APPLICATIONS OF NEXT-GENERATION SEQUENCING

Single-cell sequencing-based technologies will revolutionize whole-organism science

Ehud Shapiro[1,2], Tamir Biezuner[1,2] and Sten Linnarsson[3]

Abstract | The unabated progress in next-generation sequencing technologies is fostering a wave of new genomics, epigenomics, transcriptomics and proteomics technologies. These sequencing-based technologies are increasingly being targeted to individual cells, which will allow many new and longstanding questions to be addressed. For example, single-cell genomics will help to uncover cell lineage relationships; single-cell transcriptomics will supplant the coarse notion of marker-based cell types; and single-cell epigenomics and proteomics will allow the functional states of individual cells to be analysed. These technologies will become integrated within a decade or so, enabling high-throughput, multi-dimensional analyses of individual cells that will produce detailed knowledge of the cell lineage trees of higher organisms, including humans. Such studies will have important implications for both basic biological research and medicine.

Next-generation sequencing (NGS). High-throughput DNA sequencing of a large number of DNA molecules in parallel. There is a trade-off between read length and throughput that depends on the sequencing technology, run time and quality.

[1]*Department of Computer Science and Applied Math and* [2]*Department of Biological Chemistry, Weizmann Institute of Science, Rehovot 76100, Israel.*
[3]*Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Scheeles väg 2, 17177 Stockholm, Sweden.*
Correspondence to E.S. and S.L.
e-mails: ehud.shapiro@weizmann.ac.il; sten.linnarsson@ki.se
doi:10.1038/nrg3542
Published online 30 July 2013

DNA sequencing has undergone constant improvement since its inception in the 1970s. Today, next-generation sequencing (NGS) approaches are accelerating in speed and decreasing in cost more quickly than Moore's law[1]. DNA sequencing technologies have improved in precision and throughput, and have enabled the sequencing of entire genomes of species[2,3] and individuals[4]. An increasing number of questions can be addressed by DNA-sequencing-based technologies. In particular, transcriptomic[5], epigenomic[6] and proteomic[7] analyses are being carried out using methods that reduce a specific analysis problem to a DNA-sequencing problem, as explained in FIG. 1.

DNA sequencing technology has not only scaled up rapidly in throughput but — through advances in sample preparation — has also scaled down in terms of the amount of DNA that is required for analysis, to the point at which it is now feasible to analyse the DNA content of individual cells[8,9]. This opens up a wealth of previously impossible applications in both basic research and clinical science. Examples are: the study of microorganisms that cannot be cultured, using direct single-cell genome sequencing[10]; transcriptome analysis of rare, circulating tumour cells[11]; characterization of the earliest differentiation events in human embryogenesis; the investigation of transcriptional noise and stochastic fate choice; and the study of tumour heterogeneity[12] and microevolution[13].

Single cells can be studied and tracked using many detection technologies, including quantitative imaging and mass spectrometry. However, our Review focuses on single-cell analysis using DNA-sequencing-based technologies. Although single-cell sequencing-based analysis has been applied to both unicellular[14] and multicellular organisms[15], this Review focuses on mammalian (primarily mouse and human) single-cell analysis. We first survey current technologies for single-cell isolation, which is essential for DNA-sequencing-based single-cell analysis. We then review technologies for single-cell genomic and transcriptomic analysis, and their applications. We briefly discuss methods for sequencing-based epigenomic and proteomic analyses that have yet to be scaled to single cells. Finally, we describe the impact that the integration of these methods will have on whole-organism science (FIG. 1). We predict an era of integrated single-cell genomic, epigenomic, transcriptomic and proteomic analysis, which we believe will revolutionize whole-organism science by enabling the reconstruction of organismal cell lineage trees for higher organisms, culminating in the reconstruction of an entire human cell lineage tree[16], which will have broad implications for human biology and medicine.

Naturally, in such a diverse, rapidly developing and interdisciplinary field, we cannot possibly cover all of the work that has been carried out over the past few years.

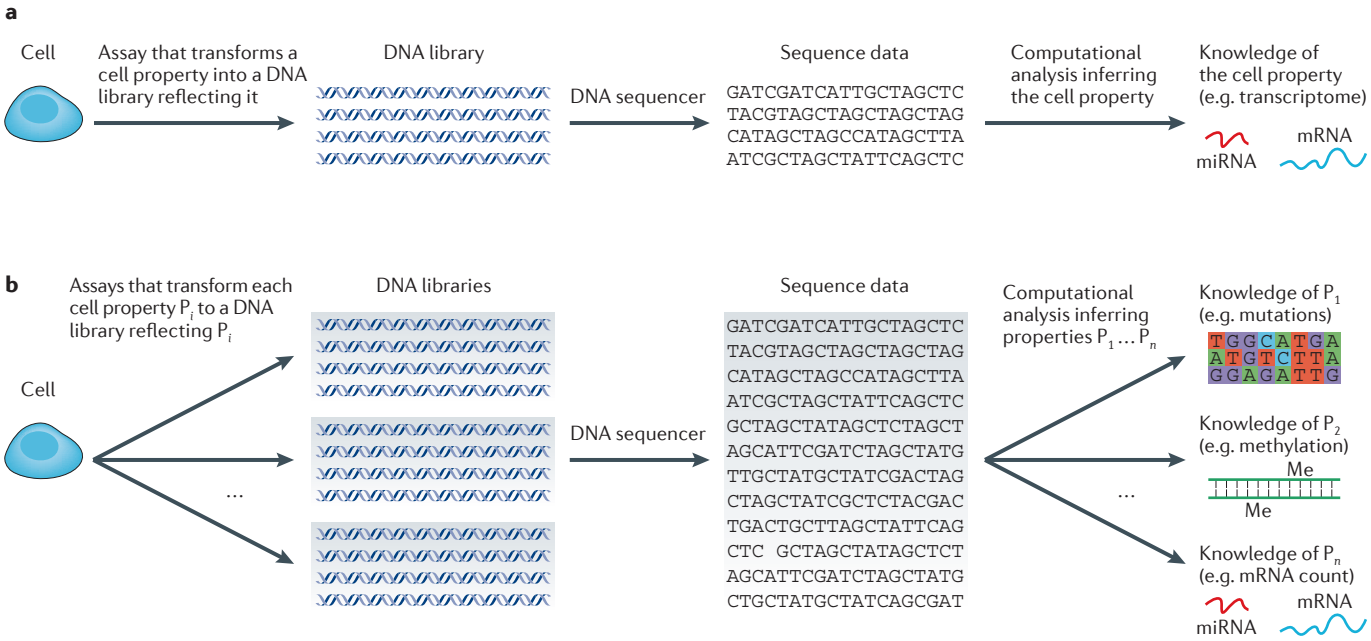


Figure 1 | **Single-cell sequencing-based analysis methods and their anticipated integration.** **a** | Architecture of single-cell DNA-sequencing-based technologies. Current implementations include single-cell genomics (targeted exome or mutational analysis[9,16,56,59], copy-number variation[8,9,58], and recombination analysis in germ cells[27,68]), transcriptomics (transcriptome analysis[11,99,102,104] and recombination analysis in the immune system[136]) and epigenomics[113]. **b** | Architecture of future integrated single-cell DNA-sequencing-based analysis. We expect that within a decade this architecture will allow the simultaneous analysis of multiple properties of an individual cell, including genomics[3,4,38,60–63,76,78–81,83,84,137–139], epigenomics (methylation[6,82,106,107,140], chromatin[108] and conformational[110,111] analysis), transcriptomics (transcriptome analysis[5,141–143], allele-specific gene expression[94] and molecule counting[93–97]) and proteomics[7,127], all of which are currently limited to bulk experiments.

Organismal cell lineage tree
A mathematical entity capturing all cell division and death events in the life of an organism up to a particular time point. The tree consists of labelled nodes, which represent all organismal cells, and directed edges, which represent progeny relationships among them. A reconstructed tree describes lineage relationships among cells sampled from an organism, and is precise only if it is a subtree of the (true) organismal cell lineage tree.

Cell type
A classification of cells by morphology, genotype, phenotype or developmental origin. There is no consensus on which properties are necessary and sufficient for this classification, nor is there general agreement on the actual number of cell types or their proper classification in any higher organism, including in humans.

Fluorescence-activated cell sorting
(FACS). A tool that enables high-speed counting and/or sorting of cells according to features detected by fluorescence.

Also, we expect that by the time this Review is published, additional progress will have been made, which we have been unable to cover. We apologize to the authors whose work we have not discussed.

Methods for single-cell isolation

Tissues are rarely homogenous, and typically consist of tens or hundreds of distinct cell types, which are often intermingled and present at widely different abundances. Single cells can be isolated from such tissues in various ways (TABLE 1), which can be classified as either unbiased (randomized) or biased (targeted) sampling. In principle, an unbiased sample better reflects the composition of the tissue, but a targeted sample may be necessary in order to isolate rare cell types.

There are two key steps in the isolation of single cells from a solid tissue. First, the tissue must be removed from the animal or plant — typically by dissection or biopsy — and dissociated into its constituent individual cells, usually using enzymatic disaggregation. Second, single cells must be placed into individual reaction chambers for lysis and further processing.

Individual cells can be isolated using micromanipulation, for example, using a simple mouth pipette[9,17] or by serial dilution[18,19]. As micromanipulation methods are easy and cheap, they are the most commonly used single-cell isolation methodologies. Their disadvantages are that they are only applicable to cells in suspension, they are low-throughput, and they are susceptible to errors, such as misidentification of a cell under a microscope. These disadvantages are partially addressed by semi-automated devices for cell isolation, with which an expert operator can isolate approximately 50–100 cells per hour[20]. A different approach, which is also classified as micromanipulation, is the optical tweezers technology, which uses a laser beam to capture cells. Although not commonly used, it allows specific cell micromanipulation and measurement[21].

Cell isolation can also be achieved by flow sorting using fluorescence-activated cell sorting (FACS), either using cell-type-specific markers for a biased, targeted sample, and/or using the light-scattering properties of cells to obtain an unbiased sample. The main advantages of FACS-based sorting are the ability to choose between biased and unbiased isolation, high levels of accuracy and high-throughput single-cell isolation[12]. However, FACS requires a large number of cells in suspension as starting material, which might affect the yield with respect to low-abundance cell subpopulations. In addition, the rapid flow in the machine might damage the cells, and care must be taken to ensure the viability of the collected cells if live cells are necessary for downstream protocols.

Table 1 | **Advantages and disadvantages of common single-cell isolation methods**

Method	Unbiased (randomized) or biased (targeted)?	Throughput	Cost	Manual or automatic isolation process?	Refs
Micromanipulation	Unbiased	Low-throughput	Low	Mainly manual	9,17–20
Fluorescence-activated cell sorting	Either biased or unbiased	High-throughput	High	Automatic	12
Laser-capture microdissection	Unbiased	Low-throughput	High	Manual	22–24
Microfluidics	Unbiased	High-throughput	High	Automatic	26–29

Laser-capture microdissection (LCM)[22,23] can be used to cut cells from fixed tissues or cryosections and is effective for collecting nuclei for genomic analyses. The great advantage of LCM is that knowledge of the spatial location of a sampled cell within a tissue is retained, unlike methodologies in which tissue disaggregation is required. There are several current disadvantages of LCM. First, it requires expert manual operation and is a low-throughput technique. Second, in our opinion it is less suitable than other methods for transcriptome analysis, because it is nearly impossible to capture all or most of the cytoplasm of a cell without also collecting material from neighbouring cells. Third, because the section to be dissected has to be of a single-cell width, DNA might be lost by partial nuclei dissection. Finally, selection may be biased owing to the misuse of markers[24]. For these reasons, LCM is less widely used than other methods for single-cell isolation.

Recently introduced microfluidic devices have opened new horizons in single-cell isolation and analysis[12,25]. These devices allow the compartmentalization and controlled management of nanolitre reactions using fabricated microfluidic chips, and they use controlled liquid streaming. The ability to accurately construct low-volume chambers and tubes makes microfluidics ideal for single-cell isolation, as well as for further downstream processes. Microfluidic devices provide inherent advantages by allowing higher throughput with less effort, reducing reagent cost and improving accuracy. In recent years several implementations of microfluidic devices have been presented for single-chromosome isolation[26] and single-cell isolation followed by analysis[27,28]. We expect microfluidic technologies and products to continue their advance and ultimately to provide a robust foundation for single-cell sequencing-based analysis[29].

Single-cell genomics

Reconstructing cell lineage trees using somatic mutations. Different cells from the same individual were initially thought to harbour identical genomes. This turns out to be false, not only for the immune system[30] and cancer cells[31] (which both undergo somatic evolution) and for germline cells that undergo recombination[27], but for all cells in our bodies. During normal mitotic cell division DNA is replicated with very high, but not absolute, precision, which leads to the incorporation of somatic mutations. These somatic mutations, accumulated since the zygotic stage, endow each cell in our bodies with a genomic signature that is unique with a very high probability[16]. As the differences in cellular genomic signatures are mostly without phenotypic effect, what would science gain by knowing them?

The answer is that knowing the unique genomic signatures of our body cells allows the reconstruction of cell lineage trees with very high precision[16]. Central unresolved problems in human biology and medicine are in fact questions about the human cell lineage tree: its structure, dynamics and variability during development, growth, renewal, ageing and disease. For example: does the oocyte pool renew during adulthood[32]? Do β-cells renew[33]? Do neural progenitor cells produce each brain cell type as needed, or do specialized progenitors each produce a single cell type[34,35]? Information about the cell lineage trees of higher organisms consists largely of data from cell fate maps[36,37], which are mostly derived from clonal-marking experiments that are not applicable to humans. Complete knowledge of the unique somatic mutations that are accumulated in each cell would allow the reconstruction of cell lineage trees with extremely high precision[16,38].

Work in this direction has focused on identifying somatic mutations in microsatellites[39] that are hypermutable in normal cells and even more so in microsatellite-instable (MSI) cells[19,40,41] and in mismatch repair (MMR)-deficient organisms[16,42,43]. Knowing only a small proportion of such mutations allowed fairly precise lineage reconstruction using standard phylogenetic algorithms, depending on cell depth[40,44,45]. By applying this approach to samples of cells from tissues of interest, key aspects of the underlying cell state dynamics were characterized. The cell lineage trees thus obtained provided information about the substructure of the population, such as the existence of small populations of stem cells. Such information has applications for developmental biology (for example, oocyte maturation, colon crypt development[18] and muscle stem cell lineages[46]) and for leukaemia[19].

Somatic mutations can be used for cell lineage reconstruction only if: the mutations do not confer a selective advantage or disadvantage, they are associated with DNA replication (rather than elapsed time, for example) and/or their dynamics is well understood and can be modelled. The accuracy of lineage reconstruction increases with the fraction of the genome analysed per cell, and there is a trade-off between accuracy and

Laser-capture microdissection
(LCM). A method that combines high-resolution microscopy and the accurate isolation of user-defined regions of a tissue slice for downstream analysis. Typically, a powerful laser is used to cut an outline of the target region, which can then be ejected into a sample tube.

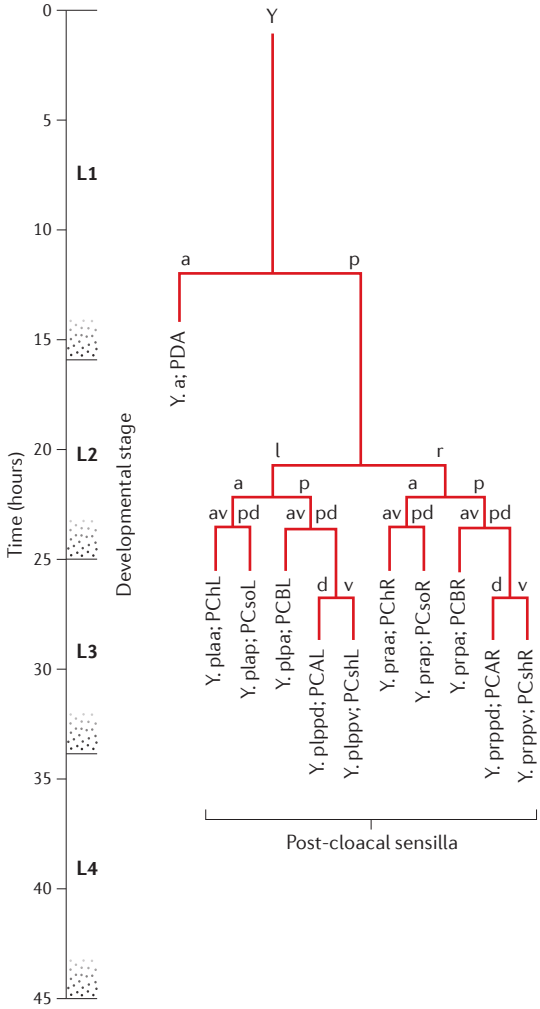
Microsatellites
Repetitive elements in the genome that consist of basic units 1–6 bp long that are repeated from a few to a few dozen times. Microsatellites occupy 3% of the human genome.

Cell depth
The number of divisions a cell underwent since the zygote.

Box 1 | **Whole-organism science: the *Caenorhabditis elegans* benchmark and the principle of biological uncertainty**

Caenorhabditis elegans is the best-studied multicellular organism and hence offers a benchmark for the systematic and integrative study of organismal biology at the molecular, cellular, organ and organismal levels, which we refer to as 'whole-organism science'. The scaffold on which the comprehensive knowledge of *C. elegans* biology is structured is its cell lineage tree[131,132], a fragment of which is shown in the figure. The structure of the tree shows the lineage relationships among all the organism's cells, past and present; the labels give the identities and types of organismal cells, and the length of tree edges represent the timing of cell division and death events. Additional knowledge that is not shown in the tree is the location within the organism of each cell. New knowledge is constantly being added to this scaffold (for example, the transcriptome of each organismal cell[99]).

Whereas the development of *C. elegans* is deemed to be deterministic, higher organisms exhibit great variability during development, renewal, ageing and disease, which is caused by genetic and environmental differences. Owing to this variability, we postulate that whole-organism science of higher organisms must deal with a 'biological uncertainty principle'. Heisenberg's uncertainty principle states that it is impossible to measure accurately and simultaneously the position and momentum of an elementary particle. Similarly, in general it is not possible to measure accurately and simultaneously the 'cellular position' (for example, the genomic, transcriptomic, epigenomic and proteomic state of a cell) and the 'cellular momentum' (for example, the next differentiation, division or degradation event of a cell) for individual cells in an organism. In order to know accurately the state of a cell, one must destroy it and analyse its content, thereby eliminating cellular momentum. Alternatively, to observe cellular momentum, one cannot interfere with the behaviour of the cell and hence must compromise on precisely knowing the state of the cell. For example, using fluorescent reporters that are minimally invasive, limited information can be obtained on both the state and momentum of single cells[133–135]. The use of external markers may still have an effect on the cell, and currently their use is limited to measuring a small number of parameters and is not applicable to humans. In non-deterministic higher organisms the structure of the cell lineage tree may be affected by nature (the genome), or nurture (the environment), as well as stochastic events such as cancerous mutations. Integrated single-cell analysis providing knowledge of the genome, epigenome, transcriptome and proteome of each sampled cell can be used to reconstruct detailed lineage trees of the sampled cells and to infer the functional state of ancestor cells. Such inferred trees can be used to predict the next differentiation or division decision of a cell on the basis of its functional state, thus overcoming (to some extent) the limitations imposed by the biological uncertainty principle. The figure is reproduced, with permission, from http://www.wormatlas.org/images/BYUFlineages.jpg © (2013) WormAtlas.



cost per cell. Given a fixed fraction of the genome to be analysed, the accuracy of its sequencing is crucial. One way to increase sequencing accuracy (up to a point) is by increasing the sequencing depth. Sequencing accuracy is decreased by the bias and infidelity introduced by the biochemical steps of preparing cellular DNA for sequencing, including whole-genome amplification (elaborated on below), library preparation and the sequencing process itself[47]. Trade-offs between cost and accuracy require fine-tuning these parameters (for example, carrying out additional sequencing runs using fewer cells per run or increasing the number of analysed loci).

A disadvantage of cell lineage reconstruction using somatic mutations is that it cannot provide, by itself, information on the state of inferred ancestor cells. It can show the depth of sampled cells and the lineage relationships among them, but not the type of the ancestor cells that are represented by internal nodes in the reconstructed cell lineage tree. In particular, the results of 'time-lapse' experiments — in which different tissue samples of the same or different organisms are analysed at different organism ages — cannot be superimposed on the same cell lineage tree, as has been done for *Caenorhabditis elegans* (BOX 1), and labels of internal nodes of a cell lineage tree can only be approximated from the properties of the sampled cells, namely the leaves of the tree. Labelling the leaves and internal nodes of the reconstructed cell lineage tree with cell type and state information requires further single-cell epigenomic and transcriptomic analyses, as explained below.

Cell lineage reconstruction of cancer will elucidate its development. Cancer patients typically do not die from the effects of the primary tumour but from those of its metastases. Yet, despite decades of research, the key question of where metastases originate from has not been fully answered[48] (FIG. 2). For example, can metastases originate from any tumour cell or only from a distinct tumour subclone (for example, circulating tumour cells[49])? In the latter case, are these subclones created early or late in the development of the tumour[23,50]? Alternatively, perhaps metastases and the primary tumour are both independent descendants of cancer stem cells[51,52]. Or maybe metastases are formed

Sequencing depth
The total amount of raw sequence mapped to a reference genome, divided by the length of the genome.

Whole-genome amplification
(WGA). Refers to methods that are used to amplify the genomic DNA of single cells to increase the number of copies of DNA for downstream processing.

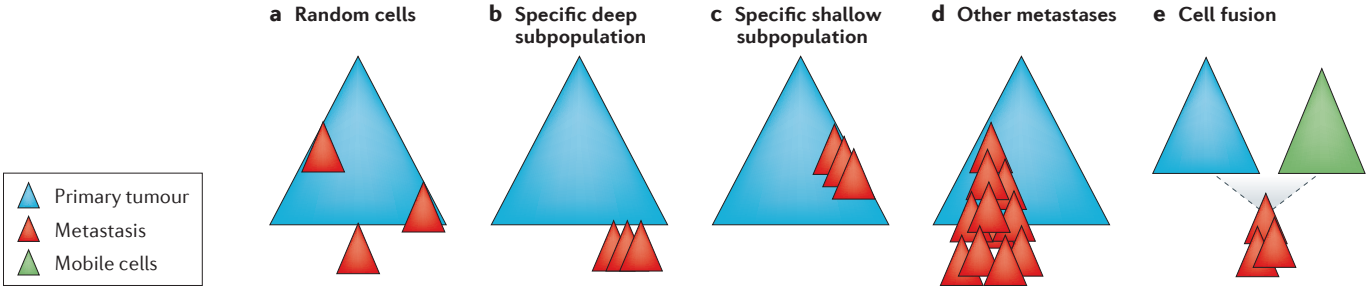


Figure 2 | **Alternative hypotheses on the origin of metastases.** **a** | Metastases originate from random cells during tumour development. **b** | Following tumour growth, metastases originate from a specific tumour subpopulation, which underwent many divisions (that is, a 'deep' subpopulation). **c** | At the initial tumour growth stages, metastases originate from a specific tumour subpopulation (that is, a 'shallow' subpopulation). **d** | Metastases originate from other metastases. **e** | Tumour cells engage metastasis by fusion with other cells, which endow a mobility property.

through the fusion of primary tumour cells and normal mobile cells such as macrophages[53,54]. As another example, consider the origin of cancer relapse after chemotherapy. Is this caused by ordinary tumour cells escaping chemotherapy stochastically, or by a distinct subpopulation of infrequently dividing cancer-initiating cells that escape chemotherapy owing to their slow division rate? The answers to these questions are encoded in the patient's cancer cell lineage tree[19,55]. Understanding the emergence and distribution of driver mutations in the context of the cancer cell lineage tree is also of prime importance[56].

Early experiments analysed a few key markers in each individual cell. In one recent example, heterogeneity and tumour origin in acute lymphoblastic leukaemia were studied by assaying the occurrence of up to eight chromosomal aberrations and their combinations in single cells using fluorescence *in situ* hybridization (FISH). This allowed an analysis of subclonal architecture during cancer progression[57]. More recently, sequencing of hundreds of single nuclei was used to generate approximated copy-number profiles for individual breast cancer cells[8,58], thus allowing the reconstruction of tumour population structure and evolutionary history. In another study[59], whole-exome single-cell sequencing in a patient with myeloproliferative neoplasm was carried out to reconstruct tumour ancestries and to identify candidate driver mutations. In a final example, single-cell DNA templates were extracted following clonal expansion and were sequenced[60] (a method that is discussed below) to study the lineage of normal cells and to determine the earliest precancerous mutations that ultimately led to the development of the tumour.

Bulk sequence analysis methods are practical and efficient predecessors to single-cell studies. They can be used to extract efficiently distributions of markers of interest (for example, somatic mutations) from a large number of cells. Recent bulk studies have used a two-tier design of low-depth, whole-genome sequencing combined with deep sequencing of loci underlying putative driver carcinogenic events. The approach can quantify the frequencies of genetic and epigenetic variants *in vitro*[38] or *in vivo*; for example, it can be used to estimate tumour cellular dynamics such as mutation penetrance, to construct models that distinguish driver and passenger mutational events and to reconstruct tumour ancestries using coalescent models[23,61–63].

Lineage tracing using NGS of somatic mutations has been demonstrated *in vivo* for bulk cell populations but not for single cells. Bulk methods do not provide accurate information on how different combinations of mutations or aberrations emerge, and precise answers to such questions await single-cell lineage analysis of cancer.

The road to single-cell genomics. Although sequencing the DNA of a cell population is now straightforward[61], sequencing DNA from single cells is still a challenge. Historically, the cost of sequencing multiple individual cells at adequate depth for genetic profiling was prohibitively high, and despite the remarkable recent increase in throughput, the sequencing cost is still a hurdle to large-scale single-cell genomics, transcriptomics and epigenomics. The current prevailing DNA sequencing approach combines WGA with the preparation of amplified, nanogram-sized DNA libraries. WGA can be achieved through multiple variants of PCR-based amplification[8,58,64–66] or isothermal amplification using multiple-displacement amplification[16,27,56,59,67]. Demand for unbiased single-cell DNA amplification has inspired the development of new techniques for WGA. These include the multiple-annealing and looping-based amplification cycles (MALBAC) method[9,68] and single-cell-specific WGA kits such as the single-cell RepliG kit, by Qiagen. Performance of the available methods varies between applications, and a comprehensive side-by-side comparison of the different methods is still much needed[14,69]. In this article we do not provide a comprehensive summary of all the available techniques. For a recent review summarizing WGA methods see REF. 14.

A high-fidelity, low-bias method for genome amplification is especially crucial for single-cell DNA analysis because the initial copy number is one, unlike for DNA sequencing from bulk cell populations or even for single-cell RNA analysis. Low-fidelity amplification can produce non-representative and biased sequencing results, which in turn may lead to incorrect single-nucleotide polymorphism calls (SNP calls), uneven sequencing coverage and

Clonal expansion
A method to retrieve representative DNA from a single cell following its proliferation. A single cell is isolated, cultured *ex vivo*, and allowed to divide several times. DNA is isolated from the bulk cell population using standard DNA extraction techniques that do not involve amplification.

Single-nucleotide polymorphism calls
(SNP calls). Following sequencing read assembly, this is the identification of single nucleotides that are different from the nucleotide at the same position in a specific reference genome. This process requires high-quality sequencing and adequate sequencing depth for statistical significance.

Sequencing coverage
In a sequencing experiment, the number of reads covering a specific nucleotide position is the coverage of that position. Increasing read depth leads to increasing coverage, and to increasing accuracy of the base calls.

Box 2 | **How many individual cells are needed for quantitative transcriptomic analysis?**

In a standard bulk RNA sequencing (RNA-seq) experiment, precision is limited only by sequencing depth. Typically, ten million reads are generated, and a threshold of 50 reads per kb per million reads (RPKM) is considered adequate to call a gene as expressed. For a gene that is 1 kb long, this corresponds to 500 reads, thus leading to a minimum coefficient of variation (CV; which is equal to the standard deviation divided by the mean) of 4%, as given by the Poisson distribution. In a fairly typical single mammalian cell containing 200,000 mRNA molecules, 50 RPKM corresponds to about ten mRNA molecules. Again, assuming a Poisson distribution across cells, the expected CV is 32%, but this can be reduced by pooling data from many cells. How many cells are needed to reduce this error to that of the bulk experiment? The answer is 50, because the pooled data from 50 cells will contain 500 mRNA molecules. These are ideal numbers, and in practice more cells will be required. For example, if the efficiency of converting mRNA to cDNA is only 10% (which is not an unrealistic assumption), then tenfold more cells will be required. Similarly, when additional noise is introduced (for example, by PCR amplification) the number of cells required will increase correspondingly. Furthermore, if the sample is heterogeneous, then enough cells must be analysed so that all representative cell types are observed. Finally, all these estimates assume that the single-cell measurements are accurate, as systematic inaccuracies (for example, due to amplification bias) will not be cured by collecting more cells.

Although necessarily simplistic, these 'back-of-the-envelope' calculations suggest that hundreds or thousands of single cells will need to be analysed to answer targeted questions in single tissues. For a whole-organism view, at least millions of cells will need to be analysed (that is, thousands of cells in thousands of tissues and time points), which is a feat that will require miniaturization, automation and further reductions in the cost of DNA sequencing.

missing loci (termed locus-dropout or allele-dropout). Such biases have less effect when sequencing bulk cell populations or even WGA products from a few hundred cells[14].

An alternative single-cell DNA extraction technique uses clonal expansion[60,70]. However, this method suffers from several drawbacks. First, the efficiency of the proliferation of a single cell *ex vivo* is dependent on the cell type, cell stress that occurs post-isolation and the availability of suitable conditioned growth media[71]. Second, cell death or decomposition prevents culturing[8]. Third, this approach is contamination-prone, laborious and more time consuming than single-cell WGA procedures, especially when a large number of cells need to be cultured and their DNA extracted independently. Finally, mutations are introduced during this procedure, especially in MMR-deficient cells. When single-cell WGA techniques mature, it will be valuable to compare them comprehensively with clonal expansion for reproducibility and for artefacts caused by polymerase bias. Nevertheless, as previously explained, subsequent biochemical steps following clonal expansion can also cause artefacts, even more than the single-cell WGA itself[32].

Single-molecule sequencing methods (often referred to as third-generation sequencing technologies)[72–75] eliminate the amplification step before sequencing. As such, they eliminate amplification bias and hold great promise for single-cell sequencing. Yet, these technologies currently suffer from high error rates, low-throughput and low sequencing efficiency, owing to slow and non-robust detection[74,75].

For some applications, analysing the entire genome (or transcriptome) is not essential, and targeting a genomic subset using genomic enrichment methods may allow higher sensitivity and lower per-sample cost. For example, genomic subsets can include exomes[76,77] or specific mutations in genes of interest[78–81]. High-throughput sequencing has the combined advantages of both high-throughput analysis and sample multiplexing using DNA barcodes in a single sequencing run, which makes it ideal for large-scale analysis of multiple single cells (BOX 2; TABLE 2). Genomic enrichment was initially approached through PCR amplification of a few to a hundred amplicons, and single-cell isolation followed by PCR is a current practical alternative to whole-genome sequencing due to technological advancements in this field[80,81]. Specifically, the ability to cost-effectively synthesize thousands of custom-designed oligonucleotides enabled the development of more-powerful genome enrichment techniques based on the hybridization of target material to oligonucleotide probes and subsequent processing (namely selective circularization methods[82,83] and hybridization-based capture methods[56,59,62,84]). These methods allow cost-efficient targeted DNA enrichment and high-throughput NGS library preparation. Further development of the sensitivity and throughput of these techniques will probably make these methods more common in single-cell genomics, as an interim step to whole-genome sequencing or as a long-term companion to such a capability.

Single-cell transcriptomics

The molecular state of cell populations. Given a heterogeneous cell population, measurement of the mean values of key factors, such as the genotype, RNA output or epigenetic state of a locus of interest, provide only a partial characterization of the state of the system. Unfortunately, most of the methods that are used for quantifying the molecular state of a cell population, from transcriptional profiling to proteomics, are based on estimating mean behaviours in ensembles of millions of cells by averaging the signal of individual cells. For example, it is impossible to determine the cell-to-cell variation of gene expression based on microarray or RNA sequencing (RNA-seq) data, or to determine whether intermediate levels of a signalling protein are a consequence of a bimodal or uniform intrapopulation distribution based on standard proteomics. Going beyond mean-based characterization of

Amplicons
DNA products of PCR amplifications.

Table 2 | **Current trade-offs in sampling heterogeneous cell populations**

	Experimental approach					
	Bulk average	**Tagged libraries**	**Multi-dimensional cell sorting**	**Deep sequencing of bulk samples**	**Small samples of single cells**	**Large samples of single cells**
Number of cells	Millions	Hundreds per marker	Millions	Millions	Tens to hundreds	Thousands to tens of thousands
Molecular markers	Any	RNA or tagged proteins	Surface markers or signalling molecules	Genetics or DNA methylation	RNA, genetics or DNA methylation	RNA, genetics or DNA methylation
Typical costs	Low	High setup cost but subsequently low	High	Low to medium	Medium, depending on the sequencing component	High, depending on the sequencing component, but low per-cell cost if samples are multiplexed
Mean?	Global	For markers (thousands)	For markers (<50)	Yes	Yes	Yes
Variance?	No	For markers (thousands)	For markers (<50)	No	Yes (of limited accuracy)	Yes
Pairwise covariance?	No	No	For profiled markers (<50)	Only linked marks	Yes (of limited accuracy)	Yes
Complex correlations and/or causal networks?	No	No	Among profiled markers (<50)	No	No	Possibly
Subpopulation structure?	No	No	Excellent, but only if markers are appropriate	model-based (for example, carcinogenesis)	Good, but only for subpopulations with significant (>10%) frequency	Excellent
Cell lineage tree?	No (averaged to most-recent common ancestor (MRCA))	No (averaged to MRCA)	No (averaged to MRCA)	No (averaged to MRCA)	Yes	Yes

cell populations requires balancing the number of sampled cells and the completeness of functional coverage on varying scales (TABLE 2).

Applications of single-cell transcriptomics. One major application for single-cell transcriptomics is in the analysis of rare cell types. For example, circulating tumour cells can be obtained from patient blood, but typically only a few cells are isolated per blood sample and these will often be contaminated by a larger number of normal cells. Single-cell RNA-seq could be used to differentiate between these cell types and simultaneously to obtain expression data from the tumour. Similarly, the early human embryo by definition contains only rare cell types, which exist only transiently. Key questions about early development could be addressed using transcriptomics. In this context, transcriptomics has the advantage of being able to use sequence polymorphisms (for example, SNPs) to distinguish transcripts that are derived from each of the two parental genomes. Another area that will benefit immensely from single-cell transcriptomics is the study of adult stem cells, which are often rare, sometimes exist only transiently, and can be intermingled with other cell types. However, by using single-cell RNA-seq, each cell type can be extensively sampled simply by taking unbiased samples of cells from the tissue.

Individual cells differ greatly in their size, morphology, developmental origin and functional properties. Yet, our current level of understanding of cell types, their origin, evolution and diversity is embarrassingly poor, despite progress in some specific cases[85,86]. Furthermore, there is no general agreement on the number of cell types in a mammalian body. In fact, there is no agreement on what defines a cell type, and finding such a definition must surely be one of the most important goals as we embark on large-scale single-cell transcriptome analysis. As a starting point, we suggest that cell types can be provisionally identified as cells for which global transcriptional states are similar. Just how similar, and just which parts of the transcriptome are relevant, will be crucial questions for the future. But this provisional concept of cell type leads immediately to an unbiased method of cell-type discovery (FIG. 3): collect a large, unbiased sample of cells from the tissue of interest, generate transcriptomes for each cell and use computational methods to find sets of similar cells. Established clustering and dimension-reduction methods — such as K-means, affinity propagation and hierarchical clustering, and principal component analysis — will be useful starting points[87]. Because some laboratories are already analysing hundreds or thousands of single-cell transcriptomes, we anticipate that the time will soon be ripe to embark on large-scale, whole-body cell-type discovery and characterization.

A further area of application for single-cell transcriptomics is the characterization of transcriptional

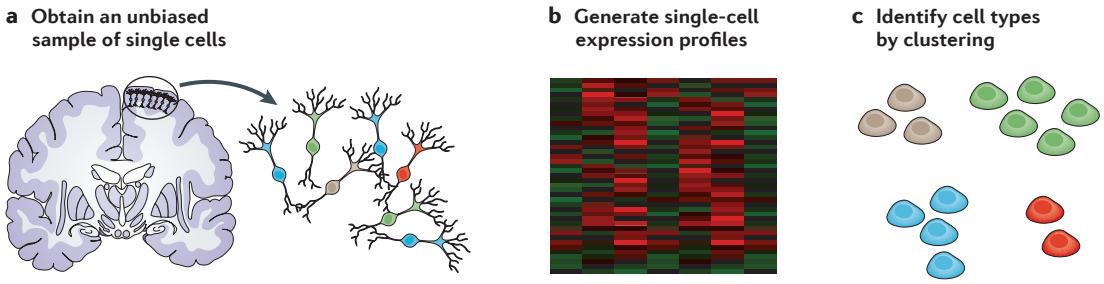


Figure 3 | **Cell-type discovery by unbiased sampling and transcriptome profiling of single cells.** **a** | A sample of cells is taken from the tissue of interest, with the aim of obtaining a representative sample of the types of cells that are present in the tissue. **b** | Each cell is profiled using single-cell RNA sequencing (RNA-seq). **c** | Subsequently, the resulting expression profiles are clustered. The result is a map of 'cell space', in which similar cells are grouped close to each other. The strategy is shown here in cartoon form, but in practice it will be necessary to collect and analyse thousands of cells in each tissue (that is, millions of cells overall) to make a comprehensive cell space map of a whole organism.

fluctuations. Dynamic changes in RNA content are associated with cyclic processes, such as the cell cycle in dividing cells and circadian rhythms. Other fluctuations are stochastic and reflect the fact that transcription is a discrete process composed of many probabilistic steps. Further heterogeneity is introduced by uneven partitioning of the cellular content at cell division (for example, REF. 88). Direct transcriptome analysis of large numbers of single cells should open up the study of oscillatory and stochastic regulatory processes in unperturbed cell populations. In a population of putatively identical cells, sets of co-regulated genes can be identified. Each set must be part of a functional process, such as an oscillator or a stochastic process. For example, genes that share a common upstream regulator would presumably show correlated expression. At present, the number of single cells that must be analysed in order to discover covariant genes is unknown, and finding first estimates of these numbers will be a key task in the near future. There is also evidence that transcription is subjected to strong intrinsic fluctuations[89,90]. Models to explain this intrinsic noise lead to predictions about the shape of the mRNA copy-number distribution, which can be tested against experimentally measured distributions[89]. Such tests cannot be carried out using bulk measurements, which do not give any information about the variance or any higher moments. Nonetheless, single-cell transcriptome analysis provides only a snapshot in time, and it will remain important to complement this view with dynamic, long-term measurements by, for example, time-lapse microscopy[91].

The road to single-cell transcriptomics. Despite advances in single-molecule DNA[72–74] and RNA[92] sequencing, it is not yet possible to sequence RNA directly from single cells. Currently, RNA needs to be converted to cDNA and amplified, and this must be achieved with minimal losses and without introducing too much quantitative bias.

There are several sources of noise in a single-cell transcriptome experiment. There are biological fluctuations, both global (that is, affecting the total amount of RNA in the cell) and local (for example due to co-regulation or large-scale chromatin modifications). There is also technical noise, for example due to pipetting errors, temperature differences, differences in sequencing depth, PCR amplification bias and differences in reverse transcription efficiency. It is important to realize that single-cell transcriptome analysis is also a single-molecule analysis, because many genes are expressed at only a few mRNA molecules per cell. Amplification from small numbers of molecules is subject to the Monte Carlo effect, in which stochastic events in the first few cycles of PCR are amplified exponentially, causing large quantitative errors.

The ultimate goal of quantitative single-cell transcriptome analysis must be to count every RNA molecule in the cell exactly, resulting in near-zero technical error. This is required, for example, if we are to use the shape of mRNA count distributions to infer the kinetics of transcription. Accurate molecule counting is in fact possible by using unique labels for molecules[93–97]. After amplification and deep sequencing, each original molecule can be identified. As long as the sample is sequenced deeply enough, so that each molecular label is observed at least once, differences in amplification efficiency do not matter. Although the use of unique molecular labels has until now been used only for bulk samples, it is a key advance that will probably enable a more quantitative analysis of single-cell transcriptomes.

Another source of error is losses, which can be severe. The detection limit of published protocols is 5–10 molecules of mRNA. If, as seems likely, the limit of detection is primarily determined by losses during sample preparation, this would indicate that 80–90% of mRNA was lost. Or, to put it the other way around, a 90% loss leads to an approximately 50% chance of failing to detect a gene that is expressed at a level of seven mRNA molecules (from the binomial distribution). These losses are especially problematic in small cells, such as stem cells, in which the mRNA content is low to begin with. But even in larger cells, such losses introduce a severe quantitative error owing to the stochastic sampling of small numbers of molecules. For example, measuring 100 molecules with a 90% loss leads to 10 ± 3 detected

Higher moments
Measures of the shape of a statistical distribution beyond mean and variance, such as skewness and kurtosis.

Table 3 | **Recently published single-cell RNA-seq methods**

Method	Principle	Strand-specific?	Positional bias?	Early multiplexing?	Ref
Tang *et al.*	Homopolymer tailing	No	Weakly 3′-biased	No	102
STRT	Template switching	Yes	5′ (TSS)	Yes	104
SMART–seq	Template switching	No	Nearly full-length	No	11
CEL–seq	*In vitro* transcription	Yes	Strongly 3′-biased	Yes	99
Quartz–seq	Homopolymer tailing	No	Weakly 3′-biased	No	144

CEL–seq, cell expression by linear amplification and sequencing; SMART–seq, switching mechanism at the 5′ end of the RNA template sequencing; STRT, single-cell tagged reverse transcription; TSS, transcription start site.

molecules, which means that the loss alone has introduced a 30% standard deviation. To mitigate the impact of technical noise, we suggest analysing large numbers of single cells (BOX 2).

The earliest single-cell transcriptomes were generated by *in vitro* transcription (IVT)[98], and recently IVT was used to produce libraries for Illumina sequencing, in a method called cell expression by linear amplification and sequencing (CEL–seq)[99]. The main advantage of IVT is the linear amplification, which should in theory be less biased than exponential amplification methods such as PCR. A disadvantage is that the resulting library is biased towards the 3′ end of genes, and this bias can be difficult to control. By contrast, PCR-based protocols are capable of amplifying full-length cDNA.

A second approach is to add a homopolymer tail to the first-strand cDNA, which allows the cDNA strand to be amplified by PCR. An early example used deoxyguanosine-tailing followed by PCR[100]. Subsequently, this protocol was optimized[101] and adapted for sequencing[102]. Like IVT, homopolymer tailing is biased towards the 3′ end.

A third approach uses 'template switching'. Common reverse transcriptases of the Moloney murine leukaemia virus family tend to add a short tail of (preferentially) cytosines to the end of the first-strand cDNA. If a helper oligonucleotide, carrying a short GGG motif, is included in the reaction, it will anneal to the cytosine motif and the reverse transcriptase will switch template and copy the helper oligonucleotide sequence[103]. The result is that an arbitrary sequence can be introduced at the 5′ end (by tailing the reverse transcription primer) and at the 3′ end (by template switching) of the cDNA, thus allowing subsequent amplification by PCR. Additionally, template switching has a preference for 5′-capped RNA, so that the resulting cDNA is enriched for full-length transcripts. This is also the main disadvantage, as template switching will occur only if reverse transcription successfully reaches the 5′ end of the mRNA; any partially reverse-transcribed mRNA will fail to be amplified, which limits the total yield. Two alternative approaches have been published for processing the full-length cDNA: single-cell tagged reverse transcription (STRT)[104], which isolates and sequences the 5′ end, corresponding to the transcription start site; and switching mechanism at the 5′ end of the RNA template (SMART)–seq[11], which fragments the cDNA and generates reads that cover the full length of each transcript.

Protocols also differ in when they introduce a barcode for multiplexing. The great advantage of already introducing barcodes at the first step is that many cells (for example, 96) can be processed together in one tube, reducing both cost and time by a considerable factor. However, no early-multiplexed protocols are currently capable of sequencing the full length of RNA, because barcodes are added only to one end of each cDNA molecule.

Several recently published protocols are compared in TABLE 3. The most important differences between them are shown, but it is also important to stress that the approaches have much in common: similar detection limits (5–10 molecules of mRNA), quantitative biases due to amplification, limitation to polyadenylated RNAs, and gene-specific biases due to GC content or secondary structure.

Through automation and the optimization of reagent consumption, the sample preparation costs of all of the published protocols are similar, and the overall cost is dominated by the cost of sequencing. For a typical mammalian cell that contains 200,000 mRNA molecules, and assuming tenfold oversampling, at least two million reads must be generated. The current minimal cost per cell, when sequencing at high-throughput on an Illumina HiSeq 2000 machine, will be approximately US$10. However, sequencing costs continue to decrease exponentially, which should make it feasible, within five years, to analyse millions of single-cell transcriptomes.

Single-cell epigenomics and proteomics

Clearly, the genome and transcriptome of a cell capture only part of its state, and much of the function of the cell is determined by its epigenome and proteome, which add to the diversity of cells in a population. The epigenomic state of a cell includes epigenomic marks such as DNA methylation and histone methylation and acetylation, the structural and regulatory proteins bound to chromatin, the spatial interactions between enhancers and promoters forming transcriptional complexes, and the three-dimensional orientation of the chromosomes.

Bulk bisulphite sequencing provides information on the average DNA methylation states for groups of clustered CpG sites at a locus. Depletion of CpG methylation is associated with transcriptional activation, and may be a consequence of the binding of regulatory proteins. Bulk experiments can provide data on the distribution of methylation within cells or alleles[105,106], or support

models for the stochastic emergence of differential methylation[107]. However, in bulk experiments it is generally impossible to determine whether two methylated sites are actually present in an individual cell, unless the methylated sites are so close that they can be detected in a single sequencing read.

Chromatin immunoprecipitation followed by sequencing (ChIP–seq) is used to study protein–DNA interactions genome-wide[108], as well as to generate genome-wide maps of histone modifications. ChIP–seq has been used to determine genome-wide patterns of transcription factor binding and their relationships to active transcription and epigenomic marks. Using chromosome conformation analysis and all its derivative methods (for example, 3C[109], 4C[110] and Hi-C[111]), it is also possible to measure the interaction between distal chromatin elements directly, thus revealing the large-scale chromosome organization within the nucleus, as well as the finer details of enhancer–promoter interactions at individual loci. Again, however, by using bulk experiments it is impossible to know if a complex chromatin conformation or a combination of bound transcription factors actually exists in an individual cell. For example, consider the analysis of a tumour sample. The observation that a transcription factor is bound to a promoter and that the corresponding gene is transcribed does not necessarily imply that these two events have occurred in the same cell. Instead, it is possible that one event occurred in the tumour and another in the infiltrating stromal cells. Combined measurements of epigenomic and transcriptomic states in single cells are required to settle the issue.

Broad applications of sequencing-based methods to single-cell epigenomics have yet to be reported. The challenges in extending epigenetics to the single-cell level are similar to those faced by single-cell transcriptomics: avoiding loss of material and minimizing quantitative bias. For this reason, widespread and largely binary marks such as DNA methylation and histone modifications should be relatively easy to detect in single cells. Indeed, proof-of-concept single-cell epigenetic analyses have already been demonstrated for both DNA methylation[112,113] and histone modification[114]. By contrast, ChIP–seq targeting transcription factors in single cells is a formidable challenge because of the small number of transcription factors that are present in any single cell, the low affinity for their target sequence and the often imperfect nature of antibodies.

Epigenetic markers were used on bulk cell populations to analyse the dynamics of colorectal cancer[41,115,116] and to construct lineage trees for colon crypt stem cells[117–119].

Proteomic analysis methods include protein arrays[120], FACS analysis[121], co-immunoprecipitation[122], pull-down assays[123] and mass spectrometry assays[124], and they reveal different protein properties in a sample (for example, protein concentration or protein–protein interactions). Methods for DNA-based proteomic analysis have also been developed — for example, immuno-PCR[125] and proximity ligation assays[126] — and these were recently applied using NGS[7,127]. As in epigenomics, broad applications of sequencing-based methods to single-cell proteomics have yet to be reported, although preliminary proof-of-concept studies have been published[128,129].

Conclusions

Single cells are the fundamental units of life. Therefore, single-cell analysis is not just one more step towards more-sensitive measurements, but is a decisive jump to a more-fundamental understanding of biology. Here we have described recent advances in sequencing-based single-cell analysis. These advances include sequencing the genomes and transcriptomes of single cells, and we predict that it will soon be possible to sequence fully all the nucleic acids in many thousands or even millions of cells. In addition, we have described how other cellular phenomena can be converted into a DNA-sequence-based readout. For example, epigenomic marks such as histone modifications can be converted into a DNA signal by ChIP–seq. Similarly, protein modifications and interactions can be converted to DNA by the proximity ligation assay.

The enormous and ever-increasing power of DNA sequencing means that many different cellular phenomena are likely to be convertible to a DNA readout. A fortuitous consequence of this convergence should allow integrated measurements of multiple modalities. The feasibility of such integration has been already demonstrated for genomic and transcriptomic analysis[100], and simultaneous DNA, RNA and protein measurements in single cells can be used to quantitatively describe the central dogma of molecular biology[130]. Nonetheless, although single-cell analysis methods for single properties (such as only DNA or only RNA) are developing at a rapid pace, there is still a long road ahead for assaying multiple properties in single-cell integrated analyses. The biochemical differences between the cellular properties lead to variations in the methods that are needed to isolate them, and modifications of current isolation methods will be needed to develop a unified single-cell multi-property analysis protocol.

Such integrated single-cell genetic, epigenetic, transcriptional and proteomic sequencing-based analyses (FIG. 1), will allow modelling of the relationships among multiple molecular markers, unbiased identification of complex cell population structure, and characterization of direct, indirect and in some cases causal dependencies among factors. Development of complex single-cell genetic analysis methods may allow for a better understanding of these cellular properties and for redefining the concept of 'cell type'. The feasibility of such integration has been already demonstrated for genomic and transcriptomic analysis[100].

Finally, the accumulation of mutations in single cells during development can be used to infer the lineage ancestry of each cell. Although cell-fate maps describe potential next states for cells in a particular state[36], they do not capture precise lineage relationships. By contrast, cell lineage trees reconstructed using somatic mutations capture the lineage relationships among the sampled cells, but do not provide information on the state of ancestor cells. *C. elegans* is the first and highest

organism with a known cell lineage tree that captures both its cell fate map and the lineage relationships among cells[131,132]. We anticipate that integrated single-cell analysis culminating from the wave of developments reviewed here will allow similarly powerful results for higher organisms such as mice and humans. If the states of the sampled cells — as determined by their transcriptomes and epigenomes, and perhaps further enhanced by their proteomes — that constitute the leaves of a reconstructed cell lineage tree, could be known with high precision, then additional assumptions about the states of ancestor cells represented by internal nodes in the tree can be formalized into a mathematical model. This would allow the reconstruction paradigm to be expanded to describe state dynamics and to integrate cell lineage trees with cell fate mapping.

Cell lineage trees of higher organisms harbour answers to many open questions in human biology and medicine, and have the potential to transform medicine towards personalized, rather than generalized, diagnosis and treatment. Almost a decade ago it was suggested[16] that advances in single-cell genomics may inspire the initiation of a "Human Cell Lineage Project," the aim of which would be to reconstruct an entire human cell lineage tree. We believe that the advances reviewed and anticipated here in single-cell sequencing-based technologies will bring us closer to achieving this goal and along the way will revolutionize whole-organism science.

1. Wetterstrand, K. *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program* [online] http://www.genome.gov/sequencingcosts (2013).
2. Walker, T. M. *et al.* Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. *Lancet Infect. Dis.* **13**, 137–146 (2013).
3. Lander, E. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
4. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
5. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5**, 621–628 (2008).
6. Gu, H. *et al.* Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nature Protoc.* **6**, 468–481 (2011).
7. Darmanis, S. *et al.* ProteinSeq: high-performance proteomic analyses by proximity ligation and next generation sequencing. *PLoS ONE* **6**, e25583 (2011).
8. Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90–104 (2011).
9. 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).
10. Kalisky, T., Blainey, P. & Quake, S. R. Genomic analysis at the single-cell level. *Annu. Rev. Genet.* **45**, 431–445 (2011).
11. Ramskold, D. *et al.* Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nature Biotech.* **30**, 777–782 (2012).
This paper described the first single-cell RNA-seq method to achieve near full-length coverage of transcripts, and demonstrated transcriptome sequencing from single circulating tumour cells.
12. Dalerba, P. *et al.* Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nature Biotech.* **29**, 1120–1127 (2011).
13. Cristofanilli, M. *et al.* Circulating tumor cells: a novel prognostic factor for newly diagnosed metastatic breast cancer. *J. Clin. Oncol.* **23**, 1420–1430 (2005).
14. Blainey, P. C. The future is now: single-cell genomics of bacteria and archaea. *FEMS Microbiol. Rev.* **37**, 407–427 (2013).
A review of single-cell genomics of microorganisms, including currently available WGA techniques.
15. Gundry, M., Li, W., Maqbool, S. B. & Vijg, J. Direct, genome-wide assessment of DNA mutations in single cells. *Nucleic Acids Res.* **40**, 2032–2040 (2012).
16. Frumkin, D., Wasserstrom, A., Kaplan, S., Feige, U. & Shapiro, E. Genomic variability within an organism exposes its cell lineage tree. *PLoS Computat. Biol.* **1**, 382–394 (2005).
A conceptual and theoretical basis for organism cell lineage tree reconstruction using the genomic variability among organismal cells. It is also a preliminary proof-of-concept demonstration of reconstructing cell lineage trees using somatic mutations in a small panel of microsatellites.
17. Kurimoto, K., Yabuta, Y., Ohinata, Y. & Saitou, M. Global single-cell cDNA amplification to provide a template for representative high-density oligonucleotide microarray analysis. *Nature Protoc.* **2**, 739–752 (2007).
18. Reizel, Y. *et al.* Colon stem cell and crypt dynamics exposed by cell lineage reconstruction. *Plos Genet.* **7**, e1002192 (2011).
19. Shlush, L. I. *et al.* Cell lineage analysis of acute leukemia relapse uncovers the role of replication-rate heterogeneity and miscrosatellite instability. *Blood* **120**, 603–612 (2012).
20. Choi, J. H. *et al.* Development and optimization of a process for automated recovery of single cells identified by microengraving. *Biotechnol. Prog.* **26**, 888–895 (2010).
21. Zhang, H. & Liu, K. K. Optical tweezers for single cells. *J. R. Soc. Interface* **5**, 671–690 (2008).
22. Frumkin, D. *et al.* Amplification of multiple genomic loci from single cells isolated by laser micro-dissection of tissues. *BMC Biotechnol.* **8**,17 (2008).
23. Yachida, S. *et al.* Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature* **467**, 1114–1117 (2010).
24. Bhattacherjee, V. *et al.* Laser capture microdissection of fluorescently labeled embryonic cranial neural crest cells. *Genesis* **39**, 58–64 (2004).
25. Guo, M. T., Rotem, A., Heyman, J. A. & Weitz, D. A. Droplet microfluidics for high-throughput biological assays. *Lab. Chip* **12**, 2146–2155 (2012).
26. Fan, H., Wang, J., Potanina, A. & Quake, S. Whole-genome molecular haplotyping of single cells. *Nature Biotech.* **29**, 51–57 (2011).
27. 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).
28. White, A. *et al.* High-throughput microfluidic single-cell RT-qPCR. *Proc. Natl Acad. Sci. USA* **108**, 13999–14004 (2011).
29. Lecault, V., White, A. K., Singhal, A. & Hansen, C. L. Microfluidic single cell analysis: from promise to practice. *Curr. Opin. Chem. Biol.* **16**, 381–390 (2012).
30. Schatz, D. G. & Swanson, P. C. V(D)J recombination: mechanisms of initiation. *Annu. Rev. Genet.* **45**, 167–202 (2011).
31. Yates, L. R. & Campbell, P. J. Evolution of the cancer genome. *Nature Rev. Genet.* **13**, 795–806 (2012).
32. Reizel, Y. *et al.* Cell lineage analysis of the mammalian female germline. *PLoS Genet.* **8**, e1002477 (2012).
33. Szabat, M. *et al.* Maintenance of β-cell maturity and plasticity in the adult pancreas: developmental biology concepts in adult physiology. *Diabetes* **61**, 1365–1371 (2012).
34. Ming, G. & Song, H. Adult neurogenesis in the mammalian brain: significant answers and significant questions. *Neuron* **70**, 687–702 (2011).
35. Chojnacki, A. K., Mak, G. K. & Weiss, S. Identity crisis for adult periventricular neural stem cells: subventricular zone astrocytes, ependymal cells or both? *Nature Rev. Neurosci.* **10**, 153–163 (2009).
36. Yona, S. *et al.* Fate mapping reveals origins and dynamics of monocytes and tissue macrophages under homeostasis. *Immunity* **38**, 79–91 (2013).
37. Schepers, A. G. *et al.* Lineage tracing reveals Lgr5$^+$ stem cell activity in mouse intestinal adenomas. *Science* **337**, 730–735 (2012).
38. Carlson, C. *et al.* Decoding cell lineage from acquired mutations using arbitrary deep sequencing. *Nature Methods* **9**, 78–80 (2012).
39. Ellegren, H. Microsatellites: Simple sequences with complex evolution. *Nature Rev. Genet.* **5**, 435–445 (2004).
40. Salipante, S. & Horwitz, M. Phylogenetic fate mapping. *Proc. Natl Acad. Sci. USA* **103**, 5448–5453 (2006).
41. Tsao, J. *et al.* Colorectal adenoma and cancer divergence - evidence of multilineage progression. *Am. J. Pathol.* **154**, 1815–1824 (1999).
42. Zhou, W. *et al.* Use of somatic mutations to quantify random contributions to mouse development. *BMC Genomics* **14**, 39 (2013).
43. Vilkki, S. *et al.* Extensive somatic microsatellite mutations in normal human tissue. *Cancer Res.* **61**, 4541–4544 (2001).
44. Wasserstrom, A. *et al.* Reconstruction of cell lineage trees in mice. *PLoS ONE* **3**, e1939 (2008).
45. Wasserstrom, A. *et al.* Estimating cell depth from somatic mutations. *PLoS Computat. Biol.* **4**, e1000058 (2008).
46. Segev, E. *et al.* Muscle-bound primordial stem cells give rise to myofiber-associated myogenic and non-myogenic progenitors. *PLoS ONE* **6**, e25605 (2011).
47. Ross, M. G. *et al.* Characterizing and measuring bias in sequence data. *Genome Biol.* **14**, R51 (2013).
48. Fidler, I. & Kripke, M. Metastasis results from preexisting variant cells within a malignant-tumor. *Science* **197**, 893–895 (1977).
49. Kim, M. Y. *et al.* Tumor self-seeding by circulating cancer cells. *Cell* **139**, 1315–1326 (2009).
50. Fidler, I. Critical determinants of metastasis. *Seminars Cancer Biol.* **12**, 89–96 (2002).
51. Eaves, C. J. Cancer stem cells: here, there, everywhere? *Nature* **456**, 581–582 (2008).
52. Frank, N. Y., Schatton, T. & Frank, M. H. The therapeutic promise of the cancer stem cell concept. *J. Clin. Invest.* **120**, 41–50 (2010).
53. Pawelek, J. M. & Chakraborty, A. K. Fusion of tumour cells with bone marrow-derived cells: a unifying explanation for metastasis. *Nature Rev. Cancer* **8**, 377–386 (2008).
54. Lazova, R. *et al.* A melanoma brain metastasis with a donor-patient hybrid genome following bone marrow transplantation: first evidence for fusion in human cancer. *PLoS ONE* **8**, e66731 (2013).
55. Blagosklonny, M. V. Target for cancer therapy: proliferating cells or stem cells. *Leukemia* **20**, 385–391 (2006).
56. Xu, X. *et al.* Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell* **148**, 886–895 (2012).
57. Anderson, K. *et al.* Genetic variegation of clonal architecture and propagating cells in leukaemia. *Nature* **469**, 356–361 (2011).
58. Baslan, T. *et al.* Genome-wide copy number analysis of single cells. *Nature Protoc.* **7**, 1024–1041 (2012).
59. Hou, Y. *et al.* Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. *Cell* **148**, 873–885 (2012).

60. Jan, M. *et al.* Clonal evolution of preleukemic hematopoietic stem cells precedes human acute myeloid leukemia. *Sci. Transl Med.* **4**, 149ra118 (2012).
61. Ding, L. *et al.* Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature* **481**, 506–510 (2012).
62. Gerlinger, M. *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* **366**, 883–892 (2012).
An exposition of the heterogeneity within different regions in a single tumour, demonstrating the importance of the integration of several analysis methods including DNA and RNA sequencing.
63. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
64. Cheung, V. & Nelson, S. Whole genome amplification using a degenerate oligonucleotide primer allows hundreds of genotypes to be performed on less than one nanogram of genomic DNA. *Proc. Natl Acad. Sci. USA* **93**, 14676–14679 (1996).
65. Arneson, N., Hughes, S., Houlston, R. & Done, S. Whole-genome amplification by improved primer extension preamplification PCR (I-PEP-PCR). *CSH Protoc.* **2008**, pdb.prot4921 (2008).
66. Klein, C. A. *et al.* Comparative genomic hybridization, loss of heterozygosity, and DNA sequence analysis of single cells. *Proc. Natl Acad. Sci. USA* **96**, 4494–4499 (1999).
67. Dean, F. *et al.* Comprehensive human genome amplification using multiple displacement amplification. *Proc. Natl Acad. Sci. USA* **99**, 5261–5266 (2002).
68. Lu, S. *et al.* Probing meiotic recombination and aneuploidy of single sperm cells by whole-genome sequencing. *Science* **338**, 1627–1630 (2012).
69. Peng, W., Takabayashi, H. & Ikawa, K. Whole genome amplification from single cells in preimplantation genetic diagnosis and prenatal diagnosis. *Eur. J. Obstet. Gynecol. Reprod. Biol.* **131**, 13–20 (2007).
70. Salipante, S. J., Kas, A., McMonagle, E. & Horwitz, M. S. Phylogenetic analysis of developmental and postnatal mouse cell lineages. *Evol. Dev.* **12**, 84–94 (2010).
71. Zaretsky, I. *et al.* Monitoring the dynamics of primary T cell activation and differentiation using long term live cell imaging in microwell arrays. *Lab. Chip* **12**, 5007–5015 (2012).
72. Harris, T. D. *et al.* Single-molecule DNA sequencing of a viral genome. *Science* **320**, 106–109 (2008).
73. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138 (2009).
74. Schadt, E., Turner, S. & Kasarskis, A. A window into third-generation sequencing. *Hum. Mol. Genet.* **19**, R227–R240 (2010).
75. Xu, M., Fujita, D. & Hanagata, N. Perspectives and challenges of emerging single-molecule DNA sequencing technologies. *Small* **5**, 2638–2649 (2009).
76. Ng, S. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272–276 (2009).
77. Teer, J. & Mullikin, J. Exome sequencing: the sweet spot before whole genomes. *Hum. Mol. Genet.* **19**, R145–R151 (2010).
78. Giulino-Roth, L. *et al.* Targeted genomic sequencing of pediatric Burkitt lymphoma identifies recurrent alterations in antiapoptotic and chromatin-remodeling genes. *Blood* **120**, 5181–5184 (2012).
79. Valencia, C. A. *et al.* Comprehensive mutation analysis for congenital muscular dystrophy: a clinical PCR-based enrichment and next-generation sequencing panel. *PLoS ONE* **8**, e53083 (2013).
80. Hollants, S., Redeker, E. & Matthijs, G. Microfluidic amplification as a tool for massive parallel sequencing of the familial hypercholesterolemia genes. *Clin. Chem.* **58**, 717–724 (2012).
81. Tewhey, R. *et al.* Microdroplet-based PCR enrichment for large-scale targeted sequencing. *Nature Biotech.* **27**, 1025–1031 (2009).
82. Li, J. *et al.* Multiplex padlock targeted sequencing reveals human hypermutable CpG variations. *Genome Res.* **19**, 1606–1615 (2009).
83. Johansson, H. *et al.* Targeted resequencing of candidate genes using selector probes. *Nucleic Acids Res.* **39**, e8 (2011).
84. Diaz-Horta, O. *et al.* Whole-exome sequencing efficiently detects rare mutations in autosomal recessive nonsyndromic hearing loss. *PLoS ONE* **7**, e50628 (2012).
85. Arendt, D. The evolution of cell types in animals: emerging principles from molecular studies. *Nature Rev. Genet.* **9**, 868–882 (2008).
In this Review, the author discusses the origin and evolution of diverse cell types in animals, an issue that has been curiously neglected by biologists.
86. Vickaryous, M. K. & Hall, B. K. Human cell type diversity, evolution, development, and classification with special reference to cells derived from the neural crest. *Biol. Rev. Cambridge Philos. Soc.* **81**, 425–455 (2006).
This paper is a careful review of all human cell types that have been given names in the literature, which is a useful starting point for future cell-type discovery experiments.
87. Gehlenborg, N. *et al.* Visualization of omics data for systems biology. *Nature Methods* **7**, S56–S68 (2010).
88. Johnston, I. G. *et al.* Mitochondrial variability as a source of extrinsic cellular noise. *PLoS Computat. Biol.* **8**, e1002416 (2012).
89. Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y. & Tyagi, S. Stochastic mRNA synthesis in mammalian cells. *PLoS Biol.* **4**, e309 (2006).
This study using single-molecule imaging of mRNAs shows that mRNA abundances vary tremendously within putatively homogenous cell populations, and provides initial estimates of transcriptional burst kinetics in mammalian cells.
90. Raj, A. & Vanoudenaarden, A. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell* **135**, 216–226 (2008).
91. Endele, M. & Schroeder, T. Molecular live cell bioimaging in stem cell research. *Ann. NY Acad. Sci.* **1266**, 18–27 (2012).
92. Ozsolak, F. *et al.* Direct RNA sequencing. *Nature* **461**, 814–818 (2009).
93. Casbon, J. A., Osborne, R. J., Brenner, S. & Lichtenstein, C. P. A method for counting PCR template molecules with application to next-generation sequencing. *Nucleic Acids Res.* **39**, e81 (2011).
94. Kivioja, T. *et al.* Counting absolute numbers of molecules using unique molecular identifiers. *Nature Methods* **9**, 72–74 (2011).
95. Shiroguchi, K., Jia, T. Z., Sims, P. A. & Xie, X. S. Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. *Proc. Natl Acad. Sci. USA* **109**, 1347–1352 (2012).
96. Fu, G. K., Hu, J., Wang, P. H. & Fodor, S. P. Counting individual DNA molecules by the stochastic attachment of diverse labels. *Proc. Natl Acad. Sci. USA* **108**, 9026–9031 (2011).
97. Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K. W. & Vogelstein, B. Detection and quantification of rare mutations with massively parallel sequencing. *Proc. Natl Acad. Sci. USA* **108**, 9530–9535 (2011).
98. Eberwine, J. *et al.* Analysis of gene expression in single live neurons. *Proc. Natl Acad. Sci. USA* **89**, 3010–3014 (1992).
99. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep.* **2**, 666–673 (2012).
100. Klein, C. A. *et al.* Combined transcriptome and genome analysis of single micrometastatic cells. *Nature Biotech.* **20**, 387–392 (2002).
This study reported a simultaneous genomic and transcriptomic analysis of individual cells using a microarray readout. This is a first example of an integrated single-cell analysis.
101. Kurimoto, K. *et al.* An improved single-cell cDNA amplification method for efficient high-density oligonucleotide microarray analysis. *Nucleic Acids Res.* **34**, e42 (2006).
102. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods* **6**, 377–382 (2009).
The first demonstration of single-cell RNA-seq with accurate detection of alternatively spliced transcripts in single mouse oocytes.
103. Maleszka, R. & Stange, G. Molecular cloning, by a novel approach, of a cDNA encoding a putative olfactory protein in the labial palps of the moth *Cactoblastis cactorum*. *Gene* **202**, 39–43 (1997).
104. Islam, S. *et al.* Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.* **21**, 1160–1167 (2011).
The first demonstration of highly multiplexed single-cell RNA-seq showing that cell types can be distinguished in an unbiased manner on the basis of unfiltered single-cell gene expression profiles.
105. Arand, J. *et al. In vivo* control of CpG and non-CpG DNA methylation by DNA methyltransferases. *PLoS Genet.* **8**, e1002750 (2012).
106. Taylor, K. H. *et al.* Ultradeep bisulfite sequencing analysis of DNA methylation patterns in multiple gene promoters by 454 sequencing. *Cancer Res.* **67**, 8511–8518 (2007).
107. Landan, G. *et al.* Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues. *Nature Genet.* **44**, 1207–1214 (2012).
108. Jothi, R., Cuddapah, S., Barski, A., Cui, K. & Zhao, K. Genome-wide identification of *in vivo* protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Res.* **36**, 5221–5231 (2008).
109. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306–1311 (2002).
110. van de Werken, H. J. *et al.* Robust 4C-seq data analysis to screen for regulatory DNA interactions. *Nature Methods* **9**, 969–972 (2012).
111. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
112. Kantlehner, M. *et al.* A high-throughput DNA methylation analysis of a single cell. *Nucleic Acids Res.* **39**, e44 (2011).
113. Denomme, M. M., Zhang, L. & Mann, M. R. Single oocyte bisulfite mutagenesis. *J. Vis. Exp.* **64**, e4046 (2012).
114. Hayashi-Takanaka, Y. *et al.* Tracking epigenetic histone modifications in single cells using Fab-based live endogenous modification labeling. *Nucleic Acids Res.* **39**, 6475–6488 (2011).
115. Tsao, J. L. *et al.* Genetic reconstruction of individual colorectal tumor histories. *Proc. Natl Acad. Sci. USA* **97**, 1236–1241 (2000).
116. Siegmund, K., Marjoram, P., Woo, Y., Tavare, S. & Shibata, D. Inferring clonal expansion and cancer stem cell dynamics from DNA methylation patterns in colorectal cancers. *Proc. Natl Acad. Sci. USA* **106**, 4828–4833 (2009).
117. Nicolas, P., Kim, K., Shibata, D. & Tavare, S. The stem cell population of the human colon crypt: analysis via methylation patterns. *PLoS Computat. Biol.* **3**, 364–374 (2007).
118. Yatabe, Y., Tavaré, S. & Shibata, D. Investigating stem cells in human colon by using methylation patterns. *Proc. Natl Acad. Sci. USA* **98**, 10839–10844 (2001).
119. Kim, K. M. & Shibata, D. Methylation reveals a niche: stem cell succession in human colon crypts. *Oncogene* **21**, 5441–5449 (2002).
120. Hodgkinson, V., ElFadl, D., Drew, P., Lind, M. & Cawkwell, L. Repeatedly identified differentially expressed proteins (RIDEPs) from antibody microarray proteomic analysis. *J. Proteom.* **74**, 698–703 (2011).
121. Bendall, S. *et al.* Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science* **332**, 687–696 (2011).
122. Lee, H. W. *et al.* Real-time single-molecule co-immunoprecipitation analyses reveal cancer-specific Ras signalling dynamics. *Nature Commun.* **4**, 1505 (2013).
123. Jain, A. *et al.* Probing cellular protein complexes using single-molecule pull-down. *Nature* **473**, 484–488 (2011).
124. Keshishian, H. *et al.* Quantification of cardiovascular biomarkers in patient plasma by targeted mass spectrometry and stable isotope dilution. *Mol. Cell Proteom.* **8**, 2339–2349 (2009).
125. Niemeyer, C., Adler, M. & Wacker, R. Detecting antigens by quantitative immuno-PCR. *Nature Protoc.* **2**, 1918–1930 (2007).
126. Fredriksson, S. *et al.* Multiplexed protein detection by proximity ligation for cancer biomarker validation. *Nature Methods* **4**, 327–329 (2007).
127. Turner, D. J. *et al.* Toward clinical proteomics on a next-generation sequencing platform. *Anal. Chem.* **83**, 666–670 (2011).
128. Salehi-Reyhani, A. *et al.* A first step towards practical single cell proteomics: a microfluidic antibody capture chip with TIRF detection. *Lab. Chip* **11**, 1256–1261 (2011).
129. Shi, Q. *et al.* Single-cell proteomic chip for profiling intracellular signaling pathways in single tumor cells. *Proc. Natl Acad. Sci. USA* **109**, 419–424 (2012).

130. Li, G. W. & Xie, X. S. Central dogma at the single-molecule level in living cells. *Nature* **475**, 308–315 (2011).
A review of the central dogma of molecular biology in terms of stochastic kinetics in single cells and of imaging-based methods for single-cell and single-molecule analysis.
131. Sulston, J. E., Schierenberg, E., White, J. G. & Thomson, J. N. The embryonic cell lineage of the nematode *Caenorhabditis elegans*. *Dev. Biol.* **100**, 64–119 (1983).
132. Sulston, J. E. & Horvitz, H. R. Post-embryonic cell lineages of the nematode, *Caenorhabditis elegans*. *Dev. Biol.* **56**, 110–156 (1977).
The first reconstruction of a complete organism cell lineage, of the *C. elegans* nematode, published almost four decades ago. Complete cell lineage trees of higher organisms are yet to be reconstructed.
133. Noctor, S. C., Martinez-Cerdeno, V., Ivic, L. & Kriegstein, A. R. Cortical neurons arise in symmetric and asymmetric division zones and migrate through specific phases. *Nature Neurosci.* **7**, 136–144 (2004).
134. Murray, J. *et al.* Automated analysis of embryonic gene expression with cellular resolution in *C. elegans*. *Nature Methods* **5**, 703–709 (2008).
135. Murray, J. *et al.* Multidimensional regulation of gene expression in the *C. elegans* embryo. *Genome Res.* **22**, 1282–1294 (2012).
136. DeKosky, B. J. *et al.* High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nature Biotech.* **31**, 166–169 (2013).
137. Peters, B. *et al.* Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature* **487**, 190–195 (2012).
138. Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
139. Timmermann, B. *et al.* Somatic mutation profiles of MSI and MSS colorectal cancer identified by whole exome next generation sequencing and bioinformatics analysis. *PLoS ONE* **5**, e15661 (2010).
140. Diep, D. *et al.* Library-free methylation sequencing with bisulfite padlock probes. *Nature Methods* **9**, 270–272 (2012).
141. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008).
142. Nagalakshmi, U. *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**, 1344–1349 (2008).
143. Cloonan, N. *et al.* Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nature Methods* **5**, 613–619 (2008).
144. Sasagawa, Y. *et al.* Quartz-Seq: a highly reproducible and sensitive single-cell RNA-Seq reveals non-genetic gene expression heterogeneity. *Genome Biol.* **14**, R31 (2013).

Acknowledgements
The work of S.L. was supported by grant 261063 from the European Research Council and by the Swedish Research Council STARGET consortium. The work of E.S. and T.B. was supported by The European Union FP7-ERC-AdG grant and by a grant from the Kenneth and Sally Leafman Appelbaum Discovery Fund. E.S. is the Incumbent of The Harry Weinrebe Professorial Chair of Computer Science and Biology. The contribution of E.S. to this Review was inspired by a research proposal prepared by E.S. in collaboration with I. Amit, A. Tanay and M. Schwarz.

Competing interests statement
The authors declare no competing financial interests.

FURTHER INFORMATION
Nature Reviews Genetics **article series on applications of next-generation sequencing:** http://www.nature.com/nrg/series/nextgeneration/index.html

ALL LINKS ARE ACTIVE IN THE ONLINE PDF