nature reviews genetics

https://doi.org/10.1038/s41576-025-00832-3

Review article

Check for updates

Advances in single-cell DNA sequencing enable insights into human somatic mosaicism

Diane D. Shao[1,2]✉, Andrea J. Kriz[2], Daniel A. Snellings[2], Zinan Zhou[2], Yifan Zhao[3], Liz Enyenihi[2,4] & Christopher Walsh[1,2,5]✉

Abstract

DNA sequencing from bulk or clonal human tissues has shown that genetic mosaicism is common and contributes to both cancer and non-cancerous disorders. However, single-cell resolution is required to understand the full genetic heterogeneity that exists within a tissue and the mechanisms that lead to somatic mosaicism. Single-cell DNA-sequencing technologies have traditionally trailed behind those of single-cell transcriptomics and epigenomics, largely because most applications require whole-genome amplification before costly whole-genome sequencing. Now, recent technological and computational advances are enabling the use of single-cell DNA sequencing to tackle previously intractable problems, such as delineating the genetic landscape of tissues with complex clonal patterns, of samples where cellular material is scarce and of non-cycling, postmitotic cells. Single-cell genomes are also revealing the mutational patterns that arise from biological processes or disease states, and have made it possible to track cell lineage in human tissues. These advances in our understanding of tissue biology and our ability to identify disease mechanisms will ultimately transform how disease is diagnosed and monitored.

Sections

[1]Department of Neurology, Boston Children's Hospital and Harvard Medical School, Boston, MA, USA. [2]Division of Genetics and Genomics, Department of Paediatrics, Boston Children's Hospital and Harvard Medical School, Boston, MA, USA. [3]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. [4]Biological and Biomedical Sciences Graduate Program, Harvard Medical School, Boston, MA, USA. [5]Howard Hughes Medical Institute, Chevy Chase, MD, USA. ✉e-mail: Diane.Shao@childrens.harvard.edu; Christopher.Walsh@childrens.harvard.edu

Introduction

Bulk genome sequencing of complex tissues primarily detects early embryonic mutations, and therefore grossly underestimates the full genomic variability that exists among single cells[1,2]. However, it is now clear from single-cell DNA sequencing (scDNA-seq) data that every cell in the body carries mosaic variants and is genetically unique[1–3] (Fig. 1a). Understanding this variability provides insight into the biological processes that drive cellular biology, has the potential to provide prognostic indicators for ageing and disease, and ultimately may highlight mechanisms or genes that can be targeted in disease states.

Unlike single-cell RNA sequencing and single-cell assay for transposase-accessible chromatin sequencing (scATAC-seq), which aim to interrogate only the transcribed and euchromatic regions of the genome, respectively, scDNA-seq aims to capture all regions of the genome evenly. As the human genome is at least 50-fold or 20-fold larger than the transcribed genome and the chromatin-accessible genome, respectively, scDNA-seq has substantially increased sequencing requirements and the associated cost. Moreover, each genomic locus is represented by only two molecules of DNA, making single-cell whole-genome amplification (scWGA) necessary. First achieved in the 1990s[4,5], scWGA involves the challenging task of amplifying the 6 pg of DNA in a diploid human cell by many hundred-fold to generate sufficient quantities for whole-genome sequencing (WGS) and, in principle, allows interrogation of genomic variability at high resolution[6,7]. However, many other barriers must also be overcome, such as limiting amplification bias and distinguishing true, fixed mutations from biological errors (for example, resulting from DNA damage)[8] and technical errors (for example, amplification or sequencing errors)[9]. In addition to being accurate and unbiased, scWGA must also be scalable to capture the true heterogeneity of a population. Early technologies, which focused on the relatively simple task of accurately detecting somatic single-nucleotide variation (sSNV) and large (>10–50 Mb) somatic copy number variants (CNVs), involved trade-offs between accuracy and scalability.

However, recent methodological improvements to scWGA approaches and bioinformatic innovations have mitigated these core issues[10,11]. Commercially available products have made single-cell genomics more broadly accessible and analytical pipelines have been built to handle the unique challenges of scWGA followed by single-cell WGS (scWGS). Most recently, duplex sequencing methods that capture sequence from both the Watson and Crick strands of DNA have been developed[12,13]; these single-molecule sequencing approaches provide single-cell level resolution even from bulk DNA (Fig. 1a). We refer to scWGA/scWGS and single-molecule sequencing techniques together as scDNA-seq, as both approaches capture the variation in the DNA of single cells. Together, these advances have made it possible to study less tractable examples of somatic mosaicism, including in non-cycling cells (such as neurons, heart and muscle), in rare cell types and early in post-zygotic development when limited numbers of cells are available. Mutational patterns and mechanisms can now be determined and high-resolution lineage tracing performed. Finally, scWGS also allows clonally complex tumours or tissues to be studied at a resolution not possible with bulk DNA sequencing. Thus, scDNA-seq technologies are catalysing a new phase of discovery across biomedical research.

In this Review, we describe experimental and computational advances for the study of single-cell level variation using scDNA-seq technologies and highlight key applications of these technologies. We discuss how single-cell sequencing is transforming our understanding of somatic mutations that arise during embryonic development at fine-scale resolution, its utility for shedding light on cell lineage relationships in human tissues, and new understanding of single-cell somatic mutations in biological process as well as disease states. This Review synthesizes key applications of single-cell genomics and future directions for the field.

Single-cell genomic technology

scDNA-seq workflow

The workflow for evaluating genomes of single cells comprises the isolation of single cells or nuclei, scWGA and scWGS, and finally computational processing to correct errors and call variants (Fig. 1b). For most single-molecule variant detection methods, such as duplex sequencing, bulk DNA is used as input, avoiding the need for isolation of single cells or nuclei. For both scWGA/scWGS and single-molecule techniques for scDNA-seq, fresh or frozen tissue is preferred, and it is critical to avoid exposure to heat or fixative or other agents of DNA damage. In our experience, small differences in temperature or buffer conditions can lead to a marked increase in false-positive variants, which easily overwhelm the low abundance true variants in the genome.

For single cell or nuclei isolation, fluorescence activated cell sorting (FACS), laser capture microdissection or other method of choice may be applied. The process of scWGA generates a library with barcodes for individual cells before sequencing. Sequencing depth is chosen based on the application, considering trade-offs between genome coverage and cost of sequencing (Fig. 1c). Although initial data processing is analogous to that of standard WGS approaches, quality control and variant calling methods specific to single cells are required to address the errors and biases of scWGA.

Single-molecule duplex approaches[3,8,14], which barcode Watson and Crick strands of DNA to specifically focus on single-nucleotide variation (SNV) detection, start from bulk DNA input (with the exception of multiplexed end-tagging amplification of complementary strands (META-CS)[10], which starts from single cells). Although duplex methods using bulk DNA cannot distinguish the precise cells from which a variant arises, the overall SNV landscape detected is fully analogous to that of the landscape from sequencing single-cell genomes (Fig. 1a). Specific informatic approaches have been developed for each duplex method to handle the nuances of the molecular barcoding strategy; however, they all rely on the same concept of using paired strands of single DNA molecules for accurate SNV detection.

scWGA

An ideal scWGA technology would maintain high genome coverage, uniformity and allelic balance at high cell throughput. Current scWGA

Fig. 1 | Abundant genetic variation is revealed by single-cell sequencing. **a**, Single-cell or single-molecule resolution technologies to detect DNA variants, together referred to as single-cell DNA sequencing (scDNA-seq), reveal the vast genetic heterogeneity that exists between cells. Bulk sequencing detects primarily clonal mutations. Single-molecule techniques assess the landscape of mutations in a population of single cells, whereas single-cell whole-genome amplification (scWGA)/single-cell whole-genome sequencing can resolve the relationship between variants within a cell. Cells are represented by grey backgrounds. **b**, Workflows for scDNA-seq. **c**, The specific application of scDNA-seq determines whether high genome coverage is required and/or how many single cells are required. Asterisks indicate approaches that use targeted sequencing of specific DNA loci rather than whole-genome sequencing. Indels, insertions and deletions; SNV, single-nucleotide variation; VAF, variant allele frequency.

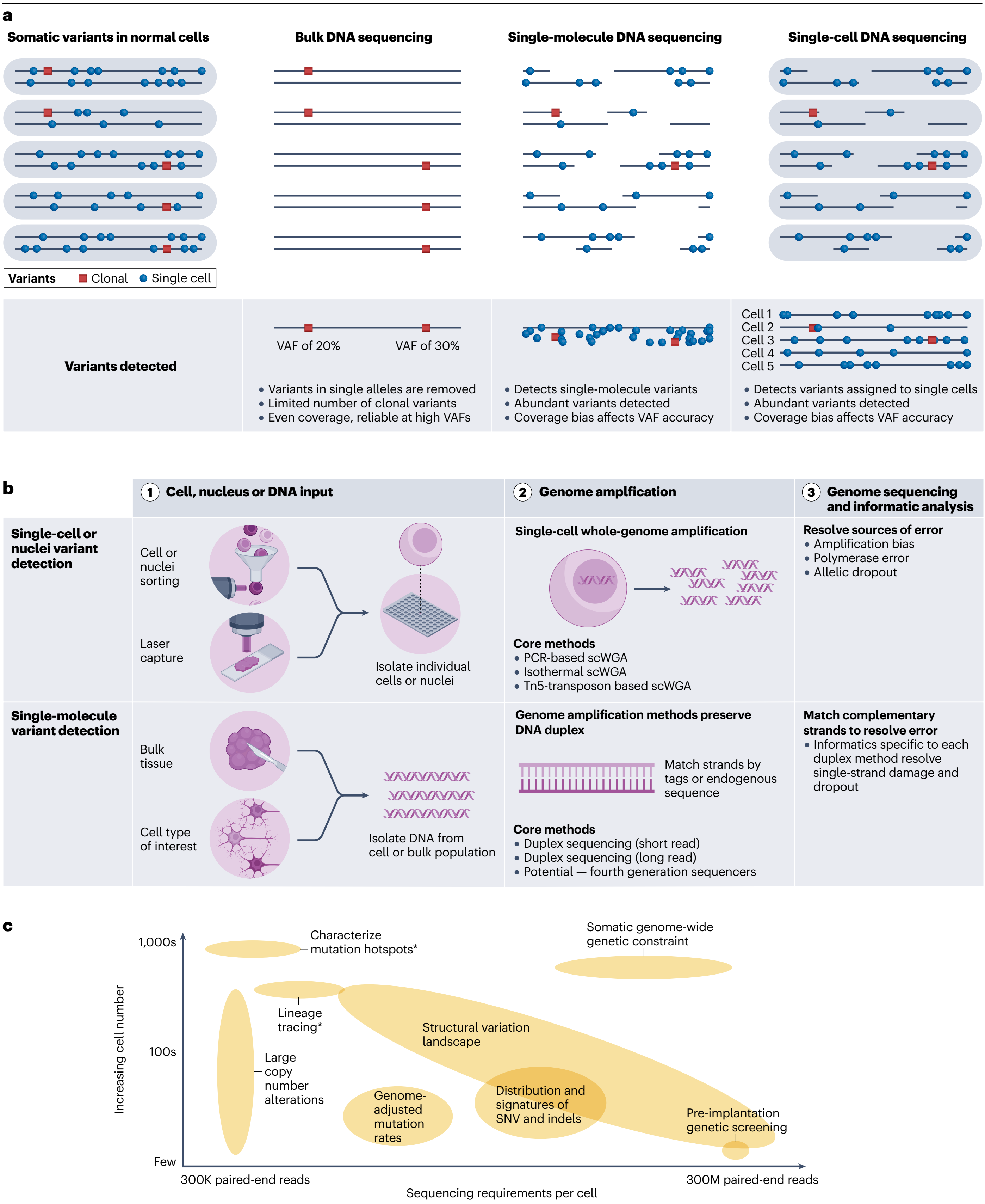
a
Somatic variants in normal cells
Bulk DNA sequencing
Single-molecule DNA sequencing
Single-cell DNA sequencing
Variants 🟥 Clonal ● Single cell
Variants detected
VAF of 20%
VAF of 30%
• Variants in single alleles are removed
• Limited number of clonal variants
• Even coverage, reliable at high VAFs
• Detects single-molecule variants
• Abundant variants detected
• Coverage bias affects VAF accuracy
Cell 1
Cell 2
Cell 3
Cell 4
Cell 5
• Detects variants assigned to single cells
• Abundant variants detected
• Coverage bias affects VAF accuracy
b
1 Cell, nucleus or DNA input
2 Genome amplfication
3 Genome sequencing and informatic analysis
Single-cell or nuclei variant detection
Cell or nuclei sorting
Laser capture
Isolate individual cells or nuclei
Single-cell whole-genome amplification
Core methods
• PCR-based scWGA
• Isothermal scWGA
• Tn5-transposon based scWGA
Resolve sources of error
• Amplification bias
• Polymerase error
• Allelic dropout
Single-molecule variant detection
Bulk tissue
Cell type of interest
Isolate DNA from cell or bulk population
Genome amplification methods preserve DNA duplex
Match strands by tags or endogenous sequence
Core methods
• Duplex sequencing (short read)
• Duplex sequencing (long read)
• Potential — fourth generation sequencers
Match complementary strands to resolve error
• Informatics specific to each duplex method resolve single-strand damage and dropout
c
1,000s
100s
Few
Increasing cell number
300K paired-end reads
300M paired-end reads
Sequencing requirements per cell
Characterize mutation hotspots*
Somatic genome-wide genetic constraint
Lineage tracing*
Structural variation landscape
Large copy number alterations
Genome-adjusted mutation rates
Distribution and signatures of SNV and indels
Pre-implantation genetic screening

Table 1 | Summary of characteristics of exemplar scWGA technology

	DOP-PCR	PicoPLEX	MALBAC	MDA	PTA	LIANTI	DLP+
Coverage	Low (20–25%)	Medium (35–45%)	Medium (55–60%)	Medium (70–75%)	High (~95%)	High (80–85%)	Very low
Uniformity (MAPD)	High (~0.2–0.4)	High (~0.2–0.4)	Medium (~0.4–0.6)	Low (>1)	High (~0.2–0.3)	High (~0.1–0.2)	Medium (ND)
Allelic balance	Low	Medium	Medium	Low	High	Medium	N/A[a]
Cell throughput	1–96 cells	1–96 cells	1–96 cells	1–96 cells	1–384 cells	1–96 cells	>10,000 cells
Time	3h	2.5h	4h	11h	2.5–10.5h	13.5h	21h
Commercial availability	Yes	Yes	Yes	Yes	Yes	No	No
Cost per reaction	~US$20	~US$20	~US$50	~US$10	~US$5 (v2), ~US$20 (v1)	N/A	N/A
Premise	PCR-based, random priming	PCR-based, looped products to encourage amplification of primary template	PCR-based, looped products to encourage amplification of primary template	Isothermal amplification	Isothermal amplification with chain terminators	Tn5-based, in vitro reverse transcription	Tn5-based, microfluidic
Reference	17	20	21	18,19	11	22	27
Advantages	High genomic uniformity; ease of the protocol; relatively low cost	High genomic uniformity; ease of the protocol	Relatively good genomic coverage, uniformity and allelic balance; ease of protocol	Good genomic coverage; low error rate; ease of protocol; relatively low cost	High genomic coverage, uniformity and allelic balance; low error rate; ease of protocol	High genomic coverage, uniformity and allelic balance; low error rate; ease of protocol	High throughput; high-resolution microscopy images allows image-based quality control and ploidy check
Limitations	Low genomic coverage	Low genomic coverage; relatively high cost	Relatively high error rate; relatively high cost	Low genomic uniformity and allelic balance	Relatively high cost	Not commercially available; protocol is relatively complex	Low genomic coverage
Applications	Copy number variant detection	Copy number variant detection	SNV and copy number variant detection	SNV and indel detection	SNV, indel and copy number variant detection	SNV and copy number variant detection	Copy number variant detection

The performance of these single-cell whole-genome amplification (scWGA) methods is summarized from our experience, prior published data[11,107] and preprint studies[108]. DLP+, DNA transposition single-cell library preparation; DOP-PCR, degenerate oligonucleotide primed PCR; indel, insertions and deletions; LIANTI, linear amplification via transposon insertion; MALBAC, multiple annealing and loop-based amplification cycles; MAPD, median absolute pairwise difference; MDA, multiple displacement amplification; N/A, not applicable; ND, no data; PTA, primary template amplification; SNV, single-nucleotide variation. [a]Typically captures one allele owing to low coverage.

methods each have distinct strengths and weaknesses with respect to these parameters (Table 1) and the choice of which technique (and sequencing depth) to use is driven by the desired application (Fig. 1c), although cost is also an important factor. For example, clinical applications of pre-implantation genome-wide genetic screening would require high genome coverage on few cells with low tolerance for error[15,16], whereas studies of large CNVs would require very low but uniform genome coverage that does not require high per-base accuracy. Implementation of distinct methods for the same purpose provides mutual benchmarking of the ground truth. Techniques for scWGA can be broadly grouped into three core categories (Table 1 and Fig. 2a): PCR-based amplification, isothermal amplification and Tn5 transposon-based amplification.

The earliest scWGA methods were based on PCR or isothermal amplification. PCR-based approaches, such as degenerate oligonucleotide primed PCR (DOP-PCR)[17], use thermostable DNA polymerases (such as Taq) and degenerate oligonucleotides to amplify the genome of single cells and provide reasonably high uniformity, but with high allelic dropout and low genome coverage. Isothermal amplification approaches, such as multiple displacement amplification (MDA)[18,19], amplify the genome using Φ29 polymerase, which achieves higher genome coverage and lower error rates than PCR-based approaches; Taq polymerase creates 10^{-4}–10^{-6} errors per nucleotide compared with 10^{-7}–10^{-8} errors per nucleotide for Φ29 polymerase. However, the exponential nature of amplification in MDA increases amplification bias and the number of propagated polymerase errors, compromising the uniformity and accuracy of the amplified genome.

To overcome these limitations, subsequent PCR-based and isothermal amplification-based techniques introduced measures to promote the use of the original copy of the genome as the template for linear pre-amplification, which in each case improves both fidelity to the original DNA template and uniformity. For example, the PCR-based methods PicoPLEX[20] and multiple annealing and loop-based amplification cycles (MALBAC)[21] generate hairpin or looped products that are less optimal substrates for subsequent random priming than the original DNA template. Similarly, primary template amplification (PTA)[11] incorporates exonuclease-resistant terminators that generate short amplicons, which encourages priming from the native DNA template because Φ29 polymerase prefers longer amplicon products.

The earliest implementation of Tn5 transposase for scWGA was in linear amplification via transposon insertion[22] and its derivatives[23]. In these approaches, Tn5 is used to insert a promoter sequence that directs linear amplification of the genome through transcription followed by reverse transcription to generate cDNA. However, the complex protocol limited the implementation of these approaches. In recent years, the widespread availability and simplicity of use of Tn5 transposase has seen it more commonly used to tagment DNA in single cells to incorporate linkers for Illumina library construction (as per the commercially available Nextera kits), followed by PCR-based scWGA[24–26]. However, although this approach is simple to implement, it cannot cover as much of the genome as random priming from PCR or isothermal amplification owing to loss of fragments with symmetric linkers (50% of fragments). The best-validated Tn5-based scWGA methods include the duplex sequencing method META-CS[10] and DNA transposition single-cell library preparation (DLP+)[27], which is implemented in a microfluidic format. META-CS has a higher accuracy of SNV detection because it is duplex based, but has a higher cost per cell; throughput of cells is increased and cost decreased with DLP+.

Despite these technological advances, errors persist in all scWGA sequencing data regardless of the method used to generate them, and these errors, which include allelic dropout, allelic imbalance, single-strand dropout (SSD) and amplification or polymerase errors,

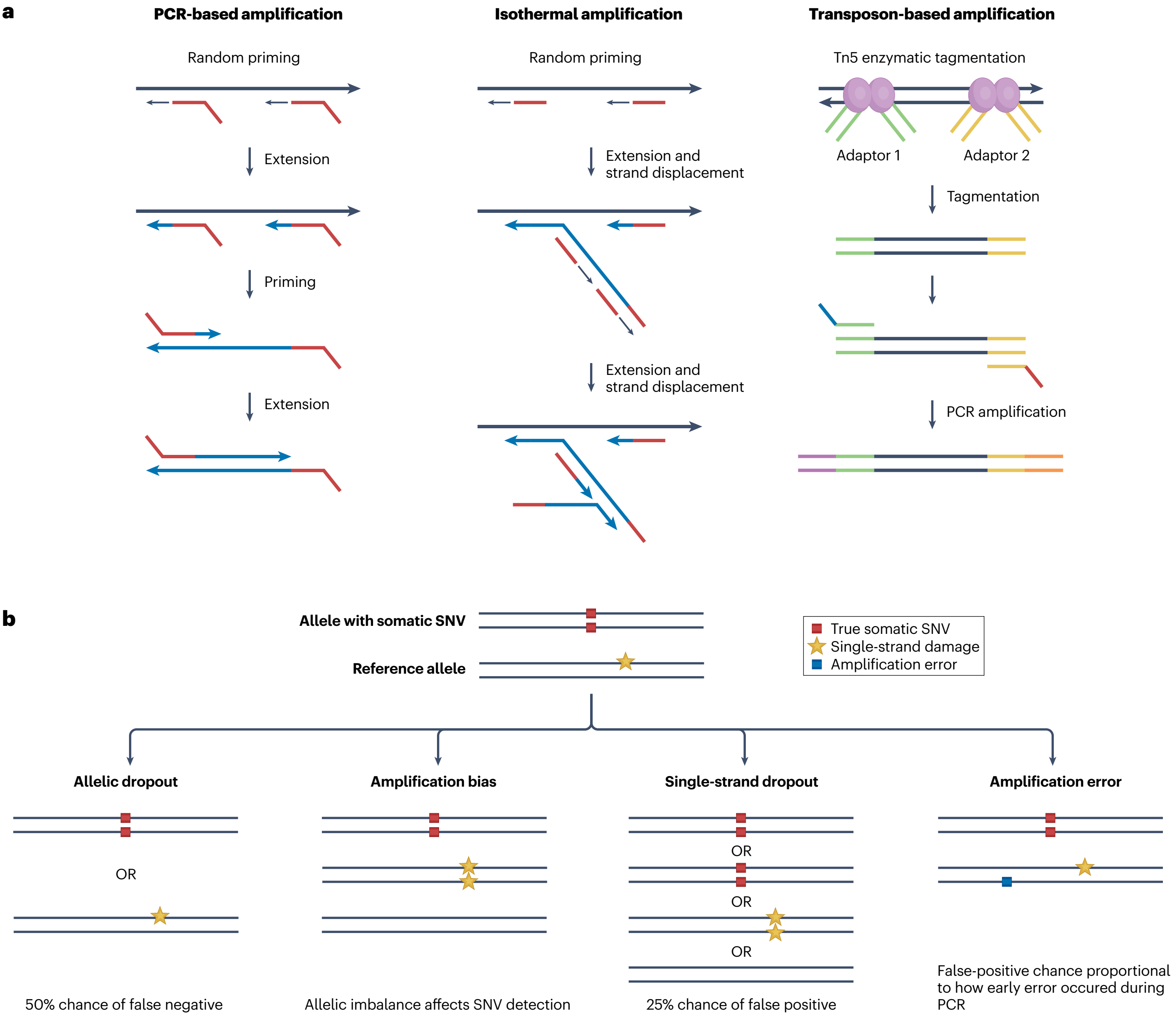

Fig. 2 | Methods of single-cell whole-genome amplification and sources of error in single-cell whole-genome sequencing. a, Schematic workflows for the three core methods for single-cell whole-genome amplification: PCR-based amplification, isothermal amplification and transposon-based amplification. Red indicates priming sites and blue indicates DNA extension. **b**, Theoretical sources and probabilities of false positives and false negatives, which obscure the true genotype of single cells after amplification and sequencing. Outcomes of the first amplification are shown after occurrence of each of the theoretical sources of error. SNV, single-nucleotide variation.

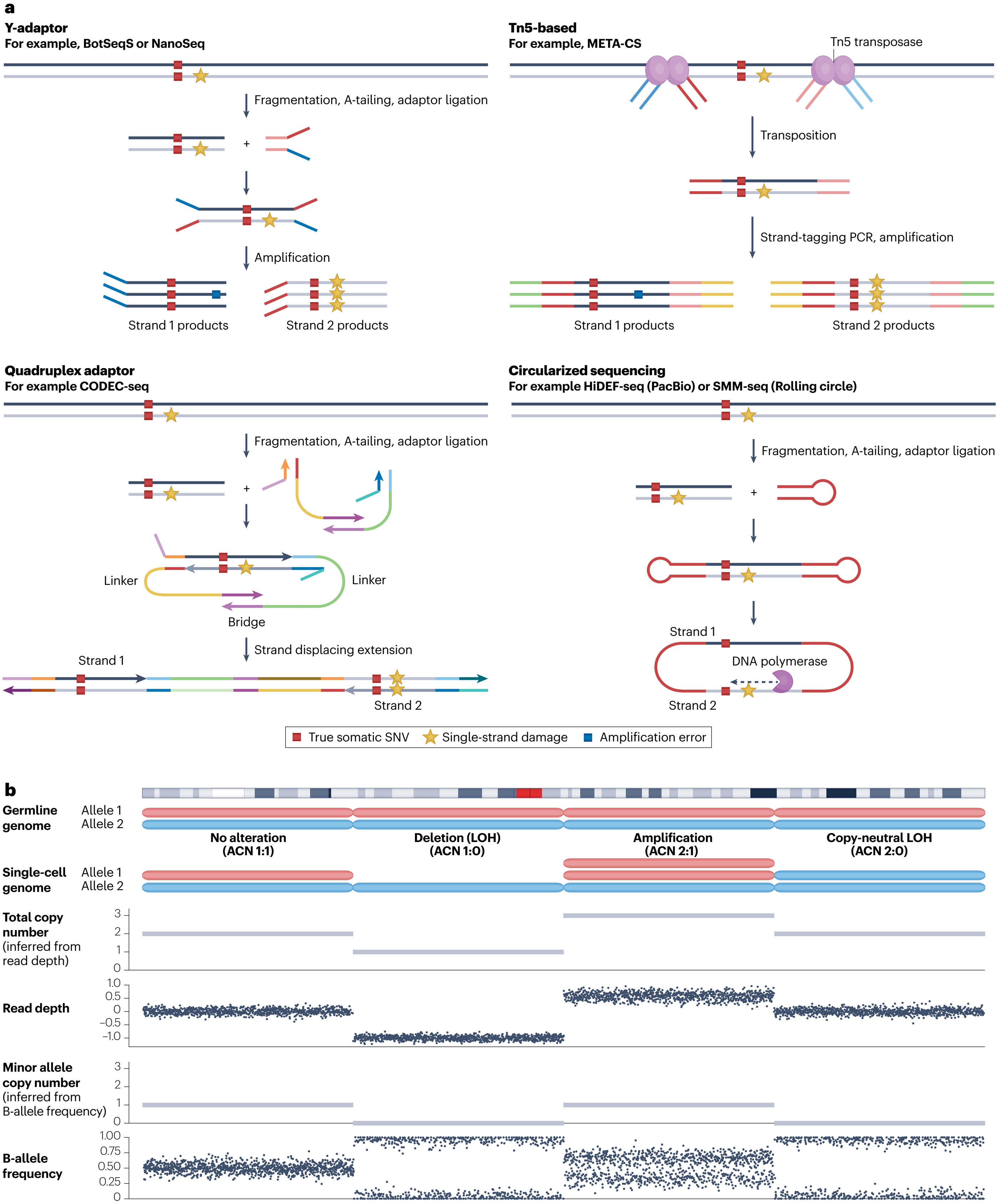
a
Y-adaptor
For example, BotSeqS or NanoSeq
Fragmentation, A-tailing, adaptor ligation
+
Amplification
Strand 1 products
Strand 2 products
Tn5-based
For example, META-CS
Tn5 transposase
Transposition
Strand-tagging PCR, amplification
Strand 1 products
Strand 2 products
Quadruplex adaptor
For example CODEC-seq
Fragmentation, A-tailing, adaptor ligation
+
Linker
Linker
Bridge
Strand displacing extension
Strand 1
Strand 2
Circularized sequencing
For example HiDEF-seq (PacBio) or SMM-seq (Rolling circle)
Fragmentation, A-tailing, adaptor ligation
+
Strand 1
DNA polymerase
Strand 2
🟥 True somatic SNV
★ Single-strand damage
🟦 Amplification error
b
Germline genome
Allele 1
Allele 2
No alteration (ACN 1:1)
Deletion (LOH) (ACN 1:0)
Amplification (ACN 2:1)
Copy-neutral LOH (ACN 2:0)
Single-cell genome
Allele 1
Allele 2
Total copy number (inferred from read depth)
3
2
1
0
Read depth
1.0
0.5
0
−0.5
−1.0
Minor allele copy number (inferred from B-allele frequency)
3
2
1
0
B-allele frequency
1.00
0.75
0.50
0.25
0

Fig. 3 | Methods of error correction in single-cell whole-genome sequencing. **a**, Duplex-sequencing approaches independently sequence the two strands of DNA to call variants, which will show consensus between strands; artefacts caused by errors or damage will be present on only one of the two strands. Four different methods of achieving duplex sequencing are shown: Y-adaptor-based, Tn5-based, quadruplex adaptor-based and circularized sequencing. **b**, Computational error correction of somatic copy number variants typically relies on two signals: total read depth, which indicates the total number of copies present, and B-allele frequency (BAF), which reveals the relative proportion of each allele. Different allele-specific copy number (ACN) states are shown for simulated single-cell whole-genome sequencing data. Combining these two signals allows distinctions to be made between copy number states that may seem similar with just one signal. For example, both diploid (that is, no alteration) and copy-neutral loss-of-heterozygosity (LOH) show no change in total read depth, indicating two total copies, but their BAF patterns differ. In the diploid state, BAF shows an equal ratio between alleles (1:1), whereas in copy-neutral LOH, BAF indicates the presence of two alleles from the same haplotype (2:0). BAF is calculated as the proportion of reads supporting the minor (less frequent) allele at heterozygous SNP sites. When the minor allele is completely lost (as is the case for both deletion (1:0) and copy-neutral LOH (2:0)), BAF is expected to be zero. CODEC-seq, concatenating original duplex for error correction sequencing; HiDEF-seq, hairpin duplex enhanced fidelity sequencing; META-CS, multiplexed end-tagging amplification of complementary strands; SMM-seq, single-molecule mutation sequencing; SNV, single-nucleotide variant.

hinder accurate SNV and/or structural variant calling (Fig. 2b). Allelic dropout or imbalance may occur when one of the two alleles at a genomic locus is lost or preferentially amplified, respectively, during scWGA; both error types result in the loss of information about variants on the lost or under-represented allele and can therefore lead to false positives or false negatives. In practice, false positives are the key source of error as overamplified errors can easily overwhelm the true biological signal. SSD occurs when one strand of a Watson–Crick paired DNA molecule is lost during scWGA; such lesions can be falsely called as a (double-stranded) SNV if the undamaged strand drops out during amplification. As each cell in the human body experiences ~70,000 single-stranded DNA lesions daily[22,28], false positives can greatly outnumber true positives at the single-cell level without appropriate error correction. The polymerase error rate depends on the enzyme used for scWGA. Polymerase errors coupled with the inherent allelic imbalance from scWGA result in unique computational needs to resolve errors in single cells compared with typical bulk sequencing, in which errors can be more reliably eliminated based on variant allele frequency.

Variant detection by duplex sequencing

The detection of variants in single molecules of DNA provides information about the single-cell mutational landscape even when variants cannot be mapped to individual cells. The most common approach to capture single molecules involves duplex sequencing methods, which 'tag' both Watson and Crick strands of the original DNA template (Fig. 3a); this makes it easier to discriminate artefacts (which will be present on only one of the two strands) from real SNVs, which will display 'duplex consensus' between Watson and Crick strands. Thus, duplex sequencing methods have the potential to reduce the false-positive rate of SNV detection to the order of $<10^{-8}$, which is the probability that a polymerase makes the exact same error in two independent amplification reactions.

The original duplex-sequencing method involves ligating Y-adaptors to DNA fragments based on sequence specificity, with each arm of the Y-adaptor containing distinct barcodes for strand discrimination[13] (Fig. 3a). Expanding this approach genome-wide required BotSeqS, which incorporated 'bottleneck' dilution to limit the number of duplexed fragments sent for sequencing, without which the sequencing cost required to identify enough complementary strands genome-wide would be prohibitive[29]. NanoSeq[3] in turn optimizes the error-prone end-repair and A-tailing processes of BotSeqS to increase accuracy by two orders of magnitude (from $<2 \times 10^{-7}$ to $<5 \times 10^{-9}$). An alternative method to improve duplex coverage is to physically link both strands of DNA, so that both strands are included in each sequencing read (Fig. 3a). To accomplish this, concatenating original duplex for error correction (CODEC) replaces the Y-adaptor with a quadruplex adaptor that directly concatenates complementary strands of DNA[14], before amplification. CODEC achieves a comparable accuracy to BotSeqS (error rate ~2.9×10^{-7}) but with ~100-fold fewer reads, and its accuracy theoretically becomes comparable to NanoSeq if a similar end-repair and A-tailing process is incorporated. Finally, all these methods then undergo library preparation through amplification with standard Illumina short-read sequencing primers.

An exciting emerging approach is to directly sequence strands of DNA molecules without amplification using long-read sequencers, as these sequencers by design read sequences from single DNA molecules. The hairpin duplex enhanced fidelity sequencing (HiDEF-seq) method[8] optimizes the PacBio long-read sequencing platform to generate a continuous read comprising multiple copies of concatenated Watson and Crick strands of unamplified DNA (estimated error rate of $<7 \times 10^{-16}$) (Fig. 3a). Before this advance, single-molecule mutation sequencing (SMM-seq) provided a similar conceptual concatenation of Watson and Crick strands, followed by a circularized amplification process that generated multiple copies of the original concatenated template, which were sequenced on Illumina platforms[30]. A preprint article describes an emerging future generation short-read sequencing platform, Ultima Genomics, that captures and sequences both strands of each unamplified DNA molecule, inherently enabling duplex sequencing for accurate base-level nucleotide detection[31]. A potential downside of this method and the HiDEF-seq method is that they require Ultima Genomics and PacBio sequencing platforms, respectively, which are currently less accessible than short-read Illumina sequencers.

All of the duplex sequencing methods described above are used on bulk DNA to detect mutations at the single-molecule level, which provides the full spectrum of mutations that occur in all cells, but cannot indicate which mutations co-occur in the same cell; they currently cannot be applied on a single-cell basis owing to their requirement for DNA fragmentation and adaptor ligation. By contrast, META-CS is a duplex method that can be applied to single cells or single nuclei, with an estimated error rate of less than ~2.4×10^{-8}. It uses Tn5 to insert sequencing adaptors into DNA, with the orientation of the adaptor differentiating the two complementary DNA strands[10] (Fig. 3a). As this approach does not require end repair, it is more efficient and avoids the errors that arise from adaptor-ligation-based methods.

Quality control of sequencing data

Computational tools can be used to assess the quality of, and resolve remaining errors in, single-cell sequencing data generated by scWGA or duplex sequencing. At the preprocessing stage, standard sequencing best practices for quality control apply, including evaluation of mapping rate, appropriate sequence diversity and expected read depth.

Table 2 | Informatic tools used to call copy number variants in single-cell whole-genome sequencing data

Method	Input		Approximate minimum bin size	Non-clonal CNV	Web portal	Year published
	Read depth	B-allele frequency				
Gingko[38]	Y	N	500 kb	Y	Y	2015
Scope[39]	Y	N	200 kb	N	N	2020
DeepCNA[40]	Y	N	500 kb	N	N	2024
CHISEL[41]	Y	Y	5 Mb	N	N	2020
Alleloscope[42]	Y	Y	–	N	N	2021
HiScanner[43]	Y	Y	100 kb	Y	Y	2024 (preprint)

CNV, copy number variant; N, no; Y, yes.

Single-cell-specific metrics include evaluating genome amplification uniformity, typically through a metric such as median absolute pairwise difference (MAPD). Although earlier scWGA methods frequently had high MAPD scores >1, the MAPD for current methods such as PTA and META-CS is in the 0.1–0.5 range; for comparison, the typical MAPD for bulk DNA sequencing is ~0.1 (Table 1). Another relevant metric is genome sensitivity, which is the proportion of a single-cell genome for which SNVs are detectable and can be often estimated based on the fraction of known heterozygous SNVs recovered. Genome sensitivity is correlated but is not equivalent to genome coverage, as some covered regions of a single cell genome may undergo too much allelic imbalance or SSD to be considered in determination of sensitivity. The required genome sensitivity will differ depending on project goals.

Analytical methods resolve error

SNV detection. For scWGA/scWGA data, one of several different computational tools can be used to process amplification errors and call true SNVs or CNVs (reviewed in ref. 32). In general, these tools utilize calling strategies, including joint calling, phylogenetic inference, infinite-sites assumption, estimation of local and global allelic dropout and allelic imbalance, and haplotype phasing[33] with nearby heterozygous single-nucleotide polymorphisms. Specific artefactual patterns of mutation can also be leveraged to reduce the number of false positives, as in the case of SCAN2 (ref. 28).

For single-molecule methods, each duplex sequencing method has its own dedicated variant caller that accounts for their unique approaches for barcoding individual strands (Fig. 3a). Subsequently, SNV calls are typically made based on a user-defined threshold; for example, a typical threshold may be two Watson and two Crick strands with a variant call, and no discordant additional reads. Although insertions and deletions (indels) or other types of variants may be amenable to detection, there has been no systematic study of detection of other variant types using duplex methods.

Structural variant detection after scWGA. The detection of structural variation such as indels and CNVs runs into challenges when single-cell genomes undergo extensive amplification, thus exacerbating problems of amplification bias. Furthermore, the typical lack of full-genome coverage and random priming often creates difficulties for breakpoint detection. Genome coverage and uniformity of a method drive the resolution of CNV detection, so the earliest scWGA methods were used to call large megabase-scale CNVs only[34,35]. So far, only SCcaller and SCAN2 callers have been developed specifically to call somatic indels from MDA and PTA data, respectively[28,36]. Advances in computational methods have enabled ever-improving resolution of CNV calling from scWGA data[37] (Table 2 and Fig. 3b).

The major class of single-cell CNV callers use changes in read depth as the primary signal to identify breakpoints to call copy number states. These methods generally bin sequencing reads into equal-sized windows, normalize read counts against a control or genome average, and then segment the data to identify potential CNVs based on changes in read depth. Read-depth-based CNV callers such as Ginkgo[38], SCOPE[39] and deepCNA[40] operate under the assumption of uniform genome sampling in single cells, and that fluctuations in read depth infer total copy number, with gains reflected by increased copy number and losses reflected by decreased copy number. A pitfall of read-depth-based CNV callers is that they predict total copy number states, but cannot distinguish which of the original two alleles was altered. For example, when the total copy number state is four, these callers cannot distinguish between whether each allele was amplified to two copies (2:2 allelic copy number (ACN)) or whether a single allele was amplified to three copies (3:1 ACN). Such distinctions are critical for understanding the underlying aetiology of CNV.

Recent informatic advances in CNV detection leverage properties in addition to read depth to enable accurate determination of ACN states. For example, the minor allele copy number (that is, the smaller of the two ACN integers) can be determined by estimating B-allele frequency (that is, the ratio of the read depth of the minor non-reference allele to the total read depth at each locus) (Fig. 3b). Allele-specific single-cell callers, including CHISEL[41], Alleloscope[42] and HiScanner[43] (an emerging method reported in a recent preprint article), aggregate allelic signals from multiple adjacent SNPs within the same haplotype block to accurately estimate B-allele frequency and thus infer allelic copy number. Although CHISEL and Alleloscope are designed to detect shared CNV from bulk samples[42] or clusters of shared CNV[41] in cancer samples, HiScanner offers robust detection of rare CNV events across a broad spectrum of sizes, including submegabase events, thereby enabling detection of CNVs reliably in non-clonal single-cell genomes. For transposon-based scWGA data, specific patterns of molecular overlap created by Tn5 tagmentation can be used to distinguish the allelic states of each genomic region[26], thus providing additional validation to current depth-based CNV detection approaches.

Applications of single-cell genomics

As discussed above, recent advances in scDNA-seq have substantially improved the accuracy and resolution of genomic analyses, unlocking new possibilities for exploring genetic variability at the single-cell level. Innovations in scWGA methods, such as PTA, have elevated genome coverage to >90%, while markedly reducing critical pitfalls such as allelic imbalance that previously hindered data reliability. Additionally, single-molecule duplex sequencing approaches now enable the interrogation of mutational landscapes with single-cell resolution, bypassing the technical challenges of single-cell isolation while maintaining exceptional accuracy. These technological strides have expanded the applications of scDNA-seq, for example, enabling the mutational architecture of rare cell populations (for example, during development) to be investigated and cell lineages to be traced in complex human tissues. In clinical settings, it promises to deliver more precise genetic diagnostics from minute tissue samples, novel biomarkers to measure environmental or drug exposure, and new tools to explore the molecular footprints of disease on the genome.

Detecting variation during human development

The study of genome-wide genetic variation in the post-fertilization zygote, which comprises only a few cells, was enabled through single-cell genomics. DOP-PCR showed that 49% of single cells in human early cleavage-stage embryos (~8 cells) are aneuploid[44], which is similar to the aneuploidy rates determined with MDA followed by array-based hybridization[45] (55%). These findings from scWGS were surprising, given that these early cells have the potential to give rise to a large proportion, or all, of the cells of an individual, and implies that cells containing deleterious genetic variants are selected against during embryogenesis[46] (Fig. 4a). The ability to detect variants in single cells during embryogenesis provides new insights and offers the potential to capture germline variants (which are present in every single cell) with minimal material. Simultaneously, the complex single-cell landscape opens new questions to be addressed, as the heterogeneity between single cells may mean that not all variants will be transmitted to the embryo[46].

Current clinical pre-implantation genetic testing is limited to aneuploidy or to targeted loci when parents are known to be carriers[47]; however, scWGS would theoretically enable unbiased genome-wide genetic pre-implantation screening even when parental carrier status is unknown, and utilize only a single cell, reducing potential risk to the embryo. scWGA methods are sensitive enough to detect aneuploidies[15], but subchromosomal CNVs and SNVs are not reliably detected. For SNVs, 26.6% of pre-existing heterozygous loci in control single blastomere samples are erroneously called as homozygous when using MDA[48], which hampers its clinical utility. A preprint article reports that more recent attempts with PTA, the scWGA method with the highest coverage and uniformity, were able to capture SNVs, aneuploid chromosomes and mitochondrial DNA from single cells of donor embryos[49]. This approach was subsequently utilized to confirm variants from clinical biopsies of extra-embryonic tissue (trophectoderm) that validated a known genetic disorder[16]. Many unanswered questions exist that will ultimately need to be answered through scWGS, such as how to interpret variants when there is high genetic heterogeneity between cells of an embryo, or even what the expected level of heterogeneity might be.

The genetic landscape of the developing human brain is also surprisingly variable. Human neuronal progenitors accumulate somatic SNVs at least 80× more rapidly (~25 somatic SNVs per progenitor cell per week during mid-gestation as assessed by sequencing clonal populations of progenitors[50]) than do human postmitotic neurons postnatally (15 somatic SNVs per postmitotic neuron per year as assessed by PTA[51]). Although there has been ongoing debate about the relative frequency of aneuploidy in the brain[52,53], a reanalysis of the data from the above studies detected extrachromosomal circular DNA and complex structural rearrangements in at least a portion of neuronal progenitors and postnatal neurons[54]. In a recent preprint using Tn5 transposon-based scWGA, we showed that human prenatal neurons frequently harbour complex CNV in as much as 40% of neurons during mid-gestation, and these abnormal genomes resolve after birth[26].

As scDNA-seq advances, we will gain clarity on the landscape of genetic variation during development, in the brain and also in other developing tissues that are affected by tissue-specific biology[55] (Fig. 4a).

Lineage tracing in human tissue

Somatic mutations that arise during normal organismal development (at a rate of ~2–4 mutations per division in humans[2,56,57]) are stably inherited by daughter cells and can thus be used as endogenous markers for tracing the lineage of individual cells, theoretically all the way back to the zygote (Fig. 4b). Lineage tracing in model organisms takes advantage of a robust suite of methods that require genetic manipulation, such as introducing a fluorescent marker into a cell[58], which precludes their use in humans. Now, scWGA techniques have made lineage tracing possible from post-mortem human tissue. Lineage-relevant mutations can be detected by either targeted or whole-genome sequencing, and these variants can then be used to reconstruct lineages from the scWGS data[59]. The increased resolution of scWGS yields a more granular lineage tree that reveals relationships between individual cells, whereas prior methods categorize large clones.

Unlike tumours, lineage markers in normal human tissues are unknown a priori and differ between individuals. To circumvent this limitation, lineage-tracing studies often sequence an initial set of single-cell genomes using scWGS to identify variants that serve as lineage markers (Fig. 4b), followed by genotyping of these variants using targeted sequencing in a larger number of cells. In human neurons, this strategy was used to track the dispersion of individual clones across the cortical surface[59,60]. One study that profiled DNA mutations in single adult neurons, with simultaneous RNA capture to classify each cell's type, found that progenitors of excitatory and inhibitory neurons in humans diverge early during development, and that excitatory neurons migrate in an 'inside out' pattern, recapitulating previous observations in animal models[61]. Recent studies have expanded on this work by using PTA, which has high coverage across the genome to detect lineage-informative variants, to track the timing at which brain regions diverge and provide evidence for a dorsal source of cortical inhibitory neurons[62,63]. Beyond the brain, similar methods have been applied to track cell lineage in the haematopoietic system[64], and a recent preprint article reports lineage tracking in the peripheral nervous system[65].

Targeted single-cell sequencing can be used when the lineage-relevant mutations are known to occur in specific regions of the genome, such as those driving tumour evolution, which largely occur in known oncogenes and tumour suppressors. By targeting sequencing to a relatively small number of genes, thousands of cells can be profiled compared with the tens of cells that are typically analysed in scWGS studies (Fig. 4b). This strategy has been leveraged in leukaemia and other blood cancers to determine the order in which the mutations that precede tumorigenesis occur and to understand the functional effect of mutations on clonal expansion[66–70]. The increased throughput of targeted sequencing has enabled longitudinal sampling studies to track tumorigenic clones throughout treatment to identify shifts in clonal composition and locate rare reservoirs of mutated cells[66,68,71]. This approach has also been applied to track the lineage of tumours in the brain[72,73], breast[34,74,75], colon[76,77] and multiple other tissues[78–80].

Although these studies provide just a limited glimpse into the advances enabled by single-cell genomics, they underscore the immediate importance of this technology to both basic and translational science. Looking ahead, an important goal will be to chart mutagenesis across cell lineage trees to understand why different lineages are vulnerable to different classes of somatic mutations[81]. Although lineage-tracing methods have elucidated developmental paths, precisely mapping genomic changes to each cell-state transition remains difficult as retrieving high-quality genomic and cell-state information simultaneously from single cells is technologically challenge. Advances in single-cell multiomics and spatial profiling are rapidly improving the quality and scale of genomic information recovered from single cells, enabling studies that connect mutational dynamics to developmental trajectories.

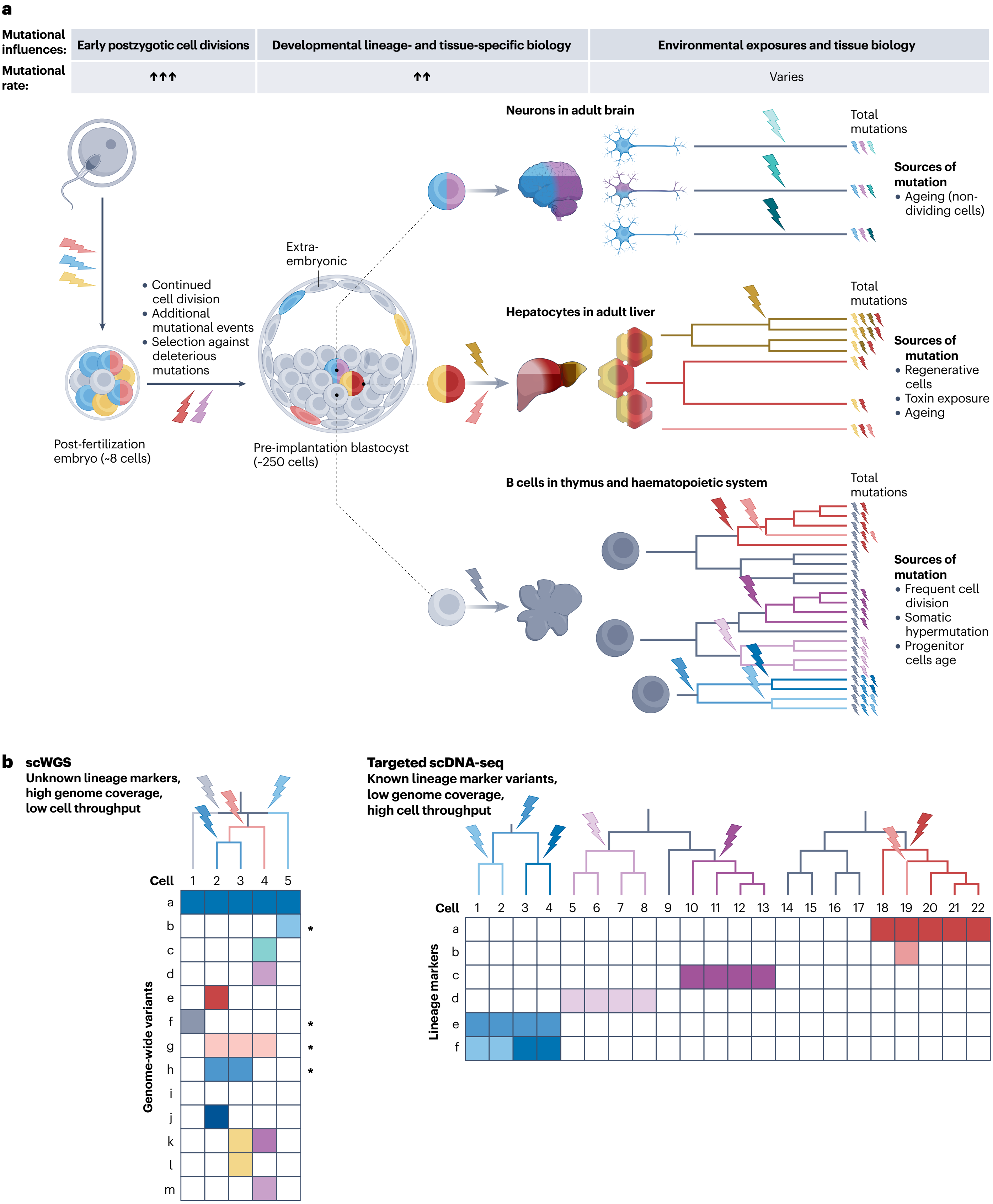
a
Mutational influences:
Early postzygotic cell divisions
Developmental lineage- and tissue-specific biology
Environmental exposures and tissue biology
Mutational rate:
↑↑↑
↑↑
Varies
Post-fertilization embryo (~8 cells)
• Continued cell division
• Additional mutational events
• Selection against deleterious mutations
Extra-embryonic
Pre-implantation blastocyst (~250 cells)
Neurons in adult brain
Total mutations
Sources of mutation
• Ageing (non-dividing cells)
Hepatocytes in adult liver
Total mutations
Sources of mutation
• Regenerative cells
• Toxin exposure
• Ageing
B cells in thymus and haematopoietic system
Total mutations
Sources of mutation
• Frequent cell division
• Somatic hypermutation
• Progenitor cells age
b
scWGS
Unknown lineage markers, high genome coverage, low cell throughput
Cell
1 2 3 4 5
Genome-wide variants
a b c d e f g h i j k l m
*
*
*
*
Targeted scDNA-seq
Known lineage marker variants, low genome coverage, high cell throughput
Cell
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Lineage markers
a b c d e f

Fig. 4 | Mutations that accrue in single cells during development and postnatal life reflect cell states and exposures and mark cellular lineages. **a**, Cell division, tissue biology, ageing and environmental exposures shape the mutational landscape in single cells throughout development and the lifespan. Cells accumulate somatic mutations (lightning bolts) that are stably inherited by daughter cells. **b**, Single-cell whole-genome sequencing (scWGS) captures all variants present in a cell (high genome coverage), but the high cost of sequencing restricts the number of cells that can be assayed (low throughput). Targeted sequencing of a smaller number of known variants (which may have been identified by scWGS) allows for many cells to be assayed but at a lower genome coverage. Asterisks denote lineage-informative markers. scDNA-seq, single-cell DNA sequencing.

Somatic mosaic variation as biomarkers

Both physiological factors (such as age, tissue characteristics and mutagenic environment) and disease processes (such as neurodegeneration and defective DNA repair) create characteristic mutational patterns in the genomes of individual cells, and these patterns can ultimately be used as biomarkers of those events (Fig. 4a). Each mutational process creates a specific mutational signature in the genome that, for SNVs, are typically defined by the trinucleotide context in which the base substitution occurs[82]. For example, spontaneous deamination of cytosine to uracil at CpG islands results in a C>T mutation, which generates a trinucleotide pattern observed as X[C>T]G. Mutational signature analysis was originally limited to cancer cells because the clonal nature of cancer allowed signatures to be detected using standard sequencing techniques[83]. The advent of scDNA-seq has opened an entire field of study by enabling the mutagenic forces acting on somatic genomes in non-cancerous settings to be ascertained accurately, shedding light on tissue biology and disease contexts.

Studies have shown that the number and/or pattern of mutations in single cells can be used as a biomarker of age. For example, single-cell genomic analysis of human brains has shown that neurons accumulate 15–18 sSNV per cell per year[3,51,84]; the total number of sSNVs in an individual neuron is therefore indicative of its age. Furthermore, the age-related accumulation predominantly follows a mutational signature previously characterized in cancer (from sequencing clonal populations of cells across all cancer types) called SBS5 (ref. 82). Moreover, the total genomic mutation burden — which includes mutations arising from cell-intrinsic processes and exposures (both pathogenic and benign)[82,85,86] — correlates with lifespan[87], posing questions about the relationship between longevity and mutations in single cells.

The genetic landscape of an individual cell is also aligned with its cellular identity and cell state. Tissues can harbour both shared and unique mutation signatures, as both the frequency and spectrum of mutations can differ between tissues[3,88–90]. These tissue-specific differences can arise early in development[50,55]; for example, distinctions in the relative abundances of the mutational patterns SBS1 and SBS5 having been discerned between individual cell divisions in the early embryo[1,56]. Even within a tissue, individual cell types can harbour distinct DNA mutational patterns. For example, two common cell types in the brain, neurons and oligodendrocytes, which originally derive from a common progenitor, differ in terms of rates and patterns of SNV, indel and CNV distribution in the genome[43,91]. An improved understanding of mutational patterns specific to cell identity has the potential to translate into new ways to identify cell-of-origin for challenging oncological diagnoses[92] and new biomarkers of tissue health, and also provides a new lens to understand forces that shape each cell and tissue.

Emerging evidence suggests that disease states may also create specific and permanent mutational signatures in DNA of single cells. Ageing in neurons is associated with an increase in C>A mutations, which are correlated with an accumulation of oxidative damage[10,51]. Alzheimer disease, a cognitive age-related neurological disorder, seems to accelerate the accumulation of oxidative damage in neurons, which have an overall higher mutational burden and higher number of functionally damaging variants than age-matched control neurons. Other forms of neurological disease, for example, those related to disorders of DNA repair such as Cockayne syndrome, create unique signatures of both SNV and small CNV in the genome[84,93]. The mutational burden and disease-specific pattern of mutation may impact risk stratification for these and other diseases, including cancers[94]. However, current barriers to interpretation of mutational patterns include a lack of appropriate large-scale references for single-cell genetic variability from which to interpret disease states, as well as the high cost and low precision of scDNA-seq. As further studies elucidate the relationship between mutation signatures and mechanistic aetiologies of disease, these findings may serve as the basis for identifying mechanistic-based interventions.

Thus, although mutational spectra promise to have diagnostic and prognostic value, achieving this goal will require the development of tools that precisely capture the sum total of health-determining physiological and non-physiological exposures to our cells.

Future perspectives

Despite important advances in the field of single-cell genomics, key questions about how somatic mutations contribute to disease remain unanswered, particularly in the context of complex tissues such as the brain. We have not yet begun to understand the relationship between the single-cell genetic landscape and disease processes, germline factors, and the environment or their implications for neurodevelopmental and neuropsychiatric disorders[6,56,95]. Similarly, the mechanistic relationship between somatic mutation burdens and disease progression in neurodegenerative conditions such as Alzheimer disease remains unclear[51,84]. Although scWGS provides information about single-cell genetic variation at unprecedented resolution, multiomics approaches that simultaneously generate, for example, transcriptomic or epigenomic data, could help infer the mechanisms that generate observable mutation signatures and elucidate the effect of DNA mutations on gene regulation and expression. Indeed, methods are emerging that generate both small nuclear RNA-sequencing and scWGS data[23,63]. A particularly exciting future development would be the integration of scWGS with spatial transcriptomics, which could provide further information about the spatial distribution of cells carrying somatic mutations in relation to different cell types in complex tissues, enabling an understanding of not only the mutations themselves, but also how they define tissue architecture and vice versa. Future work will focus on leveraging scWGS for high-throughput translational applications, such as early diagnostics and personalized treatment strategies.

A key area that remains underexplored at the single-cell level is the impact of structural variants. Although large CNVs were initially the most tractable form of genetic variation in early single-cell genomic studies, it has proven difficult to study the many varieties of structural variation at a finer scale. Ongoing efforts are underway to develop single-cell technologies for detection and evaluation of microsatellites[96], transposable elements and complex structural

Glossary

Allelic imbalance
Occurs when both alleles at a genomic locus are not amplified equally, resulting in one allele being under-represented in relation to the other.

Copy number variants
(CNVs). Gains or losses of genomic loci, typically of sizes in the range of hundreds of kilobases up to whole chromosomes.

Genome coverage
Refers to the proportion of genome covered by sequencing reads. For duplex methods, it may refer to the proportion of the genome covered by duplexes.

Indels
Gains or losses of genomic loci, typically of sizes that can be detected within a short-read sequencing read, that is, typically less than 50 nucleotides in length.

Mutational patterns
We use this term colloquially to refer to the collective variation seen in a tissue including mutation rate, distribution and various mutation signatures.

Mutational signatures
Refers to informatic reduction of all mutation calls into vector(s) that summarize the relevant features.

Single-strand dropout
(SSD). Occurs when one of the two strands of DNA that comprise each allele is lost during amplification and can lead to false-positive variant calls if a strand containing single-strand damage or base misincorporation is retained.

Structural variants
All forms of genomic changes larger than ~100 bp, which include copy number alterations, translocations and microsatellite expansion, among others.

Tn5 transposase
An enzyme that simultaneously cuts DNA randomly and inserts DNA adaptors (transposons) in a process called tagmentation.

Uniformity
Measure of evenness of genome coverage, a key quality control metric for single-cell whole-genome amplification. May include measures such a coefficient of variation, Gini index and median absolute pairwise difference (MAPD).

Variant allele frequency
A measure of the proportion of variant alleles in a genomic region; the expected variant allele frequency for variants in a diploid cell are 0%, 50% or 100%.

variation. Improved long-read sequencing technology, which has typically had limited use for single cells owing to its requirements for large amounts of input DNA, will accelerate our understanding of structural variation; long reads can span large and complex genomic regions that are problematic for standard short-read sequencing approaches. Innovative ways of determining long-read-based information have been developed to detect mosaic structural variation; for example, strand-sequencing enables long haplotype phasing by labelling newly replicated DNA during one round of cell division[97,98], but requires actively dividing cells thus limiting broad utility. In addition to improved experimental approaches, computational innovation to overcome amplification bias or correct errors will continue to be needed to provide truly comprehensive assessment of complex structural variants within individual cells.

Cellular single-strand lesions represent another source of single-cell genetic variation that often precedes double-stranded mutations, and results from endogenous cellular processes as well as exogenous exposures[99]. Above, we discussed SSD as the cause of lamentable errors that confound detection of double-strand mutations (see the ‘scWGA’ section); however, SSD also represents real intrinsic biological damage that presents an intriguing opportunity to understand the origins of genetic variation deriving from a single strand of DNA. Duplex sequencing also provides an opportunity to evaluate some forms of single-strand damage in the genome[8,10,100], particularly events that involve base misincorporation on one strand. Although there are many different types of single-strand damage for which specific methods have been developed[101], most methods have not been adapted for single-cell inputs.

Another frontier is assessing how genetic mosaicism contributes to phenotypic diversity across individuals, especially given that early developmental mutations in a few cells can contribute to the pathophysiology of complex diseases much later in life. Our understanding of the landscape of somatic variation in cancer and human germline genetic variation has benefited from vast data resources[102,103] that enable underlying mutational patterns to be determined. Large-scale population-level germline data can be leveraged to identify regions of genetic constraint[103], as well as the forces that shape the distribution of mutations across the genome[104]. Similarly large compendiums of cellular-level data will be required to understand the landscape of somatic mutations, key biological mechanisms and processes that drive mutation, and tissue- or cell-specific mutational constraints in single cells in normal tissue. This challenge is being addressed by consortia that have been established to generate single-cell genomic data for healthy tissues, such as the Brain Somatic Mosaicism Network[105] and the NIH Common Fund Somatic Mosaicism Across Human Tissues Network[106].

Conclusions

The field of single-cell genomics has evolved rapidly in recent years, providing transformative insights across diverse areas of biomedical research. Patterns of mutation in DNA are permanent, and those that are unique to cellular and disease states have the potential to be used as biomarkers that could ultimately transform disease diagnosis, risk assessment and prognosis. However, to ensure that single-cell genomics technology reaches its full potential, it needs to be made more accessible for the broader scientific community and to be integrated with other forms of single-cell data. Current limitations, such as the cost of sequencing, the computing power required for data analysis and ever-evolving benchmarks, will need to be overcome. If these goals are met over the coming decade, the opportunities for scDNA-seq data to provide foundational insights into health and disease will continue to grow.

Published online: 25 April 2025

References

1. Bizzotto, S. et al. Landmarks of human embryonic development inscribed in somatic mutations. *Science* **371**, 1249–1253 (2021).
2. Coorens, T. H. H. et al. Extensive phylogenies of human development inferred from somatic mutations. *Nature* **597**, 387–392 (2021).
3. Abascal, F. et al. Somatic mutation landscapes at single-molecule resolution. *Nature* **593**, 405–410 (2021).
   In this study, single-cell mutational landscapes were characterized using duplex-sequencing (NanoSeq) to capture mutations in single molecules of DNA, enabling the study of multiple different tissue landscapes.
4. Lichter, P., Ledbetter, S. A., Ledbetter, D. H. & Ward, D. C. Fluorescence in situ hybridization with Alu and L1 polymerase chain reaction probes for rapid characterization of human chromosomes in hybrid cell lines. *Proc. Natl Acad. Sci. USA* **87**, 6634–6638 (1990).
5. Zhang, L. et al. Whole genome amplification from a single cell: implications for genetic analysis. *Proc. Natl Acad. Sci. USA* **89**, 5847–5851 (1992).
   This paper was among the first to report whole-genome amplification from a single cell.

6. Garrison, M. A. et al. Genomic data resources of the Brain Somatic Mosaicism Network for neuropsychiatric diseases. *Sci. Data* **10**, 813 (2023).
7. Ha, Y.-J. et al. Comprehensive benchmarking and guidelines of mosaic variant calling strategies. *Nat. Methods* **20**, 2058–2067 (2023).
8. Liu, M. H. et al. DNA mismatch and damage patterns revealed by single-molecule sequencing. *Nature* **630**, 752–761 (2024).
9. Lasken, R. S. Single-cell sequencing in its prime. *Nat. Biotechnol.* **31**, 211–212 (2013).
10. Xing, D., Tan, L., Chang, C.-H., Li, H. & Xie, X. S. Accurate SNV detection in single cells by transposon-based whole-genome amplification of complementary strands. *Proc. Natl Acad. Sci. USA* **118**, e2013106118 (2021).
This article describes the development of META-CS, one of the first duplex approaches able to be applied to single cells for scWGA.
11. Gonzalez-Pena, V. et al. Accurate genomic variant detection in single cells with primary template-directed amplification. *Proc. Natl Acad. Sci. USA* **118**, e2024176118 (2021).
This paper describes the development of PTA, a current widely used approach for scWGA that is based on isothermal amplification and which has high genome coverage and evenness. This study provides a useful comparison of metrics between PTA and other scWGA methods.
12. Kennedy, S. R. et al. Detecting ultralow-frequency mutations by Duplex Sequencing. *Nat. Protoc.* **9**, 2586–2606 (2014).
13. Schmitt, M. W. et al. Detection of ultra-rare mutations by next-generation sequencing. *Proc. Natl Acad. Sci. USA* **109**, 14508–14513 (2012).
One of the first demonstrations of error reduction achieved by duplex sequencing, capturing sequence information of the two strands of DNA independently.
14. Bae, J. H. et al. Single duplex DNA sequencing with CODEC detects mutations with high sensitivity. *Nat. Genet.* **55**, 871–879 (2023).
15. Vendrell, X. et al. New protocol based on massive parallel sequencing for aneuploidy screening of preimplantation human embryos. *Syst. Biol. Reprod. Med.* **63**, 162–178 (2017).
16. Xia, Y. et al. The first clinical validation of whole-genome screening on standard trophectoderm biopsies of preimplantation embryos. *F S Rep.* **5**, 63–71 (2024).
17. Telenius, H. et al. Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer. *Genomics* **13**, 718–725 (1992).
18. Dean, F. B., Nelson, J. R., Giesler, T. L. & Lasken, R. S. Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res.* **11**, 1095–1099 (2001).
19. Zhang, D. Y., Brandwein, M., Hsuih, T. & Li, H. B. Ramification amplification: a novel isothermal DNA amplification method. *Mol. Diagn.* **6**, 141–150 (2001).
20. Langmore, J. P. Rubicon Genomics, Inc. *Pharmacogenomics* **3**, 557–560 (2002).
21. Zong, C., Lu, S., Chapman, A. R. & Xie, X. S. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science* **338**, 1622–1626 (2012).
22. Chen, C. et al. Single-cell whole-genome analyses by linear amplification via transposon insertion (LIANTI). *Science* **356**, 189–194 (2017).
23. Yin, Y. et al. High-throughput single-cell sequencing with linear amplification. *Mol. Cell* **76**, 676–690.e10 (2019).
24. Rohrback, S. et al. Submegabase copy number variations arise during cerebral cortical neurogenesis as revealed by single-cell whole-genome sequencing. *Proc. Natl Acad. Sci. USA* **115**, 10804–10809 (2018).
25. Liu, L. et al. Low-frequency somatic copy number alterations in normal human lymphocytes revealed by large-scale single-cell whole-genome profiling. *Genome Res.* **32**, 44–54 (2022).
26. Shao, D. D. et al. Perinatal reduction of genetically aberrant neurons from human cerebral cortex. Preprint at *bioRxiv* https://doi.org/10.1101/2024.10.08.617159 (2024).
27. Laks, E. et al. Clonal decomposition and DNA replication states defined by scaled single-cell genome sequencing. *Cell* **179**, 1207–1221.e22 (2019).
28. Luquette, L. J. et al. Single-cell genome sequencing of human neurons identifies somatic point mutation and indel enrichment in regulatory elements. *Nat. Genet.* **54**, 1564–1571 (2022).
29. Hoang, M. L. et al. Genome-wide quantification of rare somatic mutations in normal human tissues using massively parallel sequencing. *Proc. Natl Acad. Sci. USA* **113**, 9846–9851 (2016).
30. Maslov, A. Y. et al. Single-molecule, quantitative detection of low-abundance somatic mutations by high-throughput sequencing. *Sci. Adv.* **8**, eabm3259 (2022).
31. Cheng, A. P. et al. Whole genome error-corrected sequencing for sensitive circulating tumor DNA cancer monitoring. Preprint at *bioRxiv* https://doi.org/10.1101/2022.11.17.516904 (2022).
32. Valecha, M. & Posada, D. Somatic variant calling from single-cell DNA sequencing data. *Comput. Struct. Biotechnol. J.* **20**, 2978–2985 (2022).
33. Bohrson, C. L. et al. Linked-read analysis identifies mutations in single-cell DNA-sequencing data. *Nat. Genet.* **51**, 749–754 (2019).
This article reports one of the first applications of haplotype phasing, the basis of a core informatic strategy to distinguish true versus false positive SNV after scWGA (non-duplex methods).
34. Navin, N. et al. Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90–94 (2011).
One of the first studies to use single-cell sequencing to comprehensively analyse CNVs within a tumour, effectively demonstrating how a tumour evolves through the accumulation of genetic alterations at the individual cell level, providing insights into cancer progression pathways.
35. McConnell, M. J. et al. Mosaic copy number variation in human neurons. *Science* **342**, 632–637 (2013).
36. Dong, X. et al. Accurate identification of single-nucleotide variants in whole-genome-amplified single cells. *Nat. Methods* **14**, 491–493 (2017).
37. Zhang, C.-Z. et al. Calibrating genomic and allelic coverage bias in single-cell sequencing. *Nat. Commun.* **6**, 6822 (2015).
38. Garvin, T. et al. Interactive analysis and assessment of single-cell copy-number variations. *Nat. Methods* **12**, 1058–1060 (2015).
39. Wang, R., Lin, D.-Y. & Jiang, Y. SCOPE: a normalization and copy-number estimation method for single-cell DNA sequencing. *Cell Syst.* **10**, 445–452.e6 (2020).
40. Liu, F., Shi, F. & Yu, Z. Inferring single-cell copy number profiles through cross-cell segmentation of read counts. *BMC Genomics* **25**, 25 (2024).
41. Zaccaria, S. & Raphael, B. J. Characterizing allele- and haplotype-specific copy numbers in single cells with CHISEL. *Nat. Biotechnol.* **39**, 207–214 (2021).
42. Wu, C.-Y. et al. Integrative single-cell analysis of allele-specific copy number alterations and chromatin accessibility in cancer. *Nat. Biotechnol.* **39**, 1259–1269 (2021).
43. Zhao, Y. et al. High-resolution detection of copy number alterations in single cells with HiScanner. Preprint at *bioRxiv* https://doi.org/10.1101/2024.04.26.587806 (2024).
44. Palmerola, K. L. et al. Replication stress impairs chromosome segregation and preimplantation development in human embryos. *Cell* **185**, 2988–3007.e20 (2022).
45. Vanneste, E. et al. Chromosome instability is common in human cleavage-stage embryos. *Nat. Med.* **15**, 577–583 (2009).
46. Bolton, H. et al. Mouse model of chromosome mosaicism reveals lineage-specific depletion of aneuploid cells and normal developmental potential. *Nat. Commun.* **7**, 11165 (2016).
47. Lee, V. C. Y., Chow, J. F. C., Yeung, W. S. B. & Ho, P. C. Preimplantation genetic diagnosis for monogenic diseases. *Best Pract. Res. Clin. Obstet. Gynaecol.* **44**, 68–75 (2017).
48. Liang, D. et al. Limitations of gene editing assessments in human preimplantation embryos. *Nat. Commun.* **14**, 1219 (2023).
49. Xia, Y. et al. Genome-wide disease screening in early human embryos with primary template-directed amplification. Preprint at *bioRxiv* https://doi.org/10.1101/2021.07.06.451077 (2021).
50. Bae, T. et al. Different mutational rates and mechanisms in human cells at pregastrulation and neurogenesis. *Science* **359**, 550–555 (2018).
51. Miller, M. B. et al. Somatic genomic changes in single Alzheimer's disease neurons. *Nature* **604**, 714–722 (2022).
This study demonstrates the impact of ageing on somatic mutations and the acceleration of mutation rate in neurological disease states.
52. Cai, X. et al. Single-cell, genome-wide sequencing identifies clonal somatic copy-number variation in the human brain. *Cell Rep.* **8**, 1280–1289 (2014).
53. Chronister, W. D. et al. Neurons with complex karyotypes are rare in aged human neocortex. *Cell Rep.* **26**, 825–835.e7 (2019).
54. Sekar, S. et al. Complex mosaic structural variations in human fetal brains. *Genome Res.* **30**, 1695–1704 (2020).
55. Kuijk, E. et al. Early divergence of mutational processes in human fetal tissues. *Sci. Adv.* **5**, eaaw1271 (2019).
By evaluating two embryonically related fetal tissues, liver and intestine, this study demonstrates that tissue biology influences single-cell mutational landscapes of tissues that diverge early during gestation.
56. Rodin, R. E. et al. The landscape of somatic mutation in cerebral cortex of autistic and neurotypical individuals revealed by ultra-deep whole-genome sequencing. *Nat. Neurosci.* **24**, 176–185 (2021).
57. Park, S. et al. Clonal dynamics in early human embryogenesis inferred from somatic mutation. *Nature* **597**, 393–397 (2021).
58. VanHorn, S. & Morris, S. A. Next-generation lineage tracing and fate mapping to interrogate development. *Dev. Cell* **56**, 7–21 (2021).
59. Lodato, M. A. et al. Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science* **350**, 94–98 (2015).
This paper was among the first to demonstrate that lineage markers identified through scWGA followed by WGS can be used to track developmental trajectories — in this case, of neurons.
60. Evrony, G. D. et al. Cell lineage analysis in human brain using endogenous retroelements. *Neuron* **85**, 49–59 (2015).
61. Huang, A. Y. et al. Parallel RNA and DNA analysis after deep sequencing (PRDD-seq) reveals cell type-specific lineage patterns in human brain. *Proc. Natl Acad. Sci. USA* **117**, 13886–13895 (2020).
62. Kim, S. N. et al. Cell lineage analysis with somatic mutations reveals late divergence of neuronal cell types and cortical areas in human cerebral cortex. Preprint at *bioRxiv* https://doi.org/10.1101/2023.11.06.565899 (2023).
63. Chung, C. et al. Cell-type-resolved mosaicism reveals clonal dynamics of the human forebrain. *Nature* **629**, 384–392 (2024).
64. Weng, C. et al. Deciphering cell states and genealogies of human haematopoiesis. *Nature* **627**, 389–398 (2024).
65. Vong, K. I. et al. Genomic mosaicism reveals developmental organization of trunk neural crest-derived ganglia. Preprint at *bioRxiv* https://doi.org/10.1101/2024.09.25.615004 (2024).
66. Jan, M. et al. Clonal evolution of preleukemic hematopoietic stem cells precedes human acute myeloid leukemia. *Sci. Transl. Med.* **4**, 149ra118 (2012).

67. Gawad, C., Koh, W. & Quake, S. R. Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics. *Proc. Natl Acad. Sci. USA* **111**, 17947–17952 (2014).
68. Morita, K. et al. Clonal evolution of acute myeloid leukemia revealed by high-throughput single-cell genomics. *Nat. Commun.* **11**, 5327 (2020).
69. Albertí-Servera, L. et al. Single-cell DNA amplicon sequencing reveals clonal heterogeneity and evolution in T-cell acute lymphoblastic leukemia. *Blood* **137**, 801–811 (2021).
70. Xu, L. et al. Clonal evolution and changes in two AML patients detected with a novel single-cell DNA sequencing platform. *Sci. Rep.* **9**, 11119 (2019).
71. Maslah, N. et al. Single-cell analysis reveals selection of TP53-mutated clones after MDM2 inhibition. *Blood Adv.* **6**, 2813–2823 (2022).
72. Danilenko, M. et al. Single-cell DNA sequencing identifies risk-associated clonal complexity and evolutionary trajectories in childhood medulloblastoma development. *Acta Neuropathol.* **144**, 565–578 (2022).
73. Dogan, H. et al. Single-cell DNA sequencing reveals order of mutational acquisition in TRAF7/AKT1 and TRAF7/KLF4 mutant meningiomas. *Acta Neuropathol.* **144**, 799–802 (2022).
74. Gao, R. et al. Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. *Nat. Genet.* **48**, 1119–1130 (2016).
75. Kim, C. et al. Chemoresistance evolution in triple-negative breast cancer delineated by single-cell sequencing. *Cell* **173**, 879–893.e13 (2018).
76. Yu, C. et al. Discovery of biclonal origin and a novel oncogene SLC12A5 in colon cancer by single-cell sequencing. *Cell Res.* **24**, 701–712 (2014).
77. Tang, J. et al. Single-cell exome sequencing reveals multiple subclones in metastatic colorectal carcinoma. *Genome Med.* **13**, 148 (2021).
78. Xu, X. et al. Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell* **148**, 886–895 (2012).
79. Guo, L. et al. Single-cell DNA sequencing reveals punctuated and gradual clonal evolution in hepatocellular carcinoma. *Gastroenterology* **162**, 238–252 (2022).
80. Zhang, L. et al. Heterogeneity in lung cancers by single-cell DNA sequencing. *Clin. Transl. Med.* **13**, e1388 (2023).
81. Rockweiler, N. B. et al. The origins and functional effects of postzygotic mutations throughout the human life span. *Science* **380**, eabn7113 (2023).
82. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
Although this study was performed in clonal cancers and not in single cells, it provides key background to understanding mutational signatures and their interpretation in single-cell genomes.
83. Steele, C. D. et al. Signatures of copy number alterations in human cancer. *Nature* **606**, 984–991 (2022).
84. Lodato, M. A. et al. Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science* **359**, 555–559 (2018).
85. Kuijk, E., Kranenburg, O., Cuppen, E. & Van Hoeck, A. Common anti-cancer therapies induce somatic mutations in stem cells of healthy tissue. *Nat. Commun.* **13**, 5915 (2022).
86. Fang, H. et al. Ganciclovir-induced mutations are present in a diverse spectrum of post-transplant malignancies. *Genome Med.* **14**, 124 (2022).
87. Cagan, A. et al. Somatic mutation rates scale with lifespan across mammals. *Nature* **604**, 517–524 (2022).
88. Choudhury, S. et al. Somatic mutations in single human cardiomyocytes reveal age-associated DNA damage and widespread oxidative genotoxicity. *Nat. Aging* **2**, 714–725 (2022).
89. Yang, X. et al. Developmental and temporal characteristics of clonal sperm mosaicism. *Cell* **184**, 4772–4783.e15 (2021).
90. Kennedy, S. R., Salk, J. J., Schmitt, M. W. & Loeb, L. A. Ultra-sensitive sequencing reveals an age-related increase in somatic mitochondrial mutations that are inconsistent with oxidative damage. *PLoS Genet.* **9**, e1003794 (2013).
91. Ganz, J. et al. Contrasting somatic mutation patterns in aging human neurons and oligodendrocytes. *Cell* **187**, 1955–1970.e23 (2024).
92. Salvadores, M., Mas-Ponte, D. & Supek, F. Passenger mutations accurately classify human tumors. *PLoS Comput. Biol.* **15**, e1006953 (2019).
93. Kim, J. et al. Prevalence and mechanisms of somatic deletions in single human neurons during normal aging and in DNA repair disorders. *Nat. Commun.* **13**, 5918 (2022).
94. Huang, Z. et al. Single-cell analysis of somatic mutations in human bronchial epithelial cells in relation to aging and smoking. *Nat. Genet.* **54**, 492–498 (2022).
95. Maury, E. A. et al. Somatic mosaicism in schizophrenia brains reveals prenatal mutational processes. *Science* **386**, 217–224 (2024).
96. Murphy, Z. R., Shields, D. A. & Evrony, G. D. Serial enrichment of heteroduplex DNA using a MutS-magnetic bead system. *Biotechnol. J.* **18**, e2200323 (2023).
97. Falconer, E. et al. DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. *Nat. Methods* **9**, 1107–1112 (2012).
98. Grimes, K. et al. Cell-type-specific consequences of mosaic structural variants in hematopoietic stem and progenitor cells. *Nat. Genet.* **56**, 1134–1146 (2024).
99. Mingard, C., Wu, J., McKeague, M. & Sturla, S. J. Next-generation DNA damage sequencing. *Chem. Soc. Rev.* **49**, 7354–7377 (2020).
100. Zhu, Q., Niu, Y., Gundry, M. & Zong, C. Single-cell damagenome profiling unveils vulnerable genes and functional pathways in human genome toward DNA damage. *Sci. Adv.* **7**, eabf3329 (2021).
101. Zatopek, K. M. et al. RADAR-seq: a RAre DAmage and repair sequencing method for detecting DNA damage on a genome-wide scale. *DNA Repair* **80**, 36–44 (2019).
102. Gerstung, M. et al. The evolutionary history of 2,658 cancers. *Nature* **578**, 122–128 (2020).
103. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
104. Seplyarskiy, V. B. et al. Population sequencing data reveal a compendium of mutational processes in the human germ line. *Science* **373**, 1030–1035 (2021).
105. McConnell, M. J. et al. Intersection of diverse neuronal genomes and neuropsychiatric disease: the Brain Somatic Mosaicism Network. *Science* **356**, eaal1641 (2017).
106. Coorens, T. H. H. et al. The somatic mosaicism across human tissues network. *Nature* (in the press).
107. Biezuner, T. et al. Comparison of seven single cell whole genome amplification commercial kits using targeted sequencing. *Sci. Rep.* **11**, 17171 (2021).
108. Estévez-Gómez, N. et al. Comparison of single-cell whole-genome amplification strategies. Preprint at *bioRxiv* https://doi.org/10.1101/443754 (2018).

Author contributions

All authors researched data for the article and contributed substantially to discussion of the content. D.D.S., A.J.K., D.A.S., Z.Z. and Y.Z. wrote the article. D.A.S., L.E. and C.W. reviewed and/or edited the manuscript before submission.

Competing interests

The authors declare no competing interests.

Additional information

Peer review information *Nature Reviews Genetics* thanks Xiao Dong, Jan Korbel, Ji Won Oh 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.