood was assigned for all mice. The nine pedigrees included 6 from the Australian Phenomics Facility at the Australian National University (ANU), 2 from the MRC Harwell Centre for Mouse Genetics and one from the Beutler Group at the Scripps Research Institute. The Harwell pedigrees and one of the ANU mice were on a mixed strain background. All other mice were on a straight B6 background. To examine the proportion of this shared variation attributable to B6 reference strain mice, we excluded shared variants exclusively observed in mixed strain mice. The resultant shared variants are observed in at least 2 pedigrees, including at least one fully B6 pedigree. Since all pedigrees included from MRC Harwell are mixed strain, the MRC Harwell variants fully overlap the other laboratories.

2.9 Protein kinase C alpha experiments

2.9.1 Cloning of GFP labelled protein kinase C alpha

A mouse protein kinase C α (*Prkca*) clone was obtained from Source Bioscience (clone I920160M19). Site directed mutagenesis using PfuUltra HF DNA polymerase (Agilent) was used to with primer sequences:

sense: 5'-agaacccttcaaaatcagattggtctgtgttagcaatgacca-3'
antisense: 5'-agaacccttcaaaatcagattggtctgtgttagcaatgacca-3'

to recapitulate the I648T mutation. Then primers were designed to add restriction sites EcoRI and XhoI into WT and mutant Prkca.

Forward: 5' XhoI ccg ctc gag acc atg gct gac gtt tac ccg
Reverse: EcoRI ccg g aat tcg tac tgc act ttg caa gat tgg

The mutated clone DNA with restriction sites, and a pEGFP–N1 GFP fusion vector were double digested, products cut out on a gel and extracted (QIAquick gel extraction kit). Sanger sequencing confirmed the mutagenesis and absence of other mutations. The vector and insert ligation products were then transformed into DH5α cells and selected with kanamycin.

BE(2)–M17 cells were cultured in EMEM/ HAMS medium and transfected with WT or Mutant Prkca using polyethylenimine (PEI).

2.9.2 Flow cytometry for GFP to check transfection

48 hours post transfection, media was removed, cells washed with phosphate buffered saline, trypsinized and fresh media added. Cells were spun, supernatant removed and

FACS staining media added.

Flow cytometry was performed on a BD FACSCanto machine (BDBiosciences) with excitation at 488nm and analysed with FlowJo software.

2.9.3 Confocal Microscopy

Live cells were plated onto glass bottom petri dishes (5 x 10^4 cells/ dish), washed twice under with HANKS/HEPES buffer and examined in an incubator chamber under a Zeiss 510 MetaHead confocal microscope with excitation wavelength 488. Time-lapse images were taken before and after stimulation with GF109203X 2mM followed by DOG 400mM. Time to translocation to the cell membrane was estimated visually with blinding for WT or mutant sample type.

2.10 A hidden Markov method for mixed strain ENU

To shortlist candidate causative mutations in *nephertiti* without recourse to conventional linkage mapping it is necessary to distinguish the ENU induced homozygous mutations from the large amount of homozygous variation from the reference genome in genomic regions inherited from the CBA/J ancestor. This is achieved by analysing density of variation. Plotting the variants by genomic location across each chromosome demonstrates the densely clustered variation in CBA/J genomic intervals, and less dense variation in intervals inherited from the ENU treated B6 founder (Figure 4.9a). In order to more precisely define these intervals an algorithm was developed based on a hidden Markov model (HMM).

The standard HMM machinery is well documented (Rabiner 1989). A HMM has two main components which are model–specific: these are the state transition matrix, which specifies the probabilities of transitions between any two model states, and the

likelihood, which is the probability of the observed data given a particular model state.

The model can discriminate between genomic intervals inherited from the different founders based on differences in variant density.

Initial state probabilities are based on the expected frequencies of the 6 possible genotypes, ENU/ENU (homozygous for an ENU treated founder), CBA/CBA (homozygous for the outcross strain), WT/WT (homozygous for the B6 reference strain), ENU/WT, ENU/CBA and CBA/WT, given knowledge of the pedigree.

At each variant locus a probability for each possible state was calculated based on the mean distances between the nearest two adjacent variants. A Poisson distribution of actual distances between variants around a mean variant density in each state was assumed. Mean variant densities were estimated based on the ENU, WT and CBA variant frequencies. The ENU density used was 1.5 SNPs/Mb (section 5.5) (Bull et al. 2013). A WT background variant density of 0.2 SNPs/Mb and CBA/J to reference strain variant density of 200 SNPs/Mb were empirically derived. The estimate of the average variant density in each state, y, is specific to the known zygosity of the variant (homozygous or heterozygous). A Poisson probability density around y predicts the probability of the observed distance d given an underlying genotype or state s, for each state ($Pr(d|s)$),

$$Pr(d|s) = \frac{y^d e^{-y}}{d!}\,. \tag{2.1}$$

Figure 2.1 shows histograms of the distances between variants in *nephertiti* from all ENU/ENU homozygous or WT/WT homozygous regions, and indicates a positive skew to the actual distribution.

Bayes formula is applied to give probability of a state s given the distance d,

$$Pr(s|d) = \frac{Pr(d|s)Pr(s)}{Pr(d)}\,. \tag{2.2}$$

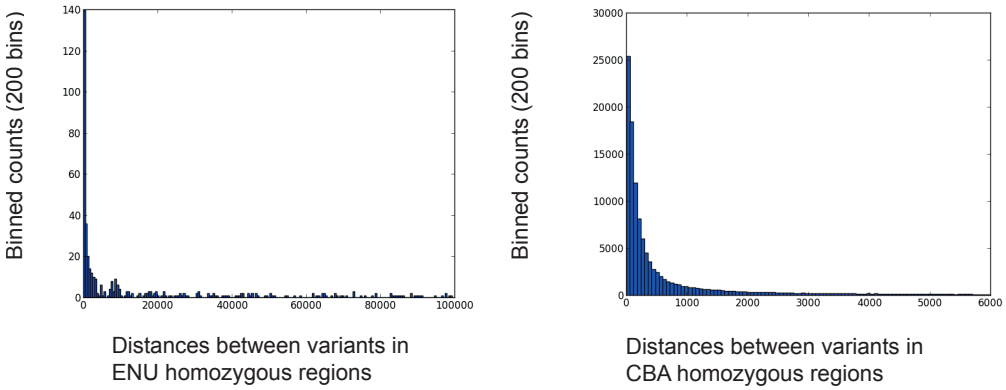


Figure 2.1: Histograms of distances between variants in nephertiti, in base pairs, for homozygous ENU and CBA (outcross strain) regions, assigned by the algorithm. Distances binned using 200 bins.

$Pr(d)$ is approximated by calculating the Poisson distribution probability of d given the overall mean distance between variants with this zygosity.

The transition probability is the probability of a recombination event leading to a different set of ancestral haplotypes occurring between two variants, and can be considered as a function of the recombination frequency, the distance between variants and the number of allelic changes. A 6 x 6 transition matrix incorporating the probability for all possible state transitions is generated for each variant. d is observed mean distance to adjacent variants and the recombination frequency is assumed to be the average mouse recombination frequency of 0.56cM/Mb per meiosis (Jensen-Seaman et al. 2004). Hence,

probability one allelic recombination = h ,

probability 2 allelic recombinations= h^2 ,

where h = recombination frequency $*d$.

A Viterbi algorithm is used to infer the most likely sequence of ancestral haplotypes across each chromosome (Figure 4.9a).

2.11 An identity by descent method using an implementation of the Lander–Green algorithm

The Lander–Green (Lander and Green 1987) based algorithm developed for this project uses a combination of genotype information from informative markers (here, the SNP genotypes called from the next-generation sequencing data), and knowledge of local recombination rates to determine the ancestral haplotypes in specific genomic intervals, and the locations at which recombinations and therefore transitions between inheritance state vectors occur.

The pedigree can be considered to originate with the G_1 pair of common ancestors who each carry a combination of the haplotypes inherited from the G_0 mice, these being ENU1, ENU2, WT1 and WT2 (Figure 5.1a). Thus for the purpose of the algorithm the pedigree consists of 5 non–originals – that is 5 individual mice with parents in the pedigree.

The Lander–Green algorithm represents the ancestral haplotypes of the 5 non–originals as a state vector with 10 binary coordinates, representing the 5 individual mice, arising from 10 gametes. Within each inheritance vector a 0 coordinate indicates a gamete carrying grand–paternal DNA at a locus, and 1 indicates grand–maternal inheritance. These are arbitrarily phased. There are 2^{10} possible state vectors.

The Lander–Green algorithm uses a state transition matrix based on the recombination rate, which encodes the probabilities of transitions between any of the ancestral state vectors, based on the number of recombinations required for the transition. In the implementation a recombination map (Cox et al. 2009) is used to compute average local recombination rates across the genome.

The whole transition matrix is not stored in memory, but matrix elements are computed on demand. This is straightforward, as all matrix entries can be expressed as powers of the recombination rate and one minus the recombination rate. This

vastly reduces the memory requirements for the algorithm, which are now linear in the number of state vectors rather than quadratic.

For each G_3, a state vector determines which two ancestral haplotypes that make up G_1, make up the local genotype, e.g. (ENU1, WT2). A probability is computed for the observed SNP genotypes, given each ancestral state vector, in 100kb windows across each chromosome. This probability has two components: a prior probability of observing a SNP in the given window, and genotype likelihoods computed by the variant caller from the sequence data. Fixed priors of observing a SNP in ENU haplotypes (2 Mb^{-1}) and WT haplotypes (0.2 Mb^{-1}) were assumed. The IBD regions inferred by the algorithm were relatively insensitive to changes in the ENU mutation density prior (Figure 5.3).

The likelihood for a particular state vector is the sum over all possible combinations of the SNP genotypes (0/0, 1/0, 1/1) for the 3 mice, of the product of the SNP priors and the relevant genotype likelihoods for the 3 G_3 mice. This incorporates the dependent relationships between the mice. In the case of multiple SNPs occurring in the same window, SNPs are assumed to be independent, and the likelihood for all mice in the window is the product of the likelihoods across all SNPs. Using the genotype likelihoods from the caller allows accommodation for errors in the WGS data; a modification to the conventional Lander-Green algorithm that has been used to infer IBD in array data (Markus, Birk, and Geiger 2011). Due to the paucity of polymorphic sites in the in-bred B6 mouse, there are many 100 kb windows that contain no SNPs. If no SNPs were called in a window, the most likely explanation is that no SNPs were present. In a small fraction of cases, a real SNP will be missed due to low coverage or variant-calling errors. To deal with windows that contain no SNPs, a set of likelihoods is supplied weighted towards 0/0. Specifically this assumes a 1/10 probability that a heterozygous SNP was missed, and a 1/100 probability that a homozygous SNP was missed.

Finally, the forward backward algorithm (Rabiner 1989) is used to compute the posterior probability of each state vector at each window. The state vector with the highest posterior in each window is selected to construct the sequence of most probable inheritance states across each chromosome. Within a three–generation pedigree, frequent recombinations, particularly from state A to state B and back to state A within a small genomic interval are likely to be artifactual. Therefore a smoothing step is performed on the output inheritance states, such that recombinations from state A to state B and back to state A within 1 Mb, within an allele, are corrected to state A.

In reporting the results we define IBD heterozygous throughout to refer to regions or variant sets in which all sequenced individuals share at least one allele from the same ENU founder but are not IBD homozygous.

Graphical plots of genotypes were generated using matplotlib (http://matplotlib.sourceforge.net).

The results are shown in Figure 5.4. The program was coded in Python (http://www.python.org) and Cython (http://cython.org), in collaboration with Andy Rimmer and Gerton Lunter.
The code is freely downloadable from http://www.well.ox.ac.uk/lgenu.

2.11.1 Density and Characteristics of ENU Mutations

The ENU16Ch17a dataset includes on average 647 mutations per mouse in homozygous regions spanning over 1,000 Mb across 3 mice (mean 405 Mb, 15.3% of the genome, per mouse). Sanger sequencing of the candidate mutations confirms the reliability of our filtered call set (Appendix A and Appendix B).

Observing the density of these mutations and subtracting the low background density from homozygous WT regions allowed estimation of the ENU mutation density. The homozygous ENU region variants were used to examine the base prefer-

ence of ENU mutations. Comparison of the transition transversion ratio was made with all the variants in the Centre for Genome Dynamics Mouse SNP Database (http://cgd.jax.org/cgdsnpdb) on 8.5.12. This database includes over 66 million SNPs from 136 inbred laboratory mouse strains predominantly imputed from the mouse diversity array (Wang et al. 2012), and is representative of the characteristics of naturally occurring (non–ENU) SNPs in an inbred mouse.

The code to handle the output from the Lander-Green based script is in Appendix J. This performs smoothing, calculates ENU mutation rate and base characteristics, and plots genotypes.

2.11.2 Simulating lower coverage depths in an empirical dataset

Random subsets of reads were generated from the *ENU16CH17a* bam file using a script that utilizes pysam, a SAMtools (Li et al. 2009b) interface for Python (http://code.google.com/p/pysam/). Variant calling and filtering on the down–sampled bam files used the same pipeline described above. Comparisons of IBD regions and variant calls between the lower coverage datasets and the full (24–fold) coverage data were made using BEDTools (Quinlan and Hall 2010) to measure intersections. The intersections of IBD regions were analysed by a per base pair (bp) comparison.

To explore the effect of possible errors in the assignment of genomic intervals to the estimation of the ENU mutation density, the pipeline to extract mutation frequencies from the Lander–Green algorithm output regions and the filtered variant call files was modified to model expansion or contraction of the ENU homozygous regions by adjusting the start and end positions of the Lander–Green output regions and recalculating the ENU mutation density based on these inputs.

2.11.3 Comparison with non–IBD approach to detect shared variation

To examine whether an IBD method performs better than a simple per SNP approach to detect shared ENU variation in the 3 *ENU16CH17a* mice, a comparison set of shared variants was generated at each simulated coverage depth by simply selecting variants that were observed in all 3 mice. In exactly the same way as with the IBD variants we included SNPs in which there was at least one homozygous or heterozygous mouse and the remaining 0, 1 or 2 mice had no genotype information (denoted ./. in the VCF file), or a reference (0/0) genotype call with at least one variant call and less than 5 supporting reference reads.

2.11.4 Calculating the proportion of mutations affecting protein sense

The distribution between missense and splice mutations was examined in the larger dataset of heterozygous mutations in *ENU16CH17a*. Across the 3 mice, 86% (21.7/25) of potentially damaging mutations were missense mutations, 6% (1.3/25) were nonsense mutations and 8% (2/25) were in splice sites. A large database (http://mutagenetix.utsouthwestern.edu/ accessed on 8.5.12) of over 5,000 incidental ENU mutations with no observed deleterious effects, identified in the course of next generation sequencing (Applied Biosystems SOLiD), reports 85.9% missense, 4.4% nonsense and 9.7% splicing mutations, and agrees broadly with the findings in *ENU16CH17a*.

2.12 Modelling expected numbers of ENU mutations in G_3 mice

To model the segregation of mutations within a pedigree, the probability of a mutation being inherited by a G_3 was calculated under the three possible situations where the G_2 parents together carry 0, 1 or 2 copies of the mutation (in 25%, 50% and 25% cases respectively; individual G_2 mice are not homozygous for any ENU mutation). We denote the chance of inheriting the mutation at G_3 under each of these situations with a given zygosity (homozygous or heterozygous) at G_3 as P_0, P_1 and P_2. Clearly some of these probabilities will be 0.

Conditional on the number of mutations carried by the G_2 parents, the number n of G_3 offspring inheriting the mutation with the required zygosity can be modeled using a binomial distribution. For a given G_2 parent pair, this distribution is denoted by $f(n)$. Assuming that each G_2 pair produces 3 litters of 4 live mice, this distribution is given by

$$f(n) = 0.25\binom{12}{n}P_0^n(1-P_0)^{12-n} + 0.5\binom{12}{n}P_1^n(1-P_1)^{12-n} + 0.25\binom{12}{n}P_2^n(1-P_2)^{12-n}\,. \tag{2.3}$$

Here $\binom{12}{n}$ or '12 choose n' denotes the binomial co-efficient indexed by 12 and n. To estimate the probability of M G_3 carrying the mutation across the 48 G_3 from 4 G_2 pairs, convolution was performed across all combinations of mice that together transmit precisely M mutations with the required zygosity from G_2 pairs, such that

$$Prob(m = M) = \sum_{i+j+k+l=M} f(i)f(j)f(k)f(l)\,. \tag{2.4}$$

Two situations were considered: one of a recessive mutation, in which a G_3 has two alleles from parents that are heterozygous for the mutant; and the situation of a dominant mutation, where homozygotes may also have the phenotype and may be indistinguishable from heterozygotes. In this way it is possible to calculate the probability of any recessive or dominant mutation carried by a founder occurring M times in the G_3 mice. The results are presented in Figure 5.12a.

Modelling Numbers of Shared IBD SNPs in Multiple Sequenced Mice

The proportion of the genome expected to be IBD in q sequenced mice was modelled without accounting for linkage to the causative mutation. The probability of M G_3 carrying a shared ancestral haplotype at any locus was calculated as described above. $Prob(m = M)$, for each M between q and 48 (the modelled number of G_3 mice), and the probability was calculated of picking q mice sharing such a locus by chance from a pool of M mice sharing the locus and $48 - M$ individuals not carrying the locus. Since each mouse can only be picked once this corresponds to a hypergeometric distribution. By summing over the product of this and the $Prob(m = M)$, for each M between q and 48, the overall probability, R, of any unlinked locus being observed in all the affected sequenced mice, can be obtained,

$$R = \sum_{M=q}^{48} Prob(m = M) \frac{\binom{M}{q}\binom{48-M}{q-q}}{\binom{48}{q}} = \sum_{M=q}^{48} Prob(m = M) \frac{\binom{M}{q}}{\binom{48}{q}} . \tag{2.5}$$

R is the proportion of the genome expected to be IBD for a specified number q of

sequenced mice. Knowledge of the ENU mutation density (1.5 mutations Mb^{-1}), and the fraction of variation affecting protein sense (1.05% missense, nonsense or splicing in *ENU16CH17a*), was used to estimate the number of homozygous or heterozygous candidate mutations shared by q affected sequenced mice,

$$number\,of\,shared\,candidates = R * 1.54 * 1.05\,.$$

One mutation is added to model the causative mutation which is always present. In phenotypically affected mice a region from the ENU founder persists around the causative mutation due to linkage, and this adds another fraction $(\frac{1}{c})(\frac{1}{m})^k$ of the genome, where c is the fractional size of the chromosome, m is the number of meioses per G_3 mouse, and k is the number of G_3 mice. This approximates to a further 0.7 mutations or $7 * 10^{-3}$ candidate coding mutations. Since this is negligible we simply approximate to 1 additional mutation. The results are presented in Figure 5.12b.

2.13 Gene saturation modelling

In order to model the chance of an ENU mutation occurring in any gene, it was necessary first to know the sizes of genes. Start and end co-ordinates for all exons in all RefSeq genes were downloaded from UCSC (http://genome.ucsc.edu) in 0 based BED format. A script was written to extract the FASTA format genomic sequence for these regions from the reference genome using pysam.

A single random mutation was then simulated in the sequence and the consequence of this mutation checked by submitting the mutated sequence to Annovar (Wang, Li, and Hakonarson 2010). This process of random mutation was repeated 10 times per exon, and the proportion of mutations causing non–synonymous, non–sense or splicing changes for each gene, plus the gene length, written out to file.

This file was used to calculate expected numbers of genes with mutations and

their types for increasing numbers of pedigrees as described in section 5.8.2.

2.14 Sanger Sequencing

Pre–Sanger Sequencing PCR for Genomic DNA All H_20 used was DNA and RNA free.

Primers were designed using Primer3 Plus (Untergasser et al. 2007) or manually, checking GC content, secondary structure, annealing temps, and blasting against the relevant genome for non-specific binding.

Dry primers were spun down in a microcentrifuge and re-suspended in appropriate volumes of TE (Tris EDTA) to 100μM stock, stored at -20°C. 1mM primer for Sanger sequencing use was made with a 1:100 dilution in H_20 and stored at -4°C.

First round of PCR First round PCR reactions were prepared as shown in table 2.3.

Reagent	Quantity
10x NH_4 Reaction Buffer	2.5μl
$MgCl_2$ solution	1.25μl
dNTP mix (at 10mM each dNTP)	0.625μl
Forward Primer (1μM)	7.5μl
Reverse Primer (1μM)	7.5μl
Biotaq (Taq polymerase) (Bioline)	0.125μl
template genomic DNA	50 –100ng
H_20	Make up to 25μl

Table 2.3: First round PCR. Quantities shown for a 25μl reaction (in PCR tubes or plate/ strip).

PCR cycles See table 2.4. Annealing temperatures were adjusted depending on the primer annealing temps.

DNA clean up Using 96 well filter plate (Millipore Multiscreen filter plates).

Step	Temperature °C	Time (seconds)
1	95	30
2	95	10
3	67	30
	Repeat 'touchdown' steps 2–3 14 times, reducing step 3 by 0.5°C per cycle	
4	72	30
5 (deanturing)	95	30
6 (annealing)	60	30
	Repeat steps 5–7 25 times	
7 (extension)	72	30
8	72	5 minutes
9	4	forever

Table 2.4: PCR thermo-cycling, first round.

25μl TE was added to each PCR product, then products were transferred to wells of the filter plate, and suctioned to dryness using a vacuum manifold. 30μl H_20 was added to the wells and then removed from the filter plate to a PCR plate or PCR tubes.

Reagent	Quantity
Template DNA (PCR product above)	1-5μl
BigDye (Applied Biosystems)	2μl
5 x BigDye sequencing buffer	4μl
Forward primer 1μM	3.2μl
H_20	make up to 20μl

Table 2.5: BigDye PCR

***BigDye* PCR**

PCR cycling - second round

Step	Temperature °C	Time (seconds)
1	96	60
2	96	10
3	50	15
4	60	4 minutes
	Repeat steps 2-4 24 times	
5	4	forever

Table 2.6: PCR thermo-cycling, second round

Ethanol DNA clean up 2μl NaAc (Sodium Acetate) and 50μl 100% ethanol were added to each well and covered with film or lids. After inverting x 4 to mix the wells were left at room temperature for a maximum of 15 mins. Next they were centrifuged at 3000g for 30 min at 4°C. The film / lids were removed immediately and the plate / tubes were inverted onto paper and centrifuged at 185g for 10-15 seconds. 70μl of 70% ethanol was added to each well, these were covered and centrifuged at 1650g for 15 min at 4°C. The covers were removed again before a final inverted spin at 185g for 1 minute. Dry wells were covered and sent for Sanger sequencing on an Applied Biosystems 3720xl machine.

2.14.1 *ENU16CH17a* Sanger sequencing

In *ENU16CH17a* the 28 candidate mutations were amplified with two rounds of PCR from genomic DNA using internal and external fully nested primers (Appendix C) and then amplified with BigDye (Applied Biosystems Ltd) before sequencing. All nested sequencing reactions were run in duplicate to check for PCR error.

2.14.2 Validation of the *Lamb2* mutation with Sanger sequencing

primers to check *Lamb2*

Forward: CTATGCTGGTGGAGCGTTCT

Reverse: TGAGTAGCGGGACTCACACA

2.14.3 Sanger sequencing of variants in the 17709 family

Primers were designed to check the WAS^{C538A} variant as follows:

Forward: GTGGCAGGGCTGTGATAACT

Reverse: GCTCGTCCATCCACATACCT

Primers were designed to check the $ANKRD45^{C385G}$ variant as follows:

Forward: TAAGTTCTATGCGCCCGAAG

Reverse: CTTGGTGTCGTCCTTGAAGC

2.15 Statistical analyses

Analysis of means, confidence intervals, standard deviations and p values for unpaired two–tailed t tests were performed using the Graphpad Prism 5 package.

Standard deviations are given, and unpaired two–tailed t tests used to compare biochemical observations made in wild type and mutant *nephertiti* mice (Figure 4.2 and Figure 4.3), assuming a Gaussian distribution for the data.

Mean values without standard deviation were given for data from the three sibling *ENU16Ch17a* mutants in Chapter 5.

95% confidence intervals are shown for ENU mutation frequencies calculated in 9 mutants from 3 pedigrees (Figure 5.7).

All other analyses were written in custom scripts and described in the Methods where used.

2.16 Histology

2.16.1 Tissue fixation

Formalin fixation of kidney sections. Mice were terminated humanely using a method approved under schedule 1 to the Animals (Scientific Procedures) Act 1986. Animals were dissected and both kidneys were removed and halved in sagittal plane. Organs were immediately placed in 10% neutral buffered formalin (Sigma Aldrich). After 24 hours the organs were removed to 70% ethanol in histology cassettes. Tissue was processed in 2 baths of 70% ethanol, 2 baths of 90% ethanol, 3 baths of 100% ethanol and 3 baths of melted wax using an automated tissue processor. Organs were then embedded in paraffin, cut into 3 - 4 μm sections by microtome and transferred to glass slides.

Fixation of frozen kidney tissue. Murine kidneys obtained as described above were snap frozen in liquid nitrogen and stored at $-80°$C. Organs were then embedded in optimal cutting temperature compound (Tissue–Tek) on dry ice. 3–4 μm sections were cut using a cryotome and fixed in acetone for 10 minutes.

2.16.2 Deparaffinisation and hydration of formalin fixed sections

Sections were incubated in $3 * 5$ minute washes of xylene (Sigma Aldrich) or histoclear (Fisher Scientific) in a fume cupboard, followed by $2 * 10$ minute incubations in 100% ethanol and $2 * 10$ minute incubations in 95% ethanol. Sections were then washed twice in distilled water for 5 minutes each.

For transmission electron microscopy tissue was fixed with gluteraldehyde, embedded with resin and sectioned at 70nm.

The author performed haematoxylin and eosin, periodic acid Schiff and methenamine

silver staining and all immunofluorescence stains. The Oxford Centre for Histopathology Research performed electron microscopy and silver staining.

2.16.3 Haematoxylin and eosin staining

Deparaffinisation and hydration as described above. Slides were immersed in neat haematoxylin for 10 seconds and rinsed in distilled water for 5 minutes. After a quick dip in alcohol–acid (5ml of concentrated hydrochloric acid added to 0.5L of 70% Ethanol) slides were rinsed again in distilled water and then immersed for 5 minutes each in increasing concentrations of ethanol (50%, 70% and 95%) followed by 1 minute in 0.5% eosin 95% ethanol. The samples were then rinsed quickly in 2 changes of 100% ethanol and transferred to xylene until mounting in DPX.

2.16.4 Periodic acid Schiff staining

Deparaffinisation and hydration as described above. Sections were placed in 0.5% Periodic Acid solution (Thermo Scientific) for 5 minutes at room temperature and then rinsed in several changes of distilled water. Sections were stained in Schiff reagent (Thermo Scientific) for 15 minutes to achieve desired contrast and rinsed in running luke–warm tap water for 1 minute. Haematoxylin I (Thermo Scientific) staining was performed for 1 minute followed by 30 seconds rinsing in distilled water. Sections were immersed in bluing reagent for 1 minute (Scott's tap water) and rinsed again in distilled water for 30 seconds. Dehydration in 2 changes of 100% ethanol for 1 minute each and clearing in 3 changes of xylene or histoclear for 1 minute each, followed by mounting with DPX mounting medium (Sigma) and No. 1.5 cover slip.

2.16.5 Methanamine silver stain

Deparaffinisation and hydration as described above. Sections were oxidised in periodic acid solution (Thermo Scientific) for 5 minutes to oxidise polysaccharides to form aldehydes and rinsed in 5 changes of distilled water. Sections were placed in freshly prepared methenamine silver solution (methenamine–borax one capsule – Thermo Scientific), distilled water 50mls silver nitrate solution 1ml (Thermo Scientific)), and warmed in a 58°C water bath. After 30 minutes sections were rinsed and checked under a microscope. Slides were returned to the methenamine silver solution for 5–minute intervals as necessary until the basement membranes were black and clearly defined. The sections were then rinsed in 4 changes of distilled water and toned by immersion for 30 seconds in gold chloride solution (Thermo Scientific). After rinsing in distilled water slides were incubated in sodium thiosulphate solution for 1 minute to remove unwanted silver, rinsed again in tap water and stained with Fast Green stain solution for 30 seconds to enhance contrast. Finally sections were dehydrated and mounted with DPX as described above.

2.16.6 Antigen retrieval methods

All antigen retrieval methods for formalin fixed paraffin embedded sections were performed on sections after deparaffinisation and hydration as described above.

Sections were fast rinsed with 20 dips in distilled water and washed in Tris Buffered Saline (TBS) at 37°C for 30 minutes. Samples were then subjected to one of the antigen retrieval methods described below.

Heat mediated retrieval Sections were placed in a heat proof dish containing either Sodium Citrate buffer or or Tris/EDTA pH 9, in a water bath pre–heated to 100°C for 20 minutes. Slides were removed and run under cold tap water for 10 minutes.

Enzymatic digestion Pepsin: Sections were incubated with RTU pepsin (Sigma Aldrich) at 37°C for 20 min (or other durations as described in the results) and rinsed in phosphate buffered saline (PBS) for 2 x 10 minutes (Ekblom et al. 1982).

Hyaluronidase: Sections were incubated with 2 mg/ml in PBS, pH 5 for 60 min at 37°C) and rinsed in PBS for 2 x 10 minutes (Zenker et al. 2004).

Pronase: Sections were washed in Tris buffered saline (TBS) at 37°C for 30 minutes. Sections were then incubated with pronase (Protease type XIV from Streptomyces griseus, Sigma St Louis, MO, USA Prod. No. P5147) 75mg/pronase per 100ml of TBS at 37°C for 60 min. Washing in TBS at 4°C for 40 minutes terminated the digestion (Nasr et al. 2006).

2.16.7 Immunofluorescence histochemistry

IgG immunofluorescence IgG immunofluorescence was performed on formalin fixed tissues. Following pronase antigen retrieval as described above slides were rinsed in phosphate buffered saline for 10 minutes. The slides were incubated in a humidified chamber at room temperature for 20 minutes with a 1/50 concentration of fluorescein isothiocyanate (FITC) conjugated conjugated goat anti–mouse antibody directed against IgG (CALTAG M30201) (Vinuesa et al. 2005). Slides were rinsed for $2 * 10$ minutes in phosphate buffered saline at 4°C. Counterstaining was performed with 4',6–diamidino–2–phenylindole (DAPI) at 300nM for 3 minutes followed by 3 rinses in TBS. Finally sections were mounted with Vectashield (Vectorlabs) and stored in the dark.

Laminin β2 Immunofluorescence staining Fixed frozen sections, or formalin fixed sections subjected to antigen retrieval (with heat based methods, pronase, pepsin or hyaluronidase) were washed with $2 * 5$ mins TBS plus 0.025% Triton X–100 with gentle agitation and blocked with 10% normal goat serum (Sigma Aldrich) with 1%

bovine serum albumin in TBS for 2 hours at room temperature.

Drained slides were incubated in a humidified chamber with primary antibody overnight at 4°C: either monoclonal rat anti mouse IgG to laminin $\beta 2$ (Millipore Anti Laminin B2 antibody clone A5 05-206) (Ishiyama et al. 2009) at 25 mcg/ml, or polyclonal rabbit anti mouse IgG to laminin $\beta 2$ at 1:50 or 1:200 concentration of 1mg/ml antibody (Novus biologicals polyclonal anti laminin NBP1-00904.

On day 2 slides were washed with $2 * 5$ mins TBS plus 0.025% Triton X–100 with gentle agitation and then secondary antibody was applied. Secondary antibody for the monoclonal primary was either Fluorescein isothiocyanate (FITC) conjugated goat anti-rat IgG (Abcam ab6840) at 1:200 concentration diluted in TBS with 1% bovine serum albumin or Alexa Fluor 633 Goat anti Rat IgG (Life Technologies A21094) at 1:200 concentration. For the polyclonal primary the secondary antibody was Alexa Fluor 633 conjugated goat anti-rabbit IgG (Life Technologies A21070) at 1:200 dilution. Sections were incubated in the dark for 2 hours with the appropriate secondary antibody. After washing $3 * 5$ minutes in TBS counterstaining was performed with 4',6–diamidino–2–phenylindole (DAPI) at 300nM for 3 minutes followed by 3 rinses in TBS. Finally sections were mounted with Vectashield (Vectorlabs) and stored in the dark.

Immunofluorescence images were captured using a Zeiss 510 metahead confocal microscope.

2.17 Whole genome sequencing and pipeline for human DNA

All human DNA used was sequenced as part of the WGS500 project at the Wellcome Trust Centre for Human Genetics. DNA from patients with early onset steroid resistant nephrotic syndrome (SRNS) was obtained from Professor Moin Saleem and

Dr Hugh MacCarthy in Bristol and had been collected as part of the UK registry for Rare Diseases (RaDaR) with prior ethics approval and patient consent for next generation sequencing techniques for the purpose of identifying genes implicated in their renal disease. DNA for patients with early onset Systemic Lupus Erythematosis was obtained from Professor Tim Vyse at King's College London and Professor Earl Silverman at the University of Toronto and Toronto Hospital for Sick Children.

As part of the WGS500 programme, sequencing was conducted on an Illumina HiSeq 2000 machine with bioinformatics analysis using a bespoke pipeline standardised across the project to detect, genotype and deeply annotate variants, identify broad scale copy number variation (CNV) and look for homozygosity. Mapping was performed using Stampy (Lunter and Goodson 2011)and variants were called using Platypus (Rimmer et al. 2014) including both individual calls and genotyping of all individuals across all variant loci identified in any sample within the project, the latter generating a union file of variants seen in one or more WGS500 individual. Annotation within the VCF files used an implementation of Annovar (Wang, Li, and Hakonarson 2010) to provide functional category information and frequency of the variant in the 1000 genomes cohort. Annotation included predicted coding consequence and effect scoring from in-silico functional prediction tools including PolyPhen–2 (Adzhubei et al. 2010), SIFT (Kumar, Henikoff, and Ng, P. C. 2009) and MutationTaster (Schwarz et al. 2010) plus UCSC 46 species conservation and segmental duplication scores (https://genome.ucsc.edu), as well as phyloP single-site conservation scores (Pollard et al. 2010). The reads were scanned for large scale CNV, and the calls for homozygosity. All the above was generated in an automated pipeline developed by members of the bioinformatics core at the Wellcome Trust Centre for Human Genetics. The author carried out all analysis of the SRNS and SLE patients downstream of this pipeline.

2.17.1 Additional filters for human sequence variants

Variant calls output from the WGS500 pipeline were filtered further to reduce the number of candidate variants. Variants were filtered for allele bias, such that homozygous variants with alternate allele frequency equal to or below 30% of reads were excluded, and heterozygous variants with less than 10% of reads carrying the variant allele were excluded. Variants seen as high frequency heterozygotes in the WGS500 union file were excluded (greater than or equal to 5%), as were variants not predicted to be deleterious by PolyPhen–2, specifically those with a score less than or equal to 0.15 (Adzhubei et al. 2010).

Homozygous and heterozygous candidate variants For homozygous variants, variants with a population frequency below 0.05 were considered, based on 1000 genomes minor allele frequency (MAF), but excluded if they were present as homozygotes in the WGS500 union file, 1000 genome database or other database of known population variation.

Additionally homozygous variants observed in more than one of the SRNS or SLE patients were excluded on the basis that a shared homozygous causative variant between these unrelated patients was unlikely, and in SRNS patient 0001 the homozygous variants were checked against a database of 108 Punjabi genomes from Lahore in Pakistani obtained from the 1000 genomes project (1000 genomes population code PJL

http://www.1000genomes.org/category/frequently-asked-questions/population).

Scripts were written to make and submit batch files of regions containing candidate variant loci in patient 0001, from all the 1000 genome Punjabi BAM files, for simultaneous variant calling across all Punjabi 1000 genome individuals at these loci.

In the SLE study variants in the two siblings, patients 26106 and 39124 were included only if called in both patients.

Heterozygous variants were filtered as above but with a 1000 genomes MAF of less than 0.01 and were excluded if present in the union file in non-disease cohort (SLE or SRNS) patients.

Compound heterozygous variants For both SRNS and SLE cohorts, compound heterozygous variants were defined as two variants occurring on different alleles in the same gene. In order to identify potentially compound heterozygous variants that could be contributing to or causative for the disease a script was written to search each individual VCF in the SLE or SRNS patients and output all pairs of rare variants within a gene, if the variants were missense, nonsense or splicing variants.

To define rare pairs of variants, the product of the 1000 genome minor allele frequency (MAF) of each variant was used. A threshold of less than or equal to 0.01 for this artificial 'product MAF' was set. This does not take into account the likelihood that many of these pairs, whilst individually rare, will be in shared haplotypes and so likely to occur together, however it provides a conservative method to exclude common variant pairs. Where the MAF was unknown the variant was assumed to be very rare and assigned an arbitrary MAF of 0.0001 to permit the product MAF calculation.

Pairs of variants were also excluded if one or both of the variants had a PolyPhen–2 score of less than 0.15, indicating it is likely to be benign, or was flagged as lying within a region of segmental duplication.

Many of these pairs of variants were found on inspection to be within the same vary large genes with multiple paralogues, such as mucin genes, and other pairs were in fact two variants within the same erroneous read. To handle these sources of likely uninformative variant pairs, the variants were further filtered to remove pairs in which the two variants were less than 50 bp apart, and to remove all variants in genes containing more than 10 unique variant pairs, such genes must contain at least

5 different rare coding variants.

Finally the candidate compound heterozygous pairs were checked against the other individuals in the WGS500 cohort. The WGS500 union file provides frequencies of homozygous and heterozygous genotypes at all variant loci within the cohort. However it does not contain information about which individuals carry each variant. Thus in order to exclude compound heterozygous variants it was necessary to search each individual WGS500 VCF file for pairs of shared variants. Variant call files from cancer tissue and the disease in question were excluded. Putative compound heterozygous candidate variants were excluded if observed together in 3 or more WGS500 individuals outside the disease group (SRNS or SLE) and annotated if observed 1 or 2 times, this conservative approach was used because the phase of the variant pairs was unknown.

Filters for known disease genes The variant files for each individual were searched for rare mutations within both exonic and intronic regions in known disease genes, and in intergenic regions if one of the two nearest genes is a SRNS / SLE gene. Variants that were annotated as in segmental duplication region or failed quality filters were excluded. Heterozygous variants were excluded if the 1000 genome MAF was greater than 0.1, or if observed in the WGS500 union file in non SRNS patients, and homozygous variants were excluded if observed as homozygous for the alternate base in in non–SRNS / SLE patients in the WGS500 union file.

The candidate variants were either exonic (synonymous variants excluded), splicing or intronic variants, or non-coding variants with one of the 32 genes as one of the two a neighbouring gene. The latter were further annotated as intergenic, upstream (variant overlaps 1–kb region upstream of transcription start site), downstream (variant overlaps 1—kb region downstream of transcription end site) or in a 3' or 5' un-translated region.

2.17.2 List of SRNS genes searched

A list of genes known to cause SRNS was generated by searching the literature (McCarthy et al. 2013; Kopp 2013) and used to filter the SRNS patients sequencing data for variants within genes. These genes are all reported causes of monogenic nephrotic syndrome, but this list is inevitably subjective. *ITGB4, LAMB2, NPHS1, NPHS2, CD2AP, PTPRO, MYO1E, ACTN4, INF2, MYH9, ARHGAP24, ARHGDIA, tRNA-Leu, MT–TL1, COQ2, COQ6, PDSS2, WT1, NEIL1, LMX1B, SMARCAL1, NXF5, PLCE1, TRPC6, SCARB2, ALG1, APOL1, COL4A3, COL4A4, COL4A5, PMM2* and *ZMPSTE24*

2.17.3 List of SLE associated genes searched

Genes associated with SLE were collated from published literature and personal communication with Professor Tim Vyse. As with the SRNS gene set, this gene list is inevitably subjective.This gene set includes genes known to cause familial forms of autoimmunity, genes reported as associated with SLE based on genome wide association studies (GWAS) and genes identified through mouse studies. Genes linked to SLE from GWAS will in some cases be merely in linkage disequilibrium with another gene or regulatory region that in influencing the disease, the 1q23 risk locus includes *LY9*, *CD244* and other SLAM family genes (Graham et al. 2008). However in some cases evidence of functional consequences for the variant indicate that the GWAS risk variant directly affects immunity, for example the rs1143679 SNP in *ITGAM* was identified as an SLE risk factor in GWAS (Harley et al. 2008) and has been shown to perturb complement functions in human monocytes. Likewise the PTPN22 SNP rs 2476601 is associated with multiple autoimmune diseases and individuals with the risk allele have increased negative regulation of T cell activation (Vang et al. 2005). The genes searched are presented in table 2.7.

2.17.4 Splice site prediction

This was made using Splice Site Prediction by Neural Network (NNSPLICE) (Reese et al. 1997). For all non–intergenic filtered variants identified as linked to the 32 SRNS/FSGS genes two FASTA files were created, comprising 50bp downstream and 50bp upstream of the variant locus and either the reference of variant base(s). To generate these, 0–based BED format files containing the start and end co-ordinates were made from the (1–based) VCF format variant files, and FASTA format sequences were generating using the human reference genome (GRCh37, the Genome Reference Consortium human genome, build 37) and BEDTools fastaFromBed command (Quinlan and Hall 2010). These were then modified by a code to include the reference or alternate bases and batch submitted to NNSPLICE. The default cut off score of 0.40 was used for reporting of both donor and acceptor splice predictions.

2.18 Genotyping in the 17709 family

This was carried out using a HumanCytoSnp–12 DNA analysis BeadChip (version 2) (Illumina) which covers 220,000 markers for cytogenetic analysis. 4μl of DNA was used for each individual. This work was in collaboration with Dr Samantha Knight's group. The data was analysed with in house scripts to plot B allele frequencies (BAF) and difference in BAF for pairs of family members. Patient 17709 was not run on the SNP array so WGS data from the VCF file was compared to SNP array data. The strand can differ between the WGS (based on he reference genome forward strand) and the SNP array. Therefore the scripting to extract and plot BAF values included a method to check the base calls and correct for strand at any locus with information from the VCF and the SNP array. This was not possible at bases where the variant allele was a C / G or A / T transversion so these ambiguous loci were excluded from the BAF analysis, this excluded 14.8% (322205 / 2171663) of all known (dbSNP)

variants called in 17709. More variants were excluded as they were not present in both the VCF file and the SNP array, leaving 66808 sites, 3.1% of known variants in 17709, or 5.7% of the array loci (66808 / 1167678), across the genome that could be compared between the WGS and array data.

2.18.1 Detection of regions of homozygosity

SNPs from the array for 17709 family members, and VCF variant calls from SRNS and SLE individuals, were submitted to an online tool to identify runs of homozygosity (Seelow et al. 2009).

SNP array data from the family members for 17709 was converted to the required format and uploaded, along with a subsample of known dbSNP VCF variants from patient 17709, and searched for genetic homogeneity, excluding homozygous stretches present in HapMap controls. Plots showing excess of found homozygosity against the expectation calculated from population values across chromosomes (see also http://www.homozygositymapper.org/technical_documentation.html) , and bed files of genes within homozygous regions were downloaded and checked against the patient VCF for variants within predicted shared homozygous regions. For other SLE and SRNS patients a random subsample of known dbSNP VCF variants was uploaded as a sorted VCF. Where possible expected population values chosen were from the ethnically appropriate populations.

MHC association	Monogenic / familial disease	Mouse Model	GWAS / candidate gene study
BTNL2, HLA-DRA, HLA-DRB1, HLA-DQB1, HLA-DQA2, HLA-DQA1, HLA-DR3, HLA-DRB1, MSH5	C1Q, TREX1, C1S, C2, FCGR3A, C3, DNASE1, DNASEIL3, AGS5, ACP5	PIKFYVE, SLAMF6, SLAMF1, SLAM, LY9 (SLAMF3), CD244 (SLAMF4)	ITGAM, KIAA1542, 3q24, ACOX2, ATG5, BANK1, BDH1, CD244 (SLAMF4), CDKL3, ETS1, IKZF1, IL10, IL12A, IRAK, IRF5, IRF8, KCTD6, KIAA0226, LRRC18, LY9 (SLAMF3), MECP2, miRNA146a, NMNAT2, NSUN3, PDHB, PHRF1, PPP2CA, PRDM1, PTTG1, PXT1, RDBP, RNF150 , SKP1, SLC15A4, TCF7, TNFAIP3, TNIP1, TNPO3, TYK2, WDFY4, BLK, CD44, FCGR2A, FCGR3B, GLS, IFIH1, IKZF1, IKZF2, IRF7, JAZF1, LYN, NAB1, NCF2, NMNAT2, PTH2R,, PTPN22, PXK, RASGRP3, RSBN1, SMG7, STAT1, STAT4, TNFSR4, UBE2L3 , UHRF1BP1, XKR6

Table 2.7: SLE genes searched in the SLE cohort, divided by main source of evidence for association with SLE

Chapter 3

Use of Whole Genome Sequencing to analyse murine *N*–ethyl–*N*–nitrosourea pedigrees

3.1 Introduction to Chapter

3.1.1 ENU mutagenesis as a forward genetic tool in mice

Initial work on mouse genetics relied upon the collection of spontaneous mutations, usually bred by mouse fanciers. Spontaneous mutants have provided vital tools for the biologist such as the severe combined immunodeficiency syndrome (SCID) mouse (Bosma, Custer, and Bosma 1983), but spontaneous mutant discovery is limited by the rate of mouse mutation. The development of suitable mutagens for faster forward genetics depended on understanding the properties and effective doses of radiation and chemical mutagens. This was made possible by use of the specific locus test (SLT), developed in 1948 by William Russell. A mouse stock was created carrying seven easily visible recessive traits (affecting coat colour and ear morphology) that could be crossed to wild type males treated with a mutagen. The rate of observed

traits in the offspring was related to the rate of mutations induced at the seven loci (A P Davis 1998; Russell 1989). Russell went on to show using the SLT that unlike previously tested mutagens such as procarbazine and diethylnitrosamine, a single dose of 250mg/kg *N*–ethyl–*N*–nitrosourea (ENU) administered to male mice induced a mutation rate in offspring 7 times that achieved by X-ray (Russell et al. 1979). This rate could be further increased by fractionated dosing to 12 times that of X-ray or over 200 times the spontaneous rate (Russell et al. 1982a; Russell et al. 1982b). ENU is an alkylating agent that induces point mutations in spematogonial stem cells. The ethyl group of ENU can be transferred to a variety of target nucleophillic nitrogen or oxygen atoms in deoxyribonucleic acid (DNA), particularly O^2 and O^4 groups of deoxythymidine (dT). B6 mice have been shown to be relatively tolerant of ENU (Justice et al. 2000) and are thus a favoured strain for mutagenesis.

Forward genetic screens in mice carrying ENU induced mutations can provide important and entirely novel insights into gene function (Justice et al. 1999; Acevedo-Arozena et al. 2008; Hoebe and Beutler 2005; Papathanasiou and Goodnow 2005). This approach does not require any prior assumption about mechanism, and by inducing random point mutations ENU generates viable phenotypes that mimic human disease. Hypomorphic and gain of function mutations are generated in addition to null mutants. In the classic approach mice treated with ENU are bred to generate pedigrees segregating thousands of mutations, which are screened for phenotypes of interest. However determining which of the many induced mutations underlies the phenotype is a significant bottleneck in the process, requiring additional generations of breeding and out-crossing to another inbred laboratory strain in order to generate a linkage map, followed by sequencing of candidate genes or regions. This process is time consuming and costly. For example conventional fine mapping to obtain a linkage region of around 3Mb (20-30 genes) requires at least 2 generations of additional

breeding and genotyping 100-200 markers in 30-60 F2 mice [1].

The need to propagate the mice requires non-lethal screens, which limits the range of assays and the scope to detect phenotypes. Furthermore, out-crossing can introduce unseen confounding variants affecting the trait, and tracking the phenotype through additional generations is complicated and can be unreliable (Wansleeben et al. 2011).

Although whole genome sequencing (WGS) and whole exome sequencing (WES) offer the prospect of accelerating discovery, current strategies remain dependent on conventional mapping (Sun et al. 2012b; Arnold et al. 2011; Fairfield et al. 2011; Leshchiner et al. 2012).

Addressing the rate-limiting step of identifying the causative mutation will also go some way towards overcoming other limitations of ENU mutagenesis.

One issue is duplicate strains. Time can be wasted characterising a phenotype only to later discover that the gene or even mutation is already well known and the new mutant provides no novel information. A rapid method to isolate causative mutations, after screening but prior to detailed phenotypic testing, would focus downstream experimental work on the most important and novel variants.

Phenotypic screening is itself potentially inefficient, particularly if the screen is overly complex, difficult to replicate consistently and efficiently in a high throughput manner, or if the sensitivity or specificity of the screen is low. This was illustrated by previous work in the Cornall lab showing that 1 in 7 ENU pedigrees generate mice with heritable anti nuclear antibodies (ANA), due to mutations in as many as one in 200 genes (unpublished work), a prerequisite for SLE diagnosis in humans, but very few of these have overt autoimmune traits.

Next generation sequencing (NGS) provides a method to rapidly identify all the variants in an individual, offering the potential for more efficient and comprehensive mutation discovery.

[1] Xin Du, Yu Xia, Bruce Beutler in Protocol 'Genetic Mapping: Whole Genome Mapping and Fine Mapping' posted 05/05/2010 on http://mutagenetix.utsouthwestern.edu/protocol/

As a first step towards a NGS method to accelerate mutation discovery, mutations caused by ENU must be distinguished from the much larger numbers of false positive or non–ENU induced genomic variation detected by NGS.

This chapter develops a pipeline to detect ENU mutations within low coverage WGS data and demonstrates that causative ENU mutations can be identified with low coverage WGS.

3.1.2 Whole genome sequencing of ENU mice

To examine the efficacy of WGS as a tool to identify ENU induced mutations in mice, WGS was performed on 5 ENU pedigrees with immune phenotypes. Conventional linkage data was available for four of these pedigrees. Professor Chris Goodnow's group at Australia National University (ANU) generated ENU mice and carried out phenotyping and mapping to a chromosome or region. Professor Warwick Britton's group at the University of Sydney performed the *007* strain phenotyping.

In 4 pedigrees, (*222*, *NIH69b*, *NIH85a* and *NIH19a*) DNA from a single mouse exhibiting the phenotype from each pedigree was sequenced. In the *007* strain two affected mice were sequenced (*007_11* and *007_12*). Thus 6 mice from 5 pedigrees were sequenced in total. Figure 3.1 shows the coverage distribution for each mouse. Mean fold read depth was 5.3, 4.9, 4.9, 3.1, 3.2 and 3.4 in *222*, *NIH69b*, *NIH85a*, *NIH19a*, *007_11* and *007_12* respectively. The *007* and *NIH19a* mice have lower mean coverage and a distribution skewed towards zero, in comparison to the other mice sequenced. This appears to be a due to the imaging on the top surface of the flow cell failing for read 2 for the lanes in which these 3 samples were run, resulting in lower coverage.

All these pedigrees were sequenced at levels well below the 30 to 50-fold coverage recommended for accurate individual variant calling and genotyping (Koboldt et al. 2010; Subramanian S Ajay 2011). In order to identify ENU induced mutations

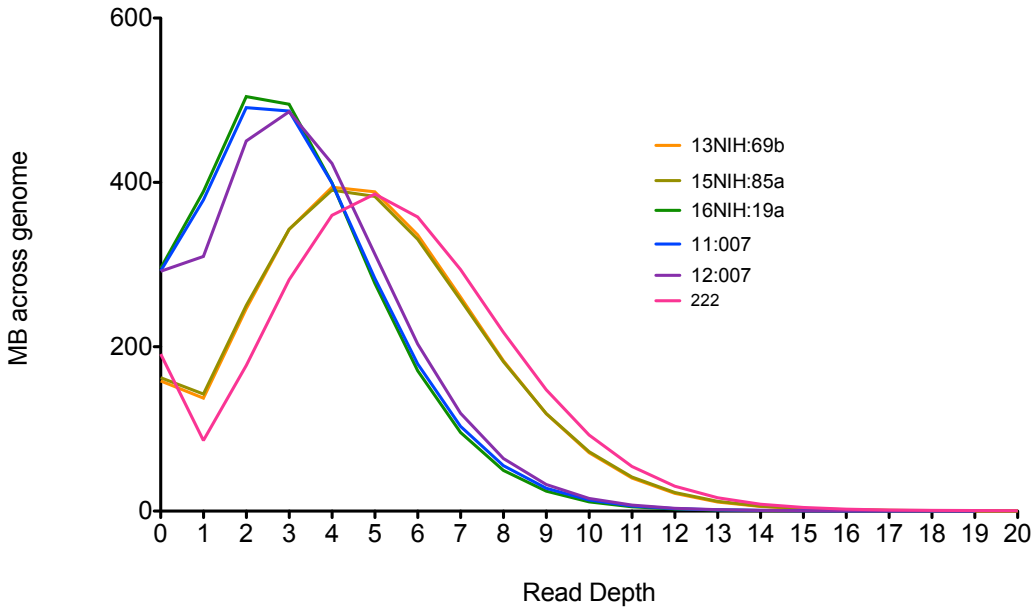


Figure 3.1: Plot of distribution of fold coverage depth, shown as number of megabases (Mb) of the genome covered at each read depth, for each sequenced mouse. 11:007, 12:007 and 16NIH:19a have lower peak read depth and a distribution that is relatively skewed toward lower coverage.

and infer pathogenicity in these mice a process to filter and prioritise variants was developed, this is briefly summarised in figure 3.2.

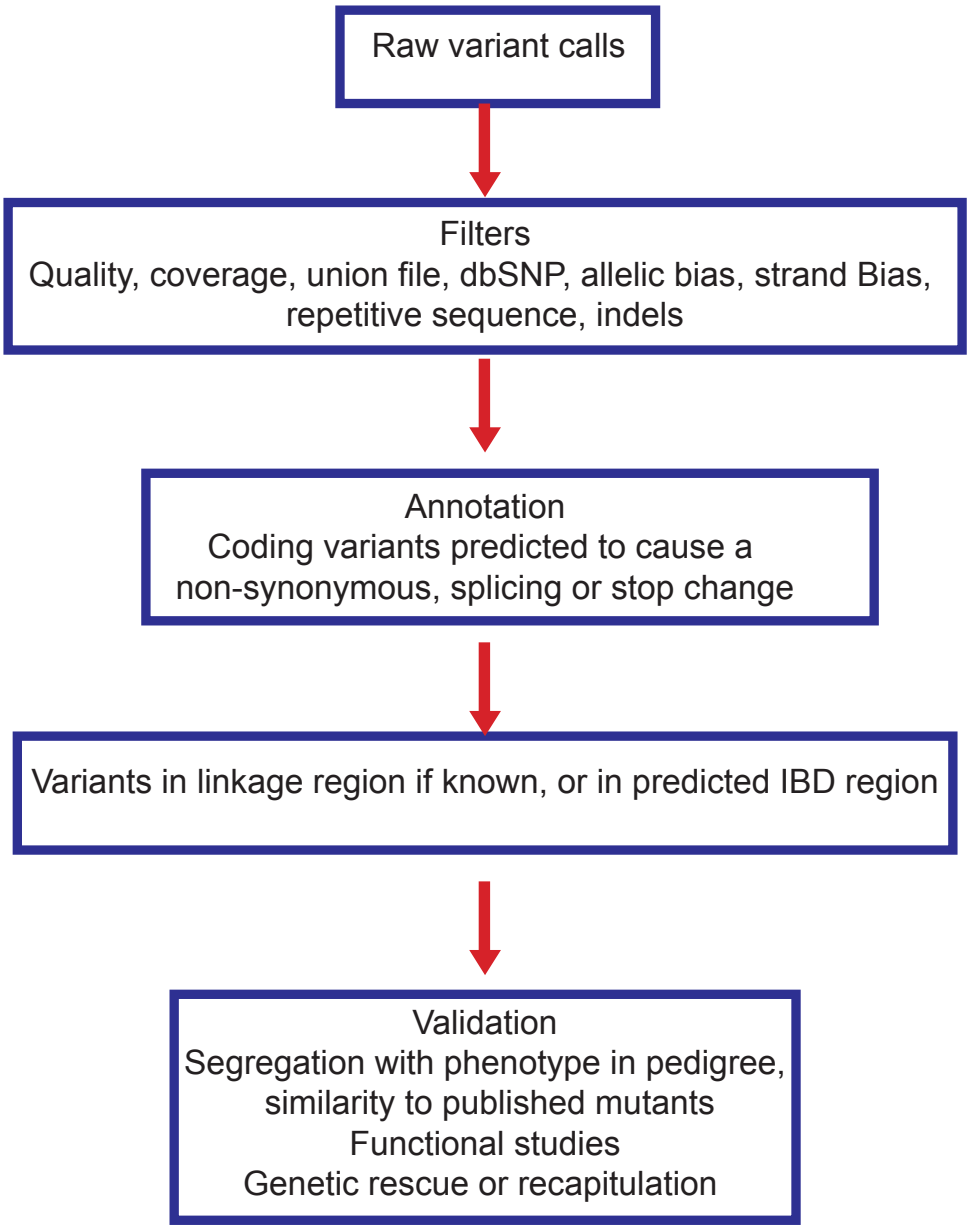


Figure 3.2: ENU variant study process. Filters are described in this chapter. Annotation was performed with Annovar. Specific validation steps depended on the novelty of the variants.

3.2 Development of filters to distinguish ENU mutations from non–ENU variation in WGS data

3.2.1 The need for filters, and the sources of non-ENU variants

The B6 strain is the same strain used to generate the mouse reference genome, against which the WGS reads were mapped. The raw variant call files contain many thousands of variants per mouse, despite an inbred B6 laboratory mouse background. ENU induces point mutations across the genome at a density previously reported to be between 0.5 Mb^{-1} and 10 Mb^{-1} (see section 5.1.2).

In this thesis the ENU mutation density is calculated to be 1.5 Mb^{-1} (CI 1.36-1.62) (section 5.5). The low coverage mice had on average 265,000 raw variant calls, but based on the calculated ENU density only 4,200 would be ENU mutations, Therefore 98% of the variation is not expected to be ENU induced. These non–ENU variant calls will be due to a combination of true variation from the reference genome, due to errors in the reference and genetic drift within breeding mouse colonies, and spurious calls due to errors in sequencing or the downstream mapping and variant calling processes. For example spurious calls can arise from mapping errors in highly repetitive or highly polymorphic regions of the genome. In order to remove errors due to these different sources a number of filters were developed. The rational for, and consequences of each filter are described below. All figures showing the absolute or percentage numbers of variants removed per filter are based on 5 mice, one from each pedigree, using only one genome from the *007* pedigree to avoid skewing due to overrepresentation of this strain. The filters were in practice applied to all the 6 sequenced individual mice.

3.2.2 Coverage depth

Accurate identification and genotyping of variants is very difficult with less than 3 reads at a locus. A single read error cannot easily be distinguished from a heterozygous variant at a locus covered by only 2 or 3 reads; but with more reads containing the reference base, the read error can be disregarded. Conversely loci with disproportionately high coverage are often due to 'pile up' of reads at repetitive regions of the genome that have been masked in the reference. Repeat masking replaces large or multiple repetitive regions with a single representative sequence. Due to variation in the multiple repeats mapped to the single region these masked regions will appear to be enriched for variants. Excluding very high coverage loci removes these uninformative regions. Therefore variants below the 2nd centile or above the 98th centile of the coverage distribution were excluded. If the lower bound was less than 3-fold read depth then 3-fold was taken as the lower bound for coverage and variant calls at loci with 2-fold coverage or less were excluded. A mean of 73,395 variants per mouse (standard deviation SD 32,395) or 29.61% of all variants per mouse (SD 15.74%) were excluded due to low coverage and 52,954 variants (SD 44,778) or 18.79% (SD 12.56%) were excluded because of coverage above the 98th centile.

3.2.3 Quality filters

The Illumina sequencing platform uses sequencing by synthesis technology. Fragments of single stranded DNA are randomly fragmented and ligated to adaptor sequences. DNA templates with universal adaptors are then distributed on a flow cell and immobilised on glass. Bridge amplification of these templates generates clusters of identical template sequences on the flow cell. Sequencing of the DNA then proceeds using reversible terminator chemistry. Amplification of the DNA extends by one single fluorescently labelled nucleotide in each cycle, after imaging the fluorecent label and blocking sequence are chemically removed ready for the next cycle (Bentley

et al. 2008).

The reads generated by the Illumina pipeline include quality information in FASTA format (Lipman and Pearson 1985). Each base within a read is assigned a quality score, which is an American Standard Code for Information Interchange (ASCII) based integer mapping of the probability that the base call is incorrect. Because fluorescently labelled bases are used to extend the amplified DNA fragments, this probability is based on the relative intensity fluorescence for the four images acquired for each read position, with some correction for phasing and pre-phasing and use of statistical learning to train the base calling software (Kircher, Stenzel, and Kelso 2009). The mapped reads are then assigned a mapping quality score by the mapper used: Stampy (Lunter and Goodson 2011). This is the Bayesian probability that the read pair is mapped incorrectly and incorporates priors based on read errors, SNPs and indels, and considers the possibility that the correct mapping was not chosen either due to read errors or variants, because of repetitive sequences or because the sequence is not present in the reference. Finally the variant caller Platypus assigns a Phred based quality score to each variant, which incorporates the base quality and mapping quality scores. Phred scores are log transformed error probabilities, such that a Phred score of q defined as:

$$q = -10log_{10}(p)$$

Where p is the error probability for a base call, or in this case a variant call, i.e. the probability that the variant call is incorrect (Ewing and Green 1998). In order to reduce false positive variant calls due to incorrectly mapped or poor quality reads, variant calls with Platypus assigned quality scores below 20 were excluded. Since the quality scores are Phred based this corresponds to probabilities of an incorrect variant of 0.01 or more (Table 3.1). This filter alone excluded a mean of 36 variants per mouse (SD 14.1), corresponding to 0.013% (SD 0.0032%) of all variants per mouse.

PHRED QUALITY SCORE	PROBABILITY OF INCORRECT CALL
10	1 in 10
20	1 in 100
30	1 in 1000
40	1 in 10000
50	1 in 100000

Table 3.1: Example Phred scores and their corresponding probabilities

3.2.4 Strand bias

The term strand bias describes a variant for which the genotype information from the forward and reverse strands is inconsistent. For example if non-reference bases are only observed on the forward strand. This phenomenon has been observed in data generated across different mapping platforms and may be due to library preparation or sequencing errors. Whilst the exact cause remains unclear, SNPs with extreme strand bias are enriched for false positives (Guo et al. 2012). Variant calls with a one sided cumulative binomial probability of less than 0.001 for the alternative forward count or the alternative reverse count were excluded (Methods). The number of variants that could be excluded with this approach was limited by the low sequencing coverage, indeed a probability of less than 0.001 for the observed bias in forward and reverse strand reads can only occur at loci with at least 9 reads. On average only 6.4% of loci had coverage of at least 9-fold. However to improve the power to detect strand bias this filter was applied to look for systematic strand bias by examining the union file of variants called simultaneously in the 6 mice, This increased the coverage at each variant locus, such that 47% of variant loci had at least 9-fold coverage. On average this filter alone excluded 23,445 variants per mouse (SD 4,523) or 8.80% of all raw variants (SD 0.60%).

3.2.5 Allelic bias

Visual inspection of variant calls revealed that some were observed only in a small proportion of the reads, much less than the theoretical 100% of reads expected for a homozygous variant or the 50% of reads for a heterozygous variant. Since ENU mutations are germline rather than somatic these are unlikely to represent true positive variants, at least not mutations attributable to ENU. Based on an average somatic cell mutation rate of 0.77 x 10^{-9} per base per cell division, an individual replicating human cell is estimated to carry 10^3 to 10^4 mutations by the time a human reaches 15 years of age (Lynch 2010). Whilst mice accumulate less somatic mutations than humans during their short lifespan, like humans, the rate of somatic mutation in mice is higher than in the germline (Hill et al. 2005).

Therefore variants with one sided cumulative binomial probability for the observed distribution of variant to reference reads of less than 0.001, assuming a probability of 0.5 for a variant read, were excluded. This filter alone excluded 27,711 variants per mouse (SD 23,565) or 9.69% of all raw variants (SD 6.60%), this parameter varied substantially between individual sequenced mice (range 5.13 to 21.11%). The variability may have been a function of low coverage resulting in inability to distinguish allelic differences rather than actual heterogeneity between the mice for false positives with bias for this parameter. Since at least one read must be non-reference for a variant to be called, the minimum coverage required to reach the threshold probability of less than 0.001 is 11-fold coverage. 11-fold coverage or more was achieved in on average 2% of individual per mouse genomic loci, however this ranged from 4.3 to 0.4% between individual genomes, consistent with low coverage explaining the variability in apparent alleic bias.

3.2.6 Duplicate reads

During the library preparation step for Illumina sequencing fragmented DNA is amplified by polymerase chain reaction (PCR). If reads generated from the same original genomic DNA fragment are sequenced in more than one individual cluster on a flow cell this will result in duplicate information, potentially leading to errors in SNP detection, particularly at low coverage. If a single fragment with an error is duplicated, and each duplicate read containing the same error is treated as an independent occurrence in the DNA then the error will be over-represented and the locus incorrectly considered to have a real variant. PCR duplicates occur more frequently if there is insufficient starting DNA, because more cycles of library amplification are needed. Exact duplicate pairs of reads mapping to the same location are unlikely to occur by chance due to the random nature of shearing during the DNA fragmentation step in library preparation. Hence PCR duplicates can be distinguished in the sequencing data as multiple read pairs with precisely the same mapping locations and orientation. Duplicate reads were removed from the .bam format mapped read file generated by Stampy, prior to variant calling (Methods).

3.2.7 Homopolymers and repetitive sequence

A genetic homopolymer is a sequence of identical bases. Homopolymers and other low complexity regions, such as di–nucleotide repeats, present difficulties for accurate variant calling. These regions evolve rapidly (Brown et al. 2002) and in some cases this has been shown to be due to replication slippage (Dieringer and Schlötterer 2003). DNA polymerase pauses during replication of a repeat region, the newly synthesized strand then separates from the template strand and re-anneals to a complementary repeat upstream, and replication resumes, resulting in expansion of the repeat region. Consequently repetitive regions are highly polymorphic in the population and vary in comparison to the reference genome. This can be a source of sequencing error or

generate large numbers of variant calls that represent true variation from the reference but may be unique and therefore not removed by comparison with other genomes. Therefore the variants were further filtered to remove variants within low complexity repetitive loci based on the neighbouring sequence.

Platypus uses a variant call format (VCF) (Danecek et al. 2011) that includes a homopolymer score, corresponding to the number of identical bases surrounding the variant, and also reports the actual sequence of 20 base pairs (bp) surrounding the variant locus. Simple homopolymer runs of 4 or more (including the variant) were excluded based on the homopolymer score, and more complex but repetitive regions were identified by excluding distinct 20 bp sequences surrounding loci that were observed more the 4 times in the raw set of variant calls for the individual mouse (Figure 3.3). This detected di-nucleotide repeat sequences not picked up by the homopolymer filter and also identified frequent occurrences of regions in which a di-nucleotide repeat appears to have a single inconsistent base, for example CACA-CACACATACACACACAC. This particular sequence was observed on average 1,149 times per mouse. Such sequences may represent real variation due to recurrent and polymorphic low complexity regions. Another frequent sequence pattern was a di-nucleotide repeat changing by one base to another di-nucleotide repeat sequence, e.g. TCTCTCTCTCTCACACACACA, observed a mean of 514 times in each variant call set. Again length variation due to slippage could account for the large number of variants called at these loci. Homopolymers of 4 bases or more comprised a mean of 47,086 variants per mouse (SD 6,166) or 17.80% of all variants called (SD 0.42%). Repetitive sequence contexts of 20 bp surrounding the variant call and seen 4 or more times in an individual mouse raw variant call set comprised a mean of 45,429 variants (SD 10,059) or 17.09% of all raw variants (SD 2.62%).

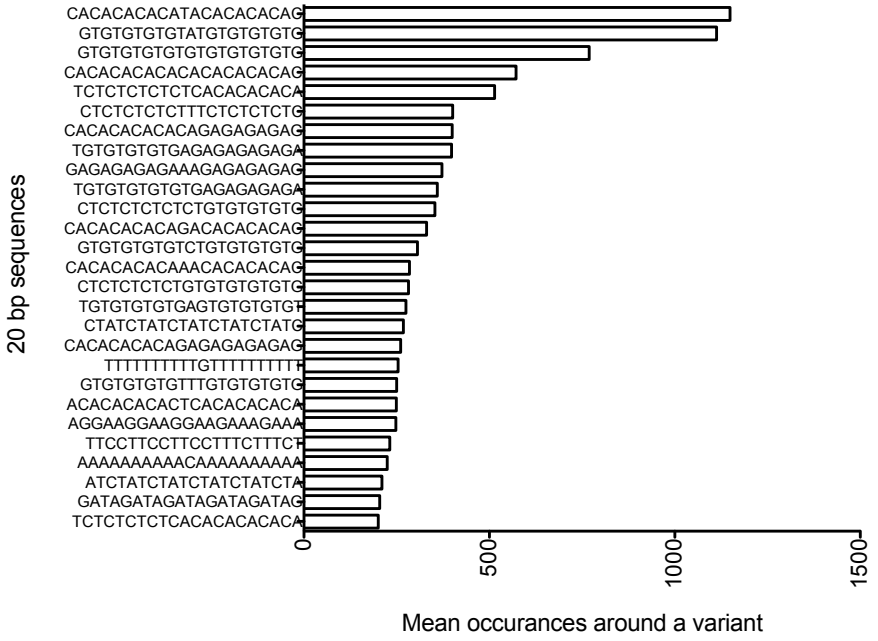


Figure 3.3: All 20 base pair sequences surrounding a variant with a mean frequency greater than 200 per genome in the 5 sequenced pedigrees

3.2.8 Indels

ENU mutagenesis overwhelmingly induces point mutations, and whilst occasional insertions and deletions have been observed in ENU programmes (Hoebe and Beutler 2005), in some cases these have been demonstrated to have arisen independently of the mutagen (Croker et al. 2007). Furthermore indels are less reliably identified than SNPs by current mapping techniques. A mean of 100,605 indels (SD 13,331), or 38.04% (SD 1.06%) of all variants per mouse were identified and excluded.

3.2.9 Published variants

Mutations induced by ENU arise as de novo germline variants within the offspring of the treated mouse, hence they are not expected to be in databases of known mouse variation. It is possible that by chance an ENU mutation will replicate the position and base substitution of a known variant but such events will be rare. The

mouse dbSNP database contains nearly 140 million variants, so the chance of a single ENU mutation replicating a dbSNP with the same base substitution is around 1.7%. Variants matching those from the mouse dbSNP database (version 128) were removed. These comprised a mean of 14,217 variants per mouse (SD 937) or 5.50% of all variants (SD 1.13%).

3.2.10 Filters against mutations in other pedigrees

Mouse dbSNP does not contain all true mouse variation. Additionally erroneous calls will exist within the set of raw variant calls. Most spurious variant calls are due to systematic errors in the sequencing, mapping and calling pipeline that are likely to be replicated in multiple sequenced individuals. In order to remove both non–ENU true variation from the reference and calls arising from systematic errors, the variants in each of the 5 pedigrees described above were checked against a union file of variants called simultaneously from whole genome sequences for 9 ENU mouse pedigrees, including the 5 described pedigrees.

Given the ENU mutation density of 1.5 mutations Mb^{-1} (section 5.5), a typical ENU G_1 carries around 4090 mutations. An individual G_3 mouse from the 2 founder ENU breeding strategy (Figure 5.1a), inherits 75% of the genome from an ENU treated founder and therefore will carry approximately 3070 homozygous or heterozygous ENU mutations. Ignoring any bias in ENU mutation sites, the chance of 2 mice from different ENU treated founders carrying the same mutation at the same site can be estimated as a binomial probability distribution. The probability of at least one mutation from mouse 1 being observed in mouse 2 is

$$x = 0$$

$$n = 3070$$

$$p = \frac{1}{2.7 \times 10^9 \times 3}$$

$$M = (1 - \binom{n}{x} p^x (1-p)^{n-x})$$

Where P is the probability of a given mutation occurring at a specific site and base substitution. In this way M is estimated as 3.79×10^{-7}.

A mean of 257,244 variants per mouse (SD 35,611) or 97.17% (SD 1.76%) of all variants were observed in more than one pedigree in the union file and thus excluded as candidate ENU mutations. Many more variants were excluded using the union file than by any of the other filters applied. All other filters combined removed only an additional mean 5,182 variants (1.95%) per mouse. The numbers (Figure 3.4a) and percentage (Figure 3.4b) of variants excluded by each filter individually are shown.

To investigate redundancy across the set of filters, the number of filters 'failed' by individual variants was measured. A variant was considered to have failed a specific filter if it would be excluded based on the thresholds described above. Figure 3.4c shows the percentage of all variants for each unique observed combination of failed filters, the 25 most frequent combinations are shown. Figure 3.4d shows the percentage of all variants excluded by only one filter, for each filter. After the union file, filtering high coverage variants removed the most variants not excluded by any other filter. Figure 3.4e shows the number of filters 'failed' against the percentage of all variants. The majority of raw variants were excluded independently by 2 or 3 filters, indicating redundancy between the filters. A mean of 2,792 variants (SD 904.2) or 1.03% of all unfiltered variants (SD 0.18%) per mouse were not excluded by any filter and these were subsequently examined for candidate variants.

3.3 The effect of laboratory and strain on shared variation

The union file of variants from 9 ENU pedigrees was used to examine the relationships between laboratory, strain, and shared variation in mice. Differences in phenotype between mice of the same background are often reported by different laboratories,

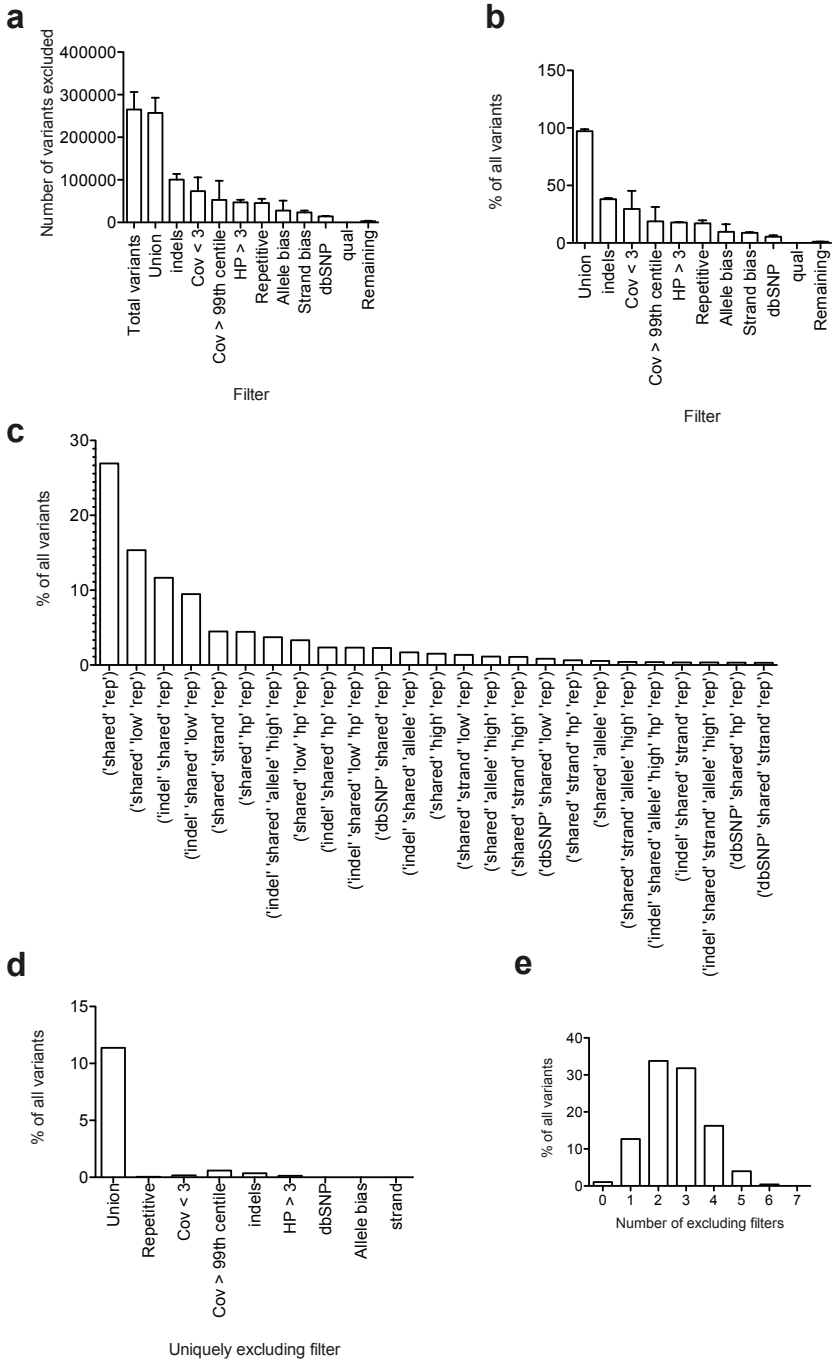


Figure 3.4: Numbers and percentages of raw variants excluded by each filter. (a) Numbers of raw variants excluded by each filter independently. Mean and SD over 5 mice from different ENU pedigrees. (b) Percentage of variants excluded by each filter independently. Mean and SD over 5 mice from different ENU pedigrees. (c) Overlap or redundancy of filters shown by percentage of variants 'failing' each combination of filters, the 25 most frequent combinations are shown (25/90 observed combinations) . (d) Percentage of all variants excluded by only one filter. (e) percentage of all variants by number of filters failed.

and sometimes attributed to genetic drift between mouse colonies. Genetic drift does occur and can result in confounded results or sub-strain divergence (Threadgill et al. 1997; Sluyter, Marican, and Crusio 1998; Christian G Specht 2001). The union file was used to explore the extent of variation that is due to inter colony genetic drift.

7,624,313 unfiltered calls were identified amongst 9 pedigrees from 3 centres independently running ENU programmes (ANU Canberra, MRC Harwell and Scripps / UT southwestern) (Figure 3.5a and Methods). The 2 mice obtained from MRC Harwell had been out–crossed to a non–B6 laboratory strain, because the current Harwell ENU linkage method includes out-crossing at an early generation.

Only shared, non–ENU variation is relevant for this analysis of inter laboratory variation, so in order to exclude ENU mutations variants observed in only one pedigree were removed. 83.6% (6,371,574 / 7,624,313) of raw variant calls were shared by more than one pedigree.

Figure 3.5b shows the degree of sharing of these variants between the 3 laboratories. Within this dataset of shared variants there were 6,371,548 unique genomic positions, 81.7% were SNPs (5,203,298) and 18.3% (1,168,250) indels. The transition:transversion ratio among SNPs was 1.9 (3,394,771 transitions, 1,805,983 transversions, 2,544 sites have more than 2 alleles).

Some of this variation will be due to the outcrossing strain used at Harwell. To examine the proportion of this shared variation attributable to B6 reference strain mice, shared variants exclusively observed in mixed strain mice were excluded. The resultant 628,973 shared variants were therefore observed in at least 2 pedigrees, including at least one fully B6 pedigree. Since all pedigrees included from MRC Harwell are mixed strain, the MRC Harwell variants fully overlap the other centres (Figure 3.5c). The large majority (91.8%) of all shared non-ENU variation in these 9 B6 mice is not laboratory specific.

The analysis of shared variants seen in at least one straight B6 was repeated using

only variants passing filters for coverage, allele or strand bias and quality as described previously (Figure 3.5d) in order to reduce the contribution of systematic error to this analysis. This dataset of 448,964 variants should have a greater proportion of true variants compared to spurious variant calls. 86.0% (386,162 / 448,964) of these variants are not laboratory specific, suggesting isolated genetic drift within individual colonies accounts for less than 15% of the observed variation. 24.1% of the shared dataset in figure 6d were present in dbSNP, setting a lower bound on the amount of the observed variation that is likely due to true variation from the reference.

3.4 Results of filters and causative or candidate variants identified in 5 pedigrees sequenced at low coverage

The 1.03% of raw variants remaining after filtering were examined to identify candidate causative ENU mutations. Variants annotated as causing non-synonymous miss-sense (NS) changes, inducing or disrupting a stop codon (non-sense), or predicted to affect splicing were prioritised as mutations that could have phenotypic consequences.

A mean per individual of 22 filtered variants across the genome were in coding regions and predicted to induce either NS, stop gain/loss or splicing changes. Using a combination of the filters described and knowledge of the linkage region from conventional mapping was sufficient to reduce the number of variants to 2 to 4 candidates for 3 of the 5 pedigrees described above (*222*, *NIH85a* and *NIH69b* (Table 3.2). This was irrespective of the mode of inheritance, genotyping variants as homozygous or heterozygous in a single mouse at very low coverage is unreliable so genotype was disregarded. A variant was considered to be a candidate if it passed all the filters

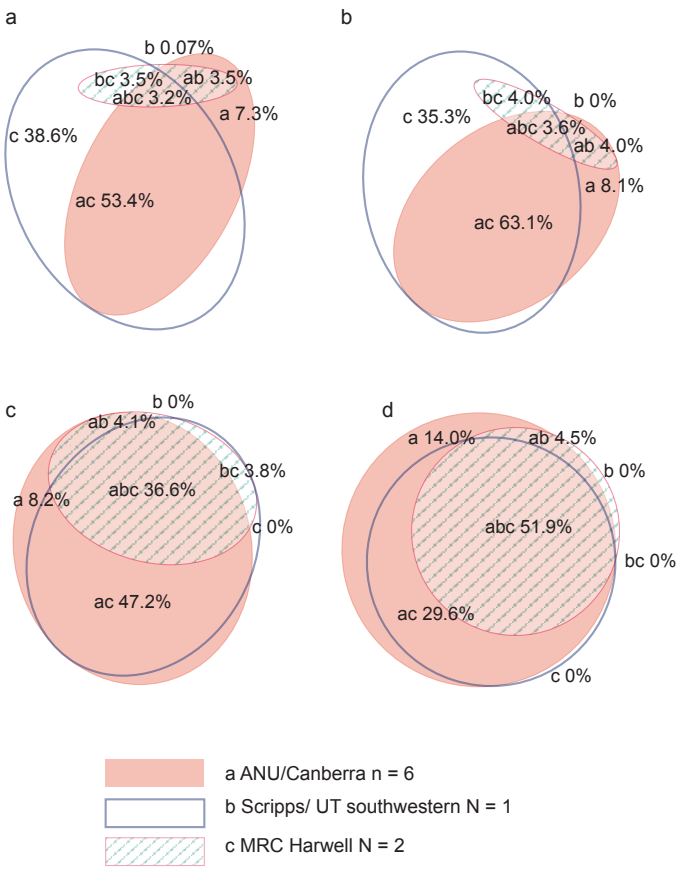


Figure 3.5: Variant Union File - distribution of variations by laboratory. (a) All variants called over 9 ENU pedigrees from 3 centres. (b) All variants observed in more than one pedigree. (c) All variants seen in more than one pedigree including at least one pedigree on a straight B6 background. (d) As (c) but after filters applied for low or high coverage, allele or strand bias and quality.

described in section 3.2, was within any known linkage region, and affected protein sense. In 2 cases (*NIH85a*, *NIH19a*) this short-listing allowed rapid identification of the causative variant based on knowledge of the phenotype and the known functions of these genes.

Pedigree	Phenotype	Mapping	Raw Variants	Filtered Variants	Protein coding variants (NS, stop or splice) across genome	Filtered Variants in mapping region	Protein coding variants in mapping region	Causative Mutation	Gene	Inheritance
222	Accumulation of peritoneal B cells	Chr 11	324789	3031	46	79	2	Unknown		recessive
NIH85a	Shifted IgD	Chr 2 93-173Mb	272898	3059	25	115	4	chr2 126746052 A:T	Sppl2a	recessive
NIH19a	Low T cells, activated T cells and B cells	Chr 11	228969	1941	13	110	3	Chr11 34414481 C:A	Dock2	recessive
NIH69b	Slightly lower IgD	not mapped	274996	2853	19	2820	19	chr4 4455038 A:T	Pax5 *	recessive
007_11	TB susceptibility	Chr6	224436	2090	8	162	0			Dominant
007_12	TB susceptibility	Chr6	34676	2456	17	222	1	chr6 125306955 A:G	Tnfrsf1a	Dominant
007 shared 11 and 12	TB susceptibility	Chr6	41338	2884	19	260	1	chr6 125306955 A:G	Tnfrsf1a	Dominant

*Table 3.2: The effect of filters and linkage region on the number of variants and causative mutation where known. The analysis focussed only on protein coding variants that were predicted to alter the amino acid sequence, either as non-synonymous substitutions (NS), gain or loss of a stop signal or alteration of splicing. 'Causative mutations' were validated by segregation with the phenotype in the pedigree (by collaborators at ANU) and either by similarity with known mutants (Dock2, Pax5 and Tnfrsf1a) or for a novel variant, experimental rescue (using retroviral transduction of wild type cDNA) and functional testing (Sppl2a). This experimental work was also performed by collaborators (Bergmann et al. 2013). * Pax5 variant was identified after further exome sequencing as it was not in the 19 candidates seen in WGS. This was due to low coverage.*

As shown in Table 3.2 the causative variant was identified using WGS in 2 pedigrees (the variants in *Pax5* and *007* were identified by collaborators using WES), in each case this was validated by Sanger sequencing in other mice within the pedigree (by collaborators at ANU).

3.4.1 *NIH19a* has a mutation in Dock2

3 candidates were identified within chromosome 11, the known linkage chromosome in pedigree *NIH19a*. A mutation at position 34,414,481 on chromosome 11, encodes a homozygous NS mutation in dedicator of cyto-kinesis 2 (*Dock2*), a Rac guanine exchange factor in lymphocytes. *Dock2* knockout mice exhibit defective T and B cell migration in response to chemokines resulting in lymphopaenia with excessive CD4+ T helper responses (Tanaka et al. 2007).The *Dock2* phenotype is well characterised and consistent with the phenotype observed in *NIH19a*.

3.4.2 *NIH85a* has a mutation in *Sppl2a*

In *NIH85a* 4 variants were identified within the linkage region, of these, the most plausible candidate was a splice site mutation on chromosome 2, position 126,746,052 in signal peptide peptidase like 2a (*Sppl2a*). The mutation disrupts splicing at exon 8, amino acid 842+T>A, resulting in a truncated protein (Figure 3.6).

Signal peptide peptidases cleave the intra-membrane fragment of trans-membrane proteins and following the identification of this mutation by WGS, our collaborators showed that *Sppl2a* is critical for proteolytic processing of CD74 in B cells and dendritic cells, and that in the ENU mutant the absence of *Sppl2a* causes an intrinsic B cell developmental block (Bergmann et al. 2013).

In mouse *222* and the *NIH69b* pedigree the variant was not identified.

3.4.3 A mutation in *Pax5* in *NIH69b*

Subsequently collaborators in the Goodnow group at ANU used exome sequencing to identify the causative variant in *NIH69b* as a mutation in *Pax5*. This locus was only covered by one read in our WGS data, illustrating a limitation of very low coverage WGS of individual mice for variant identification. From the coverage distribution in

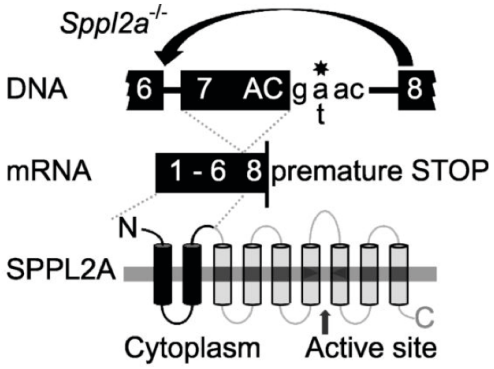


Figure 3.6: The splice site mutation in Sppl2a causes complete skipping of exon 7 in the mRNA and a frame shift and premature stop codon, resulting in a truncated protein lacking 7 of the 9 transmembrane domains including the protease site. Figure taken from Bergmann (2013) with permission.

NIH69b (Figure 3.1) it can be calculated that there was a 19.9% chance of missing the homozygous causative variant due to insufficient coverage, defined as less than 3 reads. (Assumptions: no other sources of error leading to failure to call variant at a locus with 3 or more reads, and no skewing of low coverage regions away from coding regions, in reality both of these will have an effect).

3.4.4 No causative variant is identified in *222*

Likewise we failed to identify the causative variant in mouse *222* despite linkage to chromosome 11. In *222* 7.0% of loci were not covered by any read (Figure 3.1). 16.7% of the genome was covered by less than three reads, and a mutation in these regions would be missed. *222* has a phenotype of abnormal peritoneal accumulation of B and T cells in response to immunisation. 24 candidate variants were identified after filtering across the genome, of which two were on chromosome 11, the linkage chromosome. One of these, at position 73,969,147, encodes a NS variant in olfactory receptor 402 (*Olfr402*). This variant was observed in only 3 out of 7 reads, inconsistent with the recessive mode of inheritance in *222*, with a low quality score (22) for the

variant.

The other candidate variant on chromosome 11 was a NS mutation at position 107,800,906 in the protein kinase C alpha gene (*Prkca*) (Figure 3.7a and Figure 3.7b), inducing an isoleucine to threonine substitution at amino acid 648 in the protein. This appeared consistent with the phenotype given that *Prkca* null mice exhibit impaired NFAT-dependant T cell activation (Gruber et al. 2009), and Th1 cell responses after immunisation (Pfeifhofer et al. 2006). PRKCA is a serine-threonine specific kinase that can be activate in response to calcium or diacylglycerol and is regulated by interaction with membrane phospholipid. The *Prkca* variant seen in 222 lies within the C-terminal catalytic domain, and mutations just distal to this have been shown to be critical for function (Stensman and Larsson 2007; Yeong et al. 2006).

To explore whether the observed Prkca variant in the *222* strain affects activation, green fluorescent protein (GFP) labelled wild type (WT) PRKCA and Mutant PRKCA^{I648T} expressing cell lines (Figure 3.7c and Figure 3.7d and Methods) were made and used to examine the translocation of the protein to the cell membrane using confocal microscopy (Figure 3.7e).

Mutations in the V5 domain have been reported to increase the sensitivity to 1,2-Dioctanoyl-sn-glycerol (DOG), a cell permeable analogue of the protein kinase C activating second messenger Di-acyl Glycerol (Stensman and Larsson 2007). However in both PRKCA and WT, DOG alone failed to stimulate translocation in NIH3T3, COS7 or BE(2)-M17 cell lines, consistent with the low affinity of WT PRKCA to DOG (MacEwan et al. 1993). Combinations of the protein kinase inhibitor GF109203X and DOG (Stensman, Raghunath, and Larsson 2004) allowed sustained translocation to the membrane (Figure 3.7e).

The PRKCA^{I648T} mutation had no effect on the duration of translocation but may delay the onset of the translocation following stimulation, though this was not statistically significant; the unpaired two-tailed t-test p value was 0.1897 (Figure 3.7f).

Further Sanger sequencing on other affected mice from the 222 pedigree (by Belinda Whittle at ANU) then revealed that the mutation in *Prkca* did not segregate with the phenotype. Given the coverage distribution the sequencing may have failed to adequately cover the causative locus even for the recessive *222* phenotype. *222* was sequenced on 7 lanes of an Illumina GAII machine at an early stage in the project before improvements in fluidics lead to marked increases in overall coverage and more even coverage distribution.

3.4.5 *007* has a mutation in *Tnfrsf1a*

In the *007* pedigree WGS data from one mouse (*007_11*) did not identify any candidates within the mapping chromosome. However by sequencing DNA from two affected mice more candidates were identified (Table 3.2). The variants for these two individual mice were called simultaneously by the SNP caller allowing detection of variants observed at very low individual coverage by combining the read information.

Identified variants were examined in two ways to explore the contribution of simultaneous sequencing of the affected pair. Firstly variants were counted for each individual mouse, disregarding variant call information from the other mouse. Next the variants were examined for potential candidates shared by both mice, in the latter case we also included variants called in one mouse but with insufficient information for the variant caller to ascribe a genotype in the other mouse, typically because there were no reads or only one read in that mouse.

Across the genome (disregarding the mapping information) there were 8 candidates in *007_11* alone, 17 candidates from *007_12* and 19 potential shared candidates using the variant calls in both mice (Table 3.2). The number of shared candidates is larger than either individual set of candidates due to the ability to detect variants that only reach adequate coverage to genotype in one of the two mice. This is illustrated by the only variant on the linkage chromosome, Chromosome 6, which was

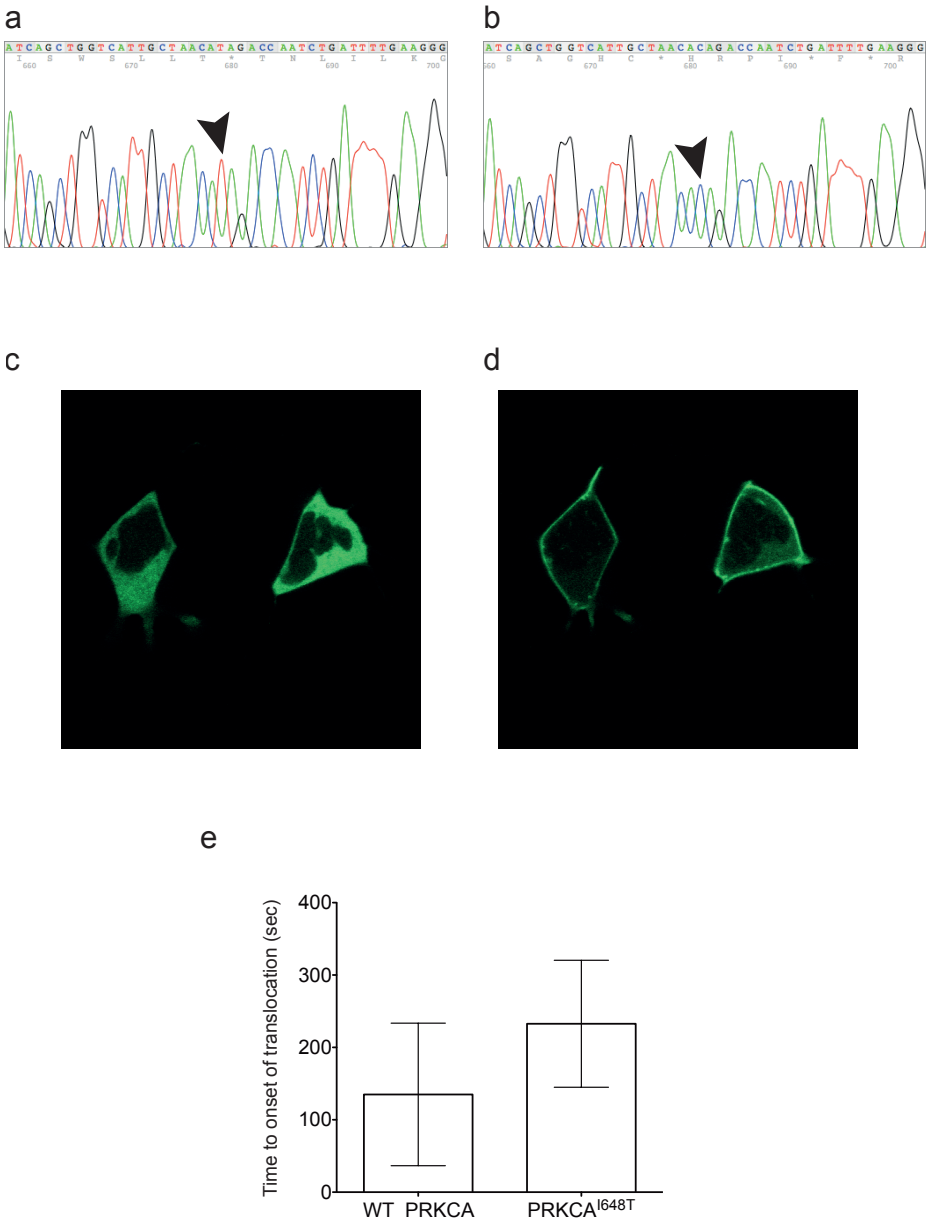


Figure 3.7: PRKCA variant in ENU pedigree 222. Site directed mutagenesis was used to replicate the mutation identified in mouse 222 within an N-terminal GFP labelled mouse Prkca DNA vector clone. Sanger sequence trace of the region surrounding the variant locus in WT (a) and Mutant (b)Prkca clones. Prkca in pEGFP-N1 vector was transfected into BE(2)-M17 cells, using polyethylenimine (PEI). Example images from a time-lapse confocal series are shown. WT Prkca expressing BE(2)-M17 cells are shown at 5 seconds (c) and 210 seconds (d) post stimulation, illustrating translocation to the cell membrane. WT and Mutant cells were stimulated with 2-[1-(3-Dimethylaminopropyl)-1H-indol-3-yl]-3-(1H-indol-3-yl)maleimide Bisindolylmaleimide (GF109203X) 2mM followed by 1,2-Dioctanoyl-sn-glycerol (DOG) 400mM. Time to onset of translocation in WT and Mutant Prkca expressing cells with standard deviation, based on 3 separate experiments for each condition (e).

observed in 2 reads in *007_12* and in the shared set but was not called if only variants for *007_11* were considered in isolation because of low coverage. The locus was only covered by one very low quality base (Phred score 2) in a single read in *007_11* .

Thus in this case, although sequencing a second affected mouse *007_12* improved coverage at the variant locus, this variant was not in the filtered shortlist, but was discovered by collaborators at ANU using exome sequencing at much higher coverage depth. The causative mutation is a non-synonymous change in tumour necrosis factor superfamily member 1a (*Tnfrsf1a*) on chromosome 6 at position 125,306,955, a histidine to arginine substitution. On re-examining the WGS data we observed the variant in only 2 reliable reads as described above. Consequently the variant caller did not call the variant, illustrating the limitations of very low coverage data (Figure 3.8).

The 007 pedigree has a phenotype of susceptibility to Mycobacterium Tuberculosis (MTB) challenge, and the identified variant in *Tnfrsf1a* is consistent with the phenotype. *Tnfrsf1a* encodes the Tumour Necrosis Factor Receptor 1 (TNFR1) protein, a member of the tumour necrosis factor (TNF) receptor super-family and the major receptor for soluble TNF-alpha induced signalling (Wajant, Pfizenmaier, and Scheurich 2003).

A monoclonal TNF–alpha neutralizing antibody causes mice to develop necrosis in addition to granulomas in response to MTB. B6 mice are relatively resistant to MTB but succumb to fatal infection after treatment with hamster monoclonal antibody to TNF-alpha. Similar effects are observed in mice with a disruption to *Tnfrsf1a* (TNFRp55-/-) (Flynn et al. 1995). TNFR1 knock out mice injected with Bacillus Calmette Gurin vaccine (attenuated Mycobacterium bovis) develop fewer and smaller liver granulomas than WT. Administration of soluble TNFR1 reduces granuloma number and size (Senaldi et al. 1996), indicating that TNFRI contributes to granuloma formation and can also induce regression. TNF may support MTB

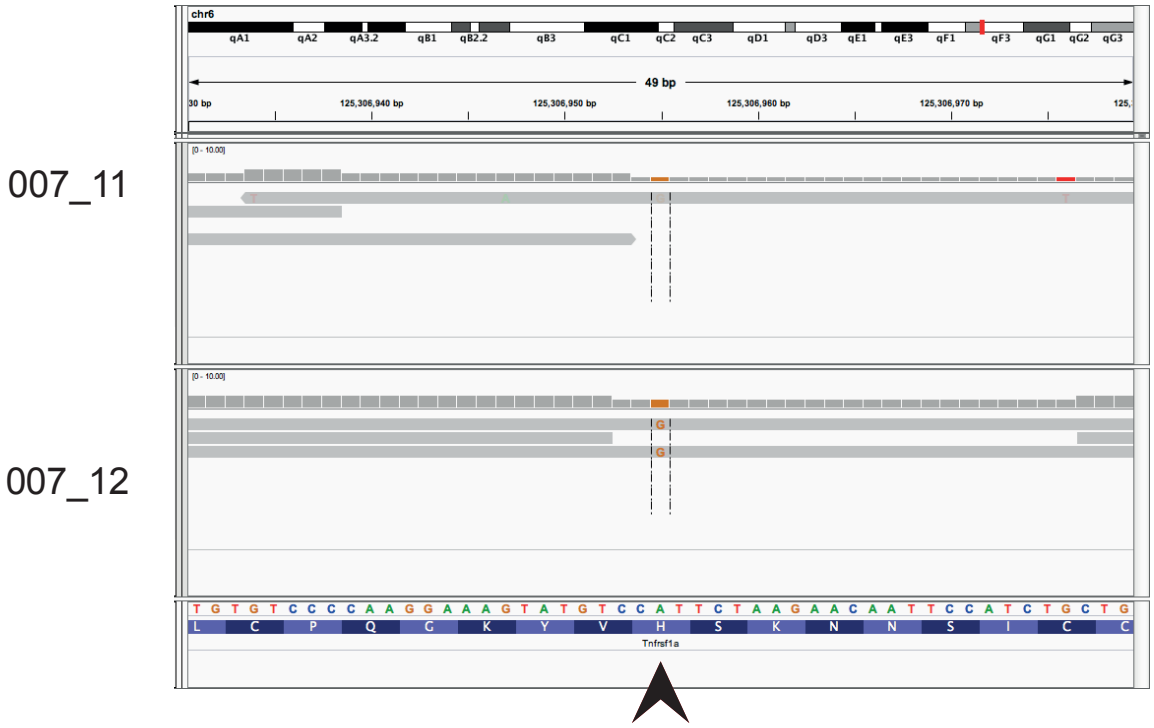


Figure 3.8: Reads covering the Tnfrsf1a mutation in 007. Integrated genomics viewer (Thorvaldsdottir, Robinson, and Mesirov 2013) image showing mapped reads for 007_11 and 007_12 at the Tnfrsf1a mutation locus. Arrowhead indicates mutated locus.

immunity via cytokine secretion, up regulation of adhesion molecules, and induction of macrophage apoptosis (Harris and Keane 2010). Clinical evidence supporting the importance of TNF-alpha comes from observations of reactivation of latent phase MTB in patients taking anti-TNF monoclonal antibodies such as infliximab for inflammatory disease such as rheumatoid arthritis or Crohn's disease (Keane et al. 2001).

3.5 Plotting variants reveals ENU genomic intervals

In the course of examining the filtered variants in these pedigrees we observed that the genomic intervals in which both alleles are inherited from the same ENU treated ancestor can be identified by plotting the homozygous variants by position on a chromosome. After filtering to remove artefacts and shared variation the density of variation in the regions inherited from the ENU ancestor should be around 1.5 mutations Mb^{-1} (Chapter 5.5), higher than the background variation from the reference genome in B6 which was observed in our datasets to be around 0.2 mutations Mb^{-1}. This increased density of variation in ENU genomic regions can be observed in the filtered data. This is illustrated for *NIH85a* in Figure 3.9. This observation shows that by using variant data across the genome we can distinguish ancestral genotypes in ENU mice.

The two sequenced *007* mice were obtained from a fixed line after multiple generations of brother–sister matings. *007_11* and *007_12* were 8 generations from the ENU treated ancestors. Plots of homozygous (Figure 3.10a) and heterozygous (Figure 3.10b) variant distribution across the genomes therefore show very similar genomic intervals inherited from the ENU ancestors, thus little additional linkage information is gained from sequencing two mice. The main benefit from the second sequenced

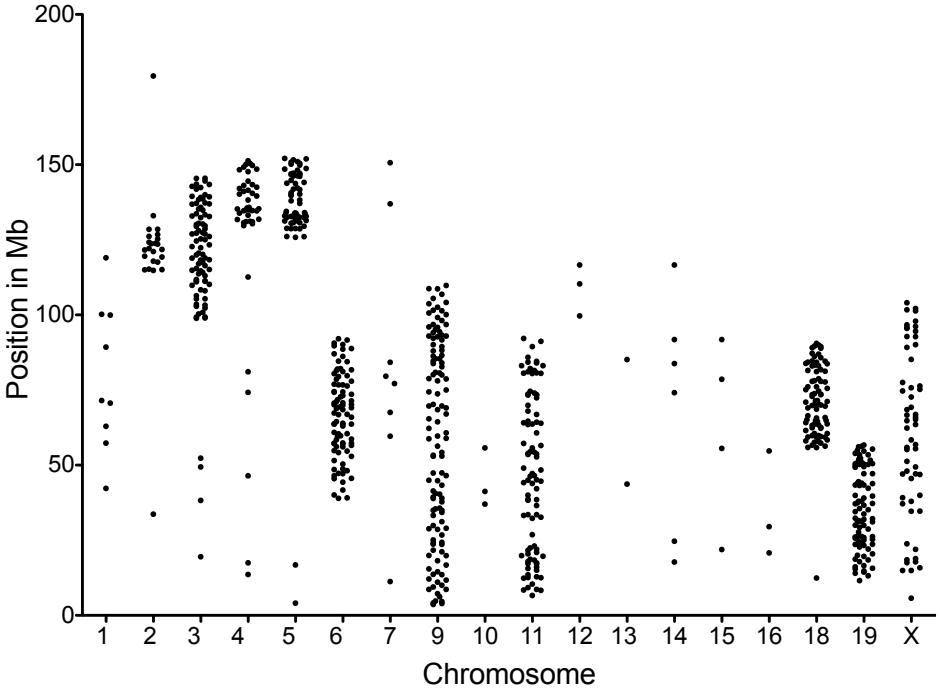


Figure 3.9: Filtered homozygous variants in NIH85a plotted by chromosomal position. Low density variants outside the ENU inherited regions are likely to represent a combination of spurious variant calls and true, non-ENU variation from the mouse reference genome.

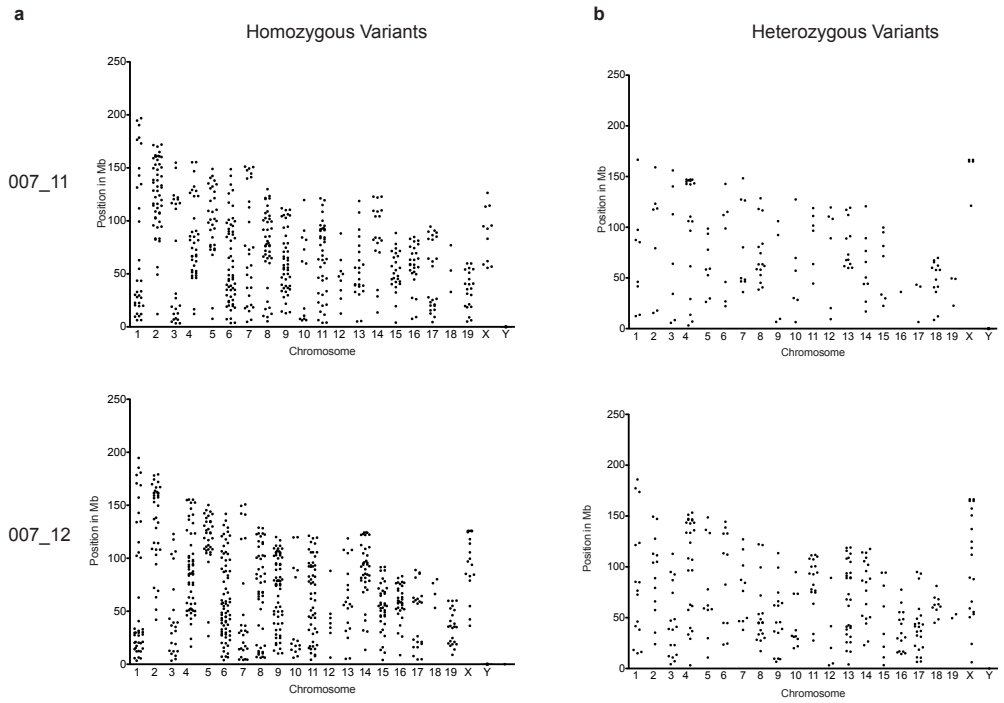


Figure 3.10: Homozygous (a) and heterozygous (b) variants by position for each chromosome in 007_11 and 007_12. Based on individual genotypes at variant loci from a filtered variant call set generated by simultaneously calling variants in both mice

individual in this pedigree is the additional coverage gained, since both 007 mice had very low mean coverage and skewing of the coverage distribution towards zero (Figure 3.1). Any distinction between homozygous and heterozygous variants in single mice, considering only information from the single variant rather than contextual information, is unreliable at these coverage depths.

3.6 Observation of a deep intronic candidate causative variant in BAFF

We used low coverage WGS in a further ENU pedigree with an immune phenotype to resolve the underlying causative variant. In this case WES had failed to identify

the variant. Strain *APFN1018* had an identical phenotype to a tumour necrosis factor (ligand) superfamily member 13b (*Tnfsf13b*) knockout mouse. *Tnfsf13b* encodes TNFSF13B, also known as B cell activating factor (BAFF), a member of the TNF family expressed on T cells and Dendritic cells and inducing proliferation and activation of B cells (Pascal Schneider 1999). Homozygous knockout BAFF mice are severely deficient in mature B cells with fewer marginal zone and follicular B cells in the spleen than WT mice (Schiemann 2001). These knockouts also exhibit decreased immunoglobulin levels and decreased T cell-dependent and T cell independent antibody responses (Mackay and Schneider 2009).

A complementation cross between the ENU mutant *APFN1018* and a known BAFF knock out mouse failed to rescue the phenotype and conventional linkage also pointed to the BAFF region, however WES did not identify any coding variant in the *Tnfsf13b* gene. Professor Chris Goodnow's group at ANU carried out the generation, phenotyping and WES of this strain.

WGS at 12.8–fold coverage, using the filtering pipeline described above, confirmed the WES findings; there were no coding variants in *Tnfsf13b*. However WGS identified 2 homozygous single bp non–coding substitutions in and near the *Tnfsf13b* gene. An A to T substitution at position 10,145,205 on chromosome 8 (GRCm38/mm10 mouse reference genome) is intergenic, downstream of BAFF by more than 10 Kilobases (Kb). However an A to G mutation at 8:10,023,621 (mm10) lies within the 4th intron, 2,192 bp from the 4th exon–4th intron boundary and 7,689 bp from the 4th intron–5th exon boundary (based on CCDS or Ensembl transcripts). A dbSNP deletion (rs2588913) includes this base, furthermore this locus is not highly conserved (the base and region are present in rat, squirrel and mouse lemur but not in other mammalian species), and so this is presumably not an important region in the wild type. However the mutation might introduce a deep intronic splice site. The non–mutated region is not predicted as a splice site by Neural Network (NNSplice) (Reese et al. 1997) but

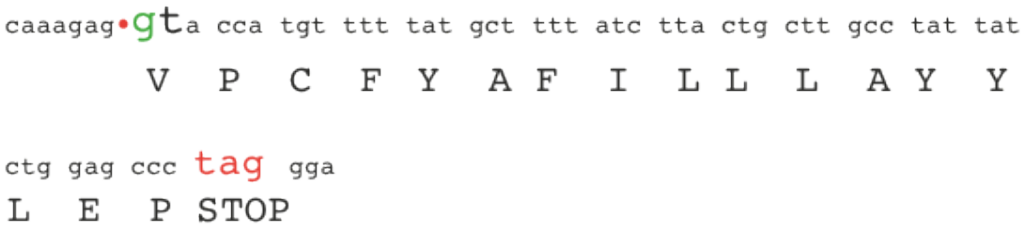


Figure 3.11: Sequence of subsection of 4th intronic region of Tnfsf13b. Bases are in lowercase and amino acids in upper case. The A to G substitution is show in green and the putative new splice site would cut at the red dot. This is followed after 17 amino acids with a TAG stop codon indicated in red.

with the A to G mutation at pos 10,023,621 (mm10) the region scores highly as a donor splice site (score 0.92). If this region did function as a donor splice site leading to 'exonisation' of the intronic region (Dehainault et al. 2007) the frame suggests that transcription would be halted by a TAG stop after 17 amino acids (Figure 3.11), potentially leading to a truncated BAFF missing the last 3 exons.

3.7 Summary of chapter

The results described in this chapter demonstrate that with sufficient filters, causative ENU mutations can be detected from low coverage WGS, including a novel variant in *Sppl2a* (Table 3.2 and Figure 3.6) (Bergmann et al. 2013). The filters most effective at removing non–ENU variation and artefacts were filters for recurrent variants observed in multiple pedigrees, and coverage depth (Figure 3.4).

The results also illustrate the limitations of sequencing single mice at very low coverage, and the *007* example shows that in later generations of inbred pedigrees there is little additional information to be gained by sequencing multiple mice (Figure 3.10).

The observation that with sufficient filtering regions inherited from the ENU an-

cestors can be discriminated by variant density (Figure 3.9 and Figure 3.10) leads to the hypothesis that sequencing multiple affected mice from an earlier generation within a pedigree would allow effective linkage analysis by elimination of regions not shared by all sequenced mice.

Chapter 4

A mutation in *Lamb2* in an ENU strain with the nephrotic syndrome models human Pierson syndrome

4.1 Introduction to chapter

4.1.1 Proteinuria

Proteinuria describes the presence of proteins in the urine, a non-specific marker of many kidney diseases. The prevalence of proteinuria increases with declining renal function (Garg et al. 2002) and proteinuria persisting for more than 3 months is a defining criteria for chronic kidney disease. It is an independent risk factor for death and end stage renal disease (ESRD), both in patients diagnosed with renal disease (Zeeuw et al. 2004; Astor et al. 2011), and the general population (Kannel et al. 1984). Pharmacological reduction of proteinuria is associated with improved outcomes for renal patients, supporting a functional role for proteinuria in the progression of kidney disease (Peterson et al. 1995; Lewis et al. 2001).

To explore accessible biological mechanisms of proteinuria with potential relevance

for the role of proteinuria in complex disease, this chapter will examine an ENU induced Mendelian model for a form of heavy proteinuria known as the nephrotic syndrome.

4.1.2 Nephrotic Syndrome

The nephrotic syndrome describes the triad of heavy proteinuria, usually defined as greater than 3.5g / 24hr in humans, hypoalbuminaemia and generalized oedema. Hyperlipidaemia can also occur.

Primary and secondary causes of nephrotic syndrome are usually described by their histology. In children primary nephrotic syndrome is most commonly due to minimal change nephropathy, whilst in adults membranous nephropathy or focal segmental glomerulosclerosis (FSGS) are the most frequent primary diagnoses. Secondary nephrotic range proteinuria can occur due systemic diseases including diabetes, obesity, malignancy, systemic lupus erythematosis, myeloma or amyloidosis.

Recent work has identified autoantibodies to the phospholipase A2 receptor in a proportion of patients with primary membranous nephropathy (Beck et al. 2009). A genetic mechanism for this is suggested by the presence of human leukocyte antigen (HLA) and phospholipase A2 receptor risk alleles which have a combined odds ratio (OR) of 78.5, and the existence of familial membranous cases (Stanescu et al. 2011). However for most of the causes of nephrotic syndrome described above the underlying mechanisms are not understood and not monogenic.

An exception to this is congenital nephrotic syndrome or steroid resistant nephrotic syndrome (SRNS) in children, which is frequently monogenic in origin (Santín et al. 2011). Nephrotic syndrome affects 2 in 100,000 children (Wong 2007), and 20% of cases fail to respond to steroid treatment (Saleem 2012). These unresponsive patients have a higher risk of progression to ESRD and a higher mortality (Barnett and Edelmann 1984; Ding et al. 2014).

Monogenic SRNS involves at least 24 known genes including nephrin (*NPHS1*) (Kestilä et al. 1998), podocin (*NPHS2*) (Boute et al. 2000), Wilms tumour 1 (*WT1*) (Barbaux et al. 1997), laminin β2 (LAMB2) (Zenker et al. 2004), CD2-associated protein (*CD2AP*) (Gigante et al. 2009), phospholipase C epsilon 1 (*PLCE1*) (Hinkes et al. 2006), actinin alpha 4 (*ACTN4*) (Kaplan et al. 2000), transient receptor potential cation channel, subfamily C, member 6 (*TRPC6*) (Winn et al. 2005) and inverted formin, FH2 and WH2 domain containing (*INF2*) (Brown et al. 2010).

The mutations causing SRNS have revealed diverse and sometimes unexpected functions ranging from ion channels to the organization of the actin cytoskeleton or mitochondrial function. In 20–40% of familial cases of childhood onset SRNS (Santín et al. 2011; Machuca, Benoit, and Antignac 2009; McCarthy et al. 2012), and at least 30% of all cases within the first year of life (Hinkes et al. 2007; Lee et al. 2011) the underlying gene is not known. Combined with the diversity of genes and pathways already identified, this suggests there are many unknown genes contributing to the filtration barrier, with potential as therapeutic targets (Brinkkoetter, Ising, and Benzing 2013).

4.1.3 The glomerulus as a filtration barrier

The genetic variants involved in SRNS are mainly expressed within podocytes or other components of the kidney glomerulus (Rood, Deegens, and Wetzels 2012). The glomerulus, consisting of capillary tufts, mesangium and Bowman's capsule, forms a semi permeable filtration barrier between the blood and the urinary space within the kidney. This barrier ensures that only small and medium weight molecular proteins, less than 69,000 daltons (Da), pass from the blood to the urine, a size selectivity that is lost in nephrotic syndrome (Deen et al. 1985). The glomerular filtration barrier is composed of three parts, the podocytes and a fenestrated endothelial cell layer, separated by the extracellular glomerular basement membrane (GBM). Nephrin

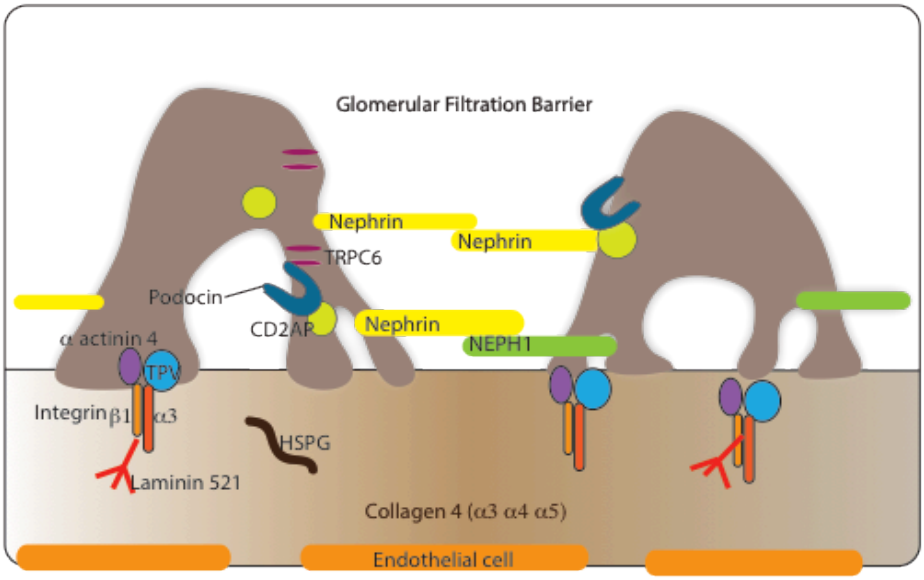


Figure 4.1: This cartoon illustrates the glomerular filtration barrier, composed of podocytes (with slit diaphragms between interdigitating podocyte foot processes), the glomerular basement membrane and endothelial cells.

and podocin form slit diaphragms between the interdigitating foot processes of the podocytes. The podocytes themselves interact with the GBM layer via integrin adhesion receptors, linking the extracellular matrix with the podocyte's actin cytoskeleton (Figure 4.1).

The genetic defects known to cause SRNS have highlighted the key role of the podocyte in maintaining the integrity of the glomerular filtration barrier (Saleem 2012). They also demonstrate the importance of the endothelial cell and the extracellular GBM, the latter composed of collagen 4, laminin, heparin sulphate proteoglycans and nidogen–1, in maintaining selective permeability (Chiang and Inagi 2010; Miner 2012). Whilst future studies using next generation sequencing will undoubtedly identify more of the genes involved in familial cases of SRNS, such families are rare, and it is therefore possible that a simultaneous and complementary approach

using animals could accelerate the discovery of new genes, and provide models for human disease. The work in this chapter explores an ENU forward genetic approach to proteinuria and reveals a novel hypomorphic *Lamb2* variant mimicking the milder spectrum of a syndromic human nephrotic disease.

4.2 Phenotyping and mapping of the *nephertiti* ENU strain

Acknowledgement: The work presented on the *nephertiti* ENU mouse in this chapter was carried out by several members of the Cornall group, including the author, and by collaborators both in Oxford and Canberra, Australia. *Nephertiti* mice were generated at the Australian Phenomics Facility, ANU, Canberra, and obtained from Chris Goodnow's group. Previous and current members of the Cornall group, including Thomas Mason, Tiphane Bouriez–Jones, Karlee Silver and Tanya Crockford, performed phenotyping and conventional mapping of the mice. Clinical chemistry of plasma samples was performed by Tertius Hough at MRC Harwell. The statistical analysis and presentation of the phenotypic and mapping data was carried out by the author. Haematoxylin and eosin staining was carried out by Thomas Mason, Tiphane Bouriez-Jones and Karlee Silver. The author performed periodic acid-Schiff staining, methenamine silver staining and immunohistochemistry, on wild type and *nephertiti* tissue. The Oxford Centre for Histopathology Research performed the electron microscopy and made further methenamine silver stains. Professor Ian Roberts (Department of Pathology, Oxford University Hospitals Trust) assisted in interpretation of all histology. All downstream analysis of WGS data for *nephertiti* including development, scripting and implementation of the hidden Markov model (HMM) approach was carried out by the author.

4.2.1 The phenotype of the *nephertiti* ENU strain

The *nephertiti* mouse strain exhibited a nephrotic phenotype. Mice examined between 17 and 25 weeks had hypoproteinaemia, hypoalbuminaemia (Figure 4.2b and Figure 4.2c), and low body weight in comparison to wild type (WT) C57BL/6J mice (Figure 4.2f). Urea and creatinine were not significantly raised in comparison to WT (Figure 4.2d and Figure 4.2e) indicating preserved renal clearance. Triglycerides (Figure 4.3b) were raised in the homozygous mutants and high-density lipoprotein (HDL) was reduced (Figure 4.3c). Cholesterol and low-density lipoprotein (LDL) showed a non-significant trend towards higher values (Figure 4.3a and Figure 4.3d). Lipid abnormalities are a prominent feature of nephrotic syndrome in humans, with triglycerides, cholesterol, and LDL typically raised and variable changes in HDL (Kronenberg et al. 2004).

Light microscopy of haematoxylin and eosin stained kidney from affected animals demonstrated protein casts within tubules and protein resorption droplets within tubular epithelium. These were not present in WT littermates (Figure 4.4a and Figure 4.4b). Periodic acid Schiff (Figure 4.4c and Figure 4.4d) and methenamine silver staining (Figure 4.4e and Figure 4.4f) revealed a coarsely thickened GBM. Silver staining also revealed widespread basement membrane spikes. Electron microscopy (EM) showed an irregular appearance to the sub-epithelial GBM with areas of thickening and spikes. The sub-endothelial GBM surface remained smooth. There was moderate effacement of the podocyte foot processes (Figure 4.5a – Figure **??**d). Basement membrane spikes are characteristic of membranous nephropathy, where spikes of GBM are visible between mainly immunoglobulin G (IgG) immune deposits. However in *nephertiti* immunofluorescence confirmed the absence of IgG antibody accumulation (Figure 4.6).

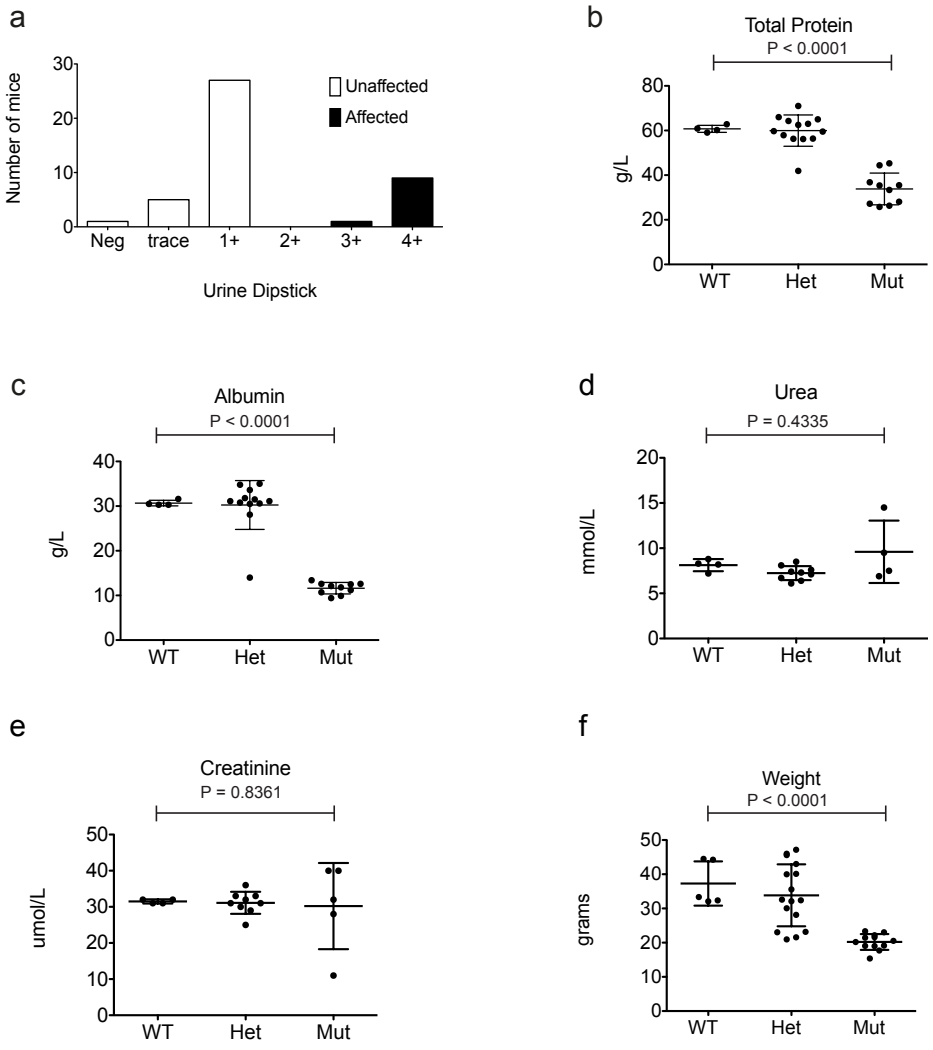


Figure 4.2: (a). Urine protein in affected nephertiti mice and unaffected littermates, where 3+ = 3–10mg/ml and 4+ > 10mg/ml. Comparison of plasma total protein (b), albumin (c), urea (d) creatinine (e) and weight (f) in age matched homozygous, heterozygous and WT sibling controls, between 17 and 25 weeks. P values are based on unpaired two tailed t-tests between WT and homozygous mutants. Error bars indicate mean and standard deviation.

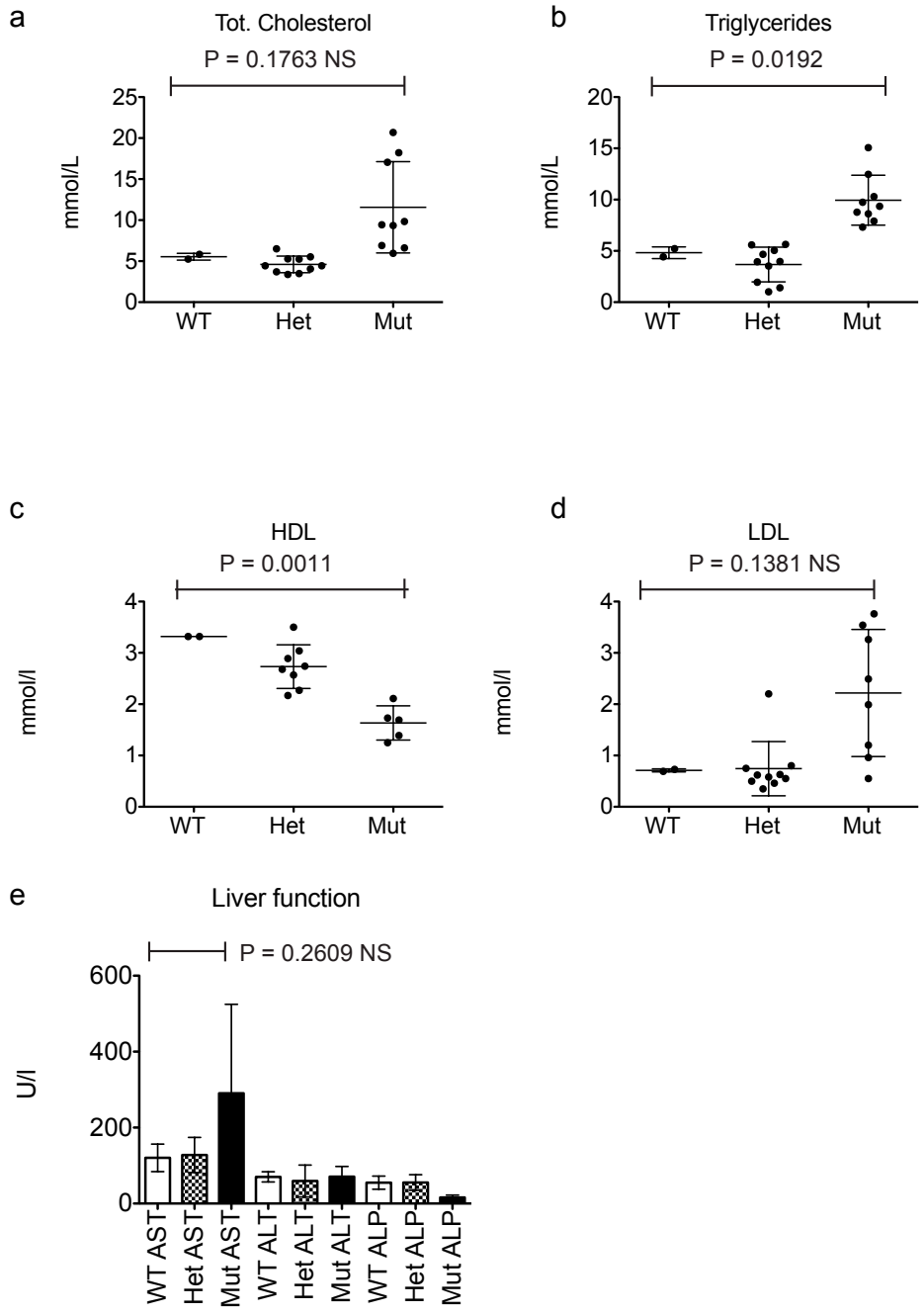


Figure 4.3: Total cholesterol (a), triglycerides (b), high-density lipoprotein (HDL) (c), low-density lipoprotein (LDL) (d), and liver function (e). Liver function is represented by aspartate transaminase (AST), alanine transaminase (ALT) and alkaline phophatase (ALP). Data shown for age matched homozygous, heterozygous and WT sibling controls, between 17 and 25 weeks. P values are based on unpaired two tailed t-tests between WT and homozygous mutants. Error bars indicate mean and standard deviation.

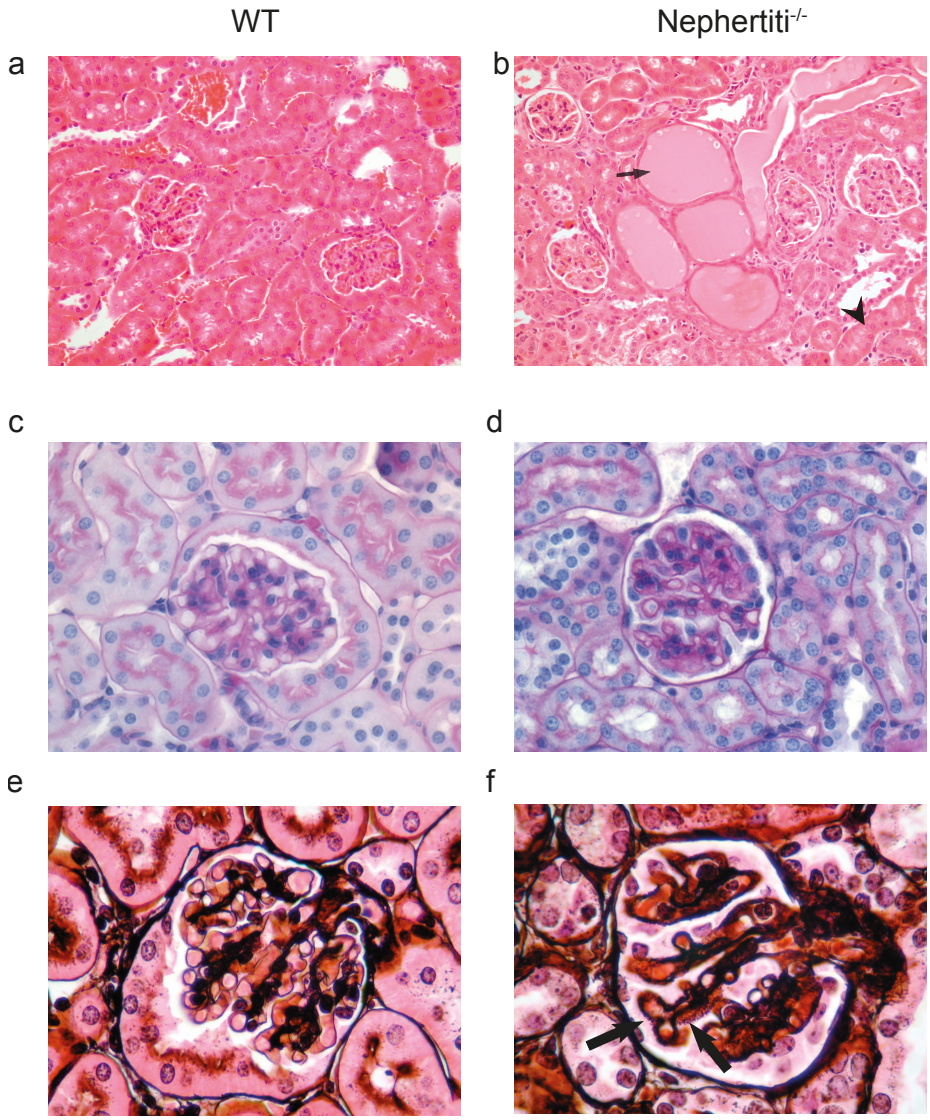


Figure 4.4: Renal light microscopy microscopy in WT (a, c and e) and nephertiti (b, d and f) mice. (a) And (b) (haematoxylin and eosin stain obj. magnification 20x). Nephertiti mice show dilated renal tubules containing protein casts (arrow) and protein resorption droplets in tubular epithelium (arrowhead). (c) And (d) Periodic acid-Schiff stain magnification 40x, showing thickened GBM in nephertiti. (e) And (f) (Methenamine silver stain obj. magnification 60x), nephertiti mice show prominent membrane spikes (arrows). All histology is shown in homozygous affected nephertiti mice and WT unaffected siblings. Histology is representative of samples from 3 affected and 3 WT for each stain.

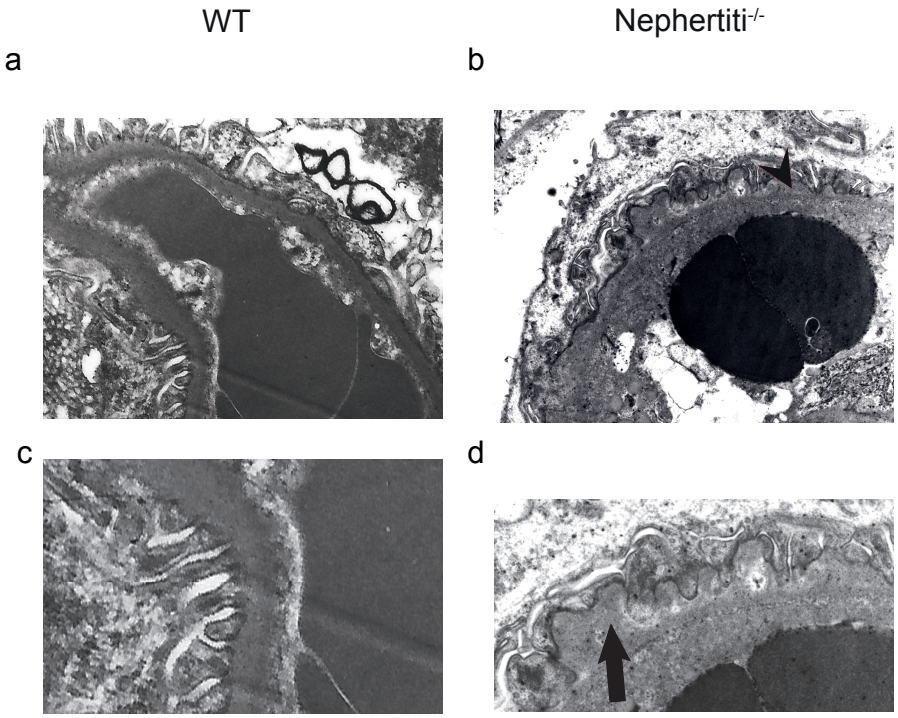


Figure 4.5: Renal electron microscopy in WT (a and c) and nephertiti (b and d) mice. (a) and (b) (electron microscopy, magnification 18,500x). Nephertiti mice show irregularly thickened glomerular basement membranes (arrowhead). (c) and (d) Higher power images of basement membrane in (a) and (b) with sub epithelial spikes and podocyte foot process effacement in nephertiti (arrow). All histology is shown in homozygous affected nephertiti mice and WT unaffected siblings. Histology is representative of samples from 3 affected and 3 WT for each stain.

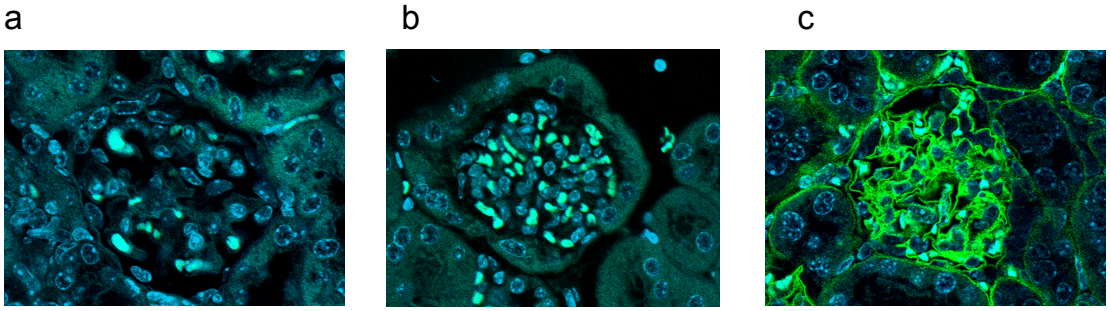


Figure 4.6: Kidney sections stained for IgG (green) and nuclei (blue). (a) Affected nephertiti mice, and (b) WT siblings do not accumulate deposits of IgG. (c) Positive control, lupus-prone roquin homozygous mutant mice exhibit IgG in glomeruli (Vinuesa et al. 2005)

4.3 Conventional linkage mapping of nephertiti

The low coverage WGS results in chapter 3 were all from ENU mice bred on a pure B6 strain background, however until now most ENU mice have been out–crossed to another inbred laboratory strain in order to carry out conventional linkage mapping using restriction fragment length polymorphisms (RFLP) or single nucleotide polymorphisms (SNPs) known to segregate between the two strains. This out–crossing can take place at the first generation of offspring from the ENU treated males, known as G_1, or more typically in a forward genetic screening program it is carried out after screening by out–crossing the G_3 mice or subsequent generations and then inter-crossing and screening offspring to track the phenotype. In a number of existing ENU pedigrees the underlying phenotypic mutation remains unknown despite a conventional linkage approach, NGS could be applied to identify these mutations, however because of the reliance on out–crossing for linkage, for many of these pedigrees only DNA from later generation mixed strain mice is available.

This was the case for the *nephertiti* pedigree. *Nephertiti* was out–crossed to the CBA strain and bred to homozygosity, tracking the phenotype by urinalysis. Coarse linkage mapping identified simple sequence length polymorphisms (SSLP) with logarithm of odds (LOD) scores (Morton 1955) of 6.84 and 2.3 on chromosome 9 (Figure 4.7a). Fine mapping narrowed the candidate region to 14.3Mb on chromosome 9 (Figure 4.7b) containing 311 RefSeq genes or 559 Ensembl genes. Despite this the causative mutation was not identified by conventional linkage methods.

The mouse reference genome is based on the widely used B6 strain, and whilst next generation sequencing has been used to generate sets of variants for many other laboratory mouse strains (Keane et al. 2011), our knowledge of variation in these strains is far from complete. Thus a key problem for identifying ENU mutations on a mixed strain background is to distinguish the ENU mutations from the large amount of variation from the reference in the non–B6 regions of the genome.

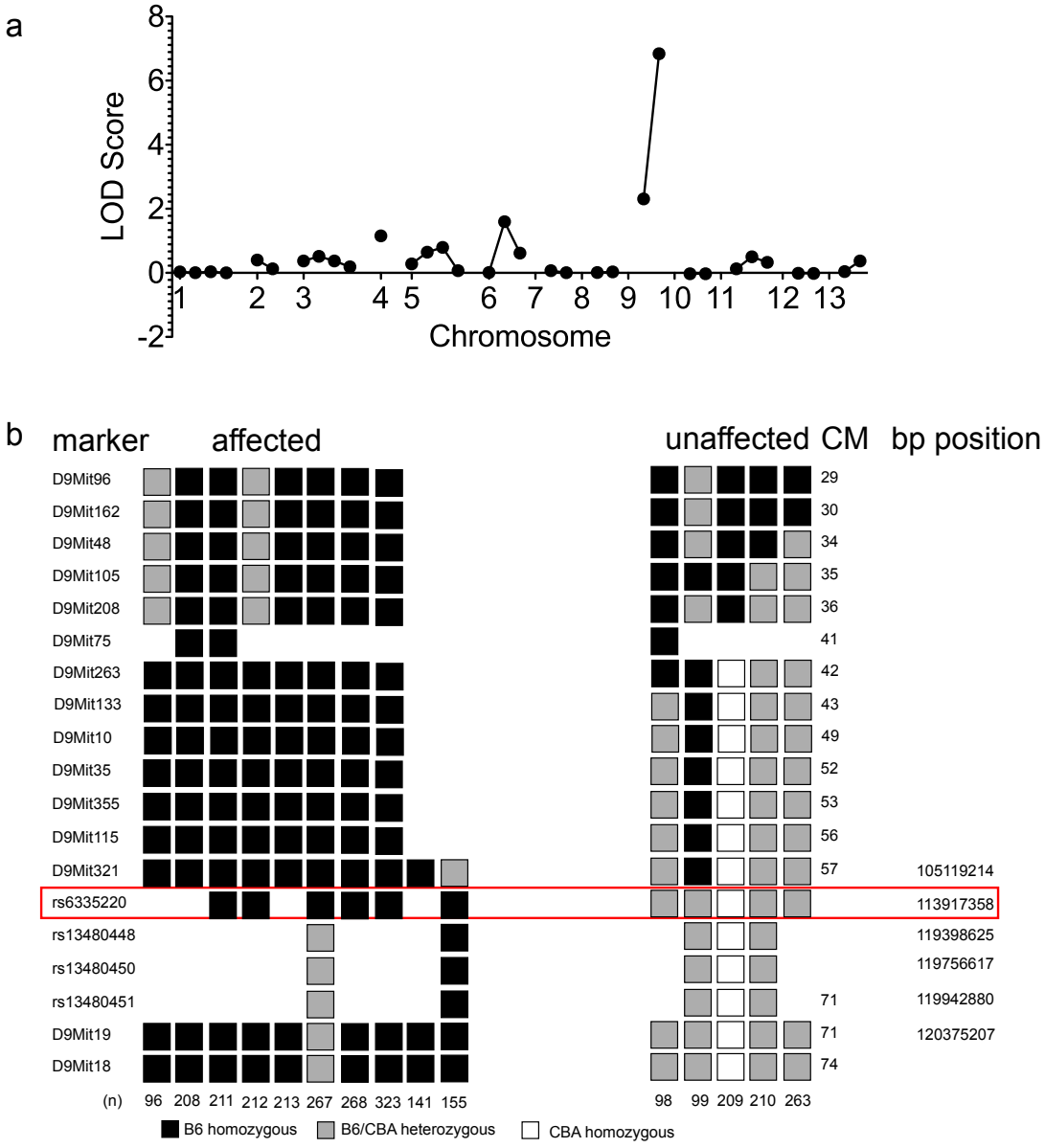


Figure 4.7: (a) Coarse mapping using SSLP. (b) Fine mapping using SSLP and SNPs. The only SNP with fully consistent linkage to the homozygous phenotype is shown boxed; the limits of the linkage region are defined by the adjacent SNPs.

4.4 *Nephertiti* WGS results

The use of a second strain introduces much larger amounts of variation from the B6 reference genome than are generated by ENU mutagenesis, and to find ENU mutations they must be distinguished from this strain specific variation.

To examine whether it is possible to identify ENU variants in such a mixed strain pedigree, and to resolve the mutation underlying this renal phenotype, WGS was performed on a single affected mouse from the *nephertiti* pedigree.

4.4.1 WGS Coverage

Low coverage WGS was performed to identify the causative mutation. DNA from one affected CBA/J B6 mixed strain *nephertiti* mouse was sequenced to mean 5.75-fold coverage across the genome, using one lane of the Illumina HiSeq machine (Figure 4.8). Based on the mutation density calculated later in this thesis, 1.5 mutations per megabase (Bull et al. 2013), a typical G_3 ENU mouse will carry 3,050 mutations. In addition to these there will be variation from the reference and spurious calls. Sequencing B6 mice at low coverage and using the same mapping and variant calling tools generated an average of $2.7\text{x}10^5$ raw variants per individual (Table 3.2). 10-fold more raw variants were observed in *nephertiti*, over $3.8\text{x}10^6$, because of greater variation from the reference due to out–crossing. However excluding known variation and filtering calls as described in chapter 3 reduced this number to 298,876.

4.4.2 Identification of a causative variant in *Lamb2* using the known linkage region

Within the 14.3Mb linkage region there were 1,680 variants, of which 8 were in coding regions or splice sites. Filtering for those affecting protein sense and present on both alleles in the affected mouse reduced this to one candidate, a G to A transition at

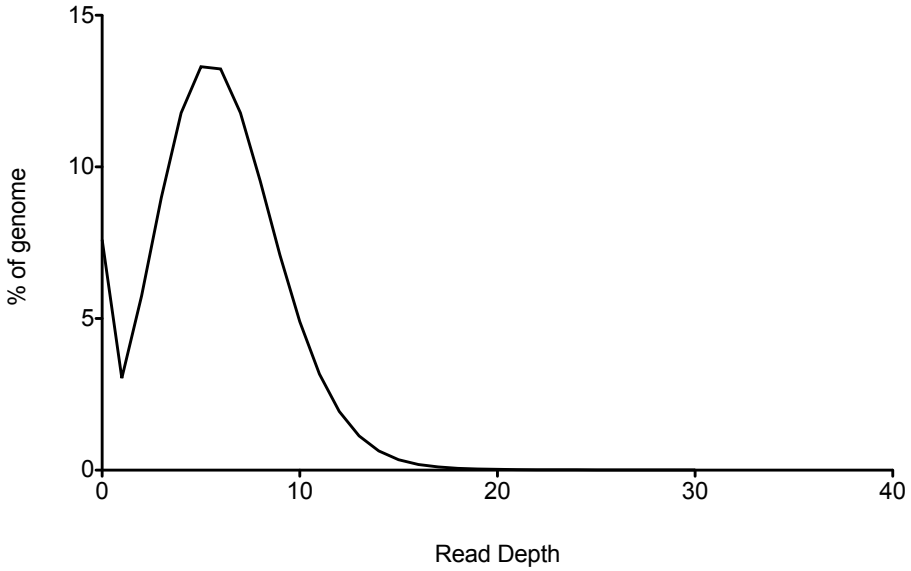


Figure 4.8: Plot of distribution of fold coverage depth, as number of reads, for nephertiti. Percentage of bases in the genome covered at each read depth, based on all mapped reads

position 108,383,650 on Chromosome 9. This was predicted to result in a C185Y amino acid substitution in exon 5 of Laminin $\beta 2$ (*Lamb2*).

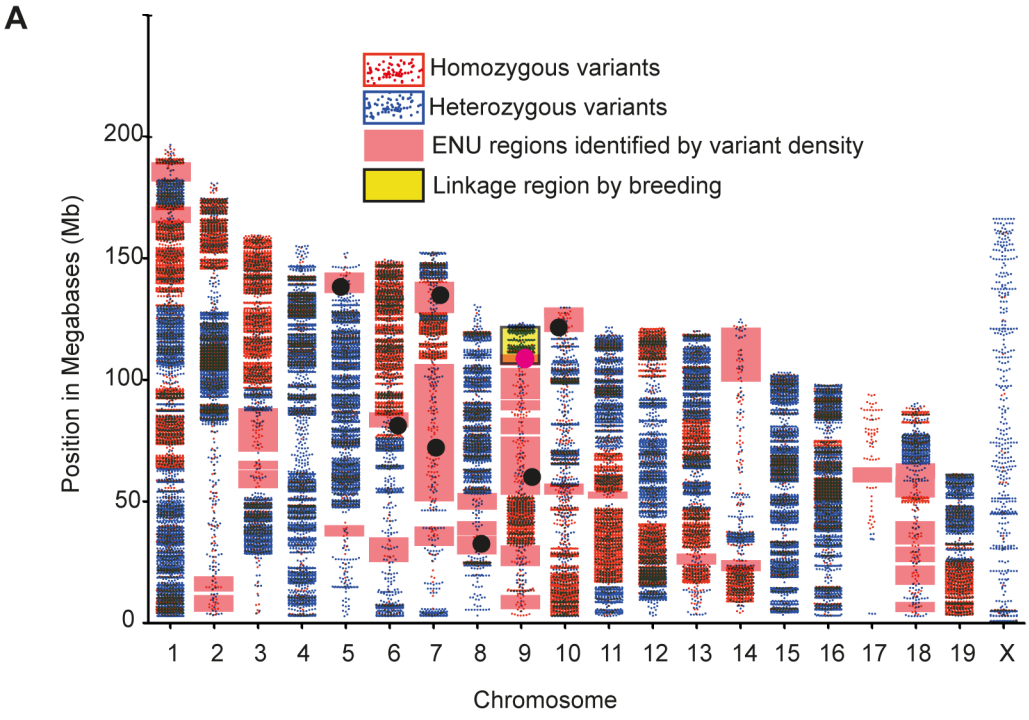
4.4.3 Isolation of candidate variants using a HMM approach

Although this result demonstrated that WGS combined with linkage data could isolate the causative mutation, it is desirable for an efficient WGS approach to be effective without utilizing conventional linkage information. To explore the efficacy of WGS in a single mixed strain mouse without using knowledge of the linkage region, an algorithm was developed to distinguish the ENU induced homozygous mutations from the large amount of homozygous variation from the reference genome in genomic regions inherited from the CBA/J ancestor by identifying all genomic intervals inherited from the B6 ENU treated founder, based on the density of variation.

Plotting the variants by genomic location across each chromosome demonstrates the densely clustered variation in CBA/J genomic intervals, and less dense variation in intervals inherited from the ENU treated B6 founder (Figure 4.9a). This contrasts

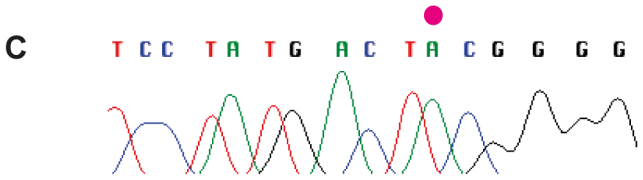
with the lower variant densities in *NIH85a* (Figure 3.9) or *007* (Figure 3.10) due to much greater density of variation in the CBA/J regions.

Based on observations in *nephertiti* and other mixed strain mice using the HMM algorithm there were approximately 260 filtered homozygous variations per Mb in homozygous out–cross strain regions. In contrast the density of homozygous variation in ENU homozygous regions was around 1.6 per Mb, consistent with later estimates, and there were 0.17 homozygous variants per Mb in wild type B6 regions.



B

	Chr	Position	Ref	Sub	Gene
●	5	139284033	C	A	Fam20c
●	6	82041963	C	G	Fam176a
●	7	73270505	G	A	Chsy1
●	7	135255783	A	G	Itgam
●	8	32346739	A	T	Fut10
●	9	58754748	T	A	Neo1
●	9	108383650	G	A	Lamb2
●	10	121277898	A	G	Srgap1



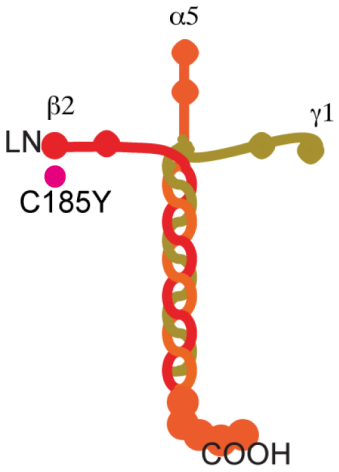


Figure 4.9 (previous page)*: (a) Scatter plots of homozygous (red) and heterozygous (blue) filtered WGS variants for each chromosome. High-density regions correspond to the CBA genomic regions, regions inheriting both alleles from the ENU treated founder have a lower density of homozygous variation than CBA regions but higher than background variation from the reference in WT genomic intervals. The region identified by conventional linkage mapping is shaded yellow. Regions identified as homozygous ENU in this mouse by variant density using the HMM algorithm are shaded pink. The yellow linkage region on chromosome 9 overlaps with a pink region identified as ENU by the algorithm. Black circles indicate candidate coding variants affecting protein sense, the Lamb2 variant is shown as a pink circle. (b) Candidate variants identified by WGS without reference to the conventional linkage region include the causative mutation. (c) Sanger sequencing confirms the Lamb2 mutation. (d) The mutation induces a C to Y substitution in the globular N-terminal domain (LN) of laminin β2, on the β2 short arm of the laminin 521 heterotrimer.*

In order to more precisely define these intervals an algorithm was developed based on a hidden Markov model (HMM). A HMM is a general framework for making inferences about unobserved states from observed data (Rabiner 1989). Here the observed data is the sequence of variant calls and their density. The underlying states correspond to the 6 possible genotypes, ENU/ENU (homozygous for an ENU treated founder), CBA/CBA (homozygous for the outcross strain), WT/WT (homozygous for the B6 reference strain), ENU/WT, ENU/CBA and CBA/WT. The algorithm is described in the Methods.

The algorithm identified regions comprising 263 Mb as having two alleles inherited from the ENU treated founder (Figure 4.9a). These regions contained 347 of the filtered variants. Selecting for homozygous variants affecting protein sense reduced the number of candidates to 8 (Figure 4.9b)

Homozygosity was determined based on the variant caller assigned genotype calls. Due to low coverage some heterozygous variants may have been incorrectly ascribed a homozygous genotype. The *Lamb2* mutation was the most likely candidate as the only homozygous mutation affecting protein sense within the linkage region, and because human *LAMB2* mutations cause nephrotic syndrome (Zenker et al. 2004).

Sanger sequencing confirmed the *Lamb2* mutation (Figure 4.9c), which is highly

conserved and predicted deleterious by PolyPhen–2 (Adzhubei et al. 2010).

To validate the accuracy of the genotyping of genomic intervals by the algorithm the density of variants in regions assigned to each genotype (CBA/CBA, CBA/WT, CBA/ENU, ENU/ENU, ENU/WT, ENU/CBA) was calculated. CBA/WT and CBA/ENU regions are indistinguishable by the algorithm, due to the overwhelming number of heterozygous CBA variants, and were therefore considered together.

Firstly the density of known CBA SNPs, not present in B6 mice (Keane et al. 2011), was considered for each set of regions (Figure 4.10a). Such SNPs should not be observed in the non–CBA regions. Since the CBA regions have a much higher density of variation from the reference (Figure 4.10b), if even a small proportion of the ENU or WT regions assigned were in fact from the CBA ancestor then the apparent proportion of CBA SNPs within these regions would rise. It can be seen that the density of CBA variants in the ENU/ENU, WT/WT and ENU/WT regions is very low (0.05, 0.08 and 0.14 SNPs/ Mb respectively) (Figure 4.10a), and assuming that the density observed in the CBA/CBA regions (4.33/Mb) is a good approximation to the true density of *known* CBA SNPs in CBA regions, it can be calculated that 1.1% of ENU/ENU regions predicted by the algorithm are miss-assigned CBA regions.

The overall density of variants in CBA/CBA regions (Figure 4.10b) is much higher than the density of known SNPs, reflecting the incomplete nature of the current SNP set for non-reference mouse strains (Keane et al. 2011).

4.4.4 A mutation in *Lamb2* mimics the milder spectrum of human Pierson syndrome

Mutations in human *LAMB2* cause Pierson Syndrome, which typically presents in childhood with severe nephrotic syndrome alongside ocular abnormalities and neuromuscular hypotonia (Zenker et al. 2004). Since *Lamb2* knock-out mice exhibit nephrotic range proteinuria (Noakes et al. 1995), both the clinical and animal data

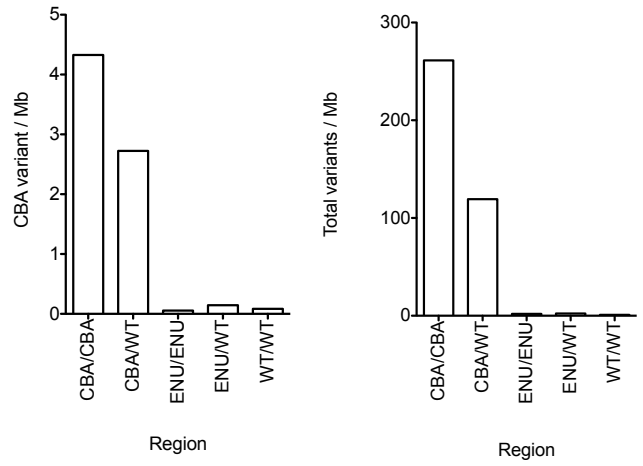


Figure 4.10: Density of CBA variants across genotype regions assigned by the algorithm. Genotype CBA/WT represents both CBA/WT and CBA/ENU regions, as the algorithm could not distinguish these. (a). Density of known CBA specific SNPs (not present in B6 mice) per Mb. (b) Density of all filtered variants per Mb by genotype of region.

suggest that this mutation is causative in *nephertiti*.

4.4.5 Fluorescence immunohistochemistry of *nephertiti* kidney

Nephertiti mimics the milder spectrum of human Pierson syndrome, in which, in contrast to null variants, miss-sense variants are associated with later onset and slower progression of kidney disease and in some cases absence of extra renal disease (Hasselbacher et al. 2006; Kagan et al. 2008; Lehnhardt et al. 2012; Mohney et al. 2011). In some of these patients expression of the laminin $\beta 2$ protein has been demonstrated (Lehnhardt et al. 2012). The underlying defect in nephertiti is a homozygous miss-sense or NS mutation; the mice have preserved renal function despite proteinuria (Figure 4.2) and survive beyond 6 months without evidence of overt neuromuscular disease. Survival, breeding, locomotor function and behaviour appear to be normal. This evidence suggested that the *nephertiti* mutation would be hypomorphic. In order to test this fluorescence immunohistochemistry (IHC) of previously prepared formalin-

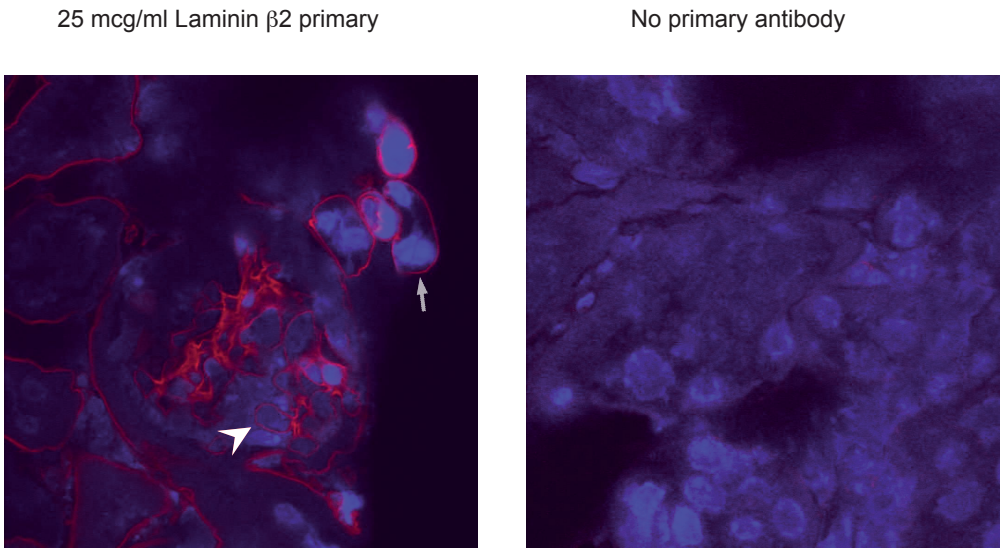


Figure 4.11: Laminin β2 shown in red in glomerular (arrowhead) and tubular (arrow) basement membranes. 4',6-diamidino-2-phenylindole (DAPI) counterstaining of nuclei shown in blue.

fixed paraffin-embedded kidney tissue was undertaken. The *nephertiti* pedigree had previously been frozen down as a sperm archive following unsuccessful conventional mapping. Therefore no fresh renal tissue could be obtained. The use of only archived formalin fixed sections was a limitation to these experiments, however this challenge led to an exploration of the most effective methods of antigen retrieval for laminin IHC in the murine kidney.

A method for fluorescence IHC for laminin $\beta 2$ was first established in fresh snap frozen WT kidney sections. This provided clear staining of the basement membranes in both the glomeruli and tubules (Figure 4.11)

However this method failed to stain laminin $\beta 2$ in freshly prepared formalin fixed paraffin embedded WT mouse kidney sections (Figure 4.12). Formalin fixation can mask antigens due to methylene cross-linking and a number antigen retrieval methods have been suggested to overcome this, with differing, antigen specific effects (Shi, Shi,

and Taylor 2011). The literature has reported variable success in applying antigen retrieval for laminin in formalin fixed tissue, with one paper finding only pepsin had any effect, allowing only marginal staining of laminin in basement membranes. No retrieval was achieved with NaCl, Trypsin, Hyaluronidase or pronase (Folkvord et al. 1989). However another study reported successful staining of laminin using a polyclonal rabbit antibody not specific to $\beta 2$ in formalin fixed kidney sections. This study found pepsin antigen retrieval effective, even after storage of paraffin blocks for 4 years at room temperature (Ekblom et al. 1982).

Therefore a range of antigen retrieval methods were tested using freshly prepared formalin fixed paraffin embedded mouse kidney sections to establish which, if any, retrieval technique would be effective to unmask laminin $\beta 2$. WT freshly prepared formalin fixed renal sections were subjected to antigen retrieval using either citrate, Tris–EDTA, hyaluronidase or pepsin antigen retrieval prior to antibody staining. Antigen retrieval with hyaluronidase (not shown), citrate or Tris–EDTA did not unmask laminin and slides were indistinguishable from negative controls (Figure 4.12).

In contrast to these negative results, pepsin did result in unmasking of the antigen. 10 min incubation with pepsin at 37°C was sufficient to unmask the laminin antigen in the freshly prepared formalin fixed slides, longer durations of exposure to pepsin at 37°C did not improve the retrieval and in fact resulted in some loss of signal, particularly at 60 minutes (Figure 4.13). Similar results with mesangial and tubular laminin staining, better with fresh slides than stored sections, were obtained using a polyclonal antibody to laminin, but the GBM was not clearly stained in any sample using the polyclonal antibody (results not shown).

The pepsin retrieval method was less effective on the stored *nephertiti* slides; laminin $\beta 2$ stained only faintly and without clear association to the basement membranes. The staining was equally poor in both *nephertiti* (Figure 4.14a) and stored WT control tissue (Figure 4.14b), indicating that this was a result of epitope degrada-

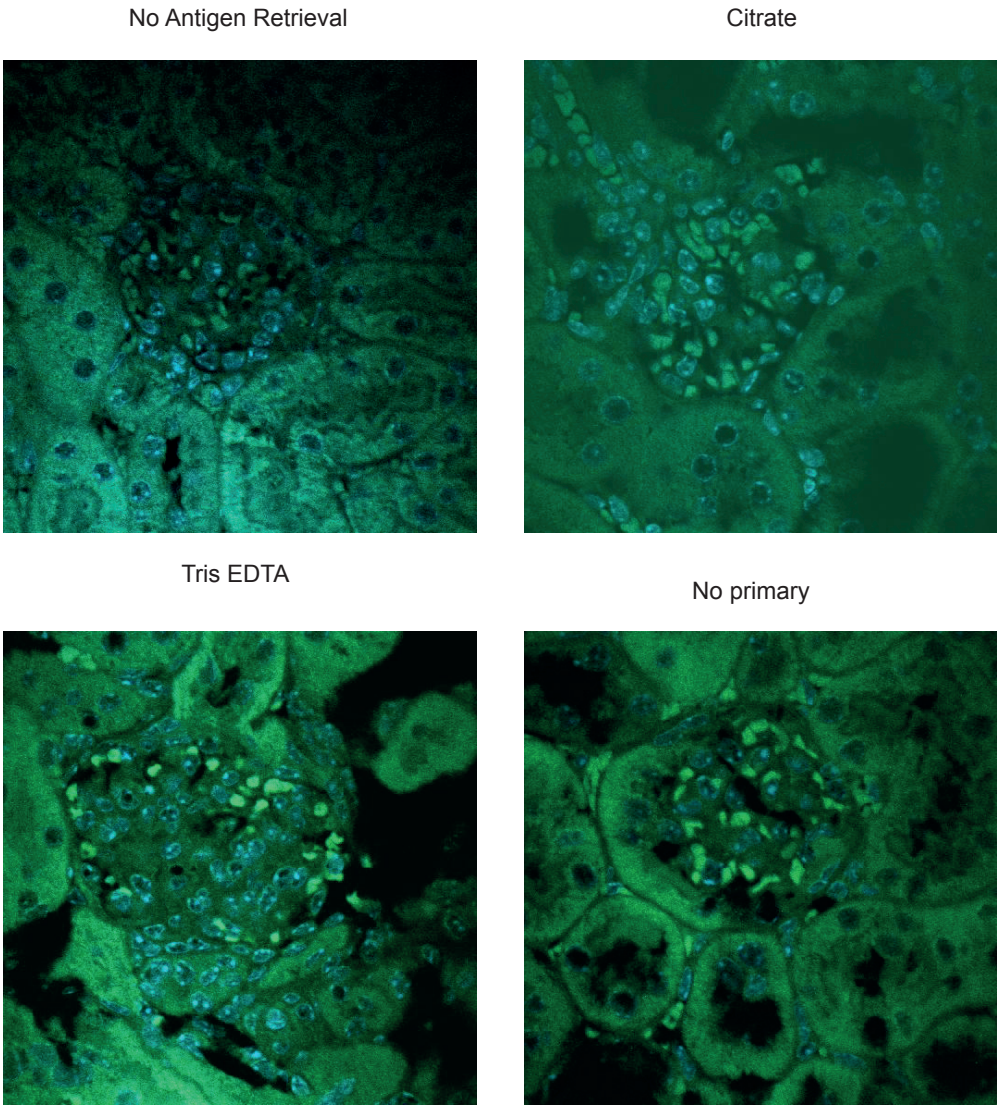


Figure 4.12: Antigen retrieval using citrate or Tris EDTA on freshly prepared formalin fixed sections. Negative controls: no antigen retrieval, no primary antibody (monoclonal rat anti mouse laminin β2). These experiments were all undertaken with a green FITC labelled secondary with DAPI counterstaining of nuclei shown in blue.

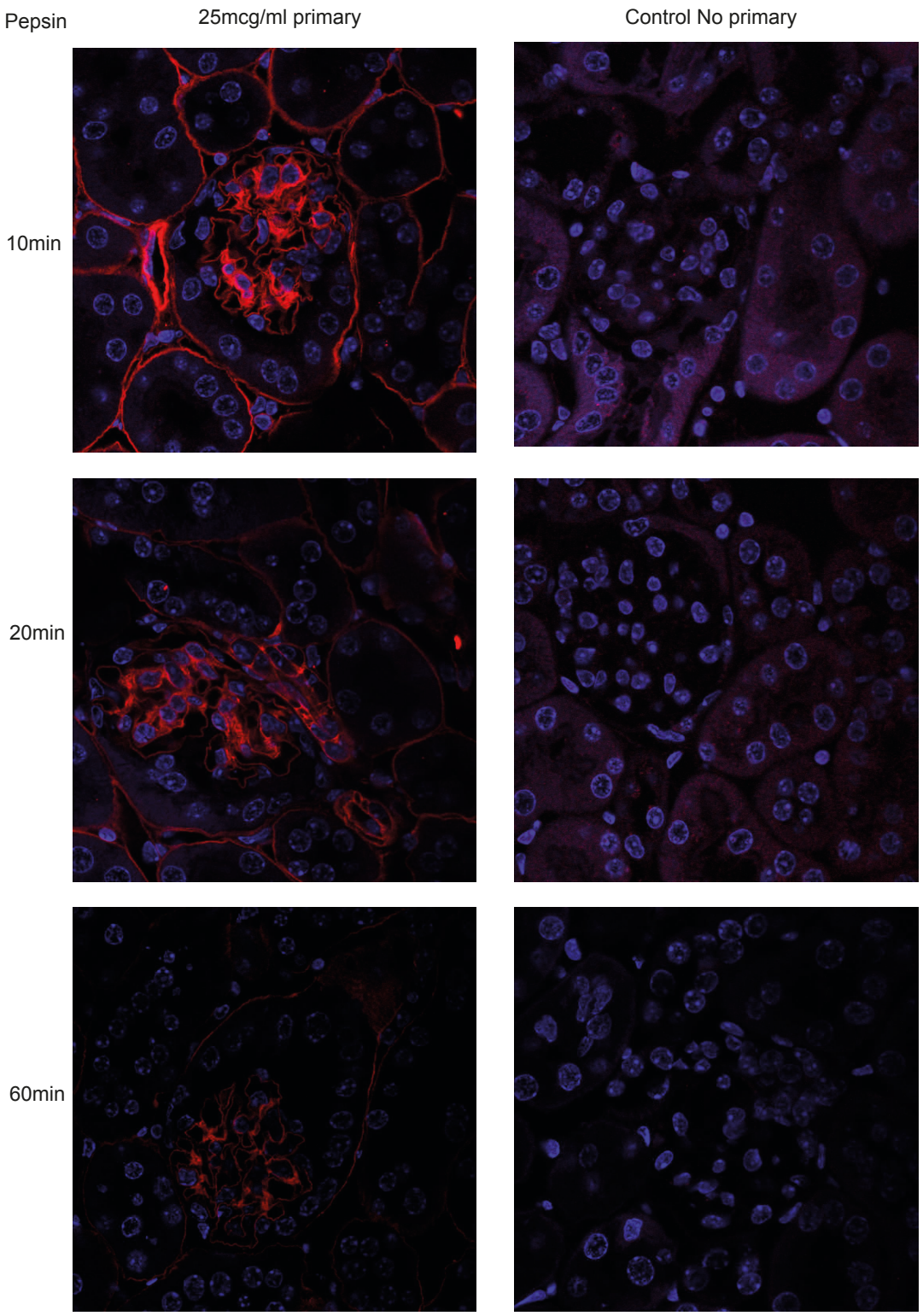


Figure 4.13: IHC of freshly prepared formalin fixed WT mouse kidney, using pepsin antigen retrieval for the durations indicated, and compared to negative controls with the same protocol excepting the primary anti Laminin β2 antibody. Laminin β2 in red and DAPI counterstaining of nuclei in blue.

tion or dissipation during storage rather than demonstrating a difference in laminin $\beta 2$ expression in the *nephertiti* mutants. Indeed the one conclusion that could be drawn is that there was comparable staining of glomeruli in both WT and nephertiti kidney sections, confirming that nephertiti is a hypomorphic mutant with expression of laminin $\beta 2$. Due to the quality of the IHC it was not possible to make more detailed study of the location of the laminin $\beta 2$ within the glomerulus with the available tissue.

4.5 Summary of chapter

This chapter has identified and characterised a novel, hypomorphic mutation in *Lamb2* in an ENU pedigree with nephrotic syndrome, providing a murine model for the milder spectrum of human Pierson syndrome. The mutation was isolated using low coverage WGS (Figure 4.8), and could be identified within a short-list of 8 candidates across the genome without use of conventional linkage data, by applying a HMM algorithm to distinguish genomic intervals based on variant density (Figure 4.9) (Bull et al. 2014). This eliminated the very large numbers of variants due to the CBA strain out-cross. Not only did this illustrate that low coverage WGS can be used to detect ENU mutations even using DNA from mixed strain mice, but the HMM method used to isolate ENU mutations in this pedigree led directly to the development of a more sophisticated dynamic programming tool for isolation of damaging ENU mutations, shown in the next chapter.

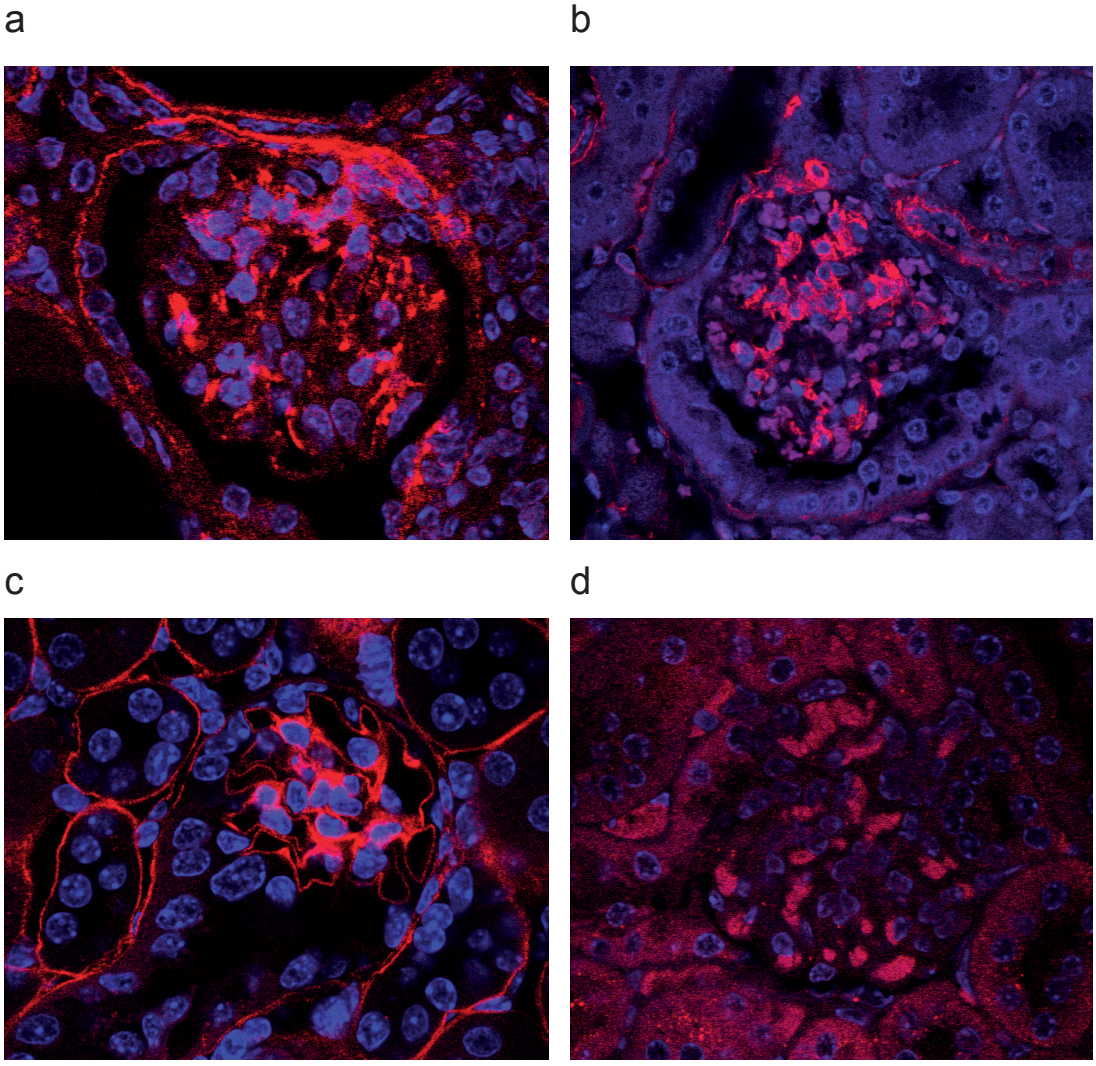


Figure 4.14: Kidney sections stained for laminin β2 (red) and nuclei (blue). Homozygous nephertiti mice (a) and WT siblings (b) express laminin β2 at the glomerulus. A fresh WT kidney section stained in parallel shows clear staining for basement membranes (c). Both WT and mutant nephertiti sections exhibit some staining in comparison with negative control without primary antibody (d). All images representative of results in 3 mice.

Chapter 5

Identity by descent analysis to isolate causative *N*–ethyl–*N*–nitrosourea mutations

5.1 Introduction to chapter

5.1.1 Identity by descent analysis

Chapter 1 and the introduction to Chapter 3 describe how the need for conventional out-breeding and linkage creates a bottleneck for ENU programmes. If NGS data could be used to directly isolate causative mutations, without the need for out–crossing, this would dramatically improve the efficiency of ENU. However, in order to do this the NGS information must be harnessed to generate and effectively filter lists of variants, identifying those due to ENU mutagenesis as shown in Chapter 3, and also to directly identify shared genomic intervals in affected mice. This in silico linkage analysis would avoid the need for out–crossing.

Two or more alleles are considered to be identical by descent (IBD) if they are identical copies of an allele inherited from a shared ancestor. This should be distin-

guished from identity by state (IBS) in which regions coincidently appear identical but are not derived from a known common ancestor. IBD is often used to infer relationships within populations, but can also be applied to identify shared regions inherited from a common ancestor. Identifying IBD regions shared by related individuals with a disease reduces the genomic search space for the causative disease variant: the smaller the IBD region, the shorter the set of candidate variants. The number of affected individuals examined, and their degrees of relatedness, will influence the size of the total IBD regions. The latter can be considered as the number of meioses that are shared between affected individuals.

An accurate approach to detecting IBD loci should consider information from multiple loci simultaneously. This avoids incorrect assignment of isolated single shared variants as IBD, which is a major potential problem with WGS data where systematic sequencing or reference errors can result in apparently shared variants. By looking at the context of the loci, putative shared variants can be checked to ensure they lie within a larger region of shared haplotypes.

Several approaches have been used to infer IBD using genetic information from SNP arrays or WES / WGS datasets. However NGS studies of Mendelian disease in humans have not always taken full advantage of sequencing data or family information to identify linkage regions directly, relying instead on genotyping SNP arrays for linkage (Sobreira et al. 2010) or filtering variants without regard to family data even in cases where DNA from family members is available (Ng et al. 2009; Züchner et al. 2011).

To identify shared IBD regions directly using NGS datasets requires an approach that accommodates error and incomplete data, as well as incorporating information from multiple loci simultaneously. A Hidden Markov Model (HMM) (Rabiner 1989) can be used to handle probabilistic information such as sequence quality, since it is based on inferring a sequence of states based on observed data with a probabilistic

relationship to the unobserved or 'hidden' true states. The hidden sequence of states is a Markov chain (Norris 1998; Eddy 2004) in which any state in the path depends only on the probabilities of transition to any state from the previous state, the observed data and the set of probabilities for any state given the observed data. Thus the most likely sequence of states, in the case of genomic data, haplotypes, is estimated using information from all loci across a chromosome.

IBD for recessive traits can be detected simply by looking for regions of homozygosity in affected individuals (Lander and Green 1987). A HMM method to infer allelic distributions in pooled populations of affected and unaffected individuals has been used to find IBD homozygous regions in out–bred ENU zebrafish, and has been applied to ENU in mice (Leshchiner et al. 2012). Similar HMM based methods can identify shared IBD recessive regions in human WES using DNA from related individuals. This can provide linkage analysis and candidate variants directly from the WES data (Chahrour et al. 2012). However simple methods looking for shared homozygous regions only work for recessive traits and fail to take advantage of knowledge of the pedigree.

An ENU–induced mutation driving a phenotype must be shared by all the affected mice and thus lie within the IBD genomic intervals, and since the ENU pedigree structure is known it is plausible that a more sophisticated HMM method which incorporates genealogy could be applied to find IBD ENU regions, both for dominant and recessive traits.

The Lander–Green algorithm is an established method for multi-locus linkage analysis that incorporates knowledge of the pedigree structure and generates probabilities for each possible flow of alleles through the genealogy (Lander and Green 1987). The algorithm was originally designed to construct genetic linkage maps in humans by calculating the most likely recombination frequencies between loci, but has been adapted for SNP array data where the genetic position of SNPs is known,

but the data is incomplete, to calculate the most likely set of haplotypes (Abecasis et al. 2002; Kruglyak et al. 1996; Gudbjartsson et al. 2005).

The Elston–Stewart algorithm is an alternative approach to calculating likelihoods for genotypes within a pedigree (Elston and Stewart 1971). However the Elston–Stewart algorithm is best applied to large pedigrees and only a few SNPs, since computational time scales exponentially for the number of loci studied. In contrast the Lander–Green algorithm scales linearly with the number of loci and exponentially with the number of individuals in the pedigree. Therefore the Lander–Green algorithm is better suited to NGS data with very large numbers of variants, but also remains computationally feasible with a pedigree of n individuals and f founders where: $2n - f \leq 25$ (Gudbjartsson et al. 2005; Abecasis et al. 2002). This suggests that an implementation of the Lander–Green algorithm could be applied to 3–generation ENU pedigrees.

Variations on the Lander–Green algorithm, developed for array data, can incorporate knowledge of the population allele frequencies of HapMap SNPs (Abecasis et al. 2002), and such methods have been applied successfully to WES data. It is also possible to incorporate local recombination rates in a non-homogeneous HMM (Rödelsperger et al. 2011; Smith et al. 2011; Guergueltcheva et al. 2012).

However such an approach has not been shown using WES or WGS for autosomal dominant traits in humans and until now no method has been demonstrated in ENU mice that eliminates the need for out–crossing. All published ENU mutants to date have been identified using some conventional out–crossing and linkage, including those identified using NGS (Sun et al. 2012b; Fairfield et al. 2011; Leshchiner et al. 2012; Arnold et al. 2011; Sheridan et al. 2011; Boles et al. 2009).

In comparison with human data, the low density of ENU variants presents challenges for the interpretation of WES or WGS, both in distinguishing true ENU variants from background variation from the reference, and in identifying enough variants

to provide adequately fine scale linkage. Furthermore the ENU mutations are de facto novel and thus no validation of true mutations from other datasets can be performed. An implementation of the Lander–Green algorithm for ENU data would require modifications to accommodate these differences.

5.1.2 The genomic density of ENU induced mutation

In order to estimate of the efficiency of an ENU programme in generating desired mutations, and to predict which genomic regions or forms of mutation may be underrepresented, it is necessary to understand the density, characteristics and biases of mutations induced by ENU.

The density of ENU mutations is dose related (Favor et al. 1990; Lewis et al. 2009) and may differ according to the mouse strain (Justice et al. 2000). Estimates of ENU mutation density in the literature vary widely from 0.5 Mb^{-1} to 10 Mb^{-1}. These estimates have been hampered by low sample size and ascertainment bias. Initial estimates based on the specific locus test (Russell et al. 1979) were inaccurate because of uncertainty about the locus size and the fraction of mutations within a locus informative for the allelism tests. When sampling from ENU offspring within a pedigree there may be underestimation of the rate of mutation due to sampling from genomic regions where both haplotypes are inherited from the WT female ancestor (Figure 5.1). To overcome this several studies have used denaturing high performance liquid chromatography (DHPLC) to examine mutations in first generation (G_1) mice generated from an ENU founder and WT mouse, which are heterozygous at all positions for an allele from the ENU treated male. One study examined five loci, an effective region of 0.37 Mb in 192 G_1 mice, and identified 6 mutations, this suggests a rate as high as 2 mutations per 10^5 bp. However the small number of variants detected makes it difficult to draw conclusions from this (Beier 2000). A DHPLC study on a BALB/c background identified 5 mutations in 9.48 Mb of sequence from

6,500 G1 mice, giving a mutation rate of 0.5 mutations per Mb (Coghill et al. 2002). Another DHPLC study estimated a mutation rate of 1 per 1.01 Mb based on 27.4 Mb of DNA (Quwailid et al. 2004). Strain specific effects and the inaccuracy of mutation detection by DHPLC limit the interpretation of these results. A study of ENU in B6 mice treated with two doses of 85–100mg/kg ENU, using Temperature Gradient Capillary Electrophoresis (TGCE), detected 130 point mutations in 181 Mb of sequence. On the basis of an 50% detection rate with TGCE, the investigators estimated a true mutation rate of 1.4 x 10^{-6} per base pair or 1.4 Mb^{-1} (Takahasi, Sakuraba, and Gondo 2007). Another TGCE based study found a similar density of 148 mutations in 197 Mb of sequence; if the 50% detection rate estimate is applied this would suggest a rate of 1.5 Mb^{-1} (Sakuraba et al. 2005). Sanger sequencing provides a more accurate method to detect mutations. A study using the B6 strain, doses of 75–100mg/kg at weekly intervals and Sanger sequencing identified 10 mutations in 51 genes covering 9.6 Mb, giving an estimate of 0.96 mutations Mb^{-1}. However this estimate had a wide 95% confidence interval of 0.52–2.0 Mb^{-1} (Concepcion et al. 2004).

5.2 Scatter plots depict IBD

With effective filters on called variants from WGS, variant density can be used to identify genomic intervals inherited from the ENU treated ancestors (Figure 3.9 and Figure 3.10). This observation led to the hypothesis that a variant density based method applied to WGS data from multiple affected mice within a pedigree would allow isolation of the causative ENU mutation within a shared haplotype block, using IBD to perform genetic linkage.

In a typical strategy for generating and screening ENU mutant mice, B6 ENU-treated founders are bred with B6 females to generate G_1 founders, establishing pedigrees in which pairs of G_2 mice produce G_3 mice segregating recessive and dominant

mutations (Figure 5.1a). To assess the utility of WGS and an IBD method in the analysis of such mice, a pedigree was chosen, identified as *ENU16CH17a*, where the recessive phenotype (identified by Prof Goodnow's group at ANU) was peripheral B cell lymphopaenia (Figure 5.1b and Figure 5.2). WGS was performed on three affected G_3 mice from a single G_2 pair to high coverage (average 24–fold mean read depth per individual).

The causative variant in *ENU16CH17a* will belong to a haplotype that is shared by all the sequenced mice and inherited from an ENU–treated ancestor. By constructing chromosomal maps of the homozygous variants in all three animals we could demonstrate clustering of mutations within haplotype blocks inherited from ENU–treated founders (Figure 5.1c), as previously shown in Chapter 3 for individual mice. By plotting the homozygous variants shared by the 3 affected, sequenced mice we could rapidly identify a linkage region on chromosome 4 (Figure 5.1d).

5.3 IBD using an implementation of the Lander–Green algorithm

Chapter 4 used the out–crossed *nephertiti* mouse to show that a HMM based method can efficiently identify ENU regions, based on the density of the filtered variants; this result suggested that a HMM based algorithm could also be applied to IBD analysis on multiple mice. A method was developed based on the Lander–Green algorithm, which uses genetic markers, knowledge of the pedigree, and recombination rates to infer the flow of alleles through the genealogy (Lander and Green 1987). The implementation developed uses probabilistic variant calls to identify haplotypes from the four founder mice (ENU1, ENU2, WT1 and WT2) across the genome of each G_3 individual (Figure 5.1a and Methods).

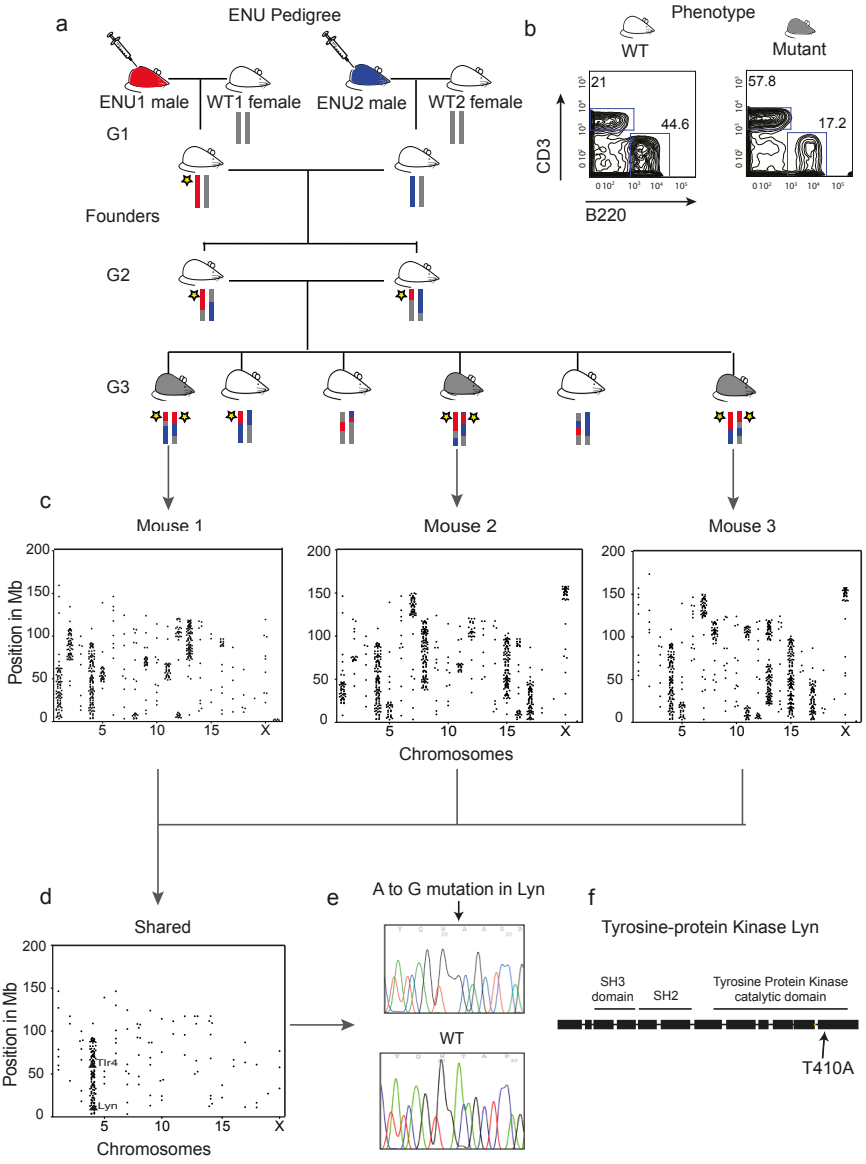


Figure 5.1: Whole genome sequencing identifies the IBD homozygous region and causative ENU mutation in ENU16CH17a. (a) The structure of an ENU pedigree: two ENU treated males paired with WT B6 females generate founder G_1 mice for the ENU16CH17a pedigree, and G_3 mice exhibiting the phenotype are selected for WGS. Thus mice within the pedigree carry 4 possible haplotypes, ENU1, ENU2, WT1 and WT2. A yellow star illustrates the segregation of a causative variant. (b) Mice homozygous for the mutation exhibit B cell lymphopaenia (here gating on blood lymphocytes). (c) Plots of homozygous filtered variants show the haplotype blocks across the chromosomes of each sequenced mouse. (d) Shared homozygous variants seen in all 3 sequenced mice cluster in an IBD region on Chromosome 4, containing exonic mutations in two genes, Lyn and Tlr4. (e) Confirmation of the Lyn A to G transition by Sanger sequencing. (f) The mutation lies in exon 12 within the catalytic domain.

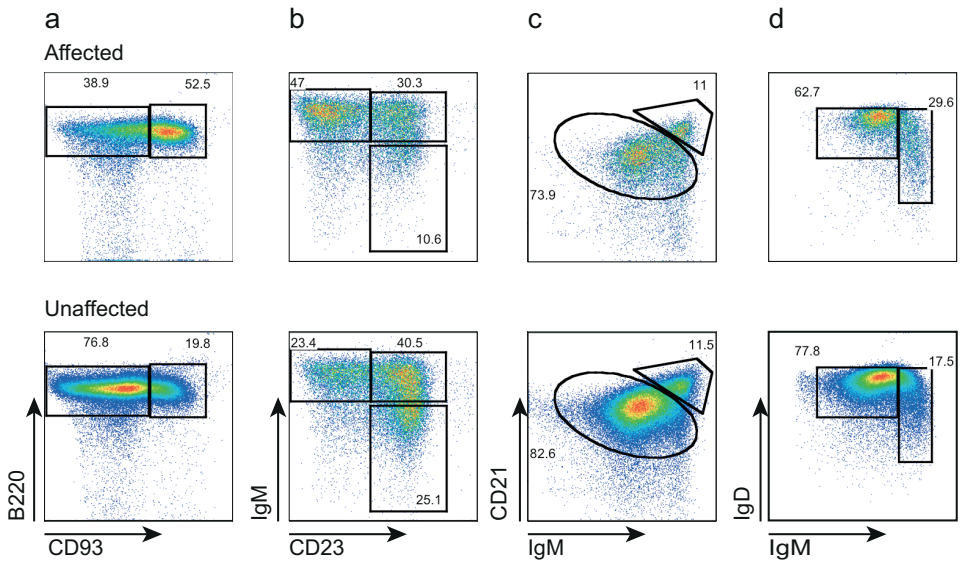


Figure 5.2: Splenic B cell populations in ENU16CH17a. (a) B220 and CD93 within the $CD45^+$ $CD19^+$ *B cell population. (b) Immature B cells gated on* $CD45^+$ $CD19^+$ $CD93^+$*, indicating the T1 (*IgM^{hi} $CD23^-$*), T2 (*IgM^{hi} $CD23^+$*) and T3 (*IgM^{lo} $CD23^-$*) populations. (c) Mature B Cells gated on* $CD45^+$ $CD19^+$ $CD93^-$*. Follicular (*IgM^{lo} $CD21^{hi}$*) and marginal zones (*IgM^{hi} $CD21^{hi}$*). (d) Follicular B cells gated on* $CD45^+$ $CD19^+$ $CD93^-$ IgM^{lo} $CD21^{hi}$ *showing* IgM *vs* IgD*, expression of* IgM *falls and* IgD *increases with maturity of follicular B cells.*

5.3.1 Modifications to the Lander-Green method specifically for ENU

In order to apply the Lander–Green algorithm to WGS data from an inbred mouse with ENU mutations, a number of modifications were made to the standard algorithm. The Lander–Green method was developed to handle fairly reliable data based on known population variants, originally RFLPs, and subsequently denser SNP maps (Abecasis et al. 2002). In the ENU / WGS implementation of the Lander–Green algorithm developed here genotype likelihoods, rather than simple presence or absence of SNPs, were incorporated to accommodate the uncertainty inherent in the WGS variant calls, compared to datasets of known variation. The genotype likelihoods computed by the variant caller for each possible variant site genotype were used, together with prior information about the probability of observing a variant (taken to be 2 Mb^{-1} in ENU regions and 0.2 Mb^{-1} for WT) to compute a probability for the observed variant genotypes across all 3 mice. This was calculated for each possible inheritance pattern of ancestral alleles (represented by the inheritance vectors), in 100kb windows across each chromosome (Methods). The predicted IBD regions were relatively insensitive to changes in the ENU mutation density prior (Figure 5.3).

Because the genotype likelihoods are used, rather than simple presence or absence of a variant, the algorithm does not absolutely exclude any ancestral state vector, though many will have extremely low probabilities. Consequently this approach does not permit reduction in the number of possible state vectors at any window below the 2^{2n} state vectors possible given n non-founder mice within the pedigree. The algorithm could be further modified to remove vectors with probabilities close to zero, to improve computing times for larger pedigrees.

A second modification uses a mouse recombination map (Cox et al. 2009) to compute average local recombination rates across the genome and incorporates this into the calculation of transition probability between ancestral state vectors between

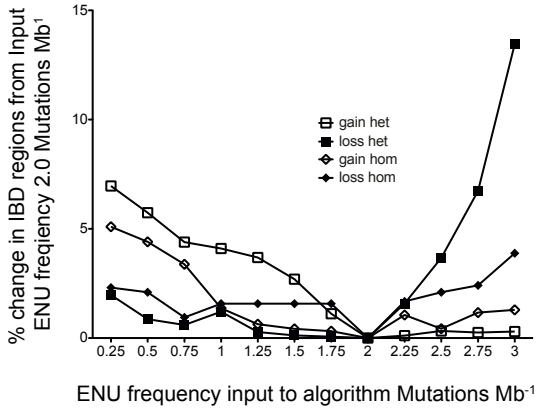


Figure 5.3: Effect of input ENU mutation density on the assigned homozygous and heterozygous regions. All regions were compared to those generated using an input rate of 2.0 Mb^{-1}. Loss of regions is % loss compared to the regions at 2.0 Mb^{-1}. Gain is % of regions assigned using each input that were inconsistent with those seen at 2.0 Mb^{-1} input.

windows. This provides correction for the presence of recombination hotspots (Smagulova et al. 2011) within which the probability of a transition to a different ancestral inheritance pattern due to recombination is higher.

Thirdly the whole transition matrix in not stored in memory, but instead matrix elements are computed on demand. This is straightforward, as all matrix entries can be expressed as powers of the recombination rate and one minus the recombination rate. This vastly reduces the memory requirements for the algorithm, which are now linear in the number of state vectors rather than quadratic.

5.4 Results from the Lander-Green based algorithm in the *ENU16CH17a* test case

The haplotypes assigned by the algorithm identify the IBD regions (Figure 5.4b). IBD homozygous regions comprise 95.3Mb (3.6% of the genome), containing 137 variants, including only two mutations in coding regions, both on chromosome 4 (Figure 5.1d).

One mutation at position 3,710,143 is an A to G mutation inducing a miss–sense change in the Src–kinase encoding gene Yamaguchi sarcoma viral (v–yes–1) oncogene homolog (*Lyn*) (Figure 5.1e). The mutation corresponds to a threonine to alanine substitution at amino acid residue 410 in exon 12 within the highly conserved Src activation loop in the protein kinase domain (Figure 5.1f).

The phenotype seen in *ENU16CH17a* has been described in ENU treated mice carrying a threonine to lysine substitution at the same codon in *Lyn* (Verhagen et al. 2009), indicating that the LYN^{T410A} mutation is causative. The other mutation, at position 66,590,107, encodes a miss-sense mutation in a single transcript of Toll-like receptor 4 (*Tlr4*) reported by Ensembl (Tlr4-004 ENSMUST00000107365), but absent from the RefSeq dataset where *Tlr4* terminates at 66,502,513. TLR4 deficient mice do not have obvious defects in B cell development (Hoshino et al. 1999).

5.4.1 Dominant mutations

The *ENU16CH17a* pedigree carries a recessive functional mutation; however, to demonstrate the wider application of our method for dominant traits, the shared IBD heterozygous ENU mutations in the same G_3 mice were examined. The 3 *ENU16CH17a* mice share one or more haplotypes from a common ENU founder across regions comprising 40.8% of the genome (1,083.8 Mb) (Figure 5.4b), containing 26 heterozygous candidate mutations shared by all 3 mice, comprising 25 miss-sense and one splicing mutation; there were no non-sense mutations. PolyPhen-2 Adzhubei:2010hi predicted 9 as benign, leaving 17 heterozygous shared mutations with possibly deleterious effects. Sanger sequencing confirmed the presence of all 28 homozygous and heterozygous IBD variants (Appendix A, Appendix B and Appendix C).

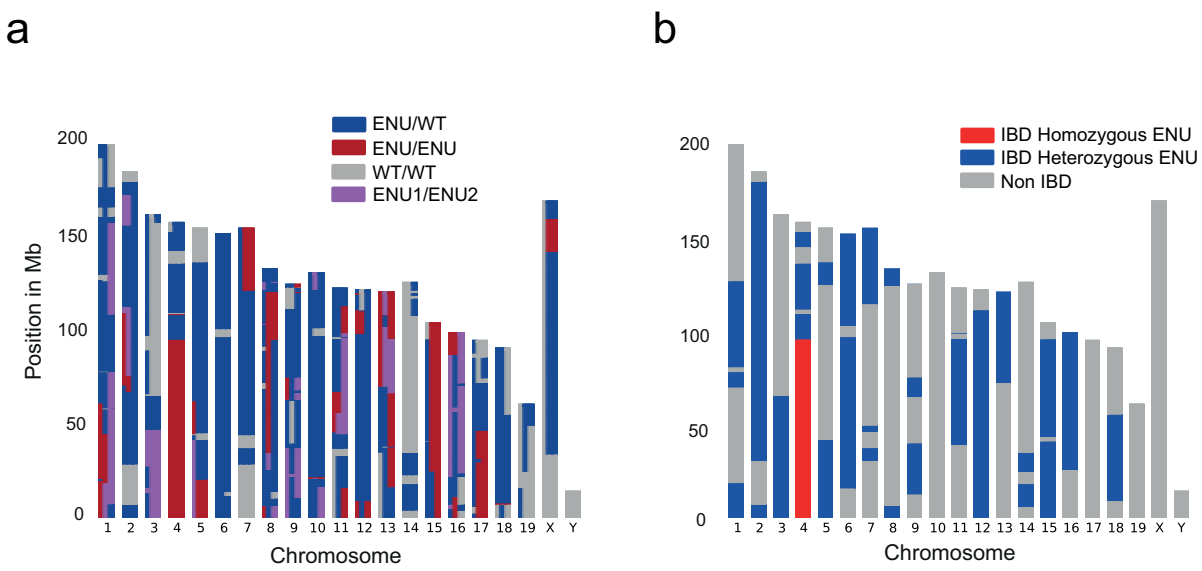


Figure 5.4: Identification of IBD regions in ENU16CH17a using a modified Lander-Green algorithm. (a) Graphical representation of the output of the algorithm, showing the genotypes for the 3 mice, based on combinations of the 4 haplotypes ENU1, ENU2, WT1 and WT2 inherited from the founder mice. WT1 and WT2 are genetically indistinguishable. Each mouse is represented by a vertical third of the plot for each chromosome, and colour blocks represent un–phased haplotype combinations for each mouse as indicated in the figure. ENU/ENU indicates homozygous ENU regions and ENU/WT indicates heterozygous regions for ENU1 or ENU2. (b) Graphical representation of the chromosomal IBD regions, showing shared heterozygous (blue) and homozygous (red) IBD regions. Regions are only IBD if all mice share alleles from a particular ENU founder, ENU1 or ENU2. Non–homozygous IBD regions in which all mice carry at least one matching ENU allele are considered IBD heterozygous. All regions are defined by the states with the highest posterior probabilities calculated in the forward backward step in the algorithm (Methods).

5.4.2 Effect of modelling lower coverage depths

For a high throughput ENU programme, an analysis based on IBD would ideally find mutations and IBD regions accurately even at low coverage per individual. To model this, a simulated lower coverage dataset was generated by randomly selecting subsets of *ENU16CH17a* reads. This was used to check the consistency of variant calls at different simulated levels of coverage compared to 24–fold per mouse. The assignment of IBD regions remained highly consistent with the complete dataset down to very low coverage levels. At 5–fold coverage per mouse, 93% of homozygous and 91% of heterozygous IBD regions seen at 24-fold were assigned (Figure 5.5a). 83% of homozygous and 77% of heterozygous variants in IBD regions overlapped with those found in IBD regions at full coverage. Within the validated set of non-synonymous coding and splice site mutations, all the homozygous IBD variants were identified (2/2), and 69% of heterozygous IBD variants (18/26) at 5–fold coverage (Figure 5.5b).

At low coverage depths false positive candidates could also accumulate due to mis-assignment of regions and miscalling of variants. As the coverage is reduced some regions are assigned as IBD that were not considered IBD at full coverage. For example 6.4% of regions ascribed as homozygous IBD and 2.5% of regions ascribed as heterozygous IBD at 5–fold coverage were absent at 24–fold coverage (Figure 5.5c). This resulted in 2.2% additional homozygous IBD variants and 1.07% additional heterozygous variants at 5-fold coverage compared to 24–fold, with no additional validated coding candidates (Figure 5.5d). These results indicate that the accumulation of false positive candidates is low at reduced coverage levels.

5.4.3 Experimental evidence for the accuracy of low coverage IBD analysis

To confirm the utility of the method at low coverage experimentally, Sanger sequencing was performed to validate shared variants from 3 affected mice in a second ENU pedigree, which had been sequenced by WGS at 4.37-fold mean coverage per mouse (range 4.05 to 4.56) (Figure 5.6a). This pedigree was phenotyped by Owen Siggs in Bruce Beutler's group at The Scripps Research Institute, California. Again the causative recessive mutation was identified (unpublished), and true-positive rates of 85.7% (24/28) for homozygous variants and 86% (48/56) for heterozygous variants were found. Within the subset of coding variants, 100% (4/4) of the homozygous variants identified were validated and 96% (24/25) of the heterozygous variants (Appendix D.1 and Figure 5.6b). As expected, Sanger sequencing of variants from non-IBD regions revealed lower true-positive rates: 57% (49/86) of called variants were confirmed by Sanger sequencing, comprising 59% (10/17) of coding variants and 57% (39/69) of non–coding variants (Appendix D.2 and Figure 5.6b), demonstrating greater accuracy of variant calling in IBD regions compared to non-IBD regions even at 4–fold coverage per mouse.

5.4.4 Comparison with a simple 'shared variant' approach

The described IBD approach was compared to a simple method of selecting all variants shared across all 3 affected mice (Methods). At lower coverage levels the simple approach identified very large numbers of homozygous shared variants compared to the IBD method; at 5–fold coverage there were 586 shared variants compared to 158 by IBD (Figure 5.5e). This error is likely to be due to miscalling of heterozygous variants as homozygous, coupled with accumulation of further shared heterozygous variants, since the overall number of heterozygous shared variants at each simulated depth is

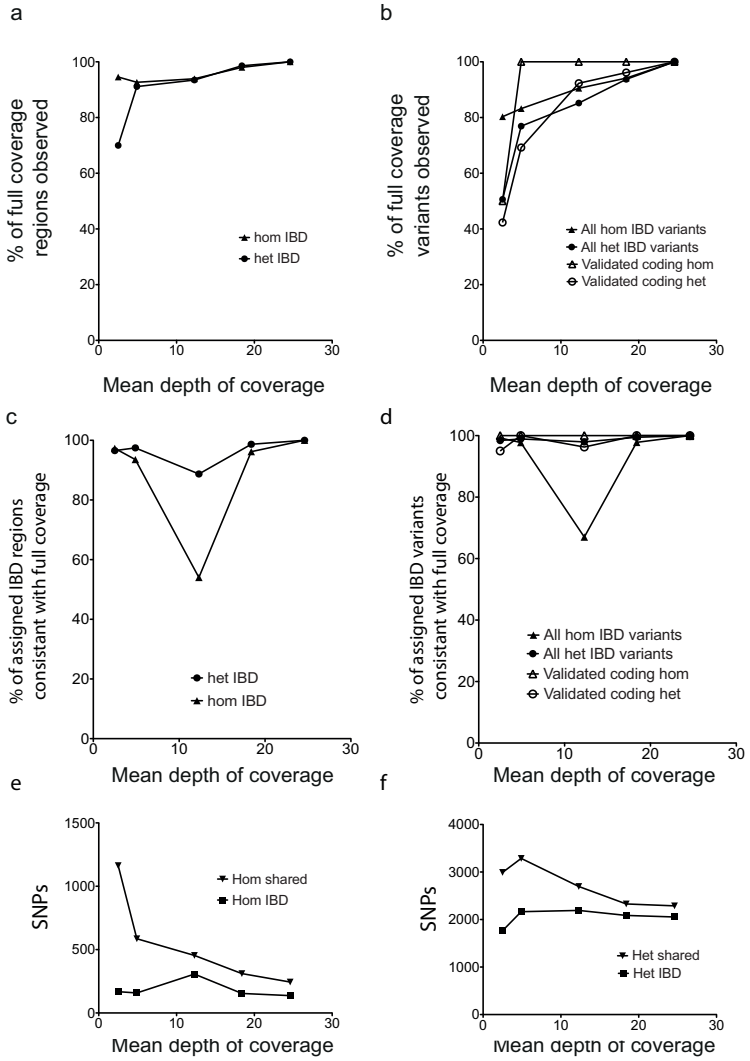


Figure 5.5: (a) The proportion of ENU homozygous and heterozygous IBD regions from the full 24–fold coverage dataset identified at simulated lower depths of coverage per mouse. (b) The proportion of homozygous and heterozygous IBD variants from the 24–fold coverage dataset identified at simulated lower depths of coverage per mouse. The validated variants are the coding or splice variants confirmed by Sanger sequencing (Appendix A). (c) The proportion of IBD regions assigned at each coverage depth that are also observed at 24–fold coverage. (d) The proportion of IBD variants assigned at each coverage depth that are also observed at 24–fold coverage. (e) The number of IBD homozygous SNPs at different simulated coverage depths compared to the number of shared homozygous SNPs across all 3 mice. (f) The number of IBD heterozygous SNPs at different simulated coverage depths compared to the number of shared heterozygous SNPs across all 3 mice

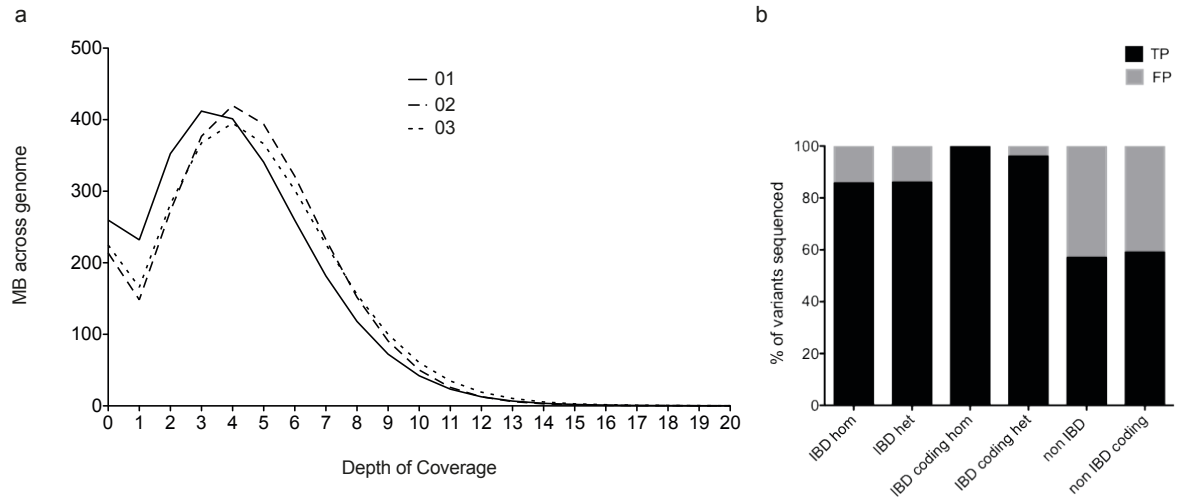


Figure 5.6: (a) Coverage distribution and (b) sequencing validation of candidates in a pedigree with 3 mice sequenced at low coverage. True positive (TP) variants are defined as those confirmed by Sanger sequencing, and false positive (FP), those not found using Sanger sequencing.

also greater than that observed using IBD (Figure 5.5f). By only considering variants in IBD genomic intervals, it is possible to distinguish homozygous from heterozygous variation more accurately and reduce the number of variants incorrectly assigned as shared, making isolation of causative mutations feasible at low coverage.

5.5 Estimation of the ENU mutation density

As described in the introduction to this chapter, previous estimates of the ENU mutation density, which have ranged from 0.5 Mb^{-1} to 10 Mb^{-1}, have been confounded by small datasets and locus specific bias (Russell et al. 1979; Beier 2000; Coghill et al. 2002; Quwailid et al. 2004; Concepcion et al. 2004; Takahasi, Sakuraba, and Gondo 2007; Nolan, Hugill, and Cox 2002).

The Lander–Green based IBD analysis identifies the genomic regions inherited from the ENU ancestor, and this knowledge can be used to measure the ENU mutation density within these regions. By observing the density of variants in the ho-

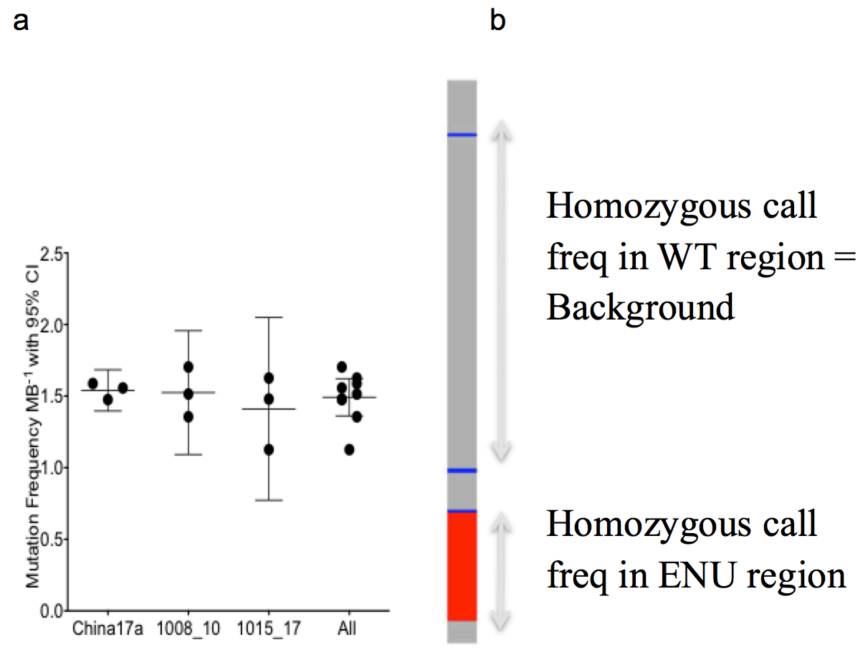


Figure 5.7: (a) Calculated ENU mutation density in 3 unrelated pedigrees. (b) ENU mutation density was calculated as mutations per Mb in ENU regions minus mutations per Mb in WT regions. Such regions are illustrated for a typical chromosome.

mozygous ENU regions and subtracting the background rate observed in homozygous WT regions (Figure 5.7b), the ENU mutation rate in the *ENU16CH17a* pedigree was calculated to be 1.54 mutations Mb^{-1}. Errors due to assignment of homozygous regions or inadequate coverage were excluded by modelling the effect of expansion or contraction of regions and of reduction in coverage (Figure 5.9a, Figure 5.9b and Methods). The estimate of mutation density was also insensitive to changes in the assumed (prior) mutation density used in the algorithm to predict IBD regions – with assumed ENU mutation frequencies in the range 0.25 to 3.0 mutations Mb^{-1}, the estimated ENU density remained between 1.52 and 1.58 mutations Mb^{-1} (Figure 5.8a). Likewise the estimate of ENU mutation density remained consistent with assumed priors for the WT mutation density below 0.8 Mb^{-1} (Figure 5.8b).

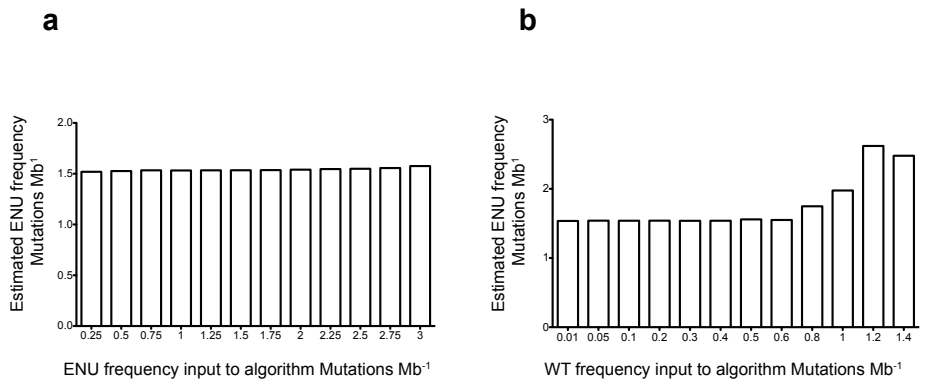


Figure 5.8: a. Effect of input ENU mutation density on the estimated ENU mutation density from the IBD homozygous regions. b. Effect of input WT variant density on the estimated ENU mutation density. The Lander-Green algorithm ran using a range of assumed ENU mutation frequencies or WT variant frequencies as inputs, and the mutation density was estimated from the data output as described in the Methods.

Subsequent sequencing of G_3 mice from further ENU pedigrees, (section 5.7) allowed comparison of ENU mutation rates across pedigrees using the same IBD method. This demonstrated consistency in the mutation rate both intra and inter pedigree, with a mean density of 1.5 mutations Mb^{-1} (confidence interval 1.36 – 1.62) (Figure 5.7a).

5.6 Other characteristics of ENU observed from the WGS

Within the homozygous ENU regions, the well-described transition:transversion ratio and AT base preference of ENU induced mutations was confirmed (Barbaric et al. 2007; Sakuraba et al. 2005).There was a 1.50:1 transition:transversion ratio in ENU mutations compared to a 2.17:1 ratio in naturally occurring mouse SNPs (Methods) (Figure 5.9c).

Analysis of bases at mutated loci confirmed the distinctive base preference sig-

nature of ENU mutations, which is mainly due to error-prone repair of O^2 and O^4 ethylthymidine. In the presence of O^2 ethylthymidine within the DNA template, DNA polymerase incorporates deoxyadenosine (dA) rather than deoxythymidine (dT) leading to AT to TA transversions (Bhanot et al. 1992) (28.5% of *ENU16CH17a* mutations). Likewise O^4 ethylthymidine induces AT to GC transitions (Klein et al. 1990) (45.0% of *ENU16CH17a* mutations) (Figure 5.9d). 78.7% of all homozygous mutations in *ENU16CH17a* were at AT sites, compared to the 58% AT content of the mouse genome (Waterston et al. 2002) and different to non–ENU variants seen in homozygous WT regions, of which 39.5% were at AT sites (Figure 5.9e).

Analysis of the 20bp either side of the full variant set in the *ENU16CH17a* strain revealed that the GC content is 42.09%, which is representative of that in the mouse genome as a whole, and shows that any GC bias in Illumina sequencing (Dohm et al. 2008) has not influenced the base substitution profile of the variants called. The nucleotides 1 or 2bp either side of the targeted base in the homozygous ENU mutations (mean 647 variants per individual) showed a slight bias towards GC and away from AT pairs at the base immediately adjacent to a SNP, with the converse observed at the base 2 bp from the SNP (Figure 5.10), the GC bias in adjacent bases is consistent with findings in published ENU variants (Barbaric et al. 2007), though the AT bias at the 2bp positions is more pronounced in our data. The average AT content at the 2 bases adjacent to each ENU SNP was 51.7% (1−) and 55.4% (1+), slightly lower than the 58% AT content of the mouse genome. The average AT content at the 2− and 2+ positions was 64.4% and 61.8% respectively.

5.7 IBD analysis in further pedigrees

The algorithm described above was applied to two additional ENU pedigrees with similar breeding strategies to that described in *ENU16CH17a*. In these pedigrees,

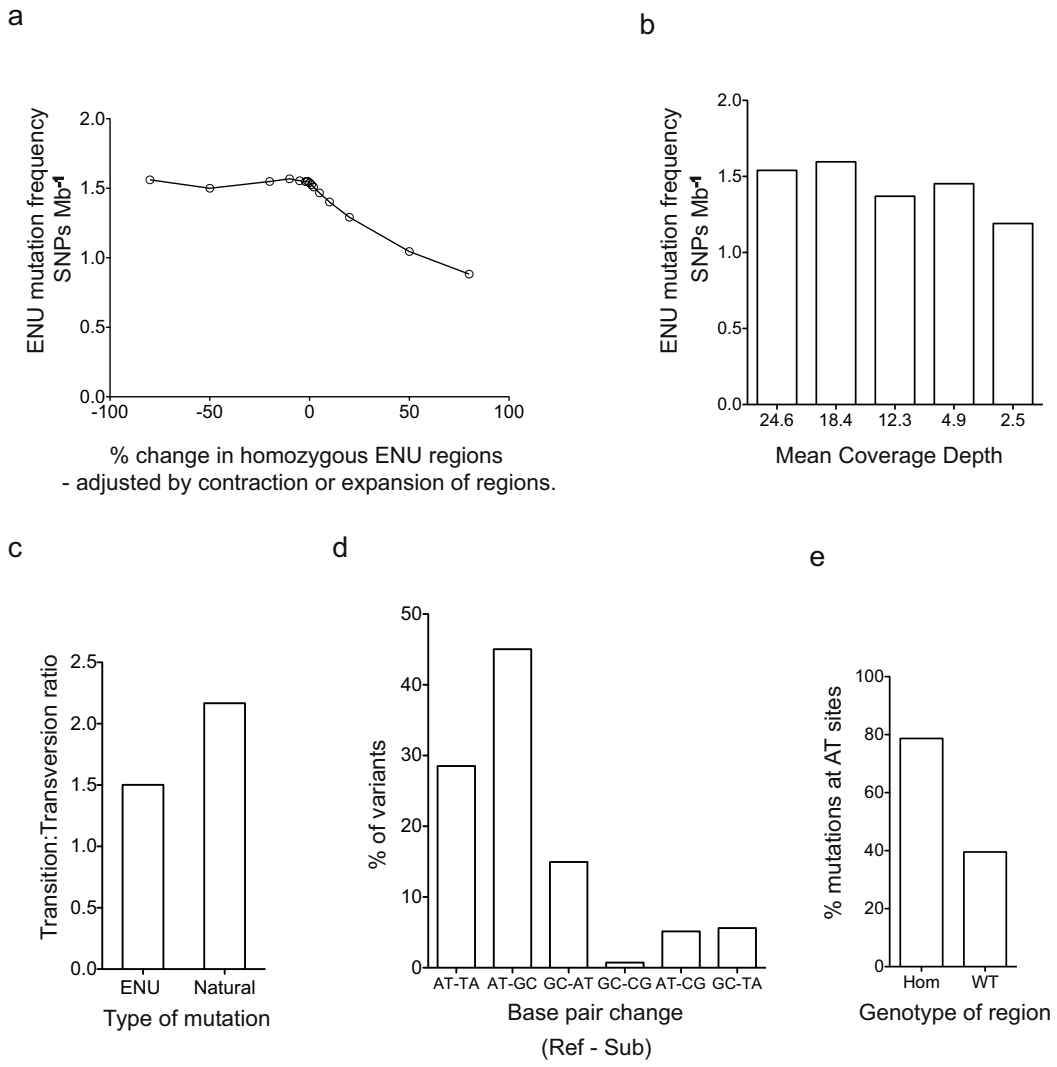


Figure 5.9: Characterisation of the ENU mutations. (a) The effect of expanding or contracting the homozygous ENU regions on the estimate of the ENU mutation density. The start and stop positions of output ENU homozygous regions from the Lander–Green algorithm were adjusted to model error in the assignment of regions by the algorithm (Methods), and the ENU mutation density recalculated as described. (b) The effect of simulated depth of coverage on the estimated ENU mutation density. (c) Transition:transversion ratio in homozygous ENU variants compared to a large dataset of non-mutagen induced laboratory mouse variation from the Centre for Genome Dynamics Mouse SNP Database. (d) The distribution of ENU mutations, showing reference base pairs and substitutions (ref–sub). (e) The proportion of homozygous mutations that occur at AT sites in homozygous ENU and homozygous WT regions. In each graph, columns or points show mean values across the 3 sequenced ENU16CH17a mice.

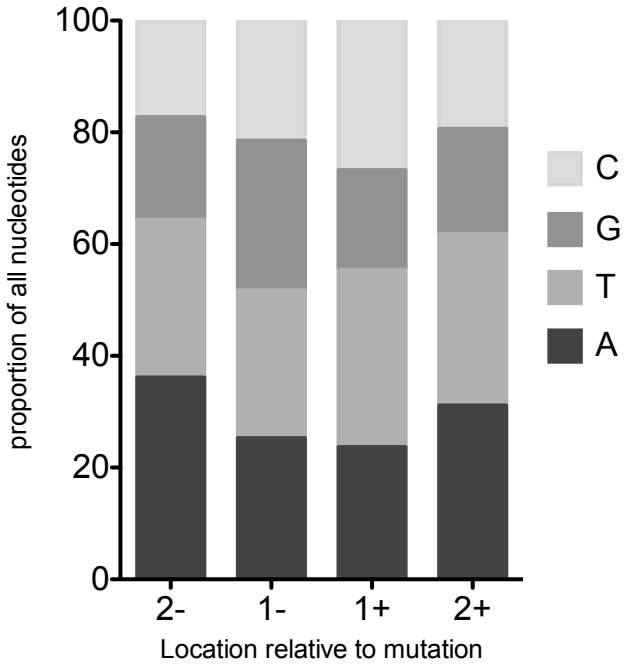


Figure 5.10: Base content of 4 bases surrounding ENU mutations. Based on the filtered set of homozygous variants across the three sequenced ENU16CH17a mice.

generated and phenotyped by Professor Goodnow's group at ANU, no linkage information was available and DNA from affected G_3 mice was sequenced at low coverage.

Based on the evidence modelling lower coverage in the *ENU16CH17a* pedigree (Section 5.4.3) all 3 mice from a single pedigree were sequenced on one lane of an Illumina Hiseq machine, at a mean total coverage of 12–fold, or approximately 4–fold per individual.

The first strain, *APFNS1015_17*, had a recessive phenotype of reduced B220 B cell numbers and low peripheral CD4 and CD8 T cells. 3 G_3 mice (Figure 5.11a) were sequenced and a homozygous IBD region of 23.9 Mb on chromosome 11 was identified, plus a smaller 12.1 Mb region on the X chromosome (Figure 5.11b). The phenotype was inherited equally by males and females in a recessive rather than X–linked fashion so the variant was suspected to lie within the IBD region on chromosome 11. Within the IBD regions two coding mutations were identified, in olfactory receptor

979 (*Olfr979*) and in dedicator of cyto-kinesis 2 (*Dock2*). The mutation in *Dock2* is a T to A substitution at position 34,258,126 on chromosome 11 (mm10 / GRCm38 mouse reference assembly). This is predicted to replace a tyrosine with a premature stop codon in exon 18. Given the phenotype *Dock2* was an obvious candidate since, as previously described in section 3.4.1, *Dock2* –/– mice have deficiencies in B and T cell migration and antigen specific T cell proliferation (Fukui et al. 2001; Sanui et al. 2003). Collaborators at ANU confirmed that the *Dock2* mutation segregated with the phenotype. Notably the *Dock2* mutation was not called in the third mouse due to low coverage, but due to the method of identifying IBD regions and examining all possible candidates within these regions the mutation was still flagged as a candidate overall despite the low coverage in one individual.

The second strain, *ENU22*, (Figure 5.11c) also had a recessive trait with loss of immunoglobulin D (IgD) expression on B cells indicating a failure of maturation (Appendix E). A 9Mb homozygous IBD region on chromosome 12 identified using the algorithm (Figure 5.11d), contains a single G to C candidate mutation in the Immunoglobulin heavy constant mu (*Ighm*). The mutation converts an AGGT splice donor sequence to AGCT, inactivating the splice donor since the GT sequence at the 5' end of the intron is almost invariant (Breathnach and Chambon 1981). This results in skipping of the CH1 region, which normally pairs with the light chain constant region to form immunoglobulin (Ig) and is also needed to associate with the endoplasmic reticulum (ER)-retention quality control chaperone, binding immunoglobulin protein (BIP). Failure of this interaction allows truncated IgM heavy chains to be expressed and function as surface and secreted Ig despite not being paired with light chains (Lee et al. 1999). Ighm -/- mice have a developmental block at the pre B cell stage, however the truncated heavy chains in *ENU22* appear to behave like normal IgM and support B cell survival and maturation and light chain exclusion, but fail to support significant expression of IgD.

The small region on the Y chromosome labelled as IBD is likely to be uninformative, all 3 sequenced mice in pedigree *ENU22* were male and the whole Y chromosome should be shared between the 3 sequenced mice.

5.8 Modelling an efficient ENU programme

5.8.1 Modelling inheritance of ENU mutations within a pedigree and the feasibility of an IBD approach

The inheritance of mutations within a typical ENU pedigree was modelled, to explore how the computationally efficient and rapid route from phenotypes to candidate genes using IBD could be applied to a large–scale ENU programme. Two key questions must be answered to assess the feasibility of applying the IBD method to identify variants in a high throughput manner within an ENU programme.

Firstly it is necessary to know how many mice in a typical ENU pedigree would carry a single detectable segregating phenotype. In the hypothetical case of complete penetrance and a fully sensitive screen this equates to the number of G_3 mice in a pedigree which carry a single shared variant.

Secondly it is necessary to know how many mice from a typical pedigree should be sequenced to isolate a linkage region containing on average only a single candidate mutation, both for a dominant or recessive trait.

To address these issues, it is necessary to know first how frequently a hypothetical fully penetrant ENU mutation causing a screened phenotype would be observed among the G_3 mice in the proposed breeding strategy. It was assumed that a G_1 pair could give rise to 4 stable G_2 pairs, each generating 3 litters with a conservative estimate of 4 live mice per litter. A single pedigree would then generate 48 G_3 mice for screening (Figure 5.1a). Using a probabilistic model incorporating all the possible inheritance patterns (Methods), it was calculated that 51% of ENU mutations

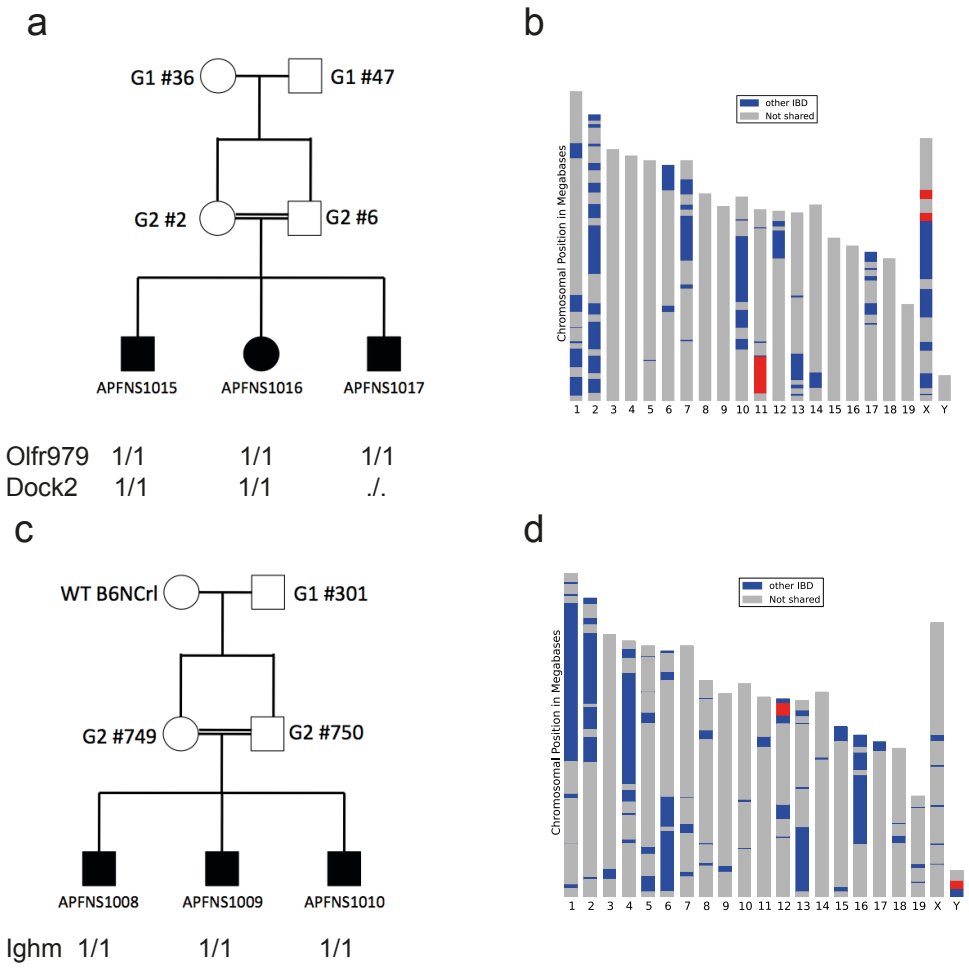


Figure 5.11: IBD using Lander-Green in additional pedigrees. (a) Pedigree in strain APFN1015-1017, the sequenced mice are shaded. The gene and genotype for the candidate mutation is shown for each sequenced individual. 1/1 indicates homozygous for mutation, ./. indicates insufficient coverage to call the genotype at that locus in an individual. (b) Plot showing IBD homozygous (red) and IBD heterozygous (blue) regions predicted by the Lander-Green based algorithm in APFN1015-1017. (c) Pedigree for strain ENU22 with genotypes for the Ighm mutation. (d) Plot showing IBD regions for ENU22.

present in the founder mice occur 3 times or more as homozygous within the set of G_3, and 62% occur twice or more. 99.6% of mutations would be present as a single allele in at least one of the 48 G_3 mice, 98% at least 6 times and 95% at least 9 times (Figure 5.12a).

Next, it was important to establish whether, under the proposed strategy, the number of non–causative candidate IBD mutations would be sufficiently small to efficiently exclude these mutations. The number of candidate mutations IBD in affected G_3 mice from a single pedigree can be estimated as a function of the number of affected G_3 sequenced, the empirical ENU mutation density of 1.5 mutations Mb^{-1} (Figure 5.7a), the relatedness of the mice and the proportion of mutations that affect protein sense. On average, 1.4% (9/647) of homozygous mutations in each *ENU16CH17a* G_3 mouse lie in exons or splice sites; 73% of this subset cause miss–sense or splice site mutations; and no homozygous stop mutations were identified in *ENU16CH17a*. Thus it was derived that 1.05% of mutations affect protein sense. These numbers are consistent with those found in larger datasets (Arnold et al. 2012). Using these parameters it is calculated that sequencing 3 or 6 mice will typically reduce the number of candidates to 1 homozygote or 2 to 3 heterozygote mutations respectively (Figure 5.12b) and Methods). This model is consistent with the empirical data from the *ENU16CH17a* pedigree.

5.8.2 Modelling saturation of genes by ENU mutagenesis

Forward genetic ENU programmes have used screens to target a broad spectrum of clinically relevant phenotypes (Angelis et al. 2000) or a specific set of biological pathways, for example those involved in immune tolerance (Hoyne and Goodnow 2006). In either case the actual number of genes involved, and the extent of functionally relevant pathways, is unknown. A key strength of ENU mutagenesis is the ability to uncover previously unknown players in genetic disease, but this also raises the

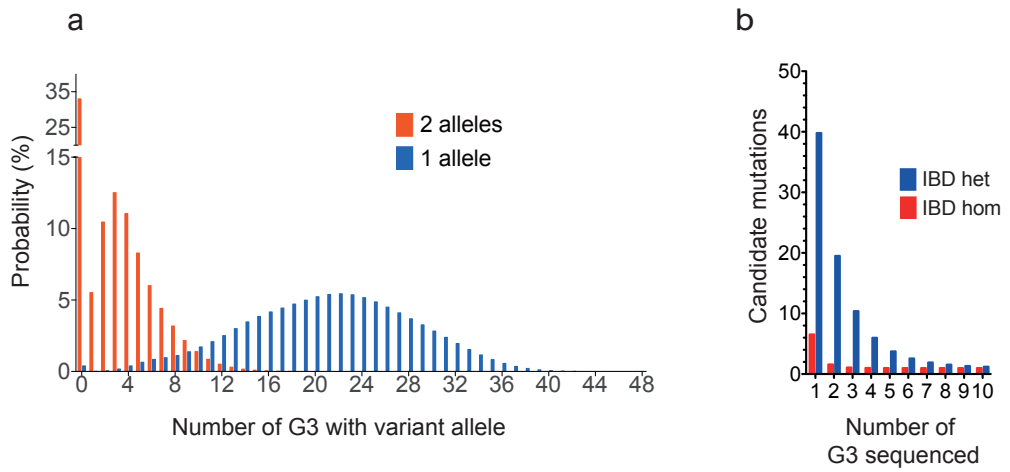


Figure 5.12: Modelling the number of mutants in a pedigree and the power to assign causation by WGS. (a) The distribution for all mutations within a pedigree at the G_3 level, based on a model pedigree of 48 G_3 arising from 4 G_2 pairs (Methods). In the specific case of mutations causing fully penetrant phenotypes, the histograms show the distribution of affected mice with recessive (2 allele) and dominant (1 allele) traits. (b) The number of IBD candidate mutations, defined as miss-sense, stop or splice-variants, as a function of the number of sequenced affected G_3 mice, based on the model.

question of how much mutagenesis must be performed to be confident of detecting all, or most of, the genes involved. Conversely one wishes to avoid wasting resources by repeatedly uncovering the same mutants having saturated the underlying genetic pathways. This modelling aims to show how many ENU pedigrees, and how many mice, would be required to generate mutations in most genes, another important parameter in developing an efficient and comprehensive ENU programme.

5.8.2.1 Assumptions for this model

The model incorporates the calculated ENU mutation density of 1.5 mutations per Mb (Figure 5.7a) and assumes 3 times 90mg/kg ENU dosing and two founder mice per B6 pedigree as used in *ENU16CH17a*, with a pedigree structure as described for the previous modelling (Section 5.8.1 and Methods).

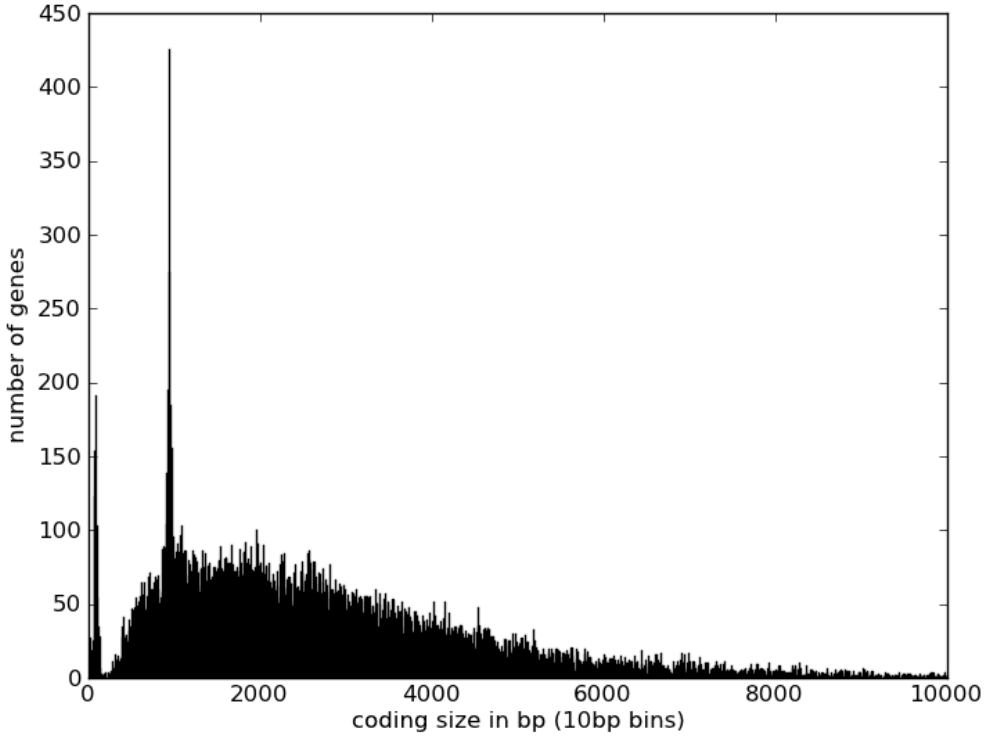


Figure 5.13: Distribution of RefSeq genes by size of coding region. The peak at around 940 bp is due to many olfactory genes with the same or very similar sizes.

5.8.2.2 Effect of gene size on mutation density

The distribution of mutations across genes will be a function of gene size. Figure 5.13 shows the distribution of RefSeq (NCBI) gene sizes. There is a peak of very small genes that will not be easily accessible by ENU, and a long tail of very large genes that are likely to be frequently mutated. The peak at around 940 bp is due to olfactory genes. Mice have over 1700 olfactory and vomeronasal genes and these are highly similar (Ibarra-Soria et al. 2014).

To model for this skewed distribution the number of mutations M for a specific gene in n pedigrees can be estimated as:

$$M = \text{ ENU mutation rate } \times 2 \times \text{ length of gene} \times n$$

Where the factor of 2 accounts for the two ENU treated founders per pedigree.

Only a proportion of these mutations will be deleterious or have a detectable phenotype. Mutations that induce a miss-sense, non-sense or splicing variant in a gene will be of most interest. Other mutations (e.g. synonymous changes or mutations in un–translated regions) will be less likely to have functional consequences.

The proportion of mutations that are miss–sense, non-sense or splicing within a gene will depend on the specific codons within that gene. To model the fraction of mutations of each type per gene, the coding sequence for each gene was extracted, and mutations randomly simulated across this sequence and checked for functional consequence in terms of mutation type using Annovar (Wang, Li, and Hakonarson 2010). Thus for each gene the fraction of mutations that will result in non–synonymous, synonymous, non-sense or splice changes can be estimated.

A simulation was scripted to incorporate the ENU mutation density, the known gene sizes and the specific fraction of mutations likely to alter protein sense for each gene (Methods).

5.8.2.3 Gene targeting

Figure 5.14 shows the result of running this simulation with increasing numbers of pedigrees. The data is for at least one mutation per gene, and shows that 242 pedigrees (approx 11,500 G_3 mice) would generate at least one coding mutation (of any kind including synonymous/ UTR) in 90% of genes (Figure 5.14, blue line). More meaningfully, 1,991 pedigrees (95,500 G_3 mice) would generate at least one miss–sense, non–sense or splice variant in 89% of genes (Figure 5.14, red line). Saturation of all genes with a stop mutation would require very large numbers of mice, but 2,000 pedigrees would generate a nonsense mutation in 20% of genes (Figure 5.14, green line).

The potential to generate allelic series, or multiple mutations within a gene, was then examined. Figure 5.15 shows the number of genes with at least 1, 2, 3, 4 or 5

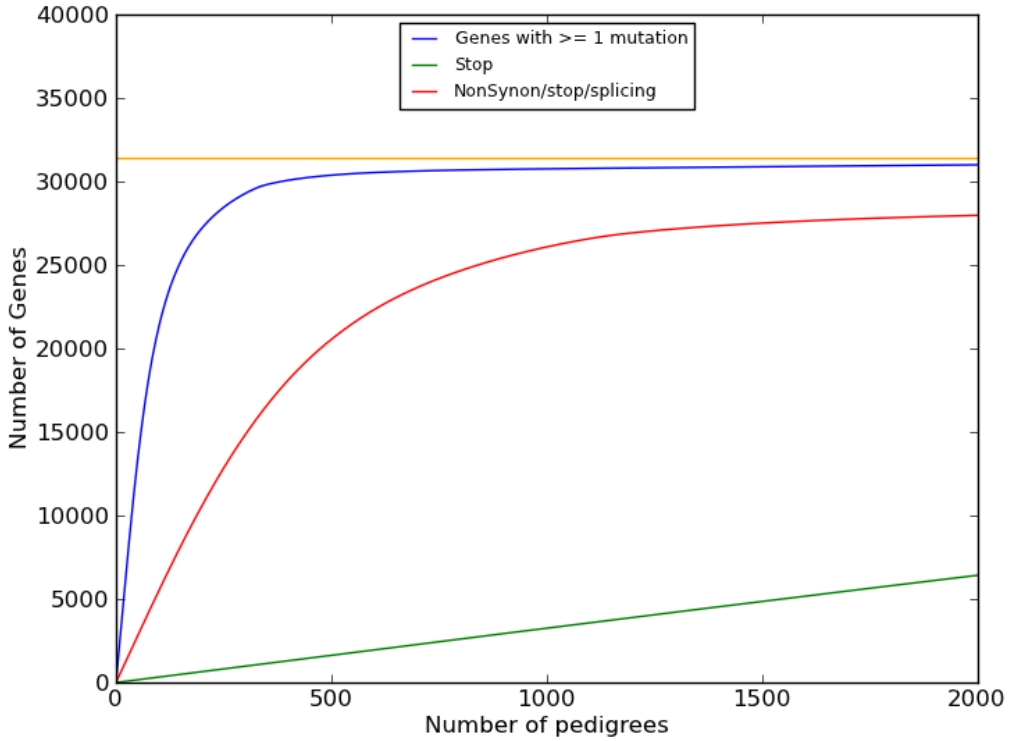


Figure 5.14: The number of genes with at least one mutation. The red line shows how an ENU programme would generate viable point mutations in genes, with potential to affect phenotype. The orange line indicates the total number of genes in the mouse genome (based on RefSeq geneset, see Methods).

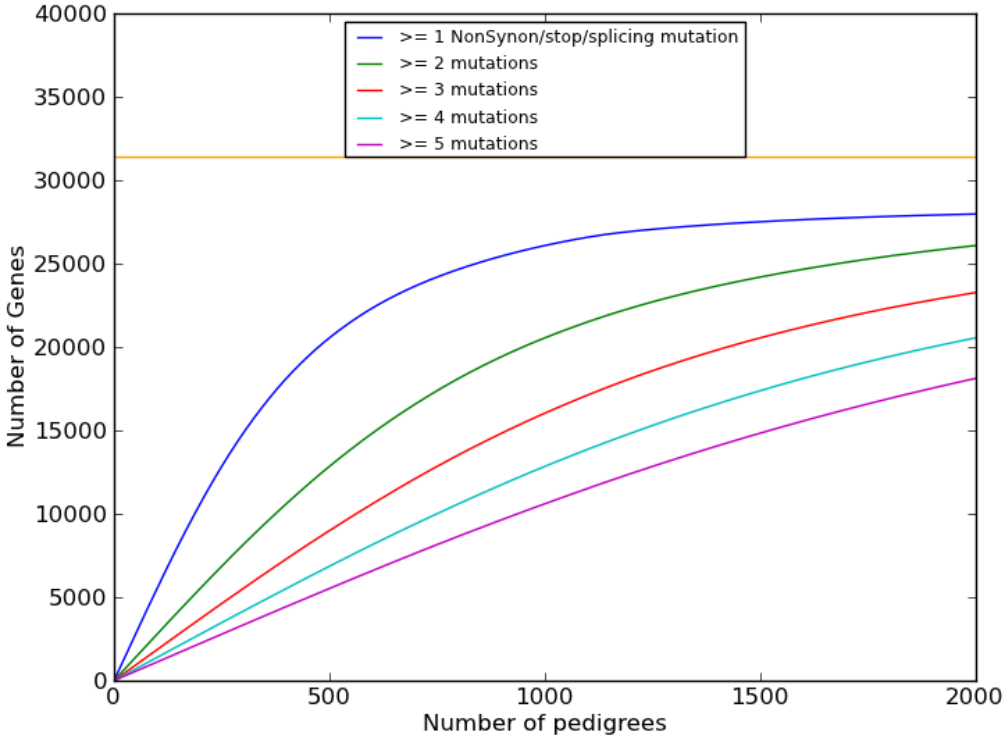


Figure 5.15: Non-synonymous, non-sense or splice site mutations occurring multiple times in the same gene, showing how allelic series of genes would be generated.

mutations as the number of pedigrees mutated is increased.

5.8.2.4 Mappable mutations

Modelling in section 5.8 showed that 51% of mutations within a pedigree will be inherited as homozygous by at least 3 G_3 mice and 98% of mutations within a pedigree will be inherited on at least one allele by 6 or more G_3 and would therefore be mappable. In the case of the non–sense (stop) mutations the observed homozygous mutations may be around 1/3 lower due to embryonic lethality. Reflected onto the scale in Figure 5.14, 1,000 pedigrees would generate 28,000 homozygous and 54,000 heterozygous mappable non-sense, miss-sense or splice mutations, distributed over 83% of all genes.

5.9 Summary of Chapter

This chapter shows how low coverage WGS of multiple mice with a phenotype can identify causative ENU mutations without the need for out-crossing, or knowledge of dominant or recessive inheritance (Section 5.3, Section 5.4 and Figure 5.1). The advantage of the linkage-based approach is that ENU-induced mutations from multiple affected mice can be used to track the IBD regions and then isolate the causative mutation. This strategy simultaneously generates linkage maps (Figure 5.4) and identifies the shared mutations with a high degree of confidence. Both modelling (Figure 5.5) and empirical results (Figure 5.6) demonstrate that the method is effective at low coverage.

The method also permitted estimation of the ENU mutation density at 1.5 mutations Mb^{-1} (Figure 5.7), based on a much larger dataset of variation that previously published estimates. Knowledge of the ENU mutation density was used to model an efficient sequencing strategy. The data show that sequencing 3 affected G_3 mice with a recessive trait or 6 mice with a dominant trait would yield on average 1 or 2 candidate IBD mutations (Figure 5.12), indicating that the method would be feasible with typical ENU pedigrees.

Chapter 6

Searching for rare variants in patients with Steroid Resistant Nephrotic Syndrome or Systemic Lupus Erythematosus

6.1 Introduction to chapter

Developing targeted therapeutic agents for common diseases is a key goal of translational biomedical research. In order to design such agents more effectively we need to understand the genetic component of common diseases in humans, which are typically genetically complex. One potentially tractable source of the heritable component for complex disease is the contribution of rare variants with large effects on phenotype.

ENU can provide models of human disease, as illustrated by *nephertiti* in Chapter 4, and can provide candidate variants of large effect to validate in patients, but ENU cannot mimic all the phenotypic or genotypic features of human disease. For example some human conditions such as neurodegenerative disease are incompletely replicated

by murine mutants (Saito et al. 2014), and experiments in inbred mice cannot model the complex interactions of environment and genetic background that contribute to human phenotypes. Furthermore ENU does not induce large insertion or deletion mutations in DNA and is an inefficient method to induce disruptions in very small genes. Therefore ENU animal studies must be used as a complement to direct forward genetic studies in humans.

This chapter explores the use of WGS to identify rare variants of large effect in humans with two diseases: Steroid Resistant Nephrotic Syndrome (SRNS) and Systemic Lupus Erythematosus (SLE). Unrelated patients with severe early onset forms of these diseases were selected with the aim of enriching for monogenic forms of the nephrotic syndrome or SLE.

SRNS in children is a predominantly monogenic disease that can provide insights into the biology of protein leak from the kidney (proteinuria). Proteinuria is a key element of more common complex renal diseases predominantly in adults, and a defining criteria for chronic kidney disease, a condition that affects 8.5% of the UK population (Stevens et al. 2007). Proteinuria, the nephrotic syndrome and the genetics of SRNS are discussed in chapter 4.1.

SLE is often considered the archetypal systemic autoimmune disease; in contrast to SRNS it is genetically and physiologically more complex and highly heterogeneous. However monogenic forms exist. Renal disease is a major cause of morbidity and mortality in SLE, affecting 50% of patients within a year of diagnosis (Seshan and Jennette 2013).

This chapter aims to explore the power of WGS to identify rare functionally significant variants for each of these diseases: one mainly monogenic but offering a window into the pathology of more complex, common disease, and the other a complex disease with a significant genetic component and rare monogenic forms.

The work is based on the hypothesis that it may be possible to identify rare

variants in isolated cases. It is an imperfect experiment because it lacks full family data, but it tests a clinical scenario , in which an individual presents de novo, without other information.

6.1.1 Systemic Lupus Erythematosus

SRNS has been described in Chapter 4.1, therefore this introductory section will focus on SLE.

A complete review of SLE is outside the scope of this work; so this section briefly discusses the clinical disease spectrum, and focuses on the current knowledge of the genetic causes of SLE, relating this to the mechanisms of autoimmunity observed in the disease.

SLE is a chronic multisystem autoimmune disease characterised by inflammation and the production of autoantibodies against intracellular self antigens. The clinical features of SLE are highly variable but can include rash, arthritis, glomerulonephritis, anaemia and neuropsychiatric disease. Diagnosis requires 4 of 11 criteria, illustrating the disease heterogeneity (Tan et al. 1982; Hochberg 1997).

6.1.1.1 Prevalence and populations

SLE has a prevalence of 28 / 100,000 in the UK (Johnson et al. 1995), but the prevalence is much higher in some ethnic groups, up to 694/100,000 in some African American populations (Chakravarty et al. 2007). Differences in prevalence, clinical manifestations and mortality in different ethnic groups may partly be attributed to socioeconomic differences (Pons-Estel et al. 2010), but admixture studies indicate that genetics explains more of the observed ethnic variance in lupus nephritis than socioeconomic status (Alarcon et al. 2006).

There is a broad spectrum of disease severity, with some patients suffering chronic severe multisystem morbidity. Although less than 10% die within 5 years of diagnosis,

the mortality rates are higher in non-white ethnicities, men, paediatric cases and the elderly (Pons-Estel et al. 2010). The disease most commonly affects women of childbearing age, but 10–20% of SLE cases are diagnosed in childhood, typically defined as below 14-16 years of age. Diagnosis is rare below the age of 5 (Cervera et al. 1993). Renal and central nervous system involvement, fever and lymphadenopathy are all more common in paediatric cases (Font et al. 1998).

6.1.2 The genetics of SLE

Siblings of individuals with SLE are more likely to develop the disease than the general population: the sibling risk ratio is estimated to be between 5 and 29 (Segovia et al. 2005), supporting the genetic or at least familial contribution to SLE susceptibility. The heritability of SLE is estimated at 66% (Guerra, Vyse, and Cunninghame Graham 2012), with concordance rates of 24 - 48% in monozygotic twins and 2-5% in dizygotic twins (Block et al. 1975; Deapen et al. 1992).

In a majority of cases the inheritance of SLE is complex and does not follow simple Mendelian rules. Multiple genetic variants each conferring a small increased risk of autoimmune disease interact with environmental factors or triggers leading to disease. Identifying these genes is important because they will highlight pathways fundamental to the development of SLE that may in turn become targets for therapeutic intervention, and individual genetic profiling is likely to be of diagnostic and prognostic benefit, enabling clinicians to identify patient groups at risk of more severe disease and predict individual responses to specific treatment.

Over 70 susceptibility genes or loci have so far been identified for SLE, in addition to strong associations with the major histocompatibility complex (Fernando et al. 2007). The majority of the known SLE associated genes have been identified since 2008 by genome wide association studies (GWAS) (Cui, Sheng, and Zhang 2013). In general these genetic variants have only small individual effects and taken together

are estimated to predict only 10-15% of the heritability of SLE (Guerra, Vyse, and Cunninghame Graham 2012; Manolio et al. 2009). Extensive genetic heterogeneity and low disease prevalence have limited the power of GWAS to identify SLE genes (Frazer et al. 2009).

6.1.2.1 Monogenic forms of SLE

An alternative approach to identifying rare variants is to focus on familial, early onset (paediatric) or syndromic SLE in which a single genetic defect of large effect may be isolated. Whilst familial monogenic SLE contributes only 1–2% of SLE cases, these families have been vital in revealing mechanisms leading to loss of self tolerance, which may be important in the aetiology of more complex multigenic, multifactorial forms of disease. Monogenic forms of SLE have been identified with diverse genetic defects including complement deficiencies, overproduction of Interferon alpha (IFN-α) and apoptosis defects.

Primary complement deficiencies in the early components of the classical complement pathway are strongly associated with SLE. In particular homozygous deficiency of C1q, C1r, C1s, C4 and to a lesser extent C2 predispose to the development of lupus. C1q carries the strongest association of any known gene with SLE, 93% of reported patients with a homozygous C1q deficiency develop a clinical syndrome consistent with SLE, although incidence of double stranded DNA (dsDNA) antibody is low (Walport, Davies, and Botto 1998). In contrast, only 10% of patients with C2 deficiency, the most common complement deficiency (prevalence 1/20,000 in Caucasians) develop SLE (Stern et al. 1976), and C3 deficiency has only been linked to SLE in 13% of cases (Manderson, Botto, and Walport 2004).

Aicardi-Goutieres syndrome (AGS) is a rare autosomal recessive disease causing inflammatory encephalopathy in infants. AGS children have high levels of IFN-α and share clinical features with SLE. Genes known to cause AGS include *AGS5*,

TREX1, and *RNASEH2* subunits. Familial chilblain lupus can be due to heterozygous mutations in *TREX1* (Rice et al. 2007) or *AGS5* (Ravenscroft et al. 2011).

Autoimmunity in the lupus prone mice C3H/HeJ–gld/gld and MRL/lpr–lpr is due to mutations in the apoptosis genes *fas* ligand (*FasL*) and *Fas* (tumour necrosis factor receptor superfamily member 6 (Tnfrsf6)) respectively. A mutation in *FASL* causing defective FasL activity has been reported in a patient with SLE (Wu et al. 2011b).

DNAse type I is an endonuclease capable of cleaving both single and double stranded DNA. A homozygous null mutation in *DNASE1L3* was identified in a consanguineous family with autosomal recessive SLE, and the authors hypothesized that the mutation leads to defective clearance of degraded DNA, triggering an immune response (Al-Mayouf et al. 2011).

6.1.3 The genetic basis of SLE pathology

Current ideas about the pathogenesis of SLE focus on complex interactions including defective clearance of apoptotic cells and immune complexes, loss of immune tolerance, increased antigenic load, T cell over–activity, defective B cell suppression and autoantibody production. The following sections discuss these mechanisms with reference to known genetic associations with SLE, focussing on non–MHC genes.

6.1.3.1 Apoptosis, ubiquitination and immune complex binding

In SLE it is proposed that defective clearance of apoptotic cells can lead to autoantibody production and immune complex mediated production of interferon (IFN), resulting in chronic inflammation and loss of tolerance (Guerra, Vyse, and Cunninghame Graham 2012).

Early classical pathway complement deficiency results in defective targeting and clearance of apoptotic cells (Bigler et al. 2009; Botto and Walport 2002), indicating a mechanism for the association of C1q deficiency and SLE, and *PRDM1–ATG5* loci

variants may influence autophagy and apoptosis (Maiuri et al. 2007).

Variants in integrin, alpha M (complement component 3 receptor 3 subunit) (*ITGAM*) are associated with SLE (Shizhong Han 2009). *ITGAM* mediates uptake of complement coated particles and an SLE variant compromises phagocytosis, immune complex clearance and leukocyte adhesion (MacPherson et al. 2011).

Ubiquitination is a process of post-translational protein modification that can tag proteins for degradation. The ubiquitin-conjugating enzyme E2L (*UBE2L3*) gene is strongly associated with SLE (Shizhong Han 2009). Variants in Tumor necrosis factor- alpha-induced protein (*TNFAIP3*), encoding the ubiquitin editing enzyme A20, and TNFAIP3-interacting protein (*TNIP1*), an adaptor protein that binds to A20, increase NF-κB signalling and are associated with increased SLE risk (Wertz et al. 2004; Musone et al. 2008; Gateva et al. 2009).

Miss–sense polymorphisms and copy number variants in Fcγ receptor (*FCGR*) genes, are associated with SLE and modulate leukocyte responses to immune complexes (Niederer et al. 2010).

6.1.3.2 Toll like receptor signalling

In SLE toll like receptors (TLRs) may be triggered by endogenous autoantigens as well as viral nuclear particles. Therefore it is suggested that activated TLRs activate both innate and adaptive immune pathways, via IFN–α production by dendritic cells and autoantibodies to DNA produced by B cells (Kontaki and Boumpas 2010). The regression of SLE in a patient with an acquired TLR signalling defect suggests that innate TLR signalling is important for disease persistence and amplification of inflammation (Visentini et al. 2009). Mice with a duplication of Toll-like receptor 7 (*Tlr7*) are predisposed to autoimmunity (Izui et al. 1995; Fossati et al. 1995; Subramanian et al. 2006) and human *TLR 3/7/8* variants are associated with SLE risk in a Taiwanese population (Wang et al. 2014).

6.1.3.3 Interferon production

It is widely proposed that type I interferons also have a central role in SLE. IFN–α/β trigger induction of immature dendritic cells, upregulation of MHC class 1, and the production of pro-inflammatory chemokines, leading to clonal expansion of activated effector CD8+ and CD4+ T cells. IFN–α/β also exert anti-apoptotic effects on B cells (Theofilopoulos et al. 2004). IFN–α levels are increased in SLE and IFN–α from SLE patients can induce maturation of monocytes into antigen-presenting dendritic cells (Blanco et al. 2001), suggesting that (pseudo)viral IFN–α responses may drive or further propagate SLE autoimmunity. SNPs in interferon regulatory factor genes, particularly interferon regulatory factor 5 (*IRF5*), increase expression of the transcription factor and IFN–α and are associated with increased SLE risk (Feng et al. 2010). SLE associated variants in interferon induced with helicase C domain 1 (*IFIH1*) and tyrosine kinase 2 (*TYK2*) may also increase IFN–α expression (Guerra, Vyse, and Cunninghame Graham 2012).

6.1.3.4 T cell Activity

Dysregulation of T cells in SLE results in aberrant T cell help to auto–reactive B cells and T cell infiltration of target organs including the kidneys (Moulton and Tsokos 2011). The MHC region encodes for proteins involved in presentation of antigen to T cells. Other genes involved in SLE that influence T cell development or T cell receptor signalling include signal transducer and activator of transcription 4 (*STAT4*), which has the strongest GWAS association with SLE (Shizhong Han 2009) and is required for Th1 cell development, protein tyrosine phosphatase non-receptor type 22 (*PTPN22*), in which an autoimmune risk SNP down regulates T cell receptor signalling (Vang et al. 2005) and tumor necrosis factor (ligand) superfamily member 4 (*TNFSF4*), which encodes the CD123 ligand, expressed on activated T cells (Cunninghame Graham et al. 2008).

6.1.3.5 B cell responsiveness

SLE patients exhibit B cell dysregulation with polyclonal hyperactivity and defects in maturation (Dörner, Giesecke, and Lipsky 2011). B cell scaffold protein with ankyrin repeats (*BANK1*) is B cell specific and increases calcium mobilisation and thus B cell activation. An SLE associated variant alters splicing efficiency and may increase downstream activity of *BANK1* (Kozyrev et al. 2008). SLE risk alleles in V-yes-1 Yamaguchi sarcoma viral-related oncogene homolog (*LYN*) (Lu et al. 2009) may increase B cell hyperactivity by modulating the activation threshold via CD22 inhibitory signalling, as shown in *Lyn–/–* mice (Hibbs et al. 1995; Nishizumi et al. 1995). Likewise B lymphoid tyrosine kinase (*BLK*) acts in the B cell signalling pathway and variants confer SLE risk (Shizhong Han 2009). Other genes that are associated with SLE and may act via modulation of B cell activity include neutrophil cytosolic factor 2 (*NCF2*) (Cunninghame Graham et al. 2011) and RAS guanyl releasing protein 3 (calcium and DAG-regulated) (*RASGRP3*), which is important for lymphocyte development and B cell proliferation and antibody production (Guerra, Vyse, and Cunninghame Graham 2012). Lymphocyte development and differentiation may be disrupted in SLE and variants have been identified in V-ETS avian erythroblastosis virus E26 oncogene homolog 1 (*ETS1*), which may function by loss of negative regulation of Th17 cells and terminal differentiation of B cells (Pan et al. 2011). SNPs in Ikaros family zinc finger 1 (*IKZF1*), a lymphoid specific transcription factor may contribute to autoimmunity by effects on lymphocyte differentiation, B cell receptor signalling and interactions with *STAT4* (Guerra, Vyse, and Cunninghame Graham 2012; Cunninghame Graham et al. 2011)

6.2 Strategy for WGS in early onset SRNS and SLE

6.2.1 Selection of SRNS Patients

To explore the possibility of identifying genes causative for nephrotic syndrome, four patients with early onset SRNS were identified from a national rare renal disease registry (www.renalradar.org) and studied in collaboration with Professor Moin Saleem's group at Bristol University. These patients had previously been studied within a larger cohort of 36 patients using a targeted exome capture panel for 24 genes associated with SRNS and a further 422 genes with known or potential involvement in nephrotic syndrome or glomerular function (McCarthy et al. 2013). In 25 of the 36 exome capture patients no definitely or probably pathogenic candidate variant had been identified. More than 30% of children with sporadic SRNS may have a monogenic aetiology (Giglio et al. 2014). Therefore it was hypothesised that these patients could harbour rare variants of large effect within genes not previously associated with SRNS, or carry variant types, such as splicing or non–coding, which fall outside the regions targeted by exome sequencing.

Exclusion criteria for WGS included recurrence of nephrotic renal disease post transplantation. Since this is rare in familial childhood SRNS, 1 in 41 cases in one series (Conlon et al. 1999) compared to rates as high as 20% in primary focal segmental glomerulosclerosis (FSGS) or 93% of patients with initial steroid sensitivity (Newstead 2003; Ding et al. 2014). Four patients without evidence of post transplant recurrence (Table 6.1) and in whom the only filtered variants observed in the 24 genes using exome capture were common and/or not predicted to be deleterious were selected for WGS (Table 6.2).

Gender	Ethnicity	Age at Onset (yrs)	Duration of follow up (yrs)	Current CKD stage (T= transplanted)	Time to ESRF (months)	Disease recurrence post transplant
M	Pakistani	1	6	5 (T)	12	No
F	Mixed	2	2	1	n/a	n/a
F	White British	4	6	2	n/a	n/a
F	White British	2	4	5	57	n/a

Table 6.1: Characteristics of 4 unrelated patients with early onset SRNS selected for WGS.

Gene	Variant	Patient	dbSNP (build 135)	MAF or ESP genotype count	HGMD	PolyPhen	MutPred
COL4A4	Ala715Val	0001	rs76636743	A=0.0014	No	Benign	0.282
COL4A4	Ala1558Val	0001	Not in dbSNP	No data	No	Unknown	0.418
INF2	Arg877Gln	0001	rs142678449	ESP: AA-0.0004 AG-0.0168 GG-0.9828	No	Unknown	0.447
NPHS1	Asn1077Ser	0001	rs4806213	C=0.0836	ConNS (but also in control)	Possibly damaging	0.269
NPHS1	Glu117Lys	0002 and 0003	rs3814995	T=0.3423	ConNS (but also in control)	Possibly damaging	0.802
PMM2	Glu197Ala	0004	rs34258285	C=0.0119	Congenital disorder of glycosylation 1a (but polymorphism)	Benign	0.826

Table 6.2: Variants within genes associated with SRNS identified using targeted exome capture in the 4 SRNS patients described in Table 6.1. All variants described were genotyped as heterozygous. None were considered causative for SRNS. MAF is minor allele frequency from 1000 Genomes Project), ESP is NHLBI Exome Sequencing Project allele frequency. HGMD is the Human Gene Mutation Database, PolyPhen is PolyPhen–2 (Adzhubei et al. 2010), ConNS is congenital nephrotic syndrome, MutPred is a random forest based pathogenicity prediction method (Li et al. 2009a; Thusberg, Olatubosun, and Vihinen 2011).

6.2.2 Selection of SLE patients

10 patients with early onset SLE were selected for whole genome sequencing through a collaboration with Professor Tim Vyse at King's College London and Professor Earl Silverman at the University of Toronto. The mean age at time of diagnosis was 7

years and 6 months (range 3 years 1 month to 10 years 6 months) (Table 6.3).

Two patients were affected siblings (sisters, patients 26106 and 39124). The 8 others were all unrelated patients. Apart from the two siblings the patients had no family history of SLE and were considered as isolated cases. Four patients had central nervous system (CNS) involvement, 8 had renal disease, and two of these had both renal and CNS involvement. The type of renal pathology is shown in Table 6.3.

Patient	Ethnicity	Gender	Age at diagnosis	Lupus nephritis class (0 = no renal disease)	CNS phenotype
16603	black	M	6.4	5	no
46154	european	F	9.1	3 and 5	no
16743	black	F	10.8	4	yes
17709	latin am	F	3.1	4	yes
22564	black	F	7.7	3	no
17571	black	F	8.6	3	no
45629	european	F	8.4	0	yes
45742	european	F	9.5	3 and 5	no
39124	black	F	8.2	0	no
26106	black	F	3.2	4	no

Table 6.3: Characteristics of 10 SLE patients selected for WGS. Lupus nephritis classed according to the International Society of Nephrology / Renal Pathology Society classification for lupus nephritis (Weening 2004). Central nervous system (CNS) indicates the presence or absence of a central neurological phenotype.

6.2.3 WGS coverage

The SRNS and SLE patients were sequenced as part of a larger WGS500 project which was a collaboration between the Oxford Biomedical Research Centre's Genomic Medicine Theme, the Wellcome Trust Centre for Human Genetics, and Illumina, with the aim of evaluating the clinical utility of WGS in 500 samples across a range of human diseases.

The individuals were sequenced to a target coverage depth of 25–fold. Figure 6.1 shows coverage depths in the SRNS patients.

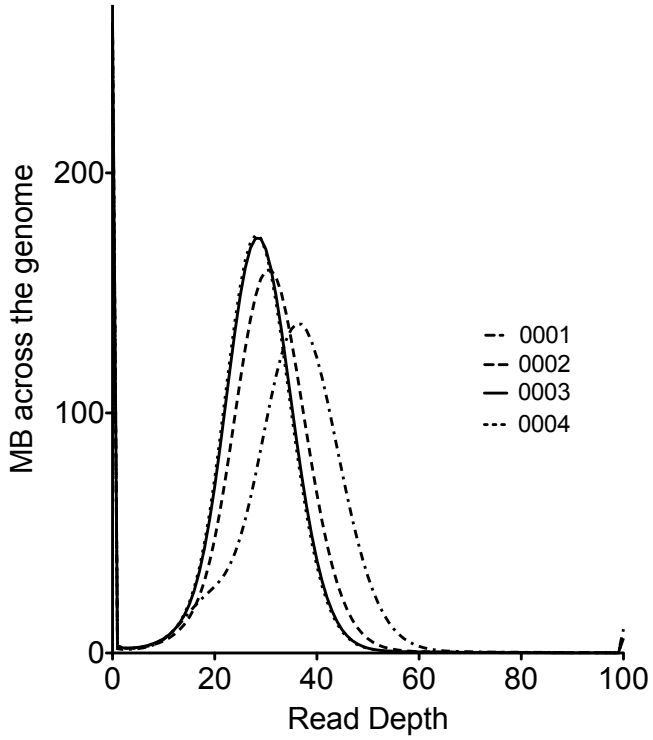


Figure 6.1: Histogram of WGS coverage depth for SRNS patients

An initial downstream pipeline developed by the bioinformatics core at the Wellcome Trust Centre, mapped the reads, called and annotated variants, and generated a 'union' file of all variants. The downstream analysis of these variants, which forms the basis of this chapter, highlights obstacles and strategies for identifying causative mutations applicable to rare forms of complex disease in humans using WGS data. This analysis process is summarised in figure **??**.

6.3 Analysis of WGS

6.3.1 General considerations

Since the affected individuals examined in this study were isolated cases without family history of renal disease and were unrelated (except the two siblings with SLE),

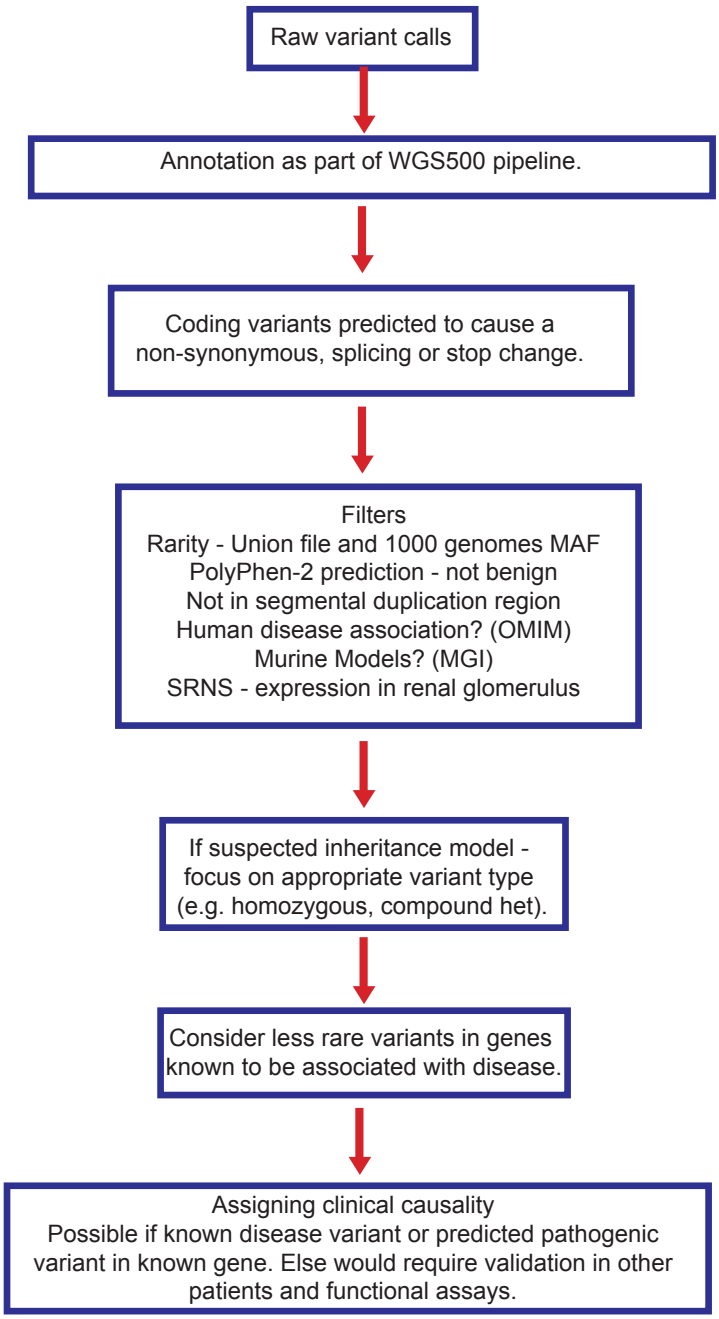


Figure 6.2: Human variant study process. Steps taken to filter and prioritise variants in the SRNS and SLE patients. Assigning a variant as disease causing must be a more stringent process than in the ENU experimental setting since a clear causative variant would be shared with the patient and their clinicians.

it was anticipated that they would not share causative genes or variants. Although we hoped to enrich for recessive alleles by choosing sporadic cases, no assumptions were made as to inheritance. For example in the case of monogenic SRNS a causative variant could be due to a recessive or compound heterozygous defect, not previously reported in the family, or to a heterozygous variant carried on a single allele. In the latter case, given the lack of family history, this must be either a de novo mutation arising in the patient or a dominant variant of incomplete penetrance. The majority of variants causing SRNS in children have an autosomal recessive mode of inheritance, including mutations in *NPHS1*, *NPHS2* and *LAMB2*, in contrast to monogenic adult onset FSGS which is predominantly autosomal dominant (Rood, Deegens, and Wetzels 2012).

Table 6.4 and Table 6.5 summarise the current progress in identifying candidate causative variants in the cohorts.

Patient	Homozygous candidates	Compound heterozygous candidates (unphased)	Heterozygous candidates	Candidates in known SRNS genes. all / coding or splice only
0001	5 (4 genes)	29	273	250 (20 homozygous) / 2
0002	2	12	144	111 / 0
0003	0	49	361	280 / 0
0004	0	17	182	131 / 0

Table 6.4: Summary of numbers and types of candidate variants in the SRNS patients

Table 6.5 summarises the candidate rare variants by type in the 10 SLE patients, and shows the numbers shared by siblings 39124 and 26106.

6.3.2 Autosomal recessive candidates

In three patients, SRNS patient 0001, SLE patient 16743 and SLE patient 17709, the WGS demonstrated evidence, shown below, of parental consanguinity. Therefore

Patient	Homozygous candidates	Compound heterozygous candidates (unphased)	Heterozygous candidates	Candidates in known SLE genes All/Coding
16603	1	32	276	98/2
46154	4	27	105	107/1
16743	18	236	331	128/0
17709	0	67	182	108/0
22564	0	32	312	117/0
17571	1	157	310	118/2
45629	0	19	132	97/1
45742	0	14	127	126/0
39124	7	240	374	115/3
26106	5	246	364	136/3
sibs 39124 and 26106	0	130	219	98/2 All/Coding

Table 6.5: Numbers of variants by type in the 10 SLE patients

in these patients an autosomal recessive cause is a more likely explanation for the disease. There is no evidence of consanguinity or family history in the other cases and so here a de novo mutation or compound heterozygote is more likely.

Any homozygous causative variant will be rare although might exist as a low frequency heterozygote within the population. Hence all homozygous variants with a population frequency below 0.05 (based on 1000 genomes minor allele frequency (MAF)) were considered, but excluded if they were present as homozygotes in the union file of all variants observed in 226 WGS500 individuals, in the 1000 genome database (McVean et al. 2012), or other database of known population variation.

The WGS500 pipeline variants were further filtered for allelic bias, such that variants genotyped as homozygous, but with alternate allele frequency equal to or below 30% were excluded. Likewise variants observed in more than one of the SRNS or SLE patients were excluded, on the basis that a shared homozygous causative variant between these unrelated patients is unlikely but systematic errors or shared benign variants are very likely. Variants seen as high frequency heterozygotes in the WGS500 union file were excluded (greater than or equal to 0.05), as were variants not

predicted to be deleterious by PolyPhen-2 (specifically those with a score less than or equal to 0.15, (Adzhubei et al. 2010)). For the two SLE siblings only shared rare homozygous variants were considered.

Additionally, the variants in patient 0001 were filtered against 108 Punjabi genomes from Lahore in Pakistan, available as part of the 1000 genomes project (see Methods). This was helpful as this ethnicity was under represented in the original 1000 genomes cohort and the WGS500 dataset, and variants in patient 0001 that appear rare may be common within individuals of the same ethnicity. Patient 0001 has many more candidate variants that the 3 other SRNS patients even after filtering against the Pakistani cohort. In patient 0001, 15.6% (7/45) of the filtered homozygous candidates could be excluded due to a minor allele frequency (MAF) of greater than 0.05 in the Punjabi cohort.

The numbers and types of rare homozygous variants for SRNS and SLE remaining after these filters are summarised in Table 6.6.

Patient	Homozygous rare	NS	Stop	Splice	Indel	Unknown
0001 SRNS	38	31	0	2	3	2
0002 SRNS	6	3	1	1	1	0
0003 SRNS	6	2	0	3	1	0
0004 SRNS	2	0	0	0	2	0
16603 SLE	9	9	0	0	0	0
46154 SLE	6	4	0	0	1	1
16743 SLE	20	16	2	0	1	1
17709 SLE	6	6	0	0	0	0
22564 SLE	0	0	0	0	0	0
17571 SLE	1	1	0	0	0	0
45629 SLE	0	0	0	0	0	0
45742 SLE	0	0	0	0	0	0
39124 SLE	9	7	0	0	1	1
26106 SLE	7	6	0	0	0	1
sibs 39124 and 26106 SLE	2	1	0	0	0	1

Table 6.6: The numbers and categories of homozygous candidate variants per individual. The homozygous rare column gives total numbers of rare recessive candidates. NS is non–synonymous.

Review of this data shoes, not unexpectedly, that the majority of rare homozygous variants are NS (miss–sense) changes. Indels, stop–gains and splicing variants are much less common. Protein truncating variants are under strong purifying selection, and homozygous variants even more so, but rare variant sets are enriched for deleterious variants including stop gains (Nelson et al. 2012; McVean et al. 2012). Stop–gain variants comprise 4% (2/55 mean per patient) of the rare coding (NS or stop) homozygous variants but only 0.7% (77 / 11792) at all population frequencies in the SLE cohort. This difference is consistent with a shortlist of homozygous variants in the SLE patients representing relatively rare, recent variants that have been subject to less purifying selection (McVean et al. 2012).

The two homozygous stop-gain variants in the SLE patients are both in patient 16743 on chromosome 6. The variant in coiled coil domain containing 28A (*CCDC28A*) has a reported 1000 genomes MAF of 0.0005 but was observed as a het in one non–SLE WGS500 individual, this variant is present in 1.5% of NHLBI exome sequencing prject African–Americans and 1% of 1000 genomes Yoruba samples. A translocation disrupting this gene has been reported in a case of megakaryoblastic leukaemia but the significance of this is unclear. The other stop gain variant is in retinoic acid early transcript 1E (*RAET1E*), a protein which functions as a ligand for NKG2D receptor (Cao et al. 2007). NKG2D is expressed on NK and CD8+ T cells, suggesting a plausible role in innate and adaptive immunity. However the *RAET1E* variant has a 1000 genome frequency is 0.01, the NHLBI Exome Sequencing Project (ESP) MAF is 0.02 and the variant is present as a heterozygote in two non–SLE WGS500 patients, suggesting that it is too common to be a single gene cause of SLE, though this does not exclude a possible contributory role in causation.

Contrasting with the data above, 17 rare homozygous variants across the SLE patients were not identified in the 1000 genomes cohort. 10 of these were X-linked, 6 in the male patient 16603. Patient 16603 has no family history suggestive of an X-

linked inheritance model so these variants appear less likely to be causal, unless due to de novo variants on the X chromosome. The remaining 11 rare homozygous variants are shown in Table 6.7. Two variants in *MMP17* were found to be on a single read in an area of low coverage and therefore were excluded. Of the remaining 9 homozygous variants, all except *ARMCX4*, seen as het in another SLE case, were observed as a heterozygote in at least one non–SLE patient, and in one case in 46 individuals. The variant in *ARMCX4* is present in dbSNP (rs113417448) with unknown MAF but on inspection was called in a region of very low coverage and allele bias and does not look reliable.

Patient	Gene	Location	Ref	Sub	Type	1000g MAF	WGS500 Union
16743	DMBT1	10:124361496	G	A	NS	unknown	HT = 1
16743	MMP17	12:132313113	T	C	NS	unknown	HT = 2
16743	MMP17	12:132313119	T	C	NS	unknown	Same read as above
16743	BMP15	X:50659210	C	CTCT	Nonframeshift Insertion	unknown	Hom in 2 SLE pt and HT in 14 inc one SLE
17571	FOXG1	14:29236616	A	C	NS	unknown	HT=3
17709	SOX21	13:95363670	C	A	NS	unknown	HT=7
17709	PAGE1	X:49452149	G	T	NS	unknown	HT=1
16603	FAM149B1	10:75000739	G	A	NS	unknown	HT=46 inc 2 in SLE
39124	BMP15	X:50659210	C	CTCT	Nonframeshift insertion	unknown	Hom in 2 SLE pt and HT in 14 inc. one SLE
46154	GDF7	2:20867208	CCCGCG CCGCGCG	C	Nonframeshift deletion	unknown	HT=5 inc 1 HT in SLE
46154	ARMCX4	X:100749037	G	A	NS	unknown	HT in SLE pt 45742

Table 6.7: Homozygous variants in SLE patients not present in 1000 genomes database. Ref is reference base, sub is variant base. 1000g MAF is the minor allele frequency reported by the 1000 genomes project. NS is non–synonymous, HT is the number of heterozygous calls in the WGS500 union file.

Lists of filtered and annotated homozygous candidates for the SRNS patients are contained in Appendix F. The homozygous rare variants in the SLE patients with gene, variant type, 1000 genome MAF and frequency in the WGS500 union file are contained in Appendix G.

Rare homozygous candidates in patients with evidence of consanguinity are discussed below.

6.3.2.1 Genetic evidence for consanguinity in SRNS patient 0001

Analysis of patient 0001 highlights potential problems associated with consanguinity and is therefore illustrative of one of the particular challenges to studying sporadic disease. In this case the evidence of consanguinity is based on regions of homozygosity within the WGS variant calls. This was initially noted in the homozygosity plot generated as part of the WGS500 pipeline. The pipeline includes an analysis of homozygosity and copy number variation (CNV). Plots of coverage and ratio of homozygous to heterozygous calls are generated in 'bins' or windows across the genome (Figure 6.3a). The precise regions of homozygosity, both size and number, were then determined by submitting known dbSNP variant calls seen in 0001 to an online tool that identifies homozygous regions (Seelow et al. 2009).

From this analysis, patient 0001 had 20 runs of homozygosity (ROH) greater than 1Mb in length, comprising 140.2 Mb of homozygous regions and equating to approximately 5.4% of the genome(Figure 6.3b). This is larger than would be expected to occur under linkage equilibrium in individuals not sharing a common ancestor (Karl W Broman 1999) (Figure 6.3c).

Theoretically offspring of second cousins should have of a coefficient of inbreeding, F, equal to 1/64, i.e. 1.5% of their genome would be expected to be homozygous and first cousins should have 1/16 or 6.3% of their genome as homozygous (Lander and Botstein 1987), suggesting that 0001 could have first cousin parents. However in practice the coefficient of inbreeding is often not clear because of multiple generations of consanguinity and incomplete knowledge of the genealogy (Woods et al. 2006).

The evidence of consanguinity points to a likely homozygous causative variant for SRNS in 0001 within the 38 rare variants identified (Table 6.6. This assumes both a

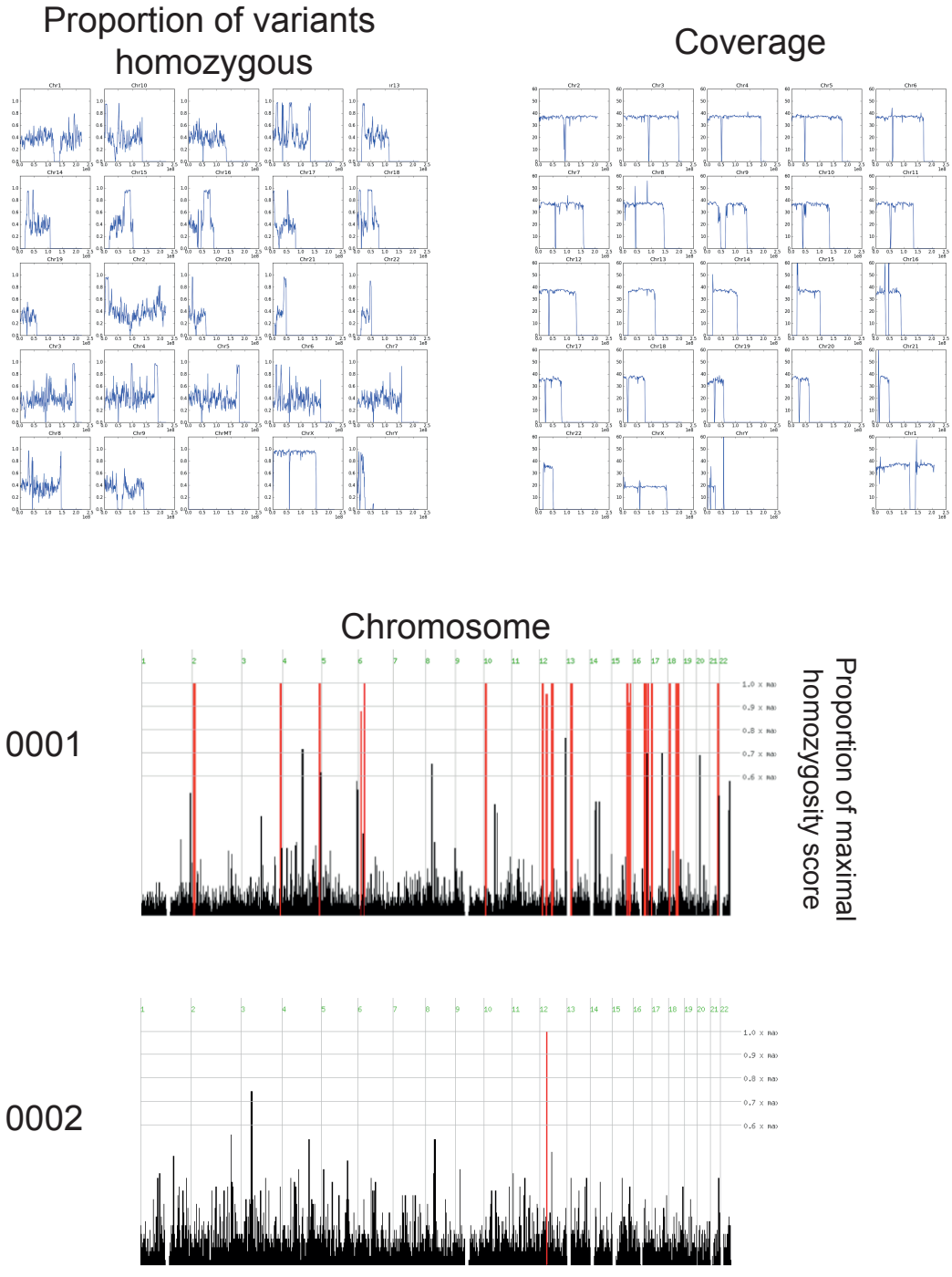


Figure 6.3: Runs of homozygosity indicate consanguineous parentage in patient 0001. Figures from the WGS500 pipeline show discrete regions with a high proportion of homozygous to heterozygous calls, without variation in coverage depth (a). These are more precisely distinguished using a homozygosity mapping tool (b) and can be contrasted with the same analysis in patient 0002, in which there is no history of consanguinity (c). Homozygosity scores in (b) and (c) represent excess of found homozygosity as a proportion of the maximum, against the expectation calculated from population values (Seelow et al. 2009).

recessive mode of inheritance, and a coding or splice site mutation.

6.3.2.2 Prioritisation of homozygous recessive candidate variants in SRNS patient 0001 narrows the short list to 6 variants

Of the 38 homozygous rare variants in patient 0001 (Table 6.6), 6 variants were in regions of segmental duplication, defined as two or more similar sites with identical flanking regions, these were excluded. 3 variants had PolyPhen-2 scores of zero indicating they were likely to have a benign effect on the protein function. 7 were not true homozygotes as they were on the X chromosome. Variants on the X chromosome in this male are not recessive variants due to consanguinity, being inherited only from the mother. Therefore they were excluded as candidates based on the assumed autosomal recessive inheritance model. One splicing variant in phospholipase C-like 2 (*PLCL2*) was observed in non-SRNS genomes as low quality calls, leaving 23 recessive candidates for patient 0001 (Appendix F.1).

For these 23 rare homozygous candidates the possible functional effects can in some cases be predicted by manually curating the shortlisted variants. This was based on available mouse model data and/or literature searches for expression data and human disease associations. Appendix F.1 describes known functional data for these 23 genes, and indicates which have a MAF below the threshold of 0.05 in the population, below 0.01, or are unique to the patient based on available datasets.

15 of the 23 candidate homozygous variants in patient 0001 are particularly rare, having a MAF less than 0.01. However 5 of these: Inter-Alpha-Trypsin Inhibitor Heavy Chain Family, Member 5 (*ITIH5*), Scribbled Planar Cell Polarity Protein (*SCRIB*), Ring finger protein 17 (*RNF17*), proline-serine-threonine phosphatase interacting protein 1 (*PSTPIP1*) and alcohol dehydrogenase 1C (*ADH1C*) have a mouse model or human disease association that does not point towards a role in glomerular

The remaining 9 candidates with MAF less than 0.01 are in *ZBED4*, *SPATA13*, *ELP5*, *PRKD1* (two variants), *NUP93*, *ADAMTSL3*, *USP15* and *ATP3A5*. These were prioritised according to available expression data. Genes causing SRNS or FSGS are in the large majority of cases expressed in the podocytes or the GBM (Rood, Deegens, and Wetzels 2012). Therefore it was hypothesised that candidates in the 4 SRNS patients should also be expressed (though not necessarily exclusively) in podocytes or glomeruli.

The Human Protein Atlas project database was searched for renal expression of each of the 8 remaining candidate genes (9 variants) in patient 0001. The Human Protein Atlas generates and publishes high throughput antibody based profiling of tissue specific expression and currently provides data for more than 75% of human coding genes (Protein Atlas Version 11: release 11.3.2013) (Uhlen et al. 2010). The Protein Atlas data does not provide podocyte or GBM specific expression data but does indicate whether the protein is detected in the renal glomerulus. Using this approach, *NUP93*, *ADAMTSL3*, *USP15* and *ATP3A5* appear less probable candidates since they were undetectable by immunohistochemistry (IHC) in normal renal glomeruli (Table 6.8).

Gene	Level of Antibody Staining human in glomerulus	Differential expression in murine podocyte
ZBED4	Strong	No (log2 fold = 0.157 but p = 0.08)
SPATA13	Weak	No
ELP5	Moderate	Unknown
PRKD1	Moderate	No
NUP93	Undetected	Under expressed
ADAMTSL3	Undetected	Unknown
USP15	Undetected	No
ATP13A5	Undetected	No

Table 6.8: Highest detected level of antibody staining reported in the Human Protein Atlas (http://www.proteinatlas.org) in normal kidney glomeruli and murine podocyte differential expression, based on array data (Boerries et al. 2013) with cut off for significance of <= 1.5 log2 fold value and p value of 0.05

.

This expression analysis therefore left 5 preferred candidate homozygous variants in 4 genes in patient 0001, although none were reported as differentially over expressed in murine podocytes compared to non podocyte renal cells (Boerries et al. 2013).

Of these 5 remaining variants, the non-synonymous substitutions in *ELP5* and *ZBED4* were each observed as a heterozygous variant in one other WGS500 individual, whilst the rest appear to be unique, at least in available sequenced populations. All are predicted as deleterious by PolyPhen–2 except the chromosome 14:30132954 variant in *PRKD1* which has a more equivocal score of 0.382 (Table 6.9), the presence of two variants in *PRKD1* could indicate an error or a compound heterozygous effect. No animal model or human disease association was found for these 4 genes, except *PRKD1* in which knock-out mice have some embryonic lethality (Fielitz et al. 2008).

Gene and Variant	Expression in Glomeruli	Polyphen2 score	Variant Type	MAF
ZBED4 22:50279061	Strong	0.99	NS	<0.01
ELP5 17:716294	Moderate	0.998	NS	<0.01
PRKD1 14:30068947	Moderate	0.995	NS	Not seen appears unique
PRKD1 14:30132954	as above	0.382	NS	Not seen appears unique
SPATA13 13:24797332	Moderate	0.995	NS	Not seen appears unique

Table 6.9: The final 5 remaining homozygous variant candidates in 0001 after in silico filtering. Since the patient has consanguineous parents it is predicted that one of these variants in likely to be causative for SRNS in this patient. NS is non-synonymous

.

Thus in patient 0001, 'in silico' filtering of WGS variants has generated a very manageable short–list of 5 prioritised candidates for further in vitro or in vivo study. The first task will be to confirm that these variants are heterozygous or absent in other unaffected members of the family, particularly any available sibling DNA.

6.3.2.3 Homozygous variants in SLE patient 17709

The case of 17709 also illustrates issues relating to consanguinity and the difficulty of assigning causation to multiple candidates. Patient 17709 has a ROH on chromosome X.

Patient 17709 is a girl who was aged 3 at the time of SLE diagnosis, with both both renal and CNS involvement. ROH analysis performed in the WGS500 pipeline and described in section 6.3.2.1, revealed a region of 74 Mb on chromosome X as having a high proportion of homozygous calls (Figure 6.4). The coverage across this region was normal (not shown), excluding a copy number variation (CNV).

Since there was no report of consanguinity within the family, no other visible homozygous regions on other chromosomes, and a single large ROH, one possibility was that the homozygous region on chromosome X had arisen due to uniparental disomy (UPD). Because the father was reportedly unaffected with SLE it was postulated that this would have arisen from a single maternal X chromosome.

UPD occurs when a child inherits both copies of a chromosome or chromosomal region from one parent. Only the central region of the X chromosome in this female had a high homozygous to heterozygous variant ratio. Partial centromeric isodiomy can arise due to an error in meiosis II resulting in a gamete with two copies of a chromosome after crossing over. This is followed by trisomy rescue that removes the allele from the single allele gamete, leaving a bialleic cell with two copies from one parent which are identical from the centromeric region to the point of cross over (Field et al. 1998). UPD is rare, but it is likely underreported as it does not always lead to disease and may not be detected on routine karyotyping. It has been previously recognised on the entire X chromosome in a case of Duchenne Muscular Dystrophy (Quan et al. 1997).

Based on this hypothesis of UPD on chromosome X, homozygous rare variants on the X chromosome were examined as possible candidate causative variants for

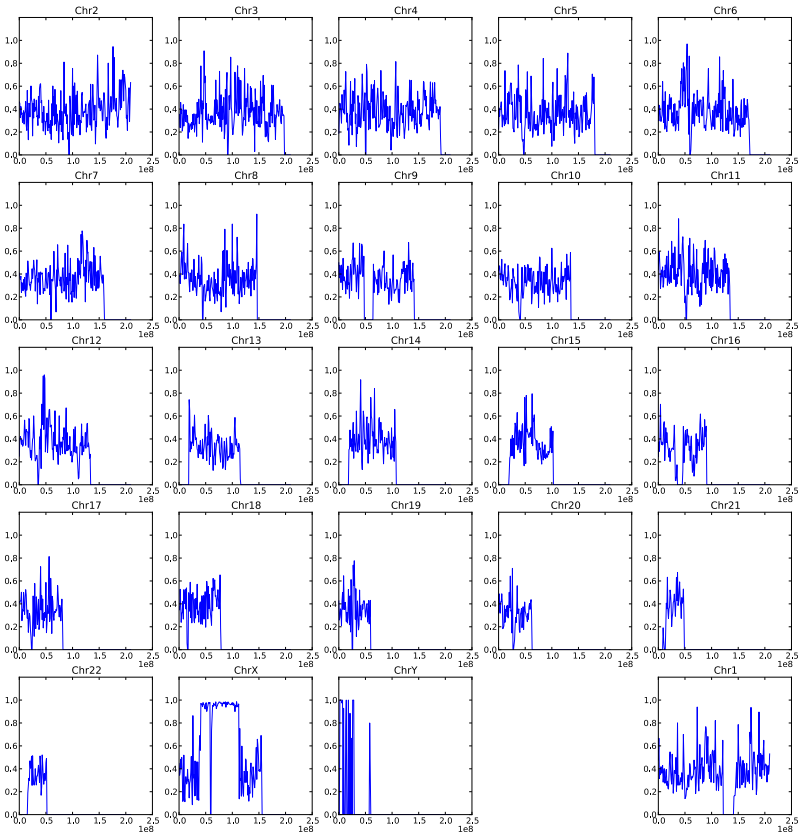


Figure 6.4: Proportion of homozygous variants across each chromosome in 17709. A central region of high homozygous to heterozygous variation can be seen on chromosome X in this female SLE patient.

monogenic SLE.

Patient 17709 has 6 homozygous rare variants (Appendix G), and 2 of these are on the X chromosome. The first is a NS homozygous variant in P antigen family member 1 (*PAGE1*). This has a PolyPhen–2 score of 0.855 and is present in one WGS500 non–SLE individual as a heterozygote, but not seen in 1000 genomes. The protein encoded by *PAGE1* may act as a tumour antigen recognised by cytotoxic T cells, but it is expressed in tumour tissue and reproductive tissues only (Chen et al. 1998; Ulrich Brinkmann 1998; Eynde et al. 1995), making it an unlikely candidate for systemic disease.

The second variant is a homozygous C to A substitution in Wiskott–Aldrich syndrome (*WAS*). This rare variant was not seen in any other WGS500 patients, has a 1000 genomes MAF of 0.0006 and a PolyPhen–2 score of 0.386 suggesting it is possibly deleterious; However the variant is reported in 4 males in the ESP cohort (NHLBI Exome Sequencing Project); although no phenotypic data is available for ESP individuals.

Wiskott–Aldrich syndrome is a rare X–linked primary immunodeficiency (Ochs and Thrasher 2006). The variant identified in 17709 induces a NS Histadine to Asparagine change at amino acid 180 in exon 6 of the WAS protein. It is N–terminal to the basic region and distal to the WIP binding domain, in a region of the protein for which no function has been described. This change is reported in a single male in a case series of *WAS* associated X–linked thrombocytopaenia. The patient had reduced WAS protein expression but no reported autoimmunity or malignancy (Albert et al. 2010). Previous reports of mutations in exon 6 of *WAS* have been deletion or splicing variants. However a family with 2 brothers carrying a G593T mutation inducing a Gly187Cys NS change has been described. Both brothers exhibited a relatively mild presentation of recurrent infections, thrombocytopaenia and eczema in childhood; one developed renal failure due to autoimmune glomerulonephritis aged 22. Levels of

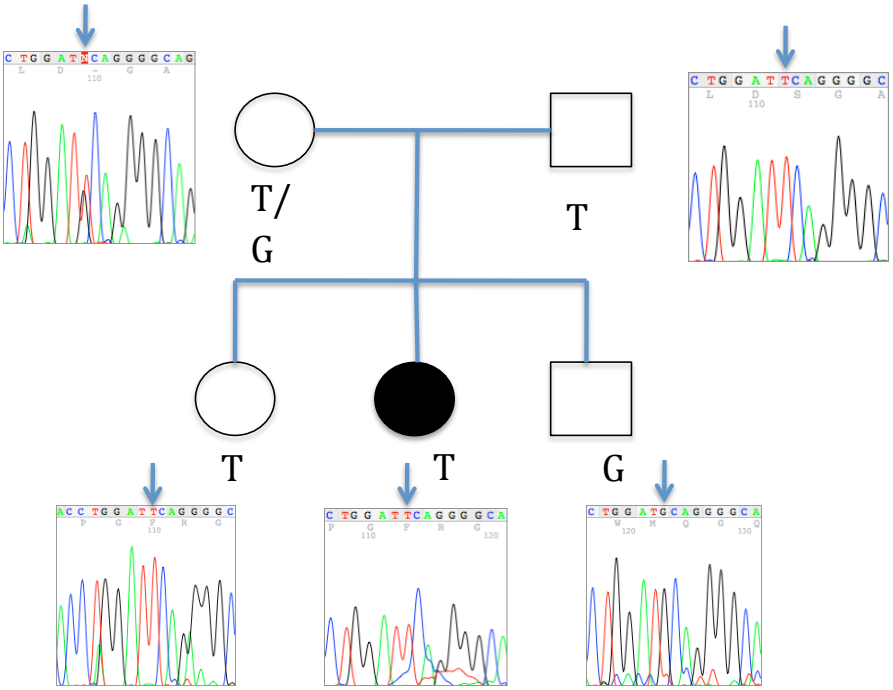


Figure 6.5: Sanger sequencing for WAS variant in 17709 family. Sanger sequencing shown on reverse strand where G is the reference allele and T the alternate allele

WASP were not assayed in this family (Derry et al. 1995).

Based on the suspected UPD and the clinical phenotype of *WAS* variants, the rare homozygous variant in *WAS* appeared a most plausible candidate causative variant in patient 17709.

6.3.2.4 Sanger sequencing for C538A WAS variant in family of patient 17709 does not fit the UPD hypothesis

Sanger sequencing of family DNA revealed that whilst the mother is heterozygote for the WAS^{C538A} allele, the father has only the alternate allele, as does both the patient and her unaffected sister (Figure 6.5).

This suggests that rather than UPD, the rare *WAS* allele may have been inherited from both parents, and is also inconsistent with WAS^{C538A} as a monogenic cause of

SLE since the sister and father are homozygous for the variant allele but unaffected by the disease.

Given the rarity of the *WAS* variant allele, parental consanguinity was now suspected and a SNP array was performed on the family of 17709 to establish the explanation for the Sanger sequencing results and the ROH on chromosome X.

6.3.2.5 Comparison of runs of homozygosity in 17709 with family using SNP array

A SNP array (Human CytoSNP–12 DNA analysis bead chip, Illumina) analysis was performed on DNA from each of the brother, sister, father and mother. This showed that the sister also has a large central ROH on chromosome X. This is the same haplotype as in the patient 17709 and is a larger ROH, sharing a 5' recombination point with 17709 but extending by a further 6 Mb at the 3' end (Figure 6.6a).

Analysis of B allele frequency (BAF) differences across individual variants between pairs of family members confirms that the mother shares the father's X haplotype heterozygously across the central region of the chromosome. Since the haplotype is rare this strongly points towards consanguinity in the family. The BAF differences also show that the chromosome X ROH in the sister is the same haplotype as the patient (Figure 6.6b).

The SNP array data showed no unexpected changes in log R ratio, ruling out copy number variation in the family members.

Further evidence supporting parental consanguinity came from plotting the BAFs from the SNP array data for the autosomes in the family of patient 17709. Small regions on chromosomes 4 and 13 in the patient's brother were noted to have BAFs of 1 or 0, not 0.5, suggesting he also inherited some identical genomic intervals from both parents (Figure 6.7a) and Figure 6.7b). To examine for runs of homozygosity in more detail, the genotypes from the family array and known SNPs from the pa-

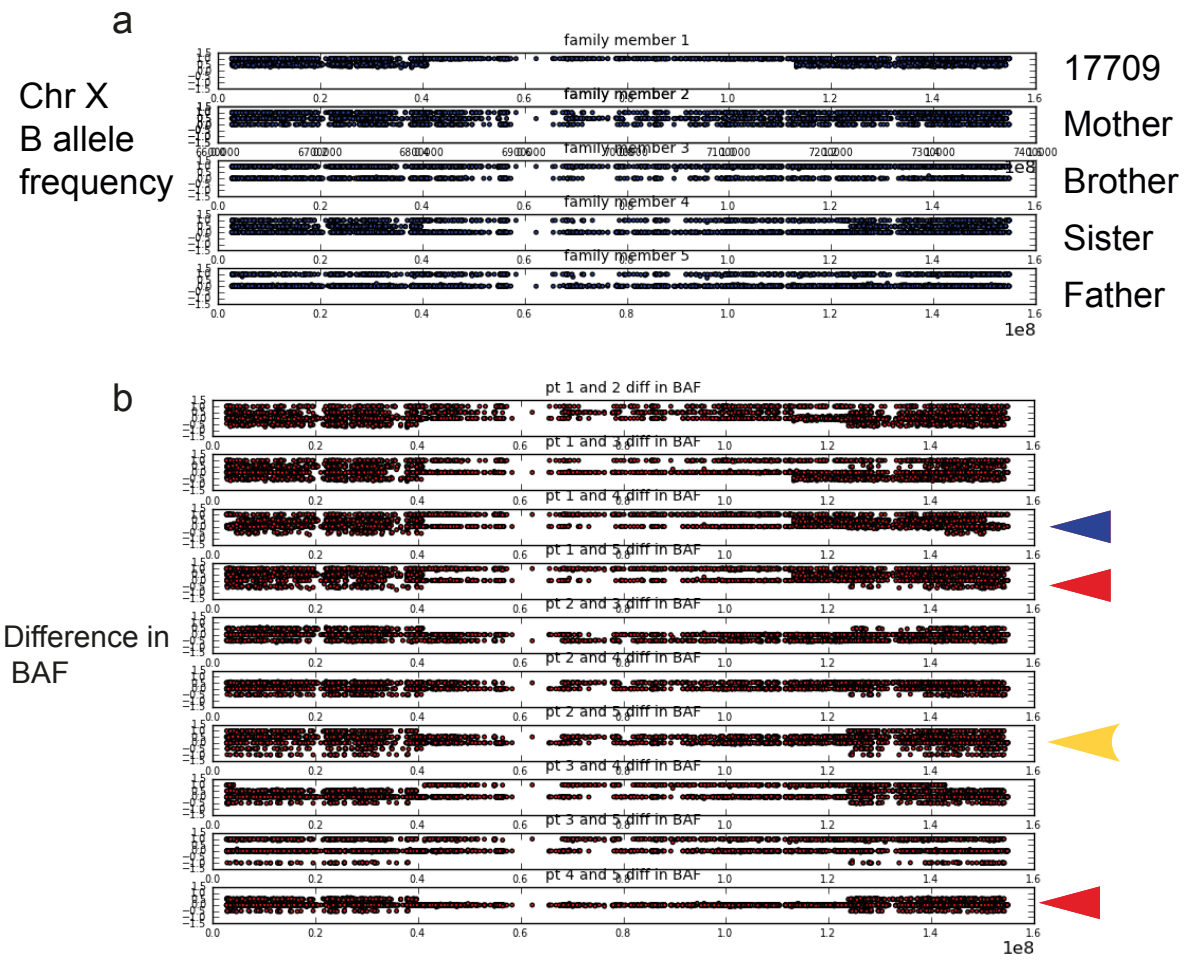


Figure 6.6: B allele frequency analysis of SNP array for 17709 family. (a) shows B allele frequency (BAF) for each variant across the X chromosome, for each family member. The data for patient 17709 comes from the VCF. The base call in the cytoSNP array is not always consistent with the forward allele as used in the VCF file, therefore only variants where the strand can be determined were included. Thus C / G or A / T transversions were excluded. In ROH BAF of 0.5 are absent, this can be seen in the males on the X chromosome and in the ROH regions in the patient and her sister. (b) shows, for individual variants in pairs of family members, the BAF in the first individual minus the BAF in the second. The patient and sister have no BAF difference of 0.5 within the ROH, and predominantly a BAF difference of 0 (blue arrow). Some SNPs with a BAF difference of 1 are seen reflecting strand differences in the VCF data compared to the SNP array data. Both patient 17709 and her sister have BAF differences of 0 (shared common / reference alleles) across the ROH when compared with the father, indicating that they share a haplotype with each other and the father (red arrows). The father and mother have BAF differences of 0 or 0.5 indicating that they share one allele across the ROH (yellow arrow).

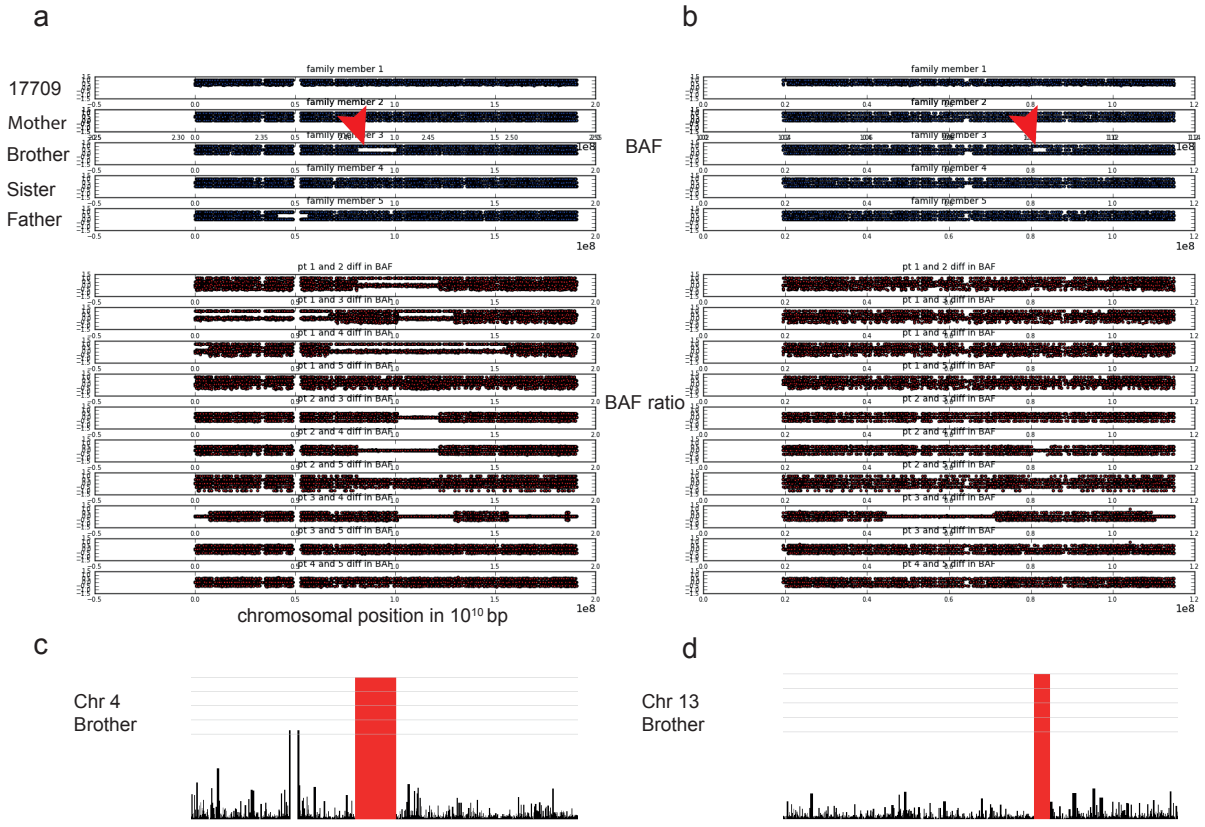


Figure 6.7: Regions of homozygosity in the brother of patient 17709, shown by B allele frequencies (BAF) of 1 or 0 without intermediate values on chromosome 4 (a) and chromosome 13 (b), these regions were also detected by homozygosity mapper as chromosome 4 81073672 to 101238386 (c) and chromosome 13 80,232,942 to 84,035,506 (d), additionally the homozygosity mapper detected 3 other small regions on chromosomes 1, 2 and 6 in the brother which were not easily distinguished by looking at theBAFplots (not shown).

tient's WGS, were input to a homozygosity mapping tool (Seelow et al. 2009). This confirmed the regions visible by BAF analysis in the brother on chromosomes 4 and 13 (Figure 6.7c) and Figure 6.7d), and identified additional regions not evident by simple visual inspection of the BAF plots.

Homozygosity mapping identified 33, 5 and 10 homozygous regions in 17709, the brother and sister respectively, a combined size per individual of 51,556, 30,143 and 31,051 kilobase pairs (kbp) (Figure 6.8a, Figure 6.8b and Figure 6.8c). This suggests that approximately one percent of the genome is homozygous in the brother. In

the sister and patient 17709 the size of the homozygous region additionally includes the 74Mb or larger region on the X chromosome, not included in the homozygosity mapper analysis, so that the total homozygous proportion of the genome is closer to 3.5%. This is larger than would be expected to occur under linkage equilibrium in individuals not sharing a common ancestor (Karl W Broman 1999) (section 6.3.2.1).

Interestingly, both parents in this family have detectable homozygous regions comprising 27,449 kbp in the mother (Figure 6.8c) and 34,710 kbp in the father (Figure 6.8f), suggesting multiple generations of both maternal and paternal consanguinity. Analysis of the WGS data from patient 17571, not known to have any ancestral consanguinity, revealed only 2 possible homozygous regions with a combined size of 1357 kbp, or 0.05% of the genome, this control indicates that the larger regions of homozygosity observed in the 17709 family are indicative of consanguinity (Figure 6.8a).

Beyond the X chromosome region, only one of the 33 homozygous regions predicted in 17709 (Figure 6.8b) could be excluded by examining regions in other family members. A homozygous region on chromosome 19 between 23,286,740 and 28,656,530 was observed in 17709, her sister and father. On average 25% of the genome in the siblings should be IBD for both alleles. The lack of consistency in regions outside the X chromosome between the patient and her siblings may be because WGS SNPs are not equivalent to the more reliable array SNPs, however the brother and sister, both analysed using the same array share only 2 small (1-2Mb) regions on chromosomes 2 and 6.

The size of homozygous regions (ROH) corresponds both to the local recombination frequency and to the type of underlying ancestral relatedness, with short regions due to founder effects or distant shared ancestry and longer regions due to more recent parental relatedness (Trevor J Pemberton 2012). It can be argued that short ROH are less likely to carry a pathogenic phenotype since they will have undergone purifying

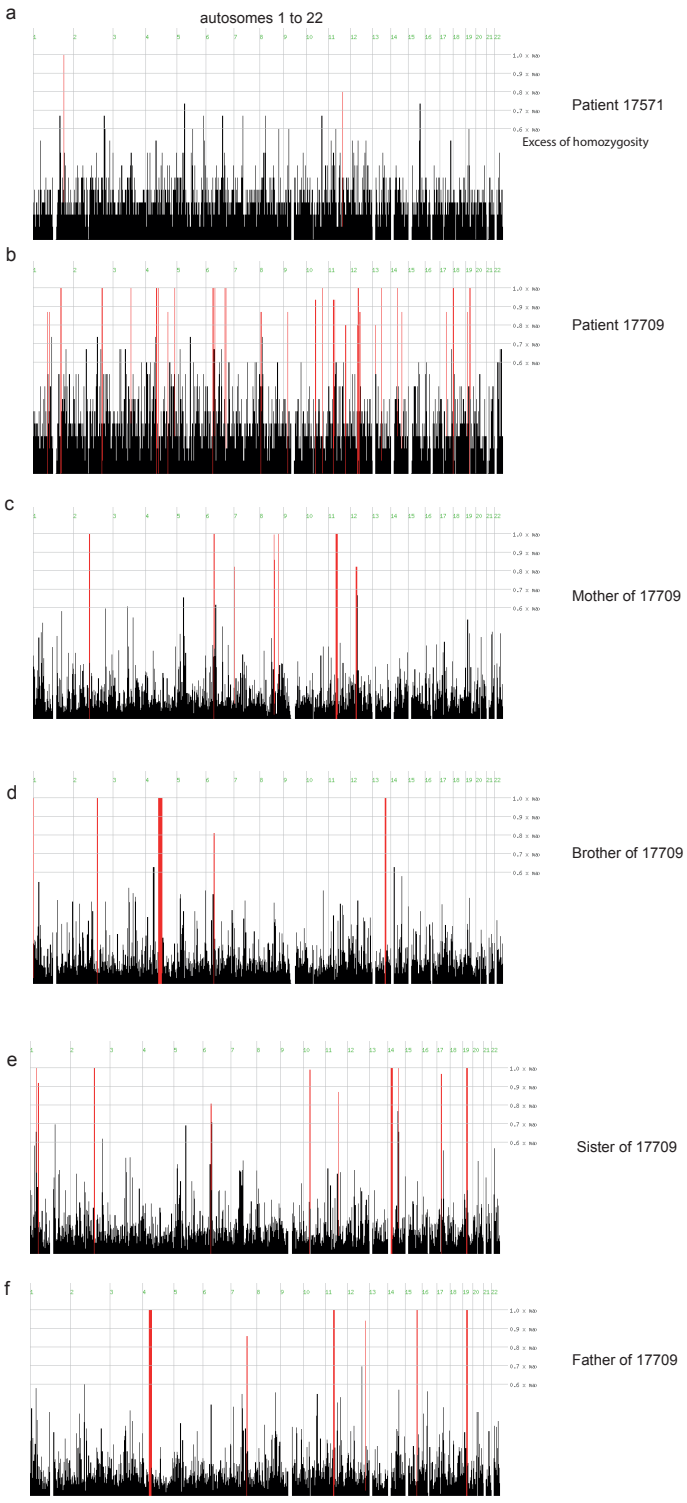


Figure 6.8: ROH across the genome in 17709 and family. Plots across all autosomal chromosomes showing excess of homozygosity against expectation calculated from population values. Regions with homozygosity exceeding the default threshold of 80% of the maximum are indicated as red bars.

selection over many generations. This has been demonstrated experimentally using WES to show that longer ROH are relatively enriched for damaging compared to neutral variants (Zachary A Szpiech 2013). Furthermore very short ROH predicted from the SNP data may reflect a common deviation from the reference haplotype or a spurious ROH due to errors in variant calling. Therefore in patient 17709 only regions of greater than one Mb were considered candidate regions for a causative mutation and searched for deleterious variants.

The homozygous autosomal regions in 17709, excluding those less than one Mb and the region on chromosome 19, comprise 3,740 Mb and contain 552 annotated genes (628 before excluding regions). The variant call file was queried for coding variants in these 552 genes. 420 exonic or splicing variants were identified, 69 had a 1000G frequency of less than 0.05. Excluding synonymous substitutions left 33 variants (Table 6.10), only 3 were homozygous. These 3 are in ankryin repeat domain containing 45 (ANKRD45), homeobox protein HOX–D12 (*HOXD12*) and sugen kinase 233 (*SGK223*).

Gene	Loc	Ref	Sub	type	MAF	pp2	genotype	NR	NV
ANKRD 45	1 173616096	G	C	NS	0.0014	0.993	1/1	46	46
ACSM4	12 7477100	G	A	NS	0.03	unknown	0/1	52	26
CSNK1A1L	13 37678865	C	T	NS	0.01	0.032	0/1	37	18
CSPG4	15 75982271	C	T	NS	0.03	0.638	1/0	41	7
CSPG4	15 75985469	A	C	NS	0.03	0.95	0/1	40	7
ACSBG1	15 78471034	C	T	NS	0.0032	0.002	0/1	51	25
PCSK6	15 102029685	GG	CG, GGCG	unknown	unknown	unknown	2/2	8,8	3,5
ACSM5	16 20442346	A	C	NS	0.0046	0.004	1/0	43	19

ACSM2B	16 20559777	T	C	NS	0.04	0.005	1/0	34	17
ACSM1	16 20635521	A	G	NS	0.01	0.558	0/1	38	14
ACSM1	16 20681245	A	C	NS	0.0046	0.774	1/0	35	18
TNRC6A	16 24801638	A	G	NS	0.0046	0.002	1/0	45	24
ACSF3	16 89178517	AAGT TCTG AAACG	A	Non frameshift deletion	unknown	unknown	1/0	43	0
ACSF3	16 89178531	C	T	NS	0.0018	0.363	1/0	47	0
TEKT1	17 6704121	C	T	NS	0.04	0	1/0	37	17
BRCA1	17 41246481	T	C	NS	0.03	0.977	1/0	74	42
SLC38A10	17 79225026	C	T	NS	0.04	0.365905	0/1	34	8
SLC38A10	17 79225040	C	G,T	NS	unknown	0.399878	0/2	36,36	1,4
SLC38A10	17 79225056	T	A	NS	0.02	0.508108	0/1	39	11
SLC38A10	17 79225330	C	A	NS	0.05	0.956	0/1	31	16
SLC38A10	17 79226264	T	C	NS	0.05	0.891	1/0	39	20
CSNK1G2	19 1979811	A	G	NS	0.04	0	0/1	26	12
ACSBG2	19 6141593	T	A	NS	0.01	0.001	0/1	62	32
TEKT4	2 95540616	T	C	NS	unknown	0.15	1/0	39	20
HOXD12	2 176964947	G	A	NS	0.01	unknown	1/1	42	42
CASP10	2 202074098	G	A	NS	0.03	0.636	0/1	47	23
BCS1L	2 219526634	G	A	NS	0.01	0	0/1	45	21

CST2	20 23804663	T	A	NS	0.05	0.005	1/0	37	21
PAICS	4 57314612	C	G	NS	0.03	unknown	1/0	47	24
FRAS1	4 79188584	C	T	NS	0.01	unknown	0/1	47	27
C6orf132	6 42074421	A	G	NS	unknown	unknown	0/1	10	0
C6orf132	6 42074442	A	G	NS	unknown	unknown	0/1	4	2
SGK223	8 8234868	C	CGC CGCT	Non frameshift insertion	unknown	unknown	1/ /1	46	46

Table 6.10: Rare variants in 17709 in the ROH not predicted to be shared with family. Homozygous variants are highlighted in bold. NS = Non-synonymous. Loc is chromosome : position , NR = number of reads NV = number of variant reads. MAF is 1000 genomes minor allele frequency, unknown MAF indicates variant not seen in 1000 genomes, pp2 = PolyPhen-2 score, unknown pp2 indicates no score available.

The insertion in *SGK233* was seen in only one read; inspection with IGV revealed that reads in this region carry multiple variant base calls and the variant T allele is present at low frequency in other non SLE individuals, strongly suggesting that the variant is not relevant for SLE.

The NS variant in *HOXD12* has a MAF of 0.01, this gene is important in morphogenesis and mice with an ENU induced NS mutation exhibit forelimb defects (Cho et al. 2008). The variant has a 1000 genome subpopulation MAF of 0.02 in Europeans, and is present as a homozygote in two individuals in the ESP6500 European–American population, indicating that this variant is insufficiently rare to be causative for SLE.

The NS mutation in ANKRD45 is rare, with a MAF from 1000 genomes of 0.0014, (never reported as homozygous) and not called in any other WGS500 patient. Polyphen–2 predicts the variant to be deleterious with a score of 0.993. Little is known about the function of this protein coding gene, the ankryin repeat domain is a common motif with a diverse range of functions (Sedgwick and Smerdon 1999).

Sanger sequencing of the ANKRD45 variant in the 17709 family members confirmed the SNP in the proband but also demonstrated that the patient's unaffected sister carries this variant as a homozygote, excluding it as a single candidate (Appendix K).

Therefore no convincing homozygous candidate remains in patient 17709.

6.3.3 Compound heterozygous variants

A loss of gene function may occur due to two different variants occurring in the same gene, with a compound effect. A phenotypic consequence due to both variants is most likely to occur if those variants are in trans –one variant carried on each allele–, since otherwise one allele is normal and only gain of function or haploinsufficiency would explain a phenotype. As with the homozygous candidate variants, these compound heterozygous variants would be expected to be rare, but could be present individually at low frequency in the population.

The variant files were searched for pairs of rare variants within the same gene observed in the same individual. To identify the subset of rare, potentially functional variants, pairs of variants encoding for splice site, non-sense or miss-sense changes in the same gene were considered if the product of the 1000 genome MAF of both variants was less than or equal to 0.01. This latter value serves as an estimate for overall rarity, but does not take into account population differences in MAF and the possibility that the two variants are in linkage within a shared haplotype. Variants that were in segmental duplication regions or had a benign PolyPhen-2 score, less than 0.15, were excluded.

Manual inspection with IGV revealed that many variant pairs arose from variants called within the same reads, often reads of low quality and with no read mate pair mapped or mapped very distantly or on other chromosomes, indicating not only that these variant pairs are in cis but also that the reads are likely miss-mapped. To eliminate such uninformative variants, pairs were excluded if they were less than 50

bp apart (in the context of 150 bp reads), this excluded a mean of 20% of variant pairs in each patient, with a mean distance of 20 bp apart.

2/3 of variant pairs exactly 50 bp apart could potentially be located on the same 150 bp read, although many will in fact be on overlapping reads. The chance that two variants occurring within a gene will be less than 50 bp apart depends on the size of the gene. Gene size has a non–normal distribution (Figure 5.13). However taking the mean average gene exon sequence length as 145 bp with 8.8 exons per gene (Lander et al. 2001), the probability that a second variant lies within the same exon is approximately 11%, and the chance that it is in the same exon and lies within 100 bp surrounding the first would be approximately 8%. The discrepancy between this figure and the 20% of apparent rare compound heterozygous pairs located within 50 bp in the patients indicates that they include spurious variants due to clustering of variant errors. This may be in the same reads or different reads.

Many rare pairs were found in very large genes or genes with multiple paralogues such as the mucin gene family. Due to gene size, and in some cases also the presence of multiple homologous genes, multiple rare variants will occur by chance within these genes in many individuals.

Therefore if more than 10 unique pairs meeting the criteria described above were found within a single gene, representing at least 5 unique rare variants, these were excluded. This filter alone removed 68% of the initial compound heterozygous candidates, with a mean of 33 pairs of variants per excluded gene.

Next, the pairs of variants were checked against the rest of the WGS500 cohort in two ways. Firstly the pairs were checked against each individual in the cohort excepting the SRNS patients and cancer genomes. Pairs of compound heterozygote variants observed together in other individuals within the project were flagged, however because the phasing of these variant pairs is unknown, only pairs observed 3 or more times in non SRNS/ SLE individuals from the WGS500 project were excluded. Pairs

observed together in one or two individuals were flagged as low frequency pairs within the cohort, since is conceivable that the variants could be in cis in the non-SRNS / SLE patients and in trans in the SRNS / SLE patient, with different phenotypic consequences.

It was noted that many variants were not called in other individuals in the individual VCF files, but were identified as being present in many individuals in the cohort in the union file of variants across all the WGS500 patients. The union file variants are genotyped simultaneously across all individuals, and thus some low frequency variants are called in many individuals, though they would not reach the threshold to be called in single genomes.

The calls observed in multiple individuals in the simultaneously genotyped union file may represent common variants or systematic errors with low allele frequency in individuals, in either case these are not candidates for rare compound heterozygous variants causing disease. Hence as a second method of comparison with the WGS500 data, the pairs of variants were checked against genotype frequencies in the union file. Variant pairs were excluded if both variants were observed at homozygotes in non-SRNS individuals, since it was assumed that neither could be pathogenic if tolerated as homozygotes. Pairs with one homozygous variant and a population MAF within the WGS500 union file of greater than or equal to 0.05 for the heterozygous variant were excluded, since the heterozygous variant was assumed to be the more deleterious, and therefore expected to be rare. Finally pairs in which the product of the WGS500 union file population MAF of each variant was greater than 0.01 were excluded, on the assumption that, although MAF will vary by population, such variants appear too common to co-exist within populations and cause rare disease.

The remaining variants were filtered, as described in chapter 3.2, for allelic bias, and were annotated if either variant had a homozygous genotype in the SRNS patient, or had failed to pass all variant quality filters in the variant caller. Unlike in the ENU

datasets, structural variants or indels could underlie disease in the patients. Because of the difficulty in mapping and calling indels these may appear to have unexpected allelic distribution within an individual. However only two indels were excluded within the compound heterozygous filtering pipeline for SRNS, one in an olfactory receptor and the other in WDR66. Both were excluded because they failed quality filters within the Platypus variant caller rather than because of allelic bias.

The numbers of putative compound heterozygous pairs per SRNS patient, after each filtering step, are shown in Table 6.11. The phasing of these variant pairs as yet remains unknown, pending verification with parental DNA. It is expected that on average 50% of the variant pairs will be excluded once phase has been established.

Patient	Number of rare putative compound heterozygous pairs	Number of rare putative compound heterozygous pairs, after filtering for >= 10 pairs / gene or < 50bp apart	After filters against WGS500 individuals and union	After allelic bias, hom genotype and PASS filters
0001	850	84	46	29
0002	863	50	20	12
0003	772	119	64	49
0004	1173	85	25	17

Table 6.11: Numbers of putative compound heterozygous pairs per individual after each filtering step. After filters for rarity, to exclude spurious pairs due to variants on the same read or in genes with multiple variants, after filtering against the union file and all non-cancer non SRNS individuals in the WGS500 cohort and after additional quality filters based on allele bias and genotype.

In SLE, compound heterozygous variant were filtered in the same way, but variants in the siblings were required to be present in both individuals for inclusion (Table 6.12).

The siblings 39124 and 26106 are expected to share a genetic cause for disease, and have no family history or evidence from the WGS data for consanguinity. A dominant mutation could be causative only if the variant was incompletely penetrant,

Patient	Compound het pairs
16603	32
46154	27
16743	236
17709	67
22564	32
17571	157
45629	19
45742	14
39124	240
26106	246
Sibs 39124 and 26106	130

Table 6.12: Numbers of putative compound heterozygous variant pairs in the SLE patients

for example if the parent carried a protective modifier not inherited by the sisters, or if the parent was mosaic for the variant, but passed it to both offspring as a germline mutation (Wallis et al. 1990).Therefore a compound heterozygous variant is the most plausible monogenic inheritance pattern in these patients. The 130 putative compound heterozygous variant pairs shared by both siblings were examined in more detail. The 130 variant pairs were prioritised by PolyPhen–2 score, type of variant, and rarity in WGS500. This highlighted 13 variant pairs in which both variants were predicted as deleterious, and / or at least one variant was a splicing, stop or indel change, and WGS500 heterozygous frequencies of the individual variants were low (below 0.05). These variants are annotated and discussed in Appendix I.

24 further variant pairs did not have PolyPhen–2 scores but were rare in WGS500 individually and cannot be excluded prior to obtaining phase information from parental DNA.

Putative (un–phased) compound heterozygous SRNS variant pairs were annotated for any mouse model or known human disease association, expression levels in human glomeruli, differential transcription in murine podocytes and Gene Ontology terms (Botstein et al. 2000). Expression data is not always reliable, for example Collagen

type IV alpha 4 (*COL4A4*), a well known component of the GBM, was reported as not detected in glomeruli. Mouse models or human disease associations may depend on the type and location of variants, thus variants were simply annotated rather than excluded based on this data (Methods and Appendix H).

6.3.4 Heterozygous rare variants

Some syndromic forms of SRNS with onset in childhood are inherited in an autosomal dominant (AD) fashion, for example AD mutations in Wilms tumour 1 (*WT1*), the cause of Denys–Drash syndrome and Frasier syndrome, manifesting with male pseudohermaphroditism, nephroblastoma (Wilms tumour) or gonadoblastoma and early onset of end stage renal failure (Pelletier et al. 2014; Barbaux et al. 1997). The dominant negative phenotype in *WT1* may be due to association of the mutant protein with the wild type protein resulting in sequestration of both forms to an alternative sub-nuclear compartment favoured by the mutant protein (Englert et al. 1995).

Other AD forms of SRNS are due to mutations in *ACTN4* causing slowly progressive renal failure typically in early adulthood (Kaplan et al. 2000), and *TRPC6* AD variants, resulting in both adult and childhood onset SRNS (Heeringa et al. 2009).

Heterozygous rare variants were examined as candidates for an AD genetic cause of SRNS or SLE with incomplete penetrance or a possible de novo mutation, pending confirmation with parental DNA. A de novo seems more probable since early onset and disease severity suggest penetrance is likely to be complete. All variants were filtered for rarity using a 1000 genomes MAF of 0.01 or less as a threshold for inclusion. Variants from all patients were filtered against the WGS500 union file, being excluded if found in any non–disease patients in the cohorts. The variants were further filtered for allele bias, such that heterozygous variants with alternate allele frequency equal to or below 10% were excluded. Variants with a benign PolyPhen–2 score of less than 0.15 were also excluded.

Table 6.13 shows the numbers of rare heterozygous candidates altering protein sense per SRNS individual. As previously, variants in patient 0001 were additionally filtered against the Punjabi 1000 genomes cohort.

Patient	heterozygous candidates	NS	Stop	Splice	Indel	Unknown	Unique to Patient
0001	323	250	14	25	27	7	309
0001 (after Punjabi 1000G filter)	273	215	11	17	24	6	248
0002	144	96	3	25	19	1	98
0003	361	293	11	33	15	9	173
0004	182	133	1	26	17	5	135

Table 6.13: Heterozygous rare variants in the 4 SRNS patients

The numbers of rare heterozygous SLE variants with their distribution by type are shown in Figure 6.9. In the case of the two siblings a de novo heterozygous causative variant is less likely since we assume that both siblings share a causative genetic defect.

Due to the large numbers of candidates isolating a causative heterozygous variant or short list of candidates, for example based on known functional information about the individual genes or variants, is not practical at this stage. Variants that appear private to the patient, not reported in any other database, would be more likely to include a de novo mutation. The final column in Table 6.13 gives the numbers of such apparently unique variants per SRNS patient. A next step in identifying de novo mutations would be to examine the parents of these patients for these rare variants, particularly the previously unreported variants.

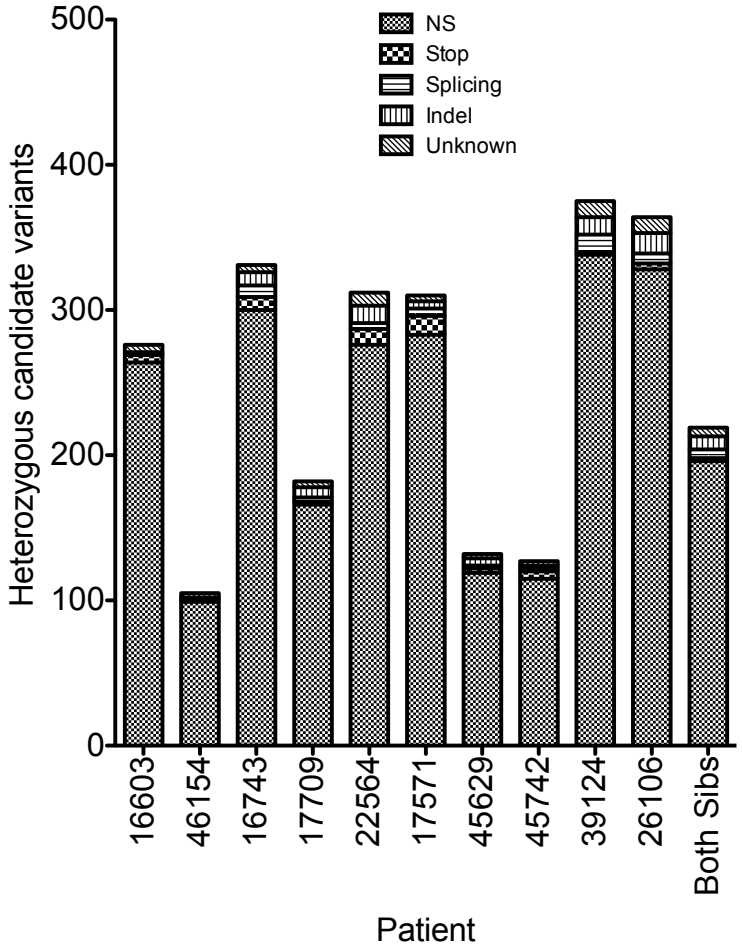


Figure 6.9: Heterozygous variants in the SLE patients by type

6.3.5 Variants in known SRNS or SLE genes

Defects in at least 32 genes have been demonstrated to cause SRNS or FSGS in humans (McCarthy et al. 2013; Kopp 2013). Coding variants in these genes should have been detected by the exome capture sequencing previously performed by collaborators (McCarthy et al. 2013) however exome sequencing does not detect non-coding variation, and can sometimes miss coding variants due to the heterogeneity of coverage (Sims et al. 2014).

A list of 89 genes linked to SLE was collated from the literature (Gateva et al. 2009; Harley et al. 2009; Shizhong Han 2009; Cui, Sheng, and Zhang 2013; Deng and Tsao 2010).

The patients were searched for coding or splicing variants in these genes; the threshold for rarity was less stringent than in the genome–wide variant searches described above.

Variants were included if they were non–synonymous, non–sense or splicing variants in one of the 89 SLE associated genes or 32 SRNS genes as applicable (Methods), with exclusions for Polyphen–2 score less than 0.15 or variants in regions of segmental duplication. Heterozygous variants were excluded if the 1000g MAF was greater than 0.1, or if observed in the WGS500 union file in non–SLE or non–SRNS patients as appropriate, and homozygous variants were excluded if observed as homozygous for the alternate base in non–SLE / SRNS patients in the WGS500 union file.

Numbers of rare variants annotated as in or near the known genes are shown in Table 6.14. Of the 6 exonic SRNS variants identified by the exome capture, all were called in the WGS dataset, but 4 (*INF* and *NPHS1* in 0001, *NPHS1* in 0002 / 0003 and *PMMS1* in 0004, see Table 6.2) were excluded as they were observed in non–SRNS individuals in the WGS500 cohort. No additional exonic coding variants were identified.

8 exonic variants passing these filters were identified in known SLE associated

genes, across 6 of the 10 SLE patients (Table 6.15). The *TNXB* variant in patient 17571 has a MAF from the 1000 genomes Asian population of 0.03, including homozygotes, and cannot be causative for monogenic SLE.

The remaining 5 variants are either 'unique', not reported in any other cohort (4 variants, in *TYK2*, *IRAK2*, and *KLK14*) or rare, in some cases reported in dbSNP without MAF information. *KLK14* was shortlisted because the Kallikrein (*KLK*) genes are linked to SLE but no SNP in *KLK14* itself has been associated with SLE (Liu et al. 2009).

Lupus patients, and lupus prone New Zealand black / New Zealand white (NZB / NZW) mice have reduced serum DNASE1 activity compared to controls, and *Dnase1* knockout mice develop glomerulonephritis, however this heterozygous splicing variant in patient 17571 has a population frequency of 0.004.

None of these rare variants in SLE are homozygous and thus an important next step will therefore be to check these variants in the parents of the sequenced patients.

Patient	Not in Union	Exonic	Splicing	Intronic	UTR	Upstream/ down-stream	Intergenic
0001 SRNS	250	2	0	91	1	0	156
0002 SRNS	111	0	0	48	0	1	62
0003 SRNS	280	0	0	104	4	3	169
0004 SRNS	131	0	0	34	1	0	96
16603 SLE	98	2	0	30	0	1	65
46154 SLE	107	1	0	32	1	7	66
16743 SLE	128	0	1	41	1	1	84
17709 SLE	108	0	0	44	1	0	63
22564 SLE	117	0	0	40	0	6	71
17571 SLE	118	2	0	44	2	2	68
45629 SLE	97	1	0	30	0	0	66
45742 SLE	126	0	0	32	0	4	88
39124 SLE	115	3	0	33	0	1	78
26106 SLE	136	3	0	56	0	1	76
sibs 39124 and 26106 SLE	98	2	0	30	0	1	52

Table 6.14: Numbers of rare variants in genes known to be associated with SRNS or SLE repectively

patient	chr	pos	zygosity	gene	type	dbSNP, MAF	PP2	Other
16603	19	10475628	1/0	TYK2	NS	not seen	0.692	
16603	3	10219595	1/0	IRAK2	NS	not seen	0.961	In death domain
17571	16	3706185	1/0	DNASE1	Splicing	rs8176928, 0.0037	1	
17571	6	32032628	1/0	TNXB	NS	rs140770834, 0.01	unknown	MAF high
26106 and 39124	19	51584856	1/0	KLK14	NS	rs373793602, 0.001	0.997	
26106 and 39124	2	213886748	1/0	IKZF2	NS	rs79632789, 0.0041	0.673	
45629	6	32713181	1/0	HLA-DQA2	NS	rs148573253 0.001	0.764	
46154	6	32038027	1/0	TNXB	NS	rs200125936, MAF unknown	unknown	

Table 6.15: Coding variants in known SLE genes in the SLE patients. Chr is chromosome, Pos is position, zygosity 1/0 is heterozygous genotype and 1/1 is homozygous genotype. NS is non-synonymous, MAF is the 1000 genomes minor allele frequency, and PP2 is Polyphen2 score, unknown if no score available.

6.3.6 Variants in non–coding regions of known SRNS genes

Non–coding variants are increasingly recognised as having potential to cause disease. For example via disruption of splicing, disruption of translation regulating regions in 5' un–translated regions of messenger ribonucleic acid (mRNA), mutations in trans regulatory RNA , or variants in promoter or enhancer regions (Ward and Kellis 2012).

Testing all the rare non–coding variants in the SRNS and SLE cohorts for such damaging variants is beyond the scope of this thesis, however an exploratory study was performed for unexpected splicing consequences of non–coding variants linked to genes implicated in SRNS. These variants were checked for disruption or generation of novel splice donor or acceptor motifs (Methods). The latter can occur beyond the expected splice sites near intron–exon boundaries due to the introduction of deep 'cryptic' splice sites (Dehainault et al. 2007).

Table 6.16 shows those non–coding variants near or within known SRNS genes that appear to disrupt or induce a splice site.

Patient	Gene	Chr	Position	Predicted splicing change	Score	Fwd/ Rev strand
0001	PLCE1	10	95923505	loss of acceptor	0.85	fwd
0001	TRPC6	11	101397502	gain of acceptor	0.51	rev
0001	MYO1E	15	59627769	loss of acceptor	0.9	rev
0001	COL4A4	2	227896135	loss of donor	0.77	rev
0001	COL4A3	5	228164875	loss of donor	0.96	fwd
0001	MYH9	22	36740669	loss of acceptor	0.81	rev
0001	ARHGAP24	4	86403549	gain of donor	0.7	fwd
0001	ARHGAP24	4	86543303	loss of donor	0.78	fwd
0001	ARHGAP24	4	86791759	gain of acceptor	0.78	fwd
0002	TRPC6	11	101328295	loss of acceptor	0.65	rev
0002	MYH9	22	36680774	loss of acceptor	0.54	rev
0003	TRPC6	11	101341584	loss of donor	0.47	rev
0003	TRPC6	11	101384216	loss of acceptor	0.43	rev
0003	TRPC6	11	101454658	loss of donor	0.71	rev
0003	MYO1E	15	59654111	loss of donor	0.5	rev
0003	ITGB4	17	73736589	gain of donor	0.5	fwd
0003	MYH9	22	36736288	gain of acceptor	0.64	rev
0003	LMX1B	9	129410972	loss of donor	0.68	fwd
0003	LMX1B	9	129410972	gain of acceptor	0.44	fwd
0004	CD2AP	6	47479356	gain of donor	0.94	fwd

Table 6.16: Non–coding variants near or within known SRNS genes predicted to disrupt or induce splicing. Based on predictions from NNSplice (Reese et al. 1997) with a default minimum score for a splice site of 0.4 (range between 0–1). Chr is Chromosome Fwd/ Rev is forward or reverse strand

6.4 Summary of Chapter

This chapter highlights the challenges of WGS in isolated cases of two diseases that have complex genetics but can also be monogenic, in cases chosen for enrichment for the single gene forms.

The results in patients 0001 (SRNS) and 17709 (SLE) demonstrate that WGS data can be used to confirm or reveal familial consanguinity, and show how this information, particularly in conjunction with family haplotype knowledge obtained from a SNP array, can point to regions of homozygosity that may harbour a causative variant (Section 6.3.2) . However 17709 also illustrates the challenges in interpreting homozygous regions in the genome, and the importance of obtaining family DNA to validate candidate variants and putative genetic models.

Use of union file of genotypes across many genomes sequenced with the same pipeline permitted exclusion of many variants that appeared rare based on published SNP databases. This is illustrated for the compound heterozygous pairs in the SRNS patients in table 6.11.

The numbers of candidate variants remain large in these unrelated individuals, particularly with non–european ethnicities (Table 6.4 and Table 6.5). Because of the phenotypic and genetic heterogeneity of the diseases studied, particularly SLE, it is not possible to prioritise variants shared between the patients.

The results shown here are short lists of possible candidate variants that generate a substrate for further analysis, in particular using parental DNA to phase variants and pick up de novo mutations. They also illustrate the limitations of WGS in isolated or unrelated cases. Assigning causation to individual variants is not possible at this stage. Knowledge of likely inheritance patterns, linkage information from additional affected family members and greater understanding of underlying biological pathways for the disease process will be critical to the success of similar strategies in the future.

Chapter 7

Discussion

7.1 Murine and Human forward genetics to identify novel disease genes

The objectives of this thesis were to explore the genetics of SLE and nephrotic syndrome using both ENU mutagenesis and extreme trait sequencing in humans. Specifically, the work set out to more efficiently harness forward genetic methods to isolate causative genetic variants or mutations using whole genome sequencing (WGS).

This discussion measures the results against these objectives, and compares two forward genetic approaches in murine mutagenesis and patients with extreme disease traits. In conclusion it proposes future genetic approaches to uncover the missing heritability of human disease.

7.1.1 The future of ENU

ENU mutagenesis is a powerful tool to identify genes and pathways of importance in human disease, generating animal models with point mutations that can be used to explore gene function. However several factors have limited the efficiency of ENU programmes. Identifying the causative variant has until now been a rate limiting and

costly step.

7.1.1.1 Filtering WGS data to pull out ENU mutations

Chapter 3 shows that filtering WGS variants can short–list candidate ENU mutations. This is an important step as it vastly reduces the number of variants under consideration. The most effective filter in terms of numbers of variants excluded is the 'union file' of variants observed in more than one ENU pedigree. This was expected since the union file will contain many variants that are either systematic errors or true fixed variation from the reference B6 genome.

Figure 3.4c and Figure 3.4d show that the 1.93% of variants, not excluded by the union file but identified using another filter, are in the vast majority of cases identified by multiple filters. Given this redundancy, as further mouse genome sequences are added to the union file, this will become an increasing comprehensive set of recurrent, non–ENU variant calls, and one approach would be to filter using only the union file, plus perhaps an indel filter and coverage thresholds.

One issue with adding more sequences to the union file is the risk of excluding a true ENU variant because an identical variant has been called in another pedigree. For two mice this was estimated as a probability of 3.79×10^{-7} (section 3.2.10).

However the chance increases with each additional sequence, so that with 100 genomes in the union file the probability of two mice from different founders sharing an ENU mutation is 3.79×10^{-5}. This is likely to be an underestimate since base preference bias in ENU mutations will increase the probability. This argues caution in using an ever–expanding union file as a filter and supports continued use of other filters as shown in Chapter 3.2 to exclude variants based on their characteristics.

Other than the union file, the indel and coverage filters picked up the most variants not excluded by any other filter (Figure 3.4d). This reflects inaccurate indel calling with current variant calling tools and unreliable variant calls at extremes of

coverage depth, and suggests that a small proportion of such variant calls occur as non-systematic errors seen in single pedigrees. The majority of raw variants were excluded by 2 or 3 filters (Figure 3.4e). The redundancy between the filters suggests that the excluded calls are likely to be spurious or non–ENU.

7.1.1.2 Faster mutation isolation

Chapter 3 shows that, with knowledge of a coarse linkage region, causative ENU mutations for a phenotype can be isolated from low coverage single mouse WGS. Using this approach 2 out of 5 causative mutations were found directly by sequencing single ENU mice, including a novel mutation in *Sppl2a* (Bergmann et al. 2013).

In two other strains the mutation was identified after further sequencing, illustrating a limitation of very low coverage WGS of individual mice for variant identification.

Finally in mouse *222* no causative mutation was conclusively identified.

These results illustrate important points that motivated later work in this thesis. Firstly, filtering WGS variant calls can effectively identify ENU mutations from background error and non–ENU variation in mice. Secondly, a proportion (3/5 in this small cohort) of causative ENU mutations are not detected using low coverage WGS in single ENU mice. Increasing coverage by additional sequencing in the same mouse may reveal these missing mutations, but by performing sequencing in additional affected mice from the same pedigree it is possible to improve mutation detection by increasing coverage and simultaneously exclude some mutations that are not shared between the affected mice, thus reducing the search space, a form of linkage analysis.

In Chapter 4 a hidden Markov model (HMM) based method was applied to identify the causative variant in an ENU mouse with a mixed strain background. This identified a mouse model for the milder spectrum of human Pierson syndrome, and shows both that a probabilistic programming algorithm can distinguish ENU inherited haplotypes, and ENU can generate mouse models of proteinuric disease (Bull

et al. 2014).

By harnessing WGS of multiple affected mice within a pedigree, and refining the HMM based approach, the Lander-Green based identity by decent (IBD) method developed and demonstrated in Chapter 5 shows that it is possible to rapidly and simultaneously identify a linkage region and isolate candidate causative variants within the region. A mutation in *Lyn* identified without recourse to conventional linkage in the *ENU16CH17a* strain is a proof of principle for this technique, and the method has subsequently been successfully applied to further ENU pedigrees.

These subsequent results confirm the modelling data, showing that the method is applicable at low coverage depths of 5–fold per genome, making it a cost effective approach for high throughput ENU programmes.

Because the method uses density of variation to identify genomic intervals inherited from the ENU treated ancestors, it does not require out-crossing to another inbred strain for mapping. This is the first demonstration that an ENU causative mutation can be identified without recourse to any outcrossing and conventional linkage mapping (Bull et al. 2013).

The approach eliminates several generations of breeding and thus can improve efficiency and reduce animal costs, as well as allowing rapid mutation identification once a phenotype has been detected in the G_3 generation. Such G_3 mice can be immediately sequenced, and the results processed automatically using the Lander–Green based algorithm.

This will allow investigators to focus their efforts on pedigrees with a novel variant or gene, and avoid time currently wasted breeding and mapping pedigrees that are subsequently found to carry a duplicate of a well established or previously reported mutation. Prompt isolation of causative mutations will provide immediate feedback on the specificity and effectiveness of phenotyping screens, screens that fail to discriminate for the biological pathway of interest or only pick up mutations in known

disease genes can be modified or redesigned.

Alternative methods to accelerate the gene discovery step in ENU mutagenesis programmes have been suggested. WES in single mice combined with either coarse linkage mapping data, or sequencing of the G_1 founder, have been used to identify causative ENU variants (Arnold et al. 2011; Andrews et al. 2012). However the first approach still requires outcrossing to another laboratory strain with the risk of confounding the phenotype, the need to propagate mice and additional breeding. The second method introduces an element of investigator bias as it relies on spotting likely candidate variants in the G_1 founder and then looking for these in affected G_3 mice. This will be most successful for well-characterised genes suspected to lead to phenotypes, but novel and poorly characterised genes are less likely to be followed in G_3, although these would be the most interesting mutations to link to phenotypes.

7.1.1.3 A more accurate estimation of the ENU mutation density

The IBD method implemented in Chapter 5 facilitated an estimation of the ENU mutation density.

This improved method of defining ENU genomic regions allows estimation of the ENU mutation density based on a reliable set of homozygous calls (Figure 5.7), rather than heterozygous calls in G_1 mice. The measurement is based on much a much larger genomic sample size than previous estimates (Russell et al. 1979; Beier 2000; Coghill et al. 2002; Quwailid et al. 2004; Concepcion et al. 2004; Takahasi, Sakuraba, and Gondo 2007; Nolan, Hugill, and Cox 2002) due to the ability to calculate an ENU density across the whole of the ENU homozygous genome.

This estimate is applicable to a dosing regime of 3 doses of 90 –100 mg/kg ENU in B6 mice (Methods). The ENU mutation density is dose related (Justice et al. 2000) and so could differ if the administration of the 3 doses was variable. However the rate estimate was consistent across 3 ENU pedigrees from different ENU treated founders,

at 1.5 mutations Mb^{-1} (CI 1.36–1.62).

This estimate is higher that the one mutation per Mb sometimes quoted in the literature (Cook, Vinuesa, and Goodnow 2006; Daxinger et al. 2013), however this estimate is based on relatively smaller studies with inaccurate methods of mutation detection (Quwailid et al. 2004) or wide confidence intervals (Concepcion et al. 2004).

7.1.1.4 Modelling an efficient ENU programme

The ENU mutation density calculated above was incorporated into a model of the distribution of mutations within a pedigree.

A fully penetrant phenotypic mutation would be observed in sufficient G_3 mice to allow identification of the causative variant by sequencing multiple affected G_3s. This modelling shows the feasibility of the IBD method.

The IBD method demonstrated in Chapter 5 and used in the modelling is based on sequencing only phenotypically affected mice. An alternative approach would be to sequence both affected and unaffected labelled mice from a pedigree.

This would have the advantage of identifying smaller IBD regions by exclusion of variants/ regions observed in any unaffecteds. However if the disease is not fully penetrant or the screen is not fully sensitive incorrect phenotyping can occur. These sources of error tend to mislabel affecteds as unaffected.

This is in contrast to the ENU zebrafish, in which the heterogeneity from the reference within breeding stocks and large numbers of individuals, both affected and unaffected, available for sequencing, has led investigators to use an HMM method utilizing pools of affected and unaffected mutants to identify regions of homozygous IBD in the affected pool (Leshchiner et al. 2012).

ENU mouse mutagenesis using the IBD method to identify causative mutations has an advantage over zebrafish ENU, because the homogeneous inbred B6 background is used, this avoids modifiers and allows detection of dominant traits.

The Lander–Green algorithm on which our IBD analysis is based, scales exponentially with the number of individuals in the pedigree, but remains computationally feasible with a pedigree up to 14 non founders (Abecasis et al. 2002; Gudbjartsson et al. 2005). The algorithm would accommodate further refinement to take into account the known characteristics of ENU mutations (Figure 5.9d).

By generating haplotype data for many ENU pedigrees, the IBD approach will also eventually lead to a fine scale map of active recombination sites in the mouse, which, unlike existing maps based on recombinations that arose historically between outbred strains of mice (Brunschwig et al. 2012) or more recently between intercrossed inbred strains (Cox et al. 2009), is unbiased by selection or strain differences. Such a map could then be used to optimise the performance of the Lander–Green based algorithm.

7.1.1.5 The IBD method to identify ENU mutations permits more powerful and efficient screening

Efficient screening is essential for a productive ENU programme. Forward genetic screens with ENU can be genome wide or region specific (Probst and Justice 2010). Region specific screens can generate allelic series within a target region, using chromosomal deletions/ balancers or inversions and detecting heterozygous ENU mutations because of allelic non-complementation (Balling 2001; Cordes 2005).

Genome wide screens allow testing for genes involved in a disease model or phenotype across the whole genome so are truly 'hypothesis free'. The simplest genome wide screens look for dominant mutations by screening G_1 offspring of an ENU treated male crossed to a WT female, and / or recessive phenotypes by screening G_3 mice generated by G_2 intercross or backcross of G_1 and G_2 mice within a pedigree. Variations on this basic approach include modifier screens, in which a strain with a pre-existing phenotype is treated with ENU, and offspring are screened for enhancement or suppression

of the phenotype, and sensitised screens, employing ENU on a genetic background predisposed to developing a phenotype (Tchekneva et al. 2007). These sensitised or modifier screens expand the scope of ENU to detect genes despite compensatory pathways or genetic redundancy.

ENU programmes screening for multiple dominant phenotypes per mouse (including congenital malformations, blood disorders and neurobehavioural phenotypes) found a detectable abnormality in 2% of first generation ENU mice (Angelis et al. 2000; Nolan et al. 2000). This rate can be improved upon, recessive screens can detect further phenotypes not evident in heterozygous state and 21% of non-synonymous ENU mutations are estimated to have a potentially detectable phenotype (Arnold et al. 2012). The efficiency of a screen in revealing novel genes will depend on the size of the pathways involved in the phenotype, as well as the proportion of unknown genes contributing to the trait that have appreciable effects in isolation.

A good screen should be specific to pathways or disease of interest to the investigator, sensitive to subtle but relevant biological changes, quantitative and reproducible, with results that can be compared across datasets from different pedigrees. ENU is well suited to the study of diseases with multiple complex and poorly understood pathways, such phenotypes are difficult to study using candidate gene methods but provide a large target region for random mutagenesis.

The need to carry out further breeding for linkage analysis has been a limiting factor on the number and type of screens in ENU programmes to date. Non–lethal screens are used, limiting the depth of physiological screens, often to blood or urine analysis and visible or behavioural traits. The number of procedures that can be performed on each animal is limited, so the breadth of screens may be narrow, failing to detect many phenotypes.

The use of WGS and IBD to perform linkage analysis will permit the use of lethal screening assays, since there is no need to propagate mice for linkage. It would also

be possible to study infertile or subfertile phenotypes, and those with very early mortality or embryonic lethality.

By archiving of sperm, pedigrees with interesting phenotypes and genes could be selectively re-derived for further phenotypic and functional studies as required. It would also be practical to identify carriers and affected mice within a pedigree, avoiding the problems of tracking a phenotype during breeding, thus reducing breeding costs and time, and improving accuracy for functional analysis.

Lethal screens could include flow cytometry based analysis of immune cell populations in multiple organs. Such screens are quantitative, reproducible and statistically robust because they measure thousands of cells, but for a high throughput ENU programme would require efficient methods to collect large amounts of data, such as automated gating. These methods are becoming available and allow more reproducible and less labour intensive data collection (Aghaeepour et al. 2013).

To fully utilise the data, screen results should be stored in an openly accessible automated and curated database that allows detection of variation and clusters of traits between as well as within ENU pedigrees, linking these to ENU gene mutations. Such a database could be mined for subtle or incompletely penetrant phenotypes and could highlight pathways linked to a phenotype by inter-pedigree comparison. Open access would allow investigators to identify genotypes or phenotypes of importance to their own research field. Accessible databases of ENU mutations currently exist (http://mutagenetix.utsouthwestern.edu) and could be enhanced to allow users to manipulate genotypic and phenotypic data across multiple pedigrees to detect associations.

7.1.2 Remaining challenges for murine ENU

The WGS method described has the potential to make ENU a more efficient tool, enabling investigators to move quickly from phenotyping to mutation and gene dis-

covery. However the scope of ENU is limited by the possible screens, the types of mutation generated, and the extent to which observations in mice correspond to human disease.

7.1.2.1 Limitations of screens

ENU identifies only phenotypes that can be detected by screening, and screens need to be simple and highly reproducible. Screens for phenotypes seen under challenge conditions have been used successfully (Richer et al. 2010; Brandl et al. 2009), and behavioural phenotypes have been examined (Oliver and Davies 2012) but the more complex or subjective the screen the more labour intensive a high throughput project. Sensitivity is important, and incompletely penetrant phenotypes may not be detected as in many cases only 2 or 3 homozygous mutants will be available for phenotyping (Figure 5.12a)

7.1.2.2 Relevance to human disease

Mice are widely used for research because of access to inbred strains and their short generation time and lifespan. The mouse is the model organism most closely related to humans, sharing many disease genes with man.

Nephertiti illustrates how the spectrum of phenotypic variation in humans can be demonstrated in the ENU mouse. Hypomorphic *nephertiti* mice carry a miss-sense point mutation and closely mimic the disease spectrum in patients with miss-sense variants in *LAMB2*, who show delayed progression of renal disease and milder or absent extra renal disease in comparison with null variants.

However as with any disease study in a model system, ENU mutants do not always fully mimic the human phenotype, and a human disease trait observed in mice may be due to disruption in genes or pathways that do not play a role in the human disease. Despite this, even in cases where the phenotype differs between mouse and

human, examining sets of orthologous genes in model organisms can provide disease gene candidates of direct relevance for human disease (McGary et al. 2010).

7.1.2.3 ENU does not mimic the full spectrum of human genetic variation

Overwhelmingly ENU induces point mutations and these mimic human variation and can generate hypomorphic mutants and allelic series. However ENU does not generate small or large insertions or deletions, copy number variants or structural rearrangements.

ENU induces mutations across the genome. The majority of published ENU mutations with phenotypic consequences are in coding regions, however ENU mutations in a microRNA and in a long range cis element have been reported (Masuya et al. 2007; Lewis et al. 2009), and this thesis includes an example of a putative cryptic splice site mutation in a deep intronic region that mimics a BAFF null mutant. There is likely to be an ascertainment bias in the current ENU literature due to difficulty in identifying these variants with existing strategies such as Sanger sequencing the exons of selected candidate genes within a linkage region, or more recently whole exome sequencing. More pathogenic non-coding mutations may be discovered in the future, particularly as resources to predict the function of these non-coding changes expand with projects such as mouse ENCODE (John A Stamatoyannopoulos 2012).

Epistatic interactions occur in complex human genetic diseases such as Alzheimer's and diabetes (Combarros et al. 2009; Wiltshire et al. 2006). ENU mutations do not mimic epistatic interactions between multiple variants due to their paucity across the genome in comparison with human genetic variation. The use of sensitised screens and the observation that genetic modifiers from outcrossing strains can confound ENU phenotypes (Moser et al. 1992) suggests a possible strategy to examine gene–gene interactions between single known phenotypic ENU mutations and genetic background. Multiple strain crosses are being examined in the Collaborative Cross (Threadgill and

Churchill 2012), but using ENU to rescue or modify the effects of strain background could provide a comparatively straightforward way to look for modifiers.

DNA Methylation patterns differ between pairs of twins discordant for SLE (Javierre et al. 2010), and epigenetic effects also play an important role in observable mouse phenotypes, the agouti mouse being perhaps the most studied example (Duhl et al. 1994; Dolinoy, Huang, and Jirtle 2007). A sensitised ENU screen for variegated GFP expression in erythrocytes has been used to detect mutations in genes required for epigenetic regulation (Daxinger et al. 2013). This is an attractive area for further ENU research, examining for example parent of origin effects, or the consequences of modification of methylation on an ENU mutant.

7.2 Whole genome sequencing to find rare variants causing complex disease in humans

Chapter 6 describes short-listing of candidate variants in patients with SRNS or SLE. The patients are sporadic unrelated cases, with the exception of a pair of sisters with SLE, with severe early onset disease. This method of extreme trait phenotyping has uncovered rare alleles with large individual affect size in a population with low levels of HDL cholesterol (Cohen et al. 2004).

The short–listing is based on the hypothesis that the patients harbour a single rare variant of large effect causing their disease. Currently all patients have multiple candidate variants. Two patients, 0001 and 17709, have clear evidence for parental consanguinity, based on runs of homozygosity (ROH) in the genomic variants. These patients are expected to have a homozygous rare variant causing disease and therefore have the smallest number of candidate variants, 5 variants in 4 genes in 0001 (Table 6.9) and no confirmed homozygous variant fitting the phenotype and segregating in the family in 17709.

The finding of ROH illustrates that the WGS can be used to examine possible modes of inheritance. This is particularly powerful when combined with SNP array data to explain inheritance and perform linkage as shown in patient 17709 (section 6.3.2.3).

The results also show that in the absence of a known mode of inheritance, examining the rare variants in a single isolated patient is insufficient to identify a single causative variant. This is true both in a disease in which known genes have well defined tissue expression and in a complex multi system autoimmune disease (Table 6.4 and Table 6.5).

Phasing of the compound heterozygous candidate variants and searching for de novo dominant mutations using parental DNA may yield a single causative variant in some of these patients. This may not be possible in the patients with non–European ethnicity, in which the power to distinguish truly rare variants is reduced, resulting in longer lists of candidate variants.

An assumption for the filtering strategy used was that pathogenic variants would be both rare and lie within coding regions. However intronic variants that disrupt or induce splice sites in known SRNS genes were considered. 19 such sites were predicted across the four SRNS patients. These could be further tested by examining complementary DNA (cDNA) to detect splice variants. Microarray based study of co-immunoprecipitation of RNA and binding proteins can identify active splicing regions (Watkins, Stewart, and Fairbrother 2009), RNA sequencing would provide a more comprehensive picture of expression and can be harnessed to detect alternative transcripts (Li et al. 2014).

The failure of the WGS approach in the SRNS and SLE individuals to identify as yet any single causative variants highlights some of the challenges of using WGS in individuals with complex or not straightforwardly Mendelian disease. These are discussed below.

7.2.1 Accurate phenotyping

The patients selected for this study were obtained via clinicians in other centres, and the study relied on the diagnostic skills of local physicians for phenotyping. In comparison with the murine model approach this introduces more subjectivity.

Phenotypic information is inevitably limited, due to the geographic distribution of the patients and the reliance on sometimes incomplete clinical work up. This was illustrated for patient 17709. Platelet size and count are usually abnormal in classical Wiskott–Aldrich syndrome (Imai et al. 2004), but only limited information about platelet count, not including morphology, was available for the patient, and there was no available haematological data for other family members.

Both these diseases, particularly SLE, are phenotypically heterogeneous, and so the patients collected may reflect different underlying physiological and genetic disease processes. Whilst by design this study did not seek a unifying genetic diagnosis in these unrelated patients, increased heterogeneity risks introducing patients with a different disease process, perhaps not monogenic or even predominantly genetic in origin.

Finally the patients, unlike the mice, experience a wide range of uncontrolled and unknown environmental factors, which may modify their phenotypes.

7.2.2 Challenges in finding genes in unrelated patients with heterogeneous disease

The variants previously linked to non–syndromic SRNS have mainly been in genes expressed in the podocyte (Rood, Deegens, and Wetzels 2012), however they reflect a variety of underlying physiological mechanisms including actin dynamics, transcriptional regulation, calcium signalling, mitochondrial dysfunction in the coenzyme Q10 pathway, or a DNA–nucleosome restructuring mediator (Saleem 2012; Benoit et al.

2010).

SLE is a highly heterogeneous autoimmune disease involving many organs and pathways. This heterogeneity is illustrated by the breadth of the SLE classification criteria (Tan et al. 1982), and the more than 70 genetic loci involved in human SLE (Cui, Sheng, and Zhang 2013; Tiffin, Adeyemo, and Okpechi 2013). In fact SLE could be viewed as a cluster of disorders sharing some clinical features, and greater understanding of SLE genetics may delineate some of these disorders and facilitate more tailored therapeutics. Due to heterogeneity the genetic 'search space' for this study was large.

In SRNS genes were prioritised if they were expressed in the glomerulus or podocyte specifically; however expression data for many genes is incomplete. Caution must be exercised in using glomerular expression as a filter since some forms of SRNS recur post–transplantation and are thought to be due to a circulating, renal extrinsic factor, possibly soluble urokinase-type plasminogen activator receptor (Ashley Jefferson and Shankland 2013).

For SLE very few genes could be excluded based on expression or functional data, only those genes with very narrow expression profiles or those with a well characterised single disease association that would not be expected to present any with features of SLE, for example a *DOCK6* mutation in siblings 39124 and 26106 (Shaheen et al. 2011).

Many gene defects with a known disease association could conceivably present with clinical signs consistent with SLE. This was the case with the Wiskott–Aldrich syndrome gene variant observed in patient 17709. *WAS* variants can present variably, with the classical Wiskott–Aldrich syndrome triad of thrombocytopenia, eczema and immunodeficiency (Aldrich, Steinberg, and Campbell 1954), X–linked thrombocytopenia (Derry et al. 1995) or neutropenia (Devriendt et al. 2001), depending on genotype. Autoimmune disease occurs in 40% of Wiskott–Aldrich syndrome cases

(Sullivan et al. 1994), and this can include renal disease. Thus Wiskott–Aldrich syndrome could mimic SLE with haematological and renal involvement.

Next generation sequencing can reveal unusual presentations of known genetic diseases (Leidenroth et al. 2012). It was postulated that this explained the disease in patient 17709; however familial DNA excluded the *WAS* variant as causative (Figure 6.5).

A specific analysis for shared rare variants was not performed in these patients, this would only detect additional variants not previously short–listed if the threshold for rarity was relaxed. This would not be informative because it would introduce variants that are moderately rare but shared between two or more SLE or SRNS patients by chance or because they arise from a systematic error in the pipeline. Most of these could be excluded by excluding variants called in patients from other WGS500 cohorts. However by chance some calls of this nature will just be seen in two affected patients, and appear to be a shared variant.

The two affected sisters with SLE, patients 26106 and 39124, have no other affected family members. A shared compound heterozygous inheritance is suspected. The sisters share 130 compound heterozygous candidate pairs. Excluding those not shared eliminated 46% and 47% of the compound heterozygote pairs in each sibling (110 / 240 and 116 / 246), in line with expected IBD sharing between sisters, but the number of compound heterozygotes remains high in comparison to other patients in the cohort. This appears to reflect ethnicity; rare variation is less shared across ethnicities than common variation (Mathieson and McVean 2012; Bustamante, De La Vega, and Burchard 2011) and African populations are typically more genetically diverse (Campbell and Tishkoff 2008).

Selecting more distantly related patients with a shared phenotype, or introducing a clearly unrelated family member could narrow the search more effectively, although unrelated relatives can be difficult to identify in SLE because of the variability in age

of onset and presentation. Simultaneous sequencing of the parents can also be used to phase compound heterozygous variants (Kamphans et al. 2013; Glazov et al. 2011).

Phasing in the SLE siblings could be achieved by examining the segregation of the variant pairs between the two parents, since there are 130 pairs this would be most efficiently be achieved by WES of the parents. At the time of writing this thesis parental DNA for patients 26106 and 39124 is being sought.

7.2.3 The importance of access to parental DNA

Apart from the sibling pair, the patients sequenced in Chapter 6 have sporadic disease, and therefore an autosomal dominant cause of SRNS or SLE is not expected. The heterozygous rare variants identified in each patient could theoretically cause the disease, due to haploinsufficiency or a gain of function, if they arose de novo in the patient.

De novo mutations have been reported in genes that are known to cause dominant monogenic forms of both nephrotic syndrome (*TRPC6*) and SLE (*TREX1*) (Gigante et al. 2011; Rice et al. 2007). Alternatively the heterozygous variants could cause disease that is not apparent in the parents due to genetic or environmental modifiers or parental germ cell mosaicism. Germ line transmission of a mosaic variant from a parent has been shown for the glomerular basement membrane disease Alport syndrome (Beicht et al. 2013). Next generation sequencing can detect low frequency mosaics, de novo mosaicism with a variant allele frequency of 18% has been detected using WES in a gene known to cause double cortex syndrome (Pagnamenta et al. 2011), however the difficulty in would be in detecting mosaics across many candidate genes while eliminating low quality variants with allelic bias.

Given the large numbers of heterozygous rare variants per individual in this study (range 105 to 361) excluding these as maternally or paternally derived in bulk using familial DNA would be most efficiently done using WGS or WES of both parents.

This trio based method to find de novo mutations can be very effective (Vissers et al. 2010), but care must be taken to distinguish errors in variant calling from true de novo mutations, and therefore adequate sequencing coverage depth and simultaneous variant calling of all individuals (multi-sample calling), as can be performed with Platypus (Rimmer et al. 2014) or GATK (McKenna et al. 2010) is useful. If WGS is performed on several related individuals, more accurate variant calls can be made if the sequence data is more fully exploited, by using linkage disequilibrium between variants and pedigree information (Zhou and Whittemore 2012) to inform genotype prediction.

On average an individual may have 1-2 coding de novo mutations (Kong et al. 2012; Sun et al. 2012a) so this approach should exclude nearly all the rare heterozygous variants.

WGS or WES of parental DNA would also allow phasing of putative compound heterozygous variants based on their segregation between the parents, to establish whether two variants within a gene affect the same or different alleles (Kamphans et al. 2013). Alternatively parental DNA could be interrogated using Sanger sequencing for each pair of compound heterozygous variants but this would be more laborious given that the patients have between 12 and 236 pairs of candidate putative compound heterozygote candidates. Sanger sequencing would of course be required to confirm candidates after WGS / WES of parental DNA.

If parental DNA is unavailable a minority of putative compound heterozygous pairs might be phased by identifying overlapping reads with additional heterozygous variants that distinguish whether the two candidate variants arise from the same allele (Delaneau, Zagury, and Marchini 2012). However using standard 150 bp paired end reads, this is only possible for variants in close proximity, within the same exon, and with an additional informative heterozygous SNP to distinguish the alleles.

In fact many pairs of variants that lie close together within an exon are on the

same reads, so can be excluded as compound heterozygotes. Pairs of variants less than 50 bp apart were excluded from the analysis. If longer reads are used then more pairs could be phased, 90% of genes can be phased using 7–9kb DNA fragments and statistical assignment of phase using a HMM. This method could be incorporated in to future WGS projects where a compound heterozygote inheritance is suspected. In a typical individual, 47% of genes have compound heterozygous SNPs, and around 2% of genes contain damaging compound heterozygous SNP pairs (Kuleshov et al. 2014).

At the time of writing only familial DNA for 17709 was available and WGS or WES of family members had not been performed on any family DNA. However for the SRNS patients ethics approval for collection and sequencing of family DNA was being obtained by collaborators and this will permit further investigation of these cases. An initial step would be to carry out Sanger sequencing of family members for patient 0001 in which a homozygous causative variant is suspected due to evidence of consanguinity.

7.2.4 How best to design future human studies

The search for genetic causative variants in the SRNS and SLE patients show that with small numbers of unrelated individuals, in diseases with multiple genetic causes, isolating a single causative variant is not straightforward. The search may have been more successful with larger numbers of individuals or by sequencing families with multiple affected members.

These are the two main approaches used to elucidate genetic disease in humans, case control studies of unrelated individuals (including large GWAS), or family studies. Case control studies can incorporate large numbers of unrelated patients with sporadic disease, but are limited by population stratification. This can be minimised by careful matching of cases to controls, for example by using relatives as controls or

by looking for stratification with genetic markers (Cardon and Palmer 2003).

Estimates of association from family based studies are similar in size to those obtained by unrelated case control studies (Evangelou et al. 2006). However in many cases of rare disease, or rare Mendelian forms of more complex disease it may be difficult or impossible to collect sufficient families, trios or relative pairs to power the family based design.

Small numbers of unrelated individuals such as the SRNS and SLE cohorts in this thesis are not sufficient to perform case control studies due to the heterogeneity of the underlying aetiology. However they serve as a pilot to explore the possible pitfalls and challenges of a larger scale sequencing project.

For SLE in particular, larger cohorts of early onset disease are available, since databases of patients have been collected for previous GWAS studies. This offers the potential to combine whole genome sequencing data with genome wide association analysis. Despite this, the power of a sequencing based case control study in SLE would be limited, due to the genetic heterogeneity of the disease and the possible impact of environmental factors (Graham 2009). This would be true even in a population enriched for monogenic disease by selecting for early onset and severe phenotypes.

Despite the challenges, genetic studies are beginning to unravel some of the phenotypic heterogeneity of SLE. For example a variant in *STAT4* has been linked to more severe disease and *STAT4*, *IRF5*, *ITGAM*, and major histocompatibility complex (MHC) SNPs are predictive of dsDNA positive disease (Taylor et al. 2008; Chung et al. 2011). Future cases can now be stratified phenotypically and / or genetically.

It seems likely that, whilst both family and case control studies will continue to generate new genetic disease loci, the paucity of family data will necessitate case control studies for many diseases in which monogenic or oligogenic aetiology is suspected in individual cases. Large numbers of individuals could be sequenced at very low

coverage and rare variants imputed by comparing to phased haplotypes from the increasingly large meta cohorts of samples available from large scale sequencing projects (International HapMap 3 Consortium et al. 2010; Khurana et al. 2013; Muddyman et al. 2013; Lomas 2013). Software tools to perform this imputation already exist (Browning and Browning 2009; Jostins, Morley, and Barrett 2011; Delaneau, Zagury, and Marchini 2012; Liu et al. 2013).

This work also highlights the need for accurate and detailed phenotypic data on all sequenced patients. This will reduce genetic heterogeneity in large scale studies or permit phenotypic subset analysis, and allow meaningful genotype-phenotype correlations. Phenotypic data also aids the prioritisation of candidate variants, for example the lack of platelet abnormalities in patient 17709 was evidence against a link to the *WAS* variant, more systematically variants could be searched based on tissue expression for example central nervous system (CNS) expression, or ability to cross the blood brain barrier, for SLE with CNS involvement.

Access to DNA and phenotypic data for other family members is particularly valuable, and provided vital information for patient 17709 using micro array data. In designing future human studies to explore rare Mendelian variants of complex disease, the availability of DNA, consent and phenotypic information from family members should be pre–specified in the study design. This will require close collaboration between investigators and local clinicians.

7.3 Limitations and advantages of murine and human forward genetics

How do murine mutagenesis experiments compare to human genetic studies as a tool to identify genetic variants responsible for rare diseases or monogenic forms of more complex disease?

A good forward genetic method to identify deleterious genetic variants should be relevant to human disease, provide a substrate for further functional assays, and permit rapid identification of causative variants for a phenotype. Other practical considerations include access to samples and the ease of obtaining accurate phenotyping data, finally ethical issues must also inform study design.

7.3.1 Relevance to human disease

Identification of a gene variant linked to disease in humans can be of immediate clinical benefit, depending on the confidence and strength of the association, but if novel will require validation in unrelated datasets with the same disease and control populations. Access to suitable datasets for validation should be planned at the outset of NGS studies for rare novel variants. If the disease itself appears novel and does not fit a known syndrome this can be difficult.

Once validated, the discovery can provide a diagnosis, prognosis and potentially influence treatment for the study participants, for example in a child with intractable inflammatory bowel disease, discovery of a novel variant in X–linked inhibitor of apoptosis (*XIAP*) led to a previously unsuspected diagnosis of X–linked lymphoproliferative syndrome and prompted treatment with a bone marrow transplant (Worthey et al. 2011). Alternatively the discovery can result in a diagnostic tool for other affected individuals. Ideally this should lead rapidly to better therapeutic options for the disease.

Moving from a mutant mouse and its underlying genetic defect to clinical research or diagnostic tools is less immediate. Classically, validation of an ENU mutant as causative in the mouse required observation of the same phenotype with other mutations in the same gene, as was known for the *ENU16Ch17a Lyn* mutant identified in Chapter 5 (Verhagen et al. 2009), or testing by complementation, phenotypic rescue or knock–out.

Mutant mice do not always fully model a human disease, a gene defect may have very different consequences in mouse compared to human, and a murine phenotype that appears to mimic a human disease may be due to a genetic defect of no clinical consequence.

Despite this, in many cases mouse models provide an excellent model for a human disease, including genotype—phenotype correlations that mirror the human disease spectrum. *Nephertiti* illustrates the power of ENU to reveal viable allelic variants that mimic the human disease under a physiological promoter. Many genetic variants identified in mice are found to have similar phenotypic consequences in humans, Fcgr2b promoter polymorphism contributes to systemic autoimmunity in NZB mice (Xiu et al. 2002), and the FCGRIIB^{T232I} SNP confers susceptibility to SLE in humans (Lee, Ji, and Song 2009).

In some cases the mouse has incomplete correlation with the human phenotype, for example an ENU mouse with dedicator of cytokinesis 8 (*Dock8*) deficiency models the immunodeficiency seen in patients with DOCK8 disease but does not explain features such as hyper–IgE in the patients (Lambe et al. 2011; Su, Jing, and Zhang 2011).

7.3.2 Models for functional assays and translation of results to clinical treatments

Although gene variant discoveries in humans can rapidly aid clinical diagnosis (one clinical centre reports a 25% molecular diagnostic success rate (Yang et al. 2013)), progress towards better therapeutics can be very slow. One limiting factor will be the need for in vitro and in vivo models for the functional experiments which are required to understand the molecular and physiological consequences of a variation and its impact on potential therapeutic modifications. Transgenic mice or gene targeted stem cells are now available for many genes due to large multi centre initiatives (Skarnes et al. 2011), these null mutants will mimic the behaviour of a human null variant.

CRISPR/Cas9 systems can be used to generate point mutations in vivo, modelling a wider range of genomic defects and accelerating the production of mouse models by generating mutations directly in the zygote (Wang et al. 2013; Mali et al. 2013; Cong et al. 2013).

7.3.3 Identification of causative variants

The inbred B6 mouse closely resembles the reference genome, and carries only small numbers of naturally occurring variants. Even the C57Bl/6N strain, separated from C57BL/6J for 200 generations, may differ at less than 40 exonic loci (Simon et al. 2013; Keane et al. 2011).

The heterogeneity of the human genome makes the task of isolating causative variants more complex than in the ENU mouse. Any individual carries around 13,500 exonic variations from the reference human genome (Tennessen et al. 2012). For comparison the ENU mouse carries 1.5 ENU mutations per Mb (Figure 5.7). This equates to approximately 4,075 ENU mutations per G_1 genome, or around 80 exonic mutations.

In humans reported rare variation from the reference genome is higher for some ethnicities. 7 non–European SLE patients carried on average 307 heterozygous candidate variants compared to an average of 121 such variants in the 3 European SLE patients. Using datasets specific for a comparable population can eliminate more variants based on population specific allele frequency. Using the Punjabi 1000 genomes dataset eliminated 15% of the apparently rare homozygous and heterozygous variants in patient 0001, since they are common in the Pakistani population. However some of the excess of rare variants in non–European ethnicities reflects greater population diversity and stratification.

One advantage of human heterogeneity is that it permits fine mapping. Linkage analysis using exome sequencing has been effectively harnessed for the study of hu-

man pedigrees. Variations on the Lander–Green method, developed for array data, incorporate knowledge of the population allele frequencies of HapMap SNPs (Smith et al. 2011; Guergueltcheva et al. 2012; Abecasis et al. 2002). Alternative approaches use other hidden Markov models (HMM) to identify regions that are IBD and common to autosomal recessive phenotypes (Chahrour et al. 2012; Rödelsperger et al. 2011).

Reducing the search space from the whole genome to the exome significantly reduces the number of informative variants (Roach et al. 2010), however this is typically several orders of magnitude larger than the number of ENU–induced exonic variants, e.g. 6,000 to 8,000 exonic HapMap variants per individual (Smith et al. 2011; Guergueltcheva et al. 2012) compared to 74 exonic variants in the *ENU16Ch17a* pedigree.

The low density of ENU coding variants does not permit fine scale linkage analysis based only on exonic variants. Exome sequencing also precludes the detection of regulatory mutations (Masuya et al. 2007), and the inefficiencies of capture have resulted in a failure to find the causative mutant in one in five ENU pedigrees using WES, even for recessive traits (Fairfield et al. 2011; Andrews et al. 2012). The ability to detect IBD regions using low coverage and the falling costs of next generation sequencing make the WGS method described in chapter 5 increasingly cost effective.

Because of human genetic heterogeneity, the search for causative variants in humans requires tight linkage regions, or knowledge of inheritance patterns, and the ability to prioritise variants based on predicted functional effect, for example based on bioinformatic prediction tools or tissue expression data. Some human studies, particularly if the causative variant is novel, will also require experimental testing of multiple candidates in vitro or in vivo. Therefore high throughput methods to test such variants in the laboratory will be required.

WGS rather than WES is required to perform linkage in ENU mutants, but this can rapidly isolate single causative mutations, and since ENU induces overwhelmingly

point mutations, which are called more reliably that insertions or deletions using currently available tools, the identification and isolation of ENU mutations are more straightforward than finding causative variants in humans.

7.3.4 Access to samples

As illustrated by the challenges of the SRNS and SLE studies, patients recruited for future sequencing studies should have detailed and accurate phenotypic information and access to family information and DNA. Many groups are now building patient databases or biobanks for specific diseases. This is particularly important for rare diseases in order to collect sufficient patient samples. The scope of such registries goes beyond sequencing, but can include collection of blood and consent for DNA analysis. The SRNS patients are part of the UK registry for Rare Kidney Diseases (https://www.renalradar.org), which enrols patients with rare renal diseases to a national database. The UK biobank (http://www.ukbiobank.ac.uk) provides data for 500,000 individuals and recently began genotyping all of these to look for genetic associations with common disease and interactions with environmental factors. Such databases provide a rich resource for identification and validation of rare variants. It is vital that family DNA and phenotypes form part of these archives.

ENU phenotype and genotype data can also be shared in openly accessible databases to provide investigators with a resource to link phenotypes and genotypes of interest.

7.3.5 Phenotyping

For a high throughput ENU programme to fully utilise the knowledge of underlying mutations linked to phenotypes provided by the IBD WGS method, comparison of mutant genes and phenotypes both within and between pedigrees is desirable. To compare quantitative phenotypes requires reproducible data and greater automation. Reliability must be demonstrated for new tools such as mass cytometry (Bodenmiller

et al. 2012) if these are to be adopted for ENU screens.

In mice, the challenge for phenotypic validity is proving that the phenotype truly models a human disease. In humans, clinical signs and symptoms may be more or less specific to underlying biology, most of the individual symptoms of SLE, such as rash, arthralgia or anaemia are non–specific. dsDNA positivity is a notable exception to this being highly specific to SLE, however dsDNA positivity has been reported in myeloma, chronic active hepatitis or rheumatoid arthritis (Isenberg et al. 2007), and it's presence in *Was*-/- mice (Nikolov et al. 2010; Stephanie Humblet-Baron 2007) fuelled suspicions in this study, eventually disproved, that a *WAS* variant was relevant for SLE patient 17709.

7.3.6 Ethical issues

Ethical return of information from next generation sequencing. Human participants (or their parents) in any research study must give informed consent prior to collection or analysis of data. The rise of next generation sequencing as a research tool and its increasingly routine clinical use has led to ethical questions about how to share sequencing data with individuals. These arise because large numbers of variants are detected, the vast majority of which will be unrelated to the research or clinical question.

These variants have been described as the 'incidentalome' and may include false positive disease risk associations due to incorrect annotations, technical errors, incorrect prior probabilities for disease risk and multiple hypothesis testing (Kohane, Hsing, and Kong 2012) as well as SNPs with possible clinical implications for the patient, but with varying degrees of certainty in the clinical association and varying implications in terms of clinical consequences and opportunity to modify risk factors or therapeutics.

Most researchers and clinicians agree that there is a duty to return incidental

results to research subjects if they are clinically relevant and may alter outcomes because treatment or preventive measures exist (Berg, Khoury, and Evans 2011). These criteria may apply to variants in only 100 or less genes at present (Evans and Rothschild 2012). A recent American College of Medical Genetics policy statement recommended searching for and returning known pathogenic, and in some cases novel but predicted pathogenic variants, in 57 genes associated with 24 life-threatening but treatable conditions (Green et al. 2013). Controversially these guidelines for clinical sequencing recommended that this be done without seeking patient permission, an approach consistent with sharing of incidental findings in other areas of clinical testing such as radiological imaging.

Whether to allow participants the option to receive less clinically relevant data, or less reliable data, remains a topic of debate. While some point out that researchers have no obligation to return such data and highlighted the dangers of 'tantalizing but essentially meaningless' results (Evans and Rothschild 2012), others have argued for lower thresholds of clinical utility and consideration of the personal meaning of results for an individual (Ravitsky and Wilfond 2006). However given our limited and often erroneous understanding of the penetrance or risk conferred by most variants and public misconceptions about genetic data a cautious approach sharing only clearly clinically relevant data (in a time limited way) is a more justifiable approach. These questions are particularly important for paediatric research participants, where consent and sharing is usually with parents or guardians. Sharing of genetic risk for later onset disease, based on respect for parental decision making, may be consistent with practice in other areas of paediatric medicine (Wilfond and Ross 2009), but undermines the child's anticipatory autonomy right (Bredenoord, Vries, and Delden 2013).

NGS studies with human participants must consider what information will be returned and make this clear to participants, with or without a degree of choice.

Return of results must be reliable and responsible and therefore requires verification of variants in a clinical genetics laboratory and sharing of results by a clinically trained individual. Planning for human NGS research should therefore include access to such labs and clinical staff. Current clinical electronic record systems are ill-equipped to handle such data and improvements in this infrastructure are a pressing need (Scheuner et al. 2009).

7.3.7 Ethical Mouse Research

ENU mutagenesis in mice avoids many of the described ethical issues of importance in human sequencing, but animal welfare issues are separate considerations in designing forward genetic experiments in mice.

The WGS sequencing method developed in chapter 5 has potential for reducing mouse numbers, both directly since no additional breeding beyond G_3 is required for mapping, and indirectly, because uninformative or duplicate strains will be detected more rapidly allowing researchers to focus breeding on novel mutants, with less strains abandoned due to failure to find a mutation. The use of lethal screens is a refinement that could increase the yield of detectable phenotypes, increasing the efficiency of ENU.

7.4 The importance of finding rare disease variants

Searching for rare variants that contribute to disease is important both to understand rare monogenic disease, and because they can be informative for genetically complex, common disease. Rare variants found in Mendelian forms of common disease may point to biological mechanisms of importance for the common form, and rare variants of large phenotypic effect may contribute directly to the heritability of the common disease.

Rare variants identified in monogenic disease can therefore provide a better understanding of of disease mechanisms and lead to more targeted treatments and improved outcomes for patients. The contribution of rare variants to heritability in common diseases is not certain, however even if this is small, the same pathways and genes highlighted by the rare variants may play a role in the complex disease, via another type of disruption, such as epigenetic changes or environmental interactions. For example *VHL* mutations contribute both to rare multiple tumour disease and common renal cell cancer (Latif et al. 1993; Barry and Krek 2004; Schödel et al. 2012), and methylation changes in clear cell renal carcinoma alter *VHL* - pathway activation in metastatic cells (Vanharanta et al. 2013).

The contribution of rare variants themselves to the heritability in common diseases is a topic of debate in the literature.

Although existing genome wide association studies have identified thousands of common variants associated with disease, the majority of these have not been functionally validated and they collectively explain only a small proportion of the heritability for any common trait. For example around 10-13% of SLE heritability is explained by known disease associated variation (Guerra, Vyse, and Cunninghame Graham 2012; So et al. 2011). Rare variants of large effect size are one explanation for the missing heritability. Evidence for rare variants causing disease includes the observation that many chronic diseases have familial forms due to rare variants of large effect, such as recurrent pyogenic infections due to rare variants in *MYD88* (George et al. 2011), and monogenic forms of hypertension due to mutations in subunits of sodium channel non neuronal 1(*SCNN1*) which encodes the amiloride-sensitive epithelial channel (EnaC). Further, many diseases can be caused by one of many rare variants in a disease gene, such as variants causing cystic fibrosis (Sosnay et al. 2013). The rare variant model is supported by evolutionary theory and empirical population genetics. Human populations often share common variants, many of which arose

before the first migration of our species out of Africa, thus they have been subject to selection pressure over more than 50-60,000 years. Highly deleterious variants will have been selected out of the population. In contrast relatively recent exponential human population growth over the last millennium has resulted in the rapid accumulation of vast numbers of rare variants (Keinan and Clark 2012). These have been subject to only weak purifying selection and so some of these variants will be highly deleterious, 86% of SNPs predicted to be deleterious arose within the last 5,000 to 10,000 years (Fu et al. 2012).

However there are a number of arguments against the rare variant model. It has been proposed that a proportion of GWAS SNPs are actually due to the combined effect of multiple causal rare variants in linkage disequilibrium (LD) with the GWAS SNP, this is known as synthetic association (Dickson et al. 2010). But quantitative genetic theory and expected linkage disequilibrium indicate that such associations are unlikely to explain many common GWAS SNPs (Wray, Purcell, and Visscher 2011). This modelling is supported by empirical data in autoimmune disease, where sequencing 25 GWAS risk genes failed to identify rare coding variants associated with disease (Hunt et al. 2013). Whilst the theoretical and empirical data indicate that synthetic association does not account for much of the missing heritability, causal rare variants that are not clustered around GWAS SNPs would not contribute to this model and very low frequency or private variants would have failed to show a signal in the autoimmunity study by Hunt et al, while still potentially having a large effect in single individuals. This study was insufficiently powered to detect variants with a MAF of 0.001 if the odds ratio for risk was less than 5.

Further arguments against the rare allele model include high sibling recurrence rates that cannot be attributed to rare variants with expected effect sizes, the lack of obviously additive effects of rare variants, and the changing prevalence of many diseases in recent history (Though this latter point argues for environmental effects

rather than supporting a common variant model). Finally the consistency of many (though not all) GWAS variants across populations is inconsistent with the idea that they are linked to rare variants that have arisen recently in populations (Gibson 2011).

In order to explain missing heritability an alternative to the rare variant model is the 'infinitesimal model' in which a large number of small effect variants across the allele frequency spectrum contribute additively to disease risk. This model is consistent with the existing GWAS findings, in that it proposes that many variants are simply too small in effect size and or allele frequency to be detected, and could also be consistent with evolutionary theory, as such variants have not been subject to purifying selection due to their weak effects. Common variants identified by GWAS have been demonstrated to be associated both with disease and with a spectrum of intermediate similar traits, for example 32 common variants are associated with normal variation in lipid levels and with extremely high lipid level phenotypes (Teslovich et al. 2010). In model organisms, quantitative trait loci (QTLs) contribute to complex phenotypes. However such QTLs are frequently not replicable in out–bred populations and so could be due to underlying rare variants, additionally many common variants associated with disease have not been functionally validated.

In summary, there is good evidence that both rare variants of large effect size and common variants with small effects on phenotype will contribute to disease risk; however, the relative contribution of each variant type in individual diseases remains unknown and, though outside the scope of this thesis, the role of the secondary but important effect of genotype by environment interactions remains poorly understood.

Bodmer and Bonilla have argued that rare variants, while an important contributor to disease, are not likely to be familial, due to low penetrance and therefore family studies to identify rare phenotypic variants are not relevant (Bodmer and Bonilla 2008). However this is in conflict with the evidence for familial clustering and familial monogenic forms of common disease. Instead a model in which multiple rare

variants with additive effects are shared within a family would lead to an increased sibling risk ratio (Schork et al. 2009), as is observed for example in SLE.

Multiple, possibly additive, rare variants may work together to lead to disease, this could account for the lack of rare monogenic candidates in the SLE and SRNS patients in chapter 6. What is the best way to find these? Sequencing large numbers of affected patients at low coverage and imputing for rare variants may reveal a number of genes with a significantly higher number of rare variants in patients in comparison to controls, enrichment for rare or de novo variants has been shown in autism and schizophrenia (Sanders et al. 2012b; Purcell et al. 2014).

It may be useful to examine affected individuals for overrepresentation of deleterious variants in specific physiological pathways. Pathway analysis requires accurately curated pathways and correction for biases due to gene or pathway size and pathway overlap (Ramanan et al. 2012).

Conventional ENU does not model complex gene interactions, however mutants can be informative for complex disease, as variants in these genes can influence disease risk in more modest ways in humans. In this way ENU mutants can highlight genes that harbour common variants of small effect size in humans. For example, the *APFN1018* mouse (section 3.6) carries a suspected null *Tnfsf13b* (BAFF) cryptic splice variant, resulting in a B cell deficient phenotype (Schiemann 2001; Mackay and Schneider 2009). Conversely, overexpression of BAFF leads to autoimmunity in mice (Mackay et al. 1999), and elevated levels of BAFF are a feature of human Sjorgren's syndrome (Groom et al. 2002). The Lyn Tyrosine Kinase mutant *ENU16Ch17a* exhibits a B cell deficient phenotype without autoimmunity consistent with a kinase dead form of the protein (Verhagen et al. 2009). The *Lyn* -/- mice has a more severe autoimmune presentation due to altered positive and negative B cell receptor induced signalling and altered *Lyn* expression or perturbations of LYN regulated pathways may also play a role in human autoimmune disease (Xu et al. 2005).

7.5 Conclusions and future directions

Mouse models are an essential complement to human research in exploring the genetics of human disease. The identity by descent based sequencing method developed in this thesis can contribute to basic research. Adoption of the approach by large–scale ENU programmes will lead to a substantial increase in the productivity of the programmes, accelerating gene discovery and eliminating the issue of unseen modifiers from an outcross strain.

Efficient ENU mutagenesis studies will advance our understanding of gene–function and the mechanisms of genetic disease. The approach will reduce the burden in animal costs and allow post mortem screens, with increased sophistication and accuracy in a broader range of tissues. With the rapidly falling costs of WGS it is possible to envisage a future in which all G_3 ENU mice are sequenced to a depth sufficient to identify and segregate all their mutations, creating a rich openly accessible dataset of allelic variation and corresponding phenotypic information, including linkage data for non–coding mutations with measurable effects. This could be achieved accurately with 4 or 5–fold sequencing due to the increased power to impute genotypes. This database could be mined for associations across pedigrees, including the detection of subtle phenotypes.

Knowledge of mutations across the genome could be combined with gene expression phenotypes from RNA-seq to identify more biological phenotypes and detect non–coding effects.

The finding of a mutation in Laminin $\beta 2$ in the nephrotic *nephertiti* pedigree illustrates the potential for ENU forward genetics to explore kidney disease. Screens for renal pathology including urinalysis, blood biochemistry for renal function, electrolytes and bone chemistry, plus imaging and renal histology could generate murine models of human renal disease, including nephrotic disease and proteinuria.

Rare diseases contribute to the burden of chronic kidney disease and the genet-

ics of many of these diseases are incompletely understood, however next generation sequencing techniques are revealing novel renal disease genes (Bredrup et al. 2011; Humbert et al. 2014).

For the SLE and SRNS datasets the next step is to obtain parental DNA, looking for de novo variants and phasing compound heterozygotes. Use of familial DNA may be particularly effective for patient 0001, with a suspected homozygous causative variant and the SLE siblings with a suspected compound heterozygous aetiology.

Any candidates identified must be validated in larger datasets of SRNS and SLE accessible through collaborators. This should be followed by functional assays to test the consequences of the variant allele, for example in 0001, after validation in family members, the variants could be expressed in a podocyte cell line and examined for effects on cell motility, structure and development (Saleem et al. 2002). This could be complemented by a mouse model generated using the CRISPR/Cas technique (Wang et al. 2013).

The work within this thesis on the SRNS and SLE patients highlights some of the challenges in using WGS or WES to find rare disease variants in humans. Experiments in isolated unrelated disease cases will be a typical scenario for many diseases, where affected families are not easily available. Careful study design, with larger cohorts, access to parental or familial samples and detailed phenotyping will overcome some of the problems identified in Chapter 6, however for many diseases issues will remain due to genetic and phenotypic heterogeneity. Larger studies based on collaborative patient datasets (for example RenalRaDaR – www.renalradar.org) are needed to reveal the genetic architecture of renal diseases and contribute to personalised genomics in clinical practice.

References

A P Davis, M J Justice (Jan. 1998). "An Oak Ridge legacy: the specific locus test and its role in mouse mutagenesis." *Genetics* 148.1, p. 7.

Abecasis, Gonçalo R, Stacey S Cherny, William O Cookson, and Lon R Cardon (Jan. 2002). "Merlin–rapid analysis of dense genetic maps using sparse gene flow trees." *Nat Genet* 30.1, pp. 97–101.

Acevedo-Arozena, Abraham, Sara Wells, Paul Potter, Michelle Kelly, Roger D Cox, and Steve D M Brown (2008). "ENU mutagenesis, a way forward to understand gene function." *Annual review of genomics and human genetics* 9, pp. 49–69.

Adams, David, Richard Baldock, Shoumo Bhattacharya, Andrew J Copp, Mary Dickinson, Nicholas D E Greene, Mark Henkelman, Monica Justice, Timothy Mohun, Stephen A Murray, Erwin Pauws, Michael Raess, Janet Rossant, Tom Weaver, and David West (Jan. 2013). "Bloomsbury report on mouse embryo phenotyping: recommendations from the IMPC workshop on embryonic lethal screening". *Disease models & mechanisms* 6.3, pp. 571–579.

Adzhubei, I. A., S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev (Apr. 2010). "A method and server for predicting damaging missense mutations". *Nat Methods* 7.4, pp. 248–249.

Aghaeepour, Nima et al. (Feb. 2013). "Critical assessment of automated flow cytometry data analysis techniques". *Nat Methods* 10.3, pp. 228–238.

Al-Mayouf, Sulaiman M et al. (Oct. 2011). "Loss-of-function variant in DNASE1L3 causes a familial form of systemic lupus erythematosus". *Nat Genet* 43.12, pp. 1186–1188.

Alarcon, G S, H M Bastian, T M Beasley, J M Roseman, F K Tan, B J Fessler, L M Vila, and G Jr McGwin (2006). "Systemic lupus erythematosus in a multi-ethnic cohort (LUMINA) XXXII: [corrected] contributions of admixture and socioeconomic status to renal involvement." *Lupus* 15.1, pp. 26–31.

Albert, M H et al. (Apr. 2010). "X-linked thrombocytopenia (XLT) due to WAS mutations: clinical characteristics, long-term outcome, and treatment options". *Blood* 115.16, pp. 3231–3238.

Aldrich, R A, A G Steinberg, and D C Campbell (Feb. 1954). "Pedigree demonstrating a sex-linked recessive condition characterized by draining ears, eczematoid dermatitis and bloody diarrhea." *Pediatrics* 13.2, pp. 133–139.

Altare, F et al. (May 1998). "Impairment of mycobacterial immunity in human interleukin-12 receptor deficiency." *Science* 280.5368, pp. 1432–1435.

Altare, Frédéric, Armin Ensser, Adrien Breiman, Janine Reichenbach, Jamila El Baghdadi, Alain Fischer, Jean-François Emile, Jean-Louis Gaillard, Edgar Meinl, and Jean-Laurent Casanova (July 2001). "Interleukin-12 Receptor β1 Deficiency in a Patient with Abdominal Tuberculosis". *Journal of Infectious Diseases* 184.2, pp. 231–236.

Andreasen, Charlotte, Jonas B Nielsen, Lena Refsgaard, Anders G Holst, Alex H Christensen, Laura Andreasen, Ahmad Sajadieh, Stig Haunsø, Jesper H Svendsen, and Morten S Olesen (Jan. 2013). "New population-based exome data are questioning the pathogenicity of previously cardiomyopathy-associated genetic variants". *European journal of human genetics : EJHG* 21.9, pp. 918–928.

Andrews, T D et al. (May 2012). "Massively parallel sequencing of the mouse exome to accurately identify rare, induced mutations: an immediate source for thousands of new mouse models." *Open biology* 2.5, p. 120061.

Angelis, M H Hrabé de et al. (Aug. 2000). "Genome-wide, large-scale production of mutant mice by ENU mutagenesis." *Nat Genet* 25.4, pp. 444–447.

Arnold, C. N., Y. Xia, P. Lin, C. Ross, M. Schwander, N. G. Smart, U. Muller, and B. Beutler (Mar. 2011). "Rapid identification of a disease allele in mouse through whole genome sequencing and bulk segregation analysis". *Genetics* 187.3, pp. 633–641.

Arnold, Carrie N et al. (2012). "ENU-induced phenovariance in mice: inferences from 587 mutations." *BMC Research Notes* 5.1, p. 577.

Ashley Jefferson, J and Stuart J Shankland (Aug. 2013). "Has the circulating permeability factor in primary FSGS been found?" *Kidney international* 84.2, pp. 235–238.

Astor, Brad C et al. (Feb. 2011). "Lower estimated glomerular filtration rate and higher albuminuria are associated with mortality and end-stage renal disease. A collaborative meta-analysis of kidney disease population cohorts". *Kidney international* 79.12, pp. 1331–1340.

Balling, R (2001). "ENU mutagenesis: analyzing gene function in mice." *Annual review of genomics and human genetics* 2, pp. 463–492.

Bamshad, Michael J, Sarah B Ng, Abigail W Bigham, Holly K Tabor, Mary J Emond, Deborah A Nickerson, and Jay Shendure (Sept. 2011). "Exome sequencing as a tool for Mendelian disease gene discovery". *Nat Rev Genet* 12.11, pp. 745–755.

Barbaric, I., S. Wells, A. Russ, and T. N. Dear (Mar. 2007). "Spectrum of ENU-induced mutations in phenotype-driven and gene-driven screens in the mouse". *Environmental and Molecular Mutagenesis* 48.2, pp. 124–142.

Barbaux, S, P Niaudet, M C Gubler, J P Grünfeld, F Jaubert, F Kuttenn, C N Fékété, N Souleyreau-Therville, E Thibaud, M Fellous, and K McElreavey (Dec. 1997). "Donor splice-site mutations in WT1 are responsible for Frasier syndrome." *Nat Genet* 17.4, pp. 467–470.

Barnett, H. L. and C. M. Edelmann (Apr. 1984). "Minimal change nephrotic syndrome in children: deaths during the first 5 to 15 years' observation. Report of the International Study of Kidney Disease in Children." *Pediatrics* 73.4, pp. 497–501.

Barry, Robert E and Wilhelm Krek (Sept. 2004). "The von Hippel-Lindau tumour suppressor: a multi-faceted inhibitor of tumourigenesis." *Trends in molecular medicine* 10.9, pp. 466–472.

Bartoloni, L, J L Blouin, Y Pan, C Gehrig, A K Maiti, N Scamuffa, C Rossier, M Jorissen, M Armengot, M Meeks, H M Mitchison, E M K Chung, C D Delozier-Blanchet, W J Craigen, and S. E. Antonarakis (July 2002). "Mutations in the DNAH11 (axonemal heavy chain dynein type 11) gene cause one form of situs inversus totalis and most likely primary ciliary dyskinesia". *Proc Natl Acad Sci U S A* 99.16, pp. 10282–10286.

Beck Jr., Laurence H, Ramon G B Bonegio, Gérard Lambeau, David M Beck, David W Powell, Timothy D Cummins, Jon B Klein, and David J Salant (July 2009). "M-Type Phospholipase A 2Receptor as Target Antigen in Idiopathic Membranous Nephropathy". *N Engl J Med* 361.1, pp. 11–21.

Beicht, Sonja, Gertrud Strobl-Wildemann, Sabine Rath, Oliver Wachter, Martin Alberer, Elke Kaminsky, Lutz T Weber, Tanja Hinrichsen, Hanns-Georg Klein, and Julia Hoefele (Sept. 2013). "Next generation sequencing as a useful tool in the diagnostics of mosaicism in Alport syndrome". *Gene* 526.2, pp. 474–477.

Beier, D. R. (July 2000). "Sequence-based analysis of mutagenized mice". *Mamm Genome* 11.7, pp. 594–597.

Benoit, G., E. Machuca, L. Heidet, and C. Antignac (Dec. 2010). "Hereditary kidney diseases: highlighting the importance of classical Mendelian phenotypes". *Ann N Y Acad Sci* 1214, pp. 83–98.

Bentley, D. R. et al. (Nov. 2008). "Accurate whole human genome sequencing using reversible terminator chemistry". *Nature* 456.7218, pp. 53–59.

Berg, Jonathan S, Muin J Khoury, and James P Evans (May 2011). "Deploying whole genome sequencing in clinical practice and public health: Meeting the challenge one bin at a time". *Genetics in medicine : official journal of the American College of Medical Genetics* 13.6, pp. 499–504.

Bergmann, Hannes et al. (Jan. 2013). "B cell survival, surface BCR and BAFFR expression, CD74 metabolism, and CD8- dendritic cells require the intramembrane endopeptidase SPPL2A." *J Exp Med* 210.1, pp. 31–40.

Bhanot, O. S., Grevatt, P. C., J. M. Donahue, C. N. Gabrielides, and J. J. Solomon (Feb. 1992). "In vitro DNA replication implicates O2-ethyldeoxythymidine in transversion mutagenesis by ethylating agents". *Nucleic Acids Res* 20.3, pp. 587–594.

Bigler, Cornelia, Monica Schaller, Iryna Perahud, Michael Osthoff, and Marten Trendelenburg (Sept. 2009). "Autoantibodies against Complement C1q Specifically Target C1q Bound on Early Apoptotic Cells". *The Journal of Immunology* 183.5, pp. 3512–3521.

Blanco, P, A K Palucka, M Gill, V Pascual, and J Banchereau (Nov. 2001). "Induction of dendritic cell differentiation by IFN-alpha in systemic lupus erythematosus." *Science* 294.5546, pp. 1540–1543.

Block, S R, J B Winfield, M D Lockshin, W A D'Angelo, and C L Christian (Oct. 1975). "Studies of twins with systemic lupus erythematosus. A review of the literature and presentation of 12 additional sets." *The American journal of medicine* 59.4, pp. 533–552.

Bodenmiller, Bernd, Eli R Zunder, Rachel Finck, Tiffany J Chen, Erica S Savig, Robert V Bruggner, Erin F Simonds, Sean C Bendall, Karen Sachs, Peter O Krutzik, and Garry P Nolan (Aug. 2012). "Multiplexed mass cytometry profiling of cellular states perturbed by small-molecule regulators". *Nat Biotechnol* 30.9, pp. 858–867.

Bodmer, Walter and Carolina Bonilla (June 2008). "Common and rare variants in multifactorial susceptibility to common diseases." *Nat Genet* 40.6, pp. 695–701.

Boerries, Melanie et al. (Jan. 2013). "Molecular fingerprinting of the podocyte reveals novel gene and protein regulatory networks". *Kidney international* 83.6, pp. 1052–1064.

Boles, Melissa K et al. (Dec. 2009). "Discovery of candidate disease genes in ENU-induced mouse mutants by large-scale sequencing, including a splice-site mutation in nucleoredoxin." *PLoS Genet* 5.12, e1000759.

Bosma, G C, R P Custer, and M J Bosma (Feb. 1983). "A severe combined immunodeficiency mutation in the mouse." *Nature* 301.5900, pp. 527–530.

Botstein, David et al. (May 2000). "Gene Ontology: tool for the unification of biology". *Nat Genet* 25.1, pp. 25–29.

Botto, Marina and Mark J Walport (Sept. 2002). "C1q, autoimmunity and apoptosis." *Immunobiology* 205.4-5, pp. 395–406.

Boute, N, O Gribouval, S Roselli, F Benessy, H Lee, A Fuchshuber, K Dahan, M C Gubler, P Niaudet, and C. Antignac (Apr. 2000). "NPHS2, encoding the glomerular protein podocin, is mutated in autosomal recessive steroid-resistant nephrotic syndrome." *Nat Genet* 24.4, pp. 349–354.

Boycott, Kym M, Megan R Vanstone, Dennis E Bulman, and Alex E MacKenzie (Sept. 2013). "Rare-disease genetics in the era of next-generation sequencing: discovery to translation". *Nat Rev Genet* 14.10, pp. 681–691.

Brandl, K, S Rutschmann, X Li, X Du, N Xiao, B Schnabl, D A Brenner, and B. Beutler (Mar. 2009). "Enhanced sensitivity to DSS colitis caused by a hypomorphic Mbtps1 mutation disrupting the ATF6-driven unfolded protein response". *Proc Natl Acad Sci U S A* 106.9, pp. 3300–3305.

Breathnach, R and P Chambon (June 1981). "Organization and Expression of Eucaryotic Split Genes Coding for Proteins". *Annual Review of Biochemistry* 50.1, pp. 349–383.

Bredenoord, Annelien L, Martine C de Vries, and Johannes J M van Delden (Mar. 2013). "Next-generation sequencing: does the next generation still have a right to an open future?" *Nat Rev Genet* 14.5, pp. 306–306.

Bredrup, Cecilie et al. (Nov. 2011). "Ciliopathies with Skeletal Anomalies and Renal Insufficiency due to Mutations in the IFT-A Gene WDR19". *The American Journal of Human Genetics* 89.5, pp. 634–643.

Brenner, Sydney et al. (June 2000). "Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays". *Nature Biotechnology* 18.6, pp. 630–634.

Brinkkoetter, Paul Thomas, Christina Ising, and Thomas Benzing (Apr. 2013). "The role of the podocyte in albumin filtration". *Nature reviews. Nephrology* 9.6, pp. 328–336.

Brown, Celeste J, Sachiko Takayama, Andrew M Campen, Pam Vise, Thomas W Marshall, Christopher J Oldfield, Christopher J Williams, and A Keith Dunker (July 2002). "Evolutionary Rate Heterogeneity in Proteins with Long Disordered Regions". *Journal of Molecular Evolution* 55.1, pp. 104–110.

Brown, Elizabeth J, Johannes S Schlöndorff, Daniel J Becker, Hiroyasu Tsukaguchi, Stephen J Tonna, Andrea L Uscinski, Henry N Higgs, Joel M Henderson, and Martin R Pollak (Jan. 2010). "Mutations in the formin gene INF2 cause focal segmental glomerulosclerosis." *Nature Genetics* 42.1, pp. 72–76.

Brown, Steve D M and Mark W Moore (May 2012). "Towards an encyclopaedia of mammalian gene function: the International Mouse Phenotyping Consortium." *Disease models & mechanisms* 5.3, pp. 289–292.

Browning, Brian L and Sharon R Browning (Feb. 2009). "A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals." *Am J Hum Genet* 84.2, pp. 210–223.

Brunschwig, H, L Levi, E Ben-David, R W Williams, B Yakir, and S Shifman (July 2012). "Fine-Scale Maps of Recombination Rates and Hotspots in the Mouse Genome". *Genetics* 191.3, pp. 757–764.

Bull, Katherine R, Thomas Mason, Andrew J Rimmer, Tanya L Crockford, Karlee L Silver, Tiphaine Bouriez-Jones, Tertius A Hough, Shirine Chaudhry, Ian Sd Roberts, Christopher C Goodnow, and Richard J Cornall (2014). "Next-generation sequencing to dissect hereditary nephrotic syndrome in mice identifies a hypomorphic mutation in Lamb2 and models Pierson's syndrome". *The Journal of pathology* 233.1, pp. 18–26.

Bull, Katherine R et al. (Jan. 2013). "Unlocking the Bottleneck in Forward Genetics Using Whole-Genome Sequencing and Identity by Descent to Isolate Causative Mutations". *PLoS Genet* 9.1, e1003219.

Bustamante, Carlos D, Francisco M De La Vega, and Esteban G Burchard (July 2011). "Genomics for the world". *Nature* 475.7355, pp. 163–165.

Calco, G N, O R Stephens, L M Donahue, C C Tsui, and B A Pierchala (Mar. 2014). "CD2-associated Protein (CD2AP) Enhances Casitas B Lineage Lymphoma-3/c (Cbl-3/c)-mediated Ret Isoform-specific Ubiquitination and Degradation via Its Amino-terminal Src Homology 3 Domains". *Journal of Biological Chemistry* 289.11, pp. 7307–7319.

Campbell, Michael C and Sarah A Tishkoff (Sept. 2008). "African Genetic Diversity: Implications for Human Demographic History, Modern Human Origins, and Complex Disease Mapping". *Annu. Rev. Genom. Human Genet.* 9.1, pp. 403–433.

Cao, Wei, Xueyan Xi, Zhiyong Hao, Wenjing Li, Yan Kong, Lianxian Cui, Chi Ma, Denian Ba, and Wei He (June 2007). "RAET1E2, a soluble isoform of the UL16-binding protein RAET1E produced by tumor cells, inhibits NKG2D-mediated NK cytotoxicity." *J Biol Chem* 282.26, pp. 18922–18928.

Cardon, Lon R and Lyle J Palmer (Feb. 2003). "Population stratification and spurious allelic association". *The Lancet* 361.9357, pp. 598–604.

Carlson, Daniel F, Scott C Fahrenkrug, and Perry B Hackett (Jan. 2012). "Targeting DNA With Fingers and TALENs". *Nat Genet* 1.1, e3.

Cervera, R, M A Khamashta, J Font, G D Sebastiani, A Gil, P Lavilla, I Doménech, A O Aydintug, A Jedryka-Góral, and E de Ramón (Mar. 1993). "Systemic lupus erythematosus: clinical and immunologic patterns of disease expression in a cohort of 1,000 patients. The European Working Party on Systemic Lupus Erythematosus." *Medicine* 72.2, pp. 113–124.

Chahrour, Maria H, Timothy W Yu, Elaine T Lim, Bulent Ataman, Michael E Coulter, R Sean Hill, Christine R Stevens, Christian R Schubert, Michael E Greenberg, Stacey B Gabriel, and Christopher A Walsh (Apr. 2012). "Whole-Exome Sequencing and Homozygosity Analysis Implicate Depolarization-Regulated Neuronal Genes in Autism". *PLoS Genet* 8.4, e1002635.

Chakravarty, Eliza F, Thomas M Bush, Susan Manzi, Ann E Clarke, and Michael M Ward (June 2007). "Prevalence of adult systemic lupus erythematosus in California and Pennsylvania in 2000: estimates obtained using hospitalization data." *Arthritis Rheum* 56.6, pp. 2092–2094.

Chen, Michael E, Sue-Hwa Lin, Leland W K Chung, and Robert A Sikes (July 1998). "Isolation and Characterization of PAGE-1 and GAGE-7 : New Genes Ex-

pressed in the LNCaP Prostate Cancer Progression Model that Share Homology with Melanoma-Associated Antigens". *Journal of Biological Chemistry* 273.28, pp. 17618–17625.

Chen, Y. M., Y. Kikkawa, and J. H. Miner (May 2011). "A missense LAMB2 mutation causes congenital nephrotic syndrome by impairing laminin secretion". *J Am Soc Nephrol* 22.5, pp. 849–858.

Chiang, Chih-Kang and Reiko Inagi (Sept. 2010). "Glomerular diseases: genetic causes and future therapeutics." *Nature reviews. Nephrology* 6.9, pp. 539–554.

Chiang, Derek Y, Gad Getz, David B Jaffe, Michael J T O'Kelly, Xiaojun Zhao, Scott L Carter, Carsten Russ, Chad Nusbaum, Matthew Meyerson, and Eric S Lander (Nov. 2008). "High-resolution mapping of copy-number alterations with massively parallel sequencing". *Nat Methods* 6.1, pp. 99–103.

Cho, Kyoung-Won, Jae-Young Kim, Jae-Woo Cho, Kyu-Hyuk Cho, Chang-Woo Song, and Han-Sung Jung (Dec. 2008). "Point mutation of Hoxd12 in mice." *Yonsei medical journal* 49.6, pp. 965–972.

Christian G Specht, Ralf Schoepfer (2001). "Deletion of the alpha-synuclein locus in a subpopulation of C57BL/6J inbred mice". *BMC Neuroscience* 2, p. 11.

Chung, Sharon A et al. (Mar. 2011). "Differential genetic associations for systemic lupus erythematosus based on anti-dsDNA autoantibody production." *PLoS Genet* 7.3, e1001323.

Church, Deanna M et al. (May 2009). "Lineage-Specific Biology Revealed by a Finished Genome Assembly of the Mouse". *PLoS Biology* 7.5, e1000112.

Cirulli, Elizabeth T and David B Goldstein (June 2010). "Uncovering the roles of rare variants in common disease through whole-genome sequencing." *Nat Rev Genet* 11.6, pp. 415–425.

Clark, Peter E and Michael S Cookson (Oct. 2008). "The von Hippel-Lindau gene". *Cancer* 113.S7, pp. 1768–1778.

Coghill, E. L., A. Hugill, N. Parkinson, C. Davison, P. Glenister, S. Clements, J. Hunter, R. D. Cox, and S. D. Brown (Mar. 2002). "A gene-driven approach to the identification of ENU mutants in the mouse". *Nat Genet* 30.3, pp. 255–256.

Cohen, Jonathan C, Robert S Kiss, Alexander Pertsemlidis, Yves L Marcel, Ruth McPherson, and Helen H Hobbs (Aug. 2004). "Multiple rare alleles contribute to low plasma levels of HDL cholesterol." *Science* 305.5685, pp. 869–872.

Combarros, Onofre, Mario Cortina-Borja, A David Smith, and Donald J Lehmann (Sept. 2009). "Epistasis in sporadic Alzheimer's disease." *Neurobiology of Aging* 30.9, pp. 1333–1349.

Concepcion, D., K. L. Seburn, G. Wen, W. N. Frankel, and B. A. Hamilton (Oct. 2004). "Mutation rate and predicted phenotypic target sizes in ethylnitrosourea-treated mice". *Genetics* 168.2, pp. 953–959.

Cong, L, F A Ran, D Cox, S Lin, R Barretto, N Habib, P D Hsu, X. Wu, W Jiang, L A Marraffini, and F Zhang (Feb. 2013). "Multiplex Genome Engineering Using CRISPR/Cas Systems". *Science* 339.6121, pp. 819–823.

Conlon, Peter J, Kelvin Lynn, Michelle P Winn, L Darryl Quarles, Mary Lou Bembe, Margaret Pericak-Vance, Marcy Speer, David N Howell, and on behalf of the International Collaborative Group for the Study of Familial Focal and Segmental Glomerulosclerosis1 (Nov. 1999). "Spectrum of disease in familial focal and segmental glomerulosclerosis". *Kidney international* 56.5, pp. 1863–1871.

Consortium, The ENCODE Project (Sept. 2012). "An integrated encyclopedia of DNA elements in the human genome". *Nature* 489.7414, pp. 57–74.

Cook, M. C., C. G. Vinuesa, and C. C. Goodnow (Oct. 2006). "ENU-mutagenesis: insight into immune function and pathology". *Curr Opin Immunol* 18.5, pp. 627–633.

Cooper, Glinda S, Milele L K Bynum, and Emily C Somers (Nov. 2009). "Recent insights in the epidemiology of autoimmune diseases: Improved prevalence es-

timates and understanding of clustering of diseases". *Journal of Autoimmunity* 33.3-4, pp. 197–207.

Cooper, Gregory M and Jay Shendure (Aug. 2011). "Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data". *Nat Rev Genet* 12.9, pp. 628–640.

Cordes, S P (Sept. 2005). "N-Ethyl-N-Nitrosourea Mutagenesis: Boarding the Mouse Mutant Express". *Microbiology and Molecular Biology Reviews* 69.3, pp. 426–439.

Cox, Allison, Cheryl L Ackert-Bicknell, Beth L Dumont, Yueming Ding, Jordana Tzenova Bell, Gudrun A Brockmann, Jon E Wergedal, Carol Bult, Beverly Paigen, Jonathan Flint, Shirng-Wern Tsaih, Gary A Churchill, and Karl W Broman (Aug. 2009). "A new standard genetic map for the laboratory mouse". *Genetics* 182.4, pp. 1335–1344.

Croker, Ben, Karine Crozat, Michael Berger, Yu Xia, Sosathya Sovath, Lana Schaffer, Ioannis Eleftherianos, Jean-Luc Imler, and Bruce Beutler (Nov. 2007). "ATP-sensitive potassium channels mediate survival during infection in mammals and insects". *Nat Genet* 39.12, pp. 1453–1460.

Cui, Yong, Yujun Sheng, and Xuejun Zhang (Mar. 2013). "Genetic susceptibility to SLE: Recent progress from GWAS". *Journal of Autoimmunity* 41, pp. 25–33.

Cunninghame Graham, Deborah S, David L Morris, Tushar R Bhangale, Lindsey A Criswell, Ann-Christine Syvänen, Lars Rönnblom, Timothy W Behrens, Robert R Graham, and Timothy J Vyse (Oct. 2011). "PLOS Genetics: Association of NCF2, IKZF1, IRF8, IFIH1, and TYK2 with Systemic Lupus Erythematosus". *PLoS Genet* 7.10, e1002341.

Cunninghame Graham, Deborah S, Robert R Graham, Harinder Manku, Andrew K Wong, John C Whittaker, Patrick M Gaffney, Kathy L Moser, John D Rioux, David Altshuler, Timothy W Behrens, and Timothy J Vyse (Jan. 2008). "Poly-

morphism at the TNF superfamily gene TNFSF4 confers susceptibility to systemic lupus erythematosus." *Nat Genet* 40.1, pp. 83–89.

Dai, Chunsun, Donna B Stolz, Lawrence P Kiss, Satdarshan P Monga, Lawrence B Holzman, and Youhua Liu (Sept. 2009). "Wnt/β-Catenin Signaling Promotes Podocyte Dysfunction and Albuminuria". *Journal of the American Society of Nephrology* 20.9, pp. 1997–2008.

Danecek, Petr, Adam Auton, Goncalo Abecasis, Cornelis A Albers, Eric Banks, Mark A DePristo, Robert E Handsaker, Gerton Lunter, Gabor T Marth, Stephen T Sherry, Gilean McVean, Richard Durbin, and 1000 Genomes Project Analysis Group (Aug. 2011). "The variant call format and VCFtools". *Bioinformatics* 27.15, pp. 2156–2158.

Davis, Erica E et al. (Jan. 2011). "TTC21B contributes both causal and modifying alleles across the ciliopathy spectrum". *Nat Genet* 43.3, pp. 189–196.

Daxinger, Lucia et al. (2013). "An ENU mutagenesis screen identifies novel and known genes involved in epigenetic processes in the mouse". *Genome Biol* 14.9, R96.

Day-Williams, Aaron G et al. (Sept. 2011). "A variant in MCF2L is associated with osteoarthritis." *Am J Hum Genet* 89.3, pp. 446–450.

Deapen, D, A Escalante, L Weinrib, D Horwitz, B Bachman, P Roy-Burman, A Walker, and T M Mack (Mar. 1992). "A revised estimate of twin concordance in systemic lupus erythematosus." *Arthritis Rheum* 35.3, pp. 311–318.

Deen, W M, C R Bridges, B M Brenner, and B D Myers (Sept. 1985). "Heteroporous model of glomerular size selectivity: application to normal and nephrotic humans." *The American journal of physiology* 249.3 Pt 2, F374–89.

Dehainault, Catherine, Dorothée Michaux, Sabine Pagès-Berhouet, Virginie Caux-Moncoutier, François Doz, Laurence Desjardins, Jérôme Couturier, Philippe Parent, Dominique Stoppa-Lyonnet, Marion Gauthier-Villars, and Claude Houdayer

(Feb. 2007). “A deep intronic mutation in the RB1 gene leads to intronic sequence exonisation”. *Nat Genet* 15.4, pp. 473–477.

Delaneau, Olivier, Jean-Francois Zagury, and Jonathan Marchini (Dec. 2012). “Improved whole-chromosome phasing for disease and population genetic studies”. *Nat Methods* 10.1, pp. 5–6.

Deng, Y. and B. P. Tsao (Dec. 2010). “Genetic susceptibility to systemic lupus erythematosus in the genomic era”. *Nat Rev Rheumatol* 6.12, pp. 683–692.

Derry, J M, J A Kerns, K I Weinberg, H D Ochs, V Volpini, X Estivill, A P Walker, and U Francke (July 1995). “WASP gene mutations in Wiskott-Aldrich syndrome and X-linked thrombocytopenia.” *Hum Mol Genet* 4.7, pp. 1127–1135.

Devriendt, K., A. S. Kim, G. Mathijs, S. G. Frints, M. Schwartz, J. J. Van Den Oord, G. E. Verhoef, M. A. Boogaerts, J. P. Fryns, D. You, M. K. Rosen, and P. Vandenberghe (Mar. 2001). “Constitutively activating mutation in WASP causes X-linked severe congenital neutropenia”. *Nat Genet* 27.3, pp. 313–317.

Dickson, Samuel P, Kai Wang, Ian Krantz, Hakon Hakonarson, and David B Goldstein (Jan. 2010). “Rare Variants Create Synthetic Genome-Wide Associations”. *PLoS Biology* 8.1, e1000294 EP –.

Dieringer, Daniel and Christian Schlötterer (Jan. 2003). “Two Distinct Modes of Microsatellite Mutation Processes: Evidence From the Complete Genomic Sequences of Nine Species”. *Genome Research* 13.10, pp. 2242–2251.

Ding, Wen Y, Ania Koziell, Hugh J McCarthy, Agnieszka Bierzynska, Murali K Bhagavatula, Jan A Dudley, Carol D Inward, Richard J Coward, Jane Tizard, Christopher Reid, Corinne Antignac, Olivia Boyer, and Moin A Saleem (June 2014). “Initial steroid sensitivity in children with steroid-resistant nephrotic syndrome predicts post-transplant recurrence.” *Journal of the American Society of Nephrology* 25.6, pp. 1342–1348.

Dohm, J. C., C. Lottaz, T. Borodina, and H. Himmelbauer (Sept. 2008). "Substantial biases in ultra-short read data sets from high-throughput DNA sequencing". *Nucleic Acids Res* 36.16, e105–e105.

Dolinoy, Dana C, Dale Huang, and Randy L Jirtle (Aug. 2007). "Maternal Nutrient Supplementation Counteracts Bisphenol A-Induced DNA Hypomethylation in Early Development". *Proceedings of the National Academy of Sciences of the United States of America* 104.32, pp. 13056–13061.

Dörner, Thomas, Claudia Giesecke, and Peter E Lipsky (2011). "Mechanisms of B cell autoimmunity in SLE". *Arthritis Research & Therapy* 13.5, p. 243.

Duhl, D M, H Vrieling, K A Miller, G L Wolff, and G S Barsh (Sept. 1994). "Neomorphic agouti mutations in obese yellow mice." *Nat Genet* 8.1, pp. 59–65.

Durbin, R. M., G. R. Abecasis, D. L. Altshuler, A. Auton, L. D. Brooks, R. A. Gibbs, M. E. Hurles, and G. A. McVean (Oct. 2010). "A map of human genome variation from population-scale sequencing". *Nature* 467.7319, pp. 1061–1073.

Eddy, Sean R (Oct. 2004). "What is a hidden Markov model?" *Nature Biotechnology* 22.10, pp. 1315–1316.

Ekblom, P, M Miettinen, J Rapola, and J M Foidart (1982). "Demonstration of laminin, a basement membrane glycoprotein, in routinely processed formalin-fixed human tissues". *Histochemistry* 75.3, pp. 301–307.

Elston, R C and J Stewart (1971). "A General Model for the Genetic Analysis of Pedigree Data". *Human Heredity* 21.6, pp. 523–542.

Englert, C, M Vidal, S Maheswaran, Y Ge, R M Ezzell, K J Isselbacher, and D A Haber (Dec. 1995). "Truncated WT1 mutants alter the subnuclear localization of the wild-type protein." *Proceedings of the National Academy of Sciences of the United States of America* 92.26, pp. 11960–11964.

Ermann, Joerg and Laurie H Glimcher (Oct. 2012). "After GWAS: mice to the rescue?" *Current Opinion in Immunology* 24.5, pp. 564–570.

Evangelou, Evangelos, Thomas A Trikalinos, Georgia Salanti, and John P A Ioannidis (Aug. 2006). "Family-Based versus Unrelated Case-Control Designs for Genetic Associations". *PLoS Genet* 2.8, e123.

Evans, James P and Barbra B Rothschild (Apr. 2012). "Return of results: not that complicated?" *Genetics in medicine : official journal of the American College of Medical Genetics* 14.4, pp. 358–360.

Ewing, Brent and Phil Green (Jan. 1998). "Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities". *Genome Research* 8.3, pp. 186–194.

Eynde, B Van den, O Peeters, O De Backer, B Gaugler, S. Lucas, and T Boon (Sept. 1995). "A new family of genes coding for an antigen recognized by autologous cytolytic T lymphocytes on a human melanoma." *The Journal of Experimental Medicine* 182.3, pp. 689–698.

Fairfield, Heather et al. (2011). "Mutation discovery in mice by whole exome sequencing." *Genome Biol* 12.9, R86.

Favor, J, M Sund, A Neuhauser-Klaus, and U H Ehling (July 1990). "A dose-response analysis of ethylnitrosourea-induced recessive specific-locus mutations in treated spermatogonia of the mouse." *Mutation research* 231.1, pp. 47–54.

Feng, Di, Rivka C Stone, Maija-Leena Eloranta, Niquiche Sangster-Guity, Gunnel Nordmark, Snaevar Sigurdsson, Chuan Wang, Gunnar Alm, Ann-Christine Syvänen, Lars Rönnblom, and Betsy J Barnes (2010). "Genetic variants and disease-associated factors contribute to enhanced interferon regulatory factor 5 expression in blood cells of patients with systemic lupus erythematosus". *Arthritis & Rheumatism* 62.2, pp. 562–573.

Fernando, Michelle M A, Christine R Stevens, Pardis C Sabeti, Emily C Walsh, Alasdair J M McWhinnie, Anila Shah, Todd Green, John D Rioux, and Timothy J Vyse (Nov. 2007). "Identification of two independent risk factors for lupus within the MHC in United Kingdom families." *PLoS Genet* 3.11, e192.

Field, L Leigh, Rose Tobias, Wendy P Robinson, Richard Paisey, and Stephen Bain (Oct. 1998). "Maternal Uniparental Disomy of Chromosome 1 with No Apparent Phenotypic Effects". *The American Journal of Human Genetics* 63.4, pp. 1216–1220.

Fielitz, Jens, Mi-Sung Kim, John M Shelton, Xiaoxia Qi, Joseph A Hill, James A Richardson, Rhonda Bassel-Duby, and Eric N Olson (Feb. 2008). "Requirement of protein kinase D1 for pathological cardiac remodeling." *Proceedings of the National Academy of Sciences of the United States of America* 105.8, pp. 3059–3063.

Flynn, J. L., M. M. Goldstein, J. Chan, K. J. Triebold, K. Pfeffer, C. J. Lowenstein, R. Schreiber, T. W. Mak, and B. R. Bloom (June 1995). "Tumor necrosis factor-alpha is required in the protective immune response against Mycobacterium tuberculosis in mice". *Immunity* 2.6, pp. 561–572.

Folkvord, J M, D Viders, A Coleman-Smith, and R A Clark (Jan. 1989). "Optimization of immunohistochemical techniques to detect extracellular matrix proteins in fixed skin specimens." *The journal of histochemistry and cytochemistry : official journal of the Histochemistry Society* 37.1, pp. 105–113.

Font, J, R Cervera, G Espinosa, L Pallarés, M Ramos-Casals, S Jiménez, M García-Carrasco, L Seisdedos, and M Ingelmo (Aug. 1998). "Systemic lupus erythematosus (SLE) in childhood: analysis of clinical and immunological findings in 34 patients and comparison with SLE characteristics in adults." *Annals of the Rheumatic Diseases* 57.8, pp. 456–459.

Fontalba, Ana, Victor Martinez-Taboada, Olga Gutierrez, Carlos Pipaon, Natividad Benito, Alejandro Balsa, Ricardo Blanco, and Jose L Fernandez-Luna (Oct. 2007). "Deficiency of the NF-κB Inhibitor Caspase Activating and Recruitment Domain 8 in Patients with Rheumatoid Arthritis Is Associated with Disease Severity". *The Journal of Immunology* 179.7, pp. 4867–4873.

Fossati, L., M. Iwamoto, R. Merino, and S. Izui (Jan. 1995). "Selective enhancing effect of the Yaa gene on immune responses against self and foreign antigens." *European journal of immunology* 25.1, pp. 166–173.

Frazer, Kelly A, Sarah S Murray, Nicholas J Schork, and Eric J Topol (Apr. 2009). "Human genetic variation and its contribution to complex traits". *Nat Rev Genet* 10.4, pp. 241–251.

Fu, Wenqing, Timothy D O'Connor, Goo Jun, Hyun Min Kang, Goncalo Abecasis, Suzanne M Leal, Stacey Gabriel, David Altshuler, Jay Shendure, Deborah A Nickerson, Michael J Bamshad, NHLBI Exome Sequencing Project, and Joshua M Akey (Nov. 2012). "Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants". *Nature* 493.7431, pp. 216–220.

Fukui, Yoshinori, Osamu Hashimoto, Terukazu Sanui, Takamasa Oono, Hironori Koga, Masaaki Abe, Ayumi Inayoshi, Mayuko Noda, Masahiro Oike, Toshikazu Shirai, and Takehiko Sasazuki (Aug. 2001). "Haematopoietic cell-specific CDM family protein DOCK2 is essential for lymphocyte migration". *Nature* 412.6849, pp. 826–831.

Garg, Amit X, Bryce A Kiberd, William F Clark, R Brian Haynes, and Catherine M Clase (June 2002). "Albuminuria and renal insufficiency prevalence guides population screening: results from the NHANES III." *Kidney international* 61.6, pp. 2165–2175.

Gateva, Vesela et al. (Oct. 2009). "A large-scale replication study identifies TNIP1, PRDM1, JAZF1, UHRF1BP1 and IL10 as risk loci for systemic lupus erythematosus". *Nat Genet* 41.11, pp. 1228–1233.

George, Julie, Precious G Motshwene, Hui Wang, Andriy V Kubarenko, Anna Rautanen, Tara C Mills, Adrian V S Hill, Nicholas J Gay, and Alexander N R Weber (Jan. 2011). "Two Human MYD88 Variants, S34Y and R98C, Interfere with

MyD88-IRAK4-Myddosome Assembly". *Journal of Biological Chemistry* 286.2, pp. 1341–1353.

Gibson, Greg (Feb. 2011). "Rare and common variants: twenty arguments." *Nat Rev Genet* 13.2, pp. 135–145.

Gibson, William T, Rebecca L Hood, Shing Hei Zhan, Dennis E Bulman, Anthony P Fejes, Richard Moore, Andrew J Mungall, Patrice Eydoux, Riyana Babul-Hirji, Jianghong An, Marco A Marra, David Chitayat, Kym M Boycott, David D Weaver, and Steven J M Jones (Jan. 2012). "Mutations in EZH2 Cause Weaver Syndrome". *The American Journal of Human Genetics* 90.1, pp. 110–118.

Gigante, M, G Caridi, E Montemurno, M Soccio, M d'Apolito, G Cerullo, F Aucella, A Schirinzi, F Emma, L Massella, G Messina, T De Palo, E Ranieri, G M Ghiggeri, and L Gesualdo (July 2011). "TRPC6 Mutations in Children with Steroid-Resistant Nephrotic Syndrome and Atypical Phenotype". *Clinical journal of the American Society of Nephrology : CJASN* 6.7, pp. 1626–1634.

Gigante, Maddalena, Paola Pontrelli, Eustacchio Montemurno, Leonarda Roca, Filippo Aucella, Rosa Penza, GIANLUCA CARIDI, Elena Ranieri, GIAN MARCO GHIGGERI, and Loreto Gesualdo (June 2009). "CD2AP mutations are associated with sporadic nephrotic syndrome and focal segmental glomerulosclerosis (FSGS)." *Nephrol Dial Transplant* 24.6, pp. 1858–1864.

Giglio, Sabrina et al. (July 2014). "Heterogeneous Genetic Alterations in Sporadic Nephrotic Syndrome Associate with Resistance to Immunosuppression". *Journal of the American Society of Nephrology.*

Glazov, Evgeny A, Andreas Zankl, Marina Donskoi, Tony J Kenna, Gethin P Thomas, Graeme R Clark, Emma L Duncan, and Matthew A Brown (Mar. 2011). "Whole-Exome Re-Sequencing in a Family Quartet Identifies POP1 Mutations As the Cause of a Novel Skeletal Dysplasia". *PLoS Genet* 7.3, e1002027.

Goldstein, David B (2009). "Common Genetic Variation and Human Traits". *New England Journal of Medicine* 360.17, pp. 1696–1698.

Graham, D S Cunninghame et al. (Jan. 2008). "Association of LY9 in UK and Canadian SLE families". *Genes Immun* 9.2, pp. 93–102.

Graham, Deborah S Cunninghame (2009). "Genome-wide association studies in systemic lupus erythematosus: a perspective". *Arthritis Research & Therapy* 11.4, p. 119.

Green, Robert C, Jonathan S Berg, Wayne W Grody, Sarah S Kalia, Bruce R Korf, Christa L Martin, Amy L McGuire, Robert L Nussbaum, Julianne M O'Daniel, Kelly E Ormond, Heidi L Rehm, Michael S Watson, Marc S Williams, Leslie G Biesecker, and American College of Medical Genetics and Genomics (July 2013). *ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing.*

Greene, T A (Aug. 2003). "Cloning and Characterization of ALX, an Adaptor Downstream of CD28". *Journal of Biological Chemistry* 278.46, pp. 45128–45134.

Griffiths, E K, O Sanchez, P Mill, C Krawczyk, C V Hojilla, E Rubin, M M Nau, R Khokha, S Lipkowitz, C c Hui, and J M Penninger (Oct. 2003). "Cbl-3-Deficient Mice Exhibit Normal Epithelial Development". *Molecular and Cellular Biology* 23.21, pp. 7708–7718.

Groom, Joanna, Susan L Kalled, Anne H Cutler, Carl Olson, Stephen A Woodcock, Pascal Schneider, Jurg Tschopp, Teresa G Cachero, Marcel Batten, Julie Wheway, Davide Mauri, Dana Cavill, Tom P Gordon, Charles R Mackay, and Fabienne Mackay (Jan. 2002). "Association of BAFF/BLyS overexpression and altered B cell differentiation with Sjögren's syndrome". *J Clin Invest* 109.1, pp. 59–68.

Gruber, Thomas, Natascha Hermann-Kleiter, Christa Pfeifhofer-Obermair, Christina Lutz-Nicoladoni, Nikolaus Thuille, Thomas Letschka, Johannes Barsig, Monika Baudler, Jianping Li, Barbara Metzler, Barbara Nüsslein-Hildesheim, Juergen

Wagner, Michael Leitges, and Gottfried Baier (June 2009). "PKC theta cooperates with PKC alpha in alloimmune responses of T cells in vivo." *Molecular immunology* 46.10, pp. 2071–2079.

Gudbjartsson, Daniel F, Thorvaldur Thorvaldsson, Augustine Kong, Gunnar Gunnarsson, and Anna Ingolfsdottir (Oct. 2005). "Allegro version 2". *Nat Genet* 37.10, pp. 1015–1016.

Guergueltcheva, Velina et al. (Sept. 2012). "Autosomal-Recessive Congenital Cerebellar Ataxia Is Caused by Mutations in Metabotropic Glutamate Receptor 1". *The American Journal of Human Genetics* 91.3, pp. 553–564.

Guerra, Sandra G, Timothy J Vyse, and Deborah S Cunninghame Graham (2012). "The genetics of lupus: a functional perspective". *Arthritis Research & Therapy* 14.3, p. 211.

Guo, Yan, Jiang Li, Chung-I Li, Jirong Long, David C Samuels, and Yu Shyr (2012). "The effect of strand bias in Illumina short-read sequencing data". *BMC Genomics* 13.1, p. 666.

Gupta, Indra Rani, Cindy Baldwin, David Auguste, Kevin C H Ha, Jasmine El Andalousi, Somayyeh Fahiminiya, Martin Bitzan, Chantal Bernard, Mohammad Reza Akbari, Steven A Narod, David S Rosenblatt, Jacek Majewski, and Tomoko Takano (Jan. 2013). "ARHGDIA: a novel gene implicated in nephrotic syndrome". *Journal of Medical Genetics* 50.5, pp. 330–338.

Gusella, James F, Nancy S Wexler, P Michael Conneally, Susan L Naylor, Mary Anne Anderson, Rudolph E Tanzi, Paul C Watkins, Kathleen Ottina, Margaret R Wallace, Alan Y Sakaguchi, Anne B Young, Ira Shoulson, Ernesto Bonilla, and Joseph B Martin (Nov. 1983). "A polymorphic DNA marker genetically linked to Huntington's disease". *Nature* 306.5940, pp. 234–238.

Haffner, Marlene E, Janet Whitley, and Marie Moses (Oct. 2002). "Outlook: Two decades of orphan product development". *Nature Reviews Drug Discovery* 1.10, pp. 821–825.

Harley, Isaac T W, Kenneth M Kaufman, Carl D Langefeld, John B Harley, and Jennifer A Kelly (May 2009). "Genetic susceptibility to SLE: new insights from fine mapping and genome-wide association studies". *Nat Rev Genet* 10.5, pp. 285–290.

Harley, J. B. et al. (Feb. 2008). "Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in ITGAM, PXK, KIAA1542 and other loci". *Nat Genet* 40.2, pp. 204–210.

Harris, J. and J. Keane (July 2010). "How tumour necrosis factor blockers interfere with tuberculosis immunity". *Clin Exp Immunol* 161.1, pp. 1–9.

Hartley, Jane Louise et al. (June 2010). "Mutations in TTC37 cause trichohepatoenteric syndrome (phenotypic diarrhea of infancy)." *Gastroenterology* 138.7, 2388–98–2398.e1–2.

Hasselbacher, K. et al. (Sept. 2006). "Recessive missense mutations in LAMB2 expand the clinical spectrum of LAMB2-associated disorders." *Kidney international* 70.6, pp. 1008–1012.

He, W, C Dai, Y Li, G Zeng, S P Monga, and Y Liu (Mar. 2009). "Wnt/ -Catenin Signaling Promotes Renal Interstitial Fibrosis". *Journal of the American Society of Nephrology* 20.4, pp. 765–776.

Heeringa, Saskia F, Clemens C Möller, Jianyang Du, Lixia Yue, Bernward Hinkes, Gil Chernin, Christopher N Vlangos, Peter F Hoyer, Jochen Reiser, and Friedhelm Hildebrandt (Nov. 2009). "A Novel TRPC6 Mutation That Causes Childhood FSGS". *PloS one* 4.11, e7771.

Hibbs, M. L., D M Tarlinton, J Armes, D Grail, G Hodgson, R Maglitto, S A Stacker, and A R Dunn (Oct. 1995). "Multiple defects in the immune system of Lyn-deficient mice, culminating in autoimmune disease." *Cell* 83.2, pp. 301–311.

Hicks, Stephanie, David A Wheeler, Sharon E Plon, and Marek Kimmel (Apr. 2011). "Prediction of missense mutation functionality depends on both the algorithm and sequence alignment employed". *Human mutation* 32.6, pp. 661–668.

Hill, Kathleen A, Asanga Halangoda, Petra W Heinmoeller, Kelly Gonzalez, Chaniga Chitaphan, Jeffrey Longmate, William A Scaringe, Ji-Cheng Wang, and Steve S Sommer (June 2005). "Tissue-specific time courses of spontaneous mutation frequency and deviations in mutation pattern are observed in middle to late adulthood in Big Blue mice." *Environmental and Molecular Mutagenesis* 45.5, pp. 442–454.

Hinkes, B. G., B. Mucha, C. N. Vlangos, R. Gbadegesin, J. Liu, K. Hasselbacher, D. Hangan, F. Ozaltin, M. Zenker, and F. Hildebrandt (Apr. 2007). "Nephrotic syndrome in the first year of life: two thirds of cases are caused by mutations in 4 genes (NPHS1, NPHS2, WT1, and LAMB2)". *Pediatrics* 119.4, e907–19.

Hinkes, Bernward et al. (Nov. 2006). "Positional cloning uncovers mutations in PLCE1 responsible for a nephrotic syndrome variant that may be reversible". *Nat Genet* 38.12, pp. 1397–1405.

Hochberg, M C (Sept. 1997). "Updating the American College of Rheumatology revised criteria for the classification of systemic lupus erythematosus." *Arthritis Rheum* 40.9, p. 1725.

Hoebe, K and B. Beutler (May 2005). "Unraveling innate immunity using large scale N-ethyl-N-nitrosourea mutagenesis." *Tissue antigens* 65.5, pp. 395–401.

Hoshino, K, O Takeuchi, T Kawai, H Sanjo, T Ogawa, Y Takeda, K Takeda, and S Akira (Apr. 1999). "Cutting edge: Toll-like receptor 4 (TLR4)-deficient mice

are hyporesponsive to lipopolysaccharide: evidence for TLR4 as the Lps gene product." *J Immunol* 162.7, pp. 3749–3752.

Hoyne, Gerard F and Christopher C Goodnow (2006). "The use of genomewide ENU mutagenesis screens to unravel complex mammalian traits: identifying genes that regulate organ-specific and systemic autoimmunity". *Immunological reviews* 210.1, pp. 27–39.

Humbert, Camille et al. (Feb. 2014). "Integrin Alpha 8 Recessive Mutations Are Responsible for Bilateral Renal Agenesis in Humans". *The American Journal of Human Genetics* 94.2, pp. 288–294.

Hunt, Karen A et al. (May 2013). "Negligible impact of rare autoimmune-locus coding-region variants on missing heritability". *Nature* 498.7453, pp. 232–235.

Ibarra-Soria, Ximena, Maria O Levitin, Luis R Saraiva, and Darren W Logan (Sept. 2014). "The Olfactory Transcriptomes of Mice". *PLoS Genet* 10.9, e1004593.

Imai, Kohsuke, Tomohiro Morio, Yi Zhu, Yinzhu Jin, Sukeyuki Itoh, Michiko Kajiwara, Jun-Ichi Yata, Shuki Mizutani, Hans D Ochs, and Shigeaki Nonoyama (Jan. 2004). "Clinical course of patients with WASP gene mutations." *Blood* 103.2, pp. 456–464.

International HapMap 3 Consortium et al. (Sept. 2010). "Integrating common and rare genetic variation in diverse human populations." *Nature* 467.7311, pp. 52–58.

International HapMap Consortium (Oct. 2005). "A haplotype map of the human genome." *Nature* 437.7063, pp. 1299–1320.

Isenberg, D A, J J Manson, M R Ehrenstein, and A Rahman (May 2007). "Fifty years of anti-ds DNA antibodies: are we approaching journey's end?" *Rheumatology (Oxford)* 46.7, pp. 1052–1056.

Ishii, H, R Baffa, S I Numata, Y Murakumo, S Rattan, H Inoue, M Mori, V Fidanza, H Alder, and C M Croce (Mar. 1999). "The FEZ1 gene at chromosome 8p22 encodes a leucine-zipper protein, and its expression is altered in multiple human

tumors." *Proceedings of the National Academy of Sciences of the United States of America* 96.7, pp. 3928–3933.

Ishiyama, Akira, Sarah E Mowry, Ivan A Lopez, and Gail Ishiyama (Aug. 2009). "Immunohistochemical distribution of basement membrane proteins in the human inner ear from older subjects". *Hearing Research* 254.1-2, pp. 1–14.

Izui, S., M Higaki, D Morrow, and R. Merino (June 1988). "The Y chromosome from autoimmune BXSB/MpJ mice induces a lupus-like syndrome in (NZW x C57BL/6)F1 male mice, but not in C57BL/6 male mice." *European journal of immunology* 18.6, pp. 911–915.

Izui, S., M. Iwamoto, L. Fossati, R. Merino, S. Takahashi, and N. Ibnou-Zekri (Apr. 1995). "The Yaa gene model of systemic lupus erythematosus". *Immunol Rev* 144, pp. 137–156.

Javierre, B M et al. (Feb. 2010). "Changes in the pattern of DNA methylation associate with twin discordance in systemic lupus erythematosus". *Genome Research* 20.2, pp. 170–179.

Jensen-Seaman, M. I., T. S. Furey, B. A. Payseur, Y. Lu, K. M. Roskin, C. F. Chen, M. A. Thomas, D. Haussler, and H. J. Jacob (Apr. 2004). "Comparative recombination rates in the rat, mouse, and human genomes". *Genome Res* 14.4, pp. 528–538.

John A Stamatoyannopoulos, Michael Snyder Ross Hardison Bing Ren Thomas Gingeras David M Gilbert Mark Groudine Michael Bender Rajinder Kaul Theresa Canfield Erica Giste Audra Johnson Mia Zhang Gayathri Balasundaram Rachel Byron Vaughan Roach Peter J Sabo Richard Sandstrom A Sandra Stehling Robert E Thurman Sherman M Weissman Philip Cayting Manoj Hariharan Jin Lian Yong Cheng Stephen G Landt Zhihai Ma Barbara J Wold Job Dekker Gregory E Crawford Cheryl A Keller Weisheng Wu Christopher Morrissey Swathi A Kumar Tejaswini Mishra Deepti Jain Marta Byrska-Bishop Daniel Blankenberg Bryan R

Lajoie1 Gaurav Jain Amartya Sanyal Kaun-Bei Chen Olgert Denas James Taylor Gerd A Blobel Mitchell J Weiss Max Pimkin Wulan Deng Georgi K Marinov Brian A Williams Katherine I Fisher-Aylor Gilberto Desalvo Anthony Kiralusha Diane Trout Henry Amrhein Ali Mortazavi Lee Edsall David McCleary Samantha Kuan Yin Shen Feng Yue Zhen Ye Carrie A Davis Chris Zaleski Sonali Jha Chenghai Xue Alex Dobin Wei Lin Meagan Fastuca Huaien Wang Roderic Guigo Sarah Djebali Julien Lagarde Tyrone Ryba Takayo Sasaki Venkat S Malladi Melissa S Cline Vanessa M Kirkup Katrina Learned Kate R Rosenbloom W James Kent Elise A Feingold Peter J Good Michael Pazin Rebecca F Lowdon Leslie B Adams (2012). "An encyclopedia of mouse DNA elements (Mouse ENCODE)". *Genome Biology* 13.8, p. 418.

Johnson, Angela E, Caroline Gordon, Robert G Palmer, and Paul A Bacon (Apr. 1995). "The Prevalence and Incidence of Systemic Lupus-Erythematosus in Birmingham, England - Relationship to Ethnicity and Country of Birth". *Arthritis Rheum* 38.4, pp. 551–558.

Jong, R de, F Altare, I A Haagen, D G Elferink, T Boer, P J van Breda Vriesman, P J Kabel, J M Draaisma, J T van Dissel, F P Kroon, J L Casanova, and T H Ottenhoff (May 1998). "Severe mycobacterial and Salmonella infections in interleukin-12 receptor-deficient patients." *Science* 280.5368, pp. 1435–1438.

Jostins, Luke, Katherine I Morley, and Jeffrey C Barrett (Mar. 2011). "Imputation of low-frequency variants using the HapMap3 benefits from large, diverse reference sets". *European journal of human genetics : EJHG* 19.6, pp. 662–666.

Justice, M. J., D A Carpenter, J Favor, A Neuhauser-Klaus, M. Hrabe de Angelis, D Soewarto, A Moser, S Cordes, D Miller, V Chapman, J. S. Weber, E M Rinchik, P. R. Hunsicker, W. L. Russell, and V C Bode (July 2000). "Effects of ENU dosage on mouse strains." *Mamm Genome* 11.7, pp. 484–488.

Justice, M. J., J. K. Noveroske, J. S. Weber, B. Zheng, and A. Bradley (1999). "Mouse ENU mutagenesis". *Hum Mol Genet* 8.10, pp. 1955–1963.

Kagan, Mikhail, Arthur H Cohen, Verena Matejas, Christopher Vlangos, and Martin Zenker (Feb. 2008). "A milder variant of Pierson syndrome." *Pediatr Nephrol* 23.2, pp. 323–327.

Kamphans, Tom, Peggy Sabri, Na Zhu, Verena Heinrich, Stefan Mundlos, Peter N Robinson, Dmitri Parkhomchuk, and Peter M Krawitz (Aug. 2013). "Filtering for Compound Heterozygous Sequence Variants in Non-Consanguineous Pedigrees". *PloS one* 8.8, e70151.

Kannel, William B, Meir J Stampfer, William P Castelli, and Joel Verter (Nov. 1984). "The prognostic significance of proteinuria: The Framingham study". *American Heart Journal* 108.5, pp. 1347–1352.

Kaplan, J M, S H Kim, K N North, H Rennke, L A Correia, H Q Tong, B J Mathis, J C Rodríguez-Pérez, P G Allen, A H Beggs, and M. R. Pollak (Mar. 2000). "Mutations in ACTN4, encoding alpha-actinin-4, cause familial focal segmental glomerulosclerosis." *Nat Genet* 24.3, pp. 251–256.

Karl W Broman, James L Weber (Dec. 1999). "Long Homozygous Chromosomal Segments in Reference Families from the Centre d'Étude du Polymorphisme Humain". *American Journal of Human Genetics* 65.6, p. 1493.

Katsanis, Sara Huston and Nicholas Katsanis (May 2013). "Molecular genetic testing and the future of clinical genomics". *Nat Rev Genet* 14.6, pp. 415–426.

Keane, J., S. Gershon, R. P. Wise, E. Mirabile-Levens, J. Kasznica, W. D. Schwieterman, J. N. Siegel, and M. M. Braun (Oct. 2001). "Tuberculosis associated with infliximab, a tumor necrosis factor alpha-neutralizing agent". *N Engl J Med* 345.15, pp. 1098–1104.

Keane, T. M. et al. (Sept. 2011). "Mouse genomic variation and its effect on phenotypes and gene regulation". *Nature* 477.7364, pp. 289–294.

Keinan, A and A G Clark (May 2012). "Recent Explosive Human Population Growth Has Resulted in an Excess of Rare Genetic Variants". *Science* 336.6082, pp. 740–743.

Kestilä, M, U Lenkkeri, M Männikkö, J Lamerdin, P McCready, H Putaala, V Ruotsalainen, T. Morita, M Nissinen, R Herva, C E Kashtan, L Peltonen, C Holmberg, A Olsen, and K Tryggvason (Mar. 1998). "Positionally cloned gene for a novel glomerular protein–nephrin–is mutated in congenital nephrotic syndrome." *Molecular cell* 1.4, pp. 575–582.

Khurana, Ekta et al. (Oct. 2013). "Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics". *Science* 342.6154, pp. 1235587–1235587.

Kircher, Martin, Udo Stenzel, and Janet Kelso (2009). "Improved base calling for the Illumina Genome Analyzer using machine learning strategies". *Genome Biol* 10.8, R83.

Klein, J. C., M. J. Bleeker, J. T. Lutgerink, W. J. van Dijk, H. F. Brugghe, H. van den Elst, G. A. van der Marel, J. H. van Boom, J. G. Westra, A. J. Berns, and et al (July 1990). "Use of shuttle vectors to study the molecular processing of defined carcinogen-induced DNA damage: mutagenicity of single O4-ethylthymine adducts in HeLa cells". *Nucleic Acids Res* 18.14, pp. 4131–4137.

Koboldt, D C, L Ding, E. R. Mardis, and R. K. Wilson (Sept. 2010). "Challenges of sequencing human genomes". *Briefings in Bioinformatics* 11.5, pp. 484–498.

Kohane, Isaac S, Michael Hsing, and Sek Won Kong (Feb. 2012). "Taxonomizing, sizing, and overcoming the incidentalome". *Genetics in medicine : official journal of the American College of Medical Genetics* 14.4, pp. 399–404.

Kondrashov, Alexey S (Dec. 2002). "Direct estimates of human per nucleotide mutation rates at 20 loci causing mendelian diseases". *Human mutation* 21.1, pp. 12–27.

Kong, Augustine et al. (Aug. 2012). "Rate of de novo mutations and the importance of father's age to disease risk". *Nature* 488.7412, pp. 471–475.

Kontaki, Elena and Dimitrios T Boumpas (Nov. 2010). "Innate immunity in systemic lupus erythematosus: Sensing endogenous nucleic acids". *Journal of Autoimmunity* 35.3, pp. 206–211.

Kopp, Jeffrey B (Jan. 2013). "An Expanding Universe of FSGS Genes and Phenotypes: LMX1B Mutations Cause Familial Autosomal Dominant FSGS Lacking Extrarenal Manifestations". *Journal of the American Society of Nephrology* 24.8, pp. 1183–1185.

Kozyrev, Sergey V et al. (Feb. 2008). "Functional variants in the B-cell gene BANK1 are associated with systemic lupus erythematosus." *Nat Genet* 40.2, pp. 211–216.

Kronenberg, Florian, Arno Lingenhel, Karl Lhotta, Barbara Rantner, Martina F Kronenberg, Paul Konig, Joachim Thiery, Michael Koch, Arnold von Eckardstein, and Hans Dieplinger (July 2004). "Lipoprotein(a)- and low-density lipoprotein-derived cholesterol in nephrotic syndrome: Impact on lipid-lowering therapy?" *Kidney international* 66.1, pp. 348–354.

Kruglyak, L., M. J. Daly, M. P. Reeve-Daly, and E. S. Lander (June 1996). "Parametric and nonparametric linkage analysis: a unified multipoint approach". *Am J Hum Genet* 58.6, pp. 1347–1363.

Kuleshov, Volodymyr, Dan Xie, Rui Chen, Dmitry Pushkarev, Zhihai Ma, Tim Blauwkamp, Michael Kertesz, and Michael Snyder (Feb. 2014). "Whole-genome haplotyping using long reads and statistical methods". *Nat Biotechnol* 32.3, pp. 261–266.

Kumar, P., S. Henikoff, and Ng, P. C. (2009). "Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm". *Nat Protoc* 4.7, pp. 1073–1081.

Lambe, Teresa, Greg Crawford, Andy L Johnson, Tanya L Crockford, Tiphaine Bouriez-Jones, Aisling M Smyth, Trung H M Pham, Qian Zhang, Alexandra F Freeman,

Jason G Cyster, Helen C Su, and Richard J Cornall (Dec. 2011). "DOCK8 is essential for T-cell survival and the maintenance of CD8+ T-cell memory." *European journal of immunology* 41.12, pp. 3423–3435.

Lamprianou, S, N Vacaresse, Y Suzuki, H Meziane, J D Buxbaum, J Schlessinger, and S Harroch (June 2006). "Receptor Protein Tyrosine Phosphatase Is a Marker for Pyramidal Cells and Sensory Neurons in the Nervous System and Is Not Necessary for Normal Development". *Molecular and Cellular Biology* 26.13, pp. 5106–5119.

Lander, E. S. and D. Botstein (June 1987). "Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children." *Science* 236.4808, pp. 1567–1570.

Lander, E. S. and P. Green (Apr. 1987). "Construction of multilocus genetic linkage maps in humans". *Proc Natl Acad Sci U S A* 84.8, pp. 2363–2367.

Lander, E. S. et al. (Feb. 2001). "Initial sequencing and analysis of the human genome." *Nature* 409.6822, pp. 860–921.

Lander, Eric S and Michael S Waterman (Oct. 1988). "Genomic mapping by fingerprinting random clones: A mathematical analysis". *Genomics* 2.3, pp. 231–239.

Latif, F, K Tory, J Gnarra, M Yao, F M Duh, M L Orcutt, T Stackhouse, I Kuzmin, W Modi, and L Geil (May 1993). "Identification of the von Hippel-Lindau disease tumor suppressor gene." *Science* 260.5112, pp. 1317–1320.

Lee, Joo Hoon, Kyoung Hee Han, HyunKyung Lee, Hee Gyung Kang, Kyung Chul Moon, Jae Il Shin, Hyewon Hahn, Young Seo Park, Ki Soo Pai, Byoung-Soo Cho, Su-Yung Kim, Seung Joo Lee, Il Soo Ha, Yong Choi, and Hae Il Cheong (Dec. 2011). "Genetic basis of congenital and infantile nephrotic syndromes." *American journal of kidney diseases : the official journal of the National Kidney Foundation* 58.6, pp. 1042–1043.

Lee, Y H, J D Ji, and G G Song (July 2009). "Fcgamma receptor IIB and IIIB polymorphisms and susceptibility to systemic lupus erythematosus and lupus nephritis: a meta-analysis." *Lupus* 18.8, pp. 727–734.

Lee, Young-Kwang, Joseph W Brewer, Rachel Hellman, and Linda M Hendershot (July 1999). "BiP and Immunoglobulin Light Chain Cooperate to Control the Folding of Heavy Chain and Ensure the Fidelity of Immunoglobulin Assembly". *Molecular Biology of the Cell* 10.7, pp. 2209–2219.

Lehnhardt, A., A. Lama, K. Amann, V. Matejas, M. Zenker, and M. J. Kemper (Jan. 2012). "Pierson syndrome in an adolescent girl with nephrotic range proteinuria but a normal GFR". *Pediatr Nephrol*.

Leidenroth, Andreas, Hanne S oslash rmo Sorte, Gregor Gilfillan, Melanie Ehrlich, Robert Lyle, and Jane E Hewitt (Sept. 2012). "Diagnosis by sequencing: correction of misdiagnosis from FSHD2 to LGMD2A by whole-exome analysis". *European journal of human genetics : EJHG* 20.9, pp. 999–1003.

Leshchiner, Ignaty et al. (July 2012). "Mutation mapping and identification by whole-genome sequencing." *Genome Res*.

Lewis, Edmund J, Lawrence G Hunsicker, William R Clarke, Tomas Berl, Marc A Pohl, Julia B Lewis, Eberhard Ritz, Robert C Atkins, Richard Rohde, and Itamar Raz (Sept. 2001). "Renoprotective Effect of the Angiotensin-Receptor Antagonist Irbesartan in Patients with Nephropathy Due to Type 2 Diabetes". *N Engl J Med* 345.12, pp. 851–860.

Lewis, M. A., E. Quint, A. M. Glazier, H. Fuchs, M. H. De Angelis, C. Langford, S. van Dongen, C. Abreu-Goodger, M. Piipari, N. Redshaw, T. Dalmay, M. A. Moreno-Pelayo, A. J. Enright, and K. P. Steel (May 2009). "An ENU-induced mutation of miR-96 associated with progressive hearing loss in mice". *Nat Genet* 41.5, pp. 614–618.

Li, Biao, Vidhya G Krishnan, Matthew E Mort, Fuxiao Xin, Kishore K Kamati, David N Cooper, Sean D Mooney, and Predrag Radivojac (Nov. 2009a). "Automated inference of molecular mechanisms of disease from amino acid substitutions." *Bioinformatics* 25.21, pp. 2744–2750.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin (Aug. 2009b). "The Sequence Alignment/Map format and SAMtools". *Bioinformatics* 25.16, pp. 2078–2079.

Li, Miao-Xin, Johnny S H Kwan, Su-Ying Bao, Wanling Yang, Shu-Leong Ho, Yong-Qiang Song, and Pak C Sham (Jan. 2013). "Predicting Mendelian Disease-Causing Non-Synonymous Single Nucleotide Variants in Exome Sequencing Studies". *PLoS Genet* 9.1, e1003143.

Li, Wenyuan, Chao Dai, Shuli Kang, and Xianghong Jasmine Zhou (June 2014). "Integrative analysis of many RNA-seq datasets to study alternative splicing". *Methods* 67.3, pp. 313–324.

Li, Yingrui et al. (Oct. 2010). "Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants". *Nat Genet* 42.11, pp. 969–972.

Lipman, D and W Pearson (Mar. 1985). "Rapid and sensitive protein similarity searches". *Science* 227.4693, pp. 1435–1441.

Liu, Eric Yi, Mingyao Li, Wei Wang, and Yun Li (Jan. 2013). "MaCH-admix: genotype imputation for admixed populations." *Genetic epidemiology* 37.1, pp. 25–37.

Liu, Kui et al. (Apr. 2009). "Kallikrein genes are associated with lupus and glomerular basement membrane–specific antibody–induced nephritis in mice and humans". *J Clin Invest* 119.4, pp. 911–923.

Lomas, David (2013). *Mapping 100,000 genomes: strategic priorities, data and ethics - Publications - GOV.UK*. Tech. rep. UK Gov.

Lu, R et al. (July 2009). "Genetic associations of LYN with systemic lupus erythematosus." *Genes Immun* 10.5, pp. 397–403.

Lunter, G. and M. Goodson (June 2011). "Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads". *Genome Res* 21.6, pp. 936–939.

Lupski, James R et al. (Apr. 2010). "Whole-Genome Sequencing in a Patient with Charcot–Marie–Tooth Neuropathy". *N Engl J Med* 362.13, pp. 1181–1191.

Lynch, M (Jan. 2010). "Rate, molecular spectrum, and consequences of human mutation". *Proc Natl Acad Sci U S A* 107.3, pp. 961–968.

MacEwan, D. J., R. Mitchell, M. S. Johnson, F. J. Thomson, E. M. Lutz, R. A. Clegg, and K. Connor (June 1993). "Evidence that protein kinase C alpha has reduced affinity towards 1,2-dioctanoyl-sn-glycerol: the effects of lipid activators on phorbol ester binding and kinase activity". *Eur J Pharmacol* 246.1, pp. 9–18.

Machuca, Eduardo, Geneviève Benoit, and Corinne Antignac (Oct. 2009). "Genetics of nephrotic syndrome: connecting molecular genetics to podocyte physiology." *Hum Mol Genet* 18.R2, R185–94.

Mackay, Fabienne and Pascal Schneider (July 2009). "Cracking the BAFF code". *Nature Reviews Immunology* 9.7, pp. 491–502.

Mackay, Fabienne, Stephen A Woodcock, Pornsri Lawton, Christine Ambrose, Manfred Baetscher, Pascal Schneider, Jurg Tschopp, and Jeffrey L Browning (Dec. 1999). "Mice Transgenic for Baff Develop Lymphocytic Disorders along with Autoimmune Manifestations". *The Journal of Experimental Medicine* 190.11, pp. 1697–1710.

MacPherson, Matthew, Hwee San Lek, Alan Prescott, and Susanna C Fagerholm (May 2011). "A systemic lupus erythematosus-associated R77H substitution in the CD11b chain of the Mac-1 integrin compromises leukocyte adhesion and phagocytosis." *Journal of Biological Chemistry* 286.19, pp. 17303–17310.

Maiuri, M Chiara, Einat Zalckvar, Adi Kimchi, and Guido Kroemer (Sept. 2007). "Self-eating and self-killing: crosstalk between autophagy and apoptosis". *Nature Reviews Molecular Cell Biology* 8.9, pp. 741–752.

Mali, Prashant, Luhan Yang, Kevin M Esvelt, John Aach, Marc Guell, James E DiCarlo, Julie E Norville, and George M Church (Feb. 2013). "RNA-guided human genome engineering via Cas9." *Science* 339.6121, pp. 823–826.

Manderson, Anthony P, Marina Botto, and Mark J Walport (2004). "The Role of Complement in the Development of Systemic Lupus Erythematosus". *Annual review of immunology* 22.1, pp. 431–456.

Manolio, Teri A et al. (Oct. 2009). "Finding the missing heritability of complex diseases." *Nature* 461.7265, pp. 747–753.

Margulies, Marcel et al. (July 2005). "Genome sequencing in microfabricated high-density picolitre reactors". *Nature Cell Biology* 437.7057, p. 376.

Markus, B., O. S. Birk, and D. Geiger (Oct. 2011). "Integration of SNP genotyping confidence scores in IBD inference". *Bioinformatics* 27.20, pp. 2880–2887.

Marth, Gabor T et al. (2011). "The functional spectrum of low-frequency coding variation". *Genome Biol* 12.9, R84.

Martínez Valle, Fernando, Eva Balada, Josep Ordi-Ros, and Miquel Vilardell-Tarres (May 2008). "DNase 1 and systemic lupus erythematosus". *Autoimmunity Reviews* 7.5, pp. 359–363.

Masuya, Hiroshi et al. (Feb. 2007). "A series of ENU-induced single-base substitutions in a long-range cis-element altering Sonic hedgehog expression in the developing mouse limb bud." *Genomics* 89.2, pp. 207–214.

Mathieson, Iain and Gil McVean (Mar. 2012). "Differential confounding of rare and common variants in spatially structured populations." *Nat Genet* 44.3, pp. 243–246.

Matthews, E, R Labrum, M G Sweeney, R Sud, A Haworth, P F Chinnery, G Meola, S Schorge, D M Kullmann, M B Davis, and M G Hanna (May 2009). "Voltage sensor charge loss accounts for most cases of hypokalemic periodic paralysis." *Neurology* 72.18, pp. 1544–1547.

McCarthy, H J, A Bierzynska, M Wherlock, G I Welsh, and M A Saleem (Jan. 2012). "Next generation sequencing (NGS) in the UK steroid resistant nephrotic syndrome (SRNS) study reveals complex genetic heterogeneity". *Archives of Disease in Childhood* 97.Suppl 1, A161–A162.

McCarthy, Hugh J et al. (Jan. 2013). "Simultaneous Sequencing of 24 Genes Associated with Steroid-Resistant Nephrotic Syndrome". *Clinical Journal of the American Society of Nephrology* 8.4, pp. 637–648.

McClellan, Jon and Mary-Claire King (Apr. 2010). "Genetic Heterogeneity in Human Disease". *Cell* 141.2, pp. 210–217.

McGary, Kriston L, Tae Joo Park, John O Woods, Hye Ji Cha, John B Wallingford, and Edward M Marcotte (Apr. 2010). "Systematic discovery of nonobvious human disease models through orthologous phenotypes". *Proc Natl Acad Sci U S A* 107.14, pp. 6544–6549.

McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo (Sept. 2010). "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data". *Genome Res* 20.9, pp. 1297–1303.

McVean, Gil A et al. (Oct. 2012). "An integrated map of genetic variation from 1,092 human genomes". *Nature* 491.7422, pp. 56–65.

Metzker, M. L. (Jan. 2010). "Sequencing technologies - the next generation". *Nat Rev Genet* 11.1, pp. 31–46.

Miner, Jeffrey H (May 2012). "The glomerular basement membrane." *Exp Cell Res* 318.9, pp. 973–978.

Mohney, Brian G et al. (June 2011). "A novel mutation of LAMB2 in a multigenerational mennonite family reveals a new phenotypic variant of Pierson syndrome." *Ophthalmology* 118.6, pp. 1137–1144.

Montserrat Moliner, A and J Waligora (2013). "The European union policy in the field of rare diseases." *Public health genomics* 16.6, pp. 268–277.

Morel, Laurence (May 2010). "Genetics of SLE: evidence from mouse models". *Nat Rev Rheumatol* 6.6, pp. 348–357.

Morrison, Alanna C et al. (June 2013). "Whole-genome sequence–based analysis of high-density lipoprotein cholesterol". *Nat Genet* 45.8, pp. 899–901.

Morton, Newton E (Sept. 1955). "Sequential tests for the detection of linkage". *American Journal of Human Genetics* 7.3, p. 277.

Moser, A R, W F Dove, K A Roth, and J I Gordon (Mar. 1992). "The Min (multiple intestinal neoplasia) mutation: its effect on gut epithelial cell differentiation and interaction with a modifier system." *The Journal of Cell Biology* 116.6, pp. 1517–1526.

Moulton, Vaishali R and George C Tsokos (2011). "Abnormalities of T cell signaling in systemic lupus erythematosus". *Arthritis Research & Therapy* 13.2, p. 207.

Muddyman, Dawn, Carol Smee, Heather Griffin, and Jane Kaye (2013). "Implementing a successful data-management framework: the UK10K managed access model." *Genome Medicine* 5.11, p. 100.

Musone, Stacy L et al. (Sept. 2008). "Multiple polymorphisms in the TNFAIP3 region are independently associated with systemic lupus erythematosus." *Nat Genet* 40.9, pp. 1062–1064.

Nasr, S H, S J Galgano, G S Markowitz, M B Stokes, and V D D'Agati (Oct. 2006). "Immunofluorescence on pronase-digested paraffin sections: A valuable salvage technique for renal biopsies". *Kidney international* 70.12, pp. 2148–2151.

Navin, Nicholas, Jude Kendall, Jennifer Troge, Peter Andrews, Linda Rodgers, Jeanne McIndoo, Kerry Cook, Asya Stepansky, Dan Levy, Diane Esposito, Lakshmi Muthuswamy, Alex Krasnitz, W Richard McCombie, James Hicks, and Michael Wigler (Apr.

2011). “Tumour evolution inferred by single-cell sequencing”. *Nature* 472.7341, pp. 90–94.

Neale, Benjamin M et al. (Apr. 2012). “Patterns and rates of exonic de novo mutations in autism spectrum disorders”. *Nature* 485.7397, pp. 242–245.

Nelms, K. A. and C. C. Goodnow (Sept. 2001). “Genome-wide ENU mutagenesis to reveal immune regulators”. *Immunity* 15.3, pp. 409–418.

Nelson, M R et al. (July 2012). “An Abundance of Rare Functional Variants in 202 Drug Target Genes Sequenced in 14,002 People”. *Science* 337.6090, pp. 100–104.

Newstead, Chas G (Aug. 2003). “Recurrent disease in renal transplants.” *Nephrol Dial Transplant* 18 Suppl 6, pp. vi68–74.

Ng, S. B., E. H. Turner, P. D. Robertson, S. D. Flygare, A. W. Bigham, C. Lee, T. Shaffer, M. Wong, A. Bhattacharjee, E. E. Eichler, M. Bamshad, D. A. Nickerson, and J. Shendure (Sept. 2009). “Targeted capture and massively parallel sequencing of 12 human exomes”. *Nature* 461.7261, pp. 272–276.

Ng, Sarah B et al. (Aug. 2010a). “Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome”. *Nat Genet* 42.9, pp. 790–793.

Ng, Sarah B, Kati J Buckingham, Choli Lee, Abigail W Bigham, Holly K Tabor, Karin M Dent, Chad D Huff, Paul T Shannon, Ethylin Wang Jabs, Deborah A Nickerson, Jay Shendure, and Michael J Bamshad (Jan. 2010b). “Exome sequencing identifies the cause of a mendelian disorder”. *Nature Genetics* 42.1, pp. 30–35.

Ng, Sarah B, Deborah A Nickerson, Michael J Bamshad, and Jay Shendure (Oct. 2010c). “Massively parallel sequencing and rare disease.” *Hum Mol Genet* 19.R2, R119–24.

Niederer, Heather A et al. (Aug. 2010). “Copy number, linkage disequilibrium and disease association in the FCGR locus.” *Hum Mol Genet* 19.16, pp. 3282–3294.

Nikolov, Nikolay P, Masaki Shimizu, Sophia Cleland, Daniel Bailey, Joseph Aoki, Ted Strom, Pamela L Schwartzberg, Fabio Candotti, and Richard M Siegel (Aug.

2010). “Systemic autoimmunity and defective Fas ligand secretion in the absence of the Wiskott-Aldrich syndrome protein.” *Blood* 116.5, pp. 740–747.

Nishizumi, H, I Taniuchi, Y Yamanashi, D Kitamura, D Ilic, S Mori, T Watanabe, and T Yamamoto (Nov. 1995). “Impaired proliferation of peripheral B cells and indication of autoimmune disease in lyn-deficient mice.” *Immunity* 3.5, pp. 549–560.

Noakes, P G, J. H. Miner, M Gautam, J M Cunningham, J R Sanes, and J P Merlie (Aug. 1995). “The renal glomerulus of mice lacking s-laminin/laminin beta 2: nephrosis despite molecular compensation by laminin beta 1.” *Nat Genet* 10.4, pp. 400–406.

Nolan, P M et al. (Aug. 2000). “A systematic, genome-wide, phenotype-driven mutagenesis programme for gene function studies in the mouse.” *Nat Genet* 25.4, pp. 440–443.

Nolan, Patrick M, Alison Hugill, and Roger D Cox (Oct. 2002). “ENU mutagenesis in the mouse: application to human genetic disease.” *Briefings in functional genomics & proteomics* 1.3, pp. 278–289.

Norris, J R (July 1998). *Markov Chains*. Cambridge University Press. ISBN: 9780521633963.

Ochs, Hans D and Adrian J Thrasher (Apr. 2006). “The Wiskott-Aldrich syndrome.” *The Journal of allergy and clinical immunology* 117.4, 725–38–quiz 739.

Oliver, Peter L and Kay E Davies (Oct. 2012). “New insights into behaviour using mouse ENU mutagenesis”. *Human Molecular Genetics* 21.R1, R72–R81.

O’Roak, Brian J et al. (Apr. 2012). “Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations”. *Nature* 485.7397, pp. 246–250.

Ozaltin, Fatih, Tulin Ibsirlioglu, Ekim Z Taskiran, Dilek Ertoy Baydar, Figen Kaymaz, Mithat Buyukcelik, Beltinge Demircioglu Kilic, Ayse Balat, Paraskevas Iatropoulos, Esin Asan, Nurten A Akarsu, Franz Schaefer, Engin Yilmaz, and Aysin

Bakkaloglu (July 2011). “Disruption of PTPRO Causes Childhood-Onset Nephrotic Syndrome”. *The American Journal of Human Genetics* 89.1, pp. 139–147.

Pagnamenta, Alistair T, Stefano Lise, Victoria Harrison, Helen Stewart, Sandeep Jayawant, Gerardine Quaghebeur, Alexander T Deng, Valerie Elizabeth Murphy, Elham Sadighi Akha, Andy Rimmer, Iain Mathieson, Samantha JL Knight, Usha Kini, Jenny C Taylor, and David A Keays (Dec. 2011). “Exome sequencing can detect pathogenic mosaic mutations present at low allele frequencies”. *Journal of Human Genetics* 57.1, pp. 70–72.

Palles, Claire et al. (Dec. 2012). “Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas”. *Nat Genet*, pp. –.

Pan, H F, R X Leng, J H Tao, X P Li, and D Q Ye (Mar. 2011). “Ets-1: a new player in the pathogenesis of systemic lupus erythematosus?” *Lupus* 20.3, pp. 227–230.

Panoutsopoulou, Kalliope, Ioanna Tachmazidou, and Eleftheria Zeggini (Oct. 2013). “In search of low-frequency and rare variants affecting complex traits”. *Human Molecular Genetics* 22.R1, R16–R21.

Papathanasiou, P. and C. C. Goodnow (2005). “Connecting mammalian genome with phenome by ENU mouse mutagenesis: gene combinations specifying the immune system”. *Annu Rev Genet* 39, pp. 241–262.

Pascal Schneider, Fabienne MacKay Véronique Steiner Kay Hofmann Jean-Luc Bodmer Nils Holler Christine Ambrose Pornsri Lawton Sarah Bixler Hans Acha-Orbea Danila Valmori Pedro Romero Christiane Werner-Favre Rudolph H Zubler Jeffrey L Browning Jürg Tschopp (June 1999). “BAFF, a Novel Ligand of the Tumor Necrosis Factor Family, Stimulates B Cell Growth ”. *J Exp Med* 189.11, pp. 1747–1756.

Pelletier, Jerry, Wendy Bruening, Clifford E Kashtan, S Michael Mauer, J Carlos Manivel, Jane E Striegel, Donald C Houghton, Claudine Junien, Renée Habib,

Laurie Fouser, Richard N Fine, Bernard L Silverman, Daniel A Haber, and David Housman (Apr. 2014). "Germline mutations in the Wilms' tumor suppressor gene are associated with abnormal urogenital development in Denys-Drash syndrome". *Cell* 67.2, pp. 437–447.

Perchonock, C E, M C Fernando, W J Quinn, C T Nguyen, J Sun, M J Shapiro, and V S Shapiro (July 2006). "Negative Regulation of Interleukin-2 and p38 Mitogen-Activated Protein Kinase during T-Cell Activation by the Adaptor ALX". *Molecular and Cellular Biology* 26.16, pp. 6005–6015.

Peterson, J C, S Adler, J M Burkart, T Greene, L A Hebert, L G Hunsicker, A J King, S Klahr, S G Massry, and J L Seifter (Nov. 1995). "Blood pressure control, proteinuria, and the progression of renal disease. The Modification of Diet in Renal Disease Study." *Annals of internal medicine* 123.10, pp. 754–762.

Pfeifhofer, Christa, Thomas Gruber, Thomas Letschka, Nikolaus Thuille, Christina Lutz-Nicoladoni, Natascha Hermann-Kleiter, Uschi Braun, Michael Leitges, and Gottfried Baier (May 2006). "Defective IgG2a/2b class switching in PKC alpha-/- mice." *J Immunol* 176.10, pp. 6004–6011.

Pickering, M C and M J Walport (Jan. 2000). "Links between complement abnormalities and systemic lupus erythematosus". *Rheumatology* 39.2, pp. 133–141.

Pisitkun, Prapaporn, Jonathan A Deane, Michael J Difilippantonio, Tatyana Tarasenko, Anne B Satterthwaite, and Silvia Bolland (June 2006). "Autoreactive B cell responses to RNA-related antigens due to TLR7 gene duplication." *Science* 312.5780, pp. 1669–1672.

Pollard, K S, M J Hubisz, K R Rosenbloom, and A Siepel (Jan. 2010). "Detection of nonneutral substitution rates on mammalian phylogenies". *Genome Research* 20.1, pp. 110–121.

Pons-Estel, Guillermo J, Graciela S Alarcon, Lacie Scofield, Leslie Reinlib, and Glinda S Cooper (Feb. 2010). "Understanding the Epidemiology and Progression of Sys-

temic Lupus Erythematosus". *Seminars in Arthritis and Rheumatism* 39.4, pp. 257–268.

Probst, Frank J and Monica J Justice (2010). "Chapter Fifteen - Mouse Mutagenesis with the Chemical Supermutagen ENU". *Guide to Techniques in Mouse Development, Part B: Mouse Molecular Genetics, 2nd Edition*. Ed. by Paul M Wassarman Soriano and Philippe M. Academic Press, pp. 297–312.

Purcell, Shaun M et al. (Feb. 2014). "A polygenic burden of rare disruptive mutations in schizophrenia". *Nature* 506.7487, pp. 185–190.

Quan, F., J. Janas, S. Toth-Fejel, D. B. Johnson, J. K. Wolford, and B. W. Popovich (Jan. 1997). "Uniparental disomy of the entire X chromosome in a female with Duchenne muscular dystrophy". *Am J Hum Genet* 60.1, pp. 160–165.

Quinlan, A. R. and I. M. Hall (Mar. 2010). "BEDTools: a flexible suite of utilities for comparing genomic features". *Bioinformatics* 26.6, pp. 841–842.

Quwailid, M. M. et al. (Aug. 2004). "A gene-driven ENU-based approach to generating an allelic series in any gene". *Mamm Genome* 15.8, pp. 585–591.

Rabiner, L. R. (Feb. 1989). "A Tutorial on Hidden Markov-Models and Selected Applications in Speech Recognition". *Proceedings of the Ieee* 77.2, pp. 257–286.

Ramanan, Vijay K, Li Shen, Jason H Moore, and Andrew J Saykin (July 2012). "Pathway analysis of genomic data: concepts, methods, and prospects for future development". *Trends in Genetics* 28.7, pp. 323–332.

Ravenscroft, Jane C, Mohnish Suri, Gillian I Rice, Marcin Szynkiewicz, and Yanick J Crow (Jan. 2011). "Autosomal dominant inheritance of a heterozygous mutation in SAMHD1 causing familial chilblain lupus." *American journal of medical genetics. Part A* 155A.1, pp. 235–237.

Ravitsky, Vardit and Benjamin S Wilfond (Dec. 2006). "Disclosing Individual Genetic Results to Research Participants". *The American Journal of Bioethics* 6.6, pp. 8–17.

Razmara, Marjaneh, Srinivasa M Srinivasula, Lin Wang, Jean-Luc Poyet, Brad J Geddes, Peter S DiStefano, John Bertin, and Emad S Alnemri (Apr. 2002). "CARD-8 protein, a new CARD family member that regulates caspase-1 activation and apoptosis." *J Biol Chem* 277.16, pp. 13952–13958.

Reese, M G, F H Eeckman, D. Kulp, and D. Haussler (1997). "Improved splice site detection in Genie." *Journal of computational biology : a journal of computational molecular cell biology* 4.3, pp. 311–323.

Rice, Gillian et al. (Apr. 2007). "Heterozygous mutations in TREX1 cause familial chilblain lupus and dominant Aicardi-Goutieres syndrome." *Am J Hum Genet* 80.4, pp. 811–815.

Richer, E, C Prendergast, D E Zhang, S T Qureshi, S M Vidal, and D Malo (Sept. 2010). "N-Ethyl-N-Nitrosourea-Induced Mutation in Ubiquitin-Specific Peptidase 18 Causes Hyperactivation of IFN- Signaling and Suppresses STAT4-Induced IFN- Production, Resulting in Increased Susceptibility to Salmonella Typhimurium". *The Journal of Immunology* 185.6, pp. 3593–3601.

Rimmer, Andy, Hang Phan, Iain Mathieson, Zamin Iqbal, Stephen R F Twigg, WGS500 Consortium, Andrew O M Wilkie, Gil McVean, and Gerton Lunter and (July 2014). "Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications".

Roach, J C, G. Glusman, A F A Smit, C D Huff, R Hubley, P T Shannon, L Rowen, K P Pant, N Goodman, M. Bamshad, J. Shendure, R Drmanac, L B Jorde, L Hood, and D J Galas (Apr. 2010). "Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing". *Science* 328.5978, pp. 636–639.

Rödelsperger, Christian, Peter Krawitz, Sebastian Bauer, Jochen Hecht, Abigail W Bigham, Michael Bamshad, Birgit Jonske de Condor, Michal R Schweiger, and Peter N Robinson (Mar. 2011). "Identity-by-descent filtering of exome sequence

data for disease-gene identification in autosomal recessive disorders." *Bioinformatics* 27.6, pp. 829–836.

Rood, Ilse M, Jeroen K J Deegens, and Jack F M Wetzels (Mar. 2012). "Genetic causes of focal segmental glomerulosclerosis: implications for clinical practice." *Nephrol Dial Transplant* 27.3, pp. 882–890.

Russell, W. L. (1989). "Reminiscences of a mouse specific-locus test addict." *Environmental and Molecular Mutagenesis* 14 Suppl 16, pp. 16–22.

Russell, W. L., P. R. Hunsicker, G D Raymer, M H Steele, K F Stelzner, and H M Thompson (June 1982a). "Dose–response curve for ethylnitrosourea-induced specific-locus mutations in mouse spermatogonia." *Proceedings of the National Academy of Sciences of the United States of America* 79.11, pp. 3589–3591.

Russell, W. L., P. R. Hunsicker, D A Carpenter, C V Cornett, and G M Guinn (June 1982b). "Effect of dose fractionation on the ethylnitrosourea induction of specific-locus mutations in mouse spermatogonia." *Proceedings of the National Academy of Sciences of the United States of America* 79.11, pp. 3592–3593.

Russell, W. L., E. M. Kelly, P. R. Hunsicker, J. W. Bangham, S. C. Maddux, and E. L. Phipps (Nov. 1979). "Specific-locus test shows ethylnitrosourea to be the most potent mutagen in the mouse". *Proc Natl Acad Sci U S A* 76.11, pp. 5818–5819.

Saito, Takashi, Yukio Matsuba, Naomi Mihira, Jiro Takano, Per Nilsson, Shigeyoshi Itohara, Nobuhisa Iwata, and Takaomi C Saido (Apr. 2014). "Single App knock-in mouse models of Alzheimer's disease". *Nature Neuroscience* 17.5, pp. 661–663.

Sakuraba, Y. et al. (Oct. 2005). "Molecular characterization of ENU mouse mutagenesis and archives". *Biochem Biophys Res Commun* 336.2, pp. 609–616.

Saleem, Moin A (July 2012). "New developments in steroid-resistant nephrotic syndrome." *Pediatr Nephrol*.

Saleem, Moin A, Michael J O'Hare, Jochen Reiser, Richard J Coward, Carol D Inward, Timothy Farren, Chang Ying Xing, Lan Ni, Peter W Mathieson, and Peter

Mundel (Mar. 2002). "A Conditionally Immortalized Human Podocyte Cell Line Demonstrating Nephrin and Podocin Expression". *Journal of the American Society of Nephrology* 13.3, pp. 630–638.

Sanders, Stephan J et al. (Apr. 2012a). "De novo mutations revealed by whole-exome sequencing are strongly associated with autism". *Nature* 485.7397, pp. 237–241.

— (Apr. 2012b). "De novo mutations revealed by whole-exome sequencing are strongly associated with autism". *Nature* 485.7397, pp. 237–241.

Santín, Sheila, Gemma Bullich, Bárbara Tazón-Vega, Rafael García-Maset, Isabel Giménez, Irene Silva, Patricia Ruíz, José Ballarín, Roser Torra, and Elisabet Ars (May 2011). "Clinical utility of genetic testing in children and adults with steroid-resistant nephrotic syndrome." *Clinical journal of the American Society of Nephrology : CJASN* 6.5, pp. 1139–1148.

Sanui, Terukazu, Ayumi Inayoshi, Mayuko Noda, Eiko Iwata, Masahiro Oike, Takehiko Sasazuki, and Yoshinori Fukui (July 2003). "DOCK2 Is Essential for Antigen-Induced Translocation of TCR and Lipid Rafts, but Not PKC-θ and LFA-1, in T Cells". *Immunity* 19.1, pp. 119–129.

Scheuner, Maren T, Han de Vries, Benjamin Kim, Robin C Meili, Sarah H Olmstead, and Stephanie Teleki (July 2009). "Are electronic health records ready for genomic medicine?" *Genetics in medicine : official journal of the American College of Medical Genetics* 11.7, pp. 510–517.

Schiemann, B (Aug. 2001). "An Essential Role for BAFF in the Normal Development of B Cells Through a BCMA-Independent Pathway". *Science* 293.5537, pp. 2111–2114.

Schödel, Johannes, Chiara Bardella, Lina K Sciesielski, Jill M Brown, Chris W Pugh, Veronica Buckle, Ian P Tomlinson, Peter J Ratcliffe, and David R Mole (Apr. 2012). "Common genetic variants at the 11q13.3 renal cancer susceptibility locus

influence binding of HIF to an enhancer of cyclin D1 expression." *Nat Genet* 44.4, 420–5–S1–2.

Schork, Nicholas J, Sarah S Murray, Kelly A Frazer, and Eric J Topol (June 2009). "Common vs. rare allele hypotheses for complex diseases." *Current opinion in genetics & development* 19.3, pp. 212–219.

Schwarz, Jana Marie, Christian Rödelsperger, Markus Schuelke, and Dominik Seelow (Aug. 2010). "MutationTaster evaluates disease-causing potential of sequence alterations". *Nat Methods* 7.8, pp. 575–576.

Sedgwick, Steven G and Stephen J Smerdon (Aug. 1999). "The ankyrin repeat: a diversity of interactions on a common structural framework". *Trends in Biochemical Sciences* 24.8, pp. 311–316.

Seelow, D, M Schuelke, F. Hildebrandt, and P Nürnberg (June 2009). "HomozygosityMapper–an interactive approach to homozygosity mapping". *Nucleic Acids Res* 37.Web Server, W593–W599.

Segovia, Donato Alarc n, Marta E Alarc n Riquelme, Mario H Cardiel, Francisco Caeiro, Loreto Massardo, Antonio R Villa, Bernardo A Pons-Estel, and on behalf of the Grupo Latinoamericano de Estudio del Lupus Eritematoso (GLADEL) (2005). "Familial aggregation of systemic lupus erythematosus, rheumatoid arthritis, and other autoimmune diseases in 1,177 lupus patients from the GLADEL cohort". *Arthritis Rheum* 52.4, pp. 1138–1147.

Senaldi, G., S. Yin, C. L. Shaklee, P. F. Piguet, T. W. Mak, and T. R. Ulich (Dec. 1996). "Corynebacterium parvum- and Mycobacterium bovis bacillus Calmette-Guerin-induced granuloma formation is inhibited in TNF receptor I (TNF-RI) knockout mice and by treatment with soluble TNF-RI". *J Immunol* 157.11, pp. 5022–5026.

Seshan, Surya V and J Charles Jennette (Jan. 2013). "Renal Disease in Systemic Lupus Erythematosus With Emphasis on Classification of Lupus Glomerulonephritis:

Advances and Implications". *Archives of Pathology & Laboratory Medicine* 133.2, pp. 233–248.

Shaheen, Ranad, Eissa Faqeih, Asma Sunker, Heba Morsy, Tarfa Al-Sheddi, Hanan E Shamseldin, Nouran Adly, Mais Hashem, and Fowzan S Alkuraya (Aug. 2011). "Recessive Mutations in DOCK6, Encoding the Guanidine Nucleotide Exchange Factor DOCK6, Lead to Abnormal Actin Cytoskeleton Organization and Adams-Oliver Syndrome". *The American Journal of Human Genetics* 89.2, pp. 328–333.

Sheridan, Rachel, Kristin Lampe, Shiva Kumar Shanmukhappa, Patrick Putnam, Mehdi Keddache, Senad Divanovic, Jorge Bezerra, and Kasper Hoebe (July 2011). "Lampe1: An ENU-Germline Mutation Causing Spontaneous Hepatosteatosis Identified through Targeted Exon-Enrichment and Next-Generation Sequencing". *PloS one* 6.7, e21979.

Shi, Shan-Rong, Yan Shi, and Clive R Taylor (Jan. 2011). "Antigen Retrieval Immunohistochemistry: Review and Future Prospects in Research and Diagnosis over Two Decades". *Journal of Histochemistry & Cytochemistry* 59.1, pp. 13–32.

Shizhong Han, Xana Kim-Howard Harshal Deshmukh Yoichiro Kamatani Parvathi Viswanathan Joel M Guthridge Kenaz Thomas Kenneth M Kaufman Joshua Ojwang Adriana Rojas-Villarraga Vicente Baca Lorena Orozco Benjamin Rhodes Chan-Bum Choi Peter K Gregersen Joan T Merrill Judith A James Patrick M Gaffney Kathy L Moser Chaim O Jacob Robert P Kimberly John B Harley Sang-Choel Bae Juan-Manuel Anaya Marta E Alarcón-Riquelme Koichi Matsuda Timothy J Vyse Swapan K Nath (Mar. 2009). "Evaluation of imputation-based association in and around the integrin-α-M (ITGAM) gene and replication of robust association between a non-synonymous functional variant within ITGAM and systemic lupus erythematosus (SLE)". *Human Molecular Genetics* 18.6, p. 1171.

Simon, Michelle M et al. (2013). "A comparative phenotypic and genomic analysis of C57BL/6J and C57BL/6N mouse strains". *Genome Biol* 14.7, R82.

Sims, David, Ian Sudbery, Nicholas E Ilott, Andreas Heger, and Chris P Ponting (Jan. 2014). "Sequencing depth and coverage: key considerations in genomic analyses". *Nat Rev Genet* 15.2, pp. 121–132.

Skarnes, William C et al. (June 2011). "A conditional knockout resource for the genome-wide study of mouse gene function". *Nature* 474.7351, pp. 337–342.

Sluyter, Frans, Charlotte C M Marican, and Wim E Crusio (Dec. 1998). "Further phenotypical characterisation of two substrains of C57BL/6J inbred mice differing by a spontaneous single-gene mutation". *Behavioural Brain Research* 98.1, pp. 39–43.

Smagulova, Fatima, Ivan V Gregoretti, Kevin Brick, Pavel Khil, R Daniel Camerini-Otero, and Galina V Petukhova (Apr. 2011). "Genome-wide analysis reveals novel molecular features of mouse recombination hotspots". *Nature* 472.7343, pp. 375–378.

Smith, Katherine R, Catherine J Bromhead, Michael S Hildebrand, A Eliot Shearer, Paul J Lockhart, Hossein Najmabadi, Richard J Leventer, George McGillivray, David J Amor, Richard J Smith, and Melanie Bahlo (2011). "Reducing the exome search space for mendelian diseases using genetic linkage analysis of exome genotypes." *Genome Biol* 12.9, R85.

So, Hon-Cheong, Allen H S Gui, Stacey S Cherny, and Pak C Sham (Mar. 2011). "Evaluating the heritability explained by known susceptibility variants: a survey of ten complex diseases". *Genetic epidemiology* 35.5, pp. 310–317.

Sobreira, Nara L M et al. (June 2010). "Whole-genome sequencing of a single proband together with linkage analysis identifies a Mendelian disease gene." *PLoS Genet* 6.6, e1000991.

Sosnay, Patrick R et al. (Aug. 2013). "Defining the disease liability of variants in the cystic fibrosis transmembrane conductance regulator gene". *Nature Genetics* 45.10, pp. 1160–1167.

Stanescu, Horia C et al. (Feb. 2011). "Risk HLA-DQA1 and PLA 2R1 Alleles in Idiopathic Membranous Nephropathy". *N Engl J Med* 364.7, pp. 616–626.

Stark, K, S Vainio, G Vassileva, and A P McMahon (Dec. 1994). "Epithelial transformation of metanephric mesenchyme in the developing kidney regulated by Wnt-4." *Nature* 372.6507, pp. 679–683.

Stensman, H. and C. Larsson (Sept. 2007). "Identification of acidic amino acid residues in the protein kinase C alpha V5 domain that contribute to its insensitivity to diacylglycerol". *J Biol Chem* 282.39, pp. 28627–28638.

Stensman, Helena, Arathi Raghunath, and Christer Larsson (Sept. 2004). "Autophosphorylation suppresses whereas kinase inhibition augments the translocation of protein kinase Calpha in response to diacylglycerol." *J Biol Chem* 279.39, pp. 40576–40583.

Stephanie Humblet-Baron, Blythe Sather Stephanie Anover Shirly Becker-Herman Debora J Kasprowicz Socheath Khim Thuc Nguyen Kelly Hudkins-Loya Charles E Alpers Steve F Ziegler Hans Ochs Troy Torgerson Daniel J Campbell David J Rawlings (Feb. 2007). "Wiskott-Aldrich syndrome protein is required for regulatory T cell homeostasis". *Journal of Clinical Investigation* 117.2, p. 407.

Stevens, P E, D J O'Donoghue, S de Lusignan, J Van Vlymen, B Klebe, R Middleton, N Hague, J New, and C K T Farmer (July 2007). "Chronic kidney disease management in the United Kingdom: NEOERICA project results." *Kidney international* 72.1, pp. 92–99.

Su, Helen C, Huie Jing, and Qian Zhang (Dec. 2011). "DOCK8 deficiency." *Ann N Y Acad Sci* 1246, pp. 26–33.

Subramanian, Srividya, Katalin Tus, Quan-Zhen Li, Andrew Wang, Xiang-Hong Tian, Jinchun Zhou, Chaoying Liang, Guy Bartov, Lisa D McDaniel, Xin J Zhou, Roger A Schultz, and Edward K Wakeland (June 2006). "A Tlr7 translocation accelerates

systemic autoimmunity in murine lupus." *Proceedings of the National Academy of Sciences of the United States of America* 103.26, pp. 9970–9975.

Subramanian S Ajay, Stephen C J Parker Hatice Ozel Abaan Karin V Fuentes Fajardo Elliott H Margulies (Sept. 2011). "Accurate and comprehensive sequencing of personal genomes". *Genome Research* 21.9, pp. 1498–1505.

Sullivan, Kathleen E, Craig A Mullen, R Michael Blaese, and Jerry A Winkelstein (Dec. 1994). "A multiinstitutional survey of the Wiskott-Aldrich syndrome". *The Journal of Pediatrics* 125.6, pp. 876–885.

Sun, James X, Agnar Helgason, Gisli Masson, Sigríur Sunna Ebenesersdóttir, Heng Li, Swapan Mallick, Sante Gnerre, Nick Patterson, Augustine Kong, David Reich, and Kari Stefansson (Aug. 2012a). "A direct characterization of human mutation based on microsatellites". *Nat Genet* 44.10, pp. 1161–1165.

Sun, Miao, Kajari Mondal, Viren Patel, Vanessa L Horner, Alyssa B Long, David J Cutler, Tamara Caspary, and Michael E Zwick (Jan. 2012b). "Multiplex Chromosomal Exome Sequencing Accelerates Identification of ENU-Induced Mutations in the Mouse." *G3 (Bethesda, Md.)* 2.1, pp. 143–150.

Sunyaev, Shamil R (Jan. 2012). "Inferring causality and functional significance of human coding DNA variants". *Human Molecular Genetics* 21.R1, R10–R17.

Supp, D M, D P Witte, S S Potter, and M Brueckner (Oct. 1997). "Mutation of an axonemal dynein affects left-right asymmetry in inversus viscerum mice." *Nature* 389.6654, pp. 963–966.

Takahasi, K. R., Y. Sakuraba, and Y. Gondo (2007). "Mutational pattern and frequency of induced nucleotide changes in mouse ENU mutagenesis". *BMC Mol Biol* 8, p. 52.

Tampe, Björn and Michael Zeisberg (Feb. 2014). "Evidence for the involvement of epigenetics in the progression of renal fibrogenesis". *Nephrology Dialysis Transplantation* 29.suppl 1, pp. i1–i8.

Tan, E M, A S Cohen, J F Fries, A T Masi, D J McShane, N F Rothfield, J G Schaller, N Talal, and R J Winchester (Nov. 1982). "The 1982 revised criteria for the classification of systemic lupus erythematosus." *Arthritis Rheum* 25.11, pp. 1271–1277.

Tanaka, Yoshihiko, Shinjiro Hamano, Kazuhito Gotoh, Yuzo Murata, Yuya Kunisaki, Akihiko Nishikimi, Ryosuke Takii, Makiko Kawaguchi, Ayumi Inayoshi, Sadahiko Masuko, Kunisuke Himeno, Takehiko Sasazuki, and Yoshinori Fukui (Oct. 2007). "T helper type 2 differentiation and intracellular trafficking of the interleukin 4 receptor-alpha subunit controlled by the Rac activator Dock2." *Nat Immunol* 8.10, pp. 1067–1075.

Taylor, Kimberly E et al. (May 2008). "Specificity of the STAT4 Genetic Association for Severe Disease Manifestations of Systemic Lupus Erythematosus". *PLoS Genet* 4.5, e1000084.

Tchekneva, Elena E, Eugene M Rinchik, Dina Polosukhina, Linda S Davis, Veronika Kadkina, Yassir Mohamed, Steve R Dunn, Kumar Sharma, Zhonghua Qi, Agnes B Fogo, and Matthew D Breyer (Jan. 2007). "A Sensitized Screen of N-ethyl-N-nitrosourea–Mutagenized Mice Identifies Dominant Mutants Predisposed to Diabetic Nephropathy". *Journal of the American Society of Nephrology* 18.1, pp. 103–112.

Tennessen, J A et al. (July 2012). "Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes". *Science* 337.6090, pp. 64–69.

Teslovich, Tanya M et al. (Aug. 2010). "Biological, clinical and population relevance of 95 loci for blood lipids". *Nature* 466.7307, pp. 707–713.

Theofilopoulos, Argyrios N, Roberto Baccala, Bruce Beutler, and Dwight H Kono (Nov. 2004). "TYPE I INTERFERONS (α/β) IN IMMUNITY AND AUTOIMMUNITY". *Annual review of immunology* 23.1, pp. 307–335.

Thorvaldsdottir, H, J T Robinson, and J. P. Mesirov (Mar. 2013). "Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration". *Briefings in Bioinformatics* 14.2, pp. 178–192.

Threadgill, David W and Gary A Churchill (Feb. 2012). "Ten Years of the Collaborative Cross". *Genetics* 190.2, pp. 291–294.

Threadgill, David W, Delia Yee, Argabin Matin, Joseph H Nadeau, and Terry Magnuson (June 1997). "Genealogy of the 129 inbred strains: 129/SvJ is a contaminated inbred strain". *Mammalian Genome* 8.6, pp. 390–393.

Thusberg, Janita, Ayodeji Olatubosun, and Mauno Vihinen (Apr. 2011). "Performance of mutation pathogenicity prediction methods on missense variants." *Human mutation* 32.4, pp. 358–368.

Tiffin, Nicki, Adebowale Adeyemo, and Ikechi Okpechi (2013). "A diverse array of genetic factors contribute to the pathogenesis of Systemic Lupus Erythematosus". *Orphanet Journal of Rare Diseases* 8.1, p. 2.

Trevor J Pemberton, Devin Absher Marcus W Feldman Richard M Myers Noah A Rosenberg Jun Z Li (Aug. 2012). "Genomic Patterns of Homozygosity in Worldwide Human Populations". *American Journal of Human Genetics* 91.2, p. 275.

Uhlen, Mathias, Per Oksvold, Linn Fagerberg, Emma Lundberg, Kalle Jonasson, Mattias Forsberg, Martin Zwahlen, Caroline Kampf, Kenneth Wester, Sophia Hober, Henrik Wernerus, Lisa Björling, and Fredrik Ponten (Dec. 2010). "Towards a knowledge-based Human Protein Atlas". *Nat Biotechnol* 28.12, pp. 1248–1250.

Ulrich Brinkmann, George Vasmatzis Byungkook Lee Noga Yerushalmi Magnus Essand Ira Pastan (Sept. 1998). "PAGE-1, an X chromosome-linked GAGE-like gene that is expressed in normal and neoplastic prostate, testis, and uterus". *Proceedings of the National Academy of Sciences of the United States of America* 95.18, p. 10757.

Untergasser, Andreas, Harm Nijveen, Xiangyu Rao, Ton Bisseling, René Geurts, and Jack A M Leunissen (July 2007). "Primer3Plus, an enhanced web interface to Primer3". *Nucleic Acids Research* 35.suppl 2, W71–W74.

Vang, Torkel, Mauro Congia, Maria Doloretta Macis, Lucia Musumeci, Valeria Orrú, Patrizia Zavattari, Konstantina Nika, Lutz Tautz, Kjetil Taskén, Francesco Cucca, Tomas Mustelin, and Nunzio Bottini (Dec. 2005). "Autoimmune-associated lymphoid tyrosine phosphatase is a gain-of-function variant." *Nat Genet* 37.12, pp. 1317–1319.

Vanharanta, Sakari, Weiping Shu, Fabienne Brenet, A Ari Hakimi, Adriana Heguy, Agnes Viale, Victor E Reuter, James J-D Hsieh, Joseph M Scandura, and Joan Massagué (Jan. 2013). "Epigenetic expansion of VHL-HIF signal output drives multiorgan metastasis in renal cancer". *Nature Medicine* 19.1, pp. 50–56.

Vecchione, Andrea, Gustavo Baldassarre, Hideshi Ishii, Milena S Nicoloso, Barbara Belletti, Fabio Petrocca, Nicola Zanesi, Louise Y Y Fong, Sabrina Battista, Daniela Guarnieri, Raffaele Baffa, Hansjuerg Alder, John L Farber, Peter J Donovan, and Carlo M Croce (Mar. 2007). "Fez1/Lzts1 absence impairs Cdk1/Cdc25C interaction during mitosis and predisposes mice to cancer development." *Cancer cell* 11.3, pp. 275–289.

Verhagen, A. M., M. E. Wallace, A. Goradia, S. A. Jones, H. A. Croom, D. Metcalf, J. E. Collinge, M. J. Maxwell, M. L. Hibbs, W. S. Alexander, D. J. Hilton, B. T. Kile, and R. Starr (Feb. 2009). "A kinase-dead allele of Lyn attenuates autoimmune disease normally associated with Lyn deficiency". *J Immunol* 182.4, pp. 2020–2029.

Vinuesa, C. G. and C. C. Goodnow (June 2004). "Illuminating autoimmune regulators through controlled variation of the mouse genome sequence". *Immunity* 20.6, pp. 669–679.

Vinuesa, C. G., M. C. Cook, C. Angelucci, V. Athanasopoulos, L. Rui, K. M. Hill, D. Yu, H. Domaschenz, B. Whittle, T. Lambe, I. S. Roberts, R. R. Copley, J. I. Bell, R. J. Cornall, and C. C. Goodnow (May 2005). "A RING-type ubiquitin ligase family member required to repress follicular helper T cells and autoimmunity". *Nature* 435.7041, pp. 452–458.

Visentini, Marcella, Valentina Conti, Maria Cagliuso, Francesca Tinti, Giulia Siciliano, Amelia C Trombetta, Anna Paola Mitterhofer, Massimo Fiorilli, and Isabella Quinti (Sept. 2009). "Regression of systemic lupus erythematosus after development of an acquired Toll-like receptor signaling defect and antibody deficiency". *Arthritis Rheum* 60.9, pp. 2767–2771.

Vissers, Lisenka E L M, Joep de Ligt, Christian Gilissen, Irene Janssen, Marloes Steehouwer, Petra de Vries, Bart van Lier, Peer Arts, Nienke Wieskamp, Marisol del Rosario, Bregje W M van Bon, Alexander Hoischen, Bert B A de Vries, Han G Brunner, and Joris A Veltman (Nov. 2010). "A de novo paradigm for mental retardation". *Nat Genet* 42.12, pp. 1109–1112.

Voss, A K, T Thomas, and P Gruss (June 1998). "Compensation for a gene trap mutation in the murine microtubule-associated protein 4 locus by alternative polyadenylation and alternative splicing." *Developmental dynamics : an official publication of the American Association of Anatomists* 212.2, pp. 258–266.

Wajant, H., K. Pfizenmaier, and P. Scheurich (Jan. 2003). "Tumor necrosis factor signaling". *Cell Death Differ* 10.1, pp. 45–65.

Waldek, Stephen and Sandro Feriozzi (2014). "Fabry nephropathy: a review – how can we optimize the management of Fabry nephropathy?" *BMC Nephrology* 15.1, p. 72.

Wallis, G A, B J Starman, A B Zinn, and P H Byers (June 1990). "Variable expression of osteogenesis imperfecta in a nuclear family is explained by somatic mosaicism

for a lethal point mutation in the alpha 1(I) gene (COL1A1) of type I collagen in a parent." *Am J Hum Genet* 46.6, pp. 1034–1040.

Walport, Mark J, Kevin A Davies, and Marina Botto (Aug. 1998). "C1q and Systemic Lupus Erythematosus". *Immunobiology* 199.2, pp. 265–285.

Wang, Chin-Man, Su-Wei Chang, Yeong-Jian Jan Wu, Jing-Chi Lin, Huei-Huang Ho, Tse-Chih Chou, Bing Yang, Jianming Wu, and Ji-Yih Chen (Jan. 2014). "Genetic variations in Toll-like receptors (TLRs 3/7/8) are associated with systemic lupus erythematosus in a Taiwanese population". *Scientific Reports* 4.

Wang, Haoyi, Hui Yang, Chikdu S Shivalila, Meelad M Dawlaty, Albert W Cheng, Feng Zhang, and Rudolf Jaenisch (May 2013). "One-Step Generation of Mice Carrying Mutations in Multiple Genes by CRISPR/Cas-Mediated Genome Engineering". *Cell* 153.4, pp. 910–918.

Wang, Jeremy R, Fernando Pardo-Manuel de Villena, Heather A Lawson, James M Cheverud, Gary A Churchill, and Leonard McMillan (Feb. 2012). "Imputation of Single-Nucleotide Polymorphisms in Inbred Mice Using Local Phylogeny". *Genetics* 190.2, pp. 449–458.

Wang, K., M. Li, and H. Hakonarson (Sept. 2010). "ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data". *Nucleic Acids Res* 38.16, e164.

Wansleeben, Carolien, Léon van Gurp, Harma Feitsma, Carla Kroon, Ester Rieter, Marlies Verberne, Victor Guryev, Edwin Cuppen, and Frits Meijlink (2011). "An ENU-mutagenesis screen in the mouse: identification of novel developmental gene functions." *PloS one* 6.4, e19357.

Ward, Lucas D and Manolis Kellis (Nov. 2012). "Interpreting noncoding genetic variation in complex traits and human disease." *Nat Biotechnol* 30.11, pp. 1095–1106.

Waterston, R. H. et al. (Dec. 2002). "Initial sequencing and comparative analysis of the mouse genome". *Nature* 420.6915, pp. 520–562.

Watkins, Katherine H, Allan Stewart, and William Fairbrother (2009). "A Rapid High-throughput Method for Mapping Ribonucleoproteins (RNPs) on Human pre-mRNA". *Journal of Visualized Experiments* 34.

Weening, J J (Feb. 2004). "The Classification of Glomerulonephritis in Systemic Lupus Erythematosus Revisited". *Journal of the American Society of Nephrology* 15.2, pp. 241–250.

Wertz, Ingrid E, Karen M O'Rourke, Honglin Zhou, Michael Eby, L. Aravind, Somasekar Seshagiri, Ping Wu, Christian Wiesmann, Rohan Baker, David L Boone, Averil Ma, Eugene V Koonin, and Vishva M Dixit (Aug. 2004). "De-ubiquitination and ubiquitin ligase domains of A20 downregulate NF-kappaB signalling." *Nature* 430.7000, pp. 694–699.

Wharram, Bryan L, Meera Goyal, Patrick J Gillespie, Jocelyn E Wiggins, David B Kershaw, Lawrence B Holzman, Robert C Dysko, Thomas L Saunders, Linda C Samuelson, and Roger C Wiggins (Nov. 2000). "Altered podocyte structure in GLEPP1 (Ptpro)-deficient mice associated with hypertension and low glomerular filtration rate". *J Clin Invest* 106.10, pp. 1281–1290.

Wilfond, Benjamin and Lainie Friedman Ross (July 2009). "From Genetics to Genomics: Ethics, Policy, and Parental Decision-making". *Journal of Pediatric Psychology* 34.6, pp. 639–647.

Wiltshire, S, J. T. Bell, C J Groves, C Dina, A T Hattersley, T M Frayling, M Walker, G. A. Hitman, M Vaxillaire, M Farrall, P Froguel, and M I McCarthy (Nov. 2006). "Epistasis between type 2 diabetes susceptibility Loci on chromosomes 1q21-25 and 10q23-26 in northern Europeans." *Annals of human genetics* 70.Pt 6, pp. 726–737.

Winn, Michelle P, Peter J Conlon, Kelvin L Lynn, Merry Kay Farrington, Tony Creazzo, April F Hawkins, Nikki Daskalakis, Shu Ying Kwan, Seth Ebersviller, James L Burchette, Margaret A Pericak-Vance, David N Howell, Jeffery M Vance,

and Paul B Rosenberg (June 2005). "A mutation in the TRPC6 cation channel causes familial focal segmental glomerulosclerosis." *Science* 308.5729, pp. 1801–1804.

Wong, William (May 2007). "Idiopathic nephrotic syndrome in New Zealand children, demographic, clinical features, initial management and outcome after twelve-month follow-up: results of a three-year national surveillance study." *Journal of paediatrics and child health* 43.5, pp. 337–341.

Woods, C Geoffrey, James Cox, Kelly Springell, Daniel J Hampshire, Moin D Mohamed, Martin McKibbin, Rowena Stern, F Lucy Raymond, Richard Sandford, Saghira Malik Sharif, Gulshan Karbani, Mustaq Ahmed, Jacquelyn Bond, David Clayton, and Chris F Inglehearn (May 2006). "Quantification of homozygosity in consanguineous individuals with autosomal recessive disease." *Am J Hum Genet* 78.5, pp. 889–896.

Worthey, Elizabeth A et al. (Mar. 2011). "Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease." *Genetics in medicine : official journal of the American College of Medical Genetics* 13.3, pp. 255–262.

Wray, Naomi R, Shaun M Purcell, and Peter M Visscher (Jan. 2011). "Synthetic Associations Created by Rare Variants Do Not Explain Most GWAS Results". *PLoS Biology* 9.1, e1000579.

Wu, Fenfen, Wentao Mi, Dennis K Burns, Yu Fu, Hillery F Gray, Arie F Struyk, and Stephen C Cannon (Oct. 2011a). "A sodium channel knockin mutant (NaV1.4-R669H) mouse model of hypokalemic periodic paralysis." *J Clin Invest* 121.10, pp. 4082–4094.

Wu, Y. L., B. P. Brookshire, R. R. Verani, F. C. Arnett, and C. Y. Yu (Oct. 2011b). "Clinical presentations and molecular basis of complement C1r deficiency in a

male African-American patient with systemic lupus erythematosus". *Lupus* 20.11, pp. 1126–1134.

Xiu, Yan et al. (Oct. 2002). "Transcriptional regulation of Fcgr2b gene by polymorphic promoter region and its contribution to humoral immune responses." *J Immunol* 169.8, pp. 4340–4346.

Xu, Yuekang, Kenneth W Harder, Nicholas D Huntington, Margaret L Hibbs, and David M Tarlinton (Jan. 2005). "Lyn Tyrosine Kinase: Accentuating the Positive and the Negative". *Immunity* 22.1, pp. 9–18.

Yang, Yaping et al. (Oct. 2013). "Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders". *N Engl J Med* 369.16, pp. 1502–1511.

Yeong, S. S., Y. Zhu, D. Smith, C. Verma, W. G. Lim, B. J. Tan, Q. T. Li, N. S. Cheung, M. Cai, Y. Z. Zhu, S. F. Zhou, S. L. Tan, and W. Duan (Oct. 2006). "The last 10 amino acid residues beyond the hydrophobic motif are critical for the catalytic competence and function of protein kinase Calpha". *J Biol Chem* 281.41, pp. 30768–30781.

Zachary A Szpiech, Jishu Xu Trevor J Pemberton Weiping Peng Sebastian Zöllner Noah A Rosenberg Jun Z Li (July 2013). "Long Runs of Homozygosity Are Enriched for Deleterious Variation". *American Journal of Human Genetics* 93.1, p. 90.

Zeeuw, Dick de, Giuseppe Remuzzi, Hans-Henrik Parving, William F Keane, Zhongxin Zhang, Shahnaz Shahinfar, Steve Snapinn, Mark E Cooper, William E Mitch, and Barry M Brenner (June 2004). "Proteinuria, a target for renoprotection in patients with type 2 diabetic nephropathy: Lessons from RENAAL". *Kidney international* 65.6, pp. 2309–2320.

Zenker, M. et al. (Nov. 2004). "Human laminin beta2 deficiency causes congenital nephrosis with mesangial sclerosis and distinct eye abnormalities". *Hum Mol Genet* 13.21, pp. 2625–2632.

Zhou, Baiyu and Alice S Whittemore (June 2012). "Improving sequence-based genotype calls with linkage disequilibrium and pedigree information". *The Annals of Applied Statistics* 6.2, pp. 457–475.

Züchner, Stephan et al. (Feb. 2011). "Whole-exome sequencing links a variant in DHDDS to retinitis pigmentosa." *Am J Hum Genet* 88.2, pp. 201–206.

Appendix A

Candidate Variants in *ENU16CH17a*

Table A.1: The total shared homozygous or heterozygous regions contained 3619 mutations from the filtered call set, including the 2 non-synonymous homozygous mutations, and 36 heterozygote missense, nonsense or putative splice site mutations, of which 28 were present in all 3 mice

Chr	Pos	Gene	Ref	Sub	AA sub	Exon	Sanger	Polyphen (Humanvar)	genotypes1 2 3	type
2	120317562	Capn3	C	T	T341M	exon9	TP	0.996/ prob	0/1 0/1 0/1	nonsynon
2	148509797	Gzf1	G	A	A71T	exon3	TP	0.009/benign	0/1 0/1 0/1	nonsynon
2	148697743	Cst3	A	G	W126R	exon3	TP	0.564/ poss	0/1 0/1 0/1	nonsynon
2	164482725	Wfdc10	T	A	M85K	exon2	TP	0/ benign	1/0 1/0 1/0	nonsynon
3	53343180	Frem21	A	T	V76D	exon3	TP	0.621/poss	1/0 1/0 1/0	nonsynon
4	98104052	Inadl	A	C	E478D	exon12	TP	0.074/ benign	1/0 1/0 1/0	nonsynon
4	141151252	Fblim1	A	T	M45K	exon3	TP	0/benign	1/0 1/0 1/0	nonsynon
5	21545383	Reln	C	A	G805V	exon19	TP	0.995/ prob	1/0 1/1 1/1	nonsynon
5	31566037	ift172	C	A	M1006I	exon28	TP	0.984/prob	0/1 0/1 0/1	nonsynon
5	134692920	Gtf2ird2	T	C	F717L	exon16	TP	0.955/prob	1/0 1/0 1/0	nonsynon
6	36473946	Chrm2	A	G	N246S	exon1	TP	0.01/ benign	0/1 0/1 0/1	nonsynon
6	70093854	Igkv8-28	A	T	V13E	exon1	TP	0.465/poss	0/1 0/1 0/1	nonsynon
6	126924519	D6Wsu163e	A	G	M505V	exon13	TP	0.539/poss	0/1 0/1 0/1	nonsynon
9	21861790	Zfp653	T	C	M466V	exon6	TP	0.788/poss	1/0 1/0 1/0	nonsynon
9	25411078	Eepd1	G	A	C532Y	exon8	TP	0.991/prob	1/0 1/0 1/0	nonsynon
11	50021016	Sqstm1	A	G	C142R	exon3	TP	0.809/prob	1/1 0/1 0/1	nonsynon
11	82883777	Slfn2	T	C	V360A	exon2	TP	0.44/benign	0/1 0/1 0/1	nonsynon
11	95246346	Slc35b1	T	C	V8A	exon1	TP	0.003/benign	0/1 0/1 0/1	nonsynon
12	25742570	Kidins220	A	G	N1688S	exon34	TP	0.967/prob	1/0 1/0 1/0	nonsynon
12	31571934	Fam150b	A	G	M94V	exon2	TP	0.137/benign	0/1 0/1 0/1	nonsynon
12	71317593	Pygl	T	A	Y71F	exon4	TP	0.998/prob	1/0 1/0 1/0	nonsynon
12	77688461	Spnb1	A	T	F533L	exon16	TP	0.001/benign	0/1 0/1 0/1	nonsynon
13	84544914	Gm17618	A	G	Y21C	exon1	TP	unknown	1/11/0 1/0	nonsynon
13	114000186	Esm1	A	G	T46A	exon1	TP	0.587/poss	1/1 0/1 1/1	nonsynon
14	35542591	Wapal	A	G	T799A	exon9	TP	0.811/poss	0/1 0/1 0/1	nonsynon
16	49759470	Ift57	T	A	W169R	exon5	TP	0.473/benign	0/1 0/1 0/1	nonsynon
4	3710143	Lyn	A	G	T410A	exon12	TP	0.942/prob	1/1 1/1 1/1	nonsynon
4	66590107	Tlr4	A	G	T146A	exon4	TP	unknown	1/1 1/1 1/1	nonsynon

Appendix B

Variants in ENU16CH17a IBD regions with inconsistent genotypes

Table B.1: 10 of the protein sense changing varaints within the IBD were rejected as causative because they were absent in one or more affected animal, despite good individual depth of coverage (threshold >5 good quality reads at the locus for the inconsistently genotyped animal.

Chr	Pos	Gene	Ref	Sub	AA sub	Exon	genotypes1 2 3	type
2	85240307	Olfr992	T	C	N128D	exon1	0/0 0/1 0/0	nonsynonymous
2	155580104	Procr	G	A	A152T	exon3	1/0 1/0 0/0	nonsynonymous
5	104906089	Pkd2	T	A	V244E	exon3	0/0 0/0 1/0	nonsynonymous
5	105243850	Zfp951	A	T	C280S	exon4	1/0 1/0 0/0	nonsynonymous
5	117691186	Taok3	G	T	A426S	exon13	0/0 0/0 1/0	nonsynonymous
5	119192305	Med13l	A	G	D834G	exon17	0/0 0/0 0/1	nonsynonymous
5	121716325	Gm15800	A	T	T127S	exon7	1/0 1/0 0/0	nonsynonymous
7	52664993	Kcna7	A	G	M445V	exon2	0/0 0/1 0/0	nonsynonymous
9	40800682	Crtam	A	T	S101R	exon3	0/1 0/0 0/1	nonsynonymous
16	87366449	N6amt1	G	T	R154S	exon5	0/1 0/0 0/1	nonsynonymous
7	53163608	Tmem143	G	A7	686+1G>A	exon5	0/1 0/0 0/1	splicing

Appendix C

Primers for candidate ENU16CH17a variants

Table C.1: forward and reverse primers used to validate candidate ENU16CH17a variants

Gene	Internal Forward	Internal Reverse	External Forward	External Reverse
Capn3	TCCTCTCTCAACCTCTCAGGA	TCTGAAGCTTATCGGACTCCA	TTCCTCAGCCTGTGTGAGTG	TCGTTTACAGACACCGTCCA
Gzf1	GCTGGAGTCTGTCCTTTTGG	CCCATTGCCCAGTCTCTCTA	AGAAGACCGTGTGCAGCAG	TTGGATGGTGTTTCTGGTGA
Cst3	TCAGCCCTTAGGCATTTTTG	ATCATCCCGGAGTGGGTATT	AGAGCCGGAGCACCCTTC	GGGAGAAGGGTTTAGCTTCC
Wfdc10	CAAGCAAATAACATCTGCTGCT	TAGTCAGAGGGAGGGGTGAC	ATTGTGAAGCCCACCAAGAC	GGACCCACTCTCTGGGTAGA
Frem21	GGTAGCCAGACACAGGTCCA	TTGCTTCTAGGCAACCAAGG	AAAGAAATGTCTTATAATTTCCCCATA	GGATTTTCCAGTCATTTTTCTGA
Inadl	CCACCCCCAACAGGACTA	CTGGTTTCGGGGATCCTG	AGTGAAGTCAGCCCATTTGC	TCCTCTGCTTCCACACTCAA
Fblim1			CCAGAGAGTGAGGCATTGGT	CGAGATGTAGCCGTGAGTGA
Reln	TTACGGAATCACTGCTGCTG	GGGAGCAAGTCTGTGCTGA	GGCCCTCCATGCTACTGTAA	TCGTGTCTAGGTTTCTTCAGTTT
ift172	AGGCGAATCATGTCGTCATAC	AGGTGAGTCTCCGGGGAAG	GTGTGTCGCTGAGGAGGTCT	GCAAGTACCGAGAGGCTGAG
Gtf2ird2	TGGCAGCCTACTGTACCACA	CGTCGTCATGTCCACTAGGA	TGGTAGACTCGGTGAACTGG	GGGAGGCATCCAGTGTATTG
Chrm2	TGCAAGGAAGAATTGTAAAGC	CTGAACGCAGTTTTCAGTCC	ACCCGGTGTCTCCGAGTCTA	TCTTTCTCCTCCCCCTGAAC
Igkv8-28	TGGTACCAGGCCAAGTAGTTC	TGATGACACAGTCTCCATCCTC	GCCCCGTAGATCAACAGTTT	AGGTACCTGTGGGGACATTG
D6Wsu163e			TTTGGTATGGAAGGCCAGAG	GTTTGGACTGGGATCTGCTC
Zfp653	CAGCTGGACCCAACAACAG	ACAGGCCCCTTCTCTCCAG	CCACTGACACCTGGTAGAGC	TTATAGCCCACTGGGGACTG
Eepd1	GCTCACAAACCCTTGGATTC	GCTCCAGTCCTTTTCCATGT	CACTGGGCTGTTGTGAGAGA	TTCCTGGGGACTTCCTTCTT
Sqstm1	AAAGGGGTTGGGAAAGATGA	CACCCCAATGTGATCTGTGA	CCCGGCTCACATCAGAGA	TCCCTGCAGAGAAGAAGGAG
Slfn2	GCGCTCAGAGTGGAGAGATT	ACCCATTCCTCTGTGGTAAAC	TCATCGAAGTGCACAAATCC	GGGCATTCATCTGGAGTTTC
Slc35b1	CGGTACCGCTACCTGCTCT	CCTGCAGGATCCCATAGTAGA	GTCGCACGCAGCTCCTAT	CCTCAAGGGTCTGCTCACAT
Kidins220	CACCAACAGAGCCAATCAGA	GCTATCTGAGGAGGCTGCAT	CCCAGCACTGTGACTCTGAA	GGATTTTCCAGTCATTTTTCTGA
Fam150b	CGATTTCTCACCCCTCGTAG	GGACCTGGAACACAGAGGAA	GTTCTCCTGTTTTATCCAGTGC	AAGTGCTTGCTGCACTTCG
Pygl	AATGCCAGCGAAATCAGTG	GCTTCCATGCCTTTGAAGAG	AAGAGCTCACGAAGGCTCAG	TGCTAAGTTCGCTGGTTCTG
Spnb1	TAGCCCTCCATCTGCACAGT	ACCACAAAGACGAGCAGGAG	CACCTGTTAGAGGCCTTCTTG	AGGCACGCTCCTCTCTTGT
Gm17618	TTGCTTTCCTTCTACCTGATGA	TTGGAAAGGAGAGGCAGAGA	TTCCTCACTTTTTATTGCTACTTTT	GCACTTCTCTCAAAAATAGTTAGATG
Esm1	TCTTGCTGCTGACCACACTC	TCCATGCCTGAGACTGTACG	CAGCTGCGAGACATGAAGAG	CTCGTCACCAAAATCATCCTC
Wapal	TGGAACAAGATGCCTCTTCAG	CCAAACCCAACCCAAACTTA	CATGATTCGGCTTTTGGAGT	GCAAAGACTTAAAGGAAGTTTTTATC
Ift57	AAATCCCATGGTTTTCTAGGG	ACCTCACTTAGCTGGGCTTG	TTTGTGGATTGGGTTGAAGAG	AAATCCTGCAGAGCATTCTTAAAT
Lyn	AACTCTGAGTAGGAGCCCACA	GGGAATCTTCCCATAGGTGAC	TTTTGATTGTATCTTTTCTATTCCAA	TCACAATGGAGTGGGGTGTA
Tlr4	TGCACATGAATTGTGTCTTCA	GAGATGGAGCGGCAGTTAAG	TTTTATTTTATACAATGATGGTTTCCA	TGGCAAAGGACAGCAATTTT

Appendix D

Variants in a second pedigree sequenced at low coverage and analysed with the Lander–Green algorithm

Table D.1: Primers for second pedigree low coverage variants. sanger: True Positive (TP), False Positive (FP) Failed Sequencing (Unknown - assumed FP)

Chr	Pos	Ref	SNP	Left (Forward) Primer	Right (Reverse) Primer	Type of mutation	Sanger:
chr10	79354457	C	T	GTGCTGCACACTGCATGG	CGGGTGGGCACTACACTCT	Coding Homozygous	TP
chr8	96929639	A	T	AAATAAACTGGAGAAATCTGAGGCT	AGATGATTTTCCCTCTGGATTTTAC	Coding Homozygous	TP
chr8	81880266	A	G	AGAGTTGGGGTAGCTGCTGTT	TCTTATTTGTCTCTCTCGCTCTCTC	Coding Homozygous	TP
chr10	94306815	G	C	GCATGCATGGAAACTCACAT	GACTGTGGAACGCTGCAATA	Splicing Heterozygous	TP
chr10	77454832	G	C	CAGGAGCTACTGACCTCTTGGT	AAGTGTAGGGCTCTGTCCTCATAAT	Splicing Heterozygous	TP
chr10	62480787	T	A	GTTTTGAAACCTCATAGGTACTGGA	AATAGGAGGAAGAGGGGGTTATTAT	Splicing Heterozygous	TP
chr10	62532691	T	A	ATGAAGAAGATGGCTGAACTACTGT	CCCTTCTAAGATGAAAAACAAGGTT	Non-coding Homozygous	TP
chr8	79028363	A	T	TGAACAAGTAAGGGTTCTTTTTCCT	TATTGACCCCTGCAAGATATAGTGT	Non-coding Homozygous	TP
chr8	68170219	A	T	TGGTGACCAAACTGACAACA	CGATGTGTGTGACAAGCAAA	Non-coding Homozygous	TP
chr10	62533295	T	A	TTCCTACAACTTGTCCTCTGACTTC	CCTGGGTTGTACCCTTTTTATCTAT	Non-coding Homozygous	Unknown
chr10	74458031	A	T	CTCTCACCTCTCCCTCCAGAC	TTTGCAAAAGTGTGCTCTGTTC	Non-coding Homozygous	Unknown
chr8	73517737	G	C	ATTGGACACATTTATAATTCGAGGA	ACTAATATGTCTACTCTGGCGCAAC	Non-coding Homozygous	TP
chr8	94762440	A	C	TATGACTTTCATCAGTGCATTCCTA	CTCTGTGACATCACTGACAATCTTT	Non-coding Homozygous	TP

chr8	75810400	A	T	TTGCTTGAGCATATAAGAAAGACCT	CAGTGTTACAGAAAACCCTGTGGT	Non-coding Homozygous	TP
chr10	100282238	C	A	CAACAGACTGTGTTATTTGTTTATTTATGA	AGAAAATGTAATCATTGGTCCTCAA	Non-coding Homozygous	TP
chr10	75736529	A	T	AAGTAGATGTCTTTACCTTGTTGCG	GCCTATGTACTGAGATGAAACCTTG	Non-coding Homozygous	Unknown
chr8	78014812	A	T	TAACTTATCAGTCGTGGAACTCCTT	TGTCCTCTGGGATTACCACTATTTA	Non-coding Homozygous	TP
chr10	99465763	C	T	CCTGGACATTAAGAATTGTAGACCT	CAGGATCTTTCCATCTTCTATTTGA	Non-coding Homozygous	TP
chr11	53008786	A	G	CAGAGACCCACAGTCAAACATTAG	CCACAAGAAGACCAAGCTATAAAAC	Non-coding Homozygous	TP
chr10	74458034	A	T	TGGCTGTGCTGGTGACTG	GCAAAAGTGTGCTCTGTTCG	Non-coding Homozygous	TP
chr8	87468436	G	A	CAGGCATTGTTAAAGTGGTTAGTTA	GTGTGTCTTGGTCCTCGGAA	Non-coding Homozygous	TP
chr8	91439253	A	G	TGCTGCTGTGAGTTGTAATGTAAGT	GACTCAGAGAAGACGAGAACAGC	Non-coding Homozygous	TP
chr8	96148228	A	G	GAGCTTTTATCGTGAATAGCAACAT	GCTTCTGATTTTTATCCAACAAAGA	Non-coding Homozygous	TP
chr8	77663410	C	T	CTCCCAAGTGCCAGGATTAAAG	GGAAGAGGAAGAAGAAGAAGAAGAA	Non-coding Homozygous	Unknown
chr8	96929639	A	T	AAATAAACTGGAGAAATCTGAGGCT	AGATGATTTTCCCTCTGGATTTTAC	Non-coding Homozygous	TP
chr8	91637313	A	T	ACCTTTCCTAATTCAAAATCAGCTT	AGCAAGGGAGAGATGGTAGAATGTA	Non-coding Homozygous	TP
chr8	72409696	A	G	CTTGCTAGATGCGGTCCTAATTAAC	GAGCTGTGCCCAGATTGTAGTAG	Non-coding Homozygous	TP
chr10	75494347	C	A	ACAGGCAAGAAGGAAAAGAATGAG	ATTCGAATTTCAGTTTTCAGTTTTG	Non-coding Homozygous	TP
chr8	70491385	A	T	AAAAAGGAAATACATCCGGTATAGG	TGACTGTTTGTGCCATCTTTACTTA	Non-coding Homozygous	TP
chr8	84771825	A	G	ATCCATGTGCAATTTAATTCTCTGT	AACACCTTTTTGCTTATTTTCACAG	Non-coding Homozygous	TP
chr10	74503812	G	T	TCTGTGCGAAAGCTGGTAGA	GGCCCAAGCTCACAGTAAAA	Non-coding Homozygous	TP
chr5	64196341	A	G	ACGAAAGCATAGAGCAGTTATTCTG	AGCCTTGCAATATACAGGTAAGGTA	Coding Heterozygous	TP
chr4	123152578	T	C	GGATACAAACTGACCTTAAGTTATTGA	TCTGGAATCCCTCTGGTCAT	Coding Heterozygous	TP
chr8	112051551	A	T	GTCTCTTCTTTGACCTCCTCAGTT	GCAGTCTATTTTGGATGTGCTTT	Coding Heterozygous	TP
chr5	38711842	T	A	AGCCTGCGAGCATCCATC	TGATGGGGTTGAGAGAGGAG	Coding Heterozygous	TP
chr10	57929942	C	T	TATTTTTGCAAACTAAGGCATTTTC	GAATATTTGGTAGGAGATGGTGTTG	Coding Heterozygous	TP
chr10	79354457	C	T	GTGCTGCACACTGCATGG	CGGGTGAGGCACTACACTCT	Coding Heterozygous	TP
chr10	121229763	A	G	GAGGTCAGGAGCTAACAATCTCAC	AGACCTACAAATTAAGGACAGCCAC	Coding Heterozygous	TP
chr4	118399428	G	T	CTTTGCAATGAATCTTCCTTACTGT	CAGATCCTTAGAATGGCTATGAAGA	Coding Heterozygous	TP
chr11	65831008	A	G	TAGAGTTCATATATTTCCTTGGGGC	TAGCTCTTCTCTCTCTCTTTCTCCC	Coding Heterozygous	TP
chr4	102102720	C	G	GTGTCTAGGTTGGTGATTTACAAGG	TAGTGCTAGAAGCAGTCTAAATGCC	Coding Heterozygous	TP
chr4	147129703	T	C	CTGATGAACGCTAAGATGAGATTTT	CAGATCCATCTTAGTATCCATCAGG	Coding Heterozygous	Unknown
chr11	33969900	T	C	CTCTGACTCTCTGTTGGGTACTGAT	AGCATTAAGATCAATGTAAAGCCAC	Coding Heterozygous	TP
chr5	145582831	A	T	AGTCCTGAACTTCTTGCTAACTGTG	AAAATGTAACTAAGGACTACGGGCT	Coding Heterozygous	TP
chr4	102915369	T	A	TGGGCATATATTTCCTTGATGTTTA	CTGTTTAGACAGAATACGCCAATC	Coding Heterozygous	TP
chr5	77830595	A	G	AACCAATTCCCAAACTAGGTAGACT	TAATACAGTTCCTCATGTTGTGGTG	Coding Heterozygous	TP
chr4	102102719	G	A	GTGTCTAGGTTGGTGATTTACAAGG	TAGTGCTAGAAGCAGTCTAAATGCC	Coding Heterozygous	TP
chr8	126957907	G	T	GGTCTCGTCCTCCTCGAGTT	GCTACGAGAGCCTGACACATC	Coding Heterozygous	TP
chr11	59598709	T	C	ACATACTCGGTGTCCTCCATAATC	TCTCTCTGTAATCTCAGTACCACCC	Coding Heterozygous	TP
chr6	55013134	G	A	AGATTTTGATCTTGTTTTTCCTTCC	CTCCCTGAGAACCTTACCTGTTC	Coding Heterozygous	TP
chr6	83002774	G	A	TGTAAGACAGGCTCACTTTTCTTTT	TCCTGGTATACAGACAGACAGACAG	Coding Heterozygous	TP
chr6	113434047	G	C	GTGCTTGAGCCTGTTTCTGAG	CTTACCCAAATGCCATACAGTTAAA	Coding Heterozygous	TP
chr10	93596769	T	C	ATAGAAGCCTACCGAAAAGATGTG	GAATTCACTTGTGCAAATCATACAC	Coding Heterozygous	TP
chr4	143401586	A	G	GGAATGTAGGAAGATGAGAAATCAC	TAACCAAAGCCACAGTCATAGAAAT	Coding Heterozygous	TP
chr11	94967924	G	T	AGCGTGTTTAGAGGAAGAAGAGAG	GCTCCTATCCTTTCCCTCTGTAG	Non-Coding Heterozygous	TP
chr2	43148044	A	G	AAAAGTGTTGGGTTTGGTGTCT	ACTTTAGCCTTAATGCTCAGAGAAG	Non-Coding Heterozygous	TP
chr4	119876618	T	C	TATCACACTGGACTTTCCTCTTTTC	ATGACTGCTGTCCCCTTGTTAC	Non-Coding Heterozygous	FP
chr6	69167249	A	T	TCTCTCTGACTCTCTCTCTCTGG	GGATACCATGGGAAATATAAATGAA	Non-Coding Heterozygous	TP

chr19	29483659	A	T	GGAACAAAGAAAGAGCTTATGAGTG	TTTTTCCACCTTATATTGCATTTTC	Non-Coding Heterozygous	TP
chr10	108727258	T	A	ACACAATCTGTCAAATAGTTCTTGG	TCGTTTACAGAAACTAGGTGTTTGC	Non-Coding Heterozygous	TP
chr9	31588417	T	G	GCATTCCAAATCCTAGATGACTATG	GTTTCTTTGGGGTTACTGGTTAGTT	Non-Coding Heterozygous	TP
chr10	65179218	A	G	TAAGATTAGTTGGGCTCTTCTGCTA	AAATTGCAGTTAATGTTGTTGGAAT	Non-Coding Heterozygous	TP
chr5	150645401	C	T	ACCCTGAAGCTACTGAACATACATC	CTGAAGACAGAACATCAGAAGTTGA	Non-Coding Heterozygous	TP
chr2	17426021	A	T	ATTTTAAATGTCTCAGATGCAGGAG	GACTGCCTTGTCTATTTATGAGGTG	Non-Coding Heterozygous	TP
chr6	100774589	A	G	AGTGAGAGACCTCGTCTCAAGG	TTCTCAACTCTCAAGGGACATAAAC	Non-Coding Heterozygous	TP
chr4	145462750	C	G	TCTTTCTCTCCCTTAAGTAGCTTCC	CAAGCCTTTAGGATTTGCTTTAATC	Non-Coding Heterozygous	FP
chr2	28619987	G	A	ACTAGGCCCCATCTTCTAGAGTTT	GAGGTAGAGACAGGAGGATCAGAAT	Non-Coding Heterozygous	TP
chr6	72034052	A	T	AAGAAGAAGAAGAAGAAGAAGCAGC	TCCCAAGATAGAGAGAGAGAGAGTG	Non-Coding Heterozygous	TP
chr6	44456662	A	G	TTCAGCCCATTATCCTTGGT	TTCCTTTCCCCACCTATCCT	Non-Coding Heterozygous	TP
chr2	31644666	A	T	CATCTCCATAGTCTTTTAGAACCCA	CTAGCTGCAAGAACAAGGCTCT	Non-Coding Heterozygous	TP
chr10	113805663	G	C	AAATCTCGTGGGGAAGGACT	AAGCAAGCAAGCAAGGAAAA	Non-Coding Heterozygous	FP
chr11	6605800	C	T	TTTCTTAATAGACATGTCCCCTGAC	GAGGTGGACCTAATGATCAATGAC	Non-Coding Heterozygous	FP
chr19	39996078	A	G	TGCACATACTAGTCTCAGCAGAATC	GCAACAGTAGGTGAGAATGGTAATC	Non-Coding Heterozygous	TP
chr8	32391933	A	T	AATTGCTAGTGCTACTGTGGTAAGG	CCCCATGATCAACATTTGTAAAGTA	Non-Coding Heterozygous	TP
chr6	100153871	A	G	CTCCCTCATTCATGCTGTCTTAG	GAGACCAGATTAGAATGGAGACAAA	Non-Coding Heterozygous	TP
chr11	28123635	T	A	TCACTGTCTTTGTATTGAGTTAAGGAA	TAAAACTGTGATTCTGTGATCCTGT	Non-Coding Heterozygous	TP
chr4	124138040	C	G	GTGTGTGTGTGTGTGTGTGTGTGTA	CACATACTCCTTCCATCATTTCTTT	Non-Coding Heterozygous	Unknown
chr10	66292950	G	A	AGGTAGGTAGGTAGGTAGGTAGGTAGA	ATCTGTCTGTCTGTCTGTCTGTCTG	Non-Coding Heterozygous	Unknown
chr10	67397252	A	T	CACCTGTACATGCACATACATCTTA	ACATCCAGGAAGTGTATGAATAAGC	Non-Coding Heterozygous	TP
chr8	3832950	A	G	AAAGGACAGTCTTAAGCCAGATACA	AGGGGAGACTACATTTCCTCTCTT	Non-Coding Heterozygous	Unknown
chr11	20336828	T	A	ATATTTCCAATATTCTTCCTCGACC	TGTTGATTGGTAGTATGAACAGGAA	Non-Coding Heterozygous	TP
chr11	5265846	T	A	AAAGTTGCTTATGCCCCAATTAAAG	TGGGGTAAGAACTAATCCTATGTGA	Non-Coding Heterozygous	TP
chr4	124954916	G	T	AGTTCCTGGTACTTTTTGGAGAGAT	CTCAACACAAGCAACTTAAGAGAGG	Non-Coding Heterozygous	TP
chr2	12336455	A	G	CACATTAAGACAATCAAATGTCAGG	GGCTAACTTGGTGTTTATTACATGG	Non-Coding Heterozygous	TP

Table D.2: Primers for second pedigree low coverage non IBD variants. showing SNP genotype from variant caller for each mouse. Sanger: True Positive (TP), False Positive (FP) Failed Sequencing (Unknown - assumed FP)

labelD.2

Chr	Pos	Ref	SNP	Left (Forward) Primer	Right (Reverse) Primer	Genotypes in mouse 1, 2,3 and Sanger:
chr10	51889817	A	T	TGAAGCTTTGAATAATTTAAGGACAA	GAATAGTGACTGAAATGCTGTTTCC	1/1 0/1 0/1 TP
chr1	46261086	A	C	ATGAAATACAGTGAAAACATGAGCA	CACTACTGTAGGAGTGCTGACACAT	1/0 1/1 1/1 FP
chr11	72731358	T	A	GATGTTGTCTGAAGAAAGGGTTGTA	TACCTTTATGACCTTTGAATGAAGC	1/0 1/0 1/1 TP
chrX	3907117	G	A	TCAGAATCTGTAGTAGAGAAACCCA	GAGGTCCATGAACATTAGAACTTGT	1/1 1/1 0/1 FP
chr11	101187980	T	C	CTATATCAACACAGCCTTCATGAGT	CTTTCTGAGCTCACTCCTTGATTT	0/1 0/1 0/1 TP
chr7	149044528	T	C	ACTCCTCTCACAGTCACCTCAGT	GACAAGGAAGCCTTTGAGGTT	0/1 0/1 0/1 TP
chr12	114018097	C	G	ATCTTGAGGCTGGGCATCT	CTGAAGACCCCTGACCTGAG	1/0 0/0 1/0 TP
chr14	43347611	G	A	CTTGGTCTGGTGTCTGTGTGTATT	GAGAACCCTAGAATCTTCTGTGATG	1/1 1/1 1/1 FP
chr2	145747631	A	C	ATACATTCAACTTGGGATCTCACC	AAAGGATACATAGAACTGGAACTGC	0/0 1/0 1/1 TP
chr2	82823507	G	T	AAAAAGTAAACGATTCCACAGAAGA	AAAGGAATTTGCCTTGACATAAAA	0/1 1/1 1/1 Unknown
chr5	139048775	A	T	ATACTGCTTTTGTGTTGAAGGATTT	GTTTATTCTGTGGAAGGACGTGTAT	0/0 0/1 1/1 TP

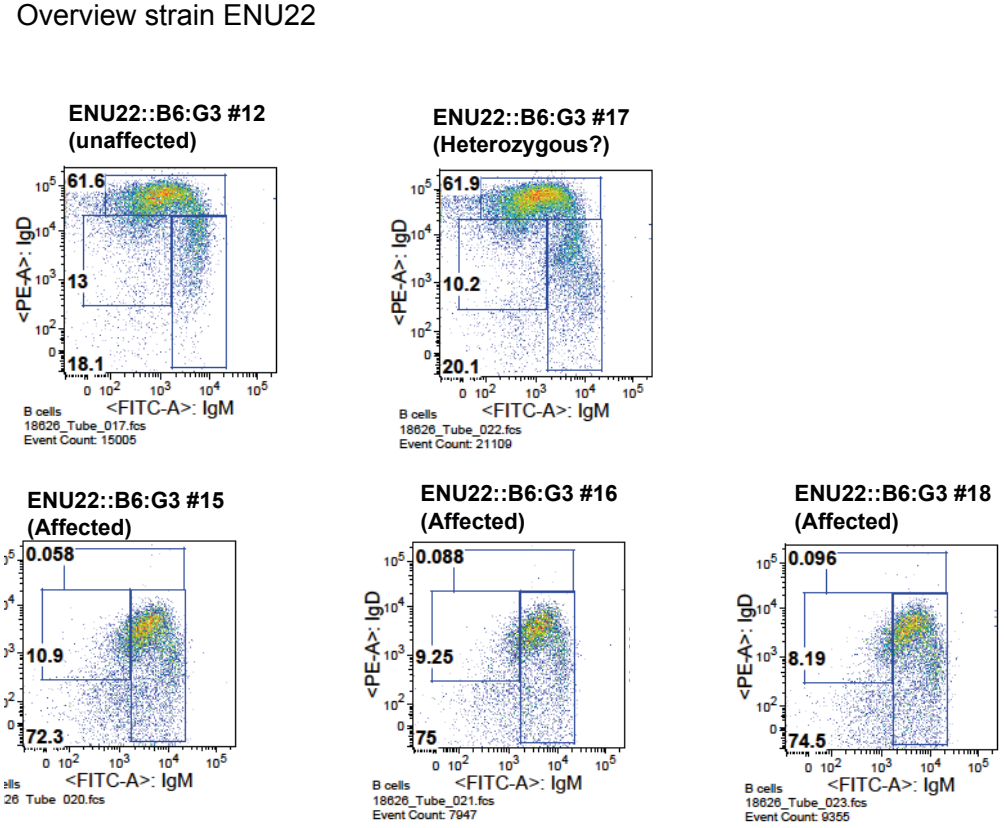
chr11	68628466	G	T	GTGGAGAGTGAGCCTTGTTTAGA	TTGACATCACTGGTCTTACTTTCTG	0/1 0/1 ./. Unknown
chr4	137104386	C	G	AATATGTGTGCAGCGTTCTGAG	CAGGTCTCAAAGAGCTTTTCTTCT	0/1 1/1 1/1 TP
chr14	66378562	A	T	GTGACATGGATTTCACAATGATTC	GTACCTTACCTCACCATCCTTACCT	0/0 0/0 1/1 FP
chr6	126688537	G	A	GATGTGGAGTCGGAAGGTAGC	ATGACCACGGTAGGTTATGGG	0/0 0/1 1/1 TP
chr11	70028825	C	T	AGAAGAGGGGCTAAAGAATGTTTG	AAAGTATCAAGGAGGAACTGCAAG	0/1 0/1 1/1 TP
chr6	32135423	A	T	TCTTATCAATGAGGTCAGCCAGTAG	AAGGTGTTCCTGCTCTCTTTCAT	0/1 0/0 0/1 FP
chr1	53549974	T	C	ATCTGGTGTAATTCTGATAGGCTTG	GAATGAAGATTTTCAACTCAAAGGA	1/1 1/1 1/1 FP
chr2	83262425	A	T	AACTGTGAAAGTCTTGGTTTAGGTG	ATCTCAGAGAATAGAATGTGAAGCC	0/0 1/1 1/1 FP
chr9	21356924	T	A	AGGTCGTAAGCTGAGGTAAAAGAAT	AGTTCAAAAAGCTACTCTATGGCAA	1/1 1/0 1/0 FP
chr4	43597914	C	T	AAAAGCACCCCAAAGATAAACTTC	CTCCTCCAGAGTCAGCAGAAAT	1/1 1/1 1/0 FP
chr4	3182839	T	C	TATTGGCTCATAATATTGCTGTTGA	TCCTCTCGAGGACTTTGAGTTATTA	1/0 1/0 1/0 Unknown
chr5	87716559	C	T	CAGCTATCATTCACTTCAATCAGAA	GGCCTGGTTGTTAATGGTAGATAAT	0/0 0/0 1/0 FP
chr1	4329977	G	A	GAAAATTTCCTGAATAGAACAGCAA	CATTTGATTTTTGATTTTCTGGAGT	0/0 0/0 1/1 FP
chr10	3042685	T	C	TTGTATTTACAACTATAGGCCAGGTG	CACACACACAGAGAGAGAGAGAGAG	0/1 0/0 0/1 FP
chr2	94795429	A	G	AAACAGGTTTTTCAGAGTTGATGAC	TGAAGTTATAATCATTCACAGGGGT	0/0 1/1 1/1 FP
chr3	82193049	C	T	GGATAGTACCATGACTTATGTCCCA	ACAGAAACGTGGATTAAAGTACAGG	0/0 1/0 1/0 TP
chr3	91370835	A	G	ATAGCTGACTGTGAGCATCCACTAC	CATGGAAGGAGTTAAAGAGACAAAA	0/0 0/0 0/1 FP
chr4	89184839	A	G	AGAATGGTTTCCAAGCTACAAGAT	TGTTCAAACATATTCCAGTAAGCAA	1/0 1/1 1/0 TP
chr2	165946667	G	A	GGTGGAGTACTAAGCATTGGAAAAG	ACAATCTCACATGCTAGATCAACAA	0/0 0/0 1/0 FP
chr6	136724323	A	T	TAAAATGTCAAGAAAACACATCCTG	AGCTTACAACTCTCGGAAACTCC	0/0 1/0 1/1 TP
chr14	66527259	C	A	AGGGACGAATACATCTTTCTAGGAT	AAAGAAAAGAGCTGAATGTTGAGAA	0/0 0/0 1/1 FP
chr16	18592638	T	C	CTCTCCTCTCCTCTCCTCTCCT	AGGATCTCTGCATTGGAAAGTTAG	1/1 ./. ./. Unknown
chr5	143163864	A	T	GCTGTAATTTTGCTCCTGTTATGAT	CTTTCAATGTGCAAGAGGTGAAG	0/0 1/0 1/1 Unknown
chr1	13025454	A	T	TCAGAAATGAAAAGGGAACAATAAA	TCAAAATAGCAAAACCTTTTTGAAG	0/1 0/1 0/1 TP
chr1	6170534	G	T	CACCTGAGATCATTGGAAAACATA	TCTTTTTACCTTTCCTGAACTTCG	0/0 ./. 1/1 TP
chr3	74125312	G	T	GAGTGTGGGCTTAGTCCTATAACAA	TGGAAAATGAACCAAAATAGTCTCT	0/0 0/1 1/1 TP
chr11	99020131	A	G	CCGAATTCTTTCTCTCCTACATTTC	GTCCAAACACTAAAGATCCTTGTCC	0/1 0/1 0/1 TP
chr3	83183859	C	T	GAGTTACAGAGACAAAATTTGGAGC	CTCCTTGGGTACTTTCTCTAGCTCT	0/0 1/0 0/0 FP
chr2	102124138	A	G	CCCTAGCCCCATTTATTTATGTTAT	TAAAGCATGAACGTAAACATTGAAA	0/0 1/1 1/1 TP
chr6	33734748	A	T	ATAAGAAGTAGAAAGAGTGCGCTTG	GGACTTCACTGGAGACATTTTACAT	1/1 0/0 1/1 TP
chr17	84861317	G	A	ACACACACACTTACTGCCCTAGTCT	ACTAGGACAGTAAGTGTGTGTGTGG	1/1 ./. 1/1 Unknown
chr16	63177786	A	G	CACTCCGTGAAACAACATACAATAG	GTAGAAGAGAAAGGAAGACCTGGTT	0/0 0/1 0/1 TP
chr10	53641804	T	A	CATTCTCATTTAATGCAACTGAACA	CATAGTTTATTTTAGGAGGCAGCAA	1/1 0/1 0/1 TP
chr6	60489599	G	T	GAATCTCACTTAGAAGGGGGAATAA	TCACTCCTATTGATTCTTAGATGCC	0/1 0/1 0/1 TP
chr3	59781263	A	G	GACTTACACACACACACACAAAA	GACTAAGACATTTGGGAAACTGAGA	0/0 1/0 1/1 TP
chr11	78281164	T	A	AAAAAGCTAAAATTCCTGCTGC	GTGTCACTACCGAGTTAGCTGAGTT	0/0 1/0 1/1 TP
chr6	122418894	A	T	TGGATGTGTCTATATTACGAGATGG	ACGGTATCTTTACCTTTGTGTCTTG	0/0 1/0 1/1 TP
chr11	81178333	C	A	AGTTTTAGCAAGTGAAAGTTGGAGA	TAAGTGGATATCCTACCCACCTGTC	0/1 0/1 1/1 TP
chr6	133915812	T	A	TGTATTGGGTTTTATGTGAGGTCTT	TAGACATAGCCATTTTACCAAAAGC	0/0 1/0 1/1 TP
chr6	38167512	A	T	CTCCCCCAGTCAGAGACTAGTTAAT	AGACTCGCTCATCTTACTGACCTAA	1/0 1/0 1/0 TP
chr17	38389993	T	C	TTTCTTTCTCAGCCTGTTTATCAAT	CTCTCTCTCTCTCTCTCTCTCCC	0/0 1/0 1/0 Unknown
chr2	115767708	T	A	CTTTTCTCGCTGAAGTCATCTAAAG	GGAGAGAAAGTGCAACTCATTTTTA	0/0 1/1 1/1 TP
chr6	34942306	A	G	ATTCAGCTGTATACATTTTTGAGGC	CGTTCTTAACTCTTCACAGTCATTG	1/0 0/0 1/0 FP
chr3	56214691	A	G	AAACCTTTAGTGTAGATGTAGGCCA	ATCAAAACAACTTTAGTGGCAAGTC	0/1 ./. 0/1 FP
chrX	84319042	T	C	CTCTTTGATGATTTGTTTTTGCTTT	TCTTTAATCTCATCACTCAGAAGGC	./. 1/1 1/1 Unknown

chr2	82823507	G	T	AAAAAGTAAACGATTCCACAGAAGA	AAAGGAATTTGCCTTGACATAAAA	0/1 1/1 1/1 FP
chr2	25471530	G	A	CAACAGACTTACACTGACGCTCTT	GGACTCAGCCAAATAGGCTTC	1/0 1/1 1/0 TP
chr10	59791020	T	C	TTATATTTAGGGAAATGGAGGGAAG	ATACATTTCACAGAAGTGCTGAAGG	1/1 1/1 1/1 TP
chr14	44082649	G	A	GAACTGGAATTAGATCTTCGTTGAA	GGCTATACTCAACAAAACTGGAAAA	1/0 1/0 0/0 FP
chr5	144295535	A	T	TATTTATTTTTAATTTAGCCGGGCA	CTAGGGTCCCCAACTCACTATGTA	0/0 0/1 1/1 TP
chr4	137271581	C	A	CATGTAATAGACCAGGACAGGACAG	AATAGTAATTGAAGTGCTCCTGTGG	1/1 1/0 1/0 TP
chrX	4767755	G	C	AGTGGACCTGTGAGTAAAAGCC	GCATCTCTATTCACACTGACAACAC	1/1 ./. 1/1 Unknown
chr12	118561783	A	G	GTTTCTTACCACTTCCTCAGACTCA	AGAAGTTCTCAGTTGTGCCTCTCT	1/1 1/0 ./. Unknown
chrX	3918573	T	G	ACTACTTTACCTTGTGCCACTATGC	CAGGAAAATCTCAACAGAACTCAAT	1/0 1/0 1/0 TP
chr6	142116493	T	A	TGTGCTTTCATGTTGCTACTTTATT	AAATGAAAGATTCTCGAATTCCC	1/1 0/1 1/1 Unknown
chr4	105692962	A	T	TTTGGGGCTCCAATAATTAAGTAAG	CTGGTCCACTTGCTTTATCATTAGT	1/0 1/1 1/0 TP
chr19	43900761	T	A	GAGCTCAGTTGTCAGTATGAGTCAG	CTGACATCCTCACACAGACATACAT	1/1 1/1 1/0 TP
chr17	30176494	A	G	TTTGAAGATCATGGATGCATAGTAA	AAAGGAGATTTACAGACGAGTTTCA	0/0 1/0 1/0 TP
chr3	60540665	A	G	GTAATTGTATGCATATTTCTTGCCC	ATGACTACTTGTGGGAATCACTTGT	0/0 0/1 1/1 Unknown
chr14	81758650	A	G	TTCTCAACAATAAAAGAACCTCTGG	AGACAGGGTTTCTCTGTGTAGTCCT	0/0 0/0 1/0 Unknown
chrX	124413795	T	C	ACAGTGAGAAAATAATGAAGCCAAG	ATCTTCTCTGAGACTCTGCTACCTG	1/1 1/1 ./. Unknown
chr1	54858008	A	T	TGTTCATTAAGATTGTTGACAGTGA	AGTAAGGCTGTTGGATATACCTGTG	0/0 1/0 1/1 TP
chr17	84843952	A	T	TGGTTAGAAGCTGCCTAGTAAAAGA	CAGTTTGGCCTTATTTGTAATTTTG	./. 1/1 1/1 TP
chr3	65616079	T	C	ACTTGCTTTTCTCCTTTTACCATTT	CAGGAGTCAAACCAATCATTTCTAT	0/0 0/1 1/1 TP
chr2	118251979	A	G	CAGCATCTGATTGTTATCTTTGGA	AAAGAGAAAAGAAAAACCAACAACA	0/1 1/1 1/1 TP
chr12	5066118	C	T	GAGTCCTTTGGTAGCCTGGTATC	AAGCAAGGCTACCAAAGTACTCC	0/1 0/1 0/1 Unknown
chr14	7559653	T	C	TGAAGAATTATGTTCCTCTTAAGGC	CTTCTCTTCACATGGTGGTCTATCT	1/1 1/1 ./. TP
chr11	107967175	G	A	CCAGACCTCACTTTTCTTTCATTTA	GGACACCTGTCCTCTGGTCC	1/1 1/1 1/1 Unknown
chr6	138241120	A	T	GAGAGAGAGAGAAAGAAAGGAAGGA	TCTTTCTTTCTTTCTTTCTTCCTCC	0/0 0/0 1/1 Unknown
chr5	92151731	A	T	GTCCCTGAACTGTGGTAAGTTAGAA	CATGCTTTACACAGCTTTGACTCTA	0/0 0/0 1/0 FP
chr5	101596050	A	C	TATGTCACAGCAATGAATAAAGCAT	TGAACCTATCTTTAAGCCCTTTTCT	0/0 0/0 1/1 FP
chr5	146494346	A	C	GAAGAATTGAGTTGGAATTTTGATG	CACAAATAATACAAAAGACCTTGGC	1/0 1/0 1/0 TP
chrX	4260792	T	G	AATCACAGGAGACAGAGATAAGCAG	ACATTATCACCAGCTTTCTGTTTTC	1/1 1/0 1/1 TP
chr4	105057916	A	G	TGAGGCTGAAATCTTATCCTTTCTA	TTTTGTAGGGTTGAGCAAAATAATG	1/0 1/1 1/0 TP
chr3	72456852	T	C	GGTACTTTCTCTAGCTCCTCCATTG	TGCATCAAATTGGATAACATCAATA	0/0 0/0 1/1 TP

Appendix E

Phenotype in ENU22

Figure E.1: Phenotype in ENU 22, showing loss of IgD+ B cells



Appendix F

Phenotypes, expression and functional terms for 23 candidate homozygous variants in SRNS patients

F.1 Phenotypes, expression and functional terms for 23 candidate homozygous variants in 0001

Gene	MAF	Mouse Model	Human Disease Association	Expression	Function	GO terms
ITIH5	<0.01	No model	Congenital uterovaginal aplasia	Female reproductive tract	Extracellular matrix protein	Peptidase inhibitor activity: serine–type endopeptidase inhibitor activity: hyaluronan metabolic process: negative regulation of peptidase activity
ZBED4	<0.01	No model	No	Broad, especially NK cells	Zinc finger	DNA binding: metal ion binding: nucleic acid binding: protein dimerization activity

SPATA13	Not seen	No model	No	Expression high in placenta, spleen, and kidney, moderate in lung, small intestine, liver, brain, and heart, and low in skeletal muscle. - may be in tubules rather than glomerulus	Guanine nucleotide exchange factor (GEFs) for CDC42. Increases both RAC1 and CDC42 activity, but decreases the amount of active RHOA	Lamellipodium assembly, cell migration, protein binding, cytoplasm, intracellular, phospholipid binding, regulation of Rho protein signal transduction, guanyl–nucleotide exchange factor activity, filopodium, nucleus, filopodium assembly, regulation of cell migration, lamellipodium, Rho guanyl–nucleotide exchange factor activity, Rac guanyl-nucleotide exchange factor activity, ruffle membrane
ELP5	<0.01	No model	No	Highest expression in heart, brain, liver, skeletal muscle, and testis	1 of 6 subunits of the elongator complex. Elongator promotes RNA polymerase II transcript elongation through histone acetylation in the nucleus and also regulates translational fidelity through tRNA modification in the cytoplasm	Positive regulation of cell migration: regulation of transcription, DNA-dependent

PRKD1	Not seen	Mice homozygous for a knock-out allele exhibit partial embryonic lethality. Mice homozygous for a knock-in allele display partial embryonic and perinatal lethality.	No	Mainly prostate? Lung and skin. Weak/ patchy expression in podocytes cytoplasm in some cases of IgA nephropathy(Sigdel et al. 2011)	Expression of constitutively active PRKD1 in an invasive rat tumour cell line enhanced phosphorylation of cofilin, blocked formation of free actin filament barbed ends, and directed cell migration.	ATP binding, Golgi apparatus, Golgi organization, Golgi vesicle transport, angiogenesis, apoptotic process, cell cortex, cell proliferation, cell-cell junction, cellular response to oxidative stress, cellular response to vascular endothelial growth factor stimulus, cytoplasm, cytosol, identical protein binding, inflammatory response, innate immune response, integral to plasma membrane, integrin–mediated signaling pathway, intracellular protein kinase cascade, intracellular signal transduction, metal ion binding, negative regulation of cell death, negative regulation of endocytosis, nucleus, peptidyl-serine phosphorylation, phospholipid binding, plasma membrane, positive regulation of CREB transcription factor activity, positive regulation of I–kappaB kinase/NF–kappaB cascade, positive regulation of NF–kappaB transcription factor activity, positive regulation of angiogenesis, positive regulation of blood vessel endothelial cell migration, positive regulation of endothelial cell chemotaxis, positive regulation of endothelial cell chemotaxis by VEGF-activated vascular endothelial growth factor receptor signaling pathway, positive regulation of endothelial cell migration, positive regulation of endothelial cell proliferation, positive regulation of histone deacetylase activity, positive regulation of neuron projection development, positive regulation of osteoblast differentiation, positive regulation of peptidyl-serine phosphorylation, positive regulation of transcription from RNA polymerase II promoter, protein

NUP93	Not seen	No model	No	Widespread	Nuclear pore structural component, critical subunit of the 120-million-Da nuclear pore complex, which functions in active transport of molecules between the nucleus and cytoplasm	mRNA transport: nuclear pore complex assembly: protein transport
SCRIB	<0.01	Circletail (crc) is a mouse phenotype characterized by craniorachischisis, a severe neural tube defect, 1-bp insertion in the Scrb1 gene, resulting in a frameshift that leads to premature termination of the protein. crc mouse mutants developed heart malformation and cardiomyopathy. The podocyte-specific knockout of the basolateral polarity protein SCRIBBLE does not induce any podocyte phenotype, suggesting that the podocyte might be dominated by apical polarity signaling	No	Includes Podocytes	SCRIB is a cytoplasmic multimodular scaffold protein targeted to epithelial adherens junctions and neuronal presynaptic compartments. SCRIB and its orthologs in vertebrates and invertebrates participate in cell polarization	Neural tube closure, positive regulation of receptor recycling, asymmetric protein localization, cell proliferation, synaptic vesicle targeting, morphogenesis of embryonic epithelium, cell-cell adhesion, cell migration, virus-host interaction, activation of Rac GTPase activity, establishment of apical/basal cell polarity, wound healing, positive regulation of apoptotic process, interspecies interaction between organisms, negative regulation of mitotic cell cycle, synaptic vesicle endocytosis, positive chemotaxis, apoptotic process involved in morphogenesis, mammary gland duct morphogenesis, protein localization to adherens junction
RNF17	Not seen	Homozygous null mice display male infertility, azoospermia, arrest of spermatogenesis, and small testis.	No	Expressed in only male germ cells - in mice	Ring finger protein essential for spermiogenesis	Protein binding, cytoplasm, nucleic acid binding, multicellular organismal development, nucleus, spermatid development, protein homodimerization activity, zinc ion binding
ADAMTSL3	<0.01	No model	No	Most adult and foetal tissues and specific brain regions examined. Slightly higher expression was found in spinal cord, and lower expression was found in spleen and testis. High levels in heart, kidney, and liver. Faint signal in podocytes and more in tubular epithelium of distal nephron	Extracellular matrix protein	Protein binding, metallopeptidase activity, peptidase activity, proteinaceous extracellular matrix, zinc ion binding

PSTPIP1	Not seen	Mice homozygous for a knock-out allele exhibit defects in immune cells with T cell hyperresponsive to antigen receptor stimulation.	Pyogenic sterile arthritis, pyoderma gangrenosum and acne	Haemopoietic tissues and cell lines, with expression up-regulated in activated T cells	Binding protein of the PEST-type protein tyrosine phosphatases	Oxidoreductase activity, protein binding, cleavage furrow, actomyosin contractile ring, cell adhesion, cytokinesis, lamellipodium, perinuclear region of cytoplasm, actin binding, stress fibre, protein phosphatase binding
PRKD1	Not seen	See above	No	See above	See above	See above
USP15	<0.01	No model	No	Expression in skeletal muscle, kidney, heart, placenta, liver, thymus, lung, and ovary	Regulation of intracellular protein breakdown, cell cycle regulation, and stress response	Cytoplasm, BMP signalling pathway, ubiquitin thiolesterase activity, transforming growth factor beta receptor signalling pathway, ubiquitin-specific protease activity, negative regulation of transforming growth factor beta receptor signalling pathway, transforming growth factor beta receptor binding, cysteine-type endopeptidase activity, protein deubiquitination, nucleus, SMAD binding, catalytic activity, cysteine-type peptidase activity, ubiquitin-dependent protein catabolic process, pathway-restricted SMAD protein phosphorylation, monoubiquitinated protein deubiquitination cytoplasm
ATP13A5	Not seen	no model	no	Under expressed in renal system	Probable cation transporting ATPase	ATPase activity, metal ion binding, cation transport, ATP binding, nucleotide binding, integral to membrane, cation-transporting ATPase activity
ADH1C	Not seen	Homozygotes for targeted null mutations exhibit impaired metabolism of (and sensitivity to) ethanol and retinol	Protection against alcohol dependence. Assoc with Parkinsons Disease	Active in intestine and kidney in foetal and early postnatal life, and persists in the stomach and liver in adult life	Catalyzes the rate-limiting step for ethanol metabolism: the oxidation of alcohol to acetaldehyde.	
NOC4L	< 0.05	No model	No	Testis, B lymphoblasts and endothelial cells	Ribosomal RNA formation and nuclear export	rRNA processing, protein binding, ribosome biogenesis, integral to membrane, binding, nucleolus, nuclear membrane

CCDC33	< 0.05	No model	No	Male germ cells (mouse), liver	May be involved in spermatogenesis	C2 calcium/lipid-binding domain, CaLB, C2 calcium-dependent membrane targeting
GRM6	< 0.05	Abnormal eye electrophysiology	Night blindness, congenital stationary type 1b	Retina, pineal	Metabotropic glutamate receptor	Plasma membrane, G-protein coupled receptor activity, glutamate receptor activity, visual perception, locomotory behaviour, synaptic transmission, integral to plasma membrane, detection of visible light, integral to membrane, signal transduction, sensory perception of light stimulus, retina development in camera-type eye, G–protein coupled receptor signalling pathway, G-protein coupled glutamate receptor signalling pathway, adenylate cyclase inhibiting G–protein coupled glutamate receptor activity, adenylate cyclase-inhibiting G-protein coupled glutamate receptor signalling pathway, new growing cell tip
BTNL9	< 0.05	No model	No	B lymphoblasts	Butyrophylin like domain	Protein binding, integral to membrane
CSMD1	< 0.05	Normal behavioural phenotype	Association with head and neck cancer	Adult brain, foetal brain, spinal cord	Possible regulator of complement activation in the central nervous system	

MESP1	< 0.05	Homozygous nulls die by embryonic day 10.5 with growth retardation and heart defects	No	Nascent mesodermal cells, not in adult tissue except testis (mouse)	Development of mesoderm including germ cells	Negative regulation of transcription, DNA-dependent, cardiac muscle cell differentiation, neurogenesis, transcription, DNA–dependent, heart morphogenesis, positive regulation of Notch signalling pathway, mesodermal cell migration, protein dimerization activity, positive regulation of transcription from RNA polymerase II promoter, nucleus, Notch signalling pathway, gastrulation, positive regulation of transcription, DNA–dependent, cardioblast anterior-lateral migration, positive regulation of Notch signalling pathway involved in heart induction, embryonic heart tube morphogenesis, positive regulation of striated muscle cell differentiation, sinus venosus morphogenesis, cardiac ventricle formation, cardioblast migration to the midline involved in heart field formation, growth involved in heart morphogenesis, cardiac cell fate determination, positive regulation of hepatocyte differentiation, lateral mesoderm development, enhancer binding, secondary heart field specification, cardioblast migration, endothelial cell differentiation, negative regulation of mesodermal cell fate specification, negative regulation of endodermal cell fate specification, heart looping, sequence–specific DNA binding transcription factor activity, signal transduction involved in regulation of gene expression, cardiac vascular smooth muscle cell differentiation, transcription regulatory region DNA binding, positive regulation of heart induction by negative regulation of canonical Wnt receptor signalling pathway,

WDR81	< 0.05	No model	Cerebellar ataxia, mental retardation and disequilibrium syndrome	Liver, spinal cord, brain	Neuronal development	Protein binding, transferase activity, transferring phosphorus-containing groups, negative regulation of phosphatase activity
ZNF592	< 0.05	No model	Spinocerebellar ataxia, autosomal recessive 5	Brain, skin, human foetal tissue, heart and skeletal muscle, liver, haematopoietic	Developmental pathway and cerebellar development	DNA binding, protein binding, regulation of transcription, DNA-dependent, transcription, DNA-dependent, nucleus, cell death, zinc ion binding

Table F.1: Animal model phenotypes, human disease associations, expression profile and functional terms for 23 candidate homozygous variants in 0001. The minor allele frequency (MAF) is determined from the highest observed frequency in 1000 genomes populations as a whole, the Pakistani 1000 genomes cohort (only for patient 0001), or the WGS500 union file. Human disease associations are based on data in Online Mendelian Inheritance in Man (OMIM http://omim.org). Mouse models were searched for in Mouse Genome Informatics (MGI www.informatics.jax.org). Expression data was obtained from multiple online sources including bioGPS (http://biogps.org) and Gene Expression Atlas (www.ebi.ac.uk/gxa). Gene Ontology (GO) functional terms were obtained from www.geneontology.org

F.2 Phenotypes, expression and functional terms for 2 candidate homozygous variants in 0002

Gene	MAF	Mouse Model	Human Disease Association	Expression	Function	GO terms
RASAL3	Not seen	No model	No	Lymph node, spleen, lymphoid cellsDetected in renal tubules but not glomeruli	Ras GTPase-activating protein	Ras GTPase activator activity, protein binding, cytoplasm, intracellular, negative regulation of Ras protein signal transduction, phospholipid binding, GTPase activator activity, regulation of small GTPase mediated signal transduction, signal transduction, positive regulation of Ras GTPase activity, intrinsic to internal side of plasma membrane

DMBT1	Not seen	Homozygous null is embryonic lethal, targeted intercalated cell deletion has distal renal tubular acidosis	Possible association with medulloblastoma and glioma	Lung, trachea, salivary gland, small intestine and stomach, most epithelial cells, not detected in kidney	Epithelial differentiation	Extrinsic to membrane, defence response to virus, protein binding, cytoplasm, intracellular, scavenger receptor activity, innate immune response, membrane, proteinaceous extracellular matrix, extracellular region, induction of bacterial agglutination, calcium-dependent protein binding, pattern recognition receptor activity, protein transport, inner cell mass cell proliferation, phagocytic vesicle membrane, zymogen granule membrane, blastocyst development, positive regulation of epithelial cell differentiation, Gram-positive bacterial cell surface binding, virus-host interaction, epithelial cell differentiation, Gram-negative bacterial cell surface binding, zymogen binding, pattern recognition receptor signalling pathway

Table F.2: Phenotypes, expression and functional terms for 2 candidate homozygous variants in 0002. The minor allele frequency (MAF) is determined from the highest observed frequency in 1000 genomes populations as a whole, the Pakistani 1000 genomes cohort (only for patient 0001), or the WGS500 union file. Human disease associations are based on data in Online Mendelian Inheritance in Man (OMIM http://omim.org). Mouse models were searched for in Mouse Genome Informatics (MGI www.informatics.jax.org). Expression data was obtained from multiple online sources including bioGPS (http://biogps.org) and Gene Expression Atlas (www.ebi.ac.uk/gxa). Gene Ontology (GO) functional terms were obtained from www.geneontology.org.

F.3 Expression and functional terms for 1 candidate homozygous variant in 0003

Gene	MAF	Mouse Model	Human Disease Association	Expression	Function	GO terms
LCE1D	not seen	no model	no major skin defect with null allele, detected in kidney by RNA-seq ? predominantly in tubules	Skin	constituent of epidemis	cytoplasm, keratinization, cytoplasmic part, perinuclear region of cytoplasm, cornified envelope, cellular response to calcium ion

Table F.3: Expression and functional terms for 1 candidate homozygous variant in 0003. The minor allele frequency (MAF) is determined from the highest observed frequency in 1000 genomes populations as a whole, the Pakistani 1000 genomes cohort (only for patient 0001), or the WGS500 union file. Human disease associations are based on data in Online Mendelian Inheritance in Man (OMIM http://omim.org). Mouse models were searched for in Mouse Genome Informatics (MGI www.informatics.jax.org). Expression data was obtained from multiple online sources including bioGPS (http://biogps.org) and Gene Expression Atlas (www.ebi.ac.uk/gxa). Gene Ontology (GO) functional terms were obtained from www.geneontology.org.

Appendix G

Homozygous rare candidate variants in the SLE patients

Patient	Gene	Location	Ref	Sub	Type	1000g MAF	WGS500 union
16743	DMBT1	10:124348678	G	A	NS	0.01	HT=0
16743	DMBT1	10:124361496	G	A	NS	unknown	HT=1
16743	CDON	11:125831691	G	A	NS	0.01	HT=0
16743	MMP17	12:132313113	T	C	NS	unknown	HT=2
16743	MMP17	12:132313119	T	C	NS	unknown	HT=0
16743	KRT27	17:38936665	T	G	NS	0.01	HT=0
16743	C19orf55	19:36257805	C	G	NS	0.01	HT=1
16743	STK11IP	2:220473139	G	C	NS	0.01	HT=0
16743	PARP15	3:122329380	C	T	NS	0.01	HT=1
16743	PARP15	3:122329508	T	A	NS	0.01	HT=1
16743	PARP15	3:122329551	A	G	NS	0.01	HT=1
16743	DST	6:56366410	T	C	NS	0.01	HT=0
16743	DST DST	6:56425220	G	A	NS	0.01	HT=0
16743	CCDC28A	6:139094917	G	T	stopgain	0.0005	HT=1
16743	RAET1E	6:150212034	G	A	stopgain	0.01	HT=2
16743	IQCE	7:2646805	C	G	NS	0.01	HT=1
16743	ABCA13	7:48520660	C	G	NS	0.01	HT=1
16743	SRRM3	7:75911007	G	T	unknown	unknown	HT=0
16743	ZNF703	8:37555683	C	G	NS	unknown	HT=0
16743	BMP15	X:50659210	C	CTCT	NF ins	unknown	HT=14
17571	FOXG1	14:29236616	A	C	NS	unknown	HT=3
17709	ANKRD45	1:173616096	G	C	NS	0.0014	HT=0
17709	SOX21	13:95363670	C	A	NS	unknown	HT=7
17709	HOXD12	2:176964947	G	A	NS	0.01	HT=16
17709	FGL1	8:17726069	C	A	NS	0.01	HT=3
17709	WAS	X:48544502	C	A	NS	0.0006	HT=0
17709	PAGE1	X:49452149	G	T	NS	unknown	HT=1
16603	FAM76A	1:28056766	C	T	NS	0.0009	HT=9
16603	FAM149B1	10:75000739	G	A	NS	unknown	HT=46
16603	FANCB	X:14882764	A	G	NS	unknown	HT=0
16603	DMD	X:32536244	C	A	NS	unknown	HT=1
16603	PLP2	X:49030688	G	A	NS	unknown	HT=0
16603	RGAG4	X:71350595	A	C	NS	unknown	HT=0
16603	XKRX	X:100177889	T	C	NS	0.0024	HT=0
16603	NRK	X:105179263	A	G	NS	unknown	HT=0
16603	CLIC2	X:154528192	C	G	NS	unknown	HT=0
26106	OBSCN	1:228465000	C	T	NS	0.01	HT=1
26106	TBCE	1:235600671	G	C	NS	0.01	HT=2
26106	THSD7B	2:138413153	G	A	unknown	0.01	HT=0
26106	CCDC158	4:77303832	C	T	NS	0.01	HT=2
26106	CDHR2	5:175995711	C	A	NS	0.01	HT=1

26106	PHF3	6:64422859	G	C	NS	0.01	HT=2
26106	CRB2	9:126125186	C	T	NS	0.0041	HT=1
39124	KPRP	1:152732786	G	A	NS	0.0032	HT=2
39124	UNC5B	10:73053556	G	A	NS	0.01	HT=1
39124	BIN2	12:51685707	T	C	NS	0.01	HT=6
39124	THSD7B	2:138413153	G	A	unknown	0.01	HT=0
39124	UNC80	2:210858145	C	T	NS	0.01	HT=3
39124	UNC80	2:210860096	C	G	NS	0.01	HT=2
39124	CDHR2	5:175995711	C	A	NS	0.01	HT=1
39124	ABCA13	7:48450173	C	T	NS	0.01	HT=2
39124	BMP15	X:50659210	C	CTCT	NF Ins	unknown	HT=14
46154	GDF7	2:20867208	CCCGCGCCGCGCG	C	NF del	unknown	HT=5
46154	C20orf132	20:35731181	G	C	unknown	0.01	HT=13
46154	C20orf132	20:35749423	T	C	unknown	0.01	HT=13
46154	VWA7	6:31737818	C	G	NS	0.01	HT=4
46154	EHMT2	6:31864538	G	A	NS	0.01	HT=3
46154	SENP6	6:76373006	A	G	NS	0.0023	HT=1
46154	ARMCX4	X:100749037	G	A	NS	unknown	HT=0

Table G.1: Homozygous rare candidate variants in the SLE patients. NF ins is nonframeshift insertion, NF del is nonframeshift deletion

Appendix H

Putative compound heterozygote pairs in the SRNS patients

H.1 Annotated putative compound heterozygote candidate rare variant pairs in SRNS patient 0001

Gene	Gene name	Chromosome	Position first variant	Position second variant	First variant type	second variant type	product of MAF for both variants	Freq hom/ het for first variant in union	Freq hom/ het for second variant in union
MUC2	Mucin 2 Multiple homologues	11	1093600	1097847	NS	NS	0.00000001	0/ 12	second not in union
MTUS1	Mitochondrial tumour supressor gene	8	17579388	17612518	NS	NS	0.000003	first not in union	0/ 10
MUC16	Mucin 16 multiple homologues	19	9048845	9060915	NS	NS	0.000003	first not in union	0/ 13
MUC16		19	9048845	9071464	NS	NS	0.00000001	first not in union	0/ 1
MUC16		19	9060915	9071464	NS	NS	0.000003	0/ 13	0/ 1
OR5A1	Olfactory receptor 5A1	11	59210795	59211348	NS	NS	0.000004	1/ 5	second not in union
OR1I1	Olfactory receptor 1—1	19	15198363	15198720	NS	NS	0.0016	0/ 7	0/ 7
TTN	278Kb gene titin	2	179439952	179515483	NS	NS	0.0004	first not in union	0/ 2
TTN		2	179439952	179589058	NS	NS	0.0004	first not in union	0/ 2
TTN		2	179515483	179589058	NS	NS	0.0016	0/ 2	0/ 2
SPAG16	Sperm associated antigen 16	2	214160817	214354811	NS	NS	0.0001	0/ 6	0/ 6
MCM10	Minichromosome maintainance complex component 10	10	13213067	13230915	NS	NS	0.000001	first not in union	0/ 5
TEP1	Telomerase associated protein 1	14	20852653	20869148	NS	NS	0.00000001	first not in union	second not in union
TNS3	Tensin 3	7	47342790	47408732	NS	NS	0.000014	0/ 2	0/ 5
WDR27	WD Repeat-Containing Protein 27	6	170043831	170068107	NS	NS	0.00000027	0/ 3	0/ 7
HAS1	hyaluronan synthase 1	19	52217128	52222482	NS	NS	0.00000027	0/ 10	second not in union

MLL3	mixed-lineage leukemia 3/ also known as lysine K-specific methyl-transferase 2C	7	151856112	151875054	NS	NS	0.00000001	first not in union	second not in union
CARD8	caspase recruitment domain family/ member 8	19	48733756	48741734	NS	NS	0.0009	0/ 2	0/ 2
OBSCN	obscurin/ cytoskeletal calmodulin and titin-interacting RhoGEF	1	228432264	228466999	NS	NS	0.000002	1/ 10	second not in union
OBSCN		1	228432264	228548290	NS	NS	0.000002	1/ 10	second not in union
OBSCN		1	228466999	228548290	NS	NS	0.00000001	first not in union	second not in union
EFCAB12	EF-hand calcium binding domain 12	3	129130087	129137188	NS	NS	0.0012	0/ 2	0/ 41
GLTSCR2	Glioma Tumor Suppressor Candidate Region Gene 2	19	48248907	48254334	NS	NS	0.000014	0/ 15	second not in union
DNAH11	dynein/ axonemal/ heavy chain 11	7	21584693	21640728	NS	NS	0.000002	first not in union	second not in union
TATDN2	TatD DNase domain containing 2	3	10290952	10311918	NS	NS	0.0001	0/ 1	0/ 1
DNHD1	dynein heavy chain domain 1	11	6565458	6568866	NS	NS	0.00000369	0/ 2	0/ 2
CPAMD8	C3 and PZP-like/ alpha-2-macroglobulin domain containing 8	19	17013546	17036024	NS	NS	0.00000001	0/ 3	0/ 1
TTC40	tetratricopeptide repeat domain 40	10	134663840	134692953	NS	NS	0.000001	first not in union	0/ 1
MYO7B	myosin VIIB	2	128364822	128393862	NS	NS	0.00000005	first not in union	second not in union

Table H.1: Annotated putative compound heterozygote candidate rare variant pairs in SRNS patient 0001, part 1. Chr is chromosome, Prod MAF is the product of the MAF for both variants. Freq 1 is the Frequency (hom, het) for the first variant in the WGS500 union file. Freq 2 is for the second variant.

Gene	PP2 score first variant	PP2 score second var	Mouse model	Human disease association	Expression in glomeruli	Expression in murine podocyte
MUC2	no score	no score	Diarrhoea and malabsorption	Downregulated in Crohn's and UC	Not detected in glomeruli or tubules	MUC2 not differentially expressed 0.069321748 0.377898996
MTUS1	0.53547	0.991	None	None known	Unknown	MTUS1 not differentially expressed 0.001645437 0.985068966
MUC16	no score	no score	null mice are viable and normal histologically	None known	Not detected in glomeruli or tubules	MUC16 not in array
MUC16	no score	no score	as above			MUC16 not in array

MUC16	no score	no score	as above			MUC16 not in array
OR5A1	0.974	0.999	None (mouse homologue is Olfr76)	None known	Low expression in glomeruli - RNAseq only	OR5A1 not in array
OR1I1	0.999	0.978	None (mouse homologue is Olfr1357)	None known	Low expression in glomeruli - RNAseq only	OR1I1 not in array
TTN	no score	no score	Widespread embryogenesis defects	Cadiomyopathy	Low expression in glomeruli	TTN not differentially expressed 0.029442444 0.658370082
TTN	no score	no score				TTN not differentially expressed 0.029442444 0.658370082
TTN	no score	no score				TTN not differentially expressed 0.029442444 0.658370082
SPAG16	0.665	0.899	defective spermatogenesis	None known	Low expression in glomeruli	SPAG16 not differentially expressed 0.155656821 0.067014337
MCM10	0.952	0.419	reduced embryonic cell proliferation and early embryonic lethality	None known	not detected in glomeruli	MCM10 not differentially expressed 0.765271801 0.004122131
TEP1	0.996	0.963	No obvious phenotype in homozygous knock out	None known	not detected in glomeruli	TEP1 not differentially expressed 0.851269755 2.10925e-06
TNS3	0.913	0.995	Mice homozygous for a null allele exhibit one third postnatal lethality, reduced body weight, growth retardation, smaller digestive tracts with defects in villi and enterocyte differentiation, abnormal lung morphology, and thinner bones with decreased chondrocyte proliferation.	None known	Low expression in glomeruli - RNAseq only	TNS3 not differentially expressed 0.35 0.00221
WDR27	no score	no score	None	None known	not detected in glomeruli	WDR27 not differentially expressed 0.215468808 0.052395217
HAS1	0.481	0.994	null mice are viable and grossly normal	None known	Low expression in glomeruli - RNAseq only	HAS1 not differentially expressed -0.01776806 0.883158664
MLL3	0.88	0.937	Mice homozygous for a knock-out allele display partial embryonic lethality, delayed eyelid opening, postnatal growth retardation, impaired fertility in both sexes, and decreased proliferation of cultured mouse embryonic fibroblasts.	Kleefstra syndrome mental retardation, hypotonia, epilepsy.	Low expression in glomeruli - RNAseq only	MLL3 not differentially expressed 0.180101421 0.088132897
CARD8	0.79	no score	None	May play a role in cancer due to inhibition of apoptosis	High expression in glomeruli	CARD8 not in array

OBSCN	no score	no score	Mice homozygous for a knock-out allele exhibit centrally localized nuclei in muscle fibers and mild myopathy in aged mice.	None known	Medium expression in glomeruli	OBSCN not differentially expressed 0.26102074 0.012335993
OBSCN	no score	no score	as above			OBSCN not differentially expressed 0.26102074 0.012335993
OBSCN	no score	no score	as above			OBSCN not differentially expressed 0.26102074 0.012335993
EFCAB12	no score	no score	None	None known	Medium expression in glomeruli	EFCAB12 not in array
GLTSCR2	0.998	1	Null is embryonic lethal	None known	not detected in glomeruli	GLTSCR2 not differentially expressed -0.805095164 0.000304006
DNAH11	no score	no score	Approximately half of live-born homozygous mutants show situs inversus indicating that this gene is no longer properly controlling left-right asymmetry.	ciliary dyskinesia	not detected in glomeruli	DNAH11 not in array
TATDN2	0.989	0.636	None	None known	High expression in glomeruli	TATDN2 not differentially expressed -0.433548025 0.000147367
DNHD1	no score	no score	None	None known	Medium expression in glomeruli	DNHD1 not in array
CPAMD8	0.323223	0.193285	None	None known	Medium expression in glomeruli	CPAMD8 not in array
TTC40	0.141	no score	None	None known	not detected in glomeruli	TTC40 not in array
MYO7B	no score	no score	None	None known	not detected in glomeruli	MYO7B not differentially expressed 0.4 9.31e-05

Table H.2: Annotated putative compound heterozygote candidate rare variant pairs in SRNS patient 0001, part 2. Expression on murine podocyte is log2 fold change (positive if higher in podocyte) and Holm–Bonferroni adjusted p–value, cut off of >= 1.5 fold change and p value <= 0.05 used for significance (Boerries et al. 2013).

H.2 Notable putative compound heterozygote pairs in SRNS patient 0001

Caspase recruitment family member 8 (*CARD8*) is thought to play a role in inhibition of apoptosis and inflammation by binding to caspases and nuclear factor kappa–B (Fontalba et al. 2007; Razmara et al. 2002), with high expression levels reported in glomeruli. However both variants were seen twice in non–SRNS individuals in the WGS500 cohort.

Dynein axonal heavy chain 11 (*DNAH11*) contains two rare NS variants not observed in any other individual in the WGS500 project. Mutations in *DNAH11* are linked to ciliary dyskinesia and situs inversus (Bartoloni et al. 2002; Supp et al. 1997). Ciliary disease can be associated with cystic renal disease but is not reported to directly cause nephrotic syndrome, furthermore *DNAH11* is not detected in glomeruli

by IHC (Human Protein Atlas data).

TatD DNase domain containing 2 (*TATDN2*) contains two NS variants observed in only a single other individual each in the WGS500 cohort. TATDN2 has high expression levels in glomeruli according to human protein atlas, but podocytes specific expression is unknown.

Myosin VIIB (*MYO7B*) contains two rare NS variants not observed in any other individual in the WGS500 project. GO terms include 'actin binding', however the protein is not detected by IHC in glomeruli nor is it differentially expressed in murine podocytes.

H.3 Annotated putative compound heterozygote candidate rare variant pairs in SRNS patient 0002

gene	Expression in glomeruli	Chromosome	Position first variant	Position second variant	First variant type	second variant type	product of MAF for both variants	Freq hom / het for first variant in union	Freq hom / het for second variant in union
MUC16	Mucin 16 multiple homologues	19	9057468	9066390	NS	NS	0.0002	0 / 16	0 / 7
RECQL4	RecQ protein-like 4	8	145737701	145739416	NS	NS	0.00000001	0 / 3	2 / 3
WNK4	WNK lysine deficient protein kinase 4	17	40947694	40948319	NS	NS	0.000041	0 / 4	second not in union
PARD3B	Par-3 partitioning defective 3 homolog B	2	205986458	206041244	NS	NS	0.00002116	0 / 5	0 / 5
C16orf46	chromosome 16 open reading frame 46	16	81087679	81094792	NS	NS	0.0016	0 / 2	0 / 2
CIDEC	cell death-inducing DFFA-like effector c	3	9912149	9918811	NS	NS	0.000092	0 / 3	0 / 15
BIRC6	baculoviral IAP repeat containing 6	2	32724715	32819010	NS	NS	0.00000046	0 / 1	0 / 8
PPP6R1	Protein phosphatase 6 / regulatory subunit 1	19	55750843	55752997	NS	NS	0.00000037	first not in union	second not in union
TLR5	toll-like receptor 5	1	223283837	223284444	NS	NS	0.0001	0 / 2	0 / 2
ANKRD53	ankyrin repeat domain 53	2	71211883	71212336	NS	NS	0.0001	0 / 6	0 / 2
LOC728819		2	43902763	43903164	NS	NS	0.000018	0 / 4	0 / 2
NEB	nebulin	2	152468776	152520258	NS	NS	0.0001	0 / 10	0 / 4

Table H.3: Annotated putative compound heterozygote candidate rare variant pairs in SRNS patient 0002 part 1. Chr is chromosome, Prod MAF is the product of the MAF for both variants. Freq 1 is the Frequency hom /het for the first variant in the WGS500 union file. Freq 2 is for the second variant.

gene	PP2 score first variant	PP2 score second var	Mouse model	Human disease association	Expression in glomeruli	Expression in murine podocyte

MUC16	no score	no score	null mice are viable and normal histologically	None known	Not detected in glomeruli or tubules	MUC16 not in array
RECQL4	no score	no score	Homozygous loss of exons 5-8 causes embryonic death. Deletion of exon 13 causes neo- and postnatal lethality, stunted growth, skin, hair and bone defects, tissue hypoplasia and tooth dysgenesis. Mice lacking exons 9-13 show palate and limb defects, aneuploidy, poikiloderma and cancer predisposition.	Rothmund-Thomson syndrome, RAPADILINO syndrome, Baller-Gerold syndrome - overlapping syndromes featuring congenital skeletal defects, short stature, abnormal pigmentation and radial hypoplasia.	low levels in glomeruli	RECQL4 not differentially expressed 0.18292801 0.025706879
WNK4	0.999	0.998	Mice homozygous for a null allele display increased Na+, K+ and Cl- urinary excretion, alkalosis and decreased plasma Cl-, K+, Mg2+ and renin levels. Mice homozygous for a point mutation exhibit acidosis, hypertension, increased circulating potassium levels and decreased potassium excretion.	Pseudohypoaldosteronism Type IIB (AD, gain of function)	low levels in glomeruli	WNK4 not differentially expressed -0.148876799 0.259408366
PARD3B	0.906	0.45	none	None known	medium expression levels in glomeruli	PARD3B differentially expressed in murine podocyte 2.36 9.63e-08
C16orf46	0.702259	0.679	none	None known	medium expression levels in glomeruli	C16orf46 not in array
CIDEC	no score	0.977	Nullizygous mice exhibit leaness, high energy expenditure, improved glucose tolerance, altered brown adipocytes, and multilocular fat droplets with enhanced mitochondrial activity and lipolysis in white adipocytes, and may show resistance to age related and diet-induced obesity and liver steatosis.	Lipodystrophy, Familial partial type 5	low levels in glomeruli	CIDEC not differentially expressed 0.17413301 0.099644486
BIRC6	no score	no score	Homozygous mice exhibit perinatal lethality and exhibit placental defects.	None known	Low expression in glomeruli - RNAseq only	BIRC6 not differentially expressed -0.153164085 0.069397468
PPP6R1	no score	no score	none	None known	not detected in glomeruli	PPP6R1 not differentially expressed 0.3 0.00181

TLR5	0.205	0.338	Mice homozygous for disruption of this gene have a generally normal phenotype. However they fail to respond immunologically to purified flagellin and are resistant to infection with Salmonella typhimurium.	Susceptibility to Legionnaire Disease, Resistance to Systemic Lupus Erythematosus	medium expression levels in glomeruli	TLR5 not differentially expressed 0.130599402 0.051483867
ANKRD53	no score	no score	none	None known	not detected in glomeruli	ANKRD53 not differentially expressed 0.060692698 0.377125034
LOC728819	no score	no score	none	None known	unknown	LOC728819 not in array
NEB	no score	no score	Homozygous inactivation of this gene leads to stunted growth, altered sarcomere structure, reduced contractility in skeletal muscle, progressive muscle weakness, and postnatal death. Observed phenotypes may include a stiff gait, blepharoptosis, kyphosis, abnormal suckling, and reduced adiposity.	congenital autosomal recessive nemaline myopathy	Not detected in glomeruli or tubules	NEB differentially expressed in murine podocyte 1.6 4.2e-06

Table H.4: Annotated putative compound heterozygote candidate rare variant pairs in SRNS patient 0002, part 2. Expression on murine podocyte is log2 fold change (positive if higher in podocyte) and Holm–Bonferroni adjusted p–value, cut off of >= 1.5 fold change and p value <= 0.05 used for significance (Boerries et al. 2013).

H.4 Notable putative compound heterozygote pairs in SRNS patient 0002

Protein phosphatase 6, regulatory subunit 1 (*PPP6R1*) contains two rare NS variants not observed in any other individual in the WGS500 project. However the protein is not detected in glomeruli or differentially expressed in murine podocytes.

Nebulin (*NEB*) encodes an actin binding protein that is differentially expressed in mouse podocytes. However the two NS variants observed in patient 0002 were observed in some reads in 10 and 4 of the non–SRNS WGS500 patients respectively, inspection of the region in IGV revealed that all reads in this region contain multiple variations from the reference strongly suggesting that there is high polymorphism or miss-mapping here (Figure **??**). Furthermore *NEB* is a 249Kb gene encoding a giant protein so pairs of benign rare variants could occur by chance.

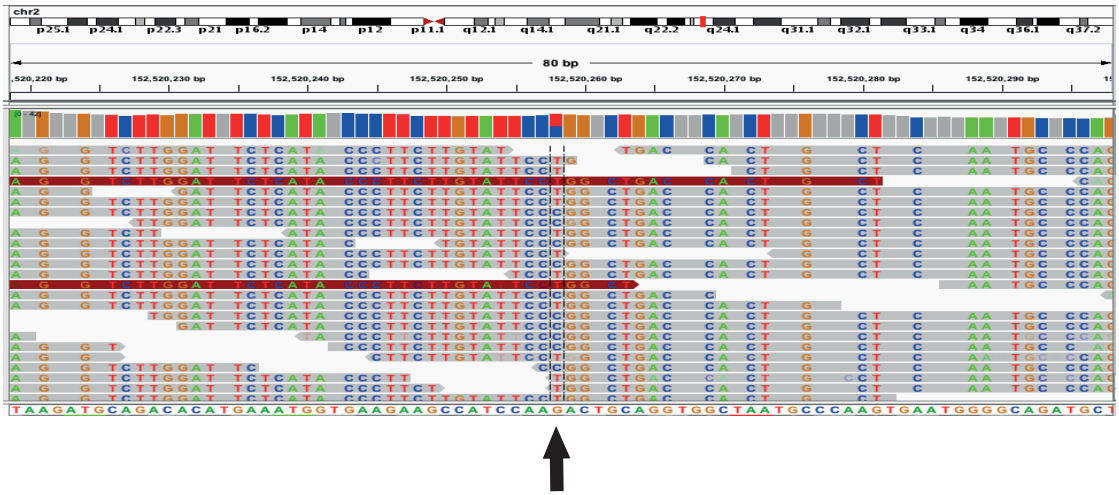


Figure H.1: Arrow indicates location of the variant call. All coloured letters within reads are locations where the base call conflicts with the reference base. The region around the first NEB variant for the putative compound heterozygous pair has a similar degree of variation.

H.5 Annotated putative compound heterozygote candidate rare variant pairs in SRNS patient 0003

Gene	Gene name	Chromosome	Position first variant	Position second variant	First variant type	second variant type	product of MAF for both variants	Freq hom /het for first variant in union	Freq hom /het for second variant in union
AGL	amylo-alpha-1 /6-glucosidase /4-alpha-glucanotransferase	1	100327088	100343310	NS	NS	0.00001	0 /4	second not in union
AGL		1	100327088	100356848	NS	NS	0.0002	0 /4	0 /3
AGL		1	100343310	100356848	NS	NS	0.000005	first not in union	0 /3
CBLC	Cbl proto-oncogene C /E3 ubiquitin protein ligase /interacts with CD2AP	19	45281211	45296806	stopgain SNV	NS	0.000064	first not in union	0 /13
STARD9	StAR-related lipid transfer START domain containing 9	15	42983017	42986471	nonframeshift deletion	NS	0.000003	first not in union	second not in union
LZTS1	leucine zipper /putative tumor suppressor 1	8	20110406	20110557	NS	stopgain SNV	0.00000018	0 /1	second not in union
ZG16B	zymogen granule protein 16B	16	2880783	2881930	nonframeshift deletion	NS	0.0001	0 /3	0 /2

LILRB3	leukocyte immunoglobulin-like receptor /subfamily B with TM and ITIM domains /member 3	19	54720967	54721272	NS	NS	0.0009	0 /2	0 /2
WDR17	WD repeat domain 17	4	177056338	177095773	NS	NS	0.0001	0 /1	0 /9
COL4A4	collagen /type IV /alpha 4	2	227872887	227912247	NS	NS	0.0001	0 /1	0 /1
COL4A4		2	227872887	227917090	NS	NS	0.0002	0 /1	0 /4
COL4A4		2	227912247	227917090	NS	NS	0.0002	0 /1	0 /4
TNK2	tyrosine kinase /non-receptor /2	3	195594494	195595374	NS	NS	0.000046	0 /2	second not in union
EFHB	EF-hand domain family /member B	3	19926053	19975497	NS	NS	0.0012	0 /7	0 /4
C9orf117	chromosome 9 open reading frame 117	9	130469384	130471764	NS	NS	0.000046	first not in union	0 /1
SCN4A	sodium channel /voltage-gated /type IV /alpha subunit /unclear how proteinuria would occur - but note example of TRCP6?	17	62019038	62050048	stopgain SNV	NS	0.00000014	first not in union	second not in union
C15orf55	NUT midline carcinoma /family member 1	15	34647875	34647972	NS	NS	0.0004	0 /3	0 /3
GPR98	G protein-coupled receptor 98	5	89948305	90006849	NS	NS	0.000123	0 /3	0 /2
ZNF77	zinc finger protein 77	19	2934625	2936535	NS	stopgain SNV	0.000015	0 /1	0 /21
DEFB119	defensin /beta 119	20	29976830	29977012	stoploss SNV	NS	0.0006	0 /6	0 /6
DNAH5	dynein /axonemal /heavy chain 5	5	13736021	13811889	NS	NS	0.00001	first not in union	0 /1
OBSCN	obscurin /cytoskeletal calmodulin and titin-interacting RhoGEF	1	228447463	228461030	NS	NS	0.000069	0 /11	0 /2
OBSCN		1	228447463	228503566	NS	NS	0.000092	0 /11	0 /3
OBSCN		1	228461030	228503566	NS	NS	0.0012	0 /2	0 /3
NLRP12	NLR family /pyrin domain containing 12	19	54313957	54314489	NS	NS	0.00001517	first not in union	second not in union
TET2	tet methylcytosine dioxygenase 2	4	106157698	106196834	NS	NS	0.00000729	0 /6	0 /6
MYH7B	myosin /heavy chain 7B /cardiac muscle /beta	20	33565890	33585205	NS	NS	0.000003	0 /2	0 /2
ZNF470	zinc finger protein 470	19	57089562	57089722	NS	NS	0.000036	first not in union	0 /4
MAGEL2	melanoma antigen /family L /2	15	23889739	23892507	NS	NS	0.0004	0 /11	second not in union

MUC22	Mucin 22	6	30996105	30996210	NS	NS	0.0001	0 /5	1 /8
PNPLA7	patatin-like phospholipase domain containing 7	9	140356012	140374799	NS	NS	0.0002	first not in union	0 /1
PNPLA7		9	140356012	140414411	NS	NS	0.0001	first not in union	second not in union
PNPLA7		9	140374799	140414411	NS	NS	0.0002	0 /1	second not in union
ADAMTS3	ADAM metallopeptidase with thrombospondin type 1 motif /3	4	73149028	73178175	NS	NS	0.0001	first not in union	0 /4
GIMAP1	GTPase /IMAP family member 1	7	150417854	150434689	NS	NS	0.0001	0 /2	0 /6
CCDC33	coiled-coil domain containing 33	15	74564054	74625055	NS	NS	0.0003	0 /2	0 /1
ACSS1	acyl-CoA synthetase short-chain family member 1	20	25038553	25038713	NS	NS	0.0009	0 /5	0 /5
RUFY4	RUN and FYVE domain containing 4	2	218937131	218940396	NS	NS	0.000003	0 /1	0 /5
CMYA5	cardiomyopathy associated 5	5	79025999	79032641	NS	NS	0.000041	0 /1	0 /1
NLRC3	NLR family /CARD domain containing 3	16	3613239	3614694	NS	NS	0.0009	0 /2	0 /2
PMS1	postmeiotic segregation increased 1	2	190719499	190728923	NS	NS	0.000001	0 /3	second not in union
CCDC88C	coiled-coil domain containing 88C	14	91739873	91755457	NS	NS	0.00000864	first not in union	second not in union
ANKK1	ankyrin repeat and kinase domain containing 1	11	113258762	113270024	NS	NS	0.0004	0 /2	0 /2
ANKRD53	ankyrin repeat domain 53-same pair seen in pt 0002	2	71211883	71212336	NS	NS	0.0001	0 /6	0 /2
ARHGEF28	Rho guanine nucleotide exchange factor GEF 28	5	73048875	73205463	NS	NS	0.000003	0 /1	0 /21
NEB	nebulin	2	152383521	152524388	NS	NS	0.000014	0 /18	0 /4
LOC285889	ncRNA	7	156231691	156234076	ncRNA splicing	ncRNA splicing	0.00000001	first not in union	second not in union
CHPF2	chondroitin polymerizing factor 2	7	150932511	150935055	NS	NS	0.00000005	first not in union	second not in union
ASXL1	additional sex combs like 1	20	31024254	31024488	NS	NS	0.000002	first not in union	0 /3

Table H.5: Annotated putative compound heterozygote candidate rare variant pairs in SRNS patient 0003 part 1. Chr is chromosome, Prod MAF is the product of the MAF for both variants. Freq 1 is the Frequency hom /het for the first variant in the WGS500 union file. Freq 2 is for the second variant.

Gene	PP2 score first variant	PP2 score second var	Mouse model	Human disease association	Expression in glomeruli	Expression in murine podocyte
AGL	0.208	0.98	none	glycogen storage disease III - hepatomegaly, hypoglycemia and growth retardation.	Low levels in glomeruli	AGL not differentially expressed 0.92 1.54e-06
AGL	0.208	0.45				AGL not differentially expressed 0.92 1.54e-06
AGL	0.98	0.45				AGL not differentially expressed 0.92 1.54e-06
CBLC	0.587898	0.958	Homozygous null mice are viable, fertile, and show no abnormalities of the epithelium or other tissues.	none known	not detected in glomeruli?-expressed in kidney (Griffith et al) and can be silenced in podocytes (calco et al)	CBLC not differentially expressed - 0.019973163 0.839806348
STARD9	no score	no score	none	none known	medium expression in glomeruli	STARD9 not in array
LZTS1	0.999	0.567769	Targeted heterozygous or homozygous inactivation of this gene results in increased incidence of both spontaneous and carcinogen-induced tumors.	Oeasaphageal squamous cell carcinoma	medium expression in glomeruli	LZTS1 not differentially expressed - 0.411440114 0.001195974
ZG16B	no score	0.255	none	none known	not detected in glomeruli or tubules	ZG16B not in array
LILRB3	0.999	0.913	Mice homozygous for disruptions of this gene (mouse homologue Pirb) display abnormalities in both B and T lymphocytes.	none known	Low expression in glomeruli - RNAseq only	LILRB3 not differentially expressed - 0.061835935 0.425201995
WDR17	0.674	0.999	none	none known	not detected in glomeruli	WDR17 not differentially expressed 0.191389496 0.139175748
COL4A4	0.491888	0.403852	Mice homozygous for an ENU-induced mutation develop an early nephritic syndrome associated with uremia, proteinuria, hematuria, leukocyturia, and focal segmental glomerulosclerosis, and die prematurely of kidney failure. Some homozygotes exhibit moderate sensorineural hearing loss.	Alport Syndrome and benign familial haematuria.	Low expression in glomeruli - RNAseq only - this is known to be expressed in podocytes!	COL4A4 differentially expressed in murine podocyte 2.34 1.1e-08
COL4A4	0.491888	0.413932	see above			COL4A4 differentially expressed in murine podocyte 2.34 1.1e-08
COL4A4	0.403852	0.413932				COL4A4 differentially expressed in murine podocyte 2.34 1.1e-08
TNK2	0.74	0.486	none	? autosomal recessive infantile epilepsy	medium expression in glomeruli	TNK2 not differentially expressed 0.391379649 0.000901212

EFHB	no score	no score	none	none known	medium expression in glomeruli	EFHB not differentially expressed 0.237530672 0.029664227
C9orf117	0.455777	0.999	none	none known	medium expression in glomeruli	C9orf117 not in array
SCN4A	no score	no score	Mice heterozygous or homozygous for a knock-in allele develop myotonia, increased myofiber damage, K+-sensitive paralysis and susceptibility to delayed weakness during recovery from fatigue. Homozygotes show perinatal lethality, low survival rate, unusual hind-limb clasping and reduced body weight.	Gain of func, het mutation leads to Hyperkalemic periodic paralysis, type 2, Myasthenic syndrome, acetazolamide-responsive, Paramyotonia congenita	not detected in glomeruli, high in tubules	SCN4A not differentially expressed 0.097330234 0.235866401
C15orf55	0.912	0.991	none	none known	Low levels in glomeruli	C15orf55 not in array
GPR98	no score	no score	Homozygotes for a spontaneous and a targeted mutation exhibit high sensitivity to audiogenic seizures. Targeted mutant mice lack the ankle links that connect growing stereocilia in the developing cochlear hair cells.	Familial Febrile seizures and Usher syndrome (hearing loss and retinitis pigmentosa)	Low expression in glomeruli - RNAseq only	GPR98 not differentially expressed -0.538634813 0.002233441
ZNF77	0.999	0.555807	none	none known	not detected in glomeruli	ZNF77 not in array
DEFB119	0.387885	0.989	none	none known	not detected in glomeruli	DEFB119 not in array
DNAH5	0.854	0.893	Mice homozygous for a disruption in this gene display postnatal lethality, hydrocephalus, respiratory infections, situs inversus and ciliary immotility.	Ciliary dyskinesia/ Kartagener's syndrome	not detected in glomeruli or tubules	DNAH5 not in array
OBSCN	no score	no score	Mice homozygous for a knock-out allele exhibit centrally localized nuclei in muscle fibers and mild myopathy in aged mice.	none known	medium expression in glomeruli	OBSCN not differentially expressed 0.26102074 0.012335993
OBSCN	no score	no score	see above			OBSCN not differentially expressed 0.26102074 0.012335993
OBSCN	no score	no score	see above			OBSCN not differentially expressed 0.26102074 0.012335993

NLRP12	1	0.985	Mice homozygous for a null allele have defects in dendritic and myeloid cell migration and a decreased susceptibility to type IV hypersensitivity reactions. Mice homozygous for a second null allele display increased susceptibility to induced colitis and to chemically-induced tumors.	Familial cold autoinflamatory syndrome	not detected in glomeruli	NLRP12 not in array
TET2	0.978	no score	Mice homozygous for a gene trapped allele die shortly after birth and exhibit a loss of acidic granules in the proximal convoluted tubules of the kidneys. Mice homozygous for a conditional allele activated in hematopoeitic compartment exhibit self-renewal and myeloid transforamtion.	association with myelodysplastic syndromes	high levels in glomeruli	TET2 not differentially expressed 0.21584806 0.080759944
MYH7B	0.317132	0.457739	none	none known	Low levels in glomeruli	MYH7B not differentially expressed 0.225126792 0.021206965
ZNF470	0.963	0.942	none	none known	medium expression in glomeruli	ZNF470 not in array
MAGEL2	no score	no score	Mice heterozygous for a null allele that is inherited paternally exhibit some postnatal lethality, reduced male fertility, abnormal circadian rhythm, and hypoactivity. Mice heterozygous for another paternal knock-out allele exhibit 50% neonatal lethality associated with weak suckling activity.	Prader-Willi-like syndrome	Low expression in glomeruli - RNAseq only	MAGEL2 not differentially expressed 0.085118297 0.333790827
MUC22	no score	no score	none	none known	Low expression in glomeruli - RNAseq only	MUC22 not in array
PNPLA7	0.67531	0.785984	none	none known	medium expression in glomeruli	PNPLA7 not differentially expressed - 0.20571551 0.015541609
PNPLA7	0.67531	0.329193	see above			PNPLA7 not differentially expressed - 0.20571551 0.015541609
PNPLA7	0.785984	0.329193	see above			PNPLA7 not differentially expressed - 0.20571551 0.015541609
ADAMTS3	0.984	0.989	none	none known	Low levels in glomeruli	ADAMTS3 not differentially expressed 0.620951129 4.67128e-05

GIMAP1	0.165	no score	Mice homozygous for a null allele have defects in the development of mature B and T lymphocytes.	none known	medium expression in glomeruli	GIMAP1 not differentially expressed -1.465861207 1.89689e-07
CCDC33	0.281769	0.722081	none	none known	not detected in glomeruli	CCDC33 not differentially expressed 0.173508116 0.047210425
ACSS1	0.748	0.371988	Mice with disruptions in this gene display abnormalities in acetate metabolism. Ability to maintain body temperature under fasting conditions is reduced.	none known	not detected in glomeruli	ACSS1 not differentially expressed -2.611542037 3.06393e-08
RUFY4	no score	no score	none	none known	not detected in glomeruli	RUFY4 not differentially expressed 0.132899471 0.184756036
CMYA5	no score	no score	none	none known	Low levels in glomeruli	CMYA5 not differentially expressed 0.75 5.39e-05
NLRC3	no score	no score	none	none known	unknown	NLRC3 not differentially expressed 0.3 0.00509
PMS1	0.999	0.738	Homozygotes for a targeted null mutation exhibit a modest increase in DNA mismatch repair errors, primarily single base pair substitutions.	colorectal cancer risk	not detected in glomeruli	PMS1 not differentially expressed 0.43 0.00677
CCDC88C	no score	no score	none	AR nonsyndromic hydrocephalus	high levels in glomeruli	CCDC88C not differentially expressed -0.29756535 0.005621606
ANKK1	no score	no score	none	Reduced brain density of dopamine recpetor D2	medium expression in glomeruli	ANKK1 not differentially expressed 0.09873578 0.445685859
ANKRD53	no score	no score	none	None known	not detected in glomeruli	ANKRD53 not differentially expressed 0.060692698 0.377125034
ARHGEF28	no score	no score	none	none known	medium expression in glomeruli	ARHGEF28 not in array
NEB	no score	no score	Homozygous inactivation of this gene leads to stunted growth, altered sarcomere structure, reduced contractility in skeletal muscle, progressive muscle weakness, and postnatal death. Observed phenotypes may include a stiff gait, blepharoptosis, kyphosis, abnormal suckling, and reduced adiposity.	congenital autosomal recessive nemaline myopathy	Not detected in glomeruli or tubules	NEB differentially expressed in murine podocyte 1.6 4.2e-06
LOC285889	no score	no score	none	none known	unknown	LOC285889 not in array
CHPF2	0.992	0.233	none	none known	not detected in glomeruli	CHPF2 not differentially expressed 0.200760587 0.245700476

ASXL1	0.253078	0.595447	Disruption of this gene causes alterations in lymphocyte development in adult mice. Mice homozygous for a different knock-out allele exhibit complete lethality. Mice heterozygous for this allele exhibit eye opacity and abnormal vertebrae morphology.	Bohring-Opitz syndrome- a severe developmental and malformation disorder	Low expression in glomeruli - RNAseq only	ASXL1 not differentially expressed 0.061482693 0.48542896

Table H.6: Annotated putative compound heterozygote candidate rare variant pairs in SRNS patient 0003, part 2. Expression on murine podocyte is log2 fold change (positive if higher in podocyte) and Holm–Bonferroni adjusted p–value, cut off of >= 1.5 fold change and p value <= 0.05 used for significance (Boerries et al. 2013).

H.6 Notable putative compound heterozygote pairs in SRNS patient 0003

CARD8 contains a stop–gain and a NS variant. The protein is expressed in kidney (Griffiths et al. 2003) and interacts with CD2AP – a known SRNS gene (Gigante et al. 2009). Silencing of podocyte *CBLC* indicates it is required for ubiquitination and degradation of the receptor tyrosine kinase Ret51 (Calco et al. 2014). The stop–gain appears very rare (1000 genomes MAF 0.0032), however the heterozygous NS variant genotype was seen in 13 non–SRNS WGS500 individual genomes.

StAR–related lipid transfer domain containing 9 (*STARD9*) contains a non–frameshift deletion and a NS variant that are both unique within the WGS500 cohort. The indel appears private to patient 0003 but the NS variant has a 1000 genomes MAF of 0.03. This transmembrane kinesin protein is expressed in glomeruli.

Leucine zipper, putative tumour suppressor 1 (*LZTS1*), contains a rare (1000 genomes MAF 0.018) NS variant predicted to be damaging and an apparently private stop–gain variant. This gene is associated with squamous cell tumour development in mice and humans (Ishii et al. 1999; Vecchione et al. 2007), and although the protein was detected in glomeruli it does not appear to be up regulated in podocytes. *Lzts1* is hypermethylated (and thus epigentically silenced) in myofibroblasts of fibrotic, compared to normal, kidneys (Tampe and Zeisberg 2014).

Two NS variants were found in Tyrosine kinase non-receptor 2 (*TNK2*), one of these was observed as a heterozygous genotype in two other WGS500 individuals and the other has an equivocal PolyPhen–2 score of 0.486. *TNK2* encodes activated CDC42 kinase 1, which has GTPase inhibitor activity, and is expressed in glomeruli. However it does not appear to be differentially expressed in podocytes (Boerries et al. 2013).

Chromosome 9 open reading frame 117 (*C9ORF117*) contains two rare NS variants and the protein is expressed in glomeruli.

Sodium channel, voltage-gated, type IV, alpha subunit (*SCN4A*) contains a stop–gain and a NS variant. Both are unique within the WGS500 cohort and the stop–

gain appears private to patient 0003. The NS variant has a 1000 genomes MAF of 0.0014. Gain of function heterozygous variants in *SCN4A* cause hypokalaemic periodic paralysis or myasthenic syndromes (Matthews et al. 2009), and this has been replicated in a knock–in mouse model (Wu et al. 2011a). The sodium channel encoded by *SCN4A* is reported to be present at high levels in renal tubules but not in glomeruli (Human Protein Atlas data). The example of *TRPC6* has shown that ion channel defects can cause nephrosis, however *TRPC6* is expressed on the podocyte (Winn et al. 2005).

Dynein axonal heavy chain 5 (*DNAH5*) contains two rare NS variants, variants in (*DNAH5* cause ciliary dyskinesia (see also texttitDNAH11 in Appendix H.2). *DNAH5* protein was not detected in the kidney in Human Protein Atlas data.

Cardiomyopathy 5 (*CMYA5*) contains two NS variants, each seen in one non–SRNS WGS500 individual as a heterozygous genotype. Low levels of the protein were reported in glomeruli and it is not differentially expressed in podocytes (Boerries et al. 2013).

Coiled-coil domain containing 88c (*CCDC88C*) contains two rare NS variants not seen in any others in the WGS500 cohort. The protein encoded by *CCDC88C* is expressed at high levels within the glomeruli (Human Protein Atlas data) and has a binding domain for Dishevelled, a central protein of the Wnt signalling pathway. The Wnt/β–catenin pathway is important for renal embryogenesis (Stark et al. 1994) and has been implicated in renal fibrosis (He et al. 2009) but more relevantly, manipulation of the Wnt pathway has been shown to influence the severity of a chemically induced model of nephrotic syndrome (Dai et al. 2009).

Chondroitin polymerising factor (*CHPF2*) contains two rare NS variants not present in any other WGS500 individual. *CHPF2* was not reported to be expressed in the glomeruli, or differentially expressed in murine podocytes.

H.7 Annotated putative compound heterozygote candidate rare variant pairs in SRNS patient 0004

Gene	Gene name	Chromosome	Position first variant	Position second variant	First variant type	second variant type	product of MAF for both variants	Freq hom/ het for first variant in union	Freq hom/ het for second variant in union
MUC16	Mucin 16 multiple homologues	19	9011412	9070264	NS	NS	0.0001	0/ 11	0/ 5
LOC554223		6	29759968	29760256	NS	NS	0.0001	0/ 6	0/ 7
LOC554223		6	29759968	29760352	NS	nonframeshift deletion	0.000001	0/ 6	2/ 2
LOC554223		6	29760256	29760352	NS	nonframeshift deletion	0.000001	0/ 7	2/ 2
EFCAB4B	EF-hand calcium binding domain 4B	12	3747496	3768803	NS	NS	0.00000023	0/ 4	second not in union
ACAA1	acetyl-CoA acyltransferase 1	3	38167095	38178499	NS	NS	0.000002	0/ 14	0/ 2
SEPT1	septin 1	16	30390417	30390818	NS	NS	0.00000081	0/ 1	0/ 1

LRRC16B	leucine rich repeat containing 16B	14	24522938	24534270	NS	NS	0.0002	0/ 4	0/ 21
PTPN14	protein tyrosine phosphatase/ non-receptor type 14	1	214576241	214588037	NS	NS	0.000001	0/ 12	second not in union
ABCA10	ATP-binding cassette/ sub-family A ABC1 / member 10	17	67146149	67151207	NS	NS	0.00001517	0/ 3	0/ 3
OSMR	oncostatin M receptor	5	38921864	38933482	NS	NS	0.000041	0/ 14	0/ 3
SLC26A10	solute carrier family 26/ member 10	12	58016602	58016690	frameshift deletion	NS	0.00000009	0/ 2	0/ 1
SLC26A10		12	58016602	58016890	frameshift deletion	NS	0.00000027	0/ 2	0/ 2
SLC26A10		12	58016690	58016890	NS	NS	0.00000243	0/ 1	0/ 2
PLEC	plectin	8	144994888	144997771	NS	NS	0.000004	0/ 35	second not in union
LOC285889	ncRNA	7	156231691	156234076	ncRNA splicing	ncRNA splicing	0.00000001	first not in union	second not in union
PTPRG	protein tyrosine phosphatase/ receptor type/ G	3	62254810	62263311	NS	NS	0.00000001	first not in union	second not in union

Table H.7: Annotated putative compound heterozygote candidate rare variant pairs in SRNS patient 0004 part 1. Chr is chromosome, Prod MAF is the product of the MAF for both variants. Freq 1 is the Frequency hom /het for the first variant in the WGS500 union file. Freq 2 is for the second variant.

Gene	PP2 score first variant	PP2 score second var	Mouse model	Human disease association	Expression in glomeruli	Expression in murine podocyte
MUC16	no score	no score	null mice are viable and normal histologically	None known	Not detected in glomeruli or tubules	MUC16 not in array
LOC554223	no score	no score	none	None known	unknown	LOC554223 not in array
LOC554223	no score	no score	see above			LOC554223 not in array
LOC554223	no score	no score	see above			LOC554223 not in array
EFCAB4B	no score	0.999	none	None known	medium levels in glomeruli	EFCAB4B not differentially expressed 0.087368264 0.421285407
ACAA1	0.267	0.168	none	None known	not detected in glomeruli	ACAA1 not in array
SEPT1	0.958	0.963	none	None known	Not detected in glomeruli or tubules	SEPT1 not in array
LRRC16B	0.580324	0.49809	none	None known	not detected in glomeruli	LRRC16B not differentially expressed 0.24306673 0.027868636
PTPN14	0.969	0.551	Mice homozygous for a gene trap allele exhibit some postnatal growth retardation, decreased body weight, periorbital and limb edema, and lymphatic vessel hyperplasia.	choanal atresia and lymphedema	low levels in glomeruli	PTPN14 not differentially expressed 1.18 2.39e-07
ABCA10	0.512	0.989	none	None known	Not detected in glomeruli or tubules	ABCA10 not in array

OSMR	0.989	0.931	Mice homozygous for a knock-out allele exhibit anemia, decreased hematocrit, and reduced erythroid progenitor, erythrocyte, platelet, and megakaryocyte cells.	primary localized cutaneous amyloidosis	not detected in glomeruli	OSMR not differentially expressed - 4.856721331 3.42218e-10
SLC26A10	no score	0.988	none	None known	low levels in glomeruli	SLC26A10 not differentially expressed - 2.83307194 1.10646e-06
SLC26A10	no score	0.536	see above			SLC26A10 not differentially expressed - 2.83307194 1.10646e-06
SLC26A10	0.988	0.536	see above			SLC26A10 not differentially expressed - 2.83307194 1.10646e-06
PLEC	0.182	0.315996	Targeted mutations of this gene result in neonatal death, skin blistering, impaired myofibril integrity, reduced hemidesmosome number, and disintegration of intercalated disks in the heart. Mice lacking isoform 1 are viable with no skin blistering but leukocyte recruitment to wounds is impaired.	epidermolysis bullosa simplex with muscular dystrophy, pyloric atresia	medium levels in glomeruli- expressed highly in podocytes (Nabet ref)	PLEC not differentially expressed - 0.448684133 0.008217677
LOC285889	no score	no score	none	none known	unknown	LOC285889 not in array
PTPRG	1	0.999	Mice homozygous for a knock-out allele are overtly normal but exhibit minor behavioral changes including specific motor deficits, reduced latency to react in the tail flick test, enhanced sensory processing for acoustic stimuli, and reduced performance with cued fear conditioning.	candidate tumor suppressor gene in renal and lung carcinoma	Low expression in glomeruli - RNAseq only	PTPRG not differentially expressed - 4.088502972 2.46102e-11

Table H.8: Annotated putative compound heterozygote candidate rare variant pairs in SRNS patient 0004, part 2. Expression on murine podocyte is log2 fold change (positive if higher in podocyte) and Holm–Bonferroni adjusted p–value, cut off of >= 1.5 fold change and p value <= 0.05 used for significance (Boerries et al. 2013).

H.8 Notable putative compound heterozygote pairs in SRNS patient 0004

EF–hand calcium binding domain 4B (*EFCAB4B*) contains two NS variants; however one was observed in 4 WGS500 patients as a heterozygous genotype despite a 1000

genomes frequency of 0.0023. The protein has GTPase activity and is expressed in the glomeruli.

Septin 1 (*SEPT1*) contains two NS variants both predicted deleterious by PolyPhen–2. However the GTP binding protein encoded was not detected in glomeruli according to human protein atlas.

Protein tyrosine phosphatase receptor type G (*PTPRG*) contains two very rare NS variants not seen in any other WGS500 individuals and both predicted deleterious by PolyPhen–2. One appears unique and the second is only present in one NHLBI exome sequencing project patient (Fu et al. 2012). *PTPRG* is a candidate tumour suppressor gene in renal cancer. Variants in Protein tyrosine phosphatase receptor type O (*PTPRO*), another gene in the same protein tyrosine phosphatase family as *PTPRG*, have been identified as a cause of childhood SRNS (Ozaltin et al. 2011). PTPRO deficient mice have reduced glomerular filtration rate and hypertension, though not albuminuria (Wharram et al. 2000), however *Ptprg* knock-out mice have a behavioural phenotype but are otherwise overtly normal (Lamprianou et al. 2006).

Appendix I

Annotated putative compound heterozygote candidate rare variant pairs shared in SLE patients 26106 and 39124

Gene	Gene name	Chr	Position first variant	Position second variant	First variant type	second variant type	PRD MAF	FREQ 1	FREQ 2	PP2 score first var	PP2 score second var	Mouse model	Human disease association
DOCK6	Dedicator of cytokinesis 6	19	11325031	11326148	NS	splicing	0.0004	(0, 3)	second not in union	no polyphen score	no polyphen score	none	Adams-Oliver syndrome - cutis aplasia congenita and limb defects
HSH2D	Haematopoietic SH2 domain containing	19	16268207	16268436	splicing	NS	0.000001	first not in union	(0, 4)	no polyphen score	no polyphen score	Knock out mice have increased Il-2 and T cell responses	none
IL12RB1	Interleukin 12 receptor beta 1	19	18174731	18193060	NS	NS	0.0002	(0, 9)	(0, 2)	no polyphen score	0.999	Knock out has reduced IFNgamma responses	Susceptibility to mycobacterium and salmonella infections
ADAM28	A disintegrin and metalloprotease 28	8	24193065	24193188	NS	NS	0.0001	(0, 3)	(0, 3)	0.999	no polyphen score	none	none
OR4D5	Olfactory Receptor, Family 4, Subfamily D, Member 5	11	123810745	123810847	NS	NS	0.000027	(0, 2)	(0, 2)	0.999	0.993	none	none

DNAH7	dynein, axonemal, heavy chain 7	2	196720647	196865579	NS	NS	0.00000322	(0, 2)	(0, 2)	0.999	0.338	none	Possible association with primary ciliary dyskinesia
TEKT5	tektin 5	16	10729653	10788598	NS	NS	0.00000162	(0, 2)	(0, 2)	0.995	0.873	none	none
LAMA3	laminin, alpha 3	18	21483972	21526138	NS	NS	0.0000007	(0, 2)	(0, 2)	0.98	0.972	Junctional epithelial skin blistering	Epidermolysis bullosa
LAMB3	laminin, beta 3	1	209791809	209806061	NS	NS	0.000001	(0, 2)	(0, 2)	0.98	0.908	none	Epidermolysis bullosa
MAP4	microtubule associated protein 4	3	47960214	47969730	NS	NS	0.0000137	(0, 3)	(0, 3)	0.976	0.956	Knock outs appear normal	none
SPG11	spastic paraplegia 11	15	44864905	44878032	NS	NS	0.00000736	(0, 2)	(0, 2)	0.956	0.992	none	spastic paraplegia with thin corpus callosum
TSSK4	testis-specific serine kinase 4	14	24676498	24677343	NS	nonframeshift deletion	0.0016	(0, 4)	(0, 5)	0.882	no polyphen score	none	none
TTC37	tetratricopeptide repeat domain 37	5	94878992	94886354	NS	splicing	0.0001	(0, 7)	second not in union	0.676	no polyphen score	none	trichohepatoenteric syndrome

Table I.1:
Annotated putative compound heterozygote candidate rare variant pairs shared in SLE patients 26106 and 39124. Chr is chromosome, Prod MAF is the product of the MAF for both variants. Freq 1 is the Frequency (hom, het) for the first variant in the WGS500 union file. Freq 2 is for the second variant.

The 13 variant pairs in Appendix I include two variants in dedicator of cytokinesis 6 (*DOCK6*) include a rare splicing variant, but the clinical phenotype associated with *DOCK6* variants is a syndrome of congenital hair and skin defects and limb developmental abnormalities that does not fit within the clinical spectrum of SLE (Shaheen et al. 2011).

Haematopoietic SH2 domain containing protein (*HSH2D*) is thought to be downstream of T cell receptor and CD28 co–stimulatory pathways and influences IL–2 promoter activation (Greene 2003). T cells from *HSH2D* deficient mice have increased Il–2 and proliferation in response to anti–CD3 and anti–CD28, indicating that *HSH2D* is a negative regulator of T cells. Splenomegaly but not glomerulonephritis was observed in the mice (Perchonock et al. 2006). Defects in a gene with role in suppression of T cell activation could lead to accumulation of activated T cells and autoimmunity.

Patients with interleukin 12 receptor beta 1 (*IL12RB1*) deficiency have increased susceptibility to mycobacterial infection (Altare et al. 1998; Jong et al. 1998). A miss–sense variant resulting in lack of cell surface IL12RB1 has been shown to prevent STAT4 nuclear translocation and interferon gamma (IFN–γ) production in response to Il–12 (Altare et al. 2001). A gain of function in the IL-12 / IFN–γ pathway could potentially be a driver for SLE.

A pair of variants including a rare splicing variant in tetratricopeptide repeat domain 37 (*TTC37*) were seen. Homozygous or compound heterozygous *TTC37*

allelic variants resulting in exon skipping or premature termination are the cause of trichohepatoenteric syndrome, which involves intractable infant diarrhoea, hair fragility and liver cirrhosis or fibrosis (Hartley et al. 2010). Low immunoglobulin levels and poor responses to childhood vaccination are a feature, suggesting an immune component to the syndrome, however the overall phenotype does not seem to fit the SLE in the siblings, one has renal involvement and the other 'chilblain lupus' features.

The other variants in Appendix I are all pairs of NS variants seen in at least 2 WGS500 genomes individually. 3 variant pairs, *LAMA3*, *LAMB3* and *SPG11* are associated with human diseases that do not fit a lupus phenotype, making them less obvious candidates.

MAP4 gene–trap mice have no discernable phenotype but this experiment did not fully block mRNA transcription (Voss, Thomas, and Gruss 1998). Therefore *MAP4*, *ADAM28*, *OR4D5*, *DNAH7*, *TEKT5*, and *TSSK4* have no known human disease association and no practical mouse model.

Appendix J

Script to analyse Lander-Green algorithm output

```
# pick hom or WT regions in any individual from the LG output genotypes,
#then get snps in these regions from the VCF for each indiv and use to calculate ENU mut rate, base profile.

import sys
import numpy
import re
from collections import defaultdict
import matplotlib

matplotlib.interactive(False)
matplotlib.use("Agg")

import pylab
from matplotlib import pyplot, mpl, figure
from mpl_toolkits.axes_grid1 import ImageGrid
from mpl_toolkits.axes_grid1 import make_axes_locatable
import numpy as np
import matplotlib.font_manager as fm


prop = fm.FontProperties(size='small')

LGoutputfile = sys.argv[1]
LGoutput = open(LGoutputfile, 'r')


vcffile1 = sys.argv[2]

Chrom = re.compile('Chrom')
error = sys.argv[6]# this allows modelling of effect of adjusting regions on the ENU mut rate

#########################################################################################################

def correct_states(list_of_states, state_length):
                corrected_list_of_states = []
                for i in range(len(list_of_states)):

                        if i <= 1:
                                corrected_list_of_states.append(list_of_states[i])
                                continue
                        elif i >= len(list_of_states)-2:
                                corrected_list_of_states.append(list_of_states[i])
                                continue
                        else:

                                if state_length[i] <= 1000000 and list_of_states[i - 1] ==
                                list_of_states[i+1]:
                                        corrected_list_of_states.append(list_of_states[i-1])
                                        # this is an unlikely recombination so replace with
                                        #prev and subsequent state
                                else:
                                        corrected_list_of_states.append(list_of_states[i])
                return corrected_list_of_states

#########################################################################################################

def state_smoother(LGoutput):


        current_start_one = None
        current_start_two = None
```

```
current_start_three = None
current_end_one = None
current_end_two = None
current_end_three = None


state_length_one = {}
state_length_two = {}
state_length_three = {}
list_of_states_one = []
list_of_states_two = []
list_of_states_three = []
count_states_one = 0
count_states_two = 0
count_states_three = 0
Current_chrom = None

state_one_count = 0
state_two_count = 0
state_three_count = 0
prev_state_one = 0
prev_state_two = 0
prev_state_three = 0
smoothedLGoutput = []

for line in LGoutput:

        if Chrom.search(line):#header
                continue
        else:
                col = line.split()
                state_one = [col[4], col[5]]
                state_two = [col[6], col[7]]
                state_three = [col[8], col[9]]
                if Current_chrom == None:# first chromosome
                        Current_chrom = col[0]

                else:
                        pass
                if current_start_one == None:# beginning of new chromosome
                        current_start_one = int(col[1])
                        current_start_two = int(col[1])
                        current_start_three = int(col[1])
                        current_end_one = int(col[2])
                        current_end_two = int(col[2])
                        current_end_three = int(col[2])
                        count_states_one += 1
                        count_states_two += 1
                        count_states_three += 1
                        list_of_states_one.append(state_one)
                        list_of_states_two.append(state_two)
                        list_of_states_three.append(state_three)
                        state_length_one[count_states_one] = current_end_one - current_start_one
                        state_length_two[count_states_two] = current_end_two - current_start_two
                        state_length_three[count_states_three] = current_end_three
                        - current_start_three

                else:
                        if state_one == list_of_states_one[count_states_one-1]:# 0 indexed dict
                                current_end_one = int(col[2])
                                state_length_one[count_states_one-1] = current_end_one -
                                current_start_one
                        elif state_one != list_of_states_one[-1]:
                        # same effect - takes last entry in dict
                                current_start_one = int(col[1])
                                current_end_one = int(col[2])
                                list_of_states_one.append(state_one)#the new state
                                count_states_one += 1

                                state_length_one[count_states_one-1] = current_end_one
                                - current_start_one
                        else:
                                print 'states cannot be checked'

                        if state_two == list_of_states_two[count_states_two-1]:
                                current_end_two = int(col[2])
                                state_length_two[count_states_two-1] = current_end_two
                                - current_start_two
                        elif state_two != list_of_states_two[-1]:
                                current_start_two = int(col[1])
                                current_end_two = int(col[2])
                                list_of_states_two.append(state_two)
                                count_states_two += 1
                                state_length_two[count_states_two-1] = current_end_two
                                - current_start_two
                        else:
                                print 'states cannot be checked'

                        if state_three == list_of_states_three[count_states_three-1]:
                                current_end_three = int(col[2])
                                state_length_three[count_states_three-1] = current_end_three
                                - current_start_three
```

```
                        elif state_three != list_of_states_three[-1]:
                                current_start_three = int(col[1])
                                current_end_three = int(col[2])
                                list_of_states_three.append(state_three)
                                count_states_three += 1
                                state_length_three[count_states_three-1] = current_end_three
                                - current_start_three
                        else:
                                print 'states cannot be checked'

if len(list_of_states_one) >= 2:

        if list_of_states_one[-1] != state_one:
                list_of_states_one.append(state_one)
                count_states_one += 1
else:
        list_of_states_one.append(state_one)
        count_states_one += 1
if len(list_of_states_two) >= 2:
        if list_of_states_two[-1] != state_two:
                list_of_states_two.append(state_two)
                count_states_two += 1
else:
        list_of_states_two.append(state_two)
        count_states_two += 1
if len(list_of_states_three) >= 2:
        if list_of_states_three[-1] != state_three:

                list_of_states_three.append(state_three)
                count_states_three += 1
else:
        list_of_states_three.append(state_three)
        count_states_three += 1

corrected_list_of_states_one = correct_states(list_of_states_one, state_length_one)

corrected_list_of_states_two = correct_states(list_of_states_two, state_length_two)
corrected_list_of_states_three = correct_states(list_of_states_three, state_length_three)
LGoutput.close()
LGoutput = open(LGoutputfile, 'r')
for line in LGoutput:
        corrected_list_of_states_one
        if Chrom.search(line):
                #smoothed.write(line)
                smoothedLGoutput.append(line)
                continue
        else:
                col = line.split()



                state_one = [col[4], col[5]]
                state_two = [col[6], col[7]]
                state_three = [col[8], col[9]]

        if prev_state_one == 0:
                prev_state_one = state_one
                prev_state_two = state_two
                prev_state_three = state_three

        else:
                if prev_state_one != state_one:
                        state_one_count += 1
                        prev_state_one = state_one
                elif prev_state_one == state_one:
                        pass

                if prev_state_two != state_two:
                        state_two_count += 1
                        prev_state_two = state_two
                elif prev_state_two == state_two:
                        pass
                if prev_state_three != state_three:
                        state_three_count += 1
                        prev_state_three = state_three
                elif prev_state_three == state_three:
                        pass



        #print col[0], col[1], col[2], col[3], corrected_list_of_states_one[state_one_count][0],
        corrected_list_of_states_one[state_one_count][1],
        corrected_list_of_states_two[state_two_count][0],
        corrected_list_of_states_two[state_two_count][1],
        corrected_list_of_states_three[state_three_count][0],
        corrected_list_of_states_three[state_three_count][1]
        smoothedLGoutput.append("%s\t%s\t%s\t%s\t%s\t%s\t%s\t%s\t%s\n" % (col[0], col[1], col[2],
        col[3], corrected_list_of_states_one[state_one_count][0],
        corrected_list_of_states_one[state_one_count][1],
        corrected_list_of_states_two[state_two_count][0],
        corrected_list_of_states_two[state_two_count][1],
```

```
                corrected_list_of_states_three[state_three_count][0],
                corrected_list_of_states_three[state_three_count][1]) )
        LGoutput.close()

        return smoothedLGoutput

#############################################################################################

def get_regions(mouse, ancestor, error, smoothedLGoutput): #this runs the script to get hom IBD regions


        Regions = []
        currentStart = None
        currentEnd = None
        currentChrom = None

        for line in smoothedLGoutput:
                col = line.split()

                if 'Chrom' in col[0]:
                        continue

                else:
                        chrom = col[0]

                        start = int(col[1])

                        end = int(col[2])
                        mouseOneHapOne = col[4]
                        mouseOneHapTwo = col[5]
                        mouseTwoHapOne = col[6]
                        mouseTwoHapTwo = col[7]
                        mouseThreeHapOne = col[8]
                        mouseThreeHapTwo = col[9]
                        allHaps = []
                        if mouse == 1:
                                allHaps.append(mouseOneHapOne)
                                allHaps.append(mouseOneHapTwo)
                        elif mouse == 2:
                                allHaps.append(mouseTwoHapOne)
                                allHaps.append(mouseTwoHapTwo)
                        elif mouse == 3:
                                allHaps.append(mouseThreeHapOne)
                                allHaps.append(mouseThreeHapTwo)
                        elif mouse == 'all':
                                allHaps.append(mouseOneHapOne)
                                allHaps.append(mouseOneHapTwo)
                                allHaps.append(mouseTwoHapOne)
                                allHaps.append(mouseTwoHapTwo)
                                allHaps.append(mouseThreeHapOne)
                                allHaps.append(mouseThreeHapTwo)
                        else:
                                print 'cannot identify which mouse'


                        if currentChrom is None:#this starts first chromosome
                                currentChrom = chrom
                        elif currentChrom != chrom:# change of chromosome
                                if currentStart is not None:
                                        if ancestor == 'hom':
                                                the_error = (float(error)*(int(currentEnd)-int(currentStart)))
                                                Regions.append( (currentChrom,(int(currentStart) - the_error),
                                                (int(currentEnd) + the_error)) )
                                                # i.e if end of last chromosome was a region
                                        elif ancestor == 'wt':
                                                the_error = 0#(float(error)*(int(currentEnd)-int(currentStart)))
                                                Regions.append( (currentChrom,(int(currentStart) + the_error),
                                                (int(currentEnd) - the_error)) )
                                        else:
                                                Regions.append( (currentChrom, currentStart, currentEnd) )
                                        currentChrom = chrom
                                        currentStart = None
                                        currentEnd = None
                                else:
                                        currentChrom = chrom
                        else:
                                pass# ie if still in same chrom this does nothing

                        if currentStart is None and ancestor == 'hom':# not in a region
                                if len(set(allHaps)) == 1 and "ENU" in allHaps[0]:
                                        currentStart = start # start a region
                                        currentEnd = end

                                else:
                                        pass
                                 # not a region
                        elif currentStart is None and ancestor == 'WT':
                                if 'WT' in allHaps[0] and 'WT' in allHaps[1]:
                                        currentStart = start
                                        currentEnd = end

                                else:
```

```
                                pass

                        elif ancestor == 'hom':
                                if len(set(allHaps)) == 1 and "ENU" in allHaps[0]:
                                        currentEnd = end
                                else:
                                        the_error = (float(error)*(int(currentEnd)-int(currentStart)))
                                        Regions.append( (currentChrom,(int(currentStart) - the_error),
                                        (int(currentEnd) + the_error)) )
                                        currentStart = None
                                        currentEnd = None

                        elif ancestor == 'WT':

                                if 'WT' in allHaps[0] and 'WT' in allHaps[1]:

                                        currentEnd = end
                                else:
                                        the_error = 0#(float(error)*(int(currentEnd)-int(currentStart)))
                                        Regions.append( (currentChrom,(int(currentStart) + the_error),
                                        (int(currentEnd) - the_error)) )
                                        currentStart = None
                                        currentEnd = None

                           # Catch last one
        if currentStart is not None:
                Regions.append( (currentChrom, currentStart, currentEnd) )
                           # this returns a set of regions with locations for homIBD.

        return Regions

##########################################################################################################

def get_hetregions(smoothedLGoutput):
        """
Simple script to pull out shared, het or mixed het and hom, excluding shared fully hom, ENU regions from the
        output of the Lander-Green algorithm.
"""


# List of hom regions shared by all mice
        sharedRegions = []
        currentStart = None
        currentEnd = None
        currentChrom = None

        for index,line in enumerate(smoothedLGoutput):

                if index == 0:
                        continue
                col = line.split()
                chrom = col[0]

                start = int(col[1])
                end = int(col[2])
                mouseOneHapOne = col[4]
                mouseOneHapTwo = col[5]
                mouseTwoHapOne = col[6]
                mouseTwoHapTwo = col[7]
                mouseThreeHapOne = col[8]
                mouseThreeHapTwo = col[9]

                allHaps = []
                allHaps.append(mouseOneHapOne)
                allHaps.append(mouseOneHapTwo)
                allHaps.append(mouseTwoHapOne)
                allHaps.append(mouseTwoHapTwo)
                allHaps.append(mouseThreeHapOne)
                allHaps.append(mouseThreeHapTwo)

                hapMouseCounts = defaultdict(int)

                if mouseOneHapOne == mouseOneHapTwo:
                        hapMouseCounts[mouseOneHapOne] += 1
                else:
                        hapMouseCounts[mouseOneHapOne] += 1
                        hapMouseCounts[mouseOneHapTwo] += 1

                if mouseTwoHapOne == mouseTwoHapTwo:
                        hapMouseCounts[mouseTwoHapOne] += 1
                else:
                        hapMouseCounts[mouseTwoHapOne] += 1
                        hapMouseCounts[mouseTwoHapTwo] += 1

                if mouseThreeHapOne == mouseThreeHapTwo:
                        hapMouseCounts[mouseThreeHapOne] += 1
                else:
                        hapMouseCounts[mouseThreeHapOne] += 1
                        hapMouseCounts[mouseThreeHapTwo] += 1

                if currentChrom is None:
                # Make sure we deal with chromosome starts/ends properly:
```

```
                #i.e. shared regions must be confined to a single chromosome.
                        currentChrom = chrom
                elif currentChrom != chrom:# Chromosome has changed
                        if currentStart is not None:
                                sharedRegions.append( (currentChrom,currentStart, currentEnd) )
                                currentChrom = chrom
                                currentStart = None
                                currentEnd = None
                        else:
                                currentChrom = chrom
                else:
                        pass

                if currentStart is None:
                        if (hapMouseCounts["ENU1"] == 3 or hapMouseCounts["ENU2"] == 3) and len(set(allHaps))
                         != 1:

                                currentStart = start
                                currentEnd = end

                        else:

                                pass

                else:
                        if (hapMouseCounts["ENU1"] == 3 or hapMouseCounts["ENU2"] == 3) and len(set(allHaps))
                        != 1:
                                #print "Extending region", hapMouseCounts, allHaps
                                currentEnd = end
                        else:
                                #print "Appending", currentChrom,currentStart,currentEnd
                                sharedRegions.append( (currentChrom,currentStart, currentEnd) )
                                currentStart = None
                                currentEnd = None

        # Catch last one
        if currentStart is not None:
                sharedRegions.append((currentChrom,currentStart, currentEnd))

        # Output to file
        #for chrom1,start1,end1 in sharedRegions:
         #    outFile.write("%s\t%s\t%s\n" %(chrom1,start1,end1))



        return sharedRegions


##############################################################################################################

def check_mutrate(mouse, vcf, error, smoothedLGoutput):

        homregions = get_regions(mouse, 'hom', error, smoothedLGoutput)

        wtregions = get_regions(mouse, 'WT', error, smoothedLGoutput)


        #homout = open(sys.argv[mouse+2], 'w')# writes out the hom regions for this mouse to file



        homsize = 0
        wtsize = 0
        homcount = 0
        wtcount = 0
        hombasemat = numpy.zeros((4,4), int)
        wtbasemat = numpy.zeros((4,4), int)
        base = 'ATCG'
        for chrom ,start, end in homregions:

                homsize += end - start

        for chrom, start, end in wtregions:

                wtsize += end - start
        vcf = open(vcf, 'r')
        for line in vcf:

                word = line.split()

                chromosomevcf = word[0]

                position = int(word[1])

                ref = word[3]
                sub = word[4]
                for chrom,start,end in homregions:

                        if chromosomevcf == chrom and position >= start and position <= end and
                        word[mouse + 8].split(":")[0] == "1/1":#check if snp in region

                                homcount +=1
```

```
                                hombasemat[base.index(ref),base.index(sub)] += 1
                                #homout.write(line)
                for chrom, start, end in wtregions:

                        if chromosomevcf == chrom and position >= start and position <= end and
                        word[mouse + 8].split(":")[0] == "1/1":#check if snp in region

                                wtcount +=1
                                wtbasemat[base.index(ref),base.index(sub)] += 1
        vcf.close()

        return homcount, homsize, wtcount, wtsize, hombasemat, wtbasemat

#############################################################################################

def perce(bases,total_point_variations):

        p = float(float(bases)/float(total_point_variations))

        return p

#############################################################################################
def basedata(basemat, total_point_variations):
                #order = 'ATCG'
        if total_point_variations != 0:

                ATTA = (basemat[0,1] + basemat[1,0])

                print 'AT-TA = %d, %s percent' % (ATTA, perce(ATTA, total_point_variations ))

                ATGC = (basemat[0,3] + basemat[1,2])

                print 'AT-GC = %d, %s percent '  % (ATGC, perce(ATGC, total_point_variations))

                GCAT = (basemat[3,0] + basemat[2,1])

                print 'GC-AT = %d, %s percent' % (GCAT, perce(GCAT, total_point_variations))

                GCCG = (basemat[3,2] + basemat[2,3])
                print 'GC-CG = %d, %s percent ' % (GCCG, perce(GCCG, total_point_variations))

                ATCG = (basemat[0,2] + basemat[1,3])

                print 'AT-CG is %d, %s percent' % (ATCG, perce(ATCG,total_point_variations))

                GCTA = (basemat[3,1] + basemat[2,0])
                print 'GC-TA is %d, %s percent' % (GCTA, perce(GCTA, total_point_variations))

                print perce(ATTA, total_point_variations),"\t", perce(ATGC, total_point_variations),
                "\t", perce(GCAT, total_point_variations),"\t", perce(GCCG, total_point_variations),
                "\t", perce(ATCG, total_point_variations),"\t", perce(GCTA, total_point_variations)
                transitions = ( ATGC + GCAT )
        else:
                print 'total number of variants is 0'
        if total_point_variations != 0 and total_point_variations-transitions != 0:
                trtr = float(transitions)/ float(total_point_variations-transitions)
                print 'trtr is %s' % (trtr)
                print 'total variants is %s' % (total_point_variations)
                print 'AT sites mutated is %s' % (ATTA + ATGC + ATCG)
        else:
                print 'cannot calculate tstv ratio'


#############################################################################################
def filtervcfbyregions(vcffile, regions, outfile):
        vcf = open(vcffile, 'r')

        out = open(outfile, 'w')
        snpcount = 0
        size = 0
        snpsinIBD = []
        for chromosome, start, end in regions:


                if Chrom.search(chromosome):
                        continue
                else:



                        size += (end - start)

        for line in vcf:

                col = line.split()
                position = int(col[1])
                chrom = col[0]
                for chromosome, start, end in regions:

                        if Chrom.search(chromosome):

                                continue
```

```
                        else:

                                if chrom == chromosome and position >= start and position <= end:
                                        snpcount += 1
                                        snpsinIBD.append(line)
                                        out.write(line)
                                else:
                                        continue


        ENUrate = float(snpcount) /float(size)
        print 'size of region is %s' % (size)
        print ENUrate
        out.close()
        return snpsinIBD

#################################################################################################
#################################################################################################

def getGenotypeFromHaps(hap1, hap2):
    """
    Convert pairs of haplotypes to genotypes: ENUhet, ENUhom, WT.
    """
    if hap1 == "WT1" and hap2 == "WT1":
        return "WT1"
    elif hap1 == "WT2" and hap2 == "WT2":
        return "WT2"
    elif hap1 == "ENU1" and hap2 == "ENU1":
        return "ENU1hom"
    elif hap1 == "ENU2" and hap2 == "ENU2":
        return "ENU2hom"
    elif hap1 == "ENU1" and hap2[:-1] == "WT" or hap2 == "ENU1" and hap1[:-1] == "WT":
        return "ENU1het"
    elif hap1 == "ENU2" and hap2[:-1] == "WT" or hap2 == "ENU2" and hap1[:-1] == "WT":
        return "ENU2het"
    elif hap1 == "ENU1" and hap2 == "ENU2" or hap2 == "ENU1" and hap1 == "ENU2" :
        return "ENU1ENU2"
    elif hap1 == "WT1" and hap2 == "WT2" or hap2 == "WT1" and hap1 == "WT2" :
        return "WT1WT2"
    else:
        print 'cannot understand this genotype %s%s' % (hap1, hap2)

#################################################################################################

def getLGDataFromLGoutput(smoothedLGoutput):
        windows = []
        for index,line in enumerate(smoothedLGoutput):
                if index == 0:
                        continue
                cols = line.strip().split("\t")
                chrom = cols[0]
                start = int(cols[1])
                end = int(cols[2])
                posterior = cols[3]
                haps1 = cols[4:6]
                haps2 = cols[6:8]
                haps3 = cols[8:10]
                genotype1 = getGenotypeFromHaps(haps1[0], haps1[1])
                genotype2 = getGenotypeFromHaps(haps2[0], haps2[1])
                genotype3 = getGenotypeFromHaps(haps3[0], haps3[1])
                windows.append( (chrom,start,end,posterior,genotype1,genotype2,genotype3))

        return windows

#################################################################################################

def plotDataForChromosome(LGData, outputfile):

        colourDict = {}
        colourDict["ENU1het"] = '#27408B'
        colourDict["ENU1hom"] = '#B0171F'
        colourDict["WT1"] = '0.7'

        colourDict["ENU2het"] = '#27408B' #prev #3A5FCD
        colourDict["ENU2hom"] = '#B0171F' #prev crimson
        colourDict["WT2"] = '0.7'# prev 0.8
        colourDict["ENU1ENU2"] = 'mediumorchid'
        colourDict["WT1WT2"] = '0.7'    #prev 0.9

        #colourDict["ENU1het"] = '#27408B'
        #colourDict["ENU1hom"] = '#B0171F'
        #colourDict["WT1"] = '0.7'

        #colourDict["ENU2het"] = '#3A5FCD' #prev
        #colourDict["ENU2hom"] = 'crimson' #prev crimson
        #colourDict["WT2"] = '0.8'# prev 0.8
        #colourDict["ENU1ENU2"] = 'mediumorchid'
        #colourDict["WT1WT2"] = '0.9'
        #grid = ImageGrid(F, (1,63,1),nrows_ncols = (1, 63), axes_pad = 0.5)
        fig = pyplot.figure()
        #fig.suptitle("Haplotype blocks for ENU mice. Green is het ENU/WT, blue is hom ENU, grey is hom WT")
```

```
        fig.subplots_adjust(wspace=0.5)
        fig.set_label('Co-ordinates')
        #c = 0
        for i in range(1,20):
                print i
                LGData_this_chrom = [x for x in LGData if x[0] == 'chr%s' %(i)]
                print LGData_this_chrom[-1]
                # Make a figure and axes with dimensions as desired.
                #LGPlotOne = grid[i+c]
                #LGPlotTwo = grid[i+1+c]
                #LGPlotThree = grid[i+2+c]

                if i == 1:
                        ax1 = fig.add_subplot(1,21,1, frameon=False, xticks=[], xticklabels=[])
                        LGPlotOne = ax1
                        pylab.xlabel('1')
                        pylab.ylabel('Chromosomal Position in Megabases')
                        pylab.setp(LGPlotOne.get_yticklabels(), visible=False)
                        #LGPlotTwo = pylab.subplot(1,63,(i+1+c))
                        #LGPlotThree = pylab.subplot(1,63,(i+2+c))
                else:
                        LGPlotOne = fig.add_subplot(1,21,i, sharey=ax1, frameon=False, yticks=[], xticks=[],
                         xticklabels=[])
                        pylab.xlabel('%s'% (i))
                        pylab.setp(LGPlotOne.get_yticklabels(), visible=False)
                        #LGPlotTwo = pylab.subplot(1,63,(i+1+c), sharey=ax1,
                        #frameon=False, xticks=[], xticklabels=[])
                        #LGPlotThree = pylab.subplot(1,63,(i+2+c), sharey=ax1,
                        #frameon=False, xticks=[], xticklabels=[])
                #c += 2
                windowcount = 0
                for window in LGData_this_chrom:
                        windowcount += 1
                        if windowcount <= (len(LGData_this_chrom)-1) or i != 18:

                                pylab.axhspan(window[1], window[2], xmin=0, xmax=0.33,
                                color=colourDict[window[4]] )
                                pylab.axhspan(window[1], window[2], xmin=0.33, xmax=0.66,
                                color=colourDict[window[5]])
                                pylab.axhspan(window[1], window[2], xmin=0.66, xmax=1,
                                color=colourDict[window[6]])
                        elif i == 18 and windowcount == len(LGData_this_chrom):
                                pylab.axhspan(window[1], window[2], xmin=0, xmax=0.33,
                                color=colourDict[window[4]], label=window[5])
                                pylab.axhspan(window[1], window[2], xmin=0.33, xmax=0.66,
                                color=colourDict[window[5]], label=window[5])
                                pylab.axhspan(window[1], window[2], xmin=0.66, xmax=1,
                                color=colourDict[window[6]], label=window[6])
                                pylab.legend(prop=prop)

        LGData_this_chrom = [x for x in LGData if x[0] == 'chrX']
        LGPlotOne = fig.add_subplot(1,21,20, sharey=ax1, frameon=False, yticks=[], xticks=[], xticklabels=[])
        pylab.xlabel('X')
        pylab.setp(LGPlotOne.get_yticklabels(), visible=False)
        for window in LGData_this_chrom:
                pylab.axhspan(window[1], window[2], xmin=0, xmax=0.33, color=colourDict[window[4]])
                pylab.axhspan(window[1], window[2], xmin=0.33, xmax=0.66, color=colourDict[window[5]])
                pylab.axhspan(window[1], window[2], xmin=0.66, xmax=1, color=colourDict[window[6]])


        LGData_this_chrom = [x for x in LGData if x[0] == 'chrY']
        LGPlotOne = fig.add_subplot(1,21,21, sharey=ax1, frameon=False, yticks=[], xticks=[], xticklabels=[])
        pylab.xlabel('Y')
        pylab.setp(LGPlotOne.get_yticklabels(), visible=False)
        for window in LGData_this_chrom:
                pylab.axhspan(window[1], window[2], xmin=0, xmax=0.33, color=colourDict[window[4]])
                pylab.axhspan(window[1], window[2], xmin=0.33, xmax=0.66, color=colourDict[window[5]])
                pylab.axhspan(window[1], window[2], xmin=0.66, xmax=1, color=colourDict[window[6]])



        pyplot.savefig("%s.png.svg" %(outputfile), format="svg")


##############################################################################################

def get_nonIBD_regions(LGOutput, Homdata, NonWTdata):
        nonIBD = []
        for line in LGOutput:
                IBD = 0
                cols = line.split()
                for hom in Homdata:

                        if cols[0] == hom[0] and int(cols[1]) >= int(hom[1]) and int(cols[2]) <= int(hom[2]):
                                IBD = 1


                for het in NonWTdata:


                        if cols[0] == het[0] and int(cols[1]) >= int(het[1]) and int(cols[2]) <= int(het[2]):
                                IBD = 1
```

```
                        else:
                                continue
                if IBD == 0:
                        nonIBD.append((cols[0],cols[1],cols[2]))
                else:
                        continue
        return nonIBD

#############################################################################################
def plotIBDDataForChromosome(Homdata, NonWTdata, NonIBDdata, outputfile):
        colourDict = {}
        colourDict["Hom_IBD"] = 'r'
        colourDict["NonWT_IBD"] = 'b'
        colourDict["WT"] = '0.7'


        fig = pyplot.figure()
        #fig.suptitle("IBD Haplotype blocks for ENU mice. blue is Non WT IBD, red is hom IBD,
        #grey is non IBD")
        fig.subplots_adjust(wspace=0.5)

        #c = 0
        for i in range(1,20):
                print i
                Homdata_this_chrom = [x for x in Homdata if x[0] == 'chr%s' %(i)]
                Hetdata_this_chrom = [x for x in NonWTdata if x[0] == 'chr%s' %(i)]
                NonIBDdata_this_chrom = [x for x in NonIBDdata if x[0] == 'chr%s' %(i)]

                if i == 1:
                        ax1 = fig.add_subplot(1,22,1, frameon=False, xticks=[], xticklabels=[])
                        #ax1.set_ylim(0, 200000000)
                        LGPlotOne = ax1
                        pylab.xlabel('1')
                        pylab.ylabel('Chromosomal Position in Megabases')


                else:
                        LGPlotOne = fig.add_subplot(1,22,i, sharey=ax1, frameon=False, yticks=[],
                        yticklabels=[], xticks=[], xticklabels=[])
                        pylab.xlabel('%s'% (i))
                        pylab.setp(LGPlotOne.get_yticklabels(), visible=False)


                homwindowcount = 0
                for window in Homdata_this_chrom:
                        homwindowcount += 1

                        if i == 14 and homwindowcount == 1:
                                pylab.axhspan(window[1], window[2], xmin=0, xmax=1,
                                color=colourDict["Hom_IBD"], label='Homozygous IBD')
                                pylab.legend(prop=prop)

                        else:

                                pylab.axhspan(window[1], window[2], xmin=0, xmax=1,
                                color=colourDict["Hom_IBD"])

                hetwindowcount = 0
                for window in Hetdata_this_chrom:
                        hetwindowcount += 1
                        if i == 14 and hetwindowcount == 1:
                                pylab.axhspan(window[1], window[2], xmin=0, xmax=1,
                                color=colourDict["NonWT_IBD"], label='other IBD')
                                pylab.legend(prop=prop)

                        else:

                                pylab.axhspan(window[1], window[2], xmin=0, xmax=1,
                                color=colourDict["NonWT_IBD"])


                nonwindowcount = 0
                for window in NonIBDdata_this_chrom:
                        nonwindowcount += 1
                        if i == 14 and nonwindowcount == len(NonIBDdata_this_chrom):
                                pylab.axhspan(window[1], window[2], xmin=0, xmax=1,
                                color=colourDict["WT"], label='Not shared')
                                pylab.legend(prop=prop)


                        else:
                                pylab.axhspan(window[1], window[2], xmin=0, xmax=1, color=colourDict["WT"])

        print 'X'
        Homdata_this_chrom = [x for x in Homdata if x[0] == 'chrX']
        Hetdata_this_chrom = [x for x in NonWTdata if x[0] == 'chrX']
        NonIBDdata_this_chrom = [x for x in NonIBDdata if x[0] == 'chrX']
        LGPlotOne = fig.add_subplot(1,22,20, sharey=ax1, frameon=False, yticks=[], yticklabels=[], xticks=[],
        xticklabels=[])
        pylab.xlabel('X')
        for window in Homdata_this_chrom:
```

```
                pylab.axhspan(window[1], window[2], xmin=0, xmax=1, color=colourDict["Hom_IBD"])
        for window in Hetdata_this_chrom:
                pylab.axhspan(window[1], window[2], xmin=0, xmax=1, color=colourDict["NonWT_IBD"])
        for window in NonIBDdata_this_chrom:
                pylab.axhspan(window[1], window[2], xmin=0, xmax=1, color=colourDict["WT"])

        print 'Y'
        Homdata_this_chrom = [x for x in Homdata if x[0] == 'chrY']
        Hetdata_this_chrom = [x for x in NonWTdata if x[0] == 'chrY']
        NonIBDdata_this_chrom = [x for x in NonIBDdata if x[0] == 'chrY']
        LGPlotOne = fig.add_subplot(1,22,21, sharey=ax1, frameon=False, yticks=[], yticklabels=[], xticks=[],
        xticklabels=[])
        pylab.xlabel('Y')

        for window in Homdata_this_chrom:
                pylab.axhspan(window[1], window[2], xmin=0, xmax=1, color=colourDict["Hom_IBD"])
        for window in Hetdata_this_chrom:
                pylab.axhspan(window[1], window[2], xmin=0, xmax=1, color=colourDict["NonWT_IBD"])
        for window in NonIBDdata_this_chrom:
                pylab.axhspan(window[1], window[2], xmin=0, xmax=1, color=colourDict["WT"])


        pyplot.savefig("%s.svg" %(outputfile), format="svg")

##############################################################################################
##############################################################################################

smoothedLGoutput = state_smoother(LGoutput)

IBDhomregions = get_regions('all', 'hom', error, smoothedLGoutput)
IBDhetregions = get_hetregions(smoothedLGoutput)
print IBDhetregions
snpsinIBDhom = filtervcfbyregions(vcffile1, IBDhomregions, sys.argv[3])
snpsinIBDhet = filtervcfbyregions(vcffile1, IBDhetregions, sys.argv[4])
LGData = getLGDataFromLGoutput(smoothedLGoutput)
plotDataForChromosome(LGData, sys.argv[5])
#nonIBD = get_nonIBD_regions(smoothedLGoutput, IBDhomregions, IBDhetregions)
#plotIBDDataForChromosome(IBDhomregions, IBDhetregions, nonIBD, sys.argv[7])
ENU_rates = []

for mouse in range(1,4):
        homcount, homsize, wtcount, wtsize, hombasemat, wtbasemat =
        check_mutrate(mouse, vcffile1, error, smoothedLGoutput)


        #print 'hom rate'
        #print float(homcount)/float(homsize)
        #print 'wt rate'
        #print float(wtcount)/float(wtsize)

        ENU_rate = (float(homcount)/float(homsize)) - (float(wtcount)/float(wtsize))
        print 'mouse, ENU rate, homcount, homsize, wtcount, wtsize, hombasemat'
        print mouse, "\n", ENU_rate, "\n",homcount, homsize, wtcount, wtsize,"\n", hombasemat

        ENU_rates.append(ENU_rate)
        basedata(hombasemat, homcount)
        print 'WTbasemat'
        basedata(wtbasemat, wtcount)

print 'ENU rates per Mb'
ENU_rates[0] = ENU_rates[0] * 1000000
ENU_rates[1] = ENU_rates[1] * 1000000
ENU_rates[2] = ENU_rates[2] * 1000000
print ENU_rates[0],"\t", ENU_rates[1],"\t",     ENU_rates[2]
```

Appendix K

ANKRD45 Sanger sequencing in the family of patient 17709

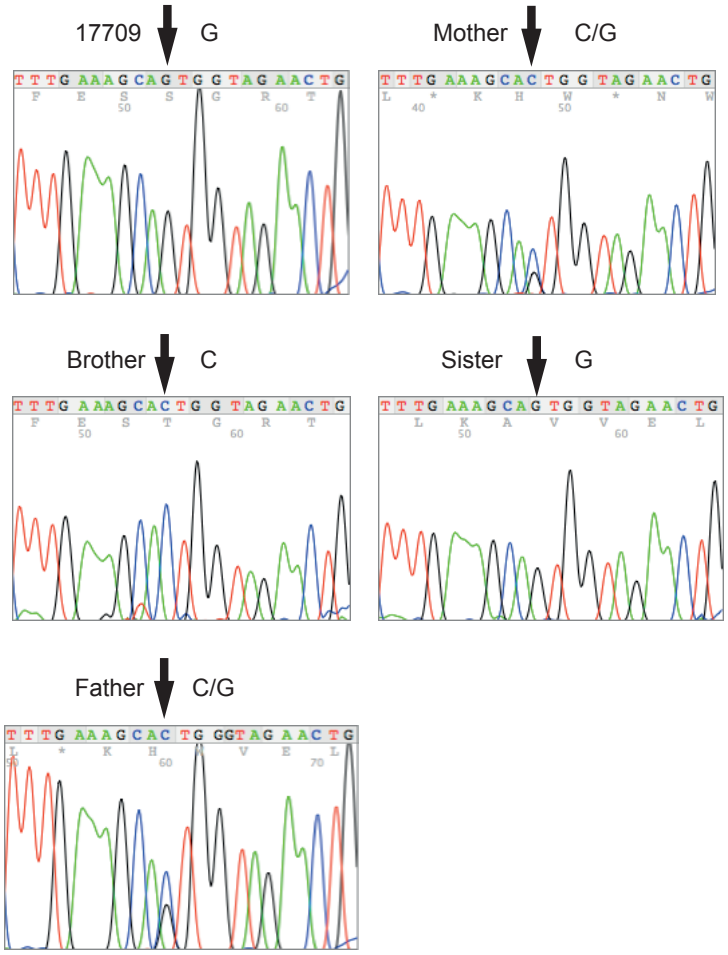


Figure K.1: ANKRD45 Sanger sequencing in the family of patient 17709. Using reverse strand primer, the reference base is C and the variant is G.