Statistical analysis of short template switch mutations in human genomes

Conor Reece Walker

European Bioinformatics Institute
Hughes Hall
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

October 2021

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Conor Reece Walker
October 2021

Abstract

Statistical analysis of short template switch mutations in human genomes

Conor Reece Walker

Many complex rearrangements arise in human genomes through template switch mutations, which occur during DNA replication when there is a transient polymerase switch to an alternate template nearby in three-dimensional space. These variants are routinely captured at kilobase-to-megabase scales in studies of genetic variation by using methods for structural variant calling. However, the genomic and evolutionary consequences of replication-based rearrangements remain poorly characterised at smaller scales, where they are usually interpreted as complex clusters of independent substitutions, insertions and deletions. In this thesis, I describe statistical methods for the detection and interpretation of short template switch mutations within DNA sequence data. I then use my methods to explore small-scale template switch mutagenesis within human genome evolution, population variation, and cancer. I show that small-scale, replication-based rearrangements are a ubiquitous feature of the germline and somatic mutational landscape of human genomes.

Acknowledgements

First and foremost, I would like to thank my supervisor Nick Goldman for his mentorship and support throughout my PhD. Nick has provided a fantastic group in which I have been able to explore ideas freely, and has helped shape those ideas with scrutiny, patience, and statistical intuition. Additionally, my co-supervisors Aylwyn Scally and Nicola De Maio have given me fantastic guidance over the past several years, and I have always left our conversations feeling energised to learn.

I would like to thank my Thesis Advisory Committee members Jan Korbel and Zamin Iqbal for their thoughtful comments and discussions on human genomic variation and variant calling methodology throughout my PhD. I would also like to thank Isidro Cortés-Ciriano for making time for some great discussions on the analysis of human cancer genomes.

My time at EMBL-EBI and the University of Cambridge has allowed me to grow both as a scientist and as an individual, and I am grateful for my time at EMBL more broadly, as it has cemented in me a strong sense of both European and global identity. I am especially thankful for the many international friendships I have made during my time here. In particular, Iain, Kai, Natalie, Veronika, Umberto, Borgthor, Holly, Aleix, Charlie, and Jack have all made these past few years more memorable.

I specifically want to thank José for his friendship throughout my time in Cambridge, for the many conversations both stimulating and silly, and for keeping me sane during the writing of this thesis. I would also like to thank Aline and Georgi for their consistent joy-filled company, and Will for the continuous encouragement and advice.

A special thank you is also owed to Pamela, who has helped me grow as a person, and has given me a much needed and consistent morale boost during a global pandemic. Finally, I want to thank my parents for their belief in me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Historical context

Deoxyribonucleic acid (DNA) molecules carry the instructions necessary to encode all observable cellular life on Earth. This remarkable molecule was first isolated as a "novel precipitate" by Friedrich Miescher in a Tübingen castle in 1869, and named "nuclein" due to its occurrence in the nuclei of cells [67, 210]. The significance of these white, swirling chemicals preciptated from the pus on surgical bandages was not fully appreciated at the time, and this work went largely unnoticed in the scientific community. The molecule was later isolated independently by Richard Altmann in 1889, and was given the name "nucleinsäuren" (or nucleic acid) [13]. This changed nomenclature would ultimately become the preferred term in the scientific record, following its use throughout the studies of Albrecht Kossel in the late 19th and early 20th century, in which the chemical composition of nucleic acids was successfully resolved. This work attracted widespread attention in the scientific community and was subsequently awarded a Nobel prize in 1910 [161].

While it was accepted in the community that this molecule must have some important biological function, its role in the transmission of genetic information was not known. This changed following a series of experiments by Oswald Avery, Colin MacLeod, and Maclyn McCarty in the 1940s [21]. Building on earlier work by Frederick Griffith in the late 1920s [110], the "Avery–MacLeod–McCarty" experiments demonstrated what is now known as the transformation principle [21]. Particles of a heat-killed virulent (smooth-coated) strain of *Streptococcus pneumoniae* were mixed with a non-virulent (rough-coated) strain and injected into mice. Smooth-coated bacteria emerged, maintaining their virulence in subsequent generations, demonstrating that some information from the heat-killed smooth-coated strain had "transformed" the initial non-virulent, rough-coated strain. The only chemical in their experiments that prevented this transformation was desoxyribonucleodepolymerase (now known as deoxyribonuclease I or DNase I), an enzyme which degrades DNA, leading to the conclusion

that DNA was responsible for the observed between-strain and subsequent transgenerational information transfer.

Shortly after DNA was established as the medium for genetic information, pioneering X-ray diffraction work led by Rosalind Franklin provided us with our first glimpse of DNA, producing the now famous "Photo 51". This subsequently led to an accurate description of DNA structure, the now ubiquitous double-stranded helix containing nucleotide base pairs [308]. The existence of base pairing between two strands provided an explanation for how the genetic code could be copied as separate "template" molecules. Replication of the strands as individual templates to produce two progeny molecules was first demonstrated in *Escherichia coli* in 1958 [208], and errors in this process could conveniently explain the presence of mutations in one of the two progeny molecules. Confirmation of a molecule which contains hereditary information and can generate new mutations through replication error tied together, at the molecular level, the early theories of genetic inheritance by Gregor Mendel and natural selection by Charles Darwin.

The information provided by the ordering of the base pairs along any given DNA molecule was soon resolved into what we now know as the genetic code [65, 224], groups of three nucleotides (codons) which redundantly encode amino acids, underlying the central dogma of molecular biology [66]. These breakthroughs in understanding the genetic code were made without the technological capacity to decipher it, however. This changed in the 1970s with the development of cloning and sequencing techniques, enabling the concept of a gene to be defined (a concept which remains somewhat debated [97]), followed by the sequencing of a complete protein-coding gene from bacteriophage MS2 [214], and four years later, the full sequencing of its entire genome [88].

The era of whole genome sequencing had begun, and refinements to sequencing methods by Fred Sanger and colleagues in the 1970s (so-called "Sanger sequencing") [263] ultimately facilitated the global scientific effort of the Human Genome Project, which produced the first full human genome sequence in 2001 [168]. Parallel private endeavours in genome sequencing also enjoyed success [300], but it is the collaborative nature of the academic scientific community which has seen the production of vast amounts of publicly available sequencing data since the initial human genome sequence was released. Processing and extracting meaningful information from these vast DNA sequence repositories has necessitated the development of novel computational methods. This computational view of DNA has allowed researchers to investigate a vast array of biological hypotheses and problems which spin out of these data repositories much like Miescher's novel precipitate two centuries earlier.

Mutations identified in these sequencing data are now an integral part of studying many areas of biology, including molecular evolution, population genetics, and medical genetics.

Arising during errors in DNA replication and DNA repair, mutations most commonly occur on small scales, changing individual nucleotides, or acting to insert or delete small numbers of nucleotides. Computationally characterising mutations as captured in these progeny sequences has given us our current understanding of the evolutionary history of life on Earth. With a focus on human genomes, this thesis seeks to explore a poorly characterised mutational process involving small-scale rearrangements which instantaneously creates apparent clustered mutations in sequence data.

1.2 Background

Human genetic variation most commonly arises as single nucleotide polymorphisms (SNPs) and small (< 50 nucleotides (nt)) insertions and deletions (indels) due to unrepaired DNA damage (see §1.2.2) and replication polymerase errors (see §1.2.3). Mutations also occur less frequently at larger scales ($\geqslant$ 50nt), but due to their size often constitute a larger proportion of between-individual genomic differences than SNPs and indels [229, 284]. Large-scale mutations are collectively referred to as structural variants, caused by DNA repair pathways responding to replication stress and DNA damage such as double-strand breaks [111]. Many structural variants occur during DNA replication (§1.2.3), and are mediated by template switching (see §1.2.4), a process in which the nascent DNA strand invades and replicates a physically-proximal alternate DNA template strand. Template switches are typically observed introducing kilobase to megabase scale genomic rearrangements, and the alternate template can be utilised for replication until a new telomere is formed [16, 120, 172, 277, 324]. The scale of these rearrangements has consequently permitted direct capture through standard computational structural variant calling pipelines [187]. As a result, large-scale template switches are routinely captured (alongside other forms of structural variation) in large-sample studies of human population genetic variation [12, 80, 180, 284], and their contribution to human genome evolution and disease is well-established [48, 126, 172, 280, 319, 324].

Although large-scale template switching is routinely considered in studies of human genetic diversity and disease [60, 284], evidence for small-scale template switching generally does not receive the same consistent scrutiny. The lack of consideration is likely because detecting template switches requires the identification an alternate template region which may have been utilised to generate the observed variation. While alternate-location mapping of sequencing reads to detect source template DNA is an intrinsic property of many structural variant calling methods [187], these are typically designed to only capture rearrangements that are defined to be $\geqslant$ 50nt in length [53, 60, 284]. Consequently, regardless of the underlying causative

mechanism, germline and somatic genetic variation at smaller scales can conventionally only be captured, represented, and interpreted as some combination of independent SNPs and/or short indels [98, 293]. If occurring at length of < 50nt, the presence of these template switches has therefore remained undetected, and their role as a process driving genetic diversity and human disease is not understood.

Assuming template switches occur at small scales and are unaccounted for, the consequences of these rearrangements will appear as clusters of SNPs and/or short indels between pairs of closely-related DNA sequences (see §1.2.5 and [185]), as there is no alternative way to capture and represent the mutation using standard methodology. I aim to provide such an alternative approach for capturing small-scale rearrangements in Chapter 2 of this thesis, where I describe methods for identifying and assigning statistical significance to small-scale template switch variants, allowing their prevalence to be explored in human genomic data. These are (to the best of my knowledge) the first statistical methods specifically designed to capture template switching at small scales, and as such permit the first statistical study of small-scale template switching in human genomes.

Attributing apparent clusters of mutations within the reference human genome to small-scale template switching was explored by Löytynoja and Goldman [185]. The primary aim of this study was to identify the presence of short template switch mutations within the reference human genome, in order to determine if this form of variation occurred during the divergence of the human and chimpanzee genome. However, due to both methodological limitations (detailed in §2.1.3) and a lack of phylogenetic resolution for individual candidate template switches, the evolutionary conclusions which could be drawn from this study were limited. The contribution of small-scale template switching to human (and by extension great ape) genome evolution therefore remains an open question. For example, it remains unclear how many apparent mutation clusters in hominid genomes are genuine, consecutive, independent mutation events, compared to statistically-assessed single mutations introduced through and error-prone DNA repair pathway involving template switching. If uncharacterised, it is possible that such clusters could be problematic for tests that involve assessing nearby variants are independent mutational events, such as when evaluating evidence for increased substitution rates in humans relative to outgroup genomes [130, 235, 236]. Additionally, no study to date (including [185]) has explored the genomic or sequence features which may influence event formation, or explored the functional constraints governing event tolerance in the genome.

I explore these topics in Chapter 3, where my methods allow me to ask: how many significant template switches are present in the (reference) human, chimpanzee, and gorilla genomes; on which branch of the hominid phylogeny did each event occur; which sequence and

genomic features are associated with event formation; where events occur; and if small-scale template switch mutations generated false signals of lineage-specific accelerated evolution in earlier studies.

Due to a lack of suitable methodology, the contribution of short template switch mutations to human population genetic variation and human genomic disorders also remains unstudied. This is an important knowledge gap to address. Understanding the population genetics of short template switches will provide a more complete picture of both the mutational processes generating human diversity, and will possibly provide insight into the selective pressures acting on regions which contain these variants. Establishing the number of apparent clustered mutations which did not arise through independent SNVs can prevent confounding impacts on estimating summary statistics, for example, it may allow more accurate estimations of human mutation rate. Furthermore, genomic rearrangements attributed to large-scale replication-based template switching have been demonstrated for the human genomic disorders Pelizaeus-Merzbacher disease [172] and Temple syndrome [51]. Perhaps most importantly, the role of these variants in human cancer is unknown. Understanding the mutational processes and signatures that are enriched in specific tumour types can be vital for early diagnosis and treatment development [68, 233]. Recently it has been shown that large-scale template switches can introduce complex rearrangements in many tumour types, and are particularly enriched in adenocarcinomas across multiple tissues [180]. If the mutational mechanisms generating these rearrangements (see §1.2.4) in human germline and somatic variation datasets are operating at small scales, but are simply not captured due to methodological limitations, key relationships with disease may be missed. It is therefore important to catalogue human germline and somatic short template switch variants, such that their relevance to population and medical genetics can be studied.

An important existing catalogue for understanding human germline variation was provided by the 1000 Genomes Project [293]. This is a widely-used [325] dataset of human germline variation, reporting SNPs, $<$ 50nt indels, and large structural variants (this project is also attributed with operationally defining the minimum size of a human genomic rearrangement for computational analysis purposes [42, 284, 293]). Note that other important human population germline sequencing studies report variants similarly, not considering short rearrangements [60, 80, 191]. This is also the case for the most recent large-sample human cancer sequencing project, the Pan-Cancer Analysis of Whole Genomes (PCAWG) study [44, 180]. The PCAWG study refers to template switch mutations as "templated insertions", but they were also unable to detect these variants at small scales (see Extended Figure 6 of [179]). This again raises the question: are small-scale template switches also driving some types of human cancer? It is currently

not possible to answer this question, as their presence is not detected due to methodological limitations. Interestingly, in both germline and somatic re-sequencing settings, large-scale template switches have been repeatedly attributed to well-defined mutational pathways [60, 111, 120, 172, 180, 284], each of which has distinctive associated genomic features (see §1.2.4). There is therefore also the scope to ask if these well-defined mutational pathways operate to introduce smaller scale templated insertions than currently understood.

By incorporating the models I detail in Chapter 2 into a larger template switch discovery pipeline, I can ask these questions by providing the small-scale resolution lacking in these large-sample studies of human genetic variation. I ask how problematic these mutations are for short-read mapping to detect; which genomic features modulate their formation; and if evidence exists for the activity of known (large-scale) template switch pathways. In Chapter 4 and Chapter 5, I investigate these topics in the human germline and in human cancer, respectively. In the former, I also ask if the identified variants are distributed within human populations as expected, and in both settings, I also assess how suitable my statistical methods are. These chapters respectively provide the first statistical assessment of replication-based genomic rearrangements at small scales in the human germline and in human cancer.

In the remainder of this chapter, I provide some necessary molecular and computational background for interpreting subsequent chapters. Specifically, I first describe DNA (§1.2.1) and well-established sources of DNA mutation (§1.2.2 and §1.2.3). This is followed by a brief overview of the major pathways involved in structural variant formation (§1.2.4); the reader is encouraged to make note of the mechanisms underlying large-scale template switching in DNA replication, as I will later assess evidence for their activity at small scales. Next, I reflect on issues with identifying and representing small-scale rearrangements using typical variant calling approaches (§1.2.5). I then provide a brief description of standard algorithms for pairwise DNA sequence alignment (§1.2.6), as these form the basis of the methodological ideas used throughout this thesis to confidently identify short template switch mutations.

1.2.1 DNA structure and packaging

DNA molecules are long, double-stranded polymers, each monomer of which consists of a deoxyribose sugar, a phosphate group, and one of four nitrogenous bases: adenine (A), cytosine (C), guanine (G), or thymine (T) [7]. These monomer units are collectively referred to as nucleotides. Each of the two DNA strands consists of a chain of nucleotides which are covalently linked together via their sugar and phosphate groups. The two strands are

antiparallel, held together by complementary hydrogen base pairing (A pairs with T, C pairs with G), and form a double helix three-dimensional structure ([308]).

Within human cells, the majority of DNA is present within nuclei (as nuclear DNA), but there is also a subset of cellular DNA present within mitochondria. For example, there are approximately $2 \times 3.1 \times 10^9$ base pairs of DNA within human cells, only 16569 of which are contained within mitochondrial DNA [17]. Mitochondrial DNA has a fascinating origin, resulting from the embosymbiosis of a proteobacterium around 2.5 billion years ago, which formed the first eukaryotic mitochondrion [109]. This DNA is uniparentally inherited (from the mother in humans), encodes just 13 proteins, and is highly conserved. As a result, it is typically less relevant than nuclear DNA when investigating mutational mechanisms from an evolutionary perspective, and I therefore focus on nuclear DNA in this thesis.

The individual base pairs of DNA each measure around 3.4Å in length [308], and the DNA polymer in which they are contained is negatively charged due to the phosphate ions in the sugar-phosphate backbone. Compacting these long, rigid molecules into the nuclei of cells has therefore necessitated the evolution of efficient packaging of DNA into chromosomes [158]. Chromosomes are composed of chromatin, the basic unit of which is a nucleosome, a structure comprised of eight histone proteins around which approximately 147 base pairs of DNA are wrapped [250]. The number of chromosomes varies in each species; humans are a diploid species, with 22 pairs of autosomes, labelled 1–22), and two sex chromosomes (or allosomes), which are labelled X and Y. The DNA packaged in chromosomes is the substrate on which mutational forces act to drive evolution. There are two broad categories of DNA mutagenesis: DNA damage, and errors in nuclear DNA replication. These forces act in combination to produce all observable genetic variation in cellular life.

1.2.2 DNA damage and pre-replicative repair

Nuclear DNA is continuously exposed to a host of endogenous and exogenous sources of DNA damage which cause the spontaneous mutagenesis which underlies genetic variation. Most endogenous damage occurs due to hydrolytic and oxidative reactions between DNA and molecules which naturally occur within cells [56]. Major sources of endogenous damage are spontaneous base deamination, the formation of apurinic and apyrimidic sites, oxidative damage caused by reactive oxygen species, and DNA methylation. Exogenous DNA damage occurs due to chemical, physical, and environmental agents which directly damage the molecule [56]. The primary sources of exogenous damage are ionising radiation, ultraviolet radiation,

alkylating agents, aromatic amines, and more rarely, environmental stress such as extreme cold or hypoxia [56].

Multiple DNA repair pathways have evolved in humans to remove damaged DNA before it is replicated, involving either excision or direct repair. For example, the nucleotide excision pathway repairs a broad range of bulky lesions by recruiting various enzymes to cleave and re-synthesise short stretches of DNA around the lesion [195]. The base excision repair pathway operates similarly, but uses an alternate set of enzymes to cleave single abasic nucleotides which do not distort the DNA helix [121]. Alternatively, alkylated bases and UV-induced lesions can respectively be directly reversed to their original state by alkyltransferases and DNA photolyases [220, 321]. Any damage which is not repaired by these pathways can interfere with replication of the DNA molecule, causing mutations and potentially inducing chromosomal instability through double-strand break formation.

1.2.3 DNA replication, polymerase errors, and replication stress

The double-stranded structure of DNA facilitates template-mediated replication, in which each strand of the existing DNA molecule is used as a template for generating a progeny strand. The two resulting progeny strands are bound to the template parent strands, producing two new double-stranded DNA molecules at each cell division, each of which contains one of the strands of the original molecule. This process is known as "semi-conservative DNA replication" [7]. The stages involved in both prokaryotic and eukaryotic replication are largely identical, and while there are some small differences, this thesis concerns itself with eukaryotic (i.e. human) replication unless stated otherwise.

A brief overview of the key stages in DNA replication (as outlined in [7]) is shown in Figure 1.1. Replication is initiated at replication origins, positions along the genome at which origin recognition complexes (ORCs) are bound to chromatin *in vivo*. During the G1 phase of the cell cycle, minichromosome maintenance (MCM) helicases combine with ORCs to form a prereplicative complex. This primes replication origins for firing during the S phase of the cell cycle, in which the prereplicative complex is phosphorylated, DNA helicases are activated to unwind the double strand, dozens of enzymes and proteins necessary for replication are recruited, and ORC rebinds. This open region of DNA in which replication is undertaken is collectively referred to as the replisome. DNA unwinding and synthesis occurs at each end of the open replisome, within structures known as replication forks (Figure 1.1b, right). The anti-parallel conformation of the two strands requires that each of the strands is synthesised differently, as replicative polymerases always read DNA in a $5' \rightarrow 3'$ orientation. Nascent

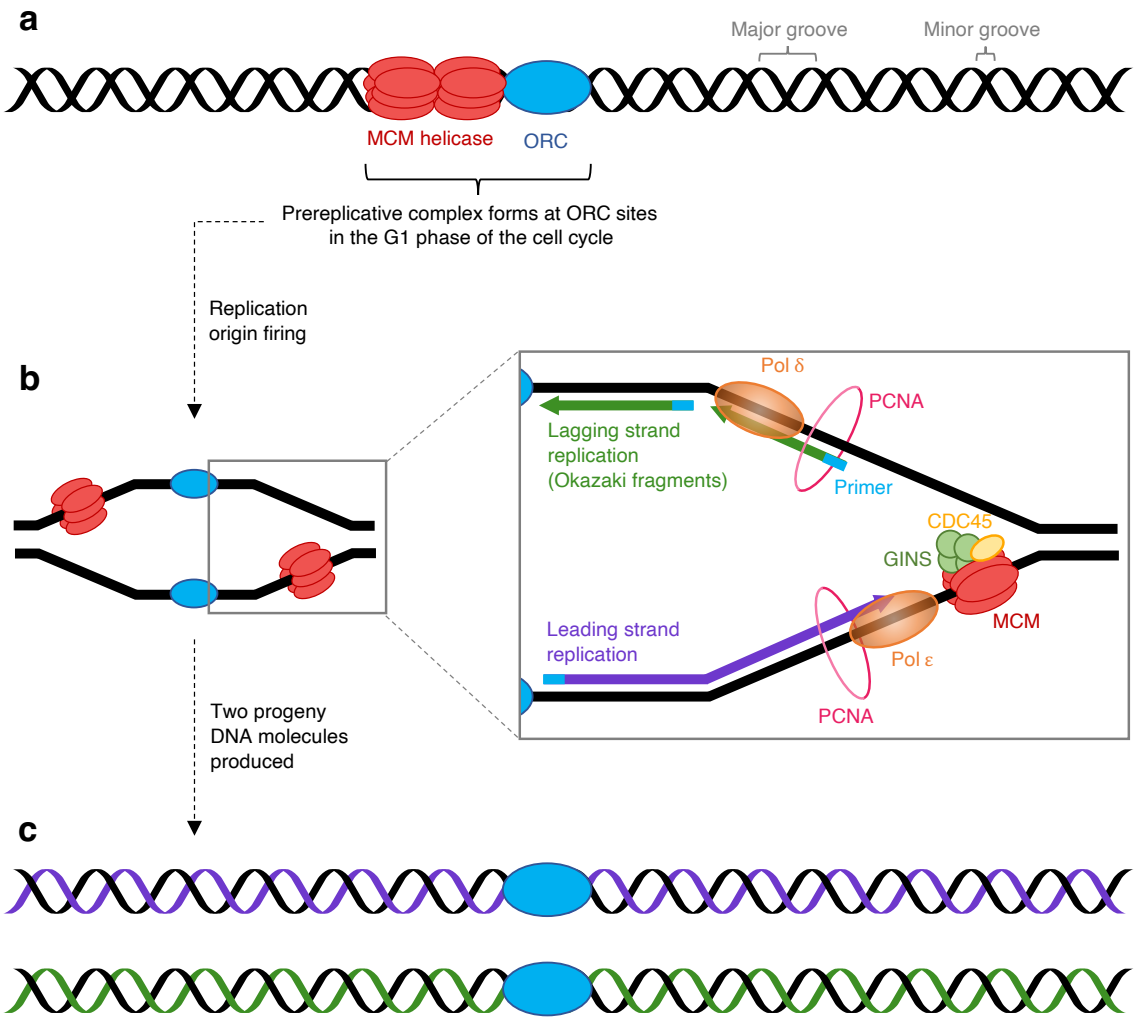

Figure 1.1: An overview of replication of a eukaryotic double-stranded DNA molecule. **(a)** A double-stranded DNA polymer composed of two chains of nucleotide monomers, with prereplicative complexes (composed of MCM helicases and ORC proteins) assembled along the chromatin associated with the DNA molecule. **(b)** Firing of the prereplicative complex gives rise to the replisome. The architecture of one replication fork within the replisome is shown on the right; note that only key proteins involved are depicted and that DNA synthesis proceeds bidirectionally in a similar fashion at both replication forks. The parental DNA is unwound by the core replicative helicase complex (composed of the MCM helicase, CDC45 protein, and the GINS protein complex). RNA-DNA hybrid primers are incorporated into open templates on each strand by Polymerase (Pol) α primase (not shown). Pol-ε and Pol-δ then catalyse the synthesis of the nascent leading (purple) and lagging (green) strands, respectively, aided by the processivity factor proliferating cell nuclear antigen (PCNA), which acts as a sliding clamp behind each polymerase. After Okazaki fragments have been synthesised, primers are removed and fragments are joined by DNA ligase I (not shown). **(c)** Two progeny molecules are generated, each of which contains one strand from the parental DNA molecule. Adapted from [7] and [32].

leading strand synthesis occurs continuously, whereas nascent lagging strand synthesis occurs discontinuously, through the construction of many short DNA pieces known as "Okazaki fragments" [227], which are ligated after their synthesis to create the continuous progeny strand. As the replication fork progresses, the phosphorylated ORCs behind the fork are displaced, and unphosphorylated ORCs bind to the replication origins in their place, yielding the two "semi-conserved" progeny DNA molecules, each of which is bound with ORCs ready for processing in the following cell cycle [7].

Focusing on humans, there are at least sixteen DNA polymerases which have been observed during replication [106, 165]. The majority of replication is carried out by just three of these: Pol-α, Pol-δ, and Pol-ε. Pol-α generates the RNA-DNA hybrid primers which are incorporated into open single-stranded DNA at the start of synthesis of the leading strand, and more frequently, the start of each Okazaki fragment. Pol-δ performs contiguous Okazaki fragment chain elongation in lagging strand replication, while Pol-ε is used for continuous chain elongation in leading strand replication. The error rates of each of these polymerases is very low, around 10^{-4}–10^{-5} per base pair per cell cycle [160, 268]. The overall fidelity of DNA replication is much higher however, at around 10^{-8}–10^{-10} errors per base pair per cell cycle [26, 199]. This is due to both the $3' \rightarrow 5'$ exonuclease activity of Pol-δ and Pol-ε, which proofreads the growing DNA chain, and the post-replicative DNA mismatch repair pathway [129, 165]. Mismatch repair operates on newly-synthesised DNA, correcting spontaneous base-base mismatches as well as small insertions and deletions which arise during both replication and recombination [177, 232]. When mutations do escape Pol-δ and Pol-ε selectivity, proofreading, and mismatch repair, the introduced mutations tend to be single nucleotide substitutions, insertions and deletions, and cause the majority of observable variation within human genomes [199].

Despite this high fidelity, obstacles encountered during DNA replication such as unrepaired DNA lesions can impede replication fork progression (for a review, see [32]), and multiple DNA-damage tolerance pathways have evolved to respond to such replication stress [32, 194, 320]. Many DNA lesions are repaired by translesion synthesis polymerases [260], some of which are error-prone and leave distinctive patterns of multinucleotide mutations in yeast and human genomes [117, 281]. Other DNA lesions are skipped during replication fork progression, leaving a single nucleotide gap of single-stranded DNA that can be repaired post-replication [32]. If stalled forks are not restarted and instead collapse, or if single-stranded DNA gaps persist post-replication, double-strand breaks can form which can cause chromosomal instability and structural variant formation [19, 49, 296].

1.2.4 Mechanisms underlying genomic rearrangements

As discussed previously, large-scale mutations are collectively referred to as structural variants, which are operationally defined as mutations which impact $\geqslant$ 50nt (although several-megabase variants are not uncommon [284]). Structural variants are often associated with genomic disorders [49], and arise through a variety of mutational pathways which frequently alter gene copy number through large-scale deletions, insertions, duplications, inversions, and translocations [123, 284]. A subset of structural variants also manifest as complex rearrangements, in which multiple large-scale mutations give rise to genomic regions which contain a combination of distinct structural variant classes [48, 50, 172, 324].

The pathways that underlie structural variant formation can be broken down into two major categories: recurrent rearrangements, which are identical in size and nucleotide composition in unrelated genomes; and non-recurrent rearrangements, which are distinct in size and composition in unrelated genomes [49]. Many recurrent structural variants are caused by the non-allelic homologous recombination (NAHR) pathway, in which ectopic crossover between low-copy repeats (i.e. segmental duplications) in either direct or inverted orientation produces reciprocal duplications and deletions. A typical human genome contains approximately 7000–9000 structural variants with respect to a reference human genome, and NAHR is thought to produce around 10% of these variants [42, 60, 284]. A further 10% of detected structural variants have been attributed to mobile element insertion in the reference genome and expansion/contraction of tandem repeats [284]. The approximately 80% of remaining structural variants per human genome have been statistically attributed to a variety of non-recurrent rearrangement mechanisms which act to repair double-strand breaks. The most common of these is non-homologous end joining (NHEJ), but increasingly replication-based pathways are attributed to non-recurrent structural variant formation, including break-induced replication (BIR), microhomology-mediated break-induced replication (MMBIR), fork stalling and template switching (FoSTeS), and serial replication slippage (SRS) [49, 284]. I will briefly cover these four non-recurrent rearrangement pathways (also see [49] for a review), as it will be useful for contextualising the small-scale rearrangements explored in subsequent chapters of this thesis. Specifically, human germline and somatic template switches underlying large non-recurrent structural variant calls are frequently attributed to the FoSTeS/MMBIR pathways [49, 180, 284], and evidence for their involvement in the generation of short template switch mutations will be considered in Chapter 4 and Chapter 5.

In humans, NHEJ is the most common pathway by which double-strand breaks are repaired, and is active throughout the cell cycle [143] (see Figure 1 of [55] for a diagrammatic overview). Repair involves initial activity by the endonuclease Artemis in complex with the kinase DNA-

PKcs to expose short regions of single-stranded DNA with shared identity ("microhomology") at the break points of each strand. These stretches of microhomology are then joined by DNA ligase IV, and nucleotides are added on each strand independently in either a template-dependent or template-independent manner by DNA polymerases Pol-μ and Pol-λ [55, 202]. This process often introduces mutations at repair junctions due to nuclease activity and the template-independent nucleotide synthesis, which often manifest as small-scale deletions, but also as larger-scale insertions and deletions if multiple rounds of resection and synthesis were involved [55, 99, 228].

The BIR and MMBIR pathways are able to respond to nicks in template strands which cause replication fork collapse when encountered during replisome progression [49, 120] (see Figure 3 of [49] for diagrams). In BIR, $5'$ resection at the single double-strand end break is followed by RecA/Rad51-mediated invasion of the homologous sequence on the sister chromatid to form a displacement loop structure in which replication proceeds until the chromosome end [259]. MMBIR operates similarly but in a RecA/Rad-51 independent manner (likely when cellular stress has caused a depletion of RecA/Rad-51), instead relying on short regions of sequence identity to facilitate repeated strand invasion of physically proximal single-stranded DNA [120].

The FoSTeS pathway operates similarly to MMBIR (and these are sometimes referred to jointly as "FoSTeS/MMBIR" [49, 324]), but is caused by a stalled fork rather than a nicked template strand. After a replication barrier such as DNA secondary structure causes fork stalling, the nascent strand becomes free from its template, and undergoes $5'$ resection of the break followed by multiple microhomology-mediated invasions of replication forks which are proximal in 3D space and may be proceeding in either $5' \rightarrow 3'$ or $3' \rightarrow 5'$ direction in relation to the originating leading strand [172, 324]. As FoSTeS/MMBIR both involve invading open single-stranded DNA (i.e. through template switching) solely based on microhomology, each has the potential to cause gross genomic rearrangements. Importantly, because microhomology mediates strand invasion in the FoSTeS/MMBIR pathway(s), evidence for their activity can be reliably found by inspecting patterns of microhomology at the site of the $5'$ resection and the site of strand invasion on the alternate template in the containing variant call data [167, 284]. Further, these pathways have been associated with late replicating regions of the genome [156, 180], where accumulation of single-stranded DNA causes increased rates of DNA damage [279]. In combination, these two signatures will later be valuable for assessing evidence for the potential activity of FoSTeS/MMBIR in generating small-scale template switch mutations (see §4.7.4, §4.7.5, §5.5.3, and §5.5.6).

The final replication-based mechanism, SRS, is an extension of the well-established replication slippage model (to which the majority of short insertions and deletions in human genomes are attributed [218]), involving the generation of large deletions and duplications through multiple slipped strand mispairing events.

1.2.5 The challenge of identifying small-scale genomic rearrangements

The purpose of the seemingly arbitrary 50nt cutoff between an observed mutation that is interpreted as a structural variant (with an associated alternate template), and a mutation which is considered as (e.g.) a short indel, is largely to distinguish between mutations which are callable from single sequencing reads and those which require paired end or split read mapping using different variant calling strategies [49, 123, 187, 284]. This cutoff has also had the benefit of simplifying the process of identifying and assigning causative mechanisms to rearrangements [49, 80, 180].

For an example, consider a sequence region containing the consequences of a replication-based rearrangement mechanism. There are two possible ways of representing this sequence: either as some combination of clustered single nucleotide variants and indels with respect to an ancestral or reference sequence, or as a single structural variant with some associated alternate template(s) elsewhere in the ancestral/reference sequence. While both of these descriptions fully define the descendant sequence as a set of differences with respect to the ancestral sequence, they imply different causes, at most one of which is correct. If the genomic rearrangement explanation is to be believed over the clustered SNV and indel explanation, we need to have confidence that the alternate template(s) possibly involved in the rearrangement was not present simply by chance (which becomes increasingly likely as genome size increases). A cut-off in structural variant definition of 50nt allows this issue to be sidestepped somewhat. It seems highly unlikely that an (unmasked) alternate template corresponding to an identified $\geqslant$ 50nt structural variant region can be found by chance anywhere else in e.g. a human-sized genome, especially when considering the co-occurence of breakpoint microhomology [49]. While this size cutoff is useful for computationally identifiying genomic rearrangmeents with confidence, it has left potential activity of rearrangement pathways unaccounted for and understudied at small scales.

1.2.6 Capturing small-scale variation using pairwise sequence alignment

To consider then how small-scale rearrangements (i.e. short template switch mutations) could be studied using DNA sequence data, it is worth examining well-established methodology

used for identifying single nucleotide substitutions and indels in these same data. The typical approach for capturing mutations involves interpreting a given DNA sequence with respect to an ancestral sequence, from which it may be possible to infer the positions in the two sequences that share an evolutionary history, and if so, which mutational operations have possibly changed these positions. This is classically performed using alignment algorithms. In the case of comparing just two sequences, pairwise alignment algorithms in particular seeks to identify regions of similarity between just two input sequences of possibly variable length. When comparing DNA sequences, this allows us to ask if the two DNA sequences are related by deciding if an alignment is more likely to have arisen through an evolutionary relationship, or simply by chance.

The standard approach for generating pairwise alignments (and possibly assessing evolutionary hypotheses) utilises dynamic programming (originally outlined by Needleman and Wunsch [221]), and a relatively efficient algorithm detailed by Gotoh [107] is widely used in computational biology. That is, given two input DNA sequences $x_1,\ldots,x_i,\ldots,x_n$ and $y_1,\ldots,y_j,\ldots,y_m$, a score matrix A (indexed by $i = 1,...,n$ and $j = 1,...,m$) is initialised with $A(0,0) = 0$ and then recursively filled using

$$A(i,j) = \max \begin{cases} A(i-1,j-1) + s(x_i,y_j), \\ A(i-1,j) - d, \\ A(i,j-1) - d. \end{cases} \tag{1.1}$$

In Equation 1.1, $s(x_i,y_j)$ is a scoring function which typically outputs a positive score for $x_i = y_j$, and a negative score for $x_i \neq y_j$, and d is a penalty for introducing a gap (corresponding to an inserted or deleted nucleotide in one sequence). As $A(i,j)$ is constructed using Equation 1.1, a pointer is recorded for each cell in the matrix to indicate the cell from which it was derived (i.e. which cell out of $A(i-1,j-1)$, $A(i-1,j)$, and $A(i,j-1)$ contributed to the max term). The optimal global alignment between sequences x and y is then traced back from $A(n,m)$ using these pointers; an example of this is shown in Figure 1.2. Note that this type of pairwise alignment is referred to as a "global" pairwise alignment, as it aligns the full length of both sequences x and y. A common alternative to this is to perform "local" pairwise alignment (most classically, by using the Smith-Waterman alignment [278]), whereby the start and end positions of the pairwise alignment are not required to include the entirety of both sequences.

This basic setup for pairwise alignment finds an optimal alignment between the input pair of sequences by using a simplistic scoring scheme. Simple scoring schemes such as this seek to find some balance of positive and negatives scores such that the maximum number of

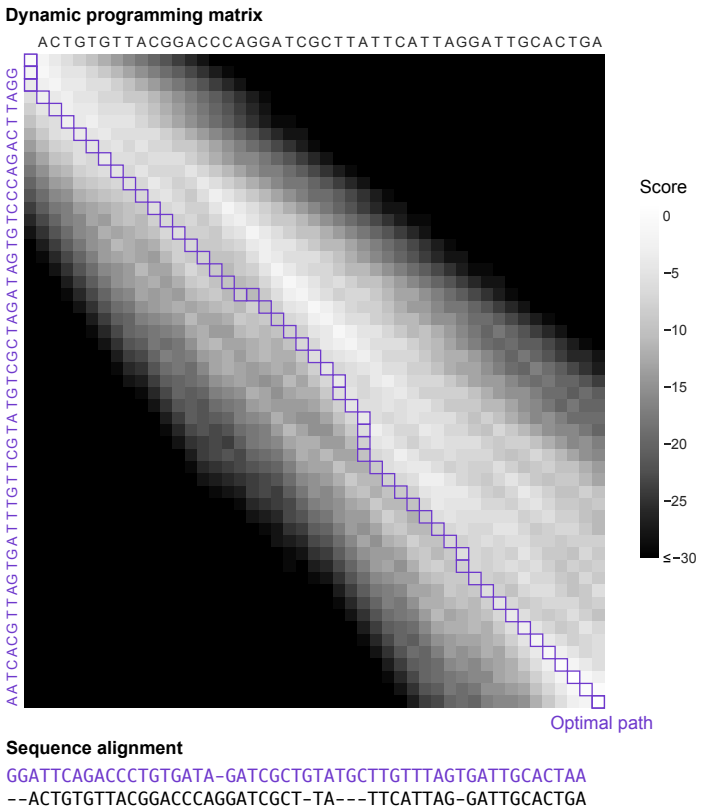

Figure 1.2: A global pairwise alignment matrix. $A(i,j)$ is constructed using Equation 1.1 between sequences x and y, respectively shown above and to the left of the matrix. Here, $s(x_i, y_j) = 1$ if $x_i = y_j$, and -1 otherwise; $d = -2$. The optimal path through this score matrix is shown, and the pairwise alignment resulting from traceback of this path is shown below the matrix. Note that in this particular case, alignment scores are consistently negative due to the high level of divergence between sequences x and y, yielding many negatively-scored mismatch and indel positions.

elements x_i and y_j which are aligned are either the same ($x_i = y_j$) or not the same ($x_i \neq y_j$), while penalising the use of gaps (corresponding to insertions or deletions). This of course is not biologically meaningful. What we actually want is to find the positions between the two sequences which share an evolutionary history, such that mutations can be identified and evolutionary hypotheses can be tested. Fortunately this problem has been well-studied, and I will describe more biologically meaningful scoring schemes later in this thesis (see §2.2.5).

Standard pairwise alignment as defined here (Equation 1.1) finds the optimal alignment between two possibly related sequences by assuming that the only mutational operations are SNPs, insertions, and deletions. As discussed previously, if one of the replication-based rearrangement pathways outlined in §1.2.4 was also involved in the divergence of the sequences undergoing pairwise alignment, the only meaningful way typical alignment algorithms have to

represent this variation is as some combination of clustered single nucleotide variants and indels within the pairwise alignment. A typical alignment-based approach for identifying small-scale variation as outlined here is therefore not able to adequately identify or represent this variation between two sequences. I will bridge this gap in Chapter 2, presenting statistical models for identifying rearrangements without a minimum length restriction by using the framework of pairwise sequence alignment. These methods will allow me to ask, for any input pair of nucleotide sequences, if the observed differences between the two sequences are more likely to have arisen through only SNPs and indels, or through any combination of SNPs, indels, and a small-scale template switch mutation.

1.3 Thesis outline and publications

1.3.1 Thesis structure

This thesis outlines statistical methods for capturing replication-based rearrangement mechanisms within a small sequence window, allowing me to model local template switches by using the framework of pairwise alignment, and then explores the human genome landscape of short replication-based rearrangements. Chapters are ordered such that the methods for identifying these rearrangements are first described, followed by the application of these methods to various datasets. Specifically:

Chapter 1: Here I have provided an introduction to the problem addressed by my thesis, and some necessary background on the molecular biology and class of algorithms which are core to the work that follows.

Chapter 2: I formally define short-range template switch mutagenesis and explore an earlier algorithmic approach for capturing this class of variant [185]. I then detail robust statistical methodology based on pair hidden Markov models for capturing short-range template switch mutations. I describe how statistical significance can be assigned to individual events using a frequentist approach in which simulations can be performed to approximate the null hypothesis distribution of my test statistic ("LPR", see §2.3.2).

Chapter 3: I apply my methods to the genomes of great apes to identify and phylogenetically interpret the prevalence of short template switch mutations in hominid evolution. I also outline the genomic features, physical properties, and sequence characteristics associated with event

initiation.

Chapter 4: I develop a pipeline which incorporates my models and methods to identify template switches in 3202 human samples from the 1000 Genomes Project [293], and discuss the ability of my approach to detect template switch variants in these resequencing data. I then describe these events in terms of human population distribution and structure, assess evidence for activity of FoSTeS/MMBIR in generating these mutations, study genomic features associated with the events, and provide evidence of *de novo* template switch mutagenesis.

Chapter 5: I explore the somatic landscape of template switch mutations by applying my models to 2658 cancer genomes produced by the Pan-Cancer Analysis of Whole Genomes (PCAWG) study [44]. I explore the challenges of suitably parameterising my models to capture template switch mutations in cancer genomes. I then detail a conservatively-selected set of identified events stratified by tumour type, inspect associations with known mutational pathways and signatures, and assess multiple biological features of interest to identify possible associations.

Chapter 6: Conclusions and discussions of possible future directions.

1.3.2 Formatting notes

Throughout this thesis, the following text formatting is used:
black text: regular text content,
(digital copy only) red text: within-document hyperlinks to references, figures, etc.,
(digital copy only) blue text: external hyperlinks to web pages,
`monospaced text`: inline code or command line tools,
`monospaced text with a grey background`: command line tools with arguments explained.

Some analysis methods report E-values and p-values that are infinitesimally small, for example, $\leqslant 10^{-100}$. The precise values carry no meaningful interpretation in these cases. I therefore report ≈ 0 in place of any values $\leqslant 10^{-10}$.

1.3.3 Published work

My main PhD project (the subject of this thesis) focused on modelling short template switch mutations in human genomes, and the work outlined in Chapter 2 and Chapter 3 resulted in a first-author publication [305]. I was also involved in additional projects throughout my PhD

related to SARS-CoV-2 sequence analysis throughout the COVID-19 pandemic. Involvement in these projects resulted in co-authorship of the following peer-reviewed and preprint/equivalent publications:

De Maio N., Boulton W., Weilguny L., **Walker C. R.**, Turakhia Y., Corbett-Detig R., Goldman N. phastSim: efficient simulation of sequence evolution for pandemic-scale datasets. *bioRxiv*, https://doi.org/10.1101/2021.03.15.435416 (2021).

De Maio N., **Walker C. R.**, Turakhia Y., Lanfear R., Corbett-Detig R., Goldman N. Mutation rates and selection on synonymous mutations in SARS-CoV-2. *Genome Biology and Evolution* 13, evab087 (2021).

Turakhia Y., De Maio N., Thornlow B., Gozashti L., Lanfear R., **Walker C. R.**, Hinrichs A. S., Fernandes J. D., Borges R., Slodkowicz G., Weilguny L., Haussler D., Goldman N., Corbett-Detig R. Stability of SARS-CoV-2 phylogenies. *PLOS Genetics* 16, e1009175 (2020).

Dellicour S., Durkin K., Hong S. L., Vanmechelen B., Martí-Carreras J., Gill M. S., Meex C., Bontems S., André E., Gilbert M., **Walker C. R.**, De Maio N., Faria N. R., Hadfield J., Hayette M., Bours V., Wawina-Bokalanga T., Artesi M., Baele G., Maes P. A phylodynamic workflow to rapidly gain insights into the dispersal history and dynamics of SARS-CoV-2 lineages. *Molecular Biology and Evolution* 38, 1608–1613 (2020).

De Maio N., **Walker C. R.**, Borges R., Weilguny L., Slodkowicz G., Goldman N. Issues with SARS-CoV-2 sequencing data. *virological.org* (2020).

De Maio N., **Walker C. R.**, Borges R., Weilguny L., Slodkowicz G., Goldman N. Masking strategies for SARS-CoV-2 alignments. *virological.org* (2020).

Chapter 2

Modelling short template switch mutations

Chapter overview

I use this chapter to introduce the four-point model of template switching [185], and explore alignment models that can be used to capture and assign significance to short template switch mutations which have occurred since any pair of DNA sequences diverged from a common ancestor. I first provide details of an existing approach for this problem and explain how this model can be improved using pair hidden Markov models. I provide a description of how these pairHMMs can be parameterised based on the organisms under study, and detail a simulation procedure which facilitates the model comparisons that provide statistical assessment of template switches identified by my models.

Declaration

The content of this chapter was adapted and expanded from a first-author publication [305]:

Walker C. R., Scally A., De Maio N., Goldman N. Short-range template switching in great ape genomes explored using pair hidden Markov models. *PLOS Genetics* 17, e1009221 (2021).

For this paper, I developed the methods and implemented the models in C++. Additionally, I performed all data collection, processing, analysis, and data visualisation. I wrote the original manuscript, which was subsequently edited and agreed upon by all co-authors.

Code availability

The unidirectional and TSA pairHMMs described in this chapter are implemented in C++, and are available from: https://gitlab.com/conorwalker/phd_thesis/tree/main/chapter_2.

2.1 Background

2.1.1 Template switching underlies many genomic rearrangements

Replication-based rearrangement mechanisms (discussed in §1.2.4) are mediated by a template switch process, involving the dissociation of the $3'$ end of the nascent DNA strand and invasion of a physically-close alternate template. A period of replication using this alternate template is then followed by either a second switch event in which the $3'$ end of the nascent strand reassociates with the original strand [111], a series of successive switch events that can generate large-scale complex rearrangements [172], or extension of the alternate template until a new telomere is formed [277]. While all of these mechanisms require a physically proximal alternate template, there is no requirement that the two regions are nearby in linear sequence space and the position of strand invasion is often mediated solely by small stretches of identity between any two genomic positions regardless of proximity [37, 50, 120].

As discussed in §1.2.4 and §1.2.6, the consequences of these rearrangement pathways are typically investigated at the $\geqslant$ 50nt scale using the framework of structural variant calling [2, 60, 187, 284]. While this 50nt cutoff for defining structural variants is useful for calling rearrangmeents with confidence, it leaves the question unanswered of how prevalent short template switch mutations are in both within-species and between-species genome comparisons. These will leave a footprint of clusters of single nucleotide substitutions and/or indels within pairwise comparisons of related genomes, as variant callers and alignment algorithms have no other way to represent these regions. However, attributing a small number of mutations to a short alternate template from any position in a large genome is a computationally intractable problem, as candidate templates with high identity to the focal mutation cluster may readily be found by chance.

Instead, a subset of clustered mutations possibly generated through template switching can be modelled by restricting the search space of potential alternate template to regions in the vicinity (tens to hundreds of nucleotides) of each mutation cluster. By restricting the search space, there is little chance of finding near-perfect alternate templates nearby which could explain a mutation cluster by chance. Modelling replication-based rearrangements locally in this way should ultimately facilitate a greater understanding of the consequences of mutational processes typically only investigated at large scales. The remainder of this chapter will provide details on how this local model of template switching can be formalised and applied to DNA sequence data.

2.1.2 A "four-point" model suitably describes template switch mutations

Löytynoja and Goldman [185] outlined a mechanism-agnostic "four-point" model for describing short-range template switch events, the computational implementation of which leverages a modified dynamic programming approach to parsimoniously explain mutation clusters between closely related species. The four-point model (Figure 2.1) assumes switch events occur locally, likely within a single replication fork, and captures the consequences of both intra-strand (Figure 2.1, left) and inter-strand (Figure 2.1, right) switch events. There are no implicit

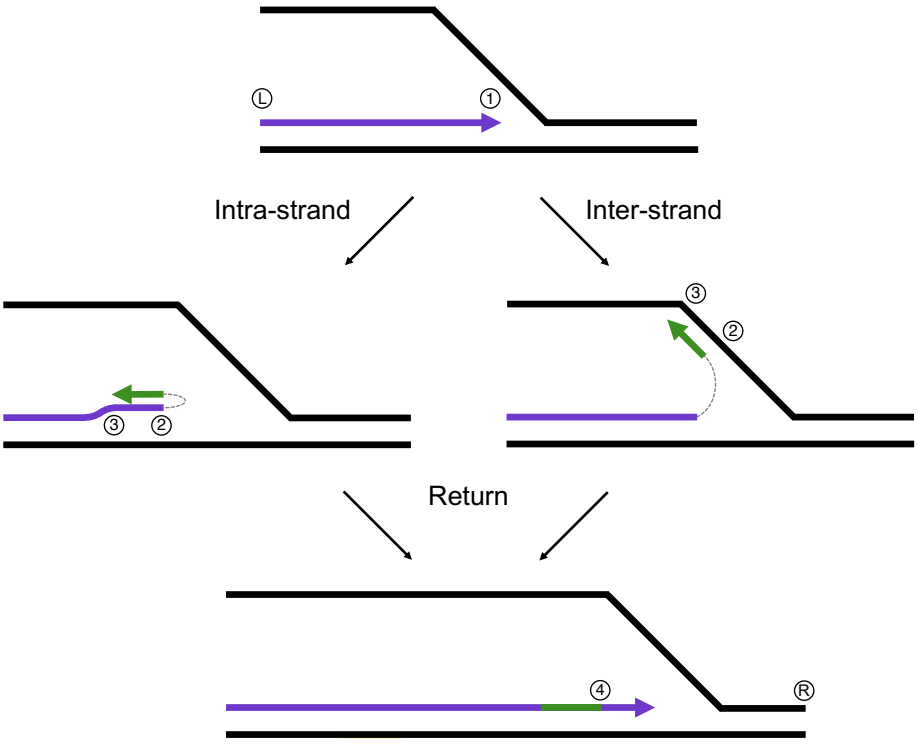

Figure 2.1: Diagrammatic representation of a short-range template switch. The template switch process projected onto a replication fork (note this is a simplified replication fork; see Figure 1.1 for a more detailed view of the replisome). DNA replication (arrow head) is shown proceeding in Ⓛ → Ⓡ orientation (Ⓛ and Ⓡ indicating the assumed direction of replication, not precise locations). A template switch event is initiated at ①; the DNA polymerase dissociates from the nascent strand and attaches at ② (left: intra-strand; right: inter-strand), and replication transiently proceeds in reverse orientation until ③. A second switch event occurs at ③, with the polymerase now detaching from the alternate template region (green lines) and reattaching at ④, from where replication proceeds as normal. This process generates three annotated fragments: the initial and final purple fragments represent the standard-replicated regions, and the central green fragment represents the reverse-replicated region from an alternate template.

assumptions made about the strandedness of events (the inter-strand switch in Figure 2.1 is depicted as a leading to lagging strand switch for simplicity), allowing the detection of switch events from either strand. Each event is described using four numbered points, assuming left (Ⓛ) to right (Ⓡ) oriented replication. Points ① and ② describe the genome coordinates of the initial switch event, with dissociation from the nascent strand at ① and strand invasion followed by alternate-template replication at ②. After this transient period of Ⓡ→Ⓛ-orientated replication from ②→③, a second switch event occurs, with dissociation at ③ and reassociation on the original strand at ④, after which replication proceeds as normal. This four-point notation provides a convenient way to represent the consequences of any single template switch event, regardless of the causative mechanism, enabling the definition of three ordered sequence fragments, Ⓛ→①, ②→③ and ④→Ⓡ, which fully describe any template switch event (e.g. Figure 2.2a).

For each template switch event, the linear ordering of the four numbered switch points ($\{①,\ldots,④\}$, referred to as an "event type", following [185]) facilitates the description of post-event rearrangement patterns and the inference of intra-strand and/or inter-strand switching. For example, the event type in Figure 2.2 is denoted ①-④-③-②, based on the linear ordering of the switch points projected onto the ancestral sequence. If the assumed ancestral sequence represents the true ancestral sequence state, it is possible to infer if an event could have arisen through intra-strand switching, inter-strand switching, or either. This follows the simple logic that for events to arise through intra-strand switching, point ② must precede point ① in the ancestral sequence; if instead ② is located ahead of point ① in linear sequence space, the necessary nascent strand has not yet been synthesised and cannot facilitate an intra-strand template switch.

The consequences of any such template switch process will present as a cluster or single nucleotide substitutions and/or indels in a typical pairwise alignment between two closely related sequences (Figure 2.2a, top), as standard alignment models assume that sequences evolve under single base substitutions and short indels and a combination of these processes is the only way in which the consequences of template switch events can be encoded. In contrast, a template switch alignment aims to model sequence evolution according to both substitutions and indels, as well as an additional single template switch event (Figure 2.2a, bottom). Assuming a template switch did indeed give rise to an apparent mutation cluster, the template switch alignment of this region will typically contain appreciably fewer substitutions and indels than the corresponding linear alignment.

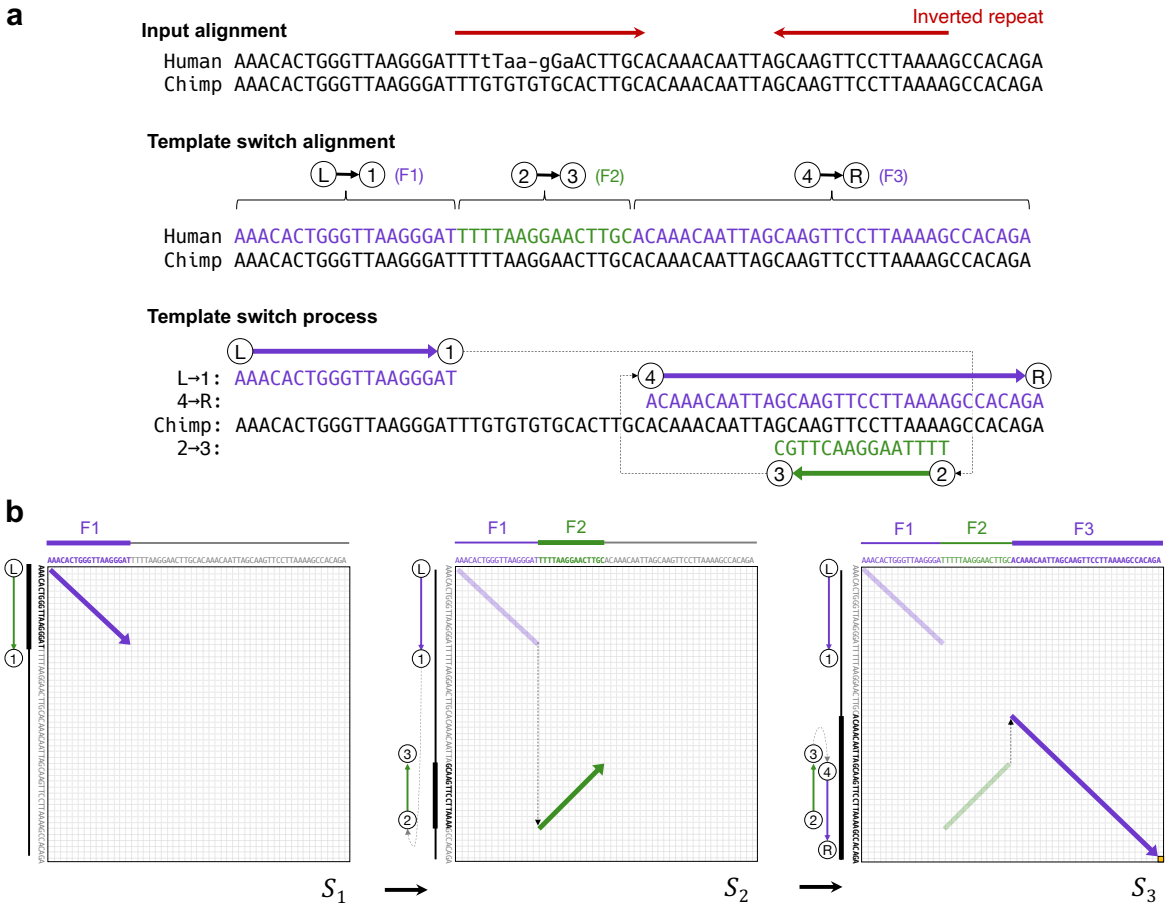

Figure 2.2: Example template switch event and linear-cost four-point alignment. (**a**) A mutation cluster containing five substitutions and a 1nt insertion (top) in the alignment between chr10:106,349,808-106,349,875 of the reference human genome and the chimpanzee genome (Ensembl v.98, EPO alignments of thirteen primates [322]). Under a model of template switching, this cluster can be explained with 100% identity by three ordered alignment fragments (middle). The sequence representation of the template switch process that generates the three fragments is shown (bottom), with purple and green sequences representing the descendant fragments and the black sequence representing the original strand. Note that the reverse-oriented replication that generates ② → ③ manifests as reverse complement sequence in the descendant with respect to the ancestral template, often generating perfect inverted repeats (red arrows above the EPO alignment). (**b**) The optimal path is found by choosing the set of moves within and between three score matrices S_1, S_2 and S_3 to maximise the alignment score. S_1 and S_3 are filled from top-left to bottom-right, and moves can be matches (diagonal moves) or indels (horizontal/vertical moves). S_2 is filled towards the top-right (because of the reversed direction of replication relative to the reference genome) and only complement matches are allowed. Jumps from S_1 to S_2 and S_2 to S_3 correspond to the template switching process. The optimal score is found in the bottom right corner of S_3 (gold square). The template switch alignment is found by back-tracking the optimal dynamic programming path, including jumps between matrices, producing three ordered alignment fragments, F1–3. Adapted from [185].

2.1.3 Template switch alignment using dynamic programming, a simple scoring scheme, and qualitative filtering

Löytynoja and Goldman [185] implement two models to produce pairwise alignments for input sequences x and y consisting of bases $x_1,...,x_i,...,x_n$ and $y_1,...,y_j,...,y_m$, respectively. Both models use the same simple linear-cost scoring scheme of +1 for matches, -1 for mismatches, and -2 for gaps.

The first model, used to assess mutation clusters as originating without any template switch, is a standard linear-cost Needleman-Wunsch algorithm [221] for pairwise global alignment as depicted in Figure 1.2. The second, used to assess mutation clusters as originating with a template switch, is a "four-point" dynamic programming algorithm characterised by three recurrence relations rather than one. An example four-point alignment under this algorithm is shown in Figure 2.2, where "four-point" refers to the location of the two switches between the three score matrices. In four-point alignment, each of the three recurrence relations is independently similar to the linear-cost alignment algorithms of Sankoff [264] and Needleman-Wunsch [221], involving additive calculation of alignment column scores. However, there are two key differences between the four-point alignment algorithm and typical pairwise aligners. First, the alignment path must start in the matrix defined by the first recursion (matrix S_1, Figure 2.2b) and finish in the matrix defined by the third recursion (matrix S_3, Figure 2.2c). This requires that the algorithm calculates not only the cost of matches, mismatches, or indels, but also the cost of jumping from S_1 when calculating matrix S_2, and the cost of jumping from S_2 when calculating matrix S_3. Second, matrix S_2 aligns backwards with respect to sequence y, only matches and mismatches are permitted, and matches are calculated using the complement of sequence y, capturing the period of reverse-orientation alternate-templated replication inherent to the short-range template switch process depicted in Figure 2.1. In both models, pointer matrices are used to trace back the highest scoring alignment path, including jumps between matrices for the four-point aligner.

To determine whether an evolutionary history involving a single template switch is significantly more likely than a combination of single base substitutions and/or indels, it is necessary to compare these two alignment models. Here, the linear alignment model represents the observed data by substitutions and indels, and the alternate template switch alignment model represents the data additionally by a single template switch. Testing the template switch alternate hypothesis requires a comparison of the optimal explanations from each of these models. This model comparison is not possible under the simple scoring scheme implemented by Löytynoja and Goldman [185], meaning the statistical significance of any particular event

cannot be established. As a replacement for statistical model comparisons, Löytynoja and Goldman required that a set of qualitative criteria must be satisfied for a template switch alignment to be labelled as convincingly more parsimonious than the corresponding linear alignment. Filtering criteria include requiring that: the region of ②→③ alignment is at least 14 nucleotides in length; 40nt upstream and downstream of the ②→③ region show $\geqslant 95\%$ sequence identity; the ②→③ region is not masked and contains all four nucleotides; and that the template switch alignment contains two fewer differences than the linear alignment, requiring that one of the linear alignment differences was a mismatch.

The problem with such a filtering approach (without a method for directly comparing the two alignment models) is that it can cause template switches, a mutation class which are *a priori* assumed to be rare, to be preferred over a small number of mismatches or indels in the linear alignment, which are far less rare. Indeed, Löytynoja and Goldman [185] identified 4.6×10^6 mismatch and indel positions in the unmasked EPO human-chimpanzee whole genome alignments, only 0.48% of which overlapped with a candidate template switch event. The only filter which acts in lieu of a model comparison for these candidate events is requiring that the template switch alignment contains two fewer mismatch/gap columns than the linear alignment, and in combination with the other filters, remaining events do generally appear visually convincing. However, two of the filters used to achieve this can be considered overly conservative: (1) a minimum ②→③ length of 14 nucleotides, and (2) requiring that the flanking sequences are near identical.

The length filter was established by scanning for local sequence matches around mutation clusters in reverse orientation, as opposed to the reverse complement orientation sequence created during ②→③ replication, and creating a length threshold at the point where the two distributions diverge. This was suitable to ensure only visually convincing candidate template switches remained in the final event set, but it came at the cost of removing all events with a short ②→③ fragment. Given that there are no estimates for the true length distribution of the ②→③ regions of short template switches, this approach may not capture the full spectrum of events which have shaped genome evolution, but is necessitated in the absence of a direct model comparison procedure.

The requirement of near-identical flanking sequence is generally reasonable, as it ensures that the only detected mutational mechanism which has impacted the realigned sequence was a single template switch which I hope to explain with an alternate model. Previously, however, it has been shown that the single-nucleotide substitution rate across a range of eukaryotic genomes increases as a function of proximity to indels [295]. As it is feasible that I may also observe elevated mutation rates within the regions flanking template switch mutations,

requiring near-identical flanking sequence can become restrictive if I want to maximise event discovery across a range of settings.

In summary, the filtering procedure of [185] aimed to produce a subset of high-confidence template switch events from the full set of four-point realignments of all small mutation clusters identified between the human and chimpanzee genomes. These filters can be regarded as conservative however, as both events with a short ②→③ region and events proximal to substitutions and indels are removed from any high-confidence set. Additionally, no direct statistical model comparison can be performed when relying on their filtering scheme, and the significance of each candidate event cannot be assessed. Regardless, the four-point pairwise alignment formalism provides an ideal method and notation for describing any single template switch mediated genomic rearrangement. The remainder of this chapter will therefore explore how the methods of [185] can be improved upon to more reliably identify template switches within pairwise DNA alignments, moving from an approach based on qualitative alignment filtering, to one based on statistical model comparison. It is worth noting that I do ultimately perform some filtering of events in subsequent chapters. These filters primarily serve to discard events found within low complexity genomic regions, but also to address a limitation of my final probabilistic model comparison which was not relevant to the models employed by [185]; see §2.3.4 for details.

2.2 Probabilistic alignment models for capturing template switch mutations

In this section, I will describe how a framework based on pair hidden Markov model (pairHMM) comparison can be used for modelling template switch mutations through pairwise alignment. Durbin *et al.* [77, p. 12] highlight three key considerations for the pairwise alignment problem: the scoring system used to rank the alignments, the algorithm used to find optimal alignments, and the statistical methods for evaluating the significance of alignment scores. To address each of these elements, first I will focus on establishing the pairHMMs for linear (§2.2.2) and template switch alignment (§2.2.3). Next, I will describe how transition (§2.2.4) and emission probabilities (§2.2.5) can be set in these models to score alignments. Then I will describe the traceback procedure used to find optimal alignments under each model (§2.2.6), and describe how boundaries are defined for each alignment to permit fair model comparisons (§2.2.7). Finally, I discuss how simulations of template switching can be used to establish significance thresholds for alignment model selection (§2.3.3).

2.2.1 Pair hidden Markov models: a brief overview

PairHMMs are probabilistic models that emit a pair of aligned sequences given two input sequences x and y which, in the context of DNA sequence alignment, consist of nucleotides $x_1,\ldots,x_i,\ldots,x_n$ and $y_1,\ldots,y_j,\ldots,y_m$ [77]. For each alignment column, the probability of emitting a particular pair of symbols is given based on the model state, determined at each column according to the distribution of transition probabilities in the previous state, and the emission probability distributions for that state of either a match/mismatch, or a gap in one sequence.

PairHMMs are specified by a set of hidden states $H = \{h_1,\ldots,h_N\}$, a transition probability matrix T with elements t_{ij} representing the probability of moving from state h_i to h_j, the two input sequences x and y, and emission probabilities $s_{h_i}(a,b)$ representing the probability of the pair $[a,b]$ being emitted from state h_i (where a and b indicate nucleotides from x and y, or gaps $-$). The symbol s is used to reflect that the logarithms of such values are often considered as emissions' (additive) scores. The transition probabilities t_{ij} for all pairHMMs (e.g. Figure 2.3) must satisfy

$$\sum_{j=1}^{N} t_{ij} = 1 \quad \forall i = 1,\ldots,N. \tag{2.1}$$

In a typical nucleotide sequence alignment, sequence homology under a pairHMM alignment is the alternate hypothesis, and the null hypothesis of no sequence homology may be rejected by comparing the global pairHMM alignment probability to that of a null alignment model in which the two sequences are emitted independently of each other [77]. In my case, the occurrence of a single template switch event is the alternative hypothesis, and the null hypothesis is that no template switch event was involved in the creation of the descendant sequence. The null hypothesis may be rejected by comparing the probability of an alignment generated under a model that emits linearly aligned sequences solely through substitutions and indels, to that of a model that emits an alignment consisting of substitutions, indels and a single template switch event (see §2.3.2).

Using the pairHMM framework, I implement two probabilistic models which facilitate this statistical testing of template switch event significance. First, for the null hypothesis, is a canonical three-state pairHMM for linear alignment (§2.2.2). The second is a seven-state pairHMM-like model for template switch alignment (described in §2.2.3).

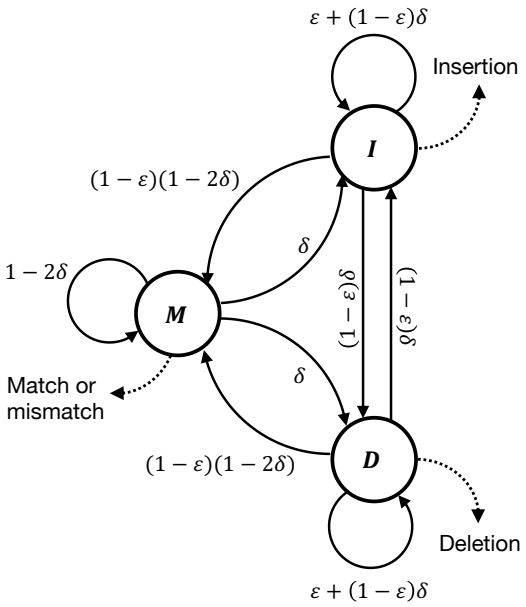

Figure 2.3: The unidirectional pairHMM. The model's three states, M, I and D, represent respectively match/mismatch, insertion and deletion alignment columns. A match/mismatch (M) column is one where both sequences have a non-gap character; an insertion (I) column has a gap character ($-$) in the sequence x; and a deletion column (D) has a gap character in sequence y. The pairHMM graph illustrates the probabilities that one type of column follows another in a pairwise alignment, with δ and ε representing gap opening and extension probabilities. For example, the directed edge from state M to state I, annotated with δ, denotes that the probability that an I column follows a M column is δ. Dashed arrows represent emissions (the observations of specific alignment columns given the corresponding state); for example, at an M column the two sequences can be either identical ("Match") or contain different nucleotides ("Mismatch"), and one nucleotide from each sequence is emitted in this case.

2.2.2 Unidirectional pair hidden Markov model structure

The first model, a three-state pairHMM, defines the probability of an alignment of two sequences that evolved undergoing only substitutions of individual nucleotides and indels. This is a standard approach for the probabilistic alignment of two biological sequences [77], and I refer to this as a unidirectional pairHMM (in contrast to the bidirectional nature of template switch alignment).

The unidirectional pairHMM (Figure 2.3) is of canonical form for pairwise alignment [77], composed of three hidden states: match (M), insertion (I) and deletion (D), giving $H = \{M, I, D\}$. M corresponds to the emission of a pair of nucleotides $[x_i, y_j]$; no gaps can be emitted. I emits a gap and a nucleotide $[-, y_j]$, and D emits a nucleotide and a gap $[x_i, -]$.

State transition probabilities T are specified using two parameters, δ and ε, where δ is the frequency of indel events expected along a pairwise alignment and ε controls their lengths. In Figure 2.3, states H are shown as nodes, non-zero elements of T are shown as directed edges (annotated with the values assigned to them in terms of probabilities δ and ε), and emissions with non-zero probabilities s are shown as annotated dashed arrows.

There are two things to note about this structure. First, the typical "Begin" and "End" states are omitted; this is because I will assume that all alignments begin in state M (for convenience) and end when a global alignment of the two input sequences has been achieved. Second, I permit transitions between insertion and deletion states. In most linear alignments of related sequences, a transition between indel states will occur with such a small probability that it is unlikely to be observed, and this transition is often not calculated in order to reduce the number of computations performed. For the types of mutational footprint left in linear alignments by template switches however, creating an insertion immediately followed by a deletion or vice versa may actually be the most appropriate way to unidirectionally align these regions. This can be interpreted as no evolutionary relationship existing between the two short sequence segments contained in this adjacent indel event, and will occur when the flanking sequences around such a region align with high identity. In these cases, the ordering of the apparent insertion-deletion event is also unimportant, so it does not matter if I appears before D in the state path (or vice versa).

2.2.3 Template switch alignment pair hidden Markov model structure

The second model is formulated similarly to a typical pairHMM (Figure 2.4); it consists of seven hidden states, each of which emits either a pair of aligned nucleotides or a nucleotide from one sequence and a gap from the other, and the probabilities of transitioning out of each state sum to 1 (satisfying Equation 2.1). Because this model is a compilation of three pairHMMs, with a period of reverse complement alignment in state M_2, and requires three combined recursions to fully decode the state path (see Algorithm 2.2), it cannot be considered a true pairHMM as classically defined by [77]. A more general description could perhaps be achieved by formulating our model as an alignment-constrained pair stochastic context-free grammar [Ian Holmes, personal communication, January 2021], such as those used for RNA gene structure and prediction [74, 127, 254]. However, given the similar statistical properties and convenient terminology provided, I opted to describe my model using a pairHMM formulation, and refer to this model as a template switch alignment pairHMM (TSA pairHMM).

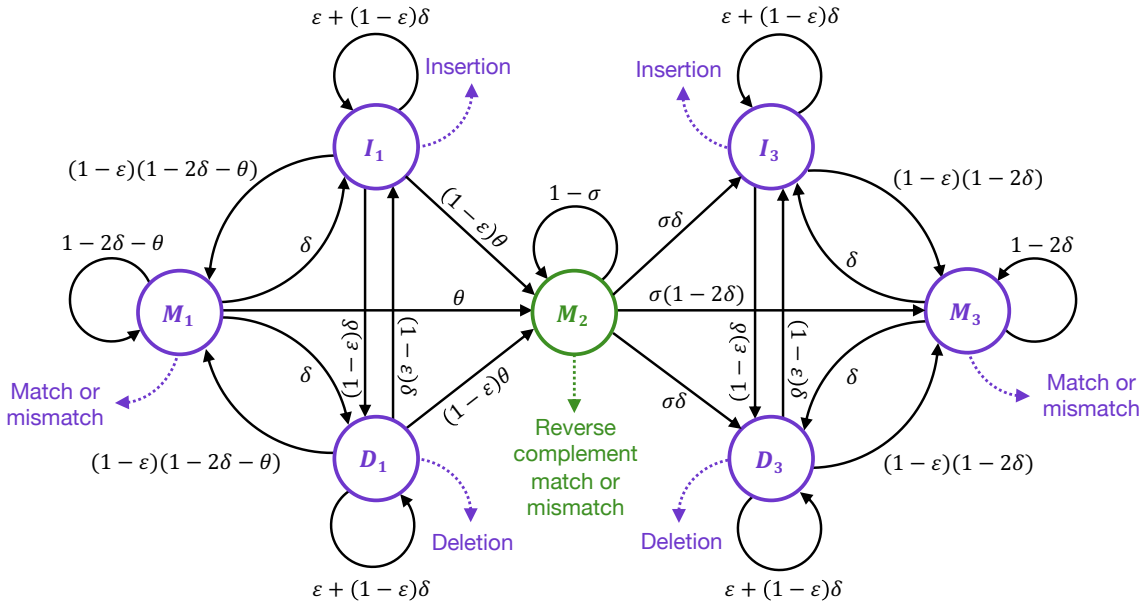

Figure 2.4: The template switch alignment pairHMM. States M_1, I_1, D_1 emit fragment Ⓛ→①; state M_2 emits fragment ②→③; and states M_3, D_3, and I_3 emit fragment ④→Ⓡ. Parameters θ and σ control the probabilities of template switch initialisation and extension, respectively. Purple states align forwards with respect to both sequences, whereas the green state aligns the two sequences in opposite directions. Emissions in state M_2 differ from M_1 and M_3 in that the emitted sequence respects the complementarity of the alternative template rather than a direct match between the two sequences at that position. Other parameters and annotations are as in Figure 2.3.

The seven hidden states ($H = \{M_1, I_1, D_1, M_2, M_3, I_3, D_3\}$) of the TSA pairHMM (Figure 2.4) are: M_1, D_1, and I_1 which emit alignment fragment Ⓛ→①, M_2 which emits fragment ②→③, and M_3, D_3, and I_3 which emit fragment ④→Ⓡ. As with the four-point aligner detailed in §2.1.2, the model is structured to capture a single template switch event per alignment by requiring a single transition into M_2 from $\{M_1, I_1, D_1\}$ (at ①, ②), and a single transition from M_2 into $\{M_3, I_3, D_3\}$ (at ③, ④). Similarly to the dynamic programming model of [185], state M_2 differs from typical pairwise aligners in that the descendant sequence y is aligned in complement and reverse orientation with respect to the ancestral sequence x, capturing the period of alternate strand-templated replication inherent to the template switch process. State transition probabilities T satisfy Equation 2.1, and are defined using the parameters δ and ε from the unidirectional pairHMM and two additional parameters: θ, the probability of initiating a template switch event, and σ, which controls the expected length of the ②→③ fragment.

As with the unidirectional pairHMM, I omit "Begin" and "End" states. I also permit transitions between indel states (between I_1 and D_1, and between I_3 and D_3) as with the unidirectional pairHMM. Additionally, I allow transitions from $\{I_1, D_1\}$ to M_2, and from M_2 to $\{I_3, D_3\}$. In all subsequent analyses, I do not observe any significant events with an optimal state path (as calculated using Algorithm 2.2) involving either an indel to template switch transition, or a template switch to indel transition, but I included their consideration for increased model flexibility.

2.2.4 Transition probabilities

In this and the next subsection I will describe how the transition probabilities are determined for use in my models. All transition probabilities are defined in terms of other parameters which are easy to estimate based on the organisms/sequences under study. All analysis-specific parameter values ($\{t, \rho, \lambda, N, C, A\}$; see below) will be detailed in the following chapters as appropriate.

In both models, I set the parameters controlling transition probabilities to $\delta = 1 - e^{-t(\rho/2)}$ and $\varepsilon = 1 - 1/\lambda$, where t is the pairwise divergence measured in expected substitutions per site, ρ is the expected number of indel events per substitution, and λ is the expected indel length. This corresponds to a Poisson process of indel formation, in which δ defines the probability of observing at least one mutation that is an indel after time t, divided by 2 to account for both insertions and deletions. Indel lengths are then defined by ε to have a geometric distribution, as this provides a reasonable enough fit to indel lengths observed in real data whilst allowing efficient use of dynamic programming alignment algorithms [69, 128, 255]. It is worth noting for Chapter 3 however that a zeta power-law model provides a slightly better description of observed great ape indel lengths [46].

For the TSA pairHMM, I set $\theta = N/CA$ where N is the expected number of template switch events in a given pairwise comparison, C is the total count of mutation clusters identified in each pairwise comparison, and A is the event-specific alignment length (determined using the procedure detailed in §2.2.7). This corresponds to the probability of initiating a template switch, normalised by the total number of alignment positions considered by the TSA pairHMM across the entire pairwise whole genome comparison. I set $\sigma = 1/L$, where L is the expected ② → ③ length, which has the effect of making ② → ③ lengths in the model match my observations (see also Figure 2.5).

For each pairwise comparison, C is calculated from the data; for example, by counting the number of observed mutation clusters in the whole genome alignment of human and chim-

Figure 2.5: Example of an event which is significant and passes all filters when using a smaller value of σ. For the chosen value of σ used in subsequent chapters (0.1, top), and a nominal small value of sigma ($\sigma = \delta = 0.001$, bottom), an event is shown which was detected in the pairwise alignments of human and chimpanzee, and gorilla and chimpanzee, in which the chimpanzee sequence was assigned to the descendant state. When using $\sigma = 0.1$, this event does not contain all four nucleotides in the ② → ③ fragment, and fails the corresponding filter. If M_2 extension is penalised less heavily, by setting $\sigma = \delta$, a longer period of ② → ③ alignment is included in the state path during Viterbi decoding, including all four nucleotides and allowing the event to be called as significant. Note that "Anc" refers to the assumed ancestral sequence and "AncC" refers to the complement of this sequence.

panzee, where a mutation cluster is defined using the procedure outlined in §2.2.7. Similarly, A can be calculated for each pairwise alignment under consideration, and simply normalises for alignment length.

Determining values for the variables N and L requires careful consideration, as unlike the values used to define indel formation, the frequency and length distribution of template switch events has not previously been well studied to provide reasonable prior estimates. I define these parameters differently for the subsequent great ape (Chapter 3) and human population/cancer

analyses (Chapter 4 and Chapter 5). For example, for the great ape comparisons, I set N to 2750 and L to 10, based on the average number of significant events found in earlier pairwise great ape comparisons and the ② → ③ length distribution of these events. These significant events were determined using an earlier version of the TSA pairHMM which did not include parameters θ or σ, simply treating transitions into and out of M_2, and emissions from M_2, as equiprobable to emissions from and transitions into/out of M_1 and M_3.

The precise value of N used likely has little impact however: because the product CA is large, θ will always correspond to a small initiation penalty for any reasonable value of N. In contrast, the value of σ can have a more substantial effect, as this parameter controls the expected length of the ② → ③ fragment. Lower values of σ lead to longer ② → ③ fragments being preferred, possibly causing some events to pass (e.g.) the 'all four nucleotides present' filter (used to assess sequence complexity, see §2.3.4 and Figure 2.5). While this may produce additional significant template switch alignments that appear convincing, I prefer to use the more natural formulation for this parameter $1/L$ in pursuit of quality of inferred events over quantity.

2.2.5 Emission probabilities

Probabilistic models of nucleotide evolution from which emission probabilities can be derived are defined using instantaneous rate matrices, conventionally denoted Q, and describe the rate of change between each unique nucleotide through time. Off-diagonal elements of Q matrices define the rate of replacement between nucleotides i and j, and the diagonal elements are chosen such that the rows all sum to zero [182]. Exponentiation of Qt produces a second matrix, conventionally denoted P. The P matrix contains the probabilities of each nucleotide i changing to each other nucleotide j after time t, where all rows must sum to 1. Many nucleotide substitution models which define these matrices have been proposed, the earliest of which was outlined by Jukes and Cantor in 1969 [138], and is commonly referred to as JC69. Under JC69, the instantaneous rate of all nucleotide changes are considered equal, specified by a nucleotide-indexed matrix

$$Q = \begin{array}{c} \begin{matrix} \mathrm{T} & \mathrm{C} & \mathrm{A} & \mathrm{G} \end{matrix} \\ \begin{bmatrix} -\frac{3}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & -\frac{3}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & -\frac{3}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & -\frac{3}{4} \end{bmatrix} \end{array} \begin{matrix} \mathrm{T} \\ \mathrm{C} \\ \mathrm{A} \\ \mathrm{G} \end{matrix} . \tag{2.2}$$

Exponentiation of Qt gives

$$P(t) = e^{Qt} = \begin{array}{c} \begin{matrix} \text{T} & \text{C} & \text{A} & \text{G} \end{matrix} \\ \begin{bmatrix} \frac{1}{4}+\frac{3}{4}e^{-t} & \frac{1}{4}-\frac{1}{4}e^{-t} & \frac{1}{4}-\frac{1}{4}e^{-t} & \frac{1}{4}-\frac{1}{4}e^{-t} \\ \frac{1}{4}-\frac{1}{4}e^{-t} & \frac{1}{4}+\frac{3}{4}e^{-t} & \frac{1}{4}-\frac{1}{4}e^{-t} & \frac{1}{4}-\frac{1}{4}e^{-t} \\ \frac{1}{4}-\frac{1}{4}e^{-t} & \frac{1}{4}-\frac{1}{4}e^{-t} & \frac{1}{4}+\frac{3}{4}e^{-t} & \frac{1}{4}-\frac{1}{4}e^{-t} \\ \frac{1}{4}-\frac{1}{4}e^{-t} & \frac{1}{4}-\frac{1}{4}e^{-t} & \frac{1}{4}-\frac{1}{4}e^{-t} & \frac{1}{4}+\frac{3}{4}e^{-t} \end{bmatrix} \end{array} \begin{matrix} \text{T} \\ \text{C} \\ \text{A} \\ \text{G} \end{matrix} \tag{2.3}$$

which can be rewritten as the well-known probabilistic scoring function with two probabilities

$$s_M(x_i, y_j) = \begin{cases} \frac{1}{4}+\frac{3}{4}e^{-t} & \text{if } x_i = y_j \\ \frac{1}{4}-\frac{1}{4}e^{-t} & \text{otherwise,} \end{cases} \tag{2.4}$$

where time t is specified by the expected number of substitutions per site between the input pair of sequences. As all nucleotides are considered to occur at equal frequency in this model, inserting or deleting any particular nucleotide assumed to occur with equal probability, giving $s_I = s_D = \frac{1}{4}$.

The assumptions of JC69 are often considered too simplistic for modelling sequence evolution. Sequenced genomes tend to not contain equal amounts of each nucleotide, transitions do not occur at the same rate as transversions, and mutation rates are not uniform across the genome. Subsequent nucleotide substitution models have expanded beyond the equal rate, equiprobable assumptions of JC69. Some noteworthy models include the two-parameter K80 model [149], which distinguishes between the rates of transitions and transversions; the four-parameter F81 model [87], which extends the JC69 model by specifying a parameter for the equilibrium frequency of each nucleotide; the five-parameter HKY85 model [119], which combines the concepts of K80 and F81 to both distinguish between transitions and transversions and allow uneqal nucleotide frequencies; and the ten-parameter GTR model [292], which specifies the rate of exchange between all nucleotide pairs as well as their individual frequencies.

There are more complex models still, which incorporate substitution rate heterogeneity across sites in the alignment. These include the discrete gamma model [317], in which sites are assigned to one of several rate categories to approximate the continuous gamma distribution, and the gamma model with a proportion of invariable sites incorporated [112].

Parameter-rich representations of typical sequence evolution, as measured by substitutions through time, are typically preferred in molecular evolution analyses. For example, phyloge-

netic inference traditionally involves computationally expensive model selection procedures, but accurately inferring tree topology and branch lengths can often be achieved simply by choosing the most parameter-rich nucleotide substitution model, GTR+I+G [1] (where I and G respectively refer to a proportion of invariable sites and a discrete gamma model, as described above). The operative phrase here is "typical sequence evolution": I cannot assume that inferring template switch alignments under these models will also benefit from the most parameter-rich model, as it is a process which definitely does not meet the definition of typical here.

Template switches introduce multiple substitutions and/or indels into the linear alignment representation per mutation event, and the nucleotide composition of the introduced "substitutions" during ② → ③ replication is determined solely by the nucleotide composition of whichever local strand acts as a reverse complement template. In some cases, I expect that the substitutions introduced by template switching could also be systematically biased by local sequence composition, such as events involving the formation of stable hairpins in the DNA secondary structure through quasipalindrome to palindrome conversion [253]. These events may involve a skewed substitution rate caused by the differences between nucleotide frequencies present in local palindromic sequences compared to a random genomic background. My general model does not exclusively capture these cases however, meaning my choice in substitution model should not be exclusively fit to the assumptions of quasipalindrome conversion.

In light of this uncertainty, I opt to use the simplest set of emission probabilities permitted in an evolutionary context, as defined by JC69, with one key exception. In both models, I set s_M according to Equation 2.4, and $s_I = \frac{1}{4}$, but set $s_D = 1$. Note that s_I is the probability of any particular nucleotide conditional on being in states $\{I_1, I_3\}$ (and that all nucleotides are equiprobable under JC69), whereas s_D is the probability of an observed gap character conditional on being in states $\{D_1, D_3\}$, which I set to necessarily be 1.

A deletion emission probability of 1 can be interpreted as not penalising the "content" of deletions and conditioning the emission probabilities only on the ancestral sequence (i.e. all deletions are equally likely, regardless of the deleted nucleotide). This is useful when scoring large deletions, as my model comparison approach cannot distinguish between the probabilities associated with true events, and the probabilities of single, large deletions in the unidirectional alignment, which are alternatively explained as small (e.g. 4nt) ② → ③ template switch events (see §2.3.4 and Figure 2.6 for further details).

Figure 2.6: The impact of deletion emission probabilities at $s_D = 0.25$ and $s_D = 1$. From top to bottom: unidirectional pairHMM alignments of a mutation cluster identified between the human and chimpanzee genomes are shown with s_D set to 0.25 (discussed in §2.2.5), then set to 1. The TSA pairHMM alignment and inferred template switch process are shown for this cluster, depicted once as each is identical under both s_D values. The log-probability and a count of non-match columns are indicated above and below each alignment, respectively. Note that although the columns emitted from each unidirectional pairHMM alignment are identical, there is a large difference in log-probability between the template switch alignment and the $s_D = 0.25$ alignment, and a much smaller difference with $s_D = 1$. Setting the deletion emission probability to $s_D = 1$ is adequate to handle such spurious events, as it ensures deletion extension is not penalised in the unidirectional alignment, causing E_U (Algorithm 2.1) to be conditional on the ancestral sequence, making these cases less likely to be significant at the 5% level under my subsequently established null Monte Carlo LPR distribution (see §2.3.2 and §2.3.3).

2.2.6 Finding optimal alignments under each pairHMM

There are several options for recovering alignments from pairHMMs. Here, I describe an approach for finding the single optimal alignment from each model, how this permits model comparison, and then I briefly discuss an alternative approach which I assert is not necessary in the case of template switch inference.

The Viterbi algorithm [90] calculates the most probable state path through a pairHMM, outputting the corresponding optimal single pairwise alignment between the two input sequences. Typically, probabilities are converted into log-space both for convenience and to prevent underflow errors. As a result, instead of multiplying the set of transition and emission probabilities (depicted in Figure 2.3 and Figure 2.4) which produce most probable state path,

the logarithms of these probabilities are instead summed. The full logarithmic Viterbi algorithm for the unidirectional pairHMM is shown in Algorithm 2.1.

As commented on in §2.2.3, the TSA pairHMM is "pairHMM-like". It consists of three separate pairHMMs, each of which requires its own set of recurrence relations to decode the state path, which are calculated sequentially. This allows calculations in state M_2 to make use of alignment probabilities from M_1, I_1, D_1, and calculations in M_3, I_3, D_3 to use probabilities from M_2. This property does not mean I cannot make use of the Viterbi algorithm for calculating the optimal TSA pairHMM state path through each pairHMM independently, but it does require me to denote any use of this procedure presented as a single algorithm as "Viterbi-like" to be

---

Algorithm 2.1: Viterbi algorithm for the unidirectional pairHMM. Given two sequences x and y of lengths n and m, respectively, I find their alignment with the highest probability using the following dynamic programming procedure. I represent the i-th entry of sequence x as x_i, and the j-th entry of sequence y as y_j. To facilitate traceback after estimating the highest probability state path, for each cell $M(i,j)$, $I(i,j)$, and $D(i,j)$, pointer matrices are used to store the moves back to the previous cell from which each cell was derived. After the termination step, the most probable alignment is recovered using the moves stored in these traceback matrices. Note that $\bullet$ indicates an index i or j ranging over all possible values from 0 to n or m, as appropriate.

Initialisation:

$$M(\bullet,0) = I(\bullet,0) = D(\bullet,0) = M(0,\bullet) = I(0,\bullet) = D(0,\bullet) = -\infty$$
$$M(0,0) = 0, \quad I(0,0) = D(0,0) = \log(0.25)$$

Recursion:

$i = 1,\ldots,n,\ j = 1,\ldots,m$:

$$M(i,j) = \log(s_M(x_i,y_j)) + \max\begin{cases} M(i-1,j-1) + \log(1-2\delta) \\ I(i-1,j-1) + \log((1-\varepsilon)(1-2\delta)) \\ D(i-1,j-1) + \log((1-\varepsilon)(1-2\delta)) \end{cases}$$

$$I(i,j) = \log(s_I) + \max\begin{cases} M(i-1,j) + \log(\delta) \\ I(i-1,j) + \log(\varepsilon + (1-\varepsilon)\delta) \\ D(i-1,j) + \log((1-\varepsilon)\delta) \end{cases}$$

$$D(i,j) = \log(s_D) + \max\begin{cases} M(i,j-1) + \log(\delta) \\ I(i,j-1) + \log((1-\varepsilon)\delta) \\ D(i,j-1) + \log(\varepsilon + (1-\varepsilon)\delta) \end{cases}$$

Termination:

$$E_U = \max(M(n,m), I(n,m), D(n,m))$$

---

consistent with my notation. With this in mind, the full Viterbi-like algorithm for finding the optimal template switch alignment is shown in Algorithm 2.2. Note that Algorithm 2.1 and Algorithm 2.2 are the fundamental methods used for the majority of template switch analyses described throughout this thesis.

The alternate approach to assessing the single optimal alignment from each model is to assess the probability that sequences x and y are related by some undefined alignment rather than unrelated, evaluated by summing over all alignments in the dynamic programming matrices. This is achieved by replacing all max terms in the Viterbi algorithm with summations, and is known as the forward algorithm (see [77] for further details). This approach is useful when concerned about the "accuracy" of the Viterbi path, as there may exist many highly similar alignments with only slightly less probable state paths. In the case of template switch alignment, I am primarily interested in events for which placement of the ② → ③ region can be unambiguously derived from the state path. I additionally want to focus on being able to visually interpret all detected events, which is greatly aided by the Viterbi path. As a result, I do not implement the forward algorithm for the TSA pairHMM, accepting that this could cause a loss of power when inferring candidate template switches with ambiguous switch point coordinates. However, my subsequent focus on establishing and applying stringent probabilistic thresholds to candidate events supports the use of Viterbi/Viterbi-like algorithms here.

2.2.7 Defining alignment boundaries to facilitate model comparison

Before considering how the optimal alignments produced by Algorithm 2.1 and Algorithm 2.2 can be compared, I need to first define how alignment boundaries are determined. The total amount of sequence given as input to the pairHMMs from from each (x,y) pair will determine the final number of alignment columns contributing to the total probability of each alignment. Careful consideration must therefore be given to ensure fair model comparison.

Input to both pairHMMs is determined by scanning a pre-existing pairwise alignment from left to right for clusters of mutations, defined as $\geqslant$2 pairwise differences within a 10nt sliding window (as in [185]). Once $\geqslant$2 pairwise differences are identified, an iterative procedure is initiated which extends the rightmost cluster boundary while additional pairwise differences are present. A 10nt region downstream of the current boundary is searched for additional differences; if any are found, the cluster boundary is updated using the position of the rightmost difference. This procedure is repeated until no additional differences are found, defining one focal mutation cluster per candidate template switch event (red/yellow sequence in Figure 2.7a).

Algorithm 2.2: Viterbi-like algorithm for the TSA pairHMM. As in the unidirectional pairHMM, given two sequences x and y of lengths n and m, respectively, I find their alignment with the highest probability using the following dynamic programming procedure. As described in §2.2.7 (and depicted in Figure 2.7a,c), $n > m$ for the TSA pairHMM, and Viterbi-like decoding must include at least one M_2 state in the state path. Traceback is facilitated using pointer matrices as above, with moves from $\{M_1,I_1,D_1\}$ to M_2 and from M_2 to $\{M_3,I_3,D_3\}$ also stored as pointers whenever a jump between these matrices produces a more probable move in the state path. Again, $\bullet$ indicates an index i or j ranging over all possible values from 0 to n or m, as appropriate.

Initialisation:

$M_1(\bullet,0) = I_1(\bullet,0) = D_1(\bullet,0) = M_1(0,\bullet) = I_1(0,\bullet) = D_1(0,\bullet) = -\infty$
$M_2(n+1,\bullet) = M_2(\bullet,0) = -\infty$
$M_3(\bullet,0) = I_3(\bullet,0) = D_3(\bullet,0) = M_3(0,\bullet) = I_3(0,\bullet) = D_3(0,\bullet) = -\infty$
$M_1(l,0) = 0, \quad I_1(l,0) = D_1(l,0) = \log(0.25)$

Recursion 1:
Find the optimal alignment of fragment Ⓛ→① by aligning x and y linearly:

$i = 1,\ldots,n,\ j = 1,\ldots,m$:

$$M_1(i,j) = \log(s_M(x_i,y_j)) + \max\begin{cases} M_1(i-1,j-1) + \log(1-2\delta-\theta) \\ I_1(i-1,j-1) + \log((1-\varepsilon)(1-2\delta-\theta)) \\ D_1(i-1,j-1) + \log((1-\varepsilon)(1-2\delta-\theta)) \end{cases}$$

$$I_1(i,j) = \log(s_I) + \max\begin{cases} M_1(i-1,j) + \log(\delta) \\ I_1(i-1,j) + \log(\varepsilon + (1-\varepsilon)\delta) \\ D_1(i-1,j) + \log((1-\varepsilon)\delta) \end{cases}$$

$$D_1(i,j) = \log(s_D) + \max\begin{cases} M_1(i,j-1) + \log(\delta) \\ I_1(i,j-1) + \log((1-\varepsilon)\delta) \\ D_1(i,j-1) + \log(\varepsilon + (1-\varepsilon)\delta) \end{cases}$$

Recursion 2:
Find the optimal alignment of fragment ②→③ by emitting y in reverse complement with respect to x, determining the best position to jump from M_1, I_1, or D_1 with c_i:

$i = 1,\ldots,n$:

$$c_i = \max\begin{cases} \max(M_1(i-1,\bullet)) + \log(\theta) \\ \max(I_1(i-1,\bullet)) + \log((1-\varepsilon)\theta) \\ \max(D_1(i-1,\bullet)) + \log((1-\varepsilon)\theta) \end{cases}$$

$j = m,\ldots,1$:

$$M_2(i,j) = \max\begin{cases} c_i + \log(s_M(x_i, comp(y_j))) \\ M_2(i-1,j+1) + \log(1-\sigma) + \log(s_M(x_i, comp(y_j))) \end{cases}$$

Recursion 3:
Find the optimal alignment of fragment ④→Ⓡ by emitting x and y linearly, determining the best position to jump from M_2 with k_i:

$i = 1,\ldots,n$:
$k_i = \max(M_2(i-1,\bullet))$
$j = 1,\ldots,m$:

$$M_3(i,j) = \log(s_M(x_i,y_j)) + \max\begin{cases} k_i + \log(\sigma(1-2\delta)) \\ M_3(i-1,j-1) + \log(1-2\delta) \\ I_3(i-1,j-1) + \log((1-\varepsilon)(1-2\delta)) \\ D_3(i-1,j-1) + \log((1-\varepsilon)(1-2\delta)) \end{cases}$$

$$I_3(i,j) = \log(s_I) + \max\begin{cases} k_i + \log(\sigma\delta) \\ M_3(i-1,j) + \log(\delta) \\ I_3(i-1,j) + \log(\varepsilon + (1-\varepsilon)\delta) \\ D_3(i-1,j) + \log((1-\varepsilon)\delta) \end{cases}$$

$$D_3(i,j) = \log(s_D) + \max\begin{cases} k_i + \log(\sigma\delta) \\ M_3(i,j-1) + \log(\delta) \\ I_3(i,j-1) + \log((1-\varepsilon)\delta) \\ D_3(i,j-1) + \log(\varepsilon + (1-\varepsilon)\delta) \end{cases}$$

Termination:

$$E_{TS} = \max(M_3(\bullet,m), I_3(\bullet,m), D_3(\bullet,m))$$

With the focal mutation cluster coordinate boundaries established, I define the total input pairwise alignment region to be realigned under my models separately for the unidirectional pairHMM and the TSA pairHMM. The unidirectional pairHMM takes as inputs x and y the sequences contained within the pairwise alignment defined by the above cluster boundaries plus a ±40nt flanking region from each sequence (Figure 2.7a,b). (The -40 position, representing the leftmost alignment boundary for the unidirectional pairHMM, is referred to as l below and in Algorithm 2.2.) This flanking region provides sufficient alignment space to interpret the mutational footprint of a putative template switch event within the context of neutrally evolving sequence that should contain few or no other differences. This has the effect of anchoring the alignment of the mutation cluster and ensures that no other locations get included in the alignments' explanation of the cluster. In contrast, the TSA pairHMM realigns this same region but includes an additional ±100nt from the assumed ancestral sequence (x), to provide (additional, local) flanking search space for the ②→③ sequence fragment (Figure 2.7a,c).

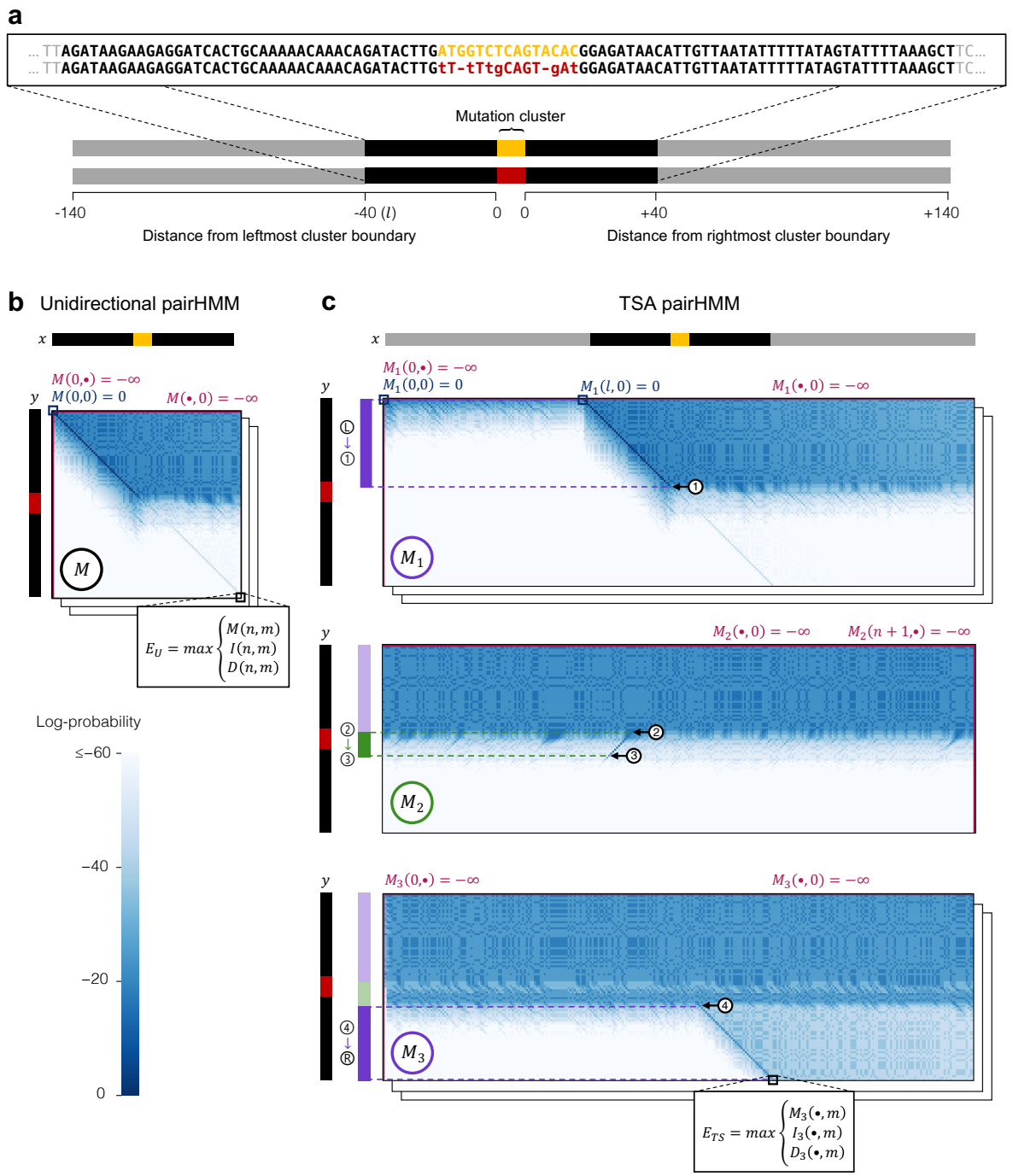

Figure 2.7: Diagrammatic overview of how sequence regions are aligned under each pairHMM. (**a**) Defining alignment boundaries for a focal mutation cluster. The focal cluster between an (x, y) pair is shown in red/yellow. In addition to the cluster, the region shown in black is used for unidirectional alignment, and the regions shown in both black and grey are used for template switch alignment (an additional 100nt both up- and downstream). (**b**) Unidirectional alignment follows Algorithm 2.1: the figure illustrates initialisation and subsequent calculation of the M matrix, simply depicting I and D matrices as white boxes hidden behind the M matrices for simplicity. (**c**) Template switch alignment follows Algorithm 2.2. For clarity, initialisations and recursive calculations are only illustrated for match state ($M_{\{1,2,3\}}$) matrices.

To make a fair comparison of the alignments emitted by each pairHMM, despite their using these two different length ancestral sequences, it is necessary to constrain the start and end positions of the TSA pairHMM alignments to match those of the unidirectional pairHMM. This ensures that the flanking region alignments are identical between the two models, and therefore contribute the same scores to each alignment. The score difference between the two models is then derived solely from the contributions of either a linearly aligned mutation cluster, or a region of reverse-orientation template switch alignment. To impose this constraint on the start position of the TSA pairHMM, I initialise matrices M_1 to 0, and I_1 and D_1 to log(0.25), at positions corresponding to y_0 (i.e. cells indexed $(l,0)$ in Algorithm 2.2). This causes all possible alignments of upstream flanking regions to have low probability, and the Viterbi-like decoding of the optimal state path should always lead back to $(l,0)$ in M_1, I_1 or D_1, facilitating score comparison between the two pairHMMs. To constrain the end TSA position, I require the Viterbi-like decoding of the TSA pairHMM state path to begin at the highest scoring alignment position for y_m (see the Termination computation E_{TS} in Algorithm 2.2).

2.3 Statistical testing and event filtering

From the Viterbi (Algorithm 2.1) and Viterbi-like (Algorithm 2.2) algorithms, a single log-probability corresponding to the highest-probability global alignment of the two input sequences under each model is captured in the termination variables E_U and E_{TS}, respectively (where subscript U and TS respectively denote unidirectional pairHMM and TSA pairHMM). It is therefore possible to compare these two variables and create a test statistic which assesses the goodness of fit of each of the pairHMMs given the observed sequence x and y. There are three major frameworks for comparing the goodness of fit of two competing models: Bayesian, information-theoretic, and frequentist. I will briefly consider possible test statistics under the two former paradigms, then outline the model comparison procedure used in all subsequent analyses, which is based on likelihood-ratio testing of non-nested models. Throughout, I will refer to the unidirectional pairHMM hypothesis as U, and the TSA pairHMM hypothesis as TS.

2.3.1 Approaches for model selection

Bayesian model comparison has been applied to assessing the goodness of fit of an alignment under a canonical three-state pairHMM to a random alignment model in which the sequences are emitted independently [77]. In my case, the Bayesian approach would assume that the sequences x and y arose under one of the competing hypotheses U or TS, with probabilities $P(x,y|U)$

and $P(x,y|TS)$, respectively where the logarithms of these calculated values are captured respectively in E_U and E_{TS}. These hypotheses would have associated prior probabilities of $P(U)$ and $P(TS) = 1 - P(U)$, and observing the sequence pair (x,y) allows the calculation of posterior probabilities $P(U|x,y)$ and $P(TS|x,y)$. The odds ratio of the posterior probabilities is then the statistic of interest, as $P(U|x,y)$ corresponds to the probability that x and y are related by only substitutions and indels, whereas $P(TS|x,y)$ is the probability that x and y are additionally related by a single template switch. The posterior odds ratio is calculated as

$$\underbrace{\frac{P(TS|x,y)}{P(U|x,y)}}_{\text{Posterior odds}} = \underbrace{\frac{P(x,y|TS)}{P(x,y|U)}}_{\text{Bayes factor}} \times \underbrace{\frac{P(TS)}{P(U)}}_{\text{Prior odds}} . \tag{2.5}$$

The value of Equation 2.5 is interpreted as the amount of evidence in support TS over U, with more positive values indicating a greater level of support for TS. For example, a posterior odds of 4 indicates that the alignment of sequences x and y is 4 times more likely to be derived from TS than from U. A qualitative label indicating the strength of evidence for each hypothesis under various values of Equation 2.5 is often additionally assigned ("substantial", "strong"' etc.) [144]. Alternatively, the posterior odds of an alignment can be interpreted as a probability by using the logistic function $\sigma(z) = e^z/(1+e^z)$ (where z is the posterior odds) which tends to 1 as z tends to infinity [77, p. 37]. The problem with any Bayesian model comparison however is that it requires selected a suitable prior, and I want to avoid making *a priori* assumptions about template switch event prevalence where possible.

Information-theoretic model comparison foregoes the need to select priors; instead it involves calculating and comparing an information theory-based performance metric for each model, taking into account the number of model parameters k. The Akaike information criterion (AIC) [5] is widely used for this purpose, and would be calculated here for each alignment model as

$$\begin{aligned} \text{AIC}_{TS} &= 2k_{TS} - 2E_{TS}, \\ \text{AIC}_{U} &= 2k_{U} - 2E_{U}. \end{aligned} \tag{2.6}$$

Assessing which competing model provides a better fit to the observed sequences x and y is performed simply by calculating the difference between the AIC of the two models [39]. Here, increasingly negative values of $\text{AIC}_{TS} - \text{AIC}_{U}$ would indicate increasingly stronger support for TS. However, AIC differences are difficult to interpret, do not permit an assessment of statistical significance for candidate template switches, and are instead contingent on rough guideline values when selecting one model over another [39].

A frequentist approach is then appealing, as it does not require assigning prior probabilities and permits interpretable assessments of statistical significance for individual events. Considering first the general formulation of model selection under a frequentist paradigm, the likelihood of observed data (x,y)[1] under a null hypothesis H_0 is directly compared to the likelihood under an alternate hypothesis H_1. Assuming parameter space Θ, H_0 specifies that (x,y) are best explained by a model with representative parameters α, taking values in $\Theta_\alpha \subset \Theta$ with probability function $f(x,y|\alpha)$ and equivalent likelihood function $L_0(\alpha|x,y)$ — often simply expressed as $L_0(\alpha)$ as data are normally considered fixed. Alternatively, H_1 specifies the data are best explained by $\beta \in \Theta_\beta \subset \Theta$ with probability function $f(x,y|\beta)$ and equivalent likelihood function $L_1(\beta|x,y)$. It is typical to assess the relative goodness of fit of the models by calculating a likelihood-ratio statistic in log-space as

$$\Lambda = -2\left[\ell_0(\hat{\alpha}|x,y) - \ell_1(\hat{\beta}|x,y)\right], \tag{2.7}$$

where $\ell_0(\hat{\alpha}|x,y) = \log\left[\sup_{\alpha\in\Theta_\alpha} L_0(\alpha|x,y)\right]$ and $\ell_1(\hat{\beta}|x,y) = \log\left[\sup_{\beta\in\Theta_\beta} L_1(\beta|x,y)\right]$ are the respective maximised likelihood functions. When H_0 is a special case of the alternate model (i.e. the two models are nested such that $\Theta_\alpha \subset \Theta_\beta$) and the null hypothesis is true, Wilks [312] showed that by multiplying the log-likelihood ratio difference by -2 (as in Equation 2.7), the Λ statistic is asymptotically χ^2_k-distributed with k degrees of freedom, where k is the difference in the number of free parameters between H_0 and H_1. This is convenient, as the Λ statistic can simply be compared to a χ^2_k distribution and H_0 can be rejected at the desired level of statistical significance.

When $\Theta_\alpha \not\subset \Theta_\beta$ however, a χ^2_k distribution cannot be used to approximate Λ under H_0 [303]. This is problematic for my purposes, as I need to perform model selection between two alignment models which are necessarily non-nested ($U \not\subset TS$). That is, the state path through the TSA pairHMM (i.e. H_1) is required to pass through all seven hidden states (Algorithm 2.2), and this means that template switch-associated parameters θ and σ (see §2.2.4) will always be included in the alternate model and no restrictions can be imposed on the parameters of TS such that $U \subset TS$. The χ^2_k approximation of Λ under the null therefore cannot be used for comparing TS and U.

For non-nested model comparisons a similar test statistic can instead be used, as was originally proposed by Cox [63, 64] and explored by others [303, 311] for investigating separate families of hypotheses. Assuming some H_0 has parameters $\alpha \in \Theta_\alpha$ and H_1 has parameters $\beta \in \Theta_\beta$, the log-likelihoods of the hypotheses can be directly compared using

[1] I refer to data as (x,y) throughout to be consistent with DNA sequences x and y assessed by my pairHMMs.

$$\Delta = \ell_1(\hat{\beta}|x,y) - \ell_0(\hat{\alpha}|x,y). \tag{2.8}$$

As with Λ, the value of Δ for some observed data can be compared to a distribution which approximates the expected value of Δ when H_0 is true. Generating the distribution of Δ under H_0 can be achieved using Monte Carlo simulations, an approach originally explored by Williams [313] and now well-established for model selection in molecular evolution analyses [103]. The general procedure (see e.g. [103, 114, 313]) involves generating through simulation many $(x,y)_1,\ldots,(x,y)_i,\ldots,(x,y)_N$ under H_0, and then calculating Δ for each $(x,y)_i$. This forms a reference Monte Carlo distribution for Δ, to which any subsequent value of Δ calculated for real (x,y) data can be compared and rejected at the desired confidence level (e.g. above the 95th percentile of the distribution).

2.3.2 LPR: the test statistic used for model selection for each candidate template switch mutation

Consider now the application of this non-nested model comparison procedure to the problem of selecting between the unidirectional pairHMM and TSA pairHMM for an observed (x,y) sequence pair. The null hypothesis U has pairwise log-likelihood function $\log[P(x,y,\hat{\pi}_U|\Theta_U)] = E_U$ which is the probability of an alignment of pair (x,y) and maximised Viterbi state path $\hat{\pi}_U$ (calculated using Algorithm 2.1) given Θ_U, the parameter space of the unidirectional pairHMM. Similarly, the alternate hypothesis TS is specified by $\log[P(x,y,\hat{\pi}_{TS}|\Theta_{TS})] = E_{TS}$, with its maximised Viterbi-like state path $\hat{\pi}_{TS}$ calculated using Algorithm 2.2. Cox's Δ test statistic (Equation 2.8, [63, 64]) for these alignment model comparisons is calculated as

$$\text{LPR} = E_{TS} - E_U. \tag{2.9}$$

I refer to this statistic as a LPR to distinguish it from a typical likelihood-ratio test statistic and to reflect that it can also be considered a logarithm of a probability ratio between the two competing alignments. As discussed in §2.2.4 and §2.2.5, parameter values used in each model are calculated from the data under investigation rather than using an iterative parameter refinement procedure such as Viterbi training (often used in place of maximum likelihood/expectation-maximisation parameter estimation for pairHMMs) [77].

2.3.3 Simulation procedure for establishing template switch alignment statistical significance

The LPR distribution under the null hypothesis can be adequately estimated using a Monte Carlo simulation approach (see above), providing that the parameters of the null model can be suitably specified. In my case, the Monte Carlo approach consists of generating many $(x,y)_1,\ldots,(x,y)_i,\ldots,(x,y)_N$ pairs of nucleotide sequences which evolved under the null hypothesis of only substitutions and indels, following the assumptions of the unidirectional pairHMM. Then I align any mutation clusters identified in each simulated $(x,y)_i$ using both the unidirectional pairHMM and TSA pairHMM, calculating a LPR between the alignment models for each cluster using Equation 2.9. I can then use this reference LPR distribution to perform null hypothesis tests using the LPRs associated with any realigned mutation cluster observed in subsequent analyses. For example, if an observed LPR falls beyond the 95th percentile of the Monte Carlo LPR distribution, the null unidirectional pairHMM can be rejected at the 5% level in favour of the TSA pairHMM. In the following subsection I will cover this computational procedure more thoroughly, and I will also describe how similar simulations under the alternate hypothesis TS are performed to estimate the statistical power of the LPR statistic.

The simulation procedure I use will differ with each subsequent analysis in terms of the model parameter values used and the number of simulations performed. Details will be given where this is the case in subsequent chapters, but the general approach is described here. I first generate the Monte Carlo LPR distribution under the null unidirectional hypothesis of no template switching. By comparing observed template switch alignments to this distribution, I can perform statistical testing for each candidate event and understand the size (probability of a false positive, i.e. a type I error) of my LPR test statistic (Equation 2.9). I also generate a Monte Carlo LPR distribution under the alternate template switch hypothesis, so as to understand the power (probability of a true positive) of my test under some additional assumptions about real template switch events (the distribution of switch points used for simulation, see below).

For the first set of simulations under the null hypothesis (without template switching), random 1kb regions are drawn from the GRCh38 reference genome, and sequence evolution is simulated in continuous time using INDELible [89] under the HKY85 substitution model [119] using nucleotide frequencies calculated genome-wide in humans. Note that nucleotide frequencies used across all simulations are fixed, as I am always working with human genome sequences. Also note that the HKY85 model is used instead of the JC69 model assumed by the pairHMMs as it provides a better approximation of human genome evolution. Evolution is performed from time $t_0 = 0$ to time t_1, where t_1 is the sequence divergence measured as the

number of substitutions per site separating the two sequences. The simulated sequence is then globally aligned back to the original sequence using the Needleman-Wunsch algorithm and a simple scoring scheme (match: 2, mismatch: -2, gap: -1) [221]. These Needleman-Wunsch alignments are then scanned with my cluster-identification approach (see Figure 2.7) and aligned under both the unidirectional pairHMM and TSA pairHMM to calculate a LPR for each identified mutation cluster, forming my Monte Carlo LPR distribution under the null hypothesis. In subsequent chapters, I set a threshold for statistical significance on this distribution at either the 95% percentile, or at the maximum observed null hypothesis LPR value. This gives the test's size, i.e. the probability of committing a type I error and inferring a false positive.

For the second set of simulations under the alternate hypothesis, I want to simulate not only sequence evolution under substitutions and indels, but also under template switch events. Template switches should be introduced into the descendant sequence at a uniformly sampled time t_{TS} (rather than at t_0 or t_1), to account for the fact that template switches could have occurred at any point in the divergence of the two sequences. To do this, first I select a uniform random time $t_{TS} \in [t_0, t_1]$. I then define a template switch event using the positioning of points ②, ③, and ④ relative to ①. For the analysis in Chapter 3, each set of relative switch points is drawn from a single high-confidence event generated between human (GRCh38) and chimpanzee (Pan_tro_3.0), using the model and filtering criteria of [185]. For human population (Chapter 4) and human cancer analyses (Chapter 5), I sample switch points using an extended set which includes both relative switch points from this initial set as well as those inferred from my hominid analysis in Chapter 3. Sequence evolution under substitutions and indels is simulated as before until t_{TS}, at which time a uniform random sequence position in the nascent sequence is selected as ① (excluding the first and last 200 bases to guarantee adequate sequence space for the template switch process). The predefined relative coordinates of points ② and ③ are used to source a sequence in reverse complement from the alternative template strand. This sequence is inserted into the sequence in a manner consistent with the template switch process, replacing the nascent sequence between points ① and ④. After this introduced templated insertion, sequence evolution continues as before under substitution and insertion/deletion, from time t_{TS} until t_1. The coordinates of the introduced event are recorded, and global alignment to the ancestral sequence is then performed. As with the null simulations, I scan for mutation clusters and realign each cluster under both pairHMMs to form my Monte Carlo LPR distribution under the alternate hypothesis. Note that this distribution may overlap the null hypothesis Monte Carlo LPR distribution. Any LPRs associated with a mutation cluster intentionally introduced by a simulated template switch that fall below the threshold on the null LPR distribution are considered type II errors (false negatives).

2.3.4 A filtering procedure is used to further increase confidence in inferred events

The LPR statistic is suitable for performing model selection and addresses the issues associated with qualitative model comparison filtering required by the approach of [185] and discussed in §2.1.2. Nevertheless, it does remain beneficial to impose some filtering on candidate template switches. Retaining a hard filter on sequence complexity is particularly useful for example, as it ensures I am not calling events for which a mutation footprint compatible with a template switch arose simply by chance due to its occurrence within a low complexity sequence region. Below I will outline additional filters which are beneficial when working with pairHMM comparisons, and provide a full list of filters applied to all events in subsequent chapters.

Because I now permit the capture of events with very short ②→③ regions, there may be candidate template switches for which the null hypothesis is a single, long deletion in the unidirectional alignment, and the alternate hypothesis is a template switch with a comparatively short ②→③ region (see Figure 2.6 for an example). My model is not able to reliably assess model fit in such cases, and it is therefore beneficial to filter these candidate events, as events with a short ②→③ region become increasingly less convincing as unidirectional deletion size increases.

A final consideration is ensuring that the LPR threshold method is not simply invoking artefactual template switch events in an attempt to correct regions of poor alignment quality or incomplete genome assembly. In such cases, the TSA pairHMM alignment may indeed be significantly more probable than the unidirectional alignment under the LPR, but the final TSA pairHMM alignment would still be a relatively low probability alignment.

To address the above concerns and improve confidence in events that I identify as significant under the LPR statistic, I therefore require events to pass the following filters in Chapter 3:

1. The alternate template sequence which donates the ②→③ fragment is not masked by RepeatMasker [276]. This disallows events being called within regions that have low sequence complexity, which often contain assembly and mapping errors [297].

2. The ②→③ sequence must contain all four nucleotides. This catches low complexity regions missed by the first filter, while explicitly disallowing calling of events within mono-, di-, and trinucleotide repeats. These microsatellite repeats show increased rates of replication slippage [166], which may be defined as a form of linear orientation template switch [37], but I have no way to distinguish these cases from the reverse orientation template switches I model here. This filter implicitly also sets a minimum

length threshold of 4 on the ② → ③ region, defining the minimum templated insertion length required to explain any focal mutation cluster (where each cluster is defined using the procedure outlined in §2.2.7).

3. Unidirectional alignments must contain fewer than 50 deletion columns. Candidate events characterised by a single deletion of around this size in the unidirectional alignment are large enough to yield a significant LPR in subsequent analyses. My statistical methods have no way to determine if these cases are true template switches, or large deletions with a small amount of reverse complement sequence (e.g. 4nt) which create false positive ② → ③ sequence matches. I therefore omit them to (conservatively) remove potential false positives events.

4. The length-normalised probability of each significant template switch event must be greater than or equal to a threshold set on a randomly sampled distribution of length-normalised unidirectional alignment probabilities. In Chapter 3, pairwise alignment regions are randomly sampled genome-wide between each pairwise great ape alignment (see §3.2.2).

This general filtering approach is retained in my human population (Chapter 4) and human cancer analyses (Chapter 5), but some analysis-specific changes are respectively detailed in §4.4.2 and §5.4.1.

2.4 Conclusions

PairHMMs offer a highly suitable framework for identifying template switch mutations from pairs of sequences which share a recent common ancestor. In this chapter, I first introduced template switching as a possible mechanism underlying small-scale genomic rearrangements, and defined a mathematical model which describes template switch events and the effect on genomic sequences. I then devised a pairHMM that can propose the most likely template switch explanation for a mutation cluster in a pre-existing linear alignment, and showed that a canonical pairHMM can act as the null hypothesis that the cluster arose solely through substitutions and indels. I then outlined a suitable test statistic to assess the statistical significance of candidate template switch events, and I provided a procedure for generating reference distributions of this test statistic under both the null and alternative hypotheses. From these Monte Carlo distributions, I have described how thresholds can be defined to make principled decisions about candidate events with measurable false positive and false negative rates to understand the size and power of my statistical tests.

Throughout the remainder of this thesis, I will apply the pairHMM comparison procedure outlined in this chapter to identify significant template switches in the contexts of hominid genome evolution, population-scale human variation, and between healthy and tumour tissues. Specifically, in Chapter 3, I detail the prevalence of template switching in hominid evolution, discuss issues surrounding calling events in this context, explore the properties of rearrangements introduced by template switches, and explore associated genomic features and motifs. In Chapter 4, I develop a pipeline to identify significant template switches in variant calls from populations of humans and parent-offspring trios across a variety of sequencing datasets. From this procedure, I delineate the extent to which template switch mutations can explain mutation clusters and short indels in human resequencing data, I delineate population structure, I assess the impact of template switching on short-read sequencing coverage, I provide mutation rate estimates, and I explore associations with an expanded set of associated genomic features for which human-specific genomic annotations are publicly available. Finally, in Chapter 5, I apply the pipeline developed in Chapter 4 to identify events in human cancer across a range of histologies, exploring the statistical issues with calling template switches from variant calls in this setting.

Before considering the application of my models, it is worth briefly reflecting on some areas of my methods which could be explored further in future. For example, an alternate formulation of my filtering procedure could perhaps be considered, as one of my early aims was to move past qualitative filtering for identifying credible template switch alignments. Because much of the human genome consists of low complexity sequence however, some level of filtering is difficult to avoid when modelling and interpreting template switch mutations, as is the case when considering credible variant calls for all classes of mutation [297]. It may also be worth considering how simulating my null Monte Carlo LPR distribution under the assumptions of HKY85 [119] impact the statistical interpretation of alignments produced under JC69 [138]. Further, it would be interesting to understand if I am under or overestimating the statistical power of the LPR statistic by simulating template switch mutations using relative switch point distributions which were obtained by applying a simpler model [185]. Finally, it would be interesting to assess if the performance of my methods could be improved using a "Forward-Backward-like" algorithm, Viterbi training or the Baum-Welch algorithm for parameter estimation, and possibly considering a pairwise stochastic context-free grammar formulation of my methods in greater detail.

Chapter 3

Template switch mutations in great ape genome evolution

Chapter overview

In Chapter 2 I described the alignment models which allow short template switch mutations to be detected and statistically evaluated using input pairwise sequence alignments. In this chapter, I describe the application of these models to pairwise great ape genome alignments, characterising the prevalence of template switch mutations throughout hominid genome evolution, as well as describing their associated genomic features.

Declaration

The content of this chapter was adapted and expanded from a first-author publication [305]:

Walker C. R., Scally A., De Maio N., Goldman N. Short-range template switching in great ape genomes explored using pair hidden Markov models. *PLOS Genetics* 17, e1009221 (2021).

For this paper, I developed the methods and wrote the C++ implementation. Additionally, I performed all data collection, processing, analysis, and data visualisation. I wrote the original manuscript, which was subsequently edited and agreed upon by all co-authors.

Code and data availability

All code underlying the analysis of this chapter and [305], as well as any associated supplementary files, are available from:
https://gitlab.com/conorwalker/phd_thesis/tree/main/chapter_3.

3.1 Background

To interpret the evolutionary forces which have acted on the human genome, potentially including short template switch mutations, it is necessary to compare the human genome sequence to those of our close relatives. It is well established that humans and other extant members of the family Hominidae (composed of humans, chimpanzees, gorillas, and orangutans, referred to as the great apes or hominids) share a recent common ancestor. Sequence divergence at orthologous sites between humans and chimpanzees, humans and gorillas, and humans and orangutans is relatively low, at approximately 1.2%, 1.6%, and 3.1%, respectively [163, 212, 266]; the species-level phylogeny is well resolved [266]; and patterns of incomplete lineage sorting (ILS), in which the local phylogeny does not reflect the species-level phylogeny, are well defined along the genome [163, 188–190, 266]. The combination of low sequence divergence between the hominids and a robust phylogeny relating alleles along their genomes has made this clade well-suited to investigating mutations specific to individual branches (though see also §3.2.1). As a result, genome evolution within the great apes is well-characterised. Robust species-specific phylogenetic mutation rates have now been estimated, which have permitted the calculation of species divergence times [265]. Functional annotations have been assigned to much of the human genome [294] (although this endeavour is not without its critics [108]) that inform our understanding of variants which may modulate disease risk. This combined body of work has facilitated the evolutionary interpretation of some genomic regions that make us distinctly human. For example, comparative genomic studies have identified functional genomic elements that regulate large brain development [100], which are often found in regions of accelerated substitutions rates specific to the human lineage [235, 236].

Underlying these comparative evolutionary genomic analyses are either pairwise or multiple sequence alignments of whole reference genomes. There are several potential confounding factors in alignment-dependent evolutionary genomic analysis which are relevant to my exploration of template switch mutations in great ape evolution: the use of reference genomes, accuracy of the whole genome multiple sequence alignment, and inference of the true mutational mechanisms which underlie the detected variation. Before focusing on template switch mutations in the evolution of great apes, I will briefly address the potential impact of the first two issues on my analysis, and outline how my investigation of template switch mutagenesis can alleviate the third.

Great ape reference genomes have historically been produced by sequencing either a single individual [212, 266], or multiple individuals but much of the assembly is contributed by a single sample [168, 269]. In both cases, the haplotypes at each locus are provided as a

haploid representation of the genome from one individual. This could be problematic when attempting to capture rare forms of variation such as template switches which are potentially only present in a single lineage. For example, assuming a template switch is present at a locus in a subset of individuals in a population, the small selection of sequenced samples used to produce that reference genome may possess the major allele that does not represent a template switch, causing it to be missed using a comparative approach. This issue is largely ignored in inter-specific comparative evolutionary genomic studies however, and only the variation that exists between species rather than within species is characterised. While not typically an issue within a phylogenetic context, it is worth noting that the reference human genome used throughout this thesis (GRCh38.p12) represents a minor human allele at approximately two million sequence positions [24], so some interspecific variation will always be mischaracterised. I will return to the topic of intraspecific variation in humans in Chapter 4, but for the remainder of the analysis of the great apes' genomes, I ignore this issue, knowing that some template switch mutations will not be detected.

Modern multiple sequence alignment pipelines that are used to align great ape reference genomes such as the Enredo, Pecan, and Ortheus (EPO) pipeline deployed by Ensembl [230, 322], typically perform well in alignment benchmarks, achieving high precision whilst accurately resolving chromosomal rearrangements [18, 79]. Additionally, comparative genomic studies typically filter input multiple sequence alignments to remove alignment blocks which are of poor quality or excessively masked [276], meaning fidelity between the mutations inferred from these alignments and the true variation in the genomes is only a small concern. It is however worth noting that the percentage of each species' genome included in these whole genome alignments varies. For example, in the Ensembl EPO alignments of 12 primates (at the time of writing), the orangutan (*Pongo abelii*) reference genome PPYG2 is covered at only 79% of positions, while the reference human genome GRCh38 is covered at 90% of positions [322]. This means that some amount of rare variation identified through comparative approaches will always be missed regardless of alignment accuracy.

This leaves the third issue, accurately inferring the causative mechanisms underlying lineage-specific substitutions, insertions, deletions, and rearrangements within these alignments. As outlined by Löytynoja and Goldman [185], template switch mutations are expected to leave footprints of complex mutation, manifesting as any combination of these mutation types clustered together in nearby linear alignment space. Evolutionary models which are informed by the patterns of substitution across an input alignment typically make the assumption that substitutions at sites not undergoing positive selection arise independently due to unrepaired DNA damage and DNA replication errors as outlined in Chapter 1, following the neutral

theory of evolution [150]. However, the instantaneous appearance of apparent mutation clusters through template switching could cause inferences of gross violations of neutrality if unaccounted for, and has the potential to generate signals of (e.g.) lineage-specific accelerated evolution in single mutational events. To understand how false signals of human evolution could be generated by this poorly characterised mutational mechanism, it is therefore necessary to assess template switch prevalence across the human genome with an evolutionary context.

Löytynoja and Goldman [185] delineated mutation clusters caused by template switch mutagenesis using pairwise alignments of the human (GRCh37) and chimpanzee (CHIMP2.1.4) reference genomes. While they were able to identify many complex mutation clusters between the human and chimpanzee genomes which were putatively caused by template switch mutations, there were several limitations of their study. First, methodological limitations (as discussed in §2.1.3) did not allow candidate events to be statistically assessed for significance. Second, investigating events solely by comparing two species does not allow template switches to be interpreted in their phylogenetic context, making it impossible to assess the impact of template switching on lineage-specific evolutionary inferences. Note that this problem is further complicated by the existence of "reversible" template switch mutations, which are detected regardless of which species is treated as the ancestral state, and which as the descendant state. See §3.3.2 and Figure 3.5. Third, questions remain about the associated functional annotations, genomic features, physical properties, and sequence motifs which may influence event formation.

In this chapter, I address all of these shortcomings, and describe the landscape of template switch mutagenesis in great ape genome evolution. Following the methods introduced in §2.3.3, a series of simulations are used to determine LPR thresholds for my pairHMMs, and sample pairwise alignments of great ape genomes to further establish baseline alignment quality thresholds (§3.2). I then apply my updated probabilistic models to pairwise alignments of the human, chimpanzee, and gorilla reference genomes, summarising the significant events identified and interpreting all events in their phylogenetic context (§3.3). I achieve greater resolution in the detection of short-range template switch events across the human reference genome than [185], identifying thousands of significant events across the great ape tree. Finally, I explore associations between event loci and human-specific genomic landmarks (§3.4). I show that event initiation may be modulated by poly(dA:dT) tracts which in turn cause an increased propensity for DNA bending and DNA double-strand break formation including features involved in transcriptional regulation, and I consider the impact of all identified events on signatures of lineage-specific evolution.

3.2 Data collection and establishing statistical thresholds

To discover and characterise template switch mutations in human genome evolution, I downloaded the Ensembl (v.98) EPO alignments [322] of thirteen primates. I extracted pairwise alignment blocks between human (GRCh38.p12) and chimpanzee (Pan_tro_3.0), human and gorilla (gorGor4), and gorilla and chimpanzee. Gap-only columns were removed for each pairwise comparison, with their positions recorded to allow me to relate the coordinates of events across comparisons to the original multiple sequence alignment coordinates later.

Before applying the models established in Chapter 2 to mutation clusters within these pairwise alignments, it is necessary to establish a threshold on the LPR between the alignment probabilities of the unidirectional and TSA pairHMMs, thus allowing me to assign statistical significance to any candidate template switch event (as detailed in §2.3.3). Additionally, to ensure the LPR threshold method is not simply invoking artefactual template switch events in an attempt to correct regions of poor alignment quality or incomplete genome assembly, I establish an average alignment quality filter by sampling these alignments (filter (4) in §2.3.4).

3.2.1 Simulations of template switching to determine a significance threshold for individual hominid events

I sought to establish a LPR threshold (Equation 2.9) that maximises the recall of true template switch events and minimises the number of false positives caused by erroneously explaining a true cluster of substitutions and indels as an apparent template switch. To this end, I used the simulation procedure outlined in §2.3.3. Robust estimates of the evolutionary distance between human-chimpanzee and human-gorilla are in the range of 1.2% to 1.6% [163]. Using the §2.3.3 Monte Carlo procedure, I simulated two sets of sequence evolution in 0.1% steps of t between 1% and 2% to cover this range, setting a LPR threshold on the null LPR distribution (no template switches) which only allows 0.5% of false positives through.

Even at small evolutionary distances, many simulated template switch events are obfuscated by surrounding neutral mutations, allowing me to capture an average of 78% of introduced events when simulating between 1–2% divergence (Figure 3.1a). Of the recaptured events, a threshold on the LPR is able to successfully discriminate between true positives (introduced events) and false positives (background mutation clusters) (Figure 3.1b). Setting a false positive rate of 0.5% still enables a high average recall (0.85 ± 0.04 SD across simulated divergences) of recaptured events, achieved at an average LPR threshold of 8.95 (Figure 3.1c). For subsequent great ape analysis, I set the LPR threshold to 9 (rounded from 8.95), forming my significance cutoff for rejecting the null hypothesis that no template switch event was involved in the creation

of an aligned descendant sequence. This threshold is fixed across pairwise comparisons to assign the same level of significance to all detected hominid events.

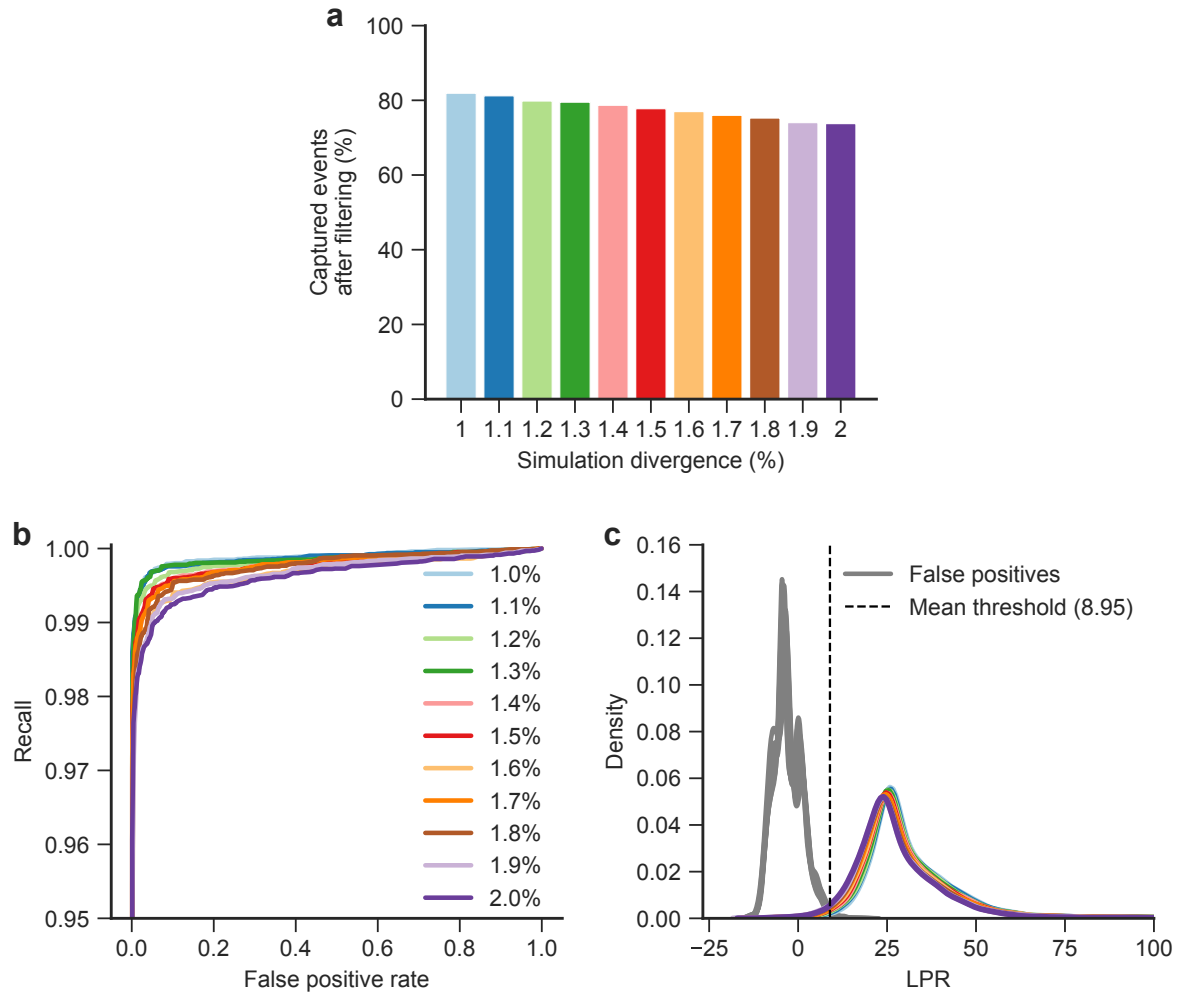

Figure 3.1: Simulated events can be distinguished from background mutation clusters. (**a**) Percentage of events recaptured from simulations of template switch events alongside substitutions and indels using INDELible across a range of divergences. (**b**) Receiver operating characteristic (ROC) curves for discriminating between simulated template switch events and background mutation clusters. Simulations using divergence t from 1–2% in 0.1% steps are shown (t value for each curve indicated by matching colour in part a). Note that the y-axis begins at 0.95 for clarity. (**c**) Density curves of LPRs for true positive (i.e. intentionally introduced) template switch events in colours corresponding to (a), and false positive events across all simulation values of t (background/chance mutation clusters) in grey. The mean LPR threshold required to achieve a FPR of 0.5% across simulations is shown as a dashed line at the value of 8.95.

Simulations at smaller evolutionary distances provide a modest improvement in recall (Figure 3.1b), which is expected, as events are obfuscated by fewer substitutions and indels. Divergence in both pairHMMs is specified using the parameter t (expected number of substitutions per codon, as detailed in §2.2.4 and §2.2.5) which, for each simulation, I set equal to the corresponding parameter value used with INDELible to represent the simulated evolutionary distance. I confirmed that my inferences are robust to misspecification of t (see Figure 3.2). While my method is able to robustly detect template switches, it is worth reflecting on the observation that sequence evolution can rapidly obfuscate the signal from past template switches. Even when simulating at small evolutionary distances of 1–2%, simulated events are often not recaptured due to background substitution and indel processes overlapping the event region (Figure 3.1a). Additionally, some events are indeed detected, but are obscured by background substitutions and indels, causing their final LPR to fall below the significance cut-off shown in Figure 3.1c. This suggests that estimates of the prevalence of short-range template switching in the evolutionary history of the hominids will underestimate the true prevalence.

3.2.2 Sampling hominid alignments to determine genome-wide alignment probabilities

To establish an average alignment quality filter (filter (4) in §2.3.4), I first sampled 100,000 random 300nt blocks from each of the human/chimpanzee, human/gorilla and chimpanzee/gorilla pairwise alignments. Each block was globally aligned under my unidirectional pairHMM (Figure 2.3a), with pairwise parameters kept identical to those used for all other analysis. I calculated a length-normalised log-probability for every sampled alignment block, by dividing each unidirectional pairHMM alignment log-probability by its corresponding alignment length. I then set the 20th percentile of the distribution of these values (Figure 3.3) as a species pair-specific threshold on the minimum length-normalised log-probability of any template switch alignment. This is assessed for each template switch alignment after subtracting the log-probability contributions of the transitions into and out of M_2 from the global event log-probability. This ensures that template switch alignments in my final event sets are as probable as the majority of linear alignments in the considered pairwise comparisons, rather than just exchanging regions of very poor alignment quality or genome assembly for a comparatively more plausible template switch alignment.

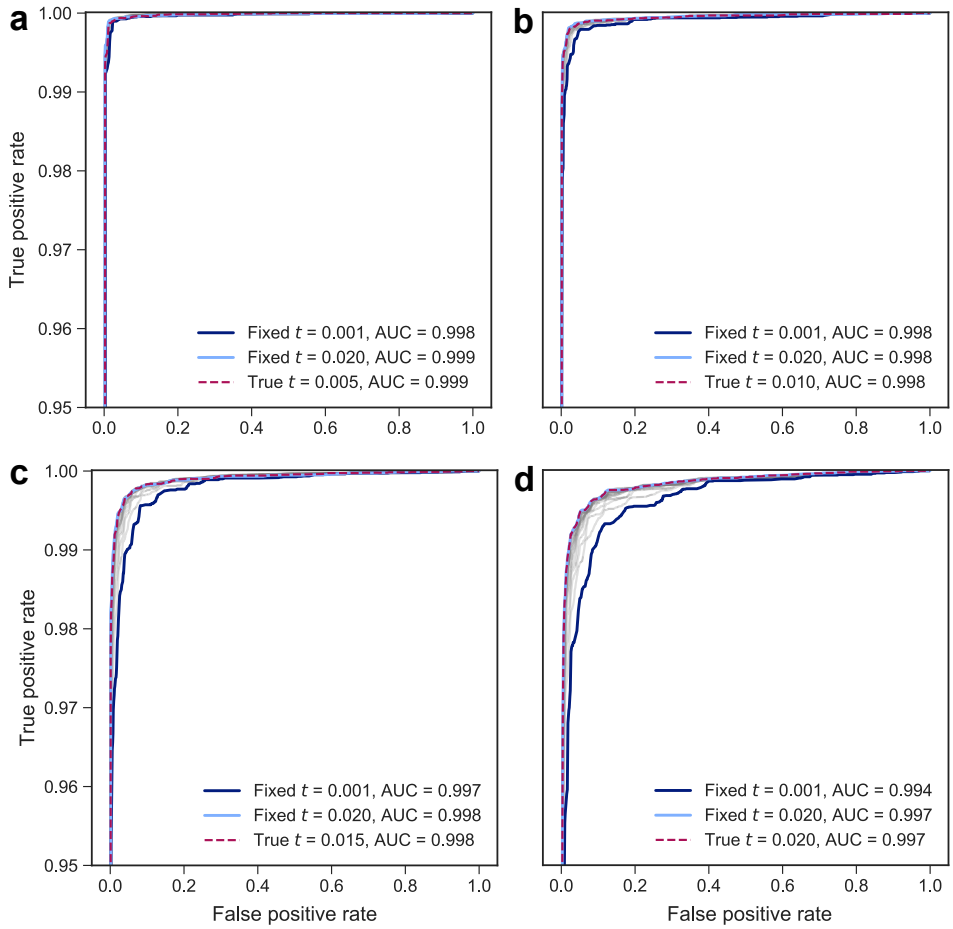

Figure 3.2: Template switch inference in great ape genomes is robust to misspecification of pairHMM parameter t. ROC curves for simulations at evolutionary distances of (**a**) 0.005, (**b**) 0.010, (**c**) 0.015, and (**d**) 0.020. At each evolutionary distance, the TSA pairHMM parameter t was set independently of the evolutionary distance used for sequence simulation, ranging from 0.001 to 0.02 in 0.001 increments. The ROC curve for the t parameter corresponding to the true evolutionary distance is shown as a dashed magenta line, the minimum and maximum fixed t values are in dark blue and light blue, respectively, and all other values of t are shown in grey. Across all fixed evolutionary distances, almost identical performance is achieved using the true t and using the highest fixed value of t, while marginally worse performance is observed when fixing t to smaller values. The performance differences are so small (as measured by the area under the ROC curve (AUC)) that any misspecification of t will have a negligible impact on model performance, indicating that my inferences are robust to my assumed values of t. Note that all y-axes start at 0.95, as the ROC curves between specified values of t would otherwise be indistinguishable.

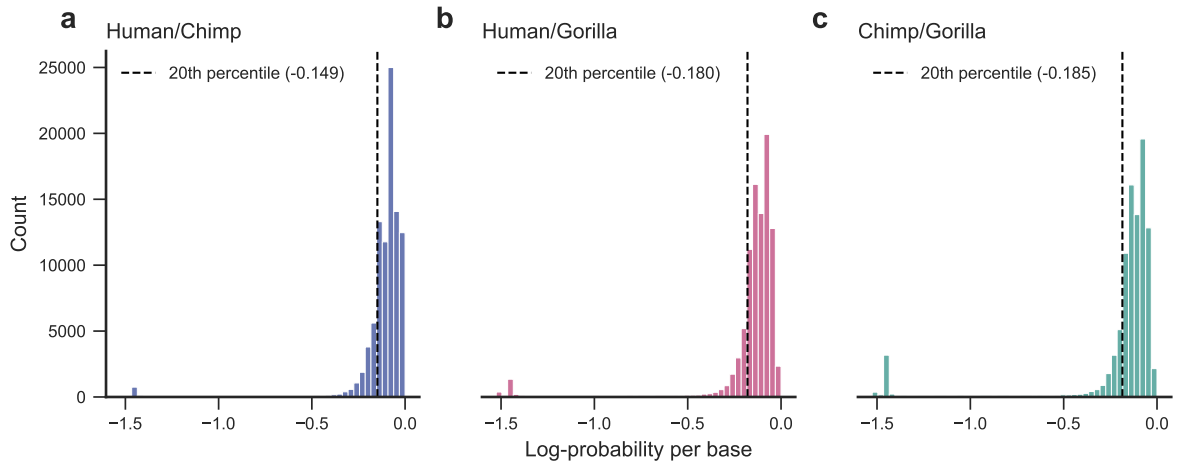

Figure 3.3: Alignment quality thresholds for candidate hominid template switches. Genome-wide samples of alignment log-probabilities under the unidirectional pairHMM for (**a**) human/chimp, (**b**) human/gorilla, and (**c**) chimp/gorilla. The derived log-probabilities of sampled alignment regions are normalised by final alignment length to produce per-base log-probabilities. Dashed lines represent the 20th percentile thresholds used as baseline alignment quality thresholds for event regions for each pairwise comparison. If both the null model and the template switch model alignments in a region fail this threshold, the region is removed from my analyses.

3.3 Short-range template switch mutations are prevalent in the genomes of great apes

3.3.1 Discovering candidate template switch mutations

Using the statistical thresholds established in §3.2.1 and §3.2.2 above, I sought to discover template switch mutations in the genomes of great apes, allowing events in the human genome to be interpreted in a phylogenetic context. This is achieved using the pairwise alignments I extracted from the EPO multiple sequence alignments (described in §3.2), where both species from each pairwise alignment are considered as being representative of the ancestral and descendant sequence states in turn. That is, for each of the retrieved pairwise species alignments in {(human, chimpanzee), (human, gorilla), (chimpanzee, gorilla)}, I assess each pairwise alignment twice by switching which of the species is specified as ancestral x and descendant y in the pairHMMs for each mutation cluster identified in that pairwise comparison. Looking in both directions like this facilitates the subsequent placement of events in their evolutionary context (see §3.3.2). As in [185], mutation clusters within each pairwise comparison are defined as any 10nt window in which two or more nonidentical bases are identified. Once $\geqslant 2$

pairwise differences are identified, the cluster itself and a small sequence region upstream and downstream of the cluster boundaries is considered for alignment by following the procedure outlined in §2.2.7 and Figure 2.7, ensuring the log probabilities of each model can be compared fairly. For each pairwise comparison, the values of the pairHMM parameters required to calculate transition and emission probabilities are specified in Table 3.1.

The statistical significance of each candidate event aligned using this procedure is assessed using the LPR threshold determined in §3.2.1. I then apply the alignment quality threshold and event filters detailed in §3.2.2 and §2.3.4. Events were removed from the event set if either the LPR was non-significant or if one of the additional filters was not passed. After this procedure, 4017 significant events were identified across the six comparisons. Unidirectional and TSA pairHMM alignments for all significant events are provided in the supplementary data files for this chapter (data/significant_template_switch_events_pairhmm_output.txt), and an annotated entry of this file is shown in Figure 3.4. The corresponding human genome (GRCh38.p12) coordinates of the mutation clusters associated with each event are also provided in the supplementary data files (data/significant_template_switch_events.grch38.bed).

3.3.2 Phylogenetic interpretation of hominid template switch mutations

With these significant events identified, accurately placing each event onto the hominid tree and determining their evolutionary direction is desirable for several reasons. It increases

Table 3.1: PairHMM parameters used in the great ape analysis.

Parameter	Value(s)	Rationale
t	(human, chimpanzee): 0.01, (human, gorilla): 0.016, (chimpanzee, gorilla): 0.016	Based on estimates from [163]
ρ	0.14	Based on estimates from [46]
λ	20	Based on estimates from [46]
N	2750	See §2.2.4
C	7.9×10^6	The average number of mutation clusters (defined using the procedure described in §2.2.7) found across the three hominid pairwise alignments
L	10	See §2.2.4

Figure 3.4: Output for each significant template switch event detected in any great ape genome. For each of the 4017 significant events included in the supplementary data (data/significant_template_switch_events_pairhmm_output.txt), I indicate: the phylogenetic resolution; the set of comparisons in which the event was detected, as well as detected and significant; the EPO alignment for this region of the human, chimpanzee, gorilla, and orangutan genomes if available (the human sequence is shown twice to allow easier comparisons with the gorilla sequence) with annotations showing the evolutionary direction and ancestral switch point coordinates; and finally, the unidirectional and TSA pairHMM alignments for each significant comparison.

confidence in events I identify as significant, as events for which an unambiguous direction cannot be established either reside in regions of poor assembly quality in one or more of the target genomes or of poor multiple sequence alignment, or are obscured by the co-occurrence of background mutational processes. It also enables the assignment of an event type (the ordering of switch point locations with respect to the ancestral sequence; see §2.1.2) to each unique event, allowing me to infer whether each one could have arisen via intra-strand template switching or inter-strand template switching (as discussed in §2.1.2). Finally, knowing the ancestral and descendant sequences allows me to investigate potential causative ancestral (and consequent descendant) features associated with events.

I first identified events which correspond to one another across pairwise comparisons. I converted the pairwise alignment coordinates of each mutation cluster associated with a significant template switch event into their corresponding multiple sequence alignment coordinates. I then checked for any overlap in the alignment coordinates of each event-associated mutation cluster identified in each pairwise comparison, recording the set of comparisons in which each significant event was found.

Using these sets of overlapping alignment coordinates, I aimed to place each significant event onto its correct branch of the hominid phylogeny. For each pairwise comparison, if the true ancestral and descendant sequences are correctly designated in my model as x and y, respectively, and post-event substitutions and indels have not excessively altered the ancestral sequence, the TSA pairHMM is able to reconstruct y from x. Assuming these loci are biallelic (presence/absence of a template switch mutation) and assembly quality is high, there should always be two of the six possible comparisons where the model reconstructs y from x. I can use these two comparisons to place an individual event onto the hominid phylogeny. For example, a significant event detected in the comparisons with each of the gorilla and chimpanzee sequences, respectively, designated as representing the ancestor (x) of human (descendant y) is denoted as being found in the gorilla $\rightarrow$ human and chimp $\rightarrow$ human comparisons and must have occurred in the human lineage. Similarly for the event shown in Figure 3.4, a template switch was identified in the pairwise comparisons in which gorilla was designated as ancestral to human (gorilla $\rightarrow$ human), and ancestral to chimpanzee (gorilla $\rightarrow$ chimp), so I can infer that the event occurred on the branch leading to human and chimpanzee. This can be further confirmed by inspecting additional outgroup genomes, such as the orangutan sequence, which in this case is identical to the gorilla sequence for this region.

However, when considering each species pair as ancestral/descendant (x/y) in turn, a subset of events are significant regardless of which species is designated x or y, allowing y to be reconstructed from x across four comparisons instead of two as above. I refer to these events as

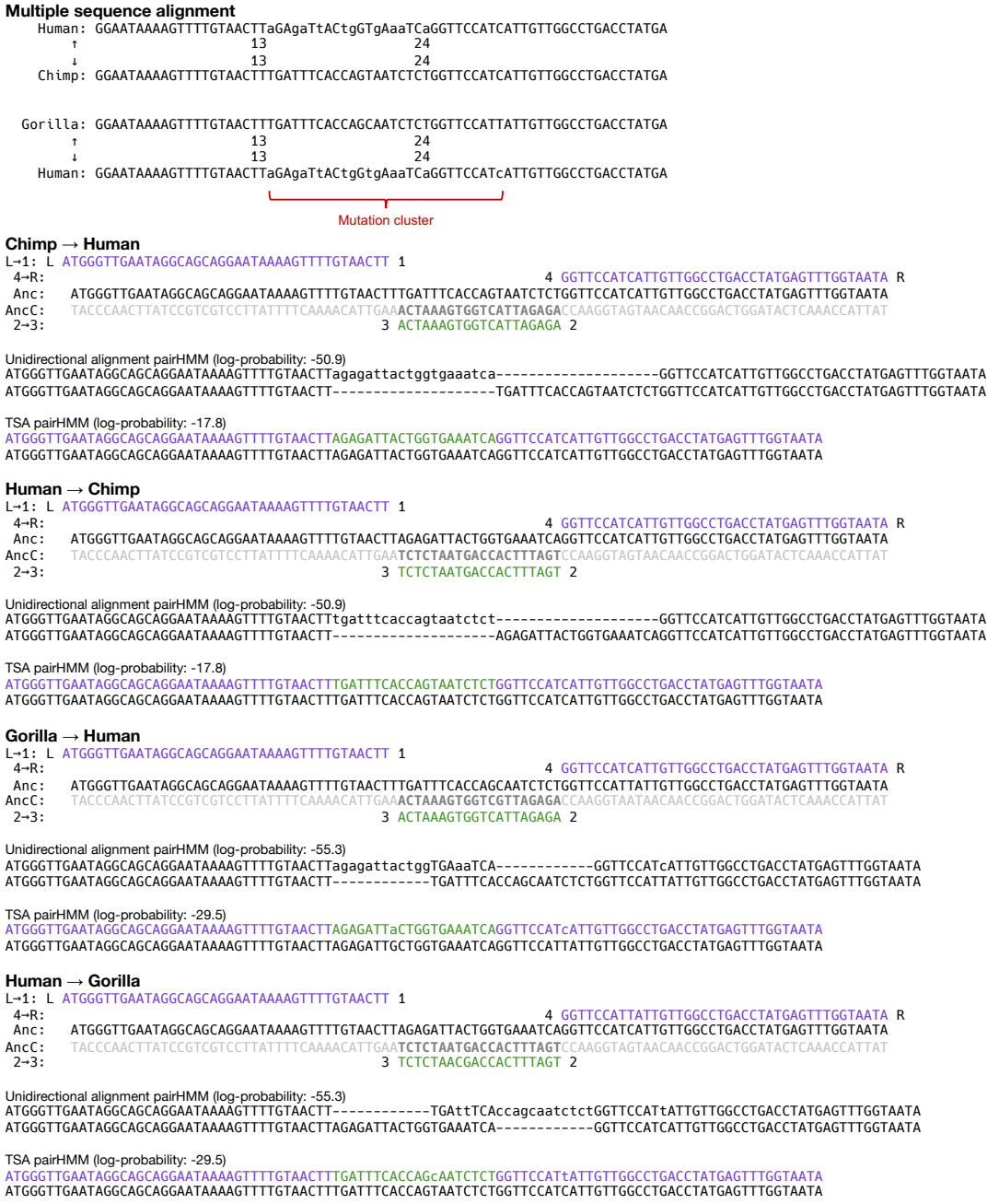

Figure 3.5: An example of a "reversible" event. A mutation cluster is observed between human/chimpanzee and human/gorilla, appearing as either a large cluster of substitutions (input multiple alignment, top), or as a large insertion and deletion event (unidirectional pairHMM alignments). Regardless of which species is specified as the ancestral sequence x or the descendant sequence y, the event is detected as significant. As I cannot tell whether this event is congruent with the species tree or represents a region of incomplete lineage sorting, I am unable to place it onto an evolutionary lineage. Coordinates here refer to positions from sequences aligned to the negative strand of GRCh38. "Anc" refers to the assumed ancestral sequence and "AncC" refers to the complement of this sequence.

"reversible", and their identification as "reversible detection", as the true ancestral sequence can be reconstructed from the true descendant sequence as well as *vice versa*. An example reversible event is shown in Figure 3.5. Event reversibility is determined by the number and length of apparent deletions introduced into the true descendant sequence. For example, if an event causes many deletions in the true descendant sequence *y*, such as a ①-③-②-④ event which replaces a larger region (between ① and ④ of *x*) with a shorter region (reverse complement of ③ to ② of *x*), too much sequence information will be lost to reversibly reconstruct *x* from *y*. Extending my previous example, consider an event that can additionally be detected in both comparisons with the human sequence designated as ancestral. This event is now denoted as gorilla $\leftrightarrows$ human and chimp $\leftrightarrows$ human. From this set of comparisons and directions, I cannot infer whether the chimpanzee and gorilla sequences correspond to the ancestral state (consistent with an event in the human lineage of the species tree), or the human sequence does (consistent with the ILS tree). In such cases, although I observe the event across a consistent set of pairwise comparisons (i.e. I have only observed two possible ancestral or descendant species), I cannot unambiguously place the event onto a single lineage.

Using these methods, I defined an annotation for each set of evolutionary directions across which individual events are discovered (Figure 3.6, dot matrix and row labels). These annotations are then used to either place events onto individual evolutionary lineages, or to demarcate ambiguous placement when assigning an event to a particular lineage is not possible without further outgroup comparisons. For each unique, significant template switch event that cannot be clearly assigned to either a set of directions which are consistent with the species tree or with ILS, I investigated the non-significant pairwise comparisons for evidence of template switches that fall marginally below the significance threshold or otherwise fail one or more of the other filters. Unique events that are significant in one comparison, but are either non-significant or fail one of my additional filters are assigned to the appropriate species tree- or ILS-consistent set, but are not used in downstream analyses (Figure 3.6, light blue bars). Remaining events retain the annotation of incomplete detection (Figure 3.6, grey bars).

Accounting for poor assembly quality, ILS [190], and event reversibility (e.g. Figure 3.5), I successfully placed almost all significant events on the hominid tree (Figure 3.6). Only six events remain unresolved (Figure 3.6, black bars), representing either regions of poor alignment quality or false positives which marginally pass the LPR threshold (Event IDs in supplementary data files: 563, 917, 1587, 1657, 1742, 3069). Of the resolved events, 1310 are consistent with the species tree and significant across all expected pairwise comparisons (Figure 3.6, dark blue bars, dark blue dots); 193 are consistent with a pattern of ILS and are significant across all expected pairwise comparisons (e.g. human appearing ancestral to both chimpanzee and gorilla;

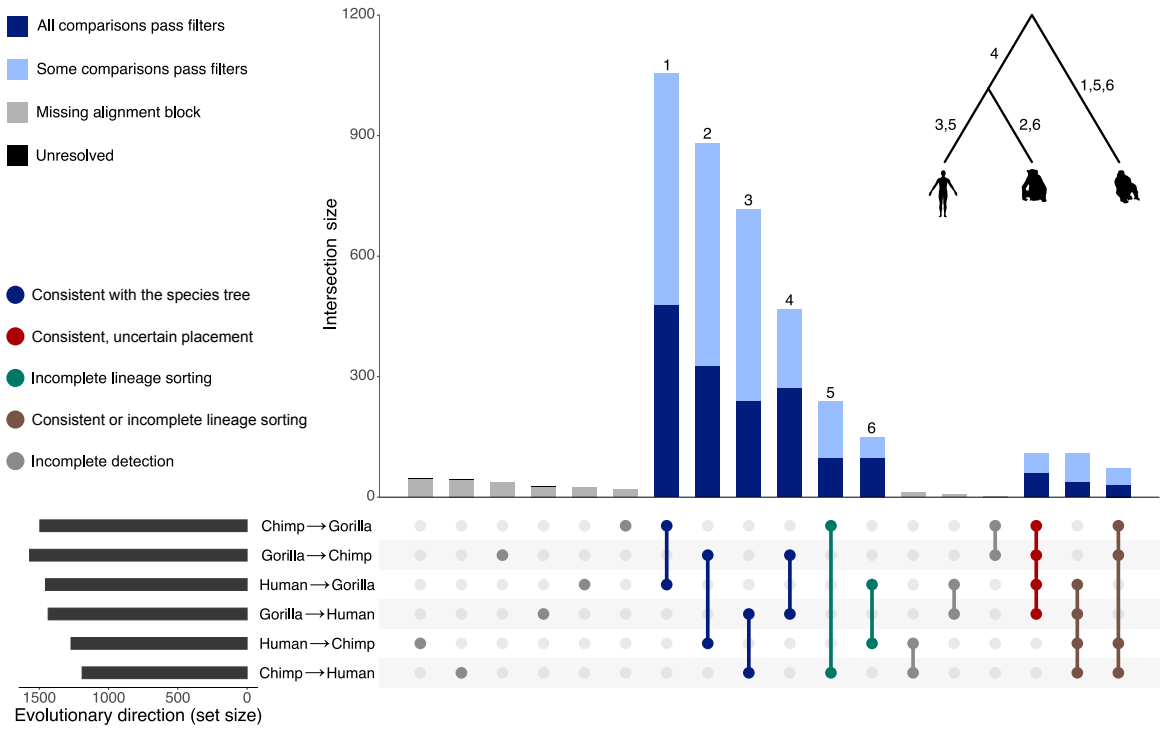

Figure 3.6: Evolutionary direction of hominid events. For each of the 4017 unique events, the intersection of pairwise genome comparisons in which it was found is indicated by the columns of bold/connected circles in the dot matrix, with corresponding intersection sizes shown above as the vertical bar plot. Detected event set sizes for the six pairwise genome comparisons are shown to the left on a horizontal bar plot. Intersections in the dot matrix are coloured according to expected direction: dark blue represents consistency with the hominid species tree, grey intersections should not be observed, teal represents incompatibility between the local tree and species tree consistent with ILS, red represents consistency with the hominid tree but uncertain branch placement, and brown represents events that are consistent either with the hominid tree or with ILS and cannot be resolved without further outgroup comparisons. Counts of evolutionarily consistent events that pass all filters are shown as dark blue bars, events with a consistent set of directions for which one or more of the comparisons has a non-significant LPR or fails an additional filter are shown in light blue, and events for which one of the genomes in this region is either absent from the alignment block or entirely gapped are shown in grey. A total of six events with unresolvable directions are shown in black at the top of the grey columns for human → chimp, chimp → human and gorilla → human comparisons; these are near-invisible due to their small numbers. Numbers above the bars of each consistent direction set indicate unambiguous placement of those events on the correspondingly numbered branch of the displayed hominid phylogeny.

Figure 3.6, dark blue bars, teal dots); 125 are significant across appropriate comparisons but could either be consistent with the species tree or with ILS, and cannot be unambiguously placed on a branch without additional outgroup comparisons (Figure 3.6, dark blue bars, red and

brown dots); 2170 are consistent with either the species tree or with ILS, but are not significant across all expected comparisons (Figure 3.6, light blue bars); and 213 cannot be placed on the hominid tree due to a missing or entirely gapped alignment block in one comparison (Figure 3.6, grey bars, grey dots). Among these event classes, it is likely that the most prevalent — those detected in an evolutionarily consistent set of comparisons, but not significant across all comparisons — is due to event obfuscation through background mutation accumulation in event regions, as demonstrated by my analysis of simulated event sets (Figure 3.1a).

For the purposes of subsequent analysis, I define two event sets of interest. First, the "unique" event set contains all 4017 of the significant events outlined above, allowing me to compare events discovered using my approach to that of [185]. Second, the "gold-standard" subset comprises events that are consistent with the species tree or with ILS and are significant across all relevant pairwise comparisons, allowing unambiguous placement on the hominid phylogeny (n=1503; Figure 3.6, dark blue bars, dark blue and teal dots). It is worth noting that while I emphasise confident placement of events onto specific branches for the gold-standard set, many significant events inferred with a high LPR are harder to place unambiguously because they are reversibly detected (see 3.5) but could be considered gold-standard if a more complete great ape phylogeny was used to facilitate lineage assignment. I use the gold-standard events to investigate genomic features associated with events' ancestral and descendant sequence contexts and physical properties of DNA surrounding event loci.

I assessed how my method compares to that of Löytynoja and Goldman [185] in terms of the number of events confidently detected, and the impact of my replacement of some non-probabilistic filters with probabilistic thresholds and statistical tests. After performing the same analysis as above but using their model and filtering scheme, I identified 3056 unique events across the three sets of pairwise comparisons (Figure 3.7a). Despite my larger unique event set, the number of events with an "unresolved" evolutionary direction drops from 8% (246/3056 unique events) using their approach (Figure 3.7b), to 0.15% (6/4017 unique events) using my approach (Figure 3.6). This demonstrates that my methods are superior in terms of both the total events recovered from pairwise alignments between closely related species and capability to interpret this larger set of events in their phylogenetic context.

3.3.3 Template switch summary statistics

Short templated insertions are a difficult class of rearrangement to capture in an evolutionary context, as many will plausibly present as a mutation cluster or short indel event in a multiple sequence alignment. Focusing on the gold-standard event set, my model largely captures

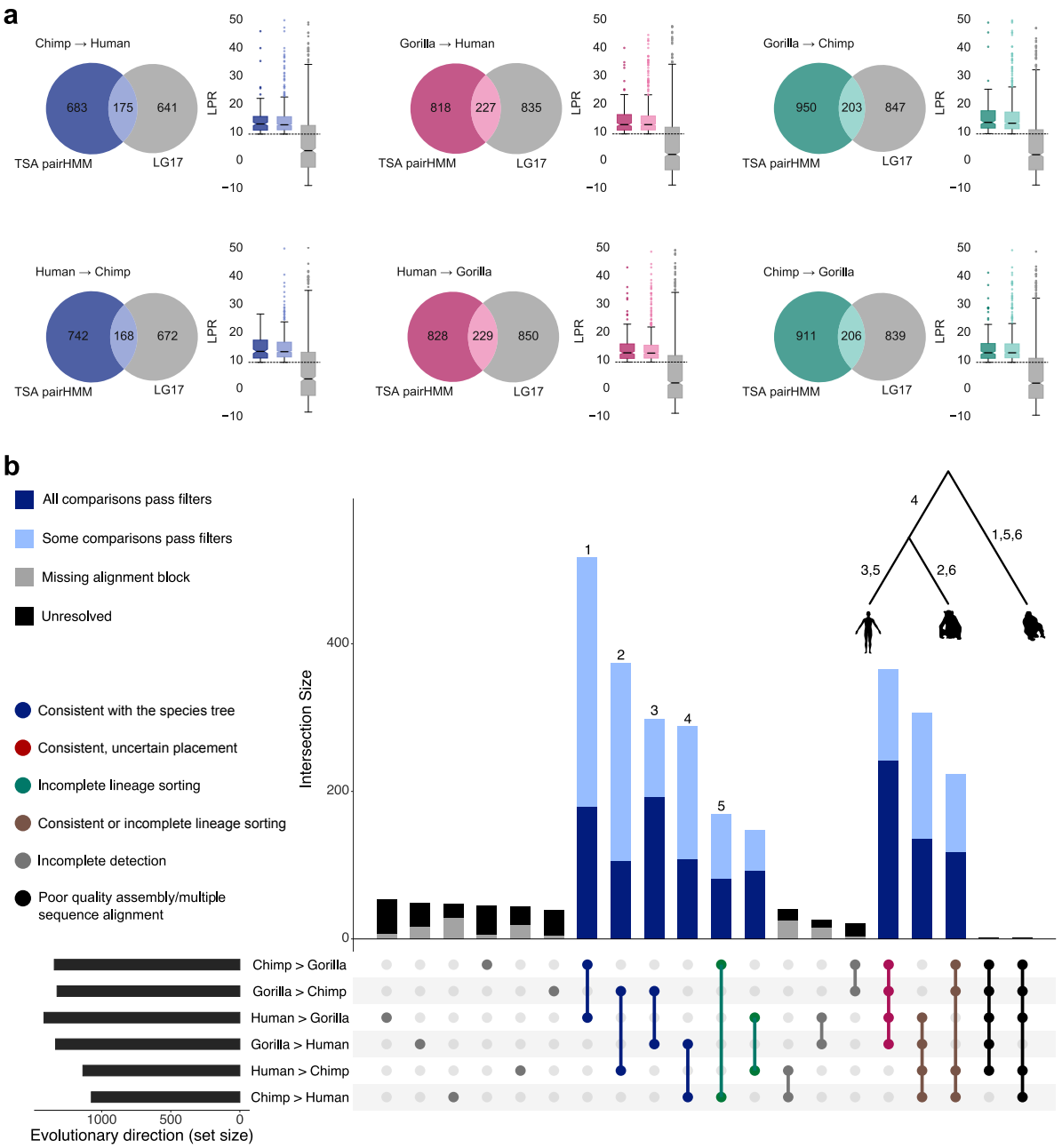

Figure 3.7: Overlap between events identified using my approach and the non-probabilistic model of [185], and the achievable resolution of direction for events identified using this previous approach. **(a)** Intersection between the set of template switch events found using the approach of [185], denoted "LG17", and the significant set of events identified using the TSA pairHMM. Box plots show log-probability ratios for each event set, as well as for candidate events found with both methods. The *y*-axes are limited to 50 for clarity. **(b)** Evolutionary direction for the LG17 event set; annotation as in Figure 3.6, but with an additional category in the dot matrix (shown in black, far right), corresponding to events that are not compatible with a three species tree, likely falling in regions of poor quality sequence assembly or erroneous multiple sequence alignment.

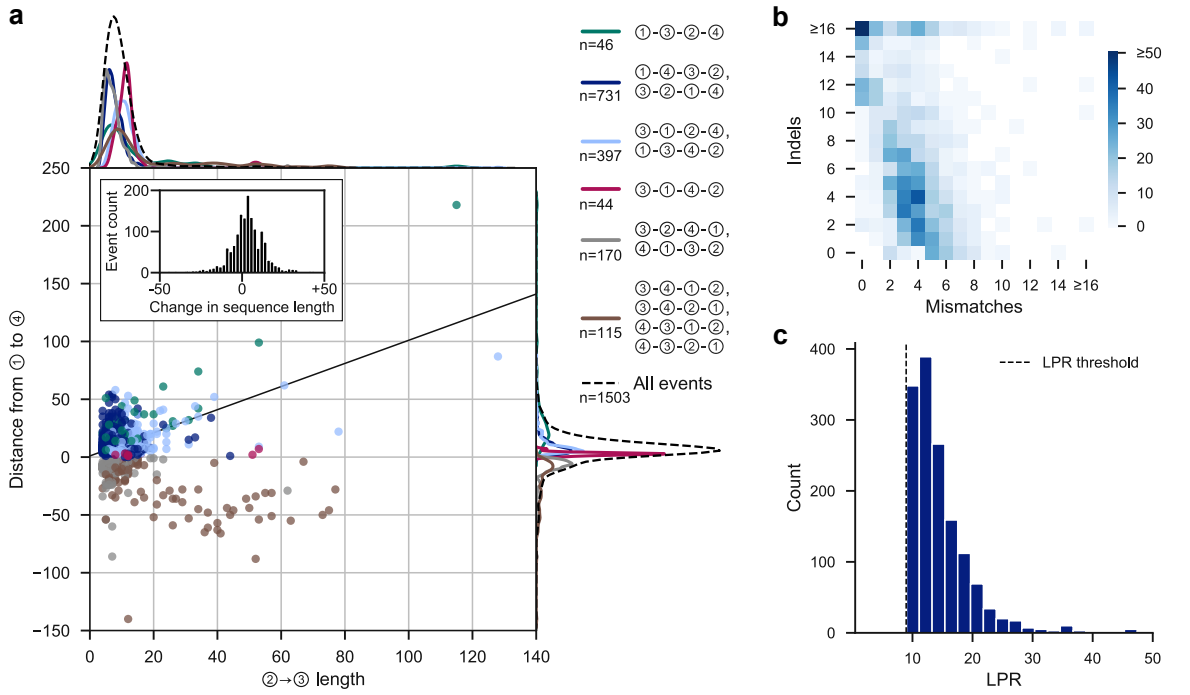

Figure 3.8: Summary statistics for template switch events in the gold-standard set. (**a**) Comparison of ②→③ lengths and the corresponding ① to ④ distances for the gold-standard events. The line $y = x + 1$ corresponds to no net change in sequence length. The inset histogram shows the change in length between the pre- and post-event sequences. Points' colours correspond to event types (legend, right), with the same colours used to show marginal densities at the top and right of the plot. The marginal densities for all gold-standard events (black dashed lines) are drawn on an enlarged scale, for clarity. (**b**) Composition of the template switch-generated mutation clusters in the unidirectional alignments in terms of mismatches and indels. Axes are capped at 16 for clarity. (**c**) LPRs of gold-standard events. The x-axis is capped at 50 for clarity; note that 60 events have a LPR greater than 50. The LPR threshold of 8.95 (Figure 3.1c) is shown as a dotted line. All summaries are derived from the 1503 events which comprise the gold-standard event set, randomly choosing the output of one pairwise comparison per event.

and confidently explains such short templated insertions in the hominids whilst maintaining the ability to capture longer templated insertions (Figure 3.8a, median ②→③ length = 12, median absolute deviation (MAD) = 4.5, max = 128). Few gold-standard template switches leave sequence length unchanged in the descendant species: 65.0% of events increase the length of the post-event sequence, 29.5% decrease the length, and 5.5% cause no net change in length (Figure 3.8a; points below, above, and on the line $y = x + 1$, respectively). Mutation clusters in the input linear alignments which I attribute to template switches events generally contain more than the minimum of two base differences required to initiate a template switch

alignment (Figure 3.8b, median of 10 differences per cluster, MAD = 4.5). Template switch events therefore plausibly explain thousands of mutation clusters and short indel events across the hominid tree that would previously have had either an incorrect or no attributed generative mechanism. The LPR distribution for these alignments indicates that a large number of events fall at the lower end of inferred LPR values (Figure 3.8a), suggesting that if the LPR threshold was relaxed slightly from my conservative choice, the number of unique events discovered could increase considerably. Additionally, many events that are not significant across all comparisons (Figure 3.6, light blue bars) fall only marginally below the LPR threshold due to my heavy penalisation of substitutions in the model. This means that post-event substitutions may have caused non-significance in one or more pairwise comparisons. I did not attempt to relax thresholds to capture more events as significant, as limiting the false positive rate in my gold-standard event set was my primary aim for all downstream analyses. However, combined with the demonstrated inability of my approach to recapture events that are obfuscated by too many additional background mutations (as in my simulations, Figure 3.1a), I further suspect that the overall rate of template switching in hominid genome evolution is greater than reported here.

As described in §2.1.2, for each event, the order of the four switch points facilitates the description of post-event rearrangement patterns and inference of intra-strand and/or inter-strand switching. As I have resolved the evolutionary direction of all events in my gold-standard set, I am able to accurately infer event types and their associated rearrangement patterns. I observed many events that can arise through both intra-strand and inter-strand switching (Table 3.2), and the majority rearrangement patterns (①-④-③-② and ③-②-①-④) generate single inverted repeats (as in Figure 2.1). I also identified many events in which point ④ precedes point ①. Whatever the precise rearrangement mechanism, under the four-point model these events require that the newly synthesised DNA double helix is opened to facilitate the return switch event from point ③ to ④ in a manner conceptually consistent with strand invasion followed by displacement-loop formation in break-induced replication [277]. These rearrangements tend to appear as a single, large insertion in the unidirectional alignment (e.g. Figure 3.9), meaning the approach of [185] cannot capture them as the template switch alignment there was required to contain at least two fewer mismatches than the corresponding unidirectional alignment. My approach of assessing significance using the LPR statistic allows me to omit this filter and facilitates capturing significant events that display these viable rearrangements.

As well as being unable to detect '④ before ①' events, Löytynoja and Goldman [185] assumed chimpanzee represents the ancestral state for every event detected in the human genome. This assumption is incorrect (Figure 3.6) and produced erroneous event type inferences.

Table 3.2: Proportions of gold-standard hominid template switch events corresponding to different event types. Template switch event types are defined by the ancestral switch point ordering, and the ensuing rearrangement patterns observed in the descendant sequences. Some pairs of event types are indistinguishable without knowledge of the direction of replication during which an event arose. I indicate these 'mirror cases' as pairs in parentheses. Events that can arise through intra-strand switching are indicated by a preceding *. See [185] for additional details.

Event type	Rearrangement pattern	Proportion of gold-standard template switch events
(①-④-③-②, *③-②-①-④)	Inverted repeat	0.49
(③-①-②-④, ①-③-④-②)	Inverted repeat with inverted spacer	0.26
(④-①-③-②, *③-②-④-①)	Inverted and direct repeat	0.11
①-③-②-④	Inverted fragment	0.03
③-①-④-②	Two inverted repeats with inverted spacer	0.03
(④-③-①-②, *③-④-②-①), ③-④-①-②, *④-③-②-①	Multiple overlapping inverted and direct repeats	0.08

These methodological artefacts led to other inferences that I now overturn, namely that template switch events occur solely via inter-strand switching and that the generation of a single inverted fragment through ①-③-②-④ events was the most common event type [185].

Using a fully probabilistic approach for template switch event discovery has enabled the identification of ~30% more significant and evolutionarily consistent events than an approach based on a constant scoring scheme coupled with conservative filtering, and has allowed me to assign statistical significance values to events in the final event sets. In addition, defining a gold-standard subset with fully resolved evolutionary directions has allowed me, for each event, to define the ordering of switch points with respect to the ancestral sequence and infer the rearrangement pattern present in the descendant sequence. Using this larger set of significant

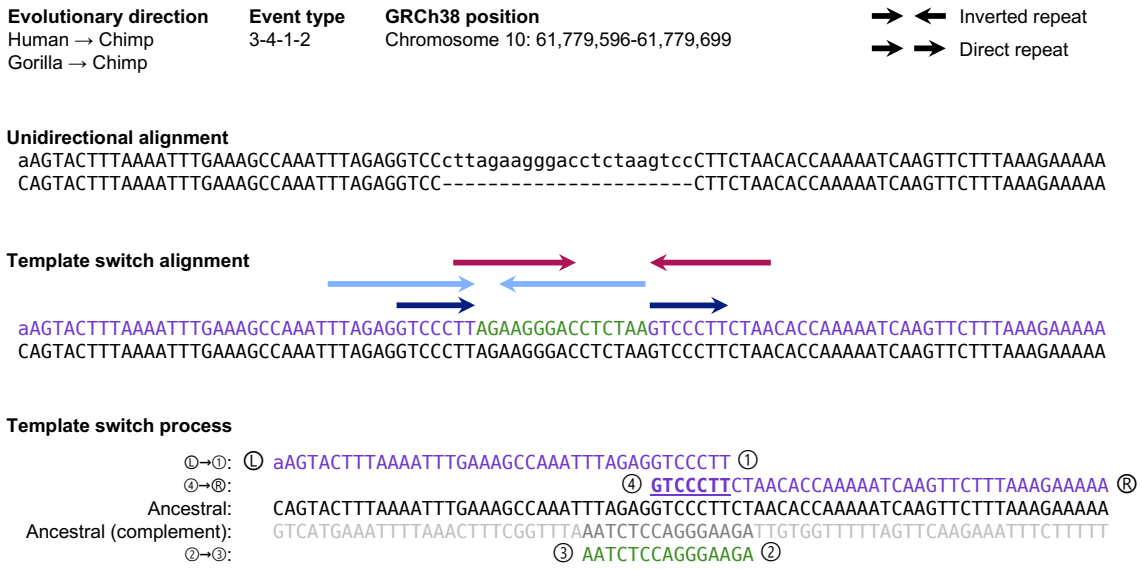

Figure 3.9: Example event in which switch point ④ precedes ①. Figure shows, top to bottom, annotation, linear alignment, template switch alignment and underlying switch process. The bold, underlined region between ④ and ① represents the nascent DNA strand prior to the initial switch event at ①, which typically forms hydrogen-bonded base pairs behind the proceeding replisome, preventing its further involvement in ongoing replication. For events in which ④ precedes ①, a direct repeat generated in the descendant sequence (dark blue arrows above the template switch alignment) indicates that this region was not sequestered from the replisome through base pairing, and facilitated the final ③ to ④ switch event through an open conformation. The mutational consequence of this event is a complicated rearrangement pattern, manifesting as a series of direct and inverted repeats at the sequence level, shown by coloured arrows above the template switch alignment (direct repeats shown as arrows in the same orientation; reverse complement regions shown with arrows in opposite orientation).

events with resolved directions, I can better assess associations between event loci and a variety of genomic features in both the ancestral and descendant species.

3.4 Genomic features associated with event loci

To begin to understand the genomic and sequence properties which may modulate template switch initiation, I investigated associations between functional genomic elements and event loci. I focused on the human coordinates of my gold-standard events, allowing me to use human-specific genomic annotations and experimental data.

3.4.1 Template switches are depleted in protein coding regions and moderately enriched in regulatory sequence regions

I created a set of 13 functional annotations to investigate enrichment and depletion at event loci, as well as processing regions of accelerated evolution in humans from the literature to check for overlaps with events (Table 3.3). As indicated in Table 3.3, several of the functional genomic feature annotations were processed using the procedures of [60]. I performed permutation tests to test for enrichment of these features intersecting gold-standard events, using the coordinate of switch point ① from each event to check for overlaps. Background distributions for each feature were generated using a set of randomly selected coordinates from the genomic background of GRCh38, selected using `bedtools random` [244]. I generated 10,000 random sets of coordinates of length equal to the size of the gold-standard event set, disallowing any coordinates that fall in GRCh38 gap locations. The $\log_2$-fold enrichment is measured with respect to the mean of the genomic background distributions. I determined significant enrichment or depletion by calculating empirical p-values as $(r+1)/(n+1)$, using the procedure of [225], where n is the number of coordinates within each randomly generated set and r is the number of these random sets that intersected with each genomic feature more than the gold-standard event coordinates.

As the apparent mutation clusters generated by single template switch events could generate a signal of species-specific accelerated evolution, I additionally checked whether any of the event-associated mutation cluster coordinates intersected with human accelerated regions (HARs) or primate accelerated regions (PARs) [34, 102, 162, 181, 239] (Table 3.3). Coordinate intersections were examined for the subset of events for which human was determined to match the descendant state using `bedtools intersect`, but I did not include this in the enrichment analysis.

I found a significant enrichment ($p < 0.01$) of events within introns, transcription factor binding sites, and super enhancers (Figure 3.10). It is unsurprising that events occur preferentially within introns whilst being depleted in protein coding regions, in line with purifying selection creating mutation intolerant regions. More interestingly, the enrichment of events within features involved in transcriptional regulation suggests that some of the gold-standard template switch events captured here may have contributed to previously observed high rates of transcription factor binding site and enhancer turnover [76].

I found five events from the unique set within the 2,438 human accelerated regions evaluated, and 11 events within the 5,124 primate accelerated regions evaluated (one and five events, respectively, from my gold-standard set). This makes it clear that template switch mutagenesis

Table 3.3: Details of human-specific genomic features used for hominid enrichment analysis.

Genomic feature	Description
Protein coding regions	Regions with a "CDS" feature annotation in GENCODE v33 [91].
Exons	Regions with an "exon" feature annotation within a protein coding region in GENCODE v33.
Untranslated regions	Regions with either a "three_prime_UTR" or "five_prime_UTR" feature annotation, within protein coding regions in GENCODE v33.
Introns	Protein coding transcripts, excluding exons, processed from GENCODE v33.
Intergenic regions	All regions not annotated as being covered by a gene, processed from GENCODE v33.
Pseudogenes	Regions with a "pseudogene" gene type annotation in GENCODE v33.
lncRNA	Regions annotated as long, non-coding RNA, requiring a "transcript" feature annotation in GENCODE v33.
Promoters	-1000nt to -1 nt upstream of the first position of "transcript" feature annotations that are protein coding, processed from GENCODE v33.
Transcription factor binding sites*	The consensus set of clustered transcription factor binding sites for 161 transcription factors across 91 cell types, released by the ENCODE Project v3 [294]. Sites were required to have a score $>$200 and be present in $\geqslant$5% of cell types ($>$4/91).
EnhancerAtlas enhancers*	Computationally predicted enhancers across 197 tissue/-cell types from the EnhancerAtlas 2.0 database [93] in GRCh37 coordinates, converted to GRCh38 coordinates using liftOver. Enhancers were required to be observed in $>$20% of tissues ($\geqslant$40), requiring a 50% reciprocal overlap of coordinates.
Super enhancers*	Computationally predicted super enhancer regions across 99 tissues from dbSUPER, requiring each region is observed in $\geqslant$5% (5/99) of tissues, converted to GRCh38 coordinates using liftOver.
Human accelerated regions	The union of human acceleration region coordinates from the supplementary information sections of [34, 102, 181, 239], converted to GRCh38 coordinates using liftOver.
Primate accelerated regions	Primate accelerated regions reported in GRCh37 coordinates from the supplementary information sections of [162, 181]. Primate regions corresponding to human, chimpanzee, and gorilla accelerated evolution were kept and converted to GRCh38 coordinates using liftOver.

*These genomic features were processed as in [60].

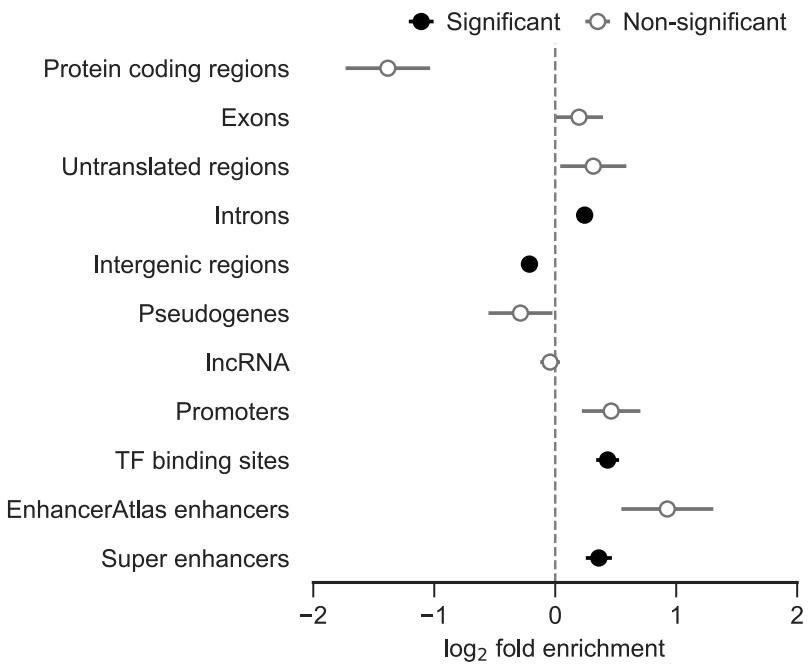

Figure 3.10: An enrichment analysis reveals that gold-standard events are depleted in protein coding regions and moderately enriched in some regulatory sequence regions. Error bars reflect standard deviations of the $\log_2$-fold changes from each test. A significance threshold was set at 0.01 for Bonferroni-corrected empirical p-values.

is not responsible for the majority of mutational patterns interpreted as accelerated evolution regions. However, the detected overlap does demonstrate that caution is required in their interpretation, as complex mutation patterns generated by either a single short-range template switch or a larger-scale replication-based rearrangement mechanism may generate a signal similar to that of lineage-specific accelerated evolution by multiple substitutions and small indels.

3.4.2 Physical properties of the DNA duplex associated with replication stress and structural variation are observed at template switch loci

Focusing on more local sequence features, the physical properties of the DNA duplex such as thermodynamic stability and localised flexibility have been shown to modulate template switch-mediated structural variant formation in larger scale mutational mechanisms [49, 167]. To investigate any such biases which may underlie short-range template switch events, I use my gold-standard event set to analyse the relationship between event loci, physical properties, and local sequence biases.

DNA sequences capable of adopting stable secondary structures such as hairpins are prevalent throughout eukaryotic genomes. These structures are particularly prone to form when DNA is exposed as a single strand during replication, and once formed can cause fork stalling and strand dissociation [215]. I therefore investigated whether the initiation of template switches at ① is biased by local DNA secondary structure stability.

For each gold-standard event, the sequence region ±500nt around switch point ① was extracted for the ancestral and descendant sequences, giving sequences of length 1001nt. I focus on ① as I assume any local genomic features will be associated with the site of the initial switch event. DNA secondary structure prediction was performed using `RNAfold` version 2.4.1 from the ViennaRNA Package [288], using a sliding window of size 50nt along these sequence regions and a step size of 1nt. Energy parameters for single-stranded DNA were used (`--noconv` and `--paramfile`), allowing G-quadruplex formation prediction (`--gquad`), but disallowing lonely (helix length 1) and GU wobble base pairing (`--noLP` and `--noGU`). Minimum free energy (MFE) secondary structure prediction was performed using the command:

```
RNAfold --noLP --noGU --gquad --noconv --paramfile=dna_mathews2004.par
```

For comparison with a genomic background, I randomly drew 10,000 1001nt regions from GRCh38 and performed the same analysis.

GC content heavily impacts the stability of potential DNA secondary structures, as the A:T base pair is less thermodynamically stable than C:G [72]. I therefore regress GC content out of calculated free energies for all MFE structures to identify regions of stable structure independent of underlying GC content. My sliding window approach assesses sequences of length 50, so each additional G or C nucleotide increases GC content in any window by 2%. Therefore I randomly generated 10,000 nucleotide sequences of length 50 for each possible GC content, 0%, 2%, 4%, …, 100%, and calculate the average MFE for each of these set of sequences. The free energies of all MFE structures in the above sliding windows are then adjusted by calculating the GC content of each window and subtracting the corresponding average GC content free energy as determined using the randomly generated sequences.

I observed two interesting signals of secondary structure stability within these regions. First, secondary structures are significantly less stable in regions flanking ① for both the ancestral and descendant sequences compared to a random genomic background (Figure 3.11a, $p < 2 \times 10^{-16}$, Wilcoxon rank-sum tests). This may be a residual effect of the greater AT content in these regions compared to the random genomic sample (see §3.4.3), as the A:T base pair is less thermodynamically stable than C:G [72]. Second, there is a striking increase in descendant secondary structure stability in the immediate vicinity of ①, and a smaller but noticeable increase in ancestral secondary structure stability across similar positions (Figure 3.11a). It is

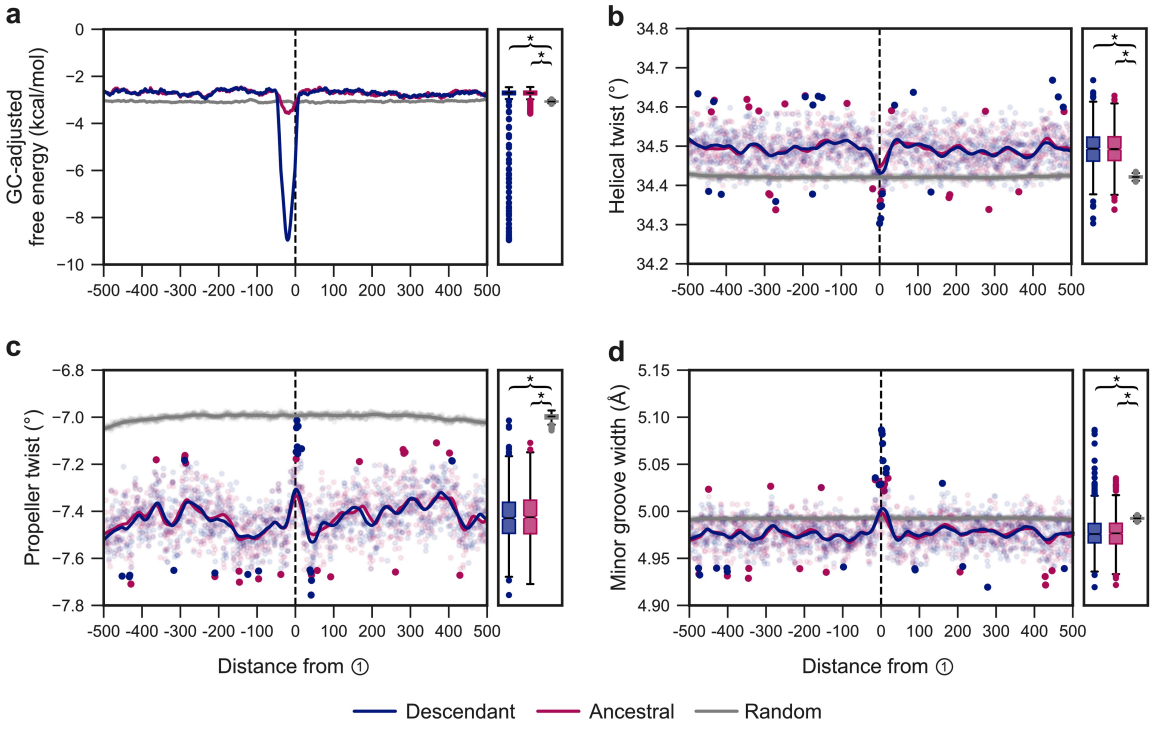

Figure 3.11: Short template switches generate non-canonical DNA structures and are associated with atypical patterns of DNA bending. (**a**) Mean GC content-adjusted free energies of the MFE secondary structures for the ancestral and descendant sequences from the gold-standard event set, compared to a random genomic background ±500nt around switch point ① using a left-aligned sliding window size of 50 in single nucleotide steps (e.g. at position -500, the MFE structure is calculated using the sequence from position -500 to -451). Marginal box plots summarise the distributions of mean values within the ±500nt region, and brackets indicate significantly different MFEs ($p < 2 \times 10^{-16}$) between groups under a Wilcoxon rank-sum test. **(b, c, d)** Mean predicted helical twist, propeller twist and minor groove width ±500nt around switch point ①. Points represent mean feature values as calculated using DNAShapeR [58], utilising a pentamer sliding window centred on each position, and a Loess fitted curve is overlaid. Additionally, the smallest and greatest 1% of mean values are shown as solid points to highlight extreme values. Box plots as in (a).

unsurprising that I observe such stable structures in the post-event descendant sequences, as the template switching process implicitly generates regions of nearby perfect inverted repeats (e.g. Figure 2.1b) which are prone to forming the hairpin and/or cruciform structures that constitute highly stable DNA secondary structures [258]. In the ancestral sequences, the smaller decrease in observed free energy around ① is reflective of pre-event potential for structural formation in a subset of events, suggesting that some events may involve hairpin-mediated quasipalindrome-to-palindrome conversion as in the original mechanism proposed

for bacteria [253]. Regardless of ancestral stability, the spontaneous creation of sequence regions capable of forming stable secondary structures is of note, as small regions of stable structure play a role in several biological processes [36, 286], and regions of similarly stable structure can cause replication fork collapse and DSB formation, and trigger genome instability [287, 302].

Regions capable of forming stable secondary structures within AT-rich sequences are abundant across chromosomal fragile sites throughout the human genome and typically display increased DNA duplex flexibility [40]. In addition, increased duplex flexibility is observed immediately at the breakpoints of some large-scale mechanisms of structural variant formation in the human genome [167], and I suspected that atypical patterns of flexibility may be observed at event loci.

Using my gold-standard events, to assess the flexibility of the DNA molecule around location ①, I calculated minor groove width, helical twist and propeller twist in these regions, as well as for 100,000 uniform random sampled 1001nt sequences from across all GRCh38 chromosomes. I use the R/Bioconductor DNAShapeR package for these calculations [58], which is based on the method of [326] for predicting DNA structural information. This approach utilises a pentamer sliding window to calculate each feature as determined through previous Monte Carlo simulations [326], which accounted for sequence context of the focal nucleotide within the window. As above, this analysis was repeated for 10,000 randomly selected regions from GRCh38 for comparison.

Helical twist angle, a measure of the inter-bp rotations with respect to the DNA helical axis, is significantly greater in both the ancestral and descendant sequence regions surrounding event loci ($p < 2 \times 10^{-16}$, Wilcoxon rank-sum tests), with a spike immediately around switch point ① (Figure 3.11b). I also observed a significant decrease in propeller twist, a measure of the inter-bp plane angles, in the vicinity of event regions ($p < 2 \times 10^{-16}$), with an increase at switch point ① that does not reach parity with genome-wide mean values (Figure 3.11c). Deviations in propeller and helical twist values from those of B-DNA is indicative of DNA bending [256]. Interestingly, DNA bending has been shown to facilitate the error-free template switching DNA damage tolerance pathway in yeast, facilitated by the high mobility group protein Hmo1 [105]. While distinct from the process I model here, the mechanistic similarity between these local template switch mechanisms coupled with my predictions of non-B DNA values of helical and propeller twist suggests that a propensity for DNA bending may indeed have helped facilitate template switch events in my gold-standard set.

Lastly, I also observed more narrow minor groove widths within the flanking sequence regions around ① compared to the genomic background level (Figure 3.11d). Decreased minor

groove width has been shown to confer resistance to DNA damage by limiting accessibility of the DNA to reactive oxygen species [45, 133]. It is conceivable that a widening of the minor groove, as observed immediately at ①, may likewise cause increased rates of DNA lesion formation that could be bypassed through a template switch process to restart a stalled replication fork, but it is difficult to confidently draw such a conclusion without supporting experimental observations.

3.4.3 Poly(dA:dT) tracts are enriched at event loci

The structural features identified around event loci consistently show the hallmarks of AT-rich and poly(dA:dT) tract DNA, which are associated with large negative values of propeller twist and a narrowing of the minor groove [75]. To identify the prevalence of poly(dA:dT) tracts, and any additional sequence motifs which may contribute to event formation, I searched for significantly enriched DNA motifs in a region ±150nt around switch point ① in the ancestral sequences of the gold-standard event set.

I generated position weight matrices for significantly enriched sequence motifs using the differential enrichment objective function (`-objfun de`) in the multiple expectation maximization for motif elicitation (MEME) suite [23]. For every event in the gold-standard event set, sequence ±150nt around switch point ① was searched for motifs in both the ancestral and descendant sequences. If more than one ancestral or descendant sequence was available, chimpanzee and human sequences were used, respectively. Event regions were compared against a global genomic background set of 30,000 301nt sequences, using 10,000 randomly sampled sequences from each of the human, chimpanzee and gorilla genomes, excluding regions containing masked bases or gaps. As I sought to identify individual putative causative motifs per sequence, I allowed one or zero occurrence of each motif per sequence (`-mod zoops`). This means that I assess the enrichment of single motifs in each sequence tested, rather than enrichment of many occurrences of (e.g.) a repetitive low-complexity element. I repeated this analysis for three ranges of window sizes: 6–10nt, 10–20nt, and 20–50nt, where window size defines the minimum (`-minw`) and maximum (`-maxw`) allowed length of the motif. The analysis was performed as follows:

```
meme event_sequences.fa -dna -nostatus -mod zoops -minw {6, 10, 20} \
     -maxw {10, 20, 50} -objfun de -neg background_sequences.fa \
     -revcomp -seed 42
```

The E-value cut-off for significant enrichment was set at $E \leqslant 10^{-6}$.

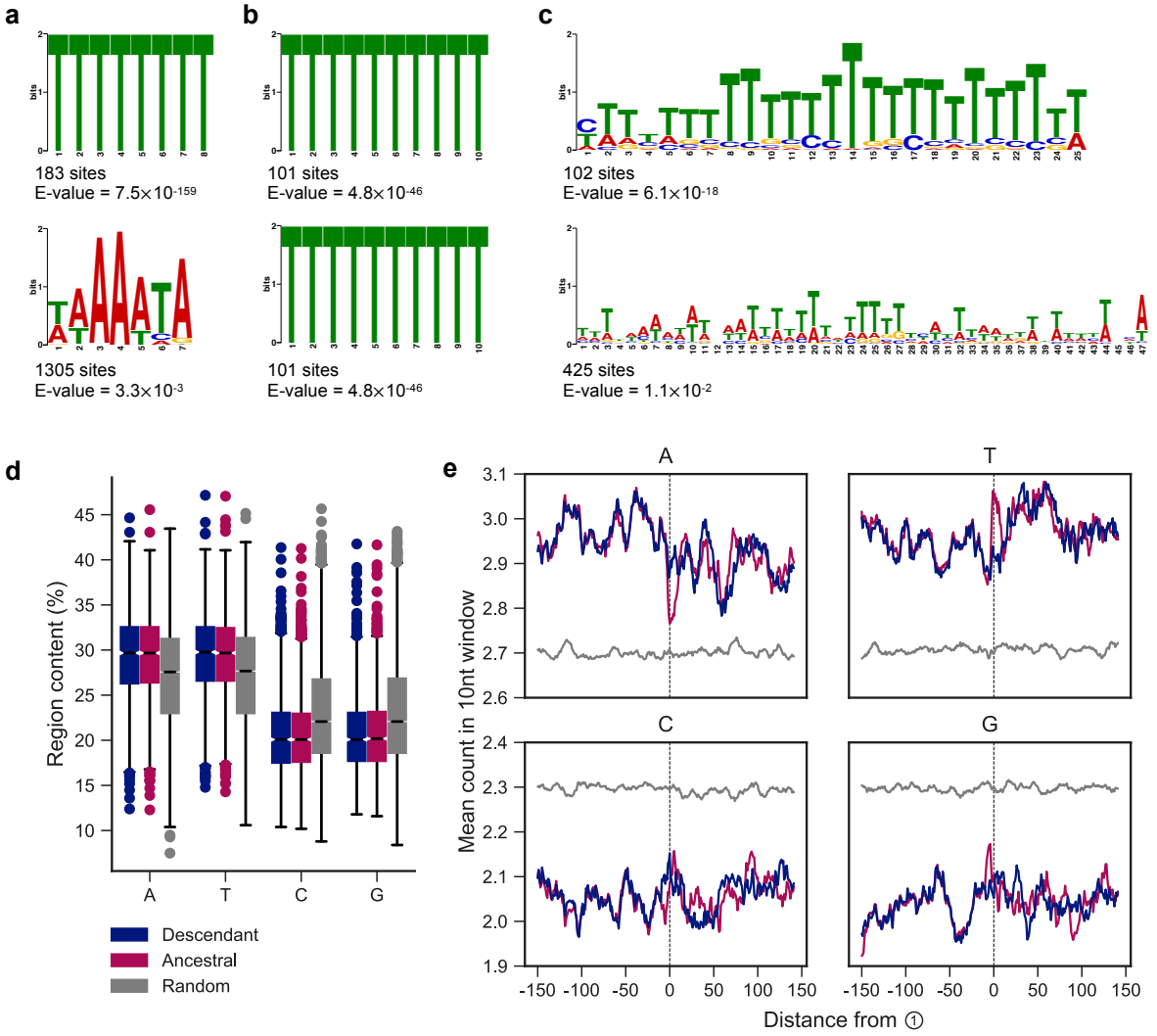

Figure 3.12: Event loci are enriched for poly(dA:dT) tracts and are observed more frequently in AT-rich genomic regions. (**a**, **b**, **c**) Enriched sequence motifs within ±150nt of switch point ① for the gold-standard events, compared to a random genomic background sampled from GRCh38. The most significantly enriched motifs (lowest *E*-value; top row) and most frequent significant motifs (bottom row) within ±150nt of ① for gold-standard events. Motifs were tested for enrichment at three motif size ranges: (a) 6–10nt, (b) 10–20nt, and (c) 20–50nt. In (b), note that for the 10–20nt motif search the same motif (T_{10}) is both most significant and most numerous. (**d**) Percentage of each nucleotide in the ancestral and descendant sequence region, compared to a random genomic background. Percentages are calculated in a region ±150nt around ① loci; to form my random background distribution, 10,000 regions of 301nt were randomly drawn from each of the human, chimpanzee, and gorilla genomes. (**e**) Counts of each nucleotide in a left-aligned single nucleotide sliding window of 10 bases, averaged across descendant, ancestral and randomly sampled sequences at each position.

Several significantly enriched A and T-dominant motifs were identified across all tested motif sizes (Figure 3.12a–c). In addition, regions surrounding template switches are generally enriched for AT content compared to a randomly sampled genomic background (Figure 3.12d,e). The most significantly enriched motifs of each size are T_8 (183 events, $E \approx 0$), T_{10} (101 events, $E \approx 0$), and YT_2YT_{21}[1] (102 events, $E \approx 0$). In no tests did a motif with greater sequence complexity appear as more significantly enriched than AT rich sequence alone, suggesting that poly(dA:dT) tract DNA plays a more important role in event initiation than any more complex template switch associated motif. It is well established that such poly(dA:dT) tract DNA consisting of $\geqslant$4–6 consecutive A:T base pairs causes intrinsic bending of the DNA molecule [25, 155]. Supported by my predictions of increased flexibility around ① in my gold-standard event set (Figure 3.11b–c), I suggest that sequence-directed bending of the DNA molecule may occur around the initial switch event, similar to that of Hmo1-mediated bending in DNA damage tolerance template switching pathway in yeast [105]. In addition, poly(dA:dT) tracts are known sites of preferential fork stalling and collapse due to elevated rates of DSB formation [298]. The enrichment of these motifs supports the notion that short-range template switching may either be involved in fork restart during DNA lesion bypass, or may occur post-replication in a similar fashion to large-scale structural variant formation in the presence of DSBs caused by persistent lesions unresolved by repair pathways [49].

3.4.4 A summary of factors influencing template switch formation in great ape genomes

In combination, the sequence biases and physical properties surrounding event loci indicate that the gold-standard events captured by my model preferentially occur in regions that are prone to replication stress, as previously outlined for well-established mechanisms of larger scale structural variant formation [49]. This validates the events identified as significant using my approach, and confirms that my method provides a previously unachievable resolution in the capture and description of small-scale replication-based rearrangements in their evolutionary context.

[1] Following the International Union of Pure and Applied Chemistry (IUPAC) ambiguity notation, Y is a pyrimidine (C or T).

3.5 Conclusions

I have identified thousands of statistically significant template switch-mediated mutations across the great ape tree, demonstrating the power of pairHMMs for confidently detecting a class of rearrangements which are traditionally difficult to model. By capturing and assigning an evolutionary direction to many of these events, I am able to explain the presence of thousands of short indels and complex mutation clusters in the evolutionary history of the hominids. My approach appears robust to selected parameter values over these timescales, and represents a methodological improvement over a previous non-probabilistic method [185] for modelling short-range template switch mutations in an evolutionary context. By shifting to probabilistic thresholds and assigning statistical significance to individual events, I have achieved superior recall and a consequent improvement in statistical power for identifying associated genomic features.

A limitation of my method is that many events that are characterised by the conversion of a near-perfect inverted repeat into a perfect inverted repeat are classed as non-significant. This quasipalindrome-mediated mutational pattern is the hallmark of a traditional prokaryotic template switch event [253]. However, such events often produce few changes in a unidirectional alignment between the pre- and post-event sequences, in many cases generating solely the minimum of two nucleotide differences that I require to initiate a local realignment under my models. Correcting two-nucleotide differences will not yield a significant LPR, regardless of the length of pre-existing reverse-complement identity (the potential ②→③ fragment) that the two nucleotides are contained within. While many such mutations may indeed have arisen through legitimate template switch processes, my statistical method cannot report these as robust, statistically supported events in preference to the null hypothesis of simple background mutation. I therefore did not attempt to incorporate these into my final event set, as my priority was to minimise the number of false positive events in my gold-standard event set, rather than maximising the total number of events discovered. To study the prevalence of template switches characterised by (e.g.) 2nt differences in future work, I would need to perform model comparisons using additional factors beyond just the LPR test statistic. For example, I could assess if candidate events fell preferentially within regions of existing local inverted repeat sequence.

Despite this conservative approach, I have described more events than have been reported previously and can be more confident that the template switches I report represent the true mutational history underlying their associated mutation clusters within linear alignments of the great apes. In future, it would be possible to increase the number of events included in

the gold-standard set which were used for downstream analyses, by resolving the subset of events which are reversibly detected (e.g. Figure 3.5). As mentioned in §3.3.2, this could be achieved through further outgroup comparisons. For example, if an event was identified in the gorilla $\leftrightarrows$ human and chimpanzee $\leftrightarrows$ human comparisons, and the gorilla and chimpanzee sequences represent the true ancestral state, we would expect no event to be identified when comparing these sequences to the orangutan genome. Similarly, if human is the true ancestral state, no event should be found between human and orangutan. As these cases only represented the minority of events characterised in this chapter, I forewent this analysis, but it would be relatively straightforward to phylogenetically place these template switches if desired in future.

It is important to emphasise however that care is always required when inferring the mutational history underlying mutation clusters such as those explored here. Other well-characterised mutational mechanisms frequently generate small mutation clusters in eukaryotic genome evolution, such as the multinucleotide substitutions caused by error-prone polymerase activity [33, 117, 201, 271]. However, my requirement for a high-homology, reverse-orientation template within 100nt of each focal mutation cluster, coupled with my strict statistical thresholds, demonstrate that a mutation involving a template switch is the most parsimonious explanation for the clusters explored here. I also suspect that the number of events reported here is an underestimate of the true extent to which short-range template switches have shaped the evolution of the hominid genomes.

My emphasis on reducing false positives has enabled more confident delineation of physical properties around event loci. It was previously reported that template switch events generate regions with greater energetic potential for DNA secondary structure formation [185], and I have shown this holds in my direction-resolved gold-standard event set. I speculate that an increased potential for fork-stalling secondary structural formation would also be observed in the ancestral species if I did not filter out many of the events involved in quasipalindrome conversion. Nonetheless, it has been demonstrated that stable secondary structures can be bypassed to restart a stalled replisome, through the recruitment of error-prone polymerases and the initiation of template switch-mediated DNA synthesis [226]. I therefore suggest that this signal should still be investigated when considering mechanisms which may underlie short-range template switch initiation in future work. More importantly for events identified using my approach, event formation appears to be associated with an excess of poly(dA:dT) tracts which are known replication barriers that can cause fork collapse [298], as well as non-B duplex geometry around event switch points and signals of helical bending which could lead to an increased potential for DSB formation at the initial disassociation site.

A consideration regarding the events I have described here is the signals such rearrangements could create in evolutionary analyses. I identified events both in curated regions of human accelerated evolution [239] and in elements involved in transcriptional regulation which are thought to be subject to high rates of evolutionary turnover [76]. In both cases, observed signals of evolutionary importance, typically interpreted as consequences of a high rate of change, could feasibly be generated by a single complex mutational event such as a template switch. I do not claim that the template switch mutations outlined here underlie regions of accelerated evolution, as I observed few intersections with such regions. However, my observation of some intersections between template switch loci and these regions still demonstrates that care is required when interpreting signatures of high turnover or accelerated evolution, and individual examples should be considered in light of this finding.

The short-range template switch events and associated features described here were identified by focusing on local template switching, as it allows me to assign enough statistical significance to individual events to distinguish candidate events from accumulated substitutions and/or short indels. While this represents a significant methodological improvement and the most comprehensive delineation of these events in the hominids to date, it does leave the characterisation of small-scale, non-local template switching unresolved. This will remain the case unless methods for the direct observation of these events are developed.

In conclusion, this chapter has demonstrated that my methodology based on pairHMM comparisons can be used to effectively identify significant template switches in a phylogenetic setting, permitting the most extensive delineation of these events in the evolution of any set of related genomes to date. In the next chapter (Chapter 4), I will show how a workflow which utilises my models can be applied to sequencing data from human populations and family trios to delineate the human germline landscape of template switch events.

Chapter 4

The human population landscape of short template switch mutations

Chapter overview

Having characterised between-species short-range template switch mutations in great ape genome evolution in Chapter 3, in this chapter I now apply the models from Chapter 2 to identify within-species template switches in population-scale human variation datasets. I characterise the prevalence of events in human populations and the associated population genetic properties, I explore summary statistics and the footprints that template switches leave in variation calls, and I outline genomic features associated with events using an expanded set of genomic features that includes human-specific experimental annotations. I also assess evidence for *de novo* template switch mutagenesis in family trios.

Declaration

The content of this chapter has not previously appeared elsewhere. I performed all data collection, processing, analysis, and data visualisation.

Code and data availability

All code underlying the analysis of this chapter, in addition to any supplementary data files, are available from:
https://gitlab.com/conorwalker/phd_thesis/tree/main/chapter_4.

4.1 Background

Germline mutations within population variation datasets are defined with respect to a reference genome, typically called as sets of SNVs, short indels that impact fewer than 50 nucleotides,

and structural variants that impact $\geqslant 50$ nucleotides [80, 191, 284, 293] (see also §1.2.4). This chapter will concern itself with germline mutations within human populations. Accurately identifying between-human variation at the population scale has enabled a greater understanding of human demographic history [282], the ancestry and relationships between diverse populations [31], ongoing selective pressure [223, 270], and identification and interpretation of causal variants underlying complex traits and diseases [14, 186, 198, 314]. These downstream analyses are only possible due to the availability of several high quality catalogues of human genetic variation, such as those released by dbSNP, the International HapMap Project, the 1000 Genomes Project, and gnomAD [131, 141, 273, 293]. Similarly, this chapter aims to catalogue short-range template switch mutagenesis within human populations, so as to understand the role of these mutations in ongoing human genome evolution. To identify short-range template switch mutations in human populations, I will primarily focus on data made available as part of the 1000 Genomes Project [42, 293]. This represents the largest publicly accessible collection of whole genome sequencing data currently available and is a standard resource in the human genomics community for interpreting genetic variation across geographically diverse populations.

A typical high-coverage, short-read whole human genome sequence contains an average of 4×10^6 SNVs, 4.2×10^5 short insertions, 4.5×10^5 short deletions, and 9.2×10^3 structural variants relative to GRCh38 when sequenced using short read technologies and assembled to current state-of-the-art standards [42]. Somewhere within this variation should exist evidence of template switch mutagenesis. This was demonstrated for a small number of candidate template switch loci by Löytynoja and Goldman [185] by first performing whole-genome alignment between a single human genome [175] and a previous reference human genome (GRCh37) each in FASTA format, and then identifying events under their simpler model (see §2.1.3). Of the 76 template switch events identified between two humans by Löytynoja and Goldman [185], 35 events were present as a combination of variants in the initial low-coverage 1000 Genomes Project variant calls. It should therefore be possible to identify complex mutation clusters and thus template switch mutations directly from these readily-available variant calls, foregoing a computationally expensive whole-genome alignment procedure and directly permitting the identification of events at a population scale.

Consider the types of footprint a template switch may leave in population-scale variant calls, as these determine how events can be identified and alternatively represented. Similar to FASTA-represented sequence alignments, there are four foreseeable event footprints within VCFs: a cluster of SNVs, a single short indel (e.g. a single insertion corresponding to the content of ② → ③), a combination of SNVs and indels, and a single structural variant. I will

in turn address the importance of each of these possible footprints as they relate to template switch identification and interpretation in between-human variation datasets.

Clustered SNVs are typically referred to as multinucleotide variants (MNVs) and within the context of human germline mutagenesis have been defined as SNVs which occur within 10 [306] or 20nt [140] of each other, appearing on the same haplotype. The mutational mechanisms underlying MNV formation are somewhat understood. MNVs impacting adjacent nucleotides have consistently been associated with an enrichment of GC→AA and GA→TT dinucleotide substitutions, characteristic of the error-prone DNA polymerase Pol-ζ mutational spectrum [117, 140, 306]. Wang *et al.* [306] also highlight that there are an excess of apparent AA→TT, AT→TA, and TA→AT MNVs within repetitive contexts in human genomes which are actually generated by consecutive deletion then insertion events. Further, MNVs in perfect linkage but separated by more than one nucleotide have been associated with the APOBEC cytosine deaminase mutational signature [140]. While my great ape analysis showed that many template switches produce indels in addition to SNVs in the unidirectional representation of the event, some template switches do indeed leave a solely MNV footprint (see Figure 3.8b). Similarly to how MNVs are typically represented in standard variant calls as a combination of SNVs (requiring the use of non-standard additional tools to identify the presence of a MNV [147, 310]), it is possible that template switch mutations could be misrepresented as MNVs. Any template switches I subsequently identify which are associated with a solely MNV cluster should therefore be considered carefully for these established signatures.

The majority of short indels in human genomes occur within homopolymer runs and tandem repeats and are primarily caused by polymerase slippage events [218]. Both homopolymer expansion and contraction (e.g. GAAAT→GAAAAAT and GAAAAAT→GAAAT) and tandem repeat expansion (e.g. GCAGCGCAGC→GCAGCGCAGCGCAGC) are well-defined, distinct from template switch mutagenesis, and most events fall within low-complexity regions that I filter out across all analyses (see §2.3.4). Around 2.5% of indels observed in human populations both occur outside of these repetitive contexts, and are not single short deletions [218] (which are again filtered from my analyses, see §2.3.4). These remaining indels are the potential template switch footprints of interest, and indeed Montgomery *et al.* [218] suggested that the fork stalling and template switching (FoSTeS) and microhomology-mediated break-induced replication (MMBIR) pathways, typically only associated with structural variant calls, may be responsible for many such small indel events detected in human genomes.

The mechanisms which give rise to clusters containing both SNVs and indels (i.e. "complex" mutation clusters) within human populations are far less studied. Excluding the small investigation into a single sample by Löytynoja and Goldman [185], I have been unable to

find any comprehensive investigation into their occurrence in human variation datasets in the literature. All complex mutation clusters are therefore of interest, given that a large proportion of the template switches observed in my great ape analysis are associated with multiple SNVs and indels in the associated unidirectional alignments (Figure 3.8b).

The remaining possible footprint involves single structural variants, which will receive less attention here as the majority of this chapter will focus solely on SNV and indel calls. Structural variant callers already seek to identify a potential alternate template that is consistent with the observed reads using strategies involving breakpoint assembly for short reads [3, 53, 167, 284]. Once identified, the potentially replication-based causative mechanisms (e.g. FoSTeS and MMBIR) can be attributed based on patterns of homology around the identified breakpoints. As in my great ape analysis however, it is important to keep the footprints of these mechanisms in mind to assess consistency with short-range template switches and thus identify possible generative pathways.

These mutational footprints will be the target of study for the remainder of this chapter. I aim to generate a robust catalogue of small-scale template switches in a large collection of human genomes and provide evidence for their occurrence in single generations. This will be the first systematic assessment of template switch activity in a large collection of human genomes down to a < 50nt resolution, challenging the operational definition of a structural variant used by methods that identify rearrangements in typical studies of human genetic variation [42, 284]. This will allow short template switches to be considered alongside traditionally-studied forms of human genetic variation at the population-scale for the first time. With this catalogue of small-scale template switches, I will investigate the population genetics of events, where I expect to observe population-level distributions in line with those observed for typically-studied classes of variation [42, 191, 284, 293]. I then want to assess evidence for the potential involvement of the FoSTeS/MMBIR pathway(s) (the primary signals for which are distinct replication timing profiles and microhomology at associated break points, see above and §1.2.4), as these are thought to cause many of the complex rearrangements that arise through template switching at a $\geqslant 50$nt scale. Finding evidence of these pathways would provide insight into the scale at which these specific, disease-associated [47, 172] pathways can operate. Finally, as in Chapter 3, I also want to understand the genomic/sequence features which may modulate event formation, such that the predisposition of short template switch initiation for any given genomic region may be better understood. Unlike in an evolutionary context, I am also able to ask if any of the variants that I attribute to template switching have any known associations with traits of interest as identified by previous GWAS investigations.

This will allow me to ask if short template switches possibly contribute to clinically-relevant phenotypes.

To achieve these aims, I will first provide details on the variant call datasets which are used to identify and characterise template switch mutations in human populations, discussing why each dataset is suitable for event discovery (§4.2). Similarly to §3.2, I then detail the procedure for establishing statistical significance threshold under my LPR test statistic (Equation 2.9) for individual events (§4.3). I then provide an overview of the pipeline used to discover events from input variant call format (VCF) files (§4.4), describing the prevalence and population genetic properties of events in human populations (§4.6), exploring the associated genomic and sequence features (§4.7), and finally considering evidence for *de novo* template switches from single-generation data (§4.8).

4.2 Datasets used for template switch event discovery

To identify events, I leverage several publicly available population and *de novo* variant call datasets which commonly act as community resources for testing hypotheses in human genomics. The datasets used are summarised in Table 4.1. The majority of my analysis makes use of data generated as part of the 1000 Genomes Project, which has been the reference for global human genetic variation since its "phase 3" release in 2015 [284, 293, 325]. As a result, this resource has continued to receive updated sequencing and analysis, and I specifically make use of datasets produced by two of these updated analyses. I additionally make use of *de novo* variant calls from parent-offspring trios in the Icelandic population.

The first dataset, denoted 1k-30x here, consists of genotyped and statistically phased SNV and indel calls from 3202 samples (containing 602 family trios), obtained via short-read sequencing to a targeted depth of 30x coverage by the New York Genome Center [42]. This variant callset is statistically phased with pedigree-based correction using SHAPEIT2 [70] for autosomal variants and Eagle2 [184] for variants on chromosome X, and singleton variants (allele count = 1) are not included in the phased callset. While I could instead opt to work with the singleton-resolved genotype calls, I use the doubleton-resolved phased callset as I want to ensure that variants observed in a mutation cluster (which I use to find evidence of template switch mutagenesis) are present on the same chromosome copy. Identifying events using this dataset is ideal as it is based on (short-read) high-coverage re-sequencing of a well-studied set of human samples, so any inferences made here are more easily contextualised in terms of established variant summary statistics and expected patterns of diversity between these populations.

Table 4.1: A summary of the human genetic variation datasets used in Chapter 4.

Dataset ID	**Description**	**Samples**	**Trios**	**Sequencing technology**	**Refs**
1k-30x	1000 Genomes Project, 30x Coverage	3202	602	Illumina NovaSeq 600	[42]
1k-HGSVC	1000 Genomes Project, Human Genome Structural Variation Consortium	35	3	Illumina NovaSeq 600, PacBio CLR, PacBio HiFi, Strand-seq	[42, 80]
Ice-Trios	deCODE genetics/Amgen 35x Icelandic Trio Sequencing	1548	1548	Illumina $GAII_x$, HiSeq 2000, HiSeq 2500, and HiSeq X	[137]

Some concerns were raised by Löytynoja and Goldman [185] about the ability of short-read, mapping-based assembly methods to detect short template switch mutations. This was likely due to the quality of variant call data released by the 1000 Genomes Project at the time, and I am not overly concerned that this is an issue with the updated 1k-30x calls for two reasons.

First, the mean ② → ③ length identified from my great ape analysis is 12nt (Figure 3.8a). The 1k-30x dataset was produced using 150bp paired-end reads, so the entirety of the novel sequence expected to be introduced by the ② → ③ region of most template switches should easily fit within a single short read with a sufficient length of flanking sequence to allow event-containing reads to map to the reference genome. In addition, the decreased coverage at variant positions associated with the events identified by [185] was observed in the 1000 Genomes Project phase 3 variant calls [293], which were produced from reads with a much lower mean sequencing depth of 7.4x, and shorter 76 or 101 bp paired-end reads [15, 293]. I expect most template switch mutations to be in non-coding regions of the genome, most of which were sequenced to an average depth of 4x or less in the phase 3 release. It is difficult to call rare variants confidently at such low levels of coverage, as they can be both easily filtered out as possible artefacts by quality control pipelines or called with erroneous genotypes [27]. The 150bp read length, along with the respective mean and minimum sequencing depths of 34x and 27x coverage used to produce the 1k-30x calls should address these issues.

Second, the pipeline used for processing the short reads in 1k-30x is highly suitable for resolving the types of complex mutation clusters (multiple SNVs and/or indels within a

small region) I expect to find as template switch footprints in variant calls. A fairly typical pipeline for variant discovery from short reads was utilised by the New York Genome Centre, involving mapping to a reference human genome (GRCh38), and using HaplotypeCaller [237] from the Genome Analysis Toolkit (GATK) to identify variants within these mapped reads. HaplotypeCaller is particularly suited to calling complex mutation clusters, as it performs *de novo* assembly of reads within genomic segments called "active regions", which are defined based on a greater than expected number of SNV and indels within a small genomic region. Within each active region, HaplotypeCaller builds a de Bruijn-like graph for reassembly to identify possible alternate haplotypes. It then aligns each read against each possible alternate haplotype using a pairHMM, producing likelihoods for each allele at each potential variant site, and then applies a Bayesian procedure to calculate sample-wise likelihoods of each genotype given the read data for that sample. In combination, there is no foreseeable reason why I should not be able to call template switch mutations from these data.

The other two datasets I will use to search for evidence of *de novo* template switch mutations. The first, denoted 1k-HGSVC here, consists of SNV, indel, and structural variant (SV) calls for 35 individuals from the 1000 Genomes Project cohort, 32 of which are unrelated, and three children from family trios [80]. This callset was generated using continuous long-read sequencing and/or high-fidelity sequencing — two long-read sequencing technologies from Pacific Biosciences (PacBio). For each sample, Strand-seq data was also produced, which is a technique for generating single-cell sequencing reads for each of the diploid DNA template strands [84, 261]. This combination of technologies permits accurate *de novo* assembly [238] and singleton-resolution, physical phasing of variants, rather than the more common approach of statistical phasing, without the need for parental or reference sequence data. As reads in 1k-HGSVC are *de novo* assembled in a reference-free manner, I do not need to be concerned about the ability of mapping techniques to handle reads containing template switch associated mutation clusters. In addition, the ample flanking sequence around complex variants has already been demonstrated to permit the resolution of much larger structural variants contained within single reads [20, 53, 80, 238], which should make short template switch events comparatively trivial to resolve for long read variant call pipelines.

I denote the third dataset Ice-Trios. It consists of *de novo* variant calls from 1,548 parent-offspring trios from the Icelandic population, sequenced to an average coverage of 35x using 76bp and 150bp paired end reads [137]. From this cohort, an average of 70.3 *de novo* mutations were identified per proband [137]. This dataset is of interest as it currently provides the largest collection of publicly-available *de novo* variant calls and has become a standard community reference for properties of human *de novo* variation when studying the human mutation

spectrum [4, 141], mutational mechanisms [176], variant calling method development [61], and indeed further studies of mutations rates amongst family trios [104]. This dataset is limited in that it is provided with no identifiable genotype information, consisting solely of information about chromosomal positions alongside reference and alternate alleles. Considering the discussion above on the importance of using phased variant calls, any events I identify using this dataset will not be fully convincing. Similar to my motivation for using a community-standard resources of human-population calls however, identifying plausible template switch variants within these calls allows me to contextualise any findings with a well-established resource.

4.3 Establishing significance and alignment quality thresholds

4.3.1 A between-human LPR threshold determined through simulations

As with my great ape analysis (Chapter 3), I establish a threshold on the statistical significance between unidirectional and template switch alignment probabilities for candidate template switch events (see §2.3.2 and §3.2) using the LPR test statistic (Equation 2.9).

I again use my evolutionary simulation approach (§2.3.3) to establish this threshold, simulating evolution both with and without template switch events to identify a suitable LPR between the unidirectional and template switch pairHMMs for any candidate event. For the between-human analysis, I perform simulations at 0.01% divergence (parameter t), which represents the average divergence between any two human samples [293]. I also change additional parameters of the TSA pairHMM to provide a better fit to the 1k-30x data. Recall in the TSA pairHMM (Figure 2.4) that λ is the mean indel length and ρ is the mean number of indels per substitution. Average values for these parameters were calculated using all 3202 samples from the 1k-30x variant calls as $\lambda = 3$ and $\rho = 0.04$. Also recall that θ corresponds to the probability of initiating a template switch, calculated as $N/(C \times A)$, where N is the expected number of events in the human sample and C is the total number of mutation clusters between the pairwise comparison of that sample and the human reference genome. Here, I set N to 200 based on the average number of significant events identified in earlier analysis of low coverage human variant calls, and C to 148032, which is the average number of mutation clusters (two substitutions within 10 nucleotides of each other and/or an indel $\geqslant$ 5nt in length) identified across 1k-30x sample VCFs. A summary of these parameter values are provided in Table 4.2 – see also §2.2.4 for a discussion on the suitability of setting parameter values in this manner.

Table 4.2: PairHMM parameters used in the human population analysis.

Parameter	Value(s)	Rationale
t	0.001	Based on [293]
ρ	0.14	Based on estimates from [46]
λ	20	Based on estimates from [46]
N	200	See §2.2.4
C	148032	The average number of mutation clusters (defined using the procedure described in §2.2.7) identified from the VCFs of all 1k-30x samples
L	10	See §2.2.4

As before, I set a threshold on the resulting Monte Carlo LPR distribution (Equation 2.9), in this case setting a threshold which removes all false positive calls which were not intentionally introduced during simulation (Figure 4.1a). Few false positives were introduced with a high LPR in the present simulations because of the low divergence separating any two human sequences (0.01% used for simulation). Despite this stringency, the selected threshold still captures nearly all of post-filtered true positive events (Figure 4.1b).

4.3.2 Pairwise alignment quality threshold

As in the hominid analysis, I establish an average alignment quality filter based on sampling genome-wide pairwise alignments (see §3.2.2). Given the low levels of divergence between any two humans, there could be an argument for only permitting alignments in my final callset which are composed entirely of matching alignment columns. I decided against this approach, as disallowing all alignments which contain additional mismatches and/or short indels would remove the possibility of capturing statistically significant template switch events for which an mismatch-containing ②→③ region is the product of an error-prone polymerase, or alignments of events which contain coinciding mutations in the flanking Ⓛ→① and ④→Ⓡ alignment regions.

To generate a between-human per-base alignment probability distribution, I repeat the following procedure for each of the 3,202 samples in the 1k-30x samples. I first generate 10000, 100nt GRCh38 genomic coordinate ranges in BED format using:

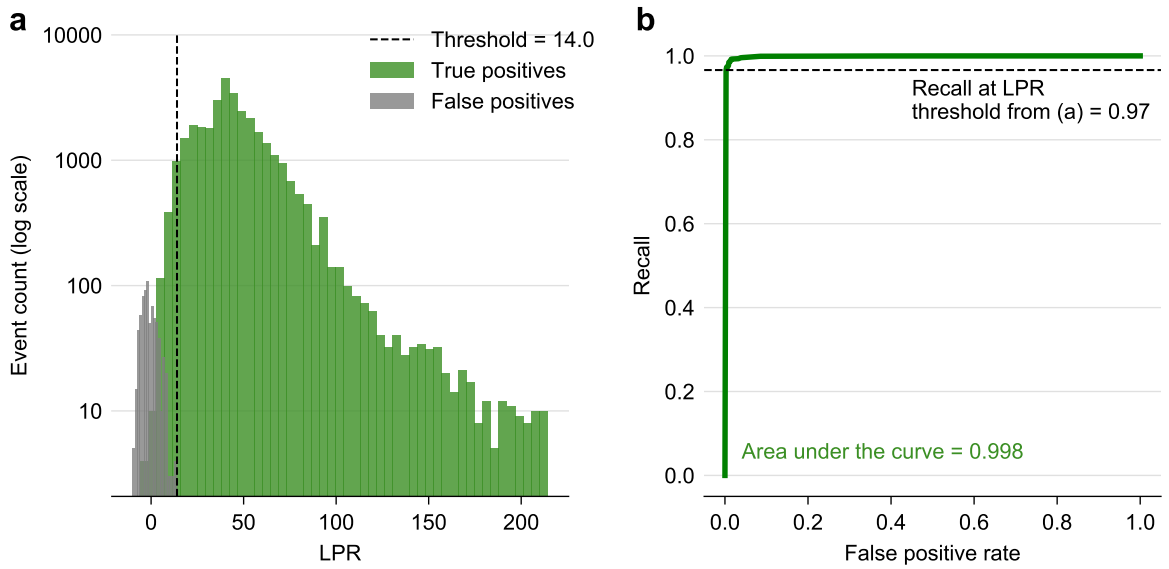

Figure 4.1: Establishing statistical significance threshold for candidate human population template switches. (**a**) Histogram of LPRs for true positive template switch events (green), and background mutation clusters which manifest as potential false positive events (grey) identified from the two sets of evolutionary simulations. The chosen LPR threshold of 14.0 (dashed line) results in having no false positives in my simulations (i.e. $p < X$, where $X = 1/N$ and N is the number of null hypothesis simulations). Note the log scale y-axis. (**b**) ROC curve for discriminating between true positive and false positive template switch events.

```
bedtools random -l 99 -n 10000 -g chromosome_lengths.tsv -seed 42
```

where `-l 99` specifies the region length (a specified value of 99 samples a region of length 100), `-n 10000` is the number of regions to sample (I redundantly sample coordinate ranges for the next step), `chromosome_lengths.tsv` is a tab-separated values file (where each line contains one chromosome ID, a tab character, and the length of that chromosome), and `-seed 42` is the random seed. As inaccessible regions of human genome assemblies such as gaps contained in coordinates sampled in this manner can confound sampling-based statistical tests [73], I filter out any of the sampled coordinates which intersect with known gap regions. Known gap regions within the GRCh38 reference genome assembly were obtained from the University of California, Santa Cruz (UCSC) Table Browser [142] in BED format and intersections were checked with `bedtools intersect`.

For each of the remaining sequence regions, I retrieve any variants within this region for the current sample from the corresponding sample VCF, left-aligning and normalising indels using `bcftools norm -m-any`. If any variants are found within this region, I create a FASTA

representation of these variants using `bcftools consensus`, which uses a sample VCF to incorporate variants into a reference sequence in FASTA format. The sample and the reference sequence are then aligned under my unidirectional pairHMM, to which I add a $M_1 \rightarrow M_2 \rightarrow M_3$ template switch penalty $(\theta + \sigma(1-2\delta))$ to allow the resulting sampled probabilities to be applicable to my template switch alignments (as in §3.2.2). This log-probability is then divided by the alignment length and recorded. If no variants are present within the region, I instead sample a "perfect" alignment probability (100% matches), which was obtained by unidirectionally aligning two arbitrary 100nt sequences (each consisting of all As), adding the template switch penalty, and dividing the resulting probability by 100 (the alignment length). Note that the nucleotide composition of this perfect alignment is unimportant, as my model is parameterised using a JC69 substitution matrix, so all nucleotides are assumed to be equally likely.

This procedure produced a total of 6,406,000 pairwise probabilities (3202 samples $\times$ 2000 alignments), the distribution of which is shown in Figure 4.2. As expected, the majority of alignment regions sampled fell within regions containing no variants, and the 100% identity per-base alignment score of -0.15 comprises 87% of alignments sampled (single tall bar, right side of Figure 4.2). If I set a threshold using the 20th percentile as previously (Figure 3.3), this would filter out every alignment that does not solely consist of match states (everything to the left of the -0.15 bar). I therefore set the per-base threshold to -0.255, which removes the majority of $< 100\%$ identity alignments whilst still allowing some mismatch and/or indel columns in the final TSA pairHMM alignments (Figure 4.2, dashed vertical line).

4.4 Identifying template switch mutations within human variation data

4.4.1 Event discovery pipeline

Based on the methods described above, I developed a pipeline to identify significant template switch events given an input multi-sample VCF and the reference genome used to generate the VCF. In my case, these VCFs are the 1k-30x, 1k-HGSVC, and Ice-Trios calls outlined in §4.2, and the reference genome is the GRCh38 genome used throughout the 1000 Genomes Project[1]. I will detail the steps involved in my event discovery pipeline, but an overview is also provided in Figure 4.3.

[1] retrieved from the 1000 Genomes Project FTP site on 07/02/21

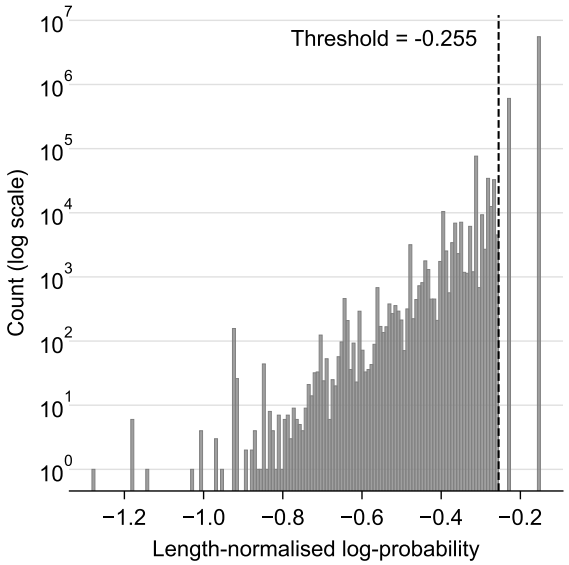

Figure 4.2: Establishing an alignment quality threshold for candidate human population template switches. Histogram of length-normalised alignment LPR values across all sampled pairwise alignments between each 1k-30x sample and the GRCh38.p12 human reference genome.

Given a set of multi-sample, population-scale chromosome VCFs, I first split each chromosome VCF by sample, normalising (left-aligning) indels and splitting multi-allelic variants into separate records. I discard the majority of variant records when creating sample VCF subsets as I am only interested in clustered variants and/or indels. This is achieved by first generating a per-sample VCF with:

```
bcftools view -c1 -a -U -I -s "$sample" ftp_vcfs/chr"$chrom".vcf.gz |
        bcftools norm -m-any
```

where `bcftools view -c1` specifies a minimum of one alternate allele is present, `-a` trims alternate alleles not seen in the subset, `-U` excludes sites with an uncalled genotype (or haplotype), `-I` turns off re-calculation of the VCF "INFO" field for improved running times, `-s` specifies the sample, and `bcftools norm -m-any` left aligns indels and splits multiallelic sites into multiple records. Variants are then retained if they either: (a) consist of a single indel $\geqslant$ 5nt in length, or (b) if the variant position is within 10nt of another variant position. Proximal variants satisfying (b) can consist of any combination of SNPs, indels, and structural variants, and the 10nt window used to define a mutation cluster was retained from my great ape analysis (§3.3.1). Each $\geqslant$ 5nt single indel is assigned a unique cluster ID and retained in the sample VCF. For proximal variants forming a mutation cluster, the cluster is assigned an ID and retained in the

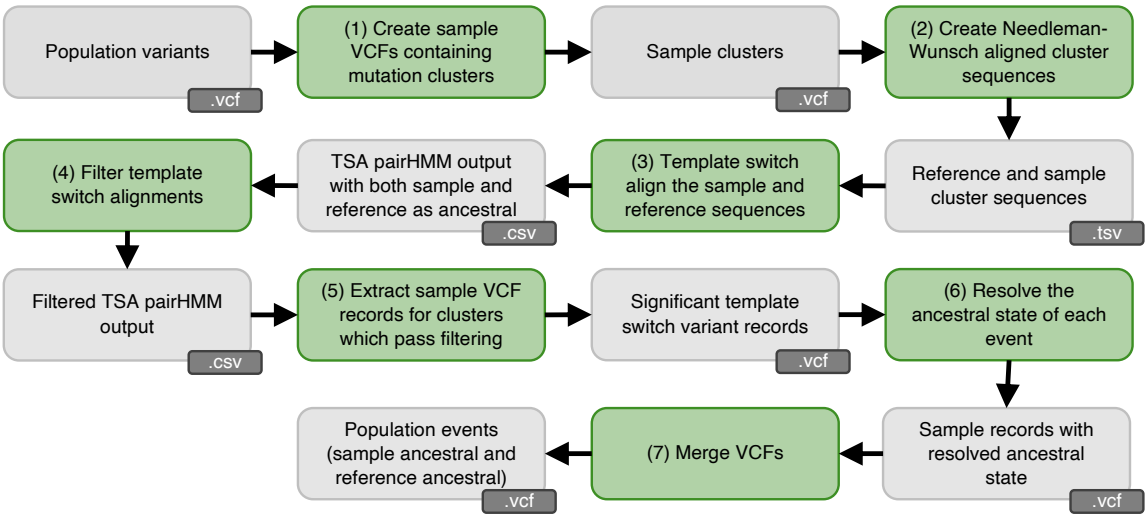

Figure 4.3: Pipeline used to identify template switch events from population-scale VCFs. As explained in the main text, variants within existing multi-sample VCFs are identifed by: (1) creating sample VCFs using `bcftools view`, assigning unique IDs to mutation clusters by identifying SNPs within 10nt of another variant or single indels of length $\geqslant$ 5nt; (2) using `bcftools consensus` to reconstruct the sample sequence from the reference FASTA file, aligning it to the corresponding reference sequence using the Needleman-Wunsch alignment algorithm, which allows the focal cluster to be identified by the TSA pairHMM; (3) template switch aligning the cluster sequences, treating both the reference and sample as ancestral in turn; (4) filtering the events by applying the thresholds outlined in §4.3 and filters described in §4.4.2; (5) extracting significant template switch events which pass all filters; (6) retrieving the ENSEMBL LastZ aligned region of the chimpanzee genome corresponding to the reference and sample sequences, realign each to the chimpanzee genome using the unidirectional pairHMM, and assigning either the reference or sample as ancestral based on a lower log-probability alignment with the chimpanzee sequence; (7) merging the sample-ancestral and reference-ancestral events into two final VCFs containing template switch associated variants. Boxes are coloured grey to indicate files, and green to indicate computational steps. Pipeline implemented using Snakemake [216].

sample VCF only if all variants which define the cluster are present on the same haplotype in fully phased datasets (i.e. all are `0|1`, `1|0`, or `1|1`), or if the variants share the same genotype for VCFs without phasing information (`0/1` or `1/1`, not relevant to the datasets here). Note that I refer to both single indels and proximal SNVs/indels retained by this procedure as "clusters" below.

I process variants assigned to each mutation cluster in the resulting VCFs separately. I retrieve the reference allele sequence associated with those positions from the GRCh38 reference FASTA file, and produce the alternate allele sequence using `bcftools consensus`.

These sequences are then aligned using a C++ implementation of the pairwise Needleman-Wunsch alignment algorithm with default parameters[2]. This initial global alignment is necessary to allow my unidirectional/TSA pairHMM code to subsequently define a focal mutation cluster for realignment (see Figure 2.7).

Using these pairwise global alignments, I then realign each mutation cluster using both the unidirectional and TSA pairHMMs (parameterised as described in Table 4.2), treating each sequence as ancestral in turn. It is necessary to consider both the reference and sample as ancestral because of the "reversibility" of template switch event detection (see Figure 3.5 and §3.3.2). I use an outgroup comparison procedure alongside these bidirectional scans to resolve the ancestral state in a subsequent step (see §4.4.2 below).

4.4.2 Filtering, ancestral state resolution, and output

For each sample, and for each candidate template switch, I apply filters in a similar manner to those outlined in §2.3.4 and applied in my great ape analysis (§3.3.1), utilising the between-human thresholds established in §4.3. To be called as significant, I require:

1. a LPR $\geqslant 14$ (see Figure 4.1a),
2. a per-base alignment probability of -0.255 (see Figure 4.1c),
3. events are not located within a low complexity region of the GRCh38 assembly,
4. the ② → ③ region contains all four nucleotides (see §2.3.4),
5. and the event is not defined solely by a single deletion (see §2.3.4 and Figure 2.6).

These filters are familiar from the previous two chapters. For between-human calls however, I also employ an additional filter of:

6. switch point ① is required to precede switch point ④ (① < ④) (see Figure 4.4).

Recall from Figure 3.9 that ④ < ① events involve the creation of complex rearrangements, including linear duplications defined by the sequence region between points ④ and ①. During the course of pipeline development and template switch mutation discovery using the initial 1k Genomes low-coverage calls, features repeatedly observed for this type of event included that: (a) the ② → ③ length of events is not strongly positively correlated with LPR (unlike the more

[2] obtained from https://github.com/noporpoise/seq-align

common ① < ④ events), and (b) LPR instead scales linearly with the number of mutations in the cluster as defined by the unidirectional pairHMM alignment. When evaluating the final callset identified from 1k-30x using my pipeline, these relationships held true (respectively see (a) and (b) of Figure 4.4). While these events are not inherently problematic for downstream analyses and are biologically feasible (see the discussion in §3.3.3), a problematic subset of ④ < ① events was repeatedly observed in which the ② → ③ region is disproportionately short compared to both the linear duplication created between ① and ④ and the number of unidirectionally-defined mutations alternately explained by the template switch model. While these events are indeed plausible under a model of template switching, it is is difficult to confidently distinguish between a template switch mechanism, and an alternate mechanisms that has generated a duplication in addition to incorporating some small amount of alternate sequence that presents as a template switch. In cases where the latter is true, a false positive template switch generated by a duplication would have an inflated LPR due to the large mutational footprint left by the duplication that can be significantly explained under my LPR test statistic. As I cannot confidently distinguish these cases, I use the above filter to discard

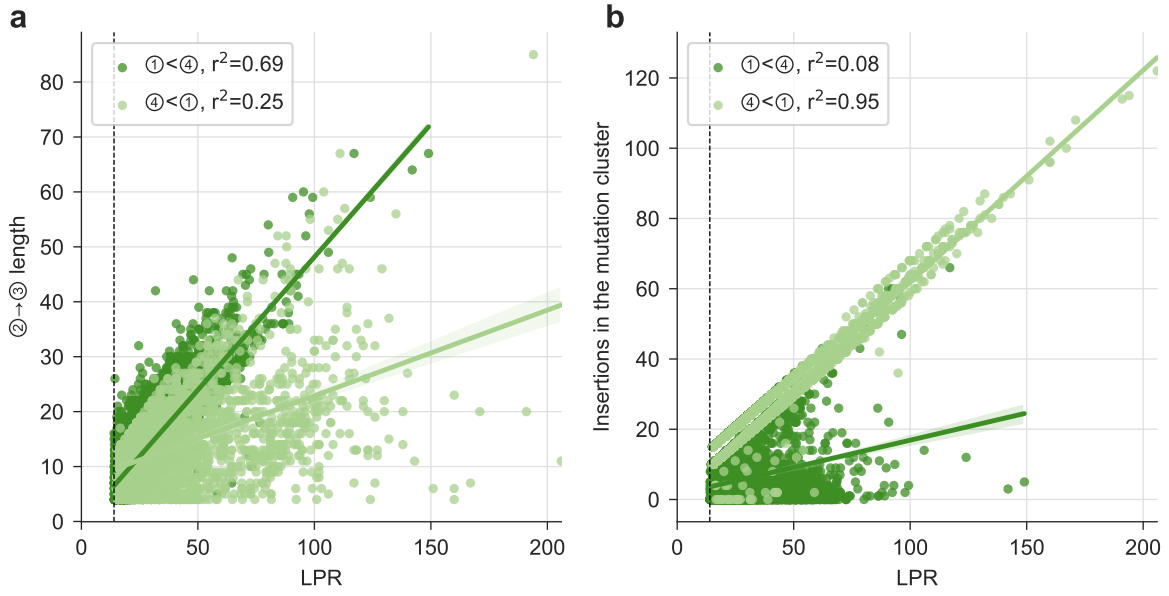

Figure 4.4: Summary statistics for statistically significant events in the 1k-30x calls for which ① precedes ④, compared to those in which ④ precedes ①. (**a**) ② → ③ length as a function of LPR. Overlaid coloured lines correspond to a linear regression model fit to the data points of each group, corresponding r^2 values are indicated in the legend. A vertical dashed line shows the LPR threshold established in §4.3.1. (**b**) Insertions in the unidirectionally aligned mutation cluster as a function of LPR. Annotations as in (a).

any ④ < ① events, accepting some loss in statistical power and possible under-reporting of events.

I retrieve any variant records associated with template switch alignments that satisfy these criteria from the sample-cluster VCFs processed above. To produce the final population template switch VCFs, I resolve the ancestral state of each event-associated mutation cluster from this significant event subset by comparing the reference and sample sequences for each cluster with a phylogenetic outgroup genome. I use the chimpanzee genome for this purpose. For each significant event, I use the GRCh38 coordinates of the event to query the ENSEMBL Rest API [318] for the corresponding region of the chimpanzee genome, provided by Ensembl [322] as a whole-genome LastZ alignment [118] between GRCh38 and the reference chimpanzee genome ("Pan_tro_3.0" at the time of writing). I then align both the reference and sample sequence to the chimpanzee sequence using the unidirectional pairHMM. I consider the pairwise alignment with the greatest log-probability to represent the ancestral state, and discard any events for which the corresponding Pan_tro_3.0 sequence region is not available. All significant events with a resolved evolutionary direction are then merged and sorted to produce three population-level VCFs: a VCF for each of the sample-ancestral and reference-ancestral template switch mutations, and a VCF containing events with an unresolved ancestral state.

These VCFs contain sets of variants that are associated with each unique template switch, identifiable from an integer-indexed, shared ID in the ID field. For example, the following VCF snippet (displaying only the first 5 VCF fields and final header line):

```
#CHROM POS     ID      REF      ALT
chr1 1456154   TS_1    G        A
chr1 1456155   TS_1    G        C
chr1 1456156   TS_1    G        C
chr1 1456164   TS_1    TGCA     T
chr1 5071719   TS_2    G        GTGCTTTT
chr1 5071720   TS_2    ACTC     A
chr1 5071724   TS_2    A        AGG
chr1 7058164   TS_3    GCACCC   G
chr1 7058171   TS_3    GCA      G
chr1 7058175   TS_3    C        CGTG
chr1 7058176   TS_3    A        ATG
```

shows the variants associated with three significant template switches on chromosome 1, each of which is indicated by IDs ∈{`TS_1, TS_2, TS_3`}. In this case, the first event is defined by a footprint of multiple SNVs and one deletion, while the second and third events are defined by

multiple insertions and deletions. Raw output from the pairHMMs, BED files defining event regions, and the printed Viterbi/Viterbi-like alignments are also generated for each event.

4.5 Overview of the template switch event callset

4.5.1 The prevalence of short-range template switch mutations within haplotype-resolved human genomes

By applying the pipeline and filtering procedures described in §4.4 to the 3202 samples in the 1k-30x dataset, I identified 3322 unique, short-range template switch mutations for which the sample genomes correspond to the derived allele. I additionally identified 122 events for which the reference genome corresponds to the derived allele, and 19 events with an unresolved ancestral state due to missing chimpanzee sequence for these regions. The remainder of this chapter will focus on the first set of 3322 events, in which samples represent the derived template switch alleles.

The population VCF, BED, and pairHMM output files for these events are provided in the supplementary data files for this chapter. The VCF file also contains annotations for each position from the Ensembl Variant Effect Predictor (release 104) [203] (performed using `ensemblorg/ensembl-vep` in Docker [207]). I will refer to this file as the 1k-30x template switch mutation cluster VCF. I additionally create a subset of this VCF which contains a single variant record per event-associated mutation cluster in the VCF, retaining the first position of each mutation cluster (or in the case of a single indel, retaining the single indel record). This VCF subset is used for any analysis below in which I need to process each mutation cluster as a single template switch variant. I will refer to this VCF as the 1k-30x template switch single-variant VCF. I will indicate when each VCF is used throughout the remainder of this chapter.

4.5.2 Apparent mutation clusters and short indels caused by template switches are not associated with poor read mapping

As discussed in §4.2, Löytynoja and Goldman [185] raised concerns that short-read variant callers may struggle to accurately map reads containing template switch mutations. I therefore inspected the read depth and mapping quality of variants in the 1k-30x template switch mutation cluster VCF (Figure 4.5). Template switch-associated variants do not display low read depth or poor mapping quality, with a median read depth of 31.5, and median mapping quality of 60.

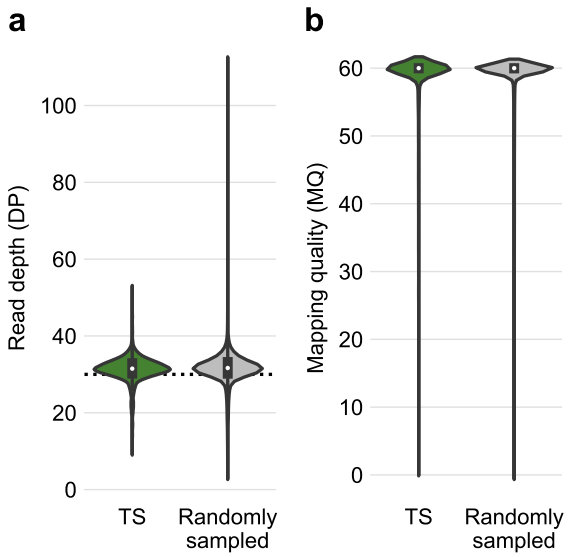

Figure 4.5: Short reads containing significant short template switches unambiguously map to the reference human genome. (**a**) The read depth of all variants associated with the 3322 significant template switch events in the 1k-30x dataset, compared with 230,000 (10,000 for each autosome and chromosome X) randomly sampled variants from the 1k-30x dataset that don't fall in regions annotated as low-complexity. The black dotted line at DP 30 indicates the mean genome-wide coverage of the 1k-30x dataset. (**b**) As in (a), but for mapping quality.

These medians are respectively greater than the genome-wide mean depth and at the maximum measurable mapping quality. I performed a Mann-Whitney U test [192] in SciPy [301] with the alternate hypothesis that the distribution of template switch depth and mapping quality are stochastically less than a random genome-wide sample of 230,000 variant positions (10,000 from autosomes 1-22 and chromosome X), sampled from regions of the genome not masked as low-complexity by RepeatMasker [276]. Mapping quality is not significantly lower for template switch associated variants than this random sample of genome-wide variants ($p \approx 1$). While the read depth is significantly less than the random sample (p=1.37×10^{-10}), this likely reflects differences in the genomic locations sampled for the random variants compared to the template switch variants, that is template switches are enriched/depleted in several genomic regions and don't occur uniformly randomly across the genome (see Figure 3.10). The median depth difference of 31.5 for template switch variants compared to 31.7 for randomly sampled variants is so negligible however that I assert that it is not a concern when considering the mappability of reads containing template switch mutations. As template switch events manifest as either single indels or clusters of SNVs and/or indels, reads containing an event will undergo local reassembly as part of the GATK HaplotypeCaller pipeline used to generate the 1k-30x

calls (discussed in §4.2). It therefore appears that local reassembly is sufficient to resolve template switch variants from short read data without an appreciable drop in mapping quality or coverage.

4.6 Population genetics of human template switch mutations

All high quality SNVs, indels, and structural variants previously identified in the 1000 Genomes Project data have consistently shown haplotype distributions amongst and within super-populations that reflect well-characterised human demographic history [42, 284, 293]. In this section, I therefore seek to both describe the catalogue of events identified by my pipeline, while interpreting template switches in their population context, expecting that short template switches should be distributed similarly to other forms of variation in these data.

4.6.1 Per-individual template switch count distributions follow population expectations

The 1k-30x calls are composed of samples from five super-populations (continental groups): Africa (AFR), America (AMR), East Asia (EAS), Europe (EUR), and South Asia (SAS). Samples are further divided into populations within each super-population, and below I refer to each population by the three-letter codes used in previous 1000 Genomes Project publications [42, 284, 293]. An average of 116 ($\pm$17 standard deviation) significant template switch events were found per sample, with African populations demonstrating the greatest average number of events per sample (Figure 4.6). This is consistent with previously observed patterns of increased SNV, indel, and structural variant diversity in African populations, attributed to a sustained larger effective population size than other continental groups due to historical non-African population bottlenecks as described by the out-of-Africa model of human origin [43]. Recall that I identified 122 events for which the reference genome represents the derived allele (§4.5), this indicates that GRCh38 is relatively diverse but similar to an ordinary genome, falling within a typical range of template switch count if one were to randomly select a 1k-30x sample. In previous studies, African populations exhibit increased levels of heterozygosity for other variant classes (followed closely by Puerto Ricans due to African admixture in their population), and high levels of homozygosity have been observed amongst East Asian populations [284]. These population-level patterns of zygosity are concordant with those observed for template switch mutations (Figure 4.7).

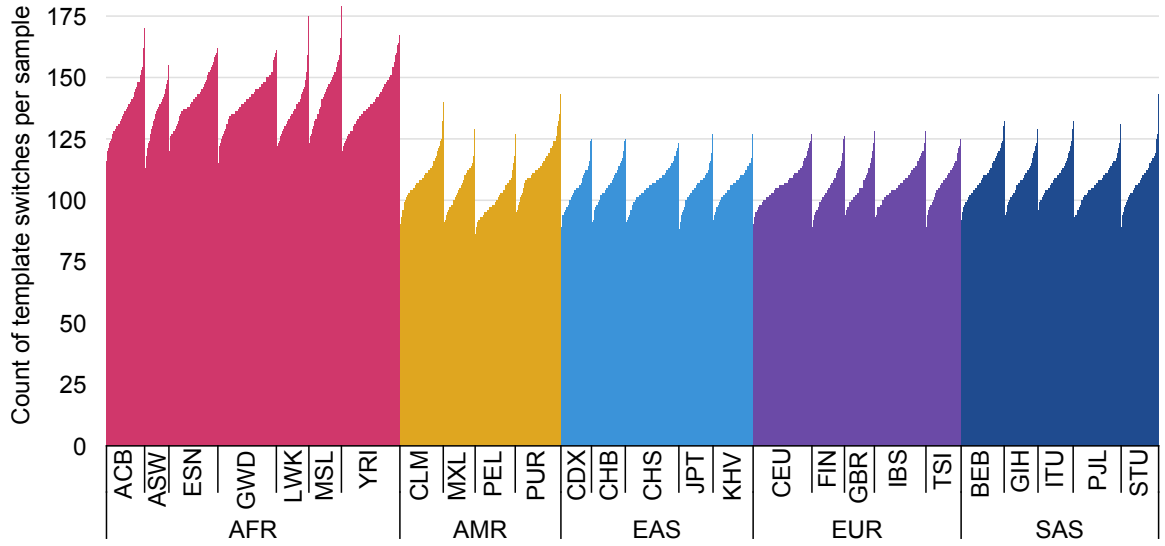

Figure 4.6: Count of template switch mutations identified per sample. Samples are ordered alphabetically by super-population, then population, and then sorted by event count in ascending order within each population. Average ($\pm$ standard deviation) events per super-population are 139 ± 11, 107 ± 10, 105 ± 8, 106 ± 8, 108 ± 8 for AFR, AMR, EAS, EUR, and SAS, respectively.

4.6.2 The population structure of variants caused by template switching is consistent with known human demographic history

It is typical to assess population structure when evaluating global human population variation callset quality [60, 284]. Principal component analysis (PCA) is a popular model-free method for identifying population structure in genetic variation datasets caused by historical demographic events [205]. For the purposes of establishing a set of variant calls as high quality, in this case short-range template switch mutations, it is expected that known continental population structuring should explain most of the variance along the first four principal components [60, 134, 191, 284]. I therefore performed a principal component analysis of the haplotype matrices associated with all 1k-30x template switch mutations. As is typical when performing PCA on human population variation datasets [191, 284], I first apply the normalisation procedure outlined by Patterson [231] (note that although I perform this step to be consistent with earlier human structural variant studies, it often has little effect on the final PCA results [205]). That is, assume a $n \times m$ matrix H with n sample-indexed rows and m (template switch) variant-indexed columns, where each entry $H(i,j)$ contains 0, 1, or 2, respectively corresponding to homozygous reference, heterozygous, or homozygous alternate. A normalised matrix is calculated as $H_{norm}(i,j) = H(i,j) - \mu(j)/\sqrt{p(j)(1-p(j))}$ where $\mu(j)$

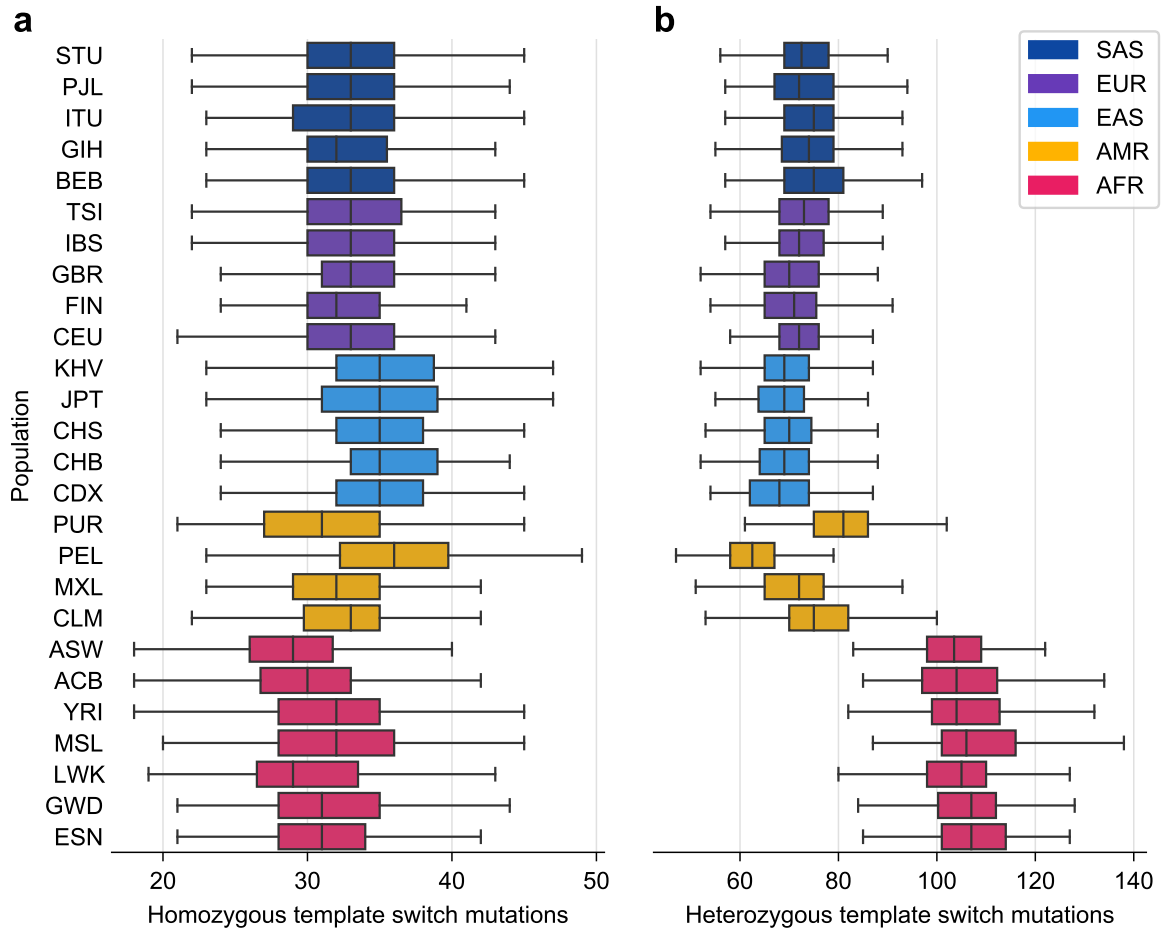

Figure 4.7: Patterns of template switch zygosity in human populations are consistent with other classes of human genetic variation. (**a**) Count of homozygous template switch mutations by population. (**b**) Count of heterozygous template switch mutations by population. For each population, boxes represent the median, Q1, and Q3 of the count distribution, and whiskers represent $\pm 1.5 \times$IQR (the interquartile range). Populations are grouped and coloured by super-population.

is the column mean given by $\sum_{n=1}^{m} H(i,j)$ and $p(j) = \mu(j)/2$ is an estimate of the underlying allele frequency (for diploid autosomal data). I generated H_{norm} for the matrix of autosomal haplotypes contained in the the 1k-30x template switch mutation cluster VCF, from which I calculated the first four principal components (PCs) via singular value decomposition using scikit-allel [213]. Population structure as captured by PCs 1–4 (Figure 4.8) is concordant with the structure observed for structural variant callsets produced by both the 1000 Genomes project [284] and the gnomAD-SV project [60]. That is, African populations separate from all other superpopulations along PC1, and East Asian populations separate along PC2, recapitulating the

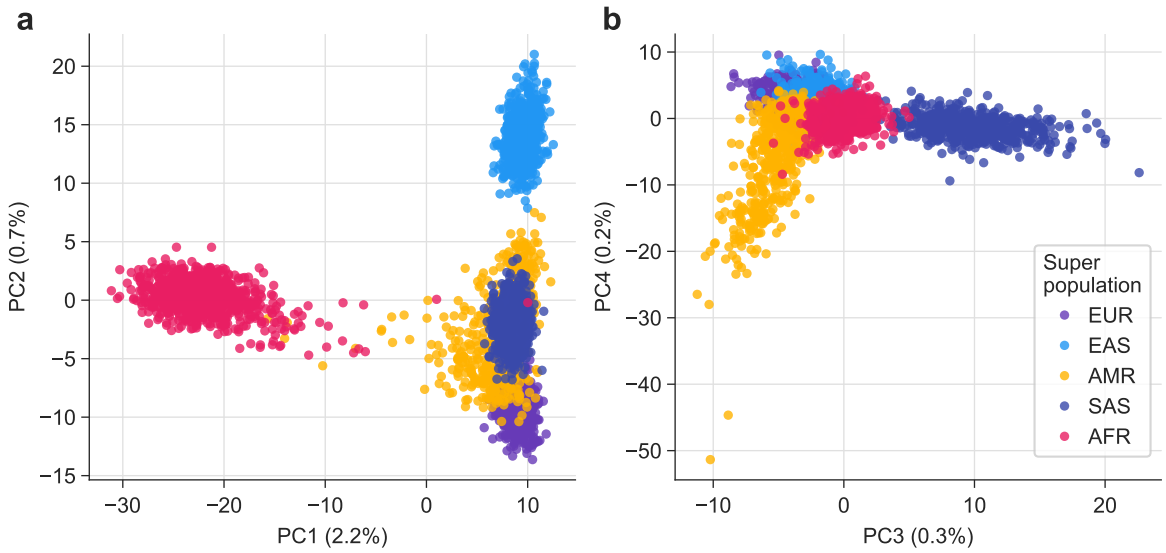

Figure 4.8: Principal component analysis of template switch haplotypes in the 1k-30x dataset captures expected continental population groupings. (**a**) Principal components 1 and 2 and (**b**) principal components 3 and 4, with superpopulations coloured as shown in the key.

distinct patterns of high heterozygosity and homozygosity unique to African and East Asian populations (shown in Figure 4.7). PC3 and PC4 further separate out South Asian and American individuals, respectively, in agreement with the initial 1000 Genomes Project structural variant callset [284].

4.6.3 Inferred template switch alleles across all samples are consistent with expected theoretical distributions

With the expected population structure of template switch mutations established (Figure 4.8), I wanted to assess event discovery power from the samples included in the 1k-30x callset. This allows me to ask if the events discovered by my methods are consistent with theoretical expectations. For each super-population, I consider the number of unique events detected as a function of the number of samples observed, and compare this to the expected coalescent tree length assuming neutral selective pressure and a population at demographic equilibrium, calculated as $2\sum_{i=2}^{n-1}\frac{1}{i}$ for $n = 3202$ diploid genomes [122, p. 27], with the summation lower bound corrected to 2 to account for no singletons in the 1k-30x dataset. African and East Asian super-populations are particularly consistent with the expected tree length (Figure 4.9), while other continental groups show varying levels of decreased event discovery power at

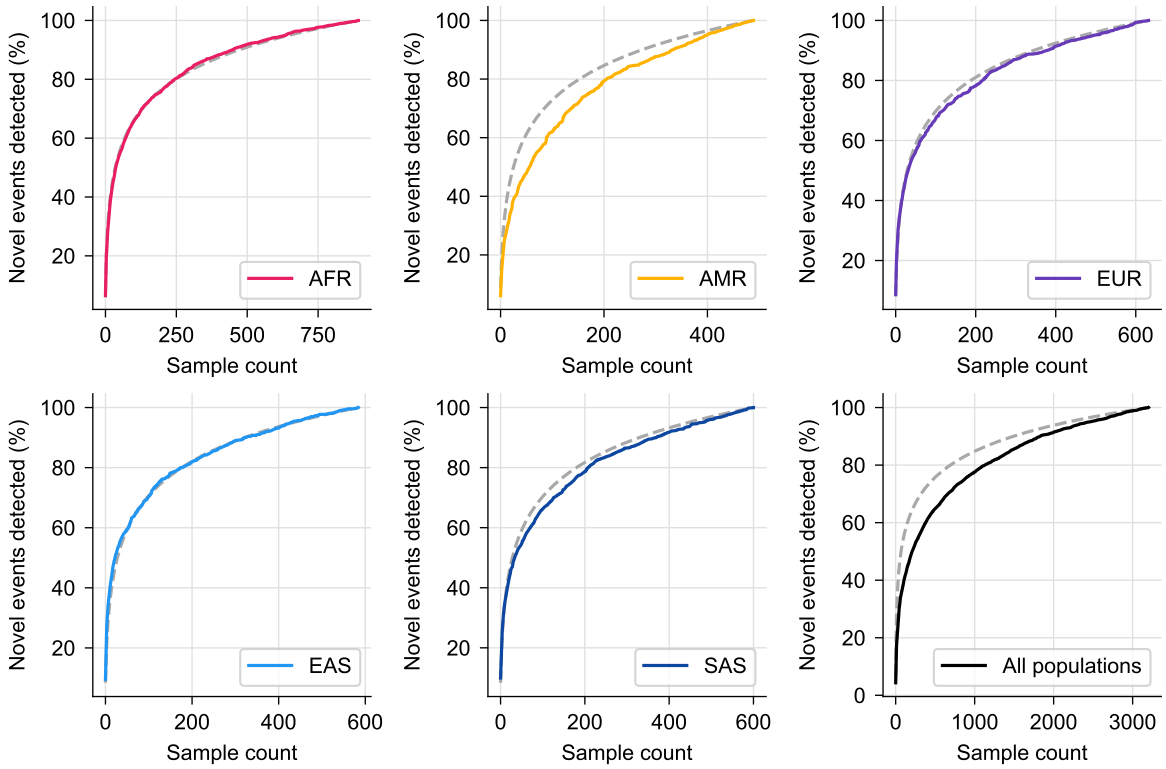

Figure 4.9: Novel template switch discovery as a function of samples observed is consistent with the expected total coalescent tree branch length. The first five subplots (read left-to-right) show the number of novel template switches discovered as randomly shuffled genomes are added for each indicated population. The dashed grey lines correspond to the expected coalescent tree length of $2\sum_{i=2}^{n-1}\frac{1}{i}$ scaled by the total number of observed template switches. The final subplot shows the same information when aggregating and randomly shuffling all samples.

small sample sizes, caused by an excess of rare variants in the population which may reflect population structure and admixture in these populations (as captured by PC1 and PC2 in Figure 4.8a). For example, amongst the American genomes, Peruvians (population code PEL) typically have an excess of rare variants as their genomes are of predominantly Native American ancestry [116, 262, 293], while Puerto Ricans (population code PUR) are a heavily admixed population with a combination of European, West African, and Native American ancestry [290, 293]. Importantly however, the cumulative event discovery distributions do not plateau for any super-population (Figure 4.9), indicating that it would be beneficial to include a larger number of samples from each super-population to maximise novel event discovery. This of course may also reflect the exclusion of singleton variants from the callset, or the need to include more demographically diverse samples. While the 1k-30x dataset consists of samples

which are broadly representative of major continental groups, samples are typically drawn from demographically large populations and rare variation (such as template switch variants) may be underrepresented. The Simons Genome Diversity Project demonstrated this for SNVs and indels [191], showing that over 10% of the the variants present in the genomes of some samples from small populations are not represented in the 1000 Genomes Project calls (although note they were comparing with the original, low-coverage callset presented in [293]).

The population-scaled mutation rate for a given variant class can be calculated as $\theta = 4N_e\mu$, where N_e is the effective population size and μ is the per-generation mutation rate [101, 284]. The effective population size for humans has been estimated as approximately 10^4 [191, 240], and has previously been used for calculating the mutation rate of distinct classes of variation within both the 1000 Genomes Project and gnomAD-SV cohorts [60, 284]. To calculate the unknown mutation rate μ, it is typical to use Watterson's estimator of θ [309], given by

$$\hat{\theta}_w = \frac{S}{\sum_{i=1}^{2n-1} \frac{1}{i}} \tag{4.1}$$

where S is the count of derived template switch alleles and n is the number of diploid samples considered. $\hat{\theta}_w$ for 1k-30x template switches is

$$\hat{\theta}_w = \frac{3322}{\sum_{i=2}^{(2\times 3202)-1} \frac{1}{i}} = 398.24, \tag{4.2}$$

where the lower bound of summation is again corrected to 2 to account for the lack of singleton variants in the callset. From Equation 4.2, I can then estimate a human per-generation short-range template switch mutation rate as

$$\mu = \frac{398.24}{4 \times 10000} \approx 0.01. \tag{4.3}$$

I note that this may be an underestimate despite the correction for no singleton mutations. Nevertheless, I can still assess if the frequency of the template switch alleles assessed here under this mutation rate are distributed approximately as expected by inspecting the allele frequency spectrum (see [284, Extended Data Figure 2] for a similar analysis across human structural variant classes). Assuming template switch mutations are selectively neutral (a reasonable assumption, given they are small in scale and primarily distributed in non-functional genomic regions, see Figure 3.10), the expected allele frequency spectrum under the coalescent for n samples $\{x_1, ..., x_{n-1}\}$ can be calculated as $x_i = \hat{\theta}_w \frac{1}{i}$ [92]. As shown in Figure 4.10, the distribution of template switch alleles in human populations is relatively consistent with the expected frequency spectrum, but with an excess of rare (doubleton) alleles. This is likely

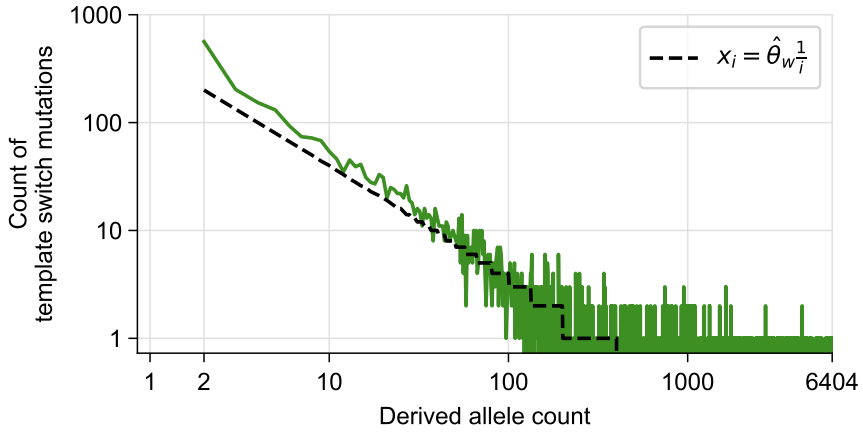

Figure 4.10: The allele frequency spectrum of short template switches indicates a slight excess of rare variants. The number of derived template switch alleles (y-axis) and the respective derived allele frequencies, represented by derived allele counts (x-axis). A black, dashed line is overlaid which represents the expected allele frequency spectrum under an estimate of the population-scaled mutation rate. The grey shaded area represents allele counts not included in the 1k-30x dataset. Note the log scale on both axes.

caused by the inclusion of 602 trios in the 1k-30x calls, which in the absence of *de novo* reversion to the reference haplotype will cause an excess of apparent doubleton template variants due to parent-offspring inheritance.

The relationship between these expected and observed frequency spectra can also be summarised using Tajima's D statistic [159, 289]. This is calculated as the difference between the average number of observed pairwise differences between n samples ($\hat{\theta}_T = \sum_{i<j} d_{ij} / \binom{n}{2}$, where d_{ij} is the count of differences between samples i and j), and the expected number of differences given by $\hat{\theta}_w$ (Equation 4.2), normalised as

$$D = \frac{\hat{\theta}_T - \hat{\theta}_w}{\sqrt{\text{var}(\hat{\theta}_T - \hat{\theta}_w)}}. \tag{4.4}$$

Details on the normalising standard deviation calculation are provided in [289]. I calculated Tajimas's D for the vector of all template switch allele counts using `scikit-allel` [213], giving a value of -1.312. A negative value of D is typically caused by an excess of rare variants compared to the expected value given by $\hat{\theta}_w$, as reflected in the allele frequency spectrum shown in Figure 4.10, and may either be attributed to the inclusion of trios in the callset, or it can be indicative of population expansion following a recent bottleneck [86].

4.6.4 Summarising the population-genetic properties of human template switch variants

I have demonstrated in this section that template switch variants are not only prevalent in human populations (Figure 4.6), but that patterns of zygosity (Figure 4.7) and population structure (Figure 4.8) are concordant with previous studies of other variant classes [60, 284, 293]. Further, I have shown that template switches display an expected (subject to some known population structure) cumulative distribution of novel events discovered as new genomes are introduced (Figure 4.9) and an allele frequency spectrum which approximately follows an expected distribution under neutrality (Figure 4.10). In combination, this shows that template switch mutagenesis is a ubiquitous feature of the human mutation spectrum that shapes ongoing human evolution, following typical patterns of inheritance one would expect for small-scale variation in neutrally evolving genomic regions. The lack of singleton variants in the 1k-30x dataset limits my ability to ascribe an accurate mutation rate to template switch mutagenesis, but it does however mean I do not need to consider the increased false discovery rates associated with genotyped singleton variants [42]. Nevertheless, this dataset has permitted the first assessment of event prevalence in human populations and allows me to investigate the genomic features associated with events in human populations.

4.7 Features associated with template switch mutations in human populations

4.7.1 Short template switch mutations explain thousands of mutation clusters and short indels within haplotype-resolved human genomes

The mutation cluster footprints left by template switching are essential for my event discovery pipeline and are the target of my alternate hypothesis. As discussed in §4.1, the mutation spectrum of clustered variants in VCFs are often used to assign causative mechanisms such as error-prone DNA lesion repair by Pol-ζ [117] and replication slippage [218] to observed population variation. Here I am specifically interested in assessing the template switch VCF footprint rather than the unidirectional pairHMM footprint (as in my great ape analysis, see Figure 3.8b), as this allows me to compare any mutation clusters attributed to template switching with features of mutational mechanisms characterised from population VCFs in previous studies.

To characterise the VCF footprint left by template switches, I tallied for each event the number of SNVs and the combined indel length associated with the event to create a two-dimensional distribution of template switch footprints (Figure 4.11). The majority of events leave a footprint of either a combination of insertions and deletions of length $\geqslant$ 5nt, or a complex mutation cluster consisting of 1-2 SNVs in addition to one or more indels (dark blue regions of Figure 4.11). The most extreme footprints create up to 14 clustered mismatches and indels with a combined length of up to 70nt — however note that any individual indels of this length would instead be called by a structural variant calling pipeline, and are not reflected in the SNV and indel calls used for study here.

Although my model comparison procedure allows me to reject that mutation clusters are created by independent, consecutive mutations within a small sequence window, I asked if these VCF footprints (Figure 4.11) share any of the mutational signatures associated with known causes of clustered mutagenesis and small indel formation which do not involve template switching. As mentioned in §4.1, the only parsimonious explanation for locally clustered complex mutations appears to be template switch mutagenesis (Chapter 2, [185]). The majority of single, short indels observed in human genomes are caused by either polymerase slippage events within repetitive sequence contexts or single, small deletions within complex sequence contexts [170, 218, 242]. Given that I remove low-complexity sequence regions and events which explain a single, small deletion (see §4.4.2 and §2.3.4), I can discard the simpler explanation of indel formation. This leaves the unlikely explanation that clusters composed solely of SNVs (which I denote MNV clusters) are potentially generated by an alternate mechanism that

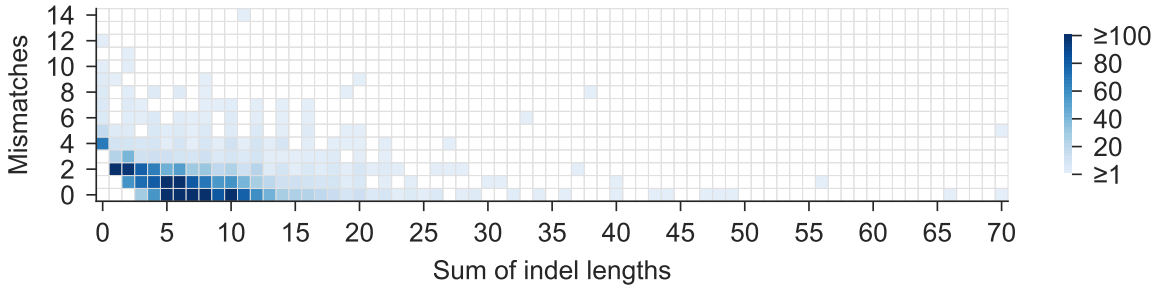

Figure 4.11: The SNV and indel mutation cluster footprint left in VCFs by template switch mutations. For each unique template switch mutation identified from the 1k-30x calls, SNVs and indels associated with the event are counted and summarised as a heatmap. The VCF records associated with each event are reported as a sum of SNV counts and a sum of indel lengths (independently of the total number of indel records).

co-occurs alongside a nearby sequence region of perfect reverse-complementarity that could falsely be identified as a ②→③ fragment.

To fully rule this out, I investigated mutation clusters associated with my events for signatures associated with MNV mutations in human populations. Causative mechanisms have typically been attributed to replication slippage [306] (which can be discarded, see above), Pol-ζ activity [117, 271, 306], and APOBEC activity [140]. Pol-ζ has previously been identified in large-scale human cohorts by assessing an enrichment of simultaneous (equal allele frequency), adjacent GC→AA and TC→AA dinucleotide mutations (and their reverse complements) compared to a random genomic background. I inspected all such dinucleotide mutations within the 1k-30x template switch VCF (n=518), and identified no such over-representation of Pol-ζ activity (Figure 4.12a). The small evidence for APOBEC activity (APOBEC activity wasn't observed in a more recent, much larger-scale investigation of 125,748 human exomes and 15,708 whole human genomes [306]) has been identified in a large collection of human trio exomes by inspecting simultaneous 2nt MNVs (2 SNVs in perfect linkage), separated by up to 20nt of spacer sequence, for enrichment of the CC→TC APOBEC motifs identified in studies of human cancer [9]. Given that MNV clusters associated with template switch mutations are composed minimally of four apparent simultaneous SNVs (Figure 4.11), a direct comparison with this study is not possible. Regardless, I inspected the SNVs contained in event-associated MNV clusters (n=620) for CC→TC signatures, and observed no obvious case for performing an enrichment analysis (Figure 4.12b). This indicates that the variant footprints which I attribute to template switch events are distinct from the only mechanisms known to create mutation clusters in human populations.

4.7.2 Many template switches are too short to permit capture by standard structural variant calling pipelines

It is useful to inspect the length distributions of the 1k-30x events, as it allows both a comparison with event lengths of those discovered in an evolutionary context (§3.3.3), as well as allowing me to assess how many of the identified events would indeed be missed by standard structural variant calling methods (refer back to §1.2.5 for a discussion of this problem).

As with events detected between great ape genomes (Figure 3.8a), template switches frequently cause a change in the post-event sequence length (Figure 4.13a) and the proportion of event types are consistent with the great ape analysis (Table 3.2), with ①-④-③-②/③-②-①-④ events (n=2046) being the most prevalent (Figure 4.13b). Recall that this event type results in an inverted repeat in the descendant sequence, and was also the most common event

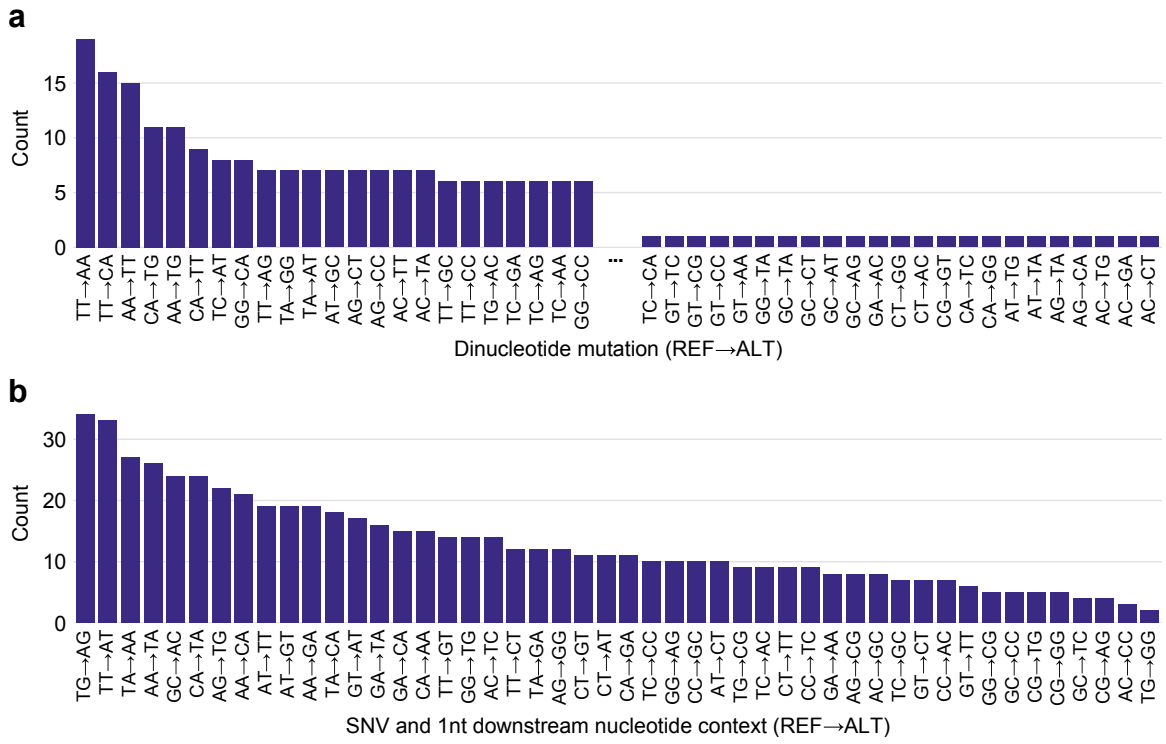

Figure 4.12: Variants attributed to template switching do not display the multinucleotide mutational signatures characteristic of APOBEC and Pol-ζ activity. (**a**) Counts of observed REF→ALT dinucleotide mutations across all 1k-30x template switch variants, sorted in descending order from left to right. The ellipsis indicates an x-axis break included for clarity — if present, dinucleotide mutational signatures indicative of alternative mutational pathways would be visible on the left of the sorted bar plot. (**b**) Count of unique SNVs and 1nt downstream contexts within MNV clusters. Labels on the x-axis indicate SNVs formatted as XZ→YZ, where X and Y are respectively the reference and alternate nucleotide, and Z is the nucleotide 1nt downstream of the SNV.

type observed in an evolutionary setting (see Table 3.2). This could suggest that a signature of a single inverted repeat alongside a change in sequence length is indeed the most common consequence of a short template switch, or that my methods have more power to detect the consequences of the mutational pathway(s) that generate these inverted repeats when compared to other rearrangement consequences.

The median ② → ③ length is 8 and the max length is 67 (x-axes of Figure 4.13b). The 2nt shorter median length compared to the great ape analysis may indicate a greater power to detect short events here due to the lower levels of divergence between two humans. That is, the longer divergence times separating great ape sequences increases the chance that proximal SNVs and indels may accumulate and obfuscate events, making the LPR contribution of very short

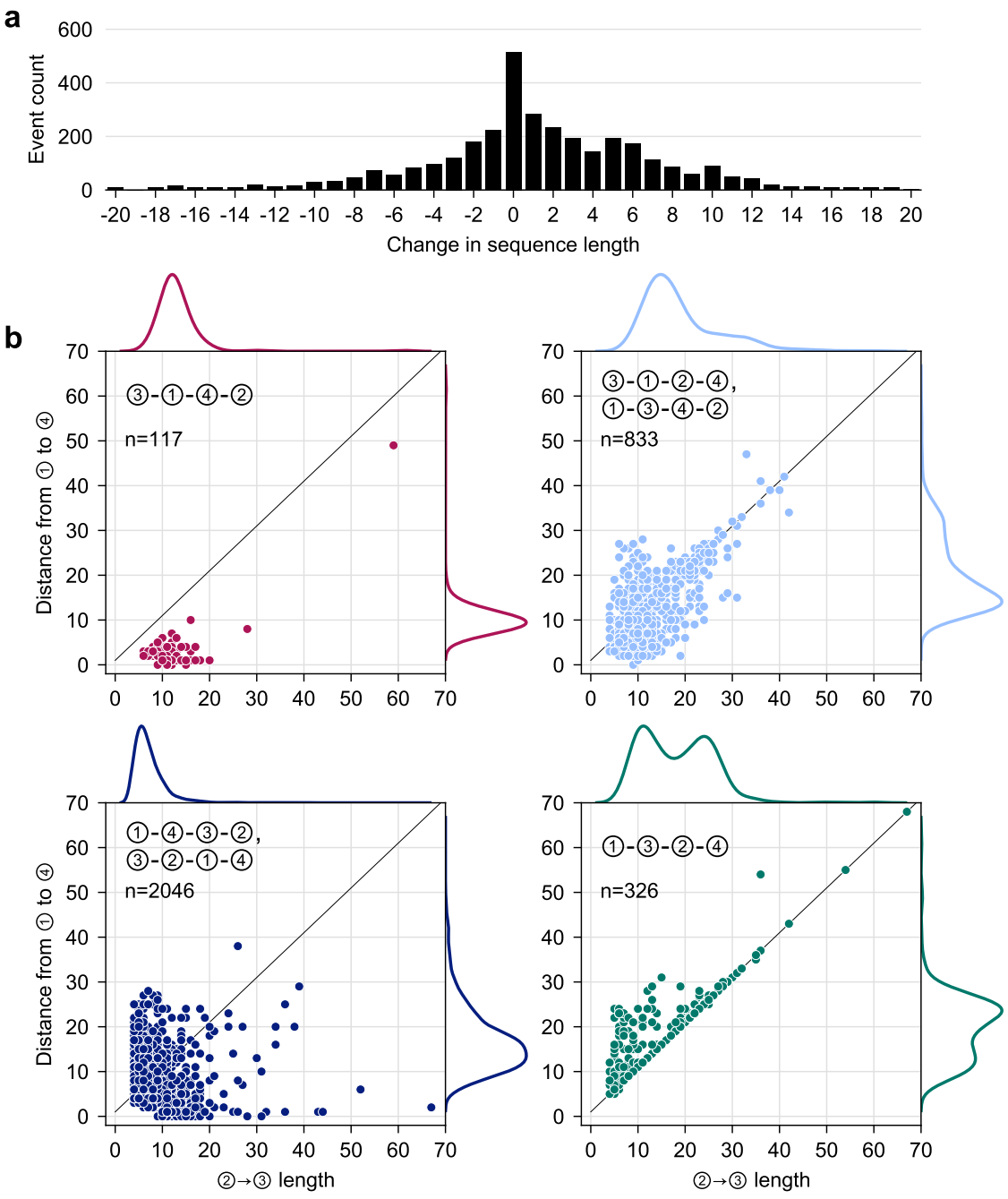

Figure 4.13: Template switches inferred in the 1k-30x callset typically result in no change to observable sequence length, and are generally too short to be detected by structural variant callers. (**a**) A histogram of net change in descendant sequence length caused by template switch mutations. (**b**) Comparison of ②→③ lengths and the corresponding ①→④ distances for all significant, unique events in the 1k-30x callset. Event types are distinguished and coloured as in Figure 3.8a, and as before the line $y = x + 1$ corresponds to no net change in sequence length.

②→③ fragments insufficient to be called as significant under the great ape LPR threshold (discussed for great apes in §3.3.3). Note however there is also less dispersion around the median ②→③ length here (median absolute deviation=2). The smaller maximum length detected is likely indicative of using solely SNV and indel calls, as larger events will be instead picked up by structural variant calling pipelines. Overall, this demonstrates that my methods are able to capture small-scale replication-based rearrangements in human resequencing data which will be missed by structural variant callers that seek to model variants $\geqslant$ 50nt or greater in length.

4.7.3 Events in the 1k-30x data are depleted in coding regions and a subset are in strong or perfect linkage with GWAS catalog variants

To assess if any genomic features are associated with human population template switch mutations and therefore may influence their formation, I performed an identical enrichment analysis to that described in §3.4 for the coordinates of great ape gold-standard events (see Figure 3.10). Again setting a significance threshold on the Bonferroni-corrected empirical p-values of 0.01 indicates a significant enrichment of events within transcription factor binding sites (mean $\log_2$-fold change$= 0.19 \pm 0.19$ SD, $p = 0.008$), and a significant depletion in protein coding regions (-1.93 ± 0.23, $p = 0.001$). Lowering this threshold to 0.05 further indicates a significant enrichment in super enhancers (0.22 ± 0.27, $p = 0.014$) and a significant depletion in exons (-0.42 ± 0.13, $p = 0.037$). These results are concordant with findings from events in the great apes.

Although template switch events are significantly depleted in coding regions, the Ensembl Variant Effect Predictor [203] annotations in the 1k-30x template switch mutation cluster VCF indicate that events introduce several potentially pathogenic frameshift variants (in genes TXNIP, OR5B21, OR6C2, ZNF223, RDH14, TEX44, NT5C1B-RDH14, TGM6, and APOL1), a splice donor variants (in NBPF25P) and a stop gained mutation (in OR6C2). The precise annotations indicated by these effect predictions need to be interpreted carefully, as although template switch events are identifiable from VCF mutation clusters, a direct representation of template switch variants in the VCFs may yield alternate annotations as chromosomal coordinates are altered. Nevertheless, I asked if variants associated with template switch mutations have a known deleterious consequence in humans.

Genome-wide association studies (GWAS) have proved successful at identifying SNVs which are significantly associated with many complex traits and diseases in humans [41, 183, 252, 315]. Ensembles of causal variants that are not directly genotyped during GWAS are

often in strong linkage disequilibrium with variants that have been significantly associated with a phenotype [57, 241, 267], and SNPs in strong linkage disequilibrium with large structural variants are enriched for GWAS hits [60, 80, 284]. These linked variants are typically located in regulatory regions of the human genome such as transcription factor binding sites [267, 323]. Given the small but significant enrichment of template switch variants within transcription factor binding sites (see above and Figure 3.10), I investigated if template switch variant coordinates either correspond to or are in strong linkage with GWAS hits. If indeed template switches either caused any GWAS variants that are associated with a phenotype or clinical outcome of interest, or are in linkage with a subset, it would certainly warrant experimental followup to further understand the genetic basis of the associated phenotypes.

The GWAS catalog [38] curates and aggregates many published GWAS results into a single resource, and at the time of writing contains a set of 276,696 significant phenotypic genome-wide associations from 5,273 publications. I downloaded the set of all GWAS catalog associations, retrieved the chromosomal coordinates of each entry, and extracted all GWAS catalog variants that are present in the 1k-30x callset. This yielded 152,197 variants from the 1k-30x dataset for linkage testing (55% of all GWAS catalog positions). I initially used `bcftools intersect` to check if any positions in the 1k-30x template switch mutation cluster VCF (i.e. the VCF containing all mutation cluster positions rather than one record per template switch) correspond directly to a position in the GWAS catalog, but found no intersections. I then used the 1k-30x template switch single-variant VCF for testing linkage as alleles within each event-associated mutation cluster identified by my pipeline are by definition in perfect linkage and on the same haplotype, so any GWAS SNPs linked to one event allele will be equally linked to the rest of the mutation cluster. This simplifies the enrichment analysis by permitting a comparison with randomly sampled single variants from a genome-wide background. I generate a single VCF for testing by merging the GWAS variants in the 1k-30x callset with the 1k-30x template switch single-variant VCF. Using this VCF, I calculated linkage disequilibrium between all pairs of variants within 1 megabase of each other using the r^2 method in `plink` [243]. This assumes two biallelic loci with alleles $\{X,x\}$ and $\{Y,y\}$, with associated allele frequencies π_X, π_x, π_Y, π_y and associated haplotype frequencies π_{XY}, π_{Xy}, π_{xY}, π_{xy}, giving $r^2 = (\pi_{XY} - \pi_X\pi_Y)^2/\pi_X\pi_Y\pi_x\pi_y$ [241, 243]. I retain only the r^2 values between template switch variants and GWAS hits (not between GWAS variants) using

```
plink --vcf template_switches_and_GWAS_hits.vcf.gz --show-tags \
      template_switch_IDs.txt --tag-kb 1000 --list-all --threads 4 \
      --tag-r2 {0.8,1.0}
```

where `--tag-r2 0.8` and `1.0` respectively retain pairs of variants in strong ($r^2 \geqslant 0.8$) and perfect linkage disequilibrium ($r^2 = 1.0$).

I identified two template switches in perfect linkage disequilibrium with GWAS variants and 49 in strong linkage disequilibrium across a range of complex phenotypes (see supplementary data). Both perfectly linked template switches are intronic variants not linked to disease phenotypes — one is weakly associated with body height [249] and the other with bilirubin levels [139]. Across a range of traits (see supplementary data), 17 of the strongly linked events are associated with a missense variant and 10 with regulatory region variants (as annotated by the Ensembl Variant Effect Predictor), which may indicate a biological link with the associated phenotypes [267].

Although all variants in perfect or strong linkage with a template switch are interesting, I asked if this represents a greater prevalence of linked GWAS variants than is expected by chance. I performed a permutation-based enrichment analysis. I sampled 10,000 random sets of variants from the 1k-30x calls, excluding regions which overlapped low-complexity annotation, GWAS variants, or template switch variants. For each of these variant sets, I generated a concatenated VCF containing both the variants belonging to the randomly sampled set, and the GWAS variants processed as above. I then used the `plink` command as above to identify and count GWAS variants in strong ($r^2 \geqslant 0.8$) and perfect ($r^2 = 1.0$) linkage disequilibrium with these randomly selected variant positions. I calculated mean $\log_2$-fold enrichment, standard deviation, and empirical p-values using the same procedure as previous enrichment analyses (see above and §3.4). An average of 59.2 ± 7.7 ($\pm$SD) and 4.6 ± 2.2 randomly sampled variants were respectively found in strong and perfect linkage disequilibrium with GWAS catalog variants. Comparing with the values of 49 and 2 for template switch variants, this represents a significant mean $\log_2$-fold depletion of -0.23 for strongly linked template switch variants (empirical p-value of 0.098), and a non-significant $\log_2$-fold depletion of -0.62 ($p = 0.328$) for perfectly linked template switch variants.

Despite the small enrichment of template switch mutations within transcription factor binding sites (which often contain an excess of linked causal variants), the depletion of linked variants compared to the randomly sampled genomic background is likely representative of the non-uniform distribution of template switches within functional genomic elements compared to the random background sample. Accounting for the density of positions within each functional element during sampling may be possible to generate marginally more informative estimates of enrichment/depletion. However, as the counts involved are so small, and any conclusions from this analysis require experimental validation to assert with confidence, I do not explore this further here.

4.7.4 Replication timing alone does not modulate event formation

Various mechanisms underlying human germline and somatic structural variant formation are associated with replication timing. For example, non-allelic homologous recombination is associated with early-replicating regions, while mechanisms such as FoSTeS, MMBIR, and non-homologous end-joining (all of which underlie large-scale templated insertions, refer back to §1.2.4) are associated with late-replicating regions [3, 156, 180]. Additionally, increasing distance from replication origins has been correlated with an overall increase in mutation rate [157]. As I am now working exclusively with human genomic variation data, I can readily relate experimental measurements associated with GRCh38 coordinates from projects such as ENCODE [294] to assess associations with template switch mutagenesis. Using these publicly available experimental data, I therefore asked if replication timing of the human genome or distance to the nearest replication origin may influence the initiation of a short template switch mutations. In particular, it would be interesting to observe any significant association with late replicating regions, as this would suggest that the large-scale template switch FoSTeS/MMBIR pathway(s) also operate at small scales.

Following [180], I retrieved (from ENCODE [294]) wavelet-smoothed signals of replication timing (measured using Repli-seq) in bedGraph format, collected from three cell lines: NHEK (normal skin), GM12878 (lymphoblastoid), and IMR90 (normal lung) [115]. I then used liftOver [145] to convert the bedGraph file into GRCh38 coordinates, and averaged the resulting signal across all three cell lines, producing an average replication time value per GRCh38 genomic coordinate, where high and low values are respectively early and late replicating. To assess distance from replication origins, I obtained a BED file containing ini-seq mapped human replication origins (cell line EJ30) from the supplementary information of [169], and converted these coordinates to GRCh38 using liftOver. To compare to these datasets, I used the BED file associated with the 1k-30x template switch mutation cluster VCF, where the start and end positions of each entry are the first and final VCF coordinates of the variants associated with the template switch. Replication timing values were then obtained for each template switch using `bedtools intersect`, and absolute distance from the nearest replication origin was calculated using `bedtools closest`. To assess the values of these experimental observations expected by chance, I generated a random background set of 230,000 loci from across GRCh38 (10,000 per autosome and chromosome X) using `bedtools random`, and retrieved associated values similarly. The comparison between template switch variants and the random genomic background is shown in Figure 4.14; a Mann-Whitney U test indicates that there is no significant difference between groups for replication timing ($p = 0.14$), but that template switch variants are significantly closer to to replication origins than the random background ($p = 3.36 \times 10^{-10}$)

(Figure 4.14). The lack of enrichment in late-replicating regions of the genome may suggests that the FoSTeS/MMBIR pathway(s) are not involved in the creation of the variants identified here; however, this does not disquality their involvement, it could simply suggest that their rearrangement consequences only become larger later in replication. It would be interesting to experimentally explore the distribution of rearrangement lengths generated by these pathways as a function of replication time. Again I note that the significant difference between proximity to replication origins may be influenced by an enrichment/depletion of template switches in some functional genomic regions compared to the functionally-agnostic random background sample. Nonetheless, it remains an interesting signal, as replication stress at loci proximal to replication origins has been shown to cause replication fork stalling and copy number variation in prokaryotes [275].

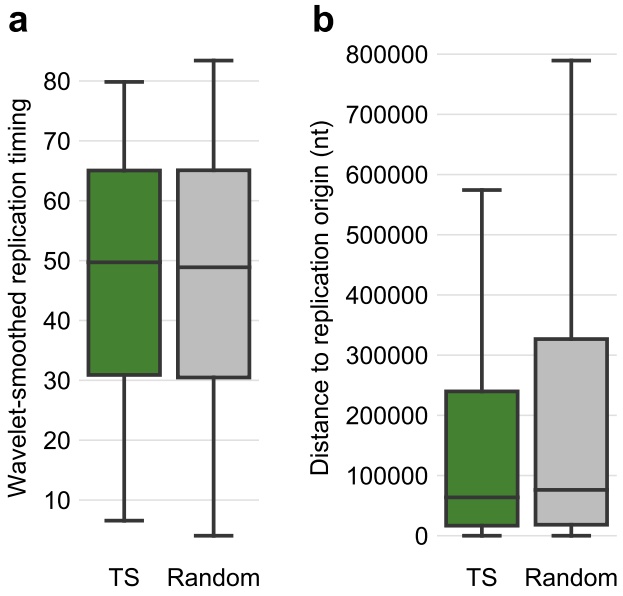

Figure 4.14: Short template switches are not associated with late replicating regions, and are significantly closer to replication origins than a randomly sampled genomic background. (**a**) The distribution of wavelet-smoothed replication timing signal for template switch (TS) loci compared to randomly sampled GRCh38 loci. Greater values on the y-axis are associated with earlier replication timing [115]. No significant difference is found between groups; $p = 0.14$, Mann-Whitney U test. (**b**) The distribution of distances to the nearest replication origin for template switch loci, compared to a randomly sampled genomic background. Template switches are typically closer to replication origins than randomly selected genomic coordinates; $p = 3.36 \times 10^{-10}$, Mann-Whitney U test. As in previous figures, boxes show the median, Q1, and Q3; whiskers show Q3/Q1 $\pm$ 1.5$\times$IQR. Outliers are hidden for clarity.

4.7.5 Short template switches are typically mediated by less than 5 nucleotides of microhomology

In human population analyses, mechanisms are often attributed to observed structural variants by investigating patterns of microhomology at their associated mapped breakpoints (i.e. the start and end coordinates) [49, 50, 80, 284]. Despite the name, micro"homology" here refers to short stretches of identity between one side of the initial breakpoint and one side of its alternately-located reciprocal breakpoint, rather than implying some shared evolutionary history between the two locations. Focusing on replication-based rearrangements that underlie large-scale (and often long-range) human structural variation, the FoSTeS and MMBIR pathways both utilise microhomology to invade an alternate location to restart a stalled or collapsed replication fork (see §1.2.4 and [120, 172]). As discussed in §4.1, these pathways have also been proposed as causative mechanisms underlying small indels in human populations [218].

To investigate patterns of microhomology associated with short-range template switch mutations, I first define microhomology for the initial switch event (① to ②) as the number of uninterrupted template switch nucleotides upstream of ① that match the equivalently located nucleotides upstream of ②, and for the return switch (③ to ④) I define microhomology similarly but instead I assess downstream identity (see Figure 4.15a for an example). Note that while this definition of microhomology is consistent with previous analyses of rearrangement breakpoints, here any inferences made about length may also be a function of the parameters used in the TSA pairHMM. That is, stretches of apparent microhomology associated with either switch event could also plausibly be a part of the ②→③ fragment under an alternate parameterisation, such as if I were to parameterise my model for maximal ②→③ length (see the discussion on this in §2.2.4 and consideration in Figure 2.5). I assess the length of microhomology tracts associated with all 1k-30x template switches by parsing the printed TSA pairHMM alignment output for all 1k-30x events directly (see supplementary data), and calculating uninterrupted identity under my definition for both the initial (① to ②) and return (③ to ④) switch events.

Many template switch events are not mediated by microhomology, but a large proportion of events have at least one nucleotide of microhomology at the associated initial and return switch sites (Figure 4.15b,c). For ① to ②: 39.1% of events have microhomology length 0, 56.2% have length 1–5, and 4.7% have length >5 (Figure 4.15b). For ③ to ④, these values are respectively 37.8%, 57.5%, and 4.7% (Figure 4.15c). Microhomology length does not appear to correlate with ②→③ length (Figure 4.15d), although an interesting signal of equal length microhomology at both the initial and return switch sites is present for some events

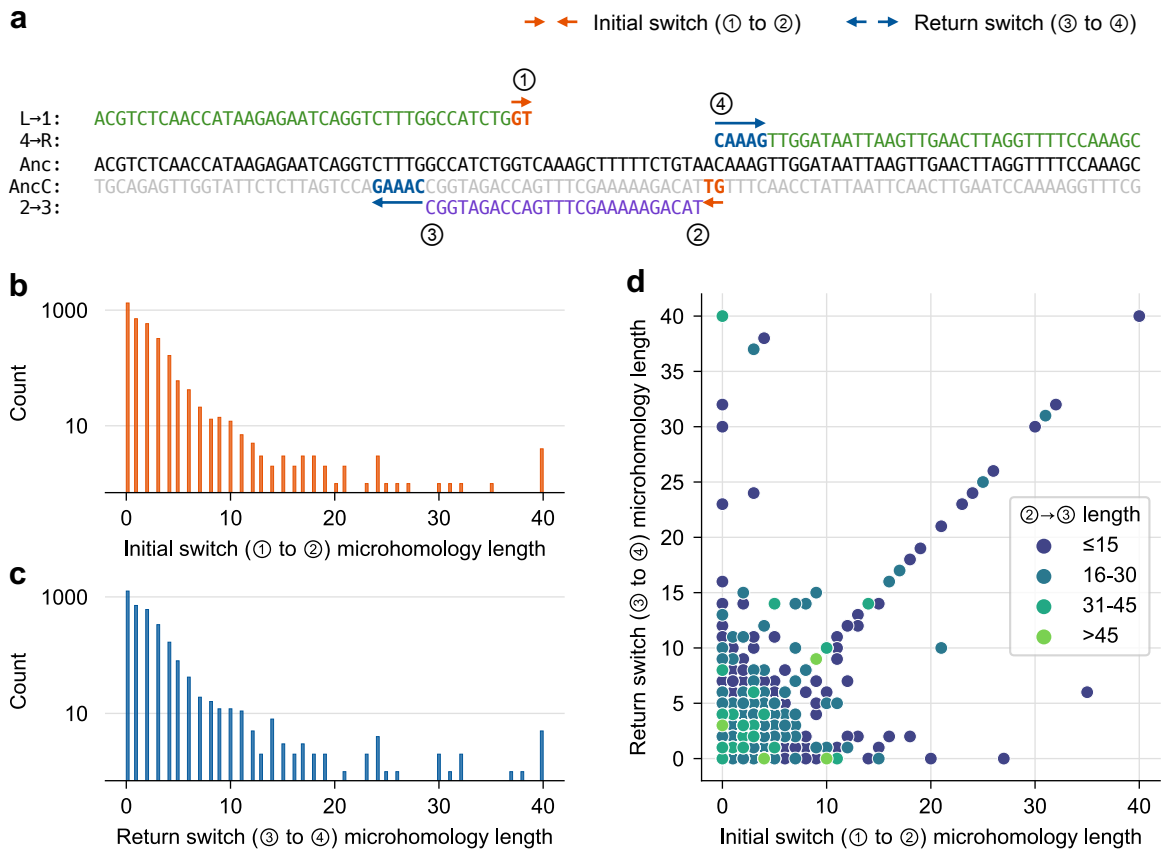

Figure 4.15: Microhomology length distributions for the initial and return switch events indicates that the FoSTeS/MMBIR pathway may not modulate many short template switch mutations. (**a**) An illustrative example of how microhomology length is calculated. For the initial template switch (the polymerase jump from ① to ②), microhomology is calculated as the number of uninterrupted template strand nucleotides upstream of ① which are equal to the equivalently located alternate template strand nucleotides upstream of ② (orange sequences, with microhomology orientation annotated using orange arrows). For the return switch (the polymerase jump from ③ to ④), microhomology is calculated similarly, but using the template sequence downstream of the direction of replication from ③ and ④ (blue sequence and blue arrows). Note here that upstream and downstream are defined according to the assumed direction of replication. For this particular example, the initial switch microhomology is 2nt, and the return switch microhomology is 5nt. (**b**, **c**) Distributions of microhomology length for the initial (b) and return (c) template switch events; note the log scale y-axes. (**d**) Microhomology length at the initial template switch positions compared to the return switch positions. Each point is coloured by the length of the ② → ③ region associated with the event.

(Figure 4.15d, $x = y$). As these cases represent sequence regions with long stretches of pre-existing reverse complement identity, stable secondary structure may have been involved in their formation (as in the signals identified for events in great ape genome evolution; Figure 3.11).

Although no association was observed solely between short templates switches and late replicating regions (§4.7.4), a hallmark of FoSTeS/MMBIR, I considered if concurrently considering the microhomology distributions in Figure 4.15b&c could reveal the involvement of these well-characterised large-scale template switch mutational pathways at small scales. FoSTeS/MMBIR have been associated with breakpoint microhomology as short as 2nt in the human germline [324]. Lengths of $\geqslant$6nt have been associated with germline short indels attributed to these pathways [218], and lengths of $\geqslant$10nt have been used to identify mutations caused through these mechanisms in human cancer ([180]). Given this broad range of microhomology lengths potentially associated with these pathways, it is difficult to ascertain their involvement solely from the length distributions shown in Figure 4.15b&c. I considered however that events which did occur later in replication may indeed have occurred through FoSTeS/MMBIR, and co-occurrence with longer microhomology tracts would help to suggest this.

To investigate this, I binned the 1k-30x events into those with microhomologies of length 0, 1, 2–5, 6–9, and $\geqslant$10nt at the initial and return switch positions (bins were chosen to cover the range of microhomologies associated with FoSTeS/MMBIR mentioned above). I then retrieved the replication timing for each event (Figure 4.16) using the data collected and processed as described in §4.7.4. I used a Kruskal-Wallis test to ask if there were differences in the resulting replication timing distributions for both the initial (Figure 4.16a) and return (Figure 4.16b) switch events, finding a non-significant difference for the initial switch events ($\chi^2(4) = 7.8$, $p = 0.1$) and a significant difference for the return switch events ($\chi^2(4) = 14.8$, $p = 0.005$). I performed follow-up Mann-Whitney U tests to compare the pairs of replication time distributions between events with no microhomology (length 0) at the return switch event, to those with at least one nucleotide of microhomology. Events in all microhomology bins except for the 6–9 bin are significantly more likely ($p < 0.05$) to occur in late replicating regions than events which occurred with no microhomology (Figure 4.16b). The greatest median shift towards later replication occurs in the $\geqslant$10 return switch event microhomology events (Figure 4.16b), and this length of breakpoint microhomology has been consistently attributed to FoSTeS/MMBIR across all previous studies investigating the mechanism(s) in large collections of human genomes [180, 218, 324]. This suggests that the FoSTeS/MMBIR pathway may indeed have generated this subset of 1k-30x short template switches. These events should be of great interest for any follow-up study that seek to investigate the scale at which these pathways can operate in the human genome.

As a final question, I asked if microhomology length distributions are distinct between event types, as this may indicate distinct causative pathways are involved in the formation of each event type. Comparing between event types, there is a significant difference in median

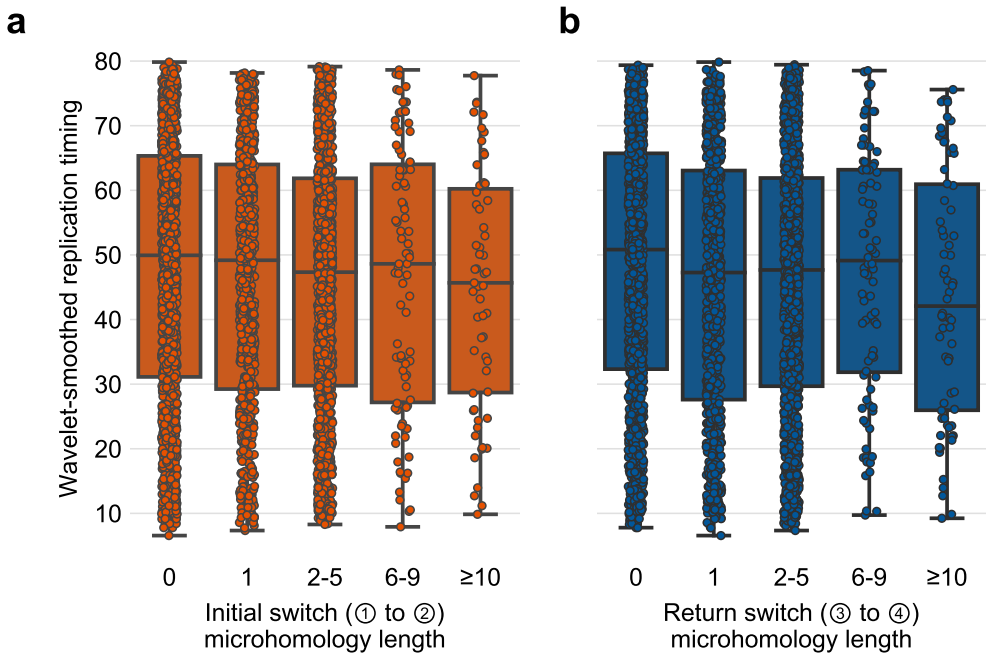

Figure 4.16: Short template switches which occur in later replicating regions frequently co-occur with return switch microhomology lengths typical of FoSTeS/MMBIR. The distribution of wavelet-smoothed replication timing signal for all 1k-30x events, binned based on microhomology lengths of 0, 1, 2–5, 6–9, and $\geqslant$ 10 at the (**a**) initial switch event, and (**b**) return switch event. Greater values on the y-axis are associated with earlier replication timing [115]. A Kruskal-Wallis test indicates a significant difference betweeng groups for the return switch events in (b). Follow-up pairwise Mann-Whitney U tests between events with no microhomology at the return switch site (the length 0 bin in (b)) were performed, indicating that events in length bin 1, 2–5, and $\geqslant$10 are significantly more likely to occur in late replicating regions ($p = 0.002$, $p = 0.001$, $p = 0.028$, respectively).

microhomology lengths for both the initial ① to ② switch event ($p \approx 0$, Kruskal-Wallis test; Figure 4.17, top) and the return ③ to ④ switch event ($p \approx 0$; Figure 4.17, bottom). The significant difference in microhomology length between types of event suggests there may be different mechanisms operating, or may reflect differential power to detect events of different types and sequence characteristics, however this would require experimental investigation in future to confirm.

4.8 *De novo* template switch mutagenesis

Throughout this thesis, I have used statistical methods to distinguish between multiple, independent, proximal mutational events (mutation clusters) and single-step mutations caused through

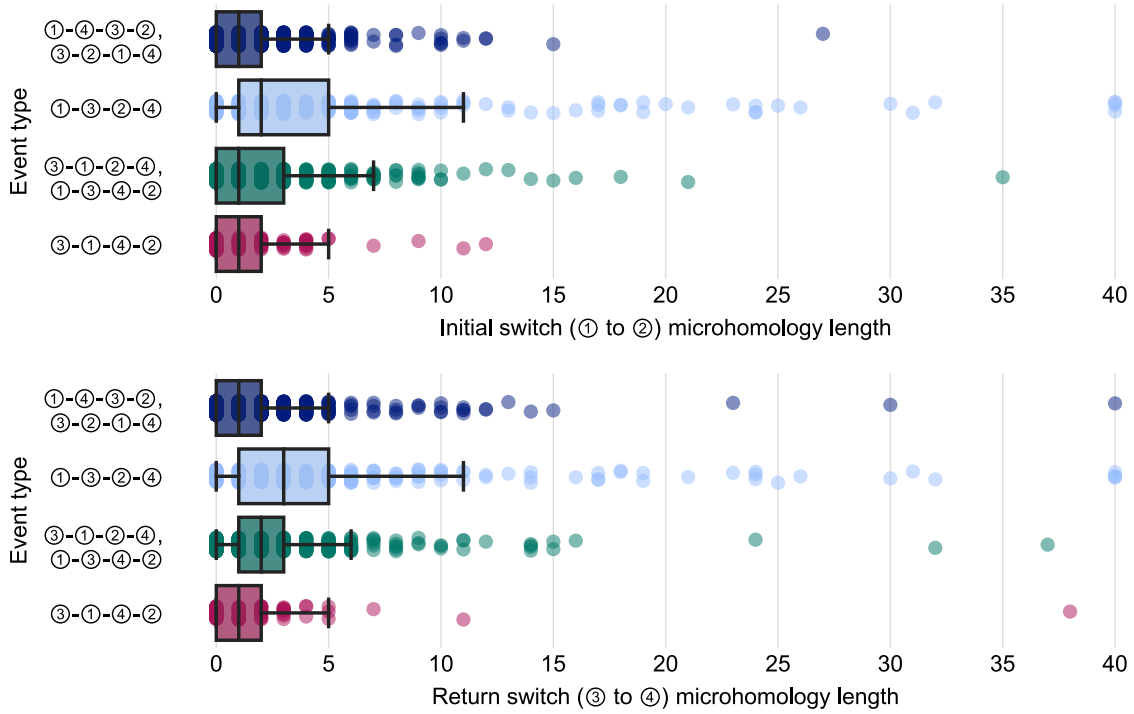

Figure 4.17: Differing microhomology length requirements per event type may suggest event type-specific causative mutational pathways. The length of microhomology for the initial switch event and the return switch event, broken down by event type. Event types are colour-coded to be consistent with Figure 4.13 and the length data is the same as shown in Figure 4.15.

template switching. This has required careful consideration, as concurrent mutational processes (SNVs and indels) may either create a false signal of template switching, or genuinely create clustered independent mutations during the divergence of the species or samples under study. In lieu of experimental observation, the most direct way I could unambiguously identify a case of template switching would be to detect *de novo* mutations which have occurred in a single generation.

A priori it is unlikely that template switch variants are present in *de novo* mutation sequencing data. Whole-genome sequencing of large population cohorts estimate that approximately 60-70 *de novo* mutations occur per meiosis in humans, the majority of which are SNVs and occur at a density of fewer than 4 SNVs per 3 megabase window [136, 154]. Besenbacher *et al.* [33] and Goldmann *et al.* [104] estimate that 2.4-3% of *de novo* mutations are generated as part of a multi-nucleotide mutation (i.e. a candidate template switch), however these estimates are based on mutations within 20kb window, and observed median distances are greater than 500nt in both studies. Multinucleotide *de novo* mutations rates are even lower and occur at greater

distances when considering population samples with developmental disorders [204, 209]. In addition, the majority of observed *de novo* insertions and deletions are 1nt in length and not present as part of a mutation cluster [33, 137, 193], and only around 3 indels ($\leqslant$20nt in length) are estimated to occur per generation [137, 152]. Indel discovery and phasing is also particularly challenging in single generations as it can be difficult to distinguish between read alignment artefacts and true variants, requiring false discovery rate to be tightly controlled possibly at a loss of power [137, 152, 246]. Combine all of this with my low mutation rate estimate (Equation 4.3; although calculated without singletons, so not directly comparable to a *de novo* setting) and unambiguously observing even a single template switch within population cohorts of sequenced family trios would be an interesting result.

I looked for evidence of *de novo* template switching in the 1k-HGSVC [80] (three trios) and Ice-Trios [137] (1548 trios) datasets described in §4.2. I applied my variant-discovery pipeline to each of these datasets in an identical manner to the 1k-30x dataset (described in §4.4), using identical filtering for candidate template switches. I call events as *de novo* if all variants within an event-associated mutation cluster are present in the child and absent in the parent genomes. Note that any issues with directionality are not an issue in a family trio setting, as the child by definition possesses the derived allele for each variant. From this procedure, I identified two candidate events in the Ice-Trios cohort (Probands 203 and 316; see supplementary data) and one candidate event in the 1k-HGSVC cohort (child: HG00514, mother: HG00513, father: HG00512; shown in Figure 4.18).

Consider the Ice-Trios dataset for contextualising these event counts compared to existing *de novo* mutation rates for SNVs and indels. There are 108,778 unique *de novo* mutations across the 1548 sequenced trios, which means that even if the variants associated with the identified template switches are located on the same chromosomal copy (recall this dataset is de-identified and excludes genotype information), template switches explain approximately 0.002% of *de novo* mutations. This is markedly lower than my estimated population-scaled mutation rate (Equation 4.3), and two orders of magnitude lower than a recent estimate of the *de novo* structural variant rate (0.16 events per generation, although note this covers all distinct classes of structural variant) estimated from 2396 family trios [28].

Overall this may indicate that the short-read *de novo* variant calling pipeline utilised by Jónsson *et al.* [137] struggled to confidently resolve template switches in the Ice-Trios dataset. This is unsurprising however, as the Ice-Trios variant calls were produced using a now-outdated GATK workflow involving the UnifiedGenotyper tool, which did not perform local *de novo* assembly for indels or clustered variants (as is now standard with the HaplotypeCaller tool that was used to produce the 1k-30x calls, discussed in §4.2). Additionally, false discovery

HG00514 chr17:47041238-47041360

VCF entries

#CHROM	POS	ID	REF	ALT
chr17	47041294	3906	T	C
chr17	47041297	3906	TTAC	T
chr17	47041302	3906	GAAATTTTT	G
chr17	47041311	3906	CT	C
chr17	47041314	3906	G	C
chr17	47041316	3906	A	G
chr17	47041319	3906	A	G
chr17	47041320	3906	A	T

pairHMM output

Template switch process

```
 L→1:                L AGTGAGGAATGGTCTAAGGCAAGTTAACACAAATCTCACTG 1
 4→R:                                                                                                4 TGCTCCTCGGGTTGCTGCAAGACTTTGGT R
 Anc:    ACTACTGAGCTAGAGAGTGAGGAATGGTCTAAGGCAAGTTAACACAAATCTCACTGTTCTTACAGAAATTTTTCTTGAATAAATGCTCCTCGGGTTGCTGCAAGACTTTGGT
AncC:    TGATGACTCGATCTCTCACTCCTTACCAGATTCCGTTCAATTGTGTTTAGAGTGACAAGAATGTCTTTAAAAAGAACTTATTTACGAGGAGCCCAACGACGTTCTGAAACCA
 2→3:  3 TGATGACTCGATCTC 2
```

Unidirectional alignment (log-probability: -58.2)

```
GTGAGGAATGGTCTAAGGCAAGTTAACACAAATCTCACTGcTCT----------------------AgctcagtagtTGCTCCTCGGGTTGCTGCAAGACTTTGGT ← HG00514
GTGAGGAATGGTCTAAGGCAAGTTAACACAAATCTCACTGTTCTTACAGAAATTTTTCTTGAATAAA----------TGCTCCTCGGGTTGCTGCAAGACTTTGGT ← GRCh38, HG00512,
                                                                                                               & HG00513
```

Template switch alignment (log-probability: -15)

```
GTGAGGAATGGTCTAAGGCAAGTTAACACAAATCTCACTGCTCTAGCTCAGTAGTTGCTCCTCGGGTTGCTGCAAGACTTTGGT
GTGAGGAATGGTCTAAGGCAAGTTAACACAAATCTCACTGCTCTAGCTCAGTAGTTGCTCCTCGGGTTGCTGCAAGACTTTGGT
```

Figure 4.18: A significant *de novo* event identified in the 1k-HGSVC calls. A heterozygous template switch identified in Southern Han Chinese (population CHS, super-population EAS) sample HG00514 which is absent in both the mother (HG00513) and father (HG00512). 3 deletion and 5 SNV records are alternatively explained by a single template switch mutation with a 15 nucleotide ②→③ region.

rate was strictly controlled and may have resulted in a loss of sensitivity for short indels. Both HaplotypeCaller and recently developed methods for accurate *de novo* variant calling that do not impose conservative hard filters on variant calls may therefore improve the resolution of single-generation event calling if applied to this dataset [61, 153]. The combination of long reads, Strand-Seq and *de novo* assembly has enabled the calling of a particularly striking event in the 1k-HGSVC however, involving a complex mutation cluster and a reasonably long ②→③ region (Figure 4.18). This suggests that this gold standard (and currently prohibitively expensive for most studies) combination of technologies for calling complex variation may also reveal larger numbers of template switches in future in both single generations and at the population scale.

4.9 Conclusions

In this chapter I have shown that short template switch variants are prevalent across human population variation datasets, and a pipeline which incorporates pairHMM realignment of

mutation clusters is sufficient for their identification from short-read data. As in hominid genome evolution, I have explained thousands of complex mutation clusters and short indels under a model of template switching, all of which were subject to stringent statistical thresholds. I have shown that these mutations are distributed within populations as expected under standard population genetics models of neutral evolution and established human population structure. This is supported by the consistent observation of event depletion within protein-coding regions of the genome. This demonstrates that short template switch mutagenesis is a ubiquitous feature in ongoing human evolution, consistently forming a part of the landscape of all mutations observed in human genomes.

All genomic feature associations tested are consistent with the hominid analysis, and I did not observe conclusive signals of association with additional features (GWAS variant linkage, replication timing and origin distance) tested here. Although I cannot directly ascribe a molecular pathway to template switch formation, I have provided evidence based on microhomology around switch points that events may be modulated by the FoSTeS [172] or MMBIR [120] pathways, which underlie many structural variants in human population [284], somatic [180], and *de novo* variation datasets [28]. This assertion of course requires that these pathways are novel in utilising microhomology for template switch-mediated rearrangements, and that short template switch mutations do not involve a distinct pathway yet to be identified. Further, assessing microhomology at switch points in a similar manner to break-point association with microhomology is complicated by possible uncertainty in the precise placement of switch points in the alignment as a function of the pairHMM parameters. Regardless, I have shown that there is no evidence of Pol-ζ, replication slippage, or APOBEC activity associated with template switch loci. This further indicates that the mechanisms modulating template switch formation are novel compared to the only known mechanisms of short indel and mutation cluster formation in human genomes.

The template switch events outlined here likely do not capture many rare events in human populations as my inferences make use of doubleton-resolution variant calls to ensure all variants belonging to each mutation cluster are present on the same chromosomal copy. In future, larger-sample datasets involving the application of multiple long-read and strand-specific sequencing technologies combined with *de novo* assembly of each chromosomal copy (the approach utilised by The Human Genome Structural Variation Consortium [80]) will likely permit better estimates of the template switch mutation rate, and greater statistical power for identifying associated genomic features and disease associations. Further, although events are callable from a linear representation of mutations in a VCF, in future it may be possible to more accurately call and represent short template switch mutations as part of a human

pangenome reference sequence [211] using now-established (but poorly adopted, in human genomics) graph-based data structures that can in principle represent any form of complex variation directly [81, 95, 132, 274].

Having considered template switch mutations in human germline variation datasets, an immediate question follows: do template switch mutations also occur somatically? As a final exploration into short template switch mutagenesis in human genomes, the next and final chapter (Chapter 5) before concluding this thesis will therefore assess event prevalence and potential acceleration in the genomes of human cancers.

Chapter 5

Exploring short template switch mutations in human cancer genomes

Chapter overview

Large-scale datasets of sequenced pairs of tumours alongside the corresponding normal tissue are for the first time becoming available through efforts such as The Pan-Cancer Analysis of Whole Genomes (PCAWG) study. Using the methods I have now established in Chapter 2 and the event discovery pipeline outlined in Chapter 4, in this chapter I seek to identify short template switch mutations in human cancer genomes. Specifying parameters in my models and generating distributions of my LPR test statistic under the null and alternate hypotheses requires more careful consideration when studying cancer genomes, and I explore these issues here. I then present a set of significant events identified from variant calls for a subset of 2703 paired normal tissue and tumour samples produced by the PCAWG study. I explore associations with genomic features, and outline template switches which may impact genes associated with human cancer.

Declaration

The content of this chapter has not previously appeared elsewhere. I performed all data collection, processing, analysis, and data visualisation.

Code and data availability

All code underlying the analysis of this chapter, in addition to any supplementary data files, are available from:
https://gitlab.com/conorwalker/phd_thesis/tree/main/chapter_5.

Data access statement

This chapter makes use of data which was produced as part of PCAWG study. In accordance with the data access policies of the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA), data produced by PCAWG which can potentially be used to identify sample donors (such as germline alleles) have restricted access. Access to a set of variant calls for 2703 matched normal-tumour samples was approved by applying through the TCGA Data Access Committee. In accordance with this access, no identifying information is included anywhere in this thesis.

5.1 Background

Somatic mutations arise within single cells of tissues and gradually accumulate throughout the lifetime of every human [10, 173, 174, 219] (and indeed all cellular life). As with the germline mutations I have considered thus far, somatic mutations arise spontaneously during cell division due to errors in DNA replication often caused by unrepaired or incorrectly repaired DNA damage [196]. The majority of somatic mutations occur as SNVs and short indels, with a smaller subset occurring as MNVs, structural variants, and in rare cases as gross chromosomal rearrangements called chromothripsis [196]. Most somatic mutations accumulate without any adverse impact on cellular functions, and even chromothripsis events are not universally harmful [200]. Occasionally, however, somatic mutations arise which provide a selective advance to single cells with respect to neighbouring cells within the microenvironment of the containing tissue. These are known as "driver mutations", as their occurrence drives clonal expansion in a manner conceptually similar to positive selection that causes an allele to tend towards fixation in germline evolution [283]. The uncontrolled cellular growth caused by driver mutations gives rise to the set of diseases collectively referred to as cancer.

Characterising the sequence variation present in the cancer genomes of tumours which arise from clonal expansion (with respect to the containing normal tissue) has facilitated a greater understanding of carcinogenesis by allowing driver mutations to be identified and distinguished from selectively neutral "passenger" mutations [283]. Recent efforts by the PCAWG consortium [44] have made use of large-scale, genome-wide sequencing of many tumours matched to their normal tissues to characterise somatic drivers underlying tumourigenesis [248], as well as permitting a description of the evolutionary history [98], patterns of structural variation and chromothripsis [62, 180], and mutational signatures [11] underlying human cancer. Throughout all of the analyses performed by the PCAWG consortium, great care was taken to call variants

accurately within each somatic tissue under study, and to attribute those mutations to specific causative pathways. This is vital in cancer genomics, as understanding the endogenous and/or exogenous sources of DNA damage which give rise to a particular cancer facilitates histology-specific clinical predictors and the identification of potential therapies [30, 44]. The aim of this chapter is therefore to identify and characterise short template switch mutations within existing matched normal tissue and tumour data, to understand the potential involvement of template switching in tumour progression.

As in my human population analysis, I am interested in capturing and explaining VCF mutation footprints composed of short indels and/or clustered SNVs under a model of template switching (see §4.1 and §4.7.1). Short indels are caused by many sources in cancer, mostly arising as single-nucleotide insertions and deletions through processes including replication slippage and defective mismatch repair, and through DNA damage caused by exogenous mutagens such as UV exposure and tobacco smoke [9, 11, 164]. Multi-nucleotide indels also arise through pathways including defective homologous recombination and error-prone double-strand break repair through non-homologous end-joining [11]. Clustered SNVs are rare in cancer, and have only been attributed to APOBEC activity (which I excluded as a source of false positives in human population events; §4.7.1) and translesion synthesis activity by the error-prone Pol-η [11, 285]. All of these processes are ascribed to observed variation based on similarity to a set of "mutational signatures", each of which has mixed levels of evidence associating it with a specific generative pathways (often with no experimental validation) [291]. Nevertheless, these mutational mechanisms and their signatures are important to be aware of, as each has the potential to generate false positive mutational footprints in my subsequent analyses.

The questions addressed by this chapter are broadly similar to those of Chapter 4, but asked in the context of human cancer. I first seek to understand the prevalence of template switching at small scales in human cancer genomes. By first evaluating the ability of my statistical methods to call significant events in human cancer, and then using a modified version of the pipeline described in §4.4, I will generate the first catalogue of short template switch variants in human cancer using the well-studied, large-sample set of tumor sequences provided by the PCAWG study [44]. As discussed in §1.2, the analogous term for a template switch mutation in the PCAWG study is a "templated insertion". These variants are suitably captured down to around 100nt in length (see Extended Figure 6 of [179]), and this chapter will provide a resolution to short templated insertions that are uncharacterised in PCAWG [179, 180]. Different tumour types are characterised by distinct distributions of genomic rearrangements including various types of complex mutations, translocation, inversions, "local

2-jumps", chromoplexies, and templated insertions [180]. In addition, tumours are typically defined by characteristic "mutational signatures" [8, 11]. I will assess if short template switch mutations are enriched or depleted in each tumour type, and ask if template switching defines any of the commonly defined mutational signatures [291]. This analysis will provide a more accurate description of the contribution of templated insertions in driving tumorigenesis across tumour types, in a setting in which understanding the exact mutational pathways which are driving disease is important for improving for early detection [68, 125, 233]. As in the two previous chapters, I also aim to assess the local sequence and genomic features that may mediate event initiation — again, many templated insertions in human cancer have been attributed to FoSTeS/MMBIR [180], and I will assess for the first time evidence for their activity at small scales in human cancer.

In the remainder of this chapter, I initially describe the set of PCAWG variant calls used to identify template switches, and the possible issues with this dataset for my purposes (§5.2). I then give considerable attention to the suitability of using simulations to parameterise my models when applied to human cancer genomes (§5.3). Next, I outline the final set of events called across all histological groups and consider if any tumour type is enriched for short template switchc mutations (§5.4). Finally, I explore potential associations with genomic features and cancer-associated genes of interest, as well as exploring the possible functional consequences for individual template switches on a case-by-basis basis (§5.5).

5.2 Overview of the PCAWG dataset

The PCAWG dataset used to identify template switches in cancer genomes throughout this chapter consists of SNV and $\leqslant$ 50nt indel calls in VCF format produced for 2703 white-listed PCAWG tumour samples (minimum 30x mean coverage) and the matched normal tissue samples (minimum 25x mean coverage) from 2658 donors [44]. These samples are grouped into 37 histologies throughout this chapter, as previously defined by the PCAWG working groups. Histologies are abbreviated to be consistent with the major PCAWG publications for convenience throughout; for example, the cervical squamous cell carcinoma cohort is referred to as Cervix-SCC. A full list of these abbreviations is provided as a supplementary data file (data/histologies.csv). 93% of samples were sequenced with 100 or 101nt reads, 2% were sequenced with shorter reads, and the remaining 5% were sequenced with longer reads (max. 151nt; see Supplementary Table S10 of [44] for a full breakdown). It is worth noting that 100nt reads should provide sufficient flanking sequence surrounding template switches for events to be called as clustered SNVs and/or short indels as previously (recall the median ②→③

lengths of 10 and 8 respectively in my hominid and human population analysis), rather than causing the read to not map. This is however shorter than the minimum read length of 150nt used to produce the 1k-30x dataset analysed throughout Chapter 4.

It is important to consider the pipelines that were used by the PCAWG consortium to call variants, as it may impact my ability to call template switches from their data. The final SNV/indel calls were generated by PCAWG for all 2703 samples using three separate pipelines: "Broad" (which uses MuTect [59] to call SNVs, and SvABA [304] to call indels), "EMBL/DKFZ" (samtools/bcftools [178] for SNVs, and Platypus [251] for indels), and "Sanger" (CaVEMan [135] for SNVs, cgpPindel for indels [245]). SNV-only calls were additionally produced using MuSE [85], and indel-only calls were additionally produced using SMuFIN [217]. A merged callset created from these five methods forms the final set of PCAWG variant calls. For the final callset, SNVs were retained if they were called by at least 2 out of 4 SNV pipelines, while indel calls were retained using more complex criteria involving a stacked logistic regression model trained on the variant calls produced by each pipeline (see [44] and [148] for full details).

I believe that the local reassembly procedure performed by HaplotypeCaller [237] at mutation clusters and short indels was important for accurately calling template switch-associated variants in human population data (see my earlier discussion in §4.2). It is therefore worth considering if local reassembly is performed by the variant calling pipelines in the PCAWG study. Local realignment is performed for both clustered mismatches and indels as part of the "GATK Best Practices" preprocessing workflow [71] used by MuTect and MuSE, and intrinsically by SvABA for indels. Platypus uses an approach similar to HaplotypeCaller, creating coloured de Bruijn graphs for reassembly of all candidate alternate haplotypes. No mention of realignment/reassembly of clustered SNVs or indels is mentioned for the other pipelines used [44]. It is interesting to note that the updated "MuTect2" (which was not used to generate PCAWG production calls) now performs graph-based haplotype reassembly at clustered mutations and indels similarly to HaplotypeCaller [237], and appears to significantly increase the precision and accuracy of both SNV and indel calling in cancer genomes [29].

Because agreement between 2 out of 4 of the callsets containing SNVs (produced by MuTect, samtools/bcftools, CaVEMan, and MuSE) is required to retain a SNV in the final callset, and two of these pipelines do involve local realignment at clusters of SNVs, there should be no issue in calling template switches which leave a footprint consisting solely of multiple point mutations. However, the VCF footprints left by template switches in human population data indicates that the majority of events leave a footprint of either single insertions, a combination of insertions and deletions, or a combination of SNVs, insertions, and/or deletions (§4.7.1).

Taking a deep dive into the performance of the machine learning methods used to merge indel callsets is beyond the scope of this section. However, the lack of consistent reassembly at indels and clustered SNVs across pipelines, combined with calling indels and SNVs separately, may mean that many template switch-associated variants are not included in the final PCAWG calls, which could impact my ability to call template switches from these data.

5.3 PairHMM parameter selection

5.3.1 Identifying candidate mutation clusters and indels in cancer

As in previous analyses, the first step in identifying events requires selecting suitable parameter values for my pairHMMs. To this end, I first identify candidate mutation clusters ($\geqslant 2$ variants within a 10nt window) and $\geqslant$ 5nt indels across all samples in the PCAWG dataset — these per-sample counts correspond to parameter C in the TSA pairHMM (see §2.2.4). As VCFs are used as the file format for storing variation information between tumours and the matched normal tissues, I can apply much of my event discovery pipeline outlined in my human population analysis to call template switch variants here (see §4.4.1 and Figure 4.3).

I apply steps (1) and (2) from the pipeline shown in §4.4.1, which first scans VCFs for mutation clusters and short indels and second produces Needleman-Wunsch alignments between each variant cluster and the corresponding region of the GRCh37 genome. In cancer genomes, I do not need to infer the ancestral state or concern myself with event directionality, as the tumour variants are always called with respect to the "ancestral" normal tissue. Note here that the matched normal tissue may itself contain a set of differences with respect to the reference genome that are proximal to the mutation cluster identified between the tumour and the matched tissue. In such rare cases, it is possible that I may misrepresent, for example, a single nucleotide in the normal tissue sequence by assuming it corresponds to the reference genome. As my PCAWG data access only provides variant calls between the matched and normal tissue, not between the normal tissue and reference genome, I am not able to assess the prevalence of such cases, although *a priori* these cases should be so rare that I deem this not to be an issue here.

5.3.2 Estimating pairHMM parameters for human cancer analysis

For each matched normal/tumour sample, I estimate values for my pairHMM parameters t (expected divergence), ρ (expected indels per substitution), λ (expected indel length), and C (mutation cluster count) from the associated VCFs. I calculate t as the sum of both SNVs and absolute indel lengths divided by the length of the GRCh37 reference human genome; ρ as

the count of both insertions and deletion variants divided by the count of substitutions in the sample, λ as the the average of those indel lengths, and C as the count of complex mutation clusters and/or single indels identified by my event discovery pipeline in §5.3.1. Note that only 2176 out of 2703 samples (81%) contain at least one mutation cluster or short indel required to satisfy my criteria for template-switch realignment, and these 2176 samples are the subject of study in the remainder of this chapter.

Divergence (Figure 5.1), indel lengths (Figure 5.2), and the ratio of insertions/deletions to SNVs (Figure 5.3) all vary drastically both within and between tumour types. Cancer genomes are substantially less diverged from normal tissue than the germline divergence levels I considered so far in my between-hominid and between-human analyses, and some samples display a several-fold increase or decrease in indel rate compared to a typical germline sample. As an extreme example, one endometrial adenocarcinoma (Uterus-AdenoCA) sample has more than one indel for every SNV (cut from the x-axis of Figure 5.3 for clarity).

As I do not have a prior expectation about the number of template switch events that will be identifiable per tumour sample, I cannot estimate N from the data. Recall that $\theta = N/CA$ is my template switch initiation penalty. In §2.2.4, I comment on the unimportance of accurately estimating N, as it is always normalised by the large product of the count of mutation clusters C and the event-specific alignment length A (C was respectively 7.9×10^6 and 1.48×10^5 in the human evolution and human population analyses). This allowed me to set N using earlier event prevalence estimates in each setting, derived by performing template switch alignments under a simpler set of parameters for the TSA pairHMM that did not include N (see §2.2.4 and §4.3.1). C is far smaller in cancer samples however, with a median of 8 mutation clusters and/or $\geqslant$5nt indels identified per sample, and a maximum of 617 (Figure 5.4). This means that inappropriately specifying N may impact my inferences in cancer, and I next explore this through simulation.

5.3.3 Establishing a LPR threshold using simulations tailored to human cancer genomes

The large variation in sample-specific model parameters t, λ, ρ, and C, combined with a lack of suitable estimates for parameter N, indicates that careful consideration is required to establish a LPR threshold in a cancer analysis setting. Unlike my previous analyses, in which I imposed an equal significance threshold on all detected events, the large variance in these possible model parameters suggests that I should consider establishing a set of histology-specific approximations to the null hypothesis LPR distribution using histology-wise simulations.

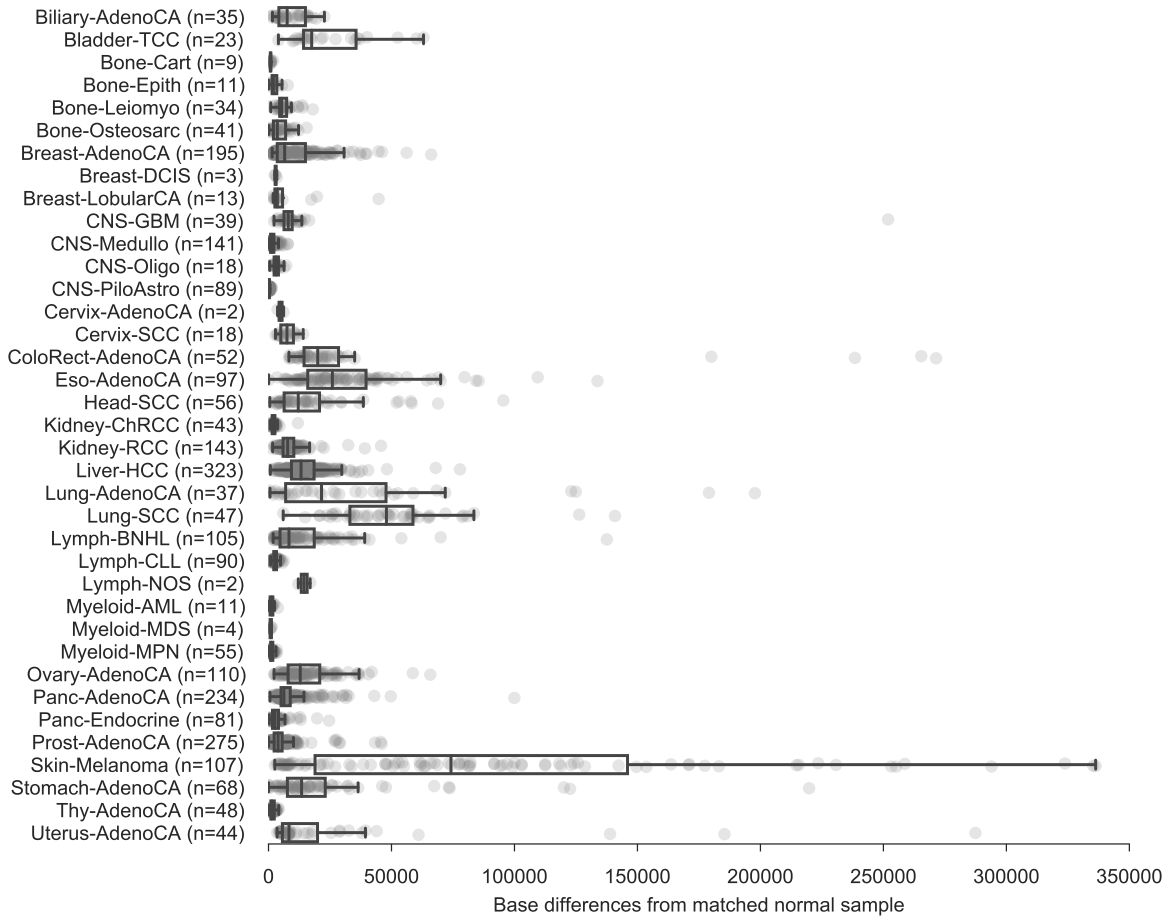

Figure 5.1: The number of nucleotide differences separating each tumour from the matched normal tissue across tumour types in the PCAWG dataset. The *x*-axis is cut at 350,000 for clarity.

Similarly, I also need to carefully consider if the sample-specific estimated parameter values should indeed be set per-histology when assessing candidate template switches, or if a cancer-wide set of parameter values are instead appropriate.

To establish a suitable LPR threshold across histologies, I performed simulations under the null model for each sample using the same procedure described previously (§2.3.3 and §4.3.1). Briefly, for each of the 2176 cancer samples with a candidate mutation cluster/indel, I simulate sequence evolution both with and without template switch mutations as in previous chapters. In the configuration files used by INDELible [89] for both sets of simulations, I specify the sample-specific estimated values of divergence, insertion rate, and deletion rates described above. Note that I now use INDELible's option to specify the insertion:SNV and deletion:SNV rates separately, as there are notable differences between the rates of each (Figure 5.3) for most

Figure 5.2: The average insertion and deletion lengths per histology group in the PCAWG dataset. The *x*-axis is cut at 20 for clarity.

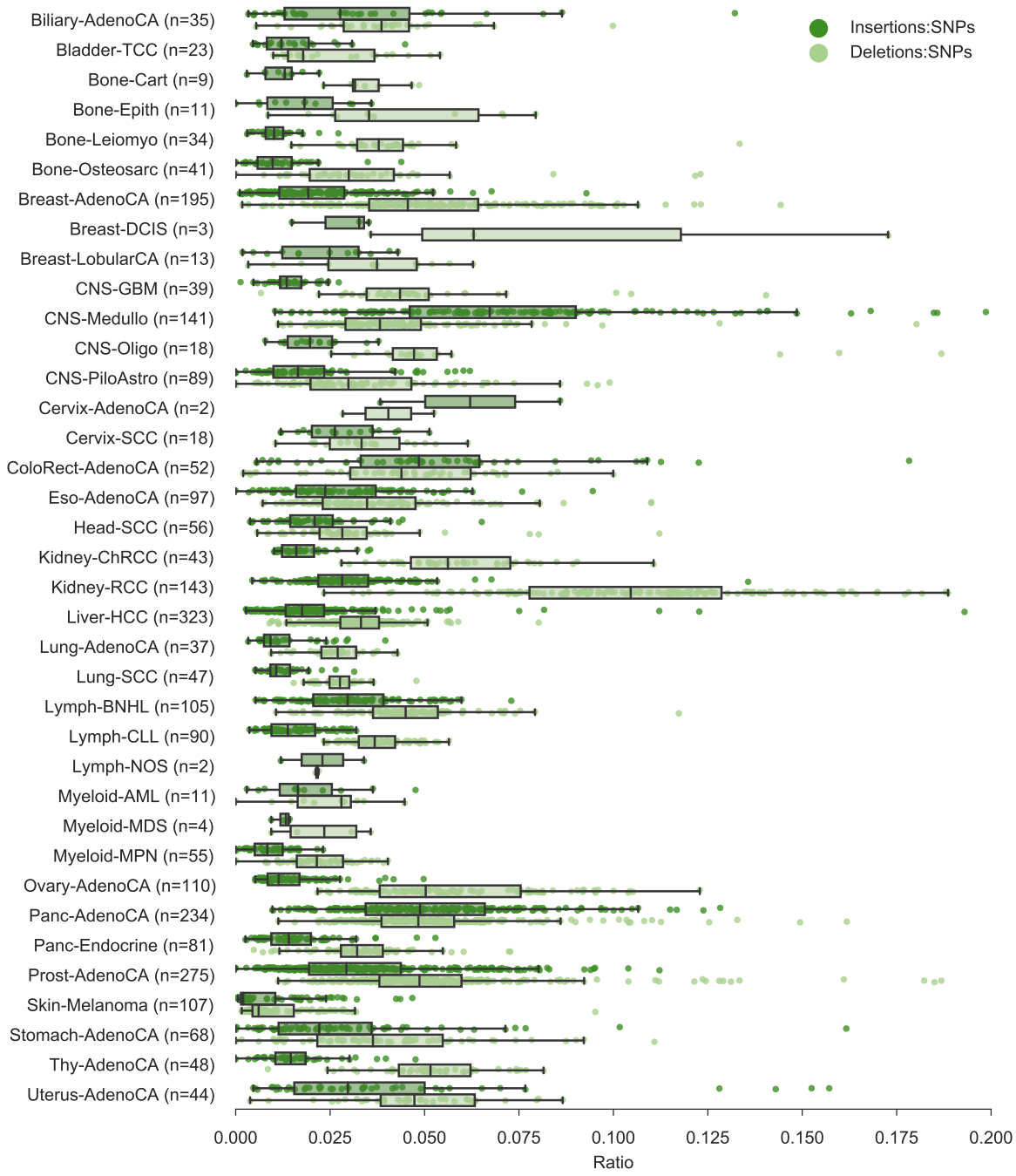

Figure 5.3: The ratio of insertions and deletions to single nucleotide polymorphisms across histologies in the PCAWG dataset. The x-axis is cut at a ratio of 0.2 for clarity,

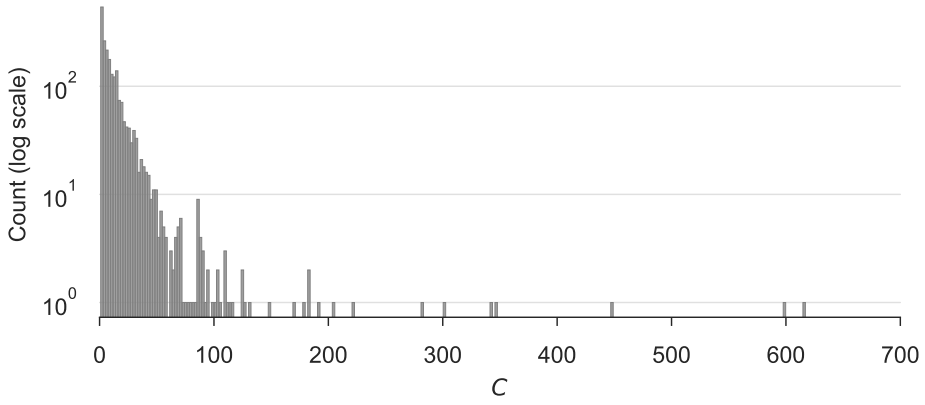

Figure 5.4: The count of mutation clusters and ⩾5nt indels (*C*) per PCAWG sample.

histological groups. I retain the power-law estimate of indel lengths from previous analyses, although I note that indel length distributions deviate more from this assumption than germline indels due to an excess of spontaneous large indels. However, when inspecting probability plots (generated using `scipy.stats.probplot` [301]) which compared random samples drawn from either a geometric or power law distribution (the two standard distributions for modelling indel lengths), to indel lengths observed in tumour samples, I consistently observed a better fit to a power law model of indel length formation in cancer (not shown).

As in previous simulations without template switching (my null hypothesis), 1000 sequences pairs of length 1000 were generated for each autosome and chromosome X, giving 23,000 simulated sequence pairs per tumour and 50,048,000 sequence pairs in total. For the simulations with template switching (my alternate hypothesis), I now simulate just 50 events per chromosome however, giving 1150 sequence pairs per sample and 2,502,400 pairs in total. This reduces the computational cost of both the simulations and downstream data processing whilst providing a sufficient estimate for the distribution of the LPR statistic under the alternate hypothesis. I used this reduced sample size as I am now performing far more simulations than in previous chapters, as I need to generate sequences under many sample-specific parameter sets rather than just under a small number of between-hominid or between-human parameter sets. It is still necessary (as demonstrated below) to simulate large number of sequences under the null however, as few mutation clusters arise simply by chance at the low levels of divergence simulated here.

For each simulated sequence pair, I align and perform model comparison for all identified mutation clusters and small indels as previously. Here, however, I use seven sets of pairHMM parameter values to realign each focal cluster/indel. The first three sets of parameters use values

of t, ρ, λ, and C fit to paired normal-tumour sample under study as described above, but are respectively defined by distinct values of $N \in \{1,5,10\}$. I assert that given the low levels of divergence associated with cancer samples, that unless short-range template switch mutagenesis was driving a particular cancer and has eluded all researchers to this date, identifying 1 event per sample is an optimistic expectation, and 5 and 10 respectively allow me to explore the importance of N whilst retaining reasonable expectations (albeit likely inflated for most samples). The second set utilise a "cancer-wide" set of parameter values, where I specify $t = 2.2 \times 10^{-6}$, $\lambda = 2.6$, $\rho = 0.06$, and $C = 8$ as the median values of these estimated parameters calculated across all 2176 samples — I again use $N \in \{1,5,10\}$ respectively to define the three parameter sets. The final set utilises identical parameters to those used in my human population analysis (Table 4.2), allowing me to assess the impact on test power when using parameters not specifically fit to the cancer mutational landscape. Performing cluster realignment under my models using each of these parameter sets will allow me to assess how important specific parameter values are for establishing a final LPR threshold. For example I may observe large variance in the type II error of my LPR test as the LPR threshold changes with each parameter set. Recall that I did assess the impact of fixing t on event discovery in §3.2.1 (see Figure 3.2), and found that my inferences are robust to misspecification of this particular parameter in a germline evolution setting.

These simulated sequences were subjected to the same filtering employed in my previous simulations, i.e. I require all four nucleotides in the ②→③ region and that a candidate event is not unidirectionally defined by a sole deletion. For each parameter set, I impose a LPR cutoff that removes all potential false positives from that set of simulations. The results of this procedure are shown in Figure 5.5. Despite simulating over 5×10^{10} nucleotides under my null hypothesis per parameter set, the low levels of divergence associated with cancer samples compared to their corresponding normal tissue means that an average of only 697 candidate false positive events (background mutation clusters/indels) were generated across parameter sets. This highlights that an alternative approach to generating a significance threshold for my model comparison procedure may be appropriate in a cancer analysis setting. Nevertheless, I am able to establish a LPR across all simulation parameter sets such that my model comparison procedure performs well at distinguishing between both sets of simulations (Figure 5.5; AUC ≈ 0.99 across all simulations).

Overall, this analysis indicates that the precise values of parameters may not be important for my inferences in cancer, as long as an appropriately conservative LPR threshold is specified. Indeed, even using parameter values from my human population analysis performed well (Figure 5.5, bottom; mean AUC 0.001 greater than sample-specific and cancer-specific parameter

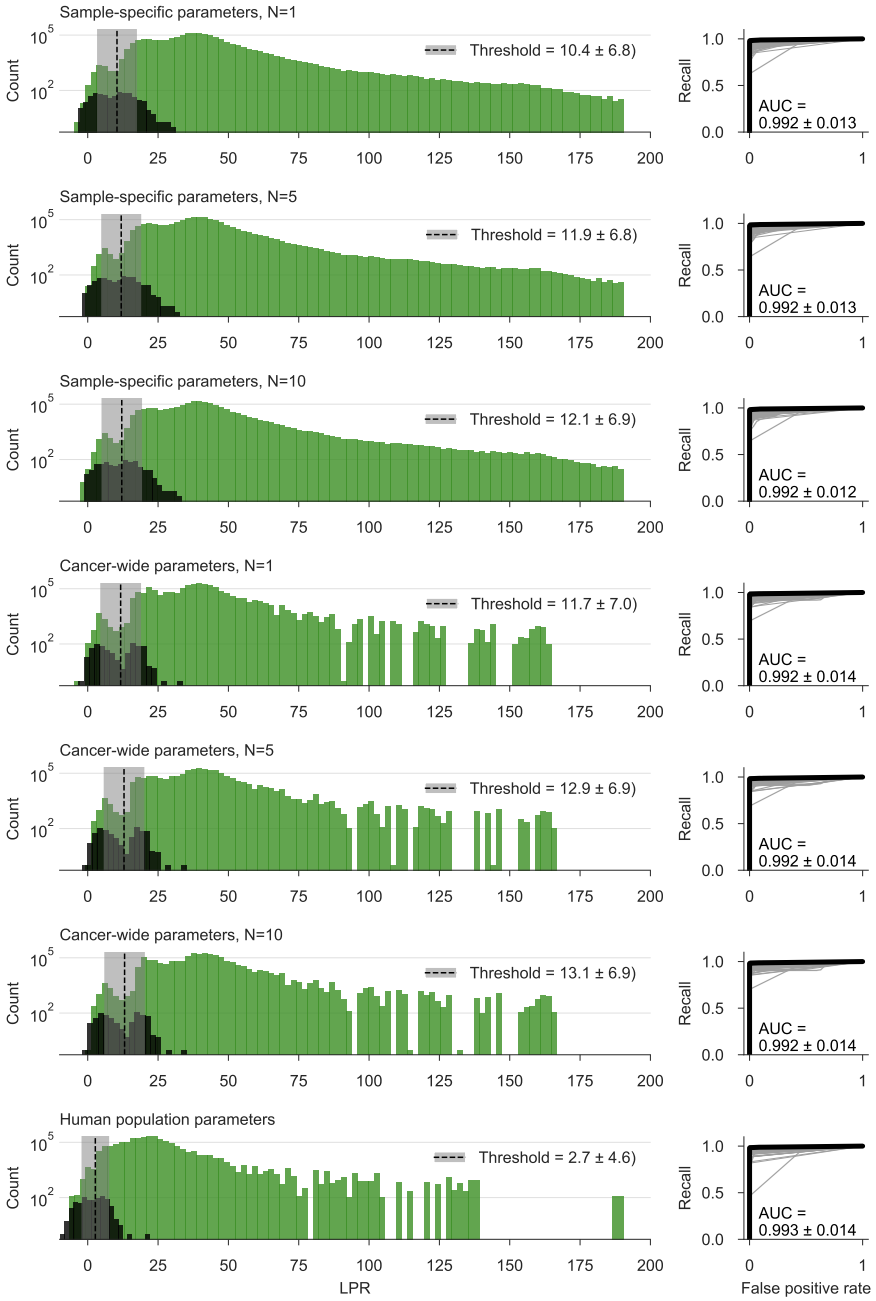

Figure 5.5: Distinguishing between the two sets of cancer evolutionary simulations under several sets of pairHMM parameters. For each parameter set (top to bottom), the left subplots show the LPR histograms associated with all post-filtered template switch events simulated across samples (true positives, green), the LPR histograms for background mutation clusters (potential false positives, black), and mean ± standard deviation (SD) LPR thresholds determined by evaluating each set of sample simulations separately and selecting a LPR that results in no false positives in these simulations (as in Figure 4.1). Note that because I am indicating the mean LPR threshold established across all sample simulations in that parameter set, some black bins appear to the right of the threshold on the *x*-axis. The right subplots show in grey the ROC curves for discriminating between the two sets of simulations when assessing each sample separately, and in black the mean value of these curves; AUC±SD is indicated.

values), differing only in the magnitude of the LPR threshold. The average LPR threshold is also fairly stable across sample-specific and cancer-wide simulations, confirming that the precise value of N is unimportant (as reasoned in §2.2.4) — and I use a fixed value of $N = 5$ for all subsequent analysis in this chapter. This means that I can specify a LPR threshold that is applicable to both cancer-wide and sample-specific parameter sets simultaneously if desired. Note however that the LPR of two background mutation clusters did cause poor performance as measured by AUC (see the count=1 black bars with LPR > 25 and corresponding grey ROC curves in e.g. the "Cancer-wide parameters" subplots in Figure 5.5). These cases are likely caused either by chance due to the large number of simulations performed, or are caused by sampling background human sequence from low complexity regions of the genome for use in simulation, and these sequences would be filtered out as masked regions in real analysis. Regardless, observing high LPR values for some genuine background mutation clusters indicates that a more conservative LPR threshold may be required here than in previous settings.

5.3.4 Event inference in cancer is reasonably robust to pairHMM parameter misspecification

Before settling on a final LPR threshold and calling significant events, I wanted to assess the impact of changing pairHMM parameters t, ρ, λ, and C on identifying events in the PCAWG dataset. My simulations showed that the LPR-based model comparison procedure is effective at distinguishing between background mutation clusters and simulated template switch events (Figure 5.5). Some individual simulation runs however produce background mutation clusters with large LPR values, indicating that background mutations could more readily manifest as false positives in cancer. Further, it is not clear if sample-specific values provide a real advantage over cancer-wide parameters that more readily allow for cross-histology comparisons of significance.

Using the mutation clusters and short indels identified in the PCAWG dataset earlier (§5.3.1), I therefore perform realignment under my pairHMMs (i.e. my template switch discovery procedure) for every identified mutation cluster under a range of values of t, λ, ρ, and C. 2880 combinations of $t \in \{8 \times 10^{-9}, 8 \times 10^{-8}, \ldots, 8 \times 10^{-4}\}$, $\lambda \in \{1, 2, \ldots, 20\}$, $\rho \in \{0.05, 0.10, 0.15, 0.20\}$, and $C \in \{1, 5, 20, 50, 100, 200\}$ were tested to cover the majority of observed sample-specific values for these parameters (estimated in §5.3.2 and respectively shown for each sample in Figure 5.1 (t), Figure 5.2 (λ), Figure 5.3 (ρ), and Figure 5.4 (C)). I maintain a fixed value of $N = 5$ for all alignments as outlined in §5.3.3. I use a LPR threshold of 13 throughout this procedure, as my simulations indicate that this is a suitable average

threshold for both sample-specific and cancer-wide parameter values (Figure 5.5). I show the mean, minimum, and maximum count of events identified in each sample across all tested parameter values in Figure 5.6, colouring samples by the divergence from the matched normal tissue, as tumours with a greater divergence may contain more false positive mutation clusters by chance.

From this procedure, 704 samples were found to contain at least one significant template switch event under at least one of the 2880 parameter value sets tested (Figure 5.6). Three features of interest arise from this procedure. First, several samples with relatively high levels of divergence show large variation in the number of events detected under different parameter values — as an extreme example, both zero and 36 significant events were found in one colorectal adenocarcinoma sample (rightmost point in the ColoRect-AdenoCA subplot of Figure 5.6). Second, many samples contain zero events depending on the parameters used; while events found in these samples (often a single event) may indeed represent true parameter-sensitive template switches, these cases may also represent mutation clusters that present as a significant template switch when using unsuitable model parameters. Third, events in some samples are consistently identified as significant regardless of the parameter values used, which is indicative of these events being true template switches.

To investigate what may cause the issue of extreme variation in the number of template switches inferred, I inspect six "extreme" samples that had a maximum of > 10 events discovered under at least one set of parameter values, and a minimum of zero under at least one set — these are the two rightmost samples from each the ColoRectAdenoCA, Lymph-BNHL, and Skin-Melanoma subplots shown in Figure 5.6. For each of these samples, I compare the values of t, ρ, λ, and C which yielded the maximum number of events, with the sample-specific estimates of each parameter value as determined in §5.3.2. Interestingly, grossly misspecifying t seemed to be the most important determining factor for inferring large numbers of significant template switches, as the maximum number of events for each sample was always found at the smallest value of t assessed (Figure 5.7). Other test parameter values showed mixed consistency with the estimated sample-specific values. Overestimating λ was required to produce these maximum event sets in two sample (Figure 5.7; ColoRect-AdenoCA-1, Skin-Melanoma-2). The specific value of ρ seems unimportant for event inference, as the maximum number of events found was consistent across values of ρ (although note that these particular samples have lower indel rates than captured in the values of ρ assessed here). Misspecifying C was also necessary to produce the maximum number of significant events, with five of the six samples assessed (excluding Lymph-BNHL-2) requiring $C \leqslant 50$ for maximised candidate event discovery.

Figure 5.6: Inspecting the min, mean, and max significant events discovered per sample across all parameters values tested indicates that the pairHMMs are mostly robust to parameter value misspecification. Each subplot includes samples from a distinct histological group; the count of samples with at least one event detected are specified. Points indicate the mean count of significant events per sample, grey error bars indicate the minimum and maximum counts; points are equally spaced on the *x*-axes and ordered by mean event count. Each point is coloured according to the number of nucleotide differences (SNV count + absolute indel lengths) between the tumour and the matched normal tissue.

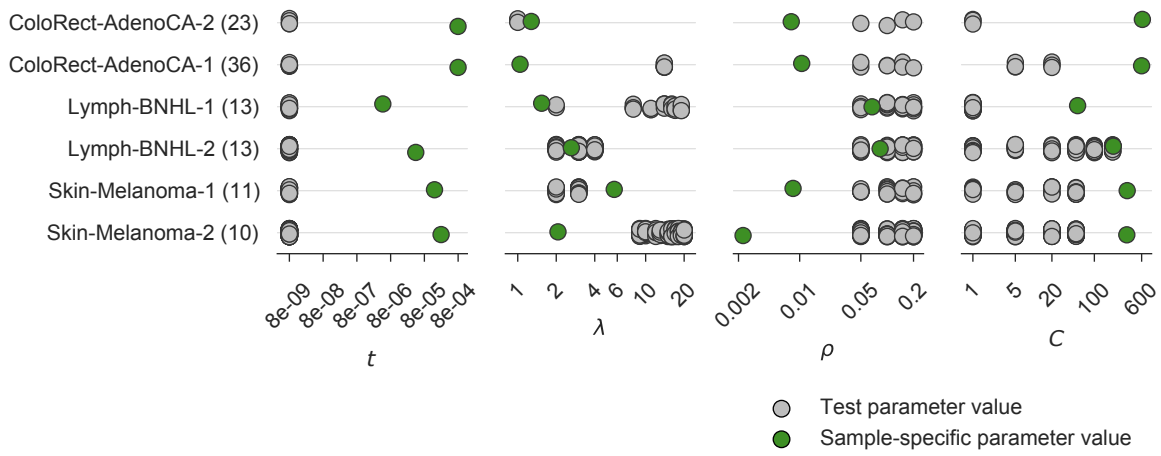

Figure 5.7: Large variation in the count of significant template switches inferred for a subset of samples under various parameter values (see Figure 5.6) only occurs when grossly misspecifying parameter values. For six samples that had a relatively large maximum number of template switch events identified as significant across tested parameter values (upper bound of the grey error bars in Figure 5.6, labelled in brackets on the *y*-axis here), each subplot shows all test parameter values (grey points) that yielded the maximum number of significant template switches for that sample. These values are compared to the estimated sample-specific parameter values (green points) estimated in §5.3.2.

Considering then how convincing the template switches in these maximal event sets are, the majority of events are defined by three SNPs in the paired normal tissue-tumour VCF, with a mean ②→③ length of 4. Recall that in my human population analysis, no events were identified as significant when defined solely by three VCF SNV records (see Figure 4.11). In combination, this indicates that misspecifying values of t, λ, and C may inflate LPR values and produce an excess of false positive events with the minimum permitted ②→③ length.

Finally, I assessed the LPR distributions for events from samples that could possess zero events depending on the parameter value set, to samples that minimally always possessed at least one event. I reasoned that events that are consistently found regardless of parameter set are more likely to be true positives than those which are only sometimes identified under various parameter sets. The LPR distributions of these event sets highlights that some events which are only identified in a subset of the parameter value space are associated with smaller LPR values (LPR < 20) compared to events which are always identified, regardless of the parameter values used (Figure 5.8).

Based on Figure 5.8, and in combination with the large LPR values associated with a subset of background mutation clusters in my simulations (Figure 5.5), I decide (by eye) on a final LPR threshold of 23 for events called across histologies, this is shown in Figure 5.8, and results

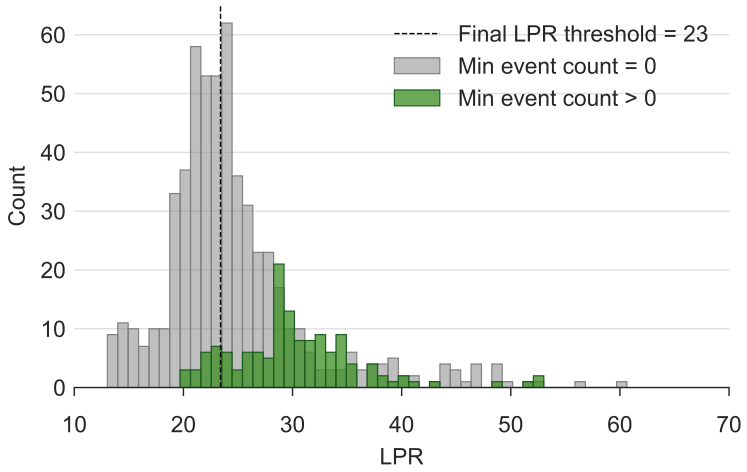

Figure 5.8: The final LPR threshold applied to candidate template switches across all PCAWG samples is chosen to reduce false positives at the cost of some recall. LPR histogram for events found across samples in a subset of the evaluated pairHMM parameter space (grey bars), compared to events consistently found regardless of the specified parameter values (green bars). The LPR threshold used in subsequent analysis is shown as a dashed black line. Note that LPRs are minimally 13 due to the threshold imposed on events as determined from my simulations shown in Figure 5.5.

in a final LPR distribution comparable to the LPR distribution associated with events found in my hominid analysis (see Figure 3.8c). This removes many of the low confidence events (Figure 5.8; grey bins to the left of the threshold) whilst retaining most of the high confidence events (Figure 5.8; grey bins to the right of the threshold). Further, to address events being found only in a subset of the parameter space (which may yield false positives, see Figure 5.7), I subsequently require that events are identified with a LPR$\geqslant$ 23 under both the sample-specific parameter values, as well as the median cancer-wide values used in my simulations (§5.3.3).

5.4 Short template switch events in human cancer genomes

5.4.1 The procedure used to determine the final set of significant events

To produce the final callset of significant template switch events, I scan the alignments of mutation clusters and indels produced by my VCF event discovery pipeline (outlined in §5.3.1) using both pairHMMs, under both the cancer-wide and sample-specific parameter sets. The full set of filters that each event must pass to be called as significant in cancer are:

1. a LPR $\geqslant$ 23 under both cancer-wide and sample-specific parameter sets,

2. events are not located within a low complexity region of the GRCh37 assembly (`bedtools intersect` is used to compare variant coordinates associated with template switch events for overlaps with GRCh37 RepeatMasker [276] annotations obtained from the UCSC table browser [142]),

3. the ② → ③ region must contain all four nucleotides,

4. the event is not defined solely by a single deletion,

5. and the final alignment contains at most one mismatch.

In cancer, I impose the final qualitative filter (5) above on alignment quality (similar to [185]) to ensure that significant template switch alignments have a high pairwise sequence identity. This was addressed in previous chapters using an approach based on sampling of genome-wide length-normalised alignments (see Figure 3.3 and Figure 4.2). In cancer however, sampling random regions of the human genome is problematic due to the low levels of divergence of each tumour, meaning almost every sampled region would simply contain no variants and the resulting expected alignment probability would be derived almost entirely from 100% identity alignments. Indeed, in my human population analysis, 87% of alignment regions sampled to derive this statistic were from regions of the genome that contain no observable variants (rightmost bin of Figure 4.2), and between-human samples are typically several orders of magnitude more diverged than a matched normal tissue and tumour sample. Also note that I do not impose a filter to remove ① < ④ events (filter 6 in my human population analysis; §4.4.2), as I did not observe any noteworthy enrichment of these events when testing preliminary parameter values.

5.4.2 Short template switch mutations are present in a subset of PCAWG samples and occur independently of tumour divergence

After filtering, I identified 128 significant template switch events in 120 samples across 20 histologies, 118 of which are unique (Figure 5.9). 115 are unique to a single sample, one event is found independently in nine prostate adenocarcinoma (Prost-AdenoCA) samples, one event is found independently in two Prost-AdenoCA samples, and one is found independently in a sample from each of kidney renal cell carcinoma (Kidney-RCC) and head and neck squamous cell carcinoma (Head-SCC). I also identified 24 and 3 events that are significant under only the cancer-wide or sample-specific parameter sets, respectively, but I do not consider these further.

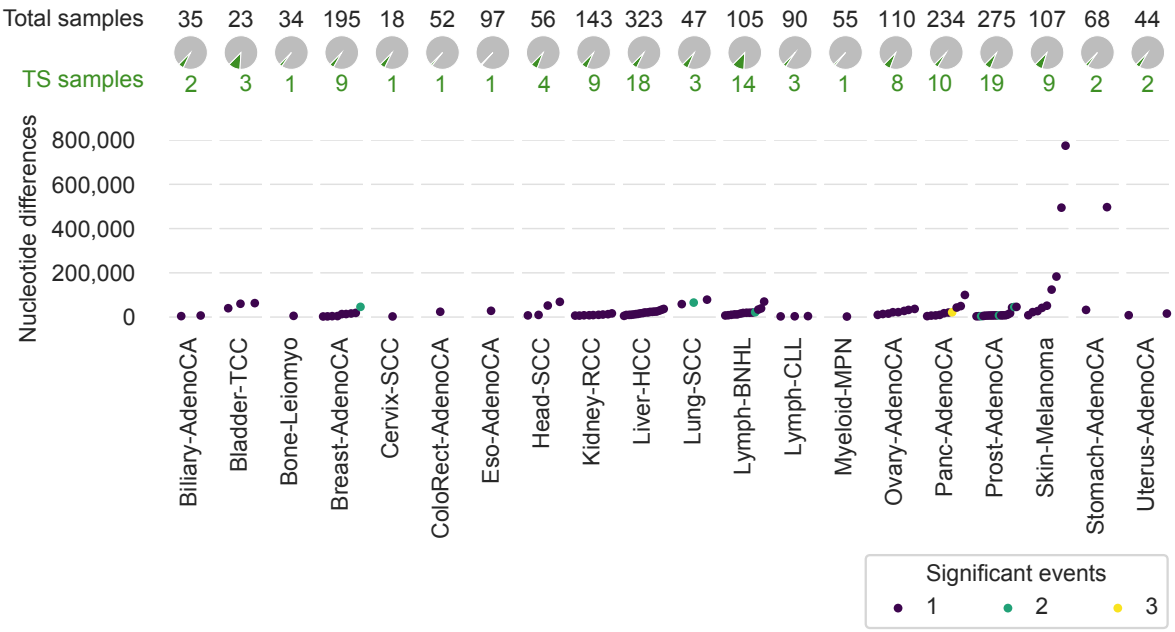

Figure 5.9: Significant short template switch mutation are found in many samples, and event count is not correlated with tumour divergence. Template switch counts are compared to tumour divergence across the PCAWG dataset. For each histological group (x-axis) in which at least at least one sample contained a significant template switch, the y-axis shows for each sample the number of nucleotide differences between the tumour and the matched normal tissue; points are then coloured according to the number of significant events that were identified in the sample (indicated by the key). Pie charts above each histology indicate the total number of samples with at least one event (green text and chart segment), samples with zero events (grey segments), and the total number of samples for that histology in the PCAWG dataset (black text).

Note that pairHMM alignment output and the VCFs associated with identified events are not provided as in the previous chapters due to PCAWG data access restrictions.

It is interesting to observe that 3 of the 118 unique events are identically observed across multiple samples, as these events must have arisen independently in each sample, rather than arising in some common ancestor (as with the shared events observed across my previous analyses). Each of these events is associated with an identical VCF footprint across each sample, and therefore the inferred template switch alignments are also identical. It is possible that some set of local sequence features makes these regions particularly prone to event formation, similar to mechanisms underlying structural variant formation, which cause independently arising, identical rearrangements in unrelated humans (e.g. the NAHR pathway, see §1.2.4 and [49]). It is however also possible that these events represent sequencing or variant calling artefacts, as *a priori* it is not expected that identical variants arise independently through replication-based

rearrangement mechanisms. In future, these events require careful consideration through local reassembly and/or experimentally validation.

I therefore use the 118 unique events for subsequent analyses, selecting one sample's alignment output and associated human reference coordinates as representative of the entire set of identical events. It may be argued that some analyses should include repeated representations of these identical events. For example, when performing enrichment analyses, treating an event found in nine samples as a single event may underestimate the prevalence of template switches within the functional regions of the genome that those nine identical events are contained within. As in previous chapters however, I choose the conservative path for my analyses, accepting some possible loss of statistical power.

Cancer genomes with at least one template switch contain more events relative to their divergence than observed in my previous germline analyses. Unsurprisingly however, template switching comprises only a small part of the mutational landscape in cancer, and no histological group is defined primarily by this form of mutagenesis (Figure 5.9). Most samples in the PCAWG dataset are characterised by no significant template switches, 20 of the 37 total considered histological groups contain a sample with a template switch, and the count of template switches observed across samples is not correlated with sample divergence (Figure 5.9). Indeed, samples from which template switches were identified largely only contained a single event, with only six samples containing two events, and one sample containing three events (Figure 5.9). This indicates that template switches do not arise simply as a function of tumour divergence, which may be informative for disqualifying some mutational pathways as being involved in template switch mutagenesis. For example, melanoma (Skin-Melanoma) contains the most diverged samples in the PCAWG dataset (Figure 5.1) due to accumulation of UV-induced photolesions throughout life [94, 124], and yet fewer than 10% (9 out of 107) of samples contain a significant event (Figure 5.9). These lesions are typically repaired by the nucleotide excision repair pathway [83, 195], and if template switching was involved in their repair in cancer, I would expect more samples to contain at least one template switch.

Some histological groups display relatively elevated proportions of samples containing template switches across both the tested parameter values (Figure 5.6) and under the final cancer-wide/sample-specific parameter values (Figure 5.9). In particular, more than 10% of lymphoid B-cell non-Hodgkin lymphoma (Lymph-BNHL) and bladder transitional cell carcinoma (Bladder-TCC) samples contain at least one template switch in the final event set (Figure 5.9). Additionally, samples which contain more than one event are only found in a few histologies, including Lymph-BNHL, lung squamous cell carcinoma (Lung-SCC), and three adenocarcinomas (Breast-AdenoCA, Panc-AdenoCA, and Prost-AdenoCA) (Figure 5.9). On

average, none of the samples from these tumour types are relatively divergent (Figure 5.1), they do not have an elevated indel to SNV ratio (Figure 5.3), and only breast adenocarcinoma stands out in terms of mean deletion length (Figure 5.2).

As divergence and patterns of short indel formation alone do not offer insight into the relative excess of template switches observed in these six tumour types, it is worth considering the patterns of structural variant formation observed in these samples as characterised by the PCAWG Structural Variation Working Group [180]. Many samples from the Bladder-TCC, Breast-AdenoCA, and Panc-AdenoCA tumour types are generally enriched for structural variants, the majority of which are tandem duplications and deletions in sample-specific varying proportions (see Fig. 2 and Extended Data Fig. 1 in [180]). Most of these structural variants are thought to arise through non-homologous end joining, as there is predominantly no microhomology at the associated breakpoints, which may indicate this pathway is involved in creating signatures of template switching (I return to the subject of template switch microhomology in §5.5.3). The other three tumour types (Lymph-BNHL, Lung-SCC, and Prost-AdenoCA) all contain only a single sample with a notable enrichment of structural variants which again are largely defined by tandem duplications and deletions (see Extended Data Fig. 1 in [180]). This may indicate that in these tumour types, short template switch mutations arise through alternate pathways to those currently characterised in cancer structural variation studies.

5.5 Features associated with short template switch mutations in human cancer

5.5.1 Significant events in cancer are characterised by a subset of possible event types and shorter ②→③ regions than in the germline

Recall that the linear ordering of the switch points associated with each event defines an "event type" (see §2.1.2). Only three event types are observed across cancer samples (Figure 5.10) — no significant ③-①-④-② or ① < ④ events are present. This may indicate either that my model comparison procedure applied to the PCAWG VCFs lacks power to detect such events, or that some causative mutational pathways are less active in cancer genomes. As in previous analyses (§3.3.3 and §4.7.2), the equivalent event types {①-④-③-②, ③-②-①-④} are the most prevalent. These event types are enriched in cancer compared to the germline analyses however, representing 81.2% of events here, compared to 49.0% and 61.5% in my hominid and human population analyses, respectively. This may be due to an over-representation of these

event types amongst events with shorter ②→③ regions (see marginal densities in Figure 3.8 and Figure 4.13). Indeed, template switches identified in human cancer genomes are defined by markedly shorter ②→③ lengths than in previous analyses, with a median length of 5 (median absolute deviation = 1.5) and a max length of 18 (Figure 5.10).

As the median divergence between a cancer genome and its matched normal tissue is $t = 2.2 \times 10^{-6}$ (substitutions/indels per site) and I only permit a single nucleotide mismatch in the flanking regions of the ②→③ fragment in significant TSA pairHMM alignments, background mutations close to the focal cluster of each event are either not present or contribute at most one additional SNV. It is therefore possible that the observed decrease in median ②→③ length is due to a genuine ability of my model to detect shorter events in such a low divergence setting without obfuscation of the focal mutation cluster or ②→③ template (also noted in §4.7.2 for human population events). However, surrounding mutations in human populations should also be near non-existent when considering that the divergence defined by SNVs and indels between any two human samples is around 0.1% [293] and I filtered all events on per-base alignment quality (Figure 4.2). I would therefore expect a consistent greater proportion of 4nt ②→③ regions in both event sets, as neither setting should suffer from complications introduced by background mutations close to the mutation cluster (which impacted my great ape analysis; see §3.2.1). The excess of events characterised by the minimum length permitted 4nt ②→③ region in cancer genomes may therefore include false positives, possibly reflecting the need for an alternate method for establishing a LPR threshold in such a low divergence setting. I opt to retain all events regardless, and I also refer the reader back to my discussion on setting parameter σ in the TSA pairHMM, as using a less natural formulation of this parameter can produce longer ②→③ state paths.

5.5.2 Microhomology lengths typical of FoSTeS/MMBIR are not observed for short-range template switches

As in studies of genomic rearrangements in human populations, mutational mechanisms in human cancer are assigned based on patterns of microhomology around the associated breakpoints [180, 228, 307, 316]. In the cancer genomes characterised by PCAWG, many structural variants are observed with no microhomology at the identified breakpoints, and are assigned to the non-homologous end joining pathway [180]. Of the rearrangements which are associated with breakpoint microhomology, lengths of 2–7nt and >10nt are respectively used to assign the observed rearrangement to either the microhomology-mediated end joining

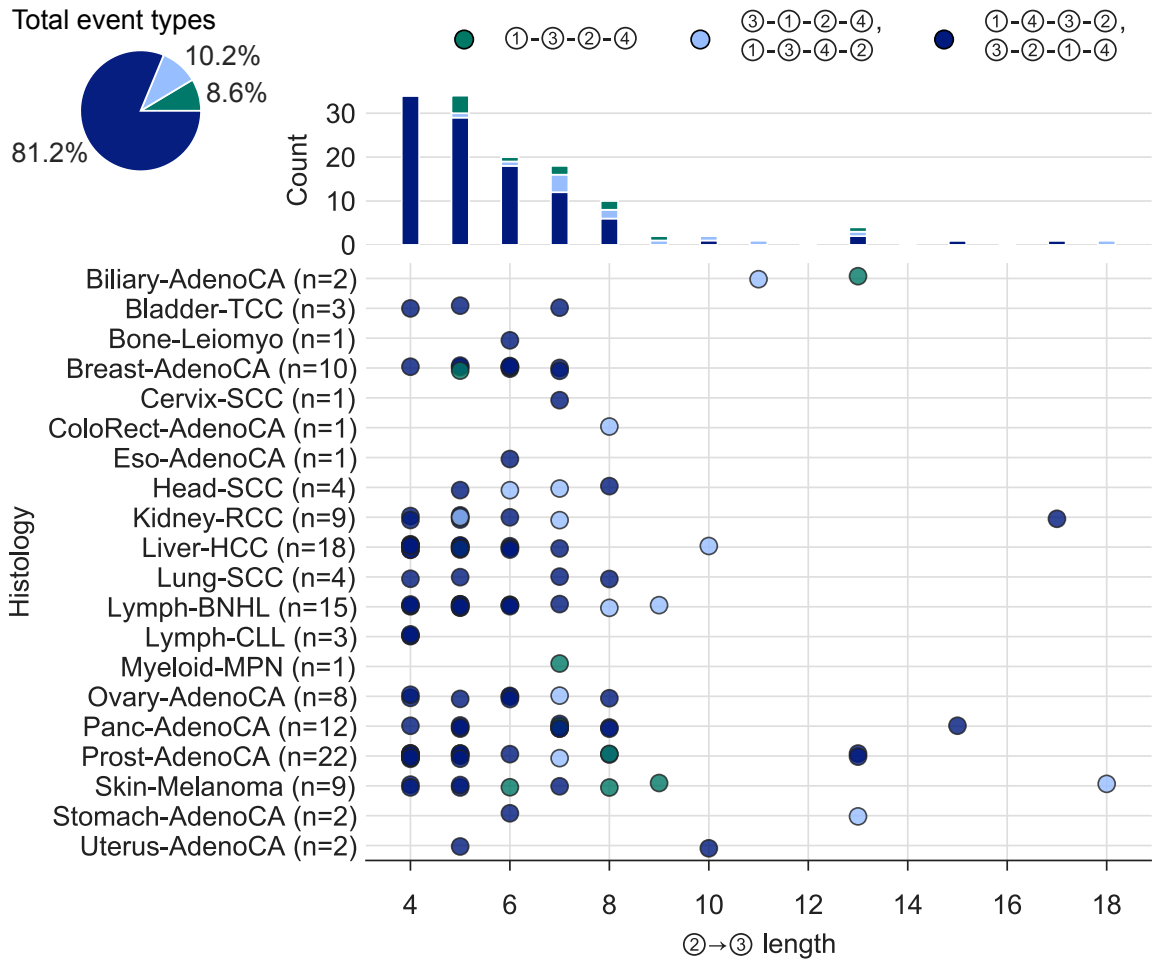

Figure 5.10: Events in human cancer genomes are relatively short and represented by a subset of possible event types. The central strip plot shows the ②→③ lengths of all 128 significant events on the *x*-axis, broken down by histological group along the *y*-axis (the number of events per histology are also indicated in brackets). Points are coloured according to their event type (indicated in the key, top), and the marginal stacked bar chart shows the total count of event types per ②→③ length. The inset pie chart in the top left aggregates the count of event types across all events.

pathway, or pathways including single-stranded annealing and the replication-based FoSTeS and MMBIR mechanisms (see §1.2.4) [180].

I assessed microhomology surrounding the initial (① to ②) and return (③ to ④) switch events in an identical manner to my analysis of human population events in §4.7.2. Of the 118 unique, significant template switches in cancer, the majority of observed events are not characterised by microhomology at the switch points (Figure 5.11). For ① to ②: 62.0% of events have microhomology length 0, 34.0% have length 1–5, and 4.0% have length >5

(Figure 5.11; orange histogram). For ③ to ④, these values are respectively 65.2%, 30.4%, and 4.4% (Figure 5.11; blue histogram). Further, no obvious correlation between microhomology length and ② → ③ length was observed, and no individual event type was noticeably enriched for either the shorter or longer microhomology lengths (not shown). Overall, this indicates that although a subset of events may possibly be mediated by the FoSTeS and MMBIR pathways as is the case for short template switch events and indel formation in the human germline (see §4.7.2 and [218]), the majority of events may either be: (a) mediated by replication-based mutational pathways to which I am unable to assign events to based solely on patterns of microhomology; or (b) mediated by alternative DNA repair pathways such as non-homologous end joining.

5.5.3 Events in human cancer may be modulated by poly(dA:dT) tracts

As I am unable to assign causative pathway to events based on patterns of homology, I next asked if any specific motifs may mediate event initiation in cancer. As in my great ape analysis

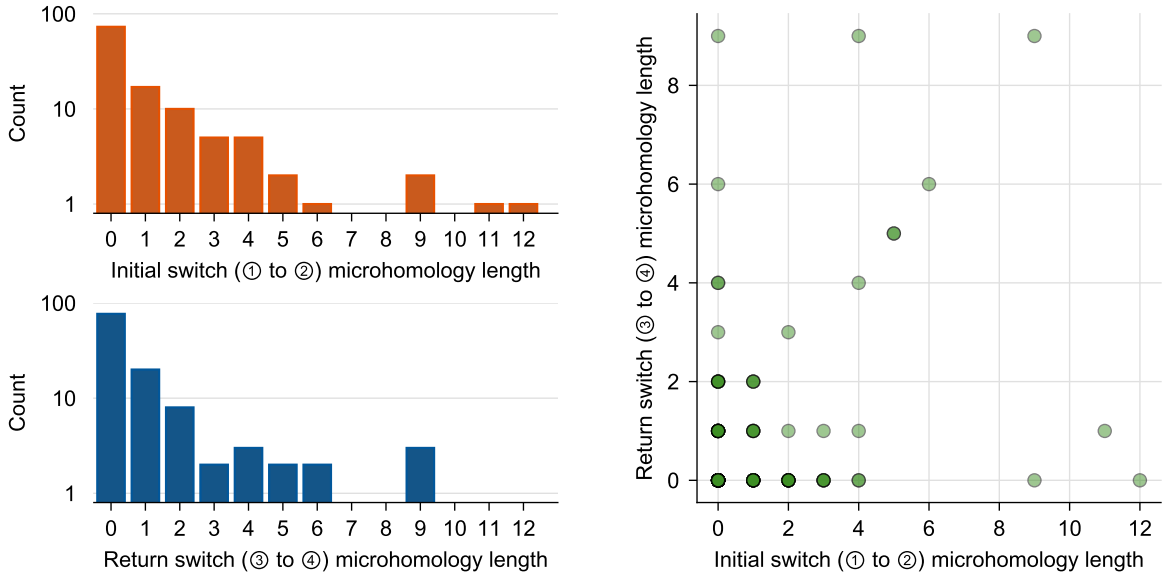

Figure 5.11: Microhomology length requirements typical of the FoSTeS/MMBIR pathways are not associated with event formation in human cancer. Histograms on the left show the distributions of microhomology length for the initial (① to ②; orange) and return (③ to ④; blue) template switch events; note the log scale on the *y*-axes of the histograms. The scatter plot on the right shows the microhomology length at the initial template switch positions compared to the return switch positions; alpha corresponds to point density. See Figure 4.15a for a depiction of how microhomology is defined for each event, and Figure 4.15b–d for similar descriptions of 1k-30x human population events.

(§3.4.3), I first used `meme` [23] to ask if sequences ±150nt surrounding switch point ① (301nt total) were enriched for a single motif per sequence. Analysis was performed using the same procedure outlined in §3.4.3, except I now sample a background set of 10,000 301nt sequences solely from the GRCh37 genome to assess motif enrichment in sequences surrounding template switches. I identified one significant motif per motif size tested (6–10nt, 10–20nt, 20–50nt): ATGGTATY ($E = 1.5 \times 10^{-2}$; 16/118 sequences), TYCCAGCACT ($E = 4.6 \times 10^{-2}$; 7/118 sequences), and RSWGRWGGRMHHANRGDRRDGAGCAAASRHRVVHG ($E = 9.9 \times 10^{-3}$; 31/118 sequences)[1]. However these motifs are either only present in a small subset of sequences evaluated, or are so ambiguous in the case of the longest motif identified that their interpretation is difficult. Further, it is possible that `meme` may lack statistical power when assessing differentially enriched motifs at the small sample sizes used here. I therefore inspected the most common motifs identified which were not significantly enriched. Of note, all 118 assessed sequences contained the motif AAAWAA (W $\in \{A,T\}$), which suggests that poly(dA:dT) tract DNA is involved in event initiation as in the events assessed in hominid genome (see §3.4.3).

5.5.4 Variants associated with template switches cannot be plausibly explained by established cancer mutational signatures

In cancer, a broad set of mutational signatures is routinely assessed to identify mutational mechanisms driving carcinogenesis in specific histological groups. Three classes of signature are typically studied: (a) single-base substitutions within specific trinucleotide contexts; (b) adjacent double-base substitutions independently of the flanking context; and (c) small indels, in which the length and microhomology of the indel, as well as the length of homopolymer or the number of repeat units in which the indel occurred are considered. A robust collection of these signatures is maintained by the Catalogue Of Somatic Mutations In Cancer (COSMIC) resource (https://cancer.sanger.ac.uk; [291]). Each signature has been established through association with a specific mutagen or mutational pathway. For example, excessive TTT→TAT SNV mutations are associated with unrepaired UV lesions, CG→TA dinucleotide substitutions are associated with defective mismatch repair, and single-nucleotide deletions within homopolymers of length 6 or greater are associated with replication slippage. Many mutational processes involve more complex signatures than one type of mutation however, consisting of a set of specific mutations each with a characteristic proportion for that process (see [9]).

[1]Note the use of the following IUPAC ambiguity codes: Y $\in \{C,T\}$; R $\in \{A,G\}$; S $\in \{G,C\}$; W $\in \{A,T\}$; M $\in \{C,A\}$; H $\in \{A,C,T\}$; D $\in \{A,G,T\}$; V $\in \{A,C,G\}$; N $\in \{A,C,G,T\}$.

In my human population analysis, I considered plausible alternatives for the generation of mutation clusters, and demonstrated that signatures of error-prone Pol-ζ, replication slippage, and APOBEC activity did not underlie the template switches identified (§4.7.1). Similarly, I asked if the variants I am attributing to template switching in cancer can instead be convincingly explained by a known mutational signature, which would indicate possible false positives in my final set of 118 significant, unique events.

To this end, I extracted mutational signatures for all SNVs and indels associated with one of the 118 unique template switches identified by my event discovery pipeline. Given the small numbers of events identified, I look for evidence of a specific mutational pathway that may underlie template switching independently of histology by considering events from all tumour types together. I called signatures for SNVs and indels using the `fit_to_signatures()` function of the R/Bioconductor package MutationalPatterns (v3.2.0) [35], which utilises a non-negative least-squares algorithm to find the set of COSMIC mutational signatures (v3.2) which best matches the input set of variants. The only notable signatures identified from this analysis were SNV signature 9 (SBS9), which was associated with 19% of observed SNVs, and indel signature 8 (ID8), associated with 67% of observed small indels. SBS9 has been attributed to the activity of the translesion polymerase Pol-η [9], it is associated with almost every possible SNV/trinucleotide context however, and its proposed aetiology is supported by unclear evidence (see SBS9 in the COSMIC web portal: https://cancer.sanger.ac.uk/signatures/sbs/sbs9; [291]) so that it is unconvincing at explaining the SNVs associated with events here. Similarly, ID8 has unclear evidence supporting non-homologous end joining activity [11] (see ID8 in the COSMIC web portal: https://cancer.sanger.ac.uk/signatures/id/id8; [291]), but has no known aetiology for most tumours in which it is observed. Under the paradigm of mutational signature analysis, there is therefore weak support for a mechanism beyond template switching generating the observed mutation clusters. However, the association with ID8 combined with no microhomology at the switch points of PCAWG template switches (Figure 5.11) suggests that some of the small indels that I attribute to short template switches may indeed have arisen through non-homologous end joining.

5.5.5 Somatic short-range template switches are not significantly depleted in coding regions

To assess genomic features associations associated with the 118 unique template switches, I performed the same enrichment analysis as in the two previous chapters, but instead use GRCh37-indexed coordinates of the identical set of features rather than the GRCh38 coordinates

used previously (otherwise processed as described in Table 3.3). No significant enrichment or depletion was was observed for any tested genomic feature (calculated as previously using a threshold of 0.01 on Bonferroni-corrected empirical p-values). This may either highlight the lack of statistical power available when using small sample sizes in permutation tests, or it could suggest that selection does not act to remove template switches from any regions of cancer genomes (unlike in germline evolution, in which template switches are significantly depleted in protein-coding regions; see §3.4.1 and §4.7.3).

5.5.6 Somatic events occur more frequently in early replicating regions, and are not mediated by non-canonical DNA structures

As with germline structural variants, rearrangements in cancer are often associated with distinct replication times. Large deletions and tandem duplications are the most common form of structural variation in cancer, and are enriched in late and early replicating regions across histological groups [180], respectively. Replication timing profiles in some tumours also deviate from this broad spectrum, and the majority of all observed structural variants sometimes occur either in almost-exclusively early or late replicating regions [180].

To assess replication timing association in cancer, as well as distance to replication origins as previously assessed (§4.7.3), I used the replication timing dataset and enrichment analysis procedure described for my human population analysis (see §4.7.3), comparing values associated with template switches to those of a randomly sampled background of 230,000 loci. (Here however I do not use liftOver to convert the replication timing and replication origin coordinates from GRCh37 to GRCh38, as PCAWG variant calls were also generated with respect to GRCh37.) As with events identified in the human population 1k-30x dataset (Figure 4.14), template switches in cancer show a moderate shift towards early replicating regions (Mann-Whitney U test, $p = 0.051$; Figure 5.12a), and a similar moderate but now significant ($p < 0.05$) association with proximity to replication origins ($p = 0.012$; Figure 5.12b). In combination with the lack of microhomology at switch points, this further suggests that the replication-based FoSTeS and MMBIR rearrangement pathways are not responsible for template switching in cancer genomes.

As a final line of inquiry into the modulators of template switch initiation, I consider known impediments of replication fork progression, reasoning as previously that fork arrest may lead to template switching as with large-scale mechanisms [49, 172]. As discussed in §3.4.2, transiently single-stranded DNA can adopt stable secondary structures during replication and cause fork stalling [215]. Consistent with this in my hominid analysis, I observed more stable predicted

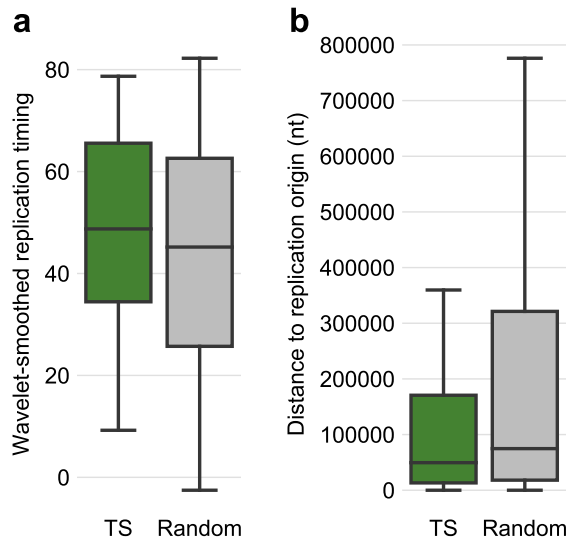

Figure 5.12: In human cancer, short template switches are observed moderately more frequently in early replicating regions and are significantly proximal to replication origins. (**a**) The distribution of wavelet-smoothed replication timing signal for significant template switch events across all cancer histological groups, compared to randomly sampled GRCh37 loci. Greater values on the *y*-axis are associated with earlier replication timing [115]. There is no significant different between groups for replication timing (Mann-Whitney U test, $p = 0.051$). (**b**) The distribution of distances to the nearest replication origin for template switch loci, compared to a randomly sampled genomic background. There is a significant different between groups for distance from replication origins (Mann-Whitney U test, $p = 0.012$). As in previous figures, boxes show the median, Q1, and Q3; whiskers show Q3/Q1 $\pm$ 1.5$\times$IQR, and outliers are hidden for clarity.

secondary structures surrounding ① in a subset of pre-event, ancestral sequences (Figure 3.11a). Further, in cancer, non-B DNA configurations including hairpins and more complex structures such as Z-DNA and guanine quadruplexes have been associated with increased localised mutation rates [96, 327]. This interpretation could erroneously be made for apparent mutation cluster footprints left by template switching if coinciding with regions capable of adopting non-B DNA conformations. I also consider patterns of nucleosome occupancy surrounding the initial switch event. Nucleosomes are displaced immediately ahead of the proceeding replisome and are rapidly recycled behind the proceeding replisome following successful synthesis of the nascent strand [6]. If nucleosomes are not successfully unbound from the DNA helix ahead of the proceeding replisome, they may act as strong barriers to fork progression [54].

I assessed DNA secondary structure formation potential for the normal tissue sequences $\pm$500nt surrounding template switch location ① for all 118 unique events using the `RNAfold` tool in the ViennaRNA package [288] as described in §3.4.2. As previously, I sampled 10,000

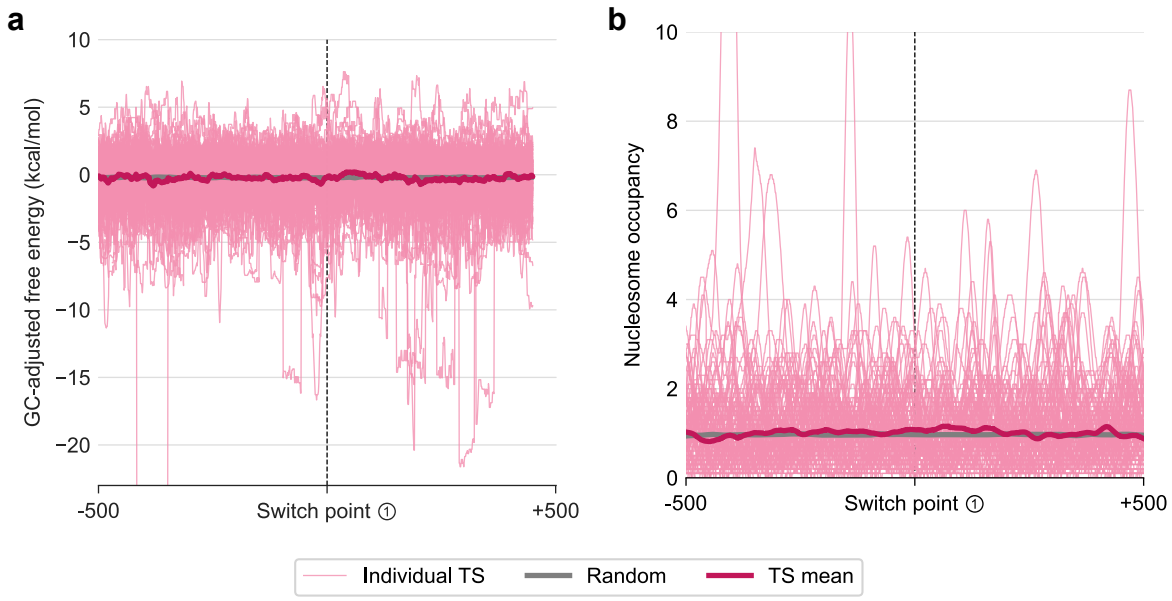

Figure 5.13: Non-canonical DNA secondary structures and stable nucleosome occupancy do not cause short template switch initiation in human cancer. (**a**) GC content-adjusted free energies of the MFE secondary structures ±500nt around switch point ① for the normal tissue (GRCh37) sequence compared to a random genomic background; calculated using a left-aligned 50nt sliding window with a step size of 1nt (as in Figure 3.11). The dark pink line ("TS mean, i.e. template switch mean") indicates the mean MFE of each window across all evaluated cancer sequences; the grey line ("Random") shows the mean MFE of each window across all evaluated randomly sampled sequences; the translucent pink lines ("Individual TS") show the MFE of each window for the individually evaluated cancer sequences. (**b**) Single nucleotide-resolution nucleosome occupancy for each of the sequences assessed in (a), determined using MNase-seq of human cell line K562 [299]. Mean, background, and individual sequence values are shown as in (a).

random background sequences of length 1001nt for comparison, now using GRCh37 to be consistent with the PCAWG reference genome. To assess nucleosome occupancy, I used single nucleotide-resolution micrococcal nuclease sequencing (MNase-seq) data for human cell line K562, retrieved in bigWig format [146] from ENCODE (accession no. ENCFF000VNN; see also [299]) and converted to BED format using the `wig2bed` tool from the BEDOPS (v2.4.36) toolkit [222]. For each position in the 118 ①±500nt sequences, I retrieved the MNase-seq measurements using `bedtools closest`. From this procedure, I did not observe any noticeable pattern of stable secondary structure formation or consistent nucleosome positioning in relation to switch point ① across any of the 118 unique events evaluated (Figure 5.13). It is still of note however that template switches by definition generate inverted repeats (refer back to Figure 2.2a) which can form stable secondary structures. If inverted repeats/hairpin DNA structures do

indeed mediate localised high rates of mutation [96, 327], the occurrence of a template switch early in the evolutionary history of a cancer genome could drive this mutagenesis through the creation of novel inverted repeat sequences.

5.5.7 Exploring individual template switches in cancer-associated genes and regulatory regions

Mutations within the coding regions of cancer genomes are often involved in driving tumourigenesis, and non-coding driver mutations have been identified in genomic regions involved in gene expression, such as in the promoters and $5'/3'$ untranslated regions of oncogenes [98, 248]. In my previous analyses, template switch mutations were significantly depleted in coding regions, and only weakly enriched in some functional regions including transcription factor binding sites (see §3.4.1 and §4.7.3). Evolutionary dynamics in human cancer genomes differ to models of germline mutation however, and it is possible that template switches may be tolerated within functional genomic regions (as is the case for SNVs and indels [206]), and potentially even drive tumourigenesis. As with my previous association analyses, I do not aim to make conclusive assertions about potential causal relationship between template switches and the features I am assessing in a purely data-driven manner, but it is nevertheless worth assessing any associated signals which may be of relevance to cancer biology.

To identify variants of interest amongst the 118 significant, unique template switches, I annotated all VCF records associated with these events using the Ensembl Variant Effect Predictor (release 104) [203] (`ensemblorg/ensembl-vep` in Docker [207]). Note that this analysis is subject to the same caution outlined in §4.7.3, in that the annotations generated for specific variant coordinates associated with template switches may differ if a direct encoding of the template switch event in the VCF was used. All variants I describe below however are associated with several hits within the same local functional region.

Focusing on coding variation, three template switches introduce missense variants into exons. Each of these variants is provided with an associated "Combined Annotation Dependent Depletion" (CADD) score, which is a metric used to assess the potential deleteriousness of coding human genetic variants [151]. One Stomach-AdenoCA sample presents several possible missense variants introduced by a 13nt ②→③ template switch in PIP4K2B, all of which are annotated as benign by CADD. Interestingly however, amplified expression of PIP4K2β kinases produced by PIP4K2B have been observed in human breast tumours [82], and PIP4K2β has been proposed as a drug target for TP53-defective cancers [82]. The other two template switches introduce missense variants which are assigned a CADD score of potentially damaging.

One potential HSPA1L missense variant is observed for a 5nt ②→③ template switch in a endometrial adenocarcinoma (Uterus-AdenoCA) sample, and multiple DRP2 missense variants are associated with a 4nt ②→③ template switch observed in a breast adenocarcinoma (Breast-AdenoCA) sample. Strong associations between cancer progression and either HSPA1L or DRP2 mutations have not been described in the literature. However inducing HSPA1L expression *in vivo* has been shown to promote cellular prion protein accumulation and drive tumour progression in colorectal cancers [171].

I next ask if any template switches may introduce driver mutations into cancer cells. Driver mutations are those that provide a selective advantage to individual cells within somatic tissues by perturbing homeostatic cellular functions, and their occurrence early in the evolutionary history of a cell causes the uncontrolled growth that leads to cancer [22]. I retrieved the curated set of cancer driver mutations provided by the Integrative OncoGenomics (IntOGen) web portal (https://www.intogen.org; [197]), identified using seven driver identification methods applied to 28,076 matched normal tissue and tumour samples across 66 histologies from 221 cohorts (including PCAWG). I then checked for genes common between my annotated VCF and the IntOGen set. I identified three driver genes (BCL2, MAML2, and MSI2) associated with template switches all of which contain a VCF footprint of 3 SNVs.

The 5′ untranslated region of BCL2 (B-cell lymphoma 2 apoptosis regulator) contains a 5nt ②→③ event in one lymphoid B-cell non-Hodgkin lymphoma (Lymph-BNHL) sample. BCL2 is a mitochondrial membrane protein that prevents apoptosis of lymphocytes, and mutation is commonly associated with of B-cell non-Hodgkin lymphomas, but this is typically in the context of translocations of the entire gene or multigenic rearrangements [78, 257]. MAML2 (mastermind-like transcriptional coactivator 2) contains an intronic 13nt ②→③ event in one biliary adenocarcinoma (Bilinary-AdenoCA) sample. MAML2 is a transcriptional coactivator of proteins involved in the Notch signalling pathway, and has been linked to tumourigenesis through activation of the Hippo signalling pathway across several cancer types as a result of gene fusions with YAP1 (Yes1-associated transcriptional regulator) [234]. MSI2 (Musashi RNA binding protein 2) contains an intronic 7nt ②→③ event in one pancreatic adenocarcinoma (Panc-AdenoCA) sample. MSI2 encodes an RNA binding protein that regulates transcription of genes involved in development and cell cycle regulation, and its overexpression is associated with poor clinical outcome in pancreatic cancer [113], possibly through down-regulation of the endocytic adaptor protein NUMB [272]. Without further experimental investigation, I am not able to establish a causal link between the template switches associated with these genes and the resulting carcinogenesis, but it is nevertheless interesting to observe these variants within or proximal to known cancer driver genes.

5.6 Conclusions

In this chapter I have described the patterns of short template switch mutagenesis across tumour types sequenced as part of the PCAWG study [44], identifying a set of 118 unique, significant events across 120 samples (Figure 5.9). I have shown that short template switch mutations comprise a small part of the human cancer mutation landscape, that rates of template switching do not correlate with tumour divergence (i.e. template switching is not causative nor accelerated by cancer), and that many tumour types do not appear to be driven by template switch mutations. Regardless, I have demonstrated that template switching is not only a human germline mutational phenomenon, but that it also occurs as part of the human somatic mutational landscape. This highlights the importance of identifying and studying small-scale rearrangements that are currently neglected in studies of human disease and cancer due to the operational definitions of structural variant.

Applying my methods in a human cancer setting presented some difficulties, namely that both between-histology and within-histology rates of divergence, indel formation, and indel length distributions vary greatly. Accounting for this variation when selecting parameters for my pairHMMs is vital, as misspecifying parameters in my models can cause high rates of type I and type II errors. While it is relatively trivial to fit these parameters to each sample, events which are then only identified under these sample-specific parameters may be unconvincing. Further, the simulation procedure used to estimate my LPR test statistic under the null hypothesis (and assess test power) was complicated by the low levels of divergence, as few mutation clusters and small indels from which this distribution is calculated arise by chance in this setting. In future, it may therefore be worth exploring alternate test statistics when performing model selection between my pairHMMs in a cancer setting, and it would be interesting to investigate a procedure for iterative parameter refinement.

Although I impose a stringent threshold on my LPR test statistic, the set of somatic template switches identified in human cancer genomes is characterised by notably shorter ②→③ regions than their germline counterparts. While this may represent a true ability of my models to detect very short events at low levels of divergence, it may also reflect that mutational signatures created by alternate mutational pathways such as non-homologous end joining are falsely inferred as template switch mutations here. Without direct experimental observation however, there is no way to distinguish between these two possibilities. Note also that it is standard when studying structural variation in human cancer genomes [180] to assign mutational mechanisms to observed variants solely by computationally inspecting patterns of microhomology at the breakpoints of the called rearrangements. It would therefore be prudent

in future work to experimentally investigate small-scale rearrangements so that the underlying mutational pathways, template switching or otherwise, may be more accurately assigned.

The existing PCAWG dataset is imperfect for event identification due to short read length and the lack of consistent *de novo* reassembly of reads at small indels and clustered SNVs. Regardless, I have identified several events of interest, including three events which impact exons, and three events which fall within or proximal to known cancer driver genes. I have also identified specific tumour types in which events appear to be over-represented relative to their divergence, indel frequency, and structural variant frequency, including lymphoid B-cell non-Hodgkin lymphoma, lung squamous cell carcinoma, and prostate adenocarcinoma. These tumour types should therefore be carefully considered in any future analysis of template switch mutagensis in a cancer setting.

Ultimately, the somatic landscape of template switching requires additional consideration in future. Read lengths were relatively short in the PCAWG study, and variant calling methodology has already improved since the production of the final PCAWG calls. It would therefore be beneficial to call variants from $\geqslant$ 150nt reads using state-of-the-art methods that perform local reassembly at all clustered SNVs and indels, such as Mutect2 [29]. Additionally, as noted in my reflections on capturing variation in a germline setting (§4.9), it may be beneficial to instead assess template switching using variant calls generated through *de novo* assembly of long reads, possibly represented in a graph-based data structure. At the time of writing, the PCAWG study has provided the most comprehensive set of high-quality, genome-wide variant calls for assessing mutagenesis across a broad range of tumour types. In future, large-cohort sequencing of cancer genomes (for example, genomes sequenced as part of the 100,000 Genomes Project [52]) assembled and variant called using gold-standard methods will permit a better assessment of template switch mutagenesis in human cancer.

Chapter 6

Concluding remarks

In this thesis, I have provided statistical methods for identifying short template switch mutations in DNA sequence data, and applied these methods to explore the involvement of template switching in human genome evolution, human population variation, and human cancer. To my knowledge, this represents the first effort to statistically and systematically explore small-scale replication-based rearrangements within these datasets.

Capturing and statistically assessing short template switch mutations using a pairHMM comparison procedure (Chapter 2) has proven to be a suitable approach across all datasets assessed in this thesis, as it ensures that the single most probable template switch is always identified within a given local sequence region. I have shown throughout that my model can be parameterised across a range of mutational settings, and demonstrated that a simulation-based approach for approximating the null hypothesis distribution of my LPR test statistic is generally appropriate (albeit computationally expensive in low divergence settings such as cancer). This thesis did not seek to statistically capture non-local short template switches; I have discussed why this may not be feasible through sequence analysis alone. It would be fascinating to instead to experimentally capture these events alongside local template switches, as there is no reason to assume that replication-based rearrangement pathways which likely underlie short-range template switch mutagenesis only operate in a local sequence context. Indeed, in structural variant formation rearrangements mediated by these pathways often occur at extreme distances in linear sequence space. Pre-pandemic aims for this thesis included pursuing experimental observations of events, and some initial work (by others) demonstrated that template switch mutagenesis can indeed be invoked *in vivo* in yeast.

In an evolutionary setting (Chapter 3), I have demonstrated that mutation clusters generated through local template switching occur routinely as part of typical genome evolution, are associated with specific genomic features, and can be identified and phylogenetically resolved through pairwise whole genome comparisons. This is an important result from an evolutionary perspective, as counting apparent clustered mutations as single events rather than as multiple

independent events can prevent their potential confounding impact on tests for positive selection or accelerated evolution. In this thesis, I concerned myself with the evolution of the human genome by considering just six pairwise genome comparisons within the hominid tree. It would certainly be worthwhile in future to test for template switch mutagenesis in a larger set of species to better understand the impact of these mutations on shaping eukaryotic and prokaryotic genome evolution more broadly. This may require some methodological improvements. The juxtaposition between related DNA sequences facilitated by pairwise alignment is ideal for identifying local rearrangements which have occurred between two evolutionarily related DNA sequences. When considering many pairwise genome alignments simultaneously however, the set of pairwise comparisons required to facilitate a phylogenetic interpretation currently scales as $O(n^2C)$ for n genomes harbouring C mutation clusters. Optimising the tests for template switching that I have developed here instead for direct consideration of multiple rather than pairwise sequence alignments would therefore be an interesting future expansion of the present methodology.

I have further shown that short template switches are a ubiquitous feature of ongoing human genome evolution at the population level (Chapter 4), distributed amongst human populations approximately as expected under a model of neutral evolution. This work challenges the operational definition that human structural variants consist of rearrangements that impact $\geqslant$ 50nt. On first consideration this may seem like an unimportant distinction, whereby I am simply arguing for "structural" to prefix a subset of small-scale genetic variants. As I have repeatedly demonstrated in this thesis, however, unless structural variant calling methods are utilised, the only alternative way such variants are represented (through read mapping and/or local reassembly) is as a cluster of SNVs and/or indels. As a result, the involvement of replication-based rearrangement pathways will go underestimated. This could have consequences for understanding the role of these pathways in human disease and any associated phenotypic consequences. The pipeline developed here was able to resolve these cases and could readily be incorporated as a post-processing step in existing human variant calling pipelines. Alternatively, TSA pairHMM alignments could be directly incorporated into the local realignment procedures already utilised by many short-read based variant calling pipelines at indels and clustered SNVs.

The PCAWG study [44] has provided the first publicly available (subject to data access approval), large-scale dataset of sequenced tumours alongside the matched normal tissue. I have leveraged this dataset (Chapter 5) to assess how well my statistical methods can detect template switches with confidence in cancer variation data. and I was able to identify a small set of significant template switches in human cancer genomes. Despite potential suitability issues with

the procedure that was used to produce the final set of variants in these data (for identifying short template switches), I have been able to show that some human cancer genomes may contain more template switches relative to their divergence than when considering human germline evolution. The same considerations apply to cancer variation datasets as with human population variation — underestimating the prevalence of short replication-based rearrangements may mean that an important disease association is missed. Several template switches detected were particularly interesting, such as those which appear to have arisen independently in multiple tumours and those which were within or proximal to known cancer driver genes. It would therefore be beneficial in future to call variants in cancer datasets using updated methodology (such as performing local reassembly at all mutation clusters and/or short indels), and it would be ideal to experimentally validate any apparently independently recurring events observed in multiple samples.

An important consideration for future work (that relates to all of my analyses) is that the variation datasets used in this thesis have not allowed me to assess the strandedness of events. That is, although I can infer whether an event may occur as an inter-strand switch or an inter-strand switch based on linear switch point ordering, I have been unable to infer the nascent strand from which events are initiated (the ① to ② switch), and the strand which donates the ② → ③ sequence region. This means that I cannot draw any conclusions about, for example, if events are initiated more frequently during lagging strand replication, where one might expect replication barriers such as secondary structure to form more readily due to the presence of single-stranded DNA. It would be particular interesting to resolve and compare the prevalence of lagging strand template switches with the mutational hotspots created during lagging strand replication by Pol-α [247].

The key takeaway from this thesis is that short template switch mutations are a ubiquitous feature of germline and somatic variation datasets, occurring alongside single nucleotide polymorphisms, insertions, deletions, and structural variants to define the human mutational landscape. This challenges the current methodological paradigm of treating small-scale mutations and structural variants as separate entities. DNA molecules change through time in complex ways. Single nucleotide changes and genomic rearrangements occur on a continuum, and classic etymology surrounding variation is not always adequate to fully describe the true complexity of observable genomic differences. Rearrangements can occur at small scales, introducing mutations that resemble traditional SNPs, while manifesting in an alignment solely as indels. This thesis has successfully described this complexity in human genomes, but the extent to which short template switch mutations have driven genetic diversity across other species in the tree of life remains to be seen. Future studies exploring small-scale mutational

complexity in other organisms may find that small rearrangements are a primary mechanism for driving evolution.

The methods I have devised here bring the routine study of short template switch mutations within the reach of many routine analyses, including population sequencing projects and comparative genomic studies (when divergence is not too high). These small-scale replication-based rearrangements have likely remained unappreciated due to the difficulty of encoding these variants in the linear representations of genomes typically assessed in genetic and evolutionary analyses. Moving forward, data structures which permit a natural encoding of small-scale rearrangements, such as graph genomes, will likely replace linear representations of large-scale genomic diversity. As non-linear genome encoding becomes standard, I believe that small-scale rearrangements will become routinely captured and studied. As a result, future studies into human genetic variation that incorporate such methods will be able to, by default, ask how non-linear, complex forms of variation contribute to evolution, population variation, and genomic disorders.

References

[1] Abadi, S., Azouri, D., Pupko, T., and Mayrose, I. (2019). Model selection may not be a mandatory step for phylogeny reconstruction. *Nature Communications*, 10:1–11. doi: 10.1038/s41467-019-08822-w.

[2] Abel, H. J., Larson, D. E., Regier, A. A., Chiang, C., Das, I., et al. (2020). Mapping and characterization of structural variation in 17,795 human genomes. *Nature*, 583:83–89. doi: 10.1038/s41586-020-2371-0.

[3] Abyzov, A., Li, S., Kim, D. R., Mohiyuddin, M., Stütz, A. M., et al. (2015). Analysis of deletion breakpoints from 1,092 humans reveals details of mutation mechanisms. *Nature Communications*, 6:1–11. doi: 10.1038/ncomms8256.

[4] Agarwal, I. and Przeworski, M. (2019). Signatures of replication timing, recombination, and sex in the spectrum of rare variants on the human X chromosome and autosomes. *Proceedings of the National Academy of Sciences of the United States of America*, 116: 17916–17924. doi: 10.1073/pnas.1900714116.

[5] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723. doi: 10.1109/TAC.1974.1100705.

[6] Alabert, C. and Groth, A. (2012). Chromatin replication and epigenome maintenance. *Nature Reviews Molecular Cell Biology*, 13:153–167. doi: 10.1038/nrm3288.

[7] Alberts, B., Johnson, A., Lewis, J., Morgan, D., Raff, M., et al. (2017). *Molecular Biology of the Cell*. Garland Science, New York, 6th edition. doi: 10.3390/ijms161226074.

[8] Alexandrov, L. B. and Stratton, M. R. (2014). Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. *Current Opinion in Genetics and Development*, 24:52–60. doi: 10.1016/j.gde.2013.11.014.

[9] Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A., Behjati, S., et al. (2013). Signatures of mutational processes in human cancer. *Nature*, 500:415–421. doi: 10.1038/nature12477.

[10] Alexandrov, L. B., Jones, P. H., Wedge, D. C., Sale, J. E., Campbell, P. J., et al. (2015). Clock-like mutational processes in human somatic cells. *Nature Genetics*, 47:1402–1407. doi: 10.1038/ng.3441.

[11] Alexandrov, L. B., Kim, J., Haradhvala, N. J., Huang, M. N., Ng, A. W. T., et al. (2020). The repertoire of mutational signatures in human cancer. *Nature*, 578:94–101. doi: 10.1038/s41586-020-1943-3.

[12] Almarri, M. A., Bergström, A., Prado-Martinez, J., Yang, F., Fu, B., et al. (7 2020). Population structure, stratification, and introgression of human structural variation. *Cell*, 182:189–199.e15. doi: 10.1016/j.cell.2020.05.024.

[13] Altmann, R. (1889). *Ueber nucleinsäuren*. Archiv für Anatomie und Physiologie, Leipzig.

[14] Altshuler, D., Daly, M. J., and Lander, E. S. (2008). Genetic mapping in human disease. *Science*, 322:881–888. doi: 10.1126/science.1156409.

[15] Altshuler, D. L., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., et al. (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467:1061–1073. doi: 10.1038/nature09534.

[16] Anand, R. P., Tsaponina, O., Greenwell, P. W., Lee, C. S., Du, W., et al. (2014). Chromosome rearrangements via template switching between diverged repeated sequences. *Genes and Development*, 28:2394–2406. doi: 10.1101/gad.250258.114.

[17] Anderson, S., Bankier, A. T., Barrell, B. G., de Bruijn, M. H. L., Coulson, A. R., et al. (1981). Sequence and organization of the human mitochondrial genome. *Nature*, 290: 457–465. doi: 10.1038/290457a0.

[18] Armstrong, J., Fiddes, I. T., Diekhans, M., and Paten, B. (2019). Whole-genome alignment and comparative annotation. *Annual Review of Animal Biosciences*, 7:41–64. doi: 10.1146/annurev-animal-020518-115005.

[19] Asaithamby, A., Hu, B., and Chen, D. J. (2011). Unrepaired clustered DNA lesions induce chromosome breakage in human cells. *Proceedings of the National Academy of Sciences of the United States of America*, 108:8293–8298. doi: 10.1073/pnas.1016045108.

[20] Audano, P. A., Sulovari, A., Graves-Lindsay, T. A., Cantsilieris, S., Sorensen, M., et al. (2019). Characterizing the major structural variant alleles of the human genome. *Cell*, 176:663–675. doi: 10.1016/j.cell.2018.12.019.

[21] Avery, O. T., Macleod, C. M., and McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from Pneumococcus type III. *Journal of Experimental Medicine*, 79:137–158. doi: 10.1084/jem.79.2.137.

[22] Bailey, M. H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., et al. (2018). Comprehensive characterization of cancer driver genes and mutations. *Cell*, 173:371–385. doi: 10.1016/j.cell.2018.02.060.

[23] Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., et al. (2009). MEME Suite: tools for motif discovery and searching. *Nucleic Acids Research*, 37:W202–W208. doi: 10.1093/nar/gkp335.

[24] Ballouz, S., Dobin, A., and Gillis, J. (2019). Is it time to change the reference genome? *Genome Biology*, 20:1–9. doi: 10.1123/att.13.3.1.

[25] Barbič, A., Zimmer, D. P., and Crothers, D. M. (2003). Structural origins of adenine-tract bending. *Proceedings of the National Academy of Sciences of the United States of America*, 100:2369–2373. doi: 10.1073/pnas.0437877100.

[26] Bębenek, A. and Ziuzia-Graczyk, I. (2018). Fidelity of DNA replication - a matter of proofreading. *Current Genetics*, 64:985–996. doi: 10.1007/s00294-018-0820-1.

[27] Belsare, S., Levy-Sakin, M., Mostovoy, Y., Durinck, S., Chaudhry, S., et al. (2019). Evaluating the quality of the 1000 Genomes Project data. *BMC Genomics*, 20:1–14. doi: 10.1101/383950.

[28] Belyeu, J. R., Brand, H., Wang, H., Zhao, X., Pedersen, B. S., et al. (2021). De novo structural mutation rates and gamete-of-origin biases revealed through genome sequencing of 2,396 families. *American Journal of Human Genetics*, 108:597–607. doi: 10.1016/j.ajhg.2021.02.012.

[29] Benjamin, D., Sato, T., Cibulskis, K., Getz, G., Stewart, C., et al. (2019). Calling somatic SNVs and indels with Mutect2. *bioRxiv*. doi: 10.1101/861054.

[30] Benson, J. D., Chen, Y.-N. P., Cornell-Kennon, S. A., Dorsch, M., Kim, S., et al. (2006). Validating cancer drug targets. *Nature*, 441:451–456. doi: 10.1038/nature04873.

[31] Bergström, A., McCarthy, S. A., Hui, R., Almarri, M. A., Ayub, Q., et al. (2020). Insights into human genetic variation and population history from 929 diverse genomes. *Science*, 367. doi: 10.1126/science.aay5012.

[32] Berti, M. and Vindigni, A. (2016). Replication stress: getting back on track. *Nature Structural and Molecular Biology*, 23:103–109. doi: 10.1038/nsmb.3163.

[33] Besenbacher, S., Sulem, P., Helgason, A., Helgason, H., Kristjansson, H., et al. (2016). Multi-nucleotide de novo mutations in humans. *PLOS Genetics*, 12:1–15. doi: 10.1371/journal.pgen.1006315.

[34] Bird, C. P., Stranger, B. E., Liu, M., Thomas, D. J., Ingle, C. E., et al. (2007). Fast-evolving noncoding sequences in the human genome. *Genome Biology*, 8:1–12. doi: 10.1186/gb-2007-8-6-r118.

[35] Blokzijl, F., Janssen, R., van Boxtel, R., and Cuppen, E. (2018). MutationalPatterns: comprehensive genome-wide analysis of mutational processes. *Genome Medicine*, 10: 1–11. doi: 10.1186/s13073-018-0539-0.

[36] Bochman, M. L., Paeschke, K., and Zakian, V. A. (2012). DNA secondary structures: stability and function of G-quadruplex structures. *Nature Reviews Genetics*, 13:770–780. doi: 10.1038/nrg3296.

[37] Branzei, D. and Foiani, M. (2007). Template switching: from replication fork repair to genome rearrangements. *Cell*, 131:1228–1230. doi: 10.1016/j.cell.2007.12.007.

[38] Buniello, A., Macarthur, J. A., Cerezo, M., Harris, L. W., Hayhurst, J., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, 47:D1005–D1012. doi: 10.1093/nar/gky1120.

[39] Burnham, K. P. and Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods and Research*, 33:261–304. doi: 10.1177/0049124104268644.

[40] Burrow, A. A., Marullo, A., Holder, L. R., and Wang, Y. H. (2010). Secondary structure formation and DNA instability at fragile site FRA16B. *Nucleic Acids Research*, 38: 2865–2877. doi: 10.1093/nar/gkp1245.

[41] Burton, P. R., Clayton, D. G., Cardon, L. R., Craddock, N., Deloukas, P., et al. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447:661–678. doi: 10.1038/nature05911.

[42] Byrska-Bishop, M., Evani, U. S., Zhao, X., Basile, A. O., Abel, H. J., et al. (2021). High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *bioRxiv*. doi: 10.1101/2021.02.06.430068.

[43] Campbell, M. C. and Tishkoff, S. A. (2008). African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. *Annual Review of Genomics and Human Genetics*, 9:403–433. doi: 10.1146/annurev.genom.9.081307.164258.

[44] Campbell, P. J., Getz, G., Korbel, J. O., Stuart, J. M., Jennings, J. L., et al. (2020). Pan-cancer analysis of whole genomes. *Nature*, 578:82–93. doi: 10.1038/s41586-020-1969-6.

[45] Cannan, W. J. and Pederson, D. S. (2016). Mechanisms and consequences of double-strand DNA break formation in chromatin. *Journal of Cellular Physiology*, 231:3–14. doi: 10.1002/jcp.25048.

[46] Cartwright, R. A. (2009). Problems and solutions for estimating indel rates and length distributions. *Molecular Biology and Evolution*, 26:473–480. doi: 10.1093/molbev/msn275.

[47] Carvalho, C. M. B., Bartnik, M., Pehlivan, D., Fang, P., Shen, J., et al. (2012). Evidence for disease penetrance relating to CNV size: Pelizaeus–Merzbacher disease and manifesting carriers with a familial 11 Mb duplication at Xq22. *Clinical Genetics*, 81: 532–541. doi: https://doi.org/10.1111/j.1399-0004.2011.01716.x.

[48] Carvalho, C. M. B., Zhang, F., Liu, P., Patel, A., Sahoo, T., et al. (2009). Complex rearrangements in patients with duplications of MECP2 can occur by fork stalling and template switching. *Human Molecular Genetics*, 18:2188–2203. doi: 10.1093/hmg/ddp151.

[49] Carvalho, C. M. and Lupski, J. R. (2016). Mechanisms underlying structural variant formation in genomic disorders. *Nature Reviews Genetics*, 17:224–238. doi: 10.1038/nrg.2015.25.

[50] Carvalho, C. M., Ramocki, M. B., Pehlivan, D., Franco, L. M., Gonzaga-Jauregui, C., et al. (2011). Inverted genomic segments and complex triplication rearrangements are mediated by inverted repeats in the human genome. *Nature Genetics*, 43:1074–1081. doi: 10.1038/ng.944.

[51] Carvalho, C. M., Coban-Akdemir, Z., Hijazi, H., Yuan, B., Pendleton, M., et al. (4 2019). Interchromosomal template-switching as a novel molecular mechanism for imprinting perturbations associated with temple syndrome. *Genome Medicine*, 11. doi: 10.1186/s13073-019-0633-y.

[52] Caulfield, M., Davies, J., Dennys, M., Elbahy, L., Fowler, T., et al. (2020). National Genomic Research Library. doi: 10.6084/m9.figshare.4530893.v7. URL https://figshare.com/articles/dataset/GenomicEnglandProtocol_pdf/4530893.

[53] Chaisson, M. J., Sanders, A. D., Zhao, X., Malhotra, A., Porubsky, D., et al. (2019). Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nature Communications*, 10:1–16. doi: 10.1038/s41467-018-08148-z.

[54] Chang, H. W., Pandey, M., Kulaeva, O. I., Patel, S. S., and Studitsky, V. M. (2016). Overcoming a nucleosomal barrier to replication. *Science Advances*, 2:24–28. doi: 10.1126/sciadv.1601865.

[55] Chang, H. H., Pannunzio, N. R., Adachi, N., and Lieber, M. R. (2017). Non-homologous DNA end joining and alternative pathways to double-strand break repair. *Nature Reviews Molecular Cell Biology*, 18:495–506. doi: 10.1038/nrm.2017.48.

[56] Chatterjee, N. and Walker, G. C. (2017). Mechanisms of DNA damage, repair, and mutagenesis. *Environmental and Molecular Mutagenesis*, 58:235–263. doi: 10.1002/em.

[57] Chiang, C., Scott, A. J., Davis, J. R., Tsang, E. K., Li, X., et al. (2017). The impact of structural variation on human gene expression. *Nature Genetics*, 49:692–699. doi: 10.1038/ng.3834.

[58] Chiu, T. P., Comoglio, F., Zhou, T., Yang, L., Paro, R., et al. (2016). DNAshapeR: an R/Bioconductor package for DNA shape prediction and feature encoding. *Bioinformatics*, 32:1211–1213. doi: 10.1093/bioinformatics/btv735.

[59] Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., et al. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*, 31:213–219. doi: 10.1038/nbt.2514.

[60] Collins, R. L., Brand, H., Karczewski, K. J., Zhao, X., Alföldi, J., et al. (2020). A structural variation reference for medical and population genetics. *Nature*, 581:444–451. doi: 10.1038/s41586-020-2287-8.

[61] Cooke, D. P., Wedge, D. C., and Lunter, G. (2021). A unified haplotype-based method for accurate and comprehensive variant calling. *Nature Biotechnology*. doi: 10.1038/s41587-021-00861-3.

[62] Cortés-Ciriano, I., Lee, J. J. K., Xi, R., Jain, D., Jung, Y. L., et al. (2020). Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. *Nature Genetics*, 52:331–341. doi: 10.1038/s41588-019-0576-7.

[63] Cox, D. R. (1961). Tests of separate families of hypotheses. In *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, pages 105–123. doi: 10.1002/9781118445112.stat05782.

[64] Cox, D. R. (1962). Further results on tests of separate families of hypotheses. *Journal of the Royal Statistical Society: Series B (Methodological)*, 24:406–424. doi: 10.2307/2334876.

[65] Crick, F. H. C., Barnett, L., Brenner, S., and Watts-Tobin, R. J. (1961). General nature of the genetic code for proteins. *Nature*, 4809:1227–1232. doi: 10.1007/s12045-019-0884-3.

[66] Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227:561–563. doi: 10.1038/227561a0.

[67] Dahm, R. (2008). Discovering DNA: Friedrich Miescher and the early years of nucleic acid research. *Human Genetics*, 122:565–581. doi: 10.1007/s00439-007-0433-0.

[68] Davies, H., Glodzik, D., Morganella, S., Yates, L. R., Staaf, J., et al. (4 2017). Hrdetect is a predictor of brca1 and brca2 deficiency based on mutational signatures. *Nature Medicine*, 23:517–525. doi: 10.1038/nm.4292.

[69] De Maio, N. (2021). The cumulative indel model: fast and accurate statistical evolutionary alignment. *Systematic Biology*, 70:236–257. doi: 10.1093/sysbio/syaa050.

[70] Delaneau, O., Zagury, J. F., and Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. *Nature Methods*, 10:5–6. doi: 10.1038/nmeth.2307.

[71] Depristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43:491–501. doi: 10.1038/ng.806.

[72] DeVoe, H. and Tinoco, I. (1962). The stability of helical polynucleotides: base contributions. *Journal of Molecular Biology*, 4:500–517. doi: 10.1016/S0022-2836(62)80105-3.

[73] Domanska, D., Kanduri, C., Simovski, B., and Sandve, G. K. (2018). Mind the gaps: overlooking assembly gaps confounds statistical testing in genome analysis. *BMC Bioinformatics*, 19:1–9. doi: 10.1101/252973.

[74] Dowell, R. D. and Eddy, S. R. (2006). Efficient pairwise RNA structure prediction and alignment using sequence alignment constraints. *BMC Bioinformatics*, 7:1–18. doi: 10.1186/1471-2105-7-400.

[75] Dršata, T., Špačková, N., Jurečka, P., Zgarbová, M., Šponer, J., et al. (2014). Mechanical properties of symmetric and asymmetric DNA A-tracts: implications for looping and nucleosome positioning. *Nucleic Acids Research*, 42:7383–7394. doi: 10.1093/nar/gku338.

[76] Dukler, N., Huang, Y. F., and Siepel, A. (2020). Phylogenetic modeling of regulatory element turnover based on epigenomic data. *Molecular Biology and Evolution*, 37:2137–2152. doi: 10.1093/molbev/msaa073.

[77] Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998). *Biological Sequence Analysis: Probabalistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge. doi: 10.1017/CBO9780511790492.

[78] Dyer, M. J., Zani, V. J., Lu, W. Z., O'Byrne, A., Mould, S., et al. (1994). BCL2 translocations in leukemias of mature B cells. *Blood*, 83:3682–3688. doi: 10.1182/blood.v83.12.3682.3682.

[79] Earl, D., Nguyen, N., Hickey, G., Harris, R. S., Fitzgerald, S., et al. (2014). Alignathon: a competitive assessment of whole-genome alignment methods. *Genome Research*, 24: 2077–2089. doi: 10.1101/gr.174920.114.

[80] Ebert, P., Audano, P. A., Zhu, Q., Rodriguez-Martin, B., Porubsky, D., et al. (2021). Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science*, 7117:eabf7117.

[81] Eggertsson, H. P., Jonsson, H., Kristmundsdottir, S., Hjartarson, E., Kehr, B., et al. (2017). Graphtyper enables population-scale genotyping using pangenome graphs. *Nature Genetics*, 49:1654–1660. doi: 10.1038/ng.3964.

[82] Emerling, B. M., Hurov, J. B., Poulogiannis, G., Tsukazawa, K. S., Choo-Wing, R., et al. (2013). Depletion of a putatively druggable class of phosphatidylinositol kinases inhibits growth of p53-null tumors. *Cell*, 155:844. doi: 10.1016/j.cell.2013.09.057.

[83] Emmert, S. and Kraemer, K. H. (2013). Do not underestimate nucleotide excision repair: it predicts not only melanoma risk but also survival outcome. *Journal of Investigative Dermatology*, 133:1713–1717. doi: 10.1038/jid.2013.72.

[84] Falconer, E., Hills, M., Naumann, U., Poon, S. S., Chavez, E. A., et al. (2012). DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. *Nature Methods*, 9:1107–1112. doi: 10.1038/nmeth.2206.

[85] Fan, Y., Xi, L., Hughes, D. S., Zhang, J., Zhang, J., et al. (2016). MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biology*, 17:178. doi: 10.1186/s13059-016-1029-6.

[86] Fay, J. C. and Wu, C. I. (1999). A human population bottleneck can account for the discordance between patterns of mitochondrial versus nuclear DNA variation. *Molecular Biology and Evolution*, 16:1003–1005. doi: 10.1093/oxfordjournals.molbev.a026175.

[87] Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17:368–376. doi: 10.1007/BF01734359.

[88] Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., et al. (1976). Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature*, 260:500–507. doi: 10.1038/260500a0.

[89] Fletcher, W. and Yang, Z. (2009). INDELible: a flexible simulator of biological sequence evolution. *Molecular Biology and Evolution*, 26:1879–1888. doi: 10.1093/molbev/msp098.

[90] Forney, D. (1973). The Viterbi algorithm. *Proceedings of the IEEE*, 61:268–278. doi: 10.1049/ic:20060556.

[91] Frankish, A., Diekhans, M., Ferreira, A. M., Johnson, R., Jungreis, I., et al. (2019). GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research*, 47:D766–D773. doi: 10.1093/nar/gky955.

[92] Fu, Y.-X. (1995). Statistical properties of segregating sites. *Theoretical Population Biology*, 48:172–197. doi: 10.1006/tpbi.1995.1025.

[93] Gao, T. and Qian, J. (2020). EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. *Nucleic Acids Research*, 48: D58–D64. doi: 10.1093/nar/gkz980.

[94] Garibyan, L. and Fisher, D. E. (2010). How sunlight causes melanoma. *Current Oncology Reports*, 12:319–326. doi: 10.1007/s11912-010-0119-y.

[95] Garrison, E., Sirén, J., Novak, A. M., Hickey, G., Eizenga, J. M., et al. (2018). Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology*, 36:875–881. doi: 10.1038/nbt.4227.

[96] Georgakopoulos-Soares, I., Morganella, S., Jain, N., Hemberg, M., and Nik-Zainal, S. (2018). Noncanonical secondary structures arising from non-B DNA motifs are determinants of mutagenesis. *Genome Research*, 28:1264–1271. doi: 10.1101/gr.231688.117.

[97] Gerstein, M. B., Bruce, C., Rozowsky, J. S., Zheng, D., Du, J., et al. (2007). What is a gene, post-ENCODE? History and updated definition. *Genome Research*, 17:669–681. doi: 10.1101/gr.6339607.

[98] Gerstung, M., Jolly, C., Leshchiner, I., Dentro, S. C., Gonzalez, S., et al. (2020). The evolutionary history of 2,658 cancers. *Nature*, 578:122–128. doi: 10.1038/s41586-019-1907-7.

[99] Ghezraoui, H., Piganeau, M., Renouf, B., Renaud, J. B., Sallmyr, A., et al. (2014). Chromosomal translocations in human cells are generated by canonical nonhomologous end-joining. *Molecular Cell*, 55:829–842. doi: 10.1016/j.molcel.2014.08.002.

[100] Gilbert, S. L., Dobyns, W. B., and Lahn, B. T. (2005). Genetic links between brain development and brain evolution. *Nature Reviews Genetics*, 6:581–590. doi: 10.1038/nrg1634.

[101] Gillespie, J. H. (1998). *Population Genetics: A Concise Guide*. The Johns Hopkins University Press, Baltimore, Maryland.

[102] Gittelman, R. M., Hun, E., Ay, F., Madeoy, J., Pennacchio, L., et al. (2015). Comprehensive identification and analysis of human accelerated regulatory DNA. *Genome Research*, 25:1245–1255. doi: 10.1101/gr.192591.115.

[103] Goldman, N. (1993). Statistical tests of models of DNA substitution. *Journal of Molecular Evolution*, 36:182–198. doi: 10.1007/BF00166252.

[104] Goldmann, J. M., Seplyarskiy, V. B., Wong, W. S., Vilboux, T., Neerincx, P. B., et al. (2018). Germline de novo mutation clusters arise during oocyte aging in genomic regions with high double-strand-break incidence. *Nature Genetics*, 50:487–492. doi: 10.1038/s41588-018-0071-6.

[105] Gonzalez-Huici, V., Szakal, B., Urulangodi, M., Psakhye, I., Castellucci, F., et al. (2014). DNA bending facilitates the error-free DNA damage tolerance pathway and upholds genome integrity. *EMBO Journal*, 33:327–340. doi: 10.1002/embj.201387425.

[106] Goodman, M. F. and Tippin, B. (2000). The expanding polymerase universe. *Nature Reviews Molecular Cell Biology*, 1:101–109. doi: 10.1038/35040051.

[107] Gotoh, O. (1982). An improved algorithm for matching biological sequences. *Journal of Molecular Biology*, 162:705–708. doi: 10.1016/0022-2836(82)90398-9.

[108] Graur, D., Zheng, Y., Price, N., Azevedo, R. B., Zufall, R. A., et al. (2013). On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. *Genome Biology and Evolution*, 5:578–590. doi: 10.1093/gbe/evt028.

[109] Gray, M. W. (2012). Mitochondrial evolution. *Cold Spring Harbor Perspectives in Biology*, 4:a011403. doi: 10.1101/cshperspect.a011403.

[110] Griffith, F. (1928). The significance of penumococcal types. *Journal of Hygiene*, 27: 113–159. doi: 10.1017/S0022172400040420.

[111] Gu, W., Zhang, F., and Lupski, J. R. (2008). Mechanisms for human genomic rearrangements. *PathoGenetics*, 1:4. doi: 10.1186/1755-8417-1-4.

[112] Gu, X., Fu, Y. X., and Li, W. H. (1995). Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. *Molecular Biology and Evolution*, 12:546–557.

[113] Guo, K., Cui, J., Quan, M., Xie, D., Jia, Z., et al. (2017). The novel KLF4/MSI2 signaling pathway regulates growth and metastasis of pancreatic cancer. *Clinical Cancer Research*, 23:687–696. doi: 10.1158/1078-0432.CCR-16-1064.

[114] Hall, P. and Wilson, S. R. (1991). Two guidelines for bootstrap hypothesis testing. *Biometrics*, 47:757–762.

[115] Hansen, R. S., Thomas, S., Sandstrom, R., Canfield, T. K., Thurman, R. E., et al. (2010). Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proceedings of the National Academy of Sciences of the United States of America*, 107:139–144. doi: 10.1073/pnas.0912402107.

[116] Harris, D. N., Song, W., Shetty, A. C., Levano, K. S., Cáceres, O., et al. (2018). Evolutionary genomic dynamics of Peruvians before, during, and after the Inca Empire. *Proceedings of the National Academy of Sciences of the United States of America*, 115: E6526–E6535. doi: 10.1073/pnas.1720798115.

[117] Harris, K. and Nielsen, R. (2014). Error-prone polymerase activity causes multinucleotide mutations in humans. *Genome Research*, 24:1445–1454. doi: 10.1101/gr.170696.113.

[118] Harris, R. S. (2007). *Improved Pairwise Alignment of Genomic DNA*. PhD thesis, The Pennsylvania State University.

[119] Hasegawa, M., Kishino, H., and Yano, T.-a. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22:160–174. doi: 10.1007/BF02101694.

[120] Hastings, P. J., Ira, G., and Lupski, J. R. (2009). A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLOS Genetics*, 5:e1000327. doi: 10.1371/journal.pgen.1000327.

[121] Hegde, M. L., Hazra, T. K., and Mitra, S. (2008). Early steps in the DNA base excision/single-strand interruption repair pathway in mammalian cells. *Cell Research*, 18:27–47. doi: 10.1038/cr.2008.8.

[122] Hein, J., Schierup, M. H., and Wiuf, C. (2005). *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford University Press, Oxford.

[123] Ho, S. S., Urban, A. E., and Mills, R. E. (2020). Structural variation in the sequencing era. *Nature Reviews Genetics*, 21:171–189. doi: 10.1038/s41576-019-0180-9.

[124] Hodis, E., Watson, I. R., Kryukov, G. V., Arold, S. T., Imielinski, M., et al. (2012). A landscape of driver mutations in melanoma. *Cell*, 150:251–263. doi: 10.1016/j.cell.2012.06.024.

[125] Hoeck, A. V., Tjoonk, N. H., Boxtel, R. V., and Cuppen, E. (5 2019). Portrait of a cancer: mutational signature analyses for cancer diagnostics. *BMC Cancer*, 19. doi: 10.1186/s12885-019-5677-2.

[126] Hollox, E. J., Zuccherato, L. W., and Tucci, S. (1 2022). Genome structural variation in human evolution. *Trends in Genetics*, 38:45–58. doi: 10.1016/j.tig.2021.06.015.

[127] Holmes, I. (2005). Accelerated probabilistic inference of RNA structure evolution. *BMC Bioinformatics*, 6:1–22. doi: 10.1186/1471-2105-6-73.

[128] Holmes, I. H. (2017). Solving the master equation for indels. *BMC Bioinformatics*, 18:255. doi: 10.1186/s12859-017-1665-1.

[129] Hsieh, P. and Yamane, K. (2008). DNA mismatch repair: molecular mechanism, cancer, and ageing. *Mechanisms of Ageing and Development*, 129:391–407. doi: 10.1016/j.mad.2008.02.012.

[130] Hubisz, M. J. and Pollard, K. S. (2014). Exploring the genesis and functions of human accelerated regions sheds light on their role in human evolution. *Current Opinion in Genetics and Development*, 29:15–21. doi: 10.1016/j.gde.2014.07.005.

[131] International HapMap Consortium. (2003). The International HapMap Project. *Nature*, 426:789–796. doi: 10.1038/nature02168.

[132] Iqbal, Z., Caccamo, M., Turner, I., Flicek, P., and McVean, G. (2012). De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nature Genetics*, 44:226–232. doi: 10.1038/ng.1028.

[133] Isabelle, V., Prévost, C., Spotheim-Maurizot, M., Sabattier, R., and Charlier, M. (1995). Radiation-induced damages in single- and double-stranded DNA. *The International Journal of Radiation Biology*, 67:169–176. doi: 10.1080/09553009514550211.

[134] Jakobsson, M., Scholz, S. W., Scheet, P., Gibbs, J. R., VanLiere, J. M., et al. (2008). Genotype, haplotype and copy-number variation in worldwide human populations. *Nature*, 451:998–1003. doi: 10.1038/nature06742.

[135] Jones, D., Raine, K. M., Davies, H., Tarpey, P. S., Butler, A. P., et al. (2016). cgpCaVEManWrapper: simple execution of caveman in order to detect somatic single nucleotide variants in NGS data. *Current Protocols in Bioinformatics*, 56:1–15. doi: 10.1002/cpbi.20.

[136] Jónsson, H., Sulem, P., Kehr, B., Kristmundsdottir, S., Zink, F., et al. (2017). Whole genome characterization of sequence diversity of 15,220 Icelanders. *Scientific Data*, 4: 170115. doi: 10.1038/sdata.2017.115.

[137] Jónsson, H., Sulem, P., Kehr, B., Kristmundsdottir, S., Zink, F., et al. (2017). Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature*, 549:519–522. doi: 10.1038/nature24018.

[138] Jukes, T. and Cantor, C. (1969). Evolution of protein molecules. In Munro, H. N. and Allison, J. B., editors, *Mammalian protein metabolism*, chapter 24, pages 22–126. Academic Press, New York. doi: 10.1016/B978-1-4832-3211-9.50009-7.

[139] Kang, T. W., Kim, H. J., Ju, H., Kim, J. H., Jeon, Y. J., et al. (9 2010). Genome-wide association of serum bilirubin levels in korean population. *Human Molecular Genetics*, 19:3672–3678. doi: 10.1093/hmg/ddq281.

[140] Kaplanis, J., Akawi, N., Gallone, G., McRae, J. F., Prigmore, E., et al. (2019). Exome-wide assessment of the functional impact and pathogenicity of multinucleotide mutations. *Genome Research*, pages 1047–1056. doi: 10.1101/gr.239756.118.

[141] Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581:434–443. doi: 10.1038/s41586-020-2308-7.

[142] Karolchik, D., Hinricks, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., et al. (2004). The UCSC table browser data retrieval tool. *Nucleic Acids Research*, 32:493–496. doi: 10.1093/nar/gkh103.

[143] Kass, E. M. and Jasin, M. (2010). Collaboration and competition between DNA double-strand break repair pathways. *FEBS Letters*, 584:3703–3708. doi: 10.1016/j.febslet.2010.07.057.

[144] Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90:773–795. doi: 10.1080/01621459.1995.10476572.

[145] Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., et al. (2002). The Human Genome Browser at UCSC. *Genome Research*, 12:996–1006. doi: 10.1101/gr.229102.

[146] Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S., and Karolchik, D. (2010). BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics*, 26:2204–2207. doi: 10.1093/bioinformatics/btq351.

[147] Khan, W., Varma Saripella, G., Ludwig, T., Cuppens, T., Thibord, F., et al. (2018). MACARON: a python framework to identify and re-annotate multi-base affected codons in whole genome/exome sequence data. *Bioinformatics*, 34:3396–3398. doi: 10.1093/bioinformatics/bty382.

[148] Kim, S. Y., Jacob, L., and Speed, T. P. (2014). Combining calls from multiple somatic mutation-callers. *BMC Bioinformatics*, 15:1–8. doi: 10.1186/1471-2105-15-154.

[149] Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16:111–120. doi: 10.1007/BF01731581.

[150] Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge University Press. doi: 10.1017/CBO9780511623486.

[151] Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., et al. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, 46:310–315. doi: 10.1038/ng.2892.

[152] Kloosterman, W. P., Francioli, L. C., Hormozdiari, F., Marschall, T., Hehir-kwa, J. Y., et al. (2015). Characteristics of de novo structural changes in the human genome. *Genome Research*, 25:792–801. doi: 10.1101/gr.185041.114.19.

[153] Kolesnikov, A., Goel, S., Nattestad, M., Yun, T., Baid, G., et al. (2021). DeepTrio: variant calling in families using deep learning. *bioRxiv*. doi: 10.1101/2021.04.05.438434.

[154] Kong, A., Frigge, M. L., Masson, G., Besenbacher, S., Sulem, P., et al. (2012). Rate of *de novo* mutations and the importance of father's age to disease risk. *Nature*, 488:471–475. doi: 10.1038/nature11396.

[155] Koo, H. S., Wu, H. M., and Crothers, D. M. (1986). DNA bending at adenine · thymine tracts. *Nature*, 320:501–506. doi: 10.1038/320501a0.

[156] Koren, A., Polak, P., Nemesh, J., Michaelson, J. J., Sebat, J., et al. (2012). Differential relationship of DNA replication timing to different forms of human mutation and variation. *American Journal of Human Genetics*, 91:1033–1040. doi: 10.1016/j.ajhg.2012.10.018.

[157] Koren, A., Handsaker, R. E., Kamitaki, N., Karlić, R., Ghosh, S., et al. (2014). Genetic variation in human DNA replication timing. *Cell*, 159:1015–1026. doi: 10.1016/j.cell.2014.10.025.

[158] Kornberg, R. D. (1977). Structure of chromatin. *Annual Review of Biochemistry*, 46:931–954. doi: 10.1146/annurev.bi.46.070177.004435.

[159] Korneliussen, T. S., Moltke, I., Albrechtsen, A., and Nielsen, R. (2013). Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data. *BMC Bioinformatics*, 14. doi: 10.1186/1471-2105-14-289.

[160] Korona, D. A., Lecompte, K. G., and Pursell, Z. F. (2011). The high fidelity and unique error signature of human DNA polymerase ε. *Nucleic Acids Research*, 39:1763–1773. doi: 10.1093/nar/gkq1034.

[161] Kossel, A. (1910). *The chemical composition of the cell nucleus; Nobel Lecture; The Nobel Prize in Physiology or Medicine 1910*.

[162] Kostka, D., Holloway, A. K., and Pollard, K. S. (2018). Developmental loci harbor clusters of accelerated regions that evolved independently in ape lineages. *Molecular Biology and Evolution*, 35:2034–2045. doi: 10.1093/molbev/msy109.

[163] Kronenberg, Z. N., Fiddes, I. T., Gordon, D., Murali, S., Cantsilieris, S., et al. (2018). High-resolution comparative analysis of great ape genomes. *Science*, 360:eaar6343. doi: 10.1126/science.aar6343.

[164] Kucab, J. E., Zou, X., Morganella, S., Joel, M., Nanda, A. S., et al. (2019). A compendium of mutational signatures of environmental agents. *Cell*, 177:821–836. doi: 10.1016/j.cell.2019.03.001.

[165] Kunkel, T. (2009). Evolving views of DNA replication (in)fidelity. *Cold Spring Harbor Symposia on Quantitative Biology*, LXXIV:91–101. doi: 10.1101/sqb.2009.74.027.

[166] Lai, Y. and Sun, F. (2003). The relationship between microsatellite slippage mutation rate and the number of repeat units. *Molecular Biology and Evolution*, 20:2123–2131. doi: 10.1093/molbev/msg228.

[167] Lam, H. Y., Mu, X. J., Stütz, A. M., Tanzer, A., Cayting, P. D., et al. (2010). Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nature Biotechnology*, 28:47–55. doi: 10.1038/nbt.1600.

[168] Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409:860–921. doi: 10.1038/35057062.

[169] Langley, A. R., Gräf, S., Smith, J. C., and Krude, T. (2016). Genome-wide identification and characterisation of human DNA replication origins by initiation site sequencing (ini-seq). *Nucleic Acids Research*, 44:10230–10247. doi: 10.1093/nar/gkw760.

[170] Leclercq, S. B., Rivals, E., and Jarne, P. (2010). DNA slippage occurs at microsatellite loci without minimal threshold length in humans: a comparative genomic approach. *Genome Biology and Evolution*, 2:325–335. doi: 10.1093/gbe/evq023.

[171] Lee, J. H., Han, Y. S., Yoon, Y. M., Yun, C. W., Yun, S. P., et al. (2017). Role of HSPA1L as a cellular prion protein stabilizer in tumor progression via HIF-1α/GP78 axis. *Oncogene*, 36:6555–6567. doi: 10.1038/onc.2017.263.

[172] Lee, J. A., Carvalho, C. M., and Lupski, J. R. (2007). A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell*, 131:1235–1247. doi: 10.1016/j.cell.2007.11.037.

[173] Lee-Six, H., Øbro, N. F., Shepherd, M. S., Grossmann, S., Dawson, K., et al. (2018). Population dynamics of normal human blood inferred from somatic mutations. *Nature*, 561:473–478. doi: 10.1038/s41586-018-0497-0.

[174] Lee-Six, H., Olafsson, S., Ellis, P., Osborne, R. J., Sanders, M. A., et al. (2019). The landscape of somatic mutation in normal colorectal epithelial cells. *Nature*, 574:532–537. doi: 10.1038/s41586-019-1672-7.

[175] Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L., et al. (2007). The diploid genome sequence of an individual human. *PLOS Biology*, 5:2113–2144. doi: 10.1371/journal.pbio.0050254.

[176] Li, C. and Luscombe, N. M. (2020). Nucleosome positioning stability is a modulator of germline mutation rate variation across the human genome. *Nature Communications*, 11:1–13. doi: 10.1038/s41467-020-15185-0.

[177] Li, G. M. (2008). Mechanisms and functions of DNA mismatch repair. *Cell Research*, 18:85–98. doi: 10.1038/cr.2007.115.

[178] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25:2078–2079. doi: 10.1093/bioinformatics/btp352.

[179] Li, Y., Roberts, N., Weischenfeldt, J., Wala, J. A., Shapira, O., et al. (2017). Patterns of structural variation in human cancer. *bioRxiv*. doi: 10.1101/181339.

[180] Li, Y., Roberts, N. D., Wala, J. A., Shapira, O., Schumacher, S. E., et al. (2020). Patterns of somatic structural variation in human cancer genomes. *Nature*, 578:112–121. doi: 10.1038/s41586-019-1913-9.

[181] Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M. F., Parker, B. J., et al. (2011). A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, 478:476–482. doi: 10.1038/nature10530.

[182] Liò, P. and Goldman, N. (1998). Models of molecular evolution and phylogeny. *Genome Research*, 8:1233–1244. doi: 10.1101/gr.8.12.1233.

[183] Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., et al. (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518:197–206. doi: 10.1038/nature14177.

[184] Loh, P. R., Danecek, P., Palamara, P. F., Fuchsberger, C., Reshef, Y. A., et al. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. *Nature Genetics*, 48:1443–1448. doi: 10.1038/ng.3679.

[185] Löytynoja, A. and Goldman, N. (2017). Short template switch events explain mutation clusters in the human genome. *Genome Research*, 27:1039–1049. doi: 10.1101/gr.214973.116.

[186] MacArthur, D. G., Manolio, T. A., Dimmock, D. P., Rehm, H. L., Shendure, J., et al. (2014). Guidelines for investigating causality of sequence variants in human disease. *Nature*, 508:469–476. doi: 10.1038/nature13127.

[187] Mahmoud, M., Gobet, N., Cruz-Dávalos, D. I., Mounier, N., Dessimoz, C., et al. (2019). Structural variant calling: the long and the short of it. *Genome Biology*, 20:1–14. doi: 10.1186/s13059-019-1828-7.

[188] Mailund, T., Dutheil, J. Y., Hobolth, A., Lunter, G., and Schierup, M. H. (2011). Estimating divergence time and ancestral effective population size of Bornean and Sumatran orangutan subspecies using a coalescent hidden Markov model. *PLOS Genetics*, 7: e1001319. doi: 10.1371/journal.pgen.1001319.

[189] Mailund, T., Halager, A. E., Westergaard, M., Dutheil, J. Y., Munch, K., et al. (2012). A new isolation with migration model along complete genomes infers very different divergence processes among closely related great ape species. *PLOS Genetics*, 8: e1003125. doi: 10.1371/journal.pgen.1003125.

[190] Mailund, T., Munch, K., and Schierup, M. H. (2014). Lineage sorting in apes. *Annual Review of Genetics*, 48:519–535. doi: 10.1146/annurev-genet-120213-092532.

[191] Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., et al. (2016). The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*, 538: 201–206. doi: 10.1038/nature18964.

[192] Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18:50–60. doi: 10.1214/aoms/1177730491.

[193] Maretty, L., Jensen, J. M., Petersen, B., Sibbesen, J. A., Liu, S., et al. (2017). Sequencing and *de novo* assembly of 150 genomes from Denmark as a population reference. *Nature*, 548:87–91. doi: 10.1038/nature23264.

[194] Marians, K. J. (2018). Lesion bypass and the reactivation of stalled replication forks. *Annual Review of Biochemstry*, 87:1–22. doi: 10.1146/annurev-biochem-062917-011921.

[195] Marteijn, J. A., Lans, H., Vermeulen, W., and Hoeijmakers, J. H. (2014). Understanding nucleotide excision repair and its roles in cancer and ageing. *Nature Reviews Molecular Cell Biology*, 15:465–481. doi: 10.1038/nrm3822.

[196] Martincorena, I. and Campbell, P. J. (2015). Somatic mutation in cancer and normal cells. *Science*, 349:1483–1489. doi: 10.1126/science.aab4082.

[197] Martínez-Jiménez, F., Muiños, F., Sentís, I., Deu-Pons, J., Reyes-Salazar, I., et al. (2020). A compendium of mutational cancer driver genes. *Nature Reviews Cancer*, 20:555–572. doi: 10.1038/s41568-020-0290-x.

[198] Mathieson, I. and McVean, G. (2012). Differential confounding of rare and common variants in spatially structured populations. *Nature Genetics*, 44:243–246. doi: 10.1038/ng.1074.

[199] McCulloch, S. D. and Kunkel, T. A. (2008). The fidelity of DNA synthesis by eukaryotic replicative and translesion synthesis polymerases. *Cell Research*, 18:148–161. doi: 10.1038/cr.2008.4.

[200] McDermott, D. H., Gao, J. L., Liu, Q., Siwicki, M., Martens, C., et al. (2015). Chromothriptic cure of WHIM syndrome. *Cell*, 160:686–699. doi: 10.1016/j.cell.2015.01.014.

[201] McDonald, M. J., Wang, W. C., Huang, H. D., and Leu, J. Y. (2011). Clusters of nucleotide substitutions and insertion/deletion mutations are associated with repeat sequences. *PLOS Biology*, 9:e1000622. doi: 10.1371/journal.pbio.1000622.

[202] McElhinny, S. A., Havener, J. M., Garcia-Diaz, M., Juárez, R., Bebenek, K., et al. (2005). A gradient of template dependence defines distinct biological roles for family X polymerases in nonhomologous end joining. *Molecular Cell*, 19:357–366. doi: 10.1016/j.molcel.2005.06.012.

[203] McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., et al. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, 17:1–14. doi: 10.1186/s13059-016-0974-4.

[204] McRae, J. F., Clayton, S., Fitzgerald, T. W., Kaplanis, J., Prigmore, E., et al. (2017). Prevalence and architecture of de novo mutations in developmental disorders. *Nature*, 542:433–438. doi: 10.1038/nature21062.

[205] McVean, G. (2009). A genealogical interpretation of principal components analysis. *PLOS Genetics*, 5. doi: 10.1371/journal.pgen.1000686.

[206] Melton, C., Reuter, J. A., Spacek, D. V., and Snyder, M. (2015). Recurrent somatic mutations in regulatory regions of human cancer genomes. *Nature Genetics*, 47:710–716. doi: 10.1038/ng.3332.

[207] Merkel, D. (2014). Docker: lightweight linux containers for consistent development and deployment. *Linux Journal*, 2014.

[208] Meselson, M. and Stahl, F. W. (1958). The replication of DNA in *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America*, 44: 671–682. doi: 10.1073/pnas.44.7.671.

[209] Michaelson, J. J., Shi, Y., Gujral, M., Zheng, H., Malhotra, D., et al. (2012). Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell*, 151:1431–1442. doi: 10.1016/j.cell.2012.11.019.

[210] Miescher-Rüsch, F. (1871). Ueber die chemische Zusammensetzung der Eiterzellen. In *Medicinisch-chemische Untersuchungen*, page 486–501. August Hirschwald, Berlin.

[211] Miga, K. H. and Wang, T. (2021). The need for a human pangenome reference sequence. *Annual Review of Genomics and Human Genetics*, 22:1–22. doi: 10.1146/annurev-genom-120120-081921.

[212] Mikkelsen, T. S., Hillier, L. W., Eichler, E. E., Zody, M. C., Jaffe, D. B., et al. (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437:69–87. doi: 10.1038/nature04072.

[213] Miles, A., Rodriguez, M., Ralph, P., Harding, N., Pisupati, R., et al. (2021). cggh/scikit-allel: v1.3.3. URL https://doi.org/10.5281/zenodo.4759368.

[214] Min Jou, W., Haegeman, G., Ysebaert, M., and Fiers, W. (1972). Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. *Nature*, 237:82–88. doi: 10.1038/237082a0.

[215] Mirkin, E. V. and Mirkin, S. M. (2007). Replication fork stalling at natural impediments. *Microbiology and Molecular Biology Reviews*, 71:13–35. doi: 10.1128/mmbr.00030-06.

[216] Mölder, F., Jablonski, K. P., Letcher, B., Hall, M. B., Tomkins-Tinch, C. H., et al. (2021). Sustainable data analysis with Snakemake. *F1000Research*, 10:33. doi: 10.12688/f1000research.29032.1.

[217] Moncunill, V., Gonzalez, S., Beà, S., Andrieux, L. O., Salaverria, I., et al. (2014). Comprehensive characterization of complex structural variations in cancer by directly comparing genome sequence reads. *Nature Biotechnology*, 32:1106–1112. doi: 10.1038/nbt.3027.

[218] Montgomery, S. B., Goode, D. L., Kvikstad, E., Albers, C. A., Zhang, Z. D., et al. (2013). The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. *Genome Research*, 23:749–761. doi: 10.1101/gr.148718.112.

[219] Moore, L., Leongamornlert, D., Coorens, T. H., Sanders, M. A., Ellis, P., et al. (2020). The mutational landscape of normal human endometrial epithelium. *Nature*, 580:640–646. doi: 10.1038/s41586-020-2214-z.

[220] Morita, R., Nakane, S., Shimada, A., Inoue, M., Iino, H., et al. (2010). Molecular mechanisms of the whole DNA repair system: a comparison of bacterial and eukaryotic systems. *Journal of Nucleic Acids*, 2010. doi: 10.4061/2010/179594.

[221] Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48:443–453. doi: 10.1016/0022-2836(70)90057-4.

[222] Neph, S., Kuehn, M. S., Reynolds, A. P., Haugen, E., Thurman, R. E., et al. (2012). BEDOPS: high-performance genomic feature operations. *Bioinformatics*, 28:1919–1920. doi: 10.1093/bioinformatics/bts277.

[223] Nielsen, R., Hellmann, I., Hubisz, M., Bustamante, C., and Clark, A. G. (2007). Recent and ongoing selection in the human genome. *Nature Reviews Genetics*, 8:857–868. doi: 10.1038/nrg2187.

[224] Nirenberg, M., Leder, P., Bernfield, M., Brimacombe, R., Trupin, J., et al. (1965). RNA codewords and protein synthesis, VII. On the general nature of the RNA code. *Proceedings of the National Academy of Sciences of the United States of America*, 53:1161–1168. doi: 10.1073/pnas.53.5.1161.

[225] North, B. V., Curtis, D., and Sham, P. C. (2002). A note on the calculation of empirical P values from Monte Carlo procedures. *The American Journal of Human Genetics*, 71:439. doi: 10.1080/13518040701205365.

[226] Northam, M. R., Moore, E. A., Mertz, T. M., Binz, S. K., Stith, C. M., et al. (2014). DNA polymerases ζ and Rev1 mediate error-prone bypass of non-B DNA structures. *Nucleic Acids Research*, 42:290–306. doi: 10.1093/nar/gkt830.

[227] Okazaki, R., Okazaki, T., Sakabe, K., Sugimoto, K., and Sugino, A. (1968). Mechanism of DNA chain growth, I: possible discontinuity and unusual secondary structure of newly synthesized chains. *Proceedings of the National Academy of Sciences*, 59:598–605. doi: 10.1073/pnas.59.2.598.

[228] Ottaviani, D., LeCain, M., and Sheer, D. (2014). The role of microhomology in genomic structural variation. *Trends in Genetics*, 30:85–94. doi: 10.1016/j.tig.2014.01.001.

[229] Pang, A. W., MacDonald, J. R., Pinto, D., Wei, J., Rafiq, M. A., et al. (2010). Towards a comprehensive structural variation map of an individual human genome. *Genome Biology*, 11:R52. doi: 10.1186/gb-2010-11-5-r52.

[230] Paten, B., Herrero, J., Beal, K., Fitzgerald, S., and Birney, E. (2008). Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Research*, 18:1814–1828. doi: 10.1101/gr.076554.108.

[231] Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PLOS Genetics*, 2:2074–2093. doi: 10.1371/journal.pgen.0020190.

[232] Pećina-Šlaus, N., Kafka, A., Salamon, I., and Bukovac, A. (2020). Mismatch repair pathway, genome stability and cancer. *Frontiers in Molecular Biosciences*, 7:1–12. doi: 10.3389/fmolb.2020.00122.

[233] Perner, J., Abbas, S., Nowicki-Osuch, K., Devonshire, G., Eldridge, M. D., et al. (12 2020). The mutread method detects mutational signatures from low quantities of cancer dna. *Nature Communications*, 11. doi: 10.1038/s41467-020-16974-3.

[234] Picco, G., Chen, E. D., Alonso, L. G., Behan, F. M., Gonçalves, E., et al. (2019). Functional linkage of gene fusions to cancer cell fitness assessed by pharmacological and CRISPR-Cas9 screening. *Nature Communications*, 10. doi: 10.1038/s41467-019-09940-1.

[235] Pollard, K. S., Salama, S. R., King, B., Kern, A. D., Dreszer, T., et al. (2006). Forces shaping the fastest evolving regions in the human genome. *PLoS Genetics*, 2:e168. doi: 10.1371/journal.pgen.0020168.

[236] Pollard, K. S., Salama, S. R., Lambert, N., Lambot, M. A., Coppens, S., et al. (2006). An RNA gene expressed during cortical development evolved rapidly in humans. *Nature*, 443:167–172. doi: 10.1038/nature05113.

[237] Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., et al. (2017). Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*. doi: 10.1101/201178.

[238] Porubsky, D., Ebert, P., Audano, P. A., Vollger, M. R., Harvey, W. T., et al. (2020). Fully phased human genome assembly without parental data using single-cell strand sequencing and long reads. *Nature Biotechnology*, 39. doi: 10.1038/s41587-020-0719-5.

[239] Prabhakar, S., Noonan, J. P., Pääbo, S., and Rubin, E. M. (2006). Accelerated evolution of conserved noncoding sequences in humans. *Science*, 314:786. doi: 10.1126/science.1130738.

[240] Prado-Martinez, J., Sudmant, P. H., Kidd, J. M., Li, H., Kelley, J. L., et al. (2013). Great ape genetic diversity and population history. *Nature*, 499:471–475. doi: 10.1038/nature12228.

[241] Pritchard, J. K. and Przeworski, M. (2001). Linkage disequilibrium in humans: models and data. *The American Journal of Human Genetics*, 69:1–14. doi: 10.1086/321275.

[242] Pumpernik, D., Oblak, B., and Borštnik, B. (2008). Replication slippage versus point mutation rates in short tandem repeats of the human genome. *Molecular Genetics and Genomics*, 279:53–61. doi: 10.1007/s00438-007-0294-1.

[243] Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, 81:559–575. doi: 10.1086/519795.

[244] Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26:841–842. doi: 10.1093/bioinformatics/btq033.

[245] Raine, K. M., Hinton, J., Butler, A. P., Teague, J. W., Davies, H., et al. (2015). cgpPindel: identifying somatically acquired insertion and deletion events from paired end sequencing. *Current Protocols in Bioinformatics*, 52:1–15. doi: 10.1002/0471250953.bi1507s52.

[246] Ramu, A., Noordam, M. J., Schwartz, R. S., Wuster, A., Hurles, M. E., et al. (2013). DeNovoGear: de novo indel and point mutation discovery and phasing. *Nature Methods*, 10:985–987. doi: 10.1038/nmeth.2611.

[247] Reijns, M. A., Kemp, H., Ding, J., De Procé, S. M., Jackson, A. P., et al. (2015). Lagging-strand replication shapes the mutational landscape of the genome. *Nature*, 518:502–506. doi: 10.1038/nature14183.

[248] Rheinbay, E., Nielsen, M. M., Abascal, F., Wala, J. A., Shapira, O., et al. (2020). Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature*, 578:102–111. doi: 10.1038/s41586-020-1965-x.

[249] Richardson, T. G., Sanderson, E., Elsworth, B., Tilling, K., and Smith, G. D. (5 2020). Use of genetic variation to separate the effects of early and later life adiposity on disease risk: Mendelian randomisation study. *The BMJ*, 369. doi: 10.1136/bmj.m1203.

[250] Richmond, T. J. and Davey, C. A. (2003). The structure of the nucleosome core particle. *Nature*, 423:145–150. doi: 10.1042/bst0140221.

[251] Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Twigg, S. R., et al. (2014). Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature Genetics*, 46:912–918. doi: 10.1038/ng.3036.

[252] Ripke, S., Neale, B. M., Corvin, A., Walters, J. T., Farh, K. H., et al. (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature*, 511:421–427. doi: 10.1038/nature13595.

[253] Ripley, L. (1982). Model for the participation of quasi-palindromic DNA sequences in frameshift mutation. *Proceedings of the National Academy of Sciences of the United States of America*, 79:4128–4132. doi: 10.1073/pnas.79.13.4128.

[254] Rivas, E. and Eddy, S. R. (2001). Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, 2:1–19. doi: 10.1186/1471-2105-2-8.

[255] Rivas, E. and Eddy, S. R. (2015). Parameterizing sequence alignment with an explicit evolutionary model. *BMC Bioinformatics*, 16:406. doi: 10.1186/s12859-015-0832-5.

[256] Rohs, R., Sklenar, H., and Shakked, Z. (2005). Structural and energetic origins of sequence-specific DNA bending: Monte Carlo simulations of papillomavirus E2-DNA binding sites. *Structure*, 13:1499–1509. doi: 10.1016/j.str.2005.07.005.

[257] Rosenthal, A. and Younes, A. (2017). High grade B-cell lymphoma with rearrangements of MYC and BCL2 and/or BCL6: double hit and triple hit lymphomas and double expressing lymphoma. *Blood Reviews*, 31:37–42. doi: 10.1016/j.blre.2016.09.004.

[258] Saini, N., Zhang, Y., Usdin, K., and Lobachev, K. S. (2013). When secondary comes first — the importance of non-canonical DNA structures. *Biochimie*, 95:117–123. doi: 10.1016/j.biochi.2012.10.005.

[259] Sakofsky, C. J. and Malkova, A. (2017). Break induced replication in eukaryotes: mechanisms, functions, and consequences. *Critical Reviews in Biochemistry and Molecular Biology*, 52:395–413. doi: 10.1080/10409238.2017.1314444.

[260] Sale, J. E., Lehmann, A. R., and Woodgate, R. (2012). Y-family DNA polymerases and their role in tolerance of cellular DNA damage. *Nature Reviews Molecular Cell Biology*, 13:141–152. doi: 10.1038/nrm3289.

[261] Sanders, A. D., Falconer, E., Hills, M., Spierings, D. C., and Lansdorp, P. M. (2017). Single-cell template strand sequencing by Strand-seq enables the characterization of individual homologs. *Nature Protocols*, 12:1151–1176. doi: 10.1038/nprot.2017.029.

[262] Sandoval, J. R., Salazar-Granara, A., Acosta, O., Castillo-Herrera, W., Fujita, R., et al. (2013). Tracing the genomic ancestry of Peruvians reveals a major legacy of pre-Columbian ancestors. *Journal of Human Genetics*, 58:627–634. doi: 10.1038/jhg.2013.73.

[263] Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74:5463–5467. doi: 10.1073/pnas.74.12.5463.

[264] Sankoff, D. (1972). Matching sequences under deletion-insertion constraints. *Proceedings of the National Academy of Sciences of the United States of America*, 69:4–6. doi: 10.1073/pnas.69.1.4.

[265] Scally, A. and Durbin, R. (2012). Revising the human mutation rate: implications for understanding human evolution. *Nature Reviews Genetics*, 13:745–753. doi: 10.1038/nrg3295.

[266] Scally, A., Dutheil, J. Y., Hillier, L. W., Jordan, G. E., Goodhead, I., et al. (2012). Insights into hominid evolution from the gorilla genome sequence. *Nature*, 483:169–175. doi: 10.1038/nature10842.

[267] Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S., and Snyder, M. (2012). Linking disease associations with regulatory information in the human genome. *Genome Research*, 22:1748–1759. doi: 10.1101/gr.136127.111.

[268] Schmitt, M. W., Matsumoto, Y., and Loeb, L. A. (2009). High fidelity and lesion bypass capability of human DNA polymerase δ. *Biochimie*, 91:1163–1172. doi: 10.1016/j.biochi.2009.06.007.

[269] Schneider, V. A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H. C., et al. (2017). Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research*, 27:849–864. doi: 10.1101/gr.213611.116.

[270] Schrider, D. R. and Kern, A. D. (2017). Soft sweeps are the dominant mode of adaptation in the human genome. *Molecular Biology and Evolution*, 34:1863–1877. doi: 10.1093/molbev/msx154.

[271] Schrider, D. R., Hourmozdi, J. N., and Hahn, M. W. (2011). Pervasive multinucleotide mutational events in eukaryotes. *Current Biology*, 21:1051–1054. doi: 10.1016/j.cub.2011.05.013.

[272] Sheng, W., Dong, M., Chen, C., Li, Y., Liu, Q., et al. (2017). Musashi2 promotes the development and progression of pancreatic cancer by down-regulating Numb protein. *Oncotarget*, 8:14359–14373. doi: 10.18632/oncotarget.8736.

[273] Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., et al. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, 29:308–311. doi: 10.1093/nar/29.1.308.

[274] Sibbesen, J. A., Maretty, L., and Krogh, A. (2018). Accurate genotyping across variant classes and lengths using variant graphs. *Nature Genetics*, 50:1054–1059. doi: 10.1038/s41588-018-0145-5.

[275] Slager, J., Kjos, M., Attaiech, L., and Veening, J. W. (2014). Antibiotic-induced replication stress triggers bacterial competence by increasing gene dosage near the origin. *Cell*, 157:395–406. doi: 10.1016/j.cell.2014.01.068.

[276] Smit, A., Hubley, R., and Green, P. RepeatMasker Open-4.0 (2013-2015). URL http://www.repeatmasker.org.

[277] Smith, C. E., Llorente, B., and Symington, L. S. (2007). Template switching during break-induced replication. *Nature*, 447:102–105. doi: 10.1038/nature05723.

[278] Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, 147:195–197. doi: https://doi.org/10.1016/0022-2836(81)90087-5.

[279] Stamatoyannopoulos, J. A., Adzhubei, I., Thurman, R. E., Kryukov, G. V., Mirkin, S. M., et al. (2009). Human mutation rate associated with DNA replication timing. *Nature Genetics*, 41:393–395. doi: 10.1038/ng.363.

[280] Stankiewicz, P. and Lupski, J. R. (2010). Structural variation in the human genome and its role in disease. *Annual Review of Medicine*, 61:437–455. doi: 10.1146/annurev-med-100708-204735.

[281] Stone, J. E., Lujan, S. A., and Kunkel, T. A. (2012). DNA polymerase zeta generates clustered mutations during bypass of endogenous DNA lesions in *Saccharomyces cerevisiae*. *Environmental and Molecular Mutagenesis*, 53:777–786. doi: 10.1002/em.

[282] Stoneking, M. and Krause, J. (2011). Learning about human population history from ancient and modern genomes. *Nature Reviews Genetics*, 12:603–614. doi: 10.1038/nrg3029.

[283] Stratton, M. R., Campbell, P. J., and Futreal, P. A. (2009). The cancer genome. *Nature*, 458:719–724. doi: 10.1038/nature07943.

[284] Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., et al. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, 526:75–81. doi: 10.1038/nature15394.

[285] Supek, F. and Lehner, B. (2017). Clustered mutation signatures reveal that error-prone DNA repair targets mutations to active genes. *Cell*, 170:534–547. doi: 10.1016/j.cell.2017.07.003.

[286] Szlachta, K., Thys, R. G., Atkin, N. D., Pierce, L. C., Bekiranov, S., et al. (2018). Alternative DNA secondary structure formation affects RNA polymerase II promoter-proximal pausing in human. *Genome Biology*, 19:1–19. doi: 10.1186/s13059-018-1463-8.

[287] Szlachta, K., Manukyan, A., Raimer, H. M., Singh, S., Salamon, A., et al. (2020). Topoisomerase II contributes to DNA secondary structure-mediated double-stranded breaks. *Nucleic Acids Research*, 48:6654–6671. doi: 10.1093/nar/gkaa483.

[288] Tafer, H., Höner zu Siederdissen, C., Stadler, P. F., Bernhart, S. H., Hofacker, I. L., et al. (2011). ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, 6:26. doi: 10.1186/1748-7188-6-26.

[289] Tajima, F. (1989). Statistical method for testing the neutral hypothesis by DNA polymorphism. *Genetics*, 123:585–595. doi: 10.1093/genetics/123.3.585.

[290] Tang, H., Choudhry, S., Mei, R., Morgan, M., Rodriguez-Cintron, W., et al. (2007). Recent genetic selection in the ancestral admixture of Puerto Ricans. *American Journal of Human Genetics*, 81:626–633. doi: 10.1086/520769.

[291] Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., et al. (2019). COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Research*, 47:D941–D947. doi: 10.1093/nar/gky1015.

[292] Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences*, 17:57–86.

[293] The 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature*, 526:68–74. doi: 10.1038/nature15393.

[294] The ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489:57–74. doi: 10.1038/nature11247.

[295] Tian, D., Wang, Q., Zhang, P., Araki, H., Yang, S., et al. (2008). Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. *Nature*, 455:105–108. doi: 10.1038/nature07175.

[296] Toledo, L. I., Altmeyer, M., Rask, M. B., Lukas, C., Larsen, D. H., et al. (2013). XATR prohibits replication catastrophe by preventing global exhaustion of RPA. *Cell*, 155: 1088. doi: 10.1016/j.cell.2013.10.043.

[297] Treangen, T. J. and Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, 13: 36–46. doi: 10.1038/nrg3117.

[298] Tubbs, A., Sridharan, S., van Wietmarschen, N., Maman, Y., Callen, E., et al. (2018). Dual roles of poly(dA:dT) tracts in replication initiation and fork collapse. *Cell*, 174: 1127–1142. doi: 10.1016/j.cell.2018.07.011.

[299] Valouev, A., Johnson, S. M., Boyd, S. D., Smith, C. L., Fire, A. Z., et al. (2011). Determinants of nucleosome organization in primary human cells. *Nature*, 474:516–522. doi: 10.1038/nature10002.

[300] Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., et al. (2001). The sequence of the human genome. *Science*, 291:1304–1351. doi: 10.1126/science. 1058040.

[301] Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17:261–272. doi: 10.1038/s41592-019-0686-2.

[302] Voineagu, I., Narayanan, V., Lobachev, K. S., and Mirkin, S. M. (2008). Replication stalling at unstable inverted repeats: interplay between DNA hairpins and fork stabilizing proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 105:9936–9941. doi: 10.1073/pnas.0804510105.

[303] Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57:307–333. doi: 10.2307/1912557.

[304] Wala, J. A., Bandopadhayay, P., Greenwald, N. F., O'Rourke, R., Sharpe, T., et al. (2018). SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome Research*, 28:581–591. doi: 10.1101/gr.221028.117.

[305] Walker, C. R., Scally, A., De Maio, N., and Goldman, N. (2021). Short-range template switching in great ape genomes explored using pair hidden Markov models. *PLOS Genetics*, 17:e1009221. doi: 10.1371/journal.pgen.1009221.

[306] Wang, Q., Pierce-Hoffman, E., Cummings, B. B., Alföldi, J., Francioli, L. C., et al. (2020). Landscape of multi-nucleotide variants in 125,748 human exomes and 15,708 genomes. *Nature Communications*, 11:1–13. doi: 10.1038/s41467-019-12438-5.

[307] Wang, W. J., Li, L. Y., and Cui, J. W. (2020). Chromosome structural variation in tumorigenesis: mechanisms of formation and carcinogenesis. *Epigenetics and Chromatin*, 13:1–17. doi: 10.1186/s13072-020-00371-7.

[308] Watson, J. D. and Crick, F. H. C. (1953). Molecular structure of nucleic acids. *Nature*, 171:1966–1967. doi: 10.1038/171737a0.

[309] Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, 276:256–276. doi: 10.1016/0040-5809(75)90020-9.

[310] Wei, L., Liu, L. T., Conroy, J. R., Hu, Q., Conroy, J. M., et al. (2015). MAC: identifying and correcting annotation for multi-nucleotide variations. *BMC Genomics*, 16:1–7. doi: 10.1186/s12864-015-1779-7.

[311] White, H. (1982). Regularity conditions for Cox's test of non-nested hypotheses. *Journal of Econometrics*, 19:301–318. doi: 10.1016/0304-4076(82)90007-0.

[312] Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9:60–62. doi: 10.1214/aoms/1177732360.

[313] Williams, D. A. (1970). Discrimination between regression models to determine the pattern of enzyme synthesis in synchronous cell cultures. *Biometrics*, 26:23–32. doi: 10.2307/2529041.

[314] Wojcik, G. L., Graff, M., Nishimura, K. K., Tao, R., Haessler, J., et al. (2019). Genetic analyses of diverse populations improves discovery for complex traits. *Nature*, 570: 514–518. doi: 10.1038/s41586-019-1310-4.

[315] Xue, A., Wu, Y., Zhu, Z., Zhang, F., Kemper, K. E., et al. (2018). Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *Nature Communications*, 9. doi: 10.1038/s41467-018-04951-w.

[316] Yang, L., Luquette, L. J., Gehlenborg, N., Xi, R., Haseley, P. S., et al. (2013). Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell*, 153: 919–929. doi: 10.1016/j.cell.2013.04.010.

[317] Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution*, 39:306–314. doi: 10.1007/BF00160154.

[318] Yates, A., Beal, K., Keenan, S., McLaren, W., Pignatelli, M., et al. (2015). The Ensembl REST API: Ensembl data for any language. *Bioinformatics*, 31:143–145. doi: 10.1093/bioinformatics/btu613.

[319] Yatsenko, S. A., Brundage, E. K., Roney, E. K., Cheung, S. W., Chinault, A. C., et al. (2009). Molecular mechanisms for subtelomeric rearrangements associated with the 9q34.3 microdeletion syndrome. *Human Molecular Genetics*, 18:1924–1936. doi: 10.1093/hmg/ddp114.

[320] Yeeles, J. T., Poli, J., Marians, K. J., and Pasero, P. (2013). Rescuing stalled or damaged replication forks. *Cold Spring Harbor Perspectives in Biology*, 5:1–16. doi: 10.1101/cshperspect.a012815.

[321] Yi, C. and He, C. (2013). DNA repair by reversal of DNA damage. *Cold Spring Harbor Perspectives in Biology*, 5:1–18. doi: 10.1101/cshperspect.a012575.

[322] Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., et al. (2018). Ensembl 2018. *Nucleic Acids Research*, 46:D754–D761. doi: 10.1093/nar/gkx1098.

[323] Zhang, F. and Lupski, J. R. (2015). Non-coding genetic variants in human disease. *Human Molecular Genetics*, 24:R102–R110. doi: 10.1093/hmg/ddv259.

[324] Zhang, F., Khajavi, M., Connolly, A. M., Towne, C. F., Batish, S. D., et al. (2009). The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. *Nature Genetics*, 41:849–853. doi: 10.1038/ng.399.

[325] Zheng-Bradley, X. and Flicek, P. (2017). Applications of the 1000 Genomes Project resources. *Briefings in Functional Genomics*, 16:163–170. doi: 10.1093/bfgp/elw027.

[326] Zhou, T., Yang, L., Lu, Y., Dror, I., Dantas Machado, A. C., et al. (2013). DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucleic Acids Research*, 41:W56–W62. doi: 10.1093/nar/gkt437.

[327] Zou, X., Morganella, S., Glodzik, D., Davies, H., Li, Y., et al. (2017). Short inverted repeats contribute to localized mutability in human somatic cells. *Nucleic Acids Research*, 45:11213–11221. doi: 10.1093/nar/gkx731.