WikiJournal of Science, 2021, 4(1):3
doi: 10.15347/wjs/2021.004
Review Article

A broad introduction to RNA-Seq

Felix Richter[1]

Abstract

RNA is a nucleic acid, like DNA, with many fundamental biological roles in when, where, and by how much genes are turned on. Classically, sections of DNA are copied to RNA which are decoded into proteins that carry out cellular functions, but RNAs also have many roles that fall outside of this framework. RNA-Seq is a technique that is used to obtain snapshots of this continuously changing RNA landscape within a cell, with broad applications across the life sciences from agriculture to medicine. RNA-Seq is typically used to analyze the amount of each gene's RNA in experimental samples (i.e., gene expression), as well as changes made during RNA processing (e.g., alternative splicing, editing, mutations, or fusions between multiple RNAs). RNA-Seq requires molecular biology and computational steps, which are described in this review. Recent advances in RNA-Seq include the ability to study single cells and entire single RNA molecules.

Keywords: microarrays, microRNA, transfer RNA, exon, intron

Introduction

RNA-Seq, named as an abbreviation of "RNA sequencing" and sometimes spelled **RNA-seq**, **RNAseq**, or **RNASeq**, uses next-generation sequencing (NGS) to reveal the presence and quantity of ribonucleic acid (RNA) in a biological sample at a given moment.[1][2]

RNA-Seq is used to analyze the continuously changing cellular transcriptome (**Figure 1**). Specifically, RNA-Seq facilitates the ability to look at alternative gene spliced transcripts, post-transcriptional modifications, gene fusion, mutations/single nucleotide polymorphisms (SNPs) and changes in gene expression over time, or differences in gene expression in different groups or treatments.[3] In addition to messenger RNA (mRNA) transcripts, RNA-Seq can look at different populations of RNA to include total RNA, small RNA, such as microRNA (miRNA), transfer RNA (tRNA), and ribosomal profiling.[4] RNA-Seq can also be used to determine exon/intron boundaries and verify or amend previously annotated 5' and 3' gene boundaries. Recent advances in RNA-Seq include single cell sequencing, in situ sequencing of fixed tissue, and native RNA molecule sequencing with single-molecule real-time sequencing.[5]

[1] Department of Biomedical Sciences, Address Icahn School of Medicine at Mount Sinai, New York, New York, United States
***Author correspondence**: felix.richter@icahn.mssm.edu
ORCID: 0000-0003-3429-9621
Supplementary material: *commons.wikimedia.org/Q100146647*
Licensed under: CC-BY
Received 15-07-2019; accepted 17-05-2021

Prior to RNA-Seq, gene expression studies were done with hybridization-based microarrays. Issues with microarrays include cross-hybridization artifacts, poor quantification of lowly and highly expressed genes, and needing to know the sequence *a priori*.[6] Because of these technical issues, transcriptomics transitioned to sequencing-based methods. These progressed from Sanger sequencing of Expressed Sequence Tag libraries, to chemical tag-based methods (e.g., serial analysis of gene expression), and finally to the current technology, next-gen sequencing of complementary DNA (cDNA), notably RNA-Seq.

Library preparation

See also: w:Library (biology)

The general steps to prepare a complementary DNA (cDNA) library for sequencing are described below, but often vary between platforms.[9][2][10]

1. *RNA Isolation:* RNA is isolated from tissue and mixed with deoxyribonuclease (DNase). DNase reduces the amount of genomic DNA. The amount of RNA degradation is checked with gel and capillary electrophoresis and is used to assign an RNA integrity number to the sample. This RNA quality and the total amount of starting RNA are taken into consideration during the subsequent library preparation, sequencing, and analysis steps.
2. *RNA selection/depletion:* To analyze signals of interest, the isolated RNA can either be kept as is, filtered for RNA with 3' polyadenylated (poly(A)) tails

to include only mRNA, depleted of ribosomal RNA (rRNA), and/or filtered for RNA that binds specific sequences (**RNA selection and depletion methods** table, below). The RNA with 3' poly(A) tails are mainly composed of mature, processed, coding sequences. Poly(A) selection is performed by mixing RNA with poly(T) oligomers covalently attached to a substrate, typically magnetic beads.[11][12] Poly(A) selection has important limitations in RNA biotype detection. Many RNA biotypes are not polyadenylated, including many noncoding RNA and histone-core protein transcripts, or are regulated via their poly(A) tail length (e.g., cytokines) and thus might not be detected after poly(A) selection.[13] Furthermore, poly(A) selection may increased 3' bias, especially with lower quality RNA.[14][15] These limitations can be avoided with ribosomal depletion, removing rRNA that typically represents over 90% of the RNA in a cell. Both poly(A) enrichment and ribosomal depletion steps are labor intensive and could introduce biases, so more simple approaches have been developed to omit these steps.[16] Small RNA targets, such as miRNA, can be further isolated through size selection with exclusion gels, magnetic beads, or commercial kits.

3. *cDNA synthesis:* RNA is reverse transcribed to cDNA because DNA is more stable and to allow for amplification (which uses DNA polymerases) and leverage more mature DNA sequencing technology. Amplification subsequent to reverse transcription results in loss of strandedness, which can be avoided with chemical labeling or single molecule sequencing. Fragmentation and size selection are performed to purify sequences that are the appropriate length for the sequencing machine. The RNA, cDNA, or both are fragmented with enzymes, sonication, or nebulizers. Fragmentation of the RNA reduces 5' bias of randomly primed-reverse transcription and the influence of primer binding sites,[12] with the downside that the 5' and 3' ends are converted to DNA less efficiently. Fragmentation is followed by size selection, where either small sequences are removed or a tight range of sequence lengths are selected. Because small RNAs like miRNAs are lost, these are analyzed independently. The cDNA for each experiment can be indexed with a hexamer or octamer barcode, so that these experiments can be pooled into a single lane for multiplexed sequencing.

Sequencing

See also: w:DNA sequencing

The cDNA library derived from RNA biotypes is then sequenced into a computer-readable format. There are many high-throughput sequencing technologies for cDNA sequencing including platforms developed by Illumina, Thermo Fisher, BGI/MGI, PacBio, and Oxford Nanopore Technologies.[17] For Illumina short-read sequencing, a common technology for cDNA sequencing, adapters are ligated to the cDNA, DNA is attached to a flow cell, clusters are generated through cycles of bridge amplification and denaturing, and sequence-by-synthesis is performed in cycles of complementary strand synthesis and laser excitation of bases with reversible terminators. Sequencing platform choice and parameters are guided by experimental design and cost. Common experimental design considerations include deciding on the sequencing length, sequencing depth, use of single versus paired-end sequencing, number of replicates, multiplexing, randomization, and spike-ins.[18]

Single-molecule real-time RNA sequencing

See also: w:Single-molecule real-time sequencing

Massively parallel single molecule direct RNA-Seq has been explored as an alternative to traditional RNA-Seq, in which RNA-to-cDNA conversion, ligation, amplifcation, and other sample manipulation steps may introduce biases and artifacts.[19] Technology platforms that perform single-molecule real-time RNA-Seq include Oxford Nanopore Technologies (ONT) Nanopore sequencing,[20] PacBio IsoSeq, and Helicos (bankrupt). Sequencing RNA in its native form preserves modifications like methylation, allowing them to be investigated directly and simultaneously.[21] Another benefit of single-molecule RNA-Seq is that transcripts can be covered in full length, allowing for higher confidence isoform detection and quantification compared to short-read sequencing. Traditionally, single-molecule RNA-Seq methods have higher error rates compared to short-read sequencing, but newer methods like ONT direct RNA-Seq limit errors by avoiding fragmentation and cDNA conversion. Recent uses of ONT direct RNA-Seq for differential expression in human cell populations have demonstrated that this technology can overcome many limitations of short and long cDNA sequencing.[22]

Single-cell RNA sequencing (scRNA-Seq)

See also: w:Single cell sequencing

Standard methods such as microarrays and standard bulk RNA-Seq analysis analyze the expression of RNAs from large populations of cells. In mixed cell populations, these measurements may obscure critical differences between individual cells within these populations.[23][24]

WikiJournal of Science, 2021, 4(1):3
doi: 10.15347/wjs/2021.004
Review Article

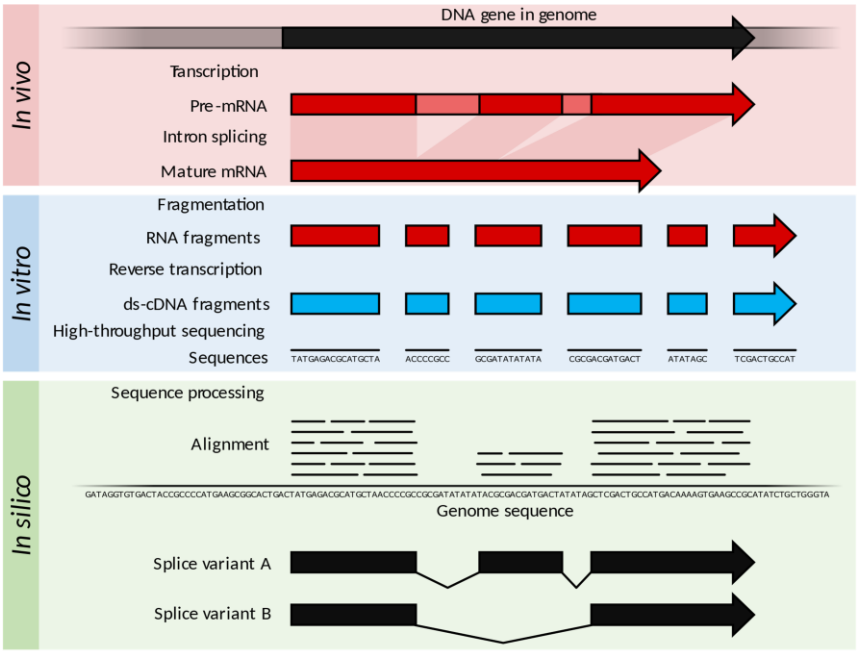

Figure 1 | *Summary of RNA-Seq.* Within the organism, genes are transcribed and (in a eukaryotic organism) spliced to produce mature mRNA transcripts (red). The mRNA is extracted from the organism, fragmented and reverse-transcribed into stable double-stranded (ds) cDNA (blue). The ds-cDNA is sequenced 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. This data can be used to annotate where expressed genes are, their relative expression levels, and any alternative splice variants.[7]
CC BY 4.0 Thomas Shafee - Own work

RNA selection and depletion methods:[9]

Strategy	Predominant type of RNA	Ribosomal RNA content	Unprocessed RNA content	Isolation method
Total RNA	All	High	High	None
PolyA selection	Coding	Low	Low	Hybridization with poly(dT) oligomers
rRNA depletion	Coding, noncoding	Low	High	Removal of oligomers complementary to rRNA
RNA capture	Targeted	Low	Moderate	Hybridization with probes complementary to desired transcripts

Single-cell RNA sequencing (scRNA-Seq) provides the expression profiles of individual cells. Although it is not possible to obtain complete information on every RNA expressed by each cell, due to the small amount of material available, patterns of gene expression can be identified through gene clustering analyses. This can uncover the existence of rare cell types within a cell population that may never have been seen before. For example, rare specialized cells in the lung called pulmonary ionocytes that express the Cystic Fibrosis Transmembrane Conductance Regulator were identified in 2018 by two groups performing scRNA-Seq on lung airway epithelia.[25][26]

Experimental procedures

Current scRNA-Seq protocols involve the following steps: isolation of single cell and RNA, reverse transcription (RT), amplification, library generation and sequencing (Figure 3). Single cells are either mechanically separated into microwells (e.g., BD Rhapsody, Takara ICELL8, Vycap Puncher Platform, or CellMicrosystems CellRaft) or encapsulated in droplets (e.g., 10x Genomics Chromium, Illumina Bio-Rad ddSEQ, 1CellBio InDrop, Dolomite Bio Nadia).[27] Single cells are labeled by adding beads with barcoded oligonucleotides; both cells and beads are supplied in limited amounts such

that co-occupancy with multiple cells and beads is a very rare event. Once reverse transcription is complete, the cDNAs from many cells can be mixed together for sequencing; transcripts from a particular cell are identified by each cell's unique barcode.[28][29] Unique molecular identifier (UMIs) can be attached to mRNA/cDNA target sequences to help identify artifacts during library preparation.[30]

Challenges for scRNA-Seq include preserving the initial relative abundance of mRNA in a cell and identifying rare transcripts.[31] The reverse transcription step is critical as the efficiency of the RT reaction determines how much of the cell's RNA population will be eventually analyzed by the sequencer. The processivity of reverse transcriptases and the priming strategies used may affect full-length cDNA production and the generation of libraries biased toward the 3′ or 5' end of genes.

In the amplification step, either PCR or in vitro transcription (IVT) is currently used to amplify cDNA. One of the advantages of PCR-based methods is the ability to generate full-length cDNA. However, different PCR efficiency on particular sequences (for instance, GC content and snapback structure) may also be exponentially amplified, producing libraries with uneven coverage. On the other hand, while libraries generated by IVT can avoid PCR-induced sequence bias, specific sequences may be transcribed inefficiently, thus causing sequence drop-out or generating incomplete sequences.[32][23] Several scRNA-Seq protocols have been published: Tang et al.,[33] STRT,[34] SMART-seq,[35] CEL-seq,[36] RAGE-seq,[37] Quartz-seq[38] and C1-CAGE.[39] These protocols differ in terms of strategies for reverse transcription, cDNA synthesis and amplification, and the possibility to accommodate sequence-specific barcodes (i.e. UMIs) or the ability to process pooled samples.[40]

In 2017, two approaches were introduced to simultaneously measure single-cell mRNA and protein expression through oligonucleotide-labeled antibodies known as REAP-seq,[41] and CITE-seq.[42]

Applications

scRNA-Seq is becoming widely used across biological disciplines including Development, Neurology,[43] Oncology,[44][45][46] Autoimmune disease,[47] and Infectious disease.[48]

scRNA-Seq has provided considerable insight into the development of embryos and organisms, including the worm *Caenorhabditis elegans*,[49] and the regenerative planarian *Schmidtea mediterranea*.[50][51] The first vertebrate animals to be mapped in this way were Zebrafish[52][53] and *Xenopus laevis*.[54] In each case multiple stages of the embryo were studied, allowing the entire process of development to be mapped on a cell-by-cell basis.[9] Science recognized these advances as the 2018 Breakthrough of the Year.[55]

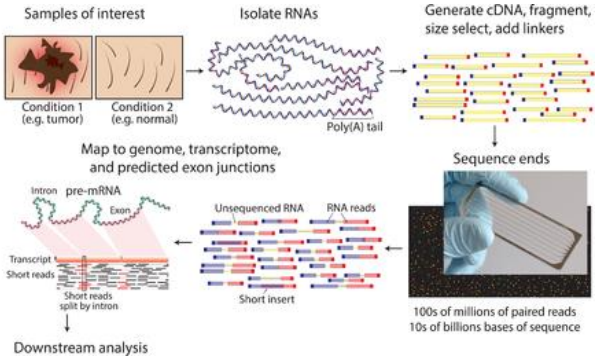

Figure 2 | *Typical RNA-Seq experimental workflow. RNA are isolated from multiple samples, converted to cDNA libraries, sequenced into a computer-readable format, aligned to a reference, and quantified for downstream analyses such as differential expression and alternative splicing.[8]*
Malachi Griffith et al., CC BY 2.5

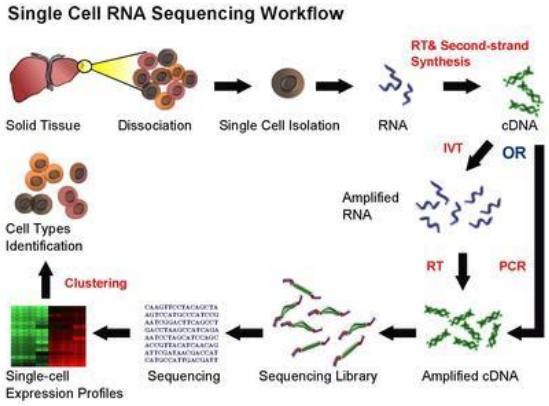

Figure 3 | *Typical single-cell RNA-Seq workflow. Single cells are isolated from a sample into either wells or droplets, cDNA libraries are generated and amplified, libraries are sequenced, and expression matrices are generated for downstream analyses like cell type identification. Yijyechern, CC BY-SA*

Experimental considerations

A variety of parameters are considered when designing and conducting RNA-Seq experiments:

- *Tissue specificity:* Gene expression varies within and between tissues, and RNA-Seq measures this mix of cell types. This may make it difficult to isolate the biological

mechanism of interest. Single cell sequencing can be used to study each cell individually, mitigating this issue.

- *Time dependence:* Gene expression changes over time, and RNA-Seq only takes a snapshot. Time course experiments can be performed to observe changes in the transcriptome.
- *Coverage (also known as depth):* RNA harbors the same mutations observed in DNA, and detection requires deeper coverage. With high enough coverage, RNA-Seq can be used to estimate the expression of each allele. This may provide insight into phenomena such as imprinting or cis-regulatory effects. The depth of sequencing required for specific applications can be extrapolated from a pilot experiment.[56]
- *Data generation artifacts (also known as technical variance):* The reagents (e.g., library preparation kit), personnel involved, and type of sequencer (e.g., Illumina, Pacific Biosciences) can result in technical artifacts that might be mis-interpreted as meaningful results. As with any scientific experiment, it is prudent to conduct RNA-Seq in a well controlled setting. If this is not possible or the study is a meta-analysis, another solution is to detect technical artifacts by inferring latent variables (typically principal component analysis or factor analysis) and subsequently correcting for these variables.[57]
- *Data management:* A single RNA-Seq experiment in humans is usually 1-5 Gb (compressed) or more when including intermediate files.[58] This large volume of data can pose storage issues. One solution is compressing the data using multi-purpose computational schemas (e.g., gzip) or genomics-specific schemas. The latter can be based on reference sequences or de novo. Another solution is to perform microarray experiments, which may be sufficient for hypothesis-driven work or replication studies (as opposed to exploratory research).

Results / Description

See also: w:List of RNA-Seq bioinformatics tools

An overview of RNA-Seq analysis techniques is illustrated in **Figure 4**.

Transcriptome assembly

See also: w:Sequence alignment software § Short-Read Sequence Alignment

Two methods are used to assign raw sequence reads to genomic features (i.e., assemble the transcriptome):

- *De novo:* This approach does not require a reference genome to reconstruct the transcriptome, and is typically used if the genome is unknown, incomplete, or substantially altered compared to the reference.[59] Challenges when using short reads for de novo assembly include 1) determining which reads should be joined together into contiguous sequences (contigs), 2) robustness to sequencing errors and other artifacts, and 3) computational efficiency. The primary algorithm used for de novo assembly transitioned from overlap graphs, which identify all pair-wise overlaps between reads, to de Bruijn graphs, which break reads into sequences of length k and collapse all k-mers into a hash table.[60] Overlap graphs were used with Sanger sequencing, but do not scale well to the millions of reads generated with RNA-Seq. Examples of assemblers that use de Bruijn graphs are Trinity,[59] Oases[61] (derived from the genome assembler Velvet[62]), Bridger,[63] and rnaSPAdes.[64] Paired-end and long-read sequencing of the same sample can mitigate the deficits in short read sequencing by serving as a template or skeleton. Metrics to assess the quality of a de novo assembly include median contig length, number of contigs and N50.[65]
- *Genome guided:* This approach relies on the same methods used for DNA alignment, with the additional complexity of aligning reads that cover non-continuous portions of the reference genome.[66] These non-continuous reads are the result of sequencing spliced transcripts (**Figure 5**). Typically, alignment algorithms have two steps: 1) align short portions of the read (i.e., seed the genome), and 2) use dynamic programming to find an optimal alignment, sometimes in combination with known annotations. Software tools that use genome-guided alignment include Bowtie,[67] TopHat (which builds on BowTie results to align splice junctions),[68][69] Subread,[70] STAR,[66] HISAT2,[71] and GMAP.[72] The output of genome guided alignment (mapping) tools can be further utilized by tools such as

Cufflinks[73] or StringTie[74] to reconstruct contiguous transcript sequences (*i.e.*, a FASTA file).The quality of a genome guided assembly can be measured with both 1) de novo assembly metrics (e.g., N50) and 2) comparisons to known transcript, splice junction, genome, and protein sequences using precision, recall, or their combination (e.g., F1 score).[65] In addition, *in silico* assessment could be performed using simulated reads.[75][76]

A note on assembly quality: The current consensus is that 1) assembly quality can vary depending on which metric is used, 2) assembly tools that scored well in one species do not necessarily perform well in the other species, and 3) combining different approaches might be the most reliable.[77][78][79]

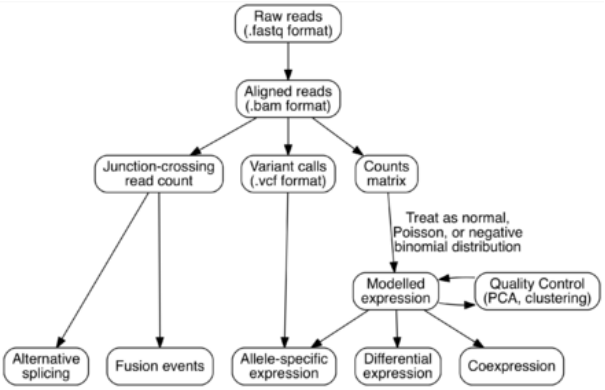

Figure 4 | *Diagram outlining a standard RNA-Seq analysis workflow.* Sequenced reads are aligned to a reference genome and/or transcriptome and subsequently processed for a variaty of quality control, discovery, and hypothesis-driven analyses. CC BY-SA 4.0 Salubrious Toxin - Own work

Gene expression quantification

Expression is quantified to study cellular changes in response to external stimuli, differences between healthy and diseased states, and other research questions. Transcript levels are often used as a proxy for protein abundance, but these are often not equivalent due to post transcriptional events such as RNA interference and nonsense-mediated decay.[80]

Expression is quantified by counting the number of reads that mapped to each locus in the transcriptome assembly step. Expression can be quantified for exons or genes using contigs or reference transcript annotations.[9] These observed RNA-Seq read counts have been robustly validated against older technologies, including expression microarrays and qPCR.[56][81] Tools that quantify counts are HTSeq,[82] FeatureCounts,[83] Rcount,[84] maxcounts,[85] FIXSEQ,[86] and Cuffquant. These tools determine read counts from aligned RNA-Seq data, but alignment-free counts can also be obtained with Sailfish[87] and Kallisto.[88] The read counts are then converted into appropriate metrics for hypothesis testing, regressions, and other analyses. Parameters for this conversion are:

- *Sequencing depth/coverage:* Although depth is pre-specified when conducting multiple RNA-Seq experiments, it will still vary widely between experiments.[89] Therefore, the total number of reads generated in a single experiment is typically normalized by converting counts to fragments, reads, or counts per million mapped reads (FPM, RPM, or CPM). The difference between RPM and FPM was historically derived during the evolution from single-end sequencing of fragments to paired-end sequencing. In single-end sequencing, there is only one read per fragment (*i.e.*, RPM = FPM). In paired-end sequencing, there are two reads per fragment (*i.e.*, RPM = 2 x FPM). Sequencing depth is sometimes referred to as library size, the number of intermediary cDNA molecules in the experiment.
- *Gene length:* Longer genes will have more fragments/reads/counts than shorter genes if transcript expression is the same. This is adjusted by dividing the FPM by the length of a feature (which can be a gene, transcript, or exon), resulting in the metric fragments per kilobase of feature per million mapped reads (FPKM).[90] When looking at groups of features across samples, FPKM is converted to transcripts per million (TPM) by dividing each FPKM by the sum of FPKMs within a sample.[91][92][93]
- *Total sample RNA output:* Because the same amount of RNA is extracted from each sample, samples with more total RNA will have less RNA per gene. These genes appear to have decreased expression, resulting in false positives in downstream analyses.[89]
- *Variance for each gene's expression:* is modeled to account for sampling error (important for genes with low read counts), increase power, and decrease false positives. Variance can be estimated as a normal, Poisson, or negative binomial distribution[94][95][96] and is frequently decomposed into technical and biological variance.

Spike-ins for absolute quantification and detection of genome-wide effects

RNA spike-ins are samples of RNA at known concentrations that can be used as gold standards in experimental design and during downstream analyses for absolute quantification and detection of genome-wide effects.

- *Absolute quantification:* Absolute quantification of gene expression is not possible with most RNA-Seq experiments, which quantify expression relative to all transcripts. It is possible by performing RNA-Seq with spike-ins, samples of RNA at known concentrations. After sequencing, read counts of spike-in sequences are used to determine the relationship between each gene's read counts and absolute quantities of biological fragments[12][97] In one example, this technique was used in *Xenopus tropicalis* embryos to determine transcription kinetics.[98]
- *Detection of genome-wide effects:* Changes in global regulators including chromatin remodelers, transcription factors (e.g., MYC), acetyltransferase complexes, and nucleosome positioning are not congruent with normalization assumptions and spike-in controls can offer precise interpretation.[99][100]

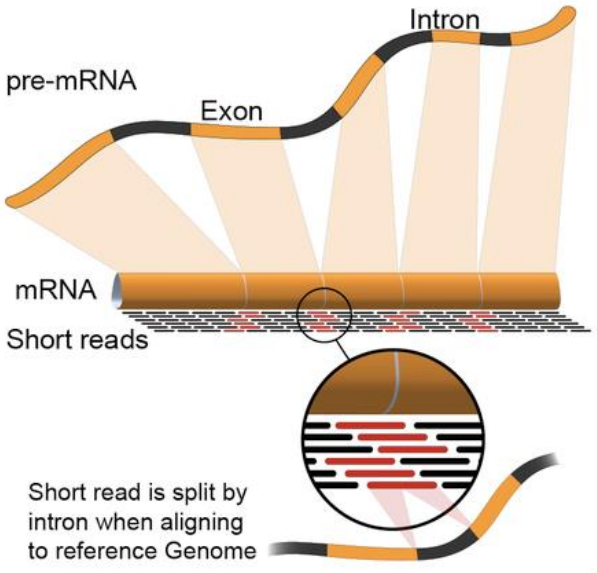

Figure 5 | *RNA-Seq alignment with intron-split short reads. In alignment of short reads to an mRNA sequence and the reference genome, alignment software has to account for short reads that overlap exon-exon junctions (in red) and thereby skip intronic sections of the pre-mRNA and reference genome. Rgocs, CC BY*

Differential expression

The simplest but often most powerful use of RNA-Seq is finding differences in gene expression between two or more conditions (e.g., treated vs not treated); this process is called differential expression. The outputs are frequently referred to as differentially expressed genes (DEGs) and these genes can either be up- or down-regulated (i.e., higher or lower in the condition of interest). There are many tools that perform differential expression. Most are run in R, Python, or the Unix command line. Commonly used tools include DESeq,[101] edgeR,[102] and voom+limma,[103][104] all of which are available through R/Bioconductor.[105][106] These are the common considerations when performing differential expression:

- *Inputs:* Differential expression inputs include (1) an RNA-Seq expression matrix (M genes x N samples) and (2) a design matrix containing experimental conditions for N samples. The simplest design matrix contains one column, corresponding to labels for the condition being tested. Other covariates (also referred to as factors, features, labels, or parameters) can include batch effects, known artifacts, and any metadata that might confound or mediate gene expression. In addition to known covariates, unknown covariates can also be estimated through unsupervised machine learning approaches including principal component, surrogate variable,[107] and PEER[108] analyses. Hidden variable analyses are often employed for human tissue RNA-Seq data, which typically have additional artifacts not captured in the metadata (*e.g.*, ischemic time, sourcing from multiple institutions, underlying clinical triats, collecting data across many years with many personnel).
- *Methods:* Most tools use regression or non-parametric statistics to identify differentially expressed genes, and are either based on read counts mapped to a reference genome (DESeq2, limma, edgeR) or based on read counts derived from alignment-free quantification (sleuth,[109] Cuffdiff,[110] Ballgown[111]).[112] Following regression, most tools employ either familywise error rate (FWER) or false discovery rate (FDR) p-value adjustments to account for multiple hypotheses (in human studies, ~20,000 protein-coding genes or ~50,000 biotypes).
- *Outputs:* A typical output consists of rows corresponding to the number of genes and at

least three columns, each gene's log fold change (log-transform of the ratio in expression between conditions, a measure of effect size), p-value, and p-value adjusted for multiple comparisons. Genes are defined as biologically meaningful if they pass cut-offs for effect size (log fold change) and statistical significance. These cut-offs should ideally be specified *a priori*, but the nature of RNA-Seq experiments is often exploratory so it is difficult to predict effect sizes and pertinent cut-offs ahead of time.

- *Pitfalls:* The raison d'etre for these complex methods is to avoid the myriad of pitfalls that can lead to statistical errors and misleading interpretations. Pitfalls include increased false positive rates (due to multiple comparisons), sample preparation artifacts, sample heterogeneity (like mixed genetic backgrounds), highly correlated samples, unaccounted for multi-level experimental designs, and poor experimental design. One notable pitfall is viewing results in Microsoft Excel.[113] Although convenient, Excel automatically converts some gene names (*SEPT1*, *DEC1*, *MARCH2*) into dates or floating point numbers.
- *Choice of tools and benchmarking:* There are numerous efforts that compare the results of these tools, with DESeq2 tending to moderately outperform other methods.[114][115][116][117][118][119][120] As with other methods, benchmarking consists of comparing tool outputs to each other and known gold standards.

Downstream analyses for a list of differentially expressed genes come in two flavors, validating observations and making biological inferences. Owing to the pitfalls of differential expression and RNA-Seq, important observations are replicated with (1) an orthogonal method in the same samples (like real-time PCR) or (2) another, sometimes pre-registered, experiment in a new cohort. The latter helps ensure generalizability and can typically be followed up with a meta-analysis of all the pooled cohorts. The most common method for obtaining higher-level biological understanding of the results is gene set enrichment analysis, although sometimes candidate gene approaches are employed. Gene set enrichment determines if the overlap between two gene sets is statistically significant, in this case the overlap between differentially expressed genes and gene sets from known pathways/databases (*e.g.*, Gene Ontology, KEGG, Human Phenotype Ontology) or from complementary analyses in the same data (like co-expression networks). Common tools for gene set enrichment include web interfaces (*e.g.*, ENRICHR, g:profiler, WEBGESTALT)[121] and software packages. When evaluating enrichment results, one heuristic is to first look for enrichment of known biology as a sanity check and then expand the scope to look for novel biology.

Alternative splicing

RNA splicing is integral to eukaryotes and contributes significantly to protein regulation and diversity, occurring in >90% of human genes.[122] There are multiple alternative splicing modes: exon skipping (most common splicing mode in humans and higher eukaryotes), mutually exclusive exons, alternative donor or acceptor sites, intron retention (most common splicing mode in plants, fungi, and protozoa), alternative transcription start site (promoter), and alternative polyadenylation (**Figure 6**).[123] One goal of RNA-Seq is to identify alternative splicing events and test if they differ between conditions. Long-read sequencing captures the full transcript and thus minimizes many of issues in estimating isoform abundance, like ambiguous read mapping. For short-read RNA-Seq, there are multiple methods to detect alternative splicing that can be classified into three main groups:[124][125][126]

- *Count-based (also event-based, differential splicing):* estimate exon retention. Examples are DEXSeq,[127], MATS[128], and SeqGSEA[129].
- *Isoform-based (also multi-read modules, differential isoform expression)*: estimate isoform abundance first, and then relative abundance between conditions. Examples are Cufflinks 2[130] and DiffSplice[131].
- *Intron excision based:* calculate alternative splicing using split reads. Examples are MAJIQ[132] and Leafcutter[133].

Differential gene expression tools can also be used for differential isoform expression if isoforms are quantified ahead of time with other tools like RSEM.[134]

Coexpression networks

Coexpression networks are data-derived representations of genes behaving in a similar way across tissues and experimental conditions.[135] Their main purpose lies in hypothesis generation and guilt-by-association approaches for inferring functions of previously unknown genes.[135] RNA-Seq data has been used to infer genes involved in specific pathways based on Pearson correlation, both in plants[136] and mammals.[137] The

main advantage of RNA-Seq data in this kind of analysis over the microarray platforms is the capability to cover the entire transcriptome, therefore allowing the possibility to unravel more complete representations of the gene regulatory networks. Differential regulation of the splice isoforms of the same gene can be detected and used to predict their biological functions.[138][139] Weighted gene co-expression network analysis has been successfully used to identify co-expression modules and intramodular hub genes based on RNA seq data. Co-expression modules may correspond to cell types or pathways. Highly connected intramodular hubs can be interpreted as representatives of their respective module. An eigengene is a weighted sum of expression of all genes in a module. Eigengenes are useful biomarkers (features) for diagnosis and prognosis.[140] Variance-Stabilizing Transformation approaches for estimating correlation coefficients based on RNA seq data have been proposed.[136]

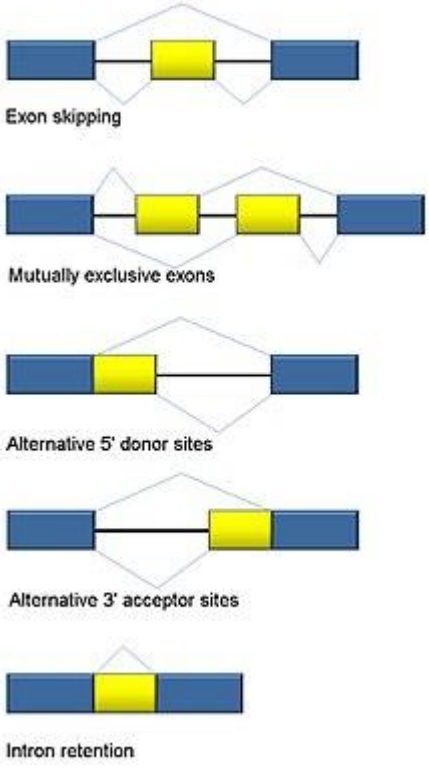

Figure 6 | *Examples of alternative RNA splicing modes. Exons are represented as blue and yellow blocks, spliced introns as horizontal black lines connecting two exons, and exon-exon junctions as thin grey connecting lines between two exons.*
Allen Gathman, CC BY-SA

Variant discovery

RNA-Seq captures DNA variation, including single nucleotide variants, small insertions/deletions. and structural variation. Variant calling in RNA-Seq is similar to DNA variant calling and often employs the same tools (including SAMtools mpileup[141] and GATK HaplotypeCaller[142]) with adjustments to account for splicing. One unique dimension for RNA variants is allele-specific expression (ASE): the vairants from only one haplotype might be preferentially expressed due to regulatory effects including imprinting and expression quantitative trait loci, and noncoding rare variants.[143][144] Limitations of RNA variant identification include that it only reflects expressed regions (in humans, <5% of the genome), could be subject to biases introduced by data processing (e.g., de novo transcriptome assemblies underestimate heterozygosity[145]), and has lower quality when compared to direct DNA sequencing.

RNA editing (post-transcriptional alterations)

See also: w:RNA editing

Having the matching genomic and transcriptomic sequences of an individual can help detect post-transcriptional edits (RNA editing).[2] A post-transcriptional modification event is identified if the gene's transcript has an allele/variant not observed in the genomic data.

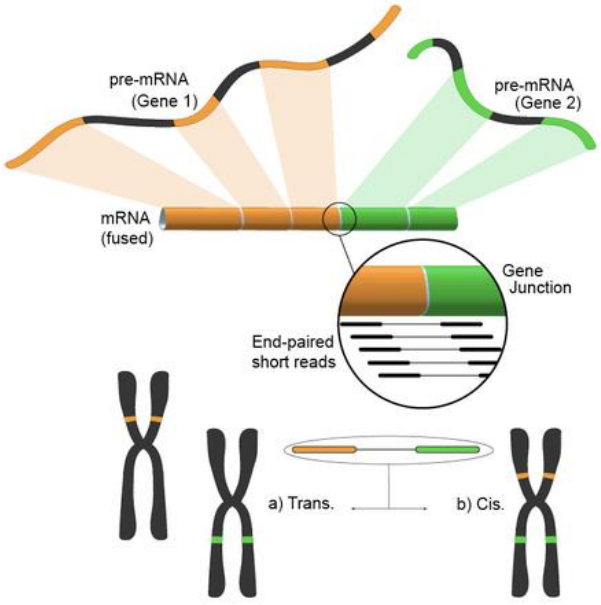

Figure 7 | *Gene fusion event. A gene fusion event and the behaviour of paired-end reads falling on both sides of the gene union. Gene fusions can occur in Trans, between genes on separate chromosomes, or in Cis, between two genes on the same chromosome.*
Rgocs , CC BY

Fusion gene detection

See also: w:Fusion gene

Caused by different structural modifications in the genome, fusion genes have gained attention because of their relationship with cancer.[146] The ability of RNA-Seq to analyze a sample's whole transcriptome in an unbiased fashion makes it an attractive tool to find these kinds of common events in cancer.[3]

The idea follows from the process of aligning the short transcriptomic reads to a reference genome. Most of the short reads will fall within one complete exon, and a smaller but still large set would be expected to map to known exon-exon junctions. The remaining unmapped short reads would then be further analyzed to determine whether they match an exon-exon junction where the exons come from different genes. This would be evidence of a possible fusion event, however, because of the length of the reads, this could prove to be very noisy. An alternative approach is to use paired-end reads, when a potentially large number of paired reads would map each end to a different exon, giving better coverage of these events (**Figure 7**). Nonetheless, the end result consists of multiple and potentially novel combinations of genes providing an ideal starting point for further validation.

Discussion

RNA-Seq and other NGS-based methods have flourished over the last decade. The number of manuscripts referring to RNA-Seq in the title or abstract (**Figure 8**, blue line) is continuously increasing with 6754 manuscripts published in 2018 (link to PubMed search). The intersection of RNA-Seq and medicine (**Figure 8**, gold line, link to PubMed search) has similar celerity.

Applications to medicine

RNA-Seq has the potential to identify new disease biology, profile biomarkers for clinical indications, infer druggable pathways, and make genetic diagnoses. These results could be further personalized for subgroups or even individual patients, potentially highlighting more effective prevention, diagnostics, and therapy. The feasibility of this approach is in part dictated by costs in money and time; a related limitation is the required team of specialists (bioinformaticians, physicians/clinicians, basic researchers, technicians) to fully interpret the huge amount of data generated by this analysis.

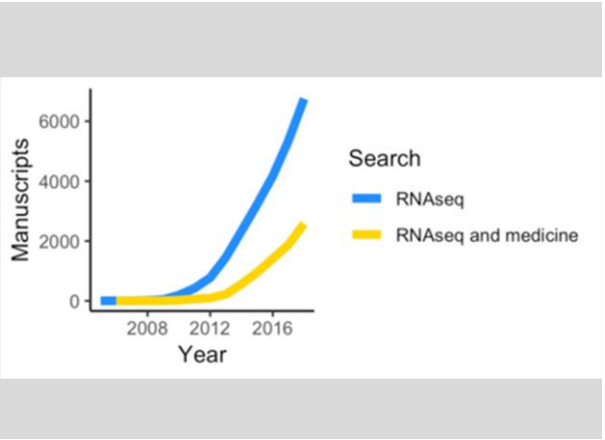

Figure 8 | *Pubmed manuscript matches highlight the growing popularity of RNA-Seq.* Matches are for RNA-Seq (blue, search terms: "RNA Seq" OR "RNA-Seq" OR "RNA sequencing" OR "RNASeq") and RNA=Seq in medicine (gold, search terms: ("RNA Seq" OR "RNA-Seq" OR "RNA sequencing" OR "RNASeq") AND "Medicine"). The number of manuscripts on PubMed featuring RNA-Seq is still increasing. CC BY-SA 4.0 Salubrious Toxin - Own work

Large-scale sequencing efforts

A lot of emphasis has been given to RNA-Seq data after the Encyclopedia of DNA Elements (ENCODE) and The Cancer Genome Atlas (TCGA) projects have used this approach to characterize dozens of cell lines[147] and thousands of primary tumor samples,[148] respectively. ENCODE aimed to identify genome-wide regulatory regions in different cohort of cell lines and transcriptomic data are paramount in order to understand the downstream effect of those epigenetic and genetic regulatory layers. TCGA, instead, aimed to collect and analyze thousands of patient's samples from 30 different tumor types in order to understand the underlying mechanisms of malignant transformation and progression. In this context RNA-Seq data provide a unique snapshot of the transcriptomic status of the disease and look at an unbiased population of transcripts that allows the identification of novel transcripts, fusion transcripts and non-coding RNAs that could be undetected with different technologies.

Acknowledgements

The authors thank the Wikipedia community and Wikijournal reviewers for their contributions. This work was supported by the Icahn School of Medicine at Mount Sinai Medical Scientist Training Program (NIH 5T32GM007280 to Felix Richter).
Competing interests: none declared.
Ethics statement: There was no human or animal subjects primary research performed for this article.

References

1. "RNA sequencing: platform selection, experimental design, and data interpretation". *Nucleic Acid Therapeutics* **22** (4): 271–4. August 2012. doi:10.1089/nat.2012.0367. PMID 22830413. PMC 3426205.
2. "RNA-Seq: a revolutionary tool for transcriptomics". *Nature Reviews Genetics* **10** (1): 57–63. January 2009. doi:10.1038/nrg2484. PMID 19015660. PMC 2949280.
3. "Transcriptome sequencing to detect gene fusions in cancer". *Nature* **458** (7234): 97–101. March 2009. doi:10.1038/nature07638. PMID 19136943. PMC 2725402.
4. "The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments". *Nature Protocols* **7** (8): 1534–50. July 2012. doi:10.1038/nprot.2012.086. PMID 22836135. PMC 3535016.
5. "Highly multiplexed subcellular RNA sequencing in situ". *Science* **343** (6177): 1360–3. March 2014. doi:10.1126/science.1250212. PMID 24578530. PMC 4140943.
6. "RNA Sequencing and Analysis". *Cold Spring Harbor Protocols* **2015** (11): 951–69. April 2015. doi:10.1101/pdb.top084970. PMID 25870306. PMC 4863231.
7. Lowe, Rohan; Shirley, Neil; Bleackley, Mark; Dolan, Stephen; Shafee, Thomas (2017-05-18). "Transcriptomics technologies". *PLOS Computational Biology* **13** (5): e1005457. doi:10.1371/journal.pcbi.1005457. ISSN 1553-7358. PMID 28545146. PMC PMC5436640.
8. Griffith, Obi L.; Ainscough, Benjamin J.; Spies, Nicholas C.; Walker, Jason R.; Griffith, Malachi (2015-08-06). "Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud". *PLOS Computational Biology* **11** (8): e1004393. doi:10.1371/journal.pcbi.1004393. ISSN 1553-7358. PMID 26248053. PMC PMC4527835.
9. "Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud". *PLoS Computational Biology* **11** (8): e1004393. August 2015. doi:10.1371/journal.pcbi.1004393. PMID 26248053. PMC 4527835.
10. *"RNA-seqlopedia". rnaseq.uoregon.edu. Retrieved 2017-02-08.*
11. "Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing". *BioTechniques* **45** (1): 81–94. July 2008. doi:10.2144/000112900. PMID 18611170.
12. "Mapping and quantifying mammalian transcriptomes by RNA-Seq". *Nature Methods* **5** (7): 621–8. July 2008. doi:10.1038/nmeth.1226. PMID 18516045.
13. Sun, Qinyu; Hao, Qinyu; Prasanth, Kannanganattu V. (2018-02). "Nuclear Long Noncoding RNAs: Key Regulators of Gene Expression". *Trends in genetics: TIG* **34** (2): 142–157. doi:10.1016/j.tig.2017.11.005. ISSN 0168-9525. PMID 29249332. PMC 6002860.
14. Sigurgeirsson, Benjamín; Emanuelsson, Olof; Lundeberg, Joakim (2014). "Sequencing degraded RNA addressed by 3' tag counting". *PloS One* **9** (3): e91851. doi:10.1371/journal.pone.0091851. ISSN 1932-6203. PMID 24632678. PMC 3954844.
15. Chen, Emily A; Souaiaia, Tade; Herstein, Jennifer S; Evgrafov, Oleg V; Spitsyna, Valeria N; Rebolini, Danea F; Knowles, James A (2014-10-23). "Effect of RNA integrity on uniquely mapped reads in RNA-Seq". *BMC Research Notes* **7**. doi:10.1186/1756-0500-7-753. ISSN 1756-0500. PMID 25339126. PMC 4213542.
16. Moll, Pamela; Ante, Michael; Seitz, Alexander; Reda, Torsten (2014-12). "QuantSeq 3′ mRNA sequencing for RNA quantification". *Nature Methods* **11** (12): i–iii. doi:10.1038/nmeth.f.376. ISSN 1548-7105.
17. Oikonomopoulos, Spyros; Bayega, Anthony; Fahiminiya, Somayyeh; Djambazian, Haig; Berube, Pierre; Ragoussis, Jiannis (2020). "Methodologies for Transcript Profiling Using Long-Read Technologies". *Frontiers in Genetics* **11**. doi:10.3389/fgene.2020.00606. ISSN 1664-8021.
18. Conesa, Ana; Madrigal, Pedro; Tarazona, Sonia; Gomez-Cabrero, David; Cervera, Alejandra; McPherson, Andrew; Szcześniak, Michał Wojciech; Gaffney, Daniel J. *et al.* (2016-01-26). "A survey of best practices for RNA-seq data analysis". *Genome Biology* **17** (1): 13. doi:10.1186/s13059-016-0881-8. ISSN 1474-760X. PMID 26813401. PMC PMC4728800.
19. "Quantitative comparison of EST libraries requires compensation for systematic biases in cDNA generation". *BMC Bioinformatics* **7**: 77. February 2006. doi:10.1186/1471-2105-7-77. PMID 16503995. PMC 1431573.
20. Garalde, Daniel R; Snell, Elizabeth A; Jachimowicz, Daniel; Sipos, Botond; Lloyd, Joseph H; Bruce, Mark; Pantic, Nadia; Admassu, Tigist *et al.* (15 January 2018). "Highly parallel direct RNA sequencing on an array of nanopores". *Nature Methods* **15** (3): 201–206. doi:10.1038/nmeth.4577.
21. Garalde, Daniel R; Snell, Elizabeth A; Jachimowicz, Daniel; Sipos, Botond; Lloyd, Joseph H; Bruce, Mark; Pantic, Nadia; Admassu, Tigist *et al.* (15 January 2018). "Highly parallel direct RNA sequencing on an array of nanopores". *Nature Methods* **15** (3): 201–206. doi:10.1038/nmeth.4577.
22. Gleeson, Josie; Lane, Tracy A.; Harrison, Paul J.; Haerty, Wilfried; Clark, Michael B. (2020-08-03). "Nanopore direct RNA sequencing detects differential expression between human cell populations". *bioRxiv*: 2020.08.02.232785. doi:10.1101/2020.08.02.232785.
23. ""Single-cell sequencing-based technologies will revolutionize whole-organism science". *Nature Reviews. Genetics* **14** (9): 618–30. September 2013. doi:10.1038/nrg3542. PMID 23897237."
24. "The technology and biology of single-cell RNA sequencing". *Molecular Cell* **58** (4): 610–20. May 2015. doi:10.1016/j.molcel.2015.04.005. PMID 26000846.
25. "A revised airway epithelial hierarchy includes CFTR-expressing ionocytes". *Nature* **560** (7718): 319–324. August 2018. doi:10.1038/s41586-018-0393-7. PMID 30069044. PMC 6295155.
26. "A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte". *Nature* **560** (7718): 377–381. August 2018. doi:10.1038/s41586-018-0394-6. PMID 30069046. PMC 6108322.
27. Valihrach, Lukas; Androvic, Peter; Kubista, Mikael (2018-03-11). "Platforms for Single-Cell Collection and Analysis". *International Journal of Molecular Sciences* **19** (3). doi:10.3390/ijms19030807. ISSN 1422-0067. PMID 29534489. PMC 5877668.
28. "Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells". *Cell* **161** (5): 1187–1201. May 2015. doi:10.1016/j.cell.2015.04.044. PMID 26000487. PMC 4441768.
29. "Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets". *Cell* **161** (5): 1202–1214. May 2015. doi:10.1016/j.cell.2015.05.002. PMID 26000488. PMC 4481139.
30. Islam, Saiful; Zeisel, Amit; Joost, Simon; La Manno, Gioele; Zajac, Pawel; Kasper, Maria; Lönnerberg, Peter; Linnarsson, Sten (2014-02). "Quantitative single-cell RNA-seq with unique molecular identifiers". *Nature Methods* **11** (2): 163–166. doi:10.1038/nmeth.2772. ISSN 1548-7105.
31. ""Methods, Challenges and Potentials of Single Cell RNA-seq". *Biology* **1** (3): 658–67. November 2012. doi:10.3390/biology1030658. PMID 24832513. PMC 4009822."
32. "The promise of single-cell sequencing". *Nature Methods* **11** (1): 25–7. January 2014. doi:10.1038/nmeth.2769. PMID 24524134.
33. "mRNA-Seq whole-transcriptome analysis of a single cell". *Nature Methods* **6** (5): 377–82. May 2009. doi:10.1038/NMETH.1315. PMID 19349980.
34. "Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq". *Genome Research* **21** (7): 1160–7. July 2011. doi:10.1101/gr.110882.110. PMID 21543516. PMC 3129258.
35. "Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells". *Nature Biotechnology* **30** (8): 777–82. August 2012. doi:10.1038/nbt.2282. PMID 22820318. PMC 3467340.
36. "CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification". *Cell Reports* **2** (3): 666–73. September 2012. doi:10.1016/j.celrep.2012.08.003. PMID 22939981.
37. "High-throughput targeted long-read single cell sequencing reveals the clonal and transcriptional landscape of lymphocytes". *bioRxiv*. 2018. doi:10.1101/424945.
38. "Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity". *Genome Biology* **14** (4): R31. April 2013. doi:10.1186/gb-2013-14-4-r31. PMID 23594475. PMC 4054835.
39. Shin, Jay W.; Plessy, Charles; Carninci, Piero; Arner, Erik; Hon, Chung-Chau; Lassmann, Timo; Kasukawa, Takeya; Suzuki, Harukazu *et al.* (2019-01-21). "C1 CAGE detects transcription start sites and enhancer activity at single-cell resolution". *Nature Communications* **10** (1): 360. doi:10.1038/s41467-018-08126-5. ISSN 2041-1723. PMID 30664627. PMC 6341120.
40. "How to design a single-cell RNA-sequencing experiment: pitfalls, challenges and perspectives". *Briefings in Bioinformatics*: bby007. January 2018. doi:10.1093/bib/bby007. PMID 29394315.
41. Klappenbach, Joel A.; Sadekova, Svetlana; McClanahan, Terrill K.; Moore, Renee; Douglas C. Wilson; Li, Lixia; Wong, Jerelyn; Kumar, Namit *et al.* (October 2017). "Multiplexed quantification of proteins and transcripts in single cells". *Nature Biotechnology* **35** (10): 936–939. doi:10.1038/nbt.3973. ISSN 1546-1696. PMID 28854175.
42. Smibert, Peter; Satija, Rahul; Swerdlow, Harold; Pratip K. Chattopadhyay; Houck-Loomis, Brian; Stephenson, William; Hafemeister, Christoph; Stoeckius, Marlon (September 2017). "Simultaneous epitope and transcriptome measurement in single cells". *Nature Methods* **14** (9): 865–868. doi:10.1038/nmeth.4380. ISSN 1548-7105. PMID 28759029. PMC 5669064.

43. "Simultaneous single-cell profiling of lineages and cell types in the vertebrate brain". *Nature Biotechnology* **36** (5): 442–450. June 2018. doi:10.1038/nbt.4103. PMID 29608178. PMC 5938111.
44. "Circulating tumour cell (CTC) counts as intermediate end points in castration-resistant prostate cancer (CRPC): a single-centre experience". *Annals of Oncology* **20** (1): 27–33. January 2009. doi:10.1093/annonc/mdn544. PMID 18695026.
45. Sims, Peter A.; Yuan, Jinzhou; Levitin, Hanna Mendes (2018-04-01). "Single-Cell Transcriptomic Analysis of Tumor Heterogeneity". *Trends in Cancer* **4** (4): 264–268. doi:10.1016/j.trecan.2018.02.003. ISSN 2405-8033. PMID 29606308. PMC 5993208.
46. Regev, Aviv; Izar, Benjamin; Yoon, Charles H.; Garraway, Levi A.; Rozenblatt-Rosen, Orit; Rotem, Asaf; Johnson, Bruce E.; Schadendorf, Dirk *et al.* (2018-11-01). "A Cancer Cell Program Promotes T Cell Exclusion and Resistance to Checkpoint Blockade". *Cell* **175** (4): 984–997.e24. doi:10.1016/j.cell.2018.09.006. ISSN 0092-8674. PMID 30388455.
47. Satija, Rahul; Swerdlow, Harold P.; Darnell, Robert B.; Orange, Dana E.; Bykerk, Vivian P.; Ivashkiv, Lionel B.; Goodman, Susan M.; Rashidfarrokhi, Ali *et al.* (2018-02-23). "Single-cell RNA-seq of rheumatoid arthritis synovial tissue using low-cost microfluidic instrumentation". *Nature Communications* **9** (1): 791. doi:10.1038/s41467-017-02659-x. ISSN 2041-1723. PMID 29476078. PMC 5824814.
48. "Pathogen Cell-to-Cell Variability Drives Heterogeneity in Host Immune Responses". *Cell* **162** (6): 1309–21. September 2015. doi:10.1016/j.cell.2015.08.027. PMID 26343579. PMC 4578813.
49. "Comprehensive single-cell transcriptional profiling of a multicellular organism". *Science* **357** (6352): 661–667. August 2017. doi:10.1126/science.aam8940. PMID 28818938. PMC 5894354.
50. "Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics". *Science* **360** (6391): eaaq1723. May 2018. doi:10.1126/science.aaq1723. PMID 29674432.
51. "Schmidtea mediterranea". *Science* **360** (6391): eaaq1736. May 2018. doi:10.1126/science.aaq1736. PMID 29674431.
52. "Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo". *Science* **360** (6392): 981–987. June 2018. doi:10.1126/science.aar4362. PMID 29700229. PMC 6083445.
53. "Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis". *Science* **360** (6392): eaar3131. June 2018. doi:10.1126/science.aar3131. PMID 29700225. PMC 6247916.
54. "The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution". *Science* **360** (6392): eaar5780. June 2018. doi:10.1126/science.aar5780. PMID 29700227. PMC 6038144.
55. *You, Jia. "Science's 2018 Breakthrough of the Year: tracking development cell by cell". Science Magazine. American Association for the Advancement of Science.*
56. "Determination of tag density required for digital transcriptome analysis: application to an androgen-sensitive prostate cancer model". *Proceedings of the National Academy of Sciences of the United States of America* **105** (51): 20179–84. December 2008. doi:10.1073/pnas.0807121105. PMID 19088194. PMC 2603435.
57. "Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses". *Nature Protocols* **7** (3): 500–7. February 2012. doi:10.1038/nprot.2011.457. PMID 22343431. PMC 3398141.
58. "Reference-based compression of short-read sequences using path encoding". *Bioinformatics* **31** (12): 1920–8. June 2015. doi:10.1093/bioinformatics/btv071. PMID 25649622. PMC 4481695.
59. "Full-length transcriptome assembly from RNA-Seq data without a reference genome". *Nature Biotechnology* **29** (7): 644–52. May 2011. doi:10.1038/nbt.1883. PMID 21572440. PMC 3571712.
60. *"De Novo Assembly Using Illumina Reads" (PDF). Retrieved 22 October 2016.*
61. Oases: a transcriptome assembler for very short reads
62. "Velvet: algorithms for de novo short read assembly using de Bruijn graphs". *Genome Research* **18** (5): 821–9. May 2008. doi:10.1101/gr.074492.107. PMID 18349386. PMC 2336801.
63. "Bridger: a new framework for de novo transcriptome assembly using RNA-seq data". *Genome Biology* **16** (1): 30. February 2015. doi:10.1186/s13059-015-0596-2. PMID 25723335. PMC 4342890.
64. Bushmanova, Elena; Antipov, Dmitry; Lapidus, Alla; Prjibelski, Andrey D. (2019-09-01). "rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data". *GigaScience* **8** (9). doi:10.1093/gigascience/giz100. ISSN 2047-217X. PMID 31494669. PMC 6736328.
65. "Evaluation of de novo transcriptome assemblies from RNA-Seq data". *Genome Biology* **15** (12): 553. December 2014. doi:10.1186/s13059-014-0553-5. PMID 25608678. PMC 4298084.
66. "STAR: ultrafast universal RNA-seq aligner". *Bioinformatics* **29** (1): 15–21. January 2013. doi:10.1093/bioinformatics/bts635. PMID 23104886. PMC 3530905.
67. "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome". *Genome Biology* **10** (3): R25. 2009. doi:10.1186/gb-2009-10-3-r25. PMID 19261174. PMC 2690996.
68. "TopHat: discovering splice junctions with RNA-Seq". *Bioinformatics* **25** (9): 1105–11. May 2009. doi:10.1093/bioinformatics/btp120. PMID 19289445. PMC 2672628.
69. "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks". *Nature Protocols* **7** (3): 562–78. March 2012. doi:10.1038/nprot.2012.016. PMID 22383036. PMC 3334321.
70. "The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote". *Nucleic Acids Research* **41** (10): e108. May 2013. doi:10.1093/nar/gkt214. PMID 23558742. PMC 3664803.
71. Kim, D; Langmead, B; Salzberg, SL (April 2015). "HISAT: a fast spliced aligner with low memory requirements.". *Nature Methods* **12** (4): 357–60. doi:10.1038/nmeth.3317. PMID 25751142. PMC 4655817.
72. "GMAP: a genomic mapping and alignment program for mRNA and EST sequences". *Bioinformatics* **21** (9): 1859–75. May 2005. doi:10.1093/bioinformatics/bti310. PMID 15728110.
73. Trapnell, Cole; Roberts, Adam; Goff, Loyal; Pertea, Geo; Kim, Daehwan; Kelley, David R.; Pimentel, Harold; Salzberg, Steven L. *et al.* (2012-03-01). "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks". *Nature Protocols* **7** (3): 562–578. doi:10.1038/nprot.2012.016. ISSN 1750-2799. PMID 22383036. PMC 3334321.
74. Pertea, Mihaela; Pertea, Geo M.; Antonescu, Corina M.; Chang, Tsung-Cheng; Mendell, Joshua T.; Salzberg, Steven L. (2015-03). "StringTie enables improved reconstruction of a transcriptome from RNA-seq reads". *Nature Biotechnology* **33** (3): 290–295. doi:10.1038/nbt.3122. ISSN 1546-1696. PMID 25690850. PMC 4643835.
75. "Simulation-based comprehensive benchmarking of RNA-seq aligners". *Nature Methods* **14** (2): 135–139. February 2017. doi:10.1038/nmeth.4106. PMID 27941783. PMC 5792058.
76. "Systematic evaluation of spliced alignment programs for RNA-seq data". *Nature Methods* **10** (12): 1185–91. December 2013. doi:10.1038/nmeth.2722. PMID 24185836. PMC 4018468.
77. "Comparative study of de novo assembly and genome-guided assembly strategies for transcriptome reconstruction based on RNA-Seq". *Science China Life Sciences* **56** (2): 143–55. February 2013. doi:10.1007/s11427-013-4442-z. PMID 23393030.
78. "Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species". *GigaScience* **2** (1): 10. July 2013. doi:10.1186/2047-217X-2-10. PMID 23870653. PMC 3844414.
79. Hölzer, Martin; Marz, Manja (2019-05-01). "De novo transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers". *GigaScience* **8** (5). doi:10.1093/gigascience/giz039. ISSN 2047-217X. PMID 31077315. PMC 6511074.
80. "Comparing protein abundance and mRNA expression levels on a genomic scale". *Genome Biology* **4** (9): 117. 2003. doi:10.1186/gb-2003-4-9-117. PMID 12952525. PMC 193646.
81. "A comparative study of techniques for differential expression analysis on RNA-Seq data". *PLOS ONE* **9** (8): e103207. August 2014. doi:10.1371/journal.pone.0103207. PMID 25119138. PMC 4132098.
82. "HTSeq--a Python framework to work with high-throughput sequencing data". *Bioinformatics* **31** (2): 166–9. January 2015. doi:10.1093/bioinformatics/btu638. PMID 25260700. PMC 4287950.
83. "featureCounts: an efficient general purpose program for assigning sequence reads to genomic features". *Bioinformatics* **30** (7): 923–30. April 2014. doi:10.1093/bioinformatics/btt656. PMID 24227677.
84. "Rcount: simple and flexible RNA-Seq read counting". *Bioinformatics* **31** (3): 436–7. February 2015. doi:10.1093/bioinformatics/btu680. PMID 25322836.
85. "Reducing bias in RNA sequencing data: a novel approach to compute counts". *BMC Bioinformatics* **15** (Suppl 1): S7. 2014. doi:10.1186/1471-2105-15-s1-s7. PMID 24564404. PMC 4016203.
86. "Universal count correction for high-throughput sequencing". *PLoS Computational Biology* **10** (3): e1003494. March 2014. doi:10.1371/journal.pcbi.1003494. PMID 24603409. PMC 3945112.
87. "Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms". *Nature Biotechnology* **32** (5): 462–4. May 2014. doi:10.1038/nbt.2862. PMID 24752080. PMC 4077321.

88. "Near-optimal probabilistic RNA-seq quantification". *Nature Biotechnology* **34** (5): 525–7. May 2016. doi:10.1038/nbt.3519. PMID 27043002.
89. "A scaling normalization method for differential expression analysis of RNA-seq data". *Genome Biology* **11** (3): R25. 2010. doi:10.1186/gb-2010-11-3-r25. PMID 20196867. PMC 2864565.
90. "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation". *Nature Biotechnology* **28** (5): 511–5. May 2010. doi:10.1038/nbt.1621. PMID 20436464. PMC 3146043.
91. *Pachter, Lior (19 April 2011). "Models for transcript quantification from RNA-Seq". arXiv:1104.3889 [q-bio.GN].* Unknown parameter |name-list-format= ignored (help)
92. *"What the FPKM? A review of RNA-Seq expression units". The farrago. 8 May 2014. Retrieved 28 March 2018.*
93. "Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples". *Theory in Biosciences = Theorie in den Biowissenschaften* **131** (4): 281–5. December 2012. doi:10.1007/s12064-012-0162-3. PMID 22872506.
94. "voom: Precision weights unlock linear model analysis tools for RNA-seq read counts". *Genome Biology* **15** (2): R29. February 2014. doi:10.1186/gb-2014-15-2-r29. PMID 24485249. PMC 4053721.
95. "Differential expression analysis for sequence count data". *Genome Biology* **11** (10): R106. 2010. doi:10.1186/gb-2010-11-10-r106. PMID 20979621. PMC 3218662.
96. Robinson, M. D.; McCarthy, D. J.; Smyth, G. K. (11 November 2009). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". *Bioinformatics* **26** (1): 139–140. doi:10.1093/bioinformatics/btp616.
97. "Quantitative analysis of fission yeast transcriptomes and proteomes in proliferating and quiescent cells". *Cell* **151** (3): 671–83. October 2012. doi:10.1016/j.cell.2012.09.019. PMID 23101633. PMC 3482660.
98. "Measuring Absolute RNA Copy Numbers at High Temporal Resolution Reveals Transcriptome Kinetics in Development". *Cell Reports* **14** (3): 632–647. January 2016. doi:10.1016/j.celrep.2015.12.050. PMID 26774488. PMC 4731879.
99. Chen, Kaifu; Hu, Zheng; Xia, Zheng; Zhao, Dongyu; Li, Wei; Tyler, Jessica K. (2015-12-28). "The Overlooked Fact: Fundamental Need for Spike-In Control for Virtually All Genome-Wide Analyses". *Molecular and Cellular Biology* **36** (5): 662–667. doi:10.1128/MCB.00970-14. ISSN 1098-5549. PMID 26711261. PMC 4760223.
100. Lovén, Jakob; Orlando, David A.; Sigova, Alla A.; Lin, Charles Y.; Rahl, Peter B.; Burge, Christopher B.; Levens, David L.; Lee, Tong Ihn *et al.* (2012-10-26). "Revisiting Global Gene Expression Analysis". *Cell* **151** (3): 476–482. doi:10.1016/j.cell.2012.10.012. ISSN 0092-8674. PMID 23101621. PMC 3505597.
101. "Differential expression analysis for sequence count data". *Genome Biology* **11** (10): R106. 2010. doi:10.1186/gb-2010-11-10-r106. PMID 20979621. PMC 3218662.
102. Robinson, M. D.; McCarthy, D. J.; Smyth, G. K. (11 November 2009). "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". *Bioinformatics* **26** (1): 139–140. doi:10.1093/bioinformatics/btp616.
103. "voom: Precision weights unlock linear model analysis tools for RNA-seq read counts". *Genome Biology* **15** (2): R29. February 2014. doi:10.1186/gb-2014-15-2-r29. PMID 24485249. PMC 4053721.
104. Ritchie, Matthew E.; Phipson, Belinda; Wu, Di; Hu, Yifang; Law, Charity W.; Shi, Wei; Smyth, Gordon K. (20 April 2015). "limma powers differential expression analyses for RNA-sequencing and microarray studies". *Nucleic Acids Research* **43** (7): e47. doi:10.1093/nar/gkv007. PMID 25605792.
105. *"Bioconductor - Open source software for bioinformatics".*
106. "Orchestrating high-throughput genomic analysis with Bioconductor". *Nature Methods* **12** (2): 115–21. February 2015. doi:10.1038/nmeth.3252. PMID 25633503. PMC 4509590.
107. Leek, Jeffrey T.; Storey, John D. (2007). "Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis". *PLoS Genetics* **3** (9): e161. doi:10.1371/journal.pgen.0030161.
108. Stegle, Oliver; Parts, Leopold; Piipari, Matias; Winn, John; Durbin, Richard (16 February 2012). "Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses". *Nature Protocols* **7** (3): 500–507. doi:10.1038/nprot.2011.457. PMC 3398141.
109. Pimentel, Harold; Bray, Nicolas L; Puente, Suzette; Melsted, Páll; Pachter, Lior (5 June 2017). "Differential analysis of RNA-seq incorporating quantification uncertainty". *Nature Methods* **14** (7): 687–690. doi:10.1038/nmeth.4324.
110. Trapnell, Cole; Hendrickson, David G; Sauvageau, Martin; Goff, Loyal; Rinn, John L; Pachter, Lior (9 December 2012). "Differential analysis of gene regulation at transcript resolution with RNA-seq". *Nature Biotechnology* **31** (1): 46–53. doi:10.1038/nbt.2450.
111. Frazee, Alyssa C; Pertea, Geo; Jaffe, Andrew E; Langmead, Ben; Salzberg, Steven L; Leek, Jeffrey T (6 March 2015). "Ballgown bridges the gap between transcriptome assembly and expression analysis". *Nature Biotechnology* **33** (3): 243–246. doi:10.1038/nbt.3172.
112. Sahraeian, Sayed Mohammad Ebrahim; Mohiyuddin, Marghoob; Sebra, Robert; Tilgner, Hagen; Afshar, Pegah T.; Au, Kin Fai; Bani Asadi, Narges; Gerstein, Mark B. *et al.* (5 July 2017). "Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis". *Nature Communications* **8** (1). doi:10.1038/s41467-017-00050-4.
113. Ziemann, Mark; Eren, Yotam; El-Osta, Assam (23 August 2016). "Gene name errors are widespread in the scientific literature". *Genome Biology* **17** (1). doi:10.1186/s13059-016-1044-7.
114. Soneson, Charlotte; Delorenzi, Mauro (2013). "A comparison of methods for differential expression analysis of RNA-seq data". *BMC Bioinformatics* **14**: 91. doi:10.1186/1471-2105-14-91. PMID 23497356. PMC 3608160.
115. Fonseca, Nuno A.; Marioni, John; Brazma, Alvis; Provart, Nicholas James (30 September 2014). "RNA-Seq Gene Profiling - A Systematic Empirical Comparison". *PLoS ONE* **9** (9): e107026. doi:10.1371/journal.pone.0107026.
116. Seyednasrollah, F.; Laiho, A.; Elo, L. L. (2 December 2013). "Comparison of software packages for detecting differential expression in RNA-seq studies". *Briefings in Bioinformatics* **16** (1): 59–70. doi:10.1093/bib/bbt086.
117. Rapaport, Franck; Khanin, Raya; Liang, Yupu; Pirun, Mono; Krek, Azra; Zumbo, Paul; Mason, Christopher E; Socci, Nicholas D *et al.* (2013). "Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data". *Genome Biology* **14** (9): R95. doi:10.1186/gb-2013-14-9-r95.
118. Conesa, Ana; Madrigal, Pedro; Tarazona, Sonia; Gomez-Cabrero, David; Cervera, Alejandra; McPherson, Andrew; Szcześniak, Michał Wojciech; Gaffney, Daniel J. *et al.* (26 January 2016). "A survey of best practices for RNA-seq data analysis". *Genome Biology* **17** (1). doi:10.1186/s13059-016-0881-8.
119. Sahraeian, Sayed Mohammad Ebrahim; Mohiyuddin, Marghoob; Sebra, Robert; Tilgner, Hagen; Afshar, Pegah T.; Au, Kin Fai; Bani Asadi, Narges; Gerstein, Mark B. *et al.* (5 July 2017). "Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis". *Nature Communications* **8** (1). doi:10.1038/s41467-017-00050-4.
120. Costa-Silva, Juliana; Domingues, Douglas; Lopes, Fabricio Martins; Wei, Zhi (21 December 2017). "RNA-Seq differential expression analysis: An extended review and a software tool". *PLOS ONE* **12** (12): e0190152. doi:10.1371/journal.pone.0190152.
121. Liao, Yuxing; Wang, Jing; Jaehnig, Eric J.; Shi, Zhiao; Zhang, Bing (2019-07-02). "WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs". *Nucleic Acids Research* **47** (W1): W199–W205. doi:10.1093/nar/gkz401. ISSN 1362-4962. PMID 31114916. PMC 6602449.
122. Keren, Hadas; Lev-Maor, Galit; Ast, Gil (8 April 2010). "Alternative splicing and evolution: diversification, exon definition and function". *Nature Reviews Genetics* **11** (5): 345–355. doi:10.1038/nrg2776.
123. Keren, Hadas; Lev-Maor, Galit; Ast, Gil (8 April 2010). "Alternative splicing and evolution: diversification, exon definition and function". *Nature Reviews Genetics* **11** (5): 345–355. doi:10.1038/nrg2776.
124. Liu, Ruolin; Loraine, Ann E; Dickerson, Julie A (16 December 2014). "Comparisons of computational methods for differential alternative splicing detection using RNA-seq in plant systems". *BMC Bioinformatics* **15** (1). doi:10.1186/s12859-014-0364-4.
125. Pachter, Lior (19 April 2011). *Models for transcript quantification from RNA-Seq* (in en).
126. Li, Yang I.; Knowles, David A.; Humphrey, Jack; Barbeira, Alvaro N.; Dickinson, Scott P.; Im, Hae Kyung; Pritchard, Jonathan K. (11 December 2017). "Annotation-free quantification of RNA splicing using LeafCutter". *Nature Genetics* **50** (1): 151–158. doi:10.1038/s41588-017-0004-9.
127. Anders, S.; Reyes, A.; Huber, W. (21 June 2012). "Detecting differential usage of exons from RNA-seq data". *Genome Research* **22** (10): 2008–2017. doi:10.1101/gr.133744.111.
128. Shen, Shihao; Park, Juw Won; Huang, Jian; Dittmar, Kimberly A.; Lu, Zhi-xiang; Zhou, Qing; Carstens, Russ P.; Xing, Yi (April 2012). "MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data". *Nucleic Acids Research* **40** (8): e61–e61. doi:10.1093/nar/gkr1291.
129. Wang, Xi; Cairns, Murray J. (15 June 2014). "SeqGSEA: a Bioconductor package for gene set enrichment analysis of RNA-Seq data integrating differential expression and splicing". *Bioinformatics* **30** (12): 1777–1779. doi:10.1093/bioinformatics/btu090.

130. Trapnell, Cole; Hendrickson, David G; Sauvageau, Martin; Goff, Loyal; Rinn, John L; Pachter, Lior (9 December 2012). "Differential analysis of gene regulation at transcript resolution with RNA-seq". *Nature Biotechnology* **31** (1): 46–53. doi:10.1038/nbt.2450.
131. Hu, Yin; Huang, Yan; Du, Ying; Orellana, Christian F.; Singh, Darshan; Johnson, Amy R.; Monroy, Anaïs; Kuan, Pei-Fen *et al.* (January 2013). "DiffSplice: the genome-wide detection of differential splicing events with RNA-seq". *Nucleic Acids Research* **41** (2): e39–e39. doi:10.1093/nar/gks1026.
132. Vaquero-Garcia, Jorge; Barrera, Alejandro; Gazzara, Matthew R; González-Vallinas, Juan; Lahens, Nicholas F; Hogenesch, John B; Lynch, Kristen W; Barash, Yoseph (1 February 2016). "A new view of transcriptome complexity and regulation through the lens of local splicing variations". *eLife* **5**: e11752. doi:10.7554/eLife.11752. ISSN 2050-084X.
133. Li, Yang I.; Knowles, David A.; Humphrey, Jack; Barbeira, Alvaro N.; Dickinson, Scott P.; Im, Hae Kyung; Pritchard, Jonathan K. (11 December 2017). "Annotation-free quantification of RNA splicing using LeafCutter". *Nature Genetics* **50** (1): 151–158. doi:10.1038/s41588-017-0004-9.
134. Merino, Gabriela A; Conesa, Ana; Fernández, Elmer A (March 2019). "A benchmarking of workflows for detecting differential splicing and differential expression at isoform level in human RNA-seq studies". *Briefings in Bioinformatics* **20** (2): 471–481. doi:10.1093/bib/bbx122.
135. "A combined algorithm for genome-wide prediction of protein function". *Nature* **402** (6757): 83–6. November 1999. doi:10.1038/47048. PMID 10573421.
136. "Comparative study of RNA-seq- and microarray-derived coexpression networks in Arabidopsis thaliana". *Bioinformatics* **29** (6): 717–24. March 2013. doi:10.1093/bioinformatics/btt053. PMID 23376351.
137. "Utilizing RNA-Seq data for de novo coexpression network inference". *Bioinformatics* **28** (12): 1592–7. June 2012. doi:10.1093/bioinformatics/bts245. PMID 22556371. PMC 3493127.
138. "Systematically differentiating functions for alternatively spliced isoforms through integrating RNA-seq data". *PLoS Computational Biology* **9** (11): e1003314. Nov 2013. doi:10.1371/journal.pcbi.1003314. PMID 24244129. PMC 3820534.
139. "The emerging era of genomic data integration for analyzing splice isoform function". *Trends in Genetics* **30** (8): 340–7. August 2014. doi:10.1016/j.tig.2014.05.005. PMID 24951248. PMC 4112133.
140. "Large-scale gene network analysis reveals the significance of extracellular matrix pathway and homeobox genes in acute myeloid leukemia: an introduction to the Pigengene package and its applications". *BMC Medical Genomics* **10** (1): 16. March 2017. doi:10.1186/s12920-017-0253-6. PMID 28298217. PMC 5353782.
141. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G. *et al.* (8 June 2009). "The Sequence Alignment/Map format and SAMtools". *Bioinformatics* **25** (16): 2078–2079. doi:10.1093/bioinformatics/btp352. PMID 19505943.
142. DePristo, Mark A; Banks, Eric; Poplin, Ryan; Garimella, Kiran V; Maguire, Jared R; Hartl, Christopher; Philippakis, Anthony A; del Angel, Guillermo *et al.* (10 April 2011). "A framework for variation discovery and genotyping using next-generation DNA sequencing data". *Nature Genetics* **43** (5): 491–498. doi:10.1038/ng.806.
143. "Genetic effects on gene expression across human tissues". *Nature* **550** (7675): 204–213. 12 October 2017. doi:10.1038/nature24277.
144. Richter, F; Hoffman, G E; Manheimer, K B; Patel, N; Sharp, A J; McKean, D; Morton, S U; DePalma, S *et al*. (23 March 2019). "ORE identifies extreme expression effects enriched for rare variants". *Bioinformatics*. doi:10.1093/bioinformatics/btz202.
145. Freedman, Adam H.; Clamp, Michele; Sackton, Timothy B. (2021-01). "Error, noise and bias in de novo transcriptome assemblies". *Molecular Ecology Resources* **21** (1): 18–29. doi:10.1111/1755-0998.13156. ISSN 1755-0998. PMID 32180366.
146. "Recurrent fusion oncogenes in carcinomas". *Critical Reviews in Oncogenesis* **12** (3–4): 257–71. December 2006. doi:10.1615/critrevoncog.v12.i3-4.40. PMID 17425505.
147. *"ENCODE Data Matrix". Retrieved 2013-07-28.*
148. *"The Cancer Genome Atlas - Data Portal". Retrieved 2013-07-28.*