SINGLE-CELL OMICS

Single-cell genome sequencing: current state of the science

Charles Gawad[1], *Winston Koh*[2,3] *and Stephen R. Quake*[2,3]

Abstract | The field of single-cell genomics is advancing rapidly and is generating many new insights into complex biological systems, ranging from the diversity of microbial ecosystems to the genomics of human cancer. In this Review, we provide an overview of the current state of the field of single-cell genome sequencing. First, we focus on the technical challenges of making measurements that start from a single molecule of DNA, and then explore how some of these recent methodological advancements have enabled the discovery of unexpected new biology. Areas highlighted include the application of single-cell genomics to interrogate microbial dark matter and to evaluate the pathogenic roles of genetic mosaicism in multicellular organisms, with a focus on cancer. We then attempt to predict advances we expect to see in the next few years.

Genetic mosaicism
Occurs when there are at least two genotypes in different cells of the same organism.

Whole-genome amplification
The use of biochemical methods to produce multiple copies of the entire genome.

Cell theory provided an entirely new framework for understanding biology and disease by asserting that cells are the basic unit of life[1]. The subsequent discovery that DNA is the heritable programme that encodes the proteins that carry out cellular functions led to the development of the fields of modern genetics and genomics[2]. Although bulk approaches for studying genetic variation have identified thousands of new unicellular species and determined genetic aetiologies for thousands of human diseases, most of that work has been done at the level of the ecosystem or organism[3,4]. However, we now know that diversity within an ecosystem of unicellular species is far greater than we can accurately measure by studying a mixed group of organisms, and that the genomes within the cells of an individual multicellular organism are not always the same.

Single-cell genomics aims to provide new perspectives to our understanding of genetics by bringing the study of genomes to the cellular level (FIG. 1). These tools are opening up new frontiers by dissecting the contributions of individual cells to the biology of ecosystems and organisms. For example, it is now possible to use single-cell genomics to identify and assemble the genomes of unculturable microorganisms[5], evaluate the roles of genetic mosaicism in normal physiology and disease[6], and determine the contributions of intra-tumour genetic heterogeneity in cancer development or treatment response[7]. However, this field rests on the ability to study a single DNA molecule from individually isolated cells, a process that is technically challenging.

In this Review, we describe the current state of the field, including approaches for cell isolation, whole-genome amplification (WGA), DNA sequencing considerations and sequence data analysis, and highlight how recent progress is addressing some of the technical challenges associated with these approaches. We then discuss how those advancements have begun to fulfil some of the ambitious aspirations for the field in applications such as identifying new features of microbial ecosystems and characterizing human intercellular genetic heterogeneity, in particular in cancer.

Technological challenges

Acquiring high-quality single-cell sequencing data has four primary technical challenges: efficient physical isolation of individual cells; amplification of the genome of that single cell to acquire sufficient material for downstream analyses; querying the genome in a cost-effective manner to identify variation that can test the hypotheses of the study; and interpreting the data within the context of biases and errors that are introduced during the first three steps. To maximize the quality of single-cell data and to ensure that the signal is separable from technical noise, each of these variables requires careful consideration when designing single-cell studies.

Cell isolation. The first step in isolating individual cells from primary samples is to produce a suspension of viable single cells. This is not trivial when working with complex solid tissues, which require mechanical or enzymatic dissociation that keeps the cells viable while not biasing for specific subpopulations. In addition,

[1]*Departments of Oncology and Computational Biology, St. Jude Children's Research Hospital, Memphis, Tennessee 38105, USA.*
[2]*Departments of Bioengineering and Applied Physics, Stanford University, Stanford, California 94304, USA.*
[3]*Howard Hughes Medical Institute, Stanford University, California 94304, USA.*

Correspondence to S.R.Q.
quake@stanford.edu

doi:10.1038/nrg.2015.16
Published online 25 Jan 2016

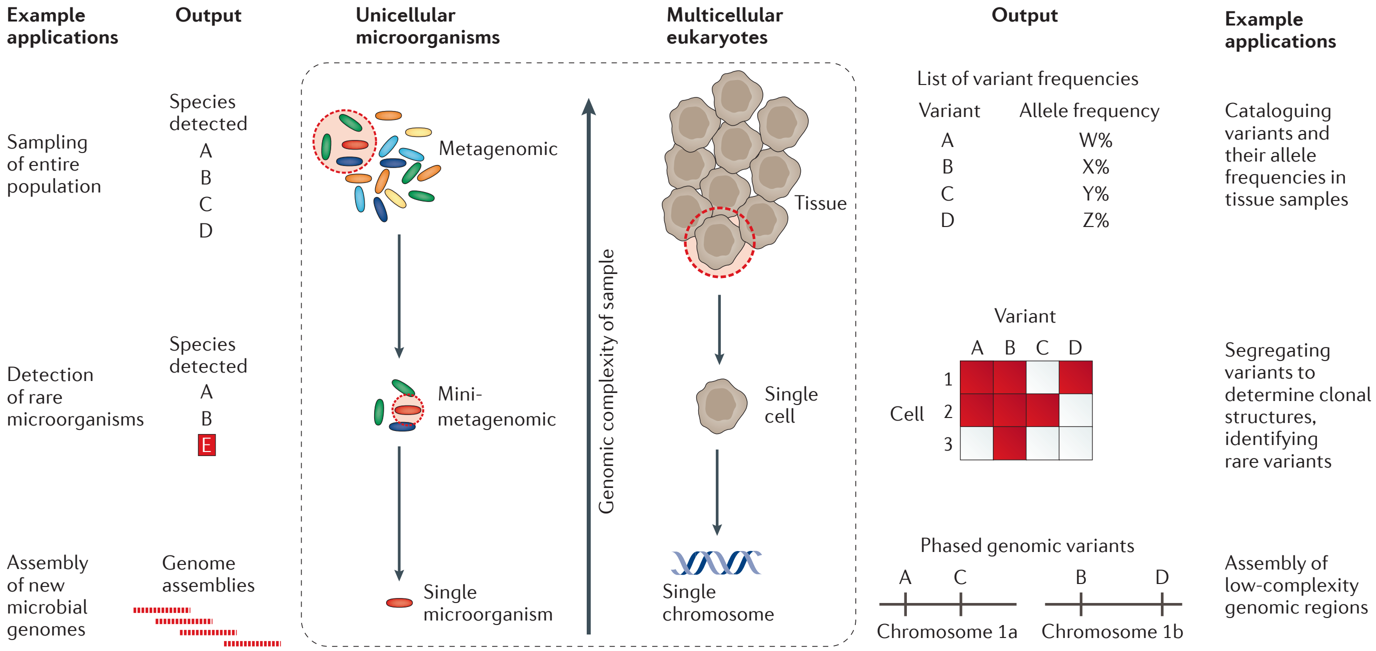

Example applications
Output
Unicellular microorganisms
Multicellular eukaryotes
Output
Example applications
Sampling of entire population
Species detected
A
B
C
D
Metagenomic
Tissue
List of variant frequencies
Variant
Allele frequency
A W%
B X%
C Y%
D Z%
Cataloguing variants and their allele frequencies in tissue samples
Genomic complexity of sample
Detection of rare microorganisms
Species detected
A
B
E
Mini-metagenomic
Single cell
Variant
A B C D
Cell
1
2
3
Segregating variants to determine clonal structures, identifying rare variants
Assembly of new microbial genomes
Genome assemblies
Single microorganism
Single chromosome
Phased genomic variants
A C B D
Chromosome 1a
Chromosome 1b
Assembly of low-complexity genomic regions

Figure 1 | **Opportunities enabled by single-cell sequencing strategies.** Single-cell approaches provide higher-resolution views of the genomic content of samples by reducing the complexity of the genomic signal through the physical separation of cells or chromosomes. Shown in the centre of the figure are schematics of the different levels of cellular complexity analysed by approaches in microorganisms and multicellular eukaryotes. Extending outwards are simplified diagrams of data outputs from these approaches and examples of applications, for unicellular microorganisms (extending to the left) and multicellular eukaryotes (extending to the right). For microbial genomics, decreasing the number of organisms enables the detection of rare microorganisms in a sample. Single-cell sequencing allows for the assembly of the entire genomes of new microorganisms. Single-cell sequencing of multicellular organisms can reveal rare genetic variants and provide information on the co-occurrence of mutations and evolutionary history of samples. Single chromosome sequencing allows for the phasing of variants across a genome.

diseased tissues can have different dissociation kinetics when compared with their normal counterparts, as well as varied dissociation between samples of the same disease. Standard digestion protocols for commonly studied tissues, as well as vigorous approaches for optimizing the dissociation of rare or diseased tissues, are areas that require further development. Laser-capture microdissection[8] provides a low-throughput way of isolating DNA from single cells in their native spatial context, but the quality of sequencing data derived from microdissected single cells has been relatively poor. Finally, microfluidic and bead-based methods have been developed to specifically enrich for single circulating tumour cells (CTCs), as reviewed in detail elsewhere[9]. Environmental microbial samples also require efficient lysis of bacteria, with requirements that can be highly variable between species[10].

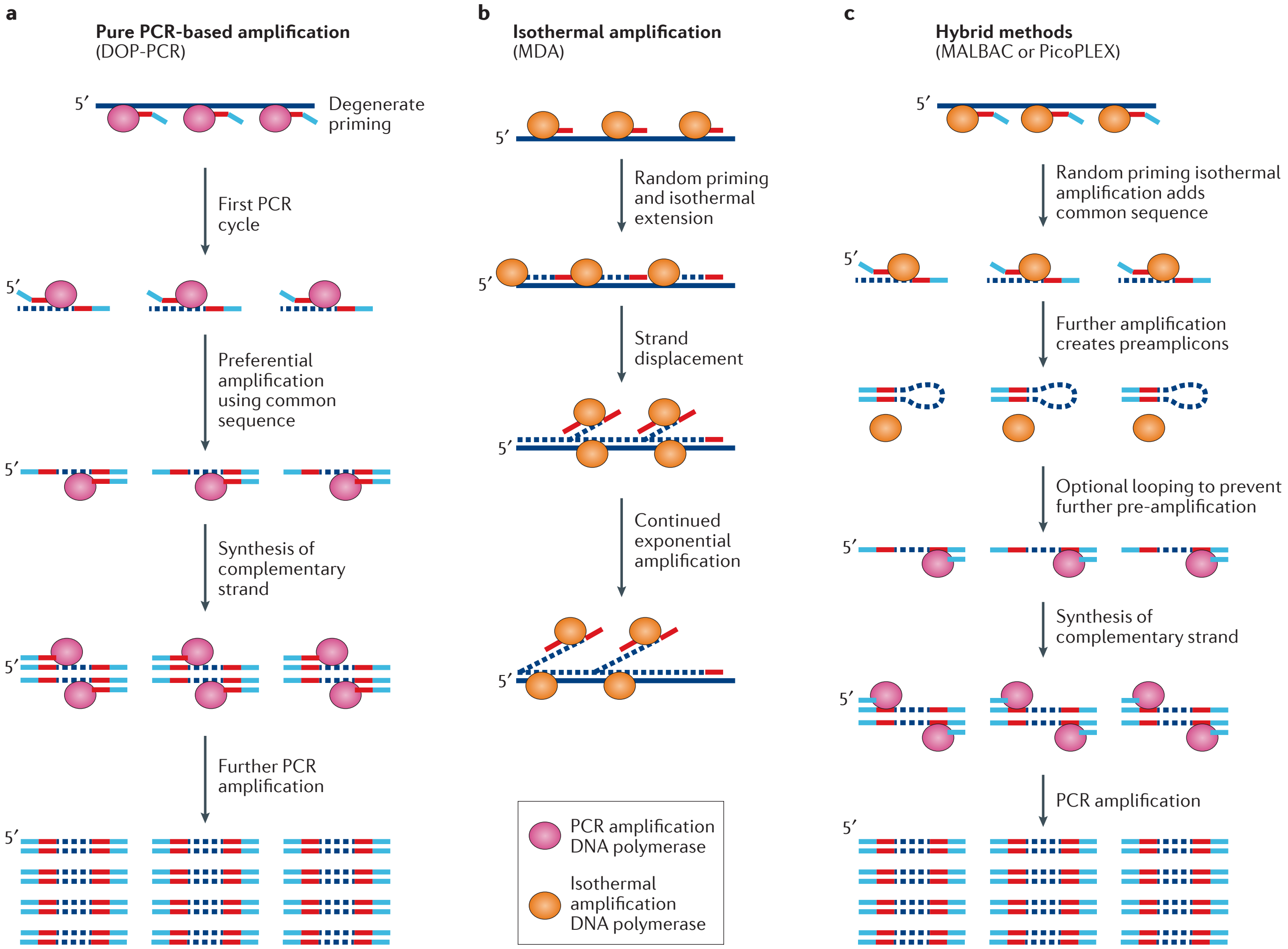
Once in suspension, several approaches have been developed to isolate single cells. These include methods that require manual manipulation, such as serial dilution[11], micropipetting[12], microwell dilution[13] and optical tweezers[14]. In addition, several protocols have been developed to isolate intact cells or nuclei using fluorescence-activated cell sorting (FACS)[15]. Nuclear isolation has the advantage of enabling single-cell sequencing on frozen tissue, which has not yet been demonstrated with other methods[16]. For microbial samples, depending on the environmental source additional sample preparation and FACS setting considerations can also be required[17]. Finally, automated micromanipulation methods that use droplets or micromechanical valves in microfluidic devices are entering mainstream use[18–20]. Regardless of the method used, it is also important to accurately confirm that a single cell has been physically isolated so that spurious biological conclusions are not made after evaluating chambers that are empty or contain multiple cells. In the ideal case this can be accomplished by obtaining microscopy data from each chamber or well containing a single cell.

Various single-cell isolation technologies have recently been reviewed, where the trade-offs in accuracy, throughput, reproducibility and ease of use were highlighted[9,21,22]. Most of the studies using these technologies have been done to illustrate feasibility using a small number of cells. Addressing many of the fundamental biological questions that are uniquely approachable with single-cell genomics will require the interrogation of thousands of cells, making it more likely that technologies that are scalable through parallelization, such as microfluidic-based approaches, are adopted for the long term. In addition, identifying scalable methods for single-cell isolation is an area of active research that is likely to continue producing innovative new tools that will improve all the capture performance metrics.

Optical tweezers
Devices that use a laser to manipulate submicron particles, such as bacterial cells or cellular macromolecules.

Chimaeras
Amplification artefacts formed when two previously disconnected genome regions are combined on the same DNA molecule.

Whole-genome amplification. Another critical component of obtaining genetic information from single cells is to amplify the single copy of a genome while minimizing the introduction of artefacts, such as amplification bias, genome loss, mutations and chimaeras. WGA has been an area with substantial progress over the past 10 years, which has been reviewed in detail elsewhere[21] (FIG. 2). Briefly, the first group of methods that attempted to amplify entire human genomes from single cells coupled PCR amplification with either common sequences interspersed throughout the genome[23], a common sequence ligated to sheared genomes[24],


a
Pure PCR-based amplification
(DOP-PCR)
5′
Degenerate priming
First PCR cycle
5′
Preferential amplification using common sequence
5′
Synthesis of complementary strand
5′
Further PCR amplification
5′
b
Isothermal amplification
(MDA)
5′
Random priming and isothermal extension
5′
Strand displacement
5′
Continued exponential amplification
5′
PCR amplification DNA polymerase
Isothermal amplification DNA polymerase
c
Hybrid methods
(MALBAC or PicoPLEX)
5′
Random priming isothermal amplification adds common sequence
5′
Further amplification creates preamplicons
Optional looping to prevent further pre-amplification
5′
Synthesis of complementary strand
5′
PCR amplification
5′

d

	PCR-based (DOP-PCR)	**Isothermal** (MDA)	**Hybrid** (MALBAC or PicoPLEX)
False-negative rate (coverage and allelic dropout)	High	Low	Intermediate
Non-uniformity	Low	High	Low
False-positive rate (amplification error rate)	High	Low	Intermediate

Figure 2 | **Overview of the three main whole-genome amplification methods. a** | Pure PCR methods such as degenerate oligonucleotide primed PCR (DOP-PCR) use random priming followed by PCR amplification, which preferentially amplifies specific sites in the genome. This results in low physical coverage of the genome, but better uniformity of amplification. **b** | Isothermal methods such as multiple displacement amplification (MDA) use random priming followed by isothermal exponential amplification using a polymerase with high processivity and strand displacement activity. These methods can cover most of the genome, but have much less uniformity. **c** | Hybrid methods such as multiple annealing and looping based amplification cycles (MALBAC) and PicoPLEX have an initial isothermal preamplification, in which common sequences are added, followed by PCR amplification using those sequences. These methods have intermediate coverage and uniformity when compared with pure PCR and isothermal methods. **d** | Summary of false-positive, non-uniformity and false-negative rates for the three main classes of amplification as taken from REFS 29,34,35.

or degenerate or random oligonucleotide priming[25,26] (FIG. 2a). In practice, these methods have resulted in loss of signal from the majority of the genome during the amplification owing to differences in the density of common sequences or variability in PCR efficiency between loci, which are exacerbated when starting with a single genome copy. Starting with two genome copies by sorting out tetraploid nuclei improves the recovery of the genome to about 10% using degenerate oligonucleotide primed PCR (DOP-PCR). However, it is unclear whether selection biases are introduced when selecting for rapidly dividing cells[15]. A recent modification of this method has extended the approach to diploid cells[16]. Of note, these methods use thermostable polymerases, which have higher error rates than thermolabile polymerases, resulting in more mutations introduced during the amplification process.

The second category of WGA is based on isothermal methods (FIG. 2b). The most commonly used approach is multiple displacement amplification (MDA), which uses isothermal random priming and extension with Φ29 polymerase, which has high processivity, a low error rate and strand displacement activity[27,28]. These methods produce greater genome coverage than the initial PCR-based methods, with lower error rates owing to the higher fidelity of Φ29 polymerase[29]. However, the exponential amplification results in overrepresentation of the loci that are amplified first, which is exacerbated by greater fold amplification[29]. It is unclear whether the overrepresentation of specific loci is due to stochastic or systematic biases. In addition, Φ29 polymerase activity results in the formation of a low level of chimeric sequence side products[30,31], which can be reduced with endonuclease treatment allowing the physical separation of the amplicons by debranching the reaction[32].

In an attempt to overcome the low coverage of PCR-based methods and lack of uniformity of isothermal approaches, two quite similar hybrid methods have been developed. Both of these methods use a limited isothermal amplification followed by PCR amplification of the amplicons generated during the isothermal step[12,33] (FIG. 2c). As the name implies, 'displacement DOP-PCR' (also known as PicoPLEX) uses degenerate primers in the first step to add a common sequence, followed by priming of the common sequence for subsequent PCR amplification[33]. Most recently, multiple annealing and looping-based amplification cycles (MALBAC) uses a similar protocol, with the exception of using random primers, as well as new common sequences and temperature cycling that are claimed to promote looping of the isothermal amplicons to inhibit further amplification before the PCR step, which may result in more uniform amplification[12].

In practice, the most commonly used WGA methods in current single-cell studies are the isothermal and hybrid methods. We recently compared these methods using serial dilutions of *Escherichia coli* DNA, as well as single bacterial cells[29]. Both MDA and MALBAC could successfully amplify genomes from single cells, but when amplification was carried out in microlitre reaction volumes in tubes, a significant amount of extraneous contaminant DNA was also amplified. This contaminant DNA was largely eliminated by moving to a microfluidic format that used nanolitre volumes. Bias in amplification was different between the two methods; for MDA bias depended strongly on the amount of gain, whereas for MALBAC it was largely independent of gain. MDA and MALBAC had roughly similar bias when the MDA gain was limited by nanolitre volumes in microfluidic devices. In addition, MALBAC was better at measuring copy number variation but MDA had a significantly lower false-positive rate. These findings suggest that the amplification method should be carefully chosen for each experiment based on the type of genetic variation that will be interrogated.

Two recent reports looked at similar performance metrics in human cells. Both reports compared DOP-PCR, MDA and MALBAC[34,35]. The first report[34] found that MDA had better coverage than MALBAC (84% versus 52%), which resulted in higher detection rates of single nucleotide variants (SNVs; 88% versus 52%). However, MALBAC and DOP-PCR had more uniform coverage, resulting in greater sensitivity and specificity for the detection of copy number variants (CNVs) of >1 Mb. Interestingly, some cells amplified by MDA had comparable CNV detection rates to MALBAC and DOP-PCR; it is unclear what variables result in the lack of reproducible MDA uniformity. The second report[35] also found that MDA had greater coverage breadth (84%) than MALBAC (72%) and DOP-PCR (39%), with MALBAC and DOP-PCR having greater uniformity (coefficient of variation 0.10 versus 0.14, respectively) than MDA (coefficient of variation 0.21). In addition, they found that the isothermal methods had lower false-positive rates, but with more false-negative variance between experiments. The authors note that MALBAC had a lower allelic dropout (ADO) rate than MDA, although MALBAC only covered 72% of the genome, so the ADO rate of 21% was probably calculated only using covered sites. Consequently, it is unclear from this analysis which method had a lower false-negative rate, as MDA covered more of the genome but lost more variant alleles at heterozygous sites owing to less uniform amplification. These studies largely confirm our conclusion from the study of single bacterial genomes that there is no clear winner in amplifier performance and that each approach has strengths depending on the metric of interest (FIG. 2d).

Previous reports had also shown a significant decrease in contamination when single-cell WGA was performed in a microfluidic device[36]. In addition, it has been shown that using nanolitre volumes of microfluidics devices results in more uniform MDA when compared to traditional microlitre reactions[31]. A microfluidic device has recently been developed to perform MALBAC[37], but it is unclear whether the performance of MALBAC will be further improved by carrying out the reactions in enclosed nanolitre amplification chambers. A recent study that used nanolitre volumes in microwells for MDA WGA (in a technique termed microwell displacement amplification system (MIDAS)) claimed to further improve amplification uniformity

Gain
The extent to which a genome undergoes amplification.

Allelic dropout
Loss of one allele of a locus that can occur during whole genome amplification.

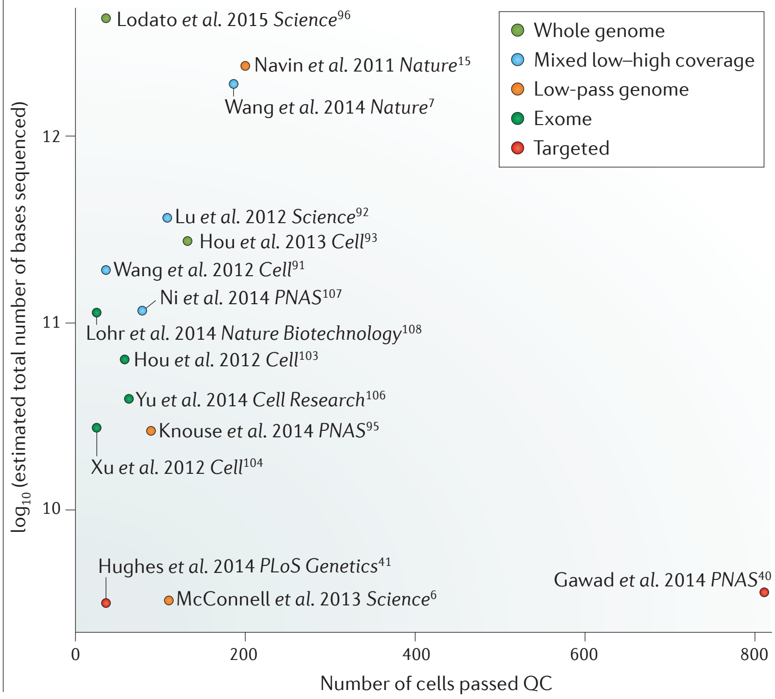
Structural variants
Variation in the genome that occurs as a result of the joining of two previously disconnected genomic locations. A subset of structural variation is copy number variation, which occurs when portions of the genome are amplified or deleted.

and reported single-cell amplification with equivalent uniformity to bulk MDA amplification; this level of performance would be striking and awaits independent confirmation[13]. Performing single-cell MDA in microfluidic emulsions seems to markedly improve the uniformity of amplification, and multiple groups have had success with this approach[38,39].

Interrogation of WGA products. The next step in single-cell genomic analyses is to determine how the amplified genomes will be interrogated. Broadly speaking, for complex eukaryotic genomes such as the human genome, one can choose to query specific loci of interest (typically <1 Mb), sequence all of the protein-coding regions (the exome; 30–60 Mb) or sequence the entire genome (3 Gb). As seen in BOX 1 and TABLE 1, each of these approaches has trade-offs in coverage, propensity for specific types of errors and cost per cell evaluated. The type of genomic interrogation also needs to be carefully considered in the context of the questions being addressed by the study, by taking the biases of the WGA method into account.

Box 1 | Cost considerations when designing single-cell sequencing experiments

Single-cell genome sequencing can provide new insights into heterogeneous human tissue samples. However, owing to the large genome size compared with microorganisms, costs need to be carefully considered. To ensure adequate and appropriate data are generated to address the hypotheses of the study, the balance between the number of cells sequenced and breadth of the genome of each cell that will be queried needs to be taken into account when designing an experiment. A comparison of total bases sequenced as an estimate of sequencing costs and number of cells interrogated (those cells that passed quality control criteria) are highlighted for several recent sequencing studies (see the figure). As noted, there are some general strategies that have been undertaken. The first is to sequence a large portion of the genome of a small number of cells. This approach would be necessary for studies that want to identify variants in non-coding regions. A second approach is to limit sequencing to the exome, which allows increasing cell numbers for performing *de novo* mutation detection. The final strategy is to limit the genomic space queried so that large numbers of cells can be queried. This method could be used to segregate mutations and determine clonal structures using mutations first detected in a bulk sample. Some have used combinations of low and high sequencing depth where low-pass sequencing of a larger number of cells is used for specific hypotheses while sequencing larger portions of the genomes of a smaller number of cells addresses complementary questions from the same samples.



Targeting specific locations of the genomes of single cells can help to focus on areas that have the greatest biological contribution to the system being studied while reducing sequencing costs and false mutation discoveries. Smaller target regions are less likely to contain errors that were introduced during the first few rounds of WGA that would be propagated to result in the erroneous identification of a genetic variant (known as a false-positive variant call). Furthermore, using the bulk sample as a reference can reduce false-positive variant calls by requiring concordance of variant calls in the bulk and single cells, although this limits the mutation discovery space to variants identified in the bulk sample. Targeting can be mediated by target-specific amplification using PCR, or target capture through hybridization. Target-specific amplification provides more uniform target coverage than capture-based methods, which is important when trying to maximize coverage of a genome that has already undergone non-uniform amplification[40]. Target capture more easily provides greater coverage breadth[41], although parallel target-specific amplification using microfluidic devices can significantly increase coverage without large increases in effort.

Single-cell exome sequencing allows broader genome interrogation, which can be used to identify variants that are unique to each of the cells. However, as the size of the genome region interrogated increases, the probability that false variants will be discovered also increases — especially when using polymerases with higher error rates with the PCR-based WGA methods (C.G. and S.R.Q., unpublished observations).

The entire genomes of single cells can also be interrogated. Again, this comes with the trade-off of increased false mutation discovery and cost with the ability to query a larger proportion of the genome. Whole-genome sequencing (WGS) of single cells also removes the additional decrease in uniformity that occurs as a result of targeted or exome capture; thus, WGS can facilitate the detection of SNVs and CNVs. In addition, WGS can catalogue non-coding and structural variants that may contribute to the biological system being studied. However, this comes at a cost of requiring roughly 30-fold more sequencing per cell relative to exome sequencing, which may become limiting if working with many cells.

Overview of single-cell sequencing errors. One of the major challenges of analysing single-cell genomics data is to develop tools that differentiate technical artefacts and noise introduced during single-cell isolation, WGA and genome interrogation from true biological variation. During single-cell isolation, the population of cells being interrogated can be biased through selection of cells based on size, viability or propensity to enter the cell cycle. Consequently, it is necessary to compare the variant alleles detected in the single cells to the bulk population to ensure there was no selection bias. This can be done by comparing the percentage of single cells with a variant to the variant allele frequency in the original bulk sample[40].

As detailed in FIG. 3, numerous errors are introduced during WGA, including loss of coverage, decreased coverage uniformity, allelic imbalance, ADO and errors during genome amplification. Most published papers have attempted to quantify rates for some or all of these

Table 1 | **Overview of technical aspects of major single-cell cancer sequencing studies**

	Isolation method	WGA method	Number of subjects	Total number of cells evaluated	Number of cells passed QC	Regions sequenced	Estimated false-positive rate	Coverage breadth of pass-QC cells at the specified minimum depth (%)	Allelic dropout of covered sites (%)	Estimated false-negative rate[¶]
Papers with high-coverage depth sequencing for SNV detection										
Xu *et al.* 2012 *Cell*[104]	Micropipetting	MDA	1	*	25	Exome	*	80 (5×)[§]	16[§]	28
Hou *et al.* 2012 *Cell*[103]	Micropipetting	MDA	1	90	58	Exome	6×10^{-5}	70 (5×)[§]	43[§]	52
Yu *et al.* 2014 *Cell Research*[106]	Micropipetting	MDA	1	78	63	Exome	*	50 (6×)[§‡]	10 (5×)[§‡]	55
Lohr *et al.* 2014 *Nature Biotechnology*[108]	CTC enrichment followed by micropipetting	MDA	2	*	25	Exome	2.5×10^{-5}	51 (10×)[§]	30[§‡]	64
Hughes *et al.* 2014 *PLoS Genetics*[41]	FACS	PicoPLEX	3	*	36	1,953 loci	*	52 (25×)[§]	12[§]	54
Gawad *et al.* 2014 *PNAS*[40]	Microfluidics	MDA	6	1,487	811	200–300 loci	*	95 (5×)[§]	20[§]	15
Lodato *et al.* 2015 *Science*[96]	Sorting nuclei	MDA	3	*	36	Genome	1.6×10^{-7}	84 (10×)[§]	8[§]	20
Papers with low-coverage depth for CNV detection or mutation phasing										
Navin *et al.* 2011 *Nature*[15]	Sorting tetraploid nuclei	DOP-PCR	2	*	200	Genome	*	6[§]	*	NA
Hou *et al.* 2013 *Cell*[93]	Micropipetting	MALBAC	8	370	132	Genome	*	32[§]	*	NA
McConnell *et al.* 2013 *Science*[6]	Sorting nuclei	DOP-PCR	3	*	110	Genome	*	*	*	NA
Knouse *et al.* 2014 *PNAS*[95]	Micropipetting nuclei	DOP-PCR	4	*	89	Genome	*	*	*	NA
Papers with mixed low–high coverage										
Wang *et al.* 2012 *Cell*[91]	Microfluidics	MDA	1	31 (low pass), 8 (higher coverage)	31 (low pass), 8 (high pass)	Genome	4×10^{-9} (5× coverage in haploid cells)	38 (1×, high pass)[§]	*	NA
Lu *et al.* 2012 *Science*[92]	Micropipetting	MALBAC	1	93 (1×), 6 (5×)	93 (1×), 6 (5×)	Genome	*	43[§]	*	NA
Wang *et al.* 2014 *Nature*[7]	Sorting tetraploid nuclei	DOP-PCR or MDA	2	*	100 DOP-PCR (low pass), 4 MDA (WGS), 91 MDA (WES)	Genome or exome	1.2×10^{-6}	81 (1×, WGS)[§], 93 (1×, WES)[§]	10 (WES)[‖]	12 (WES)
Ni *et al.* 2014 *PNAS*[107]	CTC enrichment followed by micropipetting	MALBAC	11 (CNV), 5 (SNV)	*	49 (WES), 29 (low pass)	Exome or low pass	*	*	*	NA

Table 1 (cont.) | **Overview of technical aspects of major single-cell cancer sequencing studies**

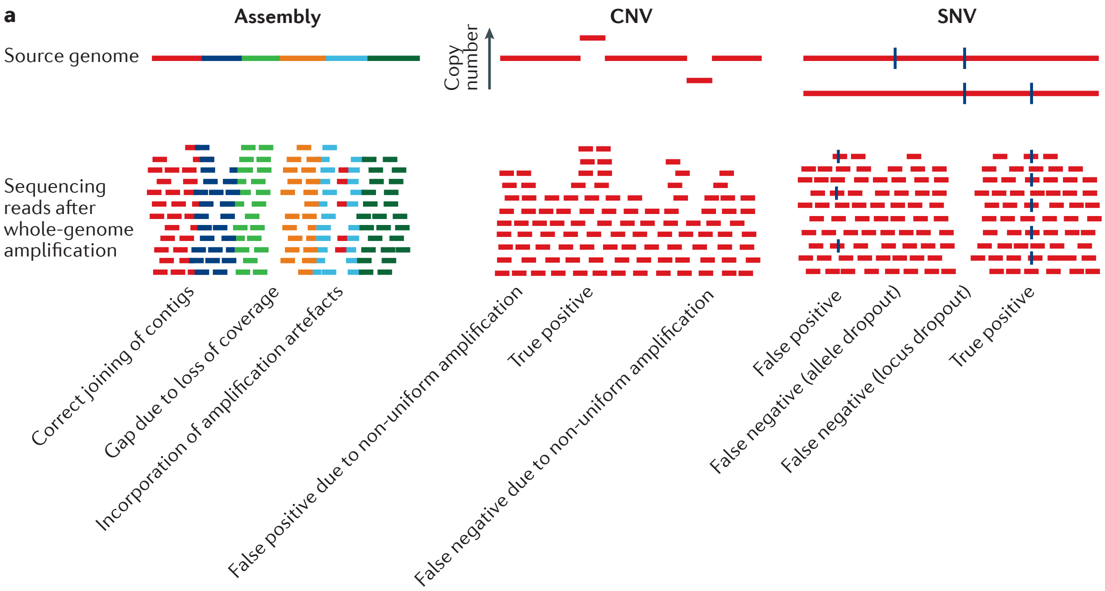
	Isolation method	WGA method	Number of subjects	Total number of cells evaluated	Number of cells passed QC	Regions sequenced	Estimated false-positive rate	Coverage breadth of pass-QC cells at the specified minimum depth (%)	Allelic dropout of covered sites (%)	Estimated false-negative rate[¶]
Papers using qPCR to detect CNV/SNV										
Potter *et al.* 2013 *Genome Research*[109]	FACS	Target-specific amplification for SNV and CNV	2	515	515	7 loci	*	*	*	NA
Papaemmanuil *et al.* 2014 *Nature Genetics*[110]	FACS	Target-specific amplification for SNV and CNV	2	269	269	6 loci	*	*	*	NA

Various strategies for combining methods for isolating cells, amplifying their genomes and interrogating the amplification products have been used to begin to obtain insights into cancer biology. A comparison of these strategies is presented, along with their associated false-positive and false-negative rates. The choice of methods used needs careful consideration to ensure that sufficient cells are analysed and that error rates are low enough to address the hypotheses. CNV, copy number variant; CTC, circulating tumour cell; DOP-PCR, degenerate oligonucleotide primed PCR; FACS, fluorescence-activated cell sorting; MALBAC, multiple annealing and looping-based amplification cycles; MDA, multiple displacement amplification; NA, not applicable; QC, quality control check; qPCR, quantitative PCR; SNV, single nucleotide variant; WES, whole-exome sequencing. *Data not presented in study. ‡Data not stated in text of studies but were estimated by plots of coverage or allele dropout. §Data taken from primary cells. ‖Data produced with cell lines. ¶False-negative rate estimated as 1-(Coverage+0.5(allele dropout)).

errors. However, many studies use cell lines to carry out quality control analyses, followed by experiments on primary samples. This makes it difficult to compare protocols between studies, as it is unclear whether similar performance can be obtained on the primary samples for each of these studies relative to the cell lines that were used for protocol optimization. One must be particularly mindful that certain cell lines or cell types may not be diploid; they can be highly aneuploid or even polyploid, and this affects experimental performance enormously. In addition, various metrics are applied in a quality control step in which the cells are categorized into a subset that meets the chosen criteria and are used to draw biological conclusions, versus a subset that is discarded. Some of the quality metrics used include visual confirmation of an individually isolated cell, WGA product qualification and/or quantification, genome coverage and ADO. Two recent studies developed methods to predict the breadth of genome coverage using low-pass sequencing, which could provide a low-cost approach for assessing cell lysate quality in larger eukaryotic genomes[42,43]. However, comparing single-cell genomics studies is currently difficult, as most studies do not report the total number of cells evaluated, the quality of the data from the discarded cells or the metrics used for the quality control categorization (TABLE 1). Finally, the definition of ADO is not uniform across studies. Some studies do not include loci where both alleles have dropped out of the single cells in their ADO calculation, which artificially reduces those ADO values relative to those studies that include all loci. In practice, the determination of clonal structures is hampered by loss of somatic variants, which occurs always when the locus drops out and occurs about 50% of the time when one of the alleles drops out. A less ambiguous term is 'false-negative rate', which would take into account both allelic and locus dropout. An additional consideration for microorganism sequencing is changes in lysis and/or amplification efficiency between species owing to cellular or genome differences. A recent study compared errors and assembly performance between species[44]. Hence, more uniform analysis and reporting methods are needed to facilitate data interpretation between single-cell studies and provide accurate performance metrics for each approach.

Single-cell variant calling. Although numerous errors are introduced during WGA, tools and strategies are now being developed to overcome the additional technical noise created with WGA, allowing the identification of true variation. SNV calling requires coverage of a variant allele at a rate that exceeds the sum of the amplification and sequencing error rates. More specifically, mutations introduced during the amplification, as well as the allelic imbalances that occur during genome amplification, must be taken into account when calling variants (FIG. 4). There are two basic strategies to overcome the false-positive variants introduced as artefacts of the amplification. First, the bulk sample can be used as a reference to reduce the false discovery rate[40]. Second, when using only the single-cell data, two or three cells can be required to have the same variant at the same location, which is unlikely to occur by chance with the several thousand mutations introduced during single-cell WGA in a 3 Gb human genome[12]. However, the actual number of cells required to call a mutation has not yet been rigorously tested based on the size of the genomic region interrogated. To overcome allelic imbalance, we need variant calling algorithms that are designed to take the technical noise into consideration. One strategy is to require that all variant calls be above the level of technical noise in control samples, which

Somatic variants
Changes in the genome of an organism that are not present in germ cells and can thus not be passed on to offspring.


a
Assembly
CNV
SNV
Source genome
Copy number
Sequencing reads after whole-genome amplification
Correct joining of contigs
Gap due to loss of coverage
Incorporation of amplification artefacts
False positive due to non-uniform amplification
True positive
False negative due to non-uniform amplification
False positive
False negative (allele dropout)
False negative (locus dropout)
True positive

b Magnitude of deleterious effects of specific genome amplification errors on single-cell applications

	Assembly	CNV detection	SNV detection
False negative rate (coverage and allelic dropout)	Large	Small	Large
Non-uniformity	Intermediate	Large	Intermediate
False positive rate (sequencing and amplification errors)	Intermediate	Small	Large

Figure 3 | **Effects of various error types on specific single-cell sequencing applications. a** | *De novo* assembly of genomes (left) is hampered by gaps that are due to loss of coverage and incorporation of amplification artefacts into contigs. Copy number variant (CNV) detection (middle) requires amplification to be sufficiently uniform so that amplification noise can be differentiated from true variants. Single nucleotide variant (SNV) detection (right) requires coverage to detect the variants while not detecting false-positive variant calls that are introduced by the amplification or sequencing. **b** | Summary of the effects that various amplification artefacts have on specific applications, as taken from REFS 29,34,35.

should not have variants[40]. Another approach is to decrease the sequencing error rate by using molecular barcoding[7]. Finally, algorithms are beginning to be developed to correct errors in single-cell sequencing data[45]. Nonetheless, more tools that incorporate all single-cell amplification errors are needed to optimally carry out variant calling in single-cell data.

CNV detection relies on algorithms such as hidden Markov models, circular binary segmentation and rank segmentation, which can normalize noisy coverage data after single-cell WGA to identify regions that are over- or under-represented compared with a diploid genome[12,46,47]. CNV detection algorithms are currently being developed to specifically address the technical artefacts introduced during specific types of single-cell WGA[47,48]. Chimaera formation can create false structural variants, although unless they occur at the beginning of the amplification they should be much less abundant than the corresponding wild-type sequences. This is important for both identifying structural variation in sequencing data and when constructing contigs for *de novo* genome assemblies. In addition, assemblies are hampered by loss of coverage and uneven coverage, which results in truncated or artefactual sequences in assembled genomes. Several assemblers have been created to specifically address these challenges[49,50], and it is likely that further progress will be made in the coming years.

Determining genetic relationships between single cells. General strategies for clustering gene expression and other large data sets have depended heavily on distance functions that provide a quantitative measurement of the differences between pairs of samples[51]. Within the context of single-cell sequencing, we require that these distance functions be robust to missing data as a result of false-negative variant detection. We have found that Jaccard distance is best suited for genotype data, as it is binary in nature[52]. However, we also observed that in general the false-negative rate can hinder statistical determination of the number of clones in a sample.

Molecular barcoding
Attaching a unique sequence to each molecule as a strategy to more accurately count nucleic acids by correcting for experimental artefacts. This approach is also used to decrease false-positive mutation call rates due to sequencing errors by creating a consensus genotype for each molecule.

An alternative to distance-based methods is to perform model-based clustering[53], which allows the inclusion of false-negative errors modelled as binomial processes. Model-based clustering is a soft clustering approach. Unlike methods that subdivide phylogenetic trees that have been generated by distance-based clustering, model-based clustering provides probabilities that a cell originates from the different clones. As seen in the example in FIG. 5, the observed single-cell data are represented as a binary matrix that is first considered to be derived from a mixture of an unknown number of clones with some data missing. Parameters in the model, such as the probability of a particular single cell originating from a specific clone, as well as the false-negative rate, can be estimated across a distinct number of possible clones using an expectation–maximization (EM) algorithm[54]. The challenge of determining the number of clones is then reduced to selecting the statistical model that best describes the observed single-cell data using Bayesian or Akaike information criteria[55]. There is also a hybrid approach based on obtaining an initial estimate of the number of clones derived from distance-based hierarchical clustering, which increases the convergence speed of the computationally intensive model-based methods[56].

After estimating the number of clones in a sample and determining which clone each cell belongs to, a consensus clonal mutation profile can be established. We have done this using mutation frequency cutoff values that exceed the false-negative rate[57], although more rigorous statistical methods could be developed. After determining the clonal genotype, the relationships between clones can be determined. There are a number of algorithms used in evolutionary biology that can be applied to establish clonal structures[58], such as those based on maximum parsimony, maximum likelihood or distance-based methods such as unweighted pair group method with arithmetic mean, neighbour joining and minimum evolution algorithms[58,59]. We prefer the modified use of directed minimum spanning trees, as they can be rooted and allow us to readily include ancestral clones as internal nodes of the evolutionary tree are identified.

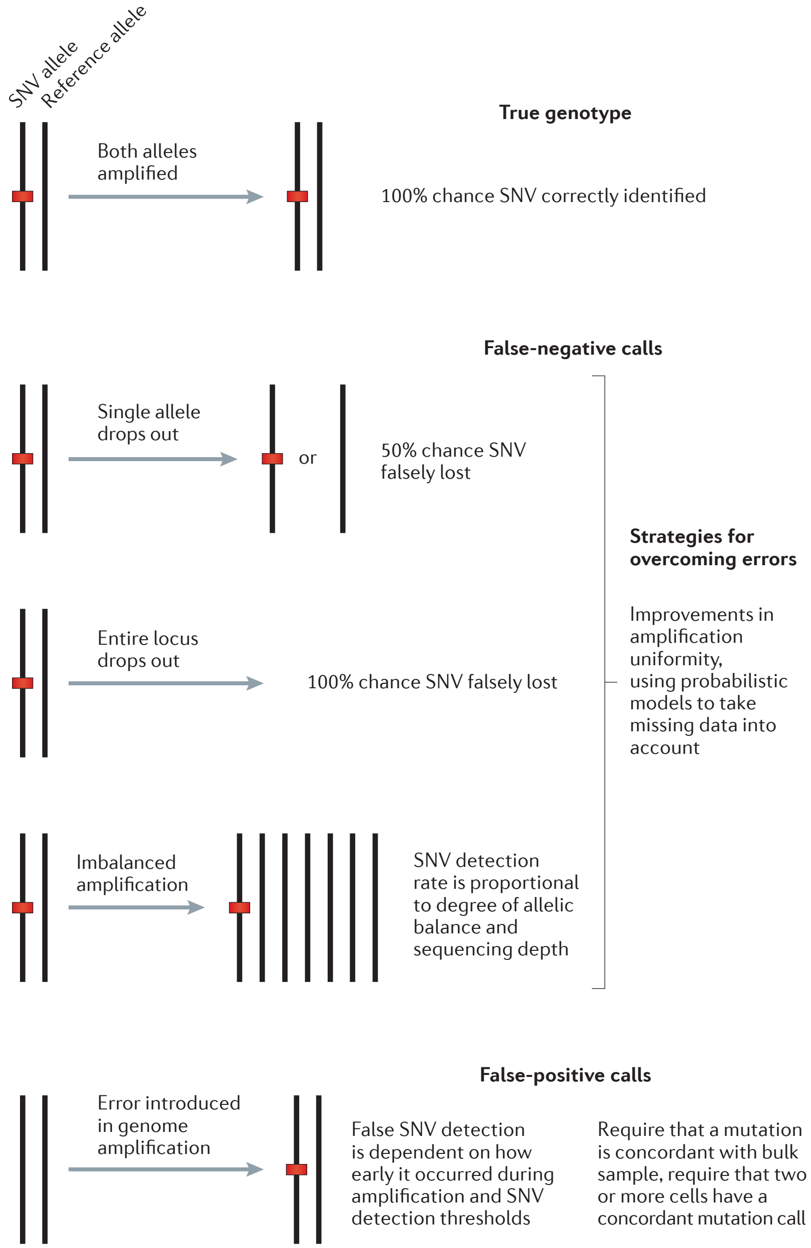
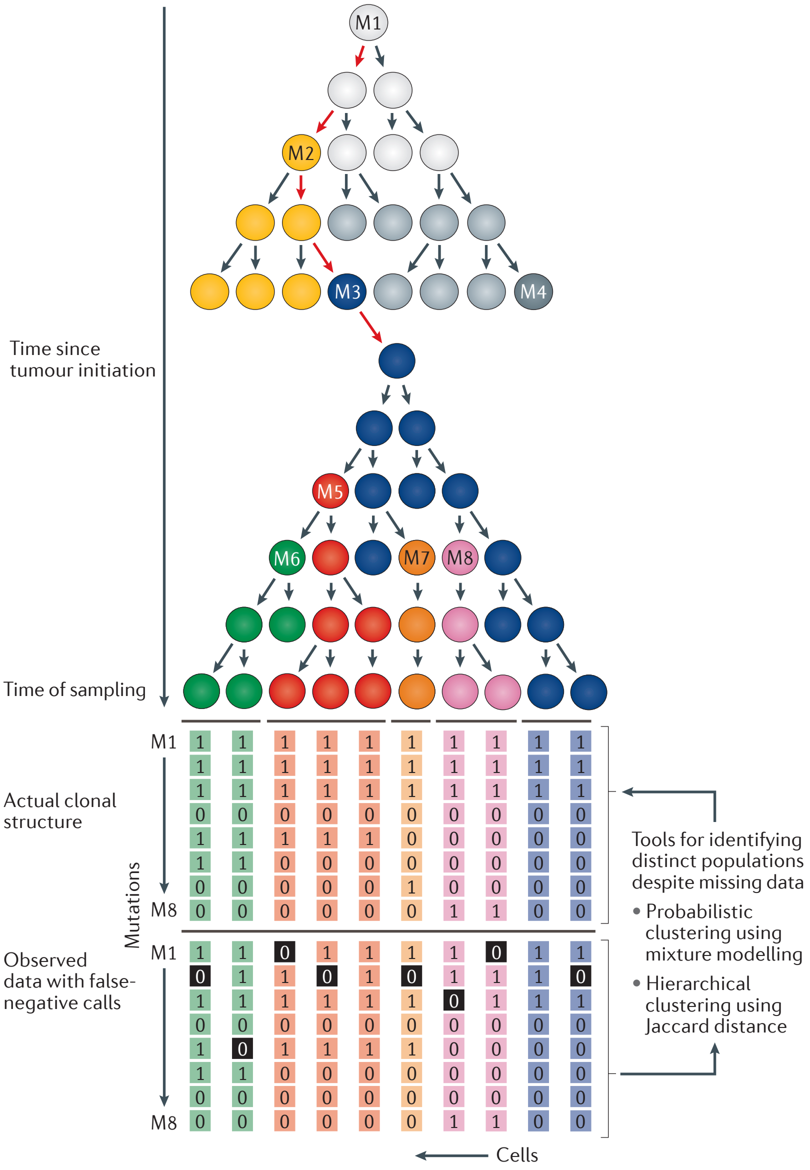

SNV allele
Reference allele
True genotype
Both alleles amplified
100% chance SNV correctly identified
False-negative calls
Single allele drops out
or
50% chance SNV falsely lost
Strategies for overcoming errors
Improvements in amplification uniformity, using probabilistic models to take missing data into account
Entire locus drops out
100% chance SNV falsely lost
Imbalanced amplification
SNV detection rate is proportional to degree of allelic balance and sequencing depth
False-positive calls
Error introduced in genome amplification
False SNV detection is dependent on how early it occurred during amplification and SNV detection thresholds
Require that a mutation is concordant with bulk sample, require that two or more cells have a concordant mutation call

Figure 4 | **Overcoming amplification artefacts when identifying SNVs in single-cell data.** Loss of an allele or entire locus, as well as biases in amplification of each allele at a single location, can result in false-negative single-nucleotide variant (SNV) calls. These errors can be overcome biochemically by improving the performance of the amplification, as well as through computational methods that use many cells to identify the missing data. Errors introduced early in whole-genome amplification (WGA) can cause false-positive SNV discovery, which can be overcome by requiring that the variant call from one cell be identified in additional cells or from a bulk sample from which it was derived.

Applications

Compartmentalizing microbial dark matter. Sequencing has the capacity to overcome the sampling bias that occurs when investigators rely on culturing methods to isolate microorganisms. Sequencing 16S ribosomal RNA has identified as-yet unculturable bacterial phyla and major archaeal groups, although the remainder of the genomes of those putative new phyla are difficult to assemble because the sequencing data are acquired from samples that are composed of multiple species. In principle, single-cell genomics has the potential to assemble the genomes of species that are present at low frequencies in these metagenomic samples[4], as well as to produce assemblies of genomes of completely uncharacterized microorganisms. Here, we focus on the sequencing of species of phyla that had been detected by 16S ribosomal RNA sequencing but had not had full genome assemblies, as these single-cell studies have shown the greatest likelihood of advancing our understanding of microbial ecosystems in the near term.

The first single-cell genome to be sequenced from the environment was a member of the TM7 phylum. In this study, species were identified from the mouths of human subjects, followed by physical isolation and MDA using a microfluidic device[5]. In a later study, cells were sorted using FACS, followed by MDA and sequencing[60]. More recently, species from the OP11 phylum isolated from an


M1
M2
M3
M4
M5
M6
M7
M8
Time since tumour initiation
Time of sampling
Actual clonal structure
Observed data with false-negative calls
Mutations
Tools for identifying distinct populations despite missing data
• Probabilistic clustering using mixture modelling
• Hierarchical clustering using Jaccard distance
Cells

M1	1	1	1	1	1	1	1	1	1	1
	1	1	1	1	1	1	1	1	1	1
	1	1	1	1	1	1	1	1	1	1
	0	0	0	0	0	0	0	0	0	0
	1	1	1	1	1	0	0	0	0	0
	1	1	0	0	0	0	0	0	0	0
	0	0	0	0	0	1	0	0	0	0
M8	0	0	0	0	0	0	1	1	0	0

M1	1	1	0	1	1	1	1	0	1	1
	0	1	1	0	1	0	1	1	1	0
	1	1	1	1	1	1	0	1	1	1
	0	0	0	0	0	0	0	0	0	0
	1	0	1	1	1	1	0	0	0	0
	1	1	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0
M8	0	0	0	0	0	0	1	1	0	0

Figure 5 | **Overview of methods used for determining the clonal structure of cancer samples despite missing data owing to false-negative variant detection.** After initiation, tumours progress through a process of mutation acquisition that results in clonal expansions and extinctions over time. When patients present with clinical disease, their tumours can be sampled to determine the clonal structure at that time using single-cell genome sequencing. However, determination of the true clonal structure is hampered by false-negative variant calls, which can be overcome by either probabilistic modelling or by distance-based clustering methods. Figure from REF. 119, Nature Publishing Group.

anoxic spring[61], SR-1 phylum derived from human oral mucosa[62], TM6 phylum from biofilm on a hospital sink[63] and OP9 phylum from a hot spring[64] have been sequenced using similar methods. The Joint Genome Institute has undertaken a project to sequence the genomes of hundreds of unculturable microorganisms from diverse environments, and has already sequenced the genomes of numerous archaeal and bacterial species of known but unsequenced phyla[65]. This large sequencing study has also identified new biological phenomena in these bacteria, including a new purine synthesis pathway[65].

The most important variable when performing *de novo* genome assemblies is the genome coverage. Almost all studies to date have used MDA. In our comparison study using raw reads from single *E. coli* cells, MDA performed better than MALBAC[29]. Much of the genome coverage for MALBAC was lost owing to contamination when the reaction was carried out in tubes. If only mapped reads are considered, MALBAC would cover a greater proportion of the genome, providing further evidence that reducing contamination using a microfluidic-based MALBAC strategy could potentially provide even better microbial genome assemblies. Tools have recently been developed to systematically assess the quality of single-cell microorganism sequencing data[66], including the presence of a contaminating sequence[67]. Another approach for improving amplification metrics and subsequent assemblies is to capture and culture individual bacteria in droplets, followed by amplification of the hundreds to thousands of cells that descended from the original bacteria[68]. Alternatively, investigators have focused on species with polyploid genomes to acquire better assemblies by starting with bacteria that have 200–900 genome copies per cell[69]. Finally, there have been several short-read assemblers that have been developed to correct for the artefacts of single-cell MDA[49,50,70].

Recent advances in single-cell genomics have enabled the description of completely new phyla, and are now beginning to provide biological insights that could not be made using metagenomics approaches[62]. In addition, a better understanding of the microbiome is creating knowledge that is leading to commercial applications. For example, new members of the Oceanospiralles order, the genomes of which contain enzymes that metabolize crude oil, were identified by single-cell sequencing of ocean samples after the Deepwater Horizon oil spill[71]. There is also promise in using single-cell genomics to identify unculturable human microbial pathogens, as well as to determine differences in pathogenicity between strains of the same species within an individual[72]. In addition, although most studies have focused on bacterial phyla with known 16S rRNA gene sequences, single-cell genomics could be used to assemble the genomes of bacteria or archaea that can be visualized, but the rRNA of which cannot be detected by PCR because of sequence divergence from the universal amplification primers.

An emerging application of single-cell genomics is to use single-cell sequencing to identify new viruses that may be difficult to assemble from metagenomic samples[73]. Several papers have highlighted the power of this approach, including the discovery of five new virus genera through single-cell interrogations of unculturable SUP05 bacteria[74]. Another study found the first viral sequences for 13 new bacterial phyla using public data sets[75]. Computational tools are being developed to improve methods for deciphering new viral sequences from their host[76]. These methods are beginning to be used to study phage–host interactions, which will probably be augmented with single-cell transcriptome sequencing. Another study has looked at virus–protist interactions using single-cell sequencing[77], and it is likely that additional studies will provide details on the relationship between a virus, phage or bacterium and its host by deconvoluting the cell-to-cell variance in that interaction which is partially lost with bulk sequencing strategies.

Still, several challenges remain to increase the throughput and quality of single-cell microbial genomes. More efficient tools for isolating and lysing single microorganisms, uniform and less error-prone amplification methods, and even more robust assembly algorithms that incorporate the additional uncertainty introduced by technical artefacts during single-cell WGA are needed to produce high-quality genome assemblies. The challenge of providing a more uniform approach for producing, analysing and assessing single-cell microorganism genomes is being addressed by the Human Microbiome Project[78], which is in the process of sequencing the genomes of 3,000 single cultured and uncultured bacteria isolated from various human anatomical sites. We have largely focused on single bacterial genomes, but investigators are also interrogating the genomes of other single-cell organisms, including protists[77,79].

Identifying genetic mosaicism in multicellular organisms. The development of cytogenetic methods in the 1950s led to the discovery that cells within the same individual can harbour different numbers of chromosomes[80]. Patients with mosaic expression of dominant Mendelian diseases were subsequently identified by unusual patterns of the stereotypical cutaneous manifestations of several diseases, including neurofibromatosis type I and hereditary haemorrhagic telangiectasia[81]. It was then shown that other diseases such as McCune–Albright Syndrome are only expressed as mosaic diseases, suggesting that germline mutations are lethal[82]. More recently, the development of variant detection methods based on microarrays and next-generation sequencing has enabled the identification of several new diseases that are the result of mosaic SNVs[83–85] or CNVs[86].

However, previous studies of human mosaicism have been limited to the identification of genetic aberrations that are present at relatively high frequencies owing to the low sensitivity of current technologies. Still, a human cell is estimated to acquire an SNV within its coding region after every 300 cell divisions[87]. As the average human body is estimated to contain 37 trillion cells[88], each position in our genomes acquires hundreds to thousands of mutations in different cells as we develop from a zygote into an adult human. In addition, studies that have sampled tissues from different sites of the same person suggest that mosaic CNV and SNV rates are higher than previously appreciated[89,90]. However, the role of that low-level genetic variation in the predisposition and pathogenesis of human diseases remains largely unexplored.

Recent studies have started characterizing mosaic genetic variation in human samples using single-cell sequencing. For example, the *de novo* mutation rate and recombination map were measured in single human sperm[91], followed by a second study on sperm recombination rates[92] and a later study on human oocytes[93]. It has also been shown that a substantial percentage of single human neurons from healthy individuals harbour megabase CNVs[6,94], although these findings have been disputed[95]. More recently, single-cell whole-genome sequencing was used to identify mosaic SNVs whereby the authors found an enrichment in mutations at sites that are actively transcribed in the brain, suggesting those locations are the main source of mutation in those cells[96]. We have also used single-cell sequencing to confirm a mosaic SNV in the sodium channel *SCN5A* as a cause of long-QT syndrome in a neonate (Euan Ashley, James Priest, C.G. and S.R.Q., unpublished observations). It is likely that low-level mosaic genetic variants will be increasingly connected with human diseases as the experimental and analytical tools continue to reduce the technical noise from single-cell WGA, which will contribute to improvements in our ability to decipher the true variants from experimental artefacts. In addition, these tools are likely to find direct clinical applications. Single-cell genomic techniques have long been used to screen embryos for *in vitro* fertilization[97] and more recently they have been used to detect aneuploidy in polar bodies before implantation[93,98,99].

Cancer. The best studied example of genetic mosaicism is cancer. Tumour initiation, maintenance and evolution are mediated by the sequential acquisition of genetic variants in single cells. The aim of the large ongoing cancer sequencing projects is to catalogue those variants to better understand tumour biology[100]. However, like other studies of genetic mosaicism, the sensitivity of detection is limited to variants that are present in about 20% of cells of a bulk sample composed of thousands of cells. The use of variant allele frequency distributions in bulk and regional sequencing studies has indicated that many cancers have considerable genetic heterogeneity[101,102]. However, those methods do not co-segregate mutations into distinct clones, which is required to unambiguously determine the clonal structure of the samples, as well as to determine the evolutionary histories of the malignancies.

Single-cell sequencing studies have now begun to dissect intra-tumour genetic heterogeneity at single-cell resolution. The first published study used DOP-PCR to identify CNVs in breast cancer nuclei[15]. Another group used isothermal amplification methods to identify SNVs in a renal cell tumour, and in a sample from a patient with a myeloproliferative disorder[103,104]. The authors of those two studies concluded that the tumours were monoclonal even though there was significant genetic heterogeneity identified between cells. Another study did identify two distinct clonal populations within a bladder carcinoma[105]. A subsequent single-cell sequencing study of colon cancer claimed that the tumour was biclonal in origin, which seems to be contradicted by the fact that the two putative unrelated clones share mutations[106]. The use of ambiguous descriptions of the clonal structures in these studies highlights the need to create common nomenclature as the field of single-cell cancer genomics matures. These initial studies provided hope that single-cell cancer sequencing would become feasible, but uncertainty of data quality owing to technical limitations prevented the investigators from making new biological insights.

Circulating tumour cells (CTCs) can be isolated and interrogated as a potential window into the genetics of a tumour through non-invasive sampling. The unique technical challenges associated with isolating and analysing the genomes of single CTCs have been detailed elsewhere[9]. Still, these studies have begun to show promise in identifying and characterizing CTCs as alternative diagnostic and disease monitoring strategies[107,108]. One of the fundamental questions that is yet to be resolved is whether CTCs will provide a representative sampling of the genomic diversity within the source tumour.

More recent studies have aimed to improve experimental and computational methods so that examinations of malignancies at single-cell resolution provide a higher-resolution understanding of the disease. Some are limited by evaluating an inadequate number of loci or have insufficient genome coverage to independently determine clonal structures based only on the single-cell sequencing data[41,109,110]. However, by sorting out haematopoietic precursor populations, one study of acute myeloid leukaemia was able to order the acquisition of mutations and provide evidence that specific mutations persisted in populations that have a phenotype similar to normal haematopoietic stem cells[111]. A more recent breast cancer study that used MDA on tetraploid nuclei inferred the clonal structure of the sample using SNVs[7]. The authors also did CNV profiling, although not on the same cells, and found that most CNVs were acquired before SNVs.

We recently used MDA to amplify the genomes of almost 1,500 acute lymphoblastic leukaemia cells[40]. With the large number of cells, we were able to develop methods to determine the clonal structures. In addition, we established general criteria that are required to accurately identify clonal structures, including: having a variant dropout rate of less than 30%, interrogating at least 20 mutations per sample and detecting at least three independent cells to accurately identify a new clone. The vigorous validation of our clonal structures using these analysis methods enabled us to confidently make new conclusions about the events that result in ALL formation, including the presence of co-dominant clones at diagnosis, the acquisition of clone-specific punctuated cytosine mutagenesis, the existence of leukaemia cells at various stages in differentiation arrest and the observation that *KRAS* mutations are acquired late in disease development but are not sufficient for clonal dominance.

Future directions in cancer research

With recent experimental and computational developments, the field of single-cell genomics is poised to begin offering important new insights into cancer development and evolution. Currently, only SNVs or CNVs can be accurately identified from a single cell with targeted or low-pass sequencing; improvements in WGA methods could further decrease sequencing requirements, which would allow more cost-effective whole-genome interrogation of all genomic variation in single cells, including SNVs and structural variants that reside in non-coding regions. The strategies used to interrogate amplified cancer genomes, such as WGS, whole-exome sequencing or targeted sequencing, should be carefully selected based on the hypotheses being tested, as well as the trade-offs between cost, throughput and the quality of the data acquired (BOX 1). Further computational method development is needed to maximize the accuracy of variant calling, as well as the clonal structures identified. Finally, more uniform definitions across cancer sequencing studies are required to allow accurate comparisons between studies. For example, cell lines should not be used to evaluate the quality of methods that are performed on primary cells, and unambiguous terms such as false-negative rate, which incorporate both the locus dropout and ADO, should be substituted for ADO, and a universal definition for a clone should be determined. The latter point is important, as more sensitive single-cell methods are beginning to identify variants that are unique to individual cells or small groups of cells (C.G. and S.R.Q., unpublished observations), and there is no consensus with regard to whether those likely incidental rare mutations should be used to establish those cells as an independent clonal population.

Conclusions

In this Review, we have presented an overview of the current state of the field of single-cell genome sequencing. Substantial progress has been made over recent years in obtaining higher quality single-cell data, which has resulted in the discovery of new biological phenomena that could not be detected with standard bulk genomic interrogations. Still, many challenges remain. Increases in the throughput of cell isolation techniques, as well as improvements in genome amplification, sequencing and computational methods will undoubtedly make the field accessible to many more groups while broadening the types of hypotheses that can be tested. In addition, single-cell genome sequencing has begun to be coupled with RNA[112–114] and/or protein measurements[115] from the same cells. The ability to correlate genotype with other cellular building blocks, as well as phenotypic measurements, will make even more biological questions accessible. Finally, incorporating intracellular[116] and intercellular[117,118] spatial information with genomic measurements will enable researchers to begin putting the cellular building blocks together by providing the surrounding cellular contexts. Many obstacles remain, but we believe the field of single-cell genomics is going to rapidly advance our understanding of microbial ecology, evolution and human disease.

1. Turner, W. The cell theory, past and present. *J. Anat. Physiol.* **24**, 253–287 (1890).
2. Avery, O. T., Macleod, C. M. & McCarty, M. Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from Pneumococcus type III. *J. Exp. Med.* **79**, 137–158 (1944).
3. Amberger, J., Bocchini, C. A., Scott, A. F. & Hamosh, A. McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res.* **37**, D793–D796 (2009).
4. Tringe, S. G. *et al.* Comparative metagenomics of microbial communities. *Science* **308**, 554–557 (2005).

5. Marcy, Y. *et al.* Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proc. Natl Acad. Sci. USA* **104**, 11889–11894 (2007).
This study shows that we can identify uncultivated microorganisms using single-cell sequencing.
6. McConnell, M. J. *et al.* Mosaic copy number variation in human neurons. *Science* **342**, 632–637 (2013).
This article provides the first evidence that mosaic CNV may be more common than previously appreciated.
7. Wang, Y. *et al.* Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* **512**, 155–160 (2014).
The study is an example of high-quality single-cell cancer sequencing data, which has enabled new insights into the pathogenesis of breast cancer.
8. Emmert-Buck, M. R. *et al.* Laser capture microdissection. *Science* **274**, 998–1001 (1996).
9. Navin, N. E. Cancer genomics: one cell at a time. *Genome Biol.* **15**, 452 (2014).
10. Zhou, J., Bruns, M. A. & Tiedje, J. M. DNA recovery from soils of diverse composition. *Appl. Environ. Microbiol.* **62**, 316–322 (1996).
11. Ham, R. G. Clonal growth of mammalian cells in a chemically defined, synthetic medium. *Proc. Natl Acad. Sci. USA* **53**, 288–293 (1965).
12. 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).
13. Gole, J. *et al.* Massively parallel polymerase cloning and genome sequencing of single cells using nanoliter microwells. *Nat. Biotechnol.* **31**, 1126–1132 (2013).
14. Landry, Z. C., Giovanonni, S. J., Quake, S. R. & Blainey, P. C. Optofluidic cell selection from complex microbial communities for single-genome analysis. *Methods Enzymol.* **531**, 61–90 (2013).
15. Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90–94 (2011).
This study provides the first evidence that single-cell sequencing can be used to dissect intratumour heterogeneity.
16. Leung, M. L., Wang, Y., Waters, J. & Navin, N. E. SNES: single nucleus exome sequencing. *Genome Biol.* **16**, 55 (2015).
17. Rinke, C. *et al.* Obtaining genomes from uncultivated environmental microorganisms using FACS-based single-cell genomics. *Nat. Protoc.* **9**, 1038–1048 (2014).
18. White, A. K. *et al.* High-throughput microfluidic single-cell RT-qPCR. *Proc. Natl Acad. Sci. USA* **108**, 13999–14004 (2011).
19. Leung, K. *et al.* A programmable droplet-based microfluidic device applied to multiparameter analysis of single microbes and microbial communities. *Proc. Natl Acad. Sci. USA* **109**, 7665–7670 (2012).
20. Macosko, E. Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
The study presents droplet-based microfluidics as a viable option for efficiently sequencing the transcriptomes of thousands of cells.
21. Blainey, P. C. The future is now: single-cell genomics of bacteria and archaea. *FEMS Microbiol. Rev.* **37**, 407–427 (2013).
22. Shapiro, E., Biezuner, T. & Linnarsson, S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat. Rev. Genet.* **14**, 618–630 (2013).
23. 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).
24. Troutt, A. B., McHeyzer-Williams, M. G., Pulendran, B. & Nossal, G. J. Ligation-anchored PCR: a simple amplification technique with single-sided specificity. *Proc. Natl Acad. Sci. USA* **89**, 9823–9825 (1992).
25. Telenius, H. *et al.* Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer. *Genomics* **13**, 718–725 (1992).
26. Zhang, L. *et al.* Whole genome amplification from a single cell: implications for genetic analysis. *Proc. Natl Acad. Sci. USA* **89**, 5847–5851 (1992).
27. Dean, F. B., Nelson, J. R., Giesler, T. L. & Lasken, R. S. Rapid amplification of plasmid and phage DNA using Phi29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res.* **11**, 1095–1099 (2001).
This paper provides the first evidence that isothermal amplification could be used to efficiently analyse whole genomes.
28. Zhang, D. Y., Brandwein, M., Hsuih, T. & Li, H. B. Ramification amplification: a novel isothermal DNA amplification method. *Mol. Diagn.* **6**, 141–150 (2001).
29. de Bourcy, C. F. *et al.* A quantitative comparison of single-cell whole genome amplification methods. *PLoS ONE* **9**, e105585 (2014).
30. Lasken, R. S. & Stockwell, T. B. Mechanism of chimera formation during the multiple displacement amplification reaction. *BMC Biotechnol.* **7**, 19 (2007).
31. Marcy, Y. *et al.* Nanoliter reactors improve multiple displacement amplification of genomes from single cells. *PLoS Genet.* **3**, 1702–1708 (2007).
32. Zhang, K. *et al.* Sequencing genomes from single cells by polymerase cloning. *Nat. Biotechnol.* **24**, 680–686 (2006).
33. Langmore, J. P. Rubicon Genomics, Inc. *Pharmacogenomics* **3**, 557–560 (2002).
34. Hou, Y. *et al.* Comparison of variations detection between whole-genome amplification methods used in single-cell resequencing. *Gigascience* **4**, 37 (2015).
35. Huang, L., Ma, F., Chapman, A., Lu, S. & Xie, X. S. Single-cell whole-genome amplification and sequencing: methodology and applications. *Annu. Rev. Genomics Hum. Genet.* **16**, 79–102 (2015).
36. Blainey, P. C. & Quake, S. R. Digital MDA for enumeration of total nucleic acid contamination. *Nucleic Acids Res.* **39**, e19 (2011).
37. Yu, Z., Lu, S. & Huang, Y. A microfluidic whole genome amplification device for single cell sequencing. *Anal. Chem.* **86**, 9386–9390 (2014).
38. Nishikawa, Y. *et al.* Monodisperse picoliter droplets for low-bias and contamination-free reactions in single-cell whole genome amplification. *PLoS ONE* **10**, e0138733 (2015).
39. Fu, Y. *et al.* Uniform and accurate single-cell sequencing based on emulsion whole-genome amplification. *Proc. Natl Acad. Sci. USA* **112**, 11923–11928 (2015).
40. 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).
This paper uses microfluidics to efficiently resequence the genomes of almost 1,500 cells, allowing new insights into the development of leukaemia.
41. Hughes, A. E. *et al.* Clonal architecture of secondary acute myeloid leukemia defined by single-cell sequencing. *PLoS Genet.* **10**, e1004462 (2014).
42. Zhang, C. Z. *et al.* Calibrating genomic and allelic coverage bias in single-cell sequencing. *Nat. Commun.* **6**, 6822 (2015).
43. Daley, T. & Smith, A. D. Modeling genome coverage in single-cell sequencing. *Bioinformatics* **30**, 3159–3165 (2014).
44. Clingenpeel, S., Clum, A., Schwientek, P., Rinke, C. & Woyke, T. Reconstructing each cell's genome within complex microbial communities-dream or reality? *Front. Microbiol.* **5**, 771 (2014).
45. Nikolenko, S. I., Korobeynikov, A. I. & Alekseyev, M. A. BayesHammer: Bayesian clustering for error correction in single-cell sequencing. *BMC Genomics* **14**, S7 (2013).
46. Baslan, T. *et al.* Genome-wide copy number analysis of single cells. *Nat. Protoc.* **7**, 1024–1041 (2012).
47. Zhang, C. *et al.* A single cell level based method for copy number variation analysis by low coverage massively parallel sequencing. *PLoS ONE* **8**, e54236 (2013).
48. Cheng, J. *et al.* Single-cell copy number variation detection. *Genome Biol.* **12**, R80 (2011).
49. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
This method overcomes some whole-genome amplification artefacts, resulting in more accurate single-cell genome assemblies.
50. Peng, Y., Leung, H. C., Yiu, S. M. & Chin, F. Y. IDBA-UD: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420–1428 (2012).
51. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* **95**, 14863–14868 (1998).
52. Jaccard, P. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bull. Société Vaudoise Sci. Naturelles* **37**, 547–579 (in French) (1901).
53. Fraley, C. & Raftery, A. E. Model-based clustering, discriminant analysis, and density estimation. *J. Am. Statist. Associ.* **97**, 611–631 (2002).
54. Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statiscal Soc.* **39**, 1–38 (1977).
55. Fraley, C. & Raftery, A. E. How many clusters? Which clustering method? Answers via model-based cluster analysis. *Computer J.* **41**, 578–588 (1998).
56. Fraley, C. & Raftery, A. E. MCLUST: software for model-based cluster analysis. *J. Classif.* **16**, 297–306 (2014).
57. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
58. Kim, K. I. & Simon, R. Using single cell sequencing data to model the evolutionary history of a tumor. *BMC Bioinformatics* **15**, 27 (2014).
59. Yang, Z. & Rannala, B. Molecular phylogenetics: principles and practice. *Nat. Rev. Genet.* **13**, 303–314 (2012).
60. Podar, M. *et al.* Targeted access to the genomes of low-abundance organisms in complex microbial communities. *Appl. Environ. Microbiol.* **73**, 3205–3214 (2007).
61. Youssef, N. H., Blainey, P. C., Quake, S. R. & Elshahed, M. S. Partial genome assembly for a candidate division OP11 single cell from an anoxic spring (Zodletone Spring, Oklahoma). *Appl. Environ. Microbiol.* **77**, 7804–7814 (2011).
62. Campbell, J. H. *et al.* UGA is an additional glycine codon in uncultured SR1 bacteria from the human microbiota. *Proc. Natl Acad. Sci. USA* **110**, 5540–5545 (2013).
63. McLean, J. S. *et al.* Candidate phylum TM6 genome recovered from a hospital sink biofilm provides genomic insights into this uncultivated phylum. *Proc. Natl Acad. Sci. USA* **110**, E2390–E2399 (2013).
64. Dodsworth, J. A. *et al.* Single-cell and metagenomic analyses indicate a fermentative and saccharolytic lifestyle for members of the OP9 lineage. *Nat. Commun.* **4**, 1854 (2013).
65. Rinke, C. *et al.* Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431–437 (2013).
This study identifies new phyla of microorganisms from diverse environments, enabling new insights into the biology of those ecosystems.
66. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
67. Tennessen, K. *et al.* ProDeGe: a computational protocol for fully automated decontamination of genomes. *ISME J.* **10**, 269–272 (2015).
68. Fitzsimons, M. S. *et al.* Nearly finished genomes produced using gel microdroplet culturing reveal substantial intraspecies genomic diversity within the human microbiome. *Genome Res.* **23**, 878–888 (2013).
69. Woyke, T. *et al.* One bacterial cell, one complete genome. *PLoS ONE* **5**, e10314 (2010).
70. Chitsaz, H. *et al.* Efficient *de novo* assembly of single-cell bacterial genomes from short-read data sets. *Nat. Biotechnol.* **29**, 915–921 (2011).
71. Mason, O. U. *et al.* Metagenome, metatranscriptome and single-cell sequencing reveal microbial response to Deepwater Horizon oil spill. *ISME J.* **6**, 1715–1727 (2012).
72. Lasken, R. S. & McLean, J. S. Recent advances in genomic DNA sequencing of microbial species from single cells. *Nat. Rev. Genet.* **15**, 577–584 (2014).
73. Tadmor, A. D., Ottesen, E. A., Leadbetter, J. R. & Phillips, R. Probing individual environmental bacteria for viruses by using microfluidic digital PCR. *Science* **333**, 58–62 (2011).
74. Roux, S. *et al.* Ecology and evolution of viruses infecting uncultivated SUP05 bacteria as revealed by single-cell- and meta-genomics. *eLife* **3**, e03125 (2014).
75. Roux, S., Hallam, S. J., Woyke, T. & Sullivan, M. B. Viral dark matter and virus–host interactions resolved from publicly available microbial genomes. *eLife* **4**, e08490 (2015).
76. Roux, S., Enault, F., Hurwitz, B. L. & Sullivan, M. B. VirSorter: mining viral signal from microbial genomic data. *PeerJ* **3**, e985 (2015).

77. Yoon, H. S. *et al.* Single-cell genomics reveals organismal interactions in uncultivated marine protists. *Science* **332**, 714–717 (2011).
This paper shows that single-cell sequencing can be used to study interactions of bacteria, protists and viruses at single-cell resolution.
78. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
79. Martinez-Garcia, M. *et al.* Unveiling *in situ* interactions between marine protists and bacteria through single cell sequencing. *ISME J.* **6**, 703–707 (2012).
80. Hirschhorn, K., Decker, W. H. & Cooper, H. L. Human intersex with chromosome mosaicism of type XY/XO. Report of a case. *N. Engl. J. Med.* **263**, 1044–1048 (1960).
81. Happle, R. Mosaicism in human skin. Understanding the patterns and mechanisms. *Arch. Dermatol.* **129**, 1460–1470 (1993).
82. Weinstein, L. S. *et al.* Activating mutations of the stimulatory G protein in the McCune–Albright syndrome. *N. Engl. J. Med.* **325**, 1688–1695 (1991).
83. Groesser, L. *et al.* Postzygotic *HRAS* and *KRAS* mutations cause nevus sebaceous and Schimmelpenning syndrome. *Nat. Genet.* **44**, 783–787 (2012).
84. Lindhurst, M. J. *et al.* A mosaic activating mutation in *AKT1* associated with the Proteus syndrome. *N. Engl. J. Med.* **365**, 611–619 (2011).
85. Lindhurst, M. J. *et al.* Mosaic overgrowth with fibroadipose hyperplasia is caused by somatic activating mutations in *PIK3CA*. *Nat. Genet.* **44**, 928–933 (2012).
86. Conlin, L. K. *et al.* Mechanisms of mosaicism, chimerism and uniparental disomy identified by single nucleotide polymorphism array analysis. *Hum. Mol. Genet.* **19**, 1263–1275 (2010).
87. Drake, J. W., Charlesworth, B., Charlesworth, D. & Crow, J. F. Rates of spontaneous mutation. *Genetics* **148**, 1667–1686 (1998).
88. Bianconi, E. *et al.* An estimation of the number of cells in the human body. *Ann. Hum. Biol.* **40**, 463–471 (2013).
89. Behjati, S. *et al.* Genome sequencing of normal cells reveals developmental lineages and mutational processes. *Nature* **513**, 422–425 (2014).
90. Piotrowski, A. *et al.* Somatic mosaicism for copy number variation in differentiated human tissues. *Hum. Mutat.* **29**, 1118–1124 (2008).
91. Wang, J., Fan, H. C., Behr, B. & Quake, S. R. Genome-wide single-cell analysis of recombination activity and *de novo* mutation rates in human sperm. *Cell* **150**, 402–412 (2012).
This study establishes the feasibility of using single-cell sequencing to identify genomic structural variants and SNVs genome-wide.
92. Lu, S. *et al.* Probing meiotic recombination and aneuploidy of single sperm cells by whole-genome sequencing. *Science* **338**, 1627–1630 (2012).
93. Hou, Y. *et al.* Genome analyses of single human oocytes. *Cell* **155**, 1492–1506 (2013).
94. 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).
95. Knouse, K. A., Wu, J., Whittaker, C. A. & Amon, A. Single cell sequencing reveals low levels of aneuploidy across mammalian tissues. *Proc. Natl Acad. Sci. USA* **111**, 13409–13414 (2014).
96. Lodato, M. A. *et al.* Somatic mutation in single human neurons tracks developmental and transcriptional history. *Science* **350**, 94–98 (2015).
97. Handyside, A. H., Kontogianni, E. H., Hardy, K. & Winston, R. M. Pregnancies from biopsied human preimplantation embryos sexed by Y-specific DNA amplification. *Nature* **344**, 768–770 (1990).
98. Geraedts, J. *et al.* Polar body array CGH for prediction of the status of the corresponding oocyte. Part I: clinical results. *Hum. Reprod.* **26**, 3173–3180 (2011).
99. Alfarawati, S., Fragouli, E., Colls, P. & Wells, D. First births after preimplantation genetic diagnosis of structural chromosome abnormalities using comparative genomic hybridization and microarray analysis. *Hum. Reprod.* **26**, 1560–1574 (2011).
100. Cancer Genome Atlas Research Network *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
101. Ding, L. *et al.* Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* **481**, 506–510 (2012).
102. Gerlinger, M. *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* **366**, 883–892 (2012).
103. Hou, Y. *et al.* Single-cell exome sequencing and monoclonal evolution of a *JAK2*-negative myeloproliferative neoplasm. *Cell* **148**, 873–885 (2012).
104. Xu, X. *et al.* Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell* **148**, 886–895 (2012).
105. Li, Y. *et al.* Single-cell sequencing analysis characterizes common and cell-lineage-specific mutations in a muscle-invasive bladder cancer. *Gigascience* **1**, 12 (2012).
106. 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).
107. Ni, X. *et al.* Reproducible copy number variation patterns among single circulating tumor cells of lung cancer patients. *Proc. Natl Acad. Sci. USA* **110**, 21083–21088 (2013).
108. Lohr, J. G. *et al.* Whole-exome sequencing of circulating tumor cells provides a window into metastatic prostate cancer. *Nat. Biotechnol.* **32**, 479–484 (2014).
109. Potter, N. E. *et al.* Single-cell mutational profiling and clonal phylogeny in cancer. *Genome Res.* **23**, 2115–2125 (2013).
110. Papaemmanuil, E. *et al.* RAG-mediated recombination is the predominant driver of oncogenic rearrangement in *ETV6–RUNX1* acute lymphoblastic leukemia. *Nat. Genet.* **46**, 116–125 (2014).
111. Jan, M. *et al.* Clonal evolution of preleukemic hematopoietic stem cells precedes human acute myeloid leukemia. *Sci. Transl Med.* **4**, 149ra118 (2012).
112. Shintaku, H., Nishikii, H., Marshall, L. A., Kotera, H. & Santiago, J. G. On-chip separation and analysis of RNA and DNA from single cells. *Anal. Chem.* **86**, 1953–1957 (2014).
113. Macaulay, I. C. *et al.* G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat. Methods* **12**, 519–522 (2015).
114. Dey, S. S., Kester, L., Spanjaard, B., Bienko, M. & van Oudenaarden, A. Integrated genome and transcriptome sequencing of the same cell. *Nat. Biotechnol.* **33**, 285–289 (2015).
115. Stahlberg, A., Thomsen, C., Ruff, D. & Aman, P. Quantitative PCR analysis of DNA, RNAs, and proteins in the same single cell. *Clin. Chem.* **58**, 1682–1691 (2012).
116. Lee, J. H. *et al.* Highly multiplexed subcellular RNA sequencing *in situ*. *Science* **343**, 1360–1363 (2014).
This study presents a method for acquiring single-cell transcriptomic data while retaining intercellular and intracellular spatial information.
117. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
118. Achim, K. *et al.* High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin. *Nat. Biotechnol.* **33**, 503–509 (2015).
119. Yachida, S. & Iacobuzio-Donahue, C. A. Evolution and dynamics of pancreatic cancer progression. *Oncogene* **32**, 5253–5260 (2013).

Acknowledgements

C.G. is supported by funding from the Burroughs Wellcome Fund, American Lebanese Syrian Associated Charities, Hyundai Foundation for Pediatric Research, American Society of Haematology, and Leukaemia and Lymphoma Society. W.K. is supported by A*STAR (Agency of Science, Technology and Research; Singapore).

Competing interests statement

The authors declare competing interests: see Web version for details.