

Transcriptomics technologies

Rohan Lowe¹, Neil Shirley², Mark Bleackley¹, Stephen Dolan³, Thomas Shafee^{1*}

1 La Trobe Institute for Molecular Science, La Trobe University, Melbourne, Australia, **2** ARC Centre of Excellence in Plant Cell Walls, University of Adelaide, Adelaide, Australia, **3** Department of Biochemistry, University of Cambridge, Cambridge, United Kingdom

* T.Shafee@LaTrobe.edu.au

Abstract

Transcriptomics technologies are the techniques used to study an organism's [transcriptome](#), the sum of all of its [RNA transcripts](#). The information content of an organism is recorded in the DNA of its [genome](#) and [expressed](#) through [transcription](#). Here, [mRNA](#) serves as a transient intermediary molecule in the information network, whilst [noncoding RNAs](#) perform additional diverse functions. A transcriptome captures a snapshot in time of the total transcripts present in a [cell](#).

The first attempts to study the whole transcriptome began in the early 1990s, and technological advances since the late 1990s have made transcriptomics a widespread discipline. Transcriptomics has been defined by repeated technological innovations that transform the field. There are two key contemporary techniques in the field: [microarrays](#), which quantify a set of predetermined sequences, and RNA sequencing ([RNA-Seq](#)), which uses [high-throughput sequencing](#) to capture all sequences.

Measuring the expression of an organism's [genes](#) in different [tissues](#), [conditions](#), or time points gives information on how genes are [regulated](#) and reveals details of an organism's biology. It can also help to infer the [functions](#) of previously [unannotated](#) genes. Transcriptomic analysis has enabled the study of how gene expression changes in different organisms and has been instrumental in the understanding of human [disease](#). An analysis of gene expression in its entirety allows detection of broad coordinated trends which cannot be discerned by more targeted [assays](#).

OPEN ACCESS

Citation: Lowe R, Shirley N, Bleackley M, Dolan S, Shafee T (2017) Transcriptomics technologies. *PLoS Comput Biol* 13(5): e1005457. <https://doi.org/10.1371/journal.pcbi.1005457>

Published: May 18, 2017

Copyright: © 2017 Lowe et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by the Australian Research Council grant DP160100309. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

This is a "Topic Page" article for *PLOS Computational Biology*.

History

Transcriptomics has been characterised by the development of new techniques which have redefined what is possible every decade or so and render previous technologies obsolete ([Fig 1](#)). The first attempt at capturing a partial human transcriptome was published in 1991 and

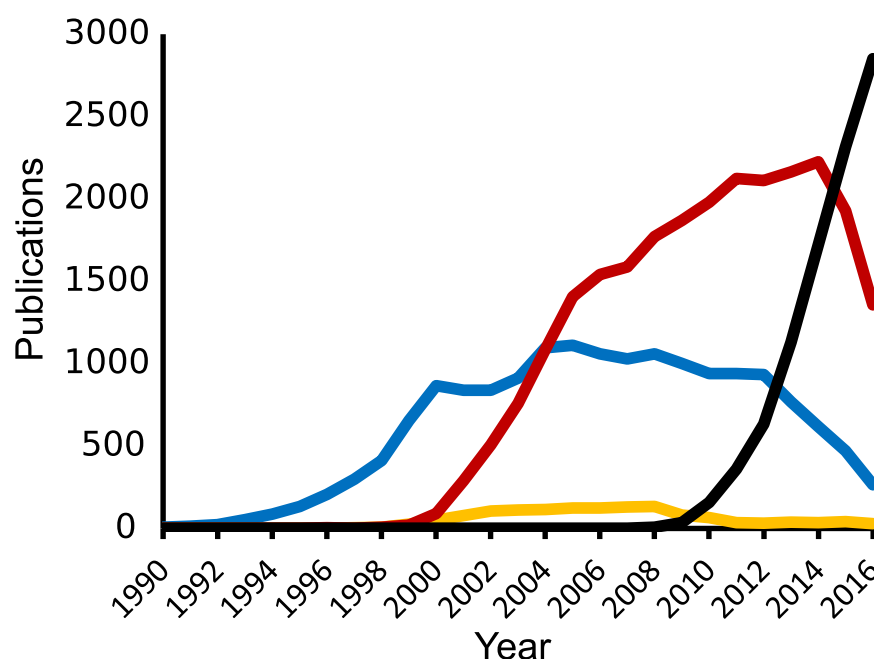


Fig 1. Transcriptomics method use over time. Published papers since 1990, referring to RNA sequencing (black), RNA microarray (red), expressed sequence tag (blue), and serial/cap analysis of gene expression (yellow)[12].

<https://doi.org/10.1371/journal.pcbi.1005457.g001>

reported 609 mRNA sequences from the human brain [1]. In 2008, two human transcriptomes composed of millions of transcript-derived sequences covering 16,000 genes were published [2][3], and, by 2015, transcriptomes had been published for hundreds of individuals [4][5]. Transcriptomes of different disease states, tissues, or even single cells are now routinely generated [5][6][7]. This explosion in transcriptomics has been driven by the rapid development of new technologies with an improved sensitivity and economy (Table 1) [8][9][10][11].

Table 1. Comparison of contemporary methods [23] [24] [10].

Method	RNA-Seq	Microarray
Throughput	High [10]	Higher [10]
Input RNA amount	Low ~ 1 ng total RNA [25]	High ~ 1 µg mRNA [26]
Labour intensity	High (sample preparation and data analysis) [10][23]	Low [10][23]
Prior knowledge	None required, though genome sequence useful [23]	Reference transcripts required for probes [23]
Quantitation accuracy	~90% (limited by sequence coverage) [27]	>90% (limited by fluorescence detection accuracy) [27]
Sequence resolution	Can detect SNPs and splice variants (limited by sequencing accuracy of ~99%) [27]	Dedicated arrays can detect splice variants (limited by probe design and cross-hybridisation) [27]
Sensitivity	10 ³ (limited by sequence coverage) [27]	10 ³ (limited by fluorescence detection) [27]
Dynamic range	>10 ⁵ (limited by sequence coverage) [28]	10 ³ (limited by fluorescence saturation) [28]
Technical reproducibility	>99% [29][30]	>99% [31][32]

RNA-Seq, RNA Sequencing

<https://doi.org/10.1371/journal.pcbi.1005457.t001>

Before transcriptomics

Studies of individual [transcripts](#) were being performed several decades before any transcriptomics approaches were available. [Libraries](#) of [silkworm](#) mRNAs were collected and converted to [complementary DNA](#) (cDNA) for storage using [reverse transcriptase](#) in the late 1970s [13]. In the 1980s, low-throughput [Sanger sequencing](#) began to be used to sequence random individual transcripts from these libraries, called [expressed sequence tags](#) (ESTs) [2][14][15][16]. The [Sanger method of sequencing](#) was predominant until the advent of [high-throughput methods](#) such as [sequencing by synthesis](#) (Solexa/Illumina, San Diego, CA). ESTs came to prominence during the 1990s as an efficient method to determine the [gene content](#) of an organism without [sequencing](#) the entire [genome](#) [16]. Quantification of individual transcripts by [northern blotting](#), [nylon membrane arrays](#), and later [reverse transcriptase quantitative PCR](#) (RT-qPCR) were also popular [17][18], but these methods are laborious and can only capture a tiny subsection of a transcriptome [12]. Consequently, the manner in which a transcriptome as a whole is expressed and regulated remained unknown until higher-throughput techniques were developed.

Early attempts

The word *transcriptome* was first used in the 1990s [19][20]. In 1995, one of the earliest sequencing-based transcriptomic methods was developed, [serial analysis of gene expression](#) (SAGE), which worked by [Sanger sequencing](#) of concatenated random transcript fragments [21]. Transcripts were quantified by matching the fragments to known genes. A variant of SAGE using high-throughput sequencing techniques, called digital gene expression analysis, was also briefly used [9][22]. However, these methods were largely overtaken by high throughput sequencing of entire transcripts, which provided additional information on transcript structure, e.g., [splice variants](#) [9].

Development of contemporary techniques

The dominant contemporary techniques, [microarrays](#) and [RNA-Seq](#), were developed in the mid-1990s and 2000s [9][33]. Microarrays that measure the abundances of a defined set of transcripts via their [hybridisation](#) to an array of [complementary probes](#) were first published in 1995 [34][35]. Microarray technology allowed the assay of thousands of transcripts simultaneously at a greatly reduced cost per gene and labour saving [36]. Both [spotted oligonucleotide arrays](#) and [Affymetrix](#) (Santa Clara, California) high-density arrays were the method of choice for transcriptional profiling until the late 2000s [12][33]. Over this period, a range of microarrays were produced to cover known genes in [model](#) or economically important organisms. Advances in design and manufacture of arrays improved the specificity of probes and allowed for more genes to be tested on a single array. Advances in [fluorescence detection](#) increased the sensitivity and measurement accuracy for low abundance transcripts [35][37].

RNA-Seq refers to the sequencing of transcript [cDNAs](#), in which abundance is derived from the number of counts from each transcript. The technique has therefore been heavily influenced by the development of [high-throughput sequencing technologies](#) [9][11]. [Massively parallel signature sequencing](#) (MPSS) was an early example based on generating 16×10^6 sequences via a complex series of hybridisations [38] and was used in 2004 to validate the expression of 10^4 genes in *Arabidopsis thaliana* [39]. The earliest RNA-Seq work was published in 2006 with 10^5 transcripts sequenced using the [454 technology](#) [40]. This was sufficient coverage to quantify relative transcript abundance. RNA-Seq began to increase in popularity after 2008 when new [Solexa/Illumina technologies](#) (San Diego, CA) allowed 10^9 transcript

sequences to be recorded [4][10][41][42]. This yield is now sufficient for accurate [quantitation](#) of entire human transcriptomes.

Data gathering

Generating data on RNA transcripts can be achieved via either of two main principles: sequencing of individual transcripts ([ESTs](#), or RNA-Seq), or [hybridisation](#) of transcripts to an ordered array of nucleotide probes (i.e., microarrays).

Isolation of RNA

All transcriptomic methods require RNA to first be isolated from the experimental organism before transcripts can be recorded. Although biological systems are incredibly diverse, [RNA extraction](#) techniques are broadly similar and involve the following: mechanical [disruption of cells](#) or tissues, disruption of [RNase](#) with [chaotropic salts](#) [43], disruption of macromolecules and nucleotide complexes, separation of RNA from undesired [biomolecules](#) including DNA, and concentration of the RNA via [precipitation](#) from solution or [elution from a solid matrix](#) [43][44]. Isolated RNA may additionally be treated with [DNase](#) to digest any traces of DNA [45]. It is necessary to enrich messenger RNA as total RNA extracts are typically 98% [ribosomal RNA](#) [46]. Enrichment for transcripts can be performed by [poly-A](#) affinity methods or by depletion of ribosomal RNA using sequence-specific probes [47]. Degraded RNA may affect downstream results; for example, mRNA enrichment from degraded samples will result in the depletion of [5' mRNA ends](#) and uneven signal across the length of a transcript. [Snap-freezing](#) of tissue prior to RNA isolation is typical, and care is taken to reduce exposure to RNase enzymes once isolation is complete [44].

EST

An [EST](#) is a short nucleotide sequence generated from a single RNA transcript. RNA is first copied as [cDNA](#) by a [reverse transcriptase](#) enzyme before the resultant cDNA is sequenced [16]. The [Sanger method of sequencing](#) was predominant until the advent of [high-throughput methods](#) such as [sequencing by synthesis](#) (Solexa/Illumina, San Diego, CA). Because ESTs don't require prior knowledge of the organism from which they come, they can also be made from mixtures of organisms or environmental samples [16]. Although higher-throughput methods are now used, [EST libraries](#) commonly provided sequence information for early microarray designs; for example, a [barley](#) GeneChip was designed from 350,000 previously sequenced ESTs [48].

Serial and Cap analysis of gene expression (SAGE/CAGE)

[SAGE](#) was a development of EST methodology to increase the throughput of the tags generated and allow some quantitation of transcript abundance ([Fig 2](#)) [21]. cDNA is generated from the RNA but is then digested into 11 bp $\frac{1}{2}$ tag fragments using [restriction enzymes](#) that cut at a specific sequence, and 11 base pairs along from that sequence. These cDNA tags are then [concatenated](#) head-to-tail into long strands (>500 bp) and sequenced using low-throughput, but long read length methods such as [Sanger sequencing](#). Once the sequences are [deconvoluted](#) into their original 11 bp tags [21]. If a [reference genome](#) is available, these tags can sometimes be aligned to identify their corresponding gene. If a reference genome is unavailable, the tags can simply be directly used as diagnostic markers if found to be [differentially expressed](#) in a disease state.

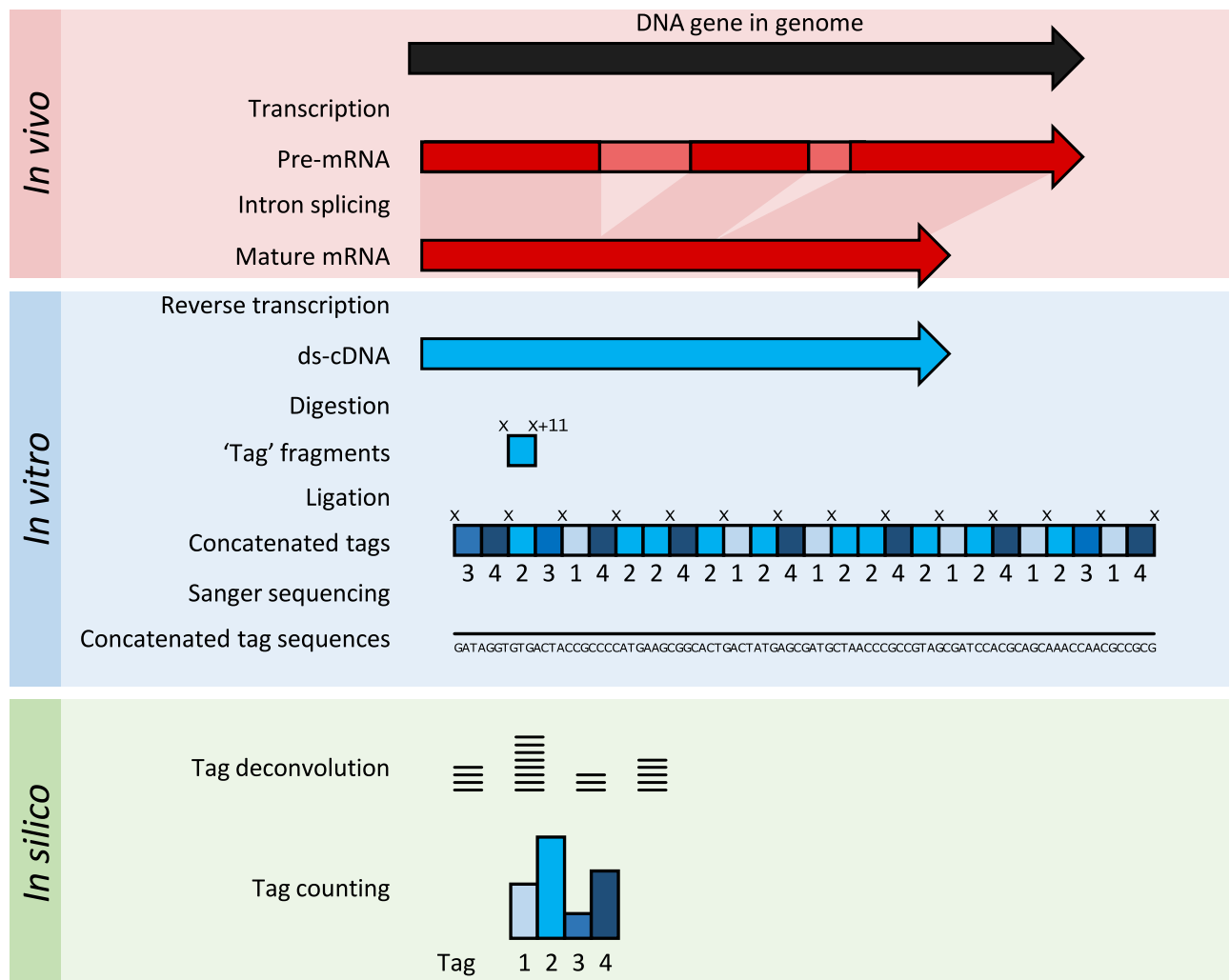


Fig 2. Summary of SAGE. Within the organisms, genes are transcribed and spliced (in eukaryotes) to produce mature mRNA transcripts (red). The mRNA is extracted from the organism, and reverse transcriptase is used to copy the mRNA into stable double-stranded cDNA (ds-cDNA; blue). In SAGE, the ds-cDNA is digested by restriction enzymes (at location x and $x+11$) to produce 11-nucleotide tags. These tags are concatenated and sequenced using long-read Sanger sequencing (different shades of blue indicate tags from different genes). The sequences are deconvoluted to find the frequency of each tag. The tag frequency can be used to report on transcription of the gene that the tag came from.

<https://doi.org/10.1371/journal.pcbi.1005457.g002>

The Cap analysis of gene expression (CAGE) method is a variant of SAGE that sequences tags from the 5' end of an mRNA transcript only [49]. Therefore, the transcriptional start site of genes can be identified when the tags are aligned to a reference genome. Identifying gene start sites is of use for promoter analysis and for the cloning of full-length cDNAs.

SAGE and CAGE methods produce information on more genes than was possible when sequencing single ESTs, but the sample preparation and data analysis are typically more labour intensive.

Microarrays

Principles and advances. Microarrays consist of short nucleotide oligomers, known as "probes," which are arrayed on a solid substrate (e.g., glass) [50]. Transcript abundance is determined by hybridisation of fluorescently labelled transcripts to these probes (Fig 3) [51].

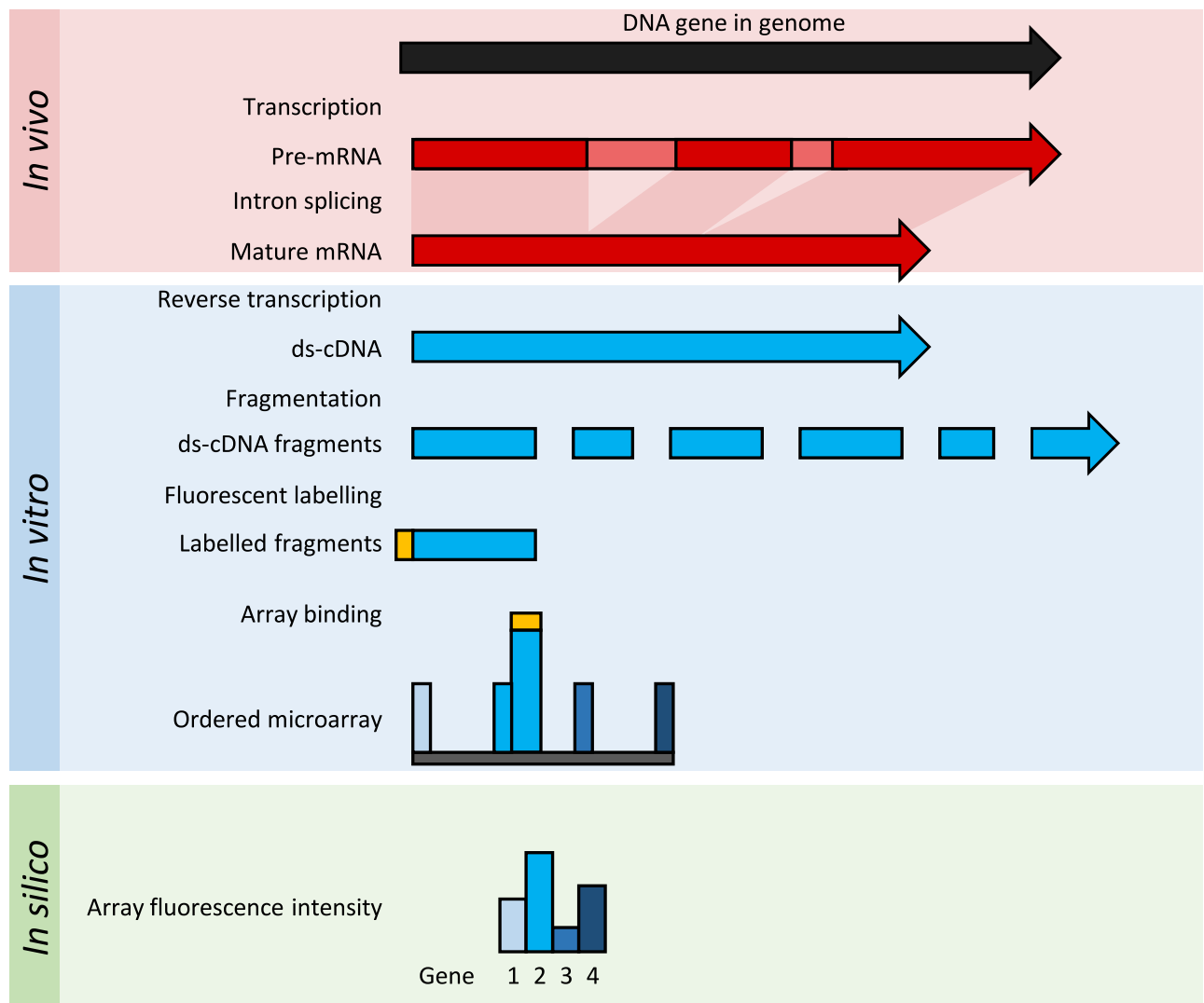


Fig 3. Summary of DNA microarrays. Within the organisms, genes are transcribed and spliced (in eukaryotes) to produce mature mRNA transcripts (red). The mRNA is extracted from the organism and reverse transcriptase is used to copy the mRNA into stable double-stranded cDNA (ds-cDNA; blue). In microarrays, the ds-cDNA is fragmented and fluorescently labelled (orange). The labelled fragments bind to an ordered array of complementary oligonucleotides, and measurement of fluorescent intensity across the array indicates the abundance of a predetermined set of sequences. These sequences are typically specifically chosen to report on genes of interest within the organism's genome.

<https://doi.org/10.1371/journal.pcbi.1005457.g003>

The [fluorescence intensity](#) at each probe location on the array indicates the transcript abundance for that probe sequence [51]. Microarrays require some prior knowledge of the organism of interest, for example, in the form of an [annotated genome](#) sequence or in a [library](#) of ESTs that can be used to generate the probes for the array.

Methods. The manufacture of microarrays relies on [micro](#) and [nanofabrication](#) techniques. Microarrays for transcriptomics typically fall into one of the following two broad categories: low-density spotted arrays or high-density short probe arrays [36]. Transcript presence may be recorded with single- or dual-channel detection of fluorescent tags.

Spotted low-density arrays typically feature [picolitre](#) drops of a range of purified [cDNAs](#) arrayed on the surface of a glass slide [52]. The probes are longer than those of high-density

arrays and typically lack the transcript resolution of high-density arrays. Spotted arrays use different [fluorophores](#) for test and control samples, and the ratio of fluorescence is used to calculate a relative measure of abundance [53]. High-density arrays use single channel detection, and each sample is hybridised and detected individually [54]. High-density arrays were popularised by the [Affymetrix](#) GeneChip array (Santa Clara, CA), in which each transcript is quantified by several short 25-[mer](#) probes that together [assay](#) one gene [55].

NimbleGen arrays (Pleasanton, CA) are high-density arrays produced by a [maskless-photochemistry](#) method, which permits flexible manufacture of arrays in small or large numbers. These arrays have hundreds of thousands of 45- to 85-mer probes and are hybridised with a one-colour labelled sample for expression analysis [56]. Some designs incorporate up to 12 independent arrays per slide.

RNA-Seq

Principles and advances. [RNA-Seq](#) refers to the combination of a [high-throughput sequencing](#) methodology with computational methods to capture and quantify transcripts present in an RNA extract ([Fig 4](#)) [10]. The nucleotide sequences generated are typically around 100 bp in length, but can range from 30 bp to over 10,000 bp, depending on the sequencing method used. RNA-Seq leverages [deep sampling](#) of the transcriptome with many short fragments from a transcriptome to allow computational reconstruction of the original RNA transcript by [aligning](#) reads to a reference genome or to each other ([de novo assembly](#)) [9]. The typical dynamic range of 5 [orders of magnitude](#) for RNA-Seq is a key advantage over microarray transcriptomes. In addition, input RNA amounts are much lower for RNA-Seq (nanogram quantity) compared to microarrays (microgram quantity), which allowed finer examination of cellular structures, down to the single-cell level when combined with linear amplification of cDNA [25]. Theoretically, there is no upper limit of quantification in RNA-Seq, and background signal is very low for 100 bp reads in nonrepetitive regions [10].

RNA-Seq may be used to identify genes within a [genome](#) or identify which genes are active at a particular point in time, and read counts can be used to accurately model the relative gene expression level. RNA-Seq methodology has constantly improved, primarily through the development of DNA sequencing technologies to increase throughput, accuracy, and read length [57]. Since the first descriptions in 2006 and 2008 [40][58], RNA-Seq has been rapidly adopted and overtook microarrays as the dominant transcriptomics technique in 2015 [59].

The quest for transcriptome data at the level of individual cells has driven advances in RNA-Seq library preparation methods, resulting in dramatic advances in sensitivity. Single-cell transcriptomes are now well described and have even been extended to [in situ](#) RNA-Seq where transcriptomes of individual cells are directly interrogated in [fixed](#) tissues [60].

Methods. RNA-Seq was established in concert with the rapid development of a range of high-throughput DNA sequencing technologies [61]. However, before the extracted RNA transcripts are sequenced, several key processing steps are performed. Methods differ in the use of transcript enrichment, fragmentation, amplification, single or paired-end sequencing, and whether to preserve strand information.

The sensitivity of an RNA-Seq experiment can be increased by enriching classes of RNA that are of interest and depleting known abundant RNAs. The mRNA molecules can be separated by using oligonucleotide probes which bind their [poly-A tails](#). Alternatively, ribo-depletion can be used to specifically remove abundant but uninformative [ribosomal RNAs](#) (rRNAs) by hybridisation to probes tailored to the [taxon's](#) specific rRNA sequences (e.g., mammal rRNA, plant rRNA). However, ribo-depletion can also introduce some bias via nonspecific

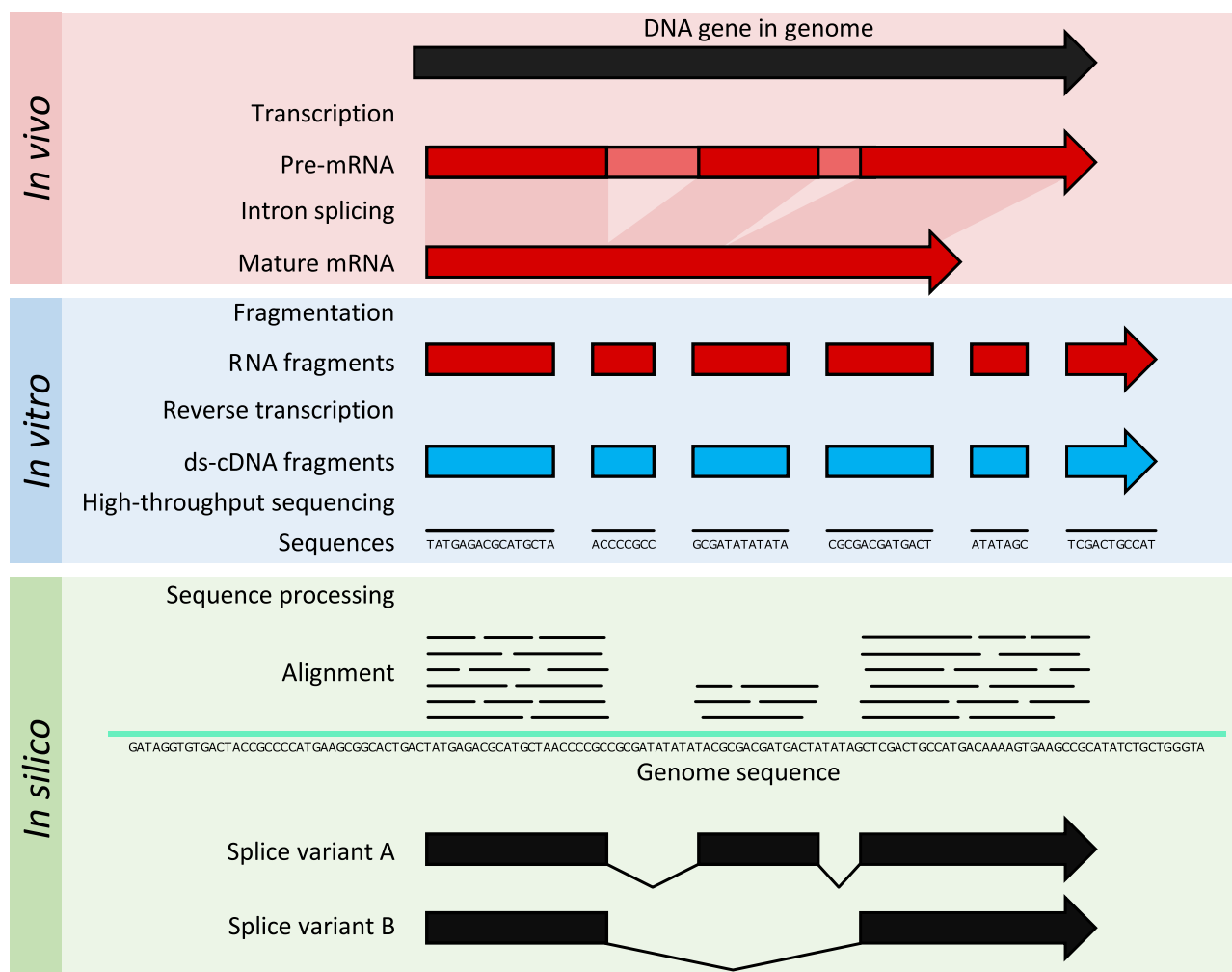


Fig 4. Summary of RNA sequencing. Within the organisms, genes are transcribed and spliced (in eukaryotes) to produce mature mRNA transcripts (red). The mRNA is extracted from the organism, fragmented and copied into stable double-stranded cDNA (ds-cDNA; blue). The ds-cDNA is sequenced using high-throughput, short-read sequencing methods. These sequences can then be aligned to a reference genome sequence to reconstruct which genome regions were being transcribed. These data can be used to annotate where expressed genes are, their relative expression levels, and any alternative splice variants.

<https://doi.org/10.1371/journal.pcbi.1005457.g004>

depletion of off-target transcripts [62]. Small RNAs such as [microRNAs](#), can be purified based on their size by [gel electrophoresis](#) and extraction.

Because mRNAs are longer than the read-lengths of typical high-throughput sequencing methods, transcripts are usually fragmented prior to sequencing. The fragmentation method is a key aspect of sequencing library construction [63]. It may incorporate chemical [hydrolysis](#), [nebulisation](#), or [sonication](#) of RNA, or utilise simultaneous [fragmentation](#), and tagging of cDNA by [transposase enzymes](#).

During preparation for sequencing, cDNA copies of transcripts may be amplified by [PCR](#) to enrich for fragments that contain the expected 5' and 3' adapter sequences [64]. Amplification is also used to allow sequencing of very low-input amounts of RNA, down to as little as 50 [pg](#) in extreme applications [65]. Spike-in controls can be used to provide quality control assessment of library preparation and sequencing, in terms of [guanine-cytosine content](#), fragment length, as well as the bias due to fragment position within a transcript [66]. [Unique molecular](#)

[identifiers](#) (UMIs) are short random sequences that are used to individually tag sequence fragments during library preparation so that every tagged fragment is unique [67]. UMIs provide an absolute scale for quantification and the opportunity to correct for subsequent amplification bias introduced during library construction and accurately estimate the initial sample size. UMIs are particularly well-suited to single-cell RNA-Seq transcriptomics, in which the amount of input RNA is restricted and extended amplification of the sample is required [68][69][70].

Once the transcript molecules have been prepared, they can be sequenced in just one direction (single-end) or both directions (paired-end). A single-end sequence is usually quicker to produce, cheaper than paired-end sequencing, and sufficient for quantification of gene expression levels. Paired-end sequencing produces more robust alignments and/or assemblies, which is beneficial for gene annotation and transcript [isoform](#) discovery [10]. Strand-specific RNA-Seq methods preserve the [strand](#) information of a sequenced transcript [71]. Without strand information, reads can be aligned to a gene locus, but do not inform in which direction the gene is transcribed. Stranded-RNA-Seq is useful for deciphering transcription for [genes that overlap](#) in different directions, and to make more robust gene predictions in nonmodel organisms [71].

Currently, RNA-Seq relies on copying of RNA molecules into cDNA molecules prior to sequencing; hence, the subsequent platforms are the same for transcriptomic and genomic data (Table 2). Consequently, the development of DNA sequencing technologies has been a defining feature of RNA-Seq [73][75][76]. Direct sequencing of RNA using [nanopore sequencing](#) represents a current state-of-the-art RNA-Seq technique in its infancy (in pre-release [beta testing](#) as of 2016) [77][78]. However, nanopore sequencing of RNA can detect [modified bases](#) that would be otherwise masked when sequencing cDNA and also eliminates [amplification](#) steps that can otherwise introduce bias [11][79].

The sensitivity and accuracy of an RNA-Seq experiment are dependent on the [number of reads](#) obtained from each sample. A large number of reads are needed to ensure sufficient coverage of the transcriptome, enabling detection of low abundance transcripts. Experimental design is further complicated by sequencing technologies with a limited output range, the variable efficiency of sequence creation, and variable sequence quality. Added to those considerations is that every species has a different [number of genes](#) and therefore requires a tailored sequence yield for an effective transcriptome. Early studies determined suitable thresholds empirically, but as the technology matured, suitable coverage is predicted computationally by transcriptome saturation. Somewhat counterintuitively, the most effective way to improve detection of differential expression in low expression genes is to add more [biological replicates](#), rather than adding

Table 2. Sequencing technology platforms commonly used for RNA-Seq [72][73].

Platform (Manufacturer)	Commercial release	Typical read length	Maximum throughput per run	Single read accuracy	RNA-Seq runs deposited in the NCBI SRA (Oct 2016) [74]
454 (Roche, Basel, Switzerland)	2005	700 bp	0.7 Gbp	99.9%	3548
Illumina (Illumina, San Diego, CA, USA)	2006	50–300 bp	900 Gbp	99.9%	362903
SOLiD (Thermo Fisher Scientific, Waltham, MA, USA)	2008	50 bp	320 Gbp	99.9%	7032
Ion Torrent (Thermo Fisher Scientific, Waltham, MA, USA)	2010	400 bp	30 Gbp	98%	1953
PacBio (Pacbio, Menlo Park, CA, USA)	2011	10,000 bp	2 Gbp	87%	160

NCBI, National Center for Biotechnology Information; SRA, Sequence Read Archive; RNA-Seq, RNA sequencing.

<https://doi.org/10.1371/journal.pcbi.1005457.t002>

more reads [80]. The current benchmarks recommended by the [Encyclopedia of DNA Elements](#) (ENCODE) Project are for 70-fold exome coverage for standard RNA-Seq and up to 500-fold exome coverage to detect rare transcripts and isoforms [81][82][83].

Data analysis

Transcriptomics methods are highly parallel and require significant computation to produce meaningful data for both microarray and RNA-Seq experiments. Microarray data are recorded as [high-resolution](#) images, requiring [feature detection](#) and spectral analysis. Microarray raw image files are each about 750 MB in size, while the processed intensities are around 60 MB in size. Multiple short probes matching a single transcript can reveal details about the [intron-exon](#) structure, requiring statistical models to determine the authenticity of the resulting signal. RNA-Seq studies can produce $>10^9$ of short DNA sequences, which must be aligned to [reference genomes](#) comprised of millions to billions of base pairs. [De novo assembly of reads](#) within a dataset requires the construction of highly complex [sequence graphs](#). RNA-Seq operations are highly repetitious and benefit from [parallelised computation](#), but modern algorithms mean consumer computing hardware is sufficient for simple transcriptomics experiments that do not require de novo assembly of reads. A human transcriptome could be accurately captured by using RNA-Seq with 30 million 100 bp sequences per sample [84][85]. This example would require approximately 1.8 gigabytes of disk space per sample when stored in a compressed fastq format. Processed count data for each gene would be much smaller, equivalent to processed microarray intensities. Sequence data may be stored in public repositories, such as the [Sequence Read Archive](#) (SRA) [86]. RNA-Seq datasets can be uploaded via the Gene Expression Omnibus.

Image processing

Microarray [image processing](#) must correctly identify the [regular grid](#) of features within an image and independently quantify the fluorescence [intensity](#) for each feature ([Fig 5](#)). [Image](#)

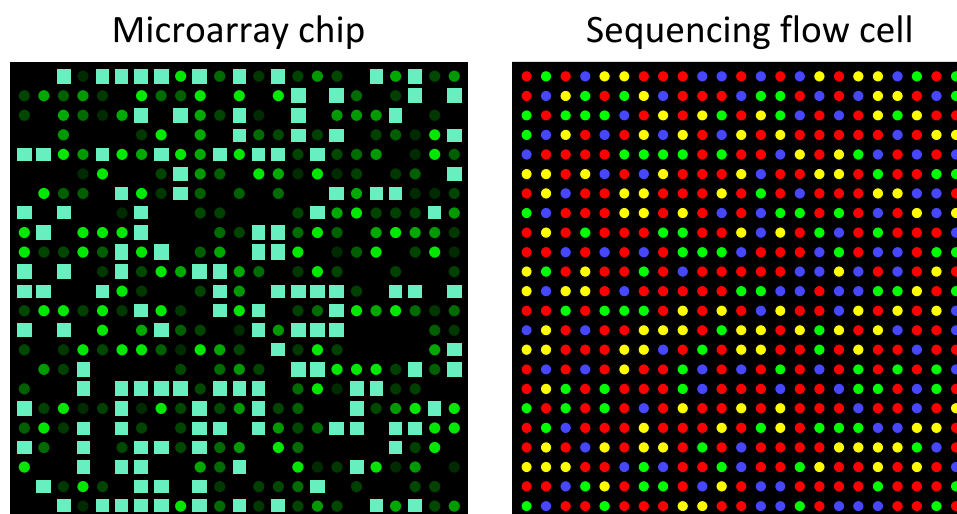


Fig 5. Microarray and sequencing flow cell. Microarrays and RNA sequencing (RNA-Seq) rely on image analysis in different ways. In a microarray chip, each spot on a chip is a defined oligonucleotide probe, and fluorescence intensity directly detects the abundance of a specific sequence (Affymetrix, Santa Clara, CA). In a high-throughput sequencing flow cell, spots are sequenced one nucleotide at a time, with the colour at each round indicating the next nucleotide in the sequence (Illumina HiSeq, San Diego, CA). Other variations of these techniques use more or fewer colour channels.

<https://doi.org/10.1371/journal.pcbi.1005457.g005>

[artefacts](#) must be additionally identified and removed from the overall analysis [87]. Fluorescence intensities directly indicate the abundance of each sequence because the sequence of each probe on the array is already known.

The first steps of RNA-seq also include similar image processing, however conversion of images to sequence data is typically handled automatically by the instrument software. The Illumina sequencing-by-synthesis method results in a random or ordered array of clusters distributed over the surface of a flow cell. The flow cell is imaged up to four times during each sequencing cycle, with tens to hundreds of cycles in total. Flow cell clusters are analogous to microarray spots and must be correctly identified during the early stages of the sequencing process. In Roche's [Pyrosequencing](#) method, the intensity of emitted light determines the number of consecutive nucleotides in a homopolymer repeat. There are many variants on these methods, each with a different error profile for the resulting data [88].

RNA-Seq data analysis

RNA-Seq experiments generate a large volume of raw sequence reads, which have to be processed to yield useful information. Data analysis usually requires a combination of [bioinformatics software](#) tools that vary according to the experimental design and goals. The process can be broken down into the following four stages: quality control, alignment, quantification, and differential expression [89]. Most popular RNA-Seq programs are run from a [command-line interface](#), either in a [Unix](#) environment or within the [R/Bioconductor](#) statistical environment [90].

Quality control. Sequence reads are not perfect, so the accuracy of each base in the sequence needs to be estimated for downstream analyses. Raw data are examined for high quality scores for base calls, guanine-cytosine content matches the expected distribution, the over representation of particularly short sequence motifs ([k-mers](#)), and an unexpectedly high read duplication rate [85]. Several options exist for sequence quality analysis, including the FastQC and FaQCs software packages [91][92]. Abnormalities identified may be removed by trimming or tagged for special treatment during later processes.

Alignment. In order to link sequence read abundance to expression of a particular gene, transcript sequences are [aligned](#) to a reference genome, or [de novo aligned](#) to one another if no reference is available. The key challenges for [alignment software](#) include sufficient speed to permit $>10^9$ of short sequences to be aligned in a meaningful timeframe, flexibility to recognise and deal with intron splicing of eukaryotic mRNA, and correct assignment of reads that map to multiple locations. Software advances have greatly addressed these issues, and increases in sequencing read length are further reducing multimapping reads. A list of currently available high-throughput sequence aligners is maintained by the [EBI](#) [93][94].

Alignment of [primary transcript mRNA](#) sequences derived from [eukaryotes](#) to a reference genome requires specialised handling of [intron](#) sequences, which are absent from mature mRNA. Short read aligners perform an additional round of alignments specifically designed to identify [splice junctions](#), informed by canonical splice site sequences and known intron splice site information. Identification of intron splice junctions prevents reads from being misaligned across splice junctions or erroneously discarded, allowing for more reads to be aligned to the reference genome and improving the accuracy of gene expression estimates. Because [gene regulation](#) may occur at the [mRNA isoform](#) level, splice-aware alignments also permit detection of isoform abundance changes that would otherwise be lost in a bulked analysis [95].

De novo assembly can be used to align reads to one another to construct full-length transcript sequences without the use of a reference genome ([Table 3](#)) [96]. Challenges particular to de novo assembly include larger computational requirements compared to a reference-based

Table 3. RNA-Seq de novo assembly software.

Software (Manufacturer)	Released	Last Updated	Resource load	Strengths and weaknesses
Velvet-Oases [100][101]	2008	2011	Heavy	The original short read assembler, now largely superseded.
SOAPdenovo-trans [102]	2011	2015	Moderate	Early short read assembler, updated for transcript assembly.
Trans-ABYSS [103]	2010	2016	Moderate	Short reads, large genomes, MPI-parallel version available.
Trinity [104][105]	2011	2017	Moderate	Short reads, large genomes, memory intensive.
miraEST [106]	1999	2016	Moderate	Repetitive sequences, hybrid data input, wide range of sequence platforms accepted.
Newbler [107]	2004	2012	Heavy	Specialised for Roche 454 sequence, homo-polymer error handling.
CLC genomics workbench (Qiagen; Venlo, Netherlands) [108]	2008	2014	Light	Graphical user interface, hybrid data.

MPI, Message Passing Interface; RNA-Seq, RNA sequencing.

<https://doi.org/10.1371/journal.pcbi.1005457.t003>

transcriptome, additional validation of gene variants or fragments, additional annotation of assembled transcripts. The first metrics used to describe transcriptome assemblies, such as [N50](#), have been shown to be misleading [97], and subsequently improved evaluation methods are now available [98][99]. Annotation-based metrics are better assessments of assembly completeness, such as contig reciprocal best hit count. Once assembled de novo, the assembly can be used as a reference for subsequent sequence alignment methods and quantitative gene expression analysis.

Quantification. Quantification of sequence alignments may be performed at the gene, exon, or transcript level. Typical outputs include a table of reads counts for each feature supplied to the software, for example, for genes in a [general feature format](#) file. Gene and exon read counts may be calculated easily using the HTSeq software package, for example [109]. Quantitation at the transcript level is more complicated and requires probabilistic methods to estimate transcript isoform abundance from short read information, for example, using cufflinks software [95]. Reads that align equally well to multiple locations must be identified and either removed, aligned to one of the possible locations, or aligned to the most probable location.

Some quantification methods can circumvent the need for an exact alignment of a read to a reference sequence all together. The kallisto method combines pseudoalignment and quantification into a single step that runs 2 orders of magnitude faster than comparable methods such as tophat/cufflinks, with less computational burden [110].

Differential expression. Once quantitative counts of each transcript are available, [differential gene expression](#) is then measured by normalising, modelling, and statistically analysing the data ([Fig 6](#)). Examples of dedicated software are described in [Table 4](#). Most read a table of genes and read counts as their input, but some, such as cuffdiff, will accept [binary alignment map](#) format read alignments as input. The final outputs of these analyses are gene lists with associated pair-wise tests for differential expression between treatments and the probability estimates of those differences.

Validation

Transcriptomic analyses may be validated using an independent technique, for example, [quantitative PCR](#) (qPCR), which is recognisable and statistically assessable [115]. Gene expression is measured against defined standards both for the gene of interest and [control](#) genes. The

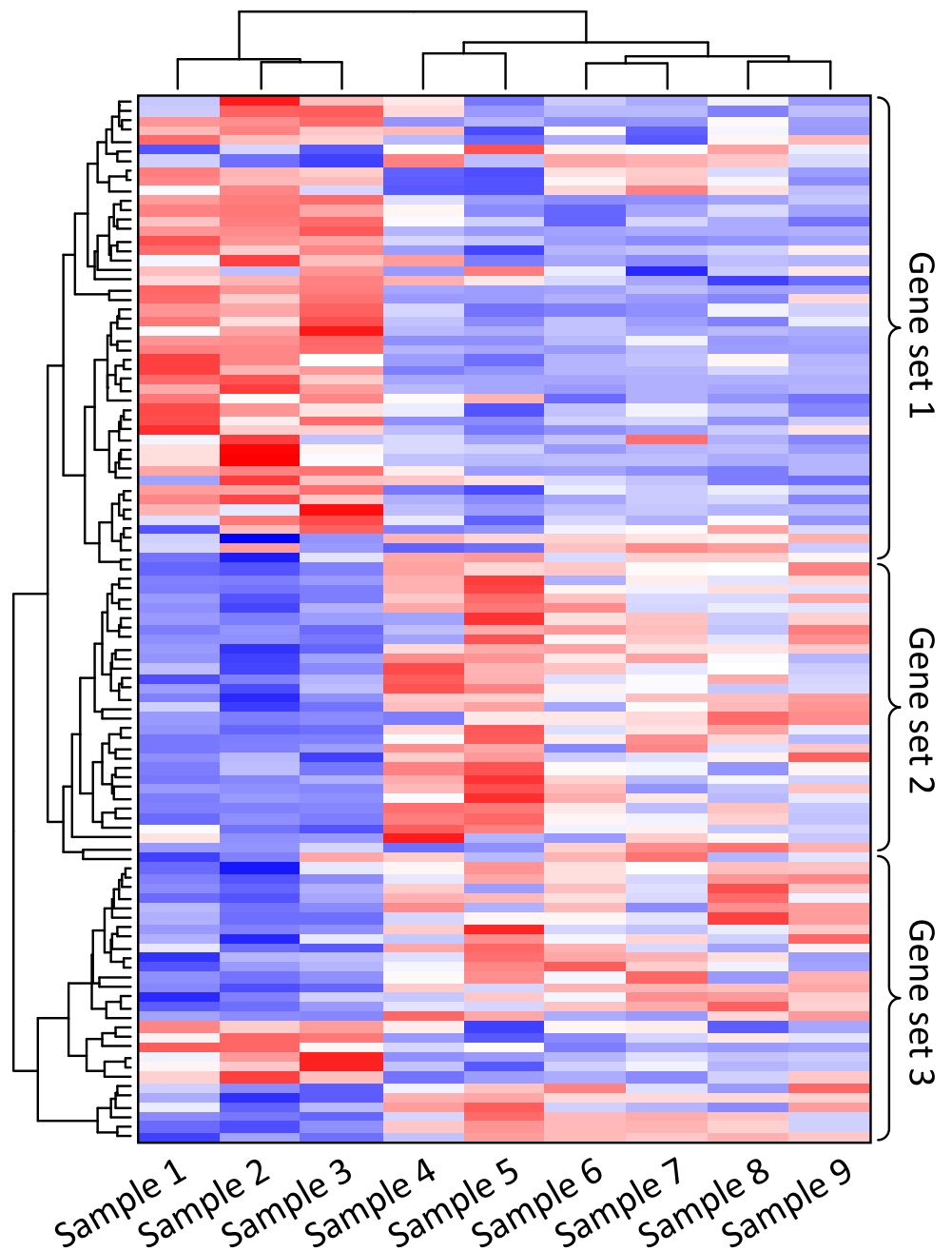


Fig 6. identification of gene co-expression patterns across different samples. **Heatmap** Each column contains the measurements for gene expression change for a single sample. Relative gene expression is indicated by colour: high-expression (red), median-expression (white) and low-expression (blue). Genes and samples with similar expression profiles can be automatically grouped (left and top trees). Samples may be different individuals, tissues, environments, or health conditions. In this example, expression of gene set 1 is high and expression of gene set 2 is low in samples 1, 2, and 3.

<https://doi.org/10.1371/journal.pcbi.1005457.g006>

Table 4. RNA-Seq differential gene expression software.

Software	Environment	Specialisation
Cuffdiff2 [111]	Unix-based	Transcript analysis at isoform-level
EdgeR [112]	R/Bioconductor	Any count-based genomic data
DEseq2 [113]	R/Bioconductor	Flexible data types, low replication
Limma/Voom [114]	R/Bioconductor	Microarray or RNA-Seq data, isoform analysis, 2^k experimental design

RNA-Seq, RNA sequencing.

<https://doi.org/10.1371/journal.pcbi.1005457.t004>

measurement by qPCR is similar to that obtained by RNA-Seq wherein a value can be calculated for the concentration of a target region in a given sample. qPCR is, however, restricted to amplicons smaller than 300 bp, usually toward the 3' end of the coding region, avoiding the 3' untranslated region (3' UTR) [16]. If validation of transcript isoforms is required, an inspection of RNA-Seq read alignments should indicate where qPCR primers might be placed for maximum discrimination. The measurement of multiple control genes along with the genes of interest produces a stable reference within a biological context [117]. qPCR validation of RNA-Seq data has generally shown that different RNA-Seq methods are highly correlated [58] [118] [119].

Functional validation of key genes is an important consideration for post transcriptome planning. Observed gene expression patterns may be functionally linked to a phenotype by an independent knock-down/rescue study in the organism of interest.

Applications

Diagnostics and disease profiling

Transcriptomic strategies have seen broad application across diverse areas of biomedical research, including disease diagnosis and profiling [10]. RNA-Seq approaches have allowed for the large-scale identification of transcriptional start sites and uncovered alternative promoter usage and novel splicing alterations. These regulatory elements are important in human disease, and therefore, defining such variants is crucial to the interpretation of disease-association studies [120]. RNA-Seq can also identify disease-associated single nucleotide polymorphisms (SNP), allele-specific expression, and gene fusions, contributing to our understanding of disease causal variants [121].

Retrotransposons are transposable elements which proliferate within eukaryotic genomes through a process involving reverse transcription. RNA-Seq can provide information about the transcription of endogenous retrotransposons that may influence the transcription of neighbouring genes by various epigenetic mechanisms that lead to disease [122]. Similarly, the potential for using RNA-Seq to understand immune-related disease is expanding rapidly due to the ability to dissect immune cell populations and to sequence T cell and B cell receptor repertoires from patients [123] [124].

Human and pathogen transcriptomes

RNA-Seq of human pathogens has become an established method for quantifying gene expression changes, identifying novel virulence factors, predicting antibiotic resistance, and unveiling host-pathogen immune interactions [125] [126]. A primary aim of this technology is to develop optimised infection control measures and targeted, individualised treatment [124].

Transcriptomic analysis has predominantly focused on either the host or the pathogen. Dual RNA-Seq has recently been applied to simultaneously profile RNA expression in both the

pathogen and host throughout the infection process. This technique enables the study of the dynamic response and interspecies [gene regulatory networks](#) in both interaction partners from initial contact through to invasion and the final persistence of the pathogen or clearance by the host immune system [127][128].

Responses to environment

Transcriptomics allows for the identification of genes and [pathways](#) that respond to and counteract [biotic](#) and [abiotic environmental stresses](#). The nontargeted nature of transcriptomics allows for the identification of novel transcriptional networks in complex systems. For example, comparative analysis of a range of [chickpea](#) lines at different developmental stages identified distinct transcriptional profiles associated with [drought](#) and [salinity](#) stresses, including identifying the role of [transcript isoforms](#) of [Apetela 2](#) and [Ethylene-Responsive Element Binding Protein](#) (AP2-EREBP) [129]. Investigation of gene expression during [biofilm](#) formation by the [fungal](#) pathogen [Candida albicans](#) revealed a coregulated set of genes critical for biofilm establishment and maintenance [130].

Transcriptomic profiling also provides crucial information on mechanisms of [drug resistance](#). Analysis of over a thousand [Plasmodium falciparum](#) isolates identified that upregulation of the [unfolded protein response](#) and slower progression through the early stages of the asexual intraerythrocytic [developmental cycle](#) were associated with [artemisinin resistance](#) in isolates from [Southeast Asia](#) [131].

Gene function annotation

All transcriptomic techniques have been particularly useful in [identifying the functions of genes](#) and identifying those responsible for particular phenotypes. Transcriptomics of [Arabidopsis ecotypes](#) that [hyperaccumulate metals](#) correlated genes involved in [metal uptake](#), tolerance, and [homeostasis](#) with the phenotype [132]. Integration of RNA-Seq datasets across different tissues has been used to improve annotation of gene functions in commercially important organisms (e.g., [cucumber](#)) [133] or threatened species (e.g., [koala](#)) [134].

Assembly of RNA-Seq reads is not dependent on a [reference genome](#) [104], and it is so ideal for gene expression studies of nonmodel organisms with nonexistent or poorly developed genomic resources. For example, a database of SNPs used in [Douglas fir](#) breeding programs was created by de novo transcriptome analysis in the absence of a [sequenced genome](#) [135]. Similarly, genes that function in the development of cardiac, muscle, and nervous tissue in lobster were identified by comparing the transcriptomes of the various tissue types without use of a genome sequence [136]. RNA-Seq can also be used to identify previously unknown [protein coding regions](#) in existing sequenced genomes.

Noncoding RNA

Transcriptomics is most commonly applied to the mRNA content of the cell. However, the same techniques are equally applicable to noncoding RNAs that are not translated into a protein, but instead, have direct functions (e.g., roles in [protein translation](#), [DNA replication](#), [RNA splicing](#), and [transcriptional regulation](#)) [137][138][139][140]. Many of these noncoding RNAs affect disease states, including cancer, cardiovascular, and neurological diseases [141].

Transcriptome databases

Transcriptomics studies generate large amounts of data that has potential applications far beyond the original aims of an experiment. As such, raw or processed data may be deposited

Table 5. Transcriptomic databases.

Name	Host	Data	Description
Gene Expression Omnibus [142]	NCBI	Microarray RNA-Seq	First transcriptomics database to accept data from any source. Introduced MIAME and MINSEQE community standards that define necessary experiment metadata to ensure effective interpretation and repeatability [143][144].
ArrayExpress [145]	ENA	Microarray	Imports datasets from the Gene Expression Omnibus and accepts direct submissions. Processed data and experiment metadata are stored at ArrayExpress, while the raw sequence reads are held at the ENA. Complies with MIAME and MINSEQE standards [144] [145].
Expression Atlas [146]	EBI	Microarray RNA-Seq	Tissue-specific gene expression database for animals and plants. Displays secondary analyses and visualisation, such as functional enrichment of Gene Ontology terms, InterPro domains, or pathways. Links to protein abundance data where available.
Genevestigator [147]	Privately curated	Microarray RNA-Seq	Contains manual curations of public transcriptome datasets, focusing on medical and plant biology data. Individual experiments are normalised across the full database, to allow comparison of gene expression across diverse experiments. Full functionality requires licence purchase, with free access to a limited functionality.
RefEx [148]	DDBJ	All	Human, mouse, and rat transcriptomes from 40 different organs. Gene expression visualised as heatmaps projected onto 3D representations of anatomical structures.
NONCODE [149]	noncode.org	RNA-Seq	ncRNAs excluding tRNA and rRNA.

DDBJ, DNA Data Bank of Japan; EBI, European Bioinformatics Institute; ENA, European Nucleotide Archive; MIAME, Minimum Information About a Microarray Experiment; MINSEQE, Minimum Information about a high-throughput nucleotide SEQuencing Experiment; NCBI, National Center for Biotechnology Information; ncRNAs, noncoding RNAs; RNA-Seq, RNA sequencing.

<https://doi.org/10.1371/journal.pcbi.1005457.t005>

into [public databases](#) to ensure their utility for the broader scientific community (Table 5). For example, as of 2016, the Gene Expression Omnibus contained millions of experiments.

Conclusions

Transcriptomics has revolutionised our understanding of how genomes are expressed. Over the last three decades, new technologies have redefined what is possible to investigate, and integration with other omics technologies is giving an increasingly integrated view of the complexities of cellular life. The plummeting cost of transcriptomics studies have made them possible for small laboratories, and large-scale transcriptomics consortia are able to undertake experiments comparing transcriptomes of thousands of organisms, tissues, or environmental conditions. This trend is likely to continue as sequencing technologies improve.

Supporting information

S1 Text. Version history of the text file.
(XML)

S2 Text. Peer reviews and response to reviews.
(XML)

References

1. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, et al. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*. 1991; 252:1651–1656. PMID: [2047873](#)
2. Pan Q, Shai O, Lee LJ, Frey BJ & Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* 2008; 40:1413–1415. <https://doi.org/10.1038/ng.259> PMID: [18978789](#)

3. Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, et al. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*. 2008; 321:956–60. <https://doi.org/10.1126/science.1160342> PMID: 18599741
4. Lappalainen T, Sammeth M, Friedländer MR, Höfer PA, Monlong J, Rivas MA, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*. 2013; 501:506–11. <https://doi.org/10.1038/nature12531> PMID: 24037378
5. Melnik AV, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, et al. Human genomics. The human transcriptome across tissues and individuals. *Science*. 2015; 348:660–6. <https://doi.org/10.1126/science.1253555> PMID: 25954002
6. Sandberg R. Entering the era of single-cell transcriptomics in biology and medicine. *Nat. Methods*. 2014; 11:22–3. PMID: 24524133
7. Kolodziejczyk AA, Kim JK, Svensson V, Marioni JC & Teichmann SA. The technology and biology of single-cell RNA sequencing. *Mol. Cell*. 2015; 58:610–20. <https://doi.org/10.1016/j.molcel.2015.04.005> PMID: 26000846
8. McGettigan PA. Transcriptomics in the RNA-seq era. *Curr Opin Chem Biol*. 2013; 17:4–11. <https://doi.org/10.1016/j.cbpa.2012.12.008> PMID: 23290152
9. Wang Z, Gerstein M & Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet*. 2009; 10:57–63. <https://doi.org/10.1038/nrg2484> PMID: 19015660
10. Ozsolak F & Milos PM. RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet*. 2011; 12:87–98. <https://doi.org/10.1038/nrg2934> PMID: 21191423
11. Morozova O, Hirst M & Marra MA. Applications of new sequencing technologies for transcriptome analysis. *Annu Rev Genomics Hum Genet*. 2009; 10:135–54. <https://doi.org/10.1146/annurev-genom-082908-145957> PMID: 19715439
12. Medline trend: automated yearly statistics of PubMed results for any query. [Internet]. Alexandru Dan Corlan [cited 2017 Apr 27]. <http://dan.corlan.net/medline-trend.html>.
13. Sim GK, Kafatos FC, Jones CW, Koehler MD, Efstratiadis A & Maniatis T. Use of a cDNA library for studies on evolution and developmental expression of the chorion multigene families. *Cell*. 1979; 18:1303–12. PMID: 519770
14. Sutcliffe JG, Milner RJ, Bloom FE & Lerner RA. Common 82-nucleotide sequence unique to brain RNA. *Proc. Natl. Acad. Sci. U.S.A.* 1982; 79:4942–6. PMID: 6956902
15. Putney SD, Herlihy WC & Schimmel P. A new troponin T and cDNA clones for 13 different muscle proteins, found by shotgun sequencing. *Nature*. 1983; 302:718–21. PMID: 6687628
16. Marra MA, Hillier L & Waterston RH. Expressed sequence tags (ESTs): linking bridges between genomes. *Trends Genet*. 1998; 14:41–7. [https://doi.org/10.1016/S0168-9525\(97\)01355-3](https://doi.org/10.1016/S0168-9525(97)01355-3) PMID: 9448457
17. Alwine JC, Kemp DJ & Stark GR. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proc. Natl. Acad. Sci. U.S.A.* 1977; 74:5350–4. PMID: 414220
18. Becker-Andre J & Hahlbrock K. Absolute mRNA quantification using the polymerase chain reaction (PCR). A novel approach by a PCR aided transcript titration assay (PATTY). *Nucleic Acids Res*. 1989; 17:9437–41. PMID: 2479917
19. Picot DG, Mariage-Samson R, Fayein NA, Matingou C, Eveno E, Houlgatte R, et al. The Genexpress IMAGE knowledge base of the human brain transcriptome: a prototype integrated resource for functional and computational genomics. *Genome Res*. 1999; 9:1951–9. PMID: 10022985
20. Velculescu VE, Zhang L, Zhou W, Vogelstein J, Basrai MA, Bassett DE, et al. Characterization of the yeast transcriptome. *Cell*. 1997; 88:243–51. PMID: 9008165
21. Velculescu VE, Zhang L, Vogelstein B & Kinzler KW. Serial analysis of gene expression. *Science*. 1995; 270:484–9. PMID: 7570003
22. Audic S & Claverie JM. The significance of digital gene expression profiles. *Genome Res*. 1997; 7:986–9. PMID: 9331369
23. Mantione KJ, Kream RM, Kuzelova H, Ptacek R, Raboch J, Samuel JM, et al. Comparing bioinformatic gene expression profiling methods: microarray and RNA-Seq. *Med Sci Monit Basic Res*. 2014; 20:1381–4. <https://doi.org/10.12659/MSMBR.892101> PMID: 25149683
24. Zhao S, Fung-Leung WP, Bittner A, Ngo K & Liu X. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLoS ONE*. 2014; 9:e78644. <https://doi.org/10.1371/journal.pone.0078644> PMID: 24454679

25. Hashimshony T, Wagner F, Sher N & Yanai I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep.* 2012; 2:666–73. <https://doi.org/10.1016/j.celrep.2012.08.003> PMID: 22939981
26. Stears RL, Getts RC & Gullans SR. A novel, sensitive detection system for high-density microarrays using dendrimer technology. *Physiol. Genomics.* 2000; 3:93–102. PMID: 11015604
27. Illumina. RNA-Seq Data Comparison with Gene Expression Microarrays. *European Pharmaceutical Review.*
28. Black MB, Parks BB, Pluta L, Chu TM, Allen BC, Wolfinger RD, et al. Comparison of microarrays and RNA-seq for gene expression analyses of dose-response experiments *Toxicol. Sci.* 2014; 137:385–403.
29. Marioni JC, Mason CE, Mane SM, Stephens M & Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* 2008; 18:1509–17. <https://doi.org/10.1101/gr.079558.108> PMID: 18550803
30. SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat. Biotechnol.* 2014; 32:903–14. <https://doi.org/10.1038/nbt.2957> PMID: 25150838
31. Chen JJ, Hsueh HM, Delongchamp RR, Lin CJ & Tsai CA. Reproducibility of microarray data: a further analysis of microarray quality control (MAQC) data. *BMC Bioinformatics.* 2007; 8:412. <https://doi.org/10.1186/1471-2105-8-412> PMID: 17961233
32. Larkin JE, Frank BC, Gavras H, Sultana R & Quackenbush J. Independence and reproducibility across microarray platforms. *Nat. Methods.* 2005; 2:337–44. <https://doi.org/10.1038/nmeth757> PMID: 15846360
33. Nelson NJ. Microarrays have arrived: gene expression tool matures. *J. Natl. Cancer Inst.* 2001; 93:492–504. PMID: 11287436
34. Schena M, Shalon D, Davis RW & Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science.* 1995; 270:467–71. PMID: 7569999
35. Pozhitkov AE, Tautz D & Noble PA. Oligonucleotide microarrays: widely applied, poorly understood. *Brief Funct Genomic Proteomic.* 2007; 6:141–52. <https://doi.org/10.1093/bfgp/elm014> PMID: 17644526
36. Heller MJ. DNA microarray technology: devices, systems, and applications. *Annu Rev Biomed Eng.* 2002; 4:129–53. <https://doi.org/10.1146/annurev.bioeng.4.020702.153438> PMID: 12117754
37. Ambrose, McLachlan Geoffrey J., Do Kim-Anh, Christopher. *Analyzing Microarray Gene Expression Data.* Hoboken: John Wiley & Sons. 2005; ISBN:9780471726128.
38. Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, et al. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays *Nat. Biotechnol.* 2000; 18:630–4. <https://doi.org/10.1038/76469> PMID: 10835600
39. Meyers BC, Vu TH, Tej SS, Ghazal H, Matvienko M, Agrawal V, et al. Analysis of the transcriptional complexity of *Arabidopsis thaliana* by massively parallel signature sequencing. *Nat. Biotechnol.* 2004; 22:1006–12. <https://doi.org/10.1038/nbt992> PMID: 15247925
40. Bainbridge MN, Warren RL, Hirst M, Romanuik T, Zeng T, Go A, et al. Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genomics.* 2006; 7:246. <https://doi.org/10.1186/1471-2164-7-246> PMID: 17010196
41. Mortazavi A, Williams BA, McCue K, Schaeffer L & Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods.* 2008; 5:621–8. <https://doi.org/10.1038/nmeth.1226> PMID: 18516045
42. Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, et al. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature.* 2008; 453:1239–43. <https://doi.org/10.1038/nature07002> PMID: 18488015
43. Chomczynski P & Sacchi N. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Anal. Biochem.* 1987; 162:156–59. <https://doi.org/10.1006/abio.1987.9999> PMID: 2440339
44. Chomczynski P & Sacchi N. The single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction: twenty-something years on. *Nat Protoc.* 2006; 1:581–8. <https://doi.org/10.1038/nprot.2006.83> PMID: 17406285
45. Grillo M & Margolis FL. Use of reverse transcriptase polymerase chain reaction to monitor expression of intronless genes. *BioTechniques.* 1990; 9:262, 264, 266–7. PMID: 1699561
46. Bryant S & Manning DL. Isolation of messenger RNA. *Methods Mol. Biol.* 1998; 86:61–74. <https://doi.org/10.1385/0-89603-494-1:61> PMID: 9664454

47. Zhao W, He X, Hoadley KA, Parker JS, Hayes DN & Perou CM. Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. *BMC Genomics*. 2014; 15:419. <https://doi.org/10.1186/1471-2164-15-419> PMID: 24888378
48. Close TJ, Wanamaker SI, Caldo RA, Turner SM, Ashlock DA, Dickerson JA, et al. A new resource for cereal genomics: 22K barley GeneChip comes of age. *Plant Physiol*. 2004; 134:960. <https://doi.org/10.1104/pp.103.034462> PMID: 15020760
49. Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, et al. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. U.S.A.* 2003; 100:15776. <https://doi.org/10.1073/pnas.2136655100> PMID: 14663149
50. Romanov V, Davidoff SN, Miles AR, Grainger DW, Gale BK & Brooks BD. A critical comparison of protein microarray fabrication technologies. *Analyst*. 2014; 139:1303. <https://doi.org/10.1039/c3an01577g> PMID: 24479125
51. Barbulovic-Nad I, Lucente M, Sun Y, Zhang M, Wheeler AR & Bussmann M. Bio-microarray fabrication techniques: a review. *Crit. Rev. Biotechnol.* 2006; 26:237. <https://doi.org/10.1080/07388550600978358> PMID: 17095434
52. Auburn RP, Kreil DP, Meadows LA, Fischer B, Matilla SS & Russell S. Robotic spotting of cDNA and oligonucleotide microarrays. *Trends Biotechnol.* 2005; 23:374. <https://doi.org/10.1016/j.tibtech.2005.04.002> PMID: 15978318
53. Shalon D, Smith SJ & Brown PO. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res*. 1996; 6:639. PMID: 8796352
54. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* 1996; 14:1675. <https://doi.org/10.1038/nbt1296-1675> PMID: 9634850
55. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B & Speed TP. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res*. 2003; 31:e15. PMID: 12582260
56. Selzer RR, Richmond TA, Pofahl NJ, Green RD, Eis PS, Nair P, Brothman AR & Stallings RL. Analysis of chromosome breakpoints in neuroblastoma at sub-kilobase resolution using fine-tiling oligonucleotide array CGH. *Genes Chromosomes Cancer*. 2005; 44:305. <https://doi.org/10.1002/gcc.20243> PMID: 16075461
57. Tachibana Chris. Transcriptomics today: Microarrays, RNA-seq, and more. *Science*.
58. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*. 2008; 320:1344. <https://doi.org/10.1126/science.1158441> PMID: 18451266
59. Su Z, Fang H, Hong H, Shi L, Zhang W, Zhang W, et al. An investigation of biomarkers derived from legacy microarray data for their utility in the RNA-seq era. *Genome Biol*. 2014; 15:523. <https://doi.org/10.1186/s13059-014-0523-y> PMID: 25633159
60. Lee JH, Daugharthy ER, Scheiman J, Kalhor R, Yang JL, Ferrante TC, et al. Highly multiplexed sub-cellular RNA sequencing in situ. *Science*. 2014; 343:1360. <https://doi.org/10.1126/science.1250212> PMID: 24578530
61. Shendure J & Ji H. Next-generation DNA sequencing. *Nat. Biotechnol.* 2008; 26:1135. <https://doi.org/10.1038/nbt1486> PMID: 18846087
62. Lahens NF, Kavakli IH, Zhang R, Hayer K, Black MB, Dueck H, et al. IVT-seq reveals extreme bias in RNA sequencing. *Genome Biol*. 2014; 15:R86. <https://doi.org/10.1186/gb-2014-15-6-r86> PMID: 24981968
63. Knierim E, Lucke B, Schwarz JM, Schuelke M & Seelow D. Systematic comparison of three methods for fragmentation of long-range PCR products for next generation sequencing. *PLoS ONE*. 2011; 6:e28240; <https://doi.org/10.1371/journal.pone.0028240> PMID: 22140562
64. Parekh S, Ziegenhain C, Vieth B, Enard W & Hellmann I. The impact of amplification on differential expression analyses by RNA-seq. *Sci Rep*. 2016; 6:25533; <https://doi.org/10.1038/srep25533> PMID: 27156886
65. Shanker S, Paulson A, Edenberg HJ, Peak A, Perera A, Alekseyev YO, et al. Evaluation of commercially available RNA amplification kits for RNA sequencing using very low input amounts of total RNA. *J Biomol Tech*. 2015; 26:4. <https://doi.org/10.7171/jbt.15-2601-001> PMID: 25649271
66. Jiang L, Schlesinger F, Davis CA, Zhang Y, Li R, Salit M, et al. Synthetic spike-in standards for RNA-seq experiments. *Genome Res*. 2011; 21:1543. <https://doi.org/10.1101/gr.121095.111> PMID: 21816910

67. Kivioja T, Vähä-Aho A, Karlsson K, Bonke M, Enge M, Linnarsson S, et al. Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods*. 2011; 9:721–4. <https://doi.org/10.1038/nmeth.1778> PMID: 22101854
68. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods*. 2009; 6:377–82. <https://doi.org/10.1038/nmeth.1315> PMID: 19349980
69. Islam S, Zeisel A, Joost S, La Manno G, Zajac P, et al. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat. Methods*. 2014; 11:1631–8. <https://doi.org/10.1038/nmeth.2772> PMID: 24363023
70. Jaitin DA, Kenigsberg E, Keren-Shaul H, Elefant N, Paul F, Zaretsky I, et al. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science*. 2014; 343:776–9. <https://doi.org/10.1126/science.1247651> PMID: 24531970
71. Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, et al. Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat. Methods*. 2010; 7:709–15. <https://doi.org/10.1038/nmeth.1491> PMID: 20711195
72. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, et al. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*. 2012; 13:341. <https://doi.org/10.1186/1471-2164-13-341> PMID: 22827831
73. Liu L, Li Y, Li S, Hu N, He Y, Pong R, et al. Comparison of next-generation sequencing systems. *J. Biomed. Biotechnol.* 2012; 2012:251364. <https://doi.org/10.1155/2012/251364> PMID: 22829749
74. SRA. [Internet]. NCBI [cited 2017 April 27]. <https://www.ncbi.nlm.nih.gov/sra>.
75. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, et al. Performance comparison of benchtop high-throughput sequencing platforms. *Nat. Biotechnol.* 2012; 30:434–9. <https://doi.org/10.1038/nbt.2198> PMID: 22522955
76. Goodwin S, McPherson JD & McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* 2016; 17:333–51. <https://doi.org/10.1038/nrg.2016.49> PMID: 27184599
77. Garalde D, Snell E, Jachimowicz D, Heron A, Bruce Mark, Lloyd J, et al. Highly parallel direct RNA sequencing on an array of nanopores. *bioRxiv*. 2016.
78. Loman NJ, Quick J & Simpson JT. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. Methods*. 2015; 12:733–8. <https://doi.org/10.1038/nmeth.3444> PMID: 26076426
79. Ozsolak F, Platt AR, Jones DR, Reifenger JG, Sass LE, McInerney P, et al. Direct RNA sequencing. *Nature*. 2009; 461:814–18. <https://doi.org/10.1038/nature08390> PMID: 19776739
80. Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, et al. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol.* 2013; 14:R95. <https://doi.org/10.1186/gb-2013-14-9-r95> PMID: 24020486
81. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012; 489:57–74. <https://doi.org/10.1038/nature11247> PMID: 22955616
82. Sloan CA, Chan ET, Davidson JM, Malladi VS, Strattan JS, Hitz BC, et al. ENCODE data at the ENCODE portal. *Nucleic Acids Res.* 2016; 44:D726–33. <https://doi.org/10.1093/nar/gkv1160> PMID: 26527727
83. ENCODE: Encyclopedia of DNA Elements. [Internet]. ENCODE [cited 2017 Apr 27]. <http://www.encodeproject.org>.
84. Hart SN, Therneau TM, Zhang Y, Poland GA & Kocher JP. Calculating sample size estimates for RNA sequencing data. *J. Comput. Biol.* 2013; 20:970–8. <https://doi.org/10.1089/cmb.2012.0283> PMID: 23961961
85. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. *Genome Biol.* 2016; 17:13. <https://doi.org/10.1186/s13059-016-0881-8> PMID: 26813401
86. Kodama Y, Shumway M & Leinonen R. The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res.* 2012; 40:D54–6. <https://doi.org/10.1093/nar/gkr854> PMID: 22009675
87. Petrov A & Shams S. Microarray Image Processing and Quality Control. *The Journal of VLSI Signal Processing-Systems for Signal, Image, and Video Technology* 38. 2004;(3): 211–22.
88. Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, et al. Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res.* 2011; 39:e90. <https://doi.org/10.1093/nar/gkr344> PMID: 21576222

89. Van Verk MC, Hickman R, Pieterse CM & Van Wees SC. RNA-Seq: revelation of the messengers. *Trends Plant Sci.* 2013; 18:1751–1759. <https://doi.org/10.1016/j.tplants.2013.02.001> PMID: 23481128
90. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods.* 2015; 12:1151–1159. <https://doi.org/10.1038/nmeth.3252> PMID: 25633503
91. FastQC: a quality control tool for high throughput sequence data. [Internet]. Babraham Institute [cited 2017 Apr 27]. <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.
92. Lo CC & Chain PS. Rapid evaluation and quality control of next generation sequencing data with FaQCs. *BMC Bioinformatics.* 2014; 15:366. <https://doi.org/10.1186/s12859-014-0366-2> PMID: 25408143
93. HTS Mappers. [Internet]. European Bioinformatics Institute [cited 2017 Apr 27]. http://www.ebi.ac.uk/~nf/hts_mappers/.
94. Fonseca NA, Rung J, Brazma A & Marioni JC. Tools for mapping high-throughput sequencing data. *Bioinformatics.* 2012; 28:3169–3177. <https://doi.org/10.1093/bioinformatics/bts605> PMID: 23060614
95. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ & Pachter L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 2010; 28:511–518. <https://doi.org/10.1038/nbt.1621> PMID: 20436464
96. Miller JR, Koren S & Sutton G. Assembly algorithms for next-generation sequencing data. *Genomics.* 2010; 95:315–327. <https://doi.org/10.1016/j.ygeno.2010.03.001> PMID: 20211242
97. O'Neil ST & Emrich SJ. Assessing De Novo transcriptome assembly metrics for consistency and utility. *BMC Genomics.* 2013; 14:465. <https://doi.org/10.1186/1471-2164-14-465> PMID: 23837739
98. Smith-Unna R, Boursnell C, Patro R, Hibberd JM & Kelly S. TransRate: reference-free quality assessment of de novo transcriptome assemblies. *Genome Res.* 2016; 26:1134–1144. <https://doi.org/10.1101/gr.196469.115> PMID: 27252236
99. Li B, Fillmore N, Bai Y, Collins M, Thomson JA, Stewart R & Dewey CN. Evaluation of de novo transcriptome assemblies from RNA-Seq data. *Genome Biol.* 2014; 15:553. <https://doi.org/10.1186/s13059-014-0553-5> PMID: 25608678
100. Zerbino DR & Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 2008; 18:821–829. <https://doi.org/10.1101/gr.074492.107> PMID: 18349386
101. Schulz MH, Zerbino DR, Vingron M & Birney E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics.* 2012; 28:1086–1092. <https://doi.org/10.1093/bioinformatics/bts094> PMID: 22368243
102. Xie Y, Wu G, Tang J, Luo R, Patterson J, Liu S, et al. SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics.* 2014; 30:1660–1661. <https://doi.org/10.1093/bioinformatics/btu077> PMID: 24532719
103. Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, et al. De novo assembly and analysis of RNA-seq data. *Nat. Methods.* 2010; 7:909–912. <https://doi.org/10.1038/nmeth.1517> PMID: 20935650
104. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 2011; 29:644–646. <https://doi.org/10.1038/nbt.1883> PMID: 21572440
105. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc.* 2013; 8:1494–1501. <https://doi.org/10.1038/nprot.2013.084> PMID: 23845962
106. Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Müller WE, Wetter T, et al. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res.* 2004; 14:1147–1156. <https://doi.org/10.1101/gr.1917404> PMID: 15140833
107. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature.* 2005; 437:376–380. <https://doi.org/10.1038/nature03959> PMID: 16056220
108. Kumar S & Blaxter ML. Comparing de novo assemblers for 454 transcriptome data. *BMC Genomics.* 2010; 11:571. <https://doi.org/10.1186/1471-2164-11-571> PMID: 20950480
109. Anders S, Pyl PT & Huber W. HTSeq: a Python framework to work with high-throughput sequencing data. *Bioinformatics.* 2015; 31:166–169. <https://doi.org/10.1093/bioinformatics/btu638> PMID: 25260700
110. Bray NL, Pimentel H, Melsted P & Pachter L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* 2016; 34:525–527. <https://doi.org/10.1038/nbt.3519> PMID: 27043002

111. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL & Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.* 2013; 31:461–467. <https://doi.org/10.1038/nbt.2450> PMID: 23222703
112. Robinson MD, McCarthy DJ & Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics.* 2010; 26:139–140. <https://doi.org/10.1093/bioinformatics/btp616> PMID: 19910308
113. Love MI, Huber W & Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 2014; 15:550. <https://doi.org/10.1186/s13059-014-0550-8> PMID: 25516281
114. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W & Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 2015; 43:e47. <https://doi.org/10.1093/nar/gkv007> PMID: 25605792
115. Fang Z & Cui X. Design and validation issues in RNA-seq experiments. *Brief. Bioinformatics.* 2011; 12:280–287. <https://doi.org/10.1093/bib/bbr004> PMID: 21498551
116. Ramsköld D, Wang ET, Burge CB & Sandberg R. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput. Biol.* 2009; 5:e1000598; <https://doi.org/10.1371/journal.pcbi.1000598> PMID: 20011106
117. Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, et al. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol.* 2002; 3:RESEARCH0034; PMID: 12184808
118. Core LJ, Waterfall JJ & Lis JT. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science.* 2008; 322:1845–1848. <https://doi.org/10.1126/science.1162228> PMID: 19056941
119. Camarena L, Bruno V, Euskirchen G, Poggio S & Snyder M. Molecular mechanisms of ethanol-induced pathogenesis revealed by RNA-sequencing. *PLoS Pathog.* 2010; 6:e1000834; <https://doi.org/10.1371/journal.ppat.1000834> PMID: 20368969
120. Costa V, Aprile M, Esposito R & Ciccodicola A. RNA-Seq and human complex diseases: recent accomplishments and future perspectives. *Eur. J. Hum. Genet.* 2013; 21:1341–1342. <https://doi.org/10.1038/ejhg.2012.129> PMID: 22739340
121. Khurana E, Fu Y, Chakravarty D, Demichelis F, Rubin MA & Gerstein M. Role of non-coding sequence variants in cancer. *Nat. Rev. Genet.* 2016; 17:931–941. <https://doi.org/10.1038/nrg.2015.17> PMID: 26781813
122. Slotkin RK & Martienssen R. Transposable elements and the epigenetic regulation of the genome. *Nat. Rev. Genet.* 2007; 8:272–283. <https://doi.org/10.1038/nrg2072> PMID: 17363976
123. Proserpio V & Mahata B. Single-cell technologies to study the immune system. *Immunology.* 2016; 147:133–140. <https://doi.org/10.1111/imm.12553> PMID: 26551575
124. Byron SA, Van Keuren-Jensen KR, Engelthaler DM, Carpten JD & Craig DW. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nat. Rev. Genet.* 2016; 17:257–271. <https://doi.org/10.1038/nrg.2016.10> PMID: 26996076
125. Wu HJ, Wang AH & Jennings MP. Discovery of virulence factors of pathogenic bacteria. *Curr Opin Chem Biol.* 2008; 12:93–101. <https://doi.org/10.1016/j.cbpa.2008.01.023> PMID: 18284925
126. Suzuki S, Horinouchi T & Furusawa C. Prediction of antibiotic resistance by gene expression profiles. *Nat Commun.* 2014; 5:5792; <https://doi.org/10.1038/ncomms6792> PMID: 25517437
127. Westermann AJ, Gorski SA & Vogel J. Dual RNA-seq of pathogen and host. *Nat. Rev. Microbiol.* 2012; 10:618–630. <https://doi.org/10.1038/nrmicro2852> PMID: 22890146
128. Durmuş S, Çelik T, Özgül A & Guthke R. A review on computational systems biology of pathogen-host interactions. *Front Microbiol.* 2015; 6:235. <https://doi.org/10.3389/fmicb.2015.00235> PMID: 25914674
129. Garg R, Shankar R, Thakkar B, Kudapa H, Krishnamurthy L, Mantri N, et al. Transcriptome analyses reveal genotype- and developmental stage-specific molecular responses to drought and salinity stresses in chickpea. *Sci Rep.* 2016; 6:19228; <https://doi.org/10.1038/srep19228> PMID: 26759178
130. García-Sánchez S, Aubert S, Iraqui I, Janbon G, Ghigo JM & d'Enfert C. *Candida albicans* biofilms: a developmental state associated with specific and stable gene expression patterns. *Eukaryotic Cell.* 2004; 3:536–545. <https://doi.org/10.1128/EC.3.2.536-545.2004> PMID: 15075282
131. Mok S, Ashley EA, Ferreira PE, Zhu L, Lin Z, Yeo T, Chotivanich K, et al. Drug resistance. Population transcriptomics of human malaria parasites reveals the mechanism of artemisinin resistance. *Science.* 2015; 347:431–435. <https://doi.org/10.1126/science.1260403> PMID: 25502316
132. Verbruggen N, Hermans C & Schat H. Molecular mechanisms of metal hyperaccumulation in plants. *New Phytol.* 2009; 181:759–766. <https://doi.org/10.1111/j.1469-8137.2008.02748.x> PMID: 19192189

133. Li Z, Zhang Z, Yan P, Huang S, Fei Z & Lin K. RNA-Seq improves annotation of protein-coding genes in the cucumber genome. *BMC Genomics*. 2011; 12:540. <https://doi.org/10.1186/1471-2164-12-540> PMID: 22047402
134. Hobbs M, Pavasovic A, King AG, Prentis PJ, Eldridge MD, Chen Z, et al. A transcriptome resource for the koala (*Phascolarctos cinereus*): insights into koala retrovirus transcription and sequence diversity. *BMC Genomics*. 2014; 15:786. <https://doi.org/10.1186/1471-2164-15-786> PMID: 25214207
135. Howe GT, Yu J, Knaus B, Cronn R, Kolpak S, Dolan P, et al. A SNP resource for Douglas-fir: de novo transcriptome assembly and SNP detection and validation. *BMC Genomics*. 2013; 14:137. <https://doi.org/10.1186/1471-2164-14-137> PMID: 23445355
136. McGrath LL, Vollmer SV, Kaluziak ST & Ayers J. De novo transcriptome assembly for the lobster *Homarus americanus* and characterization of differential gene expression across nervous system tissues. *BMC Genomics*. 2016; 17:63. <https://doi.org/10.1186/s12864-016-2373-3> PMID: 26772543
137. Noller HF. Ribosomal RNA and translation. *Annu. Rev. Biochem.* 1991; 60:191-227. <https://doi.org/10.1146/annurev.bi.60.070191.001203> PMID: 1883196
138. Christov CP, Gardiner TJ, Szűcs D & Krude T. Functional requirement of noncoding Y RNAs for human chromosomal DNA replication. *Mol. Cell. Biol.* 2006; 26:6993-7004. <https://doi.org/10.1128/MCB.01060-06> PMID: 16943439
139. Kishore S & Stamm S. The snoRNA HBII-52 regulates alternative splicing of the serotonin receptor 2C. *Science*. 2006; 311:2301-2304. <https://doi.org/10.1126/science.1118265> PMID: 16357227
140. Hüttenhofer A, Schattner P & Polacek N. Non-coding RNAs: hope or hype?. *Trends Genet.* 2005; 21:289-297. <https://doi.org/10.1016/j.tig.2005.03.007> PMID: 15851066
141. Esteller M. Non-coding RNAs in human disease. *Nat. Rev. Genet.* 2011; 12:861-874. <https://doi.org/10.1038/nrg3074> PMID: 22094949
142. Edgar R, Domrachev M & Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 2002; 30:2071-2076. PMID: 11752295
143. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.* 2001; 29:365-371. <https://doi.org/10.1038/ng1201-365> PMID: 11726920
144. Brazma A. Minimum Information About a Microarray Experiment (MIAME) - success, failures, challenges. *ScientificWorldJournal*. 2009; 9:420-429. <https://doi.org/10.1100/tsw.2009.57> PMID: 19484163
145. Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, et al. ArrayExpress update - simplifying data submissions. *Nucleic Acids Res.* 2015; 43:D1113-D1118. <https://doi.org/10.1093/nar/gku1057> PMID: 25361974
146. Petryszak R, Keays M, Tang YA, Fonseca NA, Barrera E, Burdett T, et al. Expression Atlas update - an integrated database of gene and protein expression in humans, animals and plants. *Nucleic Acids Res.* 2016; 44:D746-D751. <https://doi.org/10.1093/nar/gkv1045> PMID: 26481351
147. Hruz T, Laule O, Szabo G, Wessendorp F, Bleuler S, Oertle L, et al. Genevestigator v3: a reference expression database for the meta-analysis of transcriptomes. *Adv Bioinformatics*. 2008; 2008:420747. <https://doi.org/10.1155/2008/420747> PMID: 19956698
148. Mitsuhashi N, Fujieda K, Tamura T, Kawamoto S, Takagi T & Okubo K. BodyParts3D: 3D structure database for anatomical concepts. *Nucleic Acids Res.* 2009; 37:D782-D785. <https://doi.org/10.1093/nar/gkn613> PMID: 18835852
149. Zhao Y, Li H, Fang S, Kang Y, Wu W, Hao Y, et al. NONCODE 2016: an informative and valuable data source of long non-coding RNAs. *Nucleic Acids Res.* 2016; 44:D203-D208. <https://doi.org/10.1093/nar/gkv1252> PMID: 26586799