Probabilistic modelling of somatic alterations in bulk tissue and single cells using repeat DNA

Samer Abujudeh

Supervisor: Prof. Andy G. Lynch
Dr. Edward R. Morrissey

Cancer Research UK Cambridge Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Christ's College

March 2019

To my parents and Maria.

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text.

This dissertation does not exceed the 60,000 word limit prescribed by the Clinical Medicine and Clinical Veterinary Medicine Degree Committee.

Samer Abujudeh
March 2019

Probabilistic modelling of somatic alterations in bulk tissue and single cells using repeat DNA

Samer Abujudeh

Chromosomal instability characterises several cancer types, in which large-scale structural alterations of the genome accumulate at an increased rate. An important class of structural alterations are somatic copy number alterations (SCNAs). SCNAs have been shown to be major drivers of oncogenesis and are associated with prognosis and response to therapies.

Current sequencing and array-based methods that are used to infer SCNAs are cost-prohibitive for widespread clinical use. A low-cost, simple and more clinically applicable method to amplify and sequence more than 10,000 repeat regions across the genome was recently developed, called FAST-SeqS. However, current computational methods do not make effective use of this low-cost assay. This limits its application to clinical medicine and to biomedical research.

In this thesis, I develop conliga; a probabilistic generative model and associated inference algorithms to infer relative copy number from FAST-SeqS data at the amplicon level. I implement this method in R and C++ and provide the software as an open-source tool. By applying conliga and FAST-SeqS to oesophageal adenocarcinoma and related conditions, I show that it has similar performance to QDNAseq applied to low-coverage whole-genome sequencing, which is a more expensive and laborious alternative for SCNA profiling.

I explore several aspects of FAST-SeqS data and show that sample-specific biases can affect SCNA inferences. By extending the conliga model, I demonstrate that these biases can be jointly inferred with SCNA profiles. I validate these extensions by comparing the results to inferences obtained from whole genome sequencing in prostate cancer samples.

I show that the variants present in FAST-SeqS data can be used to infer tumour purity, ploidy and allele-specific copy number. This has potential application in large-scale cancer genome studies to identify samples with sufficient purity before performing high-coverage whole-genome sequencing. Finally, I describe preliminary data showing that the FAST-SeqS protocol can be applied to single cells, enabling further extensions of the conliga model which could lead to the inference of SCNAs in single cells.

Acknowledgements

I am grateful to my supervisors Andy Lynch and Ed Morrissey for their guidance, mentorship and support. Andy's eye for detail, expertise in statistics and genomics and his general advice have been invaluable. I'm thankful for Ed's patience and support in mentoring and teaching me about mathematical modelling and Bayesian statistics and his willingness to make me take regular coffee breaks.

I'd like to thank Simon Tavaré for his guidance, support and for prioritising time to chat whenever I needed his advice. Many thanks to the members of the Tavaré lab, past and present. A special thanks to Sarah Field for her humour, putting up with my experiments and teaching me various wet-lab skills. In addition, I'd like to thank Ed Williams, Meltem Gürel, Malvina Josephidou, Henry Farmery, Alex Tunnicliffe and Ioana Olan for their help and interesting discussions during our time in the PhD office. Thanks to the members of the Morrissey lab in Oxford, Doran Khamis and Qian Cheng. Their kindness and humour have been a great help during the final stages of my PhD.

My work wouldn't have been possible without many members of the Cancer Research UK Cambridge Institute and the surrounding Biomedical Campus. In particular, Paul Coupland and the Genomics Core have helped me to understand the wonderful world of genomics and sequencing. Ann Kaminski's support and positive spirit was invaluable and she was always available to help whenever I needed it. Charlie Massie is a great source of knowledge and his advice was useful when planning experiments.

My friends have made living in Cambridge a great experience and have been a source of support. My friends Luigi, Jen and their little one, Leo Caporaso, have been generous, kind and a source of joy. Ioana Olan's friendship has been invaluable during my time in Cambridge. Since our first days in Wolfson Court, she has always been there to listen and offer her advice. The dinner and game nights with Ioana, Omid Siddiqui, and Eugene Seah were always a great time to relax and helped me put on several kilos. My chats with Eugene were a source of amusement and his friendship has helped to keep my spirits up. Thanks to the Girtonian lads and fellow Girtonian Wolves: Francesco Brizzi, Constantine Capsaskis, Felix Leditzky, Luca Matrà, Jose Reis and Stefan Ritter. Our football sessions, on the rare occasions I wasn't injured, and our trip to Madrid were a nice break from work.

I'd like to thank Juan, Paty, Rodrigo, Pablo, Wicket and Fenri Gomez for their hospitality in Mexico and for their patience, understanding, kindness and positive wishes.

I am incredibly grateful to my parents. I'm not sure what I would do without their endless support, love and encouragement. They have always been there to pick me up in the not-so-good times and a joy to be around in the good times.

Finally, I'd like to give a special thank you to Maria. Without her, I would not be finishing a dissertation. Her patience, love and enduring support have been vital throughout my PhD. During times of stress, she managed to keep my spirits up, made me laugh and kept me on track.

Contents

List of Figures

List of Tables

Nomenclature

Acronyms / Abbreviations

BAF	B allele frequency
BAM	Binary Alignment Map
BFB	Breakage-fusion-bridge
BIR	Break-induced replication
BMI	Body mass index
BO	Barrett’s oesophagus
BWA	Burrows-Wheeler Aligner
cfDNA	Cell-free DNA
CGH	Comparative genomic hybridisation
CI	Credible interval
CIN	Chromosomal instability
CML	Chronic myeloid leukemia
CNA	Copy number alteration
cnLOH	Copy neutral loss of heterozygosity
CNV	Copy number variant
ctDNA	Circulating tumour DNA
DDR	DNA damage response
DNA	Deoxyribonucleic acid

dNTP	Deoxynucleotide triphosphate
DP	Dirichlet process
DSB	Double-stranded break
dsDNA	Double-stranded deoxyribonucleic acid
EMR	Endoscopic mucosal resection
ESD	Endoscopic submucosal dissection
FAST-SeqS	Fast Aneuploidy Screening Test Sequencing System
FFPE	Formalin-fixed and paraffin-embedded
FP	Forward primer
GAC	Gastric adenocarcinoma
GIN	Genetic instability
GORD	Gastro-oesophageal reflux disease
HC	High-coverage
HDP	Hierarchical Dirichlet process
HGD	High-grade dysplasia
HMM	Hidden Markov model
HR	Homologous recombination
HTS	High-throughput sequencing
ICGC	International Cancer Genome Consortium
IMC	Intramucosal adenocarcinoma
IM	Intestinal metaplasia
indel	Insertion or deletion
IQR	Interquartile range
LC	Low-coverage
LCR	Low copy repeat

LGD	Low-grade dysplasia
LINE	Long interspersed element
LOH	Loss of heterozygosity
MAP	Maximum a posteriori
MAPQ	Mapping quality
MCMC	Markov chain Monte Carlo
mFAST-SeqS	Modified Fast Aneuploidy Screening Test Sequencing System
MHC	Major histocompatibility complex
MM	Multiple myeloma
MPS	Massively-parallel sequencing
mRNA	Messenger ribonucleic acid
NCD	Noncommunicable disease
NDBO	Non-dysplastic Barrett's oesophagus
NGS	Next-generation sequencing
NHEJ	Non-homologous end joining
OAC	Oesophageal adenocarcinoma
ORF	Open reading frame
OSCC	Oesophageal squamous cell carcinoma
PAR	Pseudoautosomal region
PARP	Poly-ADP-ribose polymerase
PBMC	Peripheral blood mononuclear cell
PCR1	First round polymerase chain reaction
PCR2	Second round polymerase chain reaction
PCR	Polymerase chain reaction
PEM	Paired-end mapping

RCN	Relative copy number
RD	Read depth
RNA	Ribonucleic acid
RP	Reverse primer
SBS	Sequencing-by-synthesis
SCNA	Somatic copy number alteration
SD	Segmental duplication
SINE	Short interspersed element
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SR	Split read
ssDNA	Single-stranded deoxyribonucleic acid
SSR	Simple sequence repeat
SVM	Support vector machine
SV	Structural variant
TCGA	The Cancer Genome Atlas
TCN	Total copy number
TE	Transposable element
TNE	Transnasal endoscopy
TNM	Tumour-node-metastasis
TSV	Tab-separated values
UDI	Unique dual index
UID	Unique identifier
UMI	Unique molecular identifier
UPS	Universal primer sequence

UTR	Untranslated region
VAF	Variant allele frequency
WALDO	Within-Sample Aneuploidy Detection
WBC	White blood cell
WES	Whole-exome sequencing
WGD	Whole-genome doubling
WGS	Whole-genome sequencing

Chapter 1

Introduction

Cancer is one of the major noncommunicable, or chronic, diseases (NCDs) along with diabetes, cardiovascular and chronic respiratory diseases. Today, NCDs are responsible for the majority of human deaths globally. According to estimates made by the World Health Organisation (WHO) in 2015, cancer was the first or second leading cause of premature death[1] in 91 out of 185 countries. Around the world, cancer incidence and mortality are on the rise; 18.1 million new cases of cancer and 9.6 million deaths due to cancer were estimated in 2018. Currently, cancer is the cause of one in six deaths worldwide and, as we progress through the twenty-first century, cancer is expected to become the leading cause of death in every nation and a major barrier to increasing human life expectancy [1].

Cancer is characterised by uncontrolled proliferation of abnormal cells that can invade surrounding or distant tissues. These abnormal cells can originate from almost any cell type and organ in the body. Hence, cancer is considered a collection of more than a hundred related diseases with varied risk factors and epidemiology [2]. Since the draft sequence of the human genome was published in 2001 [3] and massively parallel high-throughput sequencing was introduced in 2004-2006, cancer research has rapidly advanced [4]. Current technologies allow us to study the molecular characteristics of cancer samples in great detail and at single-cell resolution. Large-scale studies have revealed that the genomes of cancer cells show particular patterns of disruption. These patterns can be used to detect cancer in its early states, determine prognosis, select targeted treatments and monitor progress over time. However, many of the technologies are impractical or currently too expensive for widespread clinical use. The development of practical, low-cost, alternatives is needed to translate these findings into standard clinical practice.

Low-cost molecular assays generate complex and noisy data. This creates a demand for new statistical models and associated inference algorithms to extract relevant biological signal

[1] where premature death is defined as death before 70 years of age

from these complex data. In this thesis, we focus on the development of statistical tools that accurately quantify copy number alterations, seen in many cancer genomes, using data from a low-cost and clinically applicable assay.

1.1 The cell and the genome

The Earth is populated by living things, chemical factories that use matter from their surroundings to survive and proliferate. There are millions of species of animals, plants, fungi, protists, archaea, and bacteria, found on this planet. To the eye, living things are incredibly diverse. Yet, common to all lifeforms is the fundamental unit of life - the cell. The cell is a microscopic, aqueous solution of chemicals, enclosed within a membrane, equipped with the ability to make copies of itself and to divide into two daughter cells. With each division, the parent cell passes down information, specifying characteristics of the cell and the instructions its progeny require to survive and proliferate.

These hereditary instructions are encoded in molecules called Deoxyribonucleic acid (DNA). DNA is a linear polymer chain made up of monomers called nucleotides. Each nucleotide consists of three parts; deoxyribose (a 5-carbon sugar), a phosphate group, and a nitrogenous base. The base may be either adenine, guanine, cytosine, or thymine (commonly referred to by their initials A, G, C and T). Linear sequences or strands of nucleotides are formed by phosphodiester bonds, linking the 5' carbon atom of one nucleotide's sugar molecule to the 3' carbon atom of the next via the phosphate group. Hence, like the words in our alphabet, a DNA molecule has directionality.

Bases from two DNA strands can hybridise together via hydrogen bonds to form base-pairs. There is strict complementarity whereby A can hybridise to T and C can hybridise to G. Double-stranded DNA (dsDNA) is made up of two complementary strands running in opposite directions. Therefore, dsDNA has redundancy of information; knowing the sequence of one strand, one could create the sequence of the other strand. This principle allows cells to replicate their DNA and in doing so, proliferate and pass their DNA to their daughter cells. The hydrogen bonds which bind the strands together are weaker than the phosphodiester bonds, allowing dsDNA to be pulled apart without the sequence of bases being destroyed. Then, a new strand can be created by using one of the strands as a template, using the cell's DNA replication machinery, to synthesise a complementary strand of nucleotides from 5' to 3'.

The complete set of an organism's DNA is called the genome. For *Homo sapiens*, the genome consists of approximately six billion nucleotides (three billion base-pairs). The genome comprises DNA found in the nucleus and the mitochondria of the cell. The nuclear genome consists of 24 linear molecules of DNA called chromosomes; the autosomes (chromosomes

1-22) and the sex chromosomes (chromosomes X and Y). Each chromosome is composed of two chromosome arms, separated by a centromere. The ends of each chromosome are protected by telomeres. In addition to the nuclear genome, the mitochondria of the cell contain the mitochondrial genome which is circular and 16,569 base-pairs in length [5].

Chromosomes contain multiple genes, which are units of information with coding regions (exons), and non-coding regions (introns). It is estimated that the human genome contains approximately 20,000 protein-coding genes [6], in addition to many thousands of genes that encode ribonucleic acid (RNA) molecules whose final product is not a protein but that have important functions in the cell. However, approximately only 1.5% of the genome contains protein-coding genes [7]. The rest is composed of regulatory elements and repeat sequences (which are described in section 1.3) and a large proportion appears to lack biological function.

Chromosomes are tightly condensed into chromatin fibres. The degree of chromatin condensation has been associated with two states: less condensed and more accessible regions are identified as *euchromatin*, while highly condensed and less accessible regions are identified as *heterochromatin*. Heterochromatin condensation levels have shown to be highly dynamic. Changes in heterochromatin condensation levels contribute to the exposure of DNA regions and regulation of gene expression, and are associated to developmental processes and responses to environmental signals [8, 9].

In sexually reproducing multicellular eukaryotes (such as *Homo sapiens*), there are two fundamentally different types of cell; *somatic* cells and *germ* cells. Germ cells are the link between the generations of the organism and are the only cells that can transmit DNA to their offspring through the *germline*. Somatic cells are those that form the body of the organism. In humans, there are hundreds of different types of somatic cells which together make up the various tissues of the human body. Human somatic cells are *diploid*, meaning they contain two sets of the 22 autosomes and two copies of the sex chromosomes. Typically, a female's somatic cells contain two copies of chromosome X while a male's somatic cells contain one copy of chromosome X and one copy of chromosome Y.

1.2 Cancer: diseases of the genome

Within healthy multicellular organisms, individual cells cooperate and organise themselves in complex tissues. Cells act as communities, with each cell deciding when to grow, divide, differentiate and die; communicating via extracellular signals for the survival of the organism. Cancer occurs when the subservient relationship of cell and organism survival is disrupted and cells proliferate and spread at the expense of the organism as a whole.

Even before the discovery of DNA as the molecular basis of inheritance [10] and its double helix structure [11], the association of the genome in the development of cancer was evident.

In the early twentieth century, Theodor Boveri noticed the presence of unusual alterations of the chromosomes visible within dividing cancer cells under the microscope [12]. Since then, evidence has accumulated that genomic alterations characterise and play a causal role in cancer development [2]; carcinogenic chemicals were also found to cause genomic alterations [13], recurrent genomic alterations were seen in patients with specific cancer types (e.g. the Philadelphia translocation in chronic myeloid leukemia) [14, 15] and the introduction of DNA from cancer cells into normal cells could convert them to cancer cells [16, 17]. These findings led to the discovery of *oncogenes* [18, 19], whose alteration or mutation can lead to a gain of function that can help drive cancer development. Meanwhile, studies on hereditary cancers led to the discovery of *tumour suppressor* genes, which help to prevent cancer development and can be inactivated by alterations in cancer cells [20, 2, 7, 21].

All cells in the human body are descendants of the fertilised egg (zygote) and inherit their diploid genome from the maternal and paternal germlines. DNA is an inherently reactive molecule that is susceptible to damage brought about by endogenous (e.g. oxidative stress due to reactive oxygen species) and exogenous agents (e.g. ultraviolet radiation or alkylating agents) or mistakes during DNA replication [22]. DNA damage or errors are usually corrected by the cell's various DNA repair mechanisms. However, some do not get repaired or are repaired incorrectly and are passed to the daughter cells as *mutations* [21, 22]. When mutations occur in the germ cells, they are called *germline mutations* and are inherited by offspring through the germline. When mutations occur in somatic cells they are called *somatic mutations* and as such, cannot be inherited by offspring. When observing mutations in a population, be that a population of genomes from individual humans or genomes from a population of cells, mutations are often referred to as *variants*. These variants are typically defined by their difference to the reference sequence of the human genome [3].

Mutations can be broadly characterised into two types; small-scale or large-scale mutations. Small-scale mutations alter small sections of the genome. The most simple small-scale mutation involve substitutions of a single base and are often referred to as *point mutations* or *single nucleotide variants* (SNVs). A germline point mutation that is present in more than 1% of a population is referred to as single nucleotide polymorphism (SNP). Other small-scale mutations involve the insertion or deletion of short sequences ($<$ 50 base-pairs), collectively referred to as *indels*. Large-scale mutations affect $\geq$ 50 base-pairs of the genome and more greatly influence its structure. As such, they are often referred to as *structural variants* (SVs). SVs can be classified as *deletions*, *duplications*, *insertions* and *rearrangements*. Deletions and duplications alter the number of copies of a sequence of DNA in the genome and are referred to as *copy number alterations* (CNAs). CNAs that are inherited through the germline are often referred to as *copy number variants* (CNVs) while those that arise in somatic cells are referred to as *somatic copy number alterations* (SCNAs). Insertions refer to DNA sequences that are inserted into the genome, such as the integration of viral genomes or transposable elements. Rearrangements include *inversions* and *translocations*. Inversions occur when the orientation of a DNA segment is changed. Translocations involve the rearrangement of segments between non-homologous chromosomes. SVs that result in CNAs are said to be unbalanced, otherwise they are balanced [23–28]. A cell is considered *euploid* if it has an equal number of copies of each of its chromosomes. A cell is considered *aneuploid* if it is not euploid. The detection and quantification of CNAs and aneuploidy is a central theme of this thesis.

Somatic alterations are not limited to the structure or sequence of DNA. DNA methylation and histone modifications regulate the structure of chromatin and can influence the observable characteristics (i.e. phenotype) of the cell. These processes are termed *epigenetic modifications* or *marks*, and are key regulators of tissue-specific gene expression and genome reprogramming during events such as embryogenesis [29]. The disruption of epigenetic modification mechanisms can lead to the development of diseases, including cancer, which can be studied by analysing epigenetic molecular landmarks [30]. The entire set of a cell's epigenetic marks is often referred to as its *epigenome*. A population of cells that are descendant from a single progenitor cell and have identical genomes and epigenomes is called a *clone*.

Cancer development (oncogenesis) is widely accepted to be an example of Darwinian evolution. As somatic cells divide, their genomes accumulate mutations compared to their zygote progenitor. Some of these mutations lead to changes in the cell's phenotype. Cellular phenotypes that confer an increased ability to survive and proliferate are selected. With continual selection and clonal expansion, cells can acquire characteristics (the hallmarks of cancer) that lead to their malignant transformation [31–34]. Cancer cells can have thousands of somatic mutations. Typically, only a handful are positively selected during development and are termed *driver* mutations. Driver mutations are vastly outnumbered by neutral or mildly deleterious mutations, which do not confer a selective advantage to the cancer cell's progenitors and do not contribute to cancer development. These mutations are termed *passenger* mutations [2, 21].

During the course of oncogenesis, cells often experience an increased rate of somatic mutation, which is termed *genetic instability* (GIN). Genetic instability is the most important enabling characteristic that allows cells to acquire the hallmarks of cancer [34]. This state can be achieved by dysfunction of the cell's genomic maintenance and repair mechanisms or increased exposure to mutagenic agents. The cell has surveillance systems which monitor its genome's integrity. If these systems are functioning properly, the detection of a disrupted genome should result in the cell entering senescence or triggering apoptosis. However, if these systems are compromised, the cell will escape these fates and continue to proliferate leading to an accumulation of somatic mutations. *TP53* (and the protein it encodes, p53) plays a central role in this process and as such, is often found to be suppressed, mutated or deleted in cancer cells [34].

Chromosomal instability (CIN) is a specific type of genetic instability, which leads to increased rates of aneuploidy, SVs and CNAs [35]. The mechanisms of CIN (and GIN more generally) are reviewed by Lee *et al.* [35], Aguilera and Garcia-Muse [36] and Janssen and Medema [37]. Briefly, CIN can be the result of a number of mechanisms. One of these is oncogene-induced replication stress. The activation of particular oncogenes (such as *MYC*) or disruption of tumour suppressor genes can lead to the activation of signalling pathways that increase cell proliferation. This can result in the stalling or collapse of replication forks during DNA replication, which if not resolved, cause double-stranded breaks (DSBs) in the genome. Subsequent dysfunctional repair of DSBs by non-homologous end joining (NHEJ) or homologous recombination (HR) can lead to various SVs. Fork collapse can also result in tandem duplication events and complex structural alterations as a result of failures in the break-induced replication (BIR) mechanism. Telomere shortening or breakage is another cause of CIN. When telomeres become short or break, the ends of sister chromatids can fuse together during mitosis. As the sister chromatids segregate during anaphase, this leads to a "bridge" formation. During segregation this bridge breaks and SVs result. This process continues in the replication cycles of daughter cells leading to what is called the breakage-fusion-bridge (BFB) cycle. Mitotic defects can also result in aneuploidy. Whole-genome doubling (WGD) has been shown to occur frequently in oncogenesis, leading to tetraploid cells. It has been shown that WGD provides increased tolerance to CIN, allowing cells to evolve more rapidly [38, 36, 37, 39, 35].

As we have discussed, genetic instability leads to the accumulation of somatic alterations. As cells divide they acquire differences from each other with respect to their DNA content (genotype) and phenotype. The resulting variation observed in a population of cancer cells is termed *intratumoural heterogeneity* [7]. It is this variability along with selective pressures, which drive Darwinian evolution and clonal expansion discussed above. This results in cancers consisting of multiple clones. Somatic alterations that appear in all clones are sometimes referred to as *clonal* alterations, while those that are present in a subset of the clones are referred to as *subclonal* alterations. Intratumoural heterogeneity may be the result of linear evolution, in which increasingly fit clones outcompete their ancestral clones, or branched evolution whereby clones emerge from ancestors and multiple clones propagate in the cancer cell population [40]. Recent evidence has shown that clones can cooperate, suggesting that the assumption of competition between clones may not always be valid [40]. Heterogeneity can manifest spatially, with different sites in the tumour harbouring different clones, and temporally, with the alterations in cancer cells and their relative frequencies varying with time. Intratumoural heterogeneity should not be confused with *intertumoural heterogeneity*, which describes the variation in alterations seen between different patients [7].

The genes that are mutated and causally implicated in the development of cancer are referred to as *cancer genes* [41]. In the current iteration of the Cancer Gene Census Database

(COSMIC, version 87) [42], there are 723 genes that are deemed to play a causal role in cancer development and this number continues to grow with each version. 680 of the 723 have been shown to be somatically altered in cancer while 107 are germline alterations which can predispose patients to the development of cancer[2]. Indeed, inherited germline mutations can increase the probability of developing certain cancer types. Famously, germline mutations which inactivate the *BRCA1* or *BRCA2* genes dramatically increase the risk of developing breast, ovarian and other types of cancer. *BRCA1* and *BRCA2* are involved in DNA repair of DSBs via HR [21, 43]. Cancers that result as a consequence of germline mutations are called *familial* or *hereditary* cancers while those that do not are referred to as *sporadic*. However, the lines that separate the two are becoming increasingly blurred as some sporadic cancers are being shown to have some inherited component [44].

Cancer types can often be dominated by particular types of somatic alterations. For example, serous ovarian cancer is dominated by SVs and CNAs while others such as kidney clear-cell carcinoma are dominated by small-scale somatic mutations. Intriguingly, the results of a study of 3,299 tumours from 12 cancer types by Ciriello *et al.* suggests that cancers tend to be driven by large-scale or small-scale somatic alterations but not both [45].

1.3 Repeat elements

While sequences of genes and known functional elements constitute 5-10% of the human genome, at least 50% of the human genome is comprised of repetitive elements [3, 46]. More recent analyses performed since the initial sequencing of the human genome suggest that more than two thirds of the human genome could be derived from repeat elements [46]. Repeat elements can be generalised into classes: (1) interspersed repeats which are transposon-derived, (2) partially retroposed and inactive copies of genes (processed pseudogenes), (3) simple sequence repeats (SSRs), also called microsatellites and minisatellites, which are contiguous repetitions of short k-mers, e.g. (CA)5 [3], (4) segmental duplications (SDs) also known as low copy repeats (LCRs) are regions of the genome (typically 10-300 kb in length) that have been copied from one region to another, (5) blocks of tandemly repeated sequences, such as those contained within centromeres and telomeres [3].

1.3.1 Transposable elements

By far the most abundant repetitive elements are transposons, also called transposable elements (TEs). Discovered by Barbara McClintock when studying the maize genome [47], they form a class of mobile genetic elements which have been referred to as "jumping

[2] Note that some genes have somatic and germline alterations associated with cancer development

[3] where (CA)5 denotes the sequence comprised of 5 repetitions of CA, i.e. CACACACACA

genes". TEs can be categorised into two types; retrotransposons, which mobilise by an RNA intermediary using a "copy-and-paste" mechanism and DNA transposons which do not have an RNA intermediary and transpose directly as DNA. Together, approximately 45% of the genome can be identified and annotated as transposons with 42% retrotransposons and 3% DNA transposons [3]. Indeed, it is thought that much of the "dark matter" of the genome which neither encodes functional elements or is annotated as repeat elements, are the relics of ancient transposable elements that have mutated and diverged beyond recognition [3, 46].

Retrotransposons

Retrotransposons can be further divided into three types; long interspersed elements (LINEs), short interspersed elements (SINEs) and LTR retrotransposons [3]. Since LINEs are utilised in the work presented in this thesis, we focus on their description. LINEs are incredibly successful members of the genome, accounting for approximately 20-21% of its content [3, 48]. LINEs are classed as autonomous transposons, meaning they encode proteins which are used to mediate their transposition in the genome [49]. The LINE1 (L1) class of LINEs are the only active autonomous elements in humans [50]. In humans, a full-length LINE1 is approximately 6 Kbp. They include a 5' untranslated region (UTR) which includes an internal polymerase II promoter, along with two open reading frames (ORF1 and ORF2) and a 3' UTR which ends with a poly (A) tail [50].

Upon transcription of a full-length LINE1 DNA sequence, the resulting messenger ribonucleic acid (mRNA) molecule moves to the cytoplasm and is translated to produce the proteins ORF1p and ORF2p. These two proteins and the LINE1 mRNA form a complex which can gain access to the cell's nucleus. Once there, ORF2p cleaves one of the DNA strands and the LINE1 mRNA is reverse transcribed from its 3' end into the genome by ORF2p [3, 51]. Typically, reverse transcription fails and does not proceed to the 5' end of the mRNA, such that only a portion of the LINE1 is copied into the genome. Indeed, the average length of LINE1 repeats is approximately 900 bp [3]. For this reason and due to inversions and mutations within the ORFs, more than 99.9% of the LINE1s present in the human genome are incapable of retrotranscription [49].

LINE1s have been evolving in mammalian genomes for more than 150 million years. In what can be considered an evolutionary arms race, host mechanisms have evolved to repress retrotransposition, while LINE1s have relentlessly evolved to avoid repression and continue to retrotranspose throughout the genome [52, 53] (see Goodier for a review [50]). There have been several LINE1 subfamilies that were simultaneously active in ancestral primates. However, in the last 40 million years, there has been a single lineage of primate-specific LINE1s that have succeeded each other [52, 53]. Of the estimated 500,000 LINE1s in the human genome, there are approximately 80-100 retrotransposition-competent human-specific

LINE1 elements (a subset of the L1HS subfamily) currently active in the human genome [54, 55, 49]. Somatic LINE1 insertions have been implicated in various diseases including autoimmunity and cancer (see Beck *et al.* [49], Hancks and Kazazian [56] for reviews).

1.4 High-throughput sequencing

Since the discovery of DNA, the desire to determine the sequence of bases of DNA from biological systems intensified as its crucial role become evident. One of the most popular first-generation DNA sequencing methods was Sanger sequencing, which was developed by Frederick Sanger during the 1970s and later commercialised by Applied Biosystems [57]. This technique was used to complete the first draft human genome in 2003 [58], which cost $3 billion and took 13 years [59, 60].

While hugely successful, Sanger sequencing is low-throughput and time consuming. Since 2004, methods have been developed to read millions of short sequences in parallel. These methods have been termed massively-parallel sequencing (MPS), next-generation sequencing (NGS), or high-throughput sequencing (HTS) and have revolutionised the sequencing of DNA by reducing costs and accelerating turnaround times.

The most popular next-generation sequencing technologies are manufactured by Illumina. Illumina sequencing technologies are based on sequencing-by-synthesis (SBS). Originally, Shankar Balasubramanian and David Klenerman developed the sequencing-by-synthesis approach in the mid 1990s. This led them to found Solexa in 1998 and to launch the Genome Analyzer in 2006. A year later, Illumina acquired Solexa and has continued to develop the technology [61]. In 2014, Illumina announced a new system able to sequence 16 human genomes in three days with an average of 30 reads per region in the genome (30X coverage), costing $1,000 per genome [62].

The methods developed by Illumina include paired-end sequencing, which involves the sequencing of DNA molecules from both ends [63]. It is also possible to sequence DNA from multiple samples in parallel with sample-specific indices (muliplexed sequencing) [63].

The Illumina DNA-sequencing workflow involves three main steps [64, 65]:

1. **Library preparation.** During this step, DNA is cleaved into short fragments with the length dependent on the platform used. Oligonucleotides, which include grafting sequences (allowing the fragments to bind to the flow cell), sample-specific indices and sequencing primers, are ligated to the 5' and 3' ends of the DNA fragments. Library preparation is performed for each sample.

2. **Cluster generation.** Libraries can then be pooled and loaded onto the flow cell. Single-stranded DNA (ssDNA) fragments in the library hybridise to millions of short

oligonucleotides which are attached to the flow cell. Next, the complementary strand of the hybridised strand is synthesised. Fragments are then denatured and the original strand is washed away, leaving complementary single-stranded molecules attached to the flow cell. The strands are then clonally amplified by bridge amplification. In this process, immobilised ssDNA molecules bend over such that their free-end hybridises to another surface oligonucleotide nearby. A complementary strand is then synthesised for each molecule, creating double-stranded bridges. Lastly, dsDNA molecules are denatured, resulting in two ssDNA molecules attached to the flow cell. This process repeats until clusters are formed consisting of single-stranded copies of each original fragment.

3. **Sequencing.** Sequencing-by-synthesis (SBS) is used to sequence the single strands of all the clusters in parallel. The principle of SBS is to synthesise the complementary strand of each single-strand attached to the flow cell. This is performed in cycles in which one nucleotide is synthesised per cycle. Each cycle consists of adding fluorescent-labelled nucleotides to the reaction. The nucleotides have terminators, which allow only one nucleotide to be assembled at a time in each growing complementary strand. After the nucleotide is assembled, lasers are used to excite its fluorescent tag and an image is taken. The signals from the molecules within each cluster are processed to report a base call per cluster together with an associated base quality (Phred) score, which provides an uncertainty measure of the base call. After imaging, the terminators and fluorescent tags are cleaved and washed away to start a new sequencing cycle. The number of cycles depends on the desired read length. The sequence of the sample-specific index in each cluster is also sequenced, which will then be used to determine which sample the reads originated from. If paired-end sequencing is performed, the reverse strand is re-synthesized and sequenced in a similar way. This results in each cluster having a forward and a reverse read.

1.5 Polymerase chain reaction

The polymerase chain reaction (PCR) [66–68], developed by Kary Mullis and colleagues at Cetus Corporation in 1983, is one of the most widely used techniques in biological and medical research [69]. PCR is an *in vitro* technique designed to amplify (make many copies of) a region of DNA. It can be used for a variety of purposes and is often used to prepare amplicon or targeted sequencing libraries, enabling a particular region or regions of the genome to be sequenced by next-generation sequencing technologies.

PCR requires the following components (reagents): DNA template, primers, deoxynucleotide triphosphates (dNTPs)[4], DNA polymerase and a buffer solution. The DNA template (also referred to as "input DNA") is dsDNA which contains the DNA target sequence to be amplified. For example, this might be DNA extracted from a tissue sample or circulating DNA that has been purified. Primers are short oligonucleotide sequences (ssDNA molecules) that are 18 to 30 nucleotides in length. They are manufactured (synthesised) to be complementary to the 3' ends of each strand of the DNA target sequence. They are termed forward and reverse primers. The dNTPs are the building-blocks that are used to synthesise copies of the DNA target sequence. DNA polymerase is an enzyme which *polymerises* new complementary DNA strands from 5' to 3' using ssDNA from the DNA target as a template. Finally, the buffer solution provides a suitable chemical environment for stable and efficient PCR.

PCR consists of multiple cycles of three sequential, temperature dependent steps: template denaturation, primer annealing and primer extension. Figure 1.1 shows an illustration of the first cycle of PCR. In the first step, the PCR mixture is heated to approximately 95 °C which is sufficient to break the hydrogen bonds between the strands of a dsDNA molecule, leaving ssDNA molecules. The mixture is cooled to around 55 °C, allowing the forward and reverse primers to anneal to their target locations on their respective template strand. The mixture is then heated to approximately 72 °C, initiating the extension of the primer-ssDNA complexes from the 3' end of the hybridised primers. The extension is facilitated by DNA polymerase, which assembles dNTPs complementary to the template strands, thereby producing dsDNA molecules (amplicons). The newly synthesised molecules, along with the original DNA starting material, act as templates for the next cycle of PCR. This process repeats in each cycle, with the original DNA starting material and DNA synthesised in all previous cycles, acting as templates for the following cycles. In this way, the DNA target is amplified exponentially. Note that at the end of the second cycle, ssDNA copies of the DNA target sequence (that are also the same length) should have been synthesised.

Typically there are many target DNA molecules in the PCR mixture. In practice, not all of these molecules are amplified in any one PCR cycle. The fraction of molecules that are copied in a cycle is referred to as the PCR efficiency [70, 71]. The PCR efficiency can vary due to a number of factors and from cycle to cycle; stable in the early cycles (exponential phase) and then decays in later cycles (linear and plateau phases) [72, 73].

During PCR, errors can be made during the synthesis of new DNA strands. The fidelity of PCR can vary depending on the DNA polymerase used and other experimental conditions. For example, Taq DNA polymerase I has been reported to have an error rate between 1×10^{-5} and 2×10^{-4} errors per base per doubling event [74]. These errors accumulate because amplicons with errors act as templates for the synthesis of further amplicons in

[4] nucleotides containing triphosphate groups

later cycles. PCR errors are problematic because they appear in sequencing data and can be misclassified as germline or somatic mutations [75, 74] or lead to the inability to determine a read's location in the genome (alignment error).

To mitigate these issues *unique identifiers* (UIDs) [75], which are also commonly referred to as *unique molecular identifiers* (UMIs) [76], can be introduced. By incorporating random nucleotides (denoted by the letter N) to the 5' end of one (or both) of the primer sequences, random sequences are introduced to the amplicons. By doing so, each dsDNA target should have two random nucleotides associated with it after two PCR cycles; one for each strand (assuming a PCR efficiency of 1). Further amplification beyond 2 cycles would result in "swapping out" a UMI sequence with another and each strand will no longer have a unique identifier. In addition to the unique identifier, the Safe-SeqS approach [75] introduces a forward and reverse universal primer sequence (UPS) on the 5' ends of the forward and reverse primers. Hence, these are also introduced to the amplicon sequence. After two cycles of the first round of PCR (PCR1), a new second round PCR (PCR2) is performed in which the UPS sequences are used as primers to synthesise copies of amplicons from PCR1. In this way, the UMI sequence is not "swapped out" and instead, amplicons represent copies of the same unique molecules with the same UMI sequence. Sample indices and Illumina grafting sequences can be introduced to the 5' ends of the second round primers. This means that, after performing several cycles of PCR2, the PCR2 product represents a sequencing library ready for sequencing on Illumina's high-throughput sequencers. This process is represented in Figure 1.2. After sequencing, true germline and somatic variants should be present in all reads sharing the same UMI. Variant sequences appearing at low frequency in reads that share the same UMI are likely to be PCR or sequencing errors and can be ignored. However, PCR errors introduced in the first cycle of PCR1 will be indistinguishable from germline or somatic mutations. In addition, identifying duplicate reads representing the same molecule can help to reduce stochastic noise associated with the process of PCR.

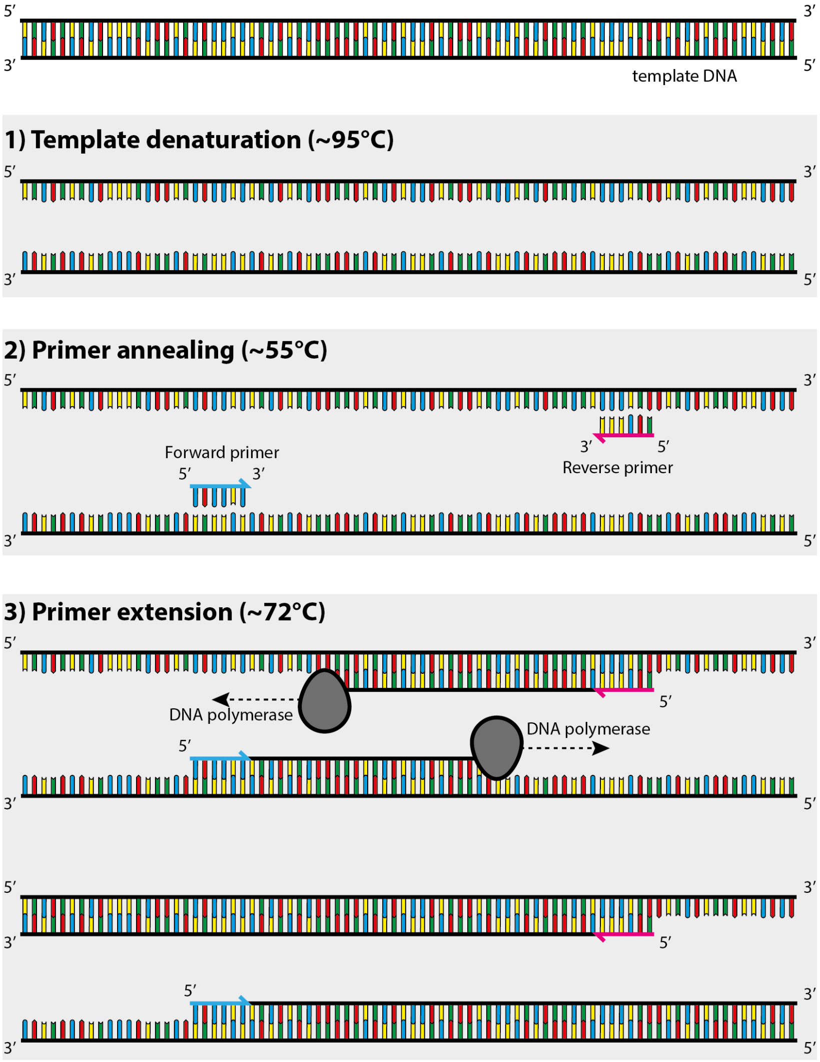

5′
3′
3′
5′
template DNA
1) Template denaturation (~95°C)
2) Primer annealing (~55°C)
Forward primer
Reverse primer
3) Primer extension (~72°C)
DNA polymerase
DNA polymerase

Figure 1.1 An illustration showing the first cycle of a polymerase chain reaction. The top of the figure depicts a molecule of double-stranded DNA which acts as the template DNA for the reaction. The first grey panel depicts the first step of PCR, *template denaturation*, where the hydrogen bonds holding the two complementary strands of DNA are broken by heating the PCR mixture to approximately 95 °C. The next step in the process, *primer annealing*, is shown in the second grey panel. The PCR mixture is cooled to approximately 55 °C (this varies depending on the primer sequences and PCR mixture components) allowing the forward and reverse primer molecules to hybridise to their complementary sequences within the template DNA. The third grey panel shows the final step in the cycle, *primer extension*. The PCR mixture is heated to approximately 72 °C, allowing DNA polymerases to assemble dNTPs to build complementary sequences to the template DNA in the 5’-to-3’ direction from the 3’ end of the primer sequences. This primer extension continues until the DNA polymerase reaches the end of the molecule, until it disassociates stochastically from the template, or until the next cycle of PCR starts with template denaturation.

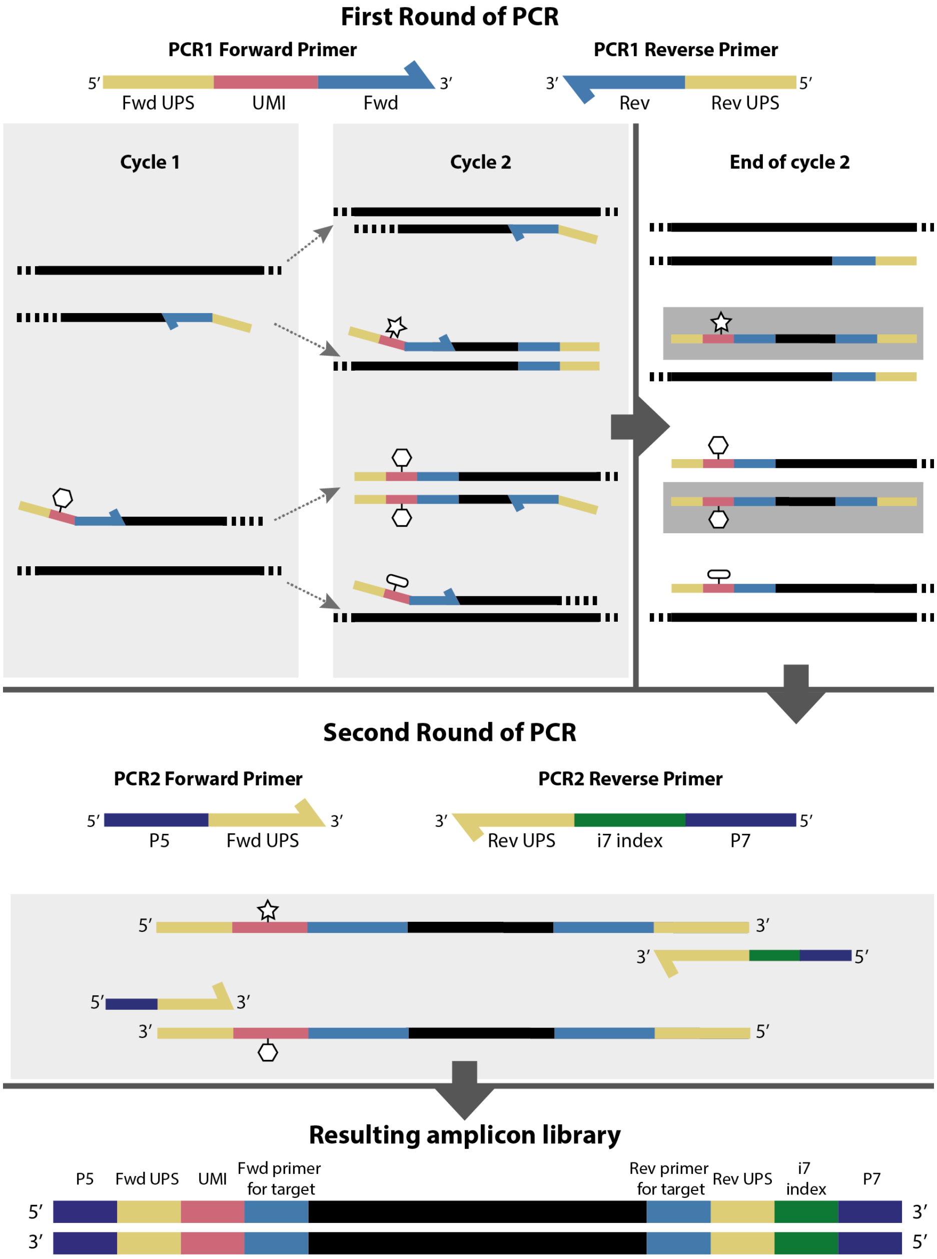

First Round of PCR
PCR1 Forward Primer
5′ 3′
Fwd UPS
UMI
Fwd
PCR1 Reverse Primer
3′ 5′
Rev
Rev UPS
Cycle 1
Cycle 2
End of cycle 2
Second Round of PCR
PCR2 Forward Primer
5′ 3′
P5
Fwd UPS
PCR2 Reverse Primer
3′ 5′
Rev UPS
i7 index
P7
5′ 3′
3′ 5′
5′ 3′
3′ 5′
Resulting amplicon library
P5
Fwd UPS
UMI
Fwd primer for target
Rev primer for target
Rev UPS
i7 index
P7
5′ 3′
3′ 5′

Figure 1.2 An illustration depicting two rounds of PCR and how this results in an amplicon library, following the Safe-SeqS approach. The figure is split into three parts, the top part depicts the first round of PCR (PCR1), the section below outlines the second round of PCR (PCR2) and the section at the bottom shows a representation of the final amplicon library. For the first round of PCR, the composition of the forward and reverse primers is shown at the top, including the nucleotide sequence (Fwd/Rev) that will hybridise to the target, the UMI sequence (UMI) which, in this example, is present only in the forward primer and the universal primer sequence (Fwd UPS/Rev UPS) which comprises the 5' end of the primers. In Cycle 1, the template DNA is denatured, the reverse primer hybridises to its target (top), the forward primer to its target (bottom) and the primers are extended. The dashed lines indicate that the sequences (the DNA template and the newly synthesised molecules) extend indefinitely beyond the target region. Note that a UMI sequence has been introduced during cycle 1 and a shape represents its uniqueness (in this case a hexagon). The four ssDNA molecules (the original DNA template and the newly synthesised molecules) act as template for cycle 2. In cycle 2, the process repeats. By the end of cycle 2, two ssDNA molecules have been synthesised that are not of indefinite length (shown in a darker grey box), each with a unique UMI sequence represented by a star and a hexagon. After a PCR clean-up step (not shown), these two ssDNA molecules act as templates in the second round of PCR (PCR2). Here, a new set of primer sequences are used to hybridise to their respective UPS. This hybridisation step is shown in the grey box. The second round primers contain a sample index (i7 index) and the Illumina grafting sequences (P5 and P7). Several cycles of PCR2 are performed (typically more than 10) resulting in many copies of the template. The final amplicon library is depicted at the bottom of the figure, showing the composition of the amplicon sequencing library.

1.6 Copy number alterations in the context of high-throughput sequencing

The inference of SVs and CNAs from high-throughput (short read) sequencing data has been reviewed by Teo *et al.* [77], Zhao *et al.* [78] and Liu *et al.* [79]. Broadly, SV and CNA inference methods can be grouped into four approaches: read depth (RD), split read (SR), paired-end mapping (PEM) and de novo assembly (AS)[5]. Read depth is the most common approach and is also known as the depth of coverage (DOC) approach. This approach is based on the assumption that, after aligning sequencing reads to the reference genome, the number of reads mapped to a genomic region is proportional to the copy number of the cells in the sample. The split read approach uses reads that fail to align or partially align. By splitting the reads and realigning them, potential breakpoints can be discovered in the genome. The paired-end mapping strategy is similar to that of the split read approach.

[5] Note that some methods use a combinations of these approaches

In the PEM approach, given DNA fragments are generally of a certain length, read pairs are expected to align to the genome with some expected distance from each other (insert length). Read pairs that align with an unexpected genomic distance (discordant read pairs) can indicate that a breakpoint has occurred between the read pairs. Finally, overlapping short reads can be assembled into longer fragments of the genome (contigs). By comparing these contigs to the reference genome SVs can be inferred. SR and PEM approaches are used to infer SVs while the RD approach is more appropriate for CNAs. The assembly approach is rarely used in practice due to the computational demands and the difficulty in assembling the genomes of eukaryotes [78].

Since the RD approach is most appropriate for the data presented in this thesis, we will briefly cover the mathematical concepts behind it.

1.6.1 Copy number of a cancer sample

As we have discussed, cancer samples consist of a mixture of cancer cell clones and normal cells. A normal diploid cell contains two copies of genetic material, one copy is inherited from the maternal germline, while the other is inherited from the paternal germline. In cancer cells, somatic copy number alterations (SCNAs) can accumulate, leading to a varying number of copies of genetic material in the cancer cells.

The mean copy number at locus l, c_l, of a sample of cells can be expressed as:

$$\begin{aligned} c_l &= \underbrace{1\left(1-\sum_{k=1}^{K} p_k\right)}_{\text{A allele from normal cells}} + \underbrace{1\left(1-\sum_{k=1}^{K} p_k\right)}_{\text{B allele from normal cells}} + \underbrace{\sum_{k=1}^{K} n_{akl} p_k}_{\text{A allele from cancer cells}} + \underbrace{\sum_{k=1}^{K} n_{bkl} p_k}_{\text{B allele from cancer cells}} \\ &= \underbrace{2\left(1-\sum_{k=1}^{K} p_k\right)}_{\text{normal cells}} + \underbrace{\sum_{k=1}^{K} p_k\left(n_{akl}+n_{bkl}\right)}_{\text{cancer cells}} \end{aligned} \tag{1.1}$$

where, p_k denotes the proportion of cells from clone k, K denotes the total number of clones, n_{akl} denotes the number of copies of allele A in clone k at locus l, n_{bkl} denotes the number of copies of allele b in clone k at locus l and $\sum_{k=1}^{K} p_k$ is the tumour *purity* or *cellularity*. Note that without any prior parental genotype information, we are unable to distinguish maternal from paternal alleles, and therefore we denote alleles arbitrarily as A and B.

Relative copy number of a cancer sample

A feature of HTS data is that the absolute number of reads does not reflect the absolute number of copies of a locus in the sample. Instead, it reflects a sample of the molecules that were prepared from the DNA sample, i.e. a sample of the sequencing library. Indeed, the absolute numbers of molecules in the sequencing library may not reflect the absolute numbers of copies of DNA in the sample. The library preparation may be composed of many steps. Each step may take a sample of DNA from the previous step. As a result, the alignment counts reflect a relative copy number measure rather than absolute.

The mean relative copy number of the sample of cells, $\hat{r}_l$ at locus l is given by the mean copy number scaled by the average copy number of the sample (referred to as the *ploidy*), D:

$$\hat{r}_l = \frac{c_l}{D} \tag{1.2}$$

where $D = \frac{1}{L}\sum_l c_l$ and L denotes the total number of loci in the genome.

Observed relative copy number from sequencing data

In practice, each step of the library preparation may introduce various biases leading to altered proportions of molecules in the sequencing libraries. Furthermore, the process of sequencing may introduce biases such that some molecules are preferentially sequenced over others. In addition, reads from particular regions of the genome may be more easily aligned than others (referred to as mappability bias). Hence, the read depth can vary across the genome not only due to copy number alterations but also because of these biases. This concept is reflected in figure 1.3.

Biases affect the observed relative copy number, which we shall denote $\hat{c}_l$. It is dependent on the mean copy number *and* the number of reads observed at each locus. This is of particular importance when there is a strong bias, leading to particular regions seeing substantially more alignments than others. Let m_l denote the expected proportion of reads to align to locus l if the sample is normal diploid, where $\sum_i m_i = 1$, and let d denote the amount by which the mean copy number is scaled, such that:

$$\hat{c}_l = \frac{c_l}{d} \tag{1.3}$$

we can show that:

$$d = \sum_i m_i c_i \tag{1.4}$$

We can see that when all m_i are equal, i.e. no bias exists, then $d = D$ and $\hat{c}_l = \hat{r}_l$.

Allele frequency

In the absence of information other than the read depth, it is difficult to estimate the absolute copy number of the cells in the cancer. Germline variants can help with this. The observed

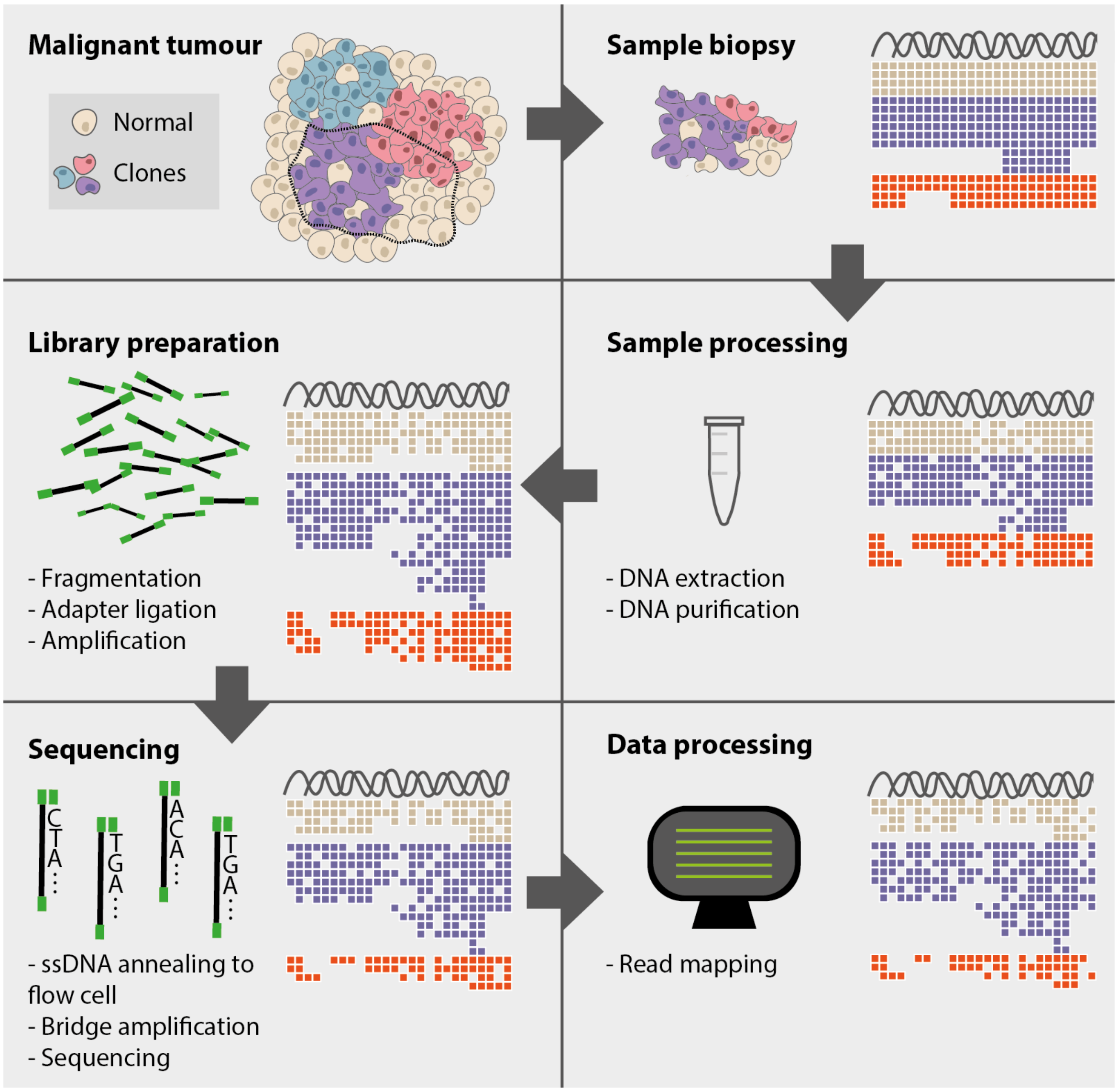


Figure 1.3 An illustration of the process of sequencing a tumour sample. Each segment represents a stage in the process from tissue biopsy through data processing. The illustration includes a representation of a region of the genome. The small blocks represent the quantity of molecules derived from each genomic locus from each clone or normal cell. The purple clone includes a duplication and the red clone includes a deletion. Noise and biases propagate through each step of the process, obscuring the biological signal.

proportion of B alleles in a region can inform how many copies of the B allele are present in comparison to the A allele. The B allele frequency (BAF) at locus l, b_l, is given by:

$$\begin{aligned} b_l &= \frac{1\left(1-\sum_{k=1}^{K} p_k\right)+\sum_{k=1}^{K} n_{bkl}p_k}{1\left(1-\sum_{k=1}^{K} p_k\right)+1\left(1-\sum_{k=1}^{K} p_k\right)+\sum_{k=1}^{K} n_{akl}p_k+\sum_{k=1}^{K} n_{bkl}p_k} \\ &= \frac{1-\sum_{k=1}^{K} p_k+\sum_{k=1}^{K} n_{bkl}p_k}{2-2\sum_{k=1}^{K} p_k+\sum_{k=1}^{K} p_k\left(n_{akl}+n_{bkl}\right)} \end{aligned} \tag{1.5}$$

Note that, b_l is only informative if the allele at locus l is heterozygous, i.e. the sequences of A and B are different. This is because if A and B are the same, i.e. the locus is homozygous, then b_l will be 0 or 1 respectively and will remain 0 or 1 regardless of the number of copies of A and B in the cancer cells.

Expressing the allele-specific copy number in terms of $\hat{c}_l$ and b_l

From equations 1.1 and 1.3 we have:

$$\hat{c}_l = \frac{2\left(1-\sum_{k=1}^{K} p_k\right)+\sum_{k=1}^{K} p_k\left(n_{akl}+n_{bkl}\right)}{d} \tag{1.6}$$

Rearranging equation 1.6 we get:

$$\sum_{k=1}^{K} p_k\left(n_{akl}+n_{bkl}\right) = \hat{c}_l d - 2 + 2\sum_{k=1}^{K} p_k \tag{1.7}$$

Substituting 1.7 into 1.5 we obtain the B allele-specific copy number for clone j, n_{bjl}:

$$\begin{aligned} b_l &= \frac{1-\sum_{k=1}^{K} p_k+\sum_{k=1}^{K} n_{bkl}p_k}{2-2\sum_{k=1}^{K} p_k+\hat{c}_l d-2+2\sum_{k=1}^{K} p_k} \\ &= \frac{1-\sum_{k=1}^{K} p_k+\sum_{k=1}^{K} n_{bkl}p_k}{\hat{c}_l d} \\ n_{bjl} &= \frac{\sum_{k=1}^{K} p_k-1+b_l\hat{c}_l d-\sum_{k=1,k\neq j}^{K} n_{bkl}p_k}{p_j} \end{aligned} \tag{1.8}$$

If we denote the A allele frequency as a_l, and by noting that $a_l = 1 - b_l$, we find that the A allele-specific copy number for clone j is:

$$n_{ajl} = \frac{\sum_{k=1}^{K} p_k-1+\left(1-b_l\right)\hat{c}_l d-\sum_{k=1,k\neq j}^{K} n_{akl}p_k}{p_j} \tag{1.9}$$

Note that, given relative copy number and B allele frequencies, multiple solutions can exist for ploidy, purity and allele-specific copy number.

1.7 Profiling cancer genomes in the clinic

The application of genomics and high-throughput sequencing in the clinic has great promise. While currently limited, these technologies are already being used to diagnose cancer, estimate prognosis and guide treatments [80]. Current successes in targeted treatments include the use of the kinase inhibitor, imatinib, to treat chronic myeloid leukemia (CML) patients that have the *BCR-ABL1* fusion gene caused by a balanced translocation (the Philadelphia chromosome) [81]. Remarkably, this treatment has allowed CML patients to achieve similar life expectancy as those from the general population [82]. Another example is the development of HER2-targeted therapies for breast cancer which have led to increased survival for those with metastatic disease (from less than 2 years to almost 5 year median survival) and the cure of up to 50% of patients for those with early stage HER2-positive breast cancer [83–85, 80]. Other targeted treatment examples include crizotinib for the treatment of *MET*-amplified non-small cell lung cancer (NSCLC) [86] and vemurafenib which targets *BRAF* V600 mutations that occur in approximately 50% of cutaneous melanoma patients [87]. As well as identifying patients for which targeted treatments may benefit, genomic profiling may be useful for identifying those that would not benefit. For example, patients with colorectal cancer (CRC) harbouring *KRAS*, *NRAS* or *BRAF* mutations would not benefit from cetuximab, a EGFR-targeted therapy [88].

Large-scale cancer genome studies are revealing the wide array of genomic alterations that afflict different cancer types. While some recurrent driver alterations are observed within cancer types, some are shared between cancer types. Therapies targeting those alterations in the particular context of one cancer type may be successfully administered to patients with a different cancer type sharing the same genomic alterations. For example, while obtaining regulatory approval for the treatment of *BRCA1/2* associated ovarian cancer, the poly-ADP-ribose polymerase (PARP) inhibitor, olaparib, has been found to have efficacy in treating patients with germline *BRCA1* and *BRCA2* mutations for breast, prostate and pancreas cancer [89]. Results like these encourage a move away from traditional pathology and tissue-of-origin classification to genomic and molecular classification [90].

Intratumoural heterogeneity complicates the targeted treatment of cancers. Genomic alterations present in some cancer cells, that are acquired either before or after cancer development, can confer resistance to therapy. Under the selection pressure of targeted therapy, these resistant clones can come to dominate the tumour landscape. An example of this is the *EGFR* T790M mutation which confers resistance to first-generation EGFR inhibitors used to treat *EGFR* mutant non-small cell lung cancer patients. Hata *et al.* observed cases in which this mutant clone pre-existed before treatment and cases where the cells acquired the mutation during treatment [91]. Profiling the genomic heterogeneity within the tumour population and applying combinations of drugs is likely to help reduce or slow acquired resistance [92, 80].

SCNAs frequently occur in cancer genomes. As we have discussed, SCNA frequencies vary between cancer types. However, the "average" cancer genome has 16% of its genome amplified and 17% deleted [93], compared with the genome of a normal sample which has CNAs affecting approximately 1.2% of the genome [94]. Therefore, just detecting such gross chromosomal abnormalities can be helpful in the diagnosis of cancer (we discuss this in the context of oesophageal adenocarcinoma in chapter 3). Short, or focal, amplifications or deletions often affect oncogenes and tumour suppressor genes respectively. Some SCNAs are associated with prognosis and response to treatment. For example, focal amplifications which increase the expression of the antiapoptotic genes *MCL1* (chromosome 1q21.2) or *BCL2L1* (chromosome 20q11.21) may protect cells from chemotherapy [93]. Other examples include amplifications of a 10 to 15 Mbp region within chromosome 1 that includes a number of genes (including the oncogene, *CKS1B*) that have abnormal expression in multiple myeloma (MM). This alteration is associated with poor prognosis and a less sensitive response to treatment [95–97]. Another example includes the amplification of *CCNE1* that occurs in approximately 25% of patients with ovarian cancer and is associated with poor prognosis. Subsequent CDK2-targeted therapy seems to select tetraploid cells which appear to confer resistance to therapy [98].

The application of genomics testing in the clinic presents a challenge. One of these is the low quantity and quality of tumour material available for testing. Patients usually receive a fine-needle aspiration or small core biopsy. Currently, much of the material collected is used for standard, often histopathology-led, diagnostic evaluation [80]. This leaves little molecular material with which to apply various HTS arrays. Whole-genome sequencing, while comprehensive and falling in cost, is still prohibitively expensive and impractical for most patients. Whole-exome sequencing (WES), can be performed at lower cost, however it is more expensive than simple gene panel assays and requires more input DNA. Gene panels are low-cost and useful for the identification of SNVs and indels. However, they typically include portions of 50-500 genes, representing a small proportion of the whole genome. As such, they have limited use for genome-wide SV and SCNA profiling. Given that some cancer types appear to be driven almost entirely by SVs and SCNAs [45, 99], there is a need to provide low-cost SCNA profiling that could be widely used in the clinic. Low-cost SCNA profiling, perhaps coupled with gene panels for SNV screening, could provide a means of diagnosis, guide treatment decisions and enable the monitoring for resistance to therapy [80, 99].

The non-invasive sampling of cells or DNA is a promising avenue for cancer care. One example is the capture of circulating cell-free DNA (cfDNA) from blood or other fluids (so-called liquid biopsies), which may include DNA originating from cancer cells (circulating tumour DNA, ctDNA). In addition to the isolation of cfDNA, other non-invasive technologies have been developed. For example, cytology-retrieval devices (discussed in the context of oesophageal adenocarcinoma in chapter 3) can be administered more easily than more expensive and more invasive traditional approaches and provide a sample of cells from a particular tissue [100]. Together, these technologies have exciting application for diagnosis, prognosis, treatment selection and disease monitoring [101].

Inherent with non-invasive approaches is the problem that a high proportion of the sample will originate from normal cells. For example, if ctDNA is present in a liquid, it usually represents a small fraction of the total cfDNA; much of the cfDNA is likely to have originated from normal cells. This is particularly true for early stage cancer and the proportion of ctDNA increases with tumour size and stage. However, quantities can vary between individuals and cancer types [102, 101]. The problem of normal contamination is not limited to non-invasive sampling. The proportion of cancer cells present in traditional biopsies is variable with different quantities of normal, stromal and immune cells seen between samples [103]. The proportion of tumour cells in a sample is referred to as the *purity* or *cellularity*, while the proportion of non-tumour cells is sometimes referred to as the *normal contamination*. Low purity samples result in a lower proportion of somatic alterations present in sequencing data (a lower signal). Low tumour purity, coupled with intratumoural heterogeneity (a mixture of signals), makes inferring the somatic alterations present in a sample more challenging. Developing principled statistical tools to distinguish between noise and signal, particularly in the extreme case of low purity and heterogeneity, is critical for these technologies to be useful in the clinical setting.

1.8 Thesis overview

The focus of this thesis is the inference of somatic copy number alterations from data generated by a low-cost and simple HTS sequencing assay (FAST-SeqS [104]) that utilises LINE1 repeats across the genome.

In **Chapter 2**, I describe the FAST-SeqS protocol and review published approaches for the analysis of FAST-SeqS data. By exploring the features of FAST-SeqS data, I motivate and describe the development of a probabilistic generative model (conliga). I go on to describe associated Markov chain Monte Carlo (MCMC) algorithms for the inference of SCNAs from FAST-SeqS data, and methods for summarising the posterior distribution. By application to normal samples derived from male and female donors, I show that conliga infers relative copy number, as expected.

In **Chapter 3**, I apply conliga to oesophageal adenocarcinoma, Barrett's oesophagus and gastric adenocarcinoma samples. I introduce oesophageal adenocarcinoma and the need for improvements to screening and surveillance strategies. I argue that SCNAs may be useful as biomarkers for detection, monitoring and treatment choices. In this chapter, I compare conliga to existing and widely used SCNA inference methods for whole-genome sequencing (WGS) data. I explore conliga's performance in determining the SCNA status of a set of recurrently altered genes in oesophageal adenocarcinoma. In addition, I explore its performance when applied to low purity samples.

In **Chapter 4**, I describe changes to the FAST-SeqS protocol with the aim of increasing the number of samples sequenced per experiment. This chapter describes a new pipeline for processing FAST-SeqS data and technical aspects of the data. Sample-specific biases are discovered when analysing prostate samples which can cause artifacts in the SCNA inferences. I describe methods to correct these biases and show that more realistic SCNA profiles can be inferred.

In **Chapter 5**, I explore the use of variants in FAST-SeqS data and how these could be used to infer purity, ploidy and allele-specific copy number. I show that conliga and FAST-SeqS could potentially be used to screen samples prior to performing WGS in large-scale cancer genome studies. Finally, I explore preliminary single-cell FAST-SeqS (scFAST-SeqS) data and briefly discuss the possibility of extending conliga to infer single cell copy number.

In **Chapter 6**, I summarise the key findings of the thesis, discuss limitations and several avenues for future work.

1.9 Original contribution

During my PhD, I made the following original contributions:

1. I developed a new probabilistic model for FAST-SeqS data and associated algorithms to infer relative copy number at the amplicon level. I implemented these methods in C++ and developed an open-source R package, conliga.

2. I applied conliga to oesophageal adenocarcinoma, gastric adenocarcinoma, Barrett's oesophagus and prostate cancer samples, showing its potential as a clinical tool and for biomedical research applications.

3. I made modifications to the FAST-SeqS primer design, which included UMIs of variable lengths to increase library diversity for higher-quality sequencing. In addition, I introduced unique dual indices to identify index swapping.

4. I showed that variants present in LINE1 sequences, in addition to inferences provided by conliga, can be used to infer the purity, ploidy, allele-specific copy number and evidence of intratumoural heterogeneity.

5. I designed experiments to apply FAST-SeqS to single cells. Using preliminary data, I showed that copy number signal is present in the data, suggesting that methods could be developed for the inference of SCNAs in single cells.

Chapter 2

conliga: a method and tool for inferring relative copy number from FAST-SeqS data

Some of the work presented in this chapter forms a subset of the work described in Abujudeh *et al.* [105] and is shared as a preprint on bioRxiv. In some sections of this chapter, parts of the paper are reproduced. This work was performed in collaboration with the Fitzgerald Laboratory at the MRC Cancer Unit at the University of Cambridge. Sebastian Zeki and colleagues performed the FAST-SeqS protocol and were responsible for processing the clinical samples through to data (FASTQ files). Jamie MJ Weaver conceived the clinical utility of FAST-SeqS for upper gastrointestinal cancers. The rest of the work presented in this chapter is my own.

In this chapter I explore the FAST-SeqS protocol; a simple and low-cost method of library preparation for DNA amplicon sequencing across the genome. I discuss previously published approaches for the analysis of FAST-SeqS data and motivate the creation of a new approach. By exploring FAST-SeqS data, I show that it has the potential to be used as a relative copy number assay. I go on to introduce a probabilistic generative model and tool, conliga[1], which I designed and implemented to infer relative copy number from FAST-SeqS data. Finally, I test the method on male and female samples to detect and quantify changes in the sex chromosomes.

[1] conliga gets its name from "copy number from LINE1 genomic amplicons" and it is also the second-person singular present active imperative of the Latin word conligō, meaning bring together, gather, deduce or infer.

2.1 The FAST-SeqS protocol

The Fast Aneuploidy Screening Test Sequencing System (FAST-SeqS) [104] is a simple protocol involving two rounds of PCR. It results in a library of amplicons ready for sequencing on Illumina platforms. The first round of PCR (PCR1) uses a single primer pair to amplify multiple genomic loci (>10,000) [104, 106–108] dispersed throughout the genome. The PCR1 primers introduce a universal primer sequence (UPS) which is used for the second round of PCR (PCR2). PCR2 uses the UPS to amplify the product produced in PCR1. The PCR2 primers include sample indices and the sequences required for the amplicons to hybridise to the flow cell on an Illumina sequencer. This process is the same as the Safe-SeqS approach described in chapter 1 (Figure 1.2). Figure 2.1 shows the composition of the resulting amplicon library.

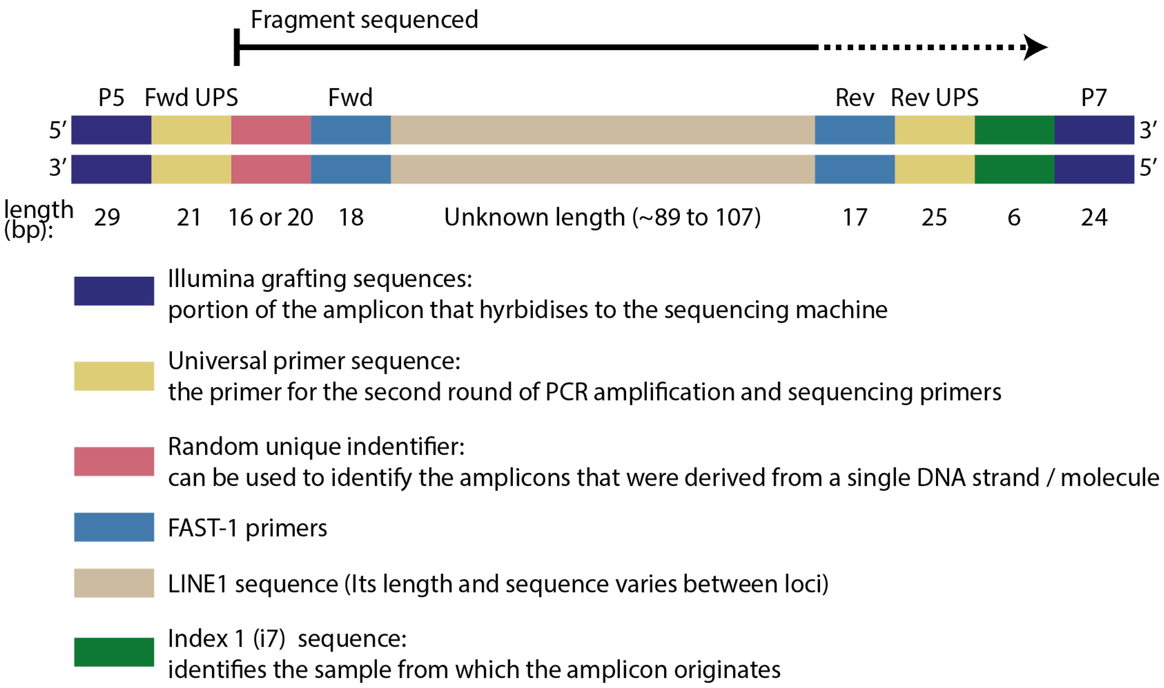

Fragment sequenced
P5
Fwd UPS
Fwd
Rev
Rev UPS
P7
5′
3′
3′
5′
length (bp):
29
21
16 or 20
18
Unknown length (~89 to 107)
17
25
6
24
Illumina grafting sequences:
portion of the amplicon that hyrbidises to the sequencing machine
Universal primer sequence:
the primer for the second round of PCR amplification and sequencing primers
Random unique indentifier:
can be used to identify the amplicons that were derived from a single DNA strand / molecule
FAST-1 primers
LINE1 sequence (Its length and sequence varies between loci)
Index 1 (i7) sequence:
identifies the sample from which the amplicon originates

Figure 2.1 A diagram showing the composition of the FAST-SeqS amplicon sequencing library. Together, the multicoloured bars represent a double-stranded amplicon produced by a FAST-SeqS experiment. Each coloured bar represents a nucleotide sequence associated with different elements of the amplicon. The beige section in the middle of the amplicon represents the target LINE1 sequence. The lengths for each element of the amplicon is listed underneath. The top of the figure indicates the portion of the amplicon that is sequenced when performing 150 bp single-end sequencing. Since the LINE1 sequences vary in length, the resulting sequencing reads include varying portions of the reverse primer (Rev), reverse UPS and sample indices. A dotted line indicates the uncertainty regarding the elements of the amplicon the reads contain.

FAST-SeqS is an attractive assay for use in research and in the clinic because of its simplicity and low-cost. PCR machines are commonly present in research laboratories and hospitals around the world. The protocol is quick and easy to perform, requiring minimal hands-on time and samples can be prepared for sequencing in approximately an hour. The simplicity of the protocol lends itself to automation, meaning the process could be easily scaled to process many samples by the use of liquid handling robots. The simplicity of FAST-SeqS is in contrast to whole-genome sequencing (WGS) library preparation which is laborious and requires approximately half a day to prepare [104, 108]. In addition, FAST-SeqS is simpler than multiplex PCR, in which multiple primer pairs are used to amplify regions within the genome with each pair designed to amplify a unique region. This is because it is technically challenging to design thousands of primer pairs which do not hybridise to each other (primer dimerisation) and have similar chemical properties (e.g. melting temperature). Furthermore, ordering thousands of primer pairs incurs a greater upfront cost compared with buying a single primer pair.

The set of primers used by Kinde *et al.* [104] (which they call the FAST-1 primer pair) amplify segments close to the 3' end of a subset of LINE1s, in particular a set of primate-specific LINE1s. The most commonly observed LINE1 subfamilies amplified by the FAST-1 primers are the most prolific subfamilies in the human genome (L1PA8 - L1PA3) which are estimated to have proliferated in the genome between 40 and 12 Mya [53]. These subfamilies make up approximately 90% of the observed FAST-SeqS alignments [104]. LINE1s accumulate mutations at a neutral rate and therefore older subfamilies have greater sequence divergence than newer ones [53]. The L1PA subfamilies have accumulated a sufficient number of mutations to allow a substantial number of FAST-SeqS amplicons to be aligned uniquely to the genome [104]. Additionally, these older L1PA subfamilies are typically not capable of retrotransposition (which is limited to the more recent human-specific LINE1s) [49].

The FAST-SeqS amplicons vary in length with a bimodal distribution observed by gel electrophoresis with modes around 124 and 142 bp. The distribution of amplicon lengths means that the FAST-1 amplicons can amplify cell-free DNA (cfDNA). cfDNA is approximately 150 bp in length [109] with a prominent mode at 167 bp [110]. As such, FAST-SeqS has promise as an assay that could be used for the detection, classification and longitudinal surveillance of patients with cancer via non-invasive sampling of ctDNA.

2.2 Published methods for FAST-SeqS data

There have been three computational methods published for FAST-SeqS data (at the time of writing). Here, I discuss their purposes, approaches and limitations. A critical examination of the computational methods is presented in section 2.2.4, after I describe the three methods.

2.2.1 Kinde *et al.*, 2012

As the name suggests, the first application of FAST-SeqS was for the detection of aneuploidy (variation in the number of whole chromosomes). Kinde *et al.* demonstrated the use of FAST-SeqS for the detection of trisomy 13 (Patau syndrome), trisomy 18 (Edwards' syndrome) and trisomy 21 (Down's syndrome) in circulating fetal DNA from maternal plasma [104].

Library preparation

33 ng of DNA was used as an input to the first round of PCR. The first round forward primer (FP) included a 20 bp UMI sequence (or 16 bp UMI in some experiments). Table 2.1 summarises the experimental protocol.

Table 2.1 FAST-SeqS experimental protocol summary

Step	Kinde *et al.*	Belic *et al.*
PCR1 input DNA	33 ng	cell lines: 20 ng cfDNA: 0.1-5 ng
UMI	FP: 16 or 20 bp	No
PCR1/PCR2 reaction	50 µl volume 0.5 µM each primer 2U PHSII*	50 µl volume 0.25 µM each primer 2U PHSII*
Bead clean-up	Yes, 1.4× volume	Yes, 1.4× volume
PCR1 thermocycler conditions	98 °C for 120 s 2 cycles of: 98 °C for 10 s 57 °C for 120 s 72 °C for 120 s	98 °C for 120 s 5 cycles of: 98 °C for 10 s 57 °C for 120 s 72 °C for 120 s
PCR2 thermocycler conditions	98 °C for 120 s 13 cycles of: 98 °C for 10 s 65 °C for 15 s 72 °C for 15 s	98 °C for 120 s 15 or 18 cycles of: 98 °C for 10 s 57 °C for 120 s 72 °C for 120 s

* PHSII: Phusion Hot Start II Polymerase

Sequencing, read processing and alignment

Kinde *et al.* sequenced the FAST-SeqS libraries using a HiSeq 2000 machine obtaining 37 bp single-end reads. They kept reads that passed the Illumina chastity filter and contained at least three terminal bases of the forward primer sequence. They masked bases that had a quality score of less than 20 with an ambiguous base *N*. Then they aligned the processed reads to the hg19 version of the human reference genome (including unresolved or unplaced contigs) using Bowtie 0.12.7 [111]. The reads were aligned allowing for up to one mismatch to the reference genome and reads mapping to multiple locations were discarded.

Kinde and colleagues sequenced a total of 49 samples and reported a mean of 20,819,781 reads per sample (range: 3,716,461-42,674,992, median: 19,063,243, interquartile range (IQR): 9,976,099-31,518,407). After alignment, they observed a mean of 7,671,928 aligned reads per sample (range: 1,343,382-16,015,347, median: 7,556,803, IQR: 3,964,677-11,526,042).

Post alignment processing and normalisation

The authors likened FAST-SeqS data to data generated from a microarray; the repeatedly amplified positions in the genome are equivalent to probes and the amplicon read counts are equivalent to microarray intensities. They observed that multiplexing samples on a sequencing lane would produce a different number of reads per sample due to what they believed to be stochastic and experimental variations. Given this observation, they reasoned that it was essential to normalise the data. They believed that expressing the reads as a fraction of the total read count per sample was too simplistic and susceptible to potential systemic biases in the data.

They filtered some of the loci in order to approximate the organisation of microarray data. They did this by including only the loci that appeared in all samples within each sequencing experiment and reported that this involved filtering less than 1% of the data. They also excluded loci mapping to unresolved or unplaced contigs in addition to the sex chromosomes.

Then, they used quantile normalisation to normalise the data in each sequencing experiment. The authors noted that the normalised data were negatively skewed and bimodal. They believed that the bimodality in normalised counts was caused by the preferential amplification of the 124 bp length amplicons with respect to 142 bp amplicons. They decided to filter the loci with the lowest counts by using kernel density estimation to obtain a smooth distribution of normalised counts, determining the inflection point between the two modes, and filtering all loci to the left of the inflection point. Following this, they performed an additional step of quantile normalisation on the filtered normalised counts. I discuss why quantile normalisation is not appropriate for this type of data in section 2.2.4.

Statistical test for evidence of aneuploidy in chromosomes 13, 18 and 21

Given normalised data, including a panel of control euploid samples, they used a simple statistical method to determine evidence of the duplication of chromosomes 13, 18 and 21 in a sample of interest. They determined a z-score for a sample i in a chromosome of interest, chrN, by using the formula:

$$z\text{-score}_{i,\mathrm{chrN}} = \frac{\mathrm{chrN}_i - \mu_{\mathrm{chrN}}}{\mathrm{sd}_{\mathrm{chrN}}} \tag{2.1}$$

here, chrN_i represents the sum of the normalised counts for chrN in sample i. μ_{chrN} and $\mathrm{sd}_{\mathrm{chrN}}$ represent the mean and standard deviation (respectively) of the sum of normalised counts for the euploid control samples in chrN. A chromosome in a sample is assumed to be aneuploid if the z-score is greater than or equal to 3.

2.2.2 Belic *et al.*, 2015

In 2015, Belic *et al.* [106] applied FAST-SeqS to cfDNA with the aim of detecting chromosome arm-level alterations in circulating tumour DNA. They named their method the modified FAST-SeqS (mFAST-SeqS) approach. In addition, they developed a score which determined if the sample had evidence of aneuploidy (genome-wide). mFAST-SeqS was additionally presented as a pre-screening tool whereby, samples with a sufficient aneuploidy score would be processed by their low-coverage WGS method “plasma-Seq” [112] for copy number profiling, while samples with insufficient scores would be screened for SNVs, indels and structural rearrangements (not SCNAs). If the genome-wide score was low, the assumption was that either SCNAs were not present in the sample or that there were insufficient mutant alleles (purity). Belic *et al.* assessed the detection limit of mFAST-SeqS to be 10% purity.

Library preparation

The library preparation steps were largely the same as Kinde *et al.* with a few modifications. They used 20 ng of input DNA for cell line experiments and 0.1 - 5 ng of DNA for cfDNA compared with 33 ng used by Kinde *et al.*. They removed the degenerate bases that were attached to the forward primer meaning that they entirely removed the UMI sequence. Although not mentioned in the main text of their paper, they replaced the custom UPS sequence with the Illumina TruSeq adapter sequence. This makes multiplexing WGS and FAST-SeqS libraries on the same lane more straightforward.

Sequencing, read processing and alignment

Libraries were sequenced on the Illumina MiSeq platform. The paper does not describe how the reads were processed. We can expect that sequencing adapters were removed but we do not know how Phred quality scores were handled; whether bases were trimmed, masked, or reads discarded based on thresholds or if the primer sequences from the reads were removed. Burrows-Wheeler Aligner (BWA) 0.7.4 was used to align reads. It is not stated whether BWA-MEM or BWA backtrack algorithms were used for alignment.

Post alignment processing and normalisation

Alignments with mapping quality (MAPQ) less than 15 were discarded[2]. The remaining alignments were counted with a custom script. The authors note that zero reads aligned to 13p, 14p, 15p, 21p, 22p, and Y and therefore they were removed from the analysis. To normalise the data and account for varying total counts between samples, read counts were divided by the sample total read count. Belic *et al.* do not mention if they filter loci so they may have differing numbers of aligned loci between samples. Post filtering, they reported a mean of 242,190 reads for the control samples (range: 106,205 - 925,843). Note that alignment counts were not presented for the other (non-control) samples.

Statistical tests

The z-score approach of Kinde *et al.* was extended to whole chromosome arms by Belic *et al.*. Rather than limiting the test to evidence of duplication, they called evidence of deletion (z-score $<$ -5) and duplication (z-score $>$ 5). Furthermore, they devised a "genome-wide z-score" to provide an overall measure of aneuploidy for a sample. They define it as the squared sum of all chromosome-specific z-scores. A genome-wide z-score of greater than 5 suggests that a sample shows evidence of aneuploidy. It is not clear from the paper whether this is the squared sum of chromosome arm z-scores or whole chromosome z-scores.

2.2.3 Douville *et al.*, 2018

A method called Within-Sample AneupLoidy DetectiOn (WALDO) has recently been published by Douville *et al.* [107]. The aim of WALDO is similar to mFAST-SeqS; to detect evidence of chromosome arm-level deletions and duplications and to provide a method to call overall aneuploidy in a test sample. As part of their study, they processed and analysed 1,678 formalin-fixed and paraffin-embedded (FFPE) tumour samples, 176 peripheral white blood cells (WBCs) samples from healthy individuals, DNA purified from the plasma of 566

[2] I explore how mapping qualities behave for LINE1 elements in chapter 4

healthy individuals and 982 individuals with cancer. The authors were able to detect allelic imbalance at chromosome arm-level resolution by using SNPs identified in LINE1 regions from the 1,000 Genomes Project [113]. By using matched control samples, they note the ability to detect somatic sequence variants.

Library preparation

Douville *et al.* [107] used the same library preparation steps as Kinde *et al.* [104]. However, they do not state the quantity of input DNA used in the study. For the plasma samples, they obtained DNA from 250 µl of plasma which is likely to yield differing quantities of DNA. The amount of DNA used for the WBC and FFPE tumour samples was not stated.

Sequencing, read processing and alignment

Depending on the experiment, the samples were sequenced using either an Illumina Miseq, HiSeq 2500 or HiSeq 4000. They do not state the length of the reads used or whether they were single or paired-end reads. The authors state that they used the term “read” to refer to “uniquely identified reads” and that, depending on the experiment, each read was sequenced 1-20 times. Presumably, by this they mean that a read’s UMI was used to determine its originating molecule in the experiment. The reads were reduced down to their original molecules and as such, the term “read” was used instead to represent a “unique molecule”. They do not state how they perform this process. The authors used Bowtie2 [114] to align the reads (presumably the UMI reduced reads) to the GRCh37 reference genome but they do not state the version or parameters that were used. For the primary tumour samples that were analysed, they reported a mean of 4,309,367 reads per sample (range: 181,124-36,367,299, median: 2,562,829, IQR: 1,718,905-4,091,714) and a mean of 2,074,172 UMIs per sample (range: 101,210-18,316,187, median: 1,144,478, IQR: 734,981-1,749,316). For plasma samples a mean of 11,537,394 reads were reported (range: 157,297-41,510,757, median: 11,844,636, IQR: 9,571,024-13,972,903) and a mean of 3,516,114 UMIs per plasma sample (range: 101,739-17,565,186, median: 3,516,114, IQR: 2,165,617-5,953,431).

Post alignment processing and normalisation

The WALDO method involves several filtering and normalisation steps to the amplicon read counts. Firstly, WALDO finds the set of seven normal (i.e. euploid) reference samples which are believed to have similar DNA fragment lengths as the test sample from a pool of previously sequenced samples. This was achieved by selecting the seven samples which minimised $D(p, q) = \sqrt{\sum_n (q_n - p_n)^2}$. Here, p_n and q_n represent the fraction of amplicons of length n bp in the test sample and a euploid sample respectively. They believed this was important due to samples with shorter fragment lengths being more likely to have an overrepresented number of shorter FAST-SeqS amplicons. These samples were selected from a panel of 677 WBC and 566 plasma samples. The methods described in the supplementary materials within Douville *et al.* [107] imply that they tested every combination of seven samples from the pool of 1,243 samples. For each set of seven samples tested, the 1% of amplicons with the highest variance of counts across the seven samples were filtered, in addition to amplicons with < 10 reads in one sample but > 50 reads in any of the remaining six euploid samples. Only the amplicons from the autosomes were considered in the analysis.

After choosing the seven reference euploid samples and filtering amplicons, the counts of the remaining amplicons were summed in 500 kb bins. The rationale for binning the genome was that it minimised the stochastic and experimental variability observed in the amplicon counts. Subsequently, the 500 kb bins were normalised for each sample. For each sample, the 500 kb binned counts were normalised by subtracting the mean and dividing by the standard deviation of the 500 kb binned counts in the sample.

The normalised 500 kb bins were then clustered together using the selected seven euploid samples. Each 500 kb bin was assigned to its own primary cluster. Then, for each primary cluster, the normalised reads in the 500 kb bin from the seven euploid samples were compared to the mean normalised read counts (of the seven samples) of all the other 500 kb bins. If the counts were not considered to be significantly different (paired t-test $p > 0.05$ and F-test $p > 0.05$) then the bins were clustered together. Essentially, they clustered bins together with similar normalised read depths across the selected euploid samples.

After clustering the bins, clusters containing fewer than ten 500 kb bins were discarded. They assumed that the normalised read counts (of the test sample) within the remaining clusters were normally distributed and fitted the mean and variance for each cluster. Then they iteratively removed outlying 500 kb bins from the clusters (if $\min(2 \cdot \Phi(\frac{y-\mu_i}{\sigma_i}), 2 \cdot (1 - \Phi(\frac{y_i-\mu_i}{\sigma_i})) < 0.01$, where Φ represents the cumulative distribution function for the standard normal distribution, y represents the summed count for the 500 kb bin being tested, μ_i and σ_i represent the fitted mean and standard deviation respectively for cluster i) and refitted the means and variances until no more outlying 500 kb bins remained.[3] These clusters were then used for determining the status of chromosome arms in the test sample.

Statistical tests

Given the remaining l 500 kb regions on a chromosome arm and their associated cluster memberships, a z-score was calculated to determine the status of the chromosome arm (amplified, deleted, normal). The normalised reads were summed over the chromosome arm.

[3] Note that any outlying 500 kb bins were removed from all the clusters they belonged to. As such, they were removed from all downstream analyses.

Then, z-scores were produced by computing $1 - \Phi((\sum_{i=1}^{l} y_i - \sum_{i=1}^{l} \mu_i)/\sum_{i=1}^{l} \sigma_i)$ where y_i is the summed normalised count for the i^{th} 500 kb bin and μ_i and σ_i represent the fitted mean and standard deviation of the cluster to which the i^{th} 500 kb bin belongs.[4] If the z-score was greater than α then the chromosome arm was determined to be gained and lost if the z-score was less than α, where α was a selected threshold.

This z-score approach detected chromosome arm gains and losses when the tumour fraction was greater than 5-10% of the sample. To attempt to detect samples with several chromosome arm losses or gains at lower tumour fractions, a two-class support vector machine (SVM) was used. The SVM was trained on synthetic samples created by adding and subtracting read counts from 5-25 chromosome arms to 63 selected WBC samples. The counts were added and subtracted to simulate tumour fractions in the range 0.5-10%.

2.2.4 Discussion of previously published approaches

The methods described above focus on the detection of whole chromosome and chromosome arm-level deletions and duplications. Common to all methods is the assumption of normally distributed normalised data. The methods involve the calculation of z-scores coupled with a threshold to decide if sufficient evidence exists for a duplication or deletion. These methods are limited to the resolution of chromosome arms and do not provide information about alterations at the sub-chromosomal arm-level. Moreover, they do not provide quantitative copy number profiles and are limited to, at most, three discrete levels; amplified, deleted, normal.

This level of resolution limits the application of FAST-SeqS data within clinical and basic research. Sub-chromosomal alterations commonly occur in cancer genomes and do so at higher frequencies than whole chromosome arm alterations [93]. In a pan-cancer study of SCNAs in 4,934 primary tumours, Zack *et al.* [115], showed that a median of 11 amplification and 12 deletion events were shorter than chromosome arm-level while there were 3 amplifications and 5 deletions of chromosome arm length or longer. Moreover, focal SCNAs have been found to be major drivers of tumour development and are associated with response to therapies and prognosis [93]. Chromosome arm-level z-scores are a poor measure for focal or sub-chromosomal arm-level alterations. For example, if half of a chromosome arm is deleted and the other half is amplified, the chromosome arm z-score may be close to zero and hence these alterations will not be detected. Furthermore, highly focal alterations, if detected at all, may be falsely characterised as chromosome arm-level alterations. Making the distinction between the focal amplifications of particular oncogenes such as *ERBB2* and

[4] Though not explicitly stated in Douville *et al.* [107], here I assume that the mean and standard deviation were taken from each 500 kb bin's primary cluster

EGFR (which have known drug targets) and a chromosome arm-level alteration may be critically important in classifying and determining optimal therapies for patients.

By failing to provide quantitative measures of copy number alterations, these methods have limited use in the study of temporally or spatially related samples. By providing quantitative measures, we might be able to infer that copies have accumulated or been lost (at least relatively) with respect to time or space in particular regions of interest. This may guide the time-course of patient's clinical treatment and be used to further the understanding of tumour evolution. Also, quantitative estimates of copy number may facilitate the estimation of the sample's tumour purity. This is itself an important piece of information and can inform further sequencing experiments from the same sample (I discuss this further in chapter 5).

mFAST-SeqS and WALDO provide a method to detect overall "aneuploidy". Here, aneuploidy is not used in the strict sense of the definition, i.e. duplications or deletions of whole chromosomes. Instead, they detect (or provide a measure of) evidence that a sample contains chromosome arm-level copy number alterations. WALDO uses chromosome arm-level z-scores as input features to a SVM to call whether a test sample is aneuploid. mFAST-SeqS uses the squared sum of all chromosome-specific z-scores to calculate an overall score. Overall scores such as these are potentially useful for the detection of cancer from low purity tumour samples such as cfDNA for example. However, by using chromosome arm-level scores as the input to these methods and not higher resolution information, they are unlikely to be as sensitive as they could be. Using sub-chromosomal or amplicon-level resolution as an input to discriminate a cancer sample from a normal sample is likely to result in greater sensitivity.

Modelling sequencing data, sampling and technical variation

As I discussed in the introduction, sequencing is fundamentally a process of sampling (or reading) a fixed number of molecules from a library of many more molecules. Because of this, it means we are not measuring absolute numbers of molecules from the sample but rather *relative* differences between molecule counts from a chromosome arm or locus with those from another chromosome arm or locus (within the same sample). This means that when one chromosome arm is duplicated, the reads originating from it will *proportionally* increase compared to those that were not duplicated. To "compensate", the reads from the other chromosome arms will proportionally decrease. In this example, other chromosome arms might be inferred to be deleted when they were not. The z-score approach used by FAST-SeqS, mFAST-SeqS and WALDO will suffer from this issue. Explicitly modelling the fact that we are measuring relative quantities between molecules would be preferable and avoids incorrectly interpreting the observed biological signal.

The process of sequencing, sampling reads and dealing with the resulting count data is well studied. Methods developed for other sequencing assays, for example RNA-seq [116, 117] and DNA-seq [118, 119], utilise count distributions such as the multinomial, binomial or Poisson (and over-dispersed versions thereof) when making inferences from sequencing data. We note that the currently published methods for FAST-SeqS data (presented above) do not use count distributions to account for sampling variation in a principled way. Indeed, Kinde *et al.* [104], Belic *et al.* [106] and Douville *et al.* [107] reported that the number of reads per sample can vary substantially, leading to varying amounts of sampling variation between samples. FAST-SeqS and mFAST-SeqS chose arbitrary z-score thresholds for all tested samples (FAST-SeqS uses > 3 and mFAST-SeqS uses > 5 or < -5). This can lead to increased false positive calls for samples with few reads and greater false negative calls for samples with high read counts. Furthermore, by using fixed z-score thresholds, these methods assume that the amount of technical variation does not vary between samples which could be the case. The WALDO method does not explicitly use read depth information in its approach but by fitting Normal distributions to 500 kb regions with similar read counts, it is indirectly doing so. However, this approach unnecessarily confounds sampling and technical error and, by fitting Normal distributions to each cluster separately, global trends in the data are ignored. If sampling and technical error in FAST-SeqS data are not sufficiently modelled, methods to calculate an overall aneuploidy score will likely amplify these errors, leading to an increased chance of misclassifying normal samples as aneuploid. This is particularly detrimental to methods such as mFAST-SeqS which use the squared sum of the z-scores. It appears that the opportunity exists to make substantial improvements in the modelling of sampling and technical error in FAST-SeqS data.

Normalisation and data processing

All three methods apply normalisation techniques to the data. The normalisation process described by Kinde *et al.* [104] includes two steps of quantile normalisation. Quantile normalisation is used to remove technical variation between samples and is based on the assumption that technical variation is the cause of global variation between samples [120]. This assumption may not hold, particularly in tumour samples with many copy number alterations. By using quantile normalisation, biological variation may be inadvertently removed or distorted. Furthermore, by forcing all samples to have the same distribution and total number of counts, we lose information about the sample's total read counts and hence the variation due to sampling error. This means we are unable to account for this source of variability when making inferences from the data. Kinde *et al.* [104] made the decision to filter a substantial amount of data by filtering below the inflection point of the two modes of the normalised counts (after the first quantile normalisation step). This choice does not appear to be justified, particularly given that they believed these two modes were a reflection of preferential amplification of shorter fragments. If this is the case, they will be removing amplicons that have a length of approximately 142 bp while keeping those around

124 bp. Unless there is a particular reason to filter longer amplicons, this appears to be an unnecessary step.

Belic *et al.* [106] normalise the counts by dividing the counts by the sample's total count. With this approach, biological signal will not be removed. However, this method relies on the assumption that technical variation is consistent between samples and therefore does not require correcting. While it is a simple approach and allows for direct comparisons between samples, it also removes the total count information and therefore the variation due to sampling error is lost.

Douville *et al.* [107] normalise 500 kb bin counts within a sample by subtracting the mean bin count and dividing by the standard deviation. Like the other two methods, this removes the total read information and hence sampling variation information is lost. The authors take an interesting approach to match the test sample to a set of seven previous sequenced normal samples. They believe that the fragment length distribution of the DNA within a sample affects the representation of amplicon lengths that are sequenced; shorter DNA fragment distributions lead to an over-representation of reads from shorter amplicons. They take a data-driven strategy to pick seven samples which they infer to have similar DNA fragment distributions. The downside to this approach is that it requires a large (>1,000) number of previously sequenced samples. These seven samples are used only to define clusters of 500 kb bins with similar normalised read counts. After these clusters are decided, Normal distributions are fitted to each cluster individually using the test sample's 500 kb normalised bin counts. Outlying 500 kb bins within each cluster are iteratively removed and Normal distributions are refitted to the clusters until no more bins are removed. Presumably, the authors consider these bins to be outlying due to noise rather than biological signal. However, it is likely that they are also filtering regions with focal amplifications and deletions, which could be informative for detecting cancer.

Conclusion

Existing methods provide low resolution and categorical inferences from FAST-SeqS data and do not quantify copy number alterations. A method to provide high-resolution, quantified copy number profiles from FAST-SeqS data would be valuable from a clinical and basic biological research perspective and would represent a significant improvement over existing methods. Previous methods have employed different techniques to normalise FAST-SeqS data and in doing so, have potentially distorted or removed biological signal in the data. Developing methods which avoid the use of normalisation procedures would be preferable. Avoiding the use of arbitrary thresholds and instead, appropriately accounting for sampling and technical variation, is critical in order to balance sensitivity with specificity. In section 2.5, I explore the characteristics of FAST-SeqS data. I show how these observations lead to the development of a new method (conliga) that can infer copy number profiles from FAST-SeqS data and addresses the various shortcomings in existing computational methods.

2.3 Generation of FAST-SeqS data

This section is largely the same as the methods section described in Abujudeh *et al.* [105], a preprint manuscript which I authored and is shared on bioRxiv. Since the methods were very similar to those presented in this thesis, I reproduce them here.

Sequencing libraries were prepared using two rounds of PCR, using a similar protocol to previously published methods [104, 106]. Each extracted DNA sample underwent a 50 µl first round PCR reaction with 10 µl 5x Phusion HF Buffer (ThermoFisher Scientific), 1 µl 10 mM dNTP (ThermoFisher Scientific), 5 µl of both the forward and reverse primers (0.5 µM) each (Sigma-Aldrich), 0.5 µl Phusion Hot Start II DNA Polymerase 2U/µl, 5-10 µl DNA template depending on the extracted concentration, and RNAse free water to make the total reaction volume. The cycling conditions for the L1PA7 primers were 98 °C for 120 s followed 2 cycles of 98 °C for 10 s, 57 °C for 120 s, and 72 °C for 120 s.

The second round was also carried out as a 50 µl sample reaction using 20 µl taken from the first round. The rest of the reaction constituents were the same as the first round reaction with the exception of primers (see Supplementary Table 12 of Abujudeh *et al.* [105]), which contained a unique index for each sample. The cycling conditions for the second round reaction were 98 °C for 120 s followed by 13 cycles of 98 °C for 10 s, 65 °C for 15 s, and 72 °C for 15 s for all the primers. After the second round, samples underwent quantification using the 2200 TapeStation (Agilent), Agilent 2100 Bioanalyser (Agilent) and Kapa quantification (KapaBiosystems) prior to submission for sequencing. The samples were then pooled in equimolar concentrations and gel extracted according to manufacturer's instructions (Qiaquick gel extraction kit, Qiagen). Finally the samples were submitted for sequencing on a MiSeq (Illumina) platform. When sequencing molecules on Illumina platforms, it is important to have a similar proportion of all four nucleotides in each cycle of sequencing. The proportion of nucleotides is often referred to as nucleotide diversity or sequencing complexity. It is important to have sequencing libraries with high nucleotide diversity (especially in the first 25 cycles) so that effective template generation is performed and high quality base calls are produced [121]. As such, all samples were sequenced with 20% PhiX control libraries to increase nucleotide diversity and sequencing complexity. Sequencing was performed as 150bp single end. Samples were run with at least three normal controls prepared at the same time and sequenced on the same platform.

2.4 Processing of FAST-SeqS data

We can broadly split the analysis of FAST-SeqS data into two parts. The first step involves processing each sample's FASTQ file into counts at genomic locations. The second step involves statistical modelling and inference from those read counts. The development of this work started in 2015, before the Belic *et al.* [106] (2015) and Douville *et al.* [107] (2018) methods were published. I decided that significant improvements could be made in the statistical modelling and inference step and marginal improvements (at least proportionally) could be made to the read processing step. For this reason, I decided to restrict the number of changes I made to Kinde *et al.*'s strategy for processing FAST-SeqS data and instead focus my energies on modelling the data and statistical inference (which are covered in sections 2.5 and 2.7). Note that I explore the processing of the data in more detail in Chapter 4.

This section is largely the same as the methods section described in Abujudeh *et al.* [105], a preprint manuscript which I authored and is shared on bioRxiv. Since the methods were very similar to those presented in this thesis, I reproduce them here.

Each sequencing run of the Illumina MiSeq platform produced a binary base call (BCL) file which was converted to FASTQ format (using Illumina's bcl2fastq tool). Sequencing reads that failed the Illumina chastity filter were removed. The FASTQ file was demultiplexed into separate FASTQ files corresponding to each sample using the demuxFQ tool [122] with the default settings. The sample barcodes are provided in Supplementary Table 12 of Abujudeh *et al.* [105]. Each sample's FASTQ file was then processed through a custom pipeline which we describe below.

2.4.1 Identifying forward primer position

For each read in the FASTQ file, the position of the forward primer sequence was detected by searching for the sequence with the minimum Hamming distance to the forward primer sequence using a sliding window. Reads with a minimum hamming distance greater than 5 were discarded.

2.4.2 Read trimming

The portion of the reads before and including the forward primer sequence were trimmed. The ends of the reads were also trimmed such that the length of the reads used for downstream analyses were 100 base pairs minus the forward primer length. Any reads shorter than 100 base pairs minus the forward primer length after trimming were discarded.

2.4.3 Quality control

After trimming, reads were discarded if they contained at least one base with a Phred quality score less than 20 and/or contained one or more ambiguous base calls (N).

2.4.4 Obtaining unique sequences and counts per unique sequence

To avoid aligning the same sequence multiple times, only unique read sequences were kept. For each unique read, the number of identical reads were recorded.

2.4.5 Alignment of unique sequences

Unique raw read sequences were aligned with Bowtie 1.0.0 [111] (using the option: -r). Three mismatches were permitted (option: -v3) and reads aligning to multiple locations were discarded (option: -m1). The reads were aligned to GRCh38 (GenBank accession: GCA_000001405.15, no alt analysis set).

2.4.6 Counts and alignments combined

Each sample's unique read alignments and its corresponding unique read counts were combined into a single file consisting of a matrix of counts. The rows corresponded to genomic positions (the union of genomic positions from the alignments in all samples) and columns corresponded to samples. The first three columns of the matrix corresponded to the chromosome, position and strand for the locus, respectively. The matrix of counts used in this analysis can be found in the conliga R package and in Supplementary Table 10 of Abujudeh *et al.* [105].

2.4.7 Selecting loci

Rows of the count matrix corresponding to genomic loci within chromosomes X, Y and within unplaced or unresolved contigs were discarded[5]. For each batch of samples, genomic loci obtaining a zero count in any one of a set of control samples were also discarded. Depending on the sequencing batch we analysed and the controls chosen to filter loci (Supplementary Table 11 of Abujudeh *et al.* [105]), this resulted in approximately 10,000 - 12,000 genomic loci across chromosomes 1 to 22.

[5] though counts from chromosomes X and Y are included in some of the data exploration that follows

2.5 Exploration of FAST-SeqS data and the development of a statistical model

In this section, we focus on a set of 13 samples which were obtained from normal adjacent tissue in the oesophagus of patients with oesophageal adenocarcinoma (OAC). These samples were taken from 5 female donors and 8 male donors.

2.5.1 LINE1 subfamilies observed in FAST-SeqS data

I used RepeatMasker annotation [48] (hg38, Dec 2013, RepeatMasker open-4.0.5, Repeat Library 20140131) and the GenomicRanges package [123] to obtain the overlaps of the FAST-SeqS alignments and the repeat annotation of GRCh38. Of the 31,653 unique alignments observed within chromosomes 1-22, X and Y across all samples, nine did not overlap with annotated repeats, 31,644 were found to overlap with annotated repeats and of these 26,992 were unique repeat elements. The reason multiple alignments can overlap with the same repeat element is that a few reads occasionally align a few bases upstream or downstream to the majority of the other alignments to the same repeat. This is likely to be because PCR or sequencing errors are introduced in a few of the molecules, leading to altered alignment positions. Indeed, the sequence following the FAST-1 forward primer is typically a short dinucleotide repeat (CA)3 which is susceptible to error in PCR and sequencing [124]. Of the 26,992 repeat elements, 26,958 originated from primate-specific LINE1 elements (the L1PA subfamilies). Figure 2.2A shows the distribution of primate-specific LINE1 elements that received an aligned read (count) in at least one of the 13 control samples. This distribution is similar to that shown in Figure 1C of Kinde *et al.* [104] and suggests that we observe alignments at similar genomic locations. By normalising by the number of elements of each subfamily in the genome, we see that the observed elements are biased towards L1PA3-L1PA11, which propagated through the genomes of ancestral primates approximately 12.5-53.3 million years ago [53] (Figure 2.2B).

Since the reads are single-end, we cannot precisely determine the length of the amplicons from the reads alone. Given the alignment position, I searched for the most likely reverse primer sequence in the reference genome. This was achieved by finding the sequence with the minimum Hamming distance (maximum of 5 mismatches) to the reverse primer up to 300 bp downstream of the alignment. For this analysis, I filtered the loci such that only those with a non-zero count in all 13 samples were considered. After filtering, 12,834 amplicons remained and the lengths of 12,364 amplicons could be inferred. The remaining amplicons either had multiple matches to the primer sequence (424) or had zero matches (46) and so were discarded for this analysis.

Kinde *et al.*. showed that the distribution of FAST-SeqS amplicon lengths is primarily bimodal, with modes observed at 124 and 142 bp (Kinde *et al.* Figure 1A). In Figure 2.2C, I show that this multimodality arises due to length differences between the L1PA subfamilies. The lengths of the targeted sequence within the L1 3' UTR in subfamilies older than L1PA8 and L1PA8A, tend to be more variable. The L1PA8A subfamily appears to have acquired a deletion and has lengths distributed around 103 bp, while the L1PA8 subfamily (which co-occurred with L1PA8A for some time [53]) has lengths distributed around 142 bp. L1PA7, L1PA6 and L1PA5 have a similar length distribution around 124 bp. An additional base is inserted sometime between the expansion of the L1PA5 and L1PA4 subfamilies, resulting in a typical length of 125 bp for L1PA4 and L1PA3 subfamilies. Figure 2.3 shows the alignment of the consensus sequences of the L1PA subfamilies and further illustrates this point. I used the Bioconductor package DECIPHER [125] to align the 3 prime consensus sequences obtained from the Repbase-derived RepeatMasker libraries (version 20181026) [48].

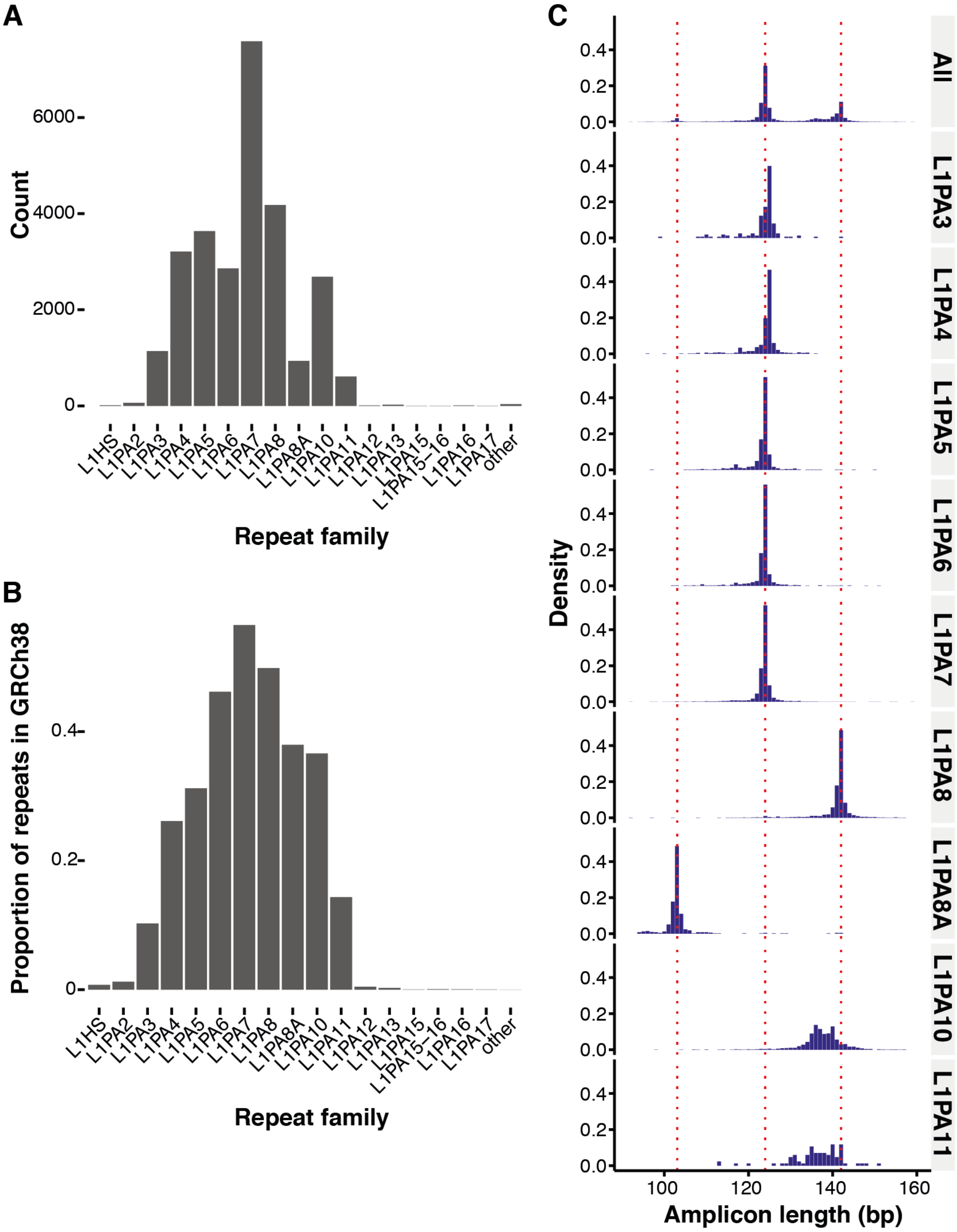


Figure 2.2 Primate-specific LINE1 subfamilies in FAST-SeqS data. A: Histogram showing the number of unique overlaps of FAST-SeqS loci with primate-specific LINE1 subfamilies and other repeats across 13 normal samples (unfiltered). The annotation was obtained from RepeatMasker for the human genome (GRCh38). B: Histogram showing the proportion of primate-specific LINE1s within the genome that overlap with FAST-SeqS loci. C: Top: Histogram of estimated amplicon length for filtered FAST-SeqS loci (filtered such that all 13 normal samples had a non-zero count). Bottom 9 histograms: Estimated length by primate-specific LINE1 subfamily (L1PA3-L1PA11).

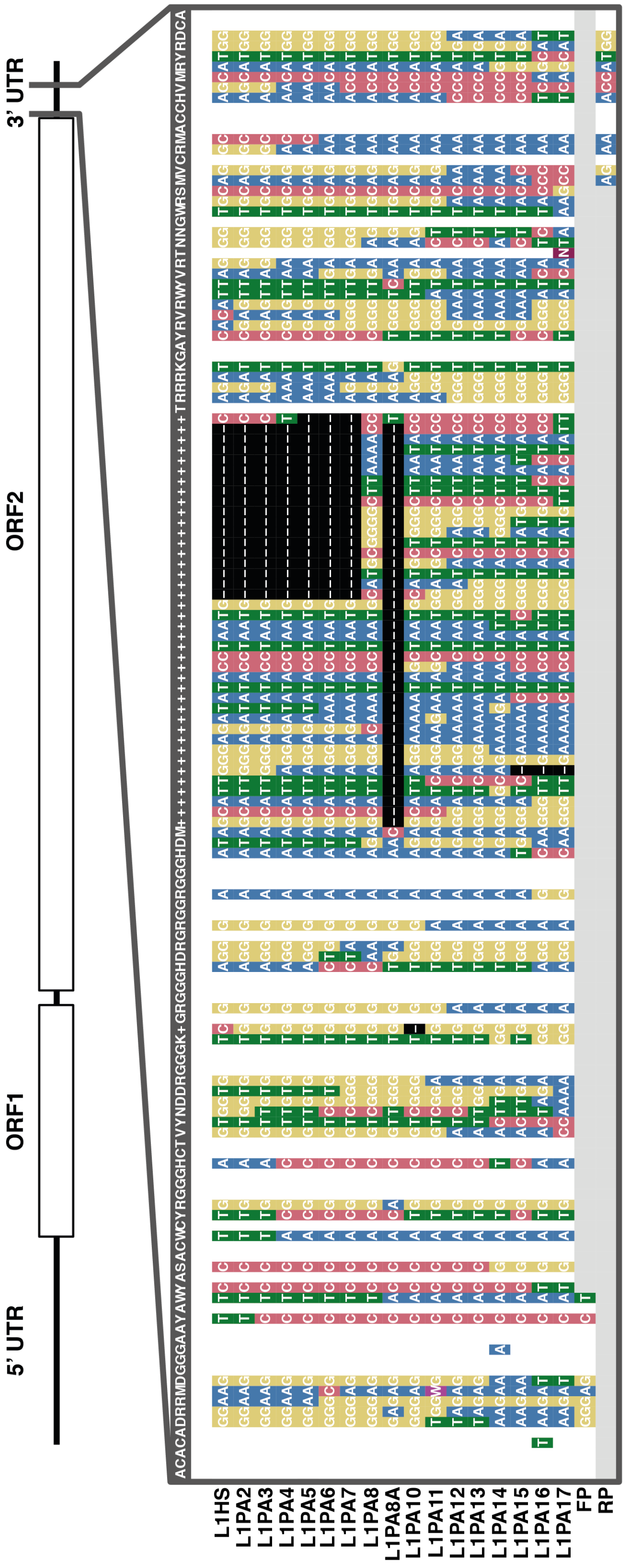


Figure 2.3 Sequence alignment of primate-specific LINE1 consensus sequences obtained from RepeatMasker annotation. Top: Annotation for L1HS consensus sequence showing the region amplified by the FAST-1 primers within the 3' UTR. The grey bar at the top shows the consensus sequence of all the primate-specific LINE1 subfamilies as called by DECIPHER [125] (using the IUPAC system). Within each subfamily consensus sequence, white spaces indicate a match to the primate-specific LINE1 consensus sequence. Dashed lines indicate a deletion to the primate-specific LINE1 consensus sequence. The two sequences at the bottom indicate the alignment of the FAST-1 forward primer (FP) and the FAST-1 reverse primer (RP).

2.5.2 Distribution of loci counts with fixed copy number

To begin with, let us explore and understand the behaviour of FAST-SeqS data in the absence of copy number alterations. The 13 normal samples are assumed to contain only diploid cells and as such, we expect equal copy numbers for every locus within each sample. Since the normal samples were obtained from male and female donors, we focus on the autosomes (chromosomes 1-22) and remove the sex chromosomes (chromosomes X and Y) and loci that are assigned to unresolved or unplaced contigs for this exploratory exercise. In addition, to limit noise introduced by alignment error, we only consider loci that had a non-zero count across all 13 control samples. Figure 2.4A shows the raw count proportions for a control sample plotted across the genome (chromosomes 1 to 22) and Figure 2.4B shows the distribution of proportions within the sample. We observe considerable variation between the loci proportions within the sample. While the read depth varies considerably, a similar range of read depths can be observed across the genome with similar mean proportions in each chromosome (figure 2.4C).

By comparing count proportions between control samples, we see that the loci proportions are similar between samples. This shows that a substantial amount of the variability in loci counts can be attributed to a systematic technical bias in the production of loci counts (from the sample preparation and PCR, through to the sequencing and alignment). This bias is likely to be the result of multiplicative biases introduced at each step in the protocol. I suspect that a large proportion of it is explained by differential amplification efficiencies between loci. Other factors such as differential mappability are also likely to play a part. We explore these biases in more detail in Chapter 4.

By computing the mean and variance for each locus' counts across the control samples, we can explore how the variability of the data relates to the mean count. If the variability observed could be explained by sampling variation only then we'd expect the variance to be approximately equal to the mean (Poisson noise). In Figure 2.4E we see that the variance is greater than the mean which shows that the there is additional variance beyond sampling error, i.e. the data are over-dispersed.

Using these important observations of the data, we can begin to explore statistical models which can describe its generation.

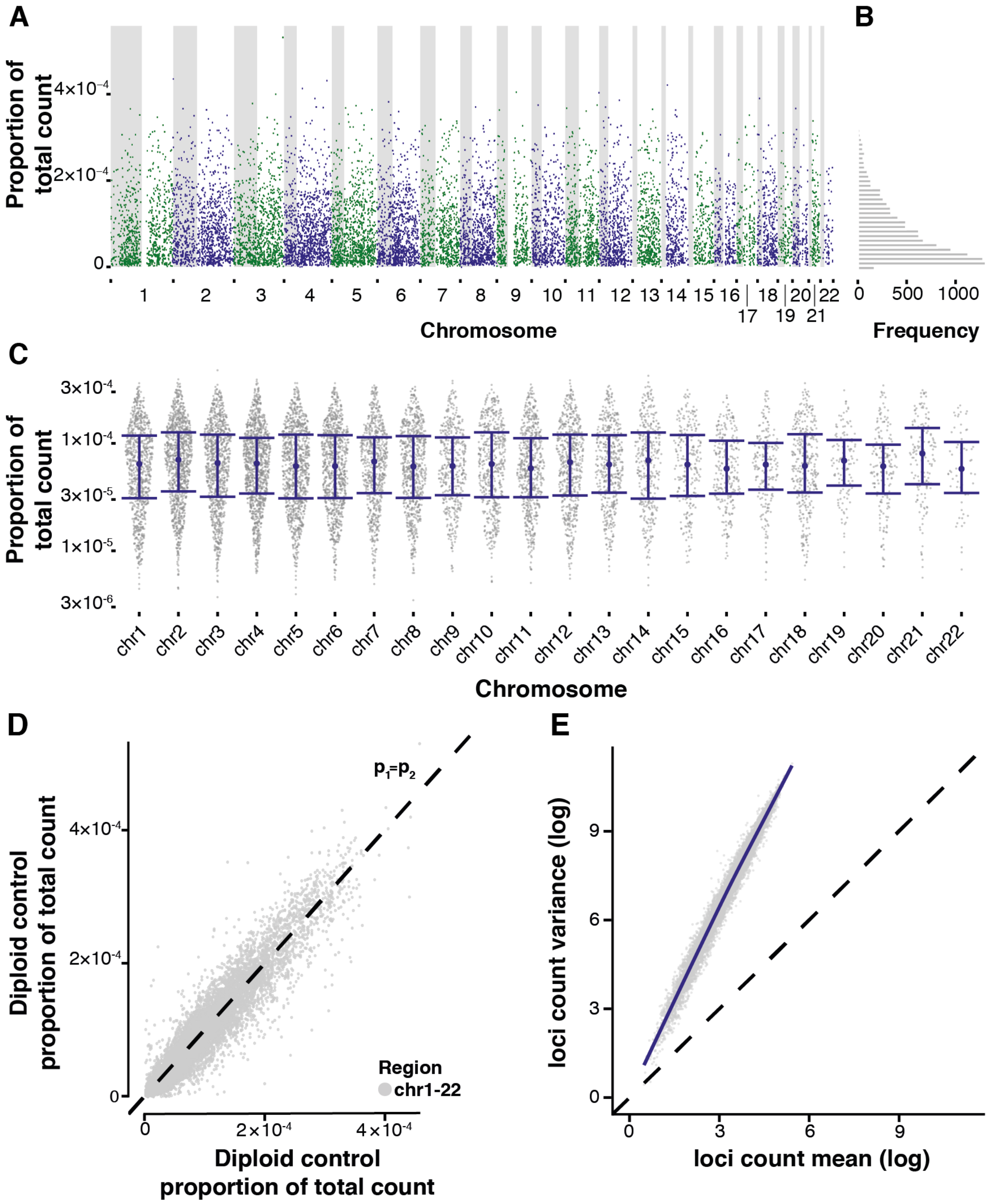


Figure 2.4 Exploration of count data within the autosomes of normal samples. A: Loci count proportions of an example sample plotted across chromosomes 1-22. B: The distribution of loci count proportions from the same sample. C: Sina plot showing the proportion of counts observed across 13 samples. Each grey point represents a locus, blue points represent the median proportion within the chromosome and blue bars indicate the interquartile range. D: Proportion of counts of one normal sample plotted against another normal sample. E: Log mean-log variance relationship in which each point represents a locus for which the mean and variance were calculated across 13 normal samples. The blue line represents a LOESS fit and the dashed black line has an intercept of 0 and slope of 1 (representing Poisson noise).

2.5.3 Modelling loci counts for samples with equal copy number

The key observations of the data which inform the modelling of the counts are:

1. Loci count proportions vary considerably within samples. We know that at least some of this variation will be introduced by the process of sampling molecules from a sequencing library (sequencing).
2. Loci proportions between control samples are similar. This shows that large part of this variation stems from a bias (or a series of biases) in the protocol, sequencing and downstream processing. These biases appear to be consistent between samples.
3. The data are over-dispersed. This means that the bias (or series of biases) together with the sampling variation introduced by sequencing cannot explain all of the variation in the data.

We wish to model the counts for a sample at a set of L loci, $1, \ldots, L$ using a probabilistic approach. I assume that the loci bias is constant between samples; that there are no sample-specific biases which would alter the observed count proportions. For a locus, l, the mean proportion of reads can be denoted by m_l, where $\sum_l m_l = 1$. We can alternatively view m_l as the probability that a mappable read aligned to locus l. In order to account for the sampling process, we might imagine that the counts, $\boldsymbol{x}_j$, for a sample, j, could be drawn from a multinomial:

$$\boldsymbol{x}_j \mid n_j, \boldsymbol{m} \sim \text{Multinomial}(n_j, \boldsymbol{m}) \tag{2.2}$$

where n_j denotes the total number of reads that aligned to loci $1, \ldots, L$ in sample j and $\boldsymbol{m} = \{m_1, \ldots, m_L\}$.

This model might be sufficient if we were not dealing with over-dispersed counts. In order to account for over-dispersion, we might imagine that due to stochastic variations in the protocol and that the protocol involves a series of sampling processes associated with each step, that the probability of aligning a read to a locus can vary between samples. As such, we could model the loci counts as follows:

$$\begin{aligned} \boldsymbol{\theta}_j \mid s_j, \boldsymbol{m} &\sim \text{Dirichlet}(s_j \boldsymbol{m}) \\ \boldsymbol{x}_j \mid n_j, \boldsymbol{\theta}_j &\sim \text{Multinomial}(n_j, \boldsymbol{\theta}_j) \end{aligned} \tag{2.3}$$

where $\boldsymbol{\theta}_j = \{\theta_{1,j}, \ldots, \theta_{L,j}\}$ and $\theta_{l,j}$ represents the probability of a mappable read aligning to locus l in sample j, s_j represents a sample-specific parameter which inversely controls the over-dispersion of the counts and $s_j > 0$. Here, we are explicitly stating that the amount of over-dispersion can vary between samples. We might intuitively imagine that this is the case, due to varying quantities of input DNA, varying DNA quality, or a combination of these factors and others. It can be shown for arbitrarily large s_j, $\boldsymbol{\theta}$ converges in probability to $\boldsymbol{m}$ and, as s_j goes to infinity, the multinomial model in equation 2.2 is recovered.

The expected value of $\theta_{l,j}$ is:

$$\mathbb{E}\left[\theta_{l,j}\right] = m_l \tag{2.4}$$

and the expected value of $x_{l,j}$ is:

$$\mathbb{E}\left[x_{l,j}\right] = n_j m_l \tag{2.5}$$

The variance of $\theta_{l,j}$ is:

$$\mathrm{Var}\left[\theta_{l,j}\right] = \frac{m_l(1-m_l)}{s_j+1} \tag{2.6}$$

and the covariance of $\theta_{l,j}$ and $\theta_{l',j}$ is:

$$\mathrm{Cov}\left[\theta_{l,j}, \theta_{l',j}\right] = \frac{-m_l m_{l'}}{s_j+1} \tag{2.7}$$

The variance of $x_{l,j}$ is:

$$\mathrm{Var}\left[x_{l,j}\right] = n_j m_l(1-m_l)\left(\frac{n_j+s_j}{1+s_j}\right) \tag{2.8}$$

and the covariance of $x_{l,j}$ and $x_{l',j}$ is:

$$\mathrm{Cov}\left[x_{l,j}, x_{l',j}\right] = -n_j m_l m_{l'}\left(\frac{n_j+s_j}{1+s_j}\right) \tag{2.9}$$

While $\boldsymbol{\theta}$ helps to understand the conceptual idea of over-dispersion in sample counts, we are not particularly interested in the values it takes. Since the Dirichlet distribution is a conjugate distribution to the multinomial, we can integrate out $\boldsymbol{\theta}$ to obtain a tractable compound distribution (Dirichlet-multinomial distribution):

$$\begin{aligned} p(\boldsymbol{x}_j \mid s_j, \boldsymbol{m}) &= \int_{\boldsymbol{\theta}_j} p(\boldsymbol{x}_j \mid \boldsymbol{\theta}_j) p(\boldsymbol{\theta}_j \mid s_j \boldsymbol{m}) d\boldsymbol{\theta}_j \\ &= \frac{(n_j!)\Gamma(s_j)}{\Gamma(n_j+s_j)} \prod_{l=1}^{L} \frac{\Gamma(x_{l,j}+s_j m_l)}{(x_{l,j}!)\Gamma(s_j m_l)} \end{aligned} \tag{2.10}$$

Intuitively, as the number of loci (L) becomes large, we can expect that the proportion of aligned reads at each locus ($\boldsymbol{m}$) will be small. From equation 2.7 we see that as L goes to infinity, the covariance of $x_{l,j}$ and $x_{l',j}$ will tend to zero. Indeed, for FAST-SeqS and many other genomics or transcriptomics assays, the number of considered loci is often large ($\geq 10,000$) and the covariance is negligible. As such, it is often convenient and more tractable to assume independence between variables. The marginal distribution of a Dirichlet-multinomial is a beta-binomial. Hence, rather than modelling loci counts by the Dirichlet-multinomial distribution, we can simply model the counts by independent beta-binomial distributions. As we will see, the assumption of independence is critical for further extensions of the model described in section 2.5.5.

Assuming independence, our model for the loci counts becomes:

$$\begin{aligned} \theta_{l,j} \mid s_j, m_l &\sim \mathrm{Beta}(s_j m_l, s_j(1-m_l)) \\ x_{l,j} \mid n_j, \theta_{l,j} &\sim \mathrm{Binomial}(n_j, \theta_{l,j}) \end{aligned} \tag{2.11}$$

Since the beta distribution is a conjugate distribution to the binomial, we can obtain a tractable compound distribution (Beta-binomial distribution) by integrating out $\theta_{l,j}$:

$$\begin{aligned} p(x_{l,j} \mid s_j m_l, s_j(1-m_l)) &= \int_{\theta_{l,j}} p(x_{l,j} \mid \theta_{l,j}) p(\theta_{l,j} \mid s_j m_l, s_j(1-m_l)) d\theta_{l,j} \\ &= \binom{n}{x_l} \frac{\mathrm{B}(x_l + a, n - x_l + b)}{\mathrm{B}(a,b)} \\ &= \frac{\Gamma(n+1)}{\Gamma(x_l+1)\Gamma(n-x_l+1)} \frac{\Gamma(x_l+a)\Gamma(n-x_l+b)}{\Gamma(n+a+b)} \frac{\Gamma(a+b)}{\Gamma(a)+\Gamma(b)} \end{aligned} \tag{2.12}$$

where $a = s_j m_l$, $b = s_j(1-m_l)$, $B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt$ (the beta function) and $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$ (Gamma function defined for positive z).

Now we can describe a generative model for loci counts with equal copy numbers. To make the model generative, we need to draw values of s_j and m_l

$$\begin{aligned} s_j \mid \psi &\sim \mathrm{Gamma}(\psi_{\text{shape}}, \psi_{\text{scale}}) \\ m_l \mid \phi &\sim \mathrm{Beta}(\phi_{c,l}, \phi_{d,l}) \\ \theta_{l,j} \mid s_j, m_l &\sim \mathrm{Beta}(s_j m_l, s_j(1-m_l)) \\ x_{l,j} \mid \theta_{l,j}, n_j &\sim \mathrm{Binomial}(n_j, \theta_{l,j}) \end{aligned} \tag{2.13}$$

In section 2.7.2 I describe an Markov chain Monte Carlo (MCMC) sampler to infer the latent variables $\boldsymbol{m}$ and $\boldsymbol{s}$ and in section 2.7.3, I briefly investigate the fit of the model using eight male samples and show that it provides a reasonable approximation to the distribution of loci counts. For the sake of the narrative, we now continue with the exploration of data and extending the count model to include copy number alterations.

2.5.4 Modelling loci counts for samples with copy number alterations

Now that we have developed a generative model for loci counts from samples with equal copy number at all loci, we can look to extend it to allow for altered copy number. How would we expect the counts to vary depending on the copy number alterations? One way to examine this is to explore the copy number differences in males and females. Here, we have a simple controlled experiment in which we can observe the change in count proportions due to the deletion of chromosome Y and duplication of chromosome X in female samples with respect to male samples.

Figure 2.5 shows the count proportions of the same two samples explored in figure 2.4D. When we considered the autosomes in isolation, the proportion of reads from the female sample was approximately equal (with some variance) to that of the males, and as such the slope of this relationship was equal to one. Now, with the addition of chromosome X and Y, we see that the slope describing the relationship between the proportion of reads within the autosomes is less than one and the slope for chromosome X is exactly double that of the autosomes. The value of this slope is the *relative* copy number of the female sample's DNA with respect to the male sample's DNA. Importantly, this relationship holds despite the loci bias; loci with high proportions are proportionally scaled by the same amount as loci with low proportions. This shows that the copy number signal is present in loci with low and high counts and that low counts (or high counts) are not simply the result of mismapping reads for example. In addition, the absence of counts from chromosome Y in the female sample further illustrates this point.

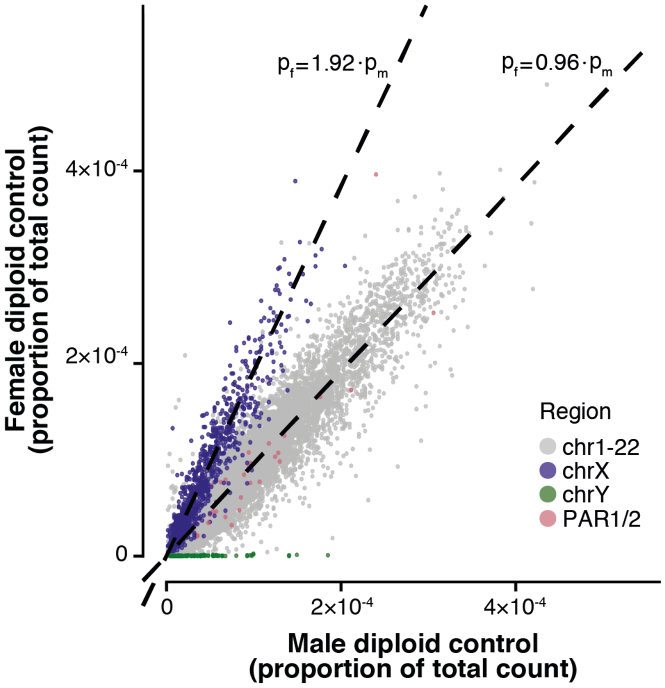


Figure 2.5 Comparison of loci count proportions between a normal male and normal female sample. Normal male loci proportions against normal female loci proportions indicating how copy number alterations affect the proportion of counts. Grey points represent loci located within the autosomes (chromosomes 1-22), blue points represent loci within chromosome X, green points represent loci within chromosome Y and pink points represent loci within the pseudoautosomal regions (PAR1 and PAR2).

Notice that the loci located within the pseudoautosomal regions (PAR1 and PAR2) are highlighted in figure 2.5. These regions are present at the termini of chromosomes X and Y and as such, cells from males and females will have two copies of these regions. PAR1 is approximately 2.7 Mbp in length and is located at the beginning of the p arms of chromosomes X and Y while PAR2 is around 330 Kbp and located at the end of the q arms of chromosomes X and Y. We see that the count proportions from these regions cluster with the autosomal chromosomes. This indicates that we may be able to use FAST-SeqS data to observe short regions of altered copy number.

I have shown there to be a difference in the proportion of counts (or mean count) between regions with altered copy number, but how do copy number alterations affect the variance? We can explore this by examining the mean-variance relationship of loci counts between male samples, comparing regions with one copy (chromosomes X and Y, excluding PAR1/2) with regions with two copies (automosomes and PAR1/2). Figure 2.6 shows that there is no discernible difference in the mean-variance relationship. This suggests that we do not need to include a dependence between the variance and relative copy number in the model.

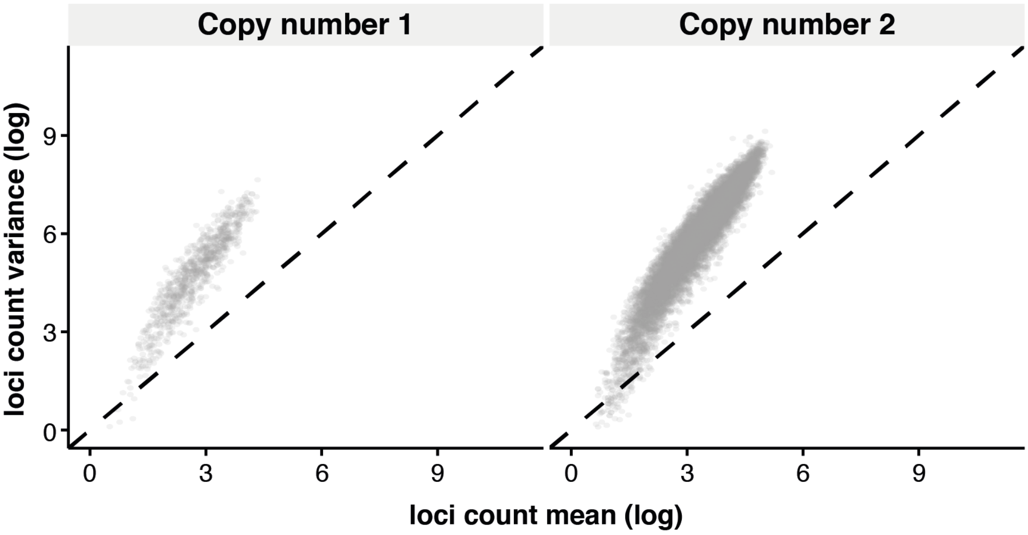


Figure 2.6 The log mean-variance relationship of loci counts with varying copy number. Each point represents a locus in which the mean and variance of the loci counts were calculated across eight normal male samples obtained from the oesophagus. Left: FAST-SeqS loci with one copy per cell; loci located within chromosome X and Y but not within the pseudoautosomal regions. Right: FAST-SeqS loci with two copies per cell; loci located within the autosomes and the pseudoautosomal regions.

Therefore, given the observations described above, we can extend the loci count model to account for copy number alterations. Here, I introduce a relative copy number value for each locus, $\hat{c}_l$:

$$\begin{aligned} s_j \mid \psi &\sim \text{Gamma}(\psi_{\text{shape}}, \psi_{\text{scale}}) \\ m_l \mid \phi &\sim \text{Beta}(\phi_{c,l}, \phi_{d,l}) \\ \theta_{l,j} \mid s_j, \hat{c}_l m_l &\sim \text{Beta}(s_j \hat{c}_l m_l, s_j(1 - \hat{c}_l m_l)) \\ x_{l,j} \mid \theta_{l,j}, n_j &\sim \text{Binomial}(n_j, \theta_{l,j}) \end{aligned} \tag{2.14}$$

Here, $\hat{c}_l$ scales the sample's expected proportion of reads, and as such equation 2.4 becomes $\mathbb{E}\left[\theta_{l,j}\right] = \hat{c}_l m_l$ and equation 2.5 becomes $\mathbb{E}\left[x_{l,j}\right] = n_j \hat{c}_l m_l$. A sample without copy number alterations can be generated if $\hat{c}_l = 1, \forall l$. With this updated generative model and given $\boldsymbol{m}$ (or the hyper-parameters $\phi_{c,1:L}$ and $\phi_{d,1:L}$), n_j, s_j (or the hyper-parameters ψ_{shape} and ψ_{scale}) and the loci relative copy numbers, $\hat{\boldsymbol{c}}$, we can generate *in silico* FAST-SeqS samples that have relative copy number alterations.

2.5.5 Linking loci using a Hidden Markov model

Given a sample's loci counts and given the model we described in 2.14, we could use a simple mixture model to infer the values of $\{\hat{c}_1, \ldots, \hat{c}_L\}$. However, by doing this we are not using all the information available to us. In particular, we are disregarding spatial information. By calculating the distances between FAST-SeqS loci within each chromosome arm, we observe that the median distance between neighbouring loci is 119,167 (IQR: 46,727-264,133, mean: 219,341) and that the distribution of these distances are similar between chromosomes (figure 2.7). The distribution of SCNA lengths (see Figure 1a of Beroukhim et al. [93]) suggest that neighbouring loci are likely to have the same copy number. Indeed, focal events are likely to involve multiple FAST-SeqS loci. Beroukhim *et al.* showed that the median length of focal SCNAs across cancer types is 1.8 Mb [93], while *Zack et al.* reported that median focal events were 19.6 Mb for telomere-bounded amplification events, 0.9 Mb for internal amplification events, 22.7 Mb for telomere-bounded deletions and 0.7 Mb for internal deletion events [115].

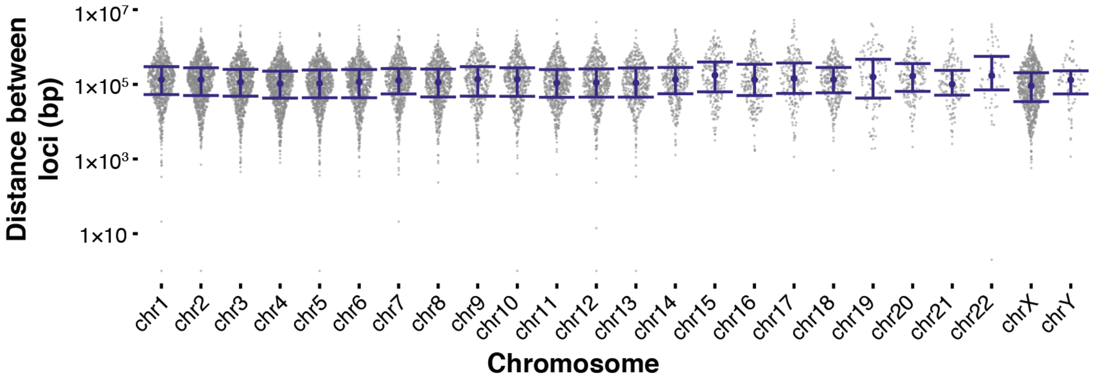


Figure 2.7 Distribution of genomic distances between neighbouring loci. Sina plot in which each grey point represents a distance between two neighbouring loci within a chromosome arm. Blue points represents the median distance between neighbouring loci for the associated chromosome and blue bars represent the interquartile range.

We can incorporate our understanding that neighbouring loci are more likely to share the same copy number through the use of a Hidden Markov model (HMM) [126, 127]. I chose to model each chromosome arm as an independent Markov chain; the relative copy number (RCN) of the first FAST-SeqS locus at the 5' end of the q arm of a chromosome is independent from that of the RCN of the last locus at the 3' of the p arm from the same chromosome (and all other chromosome arms). To continue to develop the model, we introduce an additional index to denote the chromosome arm, r. Now, a genomic locus is defined by its chromosome arm, r and the locus on the chromosome arm l, e.g. the count $x_{\text{chr1p},10}$ is associated with the tenth FAST-SeqS locus, with the loci index starting at the 5' end of each chromosome arm. Each chromosome arm, r, has a total of L_r FAST-SeqS loci within it.

Let us define the following additional variables (note that for simplicity I have dropped the sample index j):

- $z_{r,l}$ as the *hidden state* (or *copy number state*) of the Markov chain at locus l in chromosome arm r
- $\boldsymbol{\pi}^{\mathbf{0}}$ as the *initial distribution* of the first locus ($l = 1$) in chromosome arm
- $\pi_{u,v}$ as the *transition probability* from hidden state, u, to hidden state, v. The transition probabilities can be written more compactly as a *transition matrix*, $\boldsymbol{\pi}$. A row of the transition matrix, $\boldsymbol{\pi}_u$, can be thought of as a *transition distribution* from state u.
- $\hat{c}_u$ as the *relative copy number* associated with hidden state, u.

The first locus of a chromosome arm ($l = 1$) is distributed:

$$z_{r,1} \sim \boldsymbol{\pi}^{\mathbf{0}} \tag{2.15}$$

For all other loci ($l > 1$):

$$z_{r,l} \mid z_{r,l-1} \sim \boldsymbol{\pi}_{(z_{r,l-1})} \tag{2.16}$$

The count, $y_{r,l}$, at locus l in chromosome arm r is conditionally independent of the hidden states and observations of other loci:

$$\begin{aligned} \theta_{r,l} \mid \hat{\boldsymbol{c}}, z_{r,l}, m_{r,l}, s &\sim \text{Beta}(s\hat{c}_{z_{r,l}} m_{r,l}, s(1 - \hat{c}_{z_{r,l}} m_{r,l})) \\ y_{r,l} \mid \theta_{r,l}, n &\sim \text{Binomial}(n, \theta_{r,l}) \end{aligned} \tag{2.17}$$

The joint density for L_r loci in chromosome arm r is:

$$\begin{aligned} p(z_{r,1:L_r}, y_{r,1:L_r}, \theta_{r,1:L_r}) &= p(y_{r,1} \mid z_{r,1}, \theta_{r,1}) p(\theta_{r,1} \mid z_{r,1}) p(z_{r,1}) \\ &\quad \prod_{l=2}^{L_r} p(y_{r,l} \mid z_{r,l}, \theta_{r,l}) p(\theta_{r,l} \mid z_{r,l}) p(z_{r,l} \mid z_{r,l-1}) \\ &= \pi^0_{z_{r,1}} p(y_{r,1} \mid z_{r,1}, \theta_{r,1}) p(\theta_{r,1} \mid z_{r,1}) \\ &\quad \prod_{l=2}^{L_r} \pi_{z_{r,l-1}, z_{r,l}} p(y_{r,l} \mid z_{r,l}, \theta_{r,l}) p(\theta_{r,l} \mid z_{r,l}) \end{aligned} \tag{2.18}$$

where, $z_{r,1:L_r}$ denotes the sequence $\{z_{r,1}, \dots, z_{r,L_r}\}$, $y_{r,1:L_r}$ denotes $\{y_{r,1}, \dots, y_{r,L_r}\}$, and $\theta_{r,1:L_r}$ denotes $\{\theta_{r,1}, \dots, \theta_{r,L_r}\}$. The joint density for all L loci in the genome is given by:

$$p(\boldsymbol{z}, \boldsymbol{y}, \boldsymbol{\theta}) = \prod_r p(z_{r,1:L_r}, y_{r,1:L_r}, \theta_{r,1:L_r}) \tag{2.19}$$

2.5.6 The number of copy number states in a sample

In a classical finite HMM, the number of hidden states is assumed known and fixed. However, we do not know the number of copy number states present in a test sample *a priori*. Samples with equal copies of each locus will only have one copy number state. On the other hand, it is possible (although unlikely) that every locus in a sample has a unique copy number, leading to L copy number states in the sample. Furthermore, as we discussed in the introduction, samples contain a mixture of cells from multiple clones. Each clone could have a distinct, although probably phylogenetically related, SCNA profile. This heterogeneity increases the complexity in choosing the number of states; a copy number state represents a mean copy number within the sample and hence, the relative copy number states are not restricted to map to integer copy numbers.

Rather than choosing the number of states *a priori*, it would be preferable to learn the number of states from the data. Standard parametric model selection methods could be used to choose the number of states. However, this could involve testing up to L models and since $L > 10,000$ this is likely to be computationally expensive. One approach is to place a prior over the number of states in the model and use reversible jump MCMC (RJMCMC) and split-combine approaches to move between the number of hidden states using Metropolis-Hastings steps. However, these approaches can suffer from low acceptance rates in the split-combine steps and result in slow mixing [128].

Another approach is to use a Bayesian nonparametric version of the HMM [129–131]. In this approach, Teh *et al.* developed a model (the HDP-HMM) which uses a hierarchical Dirichlet process as a prior on the transition distributions and as such, allows for countably infinite numbers of states [130, 131]. Fox *et al.* showed that the HDP-HMM is inadequate for applications which involve temporal (or spatial) persistence of hidden states. In tasks such as these, the HDP-HMM can infer redundant hidden states with unrealistically rapid switching dynamics between states [131]. To account for state persistence, Fox *et al.* [131] proposed a modification to the HDP-HMM, which they called the sticky HDP-HMM. In this approach, an additional parameter was added to model self-transition bias. The sticky HDP-HMM was shown to have improved performance over the HDP-HMM in segmenting speech recordings into speaker-homogeneous regions while also inferring the number of total speakers (the speaker diarisation task) [131].

The sticky HDP-HMM provides a framework to extend our generative model for loci counts with a relative copy number profile. By using it, we are able to adequately model the spatial persistence of copy number states while simultaneously allowing for countably infinite copy number states within a sample. The sticky HDP-HMM generative model is as follows (note that I have omitted an emission distribution for now):

$$\begin{aligned}
\boldsymbol{\beta} \mid \gamma &\sim \mathrm{GEM}(\gamma) \\
\boldsymbol{\pi}^{\mathbf{0}} \mid \alpha, \boldsymbol{\beta} &\sim \mathrm{DP}\left(\alpha, \boldsymbol{\beta}\right) \\
\boldsymbol{\pi}_u \mid \alpha, \kappa, \boldsymbol{\beta} &\sim \mathrm{DP}\left(\alpha + \kappa, \frac{\alpha \boldsymbol{\beta} + \kappa \delta_u}{\alpha + \kappa}\right) \\
\hat{c}_u \mid H, \lambda &\sim H(\lambda) \\
z_{r,1} \mid \boldsymbol{\pi}^{\mathbf{0}} &\sim \boldsymbol{\pi}^{\mathbf{0}} \\
z_{r,l} \mid \{\boldsymbol{\pi}_u\}_{u=1}^{\infty}, z_{r,l-1} &\sim \boldsymbol{\pi}_{z_{r,l-1}}, \text{ for } l > 1
\end{aligned} \tag{2.20}$$

Here, DP denotes the Dirichlet process, GEM denotes the stick-breaking construction of the Dirichlet process and H is the prior base distribution of the Dirichlet process and represents a parametric distribution with parameters λ. $H(\lambda)$ can be viewed as our prior probability distribution on the relative copy number values of the hidden states. γ is a hyper-parameter of the sticky HDP-HMM and controls the number of copy number states in the sample; the greater the value of γ, the greater number of copy number states we expect in the sample. Indeed, the expected number of states is $\sum_{i=1}^{n} \frac{\gamma}{\gamma+i-1} \simeq \gamma \cdot \log\left(1 + L/\gamma\right)$ [132]. An introduction to the Dirichlet process is not provided here. A review of the Dirichlet process is provided in Teh [132] and the HDP-HMM and its sticky derivative are covered in depth by Teh *et al.* [130] and Fox *et al.* [131] respectively.

Each row of the transition matrix, π_u, is drawn from a Dirichlet process and depends on β, α and κ. It can be shown that:

$$\mathbb{E}\left[\pi_{u,v} \mid \alpha, \beta_v, \kappa\right] = \frac{\alpha \beta_v + \kappa \delta_{u,v}}{\alpha + \kappa} \tag{2.21}$$

where $\delta_{u,v}$ represents the discrete Kronecker delta function [131]. If we define $\rho = \frac{\kappa}{\alpha+\kappa}$ (as in Fox *et al.* [131]) and by noting that $\alpha = (1-\rho)(\alpha+\kappa)$, we obtain:

$$\mathbb{E}\left[\pi_{u,v} \mid \beta_v, \rho\right] = (1-\rho)\beta_v + \rho \delta_{u,v} \tag{2.22}$$

As such, we see that ρ defines how much weight is placed on self-transition within a copy number state. The vector, β, itself drawn from a Dirichlet process, represents the global transition distribution and holds information about the proportion of loci expected in each copy number state. Hence, $1-\rho$ is the weight placed on the global transition distribution.

The variance of the transition probability from copy number state u to v is given by:

$$\mathrm{Var}(\pi_{u,v} \mid \alpha, \boldsymbol{\beta}, \kappa) = \frac{\mathbb{E}\left[\pi_{u,v} \mid \alpha, \boldsymbol{\beta}, \kappa\right]\left(1 - \mathbb{E}\left[\pi_{u,v} \mid \alpha, \boldsymbol{\beta}, \kappa\right]\right)}{\alpha + \kappa + 1} \tag{2.23}$$

We see that $\alpha + \kappa$ is inversely proportional to the variance of the state transition probabilities. As such, it controls the variation between rows of the transition matrix.

2.6 The conliga probabilistic generative model

By noting observations drawn from exploring FAST-SeqS data, I have justified and described generative models for:

- loci counts with equal copy number (equation 2.13)
- loci counts with relatively altered copy number (equation 2.14)
- countably infinite copy number states with spatial state persistence (equation 2.20)

In this section, I put these generative models together into a formal and full model describing the generation of FAST-SeqS loci counts.

2.6.1 The full model

Here we describe the full generative model for the loci counts, $\boldsymbol{x}_k$, for $k \in \{1, \ldots, K\}$ control samples together with the loci counts, $\boldsymbol{y}_j$, for $j \in \{1, \ldots, J\}$ non-control samples with copy number profiles defined by $\hat{c}_{z_{j,r,l}}$ for all j, r and l, along with all other latent variables. This model is not used as a basis for inference of the latent variables due to the reasons described in the next section. Instead, this model is split into two parts for the purposes of inference. However, I include the full model for completeness. Figure 2.8 provides a graphical representation of the full model using plate notation.

$$\begin{aligned}
\boldsymbol{\beta}_j \mid \gamma &\sim \mathrm{GEM}(\gamma) \\
\boldsymbol{\pi}_j^{\mathbf{0}} \mid \alpha, \boldsymbol{\beta}_j &\sim \mathrm{DP}\left(\alpha, \boldsymbol{\beta}_j\right) \\
\boldsymbol{\pi}_{j,u} \mid \boldsymbol{\beta}_j, \alpha, \kappa &\sim \mathrm{DP}\left(\alpha + \kappa, \frac{\alpha \boldsymbol{\beta}_j + \kappa \delta_u}{\alpha + \kappa}\right) \\
\hat{c}_{j,u} \mid H, \lambda &\sim H(\lambda) \\
z_{j,r,1} \mid \boldsymbol{\pi}_j^{\mathbf{0}} &\sim \boldsymbol{\pi}_j^{\mathbf{0}} \\
z_{j,r,l} \mid \{\boldsymbol{\pi}_{j,u}\}_{u=1}^{\infty}, z_{j,r,l-1} &\sim \boldsymbol{\pi}_{j,z_{r,l-1}}, \text{ for } l > 1 \\
\tilde{s}_j \mid \boldsymbol{\omega} &\sim \mathrm{Gamma}(\omega_{\text{shape}}, \omega_{\text{scale}}) \\
s_k \mid \boldsymbol{\omega} &\sim \mathrm{Gamma}(\omega_{\text{shape}}, \omega_{\text{scale}}) \\
m_{r,l} \mid \phi_{c,r,l}, \phi_{d,r,l} &\sim \mathrm{Beta}(\phi_{c,r,l}, \phi_{d,r,l}) \\
\theta_{k,r,l} \mid s_k, m_{r,l} &\sim \mathrm{Beta}(s_k m_{r,l}, s_k(1 - m_{r,l})) \\
x_{k,r,l} \mid \theta_{r,l,k}, n_k &\sim \mathrm{Binomial}(n_k, \theta_{k,r,l}) \\
\tilde{\theta}_{j,r,l} \mid \{\hat{c}_{j,u}\}_{u=1}^{\infty}, z_{j,r,l}, m_{r,l}, \tilde{s}_j &\sim \mathrm{Beta}(\tilde{s}_j \hat{c}_{j,z_{j,r,l}} m_{r,l}, \tilde{s}_j(1 - \hat{c}_{j,z_{j,r,l}} m_{r,l})) \\
y_{j,r,l} \mid \tilde{\theta}_{j,r,l}, \tilde{n}_j, &\sim \mathrm{Binomial}(\tilde{n}_j, \tilde{\theta}_{j,r,l})
\end{aligned} \tag{2.24}$$

In this model we use a tilde to distinguish the variables that are associated with non-control samples ($\tilde{s}$ and $\tilde{\theta}$) from those associated with control samples. We note that $\boldsymbol{\beta}$, $\boldsymbol{\pi}^{\mathbf{0}}$, $\boldsymbol{\pi}_{\boldsymbol{u}}$, $\hat{c}_u$, $z_{r,l}$, $\tilde{s}$, $\tilde{\theta}_{r,l}$, $y_{r,l}$ and $\tilde{n}$ are specific to each non-control sample and include an index j. Notice that the sample inverse dispersion parameters (s and $\tilde{s}$) are drawn from the same distribution, reflecting our belief that control and non-control count observations have the same level of over-dispersion. If we were to use this model as a basis for inference of the latent variables, we may wish to place priors over the hyperparameters $\boldsymbol{\omega}$ and infer them from the data. This is also the case for the hyperparameters γ, α and κ and I exclude those priors for the sake simplicity. I provide the option to infer these hyperparameters from the data as described in the inference section.

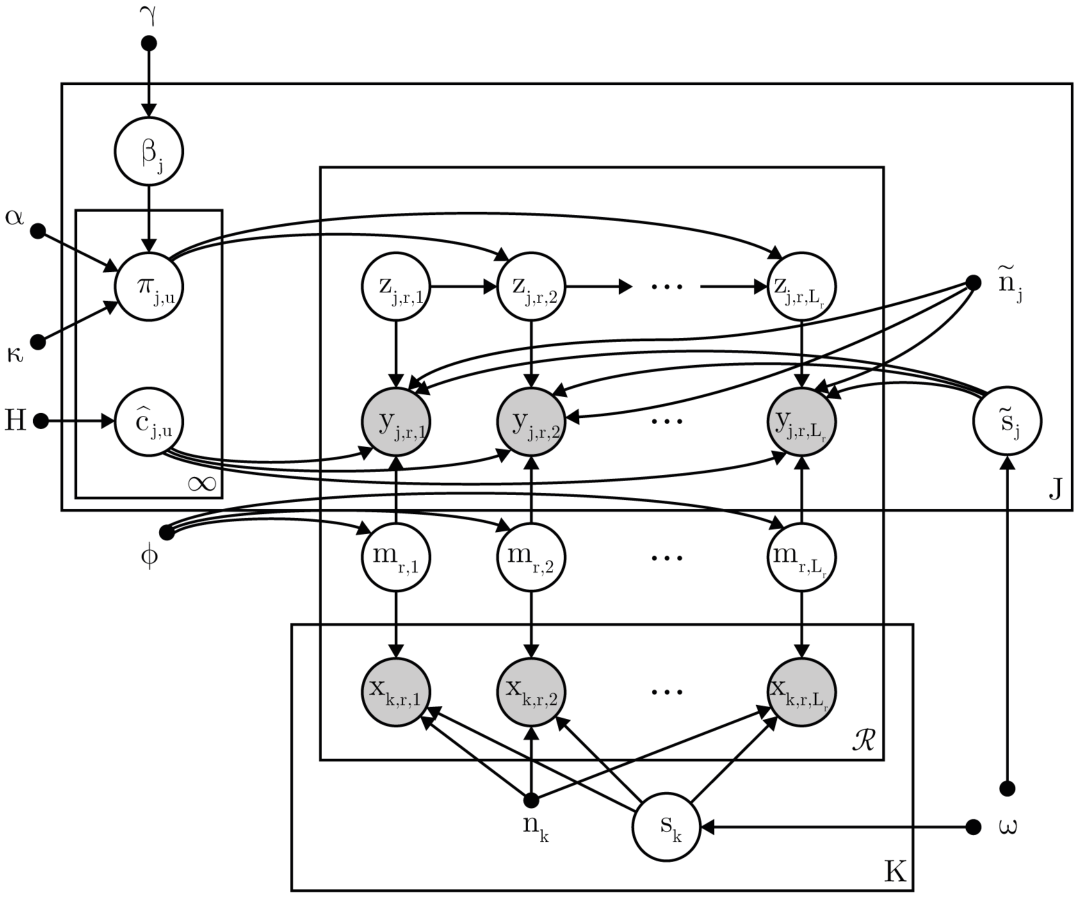


Figure 2.8 The conliga graphical model (described in section 2.6.1) represented in plate notation. Grey circles represent observed variable, white circles represent latent variables and block nodes indicate fixed hyperparameters. Note that $\boldsymbol{\theta}$ and $\tilde{\boldsymbol{\theta}}$ have been integrated out for simplicity.

2.6.2 The split model for simpler inference

I decided to split the full model in two parts. The first part describes the generation of loci counts for the control samples. The second part describes the generation of the copy number profile and loci count observations for a non-control sample, using the maximum a posteriori (MAP) estimate of $\boldsymbol{m}$. This was done to allow for simpler implementation of the inference algorithms. In this way, we could infer the MAP estimates of $\boldsymbol{m}$ and the nuisance parameters $\boldsymbol{s}$ using the counts of the control samples. Once MAP estimates were obtained, we could infer the copy number profile for each non-control sample in independent MCMC chains.

Splitting the model has the benefit that it is easier to implement and parallelise, allowing the relative copy number profiles to be inferred for each sample in parallel chains. However, splitting the model has its disadvantages. Rather than using controls and non-controls to infer $\boldsymbol{m}$, we are using only the control samples. By using less data, there will be less certainty in our inference of $\boldsymbol{m}$. In addition, by using MAP estimates in the second part of the model, we are not propagating uncertainty through the model to the inference of the relative copy number profiles. In practice, the mean proportions are likely to be well estimated by the MAP estimates and the ease of implementation outweighs the loss of statistical strength in the inference of $\boldsymbol{m}$. The split model is defined below and figure 2.9 shows a graphical representation of the split model using plate notation.

Part 1: Generative model for control counts

$$\begin{aligned}
s_k \mid \psi &\sim \text{Gamma}(\psi_{\text{shape}}, \psi_{\text{scale}}) \\
m_{r,l} \mid \phi &\sim \text{Beta}(\phi_{c,r,l}, \phi_{d,r,l}) \\
\theta_{k,r,l} \mid s_k, m_{r,l} &\sim \text{Beta}(s_k m_{r,l}, s_k(1 - m_{r,l})) \\
x_{k,r,l} \mid \theta_{k,r,l}, n_k &\sim \text{Binomial}(n_k, \theta_{k,r,l})
\end{aligned} \tag{2.25}$$

Part 2: Generative model for the counts of a sample with a relative copy number profile

$$\begin{aligned}
\boldsymbol{\beta} \mid \gamma &\sim \text{GEM}(\gamma) \\
\boldsymbol{\pi^0} \mid \alpha, \boldsymbol{\beta} &\sim \text{DP}\left(\alpha, \boldsymbol{\beta}\right) \\
\boldsymbol{\pi}_u \mid \alpha, \kappa, \boldsymbol{\beta} &\sim \text{DP}\left(\alpha + \kappa, \frac{\alpha\boldsymbol{\beta} + \kappa\delta_u}{\alpha + \kappa}\right) \\
\hat{c}_u \mid H, \lambda &\sim H(\lambda) \\
z_{r,1} \mid \boldsymbol{\pi^0} &\sim \boldsymbol{\pi^0} \\
z_{r,l} \mid \{\boldsymbol{\pi}_u\}_{u=1}^{\infty}, z_{r,l-1} &\sim \boldsymbol{\pi}_{z_{r,l-1}}, \text{ for } l > 1 \\
\tilde{s} \mid \omega &\sim \text{Gamma}(\omega_{shape}, \omega_{scale}) \\
\tilde{\theta}_{r,l} \mid \{\hat{c}_u\}_{u=1}^{\infty}, z_{r,l}, \hat{m}_{r,l}, \tilde{s} &\sim \text{Beta}(\tilde{s}\hat{c}_{z_{r,l}}\hat{m}_{r,l}, \tilde{s}(1 - \hat{c}_{z_{r,l}}\hat{m}_{r,l})) \\
y_{r,l} \mid \tilde{\theta}_{r,l}, \tilde{n}, &\sim \text{Binomial}(\tilde{n}, \tilde{\theta}_{r,l})
\end{aligned} \tag{2.26}$$

Note that $\hat{m}_{r,l}$ refers to the maximum a posteriori (MAP) value of $m_{r,l}$ and is assumed to be a known quantity in equation 2.26. For simplicity, the hyperparameters (α, κ, γ, λ and ω) are shown as fixed quantities in the model. In practice, λ and ω are treated as fixed (in addition to the total number of loci counts, $\tilde{n}$), while the model is parameterised in terms of ρ and $(\alpha + \kappa)$, with an optional Beta prior placed on ρ and optional Gamma priors placed on $(\alpha + \kappa)$ and γ as in Fox *et al.* [131]. See section 2.10 for further details on the prior distributions used.

Note that different priors are placed on the sample specific inverse dispersion parameters in part 1 and part 2. This explicit distinction is made so that we have the freedom to use the marginal posterior distribution of s_k to update our prior distribution over $\tilde{s}$. A table with all the variables used in the model can be found in table A.1 of appendix A.

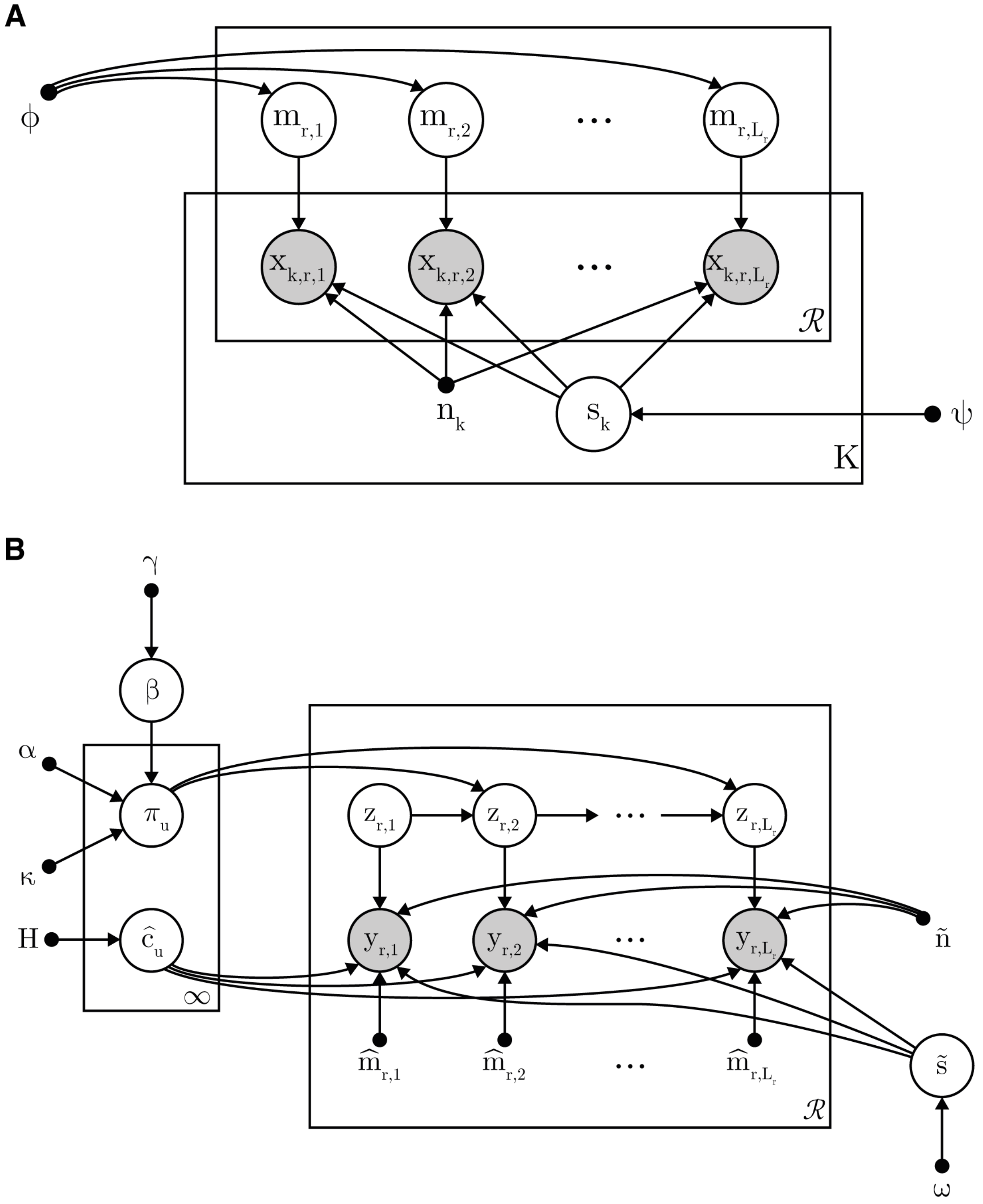


Figure 2.9 The conliga graphical model (described in section 2.6.2) represented in plate notation. Grey circles represent observed variable, white circles represent latent variables and block nodes indicate fixed hyperparameters. Note that $\boldsymbol{\theta}$ and $\tilde{\boldsymbol{\theta}}$ have been integrated out for simplicity. A: Part 1 of the model, representing the generation of FAST-SeqS data for normal diploid samples. B: Part 2 of the model, representing the generation of FAST-SeqS data for a sample with a copy number profile.

2.7 Inference

2.7.1 Notation used in MCMC algorithms

We denote the probability mass function (pmf) of the compound beta-binomial distribution as follows:

$$f_{\mathrm{BB}}(k; n, a, b) = \binom{n}{k} \frac{\mathrm{B}(k + a, n - k + b)}{\mathrm{B}(a, b)} \tag{2.27}$$

where B represents the Beta function, n represents the total number of Bernoulli trials (counts), k represents a count observation ($k \in \{0, \ldots, n\}$), and a and b are parameters ($a > 0$, $b > 0$). This is used as our likelihood function in Algorithm 1 where $k = x_{k,r,l}$, $n = n_k$, $a = s_k m_{r,l}$, and $b = s_k(1 - m_{r,l})$. It is also used as a likelihood function in Algorithm 2 where $k = y_{r,l}$, $n = \tilde{n}$, $a = \tilde{s}\hat{c}_{z_{r,l}}\hat{m}_{r,l}$, and $b = \tilde{s}(1 - \hat{c}_{z_{r,l}}\hat{m}_{r,l})$.

We denote the probability density function (pdf) of the Beta distribution as follows:

$$f_{\mathrm{Beta}}(x; c, d) = \frac{1}{\mathrm{B}(c, d)} x^{c-1}(1 - x)^{d-1} \tag{2.28}$$

where x represents an observation ($x \in [0, 1]$), and c and d are the shape parameters of the Beta distribution ($c > 0$, $d > 0$).

We denote the pdf of the Gamma distribution as:

$$f_{\mathrm{Gamma}}(x; k, \theta) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}} \tag{2.29}$$

where Γ represents the Gamma function, k is the shape parameter ($k > 0$), θ is the scale parameter ($\theta > 0$) and observation x (where $x \geq 0$).

The normal distribution is denoted $N(\mu, \sigma)$ and the uniform distribution is denoted $U(a, b)$.

We use the follow notation:

- $x_{\cdot,j} = \sum_i x_{i,j}$
- $x_{i,\cdot} = \sum_j x_{i,j}$
- $x_{i,*}$ denotes the i$^{\mathrm{th}}$ row vector of matrix, x
- $x_{*,j}$ denotes the j$^{\mathrm{th}}$ column vector of matrix, x
- $x_{i,\setminus j} = \{x_{i,1}, \ldots, x_{i,j-1}, x_{i,j+1}, \ldots x_{i,J}\}$, where J is largest value of j
- $|X|$ represents the cardinality of set X

2.7.2 Inference of loci means and sample inverse dispersion parameters

Given the model described in equation 2.25, I created a Metropolis-Hastings random walk algorithm (Algorithm 1) to sample from the joint posterior distribution $p(\boldsymbol{m}, \boldsymbol{s} \mid \boldsymbol{x})$. It consists of two steps to sample from the marginal posterior distributions $p(m_{r,l} \mid \boldsymbol{x})$ for all r, l and then to sample from $p(s_k \mid \boldsymbol{x})$ for all $k \in \{1, \ldots, K\}$.

Algorithm 1 Inferring the expected proportions of reads, $\boldsymbol{m}$, for each locus and the sample inverse dispersion values, s_k for diploid controls $1, \ldots, K$

```
procedure InferBBParams($\boldsymbol{x}, \boldsymbol{\phi}, \boldsymbol{\psi}, N, \sigma_m, \sigma_s$)
    Initialise MCMC:
    for each chromosome arm, $r$, do
        for each locus, $l$, in chromosome arm, $r$, do
            Initialise $m_{r,l}$ to $\frac{x_{\cdot,r,l}}{x_{\cdot,\cdot,\cdot}}$
    for each control sample, $k$, do
        Draw sample inverse dispersion, $s_k$, from Gamma($\psi_{\text{shape}}$, $\psi_{\text{scale}}$)
        Compute total counts for sample $k$: $n_k \leftarrow x_{k,\cdot,\cdot}$
    Begin MCMC:
    for each iteration, 1 to $N$ do
        Update expected proportions, m:
        for each chromosome arm, $r$, do
            for each locus, $l$, in chromosome arm, $r$, do
                Propose new value for expected proportion: $m^*_{r,l} \sim N(m_{r,l}, \sigma_m)$
                if $m^*_{r,l} > 0$ then
                    $p_{\text{new}} \leftarrow \sum_k (\log f_{BB}(x_{k,r,l}; n_k, s_k m^*_{r,l}, s_k(1 - m^*_{r,l}))) + \log f_{Beta}(m^*_{r,l}; \phi_c, \phi_d)$
                    $p_{\text{old}} \leftarrow \sum_k (\log f_{BB}(x_{k,r,l}; n_k, s_k m_{r,l}, s_k(1 - m_{r,l}))) + \log f_{Beta}(m_{r,l}; \phi_c, \phi_d)$
                    $p \sim U(0, 1)$
                    if $\exp(p_{\text{new}} - p_{\text{old}}) < p$ then
                        Accept new value: $m_{r,l} = m^*_{r,l}$
        Update sample inverse dispersion values, $\boldsymbol{s}$:
        for each control sample, $k$, do
            Propose new value for sample inverse dispersion: $s^*_k \sim N(s_k, \sigma_s)$
            $p_{\text{new}} \leftarrow \sum_r \sum_l (\log f_{BB}(x_{k,r,l}; n_k, s^*_k m_{r,l}, s^*_k(1 - m_{r,l}))) + \log f_{\text{Gamma}}(s^*_k; \psi_{\text{shape}}, \psi_{\text{scale}})$
            $p_{\text{old}} \leftarrow \sum_r \sum_l (\log f_{BB}(x_{k,r,l}; n_k, s_k m_{r,l}, s_k(1 - m_{r,l}))) + \log f_{\text{Gamma}}(s_k; \psi_{\text{shape}}, \psi_{\text{scale}})$
            $p \sim U(0, 1)$
            if $\exp(p_{\text{new}} - p_{\text{old}}) < p$ then
                Accept new value: $s_k = s^*_k$
```

2.7.3 Investigating the fit of the model

Now that we have described the inference procedure, we can explore how well the model fits the data. I used the set of eight male control samples (explored previously in this chapter) and their associated loci counts within chromosomes 1-22, X and Y, and the implementation of Algorithm 1 within the conliga software (the `fit_controls` function) [105, 108] to obtain a sample from the posterior distribution $p(\boldsymbol{m}, s_{1:8} \mid \boldsymbol{x})$. Within this function, the marginal distributions $p(m_{r,l} \mid \boldsymbol{x})$ for all r, l are summarised by using the KernSmooth [133] R package to smooth the marginal posterior densities and then the mode is computed. This resulted in the vector of MAP estimates, $\hat{\boldsymbol{m}}$. As we discussed previously in this chapter, $\hat{\boldsymbol{m}}$ can be interpreted as a vector describing the loci biases with $\hat{m}_{r,l}$ being the expected proportion of aligned reads to originate from locus l on chromosome arm r in a control sample. Note that these MAP estimates are unlikely to sum to one exactly, as such, we re-scale $\hat{m}_{r,l}$ so that the MAP estimates sum to 1. This means that the relative copy number for each locus in a reference male sample should be equal to 1.

I used flat Beta$(1, 1)$ priors on $m_{r,l}$ for all r, l and a Gamma$(1.5, 10^6)$ prior (parameterised by shape and scale) on $\boldsymbol{s}$. Loci were filtered such that only those loci that had a non-zero count in all 13 normal samples within the autosomes and chromosome X were kept. In addition, loci from chromosome Y were kept if all eight males samples had a non-zero count.

One way to inspect how well the model is capturing the data is to compare observed counts to our predictions using the posterior predictive distribution:

$$p\left(\tilde{x} \mid \boldsymbol{x}\right) = \int_{\tilde{s}} \int_{\boldsymbol{m}} p\left(\tilde{s}, \boldsymbol{m} \mid \boldsymbol{x}\right) p\left(\tilde{x} \mid \tilde{s}, \boldsymbol{m}\right) d\tilde{s} \cdot d\boldsymbol{m} \tag{2.30}$$

However, we carry forward the MAP estimates of $\boldsymbol{m}$ to algorithm 2 when inferring RCN profiles because we believe the marginal posterior distribution of $m_{r,l}$ is well approximated by a delta function at the MAP estimate. Therefore, rather than using the posterior predictive distribution, I used a plug-in approximation [134] in which we assume $p(\tilde{x} \mid \boldsymbol{x}) \approx p(\tilde{x} \mid \tilde{s}^{\text{MAP}}, \hat{\boldsymbol{m}})$ since if the MAP approximates the marginal posterior distributions we have:

$$\begin{aligned} p\left(\tilde{x} \mid \boldsymbol{x}\right) &\approx \int_{\tilde{s}} \int_{\boldsymbol{m}} \delta_{\tilde{s}}(\tilde{s} - \tilde{s}^{\text{MAP}}) \delta_{\boldsymbol{m}}(\boldsymbol{m} - \hat{\boldsymbol{m}}) p\left(\tilde{x} \mid \tilde{s}, \boldsymbol{m}\right) d\tilde{s} \cdot d\boldsymbol{m} \\ &\approx p(\tilde{x} \mid \tilde{s}^{\text{MAP}}, \hat{\boldsymbol{m}}) \end{aligned} \tag{2.31}$$

where $\delta(t)$ is the Dirac delta function.

By plugging in the MAP estimates of $\tilde{\boldsymbol{s}}$ and $\boldsymbol{m}$, we can compute the theoretical quantiles of the beta-binomial distribution for each locus and sample. Figure 2.10 shows the loci counts of the eight male samples, with the counts ordered on the x-axis in descending order of their associated value for $\hat{\boldsymbol{m}}$. We see that the model provides a reasonable fit to the data. However, in all samples, we see fewer counts outside the 90% prediction interval than expected (range: 5.5-7%). There are several possible reasons for this. One reason is that the assumption that all control samples have equal copies of every locus is unlikely to be true. Indeed, 4.8-12% of the human genome is thought to be variable in copy number across a population of healthy individuals [135, 136], with CNVs occurring in approximately 1.2% of an individual's genome [94]. As such, it is likely that some of the outliers will be due to CNVs. To capture these outliers, a lower inverse dispersion will be inferred leading to more loci than expected within the prediction intervals and fewer outside. Other assumptions of the model may not be strictly true. For example, the assumption that there are no sample-specific biases may be an oversimplification. Indeed, male samples 6 and 8 suggest that the assumption of a global mean proportion may not hold. In these samples, we see more outlying high counts than expected when $\hat{m}_{r,l}$ is high. We explore this observation further in Chapter 4. For now, we note that the model is not a perfect representation of the generation of FAST-SeqS data and abide by George E. P. Box's famous adage that "all models are wrong, but some are useful" [137].

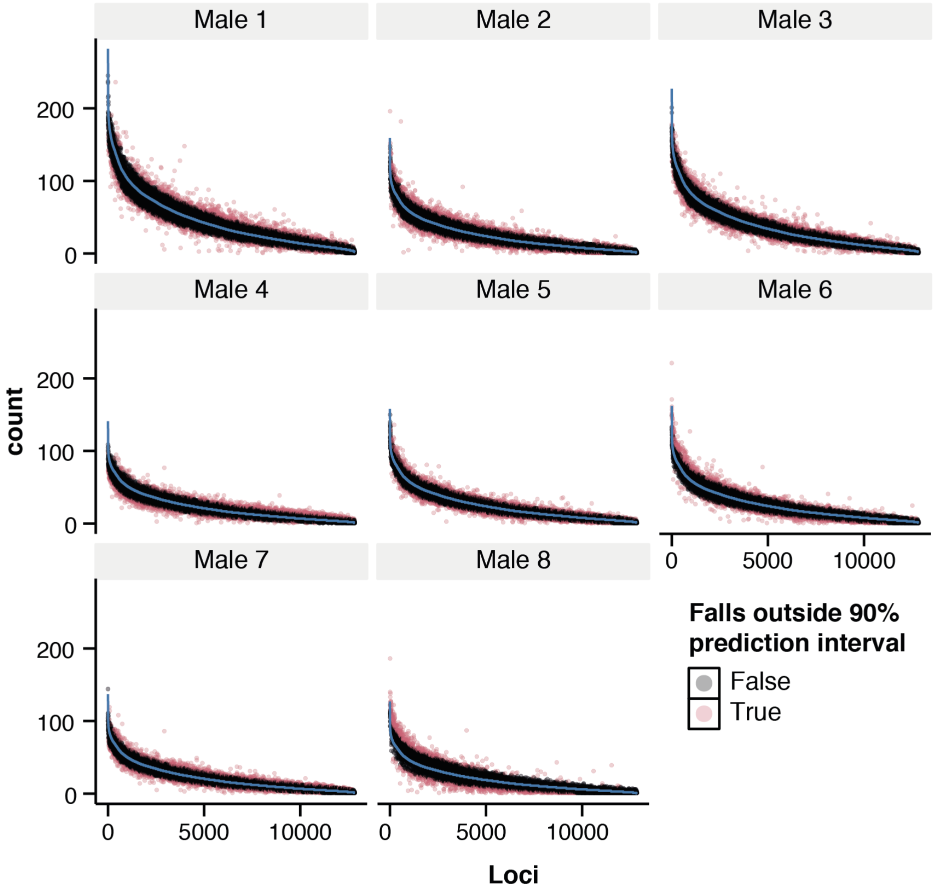


Figure 2.10 Exploration of model fit to the data. Loci counts from eight normal male samples from the oesophagus. Loci are in descending order of the inferred value of the mean proportions, $\hat{\boldsymbol{m}}$, on the x-axis. Given each sample's inferred MAP estimate for the inverse-dispersion parameter, s_k, the blue line represents the 50% percentile of the plug-in approximation to the posterior predictive distribution, $p(\tilde{x} \mid \tilde{s}^{MAP}, \hat{\boldsymbol{m}})$. Each point represents a count which is coloured black if it falls within the 90% prediction interval and red if it falls outside it.

2.7.4 Inference of relative copy number profile

Based on the generative model I described in equation 2.26 in section 2.6.2, I created a MCMC sampler to sample from the posterior distribution. I chose to do this by building upon Fox *et al.*'s blocked Gibbs sampler for the Sticky HDP-HMM, which is described in algorithm 3 of the Supplementary Materials of Fox *et al.* [131]. The blocked sampler uses the weak order K limit approximation to the Dirichlet process [138, 139, 131] which gives a prior distribution over the global transition distribution, $\boldsymbol{\beta}$:

$$\boldsymbol{\beta} \mid \gamma \sim \mathrm{Dir}(\gamma/K, \dots, \gamma/K) \tag{2.32}$$

Then the prior distribution over transition probabilities and initial probabilities are:

$$\boldsymbol{\pi^0} \mid \alpha, \boldsymbol{\beta} \sim \mathrm{Dir}(\alpha\beta_1, \dots, \alpha\beta_K) \tag{2.33}$$

$$\boldsymbol{\pi}_u \mid \alpha, \boldsymbol{\beta}, \kappa \sim \mathrm{Dir}(\alpha\beta_1, \dots, \alpha\beta_u + \kappa, \dots, \alpha\beta_K) \tag{2.34}$$

The marginal posterior distributions are:

$$\begin{aligned}
\boldsymbol{\beta} \mid \bar{\boldsymbol{M}}, \gamma &\sim \mathrm{Dir}(\gamma/K + \bar{M}_{\cdot 1}, \dots, \gamma/K + \bar{M}_{\cdot K}) \\
\boldsymbol{\pi^0} \mid \boldsymbol{z}, \alpha, \boldsymbol{\beta} &\sim \mathrm{Dir}(\alpha\beta_1 + T_1^0, \dots, \alpha\beta_K + T_K^0) \\
\boldsymbol{\pi}_u \mid \boldsymbol{z}, \alpha, \boldsymbol{\beta}, \kappa &\sim \mathrm{Dir}(\alpha\beta_1 + T_{u,1}, \dots, \alpha\beta_u + \kappa + T_{u,u}, \dots, \alpha\beta_K + T_{u,K})
\end{aligned} \tag{2.35}$$

where T_u^0 represents the number of FAST-SeqS loci at the beginning of a chromosome arm (the first amplified locus at the 5' end) in state u, $T_{u,v}$ represents the number of transitions from state u to state v across all loci in all chromosome arms. $\bar{M}_{uv}$ represents an auxiliary variable which Fox *et al.* [131] explains by analogy in which the Sticky HDP can be thought of as a Chinese Restaurant Franchise with Loyal Customers (extending Teh *et al.*'s analogy of the HDP as a Chinese Restaurant Franchise [130]). In this analogy (which can be found in the Supplementary Materials of Fox *et al.* [131]), $\bar{M}_{u,v}$ is the number of tables in restaurant u that considered dish v and therefore $\bar{M}_{\cdot,v}$ represents the sum of all tables in all restaurants that considered dish v.

The blocked Gibbs sampler utilises the forward-backward procedure [140] to jointly sample all hidden states, $\boldsymbol{z}$, given the count observations, $\boldsymbol{y}$. This is achieved by firstly computing the backward messages ($\boldsymbol{\mu}$) for each chromosome arm, where $\mu_{l,z_{l-1}}$ defines the backwards message passed from z_l to z_{l-1} within a chromosome arm and is recursively defined as (note that I have dropped the chromosome arm index r for simplicity):

$$\mu_{l,z_{l-1}} \propto \begin{cases} \sum_{z_l} p(z_l \mid \boldsymbol{\pi}_{z_{l-1}}) \cdot f_{\mathrm{BB}}(y_l \mid n, \tilde{s}\hat{c}_{z_l}\hat{m}_l, \tilde{s}(1 - \hat{c}_{z_l}\hat{m}_l)) \cdot \mu_{l+1,z_l}, & \text{for } 2 \leq l \leq L \\ 1, & \text{for } l = L + 1 \end{cases} \tag{2.36}$$

and by noting that $\mu_{l,z_{l-1}} \propto p(z_{l:L} \mid z_{l-1}, \boldsymbol{\pi}, \tilde{n}, \hat{\boldsymbol{c}}, \hat{\boldsymbol{m}}, \tilde{s})$, the conditional distribution of z_l is given by:

$$p(z_l \mid z_{l-1}, y_{1:L}, \boldsymbol{\pi}, \tilde{n}, \hat{\boldsymbol{c}}, \hat{\boldsymbol{m}}, \tilde{s}) \propto \begin{cases} p(z_l \mid \boldsymbol{\pi}_{z_{l-1}}) \cdot f_{\text{BB}}(y_l \mid n, \tilde{s}\hat{c}_{z_l}\hat{m}_l, \tilde{s}(1 - \hat{c}_{z_l}\hat{m}_l)) \cdot \mu_{l+1,z_l}, & \text{if } 2 \leq l \leq L \\ p(z_l \mid \boldsymbol{\pi}^{\mathbf{0}}) \cdot f_{\text{BB}}(y_l \mid n, \tilde{s}\hat{c}_{z_l}\hat{m}_l, \tilde{s}(1 - \hat{c}_{z_l}\hat{m}_l)) \cdot \mu_{l+1,z_l}, & \text{if } l = 1 \end{cases} \tag{2.37}$$

Note that the derivation of the forward-backward algorithm and conditional distributions for the hyperparameters (α, κ and γ) can be found in the supplementary materials of Fox *et al.* [131] and as such, I do not repeat those here.

I introduce Metropolis-Hastings random walk steps to sample the relative copy number, $\hat{c}_u$, associated with each state, $u \in \{1, \ldots, S\}$ and to sample the inverse dispersion parameter, $\tilde{s}$. The full MCMC algorithm can be found in Algorithm 2.

Algorithm 2 Inferring the relative copy number profile and inverse-dispersion, $\tilde{s}$, for a test sample

```
procedure InferRCN($\boldsymbol{y}, \boldsymbol{A}, \boldsymbol{\omega}, \boldsymbol{\lambda}, \hat{\boldsymbol{m}}, \boldsymbol{L}, \tilde{n}, N, S, \sigma_{\hat{c}}, \sigma_{\tilde{s}}$)
    Initialise MCMC:
    If $\gamma$ not fixed, draw $\gamma$ from prior: $\gamma \sim \text{Gamma}(A_{\gamma,a}, A_{\gamma,b})$
    If $(\alpha + \kappa)$ not fixed, draw $(\alpha + \kappa)$ from prior: $\alpha + \kappa \sim \text{Gamma}(A_{(\alpha+\kappa),a}, A_{(\alpha+\kappa),b})$
    If $\rho$ not fixed, draw $\rho$ from prior: $\rho \sim \text{Beta}(A_{\rho,c}, A_{\rho,d})$
    Set $\kappa \leftarrow \rho(\alpha + \kappa)$
    Set $\alpha \leftarrow (\alpha + \kappa) - \kappa$
    Draw sample inverse-dispersion, $\tilde{s}$, from prior: $\tilde{s} \sim \text{Gamma}(\omega_{shape}, \omega_{scale})$
    for each chromosome arm, $r$, do
        for each locus, $l$, in chromosome arm, $r$, do
            If not initialized, then randomly draw hidden state: $z_{r,l} \sim \text{Categorical}\left(S, \{\frac{1}{S}, \ldots, \frac{1}{S}\}\right)$
    $\boldsymbol{T}^0, \boldsymbol{T} \leftarrow$ countTransitions($\boldsymbol{z}, \boldsymbol{L}, S$)
    Define transition probabilities, $\boldsymbol{\pi}$, as a $S$ by $S$ matrix
    Define initial distribution, $\boldsymbol{\pi}^0$, as a vector of length $S$
    Set $\boldsymbol{\pi}$ and $\boldsymbol{\pi}^0$ as follows:
    for each state, $u$, from 1 to $S$, do
        Set $\pi^0_u \leftarrow \frac{T^0_u}{\sum_u T^0}$
        if $T_{\cdot u}$ is equal to 0 then
            Set $\pi_{u,u} \leftarrow 1$, $\pi_{u,\setminus u} \leftarrow 0$
        else
            for each state, $v$, from 1 to $S$, do
                Set $\pi_{u,v} \leftarrow \frac{T_{u,v}}{T_{\cdot,v}}$
    for each state, u, from 1 to S do
        If not initialized, draw relative copy number for state $u$: $c_u \sim \text{Gamma}(\lambda_{\text{shape}}, \lambda_{\text{scale}})$
    Draw $\beta$ via the stick breaking (GEM) process as follows:
    initialise $len$ to 1 and initialise $i$ to 1
    while $i \leq S$ do
        $\beta_i \sim \text{Beta}(1, \gamma) \cdot len$
        $len \leftarrow len - \beta_i$
        increment $i$
    Begin MCMC
    for each iteration, 1 to $N$ do
        $\boldsymbol{q} \leftarrow$ computeLogLikCache($\boldsymbol{y}, \hat{\boldsymbol{m}}, \hat{\boldsymbol{c}}, \tilde{n}, \tilde{s}, S$)
        for each chromosome arm, $r$, do
            $\boldsymbol{\mu} \leftarrow$ computeMessages($\boldsymbol{q}, \boldsymbol{\pi}, L_r, r, S$)
            $\boldsymbol{z_{r,*}} \leftarrow$ sampleStates($\boldsymbol{q}, \boldsymbol{\mu}, \boldsymbol{\pi}, \boldsymbol{\pi^0}, \boldsymbol{L_r}, \boldsymbol{r}$)
        $\boldsymbol{T}^0, \boldsymbol{T} \leftarrow$ countTransitions($\boldsymbol{z}, \boldsymbol{L}, S$)
        $\boldsymbol{M}, \boldsymbol{w}, \bar{\boldsymbol{M}} \leftarrow$ sampleAuxVars($\boldsymbol{\beta}, \boldsymbol{T}, \alpha, \kappa, S$)
        Sample Global Transition Distribution: $\boldsymbol{\beta} \sim Dir\left(\gamma/S + \bar{M}_{\cdot,1}, \ldots, \gamma/S + \bar{M}_{\cdot,S}\right)$
        $\boldsymbol{\pi}^0, \boldsymbol{\pi} \leftarrow$ sampleTransDists($\boldsymbol{T}^0, \boldsymbol{T}, S$)
        $\hat{\boldsymbol{c}} \leftarrow$ sampleCopyNumber($\boldsymbol{z}, \hat{\boldsymbol{c}}, \hat{\boldsymbol{m}}, \boldsymbol{y}, \boldsymbol{\lambda}, \boldsymbol{q}, S, \tilde{s}, \tilde{n}, \sigma_{\hat{c}}$)
        $\tilde{s} \leftarrow$ sampleInverseDispersion($\boldsymbol{y}, \hat{\boldsymbol{c}}, \hat{\boldsymbol{m}}, \boldsymbol{L}, \boldsymbol{\omega}, \mathcal{R}, \tilde{n}, \sigma_{\tilde{s}}$)
        $\alpha, \kappa, \gamma \leftarrow$ sampleHyperparameters($\boldsymbol{T}, \boldsymbol{M}, \boldsymbol{W}, \bar{\boldsymbol{M}}, \boldsymbol{A}, \alpha, \kappa, \gamma, S$)
```

```
function computeLogLikCache($\boldsymbol{y}, \hat{\boldsymbol{m}}, \hat{\boldsymbol{c}}, \tilde{n}, \tilde{s}, S$)
    ▷ This function is used so that we do not needlessly calculate the log likelihood function several times during an iteration of the MCMC.
    for each chromosome arm, $r$, do
        for each locus, $l$, in chromosome arm, $r$, do
            for each state, $u$, from 1 to $S$, do
                Store log likelihood for locus in state:
                $q_{r,l,u} \leftarrow \log f_{BB}(y_{r,l}; \tilde{n}, \tilde{s}\hat{c}_u\hat{m}_{r,l}, \tilde{s}(1 - \hat{c}_u\hat{m}_{r,l}))$
    return $\boldsymbol{q}$

function computeMessages($\boldsymbol{q}, \boldsymbol{\pi}, L_r, r, S$) ▷ Note that $\mu_{l,u}$ denotes the backward message passed from $z_{l+1}$ to $z_l$
    for each state, $u \in \{1, \ldots, S\}$ do
        Initialise messages:
        $\mu_{L_r,u} \leftarrow 1$
    for each locus, $l \in \{L_r - 1, \ldots, 1\}$ do
        for each state, $u \in \{1, \ldots, S\}$ do
            $\mu_{l,u} \leftarrow \sum_{v=1}^{S} (\pi_{u,v} \cdot \exp(q_{r,l,v}) \cdot \mu_{l+1,v})$
        Scale row of messages by the maximum value:
        $\boldsymbol{\mu}_{l,*} \leftarrow \boldsymbol{\mu}_{l,*} / \max_{1 \leq u \leq S} \mu_{l,u}$
     return $\boldsymbol{\mu}$

function sampleStates($\boldsymbol{q}, \boldsymbol{\mu}, \boldsymbol{\pi}, \boldsymbol{\pi}^0, L_r, r$)
    for each state, $u \in \{1, \ldots, S\}$ do
        $\log p_u \leftarrow \log \pi_u^0 + \log \mu_{1,u} + q_{r,1,u}$
    Scale $\boldsymbol{p}$ to avoid underflow: $\boldsymbol{p} \leftarrow \exp(\log \boldsymbol{p} - \max_{1 \leq u \leq S} \log p_u)$
    Make $\boldsymbol{p}$ sum to 1: $\boldsymbol{p} \leftarrow \boldsymbol{p} / \sum_{u=1}^{S} p_u$
    Sample state at the first locus of chromosome arm: $z_1 \sim \text{Categorical}\,(S, \boldsymbol{p})$
    for each locus, $l \in \{2, \ldots, L_r\}$ do
        for each state, $u \in \{1, \ldots, S\}$ do
            $\log p_u \leftarrow \log \pi_{z_{l-1},u} + \log \mu_{l,u} + q_{r,l,u}$
        Scale $\boldsymbol{p}$ to avoid underflow: $\boldsymbol{p} \leftarrow \exp(\log \boldsymbol{p} - \max_{1 \leq u \leq S} \log p_u)$
        Make $\boldsymbol{p}$ sum to 1: $\boldsymbol{p} \leftarrow \boldsymbol{p} / \sum_{u=1}^{S} p_u$
        Sample state at locus, $l$, of chromosome arm: $z_l \sim \text{Categorical}\,(S, \boldsymbol{p})$
    return $\boldsymbol{z}$

function countTransitions($\boldsymbol{z}, \boldsymbol{L}, S$)
    Set transition counts, $\boldsymbol{T}$, to a $S$ by $S$ matrix of zeros
    Set initial state counts, $\boldsymbol{T}^0$, to a vector of zeros of length $S$
    Update $\boldsymbol{T}$ and $\boldsymbol{T}^0$ as follows:
    for each chromosome arm, $r$, do
        increment $T^0(z_{r,1})$
        for locus, $l$, to $L_r - 1$, in chromosome arm, $r$, do
            increment $T(z_{r,l}, z_{r,l+1})$
    return $\boldsymbol{T}^0, \boldsymbol{T}$
```

```
function sampleAuxVars($\boldsymbol{\beta}, \boldsymbol{T}, \alpha, \kappa, S$)
    Sample $\boldsymbol{M}$:
    for each state, $u \in \{1, \ldots, S\}$ do
        for each state, $v \in \{1, \ldots, S\}$ do
            Set $M_{u,v} \leftarrow 0$
            for $i \in \{1, \ldots, T_{u,v}\}$ do
                sample: $a \sim \text{Bernoulli}\left(\frac{\alpha\beta_v + \kappa\delta(u,v)}{i + \alpha\beta_v + \kappa\delta(u,v)}\right)$
                if $a$ is equal to 1 then
                    increment $M_{u,v}$
    Sample $\boldsymbol{W}$:
    Set $\rho \leftarrow \frac{\kappa}{\alpha + \kappa}$
    for each state, $u \in \{1, \ldots, S\}$ do
        $W_{u,\cdot} \sim \text{Binomial}\left(M_{u,u}, \frac{\rho}{\rho + \beta_u(1 - \rho)}\right)$
     Set $\bar{\boldsymbol{M}}$:
    $\bar{M}_{u,v} = \begin{cases} M_{u,v}, & u \neq v; \\ M_{u,u} - W_{u,\cdot}, & u = v \end{cases}$
     return $\boldsymbol{M}, \boldsymbol{W}, \bar{\boldsymbol{M}}$          ▷ here $\boldsymbol{w}$ represents the vector $(W_{1,\cdot}, \ldots, W_{S,\cdot})$

function sampleTransDists($\boldsymbol{T}^0, \boldsymbol{T}, S$)          ▷ Sample $\boldsymbol{\pi}^0$ and $\boldsymbol{\pi}$
    Sample the initial distribution:
    $\boldsymbol{\pi}^0 \sim \text{Dirichlet}\left(\alpha\beta_1 + T^0_1, \ldots, \alpha\beta_S + T^0_S\right)$
    Sample the transition matrix:
    for each state, $u \in \{1, \ldots, S\}$ do
        Sample row $u$ of the transition matrix:
        $\boldsymbol{\pi}_{u,*} \sim \text{Dirichlet}\left(\alpha\beta_1 + T_{u,1}, \ldots, \alpha\beta_u + \kappa + T_{u,u}, \ldots, \alpha\beta_S + T_{u,S}\right)$
    return $\boldsymbol{\pi}^0, \boldsymbol{\pi}$

function sampleCopyNumber($\boldsymbol{z}, \hat{\boldsymbol{c}}, \hat{\boldsymbol{m}}, \boldsymbol{y}, \boldsymbol{\lambda}, \boldsymbol{q}, S, \tilde{s}, \tilde{n}, \sigma_{\hat{c}}$)
    Sample the copy number of each state in turn:
    for each state, $u \in \{1, \ldots, S\}$ do
        Select the loci counts and expected proportions in state $u$:
        $Y_u = \{y_{r,l} \mid z_{r,l} = u\}$
        $\hat{M}_u = \{\hat{m}_{r,l} \mid z_{r,l} = u\}$
        if $|Y_u| \neq 0$ then          ▷ If there are loci assigned to state $u$
            Propose new value for relative copy number for state $u$:
            $\hat{c}^*_u \sim N(\hat{c}_u, \sigma_{\hat{c}})$
            if $\hat{c}_u > 0$ then
                $p_{\text{old}} \leftarrow \sum_{i=1}^{|Y_u|}\left(\log f_{BB}(Y_i; \tilde{n}, \tilde{s}\hat{c}_u\hat{M}_i, \tilde{s}(1 - \tilde{s}\hat{c}_u\hat{M}_i)\right) + \log f_{\text{Gamma}}(\hat{c}_u; \lambda_{\text{shape}}, \lambda_{\text{scale}})$
                $p_{\text{new}} \leftarrow \sum_{i=1}^{|Y_u|}\left(\log f_{BB}(Y_i; \tilde{n}, \tilde{s}\hat{c}^*_u\hat{M}_i, \tilde{s}(1 - \tilde{s}\hat{c}^*_u\hat{M}_i)\right) + \log f_{\text{Gamma}}(\hat{c}^*_u; \lambda_{\text{shape}}, \lambda_{\text{scale}})$
                $p \sim U(0,1)$
                if $\exp(p_{\text{new}} - p_{\text{old}}) < p$ then
                    Accept new value: $\hat{c}_u = \hat{c}^*_u$
        else          ▷ No loci assigned to state $u$
            Draw new value from Base Distribution:
            $\hat{c}_u \sim \text{Gamma}(\lambda_{\text{shape}}, \lambda_{\text{scale}})$
    return $\hat{\boldsymbol{c}}$
```

```
function SAMPLEHYPERPARAMETERS($\boldsymbol{T}, \boldsymbol{M}, \boldsymbol{W}, \bar{\boldsymbol{M}}, \boldsymbol{A}, \alpha, \kappa, \gamma, S$) ▷ Optionally sample hyperparameters (if fixed values not provided)
    If not provided as a fixed value, sample $(\alpha + \kappa)$:
    First, sample auxiliary variables, $\boldsymbol{h}$ and $\boldsymbol{g}$:
    for each state, $u \in \{1, \ldots, S\}$ do
        $h_u \sim \text{Bernoulli}\left(\frac{T_{u,\cdot}}{T_{u,\cdot} + \alpha + \kappa}\right)$
        $g_u \sim \text{Beta}\left(\alpha + \kappa + 1, T_{u,\cdot}\right)$
    Then sample $(\alpha + \kappa)$ as follows:
    $(\alpha + \kappa) \sim \text{Gamma}\left(A_{(\alpha+\kappa),a} + M_{\cdot,\cdot} - \sum_{u=1}^{S} h_u, A_{(\alpha+\kappa),b} - \sum_{u=1}^{S} \log g_u\right)$
    If not provided as a fixed value, sample $\gamma$:
    Set $\bar{K} \leftarrow \sum_{u=1}^{S} \mathbb{1}(\bar{M}_{\cdot,u} > 0)$, where $\mathbb{1}(A)$ represents an indicator function that is 1 if event A occurs and 0 otherwise
    Sample auxiliary variables $\zeta$ and $\tau$:
    $\zeta \sim \text{Bernoulli}\left(\frac{\bar{M}_{\cdot,\cdot}}{\bar{M}_{\cdot,\cdot} + \gamma}\right)$
    $\tau \sim \text{Beta}\left(\gamma + 1, \bar{M}_{\cdot,\cdot}\right)$
    Then sample $\gamma$ as follows:
    $\gamma \sim \text{Gamma}\left(A_{\gamma,a} + \bar{K} - \zeta, A_{\gamma,b} - \log \tau\right)$
    If not provided as a fixed value, sample $\rho$:
    $\rho \sim \text{Beta}\left(\sum_{u=1}^{S} W_{u,\cdot} + A_{\rho,c}, M_{\cdot,\cdot} - \sum_{u=1}^{S} W_{u,\cdot} + A_{\rho,d}\right)$
    Set variables:
    $\alpha \leftarrow (\alpha + \kappa)(1 - \rho)$
    $\kappa \leftarrow (\alpha + \kappa)\rho$
      return $\alpha, \kappa, \gamma$

function SAMPLEINVERSEDISPERSION($\boldsymbol{y}, \hat{\boldsymbol{c}}, \hat{\boldsymbol{m}}, \boldsymbol{L}, \boldsymbol{\omega}, \mathcal{R}, \tilde{n}, \sigma_{\tilde{s}}$)
    Propose new value for inverse-dispersion:
    $\tilde{s}^* \leftarrow N(\tilde{s}, \sigma_{\tilde{s}})$
    if $\tilde{s}^* > 0$ then
        $p_{\text{old}} \leftarrow \sum_{r\in\mathcal{R}} \sum_{l=1}^{L_r} \left(\log f_{BB}(y_{r,l}; \tilde{n}, \tilde{s}\hat{c}_{z_{r,l}}\hat{m}_{r,l}, \tilde{s}(1 - \hat{c}_{z_{r,l}}\hat{m}_{r,l}))\right) + \log f_{\text{Gamma}}(\tilde{s}; \omega_{\text{shape}}, \omega_{\text{scale}})$
        $p_{\text{old}} \leftarrow \sum_{r\in\mathcal{R}} \sum_{l=1}^{L_r} \left(\log f_{BB}(y_{r,l}; \tilde{n}, \tilde{s}^*\hat{c}_{z_{r,l}}\hat{m}_{r,l}, \tilde{s}^*(1 - \hat{c}_{z_{r,l}}\hat{m}_{r,l}))\right) + \log f_{\text{Gamma}}(\tilde{s}^*; \omega_{\text{shape}}, \omega_{\text{scale}})$
        $p \sim U(0, 1)$
        if $\exp(p_{\text{new}} - p_{\text{old}}) < p$ then
            Accept new value: $\tilde{s} = \tilde{s}^*$
    return $\tilde{s}$
```

2.8 Summarising the posterior distribution

We have now covered the algorithms that allow us to sample the posterior distribution. How can one summarise the sample of the posterior distribution and provide meaningful output to the user; a relative copy number profile? The latent variables of particular interest to us are $\boldsymbol{z}$ and $\{\hat{c}_1, \ldots, \hat{c}_S\}$. Together, these define the relative copy number profile of our sample. Most of the variables in the model can be considered as nuisance variables. For instance, we are generally not interested in knowing $\boldsymbol{\pi^0}$, $\boldsymbol{\pi}$ or $\boldsymbol{\beta}$.

The most straightforward approach is to summarise the marginal posterior distributions, $P(\hat{c}_{r,l} \mid \boldsymbol{y})$ for all r, l. This involves disregarding the state assignments and using the relative copy number values which were indirectly assigned to a locus (via $z_{r,l}$) in all iterations (after burn-in). These distributions can be summarised for each locus by computing the mean, median or MAP estimate, along with various credible intervals (CIs). The second approach is to directly use the state assignments. For instance, one might first determine the MAP state estimate, $\hat{u}$, for $P(z_{r,l} \mid \boldsymbol{y})$. Then $P(\hat{c}_{\hat{u}} \mid \boldsymbol{y})$ can be summarised by the mean, median or MAP estimate and credible intervals.

In practice, the second approach is complicated by the so-called *label switching* problem caused by nonidentifiability of the hidden states under our exchangeable prior and hence the likelihood is invariant to permutations of the parameters [141, 142, 131]. This means that a state label, u, in one iteration of the MCMC may not refer to the same entity in another iteration.

One way to attempt to resolve this issue is to take a decision theoretic approach in which we find a permutation of the state labels for each iteration of the MCMC that minimises a loss function [141]. To do this, I used Stephens' Algorithm 2 [141] which is as follows:

Starting with initial values for the permutation of the states for N iterations of the MCMC $(\nu_1, \ldots, \nu_N)$, iterate step 1 and step 2 until a fixed point is reached:

Step 1: choose $\hat{Q} = (\hat{q}_{ij})$ to minimise the following:

$$\sum_{t=1}^{N} \sum_{i=1}^{n} \sum_{j=1}^{k} p_{ij}\{\nu_t(\theta^{(t)})\} \cdot \log \left[\frac{p_{ij}\{v_t(\theta^{(t)})\}}{\hat{q}_{ij}} \right] \tag{2.38}$$

Step 2: for $t = 1, \ldots, N$, choose ν_t to minimise the following:

$$\sum_{i=1}^{n} \sum_{j=1}^{k} p_{ij}\{\nu_t(\theta^{(t)})\} \cdot \log \left[\frac{p_{ij}\{v_t(\theta^{(t)})\}}{\hat{q}_{ij}} \right] \tag{2.39}$$

As Stephens points out, step 1 can be achieved by computing $\hat{q}_{ij} = \frac{1}{N} \sum_{t=1}^{N} p_{ij}\{\nu_t(\theta^{(t)})\}$.

In this algorithm, $p_{ij}(\theta)$ denotes the classification probability that locus i belongs to state j. These probabilities are computed using the forward-backward procedure in each iteration of the MCMC in order to sample the states (see Algorithm 2). Therefore, it was straightforward to include these probabilities as part of the output files of conliga. We see that $\hat{q}_{ij}$ represents the mean classification probability across all iterations after permutation. Therefore, we see that the algorithm iteratively finds the permutations of the state labels which minimises the Kullback-Leibler divergence [143] of the mean classification probabilities (permuted) over all iterations to the true state classification probabilities.

While step 1 of the algorithm can be easily computed by determining the mean permuted probabilities, step 2 requires solving a minimisation problem. For this, I implemented the Munkres assignment algorithm (Hungarian algorithm) [144] to solve it efficiently. I implemented the Stephens' Algorithm and the Munkres algorithm in C++ and they are available as part of the conliga R package (see section 2.9 for an overview of the main functions provided in the conliga package).

Since we use a Hierarchical Dirichlet process HMM, the number of states can vary between iterations. In order to use Stephens' Algorithm in our case, I implemented two approaches. The first approach (the *MAP state* approach) was to determine the MAP number of populated states across all iterations. Then, only the iterations of the MCMC with the MAP number of states were kept. Stephens' algorithm was then run using only the filtered iterations and setting k to the MAP number of states. The second approach (the *MAX state* approach) was to find the maximum number of populated states across all iterations. Stephens' algorithm was then run using all iterations of the MCMC and setting k as the maximum number of populated states. This second approach is similar to that of Fu *et al.* [145] which was applied to relabel the clusters of a Dirichlet process mixture model.

Both of these methods provide a relabelling of the hidden states. Given the relabelled states, their associated relative copy number value was summarised by the MAP estimate (using the KernSmooth R package [133]) and credible intervals. Each locus was assigned to its MAP relabelled state and therefore was assigned the summarised relative copy number values of its MAP relabelled state. conliga outputs the results of both relabelling methods as well as the marginal posterior distribution of relative copy number explained at the beginning of this section. The relabelling process does not always work faultlessly and the Stephens' algorithm is not guaranteed to find the optimal solution. This can depend on the initialised permutations which can result in local optima being returned. Given this, conliga outputs diagnostic plots from which it can be determined if the process has provided a useful and reliable result.

To summarise, I have proposed three methods to summarise the posterior distribution:

1. Using the relative copy number assigned indirectly to each locus via the hidden state but without using the hidden state assignments directly (which I will now refer to as the *indirect* method)

2. Using the hidden state assignments and the Stephens' algorithm, the states are relabelled using the *MAP state* method

3. Using the hidden state assignments and the Stephens' algorithm, the states are relabelled using the *MAX state* method

When analysing samples, I primarily use the indirect method for most purposes. For the vast majority of loci the results obtained from all three methods are the same. If the first method shows a difference in RCN between loci, it can be useful to compare the results with the other methods which make use of the state assignments. This can show that the loci were inferred to be in separate copy number states and can rule out differences due to other reasons. One reason for a difference is that MAP estimates can vary (slightly) due to numerical error introduced in smoothing the marginal posterior distributions (using KernSmooth) to calculate MAP estimates. Another reason could be that being in one state for a few more iterations than another, the mean or median RCN summary of one locus can be shifted relative to another locus even though they may share the same MAP relabelled state. In this case we are less likely to believe this difference is due to a true difference in copy number between loci. The MAP RCN estimate is more resilient to this problem than the mean or median and so this is one of the reasons I prefer to use the MAP to summarise the RCN.

When assessing focal amplifications (particularly highly focal loci involving few loci), using the state information can be informative. For example, consider the example portrayed in figure 2.11 which shows a theoretical marginal posterior distribution for the RCN of a particular locus. In this example, the locus has a probability of $\frac{1}{3}$ for being in state one and probability of $\frac{2}{3}$ for being in state two. In the indirect summarisation approach, a MAP estimate would result in an RCN of 1 being assigned to the locus. By using the state assignments, the locus would be assigned to its MAP state (state two) and would be assigned a MAP RCN of 3. In this example, the marginal posterior distribution of the RCN for state one may be more peaked due to having more loci assigned to it and therefore more data is available for its inference in comparison to state two. If a locus has an RCN marginal posterior distribution similar to this and it is located close to or within an oncogene, we may be more willing to believe the results using the relabelled state assignments.

The credible intervals can be useful in the interpretation of the relative copy number profiles. For instance, there is often uncertainty of the CNA change points and the CIs will reflect this. This means that we can take account of this uncertainty and, if a particular locus of interest falls close to a change point, we will be able to quantify the probability on which side of the change point it falls. We will see examples of change point uncertainty in figure 2.13 within section 2.10.

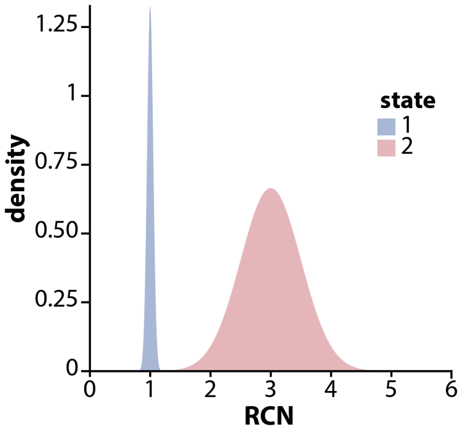


Figure 2.11 A graphical representation of a theoretical marginal posterior distribution for the RCN of a single locus. Here a locus has some probability of being in state 1 or state 2. The probability of being in state 1 is $\frac{1}{3}$, while probability of being state 2 is $\frac{2}{3}$. The *indirect* summarisation approach would result in a MAP RCN of 1 for the locus. By using the hidden state assignments (the *MAP state* or *MAX state* methods) the locus would be assigned the MAP estimate for state 2, which is an RCN of 3 in this illustration.

2.9 Implementation of conliga software package

In this section I briefly discuss the implementation of the algorithms described in this chapter within the conliga software package.

I implemented conliga as an R package [108] which is freely available under an open-source GPLv2 license at https://github.com/samabs/conliga. I implemented the two MCMC algorithms, Stephens' algorithm and the Munkres algorithm in C++, utilising Rcpp [146] and RcppArmadillo [147] to interface C++ with R. conliga expects that each sample's FASTQ file has been processed and aligned and the alignments for all samples have been summarised in an R data frame. Each row of the data frame should define a unique FAST-SeqS locus and columns define the alignment counts for each sample. Note that the first three columns define the chromosome, position and strand of each locus. conliga consists of four core functions, which are executed in order. These are:

1. `choose_loci`. This function is used to filter loci that have more than the permitted number of zero counts across a set of reference samples. For example, we may use the set of control samples as the reference samples and set `zeros_permitted` to 0. In this example, rows of the counts data frame that have a non-zero count in any of the control samples would be filtered.

2. `fit_controls`. This function implements algorithm 1. It accepts as an input: the filtered counts data frame, the set of control samples and the desired number of iterations of the MCMC (default: 20000), the number of iterations to burn (default: 5000), the

parameters of the Gamma prior on the sample specific inverse-dispersion parameter (default: 1.5, 10^6), the parameters for the Beta priors on the mean proportions for each locus (default: 1, 1 for all loci). It outputs the MAP estimate of mean proportions, i.e. $\hat{\boldsymbol{m}}$.

3. **`run_MCMC`**. This function implements algorithm 2. It takes as an input: a list of samples that the user wishes to infer the RCN profiles for, the filtered counts data frame, the MAP estimate of the mean proportions, and the cytoband information for the reference genome that was used to align the sample's reads (this can be obtained by using the `get_cytobands` function included with conliga). Optional arguments include the number of cores to use on the machine (default: 1), number of iterations (default: 50000), the thinning parameter (output every n^{th} iteration, default: 10), a data frame describing the parameters of the prior distributions (otherwise defaults are used), γ (default: 1), $\alpha + \kappa$ (default: -1, meaning to infer it from the data with a default prior distribution placed on $\alpha + \kappa$), ρ (default: -1, meaning to infer it from the data with a default prior distribution placed on ρ), maximum number of states for the weak order K limit approximation to the Dirichlet process (default: 30). The MCMC output is stored in separate files for each latent variable for each sample. Note that if the number of cores requested is more than one, the function allows each sample to be run in parallel (an instance of the algorithm for each sample) with the help of the `parallel` R package.

4. **`process_MCMC`**. This function summarises the MCMC for a given set of samples. It summarises the MCMC in the three ways described in this chapter. It outputs a tab-separated values (TSV) file for each sample, providing summarised RCN profiles. In addition, it provides relative copy number profile plots and label switching diagnostic plots.

Within the conliga software, I allow for the option to provide fixed values for the hyperparameters or provide priors and allow them to be inferred from the data. In practice, I find that it is important to fix γ and provide strong priors on ρ and $(\alpha + \kappa)$ or fix them. If we allow γ to be inferred from the data, it tends to overfit the number of states, leading to an unrealistically high number of inferred states. This is likely because the emission distribution is not an exact representation of (and only an approximation to) the true emission distribution. By allowing γ to vary, the model is able to improve the fit by introducing additional copy number states. I find that fixing γ to a value around 1 solves this issue. In addition, given the distance between loci and the typical lengths of SCNAs, we know that the probability of self transition is likely to be high. As such, values of 0.99 for ρ and 200 for $(\alpha + \kappa)$ appear to lead to sensible inferences.

2.10 Application to male and female samples

Having explained the development of the model and the observations that justified it (section 2.5), described the two inference algorithms (sections 2.7.2 and 2.7.4) and how the resulting posterior distribution can be summarised (section 2.8), we now look at a simple application. In this section, we evaluate the performance of conliga in determining the relative copy number differences between male and female samples.

For this experiment and all others described in this thesis, I used the following values for the hyperparameters:

- $\gamma = 1$
- $\text{Gamma}(2000, 10)$ prior distribution (defined by shape and scale) was placed on $(\alpha + \kappa)$
- $\text{Beta}(100000, 100)$ prior was placed on ρ
- $\text{Gamma}(3, 1)$ prior distribution (defined by shape and scale) was placed on the relative copy number value of the hidden states; the shape and scale parameters are defined by λ in equation 2.26
- $\omega_{shape} = 1.5$, $\omega_{scale} = 10^6$; where ω_{shape} and ω_{scale} define the shape and scale of the Gamma prior distribution on $\tilde{s}$, respectively

Given $\hat{\boldsymbol{m}}$, inferred using the male samples and the loci aligned to chromosomes 1-22, X and Y, and the loci counts for five female samples and the eight male samples, I ran the `run_MCMC` function from the conliga R package (which implements algorithm 2). The output of the MCMC chains were summarised (`process_MCMC` function in conliga) using the three methods described in section 2.8. Figure 2.12 shows the summarised output using the MAP estimates obtained using the indirect method to summarise the posterior distribution (the first approach discussed in section 2.8). The results show that in all female samples, the increased RCN in chromosome X is correctly identified as is the relative deletion (or absence) of chromosome Y. In five selected male samples, we observe an RCN of one in chromosomes X and Y.

In this figure, I have included the scaled counts. The scaled count is defined as $\hat{y}_{r,l} = y_{r,l}/(\tilde{n} \cdot \hat{m}_{r,l})$. Each scaled count is coloured by its associated $\hat{m}_{r,l}$ value. We see that scaled counts with low $\hat{m}_{r,l}$ values are generally more dispersed around the inferred RCN value, whereas those with high values are more closely clustered around the inferred RCN. This highlights the fact that loci with higher counts have greater statistical strength. As I have mentioned previously, conliga accounts for this varying statistical strength by the use of independent beta-binomial distributions. The HMM enables this statistical strength to be propagated to neighbouring loci. If this wasn't accounted for, the noise generated by the increased sampling error associated with low counts could lead to increased false positive calls.

In some samples (particularly the second and fifth from the top), it is noticeable that scaled counts with high or low $\hat{m}_{r,l}$ values are not symmetrically distributed around the inferred RCN. We note that this does not seem to affect the inference of the RCN. However, we shall return to this observation in Chapter 4.

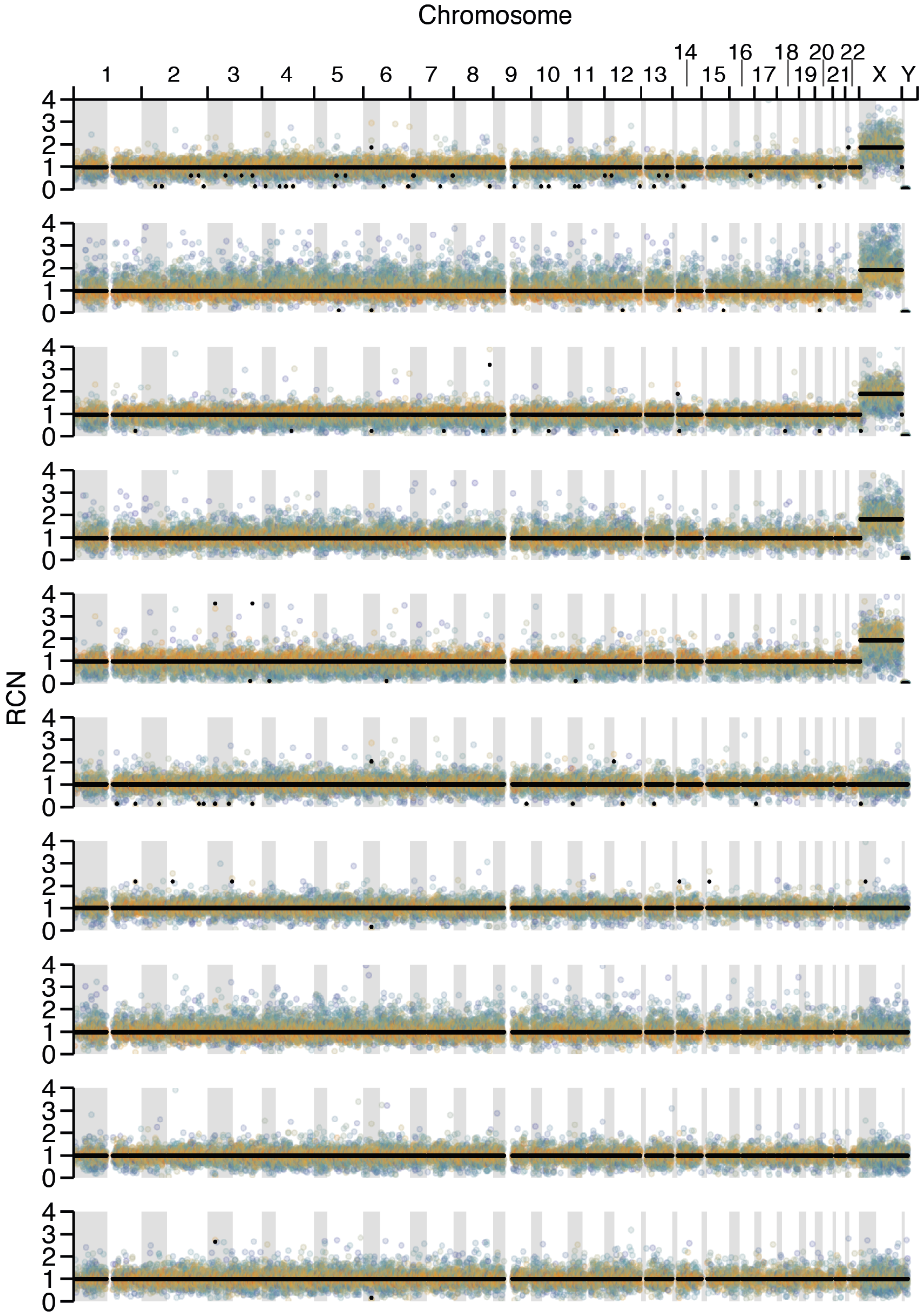
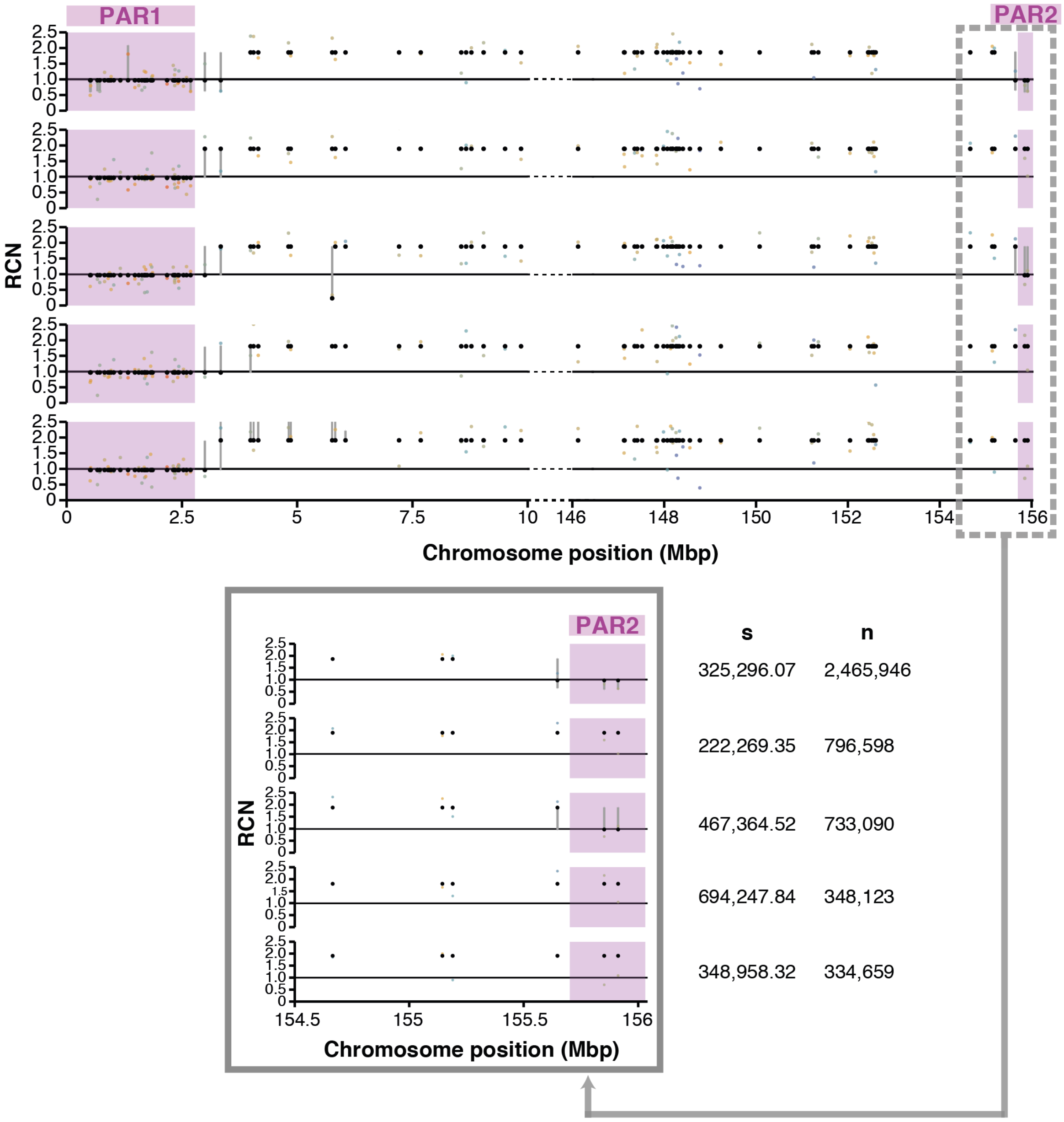


Figure 2.12 Inferred RCN profiles for 5 female samples and 5 male samples across chromosomes 1-22, X and Y. Black points represent the MAP RCN obtained using indirect summarisation method. Coloured points represent the scaled count $(y_{r,l}/(\tilde{n} \cdot \hat{m}_{r,l}))$ and are coloured by the value of $\log(\hat{m_{r,l}})$ where high values are red and low values are blue.

Looking across the inferred RCN profiles for the 10 samples, we observe short RCN changes usually involving one locus or very few loci. These may be true CNVs within normal samples, reflect false positive calls or perhaps be a mixture of the two. Indeed, short RCN segments are inferred for several samples within the p-arm of chromosome 6. These changes occur within the major histocompatibility complex (MHC) which frequently harbour variants [148]. This suggests that some of the short RCN segments may reflect true CNVs. While some of the inferred short RCN segments may be true CNVs, some could be false positive calls. There could be several possible causes of false positives. For example, the beta-binomial emission distribution is an approximation and is unlikely to reflect the true distribution of counts. As such, an imperfect emission distribution will result in false positive calls. Short changes are more frequent in samples with greater total read counts and show that with increased statistical strength, more of these regions are called. This would be the case if the short RCN segments were true CNVs or were false positives inferred due to an imperfect emission distribution. In addition, some of the false positives could be a consequence of the Markovian assumption of the HMM. The Markovian assumption implies that the lengths of copy number states (the state duration) follow a geometric distribution [149]. As such, the model encourages the inference of short RCN segments. This issue is discussed further in Chapter 6.

Figure 2.13 shows the inferred RCN profiles for the five female samples at the ends of chromosome X, including the PAR1 and PAR2 regions. Recall that the PAR1 and PAR2 regions are present on both chromosomes X and Y but are masked in chromosome Y within the reference genome. This means that male and female cells will have two copies of the PAR regions and that reads from these regions will be aligned to chromosome X. Therefore, the relative copy number within the PAR regions in the female samples should be the same as that of the autosomes.



s	n
325,296.07	2,465,946
222,269.35	796,598
467,364.52	733,090
694,247.84	348,123
348,958.32	334,659

Figure 2.13 RCN profile for 5 female samples at the 5’ and 3’ ends of chromosome X. Top: RCN profiles for 10 Mbp of the 5’ and 3’ chromosome X ends with the PAR1 and PAR2 regions highlighted in purple. Bottom: Zoomed-in region of the 3’ end of chromosome X more clearly showing the status of the two FAST-SeqS loci within PAR2. To the right of the bottom plot are the MAP estimates of the inverse dispersion parameter ($\tilde{s}$) and the total number of aligned reads for each sample ($\tilde{n}$).

We see that the PAR1 region was inferred to have a MAP RCN of just below one in all female samples. This is the same RCN as was called for the autosomes and shows that the RCN for this region has been correctly inferred. As I discussed in section 2.8, the uncertainty in the RCN change point is reflected by the 90% credible intervals. Here, the results suggest that the change point could occur within or between two or three loci close to the PAR1 region. While the PAR1 region is approximately 2.7 Mbp in length, the PAR2 region is much shorter ($\sim$ 330 kbp) and contains only two loci. Here we see that conliga was able to correctly infer the RCN for two out of the five samples. These two samples have high total counts which, together with a sufficient inverse dispersion, provides the required statistical strength to call a short RCN region. In contrast, another sample with high total count has a lower inferred inverse dispersion and, as such, conliga was unable to infer such a short RCN segment for this sample. These results suggest that there is a trade-off between the total number of reads and the sensitivity but that technical variation also plays a role. The total number of reads is a function of the sequencing cost and so greater sensitivity could be achieved by sequencing more molecules per sample, while the technical variance could be controlled by optimising the experimental protocol.

2.11 Analysis of convergence and state relabelling diagnostics

Determining whether an MCMC algorithm converges, that is that the samples are representative of the underlying stationary distribution of the Markov chain, is not straightforward and is often not possible to determine with certainty [150]. However, there are several methods which can suggest convergence or at least diagnose non-convergence [134]. One simple approach is to run several chains with random starting points and compare the samples (or summaries) of particular variables of interest across the runs. If the runs produce the same result, we have good evidence of convergence [134].

I ran four chains, that were randomly initialised, for 50,000 iterations for a female sample presented in section 2.10. The first 5,000 iterations were discarded and the chains were thinned such that every 25th iteration was kept. This left 1,800 samples from each Markov chain. Figure 2.14 shows various summaries of these four chains. We see similar marginal posterior distributions for the inverse dispersion parameters (figure 2.14A), similar posterior distributions for the number of copy number states (figure 2.14B), similar MAP RCN estimates for all loci across two randomly chosen runs using the *indirect* summarisation method (figure 2.14C) and similar quantiles of the marginal posterior distributions of the loci RCN (figure 2.14D). We find good concordance between runs and these results suggest convergence. Further evidence of convergence is shown in figure 2.15. figure 2.15A shows that the four randomly initialised chains rapidly converge to a similar log-posterior probability range. Figure 2.15B shows the log-posterior traces for the four chains after burn-in and suggests that the chains explore the same log-posterior space. Furthermore, figure 2.15C shows that the number of loci assigned to each relabelled state is consistent between the four chains. This provides further evidence that the chains are converging to the stationary distribution. How quickly a Markov chain converges and how well it mixes may depend on the sample. In general it is advisable to run several chains for each sample and check concordance. In some cases it may be necessary to run more than 50,000 iterations. However, for the majority of cases, I find 50,000 iterations to be sufficient.

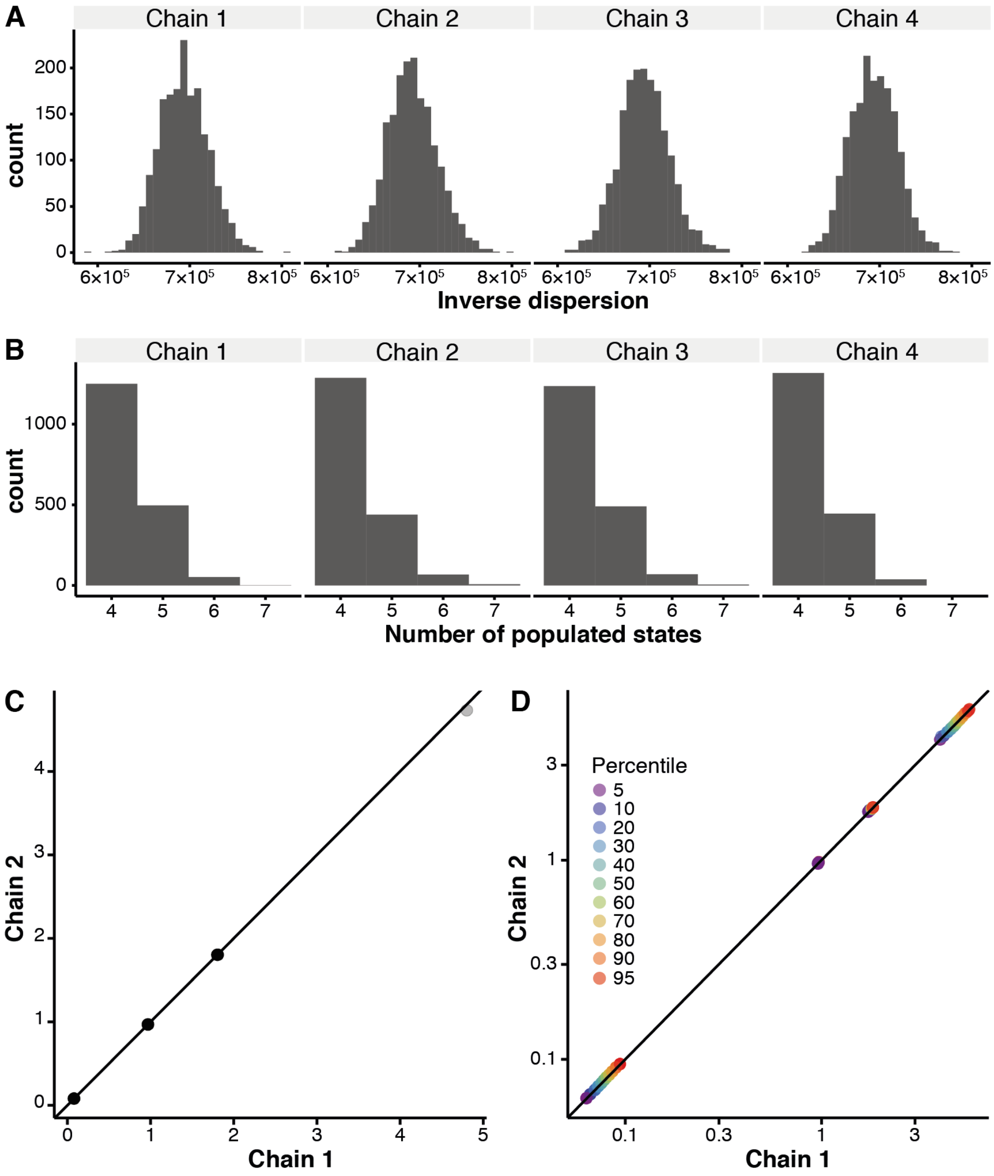


Figure 2.14 Assessing convergence: comparison of marginal posterior distributions obtained by running four chains initiated at random starting points for a female sample. A: Marginal posterior distribution of sample inverse dispersion parameter ($\tilde{s}$ from the four chains. B: Marginal posterior distribution for the number of copy number states (the number of states that had loci assigned to them over the iterations of the MCMC). C: Comparison of MAP estimates (using the indirect method) for the relative copy number of 12,834 loci between chain 1 and chain 2. D: Comparison of the percentiles of the marginal posterior distribution for the relative copy number copy number of all loci using the indirect method (the plot contains a total of $12843 \times 11 = 141174$ points).

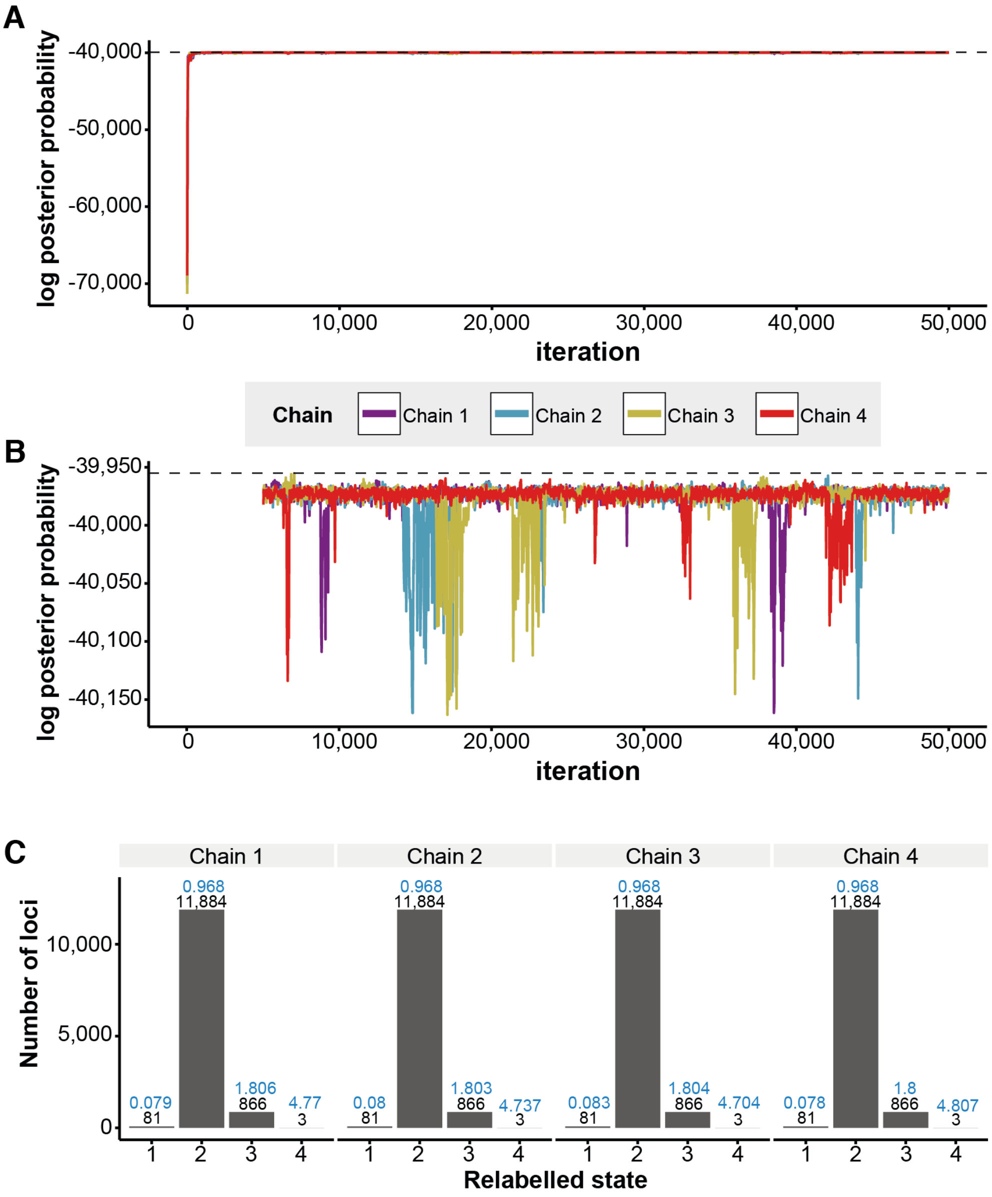


Figure 2.15 Assessing convergence: log-posterior trace plots for four chains initiated at random starting points and assessment of the number of loci assigned to each state. A: The log-posterior probability trace plots for four randomly initialised chains showing what appears to be rapid convergence to the stationary distribution. The dotted line indicates the highest log-posterior probability observed across all chains. B: The log-posterior probability trace plots (zoomed in) after discarding the first 5,000 iterations for the same four chains shown in panel A. C: Bar charts representing the number of loci assigned to each relabelled state (after relabelling the states with the *MAP state* method) for each chain. The number of loci assigned to each state is shown in black above the bars. The MAP RCN associated with each state is shown in blue. To make the bar charts comparable between chains, the states were ordered in terms of their associated MAP RCN.

Figure 2.16A shows a trace plot for the relative copy number of each populated state across the 50,000 iterations. Here we can see evidence of label-switching which we described in section 2.8. Figure 2.16B shows the relabelling of the states by the *MAP state* method and indicates a successful relabelling of the states. Figure 2.16C shows the relabelling of the states by the *MAX state* method. In this case, the relabelling of the states appears to have been unsuccessful.

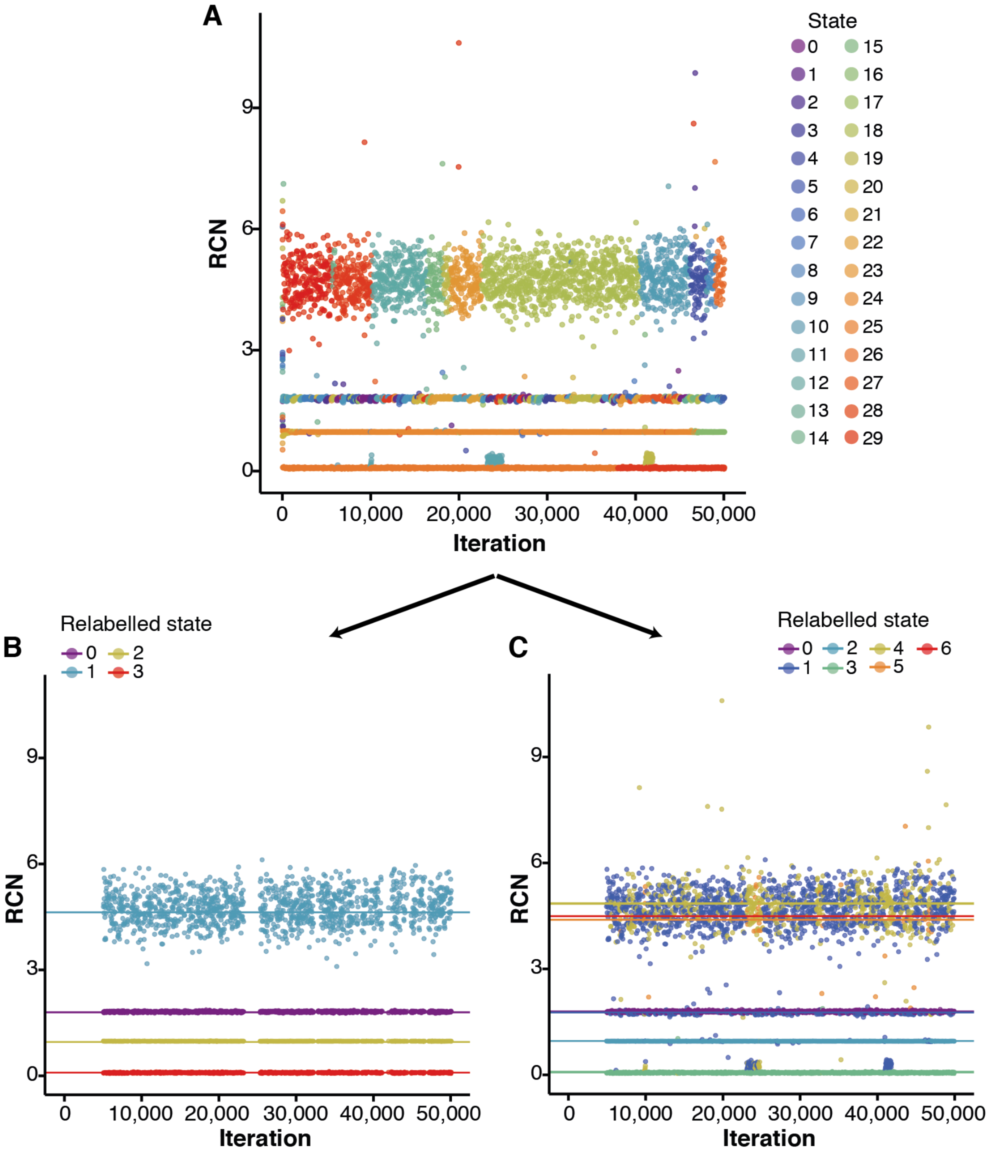


Figure 2.16 Example MCMC trace plot for the RCN of the hidden states for a female sample. A: Raw trace plot for all populated states across all 50,000 iterations. B: Trace plot of relabelled states using the *MAP state* method. C: Trace plot of relabelled states using the *MAX state* method.

2.12 Summary

In this chapter, I have described the FAST-SeqS protocol and critically examined the previously published computational methods for FAST-SeqS data. By exploring the FAST-SeqS data obtained from samples of the normal oesophageal epithelium from male and female donors, I have motivated and developed a probabilistic generative model for FAST-SeqS data. In addition, I described the associated inference procedures to infer the biases in the FAST-SeqS protocol and infer RCN profiles for each sample of interest, These algorithms are implemented in an R package called conliga, which is freely available and open-source. By using conliga and the set of male samples to infer the bias, I showed that appropriate RCN profiles were inferred for female and male samples. In the next chapter we explore the application of FAST-SeqS and conliga to oesophageal adenocarcinoma and its associated premalignant lesion, Barrett's oesophagus.

Chapter 3

Application to oesophageal adenocarcinoma and precursor lesions

Some of the work presented in this chapter forms a subset of the work described in Abujudeh *et al.* [108] and is shared as a preprint on bioRxiv. In some sections of this chapter, parts of the paper are reproduced. This work was performed in collaboration with the Fitzgerald Laboratory at the MRC Cancer Unit at the University of Cambridge. Sebastian Zeki and colleagues performed the FAST-SeqS protocol and were responsible for processing the clinical samples through to data (FASTQ files). Jamie MJ Weaver conceived the clinical utility of FAST-SeqS for upper gastrointestinal (GI) cancers. Lawrence Bower designed the pipeline for processing high-coverage WGS and obtaining allele-specific copy number calls from ASCAT. The rest of the work presented in this chapter is my own.

In this chapter, I discuss the challenges associated with oesophageal cancer and the need for its early detection. In particular, the chapter focuses on oesophageal adenocarcinoma (OAC) and its associated pre-malignant lesion, Barrett's oesophagus. I explore the limitations of the current screening and surveillance programmes, describe low-cost alternatives, and motivate the use of a low-cost SCNA assay to identify chromosomal instability, which is associated with OAC progression. As a proof of principle, FAST-SeqS and conliga are applied to OAC samples that were previously sequenced as part of the International Cancer Genome Consortium (ICGC) oesophageal adenocarcinoma project and conliga's performance is compared with ASCAT and QDNAseq. FAST-SeqS and conliga are applied to Barrett's oesophagus (BO) and gastric adenocarcinoma (GAC) samples that were not previously sequenced and regions of interest are explored.

3.1 Introduction and motivation

3.1.1 Oesophageal cancer: global incidence, mortality, development and risk factors

In 2018, oesophageal cancer accounted for an estimated 572,034 cancer cases out of 18,078,957 (3.16%) worldwide; ranking as the ninth most common cancer. It is the sixth most common cause of cancer-related death; responsible for 508,585 out of 9,555,027 (5.32%) cancer-related deaths globally [1]. Oesophageal cancer largely comprises two distinct histological types: oesophageal squamous cell carcinoma (OSCC) and oesophageal adenocarcinoma (OAC). While these two diseases affect the same organ, they have striking differences in their epidemiology and risk factors [151, 152].

OSCC comprises almost 90% of oesophageal cancer cases globally and is particularly prevalent in East Africa, Central and East Asia [151]. It develops from squamous epithelial cells, which form the lining of the oesophagus, via morphological steps; squamous hyperplasia and dysplasia (low to high-grade), carcinoma in situ through to invasive OSCC [153, 152]. Repeated physical and chemical injury to the mucosa is associated with the risk of developing OSCC, although the risk factors depend on geographic location. In areas with relatively low incidence, tobacco smoking and alcohol consumption appear to be strongly associated with developing disease. However, in areas with high incidence such as Linxian (within Shanxi province, China), smoking has mild effect (1.33-fold increase risk vs 9-fold increase risk in the United States) [154, 155, 152]. Other associations include age, low consumption of fruit and vegetables, nutritional deficiencies, positive family history of OSCC (although relatively few genetic predispositions are known) and injury caused by consumption of high-temperature beverages. Many of these factors are associated with socioeconomically deprived regions and groups. Moreover, access to water piped into the home, increased body mass index (BMI), consumption of eggs, fruits and meat were inversely associated with developing OSCC in Linxian [154, 152]. Perhaps owing to improved socioeconomic conditions and a reduction in alcohol consumption and tobacco smoking, the incidence of OSCC has fallen in many countries. For example, rates fell annually by 3.6% in the USA from 1998 to 2003 and 3.3% age-adjusted annual falls were observed in China from 1989 to 2008 [156, 157, 152].

OAC differs in its geographical incidence distribution in comparison to OSCC with relatively high incidence in North America, Northern and Western Europe and Oceania [152]. In stark contrast to OSCC, the incidence of OAC has risen rapidly in many western industrialised countries [158]. In the United States for example, the rate of increase of OAC was greater than any other cancer type between the 1970s and 1990s with a >350% rate of increase reported [159, 156]. Similar and sustained trends have been observed in other countries during the same time [158]. Indeed, OAC has overtaken OSCC to become the predominant cancer type affecting the oesophagus in Australia, Canada, Iceland, Ireland, New Zealand, Netherlands, Norway, Sweden, UK and the USA among others [151]. OAC disproportionally affects men, with 7 out of 8 patients being male [152]. Given incidence trends, it is predicted that one in 100 men in the UK and Netherlands will be diagnosed with OAC by the age of 75 in 2030 [151].

Like OSCC, OAC generally develops from a pre-malignant lesion. OAC typically develops from Barrett's oesophagus (BO), a metaplastic condition in which specialised columnar mucosa replace the native squamous cell mucosa. BO tends to occur near the gastro-oesophageal junction and is thought to form in response to inflammation and damage due to gastro-oesophageal reflux disease (GORD) [160]. Indeed, GORD is the most important risk factor for the development of Barrett's and OAC (odds ratio of 3.48 and 12 respectively [161]). Obesity, particularly central or visceral obesity, is the second most important risk factor [162, 152]. Tobacco smoking, male sex, Caucasian race, age, and dietary factors (low vegetable and fruit consumption and high red meat consumption) and genetic factors also increase the risk of OAC development [163, 152, 164].

3.1.2 The prognosis and staging of oesophageal cancer

Both OSCC and OAC carry very poor prognosis with median survival of less than 1-year [152]. Indeed, oesophageal cancer is among the cancer types with the lowest survival rates [165, 166]. Poor survival rates are seen around the world, even in developed countries, with an overall five year survival of 19% in the United States [167] and a 42% one year, 15% five year and 12% 10 year survival rate in England and Wales [168]. The five year survival rate in England and Wales is on a par with the average in Europe [165] and only modest improvements have been made in survival rates in the last 40 years (from 4% to 12% 10 year survival rates in England and Wales) [166].

For OAC and OSCC, the most critical factor in determining prognosis is the stage at which it is detected; the later the stage, the poorer the prognosis [169]. The stage classifies the progression of cancer and, in oesophageal cancer, is determined by the tumour-node-metastasis (TNM) system, in which information about the size of the primary tumour and its invasion of nearby tissue (T), the local lymphatic involvement (N) and the presence of distant metastases (M) are combined to provide an overall stage classification [170]. Staging classification depends on the mode of staging. The first mode is clinical staging which is based on imaging studies with limited resolution (cTNM), pathologic (pTMN) which includes histological information gathered from the microscopic examination of resection specimens and finally, pathological staging post neoadjuvant therapy (ypTMN) [170].

Symptoms of oesophageal cancer present relatively late in the course of the disease compared with many other cancers. It is only at a relatively advanced stage that obstruction or restriction of the oesophagus occurs. Prior to this point, oesophageal cancer is generally asymptomatic and clinical examination is unlikely to discover evidence of disease [169, 152]. Because of this, the majority of patients are diagnosed at a late stage and after symptoms present, with locally advanced or metastatic disease. For example, 70-80% are diagnosed with late stage cancer (stage III or IV) in the UK [168]. At this point in the cancer development, the majority of cancers have invaded the muscularis propria (T2) and spread to local lymph nodes (N1) [169] and around 37-42% have metastasised to distant organs (M1) [168].

3.1.3 The current stage-stratified treatment of oesophageal cancer

Current treatments are more effective at earlier stages in the disease. As such, the five year survival rate for patients caught with early stage OAC, i.e. high-grade dysplasia (HGD) or intramucosal adenocarcinoma (IMC, T1a), is approximately 90% (though lead-time bias may relatively inflate survival rates to some extent) [171, 169]. Malignant tumours confined to the mucosa (T1a) can be successfully removed with endoscopic mucosal resection (EMR) or submucosal dissection (ESD) along with radiofrequency ablation techniques [152]. Patients with HGD and IMC that are treated with these methods have low OAC recurrence; with recurrent disease seen in approximately 0.7-1.4% of cases [172]. The progression of HGD to OAC can be prevented by ablation and shows similar long-term survival and fewer complications compared with more radical surgical procedures [173, 163]. Ablation therapy is recommended for patients with low-grade dysplasia (LGD) and reduces the number that go on to develop HGD and OAC [174, 163]. However, ablation treatment for patients with non-dysplastic Barrett's oesophagus (NDBO) is not warranted given the low rates of progression to OAC with the side effects of treatment outweighing the benefit [163, 152]. Occasionally, patients with OAC invading the submucosa (T1b) can be treated by endoscopic therapies but oesophagectomy is the current treatment standard. This is because tumours that have spread to the submucosa have reasonably high risk (17%) of metastasising to lymph nodes, at which point treatment becomes more difficult and prognosis worsens [175, 172]. Indeed, the depth of invasion into the submucosa affects the risk of lymph node metastasis (varying from 9% to 50%) [176, 177]. Fewer studies exist for the prevention of disease progression or treatment of early stage OSCC with endoscopic treatments. However, higher rates of progression and complications (e.g. stricture) are seen in patients with squamous dysplasia when treated with ablation therapy compared to patients with Barrett's oesophagus [178–180, 152].

For patients with more locally advanced disease ($\geq$T2) or those that have lymph node involvement, surgery (oesophagectomy and potentially lymphadenectomy) is the main mode of treatment [152]. However, surgery alone offers poor survival; even patients at relatively early stage (T1b) have only a 50% 10 year survival when surgery is the only mode of therapy (for OAC and OSCC) [181, 152]. For these patients, adjunctive treatment is required and can involve neoadjuvant or perioperative chemotherapy, radiotherapy or chemoradiotherapy. Indeed, adjunctive treatment can be used to reduce tumour size prior to surgery in order to increase the chances of curative resection and reduce risk of disease recurrence. Treatment choices vary depending on clinical factors and disease type. For example, surgery may not be suitable for all patients given other comorbidities or old age and other treatments may be sought. Indeed, oesophagectomy can have a severe impact on quality of life and is itself associated with increased morbidity. Fortunately, some patients with OSCC can be treated with radiotherapy or chemoradiotherapy and achieve good response without surgery but typically not those with OAC [182, 152].

A substantial number of patients are diagnosed with metastatic disease (M1) for which curative therapy is not currently possible. Moreover, more than 50% of patients that are considered to have curative disease and are treated as such, will eventually have disease recurrence that is not curable. These patients will be limited to palliative care and therapy which include radiotherapy, chemotherapy and stents [182, 152].

3.1.4 The need for screening and surveillance of oesophageal cancer development and associated precursor lesions

Given the clear survival differences between those patients with early and late stage disease and that treatments are much more effective for early stage disease, progress must be made to detect oesophageal tumours earlier in their development and before the onset of symptoms. Indeed, detection and removal of precursor lesions prior to the progression to malignant disease is preferred given the risk of spread in relatively early disease.

Detection of malignancy or precursors prior to the onset of symptoms necessitates the screening and surveillance of individuals. In this context, screening is considered as a single, one-off test to detect a disease or identify individuals at risk of developing the disease within the general population. Surveillance is considered to be repeated application of a test, most often administered to individuals deemed to be at greater risk of developing the disease than the general population. In both screening and surveillance, the aim is to identify disease early so that interventions can be made to improve prognostic outcomes [169].

Successful mass screening programmes have been deployed for the early detection of breast, colorectal, lung and uterine cervical cancer and have been shown to lower mortality rates [183, 184]. Furthermore, the screening and surveillance for cervical and colorectal cancer development has shown efficacy in the detection and subsequent removal of pre-malignant lesions before they can develop to cancer [183, 184].

For a screening or surveillance programme to be successful there are multiple factors to consider. Firstly, the test employed needs to have an acceptable level of sensitivity (true positive rate) and specificity (true negative rate). In addition, the test needs to be acceptable to the individual receiving it, not be dangerous to the individual's health and be cost effective [185, 186]. Clearly then, a perfect test is 100% sensitive and specific, unnoticeable and risk free to the recipient, requiring zero cost to deliver and ultimately leading to 100% prevention or cure of the disease.

Despite OSCC and OAC having known precursor lesions that may be successfully treated with endoscopic therapy, there are currently no mass screening programmes for oesophageal cancer in the West [152]. The British Society of Gastroenterology and American College of Gastroenterology currently suggest that screening for Barrett's oesophagus could be considered for those at high-risk of developing the condition (male sex, GORD $\geq$5 years, family history and multiple other risk factors) by endoscopy [177, 163, 152]. Endoscopic surveillance of patients with a known Barrett's oesophagus diagnosis is recommended and surveillance intervals depend on a number of factors, including the length of the Barrett's segment, whether intestinal metaplasia (IM) is present and the presence and level of dysplasia [177]. If LGD or HGD is identified, endoscopic therapy (ablation) is recommended along with continued endoscopic surveillance [177, 163, 187].

The screening and surveillance process can be split into two conceptual parts. The first part involves a process by which tissue samples and/or images of the oesophagus are taken. Samples could also be taken from other sources, e.g. blood, saliva or urine. The second part involves the analysis of measurable indicators (a biomarker, or set of biomarkers) within the sample or image. These biomarkers are used to distinguish between normal and diseased tissue and to stratify patients and make treatment decisions. Currently, the first part is performed by endoscopy in which images and biopsies of the oesophagus are taken. The second part involves clinical observations (e.g. presence of salmon-coloured mucosa, length of Barrett's segment) and histopathological analysis of the biopsies (e.g. level of dysplasia, presence of IM, evidence and depth of invasive OAC) [177, 163].

3.1.5 The limitations of endoscopic and histopathological-guided screening and surveillance

There are a number of problems with the current screening and surveillance paradigm. Firstly, 90-95% of patients diagnosed with OAC do not have a previous diagnosis of Barrett's oesophagus [188, 164, 189, 163]. These patients are missed by the current surveillance strategy. Furthermore, only 0.12-0.6% of patients with Barrett's oesophagus progress to OAC [152]. Most patients with Barrett's die from other causes, suggesting that Barrett's may not be a major predictor for OAC. In a meta-analysis involving 7,930 patients with Barrett's, only 88 out of 1,271 deaths were due to OAC. Within the subset of studies that reported mortality cause, 64 out of 921 (7%) deaths were due to OAC while 93% were due to other causes.

Indeed, more patients died from other malignancies (16%) than OAC, while other deaths were due to cardiac disease (35%) and pulmonary disease (20%) [190, 163].

The presence of IM and the grade of dysplasia is used as a biomarker for progression to OAC and patient stratification [177, 163]. Meta-analyses have shown that non-dysplastic Barrett's progresses to OAC with an annual rate of 0.33% [191] and that LGD has an annual progression rate of 1.7% to HGD/OAC (combined) and 0.5% for OAC alone [192]. However, the progression of LGD varies considerably between studies and Shaheen *et al.* [163] suggest that this variation can be explained by more liberal LGD diagnoses (and perhaps overcalling LGD) in some studies. As such, the risk of progression to LGD may be overestimated [163]. Variation between progression rates are also seen in studies of progression from HGD to OAC; rates vary from less than 10% to 59% per year [164]. Crude and weighted annual incident rates of 6% and 7% respectively were estimated in a meta-analysis of patients with HGD from four studies [193]. However, incident rates of approximately 19% were found in two randomised trials that included surveillance of subjects with HGD and required confirmation of HGD by another expert pathologist [194, 195]. These results suggest that with more thorough histopathological validation, the level of dysplasia may be more predictive of OAC development [163]. However, these results also highlight that the grading of dysplasia is subjective. Indeed, studies have shown substantial variation between the assessment of pathologists [196, 197, 164]. Another issue is that endoscopic sampling error is high and a large number of endoscopic biopsies are required to minimise it [164]. For example, a retrospective study found that when four biopsies were taken IM was detected in 35% of cases but when eight biopsies were taken IM was detected in 68% [198, 163]. Even with pathologist variation and sampling error aside, there is likely to be biological heterogeneity within discrete grades of dysplasia [164] and further biomarkers are needed for more precise risk stratification [152].

Standard transoral endoscopy is the gold standard diagnostic tool for oesophageal cancer and precursor lesions [186, 177]. However, it is expensive, invasive and requires skilled operators [186, 163, 152]. In 2012, standard white-light endoscopy was estimated to cost £520 in the UK and $930 in the US per patient [186]. Histopathological analysis of at least four biopsies increases these costs further. Indeed, due to inter-pathologist variability it is recommended that dysplasia diagnoses be confirmed by independent expert pathologists [177] which increases costs and puts additional burden on healthcare systems. Given the costs and logistics, endoscopic screening and histopathological-led diagnosis of the whole population is not cost-effective and not recommended [177, 163]. Indeed, the screening of patients with GORD alone is highly controversial and the cost-effectiveness is uncertain [177]. It is for this reason that additional risk factors plus GORD are used to select patients for screening [177, 163]. Indeed, the cost-effectiveness of endoscopic surveillance for patients with Barrett's oesophagus is debated and there is insufficient data to prove its cost-effectiveness [177].

The costs and logistics of endoscopic screening limit the screening population to high-risk individuals, however the criteria used to select patients is suboptimal. Patients with Barrett's oesophagus do not necessarily have reflux symptoms. In a random sample of adults from the Swedish population, 16 (1.6%) were found to have Barrett's and of those 9 reported having reflux symptoms. The prevalence of Barrett's in patients with reflux was 2.3% compared with 1.2% with those without reflux [199]. Similar results were seen in a random sample of adults in two Italian villages with 1.3% found to have Barrett's oesophagus and 46.2% of those with Barrett's reporting no symptoms of reflux [200]. Moreover, many patients with OAC do not report typical reflux symptoms. In a nationwide case-control study in Sweden, 76 out of 189 patients with OAC (40%) reported that they did not have regular symptoms of reflux [201]. In a study in the United States, 72 out of 198 OAC patients (36%) reported never having reflux symptoms and 24 (12%) reported having infrequent reflux symptoms (1-12 times/year) [202]. These patients are likely to be missed by the current screening recommendations.

3.1.6 Low-cost alternatives to standard endoscopy and histopathological-guided screening and surveillance

Given the limitations of standard endoscopy and histopathological-guided diagnoses, alternatives are sought. Lower cost and less invasive devices to capture samples are required to screen and surveil a greater proportion of the population. Furthermore, biomarkers that can predict oesophageal cancer development with high sensitivity and specificity are also necessary along with associated low-cost assays to measure them.

Unsedated transnasal endoscopy (TNE) is a less invasive alternative to standard endoscopy. In a study of 95 randomised patients, it was shown to provide a sensitivity and specificity of 98% and 100% respectively for the detection of Barrett's oesophagus when compared with standard endoscopy [203]. However, further studies with larger patient cohorts are required before it be considered to replace standard endoscopy for screening [177]. Furthermore, biopsies taken using TNE are significantly smaller than standard endoscopy [203] which may lead to increased sampling error and difficulty in diagnosing dysplasia grades. In addition, substantial investment would be required in new equipment and skilled operators [152]. Hence, TNE is not recommended for Barrett's surveillance at this time [177].

Capsule endoscopy provides another alternative to standard endoscopy. It involves the ingestion of a small device which can record images from the oesophagus. However, capsule endoscopy has low sensitivity and specificity for the detection of Barrett's oesophagus (73% and 78% respectively) and is therefore not recommended for screening [163, 152]. In addition, tissue sampling is not possible so follow-up by standard endoscopy would be required to take biopsies of any suspicious lesions [186].

Non-endoscopic devices have shown promise for the detection of Barrett's oesophagus. A number of cytology-retrieval devices have been developed to sample tissue from the oesophagus [186, 152]. These devices are simple and low-cost technologies including inflatable balloons and sponges [204, 205, 100]. The Cytosponge is an example of one of these devices. It takes the form of a polyurethane sponge contained within a gelatin capsule that is attached to a string. After the capsule is swallowed and reaches the stomach, the capsule dissolves and the sponge expands. The sponge is then retrieved by pulling on the string and it collects cells as it makes its way back through the oesophagus [100]. It has been shown to be well tolerated by those that have used it [206]. When immunocytochemistry for trefoil factor 3 (TFF3) is performed on the specimens obtained by the Cytosponge, Barrett's oesophagus can be detected with a sensitivity and specificity of 79.9% and 92.4% [207]. Sensitivity improves for longer segments of Barrett's oesophagus and when patients take two Cytosponge tests [207].

While this test offers a low-cost alternative to endoscopy for the detection of Barrett's oesophagus, TFF3 does not enable risk-stratification of patients. In a theoretical population of 10,000 patients with reflux symptoms and an assumed 3% prevalence of Barrett's oesophagus that are given the Cytosponge plus TFF3 test, the expectation is that there would be 237 true positives, 737 false positives, 8963 true negatives and 63 false negatives [207]. Without further risk-stratification, 974 out of 10,000 patients will require follow-up endoscopy. In the majority of cases, the follow-up endoscopy will be needless. In addition to this, 63 will be missed and might go on to develop OAC. It is clear that without further risk-stratification, the number of patients requiring endoscopy following a TFF3 test would create a substantial burden on healthcare systems [152].

The expression of p53 as a biomarker for risk-stratification has been investigated. In a endoscopic surveillance study of 720 patients with BO, 49 patients developed HGD or OAC. Aberrant p53 protein expression (determined by immunohistochemistry) was associated with increased risk of progression to HGD or OAC; overexpression and loss of p53 expression were shown to have a relative risk (after adjusting for other factors) of 5.6 and 14.0 respectively. Indeed, these results suggest that aberrant p53 expression is a better predictor of progression than the histopathological diagnosis of LGD [208].

Other studies have looked at mutations within genes as markers of OAC progression. Weaver *et al.* [209] used a discovery cohort of OAC samples (using WGS and amplicon resequencing) to identify a set of recurrently mutated genes. Of the 26 genes identified, only *TP53* mutational status showed promise as a biomarker for progression; mutations in *TP53* were found in 72% of the 44 HGD samples, 69% of the 90 OAC samples and only in 1 out of 66 (2.5%) NDBO samples. Of the other 25 genes, only *SMAD4* showed significant difference in mutational frequency between the patient cohorts (NDBO, HGD, OAC); mutations were found in 13% of OAC samples but not HGD or NDBO samples. Given the higher frequency across patients and that mutation in *TP53* appears to occur prior to malignant progression, *TP53* mutational status was determined to be a biomarker with more promise than *SMAD4*. When *TP53* mutations were used as a biomarker in Cytosponge samples, *TP53* mutations were detected in 86% (19/22) of HGD samples and zero mutations were detected in normal samples (0/23) and those with NDBO (0/44). These results suggest that *TP53* mutational status has promise as a biomarker for progression. However, approximately 30% of patients with HGD or OAC do not harbour mutations in *TP53* and additional biomarkers are needed [209].

3.1.7 Genomic instability and the development of OAC: SCNAs as biomarkers for detecting and predicting OAC development?

Oesophageal adenocarcinoma is characterised by a high frequency of somatic mutations and structural rearrangements [210–212]. In a study of 149 cases of OAC and WGS and exome sequencing of tumour-normal pairs, Dulak *et al.* found a median mutation frequency of 9.9 mutations/Mb; 13.9 mutations/Mb in intergenic regions, 8.7 mutations/Mb in intronic regions and 6.5 mutations/Mb in coding exons. This places OAC among the cancers with the highest mutational frequency after lung cancer and melanoma [210]. Dulak *et al.* also found a mutational pattern unique to OAC, with high frequency of AA>AC transversions, representing 29% of all mutations [210].

In a later study, exome sequencing of paired BO and OAC samples revealed that the AA>AC mutational pattern is also present in the BO samples. Indeed, there were increased AA>AC mutations in BO samples compared with OAC samples, although not found to be statistically significant. These results suggest that these mutational patterns occur early in the progression of BO and OAC, perhaps as a consequence of damage to the mucosa by gastric acid and bile [212]. The same study compared median mutational frequencies between NDBO, dysplastic BO and OAC samples. Interestingly, NDBO was found to have a higher median mutational frequency than many invasive cancers (2.8 mutations/Mb). Dysplastic BO and OAC were found to have slightly higher median mutational frequencies than NDBO with 4.9 mutations/Mb and 4.1 mutations/Mb respectively [212]. Similar results showing high mutational frequencies in BO and OAC were seen by Ross-Innes *et al.* when studying WGS of paired Barrett's and OAC samples [211]. However, there was strikingly little overlap in the SNVs found in paired OAC and BO samples with 13 of the 23 pairs showing less than 20% overlap [211].

In contrast to somatic point mutations, structural rearrangements have been found to markedly increase with progression to OAC [213, 211, 212]. Stachler *et al.* observed that the mean number of deletions in NDBO, dysplastic BO and OAC were 10.43, 20.73 and 40.20 respectively [212]. Amplifications were seen at much higher frequency in OAC samples compared with BO with a mean of 0.42, 0.91 and 8.44 amplifications observed in NDBO, dysplastic BO and OAC samples respectively [212]. Indeed, the association between chromosomal instability and neoplastic progression of OAC has been evident since the 1980s. Reid and colleagues evaluated 317 biopsy specimens from 64 patients under endoscopic surveillance. By using flow cytometry, they found that specimens from 10 patients had evidence of non-diploid cells. Nine out of the 10 patients had dysplasia or carcinoma. Furthermore, all patients with dysplasia and carcinoma had either evidence of aneuploidy (non-diploid) or had increased number (>6%) of cells in G2 phase of the cell cycle or tetraploid cells [214]. Since then, there has been mounting evidence that OAC develops in association with increasing SCNAs [164].

In a longitudinal study of patients with BO, Li *et al.* [213] studied the genomes of BO samples from patients that went on to develop OAC (progressors) and those that did not (non-progressors). Typically, the genomes of non-progressors saw relatively stable SCNAs over time and were characterised by short deletions at fragile sites (commonly *FHIT*, *WWOX*, *CDKN2A*), loss or copy neutral loss of heterozygosity (cnLOH) of 9p and little spatial heterogeneity. More than 48 months prior to diagnosis, progressors had similar SCNA profiles to non-progressors with 18q loss the only SCNA with significantly higher frequency in progressors than non-progressors at this stage. At 24 to 48 months prior to diagnosis, larger regions of SCNAs developed in progressors, including gains of chromosomes 8 and 15q and losses of 5q, 17p, 18 and Y along with greater spatial heterogeneity in SCNA profiles. Within 24 months of OAC diagnosis, genome-doubling was often observed along with increased SCNAs involving more than 1,850 Mb of the genome [213]. Indeed, spatial heterogeneity and polyclonality were seen by Ross-Innes *et al.* in the paired BO samples from patients with OAC [211] and clonal diversity has been shown to confer risk of OAC development [215–217, 152]. These results suggest that SCNA load over time, spatial heterogeneity and the presence of particular SCNAs may be useful biomarkers to predict those that go on to develop OAC. Indeed, these biomarkers can appear two to four years before OAC diagnosis and detecting them may allow for early intervention.

Stachler *et al.* [212] have since proposed two mechanisms by which BO can progress to OAC. The first mechanism involves a gradual step-wise loss of tumour suppressor genes including *CDKN2A*, *ARID1A*, *TP53* and *SMAD4* along with alterations of chromatin-modifying enzymes. This leads to chromosomal instability, oncogenic amplifications and OAC development without a genome-doubling event. The second mechanism involves *TP53* loss followed by a catastrophic genome-doubling event. Genome-doubling leads to chromosomal instability, a greater number of oncogene amplifications than the first mechanism and perhaps a more rapid development of OAC [212]. It is suggested that the majority of OACs emerge via the second mechanism. Both mechanisms include chromosomal instability and the accumulation of SCNAs. Oncogenic amplifications appear to occur at relatively late stage in

OAC progression and Stachler *et al.* suggest it may be critical late step in the transformation to invasive OAC [212].

In a recent study, Ross-Innes *et al.* [218] investigated whether the use of the Cytosponge and a panel of biomarkers could accurately risk-stratify patients. The purpose was to identify those at low-risk of OAC progression so that these patients could be spared endoscopy. They devised a model including three binary biomarkers (p53 status, Aurora kinase A expression and histopathologist assessment of glandular atypia) along with the interaction of three clinical factors (age, length of Barrett's segment, waist:hip ratio) that split patients into low-risk, moderate-risk and high-risk. They also tested protein biomarkers for c-Myc and methylation of *MYOD1* and *RUNX3* but did not include these in the final risk-stratification model. Aurora kinase A expression was included as a surrogate marker for aneuploidy (evidence of a deviation from diploidy) because performing SCNA profiling (using SNP arrays) or cell cycle analysis was deemed to be infeasible or too expensive [218]. Interestingly, binarised Aurora kinase A expression was the most sensitive in discriminating HGD/IMC from NDBO (78%) when it was tested as an independent variable, while p53 immunohistochemistry was the most specific (96%) [218]. Increased sensitivity might be achieved by using SCNA profiles rather than using surrogate markers.

3.1.8 The landscape of SCNAs in OAC and potential drug targets

SCNAs are common in OAC and similar patterns of SCNAs alterations also occur in gastric adenocarcinoma with both cancer types marked by chromosomal instability [219, 152]. Potential therapeutic targets include amplified receptor tyrosine kinases (e.g. *ERBB2*/*HER2*, *KRAS*, *EGFR*, *MET*, *FGFR2*, *VEGFA*) cell cycle regulators (e.g. *CCND1*, *CCNE1*, *CDK6*) and transcription factors (e.g. *GATA4*, *GATA6*, *MYC*). [152]. The targeting of these genes in OAC has been largely unsuccessful to date. Only the targeting of ERBB2-positive cancers with the use of trastuzumab has shown success. However, this success has been limited to modest improvements in survival in patients with metastatic disease [220, 221].

Cell cycle regulators and members of the DNA damage response (DDR) pathway are two proposed targets for therapy. 86% of OAC cases have been shown to have cell cycle dysregulation. For example, in estrogen receptor-postive breast cancer, the inhibition of CDK4 and CDK6 has been shown to improve survival. Therapies such as ribociclib and palbociclib, which target CDK4/6 may also be useful *CDK4*/*6*-amplified OAC [222, 152]. The DDR pathway could be targeted with drugs such as poly ADP ribose polymerase (PARP) inhibitors (e.g. olaparib), ataxia telangiectasia and Rad3-related protein (ATR) inhibitors and platinum-based chemotherapy [221, 223, 152].

There are several challenges in the development and application of targeted therapies for OAC. These include the acquisition of additional tyrosine kinase amplifications (coamplification) and the presence of intratumoural heterogeneity which both confer resistance to therapy. So far, the targeting of the DDR pathway has been impeded by the lack of biomarkers to select patients that may benefit from these therapies [152, 223]. SCNA profiling may provide a means for stratifying patients for targeted therapy and monitoring the development of resistance.

3.1.9 Concluding remarks

The incidence of OAC is rapidly increasing, particularly in the modern industrialised nations and carries with it a very poor prognosis. Pre-malignant lesions and early disease can be successfully treated and treatment becomes more difficult in more advanced disease. SCNAs appear to play an important role in the onset and development of OAC and are promising biomarkers for disease detection, stratification, therapy selection and the monitoring of disease progress. FAST-SeqS and conliga could provide a clinically applicable solution for the SCNA profiling of patients.

3.2 Application of conliga to OAC and comparison with high and low-coverage WGS

The aim of this section is to assess conliga's performance as a tool for SCNA profiling in the context of its clinical utility for OAC. The performance is assessed by comparing copy number calls obtained from a set of OAC samples using high-quality and high-coverage WGS data. For comparison, I compare conliga's performance to the performance of a widely-used method applied to low-coverage WGS data. The performance is assessed at varying resolutions. Firstly, I compare the overall SCNA profiles between methods (section 3.2.3). Secondly, I compare the RCN calls across all OAC samples to assess the quantification of copy number (section 3.2.4). Thirdly, I evaluate conliga's performance at gene level resolution; assessing the associated RCN values for genes that are recurrently amplified and deleted in OAC (section 3.2.5). Lastly, by performing an *in silico* dilution experiment, I assess conliga's performance in low purity samples (section 3.2.6).

3.2.1 Sample information

We selected 15 OAC samples that were previously sequenced as part of the International Cancer Genome Consortium (ICGC) oesophageal project using WGS at $\geq$ 50X coverage (along with matched normal samples at $\geq$ 30X coverage) [221]. By performing FAST-SeqS on DNA that was previously extracted and sequenced from these 15 OAC samples, we could directly compare the performance of conliga to well-tested methods applied to WGS data.

The FAST-SeqS sample information can be found in table 3.1. Of the 15 samples selected, four samples had a low number of reads ($\leq 350,000$ reads) and these were excluded from the analysis.

Sample	Seq. ID	Reads	Counts	Counts (pf)
NA	SLX-9396.FastSeqA	27515	13108	NA
OAC1	SLX-9396.FastSeqB	2235605	1430644	1301862
OAC2	SLX-9396.FastSeqC	1771657	1167955	1049641
OAC3	SLX-9396.FastSeqD	1386224	901797	792917
OAC4	SLX-9396.FastSeqE	1424501	950871	860611
OAC5	SLX-9396.FastSeqF	4651028	3036990	2769289
OAC6	SLX-9396.FastSeqG	2126721	1413199	1287926
OAC7	SLX-9396.FastSeqH	2479560	1608219	1445539
NA	SLX-9396.FastSeqI	8945	5585	NA
OAC8	SLX-9396.FastSeqJ	753806	407586	370725
OAC9	SLX-9396.FastSeqK	1803553	1168750	1060265
NA	SLX-9394.FastSeqA	163671	84952	NA
NA	SLX-9394.FastSeqB	323995	182159	NA
OAC10	SLX-9394.FastSeqC	636616	335538	304306
OAC11	SLX-9394.FastSeqD	607073	316801	278129

Table 3.1 Oesophageal adenocarcinoma cohort: FAST-SeqS sample information. 'Reads' represents the total number of reads assigned to each sample demultiplexing the reads from the sequencing lanes. 'Counts' represents the number of counts after each sample's FASTQ file was processed using the pipeline described in section 2.4. 'Counts (pf)' shows the number of counts that remained after filtering loci with a zero count in any of a set of control samples (see section 3.5 for details).

3.2.2 Copy number tools and data used as a comparison to conliga and FAST-SeqS

ASCAT-NGS [224] is a widely used tool for SCNA analysis using matched tumour-normal WGS samples and is the preferred method used for SCNA calling as part of the ICGC oesophageal project. For example, it was used by Secrier *et al.* in their recent paper [221]. Because of this, I used ASCAT's calls as the gold standard to compare conliga's results to.

50X WGS is approximately 100-fold more expensive than FAST-SeqS data to produce and as such has limited clinical application currently. Low-coverage (LC) WGS has been suggested as a low-cost alternative to high-coverage (HC) WGS for SCNA profiling and is now widely used for this purpose [225, 99]. The cost-savings associated with LC WGS are due to the reduced sequencing required per sample. However, low-coverage is a vague term and its meaning depends on the context. For example, in the context of variant calling, low-coverage can be considered as 4-10X while 1-2X is considered as extremely low-coverage [226]. In the context of copy number profiling, low-coverage is also ill-defined; 0.1-1X coverage has been considered low-coverage [225, 227] whereas 0.01X has been considered as ultra low-coverage [228]. For our purposes, I define LC WGS as 0.1X coverage using single-end 50 bp reads. To achieve 0.1X coverage, approximately 9 million reads are required assuming that of these 9 million reads, 6 million reads will align to the genome [225]. As we calculated in Abujudeh *et al.* [108], this means that LC WGS is approximately £52-72 per sample (depending on the library preparation used) and 3.7-5.1 fold more expensive than FAST-SeqS which costs £14 per sample when using two million 150 bp single-end reads per sample [108]. Rather than preparing new sequencing libraries and sequencing the OAC samples again, LC WGS was produced by down-sampling reads from HC WGS, see section 3.5.1 for details.

QDNAseq [225] is a widely used tool for SCNA profiling using LC WGS data and has been cited more than 100 times. It uses a depth of coverage approach to infer relative copy number values for genomic bins and the genomic bin size is defined by the user. To remove read depth biases in the count data, it applies a two-dimensional LOESS correction using GC content and mappability as the two dimensions. After correcting for these biases, the noise observed in the normalised counts is primarily comprised of sampling noise [225], suggesting that GC content and mappability make up most of the technical variation in LC WGS binned read counts. Scheinin *et al.* successfully applied QDNAseq to samples with ~ 0.1X coverage (50bp single-end reads) and showed good performance [225]. For these reasons, I chose to compare conliga's performance to QDNAseq applied to LC WGS data.

3.2.3 Comparing copy number profiles

Comparing conliga's scaled and QDNAseq's normalised counts to ASCAT's total copy number calls

As an exploration of the data, we can compare the scaled counts from conliga to ASCAT's total copy number calls (TCN; that is, the sum of allele-specific copy number). In figure 3.1A we see ASCAT's TCN calls and their corresponding scaled count calculated by conliga at FAST-SeqS loci for a particular OAC sample. We see that the scaled count is proportional to ASCAT's TCN; the scaled counts proportionally increase as the TCN increases. This observation is a generalisation of what we observed when comparing male and female samples in figure 2.5 in chapter 2. When comparing QDNAseq's normalised counts (which corresponds to conliga's scaled counts) in figure 3.1B, we see the same relationship but with less variation. However, we see that FAST-SeqS loci with high $\hat{m}$ values have less variation around the ASCAT's RCN estimate and the amount of variation appears to be comparable to QDNAseq's normalised count. It is the loci with low $\hat{m}$ values that vary considerably around ASCAT's RCN estimate.

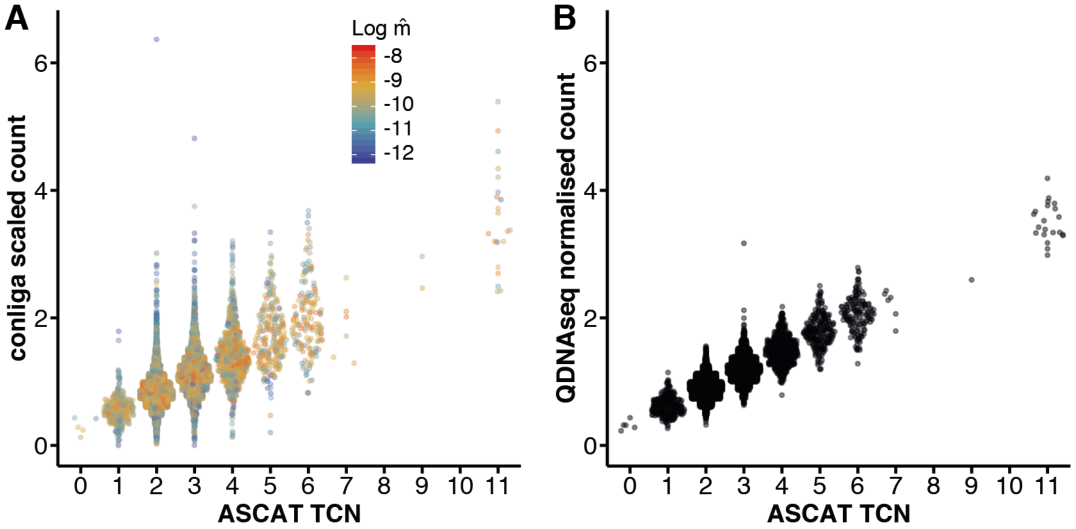


Figure 3.1 Comparison between conliga's scaled counts and QDNAseq's normalised counts with ASCAT's total copy number calls for sample OAC2. A: Sina plot showing the distribution of conliga's scaled counts that overlap with ASCAT's total copy number (TCN) calls and QDNAseq's calls. Each locus is plotted as a point and coloured by the loci's associated $\log \hat{m}$ value (higher values are red, lower values are blue). B: Sina plot showing the distribution of QDNAseq's normalised counts that overlap with ASCAT's total copy number (TCN) calls.

Figure 3.2A shows the Pearson correlation coefficient between ASCAT TCN calls and conliga's scaled and QDNAseq's normalised counts for each of the 11 OAC samples. We see that when including all FAST-SeqS loci, QDNAseq's normalised counts more strongly correlate with ASCAT's TCN calls than conliga's scaled counts correlate with ASCAT TCN. However, if we compare the 1000 loci with the highest inferred $\hat{m}$ values, we see a similar strength of correlation which confirms our observations from figure 3.1.

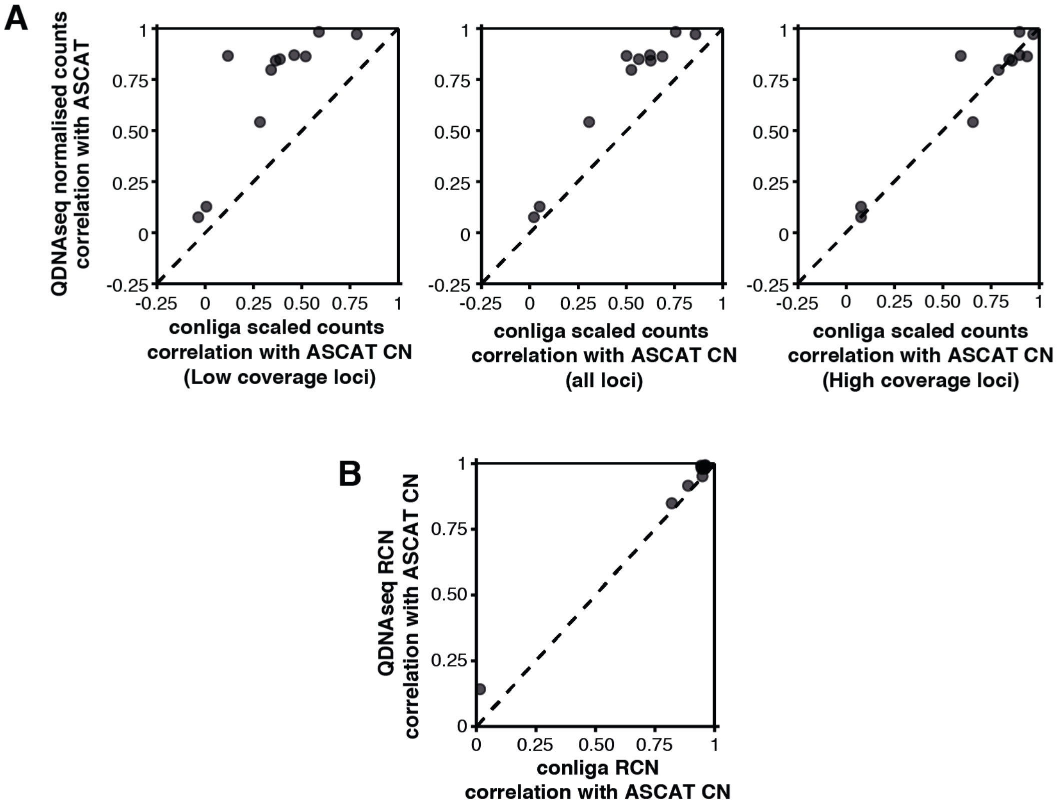


Figure 3.2 Comparison between the output of conliga and QDNAseq with ASCAT's TCN copy number profiles. A: Pearson correlation between QDNAseq's normalised counts with ASCAT's TCN profiles compared against the Pearson correlation between conliga's scaled counts with ASCAT's TCN profile. Each point represents an OAC sample. Left: comparison using conliga's scaled counts from the 1000 FAST-SeqS loci with the lowest inferred $\hat{m}$ value. Middle: comparison using all FAST-SeqS loci. Right: comparison using conliga's scaled counts from the 1000 FAST-SeqS loci with the largest inferred $\hat{m}$ value. In all cases, all the 15 Kbp bins were used in calculating QDNAseq's correlation with ASCAT TCN (see section 3.5.5 for more details). B: Pearson correlation between QDNAseq's segmented RCN profile with ASCAT TCN profile against Pearson correlation between conliga's RCN profile and ASCAT's TCN profile. Each point represents an OAC sample.

Comparing conliga's and QDNAseq's RCN calls to ASCAT's TCN calls

ASCAT produces absolute and allele-specific copy number calls along with an estimation of the proportion of normal cell contamination. If all methods provided allele-specific copy number, it would be straightforward to provide measures of sensitivity, specificity and false positive rates for conliga and QDNAseq. However, conliga and QDNAseq produce relative copy number calls. A simple alternative to provide a similarity metric between copy number profiles is to report the Pearson correlation between conliga's RCN and ASCAT's TCN and QDNAseq's RCN and ASCAT's TCN for each sample.

Another issue is that calls are made at different genomic regions; ASCAT provides calls in contiguous genomic regions, whereas conliga makes calls at FAST-SeqS loci and QDNAseq provides calls in 15 Kbp genomic bins. Indeed, in some cases conliga makes calls which do not fall in a region called by ASCAT or the corresponding QDNAseq bin was filtered by QDNAseq's `applyFilters` function. Table 3.2 shows the number of FAST-SeqS loci that do not have a corresponding ASCAT or QDNAseq call for each OAC sample. For the purposes of comparing profiles, I chose to use the overlapping FAST-SeqS loci with ASCAT calls for computing the Pearson correlation between conliga and ASCAT. For computing the Pearson correlation between QDNAseq and ASCAT, I used the overlapping 15 Kbp bins with ASCAT calls. In this case, there might be multiple ASCAT calls within a given 15 Kbp bin. In such cases, I used the length-weighted mean of ASCAT's overlapping TCN values with the QDNAseq bin to calculate the Pearson correlation.

Sample	**ASCAT**	**QDNAseq**
OAC1	6	274
OAC2	9	274
OAC3	10	274
OAC4	3	274
OAC5	14	274
OAC6	5	274
OAC7	6	274
OAC8	2	274
OAC9	12	274
OAC10	4	274
OAC11	3	274

Table 3.2 The number of FAST-SeqS loci without a corresponding copy number call in ASCAT and QDNAseq (out of a total of 11,854 loci in chromosomes 1-22 in each sample).

Figure 3.2B and table 3.3 show the Pearson correlation coefficients for each OAC sample. We see that conliga appears to have similar performance to QDNAseq with a median Pearson correlation of 0.95 (IQR: 0.92-0.96) between conliga and ASCAT compared with 0.98 (IQR: 0.93-0.99) between QDNAseq and ASCAT. Indeed, this analysis is biased in favour of QDNAseq given that the LC WGS reads were downsampled from the HC WGS reads used by ASCAT. There is a strong relationship between the Pearson correlation of both methods. We see from figure 3.3A that samples with a low value for ASCAT's goodness-of-fit have a lower correlation with conliga and QDNAseq. Indeed, these samples also have a lower inferred tumour purity (figure 3.3B). Together, these results suggest that ASCAT's solution may be less reliable for these samples or that conliga and QDNAseq have lower performance for lower tumour purity samples or a combination of both factors.

Sample	conliga/ASCAT	QDNAseq/ASCAT	conliga/QDNAseq
OAC1	0.945	0.985	0.953
OAC2	0.944	0.992	0.943
OAC3	0.961	0.993	0.939
OAC4	0.947	0.980	0.956
OAC5	0.949	0.952	0.956
OAC6	0.959	0.992	0.964
OAC7	0.889	0.916	0.921
OAC8	0.961	0.987	0.967
OAC9	0.959	0.980	0.958
OAC10	0.015	0.142	0.480
OAC11	0.820	0.849	0.818

Table 3.3 Pearson correlation coefficients between copy number calls obtained from conliga, QDNAseq and ASCAT. conliga/ASCAT represents the Pearson correlation between the overlapping ASCAT TCN calls with and conliga's RCN calls at FAST-SeqS loci. QDNAseq/ASCAT represents the Pearson correlation coefficient between ASCAT's length-weighted TCN calls with QDNAseq 15 Kbp calls. conliga/QDNAseq represents the Pearson correlation coefficient between conliga's RCN calls and QDNAseq's RCN calls.

To explore how much information is retained by sampling the genome at FAST-SeqS loci rather than sequencing the whole genome, we can compare the ASCAT calls at FAST-SeqS loci to the full set of genomic segments called by ASCAT. Figures 3.4A and C show that by sampling the genome at FAST-SeqS loci, high resolution SCNA information is maintained in ASCAT calls and the overall profile is similar.

By visually inspecting the profiles generated by QDNAseq (figures 3.4B) and conliga (figures 3.4D) we see that both provide similar profiles. This sample has a complex SCNA profile which both methods are able to recapitulate. For example, chromosome arms 1q, 5p, 8p, 8q and 13q show a high number of SCNA events which can be seen in the outputs of all methods.

For one sample (OAC10), conliga's calls more strongly correlate with QDNAseq's calls than either conliga's or QDNAseq's calls correlate with ASCAT (table 3.3). This might suggest that ASCAT's solution is incorrect. Inspection of the copy number profiles for this sample obtained by the three methods (not shown), suggest that the sample may have purity lower than that inferred by ASCAT (0.25) with indication of gains of all or part of chromosome 6, 8, 18 and a gain of 3p which were not called by ASCAT.

In this section, I have shown that for the purposes of obtaining an overall genome-wide RCN profile, conliga provides a similar level of performance to QDNAseq.

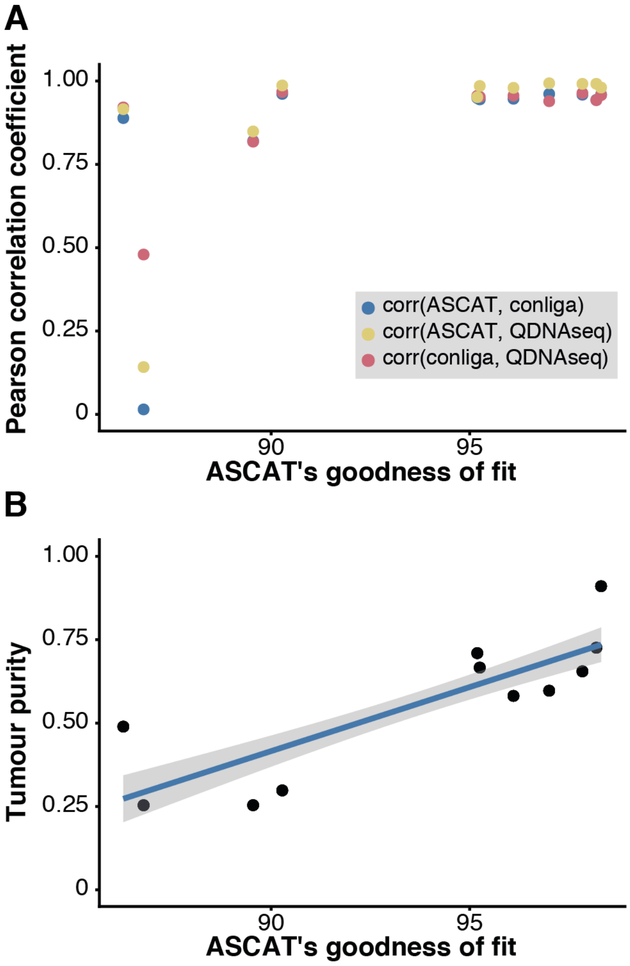


Figure 3.3 Pearson correlation coefficients compared with ASCAT's goodness-of-fit and tumour purity. A: ASCAT's goodness-of-fit compared with the Pearson correlation coefficients. B: The relationship between ASCAT's goodness-of-fit to ASCAT's inferred tumour purity.

3.2.4 Comparing RCN calls at each FAST-SeqS locus across samples

Now, rather than comparing copy number profiles, we can compare RCN calls at each FAST-SeqS locus between methods across the 11 samples. In order to do so, it is necessary to convert ASCAT's calls to RCN (see section 3.5.5). Figure 3.5A shows ASCAT's RCN values plotted against those obtained from conliga and QDNAseq. We see that RCN estimates are strongly correlated with ASCAT RCN calls; the Pearson correlation coefficient between ASCAT's RCN and conliga's RCN is 0.953 compared to 0.987 between ASCAT and QDNAseq. Figure 3.5B shows very similar residual distributions for both methods. Figure 3.5C shows that at short copy number segments, conliga and QDNAseq have greater error in their estimations (compared with ASCAT). As expected, as the copy number segment length increases conliga and QDNAseq see more accurate RCN estimates.

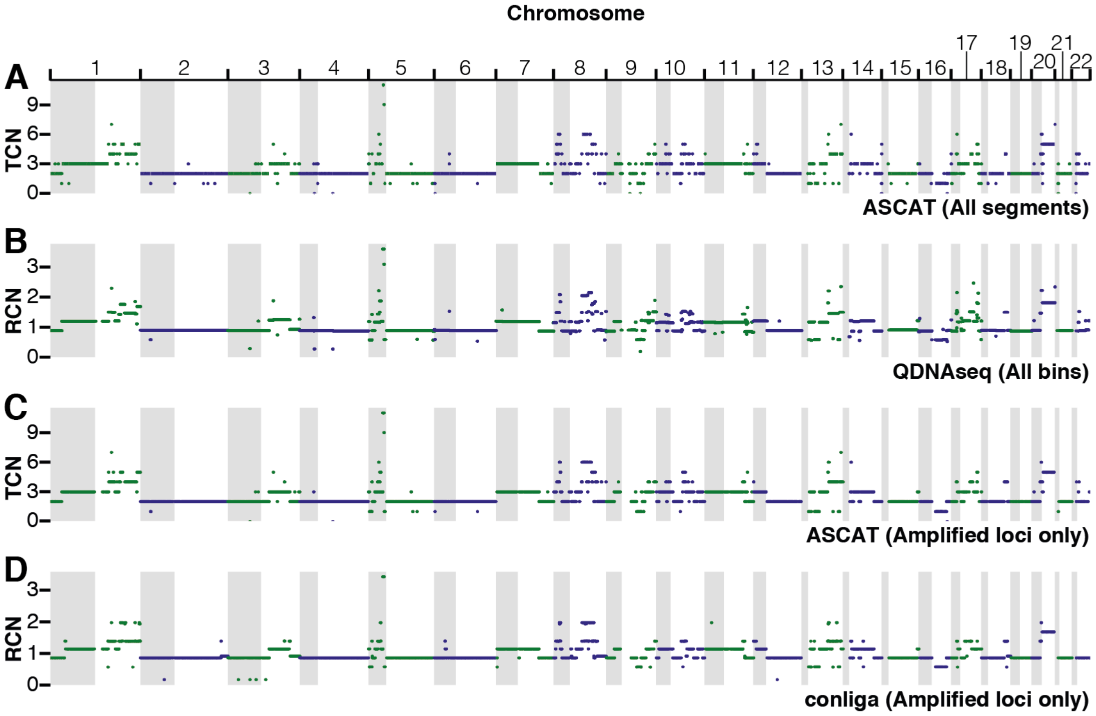


Figure 3.4 SCNA profiles for sample OAC2 obtained by ASCAT, QDNAseq and conliga in chromosomes 1-22. A: ASCAT total copy number calls for all copy number segments called by ASCAT. B: QDNAseq RCN calls for 15 Kbp bins. C: ASCAT TCN calls that intersect with FAST-SeqS loci. D: conliga RCN calls at FAST-SeqS loci.

3.2.5 Exploring performance at recurrently amplified and deleted genes

Several genes are recurrently amplified and deleted in OAC [219, 229, 211, 221]. Detecting the copy number status of these genes is useful from a clinical perspective and for basic research purposes. Since FAST-SeqS samples the genome at LINE1 elements, we might expect conliga to perform poorly in detecting and quantifying focal amplifications and deletions at the gene level of resolution. In this section, I explore conliga's performance in quantifying the RCN for a set of genes and compare its performance to QDNAseq which uses reads across the whole genome. Section 3.5.5 describes how the RCN was quantified in various ways at the gene level for each of the three methods.

I selected recurrently amplified and deleted genes that were described in Dulak *et al.* and Ross-Innes *et al.* [219, 211]. Table 3.4 lists the genes that were selected from the literature that were used in this analysis. These genes include tumour suppressor genes such as *TP53*, *CDK2NA* and *FHIT* and also recurrently amplified oncogenes. Indeed, as we discussed in section 3.1.8, amplifications involving some of these genes have therapeutic and prognostic relevance.

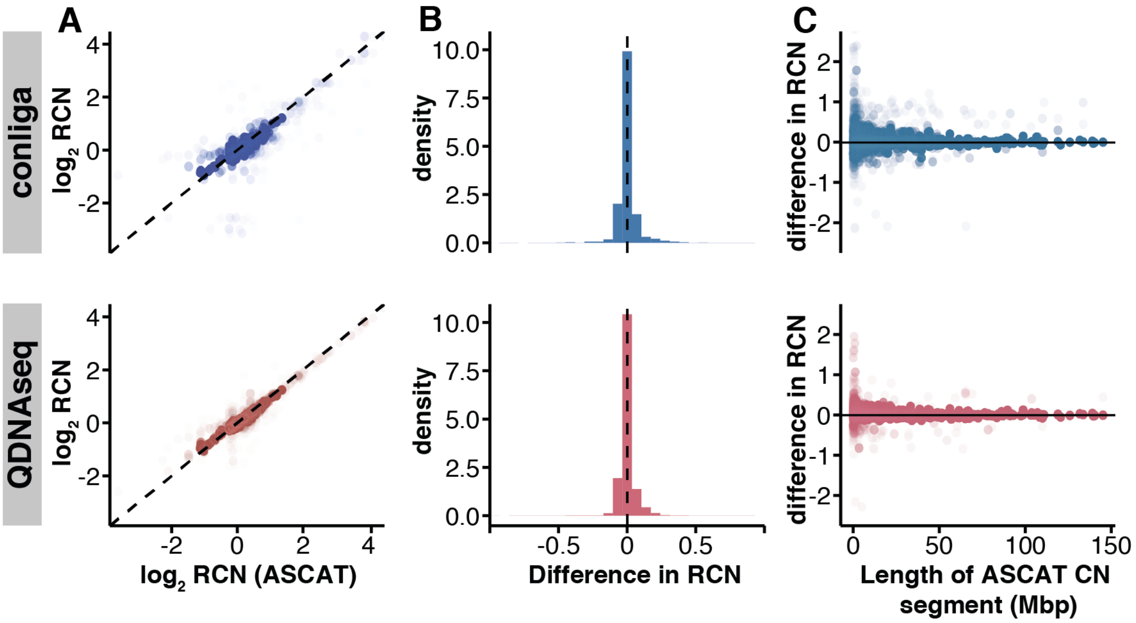


Figure 3.5 Relative copy number comparison between methods across the 11 OAC samples (compared at the intersection between ASCAT's segments, QDNAseq's bins and FAST-SeqS loci, n=127,310). A: ASCAT's converted RCN against conliga's RCN calls (top) and QDNAseq's RCN calls (bottom). B: The distribution of differences between ASCAT's converted RCN calls and conliga's and QDNAseq's RCN calls; ASCAT RCN minus conliga RCN (top) and ASCAT RCN minus QDNAseq RCN (bottom). C: Comparing the length of ASCAT's copy number segment length against the difference in RCN between ASCAT and conliga (top) and ASCAT and QDNAseq (bottom).

One important factor that affects conliga's ability to detect SCNAs involving genes is the proximity of the FAST-SeqS loci to the genes of interest. Figure 3.6A shows a graphical representation of the genomic distance of neighbouring FAST-SeqS loci to the 5' and 3' end of the genes. Interestingly, 13 out of the 36 genes have FAST-SeqS loci within them, presumably within intronic regions. Some of the tumour suppressors genes have multiple FAST-SeqS loci within them (as many as 15 in *PTPRD*). It is not only tumour suppressors that contain FAST-SeqS loci, *CDK6* (a cell cycle regulator) and *MET* (which encodes a tyrosine kinase receptor) both contain a FAST-SeqS locus. Indeed, these genes are among those that have potential therapeutic relevance [152]. Note that I have used the filtered FAST-SeqS loci (which had a non-zero count in all control samples) so there may be additional loci that fall within these genes and others. However, since these loci were not used for the inference of RCN profiles, I did not include them in this analysis.

The remaining 23 genes out of 36 do not contain FAST-SeqS loci but rather are flanked by them. If either (or both) the loci that flank these genes show evidence of amplification or deletion, we may believe that the copy number of the gene has been altered (i.e. that either of the loci have sampled the copy number segment involving the gene). Naturally, if an

Gene	**Chr**	**Start**	**End**	**Status**	**Source**
ARID1A	chr1	26696033	26782104	del	D
MCL1	chr1	150574551	150579738	amp	D
PARD3B	chr2	204545793	205620162	del	D
FHIT	chr3	59749310	61251459	del	D; R
PRKCI	chr3	170222365	170305981	amp	D; R
CCSER1	chr4	90127535	91601913	del	D
PDE4D	chr5	58969038	60522120	del	D; R
VEGFA	chr6	43770184	43786487	amp	D; R
MYB	chr6	135181315	135219173	amp	D; R
PARK2	chr6	161347420	162727771	del	D; R
EGFR	chr7	55019021	55256620	amp	D
CDK6	chr7	92604921	92836594	amp	D
MET	chr7	116672390	116798386	amp	D
EPHB6	chr7	142855061	142871094	amp	D
GATA4	chr8	11676959	11760002	amp	D; R
MYC	chr8	127735434	127741434	amp	D
PTPRD	chr9	8314246	10612723	del	D; R
CDKN2A	chr9	21967753	21995301	del	D; R
FGFR2	chr10	121478334	121598458	amp	D
CCND1	chr11	69641087	69654474	amp	D; R
FGF19	chr11	69698232	69704642	amp	R
FGF4	chr11	69771016	69775403	amp	R
FGF3	chr11	69810224	69819024	amp	R
ATM	chr11	108222484	108369102	del	D
KRAS	chr12	25204789	25250936	amp	D
MDM2	chr12	68808172	68850686	amp	D
KLF5	chr13	73054976	73077542	amp	R
RBFOX1	chr16	6019094	7713338	del	R
WWOX	chr16	78099413	79212667	del	D; R
TP53	chr17	7661779	7687550	del	D
ERBB2	chr17	39687914	39730426	amp	D
GATA6	chr18	22169443	22202528	amp	D
SMAD4	chr18	51028394	51085045	del	D
CCNE1	chr19	29811898	29824308	amp	D
MACROD2	chr20	13995369	16053197	del	D
RUNX1	chr21	34787801	36004667	del	D

Table 3.4 Recurrently amplified and deleted genes in OAC that were selected from the literature. The gene name supplied is the HGNC symbol. The coordinates given are those for the GRCh38 reference genome. Status defines the recurrent status of the gene in OAC; amp represents recurrently amplified and del represents recurrently deleted. Source represents the literature source; D represents Dulak *et al.* [219] and R represents Ross-Innes *et al.* [211].

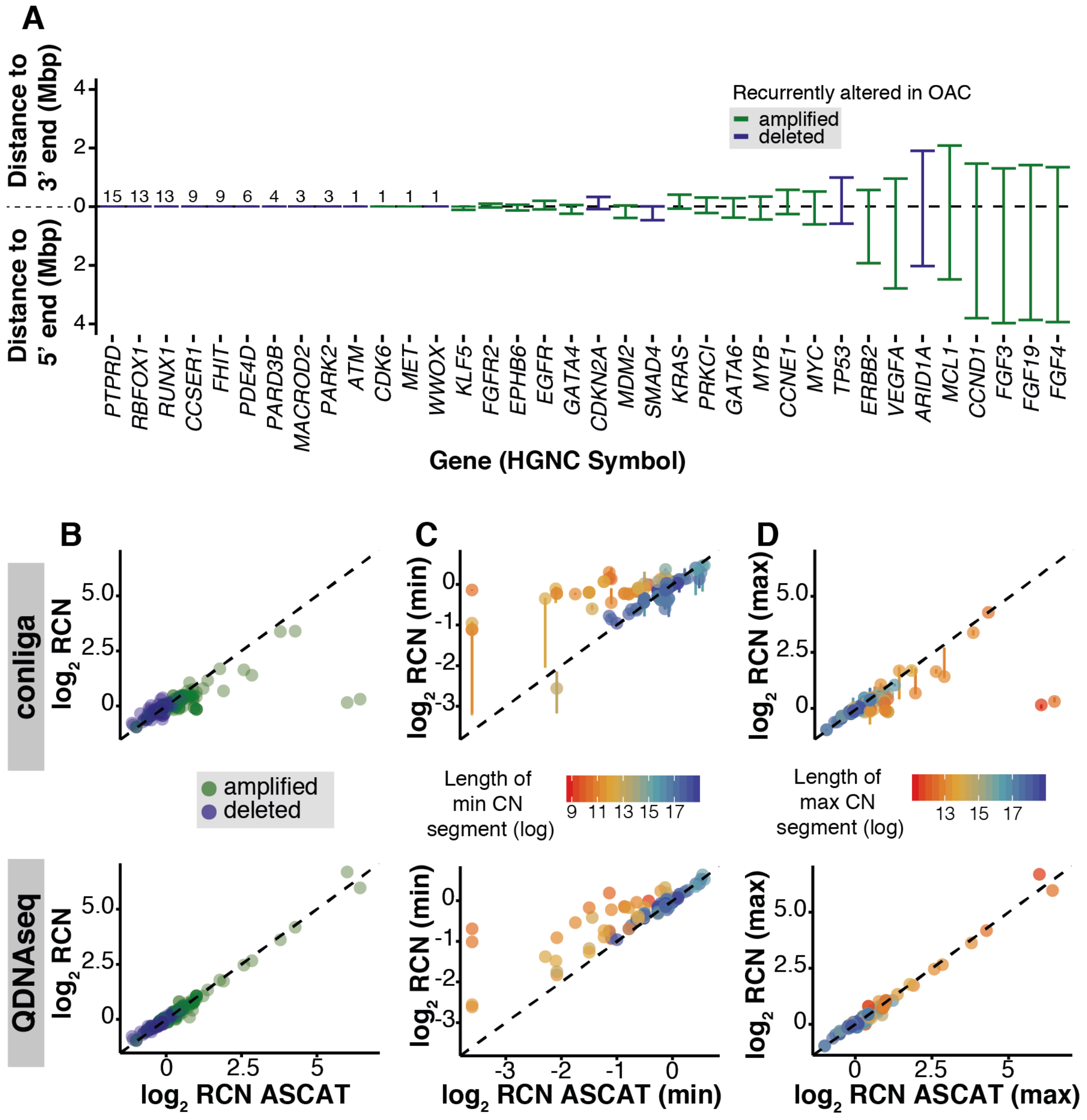


Figure 3.6 Summarising conliga's and QDNAseq's utility in determining the RCN for recurrently amplified and deleted genes in OAC. A: The number of loci within currently deleted and amplified genes or the distance of flanking FAST-SeqS loci to the 3' and 5' ends of the genes (green: recurrently amplified genes, blue: recurrently deleted genes). B: The weighted mean RCN for each gene in each sample (396 points); top: ASCAT vs conliga, bottom: ASCAT vs QDNAseq, green: recurrently amplified genes, blue: recurrently deleted genes. C: The minimum RCN associated with each recurrently deleted gene for each sample. The colour represents the length of the copy number segment with the minimum TCN associated with the gene as determined by ASCAT. D: The maximum RCN associated with each recurrently amplified gene for each sample. The colour represents the length of the copy number segment with the maximum TCN associated with the gene as determined by ASCAT.

alteration occurs between the two flanking loci, conliga will not be able to detect the gene level alteration. Indeed, the probability of detecting a copy number alteration that affects a gene will likely be inversely correlated with the genomic distance of the flanking FAST-SeqS loci. The majority of the 23 genes have flanking loci that are $\leq$ 1 Mbp of the 5' and 3' end of the gene. *CCND1*, *FGF3*, *FGF4* and *FGF19* have the greatest mean genomic distance to flanking FAST-SeqS loci. However, these four genes occur in close proximity to each other within chromosome 11. *ERBB2* (also called *HER2*) encodes a receptor tyrosine kinase that has important clinical relevance in OAC and other cancers (including breast, ovarian, endometrial and gastric) [152, 80]. Trastuzumab, which is a monoclonal antibody that targets HER2, can be used to treat patients with OAC that overexpress ERBB2/HER2 [152]. FAST-SeqS loci fall approximately 567 Kbp and 1.9 Mbp from the 3' and 5' end of *ERBB2* respectively. Other genes that encode receptor tyrosine kinases such as *FGFR2*, *EGFR* and *KRAS* have FAST-SeqS loci in relatively close proximity. Therefore we may deduce that amplifications which include these genes have greater probability of being detected compared with *ERBB2*, for example.

Figure 3.6B shows the weighted mean RCN calls obtained for each gene via the three methods in each of the 11 OAC samples (see section 3.5.5 for details of how this was calculated). For conliga and QDNAseq, we see that recurrently amplified and deleted genes show evidence of amplification and deletion (respectively) across the 11 samples. While there is clear concordance between the weighted mean RCN calls of ASCAT and conliga, QDNAseq shows greater concordance with ASCAT's weighted mean RCN values. For highly amplified genes, the associated weighted mean RCN from conliga occasionally falls below that of ASCAT. This may be due to various reasons, including at least one of the following:

1. The flanking FAST-SeqS loci do not fall within a copy number segment affecting the gene.

2. One of the flanking loci fall within the copy number segment and the other does not. As such, the weighted mean RCN will be brought down by the locus which does not.

3. There is insufficient evidence (perhaps the sample has insufficient total reads or the locus or loci receive relative few reads) for the HMM to call a change in copy number state.

4. There may be sufficient evidence to call a change in state. However, since conliga clusters loci together into copy number states, there may be insufficient statistical strength for a separate copy number state with very few loci (or perhaps only one locus) belonging to it. Therefore, several loci with similar copy numbers may be clustered into a single copy number state. In some cases, this may lead to focal amplifications being underestimated.

5. The indirect method was used to summarise the RCN and this may lead to underestimating the true RCN (this was discussed in section 2.8).

Figure 3.6C and D explore the gene level RCN values in greater detail. Figure 3.6C explores the minimum RCN value associated with the recurrently deleted genes. This value may be of interest because a focal deletion within a tumour suppressor gene may render it inactive. We see that conliga and QDNAseq often fail to detect or tend to overestimate the RCN of short deletions. For conliga, the reasons for this are similar to those described above. In some cases, it appears that the focal deletion was sampled by FAST-SeqS loci; by exploring the CI of the loci associated with the genes, we see that conliga places some probability on a lower RCN value. In a minority of cases, conliga is able to detect intra-gene deletions using the FAST-SeqS loci that fall within the genes (Figure 3.7A). This suggests that, as long as the copy number segment is sampled by a FAST-SeqS amplicon, increasing the proportion of correct focal deletions or amplifications could be achieved by increasing the total number of reads per sample or reducing the technical variation of the protocol. QDNAseq may miss or underestimate the minimum RCN for various reasons. One reason could be that a copy number alteration was not detected by the segmentation algorithm. Another reason could be that the focal deletion may cover a portion and not the entire 15 Kbp bin. By binning the genome, the associated RCN may be smoothed (and overestimated or underestimated) via an averaging effect.

Figure 3.6D shows the maximum RCN value associated with recurrently amplified genes. In two cases, we see that highly focal amplifications in sample OAC5 involving *MYC* and *ERBB2* were missed by conliga. By exploring those regions, we see that these events occur between FAST-SeqS loci and were not sampled (see figure 3.7E and 3.7F). By including the CIs, we see that in some cases, conliga places some probability on the genes having a higher maximum RCN (see for example *GATA4* in figure 3.7G). Some focal amplifications are accurately quantified (e.g. *EGFR* in figure 3.7H, albeit clustered into a single copy number state). QDNAseq tends to more accurately quantify the focal amplifications. This is in part due to it using whole genome data and therefore every copy number segment is sampled. Additionally, QDNAseq is not restricted to perform a global copy number clustering of the bins as conliga does with FAST-SeqS loci. As we discussed above, this clustering can lead to over or underestimation of the RCN for particular loci. We will see that this less restricted approach has its disadvantages when we explore each method's performance in detecting RCN profiles in low purity samples (section 3.2.6).

In summary, in this section I have shown that conliga can provide evidence of focal alterations affecting recurrently amplified and deleted genes in OAC. The ability for conliga to accurately identify and quantify these alterations depends on whether FAST-SeqS loci fall within the copy number segments affecting the genes. In some cases, conliga can detect

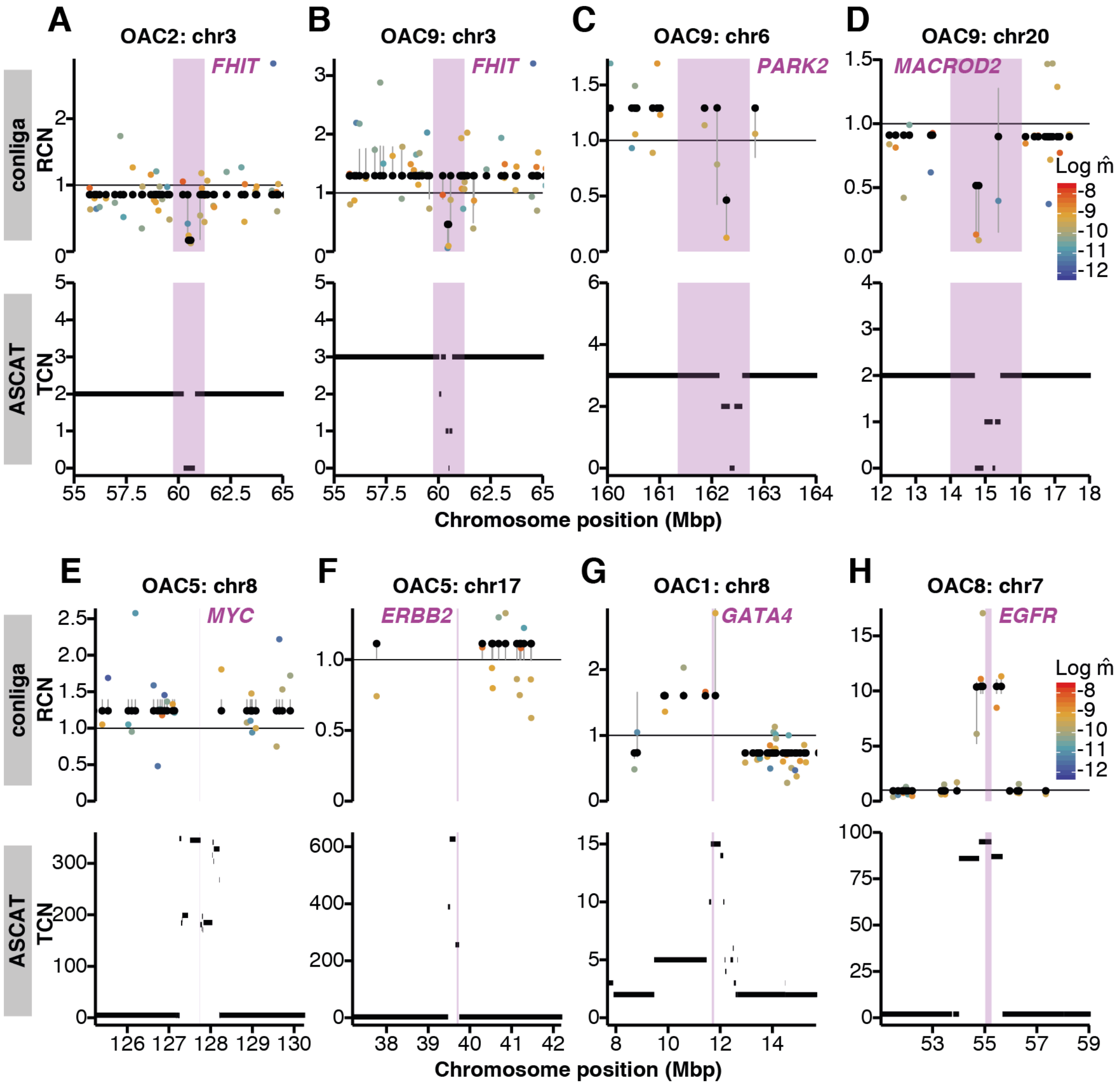


Figure 3.7 Each panel (A-H) shows a comparison between the RCN profile inferred by conliga (top) with the TCN profile inferred by ASCAT (bottom) for a specific region. Black points represent the MAP RCN inferred by conliga. The vertical grey lines indicate the 90% credible interval of the loci RCN. The coloured points show the scaled count with the colour representing the inferred value for $\log(\hat{m})$. The pink regions correspond to the genomic coordinates of the genes of interest. Intra-gene deletions inferred by conliga and ASCAT of: A-B: *FHIT*, C: *PARK2* and D: *MACROD2*. Focal amplification of regions including genes inferred by ASCAT but not by conliga due to the altered region not including FAST-SeqS loci, E: *MYC* and F: *ERBB2*. G: An alteration in the vicinity of *GATA4* is inferred by ASCAT and conliga. A focal amplification including *GATA4* is inferred by ASCAT but not shown by conliga's MAP RCN. However, the 90% credible interval indicates that there is some probability of a focal amplification. H: Amplification involving *EGFR* inferred by ASCAT and conliga.

intra-gene alterations. In other cases, focal alterations can be missed due to the events occurring between FAST-SeqS loci. If the copy number segment is sampled, more accurate quantification might be achieved by increasing the number of reads per sample or decreasing the technical variation in the protocol. Clearly, there are limitations in using FAST-SeqS data to detect gene-level alterations. Including the targeted sequencing of a gene panel alongside FAST-SeqS may be useful, particularly for genes which do not have FAST-SeqS loci in close proximity.

3.2.6 Exploring performance in low purity samples

Non-invasive sampling of tumour DNA using cytology-retrieval devices or cfDNA from liquid biopsies will typically result in low purity samples; that is, samples with low tumour to normal DNA ratio [102, 209]. Of course, sampling tissue or cfDNA from an individual without disease, should result in a sample free of DNA with large-scale structural variation. Even samples derived from direct tumour biopsies can have widely varying tumour purities [103]. As tumour purity decreases, the relative copy number signal to noise ratio decreases. This is the case for WGS and FAST-SeqS data and, at some point, distinguishing between noise and signal is challenging. In this section, I explore conliga's performance in detecting aberrant copy number profiles in samples with low tumour content and compare it to QDNAseq.

I generated *in silico* FAST-SeqS and LC WGS samples with varying tumour purity by mixing sequencing reads from OAC and normal samples. *In silico* FAST-SeqS and LC WGS samples were generated by using two million and nine million reads respectively (see section 3.5.1).

Figure 3.8A shows the performance of conliga and QDNAseq in calling RCN profiles for various dilutions of sample OAC3. Prior to performing insilico dilution, this sample had a tumour purity of 60% (as inferred by ASCAT). At 30% purity, conliga and QDNAseq are able to reproduce a similar copy number profile to that called by ASCAT on the undiluted sample. We see that QDNAseq fails to detect some of the SCNAs; it does not detect any changes in chromosome 13q and fails to detect a loss towards the 3' end of 1p. At 5% purity, QDNAseq appearingly fails to detect any intra-chromosomal SCNAs other than the focal amplifications involving chromosome 12, whereas conliga shows evidence of chromosome arm, sub-chromosomal arm (e.g. chr5q) and focal alterations (chromosome 12). At 2% purity, conliga is able to detect the more prominent chromosome arm gains (1q, 4q, 5p, 12p and 12q). The focal amplification on chromosome 12 is detected as low as 0.5% purity whereas QDNAseq fails to detect it below 1%.

It can be difficult to distinguish between signal and noise generated by the RCN segmentation in the profiles produced by QDNAseq. This problem is more clearly seen in figure 3.9 showing the dilution of sample OAC9. Since conliga globally clusters the loci into copy

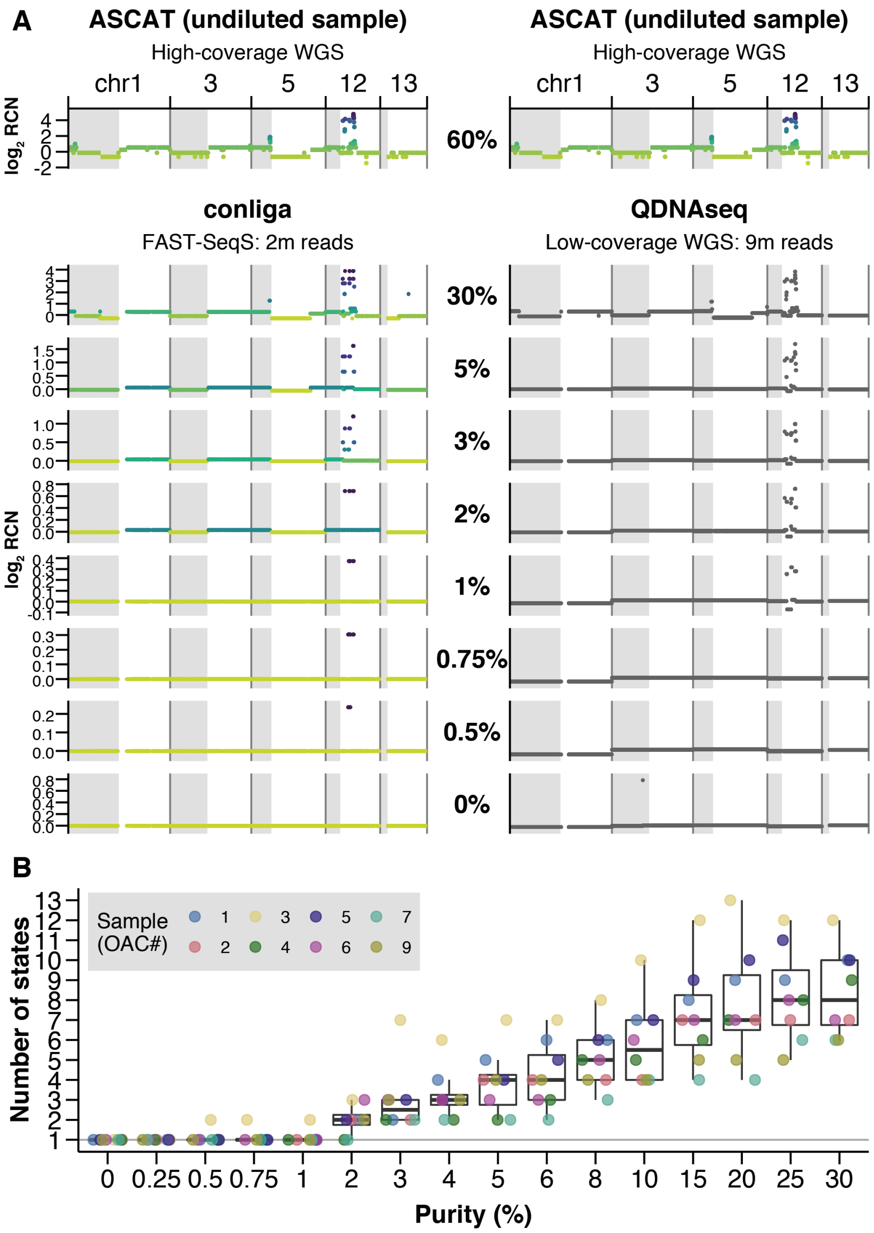


Figure 3.8 Determining the limit of SCNA detection in low tumour purity samples: comparing the performance of conliga and QDNAseq. A: left column: RCN calls using the *MAP state* method provided by conliga at various dilutions of sample OAC3, compared to ASCAT's converted RCN profile (top left). Copy number states are coloured with a gradient (light green to purple), highlighting regions with differing SCNAs. Right column: RCN calls by QDNAseq at various dilutions of sample OAC3, compared to ASCAT's converted RCN profile (top right). B: The number of copy number states detected by conliga (using the *MAP state* method) in each of eight OAC samples at differing purity levels. The limit of detection is determined by the lowest purity level in which more than one copy number state is detected.

number states, we can use the *MAP state* summarisation method to harness the hidden state assignments. This means we can discriminate between SCNAs whereas we are unable to do this using tools such as QDNAseq. This feature is particularly useful in a detection setting, in which we wish to distinguish between normal samples and those with aberrant copy number profiles.

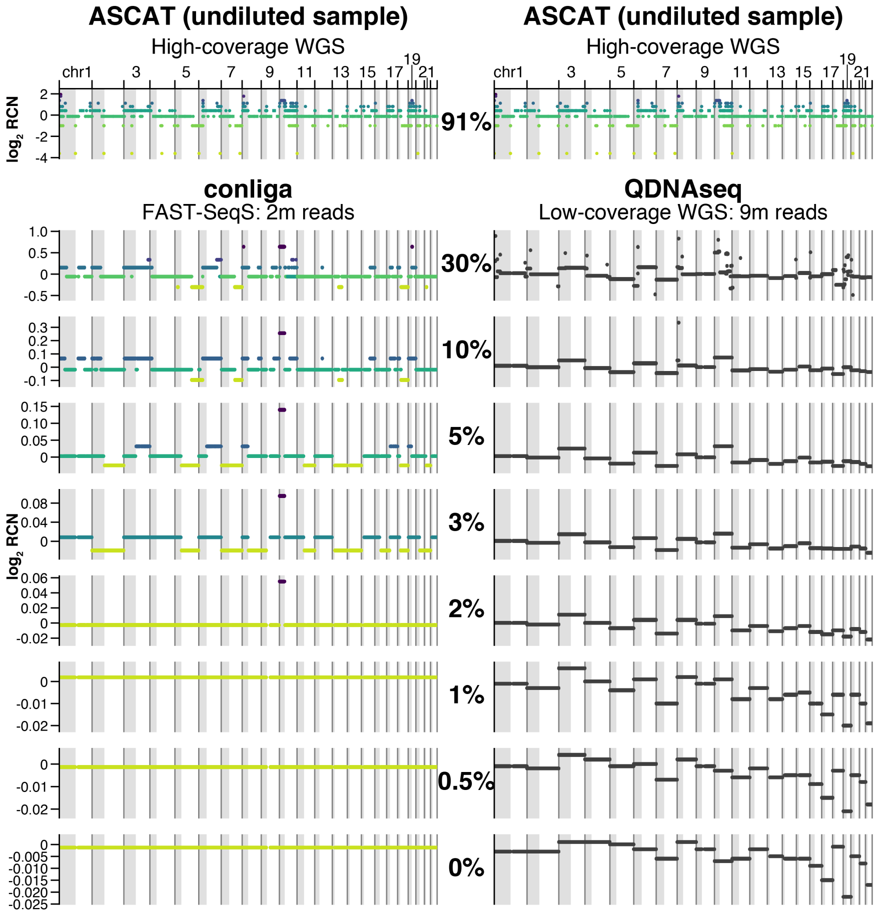


Figure 3.9 Segmentation noise in QDNAseq RCN profiles: *in silico* dilution of sample OAC9. A: left column: RCN calls using the *MAP state* method provided by conliga at various dilutions of sample OAC9, compared to ASCAT's converted RCN profile (top left). Copy number states are coloured with a gradient (light green to purple), highlighting regions with differing SCNAs. Right column: RCN calls by QDNAseq at various dilutions of sample OAC9, compared to ASCAT's converted RCN profile (top right).

In figure 3.8B, we see that conliga is able to detect SCNAs at 3% purity in all the eight diluted samples. SCNAs were detected in five out of the eight samples at 2% purity and one at 0.5% (OAC3; figure 3.8A). The limit of detection is dependent on a number of factors. These include the amplitude and lengths of the SCNAs present in a sample. By sharing and propagating statistical strength to neighbouring loci via the Markov chain, long SCNAs involving chromosome arms can be detected as low as 2-3% purity. This is particularly true for those with several gains. Increased sensitivity is achieved by sharing the statistical strength of all FAST-SeqS loci to infer global clusters of copy number states. Highly focal amplifications, particularly those involving FAST-SeqS loci with a bias towards having a high number of reads align to them (high $\hat{m}$ value), can be detected at approximately 0.5% purity. In addition, the limit of detection depends on the total number of reads per sample and the technical variability (which is sample-specific). Increasing the total number of reads beyond two million and reducing the technical variability in the protocol would assist in further improving the limit of detection.

Figures 3.8A, 3.8B and 3.9, also show that as purity reduces, fewer copy number states are inferred and they are seemingly merged together. As tumour purity reduces, the scaled count distributions centred around each true copy number state will increasingly overlap. At some purity level, there will be insufficient statistical strength for conliga to discern every true copy number state. Indeed, we have already seen this in high purity samples in which multiple focal amplifications (involving few loci) are merged together into a single copy number state. This highlights an important feature of Bayesian nonparametric models; there is an inherent preference for simpler models, often referred to as Bayesian Occam's Razor [230]. In our case, a simpler model refers to a solution with fewer copy number states. Hence, we should keep in mind that conliga's inferred number of states may not always map directly to the true number of copy states in the sample. However, this inbuilt Occam's Razor is what may allow us to distinguish normal samples from those that contain evidence of SCNAs.

The results presented in this section suggest that conliga is more sensitive than QDNAseq in detecting and discriminating between samples with normal and aberrant copy number profiles. This is despite FAST-SeqS data being technically noisier than LC WGS and with conliga using 4.5 fold fewer reads than QDNAseq. Moreover, these results suggest FAST-SeqS and conliga could have useful clinical application in the early detection of OAC (and potentially its precursor lesions). Indeed, its use combined with a non-invasive cytology-retrieval device such as the Cytosponge is promising. Using the Cytosponge, Weaver *et al.* showed that *TP53* mutations were detected in Barrett's (HGD) samples with allele fractions ranging from 0.6% to 35.7% [209]. 12 of the 22 samples had allele fractions above 2% but 7 had allele fractions around 0.06-2% (and 3 had no detectable mutations). This range of allele fractions and the results of the dilution experiment imply that the total number of reads may need to be increased to more reliably detect SCNAs in these samples.

3.3 Investigating SCNA profiles of OAC, GAC and BO samples

In this section, we briefly explore examples of SCNAs inferred across several samples from the OAC, GAC and BO cohorts. Tables 3.5 and 3.6 provide information about the samples within each cohort for GAC and BO samples respectively. Once again, we see that the total number of reads can vary considerably between samples. However, all samples had greater than 350,000 reads and were included in the analysis.

Sample	Seq. ID	Reads	Counts	Counts (pf)
GAC1	SLX-9282.FastSeqA	589050	292499	248532
GAC2	SLX-9282.FastSeqB	1479548	849084	740760
GAC3	SLX-9282.FastSeqC	1874367	1030135	898913
GAC4	SLX-9282.FastSeqD	3914819	2174418	1903357
GAC5	SLX-9282.FastSeqE	1503187	779514	643384
GAC6	SLX-9282.FastSeqF	1391575	741848	642406
GAC7	SLX-9282.FastSeqG	2935085	1549146	1310065
GAC8	SLX-9282.FastSeqH	1855764	1048785	934323

Table 3.5 Gastric adenocarcinoma cohort: FAST-SeqS sample information. ‘Reads’ represents the total number of reads assigned to each sample demultiplexing the reads from the sequencing lanes. ‘Counts’ represents the number of counts after each sample’s FASTQ file was processed using the pipeline described in section 2.4. ‘Counts (pf)’ shows the number of counts that remained after filtering loci with a zero count in any of a set of control samples (see section 3.5 for details).

Figure 3.10 shows some examples of SCNA profiles within particular chromosomes in several OAC, GAC and BO samples, harbouring alterations to several cancer genes. For example, figure 3.10A shows the amplification of *PRKCI* (a serine/threonine protein kinase) and the focal deletion of *FHIT*. PRKCI expression is implicated in the resistance to apoptosis, uncontrolled proliferation and metastasis and is a potential target for therapy in many cancer types [231]. We see several examples of tyrosine kinase receptor amplification; *EGFR* in OAC (figure 3.10B), *MET* in GAC (figure 3.10F) and *ERBB2* in BO (figure 3.10K). These receptors have potential as therapeutic targets. Interestingly, *EPHB6*, which also encodes for a tyrosine kinase receptor, is deleted in GAC8 (figure 3.10F). It has been noted that the loss of *EPHB6* in colorectal cancer contributes to metastasis and could play a similar role in OAC [232]. As, we have discussed previously, *TP53* plays a critical role as a tumour suppressor and is often deleted in cancer and pre-maligancies. Indeed, we see evidence of its deletion in an OAC and BO samples (figures 3.10D and 3.10K, respectively). We see SCNAs which may dysregulate the cell cycle pathway in a number of samples. For example, we see amplifications of *CDK6*, *CCND1* and *CCNE1* in GAC samples in figures 3.10F, 3.10G

Sample	Seq. ID	Reads	Counts	Counts (pf)
BO1	SLX-9283.FastSeqA	1748991	835963	742765
BO2	SLX-9283.FastSeqB	765189	404734	357402
BO3	SLX-9283.FastSeqC	1517421	773882	687025
BO4	SLX-9283.FastSeqD	1178002	590798	526001
BO5	SLX-9283.FastSeqE	2452075	1167884	1001412
BO6	SLX-9283.FastSeqF	1874540	933139	829790
BO7	SLX-9283.FastSeqH	2322558	1224522	1133421
BO8	SLX-9283.FastSeqI	977646	543048	484150
BO9	SLX-9279.FastSeqA	860807	297284	244491
BO10	SLX-9279.FastSeqB	750576	271747	229469
BO11	SLX-9279.FastSeqC	916640	347410	286348
BO12	SLX-9279.FastSeqD	1008211	362594	298655
BO13	SLX-9279.FastSeqE	1004855	323423	260074
BO14	SLX-9279.FastSeqF	874117	320774	269500
BO15	SLX-9279.FastSeqG	727422	257321	214561
BO16	SLX-9279.FastSeqH	976859	283664	234321

Table 3.6 Barrett's oesophagus cohort: FAST-SeqS sample information. 'Reads' represents the total number of reads assigned to each sample demultiplexing the reads from the sequencing lanes. 'Counts' represents the number of counts after each sample's FASTQ file was processed using the pipeline described in section 2.4. 'Counts (pf)' shows the number of counts that remained after filtering loci with a zero count in any of a set of control samples (see section 3.5 for details).

and 3.10H respectively. As discussed, CDK6 is a potential target for therapy. Chromosome 9p (which includes *CDK2NA* and *PTPRD*) shows evidence of deletion in a BO sample (figure 3.10J). *CDK2NA* acts as a cell cycle regulator and tumour suppressor gene which is recurrently deleted in many cancer types [233]. We see evidence amplifications of *GATA4*, *GATA6* and *MYC* which are transcription factors commonly amplified in OAC (figures 3.10C, 3.10L and 3.10C respectively). Lastly, in figure 3.10G, we see deletion of *ATM* which is a mediator of DNA damage response to double strand breaks and is frequently mutated in several cancer types [234]. Its deletion suggests ATR inhibitors could be used as a targeted therapy [235]. Other deletions including tumour suppressors are observed such as *SMAD4* (figure 3.10L) and *PARK2* (figure 3.10E).

3.3.1 Average SCNA profiles for patient cohorts

The low-cost of FAST-SeqS could enable large cohorts of patients to be profiled for SCNAs. This could help identify recurrently amplified and deleted genes that are associated with response to therapy, for example. In figure 3.11, we see the mean RCN profiles for the cohorts

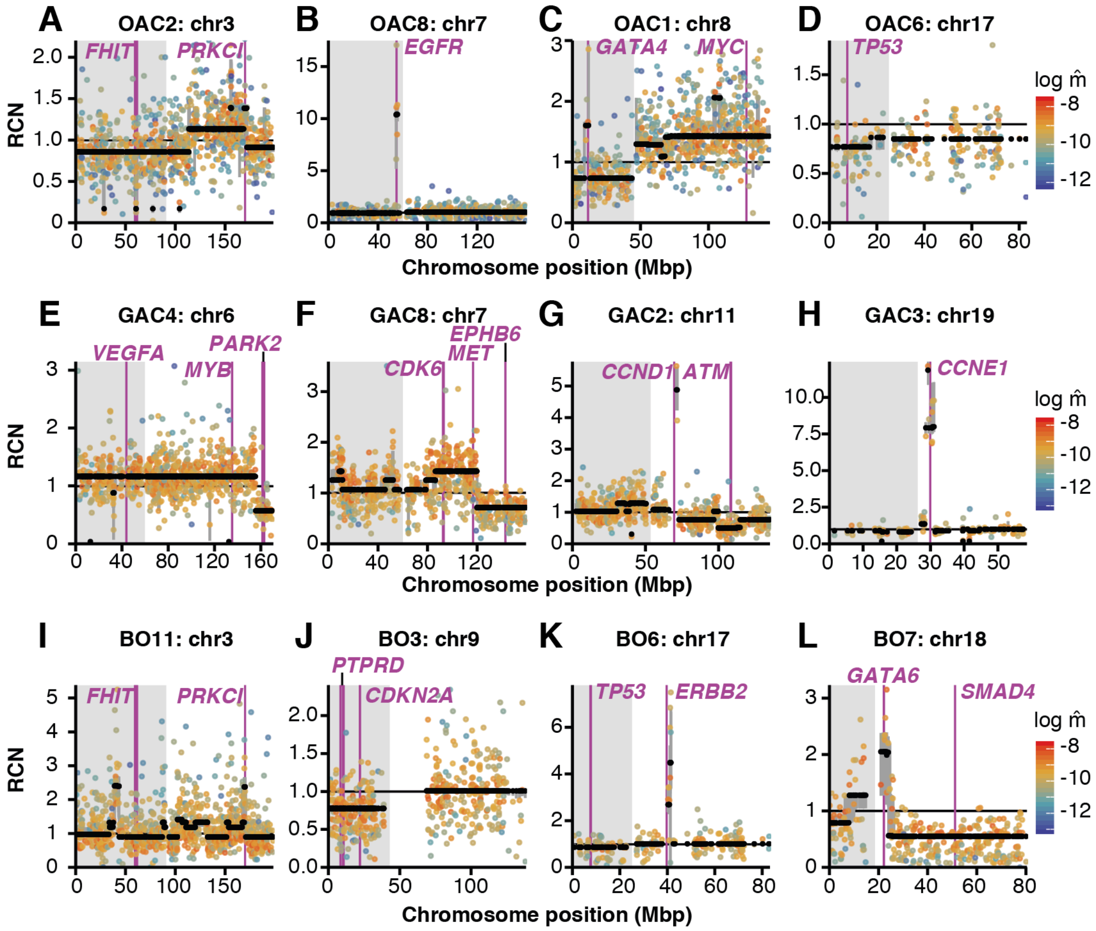


Figure 3.10 SCNA profiles inferred by conliga in several chromosomes for a selection of OAC, GAC and BO samples. Recurrently amplified and deleted genes in OAC are highlighted. Black points represent the MAP RCN inferred by conliga. The coloured points show the scaled count with the colour representing the inferred value for $\log \hat{m}$. The grey regions of each chromosome indicate the p arm while the white regions indicate the q arm.

of OAC, GAC and BO patients. These highlight amplifications of *EGFR*, *GATA4*, *MDM2* and *MYC* along with deletions of tumour suppressor genes, e.g. *FHIT*, *RUNX1*, *SMAD4* and *TP53*. MDM2 inhibits p53 and its amplification may reflect another mechanism of p53 suppression (other than *TP53* mutation or loss). Here we see that similar regions are amplified or deleted between OAC and GAC. The mean RCN profile for OAC is similar to that seen in figure 2B of Ross-Innes *et al.* [211]. As expected, we see less prominent SCNAs in the BO cohorts.

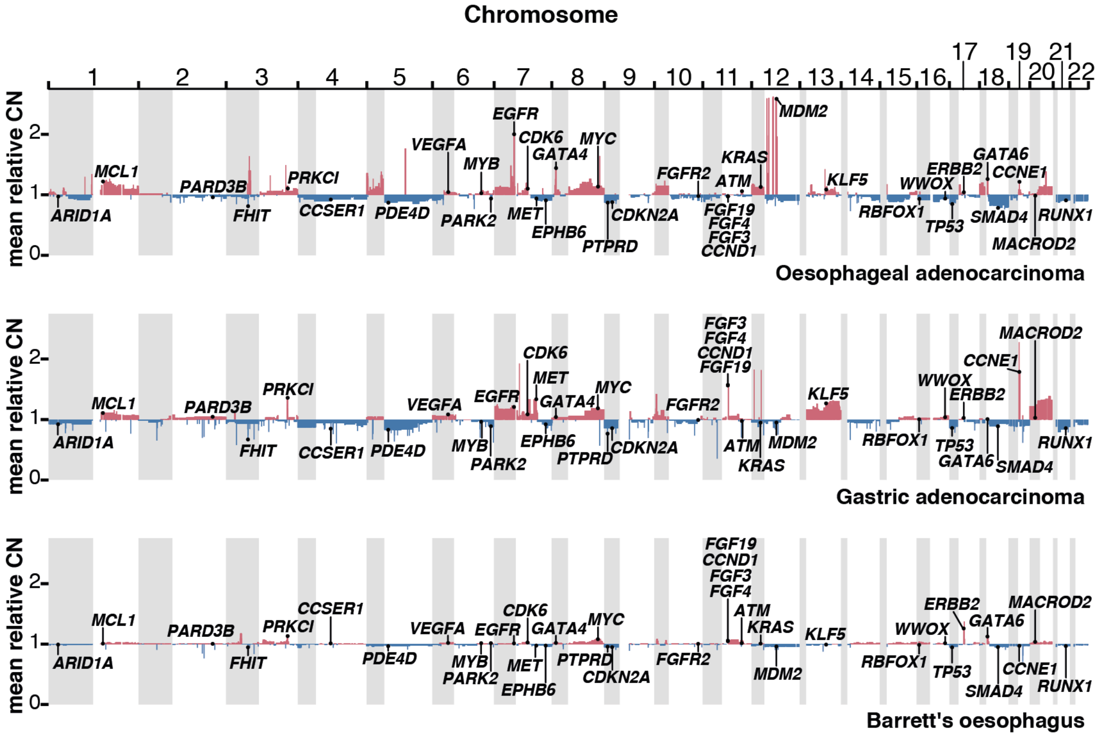


Figure 3.11 Mean inferred RCN profiles for patient cohorts (oesophageal adenocarcinoma, gastric adenocarcinoma and Barrett's oesophagus). Top: oesophageal adenocarcinoma (n=11), Middle: gastric adenocarcinoma (n=8), Bottom: Barrett's oesophagus (n=16). The location of the 36 genes that were selected as recurrently amplified or deleted in OAC are highlighted for each cohort.

3.4 Summary

In this chapter, I introduced oesophageal cancer; covering its histological types, epidemiology and its significance as a global disease. While the incidence of oesophageal squamous cell carcinoma is falling around the world overall, the rapid rise in incidence of oesophageal adenocarcinoma in the West is concerning. Oesophageal cancer has a very poor prognosis, owing to symptoms occurring late in the development of disease. This means that patients are often diagnosed with incurable metastatic disease. Treatments are much more successful in the early stages of disease and screening and surveillance programmes are necessary to improve prognostic outcomes. However, the current costs and invasive means of current screening and surveillance modalities mean they cannot be applied at scale. This means that screening and surveillance is restricted to those deemed to be at high-risk of developing disease. However, this process is suboptimal and the cost-effectiveness is debated. New, low-cost and noninvasive technologies and biomarkers are required to more successfully diagnose patients with early and curable disease. Somatic copy number alterations have been shown to accumulate in the progression to oesophageal adenocarcinoma and are promising biomarkers for early disease detection and patient stratification.

The use of FAST-SeqS has potential for widespread clinical use and SCNA profiling of large patient cohorts, owing to its low-cost and simple protocol. By comparison to ASCAT, I have shown that FAST-SeqS and conliga show comparable performance to LC WGS and QDNAseq at reduced cost and using fewer reads. In some cases, FAST-SeqS and conliga are able to detect amplifications and deletions of recurrently amplified and deleted genes, including relevant therapeutic targets. conliga appears to be more sensitive than QDNAseq at low tumour purity and may be more suitable for diagnostic purposes.

3.5 Methods

This section is largely the same as the methods section described in Abujudeh *et al.* [105], a preprint manuscript which I authored and is shared on bioRxiv. Since the methods were very similar to those presented in this thesis, I reproduce them here.

3.5.1 Sample preparation and sequencing of samples

Sample preparation and sequencing of FAST-SeqS data

FAST-SeqS samples were prepared and sequenced as described in section 2.3.

Sample preparation and sequencing of HC WGS data

WGS library preparation and sequencing was performed as previously described by Secrier *et al.* [221].

In silico generation of LC WGS data

For our purposes, LC WGS data was defined as nine million single-end 50 base pair reads per sample because this was the type of data analysed in Scheinin *et al.* [225]. Samples are typically multiplexed together and sequenced on a single Illumina sequencing lane. After processing and alignment of the reads, we expect approximately 0.1X coverage of the genome (as per analysis described in Scheinin *et al.*). I obtained LC WGS data by down-sampling reads from HC WGS Binary Alignment Map (BAM) files in the following way:

1. I selected a subset of the alignments, containing only reads sequenced on a single lane (chosen to be the lane from the first read in the BAM file), and trimmed the reads and Phred scores to the first 50 base pairs using a custom Bash script.
2. The resulting alignments were filtered (using samtools [236] version 0.1.18), excluding those that were secondary alignments (-F 256) and including only those that were first in a pair (-f 64) and output to a new BAM file.
3. This BAM file was down-sampled to 9 million reads/alignments using the DownsampleSam command from Picard tools (http://broadinstitute.github.io/picard, version 2.9.1) using the "Chained" strategy.
4. The resulting BAM file was converted to FASTQ by SamToFastq (Picard tools).
5. The FASTQ file was aligned to GRCh38 (GenBank accession: GCA_000001405.15, no alt analysis set) using BWA-backtrack (bwa samse and bwa aln, version 0.7.15-r1140), which is more suitable for reads below 70 base pairs in length.
6. In the resulting BAM file, I removed PCR duplicates and removed alignments with mapping quality below 37 as per the analysis undertaken by Scheinin *et al.* [225] using samtools (version 0.1.18).

I performed these steps for 11 oesophageal samples and their matched normal samples along with an additional four normal samples obtained from other patients (see Supplementary Table 1 from Abujudeh *et al.* [105]). This resulted in greater than seven million primary alignments per sample.

In silico generation of FAST-SeqS dilution series data

I performed an in silico dilution of FAST-SeqS data by mixing sequencing reads from control samples with reads from OAC samples. Since the number of reads in the matched controls were insufficient to create samples with two million reads, I created a pool of control reads (in silico) which were used to dilute the OAC samples. This was done by sub-sampling two million reads from 12 control samples (which were prepared and sequenced in the same batch as the OAC samples). The total number of reads from these 12 control samples was 14,405,596. To obtain a pool of 2 million reads, I used the ‘sample’ command from seqtk (https://github.com/lh3/seqtk, version: 1.2-r101) to sample a proportion (2/14.405596) of each control sample’s reads and merged these together into a single FASTQ file. The reads that were sub-sampled were removed from the control samples (using a custom python script) to avoid using the same reads to fit $\hat{\boldsymbol{m}}$.

I mixed the pool of control reads with the OAC samples in varying proportions to achieve a desired diluted tumour purity. The OAC samples did not have a tumour purity of 100%, instead they were themselves a mixture of tumour and normal DNA. The purity of these samples were determined by ASCAT-NGS (version 2.1) [224]. Based on ASCAT's purity value, I calculated the number of reads required from the OAC sample to achieve a desired dilution and total number of reads. This was calculated as follows:

$$\text{required tumour reads} = \text{round}\left(\frac{\text{desired purity proportion} \cdot \text{required total reads}}{\text{ASCAT inferred purity proportion}}\right) \tag{3.1}$$

Hence, the number of control reads required were:

$$\text{required control reads} = \text{required total reads} - \text{required tumour reads} \tag{3.2}$$

I produced in silico dilution FASTQ files in the following way:

1. I used the 'sample' command from seqtk to sample the required number of tumour reads from the OAC FAST-SeqS FASTQ file
2. I used the 'sample' command from seqtk to sample the required number of control reads from the pooled control reads FASTQ file
3. I merged the sampled tumour and control reads into a single FASTQ file

I performed these steps for each OAC sample to create diluted samples with two million total reads and the following purity values: 0.3, 0.25, 0.2, 0.15, 0.1, 0.08, 0.06, 0.05, 0.04, 0.03, 0.02, 0.01, 0.0075, 0.005, 0.0025 and 0. Here purity is defined as the proportion of tumour reads in the sample. Of the 11 OAC samples, 8 (OAC1-7 and 9, see Supplementary Table 1 Abujudeh *et al.* [105]) were of sufficient initial tumour purity to feasibly create all the desired dilution levels.

In silico generation of LC WGS dilution series data

I produced in silico diluted LC WGS tumour samples by mixing reads from tumour and matched normal LC WGS BAM files (previously downsampled and filtered as described above). I first calculated the number of reads in the tumour BAM and normal BAM files using samtools (`samtools view -F 256 -c [BAM file]`). Next, I calculated the number of reads required using equations 3.1 and 3.2. Using the DownsampleSAM command (Picard tools) and the 'HighAccuracy' strategy, I sampled the corresponding desired proportion of reads from the tumour BAM file and normal BAM file. I used samtools to merge the resulting sampled tumour BAM file with the normal BAM file into a single file representing the diluted sample. I aimed to obtain seven million filtered primary alignments per diluted sample (as this is what we expect from nine million reads after alignment and filtering) and dilution levels which matched the diluted FAST-SeqS samples. This was performed for 8 OAC samples and their matched normals (OAC1-7 and 9).

3.5.2 Analysis of copy number from FAST-SeqS data

conliga (version 0.1.0) [108] was used to analyse all FAST-SeqS samples in this study (Supplementary Table 1 from Abujudeh *et al.* [105]) using R (version 3.2.3) [237] and RcppAramdillo (version 0.6.500.4.0) [147]. Of the 15 OAC samples sequenced, four were excluded due to their obtaining fewer than 350,000 reads. Two control samples were excluded due to their inferred RCN profiles having two main hidden states incompatible with their supposed 'normal' status. The values for the priors used and MCMC settings are stated in the inference sections above. The samples used as a basis to filter loci and fit $\hat{\boldsymbol{m}}$ for each experiment are listed in Supplementary Table 9 from Abujudeh *et al.* [105].

3.5.3 Analysis of copy number from HC WGS data

High coverage WGS samples were processed and aligned using BWA-MEM [238] (version 0.5.9) and TCN profiles and normal contamination estimates were provided by ASCAT-NGS (version 2.1) using a pipeline previously described by Secrier *et al.* [221]. The only exception to this was that the reads were aligned to GRCh38 (GenBank accession: GCA_000001405.15, no alt analysis set) rather than GRCh37.

3.5.4 Analysis of copy number from LC WGS data

QDNAseq (version 1.6.1) was used to obtain relative copy number calls for all LC WGS data. The bin size used was 15Kb as per the analysis performed in Scheinin *et al.* [225] for ~ 0.1X LC WGS. The bins were created using GRCh38 (BSgenome.Hsapiens.NCBI.GRCh38) and a mappability file (bigWig format) for 50-mers was created for GRCh38 using the GEM library (GEM-binaries-Linux-x86_64-core_i3-20130406-045632) https://sourceforge.net/projects/gemlibrary/. 15 normal LC WGS samples (see Supplementary Table 1 from Abujudeh *et al.* [105]), were used to run the applyFilters and iterateResiduals functions. 11 of these 15 samples correspond to the matched normals of the oesophageal samples (Supplementary Table 1 from Abujudeh *et al.* [105]). I did not run the functions normalizeBins and normalizeSegmentedBins which scale the read counts by the median value. This was not necessary and would make the comparison between ASCAT, QDNAseq and conliga results more difficult to interpret.

3.5.5 Comparison of copy number calls between methods

ASCAT outputs TCN in contiguous genomic regions, QDNAseq outputs RCN in 15 Kb bins across the genome and conliga outputs RCN values at specific FAST-SeqS loci. To make a fair comparison between the tools, it was necessary to convert ASCAT's TCN calls to RCN for each sample as follows:

$$\mathrm{RCN}_i = \frac{(1 - \mathrm{normal}) \cdot \mathrm{TCN}_i + \mathrm{normal} \cdot 2}{\mathrm{mean\ TCN}} \tag{3.3}$$

Here, normal represents the estimated normal contamination value provided by ASCAT and i represents a contiguous genomic region or a discrete locus or fragment. In the case of a contiguous region, the mean TCN (or ploidy) was calculated as follows:

$$\mathrm{mean\ TCN} = \frac{\sum_i (\mathrm{TCN}_i \cdot \mathrm{length}_i \cdot (1 - \mathrm{normal}) + 2 \cdot \mathrm{length}_i \cdot \mathrm{normal})}{\sum_i \mathrm{length}_i} \tag{3.4}$$

and in the case of discrete loci or fragments:

$$\mathrm{mean\ TCN} = \frac{\sum_i \mathrm{TCN}_i}{L} \cdot (1 - \mathrm{normal}) + 2 \cdot \mathrm{normal} \tag{3.5}$$

where L represents the total number of loci or fragments considered.

In figure 3.5A, B and C I compared the RCN values at the intersection of genomic loci across ASCAT, QDNAseq and conliga. Since this intersection represented a subset of each method's genomic loci, the RCN values were rescaled considering only this subset. QDNAseq and conliga RCN values were rescaled by the sample's mean RCN of the considered loci. ASCAT's RCN was calculated using equations 3.3 and 3.5.

In figure 3.6B, I compared RCN values in genes of interest. Recurrently amplified and deleted genes were obtained from Dulak *et al.* [219] and Ross-Innes *et al.* [211]. Here, ASCAT's RCN values were calculated using equations 3.3 and 3.4 using all called regions for each sample. For each gene in each sample, the weighted mean of the relative copy number (weighted by the length of the overlapping called region) was computed for ASCAT and QDNAseq. This was calculated as follows:

$$\mathrm{RCN}_{\mathrm{gene}} = \frac{\sum_i \mathrm{RCN}_i \cdot l_i}{\sum_i l_i} \tag{3.6}$$

where l_i represents the length of the overlapping portion of the called region with the gene.

For conliga, if loci occurred within the gene, the mean of the RCN values within the gene was used, otherwise the loci directly upstream and downstream, i.e. either side, of the gene were used and a mean value was taken.

In figure 3.6C, I compared the minimum RCN value associated with recurrently deleted genes in OAC. Here, ASCAT's RCN values were calculated using equations 3.3 and 3.4 using all called regions for each sample. The minimum relative copy number of genomic regions that overlapped the gene was reported. For conliga, if loci occurred within the gene, the minimum RCN value was reported. Otherwise the minimum RCN of the flanking loci was reported.

In figure 3.6D, I compared the minimum RCN value associated with recurrently deleted genes in OAC. Here, ASCAT's RCN values were calculated using equations 3.3 and 3.4 using all called regions for each sample. The maximum relative copy number of genomic regions that overlapped the gene was reported. For conliga, if loci occurred within the gene, the maximum RCN value was reported. Otherwise the maximum RCN of the flanking loci was reported.

Computing Pearson correlation

For each sample, the Pearson correlation coefficient between ASCAT and conliga was calculated. I used ASCAT's TCN and conliga RCN values at the intersection of genomic loci between ASCAT and conliga. The median value of the sample's correlation coefficients was reported.

For each sample, the Pearson correlation coefficient between ASCAT and QDNAseq was calculated. I used the intersection of QDNAseq bins with ASCAT copy number regions, using the length-weighted mean of ASCAT's overlapping TCN values.

When calculating the Pearson correlation for all calls across all samples, I used the rescaled RCN value at the intersecting genomic loci between ASCAT, QDNAseq and conliga, using the rescaled RCN values described above for figures 3.5A, B and C.

3.5.6 Data availability

The WGS and FAST-SeqS data can be found at the European Genome-phenome Archive (EGA) under accession EGAD00001004289.

Chapter 4

Modifications to the primer design, processing pipeline and conliga model

The experiments that were conducted to produce the data presented in this chapter were performed by Sarah Field of the Tavaré laboratory at the Cancer Research UK Cambridge Institute. The rest of the work is my own.

4.1 Modifying the primer design for higher-throughput and paired-end sequencing

The samples analysed in the previous chapters were sequenced by an Illumina MiSeq machine. The MiSeq can sequence up to 25 million molecules per run. Since we wanted to achieve approximately 2 million reads per sample, we multiplexed 11 samples on each run of the machine. These data showed that the FAST-SeqS protocol has potential as a tool for copy number quantification and detection. However, we needed to utilise the higher-throughput instruments, such as the HiSeq 4000 and NovaSeq machines to decrease the cost per sample and make it more cost-effective.

At the time of our decision, the HiSeq 4000 provided the most cost-effective means at our disposal to scale up the FAST-SeqS protocol to more samples. The HiSeq 4000's patterned flow cell is comprised of 8 sequencing lanes and each lane can yield up to 350-400 million paired-end reads at approximately double the cost of a MiSeq run. When the machine is run, all lanes must be sequenced in the same mode. For example, if single-end 150 bp is required, all 8 lanes must be run in that mode. By far the most common mode in the institute is paired-end 150 bp. In addition, the difference in cost between single-end and paired-end was small. For these reasons, we chose to perform 150 bp paired-end sequencing rather than single-end.

Utilising the HiSeq 4000 poses a number of challenges. The HiSeq, NextSeq and NovaSeq machines are known to be less tolerant to low-diversity libraries (such as amplicon libraries) when compared to the MiSeq. To compensate for this, we would likely need to increase the proportion of PhiX spiked into our sequencing libraries to increase the diversity of the library. This would decrease the cost-effectiveness by decreasing the number of FAST-SeqS molecules we could sequence. In an attempt to mitigate this issue, I decided to modify the first round PCR primers. I included UMIs on the forward and reverse primers and allowed the length of the UMIs to vary (8 - 12 bp on either end). Including UMIs on both primers means that read1 and read2 start with random sequences which allows the sequencer to calibrate itself. Varying the length of the UMIs means that the low-diversity sequences (the L1 sequences) within the amplicons are phased/shifted, such that they would start at different positions within the forward and reverse reads. This meant that the complexity of the library would be increased. I also included UMIs on both ends of the molecule for the purposes of single cell FAST-SeqS experiments, which we discuss further in chapter 5.

Another issue was that I needed to increase the number of sample indices to enable more samples to be sequenced on the same lane. To ensure sufficient complexity of our library, we decided to spike in 20-30 % PhiX, meaning that we would sequence approximately 245 to 280 million FAST-SeqS molecules and 70 to 105 million PhiX molecules. Aiming for 2 million molecules per sample would mean approximately 120 to 140 samples per lane. With the new UMI design described above, it may be possible to reduce the amount of PhiX (perhaps removing its use entirely) while maintaining good sequencing quality (though we have not tested this). This would mean sequencing up to 175 samples on each lane.

It had been reported that free primer within the sequencing library can result in index swapping on the flow cell [239–241]. This can result in molecules from one sample being falsely assigned to another sample. It has been estimated that in some conditions, this can result in as many as 10% of the indices swapping [239] and was reported to be a particular problem for the machines that use the patterned flow cell technology, such as the HiSeq 3000/4000 and NovaSeq machines [240]. A more recent study by Macconaill *et al.* showed index swapping frequencies of 0.29% for patterned flow cells and 0.1% on non-patterned flow cells [241], suggesting index swapping is less severe than previously reported. With this uncertainty in mind, I decided to move to a unique dual index (UDI) primer design (figure 4.1) in which each sample has a unique index on either side of the molecule. These indices are incorporated into the forward and reverse primers for the second round PCR reaction. If an index swap occurs at one of the ends, we will be able to detect this and remove the molecule from downstream analyses without falsely assigning it to another sample. For a false assignment to take place, the unlikely event of two swaps would need to occur at either end of the molecule and both must switch to a unique combination identifying another sample.

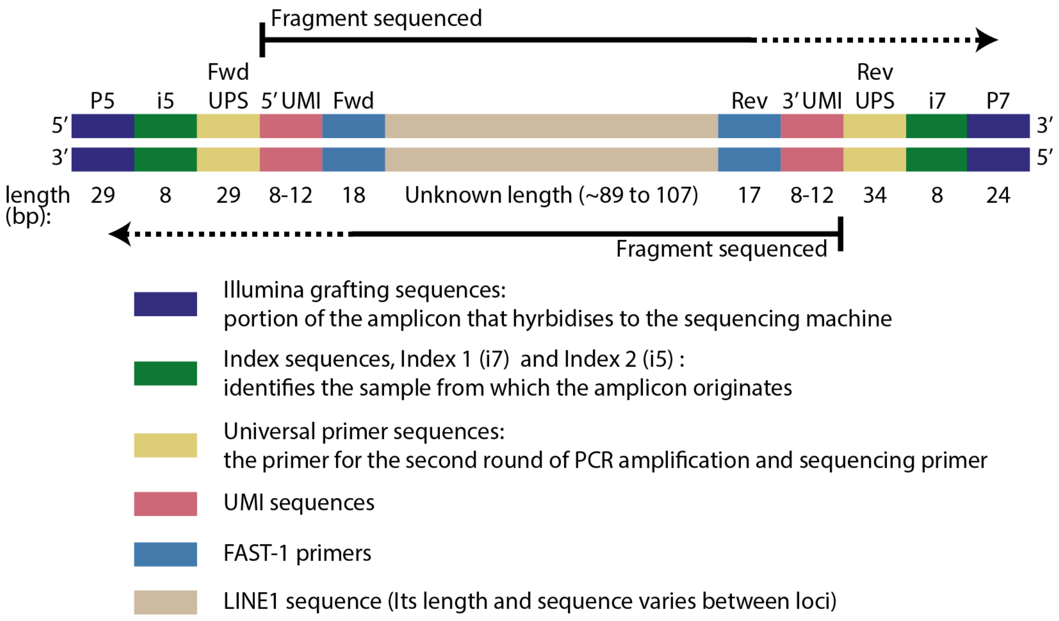

Fragment sequenced
Fwd UPS
Rev UPS
P5
i5
5′UMI
Fwd
Rev
3′UMI
i7
P7
5′
3′
3′
5′
length (bp):
29
8
29
8-12
18
Unknown length (~89 to 107)
17
8-12
34
8
24
Fragment sequenced
Illumina grafting sequences: portion of the amplicon that hyrbidises to the sequencing machine
Index sequences, Index 1 (i7) and Index 2 (i5) : identifies the sample from which the amplicon originates
Universal primer sequences: the primer for the second round of PCR amplification and sequencing primer
UMI sequences
FAST-1 primers
LINE1 sequence (Its length and sequence varies between loci)

Figure 4.1 A diagram showing the composition of the FAST-SeqS amplicon sequencing library incorporating modifications to the PCR1 and PCR2 primer sequences. Together, the multicoloured bars represent a double-stranded amplicon produced by a FAST-SeqS experiment. Each coloured bar represents a nucleotide sequence associated with different elements of the amplicon. The beige section in the middle of the amplicon represents the target LINE1 sequence. The lengths for each element of the amplicon is listed underneath. The arrowed lines above and below the amplicon indicate the portion of the amplicon that is sequenced in read1 and read2 respectively when performing 150 bp paired-end sequencing. Since the LINE1 sequences vary in length, the resulting sequencing reads include varying portions of the primers (Fwd/Rev), forward/reverse UPS and sample indices. A dotted line indicates the uncertainty regarding the elements of the amplicon the reads contain.

The danger of introducing additional sample-specific bases (i.e. indices) is that it increases the likelihood of sample-specific effects. During PCR, we might imagine that these unique sequences could interact with the L1 sequences and they may do so in different ways depending on the sequence context. This may lead to altered efficiency of loci amplification between samples which would break an essential assumption of our model and analysis. In addition to this, we may alter the chance of interactions and hybridisation between the primers (primer-dimer) during PCR. These interactions may vary between samples and result in modified amplification dynamics in the second round PCR. While this may turn out to be an issue, there might be ways to minimise the interactions by choosing index combinations that minimise primer-dimers and interactions with the L1 sequences. In addition, we may be able to model and correct the consequences of such iterations. While I started to design a way to choose UDI combinations, in the end I decided to use well-tested UDI combinations that should minimise primer-dimers. As such, I used the NEBNext UDI sequences which provide 96 unique combinations and hence allow 96 samples to be multiplexed on a single sequencing lane.

Another issue is the use of custom sequencing primers. It is common practice for independent sequencing experiments to be run on each lane of the HiSeq 4000 (and other multi-lane instruments). This is in contrast to the MiSeq which has only one lane, meaning only one sequencing experiment can be run on the machine at any time. The use of custom sequencing primers is more complicated on the HiSeq 4000 because they may interfere with other sequencing lanes. Unfortunately, the HiSeq 4000 is designed such that custom read2 sequencing primers must be delivered to all lanes on the flow cell while custom read1 sequencing primers can be fed to each lane separately. In addition, the hybridisation temperatures of the sequencing primers vary between Illumina machines with the HiSeq 4000 60 °C and the MiSeq 65 °C [242, 243]. This meant that the custom sequencing primers would need to be modified to work with the HiSeq 4000. To avoid this, I decided to replace the Universal Primer Sequence with Illumina's TruSeq adapter sequence in the same way as Belic *et al.* [106]. This meant that custom sequencing primers would not be necessary.

To summarise, I updated the primers and sequencing in the following ways:

- Paired-end sequencing instead of single-end sequencing
- UMIs of varying lengths added to both ends of the molecule
- Dual unique indices used instead of single index
- Replacement of Universal Primer Sequence with Illumina TruSeq Adapter sequence to avoid the use of custom sequencing primers

4.2 A new UMI-aware pipeline for paired-end FAST-SeqS data

The processing pipeline presented in chapter 2 was similar to that presented in Kinde *et al.* [104] with a few changes. While it seemed to work well for single-end data, it does not handle paired-end data and ignores the UMIs contained within the reads. In addition to this, Bowtie 1 [111] was used to align the processed reads to the genome. This aligner has been superseded by more modern aligners, such as Bowtie 2 [114] and BWA-MEM [238], which are aware of indels and provide more detailed information about mapping quality. These aligners are now routinely used for short-read DNA-seq analyses. Rather than update this pipeline, I decided to create a new pipeline.

The new pipeline is depicted in figure 4.2. It is built with snakemake [244] which, along with the conda package manager, enables the pipeline to be easily distributed to different environments and ensures a fully reproducible workflow. In addition to this, snakemake handles the interface to cluster environments, such as Slurm [245]. This means that hundreds or thousands of samples can be processed at the same time by scaling the pipeline to the resources available.

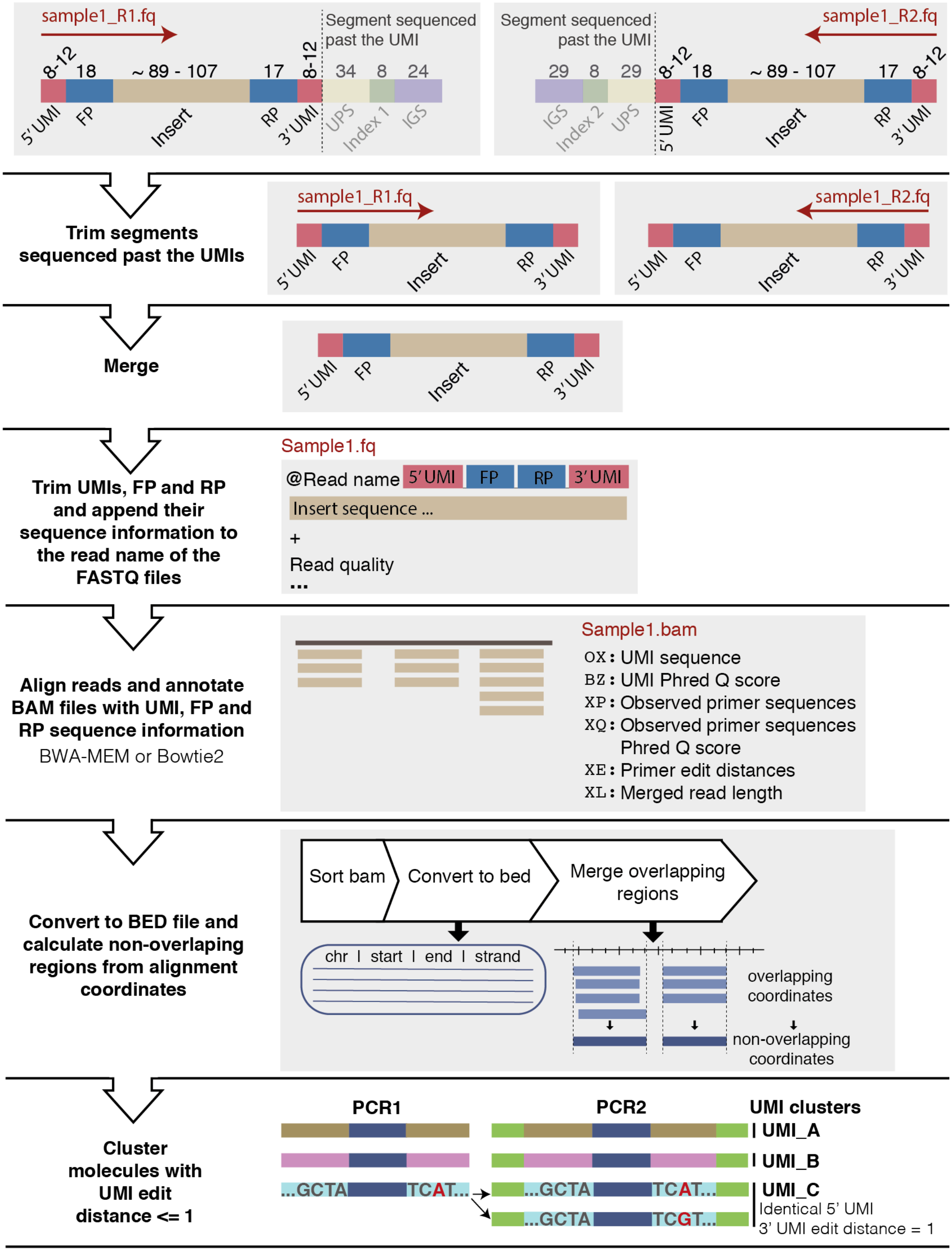

sample1_R1.fq
Segment sequenced past the UMI
8-12
18
~ 89 - 107
17
8-12
34
8
24
5'UMI
FP
Insert
RP
3'UMI
UPS
Index 1
IGS
Segment sequenced past the UMI
sample1_R2.fq
29
8
29
8-12
18
~ 89 - 107
17
8-12
IGS
Index 2
UPS
5'UMI
FP
Insert
RP
3'UMI
Trim segments sequenced past the UMIs
sample1_R1.fq
5'UMI
FP
Insert
RP
3'UMI
sample1_R2.fq
5'UMI
FP
Insert
RP
3'UMI
Merge
5'UMI
FP
Insert
RP
3'UMI
Trim UMIs, FP and RP and append their sequence information to the read name of the FASTQ files
Sample1.fq
@Read name 5'UMI FP RP 3'UMI
Insert sequence ...
+
Read quality
...
Align reads and annotate BAM files with UMI, FP and RP sequence information
BWA-MEM or Bowtie2
Sample1.bam
OX: UMI sequence
BZ: UMI Phred Q score
XP: Observed primer sequences
XQ: Observed primer sequences Phred Q score
XE: Primer edit distances
XL: Merged read length
Convert to BED file and calculate non-overlaping regions from alignment coordinates
Sort bam
Convert to bed
Merge overlapping regions
chr I start I end I strand
overlapping coordinates
non-overlapping coordinates
Cluster molecules with UMI edit distance <= 1
PCR1
PCR2
UMI clusters
UMI_A
UMI_B
...GCTA
TCAT...
...GCTA
TCAT...
...GCTA
TCGT...
UMI_C
Identical 5' UMI
3' UMI edit distance = 1

Figure 4.2 A graphical depiction of the UMI-aware and paired-end FAST-SeqS read processing pipeline. As an input, the pipeline takes demultiplexed FASTQ files for each sample. The top of the figure shows two grey boxes which represent read1 (R1) and read2 (R2) and the direction of sequencing (shown as a red arrow) with respect to the amplicon sequence, The first step in the process is to trim the segments of the reads which contain the UPS, index and IGS. Next, the trimmed reads are merged into a single consensus read. In preparation for aligning the reads, the UMI sequences, forward and reverse primer sequences are removed from the read sequence and appended to the read name, along with various statistics such as the primer edit distances. After this step, the processed reads should contain only the insert sequence. In the next step, the reads are aligned to the genome and custom scripts are run to annotate the resulting BAM files with the information contained in the read name from the FASTQ file. The SAM/BAM flag names that contain this information are shown in the figure. Next, the BAM file is coordinate sorted and the unique mapping locations are extracted into a BED file. The overlapping regions are then merged into distinct non-overlapping regions. Finally, a custom script filters reads and UMI sequences are clustered for each non-overlapping region (see main text for details). This results in UMI counts per region.

Before running the pipeline, the sequencing reads from a lane of sequencing need to be demultiplexed to separate FASTQ files for each sample. To do this, I used the demuxFQ tool provided by the CRUK Bioinformatics Core [122]. Note that to assess index switching in addition to demultiplexing the samples, all possible 96 × 96 (9216) combinations of indices need to be provided to the tool.

The pipeline takes demultiplexed FASTQ files as an input, along with the reference genome. The first step is to remove adapter sequences from the reads. Given the length of the amplicons, the ends of read1 and read2 will commonly include portions of the Universal Primer Sequence along with the indices and Illumina grafting sequences. I used bbduk.sh [246] to remove the adapters and provided it a list of adapters (along with the UDIs) for it find and remove from the reads.

The second step in the pipeline is to merge read1 and read2 into a single consensus read. I decided to merge the reads, using BBMerge [247], rather than align read1 and read2 separately to the genome because 1) given the length distribution of the amplicons, we expect almost all (if not all) read pairs to overlap, 2) by merging, we utilise the quality scores to take a consensus read, thereby filtering out many sequencing errors, 3) the task of aligning the reads to the genome should be easier, since we remove uncertainty about the insert size of the molecule we might expect increased numbers of true positive alignments, 4) it simplifies downstream analyses, for example when counting variant allele frequencies. A potential downfall of merging the reads is that sequencing errors from one read might be falsely incorporated into the merged read with the true base from the other read being discarded. We would normally expect sequencing errors to be rejected due to lower quality scores, though this may not always be the case.

The third step in the pipeline is to extract the primer and UMI sequences from the merged reads. The primers can hybridise to bases in the genome that are not an exact reverse complement to the primer. Since the sequence of the primer observed in the read may not reflect the exact sequence in the genome, we should remove the primer sequence before aligning the reads. I created a custom script in python which uses the Edlib library (which is a library to calculate edit distances written in C++) [248] to find the forward and reverse primers within the merged read. The forward primer is found by finding the sequence within the read with the minimal edit distance to the primer sequence. In the case of multiple matches with the same edit distance, the match which has the length closest to the length of the primer sequence is preferred. In the case that multiple sequences have the same edit distance and the same length, the left-most match within the read is preferred. The same process is applied to finding the reverse primer sequence, with the exception that the right-most match is preferred. If the edit distance to the forward primer and reverse primer are both 5 or below, the read is kept, otherwise the read is discarded. In the previous pipeline, we used the Hamming distance rather than the edit distance to find the primer sequence within the read. Using the edit distance is likely to decrease the number of reads discarded due to sequencing or PCR errors, which introduce indels into the sequence.

In the previous pipeline, the observed primer sequence and UMIs were discarded. I reasoned that the UMI sequences might be useful for low input DNA samples, but when the input DNA is 20 ng-50 ng, it was unlikely to help significantly reduce the number of reads down to unique molecules. This is because the number of copies of a genomic locus within the input DNA would be much greater than the average read depth at that locus. However, this was not explored further. In this pipeline, we extract the UMIs and the observed primer sequences, along with their associated phred quality scores and edit distance, and append these to the read name.

Step 4 in the pipeline is to align the reads (minus the UMI and primer sequences) to the genome. The pipeline uses BWA-MEM [238] and Bowtie 2 [114] and outputs the alignments in BAM format. The reason for using both aligners is to explore the differences between them and to decide which of them are more appropriate for aligning FAST-SeqS amplicons to the genome. We explore these results later in this chapter.

After alignment, the BAM file is annotated with the UMI and primer sequence information and stats that were included in the FASTQ header in step 3. I wrote a custom python script to do this which uses pysam [249] to access and write this information as tags to the BAM files. Specifically, the `OX`, `BZ`, `XP`, `XQ`, `XE` and `XL` tags were used to store the UMI sequences, UMI Phred quality scores, observed primer sequences, observed primer sequence Phred quality scores, primer edit distances and the merged read length respectively.

The resulting annotated BAM file is then sorted by its genomic coordinates using samtools [236]. The unique mapping locations (chromosome, start, end and strand) contained within the BAM file are then extracted for each sample using bedtools `bamtobed` [250]. Overlapping unique mapping locations are merged, using the bedtools `merge` function [250], into distinct non-overlapping regions. This is in preparation for the next step in the pipeline.

The final step in the pipeline involves summarising the alignment information for each locus in a sample. I wrote a custom script in python that iterates over the distinct non-overlapping regions and extracts alignments from these locations. The script utilises pysam to fetch the reads and the `multiprocessing` module [251] so that each region is handled in parallel by different subprocesses. For each distinct non-overlapping region, an alignment is discarded if at least one of the following conditions are true:

- it is a secondary alignment (meaning the read has a more likely alignment to another location in the genome)
- it is a supplementary alignment or the read has supplementary alignments (marked as supplementary or contains the `SA` tag) - this is specific for BWA-MEM which marks chimeric (split) reads as supplementary alignments
- it has a mapping quality less than the mapping quality threshold defined by the user.
- the read maps to the other strand (pysam does not fetch reads mapped to particular strands)
- the read does not contain a UMI tag or the length of either UMI is not valid; set by the user, for example 5-15.

For each non-overlapping region, alignments are then grouped into UMI nodes. A UMI node is defined as a group of alignments that have exactly the same 5' and 3' UMI sequence. Each UMI node is given a count which corresponds to the number of alignments associated to it. Note that even though alignments may have the same UMI sequence, due to PCR or sequencing errors, they may not be identical reads and may have different mapping qualities or slightly different alignment start and end positions.

For each non-overlapping region, reads that have the same UMI sequence are assumed to originate from the same molecule from PCR1. However, due to PCR and sequencing error, reads with similar but non-identical UMIs may have been derived from the same molecule. In order to account for PCR and sequencing error, we need to identify UMI nodes that are likely to represent the same molecule from PCR1. To do this, I use a similar approach to UMI-tools

[183] to cluster UMI nodes; determining a network of connected UMIs by creating a directed adjacency matrix for each non-overlapping region. I have two approaches for this, 1) the "combined" method which is based on clustering on the combination of 5' and 3' UMI pairs and, 2) the "separate" method which is based on clustering on the 5' and 3' UMIs separately.

In the combined method, for every possible pair of nodes A and B, the lengths of their 5' UMIs are compared and the lengths of their 3' UMIs are compared. If the difference in length between the 5' UMIs or the 3' UMIs is greater than 1, the nodes are not considered to be related. Otherwise, the edit distance between the 5' UMIs of node A and node B and the edit distance between the 3' UMIs of node A and node B are calculated. I used Edlib and the `infix` method [248] (in which gaps at the query end and start are not penalised) to calculate the edit distance. If either of the edit distances are greater than 1, the nodes are not considered related. Otherwise, if the count of node A is greater than or equal to the count of node B, then a directed edge from node A to node B is created. If the count of node B is greater or equal to the count of node A, then a directed edge from node B to node A is created. Given the directional adjacency matrix, I then use the `group_directional` method from UMI-tools to derive the UMI clusters. The "separate" method is intended for use with single cell FAST-SeqS data and is explained in chapter 5.

The script outputs (per sample):

- *Alignment counts with "combined" UMI cluster information*: including various statistics such as the UMI cluster ID, mapping quality, the forward primer observed (along with edit distance to the forward primer sequence), the reverse primer observed (along with the edit distance to the reverse primer sequence), the fragment length (whole sequence - including UMI and primers) and nucleotide composition of the fragment (number of As, Cs, Gs, Ts and Ns), the “trimmed read” length (the sequence of the alignment fragment - minus the UMIs and primers) and the nucleotide composition of the fragment (number of As, Cs, Gs, Ts and Ns).

- *Alignment counts with "separate" UMI cluster information*: including the same statistics as above but with a UMI cluster ID for the 5' and 3' UMI.

- *Base level count information*: including the chromosome, position, alignment start and end of the read, reference, base observed, Phred quality score, mapping quality and count. Note that I do not report the UMI information here (see future work section in chapter 6).

- *Filtering statistics*: The number of alignments excluded due to secondary alignments, supplementary alignments, insufficient mapping quality, mapping to the wrong strand or invalid UMI sequence lengths.

Note that this is highly detailed information which we may want to reduce in the future. The reason for outputting this level of detail was for the purposes of data exploration, to understand which aspects of the data are useful for filtering and modelling, which I explore later in this chapter.

4.3 Initial application to prostate samples

We used the new primers to prepare 96 DNA samples for sequencing using the FAST-SeqS protocol (see section 4.8.1 for methods). These DNA samples were predominately derived from prostate cancer samples and normal adjacent prostate tissue. Some of the samples were previously sequenced as part of the ICGC prostate cancer project. These samples were sequenced because they were determined to have high tumour purity by a pathologist prior to sequencing. In some cases, the estimated tumour purity using the sequencing data was much lower than the pathologist's estimate. Additionally, we included samples that were believed to have low purity according to the pathologist's estimate and hence were not sequenced. Other samples were included from benign prostate tumours which are assumed to have no SCNAs and could be used as controls. In addition, since the comparison between normal samples from male and female donors is useful for data exploration and diagnostic purposes, we prepared six replicate FAST-SeqS samples from a DNA sample that was derived from whole blood from multiple anonymous female donors (Promega G1521).

The purpose of the experiment was to:

- Prove that we could successfully scale-up the FAST-SeqS experiments to higher-throughput machines
- Understand the extent and the potential causes of index swapping in FAST-SeqS experiments using the unique dual indices (see section 4.3.1)
- Apply conliga to a new data set to see if the assumptions that were made using previous data still hold (see section 4.3.2). If the assumptions do not hold, to investigate the causes (see section 4.4)
- Investigate the application of FAST-SeqS and conliga in the research setting. In particular, using the assay and tool for the screening of samples in large-scale cancer genome studies prior to performing WGS (see chapter 5).

In the process of investigating these objectives, I found that sample-specific effects occurring within the new data set caused RCN inferences that did not represent true SCNAs (shown in section 4.3.2). From section 4.4 onwards, this chapter focuses on the identification of these sample-specific effects and methods to correct for them. The application of FAST-SeqS and conliga to large-scale genome studies is explored in chapter 5.

4.3.1 Sample demultiplexing and index cross-talk analysis

The HiSeq 4000 lane on which the samples were sequenced produced a total of 391,866,868 paired-end 150 bp reads. Initial quality control using FastQC [252] showed that the reads were typically of high quality across the tiles of the flow cell, with good sequence diversity and low levels of primer-dimer as indicated by the low proportion of adapter sequence found within the reads. I used the demuxFQ tool [122] (as previously described in section 2.4) to demultiplex the paired reads from the lane into separate FASTQ files for each sample such that each sample has a separate read1 and read2 FASTQ file. I supplied the demuxFQ tool with the 96 × 96 possible index1 and index2 combinations. As such, the reads were demuliplexed into 9,216 read1 FASTQ files and 9,216 read2 FASTQ files for each possible index1 and index2 combination.

Of the total read pairs from the sequencing lane, 95,870,090 (24.46%) were discarded as they did not match a combination (with a maximum of one mismatch to each index). The vast majority of the discarded reads will have originated from PhiX and the quantity of discarded reads coincides with our target of a library composed of 20-30% PhiX. Of the remaining 307,688,258 read pairs, 295,996,778 (96.2%) matched valid index1 and index2 combinations which represent a specific sample. The remaining 11,691,480 (3.8%) read pairs were invalid combinations, indicating that we see substantial index misassignment (index cross-talk) frequencies [241, 239, 253, 125]. The proportion of reads that had invalid index combinations are high and close to the 5-10% which Sinha *et al.* reported [239] and much higher than 0.29% reported by Macconaill *et al.* [241]. If we assume that cross-talk or swapping rates occur equally between indices from index1 as they do from index2, then a single index strategy would have resulted in approximately 1.9% of reads being misassigned to the wrong sample. A combinatorial dual-index strategy would have resulted in approximately 3.8% of reads being misassigned to the wrong sample. Using UDIs, we are able to remove the vast majority of these events and false assignment will only occur in the rare events that both indices swap to an index1 and index2 combination that represents another sample.

There are a several mechanisms that can introduce index cross-talk [241]. These mechanisms include: cross-contamination during primer oligonucleotide synthesis or purification, during the process of diluting or aliquoting the primers and other experimental handling issues, the "spreading-of-signal" phenomenon on Illumina's patterned flow cells (and to a lesser extent the non-patterned flow cells) arising from free index primers within a pooled library [239], sequence errors introduced during PCR (either during an experiment or during bridge amplification) or sequencing error, bioinformatic errors, among other mechanisms

[241]. Since all index1 indices and all index2 indices have at least an edit distance of 3 from each other, we can assume that sequence errors are very unlikely to be the cause of index cross-talk in this experiment.

Figure 4.3A shows an overview of the total number of read pairs assigned to each of the possible 9,216 combinations. Tiles on the diagonal represent valid combinations and those on the off-diagonal represent invalid combinations. We see a strong signal along the diagonal, indicating that the reads appear to be assigned to valid samples in most cases. There appears to be a varying frequency of index cross-talk between combinations of indices. Valid index combinations that have low numbers of read pairs also have fewer invalid combinations with other indices. Indeed, this occurs for both indices. This might be a reflection that fewer swapping events occur with other free indices because fewer amplicons are present for these samples. When we normalise by the total counts for index1 (figure 4.3B) and index2 (figure 4.3C), we see proportionally more swapping events (in some cases) for valid index combinations with low counts. We may hypothesise that the low counts are a result of fewer amplicons being synthesised for these samples during PCR. If this is true, it is likely there is a greater proportion of free primer corresponding to these samples (relative to the samples with higher counts) when the samples are pooled together for sequencing. As it has previously been shown, the amount of free primer within the library is directly proportional to the number of swapping events that occur on the flow cell [239]. Figure 4.3A, B and C also highlight potential errors during the experiment. In particular, the results might suggest potential mix-ups between the primers of sample 61 and sample 62 and possibly between sample 17 and 18 also. Sample 61 was assigned 610 total reads while sample 62 was assigned substantially more than expected (9,832,386).

The samples were prepared in a 96-well plate and I explored how the frequency of index cross-talk relates to the well position (figure 4.4). When index cross-talk occurs between indices from the same column, we see increased cross-talk occurring between indices in adjacent rows. Indeed, the mean frequency of index swapping within a column is proportional to the number of rows that separate the indices (figure 4.4A and C). Interestingly, we do not see such a strong behaviour between columns in the same row (figure 4.4A and D). When comparing adjacent columns (figure 4.4A), we see that cross-talk frequencies are not symmetric. Cross-talk occurs in a particular pattern that is concordant with the sample order and the order of pipetting. The highest cross-talk frequencies between columns is seen between the bottom well in a given column and the top well of the column to the right. While increased cross-talk within rows and columns were reported by Macconaill *et al.* [241], a relationship with the pipetting order was not reported. These results suggest that the process of pipetting may introduce substantial index contamination. Alternatively (or additionally) this may be explained by the order the primers were synthesised, purified, diluted or aliquoted; it is possible that they were done so in order of their sample index.

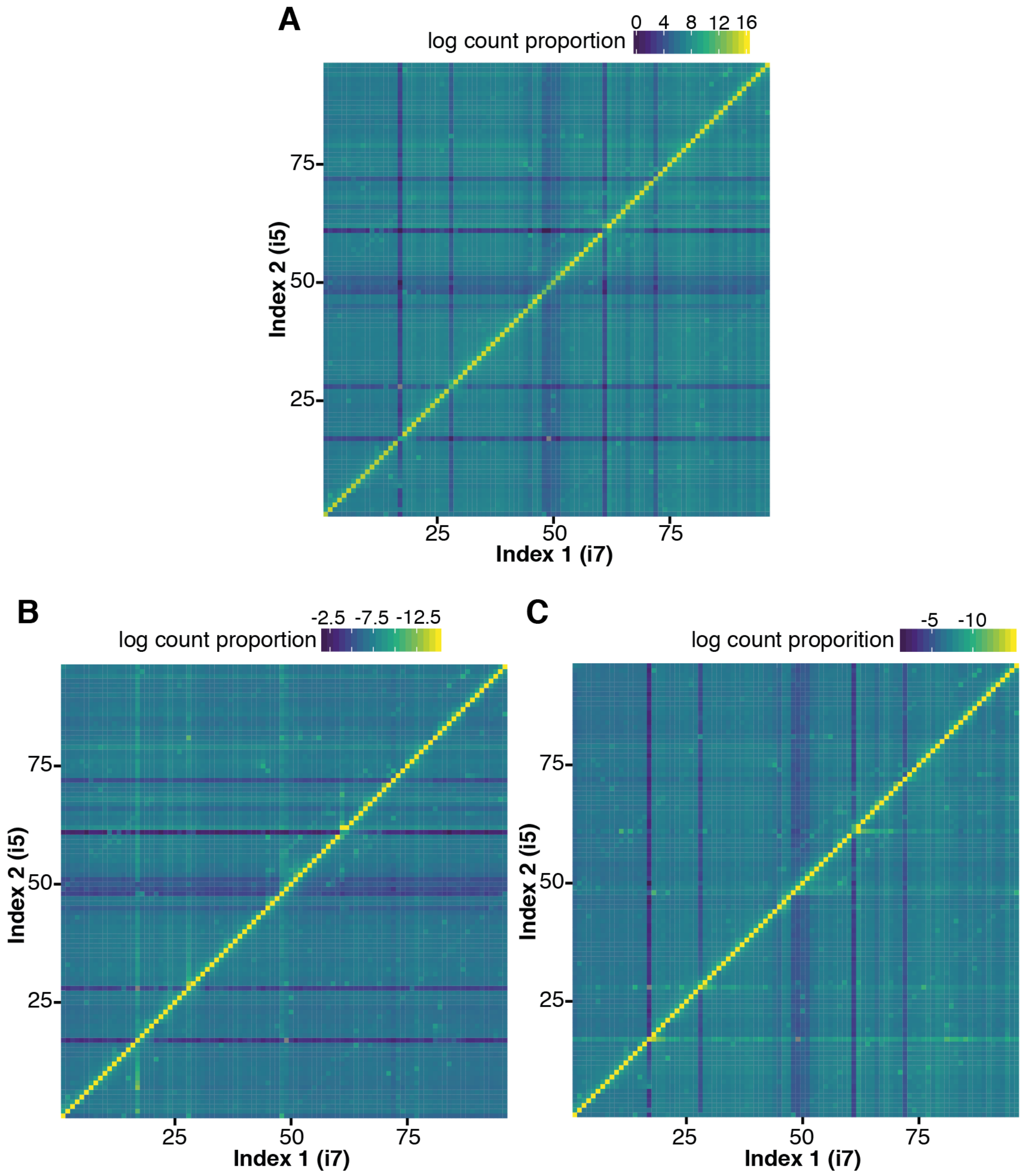


Figure 4.3 Exploring the number of paired-end reads assigned to valid and invalid combinations of sample indices (index1 and index2). A: Heatmap representing the log count of paired-end reads assigned to all 96 by 96 (9,216) possible combinations of index 1 and index2. The diagonal represents valid sample combinations while the off-diagonal represents invalid sample combinations. B: Same as A but the heatmap displays the log counts normalised by the total counts of index 1. C: Same as B but normalised by the total counts of index 2.

Further experiments in which the sample indices are randomised on the PCR plate will provide additional evidence as to the source of this issue.

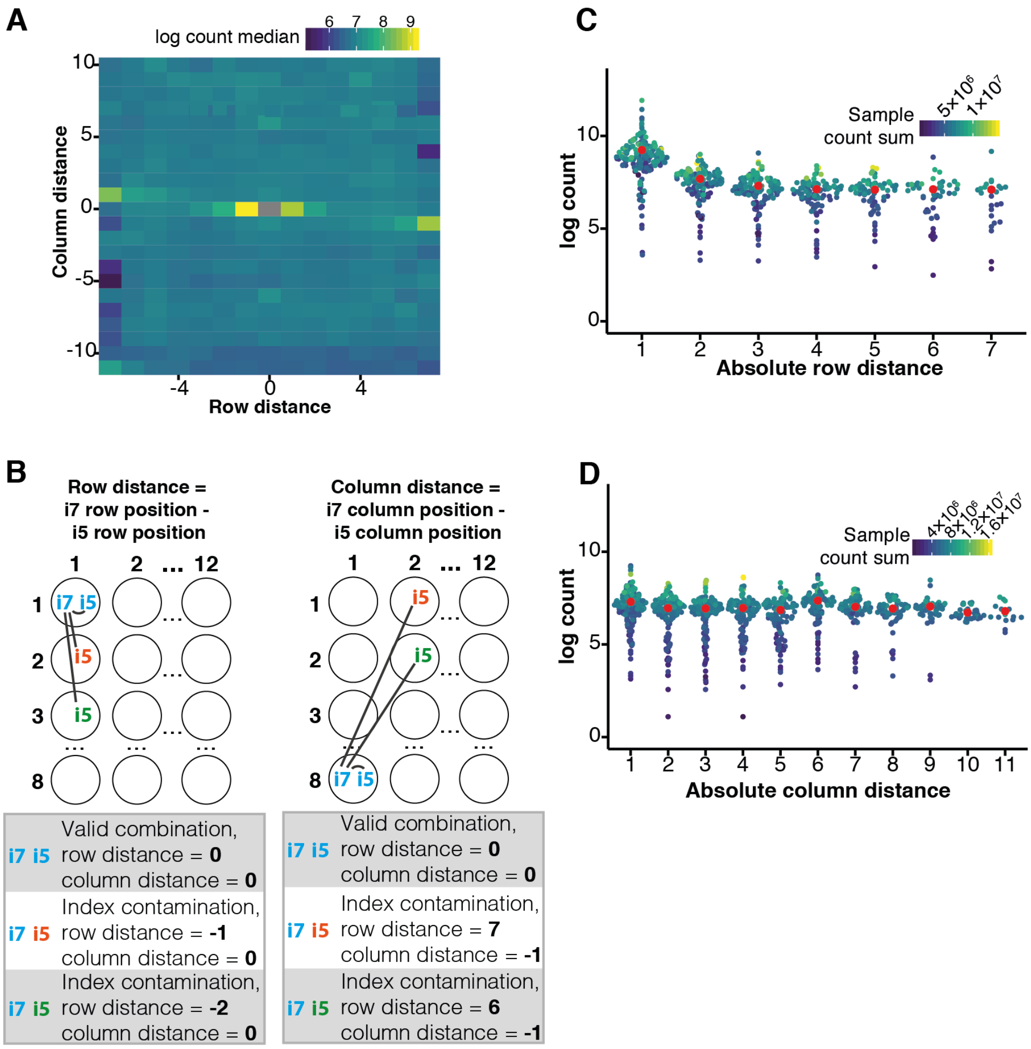


Figure 4.4 Investigating index-cross talk in the context of the 96-well plate. A: Heatmap showing the relationship between the row and column distance with the median log count (since multiple index combinations with have the same row and column distance). B: Diagram indicating how the row distance for an index combination is calculated; index1 (i7) row position - index2 (i5) row position. It also includes how the column distance is calculated; index1 (i7) row position - index2 (i5) row position. For example, i7(blue) & i5(blue) is a valid index combination with row and columns distances of 0. C: Sina plot showing the relationship between the absolute row distance (of invalid index combinations from the same column) and the log count of the index combinations. Each point represents an invalid index combination and the colour represents the sum of the corresponding valid samples. For example, if the invalid index combination is i7(blue) & i5(green), then the colour in the legend represents the sum of valid combinations i7(blue) & i5(blue) and i7(green) & i5(green). The red point represents the mean log count. D: Same as C but showing the relationship between the absolute column distance (of invalid index combinations from the same row) and the log count of the index combinations.

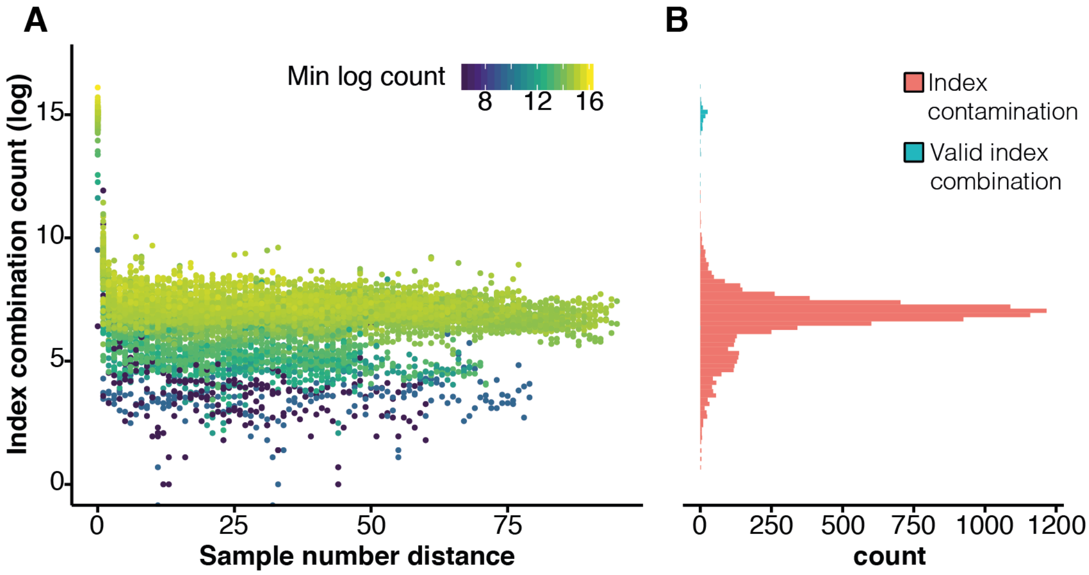


Figure 4.5 The number of paired-end reads assigned to index combinations compared with their corresponding sample number distance. A: Absolute sample number distance against index combination count (log). Each point represents an index combination and the colour represents the minimum log count of the corresponding valid samples. B: The distribution of index combination counts (log) showing invalid combinations in red and valid combinations in blue.

Given the high rates and spatial bias of index cross-talk observed, the use of a single or combinatorial dual index strategy would result in a relatively high frequency of read misassignment and a bleeding of signal between samples. Since normal and tumour samples are multiplexed on the same lane, this would have several implications. If reads from tumour samples are falsely assigned to samples used as controls, the inference of the bias ($\hat{\boldsymbol{m}}$) would be less accurate. On its own, this might lead to increased false positive and false negatives when inferring RCN profiles from test samples. Also, if reads from test samples that have SCNAs bleed into those that do not, this may mean that we falsely detect SCNAs in samples that are otherwise normal. This would be problematic for research applications but critical for clinical application. Therefore, it appears that the use of UDIs is important in order to maximise the sensitivity and specificity of FAST-SeqS and conliga.

While we can be confident that by using UDIs we are removing almost all of the cross-talk events, by reducing cross-talk events we will be able to reduce the number of reads discarded. Improved clean-up of the free primer from the samples may help to reduce the background rate of cross-talk between indices [240, 239]. Further experiments are needed to discover the cause of the spatial bias in index cross-talk frequency. If it is found that the source of the bias is not due to the synthesis or purification of the primers by the manufacturer, it may be that experimental improvements and increased handling care may reduce this effect. Sequencing using non-patterned flow cells such as those for the HiSeq 2500, NextSeq or Miseq may reduce the background level of cross-talk by approximately 3-fold [241, 240]. However, the patterned flow cell technology adopted by the HiSeq X Ten, HiSeq 3000/4000 and NovaSeq 600 decreases the sequencing costs per read and helps to make the use of FAST-SeqS and other sequencing assays more cost-effective. As such, discarding reads associated with increased cross-talk events is likely to be more cost-effective than utilising non-patterned flow cell technologies that have lower rates of cross-talk.

4.3.2 Running conliga on the samples and obtaining unusual results

After running the UMI-aware pipeline (described in section 4.2) on the demultiplexed FASTQ files, I ran conliga to infer the RCN profiles for the samples. I used the set of benign samples together with the matched normal samples (total of 16 samples) to infer the loci bias vector $\hat{\boldsymbol{m}}$ and used the same priors previously described in section 2.10. I used the summarised counts that were derived from aligning the processed FASTQ files to the genome using Bowtie 2 [114] (using the default settings). I used an arbitrary threshold to remove alignments that had a mapping quality of less than 10.

In some cases, the inferred RCN profiles seemed reasonable with what appeared to be true SCNAs and flat RCN profiles in control samples. However, in other samples, the inferred profiles were unrealistic, often with fast-switching dynamics across the whole genome, which were unlikely to represent true SCNAs. In addition, these types of profiles were seen in several of the control samples. This provided further evidence that the RCN profiles did not necessarily represent true SCNAs. Figure 4.6 shows example RCN profiles for a selection of normal samples.

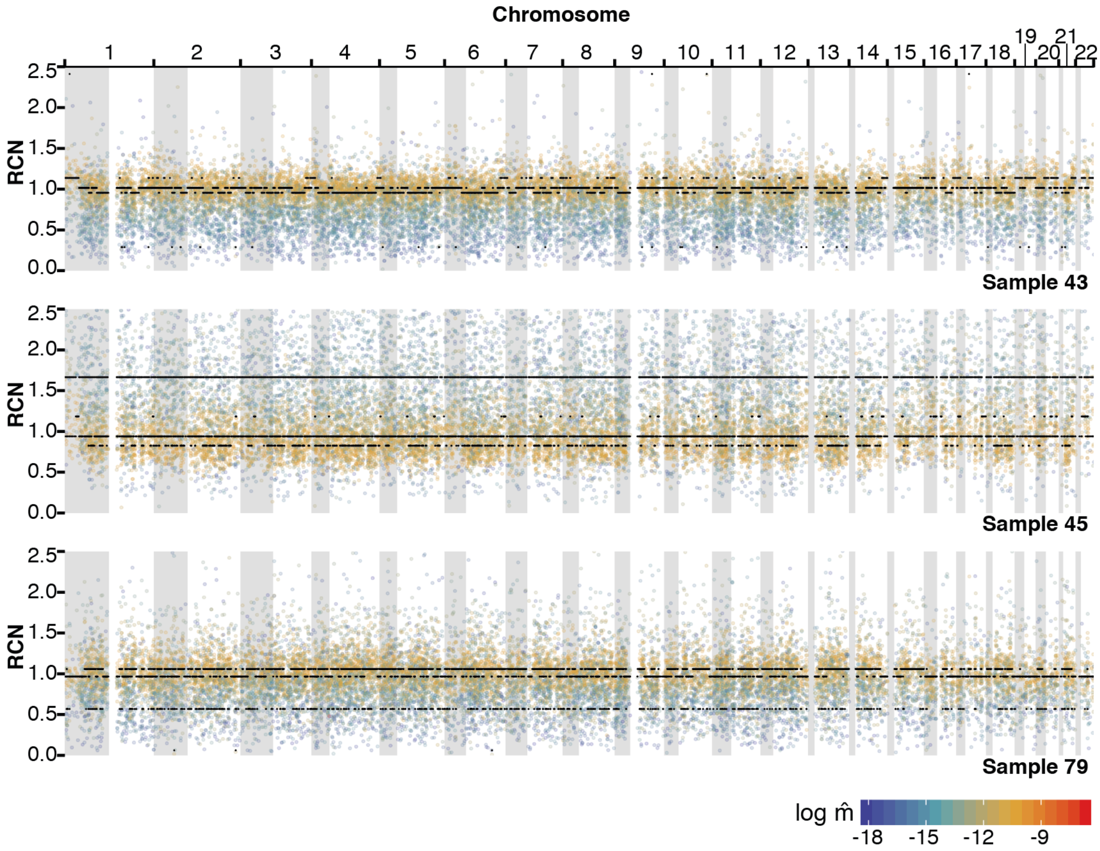


Figure 4.6 RCN profiles for three normal prostate samples showing unrealistic inferred profiles that are unlikely to represent true SCNAs. These samples were processed with the new UMI-aware pipeline (section 4.2) and reads were aligned with Bowtie 2 (mapping quality threshold >10). Each black point represents the MAP RCN inferred for each locus. The coloured points represent the scaled count for each locus coloured by the log of the inferred expected proportion of reads to align to each locus ($\log \hat{m}$).

To check that these artifacts were not introduced by the new UMI-aware pipeline, I processed the read1 FASTQ files through the original pipeline (described in section 2.4). Using the resultant counts and the same prior distributions, I re-ran conliga and compared the results. Similar, although not identical, RCN profiles were obtained by using counts derived from the old and new pipelines (figure 4.7). While the read processing may play a role, this suggests that a major source of RCN artifacts appear to originate in the DNA samples or sample preparation rather than sample-specific biases introduced by the read processing.

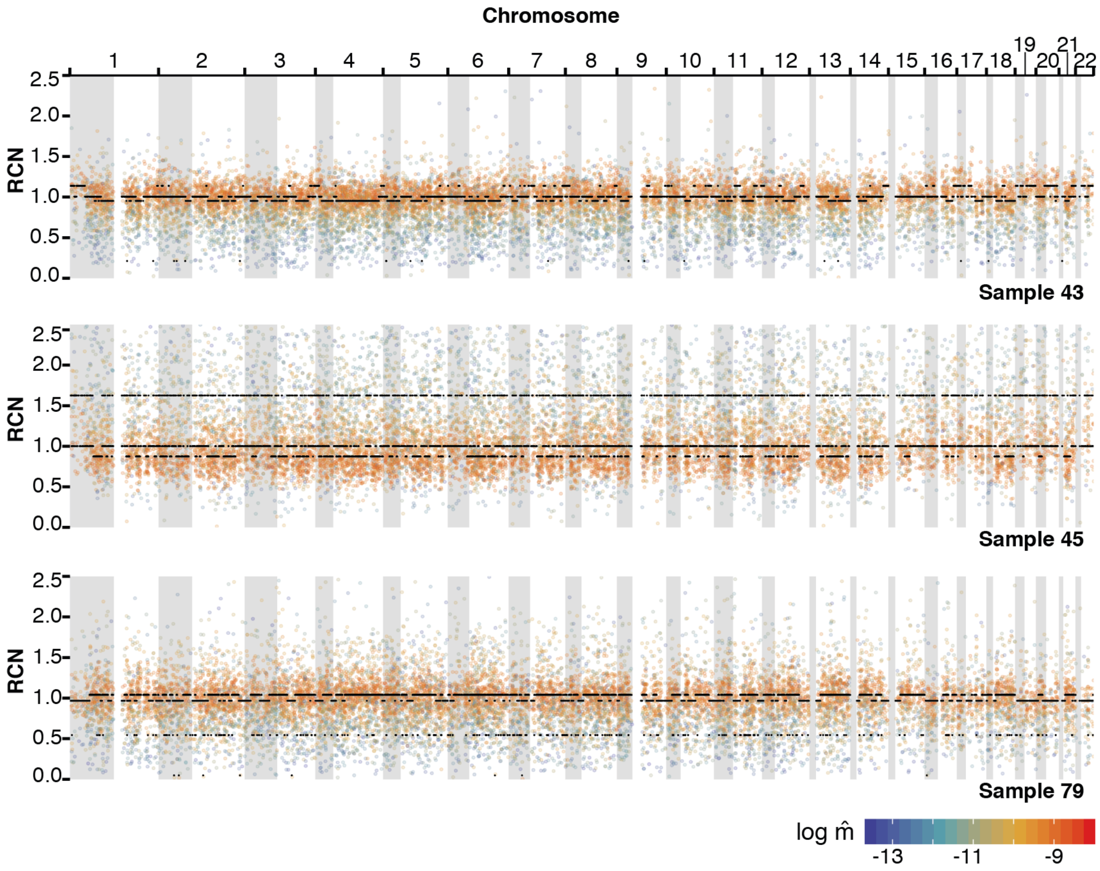


Figure 4.7 RCN profiles for three normal prostate samples showing unrealistic inferred profiles that are unlikely to represent true SCNAs. These samples were processed with the original pipeline described in section 2.4 and reads were aligned with Bowtie 1. Each black point represents the MAP RCN inferred for each locus. The coloured points represent the scaled count for each locus coloured by the log of the inferred expected proportion of reads to align to each locus ($\log \hat{m}$).

4.4 Investigating the cause of sample-specific artifacts

In this section, I explore the data in order to identify the causes of RCN profile artifacts. In section 4.4.1, I explore the data from the new UMI-aware pipeline and compare it to the data obtained from the single-end pipeline. I do this in order to discover the cause of the differences seen in profiles obtained by the old single-end read processing pipeline and the new UMI-aware pipeline. In section 4.4.2, I explore whether sample-specific biases are present in the data.

4.4.1 Investigating the data obtained from the new UMI-aware pipeline

Comparing BWA-MEM and Bowtie 2 for aligning FAST-SeqS reads

Aligning reads from repeat regions, such as LINE1 elements, to the reference genome is difficult. Since their retrotransposition into the genome, the LINE1 elements that are targeted by the FAST-1 primer pair have accumulated mutations. FAST-SeqS functions as a copy number assay on the premise that a substantial proportion of these elements have accumulated a sufficient number of mutations so that they can be uniquely identified. As such, we are able to determine the genomic origin of some, but not all, of the reads generated by the FAST-SeqS assay.

In the pipeline presented in chapter 2, I used Bowtie 1 [111] to align the reads to the genome. Bowtie 1 is a simple aligner which is unaware of indels and does not perform gapped alignment. Given a threshold on the number of mismatches allowed, it aligns reads to the genome. If, given this threshold, the read aligns to one location then the read is considered uniquely mapped (uni-read). Otherwise, the read is considered a multi-read. Hence, the greater the number of mismatches permitted, the greater the number of expected multi-reads. Considering we discard multi-reads, this means that the mismatch threshold has two effects. As the threshold increases, we are more tolerant to PCR and sequencing errors while simultaneously more stringent with our definition of what constitutes a unique genomic region.

More modern aligners such as Bowtie 2 [114] and BWA-MEM [238] are indel-aware and perform gapped alignment. With the inclusion of indels, the concept of a uniquely mappable region becomes less clear. These tools provide a mapping quality (MAPQ) which quantifies the probability that the read is incorrectly aligned [254]. The way in which this probability is calculated depends on the software used (e.g. Bowtie 2, BWA-MEM) and on the values of the parameters supplied. While it may be tempting to consider all possible alignments and use these probabilities in downstream analyses, it is common practice for reads with a mapping quality below a determined threshold to be discarded.

The use of Bowtie 1 seemed to be sufficient to align reads from the FAST-SeqS assay. We know this because we saw copy number signal in the data and because few reads aligned to chromosome Y in samples from female donors (see figure 2.5 in chapter 2). If reads are aligned to the wrong location, noise will be introduced to the data. Moreover, the number of falsely aligned reads will not scale with the copy number changes as would be expected of true alignments, leading to a dilution of the copy number signal.

It was unclear how Bowtie 2 and BWA-MEM would perform in aligning FAST-SeqS reads to the genome. Furthermore, I wished to determine an appropriate mapping quality threshold that would minimise the number of false alignments and maximise true alignments.

Although we do not know the true origin of each read, I explored the relationship between mapping quality and other features of the data which might indicate alignment performance. For example, the proportion of alignments to chromosome Y in female samples can provide a measure of false alignments. Figure 4.8A indicates that Bowtie 2 introduces a high proportion of false alignments below a mapping quality threshold of 28 (using the sensitive and very sensitive settings). Whereas, BWA-MEM sees approximately half the proportion of false alignments when no threshold is set. Furthermore, the proportion of false alignments to chromosome Y rapidly approaches zero when setting a threshold on the mapping quality.

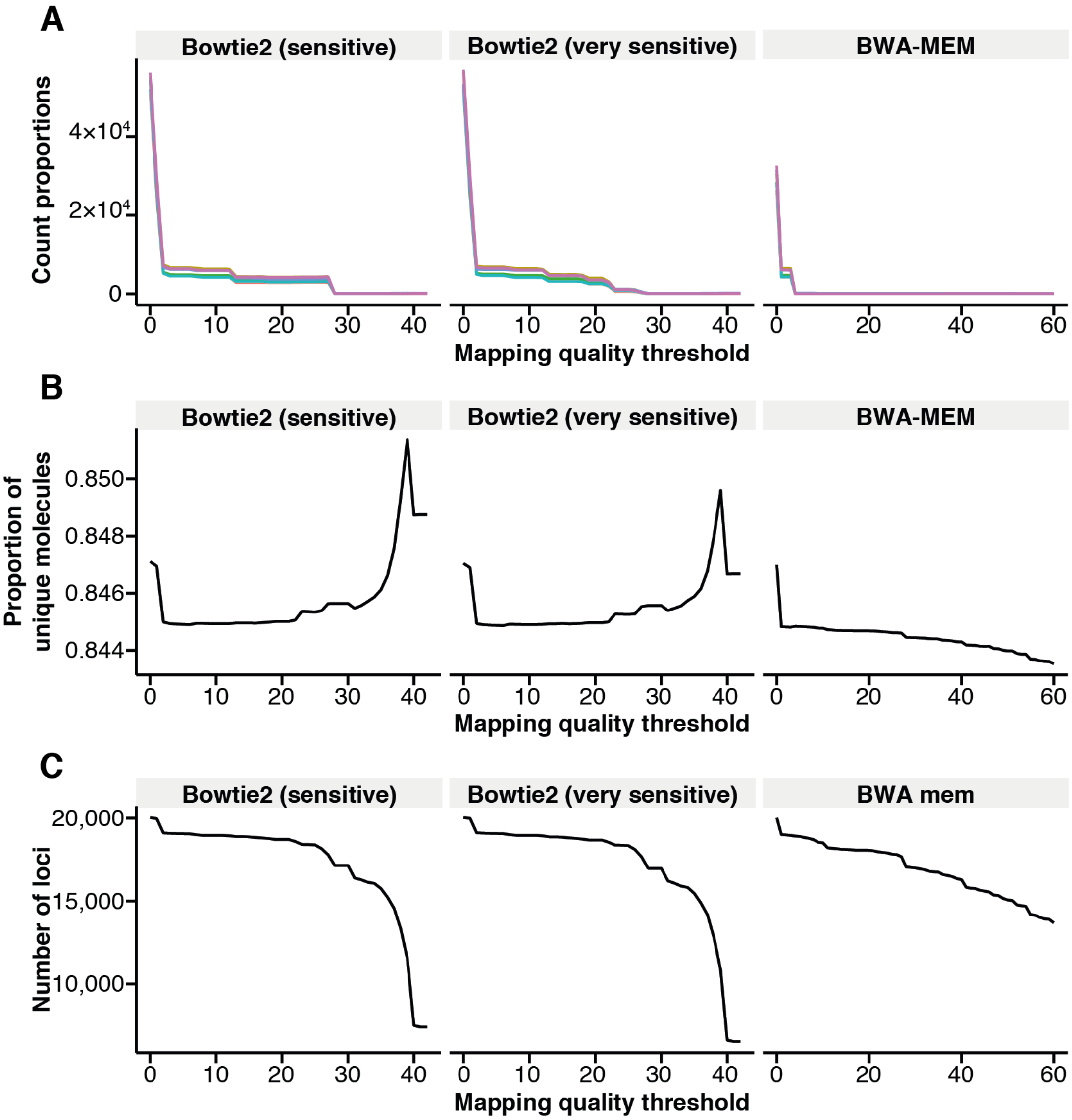


Figure 4.8 Comparing the effect of applying mapping quality thresholds for FAST-SeqS reads aligned by Bowtie 2 and BWA-MEM. The relationship between the mapping quality threshold and: A, the proportion of alignments mapping to chromosome Y in six replicates of a female sample, B, the proportion of unique molecules at each FAST-SeqS locus across all control samples (total number of unique molecules divided by the total number of reads), C, the number of loci with non-zero count in all 16 control samples.

The number of unique molecules associated with a locus as a proportion of the total read count is another indicator of performance. Because we perform a second round PCR, we occasionally expect to see multiple reads supporting the same unique molecule at each locus (duplicates). If all alignments were false positives, we would expect that at any given locus, each read would represent a unique molecule, i.e. the proportion would be 1. As false alignments decrease and true alignments increase, we expect to see proportionally fewer unique molecules at each locus. Figure 4.8B shows that, for reads mapped using Bowtie 2, the proportion of unique molecules initially decreases with the mapping quality threshold and then increases. Using the sensitive or very sensitive settings does not appear to change the outcome substantially. Importantly, at a threshold of 28, the proportion of unique molecules is on the rise even though, at this threshold, many of the false alignments to chromosome Y have been filtered out. For BWA-MEM, we see that as the mapping quality threshold increases, the proportion of unique molecules decreases. These results, plus those shown in figure 4.8C, suggest that there does not appear to be a mapping quality threshold which balances the removal of false alignments while retaining true alignments for Bowtie 2. For BWA-MEM, it seems that the ratio of true alignments to false alignments improves as the mapping quality threshold increases.

If we are to filter reads based on mapping quality scores, we would hope that the mapping qualities for the same locus are similar between samples. Figure 4.9 compares the mapping quality scores obtained across control samples with those obtained from a particular control sample. The results indicate that mapping quality scores are more variable between controls when using Bowtie 2 compared with BWA-MEM.

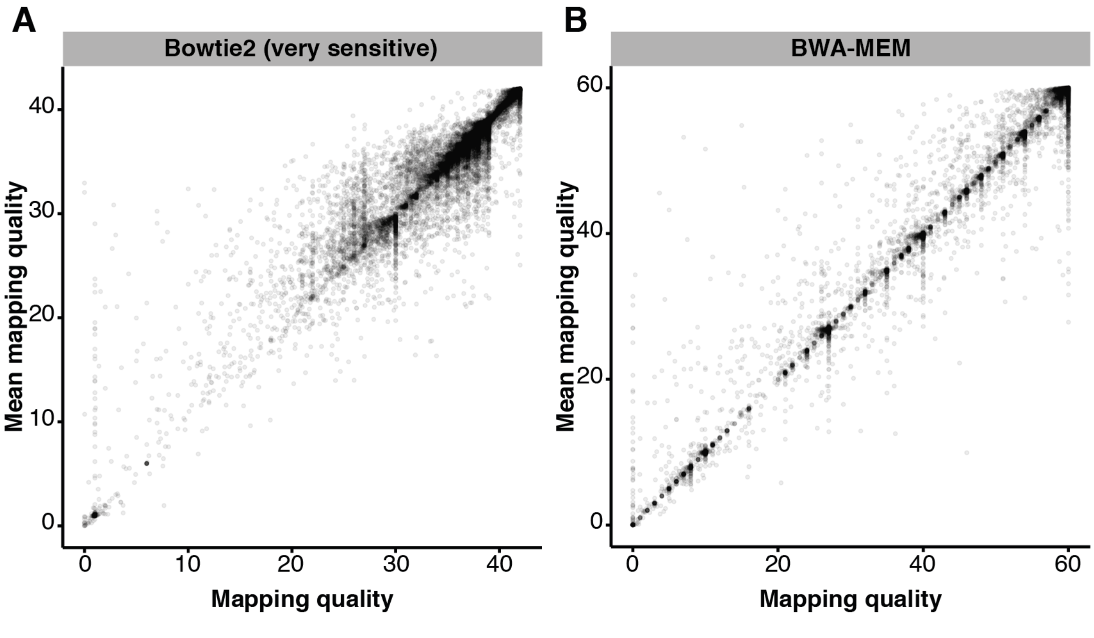


Figure 4.9 Comparing mapping quality scores between control samples using Bowtie 2 and BWA-MEM. Mean mapping quality score for each locus in control sample 73 against the mean of the mean mapping quality scores across all 16 control samples. The loci were filtered such that only those which had a non-zero count in all 16 controls were considered. A: Bowtie 2 (very sensitive settings). B: BWA-MEM.

It has been previously reported that BWA-MEM is more tolerant to mismatches and indels when compared to BWA-ALN, Bowtie 2, SOAP2, Subread and STAR [255]. We expect there to be a mixture of PCR errors and sequencing errors within FAST-SeqS reads in addition to germline and somatic variants. Given the similarity between the LINE1 sequences, there is increased difficulty in aligning FAST-SeqS reads with mismatches and indels to the reference genome compared with the majority of reads generated from standard WGS. If BWA-MEM is more tolerant than Bowtie 2 in dealing with mismatches and indels, this may partially explain why BWA-MEM sees less variability between controls in mapping qualities at the same locus.

I investigated how the two aligners performed in dealing with mismatches and indels in FAST-SeqS reads. The majority of these mismatches and indels will be caused by PCR and sequencing error. However, a small proportion will be germline and somatic variants. Figure 4.10 shows how the frequency of mismatches and indels relate to the read depth, mean mapping quality and mean base quality (Phred score). Many of the mismatches and indels that cluster around a frequency of 0.5 will be heterozygous germline or somatic variants and many of them are likely to be single nucleotide polymorphisms (SNPs). We can be increasingly confident that these are true germline and somatic variants as the read depth increases. Those mismatches and indels with frequencies close to zero are likely to be caused by PCR and sequencing error in addition to mismapped reads but could also represent somatic mutations. Those clustered around a variant frequency of 1 could be the result of homozygous non-reference germline or somatic variants in addition to mismapped reads. I should note that the reference human genome is not complete and may contain errors, particularly in repeat regions. Therefore, observed variant frequencies may be in part explained by an imperfect reference genome.

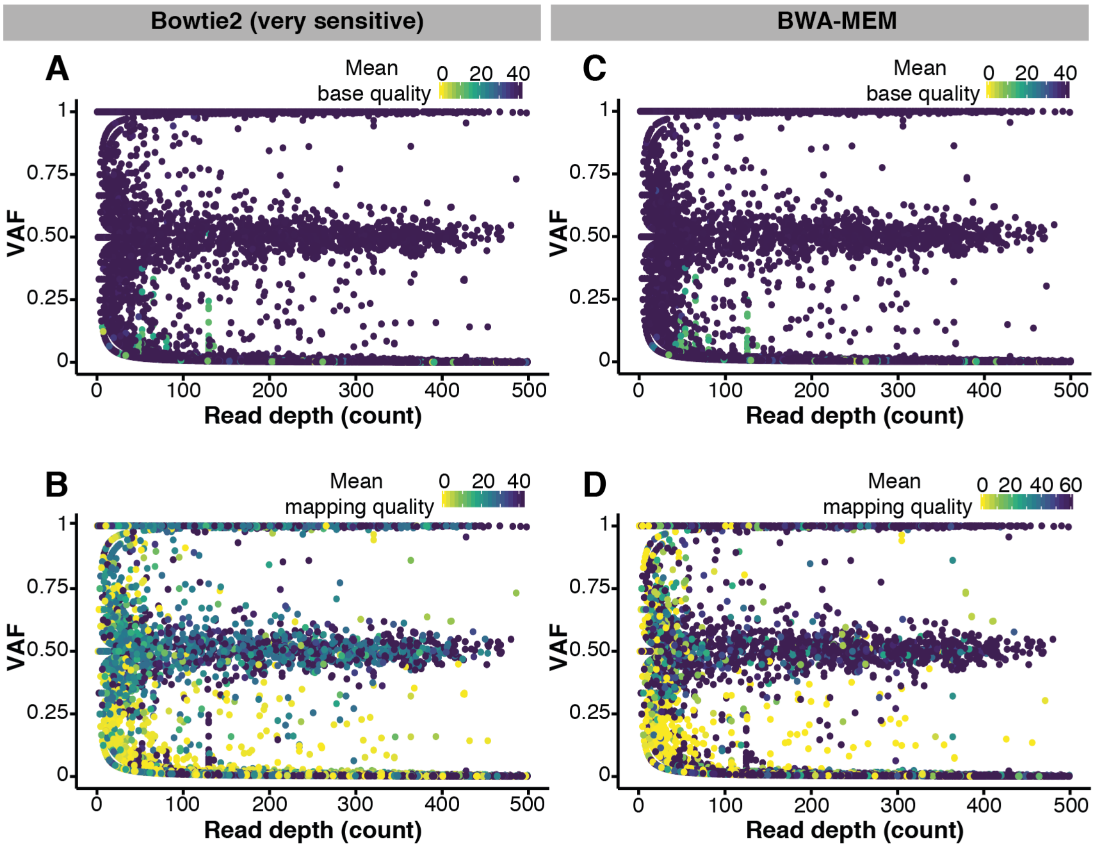


Figure 4.10 Mismatch and indel frequencies (i.e. variant allele frequencies) in aligned reads from normal sample 73 using Bowtie 2 (very sensitive settings) and BWA-MEM (default settings). A: Bowtie 2, read depth against mismatch and indel frequency. Each point represents a base which did not match the reference sequence in the set of all alignments. Each non-reference base is coloured according to the mean base quality. B: Bowtie 2, same as A but coloured according to the mean mapping quality. C: Same as A but for reads mapped by BWA-MEM. D: Same as B but for reads mapped by BWA-MEM.

In figure 4.10, we see that Bowtie 2 and BWA-MEM produce alignments with similar frequencies of mismatches and indels. By highlighting the mean base qualities, figures 4.10A and C show that relatively few of these mismatches and indels appear to be the result of sequencing error. Figures 4.10B and D suggest that a greater proportion of the mismatches and indels appear to be introduced by alignment error. Importantly, for reads aligned with what we can assume are heterozygous germline or somatic variants, Bowtie 2 frequently estimates a lower than perfect mapping quality (<42). In contrast, BWA-MEM gives a higher proportion of these reads a perfect mapping quality (60). This provides further evidence that BWA-MEM is more tolerant to variants than Bowtie 2, even in the context of repeat regions like those amplified by the FAST-SeqS assay.

Taken together, these results suggest that BWA-MEM should be used in favour of Bowtie 2 for the alignment of FAST-SeqS reads. There does not appear to be a good choice of mapping quality threshold for Bowtie 2; a threshold below 28 is likely to introduce many false alignments while a threshold above 28 is likely to remove true alignments. In contrast, it appears that BWA-MEM is not as sensitive to the choice of mapping quality threshold. Indeed, it seems that the higher the mapping quality threshold used, the ratio of true to false alignments increases. The analysis presented here is far from comprehensive. Further work should be done to evaluate a wider range of aligners and to determine how the parameters of the various aligners can be optimised to improve results. Simulating reads from FAST-SeqS loci could be used to evaluate the sensitivity and specificity of the aligners, since the ground truth is known. By varying the frequency of simulated mismatches and indels, we can better understand their performance under different conditions. Recall that Bowtie 2 was used by the authors of WALDO [107], although they did not state if mapping quality thresholds were used in their analysis. BWA was used by the authors of mFAST-SeqS [106], although they do not state whether the BWA-MEM, BWA-SW or BWA-backtrack algorithms were used in their analysis.

Investigating loci features indicative of low-quality data

In chapter 2 and figure 2.4E, we showed that there was a mean-variance relationship in FAST-SeqS count data. By visualising this relationship (figure 4.11), we can explore various features of the data that could be used to filter loci, which if not filtered, could mislead the copy number inference.

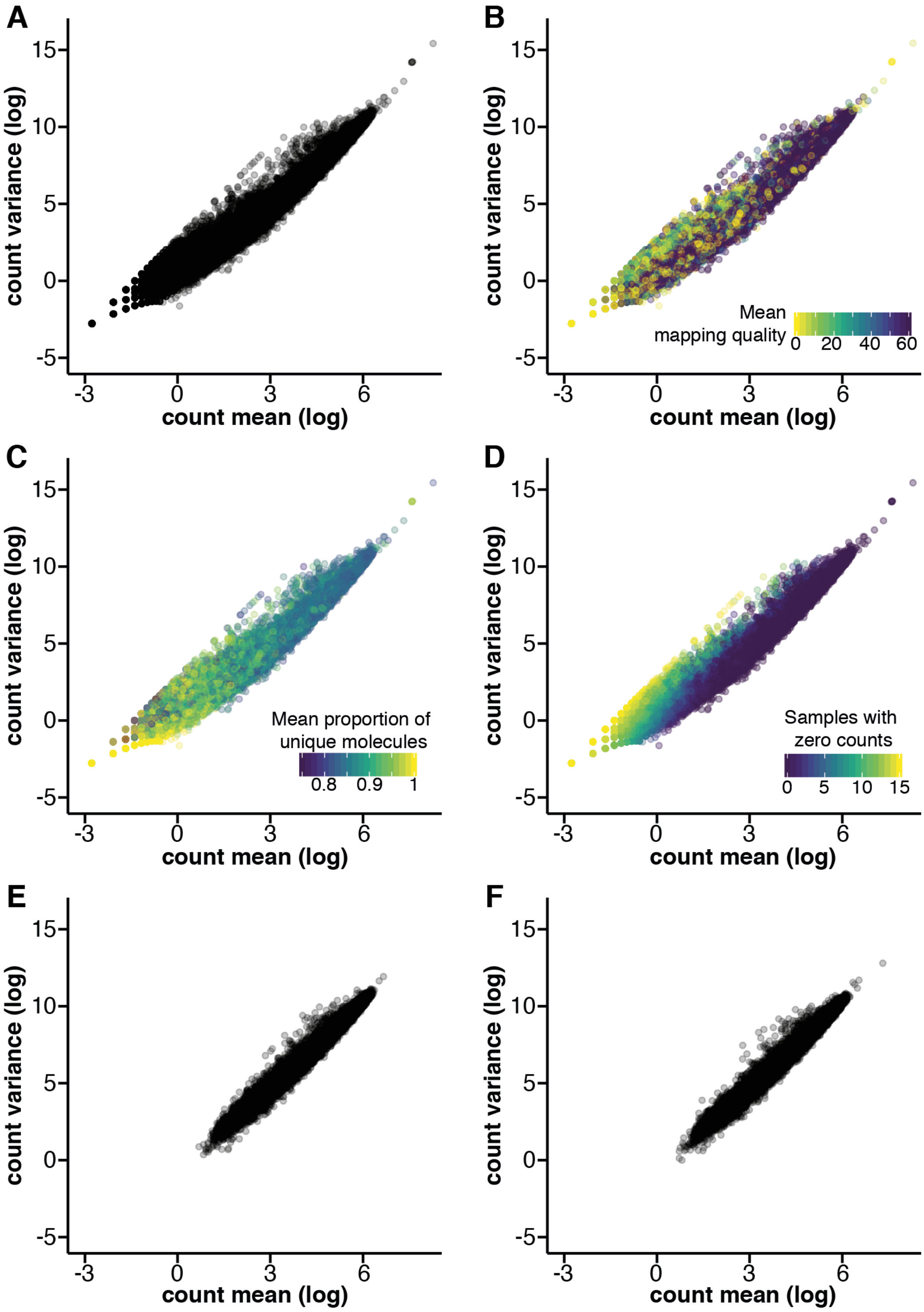

A
B
C
D
E
F
count variance (log)
count mean (log)
15
10
5
0
-5
-3
0
3
6
Mean
mapping quality
0
20
40
60
Mean proportion of
unique molecules
0.8
0.9
1
Samples with
zero counts
0
5
10
15

Figure 4.11 Exploring the mean-variance relationship between loci read counts and various loci features using 16 control samples. A: The mean of the unfiltered loci reads aligned to chromosomes 1-22 (using BWA-MEM) across 16 control samples against the variance of the loci read counts. Each point represents a unique locus defined by chromosome, align start, align end and strand. B: The mean-variance relationship shown in A but with each point (locus) coloured by the mean of the mean mapping quality across the 16 control samples. C: The mean-variance relationship shown in A but with each point (locus) coloured by the mean proportion of unique molecules across control samples. The proportion of unique molecules is calculated for each sample at every locus by computing the number of UMI clusters at the locus divided by the total number of reads aligned to the locus. D: The mean-variance relationship shown in A but with each point (locus) coloured by the number of control samples that had a zero count at the locus. E: The mean-variance relationship after filtering (keeping only alignments that had a mapping quality greater or equal to 50 and keeping only those loci that had a non-zero count in all 16 control sample). F: The mean-variance relationship when using the single-end pipeline (which uses Bowtie 1) and keeping loci that had a non-zero count in all 16 control samples.

Recall that, in the pipeline described in section 2.4, loci were filtered in two ways. Firstly, reads were filtered by Bowtie 1 if they aligned to multiple loci. This meant that FAST-SeqS loci that consisted of sequences within a certain Hamming distance of each other were automatically filtered. Secondly, loci were filtered if at least one control sample had a zero count occurring at the locus. The equivalent strategy for filtering loci obtained using the new pipeline would be to take a mapping quality threshold on the alignments, while the second step would remain the same.

Figure 4.11A shows the mean-variance relationship prior to filtering loci (using the 16 control samples). Figure 4.11B shows that some of the extreme high counts can be explained by low mapping quality. It is likely that these loci are being assigned reads that were amplified from several positions in the genome. Aside from those with extremely high mean count, the proportion of loci with high mean mapping qualities appears to increase with mean count.

Figure 4.11D highlights the number of control samples that have a zero count associated with each locus. Loci may have zero counts in control samples for various reasons, which may include:

- Sampling error; loci that have low amplification efficiencies may be present at low quantities in the sequencing library but are not "sampled" by the sequencer, i.e. do not hybridise to the flow cell and form clusters.
- Random off-target and non-recurrent amplification of a genomic locus that is sample specific. This would mean that most samples would not see alignments at these genomic loci.

- PCR or sequencing errors occur at the start or end of the read. This results in alignments that are a few bases upstream or downstream of the other alignments. Hence, they appear as unique "loci" with a new start or end position.

- CNVs in the control samples; particularly those that have homozygous deletions involving FAST-SeqS loci.

- Germline or somatic variants occurring in the amplified loci; if these occur in the sequence that the primers hybridise to then the PCR efficiency may be affected. Alternatively, if the variants occur within the sequence that is aligned to the genome, they may cause the read to be aligned elsewhere or with lower mapping quality.

- Random choice of equally good alignments. BWA-MEM and Bowtie 2 randomly select an alignment when a read maps equally well to multiple locations. This is likely to explain why there are several loci that have only a single read aligned to them in only one control sample. Note that these will also have a mapping quality of 0 in BWA-MEM.

- Finally, LINE1 insertion polymorphisms. There is variation in LINE1 insertions across the human population, especially in L1HS elements [256]. However, since we amplify relative few L1HS, this is unlikely to explain many of the zero counts.

Ideally, we would keep those loci for which sampling error was the cause of the zero counts. If we were able to identify those loci which involve CNVs, we may also wish to include them as long as we could account for the CNV when estimating the expected proportion of reads ($\hat{\boldsymbol{m}}$). However, being able to distinguish whether a CNV or some other mechanism is the cause of a zero count is likely to be difficult or impossible.

By using a combination of a mapping quality threshold of 50 (for reads aligned with BWA-MEM) and keeping loci that have a non-zero count in all controls, we can filter loci that may cause artifacts in the RCN profile. We see in Figure 4.11E that by doing this, we obtain a set of loci with a similar mean-variance relationship to those selected using Bowtie 1 and the single-end pipeline (Figure 4.11F).

In figures 4.11C and 4.12 I explore another metric which could be used to distinguish low-quality loci. As we discussed previously, the proportion of unique molecules aligned to a locus is likely to be related to the loci count. Figures 4.12A and B show the distribution of unique molecules prior to filtering. We see that those loci with low mean mapping quality tend to have high proportions of unique molecules, as expected. Once loci are filtered (using mapping quality and non-zero counts in controls), we see that the heavy tail representing loci with high mean proportion of unique molecules is mostly filtered from the distribution (figures 4.12C and D). While I do not use the proportion of unique molecules as a metric for loci filtering at the moment, potentially it could be used with other features of the data (such as the mapping quality, number of zero counts in controls, mean count, mean variance and others) to select the loci to filter. Using hard thresholds to filter on the number of non-zero counts in controls is likely to result in over-filtering the data. Instead of using hard filters, a machine learning model could be devised to select loci.

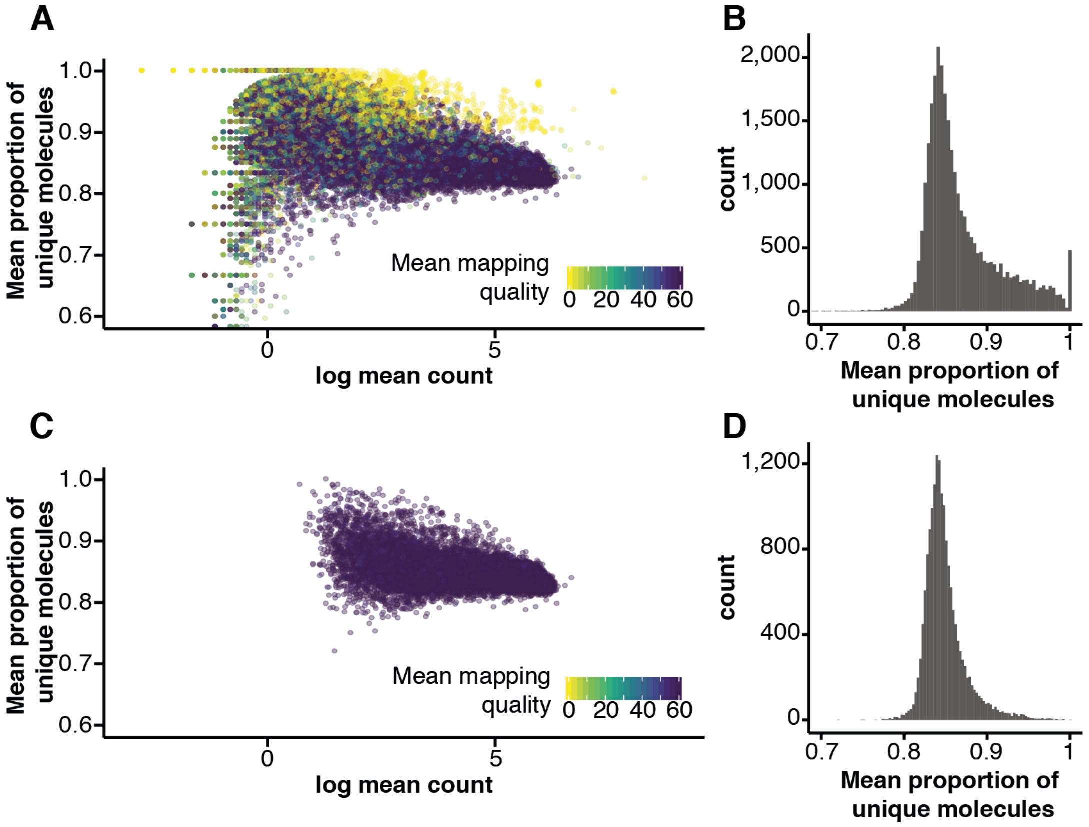


Figure 4.12 Exploring the mean proportion of unique molecules as a metric for loci quality. A: The log mean count against the mean proportion of unique molecules across the 16 control samples for each locus aligned to chromosomes 1-22 (prior to filtering). Each locus is coloured by the mean of the mean mapping quality across controls. B: The distribution of the mean proportion of unique molecules. C: Same as A but after filtering alignments with mapping quality below 50 and keeping only loci with a non-zero count in all 16 control samples. D: The distribution of the mean proportion of unique molecules after filtering.

4.4.2 Exploring sample-specific biases

In this section, I explore whether there are features of the samples or loci (other than those associated with read processing) that may introduce sample-specific biases to the data. Spurious RCN profiles are caused by read proportions being altered in a manner that does not fit with conliga's assumptions of the data generation (which I described in chapter 2).

There are two main ways in which sample-specific effects could result in spurious RCN profiles. Firstly, sample-specific effects could introduce a spatial signal or bias in the data. Spatial signal is precisely what conliga intends to detect and quantify. As such, any biases that result in spatial signal will result in spurious RCN profiles. Secondly, sample-specific effects may alter the distribution of the data. If the distribution of the data is altered such that the beta-binomial emission distribution (described in chapter 2) no longer fits the data appropriately, this can result in spurious copy number states.

GC content bias

GC content, which is the proportion of G and C bases in a sequence, is known to be associated with biases in high-throughput sequencing data [257–259]. Dohm *et al.* were the first to describe a dependence between GC content and read depth using Illumina machines [258, 259]. They showed that read depth originating from 1 Kbp bins from several bacterial genomes showed a linear relation with GC content; increasing read depth with increasing GC content [258]. Benjamini and Speed showed that the GC content of the fragment, that is the DNA fragment that is being sequenced, showed a non-linear and uni-modal relationship with read depth; with a peak observed around 50% GC. Futhermore, they claimed that the fragment GC content drives the effect seen with larger GC bins in the genome and showed that the relationship between GC content and read depth is not consistent between samples [259].

GC content biases are not limited to next-generation sequencing platforms. Technical, positive spatial autocorrelation, often described as 'wave' patterns, have been observed in array comparative genomic hybridisation (CGH) intensities and shown to correlate with regional GC content [260]. Van Heesch *et al.* carried out several technical experiments in an attempt to find the cause of the wave patterns [261]. They found that similar wave patterns were observed across technologies (array CGH and high-throughput sequencing) [261]. Like others, they found that the observed wave patterns correlated with GC content. This suggested that steps in the protocol prior to sequencing or array CGH introduce a bias that correlates with GC content. By analysing samples obtained from different tissues of the same rat, they showed that correcting for GC content was able to remove some, but not all, of the depth of coverage or intensity differences between tissues [261].

GC content correlates with many other characteristics of the genome. These include replication timing [262], gene density [3], LINE and SINE density, [3], genome isochores, gene expression and chromatin status; with GC poor regions associated with tightly packed heterochromatin and GC rich regions associated with more loosely packed euchromatin [261]. With chromatin organisation varying between tissues, Van Heesch *et al.* hypothesised that chromatin status may be the cause of tissue-specific biases; with greater quantities of DNA being isolated from euchromatin and less from heterochromatin. To test this hypothesis, they prepared multiple samples across multiple tissues with varying proteinase K lysis durations (proteinase K is used to digest the chromatin protein structures). They found that longer durations lead to more even coverage across the genome and increasingly similar coverage profiles between tissues. Furthermore, they found that tissues such as blood and liver, which are more homogeneous, had the greatest biases in read coverage [261].

Their data suggested that chromatin status explains some of the depth of coverage biases associated with GC content. Variations in DNA isolation procedures between samples, particularly in proteinase K treatment duration, may lead to spatial variation in the proportions of DNA isolated from the genome [261]. This is important for copy number tools, like conliga, that rely on depth of coverage data to make inferences. In large-scale studies such as those performed by ICGC and The Cancer Genome Atlas (TCGA), samples are often processed in several batches, by different people and potentially using different protocols. Ensuring that the DNA isolation protocols are consistent between samples and batches may help in reducing sample-specific biases. Since blood is easy to obtain, it is often used as a normal control. However, these results highlight the importance of using matched tissues as controls. If it is necessary to use blood instead of matched tissue, it appears that increasing the proteinase K lysis duration should be considered to reduce tissue-specific biases.

By fitting a thin-plate spline surface to the GC content across the genome, each FAST-SeqS locus was assigned a large-scale GC content value associated with its location in the genome which I refer to as the smoothed GC value. This was achieved by obtaining the binned GC content of the genome (GRCh38) in 100 Kbp intervals using the the bedtools `nuc` tool [250]. A thin-plate spline was then fitted to the binned GC content (excluding those bins that contained ambiguous bases) for each chromosome independently using the mgcv package [263] in R (`gc_model = gam(gc ∼ s(centre, bs="tp", k=200)`, where `centre` represents the centre of the genomic bin). Following this, each locus was assigned a smoothed GC value using the `predict` function from the mgcv package. In addition, I computed the GC content of the sequence fragment (amplicon) including the region that the primers hybridise to, which I call the amplicon GC value. The Pearson correlation between the smoothed GC and amplicon GC was low with a value of 0.12. In an attempt to not confound amplification efficiency bias (see section 4.4.2) with this exploration, I selected only those loci that had an exact match to the forward and reverse primer sequences. When comparing the amplicon GC to the count proportion (figure 4.13), we see a non-linear relationship consistent with that described by Benjamini and Speed [259]. Note that, compared with WGS fragments, the range of FAST-SeqS amplicon GC values is narrow with few values below 50% seen.

Figure 4.14 shows the relationship between the smoothed GC and count proportion. In some samples, the count proportion increases with smoothed GC whereas in others it appears to decrease or not vary with smoothed GC. Given the results of Van Heesch *et al.*, I would have expected to see the count proportions either increase with smoothed GC (i.e. increase in

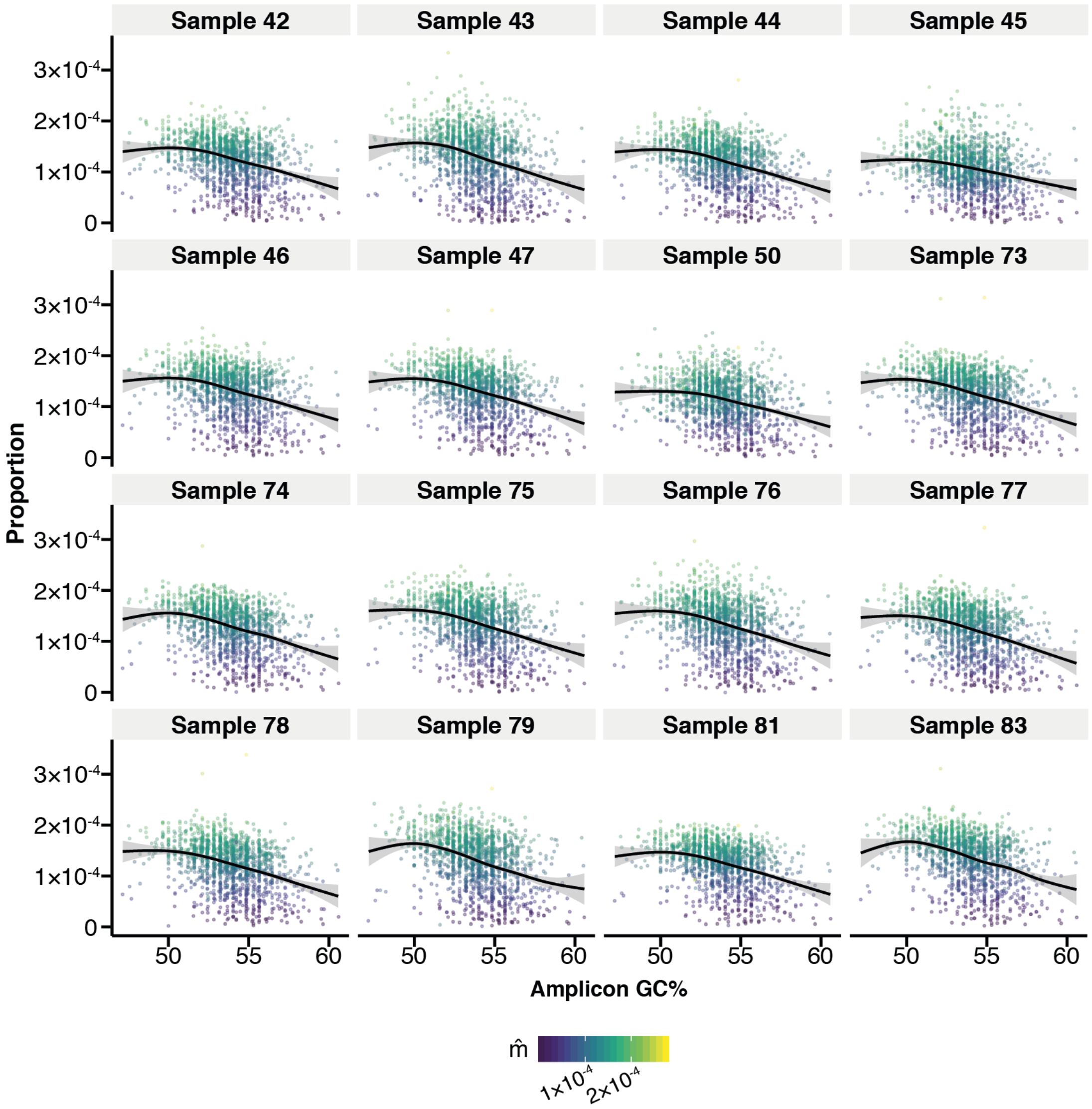


Figure 4.13 The relationship between amplicon count proportion and amplicon GC content (shown for each control sample). Each point represents a locus. Only loci on chromosomes 1-22 and those that have exact matches to the forward and reverse primer sequences are shown. Each point is coloured by the expected proportion of reads to align to the locus, $\hat{m}$.

euchromatic regions) or otherwise see no change with GC. However, we also see proportionally more counts associated with low GC regions (heterochromatic regions) in some samples which was not expected. While the relationship between amplicon GC and smoothed GC seems to differ between samples, I explored the inferred copy number profiles to see if the smoothed GC or amplicon GC could explain the multiple copy number states seen in some control samples. Figure 4.15 shows that these additional states tend to appear in regions of high and low smoothed GC. This was not seen with amplicon GC (not shown).

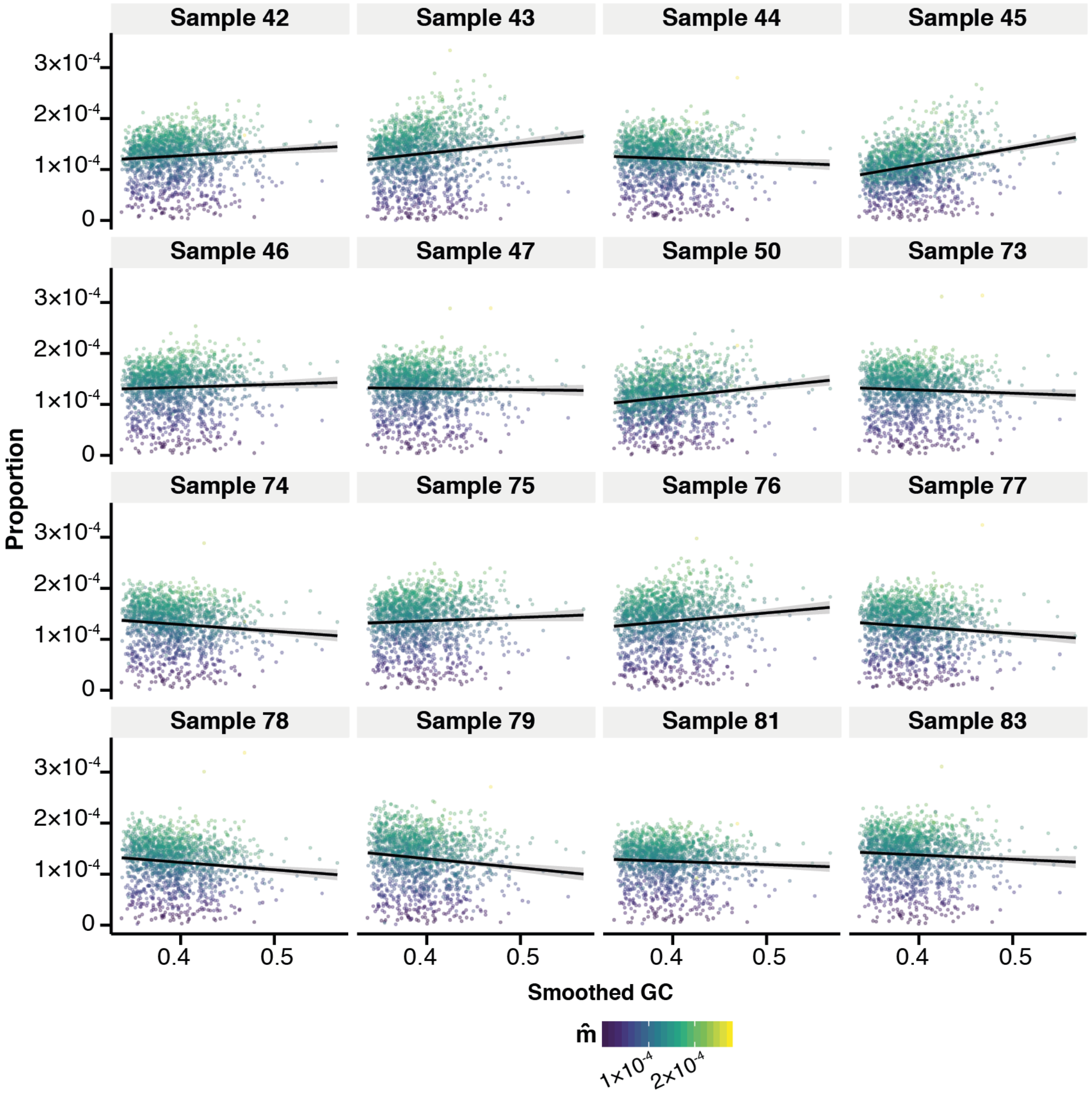


Figure 4.14 The relationship between amplicon count proportion and each amplicon's associated smoothed GC value (shown for each control sample). Each point represents a locus. Only loci on chromosomes 1-22 and those that have exact matches to the forward and reverse primer sequences are shown. Each point is coloured by the expected proportion of reads to align to the locus, $\hat{m}$

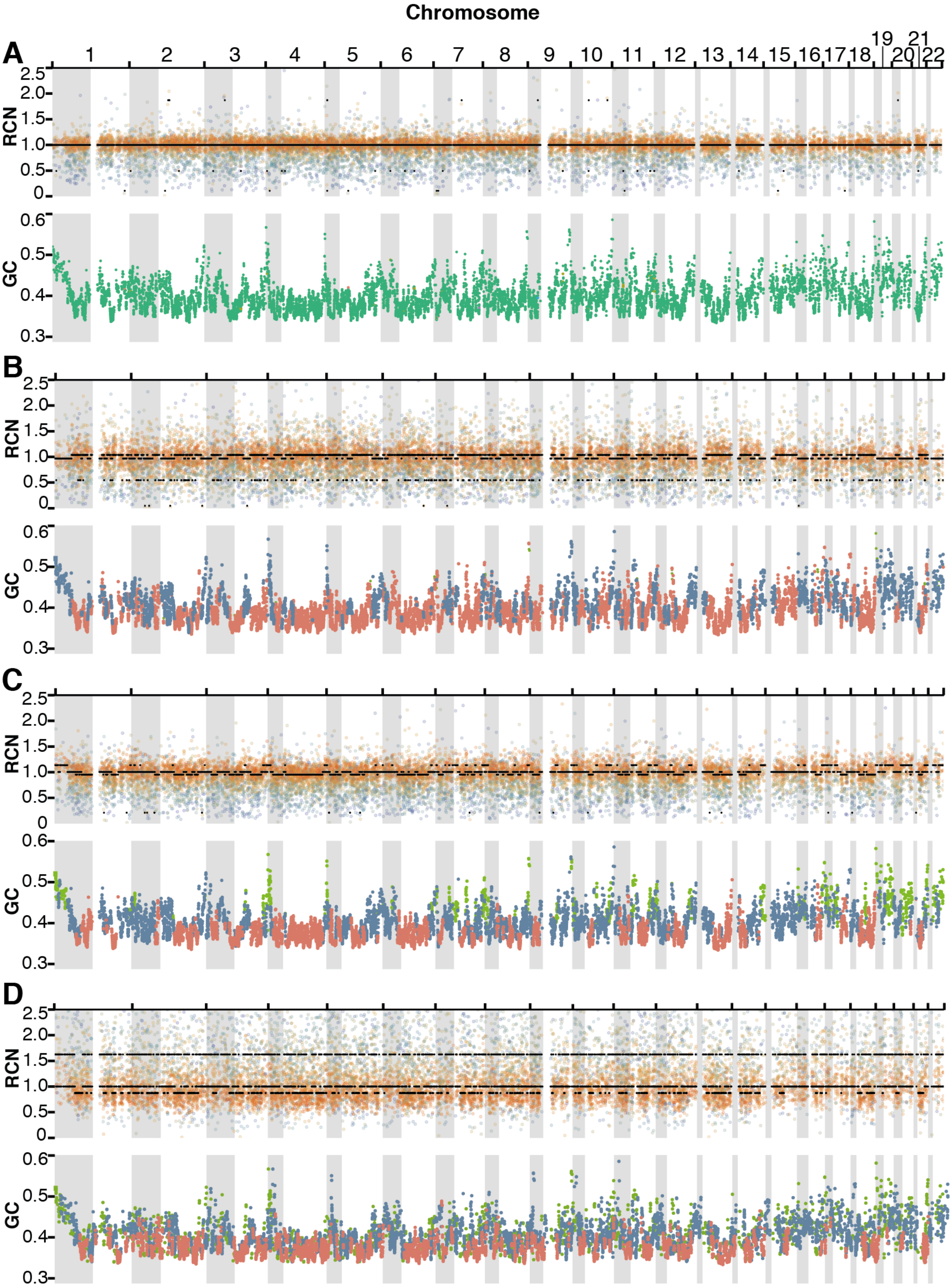

Chromosome
A
B
C
D
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
RCN
GC
2.5
2.0
1.5
1.0
0.5
0
0.6
0.5
0.4
0.3

Figure 4.15 Comparing the inferred RCN profiles of four control samples with smoothed GC. A-D show the result for four different control samples. Top: The inferred RCN profile using the *MAP state* method for the sample. Bottom: The smoothed GC value assigned to each locus, coloured by the relabelled copy number state. A: Sample 47. B: Sample 79. C: Sample 43. D: Sample 45.

Sample-specific differential loci amplification efficiencies

While exploring the effect of smoothed GC on the inferred copy number states in figure 4.15D, we see that, in sample 45, the scaled counts with low and high $\hat{m}$ values are not symmetrically distributed around 1. Recall that we saw this same observation in chapter 2 (figure 2.5) when analysing the RCN profiles of male and female samples. I noted that this did not seem to affect the RCN inference in that case. However, in sample 45, we see an additional copy number state with an RCN of approximately 1.6. Indeed, the loci with low $\hat{m}$ appear to be distributed around 1.6, while the loci with high $\hat{m}$ seem to be distributed slightly below an RCN of 1.

In figure 4.16, by comparing the observed count proportion of each control with the inferred expected proportion ($\hat{\boldsymbol{m}}$), we see that the relationship is not always linear as we would expect, e.g samples 45, 50 and 81. This figure also clearly highlights the GC effect described in the previous section. Figure 4.15 along with figure 4.16 suggest that there is a differential amplification of lowly and highly amplified loci between samples. For example, in sample 45, loci that are typically lowly amplified are seen in higher count proportions than expected and the opposite is also true. The cause of this is not clear but may be related to differing quantities of input DNA or primers. This effect is not seen in all samples and further experiments to ascertain the cause are required.

Loci which have an imperfect match of their target region to the primer sequence are frequently amplified. Indeed, a large part of the differential amplification efficiencies between loci is associated with the edit distance or Hamming distance of the primer sequence to the target region (not shown).

4.5 Updating the conliga model to account for sample-specific biases

In the conliga method presented in chapter 2, I made a central assumption that biases were introduced by the protocol (e.g. differential amplification efficiencies between loci) and by the read processing (e.g. mappability to the genome). These biases were captured in a vector,

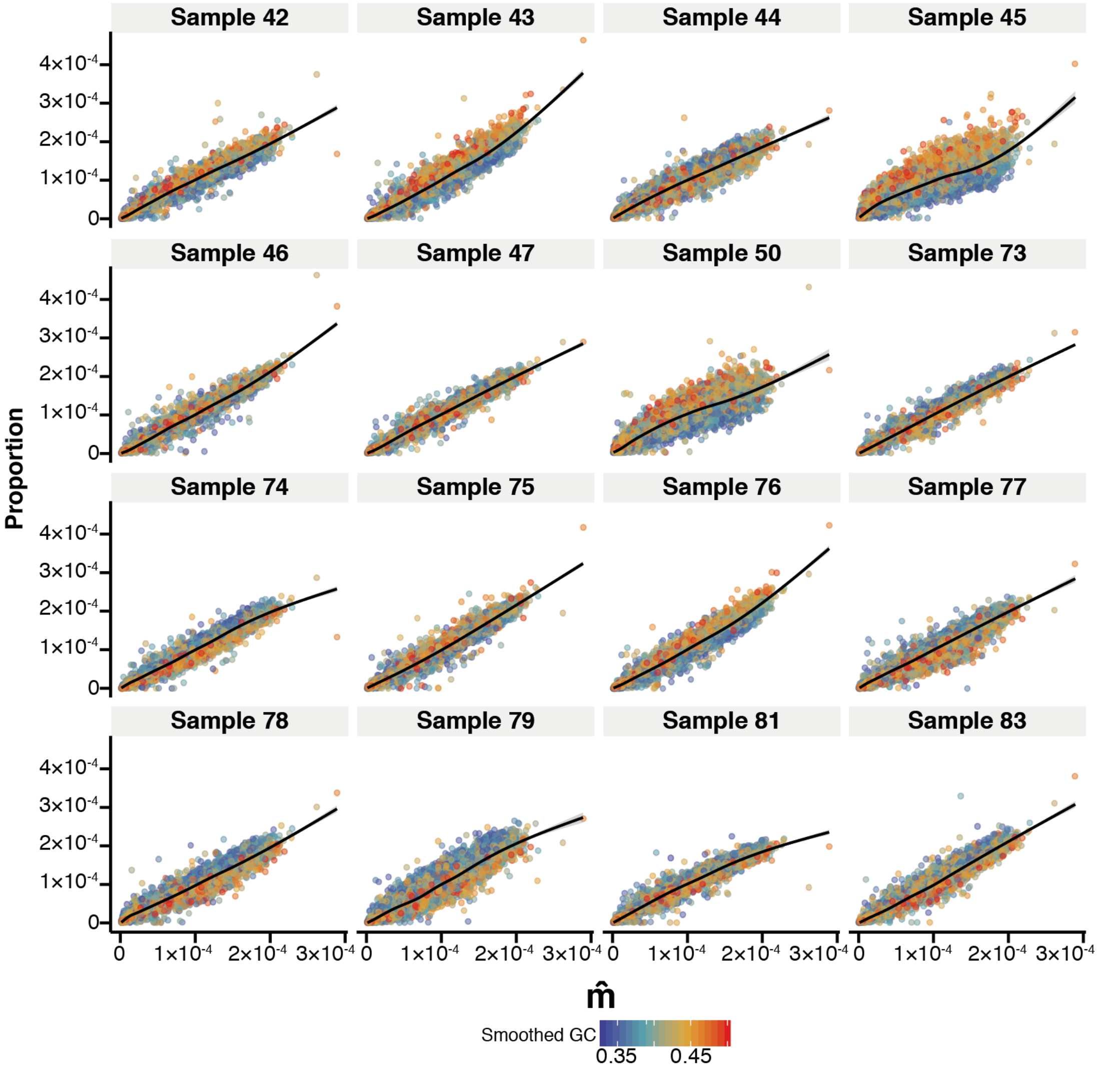


Figure 4.16 The expected proportion of aligned reads, $\hat{\boldsymbol{m}}$, against the observed count proportion for all 16 control samples. Only loci within chromosomes 1-22 are shown. Each point represents a locus and is coloured by its associated smoothed GC value.

$\hat{\boldsymbol{m}}$, describing proportions of reads we expect to align to each locus in a normal (euploid) sample. I made the assumption that sample-specific biases were not present. As we saw in chapter 3, the method produced very similar RCN profiles to those commonly applied to WGS data. As such, this assumption seemed to be sufficient in most cases. However, by exploring new data, it appears that sample-specific biases can be introduced and they can result in the inference of spurious RCN profiles.

As I have shown, the spurious RCN profiles and sample-specific biases seem to be associated with large-scale GC content (smoothed GC) and differential amplification of highly and lowly amplified loci between samples. I will now refer to these biases as GC bias and amplification bias respectively.

Rather than model a global bias $\boldsymbol{m}$, we can assume that there is some baseline or reference bias, $\boldsymbol{\eta^0}$, and some multiplicative sample-specific bias, $\boldsymbol{b_j}$, which together give the expected proportion of reads for the sample (assuming the sample is normal),

$$\eta_{l,j} = \eta_l^0 \cdot b_{l,j} \tag{4.1}$$

here, $b_{l,j}$ captures a sample-specific bias for locus, l, and sample j, which may be composed of several underlying sources.

Although I found GC and amplification biases in the data, there are potentially others. Therefore, I tried to learn these biases from the data in a flexible manner. Here, I assumed that $b_{l,j}$ was composed of several underlying biases that were multiplicative. Assuming D sources of bias and instead modelling $\log \eta_{l,j}$, this can be written as,

$$\log \eta_{l,j} = \log \eta_l^0 + B_{l,1}F_{1,j} + \cdots + B_{l,D}F_{D,j} \tag{4.2}$$

which can be written more compactly in matrix notation for all samples and loci,

$$\log \boldsymbol{\eta} = \log \boldsymbol{\eta^0} + \boldsymbol{BF} \tag{4.3}$$

Note that equation 4.3 is similar in form to a general factor model, in which $\boldsymbol{B}$ represents the loading matrix and $\boldsymbol{F}$ represents the factor weights. Since $\boldsymbol{\eta}$ now replaces $\boldsymbol{m}$ in the generative model of sample counts, instead of including an error term that is typical in factor models, the noise is captured by the beta-binomial.

I attempted to use the control samples to learn $\boldsymbol{B}$ and $\boldsymbol{F}$ from the data. I do not show the results here, however I note that I found this approach to be unsuccessful. When using few dimensions ($D \leq 3$), and by using 16 control samples, the model failed to adequately capture the biases and correct the count proportions in the controls. Even with increasing D, the model failed to capture the biases and instead, tended to over-fit to particular samples. There could be several reasons for this. Perhaps more control samples are needed to be able to learn the biases from the data. Also, given that each sample has its own inverse-dispersion parameter, it may be that the model is insufficiently constrained; some samples could be inferred to have unrealistically low or high count dispersion. Indeed, it may have been appropriate to place a hierarchical prior on the sample inverse-dispersion parameters to draw or shrink them towards a common value.

As the flexible factor model seemed to overfit, I chose to correct the two features I identified to have a strong effect. I did this using a 2-dimension B-spline. The B-spline has the advantage that they can capture non-linear effects but be expressed as a linear model. This meant that the B-spline design matrix now becomes the loading matrix, $\boldsymbol{B}$, in equation 4.3.

I used the `bs` function from the `splines` R package [237] to generate the B-spline basis matrix associated with the GC bias and amplification bias. Here, the predictor variables used were the smoothed GC values and the log mean count proportions across control samples. The knot vector for smoothed GC was defined at intervals of 0.03 for smoothed GC (in which GC is measured as a proportion), starting from the minimum smoothed GC value associated with the locus minus 0.1 until the maximum smoothed GC value plus 0.1. The knot vector for log mean proportions was defined at intervals of 0.6, starting from the minimum log mean proportion minus 1 until the maximum log mean proportion plus 1. The `tensor.prod.model.matrix` function from the `mgcv` R package was used to produce the 2-dimensional B-spline matrix given the smoothed GC and log mean proportion B-spline matrices as an input. For the prostate cancer data set this resulted in a matrix ($\boldsymbol{B}$) of size 16,403 by 156.

The `optimizing` function from the `Rstan` package [264] was used to provide a MAP estimate for $\log \boldsymbol{\eta}^{\mathbf{0}}$, $\boldsymbol{F}$ and $\boldsymbol{s}$, given the counts, $\boldsymbol{x}$, for all control samples $(1, \ldots, J)$ and the matrix $\boldsymbol{B}$. The RStan code describing the model can be found in Appendix B. This model replaces algorithm 1 that was used to infer $\boldsymbol{m}$ and $\boldsymbol{s}$ described in chapter 2. To make the model identifiable, the last column of $\boldsymbol{F}$, representing the factors associated with sample J, were set to zero. This means that $\boldsymbol{\eta}^{\mathbf{0}}$ can be interpreted as the expected proportion of reads for sample J. Therefore, sample-specific biases are measured relative to those of sample J.

Figure 4.17 and 4.18 show the smoothed GC against the scaled counts of the control samples without (algorithm 2) and with correction (Rstan model, appendix B), respectively. We see that, without correction the effect of GC and amplification bias can be seen in the scaled counts and we can see the sample-specific nature of the biases. With correction, the biases appear to be corrected; the relationship between the scaled counts and smoothed GC is no longer present and we do not see a shift in the scaled counts that is dependent on $\boldsymbol{\eta}$.

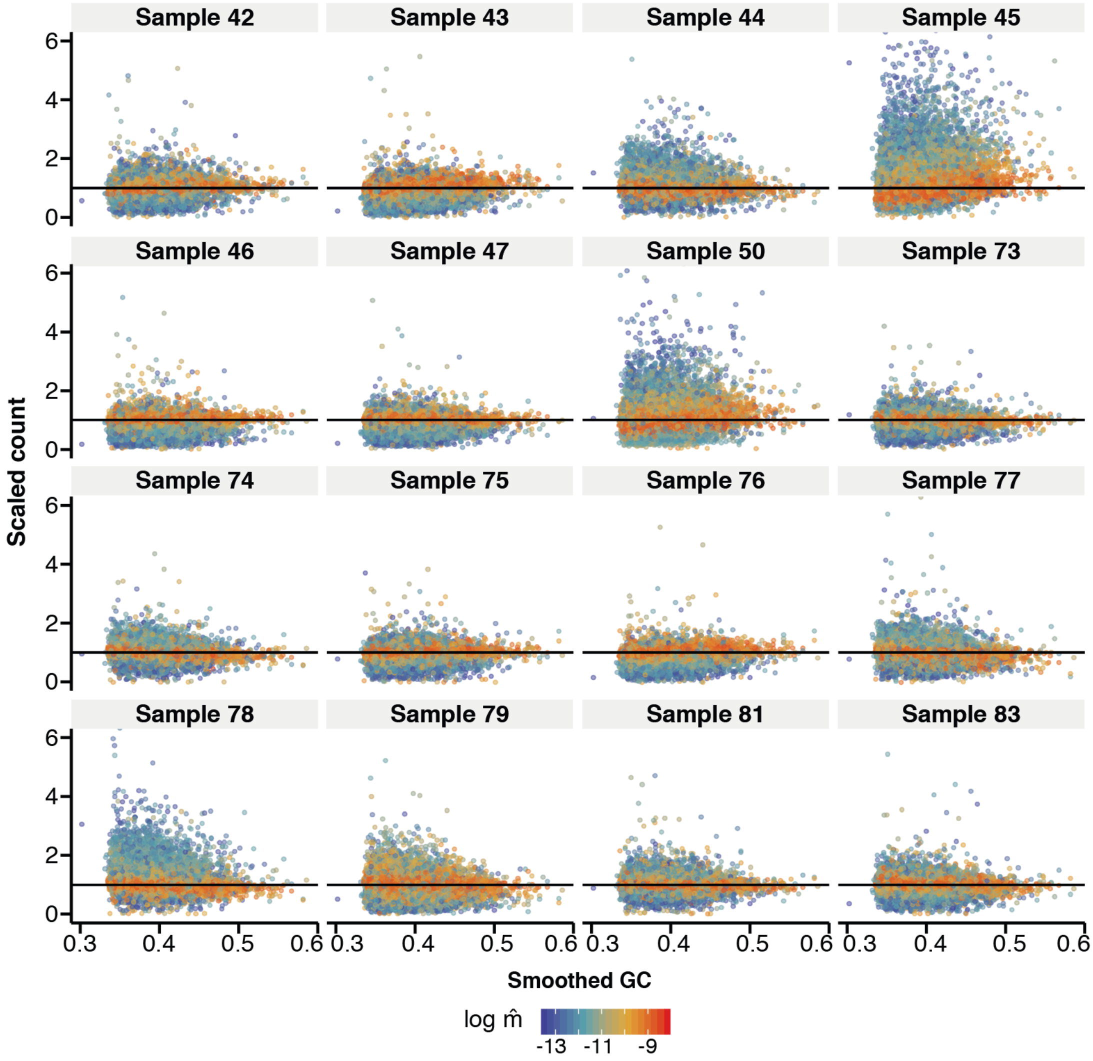


Figure 4.17 Smoothed GC against scaled count without correcting for GC and amplification biases for all control samples. Each point represents a FAST-SeqS locus within chromosomes 1-22. Each point is coloured by the log of its associated inferred expected proportion of aligned reads ($\hat{m}$) using algorithm 2.

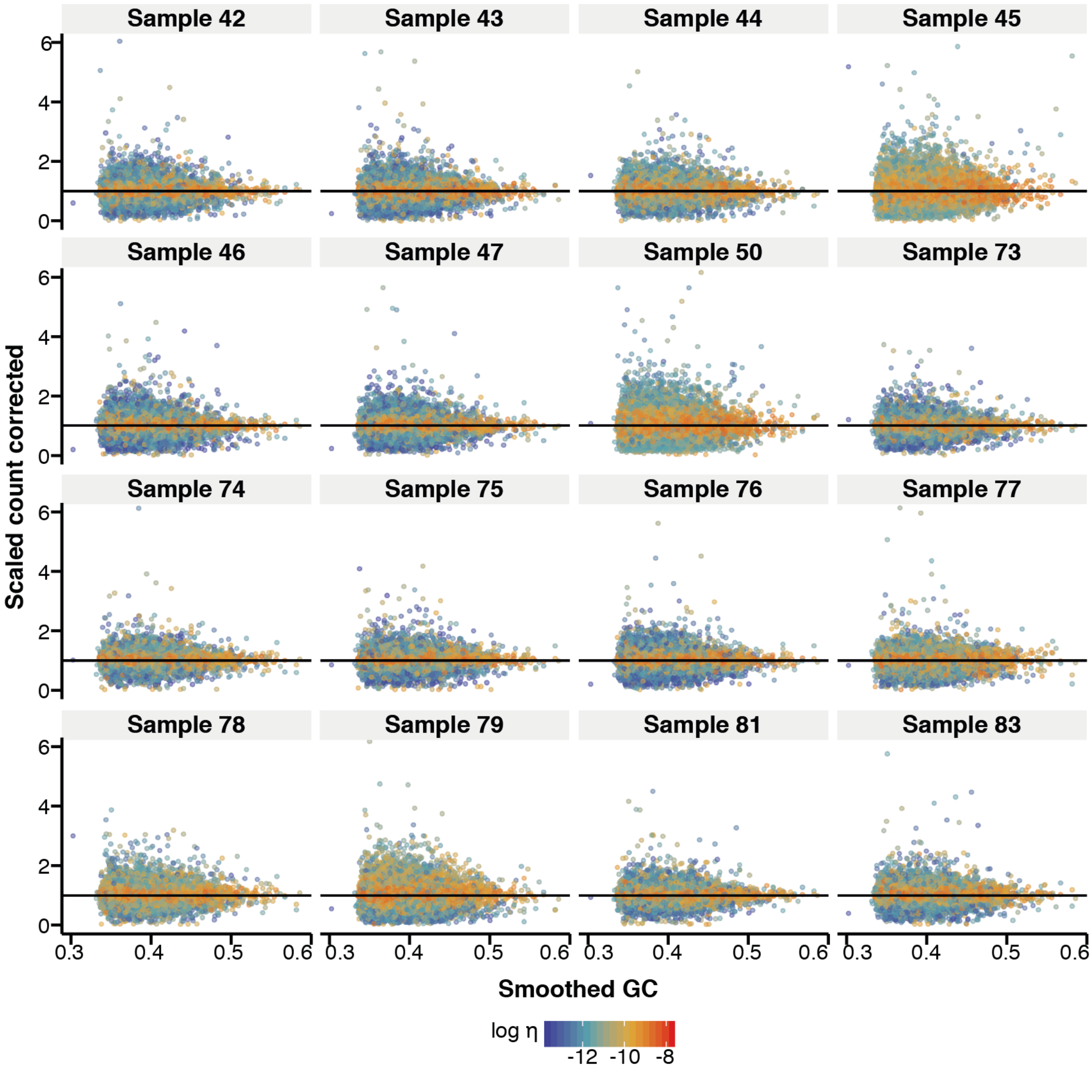


Figure 4.18 Smoothed GC against scaled count after correcting for GC and amplification biases for all control samples. Each point represents a FAST-SeqS locus within chromosomes 1-22. Each point is coloured by the log of its associated, sample-specific inferred expected proportion of aligned reads (η) using the Stan model shown in Appendix B.

Updating the MCMC algorithm to account for sample-specific biases

I updated the MCMC algorithm described in algorithm 2 (section 2.7) to jointly infer the sample-specific biases and the RCN profile for a test sample. In the updated algorithm, the $\hat{\boldsymbol{m}}$ vector is replaced and $\boldsymbol{B}$, $\log \boldsymbol{\eta}^0$ along with priors on $\boldsymbol{F}$ are passed to the algorithm instead. By jointly inferring the factors ($\boldsymbol{F}$) associated with the test sample, $\boldsymbol{\eta}$ replaces $\hat{\boldsymbol{m}}$. In order to sample $\boldsymbol{F}$, I added an additional Metropolis-Hastings step to algorithm 2. As there are a large number of correlated parameters, I used a block update step to update $\boldsymbol{F}$ rather than updating each factor independently. To do this, the proposal was drawn from a multivariate Normal distribution centred at the previous values of $\boldsymbol{F}$ and with a correlation matrix, $\boldsymbol{\Sigma}$. The value of $\boldsymbol{\Sigma}$ that I used is the same as that used in the Gibbs update of a multivariate Normal regression, i.e. $(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B})^{-1}$. While we are not performing multivariate Normal regression, the correlation structure is likely to be similar and should result in a more efficient sampling of $\boldsymbol{F}$. Algorithm 3 describes the blocked update Metropolis-Hastings step which is called before the `sampleHyperparameters` function in algorithm 2.

Algorithm 3 Metropolis-Hastings block update step for $\boldsymbol{F}$

```
function SAMPLEF($\boldsymbol{B}, \log\boldsymbol{\eta}^{\mathbf{0}}, \boldsymbol{\Sigma}, \sigma_F, \psi_\mu, \psi_{\sigma^2}$)
    Draw proposed factors from multivariate Normal:
    $\boldsymbol{F}^* \sim \text{Normal}(\boldsymbol{F}, \boldsymbol{\Sigma})$
    Calculate resulting proposed $\boldsymbol{\eta}^*$ and $\boldsymbol{\eta}$:
    $\boldsymbol{\eta}^* \leftarrow \boldsymbol{B} \cdot \boldsymbol{F}^* + \log\boldsymbol{\eta}^{\mathbf{0}}$
    $\boldsymbol{\eta} \leftarrow \boldsymbol{B} \cdot \boldsymbol{F} + \log\boldsymbol{\eta}^{\mathbf{0}}$
    Ensure resulting proposed $\boldsymbol{\eta}^*$ and $\boldsymbol{\eta}$ sum to 1:
    $\boldsymbol{\eta}^* \leftarrow \boldsymbol{\eta}^* / \sum_{r,l} \eta^*_{r,l}$
    $\boldsymbol{\eta} \leftarrow \boldsymbol{\eta} / \sum_{r,l} \eta_{r,l}$
    $p_{\text{old}} \leftarrow \sum_{r\in\mathcal{R}} \sum_{l=1}^{L_r} \big(\log f_{BB}(y_{r,l}; \tilde{n}, \tilde{s}\hat{c}_{z_{r,l}}\eta_{r,l}, \tilde{s}(1 - \hat{c}_{z_{r,l}}\eta_{r,l}))\big) + \sum_d \log f_{\text{Normal}}(F_d; \psi_\mu, \psi_{\sigma^2})$
    $p_{\text{new}} \leftarrow \sum_{r\in\mathcal{R}} \sum_{l=1}^{L_r} \big(\log f_{BB}(y_{r,l}; \tilde{n}, \tilde{s}\hat{c}_{z_{r,l}}\eta^*_{r,l}, \tilde{s}(1 - \hat{c}_{z_{r,l}}\eta^*_{r,l}))\big) + \sum_d \log f_{\text{Normal}}(F^*_d; \psi_\mu, \psi_{\sigma^2})$
    $p \sim U(0,1)$
    if $\exp(p_{\text{new}} - p_{\text{old}}) < p$ then
        Accept new value: $\boldsymbol{F} = \boldsymbol{F}^*$
        Set $\boldsymbol{\eta} \leftarrow \boldsymbol{\eta}^*$
     return $\boldsymbol{F}, \boldsymbol{\eta}$
```

Here, $f_{\text{Normal}}(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$. In practice the current values of $\log f_{BB}(y_{r,l}; \tilde{n}, \tilde{s}\hat{c}_{z_{r,l}}\eta_{r,l}, \tilde{s}(1 - \hat{c}_{z_{r,l}}\eta_{r,l}))$ for all r, l, and $\boldsymbol{\eta}$ are cached following the update of the relative copy number for each state and the sample inverse-dispersion.

4.6 Comparing results: inferred RCN profiles with and without bias correction

In this section, I compare some of the results obtained from the original conliga model, which did not correct for sample-specific biases, with those obtained with the new model. There are several things to investigate. Firstly, do we see flat RCN profiles when inferring SCNAs in normal samples? Secondly, are true SCNAs present after bias correction or are they inadvertently regressed away? Thirdly, are false positive SCNA calls introduced by the correction of biases? Finally, do we see evidence of true SCNAs when correcting biases that were not seen before correction?

Figure 4.19 shows the RCN profiles inferred for three control samples when correcting for biases. These control samples were presented in figure 4.7 as examples of samples that had inferred profiles with spurious RCN states when not correcting for biases. Sample 45 was strongly affected by GC and amplification bias and was inferred to have a flat RCN profile by jointly inferring the sample-specific biases. Recall from figures 4.19, 4.14 and 4.16 that higher smoothed GC was associated with increased read proportions in sample 43 and decreased read proportions for sample 79. In both these cases, the biases were corrected with the updated method and flat RCN profiles were obtained. In all three samples, we see loci with short RCN changes, involving a single locus or very few loci. However, note that we have seen this previously (chapter 2, figure 2.12 for example) and I believe this is unlikely to have been introduced by the bias correction.

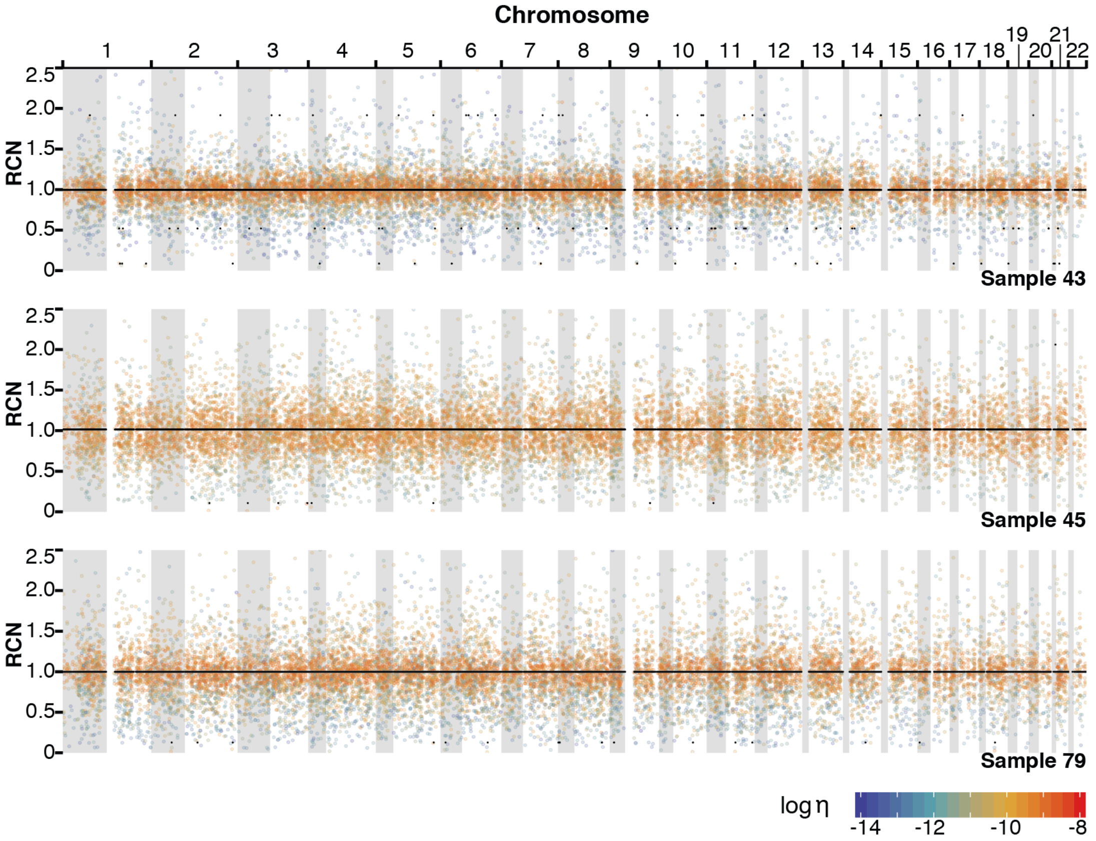


Figure 4.19 Inferred RCN profiles for three normal prostate samples using sample-specific bias correction. These samples were shown to have spurious RCN profiles in in figure 4.7. Each black point represents the MAP RCN inferred for each locus. The coloured points represent the scaled count for each locus coloured by the log of the inferred expected proportion of reads to align to each locus for each sample, $(\log \eta)$, which takes into account the sample-specific biases.

Figures 4.20 and 4.21 show examples of inferred RCN profiles in prostate cancer samples with and without bias correction. Samples from these patients were also sequenced using HC WGS as part of the prostate ICGC project. However, the samples we sequenced using FAST-SeqS were spatially related but non-identical samples taken from regions in close proximity to those sequenced by WGS. While we are able to see the loss of 8p and gain of 8q without bias correction in sample 7, the RCN profile inferred for the other chromosomes is consistent with that influenced by a strong GC bias. After bias correction is performed, the GC signal is removed and the loss of chromosome 8p and gain of 8q remains. Moreover, we are able to detect the subclonal deletions, called by Batternberg [265] using the HC WGS data, involving chromosome 5q and 12q that were not detectable by conliga without bias correction. Indeed the subclonal alteration within 12q includes 12q24.3 which is commonly deleted in prostate cancer [266]. Prior to correction, sample 9 has evidence of strong GC bias also (figure 4.21). After correction, the GC bias signal is removed. Encouragingly, several subchromosomal deletions are not regressed away in the process. These include deletions seen within 2q, 3p, 3q, 5q, 6q and 13q. Indeed, a short focal deletion which involves a deletion of both alleles on 2q is observed and even very short alterations such as these do not appear to be regressed away by the bias correction.

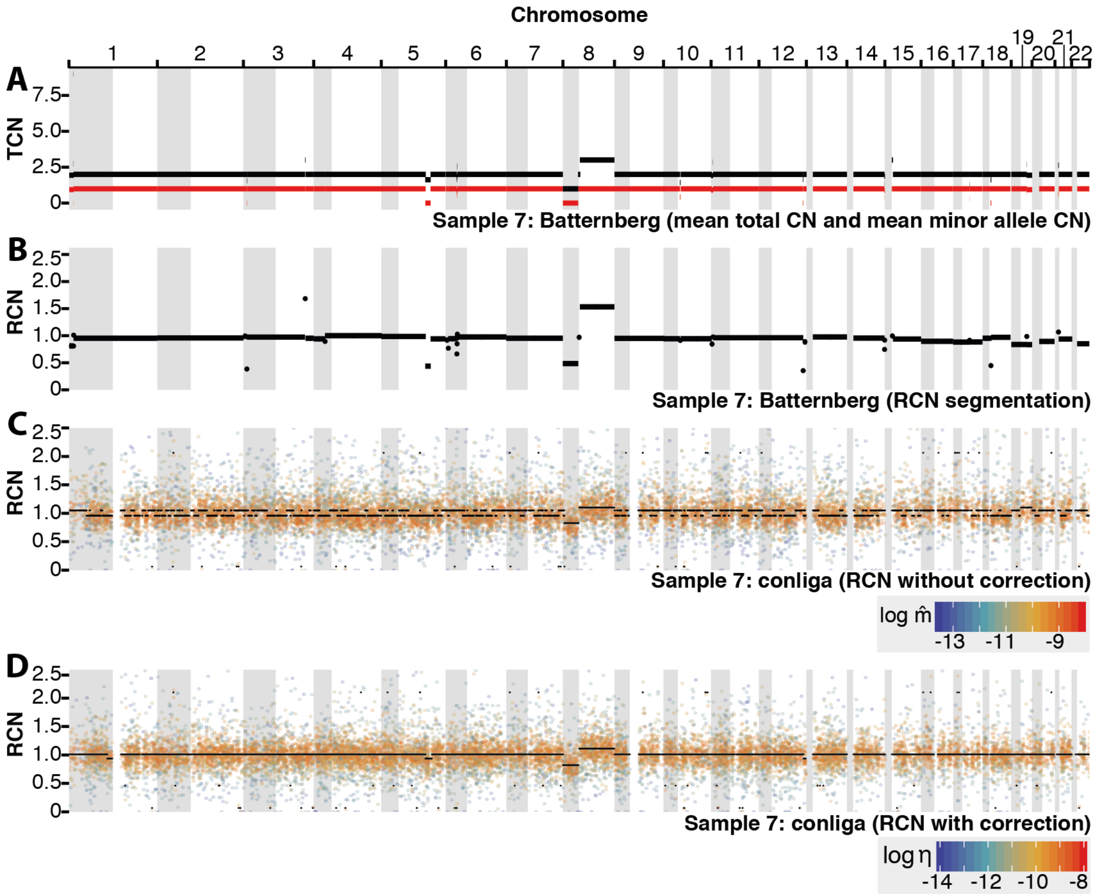


Figure 4.20 Comparison of Batternberg inferred copy number profiles (allowing for subclonality) and conliga inferred RCN profiles for prostate cancer sample 7 with and without bias correction. A: Inferred Batternberg profile showing the mean total copy number of the clones (black) and the mean minor allele copy number of the clones (red). B: The RCN segmented profile called by Batternberg prior to subclonal inference. C: RCN profile inferred by conliga without bias correction. D: RCN profile inferred by conliga with bias correction.

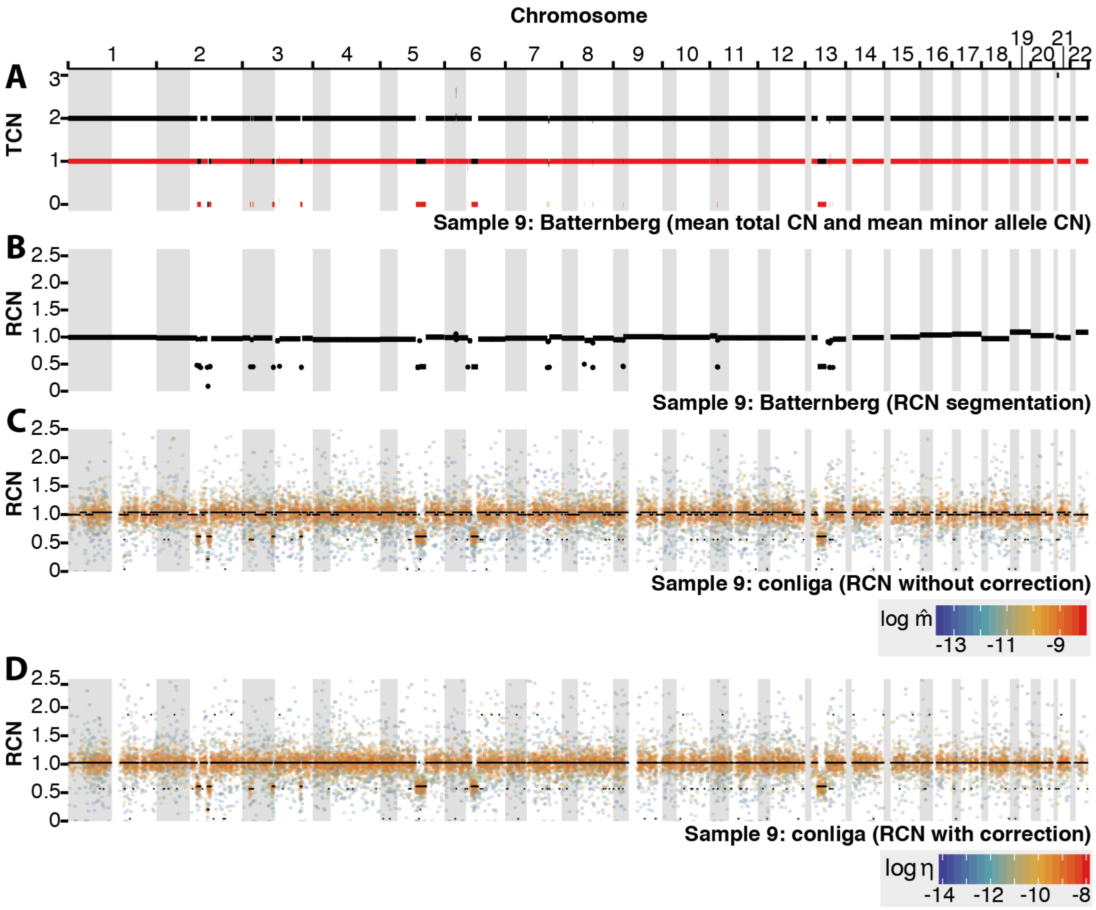


Figure 4.21 Comparison of Batternberg inferred copy number profiles (allowing for subclonality) and conliga inferred RCN profiles for prostate cancer sample 9 with and without bias correction. A: Inferred Batternberg profile showing the mean total copy number of the clones (black) and the mean minor allele copy number of the clones (red). B: The RCN segmented profile called by Batternberg prior to subclonal inference. C: RCN profile inferred by conliga without bias correction. D: RCN profile inferred by conliga with bias correction.

Figure 4.22 provides another example corresponding to sample 65. In this case, GC and amplification biases do not appear to introduce spurious copy number states prior to correction. This is despite amplification bias being present in the data, whereby lowly amplified loci typically have a lower count than expected. The profiles obtained from each tool are similar. This provides further evidence that true SCNAs do not appear to be regressed away by the bias correction. Figure 4.23 shows more detailed profiles for chromosomes 3, 5, 13 and 17 for sample 65. In chromosome 3, conliga (with and without bias correction) is able to detect a loss including 3p13, which is commonly deleted in prostate cancer [266] and a subclonal deletion occurring at the start of 3q (figure 4.23A). In addition to this, two focal deletions are detected that are also detected by Batternberg.

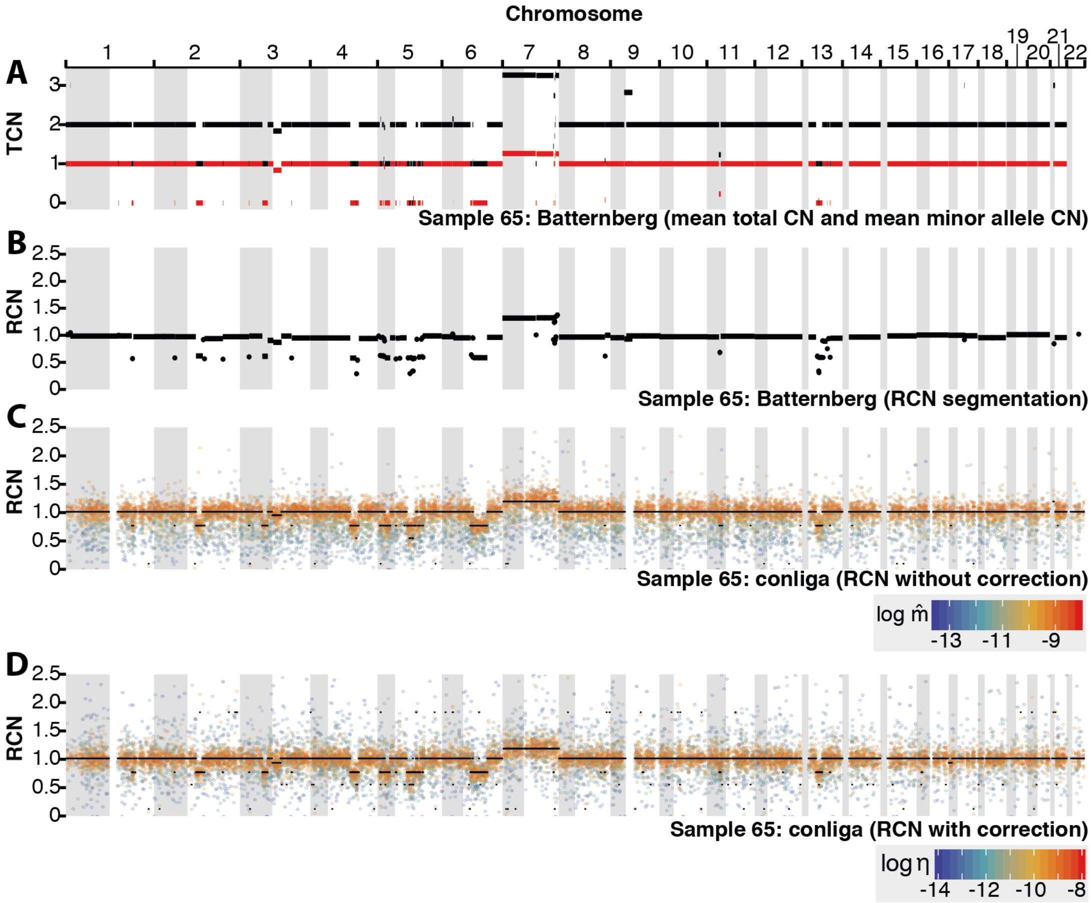


Figure 4.22 Comparison of Batternberg inferred copy number profiles (allowing for subclonality) and conliga inferred RCN profiles for prostate cancer sample 65 with and without bias correction. A: Inferred Batternberg profile showing the mean total copy number of the clones (black) and the mean minor allele copy number of the clones (red). B: The RCN segmented profile called by Batternberg prior to subclonal inference. C: RCN profile inferred by conliga without bias correction. D: RCN profile inferred by conliga with bias correction.

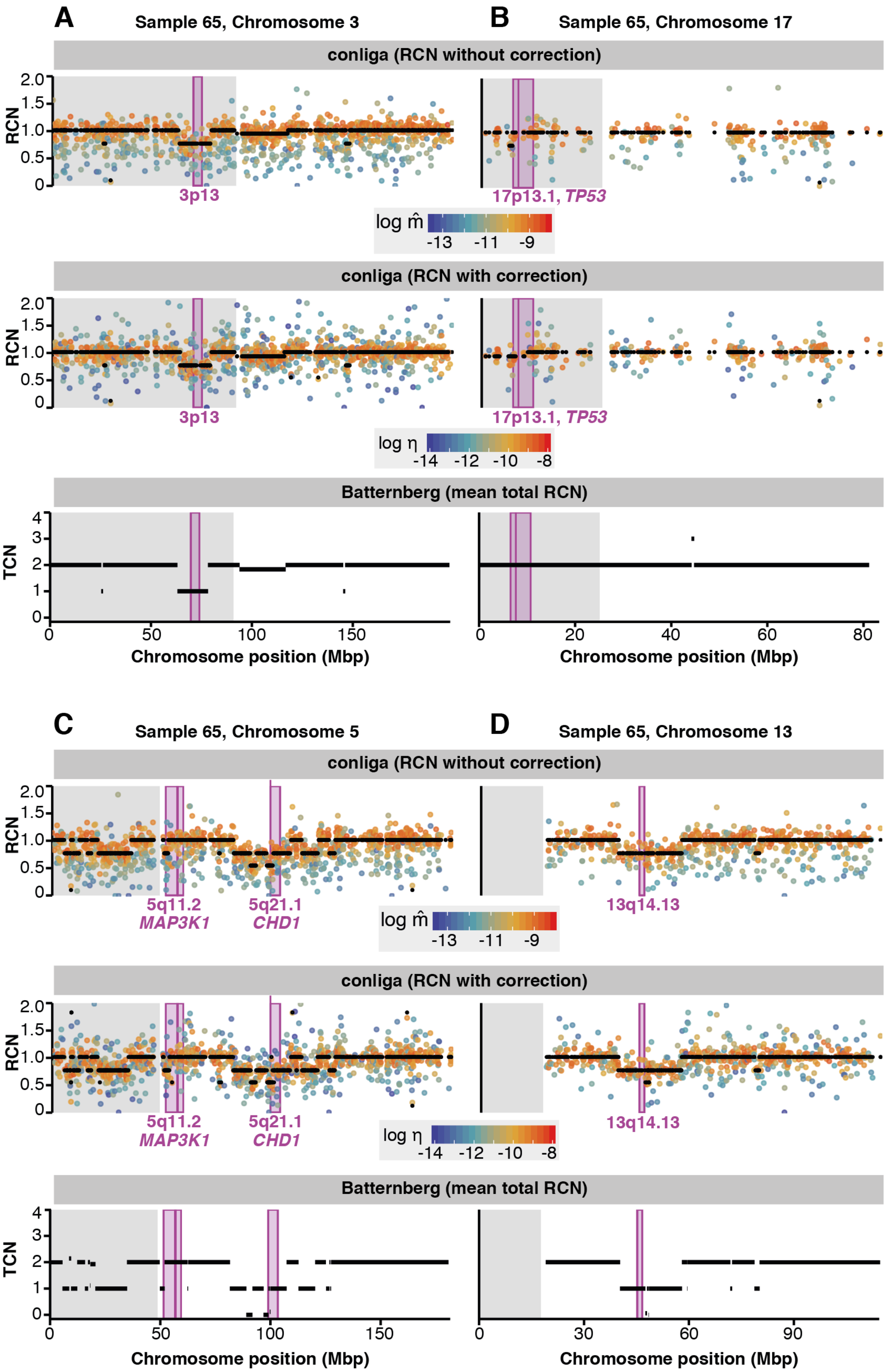

A
Sample 65, Chromosome 3
B
Sample 65, Chromosome 17
conliga (RCN without correction)
RCN
2.0
1.5
1.0
0.5
0
3p13
17p13.1, TP53
log m̂
-13 -11 -9
conliga (RCN with correction)
3p13
17p13.1, TP53
log η
-14 -12 -10 -8
Batternberg (mean total RCN)
TCN
4
3
2
1
0
0 50 100 150
0 20 40 60 80
Chromosome position (Mbp)
C
Sample 65, Chromosome 5
D
Sample 65, Chromosome 13
conliga (RCN without correction)
5q11.2
MAP3K1
5q21.1
CHD1
13q14.13
log m̂
-13 -11 -9
conliga (RCN with correction)
5q11.2
MAP3K1
5q21.1
CHD1
13q14.13
log η
-14 -12 -10 -8
Batternberg (mean total RCN)
0 50 100 150
0 30 60 90
Chromosome position (Mbp)

Figure 4.23 Comparison of the weighted mean total copy number of the clones inferred by Batternberg against conliga inferred RCN profiles for prostate cancer sample 65 with and without bias correction in particular chromosomes. A: chromosome 3. B: chromosome 17. C: chromosome 5. D: chromosome 13. Top: conliga RCN inference without correction. Black points represent the inferred MAP RCN. Coloured points represent the scaled count for each locus coloured by the log of the inferred expected proportion of reads to align to each locus ($\log \hat{m}$). Middle: conliga RCN inference with correction. Black points represent the inferred MAP RCN. Coloured points represent the scaled count for each locus coloured by the log of the inferred expected proportion of reads to align to each locus for each sample taking into account sample-specific biases ($\log \eta$). Bottom: Batternberg mean TCN weighted by the inferred clone proportions. Regions of interest are highlighted in purple. Grey sections correspond to the p arm of the chromosome while white sections correspond to the q arm of the chromosome.

Batternberg and conliga infer a complex SCNA profile within chromosome 5 (figure 4.23C), including a homozygous deletion of *CHD1* which encodes chromodomain helicase DNA-binding protein 1. Prostate tumours with *CHD1* loss comprise a distinct subtype which typically include *SPOP* mutations but lack the commonly seen *TMPRSS2-ERG* fusions and *PTEN* deletions [266, 267]. The subtype is generally characterised by increased genomic instability. CHD1 has been shown to be involved in double-strand break (DSB) repair [268, 267]. Loss of *CHD1* impairs the homologous recombination (HR) repair mechanism leading to increased use of the error-prone non-homologous end joining (NHEJ) repair mechanism [268, 267]. Recent studies suggest that PARP inhibitors and DNA damaging agents (such as olaparib, carboplatin and mitomycin C) could be effective treatments for prostate cancers harbouring *CHD1* deletions [268, 267].

By performing bias correction, conliga is able to detect focal deletions that occur close to 13q14.13 which is also commonly deleted in prostate cancer [266]. Without bias correction, conliga was unable to detect this (figure 4.23A). Furthermore, after bias correction, conliga infers what appears to be a subclonal event involving deletion of *TP53*. Before correction, a focal deletion was inferred close to *TP53* (figure 4.23B). These alterations were not detected by Batternberg and could be either false positives inferred by conliga or represent true SCNAs.

The WGS sample associated with FAST-SeqS sample 70 was inferred to have low tumour purity, with a purity of 17% determined by Batternberg. Using the single-end pipeline and without bias correction, conliga fails to detect the deletions detected by Batternberg (figures 4.24 and 4.25). However, after using the new pipeline and by correcting for biases, conliga is able to detect the deletions in chromosome 1, 2 and 21 (figures 4.25A, 4.25B, and

4.25D). The loss involving chromosome 21q occurs in a commonly deleted region, which includes *ERG* and *TMPRSS2* [266].

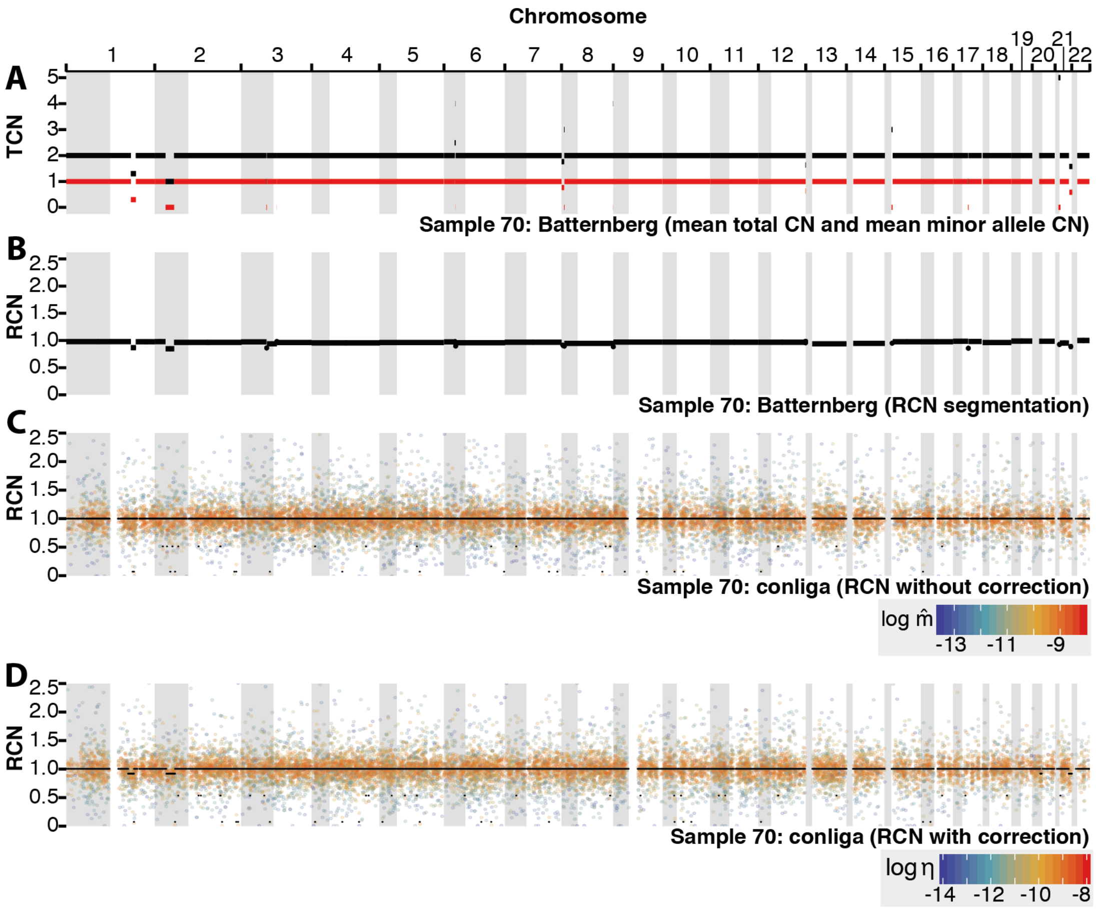


Figure 4.24 Comparison of Batternberg inferred copy number profiles (allowing for subclonality) and conliga inferred RCN profiles for prostate cancer sample 70 with and without bias correction. A: Inferred Batternberg profile showing the mean total copy number of the clones (black) and the mean minor allele copy number of the clones (red). B: The RCN segmented profile called by Batternberg prior to subclonal inference. C: RCN profile inferred by conliga without bias correction. D: RCN profile inferred by conliga with bias correction.

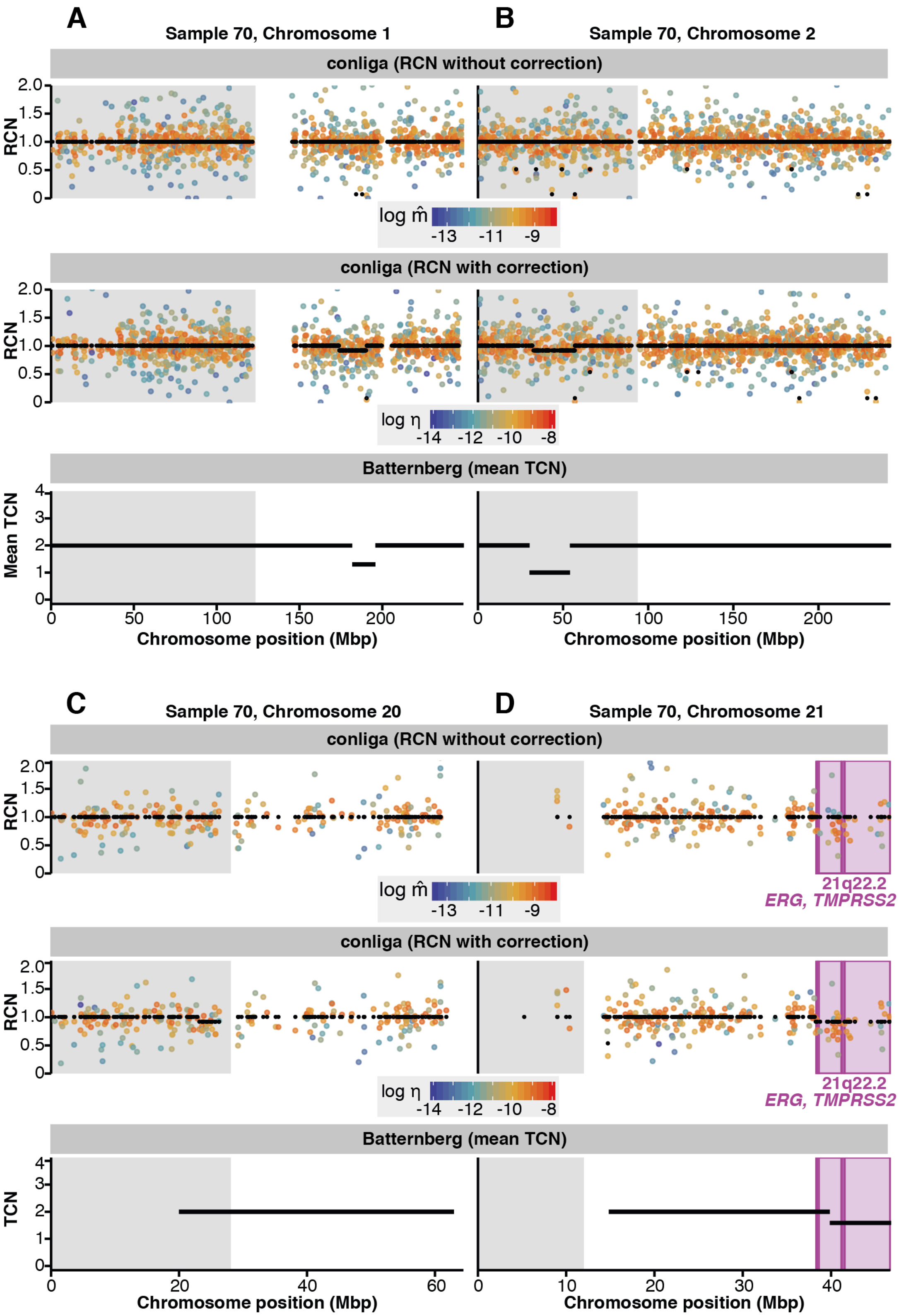

A
Sample 70, Chromosome 1
B
Sample 70, Chromosome 2
conliga (RCN without correction)
RCN
2.0
1.5
1.0
0.5
0
log m̂
-13 -11 -9
conliga (RCN with correction)
log η
-14 -12 -10 -8
Batternberg (mean TCN)
Mean TCN
4
3
2
1
0
0 50 100 150 200
Chromosome position (Mbp)
C
Sample 70, Chromosome 20
D
Sample 70, Chromosome 21
21q22.2
ERG, TMPRSS2
TCN
0 20 40 60
0 10 20 30 40

Figure 4.25 Comparison of the weighted mean total copy number of the clones inferred by Batternberg against conliga inferred RCN profiles for prostate cancer sample 70 with and without bias correction in particular chromosomes. A: chromosome 1. B: chromosome 2. C: chromosome 20. D: chromosome 21. Top: conliga RCN inference without correction. Black points represent the inferred MAP RCN. Coloured points represent the scaled count for each locus coloured by the log of the inferred expected proportion of reads to align to each locus ($\log \hat{m}$). Middle: conliga RCN inference with correction. Black points represent the inferred MAP RCN. Coloured points represent the scaled count for each locus coloured by the log of the inferred expected proportion of reads to align to each locus for each sample taking into account sample-specific biases ($\log \eta$). Bottom: Batternberg mean TCN weighted by the inferred clone proportions. Regions of interest are highlighted in purple. Grey sections correspond to the p arm of the chromosome while white sections correspond to the q arm of the chromosome.

4.7 Discussion and Summary

This chapter details the process of carrying out our own FAST-SeqS experiments within the Tavaré laboratory applied (primarily to) prostate samples. The first challenge was to adapt the primers to utilise higher-throughput Illumina machines such as the Illumina HiSeq 4000. Relative to the MiSeq, these machines tend to perform poorly in sequencing low-diversity libraries such as ours. By incorporating UMIs with varying lengths, I appeared to mitigate this issue without having to spike proportionally more PhiX control and reduce the cost-effectiveness of the assay. Further experiments should be done to explore the possibility of reducing the proportion of PhiX control used and perhaps removing its use altogether. This experiment showed that we were able to successfully scale-up the sequencing of FAST-SeqS samples, reducing the cost per sample. Hence, the experiment provides further evidence that FAST-SeqS and conliga could be used in clinical and research applications.

By including unique dual indices (UDIs), I was able to explore the problem of index swapping (also known as index hopping or index cross-talk). These results suggest that it is important to include UDIs in FAST-SeqS experiments. We saw that index cross-talk occurs at a relative high rate with 3.8% of the reads being assigned to invalid index combinations. The results show that there is a background level of cross-talk which could have various causes, one of which may be index hopping occurring on the flow-cell. We saw that index cross-talk occurs more frequently in a pattern consistent with the order of pipetting and with the order of the sample number. Further experiments are needed to determine the origin of this issue and also to explore ways to reduce the level of index cross-talk.

An exploration of the data from the new paired-end and UMI aware pipeline, showed that BWA-MEM appears to be more appropriate for the alignment of FAST-SeqS data when compared with Bowtie 2. By using the UMIs, we can calculate the proportion of unique molecules observed at each locus. This appears to be a useful metric in identifying loci that see many false positive alignments. My current strategy to filter low-quality loci, particularly selecting only those that have a non-zero count in all control samples, will lead to over-filtering. Those loci that were not observed purely because of sampling error will be unnecessarily excluded. Further work is needed in this area to use the features of the data to select the most appropriate set of loci to use.

The major finding of this chapter is that sample-specific biases were present in these data and they can result in the inference of spurious RCN profiles. These effects were seen in previous data sets but at seemingly lower frequency. Moreover, they tended not to affect the RCN inference. This may be because the sample-specific biases were not as strong in the OAC and Barrett's oesophagus data sets as they are in the prostate data set. In part, it may also be due to obtaining fewer reads per sample and the associated reduction in statistical strength meant that these biases did not result in an inferred RCN signal.

My attempts to use a factor model to learn the biases from data proved to be unsuccessful. Although, as I noted, by using more samples and modifying the model, this might be possible. Two major causes of sample-specific bias were discovered; a bias associated with large-scale genomic GC content (which I referred to as smoothed GC) and an amplification bias which saw differential amplification of lowly and highly amplified loci between samples. By using a 2-dimensional B-spline and making modifications to the conliga model, I was able to successfully correct for these biases. By jointly inferring the sample-specific effects of GC and amplification bias, control samples were inferred to have flat RCN profiles. True SCNAs did not appear to be regressed away by the bias correction and few, if any, RCN artifacts were introduced by correcting for biases. In the case of a low purity sample, the correction of biases led to seemingly increased sensitivity, with RCN profiles obtained that were similar to those inferred by Batternberg. Without correction, flat RCN profiles were inferred. Of course, Batternberg corrects for GC bias also and as such, both methods may be introducing similar spurious calls by doing so. However, this explanation seems less plausible.

In a few cases (not shown), the updated model was unable to remove sample-specific biases from the data. This was seen in samples obtained from blood but also in a few obtained from prostate tissue. This suggests that the model is not perfect and further work is needed to identify and correct all the sources of sample-specific biases. However, more importantly, further work should be done to find out where and how these sample-specific biases are introduced. Where possible, effort should be made to eliminate or substantially reduce these sources of bias.

I used a 2-dimensional B-spline which could correct for interactions between GC and amplification biases. In part, I did this because it was straightforward to implement, using code I had developed for the factor model, and that an interaction effect between GC and amplification bias seemed plausible. However, I did not test the use of two independent B-splines to correct for the effects. This model would result in considerably fewer variables and, if sufficient, would be preferable. Furthermore, I did not extensively test the number of knots for the B-splines. Increasing or decreasing the number of knots would result in a more flexible or inflexible spline, respectively. It is conceivable that improvements could be made by altering the placement and the number of knots. Moreover, the use of a penalised B-spline may be more appropriate to avoid over-fitting and regressing away true SCNAs.

While B-splines are convenient, it may be possible to use less-flexible parametric shapes instead. Looking at samples 45 and 50 in Figure 4.16, potentially the amplification bias could be modelled by a sigmoid or logistic function. This would help to constrain the model and reduce the number of variables.

In the next chapter, I explore the potential use of SNP data within the FAST-SeqS amplicons. I focus on the use of FAST-SeqS and conliga as a tool to screen and estimate the purity of samples, prior to sequencing them with more expensive methods such as WGS.

4.8 Methods

4.8.1 Preparation of FAST-SeqS libraries

The FAST-SeqS libraries were prepared by Sarah Field from the Tavaré laboratory at the Cancer Research UK Cambridge Institute. Libraries were created similarly to those previously described.

The first round PCR reaction (PCR1) was performed using 1 µl DNA (50 ng/µl), 25 µl Phusion HiFi Master Mix (ThermoFisher F531L), 0.5 µl PCR1 Forward Primer (10 µM, Sigma), 0.5 µl PCR1 Reverse Primer (10 µM, Sigma), 23 µl nuclease free water (NFW).

The cycling conditions for PCR1 were 98 °C for 2 minutes followed by 2 cycles of: 98 °C for 10 seconds, 57 °C for 2 minutes and 72 °C for 2 minutes. This was followed by 4 °C until the samples underwent bead clean-up.

Following PCR1, 1.4X AMPure bead clean-up was performed for each sample as follows. 70 µl Agencourt AMPure XP beads (Beckman Coulter A63881) were added to the mix and samples were incubated at room temperature for 2 minutes, placed on a magnetic rack until clear. Supernatant was removed and each sample was washed two times with 80% ETOH. Finally, 10 µl NFW was eluted into purified PCR1 product for each sample.

PCR2 was performed using 10 µl of PCR1 product, 25 µl Phusion HiFi Master Mix (ThermoFisher F531L), 0.5 µl PCR2 Forward Primer (10 µM, Sigma), 0.5 µl PCR2 Reverse Primer (10 µM and 14 µl nuclease free water (NFW).

The cycling conditions for PCR2 were 98 °C for 2 minutes followed by 13 cycles of, 98 °C for 10 seconds, 57 °C for 2 minutes and 72 °C for 2 minutes. This was followed by 4 °C until the samples were cleaned-up. Each PCR2 product (sample) underwent a 1X AMPure bead clean-up as follows. 50 µl Agencourt AMPure XP beads (Beckman Coulter A63881) were added to the mix, incubated at room temperature for 2 minutes and placed on a magnetic rack until clear. Supernatant was removed and the PCR2 product was washed two times with 80% ETOH. Finally, 10 µl were eluted into purified PCR2 product for each sample. Finally, purified products were quantified using the Qubit High sensitivity protocol (ThermoFisher Q32854) and quality control was performed using BioAnalyser using DNA1000 chip (Agilent). PCR2 products were then diluted accordingly such that they could be pooled in equimolar ratios. After pooling, libraries were sequenced with 30% PhiX spike-in on an Illumina HiSeq 4000 lane.

Chapter 5

Towards allele-specific and single cell copy number inference using repeat DNA

The first half of this chapter outlines more recent work regarding the exploration of SNPs in FAST-SeqS data and its implications for inferring allele-specific and subclonal copy number. I show that by using SNP frequencies as an independent source of data, the purity and ploidy of a sample can be estimated and conliga's RCN calls can be converted to absolute and allele-specific copy number. I then discuss future work and the considerations for integrating SNP data into the conliga model.

The second half of the chapter focuses on the development of single-cell FAST-SeqS (scFAST-SeqS). Inferring subclonal populations from bulk FAST-SeqS data may be possible but is likely to be limited to few clones. By performing scFAST-SeqS, we may be able to gain more insight into the subclones present in tumour populations. I discuss the experimental protocols which were developed in collaboration with Sarah Field of the Tavaré laboratory at the Cancer Research UK Cambridge Institute. I go on to explore data from scFAST-SeqS pilot experiments using flow-sorted peripheral blood mononuclear cells (PBMCs). Finally, I briefly discuss further work to extend the conliga model to infer the copy number profiles of single cells.

5.1 Somatic and germline variants in FAST-SeqS data

In the previous chapter, I showed the presence of mismatches and indels (with respect to the reference genome) in FAST-SeqS data. The vast majority of these mismatches and indels will be introduced by errors and do not represent true variants. Errors are introduced by the process of PCR, sequencing, read processing and alignment. A minority of the mismatches and indels present in FAST-SeqS data are genomic variants that may be either germline or somatic in origin. We may assume that most of the mismatches and indels that are present in low frequencies in a normal sample are errors though some may represent somatic mutations occuring in a minority of cells. Those that are present in approximately 50% and 100% of reads aligned to particular locus may be true variants. Most likely, these will be germline variants that are also present in the general population; single nucleotide polymorphisms (SNPs). Indeed, the presence of SNPs in FAST-SeqS data has been noted by Douville *et al.* also [107]. By using WGS samples from the 1,000 Genomes project [269], they identified 26,220 genomic positions within FAST-SeqS amplicons that had variants in $\geq 1\%$ of the samples [107] and therefore met the definition of a SNP. Douville *et al.* used these SNPs to detect evidence of allelic imbalance at the chromosome arm-level [107].

As I discussed in the introduction, heterozygous variants present in the germline can be informative for the inference of SCNAs. The proportion of reads that include a heterozygous variant can be used to 1) infer the number of copies of each allele in tumour cells (allele-specific copy number) and, 2) the proportion of tumour and normal cells in the sample. Also, they may help to identify subclonal tumour populations, which we discuss later.

5.1.1 The distribution of indels and mismatches

Figure 5.1 shows the distribution of indels and mismatches present in a normal prostate sample. Here, I am using BWA-MEM aligned reads that were processed by the pipeline described in the previous chapter. Alignments with a mapping quality below 50 were excluded as were indels and mismatches with a Phred quality score below 20. A total of 93,309 mismatches and indels are found within the autosomes of this sample. Of these, 83,174 are observed at a frequency of less than 0.25 and are likely to be errors. 772 were present at a frequency of 1 and have a least 10 reads supporting them. These are likely to be homozygous non-reference variants but may additionally include mixture of alignment errors and artifacts of an imperfect reference genome. There are 3,237 mismatches and indels observed with a frequency of 0.25 to 0.75. Some of these will be somatic and germline variants and we may assume the majority are heterozygous SNPs. Some of these mismatches and indels may include PCR errors and other artifacts.

Of the 3,237 mismatches and indels observed with frequencies between 0.25 and 0.75, 85.54% (2,769) fall within the 90% confidence interval of a binomial distribution; Binomial$(n, 0.5)$, where n is the total number of reads supporting the position, excluding ambiguous base calls and bases with a Phred quality score below 20. This suggests that there is additional variation in the data beyond merely sampling error. Additional variability may be introduced

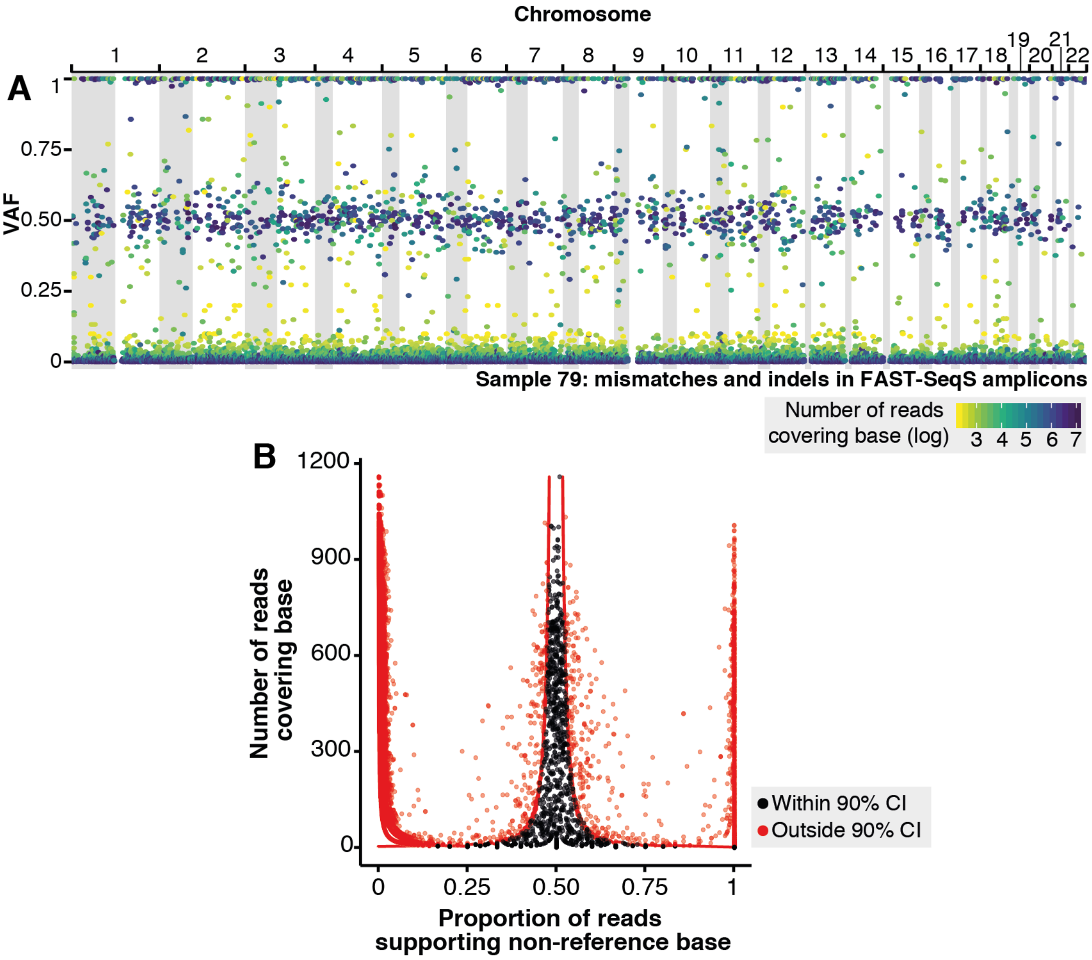


Figure 5.1 Distribution of mismatches and indels to the reference genome seen in aligned FAST-SeqS amplicons for normal prostate sample 79. A: Proportion of indels and mismatches at each genomic position plotted across autosomes. The colour of each point represents the number of reads covering the genomic position. B: The proportion of reads supporting the indels or mismatches against the number of reads covering the genomic position.

by the two-step PCR process or downstream processing. Also, the data may not be binomially distributed because there may be errors mixed with the true variants.

In section 5.3, I consider the implications of errors and SNP distributions in order to integrate SNP data into the conliga model to infer copy number from FAST-SeqS data.

5.2 Using mismatches and indels as independent supplementary information

A natural extension of the conliga model would be to integrate SNP data into the model and infer allele-specific copy number and potentially subclonal information. However, even in conliga's current form, these data can be useful as supplementary information to the relative copy number calls. Using inductive reasoning, we may be able to convert the RCN calls to absolute copy number and allele-specific copy number calls. In doing so, we will be able estimate the tumour purity and average ploidy of the sample.

From equation 1.6, we know that the relative copy number of a locus, l, is given by:

$$\hat{c}_l = \frac{2(1-p) + t_l \cdot p}{d}$$

where p represents the purity of the tumour, t_l represents the total copy number (where $t_l = n_{al} + n_{bl}$) assuming a sample consisting of one clone plus normal contamination $(1-p)$, and d represents the sample ploidy ($d = \sum_l \hat{m}_l \cdot c_l$). Therefore we have,

$$k = \frac{\hat{c}_l}{\hat{c}_{l'}} = \frac{2(1-p) + t_l \cdot p}{2(1-p) + t_{l'} \cdot p} \tag{5.1}$$

where l' represents a locus that is not l. By rearranging for p, we obtain:

$$p = \frac{2-2k}{k(t_{l'}-2) - t_l + 2} \tag{5.2}$$

where t_l and $t_{l'}$ represent the absolute copy number of the tumour population for locus l and l' respectively.

Similarly, we can estimate the sample ploidy, d,

$$d = \frac{4 - 4p + p(t_l + t_{l'})}{\hat{c}_l + \hat{c}_{l'}} \tag{5.3}$$

As such, if we are able to make an informed estimate of the absolute copy number for two loci, then we can use equations 5.2 and 5.3 as a simple method to estimate the tumour purity and sample ploidy.

Figure 5.2A shows the inferred RCN profile for a prostate cancer sample (sample 67). We know that, even copy numbers (i.e $t_l \in \{0, 2, 4, \dots\}$) are those that can have equal copies of each allele, whereas odd copy numbers cannot. Therefore, copy number states that have indel and mismatch proportions centred around 0.5 are likely to represent even copy numbers. Figure 5.2C shows the distribution of mismatch and indel proportions associated with each copy number state. In this case, the copy number state which is associated with the most loci is the only state that appears to have proportions around 0.5. Let's assume that this state represents a copy number of 2 in the tumour population, while the two other major states represent copy numbers 1 and 3. By using equations 5.2 and 5.3 we can estimate that p is 0.559 and d is 1.998. By plugging these estimates in to equation 1.8 and for every possible combination of major and minor alleles, we obtain expected B allele frequencies. We see that B allele frequencies align with the observed indel and mismatch proportions and suggests that this solution fits the data well (figures 5.2C and D).

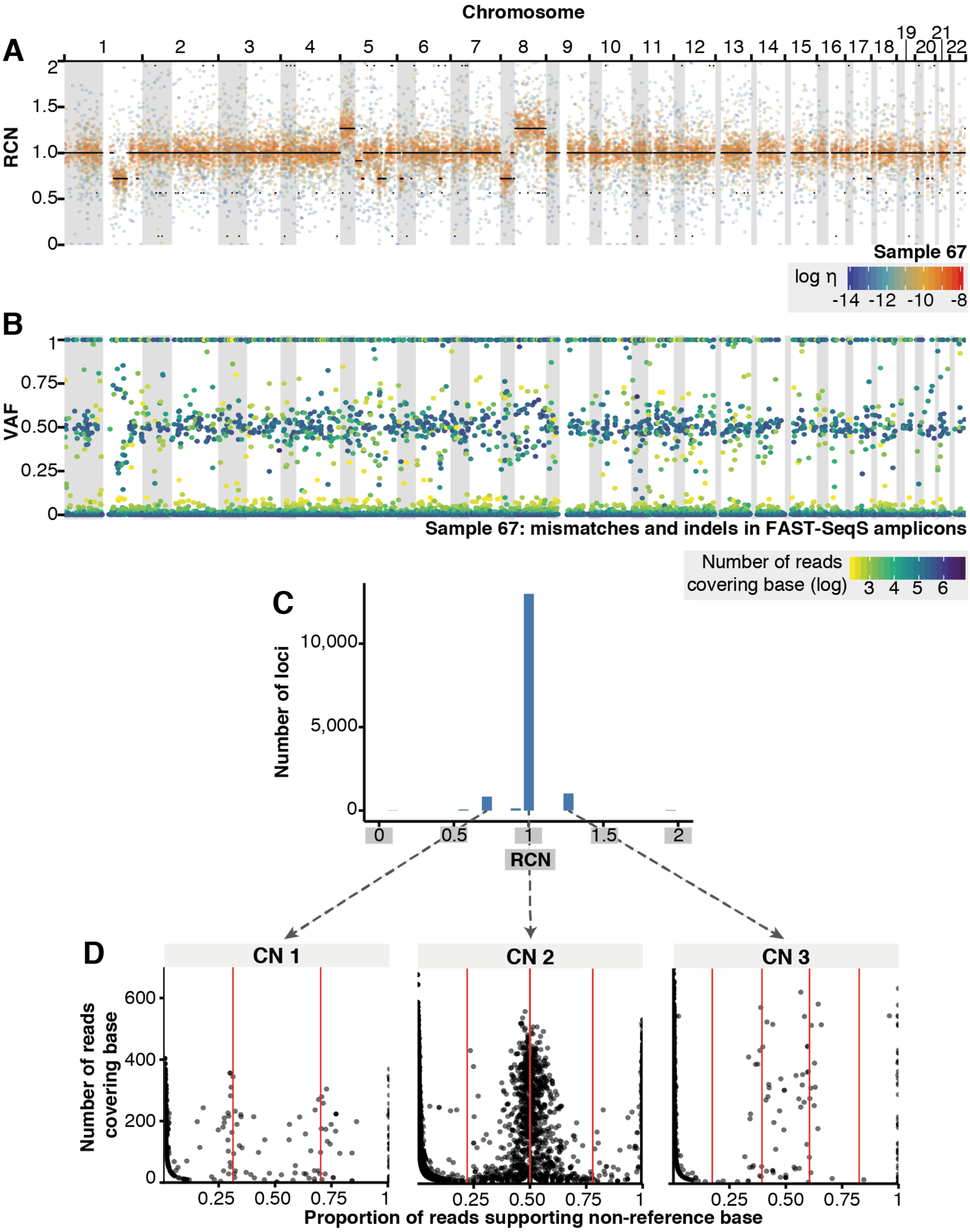

Chromosome
A
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
RCN
2
1.5
1.0
0.5
0
Sample 67
log η
-14
-12
-10
-8
B
VAF
1
0.75
0.50
0.25
0
Sample 67: mismatches and indels in FAST-SeqS amplicons
Number of reads
covering base (log)
3
4
5
6
C
Number of loci
10,000
5,000
0
0
0.5
1
1.5
2
RCN
D
CN 1
CN 2
CN 3
Number of reads
covering base
600
400
200
0
0.25
0.50
0.75
1
Proportion of reads supporting non-reference base

Figure 5.2 The inference of purity and ploidy for prostate cancer sample 67. A: Inferred SCNA profile using conliga's *MAP* state method. Black points represent the inferred MAP RCN for the loci's MAP state. Coloured points represent the scaled count for each locus coloured by the log of the inferred expected proportion of reads to align to each locus for the sample taking into account sample-specific biases ($\log \eta$). B: Mismatch and indel frequencies across every base in FAST-SeqS amplicons. Coloured points represent the scaled count for each locus coloured by the number of reads covering the base (log). C: Histogram of inferred FAST-SeqS loci RCN. D: Indel and mismatch frequencies in FAST-SeqS amplicons plotted against the number of reads supporting the base position per relabelled copy number state. CN represents our inferred absolute copy number for each relabelled copy number state.

Figure 5.3A shows the inferred RCN profile for a OAC sample (OAC4). Note that this sample was processed using the original pipeline, which uses Bowtie 1 (see section 2.4). As such, indels will not be present in the data. As I discussed in chapter 3 and section 3.1.7, a catastrophic genome-doubling event often occurs early in development. A potential solution in which the copy number state that has SNP proportions around 0.5 is assumed to represent copy number 4 in the tumour population is shown in figures 5.3C and 5.3D. By calculating the purity and ploidy, the B allele frequencies align with the observed SNP frequencies suggesting that the solution is probable. These results suggest that much of the tumour genome has undergone copy neutral LOH, perhaps following an initial genome-doubling event. In addition, the results suggest evidence of subclonal alterations affecting chromosome 10.

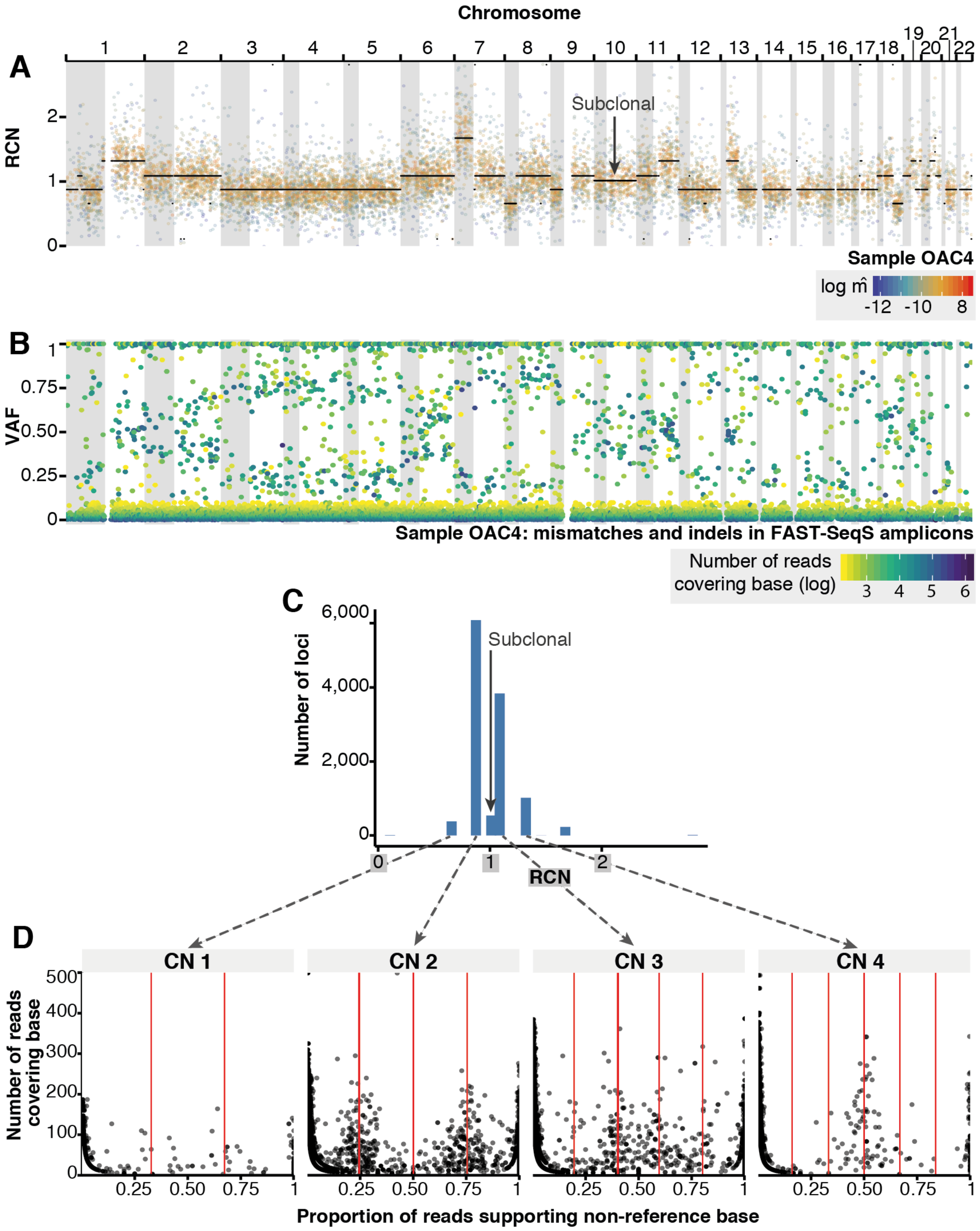

Chromosome
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
A
RCN
0
1
2
Subclonal
Sample OAC4
log m̂
-12 -10 8
B
VAF
0
0.25
0.50
0.75
1
Sample OAC4: mismatches and indels in FAST-SeqS amplicons
Number of reads covering base (log)
3 4 5 6
C
Number of loci
0
2,000
4,000
6,000
Subclonal
0 1 2
RCN
D
CN 1
CN 2
CN 3
CN 4
Number of reads covering base
0
100
200
300
400
500
0.25 0.50 0.75 1
Proportion of reads supporting non-reference base

Figure 5.3 The inference of purity and ploidy for oesophageal sample OAC4. A: Inferred SCNA profile using conliga's *MAP* state method. Black points represent the inferred MAP RCN for the loci's MAP state. Coloured points represent the scaled count for each locus coloured by the log of the inferred expected proportion of reads to align to each locus ($\log \hat{m}$). The arrow indicates a potential subclonal copy number state. B: Mismatch and indel frequencies across every base in FAST-SeqS amplicons. Coloured points represent the scaled count for each locus coloured by the number of reads covering the base (log). C: Histogram of inferred FAST-SeqS loci RCN. The arrow indicates a potential subclonal copy number state. D: Indel and mismatch frequencies in FAST-SeqS amplicons plotted against the number of reads supporting the base position per relabelled copy number state. CN represents our inferred absolute copy number for each relabelled copy number state.

5.2.1 Estimating purity for low purity samples

At low tumour purity, it is difficult to notice differences in B allele frequencies associated with different copy number states (by inspection). By itself, this information tells us that the sample is likely to be low purity. However, we can make assumptions without the use of B allele frequencies to estimate the purity and ploidy. For example, conliga might infer RCN profiles for prostate cancer samples that are predominantly flat but also includes a copy number state with slightly lower RCN (excluding states that appear to represent outliers or CNVs). In these cases, we might assume that the majority of the tumour genome is copy number 2 and the state with lower RCN represents copy number 1 or 0. With this assumption, we can estimate purity and ploidy. Table 5.1 shows the ploidy and purity estimates using conliga's RCN profiles for several samples that were determined by a pathologist to be low purity.

Large-scale cancer genome studies like those carried out by ICGC and TCGA require a pathologist to initially screen and estimate the purity of several samples taken from each patient. This initial screening is performed so that resources are spent only on those samples with sufficient purity. Given the chosen mean coverage of WGS (typically 30X-50X), a certain tumour purity threshold is required so that the experiments are sufficiently powered to detect somatic mutations, e.g. SNVs and SCNAs. Pathologist estimates often overestimate the tumour purity [270]. After sequencing, some samples are estimated to have lower tumour purity than expected and are required to be excluded from tumour genome studies. Furthermore, rather than counting tumour nuclei, the time of a pathologist is better spent on other clinical and research activities. FAST-SeqS and conliga may provide a low-cost alternative to estimate tumour purity and select samples for more in depth sequencing (such as HC WGS).

Sample	Ploidy	Purity	Pathology
18	1.990	0.118	0.05
19	1.990	0.135	0.1
20	1.986	0.137	0.15
21	1.993	0.090	0.15
22	1.998	0.102	0.1
23	2	0	0.05
24	2	0	0
25	1.989	0.087	0.15
26	2	0	0.1
27	1.999	0.183	0.1
29	1.993	0.118	0.1
30	2	0	0.05
31	1.993	0.103	0.1
32	2.009	0.188	0.1
33	2.004	0.155	0.05
34	1.992	0.314	0.2
35	1.986	0.231	0.25
36	1.994	0.088	0.05
37	2	0	0
38	1.989	0.375	0.05
39	1.980	0.370	0.1
40	2	0	0.15
41	1.995	0.086	0.1

Table 5.1 Estimates of ploidy ('Ploidy') and purity ('Purity') for 23 low purity prostate cancer samples using conliga's RCN inference compared with purity estimates provided by a pathologist ('Pathology'). Purity estimates are given as a proportion of tumour DNA within each sample. These estimates were made by assuming the most common state (i.e. the state mostly commonly assigned to loci) represents copy number 2, and that a state with lower RCN (chosen by inspection) represented copy number 1.

A study by Griffith *et al.* concluded that greater than 30-50X coverage is required to account for tumour heterogeneity, aneuploidy or for samples with high normal contamination [270]. 30-50X coverage is sufficient to detect SNVs in founding clones of tumour samples with high purity. However, they found that 30X is insufficient to infer clonal architecture and detect variants at $\leq 15\%$ variant allele frequency (VAF), even in tumour samples with at least 90% purity, and 50X provides insufficient coverage to discover variants at $< 10\%$ VAF [270]. By initially screening samples with FAST-SeqS and conliga, these factors could be estimated. Hence, researchers may be able to scale the sequencing resources to each sample individually, depending on the goals of the study.

5.3 Considerations for integrating SNP data to the conliga model

I have shown that B allele frequency data obtained from FAST-SeqS reads can be used as a supplementary source of information to translate conliga's RCN inference to allele-specific calls and estimate purity and ploidy. Instead, we could include the B allele frequencies in the model directly whereby they are emitted as observations by a copy number state. Therefore, we could jointly segment the genome using the B allele frequencies and read depth data. In doing so, we could achieve increased sensitivity and specificity in calling SCNAs as well as directly inferring absolute, allele-specific copy number, purity, ploidy and subclonal populations.

Integrating B-allele frequencies into conliga is planned future work. Further exploration of the BAF distributions would be required to sufficiently model the observations. We may wish to take a similar approach to Douville *et al.* [107] and use WGS data to create a set of known positions within FAST-SeqS amplicons that are SNPs. By doing so, we will filter out much of the PCR and sequencing errors. Alternatively, we could model the mismatch and indel emissions at every position within each FAST-SeqS amplicon. The emission distribution would need to adequately model the noise processes and homozygous SNPs. However, this approach would be computationally more expensive and if the distributional assumptions of the noise processes do not fit the data well, then spurious segmentation would be the result. As such, selecting the positions that are known to be SNPs would be the more pragmatic approach.

Currently, the pipeline does not use the UMI clusters when reporting indels and mismatches. The pipeline could be extended to find a consensus read for each UMI cluster. By doing so, this should remove some of the PCR and sequencing errors present in the data. Indeed, by using the number of clusters as n and p as the frequencies of mismatches and indels in the consensus reads, we may find that distribution of heterozygous SNPs is binomial.

There are several methods that attempt to resolve the subclonal architecture of tumour samples from bulk sequencing data [271–281], typically with the use of WGS, WES or deep targeted sequencing data. These methods can typically be divided into those that 1) use SNVs or 2) use SCNAs [278, 280]. However, some use both. The discussion of inferring subclonal architecture from FAST-SeqS data is beyond the scope of this thesis. However, FAST-SeqS data is more noisy than WGS and calling SNVs within FAST-SeqS amplicons is likely to be very challenging. As such, the inference of clonal architecture from FAST-SeqS data may be practically limited to the inference of few subclones.

5.4 Single-cell FAST-SeqS

5.4.1 Motivation

The low-cost and simplicity of the FAST-SeqS protocol makes its application to single cells appealing. Single-cell FAST-SeqS (scFAST-SeqS) may enable the low-cost SCNA profiling of hundreds or thousands of single cells. This would avoid the need to infer subclonal populations from bulk sequencing data, particularly using methods that use SCNAs as the basis of subclonal architecture inference.

5.4.2 scFAST-SeqS experiments

To investigate whether scFAST-SeqS could be used for this purpose, I designed an experiment in which we obtained PBMCs from two anonymous donors; one male donor and one female donor. These PBMCs were flow sorted into 384-well plates. Given that we were dealing with very low input DNA, it was likely that the number of cycles needed for PCR1 would need to be increased. Also, since the bead clean-up is the most technically challenging part of the protocol, we wanted to see if this could be replaced with a technically simpler solution. Therefore, we tested using ExoSAP to enzymatically purify PCR product between PCR1 and PCR2 reactions.

Our first attempt involved carrying out five rounds of PCR1, with ExoSAP clean-up and with bead clean-up. This was performed on 80 cells with the following experimental design:

1. 20 female PBMCs, 5 cycles of PCR1, ExoSAP clean-up
2. 20 male PBMCs, 5 cycles of PCR1, ExoSAP clean-up
3. 20 female PBMCs, 5 cycles of PCR1, Bead clean-up
4. 20 male PBMCs, 5 cycles of PCR1, Bead clean-up

The cells that were purified using the ExoSAP protocol produced reads that were composed almost exclusively of primer-dimer. The cells that had PCR1 product purified using beads produced reads originating from FAST-SeqS loci. However, relatively few loci appeared to be amplified per cell. These results suggest that the ExoSAP protocol degraded the PCR1 product in addition to the PCR1 primers and that we needed to increase the number of PCR1 cycles. Therefore, we attempted another pilot experiment with the following experimental design:

1. F10: 20 female PBMCs, 10 cycles of PCR1, Bead clean-up

2. M10: 20 male PBMCs, 10 cycles of PCR1, Bead clean-up
3. F15: 20 female PBMCs, 15 cycles of PCR1, Bead clean-up
4. M15: 20 male PBMCs, 15 cycles of PCR1, Bead clean-up

By increasing the cycles of PCR1, we were able to increase the number of FAST-SeqS loci amplified per cell. The results of this experiment are presented in this chapter.

5.4.3 Results

After demultiplexing the reads into samples, representing individual cells, I saw similar index swapping patterns and rates as those presented in chapter 4 (not shown). A few cells appeared to have very few reads associated with them, suggesting that the PCR reaction failed for these cells or that no cells were sorted into these wells. As such, cells with less than 100,000 reads were removed from the analysis. This left 18, 19, 19 and 20 cells for F10, M10, F15 and M15 respectively.

In chapter 4, I mentioned two methods to cluster UMIs to identify unique molecules from PCR1. The combined method worked by clustering pairs of 5' and 3' UMIs together. The separate method works in the same way but clusters each UMI separately. Therefore, reads are assigned a 5' UMI cluster and a 3' UMI cluster. In the protocols developed for single cell, more than two cycles of PCR1 were performed. This will result in UMIs being "swapped out" with further amplification after cycle two. In theory, clustering the UMIs separately could allow us to identify some cases in which the UMIs swapped. In practice, the separate clustering method failed. This is because we saw low-diversity of UMI sequences and as such, it appeared that UMIs were being clustered together erroneously. After investigation, it appears that by opting for primers that were purified using high performance liquid chromatography (HPLC), certain UMI sequences were preferentially purified [282]. As such, this reduced diversity of UMI sequences and led to erroneous clustering. This was not observed when clustering using the combined method.

The mean loci counts computed across cells from the same experimental condition could be considered to be analogous to a bulk sample (pseudo-bulk). Figure 5.4 shows the mean loci counts associated to each batch of cells (F10, M10, F15, M15). By comparing the mean loci counts across cells derived from males and females, we see a similar effect as we see in bulk samples; the raw counts or UMI counts scale proportionally with the number of copies of DNA. By comparing the mean counts to a previously sequenced normal bulk prostate sample (figure 5.5), we see that highly amplified loci in bulk samples tend to be highly amplified (on average) in single cells and vice-versa.

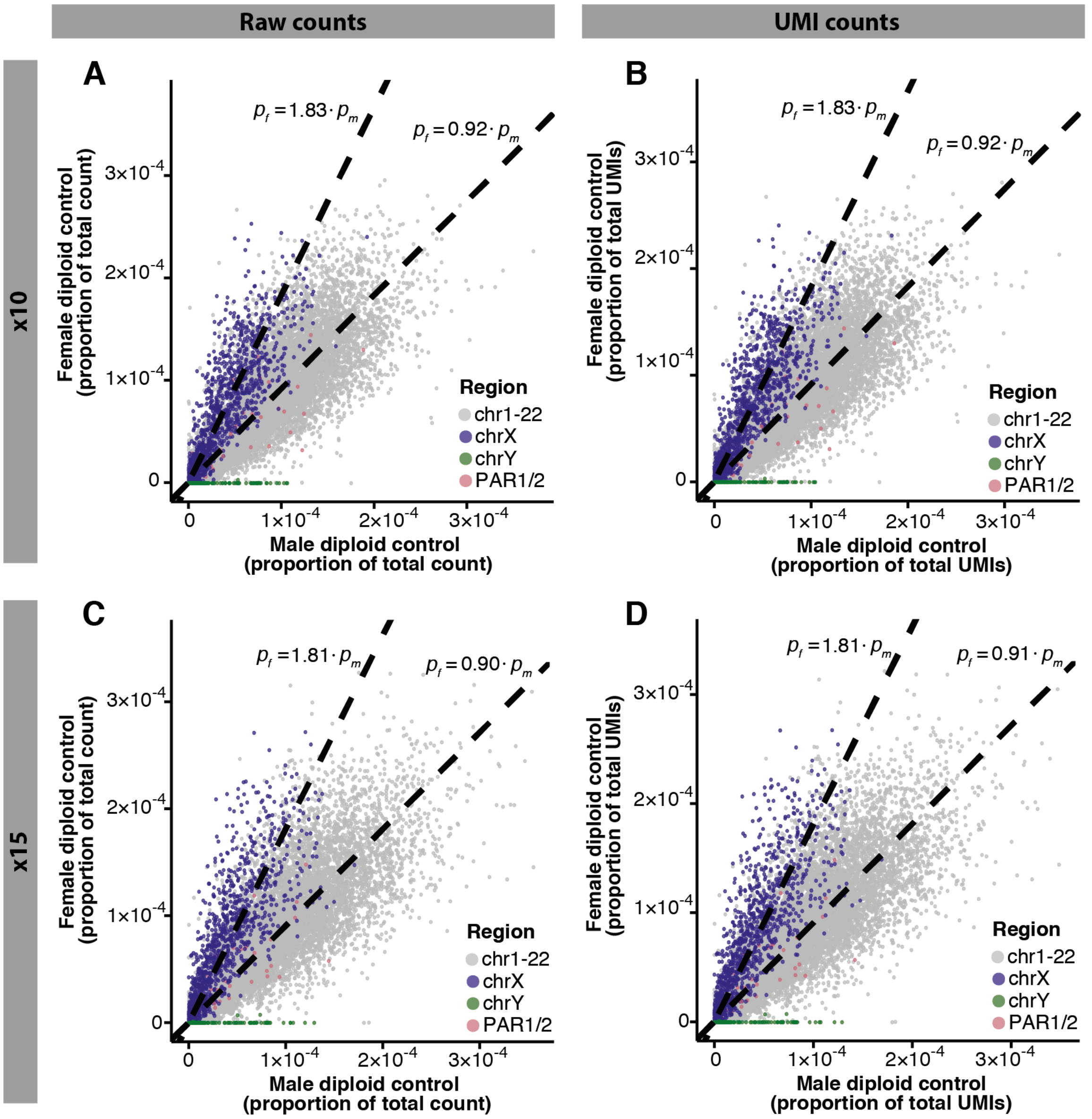


Figure 5.4 Comparison of male and female pseudo-bulk FAST-SeqS samples derived from scFAST-SeqS data. A: Raw count proportions from the PCR1 X10 experiment. B: UMI reduced count proportions from the PCR1 X10 experiment. C: Raw count proportions from the PCR1 X15 experiment. D: UMI reduced count proportions from the PCR1 X15 experiment.

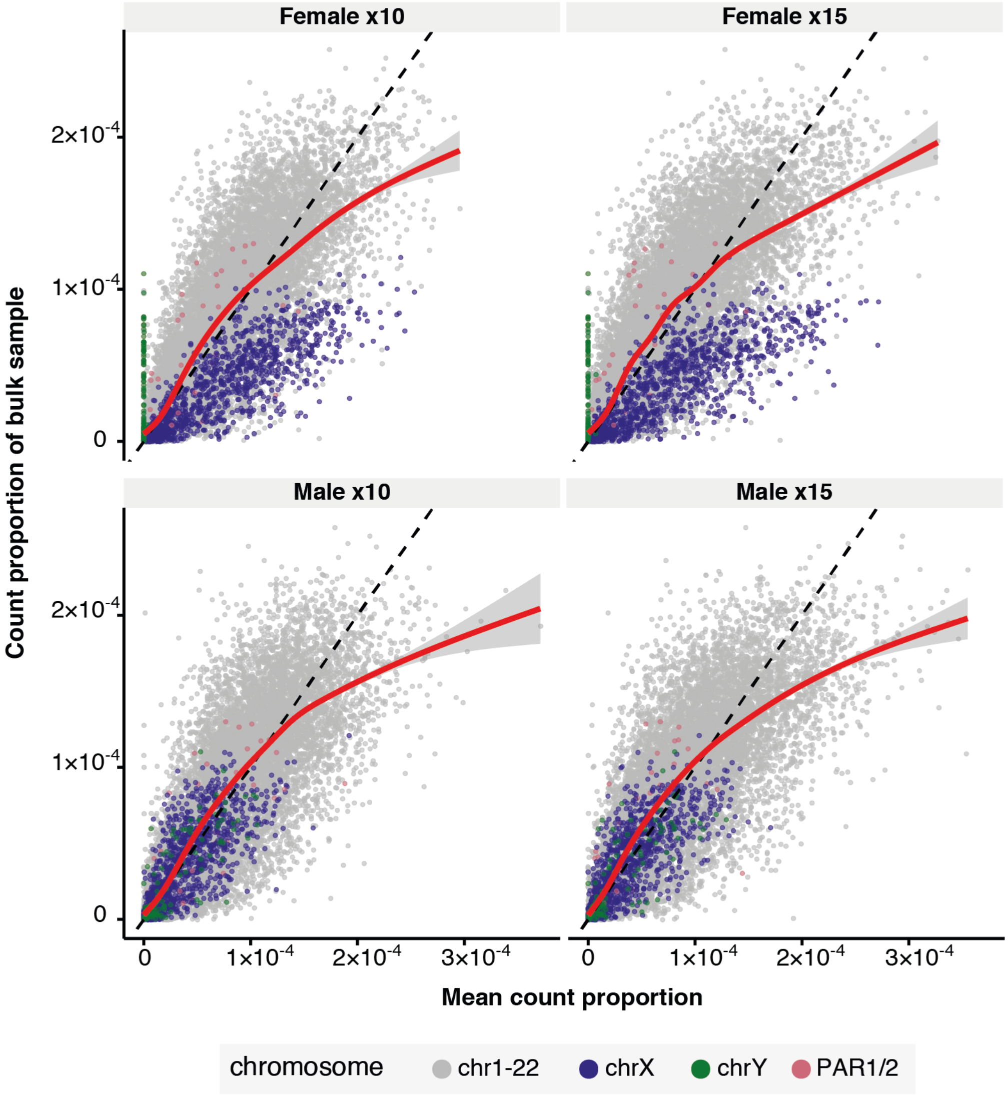


Figure 5.5 Count proportions of previously sequenced male bulk prostate sample compared with count proportions of scFAST-SeqS pseudo-bulk sample; top left: Female pseudo-bulk sample from PCR 10X experiment, bottom left: Male pseudo-bulk sample from PCR 10X experiment, top right: Female pseudo-bulk sample from PCR 15X experiment, bottom right: Male pseudo-bulk sample from PCR 15X experiment

Figure 5.6A shows the mean-variance relationship of the loci UMI counts across the experimental batches. By highlighting the proportion of cells that saw alignments at these positions, we see that there are few loci that obtain alignments across all cells. The F10 batch appears to have a greater proportion of non-zero counts at FAST-SeqS loci compared to the other batches. The reason for this is unclear. The mean-variance relationship at the single-cell level appears to be similar to the relationship seen in normal bulk samples (figure 2.4E) and is similar across experimental conditions and batches. Figure 5.6B shows how the number of DNA copies affects the mean-variance relationship. Using the PBMCs derived from males, we see that the loci counts associated with chromosome X and Y (one copy) appear to have greater variance than those with the autosomes and pseudoautosomal regions (two copies).

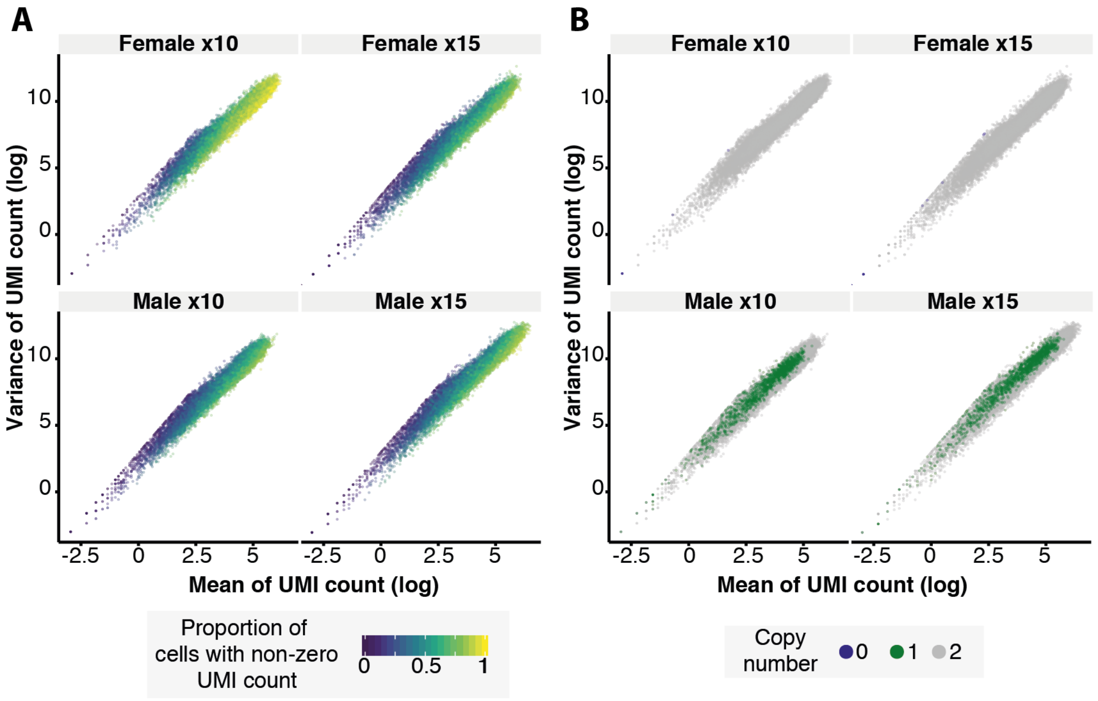


Figure 5.6 Mean UMI counts across the cells within an experimental condition plotted against the variance of the UMI counts. A: Highlighting the proportion of cells within the experimental batch that have a non-zero count at the FAST-SeqS locus. B: Highlighting the copy number of the FAST-SeqS locus.

Figure 5.6 shows that the mean loci count (computed across cells) is greatly influenced by the proportion of zero counts *across* cells. Figure 5.7 shows that the proportion of zero counts *within* cells is related to the total UMI count of the cells; the greater the total UMI counts, the fewer zeros are observed. This suggests that the average efficiency of PCR1 across loci appears to vary between cells. The results from the mean-variance analysis suggest that the proportion of zero counts at a locus across cells is related to the number of copies of the locus (figure 5.6). By comparing the proportion of zero counts in different chromosomes for each cell, we see that the number of copies affects the proportion of zeros seen; cells from the male donor have a greater number of zeros in chromosomes X and Y compared to the autosomes (figure 5.7).

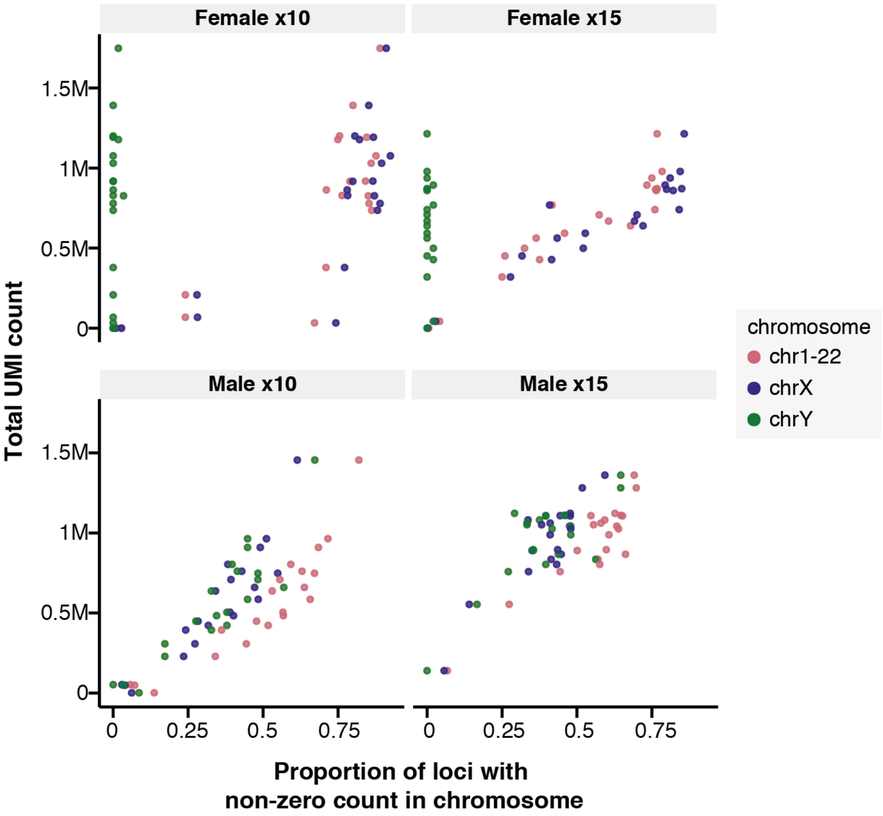


Figure 5.7 Proportion of loci within each cell that have a non-zero count within a region plotted against the total number of UMI counts from the cell. Each point represents data from chromosomes 1-22 (red), chromosome X (blue) or chromosome Y (green) from a single cell.

As I have shown, we see a similar RCN signal from the mean loci counts across cells (pseudo-bulk). However, do we see this signal within individual cells? Figure 5.8 shows the proportion of UMI counts associated with different chromosomes. Here, we see that cells derived from a male donor tend to have approximately half the proportion of counts associated with chromosome X than those cells derived from the female donor (figure 5.8 top-right). As we have also seen in bulk samples, this results in proportionally more counts in the autosomes for cells derived from the male donor (figure 5.8 top-left). Cells derived from female cells see few counts associated with chromosome Y, as expected (figure 5.8 bottom-left). The PAR1 and PAR2 regions have few loci within them and we expect the count proportions within these regions to be more variable. This is why we do not see (as clearly) the reduced proportion of reads in cells derived from females compared with males within the PAR1/2 regions as we see within chromosomes 1-22 (figure 5.8 bottom-right).

We see that the UMI count proportions from individual cells correlate with the mean UMI count proportions across cells (figures 5.9A and 5.10A). Furthermore, we see that the slope

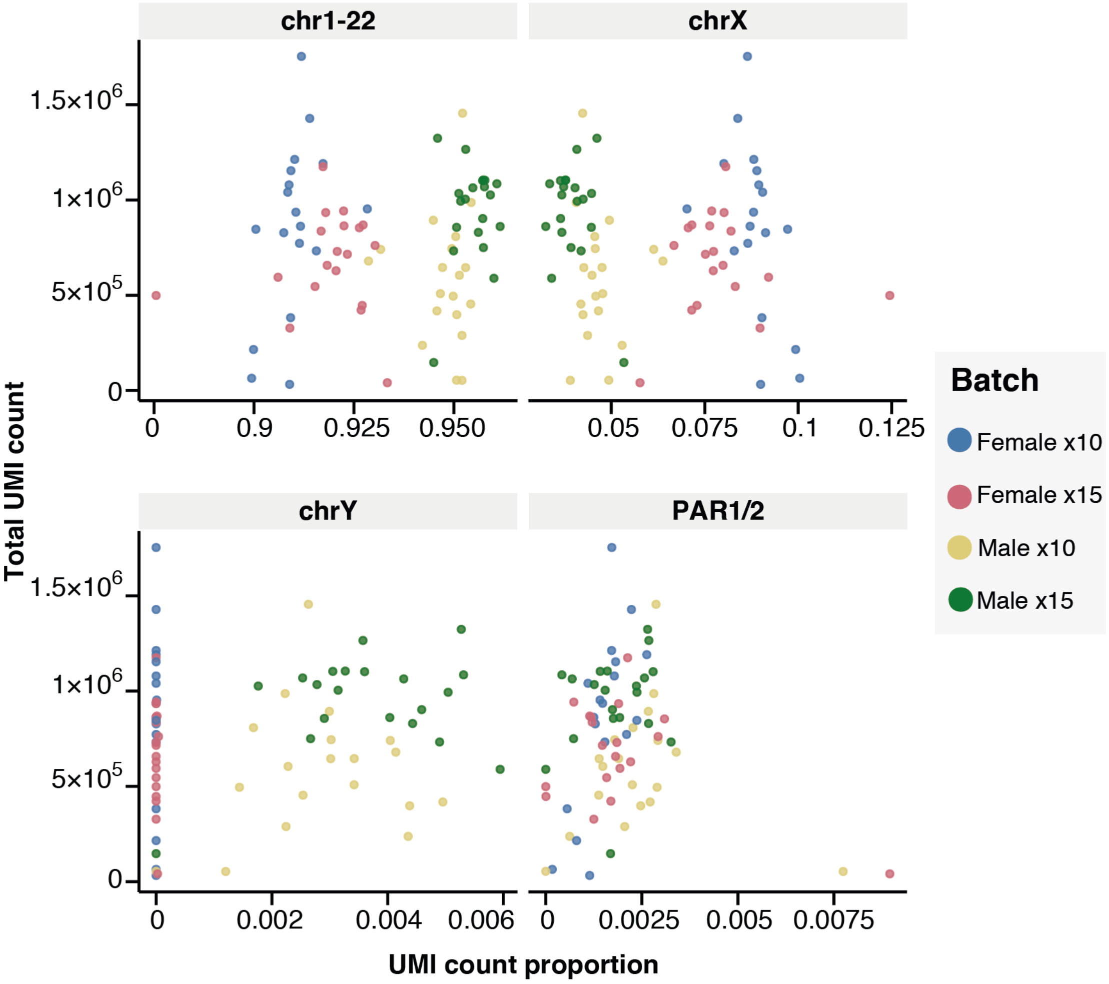


Figure 5.8 Proportion of UMI counts from each region of the genome (chromosomes 1-22, chromosome X, chromosome Y and PAR1/2) plotted against the total UMI count of the cell and coloured by the experimental condition. Each point represents data from a single cell.

of the relationship varies depending on the number of zero counts the cell has (figures 5.9 and 5.10). Intuitively this is as we would expect; cells with greater zero counts share counts across fewer loci and therefore see greater count proportions. Intriguingly, when inspecting the UMI count proportions and their distribution (figures 5.9 and 5.10) we see that in some cells, there appear to be several modes. These modes are more clearly present in the cells that have greater zero counts. These modes may represent relatively few initial molecules that were amplified in the initial cycles of PCR1 from these cells.

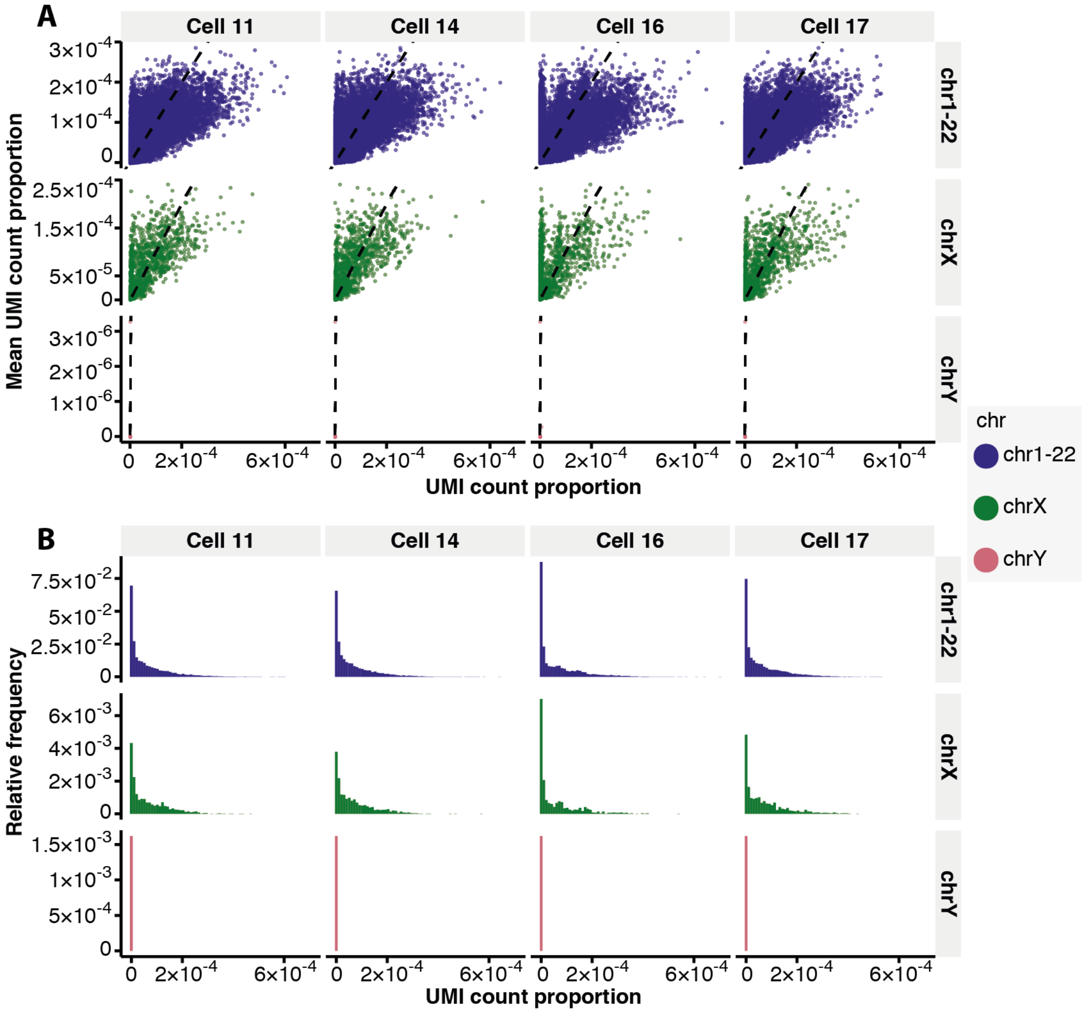


Figure 5.9 Cell-level UMI count exploration for four single cells from a female sample. A: Exploring the relationship between a cell's UMI count proportion against the pseudo-bulk UMI count proportion (in which the pseudo-bulk UMI count is the mean UMI count across all cells in the female x10 experimental condition). B: The distribution of UMI count proportions within each cell.

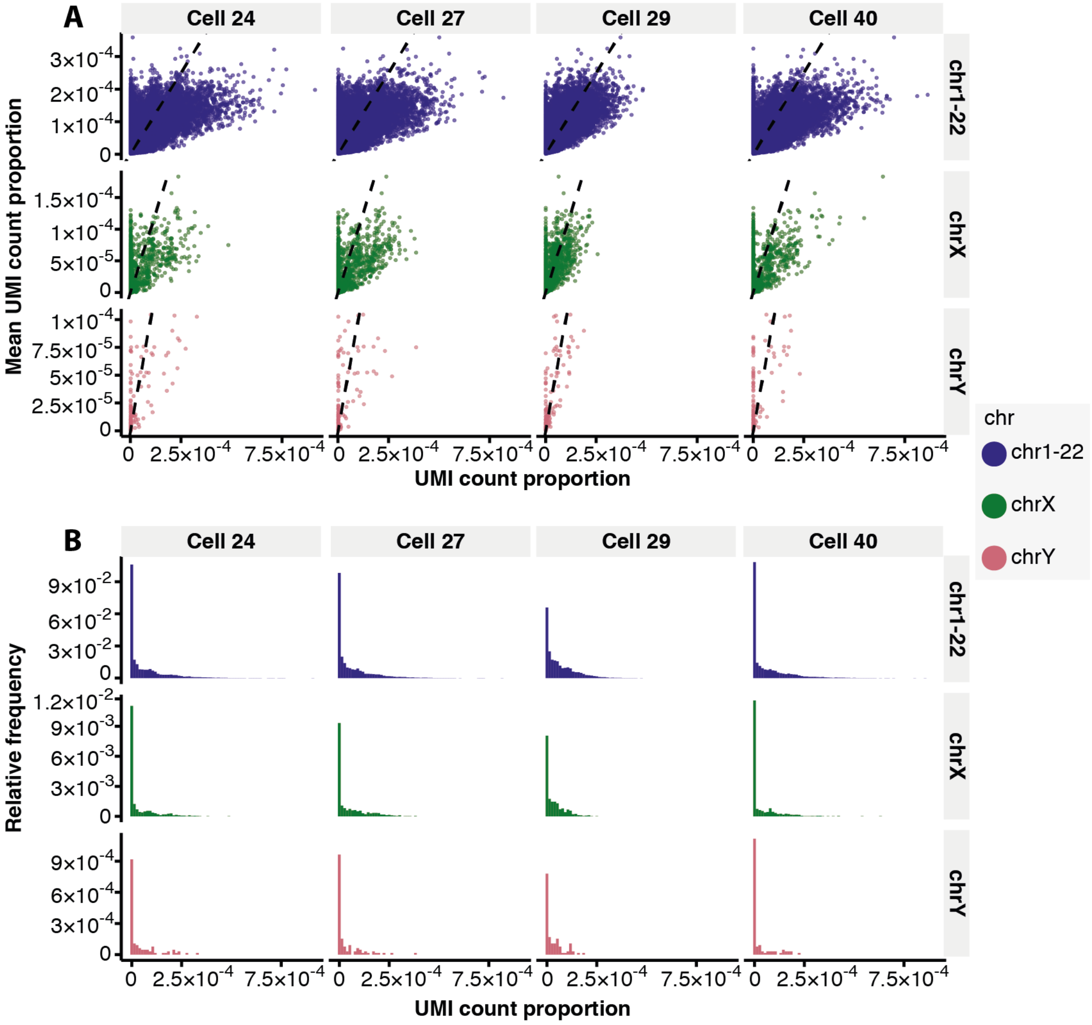


Figure 5.10 Cell-level UMI count exploration for four single cells from a male sample. A: Exploring the relationship between a cell's UMI count proportion against the pseudo-bulk UMI count proportion (in which the pseudo-bulk UMI count is the mean UMI count across all cells in the male x10 experimental condition). B: The distribution of UMI count proportions within each cell.

5.4.4 Discussion

These data suggest that scFAST-SeqS is a promising avenue for future research. The results show that copy number signal is clearly present when the mean counts are computed across as few as 20 cells. Indeed, copy number signal is present within individual cells with count proportions scaling proportionally as we expect.

To infer copy number from scFAST-SeqS data and extend the conliga model, there are several aspects of the data that would need to be modelled appropriately. The distribution of counts at a single-cell level would need to be investigated further. Similar to bulk FAST-SeqS it appears the PCR efficiency varies between loci. However, the global PCR efficiency appears to vary between cells. We may wish to explore modelling the probability of amplifying each locus. Perhaps each locus could have a probability of amplification that is scaled by a cell effect. It is clear that we will need to adequately model the loci drop outs or zeros. It may be that a probability of amplification adequately models the zero counts or we may need to use zero inflation models.

Once a count distribution is appropriately chosen for scFAST-SeqS observations, we will need to consider how to model the copy number profiles. It seems natural to model cells as mixtures (clones) whereby each clone has its own copy number profile. This way, by pooling together the statistical strength of multiple cells, we should more accurately infer the copy number profiles for each clone. The number of clones and cell memberships could be modelled by the use of a Dirichlet process.

Although I have not shown the results here, SNPs can be detected in scFAST-SeqS data. When pooling cells together, we see mismatch and indel frequencies around 0.5 which suggest the presence of heterozygous SNPs. These data could be used jointly with the loci count data to infer the allele-specific copy number profiles for each clone.

scFAST-SeqS could be combined with a scRNA-seq method to sequence the DNA and RNA from the same cell in a similar fashion to G&T sequencing [283]. The scRNA-seq data could be integrated with the scFAST-SeqS data to infer clusters of clones and additionally, assist with the inference of copy number. There are several recent methods that use scRNA-seq to infer copy number, such as InferCNV [284], HoneyBADGER [285], CONICS [286] and CaSpER [287]. Another method, clonealign [288], uses scRNA-seq applied to a set of cells and scDNA-seq applied to a separate batch of cells from the same tumour population, to map gene expression profiles from scRNA-seq to clones derived from DNA-seq. scFAST-SeqS and scRNA-seq data derived from the same cells could be used to improve on these methods.

Further work is needed to scale the protocol so that it could be applied to thousands of cells. If the protocol can be scaled to thousands of cells, the number of reads required per cell such that all the tumour clones could be accurately inferred would need to be determined. In addition, optimisation of the protocol is required to minimise the variance between cells and the number of loci that fail to amplify within cells. This could be achieved be varying the number of PCR1 cycles, altering the PCR reaction volume, varying the quantities of primer and other experimental variables.

5.4.5 Methods

scFAST-SeqS library preparation

Single cell PBMCs were sorted into 384-well PCR plates, in 3 µl 1X lysis buffer (diliuted from TaKaRa 10X Lysis Buffer 635013). Cells were incubated for five minutes at room temperature and then PCR plates were frozen at −80 °C for later use.

PCR1 reaction

5 µl Phusion HiFi Master Mix (ThermoFisher F531L), 0.1 µl PCR1 Forward Primer (10uM, Sigma), 0.1 µl PCR1 Reverse Primer (10uM, Sigma) and 1.8 µl nuclease free water (NFW) were added to the 3 µl lysate in each well.

The cycling conditions for PCR1 were 98 °C for 2 minutes followed by 5, 10 or 15 (depending on the experiment) cycles of, 98 °C for 10 seconds, 57 °C for 2 minutes and 72 °C for 2 minutes. This was followed by 4 °C until the samples underwent clean-up. For the PCR1 products that underwent purification using ExoSAP, 1 µl Exo I (NEB M0293S) and 2 µl rSAP (NEB M0371S) were added to the mix. The mixture underwent incubation at 37 °C for 15 minutes followed by heating inactivation at 80 °C for a further 15 minutes. For the PCR1 products that underwent bead clean-up, PCR1 contents were transferred to 200 µl PCR tube strips and 1.4X AMPure bead clean-up was performed as per section 4.8.1. Purified product was eluted in 13 µl NFW.

PCR2 reaction

15 µl Phusion HiFi Master Mix (ThermoFisher F531L), 0.3 µl PCR2 Forward Primer (10uM, Sigma), 0.3 µl PCR2 Reverse Primer (10uM, Sigma) and 1.4 µl nuclease free water (NFW) were added to the 13 µl purified PCR1 product.

The cycling conditions for PCR2 were 98 °C for 2 minutes followed by 25 cycles of, 98 °C for 10 seconds, 57 °C for 2 minutes and 72 °C for 2 minutes. This was followed by 4 °C until the samples underwent bead clean-up. PCR2 products from each experimental batch (e.g. PCR2 products from F10 cells) were pooled together. Pooled PCR2 product was purified using a 1X AMPure bead clean up as described in section 4.8.1 and purified product was eluted in 10 µl NFW. Each pooled batch was quantified using Qubit High sensitivity protocol (ThermoFisher Q32854) and quality control was performed on a BioAnalyser using DNA1000 chip (Agilent). Libraries representing F10, F15, M10 and M15 were then diluted accordingly such that they could be pooled in equimolar ratios.

Sequencing

Each batch of 80 cells were sequenced on a lane of an Illumina HiSeq 4000 machine using paired-end 150 bp reads.

Chapter 6

Conclusions

The profiling of somatic copy number alterations present in tumour samples has important implications for cancer research and clinical care. Somatic copy number profiling can be performed using high or low-coverage WGS, WES, or alternatively other sequencing and array-based assays. However, while the costs of high-throughput sequencing are reducing, the use of these assays is currently impractical for clinical application. In addition to the costs being prohibitive, the laborious library preparation steps associated with WGS make it challenging to perform in a clinical setting.

In comparison, PCR is relatively simple, low-cost and widely used in research and clinical laboratories around the world. Kinde *et al.* [104] developed a simple protocol that involves two rounds of PCR to amplify more than 10,000 primate-specific LINE1 elements across the genome. The protocol can be performed in an hour and lends itself to automation. It results in amplicon libraries ready for high-throughput sequencing. The cost of the entire protocol, including sequencing two million reads per sample, is currently less expensive than WGS library preparation kits alone [108]. As such, the FAST-SeqS protocol has exciting clinical application and could be a valuable tool for biomedical research.

To date, the published approaches for the analysis of FAST-SeqS data are limited. In particular, they are constrained to the detection of whole chromosome or chromosome-arm amplifications and deletions and the output of these approaches is limited to three levels; amplified, deleted and normal. Furthermore, the methods do not account for sampling variation (which can be substantial when performing multiplexed sequencing) in a principled way and some employ data normalisation approaches that are not suitable for sequencing data (e.g. quantile normalisation). As such, current methods for FAST-SeqS data have low-resolution and are unable to quantify SCNAs. The quantification and detection of SCNAs at the sub-chromosomal-arm level has important clinical relevance for the diagnosis, prognosis, treatment selection and monitoring of resistance to therapy of certain cancer types. As such, current computational approaches limit the clinical and biomedical research application of the FAST-SeqS protocol. This motivated the development of a new approach for analysing FAST-SeqS data.

In chapter 2, I described the development of a method and tool (conliga) which can be used to infer relative copy number at the amplicon level. I explored the aspects of FAST-SeqS data and, by finding an appropriate parametric distribution, I developed a probabilistic approach to model FAST-SeqS data generation. In addition, I used a hidden Markov model and a Bayesian nonparametric approach to allow for countably infinite copy number states, while modelling the fact that neighbouring FAST-SeqS loci are likely to share the same relative copy number. In this way, the statistical strength of all loci are pooled to infer global relative copy number clusters, while statistical strength is shared between spatially related loci to infer contiguous regions of altered copy number.

Based on the probabilistic generative model, I described two MCMC inference algorithms. The first infers a global bias in the FAST-SeqS protocol using a set of normal control samples. The second built upon Fox *et al.*'s [131] blocked Gibbs sampler and is used to infer a relative copy number profile for a test sample. I discussed and explored several ways in which the posterior distribution could be summarised to make use of the copy number state information. Lastly, I showed evidence that the MCMC converges and that the method worked as expected by inferring the relative copy number differences between male and female samples.

Several cancer types appear to be driven by structural alterations and somatic copy number alterations [45]. In chapter 3, we studied an example of this and motivated the use of a low-cost copy number assay for the detection of oesophageal adenocarcinoma (OAC). OAC incidence is rising rapidly in the West and carries a very poor prognosis. Its poor prognosis is due to its symptoms presenting at a late stage in the course of disease. By the time it is detected, it has often progressed to an incurable stage. Detecting OAC at its early stages (or detecting premalignant lesions that are likely to develop to OAC) when it is more easily treated is a key challenge to improve prognostic outcomes. The current strategies for the screening and surveillance of OAC development are severely limited. The development of new technologies and biomarkers are required to enable population-based screening and better patient stratification to identify those at risk for developing OAC.

The development of OAC is characterised by chromosomal instability. Li *et al.* [213] showed that SCNAs accumulate in premalignant lesions years before development of OAC, while those patients who do not progress to OAC tend to have stable SCNAs over time [213]. In contrast, small-scale somatic mutations in key driver genes (other than *TP53*) appear to offer little value in determining which patients will progress [209]. In addition, the identification of amplifications and deletions of particular oncogenes and tumour suppressors can be used to guide targeted therapies. As such, a low-cost copy number assay combined with a non-invasive tissue sampling device such as the Cytosponge [289], could be used to improve the detection and clinical care of OAC.

By using clinical OAC samples that had previously been sequenced using high-coverage WGS, we were able to see how conliga performs in comparison to other tools. By comparing conliga's results to ASCAT, we saw that conliga was able to recapitulate the RCN profiles of OAC samples. Indeed, conliga appears to offer comparable performance to a commonly used tool for low-coverage WGS (QDNAseq) even when using 4.5-fold fewer reads.

By performing an *in silico* dilution series to vary the purity of OAC samples, we saw that conliga is particularly useful in discriminating SCNAs at low purity and appears to outperform QDNAseq for this task. The results suggested that conliga is able to detect prominent chromosomal-arm alterations at 2-3% purity when using two million reads per sample. In addition, we saw that a focal amplification was detected at purities as low as 0.5%.

The *TP53* mutant fractions observed in Cytosponge samples by Weaver *et al.* [209] suggest that the proportion of Barrett's oesophagus or cancer cells sampled can vary widely. For conliga and FAST-SeqS to be applied to the clinic, the number of reads may need to be increased to reliably detect alterations. Future work is required in this area to determine how many reads are required and, by increasing the costs of sequencing, whether the protocol would be cost-effective for widespread use. Potentially, improvements to the protocol could be made to reduce the technical variance. This may also offer a way to improve sensitivity.

We know that FAST-SeqS can be applied to cell-free DNA as Kinde *et al.* [104], Belic *et al.* [106] and Douville *et al.* [107] showed. Although I do not show the results in this thesis, I have successfully applied conliga to data produced by Belic *et al.* [106] and conliga shows promise with application to circulating DNA. The challenge with ctDNA is that it often represents a very small fraction of cfDNA, particularly in early stage disease. Further work is required to see if conliga has clinical potential in the detection SCNAs from cfDNA.

I explored the potential use of conliga and FAST-SeqS in detecting the copy number status of recurrently amplified oncogenes and deleted tumour suppressors. Intriguingly, some LINE1 amplicons originate from the intronic regions of several genes of interest. I found that in some cases, FAST-SeqS and conliga could detect focal deletions that have occurred *within* recurrently altered tumour suppressor genes. Although, frequently these intra-gene alterations were not detected by conliga. I note that QDNAseq struggled to detect some of the focal deletions also. The distance of the LINE1 amplicons to genes of interest can vary and in some cases, highly focal SCNAs can be missed due to not being sampled by LINE1 amplicons. However, in a number of cases conliga was able to detect clinically relevant amplifications and deletions.

Can conliga and FAST-SeqS be used to quantify therapeutically relevant SCNAs at gene-level resolution? The answer is, it depends. For clinical application, FAST-SeqS could be supplemented by the targeted sequencing of a clinically relevant gene panel. In this way, in addition to a genome-wide copy number profile, gene-level SCNAs could be better estimated and small-scale somatic mutations could be obtained. This may provide a low-cost strategy for clinical use. The gene-panel data could act as a auxiliary source of information in addition to conliga's RCN inference. However, it would be preferable to incorporate the gene-panel data into the conliga model, to use all available information to infer a copy number profile. Further work would be required to extend conliga for this purpose.

Alternatively, or additionally, further primer pairs could be developed. The FAST-1 primer pair designed by Kinde *et al.* [104] amplify a region in the 3' UTR of L1PA3-L1PA11 elements (primarily). At least in theory, it should be possible to design other primer pairs that target other LINE1 subclasses or other repeat elements. Indeed, Alu elements (which are a subclass of SINEs and approximately 300 bp long) are the most repeated elements in the human genome [3]. Assuming that their sequences have diverged sufficiently to align them uniquely to the genome, they could be an interesting repeat element to target. Multiplexing several primer pairs which amplify several different classes of repeats should lead to a more granular sampling of the genome. In addition to this, by introducing degenerate bases or altering the primer sequences, we may be able to even out the amplification biases or increase the amplification bias towards regions of interest. This may allow us to obtain increased statistical strength and make more sensitive RCN calls in regions of interest. As such, the identification and modification of primer pairs that target repeat elements is a potentially interesting area of future work. Since cancer types can have different regions of interest, computational tools to aid with the choice of optimal primer pairs for the application to particular cancer types is an interesting area of future study. I note also that the FAST-1 primer pair will only be of use to the study of CNAs in primates and the development of other primer pairs are required to study other mammals.

In this thesis, I have focused on the inference of SCNA profiles from FAST-SeqS data. However, using these inferred profiles to provide a probability that a sample contains malignant cells (or a probability that a premalignant lesion will develop into cancer) is an area for future study. This is likely to require a large number of SCNA profiles and clinical data (including outcomes).

Further to this, if we know the patterns of SCNAs that are present in particular cancer types, can we use this prior information to assist with SCNA inference? This is an interesting topic for further work. It is not clear how this prior information could be included within the conliga model. This would require prior distributions placed over the hidden states. However, in conliga's current form the hidden states are exchangeable and have no direct meaning. An alternative would be to fix the hidden states to represent absolute or allele-specific copy numbers. In this way, meaningful priors could be placed on them. However, this approach would require assumptions about the clonal composition of the tumour sample. One of the strengths of the conliga model is that we are not constrained by clonal composition assumptions.

In chapter 4, I described the modifications I made to the PCR1 and PCR2 primer sequences. The two most important modifications were 1) the inclusion of variable length UMI sequences on either end of the amplicon and 2) the inclusion of unique dual indices (UDIs). While the LINE1 amplicons have some level of heterogeneity in their sequence content, the resulting sequencing library consists of low-diversity sequences. The variable length UMIs increased the complexity of the sequencing library and in theory, should allow us to reduce the quantity of PhiX control in further sequencing experiments. This could allow us to substantially increase the number of FAST-SeqS molecules that are sequenced per sequencing run. Future work is required to ascertain the quantity of PhiX control that is required to maintain sequencing quality and whether its use could be avoided entirely. This would allow more reads to be used for samples by either increasing the number of samples or increasing the number of reads per sample.

The inclusion of UDIs enabled the identification of sample index cross-talk. We saw a high frequency of index swapping events resulting in 3.8% of reads being assigned to invalid index combinations (invalid samples). With the use of UDIs, these reads can be excluded from analyses. However, if UDIs were not included this would result in up to 3.8% of reads being misassigned to the wrong sample (if using a combinatorial dual index strategy). Read misassignments would affect the sensitivity and specificity of downstream analyses with reads from cancer samples being falsely assigned to normal control samples and vice versa. Given these results, we recommend that future FAST-SeqS experiments use a UDI strategy. We saw that index-cross talk events occurred with increased frequency in a pattern consistent with the pipetting order and with the order of the sample number. Future experiments are required to identify the source of this issue and where possible, steps should be taken to reduce index cross-talk.

In chapter 4, I described the creation of a new pipeline to process paired-end FAST-SeqS data. I note that there are many ways to process sequencing data; different strategies and combinations of tools and thresholds can be used. The pipeline described is unlikely to be the best way of processing the data and further work is required to improve it. Ultimately, the goal of the pipeline is to maximise the number of reads that are used in downstream analyses while minimising error. The error can be reduced by 1) identifying PCR2 duplicates and 2) minimising false alignments. I presented a strategy which utilised functions from UMI-tools [183] to identify PCR2 duplicates. This appeared to work well and we saw that UMIs can be useful in identifying false alignments also. I did not show whether the use of the UMI clustering strategy reduced noise in downstream analyses. In theory, by using the

UMI reduced counts (i.e. the number of UMI clusters) rather than the raw counts, we may be able to improve the sensitivity and specificity of conliga's calls. Further work is required to investigate this.

Aligning reads to repeat regions is a notoriously difficult task and often these regions are excluded from analyses. As such, it is important to investigate how aligners perform in aligning FAST-SeqS reads to the genome. The results I presented in chapter 4 suggest that BWA-MEM should be used in favour of Bowtie 2 for this task. However, further work is required to test other aligners. Further tests of performance could include the simulation of reads from FAST-SeqS loci so that the ground truth is known when reads are aligned.

It could be explored whether we can use the mapping qualities to propagate the uncertainty of alignments into the copy number inference. Currently, a substantial number of reads are excluded from downstream analyses because we cannot be confident that they have been aligned to the correct position in the genome. By incorporating alignment probabilities, we could make use of reads that are currently excluded. However, there is a problem with this approach. For example, if 100 reads align equally well to three locations, we cannot know the true proportion of reads originating from each of these locations in control samples, i.e. we cannot estimate $\hat{m}$. The assumption that the reads are equally dispersed between the three locations may not be a good one. We could imagine doing a series of knockout experiments to resolve this whereby each ambiguous FAST-SeqS locus is deleted and the effect on the number of counts is seen. However, this is likely to be laborious and impractical. Alternatively, by using previous RCN inferences, we may be able to infer $\hat{m}$ for some loci; if we know the relative copy number of the region in the genome that an ambiguous locus originated from, we may be able to estimate $\hat{m}$. This may not work in practice, particularly if reads align equally well to many locations. However, it is an interesting avenue for future work and incorporating the uncertainty into downstream analyses would represent an interesting technical challenge.

I note that there are several potential improvements that could be made to the conliga model. Currently, the model assumes that all states have the same self-transition bias. This is controlled by the hyperparameter, κ. We know that this assumption is not true. For instance, homozygous (biallelic) deletions tend to be short as are highly focal amplifications. It would be appropriate to make κ dependent on the copy number. This is currently challenging since we are inferring relative copy number. As such, copy number is confounded with purity and how κ could be made dependent on relative copy number is unclear. Alternatively, inferring absolute or allele-specific number would make this more achievable. Additionally, the transition probabilities are not dependent on loci distance. We know from figure 2.7 that the genomic distance between loci can vary widely. As such, it would be appropriate to make the transition probabilities dependent on distance; with loci closer to each other more likely to have the same copy number.

While we saw evidence of sample-specific biases in earlier work on oesophageal samples, they became more apparent when I analysed the prostate samples. I showed that sample-specific biases could affect the RCN inference. These results were concerning and were unexpected. It was clear that my original assumption that the bias was specific to each locus and not to the sample did not hold. As such, it became apparent that I needed to correct for these biases. At first, it wasn't clear whether these artifacts were introduced by processing the data with the new pipeline. However, we saw similar results when processing the data with the original single-end pipeline. Since the possible sources of sample-specific biases could be numerous, my initial strategy was to try to learn the biases from the data using a factor model. As I described in section 4.5, this strategy was ultimately unsuccessful. However, I noted that with some tweaks to the model and with more samples it may prove successful.

Rather than learning the biases from the data, I identified two major sources of sample-specific bias; large-scale GC content and differential amplification of highly and lowly amplified loci between samples. I proposed a method to correct for these effects which involved the use of a 2-dimensional B-spline. By updating the conliga model to jointly infer sample-specific biases and the RCN profile, I showed that (in most cases) I was able to remove RCN artifacts. Indeed, by inferring sample-specific biases, increased sensitivity was achieved and we were able to infer SCNAs that were previously not detected in some cases. While the use of 2-dimensional B-splines appeared to work well and did not appear to regress away true SCNAs, there may be more appropriate ways to correct for these biases. Indeed, further work should involve investigating whether two independent splines could be used to reduce the number of parameters. Furthermore, parametric distributions may be appropriate; potentially, the amplification bias could be modelled by a sigmoid or logistic function and a parametric shape could be used to model the GC bias. I note that the sample-specific bias correction did not work in all cases and further work is required to identify and correct for all possible sources of sample-specific bias. Moreover, the experimental cause of these biases should be investigated and wherever possible their effect should be minimised.

In many samples, including those considered to be 'normal', conliga infers several short CNAs, often involving a single locus. It is not clear whether these are false positive calls or whether they represent true biological signal. Given that CNVs occur in approximately 1.2% of an individual's genome, it is conceivable that some of these inferred short segments represent CNVs present in the germline. However, some of the short RCN changes could be artifactual and may be introduced for several reasons. When comparing raw counts from normal samples obtained from different patients, we observed outlying counts (see Figure 2.5 for example). These outlying counts can result in the inference of short RCN segments. Although I have not shown this in the thesis, when comparing raw counts derived from samples from the same patient, we did not observe such outliers. This suggests that outlying counts appear to be biological in origin. However, while they could be the result of CNVs, they may also be the result of germline sequence variants in the LINE1s that affect their amplification efficiency or their ability to be aligned to the genome. The conliga model could be altered to use FAST-SeqS data obtained from matched healthy tissue from the same patient (in addition to the tumour sample) to infer the RCN profile of the tumour sample. In doing so, these outliers (whether due to CNVs or otherwise) should not lead to spurious calls in the SCNA profile of the tumour sample. However, introducing a matched healthy tissue sample would involve doubling the cost by requiring two FAST-SeqS samples per patient. Alternatively, further work could be done to identify the source of these outliers. If they are found not to be the result of CNVs, the beta-binomial emission distribution could be altered to incorporate the presence of outliers. Indeed, the beta-binomial emission distribution is an approximation to the true count distribution. An imperfect emission distribution will result in false positive RCN calls and the inference of short RCN segments. Further work to identify alterations to the emission distribution, which more closely matches the true distribution of FAST-SeqS count data, would help to reduce the false positive RCN calls.

The Markovian assumption of the HMM may also play a role in the inference of short RCN segments. This is because the Markovian assumption implies that the lengths of SCNAs follow a geometric distribution [149]. This assumption is unlikely to reflect the distribution of SCNA lengths present in real tumour samples. This is evidenced by Figure 1a of Beroukhim et al. [93], which shows that when the length of SCNAs are normalised by the length of the chromosome arms, the normalised SCNA length distribution is multi-modal; with one mode reflecting short focal amplifications and two other two modes reflecting SCNAs that affect whole chromosomes and chromosome-arms. Further work could be conducted to explore whether the use of the Hierarchical Dirichlet process Hidden Semi-Markov Model (HDP-HSMM) [149] could overcome such issues.

In chapter 5, I showed the existence of variants within the LINE1 elements, some of which are likely to be SNPs. This was previously noted by Douville *et al.* [107] and used to detect evidence of allelic imbalance of chromosome arms. I showed that this data can be used as an auxiliary source of information to infer tumour purity, ploidy and could be used to infer allele-specific copy number also. Low-coverage WGS (~ 0.1X), does not provide this information and makes the alternative use of FAST-SeqS more appealing.

Large-scale cancer genome studies require that samples contain a minimal tumour purity before sequencing such that sufficient statistical power is present to call somatic alterations. Currently, the task of screening samples for sufficient purity is placed on pathologists who have other clinical duties. FAST-SeqS and conliga could provide an alternative low-cost method to estimate the purity and ploidy of samples. We showed a comparison between pathologist purity estimates and estimates obtained using conliga's RCN inference in chapter 5, which shows this has potential. Indeed, in some cases a pathologist's purity estimate can be vastly different from purity estimates derived from sequencing data. By using a low-cost sequencing approach to estimate purity, additional high-purity samples could be identified which could be included in large-scale studies. Additionally, some samples are believed to be of high tumour purity but after performing HC WGS, they are found to have insufficient purity and are excluded from analyses. It may be possible to avoid allocating sequencing resources to these samples if they are deemed to be low purity via conliga's RCN estimates.

Rather than use alelle frequencies as auxiliary data after RCN inference, it would be preferable to use the data in the copy number inference. The most straightforward approach would be to integrate the data to jointly segment the genome based on read depth and allele frequency (and there are several tools that currently do this for other data types). As such, each hidden state would represent a minor allele frequency and a relative copy number. Inferences about ploidy and purity could be made following segmentation of the genome. This approach would avoid building in assumptions about the clonal composition of the tumour sample. Further work is required to find an appropriate distribution for the allele frequencies in order to achieve this.

In chapters 4 and 5, we saw that conliga is able to infer RCN profiles that are indicative of samples consisting of a mixture of clones. As I noted, several methods have attempted to address the inference of the clonal composition from sequencing data. With the inclusion of SNP data, similar methods could be applied to FAST-SeqS data. However, this is a difficult problem. Even with the assumption of a single clone, there are several possible solutions for purity, ploidy and allele-specific copy number. In the absence of somatic point mutations, the problem is more difficult and clonal composition inference is practically limited to few clones.

Rather than attempting to infer clonal composition from bulk FAST-SeqS data, it seemed that FAST-SeqS could be suitable for application to single cells. By adapting the protocol to single cells (in collaboration with Sarah Field), we showed we were able to obtain scFAST-SeqS data. As far as we are aware, this is the first time scFAST-SeqS data has been produced. Here, I presented a preliminary analysis showing that copy number signal is present in scFAST-SeqS data. The results are promising but they indicate that significant modifications will be required to the conliga model to infer single cell copy number from the data (which I discussed in section 5.4.4). Further work is required to find a suitable distribution to model the FAST-SeqS counts of single cells. It may be appropriate to use a Dirichlet process to cluster cells into clones and pool the statistical strength from the single cells to infer a copy number profile per clone.

Our preliminary scFAST-SeqS experiments were limited to 80 cells and provide a proof of concept. For scFAST-SeqS to be useful, the protocol will need to be scaled to potentially thousands of cells. Further work is required to explore the various ways of achieving this. Indeed, it should be possible to obtain scRNA-seq data from the same cells and incorporate these data into the single cell copy number inference.

Even in the absence of scFAST-SeqS data, the low-cost of the bulk FAST-SeqS protocol and low-input DNA required will encourage the production of several spatially or temporally related samples. This may provide another way to explore intratumoural heterogeneity. Indeed, we have recently produced FAST-SeqS data from several spatially related samples (not shown in this thesis). The initial results show that RCN profiles are similar between samples and some differences can also be seen. However, rather than inferring RCN profiles for these samples independently, it would be more appropriate to model their relatedness and jointly infer their RCN profiles. A recent preprint on bioRxiv explores the inference of SCNAs from multiple related WGS samples [290] and similar methods have been explored for the inference of SNVs from multiple related samples [119]. In this way, and by pooling the statistical strength between samples, we should be more able to infer regions that are shared and regions that differ between related samples.

To conclude, this thesis has focused on inferring somatic copy number alterations from a low-cost, simple and clinically applicable assay. By appropriate modelling of the data, I have shown that higher resolution and richer inferences could be made from the data. FAST-SeqS, along with the method and tool I developed, could have exciting clinical application and may additionally provide a low-cost tool for biomedical research. I have presented several areas for further study and hope that these will inspire improvements to the methods described here, ultimately for further clinical application and to provide greater insights into cancer biology.

Bibliography

[1] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68(6):394–424, 2018.

[2] Michael R. Stratton, Peter J. Campbell, and P. Andrew Futreal. The cancer genome. *Nature*, 458(7239):719–724, 2009.

[3] Eric S. Lander, Lauren M. Linton, Bruce Birren, Chad Nusbaum, Michael C. Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William Fitzhugh, Roel Funke, Diane Gage, Katrina Harris, Andrew Heaford, John Howland, Lisa Kann, Jessica Lehoczky, Rosie Levine, Paul McEwan, Kevin McKernan, James Meldrim, Jill P. Mesirov, Cher Miranda, William Morris, Jerome Naylor, Christina Raymond, Mark Rosetti, Ralph Santos, Andrew Sheridan, Carrie Sougnez, Nicole Stange-Thomann, Nikola Stojanovic, Aravind Subramanian, Dudley Wyman, Jane Rogers, John Sulston, Rachael Ainscough, Stephan Beck, David Bentley, John Burton, Christopher Clee, Nigel Carter, Alan Coulson, Rebecca Deadman, Panos Deloukas, Andrew Dunham, Ian Dunham, Richard Durbin, Lisa French, Darren Grafham, Simon Gregory, Tim Hubbard, Sean Humphray, Adrienne Hunt, Matthew Jones, Christine Lloyd, Amanda McMurray, Lucy Matthews, Simon Mercer, Sarah Milne, James C. Mullikin, Andrew Mungall, Robert Plumb, Mark Ross, Ratna Shownkeen, Sarah Sims, Robert H. Waterston, Richard K. Wilson, Ladeana W. Hillier, John D. McPherson, Marco A. Marra, Elaine R. Mardis, Lucinda A. Fulton, Asif T. Chinwalla, Kymberlie H. Pepin, Warren R. Gish, Stephanie L. Chissoe, Michael C. Wendl, Kim D. Delehaunty, Tracie L. Miner, Andrew Delehaunty, Jason B. Kramer, Lisa L. Cook, Robert S. Fulton, Douglas L. Johnson, Patrick J. Minx, Sandra W. Clifton, Trevor Hawkins, Elbert Branscomb, Paul Predki, Paul Richardson, Sarah Wenning, Tom Slezak, Norman Doggett, Jan Fang Cheng, Anne Olsen, Susan Lucas, Christopher Elkin, Edward Uberbacher, Marvin Frazier, Richard A. Gibbs, Donna M. Muzny, Steven E. Scherer, John B. Bouck, Erica J. Sodergren, Kim C. Worley, Catherine M. Rives, James H. Gorrell, Michael L. Metzker, Susan L. Naylor, Raju S. Kucherlapati, David L. Nelson, George M. Weinstock, Yoshiyuki Sakaki, Asao Fujiyama, Masahira Hattori, Tetsushi Yada, Atsushi Toyoda, Takehiko Itoh, Chiharu Kawagoe, Hidemi Watanabe, Yasushi Totoki, Todd Taylor, Jean Weissenbach, Roland Heilig, William Saurin, Francois Artiguenave, Philippe Brottier, Thomas Bruls, Eric Pelletier, Catherine Robert, Patrick Wincker, André Rosenthal, Matthias Platzer, Gerald Nyakatura, Stefan Taudien, Andreas Rump, Douglas R. Smith, Lynn Doucette-Stamm, Marc Rubenfield, Keith Weinstock, Mei Lee Hong, Joann Dubois, Huanming Yang, Jun Yu, Jian Wang, Guyang Huang, Jun Gu, Leroy Hood, Lee Rowen, Anup Madan, Shizen Qin, Ronald W. Davis, Nancy A. Federspiel, A. Pia Abola, Michael J. Proctor, Bruce A. Roe, Feng Chen, Huaqin Pan, Juliane Ramser, Hans Lehrach, Richard Reinhardt, W. Richard McCombie, Melissa De La Bastide, Neilay Dedhia,

Helmut Blöcker, Klaus Hornischer, Gabriele Nordsiek, Richa Agarwala, L. Aravind, Jeffrey A. Bailey, Alex Bateman, Serafim Batzoglou, Ewan Birney, Peer Bork, Daniel G. Brown, Christopher B. Burge, Lorenzo Cerutti, Hsiu Chuan Chen, Deanna Church, Michele Clamp, Richard R. Copley, Tobias Doerks, Sean R. Eddy, Evan E. Eichler, Terrence S. Furey, James Galagan, James G.R. Gilbert, Cyrus Harmon, Yoshihide Hayashizaki, David Haussler, Henning Hermjakob, Karsten Hokamp, Wonhee Jang, L. Steven Johnson, Thomas A. Jones, Simon Kasif, Arek Kaspryzk, Scot Kennedy, W. James Kent, Paul Kitts, Eugene V. Koonin, Ian Korf, David Kulp, Doron Lancet, Todd M. Lowe, Aoife McLysaght, Tarjei Mikkelsen, John V. Moran, Nicola Mulder, Victor J. Pollara, Chris P. Ponting, Greg Schuler, Jörg Schultz, Guy Slater, Arian F.A. Smit, Elia Stupka, Joseph Szustakowki, Danielle Thierry-Mieg, Jean Thierry-Mieg, Lukas Wagner, John Wallis, Raymond Wheeler, Alan Williams, Yuri I. Wolf, Kenneth H. Wolfe, Shiaw Pyng Yang, Ru Fang Yeh, Francis Collins, Mark S. Guyer, Jane Peterson, Adam Felsenfeld, Kris A. Wetterstrand, Richard M. Myers, Jeremy Schmutz, Mark Dickson, Jane Grimwood, David R. Cox, Maynard V. Olson, Rajinder Kaul, Christopher Raymond, Nobuyoshi Shimizu, Kazuhiko Kawasaki, Shinsei Minoshima, Glen A. Evans, Maria Athanasiou, Roger Schultz, Aristides Patrinos, and Michael J. Morgan. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[4] David A Wheeler and Linghua Wang. From human genome to cancer genome: The first decade. *Genome Research*, 23(7):1054–1062, 2013.

[5] S Anderson, A T Bankier, G Barrell, M H L de Brujin, A R Coulson, J Drouin, I C Eperon, D P Nierlich, B A Roe, F Sanger, P H Schreier, A J H Smith, R Staden, and I G Young. Sequence and organization of the human mitochondrial genome. *Nature*, 290(5806):457–465, 1981.

[6] Adam Frankish, Mark Diekhans, Anne Maud Ferreira, Rory Johnson, Irwin Jungreis, Jane Loveland, Jonathan M. Mudge, Cristina Sisu, James Wright, Joel Armstrong, If Barnes, Andrew Berry, Alexandra Bignell, Silvia Carbonell Sala, Jacqueline Chrast, Fiona Cunningham, Tomás Di Domenico, Sarah Donaldson, Ian T. Fiddes, Carlos García Girón, Jose Manuel Gonzalez, Tiago Grego, Matthew Hardy, Thibaut Hourlier, Toby Hunt, Osagie G. Izuogu, Julien Lagarde, Fergal J. Martin, Laura Martínez, Shamika Mohanan, Paul Muir, Fabio C.P. Navarro, Anne Parker, Baikang Pei, Fernando Pozo, Magali Ruffier, Bianca M. Schmitt, Eloise Stapleton, Marie Marthe Suner, Irina Sycheva, Barbara Uszczynska-Ratajczak, Jinuri Xu, Andrew Yates, Daniel Zerbino, Yan Zhang, Bronwen Aken, Jyoti S. Choudhary, Mark Gerstein, Roderic Guigó, Tim J.P. Hubbard, Manolis Kellis, Benedict Paten, Alexandre Reymond, Michael L. Tress, and Paul Flicek. GENCODE reference annotation for the human and mouse genomes. *Nucleic acids research*, 47(D1):D766–D773, 2019.

[7] Bert Vogelstein, Nickolas Papadopoulos, Victor E. Velculescu, Shibin Zhou, Luis A. Diaz, and Kenneth W. Kinzler. Cancer genome landscapes, 2013.

[8] Shiv I.S. Grewal and Songtao Jia. Heterochromatin revisited. *Nature Reviews Genetics*, 8(1):35–46, 2007.

[9] Jiyong Wang, Sharon T. Jia, and Songtao Jia. New Insights into the Regulation of Heterochromatin. *Trends in Genetics*, 32(5):284–294, 2016.

[10] O. T. Avery, Colin M. MacLeod, and Maclyn McCarty. Studies on The Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types: Induction of Transformation by a Desoxyribonucleic Acid. *Journal of Experimental Medicine*, 79(2):137–158, 1944.

[11] J. D. Watson and F. H. C. Crick. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, 171(4356):737–738, 1953.

[12] T Boveri. Zur Frage der Entstehung maligner Tumoren. *Gustav Fischer*, 2:1–64, 1914.

[13] Lawrence A. Loeb and Curtis C. Harris. Advances in chemical carcinogenesis: A historical review and prospective. *Cancer Research*, 68(17):6863–6872, 2008.

[14] P C Nowell and D A Hungerford. A minute chromosome in human chronic granulocytic leukemia. *Science*, 132:1497, 1960.

[15] Janet D Rowley. A New Consistent Chromosomal Abnormality in Chronic Myelogenous Leukaemia identified by Quinacrine Fluorescence and Giemsa Staining. *Nature*, 243:290–293, 1973.

[16] Theodore G Krontiris and M Geoffrey. Transforming activity of human tumor DNAs. *PNAS*, 78(2):1181–1184, 1981.

[17] Chiaho Shih, L. C. Padhy, Mark Murray, and Robert A. Weinberg. Transforming genes of carcinomas and neuroblastomas introduced into mouse fibroblasts. *Nature*, 290(5803):261–264, 1981.

[18] E. Premkumar Reddy, Roberta K. Reynolds, Eugenio Santos, and Mariano Barbacid. A point mutation is responsible for the acquisition of transforming properties by the T24 human bladder carcinoma oncogene. *Nature*, 300(5888):149–152, 1982.

[19] Clifford J. Tabin, Scott M. Bradley, Cornelia I. Bargmann, Robert A. Weinberg, Alex G. Papageorge, Edward M. Scolnick, Ravi Dhar, Douglas R. Lowy, and Esther H. Chang. Mechanism of activation of a human oncogene. *Nature*, 300(5888):143–149, 1982.

[20] A. G. Knudson. Mutation and Cancer: Statistical Study of Retinoblastoma. *Proceedings of the National Academy of Sciences*, 68(4):820–823, 1971.

[21] Inigo Martincorena and Peter J Campbell. Somatic mutation in cancer and normal cells. *Science*, 349(6255):961–968, 2015.

[22] Nimrat Chatterjee and Graham C Walker. Mechanisms of DNA Damage, Repair, and Mutagenesis. *Environmental and Molecular Mutagenesis*, 58(5):235–263, 2017.

[23] P J Hastings, James R Lupski, Susan M Rosenberg, and Grzegorz Ira. Mechanisms of change in gene copy number. *Nature Reviews Genetics*, 10(August):551–564, 2009.

[24] Aaron R. Quinlan and Ira M. Hall. Characterizing complex structural variation in germline and somatic genomes. *Trends in Genetics*, 28(1):43–53, 2012.

[25] Benjamin B. Currall, Colby Chiangmai, Michael E. Talkowski, and Cynthia C. Morton. Mechanisms for Structural Variation in the Human Genome. *Current Genetic Medicine Reports*, 1(2):81–90, 2013.

[26] Claudia M.B. Carvalho and James R. Lupski. Mechanisms underlying structural variant formation in genomic disorders. *Nature Reviews Genetics*, 17(4):224–238, 2016.

[27] Jose M C Tubio. Somatic structural variation and cancer. *Briefings in Functional Genomics*, 14(5):339–351, 2015.

[28] Kijong Yi and Young Seok Ju. Patterns and mechanisms of structural variations in human cancer. *Experimental & Molecular Medicine*, 50(8):98, 2018.

[29] En Li. Chromatin modification and epigenetic reprogramming in mammalian development. *Nature Reviews Genetics*, 3(9):662–673, 2002.

[30] Anna Portela and Manel Esteller. Epigenetic modifications and human disease. *Nature Biotechnology*, 28(10):1057–1068, 2010.

[31] John Cairns. Mutation selection and the natural history of cancer. *Nature*, 255(5505):197–200, 1975.

[32] Peter C. Nowell. The clonal evolution of tumor cell populations. *Science*, 194(4260):23–28, 1976.

[33] Douglas Hanahan and Robert A Weinberg. The Hallmarks of Cancer. *Cell*, 100(1):57–70, 2000.

[34] Douglas Hanahan and Robert A. Weinberg. Hallmarks of cancer: The next generation, 2011.

[35] June-Koo Lee, Yoon-La Choi, Mijung Kwon, and Peter J. Park. Mechanisms and Consequences of Cancer Genome Instability: Lessons from Genome Sequencing Studies. *Annual Review of Pathology: Mechanisms of Disease*, 11(1):283–312, 2016.

[36] Andrés Aguilera and Tatiana García-Muse. Causes of Genome Instability. *Annual Review of Genetics*, 47(1):1–32, 2013.

[37] A. Janssen and R. H. Medema. Genetic instability: Tipping the balance. *Oncogene*, 32(38):4459–4470, 2013.

[38] Sally M. Dewhurst, Nicholas McGranahan, Rebecca A. Burrell, Andrew J. Rowan, Eva Grönroos, David Endesfelder, Tejal Joshi, Dmitri Mouradov, Peter Gibbs, Robyn L. Ward, Nicholas J. Hawkins, Zoltan Szallasi, Oliver M. Sieber, and Charles Swanton. Tolerance of whole- genome doubling propagates chromosomal instability and accelerates cancer genome evolution. *Cancer Discovery*, 4(2):175–185, 2014.

[39] Kasey Rodgers and Mitch Mcvey. Error-Prone Repair of DNA Double-Strand Breaks. *Journal of Cellular Physiology*, 231(1):15–24, 2016.

[40] Ibiayi Dagogo-Jack and Alice T. Shaw. Tumour heterogeneity and resistance to cancer therapies. *Nature Reviews Clinical Oncology*, 15(2):81–94, 2018.

[41] P. Andrew Futreal, Lachlan Coin, Mhairi Marshall, Thomas Down, Timothy Hubbard, Richard Wooster, Nazneen Rahman, and Michael R. Stratton. A census of human cancer genes. *Nature Reviews Cancer*, 4(3):177–183, 2004.

[42] Simon A. Forbes, David Beare, Prasad Gunasekaran, Kenric Leung, Nidhi Bindal, Harry Boutselakis, Minjie Ding, Sally Bamford, Charlotte Cole, Sari Ward, Chai Yin Kok, Mingming Jia, Tisham De, Jon W. Teague, Michael R. Stratton, Ultan McDermott, and Peter J. Campbell. COSMIC: Exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Research*, 43(D1):D805–D811, 2015.

[43] Rohini Roy, Jarin Chun, and Simon N. Powell. BRCA1 and BRCA2: Different roles in a common pathway of genome protection, 2012.

[44] Yi Lu, Weronica E. Ek, David Whiteman, Thomas L. Vaughan, Amanda B. Spurdle, Douglas F. Easton, Paul D. Pharoah, Deborah J. Thompson, Alison M. Dunning, Nicholas K. Hayward, Georgia Chenevix-Trench, and Stuart Macgregor. Most common 'sporadic' cancers have a significant germline genetic component. *Human Molecular Genetics*, 23(22):6112–6118, 2014.

[45] Giovanni Ciriello, Martin L Miller, Bülent Arman Aksoy, Yasin Senbabaoglu, Nikolaus Schultz, and Chris Sander. Emerging landscape of oncogenic signatures across human cancers. *Nature Genetics*, 45(10):1127–1133, 2013.

[46] A. P Jason de Koning, Wanjun Gu, Todd A. Castoe, Mark A. Batzer, and David D. Pollock. Repetitive elements may comprise over Two-Thirds of the human genome. *PLoS Genetics*, 7(12):1–12, 2011.

[47] B. McClintock. The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences*, 36(6):344–355, jun 1950.

[48] AFA Smit, R Hubley, and P Green. RepeatMasker Open-4.0, 2019.

[49] Christine R. Beck, Pamela Collier, Catriona Macfarlane, Maika Malig, Jeffrey M. Kidd, Evan E. Eichler, Richard M. Badge, and John V. Moran. LINE-1 retrotransposition activity in human genomes. *Cell*, 141(7):1159–1170, 2010.

[50] John L. Goodier. Restricting retrotransposons: A review. *Mobile DNA*, 7(1), 2016.

[51] Sandra L Martin. Retrotransposons: On the move. *eLife*, 7:1–3, 2018.

[52] Stéphane Boissinot and Anthony V. Furano. Adaptive evolution in LINE-1 retrotransposons. *Molecular Biology and Evolution*, 18(12):2186–2194, 2001.

[53] Hameed Khan, Arian Smit, and Stéphane Boissinot. Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. *Genome Research*, 16(1):78–87, 2006.

[54] Donna M. Sassaman, Beth A. Dombroski, John V. Moran, Michelle L. Kimberland, Thierry P. Naas, Ralph J. DeBerardinis, Abram Gabriel, Gary D. Swergold, and Haig H. Kazazian. Many human L1 elements are capable of retrotransposition. *Nature Genetics*, 16(1):37–43, 1997.

[55] Brook Brouha, Joshua Schustak, Richard M Badge, Sheila Lutz-Prigge, Alexander H Farley, John V Moran, and Haig H Kazazian. Hot L1s account for the bulk of retrotransposition in the human population. *Proceedings of the National Academy of Sciences*, 100(9):5280–5285, 2003.

[56] Dustin C. Hancks and Haig H. Kazazian. Roles for retrotransposon insertions in human disease. *Mobile DNA*, 7(1), 2016.

[57] James M. Heather and Benjamin Chain. The sequence of sequencers: The history of sequencing DNA. *Genomics*, 107(1):1–8, 2016.

[58] F. S. Collins, E. S. Lander, J. Rogers, and R. H. Waterson. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, 2004.

[59] Eric D. Green, James D. Watson, and Francis S. Collins. Human Genome Project: Twenty-five years of big biology. *Nature*, 526(7571):29–31, 2015.

[60] HumanGenomeProject. The Human Genome Project Completion: Frequently Asked Questions, 2019.

[61] Illumina. History of sequencing by synthesis. https://emea.illumina.com/science/technology/next-generation-sequencing/illumina-sequencing-history.html, 2019.

[62] Erika Check Hayden. Is the $1,000 genome for real? *Nature*, 2014.

[63] Illumina. Indexed Sequencing: Overview Guide. http://emea.support.illumina.com/content/dam/illumina-support/documents/documentation/system_documentation/miseq/indexed-sequencing-overview-guide-15057455-04.pdf, 2018.

[64] Elaine R. Mardis. Next-Generation DNA Sequencing Methods. *Annual Review of Genomics and Human Genetics*, 9(1):387–402, 2008.

[65] Michael L. Metzker. Sequencing technologies the next generation. *Nature Reviews Genetics*, 11(1):31–46, 2010.

[66] Randall K. Saiki, Stephen Scharf, Fred Faloona, Kary B. Mullis, Glenn T. Horn, Henry A. Erlich, and Norman Arnheim. Enzymatic amplification of β-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science*, 230(4732):1350–1354, 1985.

[67] Kary B. Mullis and Fred A. Faloona. Specific Synthesis of DNA in Vitro via a Polymerase-Catalyzed Chain Reaction. *Methods in Enzymology*, 155(C):335–350, 1987.

[68] Randall K. Saiki, David H. Gelfand, Susanne Stoffel, Stephen J. Scharf, Russell Higuchi, Glenn T. Horn, Kary B. Mullis, and Henry A. Erlich. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, 239(4839):487–491, 1988.

[69] John M.S. Bartlett and David Stirling. A Short History of the Polymerase Chain Reaction. In *PCR Protocols*, pages 3–6. 2003.

[70] G. Stolovitzky and G. Cecchi. Efficiency of DNA replication in the polymerase chain reaction. *Proceedings of the National Academy of Sciences*, 93(23):12947–12952, 1996.

[71] Nadia Lalam. Estimation of the reaction efficiency in polymerase chain reaction. *Journal of Theoretical Biology*, 242(4):947–953, 2006.

[72] Rita S. Cha and William G. Thilly. Specificity, efficiency, and fidelity of PCR. *Genome Research*, 3(3), 1993.

[73] Weihong Liu and David A. Saint. Validation of a quantitative method for real time PCR kinetics. *Biochemical and Biophysical Research Communications*, 294(2):347–353, 2002.

[74] Vladimir Potapov and Jennifer L. Ong. Examining sources of error in PCR by single-molecule sequencing. *PLoS ONE*, 12(1), 2017.

[75] I. Kinde, J. Wu, N. Papadopoulos, K. W. Kinzler, and B. Vogelstein. Detection and quantification of rare mutations with massively parallel sequencing. *Proceedings of the National Academy of Sciences*, 108(23):9530–9535, 2011.

[76] Teemu Kivioja, Anna Vähärautio, Kasper Karlsson, Martin Bonke, Martin Enge, Sten Linnarsson, and Jussi Taipale. Counting absolute numbers of molecules using unique molecular identifiers. *Nature Methods*, 9(1):72–74, 2012.

[77] Shu Mei Teo, Yudi Pawitan, Chee Seng Ku, Kee Seng Chia, and Agus Salim. Statistical challenges associated with detecting copy number variations with next-generation sequencing. *Bioinformatics*, 28(21):2711–2718, 2012.

[78] Min Zhao, Qingguo Wang, Quan Wang, Peilin Jia, and Zhongming Zhao. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: Features and perspectives. *BMC Bioinformatics*, 14(SUPPL11), 2013.

[79] Biao Liu, Jeffrey M. Conroy, Carl D. Morrison, Adekunle O. Odunsi, Maochun Qin, Lei Wei, Donald L. Trump, Candace S. Johnson, Song Liu, Jianmin Wang, Biao Liu, Jeffrey M. Conroy, Carl D. Morrison, Adekunle O. Odunsi, Maochun Qin, Lei Wei, Donald L. Trump, Candace S. Johnson, Song Liu, and Jianmin Wang. Structural variation discovery in the cancer genome using next generation sequencing: Computational solutions and perspectives. *Oncotarget*, 6(8):5477–5489, 2015.

[80] David M. Hyman, Barry S. Taylor, and José Baselga. Implementing Genome-Driven Oncology. *Cell*, 168(4):584–599, 2017.

[81] Brian J Druker, François Guilhot, Stephen G O'Brien, Insa Gathmann, Hagop Kantarjian, Norbert Gattermann, Michael W N Deininger, Richard T Silver, John M Goldman, Richard M Stone, Francisco Cervantes, Andreas Hochhaus, Bayard L Powell, Janice L Gabrilove, Philippe Rousselot, Josy Reiffers, Jan J Cornelissen, Timothy Hughes, Hermine Agis, Thomas Fischer, Gregor Verhoef, John Shepherd, Giuseppe Saglio, Alois Gratwohl, Johan L Nielsen, Jerald P Radich, Bengt Simonsson, Kerry Taylor, Michele Baccarani, Charlene So, Laurie Letvak, and Richard A Larson. Five-Year Follow-up of Patients Receiving Imatinib for Chronic Myeloid Leukemia. *New England Journal of Medicine*, 355(23):2408–2417, 2006.

[82] Hannah Bower, Magnus Björkholm, Paul W. Dickman, Martin Höglund, Paul C. Lambert, and Therese M.L. Andersson. Life expectancy of patients with chronic myeloid leukemia approaches the life expectancy of the general population. *Journal of Clinical Oncology*, 34(24):2851–2857, 2016.

[83] Dennis J Slamon, Brian Leyland-Jones, Steven Shak, Hank Fuchs, Virginia Paton, Alex Bajamonde, Thomas Fleming, Wolfgang Eiermann, Janet Wolter, Mark Pegram, Jose Baselga, and Larry Norton. Use of Chemotherapy plus a Monoclonal Antibody against HER2 for Metastatic Breast Cancer That Overexpresses HER2. *New England Journal of Medicine*, 344(11):783–792, 2001.

[84] Sandra M. Swain, José Baselga, Sung-Bae Kim, Jungsil Ro, Vladimir Semiglazov, Mario Campone, Eva Ciruelos, Jean-Marc Ferrero, Andreas Schneeweiss, Sarah Heeson, Emma Clark, Graham Ross, Mark C. Benyunes, and Javier Cortés. Pertuzumab, Trastuzumab, and Docetaxel in HER2-Positive Metastatic Breast Cancer. *New England Journal of Medicine*, 372(8):724–734, 2015.

[85] MJ Piccart-Gebhart and M Procter. Trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer. *The New England journal of medicine*, 353(16):1659–1672, 2005.

[86] Joan Minguet, Katherine H. Smith, and Peter Bramlage. Targeted therapies for treatment of non-small cell lung cancer - Recent advances and future perspectives, 2016.

[87] Gideon Bollag, James Tsai, Jiazhong Zhang, Chao Zhang, Prabha Ibrahim, Keith Nolop, and Peter Hirth. Vemurafenib: The first drug approved for BRAF-mutant cancer, 2012.

[88] Wendy De Roock, Bart Claes, David Bernasconi, Jef De Schutter, Bart Biesmans, George Fountzilas, Konstantine T. Kalogeras, Vassiliki Kotoula, Demetris Papamichael, Pierre Laurent-Puig, Frédérique Penault-Llorca, Philippe Rougier, Bruno Vincenzi, Daniele Santini, Giuseppe Tonini, Federico Cappuzzo, Milo Frattini, Francesca Molinari, Piercarlo Saletti, Sara De Dosso, Miriam Martini, Alberto Bardelli, Salvatore Siena, Andrea Sartore-Bianchi, Josep Tabernero, Teresa Macarulla, Frédéric Di Fiore, Alice Oden Gangloff, Fortunato Ciardiello, Per Pfeiffer, Camilla Qvortrup, Tine Plato Hansen, Eric Van Cutsem, Hubert Piessevaux, Diether Lambrechts, Mauro Delorenzi, and Sabine Tejpar. Effects of KRAS, BRAF, NRAS, and PIK3CA mutations on the efficacy of cetuximab plus chemotherapy in chemotherapy-refractory metastatic colorectal cancer: A retrospective consortium analysis. *The Lancet Oncology*, 11(8):753–762, 2010.

[89] Bella Kaufman, Ronnie Shapira-Frommer, Rita K. Schmutzler, M. William Audeh, Michael Friedlander, Judith Balmaña, Gillian Mitchell, Georgeta Fried, Salomon M. Stemmer, Ayala Hubert, Ora Rosengarten, Mariana Steiner, Niklas Loman, Karin Bowen, Anitra Fielding, and Susan M. Domchek. Olaparib monotherapy in patients with advanced cancer and a germline BRCA1/2 mutation. *Journal of Clinical Oncology*, 33(3):244–250, 2015.

[90] Katherine A. Hoadley, Christina Yau, Denise M. Wolf, Andrew D. Cherniack, David Tamborero, Sam Ng, Max D.M. Leiserson, Beifang Niu, Michael D. McLellan, Vladislav Uzunangelov, Jiashan Zhang, Cyriac Kandoth, Rehan Akbani, Hui Shen, Larsson Omberg, Andy Chu, Adam A. Margolin, Laura J. Van't Veer, Nuria Lopez-Bigas, Peter W. Laird, Benjamin J. Raphael, Li Ding, A. Gordon Robertson, Lauren A. Byers, Gordon B. Mills, John N. Weinstein, Carter Van Waes, Zhong Chen, Eric A. Collisson, Christopher C. Benz, Charles M. Perou, Joshua M. Stuart, Rachel Abbott, Scott Abbott, B. Arman Aksoy, Kenneth Aldape, Adrian Ally, Samirkumar Amin, Dimitris Anastassiou, J. Todd Auman, Keith A. Baggerly, Miruna Balasundaram, Saianand Balu, Stephen B. Baylin, Stephen C. Benz, Benjamin P. Berman, Brady Bernard, Ami S. Bhatt, Inanc Birol, Aaron D. Black, Tom Bodenheimer, Moiz S. Bootwalla, Jay Bowen, Ryan Bressler, Christopher A. Bristow, Angela N. Brooks, Bradley Broom, Elizabeth Buda, Robert Burton, Yaron S.N. Butterfield, Daniel Carlin, Scott L. Carter, Tod D. Casasent, Kyle Chang, Stephen Chanock, Lynda Chin, Dong Yeon Cho, Juok Cho, Eric Chuah, Hye Jung E. Chun, Kristian Cibulskis, Giovanni Ciriello, James Cleland, Melisssa Cline, Brian Craft, Chad J. Creighton, Ludmila Danilova, Tanja Davidsen, Caleb Davis, Nathan D. Dees, Kim Delehaunty, John A. Demchok, Noreen Dhalla, Daniel DiCara, Huyen Dinh, Jason R. Dobson, Deepti Dodda, Harsha Vardhan Doddapaneni, Lawrence Donehower, David J. Dooling, Gideon Dresdner, Jennifer Drummond, Andrea Eakin, Mary Edgerton, Jim M. Eldred, Greg Eley, Kyle Ellrott, Cheng Fan, Suzanne Fei, Ina Felau, Scott Frazer, Samuel S. Freeman, Jessica Frick, Catrina C. Fronick, Lucinda L. Fulton, Robert Fulton, Stacey B. Gabriel, Jianjiong Gao, Julie M. Gastier-Foster, Nils Gehlenborg, Myra George, Gad Getz, Richard Gibbs, Mary Goldman, Abel Gonzalez-Perez, Benjamin Gross, Ranabir Guin, Preethi Gunaratne, Angela Hadjipanayis, Mark P. Hamilton, Stanley R. Hamilton, Leng Han, Yi Han, Hollie A. Harper, Psalm Haseley, David Haussler, D. Neil Hayes, David I. Heiman, Elena Helman, Carmen Helsel, Shelley M. Herbrich, James G. Herman, Toshinori Hinoue, Carrie Hirst, Martin Hirst, Robert A. Holt, Alan P. Hoyle, Lisa Iype, Anders Jacobsen, Stuart R. Jeffreys, Mark A. Jensen, Corbin D. Jones, Steven J.M. Jones, Zhenlin Ju, Joonil Jung, Andre Kahles, Ari Kahn, Joelle Kalicki-Veizer, Divya Kalra, Krishna Latha

Kanchi, David W. Kane, Hoon Kim, Jaegil Kim, Theo Knijnenburg, Daniel C. Koboldt, Christie Kovar, Roger Kramer, Richard Kreisberg, Raju Kucherlapati, Marc Ladanyi, Eric S. Lander, David E. Larson, Michael S. Lawrence, Darlene Lee, Eunjung Lee, Semin Lee, William Lee, Kjong Van Lehmann, Kalle Leinonen, Kristen M. Leraas, Seth Lerner, Douglas A. Levine, Lora Lewis, Timothy J. Ley, Haiyan I. Li, Jun Li, Wei Li, Han Liang, Tara M. Lichtenberg, Jake Lin, Ling Lin, Pei Lin, Wenbin Liu, Yingchun Liu, Yuexin Liu, Philip L. Lorenzi, Charles Lu, Yiling Lu, Lovelace J. Luquette, Singer Ma, Vincent J. Magrini, Harshad S. Mahadeshwar, Elaine R. Mardis, Marco A. Marra, Michael Mayo, Cynthia McAllister, Sean E. McGuire, Joshua F. McMichael, James Melott, Shaowu Meng, Matthew Meyerson, Piotr A. Mieczkowski, Christopher A. Miller, Martin L. Miller, Michael Miller, Richard A. Moore, Margaret Morgan, Donna Morton, Lisle E. Mose, Andrew J. Mungall, Donna Muzny, Lam Nguyen, Michael S. Noble, Houtan Noushmehr, Michelle O'Laughlin, Akinyemi I. Ojesina, Tai Hsien Ou Yang, Brad Ozenberger, Angeliki Pantazi, Michael Parfenov, Peter J. Park, Joel S. Parker, Evan Paull, Chandra Sekhar Pedamallu, Todd Pihl, Craig Pohl, David Pot, Alexei Protopopov, Teresa Przytycka, Amie Radenbaugh, Nilsa C. Ramirez, Ricardo Ramirez, Gunnar Rätsch, Jeffrey Reid, Xiaojia Ren, Boris Reva, Sheila M. Reynolds, Suhn K. Rhie, Jeffrey Roach, Hector Rovira, Michael Ryan, Gordon Saksena, Sofie Salama, Chris Sander, Netty Santoso, Jacqueline E. Schein, Heather Schmidt, Nikolaus Schultz, Steven E. Schumacher, Jonathan Seidman, Yasin Senbabaoglu, Sahil Seth, Samantha Sharpe, Ronglai Shen, Margi Sheth, Yan Shi, Ilya Shmulevich, Grace O. Silva, Janae V. Simons, Rileen Sinha, Payal Sipahimalani, Scott M. Smith, Heidi J. Sofia, Artem Sokolov, Mathew G. Soloway, Xingzhi Song, Carrie Sougnez, Paul Spellman, Louis Staudt, Chip Stewart, Petar Stojanov, Xiaoping Su, S. Onur Sumer, Yichao Sun, Teresa Swatloski, Barbara Tabak, Angela Tam, Donghui Tan, Jiabin Tang, Roy Tarnuzzer, Barry S. Taylor, Nina Thiessen, Vesteinn Thorsson, Timothy Triche, David J. Van Den Berg, Fabio Vandin, Richard J. Varhol, Charles J. Vaske, Umadevi Veluvolu, Roeland Verhaak, Doug Voet, Jason Walker, John W. Wallis, Peter Waltman, Yunhu Wan, Min Wang, Wenyi Wang, Zhining Wang, Scot Waring, Nils Weinhold, Daniel J. Weisenberger, Michael C. Wendl, David Wheeler, Matthew D. Wilkerson, Richard K. Wilson, Lisa Wise, Andrew Wong, Chang Jiun Wu, Chia Chin Wu, Hsin Ta Wu, Junyuan Wu, Todd Wylie, Liu Xi, Ruibin Xi, Zheng Xia, Andrew W. Xu, Da Yang, Liming Yang, Lixing Yang, Yang Yang, Jun Yao, Rong Yao, Kai Ye, Kosuke Yoshihara, Yuan Yuan, Alfred K. Yung, Travis Zack, Dong Zeng, Jean Claude Zenklusen, Hailei Zhang, Jianhua Zhang, Nianxiang Zhang, Qunyuan Zhang, Wei Zhang, Wei Zhao, Siyuan Zheng, Jing Zhu, Erik Zmuda, and Lihua Zou. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell*, 158(4):929–944, 2014.

[91] Aaron N. Hata, Matthew J. Niederst, Hannah L. Archibald, Maria Gomez-Caraballo, Faria M. Siddiqui, Hillary E. Mulvey, Yosef E. Maruvka, Fei Ji, Hyo Eun C. Bhang, Viveksagar Krishnamurthy Radhakrishna, Giulia Siravegna, Haichuan Hu, Sana Raoof, Elizabeth Lockerman, Anuj Kalsy, Dana Lee, Celina L. Keating, David A. Ruddy, Leah J. Damon, Adam S. Crystal, Carlotta Costa, Zofia Piotrowska, Alberto Bardelli, Anthony J. Iafrate, Ruslan I. Sadreyev, Frank Stegmeier, Gad Getz, Lecia V. Sequist, Anthony C. Faber, and Jeffrey A. Engelman. Tumor cells can follow distinct evolutionary paths to become resistant to epidermal growth factor receptor inhibition. *Nature Medicine*, 22(3):262–269, 2016.

[92] Nicholas C. Turner, Jungsil Ro, Cynthia Huang Bartlett, Sunil Verma, Maria Koehler, Sophia Randolph, Massimo Cristofanilli, Nadia Harbeck, Sibylle Loibl, Hiroji Iwata, Ke Zhang, Fabrice André, Sherene Loi, and Carla Giorgetti. Palbociclib in Hormone-

Receptor–Positive Advanced Breast Cancer. *New England Journal of Medicine*, 373(3):209–219, 2015.

[93] Rameen Beroukhim, Craig H Mermel, Dale Porter, Guo Wei, Soumya Raychaudhuri, Jerry Donovan, Jordi Barretina, Jesse S Boehm, Jennifer Dobson, Mitsuyoshi Urashima, Kevin T Mc Henry, Reid M Pinchback, Azra H Ligon, Yoon-jae Cho, Leila Haery, Heidi Greulich, Michael Reich, Wendy Winckler, Michael S Lawrence, Barbara A Weir, Kumiko E Tanaka, Derek Y Chiang, Adam J Bass, Alice Loo, Carter Hoffman, John Prensner, Ted Liefeld, Qing Gao, Derek Yecies, Sabina Signoretti, Elizabeth Maher, Frederic J Kaye, Hidefumi Sasaki, Joel E Tepper, Jonathan A Fletcher, Ming-sound Tsao, Francesca Demichelis, Mark A Rubin, Pasi A Janne, Josep Tabernero, Mark J Daly, Carmelo Nucera, Ross L Levine, Benjamin L Ebert, Stacey Gabriel, Anil K Rustgi, Cristina R Antonescu, Marc Ladanyi, Anthony Letai, Levi A Garraway, Massimo Loda, David G Beer, Lawrence D True, Aikou Okamoto, Scott L Pomeroy, Samuel Singer, Todd R Golub, Eric S Lander, Gad Getz, William R Sellers, and Matthew Meyerson. The landscape of somatic copy-number alteration across human cancers. *Nature*, 463(February):899–905, 2010.

[94] Andy W. Pang, Jeffrey R. MacDonald, Dalila Pinto, John Wei, Muhammad A. Rafiq, Donald F. Conrad, Hansoo Park, Matthew E. Hurles, Charles Lee, J. Craig Venter, Ewen F. Kirkness, Samuel Levy, Lars Feuk, and Stephen W. Scherer. Towards a comprehensive structural variation map of an individual human genome, 2010.

[95] John Shaughnessy. Amplification and overexpression of CKS1B at chromosome band 1q21 is associated with reduced levels of p27 Kip1 and an aggressive clinical course in multiple myeloma. *Hematology*, 10(SUPPL. 1):117–126, 2005.

[96] Mei Hsi Chen, Connie Qi, Donna Reece, and Hong Chang. Cyclin kinase subunit 1B nuclear expression predicts an adverse outcome for patients with relapsed/refractory multiple myeloma treated with bortezomib. *Human Pathology*, 43(6):858–864, 2012.

[97] Sweta Mishra and Johnathan R Whetstine. Different Facets of Copy Number Changes: Permanent, Transient, and Adaptive. *Molecular and Cellular Biology*, 36(7):1050–1063, 2016.

[98] Dariush Etemadmoghadam, George Au-Yeung, Meaghan Wall, Chris Mitchell, Maya Kansara, Elizabeth Loehrer, Crisoula Batzios, Joshy George, Sarah Ftouni, Barbara A. Weir, Scott Carter, Irma Gresshoff, Linda Mileshkin, Danny Rischin, William C. Hahn, Paul M. Waring, Gad Getz, Carleen Cullinane, Lynda J. Campbell, and David D. Bowtell. Resistance to CDK2 inhibitors is associated with selection of polyploid cells in CCNE1-amplified ovarian cancer. *Clinical Cancer Research*, 19(21):5960–5971, 2013.

[99] Geoff Macintyre, Bauke Ylstra, and James D Brenton. Sequencing Structural Variants in Cancer for Precision Therapeutics. *Trends in Genetics*, 32(9):530–542, 2016.

[100] P. Lao-Sirieix, B. Rous, M. O'Donovan, R. H Hardwick, I. Debiram, and R. C Fitzgerald. Non-endoscopic immunocytological screening test for Barrett's oesophagus. *Gut*, 56(7):1033–1034, jul 2007.

[101] Jonathan C.M. Wan, Charles Massie, Javier Garcia-Corbacho, Florent Mouliere, James D. Brenton, Carlos Caldas, Simon Pacey, Richard Baird, and Nitzan Rosenfeld. Liquid biopsies come of age: Towards implementation of circulating tumour DNA. *Nature Reviews Cancer*, 17(4):223–238, 2017.

[102] Frank Diehl, Meng Li, Devin Dressman, Yiping He, Dong Shen, Steve Szabo, Luis A Diaz, Steven N Goodman, Kerstin A David, Hartmut Juhl, Kenneth W Kinzler, and Bert Vogelstein. Detection and quantification of mutations in the plasma of patients with colorectal tumors. *PNAS*, 102(45):16368–16373, 2005.

[103] Dvir Aran, Marina Sirota, and Atul J. Butte. Systematic pan-cancer analysis of tumour purity. *Nature Communications*, 6(1):8971, dec 2015.

[104] Isaac Kinde, Nickolas Papadopoulos, Kenneth W. Kinzler, and Bert Vogelstein. FAST-SeqS: A simple and efficient method for the detection of aneuploidy by massively parallel sequencing. *PLoS ONE*, 7(7), 2012.

[105] S Abujudeh, S S Zeki, M C V Van Lanschot, M Pusung, J M J Weaver, X Li, A J Metz, J Bornschein, L Bower, A Miremadi, R C Fitzgerald, E R Morrissey, A G Lynch, Oesophageal Cancer Clinical Consortium, and Molecular Stratification (OCCAMS). Low-cost and clinically applicable copy number profiling using repeat DNA. *bioRxiv*, 2018.

[106] Jelena Belic, Marina Koch, Peter Ulz, Martina Auer, Teresa Gerhalter, Sumitra Mohan, Katja Fischereder, Edgar Petru, Thomas Bauernhofer, Jochen B. Geigl, Michael R. Speicher, and Ellen Heitzer. Rapid identification of plasma DNA samples with increased ctDNA levels by a modified FAST-SeqS approach. *Clinical Chemistry*, 61(6):838–849, 2015.

[107] Christopher Douville, Simeon Springer, Isaac Kinde, Joshua D. Cohen, Ralph H. Hruban, Anne Marie Lennon, Nickolas Papadopoulos, Kenneth W. Kinzler, Bert Vogelstein, and Rachel Karchin. Detection of aneuploidy in patients with cancer through amplification of long interspersed nucleotide elements (LINEs). *Proceedings of the National Academy of Sciences*, 115(8):201717846, 2018.

[108] Sam Abujudeh. samabs/conliga v0.1.0. oct 2018.

[109] S. Volik, M. Alcaide, R. D. Morin, and C. Collins. Cell-free DNA (cfDNA): Clinical Significance and Utility in Cancer Shaped By Emerging Technologies. *Molecular Cancer Research*, 14(10):898–908, 2016.

[110] Florent Mouliere, Dineika Chandrananda, Anna M. Piskorz, Elizabeth K. Moore, James Morris, Lise Barlebo Ahlborn, Richard Mair, Teodora Goranova, Francesco Marass, Katrin Heider, Jonathan C.M. Wan, Anna Supernat, Irena Hudecova, Ioannis Gounaris, Susana Ros, Mercedes Jimenez-Linan, Javier Garcia-Corbacho, Keval Patel, Olga Østrup, Suzanne Murphy, Matthew D. Eldridge, Davina Gale, Grant D. Stewart, Johanna Burge, Wendy N. Cooper, Michiel S. Van Der Heijden, Charles E. Massie, Colin Watts, Pippa Corrie, Simon Pacey, Kevin M. Brindle, Richard D. Baird, Morten Mau-Sørensen, Christine A. Parkinson, Christopher G. Smith, James D. Brenton, and Nitzan Rosenfeld. Enhanced detection of circulating tumor DNA by fragment size analysis. *Science Translational Medicine*, 10(466), 2018.

[111] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3), 2009.

[112] Ellen Heitzer, Peter Ulz, Jelena Belic, Stefan Gutschi, Franz Quehenberger, Katja Fischereder, Theresa Benezeder, Martina Auer, Carina Pischler, Sebastian Mannweiler, Martin Pichler, Florian Eisner, Martin Haeusler, Sabine Riethdorf, Klaus Pantel,

Hellmut Samonigg, Gerald Hoefler, Herbert Augustin, Jochen B. Geigl, and Michael R. Speicher. Tumor-associated copy number changes in the circulation of patients with prostate cancer identified through whole-genome sequencing. *Genome Medicine*, 5(4), 2013.

[113] Goncalo R Abecasis, Adam Auton, Lisa D Brooks, Mark A DePristo, Richard Durbin, Robert E Handsaker, Hyun Min Kang, Gabor T Marth, and Gil A McVean. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.

[114] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4):357–9, 2012.

[115] Travis I Zack, Steven E Schumacher, Scott L Carter, Andrew D Cherniack, Gordon Saksena, Barbara Tabak, Michael S Lawrence, Cheng-zhong Zhang, Jeremiah Wala, Craig H Mermel, Carrie Sougnez, Stacey B Gabriel, Bryan Hernandez, Hui Shen, Peter W Laird, and Gad Getz. Pan-cancer patterns of somatic copy number alteration. *Nature Genetics*, 45(10):1134–1140, 2013.

[116] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2009.

[117] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), 2014.

[118] Erik Garrison and Gabor Marth. Haplotype-based variant detection from short-read sequencing. *arXiv*, page 9, 2012.

[119] Malvina Josephidou, Andy G. Lynch, and Simon Tavaré. MultiSNV: A probabilistic approach for improving detection of somatic point mutations from multiple related tumour samples. *Nucleic Acids Research*, 43(9), 2015.

[120] Stephanie C Hicks and Rafael A Irizarry. When to use Quantile Normalization? *bioRxiv*, 2014.

[121] Illumina. What is nucleotide diversity and why is it important? https://support.illumina.com/bulletins/2016/07/what-is-nucleotide-diversity-and-why-is-it-important.html, 2017.

[122] Gordon Brown. demuxFQ tool, 2019.

[123] Michael Lawrence, Wolfgang Huber, Hervé Pagès, Patrick Aboyoun, Marc Carlson, Robert Gentleman, Martin T. Morgan, and Vincent J. Carey. Software for Computing and Annotating Genomic Ranges. *PLoS Computational Biology*, 9(8), 2013.

[124] L. A. Clarke, C. S. Rebelo, J. Gonçalves, M. G. Boavida, and P. Jordan. PCR amplification introduces errors into mononucleotide and dinucleotide repeat sequences. *Journal of Clinical Pathology - Molecular Pathology*, 54(5):351–353, 2001.

[125] Erik Scott Wright and Kalin Horen Vetsigian. Quality filtering of Illumina index reads mitigates sample cross-talk. *BMC Genomics*, 17(1), 2016.

[126] Leonard E Baum and Ted Petrie. Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *The Annals of Mathematical Statistics*, 37(6):1554–1563, 1966.

[127] Leonard E Baum and J A Eagon. An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. *Bulletin of the American Mathematical Society*, 73(3):360–363, 1967.

[128] Tobias Rydén. Em versus Markov chain monte carlo for estimation of hidden Markov models: A computational perspective. *Bayesian Analysis*, 3(4):659–688, 2008.

[129] Matthew J. Beal, Zoubin Ghahramani, and Carl Edward Rasmussen. The Infinite Hidden Markov Model. In *Advances in Neural Information Processing Systems 14*, pages 577–585, 2002.

[130] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[131] Emily B. Fox, Erik B. Sudderth, Michael I. Jordan, and Alan S. Willsky. A Sticky HDP-HMM with application to speaker diarization. *The Annals of Applied Statistics*, 5(2):1020–1056, 2011.

[132] Yee Whye Teh. Dirichlet Process. In *Encyclopedia of Machine Learning*, pages 280–287. Springer, 2010.

[133] Matt Wand. KernSmooth. Functions for Kernel Smoothing Supporting Wand & Jones (1995). *R package version 2.23-15.*, 2:19–22, 2015.

[134] Kevin P Murphy. *Machine Learning: A Probabilistic Perspective.* The MIT Press, 2012.

[135] Richard Redon, Shumpei Ishikawa, Karen R. Fitch, Lars Feuk, George H. Perry, T. Daniel Andrews, Heike Fiegler, Michael H. Shapero, Andrew R. Carson, Wenwei Chen, Eun Kyung Cho, Stephanie Dallaire, Jennifer L. Freeman, Juan R. González, Mònica Gratacòs, Jing Huang, Dimitrios Kalaitzopoulos, Daisuke Komura, Jeffrey R. MacDonald, Christian R. Marshall, Rui Mei, Lyndal Montgomery, Kunihiro Nishimura, Kohji Okamura, Fan Shen, Martin J. Somerville, Joelle Tchinda, Armand Valsesia, Cara Woodwark, Fengtang Yang, Junjun Zhang, Tatiana Zerjal, Jane Zhang, Lluis Armengol, Donald F. Conrad, Xavier Estivill, Chris Tyler-Smith, Nigel P. Carter, Hiroyuki Aburatani, Charles Lee, Keith W. Jones, Stephen W. Scherer, and Matthew E. Hurles. Global variation in copy number in the human genome. *Nature*, 444(7118):444–454, 2006.

[136] Mehdi Zarrei, Jeffrey R. MacDonald, Daniele Merico, and Stephen W. Scherer. A copy number variation map of the human genome. *Nature Reviews Genetics*, 16(3):172–183, 2015.

[137] G E P Box. Robustness in the strategy of scientific model building. In *Robustness in Statistics*, pages 201–236. 1979.

[138] Hemant Ishwaran and Mahmoud Zarepour. Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika*, 87(2):371–390, 2000.

[139] Hemant Ishwaran and Mahmoud Zarepour. Exact and Approximate Sum Representations for the Dirichlet Process. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 30(2):269–283, 2002.

[140] Lawrence R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[141] Matthew Stephens. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society B*, 62(4):795–809, 2000.

[142] A. Jasra, C. C. Holmes, and D. A. Stephens. Markov Chain Monte Carlo Methods and the Label Switching Problem in Bayesian Mixture Modeling. *Statistical Science*, 20(1):50–67, 2005.

[143] S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

[144] James Munkres. Algorithms for the Assignment and Transportation Problems. *Journal of the Society for Industrial and Applied Mathematics*, 5(1):32–38, 1957.

[145] Audrey Qiuyan Fu, Steven Russell, Sarah J. Bray, and Simon Tavaré. Bayesian clustering of replicated time-course gene expression data with weak signals. *Annals of Applied Statistics*, 7(3):1334–1361, 2013.

[146] Dirk Eddelbuettel and Romain François. Rcpp: Seamless R and C++ Integration. *Journal of Statistical Software*, 40(8):1–18, 2011.

[147] Dirk Eddelbuettel and Conrad Sanderson. RcppArmadillo: Accelerating R with high-performance C++ linear algebra. *Computational Statistics and Data Analysis*, 71(March):1054–1063, 2014.

[148] J. A. Traherne. Human MHC architecture and evolution: Implications for disease association studies, 2008.

[149] Matthew J. Johnson and Alan S. Willsky. The Hierarchical Dirichlet Process Hidden semi-Markov Model. In *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence, UAI 2010*, pages 252–259, 2010.

[150] Mary Kathryn Cowles and Bradley P. Carlin. Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. *Journal of the American Statistical Association*, 91(434):883–904, 1996.

[151] Melina Arnold, Isabelle Soerjomataram, Jacques Ferlay, and David Forman. Global incidence of oesophageal cancer by histological subtype in 2012. *Gut*, 64(3):381–387, 2015.

[152] Elizabeth C. Smyth, Jesper Lagergren, Rebecca C. Fitzgerald, Florian Lordick, Manish A. Shah, Pernilla Lagergren, and David Cunningham. Oesophageal cancer. *Nature Reviews Disease Primers*, 3:17048, jul 2017.

[153] Philip R. Taylor, Christian C. Abnet, and Sanford M. Dawsey. Squamous dysplasia-The precursor lesion for esophageal squamous cell carcinoma. *Cancer Epidemiology Biomarkers and Prevention*, 22(4):540–552, 2013.

[154] Gina D. Tran, Xiu Di Sun, Christian C. Abnet, Jin Hu Fan, Sanford M. Dawsey, Zhi Wei Dong, Steven D. Mark, You Lin Qiao, and Philip R. Taylor. Prospective study of risk factors for esophageal and gastric cancers in the Linxian General Population Trial cohort in China. *International Journal of Cancer*, 113(3):456–463, 2005.

[155] Neal D. Freedman, Christian C. Abnet, Michael F. Leitzmann, Traci Mouw, Amy F. Subar, Albert R. Hollenbeck, and Arthur Schatzkin. A prospective study of tobacco, alcohol, and the risk of esophageal and gastric cancer subtypes. *American Journal of Epidemiology*, 165(12):1424–1433, 2007.

[156] Katrina F. Trivers, Susan A. Sabatino, and Sherri L. Stewart. Trends in esophageal cancer incidence by histology, United States, 1998-2003. *International Journal of Cancer*, 123(6):1422–1428, 2008.

[157] J. Zhao, Y.-T. He, R.-S. Zheng, S.-W. Zhang, and W.-Q. Chen. Analysis of esophageal cancer time trends in china, 1989-2008. *Asian Pacific Journal of Cancer Prevention*, 13(9), 2012.

[158] Gustaf Edgren, Hans Olov Adami, Elisabete Weiderpass Vainio, and Olof Nyrén. A global assessment of the oesophageal adenocarcinoma epidemic. *Gut*, 62(10):1406–1414, 2013.

[159] Michelle D. Althuis, Jaclyn M. Dozier, William F. Anderson, Susan S. Devesa, and Louise A. Brinton. Global trends in breast cancer incidence and mortality 1973-1997. *International Journal of Epidemiology*, 34(2):405–412, 2005.

[160] Xia Wang, Hong Ouyang, Yusuke Yamamoto, Pooja Ashok Kumar, Tay Seok Wei, Rania Dagher, Matthew Vincent, Xin Lu, Andrew M. Bellizzi, Khek Yu Ho, Christopher P. Crum, Wa Xian, and Frank McKeon. Residual embryonic cells as precursors of a Barrett's-like metaplasia. *Cell*, 145(7):1023–1035, 2011.

[161] Lesley A. Anderson, R. G.Peter Watson, Seamus J. Murphy, Brian T. Johnston, Harry Comber, Jim McGuigan, John V. Reynolds, and Liam J. Murray. Risk factors for Barrett's oesophagus and oesophageal adenocarcinoma: Results from the FINBAR study. *World Journal of Gastroenterology*, 13(10):1585–1594, 2007.

[162] D. J. Stein, H. B. El-Serag, J. Kuczynski, J. R. Kramer, and R. E. Sampliner. The association of body mass index with Barrett's oesophagus. *Alimentary Pharmacology and Therapeutics*, 22(10):1005–1010, 2005.

[163] Nicholas J. Shaheen, Gary W. Falk, Prasad G. Iyer, and Lauren B. Gerson. ACG Clinical Guideline: Diagnosis and Management of Barrett's Esophagus. *American Journal of Gastroenterology*, 111(1):30–50, 2016.

[164] Brian J. Reid, Xiaohong Li, Patricia C. Galipeau, and Thomas L. Vaughan. Barrett's oesophagus and oesophageal adenocarcinoma: Time for a new synthesis, 2010.

[165] Roberta De Angelis, Milena Sant, M. P. Coleman, Silvia Francisci, Paolo Baili, Daniela Pierannunzio, and Annalisa Trama. Cancer survival in Europe 1999-2007: Results of EUROCARE-5-a population-based study. *The Lancet Oncology*, 15(1):23–34, 2014.

[166] Manuela Quaresma, Michel P. Coleman, and Bernard Rachet. 40-year trends in an index of survival for all cancers combined and survival adjusted for age and sex for each cancer in England and Wales, 1971-2011: A population-based study. *The Lancet*, 385(9974):1206–1218, 2015.

[167] Rebecca L Siegel and et al Miller. Cancer Statistics, 2018. *Ca Cancer J Clin*, 68(1):7–30, 2018.

[168] Cancer Research UK. Oesophageal cancer statistics.

[169] E. L. Bird-Lieberman and R. C. Fitzgerald. Early diagnosis of oesophageal cancer, 2009.

[170] Thomas W. Rice, Deepa T. Patil, and Eugene H. Blackstone. 8th edition AJCC/UICC staging of cancers of the esophagus and esophagogastric junction: application to clinical practice. *Annals of Cardiothoracic Surgery*, 6(2):119–130, mar 2017.

[171] Victor S. Wang, Jason L. Hornick, Joe A. Sepulveda, Rie Mauer, and John M. Poneros. Low prevalence of submucosal invasive carcinoma at esophagectomy for high-grade dysplasia or intramucosal adenocarcinoma in Barrett's esophagus: a 20-year experience. *Gastrointestinal Endoscopy*, 69(4):777–783, 2009.

[172] Madhav Desai, Shreyas Saligram, Neil Gupta, Prashanth Vennalaganti, Ajay Bansal, Abhishek Choudhary, Sreekar Vennelaganti, Jianghua He, Mohammad Titi, Roberta Maselli, Bashar Qumseya, Mojtaba Olyaee, Irwing Waxman, Alessandro Repici, Cesare Hassan, and Prateek Sharma. Efficacy and safety outcomes of multimodal endoscopic eradication therapy in Barrett's esophagus-related neoplasia: a systematic review and pooled analysis, 2017.

[173] Ganapathy A. Prasad, Kenneth K. Wang, Navtej S. Buttar, Louis Michel Wongkeesong, Kausilia K. Krishnadath, Francis C. Nichols, Lori S. Lutzke, and Lynn S. Borkenhagen. Long-Term Survival Following Endoscopic and Surgical Treatment of High-Grade Dysplasia in Barrett's Esophagus. *Gastroenterology*, 132(4):1226–1233, 2007.

[174] K. Nadine Phoa, Frederike G. I. van Vilsteren, Bas L. A. M. Weusten, Raf Bisschops, Erik J. Schoon, Krish Ragunath, Grant Fullarton, Massimiliano Di Pietro, Narayanasamy Ravi, Mike Visser, G. Johan Offerhaus, Cees A. Seldenrijk, Sybren L. Meijer, Fiebo J. W. ten Kate, Jan G. P. Tijssen, and Jacques J. G. H. M. Bergman. Radiofrequency Ablation vs Endoscopic Surveillance for Patients With Barrett Esophagus and Low-Grade Dysplasia. *JAMA*, 311(12):1209, 2014.

[175] Ryan P. Merkow, Karl Y. Bilimoria, Rajesh N. Keswani, Jeanette Chung, Karen L. Sherman, Lawrence M. Knab, Mitchell C. Posner, and David J. Bentrem. Treatment trends, risk of lymph node metastasis, and outcomes for localized esophageal cancer. *Journal of the National Cancer Institute*, 106(7), 2014.

[176] Kerry B. Dunbar and Stuart Jon Spechler. The risk of lymph-node metastases in patients with high-grade dysplasia or intramucosal carcinoma in Barrett's esophagus: A systematic review, 2012.

[177] Rebecca C. Fitzgerald, Massimiliano Di Pietro, Krish Ragunath, Yeng Ang, Jin Yong Kang, Peter Watson, Nigel Trudgill, Praful Patel, Philip V. Kaye, Scott Sanders, Maria O'Donovan, Elizabeth Bird-Lieberman, Pradeep Bhandari, Janusz A. Jankowski, Stephen Attwood, Simon L. Parsons, Duncan Loft, Jesper Lagergren, Paul Moayyedi, Georgios Lyratzopoulos, and John De Caestecker. British Society of Gastroenterology guidelines on the diagnosis and management of Barrett's oesophagus. *Gut*, 63(1):7–42, 2014.

[178] Jacques J G H M Bergman, Yue Ming Zhang, Shun He, Bas Weusten, Liyan Xue, David E. Fleischer, Ning Lu, Sanford M. Dawsey, and Gui Qi Wang. Outcomes from a prospective trial of endoscopic radiofrequency ablation of early squamous cell neoplasia of the esophagus. *Gastrointestinal Endoscopy*, 74(6):1181–1190, 2011.

[179] Rehan J. Haidry, Mohammed A. Butt, Jason Dunn, Matthew Banks, Abhinav Gupta, Howard Smart, Pradeep Bhandari, Lesley Ann Smith, Robert Willert, Grant Fullarton, Morris John, Massimo Di Pietro, Ian Penman, Marco Novelli, and Laurence B. Lovat. Radiofrequency ablation for early oesophageal squamous neoplasia: Outcomes form United Kingdom registry. *World Journal of Gastroenterology*, 19(36):6011–6019, 2013.

[180] Shun He, Jacques Bergman, Yueming Zhang, Bas Weusten, Liyan Xue, Xiumin Qin, Lizhou Dou, Yong Liu, David Fleischer, Ning Lu, Sanford M. Dawsey, and Gui Qi Wang. Endoscopic radiofrequency ablation for early esophageal squamous cell neoplasia: Report of safety and effectiveness from a large prospective trial. *Endoscopy*, 47(5):398–408, 2015.

[181] Thomas W. Rice, Valerie W. Rusch, Hemant Ishwaran, and Eugene H. Blackstone. Cancer of the esophagus and esophagogastric junction. *Cancer*, 116(16):3763–3773, 2010.

[182] F. Lordick, C. Mariette, K. Haustermans, R. Obermannová, D. Arnold, and clinicalguidelines@esmo org on behalf of the ESMO Guidelines Committee. Oesophageal cancer: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. *Annals of Oncology*, 27:v50–v57, 2016.

[183] Tom Smith, Andreas Heger, and Ian Sudbery. UMI-tools: Modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Research*, 27(3):491–499, 2017.

[184] Ann Goding Sauer, Rebecca L. Siegel, Ahmedin Jemal, and Stacey A. Fedewa. Updated review of prevalence of major risk factors and use of screening tests for cancer in the United States, 2017.

[185] Cancer Research UK. Why isn't screening available for all cancers?

[186] Pierre Lao-Sirieix and Rebecca C. Fitzgerald. Screening for oesophageal cancer, 2012.

[187] Massimiliano di Pietro and Rebecca C. Fitzgerald. Revised British Society of Gastroenterology recommendation on the diagnosis and management of Barrett's oesophagus with low-grade dysplasia, 2018.

[188] Gareth S. Dulai, Jeffrey Gornbein, Katherine L. Kahn, Sushovan Guha, and Wilfred M. Weinstein. Preoperative prevalence of Barrett's esophagus in esophageal adenocarcinoma: A systematic review. *Gastroenterology*, 122(1):26–33, 2002.

[189] Frederik Hvid-Jensen, Lars Pedersen, Asbjrn M. Drewes, Henrik T. Srensen, and Peter Funch-Jensen. Incidence of Adenocarcinoma among Patients with Barrett's Esophagus. *Gastroenterological Endoscopy*, 53(11):3589, 2011.

[190] Marjolein Sikkema, Pieter J.F. de Jonge, Ewout W. Steyerberg, and Ernst J. Kuipers. Risk of Esophageal Adenocarcinoma and Mortality in Patients With Barrett's Esophagus: A Systematic Review and Meta-analysis, 2010.

[191] Tusar K. Desai, Kumar Krishnan, Niharika Samala, Jashanpreet Singh, John Cluley, Subaiah Perla, and Colin W. Howden. The incidence of oesophageal adenocarcinoma in non-dysplastic Barrett's oesophagus: A meta-analysis. *Gut*, 61(7):970–976, 2012.

[192] Siddharth Singh, Palaniappan Manickam, Anita V. Amin, Niharika Samala, Leo J. Schouten, Prasad G. Iyer, and Tusar K. Desai. Incidence of esophageal adenocarcinoma in Barrett's esophagus with low-grade dysplasia: A systematic review and meta-analysis. *Gastrointestinal Endoscopy*, 79(6), 2014.

[193] Amit Rastogi, Srinivas Puli, Hashem B. El-Serag, Ajay Bansal, Sachin Wani, and Prateek Sharma. Incidence of esophageal adenocarcinoma in patients with Barrett's esophagus and high-grade dysplasia: a meta-analysis. *Gastrointestinal Endoscopy*, 67(3):394–398, 2008.

[194] Nicholas J. Shaheen, Prateek Sharma, Bergein F. Overholt, Herbert C. Wolfsen, Richard E. Sampliner, Kenneth K. Wang, Joseph A. Galanko, Mary P. Bronner, John R. Goldblum, Ana E. Bennett, Blair A. Jobe, Glenn M. Eisen, M. Brian Fennerty, John G. Hunter, David E. Fleischer, Virender K. Sharma, Robert H. Hawes, Brenda J. Hoffman, Richard I. Rothstein, Stuart R. Gordon, Hiroshi Mashimo, Kenneth J. Chang, V. Raman Muthusamy, Steven A. Edmundowicz, Stuart J. Spechler, Ali A. Siddiqui, Rhonda F. Souza, Anthony Infantolino, Gary W. Falk, Michael B. Kimmey, Ryan D. Madanick, Amitabh Chak, and Charles J. Lightdale. Radiofrequency Ablation in Barrett's Esophagus with Dysplasia. *New England Journal of Medicine*, 360(22):2277–2288, may 2009.

[195] Bergein F. Overholt, Charles J. Lightdale, Kenneth K. Wang, Marcia I. Canto, Steven Burdick, Roger C. Haggitt, Mary P. Bronner, Shari L. Taylor, Michael G.A. Grace, and Michelle Depot. Photodynamic therapy with porfimer sodium for ablation of high-grade dysplasia in Barrett's esophagus: International, partially blinded, randomized phase III trial. *Gastrointestinal Endoscopy*, 62(4):488–498, 2005.

[196] B. J. Reid, R. C. Haggitt, C. E. Rubin, G. Roth, C. M. Surawicz, G. Van Belle, K. Lewin, W. M. Weinstein, D. A. Antonioli, H. Goldman, W. MacDonald, and D. Owen. Observer variation in the diagnosis of dysplasia in Barrett's esophagus. *Human Pathology*, 19(2):166–178, 1988.

[197] Elizabeth Montgomery, Mary P. Bronner, John R. Goldblum, Joel K. Greenson, Marian M. Haber, John Hart, Laura W. Lamps, Gregory Y. Lauwers, Audrey J. Lazenby, David N. Lewin, Marie E. Robert, Alicia Y. Toledano, Yu Shyr, and Kay Washington. Reproducibility of the diagnosis of dysplasia in Barrett esophagus: A reaffirmation. *Human Pathology*, 32(4):368–378, 2001.

[198] Rebecca Harrison, Ian Perry, William Haddadin, Stuart McDonald, Richard Bryan, Keith Abrams, Richard Sampliner, Nicholas J. Talley, Paul Moayyedi, and Janusz A. Jankowski. Detection of intestinal metaplasia in Barrett's Esophagus: An observational comparator study suggests the need for a minimum of eight biopsies. *American Journal of Gastroenterology*, 102(6):1154–1161, 2007.

[199] Jukka Ronkainen, Pertti Aro, Tom Storskrubb, Sven–Erik Johansson, Tore Lind, Elisabeth Bolling–Sternevald, Michael Vieth, Manfred Stolte, Nicholas J. Talley, and Lars Agréus. Prevalence of Barrett's Esophagus in the General Population: An Endoscopic Study. *Gastroenterology*, 129(6):1825–1831, dec 2005.

[200] R M Zagari, L Fuccio, M-A Wallander, S Johansson, R Fiocca, S Casanova, B Y Farahmand, C C Winchester, E Roda, and F Bazzoli. Gastro-oesophageal reflux symptoms, oesophagitis and Barrett's oesophagus in the general population: the Loiano-Monghidoro study. *Gut*, 57(10):1354–1359, apr 2008.

[201] Jesper Lagergren, Reinhold Bergström, Anders Lindgren, and Olof Nyrén. Symptomatic Gastroesophageal Reflux as a Risk Factor for Esophageal Adenocarcinoma. *New England Journal of Medicine*, 340(11):825–831, mar 1999.

[202] Diana C. Farrow, Thomas L. Vaughan, Carol Sweeney, Marilie D. Gammon, Wong Ho Chow, Harvey A. Risch, Janet L. Stanford, Philip D. Hansten, Susan T. Mayne, Janet B. Schoenberg, Heidi Rotterdam, Habibul Ahsan, A. Brian West, Robert Dubrow, Joseph F. Fraumeni, and William J. Blot. Gastroesophageal reflux disease, use of H2 receptor antagonists, and risk of esophageal and gastric cancer, 2000.

[203] M. Kareem Shariff, Xinxue Liu, Jane Blazeby, Elizabeth L. Bird-Lieberman, Maria O'Donovan, Zarah Abdullahi, and Rebecca Fitzgerald. Randomized crossover study comparing efficacy of transnasal endoscopy with that of standard endoscopy to detect Barrett's esophagus. *Gastrointestinal Endoscopy*, 75(5):954–961, 2012.

[204] Mark J. Roth, Shu Fan Liu, Sanford M. Dawsey, Bin Zhou, Christie Copeland, Guo Qing Wang, Diane Solomon, Stuart G. Baker, Carol A. Giffen, and Philip R. Taylor. Cytologic detection of esophageal squamous cell carcinoma and precursor lesions using balloon and sponge samplers in asymptomatic adults in Linxian, China. *Cancer*, 80(11):2047–2059, 1997.

[205] G. W. Falk, R. Chittajallu, J. R. Goldblum, C. V. Biscotti, K. R. Geisinger, R. E. Petras, S. Birgisson, T. W. Rice, and J. E. Richter. Surveillance of patients with Barrett's esophagus for dysplasia and cancer with balloon cytology. *Gastroenterology*, 112(6):1787–1797, 1997.

[206] S. R. Kadri, P. Lao-Sirieix, M. O'Donovan, I. Debiram, M. Das, J. M. Blazeby, J. Emery, A. Boussioutas, H. Morris, F. M. Walter, P. Pharoah, R. H. Hardwick, and R. C. Fitzgerald. Acceptability and accuracy of a non-endoscopic screening test for Barrett's oesophagus in primary care: cohort study. *BMJ*, 341(sep10 1):c4372–c4372, jan 2010.

[207] Caryn S. Ross-Innes, Irene Debiram-Beecham, Maria O'Donovan, Elaine Walker, Sibu Varghese, Pierre Lao-Sirieix, Laurence Lovat, Michael Griffin, Krish Ragunath, Rehan Haidry, Sarmed S. Sami, Philip Kaye, Marco Novelli, Babett Disep, Richard Ostler, Benoit Aigret, Bernard V. North, Pradeep Bhandari, Adam Haycock, Danielle Morris, Stephen Attwood, Anjan Dhar, Colin Rees, Matthew D. D. Rutter, Peter D. Sasieni, and Rebecca C. Fitzgerald. Evaluation of a Minimally Invasive Cell Sampling Device Coupled with Assessment of Trefoil Factor 3 Expression for Diagnosing Barrett's Esophagus: A Multi-Center Case–Control Study. *PLOS Medicine*, 12(1):e1001780, jan 2015.

[208] Florine Kastelein, Katharina Biermann, Ewout W Steyerberg, Joanne Verheij, Marit Kalisvaart, Leendert H J Looijenga, Hans A Stoop, Laurens Walter, Ernst J Kuipers, Manon C W Spaander, and Marco J Bruno. Aberrant p53 protein expression is associated with an increased risk of neoplastic progression in patients with Barrett's oesophagus. *Gut*, 62(12):1676–1683, dec 2013.

[209] Jamie M.J. Weaver, Caryn S. Ross-Innes, Nicholas Shannon, Andy G. Lynch, Tim Forshew, Mariagnese Barbera, Muhammed Murtaza, Chin Ann J. Ong, Pierre Lao-Sirieix, Mark J. Dunning, Laura Smith, Mike L. Smith, Charlotte L. Anderson, Benilton Carvalho, Maria O'donovan, Timothy J. Underwood, Andrew P. May, Nicola Grehan, Richard Hardwick, Jim Davies, Arusha Oloumi, Sam Aparicio, Carlos Caldas, Matthew D. Eldridge, Paul A.W. Edwards, Nitzan Rosenfeld, Simon Tavaré, Rebecca C. Fitzgerald, Stephen J. Hayes, Ang Yeng, Anne Marie Lydon, Soney Dharmaprasad, Sandra Greer, Shaun Preston, Sarah Oakes, Vicki Save, Simon Paterson-Brown, Olga Tucker, Derek Alderson, Philippe Taniere, Jamie Kelly, James Byrne, Donna Sharland, Nina Holling, Lisa Boulter, Fergus Noble, Bernard Stacey, Charles Crichton, Hugh Barr, Neil Shepherd, L. Max Almond, Oliver Old, Jesper Lagergren, James Gossage, Andrew Davies, Robert Mason, Fuju Chang, Janine Zylstra, Grant Sanders, Tim Wheatley, Richard Berrisford, Tim Bracey, Catherine Harden, David Bunting, Tom Roques, Jenny Nobes, Suat Loo, Mike Lewis, Ed Cheong, Oliver Priest, Simon L. Parsons, Irshad Soomro, Philip Kaye, John Saunders, Vincent Pang, Neil T. Welch, James A. Catton,

John P. Duffy, Krish Ragunath, Laurence Lovat, Rehan Haidry, Haroon Miah, Sarah Kerr, Victor Eneh, Rommel Butawan, Laszlo Igali, Hugo Ford, David Gilligan, Peter Safranek, Andy Hindmarsh, Vijayendran Sudjendran, Andy Metz, Nick Carroll, Michael Scott, Alison Cluroe, Ahmad Miremadi, Betania Mahler-Araujo, Olga Knight, Barbara Nutzinger, Chris Peters, Zarah Abdullahi, Irene Debriram-Beecham, Shalini Malhotra, Jason Crawte, Shona MacRae, Ayesha Noorani, Rachael Fels Elliott, Xiaodun Li, Lawrence Bower, Achilleas Achilleos, Jan Bornschein, Sebastian Zeki, Hamza Chettouh, Maria Secrier, Nadeera De Silva, Eleanor Gregson, Tsun Po Yang, and J. Robert O'Neil. Ordering of mutations in preinvasive disease stages of esophageal carcinogenesis. *Nature Genetics*, 46(8):837–843, 2014.

[210] Austin M. Dulak, Petar Stojanov, Shouyong Peng, Michael S. Lawrence, Cameron Fox, Chip Stewart, Santhoshi Bandla, Yu Imamura, Steven E. Schumacher, Erica Shefler, Aaron McKenna, Scott L. Carter, Kristian Cibulskis, Andrey Sivachenko, Gordon Saksena, Douglas Voet, Alex H. Ramos, Daniel Auclair, Kristin Thompson, Carrie Sougnez, Robert C. Onofrio, Candace Guiducci, Rameen Beroukhim, Zhongren Zhou, Lin Lin, Jules Lin, Rishindra Reddy, Andrew Chang, Rodney Landrenau, Arjun Pennathur, Shuji Ogino, James D. Luketich, Todd R. Golub, Stacey B. Gabriel, Eric S. Lander, David G. Beer, Tony E. Godfrey, Gad Getz, and Adam J. Bass. Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. *Nature Genetics*, 45(5):478–486, 2013.

[211] Caryn S. Ross-Innes, Jennifer Becq, Andrew Warren, R. Keira Cheetham, Helen Northen, Maria O'Donovan, Shalini Malhotra, Massimiliano Di Pietro, Sergii Ivakhno, Miao He, Jamie M.J. Weaver, Andy G. Lynch, Zoya Kingsbury, Mark Ross, Sean Humphray, David Bentley, Rebecca C. Fitzgerald, Stephen J. Hayes, Yeng Ang, Ian Welch, Shaun Preston, Sarah Oakes, Vicki Save, Richard Skipworth, Olga Tucker, Jim Davies, Charles Crichton, Christian Schusterreiter, Tim Underwood, Fergus Noble, Bernard Stacey, Jamie Kelly, James Byrne, Annette Haydon, Donna Sharland, Jack Owsley, Hugh Barr, Jesper Lagergren, James Gossage, Andrew Davies, Robert Mason, Fuju Chang, Janine Zylstra, Grant Sanders, Tim Wheatley, Richard Berrisford, Tim Bracey, Catherine Harden, David Bunting, Tom Roques, Jenny Nobes, Suat Loo, Mike Lewis, Ed Cheong, Oliver Priest, Simon L. Parsons, Irshad Soomro, Philip Kaye, John Saunders, Vincent Pang, Neil Welch, James A. Catton, John P. Duffy, Krish Ragunath, Laurence Lovat, Rehan Haidry, Haroon Miah, Sarah Kerr, Victor Eneh, Rommel Butawan, Michael Lewis, Edward Cheong, Bhasker Kumar, Laszlo Igali, Sharon Walton, Adela Dann, Peter Safranek, Andy Hindmarsh, Vijayendran Sudjendran, Michael Scott, Alison Cluroe, Ahmad Miremadi, Betania Mahler-Araujo, Barbara Nutzinger, Chris Peters, Zarah Abdullahi, Jason Crawte, Shona MacRae, Ayesha Noorani, Rachael Fels Elliott, Lawrence Bower, Paul Edwards, Simon Tavare, Matthew Eldridge, Jan Bornschein, Maria Secrier, Tsun Po Yang, J. Robert O'Neill, Kasia Adamczuk, Pierre Lao-Sirieix, Nicola Grehan, Laura Smith, Suzy Lishman, Duncan Beardsmore, and Sarah Dawson. Whole-genome sequencing provides new insights into the clonal architecture of Barrett's esophagus and esophageal adenocarcinoma. *Nature Genetics*, 47(9):1038–1046, 2015.

[212] Matthew D. Stachler, Amaro Taylor-Weiner, Shouyong Peng, Aaron McKenna, Agoston T. Agoston, Robert D. Odze, Jon M. Davison, Katie S. Nason, Massimo Loda, Ignaty Leshchiner, Chip Stewart, Petar Stojanov, Sara Seepo, Michael S. Lawrence, Daysha Ferrer-Torres, Jules Lin, Andrew C. Chang, Stacey B. Gabriel, Eric S. Lander, David G. Beer, Gad Getz, Scott L. Carter, and Adam J. Bass. Paired exome analysis of Barrett's esophagus and adenocarcinoma. *Nature Genetics*, 47(9):1047–1055, 2015.

[213] Xiaohong Li, Patricia C. Galipeau, Thomas G. Paulson, Carissa A. Sanchez, Jessica Arnaudo, Karen Liu, Cassandra L. Sather, Rumen L. Kostadinov, Robert D. Odze, Mary K. Kuhner, Carlo C. Maley, Steven G. Self, Thomas L. Vaughan, Patricia L. Blount, and Brian J. Reid. Temporal and spatial evolution of somatic chromosomal alterations: A case-cohort study of Barrett's esophagus. *Cancer Prevention Research*, 7(1):114–127, 2014.

[214] Brian J. Reid, Rodger C. Haggitt, Cyrus E. Rubin, and Peter S. Rabinovitch. Barrett's esophagus. Correlation between flow cytometry and histology in detection of patients at risk for adenocarcinoma. *Gastroenterology*, 93(1):1–11, 1987.

[215] Patricia C. Galipeau, Laura J. Prevo, Carissa A. Sanchez, Gary M. Longton, and Brian J. Reid. Clonal expansion and loss of heterozygosity at chromosomes 9p and 17p in premalignant esophageal (Barrett's) tissue. *Journal of the National Cancer Institute*, 91(24):2087–2095, 1999.

[216] Carlo C. Maley, Patricia C. Galipeau, Jennifer C. Finley, V. Jon Wongsurawat, Xiaohong Li, Carissa A. Sanchez, Thomas G. Paulson, Patricia L. Blount, Rosa Ana Risques, Peter S. Rabinovitch, and Brian J. Reid. Genetic clonal diversity predicts progression to esophageal adenocarcinoma. *Nature Genetics*, 38(4):468–473, 2006.

[217] Pierre Martinez, Margriet R. Timmer, Chiu T. Lau, Silvia Calpe, Maria Del Carmen Sancho-Serra, Danielle Straub, Ann Marie Baker, Sybren L. Meijer, Fiebo J.W.Ten Kate, Rosalie C. Mallant-Hent, Anton H.J. Naber, Arnoud H.A.M. Van Oijen, Lubbertus C. Baak, Pieter Scholten, Clarisse J.M. Böhmer, Paul Fockens, Jacques J.G.H.M. Bergman, Carlo C. Maley, Trevor A. Graham, and Kausilia K. Krishnadath. Dynamic clonal equilibrium and predetermined cancer risk in Barrett's oesophagus. *Nature Communications*, 7, 2016.

[218] Caryn S. Ross-Innes, Hamza Chettouh, Achilleas Achilleos, Nuria Galeano-Dalmau, Irene Debiram-Beecham, Shona MacRae, Petros Fessas, Elaine Walker, Sibu Varghese, Theodore Evan, Pierre S. Lao-Sirieix, Maria O'Donovan, Shalini Malhotra, Marco Novelli, Babett Disep, Phillip V. Kaye, Laurence B. Lovat, Rehan Haidry, Michael Griffin, Krish Ragunath, Pradeep Bhandari, Adam Haycock, Danielle Morris, Stephen Attwood, Anjan Dhar, Colin Rees, Matt D. Rutter, Richard Ostler, Benoit Aigret, Peter D. Sasieni, and Rebecca C. Fitzgerald. Risk stratification of Barrett's oesophagus using a non-endoscopic sampling method coupled with a biomarker panel: a cohort study. *The Lancet Gastroenterology and Hepatology*, 2(1):23–31, 2017.

[219] Austin M. Dulak, Steven E. Schumacher, Jasper Van Lieshout, Yu Imamura, Cameron Fox, Byoungyong Shim, Alex H. Ramos, Gordon Saksena, Sylvan C. Baca, Jose Baselga, Josep Tabernero, Jordi Barretina, Peter C. Enzinger, Giovanni Corso, Franco Roviello, Lin Lin, Santhoshi Bandla, James D. Luketich, Arjun Pennathur, Matthew Meyerson, Shuji Ogino, Ramesh A. Shivdasani, David G. Beer, Tony E. Godfrey, Rameen Beroukhim, and Adam J. Bass. Gastrointestinal adenocarcinomas of the esophagus, stomach, and colon exhibit distinct patterns of genome instability and oncogenesis. *Cancer Research*, 72(17):4383–4393, 2012.

[220] Yung-Jue Bang, Eric Van Cutsem, Andrea Feyereislova, Hyun C Chung, Lin Shen, Akira Sawaki, Florian Lordick, Atsushi Ohtsu, Yasushi Omuro, Taroh Satoh, Giuseppe Aprile, Evgeny Kulikov, Julie Hill, Michaela Lehle, Josef Rüschoff, and Yoon-Koo Kang. Trastuzumab in combination with chemotherapy versus chemotherapy alone for treatment of HER2-positive advanced gastric or gastro-oesophageal junction cancer

(ToGA): a phase 3, open-label, randomised controlled trial. *The Lancet*, 376(9742):687–697, aug 2010.

[221] Maria Secrier, Xiaodun Li, Nadeera De Silva, Matthew D. Eldridge, Gianmarco Contino, Jan Bornschein, Shona Macrae, Nicola Grehan, Maria O'Donovan, Ahmad Miremadi, Tsun Po Yang, Lawrence Bower, Hamza Chettouh, Jason Crawte, Núria Galeano-Dalmau, Anna Grabowska, John Saunders, Tim Underwood, Nicola Waddell, Andrew P. Barbour, Barbara Nutzinger, Achilleas Achilleos, Paul A.W. Edwards, Andy G. Lynch, Simon Tavaré, Rebecca C. Fitzgerald, Ayesha Noorani, Rachael Fels Elliott, Jamie Weaver, Caryn Ross-Innes, Laura Smith, Zarah Abdullahi, Rachel De La Rue, Alison Cluroe, Shalini Malhotra, Richard Hardwick, Hugo Ford, Mike L. Smith, Jim Davies, Richard Turkington, Stephen J. Hayes, Yeng Ang, Shaun R. Preston, Sarah Oakes, Izhar Bagwan, Vicki Save, Richard J.E. Skipworth, Ted R. Hupp, J. Robert O'Neill, Olga Tucker, Philippe Taniere, Fergus Noble, Jack Owsley, Laurence Lovat, Rehan Haidry, Victor Eneh, Charles Crichton, Hugh Barr, Neil Shepherd, Oliver Old, Jesper Lagergren, James Gossage, Andrew Davies, Fuju Chang, Janine Zylstra, Grant Sanders, Richard Berrisford, Catherine Harden, David Bunting, Mike Lewis, Ed Cheong, Bhaskar Kumar, Simon L. Parsons, Irshad Soomro, Philip Kaye, Pamela Collier, Laszlo Igali, Ian Welch, Michael Scott, Shamila Sothi, Sari Suortamo, Suzy Lishman, Duncan Beardsmore, Hayley E. Francies, Mathew J. Garnett, John V. Pearson, Katia Nones, Ann Marie Patch, and Sean M. Grimmond. Mutational signatures in esophageal adenocarcinoma define etiologically distinct subgroups with therapeutic relevance. *Nature Genetics*, 48(10):1131–1141, 2016.

[222] G. N. Hortobagyi, S. M. Stemmer, H. A. Burris, Y. Yap, G. S. Sonke, S. Paluch-Shimon, M. Campone, K. L. Blackwell, F. André, E. P. Winer, W. Janni, S. Verma, P. Conte, C. L. Arteaga, D. A. Cameron, K. Petrakova, L. L. Hart, C. Villanueva, A. Chan, E. Jakobsen, A. Nusch, O. Burdaeva, E.-M. Grischke, E. Alba, E. Wist, N. Marschner, A. M. Favret, D. Yardley, T. Bachelot, L.-M. Tseng, S. Blau, F. Xuan, F. Souami, M. Miller, C. Germa, S. Hirawat, and J. O'Shaughnessy. Ribociclib as First-Line Therapy for HR-Positive, Advanced Breast Cancer. *New England Journal of Medicine*, 375(18):1738–1748, 2016.

[223] Yung-Jue Bang, Rui-Hua Xu, Keisho Chin, Keun-Wook Lee, Se Hoon Park, Sun Young Rha, Lin Shen, Shukui Qin, Nong Xu, Seock-Ah Im, Gershon Locker, Phil Rowe, Xiaojin Shi, Darren Hodgson, Yu-Zhen Liu, and Narikazu Boku. Olaparib in combination with paclitaxel in patients with advanced gastric cancer who have progressed following first-line therapy (GOLD): a double-blind, randomised, placebo-controlled, phase 3 trial. *The Lancet Oncology*, 18(12):1637–1651, dec 2017.

[224] P. Van Loo, B. Naume, W. Sun, V. J. Weigman, P. Marynen, A.-L. Borresen-Dale, V. N. Kristensen, C. M. Perou, A. Zetterberg, S. H. Nordgard, H. G. Russnes, I. H. Rye, and O. C. Lingjaerde. Allele-specific copy number analysis of tumors. *Proceedings of the National Academy of Sciences*, 107(39):16910–16915, 2010.

[225] Ilari Scheinin, Daoud Sie, Henrik Bengtsson, Mark A. Van De Wiel, Adam B. Olshen, Hinke F. Van Thuijl, Hendrik F. Van Essen, Paul P. Eijk, Franc¸ois Rustenburg, Gerrit A. Meijer, Jaap C. Reijneveld, Pieter Wesseling, Daniel Pinkel, Donna G. Albertson, and Bauke Ylstra. DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly. *Genome Research*, 24(12):2022–2032, 2014.

[226] Navin Rustagi, Anbo Zhou, W. Scott Watkins, Erika Gedvilaite, Shuoguo Wang, Naveen Ramesh, Donna Muzny, Richard A. Gibbs, Lynn B. Jorde, Fuli Yu, and Jinchuan Xing.

Extremely low-coverage whole genome sequencing in South Asians captures population genomics information. *BMC Genomics*, 18(1), 2017.

[227] Bo Zhou, Steve S. Ho, Xianglong Zhang, Reenal Pattni, Rajini R. Haraksingh, and Alexander E. Urban. Whole-genome sequencing analysis of CNV using low-coverage and paired-end strategies is efficient and outperforms array-based CNV analysis. *Journal of Medical Genetics*, 55(11):735–743, 2018.

[228] Daniel H. Hovelson, Chia-Jen Liu, Yugang Wang, Qing Kang, James Henderson, Amy Gursky, Scott Brockman, Nithya Ramnath, John C. Krauss, Moshe Talpaz, Malathi Kandarpa, Rashmi Chugh, Missy Tuck, Kirk Herman, Catherine S. Grasso, Michael J. Quist, Felix Y. Feng, Christine Haakenson, John Langmore, Emmanuel Kamberov, Tim Tesmer, Hatim Husain, Robert J. Lonigro, Dan Robinson, David C. Smith, Ajjai S. Alva, Maha H. Hussain, Arul M. Chinnaiyan, Muneesh Tewari, Ryan E. Mills, Todd M. Morgan, and Scott A. Tomlins. Rapid, ultra low coverage copy number profiling of cell-free DNA as a precision oncology screening strategy. *Oncotarget*, 8(52):89848–89866, 2017.

[229] Katia Nones, Nicola Waddell, Nicci Wayte, Ann Marie Patch, Peter Bailey, Felicity Newell, Oliver Holmes, J. Lynn Fink, Michael C.J. Quinn, Yue Hang Tang, Guy Lampe, Kelly Quek, Kelly A. Loffler, Suzanne Manning, Senel Idrisoglu, David Miller, Qinying Xu, Nick Waddell, Peter J. Wilson, Timothy J.C. Bruxner, Angelika N. Christ, Ivon Harliwong, Craig Nourse, Ehsan Nourbakhsh, Matthew Anderson, Stephen Kazakoff, Conrad Leonard, Scott Wood, Peter T. Simpson, Lynne E. Reid, Lutz Krause, Damian J. Hussey, David I. Watson, Reginald V. Lord, Derek Nancarrow, Wayne A. Phillips, David Gotley, B. Mark Smithers, David C. Whiteman, Nicholas K. Hayward, Peter J. Campbell, John V. Pearson, Sean M. Grimmond, and Andrew P. Barbour. Genomic catastrophes frequently arise in esophageal adenocarcinoma and drive tumorigenesis. *Nature Communications*, 5, 2014.

[230] Zoubin Ghahramani. Bayesian non-parametrics and the probabilistic approach to modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1984), 2013.

[231] Alan P. Fields and Nicole R. Murray. Protein kinase C isozymes as therapeutic targets for treatment of human cancers. *Advances in Enzyme Regulation*, 48(1):166–178, 2008.

[232] Silvia Mateo-Lozano, Sarah Bazzocco, Paulo Rodrigues, Rocco Mazzolini, Elena Andretta, Higinio Dopeso, Yolanda Fernández, Edgar Del Llano, Josipa Bilic, Luciá Suárez-López, Irati MacAya, Fernando Cartón-Garciá, Rocio Nieto, Lizbeth M. Jimenez-Flores, Priscila Guimarães De Marcondes, Yaiza Nuñez, Elsa Afonso, Karina Cacci, Javier Hernández-Losa, Stefania Landolfi, Ibane Abasolo, Santiago Ramón, John M. Mariadason, Simo Schwartz, Toshimitsu Matsui, and Diego Arango. Loss of the EPH receptor B6 contributes to colorectal cancer metastasis. *Scientific Reports*, 7, 2017.

[233] Ran Zhao, Bu Young Choi, Mee Hyun Lee, Ann M. Bode, and Zigang Dong. Implications of Genetic and Epigenetic Alterations of CDKN2A (p16INK4a) in Cancer, 2016.

[234] Anika Maria Weber and Anderson Joseph Ryan. ATM and ATR as therapeutic targets in cancer, 2015.

[235] Larry M. Karnitz and Lee Zou. Molecular pathways: Targeting ATR in cancer therapy. *Clinical Cancer Research*, 21(21):4780–4785, 2015.

[236] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, aug 2009.

[237] R Foundation for Statistical Computing. *R: A Language and Environment for Statistical Computing*, volume 1. 2011.

[238] Heng Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*, 00(00):1–3, 2013.

[239] Rahul Sinha, Geoff Stanley, GS Gulati, Camille Ezran, KJ Travaglini, Eric Wei, CKF Chan, AN Nabhan, Tianying Su, RM Morganti, SD Conley, Hassan Chaib, Kristy Red-Horse, MT Longaker, MP Snyder, MA Krasnow, and IL Weissman. Index Switching Causes "Spreading-Of-Signal" Among Multiplexed Samples In Illumina HiSeq 4000 DNA Sequencing. *bioRxiv*, page 125724, 2017.

[240] Illumina. Effects of Index Misassignment on Multiplexing and Downstream Analysis, 2017.

[241] Laura E. MacConaill, Robert T. Burns, Anwesha Nag, Haley A. Coleman, Michael K. Slevin, Kristina Giorda, Madelyn Light, Kevin Lai, Mirna Jarosz, Matthew S. McNeill, Matthew D. Ducar, Matthew Meyerson, and Aaron R. Thorner. Unique, dual-indexed sequencing adapters with UMIs effectively eliminate index cross-talk and significantly improve sensitivity of massively parallel sequencing. *BMC Genomics*, 19(1), 2018.

[242] Illumina. Spiking custom primers into the Illumina sequencing primers. https://support.illumina.com/bulletins/2016/04/spiking-custom-primers-into-the-illumina-sequencing-primers-.html, 2019.

[243] Illumina. Considerations when migrating non-Illumina libraries between sequencing platforms. https://support.illumina.com/bulletins/2016/10/considerations-when-migrating-nonillumina-libraries-between-sequencing-platforms.html, 2019.

[244] Johannes Köster and Sven Rahmann. Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, 2012.

[245] Andy B. Yoo, Morris A. Jette, and Mark Grondona. SLURM: Simple Linux Utility for Resource Management. pages 44–60. 2003.

[246] Brian Bushnell. BBDuk, 2019.

[247] Brian Bushnell, Jonathan Rood, and Esther Singer. BBMerge – Accurate paired shotgun read merging via overlap. *PLoS ONE*, 12(10), 2017.

[248] Martin Šošić and Mile Šikić. Edlib: a C/C ++ library for fast, exact sequence alignment using edit distance. *Bioinformatics*, 33(9):1394–1395, may 2017.

[249] Andreas Heger. Pysam.

[250] Aaron R. Quinlan and Ira M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, mar 2010.

[251] Michael M Mckerns, Leif Strand, Tim Sullivan, Alta Fang, and Michael A G Aivazis. Building a Framework for Predictive Science. *arXiv*, pages 1–12, 2012.

[252] Simon Andrews. FastQC - A quality control tool for high throughput sequence data, 2010.

[253] Martin Kircher, Susanna Sawyer, and Matthias Meyer. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Research*, 40(1), 2012.

[254] Heng Li, Jue Ruan, and Richard Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, 18(11):1851–1858, 2008.

[255] Mark Ziemann. Accuracy, speed and error tolerance of short DNA sequence aligners. *bioRxiv*, (October):053686, 2016.

[256] Ashfaq A. Mir, Claude Philippe, and Gäel Cristofari. euL1db: The European database of L1HS retrotransposon insertions in humans. *Nucleic Acids Research*, 43(D1):D43–D47, 2015.

[257] David R. Bentley, Shankar Balasubramanian, Harold P. Swerdlow, Geoffrey P. Smith, John Milton, Clive G. Brown, Kevin P. Hall, Dirk J. Evers, Colin L. Barnes, Helen R. Bignell, Jonathan M. Boutell, Jason Bryant, Richard J. Carter, R. Keira Cheetham, Anthony J. Cox, Darren J. Ellis, Michael R. Flatbush, Niall A. Gormley, Sean J. Humphray, Leslie J. Irving, Mirian S. Karbelashvili, Scott M. Kirk, Heng Li, Xiaohai Liu, Klaus S. Maisinger, Lisa J. Murray, Bojan Obradovic, Tobias Ost, Michael L. Parkinson, Mark R. Pratt, Isabelle M.J. Rasolonjatovo, Mark T. Reed, Roberto Rigatti, Chiara Rodighiero, Mark T. Ross, Andrea Sabot, Subramanian V. Sankar, Aylwyn Scally, Gary P. Schroth, Mark E. Smith, Vincent P. Smith, Anastassia Spiridou, Peta E. Torrance, Svilen S. Tzonev, Eric H. Vermaas, Klaudia Walter, Xiaolin Wu, Lu Zhang, Mohammed D. Alam, Carole Anastasi, Ify C. Aniebo, David M.D. Bailey, Iain R. Bancarz, Saibal Banerjee, Selena G. Barbour, Primo A. Baybayan, Vincent A. Benoit, Kevin F. Benson, Claire Bevis, Phillip J. Black, Asha Boodhun, Joe S. Brennan, John A. Bridgham, Rob C. Brown, Andrew A. Brown, Dale H. Buermann, Abass A. Bundu, James C. Burrows, Nigel P. Carter, Nestor Castillo, Maria Chiara E. Catenazzi, Simon Chang, R. Neil Cooley, Natasha R. Crake, Olubunmi O. Dada, Konstantinos D. Diakoumakos, Belen Dominguez-Fernandez, David J. Earnshaw, Ugonna C. Egbujor, David W. Elmore, Sergey S. Etchin, Mark R. Ewan, Milan Fedurco, Louise J. Fraser, Karin V. Fuentes Fajardo, W. Scott Furey, David George, Kimberley J. Gietzen, Colin P. Goddard, George S. Golda, Philip A. Granieri, David E. Green, David L. Gustafson, Nancy F. Hansen, Kevin Harnish, Christian D. Haudenschild, Narinder I. Heyer, Matthew M. Hims, Johnny T. Ho, Adrian M. Horgan, Katya Hoschler, Steve Hurwitz, Denis V. Ivanov, Maria Q. Johnson, Terena James, T. A. Huw Jones, Gyoung Dong Kang, Tzvetana H. Kerelska, Alan D. Kersey, Irina Khrebtukova, Alex P. Kindwall, Zoya Kingsbury, Paula I. Kokko-Gonzales, Anil Kumar, Marc A. Laurent, Cynthia T. Lawley, Sarah E. Lee, Xavier Lee, Arnold K. Liao, Jennifer A. Loch, Mitch Lok, Shujun Luo, Radhika M. Mammen, John W. Martin, Patrick G. McCauley, Paul McNitt, Parul Mehta, Keith W. Moon, Joe W. Mullens, Taksina Newington, Zemin Ning, Bee Ling Ng, Sonia M. Novo, Michael J. O'Neill, Mark A. Osborne, Andrew Osnowski, Omead Ostadan, Lambros L. Paraschos, Lea Pickering, Andrew C. Pike, Alger C. Pike, D. Chris Pinkard, Daniel P. Pliskin, Joe Podhasky, Victor J. Quijano, Come Raczy, Vicki H. Rae, Stephen R. Rawlings, Ana Chiva Rodriguez, Phyllida M. Roe, John Rogers, Maria C. Rogert Bacigalupo, Nikolai Romanov, Anthony Romieu, Rithy K. Roth, Natalie J. Rourke, Silke T. Ruediger, Eli Rusman, Raquel M. Sanches-Kuiper, Martin R. Schenker, Josefina M. Seoane, Richard J. Shaw, Mitch K. Shiver, Steven W.

Short, Ning L. Sizto, Johannes P. Sluis, Melanie A. Smith, Jean Ernest Sohna Sohna, Eric J. Spence, Kim Stevens, Neil Sutton, Lukasz Szajkowski, Carolyn L. Tregidgo, Gerardo Turcatti, Stephanie Vandevondele, Yuli Verhovsky, Selene M. Virk, Suzanne Wakelin, Gregory C. Walcott, Jingwen Wang, Graham J. Worsley, Juying Yan, Ling Yau, Mike Zuerlein, Jane Rogers, James C. Mullikin, Matthew E. Hurles, Nick J. McCooke, John S. West, Frank L. Oaks, Peter L. Lundberg, David Klenerman, Richard Durbin, and Anthony J. Smith. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59, 2008.

[258] Juliane C. Dohm, Claudio Lottaz, Tatiana Borodina, and Heinz Himmelbauer. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*, 36(16), 2008.

[259] Yuval Benjamini and Terence P. Speed. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research*, 40(10), 2012.

[260] John C. Marioni, Natalie P. Thorne, Armand Valsesia, Tomas Fitzgerald, Richard Redon, Heike Fiegler, T. Daniel Andrews, Barbara E. Stranger, Andrew G. Lynch, Emmanouil T. Dermitzakis, Nigel P. Carter, Simon Tavaré, and Matthew E. Hurles. Breaking the waves: Improved detection of copy number variation from microarray-based comparative genomic hybridization. *Genome Biology*, 8(10), 2007.

[261] Sebastiaan van Heesch, Michal Mokry, Veronika Boskova, Wade Junker, Rajdeep Mehon, Pim Toonen, Ewart de Bruijn, James D. Shull, Timothy J. Aitman, Edwin Cuppen, and Victor Guryev. Systematic biases in DNA copy number originate from isolation procedures. *Genome Biology*, 14(4), 2013.

[262] Tyrone Ryba, Ichiro Hiratani, Junjie Lu, Mari Itoh, Michael Kulik, Jinfeng Zhang, Thomas C. Schulz, Allan J. Robins, Stephen Dalton, and David M. Gilbert. Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Research*, 20(6):761–770, 2010.

[263] Simon N. Wood. Thin plate regression splines, 2003.

[264] Bob Carpenter, Jiqiang Guo, Matthew D. Hoffman, Marcus Brubaker, Andrew Gelman, Daniel Lee, Ben Goodrich, Peter Li, Allen Riddell, and Michael Betancourt. Stan : A Probabilistic Programming Language . *Journal of Statistical Software*, 76(1), 2017.

[265] Serena Nik-Zainal, Peter Van Loo, David C. Wedge, Ludmil B. Alexandrov, Christopher D. Greenman, King Wai Lau, Keiran Raine, David Jones, John Marshall, Manasa Ramakrishna, Adam Shlien, Susanna L. Cooke, Jonathan Hinton, Andrew Menzies, Lucy A. Stebbings, Catherine Leroy, Mingming Jia, Richard Rance, Laura J. Mudie, Stephen J. Gamble, Philip J. Stephens, Stuart McLaren, Patrick S. Tarpey, Elli Papaemmanuil, Helen R. Davies, Ignacio Varela, David J. McBride, Graham R. Bignell, Kenric Leung, Adam P. Butler, Jon W. Teague, Sancha Martin, Goran Jönsson, Odette Mariani, Sandrine Boyault, Penelope Miron, Aquila Fatima, Anita Langerod, Samuel A.J.R. Aparicio, Andrew Tutt, Anieta M. Sieuwerts, Åke Borg, Gilles Thomas, Anne Vincent Salomon, Andrea L. Richardson, Anne Lise Borresen-Dale, P. Andrew Futreal, Michael R. Stratton, and Peter J. Campbell. The life history of 21 breast cancers. *Cell*, 149(5):994–1007, 2012.

[266] Cancer Genome Atlas Research Network, Adam Abeshouse, Jaeil Ahn, Rehan Akbani, Adrian Ally, Samirkumar Amin, Christopher D. Andry, Matti Annala, Armen Aprikian, Joshua Armenia, Arshi Arora, J. Todd Auman, Miruna Balasundaram, Saianand

Balu, Christopher E. Barbieri, Thomas Bauer, Christopher C. Benz, Alain Bergeron, Rameen Beroukhim, Mario Berrios, Adrian Bivol, Tom Bodenheimer, Lori Boice, Moiz S. Bootwalla, Rodolfo Borges Dos Reis, Paul C. Boutros, Jay Bowen, Reanne Bowlby, Jeffrey Boyd, Robert K. Bradley, Anne Breggia, Fadi Brimo, Christopher A. Bristow, Denise Brooks, Bradley M. Broom, Alan H. Bryce, Glenn Bubley, Eric Burks, Yaron S.N. Butterfield, Michael Button, David Canes, Carlos G. Carlotti, Rebecca Carlsen, Michel Carmel, Peter R. Carroll, Scott L. Carter, Richard Cartun, Brett S. Carver, June M. Chan, Matthew T. Chang, Yu Chen, Andrew D. Cherniack, Simone Chevalier, Lynda Chin, Juok Cho, Andy Chu, Eric Chuah, Sudha Chudamani, Kristian Cibulskis, Giovanni Ciriello, Amanda Clarke, Matthew R. Cooperberg, Niall M. Corcoran, Anthony J. Costello, Janet Cowan, Daniel Crain, Erin Curley, Kerstin David, John A. Demchok, Francesca Demichelis, Noreen Dhalla, Rajiv Dhir, Alexandre Doueik, Bettina Drake, Heidi Dvinge, Natalya Dyakova, Ina Felau, Martin L. Ferguson, Scott Frazer, Stephen Freedland, Yao Fu, Stacey B. Gabriel, Jianjiong Gao, Johanna Gardner, Julie M. Gastier-Foster, Nils Gehlenborg, Mark Gerken, Mark B. Gerstein, Gad Getz, Andrew K. Godwin, Anuradha Gopalan, Markus Graefen, Kiley Graim, Thomas Gribbin, Ranabir Guin, Manaswi Gupta, Angela Hadjipanayis, Syed Haider, Lucie Hamel, D. Neil Hayes, David I. Heiman, Julian Hess, Katherine A. Hoadley, Andrea H. Holbrook, Robert A. Holt, Antonia Holway, Christopher M. Hovens, Alan P. Hoyle, Mei Huang, Carolyn M. Hutter, Michael Ittmann, Lisa Iype, Stuart R. Jefferys, Corbin D. Jones, Steven J.M. Jones, Hartmut Juhl, Andre Kahles, Christopher J. Kane, Katayoon Kasaian, Michael Kerger, Ekta Khurana, Jaegil Kim, Robert J. Klein, Raju Kucherlapati, Louis Lacombe, Marc Ladanyi, Phillip H. Lai, Peter W. Laird, Eric S. Lander, Mathieu Latour, Michael S. Lawrence, Kevin Lau, Tucker Lebien, Darlene Lee, Semin Lee, Kjong Van Lehmann, Kristen M. Leraas, Ignaty Leshchiner, Robert Leung, John A. Libertino, Tara M. Lichtenberg, Pei Lin, W. Marston Linehan, Shiyun Ling, Scott M. Lippman, Jia Liu, Wenbin Liu, Lucas Lochovsky, Massimo Loda, Christopher Logothetis, Laxmi Lolla, Teri Longacre, Yiling Lu, Jianhua Luo, Yussanne Ma, Harshad S. Mahadeshwar, David Mallery, Armaz Mariamidze, Marco A. Marra, Michael Mayo, Shannon McCall, Ginette McKercher, Shaowu Meng, Anne Marie Mes-Masson, Maria J. Merino, Matthew Meyerson, Piotr A. Mieczkowski, Gordon B. Mills, Kenna R.Mills Shaw, Sarah Minner, Alireza Moinzadeh, Richard A. Moore, Scott Morris, Carl Morrison, Lisle E. Mose, Andrew J. Mungall, Bradley A. Murray, Jerome B. Myers, Rashi Naresh, Joel Nelson, Mark A. Nelson, Peter S. Nelson, Yulia Newton, Michael S. Noble, Houtan Noushmehr, Matti Nykter, Angeliki Pantazi, Michael Parfenov, Peter J. Park, Joel S. Parker, Joseph Paulauskis, Robert Penny, Charles M. Perou, Alain Piché, Todd Pihl, Peter A. Pinto, Davide Prandi, Alexei Protopopov, Nilsa C. Ramirez, Arvind Rao, W. Kimryn Rathmell, Gunnar Rätsch, Xiaojia Ren, Victor E. Reuter, Sheila M. Reynolds, Suhn K. Rhie, Kimberly Rieger-Christ, Jeffrey Roach, A. Gordon Robertson, Brian Robinson, Mark A. Rubin, Fred Saad, Sara Sadeghi, Gordon Saksena, Charles Saller, Andrew Salner, Francisco Sanchez-Vega, Chris Sander, George Sandusky, Guido Sauter, Andrea Sboner, Peter T. Scardino, Eleonora Scarlata, Jacqueline E. Schein, Thorsten Schlomm, Laura S. Schmidt, Nikolaus Schultz, Steven E. Schumacher, Jonathan Seidman, Luciano Neder, Sahil Seth, Alexis Sharp, Candace Shelton, Troy Shelton, Hui Shen, Ronglai Shen, Mark Sherman, Margi Sheth, Yan Shi, Juliann Shih, Ilya Shmulevich, Jeffry Simko, Ronald Simon, Janae V. Simons, Payal Sipahimalani, Tara Skelly, Heidi J. Sofia, Matthew G. Soloway, Xingzhi Song, Andrea Sorcini, Carrie Sougnez, Serghei Stepa, Chip Stewart, John Stewart, Joshua M. Stuart, Travis B. Sullivan, Charlie Sun, Huandong Sun, Angela Tam, Donghui Tan, Jiabin Tang, Roy Tarnuzzer, Katherine Tarvin, Barry S. Taylor, Patrick Teebagy, Imelda Tenggara, Bernard Têtu, Ashutosh Tewari, Nina Thiessen, Timothy Thompson, Leigh B. Thorne, Daniela P. Tirapelli, Scott A. Tomlins, Felipe Amstalden Trevisan, Patricia Troncoso, Lawrence D. True, Maria Christina Tsourlakis, Svitlana Tyekucheva, Eliezer Van Allen, David J. Van Den Berg, Umadevi Veluvolu, Roel Verhaak, Cathy D. Vocke, Doug Voet, Yunhu Wan, Qingguo Wang, Wenyi Wang, Zhining Wang, Nils Weinhold, John N. Weinstein, Daniel J. Weisenberger, Matthew D. Wilkerson, Lisa Wise, John Witte, Chia Chin Wu, Junyuan Wu, Ye Wu, Andrew W. Xu, Shalini S. Yadav, Lixing Liming Yang, Lixing Liming Yang, Christina Yau, Huihui Ye, Peggy Yena, Thomas Zeng, Jean C. Zenklusen, Hailei Zhang, Jianhua Jiashan Zhang, Jianhua Jiashan Zhang, Wei Zhang, Yi Zhong, Kelsey Zhu, and Erik Zmuda. The Molecular Taxonomy of Primary Prostate Cancer. *Cell*, 163(4):1011–25, 2015.

[267] T. R. Shenoy, G. Boysen, M. Y. Wang, Q. Z. Xu, W. Guo, F. M. Koh, C. Wang, L. Z. Zhang, Y. Wang, V. Gil, S. Aziz, R. Christova, D. N. Rodrigues, M. Crespo, P. Rescigno, N. Tunariu, R. Riisnaes, Z. Zafeiriou, P. Flohr, W. Yuan, E. Knight, A. Swain, M. Ramalho-Santos, D. Y. Xu, J. de Bono, and H. Wu. CHD1 loss sensitizes prostate cancer to DNA damaging therapy by promoting error-prone double-strand break repair. *Annals of Oncology*, 28(7):1495–1507, jul 2017.

[268] Vijayalakshmi Kari, Wael Yassin Mansour, Sanjay Kumar Raul, Simon J Baumgart, Andreas Mund, Marian Grade, Hüseyin Sirma, Ronald Simon, Hans Will, Matthias Dobbelstein, Ekkehard Dikomey, and Steven A Johnsen. Loss of CHD 1 causes DNA repair defects and enhances prostate cancer therapeutic responsiveness. 17(11):1609–1623, 2016.

[269] 1000 Genomes Project Consortium, Adam Auton, Lisa D Brooks, Richard M Durbin, Erik P Garrison, Hyun Min Kang, Jan O Korbel, Jonathan L Marchini, Shane McCarthy, Gil A McVean, and Gonçalo R Abecasis. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

[270] Malachi Griffith, Christopher A. Miller, Obi L. Griffith, Kilannin Krysiak, Zachary L. Skidmore, Avinash Ramu, Jason R. Walker, Ha X. Dang, Lee Trani, David E. Larson, Ryan T. Demeter, Michael C. Wendl, Joshua F. McMichael, Rachel E. Austin, Vincent Magrini, Sean D. McGrath, Amy Ly, Shashikant Kulkarni, Matthew G. Cordes, Catrina C. Fronick, Robert S. Fulton, Christopher A. Maher, Li Ding, Jeffery M. Klco, Elaine R. Mardis, Timothy J. Ley, and Richard K. Wilson. Optimizing Cancer Genome Sequencing and Analysis. *Cell Systems*, 1(3):210–223, 2015.

[271] Andrew Roth, Jaswinder Khattra, Damian Yap, Adrian Wan, Emma Laks, Justina Biele, Gavin Ha, Samuel Aparicio, Alexandre Bouchard-Côté, and Sohrab P. Shah. PyClone: Statistical inference of clonal population structure in cancer. *Nature Methods*, 11(4):396–398, 2014.

[272] Andrej Fischer, Ignacio Vázquez-García, Christopher J.R. Illingworth, and Ville Mustonen. High-definition reconstruction of clonal composition in cancer. *Cell Reports*, 7(5):1740–1752, 2014.

[273] Jie Liu, John T Halloran, Jeffrey A Bilmes, Riza M Daza, Choli Lee, M Elisabeth, Donna Prunkard, Chaozhong Song, Sibel Blau, Michael O Dorschner, Vijayakrishna K Gadi, Jay Shendure, C Anthony Blau, and William S Noble. Comprehensive statistical inference of the clonal structure of cancer from multiple biopsies. *Scientific Reports*, pages 1–13, 2017.

[274] Habil Zare, Junfeng Wang, Alex Hu, Kris Weber, Josh Smith, Debbie Nickerson, Chao Zhong Song, Daniela Witten, C. Anthony Blau, and William Stafford Noble.

Inferring Clonal Composition from Multiple Sections of a Breast Cancer. *PLoS Computational Biology*, 10(7), 2014.

[275] Christopher A. Miller, Brian S. White, Nathan D. Dees, Malachi Griffith, John S. Welch, Obi L. Griffith, Ravi Vij, Michael H. Tomasson, Timothy A. Graubert, Matthew J. Walter, Matthew J. Ellis, William Schierding, John F. DiPersio, Timothy J. Ley, Elaine R. Mardis, Richard K. Wilson, and Li Ding. SciClone: Inferring Clonal Architecture and Tracking the Spatial and Temporal Patterns of Tumor Evolution. *PLoS Computational Biology*, 10(8), 2014.

[276] Layla Oesper, Ahmad Mahmoody, and Benjamin J Raphael. THetA: inferring intratumor heterogeneity from high-throughput DNA sequencing data. *Genome Biology*, 14(7):R80, 2013.

[277] Gavin Ha, Andrew Roth, Jaswinder Khattra, Julie Ho, Damian Yap, Leah M. Prentice, Nataliya Melnyk, Andrew McPherson, Ali Bashashati, Emma Laks, Justina Biele, Jiarui Ding, Alan Le, Jamie Rosner, Karey Shumansky, Marco A. Marra, C. Blake Gilks, David G. Huntsman, Jessica N. McAlpine, Samuel Aparicio, and Sohrab P. Shah. TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Research*, 24(11):1881–1893, nov 2014.

[278] Layla Oesper, Gryte Satas, and Benjamin J. Raphael. Quantifying tumor heterogeneity in whole-genome and whole-exome sequencing data. *Bioinformatics*, 30(24):3532–3540, dec 2014.

[279] Wei Jiao, Shankar Vembu, Amit G. Deshwar, Lincoln Stein, and Quaid Morris. Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC Bioinformatics*, 15(1), 2014.

[280] Amit G Deshwar, Shankar Vembu, Christina K Yung, Gun Jang, Lincoln Stein, and Quaid Morris. PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biology*, 16(1):35, 2015.

[281] Ke Yuan, Thomas Sakoparnig, Florian Markowetz, and Niko Beerenwinkel. BitPhylogeny: A probabilistic framework for reconstructing intra-tumor phylogenies. *Genome Biology*, 16(1), 2015.

[282] Eurofins Genomics. Degenerate bases. https://www.eurofinsgenomics.com/en/products/dnarna-synthesis/degenerate-bases/.

[283] Iain C. Macaulay, Wilfried Haerty, Parveen Kumar, Yang I. Li, Tim Xiaoming Hu, Mabel J. Teng, Mubeen Goolam, Nathalie Saurat, Paul Coupland, Lesley M. Shirley, Miriam Smith, Niels Van Der Aa, Ruby Banerjee, Peter D. Ellis, Michael A. Quail, Harold P. Swerdlow, Magdalena Zernicka-Goetz, Frederick J. Livesey, Chris P. Ponting, and Thierry Voet. G&T-seq: Parallel sequencing of single-cell genomes and transcriptomes. *Nature Methods*, 12(6):519–522, 2015.

[284] Anoop P. Patel, Itay Tirosh, John J. Trombetta, Alex K. Shalek, Shawn M. Gillespie, Hiroaki Wakimoto, Daniel P. Cahill, Brian V. Nahed, William T. Curry, Robert L. Martuza, David N. Louis, Orit Rozenblatt-Rosen, Mario L. Suvà, Aviv Regev, and Bradley E. Bernstein. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, 344(6190):1396–1401, 2014.

[285] Jean Fan, Hae Ock Lee, Soohyun Lee, Daeun E. Ryu, Semin Lee, Catherine Xue, Seok Jin Kim, Kihyun Kim, Nikolaos Barkas, Peter J. Park, Woong Yang Park, and Peter V. Kharchenko. Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data. *Genome Research*, 28(8):1217–1227, 2018.

[286] Sören Müller, Ara Cho, Siyuan J Liu, Daniel A Lim, and Aaron Diaz. CONICS integrates scRNA-seq with DNA sequencing to map gene expression to tumor subclones. *Bioinformatics*, 34(18):3217–3219, sep 2018.

[287] Akdes Serin Harmanci, Arif O Harmanci, and Xiaobo Zhou. CaSpER: Identification, visualization and integrative analysis of CNV events in multiscale resolution using single-cell or bulk RNA sequencing data. *bioRxiv*, 2018.

[288] Kieran R Campbell, Adi Steif, Emma Laks, Hans Zahn, Daniel Lai, Andrew Mcpherson, Hossein Farahani, Farhia Kabeer, Ciara O Flanagan, Justina Biele, Jazmine Brimhall, Beixi Wang, Pascale Walters, IMAXT Consortium, Alexandre Couchard-Cote, Samuel Aparicio, and Sohrab P Shah. clonealign: statistical integration of independent single-cell RNA & DNA-seq from human cancers. pages 1–13, 2018.

[289] Pierre Lao-Sirieix, Brian Rous, Maria O'Donovan, Richard H Hardwick, Irene Debiram, and Rebecca C Fitzgerald. Non-endoscopic immunocytological screening test for Barrett's oesophagus. *Gut*, 56(7):1033–1034, 2007.

[290] Simone Zaccaria and Benjamin J. Raphael. Accurate quantification of copy-number aberrations and whole-genome duplications in multi-sample tumor sequencing data. *bioRxiv*, page 496174, 2018.

Appendix A

Table of variables used in the conliga model

Variable	Description
$z_{r,l}$	Index of hidden state that generated observation $y_{r,l}$ in locus l of chromosome arm r
$y_{r,l}$	Count observation at locus l in chromosome arm r for sample with copy number profile $\boldsymbol{z}$
$x_{k,r,l}$	Count observation at locus l in chromosome arm r for control sample k
$m_{r,l}$	Probability of an aligned sequencing read originating from locus l in chromosome arm r in a control sample
$\hat{m}_{r,l}$	Maximum a posteriori (MAP) estimate of $m_{r,l}$
$\tilde{\theta}_{r,l}$	Probability of observing an aligned read at locus l in chromosome arm r for sample with copy number profile $\boldsymbol{z}$
$\theta_{r,l}$	Probability of observing an aligned read at locus l in chromosome arm r for control sample, k
$\tilde{n}$	Total number of counts for sample with unknown copy number profile. $\tilde{n} = \sum_{r \in \mathcal{R}} \sum_{l}^{L_r} y_{r,l}$
n_k	Total number of counts for control sample k. $n = \sum_{r \in \mathcal{R}} \sum_{l}^{L_r} x_{r,l,k}$
$\tilde{s}$	Inverse-dispersion parameter for sample with unknown copy number profile
s_k	Inverse-dispersion parameter for control sample k
$\pi_{u,v}$	Transition probability from hidden state u to hidden state v
π_u^0	Initial probability of being in state u, i.e. probability of $z_{*,1} = u$

$\hat{c}_u$	Relative copy number for hidden state u
H	Base distribution for the copy number states (Gamma distribution with parameters $\boldsymbol{\lambda}$)
$\boldsymbol{\lambda}$	Parameters of the Gamma base distribution for the relative copy number states
$\boldsymbol{\omega}$	Parameters for the Gamma distribution prior on the inverse dispersion parameter for sample with copy number profile $\boldsymbol{z}$
$\boldsymbol{\psi}$	Parameters for the Gamma distribution prior on the inverse dispersion parameter for control samples
$\boldsymbol{\phi}$	Parameters for the Beta distribution prior on the expected proportions, $\boldsymbol{m}$
α	Hyperparameter of the sticky HDP-HMM
κ	Hyperparameter of the sticky HDP-HMM: the self-transition parameter
γ	Hyperparameter of the sticky HDP-HMM: the greater the value of γ the greater expected number of hidden states
β	The global transition distribution vector
ρ	Defined as $\frac{\kappa}{\alpha+\kappa}$
$\boldsymbol{A}$	Vector representing hyperparameter priors for $(\alpha+\kappa)$, γ and ρ. $A_{(\alpha+\kappa),a}$ and $A_{(\alpha+\kappa),b}$ are the shape and rate parameters of the Gamma prior distribution on $(\alpha+\kappa)$ respectively. $A_{\gamma,a}$ and $A_{\gamma,b}$ are the shape and rate parameters of the Gamma prior distribution on γ respectively. $A_{\rho,c}$ and $A_{\rho,d}$ are the shape parameters of the Beta prior distribution on ρ
L_r	The number of loci in chromosome arm r, where $r \in \mathcal{R}$
$\mathcal{R}$	The set of chromosome arms of interest, for example: $\mathcal{R} = \{\text{chr1p}, \text{chr1q}, \ldots \text{chr22p}, \text{chr22q}, \text{chrXp}, \text{chrXq}, \text{chrYp}, \text{chrYq}\}$
K	The number of control samples
σ_m	The standard deviation for the (Normal) proposal distribution for proposed values of $m_{r,l}$ in algorithm 1
σ_s	The standard deviation for the (Normal) proposal distribution for proposed values of s_k in algorithm 1
$\sigma_{\tilde{s}}$	The standard deviation for the (Normal) proposal distribution for proposed values of $\tilde{s}$ in algorithm 1
$\sigma_{\hat{c}}$	The standard deviation for the (Normal) proposal distribution for proposed values of $\hat{c}_u$ in algorithm 1
S	The fixed truncation level for the Dirichlet approximation to the Dirichlet Process.

$\boldsymbol{T}$	Matrix of state transitions counts, where $T_{u,v}$ represents the number of transitions from state u to v in the current iteration of the MCMC
$\boldsymbol{T}^0$	Vector of state counts for locus 1 of all chromosome arms, where T_u^0 represents the number of loci (at the beginning of a chromosome arm) assigned to state u in the current iteration of the MCMC
N	The total number of iterations in the MCMC
$\boldsymbol{M}, \boldsymbol{W}, \bar{\boldsymbol{M}}$	
$\boldsymbol{h}, \boldsymbol{g}, \bar{K}, \zeta, \tau$	Auxiliary variables used to sample $(\alpha + \kappa)$, γ and ρ

Appendix B

Stan model used for GC and amplification bias correction

```
data {
  int<lower=1>        L;          // number of Loci
  int<lower=1>        J;          // number of samples
  int<lower=0>  X[L, J];          // count matrix of size [L,J]
  int<lower=1>        D;          // number of latent dimensions
  matrix[L, D]    B_mat;
}
transformed data {
  int<lower=0> Sample_total_counts[J];
  for (j in 1:J)
    Sample_total_counts[j] = sum(X[,j]);
}
parameters {
  vector<lower=0>[J] s;         // vector of inverse-dispersions
  matrix[D, J-1]     F;         // Factors
  vector<upper=0>[L] log_eta0;   // reference count proportions
}
transformed parameters{
  // Fill F matrix
  matrix[D, J] F_mat;
  F_mat = rep_matrix(0, D, J);
  // last sample is treated as reference for identifiability
  F_mat[,1:(J-1)] = F;
}
```

```
model {
  // calculate eta
  matrix[L, J] eta_l_j;
  eta_l_j = B_mat*F_mat;
  for(j in 1:J){
    eta_l_j[,j] = eta_l_j[,j] + log_eta0;
  }
  eta_l_j = exp(eta_l_j);

  // Priors
  s ~ gamma(1.5, 1e-6);
  for(j in 1:(J-1)){
    F[,j]  ~ normal(0, 1);
  }
  log_eta0    ~   normal(0, 100);

  // Likelihood
  for(j in 1:J){
    for(l in 1:L){
      X[l, j] ~ beta_binomial(Sample_total_counts[j],
                                        s[j]*eta_l_j[l, j],
                                        s[j]*(1-eta_l_j[l, j]));
    }
  }
}
```