ICQB Lecture 10:

A practical guide to RNA-seq analysis

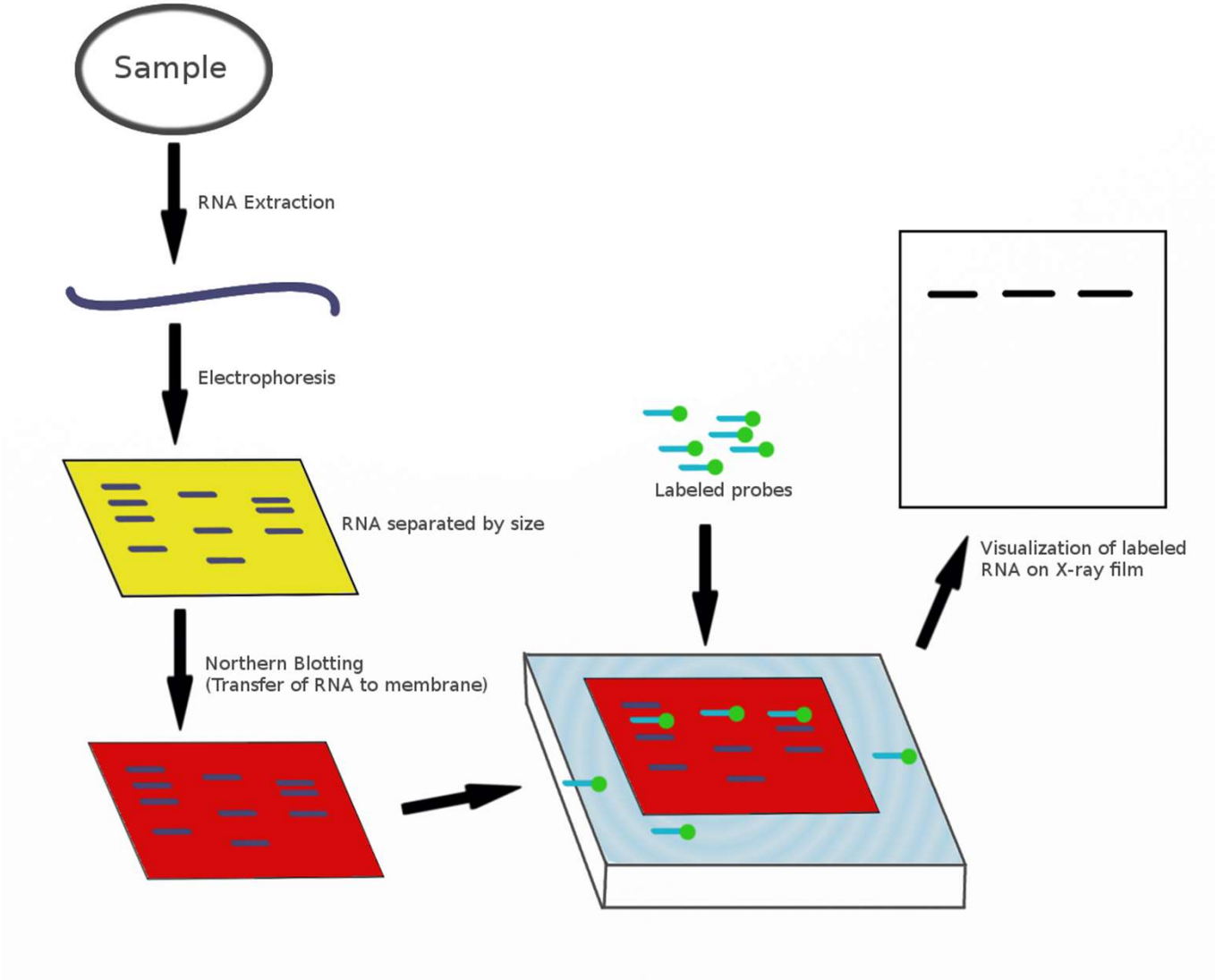
G4120

Thomas Postler

29 November, 2022

A (very) brief history of transcriptomics

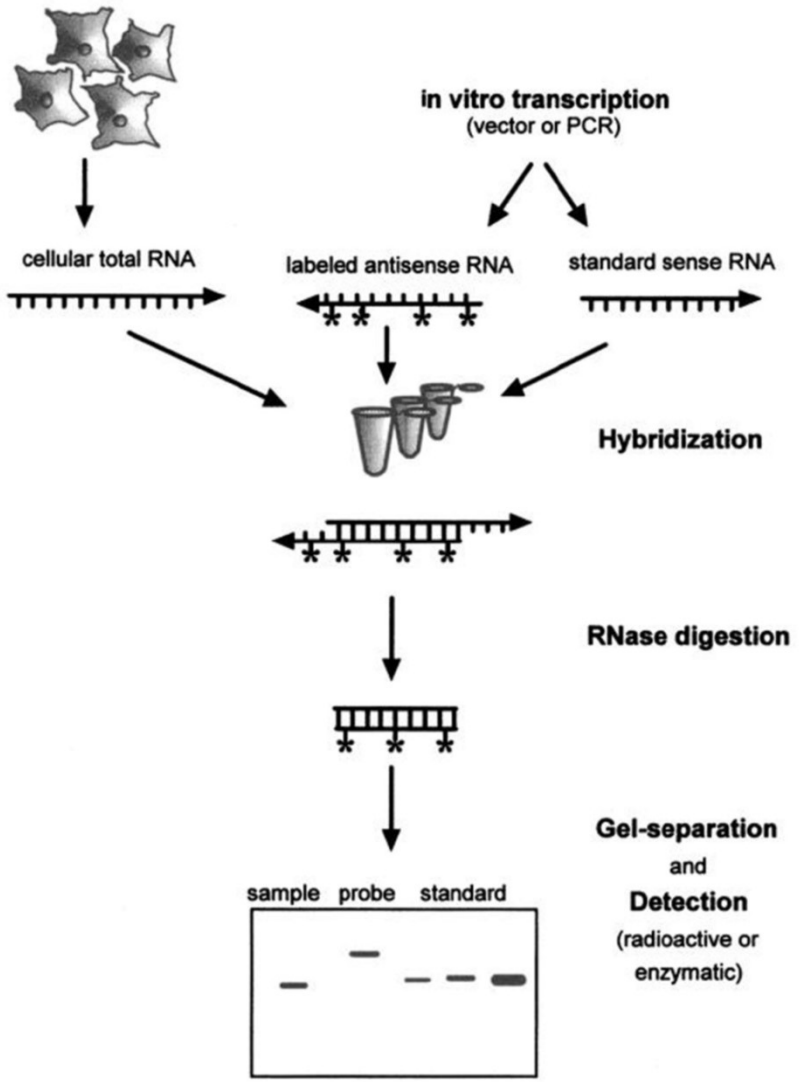
- 1977: Development of the Northern blot:



wikipedia.org

A (very) brief history of transcriptomics

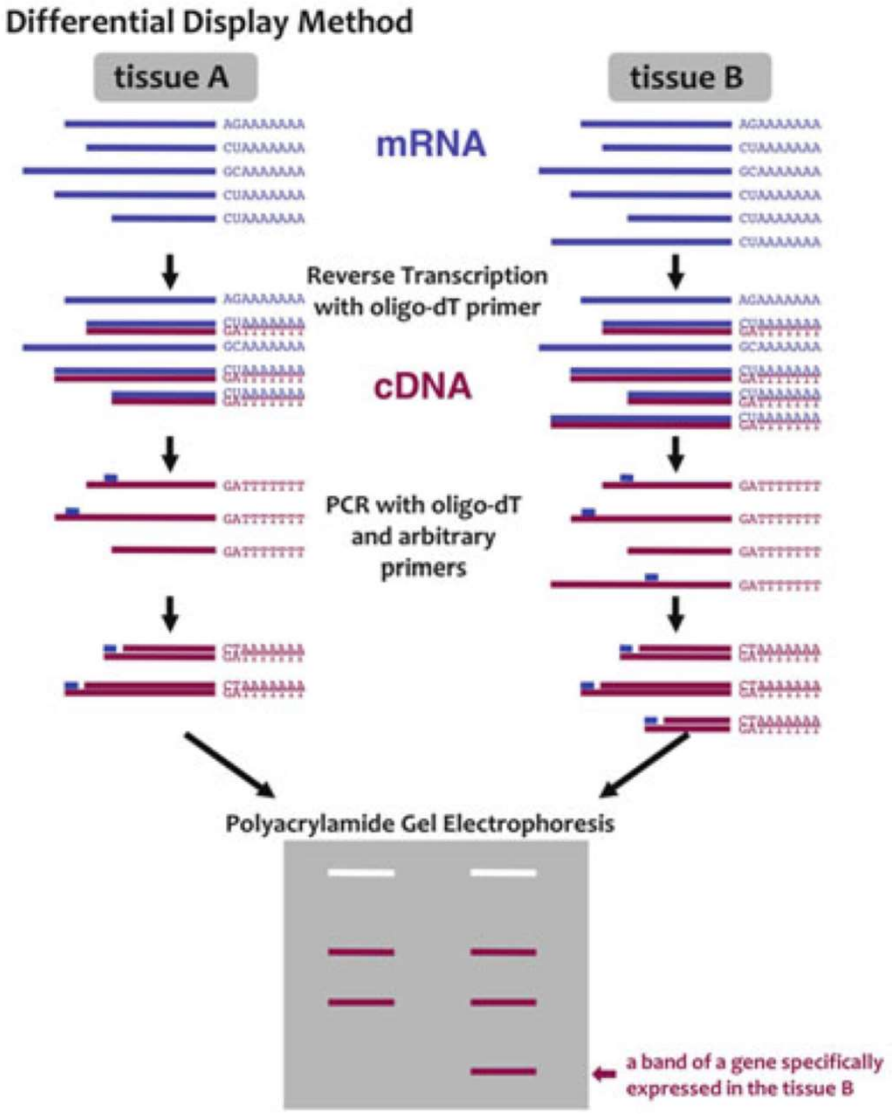
- 1977: Development of the Northern blot.
- → Alternatives targeting specific genes: Nuclease protection assay, reverse-transcription (semi-)quantitative PCR.



Rapley and Walker (eds., 1998), *Molecular Biomethods Handbook*

A (very) brief history of transcriptomics

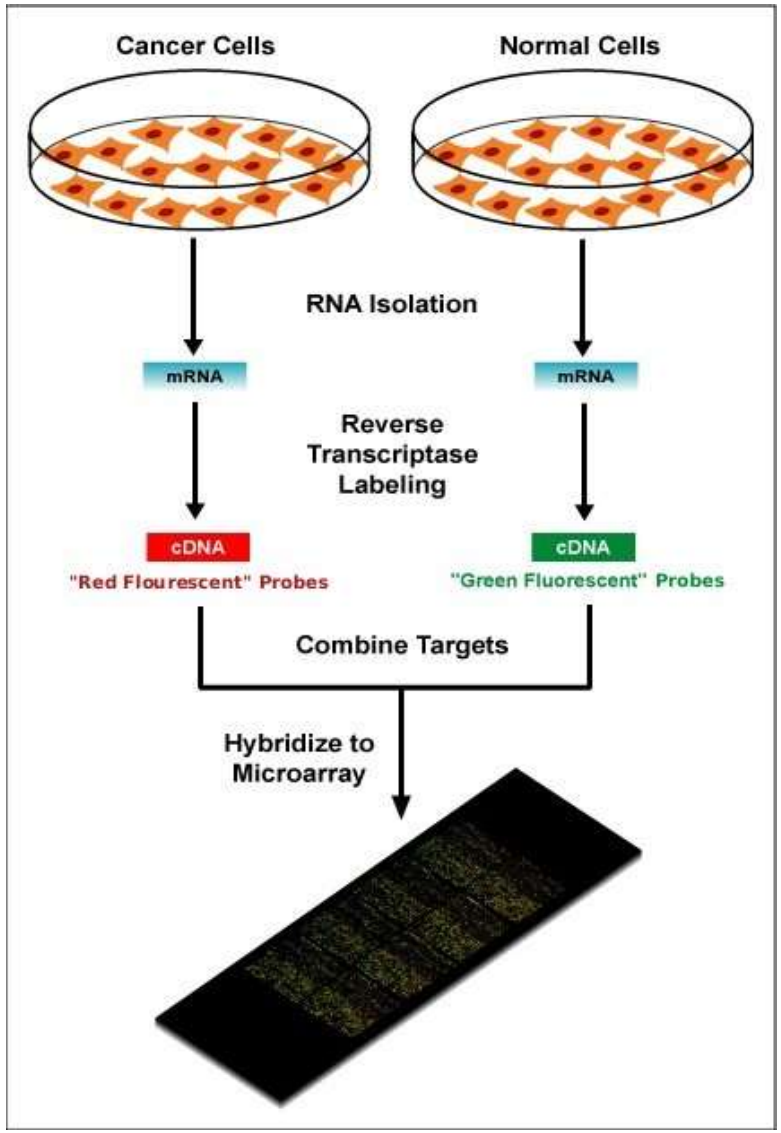
- 1977: Development of the Northern blot.
- → Alternatives targeting specific genes: Nuclease protection assay, reverse-transcription (semi-)quantitative PCR.
- 1992: Differential-display RT-PCR:
  First method to detect changes in global transcription profile.



Bignell *et al.*(eds., 2011), *Biology of Termines: A Modern Synthesis*

A (very) brief history of transcriptomics

- 1977: Development of the Northern blot.
- → Alternatives targeting specific genes: Nuclease protection assay, reverse-transcription (semi-)quantitative PCR.
- 1992: Differential-display RT-PCR:
  First method to detect changes in global transcription profile.
- Early 2000s: Microarrays become viable:
  The first truly quantitative method capturing the (known) transcriptome.



wikipedia.org

A (very) brief history of transcriptomics

- 1977: Development of the Northern blot.
- → Alternatives targeting specific genes: Nuclease protection assay, reverse-transcription (semi-)quantitative PCR.
- 1992: Differential-display RT-PCR:
  First method to detect changes in global transcription profile.
- Early 2000s: Microarrays become viable:
  The first truly quantitative method capturing the (known) transcriptome.
- 2010: Illumina HiSeq 2000 is released:
  Affordable RNA-seq has arrived!

illumina.com

A (very) brief history of transcriptomics

- 1977: Development of the Northern blot.

- → Alternatives targeting specific genes: Nuclease protection assay, reverse-transcription (semi-)quantitative PCR.

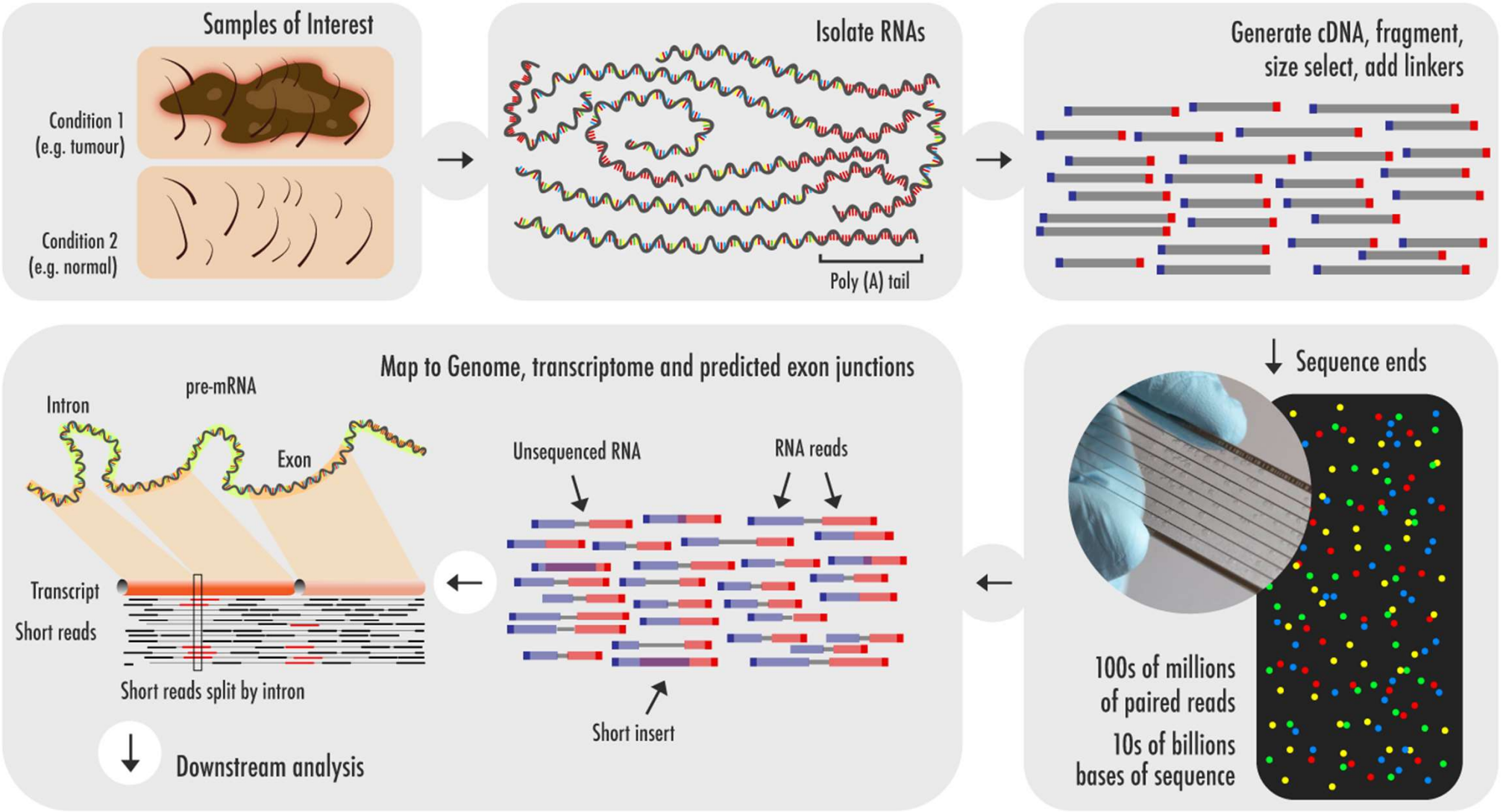
- 1992: Differential-display RT-PCR:
  First method to detect changes
  in global transcription profile.

- Early 2000s: Microarrays become viable:
  The first truly quantitative method
  capturing the (known) transcriptome.

- 2010: Illumina HiSeq 2000 is released:
  Affordable RNA-seq has arrived!

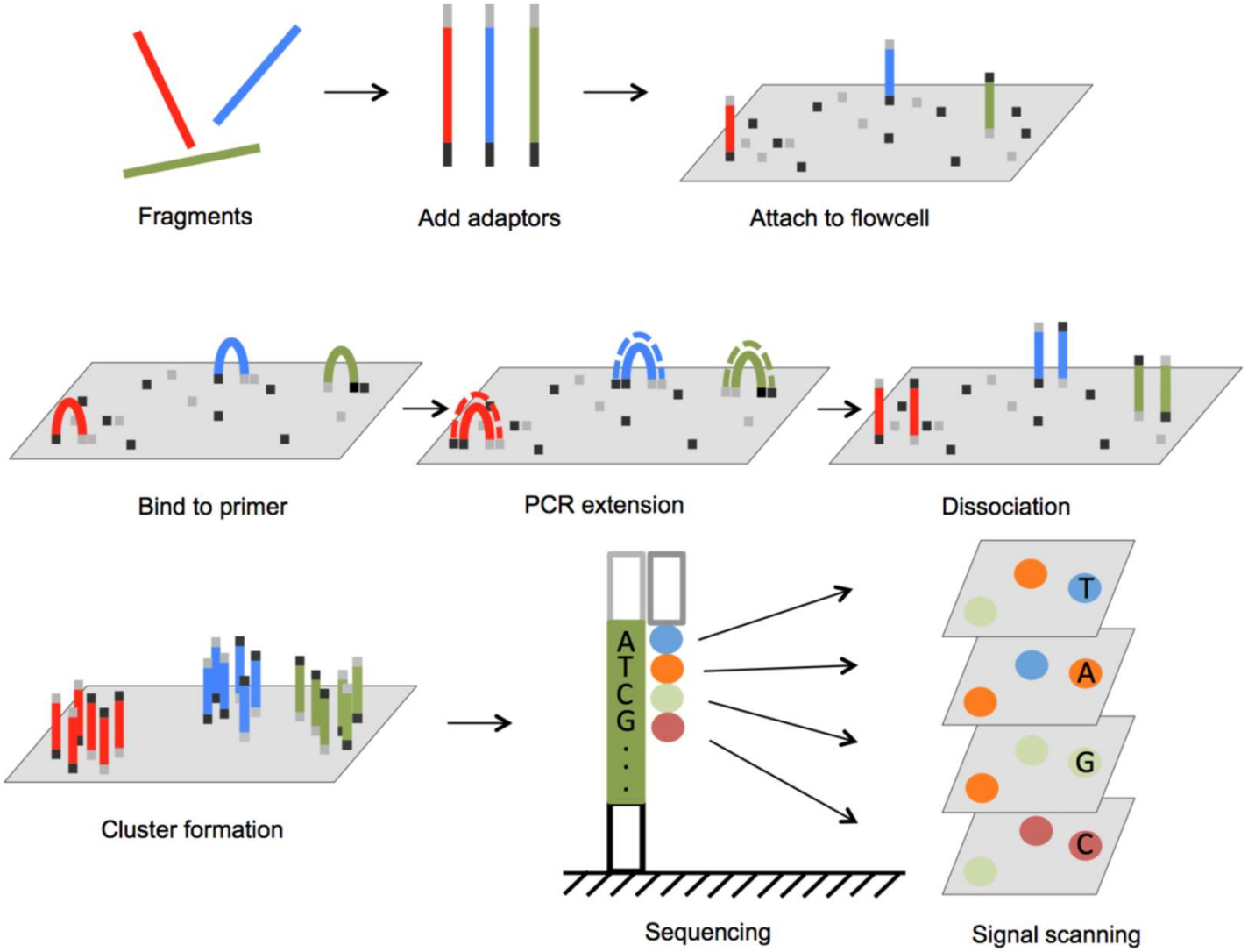
- Notable developments since:
  - PacBio allows for read lengths of > 1 kb.
  - scRNA-seq allows for analysis of single-cell transcriptomics.
  - Novoseq will further reduce the price of RNA-seq, a lot.

Overview of the RNA-seq workflow



technologynetworks.com

Basic principle of RNA-seq



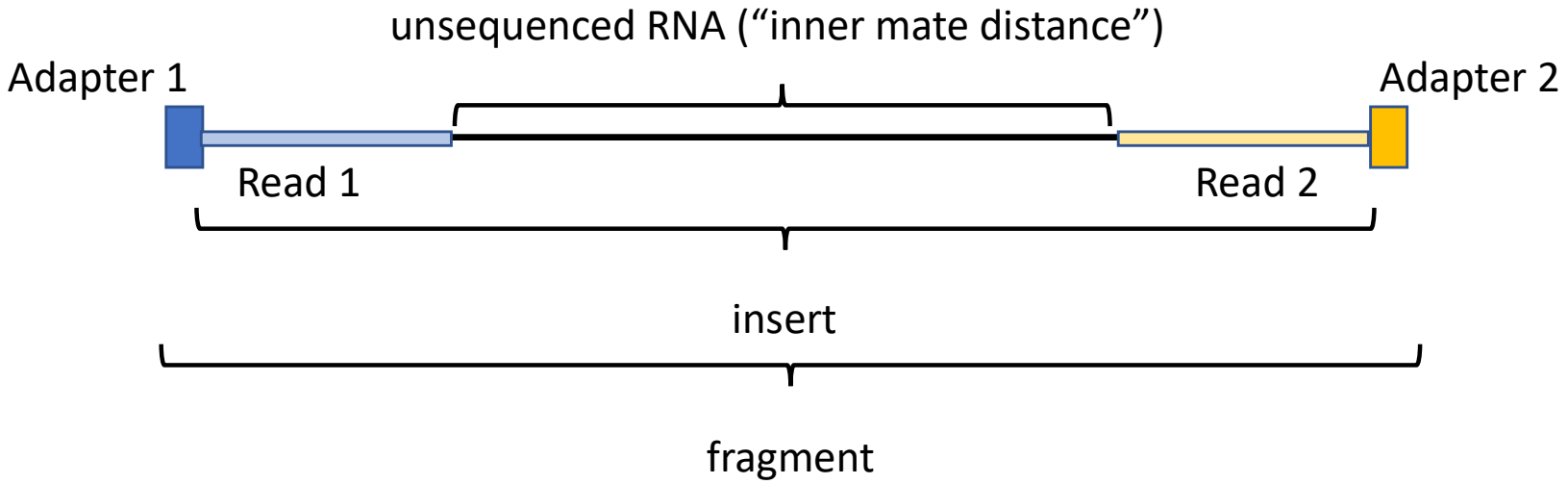
For a snazzy, animated version of this, see https://youtu.be/fCd6B5HRaZ8

Kulski (ed., 2016), *Next Generation Sequencing: Advances, Applications and Challenges*

Key concepts for sequencing

- Typical **lengths** of individual sequence reads during Illumina RNA-seq:
  50, 75, 100 or 150 nt. Longer is better for alignment, but quality drops off after 100.

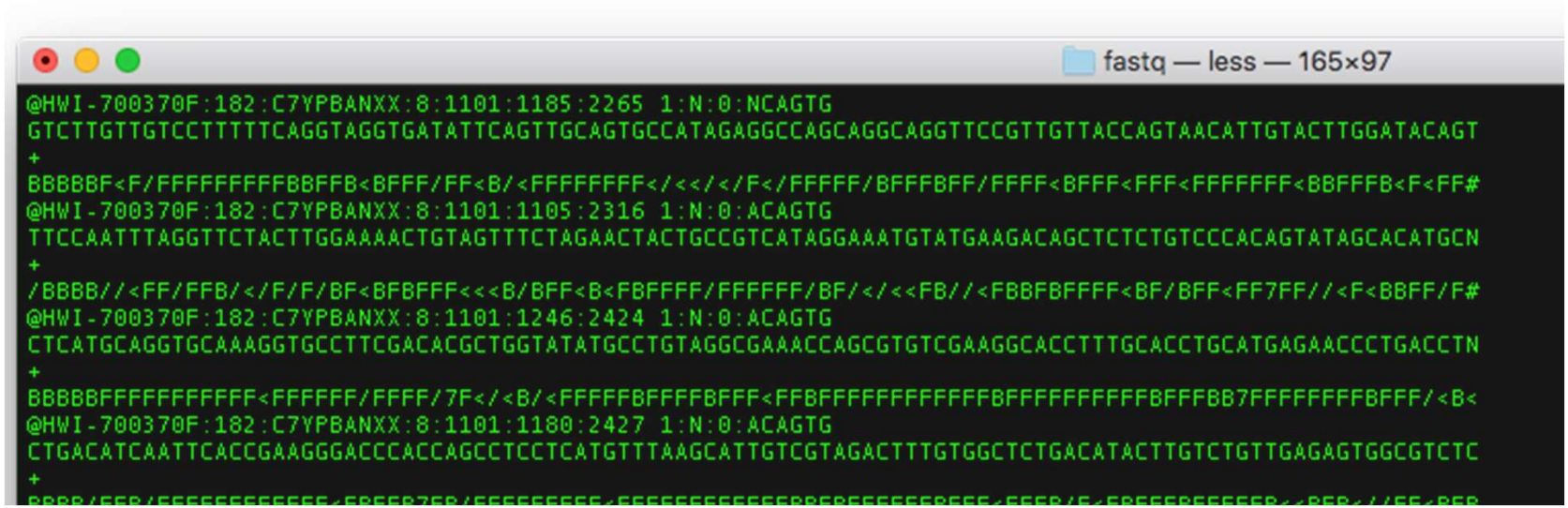
- **Single-end** (SE) vs. **paired-end** (PE) sequencing
  - SE: each (amplified) fragment is only sequenced once, from one direction.
  - PE: each (amplified) fragment is sequenced from *both* directions:



- Sequencing **depth**: the total number of reads created during RNA-seq. Should be >20 million for standard gene-expression analysis without bells and whistles.

- **Phred** (Q) score: each base is assigned a quality score indicating the probability P of the call being wrong. $Q = -10 \log_{10} P$.
  E.g., Q = 10 → P = 0.1 (poor); Q = 40 → P = 0.0001 (good)

Sequencing output: Fastq files



Line 1: ID of read.

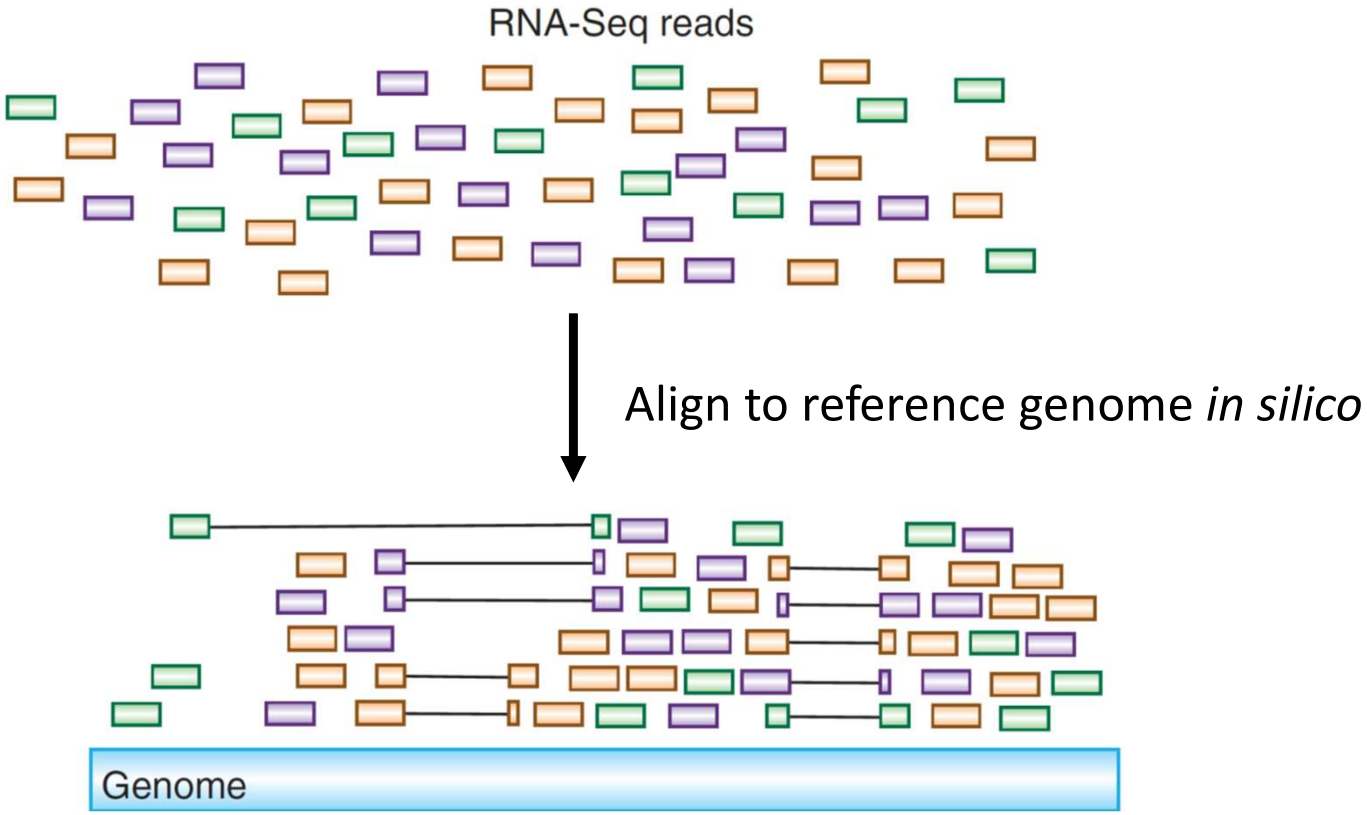
Line 2: actual nucleotide sequence of read.

Line 3: a (+) sign, because why not. (May also be followed by repeat of read ID.)

Line 4: quality of each base call, in ASCII format. Will later be converted to Phred score.

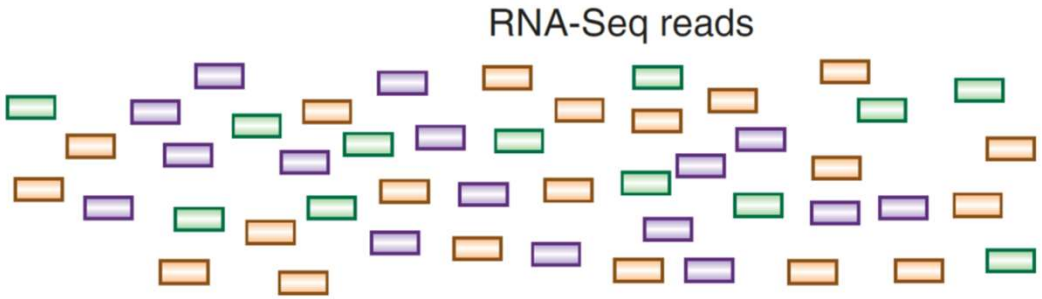
→ Repeat 20-50 million times.

From Fastq to differentially expressed genes

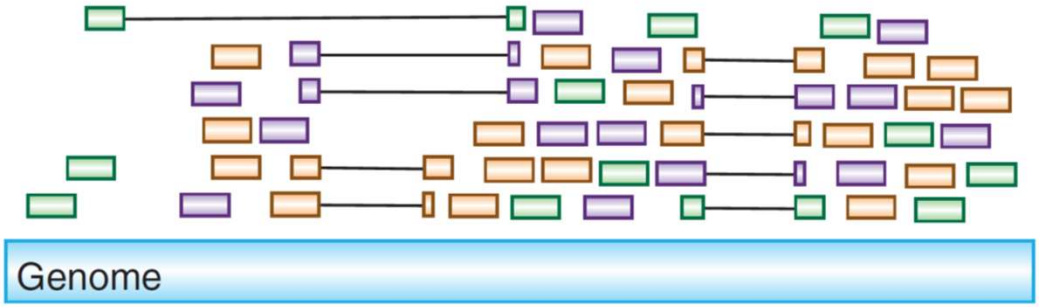


Haas & Zody (2010) *Nat Biotechnol*

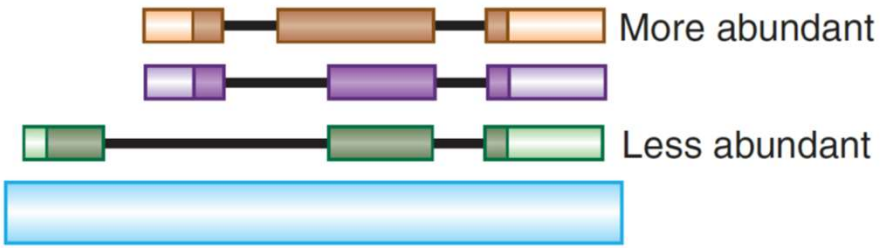
From Fastq to differentially expressed genes



Align to reference genome *in silico*



Count number of reads aligned per gene



Identify meaningful differences in expression

Haas & Zody (2010) *Nat Biotechnol*

The 9-ish steps of RNA-seq analysis

1. Choose the right computational tools.

Recommended step: Clipping and trimming

2. Prepare the reference genome.
3. Align reads to the reference genome.
4. Quality control.
5. Count how many reads align to each gene.
6. Identify differentially expressed genes.
7. Visualize results.
8. Analyze further.
9. Get help.

Step 1: Choose the right computational tools

- There are many, many different tools for each step of this task. Our pipeline uses the following:
  - (Clipping adapters and quality trimming: **Trimmomatic**.)
  - Generation of genome index and alignment: **STAR**.
  - Convert SAM files to BAM files and indexing: **SAMtools**.
  - Quality control: Integrative Genomics Viewer (**IGV**) and **QoRTs**.
  - Count aligned reads: **featureCount** of the Subread package.
  - Statistical analysis and visualization: **RStudio** with various packages.
- Do **not** use Tophat or any other part of the Tuxedo suite! They are popular for historical reasons but perform very poorly compared to more modern tools.
- **Do not, under any circumstances, use Excel – ever!!!!**
  Excel automatically converts certain gene names and other content, because it mistakes it for dates, times, currency, *etc.* This cannot be switched off. *Caveat emptor!*
- Noteworthy alternatives:
  - For the less computationally inclined: **Galaxy** (usegalaxy.org) is an online platform that allows full RNA-seq analysis from start to finish without any need for a command-line interface. Data is directly piped into the relevant programs.
  - **Kallisto** (http://pachterlab.github.io/kallisto/) is an alignment software that uses pseudoalignment, which requires far less computational power. Excellent alternative in the absence of a fast computer (32 GB+ RAM, multi-core CPU).

Recommended Step: Clipping and Trimming

- Clipping:
  Remove any adaptor sequences that may have been incorporated when insert is shorter than read, prior to alignment.

- Trimming:
  Remove low-quality bases prior to alignment.

- Modern alignment programs perform “soft-clipping”, which should eliminate the need for “hard-clipping”; but “hard-clipping” still improves mapping efficiency.

- Should be applied very conservatively! Stringent parameters introduce major biases!

- *Requires sequences of the adapters used during sequencing!*

- A commonly used tool is Trimmomatic:
  http://www.usadellab.org/cms/?page=trimmomatic

- Input: **.fastq** file
  Output: a new **.fastq** file

Step 2: Prepare the reference genome

- This requires two components:
  - The full nucleotide sequence and coordinates of the reference genome. Usually a file designated <*name of organism*>**.fa**.
  - An annotation file with information about genes, transcripts and splice sites. Usually a file designated <*name of organism*>**.gtf**.
- Typical sources:
  - UCSC (https://hgdownload.soe.ucsc.edu/downloads.html).
  - ENSEMBL (https://useast.ensembl.org/info/data/ftp/index.html).
  - Use different formats → risk of compatibility issues downstream.
  - ENSEMBL more broadly compatible and less-rigorously vetted transcripts.
- Alignment programs assemble this information into a genome index for the alignment of reads.
- Input: **.fa** file, **.gtf** file.
  Output: folder with **genome index** files.

Step 3: Align reads to the reference genome

- Input:
  - Folder with **genome index** (just generated).
  - (paired) **.fastq** files.

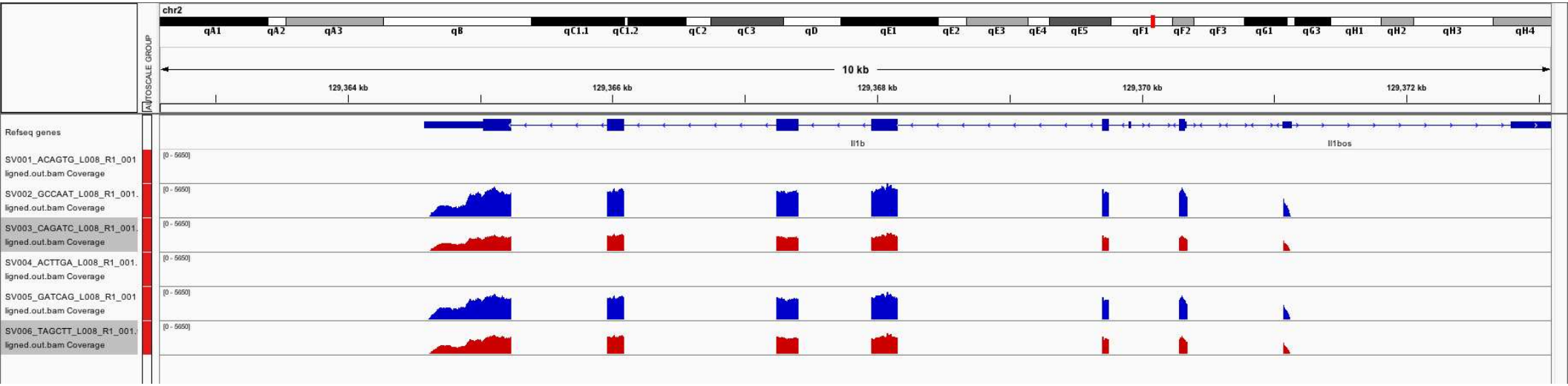
- Output: **.sam** file (for Sequence Alignment Map)
  - Contains 11 mandatory fields for each read with detailed information on sequencing and alignment, including genomic location and quality.
  - For details on each field: https://en.wikipedia.org/wiki/SAM_(file_format).

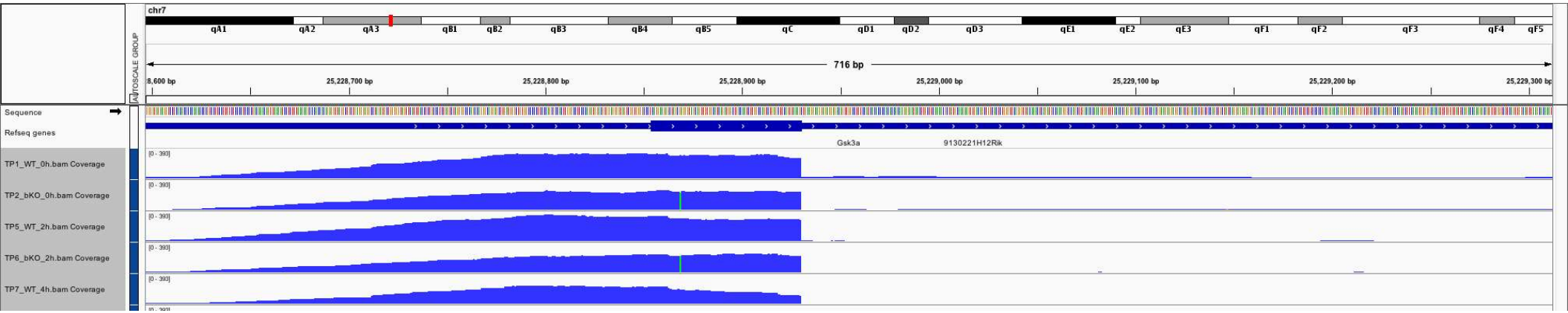
- About 70-90% of reads should be mapped to the genome.

Step 4a: Quality control with IGV

- Requires conversion of **.sam** format to binary **.bam** format (condensed) and indexing, both with SAMtools.

- IGV creates a visual representation of the reads mapped to the genome:





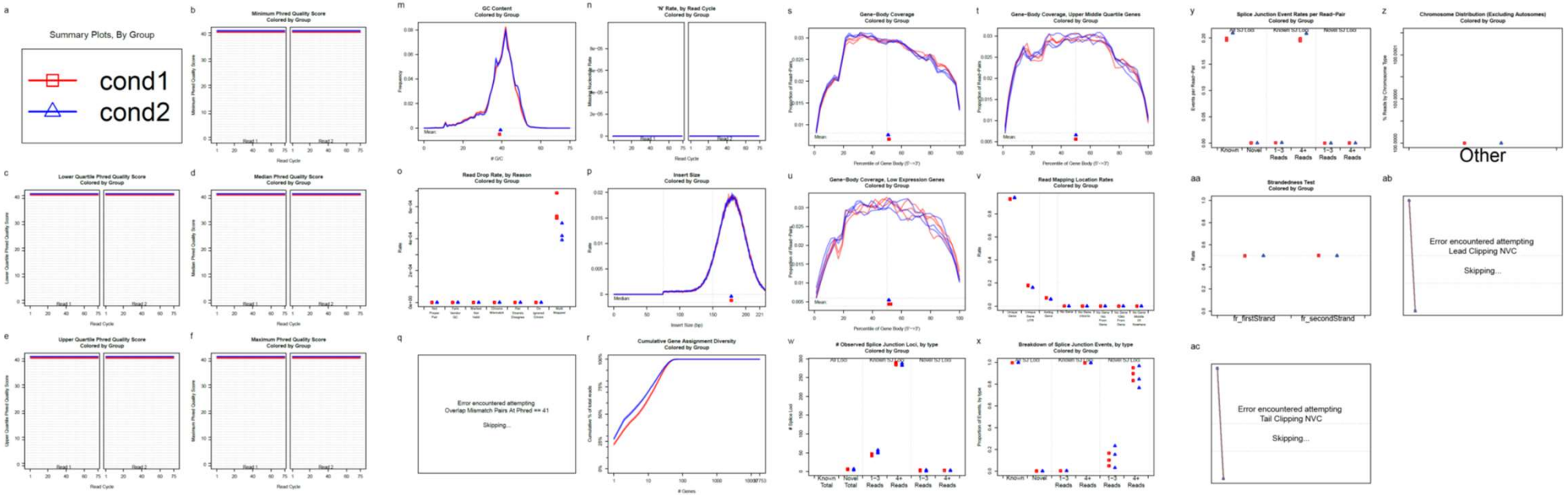
Step 4b: Quality control with QoRTs

- Input:
  - **.bam** format.
  - Requires less computational resources when files are name-sorted → conversion by SAMtools.

- Output: **.pdf** files.

- Performs calculations with Java but employs R to create graphical output.

- Provides graphical assessment of quality for both sequencing and alignment.



Step 5: Count how many reads align to each gene

- Input:
  - **.bam** files.
  - **.gtf** file (genome annotation).
- Output: **.txt** file.
- Consolidates information from all samples into single matrix of counts per gene.

Geneid	S1	S2	S3	S4	S5	S6
ENSMUSG00000103090	0	0	0	0	0	0
ENSMUSG00000025907	1710	1186	1534	1660	965	1525
ENSMUSG00000090031	74	52	25	77	30	36
ENSMUSG00000087247	0	0	0	0	0	0
ENSMUSG00000103355	0	0	0	0	1	0
ENSMUSG00000102706	0	0	0	0	0	0
ENSMUSG00000103845	4	9	17	1	14	32
ENSMUSG00000033740	757	513	924	691	461	989
ENSMUSG00000103629	0	0	0	0	0	0
ENSMUSG00000051285	1662	961	1100	1463	960	1228
ENSMUSG00000098201	0	0	1	0	0	0
ENSMUSG00000103509	37	21	5	26	14	16
ENSMUSG00000048538	1	0	0	0	0	0
ENSMUSG00000103709	0	0	0	0	0	0

Step 6: Identify differentially expressed genes

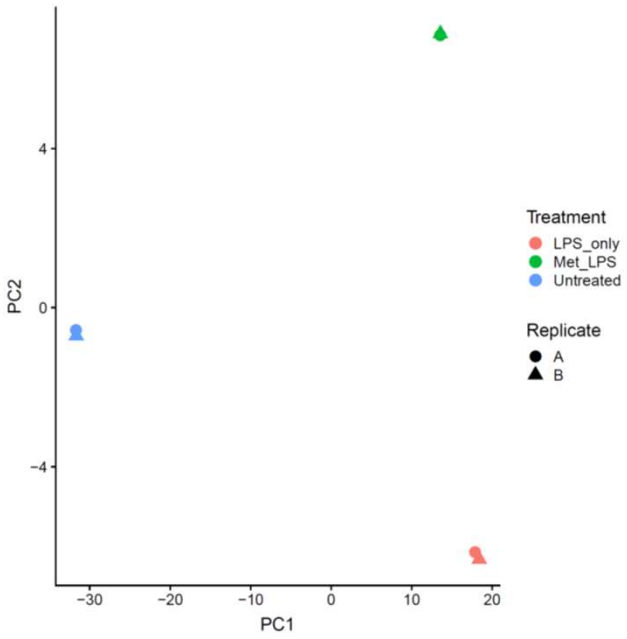
- Input: counts table (**.txt**).
- Output: data table (**.txt** or **.csv**) with fold changes and adjusted p-values.
- Most popular tools are DESeq2 and EdgeR (both R-based).
- Need at least 2 replicates; 3 are better; more are best.
- Can include batch correction.

Geneid	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
ENSMUSG00000104238	0	NA	NA	NA	NA	NA
ENSMUSG00000102269	0.195636	1.986700116	4.993	0.398	0.691	NA
ENSMUSG00000096126	0	NA	NA	NA	NA	NA
ENSMUSG00000103003	0	NA	NA	NA	NA	NA
ENSMUSG00000104328	0	NA	NA	NA	NA	NA
ENSMUSG00000102735	0	NA	NA	NA	NA	NA
ENSMUSG00000098104	1.103537	-0.097401599	2.444	-0.04	0.968	NA
ENSMUSG00000102175	0.650705	0.579256789	3.35	0.173	0.863	NA
ENSMUSG00000088000	0	NA	NA	NA	NA	NA
ENSMUSG00000103265	0	NA	NA	NA	NA	NA
ENSMUSG00000103922	0.874765	1.552357216	2.86	0.543	0.587	NA
ENSMUSG00000033845	572.9266	0.179076205	0.111	1.618	0.106	0.181
ENSMUSG00000102275	3.412223	-0.492101491	1.471	-0.33	0.738	0.827
ENSMUSG00000025903	1070.492	0.168258337	0.084	2.004	0.045	0.086

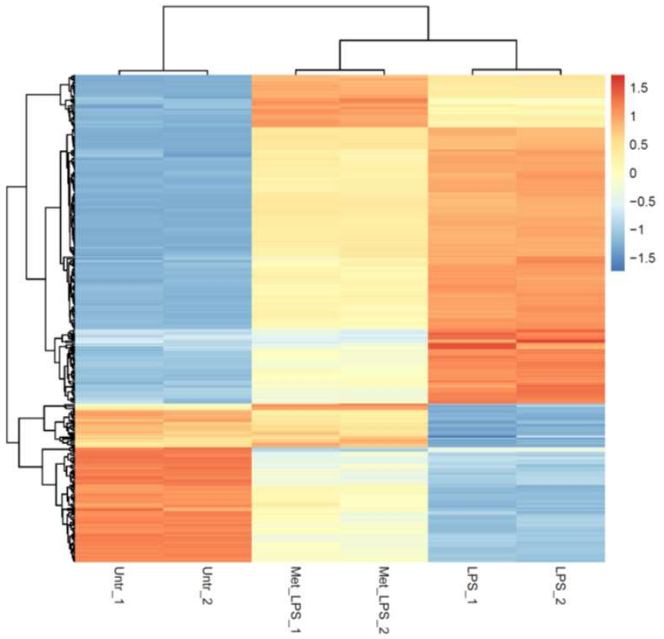
Step 7: Visualize results

- Common initial visualizations include:

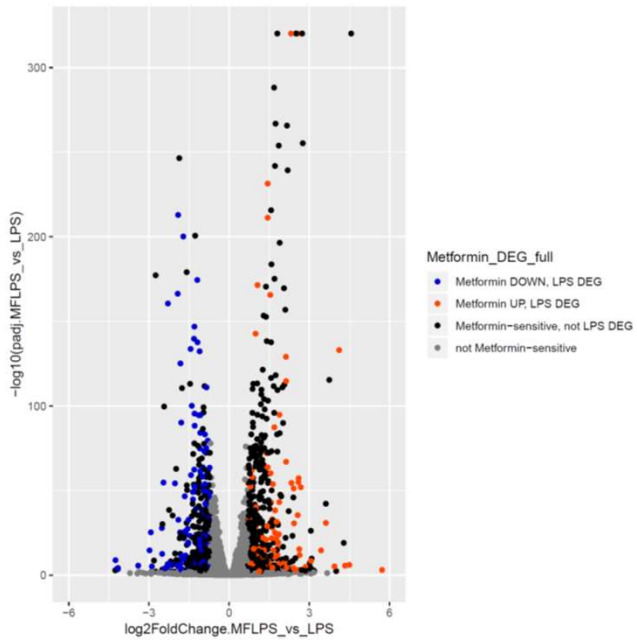
Principal-component analysis



Heatmaps



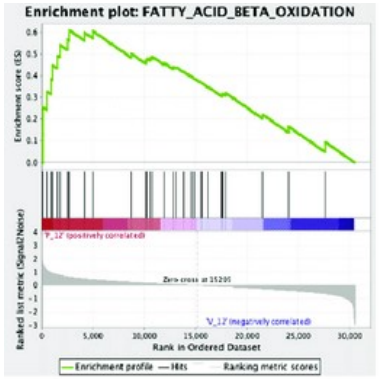
Volcano plots

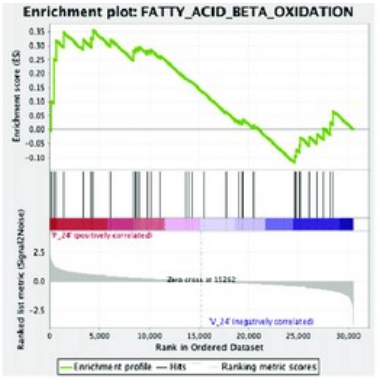


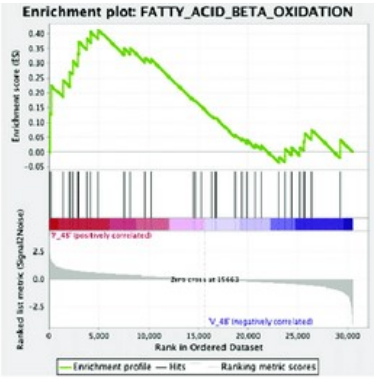
- Can all be done in R(Studio).

Step 8: Analyze further

- Gene Set Enrichment Analysis (https://www.gsea-msigdb.org/gsea/index.jsp) and Gene Ontology (https://biit.cs.ut.ee/gprofiler/gost):
  Search for pathways and functions enriched for DEGs.







- HOMER (http://homer.ucsd.edu/homer/motif/):
  Searches for DNA motifs enriched in the promoters of DEGs.

Rank	Motif	Name	P-value	log P-pvalue	q-value (Benjamini)	# Target Sequences with Motif	% of Targets Sequences with Motif	# Background Sequences with Motif	% of Background Sequences with Motif
1		IRF1(IRF)/PBMC-IRF1-ChIP-Seq(GSE43036)/Homer	1e-7	-1.625e+01	0.0000	21.0	7.29%	521.9	1.77%
2		ISRE(IRF)/ThioMac-LPS-Expression(GSE23622)/Homer	1e-6	-1.519e+01	0.0001	14.0	4.86%	244.4	0.83%
3		IRF3(IRF)/BMDM-Irf3-ChIP-Seq(GSE67343)/Homer	1e-6	-1.471e+01	0.0001	29.0	10.07%	1017.5	3.44%
		NFkB-p65(RHD)/GM12787-p65-ChIP-							

Step 9: Get help.

- If you get stuck, which you almost certainly will, you're probably not the first one who has encountered this problem. Use a search engine like DuckDuckGo or directly look at these places for help:
    - Stackoverflow:
        - https://stackoverflow.com/questions
        - Geared towards programming in general.
    - biostarS:
        - https://www.biostars.org/
        - Specializes in questions about bioinformatics.
    - R for Dummies:
        - https://rfordummies.com/
        - Swallow your pride. It's worth it.

- Affordable, local training courses:
    - New York Academy of Sciences: Introductory Coding for Researchers (1 weekend).
    - New York Genome Center: Sequencing Informatics Workshop (1 week).

A shameless plug

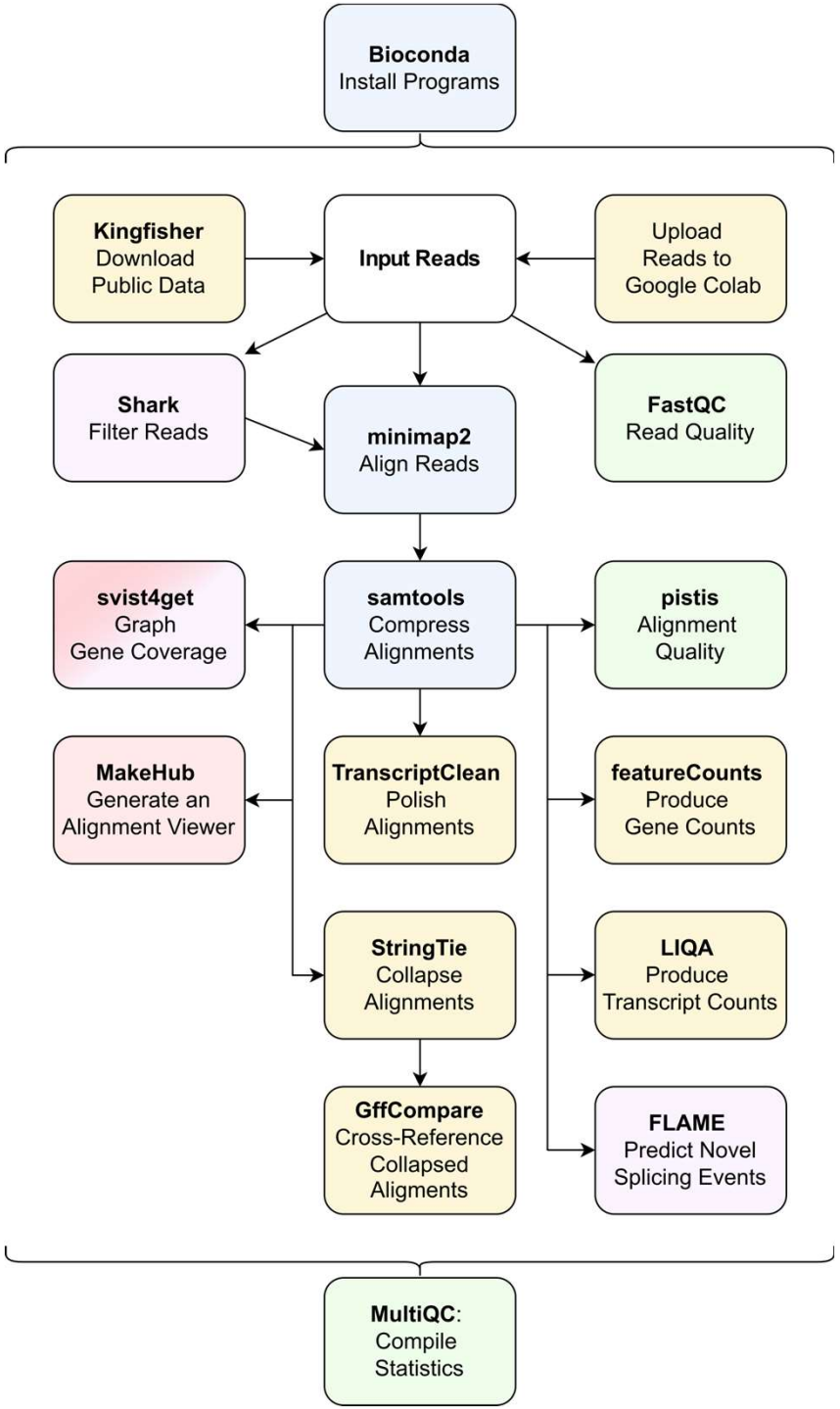
L-RAPiT: A Cloud-Computing Pipeline for the Analysis of Long-Read RNA Sequencing Data

Theodore M. Nelson, Sankar Ghosh and Thomas S. Postler*

Department of Microbiology & Immunology, Vagelos College of Physicians & Surgeons, Columbia University Irving Medical Center, New York, NY 10032, USA

ABSTRACT

Long-read sequencing (LRS) has been adopted to meet a wide variety of research needs, ranging from the construction of novel transcriptome annotations to the rapid identification of emerging virus variants. Amongst other advantages, LRS preserves more information about RNA at the transcript level than conventional high-throughput sequencing, including far more accurate and quantitative records of splicing patterns. New studies with LRS datasets are being published at an exponential rate, generating a vast reservoir of information that can be leveraged to address a host of different research questions. However, mining such publicly available data in a tailored fashion is currently not trivial, as the available software tools typically require familiarity with the command-line interface, which constitutes a significant obstacle to many researchers. Additionally, different research groups utilize different software packages to perform LRS analysis, which often prevents a direct comparison of published results across different studies. To address these challenges, we have developed the Long-Read Analysis Pipeline for Transcriptomics (L-RAPiT), a user-friendly, free pipeline requiring no dedicated computational resources or bioinformatics expertise (https://github.com/Theo-Nelson/long-read-sequencing-pipeline). L-RAPiT can be implemented directly through Google Colaboratory, a system based on the open-source Jupyter notebook environment, and allows for the direct analysis of transcriptomic reads from Oxford Nanopore and PacBio LRS machines. This new pipeline enables the rapid, convenient and standardized analysis of publicly available or newly generated LRS datasets.



Practice Run

- **Demonstration with practice data set:**
  - Computationally generated reads
    → trimming not necessary.
  - Limited to chromosome X of *Drosophila melanogaster* → can be run on standard laptop.