Introduction to Genomics

Vitor Pavinato
correapavinato.1@osu.edu

Road map

- DNA and RNA sequencing: first generations and next generation
- How to access the expression profile: microarray and RNAseq
- NGS in biology

The development of D(R)NA sequencing technology

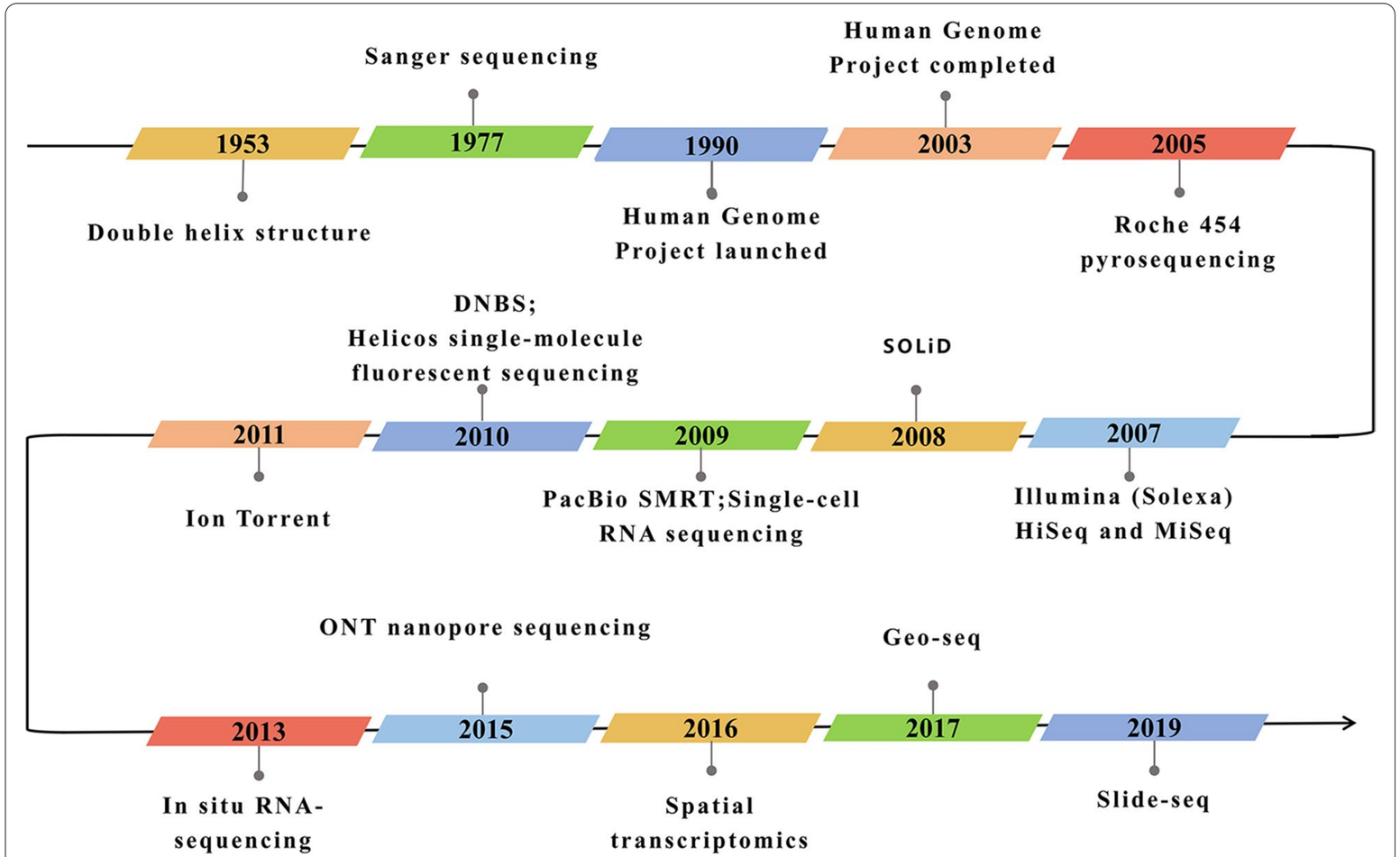
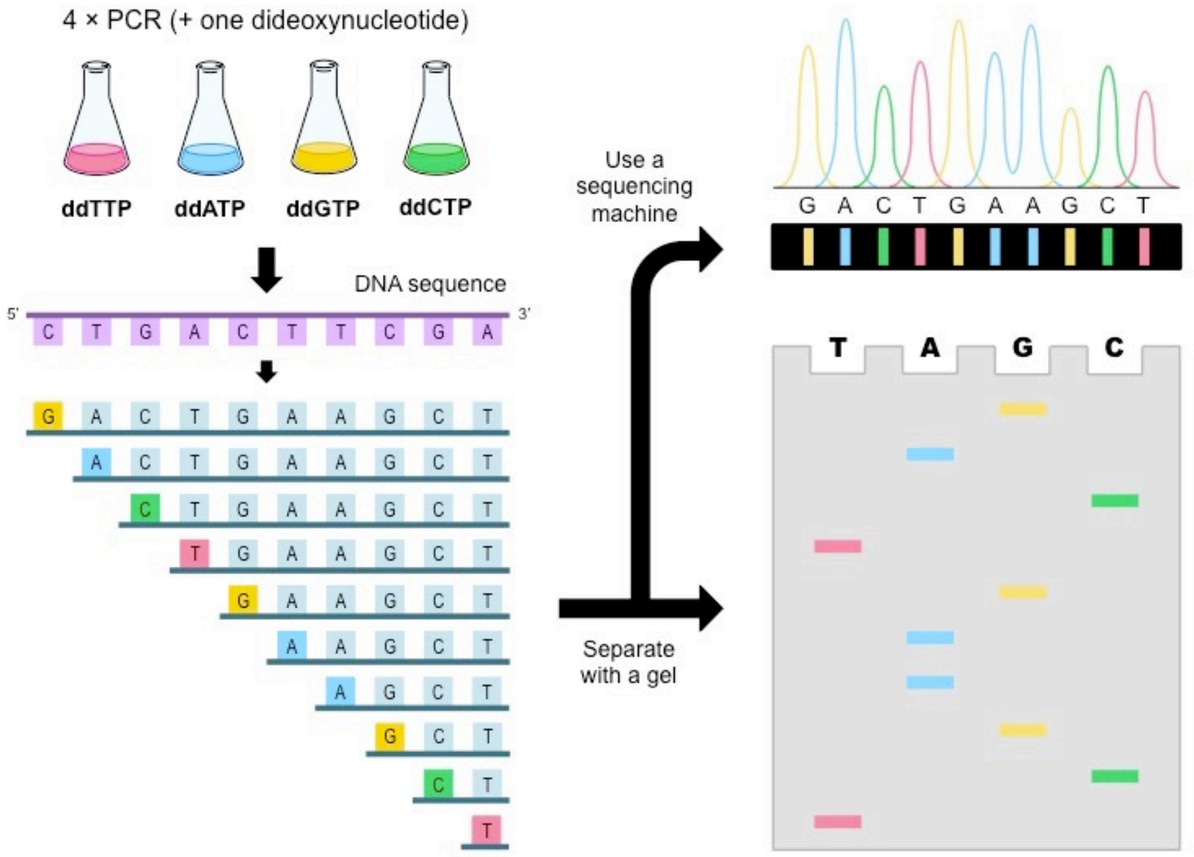


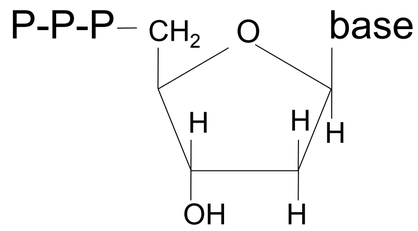
Fig. 1 The development timeline of RNA sequencing technologies

Hong M et al. (2020). RNA sequencing: new technologies and applications in cancer research. J Hematol Oncol

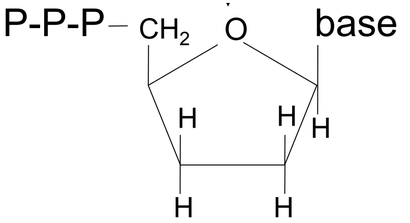
Courtesy of Dr. F. Sanger, MRC, Cambridge.
Noncommercial, educational use only.

“Sanger Method”

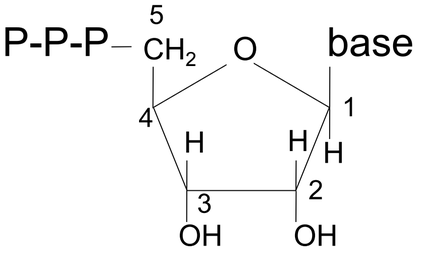




Deoxyribonucleotide



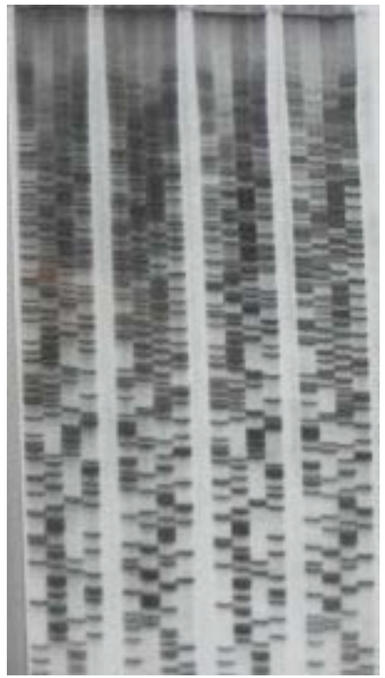
Dideoxyribonucleotide



Ribonucleotide

Evolution of Sequencing Technologies

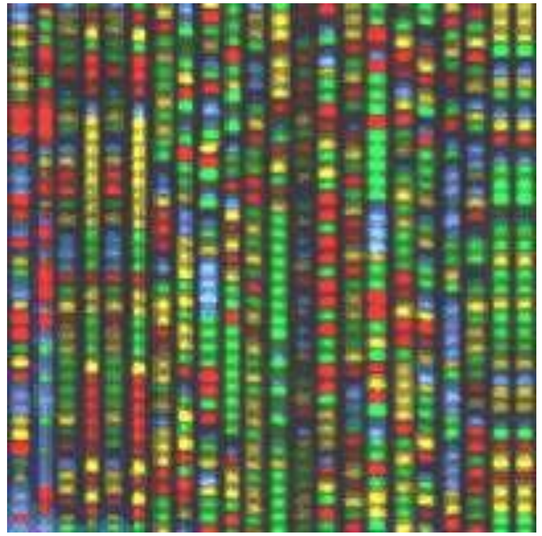
- Traditional Sanger / chain termination sequencing (70s, 80s, 90s)



A T C G ...

- Large polyacrylamide gels, radiolabeled DNA, 4 lanes per read

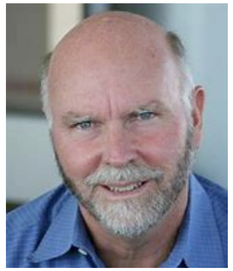
- Fluorescent-based / dye terminator sequencing (90s - present)



- Capillary electrophoresis, fluorescent tags for each base, 1 lane per read

Human Genome Project

Francis Collins
NIH



Craig Venter
Celera Genomics

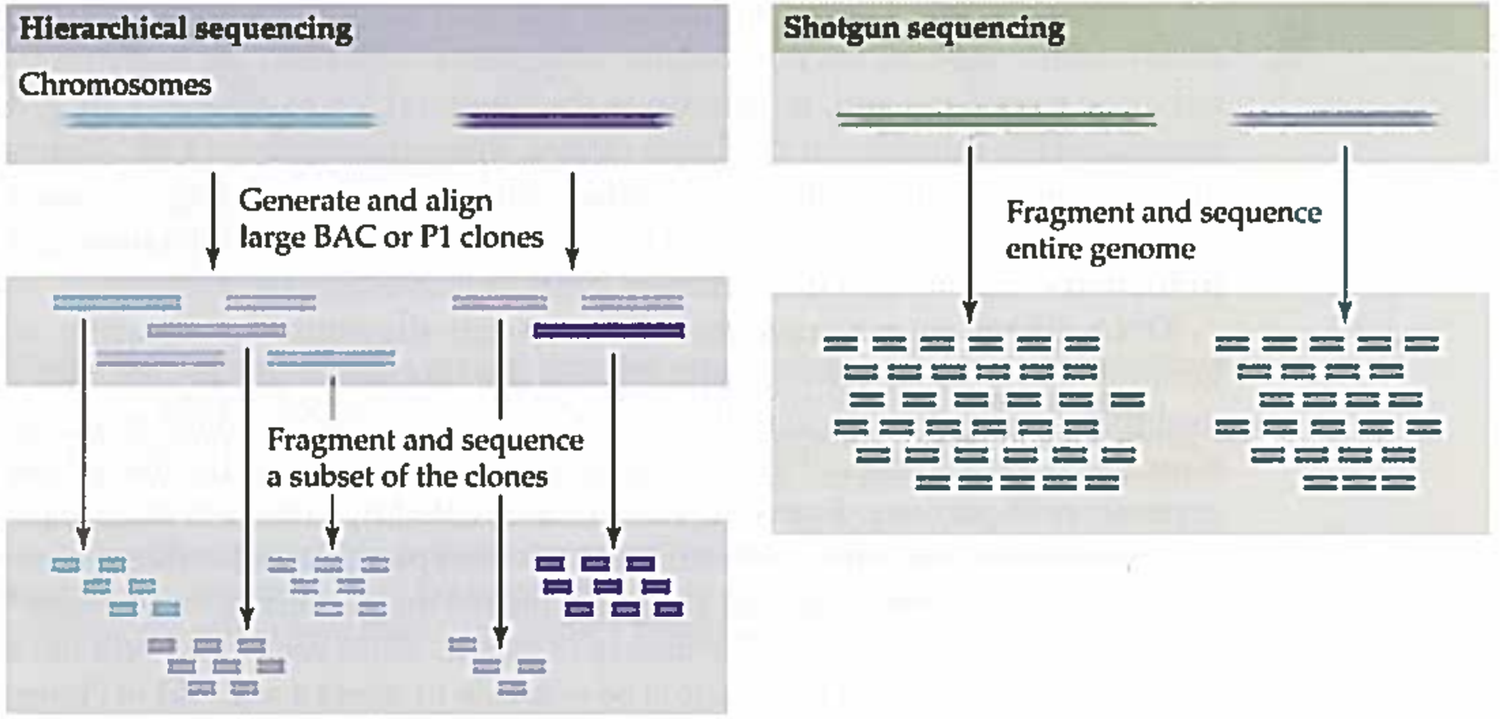
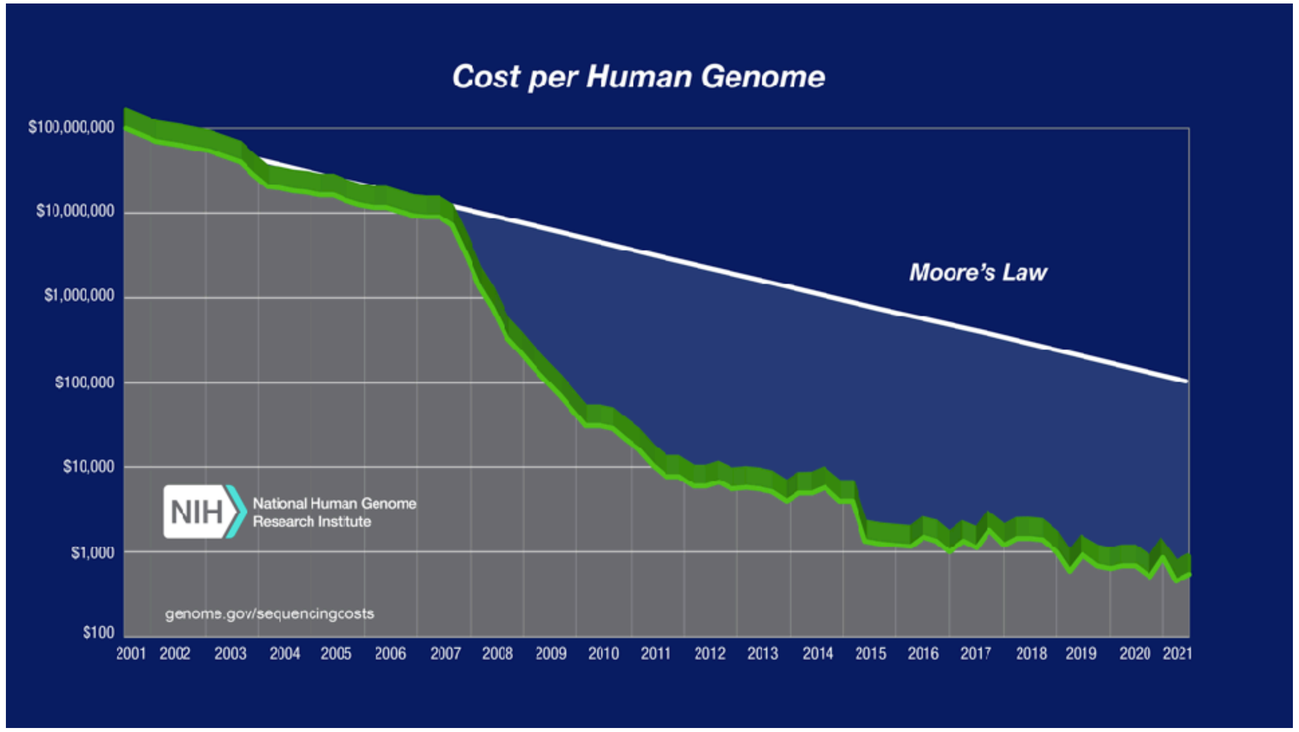
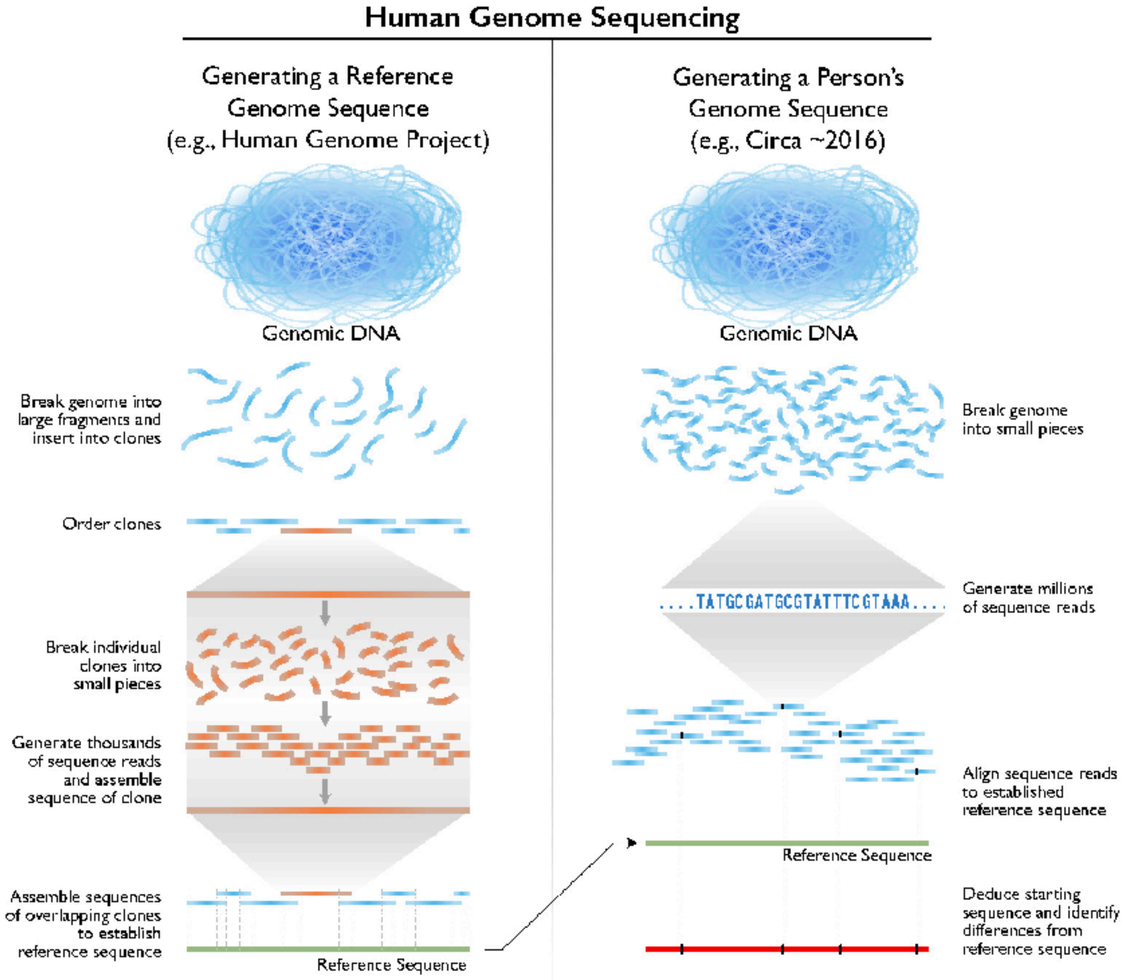


Figure 2.7 Hierarchical versus shotgun sequencing. In hierarchical sequencing, the chromosomes are first cloned as large BAC or P1 fragments up to 200 kb. These are physically ordered, and a subset that gives a minimal overlap for complete genome coverage is chosen for shotgun sequencing. In the whole-genome shotgun approach, no attempt is made to order the clones in advance. Instead, the whole genome is assembled using computer algorithms that order contigs based on their overlapping sequences.

Gibson and Muse 2009, A primer of genome science

Personalized genomics





https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

Sequence yield through time

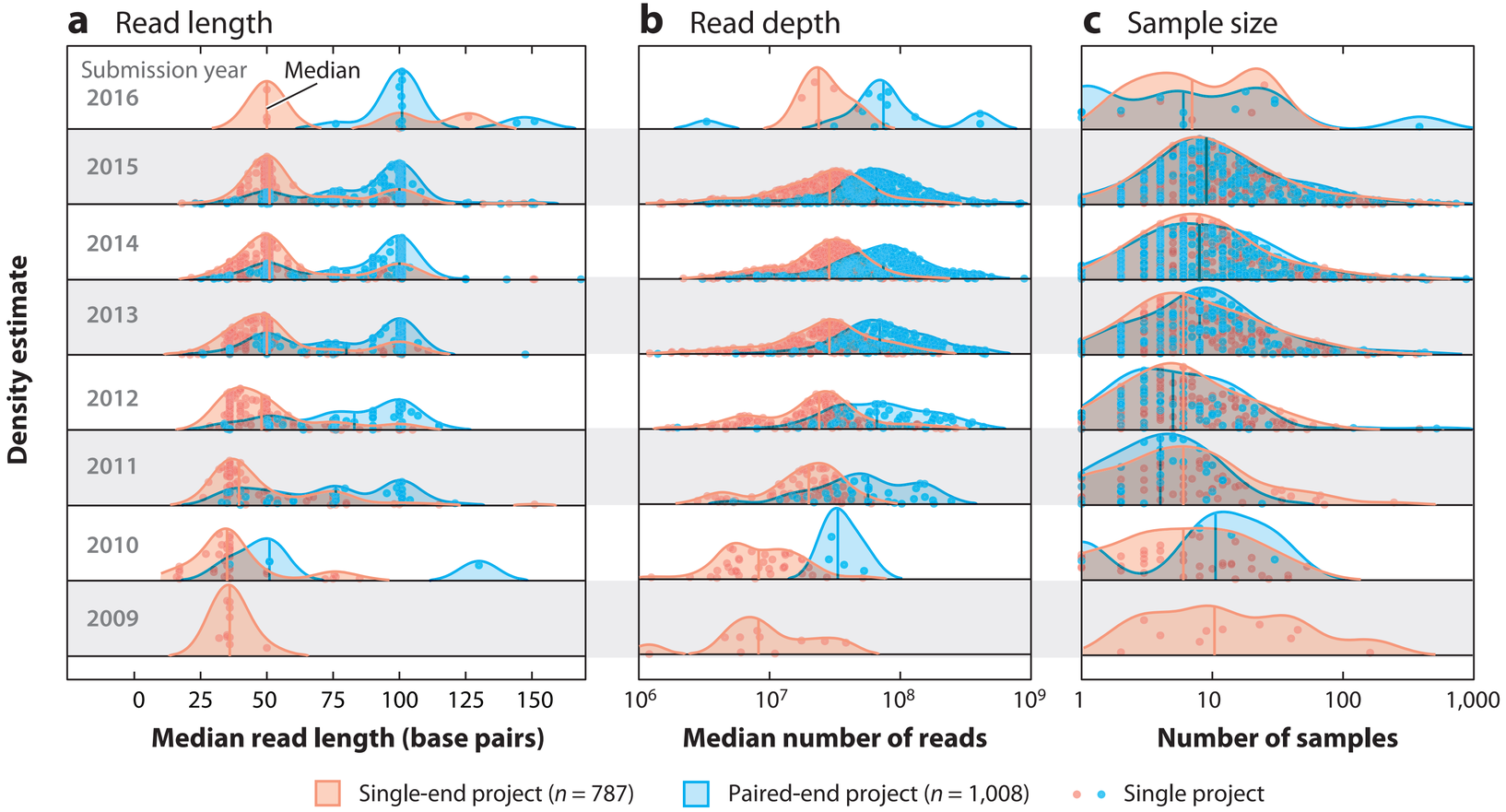


Figure 2

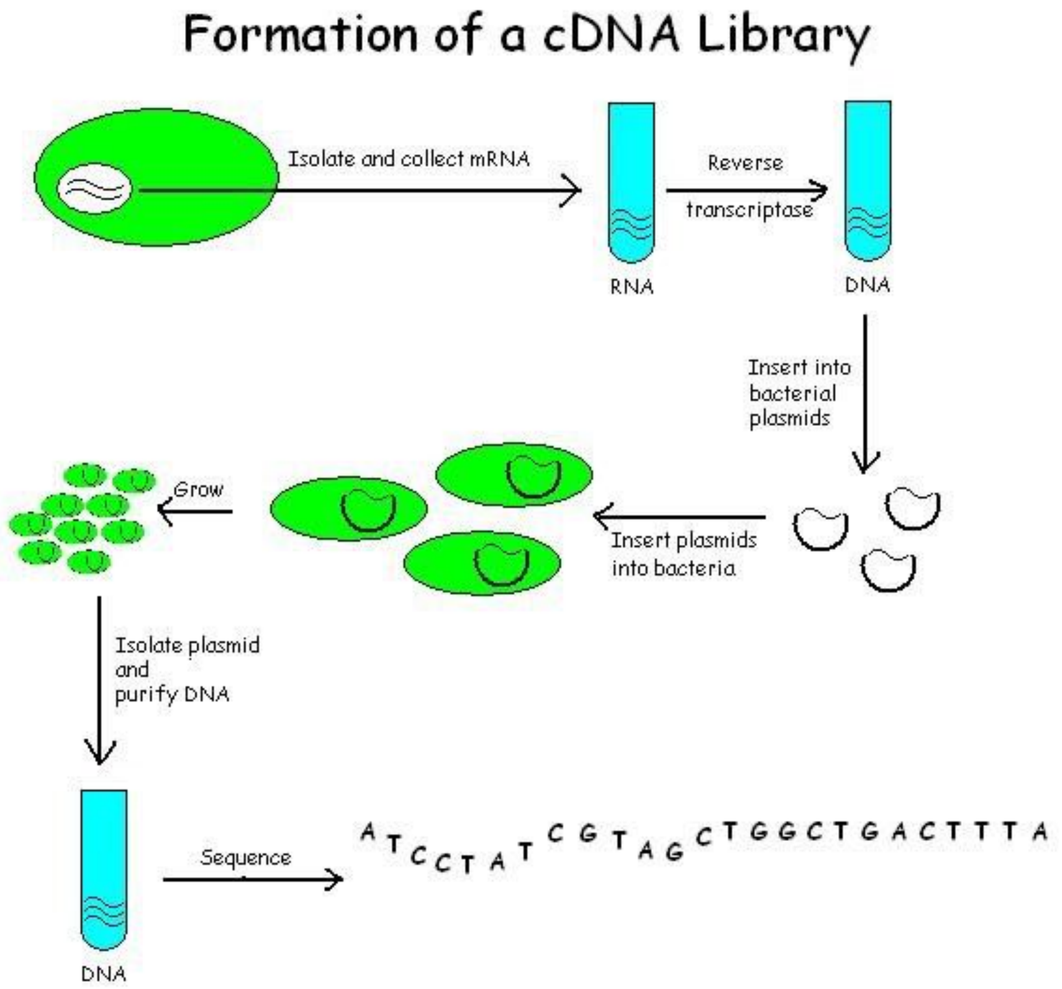
Ridge plots showing the progression of read length, depth, and sample size in Sequence Read Archive (SRA) projects using the `recount` package (15). The projects are separated by the submission year of the biosample. (*a*) Median read length of all samples per project and year. (*b*) Median number of reads across all samples per project and year. (*c*) Number of samples in each project. Each point represents one project.

Central dogma

Number	Process	Mechanism
1	Replication	DNA polymerase
2	Transcription	RNA polymerase
3	Translation	Ribosomes
4	Reverse transcription*	Reverse transcriptase
5	Reverse translation	*In silico*

* in the lab, reverse transcription is used to make cDNA (complementary DNA) from mRNA

cDNA libraries and EST



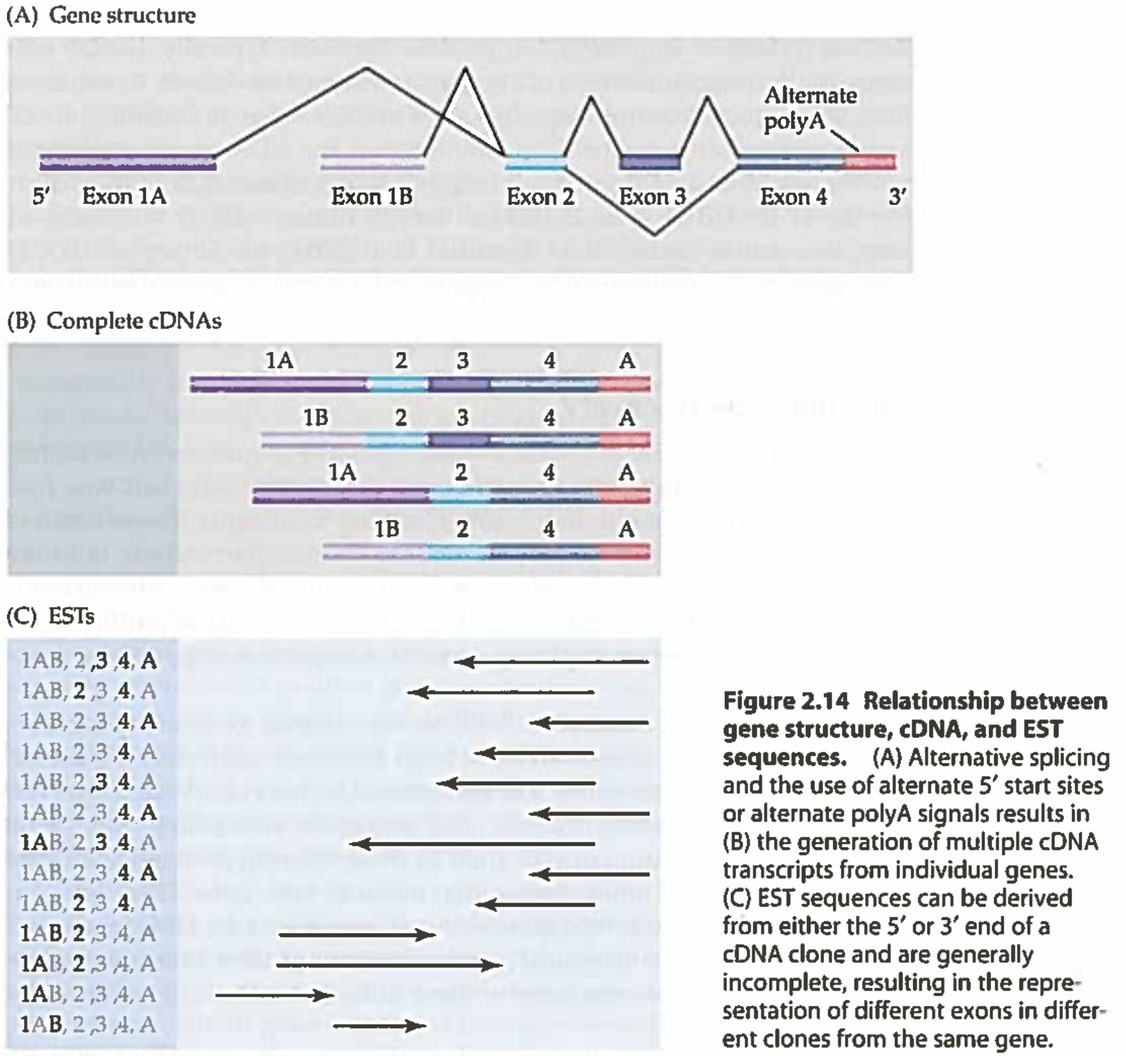


Figure 2.14 Relationship between gene structure, cDNA, and EST sequences. (A) Alternative splicing and the use of alternate 5′ start sites or alternate polyA signals results in (B) the generation of multiple cDNA transcripts from individual genes. (C) EST sequences can be derived from either the 5′ or 3′ end of a cDNA clone and are generally incomplete, resulting in the representation of different exons in different clones from the same gene.

https://en.wikipedia.org/wiki/CDNA_library

Next-generation sequencing

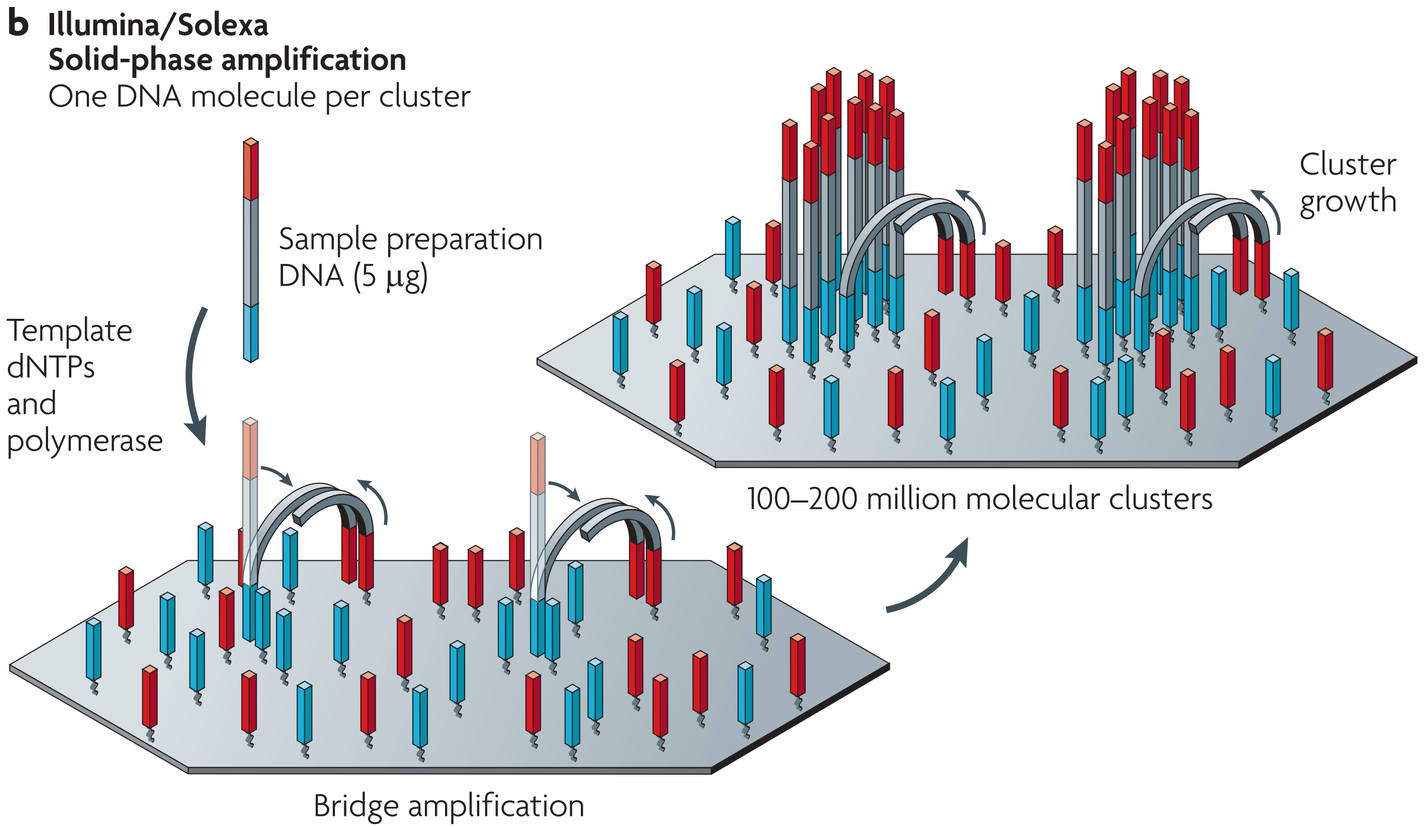
Table 1 | **Comparison of next-generation sequencing platforms**

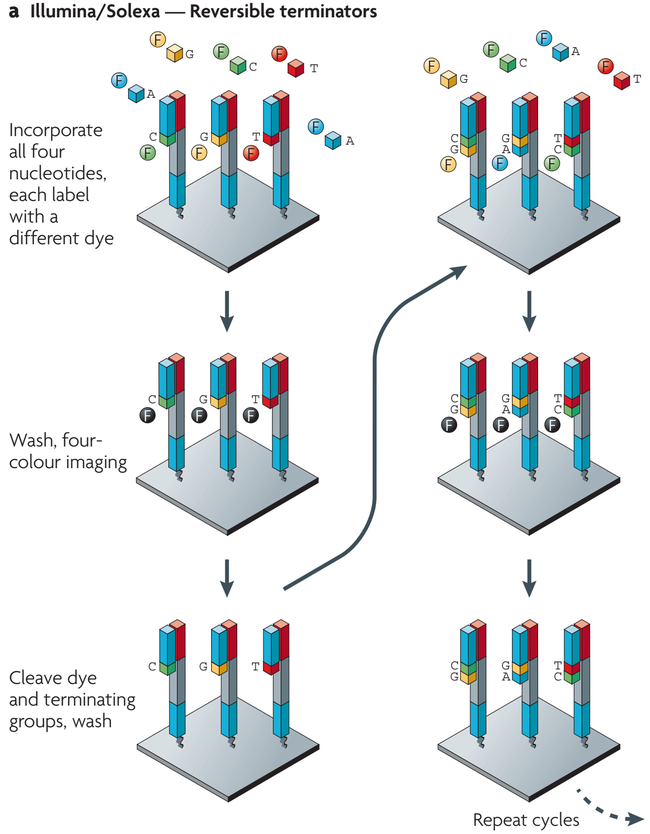
Platform	Library/ template preparation	NGS chemistry	Read length (bases)	Run time (days)	Gb per run	Machine cost (US$)	Pros	Cons	Biological applications	Refs
Roche/454's GS FLX Titanium	Frag, MP/ emPCR	PS	330*	0.35	0.45	500,000	Longer reads improve mapping in repetitive regions; fast run times	High reagent cost; high error rates in homo-polymer repeats	Bacterial and insect genome *de novo* assemblies; medium scale (<3 Mb) exome capture; 16S in metagenomics	D. Muzny, pers. comm.
Illumina/ Solexa's GA_{II}	Frag, MP/ solid-phase	RTs	75 or 100	4‡, 9§	18‡, 35§	540,000	Currently the most widely used platform in the field	Low multiplexing capability of samples	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics	D. Muzny, pers. comm.
Life/APG's SOLiD 3	Frag, MP/ emPCR	Cleavable probe SBL	50	7‡, 14§	30‡, 50§	595,000	Two-base encoding provides inherent error correction	Long run times	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics	D. Muzny, pers. comm.
Polonator G.007	MP only/ emPCR	Non-cleavable probe SBL	26	5§	12§	170,000	Least expensive platform; open source to adapt alternative NGS chemistries	Users are required to maintain and quality control reagents; shortest NGS read lengths	Bacterial genome resequencing for variant discovery	J. Edwards, pers. comm.
Helicos BioSciences HeliScope	Frag, MP/ single molecule	RTs	32*	8‡	37‡	999,000	Non-bias representation of templates for genome and seq-based applications	High error rates compared with other reversible terminator chemistries	Seq-based methods	91
Pacific Biosciences (target release: 2010)	Frag only/ single molecule	Real-time	964*	N/A	N/A	N/A	Has the greatest potential for reads exceeding 1 kb	Highest error rates compared with other NGS chemistries	Full-length transcriptome sequencing; complements other resequencing efforts in discovering large structural variants and haplotype blocks	S. Turner, pers. comm.

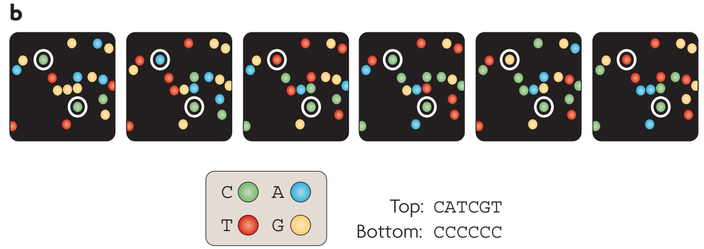
*Average read-lengths. ‡Fragment run. §Mate-pair run. Frag, fragment; GA, Genome Analyzer; GS, Genome Sequencer; MP, mate-pair; N/A, not available; NGS, next-generation sequencing; PS, pyrosequencing; RT, reversible terminator; SBL, sequencing by ligation; SOLiD, support oligonucleotide ligation detection.

Metzker et al. 2010 Nature Reviews Genetics

Illuminas/Solexa reversible terminators



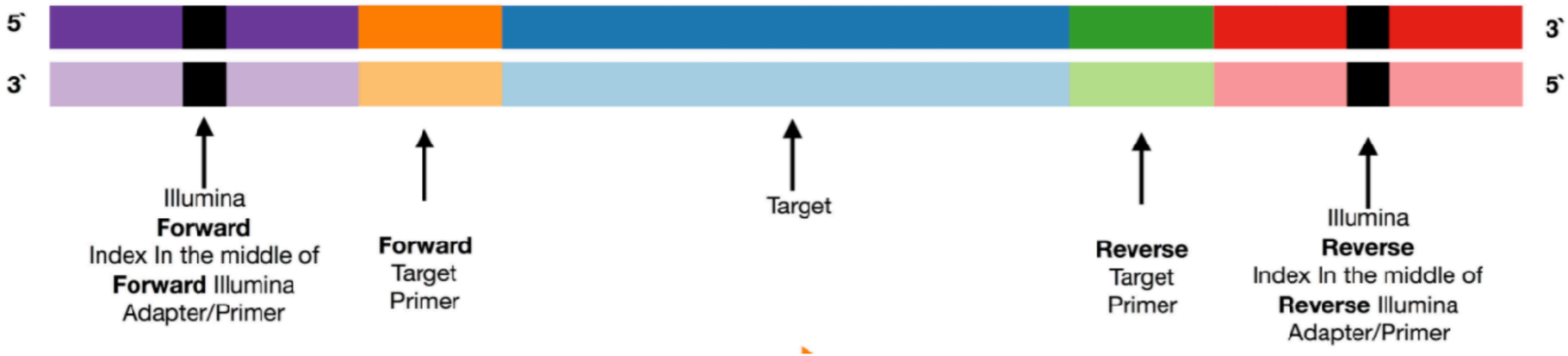




Metzker et al. 2010 Nature Reviews Genetics

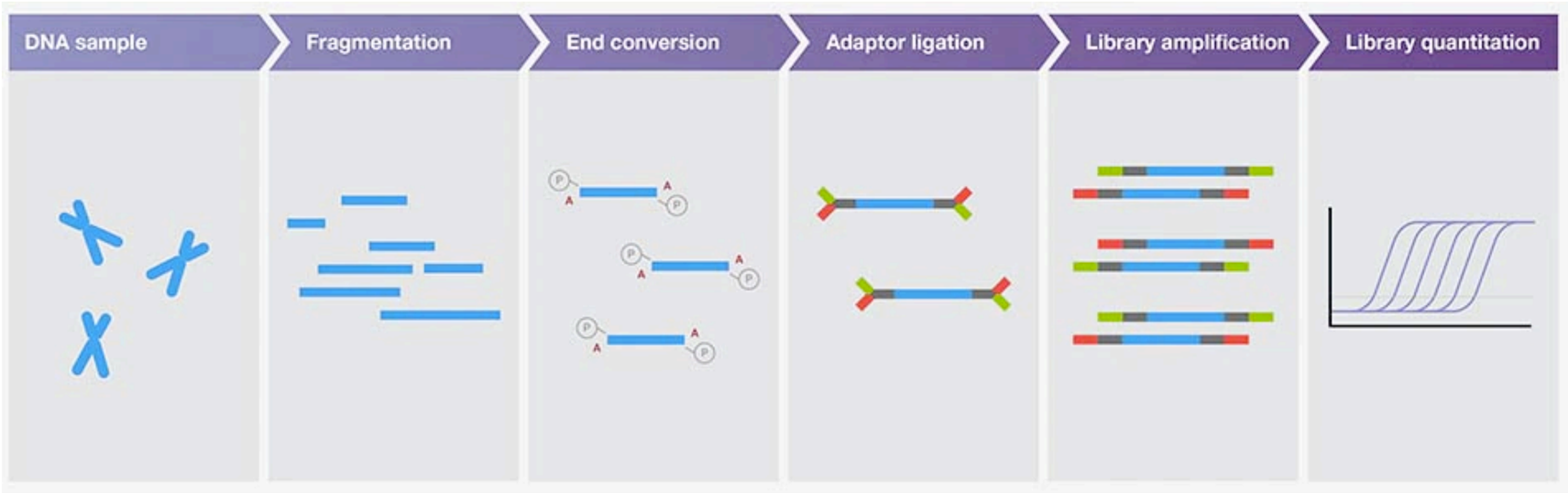
https://www.youtube.com/watch?v=fCd6B5HRaZ8&ab_channel=Illumina

Illumina Adapters



https://baileylab.brown.edu/SeekDeep/illumina_paired_info.html

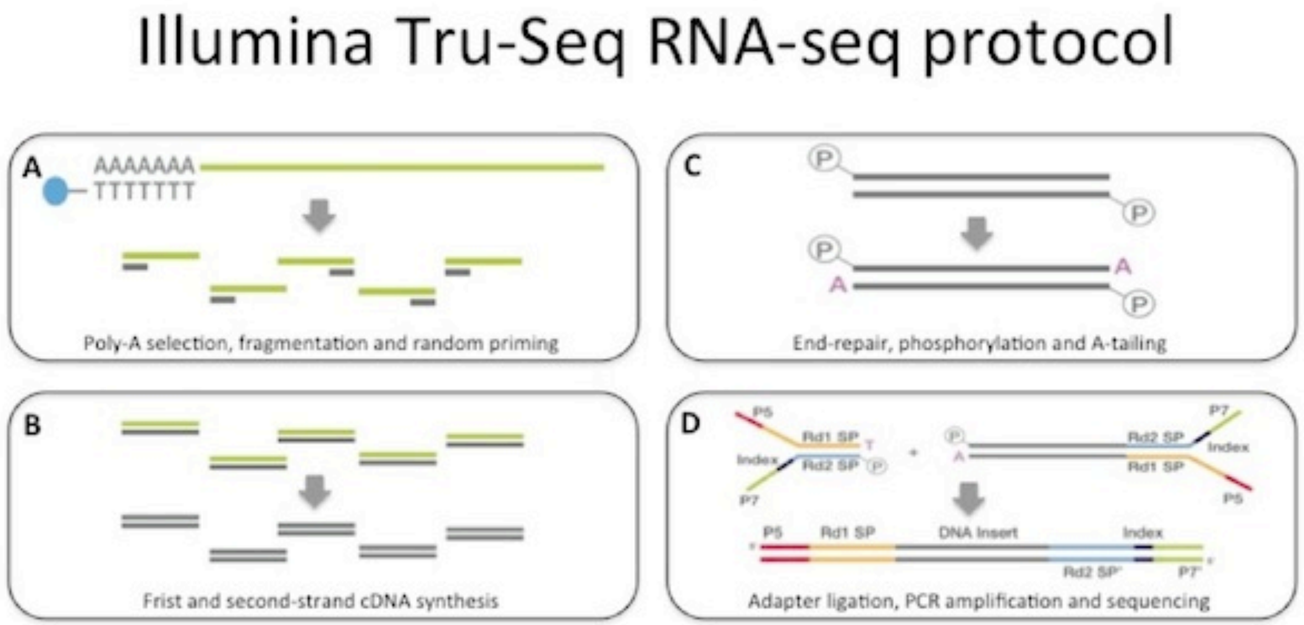
Illumina sequencing



The same principle as shotgun sequencing

Whole-genome sequencing
***De novo* sequencing**

Illumina sequencing



Library prep begins from 100ng-1ug of Total RNA which is poly-A selected (A) with magnetic beads. Double-stranded cDNA (B) is phosphorylated and A-tailed (C) ready for adapter ligation. The library is PCR amplified (D) ready for clustering and sequencing.

The same principle as ESTs

Figure 2: Unprecedented Flexibility for Multiple Applications

With 2 run modes and the ability to process 1 or 2 flow cells simultaneously, the HiSeq 2500 supports a broad range of applications and study sizes. Values here are examples based on system performance capabilities using 2 flow cells. Variation can be expected due to cluster density, sample quality, and other experimental factors. Run times are approximate and include on-board cluster generation and sequencing for rapid-run mode, and sequencing only for high-output mode. Indexing run times are not included.

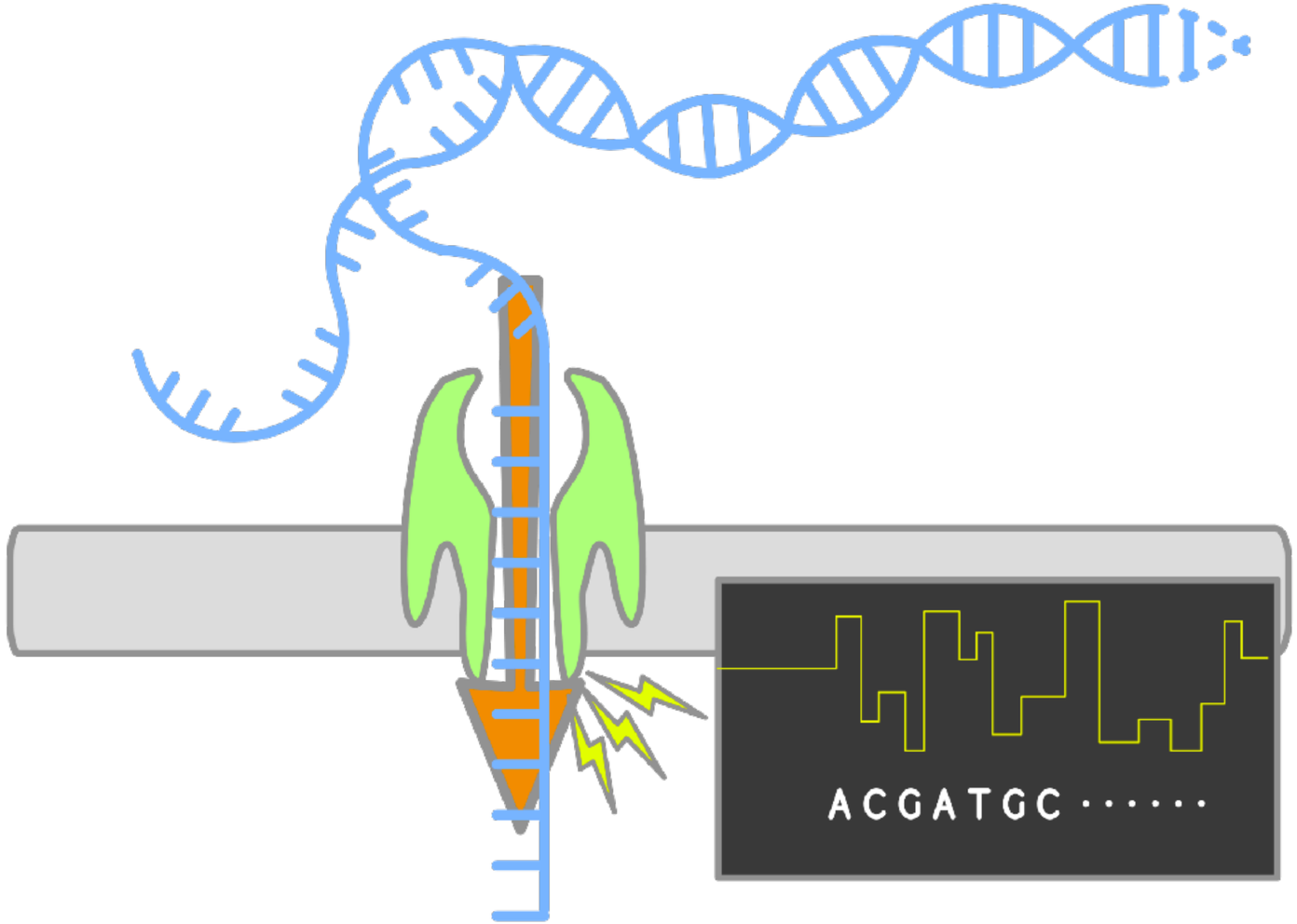
Third Generation Sequencing

Table 1 Comparison of different RNA sequencing platforms

Platform	Company	Read length(bases)	Run time	Volume per run	Cost	Template preparation	Sequencing chemistry
The first-generation sequencing							
Sanger	Life sciences	800 bp	2 h	1 read	$2400 per million bases	Bacterial cloning	Dideoxynucleosides terminator
The next-generation sequencing							
Roche 454 pyrosequencing	454 Life sciences	700 bp	< 24 h	0.7 Gb	$10 per million bases	Emulsion PCR	Sequencing by synthesis, pyrosequencing
Illumina HiSeq	Illumina	100 bp	3–10 days	120–1500 Gb	$0.02—$0.07 per million bases	Bridge PCR	Reversible terminator sequencing
Illumina MiSeq	Illumina	100 bp	1–2 days	0.3–15 Gb	$0.13 per million bases	Bridge PCR	Reversible terminator sequencing
SOLiD	Applied biosystems instruments (ABI)	50–75 bp	7–14 days	30 Gb	$0.13 per million bases	Emulsion PCR	Sequencing by ligation
DNA nanoball sequencing	Complete genomics	440–500 bp	9 days	20–60 Gb	$4400/genome	Rolling circle replication	Hybridization and ligation
Ion torrent	454 Life sciences	200–500 bp	4–5 h	660 Mb; 11 Mb	$300 to $750 per run	Emulsion PCR	Sequencing by synthesis
The third-generation sequencing							
SMRT	Pacific biosciences	> 900 bp	1-2 h	0.5–1 Gb	$2 per million bases	No need	Sequencing by synthesis
Helicos sequencing	Helicos biosciences	25–60 bp	8 days	21–35 Gb	$0.01 per million bases	No need	Hybridization and synthesis
Nanopore sequencing	Oxford nanopore technologies	Up to 98 kb	48/72 h	Up to 30 Gb	< $1 per million bases	No need	Nanopore

Hong M et al. (2020). *J Hematol Oncol*

Nanopore Sequencing



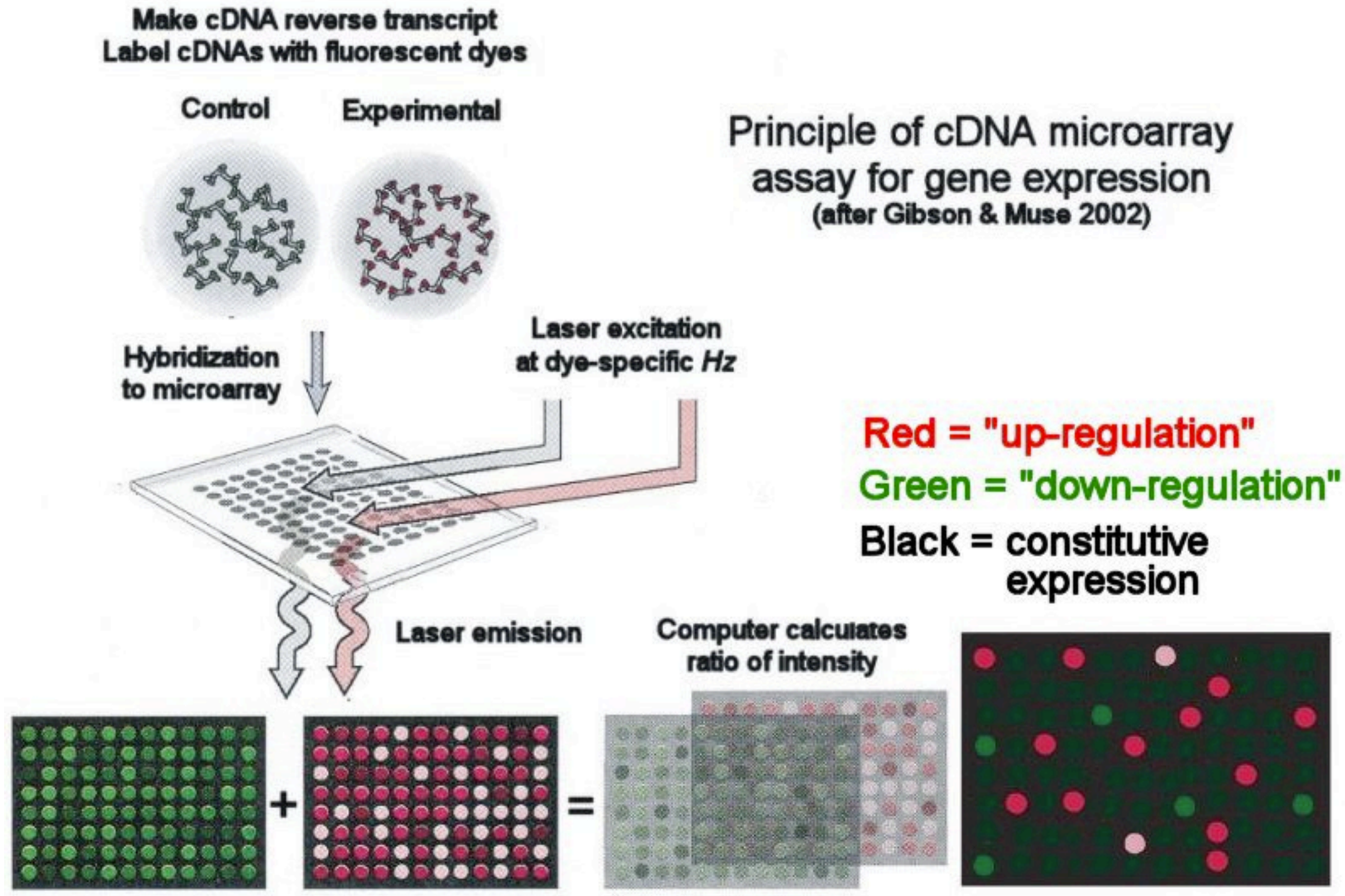
Transcriptome: genome-wide gene expression profile

The transcriptome is the complete set of transcripts and their relative levels of expression in a particular tissue or cell type in a particular condition (whole organism)

- Microarrays: they are suitable for contrasting expression levels across tissues and treatments of a **chosen subset of the genome**, but they do not provide data on **ABSOLUTE** levels of expression.

- RNAseq: not limited to known genes/transcripts. Allow the measurement of relative and absolute level of expression. It is also used for genome annotation (as cDNA libraries).

Microarrays



Gibson and Muse 2009, A primer of genome science

Microarrays workflow

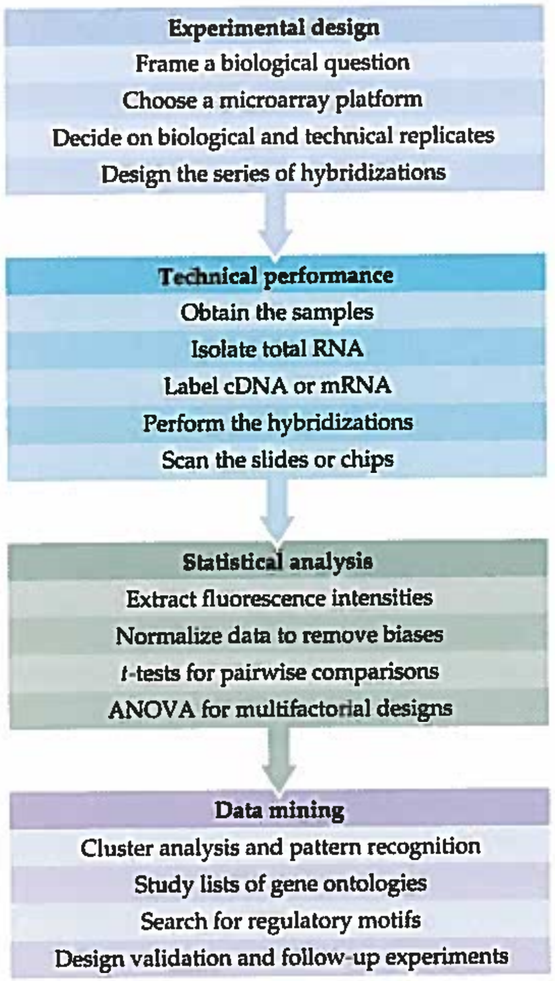


Figure 4.1 Flow diagram for gene expression profiling using microarrays.

Gibson and Muse 2009, A primer of genome science

Microarrays workflow

Common to RNAseq workflow

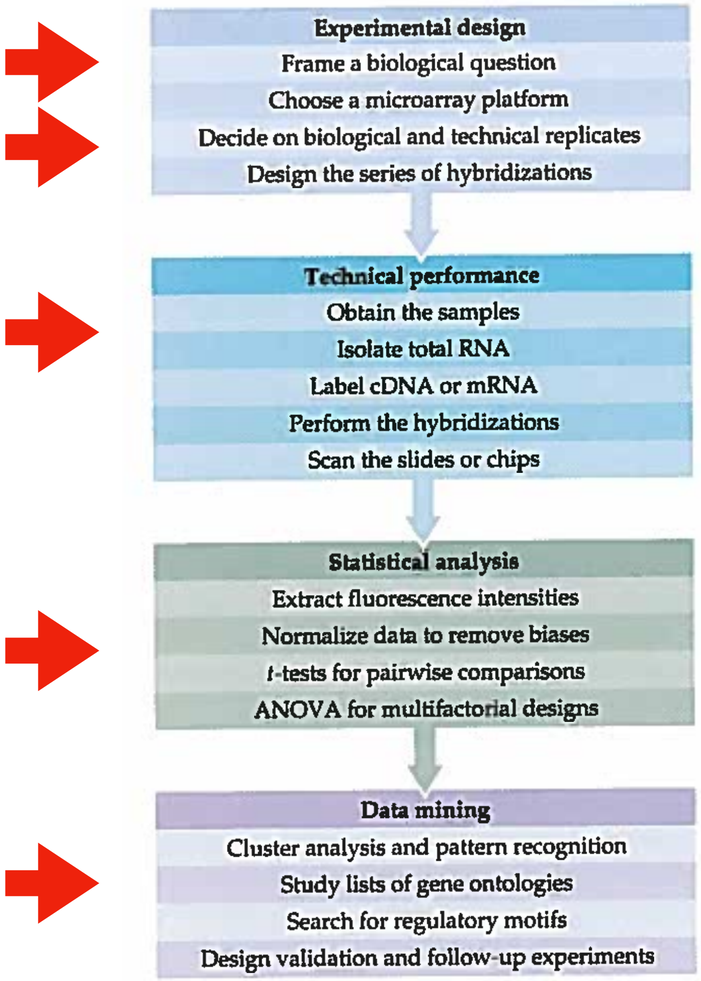
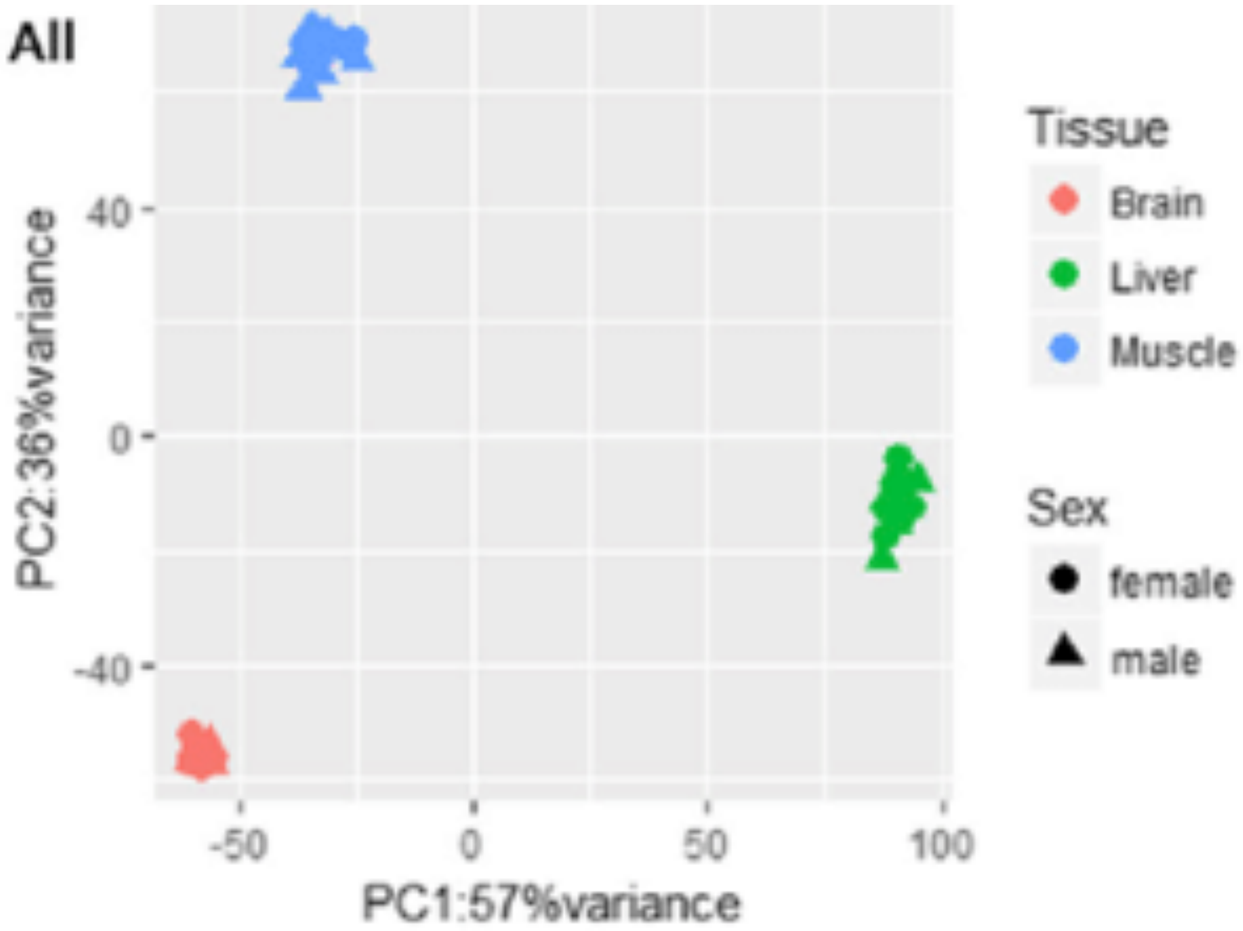


Figure 4.1 Flow diagram for gene expression profiling using microarrays.

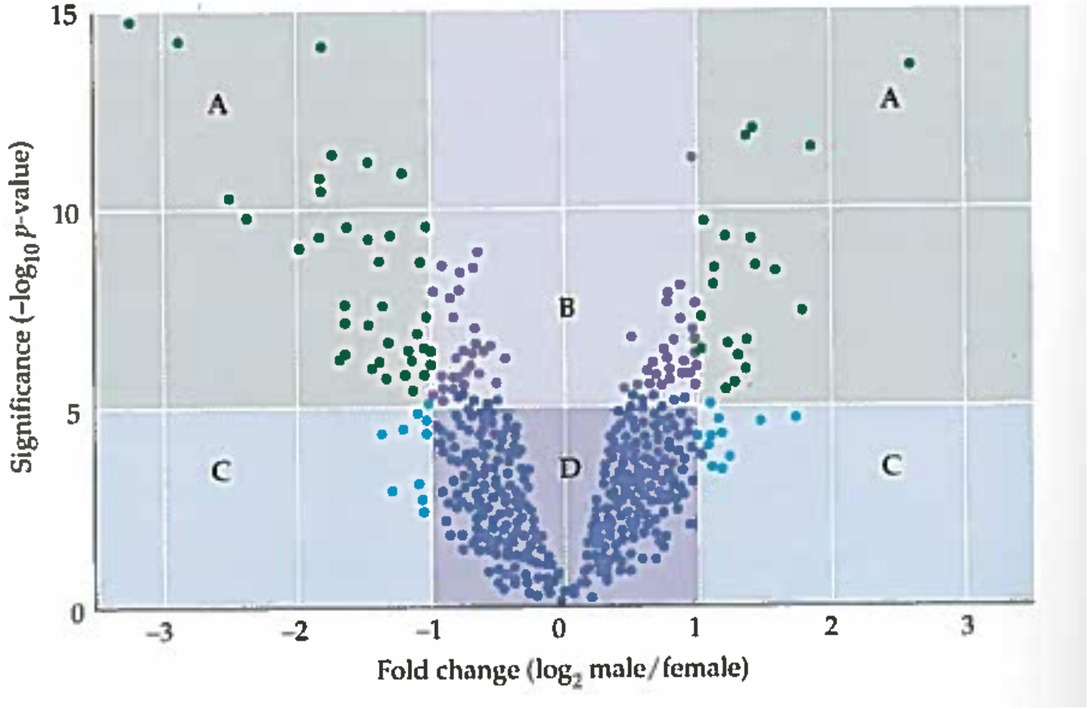
Differential gene expression with microarrays



Principal Component analysis (PCA)

One of the first steps to explore the source of variation in your data

Differential gene expression with microarrays



Volcano plot

Heatmaps

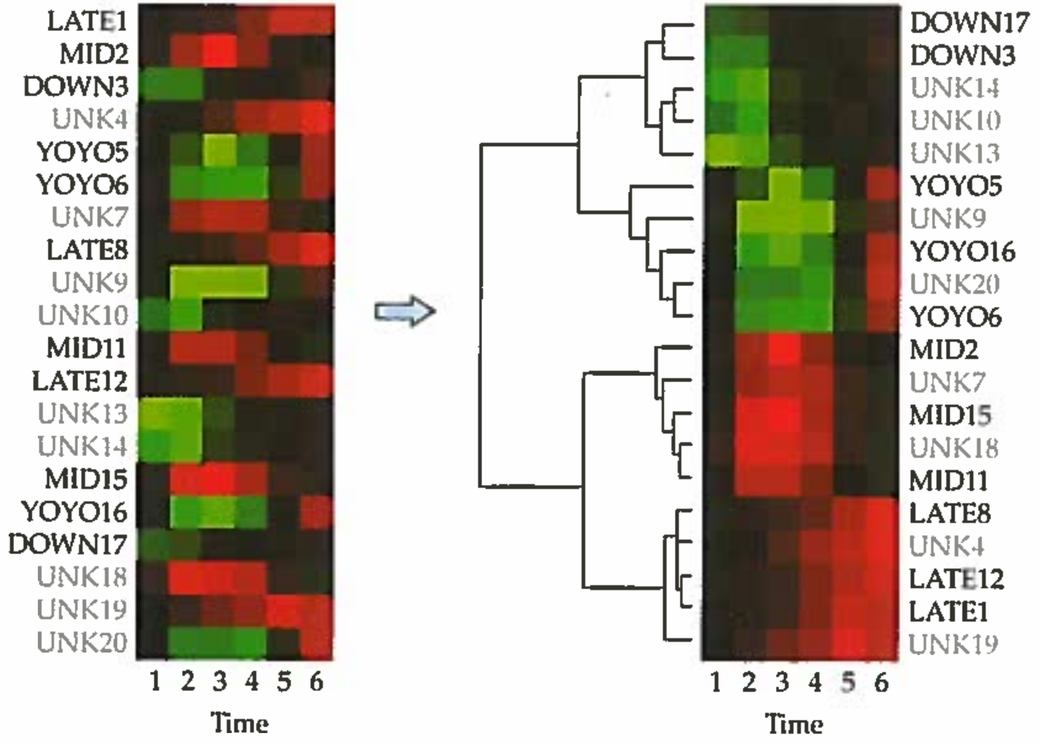


Figure 4.11 Hierarchical clustering of gene expression. An initially disordered set of gene expression profiles (left) can be converted into an immediately intelligible set of clusters by hierarchical clustering and rendering of the profiles in color, as in this hypothetical TreeView representation of a time-series with 20 genes (right). The observation that the genes of the DOWN, YOYO, MID, and LATE classes cluster together suggests that the unknown (UNK) genes may have functions of the respective groups in which they cluster.

Gibson and Muse 2009, A primer of genome science

doi: 10.1111/j.1365-294X.2005.02654.x

Gene expression patterns accompanying a dietary shift in *Drosophila melanogaster*

L. D. CARSTEN, T. WATTS and T. A. MARKOW
Department of Ecology and Evolutionary Biology, 310 Biosciences West, University of Arizona, Tucson, AZ 85721, USA

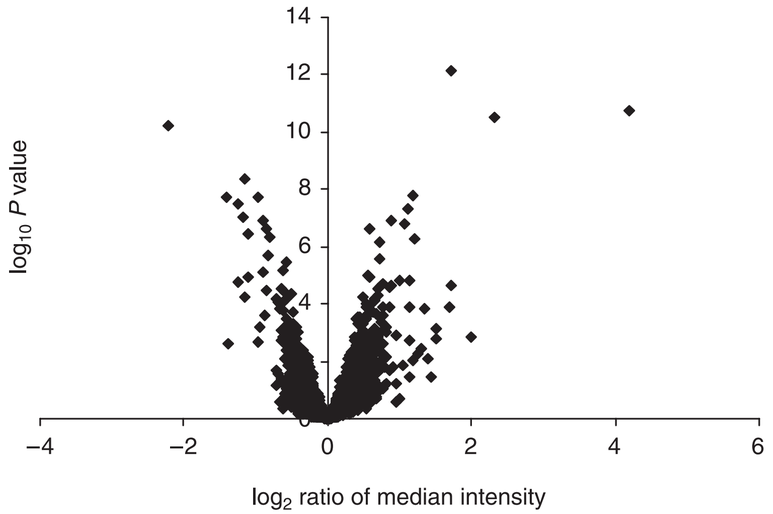


Fig. 1 Volcano plot of the 5620 genes examined in this study. Each spot represents a gene, with the $\log_{10}$ of the *P* value plotted against the $\log_2$ ratio of quantile-transformed median fluorescence intensity for each gene. Genes to the right of zero are down-regulated in the banana treatment, while genes to the left are up-regulated. A difference of one unit on the *x*-axis represents a twofold change in fluorescence intensity. Genes further away from the zero value on that axis show a higher difference in intensity of expression, while genes above the value 3 (equivalent to a *q* value ≤ 0.05) on the *y*-axis show statistically significant differences in expression.

Table 1 List of genes with expression changes in response to a dietary shift from cornmeal media to pure banana. Genes shaded grey were down-regulated, and unshaded genes were up-regulated. In total, 50 genes were down-regulated and 40 genes were up-regulated; this table shows the subset of 36 down-regulated and 19 up-regulated genes that have been named and/or have a known function

Annotation ID	$\log_2$ ratio	*P* value	*q* value	*Name*/Function	Biological process
CG4919	0.72	6.74E-07	1.90E-04	*Gclm*/glutathione gamma-glutamylcysteinyltransferase	glutathione biosynthesis
CG8782	0.56	2.31E-04	2.03E-02	*Oat*/ornithine-oxo-acid transaminase activity	amino acid biosynthesis
CG18730	1.13	1.34E-04	0.013	*Amy-p*/calcium ion binding, alpha-amylase activity, carb. metabolism	carbohydrate metabolism
CG8256	−0.68	8.33E-05	9.36E-03	*l(2)k05713*/glycerol-3-phosphate dehydrogenase	glycerol metabolism
CG15096	−0.63	1.76E-04	0.017	carb transport/metabolism, phosphate/cation transport	carbohydrate metabolism
CG1869	0.81	6.04E-04	4.24E-02	chitin binding	polysaccharide metabolism
CG10072	−0.60	3.93E-05	5.66E-03	*sgl*/UDP-glucose 6-dehydrogenase	heparan sulphate proteoglycan biosynthesis, polysaccharide chain biosynthesis
CG9466	2.32	3.15E-11	5.91E-08	alpha mannosidase	carbohydrate metabolism
CG9441	0.43	5.12E-04	0.037	*Pu*/GTP cyclohydrolase I activity	tetrahydrobiopterin biosynthesis
CG10800	−0.62	6.61E-06	1.55E-03	*Rca1*/mitosis regulation	regulation of mitosis
CG32019	0.68	2.06E-04	1.84E-02	*bt*/myosin-light-chain kinase activity	mesoderm development
CG5112	0.50	4.99E-04	3.64E-02	fatty acid amide hydrolase activity	nitrogen compound metabolism
CG7400	0.79	7.29E-04	4.71E-02	*Fatp*/long-chain fatty acid transporter	fatty acid metabolism
CG3415	0.49	4.31E-04	0.032728	estradiol 17-beta-dehydrogenase activity	fatty acid biosynthesis
GH23546	−0.49	6.18E-04	4.28E-02	*desat 1*/fatty acid biosynthesis stearoyl-CoA desaturase	fatty acid biosynthesis
CG5009	0.53	9.66E-05	1.06E-02	palmitoyl-CoA oxidase activity, fatty acid beta-oxidation	fatty acid beta-oxidation
CG3902	0.73	3.03E-04	2.45E-02	short branched chain acyl-CoA dehydrogenase	acyl-CoA metabolism
CG17597	0.77	1.30E-04	1.32E-02	acyl-CoA metabolism, lipid transport, fatty acid beta-oxidation	fatty acid beta-oxidation

RNA-seq

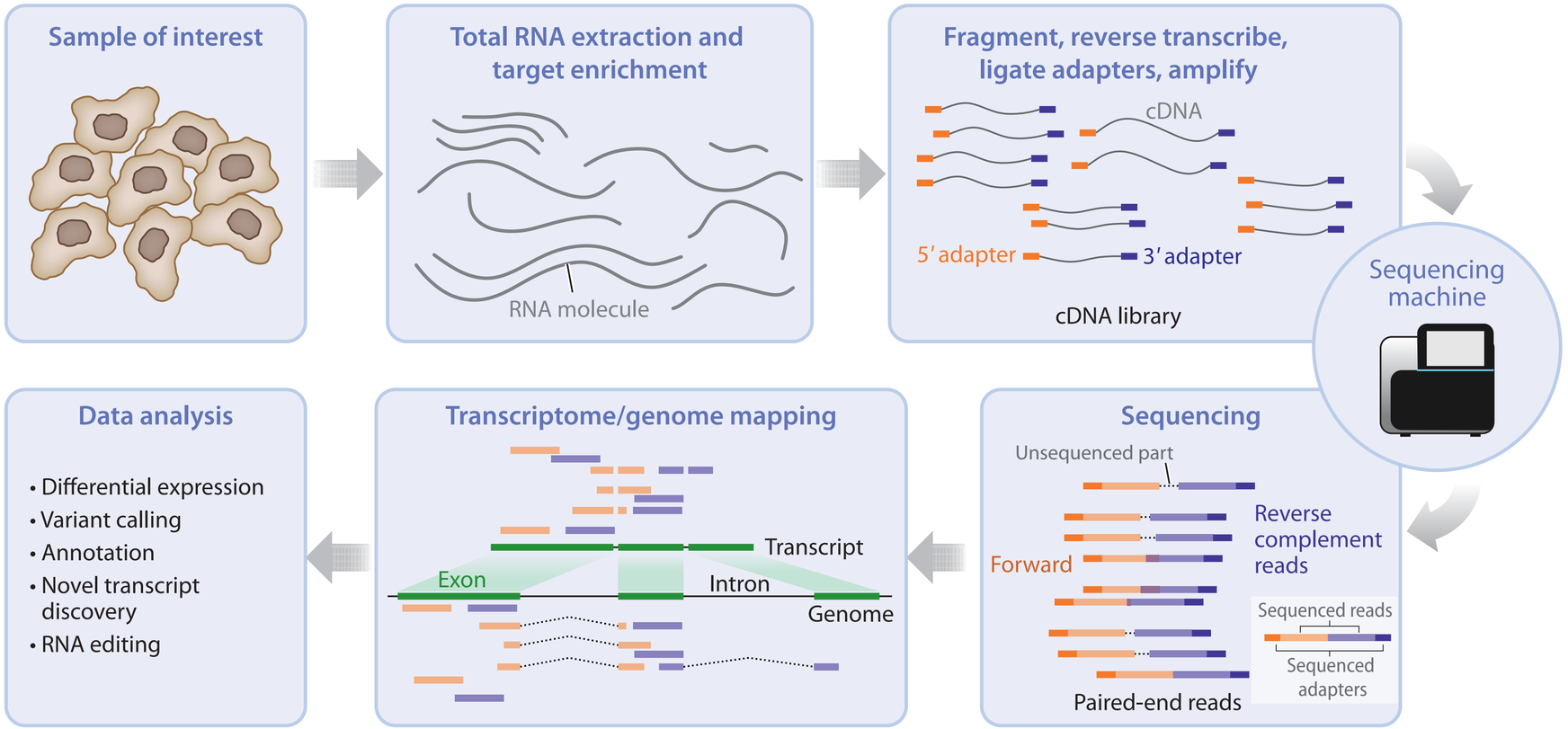


Figure 1

Overview of the experimental steps in an RNA sequencing (RNA-seq) protocol. The complementary DNA (cDNA) library is generated from isolated RNA targets and then sequenced, and the reads are mapped against a reference genome or transcriptome. Downstream data analysis depends on the goal of the experiment and can include, among other things, assessing differential expression, variant calling, or genome annotation.

Berge et al. 2019 *Annual Review of Biomedical Data Science*

RNAseq vs microarrays

Table 1 | **Advantages of RNA-Seq compared with other transcriptomics methods**

Technology	Tiling microarray	cDNA or EST sequencing	RNA-Seq
Technology specifications			
Principle	Hybridization	Sanger sequencing	High-throughput sequencing
Resolution	From several to 100 bp	Single base	Single base
Throughput	High	Low	High
Reliance on genomic sequence	Yes	No	In some cases
Background noise	High	Low	Low
Application			
Simultaneously map transcribed regions and gene expression	Yes	Limited for gene expression	Yes
Dynamic range to quantify gene expression level	Up to a few-hundredfold	Not practical	>8,000-fold
Ability to distinguish different isoforms	Limited	Yes	Yes
Ability to distinguish allelic expression	Limited	Yes	Yes
Practical issues			
Required amount of RNA	High	High	Low
Cost for mapping transcriptomes of large genomes	High	High	Relatively low

Wang Z, Gerstein M, Snyder M (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet

RNAseq vs microarrays

Table 1 | **Advantages of RNA-Seq compared with other transcriptomics methods**

Technology	Tiling microarray	cDNA or EST sequencing	RNA-Seq
Technology specifications			
Principle	Hybridization	Sanger sequencing	High-throughput sequencing
Resolution	From several to 100 bp	Single base	Single base
Throughput	High	Low	High
Reliance on genomic sequence	Yes	No	In some cases
Background noise	High	Low	Low
Application			
Simultaneously map transcribed regions and gene expression	Yes	Limited for gene expression	Yes
Dynamic range to quantify gene expression level	Up to a few-hundredfold	Not practical	>8,000-fold
Ability to distinguish different isoforms	Limited	Yes	Yes
Ability to distinguish allelic expression	Limited	Yes	Yes
Practical issues			
Required amount of RNA	High	High	Low
Cost for mapping transcriptomes of large genomes	High	High	Relatively low

Wang Z, Gerstein M, Snyder M (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet

Library type

Library design	Usage	Description
Poly-A selection	Sequencing mRNA	Select for RNA species with poly-A tail and enriches for mRNA
Ribo-depletion	Sequencing mRNA, pre-mRNA, ncRNA	Removes ribosomal RNA and enriches for mRNA, pre-mRNA, and ncRNA
Size selection	Sequencing miRNA	Selects RNA species using size fractionation by gel electrophoresis
Duplex-specific nuclease	Reduce highly abundant transcripts	Cleaves highly abundant transcripts, including rRNA and other highly expressed genes
Strand-specific	De novo transcriptome assembly	Preserves strand information of the transcript
Multiplexed	Sequencing multiple samples together	Genetic barcoding method that enables sequencing multiple samples together
Short-read	Higher coverage	Produces 50–100 bp reads; generally higher read coverage and reduced error rate compared to long-read sequencing
Long-read	De novo transcriptome assembly	Produces >1000 bp reads; advantageous for resolving splice junctions and repetitive regions

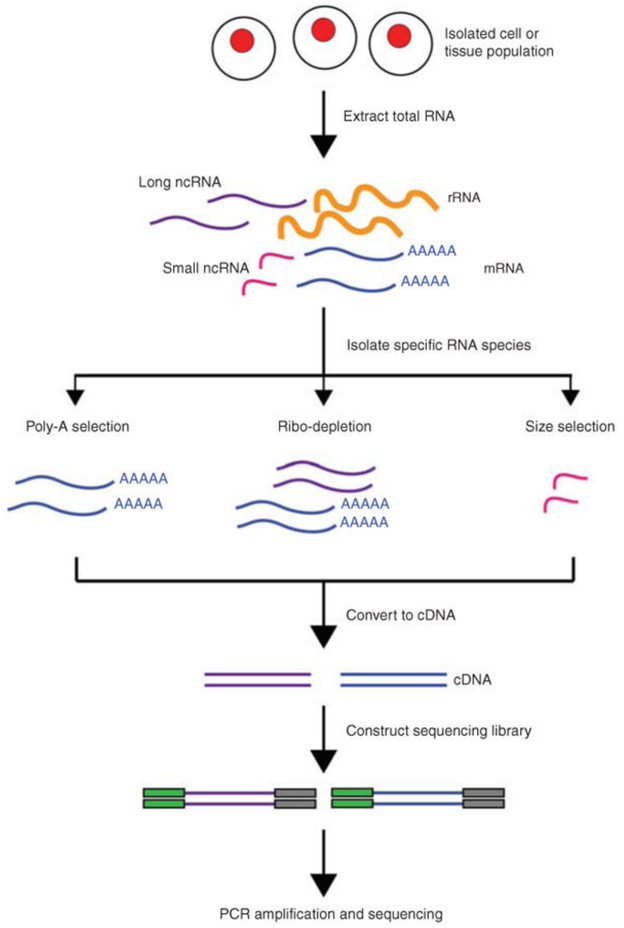
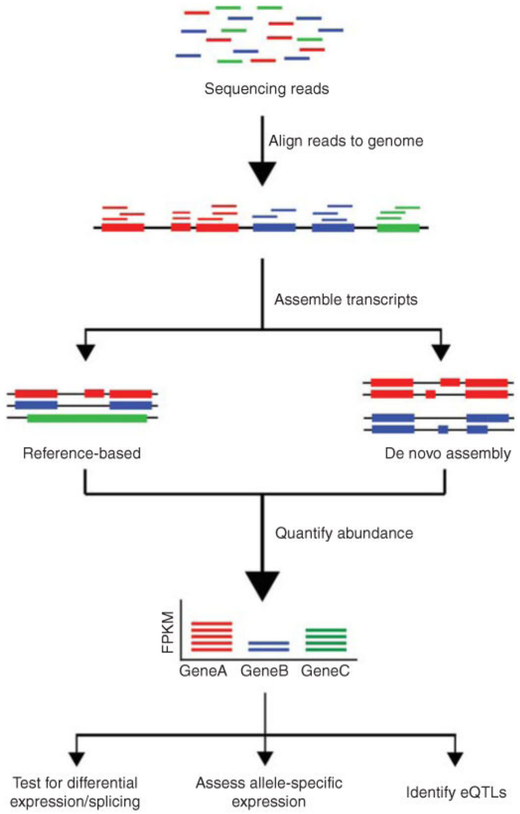


Figure 1.
Overview of RNA-Seq. First, RNA is extracted from the biological material of choice (e.g., cells, tissues). Second, subsets of RNA molecules are isolated using a specific protocol, such as the poly-A selection protocol to enrich for polyadenylated transcripts or a ribo-depletion protocol to remove ribosomal RNAs. Next, the RNA is converted to complementary DNA (cDNA) by reverse transcription and sequencing adaptors are ligated to the ends of the cDNA fragments. Following amplification by PCR, the RNA-Seq library is ready for sequencing.

Kukurba KR, Montgomery SB (2015). RNA Sequencing and Analysis. Cold Spring Harb Protoc

Analysis workflow

Reference-based
Mapping of reads



Classification-based
Pseudo-mapping of reads

Figure 2.
Overview of RNA-Seq data analysis. Following typical RNA-Seq experiments, reads are first aligned to a reference genome. Second, the reads may be assembled into transcripts using reference transcript annotations or de novo assembly approaches. Next, the expression level of each gene is estimated by counting the number of reads that align to each exon or full-length transcript. Downstream analyses with RNA-Seq data include testing for differential expression between samples, detecting allele-specific expression, and identifying expression quantitative trait loci (eQTLs).

Kukurba KR, Montgomery SB (2015). RNA Sequencing and Analysis. Cold Spring Harb Protoc

nature COMMUNICATIONS

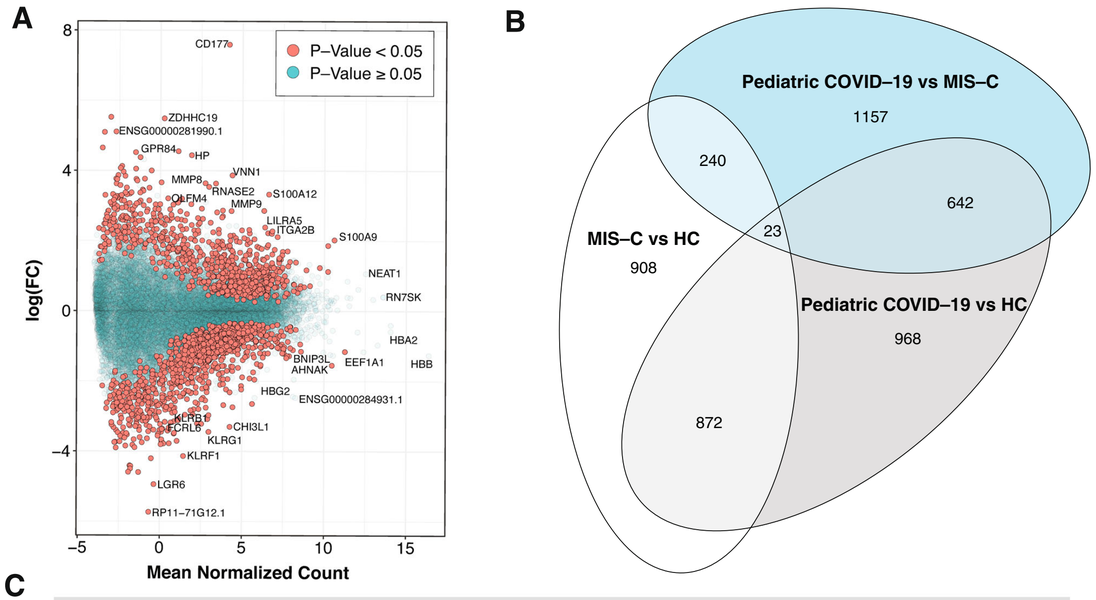
ARTICLE

https://doi.org/10.1038/s41467-021-24981-1 OPEN

Check for updates

Downregulation of exhausted cytotoxic T cells in gene expression networks of multisystem inflammatory syndrome in children

Downregulation of NK cells and cytotoxic T cell in gene expression networks



C

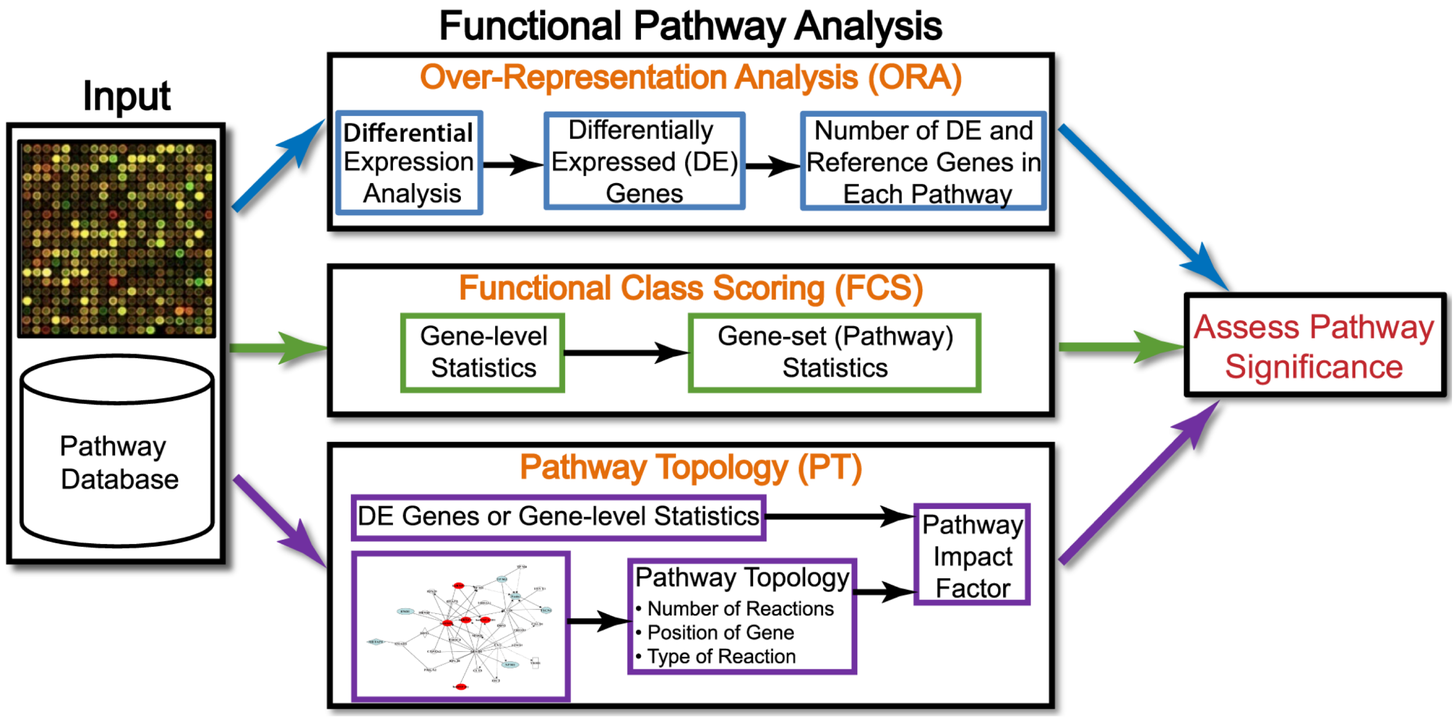
GO Term	Fold Enrichment	Adjusted P-value	DE Comparison
Regulated exocytosis	3.232	1.07E-20	Up in MIS–C
Secretion by cell	2.538	4.19E-20	Up in MIS–C
Myeloid cell activation involved in immune response	3.635	1.27E-19	Up in MIS–C
Neutrophil mediated immunity	3.751	1.27E-19	Up in MIS–C
Leukocyte degranulation	3.657	1.27E-19	Up in MIS–C
Neutrophil degranulation	3.785	1.27E-19	Up in MIS–C
Myeloid leukocyte mediated immunity	3.600	1.41E-19	Up in MIS–C
Neutrophil activation involved in immune response	3.762	1.41E-19	Up in MIS–C
Secretion	2.426	1.41E-19	Up in MIS–C
Neutrophil activation	3.692	3.17E-19	Up in MIS–C
RNA metabolic process	1.409	4.34E-11	Down in MIS–C
Nucleic acid metabolic process	1.357	3.25E-10	Down in MIS–C
Gene expression	1.327	3.04E-09	Down in MIS–C
Cellular nitrogen compound biosynthetic process	1.366	3.04E-09	Down in MIS–C
RNA biosynthetic process	1.452	3.72E-09	Down in MIS–C
Cellular macromolecule biosynthetic process	1.363	3.72E-09	Down in MIS–C
Nucleic acid-templated transcription	1.449	4.58E-09	Down in MIS–C
Transcription, DNA-templated	1.451	4.58E-09	Down in MIS–C
Regulation of transcription by RNA polymerase II	1.561	5.85E-09	Down in MIS–C
Transcription by RNA polymerase II	1.538	6.11E-09	Down in MIS–C

Fig. 2 Differential expression analyses identify the transcriptional signature of MIS-C. A Differential expression (DE) analysis for MIS-C patients versus HCs. The *x*-axis is the mean normalized count for each gene and the *y*-axis is the $\log_2$(fold-change) for the differential expression. Positive and negative $\log_2$(fold change) represent genes upregulated and downregulated in MIS-C, respectively, and the significance of association between gene expression and MIS-C status is indicated by the color of the dots as defined in the legend. **B** Overlap of MIS-C and pediatric COVID-19 transcriptional signatures: Venn diagram of the overlap of genes across DE signatures. Each comparison is labeled on the plot. **C** GO terms for MIS-C signature: GO term enrichment results for the top 10 upregulated and downregulated processes in MIS-C compared to HCs. Two-sided Fisher tests were used and *p*-values were adjusted for multi-testing as described in the "Methods" section. Full DE results and pathway enrichments for all comparisons in **B** can be found in Supplementary Data 7 and 8.

NGS across biology

- **Genome Sequencing**
- **Transcriptome Sequencing (bulk or single-cell)**
- **Protein-DNA interaction - ChIP-seq**
- **Protein-RNA interaction - CLIP-seq**
- **Translatome**
- **Methylome**
- **Open Chromatome**

Pathway analysis



- The data generated by an experiment using a high-throughput technology (e.g., microarray, proteomics, metabolomics), along with functional annotations (pathway database) of the corresponding genome, are input to virtually all pathway analysis methods.
- ORA methods require that the input is a list of differentially expressed genes
- FCS methods use the entire data matrix as input
- PT-based methods additionally utilize the number and type of interactions between gene products, which may or may not be a part of a pathway database.
- The result of every pathway analysis method is a list of significant pathways in the condition under study.