APPLICATIONS OF NEXT-GENERATION SEQUENCING

Cancer transcriptome profiling at the juncture of clinical translation

Marcin Cieślik[1,2] and Arul M. Chinnaiyan[1–5]

Abstract | Methodological breakthroughs over the past four decades have repeatedly revolutionized transcriptome profiling. Using RNA sequencing (RNA-seq), it has now become possible to sequence and quantify the transcriptional outputs of individual cells or thousands of samples. These transcriptomes provide a link between cellular phenotypes and their molecular underpinnings, such as mutations. In the context of cancer, this link represents an opportunity to dissect the complexity and heterogeneity of tumours and to discover new biomarkers or therapeutic strategies. Here, we review the rationale, methodology and translational impact of transcriptome profiling in cancer.

[1]Michigan Center for Translational Pathology, University of Michigan.
[2]Department of Pathology, University of Michigan.
[3]Comprehensive Cancer Center, University of Michigan.
[4]Department of Urology, University of Michigan.
[5]Howard Hughes Medical Institute, University of Michigan, Ann Arbor, Michigan 48109, USA.

Correspondence to A.M.C. arul@umich.edu

doi:10.1038/nrg.2017.96
Published online 27 Dec 2017

Transcriptomics is the large-scale study of RNA molecules by use of high-throughput techniques. It examines the abundance and makeup of a cell's transcriptome[1,2]. In contrast to DNA, which is largely identical across all cells of an organism, the actively transcribed RNA is highly dynamic, reflecting the diversity of cell types, cellular states and regulatory mechanisms. Because a transcriptome profile can be regarded as a signature or snapshot of the underlying cell state, the experimental profiling of samples and specimens can provide insights into their unique biology.

Depending on the specific approach, transcriptomics can not only reveal the architecture of gene expression but also provide details on the structure, modification[3] and variation of individual transcripts[4,5]. Advances in transcriptome profiling, specifically the development of genome-wide methodologies targeting diverse RNA species, have enabled us to discover the seemingly endless complexity of RNA biology and to comprehensively annotate the human genome and other eukaryotic genomes[6]. Arguably, transcriptomics is currently the most well-established modality and foundation of functional genomics, a field of study for which the goal is to synthesize large-scale data to understand the mechanisms that govern cellular and organismal phenotypes[7].

Research on the human transcriptome has identified the molecular underpinnings of many biological processes and diseases, including cancer. These novel technologies provided major insights into the aetiology and pathogenesis of several cancers as well as newfound clinical applications[8]. In particular, transcriptome-wide gene expression profiling has proved useful to better understand the molecular mechanisms underlying prognosis and drug sensitivity[9,10]. Cancer cells are characterized by altered protein function and aberrant transcriptional patterns, which are the consequence of somatic mutations and epigenetic alterations. These molecular phenotypes impinge upon the growth advantage of the cancer cells and are subject to natural selection[11]. Remarkably, the surviving cells converge upon prominent expression profiles that are concordant across experiments and samples[12] and similar to the transcriptional modules and cell states[13] in normal tissues. Importantly, gene expression continues to be among the most powerful molecular profiling data to predict drug sensitivity[14,15]. This unique capacity to describe the high-dimensional molecular state of cancer was historically one of the primary applications of transcriptomics[16]. However, with the advent of whole-transcriptome sequencing, additional readouts that ascertain the chemical modifications, sequence, interactions and even shape of transcripts became feasible. The study of alternative splicing, RNA editing, post-transcriptional modifications and various non-coding RNAs is now an essential aspect of transcriptomics. Of particular relevance to cancer, the base-pair resolution and coverage of modern techniques enabled the detection of expressed somatic mutations, including single nucleotide variants (SNVs)[17], and gene fusions[18].

Transcriptomics is now at a pivotal juncture. On the one hand, the field has been revolutionized by diverse next-generation sequencing (NGS) methodologies[19] and has expanded beyond the measurement of expression of protein-coding genes[20,21]. On the other hand, genomic discoveries are being increasingly and rapidly translated into the clinic to improve diagnosis or guide treatment[8]. Here, we begin by briefly reviewing how cancer transcriptome profiling unfolded over the past four decades. We then describe how transcriptomic data synergize with DNA-based assays by linking the genetic and

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

phenotypic aspects of tumour biology. Special attention is given to the progress in cancer transcriptome research that resulted from the introduction of high-throughput sequencing and advanced bioinformatics methodologies. We conclude by reviewing current and likely future clinical applications of transcriptomics, such as RNA profiling of single cells and liquid biopsies.

Cancer transcriptomics: four decades of progress

Although transcriptomics exploded with the advent of RNA sequencing (RNA-seq), continued progress in transcript profiling over the past four decades has expanded our understanding of the genetics and molecular biology of cancer. New experimental methodologies, in conjunction with advances in bioinformatics and efforts to catalogue and disseminate the results, have led to several fundamental discoveries in cancer biology. Behind that progress was the constant push for an increase in the breadth, depth or fidelity of measuring cellular RNA (FIG. 1). Enabled by the discovery of nucleic acid hybridization[22] and DNA sequencing[23], transcriptomics was jump-started in 1977 with the invention of the northern blot[24] and the first sequences of cloned cDNAs[25–27]. The development of expressed sequence tags (ESTs)[28] and reverse transcription quantitative PCR (RT-qPCR)[29–32] made it possible for the first time to identify cellular mRNAs in an unbiased and quantitative way, respectively.

The increasing knowledge of expressed sequences spurred the invention and design of cDNA[33] and oligonucleotide microarrays[34], which made it practical in terms of cost and labour to measure the expression levels of thousands of known genes simultaneously. Quantitative sequencing was also made more practical with the realization that very short sequences (that is, tags) are largely sufficient to identify a transcript[35], which led to the development of serial analysis of gene expression (SAGE)[36]. Although SAGE was out-competed by microarray technologies for routine expression profiling, the method has been simplified and adapted to short-read sequencing technologies and laid a foundation for tag-based high-throughput sequencing, such as digital gene expression[37], including the latest single-cell techniques[38]. This competition between hybridization-based and sequencing-based techniques continued over the next four decades and fuelled constant improvements, such as exon-tiling microarrays[6] and full-length cDNA sequencing[39]. Transcriptomics was completely transformed by the introduction of random priming to cDNA amplification, resulting in shotgun EST sequencing[40,41], the development of the first high-throughput sequencing methods[42] and the draft of the human genome[43]. Although many of the sequencing techniques, including EST sequencing[44] and SAGE[45], were adapted to high-throughput, short-read sequencing platforms (known as second-generation platforms), it was the random-primed approach followed by shotgun sequencing[46,47] that established RNA-seq as the protocol of choice. The introduction of unique molecular identifiers (UMIs)[48], PCR-free techniques[49] and cDNA hybridization-based approaches[50–52] further expanded the range of possible applications (see below). Despite these varied applications of RNA-seq, microarrays and alternative approaches, such as NanoString, continue to be popular owing to their relative simplicity and sometimes improved performance[53].

Figure 1 | **A historical timeline of transcriptomics.** Illustrated is the lockstep development of experimental and computational aspects of transcriptomics. Advances in the experimental protocols for the high-throughput profiling of RNA necessitate the development of databases to catalogue the results and trigger curation efforts to define reference transcriptomes. However, these endeavours depend on the development of accurate and scalable computational methods to search, quantify and assemble RNA molecules. Within each field, the most influential, seminal or unique references were selected. AceView, a gene annotation resource[267]; ArrayDB, a database of microarray gene expression data[268]; ArrayExpress, a public repository for microarray gene expression data[78]; BLAST, Basic Local Alignment Search Tool[73]; CAGE, cap analysis of gene expression[269]; CEL-seq, cell expression by linear amplification and sequencing[49]; CGAP, Cancer Genome Anatomy Project[270]; CIBERSORT, a tool for estimating the abundances of cell types in a mixed cell population[260]; dbEST, a database for expressed sequence tags[68]; EdgeR, a package for differential expression analysis[170]; EMBL, European Molecular Biology Laboratory; Ensembl, a genome browser for vertebrate genomes[74]; ESTs, expressed sequence tags[28]; FANTOM5, Functional Annotation of the Mammalian Genome 5 (REF. 271); FASTA, a text format for representing nucleotide or peptide sequences[72]; GenBank, the US National Institutes of Health (NIH) genetic sequence database; GENCODE, the genome annotation project of the Encyclopedia of DNA Elements (ENCODE)[272]; GenomeSpace, a cloud-based resource for integrative genomics analyses[83]; GEO, Gene Expression Omnibus[77]; GSEA, gene set enrichment analysis[273]; InsilicoDB, a database of microarray and RNA-seq data[82]; Known Genes, a resource of RNA and protein data[274]; Limma, Linear Models for Microarray Data[167]; MiTranscriptome, a human RNA-seq database[76]; Mitelman, Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer[275]; MPSS, massively parallel signature sequencing[42]; Oncomine, a cancer microarray database and integrated data mining platform[180]; qPCR, quantitative PCR[29]; RACE, rapid amplification of cDNA ends[276]; RefSeq, NCBI Reference Sequence Database[75]; RNAscope, an *in situ* hybridization assay for RNA detection[277]; RNA-seq, RNA sequencing; RNA-seq 454, RNA sequencing using the 454 (Roche) pyrosequencing platform[44]; RNA-seq SBS, RNA sequencing using sequencing-by-synthesis platforms[278]; RT-qPCR, reverse transcription quantitative PCR[30–32]; SAGE, serial analysis of gene expression[36]; SAGEmap, SAGE tag to gene mapping[279]; Sailfish, a transcript isoform quantification tool[280]; SAM, Significance Analysis of Microarrays[281]; Smith–Waterman, a local sequence alignment algorithm[70]; STAR, Spliced Transcripts Alignment to a Reference[282]; Symatlas, gene expression and annotation resource, now superseded by BioGPS[283]; TACO, Transcriptome Assemblies Combined into One (a consensus transcriptome tool)[284]; TopHat and Cufflinks, software tools for RNA-seq alignment and transcriptome assembly[5]; Trans-ABySS, Transcript Assembly By Short Sequences[285]; Trinity, a tool for *de novo* assembly of RNA-seq data[286]; UMI, unique molecular identifier[48]; Xena, a genomic data mining and analysis portal[287].

RNA sequencing
(RNA-seq). An encompassing term for all cDNA profiling techniques using high-throughput sequencing.

cDNAs
DNA molecules obtained through reverse transcription of RNAs.

Expressed sequence tags
(ESTs). Short fragments of a cDNA sequence that identify (tag) a transcript.

Microarrays
A method of cDNA profiling through hybridization and fluorescent labelling.

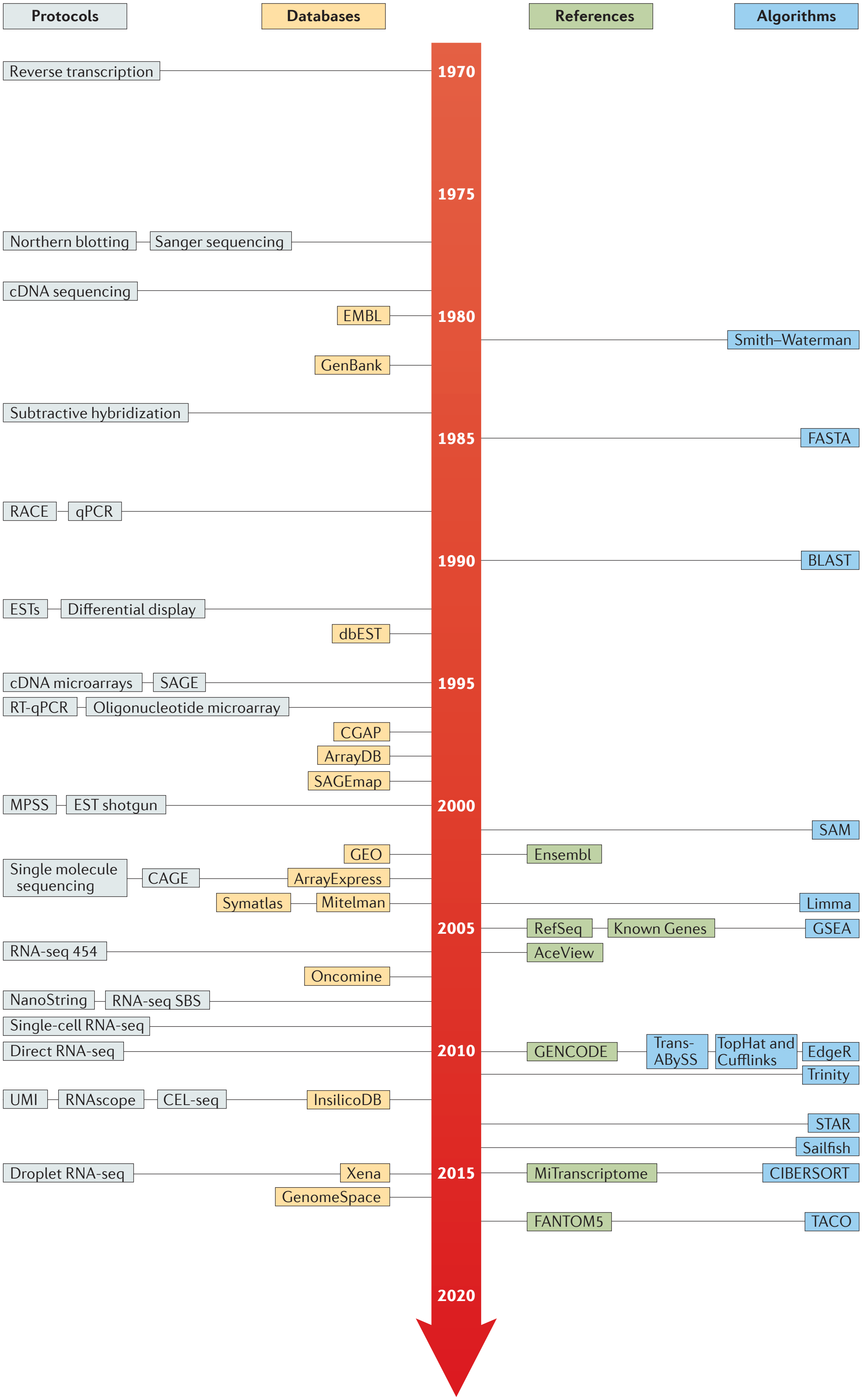
Serial analysis of gene expression
(SAGE). An economical technique for sequencing very short tags (11 nucleotides) from multiple cDNAs in one Sanger sequencing run.

Digital gene expression
A high-throughput, low-cost technique for expression profiling that involves sequencing short tags rather than the whole transcript.

Unique molecular identifiers
(UMIs). Sequences that are unique to each reverse-transcribed cDNA. PCR duplicates share the same UMI.

NanoString
A barcoding-based and imaging-based technique for the detection and quantification of hundreds of transcripts.

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Most of the transcriptomic methods were either developed for, or immediately applied to, cancer research. On the one hand, targeted hybridization and sequencing contributed to innumerable discoveries, including the first oncogene (*SRC*)[54] and tumour suppressor (*RB1*)[28]. On the other hand, high-throughput methods delivered an increasingly complete view of cancer gene expression. The first unbiased transcriptome was obtained by sequencing ESTs from the cancer cell line HepG2 (REF. 28); later, additional normal and cancer cells were profiled using SAGE[55]. The generation and sequencing of comprehensive cDNA libraries still required enormous resources, and large international projects were formed[56] to profile normal and tumour tissues by use of shotgun sequencing. With the advent of microarrays, gene expression profiling of cells[57] and tumour tissues became much less resource intensive and more routine[58]. Likewise, the dramatic reduction in sequencing costs following the introduction of NGS platforms made deep sequencing of thousands of transcriptomes feasible. In recognition of their scientific value, transcriptomic data sets became a key modality in large-scale molecular profiling efforts of cancer cell lines (such as the Encyclopedia of DNA Elements (ENCODE)[59], the Cancer Cell Line Encyclopedia (CCLE)[60] and Genentech[61]), of normal tissues (such as the Genotype–Tissue Expression (GTEx) project[62] and the Human Protein Atlas (HPA)[63]) and of tumour tissues (such as The Cancer Genome Atlas (TCGA)[64] and the Stand Up To Cancer–Prostate Cancer Foundation (SU2C–PCF) project[65]). Overall, RNA-seq has matured into the most robust and comprehensive transcriptome profiling assay, virtually subsuming all applications of expression microarrays.

The exponential growth of EST, microarray and RNA-seq data sets put particular pressure on the availability of informatics tools to store, find and compare them. Initially, all sequences were stored in the European Molecular Biology Laboratory (EMBL)[66] and GenBank[67] nucleotide databases. To capture the quantitative aspects of transcription, dedicated databases were made for EST[68] and SAGE[69] libraries. Rigorous sequence searching became possible with the Smith–Waterman algorithm[70], which was later adapted to align sequenced cDNA to the reference genome[71] and optimized for speed in the widely popular FASTA[72] and Basic Local Alignment Search Tool (BLAST)[73] programs. These algorithmic developments continue to have an imprint on the design of sequence aligners in the RNA-seq era. Over time, the high redundancy of sequence databases and the availability of the human genome sequence necessitated efforts to catalogue the human transcriptome and annotate the human genome. This need produced a number of human reference transcriptomes, including Ensembl[74] and RefSeq[75], which continue to be updated. More recent discoveries of pervasive and aberrant transcription in cancer renewed interest in more comprehensive and disease-specific transcriptome annotation[76]. Analogous computational resources were also developed for microarrays, including the Gene Expression Omnibus (GEO)[77] and ArrayExpress[78] databases, approaches for data extraction[79] and statistical methods[80]. Later, all these aspects of microarray analysis were integrated within end-to-end (often commercial or cancer-specific) data mining portals[81–83].

RNA bridges genetic causes to phenotypic effects

Tumour phenotypes are determined by the accumulation of genetic and epigenetic aberrations followed by clonal expansion of the fittest cells. The resulting tumours show intricate characteristics that reflect the diversity of the selective forces[84]. Remarkably, although tumours evolve independently, most of them ultimately exhibit similar traits that are widely regarded as the hallmarks of cancer. Many of these phenotypes require the extensive alteration of cell signalling and biochemical pathways. For example, metastasis necessitates, among other properties, the loss of E-cadherin expression and decreased cell adhesion[85], whereas immune evasion can involve the upregulation of immune checkpoints[86]. Changes in gene activity can be regarded as surrogates for many phenotypes, such as inflammation, vascularization, apoptosis[87], proliferation[88] and genomic instability[89]. The extent to which the tumour is successful in acquiring these traits influences important clinical variables, such as growth rate, metastatic potential[90] and response to drugs, and ultimately determines clinical progression and outcomes[91].

Transcriptome profiling can detect changes in gene activity and regulation by capturing quantitative expression patterns and has the capacity to describe the underlying phenotypes in great detail. Furthermore, many genetic and epigenetic events can be either directly observed or indirectly inferred from transcriptomic data. The primary readouts of modern-day cancer transcriptomics can be broadly categorized as genetic and functional (FIG. 2). Whereas functional measurements benefit mostly from the breadth of genome-wide assays, the detection of genetic events required increased depth and base pair resolution.

Functional phenotypic insights from transcriptomics. The quest for quantitative and genome-wide gene expression profiling was the motivation for the development of many of the transcriptomics techniques, such as SAGE[36], microarray[34] and RNA-seq[47]. Expression profiling of tumours and cancer cell lines was one of the first applications of each of those techniques[55,60,92] and became a routine aspect of cancer biology. Although each of those offered successive improvements in fidelity, the major advancements came from expanding the universe of surveyed genes (see below) and, in parallel, from the development of statistical and bioinformatic methods that facilitated analysis and interpretation (see below). A quantitative breakdown of gene expression levels into individual exons[93] or transcript isoforms, that is, transcript-level expression profiling[5], is also possible, but its applications in cancer remain underexplored.

Differences among transcript isoforms, such as alternative initiation[94], termination[95] and splicing[96], are effectively probed using sequencing-based methods.

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

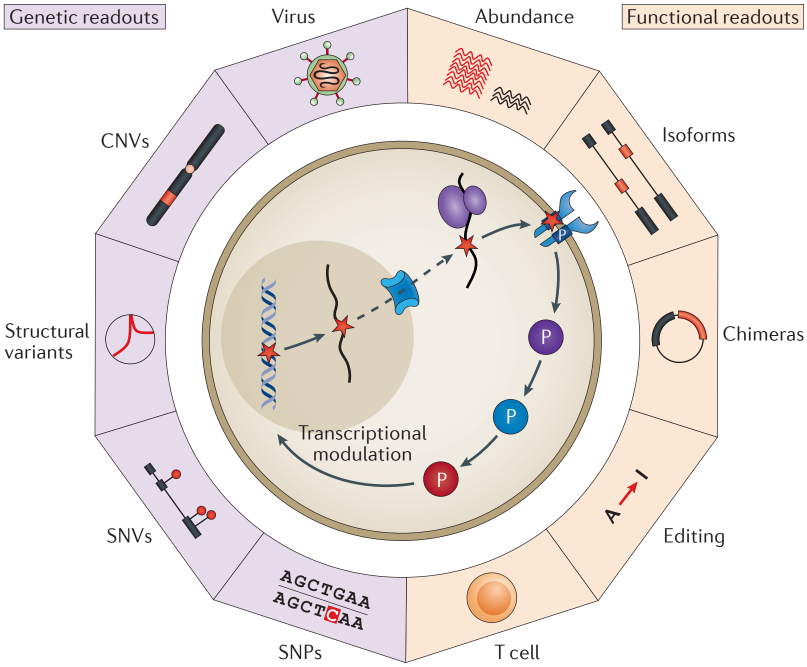


Figure 2 | **Transcriptome profiling for genetic causes and functional phenotypic readouts.** Transcriptome profiling enables researchers to link the genetic causes of cancer with their phenotypic consequences. The genotypes that can be interrogated by RNA sequencing include structural variants (for example, gene fusions), copy number variants (CNVs) (for example, amplifications), somatic mutations (for example, single nucleotide variants (SNVs)), germline variants (for example, single nucleotide polymorphisms (SNPs)) and the presence of viruses. The functional phenotypes that can be interrogated through transcriptome profiling are very broad and include quantitative estimates of expression levels and the detection of transcript isoforms, chimeric RNAs and RNA-editing sites. In addition, it is often possible to interrogate the downstream targets of genetic aberrations. In the illustrated scenario, a mutation in a receptor tyrosine kinase could be detected not only directly in the mRNA but also indirectly through a transcriptional signature resulting from the phosphorylation (P)-mediated activation of the mutant kinase downstream targets. Beyond tumour cell intrinsic features, transcriptome profiling can provide insights into the tumour microenvironment, for example, by characterizing transcripts from infiltrating T cells during an immune response.

Among those, alternative transcript initiation (ATI), or promoter usage, is particularly pertinent as it relates to epigenetic mechanisms of transcriptional regulation[97]. For example, the discovery of an oncogenic ATI for the ALK tyrosine kinase receptor demonstrated conclusively that epigenetic (non-genetic) aberrations can also drive cancer[98]. Similarly, differential expression of the ZAK (also known as MAP3K20) isoform TV1 has been associated with gastric cancer aggressiveness[99]. Beyond alternative isoforms, post-transcriptional mechanisms can fine-tune the stability and function of RNA[100]. Transcripts can either be edited (base change) or covalently linked to small molecules (base modification). Base changes can be detected directly from RNA-seq[101], whereas modifications require dedicated assays based on, for example, immunoprecipitation[102]. Overall, epitranscriptomics is an emerging field, and whether it plays a role in cancer is still unknown; however, unexpectedly, the RNA-editing enzyme APOBEC3B (also known as A3B)[103] has been shown to cause somatic hypermutation at the DNA level in a number of cancer types.

Epitranscriptomics
The study of biochemical modifications of RNA molecules.

Viral infections account for a substantial number of cancers worldwide[104], and their detection is critical for diagnosis and prognosis (for example, human papilloma virus (HPV) in cervical cancer). Viruses can be detected by a number of DNA-based or RNA-based methods[105]. In addition, the presence of chimeric human–viral RNAs can often point to their specific genomic integration sites[106]. Interestingly, RNA editing by the interferon-inducible double-stranded RNA-specific adenosine deaminase (DRADA; also known as ADAR1) is part of the innate response to virus infection. Illustratively, a single unbiased transcriptomic assay could detect the cause (exact viral mRNA), response (host enzyme expression) and consequence (viral RNA editing) of an antiviral response. Although most genetic differences, such as mutational status or the presence of a virus, are reflected in the transcriptional profile of a cancer cell, it is important to note that the discovery of such transcriptional signatures of genetic determinants typically requires a large cohort of samples owing to the heterogeneity among patients. Therefore, such associations are typically discovered retrospectively in the research setting. Additionally, as is the case for all omics-based assays, their clinical translation may become challenging because of the excessive cost, labour or sample requirements.

Detecting and interpreting genetic variants through transcriptomics. Genetic events that can be directly or indirectly detected by transcriptome profiling include SNVs[107–112], gene fusions and some structural variants and amplifications. The high coverage of RNA-seq allows both germline polymorphisms (such as single nucleotide polymorphisms (SNPs)) and somatic mutations to be called in genes with average-to-high expression levels[113,114], although the highest sensitivity and specificity are achieved by combining genomic and transcriptomic sequencing[115]. The most prominent application of transcriptome profiling beyond expression profiling is the detection of gene fusions[116]. This can be achieved indirectly from outlier gene expression[117] and differences in exon expression levels[118]. Direct evidence for the presence of gene fusions can be obtained from sequencing the chimeric RNA. This approach was first applied to ESTs[119] and later adapted to RNA-seq[18]. Contingent on the transcription around the genomic breakpoint, additional classes of structural variants can be detected that do not result in a chimeric protein. These include enhancer–promoter swaps (for example, *EVI1;* also known as *MECOM*)[120], amplicon-associated fusions (for example, *ERBB2;* also known as *HER2*)[121] or truncating fusions, which can result in loss of function for tumour suppressors (for example, *CDKN2A*)[122] or activation of oncogenes (for example, *PAX5*)[123].

Compared to variant calling from DNA, additional filters are necessary to mask RNA-editing sites and to deal with splicing-related artefacts. In particular, it is more challenging to detect small insertions or deletions (indels) than SNVs from RNA-seq data[124]. Although mutation calling from RNA is not as reliable as it is from DNA and, in general, will miss mutations in enhancers,

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

promoters or introns, it provides an inherent prioritization strategy for variants in coding regions. Coding variants that are not expressed, and hence not detected by RNA-seq, are likely to be passenger mutations. The simultaneous readout of expression levels and SNPs enabled research into allele-specific expression (ASE)[125], but not without technical challenges[126]. The major application of ASE is the study of gene imprinting[127] and epigenetic regulation[128]. Although ASE is not common in bulk normal tissues, single-cell studies revealed stochastic monoallelic expression in individual cells[129]. In the realm of cancer, ASE occurs predominantly as a consequence of copy number alterations[130], although it also occurs from loss-of-imprinting[131], and can contribute to cancer fitness. For example, the ratio of mutant to wild-type KRAS is associated with increased fitness and sensitivity to MAPK/ERK kinase (MEK) inhibitors[132].

Diverse RNA-seq protocols

Over the past several years, RNA-seq has been widely adopted by the scientific community and has become the *de facto* standard assay for many transcriptomic applications. RNA-seq is not a single protocol[46] but rather a family of related methodologies. Like most sequencing workflows, it involves sample preparation, sequencing and downstream computational analysis. This general framework can be adapted to accommodate a variety of biological questions, sample types and applications. However, the high flexibility comes at a cost of important practical considerations (BOX 1).

As for most transcriptomic protocols, the first step of RNA-seq is the disruption of cells and isolation of RNA. RNA-seq protocols have been adapted for a wide range of input materials, including cell cultures, body fluids and solid tissues. Particular challenges include RNA degradation and low input amounts[133]. A standard RNA-seq protocol requires that the sequenced RNA molecules are intact, and various strategies have been devised to achieve this. These new strategies enabled the transcriptomic profiling of clinically relevant samples, such as plasma or urine exosomes[134], platelets[135] and formalin fixed–paraffin embedded (FFPE) tumour tissues[50]. Library preparation strategies that involve multiple enzymatic reactions and purification steps are poorly suited to single-cell profiling. For single-cell[136] or low-input libraries, excessive PCR cycles can introduce biases and result in loss of complexity and information. This limitation has motivated the development of multiple single-cell RNA-seq (scRNA-seq) protocols, including CEL-seq, a clever method that replaces PCR with linear amplification[49]. As a complementary strategy, several protocols incorporate UMIs, which tag individual RNA molecules and detect PCR duplicates[48].

Although it is possible to achieve a very broad transcriptomic profile by using ‘total RNA-seq’, distinct protocols allow researchers to home in on specific RNA molecule types. Despite ribosomal RNA (rRNA) being the most abundant class of RNA molecules in a cell (up to 80%), it is of limited interest to researchers. Hence, depleting rRNA is often desirable in order to save sequencing bandwidth. A number of removal methods exist that are based on hybridization[137], duplex digestion[138] or not-so-random priming[139]. Depletion of rRNA without poly(A)-selection is necessary for the study of RNA molecules that cannot be easily enriched — such as non-polyadenylated non-coding RNAs, small nucleolar RNAs (snoRNAs), histone mRNAs and pre-mRNAs — and this approach is increasingly important as cancer transcriptomics extends beyond protein-coding gene expression. The use of RNA-seq for the detection of small RNA molecules, such as microRNAs (miRNAs), although possible[140], is marred with technical challenges. Small RNA-seq protocols require dedicated strategies to modify the native RNA termini. The efficiency of these steps is not uniform, and, ultimately, the measurements are better suited for relative abundances[141]. Similarly, targeted strategies are available for the enrichment of 5′ or 3′ ends of RNA molecules. The cap analysis of gene expression (CAGE) protocol[39], originally developed to identify and quantify 5′-capped RNAs, has been adapted to NGS platforms[142] and is particularly valuable for mapping transcription start sites (TSSs)[143]. Enrichment of 3′ ends is done predominantly for profiling gene expression and for mapping polyadenylation sites. These approaches[144] often build upon SAGE and, critical to clinical applications, can be more robust for RNA degradation than full-length transcript profiling[145].

Following RNA isolation and selection, the next steps of RNA-seq are fragmentation, cDNA synthesis and addition of sequencing adaptors. Depending on the protocol, fragmentation can be done at the RNA, single-stranded DNA or double-stranded DNA stage. RNA fragmentation is the easiest and most popular method as it does not require the use of enzymes. A major limitation of the original RNA-seq protocol was the loss of strand information during adaptor ligation following cDNA synthesis (that is, unstranded RNA-seq). A number of protocols have been developed to circumvent this issue (that is, strand-specific RNA-seq) by utilizing template-switching PCR (Peregrine)[146], deoxyuridine triphosphate (dUTP) labelling followed by enzymatic degradation[147] or end-specific RNA ligation[148]. The methods differ in terms of the required input amount, introduced biases and simplicity; therefore, the choice is very application-dependent. RNA-seq protocols can be modified once more at the cDNA library stage. Protocols have been proposed to capture and enrich specific sequences[149] or the whole exome[50,51]. Capture RNA-seq is an alternative to enrichment of poly(A) transcripts or depletion of rRNA that does not depend on intact RNA. Alternatively, depletion of sequences can be done at the single-stranded cDNA stage[150]. Selection of poly(A) mRNAs using oligo(dT) beads is currently the most popular protocol in cancer transcriptomics. Unfortunately, this approach requires largely intact RNA or is otherwise affected by technical biases or artefacts. Protocols that utilize capture[151], depletion or hybridization[152] are therefore more suitable for clinical use, where RNA obtained from frozen or FFPE tissue is generally of variable quality.

Passenger mutations
Mutations that have no measurable effect on the growth of a clone.

Allele-specific expression
(ASE). The analysis of differences in the expression from both alleles, that is, expression variation between the two haplotypes. Also known as allelic imbalance.

Cap analysis of gene expression
(CAGE). A molecular technique to sequence the 5′ end of transcripts.

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Box 1 | **Practical considerations for clinical RNA sequencing**

Can I use a gene expression signature based on prior technology?
Although most transcriptomic platforms are highly reproducible by themselves, reproducibility across platforms is limited. Unfortunately, the biggest challenge is in the measurement of absolute expression levels[263], which is the input to many biomarkers and signatures. Therefore, signatures cannot be expected to translate verbatim between platforms. Few studies have explored this topic. Fumagalli *et al.*[264] concluded that single-gene expression biomarkers and established prognostic signatures generalize well between microarrays and RNA sequencing (RNA-seq). Zhang *et al.*[265] reached a similar conclusion in that "technological platforms (RNA-seq versus microarrays) […] do not significantly affect performances of the [predictive] models." However, these studies were done on high-quality samples and did not explore whether RNA degradation or crosslinking had a detrimental effect.

Which RNA-seq protocol should I choose?
The choice of the optimal RNA-seq protocol will strongly depend on the quality and quantity of input material[50,133]. If large quantities of intact RNA can be extracted (for example, from flash-frozen tissue sections), most protocols will produce high-quality data. In that case, use of RNA-seq protocols based on poly(A)+ selection is recommended, as they will provide the best interoperability with existing resources (for example, The Cancer Genome Atlas (TCGA) and the Genotype–Tissue Expression (GTEx) project). Furthermore, these RNA-seq protocols can preserve strandedness (that is, strand-specific RNA-seq) of the library by use of the popular deoxyuridine triphosphate (dUTP) method[147] and facilitate certain clinical applications (for example, fusion detection). If RNA integrity is compromised or poly(A) selection is not possible, the popular alternatives are ribosomal RNA (rRNA) removal and cDNA capture. Capture RNA-seq is remarkably accurate and robust, but it is expensive owing to the use of capture probes. For low-input samples, unstranded protocols utilizing oligo(dT)-priming (for example, SMART-seq2) show better performance in general.

How deeply should I sequence?
The depth to which a cDNA library should be sequenced depends mostly on the application and is limited by the choice of sequencing platform. In general, if RNA-seq is used for the detection of genetic events, such as mutations or fusions, higher sequencing depth is warranted. On the other hand, the use of RNA-seq for transcriptional profiling requires only moderate amounts of sequencing (saturation at 15 million reads), and including more replicates is a substantially better strategy than more reads[266]. However, much higher read depths (100 million paired-end reads or more) are required for the study of alternative splicing or allele-specific expression.

Beyond RNA expression and sequence. The fundamentals of RNA-seq can be effectively extended to measure various aspects of RNA structure, function and biology. These advanced methods are complex and often require difficult standardization. Their utility is therefore largely limited to highly specialized laboratories, and paths to clinical use have not yet been established. RNA synthesis and degradation can be probed more directly using global run-on sequencing (GRO-seq)[153], bromouridine sequencing (Bru-seq) and 4-thiouridine sequencing (4sU-seq). These assays are important as they expand our understanding of RNA dynamics beyond the steady state and can provide mechanistic insights into, for example, oncogenic transcription factors[154] or drugs that target the epigenome[155]. Structural and conformational features of RNAs can be probed using a variety of techniques, such as parallel analysis of RNA structure (PARS)[156], *in vivo* click selective 2′-hydroxyl acylation and profiling experiment (icSHAPE)[157] or RNA G-quadruplex sequencing (rG4-seq)[158]. For example, in icSHAPE, molecular probes are preferentially attached to structurally flexible RNA fragments. Because stable RNA conformations and sequence motifs enable specific molecular interactions (which can affect RNA biology and functions), there is great interest in unbiased methods to map them. To date, several protocols have been proposed to detect all major types of interactions, including RNA–DNA[159], RNA–protein[160] and RNA–RNA[161] binding. Knowledge of RNA structure and interactions is expected to culminate in the development of small molecule drugs[162]. In particular, the structure-guided disruption of oncogenic non-coding RNAs may become a strategy for targeting oncogenic long non-coding RNAs (lncRNAs) in the future. Furthermore, the utility of RNA-seq extends beyond transcriptomics. For example, Arnold *et al.* have proposed a direct and quantitative self-reporter assay to study the functional regulatory activity of DNA[163].

Computational tools for cancer transcriptomics

With the availability of large, annotated compendia of gene expression profiles across normal tissues (GTEx and HPA), tumour tissues (TCGA and the International Cancer Genome Consortium (ICGC)) and cell lines (ENCODE and Genentech), we are beginning to understand the structure of global gene expression. However, the emerging complexity and size of the combined genetic and functional tumour molecular profiles pose great analytical challenges and opportunities (FIG. 3). Thus, the premise of using transcriptomics to elucidate cancer phenotypes is contingent upon advances in bioinformatics and computational biology. An example toolbox for the interrogation of cancer transcriptomes is listed in TABLE 1.

Types of gene expression analyses. Transcriptome-wide gene expression profiles are now available for the majority of cancer types and their corresponding tissues of origin. In general terms, there are two cancer-centric paths to analyse these data: the differential approach, which interprets tumour expression profiles relative to the patient-matched or unmatched normal tissue samples; and the relative approach, which compares transcript levels across tumours or other samples (FIG. 3). Inherently, these strategies have unique advantages and applications. Differential analyses are designed to detect cancer-specific changes, but if the normal samples are not comparable[164], the results will be difficult to interpret, for example, if the cancer cell of origin is rare or unknown. In general terms, differential analyses tend to be underpowered in the clinical setting. Comparisons at the single-patient level are often limited by the dearth of replicates due to cost and sample availability, while at the cohort level, they are often confounded by interpatient heterogeneity. Relative analyses are useful to characterize individual samples but typically depend on the availability of external knowledge or reference data sets. The validity of any relative comparison is contingent on how well a query sample is matched to the reference in terms of technical (for example, type of data processing) and biological (for example, molecular subtype) biases. Therefore, relative analyses often necessitate advanced normalization techniques[165] and batch correction[166]. Overall, the differential approach is more common in

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

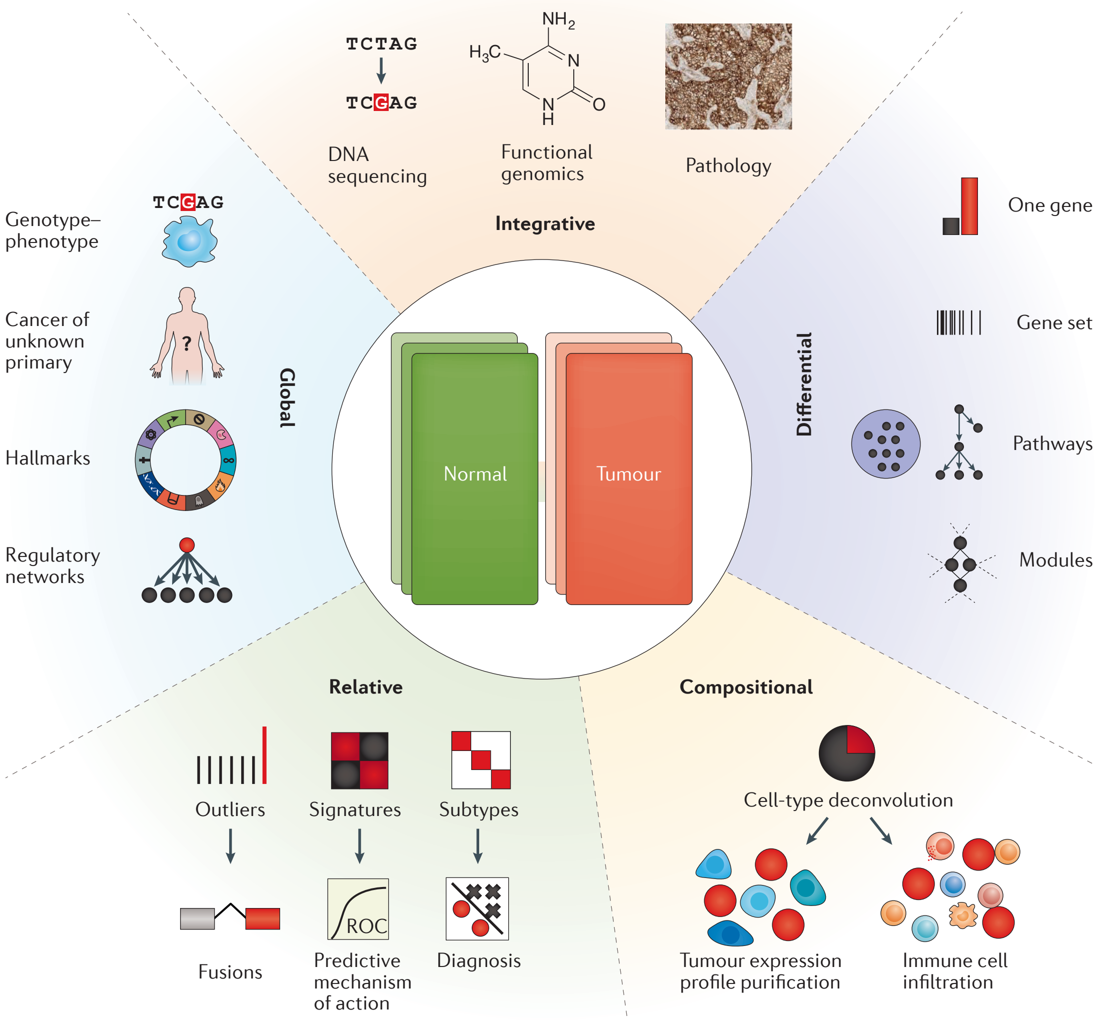


Figure 3 | **Tumour phenotypes beyond differential expression.** Analyses of transcriptomic data fall into five broad categories. Differential analyses focus on the differences between tumour and normal tissues at the gene, gene set, pathway or network level; they require at least two groups of paired or unpaired samples. Relative analyses compare a single sample or a group of samples with the whole cohort and attempt to identify transcriptional outliers that are clinically useful signatures or subtypes. Compositional analyses leverage the gene expression signatures of different cell types to assess (or control for) tumour cell purity, to deconvolute samples into constituent tumour and non-tumour cell types and to characterize immune infiltration. Global analyses compare a sample to a large reference compendium (often pan-tissue or pan-cancer) in order to characterize broad transcriptomic features, such as the accretion of cancer hallmarks, primary tissue type (if unknown) or genotype–phenotype relationships. Integrative analyses attempt to supplement transcriptomic data with other data, such as DNA sequencing, functional genomics (for example, DNA CpG methylation) or clinical data (for example, pathology).

the research setting to generate hypotheses, whereas the relative approach drives many clinical applications, such as precision medicine.

Differential approaches. The simplest type of differential analysis is the identification of genes that are upregulated or downregulated in cancer (that is, differentially expressed genes (DEGs)), and established methods to detect DEGs are available for both microarray[167] and RNA-seq data[168–170]. A typical result is a long list of DEGs that is difficult to interpret without additional functional annotation, as demonstrated by a landmark study in breast cancer[171]. Differential methods have also been proposed for splicing[93] or isoform usage[172]. Although transcriptomes have very high dimensionality, there is also substantial correlation among the genes, which can be leveraged to simplify or summarize the data[173]. A common strategy is to break down the transcriptome-wide gene expression profile into a set of modules that are less interdependent, more generalizable and simpler to understand. The specifics for each method differ substantially, but in general, it is possible to test for differential gene sets[174], pathways, gene regulatory networks or modules in co-expression networks[175]. Ideally, testing multiple related genes will improve sensitivity and yield results that are easier to understand.

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Table 1 | **The cancer transcriptomic toolbox**

Resource	Description	URL	Refs
Annotation			
RefSeq	Curated reference sequence database (transcriptome-centric, that is, defined by transcript sequence)	https://www.ncbi.nlm.nih.gov/refseq/	75
GENCODE	Curated reference gene annotation (genome-centric, that is, defined by alignment to reference genome)	http://www.gencodegenes.org/	272
MiTranscriptome	Automated reference transcriptome based on sequence assembly, includes long non-coding RNAs	http://mitranscriptome.org/	76
Reference data			
MSigDB	Collection of experimental and curated gene sets (signatures)	http://software.broadinstitute.org/gsea/msigdb	179
Human Protein Atlas	Compendium of proteomic and transcriptomic data in diverse normal tissues	http://www.proteinatlas.org/	63
CCLE	Genomic and transcriptomic data on hundreds of cancer cell lines	https://portals.broadinstitute.org/ccle/home	60
GTEx	Transcriptomic data (RNA-seq) from normal human tissues from a large number of individuals	https://gtexportal.org/home/	62
Mitelman	Database of gene fusions and chromosomal aberrations	https://cgap.nci.nih.gov/Chromosomes/Mitelman	288
COSMIC	Catalogue of somatic mutations in cancer patients and cell lines, including gene fusions	http://cancer.sanger.ac.uk/cosmic/classic#fus	289
Tool			
QoRTs	Comprehensive collection of RNA-seq quality control functions	http://hartleys.github.io/QoRTs/index.html	290
STAR	Fast and accurate splice-aware sequence aligner	https://github.com/alexdobin/STAR	282
featureCounts	Fast read counting for gene-level or exon-level expression estimates	http://bioinf.wehi.edu.au/featureCounts/	291
Kallisto	Pseudo-alignment-based quantification at the transcript level	https://pachterlab.github.io/kallisto/	292
EdgeR	Differential expression using the negative binomial distribution (see also DESeq2)	http://bioconductor.org/packages/release/bioc/html/edgeR.html	170
Limma	Flexible linear modelling and empirical Bayes moderation to assess differential expression by use of precision weights for RNA-seq data (Voom)	http://bioconductor.org/packages/release/bioc/html/limma.html	167, 168
CIBERSORT	*In silico* transcriptome deconvolution into relative abundances of different immune cell types	https://cibersort.stanford.edu/	260
MiXCR	T cell and B cell CDR3 sequences assembler; enables repertoire profiling from RNA-seq data	https://milaboratory.com/software/mixcr/	261
GSEA	Gene set enrichment analysis	http://www.broad.mit.edu/GSEA	273
PARADIGM	Computational tool for the inference of patient-specific pathway activities	https://sbenz.github.io/Paradigm	186
FusionCatcher	A sensitive and specific tool for the detection of gene fusions	https://github.com/ndaniel/fusioncatcher	293
TopHat-Fusion	A very sensitive tool for the detection of gene fusions	http://ccb.jhu.edu/software/tophat/fusion_index.shtml	294
Analysis			
Oncomine	Web application for user-friendly analysis and exploration of cancer transcriptomes	https://www.oncomine.org/resource/login.html	180
Xena	UCSC Xena: versatile genomic data mining and analysis portal	https://xenabrowser.net/	287
Data warehouse			
ENCODE	Repository of diverse functional genomics data, including RNA-seq, from the ENCODE project	https://www.encodeproject.org/	59
GDC	Genomic Data Commons: provides access to raw and harmonized data for multiple genomic projects, including RNA-seq data processed using a standard pipeline	https://portal.gdc.cancer.gov/	295
FANTOM5	Repository of CAGE data from the FANTOM5 project	http://fantom.gsc.riken.jp/5/	271
ArrayExpress	Standard repositories of functional genomic and transcriptome profiling data	http://www.ebi.ac.uk/arrayexpress/	78
GEO		https://www.ncbi.nlm.nih.gov/geo/	77

CAGE, cap analysis of gene expression; CCLE, Cancer Cell Line Encyclopedia; ENCODE, Encyclopedia of DNA Elements; FANTOM5, Functional Annotation of the Mammalian Genome 5; GENCODE, the genome annotation project of ENCODE; GEO, Gene Expression Omnibus; GTEx, Genotype–Tissue Expression Project; Limma, Linear Models for Microarray Data; MSigDB, Molecular Signatures Database; PARADIGM, Pathway Recognition Algorithm using Data Integration on Genomic Models; QoRTs, Quality of RNA-seq Toolset; RNA-seq, RNA sequencing; STAR, Spliced Transcripts Alignment to a Reference; UCSC, University of California, Santa Cruz.

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

For example, using a simple gene set method, Majeti *et al.*[176] were able to identify the dysregulation of the WNT pathway in acute myeloid leukaemia. Beyond upregulation or downregulation, methods have been developed to detect less-uniform changes in gene expression[177]; for example, detecting mechanism of action by network dysregulation (DeMAND) leverages changes in correlation to prioritize dysregulated or 'rewired' modules[178].

Relative approaches. In contrast to the differential approach, which aims to identify common features of a set of samples, the purpose of the relative approach is to identify distinct aberrations in an individual tumour. Although, in general, traits such as cancer subtype are derived from a single expression profile, the computation often requires external data, and the interpretation is relative to other samples. A simple example is the identification of outlier genes that are highly expressed in some, but not all, samples[117]. This concept can be extended to gene sets or signatures[179,180]. Analogous to differential analysis, the gene sets can be based on experimental data[181], domain-specific knowledge[182] or even clinical research[183]. This strategy was applied by Saal *et al.*[184] to first define a signature associated with the loss of *PTEN* and to show that a sample-specific signature score predicts outcomes across multiple cancer types. Dedicated computational methods, such as Gene Set Variation Analysis (GSVA)[185] or Pathway Recognition Algorithm using Data Integration on Genomic Models (PARADIGM)[186], enable comprehensive signature analyses across large numbers of samples and signatures. The application of these methods is a form of projection; expression levels of tens of thousands of genes are conveniently summarized by numeric scores for hundreds of signatures that reflect biologically relevant dimensions in the data.

Relative expression signatures and cancer subtypes. Whereas gene sets used in differential analyses typically comprise functionally related genes, such as pathways, relative signatures are sometimes designed to be less redundant in order to integrate multiple aspects of tumour biology at the same time. Somewhat arbitrarily, if a signature captures a large part of the variation in gene expression across tumours of the same type, it can be used to define and identify molecular subtypes of that cancer, for example, for posterior fossa ependymoma, a tumour type with no recurrent somatic mutations[187] and a striking epigenetic phenotype[188]. Transcriptional profiling was paramount in identifying two subtypes (A and B) that not only are delineated in terms of their pathobiology but are also associated with clinically relevant differences in outcome[187]. One of the earliest examples was the discovery of molecular classes of diffuse large-B cell lymphoma (DLBCL)[189] through microarray profiling. The two discovered subtypes of the 'germinal centre' and 'activated' DLBCL originate from different stages of B cell maturation, have distinct genetic underpinnings (most notably, *IGH–BCL2* fusions are exclusive to the germinal centre subtype) and partially explain the clinical heterogeneity of the disease (activated DLBCL has significantly worse overall survival). RNA-seq profiling of Philadelphia chromosome-like acute lymphoblastic leukaemia (ALL) identified its phenotypic similarity to *IKZF1*-deleted ALL and many actionable kinase fusions[190]. Similarly, the prediction analysis of microarray 50 (PAM50) signature, in addition to other panels such as MammaPrint[191] and Oncotype DX[192], has been developed to classify breast cancer into molecular subtypes[193]. The four intrinsic subtypes of breast cancer (luminal A, luminal B, ERBB2-enriched and basal-like) were shown to be independently associated with clinical outcomes and harbour a different set of genetic aberrations. The PAM50 expression test is among the most widely known and clinically successful cancer diagnostics. Although signatures based on smaller gene sets are more easily interpreted, reducing the number of analysed genes is not always necessary[194]. Some applications, such as unsupervised clustering[195] or modern supervised machine-learning methods, may perform best on rather large expression profiles. For example, the four major subtypes of glioblastoma were initially found from the expression levels of 1,740 genes[196].

PAM50
Prediction analysis of microarray 50. A gene expression signature to classify breast cancer into intrinsic subtypes.

Cellular composition and microenvironment. The study of the heterogeneous cellular composition of tumours is one of the most recent applications of cancer transcriptomics. Approaches typically involve either directly isolating and characterizing individual cells (using, for example, single-cell sequencing) or indirectly inferring cell compositions *in silico* from bulk expression data. From bulk expression data (which are currently more readily available from clinical samples than are single-cell data), the computational task is often referred to as sorting, or deconvolving, the gene expression profile. Deconvolution is a difficult problem that requires methodological constraints in order to converge on plausible solutions. A large number of algorithms have been proposed[197] that make different trade-offs on the basis of the available data and the desired output. In general, the methods can be divided into those that use cell-type-specific gene signatures and can be applied to a single tumour sample and those that require multiple tumour and normal samples (matched or unmatched). Currently, the most important applications are to estimate tumour clonality and purity, which are affected by intrinsic tumour cell heterogeneity or infiltration by stromal or immune cells[198]. For example, *in silico* purification of gene expression profiles has been applied to improve their performance in prognosis[199] and classification[200]. Deconvolution also provides a unique opportunity to study the tumour microenvironment, for example, to unravel tumour–stromal paracrine crosstalk[201]. In the future, single-cell transcriptomics is bound to revolutionize our understanding of the tumour microenvironment[202], heterogeneity[203] and evolution[204].

Integrative and global analyses. The true utility of transcriptomics is revealed when combined with additional DNA-based assays. In the context of clinical sequencing, the most immediate need is the prioritization of genes

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

within somatically focally amplified regions (amplicons). Expression profiling serves as an important readout to interpret the mechanistic role of these and other copy number aberrations (CNAs). As amplicons often contain multiple genes of interest, it is necessary to use additional data, such as expression levels, to pinpoint functionally important genes[205]. In the discovery setting, recurrent somatic amplification accompanied by outlier expression levels is a key characteristic of many cancer driver mutations and serves as a powerful criterion to nominate candidate oncogenes[206]. The combined use of copy number and transcriptomic data has been shown to yield the strongest predictor of outcome in breast cancer[207], illustrating the added value of data integration. A correlative approach can be used to link genetic variation with the expression of individual genes[208]. This helped in attempts to elucidate the biological mechanisms behind intergenic cancer risk loci[209,210]. Similarly, correlative analyses can point to the transcriptomic consequences of somatic mutations[211]. Beyond DNA, transcriptomic data can be integrated with many other types of omic data, such as proteomics, functional genomics, networks or phenotypic screens (reviewed in REF. 212).

Analyses that compare samples have been carried out successfully to study a single cancer type but can also be done globally across tumour types. One of the earliest applications was the prediction of tumour type using genome-wide expression patterns[213] and the identification of a pan-cancer metastasis signature[90]. Since then, a number of studies have explored cancer phenotypes that generalize across multiple primary sites. Ambitious attempts have also been made to define a single molecular taxonomy across many cancer types[214] and to identify universal genes associated either with cancer[215] or prognosis[216]. These studies confirmed that cancer is a heterogeneous disease both within and across tissues of origin. Enabled by the breadth of the available RNA-seq data, a number of pan-cancer studies have comprehensively characterized the landscapes of viral expression and integration[217], kinase fusions[218], polyadenylation[219] and gene dosage sensitivity[220], among others.

From transcriptomics to precision oncology

As illustrated in the preceding sections, tumour transcriptomes are remarkably useful for the interrogation of cancer phenotypes. Transcriptomic profiling goes beyond what can be learned from genetic testing, such as DNA sequencing or array comparative genomic hybridization (aCGH), alone. The clinical utility of RNA-seq has been demonstrated by a number of sequencing programmes where RNA-seq identified a large number of actionable genetic events[221–223]. Still, targeted DNA sequencing is currently the method of choice for many clinical applications in precision oncology. DNA is a highly stable analyte and is therefore well suited for molecular diagnostics. Genetic assays are set up to reliably detect highly actionable events, such as driver mutations, that guide patient therapy. Although results from the largest DNA panels, such as Oncoseq1500 (A.M.C *et al.*, unpublished observation) or the MSK-IMPACT test[224], are sufficient to guide the majority of patients towards US Food and Drug Administration (FDA)-approved drugs or clinical trials, many patients fail to respond to therapy or are affected by considerable side effects. Owing to the inherent limitations of DNA-based testing, a number of clinical needs remain underserved or rely on labour-intensive and low-throughput molecular techniques.

Limitations of DNA-based assays. Important limitations of DNA-based assays include the following. First, for cancers with heterogeneous progression, genetic tests often fail to identify aggressive disease[225]. Second, structural variants, such as receptor tyrosine kinase fusions, which are among the most actionable and clinically relevant classes of genetic aberrations, are largely undetected in targeted assays[226]. Third, in many cases, genetic aberrations are insufficient to predict a response to chemotherapy[227] or immunotherapy[228]. Fourth, mutation calling at the level of individual tumour cells remains challenging[229]. Fifth, DNA-based assays cannot provide detailed phenotypic characterization of the tumour microenvironment or immune responses. Finally, most genetic aberrations are not specific to a single cancer type and do not help in the diagnosis of many carcinomas of unknown primary (CUP) origin[230]. Hence, in order to improve patient care, there is a need for additional diagnostics to more fully characterize tumours. RNA-seq as a robust, high-throughput and affordable transcriptomic platform is uniquely positioned to fill many of those needs. Although several groups are working to develop predictive biomarker panels[231], such as initiating large-scale longitudinal trials that track the evolution of tumour genomes and transcriptomes (tracking cancer evolution through therapy (TRACERx)[232] and adaptive patient-oriented longitudinal learning and optimization (APOLLO)[233]) or developing machine-learning algorithms to classify CUPs on the basis of expression[234], the great potential of RNA-seq has yet to be fully realized. In the following sections, we try to highlight the main challenges, applications and opportunities of using RNA-based assays in precision oncology.

RNA as a diagnostic analyte. The major challenge of using RNA as a diagnostic analyte is the limited stability of RNA, which leads to its rapid fragmentation. RNA degradation negatively affects many quantitative assays, including RT-qPCR and RNA-seq. Rapid biospecimen-handling techniques, such as flash freezing, are necessary to preserve intact RNA. The fairly high reactivity of RNA results in extensive crosslinking with formaldehyde, the most common fixative for tumour tissues, which substantially diminishes hybridization efficiency and PCR amplification. As a result, diligent quality control (QC) of RNA integrity[235] is necessary to develop reliable assays. Recently, methods have been developed to overcome the limitations of degraded RNA, including a strategy to reverse adducts using organocatalytic chemistry[236] and the introduction of hybrid capture RNA-seq[50]. In hybrid capture RNA-seq, exome capture using RNA probes is introduced at the cDNA stage of a total RNA library. This achieves rRNA

Driver mutations
Mutations that provide the cancer with a strong selective advantage, that is, mutations that result in the clonal growth of mutant cells.

Clinical utility
Whether a test has a substantial effect on the diagnosis, prognosis or treatment of a patient.

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

depletion without depending on intact RNA and focuses sequencing bandwidth on the coding portion of the transcriptome while preserving quantitative expression levels[50,51].

The majority of blood cell-free RNA (cfRNA) comes from apoptotic and necrotic cells and can be elevated in diseases, including cancer. Therefore, the analytical use of cfRNA critically depends on the *in vitro* stability of isolated cells and requires dedicated sample handling and preservation protocols[237]. If the isolated cells become unstable, cfRNA becomes diluted by the cytosolic RNA from normal cells. Importantly, some types of RNA are particularly well suited for developing RNA-based diagnostics. Circular RNAs are resistant to exonucleases and enriched in platelets[238] and exosomes[239]. Tumour-associated miRNAs are strongly bound by proteins (for example, Argonaute 2 (AGO2)), which are believed to protect them from degradation by blood RNases[240]. Finally, lncRNAs are remarkably cancer-specific and sometimes expressed at very high levels[76].

Robust and sensitive assays based on RNA-seq herald future clinical uses. One of the primary uses of cancer transcriptomics is the development of RNA-based biomarkers (FIG. 4). Similarly to most cancer diagnostic methods, the analyte is obtained directly from invasive core biopsies or tumour resections. Both primary and metastatic tumours can be examined, with the latter posing additional analytical challenges. Among other complications, metastatic transcriptomes are derived from limiting amounts of material from needle-core biopsies, have a lower tumour content and are confounded by biopsy-site tissue with a distinct expression profile (for example, liver)[223].

RNA-based fusion detection was one of the first applications successfully translated into routine clinical diagnostics (that is, the FoundationOne Heme test). The presence of chimeric mRNA can be detected very sensitively using RT-qPCR, upon which a binary 'call' is made. This targeted approach can be applied to detect recurrent fusions with known breakpoints. For example, the FoundationOne Heme test uses RNA-seq to detect recurrent gene fusions in haematological cancers, for example, *IGH–MMSET* (*MMSET* is also known as *NSD2* and *WHSC1*) in multiple myeloma. In addition, RNA-seq remains the only cost-effective and unbiased method to detect gene fusions.

With the increased focus on circulating tumour cells, it has become increasingly important to reliably detect somatic mutations from limiting amounts of DNA[229] or RNA. Recently, progress has been made in the use of liquid biopsy samples (for example, blood) or non-invasive body fluids (for example, urine) as sources of diagnostic material. Tumour RNA has been isolated from circulating tumour cells, tumour-educated platelets[135] and exosomes[241], but it can also be found as cfRNA[242]. The isolated RNA can be used in qualitative and quantitative assays. Although potentially affected by allelic dropout, scRNA-seq has been shown to have a sensitivity and specificity that approach those of multiplex PCR[243]. For applications relying on single-cell omics, such as the monitoring of cancer progression, RNA may become the preferred analyte owing to the natural amplification (that is, transcription from two genomic copies of DNA per cell that often result in hundreds or thousands of corresponding RNA molecules). The accuracy of variant calling can be improved by combining DNA and RNA sequencing data[115]. For example, splice-site mutations and large indels can be validated by observing their consequences, which include exon-skipping, aberrant splicing patterns or exon losses[244], whereas structural variants, such as gene fusions, often result in chimeric transcripts that encode putative peptide antigens[245].

In the clinical setting, elevated expression levels are necessary to establish a rationale for the use of targeted therapeutics, such as the ERBB2-targeted monoclonal antibody trastuzumab for ERBB2-positive breast cancer or small-molecule inhibitors of hepatocyte growth factor receptor (HGF receptor; also known as MET) in non-small-cell lung cancer. It is also possible to detect smaller copy number changes by shifts in median expression levels[246]. Strikingly, this can be done even at the single-cell level[202]. Beyond genetic events, transcriptome profiling is essential for identifying protein targets for T cell receptors (TCRs) or chimeric antigen receptors (CARs), for example, the NY-ESO-1 antigen in melanoma[247].

Prognostic and predictive gene expression signatures. Over the past 20 years, gene expression profiling has been repeatedly leveraged to identify clinically useful signatures. These signatures can be developed into biomarkers if their analytical validity and clinical validity are firmly established. Potential applications of biomarkers in clinical oncology span the entire course of the disease. Specifically, biomarkers can be used for screening and early cancer detection. Diagnostic tests can help in determining the primary tissue of the cancer or identifying the disease subtype. Prognostic and predictive biomarkers can be used to assess patient risk and response to drugs and, thereby, to influence therapy selection. During the course of therapy, indicators can be used to detect early response or toxicity, which can trigger a change in treatment before severe side effects or substantial disease progression occur. Finally, sensitive tests can be used to detect disease recurrence before the presentation of other symptoms[248,249].

Many of these ideas have been commercialized and clinically validated. For example, prognostic panels are now available and are clinically used for all major cancer types, including breast (MammaPrint, Oncotype DX and Prosigna), lung (GeneFx), prostate (Prolaris) and colon (ColoPrint). Because most RNA-based biomarkers comprise multiple genes, dedicated assays have been developed for each panel, relying mostly on RT-qPCR. However, with the rapid decline in the costs of whole-transcriptome sequencing, a strategy of embedding multiple panels within a single assay has become viable. Comprehensive upfront profiling may be particularly advantageous for areas where the signatures are not yet established and for retrospective clinical trials.

Allelic dropout
When a sample is sequenced and one or more alleles are not detected.

Analytical validity
The ability to accurately detect and measure the biomarker of interest.

Clinical validity
The clinical performance of a test, that is, how well the test is able to identify the clinical variable of interest (for example, disease status).

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

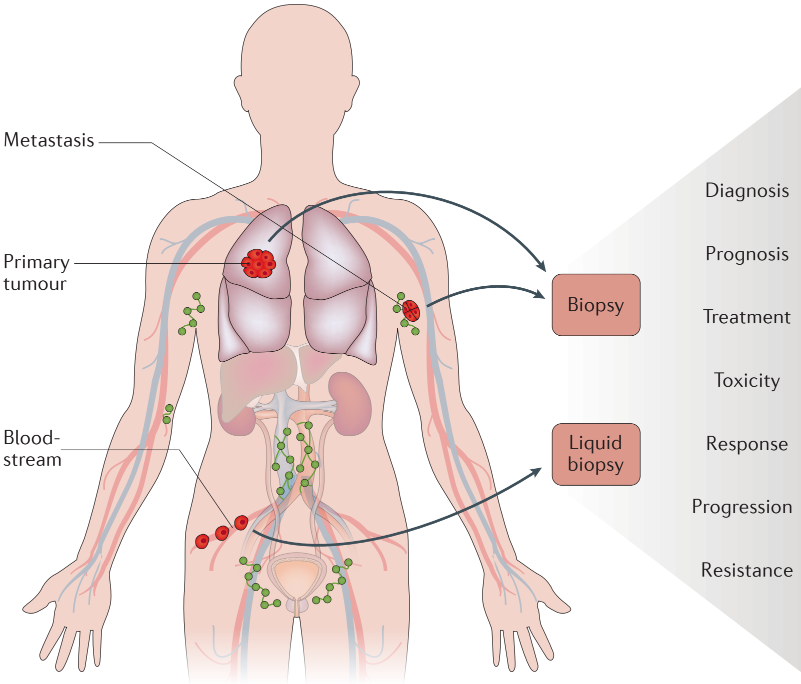


Figure 4 | **Paths to clinical translation for RNA-based assays.** For the development of biomarkers, RNA-based assays can be based on either tissue or liquid biopsies. Depending on the clinical application, RNA profiling of tissue biopsies can involve either the primary or metastatic sites. Liquid biopsies are most often blood-based but can include other fluids in select cancers (for example, urine in prostate cancer). RNA-based biomarkers are being developed for all aspects of cancer medicine, from initial diagnosis and staging (for example, prognostic biomarkers) through treatment (for example, predictive and pharmacodynamic biomarkers) to the detection of disease progression and resistance. RNA-based profiling is particularly valuable when the underlying biological mechanism is epigenetic (for example, expression of immune checkpoints as a predictive biomarker to immune checkpoint therapy).

Transcriptomics in immuno-oncology. The need for RNA-based companion diagnostics is particularly acute in immuno-oncology[250]. Cancer immune phenotypes were shown to broadly reflect the activity of the host immune system and to generalize remarkably well across cancer types. Numerous studies have investigated the association of immune infiltration with survival[251–253] and found significant correlations at the level of immune cell types, inflammation signatures and individual genes. Although immune checkpoint inhibitors are broadly beneficial across cancer types, the response rates are highly variable. It is becoming increasingly clear that positive responses to immunotherapy are associated with tumour immunogenicity and host immune infiltration[253–255]. However, given the complexity of adaptive immune responses and the dynamic nature of tumour–immune evasion, it is unrealistic to expect that a single gene will be sufficient to accurately predict outcomes or guide treatment.

The clinical utility of transcriptome profiling for immunotherapy was demonstrated in a landmark longitudinal study that demonstrated that signatures of adaptive immunity are predictive of response to immune checkpoint blockade[254]. As both prognostic and predictive approaches require the expression levels of hundreds of genes, their clinical translation will depend on the routine use of whole-transcriptome profiling or custom-targeted panels[256]. We have shown that comprehensive immunophenotypic data can be obtained from clinical transcriptomes and that they provide unique insights into the immunological heterogeneity of metastatic tumours across all major primary tissue types[223]. RNA-seq data are also particularly valuable for the development of personalized cancer vaccines[257,258], where they can be used to identify chimeric fusion proteins that contain putative mutant epitopes[245] and help in the selection of potentially highly abundant neoantigens.

The complexity of tumour–immune cell interactions is mirrored by the diversity of bioinformatics approaches to characterize them. Both data-driven[198] and knowledge-driven[253] approaches have been proposed to quantify the overall level of tumour–immune infiltration. In addition, recent methodological advances made it possible to estimate cell-type fractions from bulk tumour expression profiles in a process referred to as *in silico* cell sorting[259,260], which is similar to the 'purification' of the tumour cell expression profiles discussed above[199,200]. Finally, clonal expansion of antitumour T cells can be detected by the presence of somatically rearranged TCR sequences, that is, clonotypes[261]. An analogous strategy can be applied to B cells and immunoglobulin loci[262]. As neoantigen prediction remains a daunting problem, RNA-seq data are useful for both the detection of protein-altering genetic aberrations and their prioritization based on expression levels.

Conclusions and future perspectives

Although DNA-based assays remain the primary means of detecting genetic aberrations driving cancer, the unique readouts from sequencing RNA warrant the adoption of RNA-based cancer diagnostics in precision oncology. Constant innovation in transcriptome profiling has greatly expanded our understanding of cancer but has also transformed cancer research into one of the first data-intensive fields of biology. Methodological advances continue to remove technical barriers that limit the spatial, temporal or molecular resolution of RNA profiling, whereas decreases in sequencing cost have made routine high-throughput sequencing affordable. With the recent introduction of massively parallel scRNA-seq, we expect to see, once again, an exponential increase in the amount of data. The future success of cancer transcriptomics will be measured by how well we can turn those volumes of data into new cancer drugs and molecular diagnostics. This, in turn, will depend on our ability to identify the relevant cancer phenotypes and dissect them into concise and testable regulatory networks. Overall, the success of RNA-based diagnostics will depend on the rational choice of target RNA, continued improvements in tissue handling and RNA processing and the development and validation of computational methods.

Neoantigens
Antigens, herein short peptides, not previously recognized by the immune system. They can be formed by somatic mutations during tumorigenesis.

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

1. Velculescu, V. E. *et al.* Characterization of the yeast transcriptome. *Cell* **88**, 243–251 (1997).
2. Carninci, P. *et al.* The transcriptional landscape of the mammalian genome. *Science* **309**, 1559–1563 (2005).
This is the first study to show the transcriptional complexity of a mammalian genome.
3. Frye, M., Jaffrey, S. R., Pan, T., Rechavi, G. & Suzuki, T. RNA modifications: what have we learned and where are we headed? *Nat. Rev. Genet.* **17**, 365–372 (2016).
4. Johnson, J. M. *et al.* Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* **302**, 2141–2144 (2003).
5. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
6. Shoemaker, D. D. *et al.* Experimental annotation of the human genome using microarray technology. *Nature* **409**, 922–927 (2001).
7. Hughes, T. R. *et al.* Functional discovery via a compendium of expression profiles. *Cell* **102**, 109–126 (2000).
8. Byron, S. A., Van Keuren-Jensen, K. R., Engelthaler, D. M., Carpten, J. D. & Craig, D. W. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nat. Rev. Genet.* **17**, 257–271 (2016).
This is an excellent and complementary Review on the clinical applications of RNA-seq.
9. Chang, J. C. *et al.* Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. *Lancet* **362**, 362–369 (2003).
This study demonstrates the feasibility of predicting the therapeutic response from microarray data obtained from breast cancer biopsy samples.
10. Staunton, J. E. *et al.* Chemosensitivity prediction by transcriptional profiling. *Proc. Natl Acad. Sci. USA* **98**, 10787–10792 (2001).
This study demonstrates the feasibility of chemosensitivity prediction from microarray data obtained from cell lines.
11. Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194**, 23–28 (1976).
12. Dudley, J. T., Tibshirani, R., Deshpande, T. & Butte, A. J. Disease signatures are robust across tissues and experiments. *Mol. Syst. Biol.* **5**, 307 (2009).
13. Ma'ayan, A. Colliding dynamical complex network models: biological attractors versus attractors from material physics. *Biophys. J.* **103**, 1816–1817 (2012).
14. Costello, J. C. *et al.* A community effort to assess and improve drug sensitivity prediction algorithms. *Nat. Biotechnol.* **32**, 1202–1212 (2014).
15. Lamb, J. *et al.* The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**, 1929–1935 (2006).
16. Gerstein, M. & Jansen, R. The current excitement in bioinformatics-analysis of whole-genome expression data: how does it relate to protein structure and function? *Curr. Opin. Struct. Biol.* **10**, 574–584 (2000).
17. Goya, R. *et al.* SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. *Bioinformatics* **26**, 730–736 (2010).
18. Maher, C. A. *et al.* Chimeric transcript discovery by paired-end transcriptome sequencing. *Proc. Natl Acad. Sci. USA* **106**, 12353–12358 (2009).
19. van Dijk, E. L., Auger, H., Jaszczyszyn, Y. & Thermes, C. Ten years of next-generation sequencing technology. *Trends Genet.* **30**, 418–426 (2014).
20. Kapranov, P. *et al.* RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316**, 1484–1488 (2007).
21. Lu, C. *et al.* Elucidation of the small RNA component of the transcriptome. *Science* **309**, 1567–1569 (2005).
22. Gall, J. G. & Pardue, M. L. Formation and detection of RNA-DNA hybrid molecules in cytological preparations. *Proc. Natl Acad. Sci. USA* **63**, 378–383 (1969).
23. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl Acad. Sci. USA* **74**, 5463–5467 (1977).
24. Alwine, J. C., Kemp, D. J. & Stark, G. R. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proc. Natl Acad. Sci. USA* **74**, 5350–5354 (1977).
25. Bell, G. I. *et al.* Nucleotide sequence of a cDNA clone encoding human preproinsulin. *Nature* **282**, 525–527 (1979).
26. Nakanishi, S. *et al.* Nucleotide sequence of cloned cDNA for bovine corticotropin-β-lipotropin precursor. *Nature* **278**, 423–427 (1979).
27. Fiddes, J. C. & Goodman, H. M. Isolation, cloning and sequence analysis of the cDNA for the alpha-subunit of human chorionic gonadotropin. *Nature* **281**, 351–356 (1979).
28. Okubo, K. *et al.* Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. *Nat. Genet.* **2**, 173–179 (1992).
29. Chiang, P. W. *et al.* Use of a fluorescent-PCR reaction to detect genomic sequence copy number and transcriptional abundance. *Genome Res.* **6**, 1013–1026 (1996).
30. Gibson, U. E., Heid, C. A. & Williams, P. M. A novel method for real time quantitative RT-PCR. *Genome Res.* **6**, 995–1001 (1996).
31. Heid, C. A., Stevens, J., Livak, K. J. & Williams, P. M. Real time quantitative PCR. *Genome Res.* **6**, 986–994 (1996).
32. Higuchi, R., Fockler, C., Dollinger, G. & Watson, R. Kinetic PCR analysis: real-time monitoring of DNA amplification reactions. *Biotechnology* **11**, 1026–1030 (1993).
33. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470 (1995).
34. Lockhart, D. J. *et al.* Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* **14**, 1675–1680 (1996).
35. Sutcliffe, J. G., Milner, R. J., Bloom, F. E. & Lerner, R. A. Common 82-nucleotide sequence unique to brain RNA. *Proc. Natl Acad. Sci. USA* **79**, 4942–4946 (1982).
36. Velculescu, V. E., Zhang, L., Vogelstein, B. & Kinzler, K. W. Serial analysis of gene expression. *Science* **270**, 484–487 (1995).
37. Hanriot, L. *et al.* A combination of LongSAGE with Solexa sequencing is well suited to explore the depth and the complexity of transcriptome. *BMC Genomics* **9**, 418 (2008).
38. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
39. Carninci, P. *et al.* High-efficiency full-length cDNA cloning by biotinylated CAP trapper. *Genomics* **37**, 327–336 (1996).
40. Dias Neto, E. *et al.* Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. *Proc. Natl Acad. Sci. USA* **97**, 3491–3496 (2000).
41. de Souza, S. J. *et al.* Identification of human chromosome 22 transcribed sequences with ORF expressed sequence tags. *Proc. Natl Acad. Sci. USA* **97**, 12690–12693 (2000).
42. Brenner, S. *et al.* Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.* **18**, 630–634 (2000).
43. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
44. Bainbridge, M. N. *et al.* Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genomics* **7**, 246 (2006).
45. Nielsen, K. L., Høgh, A. L. & Emmersen, J. DeepSAGE — digital transcriptomics with high sensitivity, simple experimental protocol and multiplexing of samples. *Nucleic Acids Res.* **34**, e133 (2006).
46. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).
47. Nagalakshmi, U. *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**, 1344–1349 (2008).
48. Kivioja, T. *et al.* Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* **9**, 72–74 (2012).
49. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep.* **2**, 666–673 (2012).
50. Cieslik, M. *et al.* The use of exome capture RNA-seq for highly degraded RNA with application to clinical cancer sequencing. *Genome Res.* **25**, 1372–1381 (2015).
51. Cabanski, C. R. *et al.* cDNA hybrid capture improves transcriptome analysis on low-input and archived samples. *J. Mol. Diagn.* **16**, 440–451 (2014).
52. Mercer, T. R. *et al.* Targeted sequencing for gene discovery and quantification using RNA CaptureSeq. *Nat. Protoc.* **9**, 989–1009 (2014).
53. Git, A. *et al.* Systematic comparison of microarray profiling, real-time PCR, and next-generation sequencing technologies for measuring differential microRNA expression. *RNA* **16**, 991–1006 (2010).
54. Yamamoto, T., Jay, G. & Pastan, I. Unusual features in the nucleotide sequence of a cDNA clone derived from the common region of avian sarcoma virus messenger RNA. *Proc. Natl Acad. Sci. USA* **77**, 176–180 (1980).
55. Zhang, L. *et al.* Gene expression profiles in normal and cancer cells. *Science* **276**, 1268–1272 (1997).
56. Brentani, H. *et al.* The generation and utilization of a cancer-oriented representation of the human transcriptome by using expressed sequence tags. *Proc. Natl Acad. Sci. USA* **100**, 13418–13423 (2003).
57. DeRisi, J. *et al.* Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat. Genet.* **14**, 457–460 (1996).
58. Alon, U. *et al.* Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl Acad. Sci. USA* **96**, 6745–6750 (1999).
59. Consortium, T. E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
60. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
61. Klijn, C. *et al.* A comprehensive transcriptional portrait of human cancer cell lines. *Nat. Biotechnol.* **33**, 306–312 (2015).
62. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
63. Uhlén, M. *et al.* Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
64. The Cancer Genome Atlas Research Network *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
65. Robinson, D. *et al.* Integrative clinical genomics of advanced prostate cancer. *Cell* **161**, 1215–1228 (2015).
66. Hamm, G. H. & Cameron, G. N. The EMBL data library. *Nucleic Acids Res.* **14**, 5–9 (1986).
67. Burks, C. *et al.* The GenBank nucleic acid sequence database. *Comput. Appl. Biosci.* **1**, 225–233 (1985).
68. Boguski, M. S., Lowe, T. M. J. & Tolstoshev, C. M. dbEST — database for 'expressed sequence tags'. *Nat. Genet.* **4**, 332–333 (1993).
69. Lal, A. *et al.* A public database for gene expression in human cancers. *Cancer Res.* **59**, 5403–5407 (1999).
70. Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197 (1981).
71. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
72. Pearson, W. R. & Lipman, D. J. Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA* **85**, 2444–2448 (1988).
73. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
74. Hubbard, T. *et al.* The Ensembl genome database project. *Nucleic Acids Res.* **30**, 38–41 (2002).
75. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **33**, D501–D504 (2005).
76. Iyer, M. K. *et al.* The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.* **47**, 199–208 (2015).
77. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**, 207–210 (2002).
78. Brazma, A. *et al.* ArrayExpress — a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.* **31**, 68–71 (2003).
79. Chen, Y., Dougherty, E. R. & Bittner, M. L. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J. Biomed. Opt.* **2**, 364–374 (1997).
80. Smyth, G., Yang, Y. & Speed, T. in *Functional Genomics* (eds Brownstein, M. & Khodursky, A.) 111–136 (Humana Press, 2003).

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

81. Tomlins, S. A. *et al.* Integrative molecular concept modeling of prostate cancer progression. *Nat. Genet.* **39**, 41–51 (2007).
82. Coletta, A. *et al.* InSilico DB genomic datasets hub: an efficient starting point for analyzing genome-wide studies in GenePattern, Integrative Genomics Viewer, and R/Bioconductor. *Genome Biol.* **13**, R104 (2012).
83. Qu, K. *et al.* Integrative genomic analysis by interoperation of bioinformatics tools in GenomeSpace. *Nat. Methods* **13**, 245–247 (2016).
84. Burrell, R. A., McGranahan, N., Bartek, J. & Swanton, C. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* **501**, 338–345 (2013).
85. Onder, T. T. *et al.* Loss of E-cadherin promotes metastasis via multiple downstream transcriptional pathways. *Cancer Res.* **68**, 3645–3654 (2008).
86. Van Allen, E. M. *et al.* Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science* **350**, 207–211 (2015).
87. Chen, J.-J., Knudsen, S., Mazin, W., Dahlgaard, J. & Zhang, B. A. 71-gene signature of TRAIL sensitivity in cancer cells. *Mol. Cancer Ther.* **11**, 34–44 (2012).
88. Rosenwald, A. *et al.* The proliferation gene expression signature is a quantitative integrator of oncogenic events that predicts survival in mantle cell lymphoma. *Cancer Cell* **3**, 185–197 (2003).
89. Carter, S. L., Eklund, A. C., Kohane, I. S., Harris, L. N. & Szallasi, Z. A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. *Nat. Genet.* **38**, 1043–1048 (2006).
This paper shows that aneuploidy is associated with a gene expression signature that is associated with poor clinical outcomes.
90. Ramaswamy, S., Ross, K. N., Lander, E. S. & Golub, T. R. A molecular signature of metastasis in primary solid tumors. *Nat. Genet.* **33**, 49–54 (2003).
This study reports a signature of cancer with high metastatic potential.
91. Bild, A. H. *et al.* Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* **439**, 353–357 (2006).
92. Ross, D. T. *et al.* Systematic variation in gene expression patterns in human cancer cell lines. *Nat. Genet.* **24**, 227–235 (2000).
93. Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome Res.* **22**, 2008–2017 (2012).
94. Singer, G. A. C. *et al.* Genome-wide analysis of alternative promoters of human genes using a custom promoter tiling array. *BMC Genomics* **9**, 349 (2008).
95. Nacu, S. *et al.* Deep RNA sequencing analysis of readthrough gene fusions in human prostate adenocarcinoma and reference samples. *BMC Med. Genom.* **4**, 11 (2011).
96. Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* **40**, 1413–1415 (2008).
97. Davuluri, R. V., Suzuki, Y., Sugano, S., Plass, C. & Huang, T. H.-M. The functional consequences of alternative promoter use in mammalian genomes. *Trends Genet.* **24**, 167–177 (2008).
98. Wiesner, T. *et al.* Alternative transcription initiation leads to expression of a novel *ALK* isoform in cancer. *Nature* **526**, 453–457 (2015).
99. Liu, J. *et al.* Integrated exome and transcriptome sequencing reveals *ZAK* isoform usage in gastric cancer. *Nat. Commun.* **5**, 3830 (2014).
100. Keene, J. D. RNA regulons: coordination of post-transcriptional events. *Nat. Rev. Genet.* **8**, 533–543 (2007).
101. Bahn, J. H. *et al.* Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. *Genome Res.* **22**, 142–150 (2012).
102. Dominissini, D., Moshitch-Moshkovitz, S., Salmon-Divon, M., Amariglio, N. & Rechavi, G. Transcriptome-wide mapping of N^6-methyladenosine by m^6A-seq based on immunocapturing and massively parallel sequencing. *Nat. Protoc.* **8**, 176–189 (2013).
103. Burns, M. B. *et al.* APOBEC3B is an enzymatic source of mutation in breast cancer. *Nature* **494**, 366–370 (2013).
104. Parkin, D. M. The global health burden of infection-associated cancers in the year 2002. *Int. J. Cancer* **118**, 3030–3044 (2006).
105. Abreu, A. L. P., Souza, R. P., Gimenes, F. & Consolaro, M. E. L. A review of methods for detect human *Papillomavirus* infection. *Virol. J.* **9**, 262 (2012).
106. Li, J.-W. *et al.* ViralFusionSeq: accurately discover viral integration events and reconstruct fusion transcripts at single-base resolution. *Bioinformatics* **29**, 649–651 (2013).
107. Piskol, R., Ramaswami, G. & Li, J. B. Reliable identification of genomic variants from RNA-seq data. *Am. J. Hum. Genet.* **93**, 641–651 (2013).
108. Kim, K.-T. *et al.* Single-cell mRNA sequencing identifies subclonal heterogeneity in anti-cancer drug responses of lung adenocarcinoma cells. *Genome Biol.* **16**, 127 (2015).
109. Paul, M. R. *et al.* Multivariate models from RNA-Seq SNVs yield candidate molecular targets for biomarker discovery: SNV-DA. *BMC Genomics* **17**, 263 (2016).
110. Rubinsteyn, A. *et al.* Computational pipeline for the PGV-001 neoantigen vaccine trial. Preprint at *bioRxiv* http://dx.doi.org/10.1101/174516 (2017).
111. Sheng, Q., Zhao, S., Li, C.-I., Shyr, Y. & Guo, Y. Practicability of detecting somatic point mutation from RNA high throughput sequencing data. *Genomics* **107**, 163–169 (2016).
112. Tang, X. *et al.* The eSNV-detect: a computational system to identify expressed single nucleotide variants from transcriptome sequencing data. *Nucleic Acids Res.* **42**, e172 (2014).
113. Lopez-Maestre, H. *et al.* SNP calling from RNA-seq data without a reference genome: identification, quantification, differential analysis and impact on the protein sequence. *Nucleic Acids Res.* **44**, e148 (2016).
114. Deelen, P. *et al.* Calling genotypes from public RNA-sequencing data enables identification of genetic variants that affect gene-expression levels. *Genome Med.* **7**, 30 (2015).
115. Wilkerson, M. D. *et al.* Integrated RNA and DNA sequencing improves mutation detection in low purity tumors. *Nucleic Acids Res.* **42**, e107 (2014).
116. Maher, C. A. *et al.* Transcriptome sequencing to detect gene fusions in cancer. *Nature* **458**, 97–101 (2009).
This study shows that gene fusions can be detected from RNA-seq data.
117. MacDonald, J. W. & Ghosh, D. COPA — cancer outlier profile analysis. *Bioinformatics* **22**, 2950–2951 (2006).
118. Tomlins, S. A. *et al.* Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* **310**, 644–648 (2005).
119. Romani, A., Guerra, E., Trerotola, M. & Alberti, S. Detection and analysis of spliced chimeric mRNAs in sequence databanks. *Nucleic Acids Res.* **31**, e17 (2003).
120. Gröschel, S. *et al.* A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia. *Cell* **157**, 369–381 (2014).
121. Kalyana-Sundaram, S. *et al.* Gene fusions associated with recurrent amplicons represent a class of passenger aberrations in breast cancer. *Neoplasia* **14**, 702–708 (2012).
122. Duro, D. *et al.* Inactivation of the *P16*^{INK4}/*MTS1* gene by a chromosome translocation t(9;14)(p21–22;q11) in an acute lymphoblastic leukemia of B-cell type. *Cancer Res.* **56**, 848–854 (1996).
123. Coyaud, E. *et al.* Wide diversity of PAX5 alterations in B-ALL: a Groupe Francophone de Cytogénétique Hématologique study. *Blood* **115**, 3089–3097 (2010).
124. Sun, Z., Bhagwate, A., Prodduturi, N., Yang, P. & Kocher, J.-P. A. Indel detection from RNA-seq data: tool evaluation and strategies for accurate detection of actionable mutations. *Brief. Bioinform.* http://dx.doi.org/10.1093/bib/bbw069 (2016).
125. Pickrell, J. K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768–772 (2010).
126. DeVeale, B., van der Kooy, D. & Babak, T. Critical evaluation of imprinted gene expression by RNA–Seq: a new perspective. *PLoS Genet.* **8**, e1002600 (2012).
127. Babak, T. *et al.* Genetic conflict reflected in tissue-specific maps of genomic imprinting in human and mouse. *Nat. Genet.* **47**, 544–549 (2015).
128. Reddy, T. E. *et al.* Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. *Genome Res.* **22**, 860–869 (2012).
129. Deng, Q., Ramsköld, D., Reinius, B. & Sandberg, R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* **343**, 193–196 (2014).
130. Tuch, B. B. *et al.* Tumor transcriptome sequencing reveals allelic expression imbalances associated with copy number alterations. *PLoS ONE* **5**, e9317 (2010).
131. Anwar, S. L. *et al.* Loss of imprinting and allelic switching at the *DLK1-MEG3* locus in human hepatocellular carcinoma. *PLoS ONE* **7**, e49462 (2012).
132. Burgess, M. R. *et al.* *KRAS* allelic imbalance enhances fitness and modulates MAP kinase dependence in cancer. *Cell* **168**, 817–829.e15 (2017).
133. Adiconis, X. *et al.* Comparative analysis of RNA sequencing methods for degraded or low-input samples. *Nat. Methods* **10**, 623–629 (2013).
134. Nilsson, J. *et al.* Prostate cancer-derived urine exosomes: a novel approach to biomarkers for prostate cancer. *Br. J. Cancer* **100**, 1603–1607 (2009).
135. Best, M. G. *et al.* RNA-Seq of tumor-educated platelets enables blood-based pan-cancer, multiclass, and molecular pathway cancer diagnostics. *Cancer Cell* **28**, 666–676 (2015).
136. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).
137. Benes, V., Blake, J. & Doyle, K. Ribo-Zero Gold Kit: improved RNA-seq results after removal of cytoplasmic and mitochondrial ribosomal RNA. *Nat. Methods* **8** (2011).
138. Yi, H. *et al.* Duplex-specific nuclease efficiently removes rRNA for prokaryotic RNA-seq. *Nucleic Acids Res.* **39**, e140 (2011).
139. Armour, C. D. *et al.* Digital transcriptome profiling using selective hexamer priming for cDNA synthesis. *Nat. Methods* **6**, 647–649 (2009).
140. Linsen, S. E. V. *et al.* Limitations and possibilities of small RNA digital gene expression profiling. *Nat. Methods* **6**, 474–476 (2009).
141. Raabe, C. A., Tang, T.-H., Brosius, J. & Rozhdestvensky, T. S. Biases in small RNA deep sequencing data. *Nucleic Acids Res.* **42**, 1414–1426 (2014).
142. Valen, E. *et al.* Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE. *Genome Res.* **19**, 255–265 (2009).
143. The FANTOM Consortium and the RIKEN PMI and CLST (DGT). A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014).
144. Zhernakova, D. V. *et al.* DeepSAGE reveals genetic variants associated with alternative polyadenylation and expression of coding and non-coding transcripts. *PLoS Genet.* **9**, e1003594 (2013).
145. Sigurgeirsson, B., Emanuelsson, O. & Lundeberg, J. Sequencing degraded RNA addressed by 3' tag counting. *PLoS ONE* **9**, e91851 (2014).
146. Langevin, S. A. *et al.* Peregrine: a rapid and unbiased method to produce strand-specific RNA-Seq libraries from small quantities of starting material. *RNA Biol.* **10**, 502–515 (2013).
147. Parkhomchuk, D. *et al.* Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res.* **37**, e123 (2009).
148. Hafner, M. *et al.* Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing. *Methods* **44**, 3–12 (2008).
149. Levin, J. Z. *et al.* Targeted next-generation sequencing of a cancer transcriptome enhances detection of sequence variants and novel fusion transcripts. *Genome Biol.* **10**, R115 (2009).
This is the first study to introduce the concept of capture RNA-seq.
150. Archer, S. K., Shirokikh, N. E. & Preiss, T. Selective and flexible depletion of problematic sequences from RNA-seq libraries at the cDNA stage. *BMC Genomics* **15**, 401 (2014).
151. Eikrem, O. *et al.* Transcriptome sequencing (RNAseq) enables utilization of formalin-fixed, paraffin-embedded biopsies with clear cell renal cell carcinoma for exploration of disease biology and biomarker development. *PLoS ONE* **11**, e0149743 (2016).
152. Beltran, H. *et al.* Impact of therapy on genomics and transcriptomics in high-risk prostate cancer treated with neoadjuvant docetaxel and androgen deprivation therapy. *Clin. Cancer Res.* http://dx.doi.org/10.1158/1078-0432.CCR-17-1034 (2017).
153. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**, 1845–1848 (2008).
154. Hah, N. *et al.* A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells. *Cell* **145**, 622–634 (2011).
155. Kim, Y. J. *et al.* HDAC inhibitors induce transcriptional repression of high copy number genes in breast cancer through elongation blockade. *Oncogene* **32**, 2828–2835 (2013).

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

156. Kertesz, M. *et al.* Genome-wide measurement of RNA secondary structure in yeast. *Nature* **467**, 103–107 (2010).
157. Spitale, R. C. *et al.* Structural imprints *in vivo* decode RNA regulatory mechanisms. *Nature* **519**, 486–490 (2015).
158. Kwok, C. K., Marsico, G., Sahakyan, A. B., Chambers, V. S. & Balasubramanian, S. rG4-seq reveals widespread formation of G-quadruplex structures in the human transcriptome. *Nat. Methods* **13**, 841–844 (2016).
159. Chu, C., Qu, K., Zhong, F. L., Artandi, S. E. & Chang, H. Y. Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. *Mol. Cell* **44**, 667–678 (2011).
160. Zhao, J. *et al.* Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Mol. Cell* **40**, 939–953 (2010).
161. Engreitz, J. M. *et al.* RNA-RNA interactions enable specific targeting of noncoding rnas to nascent pre-mRNAs and chromatin sites. *Cell* **159**, 188–199 (2014).
162. Hermann, T. & Westhof, E. RNA as a drug target: chemical, modelling, and evolutionary tools. *Curr. Opin. Biotechnol.* **9**, 66–73 (1998).
163. Arnold, C. D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074–1077 (2013).
164. Wang, N. *et al.* UNDO: a Bioconductor R package for unsupervised deconvolution of mixed gene expressions in tumor samples. *Bioinformatics* **31**, 137–139 (2015).
165. Li, S. *et al.* Detecting and correcting systematic variation in large-scale RNA sequencing data. *Nat. Biotechnol.* **32**, 888–895 (2014).
166. Leek, J. T. svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Res.* **42**, e161 (2014).
167. Smyth, G. K. in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* (eds Gentleman, R., Carey, V., Huber, W., Irizarry, R. & Dudoit, S.) 397–420 (Springer, 2005).
168. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).
This study introduces a simple normalization method for RNA-seq data that made it possible to use standard linear model tools for analysis.
169. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
170. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
171. Frasor, J. *et al.* Profiling of estrogen up- and down-regulated gene expression in human breast cancer cells: insights into gene networks and pathways underlying estrogenic control of proliferation and cell phenotype. *Endocrinology* **144**, 4562–4574 (2003).
172. Frazee, A. C. *et al.* Ballgown bridges the gap between transcriptome assembly and expression analysis. *Nat. Biotechnol.* **33**, 243–246 (2015).
173. Lee, H. K., Hsu, A. K., Sajdak, J., Qin, J. & Pavlidis, P. Coexpression analysis of human genes across many microarray data sets. *Genome Res.* **14**, 1085–1094 (2004).
174. Ackermann, M. & Strimmer, K. A general modular framework for gene set enrichment analysis. *BMC Bioinformatics* **10**, 47 (2009).
175. Mitrea, C. *et al.* Methods and approaches in the topology-based analysis of biological pathways. *Front. Physiol.* **4**, 278 (2013).
176. Majeti, R. *et al.* Dysregulated gene expression networks in human acute myelogenous leukemia stem cells. *Proc. Natl Acad. Sci. USA* **106**, 3396–3401 (2009).
177. de la Fuente, A. From ‘differential expression’ to ‘differential networking’ — identification of dysfunctional regulatory networks in diseases. *Trends Genet.* **26**, 326–333 (2010).
178. Woo, J. H. *et al.* Elucidating compound mechanism of action by network perturbation analysis. *Cell* **162**, 441–451 (2015).
179. Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).
180. Rhodes, D. R. *et al.* ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia* **6**, 1–6 (2004).
181. Xiao, Y. *et al.* Gene Perturbation Atlas (GPA): a single-gene perturbation repository for characterizing functional mechanisms of coding and non-coding genes. *Sci. Rep.* **5**, 10889 (2015).
182. Lynn, D. J. *et al.* InnateDB: facilitating systems-level analyses of the mammalian innate immune response. *Mol. Syst. Biol.* **4**, 218 (2008).
183. Ulloa-Montoya, F. *et al.* Predictive gene signature in MAGE-A3 antigen-specific cancer immunotherapy. *J. Clin. Orthod.* **31**, 2388–2395 (2013).
184. Saal, L. H. *et al.* Poor prognosis in carcinoma is associated with a gene expression signature of aberrant PTEN tumor suppressor pathway activity. *Proc. Natl Acad. Sci. USA* **104**, 7564–7569 (2007).
185. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics* **14**, 7 (2013).
186. Vaske, C. J. *et al.* Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* **26**, i237–i245 (2010).
187. Witt, H. *et al.* Delineation of two clinically and molecularly distinct subgroups of posterior fossa ependymoma. *Cancer Cell* **20**, 143–157 (2011).
188. Bayliss, J. *et al.* Lowered H3K27me3 and DNA hypomethylation define poorly prognostic pediatric posterior fossa ependymomas. *Sci. Transl Med.* **8**, 366ra161 (2016).
189. Alizadeh, A. A. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511 (2000).
190. Roberts, K. G. *et al.* Targetable kinase-activating lesions in Ph-like acute lymphoblastic leukemia. *N. Engl. J. Med.* **371**, 1005–1015 (2014).
191. van ’t Veer, L. J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536 (2002).
This study demonstrates the use of microarrays to prognosticate and distinguish cancers with *BRCA1* or *BRCA2* mutations.
192. Paik, S. *et al.* A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N. Engl. J. Med.* **351**, 2817–2826 (2004).
193. Parker, J. S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **27**, 1160–1167 (2009).
194. 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).
195. Yeoh, E.-J. *et al.* Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell* **1**, 133–143 (2002).
This study discovers subtypes of ALL that differ in biology, outcomes and response to therapy.
196. Verhaak, R. G. W. *et al.* Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* **17**, 98 (2010).
197. Anghel, C. V. *et al.* ISOpureR: an R implementation of a computational purification algorithm of mixed tumour profiles. *BMC Bioinformatics* **16**, 156 (2015).
198. Yoshihara, K. *et al.* Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **4**, 2612 (2013).
199. Quon, G. *et al.* Computational purification of individual tumor gene expression profiles leads to significant improvements in prognostic prediction. *Genome Med.* **5**, 29 (2013).
200. Ciriello, G. *et al.* Comprehensive molecular portraits of invasive lobular breast cancer. *Cell* **163**, 506–519 (2015).
201. Choi, H. *et al.* Transcriptome analysis of individual stromal cell populations identifies stroma-tumor crosstalk in mouse lung cancer model. *Cell Rep.* **10**, 1187–1201 (2015).
202. Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396–1401 (2014).
203. Shaffer, S. M. *et al.* Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. *Nature* **546**, 431–435 (2017).
204. Giustacchini, A. *et al.* Single-cell transcriptomics uncovers distinct molecular signatures of stem cells in chronic myeloid leukemia. *Nat. Med.* **23**, 692–702 (2017).
205. Chin, K. *et al.* Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell* **10**, 529–541 (2006).
206. Kuijjer, M. L. *et al.* Identification of osteosarcoma driver genes by integrative analysis of copy number and gene expression data. *Genes Chromosomes Cancer* **51**, 696–706 (2012).
207. Kristensen, V. N. *et al.* Integrated molecular profiles of invasive breast tumors and ductal carcinoma *in situ* (DCIS) reveal differential vascular and interleukin signaling. *Proc. Natl Acad. Sci. USA* **109**, 2802–2807 (2012).
208. Degner, J. F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390–394 (2012).
209. Michailidou, K. *et al.* Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nat. Genet.* **47**, 373–380 (2015).
210. Bojesen, S. E. *et al.* Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer. *Nat. Genet.* **45**, 371–384 (2013).
211. Masica, D. L. & Karchin, R. Correlation of somatic mutation and expression identifies genes important in human glioblastoma progression and survival. *Cancer Res.* **71**, 4550–4561 (2011).
212. Kristensen, V. N. *et al.* Principles and methods of integrative genomic analyses in cancer. *Nat. Rev. Cancer* **14**, 299–313 (2014).
213. Ramaswamy, S. *et al.* Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl Acad. Sci. USA* **98**, 15149–15154 (2001).
214. Hoadley, K. A. *et al.* Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* **158**, 929–944 (2014).
215. Torrente, A. *et al.* Identification of cancer related genes using a comprehensive map of human gene expression. *PLoS ONE* **11**, e0157484 (2016).
216. Anaya, J., Reon, B., Chen, W.-M., Bekiranov, S. & Dutta, A. A pan-cancer analysis of prognostic genes. *PeerJ* **3**, e1499 (2015).
217. Tang, K.-W., Alaei-Mahabadi, B., Samuelsson, T., Lindh, M. & Larsson, E. The landscape of viral expression and host gene fusion and adaptation in human cancer. *Nat. Commun.* **4**, 2513 (2013).
218. Yoshihara, K. *et al.* The landscape and therapeutic relevance of cancer-associated transcript fusions. *Oncogene* **34**, 4845–4854 (2015).
219. Xia, Z. *et al.* Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3′-UTR landscape across seven tumour types. *Nat. Commun.* **5**, 5274 (2014).
220. Fehrmann, R. S. N. *et al.* Gene expression analysis identifies global gene dosage sensitivity in cancer. *Nat. Genet.* **47**, 115–125 (2015).
221. Mody, R. J. *et al.* Integrative clinical sequencing in the management of refractory or relapsed cancer in youth. *JAMA* **314**, 913–925 (2015).
This is one of the first studies to demonstrate the feasibility and utility of RNA-seq in the real-time management of paediatric tumours.
222. Oberg, J. A. *et al.* Implementation of next generation sequencing into pediatric hematology-oncology practice: moving beyond actionable alterations. *Genome Med.* **8**, 133 (2016).
223. Robinson, D. R. *et al.* Integrative clinical genomics of metastatic cancer. *Nature* **548**, 297–303 (2017).
This is the first study to demonstrate the broad utility of transcriptomic data in characterizing metastatic tumours.
224. Cheng, D. T. *et al.* Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): a hybridization capture-based next-generation sequencing clinical assay for solid tumor molecular oncology. *J. Mol. Diagn.* **17**, 251–264 (2015).
225. Shukla, S. *et al.* Identification and validation of PCAT14 as prognostic biomarker in prostate cancer. *Neoplasia* **18**, 489–499 (2016).
226. Yang, L. *et al.* Analyzing somatic genome rearrangements in human cancers by using whole-exome sequencing. *Am. J. Hum. Genet.* **98**, 843–856 (2016).
227. Hutchins, G. *et al.* Value of mismatch repair, *KRAS*, and *BRAF* mutations in predicting recurrence and benefits from chemotherapy in colorectal cancer. *J. Clin. Oncol.* **29**, 1261–1270 (2011).
228. Meng, X., Huang, Z., Teng, F., Xing, L. & Yu, J. Predictive biomarkers in PD-1/PD-L1 checkpoint blockade immunotherapy. *Cancer Treat. Rev.* **41**, 868–876 (2015).
229. Gawad, C., Koh, W. & Quake, S. R. Single-cell genome sequencing: current state of the science. *Nat. Rev. Genet.* **17**, 175–188 (2016).

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

230. Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014).
231. Cardoso, F. *et al.* 70-gene signature as an aid to treatment decisions in early-stage breast cancer. *N. Engl. J. Med.* **375**, 717–729 (2016).
This is a large-scale, multi-institutional study to evaluate the clinical utility of MammaPrint.
232. CRUK Lung Cancer Centre of Excellence. TRACERx. *CRUK Lung Cancer Centre of Excellence* http://www.cruklungcentre.org/Research/TRACERx (2017).
233. MD Anderson Cancer Center. APOLLO. *MD Anderson Cancer Center* https://www.mdanderson.org/cancermoonshots/research_platforms/apollo.html (2017).
234. Wei, I. H., Shi, Y., Jiang, H., Kumar-Sinha, C. & Chinnaiyan, A. M. RNA-Seq accurately identifies cancer biomarker signatures to distinguish tissue of origin. *Neoplasia* **16**, 918–927 (2014).
235. Feng, H., Zhang, X. & Zhang, C. mRIN for direct assessment of genome-wide and gene-specific mRNA integrity from large-scale RNA-sequencing data. *Nat. Commun.* **6**, 7816 (2015).
236. Karmakar, S. *et al.* Organocatalytic removal of formaldehyde adducts from RNA and DNA bases. *Nat. Chem.* **7**, 752–758 (2015).
237. Fernando, M. R., Norton, S. E., Luna, K. K., Lechner, J. M. & Qin, J. Stabilization of cell-free RNA in blood samples using a new collection device. *Clin. Biochem.* **45**, 1497–1502 (2012).
238. Alhasan, A. A. *et al.* Circular RNA enrichment in platelets is a signature of transcriptome degradation. *Blood* **127**, e1–e11 (2016).
239. Li, Y. *et al.* Circular RNA is enriched and stable in exosomes: a promising biomarker for cancer diagnosis. *Cell Res.* **25**, 981–984 (2015).
240. Arroyo, J. D. *et al.* Argonaute2 complexes carry a population of circulating microRNAs independent of vesicles in human plasma. *Proc. Natl Acad. Sci. USA* **108**, 5003–5008 (2011).
241. Huang, X. *et al.* Characterization of human plasma-derived exosomal RNAs by deep sequencing. *BMC Genomics* **14**, 319 (2013).
242. Chen, X. Q. *et al.* Telomerase RNA as a detection marker in the serum of breast cancer patients. *Clin. Cancer Res.* **6**, 3823–3826 (2000).
243. Wu, A. R. *et al.* Quantitative assessment of single-cell RNA-sequencing methods. *Nat. Methods* **11**, 41–46 (2014).
244. Kong-Beltran, M. *et al.* Somatic mutations lead to an oncogenic deletion of met in lung cancer. *Cancer Res.* **66**, 283–289 (2006).
245. Zhang, J., Mardis, E. R. & Maher, C. A. INTEGRATE-neo: a pipeline for personalized gene fusion neoantigen discovery. *Bioinformatics* **33**, 555–557 (2016).
246. Mehra, R. *et al.* Biallelic alteration and dysregulation of the Hippo pathway in mucinous tubular and spindle cell carcinoma of the kidney. *Cancer Discov.* **6**, 1258–1266 (2016).
247. van Rhee, F. *et al.* NY-ESO-1 is highly expressed in poor-prognosis multiple myeloma and induces spontaneous humoral and cellular immune responses. *Blood* **105**, 3939–3944 (2005).
248. Ludwig, J. A. & Weinstein, J. N. Biomarkers in cancer staging, prognosis and treatment selection. *Nat. Rev. Cancer* **5**, 845–856 (2005).
249. Kulasingam, V. & Diamandis, E. P. Strategies for discovering novel cancer biomarkers through utilization of emerging technologies. *Nat. Clin. Pract. Oncol.* **5**, 588–599 (2008).
250. Topalian, S. L., Taube, J. M., Anders, R. A. & Pardoll, D. M. Mechanism-driven biomarkers to guide immune checkpoint blockade in cancer therapy. *Nat. Rev. Cancer* **16**, 275–287 (2016).
251. Aran, D. *et al.* Widespread parainflammation in human cancer. *Genome Biol.* **17**, 145 (2016).
252. Gentles, A. J. *et al.* The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nat. Med.* **21**, 938–945 (2015).
253. Charoentong, P. *et al.* Pan-cancer immunogenomic analyses reveal genotype-immunophenotype relationships and predictors of response to checkpoint blockade. *Cell Rep.* **18**, 248–262 (2017).
254. Chen, P.-L. *et al.* Analysis of immune signatures in longitudinal tumor samples yields insight into biomarkers of response and mechanisms of resistance to immune checkpoint blockade. *Cancer Discov.* **6**, 827–837 (2016).
This is one of the first longitudinal studies involving RNA-seq profiling.
255. Roh, W. *et al.* Integrated molecular analysis of tumor biopsies on sequential CTLA-4 and PD-1 blockade reveals markers of response and resistance. *Sci. Transl Med.* **9**, eaah3560 (2017).
256. Paluch, B. E. *et al.* Robust detection of immune transcripts in FFPE samples using targeted RNA sequencing. *Oncotarget* **8**, 3197–3205 (2017).
257. Ott, P. A. *et al.* An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature* **547**, 217–221 (2017).
258. Carreno, B. M. *et al.* A dendritic cell vaccine increases the breadth and diversity of melanoma neoantigen-specific T cells. *Science* **348**, 803–808 (2015).
259. Gong, T. & Szustakowski, J. D. DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. *Bioinformatics* **29**, 1083–1085 (2013).
260. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 (2015).
261. Bolotin, D. A. *et al.* MiXCR: software for comprehensive adaptive immunity profiling. *Nat. Methods* **12**, 380–381 (2015).
262. Mose, L. E. *et al.* Assembly-based inference of B-cell receptor repertoires from short read RNA sequencing data with V'DJer. *Bioinformatics* **32**, 3729–3734 (2016).
263. Seqc/Maqc-Iii Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat. Biotechnol.* **32**, 903–914 (2014).
264. Fumagalli, D. *et al.* Transfer of clinically relevant gene expression signatures in breast cancer: from Affymetrix microarray to Illumina RNA-Sequencing technology. *BMC Genomics* **15**, 1008 (2014).
265. Zhang, W. *et al.* Comparison of RNA-seq and microarray-based models for clinical endpoint prediction. *Genome Biol.* **16**, 133 (2015).
266. Schurch, N. J. *et al.* How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA* **22**, 839–851 (2016).
267. Thierry-Mieg, D. & Thierry-Mieg, J. AceView: a comprehensive cDNA-supported gene and transcripts annotation. *Genome Biol.* **7** (Suppl. 1), S12 (2006).
268. Ermolaeva, O. *et al.* Data management and analysis for gene expression arrays. *Nat. Genet.* **20**, 19–23 (1998).
269. Shiraki, T. *et al.* Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl Acad. Sci. USA* **100**, 15776–15781 (2003).
270. Strausberg, R. L. Cancer Genome Anatomy Project. *eLS* http://dx.doi.org/10.1038/npg.els.0006070 (2006).
271. Hon, C.-C. *et al.* An atlas of human long non-coding RNAs with accurate 5′ ends. *Nature* **543**, 199–204 (2017).
272. Searle, S. *et al.* The GENCODE human gene set. *Genome Biol.* **11** (Suppl. 1), P36 (2010).
273. Subramanian, A., Kuehn, H., Gould, J., Tamayo, P. & Mesirov, J. P. GSEA-P: a desktop application for Gene Set Enrichment Analysis. *Bioinformatics* **23**, 3251–3253 (2007).
274. Hsu, F. *et al.* The UCSC known genes. *Bioinformatics* **22**, 1036–1046 (2006).
275. Mitelman, F., Johansson, B., & Mertens, F. Mitelman database of chromosome aberrations in cancer. *National Cancer Institute* https://cgap.nci.nih.gov/Chromosomes/Mitelman (2001).
276. Frohman, M. A., Dush, M. K. & Martin, G. R. Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. *Proc. Natl Acad. Sci. USA* **85**, 8998–9002 (1988).
277. Wang, F. *et al.* RNAscope: a novel *in situ* RNA analysis platform for formalin-fixed, paraffin-embedded tissues. *J. Mol. Diagn.* **14**, 22–29 (2012).
278. Lister, R. *et al.* Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* **133**, 523–536 (2008).
279. Lash, A. E. *et al.* SAGEmap: a public gene expression resource. *Genome Res.* **10**, 1051–1060 (2000).
280. Patro, R., Mount, S. M. & Kingsford, C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat. Biotechnol.* **32**, 462–464 (2014).
281. Tusher, V. G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA* **98**, 5116–5121 (2001).
282. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
283. Wu, C. *et al.* BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol.* **10**, R130 (2009).
284. Niknafs, Y. S., Pandian, B., Iyer, H. K., Chinnaiyan, A. M. & Iyer, M. K. TACO produces robust multisample transcriptome assemblies from RNA-seq. *Nat. Methods* **14**, 68–70 (2017).
285. Robertson, G. *et al. De novo* assembly and analysis of RNA-seq data. *Nat. Methods* **7**, 909–912 (2010).
286. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
287. Goldman, M. *et al.* The UCSC Xena system for integrating and visualizing functional genomics [abstract]. *Cancer Res.* **76** (Suppl.), 5270 (2016).
288. Mitelman, F., Johansson, B. & Mertens, F. Fusion genes and rearranged genes as a linear function of chromosome aberrations in cancer. *Nat. Genet.* **36**, 331–334 (2004).
289. Forbes, S. A. *et al.* The Catalogue of Somatic Mutations in Cancer (COSMIC). *Curr. Protoc. Hum. Genet.* **57**, 10.11 (2008).
290. Hartley, S. W. & Mullikin, J. C. QoRTs: a comprehensive toolset for quality control and data processing of RNA-Seq experiments. *BMC Bioinformatics* **16**, 224 (2015).
291. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2013).
292. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
293. Nicorici, D. *et al.* FusionCatcher — a tool for finding somatic fusion genes in paired-end RNA-sequencing data. Preprint at *bioRxiv* http://dx.doi.org/10.1101/011650 (2014).
294. Kim, D. & Salzberg, S. L. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol.* **12**, R72 (2011).
295. Grossman, R. L., Heath, A. P., Ferreti, V., Varmus, H. E., Lowy, D. R., Kibbe, W. A. & Staudt, L. M. Toward a shared vision for cancer genomic data. *N. Engl. J. Med.* **375**, 1109–1112 (2016).

Acknowledgements
The authors thank S. Ellison for assistance in writing, editing and preparing this manuscript. A.M.C. is a Howard Hughes Medical Institute investigator and American Cancer Society professor. M.C. is a Prostate Cancer Foundation Young Investigator.

Author contributions
Both authors made substantial contributions to the discussion of content and reviewing and editing the manuscript before submission. M.C. was primarily involved in researching data for the article, and A.M.C. was involved in writing the manuscript.

Competing interests statement
The authors declare no competing interests.

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

© 2017 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.