CellPress REVIEWS

Review

Detecting Somatic Mutations in Normal Cells

Yanmei Dou,[1,4] Heather D. Gold,[1,2,4] Lovelace J. Luquette,[1,2,4] and Peter J. Park[1,3,*]

Somatic mutations have been studied extensively in the context of cancer. Recent studies have demonstrated that high-throughput sequencing data can be used to detect somatic mutations in non-tumor cells. Analysis of such mutations allows us to better understand the mutational processes in normal cells, explore cell lineages in development, and examine potential associations with age-related disease. We describe here approaches for characterizing somatic mutations in normal and non-tumor disease tissues. We discuss several experimental designs and common pitfalls in somatic mutation detection, as well as more recent developments such as phasing and linked-read technology. With the dramatically increasing numbers of samples undergoing genome sequencing, bioinformatic analysis will enable the characterization of somatic mutations and their impact on non-cancer tissues.

Highlights

Somatic mosaicism resulting from post-zygotic mutations has been shown to contribute to many diseases including brain-related disorders, in addition to cancer. Emerging data also suggest that mosaicism is common in healthy individuals.

Mutations occurring late in development have very low allele fractions, and their detection requires specialized algorithms and filters that can remove artifacts that arise in sample handling, DNA sequencing, and analysis.

Emerging technologies, such as single-cell sequencing and linked-read sequencing, allow improved phasing of variants, thus increasing detection accuracy.

Somatic Mosaicism and Challenges in Detecting Mosaic Variants

Genomes from individuals of the same species differ from one another because of a constant influx of genetic mutation and recombination. **Single-nucleotide variants** (SNVs; see Glossary), **copy-number variants** (CNVs), **transposable element** (TE) insertions, and other **structural variants** (SVs) are common types of genetic variation. Population-level heterogeneity generally arises due to germline mutations that occur before the formation of the zygote, and are inherited by all cells in the offspring. However, heterogeneity within an individual may also exist due to **somatic mutations** that occur post-zygotically and exist only in a subpopulation of cells. The genetic heterogeneity resulting from somatic mutations is known as **somatic mosaicism**. Recent papers have attempted to characterize somatic mosaicism [1], but the extent to which it exists, whether specific regions of the genome and nucleotide contexts are more susceptible to it, and how it impacts on normal cellular function remain open questions.

In **bulk sequencing** data, somatic mutations have **variant allele fractions** (VAFs) that deviate from those typical of germline mutations (~0.5/1 for heterozygous/homozygous). The VAF of a somatic mutation depends both on the prevalence of the mutation, which is largely driven by how early the mutation occurs in development, and on the heterogeneity of the tissue selected for sequencing. For example, if a mutation occurs during the first cell division, and every cell produces the same number of descendants, the VAF would be ~0.25 in an unbiased sample (Figure 1, Key Figure). At the other extreme, if a mutation is uniquely acquired in a post-mitotic cell, the VAF would be infinitesimal (if bulk sequencing with 1 million cells, the VAF would be $\sim 0.5 \times 10^{-6}$). In general, somatic mutations occurring earlier during development attain higher VAFs than those occurring later. However, asymmetry in the developmental cell-lineage tree [2], heterogeneity in selective pressure across tissues [3], and technical factors (such low read depth, sequencing errors, and misalignment) can violate this principle.

A great deal of work has been done to develop algorithms for detecting somatic mutations in cancer. However, the VAFs of functionally relevant cancer mutations tend to be higher than those in normal cells because of the selective advantage conferred by those mutations in

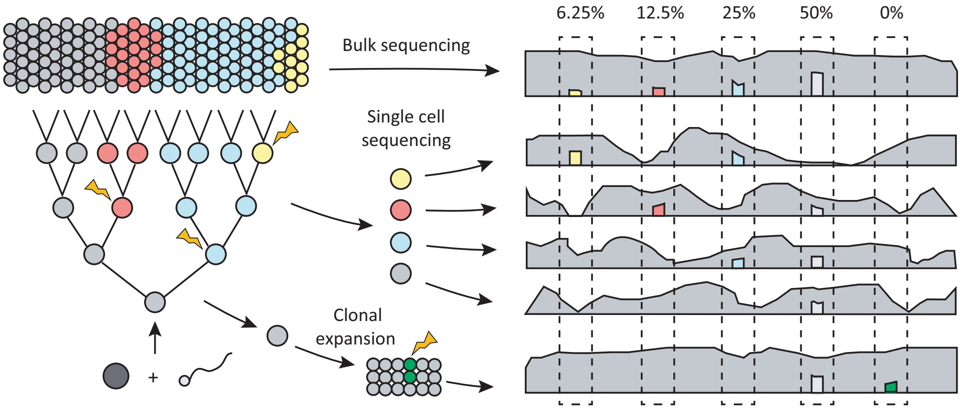
[1] Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA
[2] Bioinformatics and Integrative Genomics PhD Program, Harvard Medical School, Boston, MA, USA
[3] Division of Genetics, Brigham and Women's Hospital, Boston, MA, USA
[4] Equal contributions

*Correspondence: peter_park@hms.harvard.edu (P.J. Park).

https://doi.org/10.1016/j.tig.2018.04.003

Key Figure

Detecting Somatic Mosaicism in the Genome through Various Sequencing Strategies



Trends in Genetics

Figure 1. Somatic mutations arise during development and propagate to a subpopulation of cells (blue, 50% of cells; red, 25%; yellow, 12.5%). With bulk sequencing these somatic mutations are expected to be approximately half of the subpopulation frequency. Lower-frequency somatic mutations require higher sequencing depth to maintain detection sensitivity. With single-cell sequencing, somatic mutations can be detected as heterozygous variants that occur in a subset of cells. The ability to detect variants is dependent on uniformity of coverage and allelic balance in genome amplification, as well as on picking cells that contain variants. Clonal expansion followed by bulk sequencing does not suffer from the problems associated with single-cell sequencing, but artifactual mutations that occur early during expansion (green) can be difficult to distinguish from mutations in the original cell.

proliferating cells. Thus, many popular algorithms for cancer are not focused on detecting very low VAF events (e.g., <5% [4]), and comprehensive detection of somatic mutations at arbitrarily small VAFs in normal cells requires alternative methods. In addition, somatic mutations in cancer are typically identified by the tumor–normal design in which tumor tissue is compared to non-cancerous (‘normal’) tissue from the same individual to determine the mutations unique to the tumor. For non-cancer samples, mosaic variants arising early in embryogenesis are often shared among many tissues. This makes it difficult to identify a clear normal cellular subpopulation that can serve as a matched control. With careful selection of tissue specimens, however, it is possible to derive an accurate list of mosaic mutations that allows lineage analysis of cells in an individual. For example, Lodato *et al*. [5] analyzed heart and brain tissues, which develop from the mesoderm and ectoderm, respectively, to find mosaic mutations informative of brain cell lineage; Behjati *et al*. [6] compared endoderm-derived gastrointestinal tissues to mouse tail, which consists of both mesodermal and ectodermal tissues, to find early embryonic mutations. The locations of selected specimens within a larger tissue can also be relevant: Martincorena *et al*. [7] utilized ultra-deep sequencing of multiple nearby fine biopsies to infer spatial patterns and rates of mosaicism in human skin.

In this review we provide an overview of somatic mutation analysis in normal cells. We first cover the various platforms and experimental designs including bulk sequencing and single-cell

Glossary

Allelic imbalance/allele dropout: a difference in the sequencing of two alleles caused by differential amplification (allelic imbalance) or amplification failure (allele dropout) of one allele. This is a frequent source of error in single-cell DNA sequencing.

Amplification bias: the differential amplification of a region of DNA relative to another, resulting in unequal coverage across the genome.

Bulk sequencing: the sequencing of DNA extracted from a large number of cells from the same individual.

Copy-number variant (CNV): a section of the genome that has a different number of repeats or copies than the reference genome.

Coverage: the number of reads overlapping a region in DNA sequencing, also known as depth. Sequence coverage can also refer to the average number of reads covering loci across the entire genome.

Haplotype: a segment of DNA that is inherited as a block from a single parent.

Indel: the insertion or deletion of a small sequence of DNA (1–50 bp) in the genome; affects fewer bases than a structural variant.

Phasing: the process of statistical estimation of the haplotypes of an individual by using the variants in their genome, also called haplotype estimation.

Single-nucleotide variant (SNV): a single nucleotide in the genome of an individual that differs from the nucleotide in the reference sequence.

Somatic mutation: a change to the genome of an individual that arises during its lifetime as opposed to being inherited; also called a post-zygotic mutation.

Somatic mosaicism: the presence in an individual of at least two genetically distinct populations of cells that arise from somatic mutation (s).

Structural variant (SV): a rearrangement of the genome that affects a region greater than 50 bp. Structural variation comprises many types of variants with varying length scales, including deletions, insertions, sequencing. We then describe strategies for detecting variants such as **phasing** of **haplotypes**, as well as common pitfalls encountered in these analyses.

Strategies for Profiling Mosaic Variants

Whole-genome sequencing (WGS), whole-exome sequencing (WES), and targeted panels offer tradeoffs between the types of detectable variants and the range of detectable VAFs. WGS produces the most uniform read depth across the genome and enables the detection of most types of somatic mutations, including structural variants. However, detection is limited to relatively high VAF mutations because the high sequencing depth required to detect low VAF mutations remains prohibitively expensive [8]. If attention can be restricted to specific loci, a customized panel can be constructed (e.g., amplicon-seq or targeted hybridization methods) and sequenced at very high depth (e.g., >100 000×). WES offers a compromise between WGS and small panels by targeting the ~1–2% of the genome that codes for proteins and does not need to be custom-designed.

Characterizing Variants at the Single-Cell Level

Unlike bulk sequencing strategies that pool DNA from thousands or millions of cells, single-cell sequencing attempts to sequence the DNA of only one cell. The advantage is that rare mosaic mutations can be more easily detected: if present in a diploid region of the chosen cell, the mutation will be present on one of two alleles, regardless of its frequency in the surrounding tissue (Figure 1). This shifts the technical difficulties associated with low frequency away from variant detection and onto the cell selection process. To estimate the overall frequency of each mutation in the tissue, multiple single cells must be sequenced, which can be expensive, laborious, and confounded by sampling bias. Hybrid experimental designs integrating bulk (either WGS or targeted) and single-cell approaches can address many of these issues. For example, somatic mutations discovered in bulk can be confirmed by single-cell data, and frequencies for somatic mutations discovered in single cells can be estimated from bulk sequencing.

A common strategy to produce sufficient input DNA for next-generation sequencing from a single cell is clonal expansion, in which a cell is expanded in culture until there are sufficient cells to perform standard bulk sequencing [6,9–13]. However, additional mutations – especially SNVs – are continuously acquired during expansion, and these must be differentiated from mutations that existed in the founding cell. This is often addressed by discarding low VAF candidate mutations because *in vitro* mutations acquired after the first mitosis should be present at <25% VAF if cell division in culture is approximately symmetric. However, this symmetry assumption could be violated by variability in cell cycle length and the potential for selectively advantageous mutations *in vitro*, and careful analysis is therefore warranted. It has also been shown that *in vitro* mutations can be characterized by mutational signatures that correlate with increasing culture time [6,10]. An additional concern is that only a subset of the isolated single cells may successfully expand into colonies, possibly reflecting differences in cell fitness, tolerance to handling and cell culture, or stochastic effects. Thus, studies relying exclusively on clonal expansion might not provide an accurate picture of tissue heterogeneity owing to biased loss of specific cell types. For post-mitotic cell types (e.g., neurons), clonal expansion is not directly applicable. Encouragingly, a recent study demonstrated that adult neurons in mice could be clonally expanded and sequenced after inducing totipotency via single-cell nuclear transfer (SCNT) [14]. However, SCNT is labor-intensive, notoriously inefficient, and may be even further affected by selection biases.

Another widely used approach to produce enough DNA from a single cell is to apply **whole-genome amplification** (WGA) [15–17] followed directly by sequencing. This approach has been used both in cancer [18–20] and in development [5,21–24]. Several methods for WGA are

translocations, and TE insertions. SVs encompass events that result in CNVs and copy-neutral events.
Transposable element (TE): a sequence of DNA that, either via an RNA intermediate (retrotransposons) or a DNA intermediate (DNA transposons), can relocate within a genome. Active TEs include L1 and *Alu* elements.
Variant allele fraction (VAF): the fraction of sequencing reads in a sample corresponding to the non-reference allele. For bulk sequencing data, this is an estimate of the frequency of DNA molecules carrying the variant.
Whole-genome amplification (WGA): the amplification of a single genome, or a similarly limited amount of DNA, to generate sufficient DNA for sequencing. WGA is necessary for single-cell DNA sequencing.

available, and represent different tradeoffs between genomic **coverage**, amplification uniformity, and artifact load, and are reviewed elsewhere [15]. Because cell culture is unnecessary, WGA-based methodologies enjoy significant cost savings in both labor and reagents, and can be directly applied to post-mitotic cells (such as neurons) and cells that are difficult to culture. The technical simplicity of WGA has also made it an attractive technology for scaling to handle hundreds or thousands of cells simultaneously [25,26]. However, the disadvantage of WGA is the introduction of considerable **amplification bias** and **allelic imbalance/allele dropout**, which can produce artifacts that can be difficult to distinguish from true mutations. Research to improve variant calling despite these amplification artifacts is ongoing. It was recently demonstrated [27] that good specificity can be achieved for SNV detection for candidate somatic mutations that can be linked to nearby germline heterozygous variants (~20% of the candidates, if using standard Illumina sequencing; discussed further below).

Mosaic Mutation Calling

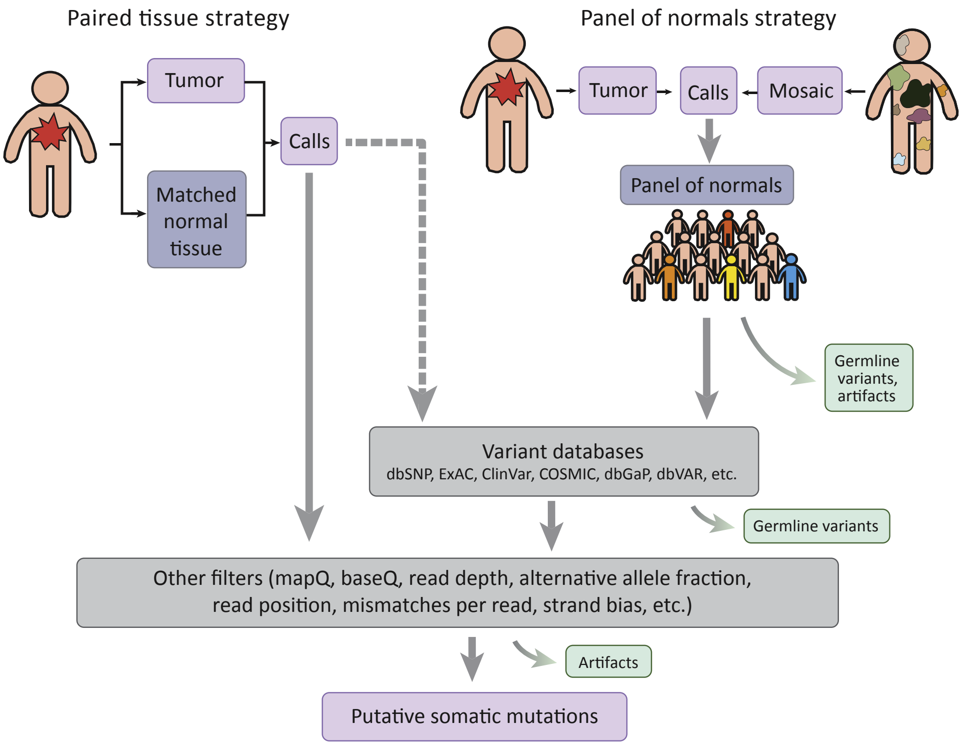
Approaches for Filtering Germline Variants

In cancer applications, mutation callers are often designed to simultaneously evaluate data from tumor and matched normal tissue from the same individual to discard mutations with any support in the normal tissue [4,28–31]. Germline variants can also be filtered out by querying public variation databases or by using a ‘panel of normals’ (PON) consisting of unrelated individuals (Figure 2). A recent study estimated that common variants in the public dbSNP database account for ~95% of germline SNVs in a typical human genome [32–34]; however, there is also evidence that aggressive exclusion of all polymorphic sites in dbSNP could lead to considerable false negative rates [35]. If the PON samples are processed and analyzed in the same way as the tumor, the PON approach can better control for systematic artifacts, such as those due to misalignment. For removing germline variants, it has been estimated that a PON consisting of at least 400 individuals would be necessary to reach the accuracy of having a matched normal sample [36]. Matched normal tissue sequencing, PON approaches, and population databases are often combined to achieve high specificity.

Applying these same strategies to detecting somatic mutations in non-tumor samples is problematic because there is no clear ‘normal’ tissue to use as a reference. When another tissue from the same individual is used as a reference, a true somatic mutation can be present in the reference sample if the mutation occurred in a common ancestor to both selected tissues. A large panel of other individuals may be used, with all samples being processed in the same way as the sample of interest. However, because the somatic mutation rate in non-cancer samples is much lower than in cancer [37–39], studies of somatic mosaicism are substantially less tolerant of false positives. More sophisticated algorithms and a series of stringent filters are necessary for detecting somatic mutations with higher accuracy. One example is MosaicHunter [40], which aims to detect mosaic SNVs without a matched normal by using a Bayesian approach.

Some have applied germline variant callers, such as the Genome Analysis Toolkit (GATK) HaplotypeCaller, to detect mosaic SNVs [41]. One approach is to search for ‘heterozygous’ mutations and then to distinguish somatic mutations from germline mutations using a VAF threshold or other *ad hoc* heuristics. To increase sensitivity for low VAF variants, one could set the ‘ploidy’ in GATK HaplotypeCaller to be high, which lowers the expected VAF for a heterozygous variant [42]. However, a straightforward application of a germline caller is unlikely to yield sufficient sensitivity.

A parent–offspring trio analysis greatly increases the accuracy of variant detection because mosaic mutations arising post-zygotically in a child are unlikely to be shared by the parents.


Paired tissue strategy
Tumor
Matched normal tissue
Calls
Panel of normals strategy
Tumor
Calls
Mosaic
Panel of normals
Germline variants, artifacts
Variant databases
dbSNP, ExAC, ClinVar, COSMIC, dbGaP, dbVAR, etc.
Germline variants
Other filters (mapQ, baseQ, read depth, alternative allele fraction, read position, mismatches per read, strand bias, etc.)
Artifacts
Putative somatic mutations

Trends in Genetics

Figure 2. Different Strategies for Detecting and Filtering Somatic Mutations. Somatic variant callers for a tumor tissue often require a matched normal tissue from the same individual. However, this strategy is not possible when matched normal tissue is unavailable. For somatic mosaicism in a non-tumor tissue, a matched 'normal' may not exist because mutations of interest can be shared across tissues. Whenever matched normal tissue is unavailable, germline variants as well as some artifacts can be removed by querying public variation databases or by constructing a 'panel of normals' from sequencing data of unrelated individuals. Additional filters can be applied to further remove artifacts, and both PON and variant databases are frequently applied when matched normal tissue is available to improve specificity.

Recently, four groups [41–44] studying autism spectrum disorders successfully detected mosaic SNVs by WES of parent–offspring trios using various approaches. However, even after removing variants present in either parent, mosaic SNV validation rates remained modest (~10–40%; validation is discussed further below). Each study found it necessary to apply additional filters to reduce false positives, and in some cases it was necessary to exclude families with excess candidate mutations altogether. It was also apparent that the detection sensitivity and accuracy of many tools were diminished for mosaic SNVs with VAF <0.10.

Detection of Mosaic Structural Variations

Somatic CNV detection in cancer is complicated by clonal heterogeneity as well as by experimental and technical noise, thus requiring sophisticated computational approaches [45]. Detecting mosaic CNVs in non-tumor samples is challenging because the amplitude of the copy-number change may be small, and matched normal samples are frequently unavailable. Some success in identifying mosaic CNVs has been achieved in single-cell sequencing data [21,22] or WES data [46], but they are limited to large, Mb-scale CNVs. Combined haplotyping (described later) of CNVs and SNVs or pedigree-based analyses appear to be the most promising strategies for detecting mosaic CNVs [47,48].

Structural variation consists of many types of variants with varying length-scales, including deletions, insertions, translocations, and TE insertions (SVs encompass events that result in CNVs and copy-neutral events). A survey of existing SV callers can be found elsewhere [49].

Less progress has been made in detecting mosaic SVs because most somatic SV methods in cancer require matched normal data [50–54]. A recent method called MrMosaic can detect mosaic SVs without any matched normal by using deviations in coverage and allele fraction at polymorphic SNV sites [46]. However, MrMosaic can only detect insertions, deletions, and loss-of-heterozygosity events, and does not identify specific breakpoints.

TEs are DNA sequences that can be copied and reinserted into the genome. Although most TE activity in somatic tissue is repressed, some TEs are active during early embryogenesis and in germline cells. TEs have been shown to play important roles in many cancers [55–57], and there is some evidence that TEs may contribute to neuronal diversity, although the rate of such insertions has been shown to be much lower than was initially proposed [58]. Detection of mosaic TEs from bulk data is difficult unless the insertion occurred early in development and has a high VAF. Alternative approaches include L1-insertion profiling [59] or WGS [23] for single cells.

Alternative Technologies

Although short-read sequencing has matured considerably, it still suffers from alignment issues (especially in repetitive regions) and has limited power to detect complex structural variants. In particular, detecting very low VAF variants requires relying on as few supporting reads as possible, and even the smallest error rate in sequencing introduces potential artifacts. Read misalignments, which can create artifacts with many supporting reads, are often very difficult to differentiate from true somatic mutations. One recent technology with potential to alleviate issues related to short-read alignment is linked-read sequencing, in which fragments derived from the same long DNA molecule share a unique barcode [60]. These short fragments are then sequenced using standard short-read platforms, and the barcodes are used to stitch the reads into long sequences representing the original DNA molecule. Linked-read sequencing incurs additional cost for library construction but provides new opportunities for haplotype construction, detecting complex structural variants and extending mutation detection into repetitive regions of the genome. Single-molecule sequencing chemistries from PacBio and Oxford Nanopore also offer similar advantages, but their relatively high per-base error rate and cost do not make them competitive for large-scale profiling at this point.

Methods have also been designed to reduce the rate of sequencing artifacts to an order of magnitude below the expected range of somatic mutation rates by sequencing both the forward and reverse DNA strands [61,62]. The mutations identified represent a random sampling of mosaics from the cell population and can provide estimates for somatic mutation rates and spectra. In theory, these methods can detect mutations present on only a single DNA molecule with reasonable specificity. In practice, a small fraction of the genome can be assayed and higher VAF mutations are more likely to be sampled.

Increasing Accuracy by Haplotype Phasing

A haplotype is the sequence of alleles on one chromosome that are inherited from a single parent and haplotype phasing – sometimes simply referred to as haplotyping or phasing – is the process of identifying alleles that are colocated on the same chromosome. Haplotype phasing is informative in several applications, including correlating genetic variation with disease, detecting genotyping error, inferring evolutionary history, and examining the effect of *cis*-regulatory elements on gene expression [63]. Phasing is beneficial for somatic mutation detection because true mosaic events create a new haplotype with a consistent allele sequence, whereas artifacts often associate with haplotypes non-specifically.

Read-Based Phasing for Mosaic SNVs

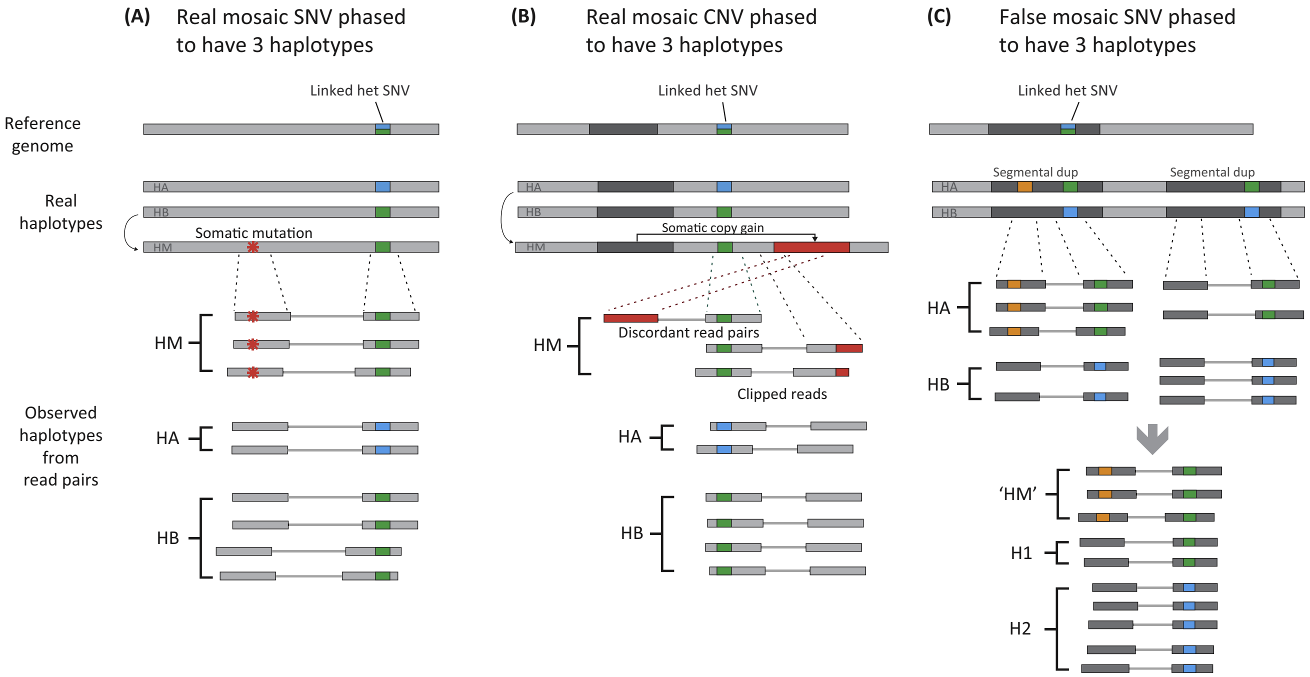
Traditional germline phasing methods infer haplotypes by taking advantage of segregation patterns in related individuals [63] or models of genetic recombination and mutation in a large population [63,64]. However, these methods require genotype data from several individuals and depend on genetic inheritance, and they are therefore of little use when phasing *de novo* somatic mutations. Sequencing data enable a different approach to phasing by exploiting the direct physical evidence of linkage provided by reads (or read pairs) that span multiple variants. This 'read-based' phasing approach does not rely on inheritance and can be easily applied to data from a single individual; however, it is only effective when consecutive variant loci are close enough to be covered by a single sequencing read (or read pair). Because the read (or library fragment) length determines the maximum linkable inter-variant distance, the effectiveness of this approach depends considerably on the choice of sequencing platform [65,66].

Spontaneously arising mosaic mutations are extremely unlikely to affect more than one haplotype, and true mosaics that can be linked to a nearby germline heterozygous variant should therefore be associated with only one of the two germline alleles. In bulk sequencing data of diploid organisms, a pair of SNVs consisting of a mosaic mutation and a germline heterozygote should therefore produce three haplotypes (Figure 3A,B), whereas some types of artifacts (e.g., misalignment, sequencing errors at homopolymers, sample contamination) would associate with both alleles, generating additional haplotypes. Most candidate mosaic mutations with two or four apparent haplotypes can be safely rejected. This approach has been successfully applied in various studies [2,42,43], and specialized mutation-and-linkage callers have been developed [67]. Although only a small set (~10–30%) of candidate mosaic events are sufficiently close to be linked to germline SNPs, the retained mutations are typically of higher quality. However, it is important to note that a significant fraction of variants with three haplotypes may still be false positives [42], most likely due to misalignment (Figure 3C). This is most prominent in repetitive regions, but must also be guarded against in nonrepetitive regions.

The linkage principle can be extended to two or more nearby germline heterozygotes to reduce the probability that an artifact associates only with a single germline allele by chance. However, the number of potential haplotypes increases exponentially with the number of heterozygotes considered, which quickly leads to computational issues. An algorithm called LocHap [68] models the number of haplotypes at several SNVs in small genomic regions, and defines regions with three or more haplotypes to contain mosaic events. However, because consideration of all possible haplotypes is computationally expensive, it disregards regions with more than three SNVs.

Phasing for Single Cells, Structural Variants, and the Use of Linked Reads

For single-cell data, read-based phasing is particularly attractive because standard variant callers have difficulty distinguishing true mutations from the relatively large number of artifactual mutations that arise in genome amplification. Although only ~20% of the total candidate mutations in single-cell WGS data can be phased, that subset can be used to infer the genome-wide mutation rate and to characterize the sequence features of the mutational processes. Recently, a method called LiRA was developed based on this idea [27] and was applied to neuron WGS data to demonstrate that aging and neurodegeneration are associated with an increased rate of mutation in the brain and to infer the source of those mutations [24].


(A)
Real mosaic SNV phased to have 3 haplotypes
(B)
Real mosaic CNV phased to have 3 haplotypes
(C)
False mosaic SNV phased to have 3 haplotypes
Linked het SNV
Reference genome
Real haplotypes
HA
HB
HM
Somatic mutation
Somatic copy gain
Segmental dup
HM
Discordant read pairs
Clipped reads
Observed haplotypes from read pairs
HA
HB
'HM'
H1
H2

Trends in Genetics

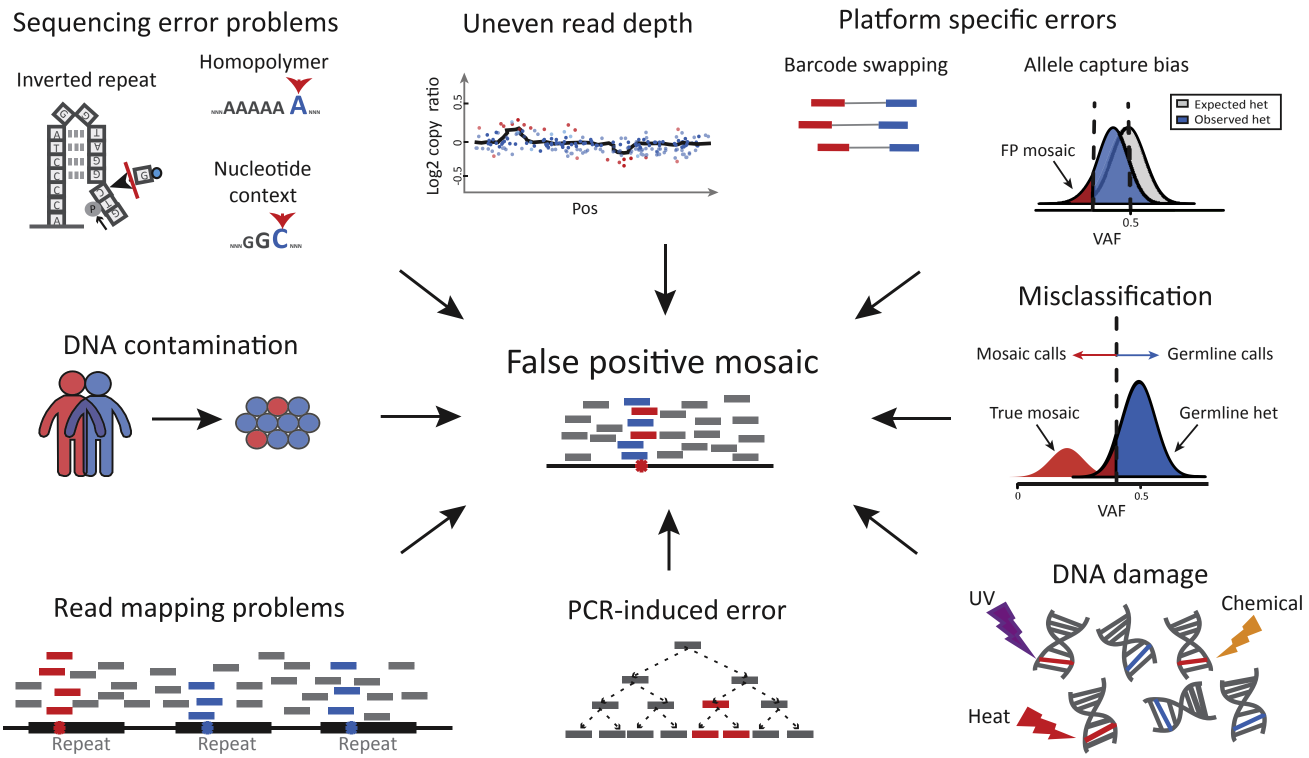
Figure 3. Overview of Read-Based Mosaic Phasing Scenarios. Read-based phasing can help to identify true somatic mosaic mutations by examining the relationship between germline heterozygous (het) variants and putative somatic mutations. However, some patterns of false positives can confound this method. (A) If a real mosaic single-nucleotide variant (SNV, red star) arises near a heterozygous single-nucleotide polymorphism (SNP) it will always be found in conjunction with one of the two SNP alleles (green) and will never appear on reads with the other allele (blue). This generates three haplotypes in bulk sequencing (HM for mosaic haplotype, in addition to HA and HB). (B) Similarly, a true mosaic CNV will phase with one allele of a nearby heterozygous SNP, resulting in three haplotypes. (C) A segmental duplication (dup) can cause a germline variant (orange) occurring on one duplicated segment to phase to a nearby heterozygous SNP occurring on both segments as if it were somatic, resulting in a false positive identification.

The prevalence of SVs in healthy individuals is still under active investigation [69]. Current SV phasing methods are limited to germline events and often rely on data from multiple sequencing strategies [70]. Mosaic SVs can, in principle, be phased to a germline SNV in a manner analogous to mosaic SNVs by regarding the inferred SV breakpoint as a point event. As in SNV detection, artifactual mosaic SVs are likely to link to both of the germline alleles (Figure 3B).

Longer sequencing reads increase the power of read-based haplotyping by increasing the fraction of the genome that can be physically linked to germline heterozygous sites [65,71], and by improving alignment to repetitive regions. Currently, the applicability of long-read platforms to mosaic mutation detection is limited owing to high cost and low per-basepair accuracy. However, an effective compromise may be provided by recent linked-read sequencing platforms that retain much of the long-range linkage information while achieving error rates similar to standard short-read sequencing. Several programs specializing in the phasing of linked reads are now available [60,72,73], and additional developments are likely to play an important role in future investigations of mosaic mutations.

Pitfalls in the Detection of Mosaic Variants

The search for mosaic mutations can be confounded by many factors, and claims of mosaic mutation discovery should be made cautiously. Several artifact sources that may lead to false positive mosaic calls are discussed below and summarized in Figure 4.



Trends in Genetics

Figure 4. False Positive Mosaic Calls Can Arise from Multiple Sources. Clockwise from top left: inverted repeats, homopolymers, and some specific nucleotide contexts are common locations of sequencing error. Uneven read coverage can cause false positive calls of mosaic copy-number variants (CNVs). Platform-specific errors from targeted sequencing methods may result in underestimating the variant allele fraction (VAF) for germline variants; barcode swapping can lead to the spread of false positive signals in multiplexed samples. Germline mutations with low VAF as a result of read sampling bias can be misclassified as somatic. DNA damage can induce artifactual single-base substitutions during sample handling and library preparation. PCR errors are also common and will propagate in subsequent PCR steps. Misalignment, especially within repetitive regions of the genome, contributes to a large proportion of false positive calls. Cross-individual contamination may lead to false positives. Abbreviations: FP, false positive; het, heterozygote; pos, position.

DNA Contamination

DNA contamination – whether by other samples or artificial constructs – can occur at several steps during sample handling and sequencing. DNA contamination by other human subjects is perhaps the most dangerous: it was recently estimated that 1.5% contamination by another human source is a common occurrence, and this produces roughly 0.2 erroneous somatic mutation calls per Mb in tumor–normal experiments [74], a considerable burden given that somatic mutation rate estimates in various cancers roughly range from 0.1 to 100 SNVs per Mb [1]. In principle, if genotypes for the contaminating individual are known, then putative somatic mutations coinciding with known genotypes in the contaminant should be treated with suspicion; if the contaminant is unknown, common variants from population databases can serve as an approximate substitute. Several algorithms can quantify contamination from sequencing data when the source is unknown [74–76] or known [77–80]. Some somatic mutation callers can be adjusted to compensate for contamination [4,29,75,81], but it is also reasonable to remove candidate mutations at known polymorphic sites (with the associated loss of sensitivity in mind) or to exclude highly contaminated samples altogether.

DNA Damage

Low levels of DNA damage frequently occur during routine sample handling and storage. Many sources of DNA damage have been identified: for example, UV radiation can create pyrimidine dimers [82], high temperature increases the rate of spontaneous cytosine deamination resulting in C > T transitions [83], reactive oxygen species can induce 7,8-dihydro-8-oxo-2′-deoxyguanosine (8-oxoG) which can mis-pair with A [84], and ionizing radiation can cause double-stranded DNA breaks [85]. These types of infrequent damage often go unnoticed in germline variant analyses, but they become much more prominent when low VAF somatic mutations are of interest. It was recently found that the majority of low VAF G > T/C > A somatic mutations in an exome dataset were likely caused by oxidative damage during library construction [43,61,86], and that similar damage is widespread in WGS samples [87]. Single-cell sequencing experiments are especially vulnerable to DNA damage before amplification because a single-base lesion affects a quarter of the original DNA strands. Indeed, pronounced effects have been observed when single cells are lysed by heat treatment [88]. Added care in sample handling and during routine benchwork may help to prevent damage to some extent, but investigators should remain wary because the full spectrum of damage-inducing processes is unknown.

Read-Mapping Problems

Improperly aligned reads are responsible for a large fraction of false positive variant calls, especially for the low VAF cases. Misalignment or non-unique alignment often occurs near an **indel** or in repetitive regions of the genome, such as centromeres or telomeres. Although repetitive regions are estimated to account for nearly half of the human genome [89], they pose such a great a challenge for mutation detection that they are often excluded from analysis [2,43,90]. Reads can also be misplaced due to limitations of the reference genome, which lacks any representation of genetic variation. Emerging long- and linked-read technologies will be needed to mitigate alignment issues. Ultimately, *de novo* assembly that does not rely on a reference genome will be needed; however, it is not yet feasible for routine analysis [72,73,91–93].

Sequencing Artifacts

While tolerable for germline variant calling, the per-base error rates intrinsic to sequencing platforms (~0.3% are miscalls according to one estimate [94]) are high relative to the rate of somatic mutation. If miscalls were produced independently, they would essentially be supported by only a single sequencing read, and thus be removed. However, artifacts are frequently reproduced by factors that increase local error density, such as homopolymer runs and high GC content [94–98], early amplification errors [99–101], uneven capture efficiency [102], and incorrect sample assignment in multiplexed sequencing runs [103]. Technical replicates can provide a modicum of internal control [104], but true low VAF mutations may also be less reproducible because of sampling bias.

Validation Methods for Mosaic Mutations

Because false positive mosaic mutations can arise from very many sources, confirmation using an orthogonal technology is essential. Available methodologies offer tradeoffs in cost, effort, and scalability [105,106]. A popular method is droplet digital PCR (ddPCR), which can achieve sensitivity as low as 0.001% VAF by performing millions of fluorescently labeled PCR reactions in nanoliter-sized droplets and measuring the fraction of fluorescent droplets [107]. A disadvantage of ddPCR, however, is that it is less scalable because PCR probes must be designed for every candidate mutation, and options for target multiplexing are currently limited [108]. Another approach is multiplexed confirmation of many candidates using deep sequencing, either through unique molecular barcodes that aid in artifact removal [109] or by sheer sequencing depth [110–113]. Mutations with VAF as low as 0.1% have been confirmed using these techniques [109,110]; similarly, mutations with relatively high VAFs can be distinguished from heterozygous germline mutations when high sequencing depth allows a more precise estimate of their VAFs. Ideally, multiple tissues from the same individual should be examined to confirm the somatic nature of a mutation. Single-cell sequencing may also provide confirmation for candidates identified in bulk sequencing; however, given sampling noise a large number of cells may be necessary to capture the cells carrying the mutation of interest for low-VAF mutations.

Concluding Remarks and Future Perspectives

Somatic mutations are being implicated in a growing number of diseases. As our understanding of mutagenic processes in normal cells increases, we will be able to better delineate the extent of somatic mosaicism in healthy individuals and their potential contribution to a wide range of diseases (see Outstanding Questions).

Although methods for the detection and validation of somatic mutation have long been studied in cancer research, characterization of mutation in non-tumor cells presents new challenges due to (i) the orders-of-magnitude lower mutation rates, and (ii) the extremely low frequency of the majority of variants in the absence of selection. Many of the artifacts we have described – sample contamination, damage to DNA *in vitro*, read misalignment, sequencing instrument errors, and platform biases – tend to occur at low allele frequencies and vastly outnumber mosaic mutations. Whereas germline sequencing is typically done at ~30× (thus an average of ~15 reads supporting a heterozygous variant), the same level evidence for a low-frequency somatic variant would require an amount of sequencing that is currently impractical unless confined to a small region.

Thus, bioinformatics algorithms that incorporate refined filtering criteria will be key for improved sensitivity and specificity in mutation detection. Recent advances in machine-learning algorithms, for instance, offer the possibility that various features related to the supporting reads and their configurations could be combined more efficiently for higher prediction accuracy. Experimental and computational methods are still being developed for single-cell approaches, but they will be essential for detailed analysis of how mutations arise *de novo*.

Outstanding Questions

What is the role of somatic mosaicism in human evolution and human diseases?

What are the best bioinformatic approaches for identifying somatic mutations, especially when matched controls are not available? Could we use haplotype phasing to improve variant identification?

What are the common artifacts that confound detection of mosaic variants, and how do we mitigate their effect? Which methods should be used to validate mosaic mutations?

References

1. Martincorena, I. and Campbell, P.J. (2015) Somatic mutation in cancer and normal cells. *Science* 349, 1483–1489
2. Ju, Y.S. *et al.* (2017) Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. *Nature* 543, 714–718
3. Martincorena, I. *et al.* (2017) Universal patterns of selection in cancer and somatic tissues. *Cell* 171, 1029–1041
4. Cibulskis, K. *et al.* (2013) Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* 31, 213–219
5. Lodato, M.A. *et al.* (2015) Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science* 350, 94–98
6. Behjati, S. *et al.* (2014) Genome sequencing of normal cells reveals developmental lineages and mutational processes. *Nature* 513, 422–425
7. Martincorena, I. *et al.* (2015) High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* 348, 880–886
8. Sims, D. *et al.* (2014) Sequencing depth and coverage: key considerations in genomic analyses. *Nat. Rev. Genet.* 15, 121–132
9. Welch, J.S. *et al.* (2012) The origin and evolution of mutations in acute myeloid leukemia. *Cell* 150, 264–278
10. Blokzijl, F. *et al.* (2016) Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* 538, 260–264
11. Abyzov, A. *et al.* (2017) One thousand somatic SNVs per skin fibroblast cell set baseline of mosaic mutational load with patterns that suggest proliferative origin. *Genome Res.* 27, 512–523
12. Bae, T. *et al.* (2018) Different mutational rates and mechanisms in human cells at pregastrulation and neurogenesis. *Science* 359, 550–555
13. Hazen, J.L. *et al.* (2016) The complete genome sequences, unique mutational spectra, and developmental potency of adult neurons revealed by cloning. *Neuron* 89, 1223–1236
14. Mizutani, E. *et al.* (2015) Generation of cloned mice from adult neurons by direct nuclear transfer. *Biol. Reprod.* 92, 81

15. Gawad, C. *et al.* (2016) Single-cell genome sequencing: current state of the science. *Nat. Rev. Genet.* 17, 175–188
16. Hou, Y. *et al.* (2015) Comparison of variations detection between whole-genome amplification methods used in single-cell resequencing. *Gigascienc* 4, 37
17. Huang, L. *et al.* (2015) Single-cell whole-genome amplification and sequencing: methodology and applications. *Annu. Rev. Genom. Hum. Genet.* 16, 79–102
18. Navin, N. *et al.* (2011) Tumour evolution inferred by single-cell sequencing. *Nature* 472, 90–94
19. Wang, Y. *et al.* (2014) Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* 512, 155–160
20. Zhang, C.Z. *et al.* (2015) Chromothripsis from DNA damage in micronuclei. *Nature* 522, 179–184
21. Voet, T. *et al.* (2013) Single-cell paired-end genome sequencing reveals structural variation per cell cycle. *Nucleic Acids Res.* 41, 6119–6138
22. McConnell, M.J. *et al.* (2013) Mosaic copy number variation in human neurons. *Science* 342, 632–637
23. Evrony, G.D. *et al.* (2015) Cell lineage analysis in human brain using endogenous retroelements. *Neuron* 85, 49–59
24. Lodato, M.A. *et al.* (2018) Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science* 359, 555–559
25. Vitak, S.A. *et al.* (2017) Sequencing thousands of single-cell genomes with combinatorial indexing. *Nat. Methods* 14, 302–308
26. Lan, F. *et al.* (2017) Single-cell genome sequencing at ultra-high-throughput with microfluidic droplet barcoding. *Nat. Biotechnol.* 35, 640–646
27. Bohrson, C.L. *et al.* (2017) Linked-read analysis identifies mutations in single-cell DNA sequencing data. *bioRxiv* Published online October 30, 2017. http://dx.doi.org/10.1101/211169
28. Koboldt, D.C. *et al.* (2009) VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* 25, 2283–2285
29. Saunders, C.T. *et al.* (2012) Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* 28, 1811–1817
30. Ewing, A.D. *et al.* (2015) Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nat. Methods* 12, 623–630
31. Alioto, T.S. *et al.* (2015) A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nat. Commun.* 6, 10001
32. Sherry, S.T. *et al.* (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29, 308–311
33. 1000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* 491, 56–65
34. 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. *Nature* 467, 1061–1073
35. Jung, H. *et al.* (2013) Systematic investigation of cancer-associated somatic point mutations in SNP databases. *Nat. Biotechnol.* 31, 787–789
36. Hiltemann, S. *et al.* (2015) Discriminating somatic and germline mutations in tumor DNA samples without matching normals. *Genome Res.* 25, 1382–1390
37. Lynch, M. (2010) Evolution of the mutation rate. *Trends Genet.* 26, 345–352
38. Rahbari, R. *et al.* (2016) Timing, rates and spectra of human germline mutation. *Nat. Genet.* 48, 126–133
39. Watson, I.R. *et al.* (2013) Emerging patterns of somatic mutations in cancer. *Nat. Rev. Genet.* 14, 703–718
40. Huang, A.Y. *et al.* (2017) MosaicHunter: accurate detection of postzygotic single-nucleotide mosaicism through next-generation sequencing of unpaired, trio, and paired samples. *Nucleic Acids Res.* 45, e76
41. Lim, E.T. *et al.* (2017) Rates, distribution and implications of postzygotic mosaic mutations in autism spectrum disorder. *Nat. Neurosci.* 20, 1217–1224
42. Freed, D. and Pevsner, J. (2016) The contribution of mosaic variants to autism spectrum disorder. *PLoS Genet.* 12, e1006245
43. Dou, Y. *et al.* (2017) Postzygotic single-nucleotide mosaicisms contribute to the etiology of autism spectrum disorder and autistic traits and the origin of mutations. *Hum. Mutat.* 38, 1002–1013
44. Krupp, D.R. *et al.* (2017) Exonic mosaic mutations contribute risk for autism spectrum disorder. *Am. J. Hum. Genet.* 101, 369–390
45. Xi, R. *et al.* (2012) A survey of copy-number variation detection tools based on high-throughput sequencing data. *Curr. Protoc. Hum. Genet.* Chapter 7, Unit 7.19
46. King, D.A. *et al.* (2017) Detection of structural mosaicism from targeted and whole-genome sequencing data. *Genome Res.* 27, 1704–1714
47. Su, S.Y. *et al.* (2010) Inferring combined CNV/SNP haplotypes from genotype data. *Bioinformatics* 26, 1437–1445
48. Palta, P. *et al.* (2015) Haplotype phasing and inheritance of copy number variants in nuclear families. *PLoS One* 10, e0122713
49. Guan, P. and Sung, W.K. (2016) Structural variation detection using next-generation sequencing data: a comparative technical review. *Methods* 102, 36–49
50. Rausch, T. *et al.* (2012) DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* 28, i333–i339
51. Yang, L. *et al.* (2013) Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell* 153, 919–929
52. Ye, K. *et al.* (2016) Systematic discovery of complex insertions and deletions in human cancers. *Nat. Med.* 22, 97–104
53. Wang, J. *et al.* (2011) CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nat. Methods* 8, 652–654
54. Lai, Z. *et al.* (2016) VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res.* 44, e108
55. Lee, E. *et al.* (2012) Landscape of somatic retrotransposition in human cancers. *Science* 337, 967–971
56. Tubio, J.M. *et al.* (2014) Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. *Science* 345, 1251343
57. Helman, E. *et al.* (2014) Somatic retrotransposition in human cancer revealed by whole-genome and exome sequencing. *Genome Res.* 24, 1053–1063
58. Evrony, G.D. *et al.* (2016) Resolving rates of mutation in the brain using single-neuron genomics. *Elife* 5, e12966
59. Evrony, G.D. *et al.* (2012) Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell* 151, 483–496
60. Zheng, G.X. *et al.* (2016) Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat. Biotechnol.* 34, 303–311
61. Schmitt, M.W. *et al.* (2012) Detection of ultra-rare mutations by next-generation sequencing. *Proc. Natl. Acad. Sci. U. S. A.* 109, 14508–14513
62. Hoang, M.L. *et al.* (2016) Genome-wide quantification of rare somatic mutations in normal human tissues using massively parallel sequencing. *Proc. Natl. Acad. Sci. U. S. A.* 113, 9846–9851
63. Browning, S.R. and Browning, B.L. (2011) Haplotype phasing: existing methods and new developments. *Nat. Rev. Genet.* 12, 703–714
64. Delaneau, O. *et al.* (2011) A linear complexity phasing method for thousands of genomes. *Nat. Methods* 9, 179–181

65. Edge, P. *et al.* (2017) HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Res.* 27, 801–812

66. Delaneau, O. *et al.* (2013) Haplotype estimation using sequencing reads. *Am. J. Hum. Genet.* 93, 687–696

67. Ramu, A. *et al.* (2013) DeNovoGear: de novo indel and point mutation discovery and phasing. *Nat. Methods* 10, 985–987

68. Sengupta, S. *et al.* (2016) Ultra-fast local-haplotype variant calling using paired-end DNA-sequencing data reveals somatic mosaicism in tumor and normal blood samples. *Nucleic Acids Res.* 44, e25

69. Sudmant, P.H. *et al.* (2015) An integrated map of structural variation in 2,504 human genomes. *Nature* 526, 75–81

70. Chaisson, M.J. *et al.* (2017) Multi-platform discovery of haplotype-resolved structural variation in human genomes. *bioRxiv* Published online September 23, 2017. http://dx.doi.org/10.1101/193144

71. Kuleshov, V. *et al.* (2014) Whole-genome haplotyping using long reads and statistical methods. *Nat. Biotechnol.* 32, 261–266

72. Weisenfeld, N.I. *et al.* (2017) Direct determination of diploid genome sequences. *Genome Res.* 27, 757–767

73. Seo, J.S. *et al.* (2016) De novo assembly and phasing of a Korean human genome. *Nature* 538, 243–247

74. Cibulskis, K. *et al.* (2011) ContEst: estimating cross-contamination of human samples in next-generation sequencing data. *Bioinformatics* 27, 2601–2602

75. Flickinger, M. *et al.* (2015) Correcting for sample contamination in genotype calling of DNA sequence data. *Am. J. Hum. Genet.* 97, 284–290

76. Jun, G. *et al.* (2012) Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am. J. Hum. Genet.* 91, 839–848

77. Kim, S. *et al.* (2013) Virmid: accurate detection of somatic mutations with sample impurity inference. *Genome Biol.* 14, R90

78. Su, X. *et al.* (2012) PurityEst: estimating purity of human tumor samples using next-generation sequencing data. *Bioinformatics* 28, 2265–2266

79. Bergmann, E.A. *et al.* (2016) Conpair: concordance and contamination estimator for matched tumor-normal pairs. *Bioinformatics* 32, 3196–3198

80. Lee, S. *et al.* (2017) NGSCheckMate: software for validating sample identity in next-generation sequencing studies within and across data types. *Nucleic Acids Res.* 45, e103

81. Koboldt, D.C. *et al.* (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* 22, 568–576

82. Sinha, R.P. and Hader, D.P. (2002) UV-induced DNA damage and repair: a review. *Photochem. Photobiol. Sci.* 1, 225–236

83. Fryxell, K.J. and Zuckerkandl, E. (2000) Cytosine deamination plays a primary role in the evolution of mammalian isochores. *Mol. Biol. Evol.* 17, 1371–1383

84. Evans, M.D. *et al.* (2004) Oxidative DNA damage and disease: induction, repair and significance. *Mutat. Res.* 567, 1–61

85. Helleday, T. *et al.* (2014) Mechanisms underlying mutational signatures in human cancers. *Nat. Rev. Genet.* 15, 585–598

86. Costello, M. *et al.* (2013) Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res.* 41, e67

87. Chen, L. *et al.* (2017) DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. *Science* 355, 752–756

88. Dong, X. *et al.* (2017) Accurate identification of single-nucleotide variants in whole-genome-amplified single cells. *Nat. Methods* 14, 491–493

89. Treangen, T.J. and Salzberg, S.L. (2011) Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.* 13, 36–46

90. Huang, A.Y. *et al.* (2014) Postzygotic single-nucleotide mosaicisms in whole-genome sequences of clinically unremarkable individuals. *Cell Res.* 24, 1311–1327

91. Simpson, J.T. and Durbin, R. (2012) Efficient de novo assembly of large genomes using compressed data structures. *Genome Res.* 22, 549–556

92. Iqbal, Z. *et al.* (2012) De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nat. Genet.* 44, 226–232

93. Snyder, M.W. *et al.* (2015) Haplotype-resolved genome sequencing: experimental methods and applications. *Nat. Rev. Genet.* 16, 344–358

94. Ross, M.G. *et al.* (2013) Characterizing and measuring bias in sequence data. *Genome Biol.* 14, R51

95. Dohm, J.C. *et al.* (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* 36, e105

96. Meacham, F. *et al.* (2011) Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinform.* 12, 451

97. Allhoff, M. *et al.* (2013) Discovering motifs that induce sequencing errors. *BMC Bioinform.* 14 (Suppl. 5), S1

98. Nakamura, K. *et al.* (2011) Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res.* 39, e90

99. Keohavong, P. and Thilly, W.G. (1989) Fidelity of DNA polymerases in DNA amplification. *Proc. Natl. Acad. Sci. U. S. A.* 86, 9253–9257

100. Walsh, P.S. *et al.* (1992) Preferential PCR amplification of alleles: mechanisms and solutions. *PCR Methods Appl.* 1, 241–250

101. Brodin, J. *et al.* (2013) PCR-induced transitions are the major source of error in cleaned ultra-deep pyrosequencing data. *PLoS One* 8, e70388

102. Lelieveld, S.H. *et al.* (2015) Comparison of exome and genome sequencing technologies for the complete capture of protein-coding regions. *Hum. Mutat.* 36, 815–822

103. Sinha, R. *et al.* (2017) Index switching causes 'spreading-of-signal' among multiplexed samples in Illumina HiSeq 4000 DNA sequencing. *bioRxiv* Published online April 9, 2017. http://dx.doi.org/10.1101/125724

104. Robasky, K. *et al.* (2014) The role of replicates for error mitigation in next-generation sequencing. *Nat. Rev. Genet.* 15, 56–62

105. McConnell, M.J. *et al.* (2017) Intersection of diverse neuronal genomes and neuropsychiatric disease: The Brain Somatic Mosaicism Network. *Science* 356, eaal1641

106. Campbell, I.M. *et al.* (2015) Somatic mosaicism: implications for disease and transmission genetics. *Trends Genet.* 31, 382–392

107. Hindson, C.M. *et al.* (2013) Absolute quantification by droplet digital PCR versus analog real-time PCR. *Nat. Methods* 10, 1003–1005

108. McDermott, G.P. *et al.* (2013) Multiplexed target detection using DNA-binding dye chemistry in droplet digital PCR. *Anal. Chem.* 85, 11619–11627

109. Hiatt, J.B. *et al.* (2013) Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. *Genome Res.* 23, 843–854

110. Xu, X. *et al.* (2015) Amplicon resequencing identified parental mosaicism for approximately 10% of 'de novo' SCN1A mutations in children with Dravet syndrome. *Hum. Mutat.* 36, 861–872

111. Froyen, G. *et al.* (2016) Validation and application of a custom-designed targeted next-generation sequencing panel for the diagnostic mutational profiling of solid tumors. *PLoS One* 11, e0154038

112. Nikiforova, M.N. *et al.* (2013) Targeted next-generation sequencing panel (ThyroSeq) for detection of mutations in thyroid cancer. *J. Clin. Endocrinol. Metab.* 98, E1852–E1860

113. Izawa, K. *et al.* (2012) Detection of base substitution-type somatic mosaicism of the NLRP3 gene with >99.9% statistical confidence by massively parallel sequencing. *DNA Res.* 19, 143–152