Computational analysis of cancer genome sequencing data

Isidro Cortés-Ciriano[1], Doga C. Gulhan[2], Jake June-Koo Lee[2], Giorgio E. M. Melloni[2] and Peter J. Park[2] ✉

Abstract | Distilling biologically meaningful information from cancer genome sequencing data requires comprehensive identification of somatic alterations using rigorous computational methods. As the amount and complexity of sequencing data have increased, so has the number of tools for analysing them. Here, we describe the main steps involved in the bioinformatic analysis of cancer genomes, review key algorithmic developments and highlight popular tools and emerging technologies. These tools include those that identify point mutations, copy number alterations, structural variations and mutational signatures in cancer genomes. We also discuss issues in experimental design, the strengths and limitations of sequencing modalities and methodological challenges for the future.

'Driver' mutations
Somatic alterations of the DNA sequence that confer a selective advantage to cells harbouring it.

Mutational signatures
Distinct patterns of mutational spectra, often associated with specific mutational processes.

[1]*European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, UK.*

[2]*Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA.*

✉*e-mail: peter_park@hms.harvard.edu*

https://doi.org/10.1038/s41576-021-00431-y

Analysis of cancer genomes using high-throughput sequencing technologies can provide a comprehensive view of the mutational landscape of human tumours, from single base mutations to chromosomal or whole genome-scale events[1,2]. Exome sequencing and whole-genome sequencing (WGS) have become commonplace for characterizing genomic alterations in research studies, whereas targeted sequencing of selected genes is routine for detecting therapeutically relevant hotspot mutations in clinical practice.

With an increasingly large number of tumour genomes sequenced, researchers have expanded the catalogue of 'driver' mutations across multiple cancer types (reviewed in REF.[3]) including primary[1,2,4–6] and metastatic[7–9] tumours, described a wide range of complex structural variations[2,10], unravelled the mutational patterns shaping tumour evolution and heterogeneity[11–15], examined the role of non-coding mutations in cancer (reviewed in REF.[16]) and elucidated the molecular mechanisms underlying tumour immune evasion and resistance to anticancer therapy[17]. In a recent compendium of papers, investigators in the Pan-Cancer Analysis of Whole Genomes (PCAWG) project[2] within the International Cancer Genome Consortium reported the analysis of WGS data for 2,658 primary tumours spanning 38 tumour types. These papers highlighted the diversity of somatic rearrangements[18–20], extracted a comprehensive set of mutational signatures[21], reconstructed evolutionary histories[22], delineated RNA-level alterations[23], characterized mutations in mitochondrial DNA[24], examined somatic retrotransposition events[25], detected associations with viruses[26] and identified chromothripsis events[19] (FIG. 1).

Accurate characterization of somatic alterations from genome sequencing data is critical for deriving biological insights. Despite the advances in algorithm development[27,28], applications of different methods often give discordant results, especially for variants present in a small fraction of cells or variant types other than single-nucleotide substitutions. Several fundamental challenges are also associated with current genome sequencing technologies. For example, short-read sequencing (reads of 100–150 bp) on the dominant Illumina platform poses inherent constraints in reconstructing the tumour genome, especially in repetitive regions and for complex structural alterations[29,30]. Given the cost of sequencing, current assays also impose a trade-off between the depth of sequencing and genome coverage — WGS offers whole-genome coverage but with limited depth (typically 30–60×), making variants present in a small fraction of cells hard to detect, whereas targeted (or 'panel') sequencing of a subset of genes offers high depth (typically 500–1,000×) but low sensitivity for copy number variants (CNVs) and misses nearly all structural variants (SVs). Exome sequencing offers a compromise (typically a depth of ~100–200×), although it is becoming less common as the cost of WGS has decreased.

In this Review, we provide an overview of the algorithms designed for detecting somatic alterations using genome sequencing data, highlighting innovations made in recent years. We discuss the limitations of current sequencing technologies and the extent to which these can be addressed by the algorithmic paradigms utilized today. Several topics specific to clinical applications, such as the interpretation of germline and somatic variants and techniques for detecting predetermined hotspot mutations, and emerging areas such as computational immuno-oncology, single-cell technologies, long-read

Mapping quality
A measure of confidence that a sequencing read originated from the aligned position.

Single-nucleotide variants
(SNVs). Changes in the sequence of the DNA involving one base pair.

Variant allele fraction
(VAF). The number of reads supporting a candidate mutation divided by the read depth at that position.

Read depth
Number of sequenced reads at a genomic position.

Tumour purity
The fraction of cancer cells in the sequenced sample.

sequencing and cell-free DNA analysis, are not discussed as they have been recently reviewed elsewhere[31–38].

Preprocessing sequencing reads

The first step in sequencing data analysis consists of mapping (aligning) the sequenced reads, usually delivered from sequencing facilities in FASTQ or BAM format[39], to the human reference genome. To reduce the storage footprint (which is ~100 GB for a 30× genome), reference-based compression (which reduces the footprint to 30–50% of the original size) using an open format called CRAM[40] is becoming common. The quality of sequencing data has improved substantially with the successive generations of sequencing platforms, with more consistency between machine runs and lower per-base error rates. One technical problem on some sequencers was the occurrence of 'index hopping' with multiplexed libraries, in which a tiny fraction of DNA fragments were assigned to incorrect samples, but its effect can be mitigated by dual indexing[41]. The quality of sequencing runs is assessed by inspecting biases in the distribution of base quality scores and base composition along the sequencing reads, using tools such as FastQC[42]. To assess mapping quality, the fraction of reads that align in the expected orientation to the reference genome, the insert size and read length distributions, and the fraction of duplicate reads (for example, due to PCR amplification artefacts and low library complexity) are among the common metrics. Several tools, including Alfred[43] and Qualimap 2 (REF.[44]), can assess sequencing and mapping quality for large sequencing data sets.

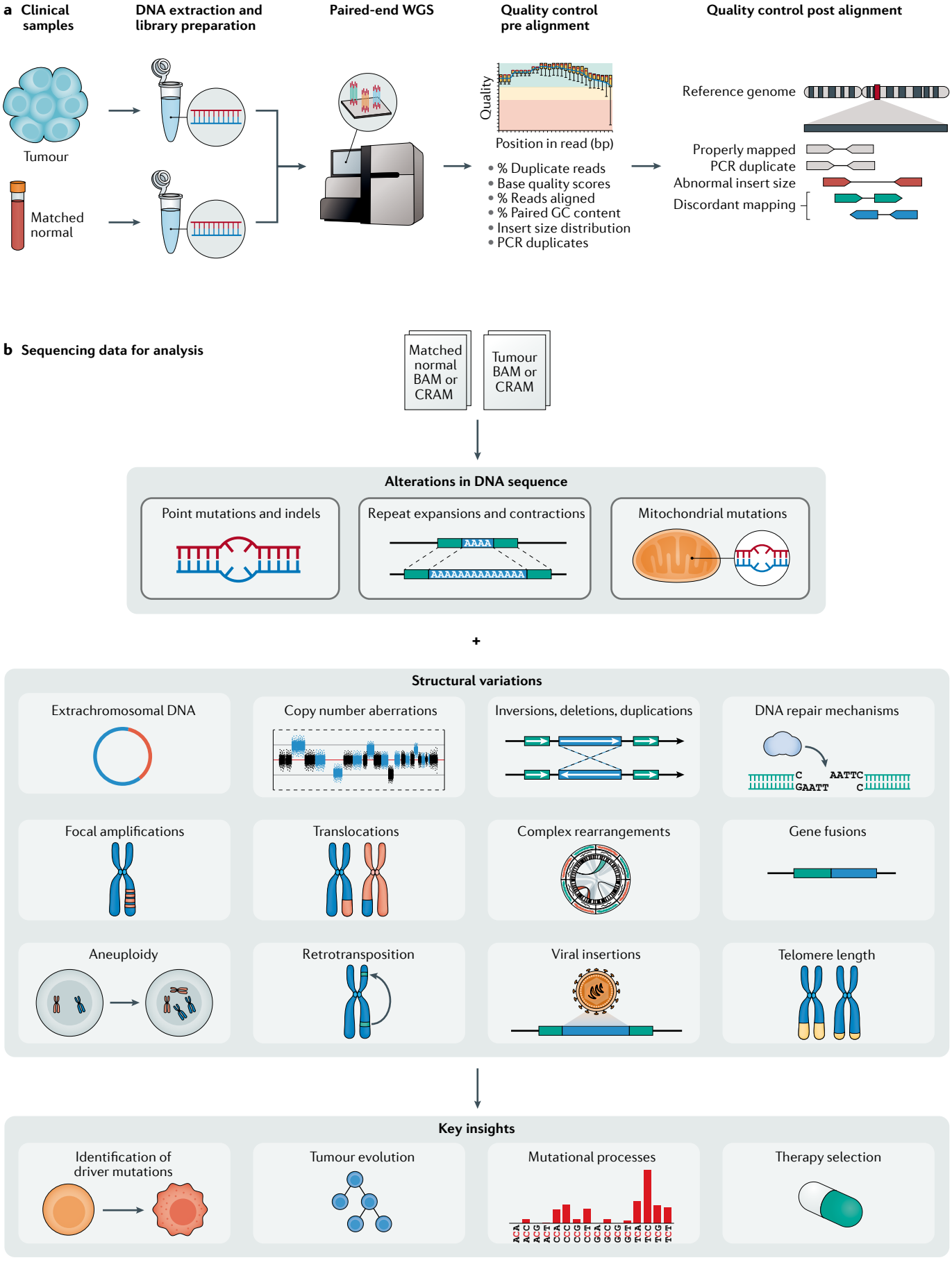
Mapping short reads to the human genome sequence is complex and requires heuristic approaches. The most popular aligner for cancer genome analysis is BWA-MEM[45]. This algorithm can efficiently align relatively long reads (from 70 bp to a few hundred base pairs) against the human genome, supporting paired-end reads and chimeric alignment while being robust to mismatches. As mapping reads from diverse individuals to a single linear reference genome does not adequately account for genetic variation in the population, the next generation of alignment strategies should involve graph-based representations of polymorphic regions across individuals and de novo reconstruction of regions that are altered in the cancer genome, ideally assisted by long-read technologies[46]. For RNA sequencing (RNA-seq) data, splicing events and diverse isoforms must be considered when mapping to the human genome or transcriptome, and STAR[47] and HISAT[48] are among the popular aligners. Furthermore, with improved accuracy and reduced cost, long-read sequencing is becoming more popular for characterizing SVs, and minimap2 (REF.[49]) is a common aligner for mapping long reads.

Alignment can be confounded by biological variability (for example, polymorphisms), sequencing errors, segmental duplications, repetitive sequences and an incomplete reference genome, all of which could result in ambiguous or mistaken mapping. For example, when mapping sequencing reads to a version of the human reference genome lacking the 'decoy' sequences (which include parts of the human genome that were not assembled into the reference and viral sequences such as the Epstein–Barr virus genome), the fragments corresponding to the decoy segments are often misaligned to the reference. When reads are misaligned, the mismatched bases may be incorrectly identified by a detection algorithm as a somatic mutation. Sophisticated post-alignment filtering steps are needed to reduce such false positives (see below).

Genome alignment requires choosing a reference genome version. The Genome Reference Consortium released the most recent build of the human reference genome, GRCh38, in 2013. GRCh38 is a more complete reference than previous versions, and includes the refinement of thousands of nucleotides, a greater number of alternative contigs for regions that are too complex to be represented by a single sequence (for example, HLA loci), modelled centromeric sequences and gap closures[50]. These enhancements improve alignment quality, including for some transcripts in clinically relevant genes[50]. Yet adoption of each new assembly is slow, due to possible incompatibility with previous work and the amount of effort required to reprocess existing data and update databases. The PCAWG, for instance, used the hs37d5 assembly that was developed by the 1000 Genomes Project[51], whereas The Cancer Genome Atlas (TCGA) utilized GRCh37, although the official NIH (National Institutes of Health) repository (Genomic Data Commons) realigned TCGA data to GRCh38 recently. Differing versions of genomes can lead to differences in analysis output, especially in repetitive regions, but these are relatively minor[52] compared with the impact of different analytical tools. With the availability of gapless, telomere-to-telomere assemblies[53], it may be possible to reach a stable human reference genome in the near future.

Detecting SNVs and indels

Algorithms for variant detection. Single-nucleotide variants (SNVs) and small (<50 bp) insertions and deletions (indels) are the most common alterations found in tumours[54,55]. In this article, we use the term SNVs to refer to somatic SNVs, although the term 'variant' is used in the literature for somatic or germ line depending on the variant type and the context. In a typical tumour–normal paired design, candidate somatic mutations are identified as genomic positions for which an alternate allele supported by tumour reads is not present in the matched normal sample (in the absence of tumour-in-normal contamination). The variant allele fraction (VAF), which is the number of reads supporting a candidate mutation divided by the read depth at that position, is a key determinant in finding a somatic variant. For germline variants, the VAF is close to 0.5 for heterozygous variants, and ~1 for homozygous variants. For somatic variants, the VAF is often substantially lower than 0.5 and depends on the tumour purity, copy number at the candidate location and intra-tumour heterogeneity (FIG. 2). For instance, a subclonal mutation present in 20% of the diploid tumour cells in a 60× sample should have ~6 reads containing the variant, but this would be halved to ~3 reads if the tumour purity is 50%. Depending on the cancer type and mode of tissue extraction (for


a Clinical samples
DNA extraction and library preparation
Paired-end WGS
Quality control pre alignment
Quality control post alignment
Tumour
Matched normal
Quality
Position in read (bp)
• % Duplicate reads
• Base quality scores
• % Reads aligned
• % Paired GC content
• Insert size distribution
• PCR duplicates
Reference genome
Properly mapped
PCR duplicate
Abnormal insert size
Discordant mapping
b Sequencing data for analysis
Matched normal BAM or CRAM
Tumour BAM or CRAM
Alterations in DNA sequence
Point mutations and indels
Repeat expansions and contractions
AAAA
AAAAAAAAAAAAAA
Mitochondrial mutations
+
Structural variations
Extrachromosomal DNA
Copy number aberrations
Inversions, deletions, duplications
DNA repair mechanisms
C
AATTC
GAATT
C
Focal amplifications
Translocations
Complex rearrangements
Gene fusions
Aneuploidy
Retrotransposition
Viral insertions
Telomere length
Key insights
Identification of driver mutations
Tumour evolution
Mutational processes
Therapy selection

◀ Fig. 1 | **Workflow for cancer genome analysis. a** | Detection of somatic mutations in human cancers involves the extraction of DNA from a tumour and, ideally, a matched normal sample. Sequencing is usually performed in the paired-end mode with read lengths of 100–150 bp. Quality control is performed on the sequencing reads by assessing several metrics before and after alignment to the reference genome. **b** | Computational analysis of whole-genome sequencing (WGS) data, provided in BAM format or in a compressed format called CRAM, enables the detection of not only sequence alterations (such as point mutations and indels) but also many types of structural variations. Additional analysis of somatic alterations allows the identification of driver events, the mutational processes operative in cancer, patterns of tumour evolution and biomarkers of response to therapy.

example, needle biopsy), tumour purity could be as low as 10–20%. Thus, high-depth sequencing or a sophisticated algorithm (preferably both) is needed for identifying mutations with a low VAF in tumour samples. For a comprehensive discussion on experimental design considerations and analytical strategies for SNV detection, see REF.[56].

Assessing whether alternate alleles supported by sequencing reads represent true mutations or artefacts is the core task in variant calling. The large majority of tools (reviewed in REF.[57]) make use of a matched normal sample, typically blood, as a control. Many of the initial variant calling methodologies were heuristic approaches utilizing hard thresholds, for example, requiring a minimum number of reads or an allele fraction over a specific value. More advanced algorithms such as MuTect2 (REF.[58]), CaVEMan[59], Strelka2 (REF.[60]), VarDict[61] and MuSE[62] estimate the genotype of both the normal and the tumour sample jointly, with prior probabilities for the genotypes or allele frequencies adjusted for factors such as expected mutation rates or mutation types (that is, transitions versus transversions), tumour purity and tumour ploidy, and local features such as copy number. Although many research studies have a tumour–normal paired design, a normal tissue is often not available in clinical applications. In such cases, variant calls are made using the tumour sample only and efforts are made afterwards to remove germline variants. A standard approach to remove germline variants is to filter annotated single-nucleotide polymorphisms (SNPs) found in dbSNP[63] (the NCBI (National Center for Biotechnology Information) database of genetic variation) as well as additional population variants in a large set of exomes or genomes such as those in the Genome Aggregation Database (gnomAD)[64]. However, this database-based filtering approach is often inadequate because it does not capture SNVs incorrectly identified due to the particularities of the specific analytical workflow used (for example, misalignment due to the aligner used), or population-specific germline variants not contained in the database. Thus, using a 'panel of normals' collected from other studies and processed in the same way as the tumour samples is an effective approach for removing germline variants. Identical sequencing protocols and data processing of the samples is essential for removing false positives — on one occasion, we traced some false positive variants to subtle alignment differences that arose between 150 bp reads from tumour samples and 100 bp reads from normal samples (P.J.P., unpublished observations). Without matched normals, some individual-specific germline variants will still be misclassified as somatic variants[65], but the number of such variants should be small for a sufficiently large panel of normals.

For stringent filtering, sites with any reads supporting alternate alleles in a normal sample may be filtered as potential germline variants or artefacts. On rare occasions, true SNVs may overlap with known polymorphic sites. To avoid such variants from getting filtered, one could keep those with sufficient coverage in the tumour but no support in the normal sample[2]. It is also possible to distinguish between somatic and germline variants by employing a machine learning model that integrates features gathered from large collections of known somatic (for example, COSMIC)[66] and germline (for example, dbSNP)[63] variants.

Comparison of algorithms and variant filtering. Numerous studies have found variable agreement when comparing the performance of existing mutation calling methods[67–71]. The ICGC-TCGA DREAM Somatic Mutation Calling challenge used simulated cancer genomes to benchmark pipelines[70]; other benchmarking efforts focused on sequencing data from clinical samples[2,71]. Analysis of high-depth (~300×) paired tumour–normal sequencing data showed that precision and recall strongly varied depending on the combination of aligner and caller used[71]. The library preparation protocol used also profoundly influenced the accuracy of the mutation calls; PCR-free libraries performed best, as they produced the most uniform coverage across the genome and were less affected by GC content bias than PCR-based libraries. Filtering out variants supported by soft-clipped or low mapping quality reads, or that showed positional or strand bias, improved the detection of SNVs. To accurately detect subclonal mutations, reaching a sequencing depth of at least 100× for the tumour sample was recommended[67,71]. The validation of 13 SNV calling pipelines on 50 tumour genomes, through high-depth targeted sequencing of thousands of candidate mutations, showed[2] that mutation detection algorithms generally exhibit high specificity but variable sensitivity (ranging from 0.10 to >0.95), especially for variants with a low VAF.

Several studies showed that combining the output of multiple callers decreases the number of false positives[71,72]. Ensemble methods based on simple rules (for example, keep only mutations called by at least a minimum number of algorithms[2,73] or majority vote[70]) or weighted schemes have been used to derive consensus call sets in large cancer genomics projects[1,2] and in benchmarking studies[70,72]. For instance, the consensus call set in the PCAWG consisted of SNVs detected by at least two out of four algorithms: CaVEMan, MuSE, MuTect2 and SAMtools[45]. This 'wisdom of crowds' approach is reasonable for gathering a high-confidence set of variants, but may be too conservative, thus sacrificing sensitivity for increased specificity. To improve both sensitivity and specificity, more complex strategies using machine learning have been proposed[74–78].

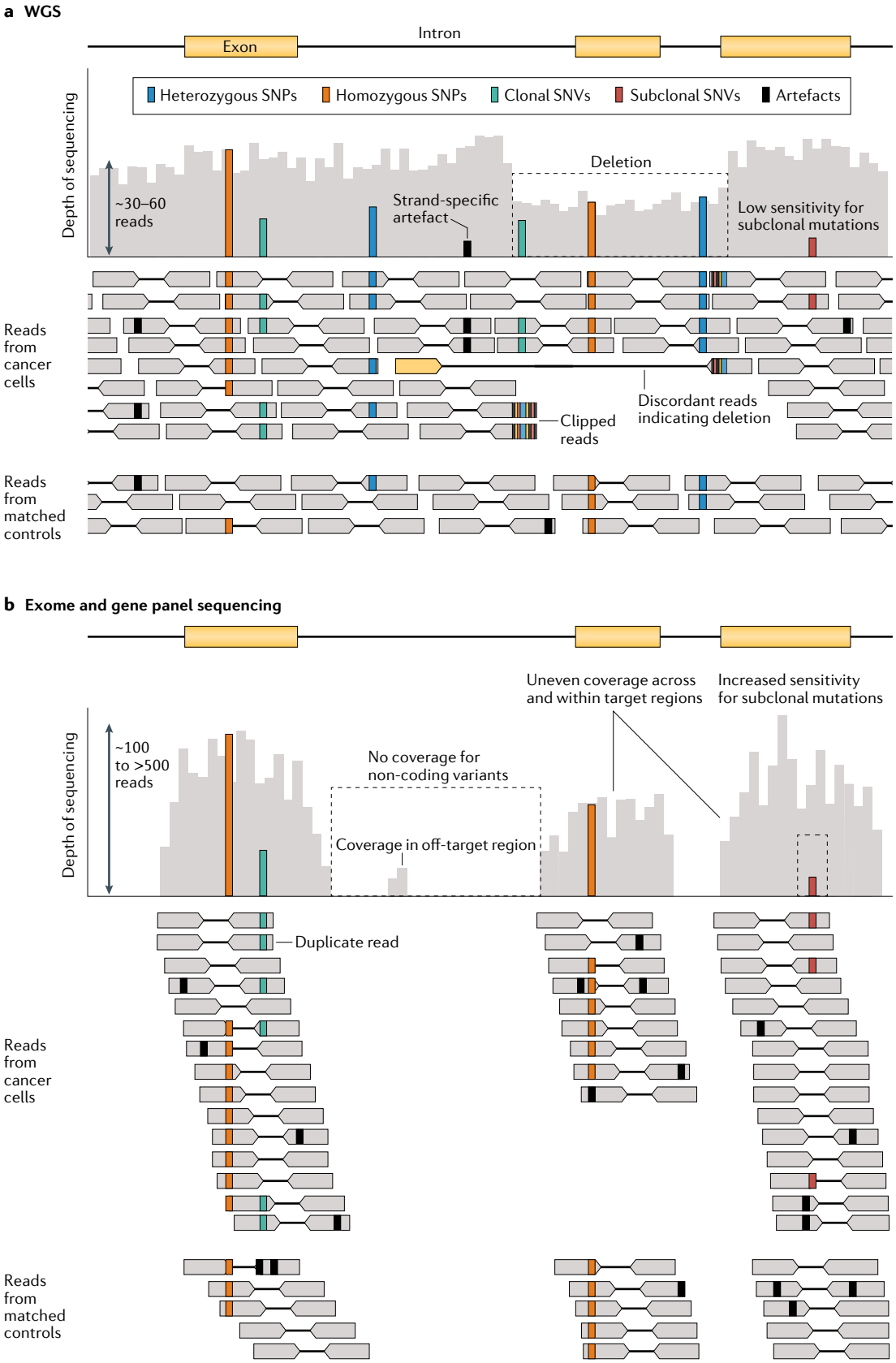
Despite the increased sophistication in the statistical models, an essential component of the variant identification process is a set of ad hoc filters designed to

Tumour ploidy
The amount of DNA that cancer cells contain, usually estimated for the major clone in a tumour sample.

Panel of normals
A set of 'normal' samples that are used as a control to remove germline variants in a population as part of somatic variant calling.

GC content
Percentage of bases in a genomic region that are either guanine (G) or cytosine (C).

Strand bias
In the context of variant calling, the presence of the variant allele in either forward or reverse reads with a frequency higher than expected for binomial sampling.



remove the likely false positive calls. These filters include strand bias (for example, supporting reads are all in the reverse strand) (FIG. 2), the mapping quality of supporting reads, the presence of nearby indels or multiple alternative alleles and whether the read alignment patterns in the genomic region are 'noisy'. These factors cannot be captured effectively in a standard linear model. As a result, visual inspection of the aligned read patterns in a

◀ Fig. 2 | **Identifying variants and artefacts in sequencing reads. a** | Whole-genome sequencing (WGS) provides nearly uniform depth of coverage across the genome. By comparing WGS reads from cancer cells and matched controls, clonal single-nucleotide variants (SNVs) can be reliably detected both in coding and non-coding regions, while filtering out germline variants including common single-nucleotide polymorphisms (SNPs). However, subclonal SNVs may be missed when the number of reads containing the mutation is too small or comparable with the number of reads containing artefacts. The discordant reads are mapped further apart than expected, indicating a deletion; in the 'clipped' or 'split' reads, a portion of the read (coloured bars) maps to another location. **b** | Exome sequencing (~100–200×) and gene panel sequencing of cancer-related genes (~500–1,000×) have higher depth of sequencing than WGS, resulting in increased sensitivity to detect subclonal mutations. However, the coverage within and across target regions is more variable than with WGS, making copy number estimation more difficult.

genome browser and/or experimental validation remains a time-consuming but often necessary step for obtaining a comprehensive set of mutations[71]. Given the effectiveness of visual review, a natural approach is to utilize machine learning techniques to incorporate the features that are informative to the reviewer but not easily captured by simple models. In DeepVariant, a read 'pile-up' image of the candidate site is fed into a convolutional neural network (CNN) to detect germline variants and indels[79]. Similar approaches have been developed subsequently for detecting somatic mutations[80,81]. The main difficulty with CNN or other machine learning methods is obtaining a large amount of high-quality training data. Whereas known SNPs can be used to train a model for germline variants, the training data for SNVs are harder to obtain. In some cases, read-based phasing (for example, as used in MosaicForecast[82]) or other schemes may be needed to procure training data. Nonetheless, these data-driven formulations permit learning of complex patterns of errors and biases and should increase sensitivity and specificity[79].

Detecting indels is substantially more difficult than detecting SNVs, as alignment issues are more intricate[83]. The preferred strategy for detecting indels, therefore, is to reassemble and realign the reads based on the most likely local haplotype, as done in GATK HaplotypeCaller[84] (for SNPs) or MuTect2 (for SNVs), SvABA[85], Platypus[86] and other methods. Although the integration of indel calls from multiple methods increases performance, the accuracy remains low. In the PCAWG Consortium, for instance, consensus indel calls had precision of 91% and sensitivity of 60%; in comparison, consensus SNV calls had precision and sensitivity of 95% and 97.5%, respectively.

Identification of cancer drivers. The availability of large collections of cancer genomes permits the identification of cancer driver mutations using data-driven methods, as recently reviewed[3]. A higher than expected frequency of a mutation across samples is generally interpreted as evidence for positive selection, and such alterations are considered driver events. Over the years, the calculation of the background mutation frequency has improved considerably. The earliest, most naïve model had assumed a constant background mutation rate along the genome; an improved model used a synonymous mutation count on each gene to estimate the background[87]. Currently, more sophisticated methods, such as MutSig2CV[88] and dNdScv[14], utilize statistical models that consider several variables that correlate with mutation frequency, including transcriptional activity, the timing of DNA replication[88], epigenetic context (that is, open versus closed chromatin)[89,90], DNA repair activity[91] and the sites of DNA-binding proteins[92–94]. Other tools, such as MutPanning[95], also utilize the identification of mutational processes, which have variable preferences for different trinucleotides in specific contexts, to inform driver-gene identification[95–97]. In addition to mutation recurrence, other signals of positive selection for identifying driver genes include unexpected clustering of mutations in the linear amino acid sequence[98], and the enrichment of mutations in 3D positions that are functionally important for the protein[99,100], in regulatory elements[101,102] and in evolutionarily conserved regions[103]. Recently, the cytidine deaminase APOBEC3A was shown to generate up to 200 times more mutations at DNA hairpin substrates than at non-hairpin sites, causing 'hotspot' mutations that are not driver events[104]. Additional work is needed to identify other sources of recurrent but non-driver mutations in cancer. For all statistical models, proper examination of the *p*-value distribution using quantile–quantile plots and accurate estimation of statistical significance using false discovery rates are essential.

Prioritizing the observed somatic mutations based on their potential functional consequences is challenging. The choice of the transcript database used (for example, RefSeq, UCSC, GENCODE or ENSEMBL) can have a major impact on annotation because the same mutation can have different effects on different transcripts. Variant Effect Predictor (VEP)[105], SnpEff[106] and ANNOVAR[107] are among the most popular annotators and report the effects of mutations on transcripts, although discrepancies exist among annotators especially for mutations that fall within splice sites and non-coding DNA[108,109]. These tools often check databases such as the Catalogue of Somatic Mutations In Cancer (COSMIC)[66], ClinVar[110] and OncoKB[111] to determine whether the mutation was found in previous cancer studies or whether it has been shown to be pathogenic and potentially actionable. The functions of non-coding variants are much more difficult to infer than those of coding variants, but are informed by cell type-specific genome annotations such as open chromatin regions (for example, from DNase I hypersensitive sites or peaks identified using ATAC-seq (assay for transposase-accessible chromatin using sequencing)), regulatory regions (for example, from the acetylated histone H3 lysine 27 (H3K27ac) histone mark for enhancers) and matched transcriptome data[112,113].

Sources of errors in SNV and indel identification. In addition to the misclassification of germline variants as somatic variants, there are other potential sources of artefacts in identifying SNVs and indels. First, DNA amplification errors can result in artefacts. Until the PCR-free library preparation technique became available, sequencing library preparation often involved PCR amplification to generate a sufficiently large amount of DNA for sequencing. When the amount of DNA extracted is very small (<0.1 µg, for instance), PCR is still used. PCR amplification leads to increased levels of stutter noise at repetitive regions and additional GC bias, and can create single-nucleotide artefacts[114]. In other

Read 'pile-up'
Text-based format that represents the base calls in sequencing reads aligned to a reference genome.

Quantile–quantile plots
A graphical method used to compare two probability distributions by plotting the quantiles of one distribution against the same quantiles of a second distribution.

cases, whole-genome amplification (WGA) by multiple displacement amplification or PCR-based techniques was used prior to sequencing. Early TCGA data sets suffer from such WGA-related artefacts that must be filtered out carefully[115].

Second, oxidative damage, such as the oxidation of guanine to 8-oxoguanine during DNA shearing, also generates artefacts such as the low-frequency transversions C > A within CCG > CAG[116]. These lesions might be mistakenly identified as true mutations if not flagged by computational tools specifically designed for their detection, such as D-ToxoG[116], or by the requirement for a candidate mutation to be supported by both strands.

Third, although many research studies, such as TCGA, were based on fresh-frozen tissues, many clinical samples are archived by embedding in paraffin and tissue fixation can generate artefacts. Indeed, in such formalin-fixed paraffin-embedded (FFPE) tissues, artefacts may arise from DNA fragmentation and base changes induced by formaldehyde, especially the deamination of cytosine into thymines at methylated CpG sites[117]. For FFPE samples, specialized experimental protocols and algorithms must be used for variant detection[118–121].

Finally, whereas the inclusion of non-cancerous cells from the same individual in the tumour sample reduces sensitivity for somatic variant detection, cross-individual contamination — such as a tumour sample being mistakenly paired with a normal sample from a different individual — is more serious in terms of false positive calls. Specific tools to assess cross-individual contamination, including ConEst[122], ART-DeCO[123] and Conpair[124], estimate the probability of contamination based on the allele fraction of homozygous polymorphisms. Algorithms to identify sample swaps based on the concordance of germline polymorphisms between samples include BAMixChecker[125], NGSCheckMate[126] and HYSIS[127]. In some cancers, a matched 'normal' sample may be contaminated by tumour cells that have invaded healthy tissue. Algorithms such as DeTiN[128] can be used to avoid erroneous filtering of true SNVs in these cases.

Mutational signature analysis

Somatic mutations are caused by specific mutational processes, such as defective DNA repair, imperfect DNA replication and exposure to various mutagens. Many such processes induce specific nucleotide changes, for example, lung cancer in smokers is characterized by the abundance of G > T transversions, and melanoma is characterized by C > T transitions at dipyrimidine sites[21]. Cancer genomes share the footprints of these mutational processes at different relative abundances. From a sufficiently large set of cancer exomes or genomes, it is possible to perform de novo discovery of independent components that underlie the mutational spectrum, with each component referred to as a 'mutational signature' (FIG. 3a). This decomposition is similar to how orchestral sound can be separated into specific components corresponding to musical instruments if the sound is heard long enough with sufficient variation.

In their pioneering work[129], the cancer group at the Sanger Institute represented each single base substitution (SBS) spectrum by a 96-dimensional vector: the 6 canonical substitutions (C > A/G/T, T > A/C/G; their reverse complements are indistinguishable), combined with flanking 5′ and 3′ bases (6 base changes × four 5′ bases × four 3′ bases = 96) (FIG. 3b). Application of a matrix decomposition technique called non-negative matrix factorization (NMF) to the mutation type by sample matrix of thousands of samples enabled the decoupling of the distinct patterns of mutations that are attributable to different underlying mutational processes, along with their relative contribution in each sample (FIG. 3c). In this way, the COSMIC Mutational Signatures catalogue was constructed[130]. This catalogue initially consisted of 30 signatures and has since been extended to more than 50 distinct SBS processes[21]. A similar concept was applied to other types of mutations, including indels[21], rearrangements[131] and copy number alterations (CNAs)[132,133]. Various tools for such de novo signature discovery are available now, utilizing NMF[129,130] or its Bayesian formulation[134] as well as expectation maximization[135], topic models[136] and others. Although NMF is a popular dimensionality reduction technique, it suffers from non-uniqueness; thus, a modified version of NMF with additional constraints should be considered in some cases.

Although the mechanistic origins of many signatures remain to be elucidated, mutational signature analysis has proved invaluable in providing insights into the mechanisms behind observed mutations. Indeed, the mutational signatures of a large variety of exogenous agents, DNA repair deficiencies and therapies have been described[21,137,138]. Characterization of mutational signatures in different tumour types allows a better understanding of the underlying mechanisms of cancer in different tissues. Some SBS signatures, such as the 'clock-like' signatures, have been observed in both normal and malignant tissues[139–142]. Others have a high prevalence in specific tumour types, for example, signatures corresponding to the activation of APOBEC cytidine deaminases are frequent in breast, cervical and bladder cancers, and signatures of exogeneous agents are specific to tissues that are directly exposed to them[21].

Prior knowledge about the mutational signatures in cancer genomes, as summarized in a signature catalogue, allows signatures in a new tumour sample to be estimated without necessitating de novo signature extraction with a large cohort. For this 'refitting' step, restricting the set of signatures under consideration to those known to be active in specific tumour types is helpful, because several signatures in the full catalogue have similar distributions and can result in incorrect assignments. Further measures should also be taken to increase the sparsity of signatures of interest so that signatures can be assigned parsimoniously[143,144]. Several tools are available to perform refitting and use algorithms such as non-negative least squares (NNLS)[145–147]. An obvious limitation of refitting is that the analysis is restricted to known signatures. To overcome this challenge, other methods perform de novo discovery and refitting using known signatures as priors[148–150].

One limitation of the above approaches is that a large number of mutations (at least hundreds) must be

'Clock-like' signatures
Mutational signatures that correspond to those mutations that accumulate in normal somatic cells at a steady rate.

Non-negative least squares
(NNLS). A method for finding the optimal non-negative coefficients for a set of predefined vectors such that their weighted sum is as close as possible to another given vector. In signature analysis, it is used to calculate the contribution of each signature to the mutational spectrum of a sample.

observed per sample to enable signature decomposition to work. Therefore, exome or WGS is required, unless the tumour has an exceptionally large number of mutations (for example, unless it is positive for microsatellite instability or has a mutation in DNA polymerase epsilon). A recent method by our group, SigMA[151], enables signature analysis to be performed from a much smaller number of mutations, thus extending mutational signature

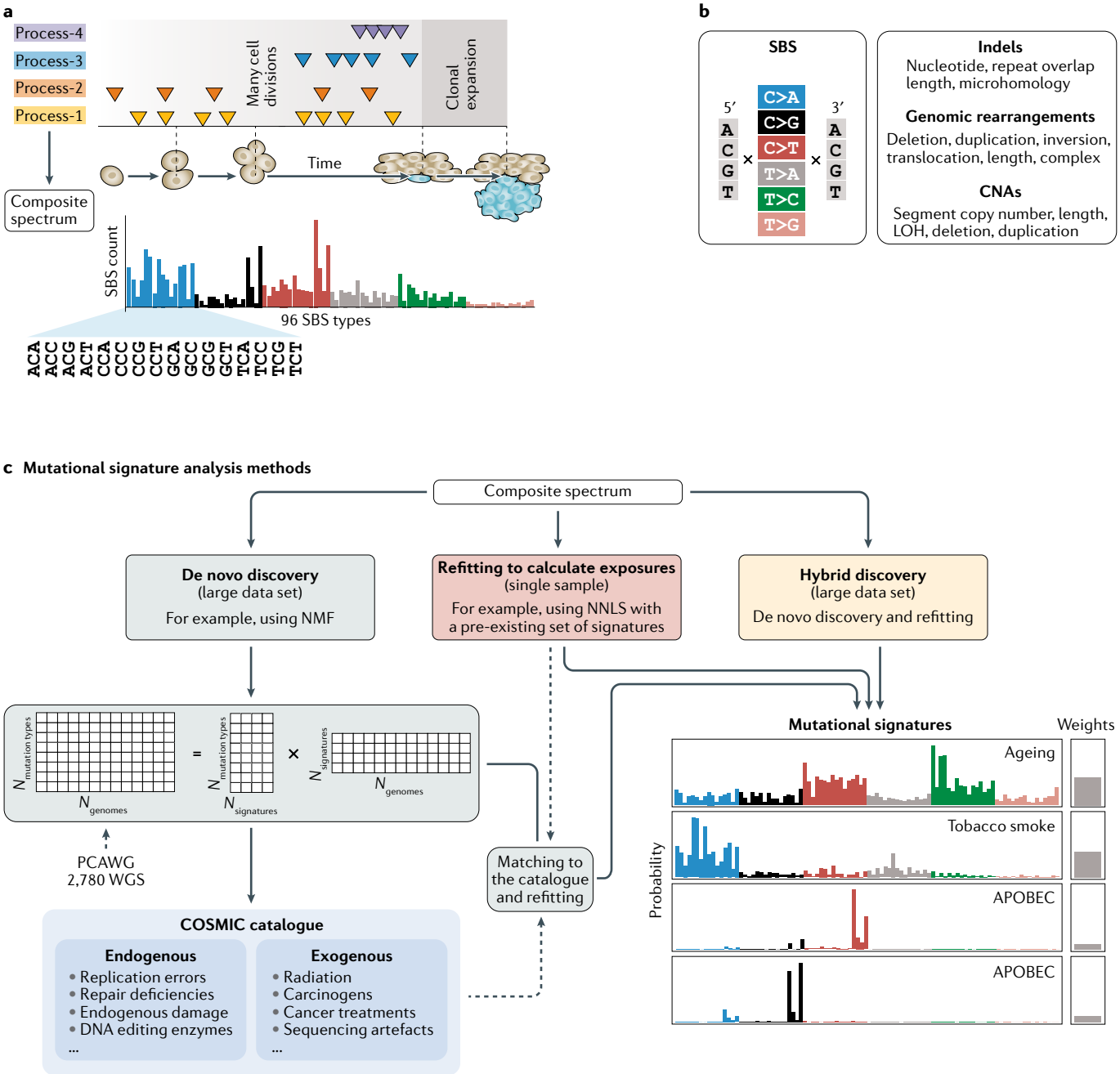


Fig. 3 | **Mutational signature analysis of cancer genomes. a** | Mutations accumulate in the DNA starting from the first cell division. Therefore, mutations detected in a cancer sample using bulk sequencing data reflect the composite effect of multiple mutational processes. The composite mutational spectrum can be described by the counts of single base substitution (SBS) types and their genomic contexts. **b** | An SBS spectrum is commonly represented by a 96-dimensional vector, with the six canonical substitutions (C > A/G/T, T > A/C/G; their reverse complements are indistinguishable) and their flanking 5′ and 3′ bases. Additional spectra based on indels, genomic rearrangements or copy number alterations (CNAs) are also becoming more common. **c** | The mutational signatures associated with distinct mutational processes can be discovered de novo through joint analysis of a large number of cancer genomes. For example, application of non-negative matrix factorization (NMF) on 2,780 whole-genome sequencing (WGS) samples from the Pan-Cancer Analysis of Whole Genomes (PCAWG) project established the latest Catalogue of Somatic Mutations In Cancer (COSMIC) Mutational Signatures database[21]. Identification of mutational processes from a few genomes is underpowered. Thus, mutational signature analysis on a small set of samples is generally performed by estimating the relative contribution for a set of signatures previously defined ('refitting') using methods such as non-negative least squares (NNLS). Hybrid models perform both de novo discovery and refitting using existing signatures as priors. Dashed lines indicate that the refitting procedure and the catalogue are utilized in the matching step. LOH, loss of heterozygosity.

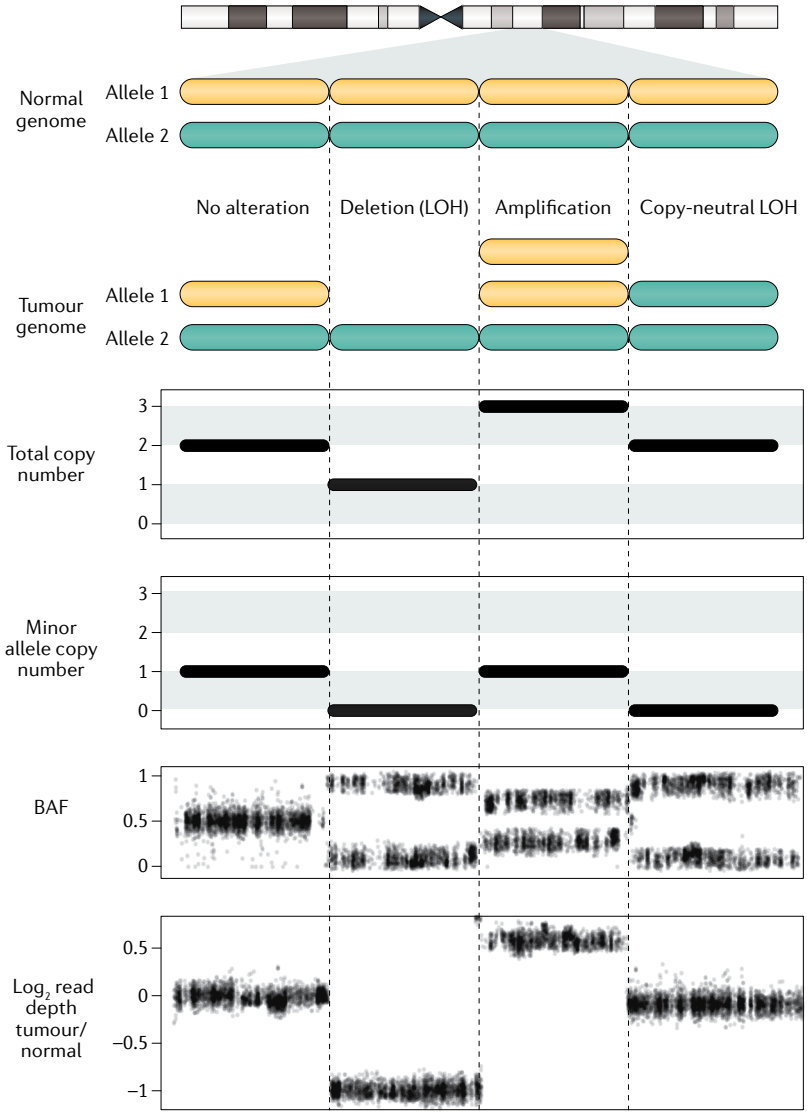


Fig. 4 | **Impact of different copy number alterations on read depth and BAF profiles.** In a diploid region with no alteration, there is one copy of each of the maternal and paternal alleles (alleles 1 and 2), so the total copy number is 2, the minor allele copy number is 1, the B-allele frequency (BAF) profile is centred on 0.5 and the read depth is approximately the same between the tumour and normal samples. In the event of a deletion, the total copy number decreases to 1 and the minor allele copy number to 0. The coverage of that region in the tumour sample drops, and the BAF of heterozygous single-nucleotide polymorphisms becomes either 0 or 1, with deviation towards 0.5 due to the infiltration of normal tissue. An amplification leads to an increase in the total copy number to 3, but the copy number of the minor allele remains 1 because the non-amplified allele is not altered. The BAF profile shifts towards 1/3 and 2/3, corresponding to the fractions of reads covering the non-amplified or amplified allele, respectively. Finally, the total copy number for copy-neutral loss of heterozygosity (LOH) events, where one allele is amplified and the other is lost, is 2, and the minor copy number is 0, reflecting the loss of one allele. Thus, the BAF profile is similar to that of a deletion, although the read depth of the tumour relative to the normal sample is not altered. These examples illustrate the importance of integrating depth of coverage and BAF information to fully characterize copy number alterations.

Enhancer hijacking
Juxtaposition of an active enhancer element from a distant locus into the proximity of another gene, usually caused by a genomic rearrangement, leading to gene activation.

analysis to targeted cancer panels for clinical applications. This method was used to identify patients displaying the signature of a deficiency in homologous recombination, which may serve as an indication for favourable response to poly(ADP-ribose) polymerase (PARP) inhibitors[152].

Analysing somatic structural variations

With the availability of WGS, a great deal of progress has been made in the characterization of somatic SVs in human cancers. The SVs inferred from WGS data range from simple deletions, insertions, duplications, inversions and translocations to copy number changes, transposable element insertions, viral integrations, telomere length variation and complex rearrangements such as chromothripsis. Small deletions and insertions (<50–100 bp) can be identified using individual reads, for example, by gapped alignment, but other SVs require different strategies. Although not discussed in this Review, a large body of work exists on how each type of variant may contribute to cancer development. For example, integrative analysis of translocations and epigenetic data has revealed the phenomenon of enhancer hijacking[153–155] and enhancer amplification[156,157].

As the terms CNVs and SVs are used to refer to either germline or somatic variants in the literature, here we use CNAs (also known as copy number aberrations) to refer to somatic CNVs and 'somatic SVs' (when contrasting with 'germline SVs') or simply 'SVs' to refer to somatic structural alterations. Approaches for investigating germline SVs have been recently reviewed elsewhere[29].

Detecting copy number alterations. Conventional cytogenetic techniques such as fluorescent in situ hybridization (FISH) and spectral karyotyping are useful for the routine diagnosis of genetic disorders and large chromosomal alterations, but their spatial resolution is in the order of megabases. Use of array comparative genomic hybridization (aCGH)[158] and SNP arrays increased spatial resolution to the order of 100 kb[159]. Some of the array-based platforms also offer information on copy-neutral loss of heterozygosity (LOH) events, tumour purity and ploidy[160,161]. WGS can identify CNAs and their underlying SVs, with breakpoints at single-nucleotide resolution.

One class of methods for identifying CNAs from sequencing data adapts the techniques developed for aCGH and SNP arrays, utilizing the 'read depth' along the genome as the main component. These methods segment the genome into regions with distinct copy numbers using hidden Markov models, circular binary segmentation[162], piecewise constant fitting regression[163] or other statistical techniques[164]. Even for a diploid genome, the read depth varies along the genome due to mappability, GC bias and other factors; thus, either a matched control (so that the same biases could be subtracted) or normalization using a statistical model is necessary[164]. More advanced methods for detecting CNAs incorporate the frequencies of minor alleles, termed the B-allele frequency (BAF), which are inferred from heterozygous SNPs, for segmentation, and also detect allele-specific CNAs and copy-neutral LOH events (FIG. 4). The BAF profile, for instance, can help identify copy-neutral LOH events, which have comparable read depth with the non-altered regions of the genome (FIG. 4).

Most segmentation methods can detect large, chromosome-scale CNAs, but often give conflicting copy number profiles at higher resolution[165]. This inconsistency is, in part, because the parameters in each

Enhancer amplification
Increased copy number of regulatory regions (enhancers) that leads to the overexpression of target genes.

Loss of heterozygosity
(LOH). Loss of one allele in biallelic regions, which often results from a somatic deletion.

B-allele frequency
(BAF). The fraction of sequencing reads supporting one allele at a heterozygous single-nucleotide polymorphism (SNP) with respect to the total read depth at that position.

Split reads
Reads containing two contiguous DNA sequences mapping to non-adjacent regions in the reference genome.

Discordant read pairs
Pairs of sequencing reads that do not map to the reference genome with the expected forward–reverse orientation or insert size, suggesting the presence of structural variation.

method are tuned based on specific data sets with certain assumptions, for example, on sequencing coverage, availability of matched controls, tumour purity or the size of the CNA. Careful testing and parameter adjustments for the data set of interest appear to be as important as the underlying algorithm. Given that a change in copy number is a result of a genome rearrangement, incorporating SV information helps identify CNAs more accurately[166]. In the PCAWG Consortium, consensus copy number profiles for WGS were derived from integrating the output of six algorithms (ABSOLUTE, ACEseq, Battenberg, cloneHD, JaBbA and Sclust)[161,163,167–170] using an ad hoc procedure that incorporates SV breakpoints to improve the accuracy of segment boundaries. Methods such as Weaver[171], JaBbA[172] and Reconstructing Cancer Karyotypes[173] use graph-based approaches to jointly model breakpoint, read depth and BAF data, allowing allele-specific reconstruction of cancer genome configurations and of the timing of SVs.

Although WGS offers the best platform for detecting CNAs, CNAs can also be inferred using exome[174–180] or targeted sequencing data[7], albeit with lower accuracy and resolution. Not surprisingly, benchmarking studies often show low concordance of exome-derived copy number profiles between algorithms and in comparison with matched WGS profiles[181,182]. The main challenge for capture-based platforms is the variable read coverage across the genome introduced by the uneven capture efficiency of the probes. Many methods, such as ExomeCNV[177] and CONTRA[179], perform normalization of the read coverage and carry out a segmentation algorithm on the tumour–normal ratio profile. Other methods, such as CopywriteR[183] and CNVkit[184], improve genome-wide profiles by incorporating off-target reads. More recent algorithms, such as Sequenza[185] and FACETS[162], infer allele-specific copy numbers, tumour purity and ploidy simultaneously by fitting both sequencing depth and BAF data. It is also possible to identify chimeric reads from genomic rearrangements in exome data but with low sensitivity[186].

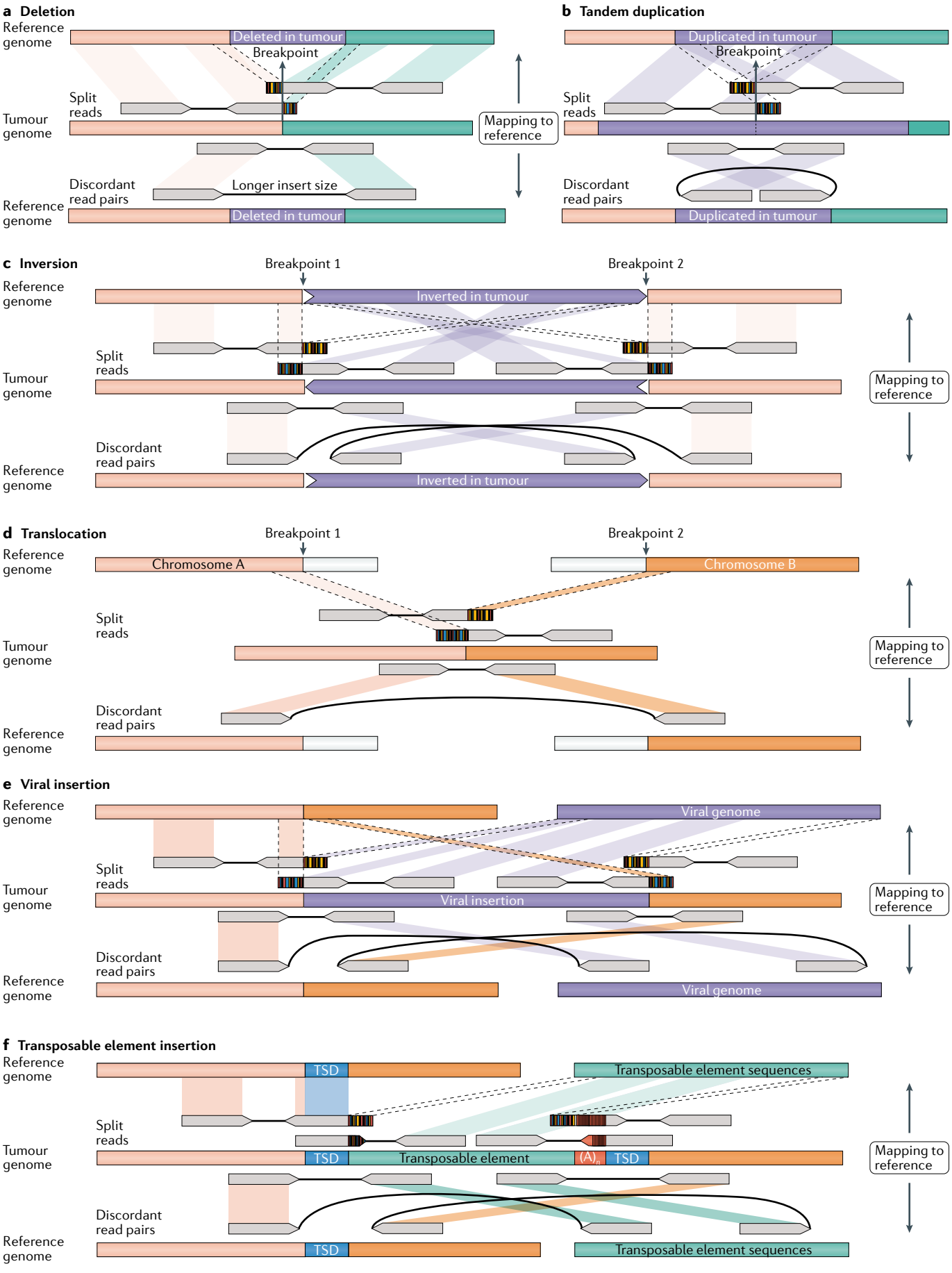
Once CNAs are found in a set of samples, regions altered at a higher than expected frequency are identified using algorithms such as GISTIC[187], as such regions are likely to be driver events under positive selection. To infer the timing of some key events in tumour evolution (see below), the precise copy number for each CNA region must also be determined, based on accurate estimation of tumour purity and ploidy. Although such estimates derived from sequencing data[188] are generally concordant with those derived from histopathology, concomitant inference of purity, ploidy and genome-wide copy number profiles for genomes with complex karyotypes may not lead to a unique mathematical solution. Thus, estimates of tumour purity and ploidy made by an algorithm should be checked with experimental data or the output of multiple algorithms should be examined[2,189].

Detecting structural variants. Upon mapping paired-end sequencing reads from the tumour genome to the reference genome, SVs are identified by the presence of split reads and clusters of discordant read pairs (FIG. 5). Nearly all SV detection algorithms, such as DELLY[190], Lumpy[191,192] and Meerkat[193], rely on split read and discordant read pair information to detect SVs. However, breakpoint junctions often show complex patterns, such as microhomology tracts as a consequence of DNA repair and insertions generated through template-switching events, that can result in poor alignment. To mitigate this problem, methods such as CREST[194], SvABA[85], BRASS[131] and Manta[195], also include a local assembly step. Although computationally intensive, contigs assembled from raw reads improve read mapping and the characterization of insertion sequences at the breakpoints[196]. Read depth data can provide additional information to improve the detection of deletions and amplifications[191,192].

Detecting somatic SVs is substantially more difficult than detecting germline SVs because of the low VAF of some somatic SVs. For example, if the tumour is sequenced at 60×, a subclonal SV present in 33% of diploid cancer cells in a sample with 30% tumour purity might only have, on average, 3–4 discordant and/or split reads supporting the insertion. In addition, the number of supporting reads will fluctuate due to non-uniform read coverage across the genome and sampling variation. Thus, the dynamic determination of appropriate thresholds (for example, number of supporting split reads) depending on the local context, as well as various filters to increase detection sensitivity, are key to a successful algorithm. Not surprisingly, benchmarking of SV detection algorithms shows discrepancies in specificity and recall, even across callers based on the same strategy[191]. Although improvement is not guaranteed[197], combining the output of multiple callers to increase specificity is a common strategy, both for germline[198,199] and somatic[2] events. For instance, in the PCAWG project, only SVs detected by at least two out of four algorithms were considered[2]. Similarly, germline SVs are filtered out by comparing them against a matched normal sample or a catalogue of germline SVs, similar to the 'panel of normals' strategy for SNV and indel calls. Several artefacts can still lead to false positive SV calls, such as chimeric reads originating from whole-genome amplification or library preparation.

Analysis of sequence homology at SV breakpoints can reveal the DNA repair mechanisms operative in human cancers[193]. For instance, microhomology and templated insertions at the breakpoints suggest the involvement of replication-associated mechanisms, such as microhomology-mediated break-induced replication[200]. Examining the number of SVs categorized by type (deletion, insertion, inversion or translocation), size and whether they are clustered or not can also give mechanistic insights[131]. Similar to mutational signature analysis using SNVs, the observed spectrum can be decomposed into signatures that might be linked to specific biological processes. Other features, such as background SV rates and copy number information[18,201], can also be incorporated into SV signature analysis. Some of the SV signatures have been shown to be relevant clinically, for example, the presence of indels showing microhomology indicates defects in homologous recombination[202].

Characterizing complex rearrangements. WGS studies enabled the discovery of additional types of complex genomic rearrangements[2,8]. These include


a Deletion
Reference genome
Deleted in tumour
Breakpoint
Split reads
Tumour genome
Discordant read pairs
Longer insert size
Reference genome
Deleted in tumour
Mapping to reference
b Tandem duplication
Duplicated in tumour
Breakpoint
Split reads
Discordant read pairs
Duplicated in tumour
c Inversion
Breakpoint 1
Breakpoint 2
Reference genome
Inverted in tumour
Split reads
Tumour genome
Discordant read pairs
Reference genome
Inverted in tumour
Mapping to reference
d Translocation
Breakpoint 1
Breakpoint 2
Reference genome
Chromosome A
Chromosome B
Split reads
Tumour genome
Discordant read pairs
Reference genome
Mapping to reference
e Viral insertion
Reference genome
Viral genome
Split reads
Tumour genome
Viral insertion
Discordant read pairs
Reference genome
Viral genome
Mapping to reference
f Transposable element insertion
Reference genome
TSD
Transposable element sequences
Split reads
Tumour genome
TSD
Transposable element
(A)n
TSD
Discordant read pairs
Reference genome
TSD
Transposable element sequences
Mapping to reference

◀ Fig. 5 | **Examples of detecting somatic structural variants from patterns of paired-end reads. a** | Detecting deletions. A somatic deletion bridges together two DNA segments (orange and green) that are not adjacent in the reference genome. When tumour reads are mapped to the reference genome, reads spanning the breakpoint between these two segments show a clipped (or split) alignment. These split reads contain two contiguous DNA segments that map to distant regions in the genome, corresponding to the two segments bridged by the deletion. Discordant read pairs spanning, but not mapping to, the breakpoint show longer than expected insert sizes. **b** | Detecting duplications. Tandem duplications also generate one novel adjacency, revealed by the presence of split reads. Discordant read pairs spanning the breakpoint map to the reference genome in an unexpected orientation and distance. **c** | Detecting inversions. Inversion of a genomic segment generates two new adjacencies. Reads mapping to these breakpoints show split patterns; discordant read pairs also occur as shown. **d** | Detecting translocations. A translocation bridges two genomic regions located in distant regions to generate a novel adjacency, which is revealed by the presence of split reads and clusters of discordant read pairs. In the inter-chromosomal translocation example shown, one read of the discordant read pair maps to one chromosome and the mate read maps to another chromosome. **e** | Detecting viral insertions. The insertion of viral sequences in the genome generates two adjacencies, as shown. A split read has one side mapping to the viral genome and the other side mapping to the human genome; a discordant read pair has one read mapping to the viral genome and the other read mapping to the human genome. Identification of the viral species inserted in the genome requires a database of viral genomes. **f** | Detecting transposable element insertions. Transposable element insertions are flanked by target site duplications (TSDs; shown in blue) at both ends of the insertion and by poly-A tails (shown in red) on one side. Split reads and discordant read pairs have similar patterns to those seen during viral insertion, but with transposable element consensus sequences rather than viral sequences. For all panels, depending on the sequencing depth and clonality of the alteration, only a subset of the read patterns may be observed in the data. There may be increased ploidy and multiple alterations in the same region, further complicating the detection and classification of these structural variants (SVs).

chromothripsis[19,203], which is characterized by massive de novo rearrangements of one or multiple chromosomes; chromoplexy[204,205], which involves balanced translocations across multiple chromosomes; and chromoanasynthesis[206], which is characterized by low-level copy number gains originating from template-switching events. WGS data also enabled the detailed characterization of known mechanisms underlying SV formation, such as breakage–fusion–bridge cycles (BFB cycles; characterized by copy number gains associated with multiple fold-back inversions)[207] and the generation of extrachromosomal DNA elements[208], which constitute self-replicating, circular DNA structures amplified to high copy numbers and often contain oncogenes[209].

The diversity, complexity and overlapping features of complex genomic rearrangements, coupled to their co-localization in highly rearranged tumours, make their detection and classification challenging[19,210], often requiring extensive manual curation. CTLPScanner[211] and ShatterProof[212] detect chromothripsis in array data based on the clustering of breakpoints and oscillating copy number patterns; ShatterSeek[19] and chromAL[210], developed for analysing WGS, utilize intra-chromosomal and inter-chromosomal SVs and copy number data to increase sensitivity and specificity and can detect events involving multiple chromosomes. Detection of chromoanasynthesis and BFB cycles relies on ad hoc solutions, whereas ChainFinder[204] and AmpliconArchitect[208] are specifically designed to detect chromoplexy and extrachromosomal circular DNA elements respectively. Reconstruction of the genomic sequence generated by these complex rearrangements is generally intractable using short reads only, as breakpoints falling in repetitive or low-complexity regions cannot always be reliably detected[213]. To characterize the full spectrum of somatic variation in a haplotype-resolved manner, data from multiple platforms must be integrated, for example, combining short reads with optical mapping, long-read sequencing and conformation capture information[30,214,215].

Detecting gene fusions using RNA-seq data. Although limited in sensitivity for transcripts expressed at low levels, RNA-seq data can be used to identify fusion oncogenes[216,217]. A standard approach consists of mapping sequencing reads to the transcriptome to identify discordantly mapped reads or reads mapping to the fusion junction. Tools such as Arriba[218], TopHat-Fusion[219], STAR-Fusion[220] and deFuse[221] follow this strategy. Other methods, such as TrinityFusion[220], CICERO[222] and JAFFA-Assembly[223], perform de novo assembly of sequencing reads to identify chimeric fusions. Methods relying on aligned reads show better performance than assembly-based callers, likely due to the difficulty in assembling de novo fusion transcripts when few reads spanning the junction are present[220]. However, assembly-based methods are more sensitive than those relying on aligned reads for reconstructing complex breakpoint junctions, such as those involving non-templated insertions[222] or viral and bacterial sequences[220]. Overall, the sensitivity, recall and computational efficiency of these tools are highly variable, with fusion calls often showing little overlap between tools[216,224]. Thus, when detecting fusion oncogenes, as for other variant types, considering calls made by multiple algorithms improves specificity[23,220,224], and filtering calls detected in normal samples helps to further reduce the false positive rate[23,225]. Profiling both DNA and RNA, or enriching for transcripts of interest showing low expression levels, also increases the sensitivity of detection for clinically relevant fusions[225,226].

Clonal composition and evolution

Cancer progression is an evolutionary process characterized by clonal competition that is fuelled by the accumulation of somatic mutations. As a result, tumours are complex mixtures of cells with different morphological and molecular profiles. Intra-tumour heterogeneity underpins drug resistance and relapse[17,227], and is associated with poor prognosis. Thus, dissecting clonal structure based on mutational data is important for understanding the molecular underpinnings of cancer evolution.

Clonal mutations are accrued in cancer development and are thus present in most cancer cells, whereas subclonal mutations are detected in a subset of cancer cells. Most algorithms discussed above can be used to detect subclonal mutations, but specialized methods have been developed to infer the subclonal architecture, especially when multiple samples (either across time or space) from the same individual are available (reviewed in REFS[228–230]). Popular tools such as SciClone[231], PyClone[232] and EXPANDS[11] were originally designed for high-coverage exomes, whereas other tools are focused on WGS-based inference[165,167,233,234]. Central to these methods is the notion that mutations from the

same subclone should have the same cancer cell fraction. Assignment of somatic mutations to subclones is achieved by clustering their VAFs adjusted for normal contamination and copy number.

Although WGS with standard sequencing depth is limited in detecting subclonal mutations, its genome-wide coverage of SNVs and more accurate characterization of CNAs allow for inference of the temporal order of CNAs relative to SNVs[22,168]. In the case of an early copy number gain, for example, most somatic mutations will be present in just one copy as they would have been acquired after the copy number gain; for a copy number gain late in tumour evolution, the majority of somatic mutations, assuming a comparable mutation rate, will be present in two out of three copies, because the mutations would have been amplified in the same event. Similarly, point mutations occurring before whole-genome doubling would be present in two out of four copies, and those accrued after whole-genome doubling would be present in only one copy. By applying this rationale to the set of somatic mutations detected in a cancer genome, the temporal order of somatic mutations during cancer evolution can be established[235]. Recent tumour evolution analysis of the PCAWG cohort uncovered common patterns of tumour evolution across 39 cancer types, revealing that some driver alterations, such as the formation of isochromosome 17q in medulloblastomas, date back to early development[22]. In other studies, the timing of CNAs and point mutations revealed a latency of years to decades between the acquisition of the early driver alterations and the major clonal expansion in lung adenocarcinomas[236] and renal cell carcinomas[237].

Although nearly all genomic analyses of tumour samples are based on a single biopsy, multiregional sampling is needed to reveal spatially complex patterns of clonal distribution. Such multiregional studies have proven powerful in delineating the heterogeneity of tumour clones and their evolutionary trajectories, underscoring the commonality of evolutionary branching and the role of subclonal drivers and chromosomal instability[238] (recently reviewed in REFS[17,239]). Analysis of longitudinally collected samples is ideal for understanding the effect of clonal dynamics over time, but most such studies focused on haematologic malignancies and selected solid tumour types that are easily sampled[240,241]. These studies showed the impact of treatments on the mutational landscape of tumours and their clonal compositions[242]. Although cancer genome analyses including TCGA and the PCAWG have been largely restricted to primary tumours, more recent efforts have collected large numbers of clinically annotated WGS and whole-exome sequencing data sets for metastatic samples[8,9], permitting the characterization of the effect of chemotherapeutic agents[138] and radiotherapy[243] on the genomes of cancer cells.

Visualizing and exploring cancer genomes

A critical component at all stages of cancer genome analysis is data visualization and exploratory analysis. At the variant calling step, visual inspection of the read-level data for candidate mutations is customary and performed using tools such as the Integrative Genomics Viewer (IGV)[244]. To investigate the functional and therapeutic relevance of somatic mutations, cBioPortal[245] provides easy access to a comprehensive set of genomic and clinical data from large-scale cancer genomics projects. Other portals include Genomic Data Commons by the US National Cancer Institute[246], GenomePaint (paediatric cancers)[247] and the International Cancer Genome Consortium data portal[248]. These portals offer many tools for interactively exploring genomic alterations across genes, samples and pathways, as well as for correlating them with clinical attributes. Currently, most tools are focused on SNVs and CNAs; to take full advantage of WGS data, tools for visualizing various types of SVs and for the integrative analysis of multiple data types will be helpful.

Conclusions and perspectives

In the last decade, we have witnessed tremendous growth in sequencing capability, accompanied by growing sophistication in computational tools. As DNA sequencing is now a commodity, the amount of data generated by the cancer research community will continue to increase. To harness the potential of these data, effective consolidation of data and their integrative analysis using efficient cloud-based tools will be needed. The recent PCAWG Consortium efforts were exemplary, yet there are tens or hundreds of thousands of genomes that have been profiled but are not being fully utilized. Data access and utilization are often limited due to, for example, restrictive language on patient consent forms, government regulations, especially for sharing across countries, and incomplete annotation of the data by researchers that submit them. Improved data sharing practices and the development of scalable cloud-based infrastructure for storage, analysis and sharing of genomic data[249,250] will be essential for researchers to utilize all of the available sequencing data.

For analytical tools, we envision exciting new insights from integration of genomics data with other data modalities, such as imaging and histopathology data[251–254], as well as an expanding role for methods based on deep learning. The development of new methods for mining and extracting clinical information from electronic health records will also add new dimensions to the interpretation of genomics data in the context of disease progression. New technologies that will be prominent in the next few years include non-invasive techniques based on circulating tumour DNA to characterize tumour heterogeneity and monitor disease[255]; long-read technologies — assuming a substantial decrease in their cost — that can overcome many of the limitations discussed above, especially in characterization of the repetitive regions and SVs; and single-cell DNA sequencing technologies, aided by improved DNA amplification techniques[256,257] and with concurrent profiling of the RNA or epigenome of the same cells, to interrogate cellular heterogeneity and clonal evolution at high resolution. All of these technologies will require the development of innovative bioinformatic tools for analysis, visualization and interpretation.

Published online 8 December 2021

1. Bailey, M. H. et al. Comprehensive characterization of cancer driver genes and mutations. *Cell* **173**, 371–385.e18 (2018).
This study reports the analysis of nearly 10,000 exomes from TCGA, identifying ~300 cancer driver genes and finding that more than half of the samples have potentially actionable events.
2. Campbell, P. J. et al. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
This is the flagship paper for an international effort to analyse WGS data from 2,658 primary tumours, describing the consortium's variant calling steps as well as reporting the landscape of somatic mutation especially for structural variation.
3. Martínez-Jiménez, F. et al. A compendium of mutational cancer driver genes. *Nat. Rev. Cancer* **20**, 555–572 (2020).
4. Rheinbay, E. et al. Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature* **578**, 102–111 (2020).
5. Ma, X. et al. Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours. *Nature* **555**, 371–376 (2018).
6. Gröbner, S. N. et al. The landscape of genomic alterations across childhood cancers. *Nature* **555**, 321–327 (2018).
7. Zehir, A. et al. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nat. Med.* **23**, 703–713 (2017).
This study describes the analysis of panel sequencing data from a prospective clinical sequencing initiative to demonstrate the clinical utility of tumour molecular profiling.
8. Priestley, P. et al. Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature* **575**, 210–216 (2019).
This paper reports the mutational landscape of >2,500 metastatic tumours, finding genetic variants that may be used to stratify patients towards therapies for >60% of the cases.
9. Pleasance, E. et al. Pan-cancer analysis of advanced patient tumors reveals interactions between therapy and genomic landscapes. *Nat. Cancer* **1**, 452–468 (2020).
10. Koche, R. P. et al. Extrachromosomal circular DNA drives oncogenic genome remodeling in neuroblastoma. *Nat. Genet.* **52**, 29–34 (2020).
11. Andor, N. et al. Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nat. Med.* **22**, 105–113 (2016).
12. Weghorn, D. & Sunyaev, S. Bayesian inference of negative and positive selection in human cancers. *Nat. Genet.* **49**, 1785–1788 (2017).
13. Marty, R. et al. MHC-I genotype restricts the oncogenic mutational landscape. *Cell* **171**, 1272–1283.e15 (2017).
14. Martincorena, I. et al. Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029–1041.e21 (2017).
This paper examines the selection pressures on somatic single-nucleotide mutations, finding near-complete absence of negative selection.
15. Hu, Z. et al. Quantitative evidence for early metastatic seeding in colorectal cancer. *Nat. Genet.* **51**, 1113–1122 (2019).
16. Zhang, X. & Meyerson, M. Illuminating the noncoding genome in cancer. *Nat. Cancer* **1**, 864–872 (2020).
17. McGranahan, N. & Swanton, C. Clonal heterogeneity and tumor evolution: past, present, and the future. *Cell* **168**, 613–628 (2017).
18. Li, Y. et al. Patterns of somatic structural variation in human cancer genomes. *Nature* **578**, 112–121 (2020).
This study describes comprehensive identification and classification of SVs based on WGS data from >2,600 tumours, and reports 16 structural variation signatures and their characteristics.
19. Cortés-Ciriano, I. et al. Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. *Nat. Genet.* **52**, 331–341 (2020).
20. Sieverling, L. et al. Genomic footprints of activated telomere maintenance mechanisms in cancer. *Nat. Commun.* **11**, 733 (2020).
21. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
22. Gerstung, M. et al. The evolutionary history of 2,658 cancers. *Nature* **578**, 122–128 (2020).
This paper reports a large-scale analysis of the timing of point mutations and CNAs, and describes the common trajectories of tumour development across multiple tumour types.
23. Calabrese, C. et al. Genomic basis for RNA alterations in cancer. *Nature* **578**, 129–136 (2020).
24. Yuan, Y. et al. Comprehensive molecular characterization of mitochondrial genomes in human cancers. *Nat. Genet.* **52**, 342–352 (2020).
25. Rodriguez-Martin, B. et al. Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. *Nat. Genet.* **52**, 306–319 (2020).
26. Zapatka, M. et al. The landscape of viral associations in human cancers. *Nat. Genet.* **52**, 320–330 (2020).
27. Meyerson, M., Gabriel, S. & Getz, G. Advances in understanding cancer genomes through second-generation sequencing. *Nat. Rev. Genet.* **11**, 685–696 (2010).
28. Ding, L., Wendl, M. C., McMichael, J. F. & Raphael, B. J. Expanding the computational toolbox for mining cancer genomes. *Nat. Rev. Genet.* **15**, 556–570 (2014).
29. Ho, S. S., Urban, A. E. & Mills, R. E. Structural variation in the sequencing era. *Nat. Rev. Genet.* **21**, 171–189 (2020).
30. Dixon, J. R. et al. Integrative detection and analysis of structural variation in cancer genomes. *Nat. Genet.* **50**, 1388–1398 (2018).
31. Liu, X. S. & Mardis, E. R. Applications of immunogenomics to cancer. *Cell* **168**, 600–612 (2017).
32. Gawad, C., Koh, W. & Quake, S. R. Single-cell genome sequencing: current state of the science. *Nat. Rev. Genet.* **17**, 175–188 (2016).
33. Castro, L. N. G., Tirosh, I. & Suvà, M. L. Decoding cancer biology one cell at a time. *Cancer Discov.* **11**, 960–970 (2021).
34. Lim, B., Lin, Y. & Navin, N. Advancing cancer research and medicine with single-cell genomics. *Cancer Cell* **37**, 456–470 (2020).
35. Chakravarty, D. & Solit, D. B. Clinical cancer genomic profiling. *Nat. Rev. Genet.* **22**, 483–501 (2021).
36. Logsdon, G. A., Vollger, M. R. & Eichler, E. E. Long-read human genome sequencing and its applications. *Nat. Rev. Genet.* **21**, 597–614 (2020).
37. Amarasinghe, S. L. et al. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* **21**, 1–16 (2020).
38. Cescon, D. W., Bratman, S. V., Chan, S. M. & Siu, L. L. Circulating tumor DNA and liquid biopsy in oncology. *Nat. Cancer* **1**, 276–290 (2020).
39. Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L. & Rice, P. M. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.* **38**, 1767–1771 (2009).
40. Fritz, M. H. Y., Leinonen, R., Cochrane, G. & Birney, E. Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Res.* **21**, 734–740 (2011).
41. Costello, M. et al. Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. *BMC Genomics* **19**, 332 (2018).
42. Andrews, S. FastQC: a quality control tool for high throughput sequence data. *Babraham Bioinformatics* http://www.bioinformatics.babraham.ac.uk/projects/fastqc (2010).
43. Rausch, T., Hsi-Yang Fritz, M., Korbel, J. O. & Benes, V. Alfred: interactive multi-sample BAM alignment statistics, feature counting and feature annotation for long- and short-read sequencing. *Bioinformatics* **35**, 2489–2491 (2019).
44. Okonechnikov, K., Conesa, A. & García-Alcalde, F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* **32**, 292–294 (2016).
45. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
46. Garrison, E. et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.* **36**, 875–879 (2018).
47. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
48. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
49. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
50. Schneider, V. A. et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* **27**, 849–864 (2017).
51. Auton, A. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
52. Gao, G. F. et al. Before and after: comparison of legacy and harmonized TCGA Genomic Data Commons' data. *Cell Syst.* **9**, 24–34.e10 (2019).
53. Logsdon, G. A. et al. The structure, function and evolution of a complete human chromosome 8. *Nature* **593**, 101–107 (2021).
54. Martincorena, I. & Campbell, P. J. Somatic mutation in cancer and normal cells. *Science* **349**, 1483–1489 (2015).
55. Cortes-Ciriano, I., Lee, S., Park, W.-Y. Y., Kim, T.-M. M. & Park, P. J. A molecular portrait of microsatellite instability across multiple cancers. *Nat. Commun.* **8**, 15180 (2017).
56. Griffith, M. et al. Optimizing cancer genome sequencing and analysis. *Cell Syst.* **1**, 210–223 (2015).
This study examines the impact of different experimental and computational strategies in characterization of a complex tumour and provides a resource of validation data for 200,000 SNVs.
57. Xu, C. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Comput. Struct. Biotechnol. J.* **16**, 15–24 (2018).
58. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
59. Jones, D. et al. cgpCaVEManWrapper: simple execution of CaVEMan in order to detect somatic single nucleotide variants in NGS data. *Curr. Protoc. Bioinformatics* **56**, 15.10.1–15.10.18 (2016).
60. Kim, S. et al. Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods* **15**, 591–594 (2018).
61. Lai, Z. et al. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res.* **44**, e108 (2016).
62. Fan, Y. et al. MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biol.* **17**, 178 (2016).
63. Sherry, S. T. et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
64. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
65. Jones, S. et al. Personalized genomic analyses for cancer mutation discovery and interpretation. *Sci. Transl Med.* **7**, 283ra53 (2015).
66. Forbes, S. A. et al. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.* **45**, D777–D783 (2017).
67. O'Rawe, J. et al. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Med.* **5**, 28 (2013).
68. Krøigård, A. B., Thomassen, M., Lænkholm, A. V., Kruse, T. A. & Larsen, M. J. Evaluation of nine somatic variant callers for detection of somatic mutations in exome and targeted deep sequencing data. *PLoS ONE* **11**, e0151664 (2016).
69. Wang, Q. et al. Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers. *Genome Med.* **5**, 91 (2013).
70. Ewing, A. D. et al. Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nat. Methods* **12**, 623–630 (2015).
71. Alioto, T. S. et al. A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nat. Commun.* **6**, 10001 (2015).
72. Ellrott, K. et al. Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. *Cell Syst.* **6**, 271–281.e7 (2018).
73. Callari, M. et al. Intersect-then-combine approach: improving the performance of somatic variant calling in whole exome sequencing data using multiple aligners and callers. *Genome Med.* **9**, 35 (2017).
74. Huang, W. et al. SMuRF: portable and accurate ensemble prediction of somatic mutations. *Bioinformatics* **35**, 3157–3159 (2019).
75. Wood, D. E. et al. A machine learning approach for somatic mutation discovery. *Sci. Transl Med.* **10**, eaar7939 (2018).
76. Ding, J. et al. Feature-based classifiers for somatic mutation detection in tumour-normal paired sequencing data. *Bioinformatics* **28**, 167–175 (2012).

77. Cantarel, B. L. et al. BAYSIC: a Bayesian method for combining sets of genome variants with improved specificity and sensitivity. *BMC Bioinformatics* **15**, 104 (2014).
78. Fang, L. T. et al. An ensemble approach to accurately detect somatic mutations using SomaticSeq. *Genome Biol.* **16**, 197 (2015).
79. Poplin, R. et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983 (2018).
80. Sahraeian, S. M. E. et al. Deep convolutional neural networks for accurate somatic mutation detection. *Nat. Commun.* **10**, 1041 (2019).
81. Torracinta, R. et al. Adaptive somatic mutations calls with deep learning and semi-simulated data. Preprint at *bioRxiv* https://doi.org/10.1101/079087 (2016).
82. Dou, Y. et al. Accurate detection of mosaic variants in sequencing data without matched controls. *Nat. Biotechnol.* **38**, 314–319 (2020).
83. Li, H. & Wren, J. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* **30**, 2843–2851 (2014).
84. Poplin, R. et al. Scaling accurate genetic variant discovery to tens of thousands of samples. Preprint at *bioRxiv* https://doi.org/10.1101/201178 (2017).
85. Wala, J. A. et al. SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome Res.* **28**, 581–591 (2018).
86. Rimmer, A. et al. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* **46**, 912–918 (2014).
87. Greenman, C. et al. Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153–158 (2007).
88. Lawrence, M. S. et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
This study introduces a computational framework for the discovery of driver genes that accounts for the variable mutation rates across the genome.
89. Schuster-Böckler, B. & Lehner, B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* **488**, 504–507 (2012).
90. Polak, P. et al. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* **518**, 360–364 (2015).
91. Supek, F. & Lehner, B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature* **521**, 81–84 (2015).
92. Katainen, R. et al. CTCF/cohesin-binding sites are frequently mutated in cancer. *Nat. Genet.* **47**, 818–821 (2015).
93. Sabarinathan, R., Mularoni, L., Deu-Pons, J., Gonzalez-Perez, A. & Lopez-Bigas, N. Nucleotide excision repair is impaired by binding of transcription factors to DNA. *Nature* **532**, 264–267 (2016).
94. Gonzalez-Perez, A., Sabarinathan, R. & Lopez-Bigas, N. Local determinants of the mutational landscape of the human genome. *Cell* **177**, 101–114 (2019).
95. Dietlein, F. et al. Identification of cancer driver genes based on nucleotide context. *Nat. Genet.* **52**, 208–218 (2020).
96. Nissim, S. et al. Mutations in RABL3 alter KRAS prenylation and are associated with hereditary pancreatic cancer. *Nat. Genet.* **51**, 1308–1314 (2019).
97. Hess, J. M. et al. Passenger hotspot mutations in cancer. *Cancer Cell* **36**, 288–301.e14 (2019).
98. Tamborero, D., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics* **29**, 2238–2244 (2013).
99. Niu, B. et al. Protein-structure-guided discovery of functional mutations across 19 cancer types. *Nat. Genet.* **48**, 827–837 (2016).
100. Reimand, J. & Bader, G. D. Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Mol. Syst. Biol.* **9**, 637 (2013).
101. Zhu, H. et al. Candidate cancer driver mutations in distal regulatory elements and long-range chromatin interaction networks. *Mol. Cell* **77**, 1307–1321.e10 (2020).
102. Khurana, E. et al. Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* **342**, 1235587 (2013).
103. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* **39**, e118–e118 (2011).
104. Buisson, R. et al. Passenger hotspot mutations in cancer driven by APOBEC3A and mesoscale genomic features. *Science* **364**, eaaw2872 (2019).
105. McLaren, W. et al. The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
106. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* **6**, 80–92 (2012).
107. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
108. McCarthy, D. J. et al. Choice of transcripts and software has a large effect on variant annotation. *Genome Med.* **6**, 26 (2014).
109. Yen, J. L. et al. A variant by any name: quantifying annotation discordance across tools and clinical databases. *Genome Med.* **9**, 7 (2017).
110. Landrum, M. J. et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42**, D980–D985 (2014).
111. Chakravarty, D. et al. OncoKB: a precision oncology knowledge base. *JCO Precis. Oncol.* **1**, 1–16 (2017).
112. Khurana, E. et al. Role of non-coding sequence variants in cancer. *Nat. Rev. Genet.* **17**, 93–108 (2016).
113. Liu, Y. et al. Discovery of regulatory noncoding variants in individual cancer genomes by using *cis*-X. *Nat. Genet.* **52**, 811–818 (2020).
114. Kanagawa, T. Bias and artifacts in multitemplate polymerase chain reactions (PCR). *J. Biosci. Bioeng.* **96**, 317–323 (2003).
115. Buckley, A. R. et al. Pan-cancer analysis reveals technical artifacts in TCGA germline variant calls. *BMC Genomics* **18**, 458 (2017).
116. Costello, M. et al. 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 (2013).
117. Do, H. & Dobrovic, A. Sequence artifacts in DNA from formalin-fixed tissues: causes and strategies for minimization. *Clin. Chem.* **61**, 64–71 (2015).
118. Frampton, G. M. et al. Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nat. Biotechnol.* **31**, 1023–1031 (2013).
119. Kerick, M. et al. Targeted high throughput sequencing in clinical cancer settings: formaldehyde fixed-paraffin embedded (FFPE) tumor tissues, input amount and tumor heterogeneity. *BMC Med. Genomics* **4**, 68 (2011).
120. Van Allen, E. M. et al. Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. *Nat. Med.* **20**, 682–688 (2014).
121. Robbe, P. et al. Clinical whole-genome sequencing from routine formalin-fixed, paraffin-embedded specimens: pilot study for the 100,000 Genomes Project. *Genet. Med.* **20**, 1196–1205 (2018).
122. Cibulskis, K. et al. ContEst: estimating cross-contamination of human samples in next-generation sequencing data. *Bioinformatics* **27**, 2601–2602 (2011).
123. Fiévet, A. et al. ART-DeCo: easy tool for detection and characterization of cross-contamination of DNA samples in diagnostic next-generation sequencing analysis. *Eur. J. Hum. Genet.* **27**, 792–800 (2019).
124. Bergmann, E. A., Chen, B. J., Arora, K., Vacic, V. & Zody, M. C. Conpair: concordance and contamination estimator for matched tumor–normal pairs. *Bioinformatics* **32**, 3196–3198 (2016).
125. Chun, H. & Kim, S. BAMixChecker: an automated checkup tool for matched sample pairs in NGS cohort. *Bioinformatics* **35**, 4806–4808 (2019).
126. Lee, S. S. et al. NGSCheckMate: software for validating sample identity in next-generation sequencing studies within and across data types. *Nucleic Acids Res.* **45**, e103 (2017).
127. Schröder, J., Corbin, V. & Papenfuss, A. T. HYSYS: have you swapped your samples? *Bioinformatics* **33**, 596–598 (2017).
128. Taylor-Weiner, A. et al. DeTiN: overcoming tumor-in-normal contamination. *Nat. Methods* **15**, 531–534 (2018).
129. Nik-Zainal, S. et al. Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
130. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
This is the first comprehensive study on mutational signatures, describing >20 mutational processes operative in >7,000 tumours using mutational signature analysis.
131. Nik-Zainal, S. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).
This study identifies mutational signatures in breast cancers, including the rearrangement signatures associated with BRCA1/2 mutations that can serve as a biomarker of homologous recombination deficiency.
132. Macintyre, G. et al. Copy number signatures and mutational processes in ovarian carcinoma. *Nat. Genet.* **50**, 1262–1270 (2018).
133. Steele, C. D. et al. Signatures of copy number alterations in human cancer. Preprint at *bioRxiv* https://doi.org/10.1101/2021.04.30.441940 (2021).
134. Kasar, S. et al. Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nat. Commun.* **6**, 8866 (2015).
135. Fischer, A., Illingworth, C. J. R., Campbell, P. J. & Mustonen, V. EMu: probabilistic inference of mutational processes and their localization in the cancer genome. *Genome Biol.* **14**, R39 (2013).
136. Funnell, T. et al. Integrated structural variation and point mutation signatures in cancer genomes using correlated topic models. *PLoS Comput. Biol.* **15**, e1006799 (2019).
137. Kucab, J. E. et al. A compendium of mutational signatures of environmental agents. *Cell* **177**, 821–836.e16 (2019).
138. Pich, O. et al. The mutational footprints of cancer therapies. *Nat. Genet.* **51**, 1732–1740 (2019).
139. Lee-Six, H. et al. The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532–537 (2019).
140. Lodato, M. A. et al. Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science* **359**, 555–559 (2018).
141. Alexandrov, L. B. et al. Clock-like mutational processes in human somatic cells. *Nat. Genet.* **47**, 1402–1407 (2015).
142. Martincorena, I. et al. Somatic mutant clones colonize the human esophagus with age. *Science* **362**, 911–917 (2018).
143. Li, S., Crawford, F. W. & Gerstein, M. B. Using sigLASSO to optimize cancer mutation signatures jointly with sampling likelihood. *Nat. Commun.* **11**, 3575 (2020).
144. Peharz, R. & Pernkopf, F. Sparse nonnegative matrix factorization with ℓ 0-constraints. *Neurocomputing* **80**, 38–46 (2012).
145. Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S. & Swanton, C. deconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol.* **17**, 31 (2016).
146. Blokzijl, F., Janssen, R., van Boxtel, R. & Cuppen, E. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. *Genome Med.* **10**, 33 (2018).
147. Omichessan, H., Severi, G. & Perduca, V. Computational tools to detect signatures of mutational processes in DNA from tumours: a review and empirical comparison of performance. *PLoS ONE* **14**, e0221235 (2019).
148. Riva, L. et al. The mutational signature profile of known and suspected human carcinogens in mice. *Nat. Genet.* **52**, 1189–1197 (2020).
149. Baez-Ortega, A. et al. Somatic evolution and global expansion of an ancient transmissible cancer lineage. *Science* **365**, eaau9923 (2019).
150. Cartolano, M. et al. CaMuS: simultaneous fitting and de novo imputation of cancer mutational signature. *Sci. Rep.* **10**, 1–10 (2020).
151. Gulhan, D. C., Lee, J. J.-K., Melloni, G. E. M., Cortés-Ciriano, I. & Park, P. J. Detecting the mutational signature of homologous recombination deficiency in clinical samples. *Nat. Genet.* **51**, 912–919 (2019).
152. Färkkilä, A. et al. Immunogenomic profiling determines responses to combined PARP and PD-1 inhibition in ovarian cancer. *Nat. Commun.* **11**, 2543 (2020).
153. Weischenfeldt, J. et al. Pan-cancer analysis of somatic copy-number alterations implicates IRS4 and IGF2 in enhancer hijacking. *Nat. Genet.* **49**, 65–74 (2017).

154. Northcott, P. A. et al. Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. *Nature* **511**, 428–434 (2014).
155. Herranz, D. et al. A NOTCH1-driven MYC enhancer promotes T cell development, transformation and acute lymphoblastic leukemia. *Nat. Med.* **20**, 1130–1137 (2014).
156. Quigley, D. A. et al. Genomic hallmarks and structural variation in metastatic prostate cancer. *Cell* **174**, 758–769.e9 (2018).
157. Takeda, D. Y. et al. A somatically acquired enhancer of the androgen receptor is a noncoding driver in advanced prostate cancer. *Cell* **174**, 422–432.e13 (2018).
158. Kallioniemi, A. et al. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* **258**, 818–821 (1992).
159. Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* **12**, 363–376 (2011).
160. Van Loo, P. et al. Allele-specific copy number analysis of tumors. *Proc. Natl Acad. Sci. USA* **107**, 16910–16915 (2010).
161. Carter, S. L. et al. Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30**, 413–421 (2012).
162. Shen, R. & Seshan, V. E. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Res.* **44**, e131 (2016).
163. Raine, K. M. et al. ascatNgs: identifying somatically acquired copy-number alterations from whole-genome sequencing data. *Curr. Protoc. Bioinformatics* **56**, 15.9.1–15.9.17 (2016).
164. Xi, R., Lee, S., Xia, Y., Kim, T. M. & Park, P. J. Copy number analysis of whole-genome data using BIC-seq2 and its application to detection of cancer susceptibility variants. *Nucleic Acids Res.* **44**, 6274–6286 (2016).
165. Dentro, S. C. et al. Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. *Cell* **184**, 2239–2254.e39 (2021).
166. Chen, X. et al. CONSERTING: integrating copy-number analysis with structural-variation detection. *Nat. Methods* **12**, 527–530 (2015).
167. Fischer, A., Vázquez-Garcïa, I., Illingworth, C. J. R. R. & Mustonen, V. High-definition reconstruction of clonal composition in cancer. *Cell Rep.* **7**, 1740–1752 (2014).
168. Nik-Zainal, S. et al. The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
169. Cun, Y., Yang, T.-P., Achter, V., Lang, U. & Peifer, M. Copy-number analysis and inference of subclonal populations in cancer genomes using Sclust. *Nat. Protoc.* **13**, 1488–1501 (2018).
170. Kleinheinz, K. et al. ACEseq — allele specific copy number estimation from whole genome sequencing. Preprint at *bioRxiv* https://doi.org/10.1101/210807 (2017).
171. Li, Y. et al. Allele-specific quantification of structural variations in cancer genomes. *Cell Syst.* **3**, 21–34 (2016).
172. Hadi, K. et al. Distinct classes of complex structural variation uncovered across thousands of cancer genome graphs. *Cell* **183**, 197–210.e32 (2020).
173. Aganezov, S. & Raphael, B. J. Reconstruction of clone- and haplotype-specific cancer genome karyotypes from bulk tumor samples. *Genome Res.* **30**, 1274–1290 (2020).
174. Amarasinghe, K. C. et al. Inferring copy number and genotype in tumour exome data. *BMC Genomics* **15**, 732 (2014).
175. Boeva, V. et al. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics* **28**, 423–425 (2012).
176. Magi, A. et al. EXCAVATOR: detecting copy number variants from whole-exome sequencing data. *Genome Biol.* **14**, R120 (2013).
177. Sathirapongsasuti, J. F. et al. Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics* **27**, 2648–2654 (2011).
178. Xu, H., DiCarlo, J., Satya, R. V., Peng, Q. & Wang, Y. Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. *BMC Genomics* **15**, 244 (2014).
179. Li, J. et al. CONTRA: copy number analysis for targeted resequencing. *Bioinformatics* **28**, 1307–1313 (2012).
180. Bao, L., Pu, M. & Messer, K. AbsCN-seq: a statistical method to estimate tumor purity, ploidy and absolute copy numbers from next-generation sequencing data. *Bioinformatics* **30**, 1056–1063 (2014).
181. Nam, J. Y. et al. Evaluation of somatic copy number estimation tools for whole-exome sequencing data. *Brief. Bioinform.* **17**, 185–192 (2016).
182. Zare, F., Dow, M., Monteleone, N., Hosny, A. & Nabavi, S. An evaluation of copy number variation detection tools for cancer using whole exome sequencing data. *BMC Bioinformatics* **18**, 286 (2017).
183. Kuilman, T. et al. CopywriteR: DNA copy number detection from off-target sequence data. *Genome Biol.* **16**, 1–15 (2015).
184. Talevich, E., Shain, A. H., Botton, T. & Bastian, B. C. CNVkit: genome-wide copy number detection and visualization from targeted DNA sequencing. *PLoS Comput. Biol.* **12**, e1004873 (2016).
185. Favero, F. et al. Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann. Oncol.* **26**, 64–70 (2015).
186. Yang, L. et al. Analyzing somatic genome rearrangements in human cancers by using whole-exome sequencing. *Am. J. Hum. Genet.* **98**, 843–856 (2016).
187. Mermel, C. H. et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, 1–14 (2011).
188. Haider, S. et al. Systematic assessment of tumor purity and its clinical implications. *JCO Precis. Oncol.* **4**, 995–1005 (2020).
189. Aran, D., Sirota, M. & Butte, A. J. Systematic pan-cancer analysis of tumour purity. *Nat. Commun.* **6**, 8971 (2015).
190. Rausch, T. et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).
191. Kosugi, S. et al. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol.* **20**, 117 (2019).
192. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.* **15**, R84 (2014).
193. Yang, L. et al. Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell* **153**, 919–929 (2013).
194. Wang, J. et al. CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nat. Methods* **8**, 652–654 (2011).
195. Chen, X. et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–1222 (2016).
196. Cameron, D. L. et al. GRIDSS2: comprehensive characterisation of somatic structural variation using single breakend variants and structural variant phasing. *Genome Biol.* **22**, 1–25 (2021).
197. Lee, A. Y. et al. Combining accurate tumor genome simulation with crowdsourcing to benchmark somatic structural variant detection. *Genome Biol.* **19**, 188 (2018).
198. Sudmant, P. H. et al. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
199. Mills, R. E. et al. Mapping copy number variation by population-scale genome sequencing. *Nature* **470**, 59–65 (2011).
200. Carvalho, C. M. B. & Lupski, J. R. Mechanisms underlying structural variant formation in genomic disorders. *Nat. Rev. Genet.* **17**, 224–238 (2016).
201. Glodzik, D. et al. A somatic-mutational process recurrently duplicates germline susceptibility loci and tissue-specific super-enhancers in breast cancers. *Nat. Genet.* **49**, 341–348 (2017).
202. Davies, H. et al. HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nat. Med.* **23**, 517–525 (2017).
203. Stephens, P. J. et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27–40 (2011).
This paper reports the discovery of a mutational process, termed chromothripsis, characterized by tens to hundreds of structural rearrangements acquired in a single cell division.
204. Baca, S. C. et al. Punctuated evolution of prostate cancer genomes. *Cell* **153**, 666–677 (2013).
By examining the patterns of structural variation, this study finds 'chromoplexy', a large chain of rearrangements that affect multiple chromosomes and may drive prostate carcinogenesis.
205. Anderson, N. D. et al. Rearrangement bursts generate canonical gene fusions in bone and soft tissue tumors. *Science* **361**, eaam8419 (2018).
206. Liu, P. et al. Chromosome catastrophes involve replication mechanisms generating complex genomic rearrangements. *Cell* **146**, 889–903 (2011).
207. Campbell, P. J. et al. The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature* **467**, 1109–1113 (2010).
208. Deshpande, V. et al. Exploring the landscape of focal amplifications in cancer using AmpliconArchitect. *Nat. Commun.* **10**, 392 (2019).
209. Turner, K. M. et al. Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. *Nature* **543**, 122–125 (2017).
210. Notta, F. et al. A renewed model of pancreatic cancer evolution based on genomic rearrangement patterns. *Nature* **538**, 378–382 (2016).
211. Yang, J. et al. CTLPScanner: a web server for chromothripsis-like pattern detection. *Nucleic Acids Res.* **44**, W252–W258 (2016).
212. Govind, S. K. et al. ShatterProof: operational detection and quantification of chromothripsis. *BMC Bioinformatics* **15**, 78 (2014).
213. Wang, S. et al. HiNT: a computational method for detecting copy number variations and translocations from Hi-C data. *Genome Biol.* **21**, 73 (2020).
214. Harewood, L. et al. Hi-C as a tool for precise detection and characterisation of chromosomal rearrangements and copy number variation in human tumours. *Genome Biol.* **18**, 125 (2017).
215. Chaisson, M. J. P. P. et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* **10**, 1784 (2019).
216. Kumar, S., Vo, A. D., Qin, F. & Li, H. Comparative assessment of methods for the fusion transcripts detection from RNA-seq data. *Sci. Rep.* **6**, 21597 (2016).
217. Liu, S. et al. Comprehensive evaluation of fusion transcript detection algorithms and a meta-caller to combine top performing methods in paired-end RNA-seq data. *Nucleic Acids Res.* **44**, e47 (2015).
218. Uhrig, S. et al. Accurate and efficient detection of gene fusions from RNA sequencing data. *Genome Res.* **31**, 448–460 (2021).
219. Kim, D. & Salzberg, S. L. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol.* **12**, R72 (2011).
220. Haas, B. J. et al. Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. *Genome Biol.* **20**, 213 (2019).
221. McPherson, A. et al. deFuse: an algorithm for gene fusion discovery in tumor RNA-seq data. *PLoS Comput. Biol.* **7**, e1001138 (2011).
222. Tian, L. et al. CICERO: a versatile method for detecting complex and diverse driver fusions using cancer RNA sequencing data. *Genome Biol.* **21**, 1–18 (2020).
223. Davidson, N. M., Majewski, I. J. & Oshlack, A. JAFFA: high sensitivity transcriptome-focused fusion gene detection. *Genome Med.* **7**, 43 (2015).
224. Picco, G. et al. Functional linkage of gene fusions to cancer cell fitness assessed by pharmacological and CRISPR–Cas9 screening. *Nat. Commun.* **10**, 2198 (2019).
225. Gao, Q. et al. Driver fusions and their implications in the development and treatment of human cancers. *Cell Rep.* **23**, 227–238.e3 (2018).
226. Heyer, E. E. et al. Diagnosis of fusion genes using targeted RNA sequencing. *Nat. Commun.* **10**, 1388 (2019).
227. Burrell, R. A., McGranahan, N., Bartek, J. & Swanton, C. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* **501**, 338–345 (2013).
228. Tarabichi, M. et al. A practical guide to cancer subclonal reconstruction from DNA sequencing. *Nat. Methods* **18**, 144–155 (2021).
229. Dentro, S. C., Wedge, D. C. & Van Loo, P. Principles of reconstructing the subclonal architecture of cancers. *Cold Spring Harb. Perspect. Med.* **7**, a026625 (2017).
230. Salcedo, A. et al. A community effort to create standards for evaluating tumor subclonal reconstruction. *Nat. Biotechnol.* **38**, 97–107 (2020).
231. Miller, C. A. et al. SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS Comput. Biol.* **10**, e1003665 (2014).
232. Roth, A. et al. PyClone: statistical inference of clonal population structure in cancer. *Nat. Methods* **11**, 396–398 (2014).
233. Deshwar, A. G. et al. PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol.* **16**, 35 (2015).

234. Caravagna, G. et al. Subclonal reconstruction of tumors by using machine learning and population genetics. *Nat. Genet.* **52**, 898–907 (2020).
235. Yang, L. et al. An enhanced genetic model of colorectal cancer progression history. *Genome Biol.* **20**, 168 (2019).
236. Lee, J. J.-K. et al. Tracing oncogene rearrangements in the mutational history of lung adenocarcinoma. *Cell* **177**, 1842–1857.e21 (2019).
237. Mitchell, T. J. et al. Timing the landmark events in the evolution of clear cell renal cell cancer: TRACERx Renal. *Cell* **173**, 611–623.e17 (2018).
238. Watkins, T. B. K. et al. Pervasive chromosomal instability and karyotype order in tumour evolution. *Nature* **587**, 126–132 (2020).
239. Schwartz, R. & Schäffer, A. A. The evolution of tumour phylogenetics: principles and practice. *Nat. Rev. Genet.* **18**, 213–229 (2017).
240. Ding, L. et al. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* **481**, 506–510 (2012).
241. Landau, D. A. et al. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell* **152**, 714–726 (2013).
242. Liu, D. et al. Mutational patterns in chemotherapy resistant muscle-invasive bladder cancer. *Nat. Commun.* **8**, 2193 (2017).
243. Behjati, S. et al. Mutational signatures of ionizing radiation in second malignancies. *Nat. Commun.* **7**, 1–8 (2016).
244. Robinson, J. T. et al. Integrative Genomics Viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
245. Cerami, E. et al. The cBio Cancer Genomics Portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–404 (2012).
246. Grossman, R. L. et al. Toward a shared vision for cancer genomic data. *N. Engl. J. Med.* **375**, 1109–1112 (2016).
247. Zhou, X. et al. Exploration of coding and non-coding variants in cancer using genomepaint. *Cancer Cell* **39**, 83–95.e4 (2021).
248. Zhang, J. et al. The International Cancer Genome Consortium data portal. *Nat. Biotechnol.* **37**, 367–369 (2019).
249. Saunders, G. et al. Leveraging European infrastructures to access 1 million human genomes by 2022. *Nat. Rev. Genet.* **20**, 693–701 (2019).
250. Molnár-Gábor, F., Lueck, R., Yakneen, S. & Korbel, J. O. Computing patient data in the cloud: practical and legal considerations for genetics and genomics research in Europe and internationally. *Genome Med.* **9**, 1–12 (2017).
251. Chen, P. H. C. et al. An augmented reality microscope with real-time artificial intelligence integration for cancer diagnosis. *Nat. Med.* **25**, 1453–1457 (2019).
252. Fu, Y. et al. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nat. Cancer* **1**, 800–810 (2020).
253. Kather, J. N. et al. Pan-cancer image-based detection of clinically actionable genetic alterations. *Nat. Cancer* **1**, 789–799 (2020).
254. Coudray, N. et al. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nat. Med.* **24**, 1559–1567 (2018).
255. Parikh, A. R. et al. Liquid versus tissue biopsy for detecting acquired resistance and tumor heterogeneity in gastrointestinal cancers. *Nat. Med.* **25**, 1415–1421 (2019).
256. Laks, E. et al. Clonal decomposition and DNA replication states defined by scaled single-cell genome sequencing. *Cell* **179**, 1207–1221.e22 (2019).
257. Sanders, A. D. et al. Single-cell analysis of structural variations and complex rearrangements with tri-channel processing. *Nat. Biotechnol.* **38**, 343–354 (2020).

Acknowledgements
This work was supported by grants from EMBL (to I.C-C.) and the Harvard Ludwig Center (to P.J.P.) and an award from the Cancer Research UK Grand Challenge and the Mark Foundation for Cancer Research to the SPECIFICANCER team.

Author contributions
All authors contributed to all aspects of the article.

Competing interests
D.C.G. and P.J.P. have filed a patent application on SigMA. I.C-C., J.J.-K.L. and G.E.M.M. declare no competing interests.

Peer review information
Nature Reviews Genetics thanks M. Peifer, J. Zhang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Publisher's note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

RELATED LINKS
1000 Genomes Project: https://www.internationalgenome.org/home
Catalogue of Somatic Mutations In Cancer: https://cancer.sanger.ac.uk/cosmic
cBioPortal: http://www.cbioportal.org/
ClinVar: https://www.ncbi.nlm.nih.gov/clinvar/
COSMIC Mutational Signatures: https://cancer.sanger.ac.uk/signatures
OncoKB: https://www.oncokb.org/
Pan-Cancer Analysis of Whole Genomes (PCAWG) project: https://dcc.icgc.org/pcawg
The Cancer Genome Atlas (TCGA): https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga