Machine Learning Methods for Next Generation Sequencing Data: Applications to MLL-AF4 Leukemia and Demographic Inference

Christopher B. Cole

Exeter College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Michaelmas 2021

Abstract

As next generation sequencing technologies continue to mature and find applications across genomics, it has become clear that the scale and scope of generated data far exceeds our ability for manual interpretation. Machine learning has shown remarkable success in finding patterns in this data and generating biologically testable hypotheses. In this thesis, I develop and apply machine learning methods which use NGS data to answer outstanding questions in population and functional genomics.

An understanding of the genetic history of global populations has been hindered by a lack of methods capable of inferring directional migration over time. I use a sequential Monte Carlo approach (a particle filter) to sample from the posterior distribution of ancestral recombination graphs and infer likely population size and migration histories from whole genome sequencing data. I apply this particle filter to global sequencing biobanks and uncover an abundance of migration from the ancestors of non-Africans into Africa between 40 and 70 thousand years ago. I show that latent directional migration has broader implications for the inference of population size in gold-standard approaches and explore this migration in the context of African pre-history.

On a cellular rather than population scale, I apply latent Dirichlet allocation to NGS-based chromatin accessibility assays in order to model shared and distinct regulatory pathways between different cell types. I demonstrate the method's utility by recovering known regulatory biology in erythroblast development. I apply this topic modelling approach to understand cis-regulatory element usage in a treatment-resistant leukemia caused by the MLL-AF4 oncoprotein. The results highlight a previously uncharacterized class of enhancer elements depleted in DOT1L-deposited H3 lysine 79 methylation and enriched for PAF1c binding.

In this thesis, I have developed and applied machine learning approaches to identify patterns in large genomics databases and answer biological questions on both population and cellular levels.

Machine Learning Methods for Next Generation Sequencing Data: Applications to MLL-AF4 Leukemia and Demographic Inference

Christopher B. Cole

Exeter College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2021

Acknowledgements

Personal

I owe a massive debt of gratitude to my supervisors Professors Gerton Lunter and Thomas Milne for their persistent support and encouragement. Gerton has been both an invaluable source of knowledge and a staunch optimist throughout the many challenges of this degree. His guidance through adversity has taught me to be a better scientist and a more mature individual. Tom became an unexpected and invaluable mentor in the second half of my DPhil. Endlessly generous with his intricate knowledge of cancer biology, Tom's unique combination of kindness, enthusiasm, and generosity helped me to discover my own passion for research. Special thanks to Professor Jim Hughes who took on an unofficial advisory role. His knowledge of both machine learning and functional genomics, as well as his sense of humour, laid the foundation for an extremely beneficial collaboration.

My fellow labmates have been crucial to any success that I've had throughout this degree. Perhaps none more so than Alastair Smith, whose creativity and mentorship extensively contributed to our work on MLL-AF4 leukemia. Joe (Sha) Zhu has always been generous with his time and knowledge, and along with Donna Henderson laid the groundwork for the SMC2 method and the applications in this thesis. The Lunter group has been my home away from home, with special thanks to Richard Brown, Ron Schwessinger, and Daniel Cooke for the endless hours of table tennis at the Wellcome Centre.

I was lucky enough to form many lasting friendships in Oxford, particularly in the Genomic Medicine and Statistics cohort with Xilin, Dan, Chris, and Ryan. The pandemic spread us across the globe but they always found time to offer help, advice, or a joke. Thank you also to my friends in Canada, who found time to suffer through my endless and changing passions in genetics. My family and soon to be in-laws in Ottawa and beyond have provided me with both a solid support structure and the opportunity to explore my interests overseas.

However, none of the above would have been in any way possible without the unending support and patience of my fiancée Kennedy (Ning) Hao. She has been both my rock and my best friend for the last decade, and I hope to be able to provide the same support to her.

Institutional

I am incredibly grateful to my funders and institutional supporters for the opportunity to pursue this research. Firstly to the Wellcome Trust and the Clarendon Scholarship for allowing me to attend the University of Oxford and removing any barriers to pursuing my interests. Secondly to the Maple Leaf Fund and the Victor Dahdaleh Foundation for their support through the Canadian Centennial Scholarship Fund. Thirdly to the Wellcome Centre for Human Genetics and the Weatherall Institute of Molecular Medicine, with particular thanks to the Centre for Computational Biology and the staff involved in maintaining their compute cluster as well as the biomedical research computing (BMRC) system.

Clarification on Contributions

The methods developed and analyses performed in this thesis represent my original work. Due to its highly collaborative nature, I use "I" and "we" pronouns interchangeably throughout. Contributions of colleagues are outlined explicitly in the "Acknowledgements" sections of each "Results" chapter.

Previously Published Work

Some of the text in this thesis has been previously published. Specifically, some expositional text in the introduction is published in Donna Henderson et al. "Demographic inference from multiple whole genomes using a particle filter for continuous Markov jump processes". In: *PLOS ONE* 16.3 (2021), pp. 1–24. URL: `https://doi.org/10.1371/journal.pone.0247647`, and the majority of Chapter 3 is published in Christopher B. Cole et al. "Ancient Admixture into Africa from the ancestors of non-Africans". In: *bioRxiv* (2020).

Abstract

As next generation sequencing technologies continue to mature and find applications across genomics, it has become clear that the scale and scope of generated data far exceeds our ability for manual interpretation. Machine learning has shown remarkable success in finding patterns in this data and generating biologically testable hypotheses. In this thesis, I develop and apply machine learning methods which use NGS data to answer outstanding questions in population and functional genomics.

An understanding of the genetic history of global populations has been hindered by a lack of methods capable of inferring directional migration over time. I use a sequential Monte Carlo approach (a particle filter) to sample from the posterior distribution of ancestral recombination graphs and infer likely population size and migration histories from whole genome sequencing data. I apply this particle filter to global sequencing biobanks and uncover an abundance of migration from the ancestors of non-Africans into Africa between 40 and 70 thousand years ago. I show that latent directional migration has broader implications for the inference of population size in gold-standard approaches and explore this migration in the context of African pre-history.

On a cellular rather than population scale, I apply latent Dirichlet allocation to NGS-based chromatin accessibility assays in order to model shared and distinct regulatory pathways between different cell types. I demonstrate the method's utility by recovering known regulatory biology in erythroblast development. I apply this topic modelling approach to understand cis-regulatory element usage in a treatment-resistant leukemia caused by the MLL-AF4 oncoprotein. The results highlight a previously uncharacterized class of enhancer elements depleted in DOT1L-deposited H3 lysine 79 methylation and enriched for PAF1c binding.

In this thesis, I have developed and applied machine learning approaches to identify patterns in large genomics databases and answer biological questions on both population and cellular levels.

Contents

List of Figures

List of Tables

Glossary

ALL acute lymphoblastic leukemia. 115, 117, 119, 124, 127, 143, 147, 150

AMH anatomically modern human. 6

AML acute myeloid leukemia. 149

ARG Ancestral Recombination Graph. 7, 21, 64, 154

ATAC-seq Assay of Transposase Accessible Chromatin sequencing. 2, 14, 75, 76, 81, 85, 112, 115

BCP B-cell precursor. 119, 120, 121, 123, 125, 127, 130, 138, 139, 141, 144, 146, 148

BLDA bulk latent dirichlet allocation. 21, 22, 158, 160

CAHG Central African Hunter Gatherer. 28, 157

ChIP-seq Chromatin immunoprecipitation followed by sequencing. 2, 118, 133, 134

CRE cis-regulatory element. 16, 74, 133, 142, 143

CwR coalescent with recombination. 7, 9, 155

DNA deoxyribonucleic acid. 2, 74

DOT1L disruptor of telomeric silencing 1-like. 18

EBV Epstein-Barr Virus. 89

HGDP Human Genome Diversity Panel. 8, 13, 21

HMM hidden Markov model. 9, 10, 20

HSC hematopoietic stem cell. 75, 77, 112, 116, 140

IBD Identity by Descent. 25

IICR inverse instantaneous coalescence rate. 9

IMF Integrated Migration Fraction. 32

InDel insertion or deletion. 5

kya kiloanni (thousands of years) before present. 56

LD linkage disequilibrium. 7

LDA Latent Dirichlet Allocation. 20, 80, 81, 86, 87, 112, 158, 160

lncRNA long non-coding RNA. 133, 152

LSC leukemia stem cell. 145

MCMC Markov chain Monte Carlo. 10, 83

MLL mixed lineage leukemia gene. 18, 115

MLL-FP MLL fusion protein. 115, 117

MRCA most recent common ancestor. 6

MSMC Multiply Sequentially Markovian Coalescent. A program developed in [3].. 10, 13

mtDNA Mitochondrial DNA. 24, 156

N_e Effective Population Size. 51

NGS Next Generation Sequencing. 24, 25, 28, 154

OHE One-hot Encoding. 96

OoA Out of Africa. 156, 157

PCR polymerase chain reaction. 7

PSMC pairwise sequentially Markovian coalescent. 9

PWM Position weight matrix. 106

RNA ribonucleic acid. 3

SAHG South African Hunter Gatherer. 28

SBL sequencing by ligation. 4

SBS sequencing by synthesis. 3

scATAC-seq Single cell ATAC-seq. 80, 84, 85, 112

SCRM sequential coalescent with recombination model. 11

SFS site frequency spectrum. 10

SGDP Simons Genome Diversity Panel. 13, 21

SMC sequentially Markovian coalescent. 9

SMCSMC Sequential Monte Carlo for the Sequentially Markovian Coalescent. 13, 21, 25, 26, 28, 31, 154

SNP single nucleotide polymorphism. 7

SRS short-read sequencing. 3

TF Transcription factor. 76

TFBS Transcription factor binding site. 74, 106

TMRCA Time to Most Recent Common Ancestor. 25

TSS transcription start site. 17

WGS whole genome sequencing. 3, 5, 7, 8, 21, 24

In the beginning there was nothing, which exploded.

— Terry Pratchett, *Lords and Ladies*

1

Machine Learning and Next Generation Sequencing

Genomics has taken front and center stage recently in the search for knowledge about our physiology and our humanity. Much of this growth from the time of Mendel's peas to the current era of rapid development and global vaccination against deadly disease with messenger RNAs has come at the heels of technological developments in sequencing technologies. As we continue to explore the questions raised by the first efforts to understand what it is that makes up humanity at a molecular level, the availability of massively high-throughput sequencing has allowed for the study of our species, from its history to its pathology. Cost lowers every year, and with it, so too does the barrier to entry into genomics research. The resulting explosion in sequencing data, concurrent with a famously exponential increase in computational processing power, has changed the way that we conduct research and philosophically approach hypothesis testing. Algorithms designed to learn patterns directly from petabytes of freely available data may be used to formulate biological hypotheses. This thesis broadly aims to extend the use of machine learning for next generation sequencing data in two different ways. Firstly, by learning demographic parameters from whole genome sequencing data, and secondly by learning cellular regulatory programs from collections of accessible

chromatin experiments. We apply the first method to detect a substantial back migration in the ancient past, and the second to identify a subset of novel enhancer elements active in childhood MLL-AF4 leukemia patients. These methods both address open questions in the field, and provide novel insight into human origins and the dysregulation of functional genomics leading to cancer.

This chapter is structured as follows. I begin by introducing next generation sequencing and its development. Next, I give context to the long-standing problem of demographic inference and how the NGS revolution has set the stage for the use of large-scale machine learning algorithms. I also briefly describe the mathematical model underpinning these methods and other comparable approaches. We then pivot to discussing the application of next generation sequencing to functional genomics and the development of Assay of Transposase Accessible Chromatin sequencing (ATAC-seq), Chromatin immunoprecipitation followed by sequencing (ChIP-seq). I describe previous approaches for learning from these data applied to identifying signal enrichment and differentially active regulatory regions.

1.1 The Evolution of Next Generation Sequencing

The order of bases in the human genome simultaneously encodes the instructions for protein synthesis and all information about the genealogical history of an individual. Despite an understanding of the three-dimensional structure of deoxyribonucleic acid (DNA) in the early 1950s, the first protein-coding gene sequence was not completed until 1972 [4, 5]. By this point, a sequencing method developed by Fred Sanger and colleagues was giving rise to the first generation of sequencing technologies [6]. Sanger sequencing, and its more modern incarnation Sanger sequencing by capillary electrophoresis, remains an important technique for clinical genomics and is the gold standard for accuracy in regard to base calling [7]. This thesis is concerned with the analysis of data resulting from sequencing technologies, and not the mechanism of sequencing itself, so descriptions here will be brief. Sanger sequencing combines denatured single-stranded DNA molecules and a solution of nucleotides with small amounts of chain-terminating dideoxynucleotides to produce fragments of varying

lengths, which can then be separated by molecular weight through electrophoresis and read either manually from a slab gel or automatically through a capillary [8, 9].

In many ways, modern NGS techniques are natural extensions of the original Sanger sequencing method. The first step in an NGS workflow is library preparation [10]. This step involves gathering DNA or ribonucleic acid (RNA) from the experimental design being employed, such as whole genome sequencing (WGS), fragmenting the product to a desired length and attaching oligonucleotide adaptors to both the 3' and 5' ends of the sequence. Adaptors are attached either through ligation, where adaptors are ligated to end-repaired inserts, or by "tagmentation" where a single transposase enzyme fragments and ligates adaptors in a single step [11]. Preparing and assessing the quality of a library is a crucial step in the process of NGS.

1.1.1 Next generation sequencing methods

There are two main methods for short-read sequencing (SRS) currently applied. They include:

1. Sequencing by synthesis (SBS) is used by technologies like Illumina Novaseq 6000. After library preparation, fragmented sequences with annealed primers are allowed to hybridize to fixed anchor sequences on the surface of a glass sequencing chip. In a process called bridge amplification the reverse primer anneals to an adjacent anchor sequence and polymerase enzymes synthesize the original fragment of DNA. The bridges are broken to form single stranded sequences, and the process is repeated to form clusters of clones, each representing a single fragment of the original fractionated DNA. Reversibly-terminated fluorescently labeled nucleotides are incorporated to the solution, and the sequencing chip is imaged to determine the consensus nucleotide incorporated for each cluster of fragments. The termination domain is washed away and the process repeated for a given insert length. Additions to this protocol can include a subsequent step where partial bridge amplification followed by the synthesis reaction sequences the insert from the 5' rather than the 3' end and the incorporation of sample specific or microfluidics-based

barcodes which encode the source of the insert. Protocols for sequencing by synthesis are taken from Illumina [12].

2. Sequencing by ligation (SBL) is primarily used in SOLiD sequencing from Life Technologies and shares many similarities with sequencing by synthesis, including the preliminary fractionation, annealing of sequencing primers, and hybridization with anchors on the surface of the sequencing chip. Rather than allowing individual bases to join the single stranded insert, however, sequencing by ligation allows DNA ligase enzymes to join entire oligonucleotides, relying on the reduced efficiency of the ligase for mismatching sequences. These labelled oligos are fluorescently dyed and when imaged provide information on several of the first bases in the unknown sequence of the insert. Depending on the specific protocol, oligose are cleaved and the process repeated outwards from the anchor sequence. Details of the sequencing protocol were taken from Slatko, Gardner, and Ausubel [13].

Other sequencing methods exist for specific situations, and have been categorized into "third" and "fourth" waves depending on their capabilities [13]. A notable example is Oxford Nanopore sequencing, which is able to provide read lengths up to theoretically hundreds of kilobases; currently, the technology is suitable for resolving structural variants [14, 15]. The per-base error rate is higher than with SBS or SBL techniques, but is steadily improving year over year [16]. These new and evolving techniques have an extensive literature dedicated to their uses and applications, which are not within the scope of this brief review or this thesis, but show huge promise for future applications in population and functional genomics.

After imaging each of the clusters on the flow cells, both protocols must identify nucleotides and their associated confidence in a process known as base calling. The result is a collection of similarly sized contiguous "reads" from the unknown DNA, each annotated with its sequencing primer and optionally a barcode with additional information about the sample that it derives from (in the case of library multiplexing). The next steps in analyzing NGS data are almost entirely computational, and differ

between sequencing applications. In general, these steps aim to filter and control the resulting base calls such that errors in analysis due to sequencing artifacts are minimized. These include, but are not limited to, errors in base calling and small insertion or deletions (InDels), poor quality reads, and adaptor or sequencing primer contamination in the final data [13]. Tools such as FastQC provide an accessible interface for performing common quality control procedures on raw data from sequencing platforms, assessing and marking reads which fail per-base and per-sequence error-rate, quality metrics, GC-content, length, segment duplication, and kmer counts [17]. These marked reads can be removed along with sequencing adaptors using Trimmomatic or CutAdapt among others [18, 19]. This previously esoteric process has been greatly simplified in recent years due to large amounts of work from software developers, expanding the accessibility of NGS technologies to areas which have otherwise not been focused on the use of command line tools. The last major step before the procedure branches off to application specific analysis involves identifying the likely origin of each fragment of DNA from the human genome. The combination of thorough statistical models as well as software implementations such as the burrows-wheeler aligner (BWA) and Bowtie, among others, has made paired read mapping (notwithstanding mutations) a largely solved problem [20, 21]. These computational steps represent a minimal path to usable sequencing reads for further analysis, however each one of these steps involves critical analysis of the data and interpretation that is not the subject of this review.

From this point onwards, we separate the discussion of sequencing technologies and associated machine learning techniques into two sections relevant to the two applications discussed in this thesis.

1.2 Modern Human Origins and Whole Genome Sequencing

Recently, the decreased cost and increased availability of WGS technologies has allowed for the study of populations previously unrepresented in the genomic record [22–26]. At the same time, mathematical descriptions of how mutations arise

and spread within populations have allowed for the development of algorithms to infer influential aspects of the origins, dispersal, and interactions of anatomically modern humans (AMHs) throughout history [27]. Here I introduce the background necessary to understand how the mathematical modelling of ancestry has allowed for demographic inference from whole genomes.

1.2.1 A Model to Describe the Evolution of a Sequence

The transmission of genetic material from parent to offspring encodes a record of all ancestral events written in a string of four bases. This process can be modelled by the coalescent, a mathematical formulation of genetic ancestry that makes predictions about patterns of variation as a consequence of a series of unobserved trees along the genome of an individual called the *gene genealogy*. Introduced in a series of seminal papers by Kingman, Hudson, and Tajima in the early 1980s, the coalescent is a stochastic process describing the $n-1$ *coalescent* events where lineages from a population of size n find their common ancestor backwards in time until only a single lineage called the most recent common ancestor (MRCA) remains [28–31]. In contrast to the classical population genetics that dominated the early 20th century, coalescent theory is mostly concerned with a retrospective view of ancestry backwards from the current generation [32]. In the simplest case of the original coalescent formulation, it can be shown for reasonable assumptions that in the limit of population size N, a time to coalescence for a pair of lineages is exponentially distributed with a mean of 1. The times for each lineage to coalesce (in units of generations) are independent and exponentially distributed as

$$f_{T_i}(t_i) = \binom{i}{2} e^{-\binom{i}{2}} \quad t_i \geq 0$$

for $i = 2, \ldots, n$ and approximates the Wright-Fisher forward-in-time models in this same limit (derivations found in Wakeley [33, Chapter 3.2]). The coalescent provides an elegant framework to describe the relative likelihood of the marginal trees at each position of the genome given information about genetic variation in a population, however the standard coalescent has proved not powerful enough

for many applications. Motivated by the observation of autocorrelation in allele frequencies along the genome, also known as linkage disequilibrium, the importance of recombination in the human genome motivated Griffiths and Marjoram [34] to extend the gene genealogy to include breakpoints where the marginal trees are permitted to change due to potentially differing ancestral origin for each haplotype in a model called the coalescent with recombination (CwR). The resulting latent data structure of piece wise constant genealogical trees along the genome, broken up with ancestral recombination breakpoints, is known as the Ancestral Recombination Graph (ARG). The ARG is an explicit representation of the structure underlying empirical linkage disequilibrium (LD) and both represents a record of the ancestral process for each marginal tree and provides the scaffold against which mutational process plays out. Developing increasingly accurate methods to understand the ancestral events leading to present day population structure is fundamental to modern population genomics.

1.2.2 Whole Genome Sequencing for Population Genomics

For the majority of the twentieth century, population genetics was driven by developments in mathematical modelling rather than insights from real data. This situation changed radically with exponentially increasing data size as polymerase chain reaction (PCR) and small-scale DNA-sequencing, polymorphism data, and finally population level sequencing experiments shifted the focus of the field to be primarily driven by data [35–37]. Though analyses based on single nucleotide polymorphisms (SNPs) are still a mainstay in the field (e.g. many studies have used, adapted, and extended the formal statistical methods for the analysis of genome-wide polymorphism data presented in Patterson et al. [38] such as [39–47]), many recent efforts focus on the use of WGS to accurately survey the distribution of genetic variation across global populations and even ancient humans such as Neanderthals [26, 48–51]. WGS presents several attractive properties over polymorphism data for population level genetics [52]. Most obviously, the ability of WGS to discover not only selected and imputed variants allows for a better understanding of the distribution

of variation across the genome. An additional benefit is the ability for researchers to move beyond the standard reference panel and identify structural variation and private polymorphisms in understudied population such as South Africans [53]. Though these analyses are technically challenging, a clear understanding of the actual structure of the genome for specific populations can help to avoid unnecessary read mapping artefacts and better represent the underlying evolutionary process.

Methodologically, WGS typically follows the steps laid out above Section 1.1.1, however an important aspect to consider in regard to population genomics analyses is the phasing strategy employed. Humans are diploids, meaning that one set of chromosomes is inherited paternally and the other maternally. At its core, the NGS protocol does not phase, or delineate the nucleotide content of the two sets of chromosomes. Estimating the phase of sequence variation is a critical step in many population genetics pipelines as haploid pieces of individual genomes (haplotypes or haploblocks) are passed down from generation to generation. There are many strategies for estimating the phase of a sequence (Choi et al. [54] studied and compared eleven such methods in regard to a single well-characterized genome), however experimental phasing strategies such as those employed in a subset of the dataset presented in Bergström et al. [50] (the Human Genome Diversity Panel (HGDP)), still represent the gold standard. Bergström et al. [50] employed a linked-read sequencing technology based on microfluidics to computationally reconstruct long haplotypes and structural variants (method is adapted from [55]). As the authors note, however, long read sequencing may enable cheaper, faster, and more accurate resolution of phase in diverse populations in the future. Other studies have opted for the approach of using statistically phased genomes in their analyses despite the known bias in parameter estimates as a consequence of errors in statistical phasing (see Steinrücken et al. [56, Supplemental section 7] and Raghavan et al. [57] for a simulation study which explores this). Choi et al. [54] found that a combination of computational strategies are able to approach experimental levels of accuracy without the need for expensive re-sequencing, so it is likely that in the near future a combination of innovation in long read

sequencing and statistical phasing methods will ameliorate the systematic issues caused by inaccurate phasing in demographic inference. Phasing and its accuracy are important factors to consider when interpreting the inferred parameter values from machine learning methods which aim to infer demographic parameters and the ancestral recombination graph as a whole.

1.2.3 Using Machine Learning to Infer the Ancestral Recombination Graph from WGS Data

Stochastic simulation under the CwR model backwards in time is computationally efficient, as the per-generation coalescent and recombination rates are simply a Markovian function of the ancestral lineages present in that generation [58]. However, for reasons of computational efficiency, in many applications it is desirable to simulate trees along the genome rather than backwards in time. Under the CwR model, the marginal trees along the genome have a complex, non-Markovian relationship, with each tree depending on every tree preceding it in the sequence [59]. A desire for a computationally tractable method for simulating genealogies along the genome lead McVean and Cardin [60] to extend the work of Wiuf and Hein [59] to develop a Markovian approximation of this process called the sequentially Markovian coalescent (SMC), greatly simplifying the problem of demographic inference along the genome and leading to the first true inference algorithm for piece wise constant population size inference known as the pairwise sequentially Markovian coalescent (PSMC) [61]. The key innovation put forward by PSMC was to approximate the continuous state space for the Markov chain with a more reasonably size finite set of topologies and discretised time intervals where coalescent events may occur allowing for description with a discrete-state hidden Markov model (HMM). Inference is then performed using the Baum-Welch algorithm, a specialization of the expectation maximization (EM) algorithm for HMMs, estimating the piece wise constant ancestral coalescent rates given a fixed mutation and recombination rate. The coalescent rate in a particular era is inversely proportion to the ancestral population size at that point in time, making the inverse instantaneous coalescence

rate (IICR) (the inverse of the coalescent rate) a reasonable estimator for the effective population size (N_e) in panmictic populations [62]. In this way, the HMM functions as a machine learning model using the observed mutations to maximize the probability of generating the inferred tMRCAs given a specific set of coalescent rates discretized over time. With an explicit and tractable likelihood function to work with, modern machine learning and statistical inference methods such as MSMC (with extensions such as MSMC2 and MSMC-IM) have become gold standard approaches for studying the deep histories of global populations through their inferred N_e [63–65]. At the same time, machine learning methods are moving beyond summaries of the underlying data (such as the tMRCA, site frequency spectrum (SFS), principal component analyses, and haplotype maps) to inferring the entire underlying ARG for a group of samples. The first method to do so on a genome-wide scale was ArgWeaver, which used a Markov chain Monte Carlo (MCMC) sampler to asymptotically sample from the posterior distribution of single panmictic population ARGs conditioned on observed genetic variation [66]. An extension of this method, ArgWeaver-D is also capable of inferring ARGs from WGS data conditioned on a user-specified topology, including migration, however it requires a fixed topology from which to effectively infer branch lengths [67]. Relate uses a modified version of the N and M [68] HMM to efficiently construct a distance matrix between samples in the dataset; a tree is constructed from this non-symmetrical distance matrix and branch lengths are estimated using a coalescent-aware MCMC [69]. A different approach is taken by tsInfer, which uses a heuristic algorithm to infer ancestral haplotypes and their most likely path to the current haplotypes through an error-prone copying process [70]. These three examples represent the current gold standard machine learning algorithms to maximize the likelihood of a set of marginal trees given observations about genetic variation.

We here focus on the general problem of inferring demography from several whole-genome sequences, which is informative about demographic events in all but the most recent epochs [61, 71]. A promising approach which so far has not been applied to this problem is to use a particle filter. Particle filters have many desirable

properties [72–74], and applications to a range of problems in computational biology have started to appear [75–78]. Like MCMC methods, particle filters converge to the exact solution in the limit of infinite computational resources, are computationally efficient by focusing on realisations that are supported by the data, do not require the underlying model to be approximated, and generate explicit samples from the posterior distribution of the latent variable. Unlike MCMC, particle filters do not operate on complete realisations of the model, but construct samples sequentially, which is helpful since full genealogies over genomes are cumbersome to deal with.

To use particle filters, we use a formulation of the coalescent model in which the state is a genealogical tree at a particular genome locus, which "evolves" sequentially along the genome, rather than in evolutionary time. To avoid confusion, in this context "time" by itself refers to the variable along which the model evolves, while evolutionary (coalescent, recombination) time refers to an actual time in the past on a genealogical tree. For this purpose, we use the sequential coalescent with recombination model (SCRM) (Section 1.2.4) [79].

Originally, particle filters were introduced for models with discrete time evolution and with either discrete or continuous state variables [74, 80]. In our model, the latent variable is a piece wise constant sequence of genealogical trees along the genome, with trees changing only after recombination events that, in mammals, occur once every several hundred nucleotides. The observations of the model are polymorphisms, which are similarly sparse. Realizations of the discrete-time model of this process (where "time" is the genome locus) are therefore stationary (remain in the same state) and silent (do not produce an observation) at most transitions, leading to inefficient algorithms. Instead, it seems natural to model the system as a Markov jump process (or purely discontinuous Markov process, [81]), a continuous-time stochastic process with as realisations piece wise constant functions $x : [1, L] \mapsto \mathbb{T}$, where $\mathbb{T}$ is the state space of the Markov process (the space of genealogical trees over a given number of genomes) and L the length over which observations are made (here the genome size).

1.2.4 The Sequential Coalescent with Recombination Model

The first model of the CwR process that evolves sequentially rather than in the evolutionary time direction was introduced by Wiuf and Hein [59], opening up the possibility of inference over very long sequences. Like Griffiths' process, the Wiuf-Hein algorithm operates on an ARG-like graph, but it is more efficient as it does not include many of the non-observable recombination events included in Griffiths' process. The Sequential Coalescent with Recombination Model (SCRM) [79] further improved efficiency by modifying Wiuf and Hein's algorithm to operate on a local genealogy rather than an ARG-like structure. Besides the "local" tree over the observed samples, this genealogy includes branches to non-contemporaneous tips that correspond to recombination events encountered earlier in the sequence. Recombinations on these "non-local" branches can be postponed until they affect observed sequences, and can sometimes be ignored altogether, leading to further efficiency gains while the resulting sample still follows the exact CwR process. An even more efficient but approximate algorithm is obtained by culling some non-local branches. In the extreme case of culling *all* non-local branches the SCRM approximation is equivalent to the SMC' model [82, 83]. With a suitable definition of "current state" (i.e., the local tree including all non-local branches) these are all Markov processes, and can all be used in the Markov jump particle filter; here we use the SCRM model with tunable accuracy as implemented in [79].

The state space $\mathbb{T}$ of the Markov process is the set of all possible genealogies at a given locus. The probability measure of a complete realisation x can be written as

$$\pi_x(x) = \exp\left\{-\int B(x_s)\rho(s)\mathrm{d}s\right\}\left[\prod_{j=1}^{|x|}\exp\left\{-\int_{\nu_j}^{\tau_j} b_u(x_{s_j})C(u)\mathrm{d}u\right\}\rho(s_j)C(\tau_j)\right](\mathrm{d}s)^{|x|}(\mathrm{d}u)^{2|x|}. \tag{1.1}$$

Here x is the sequence of genealogies along the genome; $|x|$ is the number of recombinations that occurred on x; $b_u(x_s)$ is the number of branches in the genealogy at locus s at evolutionary time u; $B(x_s) = \int_{u=0}^{\mathrm{root}(x_s)} b_u(x_s)\mathrm{d}u$ is the total branch length of x_s; $\rho(s)$ is the recombination rate per nucleotide and per generation at locus s, so that $\rho(s)B(x_s)$ is the exit rate of the Markov process in state x_s; (s_j, ν_j)

is the locus and recombination time of the jth recombination event; $\tau_j > \nu_j$ is the coalescence time of the corresponding coalescence event; and $C(u) = 1/2N_e(u)$ is the coalescence rate in generation u. process") for more details. The distribution $\pi_x(x)$ has a density with respect to the Lebesgue measure $(\mathrm{d}s)^{|x|}(\mathrm{d}u)^{2|x|}$, because each of the $|x|$ recombination events is associated with a sequence locus, a recombination time, and a coalescent time.

Mutations follow a Poisson process whose rate at s depends on the state x_s via $\mu(s)B(x_s)$ where $\mu(s)$ is the mutation rate at s per nucleotide and per generation. Mutations are not observed directly, but their descendants are; a complete observation is represented by a set $y = \{(s_j, A_j)\}_{j=1,\ldots,|y|} \in \mathcal{Y}$ where $s_j \in [1, L)$ is the locus of mutation j, and $A_j \in \{0,1\}^S$ are the wildtype (0) and alternative (1) alleles observed in the S samples. The conditional probability measure of the observations y given a realisation x is

$$\pi(y|X = x) = \frac{1}{|y|!} \exp\left\{-\int B(x_s)\mu(s)\mathrm{d}s\right\} \left[\prod_{j=1}^{|y|} P(A_j|x_{s_j}, \mu(s_j))\right] (\mathrm{d}s)^{|y|} \quad (1.2)$$

where $P(A|x_s, \mu)$ is the probability of observing the allelic pattern A given a genealogy x_s and a mutation rate μ per nucleotide and per generation; this probability is calculated using Felsenstein's peeling algorithm [84].

1.2.5 Directional Migration Inference using a Particle Filter

Full details of the formulations necessary for the implementation of a particle filter for demographic inference may be found in [1]. The focus of this thesis, and an aspect of this work which has been previously unexplored in [1] is the use of this particle filter for inferring directional migration. In this thesis, we characterize the use of Sequential Monte Carlo for the Sequentially Markovian Coalescent (SMCSMC) for this purpose in two large biobanks, the Simons Genome Diversity Panel (SGDP) and HGDP [48, 50]. We discuss the inferences in the context of African pre-history and make explicit comparisons to the MSMC model [3].

1.3 Functional Genomics and Machine Learning

Another aspect of the field impacted by advancements in NGS technologies is functional genomics. Concerned with deciphering the genomic code by which genes and their networks are regulated, NGS techniques have allowed for the efficient profiling of transcription and epigenetic processes with unprecedented resolution [85–89]. As research groups and consortia are increasingly sharing their data publicly using tools such as the Gene Expression Omnibus (`https://www.ncbi.nlm.nih.gov/geo/`) and the Sequence Read Archive (`https://www.ncbi.nlm.nih.gov/sra`), massive amounts of data on many cellular systems are now freely available. Tools are needed to simplify and reduce the dimensionality of these datasets so that expert knowledge can be used to understand the biological processes underlying them. This section introduces the fundamental NGS techniques used in functional genomics (as relevant for this thesis) and how machine learning is being used to draw biological insight from them.

1.3.1 NGS-based Functional Assays

NGS techniques have been adopted to study many facets of a cell's genome, usually by strategically fragmenting the sequence to predominantly represent regions of interest. These techniques are frequently applied to examine the epigenome, the collection of all heritable changes to chromatin that do not affect the underlying nucleotide sequence [90]. Here I discuss two individuals techniques that detect chromatin accessibility and protein-DNA interactions. These assays have direct relevance to the analyses performed in this thesis. There exist more techniques that are not discussed here, or which are introduced in this thesis where relevant.

1.3.1.1 Assay for transposase-accessible chromatin with high throughput sequencing

In 2015 Buenrostro et al. [91] introduced ATAC-seq (Figure 1.1). The goal of the method is to assess the substantial heterogeneity in physical compaction of the genomic sequence, referred to as chromatin accessibility [92]. Chromatin is a

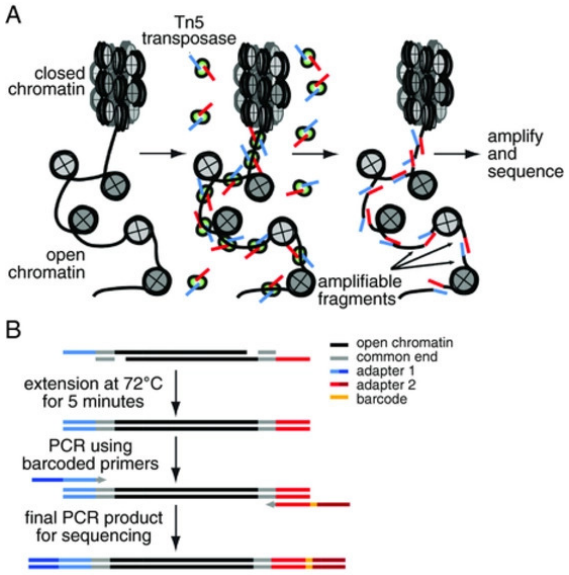


Figure 1.1: ATAC-seq protocol schematic adopted from [91]

complex of DNA and proteins used to organize the genome, and its fundamental unit is the nucleosome [93]. The nucleosome is a complex of DNA wrapped around a protein octamer of four core histones numbered as H2A, H2B, H3, and H4 [93]. These histones are frequently the target of post-translational modifications which confer additional functional information about the sequence, as described in the next section on ChIP-seq. In general, a higher density of nucleosomes along a sequence leads to more compact DNA within the nucleus [94]. The physical compaction of DNA, a necessary process due to the almost two meters of DNA packed into a five micron nucleus, has been shown to be a key factor in the regulation of transcription by either allowing or blocking transcriptional machinery physical access to the sequence [95, 96]. A more thorough exploration of chromatin accessibility as it relates to the topics explored in this thesis is presented in the introduction of Chapter 3 in Section 3.1.1.

The ATAC-seq assay is conceptually a simple adaptation of the typical NGS procedure, and as previously mentioned relies on strategic sheering of the genome (in this case, targeted amplification) followed by sequencing. Here, a hyperactive Tn5 transposase enzyme preferentially accesses the sequence not wrapped in nucleosomes, inserting sequencing adaptors that can be directly amplified. This

results in fragments of DNA whose length follows a bimodal distribution, with the majority derived from the space between nucleosomes and tending to be short, and the minority spanning a nucleosome and thus having a fragment length of at least ∼150 base pairs [97]. Since the amplified pieces of sequence predominantly appear in regions without nucleosomes, identifying "accessible" regions (for a given threshold) is an exercise in statistical detection of regions with signal enriched over the background.

The concept behind ATAC-seq is similar to previous assays for chromatin accessibility such as MNase-seq and DNAse-seq which use micrococcal nucleases, and DNAse 1 enzymes to fragment the sequence respectively [93]. FAIRE-seq, a successor to DNAse-seq introduced in Giresi et al. [98], performs a fundamentally different fragmentation by isolating DNA that is not able to be cross-linked to nucleosomes. In contrast to MNAse-seq and DNAse-seq, ATAC-seq requires far fewer experimental steps and isolated cells, allowing for efficient study of small populations of cells (as shown by Buenrostro et al. [91]) while comparisons with FAIRE-seq show substantial bias in the latter towards enhancer and intronic elements and a smaller enrichment of signal over the sequencing background [99]. This makes ATAC-seq an experimentally efficient procedure for generating high coverage chromatin accessibility tracks in low to moderate numbers of cells. DNAse-seq was adopted enthusiastically by consortia such as ENCODE Project Consortium [100] and substantial effectively-legacy data exists from this assay, however more recent efforts in the same groups (i.e. Moore et al. [101]) are focusing on ATAC-seq for this task. The analysis of ATAC-seq data therefore remains a valuable task, especially as the number of generated datasets continues to increase year over year (Yan et al. [97] demonstrates this up until 2019).

1.3.1.2 Chromatin immunoprecipitation followed by sequencing

Both the combinatorial binding of transcription factors and post-translational modification of histone proteins are key players in the complex logic of gene expression [102]. Transcription is regulated by a large number of cis-regulatory

elements (CREs) categorized by function and distance to the transcription start site (TSS). These include promoters and promoter-proximal elements, enhancers, silencers, insulators, and boundary elements for topologically associated DNA domains (reviewed in Wittkopp and Kalay [103]). Many transcription factors bind to specific patterns in DNA called "motifs", however the majority have either a non-specific binding sequence or none at all, leading to the desire for an experimental approach to determine where specific transcription factors are bound in specific cellular contexts [104]. At the same time, specific covalent modifications to histone proteins have been shown to be enriched at active enhancer and promoter elements, indicating their involvement in the recruitment of proteins for transcriptional activation. For the purposes of this thesis, only a handful of the many histone modifications are necessary for the putative identification of enhancer and promoter regions (Figure 1.2). In the following and throughout, the nomenclature for histone modifications is the histone that is modified (Histone H3 = H3), the residue that is modified (lysine residue 4 = K4), and the actual modification (me1 = mono-methylation, me3 = tri-methylation, ac = acetylation) such that, for example, mono-methylation of the fourth lysine residue on histone protein H3 would be denoted as H3K4me1 [105]. A summary of important histone modification follows.

H3K4me1 is found predominantly at enhancer elements [105]. Recent evidence suggests that H3K4me1 can also be found in promoter elements, with a bimodal pattern flanking H3K4me3 in active promoters and unimodal peak, coinciding with H3K4me3 and H3K27me3, proximal to the TSS in poised promoters [106].

H3K4me3 is typically believed to be the distinguishing mark of promoter elements [105]. Typically, the dynamic regulation of methylation between DNA methylation, H3K4me1, and H3K4me3 is able to alter to state of a region from functionally inactive to enhancer or promoter elements respectively [107]. Interestingly though, H3K4me3 occasionally is observed on putative "super enhancers", especially in some cancers, though this relationship is the subject of much debate [108].

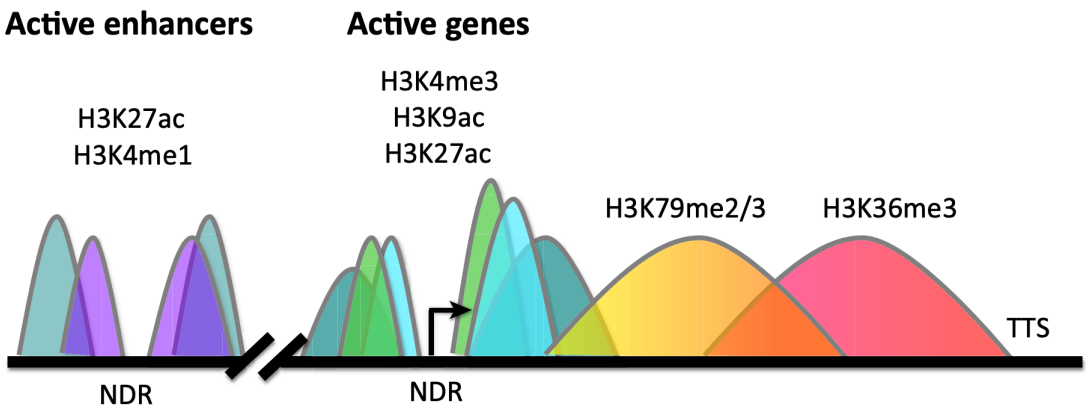


Figure 1.2: Typical eukaryotic ChIP-seq defined active enhancer, promoter, and gene body histone modifications from Gates, Foulds, and O'Malley [105]. H3K4me3 and H3K9ac typically load onto active promoter regions, while H3K79me2/3 and H3K36me3 are found on the actual gene body being transcribed. H3K27ac loads onto both active promoters and enhancers, while H3K4me1 is found predominantly on the latter. These are all generalizations, and represent typically observed patterns. NDR = nucleosome-depleted (accessible) regions, NDR = nucleosome depleted region (i.e. accessible chromatin).

H3K27ac is an activity marker found at both active enhancers and promoters [105].

H3K79me2/3 as in, either the di or tri methylation of H3K79, is uniquely deposited by the disruptor of telomeric silencing 1-like (DOT1L) protein, and is often found within the body of actively transcribed genes and is associated with transcriptional elongation [109, 110]. While not usually found at enhancer elements, recent evidence strongly suggests a role for DOT1L mediated H3K79 methylation in enhancer-promoter interactions in cancers caused by mixed lineage leukemia gene (MLL) fusion proteins, which is of interest to results later in this thesis [111].

H3K27me3 functions as a repressive mark for gene expression and is found at silencer elements [105, 112].

These marks vary across the genome depending on a large number of biological factors; ChIP-seq provides an empirical tool to measure their presence.

ChIP-seq combines chromatin immunoprecipitation (ChIP) with NGS sequencing to identify protein-DNA interactions on a genome-wide scale [113]. ChIP involves the covalent crosslinking of DNA to any associated proteins followed by the random

fragmentation of the genome; specific antibodies are used to "pull down" regions bound by a specific protein or histone modification, and the remainder are washed away [114, 115]. The remaining fragments of DNA are annealed to sequencing primers and amplified. The effect of this procedure is to enrich sequenced regions for areas of the genome bound by a specific protein. Typically, ChIP-seq experiments are paired with an "input" run, in which the pull down step is not performed, in order to assess the genomic background of enriched regions. These are typically removed from the analysis and peak calling as they represent technical artefacts and not biologically relevant DNA-protein complexes.

1.3.2 Specific Machine Learning Applications for Functional Genomics

1.3.2.1 Identifying Signal Enrichment (Peak Calling) with Machine Learning

One of the key tasks after performing ATAC-seq or ChIP-seq is to determine regions which are enriched for signal over the background [97]. Many approaches have been developed for this task (partially reviewed and benchmarked in R et al. [116]) based on statistical models typically modelling signal-to-noise ratios according to a Poisson distribution. This model appears to correctly discriminate visually enriched peak regions, though both the false positives and false negative identifications are frequently thought to be bottlenecks in analyses that rely on discretized region sets. Recently, a deep learning based peak called was proposed which appears to exceed the performance of the gold standard approach, MaCS2 [117–119]. The model relies on a wide-and-deep convolutional neural network trained on a manually curated set of 8463 peaks from various ENCODE Project Consortium [100] datasets and 8503 noise regions [117, 120].

1.3.2.2 Chromatin State Annotation

After sequencing and discretizing histone modifications into peak regions, a researcher may wish to annotate the genome of their cell type of interest with putative states (i.e. active or poised enhancer or promoter, heterochromatin, etc.). Due to

the high dimensionality of the data (a number of chromatin marks each with a genome-scale coverage track), this is a difficult task to perform manually. Ernst and Kellis [121] discovered that by modelling the chromatin state along the genome as a Markovian process, inference of putative states can be efficiently performed with standard approaches such as HMMs. ChromHMM and extensions have remained standard approaches for chromatin state annotation and the identification of putative enhancer regions across cell types of interest [122]. However, latent states require manual annotation which is often non-trivial, and in the case of niche cell types it is often difficult to generate the required ChIP-seq data for manual annotation. The alternative approach, using an imputation program such as ChromImpute leads to variable quality of input data heavily depending on the sequence specificity of the given mark [122]. Topic modelling represents an attractive alternative that relies solely on accessible chromatin to annotate shared and distinct regulatory elements, and has been successfully used for single cell ATAC-seq data [123]. ATAC-seq is comparably cheaper and easier to assay in niche cell populations, and a wide array of public reference data is available from consortia. However, a satisfying adaptation of the cisTopic Latent Dirichlet Allocation (LDA) approach has never been attempted for bulk ATAC-seq in specific populations of cell types. In order to determine shared and distinct regulatory elements in closely related cell types, topic modelling represents a viable approach which has previously been demonstrated effective in a single cell context.

1.4 Thesis Aims

The overarching goal of this thesis is to develop machine learning methods to interpret next generation sequencing data and apply them to two specific questions. The two sections above give background information on the motivation for solving the problem of inferring directional migration from whole genome sequencing data (Section 1.2.2) and identifying shared and distinct regulatory elements in closely related celltypes (Section 1.3). For this second question, we develop a topic modelling approach based on LDA for bulk ATAC-seq data. We apply this approach

to a specific form of leukemia with a poorly understood regulatory landscape and uncover a specific set of enhancer elements with novel histone modification profiles. This thesis both represent methodological advances in two different subfields and the contribution of novel results, namely the identification of an ancient back-migration and a subset of PAF1c bound enhancer elements in MLL-AF4 leukemia. A summary of the chapters follows.

Chapter 2 describes the SMCSMC method for sampling from the posterior distribution of ARGs conditioned on genomic variation from WGS data and inferring demographic parameters including population-specific effective population size and directional migration rates over time. I use simulation to demonstrate the accuracy and limitations of inference and apply the method to analyse individuals from the SGDP and HGDP. In doing so, I discover a substantial back-migration from the ancestors of non-Africans to the ancestors of present day Africans between 40 and 70 thousand years before present. The remainder of the chapter investigates the dynamics and magnitude of this migration, and explores its ramifications on human origins.

Chapter 3 introduces the bulk latent dirichlet allocation (BLDA) method for performing topic modelling on bulk ATAC-seq data. This method is a direct extension of the cisTopic LDA approach taking into account a quantitative rather than binary signal for read density at peak regions. I show that BLDA is as effective in statistically pseudobulked ATAC-seq samples as cisTopic is in single cell ATAC-seq data. Additionally, I show that LanceOTron's deep learning peak caller results in cleaner inference in a range of topic modelling applications. I use the developed approach on a curated set of sorted cell types representing the differentiation pathway of hematopoiesis and erythopoesis and show that BLDA is superior to a naive implementation of cisTopic, and additionally recovers known regulatory programs active in these cell types.

Chapter 4 applies the previously developed BLDA method to a set of MLL-AF4 driven leukemia patients and cell lines and demonstrates a unique and shared set of differentially accessible chromatin regions when compared to a specifically selected set of closely related B cell progenitors. These differentially accessible regions are hypothesized to represent unique regulatory programs that distinguish the cells. I isolate key regions of these regulatory topics and show that they are highly robust against the stochastic inference procedure. I model these cell types alongside blood cells from the ENCODE consortium and demonstrate their distinctiveness. I compare these regions against ChIP-seq for known histone modifications and show that they are marked as putative enhancers without the DOT1L signature typical of enhancers in MLL-AF4 leukemia, and additionally bound by PAF1c. I show that these regions share activity profiles with certain hematopoietic progenitor cells and elaborate on future experimental validation to elucidate their exact function in MLL-AF4 leukemia.

Chapter 5 discusses the contribution of these results to their respective areas and suggests further work.

"I kind of lost track of time..."
"For two hours?"
Elend nodded sheepishly. "There were books involved."

— Brandon Sanderson, *The Well of Ascension*

2

Ancient Admixture into Africa from the Ancestors of non-Africans

Contents

2.1 Introduction

The history of a population shapes its patterns of genetic diversity. A nuanced understanding of the historical relationships between global populations has been hindered by both the lack of availability genomic sequencing data and methodological barriers. With the recent completion of several repositories of genomic variation across the globe, and the advances presented in the previous chapter allowing for tractable inference of directional migration rates, here I aim to investigate ancient contributors to extant genomic variation using machine learning and sequence data.

Before the era of Next Generation Sequencing (NGS), an understanding of diverse populations was mostly based on combinations of variants in Y chromosome and Mitochondrial DNA (mtDNA). Genotype frequency data from microarrays has been used more recently, and will be discussed in the following section. As the Y chromosome and mtDNA do not recombine, each comprise a single genealogical tree dating back to a common ancestor. This attractive analytical property allowed geneticists in the late twentieth century to study the evolution of these two lineages in great depth, leading to an extensive catalog of variation within so-called haplogroups (see `https://isogg.org/tree/index.html`). As mtDNA is inherited maternally and Y chromosome DNA is inherited paternally, they represent unique viewpoints into sexually dimorphic anthropological processes (for a review, see Kivisild [124]). Results from haplogroup analyses have motivated many important discoveries, however caution must be taken when interpreting the results of a single tree in light of the whole of human history [125]. While these single trees are constrained by the overall demographic process, they are not representative of it due to incomplete lineage sorting. With recent access to high quality WGS data, we are able to efficiently sample many more of the marginal trees along the genome. Therefore,

for the general case of investigating population structure across the globe in light of human history, the nuclear genome must be focus of our study.

Before the advent of NGS, patterns in genomic diversity have been investigated through differences in allele frequencies between subpopulations. Metrics such as the fixation index are often used to summarize differences between subsets of the population, either in terms of their relative variances, their probability of Identity by Descent (IBD) or their Time to Most Recent Common Ancestor (TMRCA) [126]. Modern adaptations and extensions of Wright's F statistics include so-called drift statistics, which quantity the degree to which two populations share genetic drift [127]. These statistics can be built up to create intricate descriptions and statistical tests for shared drift, (f_3) admixture (D, f_4), number of introgression events (`qpWave`) and even entire global topologies of population differentiation with branch lengths (`qpAdm`) [128]. These approaches form a highly useful toolkit for studying easy to gather polymorphism data, but do not attempt to recapitulate the processes by which this variation is produced. Critically, they do not attempt to reconstruct demographic parameters over time. In order to explore this much larger parameter space, machine learning is necessary.

How a given set of individuals are related at a particular locus is represented by a phylogenetic tree. Because of past recombination events, this tree changes along the genome, resulting in a series of trees that may be represented in a data structure known as the ancestral recombination graph. [129]. All information about the genealogical history of a set of individuals can be encoded in their ARG, and having access to the ARG would help immensely when making statements of a population's demographic history. However, the ARG cannot be observed directly [130]. Recently, methods which use stochastic inference techniques have been able to directly sample from the posterior distribution of marginal genomic trees and produce representations of probable ancestral recombination graphs [129–131].

Currently, no analytical methods for inferring ARGs are known, and it seems likely that none exist. Therefore, several researchers have developed approximate statistical inference techniques, such as ArgWeaver-D, tsInfer, Relate, and SMCSMC

[1, 129–131]. I extend the approach taken by SMCSMC to infer directional migration rates simultaneously with effective population size, enabling a unique viewpoint into the ancient past.

A particularly interesting time in human history is the end of the Middle Paleolithic approximately ∼60 ka BP, which saw the divergence of the most deeply sampled lineages of human genetic variation, introgression from multiple archaic sources, and the expansion of anatomically modern humans Out of Africa. As archaeological evidence and ancient DNA from this period are scarce, inference of demography from present-day genetic data is potentially very informative, though technically challenging. Here, we use the previously developed approach, SMCSMC, to infer population size and directional migration in a unique and dynamic period of ancient human development.

This chapter is structured as follows. Firstly, I give an introduction to pertinent historical and anthropological theories relating to this period of time to orient the reader. Secondly, I introduce competing approaches for the inference of ancestral recombination graphs and motivate the usage of SMCSMC for this application. Following this, I outline my contributions to this area of research, which involve the identification and characterization of a putative directional migration from the ancestors of modern day Eurasians to the ancestors of modern day Africans.

2.1.1 Out of Africa and the peopling of Eurasia

An abundance of archaeological and genetic evidence has shown that the continent of Africa is the historical source of all modern humans [132]. It contains more genetic sequence diversity than any other region of the world, so much so that on average, the two haplotypes that comprise a single African genome are less similar than two haplotypes taken anywhere outside the continent. [48]. Evidence from climate science suggests that a combination of a gradual shift away from aridity in Northern Africa as well as short term dry-wet cycles may have motivated both global and local range expansions [133, 134]. The most pertinent of these range expansions is the migration Out of Africa and into Eurasia, an event which has

formed the basis of modern population structure around the globe. The individual migrants involved in this event probably diverged from sister populations within the continent for many tens of thousands of years before their eventual dispersal into the Levant and beyond [3, 50]. The population which formed this successful range expansion out of Africa experienced at least one, and potentially multiple, breeding events with Neanderthals [135]. Eventually, the group split into two distinct subpopulations around the time of the Ust'-Ishim individual, approximately 45 ka BP [136]. One of these populations went East, forming the basis of for East Asians and Aboriginal Australians, while the other went West, forming the initial Upper Paleolithic European hunter-gatherers [39, 41]. These were not the only derivative groups from the original Out of Africa, as another earlier diverged population popularly known as "Basal Eurasians" are thought to have branched before the initial contact with Neanderthal populations and contribute to later European population structure [137]. From this point on, diversification occurred on a highly regional basis, beyond the scope of this brief review.

Little is known about population structure within Africa prior to the expansion of agriculturalists and pastoral groups [138, 139]. Recent evidence from the handful of successfully sequenced ancient African genomes hint at large-scale population movements and admixture from multiple highly divergent, extinct populations, with complex affinities to current groups [140–142]. The majority of structure in the continent is derived from events in the Holocene, including the spread of Bantu languages from Western Central Africa both East and South, as well as admixture from pastoralists in the Near East and Western Eurasia [138]. Eastern Africans are the most closely related group to the ancestral Out of Africa migrants, though they show particularly high levels of ancestry related to neolithic populations from Iran and the Levant consistent with multiple waves of back-migration in the Holocene [39]. Evidence for recent admixture from Eurasian sources is well established, however the lack of ancient African DNA from the Pleistocene has confounded efforts to uncover interactions between the earliest inhabitants of the continent. While the migration event associated with establishing current global

population structure has been confidently dated between 60-80 ka BP, Central African Hunter Gatherer (CAHG) and South African Hunter Gatherer (SAHG) such as the Mbuti and KhoeSan (without implying linguistic unity, defined as southern African hunter-gatherers who speak non-Bantu languages which include a click consonant) may have diverged from other groups 200-250 ka BP [141, 143]. In the intervening millennia, fossils identified as AMH have been found in China about 80-120kya, Sumatra about 63-73 ka BP, and artifacts from Australia 65 ka BP [144–146]. Support for multiple migrations across Eurasia additionally comes from climate science, where four distinct periods of warming may have provided vegetated migration routes out of Africa (OoA) as early as 120 ka BP [134]. The extent of contributions to modern day populations from "ghost" populations is unknown, though controversially suggested in Australasia and South East Asia [42, 48, 147–149] and Africa [130, 141, 150–153]. To a large degree, the fate of these anciently diverged populations and their contributions, if any, to modern day populations remains an open question. Using an extension to SMCSMC, I aim to use machine learning to investigate ancient contributions to modern day population structure within global NGS datasets.

2.2 Methods

2.2.1 A Particle Filter for Demographic Inference

Details of the Sequential Monte Carlo for the Sequentially Markovian Coalescent (SMCSMC) algorithm have been previously published [154] (see the URLs for an implementation). Briefly, SMCSMC builds an approximation of the posterior distribution of genealogical trees conditional on observed mutations along the genome using a particle filter, a method also known as sequential Monte Carlo sampling. It does so by simulating a number of sequences of genealogical trees (particles) under a fixed set of demographic parameters θ using the sequential coalescent sampler `SCRM` [79]. Simulated recombination events may change the local trees along the sequence. Particles are then weighted according to their conditional likelihood given observed polymorphisms. To avoid sample depletion, the set of

particles is regularly resampled, which tends to remove and duplicate particles with low and high weight respectively. To further increase the efficiency of the procedure, the resampling procedure targets not the partial posterior distribution that includes polymorphisms up to the current location, but also includes a "look-ahead likelihood" term that approximates a particle's likelihood's dependence on subsequent polymorphisms, while ensuring that the estimate of the posterior tree distribution remains asymptotically exact. From a sample of trees from the posterior distribution, Variational Bayes (VB) or Stochastic Expectation Maximization (SEM) is used to update the estimates of demographic parameters $\hat{\theta}$. This is repeated over a given number of iterations, or until the estimate of θ has converged.

To add the ability to infer time-varying migration rates, we exploit the capabilities of SCRM to simulate ARGs under complex demographic scenarios, and collect sufficient statistics (migration opportunity, and number, time and direction of simulated migration events) for each particle.

We use SMCSMC to infer effective population sizes and migration matrices in pairs of unrelated individuals from the phased release of the Simons Global Diversity Panel. We set a uniform recombination rate of 3×10^{-9} and a neutral mutation rate of 1.25×10^{-8}, both in units of events per nucleotide per generation; previous results indicate that modeling recombination hotspots minimally affects results [155]. To reduce the number of iterations to convergence, we initialise the particle filter with an approximation of human demographic history (Figure 2.1). We seed the model with an initial constant symmetric migration rate of 0.0092 ($M_{i,j}$; proportion per generation of the sink population replaced by migrants from the source backwards in time). We arrive at this value through simulation (data not shown).

Unless otherwise noted, the directionality of migration is given forwards in time. That is, from population A in the past to population B in the future.

2.2.2 Multiply Sequential Markovian Coalescent

We use MSMC2 to estimate the effective population size of pairs of African and Eurasian individuals using default configurations and scripts provided in `msmc-tools`

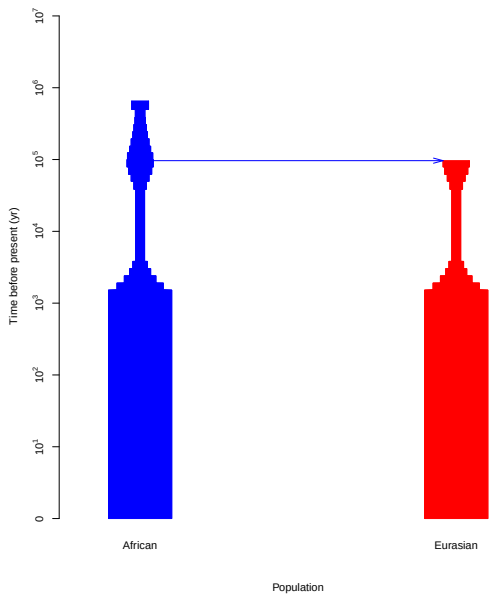


Figure 2.1: Demographic model used as initialisation for SMCSMC analysis visualised using PopDemog [156]. The width of the coloured blocks represents the effective population size at that time.

(see URLs) [3, 157]. We use a fixed recombination rate in line with our SMCSMC analysis and skip ambiguously phased sites. Twenty iterations are performed by default. We additionally compute the relative cross-coalescent rate to examine relative gene flow by transforming the coalescent rates generated by MSMC2 as indicated in the software documentation.

A brief example is given below, where I estimate coalescence rates within population 1 (Eurasians), population 2 (Africans), and between them given properly formatted input.

```
msmc2 -I 0,1 --fixedRecombination --skipAmbiguous -t 1 -o {
  output_prefix} {input_string}
msmc2 -I 2,3 --fixedRecombination --skipAmbiguous -t 1 -o {
  output_prefix} {input_string}
msmc2 -I 0-2,0-3,1-2,1-3 --fixedRecombination --skipAmbiguous -t
   1 -o {output_prefix} {input_string}
python combineCrossCoal.py {input[2]} {input[0]} {input[1]} > {
  output[0]}
```

In order to convert between SMCSMC style input (seg files) and MSMC style input files, we use a custom script found here.

2.2.3 Inferring population size and migration rates in the Simons Genome Diversity Panel

This section describes analysis of the SGDP with both SMCSMC and MSMC. SMCSMC version 1.0.1 was installed from the conda package manager (also found at `https://github.com/luntergroup/smcsmc/releases/tag/v1.0.2`), MSMC2 version 2.1.2 was installed from GitHub (found at `https://github.com/stschiff/msmc2/releases/tag/v2.1.2`) and all analyses were performed on the Oxford Biomedical Research Computation cluster.

We download pre-phased sequencing data from `https://sharehost.hms.harvard.edu/genetics/reich_lab/sgdp/phased_data/` and mask for the strict accessibility mask from the 1000 genomes project. We additionally mask for any sites absent Chimpanzee ancestry due to a known issue with the phasing algorithm [157]. We perform this masking in `vcftools`. We use the SMCSMC python package function `smcsmcv̇cf_to_seg` to convert the sequence data from VCF to seg file format, a format very similar to MSMC format. Furthermore, we provide a script to convert from seg file format to MSMC file format as well. Unless otherwise noted, the names of individuals used in this paper are the first in their population (i.e. an individual named Yoruban is `S_Yoruba1` in the SGDP nomenclature, full list in Table 2.1). We select two diploid individuals from each population in Africa and infer piece wise constant population size and directional migration rates. Specifically, we use the following options for SMCSMC:

```
smc2 -c -chunks 100 -no_infer_recomb -nsam 4 -I 2 2 2 -mu 1.25e-8
  -rho 3e-9 -calibrate_lag 1.0 -EM \${EM} -tmax 3.5 -alpha 0.0 \
  -apf 2 -N0 14312 -Np ${Np} -VB ${DEMOGRAPHIC_MODEL}
  -P 133 133016 31*1 -arg -o ${OUTPUT} -segs ${SEGS}
```

In order, we invoke the use of a QSUB cluster with `-c` and split our analysis into 100 chunks. We do not infer recombination sites along with the demographic model in order to reduce runtime. Four haploid samples, two from each population, are analyzed with a fixed mutation rate of 1.25 $\times 10^{-8}$, a fixed recombination rate of 3 $\times 10^{-9}$, and accumulating events for one unit of survival time along the sequence. We use a given number of epochs for parameter units, and bound the

upper limits of the trees at 3.5 times the effective population size (set to 14312). We use the look-ahead likelihood to guide the resampling process for a given number of particles `Np` and use variational Bayes in place of the default stochastic expectation maximization algorithm. Parameters are inferred over 31 equally spaced intervals from 133 to 133016 generations in the past, and the sampled posterior ARGs are reported. The choice of these parameters is discussed in depth in [1].

The demographic model used to seed inference is given on page 165. We visualise this demographic model in the POPdemog package in Figure 2.1 [156]. This demographic model has been designed to roughly mimic human population size history without overly biasing the results from inference.

Each SMCSMC analysis gives a final output file detailing migration and coalescent events, their rates, and their opportunities which denote the total opportunity for an event to occur during a particular epoch. Output files are trimmed to only visualise the final iteration of variational Bayes inference and assessed for convergence. Times and rates are interpreted differently than `scrm` output. Rates are in units of $4N_0$ per generation (defined here as 29 years as per [158]), while times are given in generations.

We implement the above in a `Snakemake` pipeline. Sample size and relative cross-coalescent rates are transformed as described in the documentation using the same parameter values for mutation rate and generation time used for SMCSMC analysis.

Migration during the last 100ky is integrated into a metric we call the Integrated Migration Fraction (IMF). This is related to the cumulative migration fraction (CMF) as introduced in MSMC-IM [157], except the quantity is integrated in a particular epoch. IMF is calculated as a function of time $F(t) = e^{-\int_{t=0}^{T} \rho(t)dt}$ given an upper bound T. A practically identical solution can be found from first principles. Consider p proportion of the population are replaced every generation. Start with 0 individuals from the source N_{source} population in the sink population N_{sink}, each generation replace p proportion of the sink population with the source. We track the proportion of the population which are replaced by the source P.

$$
\begin{aligned}
P_0 &= 0 \\
P_1 &= pN_{sink} \\
P_2 &= pN_{sink} + p(N_{sink} - pN_{sink}) \\
&= pN_{sink} + pN_{sink}(1-p) \\
P_3 &= pN_{sink} + pN_{sink}(1-p) + p((N_{sink} - pN_{sink}) - p(N_{sink} - pN_{sink})) \\
&= pN_{sink} + pN_{sink}(1-p) + p(N_{sink}(1-p) - pN_{sink}(1-p)) \\
&= pN_{sink} + pN_{sink}(1-p) + pN_{sink}(1-p)(1-p) \\
&\dots \\
P_n &= N_{sink}p(1-p)^n
\end{aligned}
$$

In practice, both methods give essentially identical proportions for all considered questions.

2.2.4 Statistical Analysis of Migrated Segments

We run SMCSMC with the `-arg` flag to report the posterior estimate of the ancestral recombination graph. We use this to isolate segments of the African genome where predicted migration events occurred between 50 and 70kya and used these segments to calculate drift statistics. The isolation procedure is implemented in `smcsmc.find_segments`, and involves sequentially reconstructing marginal trees and keeping track of which contain migration events in a particular epoch. We isolate segments from the marginal trees of all SGDP comparisons.

2.2.5 Length Distribution of Isolated Segments

Under the Markovian model of the SMC', the length of admixed tracts L is an exponential process with scale factor $2N(1-m)\left(1-e^{-T/2N}\right)$, with a proportion m of the sink population being replaced with the source T generations in the past and an effective population size of N [83, 159]. This gives an approximate mean length $[(1-m)r(T-1)]^{-1}$ with recombination rate r in units of Morgans, which is well approximated by $(rT)^{-1}(1-m)$ [160]; we use this approximation

to derive expected distribution of fragment sizes. When analysing populations with SMCSMC, we fix the recombination rate at 3×10^{-9} uniformly across the genome, in line with that used by tt MSMC in simulations [3]. This value is a conservative underestimate, accounting for the presence of recombination hotspots and SMCSMC's inability to deconvolve recombinations in these areas, effectively underestimating the true r. For estimates of ancestral tract lengths, we use the more universally accepted value of 1×10^{-8}, equivalent to a one percent chance of a cross-over per megabase and per generation [161].

2.2.6 Drift Statistics

Patterson's drift statistics were calculated with `ADMIXTOOLS` [127] and the `admixr` package [128] in `R`. We converted the above sequence data to Eigenstrat format with `vcf2eigenstrat` formerly distributed with `admixr`. We merged SGDP and archaic Eigenstrat datasets with `convertf` and `mergeit` implemented in `ADMIXTOOLS`.

Here, Yoruba-1 is used as a representative of Western African groups, and used for ascertaining putatively migrated segments. Yoruba-2 is used as a comparison individual from the same populations. In this way, we look for evidence above another individual in the same population of similarity to Eurasians.

2.2.7 Simulation procedure

Coalescent simulations were performed under the sequential coalescent with recombination model (`SCRM`) [79]. 1 gigabase (Gb) of sequence was simulated. In addition to branches in local genealogical trees, `SCRM` retains non-local branches in the ancestral recombination graph (ARG) within a user-specified sliding window. In the limit of a chromosome-sized windows `SCRM` is equivalent to the coalescent with recombination, while for a zero-length window it is equivalent to the sequentially Markovian coalescent (SMC') [82, 83]; we use a 100kb sliding window to approximate the CwR and improve accuracy over SMC' while retaining tractable inference.

We modelled migration as a 10ky pulse of constant migration rate resulting in an integrated migration fraction (IMF) of 0 to 0.593. The migration pulse

was centered at various times between 40 and 70 kya. Due to the amount of compute required, we then used SMCSMC to infer the demographic parameters using a reduced set of 5000 particles and 5 iterations of the VB procedure. To aid convergence, we started inference at a reasonable approximation of human demographic history (Figure 2.2). We modelled N_e and migration rates as piecewise continuous functions and set 32 exponentially spaced epochs from 133 to 133016 generations in the past. To convert evolutionary rates to years we set a generation time of 29 years [158]. For computational efficiency, individual genomes were split into 120 chunks and processed in parallel, with sufficient statistics collected and processed together in the VB steps.

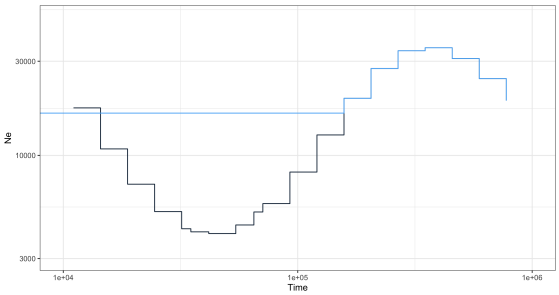


Figure 2.2: Effective Population size model used for simulations. Following the simulation procedure in Section 2.2.7, the population size model is plotted per epoch, with the effectively African population plotted in blue and the effectively non-African population in black.

Times are in units of $4N_0g$ while population sizes are in units of N_0. For $g = 29, N_0 = 14312$, the demographic model is as shown in Figure 2.2. Exact specifications are given in Appendix A

The demographic model which we have assumed for both population's effective sizes has been shown to recapitulate similar inference to real data (data not shown). The migration parameter must be initiated at a given magnitude; further back in time, the particle filter is less able to identify lineage's true populations, and the inference of migration rates becomes essential uniform. Thus, we see a "drop-off" effect, where in the ancient past, the inference remains at the initiation value, and as more certainty about different histories is obtained, the migration values recapitulate real information. Thus, the choice of an appropriate parameter for

the initial migration rate is a crucial step in SMCSMC analysis, and here we chose to arrive at this value through simulation.

We simulate back-migration scenarios of varying total migration proportions from 0 (no migration) up to 60% population replacement. For each simulation, we initiate the particle filter at either 0, 1, or 5 units of $\frac{1}{4N_0}$ proportion replaced per generation (which are the units used internally by `scrm` and `ms` for simulation). SMCSMC is then used to infer effective population size and migration histories in five iterations with 5000 particles. As a cautionary note, these simulations are almost certainly not fully converged, and are used as an indication of power. However, these low resolution attempts are indicative of a "quick" overview of the abilities of the algorithm. With 600 cores available, each of the cases (forward, backward, or bidirectional) was able to run in approximately 20 hours.

Generally, beginning with a higher migration rate seems to recover a higher proportion of the simulated migration. However, as in the case of a 60% replacement simulated 40kya, beginning with 5 $4N_0$ rather than 1 $4N_0$ recovers similar proportions of backwards migration (0.502 vs 0.52) yet the higher migration rate finds 0.301 Eurasian migration rather than 0.195. The higher initial migration rates thus slightly reduce power (though, not in all cases, and for fully converged solutions, we would expect both proportions to be similar up to noise) while additionally finding an increased migration in the opposite direction. Beginning with a zero rate leads to highly unstable estimates of the migration rate and effective population size, and we exclude it from our analysis.

2.2.8 Isolating Anciently Admixed Segments

To investigate shared genetic drift in sequences putatively inherited from an ancient migration, we study a single sample from the posterior in detail. We sample genealogical trees with migration events from the posterior distribution estimated by the particle filter under the final, converged, demographic parameters. With the caveat that no single instance of an evolutionary history will be definitively "true", inferred coalescence and migration events within inferred marginal trees are

representative of the underlying demographic processes [67]. This makes individual events in segments of the genome good candidates for analysis in aggregate. We scan along the sequence and identified marginal trees with migration events from the source (Eurasian) population to the sink (African) population (forward in time) within the desired time period along with the beginning and end position of that tree in the genome sequence. In this process, we ignore recombination events that alter a tree in such a way that the migration event is retained.

2.2.9 Sequence Data and Preparation

We downloaded whole genome sequence (WGS) data from the phased release of the Simons Genome Diversity Panel and converted it to `.seg` file format using scripts provided (See URLs). We apply two masks to the data. First, we mask the data with the strict accessibility mask provided by the 1000 genomes project (see URLs). Second, we mask any sites absent chimpanzee ancestry, to address a known variant issue in the data that resulted in artificially long runs of homozygosity [157]. We develop a `Snakemake` [162] pipeline for efficiently analysing sequence data with both SMCSMC and MSMC2. We assume a mutation rate of 1.25×10^{-8} and a recombination rate of 3×10^{-9} (events per nucleotide per generation), in line with recent literature [163, 164]. The number of particles, and the number of VB iterations, are set per analyses, and are reported in figure captions. Unless otherwise noted, the names of individuals used in this paper are the first in their population (e.g. an individual named Yoruban is `S_Yoruba-1` in the SGDP nomenclature); a complete list of sample identifiers is provided in Table 2.1.

2.2.10 Integrated Migration Fraction

The IMF, the total fraction of a particular population A replaced during a particular time period from T_0 to T_1 generations in the past is found as follows. Let $\rho(t)$ be the instantaneous rate of migration out of A per unit of time in the backward direction (i.e. into A forwards in time), and $F(t)$ the fraction not migrated in the epoch $[T_0, t]$,

then $\frac{d}{dt}F(t) = -\rho(t)F(t)$ with solution $F(t) = e^{-\int_{T_0}^{T_1}\rho(t)\mathrm{d}t}$, so that the IMF is given by $1 - F(T_1)$. The integral is calculated as a finite sum since ρ is piece-wise constant.

2.3 Results

2.3.1 Substantial Migration from Eurasian to African Ancestors

We use SMCSMC to analyse pairs of individuals from the SGDP and simultaneously infer migration rates and effective population sizes (N_e) under a two-island model with directional migration. Population sizes and migration rates are modeled as piece-wise constant across 32 exponentially spaced epochs from 133 to 133016 generations in the past, corresponding to 3.8 thousand to 3.8 million years ago (3.8kya–3.8Mya) using a generation time $g = 29$ years [158]. We find that the method infers high rates of migration from descendants of the OoA event ('non-Africans') to Africans, but not in the opposite direction, in the period 30–70kya corresponding to the Late Middle Paleolithic (Figure 2.3). In populations from the Niger-Kordofanian and Nilo-Saharan language groups, comprising the majority of the population on the African continent, the peak inferred migration rate from Eurasian populations (2.5–3.0 $\times 10^{-4}$ and 3.5–4.0 $\times 10^{-4}$, in units of proportion of the target (ancestral African) population replaced per generation) most frequently falls in the epochs spanning 35–45kya, while peak migration rates in the opposite direction are substantially lower (0.5-1.0 $\times 10^{-4}$) and occur earlier, in the epochs spanning 55–70kya (Figure 2.4). Populations in the Afroasiatic language group show evidence of large amounts of directional migration in the Holocene (Figure 2.5), which is consistent with previous findings of relatively recent European introgression into these populations [138, 165].

We track the overall peak of migration rate in different populations (Figure 2.4a,b). The most common backwards migration peak falls in the epoch between 35–45kya in the Nilo-Saharan and Niger-Kordofanian groups. Forwards migration has an earlier peak, in the epoch spanning 55–70kya. This result must be interpreted in light of the simulation results presented below.

We model the migration adjusted N_e in Eurasian populations, averaged over African partners, and African populations averaged over Eurasian partners (Figure 2.11). The resulting curves largely represent our prior knowledge of world history, with an early divergence of Papuans consistent with the timing proposed in [147], and a second bottleneck of populations inhabiting North America such as the Karitiana and Pima. Because we do not explicitly infer population split times, and there is no clear summary metric as proposed by MSMC2 in the case of directional migration, more fine-scale trends are difficult to identify. The African population size models show more discrepancy between populations, including an OoA-like bottleneck in Afroasiatic populations, and a large historical population size in hunter-gatherer groups such as proposed in [141].

To assess the impact of errors introduced by statistical phasing, as is the case for the SGDP, we repeated the analyses above on a subset of physically phased individuals from the Human Genome Diversity Project (HGDP) [48]

2.3.2 Validation in a physically phased subset of the Human Genome Diversity Panel (HGDP)

Phased data is not essential for demographic inference using SMCSMC; however, the use of phase alongside the look-ahead likelihood allows for more efficient convergence. The Human Genome Diversity Project collected 929 genomes from a diverse collection of human populations [50]. 36 of these genomes, two each from nine Eurasian and four African populations, were physically phased by use of linked-read sequencing technologies. This resource allows us to validate our inference both in an independent dataset, and evaluate the effect of phasing errors on SMCSMC inference.

To analyse these data, the same `Snakemake` pipeline was used with minor adjustments in wildcard constrains to account for differences in sample names. 120 chunks of the genome were run in parallel for reasons of computational efficiency, while fixed recombination rate and mutation rates were held at the same values as the SGDP analysis, and an identical demographic model was used to initiate the analysis. Three replicates of the analysis were performed to assess the impact

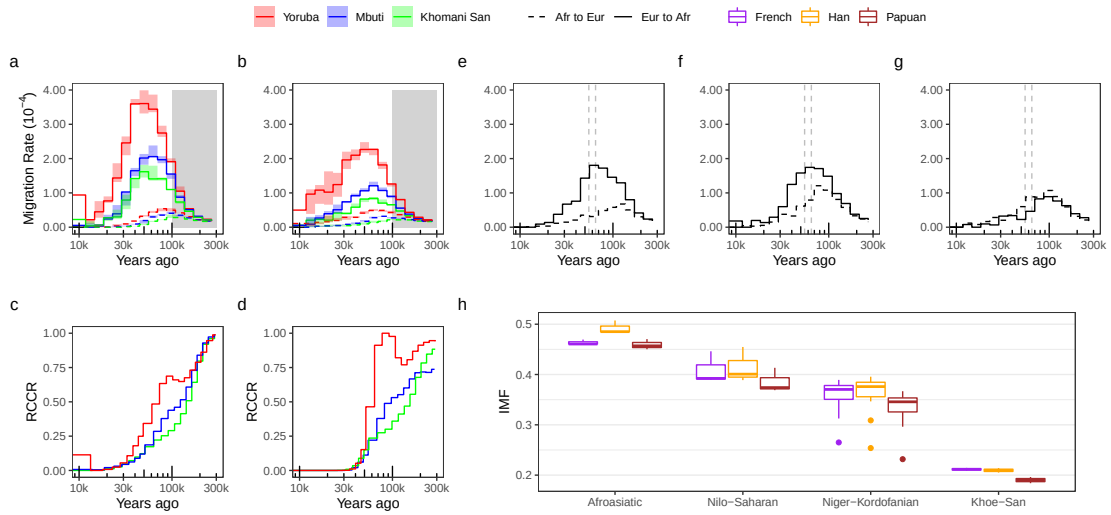


Figure 2.3: Migration rate inference. **a**. Inferred migration between an African individual and a Han Chinese individual in the SGDP. Three replicates were performed, with the median estimate plotted and the range shaded. Solid lines show inferred migration from Eurasians to Africans (forward in time) while dotted lines show the reverse migration. The SMCSMC analysis used 10000 particles to estimate the posterior distribution of marginal trees, and 25 iterations of variational Bayesian inference to achieve converged parameter estimates. The shaded grey regions represent a time period where simulation shows SMCSMC has very little power to infer migration (details found in Section 2.2.7). **b**. The same analysis as in a. except using individuals from the physically phased subset of the HGDP, showing similar differences between populations but systematically lower migration overall. Three replications were performed to estimate error and the standard deviation is shaded. The same SMCSMC settings were used as in a. **c**. Relative cross-coalescence rate (RCCR) estimated by MSMC in three different populations in the SGDP, supporting gene flow between Eurasians and Yorubans not shared by Mbuti or Khoe-San. 40 iterations were used to achieve parameter convergence. **d**. The same analysis as in c. but performed on individuals in the physically phased subset of the HGDP, similarly supporting shared gene flow between the Yoruban and Eurasians not shared by Mbuti or Khoe-San. **e**, **f** and **g**. Inferred migration rates from data simulated under a two-island model with, from left to right, a backward Eurasia-to-Africa, a bidirectional, and a forward migration pulse lasting 10 ka BP (dashed vertical lines) and replacing 40% of the recipient population(s) approximately 60kya. The migration rate from Africa to Eurasia is not well estimated by SMCSMC (Figure 2.13–Figure 2.15 and Section 2.2.7), but SMCSMC is well powered to infer migration from Eurasia to Africa in this period. **h**. Integrated total migration fraction (IMF) over the last 100 thousand years stratified by language phyla in the SGDP and comparison Eurasian population used to estimate migration. Afroasiatic (Mozabite, Saharawi, and Somali), Nilo-Saharan (Dinka, Luo, and Masai), Niger-Kordofanian (BantuHerero, BantuKenya, BantuTswana, Biaka, Esan, Gambian, Luhya, Mandenka, Mbuti, and Mende), and San (Khomani San and Ju hoan North) are grouped as in [165]. Similar levels of migration are inferred from French and Han Chinese to all language groups, with significantly less migration from Papuan groups ($p \leq 0.05$, two-tailed paired t-test, Table 2.3). Outliers in the Niger-Kordofanian group are the Mbuti.

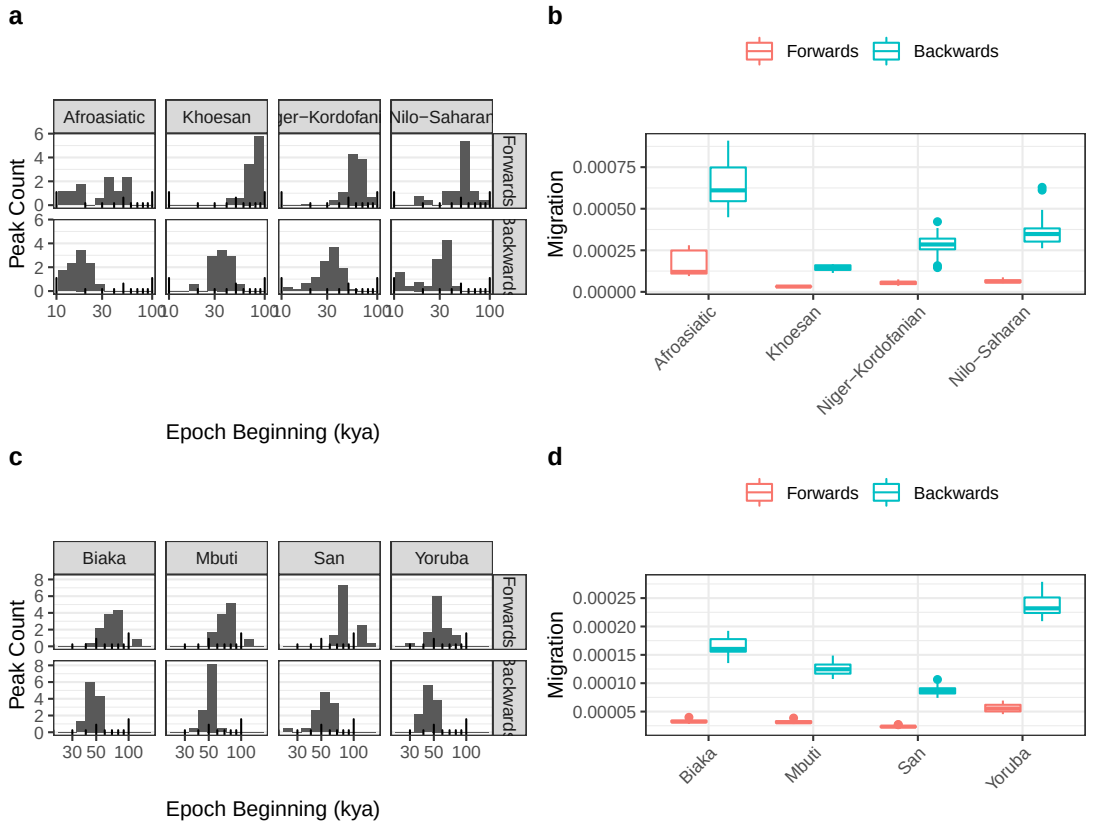


Figure 2.4: Timing and average maximum rate of directional migration in HGDP and SGDP. **a** Migration is inferred in evenly spaced epochs on the log scale from 3.8 thousand to 3.8 million years ago. For each population in the SGDP, we record the epoch with the highest inferred directional migration rate (the "peak" of migration) and plot this as a histogram. Backwards migration refers to migration from Eurasians to Africans, whilst forward represents the reverse. **b** In the epochs of highest migration identified in a., we record the inferred rate per population and plot these as a boxplot. Whiskers represent 1.5 times the interquartile range. The migration rate is given in proportion of the population replaced per generation. **c** and **d** represent the same analyses as in a. and b. calculated for the Human Genome Diversity Panel, rather than the SGDP.

of stochastic sampling variation on inference. We infer both effective population size and directional migration in each of these 9x4 comparisons between Eurasian and African populations (Figure 2.9a). The resulting inference allows us to verify and validate many observations from the SGDP.

Firstly, we calculate the timing of the migration peak and its magnitude, and find the estimates largely in line with the SGDP inference (Figure 2.4c,d). For instance, inferred backwards migration in the Yoruban and Biaka populations peak at 40-50kya, while the Mbuti and San show earlier migration peaks around 50-60kya

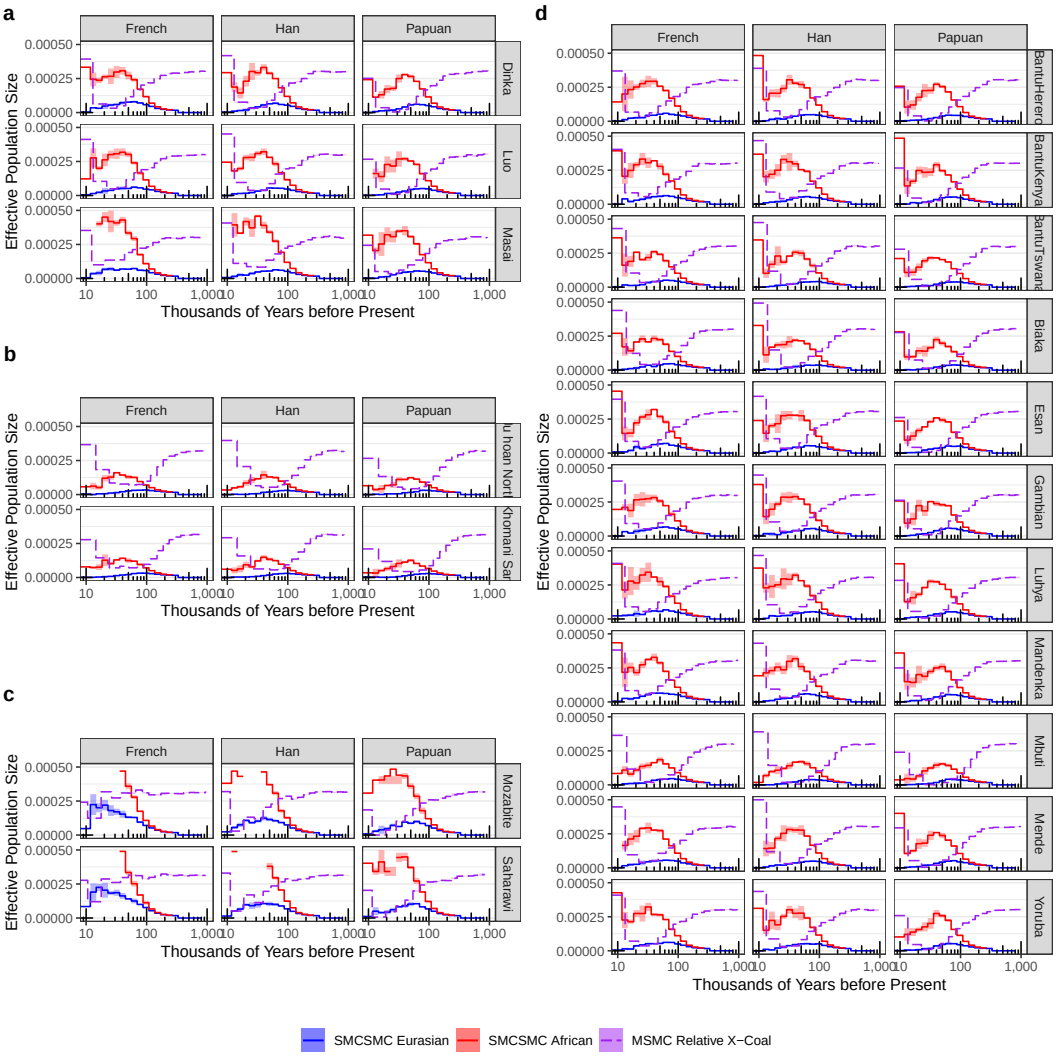


Figure 2.5: Inference of directional migration in the Simons Genome Diversity Project. The SMCSMC particle filter was used to infer directional migration rates in both directions from one of three Eurasian populations (French, Han, and Papuan) to one of 18 African populations. 5000 particles were used to approximate the ancestral recombination graph with 10 iterations of variational Bayesian inference to update demographic parameter values. Panels represent a. Nilo-Saharan, b. KhoeSan, c. Afroasiatic, and d. Niger-Kordofanian language families. Alongside the SMCSMC inference, we use MSMC2 to infer the relative cross coalescent rate (RCCR) with default settings and 20 iterations for convergence.

(Figure 2.4c). The migration rate at the peak shows the same qualitative trends as the SGDP, with the peak in the Yoruban (approximately 2.5×10^{-4}) far exceeding the peaks in the Biaka, Mbuti, or San (between $0.1 - 0.175 \times 10^{-4}$) (Figure 2.4d). This replication in the HGDP confirms the presence of a large directional migration in the Late Middle Pleistocene, and demonstrates that statistical errors in phasing the SGDP are not large contributors to the qualitative trends observed.

We integrate migration between 40–70kya to obtain the inferred IMF for each of the comparisons in the HGDP (Figure 2.9b). Differences amung the African populations mirror those in the SGDP, though the proportions are uniformly smaller (main text). However, the migration rates backwards into Africa are apparent in all comparisons, and the order of populations IMF remains the same. Differences between individual Eurasian donor populations are small and, with the exception of the Papuan, insignificant. A discussion of the Papuan comparisons appears in the subsequent section.

To compare the HGDP inference with the SGDP inference, we construct a set of the SGDP with the same donor populations as the HGDP.

2.3.3 Comparisons between the HGDP and a subset of the SGDP

Previously, inference in the SGDP has relied on three candidate Eurasian donor populations. However, the physically phased subset of the HGDP provides a higher resolution view into global migration patterns with nine Eurasian populations represented. In order to compare effectively between the inferences made in these two datasets, we find representatives from these nine Eurasian populations in the SGDP dataset and use them as donor populations to the same four African populations (Yoruban, San, Mbuti, and Biaka), effectively recreating the analysis done in the physically phased subset of the HGDP. We select the Khomani San as a representative of the San, and only use one of the Papuan populations in the HGDP to compare (Highlands, as opposed to Sepik), creating the same 8x4 analysis table for both data sets. We infer the effective population sizes and migration

rates using both SMCSMC and MSMC, with analysis details effectively identical to the original comparisons listed above (Figure 2.6). We average over inferences to visually compare trends between the two datasets, in the same populations and compute the inferred IMF between 40–70kya (Figure 2.9).

In both the HGDP and the SGDP, MSMC estimates of African population size are higher than SMCSMC estimates in the ancient past (80 – 300kya) (Figure 2.9a). By modelling directional migration, we are able to account for excess genetic diversity in the ancestral African population in both datasets. Uncertainty in the estimates increases substantially nearer to the present, as would be expected with the SMCSMC method.

We summarise migration from 40–70kya in the HGDP similarly to the SGDP. The total inferred migration is lower in the HGDP than in the SGDP (Figure 2.9b). We use this comparison setup to additionally test the differences between Papuan donors and the remainder of Eurasians. We construct a linear model predicting IMF based on an indicator variable of Papuan/not Papuan and the receptor donor population, and find that in both the SGDP and the HGDP, Papuans show approximately 2% less IMF than other donor populations (Table 2.5, Table 2.4). While this difference is small, it is highly significant. However, the demographic scenario causing this difference in inferred IMF is not obvious; it is possible that the Papuan group had begun to diverge from the donating population prior to the admixture event, or alternatively that differences in archaic admixture between Eurasian and Papuan groups make up the difference in affinity.

However, the qualitative patterns in inferred directional migration rates between populations are similar in both datasets (Figure 2.9c). In both datasets, the highest rates are found in the Yorubans, follow by the Biaka, then the Mbuti and San. The MSMC curves are interestingly dissimilar between the different data sets, with a much steeper ascent around the period of our inferred migration in the HGDP than the SGDP.

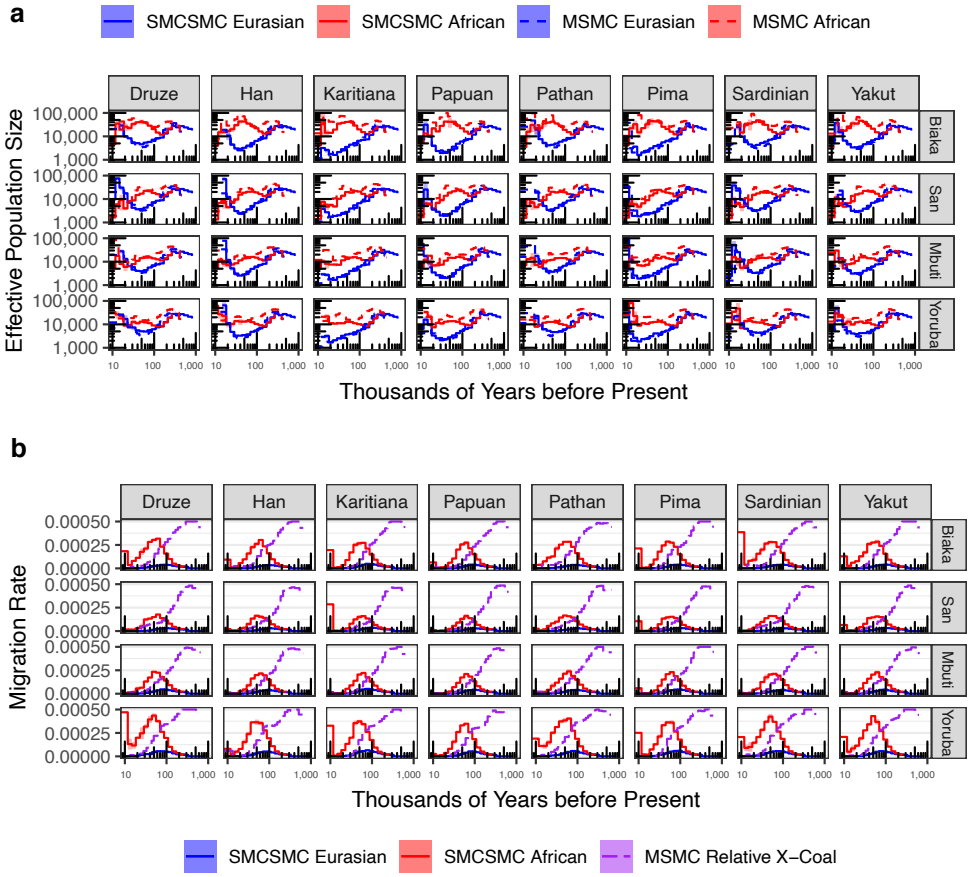


Figure 2.6: Demographic inference in a matched subset of the Simons Genome Diversity Panel. **a.** SMCSMC and MSMC2 inferred effective population size of several populations in the Simons Genome Diversity Panel. These samples were selected to match, as closely as possible, those in the physically phased subset of the Human Genome Diversity Project panel. b. Inferred migration using SMCSMC in the Simons Genome Diversity Panel along with the scaled relative cross-coalescent rate estimated by MSMC2. 10,000 particles were used to approximate the ancestral recombination graph in the SMCSMC particle filter and 25 iterations were used to update demographic parameters. 20 iterations were used for MSMC2.

2.3.4 Simulation demonstrates power to infer large directional migration pulses

We asked whether SMCSMC has power to detect a large back-migration event in the Late Middle Paleolithic and distinguish it from other demographic scenarios. To answer this we used `SCRM` [79] to simulate a gigabase of sequence data under a two-island demographic model, with effective population sizes chosen to be comparable to typical African and Eurasian populations as inferred from real data.

To this we added a 10ky pulse of forward, backward or bidirectional migration of varying strengths, with the midpoint of the migration pulse within the range 40 to 70kya. To quantify the inferred amount of migration we calculate the integrated migration fraction (IMF), defined as one minus the probability that a lineage in the destination (e.g. African) population traced backwards in time remains in that population across a given epoch according to the migration model (see Methods). For the simulations, we chose the most recent 100kya as epoch, and used scenarios with IMFs ranging from 0 to 0.593. For each simulation we report the inferred IMF in both the forward and backward direction (Figure 2.7). We find that SMCSMC has good power to detect backward migration pulses up to 60kya (median ratio of inferred and true IMF, 0.91), while power drops off at 70kya (IMF ratio 0.46). In the pure backward migration case, some forward migration is falsely inferred, but this is always substantially less than the inferred backward migration (median ratio inferred forward to true backward IMF, 0.37; true migration peak $\leq$ 60kya). However, in the case of true forward migration as well as bidirectional migration, roughly equal mixtures of forward and backward migration are inferred (Figure 2.7). We conclude that in the epoch 40–70kya the forward and bidirectional scenarios are difficult to distinguish from each other, but both can be distinguished from backward migration, the only scenario resulting in substantially different inferred backward and forward migration.

To validate the existence of the migration pulse, though not its direction, we next analyzed the same data using MSMC, which is widely used to estimate gene flow in the ancient past by estimating the relative cross-coalescent rate (RCCR) between two populations [3, 165–167]. We use the updated implementation MSMC2 recommended by the authors and first published in [147]. Each of the SMCSMC analyses are repeated using MSMC2 to estimate effective population size and RCCR (Figure 2.5, Figure 2.9, Figure 2.8). Consistent with previous analyses conducted with MSMC2, our estimates show high RCCR in the Late Middle Pleistocene in both the SGDP and the HGDP (Figure 2.3c,d) [50, 165]. These observations

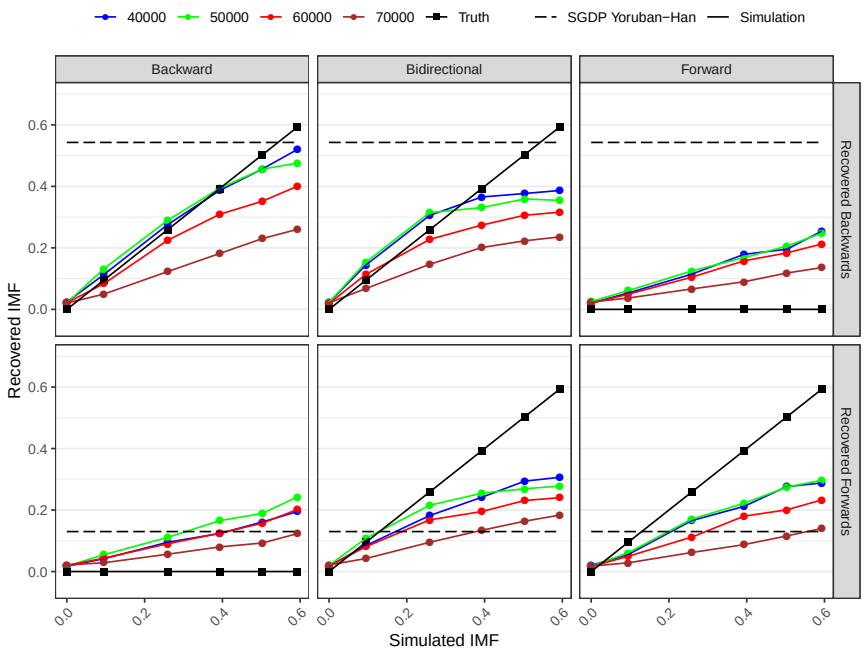


Figure 2.7: Simulation study. `SCRM` was used to simulate 1 gigabase sequence data for two diploid individuals under three different migration models. Migration was simulated backwards (from a Eurasian-like population to an African-like population), forward (the reverse), and symmetrically (equal migration in both directions). The amount of migration indicates the proportion of the sink population replaced by the source over a 10ky period centered at 40, 50, 60, or 70kya. The total IMF inferred by `SMCSMC` over the last 100ky is plotted and compared to the true simulated amount. For reference, the inferred IMF in either direction across 0-100kya for a Yoruban and Han individual is given in dashed lines. 5 iterations of variational Bayes and 5000 particles were used for inference. The effective population size model and additional details are given in Section 2.2.7.

confirm the existence of a substantial pulse of ancient gene flow between Eurasians (Han Chinese) and Africans.

2.3.5 Migration Pre-dates East-West Eurasian Divergence

To assess whether the inferred back-migration shows variation across the descendants of the OoA event, we repeated the analyses using three representative non-African groups in the SGDP: Han Chinese, French European, and Papuans. Since simulations show that SMCSMC has little power to detect migration predating 70kya, and to exclude Holocene migration, the epoch we use to calculate real-data IMFs comprise the period of peak inferred migration up to the period of diminishing power (30–70kya); we use this epoch for all subsequent analyses. Inferred IMFs

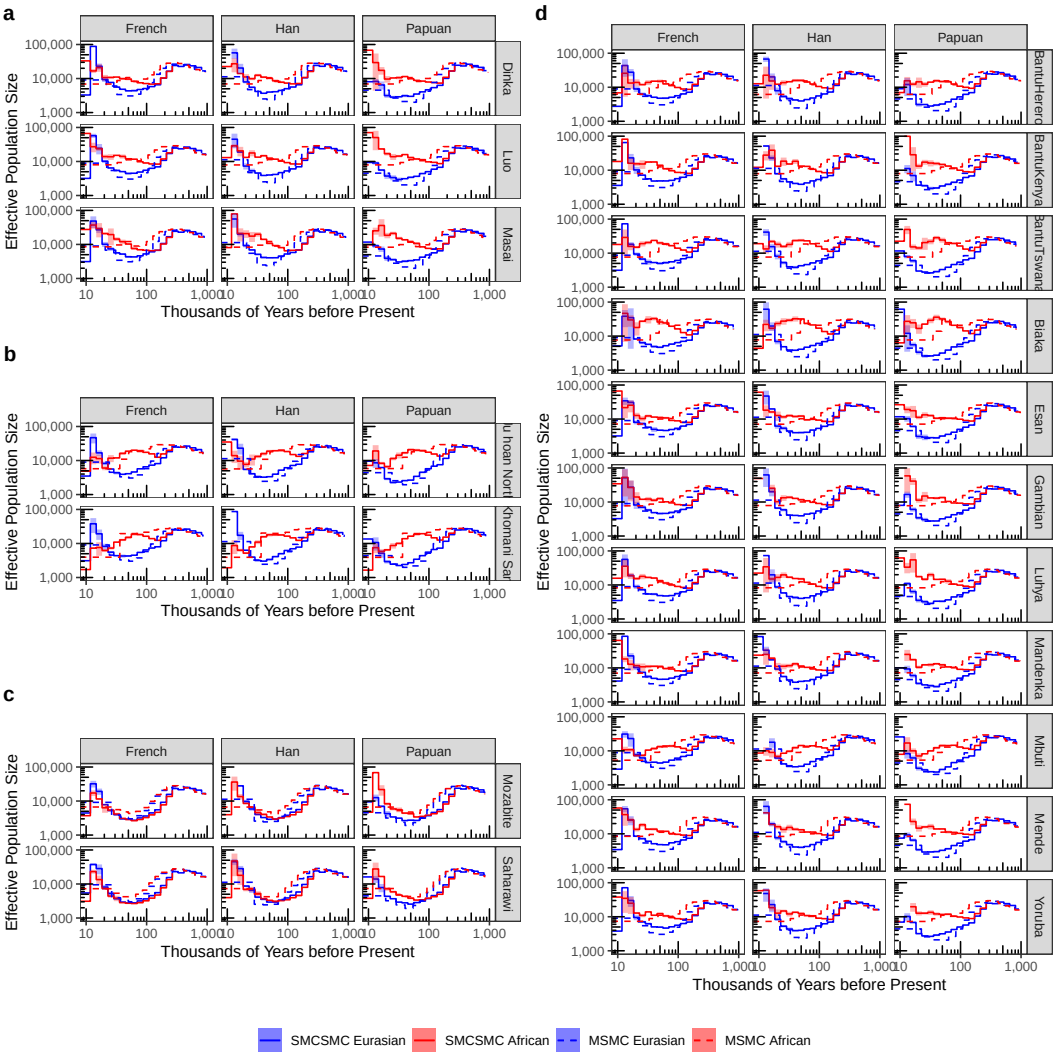


Figure 2.8: Inference of historical effective population size in the Simons Genome Diversity Project. The SMCSMC particle filter was used to infer directional migration rates and effective population size in both directions from one of three Eurasian populations (French, Han, and Papuan) to one of 18 African populations. 5000 particles were used to approximate the ancestral recombination graph with 10 iterations of variational Bayesian inference to update demographic parameter values. Panels represent a. Nilo-Saharan, b. KhoeSan, c. Afroasiatic, and d. Niger-Kordofanian language families. Alongside the SMCSMC inference, we use MSMC2 to infer the same values with default settings and 20 iterations for convergence.

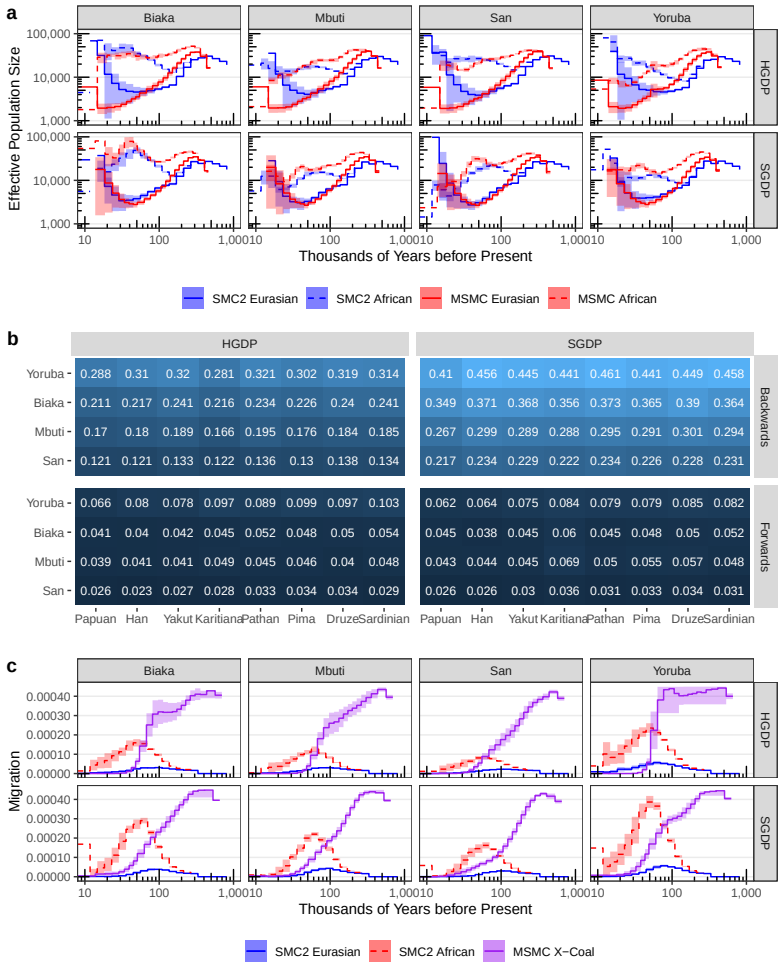


Figure 2.9: Inference of directional migration is comparable between data sets and phasing strategies. We used SMCSMC to simultaneously infer directional migration rates and effective population size in the 36 genome physically phased subset of the Human Genome Diversity Panel. We match these 36 genomes with comparable individuals in the Simons Genome Diversity Panel (with the exception of one Papuan population, which has no comparable population in the SGDP) and perform an identical analysis. **a.** Average N_e estimate across four populations in the physically phased subset of the HGDP and the subset of SGDP used to compare with HGDP inference. Inference of population size is averaged over eight Eurasian populations, with the bars representing standard deviation. For MSMC2, the time indexes were averaged to have consistent start and stop times for the steps. **b.** Inferred integrated migration fraction (IMF) from Africa to Eurasians (forwards) and from Eurasians to Africans (backwards) between 40 and 70 kya (see Methods). **c.** Directional migration inference in African populations averaged over Eurasian partners in the two data sets. Shaded regions denote standard deviations. For MSMC, the time indexes were averaged to have consistent start and stop times for plotting.

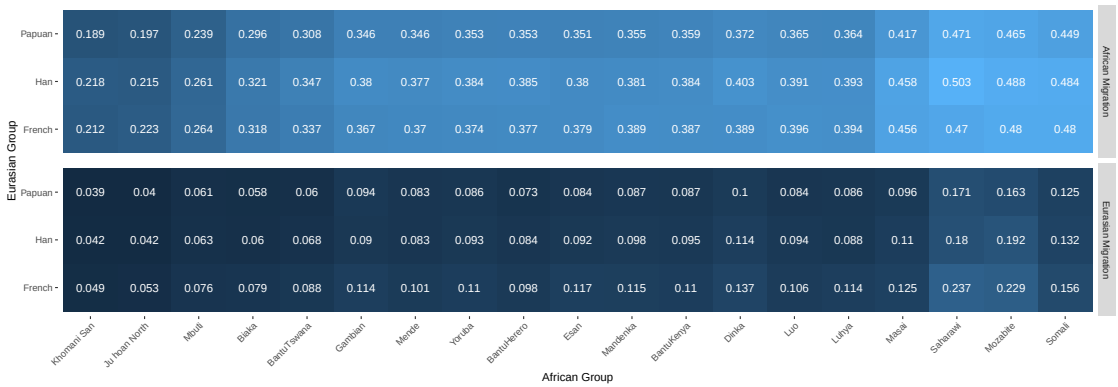


Figure 2.10: Integrated migration fraction 40–70kya in `SMCSMC` analysed SGDP populations. Directional migration was integrated by finding the cumulative probability of an individual migrating during the specified epoch. Directional migration backwards from Eurasia to Africa and forwards from Africa to Eurasia (both forward in time) are reported separately. Displayed is the average values from three technical replicates.

are not significantly different between Han Chinese and European populations in non-Afroasiatic populations (p=0.14, two-tailed paired t-test; Figure 2.3h and Figure 2.10, Table 2.3), consistent with migration occurring before the European-East Asian split approximately 40kya [168]. The contribution of this admixture event to extant African genetic variation is substantial; the estimated IMFs indicate that for individuals in the major African language groups, approximately a third of ancestral lineages trace their ancestry through the proto-Eurasian population (Niger-Kordofian group, 0.35 ± 0.04; Nilo-Saharan groups, 0.41 ± 0.03; Table 2.3). When we estimate these proportions using a Papuan sample to represent non-African descendants we find slightly but significantly smaller values compared to estimates using either the Han Chinese or European populations (mean difference of 0.029 ± 0.002, p=9.2×10^{-15}, and 0.025 ± 0.004, p=2.3×10^{-10}, paired t-tests, Table 2.3, Table 2.4). Similarly, in the HGDP, inferred migration in both Papuan groups (Sepik and Highlands) was 0.025 ± 0.004 (p=1.4×10^{-6}) lower than French and Han (Table 2.5). We comment on this observation in the Discussion.

2.3.6 Directional Migration Explains Excess Inferred African Genetic Diversity 100kya

Previous studies looking at Effective Population Size (N_e) in human ancestral populations have consistently reported inflated inferences in African populations approximately 100kya, often hypothesized to be due to unaccounted-for population substructure within Africa [3, 155]. We use SMCSMC to analyze African individuals paired with an individual from one of three non-African populations (Han Chinese, French European, and Papuans) and infer N_e for the African ancestral population under a two-island model with directional migration. Each analysis was repeated three times to assess the contribution of stochastic sampling to the inferences (Figure 2.12, Figure 2.8, per population N_e in Figure 2.11). SMCSMC infers substantially lower African N_e than MSMC in the period 80kya–300kya. In addition, while MSMC inferences show convergence of African and Eurasian ancestral N_e estimates only around 300kya, inferences from SMCSMC indicate convergence at 150kya (Fig. Figure 2.12a), closer to the hypothesized time of the diversification of the ancestral lineages prior to the main out-of-Africa migration episode [134, 147]. The same analysis on physically phased samples from HGDP show that these results are not driven by errors due to statistical phasing (Figure 2.9 and Section 2.3.2). When we used SMCSMC to infer both African and European N_e under a single-population model without migration, N_e estimates were comparable to those from MSMC (Figure 2.12b), indicating that the SMCSMC inferences are not driven by methodological biases particular to SMCSMC.

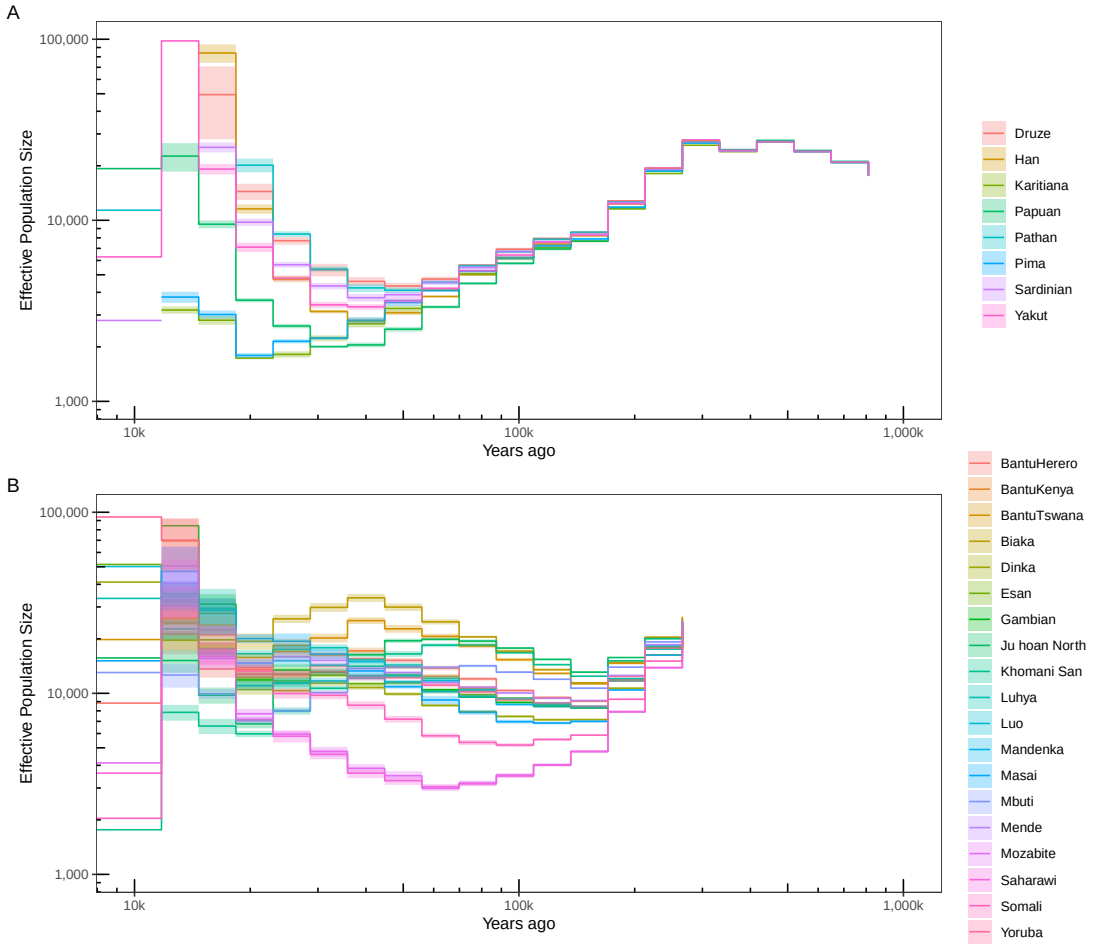


Figure 2.11: Estimates of individual population sizes incorporating directional migration. Using `SMCSMC` the effective population size of global populations in the Simons Genome Diversity Panel is inferred while simultaneously fitting directional migration estimates. Averages are plotted by epoch, with shaded regions denoting the standard deviation. **a.** Estimates of Eurasian population sizes when averaged over Eurasian donor populations. This analysis uses the eight Eurasian populations matched to HGDP populations averaged over the four matched African populations. B. Estimates of African population sizes when averaged over Eurasian recipient populations. This analysis uses the three donor Eurasian populations used for the majority of the analyses in the main text (French, Han, and Papuan) along with the given African populations. Before approximately 250kya, the populations share the same population size within the model, and are not plotted. 10,000 particles are used to approximate the ancestral recombination graph in the SMCSMC particle filter and 15 iterations are used to update demographic parameter values.

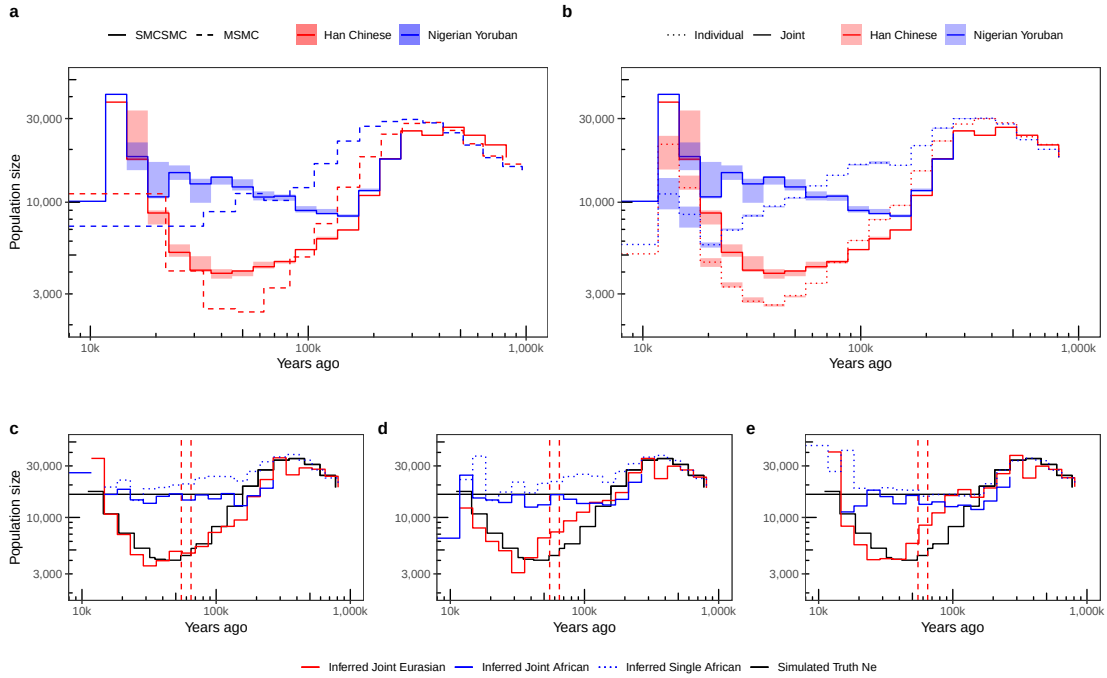


Figure 2.12: Effective population size inference. **a.** Analyzing a Nigerian Yoruban and a Han Chinese individual from the Simons Genome Diversity Panel jointly in a two-island model with directional migration using `SMCSMC` yields markedly lower N_e estimates and a more recent apparent split time, than when the same data are analyzed using `MSMC` with a model that does not explicitly include migration. Analyses for `SMCSMC` repeated three times; range of the estimates shaded. **b.** When each individual is analysed separately, using a model not including migration, N_e estimates from `SMCSMC` are similar to those of `MSMC2`. (Joint estimate from **a** included for comparison.) **c, d,** and **e.** Inferred Eurasian and African N_e from data simulated under a two-island model with, from left to right, a backward Eurasia-to-Africa, a bidirectional, and a forward migration pulse lasting 10ky (dashed vertical lines; same data as for Figure 2.3**e-g**). Particularly for the backward migration case, inferred N_e under a two-island model tracks the true values (black) well, while inferred N_e under a single-population model are inflated around the split time. All `SMCSMC` analyses used 10000 particles and 25 variational Bayesian iterations; `MSMC` analyses used 40 iterations (Section 2.2.7).

To more directly support the interpretation that the lower African N_e inferred by SMCSMC is due to appropriate modeling of directional migration, we again used coalescent simulation with `SCRM` to investigate various migration scenarios and their effects on inferred African N_e. Using the simulation framework as above, we examine N_e estimates inferred under a two-island model with migration, and in addition N_e separately inferred for each of the two simulated populations under a single-population model (Section 2.2.7). Focusing on single-population inferences, we

found that for simulated African populations that had received substantial migration from the simulated Eurasian population either through backward or bidirectional migration, inferred N_e values indeed were substantially inflated compared to true values (Fig. Figure 2.12c,d), while this effect was not seen when forward (African-to-Eurasian) migration was simulated (Fig. Figure 2.12e). Similarly, single-population Eurasian N_e estimates were inflated in the presence of forward and bidirectional migration, but not backward migration (Figure 2.13–Figure 2.15). In contrast, when using a model that includes migration, inferred African N_e do not show inflation in any of the three scenarios (Fig. Figure 2.12c-e). We conclude that the inferences from SMCSMC and MSMC are compatible with substantial back-migration from ancestral Eurasians into Africans, but not substantial bidirectional or forward migration.

2.3.7 Less Gene Flow to Central and South African Hunter-Gatherers

We infer substantial Eurasian back-migration into all African groups, however the inferred IMFs for individuals from Khoe-San populations are significantly lower than for any other group (difference with Niger-Kordofians, 0.14 ± 0.02, $p = 4.4 \times 10^{-14}$; difference with Nilo-Saharans, 0.20 ± 0.03, $p = 6.9 \times 10^{-9}$, two-tailed t-test, Table 2.2). To further support this observation we used MSMC to estimate the relative cross-coalescent rate (RCCR) for several populations, and find evidence for gene flow between Yorubans and Eurasians that is not shared with the Khoe-San individuals in either the SGPD and the HGDP (Figure 2.3c,d). These results are consistent across Eurasian donor populations (Figure 2.6). The Khoe-San individuals are particular outliers, whose ancestors are inferred to have experienced approximately half the amount of admixture seen in Nilo-Saharan and Niger-Kordofanian groups (Figure 2.10). In addition, we find that the Mbuti and Biaka, both Central African hunter-gatherer populations, show levels of Eurasian gene flow that are intermediate between levels observed in the Khoe-San and Yorubans (Figure 2.3a,b, Table 2.3). This is mirrored by inferred IMFs for Central African Hunter Gatherers, which are significantly lower than other Niger-Kordofanian groups (difference -0.08 ± 0.03,

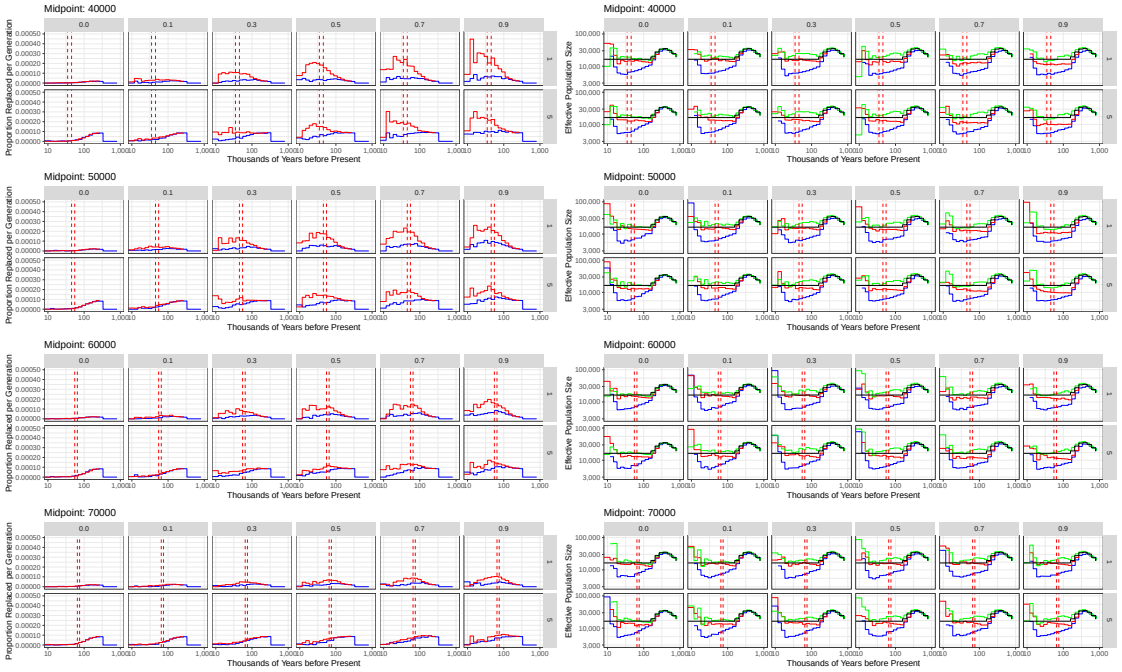


Figure 2.13: Simulated migration backwards from effectively Eurasian to effectively African populations. One gigabase of simulated sequence data was generated with `SCRM` for two diploid individuals from different populations using a 100kb sliding window approximation of the coalescent with recombination and a demographic model similar to Eurasians and Africans specified in Section 2.2.7. The timing, representing the midpoint of a 10ky migration episode, and the integrated migration fraction (IMF) were systematically varied. In this scenario, only backwards migration was simulated. The left panel displays the recovered directional migration, with red lines representing migration backwards from Eurasia and blue representing migration forwards to Eurasia. The timing of the migration episode is demarcated with dotted red lines. The right panel displays the inferred effective population size of both populations. Additionally, the effective population size of the African-like population was modelled separately, without simultaneously inferring migration. 5000 particles were used in the SMCSMC particle filter to approximate the ancestral recombination graph and 5 iterations of variational Bayesian inference were used to updated demographic parameter values.

$p = 1.2 \times 10^{-3}$, Table Table 2.2), possibly reflecting the proposed early split times of the Mbuti and Biaka from the remainder of ancestral African populations between 60 and 200kya [139, 141].

2.3.8 No Evidence for Excess Neanderthal Ancestry

Previous studies have proposed that a backflow from Eurasia may have brought Neanderthal ancestry into African populations [169]. To assess whether the proposed Late Middle Paleolithic back migration might have introduced Neanderthal material,

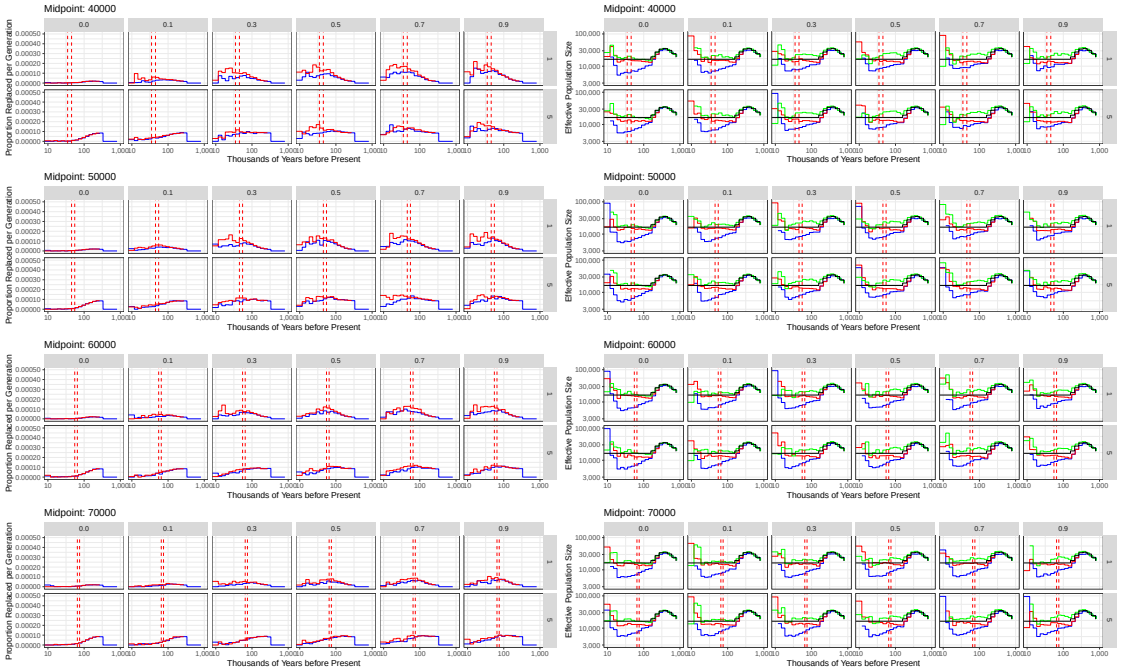


Figure 2.14: Simulated bidirectional migration between effectively Eurasian and effectively African populations. One gigabase of simulated sequence data was generated with `SCRM` for two diploid individuals from different populations using a 100kb sliding window approximation of the coalescent with recombination and a demographic model similar to Eurasians and Africans specified in Section 2.2.7. The timing, representing the midpoint of a 10ky migration episode, and the integrated migration fraction (IMF) were systematically varied. In this scenario, migration between these two populations was simulated with equal rates. The left panel displays the recovered directional migration, with red lines representing migration backwards from Eurasia and blue representing migration forwards to Eurasia. The timing of the migration episode is demarcated with dotted red lines. The right panel displays the inferred effective population size of both populations. Additionally, the effective population size of the African-like population was modelled separately, without simultaneously inferring migration. 5000 particles were used in the SMCSMC particle filter to approximate the ancestral recombination graph and 5 iterations of variational Bayesian inference were used to updated demographic parameter values.

we analyzed a Yoruban and a French individual using SMCSMC to draw a sample from the posterior distribution of ARGs, isolated the marginal trees containing an inferred back-migration event in the epoch 30–70 kiloanni (thousands of years) before present (kya), and reported the inferred admixture tracts. This analysis relies on a sample from the distribution of likely ancestral events, which makes interpreting particular genomic segments challenging. However, along the genome we expect that segments which contain inferred migration events in this period

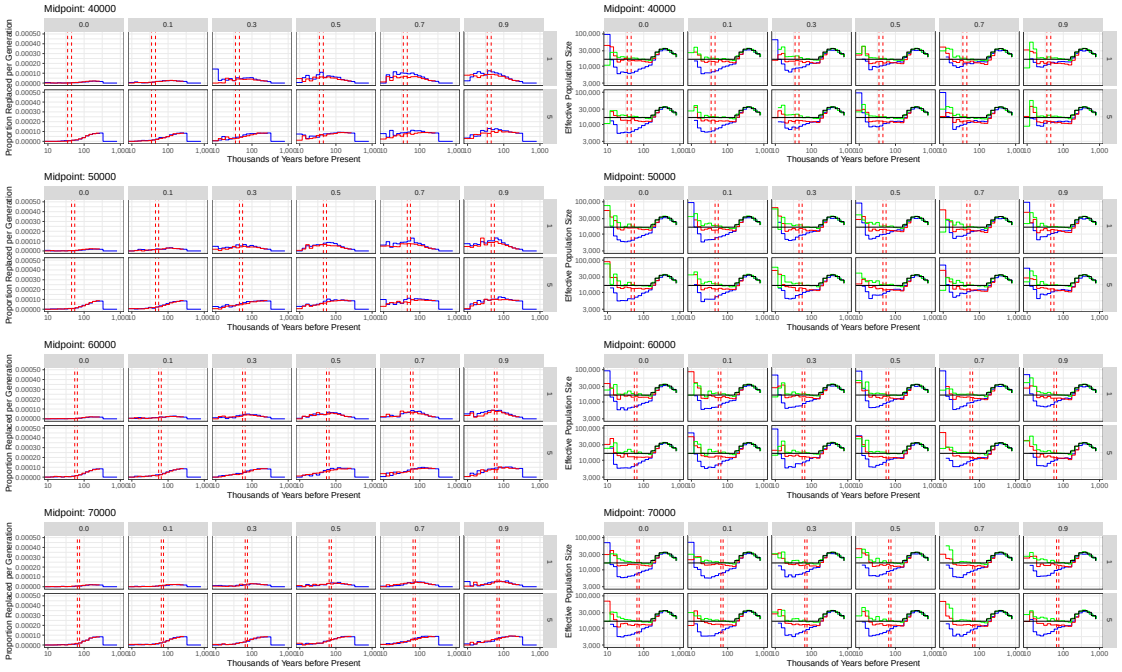


Figure 2.15: Simulated migration forwards from effectively African to effectively Eurasian populations. One gigabase of simulated sequence data was generated with `SCRM` for two diploid individuals from different populations using a 100kb sliding window approximation of the coalescent with recombination and a demographic model similar to Eurasians and Africans specified in Section 2.2.7. The timing, representing the midpoint of a 10ky migration episode, and the integrated migration fraction (IMF) were systematically varied. In this scenario, only forwards migration was simulated. The left panel displays the recovered directional migration, with red lines representing migration backwards from Eurasia and blue representing migration forwards to Eurasia. The timing of the migration episode is demarcated with dotted red lines. The right panel displays the inferred effective population size of both populations. Additionally, the effective population size of the African-like population was modelled separately, without simultaneously inferring migration. 5000 particles were used in the SMCSMC particle filter to approximate the ancestral recombination graph and 5 iterations of variational Bayesian inference were used to updated demographic parameter values.

will be enriched for segments truly descended from ancient admixture. Here we study drift within these segments across global populations.

To assess whether the identified segments are plausible under the demographic model in question, we confirmed that their length distribution is consistent with the IMF and timing of the migration inferred by SMCSMC. We take the isolated segments (Figure 2.16a) and compute the mean track length (Table 2.6). We use the approximation that the mean segment length should be approximately equal to $((1-m)r(T-1))^{-1}$ to determine that, if the migration happened in one pulse, our

empirical distribution would suggest either a recent timing or a very large pulse (Figure 2.16b). However, we heavily caveat any interpretation of these data with the fact that they are explicitly generated under a model of a given migration proportion. The fact, therefore, that they are of a consistent length with a large migration is more evidence for the model producing internally-consistent tracts than any external validation of the results in this article.

The assumption that migration has occurred in a single wave is largely unrealistic. We used expectation maximization to investigate if a mixture of exponential distributions explained the observed tract lengths better than a single distribution. We found that in some cases, two or three exponential distributions were better supposed by the data, however the differences in log likelihoods was negligible and the support for the different distributions was approximately inversely proportional to their number (data not shown). We found no strong support for multiple waves of migration from this analysis.

As expected, we found that these African segments with putative Eurasian ancestry tend to be more closely related to a Eurasian sample than another representative of the same African population (Table 2.7, Figure 2.17, Figure 2.18) in a global dataset of modern and ancient individuals compiled by the Reich group (see URLs). Within these African segments that are likely enriched for material with Eurasian ancestry, we then used D statistics [127] to identify enrichment for Neanderthal material compared to an African background.

We first use f_3 statistics to look for evidence of admixture between the African and various Eurasian groups. Thus, we calculate f_3(Yoruba-1, Eurasian group, Yoruba-2) for Papuans, French, Han Chinese, and the Vindija Neanderthal. We calculate this statistic in all available markers, and additionally for the segments isolated from the three Eurasians separately (Figure 2.17). These statistics show, firstly, that ascertaining in a particular group influences the shared drift with that group. This is exemplified by the non-significant shared drift with Papuans in French and Han ascertained segments. Secondly, these statistic show significant levels ($|Z| > 2$) of drift between the test individual and Eurasian populations, while

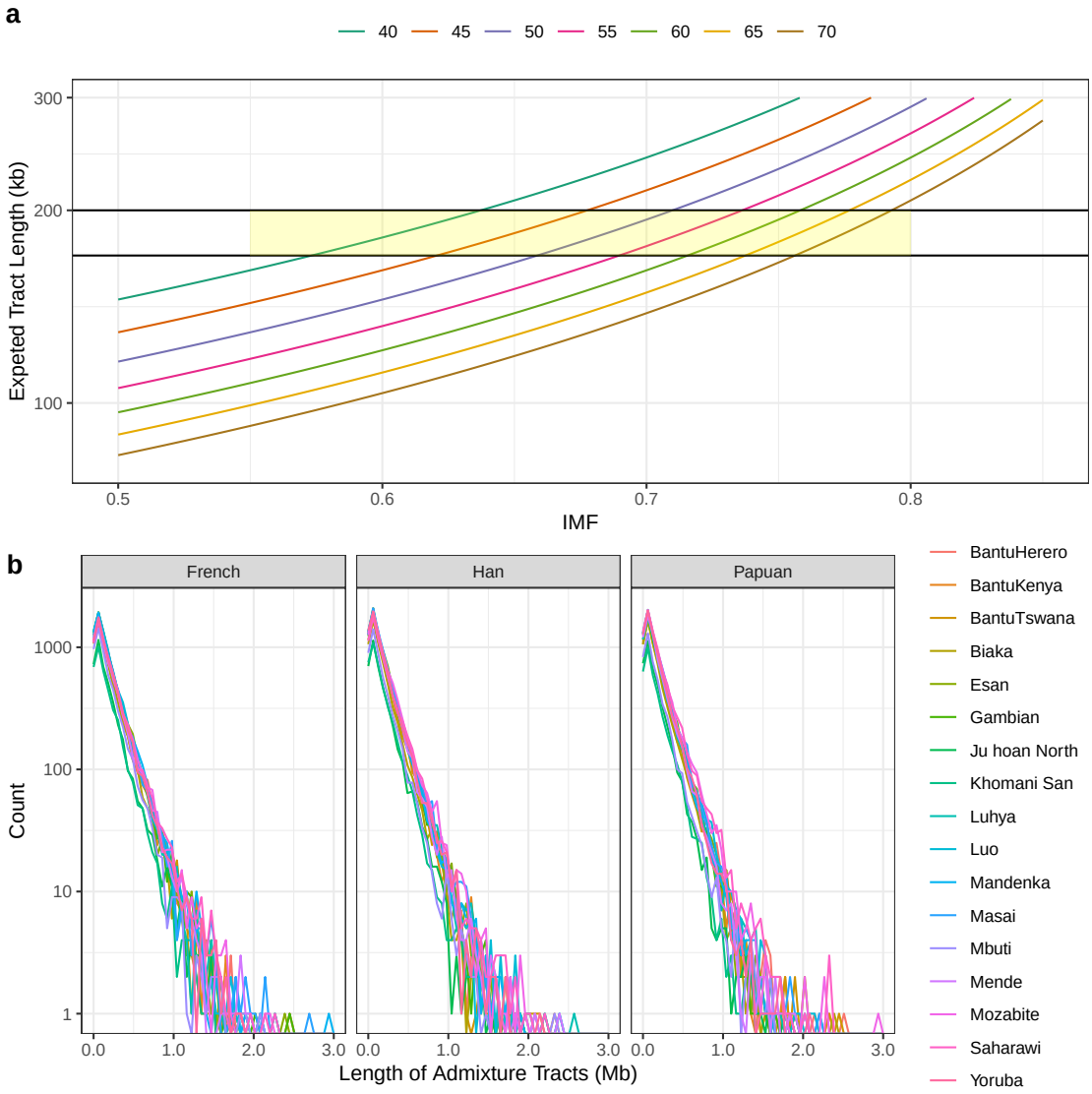


Figure 2.16: Analysis of the length of putatively migrated segments. **a.** Theoretical length distribution of admixture tract rate parameter under varying migration and admixture timing assuming that $L = ((1-m)r(T-1))^{-}1$, given L length, m symmetrical migration rate, r recombination rate in events per nucleotide, and T time in generations [159]. Shaded region denotes empirical range (with San at the bottom end, and Yoruban at the top end, Table 2.6) of fragments observed in the Simons Genome Diversity Panel. **b.** Following the reconstruction of the ancestral recombination graph using different African and non-African individuals using the `SMCSMC` particle filter, we use a sample of the posterior distribution of marginal trees to reconstruct putatively migrated segments (see Methods). We plot the length distribution of admixture tracts between individuals in the SGDP using 50 bins. Length is given in megabases (Mb). Isolated from `SMCSMC` estimated ancestral recombination graphs. Migration rate is given in terms of proportion of the sink population replaced per generation.

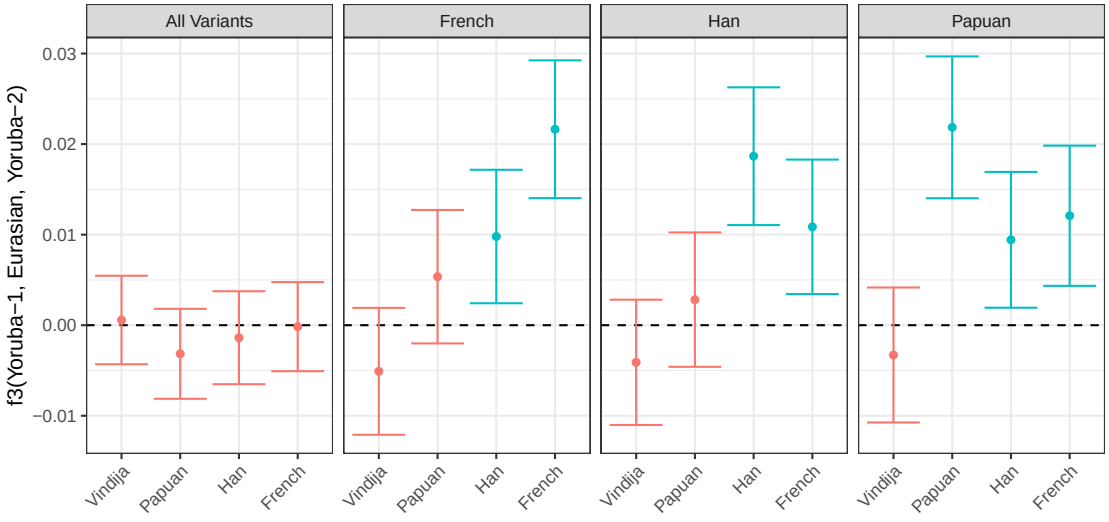


Figure 2.17: f_3 statistics show evidence for shared drift with Eurasians. Following the reconstruction of putatively migrated segments from the inferred ancestral recombination graph, we use the Reich Human Origins data set to investigate admixture using drift statistics using `ADMIXTOOLS` and `admixr` (see URLs). Tests are separated into all markers, and those ascertained through `SMCSMC` runs with the French, Han, and Papuans as a comparison Eurasian group. The f_3 statistics estimate is plotted, along with 95% C.I. computed via a block jackknife. Statistics which are significantly larger than zero are coloured blue, indicating shared drift.

also showing no increase in Neanderthal allele sharing ($f_3 = 0$). To find statistical evidence of admixture, we compute f_3(Yoruba-2, Eurasian, Yoruba-1) for the same Eurasian groups. We find statistical evidence for admixture in each of the groups examined, for all ascertainment schemes (Figure 2.18).

We use D statistics to examine more nuanced scenarios. We find that the two Yorubans share more alleles than other groups in Africa (D(African group, Yoruba-1; Yoruba-2, Chimp) is significantly negative with $|Z| > 3$), but the individual of interest is closer to Out of Africa (OoA) groups such as the Han, French, and Papauns (D(OoA, Yoruba-2; Yoruba-1, Chimp) is significantly negative with $|Z| < 3$) than to its partner Yoruban (Table 2.7). This implies that SMCSMC has identified segments of the African Genome which are more closely related to OoA populations than to fellow Africans.

In summary, we find no evidence for gene flow with a Vindija Neanderthal on the Mbuti baseline, or when compared to a different Yoruban (Table 2.8, Table 2.9). We

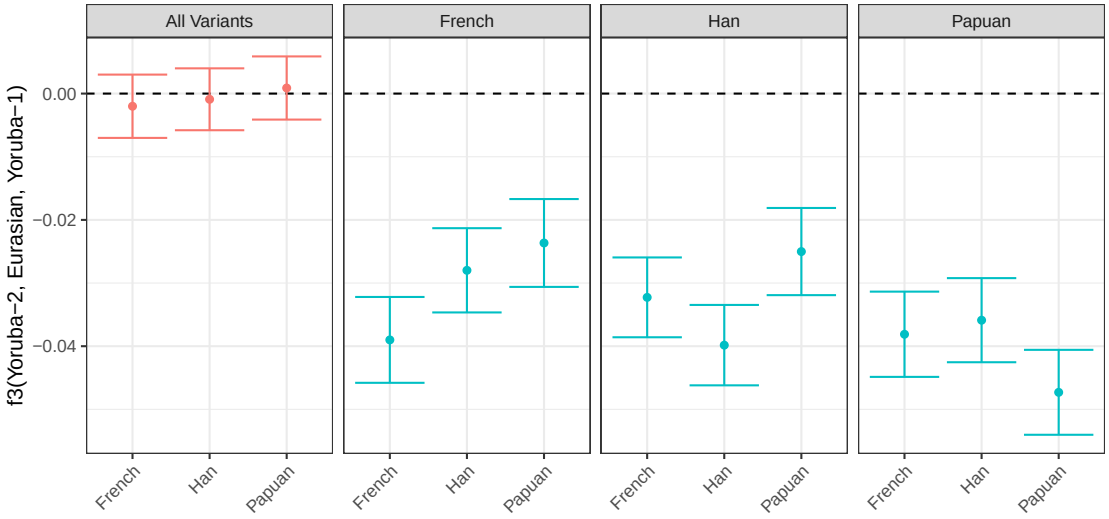


Figure 2.18: f_3 statistics show evidence for Eurasian admixture. Following the reconstruction of putatively migrated segments from the inferred ancestral recombination graph, we use the Reich Human Origins data set to investigate admixture using drift statistics using `ADMIXTOOLS` and `admixr` (see URLs). Tests are separated into all markers, and those ascertained through `SMCSMC` runs with the French, Han, and Papuans as a comparison Eurasian group. The f_3 statistics estimate is plotted, along with 95% C.I. computed via a block jackknife. Statistics which are significantly less than zero indicate statistical evidence for admixture, and are coloured blue.

additionally find no evidence for increased affinity to the Vindija Neanderthal when compared to the Altai, as would be expected if the material were descended from admixing Eurasians (Table 2.10). However, we find that restricted to the identified segments, D statistics have power to detect evidence for the known admixture from Vindija into a French individual (Figure 2.19), suggesting that lack of power does not explain the lack of evidence we find for Neanderthal admixture into Africans. In addition, we find no differences in affinity to Neanderthals or Denisovans between the variants which fall in segments and the whole genome (Figure 2.19d). Taken together, this suggests that Eurasian-derived segments of the African genomes are not enriched with Neanderthal material.

2.4 Discussion

We have developed an approach for estimating demographic parameters and ARGs from whole genome sequence data, which can handle inference in complex demo-

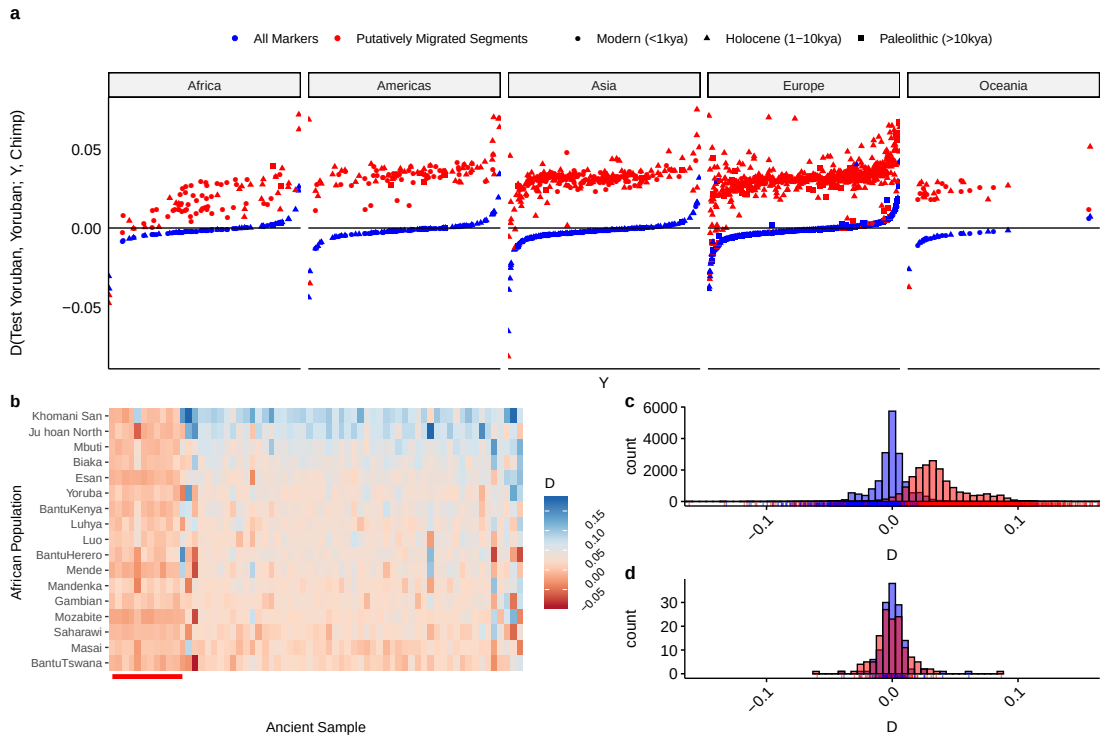


Figure 2.19: African introgressed segments are more similar to Eurasians but show no Neanderthal or Denisovan enrichment. **a.** D(Test Yoruban, Comparison Yoruban; Test population, Chimpanzee) calculated for all populations in the Reich Human Origins dataset (see URLs). D statistics in the putatively migrated segments are higher across the board in 3589 ancient, 6472 present day individuals. **b.** The same D statistic but computed for all African populations and individuals sampled from the Paleolithic. Neanderthal and Denisovan samples (marked with red bar) show low affinity to a Yoruban in putatively migrated segments. **c.** Histogram of D statistics computed in a. showing clear inflation of statistics calculated in segments (red) versus all markers (blue). **d.** Subset of individuals from a. involving Neanderthal ($n = 6$), Denisovan ($n = 1$), and a unique mixture individual ($n = 1$) with statistics calculated in segments (red) and all markers (blue) for all $n = 17$ African individuals indicating no difference in this population.

graphic models, and implemented this in the software program SMCSMC [154]. We used SMCSMC to investigate ancient migration rates and population substructure, and found evidence for a substantial admixture from ancestors of present-day Eurasian populations into African populations in the Late Middle Paleolithic.

Our analysis suggests that a population ancestral to present-day Eurasians contributed as much as a third of the genetic material in many modern African populations. We find no difference in inferred admixture proportions when using French Europeans or Han Chinese as extant representatives of the donor population,

indicating that the admixing population must have split from the out-of-Africa population before the East/West Eurasian divergence, implying a lower bound on the timing of the admixture of approximately 40kya [168]. It appears that our results suggest that the migrating population was more similar to present-day French and Chinese populations than to Papuans. However, up to 5% of the genomes of some present-day Papuans have been suggested to derive from archaic introgressions [170], and these contributions will have reduced the inferred levels of admixture into Africans when using Papuans as a representative of the Eurasian ancestors. The alternative explanation, of an earlier divergence of Papuans and Eurasian ancestors, is possible but contested; in light of documented Eurasian admixture into Oceania, the effects of this early isolation are likely to be small relative to the large confounding effects of Denisovan admixture [45, 147].

The proposed period of admixture has biased previous inferences of the African population sizes. We show that including directional migration into the model resolves previously unexplained high inferred N_e in the period 80 to 300kya. It is well known that effective population size estimates are biased in the presence of population substructure and migration [62, 155]. We use simulations to show that the proposed admixture event indeed causes an increase in estimated N_e in analyses that do not explicitly model migration. Correctly modeling of directional migration recovers the correct N_e, and allows us to infer a more recent split time between the two populations than indicated by previous analyses, although we did not attempt to formally estimate this time of divergence.

We found that not all populations in Africa have been equally affected by the proposed migration event. While the ancestors of Niger-Kordofanian and Nilo-Saharan populations show evidence of similar levels of Eurasian admixture, the ancestors of Central African and South African hunter-gatherer populations show markedly lower levels. The date of genetic diversification of both the Central Hunter Gatherers and Khoe-San (SAHG) is contested [141], but a date of 100kya has been proposed [171], providing a putative upper bound on the main admixture event. Our simulations indicate that SMCSMC has little power to detect the impact of

migration events occurring more than 70kya, providing an additional upper bound on the time of the migration episode, or the fraction of it that left a sufficiently distinct imprint on extant genetic material.

Compared to the remainder of the Niger-Kordofanians and Nilo-Saharans on the one hand, and the SAHG populations on the other, the Mbuti and Biaka show intermediate levels of admixture. Of these populations, the Biaka show slightly higher levels of admixture than the Mbuti, which is likely due to the well-documented admixture from Western African groups not shared with the Mbuti [172]. The lower levels of admixture in Mbuti and Biaka compared to Niger-Kordofian and Nilo-Saharan populations imply at least partial diversification of the former at the time of the migration, placing an upper bound on the timing. However, dating the diversification of these groups is difficult. Recent estimates using f statistics place the split concurrent with the San in a large-scale early expansion 200-250kya [141], while older data consistently report an earlier split time between 50 and 90 kya [173]. Further clarity on the early structure and diversification of hunter-gatherer populations are necessary to interpret their interactions with Eurasian migrants. The Afroasiatic populations on the other hand show high levels of admixture, which also appears to be of much more recent origin, and it appears likely that this is the result of extensive admixture from Eurasian populations during the Holocene [138, 165].

It has previously been suggested that Eurasian back-migration may be responsible for Neanderthal material in Africans [169]; however, we find no evidence for enrichment of Neanderthal-like material in putatively Eurasian-derived genomic segments in Africans based on a sample from the posterior distribution of ARGs. We expect that this sample ARG is representative of the underlying demographic process across the whole genome; we find no evidence of shared drift between African populations and representative Neanderthal samples in isolated admixture segments across the entire genome. This observation indicates that Neanderthal introgression into Eurasians occurred after the African introgression event we study here, or that further population structure in the Eurasian ancestral population precluded substantial transmission of Neanderthal material into Africa.

Our findings are consistent with several other published observations. Migration rate estimates using `MSMC-IM` revealed high levels of admixture at times comparable to our results [157]. The coalescent intensity function additionally shows similar histories between sub-Saharan African and Eurasian groups with high coalescent intensity in epochs consistent with our inference and those of `MSMC-IM`, supporting both an early split between the groups and a substantial replacement of genetic material more recently than $\sim$ 100kya [174]. Evidence has been mounting for multiple migrations into the Eurasian continent, possibly mediated by climatic drivers [134, 148]. Eurasian backflow during the Holocene has been well established [132, 175], but earlier migrations have also been proposed before based on observations of the spatial distribution of Y chromosome and mitochondrial haplogroups [176–182]. At the same time, evidence has been mounting for extreme heterogeneity in the history of sub-Saharan Africans, with several unsampled populations theorised to have contributed at various points in the past [130, 141, 150]. In light of these recent studies, the observations in this paper add to a growing body of evidence for complex population structure and migration surrounding the Out of Africa event leading to a substantial replacement of the African population in the Late Middle Paleolithic.

2.5 Acknowledgments

The SMCSMC project began in 2013, more than four years before I began my DPhil with Gerton Lunter. At that time, Sha (Joe) Zhu and Gerton Lunter developed the core of the analytical method and derived many of the mathematical innovations presented in the article (deriving waypoints, the extension of particle filters for continuous Markov jump processes). Donna Henderson devoted the majority of her DPhil to fine-tuning the method and performing many core aspects of quality control and validation. She mainly applied her work to an understanding of Neanderthal introgressions throughout history. My place in this project comes after Donna and Joe both completed their DPhils and postdoctoral fellowships respectively. I contributed substantially to the software implementation of the SMC2 project

as well as validation of the method, as well as the work presented in this thesis describing its application to inferring directional migration rates.

Name	ID	Source
French	S_French-1	SGDP
Han	S_Han-1	SGDP
Papuan	S_Papuan-1	SGDP
BantuHerero	S_BantuHerero-1	SGDP
BantuKenya	S_BantuKenya-1	SGDP
BantuTswana	S_BantuTswana-1	SGDP
Biaka	S_Biaka-1	SGDP
Dinka	B_Dinka-3	SGDP
Esan	S_Esan-1	SGDP
Gambian	S_Gambian-1	SGDP
Ju hoan North	S_Ju_hoan_North-1	SGDP
Khomani San	S_Khomani_San-1	SGDP
Luhya	S_Luhya-1	SGDP
Luo	S_Luo-1	SGDP
Mandenka	S_Mandenka-1	SGDP
Masai	S_Masai-1	SGDP
Mbuti	S_Mbuti-1	SGDP
Mende	S_Mende-1	SGDP
Mozabite	S_Mozabite-1	SGDP
Saharawi	S_Saharawi-1	SGDP
Somali	S_Somali-1	SGDP
Yoruba	S_Yoruba-1	SGDP
Druze	HGDP00562	HGDP
Han	HGDP00774	HGDP
Karitiana	HGDP01013	HGDP
PapuanHighlands	HGDP00549	HGDP
PapuanSepik	HGDP00542	HGDP
Pathan	HGDP00224	HGDP
Pima	HGDP01043	HGDP
Sardinian	HGDP00670	HGDP
Yakut	HGDP00946	HGDP
Yoruba	HGDP00930	HGDP
San	HGDP01029	HGDP
Mbuti	HGDP00450	HGDP
Biaka	HGDP00460	HGDP
Vindija	Vindija.DG	Prufer et al. 2017
Altai	Altai_published.DG	Pruefer et al. 2013
Denisovan	Deniosva_published.DG	Myers et al 2012

Table 2.1: Sample IDs of the individuals used in this article and relevant resources.

Language Family	Comparison Family	Mean Difference (95% CI)	Adjusted P
Niger-Kordofanian	Nilo-Saharan	-0.053 (-0.081–0.025)	1.11e-03
Niger-Kordofanian	Khoesan	0.143 (0.125-0.161)	4.39e-14
Niger-Kordofanian	Afroasiatic	-0.125 (-0.142–0.107)	1.67e-15
Nilo-Saharan	Niger-Kordofanian	0.053 (0.025-0.081)	1.11e-03
Nilo-Saharan	Khoesan	0.196 (0.169-0.223)	6.91e-09
Nilo-Saharan	Afroasiatic	-0.072 (-0.098–0.045)	1.23e-04
Khoesan	Niger-Kordofanian	-0.143 (-0.161–0.125)	4.39e-14
Khoesan	Nilo-Saharan	-0.196 (-0.223–0.169)	6.91e-09
Khoesan	Afroasiatic	-0.268 (-0.284–0.252)	3.62e-13
Afroasiatic	Niger-Kordofanian	0.125 (0.107-0.142)	1.67e-15
Afroasiatic	Nilo-Saharan	0.072 (0.045-0.098)	1.23e-04
Afroasiatic	Khoesan	0.268 (0.252-0.284)	3.62e-13
CAHG	Niger-Kordofanian	-0.084 (-0.119–0.049)	1.18e-03

Table 2.2: Differences in African IMF in the SGDP. The integrated migration fraction (IMF) in the epoch 30–70kya is calculated as per the Methods section for all comparisons in the Simons Genome Diversity Project (SGDP), and a two tailed t-test is used to statistically test for differences between the inferred migration in African language groups. Averaged over three technical replicates to account for the influence of stochastic sampling variation. P values corrected for multiple testing using the Bonferroni method. Abbreviations: CAHG, Central African Hunter-Gatherers (include Mbuti and Biaka); CI, Confidence Interval.

Language Family	Partner Population	Mean AFR IMF (SD)	Mean EUR IMF (SD)
Afroasiatic	All	0.477(0.015)	0.176(0.038)
Afroasiatic	French	0.477(0.006)	0.207(0.044)
Afroasiatic	Han	0.492(0.01)	0.168(0.032)
Afroasiatic	Papuan	0.462(0.011)	0.153(0.024)
Khoesan	All	0.209(0.013)	0.044(0.006)
Khoesan	French	0.217(0.007)	0.051(0.003)
Khoesan	Han	0.217(0.002)	0.042(0)
Khoesan	Papuan	0.193(0.006)	0.04(0)
Niger-Kordofanian	All	0.352(0.04)	0.088(0.017)
Niger-Kordofanian	French	0.36(0.039)	0.102(0.015)
Niger-Kordofanian	Han	0.363(0.04)	0.083(0.013)
Niger-Kordofanian	Papuan	0.334(0.038)	0.078(0.013)
Nilo-Saharan	All	0.405(0.033)	0.107(0.016)
Nilo-Saharan	French	0.414(0.037)	0.123(0.016)
Nilo-Saharan	Han	0.417(0.036)	0.106(0.01)
Nilo-Saharan	Papuan	0.385(0.029)	0.093(0.008)

Table 2.3: Integrated Migration Fraction (IMF) in either direction averaged over African language groups. SMCSMC was used to infer directional migration and effective population size between populations in the Simons Genome Diversity Project. The total migration between each African and non-African population was integrated in the epoch 30–70kya (see Methods) and averaged over language family. Abbreviations: Integrated Migration Fraction (IMF), Standard Deviation (SD), EUR (Eurasian), AFR (African)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.3807	0.0036	105.42	1.79e-47
Papuan	-0.0271	0.0017	-15.60	7.44e-18
BantuKenya	0.0051	0.0050	1.02	3.16e-01
BantuTswana	-0.0410	0.0050	-8.13	9.49e-10
Biaka	-0.0598	0.0050	-11.87	3.53e-14
Dinka	0.0160	0.0050	3.17	3.05e-03
Esan	-0.0016	0.0050	-0.32	7.51e-01
Gambian	-0.0073	0.0050	-1.44	1.58e-01
Ju hoan North	-0.1603	0.0050	-31.79	1.82e-28
Khomani San	-0.1653	0.0050	-32.80	5.96e-29
Luhya	0.0118	0.0050	2.34	2.46e-02
Luo	0.0123	0.0050	2.45	1.94e-02
Mandenka	0.0032	0.0050	0.63	5.34e-01
Masai	0.0719	0.0050	14.25	1.32e-16
Mbuti	-0.1171	0.0050	-23.22	1.17e-23
Mende	-0.0071	0.0050	-1.42	1.65e-01
Mozabite	0.1060	0.0050	21.02	3.63e-22
Saharawi	0.1097	0.0050	21.75	1.12e-22
Somali	0.0995	0.0050	19.73	3.12e-21
Yoruba	-0.0013	0.0050	-0.26	7.97e-01

Table 2.4: Linear model predicting integrated migration fraction in the SGDP. The integrated migration fraction (IMF) in the epoch 30–70 kya is obtained as per the Methods section in the Simons Genome Diversity Project. A binary variable representing Papuan / not Papuan Eurasian donor and categorical variable representing African population were used to predict the IMF in a simple linear model. When adjusted for the different African populations, Papuans contribute less IMF than do other Eurasian partners (French and Han).

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.3392	0.0037	92.92	1.72e-39
Papuan	-0.0252	0.0042	-5.95	1.43e-06
San	-0.1942	0.0050	-38.94	6.85e-28
Mbuti	-0.1410	0.0050	-28.27	1.07e-23
Biaka	-0.0883	0.0050	-17.69	9.33e-18

Table 2.5: Linear model predicting integrated migration fraction in the HGDP. The integrated migration fraction (IMF) in the epoch 30–70 kya is obtained as per the Methods section in the physically phased subset of the Human Genome Diversity Project. A binary variable representing Papuan / not Papuan Eurasian donor and categorical variable representing African population were used to predict the IMF in a simple linear model. When adjusted for the different African populations, Papuans contribute less IMF than do other Eurasian partners (French and Han).

African Population	Mean (SD)	Total (Mb)	Mean (SD)	Total (Mb)	Mean (SD)	Total (Mb)
Mbuti	172.204 (189.255)	939.371	165.822 (182.549)	850.335	163.352 (181.607)	756.320
Biaka	172.332 (184.183)	1057.083	169.296 (185.457)	1052.004	170.149 (185.853)	1044.036
Khomani San	173.06 (188.492)	671.125	169.005 (183.871)	706.777	166.33 (175.899)	597.125
Ju hoan North	175.415 (195.979)	747.618	171.909 (187.835)	711.186	161.146 (175.476)	656.024
Luhya	177.303 (193.164)	1370.729	181.025 (197.94)	1403.123	175.821 (192.259)	1211.762
Esan	177.483 (191.148)	1364.132	178.429 (190.819)	1426.180	170.451 (183.886)	1270.204
Gambian	177.804 (195.646)	1333.529	177.099 (191.542)	1304.155	173.892 (185.544)	1213.071
Luo	178.653 (192.447)	1368.481	175.618 (192.833)	1361.917	180.635 (201.159)	1348.798
BantuHerero	179.691 (194.451)	1354.869	176.473 (188.855)	1327.779	178.544 (198.496)	1310.511
BantuTswana	180.309 (195.176)	1150.551	174.826 (188.153)	1159.796	172.542 (192.6)	1096.674
Mende	182.087 (200.591)	1256.580	177.24 (197.23)	1288.533	169.901 (182.944)	1191.515
Yoruba	184.255 (199.949)	1283.151	176.276 (193.69)	1361.031	169.867 (181.948)	1275.701
BantuKenya	184.419 (200.395)	1265.297	173.485 (191.175)	1269.739	179.415 (189.784)	1214.104
Mandenka	185.35 (206.972)	1411.440	176.908 (194.337)	1328.754	172.03 (182.984)	1265.625
Masai	186.769 (209.675)	1353.890	182.366 (198.06)	1483.361	181.22 (199.255)	1438.160
Mozabite	194.378 (214.029)	1318.273	193.469 (211.945)	1557.040	188.867 (210.994)	1532.842
Saharawi	197.668 (218.662)	1383.280	196.156 (217.233)	1406.245	195.045 (217.263)	1585.907

Table 2.6: Summary of the length distribution for putatively migrated segments in different African individuals. Means and standard deviations are given in kilobases (kb) while the total length of all segments is given in megabases (Mb).

Statistic	D	Z
D(KhomaniSan-1, Yoruba-1, Yoruba-2, Chimp)	-0.181	-28.403
D(Mbuti-1, Yoruba-1, Yoruba-2, Chimp)	-0.135	-19.554
D(Papuan-1, Yoruba-1, Yoruba-2, Chimp)	-0.026	-3.422
D(French-1, Yoruba-1, Yoruba-2, Chimp)	-0.006	-0.866
D(Han-1, Yoruba-1, Yoruba-2, Chimp)	0.001	0.072
D(KhomaniSan-1, Yoruba-2, Yoruba-1, Chimp)	-0.187	-28.109
D(Mbuti-1, Yoruba-2, Yoruba-1, Chimp)	-0.130	-19.323
D(Papuan-1, Yoruba-2, Yoruba-1, Chimp)	-0.008	-1.003
D(French-1, Yoruba-2, Yoruba-1, Chimp)	0.030	4.355
D(Han-1, Yoruba-2, Yoruba-1, Chimp)	0.056	8.037

Table 2.7: Putatively migrated segments of a Yoruban are closer to Out of Africa groups than a comparable Yoruban.

Statistic	D	Z
D(Mbuti-1, Yoruba-1, Vindija, Chimp)	-0.001	-0.141
D(Mbuti-1, Yoruba-2, Vindija, Chimp)	-0.003	-0.306

Table 2.8: No difference in allele sharing with Vindija Neanderthal over Mbuti baseline.

Statistic	D	Z
D(Yoruba-2, Yoruba-1, Vindija, Chimp)	0.000	0.012

Table 2.9: No difference in allele sharing with Vindija Neanderthal.

Statistic	D	Z
D(Vindija, Altai, Yoruba-1, Chimp)	0.024	1.095
D(Vindija, Altai, Yoruba-2, Chimp)	0.034	1.526
D(Vindija, Altai, Mbuti-1, Chimp)	0.002	0.103
D(Vindija, Altai, KhomaniSan-1, Chimp)	0.023	1.008

Table 2.10: No increased affinity to Vindija Neanderthal over Altai, as would be expected if the source of any Neanderthal ancestry was Eurasian.

It's easier to be terrified by an enemy you admire.

— Frank Herbert, *Dune*

3

Identifying co-accessible regulatory regions using topic modelling

Contents

3.1 Introduction

The physical accessibility of CREs in part determines which and how many transcription factor proteins are able to bind to the DNA. On the other hand, the binding of certain transcription factors, especially pioneer factors, has the ability to alter the accessibility of the chromatin [183]. The complex and dynamic interplay between these opposing forces regulates the process of transcription, controlling the expression of genes in a dynamic and contextual way [184, 185]. Recent methods have permitted the profiling of chromatin accessibility on a genome-wide scale, however the role of the physical compaction of the genome and specific CREs remains a poorly understood predictor of cell identity outside of niche model systems [186]. Understanding the dynamics of chromatin accessibility both between and within cell systems is of relevance to understanding the effect of sequence mutations that disrupt Transcription factor binding sites (TFBSs) and which alter the availability of the entire CRE. Additionally, a thorough catalog of relevant and distinct accessible elements within a pathological cell system would allow for improved prioritization of dysregulated transcription factors and their genomic consequences. In this chapter, I aim to improve the characterization of regulatory programs through modelling chromatin accessibility with high throughput sequencing. The resulting method, BLDA, represents a viable approach to identify key regions of accessible chromatin that are both shared between similar cell types and discriminatory of others. I show that this method has similar power to identify

important regions when compared to single cell ATAC-seq, and demonstrate its use on the well-characterised developmental stages between hematopoietic stem cells (HSCs) and mature erythrocytes. The results here demonstrate that topic modelling is a reliable method for general purpose discrimination of important accessible regions in arbitrary collections of bulk ATAC-seq experiments.

3.1.1 Transcriptional regulation through chromatin accessibility

Beginning with a single copy of the diploid genome, sequential cellular differentiation creates upwards of ∼40 trillion individual cells with unique functions across organ systems and functional niches [187]. Each of these cells contains essentially the same genetic code yet performs entirely different roles in the body, indicating the presence of an immensely intricate regulatory process dictating how genes are expressed in certain cellular contexts. As explored in Section 1.3.1.1, a portion of this cell type specific regulation of expression is explained by DNA binding proteins, yet where and how these proteins are able to bind is dictated by the physical accessibility of the underlying sequence [95]. Within the nucleus, DNA is packaged into a highly compact and organized structure known as chromatin. The construction of chromatin involves wrapping molecules of DNA around histone proteins. Typically, 147 base pairs of DNA wrap around an octomer of histone proteins to form the fundamental subunit of chromatin called a nucleosome [183]. Nucleosomes can act as a barrier to RNA polymerase II mediated transcription as well as many transcription factors [99]. The exception to this rule are pioneer factors, which bind instead to closed chromatin and recruit chromatin remodellers to alter its accessibility, priming it for functional activities and the binding of other transcription factors [183, 188]. Other factors such as the post-translational modifications of histone proteins and higher order organization of chromatin also contribute to functional chromatin accessibility [105]. As a consequence of these facts, nucleosomes tend to be found at lower densities within active regulatory regions; this makes the relative accessibility of a region an indicator of its regulatory capacity [99].

The accessible genome comprises between 2 and 3 percent of the actual sequence in humans, but represents approximately 90 percent of regions bound by Transcription factor (TF) [189]. The accessibility of a TF binding site does not in and of itself dictate that it will bind, but represents a regulatory capacity at the accessible site [99]. TFs bind to sequence competitively with histones and other chromatin binding proteins to dynamically modulate the organization and placement of nucleosomes [190]. While closed chromatin is generally not amenable to binding by TFs other than the previously described pioneer factors, permissive chromatin is sufficiently dynamic to allow for some binding [93]. Accessibility therefore exists on a continuum that is dynamically reorganized in part based on the cellular context and its battery of expressed transcription factors. Profiling the accessible regions of the genome in a particular cellular context provides a window into active regulatory programs.

Several methods exist to experimentally determine chromatin accessibility. The majority of recent methods use an enzymatic reaction to selectively fragment the genomic sequence and next-generation sequencing to comprehensively survey the genome for the enrichment of fragments. Examples of these approaches include DNAse-seq, ATAC-seq, MNase-seq, FAIRE-seq, and NOMe-seq, reviewed in Klemm, Shipony, and Greenleaf [185] and Meyer and Liu [191]. Of these, DNAse-seq is the most sensitive but requires a large number of cells to generate reliable libraries for sequencing [192]. ATAC-seq is a reliable method for characterising accessibility, applicable to systems with as few as 500 cells for bulk, or even on a per-cell basis [193, 194]. Datasets generated with ATAC-seq have been growing exponentially year over year, representing in 2019 several times more data generated than any other approach for assaying chromatin accessibility [195]. Thus, methods for the interpretation of large compendiums of ATAC-seq form a useful complement to the method's growing adaptation.

3.1.2 Regulation of key stages within hematopoiesis and erythropoiesis

Of the ∼40 trillion cells in a typical human body, approximately 90% derive from the hematopoietic lineage [187]. The process of differentiation from HSCs to mature blood cells is therefore of immense importance to health and disease. In adults, hematopoiesis begins from HSCs which exist in a relatively quiescent state within the bone marrow, maintaining the capability for self-renewal and multipotenency [196]. Because this differentiation process occurs throughout life, hematopoiesis represents an ideal model system to understand the properties of stem cell biology, the dysregulation of which is intricately linked with oncogenesis [197]. These cells give rise to all other phenotypically distinct blood cells through a hierarchical cascade of differentiation (Figure 3.1). Several intermediary points in this process are relevant to this chapter, including the initial priming of the stem cells into multipotent progenitor cells (MPPs). This involves, among other things, active polycomb repression of lineage specification genes such as *Ebf1* and *Pax5* in MPP cells [198]. Lineage commitment represents a decision point, where continuing specification will exclusively occur within a specific branch of the hematopoietic differentiation tree. In this chapter, we use erythropoiesis as a model system to biologically validate the model we introduce.

The regulation of chromatin accessibility plays a central role in the differentiation trajectory of erythrocytes [186, 199]. Erythropoiesis refers to the process of successive differentiation from pluripotent stem and progenitor cells through several morphologically and functionally distinct stages to form enucleated erythrocytes (also known as red blood cells). Here we focus on the process up until erythroblasts, the penultimate step in erythropoiesis and the last before enucleation. Ludwig et al. [200] used FACS sorting on CD71, CD235a, CD49d, and BAND3 surface markers to identify eight distinct stages of development. Their analysis using both ATAC-seq and RNA-seq is the most complete representation of this differentiation trajectory to date. The authors identify myeloid progenitors (MyP), colony forming units - erythroid (CFU-E), pro-erythoblasts with two stages (ProE1,

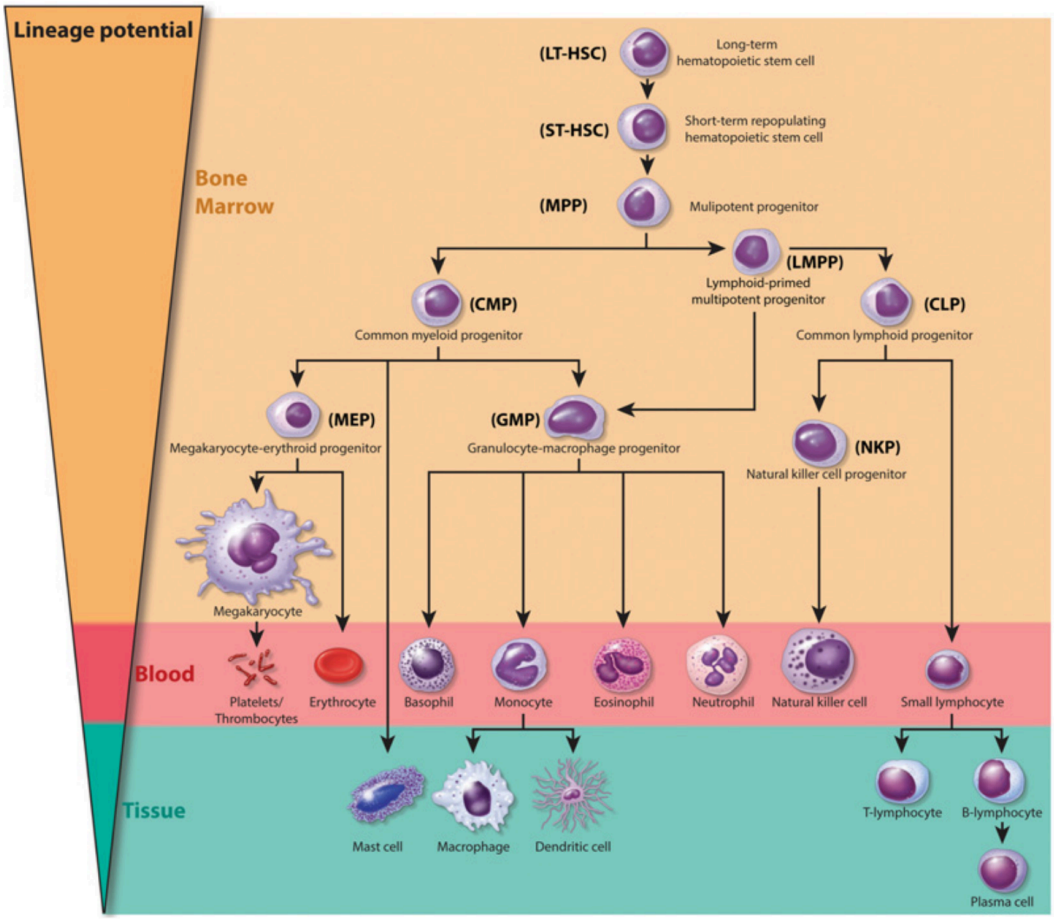


Figure 3.1: Summary of important cell types arising during hematopoiesis from HSCs to phenotypically and functionally distinct cell types. Adapted from Hu and Shilatifard [199]. HSC = hematopoietic stem cell.

ProE2), basophilic erythoblasts (Baso-E), polychromatic erythroblasts (Poly-E), orthochromatic erythoblasts (OrthoE), and orthochromatic reticulocyte (Ortho/Ret) as key cell stages. A detailed explanation of the molecular biology of each stage is not the focus of this thesis, however interested readers may consult texts such as Sinclair [201]. After augmenting their dataset with hematopoietic progenitor cells created by Corces et al. [202], the authors found that accessible elements clustered into several groupings. Some were predominantly active in the earliest stages of hematopoiesis, while others were broadly accessible across lineage commitment and intermediate erythropoiesis, while others still acted primarily in terminal erythropoiesis. These broad groupings allowed the authors to identify several factors of interest which

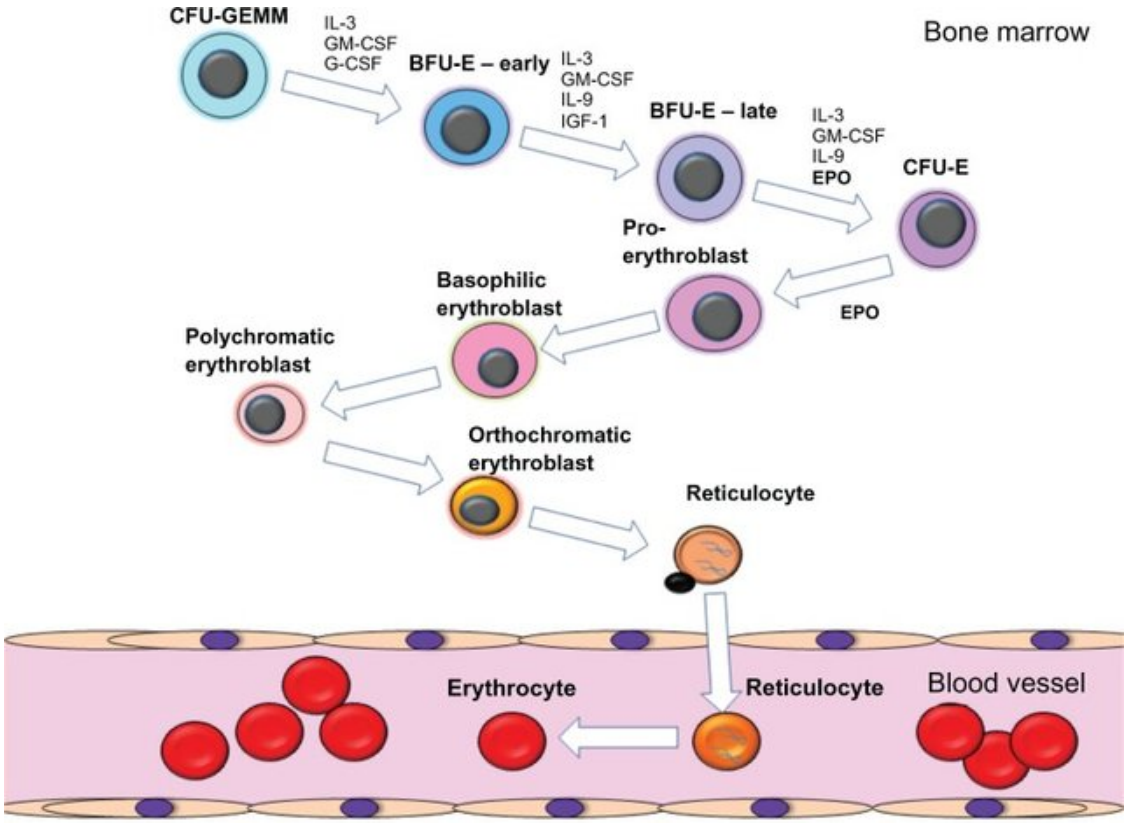


Figure 3.2: Key stages of erythropoiesis from Sinclair [201]. Blast forming units are not represented in the Ludwig et al. [200] dataset, while erythroblasts replace erythrocytes as the end point of the process for our purposes.

may act differentially; we investigate these factors such as *TMCC2, UROS*, and *RHAG* as well as other well known markers like *TAL1, KLF5* and *GATA1* later in this chapter. In this chapter, we recreate this dataset by augmenting the eight erythroid cell type presented in Ludwig et al. [200] with the hematopoietic stem and progenitor cells from Corces et al. [202] to create a dataset of chromatin accessibility throughout erythropoiesis.

3.1.3 Identifying regulatory programs in large databases

The task of identifying co-accessible sets of regulatory elements usually falls on differential peak accessibility analyses [195]. Many methods exist for this task, including MACS2, DiffBind, csaw, voom, limma, edgeR, and DESeq2 [118, 203–208]. A comprehensive review of their relative performances was undertaken by Reske, Wilson, and Chandler [209]. In general, these methods work by grouping similar

experiments, usually taken to be biological replicates, and finding peak regions whose deviation overcomes some level of statistical significance. These tools are highly refined and benchmarked for applications with two cell types of interest. However, using differential accessibility in large systems with multiple consistent subpopulations poses technical issues. As an example, edgeR is able to estimate coefficients for a general linear model and find significant differences in accessibility for a particular cluster. However, the estimation of statistical significance relies on estimating variance with biological replicates, and the user is warned not to attempt significance testing without performing replicates [206]. Though this places no hard constraint on the data to be analysed, clustering in large datasets is imperfect. The estimation of the variance is therefore critically reliant on the homogeneity of the cluster, which may vary between clusters. This makes the interpretation of differentially accessible elements between clusters in large datasets difficult. Secondly, it is difficult to use differential accessibility testing to find patterns unique to combinations of clusters, as the estimated coefficients tend to relate to enrichment in a specific cluster, or between all of them. One alternative is to look at the pairwise differential accessibility between clusters and search for patterns in the identified regions. This is, however, onerous and to our knowledge no dedicated method exists for the task.

Recently, Bravo González-Blas et al. [123] proposed the use of topic modelling to study collections of accessible regions in single cell Single cell ATAC-seq (scATAC-seq). Topic modelling, specifically LDA represents a viable method for the unsupervised identification of key regions of accessible sequence in large databases, while simultaneously identifying the distribution of their associated regulatory programs. The method learns groupings of differentially accessible elements and where they tend to be active, in aggregate. However, the use of LDA in large collections of ATAC-seq data has not been explored. As the amount of available ATAC-seq data grows year over year, analyses of large compendiums of cellular variation in accessibility form a useful complement to sparse single cell analyses. On the one hand, single cell analyses are well powered to identify fine-scale groupings of regulatory elements

active in sub-populations of similar cell types. On the other, analysis of dense bulk ATAC-seq with high read depths allows for a thorough investigation of pathways active across large groupings of cells. In this chapter, we investigate the use of LDA for identifying regulatory programs and their distribution in bulk ATAC-seq.

3.1.4 Latent Dirichlet Allocation

A detailed methodological motivation of LDA is beyond the scope of this chapter, however here I present in general terms a typical formulation of the generative model and inference from real data.

3.1.4.1 The generative model

LDA is interested in the generation of a corpus of D documents, composed of n words. It was introduced in Blei, Ng, and Jordan [210]. In brief, each document $d = 1 \dots D$ can be described as a collection of weights on k topics. An example in the realm of natural language processing would be books and their respective genres (i.e. fiction, science, etc.). In the specific application here presented, the corpus represents the dataset of sequencing experiments, with each document representing a single ATAC-seq sample. The words are individual accessible regions, which are determined and discretized through peak calling on respective samples. The inference procedure is unsupervised, and there is no annotation or metadata given to discriminate amongst samples, unlike later adaptations like structured topic modelling [211].

The generative process is as follows. For a fixed number of topics k, each document d is given a distribution on its topic weights θ_d

$$\theta_d \sim Dir(\alpha), d = 1, \dots, D$$

Each word n in document d is generated from a particular topic z_{dn} where z_{dn}

$$z_{dn} \sim Discrete(\theta_d)$$

and the word itself is chosen from

$$p(w_{dn}|z_{dn}, \zeta)$$

which is a multinomial probability conditional on the topic allocation z_{dn}. The other dependency is on ζ, here defined as a $k \times N$ matrix where $\zeta_{ij} = P(w_j = 1|z_i = 1)$, or the conditional probability of a word given a topic allocation. This is a fixed quantity that will be estimated.

The Dirichlet distribution is convenient in this case, as draws from $Dir(\alpha)$ represent points on the $(k-1)$ simplex where parameter α is a k-vector of positive reals. A simplex is a multidimensional generalization of the triangle. The parameter vector α is also known as the concentration parameter which, in the case of symmetric Dirichlet distributions where $\alpha_i = \alpha_j \forall i, j \in 1 \dots k$ determines the sparsity of the draws from the distribution. In the limit of extremely high positive values of the concentration parameter, draws from the distribution tend to be approximately uniform draws over the probability space, while at values less than one and approaching zero the distribution is degenerate [212]. In addition, the Dirichlet distribution is the conjugate prior to the multinomial distribution, a fact which is exploited for efficient inference algorithms. Explicitly, the probability density function of the Dirichlet distribution here considered is given by

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \dots \theta_k^{\alpha_k - 1}$$

which leads to an overall joint distribution and probability of an overall corpus as given in Blei, Ng, and Jordan [210].

This traditional set up for LDA does not allow for any control over the sparsity of word allocations to topics. For that, an extension called Smoothed LDA replaces the ζ matrix of conditional word probabilities with a second Dirichlet distribution, parameterized by β. As a side note, all the applications of LDA in this chapter are concerned with symmetric Dirichlet distributions, where each of the entries of the α and β k-vectors are the same. This modified generative procedure proceeds identically, except that a seperate draw is made for each topic representing the distribution on its words from

$$\psi_k \sim Dir(\beta), k = 1, \dots, K$$

, and drawing a word conditional on its topic assignment is instead generated as

$$w_{dn} \sim Discrete(\psi_{z_{dn}})$$

where $\psi_{z_{dn}}$ intuitively represents the topic-word distribution. In the case here considered, it represents the importance of a particular region of accessible chromatin to a regulatory program z.

3.1.4.2 Parameter inference

The joint distribution of latent parameters θ, z conditional on w, α, β does not admit analytical inference as the posterior probability is intractable, as noted in Blei, Ng, and Jordan [210]. They proposed an approach based on variational Bayesian inference, however it has become commonplace to integrate out θ and approach the problem of inference using collapsed Gibbs sampling [213–215]. Gibbs sampling is an extension of the MCMC approach for sampling from the posterior distribution when, as is the case here, analytical inference is not possible. Gibbs sampling is closely related to the Metropolis-Hastings algorithm, and relies on constructing a Markov chain with the previously described posterior as its equilibrium distribution. In this way, inference about parameters is deduced from direct samples of the posterior. This is philosophically very similar to the approach taken in Chapter 3, where samples of the posterior distribution of genealogies conditioned on mutations were used to draw inference about demographic parameters such as directional migration. The details of the inference procedure are not of direct relevance to the results presented in this chapter, and a more detailed description can be found in Qiu et al. [213] and others.

Here, the posterior distribution of interest concerns the most likely values of θ and ψ, that is the topic weight vectors on each of the different ATAC-seq experiments and the topic distribution on each of the constituent accessible regions, conditioned on observed regions within ATAC-seq experiments and given parameters α, β.

3.1.5 The LDA algorithm and cisTopic

The basis of this chapter is the cisTopic algorithm, and some expansion on its basic formulation is necessary. cisTopic is introduced in Bravo González-Blas et al. [123], and is primarily intended for use on scATAC-seq data resulting from a combination of sequencing data and associated peak calls. In general, this data is either encoded in a count matrix or it is internally converted to one. A count matrix M, for the purposes of this chapter, refers to a $C \times R$ matrix where element M_{cr} equals the number of reads (or fragments) overlapping region r in cell c, for regions $r = 1, \ldots, R$ and cells $c = 1, \ldots, C$. Each region r is selected on the basis of statistical peak calls, typically performed with software such as Macs2 or similar, on the aggregated scATAC-seq signal. This count matrix M is then subjected to binarization on the basis of some threshold T, where T is the minimum number of counts necessary for a region to be declared accessible. This threshold reflects the low sequencing depth of scATAC-seq data, and the difficulty in comparing cells quantitatively based on read counts alone without some correction for depth. This threshold is set to 1 by default in the stable version of the algorithm implementation. It is necessary to provide symmetric hyperparameters α and β, which cisTopic prefers to normalize by the total number of topics and provide an α-per-topic value rather than an α-per-sey value with which to do inference. We follow this convention in this thesis, providing unnormalized α values whenever relevant and raw β values.

The corrected count matrix is used for the inference of the cell-topic distribution and the region-topic distribution, previously represented by θ and ψ respectively as the distribution of topics over documents and the distribution of words over topics. Topic loadings for both the cell and region loadings are normalized to the range [0, 1]. Key regions which are important to the topic may be selected in one of two ways. Firstly, by fitting a gamma distribution to the normalized region-topic loadings on a per-topic basis and selecting a percent point threshold of the resulting density's tail (i.e. the top 1 percentile of the fit gamma distribution). Alternatively, a given number of top regions may be selected based on the rank of the region-topic loadings.

Selecting the right number of topics for a particular analysis is not straight forward. cisTopic implements an optimized version of collapsed Gibbs sampling by using WarpLDA, an algorithm for constant time inference [216]. An advantage of WarpLDA is that it returns the second derivative of each value for k. As log-likelihood values increase with increasing k, cisTopic makes use of these values to automatically select a value for k based on a range given by the user. A proof that this procedure produces optimal values of k is not readily available.

3.1.6 Aims of this chapter

The overarching aim of this chapter is to investigate the use of LDA for bulk ATAC-seq. As noted above, the method represents a significant advancement above differential accessibility, and shows theoretical promise for the analysis of the growing collections of sequencing data in diverse cell systems. In doing so, I adapt the existing cisTopic method for bulk samples. Specifically, I investigate the performance of bulk LDA when compared to established scATAC-seq inference using cisTopic. I also investigate whether the method infers meaningful topic loadings on a well understood system with ground truth values to compare against, erythropoiesis.

3.2 Methods

3.2.1 Single Cell ATAC-seq Dataset Generation

A single cell ATAC-seq dataset was compiled from data generated in Buenrostro et al. [194] for three cell types: K562, GM12878, and H1ESC. These cell types were selected as a subset which represents maximal diversity in accessible chromatin within the larger dataset.

For each labelled cell type, we collected accession numbers within the NCBI sequence read archive. These include records SRR1780163 through SRR1780354 (K562), SRR1779683 through SRR1779778 (GM12878), and SRR1779589 through SRR1779683 (H1ESC). Sequencing reads were merged and adapters were trimmed with cutadapt v 2.10 using the following adapter sequences (`-a CTGTCTCTTATACACATCT -A CTGTCTCTTATACACATCT`) [217]. Quality of the merged dataset was verified with

fastqc and aligned to hg19 using bowtie2 [218, 219]. Cell specific barcodes were added to the resulting alignment file within the `CB` tag using pySam [220]. Peak calling was performed using MACS2 and LanceOTron using default parameters [119, 221] as detailed in subsection 3.2.3 and count matrices were constructed as detailed in subsection 3.2.4. A more thorough discussion of our choice to compare MACS2 and LanceOTron may be found in Section 1.3.2.1.

3.2.2 Construction of Pseudo-bulk ATAC-seq Dataset

In order to construct a pseudo-bulk dataset, entries for each of the single cells were merged into a single alignment file. Cell barcodes were replaced with a cell-type marker and peak calling was similarly conducted with macs2 and LanceOTron [119, 221].

3.2.3 Peak Calling from Coverage Data

We compare two methods of peak calling.

A public implementation of the LanceOTron can be found at `https://github.com/chris1221/lanceotron`. LanceOTron relies on a pre-trained neural network. We use the supplied weights in `wide_and_deep_fully_trained_v5_03.h5` along with the standard scaler values from the same implementation. We allow the network to identify candidate peaks and assign a peak score, thresholding our selected peaks on a peak score of at least 0.5 as in the implementation of LanceOTron at `https://lanceotron.molbiol.ox.ac.uk/`.

3.2.4 Running LDA with cisTopic

We used the implementation of LDA in cisTopic [123]. cisTopic is intended for use on single cell ATAC-seq experiments, however the input format is amenable to any quantitative data observed on a cell by region basis. To construct the input data to cisTopic, we create a bespoke pipeline that firstly calls peaks from ATAC-seq experiments, and secondly harmonizes these peaks while constructing a count matrix. A count matrix is a sparse matrix where the rows are cells, or

in the case of this thesis, cell types in the form of ATAC-seq samples, and the columns are individual peak regions.

3.2.5 Bayesian Hyper-parameter Optimization

LDA requires a set of three hyperparameters to be supplied, alpha, beta, and the number of topics. This chapter investigates applications of LDA to problems of several scales, so the choice of hyperparameters is not easily chosen, at least in an unbiased way. We approach this issue by using a Bayesian Optimization approach to learn the best set of hyperparameters for a given set of ATAC-seq experiments. We wrote a typical cisTopic analysis in python via rpy2 (`https://rpy2.github.io/doc/latest/html/index.html`) and used the BayesianOptimization library (`https://github.com/fmfn/BayesianOptimization`) to optimize a target function for a given set of hyperparameters. This task was facilitated by the use of Dask to compute several hundred possible combinations simultaneously [222]. The target function to optimise is based on the specific application, as discussed in section 3.3 section.

3.2.6 Bulk LDA (BLDA) Method

The BLDA method is a small extension to cisTopic which aims to incorporate a proxy of how accessible each peak region is in a particular experiment. This makes it in some ways more similar to the traditional approach to LDA within natural language processing, which gives an integer value for the number of times a particular word appears in a text. To create the RPKM normalized count matrix and run the inference, we use functions from the `blda` python package available on Github at `https://github.com/Chris1221/blda`. The construction of the count matrix requires linked coverage files and their associated peak calls. The pipeline is agnostic to the type of coverage file (either BAM files or RPKM normalised bigWig tracks can be provided) and peak calls can be made by any software which outputs delineated, non-overlapping genomic regions. Hyperparameter optimization is conducted according to Section 3.2.5 and a modified version of cisTopic is run using

`rpy2`. The modification to cisTopic, found at `https://github.com/Chris1221/cisTopic`, relaxes a constraint to provide the `lda` package's collapsed Gibb's sampler with a binary sparse matrix. Several other functions from the cisTopic function are applicable directly to the situation of bulk LDA after the data has been altered.

3.2.7 Computing the average fold enrichment of topics for groups

To get a measure for the number of topics to use in a particular model, we compute the average fold enrichment. This is done with a two-step process. First, a one-tailed student's T test is performed for each of the grouping of cells. The cells are split into one of the cell types, with the remainder forming the comparison group. The null hypothesis here is that the means do not differ, or that the comparison group is significantly smaller than the group under question. The P values are corrected for the number of comparisons using Bonferroni correction (multiplying by the number of effective tests, which in this case is the number of tests) and the significant associations are reported. Second, the differences in the estimated means are divided to form a fold change metric amongst the subset of significant topic associations. We take the median of these values to form the average fold change metric for a particular set of topics being modelled. The median is selected to avoid extreme results overly biasing the statistic.

3.2.8 Co-enrichment of topics

The number of times two topics i and j are co-enriched is defined to be the count of occurrences where a single cell type has a topic loading for i and j exceeding some pre-defined threshold. After some experimentation, we decide to use a threshold of 0.25, where the two co-enriched topics collectively form half of the normalised topic loadings.

3.3 Results

LDA has previously been shown to be highly effective at detecting true patterns in shared accessibility in single cell data. Here, I create a dataset of known cell types in order to set baseline expectations about the capabilities of LDA in a single cell system before moving on to studying bulk samples.

3.3.1 Bulk LDA recapitulates patterns from single cell ATAC-seq

I generated a dataset of three labelled cell types, as per Section 3.2.1. This dataset comprises three cell types which have substantial differences in their accessible DNA. K562 is an erythro-leukemic cell model of chronic myelogenous leukemia from a fifty-three-year old female donor [223]. GM12878 is an Epstein-Barr Virus (EBV) immortalized B-lymphocyte cell line cultured from a female member of the international HapMap project CEPH panel [224]. Finally, H1ESC are a human embryonic stem cell line collected from a healthy male donor of unknown age [225].

The dataset comprises 96 GM12878 cells with a total of 1928090 reads, 96 H1ESC cells with 4192100 reads in total, and 192 K562 cells from two biological replicates with a total of 5269388 reads.

We generate coverage tracks for the merged single cell dataset using deepTools and peaks are called using both MACS2 and LanceOTron as per Section 3.2.3 [226]. MACS2 identifies 10,075 statistically significant peak regions, while MACS2 finds 12,994, 8337 of which are common between the two approaches.

3.3.1.1 Identifying differentially accessible peaks with EdgeR

The R package `edgeR` is used to set a baseline expectation for differentially accessible regions of the genome between the clusters. `edgeR` is selected as it was recently shown to have the overall highest sensitivity for identifying differential peaks from properly normalized ATAC-seq data. We took the single cell dataset and created a count matrix with all true counts across different cells. We inputted these into edgeR and estimated dispersion coefficients as recommended in the tutorial, fitting

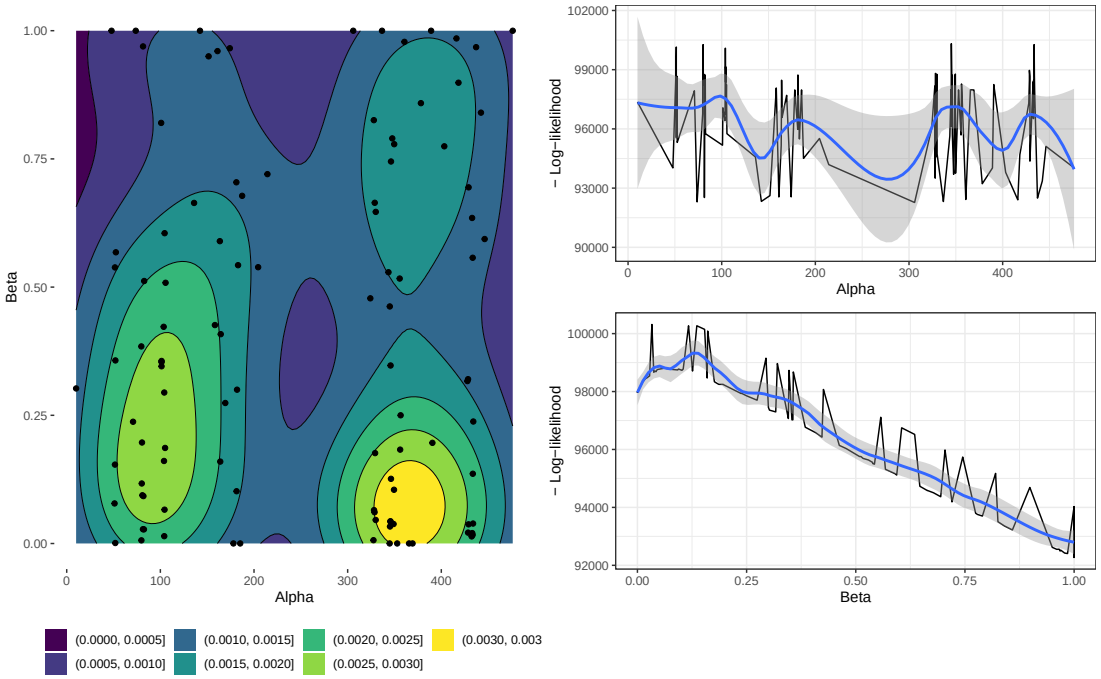


Figure 3.3: Hyperparameter log-likelihood surface for two hyperparameters. Alpha and beta parameters are given as symmetric hyperparameters of the Dirichlet distribution, where alpha is normalized according to the number of topics.

a generalized linear model according to the true cell identities which assume are known in this case. We identify the top 100 differentially accessible regions based on this analysis, and use these to study the results from our topic modelling approaches.

We investigate topic modelling in three separate scenarios. Firstly, we ran cisTopic as intended, using the single cell data. Secondly, we created pseudo-bulk alignment files for each of the clusters and run cisTopic normally. Thirdly, we adapted cisTopic to accept normalized read count. This allowed us to validate our baseline assumption that LDA is able to capture similar information within bulk sequencing experiments.

3.3.1.2 LDA captures realistic regulatory patterns in known single cell systems

For the first case, we model the system of 384 cells with 4, 5, 6, 7, and 8 topics to assess the best biological fit to the system. For each pre-set number of topics, hyperparameters are selected using Bayesian optimization. The best value for α

was not well estimated, with considerable noise and no clear peak; this is in contrast to the β value, which showed higher log-likelihoods at higher values approaching 1 (Figure 3.4). Consistent with this, a high value is selected for β in each case. β controls the number of regions loaded to a topic. With a number of cells much larger than the number of topics, especially in the present scenario, a high value of β represents a prior on more regions loaded to each of the small number of topics. This indicates that a relatively high number of regions are differentially accessible between the different clusters. The relatively uncertain distribution of alpha log-likelihoods represents an ambiguity in the likelihood from different Dirichlet distributions. Draws with a low alpha tend to represent topic loadings which are specific to individual cells, and accordingly the flat likelihood surface may indicate that there are multiple ways of assigning the topics to cells with relatively equal likelihoods.

We investigate the loading of these topics on each of the cells in the collection (Figure 3.5). As the number of topics modelled increases, some cell systems split more readily than others. H1ESC, for instance, has two enriched topics in even the smallest four topic model. For each of the number of topics, we calculate the median fold enrichment (Section 3.2.7) and find that five and six topics produce very similar median fold enrichments (5.81 versus 5.82 respectively). To reduce the complexity of the model, and acknowledging that a model with more parameters will tend to fit data better, we choose the study the case with five topics.

Choice of peak caller influences the inference of topic loadings We additionally investigate the role that peak calling has on the inference of topics.

Peaks were called with Macs2 instead of LanceOTRon, resulting in a larger number of peaks identified in general, with 4657 being uniquely identified by Macs2 (Figure 3.6A). We repeated the above analysis, including the hyperparameter search, topic inference, and calculation of median fold enrichment was conducted using MACS2 instead of LanceOTron, as detailed in Section 3.2.3. Several differences are apparent in both the optimized hyperparameters and the inferred topic loadings (Figure 3.6). Slight bias towards lower α values is observed in this dataset, and the

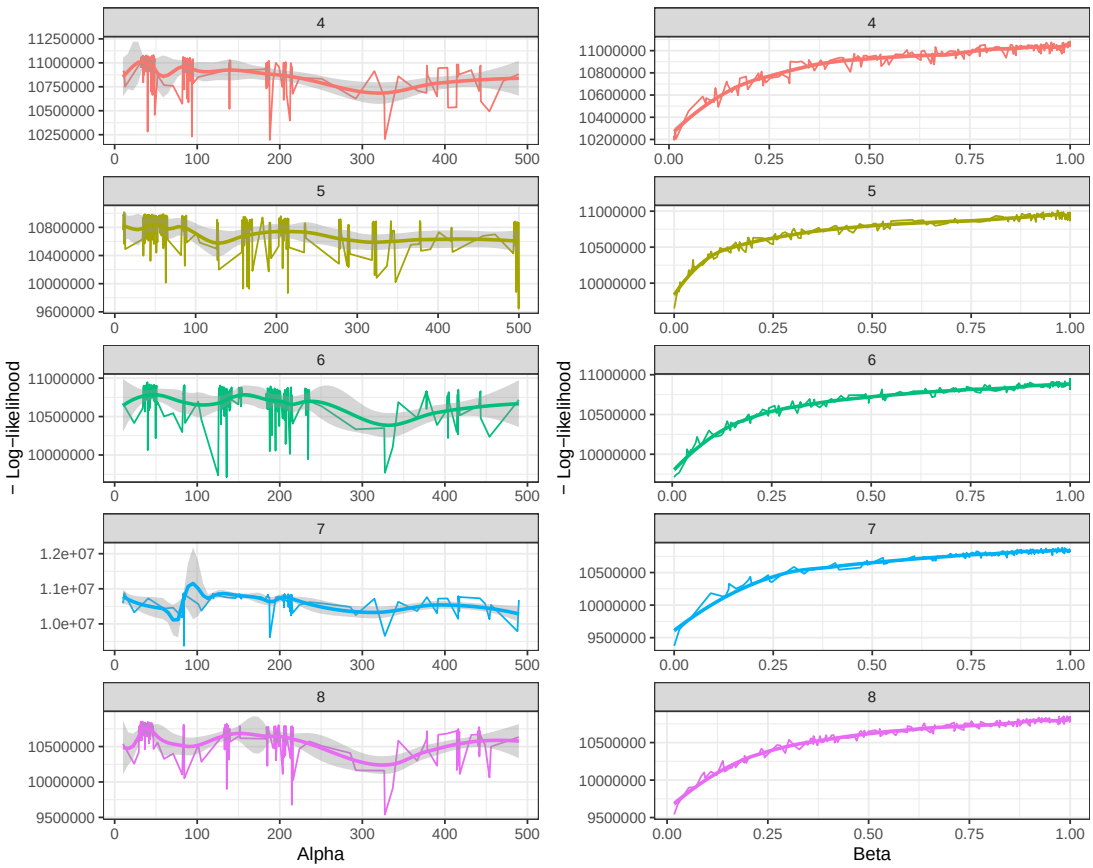


Figure 3.4: Single cell log likelihood for different values of the topic modelling hyperparameters α and β for various numbers of topics being modelled. Alpha and beta parameters are given as symmetric hyperparameters of the Dirichlet distribution, where alpha is normalized according to the number of topics

same trend towards preferring higher values of β is replicated here Figure 3.6B, C. However, when looking at the inferred topics, we find lower median fold enrichment in general (Figure 3.6D, E). From this we conclude that LanceOTron has identified a more specific set of accessible regions for the purposes of this study, and use it in all analyses going forward.

LDA produces both a distribution of topic loadings on cells as well as topic loadings on regions. In order to identify regions putatively identified with the cell types through the topic loadings, a decision must be made about how to discretise the continuous loadings to a set of enriched regions. The cisTopic package accomplishes this by fitting a gamma distribution to the topic loadings, and identifying regions within the top user-defined percentile of this distribution. By default, this is set

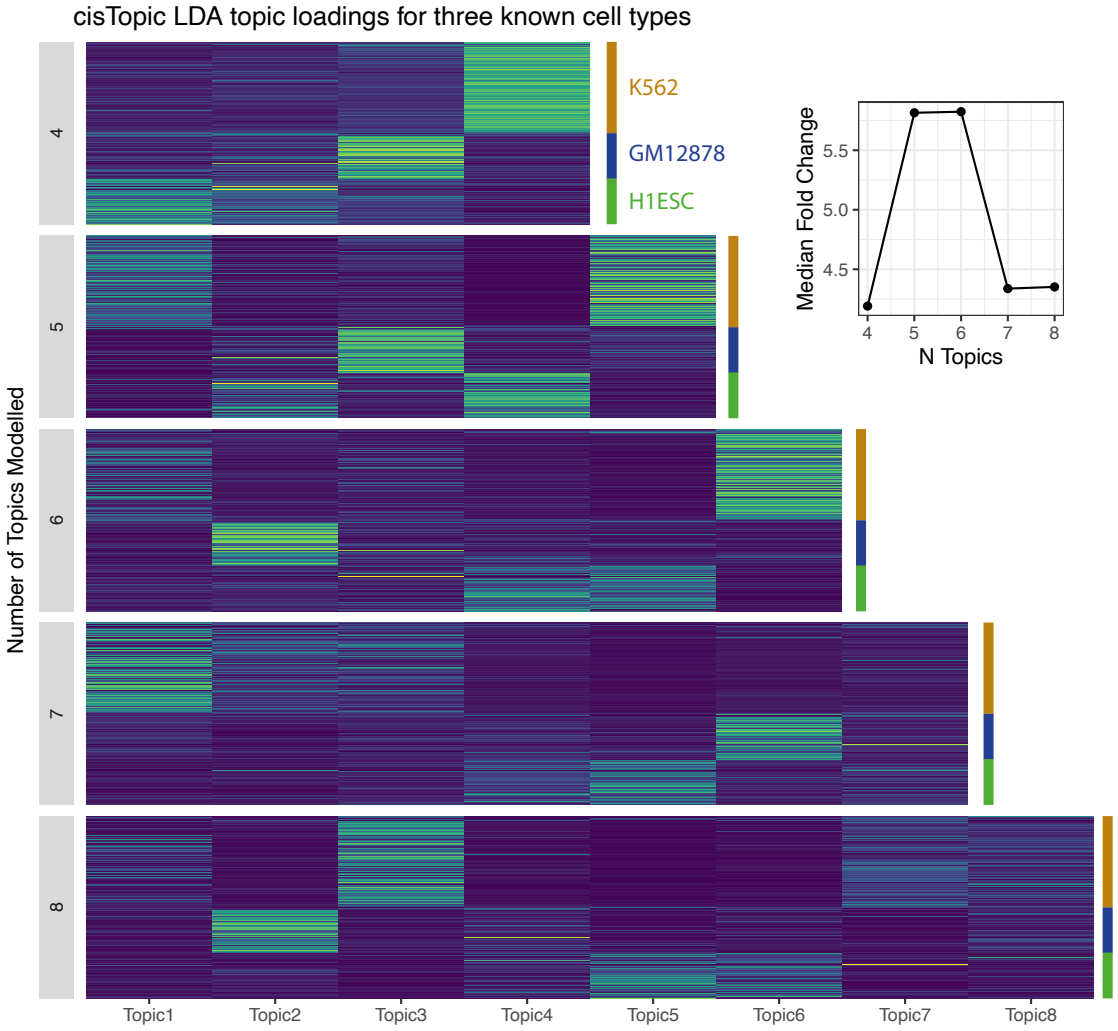


Figure 3.5: Topic loadings for 4, 5, 6, 7, and 8 topic instances of the LDA inference procedure using optimal hyperparamters as decided in Figure 3.4. Inset shows average fold change amongst enriched topics for each of the number of topics modelled. Enriched topics are identified by a one-tailed student's t test for difference between a known cell cluster and the remainder of cells.

to the top 5% of the distribution. Here we investigate how to set this parameter value and the influence it has on the resulting post-hoc analyses.

Enrichment of differentially accessible regions within identified topics To assess whether the identified topic loadings represent similar trends to the baseline differential accessibility analysis, we fit a gamma distribution to the shape of each topic loading and take the top 2.5 percentile of the distribution as representative so-called "keywords" or key regions. We use bedtools to overlap these key regions

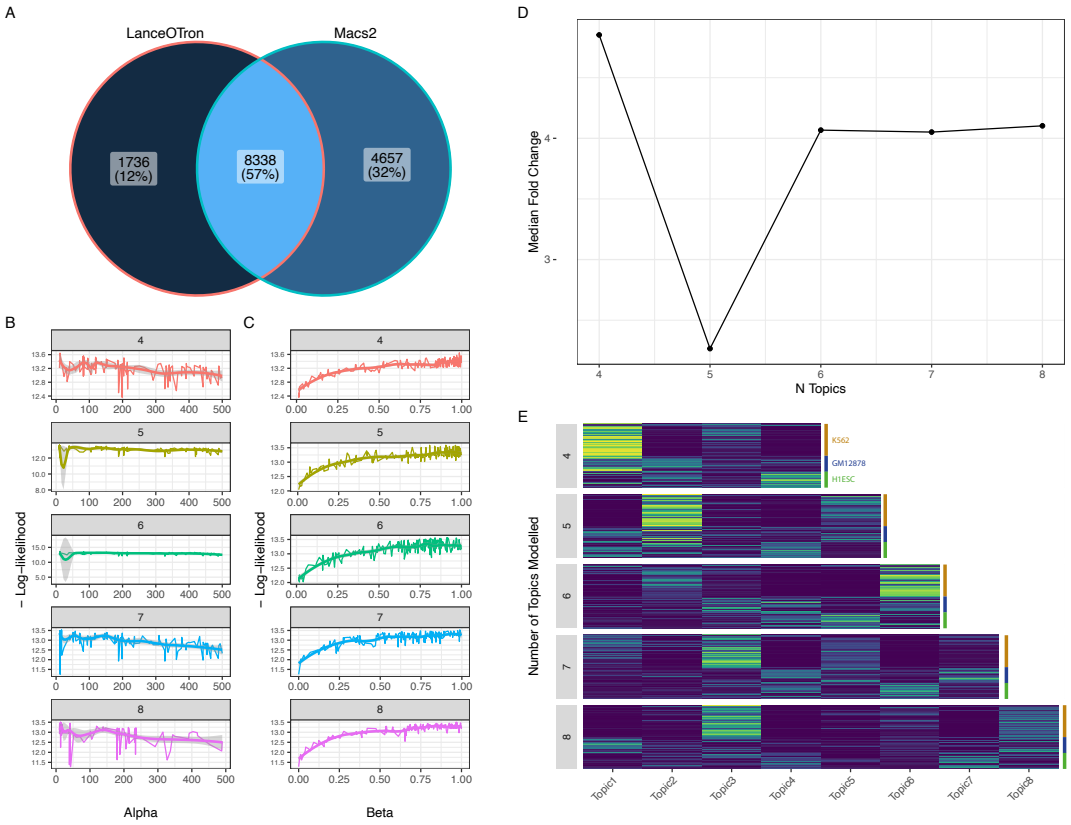


Figure 3.6: Analysis of single cell dataset with peaks called by MACS2. A. The number of peak calls made by both callers, with their overlapping portion displayed in the center of the Venn diagram. B. Likelihood surface for the optimization of the α hyperparameter. C. Likelihood surface for the optimization of the β hyperparameter. D. Median fold enrichment for topics inferred under optimal hyperparameters. E. Inferred topic loadings on single cells using optimal hyperparameters.

with the 100 differentially accessible regions from the `edgeR` analysis and find that 61 of the 100 regions are shared with the top 415 selected key regions (Table 3.1). Given that there are 10074 regions in total, this represents a significant ($P < 0.00001$, $\chi^2 = 827.33$) enrichment of differentially expressed regions in the selected regions from the LDA procedure.

3.3.1.3 Extending LDA to Pseudo-bulked Single Cells

We investigate whether the approach implemented in cisTopic is appropriate for use in bulk ATAC-seq experiments. This change represents a deviation in the intended system for the approach. Practically, the difference between the single cell case and the bulk one is the difference between a wide, sparse count matrix and a dense one

Topic	Total Selected Regions	Intersecting Regions
1	104	15
2	3	0
3	194	29
4	3	1
5	111	19

Table 3.1: Five topic single cell inferred topic loadings and the proportion of their selected regions which overlap 100 established differentially accessible regions. Overall, 61 of the 100 regions were found in the top 415 regions.

with fewer observations. An attractive property of using Gibb's sampling to infer the posterior cell-topic distribution is its ability to efficiently deal with sparsity in the region-topic distribution. However, it is not clear to what degree this sparsity is a requirement for the procedure to obtain biologically relevant topics, each representing a proxy of co-accessible regulatory elements. In this section, I take the well-characterized single cell dataset from the previous section and study the analogous bulk sequencing case by artificially creating bulk samples from the individual known cell types. We denote these new bulk cells as pseudobulked samples.

We combine all reads from each known cell type into pseudo-bulked alignment files using pySam. Peak calling is performed separately on each of the pseudo-bulked samples using LanceOTron, as described in the single cell case. We additionally explore the role of thresholding on LanceOTRon peak calls by creating a subset of calls with at least 0.5 peak score. This is a recommendation made by the web interface to the peak caller. Before thresholding, 517620, 395953, and 455886 peaks were identified in the H1ESC, GM12878, and K562 cells respectively. After merging, 1063153 regions were used for the complete analysis. After thresholding on a peak score of 0.5, 7524, 4945, and 6722 peaks were found in the GM12878, H1ESC and K562 cells respectively. After merging, there were 16210 regions in total used for the thresholded case.

We take two approaches to constructing the count matrix. The first is the already described cisTopic method, where peak regions are merged and annotated by contributing cell types, leading to a binary sparse matrix where the entries i, j represent overlapped peak j in cell type i. We denote this case as "one-hot encoded"

Method	Number of topics	Best Alpha	Best Beta
One-hot encoding	4	123.68822	0.0382351
One-hot encoding	5	466.58752	0.0665157
One-hot encoding	6	450.96735	0.0756824
One-hot encoding	7	337.58572	0.0769945
One-hot encoding	8	100.35205	0.0872716
RPKM Normalization	4	40.38505	0.2350854
RPKM Normalization	5	285.51986	0.3011584
RPKM Normalization	6	285.70995	0.4163319
RPKM Normalization	7	366.02309	0.3825605
RPKM Normalization	8	101.26750	0.3455916

Table 3.2: Optimal LDA hyperparameters for pseudo-bulked scATAC-seq parameterized by *a priori* defined topic numbers and two different read quantification methods.

as the representation is analogous to one-hot encoding factor levels within a design matrix. This approach is justified with the single cell case as read depth is much lower per cell, and a quantification of relative read support for a particular peak region is highly variable. In the case of bulk ATAC-seq however, a difference in read support for different regions can imply varying degrees of accessibility, a key feature of closely related cell types or differentiation processes. To reflect this in the model, we propose an extension of the cisTopic method which we call bulk LDA (BLDA). In this extension, we normalize read counts according to the floor of their reads per kilobase pair and million reads (RPKM) normalized value, giving an integer value of effective number of times the "word", or region, is represented in the particular dataset. This approach has the advantage of incorporating quantitative information about differential peak strengths across experiments, and more closely mimics typical applications of LDA within natural language processing. Here we investigate whether the quantitative extension of cisTopic for bulk ATAC-seq is able to better recapitulate the single cell case.

We decided hyperparameters through Bayesian Optimization with four through eight topics, the same as the single cell case for consistency for both approaches separately. We have omitted the likelihood surfaces and provided the optimal parameters after 500 iterations in Table 3.2. We run classic LDA for the One-hot

Encoding (OHE) and our extended BLDA cases using the selected hyperparameters and compare the inferred topic loadings between the approaches (Figure 3.7). Given the small number of dense cells which share a high proportion of peak regions, in both the entire peak dataset and the thresholded one the BLDA method produces sharper and more defined topic loadings onto the individual pseudo-bulk cells. The cell-topic distribution is not noticeably different between the thresholded and non-thresholded peak calls (A versus B, Figure 3.7). The OHE case for this data was unable to produce topics which differentiated between the difference cell types, while the RPKM normalization may have allowed the inference procedure to select regions which varied in their accessibility (left versus right, Figure 3.7).

To investigate to what degree the BLDA procedure found regions which were differentially accessible across pseudo cell-types, we compare the "key word" regions to the regions selected by edgeR in Section 3.3.1.1. We use cisTopic's built in procedure for selecting important contributory regions to topic definitions, fitting a gamma distribution to the shape of the region-topic distribution for each topic and selecting regions which lie in the top 1% of that distribution. Doing this, we find a difference in the number of regions identified between the thresholded and non-thresholded LanceOTron peak calls, which we expect given the selection procedure (Figure 3.8). More peaks are also uniformly selected for the BLDA method, though we expect that this is due to the lack convergence in the classic LDA set up, rather than a systematic difference based on the method.

However, we find that the key-word regions are not equally specific with regards to the "ground-truth" differentially accessible regions from edgeR. Taking the top 100 regions from the edgeR analysis, we overlap them with all key-word regions in a given analysis (Figure 3.9).

Firstly, we find a clear difference between OHE and BLDA in their ability to identify truly differentially accessible regions (First versus second panel of Figure 3.9). Considering the case of thresholded peak calls, OHE identifies almost none of the known differentially accessible peaks, while the BLDA method identifies slightly fewer than the original single cell experiment. This supports the assertion that

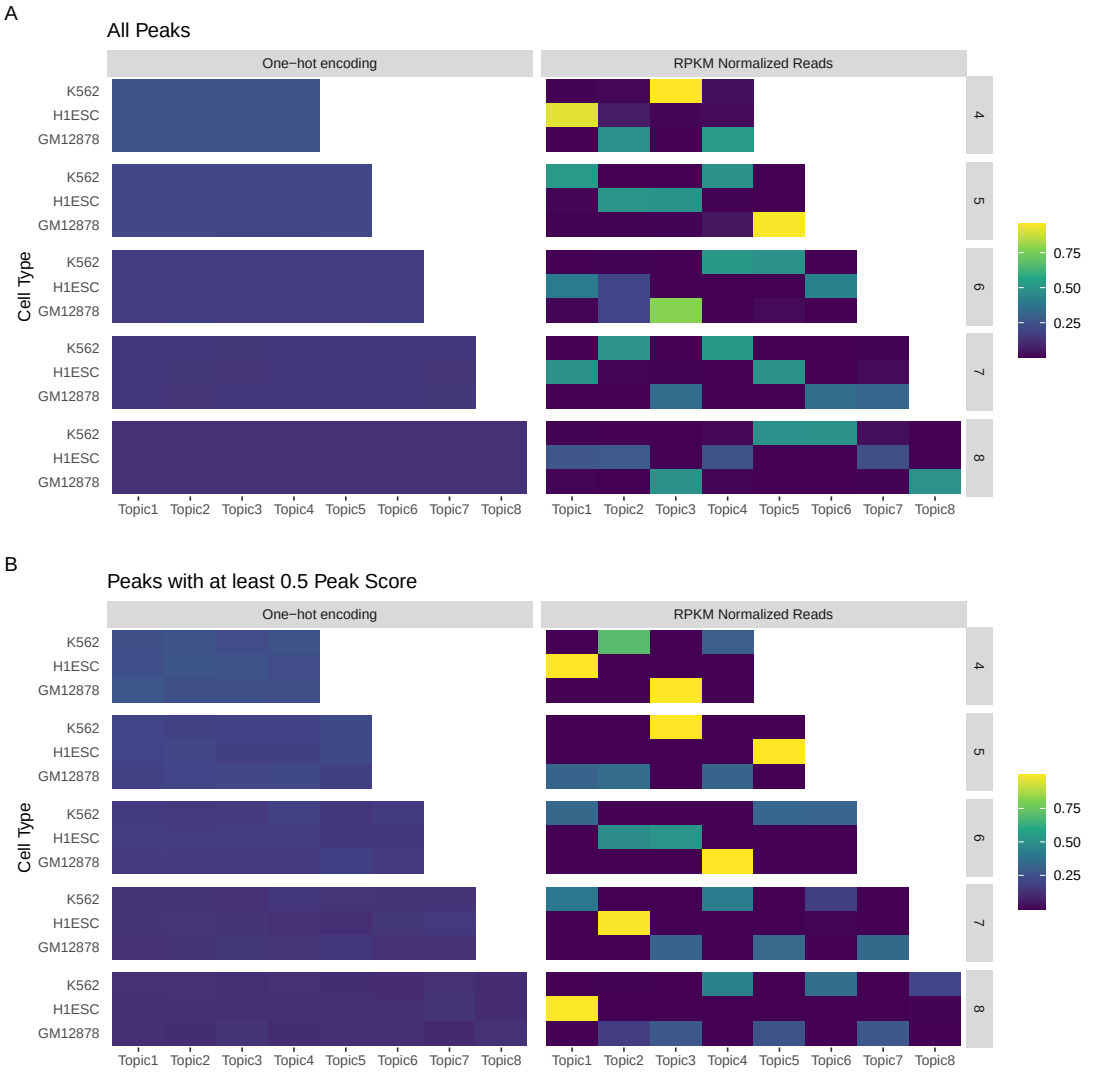


Figure 3.7: Inferred topic loadings using optimized hyperparameters and *a priori* defined for both the OHE and BLDA methods using pseudo-bulked scATAC-seq data.

bulk LDA is a comparable approach to single cell LDA for datasets consisting of a small number of dense cells.

3.3.2 Bulk LDA describes Erythropoiesis

Having established that BLDA is able to identify realistic patterns in bulk data, we investigate a well-characterized biological system. Erythropoiesis, the process by which red blood cells are produced, involves known differentiation stages with defined marker genes. This allows us to compare the topics inferred from BLDA

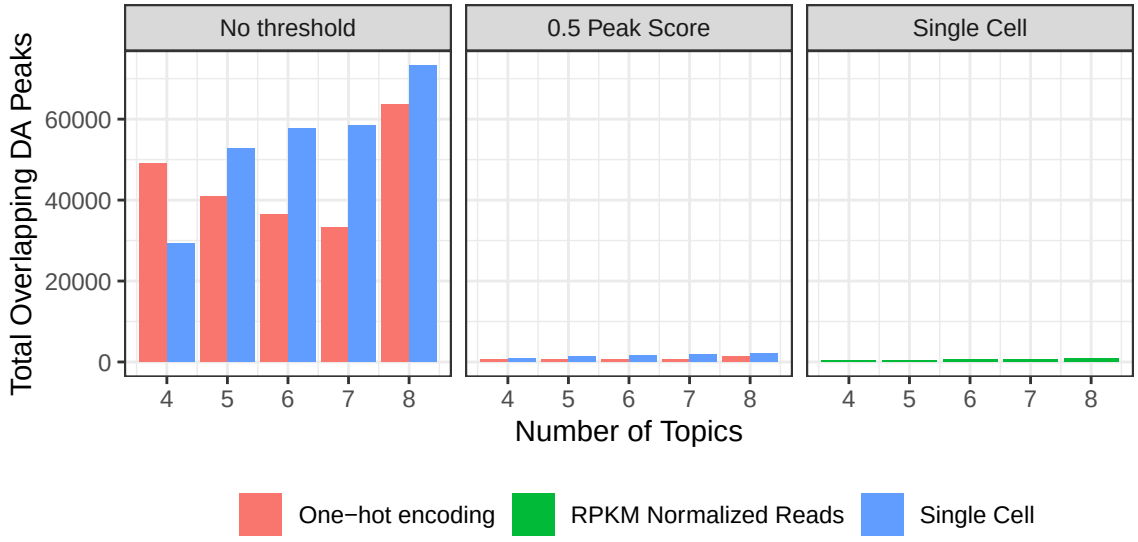


Figure 3.8: Total number of selected "key-word" regions for a given topic number between the two pseudo-bulk approaches, the two peak calling methods (thresholded and non-thresholded LanceOTron) and single cell analyses.

Figure 3.9: Number of overlapping regions between the top 100 differentially accessible regions determined by EdgeR and key-word regions selected by taking the top 1% of a fitted gamma distribution for several LDA analyses.

with expectations based on the biology of the system.

To create a dataset of bulk ATAC-seq data to study erythropoiesis, we download and process raw sequencing data from Ludwig et al. [200] and Corces et al. [202] using gene expression omnibus datasets GSE115684 and GSE74912 respectively. Raw reads were assessed for quality with FastQC and aligned using. Coverage tracks were created using DeepTools and peak calling was performed using LanceOTron using the same score cutoff as previously described. The number of peaks identified this way varied considerably across cell types (Figure 3.10A). Accessible regions of the genome generally increase from hematopoietic stem cells (HSCs) to common myeloid progenitors before decreasing in Megakaryocyte–erythroid progenitor cell. Erythrocyte colony forming units (CFUE) and orthochromatic erythroblasts (OrthoE cells) have especially accessible chromatin, with nearly 50,000 accessible regions. As it is known that nuclear chromatin condenses in preparation for enucleation in terminally differentiated immature erythroblasts, accesibility and the number of identified peak regions are greatly decreased in erythroblasts here as well [186]. There is relatively low sharing of accessible regions outside of a central differentiation "block" between ProE and BasoE cells (Figure 3.10B). This too is consistent with our expectation, as terminal differentiation is known to significantly alter the expression of many hundreds of genes [186]. It is now well understood to what degree the differential usage versus differential accessibility of key regulatory elements shapes normal differentiation of human cells [227, 228]. Here we use the previously characterized bulk LDA to identify groupings of accessible regions which are associated with specific stages in erythropoiesis.

Peak calls were merged to create two count matrices. The first represents the typical cisTopic analysis method, one-hot encoding each peak region from its derivative cell type. The second is our BLDA method, using an integer value of RPKM normalised read count for each identified accessible region. We optimise the hyperparameters as before, and run inference for between 8 and 20 topics. The number of topics was chosen to demonstrate, on the lower end, how LDA will group similar cells when it is forced to, and on the upper end if there is any

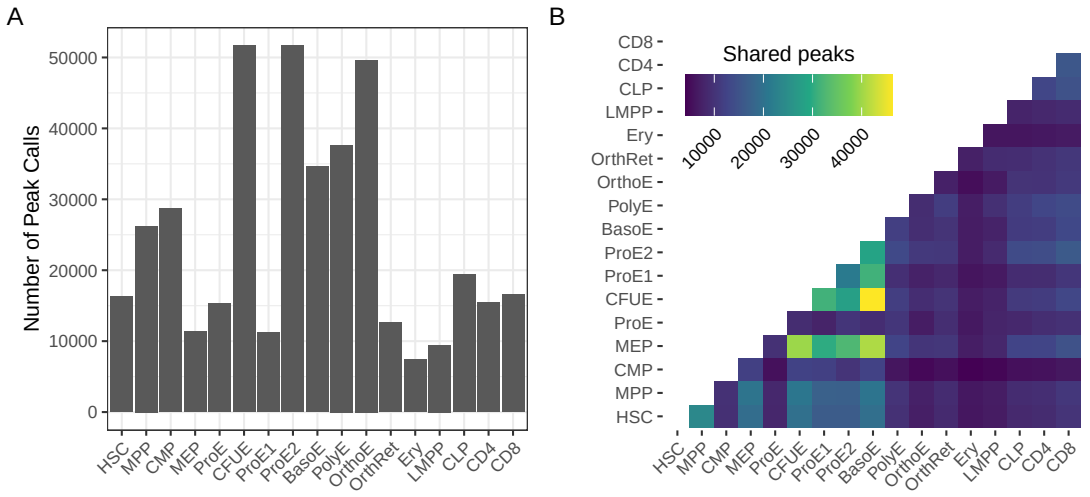


Figure 3.10: Peak calling in erythropoiesis dataset. A. Number of identified peak regions by cell type using LanceOTron and a score threshold of 0.5. B. The number of peaks shared between cell types.

remaining structure after each cell is allowed its own topic. We find that, similar to pseudo-bulked scATAC-seq data, the BLDA method consistently produces cleaner and better defined topic loadings for individual cell types (Figure 3.11). While this larger and more realistic dataset allows the OHE strategy to pick out some topics which differentiate between cell types, the patterns tend to be strongly constrained to certain grouping such as the erythroid precursors or lineage committed cells. The BLDA method on the other hand distinguishes highly cell-specific topics. The number of topics which are shared across multiple cell types varies considerably as the model is given more freedom with increasing number of topics. Even at low numbers, highly differentiated cell types like CD4 and CD8 T cells show distinct, highly enriched topics. Conversely, topics are almost always identified which are enriched in closely related intermediary cell types. We quantify the number of times two cell types share an enriched topic, taking a cell-normalised topic loading threshold of 0.25 to indicate enrichment (Figure 3.12). Two blocks of co-enriched cell types are obvious, one beginning from erythroid progenitors (ProE) and ending with basophilic erythroblasts (BasoE), the second beginning with BasoE and ending with Erythroblasts (Ery). It is rare for a cell type in either

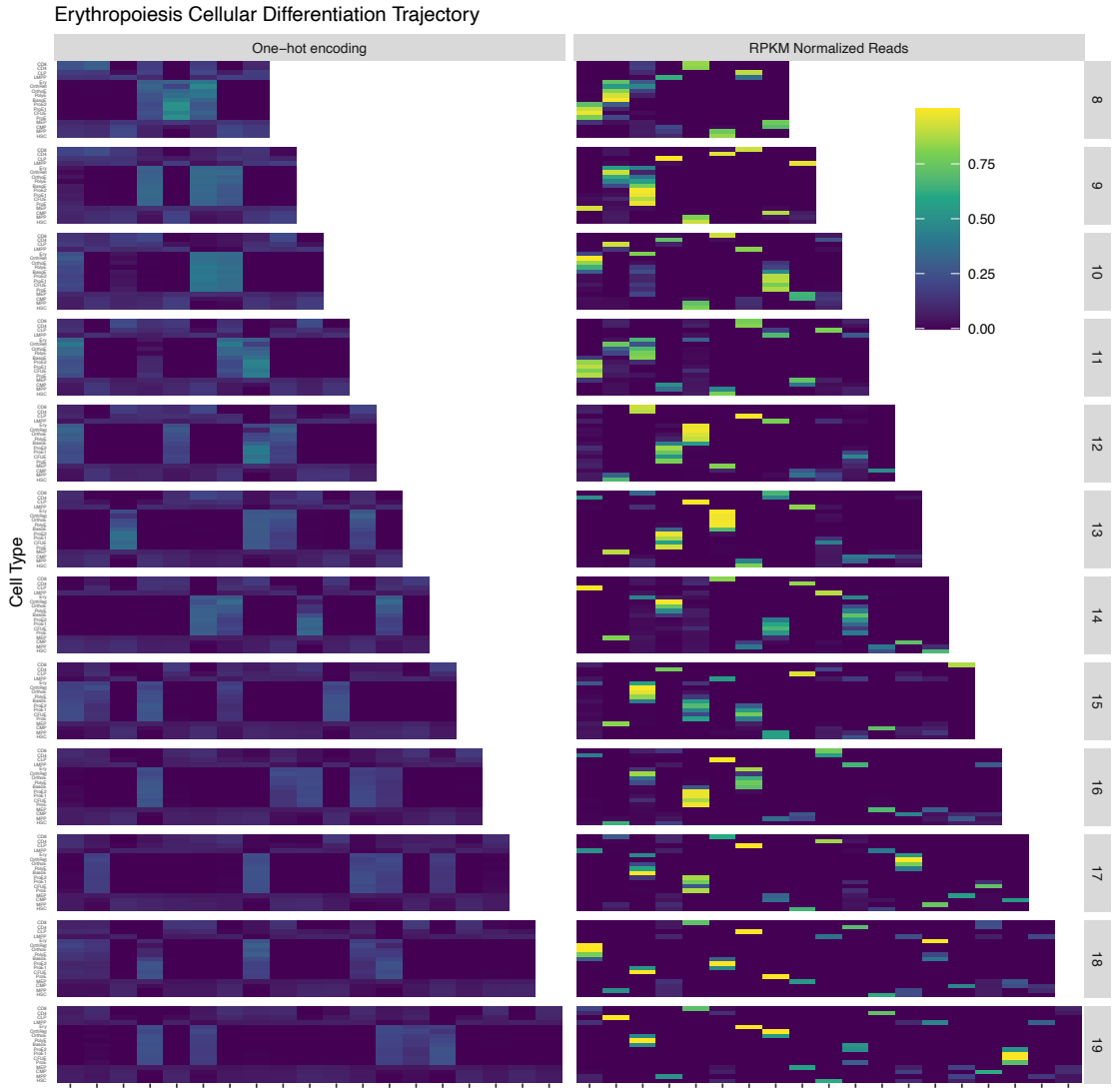


Figure 3.11: Inferred topic loadings for erythropoiesis dataset. On the left hand side, One-hot encoded LDA, on the right bulk LDA with RPKM normalization. Facets indicate the number of modelled topics. Cells are indicated in reverse differentiation pseudotime, with each facet ordered on the Y axis as CD8, CD4, CLP, LMPP, Erythroblast, Ortho/Ret, OrthoE, PolyE, BasoE, ProE2, ProE1, CFUE, ProE, MEP, CMP, MPP, and HSC. See Figure 3.15 for a detailed exploration for $k = 8$.

one of these clusters to share topic loadings with the other. Additionally, topics are occasionally shared between sequential stages of early differentiation, i.e. between hematopoietic stem cells and multi-potent progenitors (MPPs), but these topics never overlap with lineage committed cell stages.

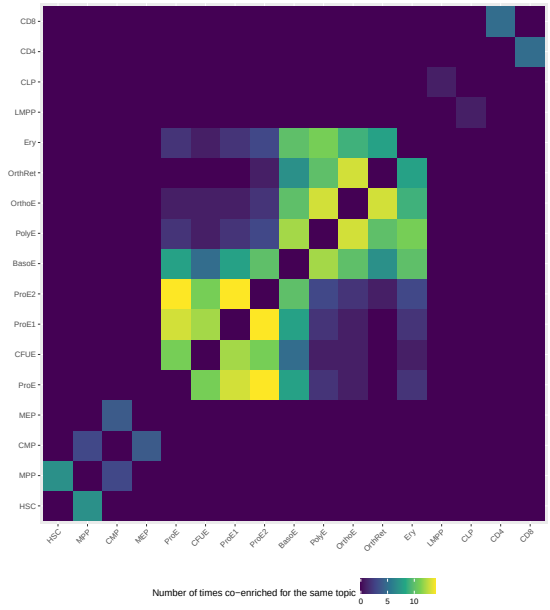


Figure 3.12: Similarity of cell types based on the number of times they are co-enriched for a topic, summarized over all numbers of inferred topics. Topics inferred using BLDA for the erythropoiesis dataset as described in the Methods section. For a description of the cell types, see Section 3.1.2

3.3.2.1 Isolating key-word regions from region-topic loadings

Certain regions are inferred to be uniquely important for a topic. We follow the convention within the LDA literature in calling these samples key-words, as the observations within a document are typically denoted as words. There is no theoretical definition of a key-word region, therefore a threshold on the region-topic distribution must be empirically chosen to represent the most important regions. The cisTopic method normalizes the distribution of loadings within a topic such that it falls in the range 0-1 and theoretically follows a Gamma distribution, though the parameters for this distribution must again be inferred empirically. In this section, we attempt to identify sensible thresholds for key-word regions based on topic loadings.

To decrease the number of variables that need to be studied, for this section we will focus on the BLDA inference. Based on the qualitative results from the previous section, quantitative input to the LDA inference algorithm produces more specific topics which are shared amongst related cell types in realistic ways. This specificity is important for the identifications of key-word regions.

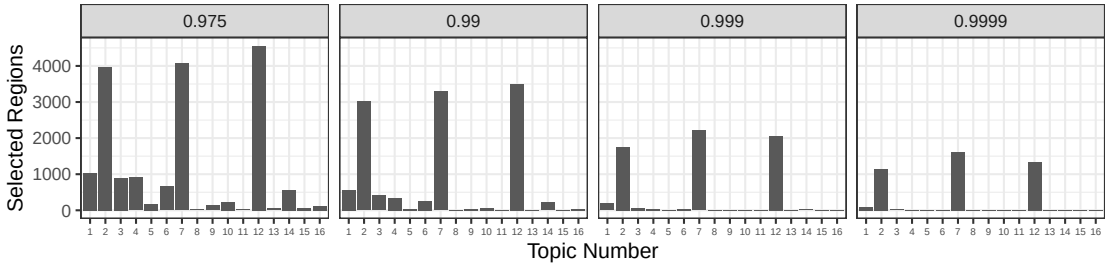


Figure 3.13: Number of identified regions above the facetted percent point function of a Gamma distribution with inferred parameters using SciPy.

Additionally, we begin our analysis by focusing on only one value for the number of topics. For the analysis with $k = 16$ topics, SciPy was used to estimate the parameters of a Gamma distribution for the region-topic distribution for each of the 16 topics. Four thresholds were used to select the top 2.5, 1, 0.1, and 0.01% of the inferred Gamma distribution. The number of selected regions ranged considerably across the topics (Figure 3.13). From this we conclude that a gamma distribution may not fit each of the topics equally well, as the proportion of peaks that were selected based on the threshold does not reflect the theoretical expectation.

To investigate further, descriptive statistics are used to identify candidate distributions. We use a Cullen a Frey graph to examine one of the poorly fit topics from Figure 3.13, topic 8 (Figure 3.14). 1000 bootstraps of the data indicate two potential distributions, gamma and log-normal (top left of Figure 3.13). Maximum likelihood is used to estimate parameters for these distributions, and the empirical versus theoretical density, percentiles, and cumulative distribution functions are plotted (top right, bottom left, and bottom right of Figure 3.13 respectively). Though it appears as though the density at least visually matches the gamma distribution, the larger empirical percentiles of the data do not match either a gamma or log-normal distribution (bottom left of Figure 3.14). It seems likely that the skew of the data is causing the under-representation of selected key-word regions.

This lack of fit to a theoretical gamma distribution presents several avenues for exploration. One option would be to fit a mixture distribution, explaining different portions of the data with different parameters for the distributions. However, this would make it difficult to estimate a certain proportion of the overall distribution,

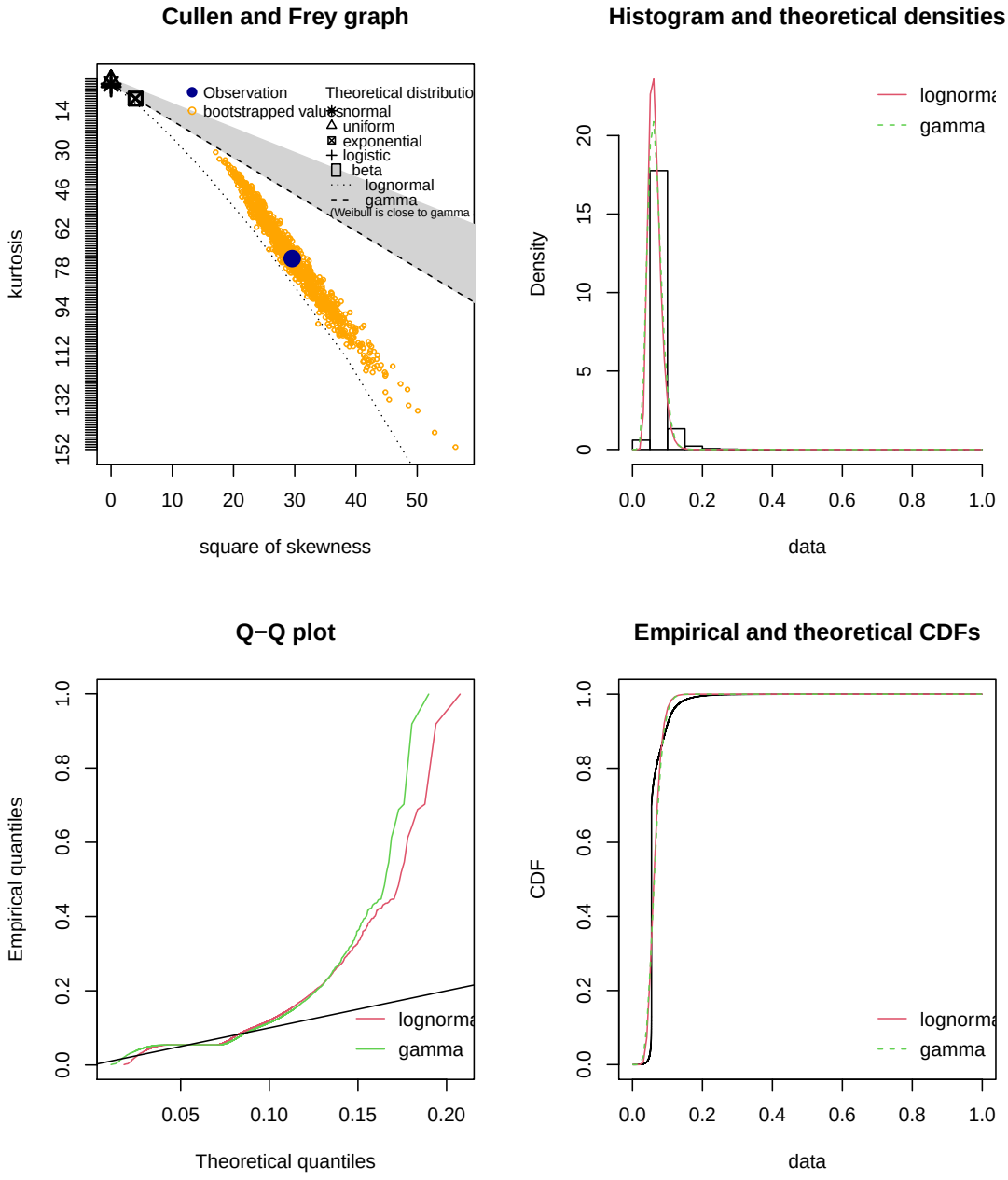


Figure 3.14: Distribution of the region-topic distribution from a candidate topic from Figure 3.13, topic 8. Top left shows a comparison of 1000 bootstrapped values of the data against descriptive statistics for several common distributions. Top right shows the empirical histogram of the data compared against the two distributions nominated from the Cullen and Frey graph fitted via maximum likelihood estimation. Top left shows empirical versus theoretical percentiles from the two fitted distributions, while bottom right shows the empirical cumulative distribution function.

which is the goal of the exercise. Instead, we opt to take the simpler route and simply take a fixed number of the highest region-normalised topic loadings. This guarantees a sufficient sample size for further analyses like motif identification and pathway enrichment, while not making the analysis overly complex for replication across the different k values of topics.

3.3.2.2 Motif enrichment within key-word regions

The top 100, 250, and 500 regions from each region-topic distribution are selected and MotifScan (`https://motifscan.readthedocs.io/en/latest/`) is used along with the JASPAR database of TFBS motifs to identify regions matching known Position weight matrix (PWM) and find over-representation in the set of regions. A control set of regions is constructed by taking the negative intersection between the original merged dataset of regions and the selected regions, in order to better represent both accessible DNA and relevant common motifs for the system. A second control set is constructed by taking the union of all regions selected as key-word regions. This second set investigates motifs which are enriched relative to other selected regions.

The smallest k topics forces sharing between similar cell types (Figure 3.15). For each topic, the enrichment of all PWM in the JASPAR database is calculated for each count of top regions, and each control strategy (Figure 3.16). Some topics in this particular instance are of interest. Topic 6 spans the early developmental stages between hematopoietic stem cells and multi-potent progenitors, and shows enrichment for known factors like MEIS1, whose expression is required for erythropoiesis in HSCs [229–231].

3.3.2.3 BLDA identifies relevant pathways active in Erythropoiesis

A combination of the motifs identified as well as the groupings of regions allows for a deeper interrogation of the uncovered biology.

Ludwig et al. [200] identified several regions of the genome with substantial impacts on the process of Erythropoiesis. Of these, there are no overlaps between UROS, CCND3, VEGFA, and TMCC2 in the set of selected keyword regions. However, the promoter region for the Rh associated glycoprotein (RhAG), which

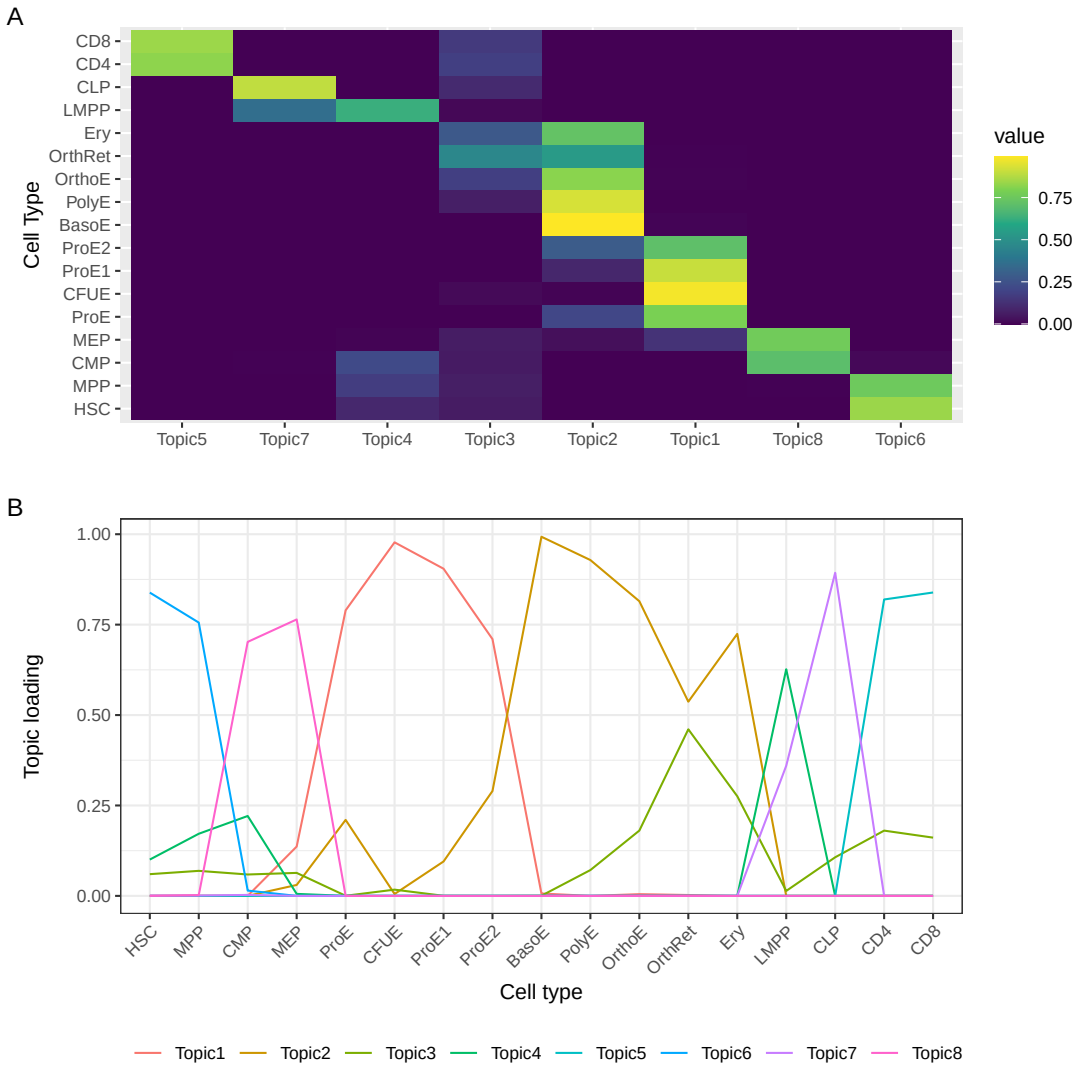


Figure 3.15: Cell-topic distribution for $k = 8$ topic. A. Loadings across cell types ordered by differentiation trajectory. B. Loadings by topic across differentiation pseudo-time.

is not accessible in HSC, MPP, or CMP cells, shows a progressive increase in accessibility throughout erythropoiesis, as shown in Figure 3.17a. A keyword region for topic 2 is located in its promoter. Topic 2 mirrors the accessibility patterns of RhAG, showing which also is shown to be inactive in early hematopoiesis and increasing in accessible after lineage commitment, peaking in BasoE cells. Other examples of individual regions which mirror topic loading patterns are readily apparent, such as the R3HDM4 locus, which is also a key region for topic 2 (Figure 3.17b).

	All Reference Regions			All Selected Regions		
Topic Numbers	100 Regions	250 Regions	500 Regions	100 Regions	250 Regions	500 Regions
1	NA	GATA3, ZNF354C	ZNF354C, GATA1::TAL1, GATA3, Klf1, ZNF460, GATA2, Klf12	NA	Arid3a	GATA1::TAL1, GATA3, Klf1, FOXI1, FOXL1, FOXO4, FOXO6, FOXP3, NFATC2, Arid3a
2	Bach1::Mafk, MEIS1, GATA1::TAL1, ZNF354C, MEIS3	GATA1::TAL1, Bach1::Mafk, ZNF354C, MGA, TBX1, TBX15, TBX4, TBX5, MEIS1, Nfe2l2, VDR, Arnt, GATA3, Zfx, NFE2L1, RHOXF1, HLTF, HIC2	GATA1::TAL1, ZNF354C, GATA3, Bach1::Mafk, MEIS1, VDR, Zfx, GATA5, MGA, TBX1, TBX15, TBX4, TBX5, Nfe2l2, Arnt, SP2, Nr2f6(var.2), Rarg, Klf1, SP4, ZNF460, NFE2L1, NR4A2, MYC, HLTF, KLF10, MAX, HIC2	NA	GATA1::TAL1, Bach1::Mafk, Zfx, Nfe2l2, MGA, TBX1, TBX15, TBX4, TBX5	GATA1::TAL1, Bach1::Mafk, Zfx, SP4, SP2, Nfe2l2, GATA5, KLF10
3	Ahr::Arnt, THAP11, SP4, SP2, SP1, ZBTB14, KLF16, TFDP1, KLF5, NRF1, KLF15, SP8, SP3, NFYC, NFYA, KLF11, HINFP, TFAP2E, Gmeb1, KLF14, FOSL1::JUND(var.2), KLF3, EGR3, ZNF354C, ZBTB7A, Arnt, E2F6, EGR1, ZNF263, FOSL2::JUND(var.2), CREM, ZBTB33, NFYB, FOXI1, FOXL1, FOXO4, FOXO6, FOXP3, JUNB(var.2), NR2C1, KLF13, ELK4, Pax2, Klf12, Zfx, YY2, TCFL5, EGR4, JDP2(var.2), FOS::JUN(var.2), CREB3L4(var.2), BARHL1, ZBTB7B, ZBTB7C, KLF6, INSM1, ATF3	Ahr::Arnt, THAP11, KLF5, SP4, NRF1, SP2, TFDP1, SP1, ZNF354C, ZBTB14, KLF16, TFAP2E, ELK4, Gmeb1, SP3, Arnt, KLF15, EGR1, Pax2, NFYA, ZBTB7A, NFYC, HINFP, ETV3, SP8, NFYB, Zfx, EGR2, YY2, EGR3, CREM, KLF11, YY1, ETV5, JUNB(var.2), ZBTB33, E2F6, KLF3, KLF14, Klf12, ZNF263, SP9, TCF12(var.2), TCF3, ZEB1, FOSL1::JUND(var.2), ELK1, FOXI1, FOXL1, FOXO4, FOXO6, FOXP3, FOSB::JUNB(var.2), BHLHE22(var.2), FOSL1::JUN(var.2), ZBED1, CREB3L4(var.2), GLIS2, ERF, FEV, NR2C1, EGR4, JDP2(var.2), ETS1, ELK3, ELF2, SPDEF, ETV6, CREB3, FOSL2::JUND(var.2), Wt1, E2F4, JUN::JUNB(var.2), KLF10, ELF4, KLF13, ZBTB7B, RFX7, EBF1, TFE3, BARHL1, Atf1, MEIS1, ETS2, MAZ, ARNT::HIF1A, ELF5, NFATC2, ZIC3, KLF6, ATF3, ATF7, FOSL2::JUN(var.2), TFEB, HOXA9, ATF6, TCFL5, Mlxip, NHLH1, INSM1, BHLHE41, ZBTB7C, FOS::JUN(var.2), BSX, EN1, ISX, MSX1, MSX2, Msx3, RAX2, SHOX, Shox2, TLX2, UNCX	Ahr::Arnt, THAP11, SP4, SP1, KLF5, SP2, NRF1, KLF16, ZBTB14, TFDP1, KLF15, SP8, SP3, HINFP, KLF3, FOSL1::JUND(var.2), KLF11, NFYA, NFYC, KLF14, ELK4, TFAP2E, KLF13, ZBTB7A, Gmeb1, CTCFL, EGR3, E2F6, NR2C1, NFYB, ZBTB33, ZNF263, FOSL2::JUND(var.2), CREM	Ahr::Arnt, THAP11, KLF5, NRF1, SP4, SP2, SP1, TFDP1, KLF16, ELK4, ZBTB14, SP3, TFAP2E, KLF15, HINFP, ZBTB7A, Gmeb1, ETV3, EGR1, ETV5, SP8, Pax2, KLF3, EGR3, YY1, NFYA, NFYB, NFYC, KLF14, Zfx, EGR2, Arnt, KLF11, CREM, JUNB(var.2), ELK1, YY2, ZNF354C, SP9, ZBTB33, ZEB1, E2F6, ELF2, Klf12, CREB3L4(var.2), ERF, FEV, TCF12(var.2), FOSL1::JUND(var.2), TCF3, ETS1, ZNF263, ELK3, FOSL2::JUND(var.2), KLF13, ELF4, ETV6, FOSB::JUNB(var.2), FOSL1::JUN(var.2), ATF3, SPDEF, RFX7, NR2C1, GLIS2, E2F4, ZBTB7B, EGR4, JDP2(var.2), ETS2, BHLHE41, Wt1, ATF7, TFEB, KLF10, MAZ, FOSL2::JUN(var.2), ZBED1, CTCFL, BHLHE22(var.2)	Ahr::Arnt, NRF1, THAP11, SP4, KLF5, TFDP1, HINFP, ELK4, SP1, ZBTB14, SP2, SP3, KLF16, ETV5, ZBTB7A, ETV3, Gmeb1, Arnt, EGR2, KLF3, TFAP2E, ZNF354C, KLF15, Pax2, Zfx, NFYC, YY1, NFYB, EGR1, ELK1, NFYA, SP8, Klf12, KLF14, CREM, E2F6, ERF, FEV, ZBTB7B, ATF3, KLF11, ELK3, ETS1, CREB3L4(var.2), EGR3, JUNB(var.2), FOSL2::JUND(var.2), ZBTB33, SP9, ATF7, ZNF263, NR2C1, ELF2, YY2, ELF4, SPDEF, ZIC3, FOSB::JUNB(var.2), E2F4, FOSL1::JUN(var.2), ETV6, FOSL1::JUND(var.2), BARHL1, TFE3, TFEB, GLIS2, KLF13, FOSL2::JUN(var.2), JDP2(var.2), TCF12(var.2), KLF10, ZBTB7C, ARNT::HIF1A, TFAP2C(var.2), BHLHE22(var.2), CENPB, E2F1, CTCFL, BHLHE41, EGR4, ZEB1, JUN::JUNB(var.2), Wt1, FOS::JUN(var.2), CREB3, MAZ, Arntl, HOXA9, FLI1, MLX, TCF3, KLF6, Creb5, MLXIPL, ZBED1, ETS2, FOSL2::JUNB(var.2), ELF5, FOSB::JUN, ERG, RFX7, NHLH1, Dux, ZBTB6, SREBF2(var.2), KLF2, THAP1, MITF, Atf1, CREB3L4, ZFP42, GLIS1, TFEC, EBF1, ARNT2, XBP1, ZNF143, RELB, MZF1(var.2), Znf281, GATA1::TAL1, GATA3, BATF::JUN, JDP2
4	NA	NFYA, SOX4	Arid3a, NFYA, Ahr::Arnt	NA	Ahr::Arnt, TFAP2E	NFATC2, ZNF354C, EGR2, Arnt, ELK4, SP3, KLF3, THAP11, KLF15, Gmeb1, KLF16, SP1, TFDP1, TFE3, ZBTB7A, SP2, NRF1, GATA1::TAL1, KLF5, SP4, Ahr::Arnt
5	ZNF354C, Arid3a, MEIS1, NFATC2, NFKB2, FOXI1, FOXL1, FOXO4, FOXO6, FOXP3	Arid3a, ZNF354C, FOXI1, FOXL1, FOXO4, FOXO6, FOXP3, NFATC2, GABPA, ELF1, MEIS1, SOX15, Arnt, Sox5, ELK4, NFKB2	ZNF354C, Arid3a, FOXI1, FOXL1, FOXO4, FOXO6, FOXP3, NFATC2, GABPA, MEIS1, ELK4, FOXG1, Foxj2, FOXN3, FOXO3, Ahr::Arnt, ELF1, Sox5, NFKB2, SOX15, Arnt, EHF, MGA, TBX1, TBX15, TBX4, TBX5, EGR1, ELF3, Nr2e3, Foxj3, Nobox, ETV4, ZBTB33	NA	GABPA, ELF1	GABPA, ELF1, FOXG1, Foxj2, FOXN3, FOXO3, ETV4, GATA1::TAL1
6	Arid3a	Arid3a, GATA3, MEIS1	Arid3a, GATA3, NFATC2, Sox5, MEIS1, SP4	Ahr::Arnt	SP1, Ahr::Arnt, SP4, THAP11	Arid3a, SP2, TFDP1, KLF3, ZBTB7A, NFYB, ETV5, Zfx, KLF16, THAP11, KLF10, Klf12, ELK4, SP3, ETV3, KLF5, NRF1, SP1, Ahr::Arnt, SP4
7	MEIS1	ZNF354C, MEIS1, HAND2, Ascl2	MEIS1, Ascl2, ZNF354C, MYOG, GABPA, Arid3a, HAND2, SNAI3, POU6F2, NR4A2	NA	SP4	Ascl2, HINFP, SP3, KLF15, TFDP1, Klf12, SP1, KLF16, SP2, THAP11, NRF1, KLF5, SP4, Ahr::Arnt
8	GATA1::TAL1	Arid3a, GATA3, SOX18, Sox5	Arid3a, GATA3, Sox5, GATA1::TAL1, POU3F4, POU5F1B, NFATC2, SOX18, KLF5, SP4	NA	SP4, Ahr::Arnt, KLF5	GATA3, ZBTB7A, KLF15, ZBTB14, HINFP, SP8, SP2, TFDP1, ELK4, KLF16, SP1, Zfx, NRF1, THAP11, Arnt, ZNF354C, SP3, KLF5, Ahr::Arnt, SP4

Figure 3.16: All identified motifs for $k = 8$ topic analysis. The top 100, 250, and 500 topics were selected and enrichment was calculated either against all peak regions, or against all selected key-word regions.

To comprehensively survey the closest annotated protein coding genes, bedtools was used to find the nearest genic regions to each of the key regions for the different topics. Mello et al. [232] recently surveyed differentially accessible genes throughout different phases of erythoiesis. These genes are compared against the nearest genic regions for the top 500 keyword regions from each of the topics. The results illustrate a close match between topic activation patterns and the underlying biology of blood cell differentiation.

Topic 1, active in early lineage commitment, include critical proteins for erythropoiesis such as STAT5A and STK3. STAT5a is predominantly expressed in early erythropoiesis [233]. The regions are enriched for the binding motifs of GATA2, GATA3, KLF1, and KLF12. GATA3 is typically associated with lymphoid precursors and committed T cells, recent results indicate a more ubiquotous role within hematopoiesis [234]. The topic shows no overlap with genes involved in

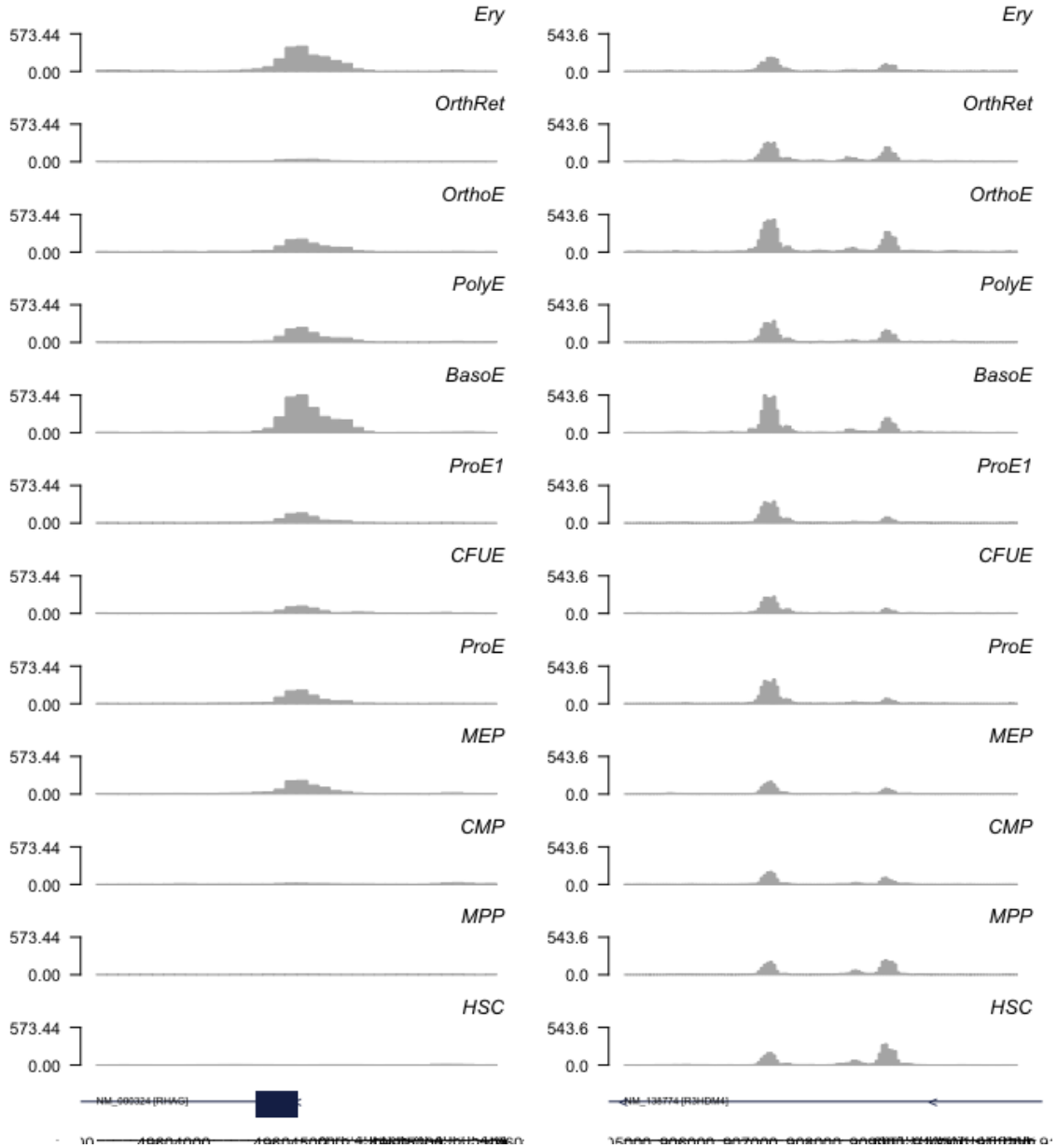
apoptosis, unlike later topics.

Topic 2 represents the main inferred mid-to-late erythropoiesis grouping of regulatory elements. It contains the majority of important proteins involved in iron homeostasis and mitochondrial transportation (SLC25A38), heme production (ALAS2), regulation of cellular differentiation (SP1). In addition, SP1 represents the first protein involved with enucleation. An gene set enrichment analysis showed subenrichment for "Positive regulation of erythrocyte differentiation", indicating that a significant portion of the genes identified relate to erythropoiesis. Additional terms identified include "tetrapyrrole biosynthetic process", "tetrapyrrole metabolic process", and "pigment metabolic process".

Topic 3 peaks in activity amongst orthochromatic erythroblasts, and also has some representation in early hematopoiesis, as well as committed T cells. Topic 3 includes NRF1, a knockout of which leads to embryonic lethality due to impaired fetal liver erythropoiesis [234]. GLRX5, predominantly expressed in late erythropoiesis by poly and orthochromatic erythroblasts (Fig 1c [233]) is also amongst the nearest key genes, along with FOXO3 and PPP2R1A [232]. Topic 3 also shows enrichment for a large number of motifs, indicative that a number of identified peak regions are not directly involved in protein coding but are intergenic enhancer elements (Figure 3.16).

Early hematopoiesis is represented by topics 6 and 8, which are only associated with a handful of genes associated with erythropoiesis (Table 3.3). HIP1 is known to be crucial to successful hematopoiesis and LYN is involved in the successful proliferation of hematopoetic cells [235, 236]. Motifs in these regions include ARID3a, known to be expressed in HSCs, and MEIS1, also involved in early hematopoiesis [229, 230, 237].

These results indicate that BLDA has identified regions which correspond to realistic regions of differential accessibility throughout the process of erythropoiesis. Though the method is not specific, in that many important regions are not represented amongst the results.



(a) The RhAG promoter **(b)** R3HDM4

Figure 3.17: Accessibility at two keyword regions for topic 2 among cells in the erythropoietic differentiation trajectory.

Category	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
Anti-apoptosis	CTSB	MKL1, BCL2L1, CTSB, NR3C1	CITED2, PGAP2, TNFAIP3	PDPK1	HSPA9, NFKB1			
Cellular component involved in apoptosis	ACTN4, KPNB1	CDH1, KPNB1, GSN	PSMB1, RB1CC1	ACTN4	VIM	PSMB1	DBNL	
Erythroblast enucleation		SP1						
Erythrocyte differentiation	STAT5A	ALAS2, SLC25A38, BPGM, ERCC2	CITED2, BPGM		HCLS1, NCKAP1L	LYN	BPGM, DYRK3	
Hematopoiesis & regulation of cell differentiation	GFI1B, ZBTB16, TXNRD2	BCL11A, SP1, TXNRD2	GLRX5, PPP2R1A	TTC7A	IKZF1	ZBTB16, LYN	IKZF1, CDK6, TTC7A	ZBTB16
Heme metabolism	ALAD, BLVRB, TMEM14C	ALAD, ALAS2, BLVRB, FECH, HMBS, SLC25A38, SPTA1, UROD	ALAD	FECH				
Induction of apoptosis	LGALS1, MAPK1, SPN, DAPK2, SAP30BP, VAV1	KAT2B, RBM38, TP53BP2, AKAP13, ERCC2	BCLAF1, PPP2R1A, STK17B, TP53BP2, BTG1, NMT1	CUL3, KAT2B, HIP1	RBM38, MAP3K5, APP	HIP1	HIP1	MAP3K5
Iron homeostasis		UROD	FBXL5, SLC25A28					
Oxygen homeostasis/ response to hypoxia or to oxidative stress	ACTN4, MLH1	HMBS, STAT5B, ERCC2, NRF1, PTK2B	CAT, CITED2, MLH1, OXSR1, PARK7	ACTN4, PTK2B	ECE1, NFKB1, NR4A2	IPCEF1		CAT, IPCEF1
Primitive hematopoiesis	STK3							

Table 3.3: Genes from Mello et al. [232] represented in the closest genes set of 500 keyword regions for each of the eight BLDA topics grouped by function.

3.4 Discussion

Recently, the use of ATAC-seq to characterize chromatin accessibility in varied cell systems has resulted in an excess of high quality data. However, it is currently difficult to integrate these data and perform inference for common regulatory programs. In this chapter, I adapted the cisTopic approach for using LDA to simultaneously infer weighted collections of co-accessible regions and the cell types in which they are active. I show that the method, BLDA, is able to identify a similar number of differentially expressed regions (as defined by EdgeR) as the established scATAC-seq pipeline in a collection of pseudobulked single cell experiments with known cell types. The modification was essential to its comparative success, as a naive implementation of the cisTopic algorithm using region thresholding and one-hot encoding did not identify any of the same differentially accessible regions as the single cell analysis. BLDA additionally identified topics which were much more specific to target cell types than the naive implementation, as is expected in a purpose-built dataset of dissimilar cell types (Figure 3.7). This simple dataset shows that BLDA is able to recreate some of the success of the single cell method in bulk samples, so I next turn to a well understood system of differentiating cell types. Erythropoiesis, the process by which HSCs differentiate into red blood cells, or here more specifically immature erythroblasts which have not yet undergone enucleation, has an extensive literature documenting genes which are up and down regulated at various check points [200]. We take sequencing data from Ludwig et al. [200] and Corces et al. [202] and create a pseudo-timed differentiation trajectory incorporating hematopoietic precursors and erythropoietic check points, as well as stages of the lymphoid differentiation pathway incorporating CD4 and CD8 positive T cells. I show that in this system as well, the BLDA method out performs a naive implementation of cisTopic, identifying specific topic loadings that are enriched for the regulatory biology of erythropoiesis. Inferred topics identify regions of the genome which are shown to be differentially accessible in similar patterns to topic loadings across pseudo-time such as Figure 3.17a and Figure 3.17b.

Having established that BLDA has power to find biologically enriched pathways within topic loadings, the next chapter applies the algorithm to large scale datasets of ATAC-seq experiments like ENCODE and uses them to investigate cell types with relatively unknown regulatory grammar.

3.5 Data and Code Availability

Single cell sequencing experiments from [194] are available via SRA, and scripts for performing peak calling and pseudobulking are distributed with the BLDA package at `https://github.com/Chris1221/BLDA`. This package also contains functionality for creating count matrices from BAM files and bigWig tracks in order to run the adapted analysis pipeline. A modified version of cisTopic necessary for running these analyses is available at `https://github.com/Chris1221/cisTopic_bulk`.

Data for the erythropoiesis developmental trajectory is available from GEO, with more details about specific file formats available from the original publications Corces et al. [202] and Ludwig et al. [200].

An implementation of LanceOTron peak caller for batch data is available from `https://github.com/Chris1221/LanceOTron`.

A Snakemake pipeline for replicating the entirety of the analyses presented in this chapter is available in the chapter_3 subdirectory of `https://github.com/Chris1221/thesis_lda` repository.

3.6 Acknowledgments

The original conception of the BLDA extension, that is incorporating read counts rather than binary accessibility into the count matrix, came from Alastair Smith. Single cell experiments were downloaded and combined into a single BAM file by Emine Ravza Gur. Alignment and quality control of the datasets from Corces et al. [202] and Ludwig et al. [200] was performed by Damien Downes.

Once you've got a task to do, it's better to do it than live with the fear of it.

— Joe Abercrombie, *The Blade Itself*

4

Topic modelling identifies novel PAF1c-bound enhancer elements in MLL-AF4 leukemia patients

Contents

4.1 Introduction

Substantial advances in the treatment and supportive care of children with acute lymphoblastic leukemia (ALL) have resulted in over 90% of patients surviving beyond five years of diagnosis [238]. Despite this progress, several subsets of high-risk paediatric ALLs have remained mostly resistant to current treatment regimes [238]. Chromosomal translocations of the MLL (also known as MLL1 or more recently k-methyl transferase 2A (KMT2A)) gene are a major cause of incurable paediatric and infant leukemias, and represent between 5-10% of ALL cases across all age groups [239]. MLL translocations fuse the MLL gene in frame with over 100 different partner genes, creating novel fusion proteins (MLL-FP) such as MLL-AF4 [240]. The MLL fusion protein (MLL-FP) changes the transcriptome sufficiently to cause leukemogenesis. This disease represents one of the rare instances where a single genetic change is sufficient to cause cancer, in essence a single oncoprotein directing the performance of all ten canonical hallmarks of cancer [241, 242]. An understanding of this minimal set of carcinogenic regulatory pathways is therefore of generalisable interest, and a study of this system has already yielded actionable insights into other, more prevalent, diseases. Examples include BET, LSD1, and Menin inhibitors, each originally shown to be effective in cellular and xenographed murine models of MLLr leukemia [243–246]. However, the active regulatory programs involved in this disease are only just beginning to be elucidated, and the extent to which MLL-FPs create or reuse existing pathways is not well understood. One method to assess the complement of active regulatory elements defining a certain cell type is through NGS-based chromatin accessible assays such as ATAC-seq [184,

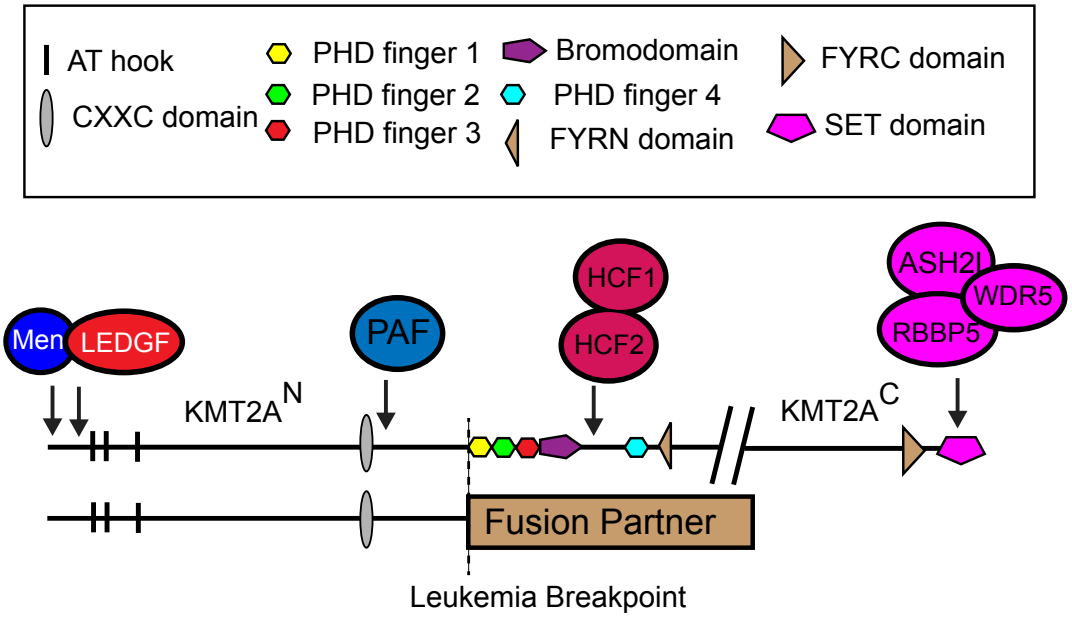


Figure 4.1: Major domains and protein interactions of MLL (KMT2A) and their loss during in the fusion protein.

247]. In this chapter, I adopt the previously developed machine learning topic modelling approach to study and subsequently character patterns in accessible regulatory elements within MLL-AF4 patients.

MLL is expressed ubiquitously throughout haematopoiesis, playing an crucial role in maintaining normal numbers of progenitors and ensuring the availability of adult HSCs [240, 248, 249]. Its deletion is embryonically lethal in mice, illuminating its role in stage specific control of gene expression [250]. Much of the original work of characterizing its function as a transcriptional regulator came from model systems due to a close homology to Trithorax in *Drosophila* and similar activity to SET1 (aka COMPASS) in yeast [251]. In wild-type MLL, a C-terminal SET domain confers histone 3 lysine 4 (H3K4) methyltransferase activity while the N-terminal contains a CXXC domain which binds to unmethylated CpG islands [252–256]. In the chromosomal translocation event, many of the functional domains of the MLL protein are lost, including four PHD fingers and the SET domain, leaving three AT hooks and the CXXC to determine the binding preferences (Figure 4.1)

In MLL-FP driven leukemias, the N-terminus instead associates with one of up to one hundred and twenty C-terminal proteins, conferring a huge range of functional activities and interrupting its normal function as a H3K4 methyltransferase [253]. The fusion (MLL-FP) is able to recruit new co-factors such as DOT1L, whose aberrant methylation profile at H3K79 sites is highly discriminatory for genes activated by the fusion protein and believed to be crucial for the transforming activity of MLL-FPs [257, 258]. Identifying distinct pathways that the fusion protein activates requires the identification of a potential cell of origin to compare against. Previous research has shown that, especially in infants, ALLs caused by MLL-FPs represent an early B cell developmental block [259, 260]. Despite initially promising results in cell lines, clinically targeting DOT1L has not been generally effective in leukemia patients, though some early evidence suggests that targeting DOT1L alongside other key cofactors like menin may show more success [261]. This demonstrates the resiliency and redundancy in MLL-FP driven leukemias, and the need to understand the full complement of epigenomic changes involved in leukemia maintenance against the backdrop of developmentally normal early B cell progenitors. An understanding of these pathways may also lead to a better understanding of oncogenic pathways more generally.

Recent research points to disturbances in the epigenetic regulation of gene expression as a potential mechanism for leukemogenesis [262]. The amount of expressed gene product in a specific cellular system represents a complex interplay between DNA binding transcription factors occupying sequences at enhancers and promoter elements [254]. The ability of these transcription factors to bind is due in part to the physical accessibility of the chromatin at these elements [99]. The fundamental unit of eukaryotic chromatin, the nucleosome, is regulated and enzymatically modified by a variety of protein complexes that both promote the recruitment of additional regulatory proteins and mobilize nucleosomes to allow polymerase direct access to the sequence [199, 263]. These nucleosomes consist of octomers of histone proteins, whose terminal regions (known as “tails”) are frequently post-translationally modified through methylation, ubiquitylation,

phosphorylation, and acetylation [105]. A summary of relevant histone modifications for this investigation is presented in Section 1.3.1.2. Emerging evidence suggests an intricate relationship between the histone modifications present at a genomic region and its ability to recruit proteins involved in many stages of transcription [104, 107, 190, 264]. The deposition and removal of these histone marks is orchestrated in a cell type specific manner, such that enhancer and promoter elements active in a sample may be identified at scale by examining the enrichment of specific histone modifications. For instance, enhancer elements are stereotypically marked with H3K27ac (acetylation of the 27th lysine residue on the H3 histone) and H3K4me1, while promoters show the same H327ac but higher enrichment of H3K4me3 [265]. In theory, the presence of both transcription factors and regulatory histone modifications at a particular genomic location can be experimentally assessed by ChIP-seq (as described in Section 1.3.1.2). However, a full investigation into differentially active enhancers and promotors would require performing ChIP-seq independently for each transcription factor and chromatin modification in each cell type of interest. This is infeasible for all but the most heavily studied cellular models. However, a useful alternative exists in the form of ATAC-seq (see Section 1.3.1.1). A cell's accessible chromatin is reflective of its active regulatory landscape, making ATAC-seq a powerful method to identify potential cell type specific enhancer and promoter elements which can be followed up with more careful study through ChIP-seq for relevant histone modifications [186, 189, 266].

In this chapter, I concentrate on the most frequent fusion partner of MLL, AF4 (also known as AFF1) [240]. AF4 belongs to a higher-order protein complex implicated in transcriptional regulation and involving many of the most frequently-fused MLL partners [267]. Recent evidence has suggested that the MLL-AF4 fusion protein is involved in a feed-forward gene regulatory network (GRN) involving transcription factors such as RUNX1 alongside some which have no previously-described function [268, 269]. Outside of well-characterized direct binding events where MLL-AF4 is able to regulate the expression of key genes such as proto-oncogene MyC and BCL2, the direct and indirect regulation of transcription by

the fusion protein remains an active area of research [268]. An investigation into the patterns of accessible chromatin in leukemic samples would provide insight into the differential usage of enhancer and promoter elements as well as the pathways which may drive their use. Recently, a well-characterized set of $CD19^+$ lymphoid progenitors was identified in infants [270]. These include a $CD10^+$ ProB (PB) cell type resembling adult progenitors and a completely novel Pre-ProB (PPB) $CD10^-$ cell type almost undetectable in adult bone marrow and with a gene expression profile resembling ALL blasts [270]. These cells represent the earliest identified committed lymphoid progenitors. The authors suggest that these cells are a candidate for the cell of origin for this leukemia. As such, they represent the closest known reasonable system to represent developmentally "normal" lymphoid cells early in their differentiation.

This chapter aims to use topic modelling to describe regulatory programs within four MLL-FP driven leukemia patients, contrasting them against these two closely related healthy fetal cell types, PB, and PPB. We also compare these samples against two cellular models of MLL-AF4, RS4;11 and SEM, both derived from relapsed lymphoblastic leukemias in thirty-two and five-year-old patients respectively [271]. I use the previously developed BLDA method to identify key-word regions discriminative of the different groupings of samples, and study their epigenomic properties. This amounts to an investigation into the accessible regulatory regions and pathways active in this uniquely difficult to treat form of leukemia.

4.2 Results

The focus of this chapter is an in depth investigation into using topic modelling for MLL-AF4 cells. Here, we aim to discover novel pathways that differentiate MLL-AF4 cells and normal B-cell precursor (BCP). We do this by firstly identifying a reasonable number of topics to analyse further based on patterns in topic sharing that we observe. Secondly, we select a specific number of topics and replicate the analysis a number of times. We do this to assess the contribution of stochasticity to the inference procedure. Gibbs sampling is an approach based on Monte Carlo

sampling, and as such the quality of the sampled posterior distribution can depend on the starting conditions and the convergence of the algorithm. By explicitly taking into account several repetitions of the inference algorithm we attempt to minimize the effect of chance on identified key regions.

We created a unified dataset of open chromatin from cancerous and normal ATAC-seq experiments. We collect four MLL-AF4 patient samples along with two MLL-AF4 cell lines and use latent Dirichlet allocation through BLDA to infer latent weighted collections of regions which may be distinct from the regulatory regions active in six normal lymphoid precursor cells.

4.2.1 Data Processing and Peak Calling

Motivated by its superior performance compared to MACS2 for LDA in Chapter 3, we use LanceOTron to define peaks representing accessible sequence from the different samples [221]. The number of called peaks differed significantly between experiments (Figure 4.2A). The most peaks were found in SEM and RS4;11, and the fewest were identified in Patient 2. We suspected that the low number of peaks identified in Patient 2 was the result of poor sequencing quality. To confirm this, we use `megadepth` to calculate the average read coverage under identified peak regions [272] (Table 4.1). Unexpectedly, patient 2 showed comparable read coverage to the other samples, while Patient 3 had anomalously low read coverage under the identified peak regions. The number of peaks in Patient 3 was comparable to those found in SEM and RS4;11. The remainder of the samples have high coverage values sufficient to identify high quality peaks. Because of the lower quality, we interpret the results of analyses involving Patients 2 and 3 with caution.

We compared two methods, BLDA and cisTopic. cisTopic uses as its input a binary accessibility matrix, while BLDA uses a quantitative measure of relative coverage. To understand the structure of input to cisTopic, we plotted the number of shared peak regions between the samples (Figure 4.2). Patient samples were well differentiated from BCP and the cell lines. Consistent with the number of available peak calls, patient 2 shared the fewest peaks with the other samples. This

Sample Identifier	Sample Alias	Coverage
Patient 11911	Patient 1	884.3
Patient 21940	Patient 2	801.0
Patient 26754	Patient 3	478.8
Patient 27800	Patient 4	809.2
PPB_95R	PPB 1	1361.5
PPB_75	PPB 2	1610.6
PPB_23	PPB 3	1295.9
PB_95R	PB 1	1776.3
PB_75	PB 2	1242.6
PB_23	PB 3	1405.6
RS4;11	RS4;11	701.8
SEM	SEM	725.3

Table 4.1: Average coverage in peak regions for each sample calculated with megadepth.

is explainable by the low numbers of accessible regions in patient 2 in general. Between the two cell lines, SEM shared significantly more peaks with the BCP than RS4;11 does (paired T-test P=2.56e-4). Surprisingly, there was no evidence for increased sharing within either PPB or PB cells (T-test P=0.595). These patterns of sharing will be important when interpreting the raw differential peak topic modelling versus the qualitative BLDA results.

4.2.2 An MLL-AF4 specific accessibility blacklist

Cell lines maintained in culture accumulate genetic differences over time, some of which may be advantageous to their survival in culture [273]. Latent structural differences between the cell lines in question and the reference genome pose a particularly concerning confounding factor to the interpretation of topic modelling; differences in mappability between regions reflected in abnormal peaks unrelated to biological differences between cell types would contaminate pathway and motif enrichment analyses. To address concerns associated with structural diversity in the RS4;11 and SEM cell lines that were used in our analyses, we constructed a list of regions whose enrichment is solely due to technical issues such as biases in mappability.

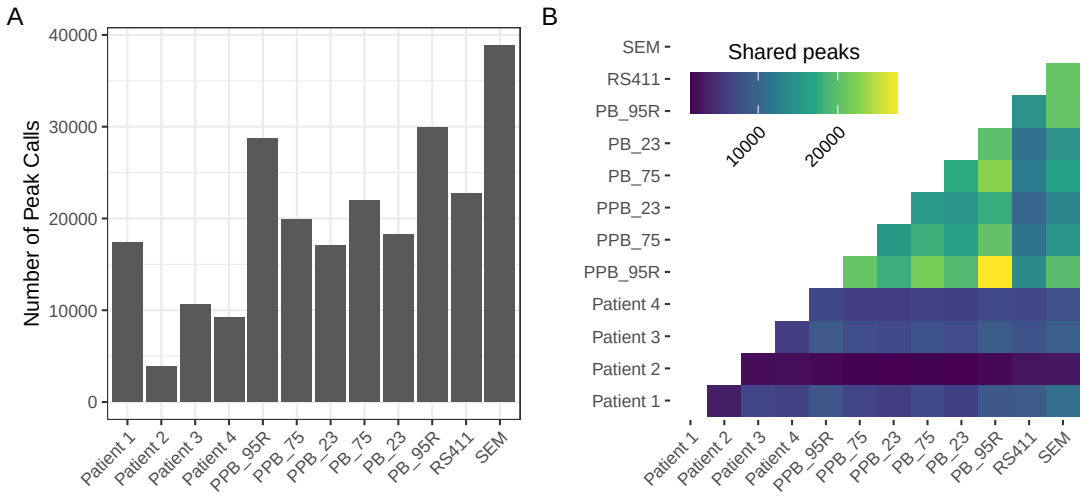


Figure 4.2: Peak calling of MLL-AF4 and B Cell Precursor cells with LanceOTron. A. Raw number of peak calls per cell. B. Number of shared peak calls between pairs of cells.

As a proxy for regions with mappability and other technical issues, we collected reference tracks for ChIP-seq experiments conducted on these cells. Briefly, these input tracks represent sonicated DNA that has not been pulled down with an antibody, and therefore devoid of biological meaning. Peak calling is performed using Macs2 with an extremely stringent Q value cutoff, 0.00001, in order to preserve as much of the accessible genome as possible while eliminating the most obvious signals of technical artifacts. The resulting merged blacklist contains 21068 short regions (average $\pm$ standard deviation length = 483 $\pm$ 342 base pairs) covering 10.1 megabases of sequence in total. This blacklist is combined with the blacklist of accessible regions from the ENCODE project [274]. The combined list covers 21.3 megabases of sequence, which we exclude from all subsequent analyses.

To justify this approach and the necessity of removing blacklisted regions, we perform the entire LDA analysis using unfiltered data, inferred topic loadings not shown. We performed region selection using the top 100, 250, and 500 regions in each topic for $k = 5, \ldots, 10$ topics and found the strict overlap between the identified regions and those identified as being a part of either the ENCODE blacklist, or the ENCODE blacklist plus our custom MLL-AF4 blacklist (Figure 4.3). Overall, between 20.1% and 48.8% of the selected regions for different values of k and the

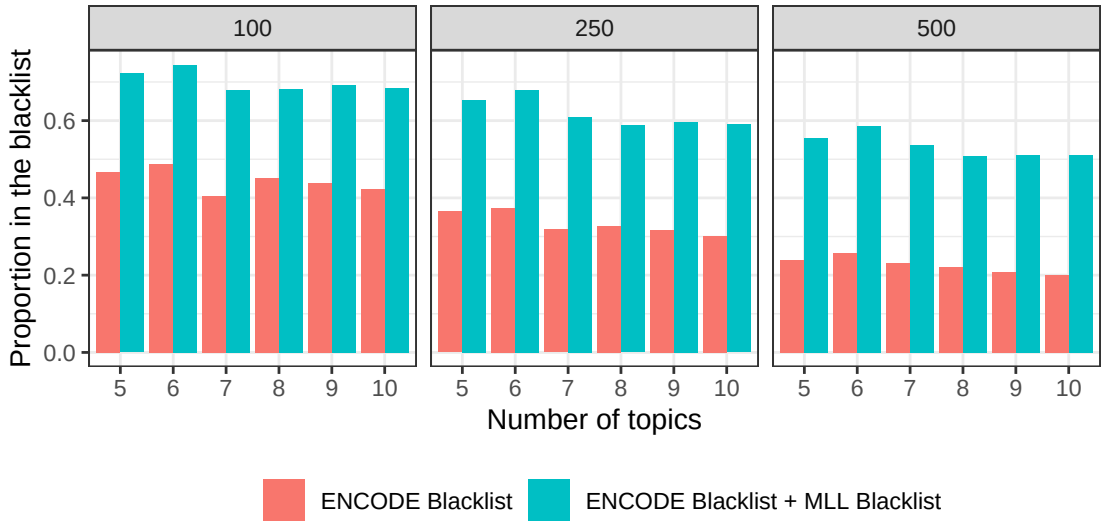


Figure 4.3: In an unfiltered LDA analysis of MLL-AF4 and B cell precursor cells, blacklisted regions made up the majority of identified keyword regions. Plotted is the total number of regions overlapping either of the two blacklists for a given k value and the top number of regions indicated in the facet label.

thresholds overlapped with ENCODE blacklisted regions. Between 51.1% and 74.5% of the same additionally overlapped with the custom-made MLL-AF4 blacklist. This is compared to an overall overlap of 5998 ENCODE blacklist regions from 249903 total regions for the analysis (2.4%) and 83434 total regions overlapping with the MLL-AF4 blacklist (33.3%). The blacklisted regions were therefore over-represented in the keyword regions, as expected. This analysis reaffirms the need to filter blacklisted regions.

4.2.3 Differential Accessibility between B cell Precursors and MLL-AF4 cells

After filtering out regions which overlapped with either the ENCODE or MLL-AF4 blacklist, we construct a count matrix using RPKM normalized read counts to test for baseline differential accessibility. We initially test for differential accessibility between the MLL-AF4 samples (Patients 1 through 4, RS4;11, and SEM) and the BCP (PPB 1-3, PB 1-3). edgeR is used to identify differentially accessible regions between the two groups based on read counts. Overall, 27970 of the total 64162 regions were differentially accessible at a Q value threshold of 0.05. This is,

however, too large a number to reasonably analyse in depth, so we select the top 100, 250, and 500 differentially accessible regions, all of which achieved statistical significance of Q-value less than 0.05. We use GREAT to conduct an association analysis between these regions and pre-defined biological pathways [275]. The background against which enrichment should be calculated is set to be union of all peak regions from the samples, in order to mitigate the inherent bias of analyzing accessible regions in the genome. All three subsets were enriched for Gene Ontology pathways, including peptidyl-threonine phosphorylation (4.97 fold enrichment, Q = $2.0e-7$), peptidyl-threonine modification (4.31 fold enrichment, Q = 2.4×10^{-6}), peptidyl-serine phosphorylation (2.59 fold enrichment, Q = 0.019) and nucleic acid phosphodiester bond hydrolysis (2.62 fold enrichment, Q = 0.017). The regions were also enriched for proximity to several genes, all of which with a corrected FDR Q value lower than 0.05 are shown in Table 4.2. Some of these genes are of immediate interest, including RHOU which promotes adhesion of T-cell ALL cells via β_1 integrin, potentially implicating the protein in known impaired transendothelial migration potential of various kinds of leukemia cells [276, 277]. It is difficult to comprehensively survey this list of genes because their relationship with the identified regions is purely statistical. We have no mechanistic reason to believe that they are truly associated with disease.

4.2.4 Topic modelling for MLL-AF4 leukemia

Similar to the above cases, peak calling is performed with LanceOTron and a score threshold of 0.5, as recommended. We construct a count matrix using the normalized read counts under each peak with BLDA as well as a one-hot encoded matrix for comparison. Hyperparameters alpha and beta are set for a given number of topics k using Bayesian optimization and loadings are inferred using cisTopic and BLDA.

We infer topic loadings for $k = 5, \ldots, 12$ topics (Figure 4.4). The upper end of this range was chosen as the number of cells in the analysis. Increase in topic number above this point are of questionable utility, as they force more structure on the data than is present naturally. In both the OHE and BLDA approaches, topics

Gene	Q Value	Fold Enrichment
REXO1L1P	1.41e-33	62.85
PSKH2	3.70e-30	67.37
FOXD4L5	1.29e-15	61.60
CBWD6	2.08e-10	52.50
FRG2B	4.75e-09	74.86
RNF187	4.61e-09	51.33
RHOU	8.52e-07	28.52
FAM27E3	8.54e-07	39.06
DUX4L3	7.51e-06	128.32
FOXD4L4	3.36e-05	102.66
FOXD4L2	7.39e-05	29.61
SERF1A	1.94e-04	73.33
SMN2	3.55e-04	64.16
DUX4L4	6.23e-04	128.32
SPATA31A5	2.12e-03	42.77
ANKRD30B	1.19e-02	28.52

Table 4.2: Enriched genic regions along with their corrected FDR P value (Q-value) from the top 500 regions differentially accessible between MLL-AF4 cells and BCP using edgeR.

are identified that differentiate between the MLL-AF4 and BCP cells. However, the topics identified by BLDA tend to be more specific, loading onto some specific cells like Patients 1 and 2, SEM/RS4;11, PPB, or PB predominantly. No such structure is identified in the cisTopic (right hand) case. For all values of k however, at least one topic is identified for the OHE case which is predominantly active in the MLL-AF4 cells over the BCP.

Higher values of k appear to impose too much structure on the data. This is evident for the BLDA case, which identifies one topic enriched for each cell type (except patient 2) at $k = 12$. The OHE case interestingly deals with the over-parameterization problem differently, preferring to retain the structure seen in lower values of k but assign many different topics with very low loadings within these groups. Though neither of these results are unexpected, the level of granularity for BLDA and non-specificity for OHE make it difficult to study the co-accessible regulatory elements shared between similar cell types. For that reason, we primarily

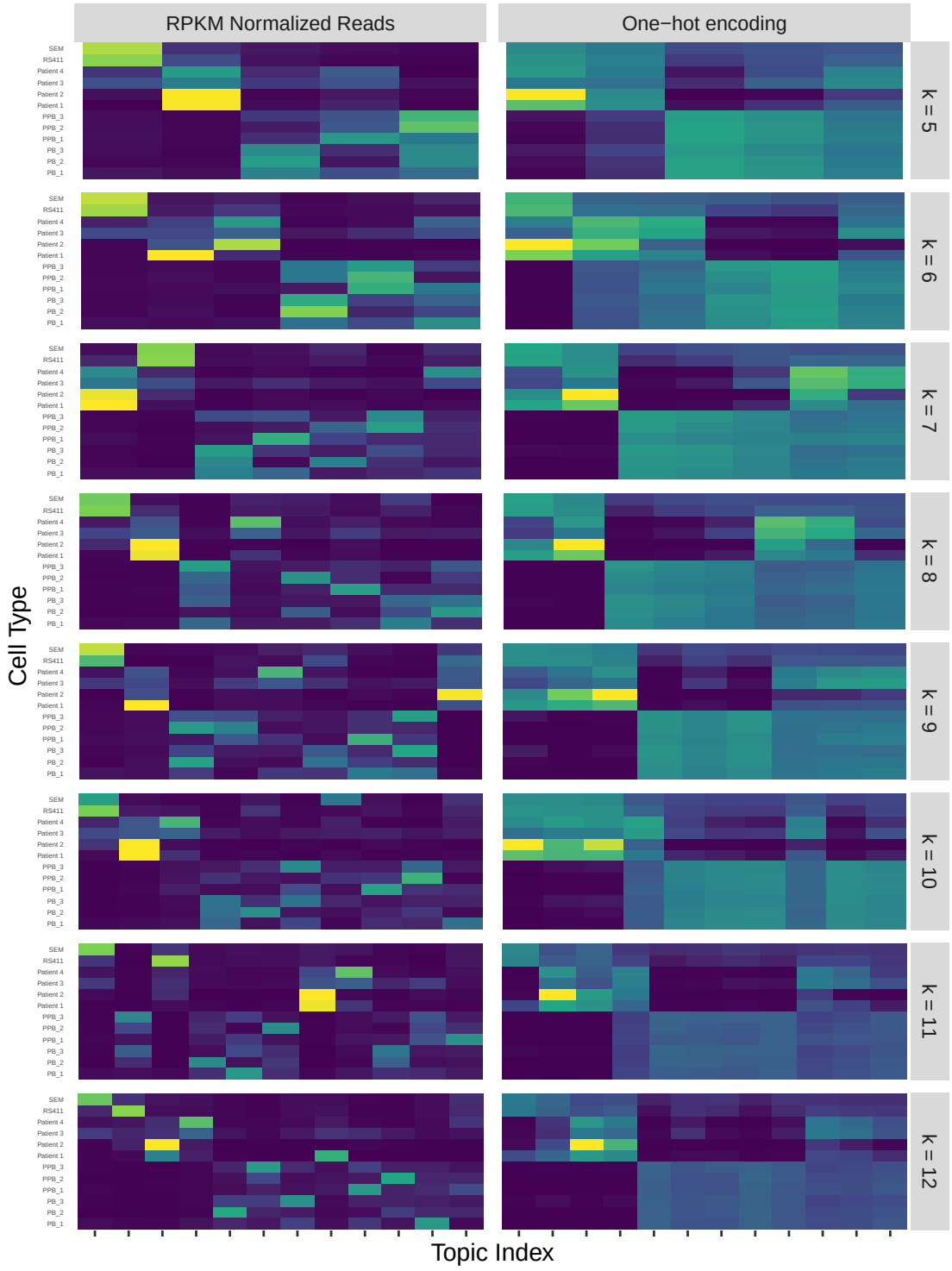


Figure 4.4: Inferred topic loadings for $k = 5, \ldots, 12$ topics comparing MLL-AF4 cells to B cell precursors. RPKM normalised refers to the count matrix used to infer the topic loadings by BLDA, while one-hot encoding simply annotates which regions are called as peaks by LanceOTron. Topic loadings are normalised such that the sum of all topics within a cell equals one.

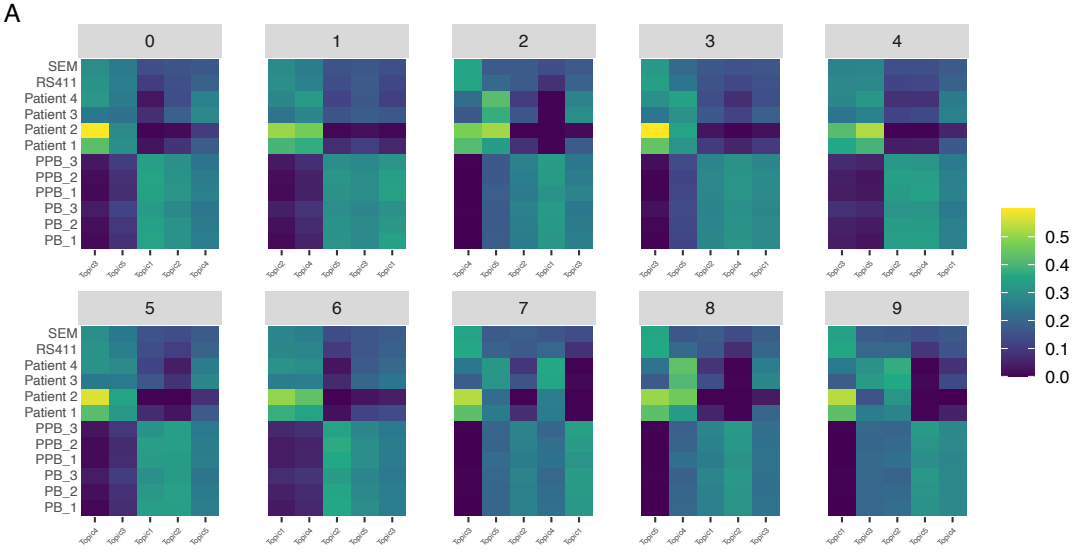
are interested in the analyses for smaller values of k.

We select the simplest model for further analysis. We base our decision on the extremely specific topic loadings identified in the BLDA case (a single topic representing the cell lines, another involved in patients, with some sharing between them) as well as the common structure identified in PB cells separate from PPB cells with a single topic explaining the shared accessibility profile between them both. We repeat the analysis ten times for both BLDA and OHE.

4.2.4.1 Five topic one-hot encoding replication

The $k = 5$ OHE analysis consistently identifies at least one topic which is active in all of the MLL-AF4 cells and none of the BCP samples (Figure 4.5). Patient 2 is usually the most active representative of this group within that topic. Some replicates (i.e. replicates 2, 5, 6, 7, 8, 9) also find topics active in all of the BCP and few of the MLL-AF4 samples. There is no obvious substructure within the BCP samples. We manually annotate each topic as "enriched in MLL-AF4 cells", "enriched in BCP", or neither and perform enrichment for each of the two specific sets relative to the other topics. Motifs found in the MLL-AF4 specific topics in the top 100, 250, and 500 regions include SP4, KLF5, EGR1, and KLF15 (Figure 4.5B, C, D). Ahr::Arnt, ZNF354C, Arid3a, and Arnt are identified but found broadly between MLL-AF4 and BCP regions. The latter were additionally enriched for BHLHE22 and EBF1 which is a marker for B cell lineage commitment [277] along with PAX5 which is occasionally identified in BCP specific regions (data not shown). The list of motifs presented is reasonably unspecific, though there is evidence that EGR1 has deep functional relationships with many hematological malignancies including ALL [278]. The fact that there is at least one known system specific transcription factor whose motif is enriched for each set of samples is reassuring. It is additionally reassuring that these important motifs were identified in each of the ten replicates in their respective sets. Even in this low-resolution OHE setup, the technique has some power to robustly identify relevant biology across replicates.

A



B

Motifs	N
Ahr::Arnt	6
KLF5	5
SP4	5
ZNF354C	5
Arnt	4

C

Motifs	N
EGR1	10
KLF5	10
ZNF354C	10
Ahr::Arnt	9
Arnt	9

D

Motifs	N
Ahr::Arnt	10
EGR1	10
Arnt	10
Arid3a	10
KLF15	10

E

Motifs	N
SP4	5
Ahr::Arnt	4
ZNF354C	4
Arid3a	3
Arnt	2

F

Motifs	N
ZNF354C	10
Ahr::Arnt	10
Arid3a	8
Arnt	8
BHLHE22(var.2)	6

G

Motifs	N
ZNF354C	10
Ahr::Arnt	10
Arid3a	10
Arnt	10
EBF1	10

Figure 4.5: Ten replicates of topic modelling using one-hot encoded input. A. Inferred topic loadings on each of the samples for all 10 replicates, sorted by median topic activation. B, C, and D show the top 5 motifs identified in N of the replicates in the top 100, 250, and 500 regions of manually annotated MLL-AF4 specific topics. E, F, and G show an equivalent for BCP specific topics.

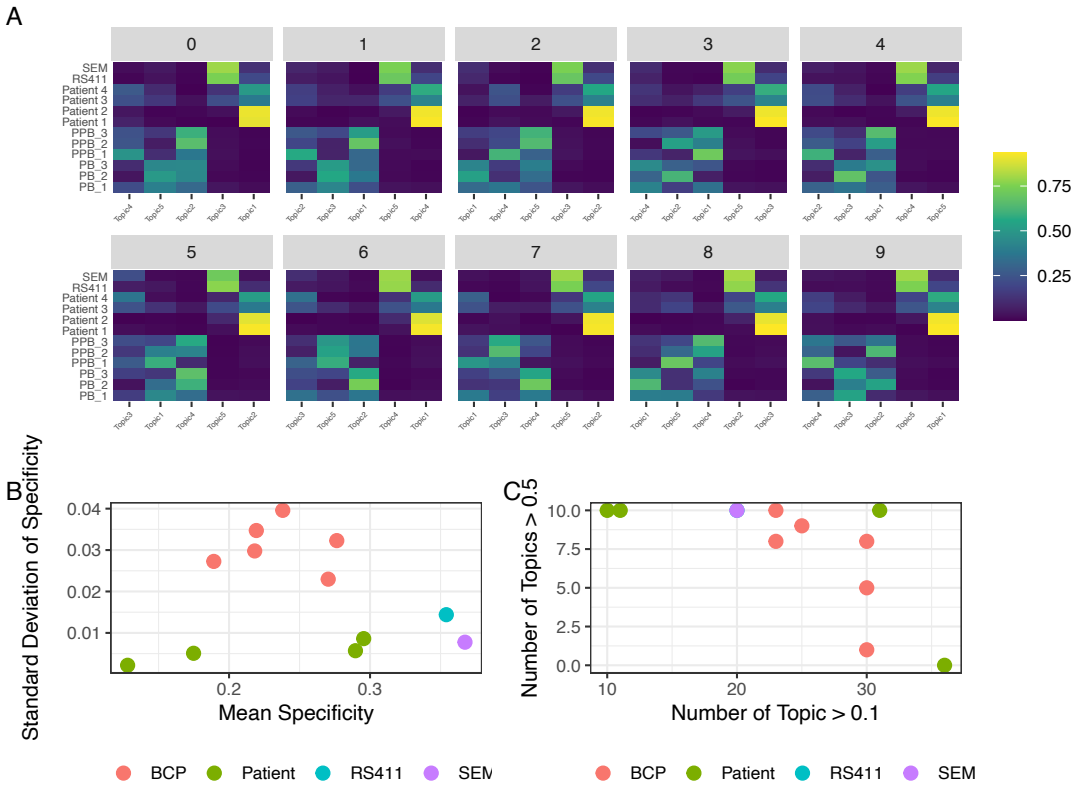


Figure 4.6: Ten replicates of topic modelling using RPKM normalized input, aka BLDA. A. Inferred topic loadings for each of the 12 samples and all 10 replicates, sorted by standard deviation of topic loading. B. Mean versus standard deviation of specificity, here defined as the contribution of the most enriched topic to that sample divided by the sum of that topics enrichment to all other samples. Averaged over replicate. C. The number of instances where a sample has a topic identified above a loading value of 0.1 versus the number of instances where any topic is annotated above 0.5. Values represent contribution of a particular topic to a particular cell, such that the sum of topic loadings within a particular cell equals 1.

4.2.4.2 BLDA replication with five topics

The inferred topic loadings in the BLDA case are also reproducible across replicates (Figure 4.6A). In each of the 10 replicates, one topic loads preferentially onto Patients 1 and 2, and to a lesser degree onto patients 3 and 4. Another loads preferentially onto RS4;11 and SEM, which are also models of the same MLL-AF4 leukemia but do not demonstrate the exact same patterns as the patients. The remainder are somewhat divided between a common co-accessibility program amongst all samples (i.e. topic 3 in replicate 5), specifically loading onto PPB (i.e. topic 5 in replicate

6), or PB (Topic 5 in replicate 0). In general, each of the replicates has at least one topic that falls into each of these categories, though the loadings onto the BCP are less consistent than the MLL-AF4 cells. The topics loading onto RS4;11 and SEM were highly specific to those two samples, in contrast to the BCPs whose specificity was much lower, and additionally much more variable (Figure 4.6B). The BCPs additionally had many topics annotated to them at low levels, but very few with a higher annotation level of 0.5 (Figure 4.6C). This is in contrast to Patients 1 and 2, who had both a large number of highly specific topics across the replicates, and also a very low number of non-specific topics. From this we conclude that similar topics are able to be found consistently across replicates.

Having established a high degree of specificity for individual topics to the expected annotations of the samples, we investigate the regions making up the topics. We briefly focus on a single replicate, indexed as 0 in Figure 4.6. An interesting avenue for exploration is the degree to which the topics are made up of similar regions. If they were, it would represent a shared core of co-accessible elements. The alternative, of completely separate underlying regions, is equally interesting. An understanding of the region-topic distribution over cell types will aid in the interpretation of the key regions for particular topics. We transform each region-topic vector to Z-scores by subtracting the mean and dividing by the standard deviation of the distribution, and count the number of regions with a Z-score over 3 (98.8 percentile of a theoretical standard normal distribution) in each of the topics (diagonal of Figure 4.7A). We additionally count the regions with a Z score of 3 or higher in both of two different topics and find generally low region sharing overall, except in the case of topics 2 and 5 where 330 regions were enriched for both(off-diagonal elements of Figure 4.7A). Topics 2 and 5 in this replicate were enriched for PB cells and all BCP respectively. The regions selected as a part of this set had higher read counts in BCP samples than MLL-AF4 samples (Figure 4.7B). This indicates a high degree of co-accessibility between important regions in PB and PPB cells. In contrast to this, few regions were shared between the two MLL-AF4 topics. The regions with topic 1 loadings over a Z score of three

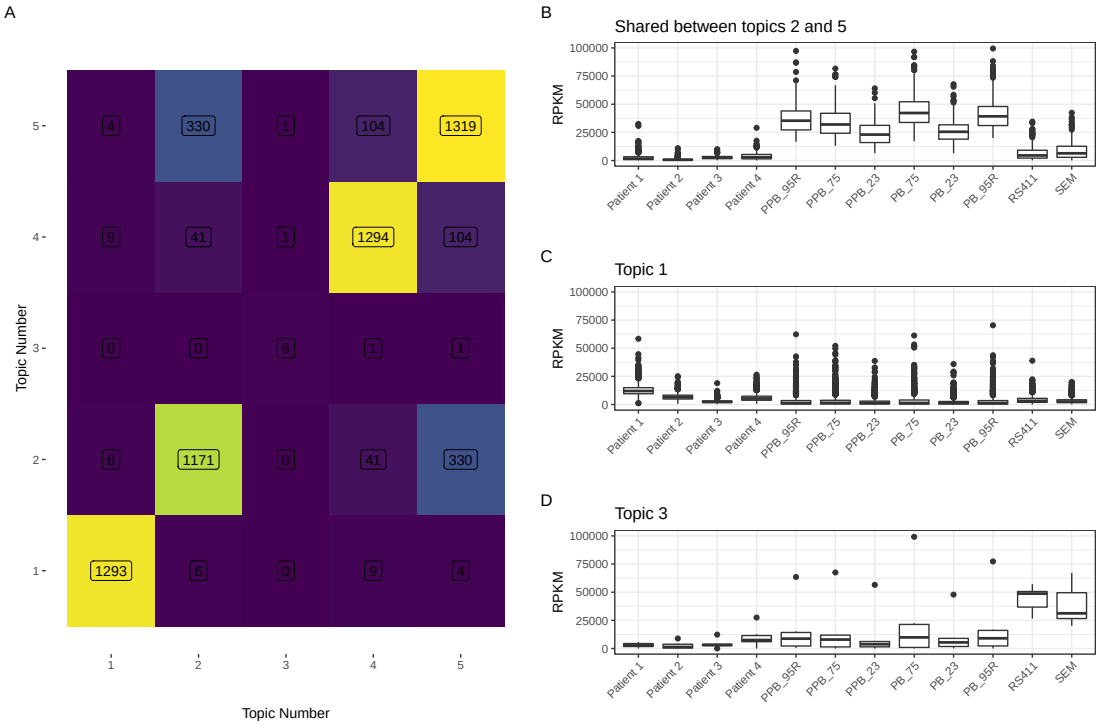


Figure 4.7: Regions highly annotated on each of the five topics in replicate 0 for the BLDA topic modelling. A. Region-topic loadings were converted to Z scores by subtracting the mean of the distribution and dividing by the standard deviation. Regions with a Z score of 3 or more were selected, and the number of shared highly loaded regions is plotted. The number identified per topic is plotted on the diagonal. B. RPKM values for the 330 regions highly loaded for both topics 2 and 5. In this replicate, these topics were enriched in BCPs. C. RPKM values for regions highly loaded with topic 1, which primarily loads onto patient samples. D. RPKM values for the 6 regions highly loaded with topic 3, which is enriched in RS4;11 and SEM samples.

were more accessible in the patient samples, though this was not a statistically significant difference (Figure 4.7C). However, the small number of important regions for the RS4;11/SEM topic were more accessible in these cell types (Figure 4.7).

4.2.4.3 Sample-specific key-word regions across replications

We sought to understand how reproducibly these regions were identified between replicates. Regions robustly identified as being associated with a topic that loads highly onto a desirable set of samples represent key candidates for further investigation. We identified the top 100, 250, and 500 regions for each topic in each replicate, and found how many of these regions are shared between 8, 9, or all of the replicates (Figure 4.8A). Patients have the highest key-word topic

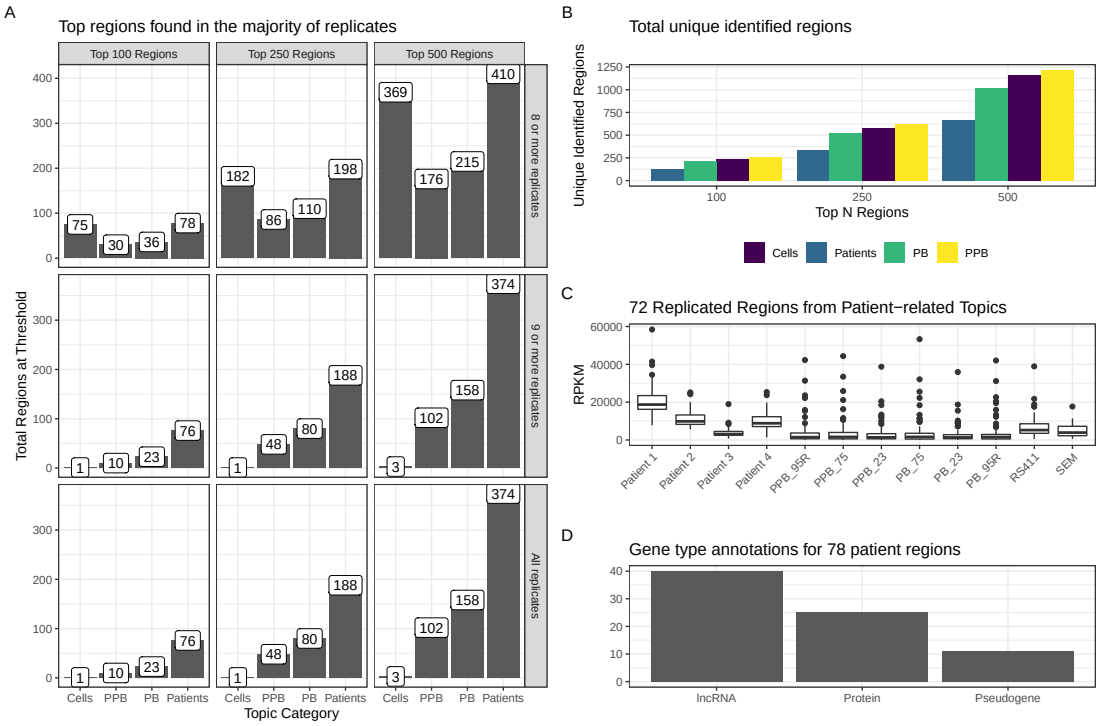


Figure 4.8: A subset of patient specific regions are highly reproducible across replicates and are enriched for lncRNA genes. A. For each of the 100, 250, or 500 top regions per manually annotated sample specific topic, the number of regions occurring in eight or more, nine or more, or all of the ten replicates is plotted. B. The intersection of all regions identified across replicates for each set of sample-specific topics. C. RPKM profile for the 76 regions identified in A as being highly reproducible. D. Gene type annotation from refseq for closest annotated genes for the same regions.

reproducibility across replicates, with 76 of the top 100 identified in every single replicate. The patient group has the lower number of unique regulatory elements identified for any threshold (Figure 4.8B). Amongst the 76 reproducibly identified regions, average expression is almost threefold higher in patients than the remainder of samples (mean of patients = 11321.8, mean of other = 4291.5, Student's T-test P value for difference < 2.2e-16). These regions therefore represent a prioritized set that is both important for the highly-specific topic modelling approach and identified in each of the ten replicates.

4.2.5 Annotating reproducibly identifiable patient-related regions

We begin our characterization of these regions by finding their closest annotated gene body. 32 of the 76 regions lie directly within genic regions. The remainder tend to be found within 50kb of the nearest gene, but the distance is highly variable and some regions are as far as 200kb from their closest gene (Figure 4.9A). This distance distribution fits well to our expectations if the regions represented intergenic enhancer elements. The closest annotated genes are, interestingly, long non-coding RNAs (lncRNAs), and are found in a much higher frequency than expected in 76 randomly samples genes (Figure 4.9B,C). Though this result is suggestive, lncRNAs are very challenging to correctly annotate, with efforts being confounded by both technical artifacts and methodological complications [279, see Figure 1 for an overview of challenges associated with lncRNA annotation]. There is, however, growing evidence that cancer cells, including leukemic blasts, may hijack the large and poorly understood lncRNA transcriptome to influence differentiation, energy metabolism, malignant proliferation, apoptosis, and the drug resistance of leukemia cells [280, Table 1]. It is possible, therefore, that the regions identified represent CREs influencing the expression of functional lncRNAs involved in each individual patient's leukemia. However, none of the lncRNAs identified match with those in Table 1 of Gao et al. [280], so their potential functions will remain the subject of further investigation.

4.2.5.1 Histone modifications and transcription factor binding in the patient-related regions

Histone proteins within nucleosomes may be post-translationally modified by the addition of several chemical groups which collectively serve to regulate transcription and the recruitment of transcription factors. The presence of specific histone modifications at a particular location of the genome can be determined experimentally using ChIP-seq, where peak regions give an estimate of the presence or absence of a particular modification. There are many histone modifications, and their

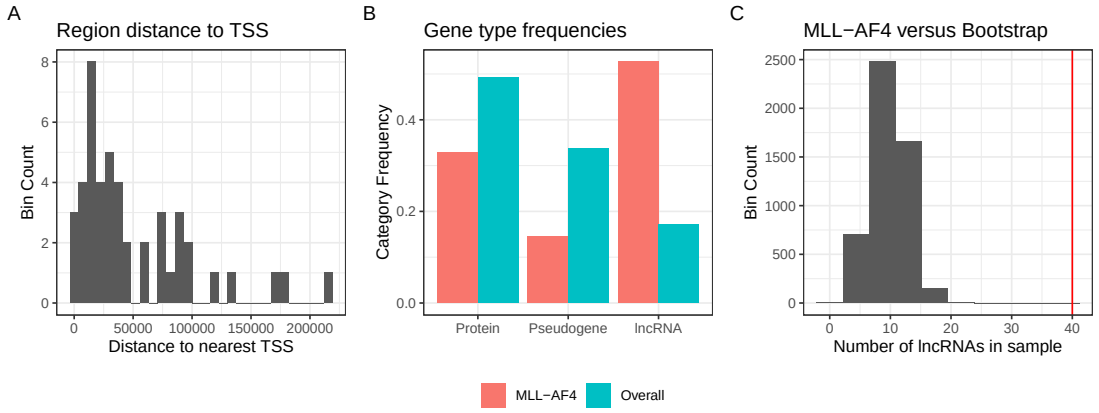


Figure 4.9: MLL-AF4 regions compared to reference genes. A. Distance of the 76 reproducible patient regions to their nearest annotate TSS, minus the regions which are sitting in genic regions. B. Frequencies of gene categories in refSeq versus the sample of closest patient genes. C. Frequency of lncRNAs in 5000 random samples of 76 regions from refSeq, showing that the observed frequency of lncRNA is in the far tail of this distribution.

combinatorial method of action makes it difficult to interpret the exact ramifications of their presence [281]. However, specific combinations of histone modifications have been successfully used to demarcate the genome and identify putative active and repressed enhancers and promoters. A small number of important histone modifications and transcription factor binding sites were characterized in these samples using ChIP-seq (see Methods).

We called peaks on each of the ChIP-seq tracks using LanceOTron and intersected the peaks with the reproducible set of patient regions. We additionally drew 1000 random samples from the set of accessible regions used in this section, creating an empirical P-value, to understand the relative enrichment of these ChIP-seq signals versus the genomic background of accessible sequence. Within the available ChIP-seq data, regions are relatively enriched for H3K27ac, H3K4me1, and H3K4me3 (Table 4.3). They are depleted in K379me2, and there is evidence that we see more binding of MLL to these regions than we would expect by chance, especially in patient 11911, where we additionally see increased binding of AF4, PAF1c, and RUNX1. The relative abundance of H3K27ac and H3K4me1 indicate that these elements may be acting as enhancers. Interestingly, H3K4me3 also appears to be

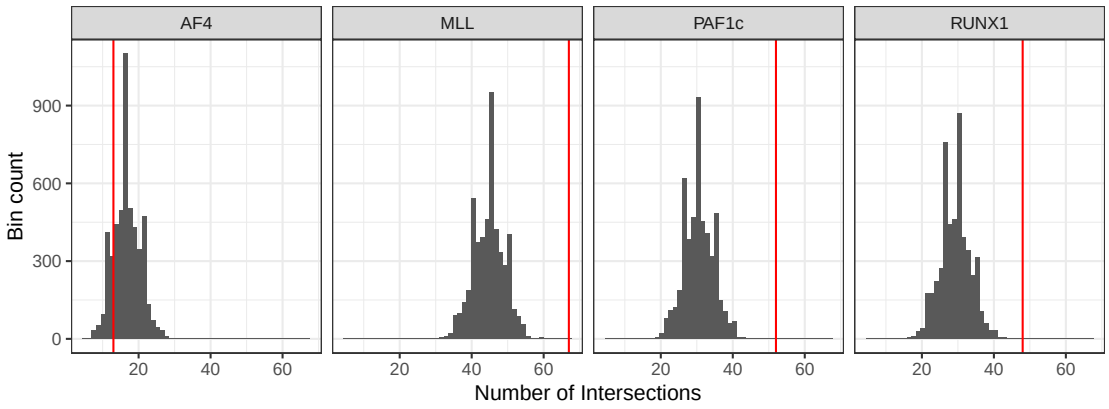


Figure 4.10: Number of intersections between 76 randomly selected putative enhancer sites (accessible regions which overlapped with both H3K4me1 and H3K27ac chromatin peaks) in 5000 random samples compared to the observed quantities in the reproducible patient regions (plotted in red as a vertical line).

statistically enriched in this sample, albeit to a lower degree. This may indicate a mix of enhancer and promoter elements, or the possible implications of controversial "super-enhancers". We describe this result in more depth in the Discussion. However, it is not clear whether the high abundance of PAF1c, RUNX1, and MLL marks are due to general enrichment in enhancer regions or are somehow associated with these regions functionally. As patient 11911 is the only available sample with all of these chromatin marks and transcription factors available, we elect to study this patient specifically. We create a set of putative enhancer elements by selecting 18101 regions from the total 64162 accessible regions that overlap with both H3K27ac and H3K4me1 peaks. We then perform a similar analysis to before, selecting 5000 sets of 76 regions at random and observing the proportion of the putative enhancer elements that are additionally bound by MLL, AF4, PAF1c, and RUNX1. The resulting distributions show that none of the 5000 random samples are as highly enriched for MLL, PAF1c, or RUNX1 as the reproducible patient regions (Figure 4.10). MLL binding is enriched by approximately 50% over the mean of the distribution (67 versus 44.8), RUNX1 binding is 63% higher than the mean (48 versus 29.3), and PAF1c is enriched by nearly 70% (52 versus 30.6). We discuss the relevance of these factors in the Discussion section.

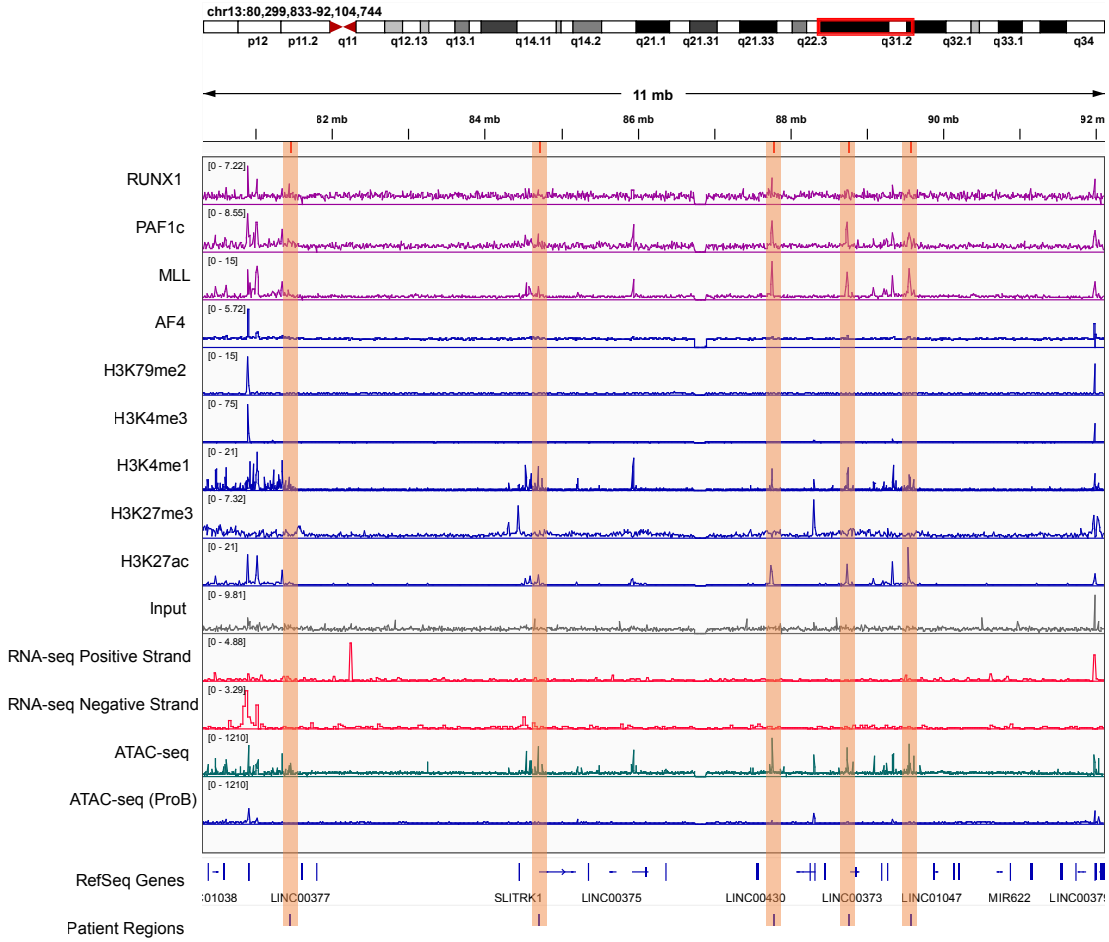


Figure 4.11: Coverage tracks for ChIP-seq marks, RNA-seq, and ATAC-seq for patient 11911 as well as PB cell for comparison. Five regions are in close proximity to each other, each marked with H3K4me1, H3K27ac, MLL, and PAF1c. Regions of interest are highlighted in orange, with the highlight extending slightly beyond their borders for visual clarity.

Patient ID	H3K27ac	H3K4me1	H3K4me3	H3K79me2	MLL	AF4	H3K27me3	PAF1c	RUNX1
26754	57 (0)	60 (0)	31 (0)	2 (0.968)	42 (0)	4 (0.631)			
11911	66 (0)	72 (0)	32 (0)	2 (0.975)	67 (0)	13 (0.004)	1 (0.933)	52 (0)	48 (0)
27800	60 (0)	61 (0)	38 (0)	3 (0.956)					
25911	23 (0)				24 (0)				

Table 4.3: ChIP-seq peak overlaps with reproducible patient regions. The number shown in the column represents the number of the regions overlapping with the specified ChIP-seq mark (out of a total of 76). We construct an empirical P-value for this overlap value (displayed in brackets for each value) by taking 5000 random samples of 76 accessible regions from the complete set of peaks and finding the proportion exceeding the observed count for each sample. Zero values indicate that none of the random samples overlapped with the ChIP-seq mark more than the observation.

Region	Function	Gene	Cell line reference
chr1:239881927-239883919	Promoter	CHRM3-AS2	
chr2:133023823-133025309	Enhancer	AC098826.5;	GM19238;
		CDC27P1;	GM19238,Pancreatic_islet;
		GPR39;	Pancreatic_islet;
		LYPD1	Pancreatic_islet
chr3:97628722-97630587	Enhancer	ARL6;	HT29;
		AC110491.1;	HT29;
		CRYBG3;	HT29,PC3,th1,Thymus;
		MINA	HT29,Mesendoderm,NB4,PC3
		IFT80;	HT29;
		RP11;	HT29,NB4;
		TRIM59;	HT29;
		SCARNA7;	HT29;
chr3:160554953-160556142	Enhancer	ARL14;	HT29;
		RP11;	HT29,NB4;
		PPM1L;	HT29,NB4;
		RP11;	HT29,NB4;
		NMD3	NB4
chr3:190303470-190305623	Enhancer	IL1RAP;	HFF,hMADS-3,HT29,Keratinocyte,MCF10A, melanoma,Mesendoderm,NB4,PC3,T98G,ZR75-30;
		GCNT1P3;	hMADS-3,Keratinocyte,MCF10A,melanoma,NB4;
		7SK;	hMADS-3,Keratinocyte,MCF10A,melanoma,NB4;
		MTAPP2;	HT29,Keratinocyte,Mesendoderm,T98G;
		CLDN1;	HT29,Keratinocyte,MCF10A,PC3;
		LEPREL1;	Keratinocyte,MCF10A,Mesendoderm;
		CLDN16;	Keratinocyte,NB4;
		CCT6P4;	Keratinocyte,NB4;
		TMEM207;	NB4;
chr4:54721244-54722456	Enhancer	RP11	NB4
		RP11;	HFF,ZR75-30;
		LNX1;	HFF;
		RP11;	HFF,ZR75-30;
		CHIC2;	HFF,ZR75-30;
		RP11;	HFF,ZR75-30;
		PDGFRA;	HFF;
		RPL21P44;	ZR75-30;
		MORF4L2P1	ZR75-30
chr4:177823634-177824977	Enhancer	SPCS3;	HFF;
		RP11;	HFF,LHCN-M2,SK-N-SH_RA;
		NEIL3	Keratinocyte,MCF10A
chr5:82663769-82665217	Enhancer	TMEM167A	ZR75-30
chr5:119723091-119724865	Enhancer	PRR16;	HFF;
		CTD;	HFF,LHCN-M2;
chr6:141130170-141131796	Enhancer	RP11	melanoma
		RP3;	ME-1;
		RP11	ME-1
chr6:161161634-161163661	Enhancer	PLG	GM12891,Namalwa
chr7:11870320-11872428	Promoter	THSD7A	
chr7:81474664-81476523	Enhancer	MIR1255B1	NB4
chr8:77492793-77494807	Enhancer	MRPL9P1;	ESC_neuron,SK-N-SH_RA;
		ZFHX4;	ESC_neuron;
		RP11	ESC_neuron,SK-N-SH_RA
chr9:42018481-42020192	Promoter	LOC102724238,LOC554249,KGFLP2	
chr9:46686996-46688539	Promoter	KGFLP1,LOC554249,LOC102724238	
chr10:33654161-33655412	Enhancer	AK3P5;	FT246,HFF,hMADS-3,HT29,LHCN-M2,ZR75-30;
		RP11;	FT246,FT33,HFF,hMADS-3,HT29,Keratinocyte,LHCN-M2,ZR75-30;
		RP11;	FT246,FT33,HFF,hMADS-3,HT29,Keratinocyte,LHCN-M2,ZR75-30;
		ITGB1;	FT246,FT33,HFF,hMADS-3,HT29,Keratinocyte,LHCN-M2,ZR75-30;
		RP11;	FT246,FT33,HFF,hMADS-3,HT29,Keratinocyte,LHCN-M2,ZR75-30;
		RP11;	FT246,FT33,HFF,hMADS-3,HT29,Keratinocyte,LHCN-M2,ZR75-30;
		NRP1;	FT246,FT33,HFF,hMADS-3,LHCN-M2;
		AL353600.1	hMADS-3
chr10:58119756-58122125	Promoter	ZWINT	
chr12:24531253-24532422	Enhancer	BCAT1	hMADS-3,Mesendoderm
chr13:54602807-54604709	Enhancer	LINC00458;	Mesendoderm;
		RPL13AP25	Mesendoderm
chr13:84710151-84712348	Promoter	LINC00333	
chr13:103658908-103660657	Enhancer	TPP2;	GM19238,GM19239,HFF,Namalwa;
		KDELC1;	GM19238,GM19239,HFF,Namalwa;
		BIVM;	GM19238,GM19239,HFF,Namalwa;
		ERCC5;	GM19238,GM19239,HFF,Namalwa;
		TEX30	HFF,Namalwa
chr14:38593423-38595439	Enhancer	CTD;	Mesendoderm;
		SSTR1	Mesendoderm

Table 4.4: EnhancerAtlas 2.0 results for the 76 reproducibly identified patient regions.

4.2.5.2 Known annotations in EnhancerAtlas

We searched EnhancerAtlas 2.0 (Citation) for regions matching ours, and found that 17 out of the 76 regions were known enhancers, and 6 were known promoters (Table 4.4). The functionality validated active cell types varied considerably, from enhancers for GPR39 and LYPD1 in pancreatic islet cells to more similar TMEM207 and RP11 enhancers in pro-myelotic leukemia cell line NB4. In this list as well, non-coding RNA targets are also over-represented. RP11, for example, was identified as functionally validated target of 10 of the 17 enhancers. Though RP11 represents a class of otherwise unannotated genes, rather than a protein of its own, many subtypes of RP11 are in fact non-coding RNAs (CITATION). As previously mentioned, these regions are clearly marked as putative enhancer regions, though it is not immediately obvious which genes they may be regulating. Interestingly, some regions co-occur in clusters such as the one on Chromosome 13 in Figure 4.11. Here, five regions are found together, each marked with enhancer marks H3K4me1 and H3K27ac as well as the transcription factors MLL and PAF1c. The second of these five regions is annotated as an enhancer for TPP2, KDELC1, BIVM, ERCC5, and TEX30 in GM EBV-immortalized cells as well as a fibroblast cell line HFF-1. Whether the remainder of the enhancers are novel in the MLL-AF4 patients remains to be shown experimentally, however there is no clear accessibility in these regions for BCPs.

4.2.6 Accessibility of patient-related regions within the ENCODE Consortium Blood Cell Collection

To investigate this further, we collect all of the blood related cell types in the ENCODE consortium's list of available ATAC-seq experiments. We exclude any DNAse-seq experiments due to known systematic differences as well as the unproven ability of the method to transfer between accessibility protocols [263]. Without an obvious way to show that differences between DNase-seq and ATAC-seq are not a result of modelling systematic differences, we restrict the analysis to 143 samples within the ENCODE database with available coverage tracks in BigWig format and which are annotated as being associated with the blood tissue system. This

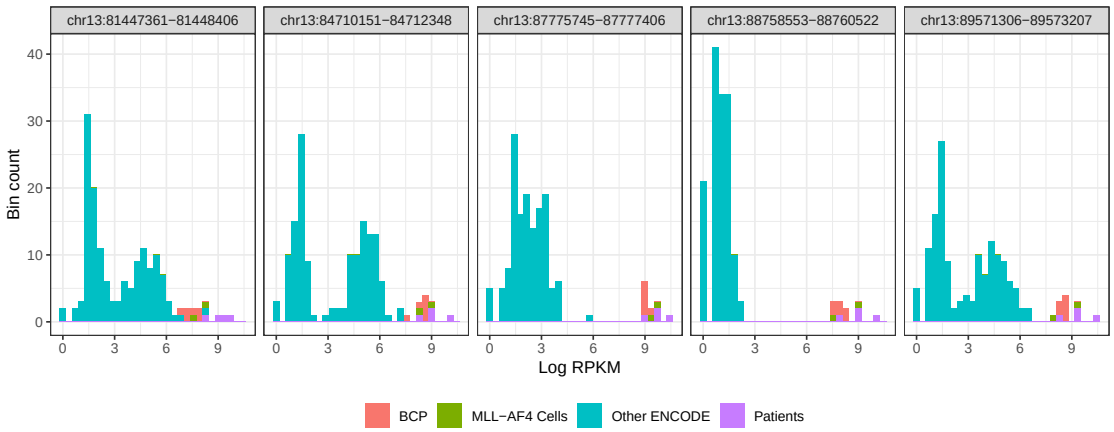


Figure 4.12: Log RPKM for five regions in a putative enhancer cluster identified through topic modelling in ENCODE blood cells versus patients and MLL-AF4 cell types.

data is downloaded and peak called using LanceOTron as before, excluding regions in the ENCODE blacklist as well as the MLL-AF4 specific blacklist developed in Section 4.2.2. We additionally use liftOver to transfer genomic coordinates from the native GRCh38 to hg19 to match the other samples' build. The dataset contained, in total, 326427 regions and 155 ATAC-seq experiments. We begin our validation by examining the previously identified regions in the entire ENCODE blood data set. Interestingly, the five regions contained in the previously described putative enhancer cluster are all most accessible in patients compared to any other cell type in the ENCODE blood cell type collection, suggesting that their function may be new in these patient samples (Figure 4.12). In fact, 74 of the 76 patient regions were the most accessible in patient samples, with the other two being more accessible in BCPs and MLL-AF4 cells rather than the ENCODE dataset.

We perform topic modelling on the entirety of the ENCODE dataset alongside the MLL-AF4 and BCP samples. Hyperparameter optimization is difficult in this system, purely due to computational considerations. Therefore we select reasonable hyperparameters based on the defaults for single-cell analyses, setting alpha to 50 and beta to 0.1, as is the recommended default for cisTopic. We run the analysis modelling $k = 10, 15, 20, 25, 30$ and examine the output (Figure 4.13). The relative sparsity of the topic loadings above $k = 15$ leads to each of the patients and BCP cells essentially being enriched for a seperate topic, which makes interpretation

difficult. The $k = 10$ and $k = 15$ cases both model the structure of the system well, however, especially amongst the most closely related cell types (left-hand side of Figure 4.14). The one-hot encoding does not decipher any different structure within the MLL-AF4 and BCP population for either of the two k values, only loading a single topic strongly and a second weakly to the entirety of the set (right-hand side of Figure 4.14). Some interesting discrepancies between the $k = 10$ and $k = 15$ case include an active B cell topic in $k = 15$ which loads weakly onto BCP cells but not MLL-AF4 samples. Additionally, patients 3 and 4 share a topic with MLL-AF4 cell lines in the $k = 10$ analysis but not in the $k = 15$ version.

One of the key questions that remains to be answered is whether the unique regulatory regions identified in the previous section are still enriched in the patient topics. We took the top 100 annotated regions from topic 6 in the $k = 15$ case and found that the top 100 regions in the ENCODE analysis included 55 of the 76 reproducible patient regions. The trend across all topics for these regions was consistent with this, most of which were predominantly involved in topic 6 (Figure 4.15A). Of the 21 regions not within the top 100 regions for topic 6, their accessibility remained strongly associated with topic 6 (Figure 4.15B). However, three other topics were more represented in this set, namely topics 4, 5, and 12. These topics are involved in a variety of cell types, and interestingly Topic 4 is more active in patients 3 and 4 than in patients 1 and 2, which is the opposite of the trends previously observed. Topic 5, of which at least one of the patient-related regions has a Z-score over 30, is active in CD4 negative natural killer and T-cells, as well as K562 cancer cells. Together, these results imply that the regulatory regions previously identified are a mix of regions uniquely active in MLL-AF4 patients and a minority which may be re-used from other cell types.

4.2.7 Accessibility of patient-related topics within lymphopoiesis

To determine the similarity of patient-related regions to other lymphoid progenitors, we collected a dataset of cells undergoing lymphopoiesis from HSC to activated B cells spanning the dataset from Corces et al. [202], the data that we collected

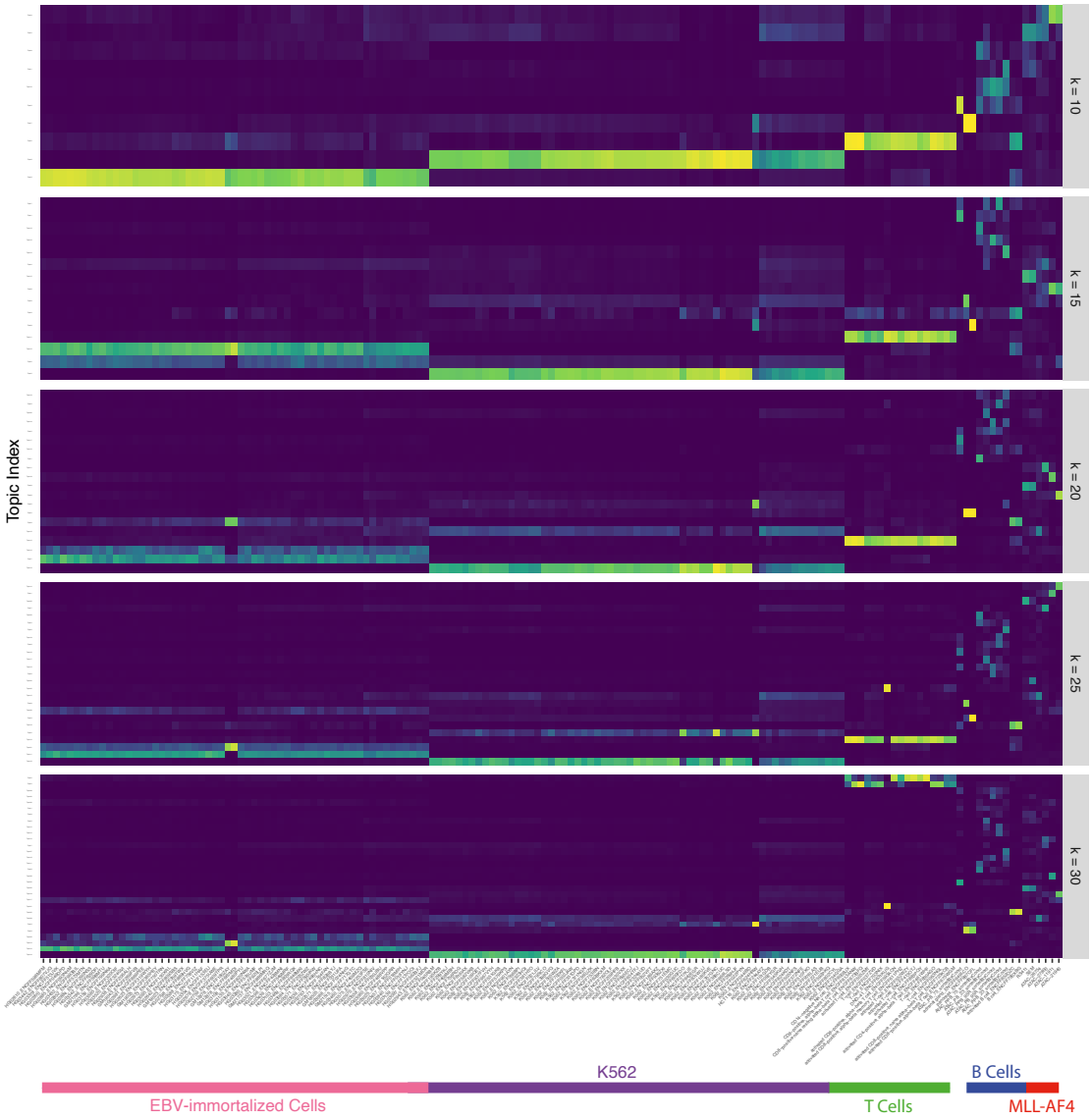


Figure 4.13: Topic modelling for ENCODE blood cell collection with MLL-AF4 and BCP samples for $k = 10, 15, 20, 25, 30$. Lines delineating classes of cells are approximate. See Figure 4.14 for details. EBV = Epstein-Barr Virus.

from patients and BCP as well as two samples from the ENCODE consortium. We find the average normalised read counts under the patient-related regions to be highly enriched for the patient samples, with some small enrichment for BCP and early hematopoietic progenitors. Individual regions such as the putative enhancer cluster described in Figure 4.11 show a similar pattern, with some accessibility in early progenitor cells which is mostly lost by terminal differentiation (Figure 4.17). Some regions appear to be accessible relative to the background in early progenitors,

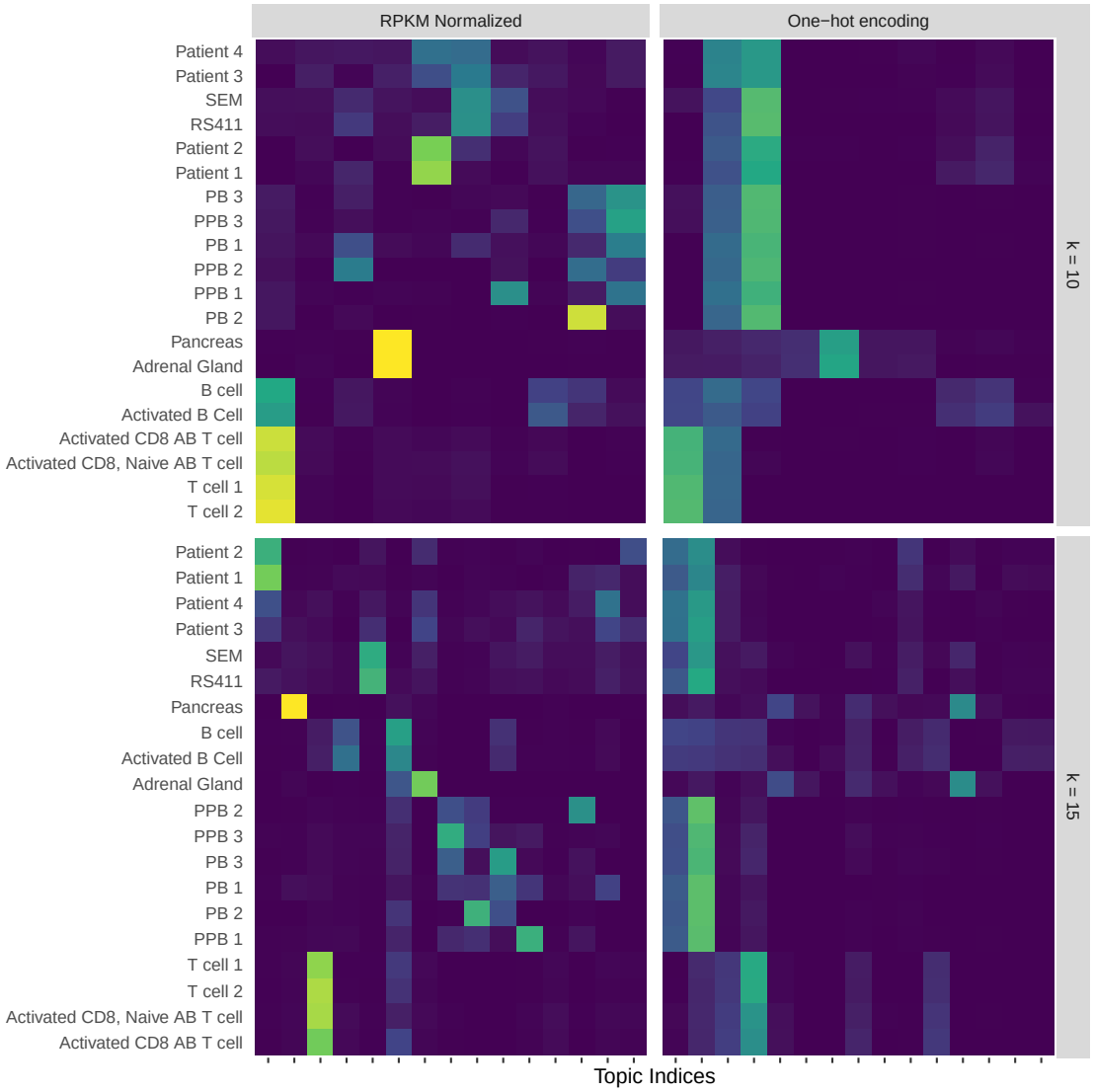


Figure 4.14: Zoomed in plot looking at only the most related cell types to the MLL-AF4 and BCP in the ENCODE blood cell collection for $k = 10$ and $k = 15$ using both one-hot encoding and RPKM normalization.

though the read coverage under the peak is much higher in the patient samples than in any other cell type.

4.3 Discussion

This chapter proposes the use of topic modelling to prioritise accessible regulatory regions in a poorly understood but difficult to treat subsets of leukemias. One of the primary goals of this research was to learn more about CREs in MLL-

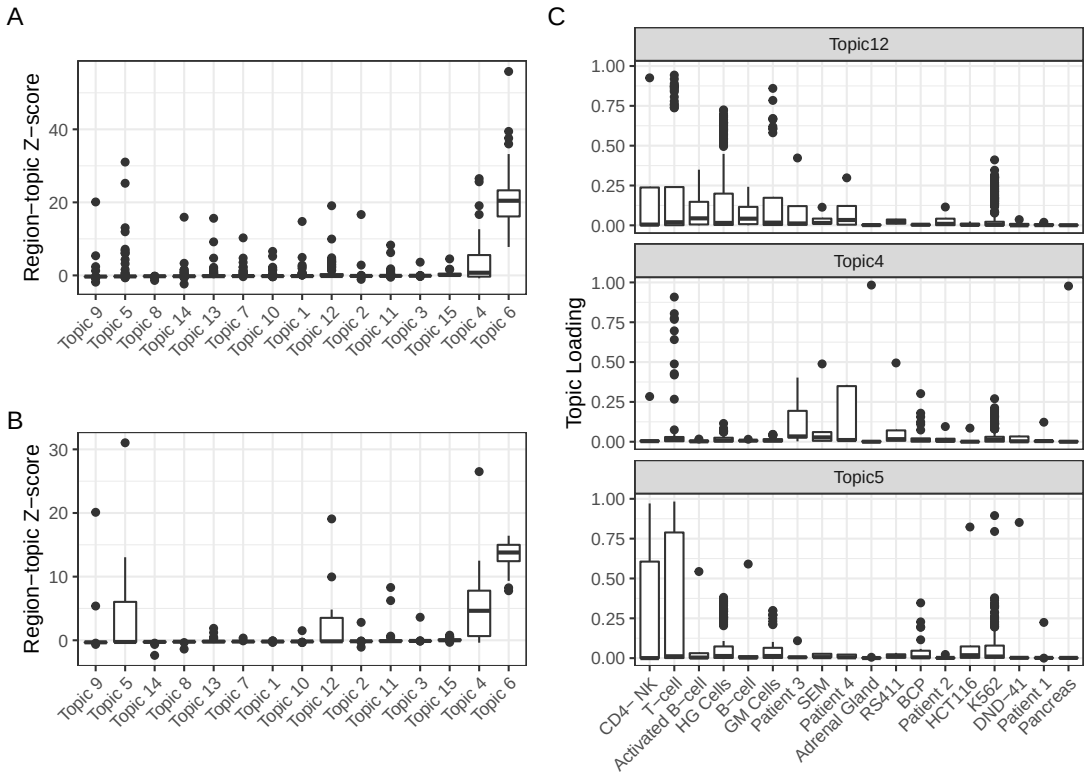


Figure 4.15: Activity of 75 reproducibly identified patient-related regions within the inferred $k = 15$ BLDA analysis in all of the ENCODE blood cell collection.

AF4 leukemias. It is currently unknown whether MLL-AF4 leukemias entirely reuse available regulatory regions in different ways, create novel CREs which are unknown in other cellular texts, or if they co-opt enhancers and promoters from different cellular contexts. Four childhood MLL-AF4 patients were studied alongside two cell models of the same leukemia. To contrast these cancerous cells and model "normal", non-cancerous, developmental biology, we include a newly identified lymphoid committed fetal-specific cell type whose transcriptional profile closely matches known infant ALLs (pre-proB cells). We additionally include a transcriptionally distinct but closely related lymphoid progenitor, ProB (PB) cells, which have known analogues in adults, but are derived from fetal sources. We show that there are distinct accessibility patterns in the four different subsets of cells, each represented by a topic which loads uniquely onto that subset. We show that these topics encode reproducibly identified accessible regions that differentiate

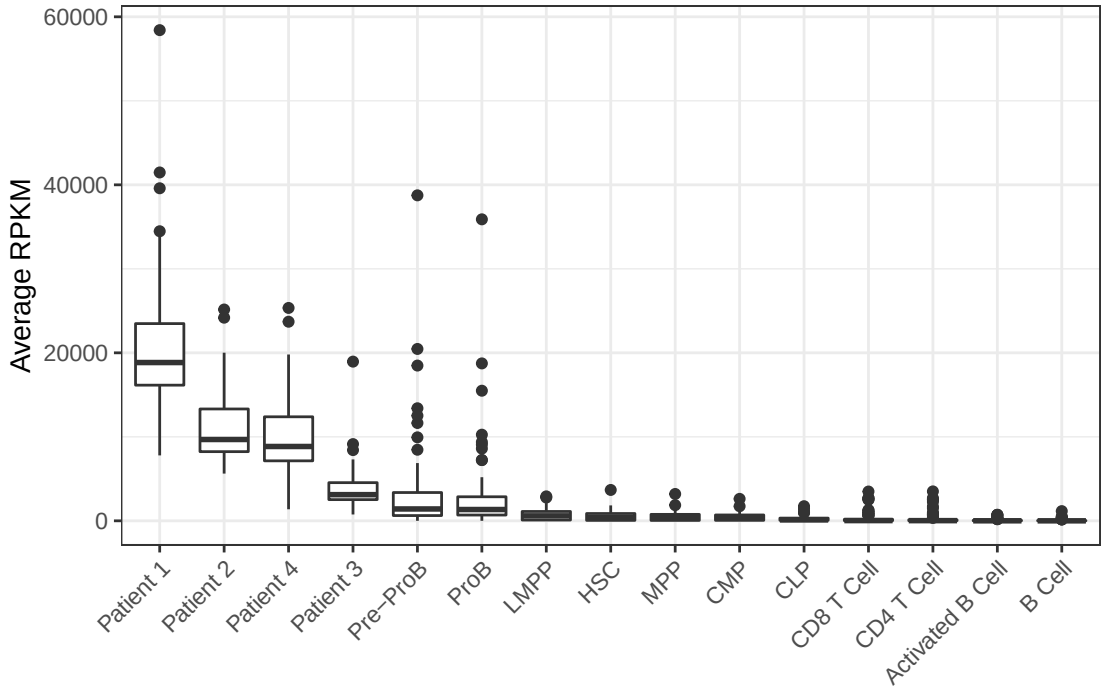


Figure 4.16: Average normalised read count under 76 patient-related regions in a collection of cells from early hematopoiesis to terminal lymphopoiesis. Patient and ProB/PreProB cells are described in this chapter, whereas B cell and Activated B Cell samples are lifted over from GRCh38 in the Encode Consortium with CrossMap. The remainder of the samples are from Corces et al.

between MLL-AF4 cancerous samples and developmentally normal BCP, even at the stringent threshold of 100 individual regions per topic. The alternative topic modelling approach tested against, the one-hot encoding input method used by cisTopic, did not demonstrate any ability to differentiate between cell models and patient samples or PB versus PPB but did broadly differentiate between the cancerous and normal cell types, indicating at least a partial role of regions which are entirely accessible or not in the accessibility program of leukemia samples (Figure 4.4). This bears relevance to the question of differential regulatory element usage versus differential accessibility at the same regulatory element. The fact that both methods are able to distinguish between the samples means that, on a broad level, there is at least some differential regulatory element usage between fetal MLLr leukemia and the fetal PPB cells.

Interestingly, BLDA also identified topics of regulatory elements accessible

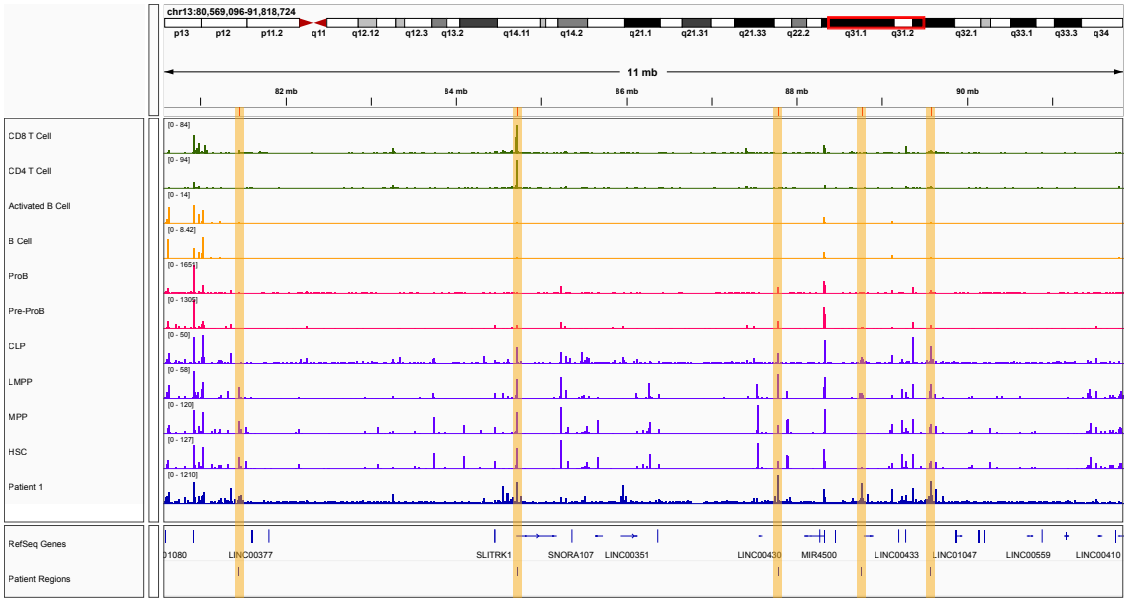


Figure 4.17: Read coverage over 76 patient-related regions in a collection of ATAC-seq experiments from early hematopoiesis to terminal lymphopoiesis. Patient and ProB/PreProB cells are described in this chapter, whereas B cell and Activated B Cell samples are lifted over from GRCh38 in the Encode Consortium with CrossMap. The remainder of the samples are from Corces et al. Track specific normalisation values are displayed on the left-hand side of each track, where patient 1 the maximum value represents 1210 RPKM, HSCs represent 127, MPPs 120, etc.

in Pre-ProB cells not shared with ProB cells, helping to illuminate fetal-specific lymphopoiesis trajectories (Figure 4.6, Figure 4.8A). A large number of these regions (102 for PPB and 158 for PB) were consistently identified across every replicate in the top 500 regions. These regions will be the focus of more mechanistic studies trying to understand chromatin remodelling in fetal lymphopoiesis and how it differs from the same differentiation process in adults. Understanding this process in detail is crucial to the study of fetal leukemiagenesis. These results will be particularly applicable to the study of leukemia stem cells (LSCs) and how they are transformed from developmentally normal cells; is it unclear whether the pathways involved in self-renewal and maintenance of LSCs may be shared with PPB cells. In this case the topic modelling approach allows for a fine-grained dissection of known pathways and their relative contributions to different cell types. However, the aim for this chapter was explicitly to examine the regulatory regions involved in patients. The remainder of the fine-grained results will be examined in depth as time allows.

BLDA identified a topic in each replicate that was uniquely enriched in the patient samples, most strongly in patients 1 and 2. For values of k larger than 10, this topic starts to split in two, with one represented strongly in patients 1 and 2 as before but the second represented more strongly in patients 3 and 4 (Figure 4.4). Values of k this large, however, seem extremely unreliable in practice, and it is unlikely that regions would be consistently prioritized after replication. This is demonstrated by the extremely dispersed nature of the topics in BCP samples and the lack of a central topic explaining core regulatory features shared by the entire group, which we expect to see in this case. The reasons for the consistent lack of enrichment for this topic in patients 3 and 4 are difficult to explain, but may revolve around either patient heterogeneity or sample quality. The first is an area of active research, with well known results showing very little consensus between patient derived models for different MLL-FPs even in terms of distinct chromatin binding [282]. These previous results urge caution when interpreting observations within samples deriving from a single fusion partner. The heterogeneity within the patient population with regards to chromatin accessibility and binding efficiency of the MLL-AF4 fusion protein as well as important downstream chromatin remodellers like RUNX1 is not well studied. However, in the case that substantial heterogeneity exists, and if it is important for treatment efficacy and differential prognosis, topic modelling appears to be a promising approach for dissecting these differences. The alternative explanation is differences in patient sample quality, which is difficult to describe quantitatively in this case. Though the sequencing appears to have reasonable coverage under each of the peak regions, known difficulties with preparing and isolating the necessary cells from the patients may have prevented a completely unbiased view into their accessible chromatin. The third possibility for these differences may be differential genome instability and mutational background within the patients. Though well established that MLLr cancers tend to not carry with them many functional cooperating mutations, mutations in key pathways such as FLT3 (frequently up-regulated in MLLr leukemias) have been described in pediatric and adult AML leading to extremely poor prognosis [283–285]. The genetic background of these

samples is unknown, and further work is necessary to understand the contribution of genetic differences to the epigenetic profile of MLL-AF4 driven leukemias.

Patient heterogeneity makes differential accessibility analysis such as the one we attempted to perform very challenging. The genes identified with the edgeR analysis bear no strong relationship to known MLLr pathways, though some such as RhoU and SERF1A from Table 4.2 may have suggestive links to ALL [286–288]. The fundamental strength with the topic modelling approach is the ability to look at variation in accessibility both collectively, between the cell lines and each individual patient, as differentially. Few regions, it seems, are uniformly different in accessibility between the different categories. This is reinforced by the inferred topic loadings, which separate well on the different categories and suggest that fundamentally different regulatory biology underlies the cell lines versus the patients, and that substantial heterogeneity exists even within each of those subclasses. Within the context of this sample, understanding the role of these regions is additionally confounded by the fact that they are derived from blastic, malignant cancer cells. Practically, this means that they may contain somatic mutations not represented in the reference panel. At present, there is no reliable, published, method for calling mutations from peak-based NGS experiments. Even if this data were available, it would be difficult to differentiate between accessibility as a consequence of cellular environment and differential evolution from that caused by genetic differences. In the future, this may be improved with the incorporation of more patient samples with similar demographics and disease profiles.

The regions picked out with the topic modelling approach appear to be mostly functional in this cellular context. Many of them have been previously annotated as enhancer or promoter elements in distantly related cell types, indicating their functional potential (Table 4.4). Additionally, the profile of histone modifications and bound transcription factors within one of the specific patients indicates a large over-abundance of MLL binding alongside RUNX1 and PAF1c even when compared with other regions enriched for H3K4me1 and H3K27ac (Table 4.3, Figure 4.10). Interestingly, these regions are not enriched for AF4 binding, indicating that they

may in fact be bound by wild-type MLL, though to different levels than the comparable BCP. The alternate possibility is that AF4, being a difficult mark to pull down with antibodies, is being found less frequently due to technical bias. To our knowledge, the former possibility has not been previously observed, and raises the possibility that the MLL-FP may alter the binding profile of wild-type MLL within the genome. These regions are depleted for H3K79me2 marks, which is unexpected given that MLL typically controls the methylation of K79 through the recruitment of DOT1L. Recent work has shown that H3K79 methylation conferred by DOT1L is essential for a subset of enhancer elements to form promoter interactions in MLL-AF4 leukemia, and that inhibition of DOT1L destroys these essential connections [289]. As the only known H3K79 methyltransferase, we can be confident that the function of regions is not dependent on DOT1L recruitment, and therefore representing a completely separate class of MLL-bound but DOT1L-depleted enhancer elements that have, to our knowledge, not been previously described [290]. In addition, the majority of these regions of the genome are still specifically associated with the MLL-AF4 patients against the backdrop of a large set of blood cells from the ENCODE project, and they remain the most accessible cell types for all but two of the regions. This is suggestive evidence that these regions may be specifically active in these patients, at least within the hematopoietic niche. Within the context of lymphopoiesis, a portion of these regions appear to be enriched above the background, and may be functional within certain hematopoietic precursors (Figure 4.17). However, the actual reads under each region are predominantly found in patients, indicating that even if these regions are functional in hematopoetic precursors, they may be more accessible and active in the patients. Further work will investigate these regions in the context of the entire ENCODE accessibility dataset. Though the initial results do not suggest that this is the case, as the set of cell types that these regions are previously annotated in are extremely diverse, evidence for a shared or co-opted regulatory program from a distantly related cell type would be extremely exciting.

The next steps to validate these regions and their roles in the regulatory biology of MLL-AF4 will be necessarily experimental. In order to understand which, if any, genes are being regulated by these regions, Capture-C assays will need to be performed within these specific samples. Motif enrichment within these regions, and indeed within most of the annotated sets of regions, has not been particularly illuminating as to their pathways. To understand their role in a DOT1L-independent regulatory program, their enhancer-promoter interactions should be assessed with and without the use of inhibitors such as EPZ-5676. Though the possibility that these regulatory regions work completely independent of this key co-factor is remote, the implications for current research into therapeutic DOT1L suppression make them key candidates to understand the resistance mechanisms employed by cancer cells. The other crucial and specific aspect of these enhancers is their association with PAF1c, more typically understood as an elongation factor associated with RNA polymerase 2 [291–294]. Very recent evidence from mouse models indicates that PAF1c occupancy may be associated with super-enhancers and is crucial for maintaining the self-renewal capacity of mouse embryonic stem-cells [295]. Additionally, it is known that PAF1c is involved in the regulation of key genes involved in acute myeloid leukemia (AML) such as HOXA9 and MEIS1, so its coordinated activity across these enhancers is of interest to the central regulatory axis of leukemia.

In summary, this chapter has adopted a topic modelling approach that we previously developed to model chromatin accessibility patterns in MLL-AF4 patients when compared to healthy fetal-specific lymphoid precursors. We identified a subset of regions bound by wildtype MLL but not involving DOT1L recruitment. These regions were uniquely accessible in these patients against the background of the hematopoietic niche, but there is evidence that a subset function as enhancers and promoters in distantly related cell types. This result demonstrates that that oncogenic regulatory pathways active in MLL-AF4 leukemia may be a mixture of existing pathways from distantly related cell types along with entirely novel enhancers. Further work will focus on characterizing the interactions of these regions with genes and ways in which their leukemia-specific functions may be interrupted.

These regulatory elements represent key targets for investigation in the search for therapeutic targets to a deadly and impactful disease.

4.4 Methods

4.4.1 Sequencing data

Patient samples were acquired from the United Kingdom BioBank as liquid biopsies from four confirmed MLL-AF4 ALL patients.

Two commercial cell lines were used in this analysis. SEM cells, a MLL-AF4 B cell ALL cell line were purchased from DSMZ (`https://www.dsmz.de`). SEM cells were cultured in Iscove's modified Dulbecco's medium (IMDM) with 10% fetal bovine serum (FBS) and 1x GlutaMAX, with cell density maintained between 5x105/mL and 2x106/mL. RS4;11 cells were purchased from ATCC (`https://www.atcc.org`) and cultured in RPMI-1640 with 10% FBS and 1x GlutaMAX, with cell density maintained between 5x105/mL and 1.5x106/mL. Cells were confirmed to be free of mycoplasma.

ATAC-seq was performed on these samples following the protocol laid out in Buenrostro et al. [194]. ChIP-seq was generated by Alastair Smith in the Milne Group.

Three instances of pre-proB and three proB cells were isolated by O'Byrne et al. [270]. We downloaded ATAC-seq data as an RPKM normalised coverage track from the Gene Expression Omnibus (accession number GSE122989).

4.4.2 Preparation of the ENCODE blood cell dataset

Using the online ENCODE consortium web portal, all available data was subsetted into released human ATAC-seq samples which were a part of the "blood" organ system (`https://www.encodeproject.org/matrix/?type=Experiment&status=released&assay_title=ATAC-seq&replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&biosample_ontology.organ_slims=blood`) resulting in 161 cell types. Data was downloaded as normalised coverage tracks and lifted over to hg19 from GRCh38 using CrossMap [296].

4.4.3 Blacklist construction

The ENCODE blacklist was downloaded from `https://github.com/Boyle-Lab/Blacklist/tree/master/lists` as described in Amemiya, Kundaje, and Boyle [274].

To account for additional technical artefacts in MLL-AF4 cell lines, we collected input tracks for ChIP-seq experiments and performed peak calling with MACS2 to identify regions enriched for technical artifacts [119]. To reduce the amount of excluded sequence, we selected a stringent P-value threshold of 1×10^{-6}, resulting in approximately 10 megabases of additional sequence to exclude.

4.4.4 Peak Calling with LanceOTron

Peak calling on coverage data for ATAC-seq and ChIP-seq was performed with LanceOTron v1.0.1 [221]. LanceOTron was installed from Pypi here `https://pypi.org/project/lanceotron/1.0.1/` and used with the flags *-c 0.5* and *–format bed* to select only regions exceeding a peak score of 0.5 and outputting the data as a bed file for further analysis.

4.4.5 Coverage Metrics

Megadepth was used to construct coverage metrics [272]. Peak regions were used as an annotation (flag *-a*) in order to only consider coverage under regions expected to be well represented.

4.4.6 Topic modelling

Count matrices were constructed from paired coverage data and LanceOTron peak calls using the BLDA python package, available at `https://github.com/Chris1221/blda`. For BLDA analyses, the format was given as "bigwig", while OHE matrices were constructed with the "dummy" option. Matrices were used directly for topic modelling with a modified version of cisTopic available at `https://github.com/Chris1221/cisTopic_bulk`. Bayesian hyperparamter optimization is described in Section 3.2.5 and more detail on the BLDA method is given Section 3.2.6.

4.4.7 Differential accessibility analysis

edgeR was used to perform a naive differential accessibility analysis [206]. The count matrix previously constructed for BLDA was imported and samples were grouped into either healthy (PPB/PB) or MLL-AF4 (patients/cell lines). We followed the edgeR user guide, and make our analysis code public at `https://github.com/chris1221/thesis_lda`.

4.4.8 Motif enrichment with motifscan

Motifscan was used to identify enriched motifs [297]. We follow the usage instructions and create a local instance of the latest version of the JASPAR database [298]. After experimentation (data not shown) we select the most stringent selection threshold (P-value $< 1 \times 10^{-6}$) for position weight matrix identification within sequences.

4.4.9 Bootstrapping statistical significance

Bootstrapping was used to construct empirical distributions in several instances throughout the chapter. Firstly, to identify the number of lncRNAs we downloaded the list of annotated transcripts from RefSeq (from `https://hgdownload.cse.ucsc.edu/goldenpath/hg19/database/`) and discarded transcript level annotations, leaving all remaining genes and pseudogenes. We sampled from this list 5000 times with replacement and counted the number of time a selected gene was annotated as a lncRNA. This formed an empirical distribution of annotations, against which we compared the observed count.

Secondly, we used bootstrapping to set our expectation for ChIP-seq peak overlaps in Table 4.3 by drawing n=76 samples without replacement from the total union set of accessible regions (n=61,739) and use bedtools with the *-u* flag to count the number of the samples regions overlapping with a given ChIP-seq peak file [187]. To calculate an empirical P-value from this distribution, we count the number of random samples with intersections meeting or exceeding the observed value and divide by the total number of samples (5000 in this case).

A similar procedure is followed for Figure 4.10, except pre-filtering on regions that strictly overlap with both H3K4me1 and H3K27ac (n=30,165).

4.5 Acknowledgments

Alignment and quality control of the datasets from Corces et al. [202] and Ludwig et al. [200] was performed by Damien Downes. All experimental work, including cell culture and sequencing was performed by Alastair Smith. Alignment of the cell line and patient data ATAC-seq and ChIP-seq were performed by Alastair Smith.

"I've made peace with myself."
"Good for you. That's the hardest war of all to win."
"Didn't say I won. Just stopped fighting."

— Joe Abercrombie, *Best Served Cold*

5 Conclusion

With a year-over-year increase in the amount of publicly available NGS data, there is a need for machine learning methods to interpret biologically meaningful patterns and generate hypotheses for further study. This thesis introduces two new methods, SMCSMC and BLDA. I use these two new methods to analyse existing NGS data and draw novel conclusions about human history and the regulatory landscape of MLL-AF4 leukemia respectively. In this chapter, I discuss the implications of both the methods developed and the results generated. I finish with a statement on the overall applicability of machine learning to sequencing data and what I believe to be fruitful directions for further study.

5.1 Extending the SMC2 particle filter

Inferring the ARG from a sample of individuals has been a longstanding goal in population genetics. Recently, four approaches have been developed that directly tackle this problem using modern methodologies and data. The SMCSMC method, introduced in Henderson et al. [1] with further exposition in this thesis, is unique in its ability to infer time dependent directional migration rates due to its "first-class" treatment of migration as a distinct node in each marginal tree. Though this is certainly an attractive property of the algorithm, its distinct and incredibly

flexible implementation can allow for the inference of essentially any process which admits simulation along the sequence. This includes, for example, background selection. This process of reducing a loci's diversity as a consequence of linkage with selected alleles has the potential to directly impact demographic inference due to the general assumption of selective neutrality. Johri et al. [299] starkly demonstrate this principle in regard to MSMC and *fastsimcoal2* and there is no reason why SMC2 would fare better in these comparisons. With modifications of the CwR which allow for the simulation of purifying selection (e.g. [300]) and unifying programs for performing population genetics simulations such as Adrion et al. [301], inference from a sample of ARGs generated with the particle filter could be expanded to include additional parameters for selection.

Increasing the number of populations in the model would also massively increase its applicability to real world problems. This is possible with the method as it is, though the implementation would be non-trivial. In my opinion, this and other related problems could be solved by incorporating a standard backend solution for simulating and storing ARGs such as tskit (`https://github.com/tskit-dev/tskit`). Along with reducing the computational task of implementing different simulation models on the back ends, this would provide interoperability with the current gold-standard format for inferred ARGs. I have made preliminary efforts in this direction by providing a script to convert a sample of the inferred ARGs to the tskit format (and Speidel et al. [69] have done the same), however issues with compatibility between the encoding of migration nodes in SMC2's trees has limited the usefulness of this approach.

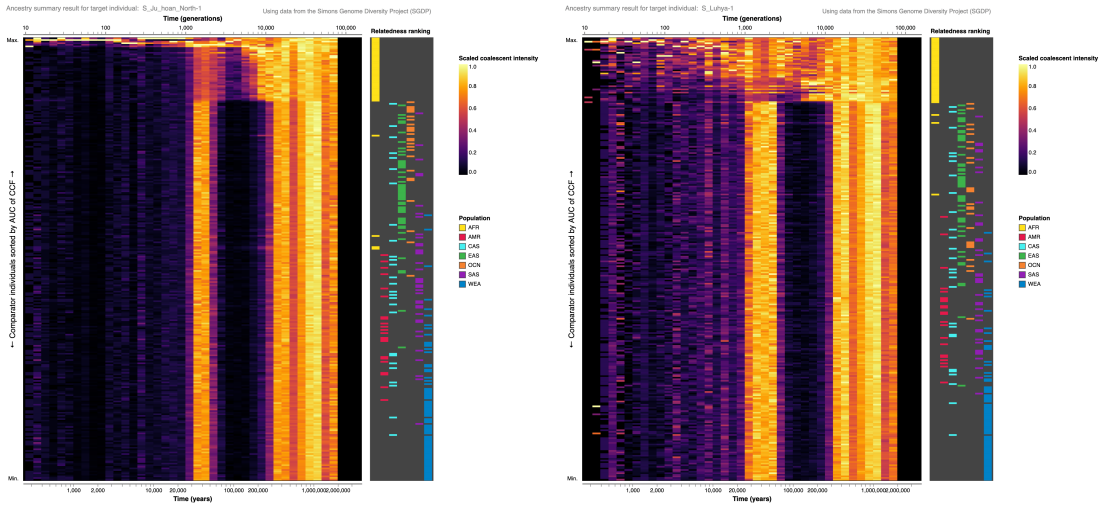
Another beneficial implementation-specific extension would be offloading computationally expensive steps to a graphical processing unit (GPU) implementation of a particle filter. Recent work has shown incredible computational gains in time-sensitive industrial applications such as robotics and manufacturing on the order of 10x runtime improvements simply by switching architectures [302–304]. As SMC2 is an asymptotically exact inference procedure in the limit of compute time, performing 10x the iterations in the same amount of time would improve (i) the resolution of

existing analyses, (ii) the ability to infer from a larger number of haplotypes, (iii) the ability to infer a larger number of populations as stated above, and (iv) the ability to infer more recent events than 10 kya. Though this represents a significant amount of opportunity cost in terms of the time spent implementing the algorithm within the context of GPU libraries, the particle filter methodology is an excellent fit for many problems in population genetics and significantly speeding up the inference procedure would greatly simplify many outstanding problems in the field.

5.2 An ancestral back-migration in the context of African pre-history

In Chapter 2, I applied SMC2 to two databases of global genetic variation and detected a large surplus of directional migration from populations deriving from the Out of Africa (OoA) migration event backwards to those still living in Africa. The observation of this event is not unprecedented, as explored in the Discussion section of that chapter, however the data from single-tree inference in Y chromosome and mtDNA is only able to state lower-bounds on the timing and is completely unable to estimate the proportion of the genome impacted. Since posting the study as a preprint, two independent groups (Montinaro et al. [305] and Wang et al. [306]) have replicated our findings. Replication with independent methodologies lends confidence as to the robustness of the results. The estimates of magnitude from SMC2 are imprecise as demonstrated by simulation, however we expect a large scale back migration to become a commonly modelled feature of African pre-history in future demographic simulations.

One of the most interesting questions arising from this investigation is the identity of the participants in the migration. One suggestion is that the participants may be descended from the unobserved ancestral basal Eurasian branch. These individuals are theorized to have diverged from the OoA population around the same time as Neanderthal admixture was taking place approximately 55 kya before contributing to several Western Eurasian lineages [307, 308]. This would explain the lack of similarity that we see in the isolated segments to Neanderthals, despite



(a) Ju hoan North **(b)** Luhya

Figure 5.1: Inferred coalescent intensity function as calculated in Albers and McVean [174] and exported from the web interface at `https://human.genome.dating/ancestry/` for Ju hoan North and Luhya individuals from the Simons Genome Diversity Panel. Shared coalescent intensity with other African individuals is found at the top of each figure within the yellow bar on the right hand side.

their putatively Eurasian identity, and line up well with the assumed timeline of the migration. The actual dynamics of the migration itself is better phrased as an anthropological or archaeological question, and fine-tuning the resolution of the timing, magnitude, and participants aside, an interdisciplinary approach will be fruitful when investigating this event in the future.

The deep pre-history of Africa is poorly understood. Two populations are theorized to have diverged from the remainder of extant groups in the ancient past, though the actual timing of their divergences is contentious and similar to accepted estimates of the OoA divergence ∼120kya [141]. The degree to which large-scale events such as this migration and other previously unmodelled events such as archaic introgressions (such as that observed by Durvasula and Sankararaman [150]) have contributed to inferred population structures is not well understood. The coalescent intensity function as proposed in Albers and McVean [174], for example, shows similarities between essentially all African groups including the San and CAHGs with regards to their coalescent histories around the time of our identified migration event. It is unlikely that this dip in coalescent intensity before

the migration is due to mass population bottlenecks from super volcano activities, as was previously suspected [309]. Populations identified here as receiving less genetic material from the migration (Khomani San, Ju Hoan, Mbuti, Biaka) show much attenuated coalescent intensity with other African populations between their supposed diversification and this migration. It is not clear how the San and other distantly related populations became recipients of a smaller amount of introgressed material. It is also not clear whether their differential acquisition of OoA genetic material may be sufficient to cause the entirety of African similarities before this period, implying a much larger role for population migrations in human history and explaining the significance of identifying a large portion of the genome as being derived from the migration.

Much more work is needed to understand the place of this migration in African pre-history and to unravel the tangled web of deep relationships within the continent.

5.3 Better discrimination between closely related cell types with topic modelling

Identifying shared groupings of active and accessible regulatory elements between closely related cell types is a difficult problem in functional genomics. It is especially relevant for cancers which have an unknown cell of origin, such as MLL-fusion driven leukemias. In Chapter 3, we adapt the cisTopic LDA model for the case of bulk ATAC-seq data. NGS data is becoming increasingly available, and large panels of cell type specific accessibility variation are available from large consortia such as ENCODE Project Consortium [100]. The approach, which we called BLDA, uses a quantitative value for each peak region rather than a binary activity score as is appropriate for the single cell case. I demonstrate that the approach is superior to a naive implementation of cisTopic in several applications including pseudobulked data and a bespoke erythropoesis dataset. Though these comparisons are reassuring, and lend confidence to the results being generated, they do not constitute adequate benchmarking against gold-standard approaches. Identifying such approaches was difficult, as to our knowledge topic modelling has never

been applied to bulk ATAC-seq data, and the closest class of approaches such as ChromHMM is not appropriate to answer the problems that we set out to address. This is true for two reasons. Firstly that chromHMM infers the most likely state of the genome but does not identify which states are statistically shared or distinct between different cell types. Secondly, because the performance of chromHMM to call distinct state paths in extremely closely related cell types (such as those studied in Chapter 4) is not well understood; extensions such as Marco et al. [310] are evaluated on datasets such as cell lines from the ENCODE project which are known to be extremely diverse. Explicit comparisons between these methods and their ability to identify differentially active regions (most directly assayed through simulation studies) should be performed to establish topic modelling as a viable approach for this class of problem.

Though empirically, our results are consistent with biological expectations, there remain many unanswered questions about the fine-scale details of the implementation. One such question is the choice of normalization method for read data, which has been shown to starkly impact the ability of peak-calling algorithms to identify enriched regions in ATAC-seq data [**Reske2020a**]. We chose RPKM as a respected and reasonable first attempt, however experimentation is needed to determine the optimal method to compare amongst experiments from different centers and different protocols. This leads directly to the unanswered question of the effects of potential latent batch effects, for which I suggest a solution in the following paragraph. The inclusion of many cell types in analyses such as this necessarily increases the complexity of the topic model involved, however another unanswered question is the optimal granularity of a model for a specific purpose; is it better, for instance, to include all available cell types, or to select a few reasonable candidates for comparison? Part of the answer to this question lies in the number of topics that the model is able to reproducibly identify, which is a facet of these investigations not currently appreciated in the text mining literature. I suggest that reproduction of topics across many stochastic inference instantiations may be a reliable way of increasing the signal-to-noise in key-word region identification. However, the

methods for actually selecting the number of topics and hyperparameter values were rudimentary in this thesis, and a more complete solution would involve inferring these values simultaneously with topic loadings.

There have been many innovations in the field of topic modelling in recent years. The performance of BLDA on pseudo-bulked and real data is encouraging, and encourages the use of improved methodologies with the ability to incorporate more data and address the issues raised above. One such class of model is hierarchical LDA, which is a simple extension of the LDA method that treats the number of topics as an unknown variable to be inferred along with topic loadings. The choice of k for the applications in this thesis was highly non-trivial, and an automated procedure guaranteed to select a reasonable k would be extremely beneficial for situations where the similarity structure of data is unknown. Additionally, structured topic models introduce the ability to intelligently learn associations between provided metadata and the probability of a topic occurring in a document. This has natural application to controlling for batch effects between disparate experimental sources as well as introducing prior information on the biological relationship of samples. This structured model would also be an ideal environment to combine ATAC-seq and DNAse-seq experiments while acknowledging and controlling for systematic differences. Another innovation can be drawn from the world of single cell analysis, where the ArchR pipeline chooses to discard entirely the concept of peak regions of accessibility and instead works directly with a windowed view of the genome. This approach has the benefit of removing any ambiguity involved in estimating enriched regions and working directly with the underlying data, though at the cost of significant compute cost. For smaller analyses however, this may lead to more refined estimates of important regions, edge conditions notwithstanding.

The BLDA method represents a preliminary attempt to use topic modelling for a difficult problem in functional genomics. The work in this thesis provides a foundation for further study into the potential for this kind of methodology in determining shared and distinct regulatory regions between similar cell types.

5.4 Novel Enhancers in MLL-AF4 Leukemia

Despite recent progress in treatment options for childhood and infant leukemias, cancers driven by MLL translocations remain mostly incurable and show far worse outcomes than similar cancers without the MLL driver. In Chapter 4, I apply the BLDA method to a collection of ATAC-seq experiments from MLL-AF4 patients and cell lines alongside closely related healthy pre-proB (PPB) and proB (PB) cells. I show highly enriched topics consistently load onto the cancerous cells and that across stochastic replication they consistently contain a subset of associated regions for any threshold. ChIP-seq association shows that these regions are active enhancers in the MLL-AF4 patients, some of which have been previously annotated in distantly related cell types and some of which have not. I explore the relationship to other bound transcription factors such as RUNX1 and PAF1c in the discussion of Chapter 4.

Due to the impacts of COVID, there are many aspects of this project which remain as future work. The first concerns computational analyses which I was unable to complete. These include full-scale analysis of the entirety of the ENCODE dataset and an in depth study of the identified regulatory patterns across cell types. Though this analysis would be of broad interest, it was never performed due to the weeks-long runtimes required. The second concerns experimental validation. In order to pursue these associations, capture-C experiments must identify first which promoters are interacting with these enhancers in these particular cell types. Additionally, if a sufficiently encouraging pathway involving these enhancers can be identified, experimental deletion can identify if their presence is necessary for leukemogenesis. One interesting pathway for experimental validation with clinical applications concerns the possibility that these enhancer elements represent a DOT1L rescue pathway. It has been previously shown that DOT1L inhibition is not sufficient to eliminate leukemia blasts in vitro, despite strong interaction between the fusion protein and the DOT1L protein and a proven subset of enhancers whose promoter interactions are dependent on H3K79me2. The enhancers identified in this thesis appear to be statistically depleted for H3K79me2, indicating that DOT1L is

not necessary for their function. In order to study their use in a putative DOT1L rescue pathway, their accessibility should be assayed before and after treatment with a DOT1L inhibitor such as EPZ-5676. This would demonstrate their utility and motivate further study in possibly novel mechanisms of action in this leukemia. One interesting possibility is that the enhancer regions are being recruited not by the fusion protein but by wild-type MLL. This is suggested by the ChIP-seq associations, which show high enrichment of MLL binding at these sites without any enrichment for the fusion protein C terminus AF4. Though it is possible, and indeed likely, that this difference is due to inefficient antibodies for AF4 pull down, the independent action of the wild type MLL protein on the regulatory landscape of cancer cells represents an interesting possibility for further study, especially as the binding sites of the fusion protein are typically seen to represent a subset of the binding sites of the wild-type protein.

Overall this analysis represents an initial investigation of distinct regulatory elements in MLL-AF4 patients. I used the BLDA method to nominate potential regions and used external data to show that they are functional within this cellular context. These regions represent strong candidates for functional validation. Furthermore, this analysis demonstrates the potential of the BLDA method to identify distinct and interesting regulatory regions in extremely similar cell types.

5.5 Concluding Remarks

This thesis has introduced, developed, and applied new methods for the analysis of next generation sequencing data. Two different problems are approached through the lens of machine learning. The first method, SMC2, represents an algorithmic implementation of an intricately designed statistical model with decades of motivation in population genetics. Despite this, our results demonstrate that large-scale events can be missed by conventional analyses which fail to fully parameterize migration. The second method, BLDA, represents an incremental upgrade on a general method which contains no deep relationship to the biology which it models, yet still yields actionable insights into the pathology of an incurable leukemia. The

diversity of approaches here demonstrated show the potential for machine learning algorithms of all kinds in this growing field.

Appendices

A Demographic Models

Generally, these models can be implemented in either `scrm` or `ms` through they have been written with the former in mind.

A.1 Seed model for SMCSMC Inference

We seed the particle filter with a demographic model of population size and uniform symmetric migration rate, given by the following `scrm` command:

```
-ej 0.2324 2 1 -eM 0 1 -eN 0.0 6 -eN 0.0037 4.4 -eN 0.0046 3 -eN 0.0058 2 -eN 0.0073 1.4
-eN 0.0092 0.85 -eN 0.093 1.2 -eN 0.12 1.7 -eN 0.15 2.2 -eN 0.19 2.5 -eN 0.24 2.4
-eN 0.30 2.0 -eN 0.37 1.7 -eN 0.47 1.4 -eN 0.59 1.2 -eN 0.74 1.0 -eN 0.93 0.91 -eN 1.2 1.6
```

A.2 Migration Simulations

The following models were used for population sizes:

A.2.1 African Population Size

```
-en 0.00000000 1 36.9124479 -en 0.00229999 1 14.8978177 -en 0.00299994 1 7.04453213
-en 0.00391291 1 3.68961222 -en 0.00510371 1 2.06587476 -en 0.00665692 1 1.21617010
-en 0.00868280 1 0.75362392 -en 0.01132521 1 0.49927968 -en 0.01477178 1 0.36258332
-en 0.01926724 1 0.29687253 -en 0.02108190 1 0.28637149 -en 0.02513079 1 0.28071694
-en 0.03277878 1 0.31028768 -en 0.03915210 1 0.36107482 -en 0.04275426 1 0.39815181
-en 0.05576555 1 0.57528787 -en 0.07273654 1 0.88701054 -en 0.09487226 1 1.36014053
-en 0.12374449 1 1.92573639 -en 0.16140334 1 2.36832894 -en 0.21052280 1 2.45284038
-en 0.27459066 1 2.16222564 -en 0.35815613 1 1.71146032 -en 0.46715286 1 1.32388966
-en 0.60932028 1 1.09778746 -en 0.79475315 1 1.04669123 -en 1.03661833 1 1.16969768
-en 1.35208972 1 1.45788656 -en 1.76356769 1 1.80077313 -en 2.30026970 1 1.89942369
```

A.2.2 Eurasian Population Size

```
-en 0.00000000 2 1.14422216 -en 0.00229999 2 1.14422216 -en 0.00299994 2 1.14422216
-en 0.00391291 2 1.14422216 -en 0.00510371 2 1.14422216 -en 0.00665692 2 1.14422216
-en 0.00868280 2 1.14422216 -en 0.01132521 2 1.14422216 -en 0.01477178 2 1.14422216
-en 0.01926724 2 1.14422216 -en 0.02108190 2 1.14422216 -en 0.02513079 2 1.14422216
-en 0.03277878 2 1.14422216 -en 0.03915210 2 1.14422216 -en 0.04275426 2 1.14422216
-en 0.05576555 2 1.14422216 -en 0.07273654 2 1.14422216 -en 0.09487226 2 1.36014053
-en 0.12374449 2 1.92573639 -en 0.16140334 2 2.36832894 -en 0.21052280 2 2.45284038
-en 0.27459066 2 2.16222564 -en 0.35815613 2 1.71146032 -en 0.46715286 2 1.32388966
-en 0.60932028 2 1.09778746 -en 0.79475315 2 1.04669123 -en 1.03661833 2 1.16969768
-en 1.35208972 2 1.45788656 -en 1.76356769 2 1.80077313 -en 2.30026970 2 1.89942369
```

References

[1] Donna Henderson et al. “Demographic inference from multiple whole genomes using a particle filter for continuous Markov jump processes”. In: *PLOS ONE* 16.3 (2021), pp. 1–24. URL: https://doi.org/10.1371/journal.pone.0247647.

[2] Christopher B. Cole et al. “Ancient Admixture into Africa from the ancestors of non-Africans”. In: *bioRxiv* (2020).

[3] Stephan Schiffels and Richard Durbin. “Inferring human population size and separation history from multiple genome sequences”. In: *Nature Genetics* 46.8 (2014), pp. 919–925. arXiv: 005348 [10.1101].

[4] WATSON JD and CRICK FH. “Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid”. In: *Nature* 171.4356 (1953), pp. 737–738. URL: https://pubmed.ncbi.nlm.nih.gov/13054692/.

[5] W. MIN JOU et al. “Nucleotide Sequence of the Gene Coding for the Bacteriophage MS2 Coat Protein”. In: *Nature 1972 237:5350* 237.5350 (1972), pp. 82–88. URL: https://www.nature.com/articles/237082a0.

[6] Shashikant Kulkarni and John Pfeifer. “Clinical genomics”. In: *Clinical Genomics* (Nov. 2014), pp. 1–470.

[7] Jay Shendure et al. “DNA sequencing at 40: past, present and future”. In: *Nature 2017 550:7676* 550.7676 (Oct. 2017), pp. 345–353. URL: https://www.nature.com/articles/nature24286.

[8] Sanger F, Nicklen S, and Coulson AR. “DNA sequencing with chain-terminating inhibitors”. In: *Proceedings of the National Academy of Sciences of the United States of America* 74.12 (1977), pp. 5463–5467. URL: https://pubmed.ncbi.nlm.nih.gov/271968/.

[9] Lin Liu et al. “Comparison of next-generation sequencing systems”. In: *Journal of Biomedicine and Biotechnology* 2012 (2012).

[10] Steven R. Head et al. “Library construction for next-generation sequencing: Overviews and challenges”. In: *https://doi.org/10.2144/000114133* 56.2 (Apr. 2018), pp. 61–77. URL: https://www.future-science.com/doi/abs/10.2144/000114133.

[11] Marine R et al. “Evaluation of a transposase protocol for rapid generation of shotgun high-throughput sequencing libraries from nanogram quantities of DNA”. In: *Applied and environmental microbiology* 77.22 (Nov. 2011), pp. 8071–8079. URL: https://pubmed.ncbi.nlm.nih.gov/21948828/.

[12] Illumina. “An Introduction to Next-Generation Sequencing for Cardiology”. In: (). URL: www.illumina.com/technology/next-generation-sequencing.ilmn.

[13] Barton E. Slatko, Andrew F. Gardner, and Frederick M. Ausubel. "Overview of Next-Generation Sequencing Technologies". In: *Current Protocols in Molecular Biology* 122.1 (Apr. 2018), e59. URL: https://onlinelibrary.wiley.com/doi/full/10.1002/cpmb.59%20https://onlinelibrary.wiley.com/doi/abs/10.1002/cpmb.59%20https://currentprotocols.onlinelibrary.wiley.com/doi/10.1002/cpmb.59.

[14] C Vollmers A Byrne C Cole R Volden. "Realizing the potential of full-length transcriptome sequencing". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 374.1786 (Nov. 2019), p. 20190097.

[15] A Bayega. "Transcript profiling using long-read sequencing technologies". In: *Methods Mol. Biol.* 1783 (2018), pp. 121–147.

[16] Kristoffer Sahlin and Paul Medvedev. "Error correction enables use of Oxford Nanopore technology for reference-free transcriptome analysis". In: *Nature Communications 2021 12:1* 12.1 (Jan. 2021), pp. 1–13. URL: https://www.nature.com/articles/s41467-020-20340-8.

[17] Simon Andrews et al. *FastQC. A quality control tool for high throughput sequence data. Babraham Bioinformatics.* 2015. URL: https://www.bibsonomy.org/bibtex/f230a919c34360709aa298734d63dca3%20https://www.bioinformatics.babraham.ac.uk/projects/fastqc/%7B%5C%%7D0Ahttp://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/ (visited on 10/04/2021).

[18] Marcel Martin. "Cutadapt removes adapter sequences from high-throughput sequencing reads". In: *EMBnet.journal* 17.1 (May 2011), pp. 10–12. URL: https://journal.embnet.org/index.php/embnetjournal/article/view/200/479%20https://journal.embnet.org/index.php/embnetjournal/article/view/200.

[19] Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. "Trimmomatic: a flexible trimmer for Illumina sequence data". In: *Bioinformatics* 30.15 (Aug. 2014), p. 2114. URL: /pmc/articles/PMC4103590/%20/pmc/articles/PMC4103590/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4103590/.

[20] Ben Langmead et al. "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome". In: *Genome Biology 2009 10:3* 10.3 (Mar. 2009), pp. 1–10. URL: https://genomebiology.biomedcentral.com/articles/10.1186/gb-2009-10-3-r25.

[21] Heng Li and Richard Durbin. "Fast and accurate short read alignment with Burrows–Wheeler transform". In: *Bioinformatics* 25.14 (July 2009), p. 1754. URL: /pmc/articles/PMC2705234/%20/pmc/articles/PMC2705234/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2705234/.

[22] Wu D et al. "Large-Scale Whole-Genome Sequencing of Three Diverse Asian Populations in Singapore". In: *Cell* 179.3 (Oct. 2019), 736–749.e15. URL: https://pubmed.ncbi.nlm.nih.gov/31626772/.

[23] Lai Ping Wong et al. "Deep whole-genome sequencing of 100 southeast Asian malays". In: *American Journal of Human Genetics* 92.1 (Jan. 2013), pp. 52–66.

[24] Degang Wu et al. "Genetic Admixture in the Culturally Unique Peranakan Chinese Population in Southeast Asia". In: *Molecular Biology and Evolution* 38.10 (Sept. 2021), pp. 4463–4474.

[25] Jacobs GS et al. "Multiple Deeply Divergent Denisovan Ancestries in Papuans". In: *Cell* 177.4 (May 2019), 1010–1021.e32. URL: https://pubmed.ncbi.nlm.nih.gov/30981557/.

[26] Shaohua Fan et al. "African evolutionary history inferred from whole genome sequence data of 44 indigenous African populations". In: *Genome Biology* 20.1 (Apr. 2019).

[27] Jason A. Hodgson and Todd R. Disotell. "Anthropological Genetics: Inferring the History of Our Species Through the Analysis of DNA". In: *Evolution: Education and Outreach 2010 3:3* 3.3 (Aug. 2010), pp. 387–398. URL: https://evolution-outreach.biomedcentral.com/articles/10.1007/s12052-010-0262-9.

[28] Richard R. Hudson. "TESTING THE CONSTANT-RATE NEUTRAL ALLELE MODEL WITH PROTEIN SEQUENCE DATA". In: *Evolution* 37.1 (Jan. 1983), pp. 203–217. URL: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1558-5646.1983.tb05528.x%20https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1558-5646.1983.tb05528.x%20https://onlinelibrary.wiley.com/doi/10.1111/j.1558-5646.1983.tb05528.x.

[29] J. F. C. Kingman, Kingman, and J. F. C. "The coalescent". In: *Stochastic Processes and their Applications* 13.3 (1982), pp. 235–248. URL: https://econpapers.repec.org/RePEc:eee:spapps:v:13:y:1982:i:3:p:235-248.

[30] J. F. C. Kingman. "On the genealogy of large populations". In: *Journal of Applied Probability* 19.A (1982), pp. 27–43. URL: https://www.cambridge.org/core/journals/journal-of-applied-probability/article/abs/on-the-genealogy-of-large-populations/539757AA0FCA763216F502567CD01796.

[31] Tajima F. "Evolutionary relationship of DNA sequences in finite populations". In: *Genetics* 105.2 (1983), pp. 437–460. URL: https://pubmed.ncbi.nlm.nih.gov/6628982/.

[32] W. J. Ewens. "Population Genetics Theory - The Past and the Future". In: *Mathematical and Statistical Developments of Evolutionary Theory* (1990), pp. 177–227. URL: https://link.springer.com/chapter/10.1007/978-94-009-0513-9%7B%5C_%7D4.

[33] John Wakeley. *Coalescent Theory: An Introduction*. 1st Editio. W. H. Freeman, 2009.

[34] Robert C. Griffiths and Paul Marjoram. "An Ancestral Recombination Graph". In: 1997.

[35] Martin Kreitman. "Nucleotide polymorphism at the alcohol dehydrogenase locus of Drosophila melanogaster". In: *Nature 1983 304:5925* 304.5925 (1983), pp. 412–417. URL: https://www.nature.com/articles/304412a0.

[36] Pool JE et al. "Population genetic inference from genomic sequence variation". In: *Genome research* 20.3 (Mar. 2010), pp. 291–300. URL: https://pubmed.ncbi.nlm.nih.gov/20067940/.

[37] Jacob Evan Crawford and Brian P Lazzaro. "Assessing the Accuracy and Power of Population Genetic Inference from Low-Pass Next-Generation Sequencing Data". In: *Frontiers in Genetics* 0.APR (2012), p. 66.

[38] Nick Patterson et al. "Ancient Admixture in Human History". In: *Genetics* 192.3 (Nov. 2012), pp. 1065–1093. URL: https://www.genetics.org/content/192/3/1065%20https://www.genetics.org/content/192/3/1065.abstract.

[39] Pontus Skoglund et al. "Reconstructing Prehistoric African Population Structure". In: *Cell* 171.1 (2017), 59–71.e21.

[40] Eric Y. Durand et al. "Testing for ancient admixture between closely related populations". In: *Molecular Biology and Evolution* 28.8 (2011), pp. 2239–2252.

[41] Mark Lipson, David Reich, and Jeffrey P. Townsend. "A working model of the deep relationships of diverse modern human genetic lineages outside of Africa". In: *Molecular Biology and Evolution* 34.4 (2017), pp. 889–902.

[42] Pontus Skoglund et al. "Genetic evidence for two founding populations of the Americas". In: *Nature* 525 (2015), p. 104.

[43] Joseph K. Pickrell et al. "The genetic prehistory of southern Africa". In: *Nature Communications* 3 (2012), pp. 1–6. arXiv: 1207.5552.

[44] Jeffrey D Wall. "Detecting Ancient Admixture in Humans Using Sequence Polymorphism Data". In: (2000).

[45] Rasmus Nielsen et al. *Tracing the peopling of the world through genomics*. 2017.

[46] Shannon L. Carto et al. "Out of Africa and into an ice age: on the role of global climate change in the late Pleistocene migration of early modern humans out of Africa". In: *Journal of Human Evolution* 56.2 (2009), pp. 139–151.

[47] Ashot Margaryan et al. "Population genomics of the Viking world". In: *Nature* 585.7825 (2020), pp. 390–396. URL: http://dx.doi.org/10.1038/s41586-020-2688-8.

[48] Swapan Mallick et al. "The Simons Genome Diversity Project: 300 genomes from 142 diverse populations". In: *Nature* 538.7624 (2016), pp. 201–206.

[49] Kay Prüfer et al. "The complete genome sequence of a Neanderthal from the Altai Mountains". In: *Nature* 505.7481 (2014), pp. 43–49.

[50] Anders Bergström et al. "Insights into human genetic variation and population history from 929 diverse genomes". In: *bioRxiv* (2019), p. 674986.

[51] Adam Auton et al. "A global reference for human genetic variation". In: *Nature 2015 526:7571* 526.7571 (Sept. 2015), pp. 68–74. URL: https://www.nature.com/articles/nature15393.

[52] Julia Höglund et al. "Improved power and precision with whole genome sequencing data in genome-wide association studies of inflammatory biomarkers". In: *Scientific Reports 2019 9:1* 9.1 (Nov. 2019), pp. 1–14. URL: https://www.nature.com/articles/s41598-019-53111-7.

[53] Ananyo Choudhury et al. "Whole-genome sequencing for an enhanced understanding of genetic variation among South Africans". In: *Nature Communications 2017 8:1* 8.1 (Dec. 2017), pp. 1–12. URL: https://www.nature.com/articles/s41467-017-00663-9.

[54] Yongwook Choi et al. "Comparison of phasing strategies for whole human genomes". In: *PLOS Genetics* 14.4 (Apr. 2018), e1007308. URL: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1007308.

[55] Zheng GX et al. "Haplotyping germline and cancer genomes with high-throughput linked-read sequencing". In: *Nature biotechnology* 34.3 (Mar. 2016), pp. 303–311. URL: https://pubmed.ncbi.nlm.nih.gov/26829319/.

[56] Matthias Steinrücken et al. "Inference of complex population histories using whole-genome sequences from multiple populations". In: *Proceedings of the National Academy of Sciences* 116.34 (Aug. 2019), pp. 17115–17120. URL: https://www.pnas.org/content/116/34/17115%20https://www.pnas.org/content/116/34/17115.abstract.

[57] Maanasa Raghavan et al. "Genomic evidence for the Pleistocene and recent population history of Native Americans". In: *Science* 349.6250 (Aug. 2015). URL: https://www.science.org/doi/abs/10.1126/science.aab3884.

[58] Hudson RR. "Properties of a neutral allele model with intragenic recombination". In: *Theoretical population biology* 23.2 (1983), pp. 183–201. URL: https://pubmed.ncbi.nlm.nih.gov/6612631/.

[59] Carsten Wiuf and Jotun Hein. "Recombination as a point process along sequences". In: *Theoretical Population Biology* 55.3 (1999), pp. 248–259. URL: https://pubmed.ncbi.nlm.nih.gov/10366550/.

[60] Gilean A.T McVean and Niall J Cardin. "Approximating the coalescent with recombination". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 360.1459 (2005), p. 1387. URL: /pmc/articles/PMC1569517/%20/pmc/articles/PMC1569517/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1569517/.

[61] Heng Li and Richard Durbin. "Inference of human population history from individual whole-genome sequences". In: *Nature 2011 475:7357* 475.7357 (July 2011), pp. 493–496. URL: https://www.nature.com/articles/nature10231.

[62] Lounès Chikhi et al. "The IICR (inverse instantaneous coalescence rate) as a summary of genomic diversity: insights into demographic inference and model choice". In: *Heredity* 120.1 (2018), pp. 13–24.

[63] Ke Wang et al. "Tracking human population structure through time from whole genome sequences". In: *PLOS Genetics* 16.3 (Mar. 2020), e1008552. URL: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1008552.

[64] Shaohua Fan et al. "African evolutionary history inferred from whole genome sequence data of 44 indigenous African populations". In: *Genome Biology 2019 20:1* 20.1 (Apr. 2019), pp. 1–14. URL: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1679-2.

[65] Stephan Schiffels and Richard Durbin. "Inferring human population size and separation history from multiple genome sequences". In: *Nature Genetics 2014 46:8* 46.8 (June 2014), pp. 919–925. URL: https://www.nature.com/articles/ng.3015.

[66] Matthew D. Rasmussen et al. "Genome-Wide Inference of Ancestral Recombination Graphs". In: *PLOS Genetics* 10.5 (2014), e1004342. URL: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1004342.

[67] Melissa J. Hubisz, Amy L. Williams, and Adam Siepel. "Mapping gene flow between ancient hominins through demography-aware inference of the ancestral recombination graph". In: *PLOS Genetics* 16.8 (Aug. 2020), e1008895. URL: https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1008895.

[68] Li N and Stephens M. "Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data". In: *Genetics* 165.4 (Dec. 2003), pp. 2213–2233. URL: https://pubmed.ncbi.nlm.nih.gov/14704198/.

[69] Leo Speidel et al. "A method for genome-wide genealogy estimation for thousands of samples". In: *Nature Genetics 2019 51:9* 51.9 (Sept. 2019), pp. 1321–1329. URL: https://www.nature.com/articles/s41588-019-0484-x.

[70] Jerome Kelleher et al. "Inferring whole-genome histories in large population datasets". In: *Nature Genetics 2019 51:9* 51.9 (Sept. 2019), pp. 1330–1338. URL: https://www.nature.com/articles/s41588-019-0483-y.

[71] Stephan Schiffels and Richard Durbin. "Inferring human population size and separation history from multiple genome sequences". In: *Nature Genetics* (2014).

[72] M. Sanjeev Arulampalam et al. "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking". In: *IEEE Transactions on Signal Processing* 50.2 (Feb. 2002), pp. 174–188.

[73] Arnaud Doucet, Simon Godsill, and Christophe Andrieu. "On sequential Monte Carlo sampling methods for Bayesian filtering". In: *Statistics and Computing 2000 10:3* 10.3 (2000), pp. 197–208. URL: https://link.springer.com/article/10.1023/A:1008935410038.

[74] N. J. Gordon, D. J. Salmond, and A. F.M. Smith. "Novel approach to nonlinear/non-gaussian Bayesian state estimation". In: *IEE Proceedings, Part F: Radar and Signal Processing* 140.2 (1993), pp. 107–113.

[75] Wang L, Wang S, and Bouchard-Côté A. "An Annealed Sequential Monte Carlo Method for Bayesian Phylogenetics". In: *Systematic biology* 69.1 (Jan. 2020), pp. 155–183. URL: https://pubmed.ncbi.nlm.nih.gov/31173141/.

[76] Fourment M et al. "Effective Online Bayesian Phylogenetics via Sequential Monte Carlo with Guided Proposals". In: *Systematic biology* 67.3 (May 2018), pp. 490–502. URL: https://pubmed.ncbi.nlm.nih.gov/29186587/.

[77] Smith RA, Ionides EL, and King AA. "Infectious Disease Dynamics Inferred from Genetic Data via Sequential Monte Carlo". In: *Molecular biology and evolution* 34.8 (Aug. 2017), pp. 2065–2084. URL: https://pubmed.ncbi.nlm.nih.gov/28402447/.

[78] Simon Taylor et al. "Particle Learning Approach to Bayesian Model Selection: An Application from Neurology". In: (2014), pp. 165–167. URL: https://link.springer.com/chapter/10.1007/978-3-319-02084-6%7B%5C_%7D32.

[79] Paul R. Staab et al. "SCRM: efficiently simulating long sequences using the approximated coalescent with recombination". In: *Bioinformatics* 31.10 (2015), pp. 1680–1682.

[80] Marshall N. Rosenbluth and Arianna W. Rosenbluth. "Monte Carlo Calculation of the Average Extension of Molecular Chains". In: *The Journal of Chemical Physics* 23.2 (Dec. 2004), p. 356. URL: https://aip.scitation.org/doi/abs/10.1063/1.1741967.

[81] Willy Feller. "On the integro-differential equations of purely discontinuous Markoff processes". In: *Transactions of the American Mathematical Society* 48.3 (Mar. 1940), pp. 488–515. URL: https://www.ams.org/tran/1940-048-03/S0002-9947-1940-0002697-3/.

[82] Gilean A.T. McVean and Niall J. Cardin. "Approximating the coalescent with recombination". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* (2005).

[83] Paul Marjoram and Jeff D Wall. "Fast "coalescent" simulation". In: *BMC Genetics* 7.1 (2006), p. 16.

[84] Joseph Felsenstein. "Evolutionary trees from DNA sequences: A maximum likelihood approach". In: *Journal of Molecular Evolution 1981 17:6* 17.6 (Nov. 1981), pp. 368–376. URL: https://link.springer.com/article/10.1007/BF01734359.

[85] Olena Morozova and Marco A. Marra. "Applications of next-generation sequencing technologies in functional genomics". In: *Genomics* 92.5 (Nov. 2008), pp. 255–264.

[86] Rebecca Cullum, Olivia Alder, and Pamela A. Hoodless. "The next generation: Using new sequencing technologies to analyse gene regulation". In: *Respirology* 16.2 (2011), pp. 210–222.

[87] Olena Morozova and Marco A. Marra. "Applications of next-generation sequencing technologies in functional genomics". In: *Genomics* 92.5 (Nov. 2008), pp. 255–264.

[88] Paul J. Hurd and Christopher J. Nelson. "Advantages of next-generation sequencing versus the microarray in epigenetic research". In: *Briefings in Functional Genomics and Proteomics* 8.3 (2009), pp. 174–183.

[89] Werner T. "Next generation sequencing in functional genomics". In: *Briefings in bioinformatics* 11.5 (May 2010), pp. 499–511. URL: https://pubmed.ncbi.nlm.nih.gov/20501549/.

[90] C. David Allis and Thomas Jenuwein. "The molecular hallmarks of epigenetic control". In: *Nature Reviews Genetics 2016 17:8* 17.8 (June 2016), pp. 487–500. URL: https://www.nature.com/articles/nrg.2016.59.

[91] Jason Buenrostro et al. "ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide". In: *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]* 109 (2015), p. 21.29.1. URL: /pmc/articles/PMC4374986/%20/pmc/articles/PMC4374986/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4374986/.

[92] Robert E. Thurman et al. "The accessible chromatin landscape of the human genome". In: *Nature* 489.7414 (Sept. 2012), pp. 75–82.

[93] Oliver Bell et al. "Determinants and dynamics of genome accessibility". In: *Nature Reviews Genetics* 12.8 (Aug. 2011), pp. 554–564.

[94] RD Kornberg. "Chromatin structure: a repeating unit of histones and DNA". In: *Science.* 184.4139 (1974), pp. 868–871.

[95] Oliver Bell et al. "Determinants and dynamics of genome accessibility". In: *Nature Reviews Genetics 2011 12:8* 12.8 (July 2011), pp. 554–564. URL: https://www.nature.com/articles/nrg3017.

[96] David S. Gross and William T. Garrard. "NUCLEASE HYPERSENSITIVE SITES IN CHROMATIN". In: *https://doi.org/10.1146/annurev.bi.57.070188.001111* 57 (Nov. 2003), pp. 159–197. URL: https://www.annualreviews.org/doi/abs/10.1146/annurev.bi.57.070188.001111.

[97] Feng Yan et al. "From reads to insight: a hitchhiker's guide to ATAC-seq data analysis". In: *Genome Biology 2020 21:1* 21.1 (Feb. 2020), pp. 1–16. URL: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-1929-3.

[98] Paul G. Giresi et al. "FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin". In: *Genome Research* 17.6 (June 2007), p. 877. URL: /pmc/articles/PMC1891346/%20/pmc/articles/PMC1891346/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1891346/.

[99] Maria Tsompana and Michael J Buck. "Chromatin accessibility: a window into the genome". In: *Epigenetics & Chromatin 2014 7:1* 7.1 (Nov. 2014), pp. 1–16. URL: https://epigeneticsandchromatin.biomedcentral.com/articles/10.1186/1756-8935-7-33.

[100] The ENCODE Project ENCODE Project Consortium. "An integrated encyclopedia of DNA elements in the human genome." In: *Nature* 489.7414 (Sept. 2012), pp. 57–74. URL: http://www.ncbi.nlm.nih.gov/pubmed/22955616%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3439153.

[101] Jill E. Moore et al. "Expanded encyclopaedias of DNA elements in the human and mouse genomes". In: *Nature 2020 583:7818* 583.7818 (July 2020), pp. 699–710. URL: https://www.nature.com/articles/s41586-020-2493-4.

[102] Peggy J. Farnham. "Insights from genomic profiling of transcription factors". In: *Nature Reviews Genetics 2009 10:9* 10.9 (Aug. 2009), pp. 605–616. URL: https://www.nature.com/articles/nrg2636.

[103] Patricia J. Wittkopp and Gizem Kalay. "Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence". In: *Nature Reviews Genetics 2011 13:1* 13.1 (Dec. 2011), pp. 59–69. URL: https://www.nature.com/articles/nrg3095.

[104] François Spitz and Eileen E. M. Furlong. "Transcription factors: from enhancer binding to developmental control". In: *Nature Reviews Genetics 2012 13:9* 13.9 (Aug. 2012), pp. 613–626. URL: https://www.nature.com/articles/nrg3207.

[105] Leah A. Gates, Charles E. Foulds, and Bert W. O'Malley. "Histone Marks in the 'Driver's Seat': Functional Roles in Steering the Transcription Cycle". In: *Trends in Biochemical Sciences* 42.12 (Dec. 2017), pp. 977–989. URL: http://www.cell.com/article/S0968000417301895/fulltext%20http://www.cell.com/article/S0968000417301895/abstract%20https://www.cell.com/trends/biochemical-sciences/abstract/S0968-0004(17)30189-5.

[106] Sunhee Bae and Bluma J. Lesch. "H3K4me1 Distribution Predicts Transcription State and Poising at Promoters". In: *Frontiers in Cell and Developmental Biology* 0 (May 2020), p. 289.

[107] Ali Sharifi-Zarchi et al. "DNA methylation regulates discrimination of enhancers from promoters through a H3K4me1-H3K4me3 seesaw mechanism". In: *BMC Genomics 2017 18:1* 18.1 (Dec. 2017), pp. 1–21. URL: https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-017-4353-7.

[108] Qing-Lan Li et al. "The hyper-activation of transcriptional enhancers in breast cancer". In: *Clinical Epigenetics 2019 11:1* 11.1 (Mar. 2019), pp. 1–17. URL: https://clinicalepigeneticsjournal.biomedcentral.com/articles/10.1186/s13148-019-0645-x.

[109] Steger DJ et al. "DOT1L/KMT4 recruitment and H3K79 methylation are ubiquitously coupled with gene transcription in mammalian cells". In: *Molecular and cellular biology* 28.8 (Apr. 2008), pp. 2825–2839. URL: https://pubmed.ncbi.nlm.nih.gov/18285465/.

[110] Feng Q et al. "Methylation of H3-lysine 79 is mediated by a new family of HMTases without a SET domain". In: *Current biology : CB* 12.12 (June 2002), pp. 1052–1058. URL: https://pubmed.ncbi.nlm.nih.gov/12123582/.

[111] Laura Godfrey et al. "DOT1L inhibition reveals a distinct subset of enhancers dependent on H3K79 methylation". In: *Nature Communications 2019 10:1* 10.1 (June 2019), pp. 1–15. URL: https://www.nature.com/articles/s41467-019-10844-3.

[112] Yichao Cai et al. "H3K27me3-rich genomic regions can function as silencers to repress gene expression via chromatin interactions". In: *Nature Communications 2021 12:1* 12.1 (Jan. 2021), pp. 1–22. URL: https://www.nature.com/articles/s41467-021-20940-y.

[113] Peter J. Park. "ChIP-seq: Advantages and challenges of a maturing technology". In: *Nature Reviews Genetics* 10.10 (Oct. 2009), pp. 669–680.

[114] Ryuichiro Nakato and Toyonori Sakata. "Methods for ChIP-seq analysis: A practical workflow and advanced applications". In: *Methods* 187 (Mar. 2021), pp. 44–53.

[115] Terrence S. Furey. "ChIP-seq and beyond: New and improved methodologies to detect and characterize protein-DNA interactions". In: *Nature Reviews Genetics* 13.12 (Dec. 2012), pp. 840–852.

[116] Thomas R et al. "Features that define the best ChIP-seq peak calling algorithms". In: *Briefings in bioinformatics* 18.3 (May 2017), pp. 441–450. URL: https://pubmed.ncbi.nlm.nih.gov/27169896/.

[117] Lance D. Hentges et al. "LanceOtron: a deep learning peak caller for ATAC-seq, ChIP-seq, and DNase-seq". In: *bioRxiv* (Aug. 2021), p. 2021.01.25.428108. URL: https://www.biorxiv.org/content/10.1101/2021.01.25.428108v3%20https://www.biorxiv.org/content/10.1101/2021.01.25.428108v3.abstract.

[118] Yong Zhang et al. "Model-based analysis of ChIP-Seq (MACS)". In: *Genome Biology* 9.9 (2008).

[119] John M. Gaspar. "Improved peak-calling with MACS2". In: *bioRxiv* (2018).

[120] Heng-Tze Cheng et al. "Wide & Deep Learning for Recommender Systems". In: (June 2016). arXiv: 1606.07792. URL: https://arxiv.org/abs/1606.07792v1.

[121] Jason Ernst and Manolis Kellis. "ChromHMM: automating chromatin-state discovery and characterization". In: *Nature Methods 2012 9:3* 9.3 (Feb. 2012), pp. 215–216. URL: https://www.nature.com/articles/nmeth.1906.

[122] Jason Ernst and Manolis Kellis. "Chromatin-state discovery and genome annotation with ChromHMM". In: *Nature Protocols 2017 12:12* 12.12 (Nov. 2017), pp. 2478–2492. URL: https://www.nature.com/articles/nprot.2017.124.

[123] Carmen Bravo González-Blas et al. "cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data". In: *Nature Methods* (2019).

[124] Toomas Kivisild. *Maternal ancestry and population history from whole mitochondrial genomes*. 2015.

[125] Daniel Rubinoff and Brenden S. Holland. "Between two extremes: mitochondrial DNA is neither the panacea nor the nemesis of phylogenetic and taxonomic inference." In: *Systematic biology* 54.6 (2005).

[126] Kent E. Holsinger and Bruce S. Weir. *Genetics in geographically structured populations: Defining, estimating and interpreting FST*. 2009.

[127] Nick Patterson et al. "Ancient Admixture in Human History". In: 192.November (2012), pp. 1065–1093.

[128] Martin Petr, Benjamin Vernot, and Janet Kelso. "admixr — R package for reproducible analyses using ADMIXTOOLS". In: *Bioinformatics* January (2019), pp. 1–2.

[129] Matthew D. Rasmussen et al. "Genome-Wide Inference of Ancestral Recombination Graphs". In: *PLoS Genetics* (2014).

[130] Leo Speidel et al. "A method for genome-wide genealogy estimation for thousands of samples". In: *Nature Genetics* 51.9 (2019), pp. 1321–1329.

[131] Jerome Kelleher et al. "Inferring whole-genome histories in large population datasets". In: *Nature Genetics* 51.9 (2019), pp. 1330–1338.

[132] Saioa López, Lucy van Dorp, and Garrett Hellenthal. "Human Dispersal Out of Africa: A Lasting Debate." In: *Evolutionary bioinformatics online* 11.Suppl 2 (2015), pp. 57–68.

[133] Frank Schaebitz et al. "Hydroclimate changes in eastern Africa over the past 200,000 years may have influenced early human dispersal". In: *Communications Earth & Environment* 2.1 (2021), pp. 1–10. URL: http://dx.doi.org/10.1038/s43247-021-00195-7.

[134] Axel Timmermann and Tobias Friedrich. "Late Pleistocene climate drivers of early human migration". In: *Nature* 538.7623 (2016), pp. 92–95.

[135] Sriram Sankararaman et al. "The Date of Interbreeding between Neandertals and Modern Humans". In: *PLoS Genetics* (2012).

[136] Qiaomei Fu et al. "Genome sequence of a 45,000-year-old modern human from western Siberia". In: *Nature* 514.7253 (2014), pp. 445–449.

[137] Iosif Lazaridis et al. "Ancient human genomes suggest three ancestral populations for present-day Europeans". In: *Nature* 513.7518 (2014), pp. 409–413. arXiv: 1312.6639. URL: http://dx.doi.org/10.1038/nature13673.

[138] George Bj Busby et al. "Admixture into and within sub-Saharan Africa". In: *eLife* (2016).

[139] Etienne Patin et al. "Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America". In: *Science* 356.6337 (2017), pp. 543–546.

[140] Pontus Skoglund and Iain Mathieson. "Ancient Genomics of Modern Humans: The First Decade". In: *Annual Review of Genomics and Human Genetics* 19.1 (2018), pp. 381–404.

[141] Mark Lipson et al. "Ancient West African foragers in the context of African population history". In: November 2018 (2019).

[142] M. Gallego Llorente et al. "Ancient Ethiopian genome reveals extensive Eurasian admixture throughout the African continent". In: *Science* 350.6262 (2015), pp. 820–822. arXiv: arXiv:1011.1669v3.

[143] Carina M. Schlebusch et al. "Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago". In: *Science* (2017).

[144] Chris Clarkson et al. "Human occupation of northern Australia by 65,000 years ago". In: *Nature* 547.7663 (2017), pp. 306–310.

[145] Wu Liu et al. "The earliest unequivocally modern humans in southern China". In: *Nature* 526.7575 (2015), pp. 696–699.

[146] K E Westaway et al. "An early modern human presence in Sumatra 73,000-63,000 years ago." In: *Nature* 548.7667 (2017), pp. 322–325.

[147] Anna Sapfo Malaspinas et al. "A genomic history of Aboriginal Australia". In: *Nature* 538.7624 (2016), pp. 207–214. arXiv: NIHMS150003.

[148] Luca Pagani et al. "Genomic analyses inform on migration events during the peopling of Eurasia". In: *Nature* 538.7624 (2016), pp. 238–242.

[149] Morten Rasmussen et al. "An aboriginal Australian genome reveals separate human dispersals into Asia". In: *Science* (2011).

[150] Arun Durvasula and Sriram Sankararaman. "Recovering signals of ghost archaic introgression in African populations". In: *bioRxiv* (2019), p. 285734.

[151] M. F. Hammer et al. "Genetic evidence for archaic admixture in Africa". In: *Proceedings of the National Academy of Sciences* 108.37 (2011), pp. 15123–15128.

[152] Vincent Plagnol and Jeffrey D Wall. "Possible Ancestral Structure in Human Populations". In: 2.7 (2006).

[153] Aaron P. Ragsdale and Simon Gravel. "Models of archaic admixture and recent history from two-locus statistics". In: *PLoS genetics* (2019).

[154] Donna Henderson, Sha (Joe) Zhu, and Gerton Lunter. "Demographic inference using particle filters for continuous Markov jump processes". In: *bioRxiv* (2018), p. 382218.

[155] Heng Li and Richard Durbin. "Inference of human population history from individual whole-genome sequences". In: *Nature* 475.7357 (2011), pp. 493–496. arXiv: `arXiv:1011.1669v3`.

[156] Ying Zhou et al. "POPdemog: visualizing population demographic history from simulation scripts". In: *Bioinformatics (Oxford, England)* (2018).

[157] Ke Wang et al. "Tracking human population structure through time from whole genome sequences". In: *bioRxiv* (2019), pp. 1–21.

[158] Jack N. Fenner. "Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies". In: *American Journal of Physical Anthropology* 128.2 (2005), pp. 415–423.

[159] Mason Liang and Rasmus Nielsen. "The Lengths of Admixture Tracts". In: *Genetics* 197.3 (2014), pp. 953–967.

[160] Fernando Racimo et al. "Evidence for archaic adaptive introgression in humans". In: *Nature Reviews Genetics* 16 (2015), p. 359.

[161] Beth L Dumont and Bret A Payseur. "Evolution of the genomic rate of recombination in mammals". In: *Evolution* 62.2 (2008), pp. 276–294.

[162] Johannes Köster and Sven Rahmann. "Snakemake-a scalable bioinformatics workflow engine". In: *Bioinformatics* (2012).

[163] Aylwyn Scally and Richard Durbin. "Revising the human mutation rate: Implications for understand ing human evolution". In: *Nature Reviews Genetics* 13.10 (2012), pp. 745–753.

[164] Stephan Schiffels and Richard Durbin. "Inferring human population size and separation history from multiple genome sequences". In: *Nature Genetics* 46.8 (2014), pp. 919–925. arXiv: `005348 [10.1101]`.

[165] Shaohua Fan et al. "African evolutionary history inferred from whole genome sequence data of 44 indigenous African populations". In: *Genome Biology* (2019).

[166] Luca Pagani et al. "Tracing the Route of Modern Humans out of Africa by Using 225 Human Genome Sequences from Ethiopians and Egyptians". In: *American Journal of Human Genetics* 96.6 (2015), pp. 986–991.

[167] Maanasa Raghavan et al. "Genomic evidence for the Pleistocene and recent population history of Native Americans". In: *Science* 349.6250 (2015).

[168] Iain Mathieson and Gil McVean. "Demography and the Age of Rare Variants". In: *PLoS Genetics* 10.8 (2014). arXiv: 1401.4181.

[169] Lu Chen et al. "Identifying and Interpreting Apparent Neanderthal Ancestry in African Individuals Article Identifying and Interpreting Apparent Neanderthal Ancestry in African Individuals". In: *Cell* (2020), pp. 1–11.

[170] Sriram Sankararaman et al. "The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans". In: *Current Biology* 26.9 (2016), pp. 1241–1247. URL: http://dx.doi.org/10.1016/j.cub.2016.03.037.

[171] Carina M. Schlebusch et al. "Genomic Variation in Seven Khoe-San". In: 1187.October (2012), pp. 374–379.

[172] Chiara Batini et al. "Insights into the demographic history of African pygmies from complete mitochondrial genomes". In: *Molecular Biology and Evolution* 28.2 (2011), pp. 1099–1110.

[173] Etienne Patin and Lluis Quintana-Murci. "The demographic and adaptive history of central African hunter-gatherers and farmers". In: *Current Opinion in Genetics and Development* 53.August (2018), pp. 90–97.

[174] Patrick K. Albers and Gil McVean. "Dating genomic variants and shared ancestry in population-scale sequencing data". In: *bioRxiv* (2019), p. 416610.

[175] M Gallego Llorente and A Manica. "Ancient Ethiopian genome reveals extensive Eurasian admixture in Eastern Africa". In: *Science* 350.October (2015), pp. 820–825.

[176] T K Altheide and M F Hammer. "Evidence for a possible Asian origin of YAP + Y chromosomes." In: *American journal of human genetics* 61.2 (1997), pp. 462–6.

[177] M. F. Hammer et al. "Out of Africa and back again: nested cladistic analysis of human Y chromosome variation". In: *Molecular Biology and Evolution* 15.4 (1998), pp. 427–441.

[178] Fulvio Cruciani et al. "A Back Migration from Asia to Sub-Saharan Africa Is Supported by High-Resolution Analysis of Human Y-Chromosome Haplotypes". In: *The American Journal of Human Genetics* 70.5 (2002), pp. 1197–1214.

[179] A. Chandrasekar et al. "YAP insertion signature in South Asia". In: *Annals of Human Biology* 34.5 (2007), pp. 582–586.

[180] Vicente M. Cabrera et al. "Carriers of mitochondrial DNA macrohaplogroup L3 basal lineages migrated back to Africa from Asia around 70,000 years ago". In: *BMC Evolutionary Biology* 18.1 (2018), p. 98.

[181] M Hervella et al. "The mitogenome of a 35,000-year-old Homo sapiens from Europe supports a Palaeolithic back-migration to Africa". In: *Scientific Reports* 6 (2016), p. 25501.

[182] Marc Haber et al. "A Rare Deep-Rooting D0 African Y-Chromosomal Haplogroup and Its Implications for the Expansion of Modern Humans out of Africa". In: *Genetics* (2019), genetics.302368.2019.

[183] Kenneth S. Zaret and Jason S. Carroll. "Pioneer transcription factors: establishing competence for gene expression". In: *Genes & Development* 25.21 (Nov. 2011), p. 2227. URL: /pmc/articles/PMC3219227/%20/pmc/articles/PMC3219227/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3219227/.

[184] Liesbeth Minnoye et al. "Chromatin accessibility profiling methods". In: *Nature Reviews Methods Primers* 1.1 (2021).

[185] Sandy L Klemm, Zohar Shipony, and William J Greenleaf. "Chromatin accessibility and the regulatory epigenome". In: *Nature Reviews Genetics* (), pp. 29–35. URL: http://dx.doi.org/10.1038/s41576-018-0089-8.

[186] Vincent P. Schulz et al. "A Unique Epigenomic Landscape Defines Human Erythropoiesis". In: *Cell Reports* 28.11 (2019).

[187] Aaron R. Quinlan and Ira M. Hall. "BEDTools: a flexible suite of utilities for comparing genomic features". In: *Bioinformatics* 26.6 (Mar. 2010), pp. 841–842. URL: https://academic.oup.com/bioinformatics/article/26/6/841/244688.

[188] King HW and Klose RJ. "The pioneer factor OCT4 requires the chromatin remodeller BRG1 to support gene regulatory element function in mouse embryonic stem cells." In: *Elife* 6 (Mar. 2017). URL: https://europepmc.org/articles/PMC5400504%20https://europepmc.org/article/med/28287392.

[189] Robert E. Thurman et al. "The accessible chromatin landscape of the human genome". In: *Nature* 489.7414 (2012).

[190] Eliezer Calo and Joanna Wysocka. "Modification of enhancer chromatin: what, how and why?" In: *Molecular cell* 49.5 (Mar. 2013), pp. 825–837. URL: /pmc/articles/PMC3857148/%20/pmc/articles/PMC3857148/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3857148/.

[191] Clifford A. Meyer and X. Shirley Liu. *Identifying and mitigating bias in next-generation sequencing methods for chromatin biology*. 2014.

[192] Alan P. Boyle et al. "High-Resolution Mapping and Characterization of Open Chromatin across the Genome". In: *Cell* 132.2 (2008).

[193] M. Ryan Corces et al. "An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues". In: *Nature Methods* 14.10 (2017).

[194] Jason D. Buenrostro et al. "Single-cell chromatin accessibility reveals principles of regulatory variation". In: *Nature* 523.7561 (2015).

[195] Feng Yan et al. *From reads to insight: A hitchhiker's guide to ATAC-seq data analysis*. 2020.

[196] Margaret H. Baron, Joan Isern, and Stuart T. Fraser. "The embryonic origins of erythropoiesis in mammals". In: *Blood* 119.21 (May 2012), pp. 4828–4837. URL: http://ashpublications.org/blood/article-pdf/119/21/4828/1351830/zh802112004828.pdf.

[197] Stuart H. Orkin and Leonard I. Zon. "Hematopoiesis: An Evolving Paradigm for Stem Cell Biology". In: *Cell* 132.4 (Feb. 2008), pp. 631–644. URL: http://www.cell.com/article/S0092867408001256/fulltext%20http://www.cell.com/article/S0092867408001256/abstract%20https://www.cell.com/cell/abstract/S0092-8674(08)00125-6.

[198] Hideyuki Oguro et al. "Poised Lineage Specification in Multipotential Hematopoietic Stem and Progenitor Cells by the Polycomb Protein Bmi1". In: *Cell Stem Cell* 6.3 (Mar. 2010), pp. 279–286. URL: https://pubmed.ncbi.nlm.nih.gov/20207230/.

[199] Deqing Hu and Ali Shilatifard. "Epigenetics of hematopoiesis and hematological malignancies". In: (2016). URL: http://www.genesdev.org/cgi/doi/10.1101/gad.284109..

[200] Leif S. Ludwig et al. "Transcriptional States and Chromatin Accessibility Underlying Human Erythropoiesis". In: *Cell Reports* 27.11 (2019), 3228–3240.e7. URL: https://doi.org/10.1016/j.celrep.2019.05.046.

[201] Angus M. Sinclair. "Erythropoiesis stimulating agents: Approaches to modulate activity". In: *Biologics: Targets and Therapy* 7.1 (2013), pp. 161–174.

[202] M. Ryan Corces et al. "Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution". In: *Nature Genetics* 48.10 (2016).

[203] Matthew E. Ritchie et al. "Limma powers differential expression analyses for RNA-sequencing and microarray studies". In: *Nucleic Acids Research* 43.7 (2015).

[204] Charity W. Law et al. "Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts". In: *Genome Biology* 15.2 (2014).

[205] Michael I. Love, Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome Biology* 15.12 (2014).

[206] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. "edgeR: A Bioconductor package for differential expression analysis of digital gene expression data". In: *Bioinformatics* 26.1 (2009).

[207] G Stark,
R and Brown. "DiffBind: differential binding analysis of ChIP-Seq peak datatle". In: *http://bioconductor.org/packages/release/bioc/vignettes/DiffBind/inst/doc/DiffBind.pdf.* (2016).

[208] Aaron T.L. Lun and Gordon K. Smyth. "Csaw: A Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows". In: *Nucleic Acids Research* 44.5 (2015).

[209] Jake J. Reske, Mike R. Wilson, and Ronald L. Chandler. "ATAC-seq normalization method can significantly affect differential accessibility analysis and interpretation". In: *Epigenetics and Chromatin* 13.1 (2020).

[210] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet allocation". In: *Journal of Machine Learning Research* 3.4-5 (2003).

[211] Margaret E. Roberts, Brandon M. Stewart, and Dustin Tingley. "Stm: An R package for structural topic models". In: *Journal of Statistical Software* 91 (2019).

[212] Thomas S. Ferguson. “A Bayesian Analysis of Some Nonparametric problems”. In: *Annals of Statistics* (1991).

[213] Zhuolin Qiu et al. “Collapsed Gibbs sampling for latent Dirichlet allocation on spark”. In: *Journal of Machine Learning Research* 36 (2014).

[214] Måns Magnusson et al. “Sparse Partially Collapsed MCMC for Parallel Inference in Topic Models”. In: *Journal of Computational and Graphical Statistics* 27.2 (2018).

[215] Hongju Park, Taeyoung Park, and Yung Seop Lee. “Partially collapsed Gibbs sampling for latent Dirichlet allocation”. In: *Expert Systems with Applications* 131 (2019).

[216] Jianfei Chen et al. “WarpLDA: A cache efficient O(1) algorithm for latent dirichlet allocation”. In: *Proceedings of the VLDB Endowment*. Vol. 9. 10. 2016.

[217] Marcel Martin. “Cutadapt removes adapter sequences from high-throughput sequencing reads”. In: *EMBnet.journal* 17.1 (2011).

[218] Simon Andrews. “FastQC”. In: *Babraham Bioinformatics* (2010).

[219] Ben Langmead and Steven Salzberg. “Bowtie2”. In: *Nature methods* 9.4 (2013).

[220] Andreas Heger. *Pysam*. 2009.

[221] Lance D Hentges et al. “LanceOtron: a deep learning peak caller for ATAC-seq, ChIP-seq, and DNase-seq”. In: *bioRxiv* (2021).

[222] Matthew Rocklin. “Dask: Parallel Computation with Blocked algorithms and Task Scheduling”. In: *Proceedings of the 14th Python in Science Conference*. 2015.

[223] C. B. Lozzio and B. B. Lozzio. “Human chronic myelogenous leukemia cell line with positive Philadelphia chromosome”. In: *Blood* 45.3 (1975).

[224] John W. Belmont et al. “A haplotype map of the human genome”. In: *Nature* 437.7063 (2005).

[225] James A. Thomson. “Embryonic stem cell lines derived from human blastocysts”. In: *Science* 282.5391 (1998).

[226] Fidel Ramírez et al. “DeepTools: A flexible platform for exploring deep-sequencing data”. In: *Nucleic Acids Research* 42.W1 (2014).

[227] Lingyun Song et al. “Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity”. In: *Genome Research* 21.10 (2011).

[228] Jason A. West et al. “Nucleosomal occupancy changes locally over key regulatory regions during cell differentiation and reprogramming”. In: *Nature Communications* 5 (2014).

[229] Michelle Erin Miller et al. “Meis1 is required for adult mouse erythropoiesis, megakaryopoiesis and hematopoietic stem cell expansion”. In: *PLoS ONE* 11.3 (2016).

[230] Sabrina Zeddies et al. “MEIS1 regulates early erythroid and megakaryocytic cell fate”. In: *Haematologica* 99.10 (2014).

[231] Zeenath Unnisa et al. “Meis1 preserves hematopoietic stem cells in mice by limiting oxidative stress”. In: *Blood* 120.25 (2012).

[232] Fabiana V. Mello et al. "Maturation-associated gene expression profiles during normal human bone marrow erythropoiesis". In: *Cell Death Discovery* 5.1 (2019). URL: http://dx.doi.org/10.1038/s41420-019-0151-0.

[233] Novalia Pishesha et al. "Transcriptional divergence and conservation of human and mouse erythropoiesis". In: *Proceedings of the National Academy of Sciences of the United States of America* 111.11 (2014).

[234] Dan Chen and Gu Zhang. "Enforced expression of the GATA-3 transcription factor affects cell fate decisions in hematopoiesis". In: *Experimental Hematology* 29.8 (2001).

[235] Katherine I. Oravecz-Wilson et al. "Huntingtin Interacting Protein 1 mutations lead to abnormal hematopoiesis, spinal defects and cataracts". In: *Human Molecular Genetics* 13.8 (2004).

[236] Bridget O'Laughlin-Bunner et al. "Lyn is required for normal stem cell factor-induced proliferation and chemotaxis of primary hematopoietic cells". In: *Blood* 98.2 (2001).

[237] Michelle L. Ratliff et al. "ARID3a expression in human hematopoietic stem cells is associated with distinct gene patterns in aged individuals". In: *Immunity and Ageing* 17.1 (2020).

[238] Hiroto Inaba and Ching-Hon Pui. "Advances in the Diagnosis and Treatment of Pediatric Acute Lymphoblastic Leukemia". In: *Journal of Clinical Medicine 2021, Vol. 10, Page 1926* 10.9 (Apr. 2021), p. 1926. URL: https://www.mdpi.com/2077-0383/10/9/1926/htm%20https://www.mdpi.com/2077-0383/10/9/1926.

[239] Siobhan Rice and Anindita Roy. "MLL-rearranged infant leukaemia: A 'thorn in the side' of a remarkable success story". In: *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* 1863.8 (Aug. 2020), p. 194564.

[240] C Meyer et al. "The MLL recombinome of acute leukemias in 2017". In: *Leukemia 2018 32:2* 32.2 (July 2017), pp. 273–284. URL: https://www.nature.com/articles/leu2017213.

[241] Rajesh C. Rao and Yali Dou. "Hijacked in cancer: the KMT2 (MLL) family of methyltransferases". In: *Nature Reviews Cancer 2015 15:6* 15.6 (May 2015), pp. 334–346. URL: https://www.nature.com/articles/nrc3929.

[242] Douglas Hanahan and Robert A. Weinberg. "Hallmarks of Cancer: The Next Generation". In: *Cell* 144.5 (Mar. 2011), pp. 646–674. URL: http://www.cell.com/article/S0092867411001279/fulltext%20http://www.cell.com/article/S0092867411001279/abstract%20https://www.cell.com/cell/abstract/S0092-8674(11)00127-9.

[243] Ghayas C. Issa et al. "Therapeutic implications of menin inhibition in acute leukemias". In: *Leukemia 2021 35:9* 35.9 (June 2021), pp. 2482–2495. URL: https://www.nature.com/articles/s41375-021-01309-y.

[244] Yuan Fang, Guochao Liao, and Bin Yu. "LSD1/KDM1A inhibitors in clinical trials: advances and prospects". In: *Journal of Hematology & Oncology 2019 12:1* 12.1 (Dec. 2019), pp. 1–14. URL: https://jhoonline.biomedcentral.com/articles/10.1186/s13045-019-0811-9.

[245] Tatiana Shorstova, William D. Foulkes, and Michael Witcher. "Achieving clinical success with BET inhibitors as anti-cancer agents". In: *British Journal of Cancer 2021 124:9* 124.9 (Mar. 2021), pp. 1478–1490. URL: https://www.nature.com/articles/s41416-021-01321-0.

[246] Sameem M Abedin, Craig S Boddy, and Hidayatullah G Munshi. "BET inhibitors in the treatment of hematologic malignancies: current insights and future prospects". In: *OncoTargets and therapy* 9 (Sept. 2016), p. 5943. URL: /pmc/articles/PMC5047722/%20/pmc/articles/PMC5047722/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5047722/.

[247] Yuexin Liu. "Clinical implications of chromatin accessibility in human cancers". In: *Oncotarget* 11.18 (2020).

[248] Eric T.B. Antunes and Katrin Ottersbach. "The MLL/SET family and haematopoiesis". In: *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms* 1863.8 (Aug. 2020).

[249] Patricia Ernst et al. "An Mll-dependent Hox program drives hematopoietic progenitor expansion". In: *Current Biology* 14.22 (Nov. 2004), pp. 2063–2069.

[250] Yu BD et al. "Altered Hox expression and segmental identity in Mll-mutant mice". In: *Nature* 378.6556 (Nov. 1995), pp. 505–508. URL: https://pubmed.ncbi.nlm.nih.gov/7477409/.

[251] Relja Popovic and Nancy J. Zeleznik-Le. "MLL: How complex does it get?" In: *Journal of Cellular Biochemistry* 95.2 (May 2005), pp. 234–242.

[252] Hsieh JJ, Cheng EH, and Korsmeyer SJ. "Taspase1: a threonine aspartase required for cleavage of MLL and proper HOX gene expression". In: *Cell* 115.3 (Oct. 2003), pp. 293–303. URL: https://pubmed.ncbi.nlm.nih.gov/14636557/.

[253] James J.-D. Hsieh et al. "Proteolytic Cleavage of MLL Generates a Complex of N- and C-Terminal Fragments That Confers Protein Stability and Subnuclear Localization". In: *Molecular and Cellular Biology* 23.1 (Jan. 2003), p. 186. URL: /pmc/articles/PMC140678/%20/pmc/articles/PMC140678/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC140678/.

[254] Tomasz Cierpicki et al. "Structure of the MLL CXXC domain–DNA complex and its functional role in MLL-AF9 leukemia". In: *Nature Structural and Molecular Biology* 17.1 (2010), pp. 62–69.

[255] Paul M. Ayton, Everett H. Chen, and Michael L. Cleary. "Binding to Nonmethylated CpG DNA Is Essential for Target Recognition, Transactivation, and Myeloid Transformation by an MLL Oncoprotein". In: *Molecular and Cellular Biology* 24.23 (Dec. 2004), pp. 10470–10478.

[256] Risner LE et al. "Functional specificity of CpG DNA-binding CXXC domains in mixed lineage leukemia". In: *The Journal of biological chemistry* 288.41 (Oct. 2013), pp. 29901–29910. URL: https://pubmed.ncbi.nlm.nih.gov/23990460/.

[257] Y Okada And et al. "hDOT1L links histone methylation to leukemogenesis". In: *Cell* 121.2 (Apr. 2005), pp. 167–178.

[258] KM Bernt and N Zhu and AU Sinha and S Vempati and J Faber and AV Krivtsov. "MLL-rearranged leukemia is dependent on aberrant H3K79 methylation by DOT1L". In: *Cancer Cell* 20.1 (2011), pp. 66–78.

[259] Siobhan Rice et al. "A novel human fetal liver-derived model reveals that MLL-AF4 drives a distinct fetal gene expression program in infant ALL". In: *bioRxiv* (Nov. 2020), p. 2020.11.15.379990. URL: https://www.biorxiv.org/content/10.1101/2020.11.15.379990v2%20https://www.biorxiv.org/content/10.1101/2020.11.15.379990v2.abstract.

[260] Rahul Nahar and Markus Müschen. "Pre-B cell receptor signaling in acute lymphoblastic leukemia". In: *Cell cycle (Georgetown, Tex.)* 8.23 (Dec. 2009), p. 3874. URL: /pmc/articles/PMC4047560/%20/pmc/articles/PMC4047560/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4047560/.

[261] C Dafflon et al. "Complementary activities of DOT1L and Menin inhibitors in MLL-rearranged leukemia". In: *Leukemia 2017 31:6* 31.6 (Nov. 2016), pp. 1269–1277. URL: https://www.nature.com/articles/leu2016327.

[262] Andrei V. Krivtsov and Scott A. Armstrong. "MLL translocations, histone modifications and leukaemia stem-cell development". In: *Nature Reviews Cancer 2007 7:11* 7.11 (Nov. 2007), pp. 823–833. URL: https://www.nature.com/articles/nrc2253.

[263] Sandy L. Klemm, Zohar Shipony, and William J. Greenleaf. "Chromatin accessibility and the regulatory epigenome". In: *Nature Reviews Genetics* 20.April (2019), pp. 29–35. URL: http://dx.doi.org/10.1038/s41576-018-0089-8.

[264] Sunhee Bae and Bluma J. Lesch. "H3K4me1 Distribution Predicts Transcription State and Poising at Promoters". In: *Frontiers in Cell and Developmental Biology* 0 (May 2020), p. 289.

[265] Raphael Margueron, Patrick Trojer, and Danny Reinberg. "The key to development: interpreting the histone code?" In: *Current Opinion in Genetics & Development* 15.2 (Apr. 2005), pp. 163–176.

[266] Kai Zhang et al. "A cell atlas of chromatin accessibility across 25 adult human tissues". In: *bioRxiv* (2021).

[267] Akihiko Yokoyama et al. "A higher-order complex containing AF4- and ENL-family proteins with P-TEFb facilitates oncogenic and physiologic MLL-dependent transcription". In: *Cancer cell* 17.2 (Feb. 2010), p. 198. URL: /pmc/articles/PMC2824033/%20/pmc/articles/PMC2824033/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2824033/.

[268] Joe R. Harman et al. "A KMT2A-AFF1 gene regulatory network highlights the role of core transcription factors and reveals the regulatory logic of key downstream target genes". In: *Genome Research* 31.7 (June 2021), gr.268490.120. URL: https://genome.cshlp.org/content/early/2021/06/04/gr.268490.120%20https://genome.cshlp.org/content/early/2021/06/04/gr.268490.120.abstract.

[269] Adam C. Wilkinson et al. "RUNX1 Is a Key Target in t(4;11) Leukemias that Contributes to Gene Activation through an AF4-MLL Complex Interaction". In: *Cell Reports* 3.1 (Jan. 2013), pp. 116–127.

[270] Sorcha O'Byrne et al. "Discovery of a CD10-negative B-progenitor in human fetal life identifies unique ontogeny-related developmental programs". In: *Blood* 134.13 (Sept. 2019), pp. 1059–1071. URL: www.hdbr.org.

[271] Denise Ragusa et al. "The RS4;11 cell line as a model for leukaemia with t(4;11)(q21;q23): Revised characterisation of cytogenetic features". In: *Cancer Reports* 2.5 (Oct. 2019), e1207. URL: https://onlinelibrary.wiley.com/doi/full/10.1002/cnr2.1207%20https://onlinelibrary.wiley.com/doi/abs/10.1002/cnr2.1207%20https://onlinelibrary.wiley.com/doi/10.1002/cnr2.1207.

[272] Christopher Wilks et al. "Megadepth: efficient coverage quantification for BigWigs and BAMs". In: *Bioinformatics* (2021).

[273] Yansheng Liu et al. "Multi-omic measurements of heterogeneity in HeLa cells across laboratories". In: *Nature Biotechnology* 37.3 (2019).

[274] Haley M. Amemiya, Anshul Kundaje, and Alan P. Boyle. "The ENCODE Blacklist: Identification of Problematic Regions of the Genome". In: *Scientific Reports 2019 9:1* 9.1 (June 2019), pp. 1–5. URL: https://www.nature.com/articles/s41598-019-45839-z.

[275] McLean CY et al. "GREAT improves functional interpretation of cis-regulatory regions". In: *Nature biotechnology* 28.5 (May 2010), pp. 495–501. URL: https://pubmed.ncbi.nlm.nih.gov/20436461/.

[276] Eva M Trinidad et al. "An impaired transendothelial migration potential of chronic lymphocytic leukemia (CLL) cells can be linked to ephrin-A4 expression". In: (2009). URL: http://ashpublications.org/blood/article-pdf/114/24/5081/1321092/zh804909005081.pdf.

[277] Elvira Infante and Anne J. Ridley. "Roles of Rho GTPases in leucocyte and leukaemia cell transendothelial migration". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 368.1629 (Nov. 2013). URL: /pmc/articles/PMC3785963/%20/pmc/articles/PMC3785963/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3785963/.

[278] Jing Tian et al. "The progress of early growth response factor 1 and leukemia". In: *Intractable & Rare Diseases Research* 5.2 (2016), p. 76. URL: /pmc/articles/PMC4869586/%20/pmc/articles/PMC4869586/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4869586/.

[279] Huifen Cao, Claes Wahlestedt, and Philipp Kapranov. "Strategies to Annotate and Characterize Long Noncoding RNAs: Advantages and Pitfalls". In: *Trends in Genetics* 34.9 (Sept. 2018), pp. 704–721. URL: http://www.cell.com/article/S0168952518301094/fulltext%20http://www.cell.com/article/S0168952518301094/abstract%20https://www.cell.com/trends/genetics/abstract/S0168-9525(18)30109-4.

[280] Jie Gao et al. "Aberrant LncRNA Expression in Leukemia". In: *Journal of Cancer* 11.14 (2020), p. 4284. URL: /pmc/articles/PMC7196264/%20/pmc/articles/PMC7196264/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7196264/.

[281] Jenuwein T and Allis CD. "Translating the histone code". In: *Science (New York, N.Y.)* 293.5532 (Aug. 2001), pp. 1074–1080. URL: https://pubmed.ncbi.nlm.nih.gov/11498575/.

[282] Shan Lin et al. "The Transcriptome Heterogeneity of MLL-Fusion ALL Is Driven By Fusion Partners Via Distinct Chromatin Binding". In: *Blood* 128.22 (Jan. 2016), p. 576.

[283] Annesley CE and Brown P. "The Biology and Targeting of FLT3 in Pediatric Leukemia". In: *Frontiers in oncology* 4.SEP (2014). URL: https://pubmed.ncbi.nlm.nih.gov/25295230/.

[284] David W. Sternberg and Jonathan D. Licht. "Therapeutic intervention in leukemias that express the activated fms-like tyrosine kinase 3 (FLT3): Opportunities and challenges". In: *Current Opinion in Hematology* 12.1 (Jan. 2005), pp. 7–13.

[285] Amy N. Sexauer and Sarah K. Tasian. "Targeting FLT3 signaling in childhood acute myeloid leukemia". In: *Frontiers in Pediatrics* 5 (Nov. 2017).

[286] Elvira Infante and Anne J. Ridley. "Roles of Rho GTPases in leucocyte and leukaemia cell transendothelial migration". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 368.1629 (Nov. 2013). URL: /pmc/articles/PMC3785963/%20/pmc/articles/PMC3785963/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3785963/.

[287] Ding Z et al. "Leukemia-Associated Rho Guanine Nucleotide Exchange Factor and Ras Homolog Family Member C Play a Role in Glioblastoma Cell Invasion and Resistance". In: *The American journal of pathology* 190.10 (Oct. 2020), pp. 2165–2176. URL: https://pubmed.ncbi.nlm.nih.gov/32693062/.

[288] Jennifer L. Wilson et al. "Pathway-based network modeling finds hidden genes in shRNA screen for regulators of acute lymphoblastic leukemia". In: *Integrative Biology* 8.7 (July 2016), pp. 761–774. URL: https://academic.oup.com/ib/article/8/7/761/5115211.

[289] Laura Godfrey et al. "DOT1L inhibition reveals a distinct subset of enhancers dependent on H3K79 methylation". In: *Nature Communications 2019 10:1* 10.1 (June 2019), pp. 1–15. URL: https://www.nature.com/articles/s41467-019-10844-3.

[290] Feng Q et al. "Methylation of H3-lysine 79 is mediated by a new family of HMTases without a SET domain". In: *Current biology : CB* 12.12 (June 2002), pp. 1052–1058. URL: https://pubmed.ncbi.nlm.nih.gov/12123582/.

[291] S. B. Van Oss, C. E. Cucinotta, and K. M. Arndt. "Emerging insights into the roles of the Paf1 complex in gene regulation". In: *Trends Biochem. Sci.* 42.10 (Oct. 2017), pp. 788–798.

[292] Y. Yang. "PAF complex plays novel subunit-specific roles in alternative cleavage and polyadenylation". In: *PLoS Genet.* 12.1 (2016), e1005794.

[293] Liming Hou et al. "Paf1C regulates RNA polymerase II progression by modulating elongation rate". In: *Proceedings of the National Academy of Sciences* 116.29 (July 2019), pp. 14583–14592. URL: https://www.pnas.org/content/116/29/14583%20https://www.pnas.org/content/116/29/14583.abstract.

[294] J. A. Jaehning. "The Paf1 complex: platform or player in RNA polymerase II transcription?" In: *Biochimica et Biophysica Acta. Protein Structure and Molecular Enzymology* 1799.5-6 (2010), pp. 379–388.

[295] Li Ding et al. "The Paf1 complex positively regulates enhancer activity in mouse embryonic stem cells". In: *Life Science Alliance* 4.3 (Dec. 2021). URL: /pmc/articles/PMC7772781/%20/pmc/articles/PMC7772781/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7772781/.

[296] Zhao H et al. "CrossMap: a versatile tool for coordinate conversion between genome assemblies". In: *Bioinformatics (Oxford, England)* 30.7 (Apr. 2014), pp. 1006–1007. URL: https://pubmed.ncbi.nlm.nih.gov/24351709/.

[297] Hongduo Sun et al. "Quantitative integration of epigenomic variation and transcription factor binding using MAmotif toolkit identifies an important role of IRF2 as transcription activator at gene promoters". In: *Cell Discovery 2018 4:1* 4.1 (July 2018), pp. 1–4. URL: https://www.nature.com/articles/s41421-018-0045-y.

[298] Oriol Fornes et al. "JASPAR 2020: update of the open-access database of transcription factor binding profiles". In: *Nucleic Acids Research* 48.D1 (Jan. 2020), pp. D87–D92. URL: https://academic.oup.com/nar/article/48/D1/D87/5614568.

[299] Parul Johri et al. "The Impact of Purifying and Background Selection on the Inference of Population History: Problems and Prospects". In: *Molecular Biology and Evolution* 38.7 (June 2021), pp. 2986–3003. URL: https://academic.oup.com/mbe/article/38/7/2986/6137841.

[300] K Zeng. "A coalescent model of background selection with recombination, demography and variation in selection coefficients". In: *Heredity 2013 110:4* 110.4 (Nov. 2012), pp. 363–371. URL: https://www.nature.com/articles/hdy2012102.

[301] Jeffrey R. Adrion et al. "A community-maintained standard library of population genetic models". In: *eLife* 9 (June 2020), pp. 1–39.

[302] Anna Gelencsér-Horváth et al. "Fast, parallel implementation of particle filtering on the GPU architecture". In: *EURASIP Journal on Advances in Signal Processing 2013 2013:1* 2013.1 (Sept. 2013), pp. 1–16. URL: https://asp-eurasipjournals.springeropen.com/articles/10.1186/1687-6180-2013-148.

[303] Lawrence Murray. "GPU acceleration of the particle filter: the Metropolis resampler". In: (Feb. 2012). arXiv: 1202.6163. URL: http://arxiv.org/abs/1202.6163.

[304] Felipe Lopez et al. "Particle filtering on GPU architectures for manufacturing applications". In: *Computers in Industry* 71 (Aug. 2015), pp. 116–127.

[305] Francesco Montinaro et al. "Revisiting the Out of Africa event with a novel Deep Learning approach". In: *bioRxiv* (Dec. 2020), p. 2020.12.10.419069. URL: https://www.biorxiv.org/content/10.1101/2020.12.10.419069v2%20https://www.biorxiv.org/content/10.1101/2020.12.10.419069v2.abstract.

[306] Zhanpeng Wang et al. "Automatic inference of demographic parameters using generative adversarial networks". In: *Molecular Ecology Resources* (2021). URL: https://onlinelibrary.wiley.com/doi/full/10.1111/1755-0998.13386%20https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.13386%20https://onlinelibrary.wiley.com/doi/10.1111/1755-0998.13386.

[307] Iosif Lazaridis et al. "Genomic insights into the origin of farming in the ancient Near East". In: *Nature 2016 536:7617* 536.7617 (July 2016), pp. 419–424. URL: https://www.nature.com/articles/nature19310.

[308] Iosif Lazaridis et al. "Paleolithic DNA from the Caucasus reveals core of West Eurasian ancestry". In: *bioRxiv* (Sept. 2018), p. 423079. URL: https://www.biorxiv.org/content/10.1101/423079v1%20https://www.biorxiv.org/content/10.1101/423079v1.abstract.

[309] Eugene I. Smith et al. "Humans thrived in South Africa through the Toba eruption about 74,000 years ago". In: *Nature* 555.7697 (2018), pp. 511–515.

[310] Eugenio Marco et al. "Multi-scale chromatin state annotation using a hierarchical hidden Markov model". In: *Nature Communications 2017 8:1* 8.1 (Apr. 2017), pp. 1–9. URL: https://www.nature.com/articles/ncomms15011.