8-2014

Utilizing Haplotypes For Sensitive Snp Array-Based Discovery Of Somatic Chromosomal Mutations

Selina M. Vattathil

Recommended Citation

UTILIZING HAPLOTYPES FOR SENSITIVE SNP ARRAY-BASED

DISCOVERY OF SOMATIC CHROMOSOMAL MUTATIONS

by

Selina Maria Vattathil, B.S., B.A.

APPROVED:

_____

Paul Scheet

_____

Swathi Arur

_____

Ralf Krahe

_____

Jeffrey Morris

_____

Nicholas Navin

APPROVED:

_____

Dean, The University of Texas
Graduate School of Biomedical Sciences at Houston

UTILIZING HAPLOTYPES FOR SENSITIVE SNP ARRAY-BASED

DISCOVERY OF SOMATIC CHROMOSOMAL MUTATIONS

A

DISSERTATION

Presented to the Faculty of
The University of Texas
Health Science Center at Houston
and
The University of Texas
MD Anderson Cancer Center
Graduate School of Biomedical Sciences
in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

by

Selina Maria Vattathil, B.S., B.A.
Houston, Texas

August 2014

ACKNOWLEDGEMENTS

ABSTRACT

Utilizing haplotypes for sensitive SNP array-based

discovery of somatic chromosomal mutations

Selina Maria Vattathil, B.S., B.A.

Supervisory Professor: Paul Scheet, Ph.D.

Somatic copy-number (CN) gains and losses and copy-neutral loss of heterozygosity (CNLOH) frequently occur in tumors and play a major role in the progression of disease by altering gene dosage and unmasking deleterious recessive variants. Characterizing these mutations in an individual tumor sample is therefore critical for research on the relationship of specific mutations to disease outcome and for clinical decision-making based on mutations with known impact. A pervasive hindrance to sensitive detection of these mutations is genetic heterogeneity and high levels of contaminating normal cells in tumor samples, which limit the fraction of cells carrying informative mutations. The method presented here is the first method to utilize population-based haplotype estimates to discover low-frequency somatic kilobase- to megabase-size CN alterations and CNLOH mutations using DNA microarrays. The major innovation of the method is the use of phase concordance as a robust metric to measure evidence of allelic imbalance in the face of sporadic phasing errors in the statistical haplotype estimates and stochastic variation in the microarray data. In addition to presenting a hidden Markov model that uses the phase concordance data to perform agnostic whole-genome discovery of imbalanced regions, we also describe how to test candidate regions, and to infer the haplotype of the major chromosome. We demonstrate through controlled experiments using lab-created tumor-normal mixture samples and *in silico* simulated data that the sensitivity is higher than that of existing methods, detecting specific imbalance events in samples with 7% tumor or less, while maintaining specificity. We also demonstrate the potential of the method via a real-data analysis of genomic mosaicism in the general population using over 30,000 samples that were previously analyzed using another method. We made nearly three times

as many calls in these samples as the previous analysis (1,119 vs. 379), most of which appear to exist at low frequencies. These findings validate recent hypotheses that somatic variation in healthy tissues is more prevalent than had previously been reported, and provides valuable observations of *in vivo* mutations that can be studied to make inference on genetic robustness and how these mutations impact cell fitness.

Table of Contents

List of Illustrations

List of Tables

CHAPTER 1
Introduction to mosaicism, heterogeneity, and SNP array-based detection methods

1.1 Mosaicism and heterogeneity

Mosaicism is the presence of two or more genetically-distinct cell lineages within an individual derived from a single zygote. Novel cell lineages are the products of somatic genome-altering events which may affect entire chromosomes, interstitial or telomeric chromosomal segments, or one or a few nucleotides. Mosaicism is considered separate from chimerism, which exists when an individual carries cell lineages from multiple zygotes, for example when an individual has received an organ transplant or received cells from a fraternal twin *in utero*. The average adult human body is made up of about 10^{13} cells (1), each of which can be traced back along its developmental lineage to the single-cell zygote, and in most organs there is some level of constant ongoing cell renewal from stem cells. So, the number of possible cell divisions at which a somatic mutation can occur is substantial. Assuming there are 10^6 independent stem cells in the intestinal epithelium which each give rise to terminal daughter cells every few weeks, nearly every site in the genome will have been mutated in some cell in the intestine alone by the end of a person's lifetime (2). Depending on the point in an individual's development at which a mutation occurs, the cell population carrying the mutation may exist in only a part of a tissue, in multiple tissues or organs, or in almost all cells in the individual.

In the field of cancer genetics, the term 'heterogeneity' is usually used instead of 'mosaicism' to describe the existence of multiple genetically-distinct cell populations. Since tumor cells are already expected to carry somatic mutations that make them genetically distinct from normal cells, the term usually refers to multiple genetically-distinct populations existing within the subset of an individual's cells that are considered tumor cells.

1.1.1 Somatic mutation mechanisms

The variety in event types is owing to the variety of different mechanisms or processes by which the mutations are generated. I will be focusing here on mutations that generate allelic imbalance, or a departure from the 1:1 ratio of the maternal and paternal alleles.

Chromosome mis-segregation during mitosis results in whole-chromosome loss and gain. The mitotic chromosome mis-segregation rate was estimated at about 0.01 per cell division *in vitro* (3). A cell normally executes processes to prevent mis-segregation during mitosis, but centrosome amplification, hyperstability of kinetochore-microtubule attachments, sister chromatid cohesion defects, and failure of spindle assembly checkpoints can all result in mis-segregation (3). Chromosome loss followed by reduplication of the remaining homologue will result in whole-chromosome CNLOH. Although the loss and gain may occur in either order, it is considered more likely that in these cases the loss occurs first, and then the diploid cell line is strongly selected for over the aneuploid cell line (4). Trisomy rescue is a specific example of chromosome mis-segregation and is unusual in that the inherited genome is more aberrant than the one that results after the mutation. In trisomy rescue, the individual inherits two copies of a chromosome from one parent (usually the mother) and one copy from the other parent. At some point after fertilization, one of the extra chromosomes is not passed on to one daughter cell, leaving a diploid cell. If the chromosome that is lost is from the parent that contributed the extra copy, then the new genome is essentially normal, with the usual level of heterozygosity and proper imprinting. If the lost chromosome is the one from the parent who contributed exactly one copy, then the homologous chromosomes will be of uniparental inheritance, and will have segments of uniparental heterodisomy (homologous segments derived from each of the parent's homologues) and segments of uniparental isodisomy (homologous segments both derived from the same parental homologue, and therefore completely identical, barring any other mutations). Segments of both types of disomy will exist as a result of meiotic recombination during gametogenesis in the parent. Since meiotic non-disjunction is much more likely in females than males, trisomy rescue creates maternal uniparental

disomy (UPD) more often than it does paternal UPD. Mosaicism will exist in the case of trisomy rescue only if the trisomic cell lineage survives. Studies of X-inactivation suggest that trisomic cells are subject to severe negative selection in early embryonic development (5), so the mosaicism in these cases may be very short-lived. This expectation of transiency of mosaicism due to strong negative selection against cells with an aneuploid genome applies even more so to monosomy rescue.

Double-strand DNA break repair is another class of mechanisms that can result in somatic allelic imbalance. The two major subcategories of mechanisms are non-homologous end joining (NHEJ) and homologous recombination. The canonical NHEJ is largely error-free, but when multiple DSBs are present can result in deletions or insertions (6). The homologous recombination mechanisms include the canonical pathway which involves resolution of double Holliday junctions, and several alternative mechanisms including single-strand annealing (SSA), synthesis-dependent strand annealing (SDSA), and break-induced replication (BIR). When recombination occurs during G1 phase, the homologous chromosome is the appropriate template; in G2 phase, the sister chromatid is the preferred template but recombination can also occur with the homologous chromosome. Canonical homologous recombination can result in a crossover product or a non-crossover product. Crossover between homologous chromosomes in G1 results in daughter cells with translocations. Crossover between homologous chromosomes in G2 will either create two daughter cells with translocations or one daughter cell with maternal UPD and one daughter cell with paternal UPD, depending on how the chromosomes segregate. In the latter case, a single crossover event will result in the two cells having CNLOH segments from the crossover point to the end of the chromosome (i.e. terminal CNLOH), while a double crossover will result in an interstitial segment of CNLOH. Mitotic recombination is a major source of mosaicism in normal cells (7), and is often cited as a possible mechanism for the second hit to a tumor suppressor gene in cancer. Studies in mouse hybrids of distantly related strains have shown that the rate of mitotic recombination is positively correlated with the level of sequence identity between homologous chromosomes, and also varies across tissues (8). Crossover results in observable allelic imbalance in the heterogeneous mixture only if the two new populations come to different frequencies, which is possible

by chance but is more likely if the genetic divergence gives one population an advantage over the other. The alternative pathways do not involve crossover, but most pathways including the canonical pathway have the potential for gene conversion, or the unidirectional copying of a segment of a one homologous chromosome to repair the broken chromosome. Gene conversion involves small segments of DNA, from a few base pairs up to 1-2 kb. Another potential source of allelic imbalance is the 5'-3' resection step that occurs during homologous recombination, and which can be so extensive in some mechanisms that large segments or even whole arms can be deleted (6). BIR involves replication using the homologous chromosome of the entire terminal segment of a damaged chromosome, so can result in terminal CNLOH.

Stalled replication forks also initiate repair mechanisms that may leave behind mutations. Mechanisms for dealing with a stalled replication fork include NHEJ, microhomology-mediated end joining (MMEJ), and template switching, with the most popular working models being the fork stalling and template switching (FoSTeS) model of Lee *et al.* and the microhomology-mediated break-induced replication (MMBIR) model by Hastings *et al.* (9). Imperfect DNA repair and inaccurate restart of a stalled replication fork are hypothesized to be the common mechanisms for generation of *de novo* simple deletions and tandem duplications, which are the types of aberrations that make up the bulk of non-recurrent *de novo* copy number variants (9). Although these hypotheses are based on studies of constitutive variants, the observed mutations indicate that these types of *de novo* mutations occurred during DNA replication. Somatic copy number variation appears to occur at a higher rate at segmental duplications and other repeat-rich sequences and therefore may occur by some of the same mechanisms that generate inherited variation (10). Other classes of mutations that have been seen in mosaic form include balanced and unbalanced translocations, ring chromosomes, isochromosomes (11) and single nucleotide substitutions (12).

1.1.2 Chromosomal instability in cancer

The rates of generation of different types of DNA aberrations and the control of which repair mechanism is employed in response to an error together impact the observed rates of different types of mutations. Tumor cell populations, since they often have DNA repair deficiencies and are hyperproliferative, can be expected to produce a different spectrum of mutations (in terms of type and frequency) than normal cells. Chromosome instability, or a high rate of numerical or structural chromosomal aberration, is, in fact, common in most cancer types due to dysregulation of DNA repair and mitosis. The most common cause of whole-chromosome and arm-level losses and gains is defective assembly of the mitotic spindle, especially transient multipolarity and merotelic attachments. Transient multipolarity describes the situation in which a multipolar mitotic spindle is formed in early mitosis due to presence of supernumerary chromosomes, but is converted to a bipolar spindle before anaphase by clustering of the centrosomes into two groups. It is distinct from permanent multipolarity, in which the supernumerary centrosomes are not clustered and the cell therefore undergoes multipolar cell division. Whereas permanent multipolarity causes severe chromosome mis-segregation and daughter cells are unlikely to survive, transient multipolarity creates less severe aneuploidy and can produce viable daughter cells with a new aneuploid genome (13). Transient multipolarity has been observed only in chromosomally instable cancer cells, which emphasizes its importance as a mechanism for generating instability (14). Merotelic attachment describes the attachment of one kinetochore to spindle fibers from both poles, and can lead to aneuploidy (when it creates anaphase lag) as well as mis-segregation of chromosome arms (when the stress from spindle fibers pulling in opposite directions distorts the kinetochore and causes the DNA molecule to shear and break at the centromere). Merotelic attachments occur with increased frequency in cells with spindle assembly checkpoint defects, but can also occur in normal bipolar cells (15). Failure of the mitotic checkpoint increases the rate of uncorrected merotelic attachments and therefore the rate of whole-chromosome and arm-level mis-segregation.

An important mechanism of structural mutations in chromosomally instable cells is breakage-fusion-bridge (BFB) cycles. BFB cycles are initiated when a chromosome with a double-strand break or loss of telomere function is replicated, creating sister chromatids which each have ends unprotected by telomeres. These sticky ends are likely to fuse, creating one dicentric molecule. Pulling apart of the two centromeres during anaphase will cause the chromosome to break into two fragments, each of which segregates to a daughter cell. Since the break in anaphase does not necessarily occur at the site of the initial fusion, the daughter cells may inherit a deletion or a reciprocal terminal duplication. This process repeats in subsequent mitoses until the chromosome acquires a telomere, usually by translocation from another chromosome (16).

Not all tumor cell populations suffer from chromosome instability, nor is chromosome instability a prerequisite for structural aberrations and aneuploidy. Aberrations may arise as one-off events and be transmitted as stable variation to daughter cells. Large-scale aberrations and aneuploidy are expected to occur with low frequency in normal cells and may serve as tumor-initiating events, and especially as the 'second hit' that unmasks existing functional variation. However, they do not inevitably lead to malignant transformation (13).

1.1.3 Clonal expansion of somatic mutations

In the strictest sense, the existence of one genetically mutated cell is enough to qualify an individual as mosaic, but a single mutant cell among thousands or millions of normal cells is phenotypically irrelevant. Once a mutation occurs, many factors come into play to determine whether that cell will divide and give rise to a clone of mutant cells, what the size of that clonal population will be, how long it will persist, and what difference it will make to the individual. What is the nature of the mutation and what genes or other functional elements are impacted? Are those elements functionally relevant to the specific cell given the tissue type and the developmental stage of the individual? Does any functional alteration impact the proliferation or growth potential of the cell relative to other cells that are competing for the same resources? Does the mutation's impact on genome structure impact the

stability of the cell? Will the cell be passing the mutation on to daughter cells or is it a terminal cell in a lineage? Some mutations will be deleterious and cause the cell to malfunction and the clone to be eliminated by negative selection. Many, however, will have no net effect on the cell either because they are functionally neutral or affect a gene for which there are redundant or compensatory mechanisms, and may come to detectable frequency by genetic drift. Still other mutations will be advantageous to the cell and may experience positive selection.

1.1.4 Systemic consequences of mosaicism

Whether or not a somatic mutation persists depends primarily on how it impacts fitness at a cellular level. The impact of the mutation on an individual level is a different evaluation altogether. The phenotypic impact of mosaicism on the individual can take a wide range. Although it can be best described as a spectrum of consequences, here I will present examples of phenotypes caused by mosaicism in the following four categories: cancer, non-malignant disease, negligible health impact, and restoration of health.

Cancer

Cancer is a highly heterogeneous set of diseases whose single most defining characteristic is the acquisition of somatic mutations that transform a normal cell into a malignant population. Cancer is by far the most prevalent known example of mosaicism-associated disease, and is special among mosaicism-associated phenotypes for several reasons. First, by definition, cancer cells acquire accelerated proliferative capacity and immortality, which allows the mutant clone to increase greatly in size. This characteristic makes the presence of mosaicism easier to detect. Second, cancer is exceptional in the multitude of somatic genomic alterations that have been associated with it. For example, reports of whole-chromosome deletion and duplication exist in the Mitelman database for every chromosome (17). Not only do cancer cells exhibit a range of events, most cancer cells carry multiple somatic mutations. Third, a cancer cell population may comprise multiple genetically distinct subclones due to a driver mutation that causes chromosomal instability. Often, the sheer number of mutations that are

observed in a given cancer cell population and genetic heterogeneity within the population interfere with the identification of the somatic mutations that drive the cancer initiation and progression. Despite its prevalence and the obvious causal nature of somatic mutations, the complexity of the mechanisms in any individual's cancer and the variation among individuals makes understanding cancer very difficult. Interestingly, the extreme level of genomic aberration observed in many cancers highlights the high level of robustness of the human genome and in some ways help to support the notion that sporadic random somatic mutations can be of little consequence and should be expected at a low frequency in normal tissues.

Non-malignant disease

Some inherited genetic diseases are known to also exist in mosaic form. In some diseases the mosaic form is phenotypically less severe than the constitutional form, with the severity proportional to the number of cells carrying the somatic mutation. For example, mosaic versions of all of the inherited trisomies (chromosomes 13, 18, 21) are less severe than the inherited forms. In other diseases, the inherited and mosaic forms produce qualitatively different phenotypes (11). For example, inherited *HRAS* G12S mutations are associated with Costello syndrome, which includes developmental delay and specific skin phenotypes, while mosaic *HRAS* G12S mutation in epidermal tissue has been associated with sebaceous nevi, a different skin phenotype (18). The difference in effects is likely to be related to the specific tissues affected in the mosaic case and the timing of the mutation.

A number of genetic diseases can only be caused by somatic mutations and cannot be inherited or passed to offspring, either because the mutation disrupts gametogenesis or is embryonic lethal. Examples of this class of mosaic disorders include McCune-Albright Syndrome, Pallister-Killian Syndrome, and Beckwith-Wiedemann Syndrome.

Another important category of mosaic diseases is skin diseases, since they have visually obvious manifestations and have historically been easy to identify as mosaic. The observation that the patterning

8

of the mutant cells followed the known patterns of cell migration prompted the early models of clonal mosaicism (11, 19).

Mosaicism may have a particularly important role in the brain, both in the context of brain disorders and perhaps also in normal brain function. Somatic mutation has been observed in monozygotic twin pairs discordant for neurodegenerative disease, and has been hypothesized as a potential common cause of sporadic cases of neurodegenerative disease (20). One study (21) used a read depth binning method applied to single-cell sequence data from 110 human frontal cortex neurons collected posthumously from 3 individuals and found that 45 cells (41%) exhibited at least one somatic CNV. Some cells had low-level aneuploidy, whereas others were observed to have only one or two small or subchromosomal CNVs. LINE-1 transposable elements have already been suggested as a mechanism for generation of neuronal plasticity (22). The high rate of somatic CNVs observed in the study suggests another mechanism for genetic diversification in neuronal cells, and the presence of mutations in all three individuals examined lends credence to the hypothesis that somatic mutation is part of normal neuronal function. The authors further argue that maintenance of genomic plasticity in human neurons may actually have been selected for since it allows for an individual to express a greater range of behavioral phenotypes from a single inherited genome (22).

The possibility of cryptic (undetected) mosaicism has been gaining ground recently as a partial explanation for genetic diseases observed to have unusual inheritance patterns. For example, it has been reported that an increasing number of individuals diagnosed with diseases that are presumed to be caused by *de novo* mutation have actually been found to carry mosaic somatic mutations (7). Somatic mutation is an especially plausible explanation in diseases where sporadic cases are associated with a later age of onset compared to cases with familial inheritance (20).

Mosaicism has been associated with complex disease as well. A recent study found a five-fold higher rate of large (>2 Mb) mosaic mutations in blood samples from individuals with Type 2 diabetes (T2D) compared to individuals without T2D. Since almost all of the mosaic mutations they discovered were also reported in studies of mosaicism in the general population, the authors interpret the results as

evidence that T2D increases the risk of blood mosaicism, not that mosaicism increases risk of T2D. Combined with the idea that age-related genetic instability increases disease risk by causing unbalanced expression or unmasking of deleterious recessive alleles (23), this conclusion suggests the avenue by which T2D creates an 'accelerated aging' phenotype, conferring higher risk for age-related diseases such as cancer and cardiovascular disease.

Mosaicism in healthy individuals

An increasing amount of observational evidence from primary tissue samples indicates that base substitutions and large chromosomal changes occur with low but non-zero frequency during normal tissue maintenance and do not necessarily lead to disease. For example, Tomasetti *et al.* (24) used somatic point mutation data from tumor whole-exome sequencing to test the relationship between number of mutations and patient age, and to estimate what fraction of the observed mutations occurred during tumor progression and what fraction occurred before tumor initiation. They found that patient age strongly correlated with the number of mutations in tumors originating from self-renewing tissues (chronic lymphocytic leukemia, uterine cancer, and colorectal cancer), but not in tumors originating from non-self-renewing tissue (pancreatic ductal adenocarcinoma). By assuming that the time from tumor initiation to diagnosis was approximately the same across individuals, the authors were able to estimate the number of mutations that occurred during tumor progression and subtract this number from the total number to estimate the number of mutations that occurred before tumor initiation, and therefore represent neutral mutations that occurred during normal tissue self-renewal. They combined these numbers with patient age information to estimate somatic substitution rates of 6.4×10^{-10} and 7.6×10^{-10} mutations per base per cell division in lymphocytes and colorectal epithelial cells, respectively.

Several recent large-scale studies used SNP array data to directly analyze the rate of somatic segmental copy-number variation and CNLOH in the general population. Laurie *et al.* (25) analyzed 50,222 samples from 15 different case-control GWA studies collected as part of the Gene-Environment Association Studies consortium. The study phenotypes included cancers and non-cancer conditions, and

the subjects had a wide age range, from newborns to subjects more than 80 years old. They found that the prevalence of a detectable mosaic event was low for younger age groups but increased to 2-3% for subjects more than 80 years old. They also found a tenfold increased risk of incident hematological cancer for individuals with detectable blood mosaicism compared to subjects without detectable blood mosaicism, suggesting that mosaic mutations may be a biomarker for cancer risk (although the absolute increase in risk was low since the cancer incidence was low, as the authors point out). Jacobs *et al.* (26) also found that the prevalence of mosaicism increased to about 2% in cancer-free individuals older than 75 years, and found an even stronger relationship between blood mosaicism and incident hematological cancer (odds ratio 35.4).

Another study (27) also sought to investigate the relationship between mosaicism and age, but collected SNP array data expressly for the purpose instead of using existing GWAS data. The Forsberg *et al.* study was also unique in that it used comparison of monozygotic twin pairs to confirm the somatic origin of mutations and performed orthogonal validation (using qPCR and custom tiling-path oligonucleotide arrays) of mutations identified using Illumina arrays. They found that both large megabase-range mutations and smaller kilobase-range variants were more common in older individuals. The samples they used were from a longitudinal cohort study, and for a subset of samples they analyzed samples taken at different timepoints taken several years apart. They observed from these intra-individual comparisons that the intra-individual frequency of clonal populations sometimes increased and sometimes decreased over time, suggesting that the mutations increased the proliferative capacity of the cells enough to bring them to observable frequency but that the mutant cells were not immortalized.

Somatic reversion

At the positive end of the spectrum are somatic mutations that result in complete or partial restoration of a normal phenotype in individuals with an inherited disease variant and associated disease. Somatic reversion includes back mutations which exactly reverse the inherited variant (for example, point substitution of an abnormal allele to the wild-type allele), or compensatory mutations such as an insertion

11

that corrects a frameshift caused by a deletion elsewhere in the gene. At first blush, somatic reversion

may seem like lightning striking the same location twice. However, Davis and Candotti (28) suggest that

the rate of somatic mutation is high enough that these types of mutations happen often, and the timing of

the mutation early enough in the developmental pathway on the cellular level is the most important factor

influencing whether or not restoration of normal phenotype is detected on the organismal (human) level.

Implicitly, reversion seems more likely if the mutation that confers a normal phenotype also confers a

selective advantage to the cell. This is true in the example of Wiskott-Aldrich syndrome (WAS). WAS

is caused by a loss of function of the WAS gene, which impairs T-cell proliferation. T-cell progenitors

with reversions will produce more T-cells than non-revertant progenitors, and the abundance of T-cells

will in turn restore immune function. An alternate hypothesis is that cases of reversion are associated

with higher mutation rate, either globally or at the site of the causative mutation.

1.2 Detecting somatic mutations -- SNP arrays and other methods

Many techniques exist for detecting mosaic mutations, each with unique strengths and weaknesses.

There is no single strategy that meets all of the criteria one would consider, such as sensitivity for events

of all types and genomic sizes and all cell fractions, ease of execution, ease of interpretation, and

requirement for *a priori* knowledge of genomic location. In the remainder of this chapter I will first

summarize the strengths and weaknesses of the most common techniques. Then I will provide a more

thorough explanation of how SNP arrays are used for mosaic mutation detection.

1.2.1 Summary of common techniques

G-banded karyotyping is a common technique for perinatal genetic testing of individuals suspected to

carry chromosomal aberrations, i.e. because of observation of phenotypes known to be associated with

chromosomal aberrations. Karyotyping has been used for decades, so is well established and remains the

most reliable and trusted method for detecting large constitutional abnormalities, and is among the subset

of methods that can reveal balanced rearrangements such as inversions and translocations. However,

12

making and reading karyotypes is time-consuming and labor-intensive and therefore usually only a modest number of cells are examined, at least in the initial screening of an individual; commonly 20 cells are assayed (29). Since mutations can only be detected if they are present in the sampled cells, a limited number of cells may translate to limited power to detect low-frequency mosaicism. For example, analysis of 20 cells has a 95% chance of detecting a mosaic mutation present at a proportion of 14% (30). Of course, follow-up may be conducted on individuals for whom alternative lines of evidence suggest low-level mosaicism. G-banding is too low-resolution to be useful for detecting mutations smaller than several megabases (31), and will not provide evidence of CNLOH.

Fluorescence *in situ* hybridization (FISH) is another microscopy technique. FISH involves designing probes to target specific genetic sequences or structures, such as particular genes or the centromere or telomere. Probes labelled with differently colored fluorescent markers can be combined in a single experiment to interrogate the copy number and chromosomal location of multiple targets at once, and can therefore be used to identify copy number variation as well as translocations and inversions. Like G-banding, usually only a limited number of cells is assayed at a time, although more cells may be visualized at once especially if the interest is only in copy number of one or a few regions (since the number of fluorescent spots per cell can be counted without requiring very high image resolution). FISH can reveal variation involving small genetic regions (a few kilobases (32)). Spectral karyotyping (SKY) is related to FISH in that it involves combinations of colored probes, but instead of targeting a few specific regions it involves 'painting' of each chromosome or even arm in a unique color, so karyotypes can be assembled automatically using computer software that recognizes the signal for each chromosome. Like G-banded karyotypes, the sensitivity of both FISH and SKY analysis for discovering low-frequency mosaic or heterogeneous mutations can be low depending on how many cells are assayed. Another factor is that, since these techniques require growing and stimulating cells in culture, estimates of the mosaic cell fraction may be inaccurate if the mutant cells are subject to different negative or positive selection pressures in culture than *in vivo* (31).

Several innovative methods based on PCR have recently been developed and are useful for assessing copy number of specific genetic regions. Real-time quantitative PCR (qPCR) is commonly used to detect single nucleotide polymorphisms and can also be used to detect constitutive copy number variation using a protocol that involves comparison of signal from test and reference copy number probes (32, 33). However, the precision of the test is not sufficient to detect mutations if the mutant cell fraction is very small or to discriminate mosaic from constitutive mutations if the mutant cell fraction is very high and there is no germline sample available. Other allele-specific qPCR assays have high sensitivity and specificity for mutations that are rare within a sample and could be used for validation of suspected heterogeneous CNVs using inherited heterozygous sites. These would not be efficient methods for discovering variation, but qPCR can be useful for validating mosaic mutations (27).

Multiplex ligation-dependent probe amplification (MLPA) for detecting copy number variation involves PCR using fluorescently-labeled ligation-dependent probes targeting suspected CNV loci followed by capillary gel electrophoresis of the PCR products. Probes for different loci are designed to have different sizes of amplicons, so the height of peaks at different locations in the electropherogram will be proportional to the concentration of copy number alleles in the sample. By designing probe sets to have different amplicon sizes for different target regions, multiple regions can be tested at once (32). Once a probe set has been designed, MLPA is an efficient experiment for inferring copy number for tens of loci of interest with one reaction per genomic sample and has enough precision to identify mosaic mutations (34) with sizes ranging from single nucleotides to multi-megabase events (by using multiple probes) (34), but is not suited for mutation discovery without a priori knowledge of variant loci of interest.

Whole-genome sequencing (WGS) can detect heterogeneous copy-number variation, CNLOH, and balanced translocations. Mate-pair and paired-end sequencing are especially useful (compared to single-end sequencing) for detecting translocations and small insertions and deletions, since they improve read mapping. Analysis of read depth can reveal larger copy number changes. WGS has the potential to provide very complete resolution of a heterogeneous sample. On the other hand, several

14

factors make it less than ideal. First, if the goal is detection of mosaic mutations, high coverage of the genome must be achieved, which will make the experiments expensive. Second, the scale of the data and the complexity and sheer amount of computer processing required to transform the raw data means that sequence analysis is a major undertaking unless a pipeline has been set up (and maybe even then as well).

High-density whole-genome SNP arrays provide somewhat of an intermediate option between traditional cytogenetics and targeted PCR techniques and WGS. The sample preparation is relatively straightforward and akin to preparation protocols for PCR (DNA must be extracted and purified from tissue, but no cell culturing or library preparation required), it does not require *a priori* knowledge of the locations of mutations, and some data processing and informatics is required to obtain interpretable results, but requires computing power on the order of a desktop machine instead of a computing cluster. During the past decade, SNP arrays and array comparative genomic hybridization have supplanted cytogenetics techniques as the first-line tests in prenatal and neonatal genetic diagnostic laboratories (11) and have immensely increased the rate of detection for mosaic segmental and whole-chromosome abnormalities in that setting. SNP arrays have also made it possible to conduct large-scale surveys of somatic mosaicism, as discussed above, and of chromosomal abnormalities in cancer. SNP arrays are useful for detecting mosaic whole-chromosome or segmental aberrations, but generally not point mutations, since only a small fraction of sites are measured directly. The two major manufacturers of SNP arrays are Affymetrix and Illumina. They employ different chip manufacturing techniques and probe designs, and so the native data types are unique and have different error characteristics and of course each manufacturer has unique proprietary normalization procedures. Illumina is now providing arrays for targeting 2.5 million or even 5 million SNP and CNV loci.

SNP arrays were originally designed for the purpose of calling genotypes at each marker, and the basic calling algorithms in use today still assume that the majority of the sample is diploid, which is a reasonable except for studies of tumor samples. A number of publically-available methods have been developed for inferring constitutive and mosaic copy number aberrations. These usually rely not only the

15

called genotypes but also on intermediate-level data, which may be subject to bias and miscalibration and always display some imprecision. Some of the common sources of bias include GC content of the target region and inherent differences in signal intensity from the different fluorescent tags. Miscalibration can arise when the sample is highly aneuploid, since the baseline values are calculated from the sample itself using an algorithm that assumes that most of the genome is diploid. The random and non-random non-biological variation in SNP array data makes the detection of mosaic events from them a non-trivial endeavor. Much research has been aimed in recent years at improving normalization strategies such that the observed data accurately represents the underlying sample characteristics, and at interpretation strategies for extracting the information of interest from the normalized data. In section 1.2. I will review the historically important and state-of-the-art methods for inferring copy number changes and CNLOH in heterogeneous samples; in order to understand the basis of these methods, I will first provide some background on two intermediate data types, the B allele frequency and the log R ratio, that are commonly used as observed data in these methods. These two data types are the result of normalization and processing of even lower level observations; I will not provide details on the normalization methods (unique methods for Affymetrix and Illumina arrays), although they are also a field of ongoing research. The BAF and LRR data types are native to Illumina arrays; they are not native to Affymetrix data but are a natural way to represent the underlying state at each marker and Affymetrix data can be coerced to produce them.

1.2.2 SNP array data

Each SNP marker on a SNP microarray is designed to target a biallelic variant site. The two alleles are labeled as 'A' and 'B' (arbitrarily in the case of Affymetrix arrays, and according to an algorithm based on the specific alleles and sometimes on sequence context in the case of Illumina arrays (35)). At each marker, the raw intensity data are transformed to make genotype calls. In addition to the genotypes, two intermediate data types, the B allele frequency (BAF) and the log R ratio (LRR), can be calculated from the raw data. The BAF represents the proportion of sampled chromosomes that carry the B allele. The

16

LRR is a measure of the average per-cell copy number. These data types are native to the Illumina

platform and are easy to output using Illumina's proprietary GenomeStudio software. With Affymetrix

SNP arrays, the BAFs and LRRs may be calculated from .CEL files using Affymetrix Power Tools.

        With Illumina arrays, genotype calls are made for each marker by comparing the observed BAF

and LRR values to expected values derived using data from control samples. Illumina provides sets of

expected values for each array that have been calculated using HapMap data. For a copy number of 2,

the expected LRR value is 0. For the three diploid genotypes at a biallelic marker AA, AB, and BB, the

expected BAF values are 0, 0.5, and 1, respectively. Visualization of the BAFs and LRRs for a normal

region emphasizes a few important points (Figure 1). First, the points in the BAF plot form three distinct

bands roughly located at 0, 0.5, and 1, one for each of the three possible genotypes. In other words, the

data for a good-quality DNA sample that has been properly processed is quite predictable, and outlying

points are usually few and easy to identify. Second, all three bands exhibit some noise. The two

homozygote bands tend to be less noisy than the heterozygote band, probably because of an explicit

shrinkage that is performed on BAFs that are close to the extremes of the value range (36). While the

imprecision in the heterozygote BAFs in a normal sample has little to no impact on the genotype

determination, we will see that it is the reason that SNP array-based strategies for detecting allelic

imbalance encounter a loss of sensitivity for low-proportion events.
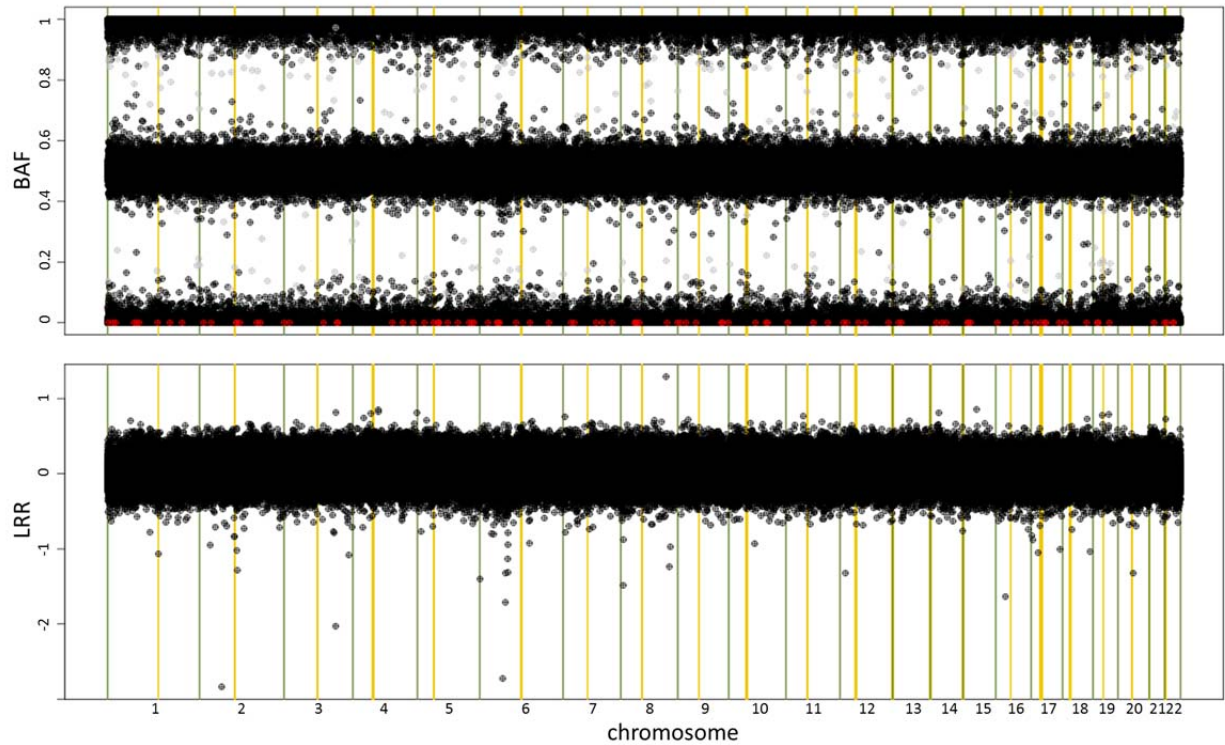
Figure 1. Examples of whole-genome B allele frequency (BAF) and log R ratio (LRR) data. Top plot

shows BAFS, bottom plot shows LRRs. Green lines indicate chromosome boundaries and yellow lines

indicate centromere regions. Points are plotted by marker index, so chromosomes are scaled by marker

count, not genomic size. In the BAF plot, grey points indicate that no genotype call was made at the

marker, and red points indicate markers where no BAF value was produced (they are arbitrarily plotted

at 0).

The data for a region of a genome that is in imbalance or has undergone a copy number change

will be similarly predictable. When there is complete loss of heterozygosity, all of the BAFs will be

close to 0 or 1, and there will be no middle band of points. When the sample contains a mixture of

normal cells and cells with allelic imbalance, the BAFs at inherited heterozygous markers will be shifted

away from the expected heterozygote value (generally 0.5), but less so than in a 100% aberrant sample.

Figure 2 presents examples of BAF plots expected for a samples that contain mixtures of normal cells

and mutant cells with a hemizygous deletion (top row), CNLOH (middle row), and duplication (bottom

row) at mutant cell fractions 0%, 5%, 10%, 15%, 20%, 50%, and 100%. Even though both hemizygous

18

deletion and CNLOH create loss of heterozygosity, the two types of mutation create unique BAF patterns even when they occur at the same frequency because they result in different copy numbers. Duplications do not create LOH, and the shifts in the BAF pattern are smaller than either hemizygous deletions or CNLOH occurring at the same frequency. In other words, the expected distance between the observed BAFs and the expected BAF at heterozygote markers (generally 0.5) is a function of both the mutation type and the fraction of the sampled cells that carry that mutation. A similar relationship exists for the LRRs, but note that LRRs in CNLOH mutation regions have the same expected LRR as those in normal regions. The theoretical expected magnitude of the observed deviations given the event type and mutant cell fraction are given by

$$dev_{BAF} = \frac{2(1-f)(0.5) + x_t(f)(\frac{x_n}{x_t})}{2(1-f) + x_t(f)} \text{ and } dev_{LRR} = \log_2\left(\frac{2(1-f) + x_t(f)}{2}\right),$$

where f is the fraction of mutant cells, x_t is the total copy number in the mutant cells, and x_n is the maximum of the allele-specific copy numbers in the mutant cells (so $\frac{x_n}{x_t}$ will be $\frac{1}{1}$, $\frac{2}{2}$, and $\frac{2}{3}$ for hemizygous deletions, CNLOH, and duplications, respectively). Since the normalization procedure that generates the BAFs and LRRs from the raw measured data may not be perfectly tuned, other factors such as scale or shift parameters may be included in these equations when they are applied in practice. These equations are for the simplest case of heterogeneity, where there is a normal cell population and exactly one mutant cell population. In more complex situations the same principle of the observed values being a function of the weighted average of the expected values for each population applies, but of course all of the populations would need to be factored in order to make accurate inference.
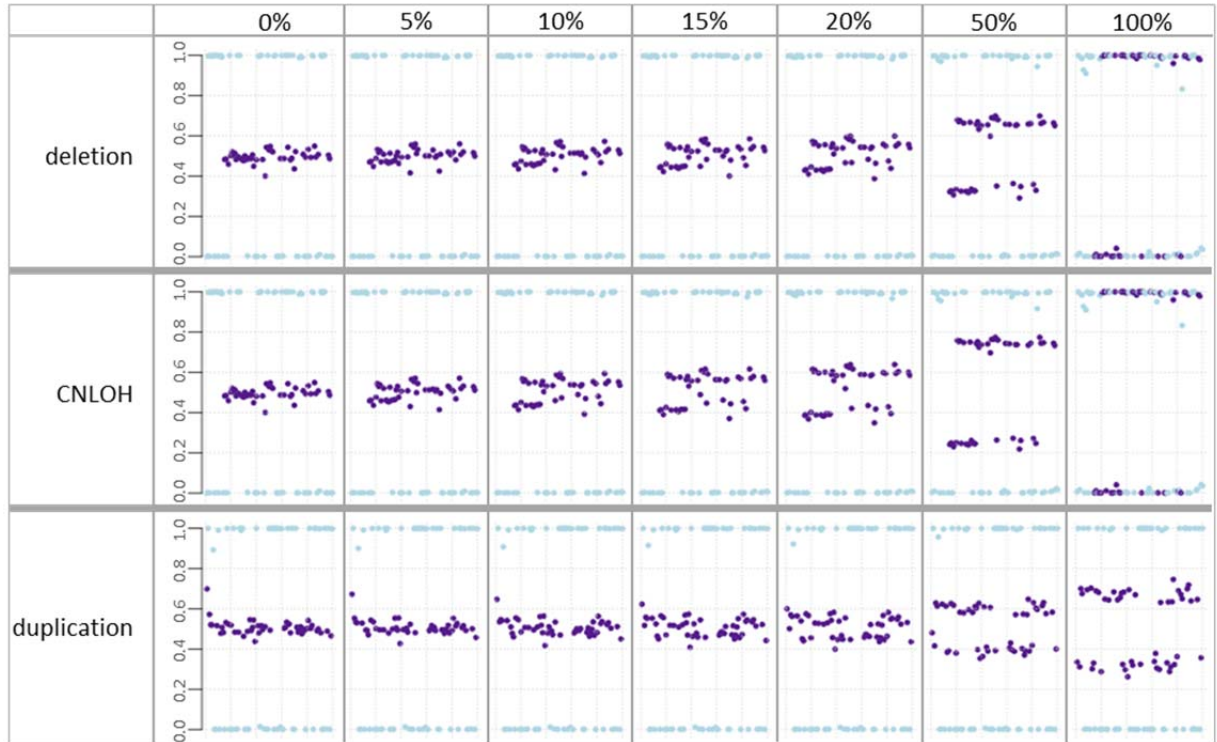
Figure 2. Examples of BAF plots for mosaic deletions, CNLOH, and duplications. Data were simulated for proportions other than 0% and 100% by weighted averaging of the observed heterozygous BAFs from constitutively normal and constitutive homozygous (for deletion and CNLOH) and truly duplicated (for duplication) regions. Purple points are BAFs at heterozygous markers and light blue points are BAFs at homozygous markers based on genotype calls from the normal (0%) sample.

1.2.3 Review of existing methods for detecting low-proportion chromosomal aberrations

All existing methods depend on 1) there being a consistent pattern in the data from within a region affected by a single allelic imbalance event and 2) the pattern in abnormal regions being distinguishable from the pattern in normal regions. So, genomewide characterization of allelic imbalance status can be thought of in terms of the need to identify the breakpoints between normal and abnormal regions (or between adjacent abnormal regions arising from distinct mutations), and the need to interpret the observed data in abnormal regions to infer the specific chromosomal aberration (i.e. copy number,

fraction of sampled cells affected, the specific haplotype affected). Some methods perform first

breakpoint detection and then characterization, while other methods perform the two procedures jointly

or iteratively. Existing methods for the most part employ one of two breakpoint detection strategies –

circular binary segmentation (CBS) or hidden Markov models (HMMs). CBS was introduced by Olshen

and colleagues (37) in the context of detecting copy number variation using aCGH data. In this method,

the data (i.e. LRR values) across markers in a normal region are considered to be independent and

identically distributed (IID) following a Gaussian distribution, and the data in abnormal regions will

follow a Gaussian distribution with a different mean than normal regions. For each possible breakpoint

(usually the breakpoints will occur between a consecutive pair of data points), a likelihood ratio is

calculated to assess whether the observations to the left and the right of the border have different means,

and a breakpoint is assigned to the segmentation that produces the highest likelihood ratio above a

specified threshold. The procedure is then repeated recursively within each newly defined segment. The

advantage of CBS is that the only assumption is that the data are normal, and it is robust to that

assumption (which is useful, since BAF and LRR data often have a more heavy-tailed distribution). CBS

is therefore a good option for segmenting highly complex samples, such as tumor samples. One major

drawback is that, since BAFs in an abnormal region are likely bimodal, BAF data must be transformed to

obtain something approximately normal before CBS can be used. Most often, a mirroring transformation

is applied. This transformation step greatly reduces the sensitivity for detecting low-frequency mutations

(38). Another characteristic that could be considered a drawback is that it can only be applied to one-

dimensional data, so BAFs and LRRs must be analyzed separately and the results somehow reconciled.

A joint analysis offers potential for greater power.

Hidden Markov models underlie the second major category of segmentation algorithms. In the

case of SNP array data that has been successfully normalized, it is usually reasonable to assume that the

observations at each marker are independent conditional on the underlying mutation state, making it

possible to use an HMM where the BAF and LRR data are the observed data and some representation of

the genomic state is the hidden state. Most HMM methods use the possible allele-specific copy numbers

21

in the mutant cells as discrete hidden states, and include a parameter representing the frequency of mutant cells. Examples include OncoSNP (39) and GPHMM (40). One method, PSCN (41), uses a continuous-state HMM, where the latent state is a two-dimensional variable representing the average frequency in the sample of each of the two inherited alleles at an inherited heterozygous site. A summary of the input data and segmentation methods for the aforementioned implementations and others is presented in Table 1.

Name	Year	segmentation strategy	input data	paired/unpaired/trio	CNLOH
DNAcopy (37)	2004	CBS	LRRs	unpaired	no
AsCNAR (42)	2007	HMM	Allelic intensities	unpaired or paired	yes
SOMATICs (43)	2008	Adaptive Weights Smoothing	BAFs, LRRs†	unpaired	yes
BAFsegmentation (44)	2008	CBS	BAFs	unpaired or paired	yes
genoCNA (45)	2009	HMM	BAFs, LRRs, population allele frequencies	unpaired or paired	yes
GAP (46)	2009	CBS	BAFs, LRRs, genotypes	unpaired	yes
gBPCR (47)	2010	Bayesian PCR	genotypes, LRRs	unpaired	yes
OncoSNP (39)	2010	Bayesian HMM	BAFs, LRRs	unpaired	yes
ASCAT (48)	2010	PCR	BAFs, LRRs	unpaired or paired	yes
MAD (49)	2011	Sparse Bayesian Learning	BAFs, LRRs	unpaired	yes
GPHMM (40)	2011	HMM	BAFs, LRRs	unpaired	yes
PSCN (41)	2011	Continuous-state HMM	BAFs, total intensities	unpaired or paired	yes
TAPS (50)		--	Allelic intensities	unpaired	yes
HAPSEG (51)	2011	CBS	Allelic intensities, haplotype estimates†	unpaired	no
Battenburg algorithm (52)	2012	PCR	BAFs, haplotypes from matched normal sample†	paired	yes
POD (53)	2013	Binomial testing in overlapping windows	genotypes, BAFs, LRRs†	trio	yes

Table 1. Existing SNP array-based methods for allelic imbalance in heterogeneous samples. CBS-circular binary segmentation. HMM-hidden Markov model. PCR-piecewise constant regression. BAF-B allele frequency. LRR-log R ratio. †data type is not used for segmentation, only for post-segmentation characterization.

So far, I have emphasized that the BAFs at inherited heterozygous markers in abnormal regions exhibit an increased deviation from 0.5, and the magnitude of the deviation is predictable given the mutation type and affected cell fraction. The deviations can be described not only by their magnitude, but also by their direction, i.e. towards 0 or 1. The direction is determined by the allele configuration (haplotypes) of the inherited chromosomes, and which of them has been lost or gained in the mutant cells. To use the direction of deviations to assist in the detection or characterization of allelic imbalance regions requires knowledge about the inherited haplotypes. Until recently, no methods used the direction information, either ignoring it by taking the absolute value of the deviations or integrating over the two possible directions. Two recently published strategies make use of best-guess haplotypes estimated from population genetics-based statistical methods to improve post-segmentation characterization of abnormal regions, and one uses highly accurate haplotypes inferred from trio data to perform segmentation.

The goal of the first method, HAPSEG (51), is to estimate for each abnormal region the average ratio of the copy number of each of the haplotypes, or homologues, per cell relative to the inherited copy numbers (which would be 1 for each haplotype in a normal diploid region). The algorithm involves first performing segmentation using summed allele-specific intensities, and then uses individual allele-specific intensities (similar to BAFs) to make an initial definition of the inherited haplotypes. The assumption is that in regions where the level of imbalance is high, the two BAF bands in the germline heterozygous regions will be sufficiently diverged to provide accurate observation of the true haplotypes. In more subtly imbalanced regions, where the two BAF bands overlap, the statistically-estimated haplotypes will provide a more accurate estimate of the true haplotypes. The two sets of haplotypes (BAF-estimated and statistically-estimated) are compared in the called abnormal regions, and where it appears that the BAF-estimated haplotypes are noisy, they are corrected using the statistically-estimated haplotypes. Homologue-specific copy ratios are then calculated for each segment by smoothing the values at each marker.

The second strategy was published in a study of breast tumor evolution that collected SNP array and sequence data from tumor and matched normal samples (52). The first step in the procedure, which

the authors refer to as the Battenburg algorithm, is to estimate inherited haplotypes using the genotypes from the normal sample and a population genetics-based statistical method (Beagle (54) was used in the published experiment). Then, if the frequency of each haplotype at each marker is plotted in a different color (for example red for one haplotype and blue for the other haplotype), in regions of imbalance the two haplotype frequencies will form two bands that will switch whenever there has been a phasing error, creating a plot that resembles a 'Battenburg marking', or a two-color checkerboard pattern. In normal regions the two bands will overlap. The evidence for distinct versus overlapping bands is then assessed. Neither the main text nor the supplemental materials explicitly mention the specific testing strategy, although the authors mention comparing observed haplotype frequencies to the expected frequency of 0.5, and this could be done using a t-test if a region of interest is defined. It appears that in the published analysis the approach was applied to each chromosome, which would be a powerful approach for discovering low frequency aneuploidy.

Both of the above methods first define segments of interest, and then use haplotype information to perform characterization within the segment. Only one existing strategy uses haplotype information to perform segmentation. The POD method (53) first uses trio (mother, father, offspring) genotypes to estimate haplotypes for the offspring, then identifies potentially abnormal (over-represented) alleles by comparing the BAF at each marker to a threshold, then infers the parental source for any abnormal alleles. Using markers with abnormal alleles, POD discovers segments of over-representation from one parent by applying two-sided binomial tests in overlapping windows and subsequently annotates each segment using LRRs. Trio-based phasing provides highly accurate haplotypes at informative loci and this method is very useful in cases where trio genotypes are available.

1.3 Relevance of new SNP array methods for sensitive detection of low-frequency mutations

Decades worth of case studies and research has demonstrated that genomic mosaicism in humans is not an uncommon phenomenon and that the aberrations that can be tolerated in an individual sometimes defy conventional wisdom. Probably the situation in which the need for sensitive detection is most obvious is

in genetic studies of cancer. One pervasive hindrance to the identification of tumor-associated mutations is the presence of large amounts of contaminating normal tissue in a tumor sample. The tumor cells carrying informative mutations, then, make up only a fraction of the cells in the sample. In some cases, samples contain so little tumor cells that they are unusable. Tumor-normal mixture was the motivating context for our method. The ongoing generation of subclonal mutations, which start out with low frequency, is another reason that relevant mutations may be present in only a fraction of the sampled cells, even in a sample with high tumor content. Therefore, sensitive methods will allow analysis of samples that may previously have been discarded for having low purity, and will provide a more complete characterization than lower-resolution methods. These are important benefits both in the research setting to properly characterize the genetic mutations in a sample so that any potential relationship to phenotype can be accurately described, and in the clinical setting to make sure informative genetic markers are not overlooked.

Outside of cancer, mosaicism and genetic heterogeneity have been most thoroughly studied in individuals that were highly suspected to carry somatic or *de novo* germline mutations because they exhibited phenotypes such as intellectual disability or skin diseases with distinctive spatial patterning known to often be caused by these types of mutations. The landscape of somatic mutations in other populations remains to be explored. Whole-genome sequencing performed as part of the 1000 Genomes Project revealed that phenotypically healthy individuals carry an average of 250-300 inherited variants that are predicted to cause loss of function and several hundred inherited variants that have been associated with disease (55), further underlining the complexity of genotype-phenotype relationships and genetic robustness. Acknowledging that variants that appear to be deleterious may not actually have a significant effect has now become vital to proper interpretation of the genetic variation that is discovered in whole-genome and whole-exome analyses performed to find variants associated with disease. Likewise, a better understanding of the range and impact of somatic variation will undoubtedly provide context that will be valuable for making informed decisions when it comes to studying and interpreting an individual's genetic risk for disease – for example, is organ- or tissue-specific sampling prescribed?

In the context of fundamental biology, characterization of natural somatic mutation is a way to capture the routine maintenance processes that have been executed in a cell population, and looking at the types of mutations observed across samples will show what variation exists in these processes, perhaps across individuals, across tissues, across environments, or over time as individuals age.

The specific benefits of SNP arrays is that they are relatively cheap and efficient, can be applied agnostically, and can potentially detect a wide range of allelic imbalance and balanced copy number changes with sizes ranging from kilobases to whole chromosomes. A number of analysis methods have been developed for using SNP array data for mutation detection from SNP array data from heterogeneous samples, and several publications have acknowledged the potential sensitivity improvement that can be gained by using haplotype information. So far, however, only two strategies have been published for using haplotype information to characterize predefined segments of interest, and only the POD method has used haplotype information as a source of signal for defining segments of allelic imbalance and works only in the special case where trio data are available to infer highly accurate haplotypes.

1.4 Dissertation outline

My dissertation project has focused on the development, implementation, testing, and real-world application of a novel algorithm for using estimates of inherited haplotypes to discover regions of low-frequency allelic imbalance from whole-genome SNP array data. The remainder of this dissertation is organized as follows. Chapter 2 describes the algorithm, beginning with a discussion of the motivation for incorporating inherited haplotypes into a method for detecting somatic allelic imbalance, and how phase concordance allows the use of imperfect haplotype estimates. The subsequent section gives a detailed description of the algorithm we have developed based on this motivation, including notes on the specific implementation available in the hapLOH software that we have written. The final section presents guidelines for the pre-processing and post-processing quality control measures that should be applied to achieve high-quality results from hapLOH, and explains the necessity of each step. Chapter 3

demonstrates the performance of the method using simulated and real data. The first section describes results from a set of computationally simulated data which demonstrates the performance of the method compared to other state-of-the-art SNP array-based methods and touches upon the parameter settings and their impact on results. The second section describes results from analysis of real heterogeneous tumor samples which demonstrates that hapLOH works for data from Affymetrix as well as Illumina arrays and provides an example of how being able to detect low-frequency imbalance changes the interpretation of results compared to a study that employed less sensitive methods. Chapter 4 details an analysis of SNP array data from over 30,000 samples to characterize somatic mosaicism in the general population. Chapter 5 lays out potential short-term and long-term improvements and extensions to the method, and outlines additional applications of the method outside of the ones that are demonstrated here.

CHAPTER 2
A method for combining SNP array data and haplotypes

Portions of this chapter are based upon "Haplotype-based profiling of subtle allelic imbalance with SNP arrays", Selina Vattathil and Paul Scheet (2013), *Genome Research* 23(1): 152-158.

2.1 Use of haplotype information

2.1.1 Biological and statistical rationale

As was mentioned in the previous chapter, all of the existing methods except HAPSEG (51), the Battenburg algorithm (52), and POD (53) consider the magnitude characteristic of BAFs only, ignoring or summing over possible values for the direction characteristic. In our method, we consider the direction of the observed BAF from the expected heterozygote value. Within an imbalanced region, the direction of the deviation at each marker correlates with the allele on the excess haplotype at that marker. In the case of allelic imbalance resulting from somatic mutation, we know that the excess haplotype is one of the germline haplotypes, which means that the directions correlate with one of the germline haplotypes. In a region where there is subtle imbalance, if we pick one of the two germline haplotypes and divide the heterozygous markers into two groups according to which allele is on that haplotype, we would see that the points within a group tend to shift in the same direction (Figure 3, left panel). If we do the same in a normal region, where the deviations are due to measurement error, not imbalance, there is no correlation between group membership and direction (Figure 3, right panel). This means that one way to test for imbalance is to test for independence between the direction of BAF shifts and the group membership that can be defined using the inherited haplotypes. Even the slight correlation created at low aberrant cell fractions is strong enough evidence to reject the null hypothesis of independence. In fact, this is exactly the statistical basis for the POD method. The primary limitation of this strategy is that defining the group memberships requires perfect knowledge of the inherited haplotypes. In cases where trio data is available, perfect haplotypes can be inferred by assuming Mendelian inheritance (and

filtering sites at which there are Mendelian inconsistencies), but only for the subset of markers where the genotypes in the three individuals allows unambiguous phasing.
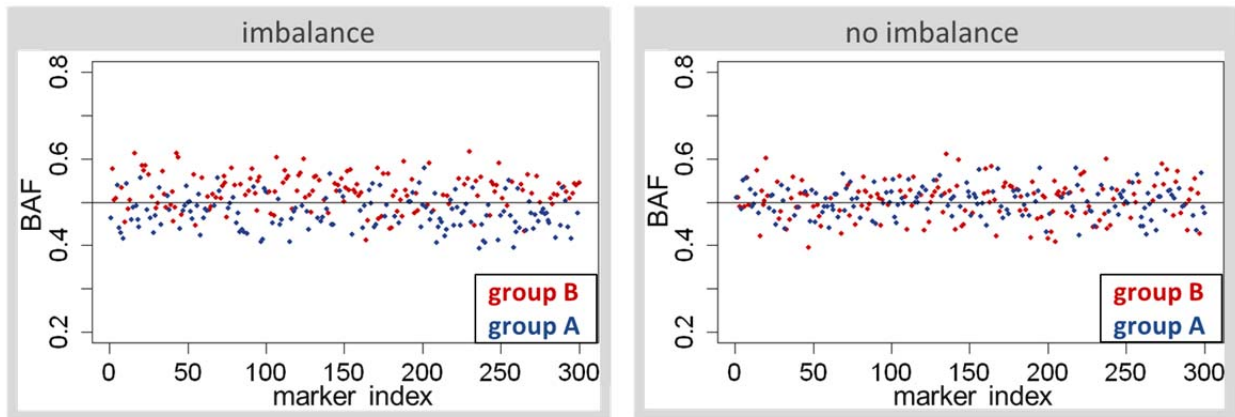
Figure 3. Grouping using haplotype information reveals marker clustering in imbalance regions. The left panel shows a simulated region with CNLOH in 5% of cells, and the right panel shows a normal region. Points were colored by choosing one of the inherited haplotypes and determining the allele on that haplotype at each marker. Only inherited heterozygous markers are shown.

In cases where trio phasing is not an option (and until long-range sequencing or other forms of chromosome separation and typing become scalable), the only way to efficiently make whole-genome estimates of inherited haplotypes is using population-based statistical phasing methods. These methods have high accuracy, but will make occasional phasing errors, such that the best-guess haplotypes will be a mosaic of the true haplotypes with chunks of markers that are perfectly phased separated by switch errors (Figure 4). The rate of switch errors depends on multiple factors including the local recombination rate at the locus, the quality of the reference population, and the specific phasing method, but one can expect accuracies of something like 95%, which translates to an average of one switch error per 20 heterozygous markers.
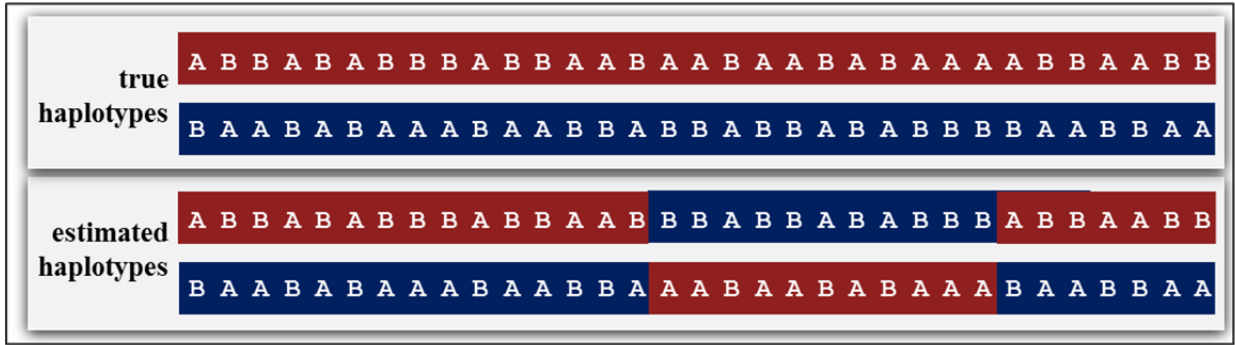
Figure 4. Example of switch errors in haplotype estimates. In this cartoon, the estimated haplotypes contain two switch errors.

2.1.2 Choice of phase concordance as metric

In order to use these imperfect haplotypes to assess the observed directions of BAF deviations for evidence of imbalance, we borrow a metric called switch error rate that is commonly used to assess the accuracy of a set of unordered haplotype estimates. In unordered haplotype estimates, an error occurs when the alleles at one marker are incorrectly phased with respect to an adjacent marker. Only heterozygous markers are considered in this assessment, since it is impossible to make a phasing error at a homozygous marker (because there is only one possible configuration). The switch error rate is then the ratio of the number of switch errors made over the number of consecutive heterozygous pairs. In the example in Figure 4, two switch errors were made out of a total of 35 heterozygous marker pairs, so the switch error rate is 2/35 = 0.04. This could also be expressed in terms of the switch accuracy rate, which in this example would be 0.96.

In our method, we use the switch accuracy metric described above to compare two sets of haplotypes, the statistically estimated set of haplotypes and a set constructed using the BAF deviations that we anticipate will reflect actual allele frequencies in the sample (see next section). We assess phase concordance between these sets by evaluating agreement at each consecutive two-marker haplotype, producing a binary score for every marker pair. The metric is more accurately described as 'switch consistency' than 'switch accuracy', since we are comparing two sets of estimated haplotypes and the

31

true haplotypes are unknown. This local scoring is the critical characteristic that makes this transformation strategy so valuable in the setting we consider, because it allows subsequent testing to be sensitive in the face of both switch errors inherent to statistically-estimated haplotype estimates, and to high levels of noise that will exist in the BAF-based haplotype estimates when the aberrant cell fraction is low. We use the term 'phase concordance' to describe the average switch consistency rate for a region.

2.2 Algorithm and implementation details

2.2.1 Calculation of phase concordance

We begin with two pieces of SNP array data for each marker – the genotype call with alleles labeled A and B and the BAF. Since our method is designed for mixtures with a high proportion of normal cells, we use genotype calls obtained from the mixed sample and assume they are representative of the germline, since small deviations in the BAFs and the LRRs will not alter the genotype call.

The first step in the procedure is to estimate the two sets of haplotypes. Let us call the estimates made using the population-genetics based statistical models $h^{(g)}$ and $h^{(g)'}$. A number of software packages are available that can be used to produce these estimates. The BAF-based haplotypes, which we call $h^{(b)}$ and $h^{(b)'}$, are defined by the allele configurations at the heterozygous markers only. Note that the pair of haplotypes are unordered, so the $h^{(g)}$ and $h^{(g)'}$ labels may be assigned arbitrarily. The BAF-based haplotype estimates are constructed as follows. Let M denote the number of heterozygous markers, and b_m denote the BAF at heterozygous marker m ($m = 1, \dots, M$). We define $\boldsymbol{h}^{(b)}$ by calling each allele $h_m^{(b)}$ independently at each marker:

$$h_m^{(b)} = \begin{cases} B, & b_m > \tilde{b} \\ A, & \text{otherwise,} \end{cases}$$

32

where $\tilde{b}$ is some threshold. When $h_m^{(b)}$ is equal to $\tilde{b}$, the alleles are assigned with equal probability. The haplotype $h^{(b)}$, then, is an estimate of the overrepresented haplotype as indicated by the BAFs. For the optimal interpretation of the phase concordance, we should define the threshold $\tilde{b}$ at each marker by the median BAF for a diploid heterozygous genotype at that marker; since in practice we rarely know or have enough samples to estimate this value with precision, we instead use the median of observed BAFs across all heterozygous loci for the sample. Since we consider only heterozygous sites, $h_m^{(b)\prime}$ can be defined for each marker m by the complementary allele to $h_m^{(b)}$, and is the BAF-based estimate of the underrepresented haplotype.

Once the two haplotype estimates are constructed, we can assess phase concordance using switch consistency. Formally, let x_i be an indicator of consistency between the two sets of 2-site haplotypes defined by $\left(h_i^{(b)}, h_{i+1}^{(b)}\right)$ and $\left(h_i^{(g)}, h_{i+1}^{(g)}\right)$ ($i = 1, \ldots, M-1$), i.e.

$$
x_i = \begin{cases} 1, & h_i^{(b)} = h_i^{(g)}, h_{i+1}^{(b)} = h_{i+1}^{(g)} \quad or \quad h_i^{(b)} \neq h_i^{(g)}, h_{i+1}^{(b)} \neq h_{i+1}^{(g)} \\ 0, & \text{otherwise.} \end{cases}
$$

Note that comparing one haplotype from each set is sufficient, since in each set one haplotype is sufficient information to define the other.

This strategy makes several assumptions. First, by using the same threshold value across markers we are assuming no marker-specific bias in the BAFs toward alleles that tend to cosegregate in the population, which could possibly arise if "B" allele designations were made based on population frequencies, intentionally or otherwise, and there existed some biased intensity for alleles of the same labels. Additionally, we assume that the observed BAF values are accurate (if not precise) estimates of the true proportion of B alleles in the sample of interest, and not majorly distorted by DNA quality issues (see section 2.3).

2.2.2 Hypothesis testing

The result from above is a vector x of indicators of switch consistency. The most straightforward way to use this data is to test for imbalance in a specific region of interest, for example a chromosome arm or gene region. To do this, first identify the consecutive heterozygous markers $(j, ..., k)$ that map to the region of interest. The length and specific markers in the region may vary across individuals, since each will be heterozygous at a unique set of loci. Then,

$$\sum_{i=j}^{k-1} x_i \sim Binom\big((k-j), p\big),$$

where p is the true concordance rate in the region. We can therefore test for evidence of imbalance by performing a one-sided binomial test of $p > 0.5$ against the null hypothesis, $p = 0.5$. Note that, since each element of x corresponds to a pair of heterozygous markers, we use $x_{(j,...,k-1)}$ instead of $x_{(j,...,k)}$ so that we only use values for which both markers in the pair are within the region of interest. The power of this test depends on the number of observed heterozygous markers in the test region and the true phase concordance for the region, which in turn depends on the biological factors including type of imbalance event, the proportion of tumor cells in the sample, the size of the aberration, and technical factors including the amount of noise in the array data and the accuracy of the statistically-estimated haplotypes. In Appendix Figure 1 we present power to detect events across different event types and tumor proportions.

2.2.3 Whole-genome profiling

A second application for these transformed data is for characterizing allelic imbalance status along the genome. Due to the segmental nature of AI events, signal of imbalance in the data will not be distributed uniformly across the genome, but rather in clusters of phase-concordant marker pairs (subsets of **x** where 1s are observed at frequency higher than 0.5). We use a simple discrete-state HMM to model the observed switch consistency and underlying imbalance states.

Structure and Parameters. Let L_i be an indicator for the imbalance level of the interval between

heterozygous loci i and $i + 1$ ($i = 1, ..., M - 1$). For ease of exposition, let us assume 3 states, defined

as follows: 0, no AI; 1, low level of AI; 2, higher level of AI. Each non-zero state may represent a

different copy number in the aberrant cells, or may capture one event type occurring in different

proportions of the sample; that is, they are defined in terms of imbalance level, not explicitly in terms of

underlying mutation characteristics. Then $L_1, ..., L_{M-1}$ form a Markov chain on the 3 underlying

imbalance states. We let α_l ($l = 0,1,2$) denote the emission probability $Pr(x_i = 1 \mid L_i = l)$. The

transition probability matrix is constructed as

	0	1	2
0	$1-\lambda_0$	$\dfrac{\lambda_0}{2}$	$\dfrac{\lambda_0}{2}$
1	λ_1	$1 - \lambda_1$	0
2	λ_1	0	$1 - \lambda_1$

where λ_0 and λ_1 are assumed to be constant across marker intervals. This assumption could be relaxed,

for example to reflect known mutation hotspots. States 1 and 2 cannot directly communicate; the process

must pass through the non-AI state to transition between any AI states. The definitions also imply that

all AI states have identical distributions on their event lengths. Known differences in the sizes or

frequencies of different types of copy number changes would be a potential reason to use different

distributions for different states, but only if the states were explicitly defined in terms of copy number.

In any case, the model is flexible enough to accommodate departures from the assumptions.

Parameter estimation. The emission probability for each state is estimated from the data using an

expectation-maximization (EM) algorithm. At each iteration the parameter α_l is updated to

$$\hat{\alpha}_l = \sum_{i=1}^{M-1} \frac{w_{i,l} x_i}{w_{i,l}},$$

where $w_{i,l}$ is the marginal conditional probability that the process is in state l at marker interval i and is

calculated using standard forward and backward algorithms (56).

The transition probabilities λ_0 and λ_1 may either be fixed or estimated via a stochastic-EM

algorithm that incorporates pseudocounts for a combination of flexibility and stability. The weight

parameter γ indicates how much weight should be given to the 'prior' knowledge relative to the data.

The reason we include pseudocounts in the estimation procedure is because it can otherwise be

unpredictable when the sample does not exhibit strong evidence of imbalance. In this procedure, at each

iteration, we first sample L given the data and the current parameter values, then combine these

stochastic samples with the pseudocounts in a maximization step before re-estimating parameters.

Specifically, the transition probabilities are updated to

$$\lambda_s = \frac{\gamma \lambda_s^i + x_s}{\gamma + n_s}$$

where λ_s^i is the transition probability calculated using the user-specified parameters for the event length

and event prevalence, and n_s and x_s are the number of sampled intervals in state category s, and the

number of transitions out of state category s in the stochastic sampling. State 0 (i.e. the normal state) is

the only state in state category 0; all of the non-normal states make up state category 1.

Profiling. The evidence of the underlying imbalance level is represented by the conditional marginal

probability of each event state at each marker interval, which is calculated using forward and backward

algorithms. One way to interpret these data is to sum the probabilities for the imbalance states at each

marker, and define imbalance regions by groups of consecutive markers for which the probability of

imbalance exceeds some threshold. The probabilities for each state may also be assessed separately.

2.2.4 Identification of the excess haplotype

The third application that we present is identification of the excess haplotype in a region of imbalance,

which we denote by h^*. We assume the region covers a single event, so that h^* is one of the individual's

true haplotypes. To motivate the technique, note that statistical estimation provides an unordered pair of

haplotypes ($h^{(g)}$ and its complement $h^{(g)'}$) with a low but non-zero switch error rate. Thus, h^* is a mosaic of $h^{(g)}$ and $h^{(g)'}$, switching between the two when there has been a phasing error. Also note that in AI regions the haplotype $h^{(b)}$ represents a naive guess at the allele in excess at each marker, although with low accuracy at the tumor proportions we consider. Nevertheless, when there is at least subtle imbalance, $h^{(b)}$ contains information useful for determining which of the two haplotypes $h^{(g)}$ and $h^{(g)'}$ is the source for h^* at each marker. We assume a series of hidden states at M heterozygous loci, denoted by $H_1, \dots, H_M$, form a 2-state Markov chain on *XX replace* $(0,1)$. The transition probabilities are specified by locus-specific switch accuracy estimates for $h^{(g)}$ (from the software used to estimate haplotypes), if available, or an estimate of the average accuracy otherwise. Locus-specific estimates are a way to propagate "prior" information from the statistical phasing to inform distributions of the size and locations of "chunks" of $h^{(g)}$ and $h^{(g)'}$ that likely represent the actual excess chromosome. We let the observed data for our HMM consist of a series of indicators $y_1, \dots, y_M$, where

$$
y_m = \begin{cases} 0 & , h_m^{(b)} = h_m^{(g)} \\ 1 & , h_m^{(b)} = h_m^{(g)'}. \end{cases}
$$

Finally, the emission probabilities are specified as

$$
p(y_m = 1 \mid a, H_m) = a^{I\{y_m = H_m\}}(1-a)^{I\{y_m \neq H_m\}},
$$

where $I_{\{C\}}$ is 1 if C is true and 0 otherwise, and a may be estimated from the data but has a natural relationship with the emission probabilities α specified above. In practice, we simply substitute $\hat{\alpha}_{\hat{s}}$, where $\hat{s}$ is the maximum *a posteriori* estimate of the imbalance state for the interval to the right of the marker (or to the left for the last marker). The final step in our algorithm is to summarize the evidence that a particular allele is in excess by making maximum *a posteriori* probability estimates of H_m,

$(m = 1, \dots, M)$, which yields a best-guess estimate of h^*. hapLOH produces best-guess estimates of the overrepresented haplotype for the entire genome. Of course, these estimates are meaningless in normal regions, since there is no overrepresented haplotype. The user should only consider these estimates in regions for which the profiling HMM produces evidence of imbalance.

2.3 Implementation

The method has been implemented using Perl and Python for UNIX and Mac platforms. Various parameter and options may be specified via command-line arguments. Some of the most important arguments are those controlling the transition probabilities. Since transition probabilities are not a natural quantity for a user to think about, the input that is taken from the user is the expected genomic size of imbalance events and the expected genome-wide prevalence of imbalance (in other words, what proportion of the genome is expected to be affected by allelic imbalance). The transition probabilities are then calculated using these values and the observed heterozygote density for the sample. The user has the option to either use fixed transition probabilities or to estimate them for a sample. When the option is selected to estimate the transition probabilities, in addition to the expected genomic size and expected prevalence, the user also specifies a weight parameter γ, and the three values are used to calculate pseudocounts.

2.4 Unique sources of error and appropriate quality control

The source of signal for hapLOH is slightly different than the source of signal for other methods, which necessitates specific quality control criteria to prevent false positive results.

2.4.1 Cell line samples

DNA samples from tumor cell lines or immortalized non-malignant tissue may acquire mutations *in vitro* that were not present in the primary sample and are not of interest. These cell line artifacts are impossible to distinguish from the mutations of interest, especially since they may preferentially affect

regions that confer a positive advantage to the resulting lineage, just as we would expect to see in

detectable somatic mutations occurring *in vivo*. Because of this severe potential for spurious results,

DNA samples derived from cell lines should be excluded from hapLOH analyses.

2.4.2 Amplified DNA samples

Whole-genome amplification (WGA) is sometimes applied to samples when the DNA quantity is low.

There are multiple methods for WGA, each with its own strengths and drawbacks. Multiple

displacement amplification using Phi-29 polymerase has been widely touted as a more efficient and less

bias-prone WGA method compared to PCR-based methods, especially for DNA that is to be applied to

oligonucleotide arrays (including comparative genomic hybridization (CGH) arrays and SNP arrays). In

the case of application of SNP genotyping arrays and inference of allelic imbalance, an important

characteristic for WGA methods is the level of allele-specific amplification bias. This type of bias (and

the most extreme example of bias, allelic dropout) are especially a concern in the case of very low

amounts of input DNA, since stochastic differences in the early rounds of amplification are magnified in

subsequent rounds (57). Degraded DNA has also been seen to result in high levels of allele-specific

amplification bias and allelic dropout after MDA (58). In a set of 3,838 samples genotyped on the

Illumina 660W array, we observed unusually high overall phase concordance for WGA samples

compared to unamplified samples (Figure 5). The methods descriptions for this study do not specify the

method of WGA. Several sources of DNA (Buffy coat, blood spot) were used for this genotyping, and

we also noticed an association between the DNA types and genomewide average. For the most

conservative strategy, data from samples that have undergone WGA should be excluded, especially if the

samples were amplified from sparse or low-quality input DNA; at the very least, data from WGA

samples should be interpreted with caution.
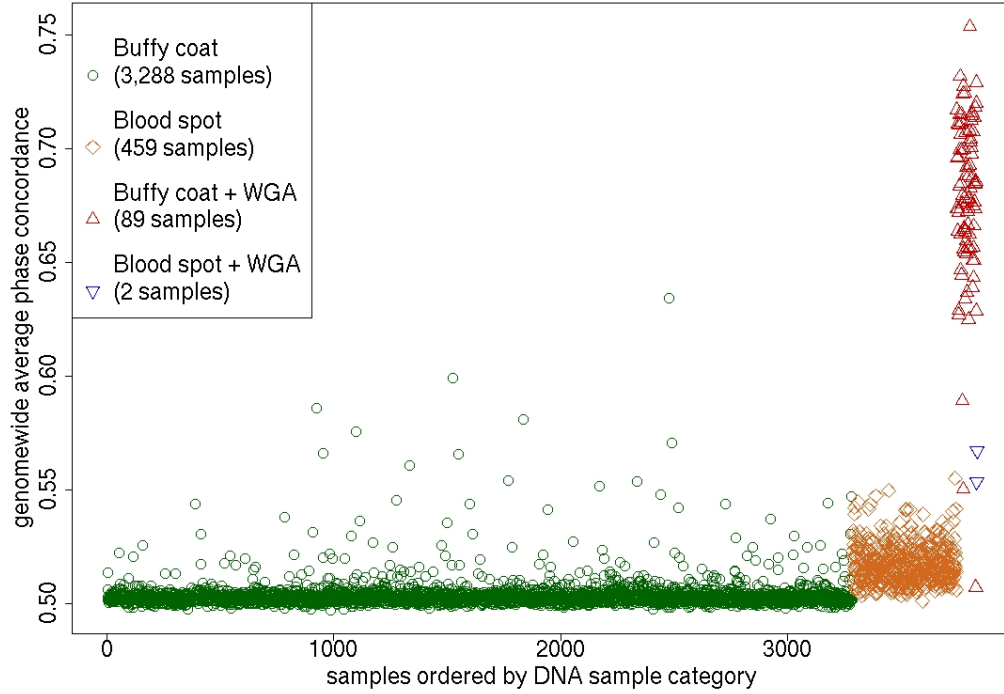
Figure 5. Effect of DNA type and WGA.   The WGA samples that were derived from Buffy coat samples (red triangles) generally had higher genomewide phase concordance than other sample types. Amongst the non-WGA samples, the blood spot samples had slightly higher genomewide phase concordance than the Buffy coat samples.

2.4.3 Inter-sample contamination

Low levels of inter-sample contamination do not impact genotype calls, but create minor shifts in observed allele frequencies at markers for which the intended target individual and the contaminating individual have different genotypes. These small perturbations, in fact, have been used to detect low levels of sample contamination in SNP array and sequence data (59, 60). Individuals may share identical haplotypes for some stretches and may exhibit substantial haplotype similarity for long stretches. These perturbations, then, may create a false signal of imbalance in our method. We have found that samples that were estimated to have >1% inter-sample contamination using ContaminationDetection (59) tended to have higher overall phase concordance than samples that had estimated contamination < 1% . We

40

suggest excluding results from samples that have an estimated α_0 value below 0.52, or a similar threshold.

2.4.4 Genotyping errors

The method utilizes genotype calls to establish the set of informative sites and in the construction of the truth-surrogate haplotypes. Any type of genotyping error potentially may impact the phasing accuracy, although phasing algorithms are generally robust to sporadic errors and any potential impact will be local. Homozygous-to-heterozygous miscalls are a potential source of inflation of the phase concordance, although sporadic errors are likely to have any measurable impact only if they occur at a very high rate or are aggregated by location due to chance. Heterozygous-to-homozygous errors will decrease the informative marker count and therefore potentially result in a loss of power, but again the effect of sporadic errors will be minimal and will not bias results. Genotype calls should be filtered based on quality scores or other appropriate filters to minimize miscalls. It is worth noting that in a region with a very high level of imbalance, the BAF bands may deviate so strongly from the expected heterozygote and homozygote values that the genotype calling software will produce a 'no-call' (will not assign a genotype) or will produce a homozygote call at a germline heterozygous site. The power of the method is severely or completely reduced in these regions, but could be restored by using BAFs from the sample of interest with legitimate germline genotypes from a matched normal sample.

2.4.5 Log R ratio waviness

Using a higher or lower DNA concentration than specified by the genotyping protocol has been shown to cause 'genomic waves', or visible oscillation of the LRRs about the expected value (61). The wave pattern is related to the GC content of the target genomic sequence, so wave periods can vary along the genome but will cover multiple markers. So, like copy number variation, genomic waves create increases and decreases in the LRR, although they generally have a sinusoidal pattern whereas somatic or inherited CNVs create a step-function-like pattern. The effect of GC content will be the same for

41

chromosomes carrying either allele, and there is no evidence that BAFs are affected.   Since hapLOH

currently only considers BAFs, not LRRs, genomic waves should not interfere with any of the hapLOH

functions such as specific region testing or whole-genome characterization. However, once a region has

been identified, the LRRs become very useful for interpreting the mutation, and at that point genomic

waves can severely impair correct interpretation. The LRR data should be assessed for genomic waves.

The software program PennCNV includes a function to quantify the severity of GC-associated genomic

waves using a 'waviness factor' (61) which is helpful, and even visual inspection can identify samples

with low quality data.

CHAPTER 3
Proof of principle demonstrations

Portions of this chapter are based upon "Haplotype-based profiling of subtle allelic imbalance with SNP arrays", Selina Vattathil and Paul Scheet (2013), *Genome Research* 23(1): 152-158.

3.1 Application to computationally diluted tumor samples

3.1.1 Introduction

The goal of the experiments presented in this section was to characterize the performance of the method under various scenarios and especially to test performance in samples with low tumor fractions, since this is the scenario in which existing SNP array methods fare poorly and where we expect the addition of haplotype information to be most useful. The first proof of principle test that we conducted was to apply hapLOH to samples with known somatic allelic imbalance events, for which we could define true positive and true negative regions and therefore assess sensitivity and specificity. We did not find any existing test datasets with somatic mutations in the low frequencies that we were interested in testing. We decided to use an existing dataset to generate what we refer to as 'computational dilutions' (details in next section). Compared to simulating BAFs completely from scratch using theoretically expected shifts, the benefit of the computational mixing approach is that we preserve not only the actual level of random noise in the data but also any existing correlation that exists among BAFs at individual markers. Possible sources of correlation include intensity variation that is associated with GC content or with dye bias and variation in the effectiveness of intensity calibration across markers. Since hapLOH uses BAFs only, not LRRs, we did not simulate LRRs. We chose to compare the hapLOH results to the results from applying BAFsegmentation (44), which also considers BAFs but not LRRs.

3.1.2 Materials and methods

BAF and genotype simulation

To test their own detection algorithm (BAFsegmentation), Staaf and colleagues (44) created a set of tumor-normal mixture samples by mixing DNA from a breast cancer cell line (ATCC CRL-2324) and the matched lymphoblastoid (normal) cell line (ATCC CRL-2325) in nine proportions (10%, 14%, 21%, 23%, 30%, 34%, 45%, 47%, and 50%). They also had samples of the pure tumor and pure normal DNA, for a total of eleven samples. They genotyped these samples on the Illumina HumanCNV370-Duov1 BeadChip array. We downloaded the genotypes, BAFs, and LRRs for these assays from GEO (accession GSE11976).

The lowest dilution created in this series was 10% tumor, and our goal was to test hapLOH at lower tumor proportions. We therefore constructed "computational dilutions" using the data from the pure cancer and pure normal samples as follows. First, we visually inspected chromosome plots of the BAFs and LRRs from the pure tumor sample and identified regions that exhibited complete CNLOH or hemizygous deletion. We did not include any regions that looked like they harbored subclonal mutations (cell line artifacts). We also visually inspected and applied BAFsegmentation to plots of the pure normal sample data and excluded 7 regions from analysis due to their exhibiting cell line artifacts that would trigger false positive calls (Appendix Table 1).

From the intersection of this set of complete LOH regions in the tumor sample and normal regions in the normal sample, we chose 15 segments to simulate as hemizygous deletions and 10 segments to simulate as CNLOH. Segments ranged in size from 2.39 Mb to 95 Mb and covered 25% of the genome (complete details for each event are presented in Appendix Table 2). To create each simulated sample, we started with the pure normal sample and replaced the BAFs in the 25 regions with weighted averages of the BAFs from the pure tumor and pure normal samples, with the weights determined by the target tumor proportion for the sample and the type of event, i.e. the mixture BAF at marker i, denoted by b_i^m, is given by

$$b_i^m = \frac{2(1-f)(b_i^n) + x_t(f)(b_i^t)}{2(1-f) + x_t(f)},$$

where f is the fraction of mutant cells, x_t is the total copy number for the simulated event (i.e. 2 for

CNLOH and 1 for hemizygous deletion), and b_i^n and b_i^t are the observed BAFs in the normal and tumor samples, respectively.

We created simulations at each of the proportions targeted in the lab dilution series, and also for proportions from 1% to 10%. For tumor proportions below 10%, we used the genotypes from the normal sample since imbalance levels this small have little to no impact on the genotype calls (in fact, even the genotype calls from the 10% lab dilution are nearly identical to those from the pure normal sample). For each of the higher proportions, we assembled the set of genotypes for the simulated sample by replacing the normal genotypes in the AI regions with the genotypes from the lab sample with that tumor proportion.

To generate the test data for the detection of the presence of tumor genome, we averaged the normal and tumor BAFs as described above, with one change. Above, we averaged only BAFs at markers within the 25 event regions that we chose to simulate. Here, since we wanted to test the ability to detect imbalance at the genome-wide level (without having to test a specific region), we instead averaged BAFs at all markers, with weights determined by assuming hemizygosity for the odd chromosomes and CNLOH for the even chromosomes. We thought this reasonable based on the observation that these two event types constituted the majority of events and occurred in roughly equal proportions. Empirical comparisons with the lab-based dilution series indicate this procedure is well calibrated (Appendix Figure 2).

To test hapLOH estimation of the overrepresented haplotype, we inferred the true overrepresented haplotype using the pure tumor sample. Since the pure tumor sample exhibits true imbalance at a high (100%) frequency, the BAFs at germline heterozygous sites have moved to either 0 or 1, and unequivocally indicate the overrepresented allele at each marker. For each marker within the simulated event regions of our curated data set, we called a "B" if the pure tumor BAF was greater than 0.8 and an "A" if the BAF was less than 0.2.

hapLOH settings

We performed statistical phasing with fastPHASE (62) using default settings. We use the 120 haplotypes from the HapMap CEU panel as our reference panel. We first determined the intersection set between the markers in the reference panel and those in the sample data. We used the data from these 319,000 markers to fit the fastPHASE model using the reference samples only, and then applied the fitted model to the sample data to estimate haplotypes.

We applied the hapLOH profiling HMM using an expected event length of 660 informative markers, or about 20 Mb, and assuming expected genomewide event prevalence of 10%. The transition probability matrix was estimated for each sample using $\gamma=1,000$. For the hapLOH haplotype estimation HMM, we set the transition probabilities for each sample using the local switch probability estimates output by fastPHASE for that sample.

BAFsegmentation settings

We applied BAFsegmentation (44) to all samples with the mBAF threshold (a parameter of the method) set to 0.526. Setting the mBAF to the default value of 0.56 resulted in very low sensitivity; the value of 0.526 corresponds roughly to the value that should allow detection of deletions in 10% of the cells and provided better sensitivity. We found that using lower thresholds resulted in highly unlocalized calls, usually covering entire chromosomes (data not shown).

3.1.3 Results

In this section I will present results for the three hapLOH functions described in sections 2.2.2-2.2.4 – hypothesis testing, whole-genome profiling, and identification of the excess haplotype.

Hypothesis testing. Application of the method to the 10% tumor sample data (entire dataset, which included 103,556 heterozygous sites) results in a phase concordance of 0.65, which differs significantly from the expected null concordance of 0.5 (p-value=10^{-2140}).

We also extrapolated the results from the observed 370K data to predict the results that would be observed from application of 1M and 5M SNP chips with 30% and 25% heterozygous markers, respectively. We present these results in Table 2. We observe that in the pure normal sample the phase

concordance is 0.5005, which is slightly higher than the expected null concordance rate of 0.5. This

deviation could be due to somatic non-tumor associated events in the individual, mutation during the

growth of the cell line, or inter-sample contamination during sample preparation. To obtain the correct

type I error rate for detecting tumor cells, we show power results using both the expected normal phase

concordance rate of 0.5 and the observed 'normal' rate of 0.5005. Our results indicate potential to detect

aberrant cells for this particular breast cancer genome at concentrations on the order of 2 or 3 in a

thousand using a hypothetical 5M SNP array (power > 50%). It appears that below these levels the

signal from imbalance is overcome by the stochastic variation in the BAF data, since we start to observe

slightly erratic measures (phase concordance at .05% tumor is higher than at .10% tumor).

tumor content	phase concordance	Power with type I error rate = .05		
		Null phase concordance 0.5 (and 0.5005)		
		370K array	1M SNPs	5M SNPs
0% (no tumor)	0.5005	.09 (.05)	.13 (.05)	.28 (.04)
0.05%	0.5010	.16 (.09)	.30 (.14)	.74 (.31)
0.10%	0.5007	.12 (.07)	.20 (.08)	.51 (.14)
0.25%	0.5012	.20 (.12)	.39 (.20)	.87 (.51)
0.50%	0.5014	.22 (.14)	.44 (.24)	.92 (.61)
0.75%	0.5025	.49 (.36)	.87 (.72)	1 (1)
1.00%	0.5035	.73 (.61)	.99 (.95)	1 (1)
2.00%	0.5071	1 (1)	1 (1)	1 (1)

Table 2. Power to detect presence of CRL-2324 cells. We compare the expected power across tumor

proportions (first column) and genotyping array densities. The primary power results were calculated

assuming a null concordance of 0.5; in parentheses we give the power assuming a null concordance of

0.5005. For the 370K results we used the observed phase concordance rates from the computational

dilution data and the heterozygous marker count from the pure normal sample. For the 1M and 5M array

results, we used the same concordance rates and assumed 30% and 25% of markers would be

heterozygous, respectively.

Whole-genome profiling

At each marker, we summarize the pointwise evidence of imbalance by the conditional probability of being in an AI state, given the data and parameters (conditional marginal probabilities, or "posterior probabilities"). The results for a few representative samples (0%, 4%, 7%, 10%, and 14% tumor) are presented in Figure 6. The posterior probabilities for each of the two AI states, which we can call S1 and S2, are considered separately. As we would expect, in the 0% tumor sample the posterior probabilities are at 0 for the entire genome. At 4% tumor, signal is obvious at the CNLOH events, which create stronger imbalance than hemizygous deletions at the same proportion. There is some increased posterior probability at some of the deletion regions, but there is also some background increased probability genomewide. Since only one level of imbalance is obvious (imbalance from the CNLOH events, not the deletion events), having two event states is superfluous and the posterior probability at CNLOH events is split evenly between S1 and S2. In the 7% tumor sample, hapLOH detects hemizygous deletions in addition to the CNLOH events and discriminates between the two event types, giving the hemizygous deletions high probability for S1 and the CNLOH events high probability for S2. BAFsegmentation begins to identify the stronger (CNLOH) signals at this tumor proportion. All events in the 14% tumor sample are picked up by both methods except for one small true event (1.52 Mb on chromosome 21) for which hapLOH signal is very low. The background noise is also almost completely mitigated, since the emission probabilities for the event states will be very different from that for the normal state. At higher concentrations of tumor ($\geq$34%), the BAFs for both deletions and CNLOH regions diverge so strongly from the expected normal value that the phase concordance reaches the upper limit determined by the statistical phasing accuracy, for example about 93% in this analysis. At these proportions hapLOH no longer distinguishes the two types of simulated events.
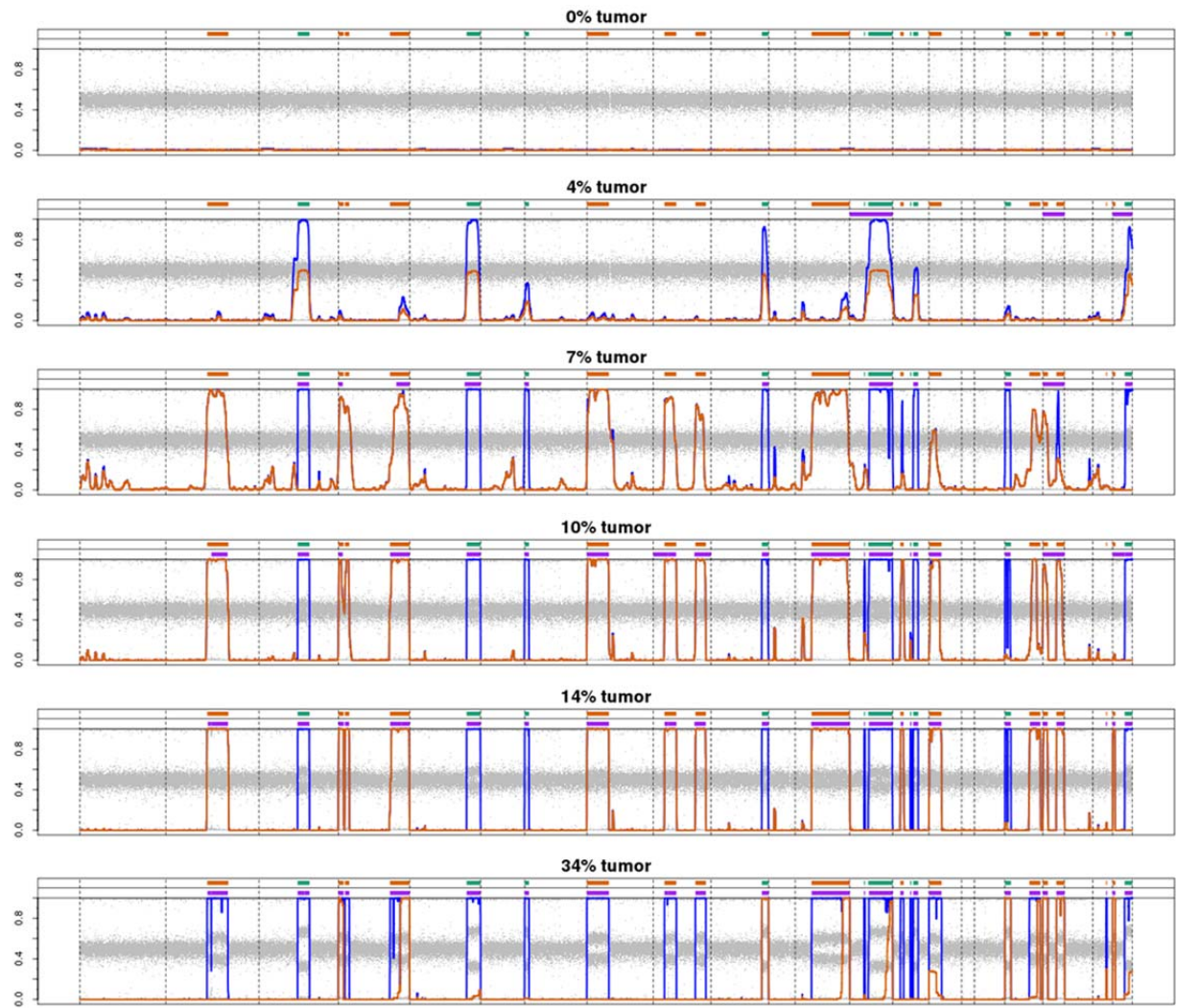
Figure 6. Local posterior probabilities of allelic imbalance at various dilutions. Vertical axes range from 0 to 1 for both the BAFs (grey points) and posterior probabilities (orange lines for probability of deletion, blue lines for summed probability of deletion or CNLOH). Horizontal lines at the top of each plot show the locations of simulated deletions (orange) and CNLOH (green). Below these, purple bars show the regions identified by BAFsegmentation to contain AI.

Identification of the excess haplotype

We evaluated the overrepresented haplotype constructed using our HMM by comparing it to the 'true' overrepresented haplotype (determined by applying a threshold to the pure tumor data) at a range of tumor proportions (Figure 7). Accuracy was calculated as the proportion of correct calls at sites where

49

we could confidently call LOH using the BAFs from the pure tumor sample – about 18,000 markers for deletions and 8,000 markers for CNLOH. To provide a baseline, we also constructed a naïve estimate of the overrepresented haplotype by simply taking the allele with the highest frequency at each site, and calculated its accuracy. The method implemented in hapLOH improves accuracy over the naive estimate at the lower tumor proportions. For example, at 5% tumor content and within our simulated events, hapLOH achieved accuracies of 80% and 89% for deletions and CNLOH, respectively, compared to 64% and 75% using the naive method. As expected, the accuracy is higher in CNLOH regions than in deletion regions, since the BAFs are more strongly imbalanced. As tumor proportion increases, the BAFs contain increasingly perfect information about the excess haplotype and incorporating the statistical haplotype estimates actually gives lower accuracy than the naïve BAF-only method.
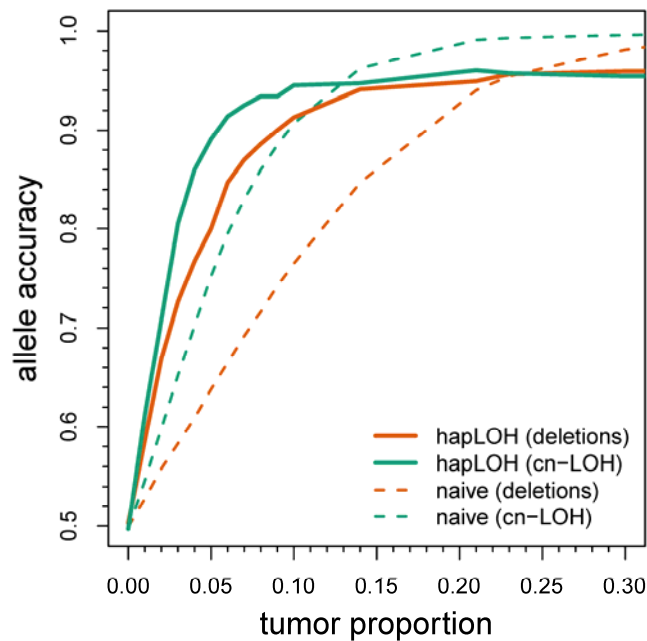

Figure 7. Identification of the overrepresented haplotype. Solid lines indicate the accuracy of hapLOH's haplotype calls at deletions (orange) and CNLOH (green) in the curated dataset at various proportions. Dotted lines indicate accuracy of the naive BAF-based calls.

3.1.4 Robustness tests

Robustness to parameter specification

We have implemented hapLOH with two options for setting the transition probability matrix (TPM) for the profiling HMM. For both, the user must specify two values, the expectation on the length of mutations and the expectation on the fraction of the genome that is imbalanced (which we refer to as prevalence). In the 'fixed' option, the TPM is defined to satisfy the user-specified values and is left constant while the EM to estimate the emission probabilities is performed. In the 'estimate' option, the initial TPM is defined using the user-specified values and then is iteratively updated via a stochastic EM. The true values for the mean event length and prevalence are unknown before analyzing a sample, and in any case it not necessarily true that specifying the correct values will result in the most desirable set of results from the profiling HMM. In order to understand how different parameter values and different TPM specification strategies impact the profiling results, we applied hapLOH to the curated dataset under a range of conditions (Table 3).

event length (Mb)	prevalence	option
10	0.10	estimate
15	0.10	estimate
25	0.10	estimate
10	0.01	fixed
10	0.05	fixed
10	0.20	fixed
10	0.25	fixed
10	0.50	fixed

Table 3. Settings for testing of robustness to parameter specification. For all of the runs using TPM estimation, the gamma value was set to 1,000.

The posterior probability curves tend to be noisier with smaller specified mean event lengths, especially at the lower tumor proportions (data not shown). Also, using fixed TPM (versus estimating the TPM using the data) can result in considerable and constant background noise, again most prominently

51

observable at low tumor proportions (presumable because the data is less informative and the parameters have a poorer fit to the data). However, perhaps most importantly, ROC curves comparing these results (Figure 8) emphasize that the differences between event calls using different parameter sets is almost non-existent at most threshold levels. Based on these observations, a practical strategy may be to choose parameters based on the targeted event length and considering the expected prevalence, but to err on the side of lower prevalence and larger events.
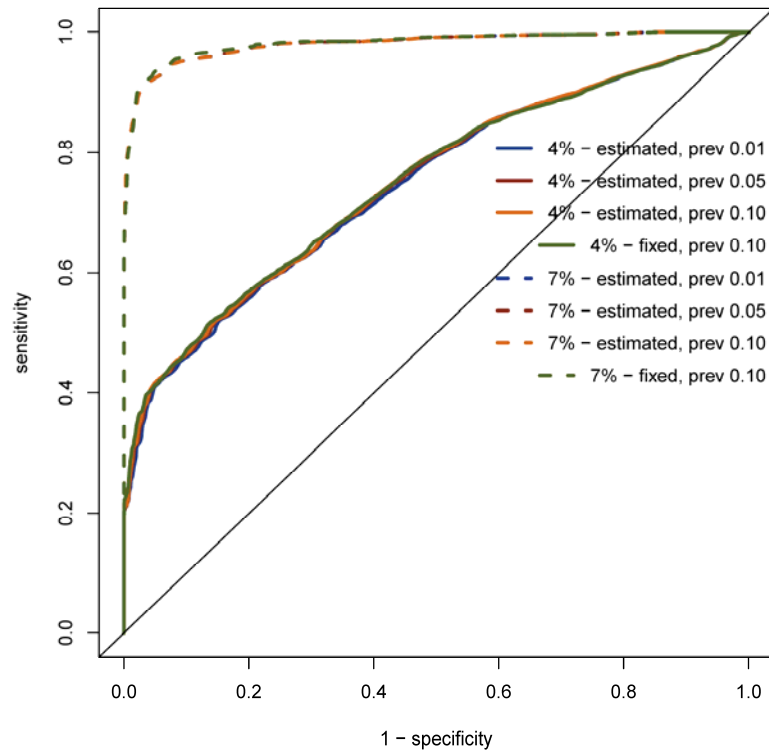
Figure 8. Robustness to prevalence specification and to using fixed or estimated TPM. The ROC curves within a sample are actually or nearly overlaid, indicating that parameter choice does not have a strong impact for these samples.

Robustness to statistical phasing errors

The accuracy of statistical phase estimates plays a role in the sensitivity of the hypothesis test and of the profiling HMM. We expected the impact to be small at the tumor proportions that are our target, as

state-of-the-art statistical phasing packages have accuracies upwards of 90% and the majority of the

noise that will interfere with signal will be contributed by the BAFs. The fastPHASE statistical

haplotype estimates for this dataset have accuracy rates of about 93% (as estimated by fastPHASE). To

quantify the impact of phasing errors on sensitivity of profiling, we randomly introduced switch errors at

rates of 0.05, 0.10, and 0.15 per interval into the haplotype estimates of the curated samples, and then

applied hapLOH using the same parameters used for the results in section 3.1. We then calculated ROC

curves for each set of results. The ROC curves for the pure normal sample and the 4%, 7%, and 10%

tumor samples are plotted in Figure 9. The signal from the BAFs in the 10% tumor sample appears to be

strong enough that even severe inaccuracy in the phase estimates has minor impact. In the 4% sample,

for which hapLOH picked up the CNLOH events but had less signal for the more weakly imbalanced

deletions, it appears that phasing errors have a small effect but surprising have less impact than in the 7%

tumor sample, for which hapLOH had strong signal at almost all simulated events. To interpret this

observation, recall that these results are from considering average sensitivity over the whole genome. By

comparing the posterior probabilities across added switch rates for the 4% and 7% tumor samples

(Figure 10), we can begin to understand this observation. First, many more true positive events had

strong signal in the 7% sample compared to the 4% sample, so it had more potential to lose sensitivity.

Second, the signal for the CNLOH events largely remains intact across switch rates in the 7% sample.

The signal for the hemizygous deletions, however, does decrease. These events had less signal to begin

with in the 4% samples. In summary, the impact of phasing accuracy will depend not only on the

particular accuracy level but also on the specific level of allelic imbalance in a given event region and on

the total set of mutations (since that will influence the final parameter estimates). We note that errors at

an additional 5% of intervals, which is the lowest rate we simulated here, will give these estimates an

error rate well below that rates that are commonly produced from statistical phasing algorithms. So,

these results indicate that standard statistical phasing protocols are appropriate for the purposes of

estimating haplotypes for hapLOH, and it is not necessary to allocate extra resources to improve phasing

accuracy because small differences in accuracy have little impact on the results.
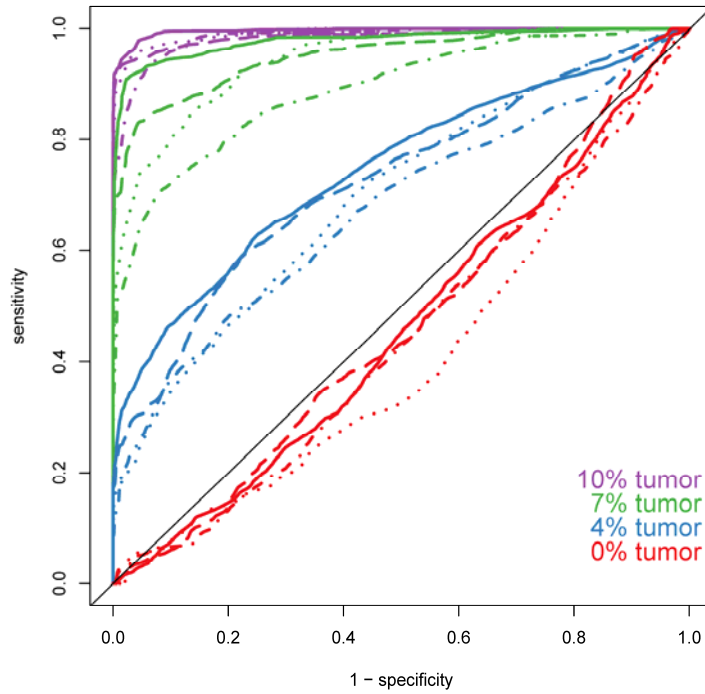
Figure 9. ROC curve for samples with added switch errors. Results are shown for runs using the original data (solid lines), and for runs using the fastPHASE estimates with switches introduced independently at each marker interval at rates of .05 (dashed lines), .10 (dotted lines), and .15 (dashed-dotted lines).
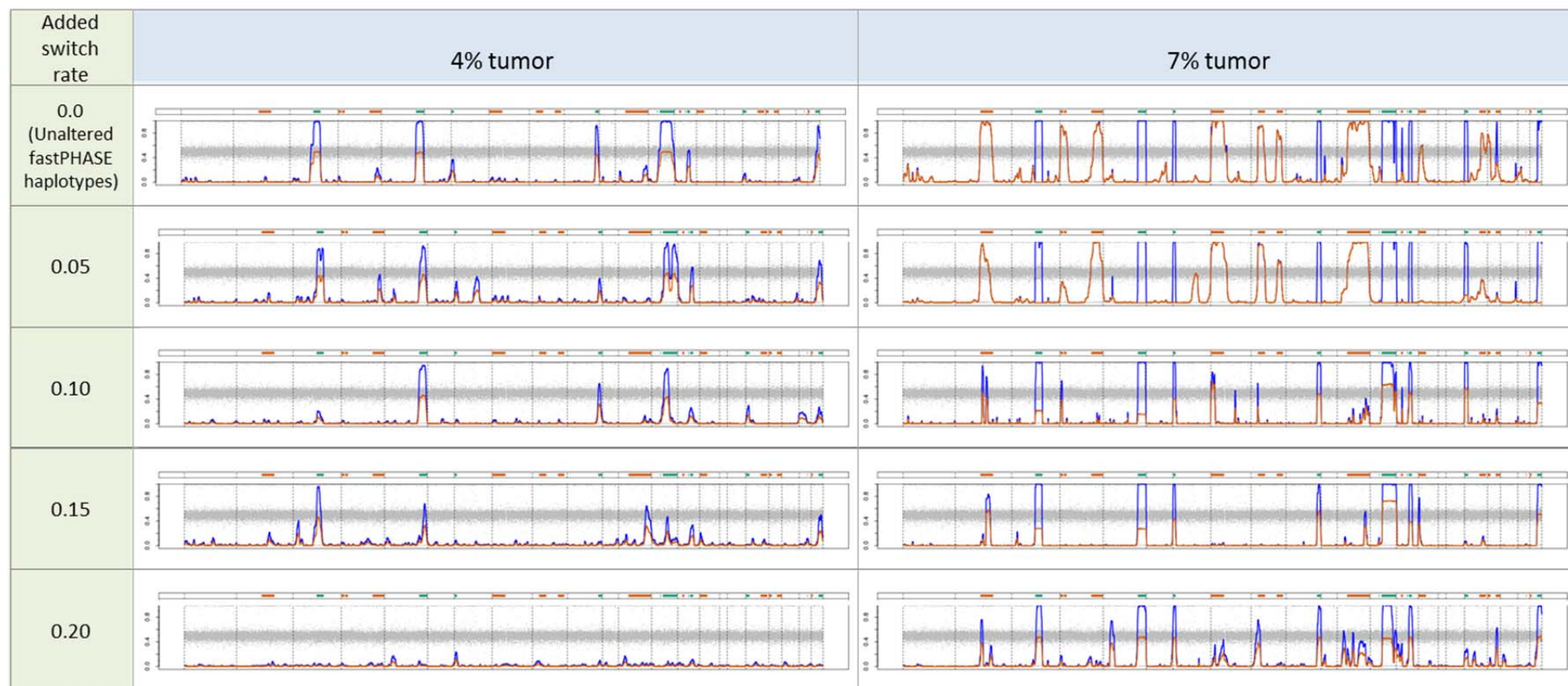
Figure 10. Posterior probability curves demonstrating effect of decreased phasing accuracy. Vertical axes range from 0 to 1 for both the BAFs (grey points) and posterior probabilities (orange lines for probability of deletion, blue lines for summed probability of deletion or CNLOH). Horizontal lines at the top of each plot show the locations of simulated deletions (orange) and CNLOH (green).

3.1.5 Discussion

The event states in hapLOH do not explicitly indicate event type, although they can differentiate between events with different levels of imbalance. In this set of results, since we knew precisely the complement of event types that existed in the samples (that is, hemizygous deletion and CNLOH) and we knew that each event existed in the same proportion, it was reasonable to interpret the two event states as corresponding to different event types. The results from this dataset indicate that the contrast between the multiple levels of imbalance will influence how well the profiling HMM will recognize them as being generated from different underlying states. In other studies, evaluating the log R ratios in event regions detected by hapLOH may be helpful for determining event type (see Chapter 4 for an algorithm to interpret event-specific LRRs), and given the event type it may be possible to infer the mutant cell fraction. Other methods do this, but miscalibration of the BAFs and LRRs can cause biased interpretation and intra-tumor heterogeneity interferes with the ability to borrow strength across events.

We use a time-homogeneous Markov chain to describe the underlying imbalance states. This puts a geometric distribution on event length, with event length measured by the number of informative markers. This distribution allows events from a wide range of lengths to be discovered, while naturally penalizing very short regions with slightly increased phase concordance that may occur due to chance. However, true small events with higher levels of imbalance may also go undetected. For example, a 1.52 Mb event in the simulated samples was not picked up by hapLOH even in the 14% tumor sample. Very large events may be truncated or split. When the transition probabilities are estimated from the data, the probabilities for any given event are affected by the length and level of imbalance for each of the other events in the sample, since they will all influence the estimation of the HMM parameters. Other segmentation procedures that do not have these distributional assumptions may do a better job at picking up small events.

3.2 Application to Affymetrix 6.0 data

3.2.1 Introduction

The goal of the work described in this section was to demonstrate that hapLOH could be applied to Affymetrix data, for which BAF is not a native data type. To do this, we again looked for publicly available data with known true positive regions so that we could assess our hapLOH results in terms of sensitivity and specificity. After searching the list of datasets available in GEO with Affymetrix 6.0 data (the state-of-the-art genomewide chip from Affymetrix), we chose the dataset presented in a 2010 publication by Barresi *et al.* (63). The focus of the original study was to identify tumor-associated copy-number aberrations or CNLOH to assess the submicroscopic instability in normal-karyotype acute myeloid leukemia (NK-AML) genomes. NK-AML is defined as 'karyotypically normal' on the basis of cytogenetic tests, but is known to be a genetically heterogeneous subgroup and accurate prognostication is difficult. The dataset included bone marrow DNA taken at diagnosis (i.e. tumor DNA) from 19 individuals and matched normal DNA for 11 of the individuals (obtained by sampling bone marrow at remission, defined as <5% blasts). Normal DNA was not available for 8 of the patients. A particular emphasis of the study was using the normal samples to distinguish inherited from somatic variants. Inherited variants were identified by comparing calls in the normal samples to the Database of Genomic Variants (64) and by comparing the calls in the tumor and matched normal samples. They also used empirically-defined minimum size thresholds to control their false positive rate by using discordant calls between replicate samples to identify potential false positives and choose a minimum size that allowed a 5% false positive rate. Thresholds of 50 Kb and 80 Kb were employed for losses and gains, respectively.

In their investigation, Barresi *et al.* found an average of 50 CNLOH events larger than 1 Mb per tumor sample. These included 3 CNLOH events larger than 10 Mb that are clearly visible in the BAF plots and were not present in the remission sample, and two large (22 Mb and 16 Mb) CNLOH events present in both the tumor and normal samples for the respective patient that were deemed not to be tumor-associated. All CNLOH events smaller than 10Mb were present in both tumor and matched

normal samples and were also deemed not to be tumor-associated. They also discovered on the order of 10 deletion events and 10 gain events per sample, with a median of one tumor-associated deletion and one tumor-associated gain per sample and only one large copy number change (11 Mb deletion). The authors interpreted these results as support for the relative genomic stability of NK-AML cells, while highlighting the potential relevance of large CNLOH events and small deletions in NK-AML. We sought to investigate whether identification of low-frequency chromosomal alterations would help to provide additional characterization of genomic stability in NK-AML cells.

3.2.2 Materials and Methods

We downloaded Affymetrix 6.0 CEL files for 19 normal-karyotype acute myeloid leukemia (NK-AML) samples (>90% blasts), and matched remission samples (<5% blasts) for 11 of them (GEO accession GSE21780). The samples are described in detail in (63). BAFs and LRRs were extracted from the CEL files using PennCNV (with 77 additional CEL files from the HapMap2 CEU data to help train the algorithm), and genotypes were called using Birdseed v2 (65). We applied hapLOH to each sample, setting transition parameters for a mean event length of 2.5 Mb and a 10% genomewide prevalence of AI. To call discrete events, we summed at each interval, the probabilities of being in an imbalance state and defined an event as any stretch of consecutive markers for which the probability exceeded the threshold value of 0.5 (i.e. the evidence favored imbalance over imbalance).

3.2.3 Results

We identified 46 imbalanced regions in 11 samples, with a median event size of 1.55 Mb. Appendix Table 3 presents details for each called event. In our analysis, 4 out of 11 paired sets harbored AI events in the diagnosis sample only, including two events larger than 10 Mb in two distinct samples. I will focus on these large events. One of these covered 11 Mb on chromosome 3 in the diagnosis sample of Patient 9. The LRR indicates that it is a hemizygous deletion. This loss was also reported in the original publication. The other large event covered 60 Mb in the terminal region of 11q in the

58

diagnosis sample of Patient 61 (Figure 11). Inspection of the LRRs in this region suggests CNLOH. This event includes 5,091 heterozygous genotypes (16,828 total markers) and had a phase concordance of 0.86. This event was not reported in the original publication. Interestingly, a small (88 Kb) event in the remission sample of the same patient overlaps the large event in the diagnosis sample, but the observed BAF and phase concordance indicate a much less prevalent event.
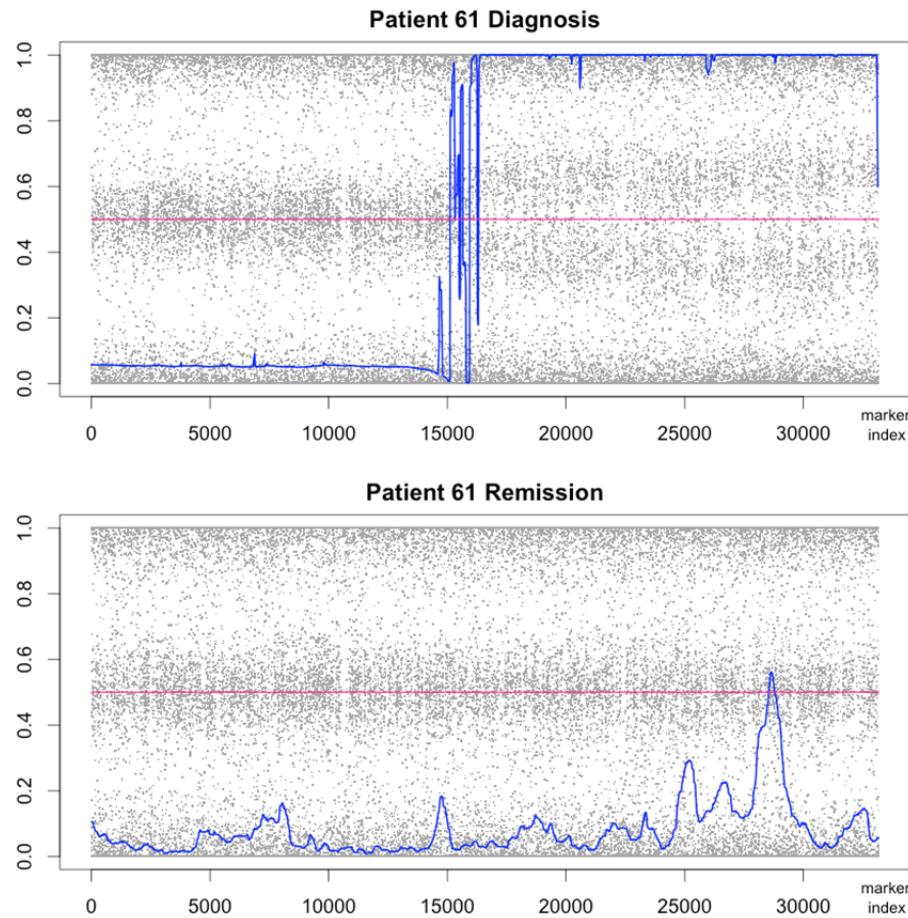

Figure 11. Large partial CNLOH event. BAFs (grey points) and posterior probability of AI (blue curve) for chromosome 11. The pink line at 0.5 indicates the threshold posterior probability that was used to call events.

3.2.4 Discussion

We did not call the 3 large CNLOH events that appeared in the Barresi *et al.* analysis, and called many fewer smaller events than they did. There are several possible explanations for the discrepancy in the number of event calls in the hapLOH analysis and the previous study. First, many of the calls made by Barresi *et al.* are inherited variation rather than somatic variation. In fact, all of the CNLOH events in the 1-10 Mb range occurred in both the diagnosis and remission sample when paired samples were available, so they were all interpreted as inherited variation. hapLOH does not detect inherited losses and CNLOH because they contain no informative (i.e. heterozygous) genotypes. Somatic mutations carried by a high proportion of the sample would also be missed by hapLOH for the same reason. It is important to note that our method is intended to be complementary to methods that identify events occurring in high proportion.

We called one very large event, a CNLOH of 11q in Patient 61, that was not reported in the initial analysis of this data. Importantly, chromosome 11 aberrations are frequently observed in hematological cancers, and 11q CNLOH specifically is known to be associated with disease progression and poor prognosis (66). The change in frequency of an 11q CNLOH-harboring clone during AML development was observed directly by a study (67) that used Affymetrix SNP 6.0 genotyping at multiple timepoints during the progression of a patient through early disease (refractory anemia with excess blasts [RAEB-2]), remission, and late disease (acute myeloid leukemia). They observed a subclone of cells harboring 11q CNLOH soon after RAEB-2 diagnosis, apparent disappearance of the 11q CNLOH clone during remission, and complete or near-complete expansion of the clone during late disease.

CNLOH at 11q is associated with mutations in the gene Casitas B-lineage lymphoma (CBL) (68). The Cbl family of proteins (Cbl, Cbl-b, and Cbl-c) are E3 ubiquitin ligases that negatively cell signaling through protein tyrosine kinase pathways (69). Loss of *CBL* function has been associated with hyperproliferation and increased potency of hematopoietic stem cells in cell counting experiments and competitive repopulation experiments comparing double-knockout and wild-type mice (70). In humans,

CBL mutations are usually seen with CNLOH, that is the initial loss of function CBL mutation (usually either a missense point mutation or deletion) is followed by CNLOH that results in loss of the remaining wild-type allele (69). The interpretation of this is that a single mutant allele is insufficient to promote malignancy, and the CNLOH event is an oncogenic driver. The identification of a missense mutation or small deletion in the CBL gene in Patient 61 would provide strong evidence that CBL is the oncogenic driver mutation in this patient.

AML patients are currently subcategorized partly on the basis of the genetic mutations that are observed in bone marrow DNA samples, and a specific translocation involving a gene on 11q23 is on the list of mutations used to make the categorization, but the criteria are very specific (since the guidelines are based on very well established evidence only) and the 11q CNLOH that we observed the Patient 61 sample would not be sufficient to change it from the 'AML not otherwise specified' category into which NK-AML patients are placed. Nevertheless, the identification of low-prevalence 11q CNLOH, a known common oncogenic driver mutation, is valuable information. The 11q CNLOH clone may expand, based on observations from previous cases. Since the mechanism of action of Cbl is known to some degree (that is, it is known that wild-type Cbl attenuates protein tyrosine kinase signaling and most Cbl mutations involve loss of E3 function), the mutation indicates that the patient may respond to existing or potential therapies that are effective in cases who have other mutations that also target the same pathways (69).

CHAPTER 4
Analysis of genomic mosaicism in non-cancerous tissue

4.1 Introduction

Several recent studies used SNP array data to investigate chromosomal abnormalities in blood and buccal samples from phenotypically normal individuals or individuals who do not have a phenotype known to be associated with mosaicism of the sampled tissue (25-27). These and other studies (21, 71) have highlighted the potential offered by high-resolution technologies such as SNP arrays and sequencing for better understanding of the role of somatic mutation to complex disease and for the spectrum of somatic mutation in normal tissues (72, 73). All three studies explicitly noted the limited sensitivity to detect low-frequency somatic mutations, with the Jacobs *et al.* study reporting events with a minimum frequency of 7% and the Laurie *et al.* study estimates that a minimum of 5% of mutant cells is required for detection. The initial motivation for hapLOH was discovery of tumor mutations in tumor-normal mixture samples, but the method is not limited to this context. Given the limitations cited by these other SNP array analyses, we expected that hapLOH could augment the existing observations of somatic clonal mutations in normal tissues, especially for mutations at low frequency. To explore this, we applied hapLOH to the dataset previously analyzed by Laurie *et al.* (25).

4.2 Materials and Methods

4.2.1 Samples and data preparation

All of the data that we analyzed were collected for GWA studies conducted by the Gene Environment Association Studies (GENEVA) Consortium. These were case-control studies investigating the role of genetic variation and gene-environment interaction in a wide range of disease phenotypes, including both cancer and non-cancer phenotypes. The DNA samples were collected from blood or buccal cells, or

from blood-derived cell lines. All samples were genotyped using Illumina genotyping arrays (specific array varied across studies, Appendix Table 4). We analyzed both case samples and control samples.

Genotypes, B allele frequencies, and Log R Ratios were downloaded from dbGaP (study accession numbers, Appendix Table 4). A subset of the samples from the Prostate study did not have BAFs and LRRs in dbGaP; for these we used the 'BAFfromGenotypes' function (specifying the 'by.study' option) from the R package GWASTools (74) to calculate BAFs and LRRs from the θ and R data.

Each study was processed independently, and in case of studies with related subjects or subjects from various ethnicities some steps were performed on study subsets separately (as noted below). First, bi-allelic SNP markers were selected using the chip-specific manifest file and criteria 'Intensity_only=0' (to exclude CNV markers) and 'Exp_Clusters=3' (to keep only SNP markers for biallelic sites). Miscalled genotypes have the potential to create false positives, although the effect of sporadic errors should be minor. We applied loose filters to control the effect. Specifically, we removed markers with missingness rate > 10% and markers that showed departure from Hardy-Weinberg proportions (Chi-square or exact test p-value < 10^{-5}).

The samples for each study were phased in one group using either fastPHASE (62) (Melanoma, Preterm, and Glaucoma studies) or Beagle (54) (remainder of studies), except the Prostate study, for which the samples genotyped on the 1M and the 660 were phased separately. The choice of phasing software was arbitrary. Any genotypes that were imputed during Beagle phasing were subsequently masked.

4.2.2 hapLOH HMM parameters and event calling

The hidden Markov model (HMM) was set up to have two states, the minimum possible number. A two-state model was chosen since we expect these samples to have very low complexity, with only one event in most samples that have any events, and using an excessive number of states could lead to false positives. Transition parameters for each study were calculated using the total marker count and

approximate heterozygote rate to specify an average imbalance event size of 20 Mb and average genome-wide imbalance rate of 0.1%.

We performed two runs of the EM with starting values for the emission probabilities defined as $(p_n, p_n + 0.05)$ and $(p_n, 0.95)$, where p_n is the sample-specific average phase concordance rate (calculated using all informative markers). We tested higher numbers of EM runs with different starting values, and found that the strategy we chose produces results very similar to those from running 10 EM starts with starting values spread throughout the parameter space. Each EM run continued until the log likelihood increase was smaller than 0.0001 (usually between 4 and 20 iterations), and the parameter set with the highest log likelihood was used to calculate posterior probabilities.

The unit of inference for hapLOH is a pair of consecutive heterozygous markers. hapLOH outputs the posterior probability of each underlying state (in this case normal state, imbalance state) at each pair, or in other words at each inter-heterozygote interval. This is effectively a continuous scoring of imbalance across the genome for each sample. To create a list of discrete events, we applied a threshold of 0.95 to the probability of being in the imbalance state and defined an event as a run of intervals with probabilities exceeding the threshold. The starting and ending base positions are define by the left-side marker of the first interval and the right-side marker of the last interval of the run.

4.2.3 Quality control

DNA quality

An elevated value for α_0 indicates a potential sample quality issue, such as a low level of inter-sample contamination, which may create a false positive signal. To control for these effects, we exclude samples with estimated $\alpha_0 > 0.52$.

Genomic waves

Excessive LRR waviness interferes with proper filtering of inherited duplications and classification of events by copy number. We calculated a LRR waviness wf score using PennCNV (75) for each sample, and samples with scores $|wf| > 0.04$ were excluded.

HLA region

The initial results included a large number of calls overlapping the HLA region in the Melanoma study results. The Melanoma study was genotyped on the Omni1-Quad array, which has much higher density in the HLA region than previous arrays. We suspect the inflated call rate to reflect possible marker specificity issues or inherited polymorphism and chose to take a conservative approach, excluding any events overlapping the HLA region (genomic coordinates chr6:29,677,984-33,485,677, taken from (76)).

Manual exclusions

We excluded all DNA samples that were taken from cell lines or went through whole-genome amplification, as these samples have a higher probability of carrying imbalance acquired after sample collection that may be confused for true *in vivo* imbalance. For one sample in the Venous Thromboembolism study, over 75% of the genome was called as imbalance. This sample is likely a case of inter-sample contamination, but did not fail the α_0 threshold. We excluded this sample from analysis.

Boundary editing

For some analyses (i.e. analyses that used genomic sizes and for the concordance analysis with the Laurie *et al*. calls), we forced all event calls in our dataset to be contained within a single chromosome by trimming event calls that crossed into an adjacent chromosome by <30 markers and splitting event calls that had >=30 markers on two adjacent chromosomes.

4.2.4 Event classification

We used a simple thresholding procedure to call events. Plotting the BAF and LRR deviations for all events shows the cluster corresponding to inherited duplications. Based on the observed clustering, we chose a LRR deviation threshold of 0.08 and a BAF deviation threshold of 0.10, and any events in the upper right quadrant defined by these thresholds were removed as likely inherited duplications. We expect these parameters to result in some events that are true high-frequency somatic events to be called as germline duplications; we accept this loss of sensitivity to maintain specificity.

The remaining calls were interpreted as mosaic events. Those with LRR deviation > 0.05 were classified as gains and those with LRR deviation < 0.05 were classified as losses. After this filter, the remaining calls include the CNLOH events and events involving very low cell fractions, for which we expect the LRR deviation will be small even if there is a copy number change. Events with BAF deviation > 0.1 were classified as CNLOH, and the type for the remaining events (with small LRR deviation and small BAF deviation) was considered 'undetermined'.

4.2.5 Comparison with Laurie et al. calls

The sample identifiers for the 514 mosaic events identified by Laurie *et al.* were provided to us upon request (C. Laurie, personal communication). We used the genomic positions to define the extent of overlap between hapLOH events and Laurie *et al*. events. We considered only location, not event classification, when defining concordance. Within the subset of events with any overlap, the vast majority of events had more than 80% overlap with an event in the other analysis. For simplicity, we deemed calls to be concordant if they had any overlap with calls in the other analysis. All calls in the Laurie *et al*. analysis were within chromosomes because segmentation was performed separately for each chromosome.

4.3 Results

4.3.1. Quality control summary

We started with 36,293 unique samples after removing duplicates, whole-genome amplified samples, and cell line samples. Applying a threshold to the *a posteriori* probability of imbalance resulted in an initial set of 3,478 events in 2,903 samples. 809 samples were flagged based on α_0 values, and 4,965 samples failed the genomic waves filter, including 583 samples that failed both of these criteria. In total, 31,101 samples (86% of initial set) passed sample-level QC, and these contained 2,657 calls. From these, 1,471 events in 1,426 samples were removed as likely duplications, and another 57 calls were removed because

they overlapped the HLA region. The final set of mosaic calls includes 1,129 events in 895 samples. A breakdown of the filtering per dataset is presented in Table 4 and Appendix Table 5.

Forty-six mosaic events were called as covering multiple chromosomes. All chromosomes for a sample are considered together (by simply concatenating the data for all chromosomes, in order). It is possible that consecutive chromosomes have allelic imbalance at arm termini that are adjacent when they are lined up, but these are more likely due to imprecise boundary definition. We expect the latter to occur especially for imbalanced regions with low phase concordance, where there in not much contrast in phase concordance with normal regions. Using the criteria defined in Materials and Methods, 14 multi-chromosome events were trimmed and 32 multi-chromosome events were split.

Genomic waves associated with GC content are expected to manifest only in LRRs, not BAFs, since GC content will affect variation of probe binding across markers but will not affect relative allele intensities within a marker. The observed BAF and LRR patterns support this. Since hapLOH signal is generated based on BAFs, not LRRs, GC-associated genomic waves are not expected to be a direct cause of false positive imbalance calls. The waviness filter, then, may be expected to result in filtering of true somatic mosaic events. This filter is still important, however, because accurate LRRs are critical for identifying inherited duplications, which if not removed would be false positives.

The Glaucoma dataset had a large number of samples (1,155/2,185, 53%) that failed the filter for LRR waviness. We also noticed that the inherited duplication call rate in this dataset was two to three times smaller than the rates in the other projects, and the mosaic call rate is among the highest across datasets, provoking us to hypothesize that the LRR waviness filter was not completely effective at removing samples with inaccurate LRR and that the set of mosaic calls may contain events that are truly inherited duplications. However, upon manual examination of the BAF and LRR data for all 53 of the Glaucoma events that passed QC and were classified as mosaic, we noticed that all of the events had a lower level of BAF divergence than expected from inherited duplications, or displayed a sharp LRR decrease that indicated a copy-number loss event. Further investigation is necessary to explain the low rate of duplication calls in the dataset.

67

	sample counts						events classified as mosaic				events classified as inherited duplication		
phenotype/ study	pre-QC	failed α_0		failed wf		post-QC	number of samples	count	frequency	rate (average count/sample)	number of samples	count	rate (average count/sample)
Addiction	2,801	9	<1%	131	5%	2,661	22	23	0.8%	0.009	129	136	0.051
Melanoma	3,033	79	3%	124	4%	2,871	83	101	2.9%	0.035	165	170	0.059
Lungcancer	1,629	4	<1%	79	5%	1,546	63	83	4.1%	0.054	84	88	0.057
Preterm	3,747	207	6%	608	16%	3,109	32	32	1.0%	0.010	126	126	0.041
Glaucoma	2,185	132	6%	1,155	53%	1,003	53	63	5.3%	0.063	18	19	0.019
Prostate_1M	4,770	19	<1%	545	11%	4,211	190	270	4.5%	0.064	249	259	0.062
Craniofacial	7,089	230	3%	1,332	19%	5,710	55	59	1.0%	0.010	266	277	0.049
Prostate_660	4,346	35	1%	212	5%	4,106	250	313	6.1%	0.076	169	171	0.042
Vte	2,594	12	<1%	60	2%	2,521	75	103	3.0%	0.041	105	110	0.044
Lunghealth	4,099	82	2%	719	18%	3,363	68	82	2.0%	0.024	115	115	0.034
	36,293					31,101	891	1,129	2.9%		1,426	1,471	

Table 4. hapLOH call summary. The 'failed α_0' and 'failed wf' categories are not mutually exclusive. One sample from the Vte dataset was also excluded based on manual inspection.

4.3.2 Event classification

For each event, we attempted to interpret the BAF and LRR data together to classify the event as mosaic copy-number loss (hemizygous deletion), mosaic copy-number gain, mosaic copy-neutral heterozygosity, or inherited duplication. Inherited duplications, unless they include markers with uncalled genotypes, will create a strong signal of imbalance in the BAFs and are likely to make up a non-negligible portion of the called events. The inherited duplications are expected to have characteristically increased LRR deviation and increased BAF deviation.

The observed relationship between BAF deviations and LRR deviations generally matches well with the theoretical values expected for mosaic hemizygous deletion, CNLOH, and duplication events and inherited duplications (Figure 12). The plot emphasizes that imprecision in the observed deviations creates difficulty in discriminating different event types when the BAF deviation (and mutant cell fraction) is low. Using the criteria described in the previous section, we classified 202 events as mosaic losses, 67 as mosaic gains, 30 as mosaic CNLOH, 1471 as inherited duplications, and left 830 as unclassified.
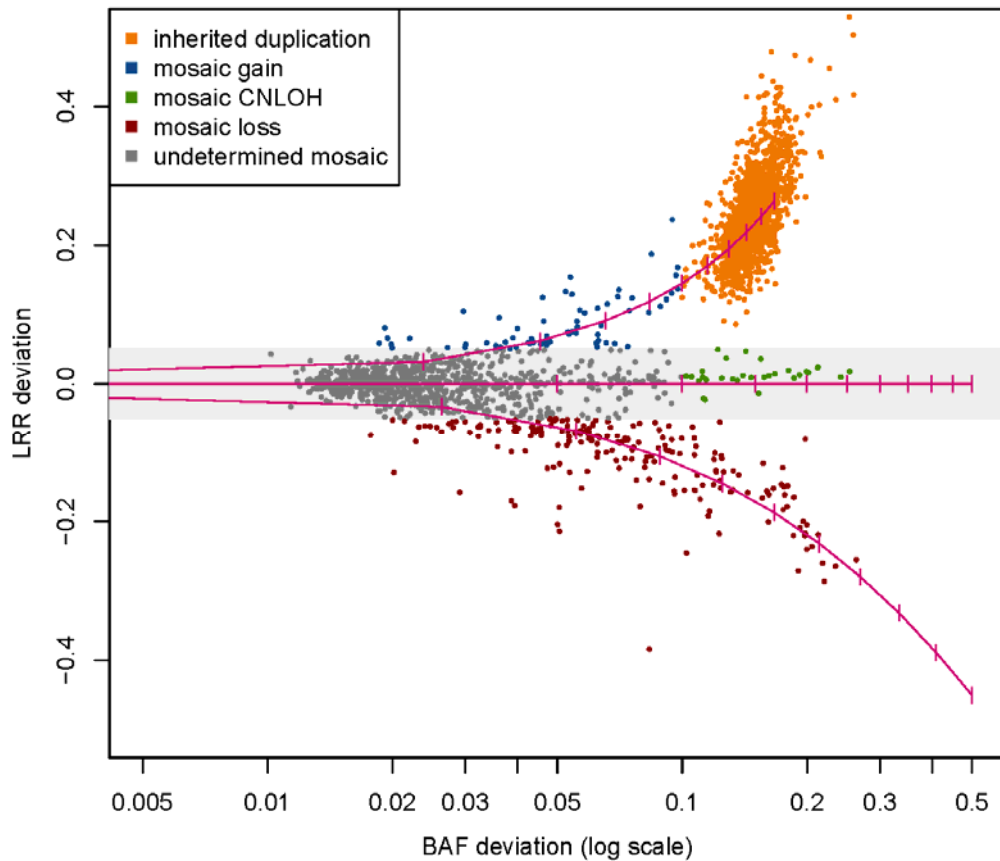
Figure 12. Event deviations colored by event type. The grey shaded region indicates the area within the thresholds used to define mosaic copy number loss or gain. The pink lines indicate the expected values for mosaic losses (lower line), mosaic CNLOH (middle line), and mosaic gains (upper line) for events occurring in from 10% to 100% of the sampled cells (dashes at 10% increments).

Inherited duplications. Other studies (25) have used large size as an indicator that an event must be a somatic, not inherited, mutation, based on the assumption that such large events will not be present in phenotypically normal individuals. We made our initial classifications based only on BAF and LRR deviations, but did want to investigate large events in the inherited duplication category. Out of the events classified as inherited duplications, 18 were larger than 10 Mb and 10 were larger than 20 Mb. Three of these cover the whole of chromosome 8, one covers the whole of chromosome 9, and one covers the whole of chromosome 12. The next two largest events cover arms 1q (102 Mb) and 2p (89 Mb) and the terminal portion of 13q (58 Mb). All of these occurred in blood samples. The whole-

chromosome events likely had a somatic origin since trisomies for both of these chromosomes are thought to be incompatible with life or survival past infancy. We do not have enough information to make any comparable inference about the other large arm-level events classified as duplications.

One large 58 Mb event at the terminal portion of 13q and five other calls between 10Mb and 25 Mb occurred in samples from the craniofacial study, in which case-parent trios were collected. All of these events occurred in children, and the data for parents was available for five of them. In all five of these trios, both parents' genomes looked normal in the region that was classified as an inherited duplication in the child. Assuming they are true duplication regions, the possible explanations for the generation of these events, then, are *de novo* duplication or mosaicism in one parent or somatic mutation in the child. We noticed that some of these events had lower phase concordance than expected for somatic events with such strongly diverged BAFs. In fact, the pattern of phase concordance in the inherited duplication category is qualitatively different from the pattern in the mosaic category (Figure 13). When the two BAF bands in a mosaic region are completely diverged, the BAF-based phasing will provide very accurate inference of the true haplotypes and the observed phase concordance will be limited only by the accuracy of the statistical haplotype estimates, so should be well above 0.9. Only one of the calls classified as mosaic with BAF deviation > 0.1 have phase concordance less than 0.9, and that call has such diverged BAF bands that there are many missing genotypes and what look to be heterozygote-to-homozygote genotyping errors that may have interfered with the statistical phasing; in the set of calls classified as inherited duplications, which have deviation > 0.1 by definition, 16% percent of events have phase concordance less than 0.9.
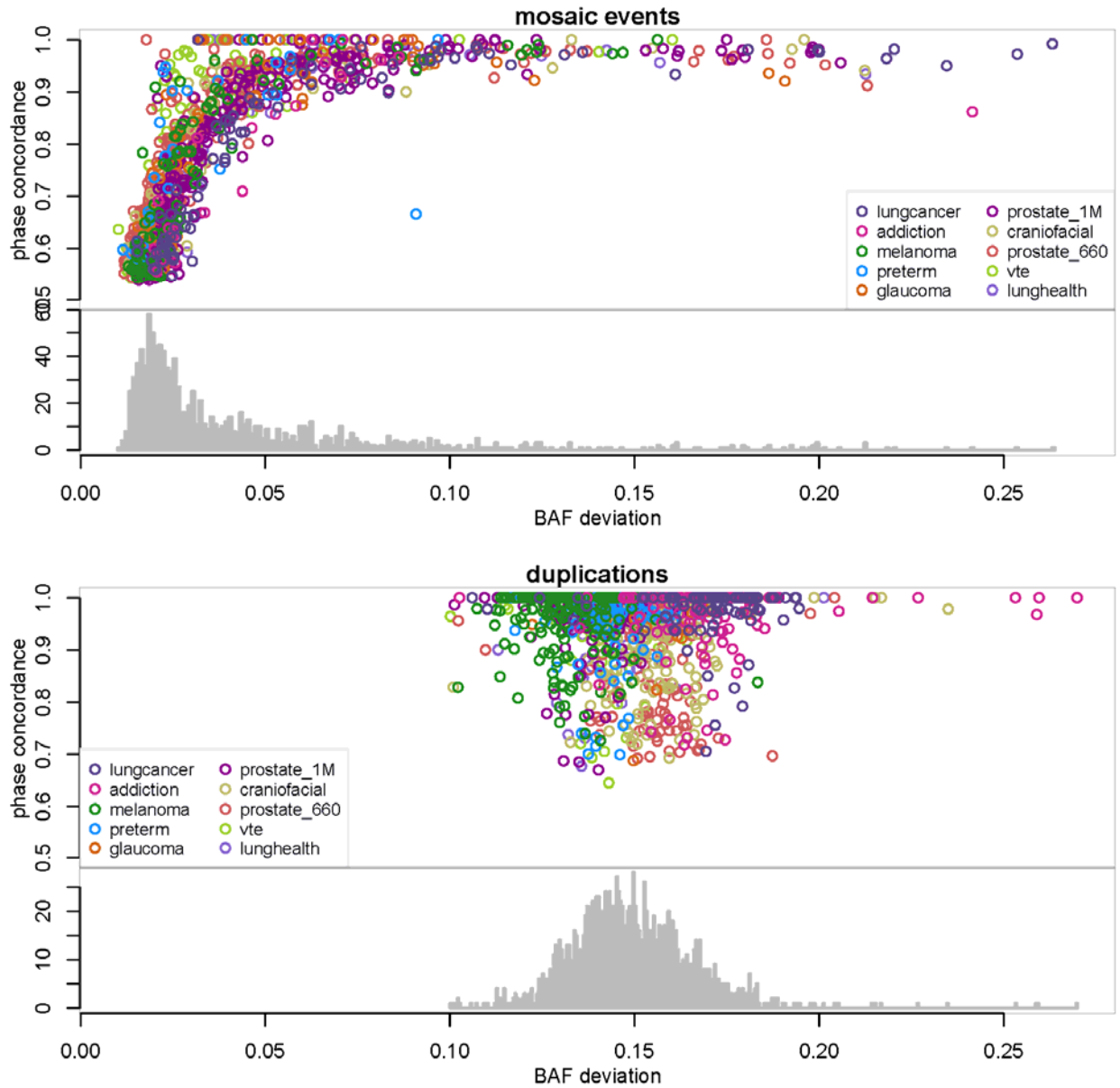
Figure 13. Comparison of patterns of BAF deviation versus phase concordance for mosaic and duplication calls.

We hypothesize that these low phase concordance events in the duplication category represent duplications for which the three copies are unique, which can only occur as an inherited duplication (with one parent passing on both homologous copies in the region, and the other parent contributing a different copy), since in a somatic duplication there are only two source haplotypes to begin with. We

used the five large duplication calls in the children from the craniofacial trios to investigate this hypothesis. We applied an algorithm under development in the lab (77) that uses the B allele frequencies from the child and statistical haplotype estimates for each parent to compare the likelihood that the child carries both homologues from one parent and one homologue from the other parent versus the likelihood that the child carries two identical copies of one homologue from one parent and one homologue from the other parent. The results are presented in Table 5. The events with high phase concordance had low evidence for presence of three copies, while the two events with the lowest phase concordance had high evidence for three unique copies. One event had an intermediate phase concordance level (0.82) and low evidence for three unique copies. Upon looking at the data for this event we noticed that one segment appeared to have a higher rate of heterozygotes, which prompted us to suspect that it may be a complex event in terms of allele identity. We re-analyzed this sample with a 2-state hapLOH model (keeping all other parameters the same as in the initial run) and found that the event was split into two distinct events, the left one with low phase concordance (0.76) and the right one with high phase concordance (0.95). We ran the likelihood calculation algorithm again separately for each segment and found that the left segment was found to have high probability of three unique copies. From these results we conclude that at least three of the events larger than 10Mb are inherited duplications, and that unexpectedly low phase concordance in events with high BAF deviation is a useful indicator of non-somatic origin. Again, these three events appear to be *de novo* and occurred in individuals with congenital physical abnormalities, so inherited duplications this large in phenotypically normal individuals may be very rare. Nevertheless, we did not use size or assumptions on biological impact to reclassify any events that met the BAF and LRR deviation criteria we defined for inherited duplications. We chose these thresholds to be conservative about calling mosaic events, and it may be that the inherited duplications set includes events that are high-frequency somatically acquired mutations (e.g. the whole-chromosome 8 and 12 events).

sample	chr	start	end	size (Mb)	location	phase concordance	evidence for unique homologues?
lungcancer_897	8	199,152	146,244,526	146.0	whole chromosome 8	0.98	-
vte_2212	8	166,818	146,106,670	145.6	whole chromosome 8	0.99	-
prostate_660_J_484	8	229,189	145,647,819	145.4	whole chromosome 8	0.96	-
vte_970	9	36,587	139,814,152	139.8	whole chromosome 9	0.96	-
prostate_1M_597	12	120,500	132,272,282	132.2	whole chromosome 12	0.99	-
vte_1517	1	144,155,913	246,560,134	102.4	whole arm 1q	0.97	-
prostate_1M_1950	2	19,443	88,905,353	88.9	whole arm 2p	0.99	-
craniofacial_children_1634	13	55,173,808	113,595,154	58.4	terminal 13 q	0.97	no
craniofacial_children_1611	3	79,972	24,984,831	23.2	terminal 3p	0.72	yes
craniofacial_children_873	4	133,669,279	156,775,006	23.1	interstitial	0.74	yes
craniofacial_children_2321	11	15,077,706	35,343,994	20.3	interstitial	0.82	after splitting only
glaucoma_1208	20	46,015,978	62,382,907	16.4	complex terminal 20q	0.95	-
craniofacial_children_408	20	47,690,212	62,199,875	14.5	centromere-adjacent 20	0.72	-
craniofacial_children_1919	8	36,539,160	50,382,228	13.8	straddles centromere	0.97	no
preterm_2152	11	90,028,983	103,605,932	13.6	interstitial	0.97	-
craniofacial_children_469	8	127,771,130	138,008,258	10.2	interstitial	0.95	-

Table 5. Calls classified as inherited duplications with size >10Mb. The likelihood calculation was performed only for samples from individuals for whom parental genotypes were available.

Another possible scenario is that we have classified some inherited copy-number variants as mosaic events. This may occur if the BAF deviation, LRR deviation, or both are not representative of the true allelic imbalance level and copy number, which in turn is possible if the data normalization was not effective or the data quality is low or the called event boundaries do not match the true event boundaries. To investigate the possibility that some of our mosaic calls may actually be inherited copy-number variation, we sought to compare the locations of our calls to locations of known copy-number polymorphisms. We downloaded the variants catalogued in the Database of Genomic Variants (DGV) (64), a curated database of copy number variants larger than 50 bp that have been identified in healthy individuals. The download included about 150,000 entries. Less than 0.4% of the reported variants are larger than 1 Mb; 84% are smaller than 100 Kb. By contrast, 85% of the hapLOH mosaic calls are larger than 1Mb and less than 0.7% are smaller than 100 Kb. Since the hapLOH calls do not necessarily have precise breakpoint definitions, we chose to compare the calls to the DGV variants by looking for >80% between a hapLOH call and a known polymorphism (each must have had > 80% overlap with the other), and 29 (2.6%) of the mosaic calls met this criterion. Further inspection of the breakpoints and event

types for these events will help to make an inference on whether these events are somatic or inherited variants.

Mosaic call summary. Mosaic calls ranged in size from 46 kb to 146 Mb, with a median size of 21 Mb. The smallest observed phase concordance value was 0.53. A histogram of the event sizes and a plot of size against phase concordance are presented in Figure 14. The vast majority of events with low phase concordance are large. This observation reiterates that both the size of the event and the phase concordance together impact the detection power of the hapLOH HMM, and that the HMM is not well suited for detecting small events with low phase concordance.
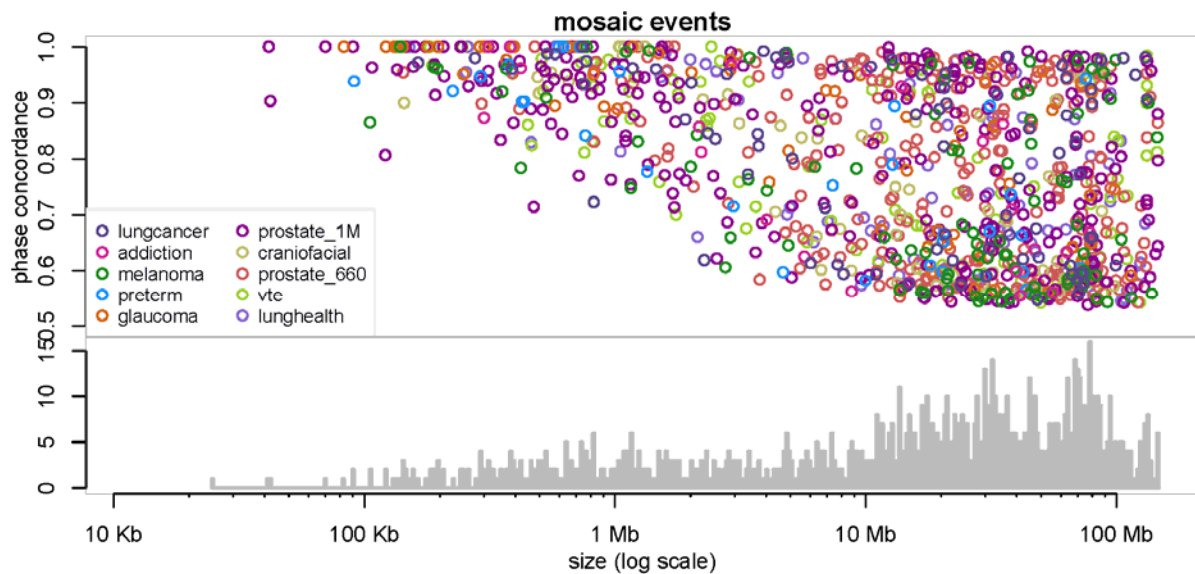

Figure 14. Comparison of size to phase concordance for each mosaic call. Histogram below plot represent distributions for size.
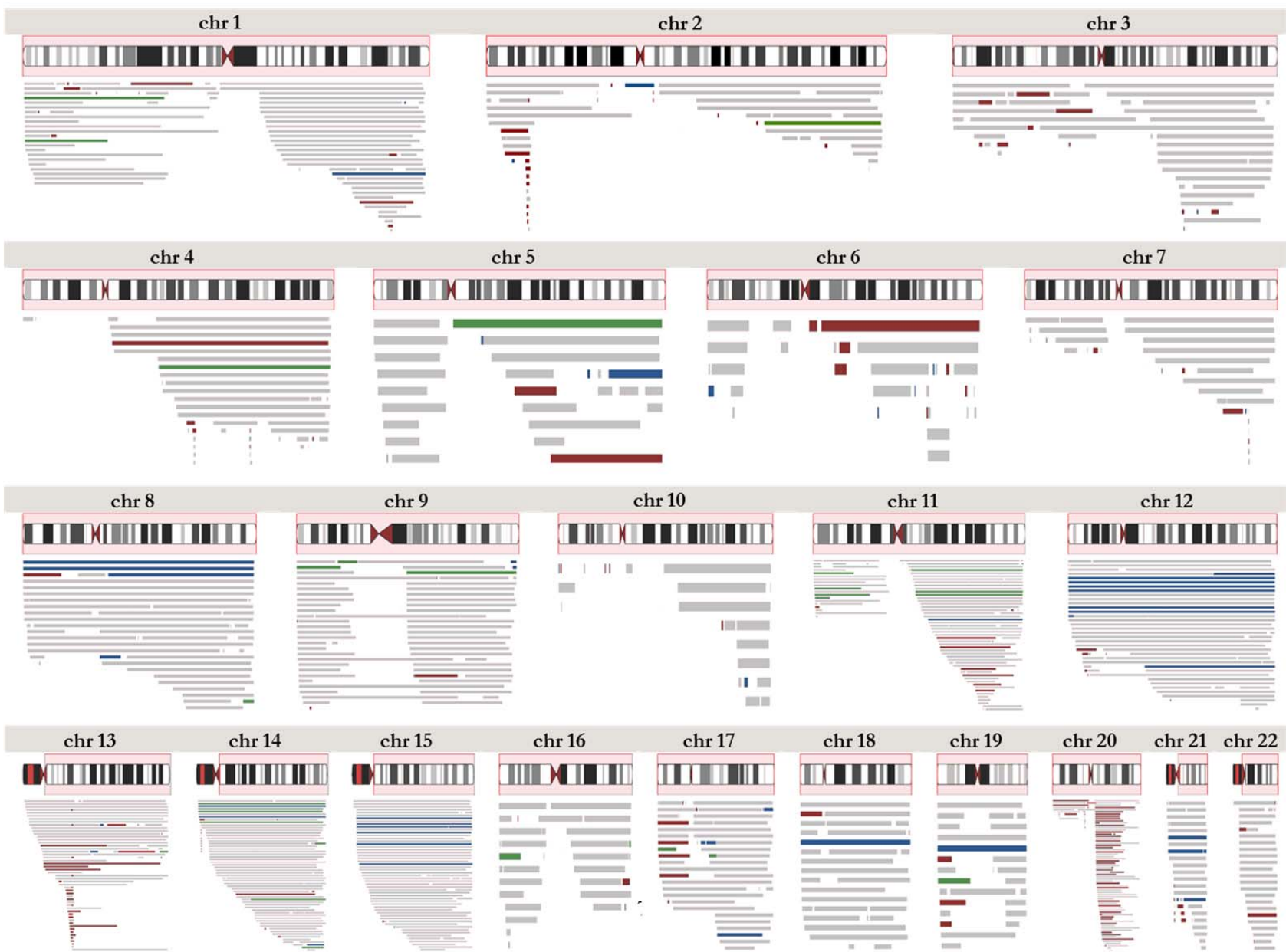
Figure 15. Mosaic calls by chromosome.   The red shading on each chromosome ideogram indicates the region represented in the plot space below the

ideogram. Calls are represented by horizontal bars, colored by event type: red-loss, green-CNLOH, blue-gain, grey-undetermined. Larger versions of

the plots are presented in Appendix Figure 5.

The phase concordance in an imbalance region is affected by multiple factors including the event type, fraction of cells affected, level of BAF noise, and accuracy of BAF threshold specification. The first two of these factors also affect the BAF deviation. As expected, phase concordance and BAF deviation have a positive relationship until the deviations reach about 0.10 (Figure 13), at which point the phase concordance has reached a maximum (governed by the accuracy of the statistical haplotype estimates) and the relationship plateaus. The obvious outlier visible in this plot is a 20.4 Mb event from the Preterm study with BAF deviation 0.09 and phase concordance 0.665. The BAF plot for this region exhibits 7 bands and the LRR data suggest an average diploid copy number. The data pattern is consistent with an inheritance of three unique copies of the region followed by loss of one chromosome in one subclone and loss of a different chromosome in another subclone, with loss of all cells carrying the duplication. The locations of the BAF bands indicate subclone fractions of approximately 1/3 and 2/3.
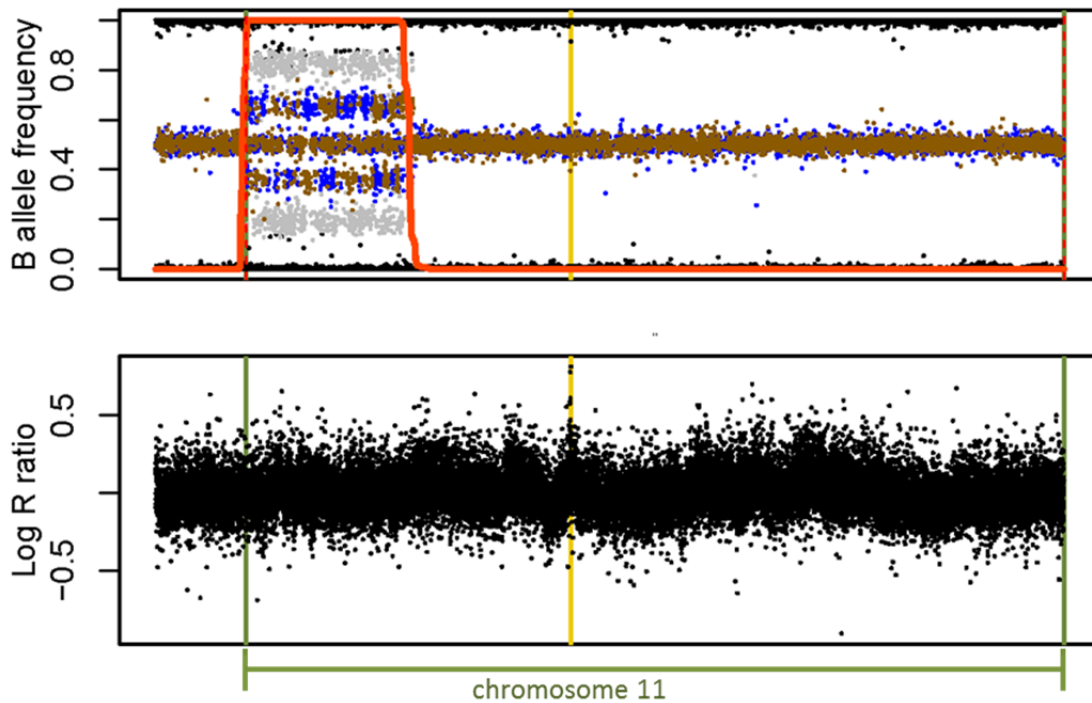

Figure 16. Complex 20 Mb event. The BAF pattern suggests two independent partial trisomy rescue events in a region in which one parent transmitted both homologues in the region and the other parent

transmitted a unique homologue. In the BAF plot, blue and brown points indicate markers with

heterozygous genotype calls, black points indicate homozygous genotype calls, and grey points indicate

genotype no-calls.

4.3.3 Comparison with Laurie et al. results

We used different sample level filters than were used in the Laurie *et al.* analysis, and also did not gain

access to all of the GENEVA datasets. The intersection set of samples that passed QC and were included

in both analyses contained 30,334 samples. In these samples, hapLOH made almost three times as many

calls as Laurie *et al*. (1,119 v. 379). The locations and event classification of all calls in the intersection

sample set are presented in Appendix Figure 5. For overlapping calls, the breakpoints are, for the most

part, very similar. We note that Laurie *et al.* manually edited breakpoints by visual inspection of the

BAF and LRR plots for each putative event that was called by their automated segmentation algorithm,

whereas we performed no editing of breakpoints other than the trimming or splitting of multi-

chromosome calls. Laurie *et al.* also manually removed some event calls if the BAF and LRR plots did

not show visually apparent shifts.

According to the concordance criteria described in the previous section, there were 295 hapLOH

calls that were also made by Laurie *et al.*, 88 calls made by Laurie *et al.* only, and 824 calls made by

hapLOH only. Another 111 calls made by Laurie *et al.* were in samples that failed hapLOH sample-

level QC, and 42 hapLOH calls were in samples that were excluded from the Laurie *et al.* analysis,

probably due to failure to meet the sample-level QC criteria employed in that analysis. Concordance

results by dataset are presented in Table 6.

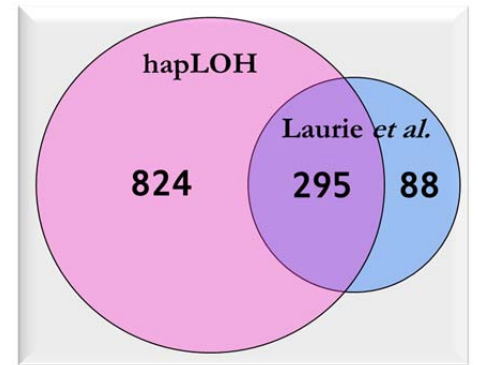|	hapLOH calls				Laurie *et al.* calls			
	not found by Laurie *et al.*		found by Laurie *et al.*		found by hapLOH		not found by hapLOH	
addiction	19	79%	5	21%	5	83%	1	17%
melanoma	83	78%	23	22%	24	75%	8	25%
lungcancer	53	62%	32	38%	32	80%	8	20%
preterm	27	84%	5	16%	5	56%	4	44%
glaucoma	41	68%	19	32%	19	76%	6	24%
prostate_1M	178	71%	72	29%	69	79%	18	21%
craniofacial	42	71%	17	29%	17	77%	5	23%
prostate_660	230	73%	87	27%	85	79%	22	21%
vte	86	83%	17	17%	17	57%	13	43%
lunghealth	65	78%	18	22%	18	86%	3	14%
Overall	824	74%	295	26%	291	77%	88	23%

Table 6. Concordance between Laurie *et al.* and hapLOH calls. Concordance was calculated using only the 30,334 samples included in both the hapLOH and Laurie *et al.* analyses. The columns give the numbers and fraction (out of the total made in the dataset in that analysis) of concordant calls (purple columns) and discordant calls (pink and blue columns). Discrepancies occur between the counts in the purple columns when a call in one analysis overlaps multiple calls in the other analysis. The Venn diagram summarizes the concordance between the analyses.

The events that were called only by Laurie *et al.* fall into three categories. The first category comprises 44 events that include few to no heterozygous calls. These may include regions of inherited homozygosity (either by chance identity-by-descent or germline CNLOH or deletion), or regions in which a large enough fraction of cells have imbalance that the BAF and/or LRR values are so aberrant that the Illumina genotyping algorithm could not make genotype calls or calls homozygous genotypes. hapLOH relies on heterozygous genotypes to calculate phase concordance, and also relies on the genotype calls accurately representing the inherited genotypes even in regions of allelic imbalance. The second category is short events. Twenty-two of the events called only by Laurie et al. had fewer than 100 markers, and another 11 had between 100 and 200 markers. With the parameter settings that we used for this analysis, hapLOH is not well suited to pick up events of this size. The third category of

events is large events that were detected by hapLOH but were excluded from the set of mosaic calls. This category included 6 large events that we excluded as inherited duplications, and 5 calls that overlapped the HLA region.
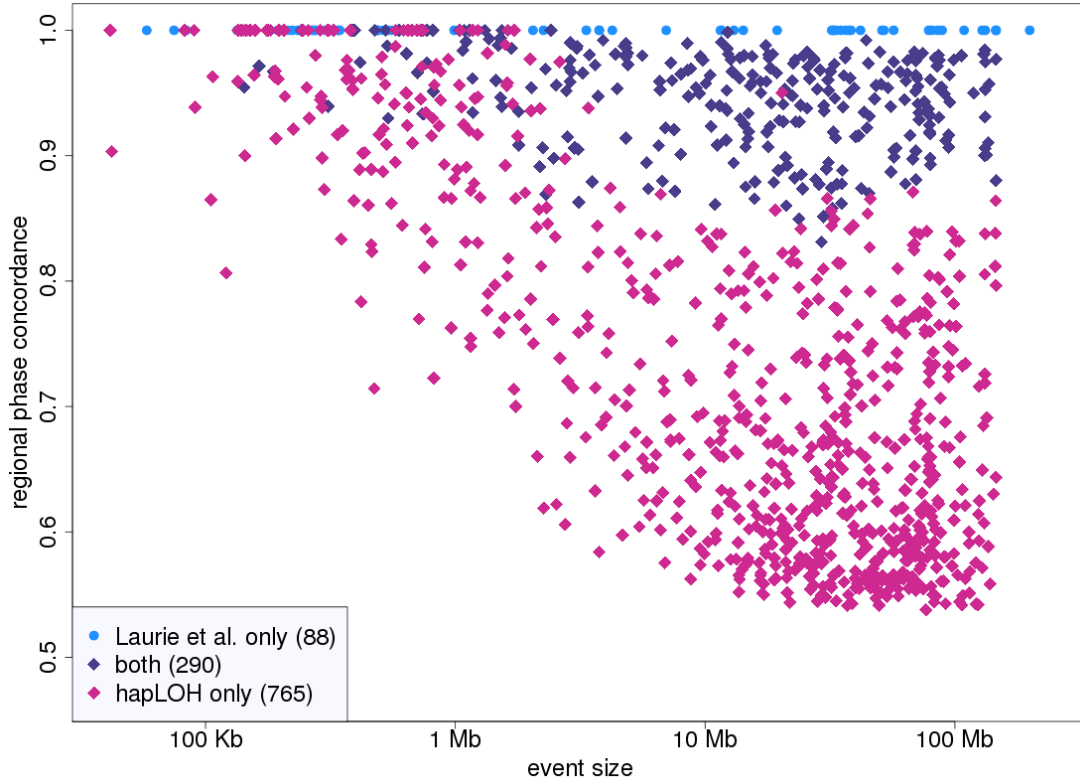

Figure 17. Genomic size versus phase concordance. Each point represents a hapLOH or Laurie *et al.* call, and colors indicate the concordance status. Most of the Laurie et al. calls had little or no heterozygous markers so phase concordance could not be calculated; all of these are plotted at phase concordance 1.

The genomic size and phase concordance for events called by both analyses and the events called by hapLOH only are summarized in Figure 17. This plot is similar to that in Figure 14, except only includes hapLOH event calls in the 30,334 intersection samples with Laurie *et al.* Within events with phase concordance >0.8, the rate of concordance is 71% for events larger than 1 Mb and only 18% for events smaller than 1 Mb.   No events with phase concordance <0.8 were called by Laurie *et al*. Although we do not know if all the hapLOH calls represent true events, our results indicate that hapLOH has the

potential for more sensitive detection of low-cell-fraction mosaic events than the method employed by Laurie *et al*.

4.4 Discussion

The results presented here suggest that somatic mosaicism in phenotypically normal individuals is higher than was reported in recent large-scale studies. This is not surprising, since these studies note that they used methods that lose sensitivity for low-frequency mutations, of which we would expect there to be many. These results further add to the observational evidence of somatic mosaicism in normal tissue. We have not calculated the fraction of mutant cells for each of the calls we made because that calculation relies on knowing the copy number of the mutation, which we cannot reliable infer for the majority of the calls. We are working on an improved strategy for classifying calls by copy number that will take into account the length of the called region (to account for the level of precision in the deviation measurements) and the background noise for the specific sample carrying the call. This information combined with breakpoint analyses may be useful for learning about the mechanisms that generated the mutations (78).

The filtering method we used here for removing inherited duplications is crude, but we believe it is effective at reducing the number of inherited variants falsely called as mosaic. Large inherited duplications certainly occur but are rare, so given the sizes of the hapLOH mosaic event calls it is unlikely that many of these are inherited duplications. Inherited deletions and homozygosity by descent are not detected by hapLOH since they include no heterozygous genotype calls, so we did not need to take any measures to remove them.

We were not able to experimentally validate the mosaic calls we made, because we did not have access to the DNA samples. Previous tests of the method using lab-created cell line mixtures and *in silico* simulations suggest that the hapLOH method has high sensitivity, and we used sample-level filters to reduce the number of false positive calls generated from samples with low-quality data. Several other studies that used SNP array methods for detecting somatic mosaicism in non-malignant tissues have

conducted experimental validation, and provide support for the specificity of SNP array methods in general in this context. Rodgríquez-Santiago *et al.* (78) confirmed blood mosaicism for 42 out of 42 events using MLPA and microsatellite analysis on blood DNA from the same source sample as was used for SNP array discovery analysis. They also used FISH to confirm the events in 3 out of 4 matched bladder tumor samples for which the tissue was available; a mosaic trisomy of chromosome 22 found in the blood using SNP array was not detected in the tumor sample using FISH. Forsberg *et al.* used qPCR to validate calls (27). Many of the mutations we called occur in lower frequency than those detectable by the other SNP array methods, as evidenced by the very small BAF and LRR deviations; MLPA and qPCR may not provide conclusive evidence of a false positive for such low-frequency events. FISH would be more appropriate, but is only possible if an appropriate sample of cells was available. Evaluation of recurrent events across multiple datasets will help to provide support for the calls.

We used a 2-state HMM here for all samples, which was reasonable given that most samples should have at most one event, but using such a reduced model means that higher complexity samples may not fit. This may result in false negatives in samples carrying an event with strong imbalance (either a somatic event or an inherited duplication) and also another event with a lower level of imbalance. One way to investigate this possibility is to re-analyze the samples with an HMM with an additional event states, or censoring the data from the previously called event, or fixing the emission probability corresponding to a phase concordance value intermediate to that of the previously called event and the baseline.

The boundary definition for events may be imprecise, especially at events with low phase concordance. The most obvious examples of events with imprecise boundaries are those that extend into adjacent chromosomes, for which the called boundaries are outside of the true boundaries. We also see examples of higher-phase concordance events where the posterior probability drops below the calling threshold (i.e. 0.95) at one or both ends of the called region and it is apparent that the called boundaries are within the true boundaries. An alternative event calling strategy that would improve the boundary

definitions for these events would be to use a high threshold (i.e. 0.95) to discover events and a slightly lower threshold (between 0.5 and 0.95) to define the boundaries for the discovered event.

Our analysis was designed for discovery of low-frequency events. Since small events with low frequency create only weak signal, we implicitly were aiming for large events (and also explicitly, by use of a mean event size of 20 Mb). Small events with high frequency do create a strong enough signal that they are picked up using this parameter setting as well. This bias for events in certain size and phase concordance ranges must be kept in mind when we try to interpret the observed distribution of events. There is a biological rationale for expecting to observe large events mostly at low frequencies, since they might be more likely to have a negative impact on cell fitness and therefore keep a clone from ever reaching high frequencies. However, we do see a number of large events with high phase concordance, suggesting mutant cell frequencies of at least 15%; these may carry mutations that increase cell fitness or are neutral and were observed due to drift. The lack of observed events with small size and low frequency is clearly due to the lack of power to detect this category of events; we expect there to be a substantial number of true events in this category. One way to get a more comprehensive discovery of somatic events would be to combine a hapLOH analysis with an analysis using a different method that is more suited for events in this category.

One of the major observations of the Laurie *et al.* study was a sharp increase in the rate of detected mosaicism in elderly individuals compared to younger individuals. This observation may indicate a higher rate of somatic mutation in the elderly, which is consistent with the hypothesis that mutation rate increases with age due to reduction in DNA repair activity or increase in the incidence of errors (for example, increased incidence of structural rearrangements from loss of telomere function). An alternative explanation is that the mutation rate is largely constant over time, but detectable mosaicism is associated with age because in older individuals there has been more time for mutations to accumulate and the subsequent mutant clone to expand to a detectable frequency. Given the latter hypothesis, the strength of the observed association will be highly dependent on the minimum detectable mutation frequency of the method used to generate observations. Our analysis made many more

observations of low-frequency mutations than the Laurie *et al.* analysis. Comparison of age association

results between the two anal may help to define the relative contribution of each of the above

explanations for the higher mosaicism rate observed in the elderly.

Genes or variants within somatically imbalanced regions may be relevant to proliferation, given

that the mutant clones increased to detectable frequency. We so far have performed an initial

investigation of recurrently imbalanced regions, with chromosomes 15 and 20 carrying the regions with

the highest number of calls. Several of the recurrently imbalanced regions include genes that have been

associated with cancer. All of the blood samples in the study were collected from individuals without

diagnosed hematological cancer. We can therefore conclude that any observed event is not sufficient to

initiate transformation, but how important is its potential impact on cellular proliferation? Is it simply a

commonly occurring clonal mutation that has neutral impact at the systemic level, or does it promote the

risk of cancer and represent a valuable early disease biomarker? Follow-up on the hematological cancer

status and on any further accumulation of somatic mutations in individuals with blood mosaicism

involving cancer-associated genes will be helpful for making this distinction. We would expect the

landscape of tolerated and functional somatic mutations to vary by tissue, so similar studies using

samples from other tissues would provide complementary information. These results have great

potential to improve the interpretation of mutations observed in tumors. Regions with no observed

somatic imbalance may indicate the presence of genes that are sensitive to aberrant gene dosage and are

critical to cell function; we are also trying to characterize regions with very few imbalance calls.

CHAPTER 5
Conclusions and future directions

5.1 Overall significance of the project

The method presented here is the first method to utilize imperfect population-based haplotype estimates to discover low-frequency somatic chromosomal mutations from SNP array data. The major innovation of the method is the use of phase concordance as a robust metric to measure evidence of allelic imbalance in the face of sporadic phasing errors in the statistical haplotype estimates and stochastic variation in the B allele frequencies. In addition to describing a hidden Markov model that uses the phase concordance data to perform agnostic whole-genome discovery of imbalanced regions, we also describe how to test candidate regions and infer the haplotype of the major chromosome. We have demonstrated through controlled experiments that the sensitivity is higher than other existing methods while maintaining specificity, and that the strategy is applicable to both Affymetrix and Illumina data. The hidden Markov model has been implemented in publically-available software with a user-friendly command line interface and online documentation so that researchers may easily apply the method to their own data. The method is complementary to existing SNP array methods for detecting mosaic kilobase- to megabase-size CNV and CNLOH events, since it targets a lower range of the allelic imbalance spectrum (< 10% mutant cells) while still detecting mutations with mid-level imbalance. The superior performance in the low-imbalance range makes it especially useful for detecting mutations in tumor samples with high proportions of contaminating normal cells. We also demonstrate the potential of the method via a real-data analysis of over 30,000 samples from phenotypically normal individuals that were previously analyzed using another method. We made nearly three times as many calls in these samples, most of which appear to exist at low frequencies. These findings validate recent hypotheses that somatic variation in healthy tissues is more prevalent than had previously been reported due to

undercalling of low-frequency mutations, and provides valuable observations of *in vivo* mutations that can be studied to make inference on genetic robustness and how these mutations impact cell fitness.

5.2 Potential model extensions and software improvements

5.2.1 Alternative segmentation strategies and strategies to accommodate higher-complexity samples

Fitting a hidden Markov model to the switch accuracy data and then calculating conditional marginal probabilities (to which I have been referring as posterior probabilities) via the forward-backward algorithm provides a continuous measure of the local evidence of imbalance across the genome, which can in turn be transformed to define discrete regions of balance or imbalance. This segmentation approach has several advantages. For example, it is computationally efficient. The computational complexity of the forward-backward algorithm is $O(L^2N)$, where L is the number of hidden states and N is the number of markers. Basic CBS has complexity $O(N^2)$, but DNAcopy, the R package implementing CBS, uses a hybrid approach for obtaining p-values (combination of permutation to define empirical null distribution and approximation using standard normal distribution when number of markers in segment is large enough), and has rules for early stopping when the evidence for a segment is high, two strategies which improve computational complexity (79). So, the HMM is at least comparable or better in terms of computation time.

On the other hand, we have noticed several examples of suboptimal performance that may be ameliorated by adjusting the segmentation procedure. First, the success in calling events depends not only on how much the phase concordance for the event differs from the normal phase concordance, but also how it relates to the phase concordance of any other events in the sample, especially when a small number of event states is specified. The reason for this is that, in the default procedure, the emission probabilities for the normal state and each of the event states is iteratively updated to the MLE based on the observed data from the whole genome. The consequence of this is that, in samples with multiple events of different levels of imbalance, the emission probabilities are likely to fit the more extreme

values, and the more subtle events may go undetected if the phase concordance is closer to the normal phase concordance than the phase concordance for the more extreme events. The sizes of the true segments and the choice of parameters controlling the transition probabilities will also play a role. As a demonstration, we performed two runs of the hapLOH HMM on a series of lab-diluted tumor-normal mixture samples in which we changed only the number of hidden states (we ran a 2-state HMM and 4-state HMM). The tumor cells were complex, with multiple events of different types. There are a few ways to address this issue. The most obvious solution is to simply run the HMM with a larger number of hidden states. This may work well for more complex samples, but the downside of this is that if the specified number of states is larger than the number of states that hapLOH can distinguish (i.e. the sample is less complex than predicted), then some states may have very similar phase concordance values and the evidence for an event will be split between multiple states. If there are multiple states that end up with low emission probability, only one of which will be defined as the 'normal' state, then this strategy may result in false positives because segments that have slightly elevated phase concordance by chance may be called as events. An iterative procedure might work well, where the HMM is first run with a lower number of states and samples with event calls are re-run with a higher number of states to capture any lower-imbalance events that were missed. Another option is to mask the data from any called events and perform a second run of the HMM using the remaining data. These strategies can be expected to recover events with low phase concordance that were missed in the first run, and potentially also recover short events in samples that also have longer events (which would probably drive the transition probabilities to favor longer events). A third option is to fix the emission probabilities to ensure that there is an event state for different levels of imbalance.

A second issue that we observe in the hapLOH results is that the boundary definition can be imprecise. There are a few different categories of imprecise boundaries. For some events the boundaries are obviously conservative, and are within a region with visibly diverged BAF bands, because the posterior probability tapers off at the boundaries and falls below the event calling threshold. This slight decrease in posterior probability also sometimes results in one event being split into multiple events.

One option to address this is to use one threshold to identify locations of events, and a lower threshold to define the set of markers belonging to the event. For example, we could use 0.95 to identify a peak that indicates an event region and 0.5 to define the boundaries of that specific event. Another option is to use the endpoints indicated by the HMM segmentation as initial boundaries to conduct CUSUM-based (80) boundary refinement using the observed BAF values; this is similar to the strategy implemented in POD (53). I expect this will do well for events with moderate to high imbalance and will be particularly useful to identify and properly define overlapping events and adjacent events, and should provide some improved precision for low-level imbalance events as well. Another automated method that might be helpful for refining boundaries at low-frequency events is to perform a second pass of the HMM using emission probabilities fixed at low values.

5.2.2 Adaptation to sequence data

We have designed the hapLOH software for genotypes and B allele frequencies from SNP array data, but the principles underlying the method may also be used to detect allelic imbalance from whole genome or whole exome sequencing data. The most important difference is that sequence data provides discrete allele counts at each marker, while SNP array data provides a continuous allele frequency. The sampling variation in these counts should be taken into account, especially when the sequencing depth is low. Other important differences include mapping bias that may skew allele counts toward the reference allele, and unevenly spaced markers especially when using exome sequence data. All of these differences can be accommodated, and in fact a version of the method for sequence data has been implemented and applied to exome sequence data by a colleague in the lab (F. Anthony San Lucas, unpublished).

5.2.3 Automated post-processing

As is the case for many experiments, the sample preparation and results interpretation steps of a hapLOH analysis are much more time-consuming and challenging than the generation of results. Thus, utility

scripts for pre-processing and post-processing are among the most efficient ways to make hapLOH more easy to use.

The most onerous pre-processing step is statistical phasing. It is somewhat counter-intuitive, but this step is more complicated when sample size is low, because in that case a separate reference dataset must be used to fit the statistical model, which has multiple implications First, if the test and reference data use different sets of markers, then either the user must filter and/or augment each of them so that they contain exactly the same set of markers (fastPHASE), or the markers in the test dataset that are not included in the reference dataset will be dropped during phasing and the results files will not match the input files (BEAGLE, MaCH), and then the user will need to identify the dropped markers and modify the BAF files accordingly. Another issue is that the user must ensure that the allele designation scheme and genome reference version used to characterize the markers positions and polymorphisms are the same for the test and reference datasets, and convert one dataset or the other if they do not match, and then convert the test dataset back to A/B labeling since this is the allele designation scheme is required by hapLOH. One way to make this step easier is to provide ready-to-use or partially prepared reference data. fastPHASE provides an option to save parameter values after fitting the model. So we could fit the model to a set of reference samples using all of the SNP markers on, for example, the Illumina 2.5 array, and then output the parameters to files. Users would then need only to specify these files to rebuild the model and phase their samples of interest, without having to re-label alleles or filter the marker set. Other popular methods, including BEAGLE, MaCH, and IMPUTE2, and SHAPEIT, do not have the option of saving models.

The 1000 Genomes Project data is currently the only public reference dataset available that includes enough sites to be useful for phasing Illumina 2.5 array datasets, and the documentation pages for each of the aforementioned softwares include links to download 1000 Genomes Project data in the correct format for use as a reference dataset. However, the files are very large, much larger than necessary since for the purposes of running hapLOH it is not helpful to impute additional genotypes that are not on the array (since they will not have a B allele frequency value anyway). Ideally we would

provide reference files including just the markers on the Illumina 2.5 array, already in A/B format. A

single reference panel including individuals from diverse genetic backgrounds would be sufficient to

provide high-quality phasing for test samples from any genetic background (81).

Currently, there is no option in the hapLOH software to automatically call discrete events based

on applying thresholds to the posterior probabilities (the method I used in Chapter 4). This should be

one of the next functions to make available, since a list of discrete calls is generally easier to interpret

than the posterior probabilities themselves. Also, since it is highly advisable to visually follow up on

event regions to identify potential data quality issues that may create false positives or bias event-specific

statistics such as median BAF or LRR deviation and phase concordance, automatic generation of plots

for called events is also important. Both of these steps could be added as functions that could be turned

on with hapLOH command line flags, or could be provided as separate utility scripts.

5.3 Additional applications

In this work I presented two possible scenarios in which hapLOH is a relevant choice of method for

identifying low-frequency somatic segmental imbalance: identifying tumor mutations from samples with

high levels of normal contamination, and detecting non-malignant intra-individual variation in normal

samples. We envision several other applications in which hapLOH may be valuable. For example,

hapLOH may be useful for identification of low-frequency clones in samples with high tumor cell

fractions. In these cases it might be necessary to use a matched normal sample to infer the inherited

genotypes, since high-frequency imbalance in the tumor cells may affect the genotype calls. Another

potential application is multiple-timepoint analysis of mutation dynamics. This is feasible since

preparing samples for and running SNP arrays is cost-effective and not too laborious, especially if

multiple samples are prepared in parallel, and because the hapLOH method is efficient.

The usefulness of the method will be further improved by using custom-designed arrays such as

the Oncochip (82), which includes probes for candidate functional cancer mutations discovered through

sequence analysis on a backbone of markers tagging common variation that could be used for phasing.

This way somatic copy number and CNLOH mutations can be analyzed simultaneously with inherited sequence variants in an efficient manner.

Appendix

Tables

chr	start bp	end bp	marker count	size (Mb)
2	31,982,105	33,266,534	131	1.28
2	38,755,294	38,887,094	16	0.13
6	99,536	68,754,442	9,968	68.65
11	71,640,522	134,435,899	7,795	62.8
12	28,466,092	28,491,511	8	25.42
12	31,157,554	31,298,174	20	0.14
16	30,423,993	88,690,776	5,604	58.27

Appendix Table 1. Regions excluded from simulated sample data. Upon visual examination of the tQN-normalized array data from the normal cell line sample, we noticed several obvious structural variants, likely cell-line artifacts. In order to assess the sensitivity and specificity rates for our methods and the other methods, since these would persist in the computational dilution data, we decided to mask these regions of the genome to prevent them from creating false positive calls. We ran BAFsegmentation on the normal cell line sample using default parameters, which identified 8 segments on 4 chromosomes. One segment corresponded to a visible deletion on chromosome 16, and two contiguous segments corresponded to a visible deletion on chromosome 6. In these cases the segments did not cover the entire region that could be visually identified by looking carefully at the BAFs and LRRs. For these events, we excluded a region that included both the BAFsegmentation segments and any additional contiguous loci that appeared (by inspection) to be part of the same event. Another region on chromosome 6 was identified (11 SNPs only), approximately 700 SNPs downstream from the other segments. The exclusion region on chromosome 6 was extended to include this event. A segment on chromosome 2 corresponds to an obvious increase in copy number. Another small segment (16 SNPs) was identified about 800 SNPs downstream. These two regions were excluded according to the BAFsegmentation segment coordinates. A segment on chromosome 12 corresponds to a small duplication event. BAFsegmentation also identified another small event about 400 SNPs upstream. Both of these regions were excluded according

to the BAFsegmentation segment coordinates. An apparently heterogeneous region on chromosome 11 was also excluded.

chr	start index	end index	start bp	end bp	marker count	size (Mb)	type
2	11,800	17,750	102,083,769	162,645,233	5,951	60.56	del
3	10,850	14,200	95,977,726	128,522,090	3,351	32.54	CNLOH
4	1	1,450	63,508	11,233,183	1,450	11.17	del
4	1,800	3,050	14,085,628	24,243,339	1,251	10.16	del
4	14,450	19,985	141,764,847	191,164,126	5,536	49.40	del
5	16,000	19,916	149,859,181	180,642,521	3,917	30.78	CNLOH
7	1	1,250	149,081	10,030,213	1,250	9.88	CNLOH
8	1	6,200	166,818	47,383,928	6,200	47.22	del
9	3,100	6,450	15,879,411	38,350,333	3,351	22.47	del
9	11,800	14,750	109,184,150	130,054,427	2,951	20.87	del
10	14,300	16,209	122,528,909	135,284,293	1,910	12.76	CNLOH
12	4,650	15,374	37,245,320	132,288,869	10,725	95.04	del
13	4,000	4,300	48,985,143	51,376,986	301	2.39	CNLOH
13	5,200	12,029	60,650,097	114,108,121	6,830	53.46	CNLOH
14	2,150	3,100	35,794,972	45,110,909	951	9.32	del
14	4,950	5,300	61,557,191	64,441,966	351	2.88	CNLOH
14	5,800	7,200	69,548,910	81,198,589	1,401	11.65	CNLOH
15	1	3,550	18,421,386	53,338,047	3,550	34.92	del
18	1	1,700	2,842	11,531,256	1,700	11.53	CNLOH
18	6,850	9,950	53,442,584	72,398,882	3,101	18.96	del
19	1	1,450	217,034	11,590,705	1,450	11.37	del
19	3,800	6,115	43,854,070	63,776,118	2,316	19.92	del
21	3,700	3,950	37,537,165	39,062,090	251	1.52	del
22	1	800	15,276,762	21,136,795	800	5.86	del
22	3,400	5,540	36,380,965	49,524,956	2,141	13.14	CNLOH

Appendix Table 2. Details of events simulated in computational dilution dataset. Start and end indices are 1-based; Start and end chromosomal base pair positions (BP) are 1-based and given according to position annotations in the array manifest (which used the hg18 genome build).

Sample	start SNP	chromosome	start bp	end SNP	end bp	Number of informative markers	Phase concordance
samp1	rs17025785	3	30642429	rs9813622	30723944	29	0.931
samp1	rs7630256	3	67867664	rs12486635	79280400	1001	0.891
samp2	rs17649641	17	41353200	rs8080254	42702891	120	0.750
samp2	rs2825523	21	19627751	rs2826530	21108259	111	0.667
samp10	rs7528118	1	56741343	rs1022636	58133153	208	0.572
samp10	rs7642123	3	153610044	rs6797289	157080481	249	0.602
samp10	rs7692447	4	135190819	rs13151254	139767513	298	0.587
samp10	rs1461349	11	37157502	rs11821682	41821239	296	0.618
samp10	rs7488279	12	5438481	rs17728942	8179670	193	0.606
samp10	rs1342606	13	82759884	rs418853	87877460	328	0.576
samp10	rs9559492	13	87991344	rs1948851	89058641	14	0.786
samp10	rs1409911	13	89165586	rs4773607	90412505	129	0.589
samp10	rs1906160	15	37813417	rs2291620	38116265	16	0.750
samp10	rs3107997	18	26174242	rs8093901	27288097	146	0.596
samp10	rs6025034	20	54655431	rs13038808	54878739	41	0.610
samp11	rs478410	13	29554081	rs7325798	30495592	136	0.632
samp11	rs2911851	15	29118254	rs951443	31705123	229	0.620
samp12	rs13191136	6	167606330	rs9283859	169721000	190	0.616
samp14	rs494723	6	10854213	rs4574630	11726173	90	0.633
samp14	rs11794457	9	7358850	rs12235266	8292479	161	0.621
samp14	rs2507903	11	114504525	rs592525	118889063	355	0.594
samp14	rs3741808	12	67011218	rs4019400	74096025	625	0.602
samp14	rs785450	15	27928636	rs2596210	31579227	316	0.674
samp14	rs874187	19	49509008	rs12460033	51098538	76	0.776
samp15	rs774381	12	65726355	rs10878795	66911125	119	0.622
samp19	rs882311	10	1219548	rs622898	3512302	330	0.664
samp20	rs11166104	1	98989371	rs945748	103256102	374	0.602
samp20	rs6880142	5	149105642	rs153478	150619632	156	0.635
samp20	rs7758899	6	116374038	rs4946399	119470457	185	0.632
samp20	rs7077992	10	934496	rs7904349	3411055	335	0.687
samp20	rs2356535	14	50445000	rs2069002	51320440	87	0.621
samp21	rs6442749	3	2730983	rs4493418	2811418	17	1.000
samp21	rs9422881	10	127110624	rs7075452	127340239	33	0.879
samp21	rs4930561	11	67688337	rs7481750	68904167	82	0.780
samp21	rs6606662	11	68915925	rs9630218	69588488	38	0.789
samp21	rs674374	11	69817313	rs3017478	70375682	44	0.864
samp21	rs1028050	11	71929039	rs6592575	74000654	88	0.818
samp21	rs605954	11	74258861	rs1965277	134355825	5091	0.863
samp21	rs1556999	13	46870340	rs11148069	47117281	15	0.933
samp22	rs2714298	2	14769944	rs6717502	16858388	198	0.616
samp22	rs354707	2	143602907	rs1580063	148794127	287	0.603
samp22	rs17341291	3	133383722	rs9843725	137185555	403	0.600
samp22	rs6879951	5	123873768	rs17517907	124554764	83	0.711
samp22	rs513870	6	11616840	rs12211264	16021001	431	0.587
samp22	rs12701976	7	42474899	rs12375125	47349043	352	0.571
samp22	rs11218071	11	120322070	rs1238553	121207450	121	0.620

Appendix Table 3. hapLOH AI calls made in Barresi data set by applying threshold of 0.5 to posterior

probabilities.

full name	short name	dbGaP accession	array	DNA sources
Study of Addiction: Genetics and Environment (SAGE)	Addiction	000092.v1.p1	Human1M	blood cell lines
High Density SNP Association Analysis of Melanoma	Melanoma	000187.v1.p1	Omni1-Quad	blood
A Genome Wide Scan of Lung Cancer and Smoking	Lung Cancer	000093.v2.p2	HumanHap550	blood
Genome-Wide Association Studies of Prematurity and its Complications	Preterm	000103.v1.p1	660W-Quad	blood spot blood spot (WGA) Buffy coat Buffy coat (WGA)
The Primary Open-Angle Glaucoma Genes and Environment (GLAUGEN) Study	Glaucoma	000308.v1.p1	660W-Quad	blood cheek
A Multi-ethnic Genome-wide scan of Prostate Cancer, with Japanese and Latino substudies	Prostate_1M	000306.v3.p1	Human1M	blood
International Consortium to Identify Genes and Interactions Controlling Oral Clefts	Craniofacial	000094.v1.p1	610-Quad	blood

Appendix Table 4. GENEVA dataset dbGaP accession numbers and array types.

DATASET	FILTERS FAILED							CALLED AS MOSAIC
	sample QC	sample QC duplication	sample QC HLA	sample QC duplication HLA	duplication	HLA	duplication HLA	
addiction	4	10	0	0	136	1	1	23
melanoma	19	10	78	0	170	34	5	101
lungcancer	4	5	0	0	88	2	0	83
preterm	35	24	0	0	126	0	0	32
glaucoma	179	20	2	0	19	1	0	63
prostate_1M	63	28	1	0	259	1	5	270
craniofacial	122	53	33	2	277	1	1	59
prostate_660	25	5	0	0	171	4	0	313
vte	38	7	1	0	110	0	1	103
lunghealth	33	19	1	0	115	0	0	82

Appendix Table 5. Summary of filtered events. Each event may have been in a sample that failed QC, may have met the BAF and LRR criteria and been marked as a likely inherited duplication, may have overlapped the HLA region, or may have met failed multiple of these filters. Calls that failed none of these filters were categorized as mosaic events.

Figures
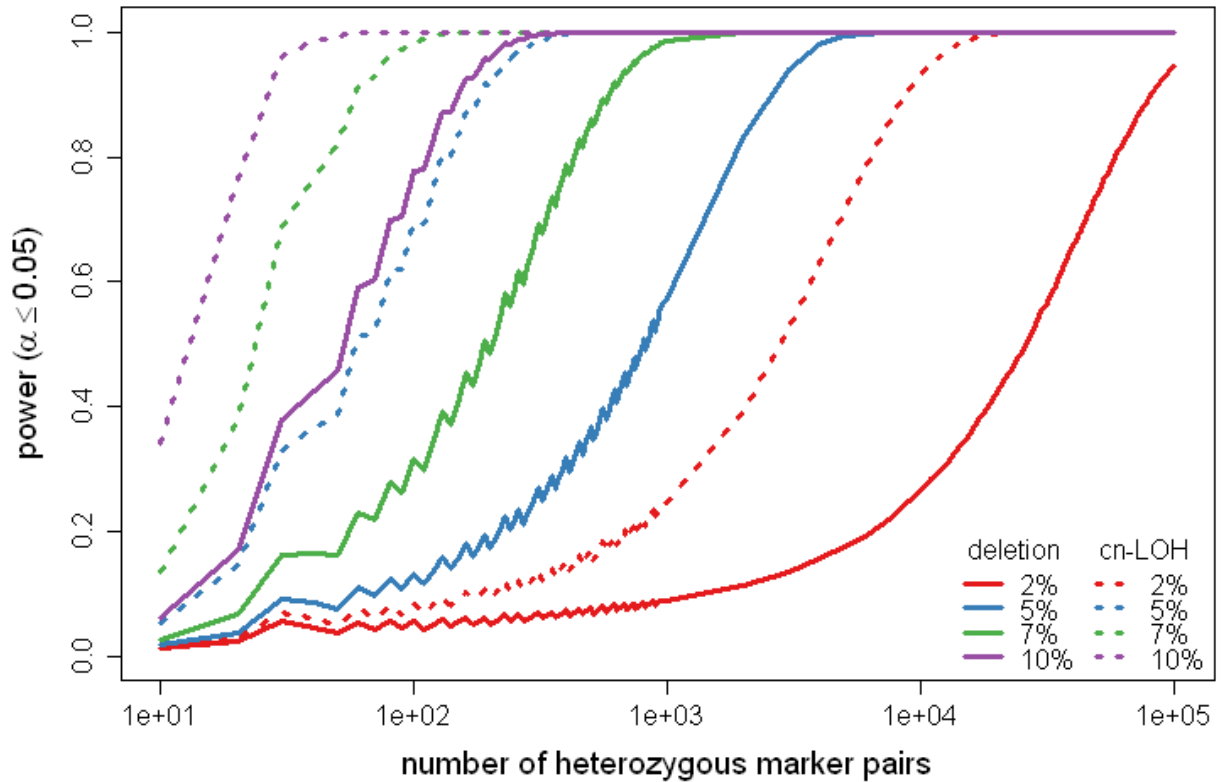
Appendix Figure 1. Power to detect allelic imbalance in a specific region of interest. We show power as a function of the number of informative sites in the tested region, assuming the entire tested region was affected by AI, for hemizygous deletion and NCLOH at four tumor proportions. Power will be lower if imbalance only affects a portion of the tested region. The results were calculated using the phase concordance rate per event type and tumor proportion observed from the simulated CRL-2324 data described in Chapter 3 of the main text (p.43). The maximum false positive rate was set to 5%; the sawtooth appearance of the curve at the lower marker counts is due to fluctuation in the actual false positive rate due to the discreteness of the binomial probability distribution.
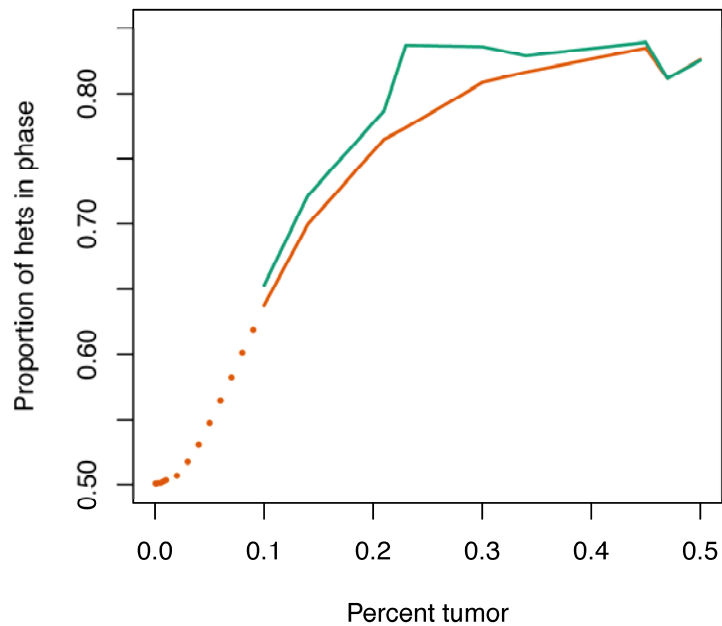
Appendix Figure 2. Calibration of computations simulations of BAFs. The genomewide phase concordance (using the heterozygous markers defined using the pure normal sample) are plotted against tumor proportion. At the proportions at which we have lab dilutions, the phase concordance from the lab dilution samples (green) and that from the simulated data (orange) show similar patterns. The simulated data are slightly conservative, indicating that perhaps more of the events are CNLOH than what we assumed for these dilutions.
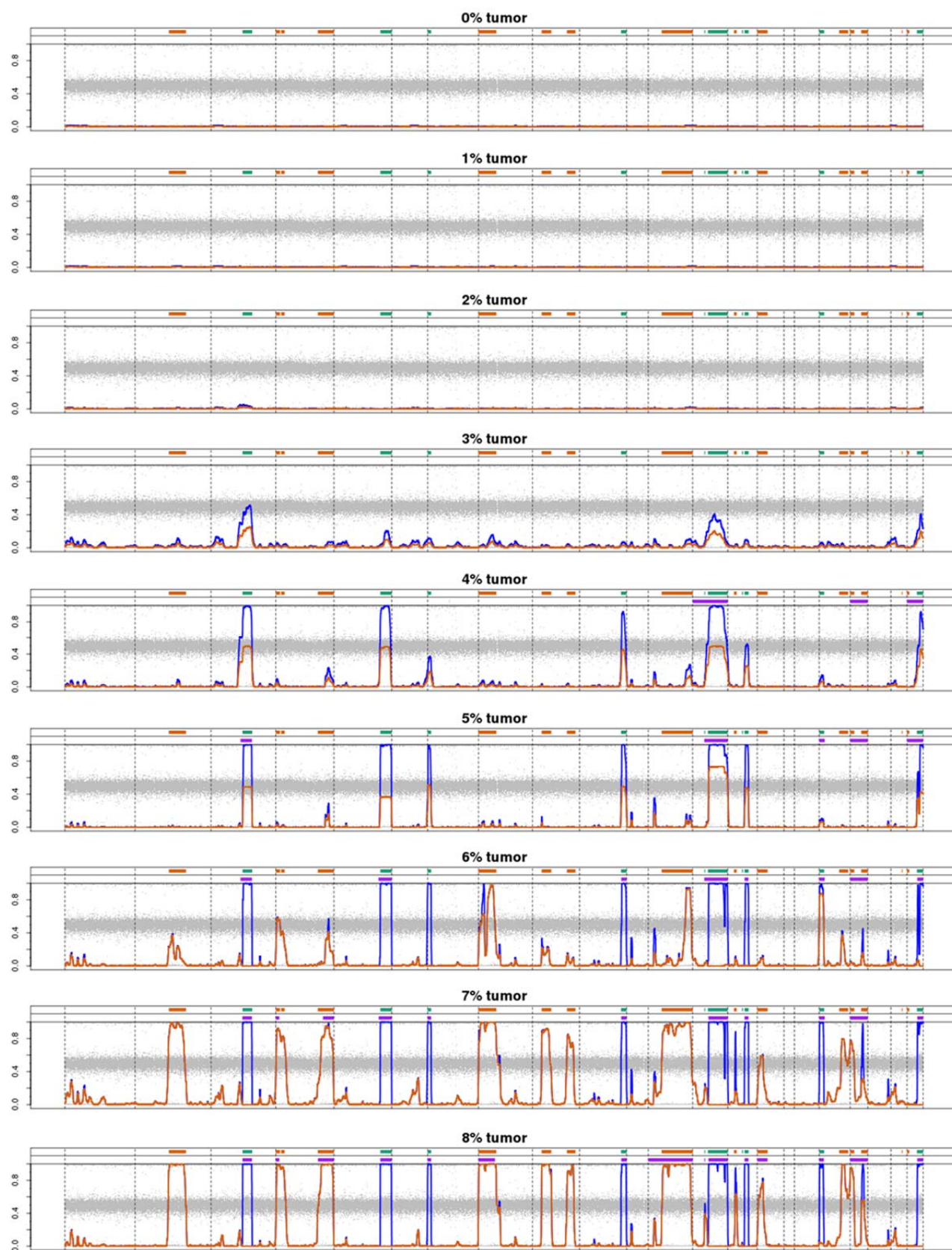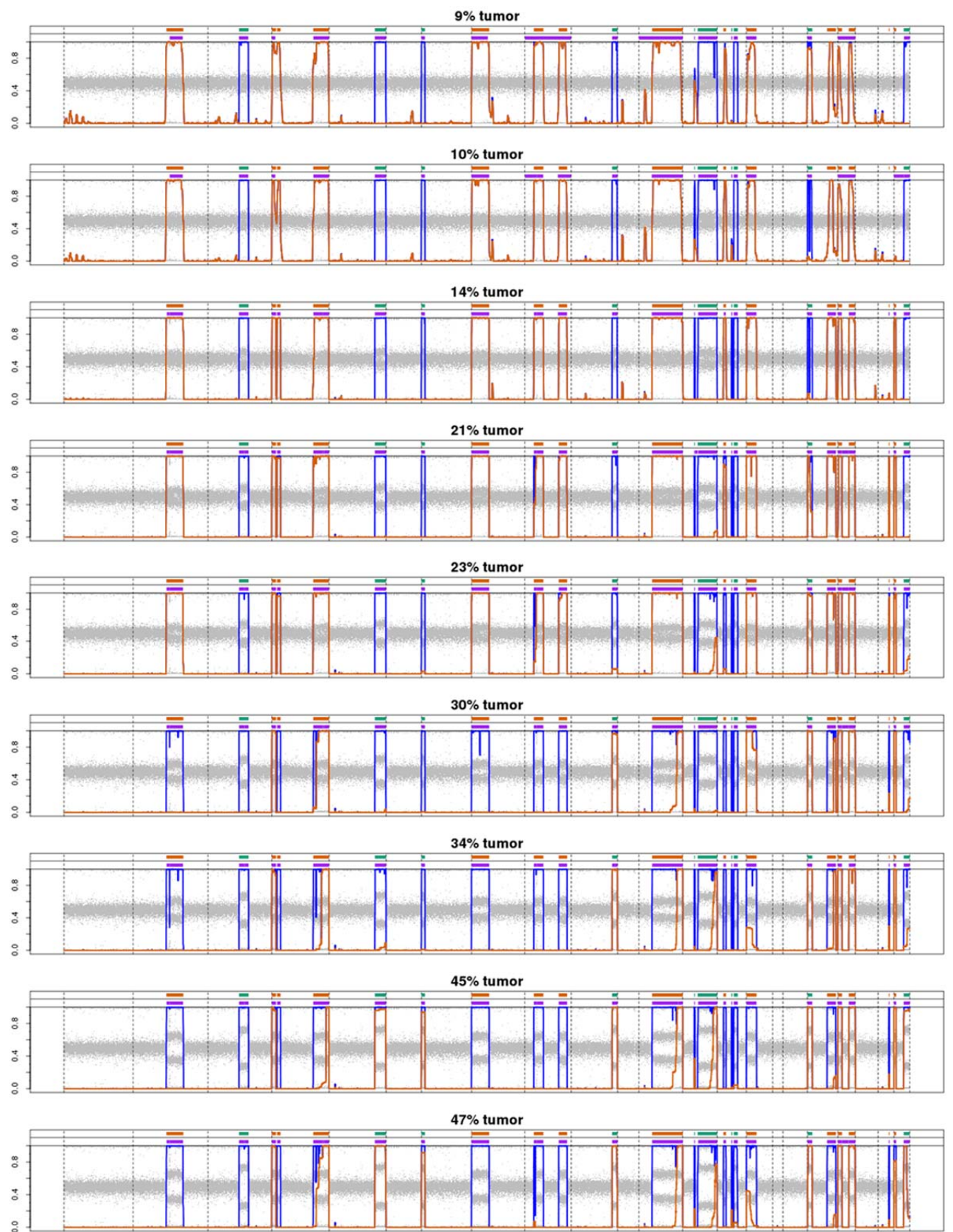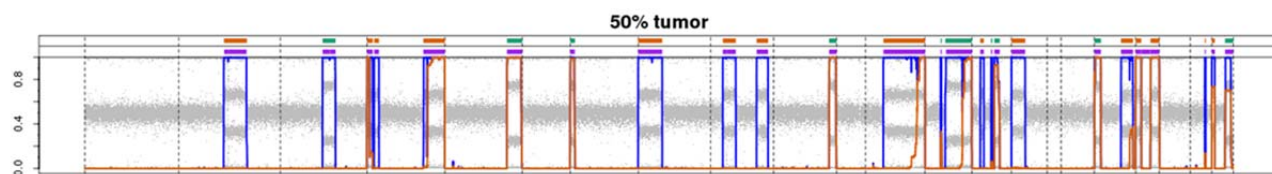
Appendix Figure 3. hapLOH results for computational dilutions of CRL-2324/CRL-2325. Vertical axes range from 0 to 1 for both the BAFs (grey points) and posterior probabilities (orange lines for probability of deletion, blue lines for summed probability of deletion or CNLOH). Horizontal lines at the top of each plot show the locations of simulated deletions (orange) and CNLOH (green). Below these, purple bars show the regions identified by BAFsegmentation to contain AI.
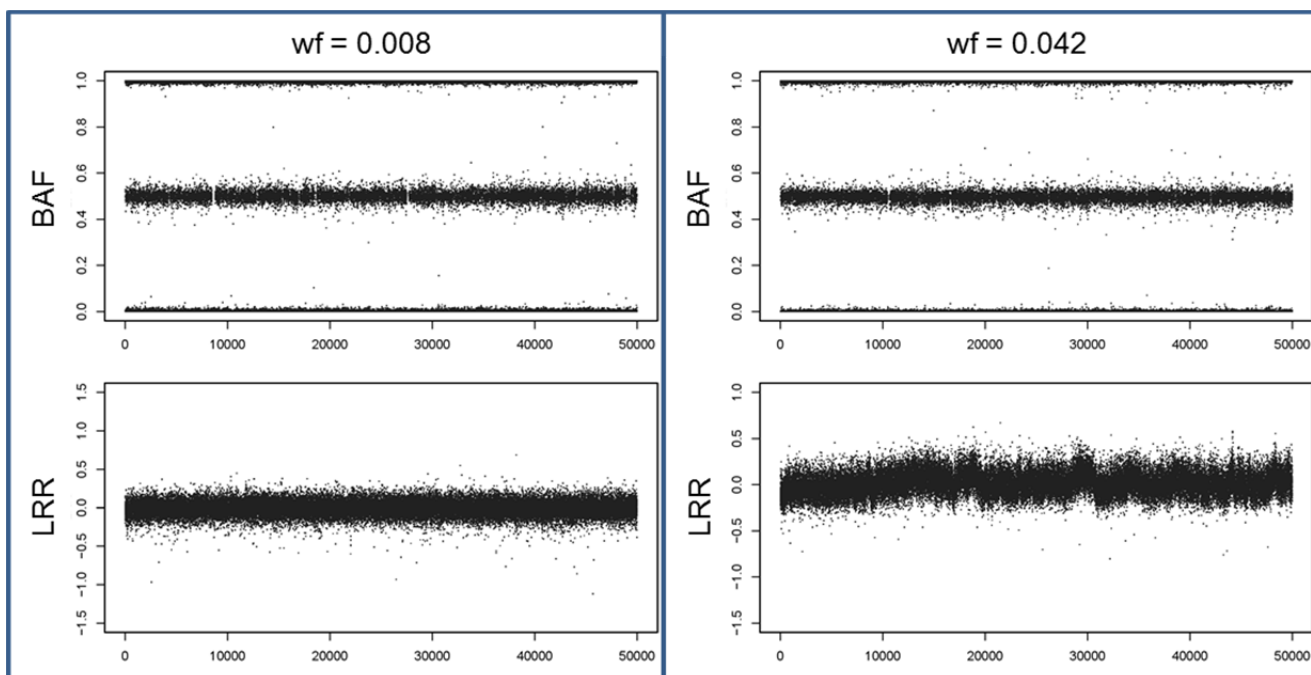
Appendix Figure 4. Genomic waves in LRRs. Left panel shows BAFs (top) and LRRs (bottom) for a 50,000 marker region from a sample that passed the waviness filter; right panel shows the same from a sample that failed the waviness filter.
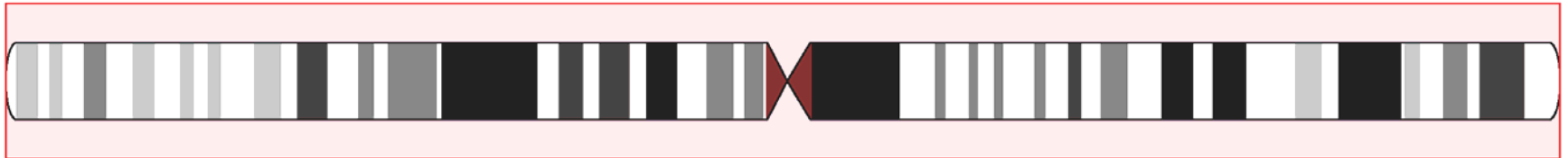
Chromosome 1
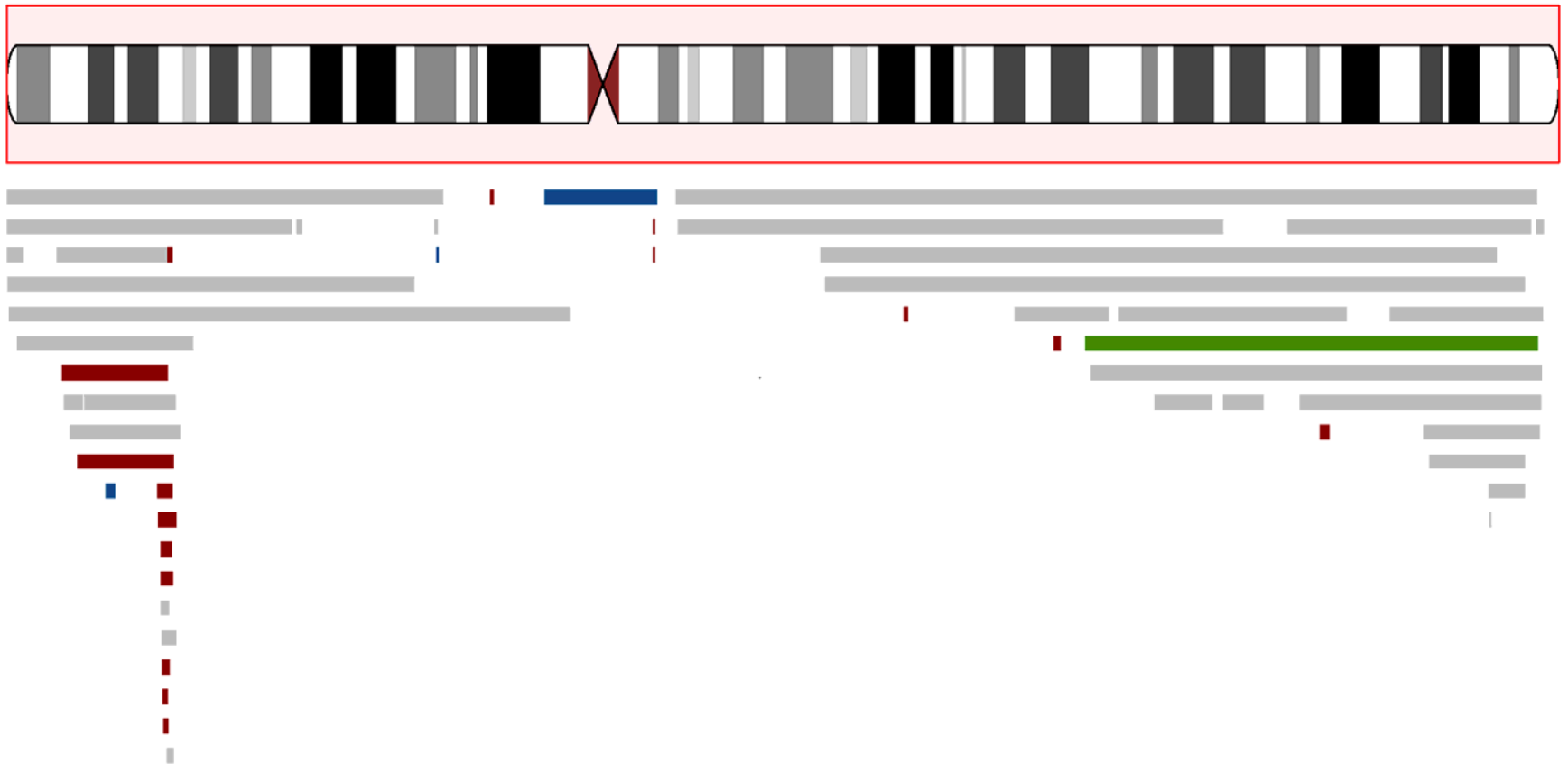
Chromosome 2

Chromosome 3

104
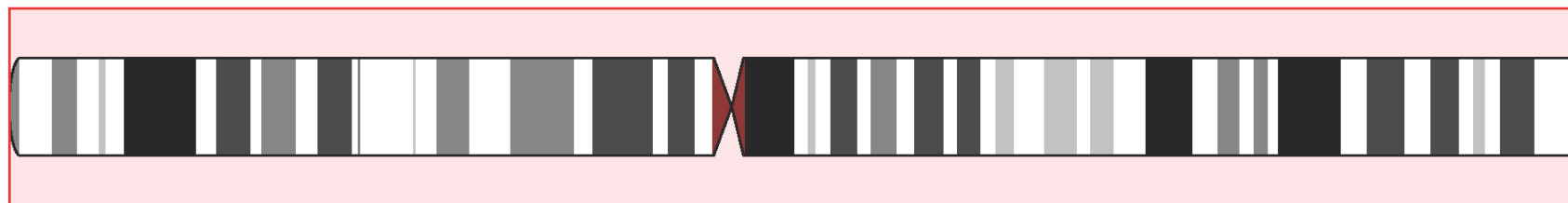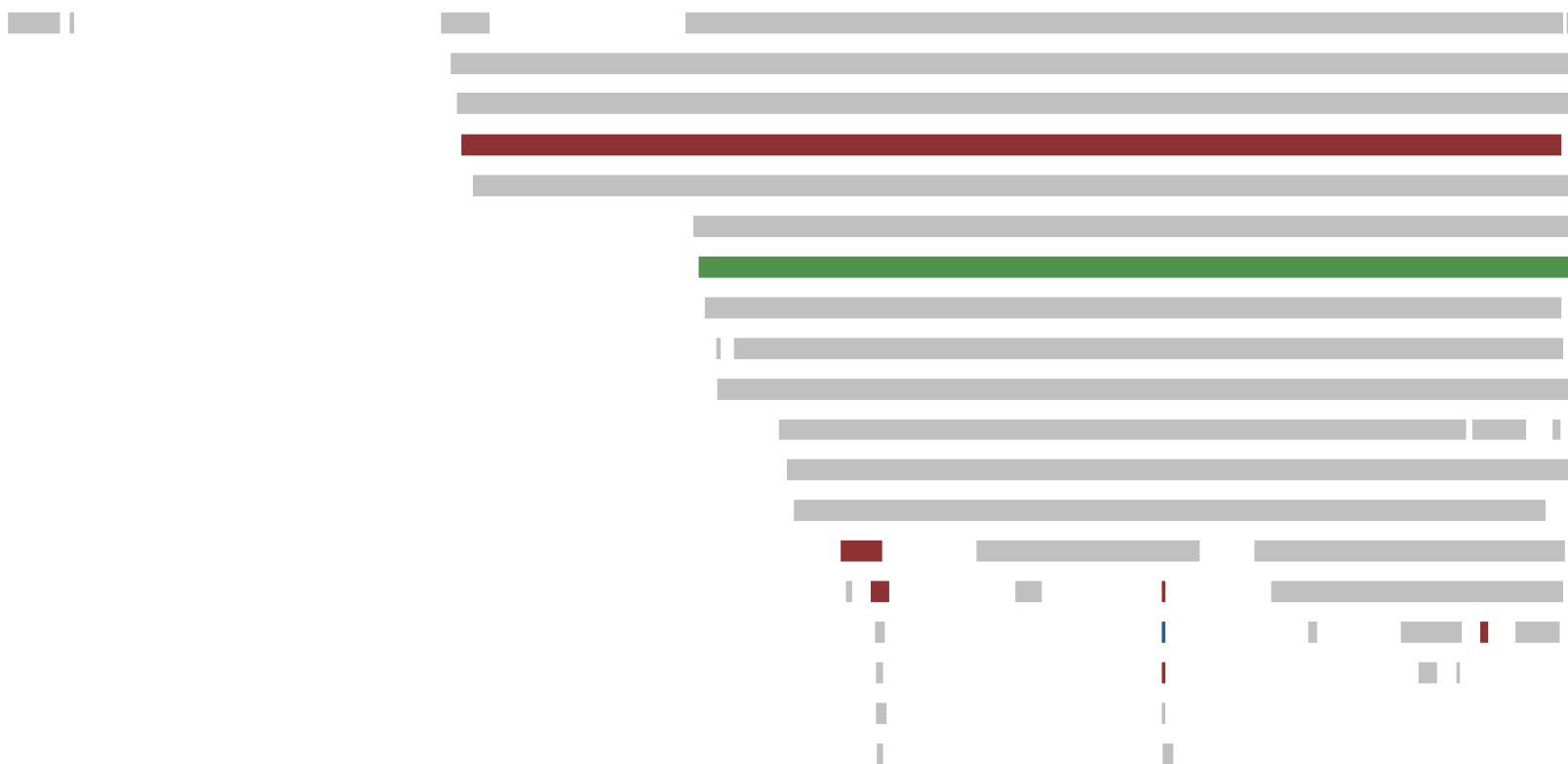
Chromosome 4

105
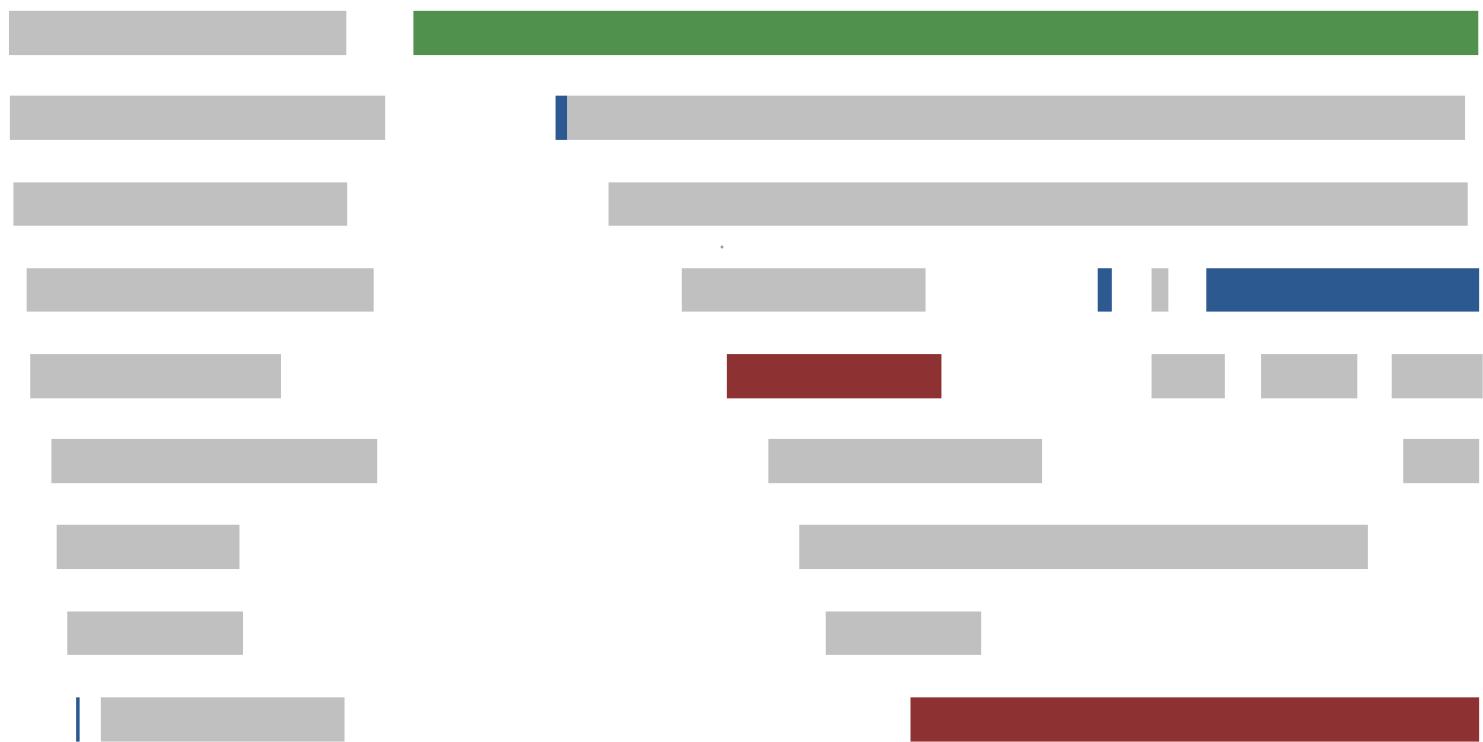
Chromosome 5
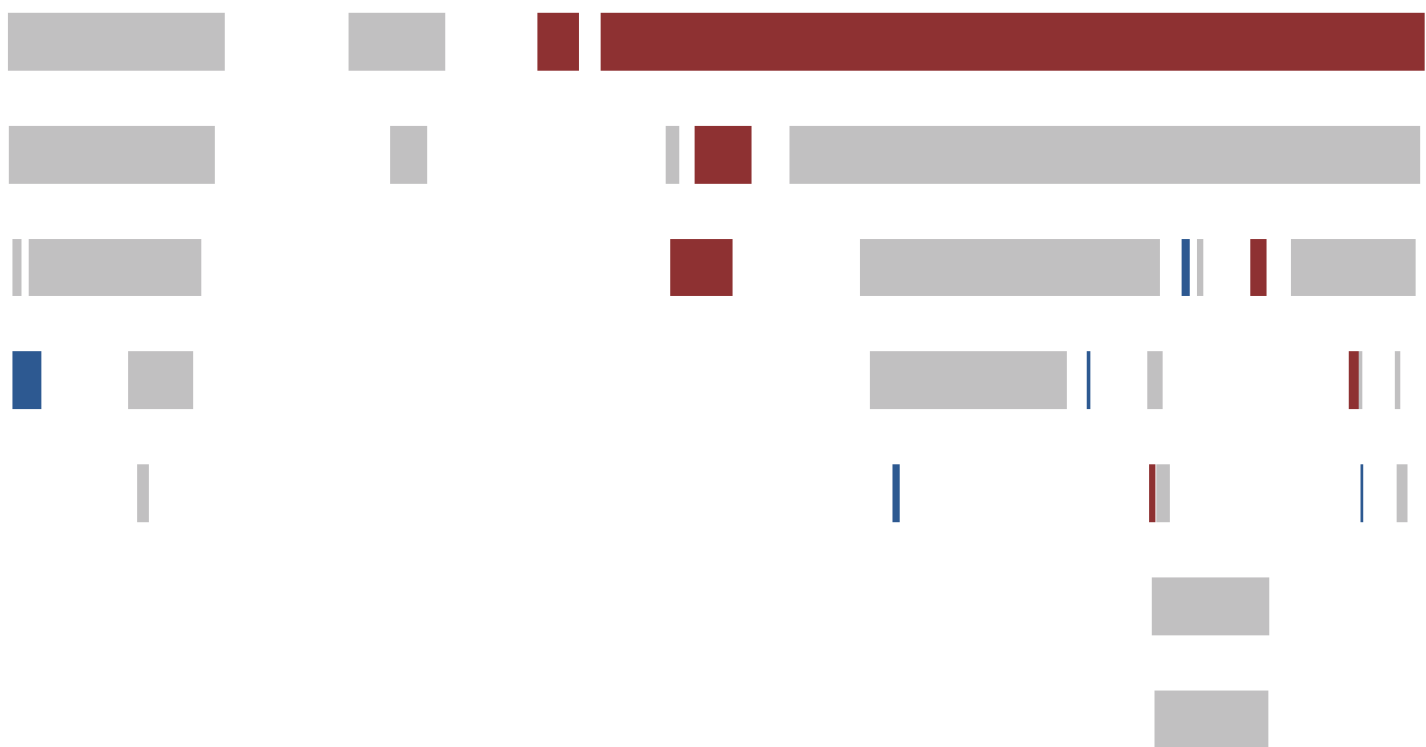
Chromosome 6

107

Chromosome 7

108
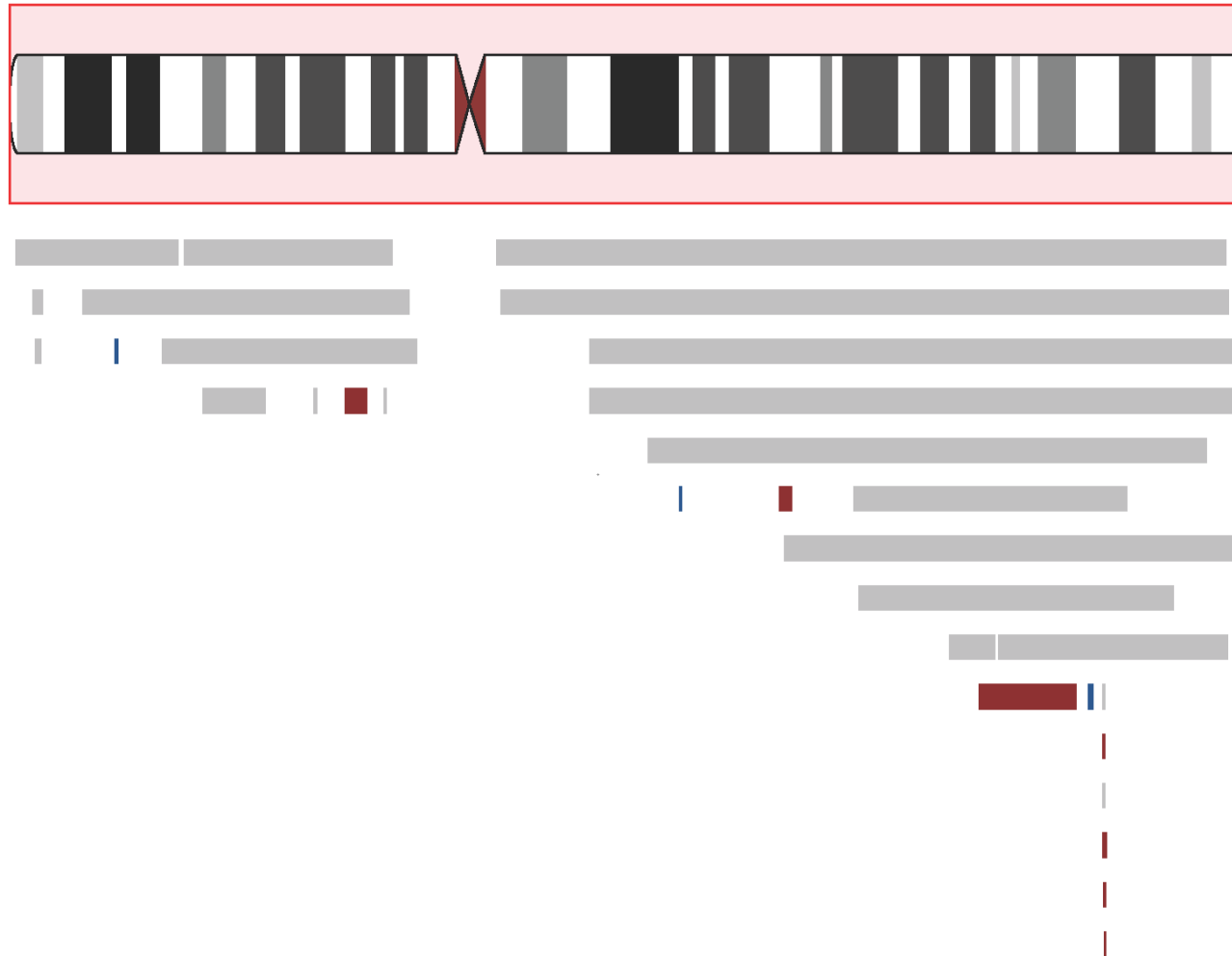
Chromosome 8

Chromosome 9

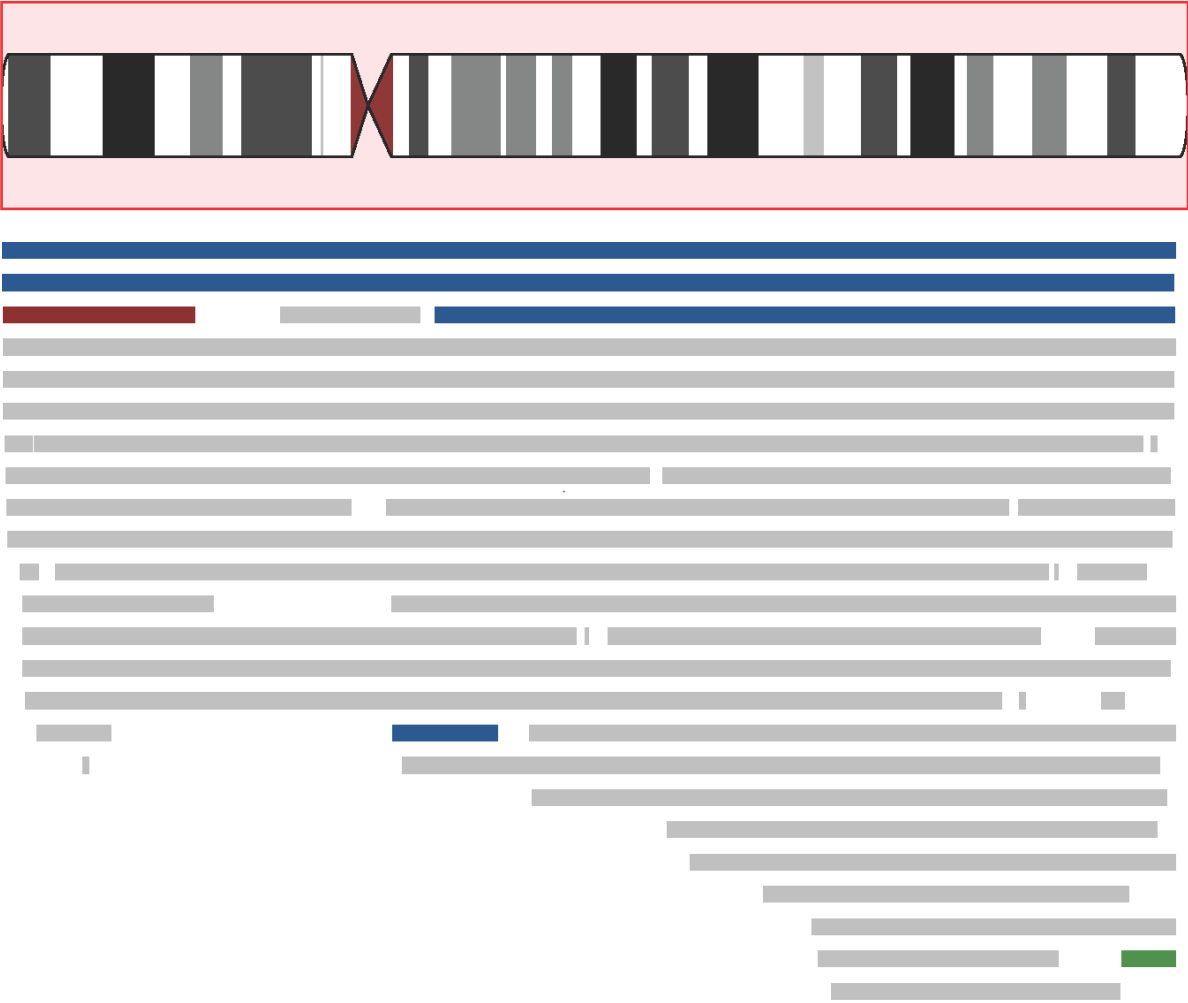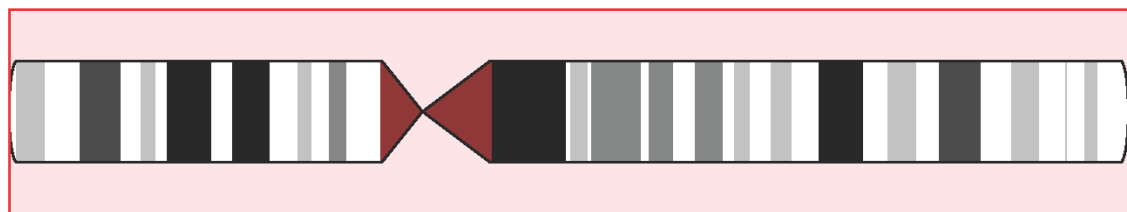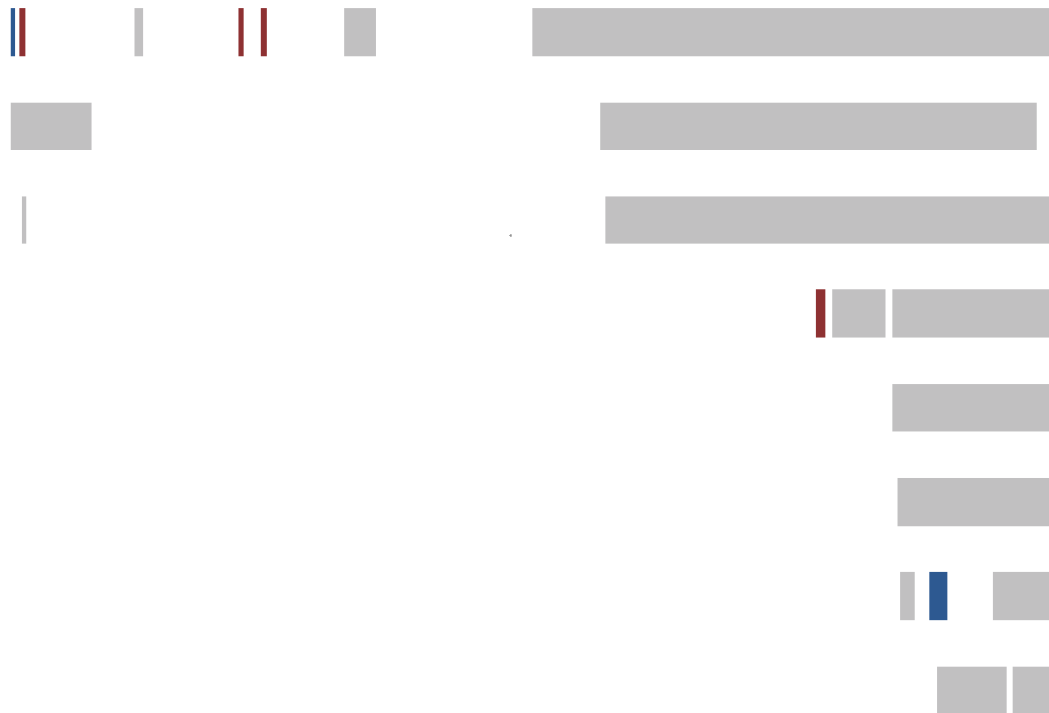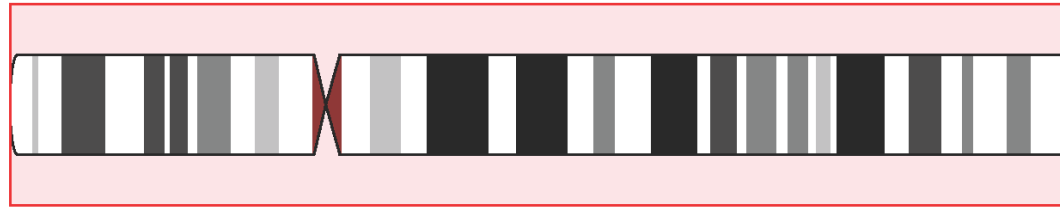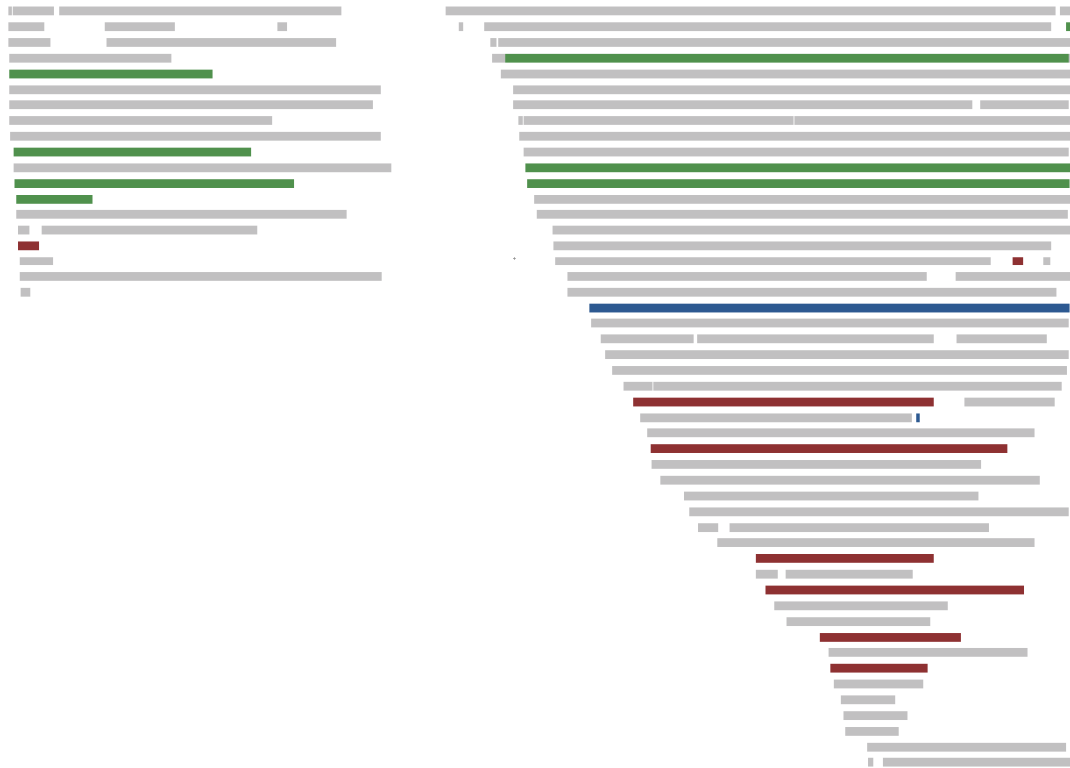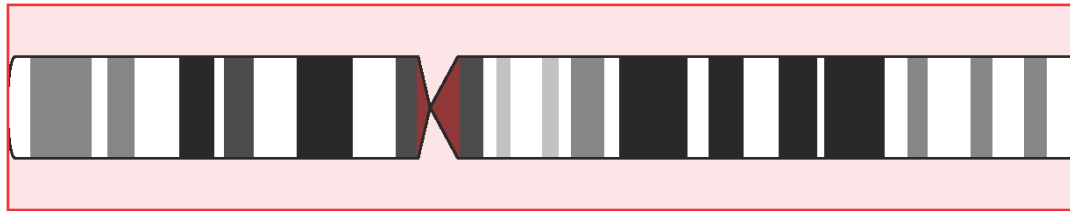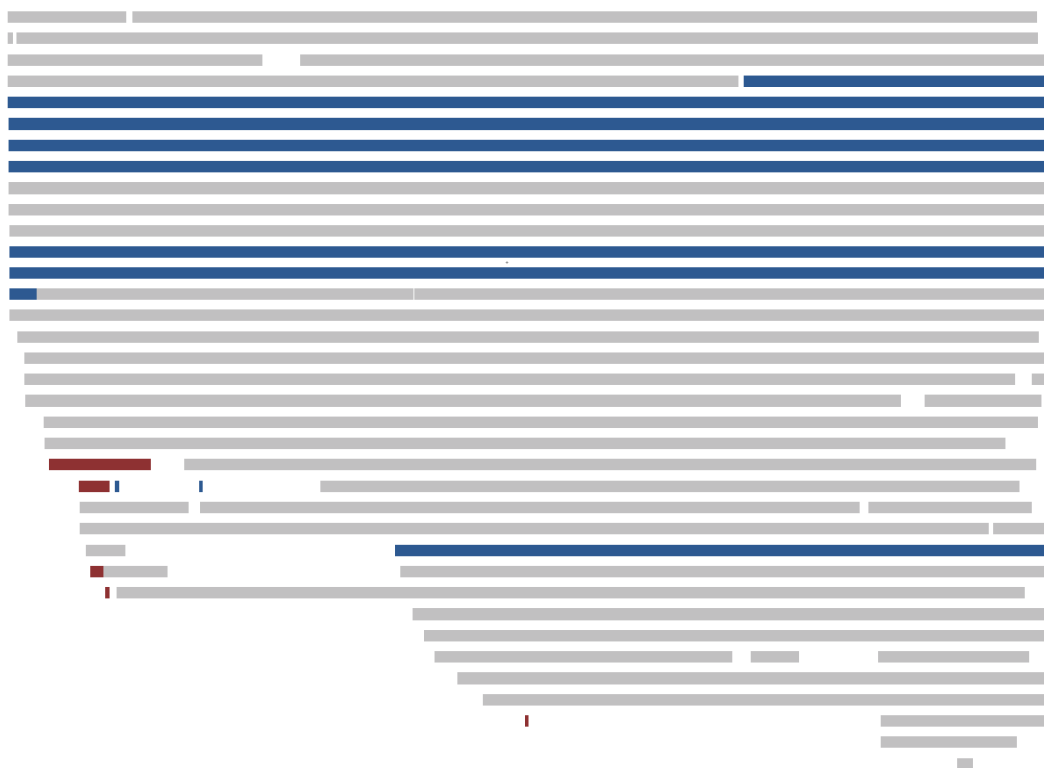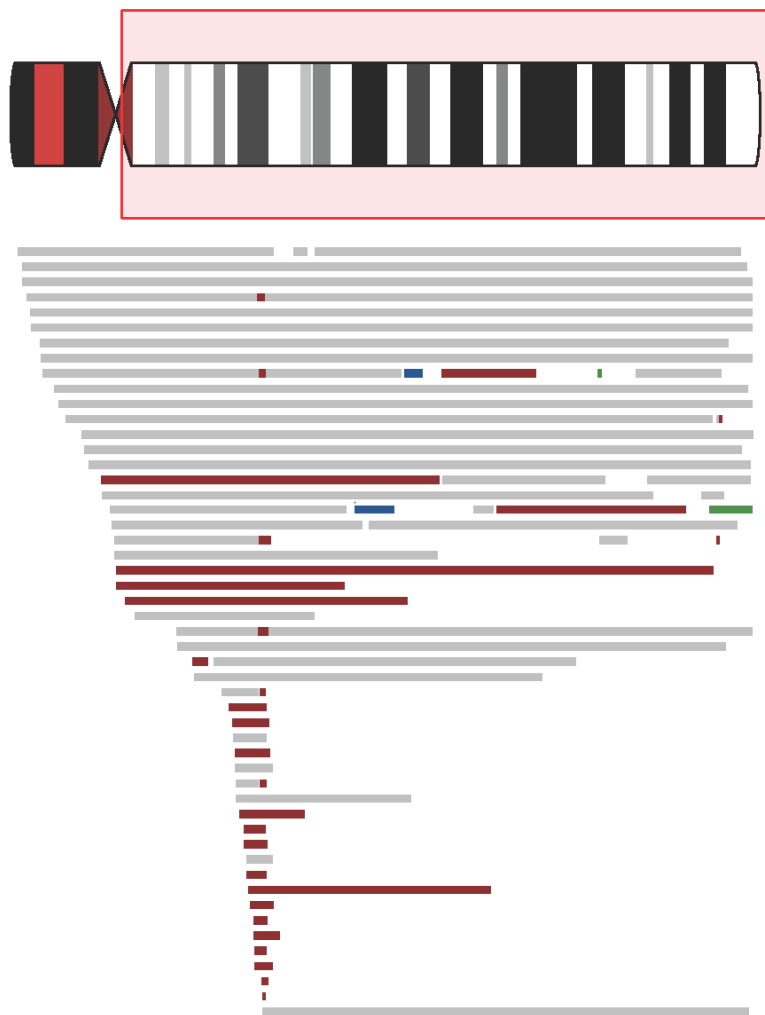Chromosome 10

Chromosome 11

Chromosome 12

113

Chromosome 13

Chromosome 14

Chromosome 15

Chromosome 16

Chromosome 17
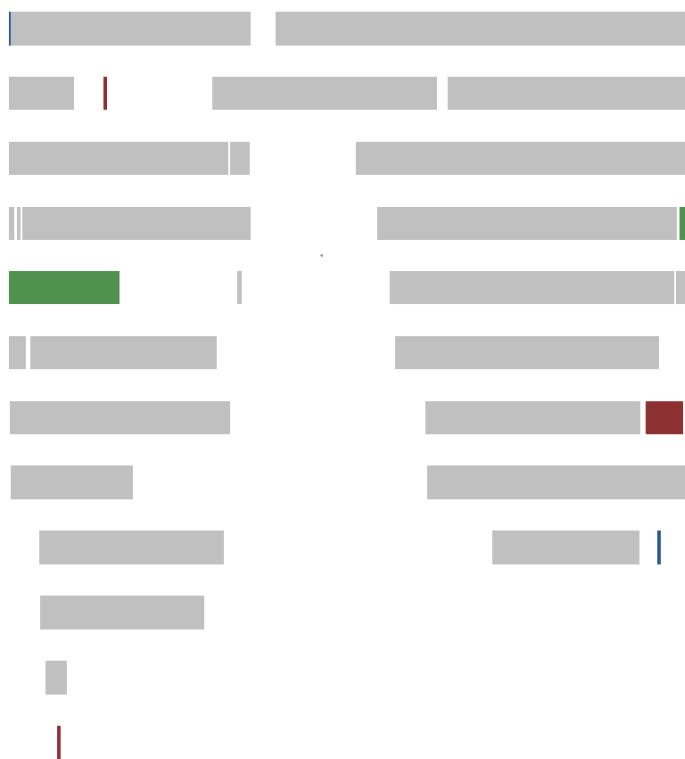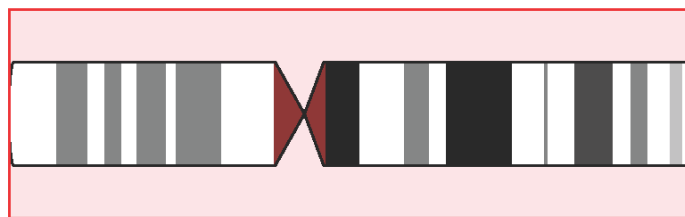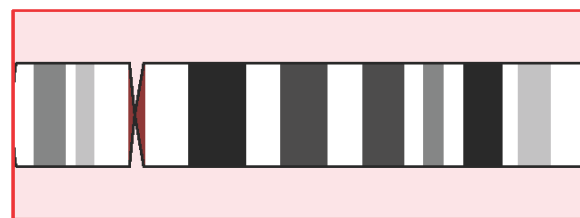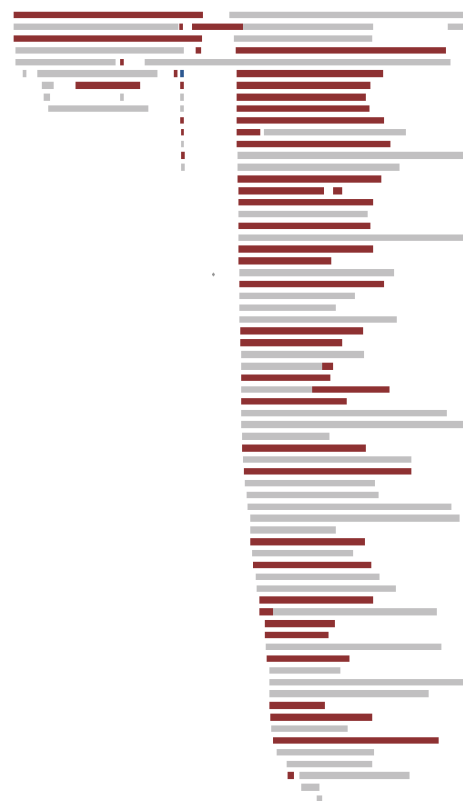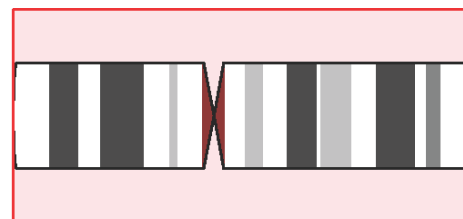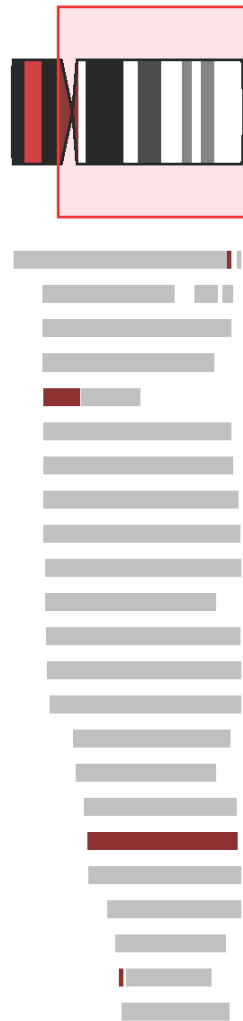
Chromosome 18
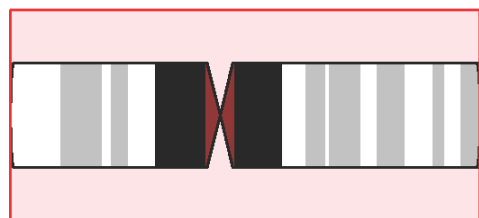
Chromosome 19

Chromosome 20

Chromosome 21

Chromosome 22

Appendix Figure 5. Per-chromosome plots of mosaic event calls in the GENEVA analysis. The red shading on each chromosome ideogram indicates the region represented in the plot space below the ideogram. Calls are represented by horizontal bars, colored by event type: red-loss, green-CNLOH, blue-gain, grey-undetermined.

Bibliography

1.   Bianconi, E., A. Piovesan, F. Facchin, A. Beraudi, R. Casadei, F. Frabetti, L. Vitale, M. C. Pelleri, S. Tassani, F. Piva, S. Perez-Amodio, P. Strippoli, and S. Canaider. 2013. An estimation of the number of cells in the human body. Annals of Human Biology 40:463-471.

2.   Lynch, M. 2010. Rate, molecular spectrum, and consequences of human mutation. Proceedings of the National Academy of Sciences of the United States of America 107:961-968.

3.   Gordon, D. J., B. Resio, and D. Pellman. 2012. Causes and consequences of aneuploidy in cancer. Nature Reviews Genetics 13:189-203.

4.   Robinson, W. P. 2000. Mechanisms leading to uniparental disomy and their clinical consequences. Bioessays 22:452-459.

5.   Lau, A. W., C. J. Brown, M. Penaherrera, S. Langlois, D. K. Kalousek, and W. P. Robinson. 1997. Skewed X-chromosome inactivation is common in fetuses or newborns associated with confined placental mosaicism. American Journal of Human Genetics 61:1353-1361.

6.   Kasparek, T. R., and T. C. Humphrey. 2011. DNA double-strand break repair pathways, chromosomal rearrangements and cancer. Seminars in Cell & Developmental Biology 22:886-897.

7.   Youssoufian, H., and R. E. Pyeritz. 2002. Mechanisms and consequences of somatic mosaicism in humans. Nature Reviews Genetics 3:748-758.

8.   Shao, C. S., P. J. Stambrook, and J. A. Tischfield. 2001. Mitotic recombination is suppressed by chromosomal divergence in hybrids of distantly related mouse strains. Nature Genetics 28:169-172.

9.   Arlt, M. F., T. E. Wilson, and T. W. Glover. 2012. Replication stress and mechanisms of CNV formation. Current Opinion in Genetics & Development 22:204-210.

10.   Piotrowski, A., C. E. G. Bruder, R. Andersson, T. D. de Stahl, U. Menzel, J. Sandgren, A. Poplawski, D. von Tell, C. Crasto, A. Bogdan, R. Bartoszewski, Z. Bebok, M. Krzyzanowski, Z. Jankowski, E. C. Partridge, J. Komorowski, and J. P. Dumanski. 2008. Somatic mosaicism for copy number variation in differentiated human tissues. Human Mutation 29:1118-1124.

11.   Biesecker, L. G., and N. B. Spinner. 2013. A genomic view of mosaicism and human disease. Nature Reviews Genetics 14:307-320.

12.   Davis, B. R., and F. Candotti. 2009. Revertant somatic mosaicism in the Wiskott-Aldrich syndrome. Immunologic Research 44:127-131.

13.   Holland, A. J., and D. W. Cleveland. 2009. Boveri revisited: chromosomal instability, aneuploidy and tumorigenesis. Nature Reviews Molecular Cell Biology 10:478-487.

14.   Silkworth, W. T., and D. Cimini. 2012. Transient defects of mitotic spindle geometry and chromosome segregation errors. Cell Division 7.

15.   Alonso Guerrero, A., C. Martinez-A, and K. H. M. van Wely. 2010. Merotelic attachments and non-homologous end joining are the basis of chromosomal instability. Cell Division 5.

16.   Bailey, S. M., and J. P. Murnane. 2006. Telomeres, chromosome instability and cancer. Nucleic Acids Research 34:2408-2417.

17.   Ozery-Flato, M., C. Linhart, L. Trakhtenbrot, S. Izraeli, and R. Shamir. 2011. Large-scale analysis of chromosomal aberrations in cancer karyotypes reveals two distinct paths to aneuploidy. Genome Biology 12:R61.

18.   Hafner, C., and L. Groesser. 2013. Mosaic RASopathies. Cell Cycle 12:43-50.

19.   Traupe, H. 1999. Functional X-chromosomal mosaicism of the skin: Rudolf Happle and the lines of Alfred Blaschko. American Journal of Medical Genetics 85:324-329.

20.   Proukakis, C., H. Houlden, and A. H. Schapira. 2013. Somatic alpha-synuclein mutations in Parkinson's disease: Hypothesis and preliminary data. Movement Disorders 28:705-712.

21. McConnell, M. J., M. R. Lindberg, K. J. Brennand, J. C. Piper, T. Voet, C. Cowing-Zitron, S. Shumilina, R. S. Lasken, J. R. Vermeesch, I. M. Hall, and F. H. Gage. 2013. Mosaic Copy Number Variation in Human Neurons. Science 342:632-637.

22. Singer, T., M. J. McConnell, M. C. N. Marchetto, N. G. Coufal, and F. H. Gage. 2010. LINE-1 retrotransposons: mediators of somatic variation in neuronal genomes? Trends in Neurosciences 33:345-354.

23. Aviv, A., and H. Aviv. 1998. Telomeres, hidden mosaicism, loss of heterozygosity, and complex genetic traits. Human Genetics 103:2-4.

24. Tomasetti, C., B. Vogelstein, and G. Parmigiani. 2013. Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. Proceedings of the National Academy of Sciences of the United States of America 110:1999-2004.

25. Laurie, C. C., C. A. Laurie, K. Rice, K. F. Doheny, L. R. Zelnick, C. P. McHugh, H. Ling, K. N. Hetrick, E. W. Pugh, C. Amos, Q. Wei, L.-E. Wang, J. E. Lee, K. C. Barnes, N. N. Hansel, R. Mathias, D. Daley, T. H. Beaty, A. F. Scott, I. Ruczinski, R. B. Scharpf, L. J. Bierut, S. M. Hartz, M. T. Landi, N. D. Freedman, L. R. Goldin, D. Ginsburg, J. Li, K. C. Desch, S. S. Strom, W. J. Blot, L. B. Signorello, S. A. Ingles, S. J. Chanock, S. I. Berndt, L. Le Marchand, B. E. Henderson, K. R. Monroe, J. A. Heit, M. de Andrade, S. M. Armasu, C. Regnier, W. L. Lowe, M. G. Hayes, M. L. Marazita, E. Feingold, J. C. Murray, M. Melbye, B. Feenstra, J. H. Kang, J. L. Wiggs, G. P. Jarvik, A. N. McDavid, V. E. Seshan, D. B. Mirel, A. Crenshaw, N. Sharopova, A. Wise, J. Shen, D. R. Crosslin, D. M. Levine, X. Zheng, J. I. Udren, S. Bennett, S. C. Nelson, S. M. Gogarten, M. P. Conomos, P. Heagerty, T. Manolio, L. R. Pasquale, C. A. Haiman, N. Caporaso, and B. S. Weir. 2012. Detectable clonal mosaicism from birth to old age and its relationship to cancer. Nature Genetics 44:642-650.

26. Jacobs, K. B., M. Yeager, W. Zhou, S. Wacholder, Z. Wang, B. Rodriguez-Santiago, A. Hutchinson, X. Deng, C. Liu, M.-J. Horner, M. Cullen, C. G. Epstein, L. Burdett, M. C. Dean, N.

Chatterjee, J. Sampson, C. C. Chung, J. Kovaks, S. M. Gapstur, V. L. Stevens, L. T. Teras, M. M. Gaudet, D. Albanes, S. J. Weinstein, J. Virtamo, P. R. Taylor, N. D. Freedman, C. C. Abnet, A. M. Goldstein, N. Hu, K. Yu, J.-M. Yuan, L. Liao, T. Ding, Y.-L. Qiao, Y.-T. Gao, W.-P. Koh, Y.-B. Xiang, Z.-Z. Tang, J.-H. Fan, M. C. Aldrich, C. Amos, W. J. Blot, C. H. Bock, E. M. Gillanders, C. C. Harris, C. A. Haiman, B. E. Henderson, L. N. Kolonel, L. Le Marchand, L. H. McNeill, B. A. Rybicki, A. G. Schwartz, L. B. Signorello, M. R. Spitz, J. K. Wiencke, M. Wrensch, X. Wu, K. A. Zanetti, R. G. Ziegler, J. D. Figueroa, M. Garcia-Closas, N. Malats, G. Marenne, L. Prokunina-Olsson, D. Baris, M. Schwenn, A. Johnson, M. T. Landi, L. Goldin, D. Consonni, P. A. Bertazzi, M. Rotunno, P. Rajaraman, U. Andersson, L. E. B. Freeman, C. D. Berg, J. E. Buring, M. A. Butler, T. Carreon, M. Feychting, A. Ahlbom, J. M. Gaziano, G. G. Giles, G. Hallmans, S. E. Hankinson, P. Hartge, R. Henriksson, P. D. Inskip, C. Johansen, A. Landgren, R. McKean-Cowdin, D. S. Michaud, B. S. Melin, U. Peters, A. M. Ruder, H. D. Sesso, G. Severi, X.-O. Shu, K. Visvanathan, E. White, A. Wolk, A. Zeleniuch-Jacquotte, W. Zheng, D. T. Silverman, M. Kogevinas, J. R. Gonzalez, O. Villa, D. Li, E. J. Duell, H. A. Risch, S. H. Olson, C. Kooperberg, B. M. Wolpin, L. Jiao, M. Hassan, W. Wheeler, A. A. Arslan, H. B. Bueno-de-Mesquita, C. S. Fuchs, S. Gallinger, M. D. Gross, E. A. Holly, A. P. Klein, A. LaCroix, M. T. Mandelson, G. Petersen, M.-C. Boutron-Ruault, P. M. Bracci, F. Canzian, K. Chang, M. Cotterchio, E. L. Giovannucci, M. Goggins, J. A. H. Bolton, M. Jenab, K.-T. Khaw, V. Krogh, R. C. Kurtz, R. R. McWilliams, J. B. Mendelsohn, K. G. Rabe, E. Riboli, A. Tjonneland, G. S. Tobias, D. Trichopoulos, J. W. Elena, H. Yu, L. Amundadottir, R. Z. Stolzenberg-Solomon, P. Kraft, F. Schumacher, D. Stram, S. A. Savage, L. Mirabello, I. L. Andrulis, J. S. Wunder, A. Patino Garcia, L. Sierrasesumaga, D. A. Barkauskas, R. G. Gorlick, M. Purdue, W.-H. Chow, L. E. Moore, K. L. Schwartz, F. G. Davis, A. W. Hsing, S. I. Berndt, A. Black, N. Wentzensen, L. A. Brinton, J. Lissowska, B. Peplonska, K. A. McGlynn, M. B. Cook, B. I. Graubard, C. P. Kratz, M. H. Greene, R. L. Erickson, D. J. Hunter, G. Thomas, R. N.

123

Hoover, F. X. Real, J. F. Fraumeni, Jr., N. E. Caporaso, M. Tucker, N. Rothman, L. A. Perez-Jurado, and S. J. Chanock. 2012. Detectable clonal mosaicism and its relationship to aging and cancer. Nature Genetics 44:651-658.

27.   Forsberg, L. A., C. Rasi, H. R. Razzaghian, G. Pakalapati, L. Waite, K. S. Thilbeault, A. Ronowicz, N. E. Wineinger, H. K. Tiwari, D. Boomsma, M. P. Westerman, J. R. Harris, R. Lyle, M. Essand, F. Eriksson, T. L. Assimes, C. Iribarren, E. Strachan, T. P. O'Hanlon, L. G. Rider, F. W. Miller, V. Giedraitis, L. Lannfelt, M. Ingelsson, A. Piotrowski, N. L. Pedersen, D. Absher, and J. P. Dumanski. 2012. Age-Related Somatic Structural Changes in the Nuclear Genome of Human Blood Cells. American Journal of Human Genetics 90:217-228.

28.   Davis, B. R., and F. Candotti. 2010. Mosaicism-Switch or Spectrum? Science 330:46-47.

29.   Conlin, L. K., B. D. Thiel, C. G. Bonnemann, L. Medne, L. M. Ernst, E. H. Zackai, M. A. Deardorff, I. D. Krantz, H. Hakonarson, and N. B. Spinner. 2010. Mechanisms of mosaicism, chimerism and uniparental disomy identified by single nucleotide polymorphism array analysis. Human Molecular Genetics 19:1263-1275.

30.   Hook, E. B. 1977. Exclusion of chromosomal mosaicism: tables of 90%, 95% and 99% confidence limits and comments on use. American Journal of Human Genetics 29:94-97.

31.   Pham, J., C. Shaw, P. Hixson, A. Ester, A. Pursley, S. S. H. Kang, W. Bi, S. Lanani, C. Bacino, P. Stankiewicz, A. Patel, and S. W. Cheung. 2012. Somatic Mosaicism Detected by Exon-Targeted, High-Resolution aCGH in 10,362 Consecutive Cases. Cytogenetic and Genome Research 136:333-333.

32.   Vandeweyer, G., and R. F. Kooy. 2013. Detection and interpretation of genomic structural variation in health and disease. Expert Review of Molecular Diagnostics 13:61-82.

33.   TaqMan Copy Number Assays. In Product Bulletin. Life Technologies.

34.   Eijk-Van Os, P. G. C., and J. P. Schouten. 2011. Multiplex Ligation-dependent Probe
        Amplification (MLPA) for the detection of copy number variation in genomic sequences.
        Methods in molecular biology (Clifton, N.J.) 688:97-126.

35.   Illumina. "TOP/BOT" Strand and "A/B" Allele.

36.   Peiffer, D. A., J. M. Le, F. J. Steemers, W. Chang, T. Jenniges, F. Garcia, K. Haden, J. Li, C. A.
        Shaw, J. Belmont, S. W. Cheung, R. M. Shen, D. L. Barker, and K. L. Gunderson. 2006. High-
        resolution genomic profiling of chromosomal aberrations using Infinium whole-genome
        genotyping. Genome Research 16:1136-1148.

37.   Olshen, A. B., E. S. Venkatraman, R. Lucito, and M. Wigler. 2004. Circular binary segmentation
        for the analysis of array-based DNA copy number data. Biostatistics 5:557-572.

38.   Xia, R., S. Vattathil, and P. Scheet. 2014. Identification of allelic imbalance with a statistical
        model for subtle genomic mosaicism. PLOS Computational Biology:(in press).

39.   Yau, C., D. Mouradov, R. N. Jorissen, S. Colella, G. Mirza, G. Steers, A. Harris, J. Ragoussis, O.
        Sieber, and C. C. Holmes. 2010. A statistical approach for detecting genomic aberrations in
        heterogeneous tumor samples from single nucleotide polymorphism genotyping data. Genome
        Biology 11.

40.   Li, A., Z. Liu, K. Lezon-Geyda, S. Sarkar, D. Lannin, V. Schulz, I. Krop, E. Winer, L. Harris,
        and D. Tuck. 2011. GPHMM: an integrated hidden Markov model for identification of copy
        number alteration and loss of heterozygosity in complex tumor samples using whole genome
        SNP arrays. Nucleic Acids Research 39:4928-4941.

41.   Chen, H., H. Xing, and N. R. Zhang. 2011. Estimation of Parent Specific DNA Copy Number in
        Tumors using High-Density Genotyping Arrays. Plos Computational Biology 7.

42.   Yamamoto, G., Y. Nannya, M. Kato, M. Sanada, R. L. Levine, N. Kawamata, A. Hangaishi, M.
        Kurokawa, S. Chiba, D. G. Gilliland, H. P. Koeffler, and S. Ogawa. 2007. Highly sensitive
        method for genomewide detection of allelic composition in nonpaired, primary tumor specimens

by use of affymetrix single-nucleotide-polymorphism genotyping microarrays. American Journal of Human Genetics 81:114-126.

43.   Assie, G., T. LaFramboise, P. Platzer, J. Bertherat, C. A. Stratakis, and C. Eng. 2008. SNP Arrays in heterogeneous tissue: Highly accurate collection of both germline and somatic genetic information from unpaired single tumor samples. American Journal of Human Genetics 82:903-915.

44.   Staaf, J., D. Lindgren, J. Vallon-Christersson, A. Isaksson, H. Goransson, G. Juliusson, R. Rosenquist, M. Hoglund, A. Borg, and M. Ringner. 2008. Segmentation-based detection of allelic imbalance and loss-of-heterozygosity in cancer cells using whole genome SNP arrays. Genome Biology 9.

45.   Sun, W., F. A. Wright, Z. Tang, S. H. Nordgard, P. Van Loo, T. Yu, V. N. Kristensen, and C. M. Perou. 2009. Integrated study of copy number states and genotype calls using high-density SNP arrays. Nucleic Acids Research 37:5365-5377.

46.   Popova, T., E. Manie, D. Stoppa-Lyonnet, G. Rigaill, E. Barillot, and M. H. Stern. 2009. Genome Alteration Print (GAP): a tool to visualize and mine complex cancer genomic profiles obtained by SNP arrays. Genome Biology 10.

47.   Rancoita, P. M. V., M. Hutter, F. Bertoni, and I. Kwee. 2010. An integrated Bayesian analysis of LOH and copy number data. Bmc Bioinformatics 11.

48.   Van Loo, P., S. H. Nordgard, O. C. Lingjaerde, H. G. Russnes, I. H. Rye, W. Sun, V. J. Weigman, P. Marynen, A. Zetterberg, B. Naume, C. M. Perou, A.-L. Borresen-Dale, and V. N. Kristensen. 2010. Allele-specific copy number analysis of tumors. Proceedings of the National Academy of Sciences of the United States of America 107:16910-16915.

49.   Gonzalez, J. R., B. Rodriguez-Santiago, A. Caceres, R. Pique-Regi, N. Rothman, S. J. Chanock, L. Armengol, and L. A. Perez-Jurado. 2011. A fast and accurate method to detect allelic genomic imbalances underlying mosaic rearrangements using SNP array data. Bmc Bioinformatics 12.

50.   Rasmussen, M., M. Sundstrom, H. G. Kultima, J. Botling, P. Micke, H. Birgisson, B. Glimelius, and A. Isaksson. 2011. Allele-specific copy number analysis of tumor samples with aneuploidy and tumor heterogeneity. Genome Biology 12.

51.   Carter, S. L., M. Meyerson, and G. Getz. 2011. Accurate estimation of homologue-specific DNA concentration-ratios in cancer samples allows long-range haplotyping. Nature Precedings.

52.   Nik-Zainal, S., P. Van Loo, D. C. Wedge, L. B. Alexandrov, C. D. Greenman, K. W. Lau, K. Raine, D. Jones, J. Marshall, M. Ramakrishna, A. Shlien, S. L. Cooke, J. Hinton, A. Menzies, L. A. Stebbings, C. Leroy, M. Jia, R. Rance, L. J. Mudie, S. J. Gamble, P. J. Stephens, S. McLaren, P. S. Tarpey, E. Papaemmanuil, H. R. Davies, I. Varela, D. J. McBride, G. R. Bignell, K. Leung, A. P. Butler, J. W. Teague, S. Martin, G. Joensson, O. Mariani, S. Boyault, P. Miron, A. Fatima, A. Langerod, S. A. J. R. Aparicio, A. Tutt, A. M. Sieuwerts, A. Borg, G. Thomas, A. V. Salomon, A. L. Richardson, A.-L. Borresen-Dale, P. A. Futreal, M. R. Stratton, P. J. Campbell, and C. Int Canc Genome. 2012. The Life History of 21 Breast Cancers. Cell 149.

53.   Baugher, J. D., B. D. Baugher, M. D. Shirley, and J. Pevsner. 2013. Sensitive and specific detection of mosaic chromosomal abnormalities using the Parent-of-Origin-based Detection (POD) method. Bmc Genomics 14.

54.   Browning, S. R., and B. L. Browning. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. American Journal of Human Genetics 81:1084-1097.

55.   Altshuler, D. L., R. M. Durbin, G. R. Abecasis, D. R. Bentley, A. Chakravarti, A. G. Clark, F. S. Collins, F. M. De la Vega, P. Donnelly, M. Egholm, P. Flicek, S. B. Gabriel, R. A. Gibbs, B. M. Knoppers, E. S. Lander, H. Lehrach, E. R. Mardis, G. A. McVean, D. Nickerson, L. Peltonen, A. J. Schafer, S. T. Sherry, J. Wang, R. K. Wilson, D. Deiros, M. Metzker, D. Muzny, J. Reid, D. Wheeler, J. Li, M. Jian, G. Li, R. Li, H. Liang, G. Tian, B. Wang, J. Wang, W. Wang, H. Yang, X. Zhang, H. Zheng, L. Ambrogio, T. Bloom, K. Cibulskis, T. J. Fennell, D. B. Jaffe, E. Shefler,

C. L. Sougnez, N. Gormley, S. Humphray, Z. Kingsbury, P. Koko-Gonzales, J. Stone, K. J. McKernan, G. L. Costa, J. K. Ichikawa, C. C. Lee, R. Sudbrak, T. A. Borodina, A. Dahl, A. N. Davydov, P. Marquardt, F. Mertes, W. Nietfeld, P. Rosenstiel, S. Schreiber, A. V. Soldatov, B. Timmermann, M. Tolzmann, J. Affourtit, D. Ashworth, S. Attiya, M. Bachorski, E. Buglione, A. Burke, A. Caprio, C. Celone, S. Clark, D. Conners, B. Desany, L. Gu, L. Guccione, K. Kao, A. Kebbel, J. Knowlton, M. Labrecque, L. McDade, C. Mealmaker, M. Minderman, A. Nawrocki, F. Niazi, K. Pareja, R. Ramenani, D. Riches, W. Song, C. Turcotte, S. Wang, D. Dooling, L. Fulton, R. Fulton, G. Weinstock, J. Burton, D. M. Carter, C. Churcher, A. Coffey, A. Cox, A. Palotie, M. Quail, T. Skelly, J. Stalker, H. P. Swerdlow, D. Turner, A. De Witte, S. Giles, M. Bainbridge, D. Challis, A. Sabo, F. Yu, J. Yu, X. Fang, X. Guo, Y. Li, R. Luo, S. Tai, H. Wu, H. Zheng, X. Zheng, Y. Zhou, G. T. Marth, E. P. Garrison, W. Huang, A. Indap, D. Kural, W.-P. Lee, W. F. Leong, A. R. Quinlan, C. Stewart, M. P. Stromberg, A. N. Ward, J. Wu, C. Lee, R. E. Mills, X. Shi, M. J. Daly, M. A. DePristo, A. D. Ball, E. Banks, B. L. Browning, K. V. Garimella, S. R. Grossman, R. E. Handsaker, M. Hanna, C. Hartl, A. M. Kernytsky, J. M. Korn, H. Li, J. R. Maguire, S. A. McCarroll, A. McKenna, J. C. Nemesh, A. A. Philippakis, R. E. Poplin, A. Price, M. A. Rivas, P. C. Sabeti, S. F. Schaffner, I. A. Shlyakhter, D. N. Cooper, E. V. Ball, M. Mort, A. D. Phillips, P. D. Stenson, J. Sebat, V. Makarov, K. Ye, S. C. Yoon, C. D. Bustamante, A. Boyko, J. Degenhardt, S. Gravel, R. N. Gutenkunst, M. Kaganovich, A. Keinan, P. Lacroute, X. Ma, A. Reynolds, L. Clarke, F. Cunningham, J. Herrero, S. Keenen, E. Kulesha, R. Leinonen, W. McLaren, R. Radhakrishnan, R. E. Smith, V. Zalunin, X. Zheng-Bradley, J. O. Korbel, A. M. Stuetz, M. Bauer, R. K. Cheetham, T. Cox, M. Eberle, T. James, S. Kahn, L. Murray, K. Ye, F. M. De La Vega, Y. Fu, F. C. L. Hyland, J. M. Manning, S. F. McLaughlin, H. E. Peckham, O. Sakarya, Y. A. Sun, E. F. Tsung, M. A. Batzer, M. K. Konkel, J. A. Walker, M. W. Albrecht, V. S. Amstislavskiy, R. Herwig, D. V. Parkhomchuk, R. Agarwala, H. Khouri, A. O. Morgulis, J. E. Paschall, L. D. Phan, K. E. Rotmistrovsky, R. D. Sanders, M. F. Shumway, C.

Xiao, A. Auton, Z. Iqbal, G. Lunter, J. L. Marchini, L. Moutsianas, S. Myers, A. Tumian, J. Knight, R. Winer, D. W. Craig, S. M. Beckstrom-Sternberg, A. Christoforides, A. A. Kurdoglu, J. Pearson, S. A. Sinari, W. D. Tembe, D. Haussler, A. S. Hinrichs, S. J. Katzman, A. Kern, R. M. Kuhn, M. Przeworski, R. D. Hernandez, B. Howie, J. L. Kelley, S. C. Melton, Y. Li, P. Anderson, T. Blackwell, W. Chen, W. O. Cookson, J. Ding, H. M. Kang, M. Lathrop, L. Liang, M. F. Moffatt, P. Scheet, C. Sidore, M. Snyder, X. Zhan, S. Zoellner, P. Awadalla, F. Casals, Y. Idaghdour, J. Keebler, E. A. Stone, M. Zilversmit, L. Jorde, J. Xing, E. E. Eichler, G. Aksay, C. Alkan, I. Hajirasouliha, F. Hormozdiari, J. M. Kidd, S. C. Sahinalp, P. H. Sudmant, K. Chen, A. Chinwalla, L. Ding, D. C. Koboldt, M. D. McLellan, J. W. Wallis, M. C. Wendl, Q. Zhang, C. A. Albers, Q. Ayub, S. Balasubramaniam, J. C. Barrett, Y. Chen, D. F. Conrad, P. Danecek, E. T. Dermitzakis, M. Hu, N. Huang, M. E. Hurles, H. Jin, L. Jostins, T. M. Keane, S. Q. Le, S. Lindsay, Q. Long, D. G. MacArthur, S. B. Montgomery, L. Parts, C. Tyler-Smith, K. Walter, Y. Zhang, M. B. Gerstein, M. Snyder, A. Abyzov, S. Balasubramanian, R. Bjornson, J. Du, F. Grubert, L. Habegger, R. Haraksingh, J. Jee, E. Khurana, H. Y. K. Lam, J. Leng, X. J. Mu, A. E. Urban, Z. Zhang, C. Coafra, H. Dinh, C. Kovar, S. Lee, L. Nazareth, J. Wilkinson, H. M. Khouri, A. Coffey, C. Scott, N. Gharani, J. S. Kaye, A. Kent, T. Li, A. L. McGuire, P. N. Ossorio, C. N. Rotimi, Y. Su, L. H. Toji, L. D. Brooks, A. L. Felsenfeld, J. E. McEwen, A. Abdallah, C. R. Juenger, N. C. Clemm, A. Duncanson, E. D. Green, M. S. Guyer, J. L. Peterson, and C. Genomes Project. 2010. A map of human genome variation from population-scale sequencing. Nature 467.

56.   Rabiner, L. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE 77:257-286.

57.   Lasken, R. S. 2009. Genomic DNA amplification by the multiple displacement amplification (MDA) method. Biochemical Society Transactions 37:450-453.

58.   Yan, J., J. N. Feng, S. Hosono, and S. S. Sommer. 2004. Assessment of multiple displacement
        amplification in molecular epidemiology. Biotechniques 37:136-+.

59.   Jun, G., M. Flickinger, K. N. Hetrick, J. M. Romm, K. F. Doheny, G. R. Abecasis, M. Boehnke,
        and H. M. Kang. 2012. Detecting and Estimating Contamination of Human DNA Samples in
        Sequencing and Array-Based Genotype Data. American Journal of Human Genetics 91:839-848.

60.   Cibulskis, K., A. McKenna, T. Fennell, E. Banks, M. DePristo, and G. Getz. 2011. ContEst:
        estimating cross-contamination of human samples in next-generation sequencing data.
        Bioinformatics 27:2601-2602.

61.   Diskin, S. J., M. Li, C. Hou, S. Yang, J. Glessner, H. Hakonarson, M. Bucan, J. M. Maris, and K.
        Wang. 2008. Adjustment of genomic waves in signal intensities from whole-genome SNP
        genotyping platforms. Nucleic Acids Research 36.

62.   Scheet, P., and M. Stephens. 2006. A fast and flexible statistical model for large-scale population
        genotype data: Applications to inferring missing genotypes and haplotypic phase. American
        Journal of Human Genetics 78:629-644.

63.   Barresi, V., A. Romano, N. Musso, C. Capizzi, C. Consoli, M. P. Martelli, G. Palumbo, F. Di
        Raimondo, and D. F. Condorelli. 2010. Broad Copy Neutral-Loss of Heterozygosity Regions and
        Rare Recurring Copy Number Abnormalities in Normal Karyotype-Acute Myeloid Leukemia
        Genomes. Genes Chromosomes & Cancer 49:1014-1023.

64.   MacDonald, J. R., R. Ziman, R. K. C. Yuen, L. Feuk, and S. W. Scherer. 2014. The Database of
        Genomic Variants: a curated collection of structural variation in the human genome. Nucleic
        Acids Research 42:D986-D992.

65.   Korn, J. M., F. G. Kuruvilla, S. A. McCarroll, A. Wysoker, J. Nemesh, S. Cawley, E. Hubbell, J.
        Veitch, P. J. Collins, K. Darvishi, C. Lee, M. M. Nizzari, S. B. Gabriel, S. Purcell, M. J. Daly,
        and D. Altshuler. 2008. Integrated genotype calling and association analysis of SNPs, common
        copy number polymorphisms and rare CNVs. Nature Genetics 40:1253-1260.

66.   Klampfl, T., J. D. Milosevic, A. Puda, A. Schoenegger, K. Bagienski, T. Berg, A. S. Harutyunyan, B. Gisslinger, E. Rumi, L. Malcovati, D. Pietra, C. Elena, M. G. Della Porta, L. Pieri, P. Guglielmelli, C. Bock, M. Doubek, D. Dvorakova, N. Suvajdzic, D. Tomin, N. Tosic, Z. Racil, M. Steurer, S. Pavlovic, A. M. Vannucchi, M. Cazzola, H. Gisslinger, and R. Kralovics. 2013. Complex Patterns of Chromosome 11 Aberrations in Myeloid Malignancies Target CBL, MLL, DDB1 and LMO2. Plos One 8.

67.   Barresi, V., G. A. Palumbo, N. Musso, C. Consoli, C. Capizzi, C. R. Meli, A. Romano, F. Di Raimondo, and D. F. Condorelli. 2010. Clonal selection of 11q CN-LOH and CBL gene mutation in a serially studied patient during MDS progression to AML. Leukemia Research 34:1539-1542.

68.   Dunbar, A. J., L. P. Gondek, C. L. O'Keefe, H. Makishima, M. S. Rataul, H. Szpurka, M. A. Sekeres, X. F. Wang, M. A. McDevitt, and J. P. Maciejewski. 2008. 250K Single Nucleotide Polymorphism Array Karyotyping Identifies Acquired Uniparental Disomy and Homozygous Mutations, Including Novel Missense Substitutions of c-Cbl, in Myeloid Malignancies. Cancer Research 68:10349-10357.

69.   Naramura, M., S. Nadeau, B. Mohapatra, G. Ahmad, C. Mukhopadhyay, M. Sattler, S. M. Raja, A. Natarajan, V. Band, and H. Band. 2011. Mutant Cbl proteins as oncogenic drivers in myeloproliferative disorders. Oncotarget 2:245-250.

70.   Rathinam, C., C. B. F. Thien, W. Y. Langdon, H. Gu, and R. A. Flavell. 2008. The E3 ubiquitin ligase c-Cbl restricts development and functions of hematopoietic stem cells. Genes & Development 22:992-997.

71.   Abyzov, A., J. Mariani, D. Palejev, Y. Zhang, M. S. Haney, L. Tomasini, A. F. Ferrandino, L. A. R. Belmaker, A. Szekely, M. Wilson, A. Kocabas, N. E. Calixto, E. L. Grigorenko, A. Huttner, K. Chawarska, S. Weissman, A. E. Urban, M. Gerstein, and F. M. Vaccarino. 2012. Somatic

copy number mosaicism in human skin revealed by induced pluripotent stem cells. Nature 492:438-442.

72.   Lupski, J. R. 2013. Genome Mosaicism-One Human, Multiple Genomes. Science 341:358-359.

73.   Macosko, E. Z., and S. A. McCarroll. 2012. Exploring the variation within. Nature Genetics 44:614-616.

74.   Gogarten, S. M., T. Bhangale, M. P. Conomos, C. A. Laurie, C. P. McHugh, I. Painter, X. Zheng, D. R. Crosslin, D. Levine, T. Lumley, S. C. Nelson, K. Rice, J. Shen, R. Swarnkar, B. S. Weir, and C. C. Laurie. 2012. GWASTools: an R/Bioconductor package for quality control and analysis of genome-wide association studies. Bioinformatics 28:3329-3331.

75.   Wang, K., M. Li, D. Hadley, R. Liu, J. Glessner, S. F. A. Grant, H. Hakonarson, and M. Bucan. 2007. PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Research 17:1665-1674.

76.   Shiina, T., K. Hosomichi, H. Inoko, and J. K. Kulski. 2009. The HLA genomic loci map: expression, interaction, diversity and disease. Journal of Human Genetics 54:15-39.

77.   Liu, Y., S. Vattathil, L. Huang, X. Xiao, G. Davies, E. Ehli, J. Hottenga, A. Abdellaoui, I. Ruczinski, S. Arur, D. Boomsma, T. Beaty, and P. Scheet. 2014. Estimating the parental haplotype source of germline-transmitted de novo duplications; Program # 610S. Presented at the Annual Meeting of The American Society of Human Genetics, October 2014, San Diego, California.

78.   Rodriguez-Santiago, B., N. Malats, N. Rothman, L. Armengol, M. Garcia-Closas, M. Kogevinas, O. Villa, A. Hutchinson, J. Earl, G. Marenne, K. Jacobs, D. Rico, A. Tardon, A. Carrato, G. Thomas, A. Valencia, D. Silverman, F. X. Real, S. J. Chanock, and L. A. Perez-Jurado. 2010. Mosaic Uniparental Disomies and Aneuploidies as Large Structural Variants of the Human Genome. American Journal of Human Genetics 87:129-138.

79. Venkatraman, E. S., and A. B. Olshen. 2007. A faster circular binary segmentation algorithm for the analysis of array CGH data. Bioinformatics 23:657-663.

80. Page, E. S. 1954. Continuous inspection schemes. Biometrika 41:100-&.

81. Howie, B., J. Marchini, and M. Stephens. 2011. Genotype Imputation with Thousands of Genomes. G3-Genes Genomes Genetics 1:457-469.

82. Amos, C., A. Antoniou, A. Berchuck, G. Chenevix-Trench, C. FJ, R. Eeles, L. Esserman, S. Gayther, C. Goh, D. Goldgar, S. Gruber, C. Haiman, P. Hall, D. Hunter, Z. Kote-Jarai, P. Lepage, S. Lindstrom, J. McKay, R. Milne, U. Peters, P. Pharoah, C. Phelan, F. Schumacher, T. Sellers, J. Simard, Z. Wang, D. Seminara, S. Chanock, D. Easton, and B. Henderson. 2013. A comprehensive genetic analysis of common cancer risk through the development of the Oncochip. Presented at the Annual Meeting of The American Society of Human Genetics, October 2013, Boston, Massachusetts.

Vita

Selina Maria Vattathil was born in Houston, Texas on December 29, 1982. She graduated from Alief Hastings High School in 2001 and then studied at The University of Texas at Austin, earning bachelor's degrees in Biology and Plan II in 2005. After three years as a research technician at the Human Genome Sequencing Center at Baylor College of Medicine in Houston, Texas, she began her graduate studies at the Graduate School of Biomedical Sciences in autumn of 2008. She will continue her research training as a postdoctoral fellow in the Department of Genome Sciences at the University of Washington.