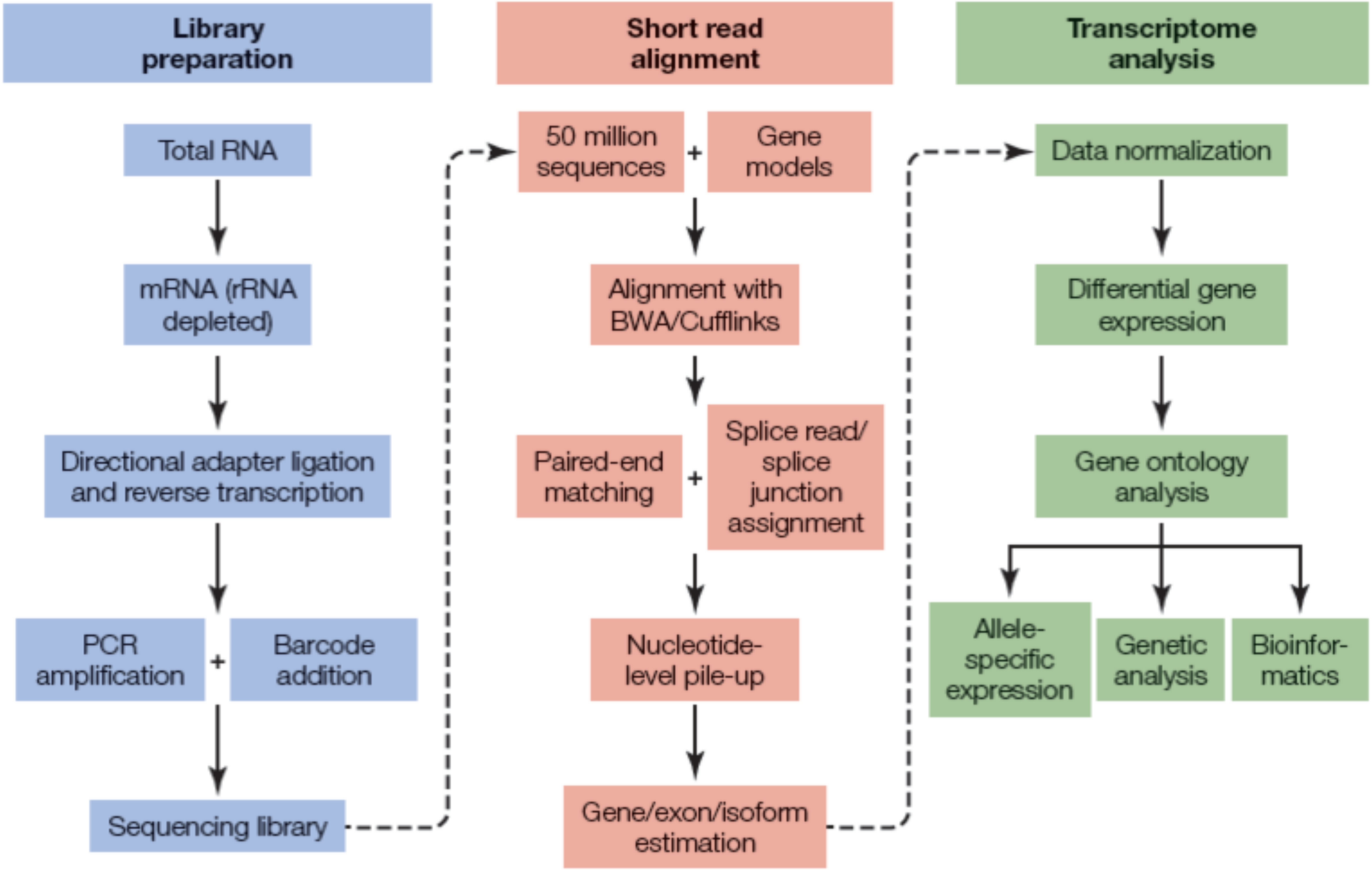
Introduction to RNAseq

Experimental design and data analysis

Vitor Pavinato

RNAseq in summary



Library preparation

Total RNA

mRNA (rRNA depleted)

Directional adapter ligation and reverse transcription

PCR amplification + Barcode addition

Sequencing library

Short read alignment

50 million sequences + Gene models

Alignment with BWA/Cufflinks

Paired-end matching + Splice read/ splice junction assignment

Nucleotide-level pile-up

Gene/exon/isoform estimation

Transcriptome analysis

Data normalization

Differential gene expression

Gene ontology analysis

Allele-specific expression

Genetic analysis

Bioinfor-matics

Pieces you’ll need to put together

Experimental Design

Biological vs Technical Replicates

How many (sample size)?

Desired sequencing yield (coverage)

How many lanes?

Hoe many samples / lane?

Single-end vs Paired-end sequencing

Sequence length (75, 100, 300 bp)?

Experimental design

Technical vs biological replicates

Biological replicates are necessary because there is a large variance between samples in the same condition.

Technical replicates are necessary if there were factors during sample and library preparation and sequencing that may increase the variance among replicates

Examples of biological and technical replicates

Often you will have a fixed budget that constrains how many arrays can be processed. So your first task is to determine what levels of replication you can afford, and how they will impact statistical power.

Technical Replication:

- RNA preparation (eg. from adjacent biopsies)
- cDNA synthesis (pooling minimizes outlier effects)
- library preparation
- sequencing lane or array hybridization (usually a minimal effect)

Biological Replication:

Fixed effects:

- sex
- treatment (drug, growth regimen, tissue)
- time of sampling (repeated measures in some cases)
- genotype (IF specifically chosen and resampled)

Random effects - individual from a population

- field plot

Contrast of interest

At the same time, you need to be aware of the contrasts you wish to make since by tweaking the design you may gain a lot in terms of what you can infer.

Suppose you want to compare B cells and T cells from Healthy controls and COVID-19 patients, and you have the funds to generate 24 RNASeq profiles

What is the best design?

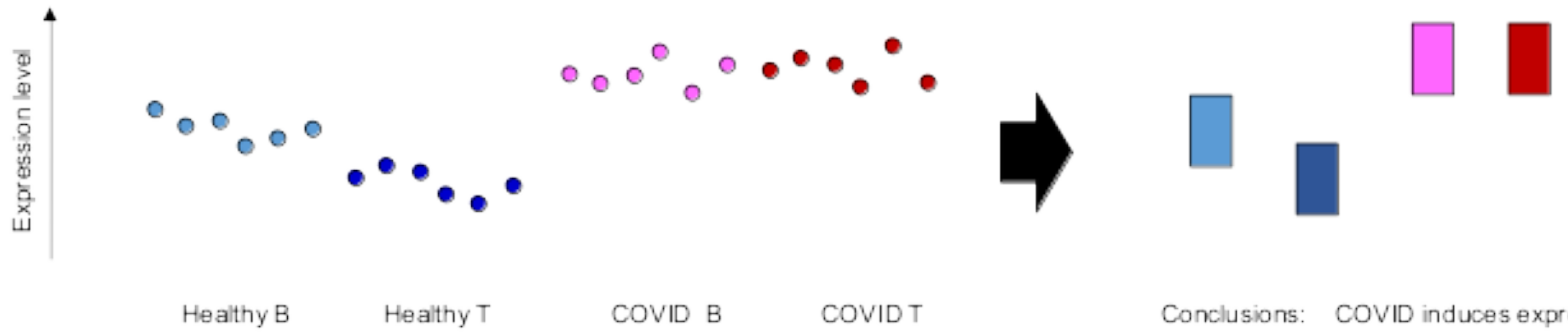
- 6 controls and 6 patients, each donating both a B and a T cell sample
- 12 controls and 12 patients, each donating either a B or a T cell sample
- 3 controls and 3 patients, each donating a B and a T cell sample, processed twice
- 3 controls and 3 patients, each donating 2 B and 2 T cell samples, on separate days
- same as above, but only men or only women
- 12 controls and 12 patients, each donating either a B or a T cell sample, but pooling two visits

Main effects can only be contrasted if you have biological replicates:

reducing the number of individuals may allow you to address intra-individual variability

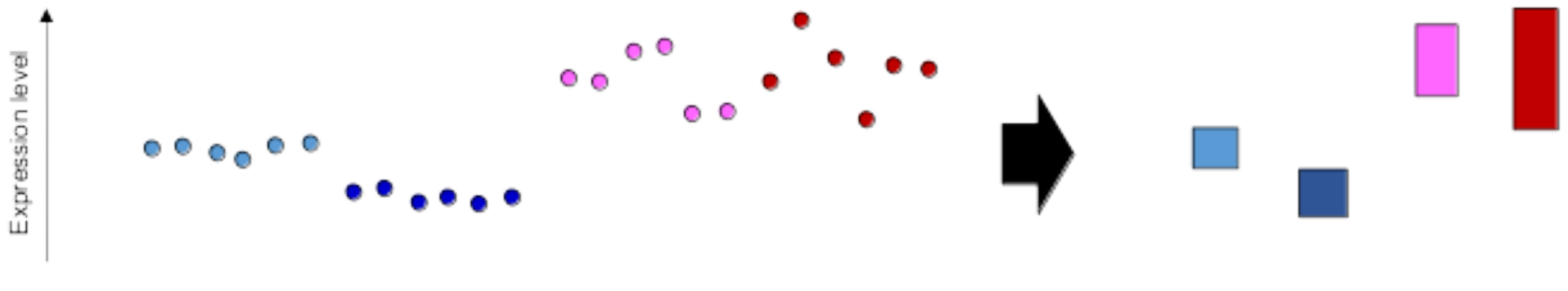
Interaction effects allow you to ask questions like whether B cells and T cells differ more between healthy volunteers or patients

Two Hypothetical Sets of Results Illustrating Design Principles



Conclusions: COVID induces expression

T < B only in healthy people



Additional
Conclusions: Variability is low in Healthy controls

Flu B cell response is individualized

Flu T cell response is hypervariable

https://cqs-vumc.shinyapps.io/rnaseqsamplesizeweb/

Pieces you’ll need to put together

How to process your data

Is there a reference genome that I can use?

If not, what can I do?

If I have a reference genome and transcriptome,
Which approach is the best?

DEG analysis

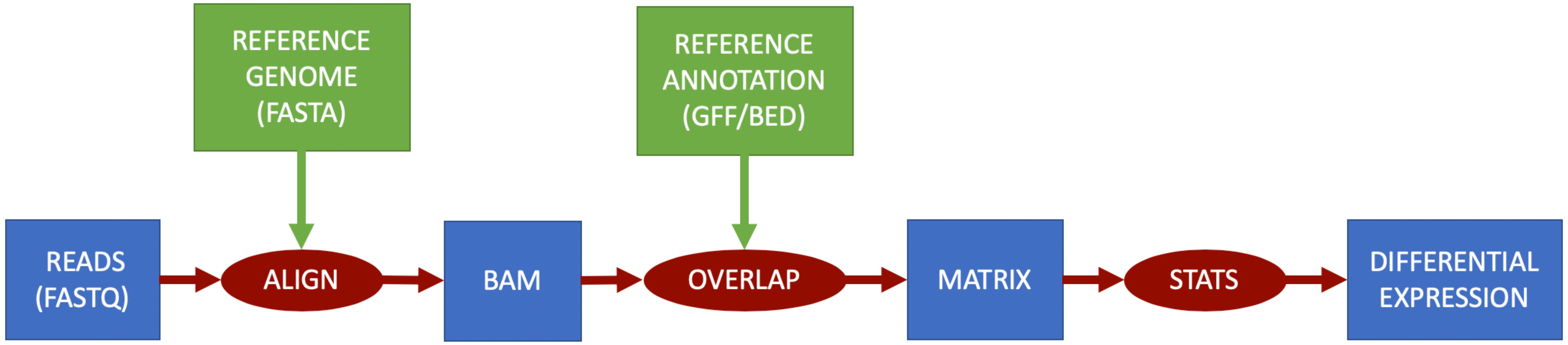
Features:
Genes
Exons
Transcripts

name	control	shock	fold_change	pvalue
Gene A	100	200	2.0	0.000035
Gene B	80	60	0.75	0.234
Gene C	120	180	1.5	0.013
...				

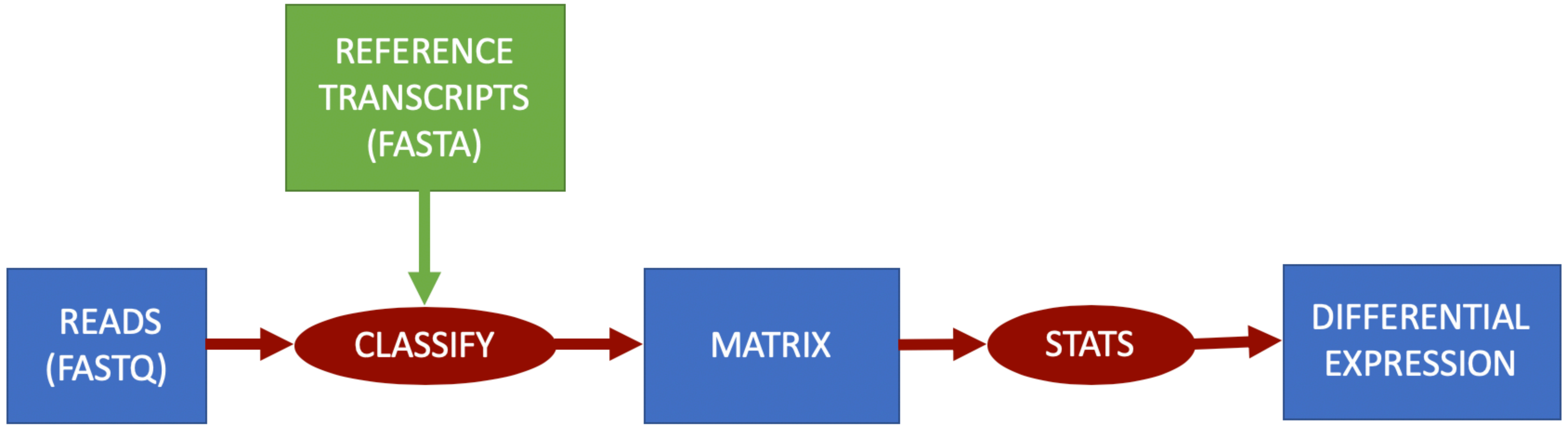
How do I get this?

Methods to quantify expression

Reference-based



Classification-based



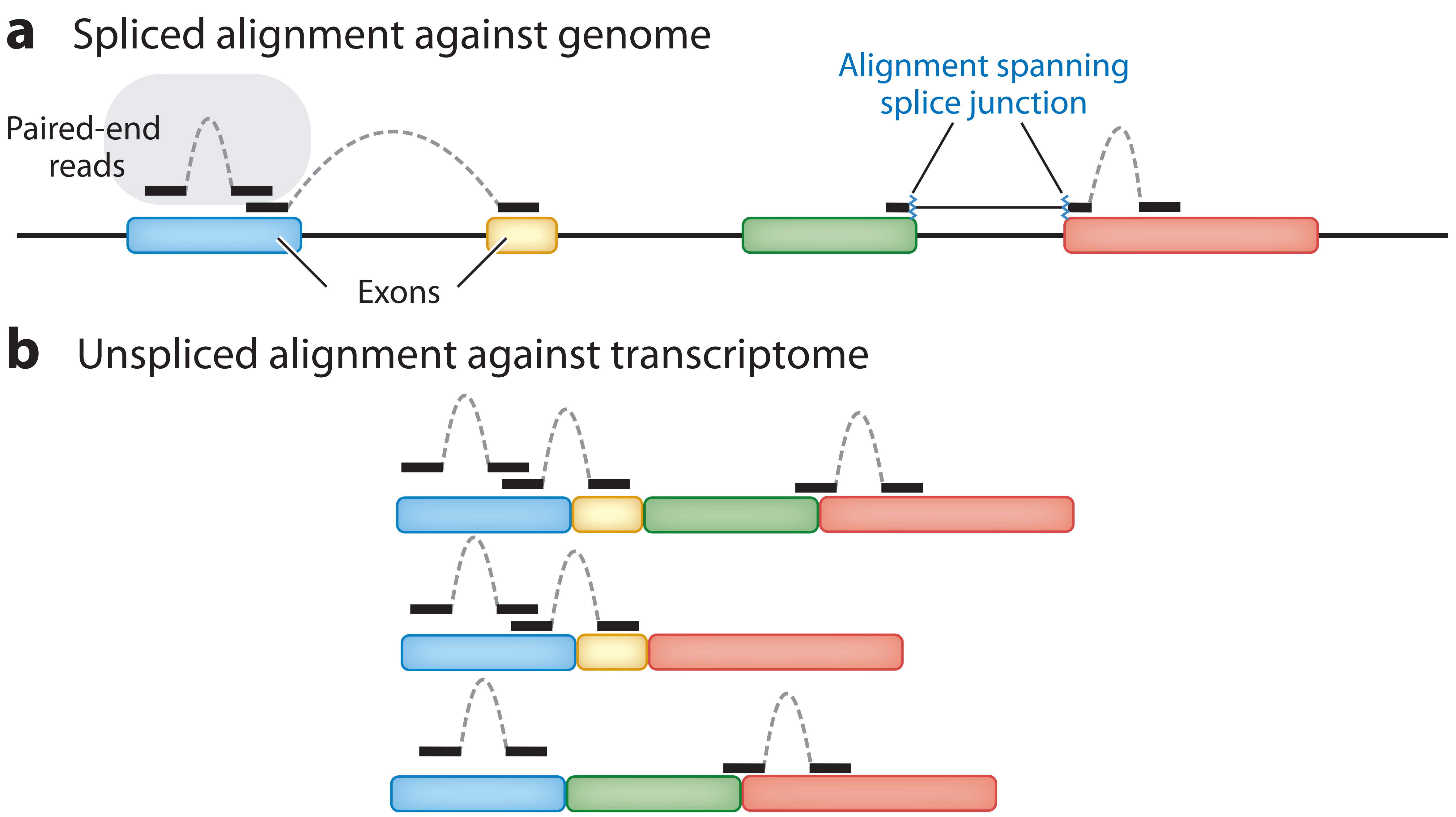
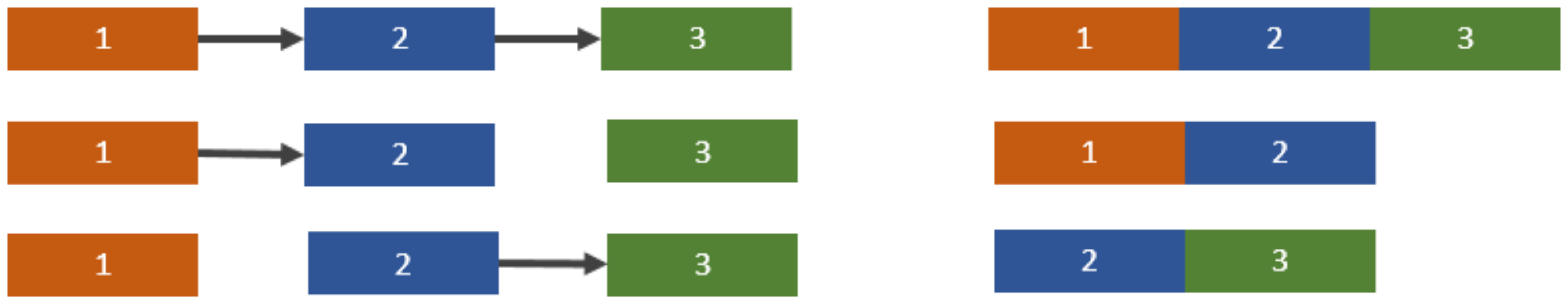

a
Spliced alignment against genome
Paired-end reads
Alignment spanning splice junction
Exons
b
Unspliced alignment against transcriptome

Figure 4

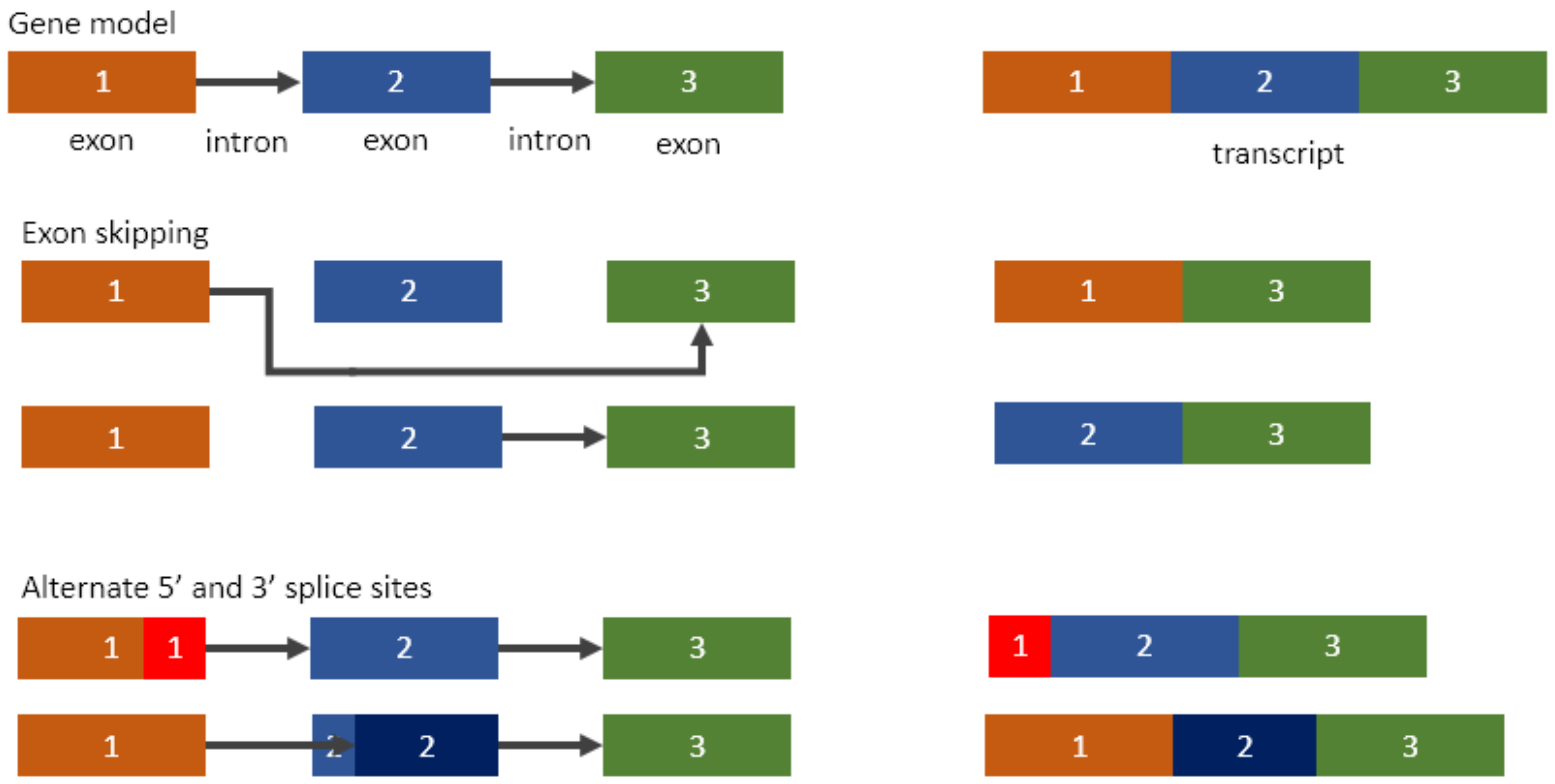
An illustration of spliced alignment of RNA sequencing (RNA-seq) fragments to a genome (*a*) and direct alignment to a transcriptome (*b*). Reads are designated by thick solid lines, while dashed arcs represent the pairing relationship between paired-end reads. This illustration depicts alignment to a single four-exon gene consisting of three distinct transcripts. In the spliced alignment (*a*), the left read of the rightmost pair is a junction-spanning alignment to the red–green exon boundary. In the direct alignment to the transcriptome (*b*), one observes how the same alignment (e.g., the alignment to the blue exon) is repeated for each transcript.

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

What are gene isoforms

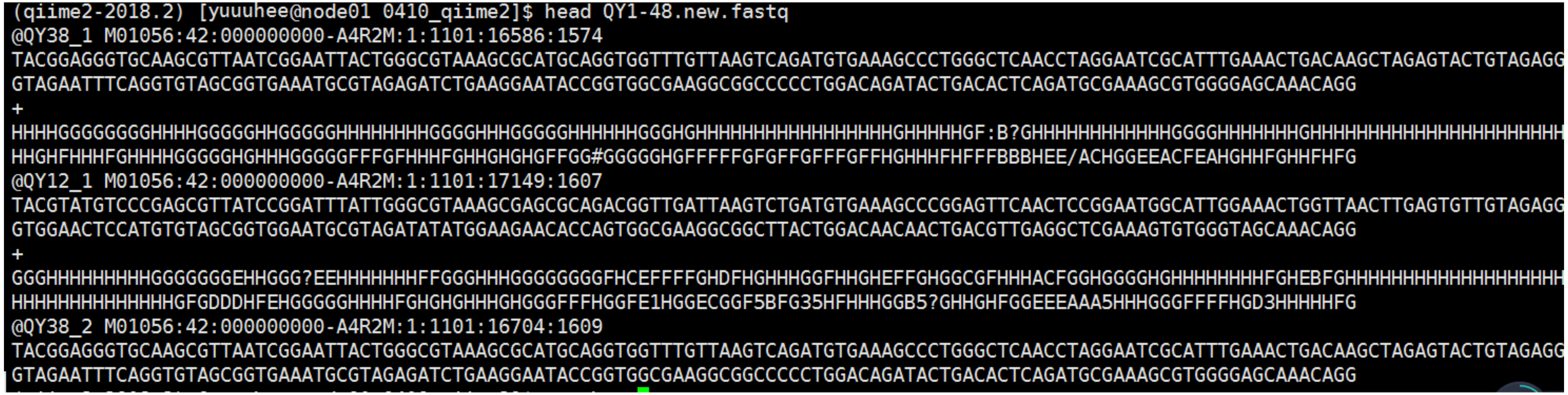


Alternative splicing

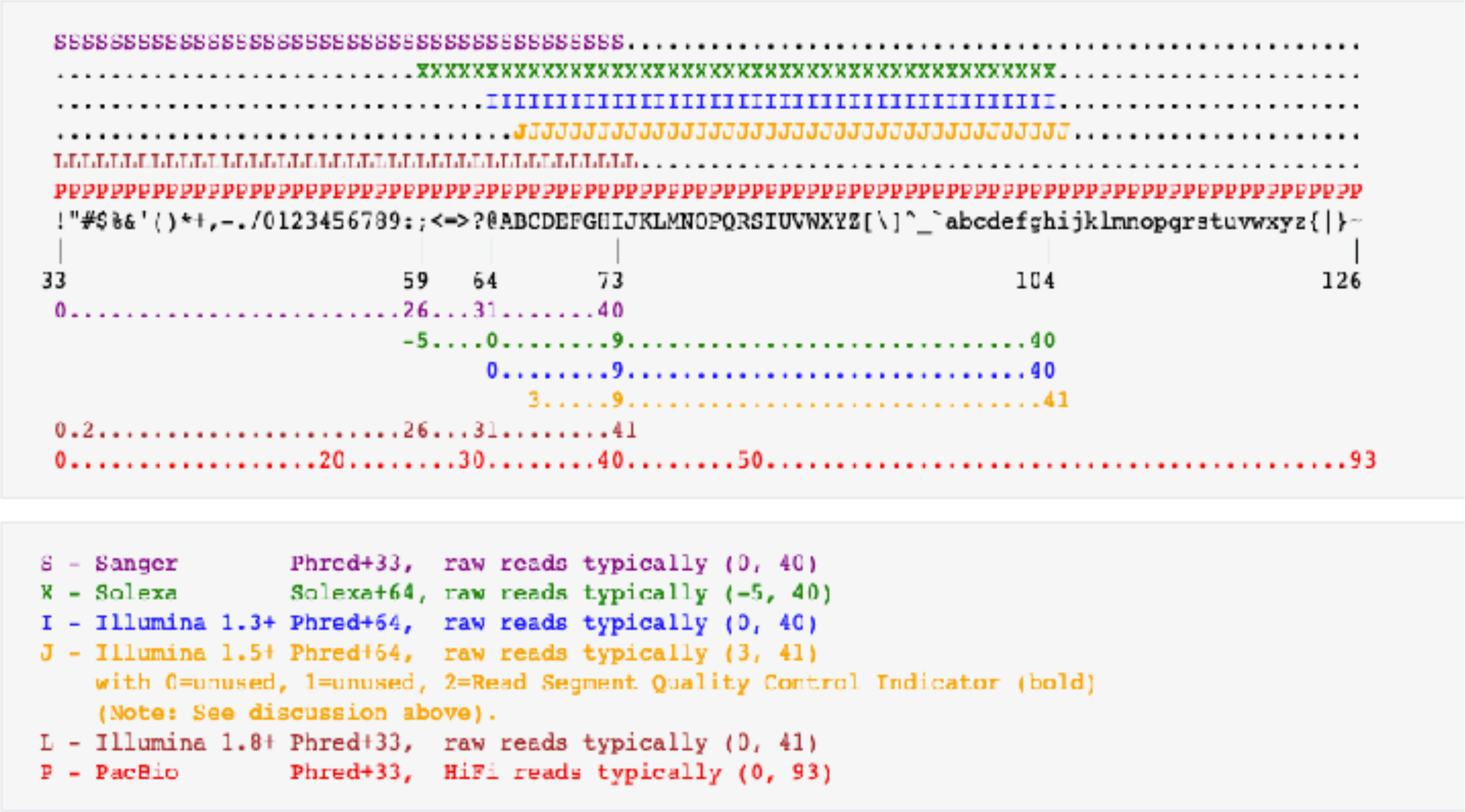


How the data looks like

Example FASTQ file



```
(qiime2-2018.2) [yuuuhee@node01 0410_qiime2]$ head QY1-48.new.fastq
@QY38_1 M01056:42:000000000-A4R2M:1:1101:16586:1574
TACGGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGCATGCAGGTGGTTTGTTAAGTCAGATGTGAAAGCCCTGGGCTCAACCTAGGAATCGCATTTGAAACTGACAAGCTAGAGTACTGTAGAGG
GTAGAATTTCAGGTGTAGCGGTGAAATGCGTAGAGATCTGAAGGAATACCGGTGGCGAAGGCGGCCCCCTGGACAGATACTGACACTCAGATGCGAAAGCGTGGGGAGCAAACAGG
+
HHHHGGGGGGGGHHHHGGGGGHHGGGGGHHHHHHHHGGGGHHHGGGGGHHHHHHGGGHGHHHHHHHHHHHHHHHHHHGHHHHHGF:B?GHHHHHHHHHHHHGGGGHHHHHHHGHHHHHHHHHHHHHHHHHHHHHH
HHGHFHHHFGHHHHGGGGGHGHHHGGGGGFFFGFHHHFGHHGHGHGFFGG#GGGGGHGFFFFFGFGFFGFFFGFFHGHHHFHFFFBBBHEE/ACHGGEEACFEAHGHHFGHHFHFG
@QY12_1 M01056:42:000000000-A4R2M:1:1101:17149:1607
TACGTATGTCCCGAGCGTTATCCGGATTTATTGGGCGTAAAGCGAGCGCAGACGGTTGATTAAGTCTGATGTGAAAGCCCGGAGTTCAACTCCGGAATGGCATTGGAAACTGGTTAACTTGAGTGTTGTAGAGG
GTGGAACTCCATGTGTAGCGGTGGAATGCGTAGATATATGGAAGAACACCAGTGGCGAAGGCGGCTTACTGGACAACAACTGACGTTGAGGCTCGAAAGTGTGGGTAGCAAACAGG
+
GGGHHHHHHHHHHGGGGGGGEHHGGG?EEHHHHHHHHFFGGGHHHGGGGGGGGFHCEFFFFGHDFHGHHHGGFHHGHEFFGHGGCGFHHHACFGGHGGGGHGHHHHHHHHFGHEBFGHHHHHHHHHHHHHHHHHHH
HHHHHHHHHHHHHHHGFGDDDHFEHGGGGGHHHHFGHGHGHHHGHGGGFFFHGGFE1HGGECGGF5BFG35HFHHHGGB5?GHHGHFGGEEEAAA5HHHGGGFFFFHGD3HHHHHFG
@QY38_2 M01056:42:000000000-A4R2M:1:1101:16704:1609
TACGGAGGGTGCAAGCGTTAATCGGAATTACTGGGCGTAAAGCGCATGCAGGTGGTTTGTTAAGTCAGATGTGAAAGCCCTGGGCTCAACCTAGGAATCGCATTTGAAACTGACAAGCTAGAGTACTGTAGAGG
GTAGAATTTCAGGTGTAGCGGTGAAATGCGTAGAGATCTGAAGGAATACCGGTGGCGAAGGCGGCCCCCTGGACAGATACTGACACTCAGATGCGAAAGCGTGGGGAGCAAACAGG
```

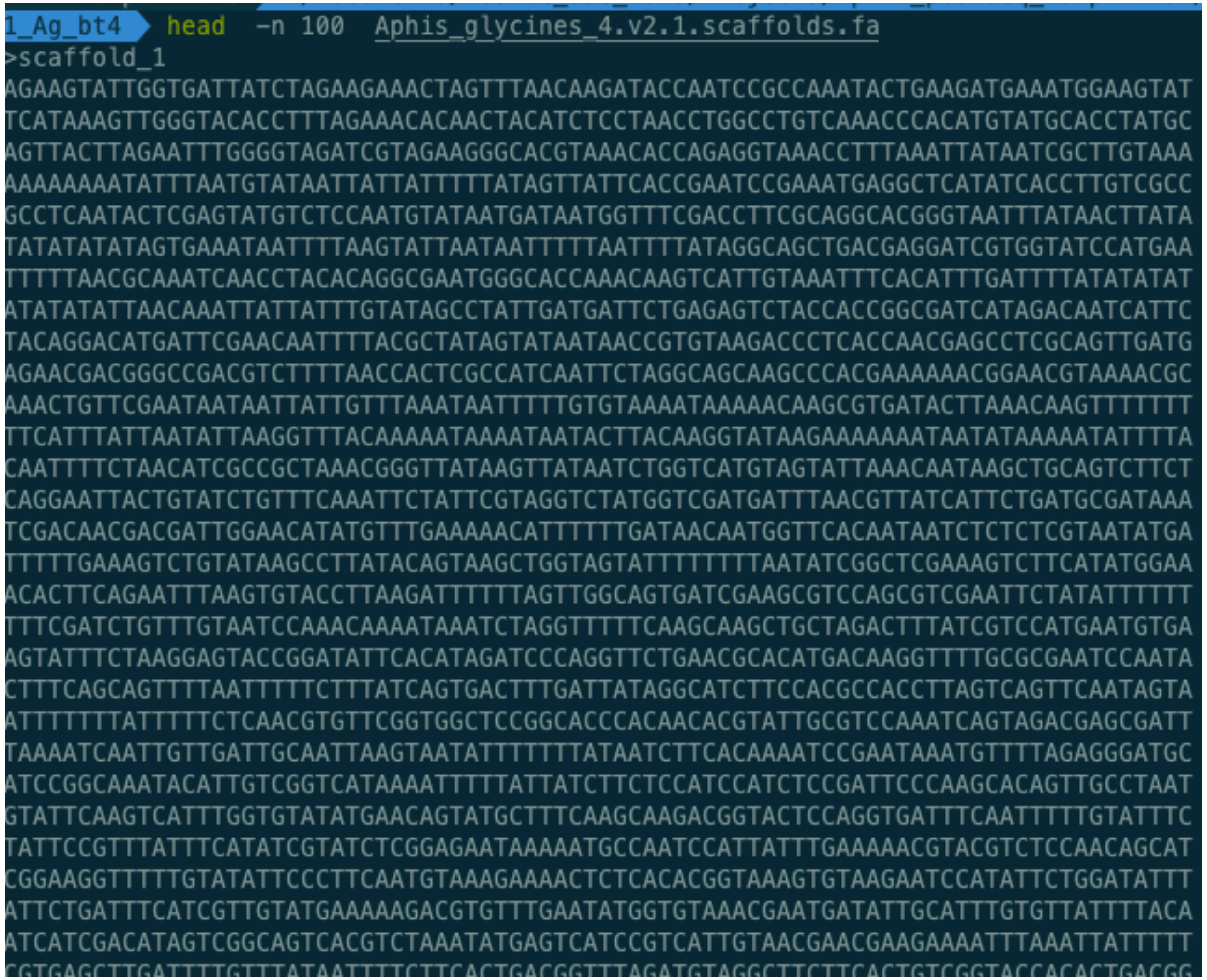


FASTQ scoring

```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
 ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
 .................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
 PPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPPP
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126
 0........................26...31.......40
                          -5....0........9.............................40
                                0........9.............................40
                                   3.....9..............................41
 0.2......................26...31........41
 0..................20........30........40........50..........................................93
```

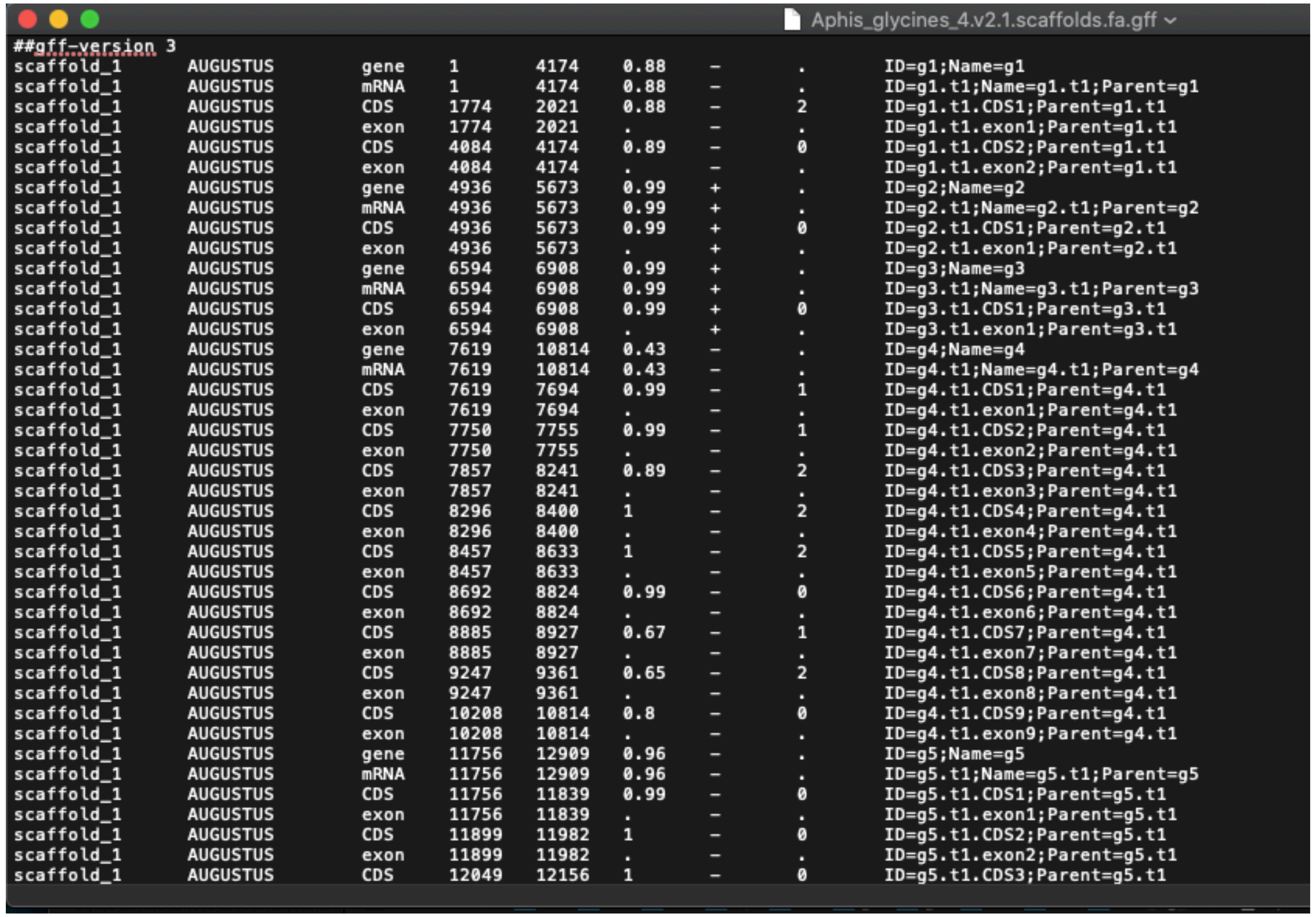
```
S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 41)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
P - PacBio        Phred+33,  HiFi reads typically (0, 93)
```

Example FASTA file (Genome)



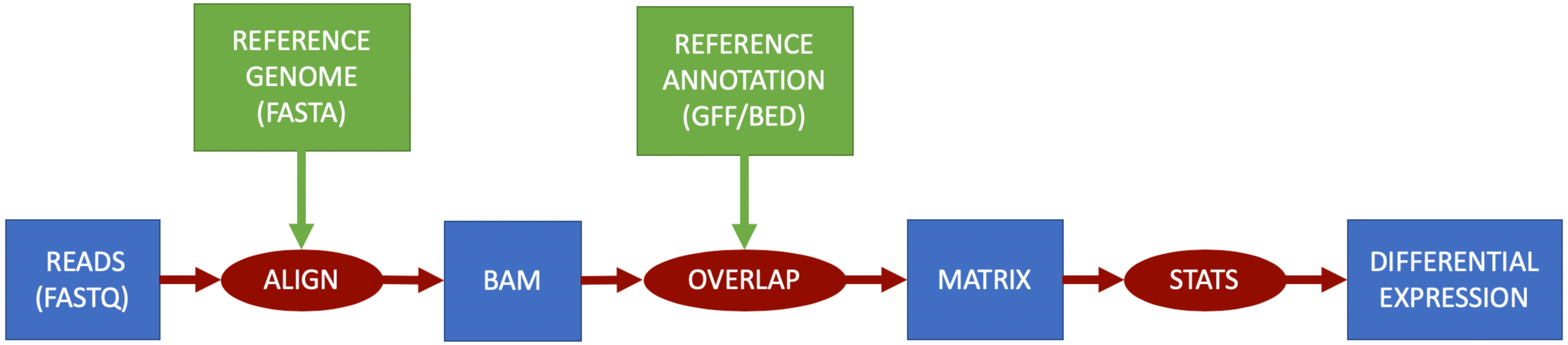
```
1_Ag_bt4  head  -n 100  Aphis_glycines_4.v2.1.scaffolds.fa
>scaffold_1
AGAAGTATTGGTGATTATCTAGAAGAAACTAGTTTAACAAGATACCAATCCGCCAAATACTGAAGATGAAATGGAAGTAT
TCATAAAGTTGGGTACACCTTTAGAAACACAACTACATCTCCTAACCTGGCCTGTCAAACCCACATGTATGCACCTATGC
AGTTACTTAGAATTTGGGGTAGATCGTAGAAGGGCACGTAAACACCAGAGGTAAACCTTTAAATTATAATCGCTTGTAAA
AAAAAAAATATTTAATGTATAATTATTATTTTTATAGTTATTCACCGAATCCGAAATGAGGCTCATATCACCTTGTCGCC
GCCTCAATACTCGAGTATGTCTCCAATGTATAATGATAATGGTTTCGACCTTCGCAGGCACGGGTAATTTATAACTTATA
TATATATATAGTGAAATAATTTTAAGTATTAATAATTTTTAATTTTATAGGCAGCTGACGAGGATCGTGGTATCCATGAA
TTTTTAACGCAAATCAACCTACACAGGCGAATGGGCACCAAACAAGTCATTGTAAATTTCACATTTGATTTTATATATAT
ATATATATTAACAAATTATTATTTGTATAGCCTATTGATGATTCTGAGAGTCTACCACCGGCGATCATAGACAATCATTC
TACAGGACATGATTCGAACAATTTTACGCTATAGTATAATAACCGTGTAAGACCCTCACCAACGAGCCTCGCAGTTGATG
AGAACGACGGGCCGACGTCTTTTAACCACTCGCCATCAATTCTAGGCAGCAAGCCCACGAAAAAACGGAACGTAAAACGC
AAACTGTTCGAATAATAATTATTGTTTAAATAATTTTTGTGTAAAATAAAAACAAGCGTGATACTTAAACAAGTTTTTTT
TTCATTTATTAATATTAAGGTTTACAAAAATAAAATAATACTTACAAGGTATAAGAAAAAAATAATATAAAAATATTTTA
CAATTTTCTAACATCGCCGCTAAACGGGTTATAAGTTATAATCTGGTCATGTAGTATTAAACAATAAGCTGCAGTCTTCT
CAGGAATTACTGTATCTGTTTCAAATTCTATTCGTAGGTCTATGGTCGATGATTTAACGTTATCATTCTGATGCGATAAA
TCGACAACGACGATTGGAACATATGTTTGAAAAACATTTTTTGATAACAATGGTTCACAATAATCTCTCTCGTAATATGA
TTTTTGAAAGTCTGTATAAGCCTTATACAGTAAGCTGGTAGTATTTTTTTTAATATCGGCTCGAAAGTCTTCATATGGAA
ACACTTCAGAATTTAAGTGTACCTTAAGATTTTTTTAGTTGGCAGTGATCGAAGCGTCCAGCGTCGAATTCTATATTTTTT
TTTCGATCTGTTTGTAATCCAAACAAAATAAATCTAGGTTTTTCAAGCAAGCTGCTAGACTTTATCGTCCATGAATGTGA
AGTATTTCTAAGGAGTACCGGATATTCACATAGATCCCAGGTTCTGAACGCACATGACAAGGTTTTGCGCGAATCCAATA
CTTTCAGCAGTTTTAATTTTTCTTTATCAGTGACTTTGATTATAGGCATCTTCCACGCCACCTTAGTCAGTTCAATAGTA
ATTTTTTTATTTTTCTCAACGTGTTCGGTGGCTCCGGCACCCACAACACGTATTGCGTCCAAATCAGTAGACGAGCGATT
TAAAATCAATTGTTGATTGCAATTAAGTAATATTTTTTTATAATCTTCACAAAATCCGAATAAATGTTTTAGAGGGATGC
ATCCGGCAAATACATTGTCGGTCATAAAATTTTTATTATCTTCTCCATCCATCTCCGATTCCCAAGCACAGTTGCCTAAT
GTATTCAAGTCATTTGGTGTATATGAACAGTATGCTTTCAAGCAAGACGGTACTCCAGGTGATTTCAATTTTTGTATTTC
TATTCCGTTTATTTCATATCGTATCTCGGAGAATAAAAATGCCAATCCATTATTTGAAAAACGTACGTCTCCAACAGCAT
CGGAAGGTTTTTGTATATTCCCTTCAATGTAAAGAAAACTCTCACACGGTAAAGTGTAAGAATCCATATTCTGGATATTT
ATTCTGATTTCATCGTTGTATGAAAAAGACGTGTTTGAATATGGTGTAAACGAATGATATTGCATTTGTGTTATTTTACA
ATCATCGACATAGTCGGCAGTCACGTCTAAATATGAGTCATCCGTCATTGTAACGAACGAAGAAAATTTAAATTATTTTT
CGTGAGCTTGATTTTGTTTATAATTTTCTTCACTGACGGTTTAGATGTAGGCTTCTTCACTGTCGGTACCACACTGACGG
```

Example GFF file (Annotation)



```
##gff-version 3
scaffold_1    AUGUSTUS    gene  1      4174   0.88  -  .  ID=g1;Name=g1
scaffold_1    AUGUSTUS    mRNA  1      4174   0.88  -  .  ID=g1.t1;Name=g1.t1;Parent=g1
scaffold_1    AUGUSTUS    CDS   1774   2021   0.88  -  2  ID=g1.t1.CDS1;Parent=g1.t1
scaffold_1    AUGUSTUS    exon  1774   2021   .     -  .  ID=g1.t1.exon1;Parent=g1.t1
scaffold_1    AUGUSTUS    CDS   4084   4174   0.89  -  0  ID=g1.t1.CDS2;Parent=g1.t1
scaffold_1    AUGUSTUS    exon  4084   4174   .     -  .  ID=g1.t1.exon2;Parent=g1.t1
scaffold_1    AUGUSTUS    gene  4936   5673   0.99  +  .  ID=g2;Name=g2
scaffold_1    AUGUSTUS    mRNA  4936   5673   0.99  +  .  ID=g2.t1;Name=g2.t1;Parent=g2
scaffold_1    AUGUSTUS    CDS   4936   5673   0.99  +  0  ID=g2.t1.CDS1;Parent=g2.t1
scaffold_1    AUGUSTUS    exon  4936   5673   .     +  .  ID=g2.t1.exon1;Parent=g2.t1
scaffold_1    AUGUSTUS    gene  6594   6908   0.99  +  .  ID=g3;Name=g3
scaffold_1    AUGUSTUS    mRNA  6594   6908   0.99  +  .  ID=g3.t1;Name=g3.t1;Parent=g3
scaffold_1    AUGUSTUS    CDS   6594   6908   0.99  +  0  ID=g3.t1.CDS1;Parent=g3.t1
scaffold_1    AUGUSTUS    exon  6594   6908   .     +  .  ID=g3.t1.exon1;Parent=g3.t1
scaffold_1    AUGUSTUS    gene  7619   10814  0.43  -  .  ID=g4;Name=g4
scaffold_1    AUGUSTUS    mRNA  7619   10814  0.43  -  .  ID=g4.t1;Name=g4.t1;Parent=g4
scaffold_1    AUGUSTUS    CDS   7619   7694   0.99  -  1  ID=g4.t1.CDS1;Parent=g4.t1
scaffold_1    AUGUSTUS    exon  7619   7694   .     -  .  ID=g4.t1.exon1;Parent=g4.t1
scaffold_1    AUGUSTUS    CDS   7750   7755   0.99  -  1  ID=g4.t1.CDS2;Parent=g4.t1
scaffold_1    AUGUSTUS    exon  7750   7755   .     -  .  ID=g4.t1.exon2;Parent=g4.t1
scaffold_1    AUGUSTUS    CDS   7857   8241   0.89  -  2  ID=g4.t1.CDS3;Parent=g4.t1
scaffold_1    AUGUSTUS    exon  7857   8241   .     -  .  ID=g4.t1.exon3;Parent=g4.t1
scaffold_1    AUGUSTUS    CDS   8296   8400   1     -  2  ID=g4.t1.CDS4;Parent=g4.t1
scaffold_1    AUGUSTUS    exon  8296   8400   .     -  .  ID=g4.t1.exon4;Parent=g4.t1
scaffold_1    AUGUSTUS    CDS   8457   8633   1     -  2  ID=g4.t1.CDS5;Parent=g4.t1
scaffold_1    AUGUSTUS    exon  8457   8633   .     -  .  ID=g4.t1.exon5;Parent=g4.t1
scaffold_1    AUGUSTUS    CDS   8692   8824   0.99  -  0  ID=g4.t1.CDS6;Parent=g4.t1
scaffold_1    AUGUSTUS    exon  8692   8824   .     -  .  ID=g4.t1.exon6;Parent=g4.t1
scaffold_1    AUGUSTUS    CDS   8885   8927   0.67  -  1  ID=g4.t1.CDS7;Parent=g4.t1
scaffold_1    AUGUSTUS    exon  8885   8927   .     -  .  ID=g4.t1.exon7;Parent=g4.t1
scaffold_1    AUGUSTUS    CDS   9247   9361   0.65  -  2  ID=g4.t1.CDS8;Parent=g4.t1
scaffold_1    AUGUSTUS    exon  9247   9361   .     -  .  ID=g4.t1.exon8;Parent=g4.t1
scaffold_1    AUGUSTUS    CDS   10208  10814  0.8   -  0  ID=g4.t1.CDS9;Parent=g4.t1
scaffold_1    AUGUSTUS    exon  10208  10814  .     -  .  ID=g4.t1.exon9;Parent=g4.t1
scaffold_1    AUGUSTUS    gene  11756  12909  0.96  -  .  ID=g5;Name=g5
scaffold_1    AUGUSTUS    mRNA  11756  12909  0.96  -  .  ID=g5.t1;Name=g5.t1;Parent=g5
scaffold_1    AUGUSTUS    CDS   11756  11839  0.99  -  0  ID=g5.t1.CDS1;Parent=g5.t1
scaffold_1    AUGUSTUS    exon  11756  11839  .     -  .  ID=g5.t1.exon1;Parent=g5.t1
scaffold_1    AUGUSTUS    CDS   11899  11982  1     -  0  ID=g5.t1.CDS2;Parent=g5.t1
scaffold_1    AUGUSTUS    exon  11899  11982  .     -  .  ID=g5.t1.exon2;Parent=g5.t1
scaffold_1    AUGUSTUS    CDS   12049  12156  1     -  0  ID=g5.t1.CDS3;Parent=g5.t1
```

Reference-based



REFERENCE GENOME (FASTA)

REFERENCE ANNOTATION (GFF/BED)

READS (FASTQ)

ALIGN

BAM

OVERLAP

MATRIX

STATS

DIFFERENTIAL EXPRESSION

Hisat2

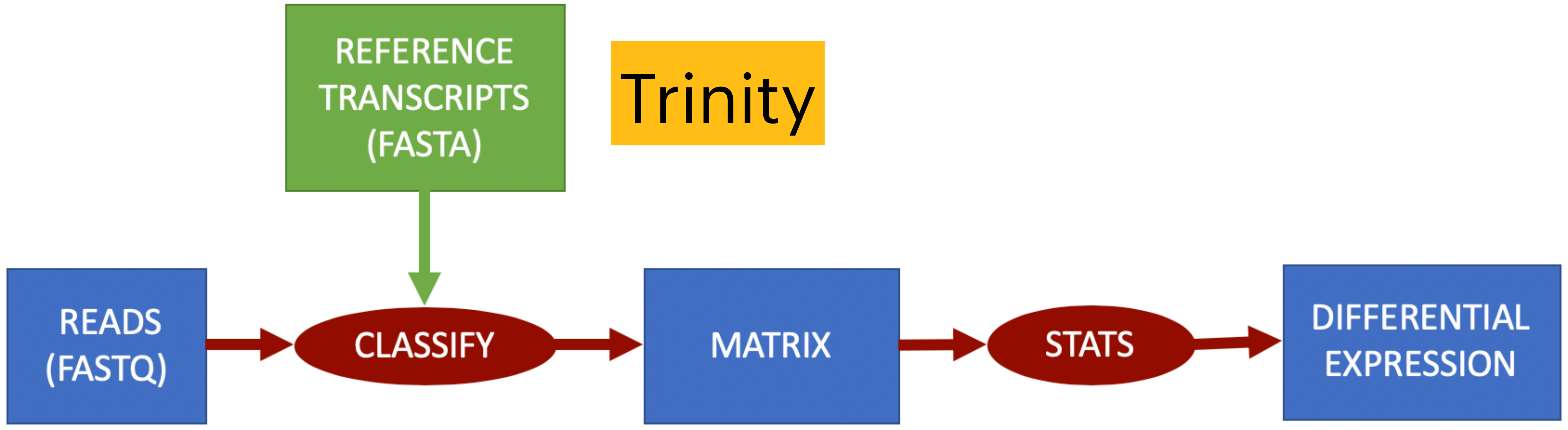
Splice-aware mapper

FeatureCounts

mRNA
cDNA
Gene

Exons?

Classification-based



Trinity

Kallisto

Features: Transcripts isoforms

Transcripts or genes

Features

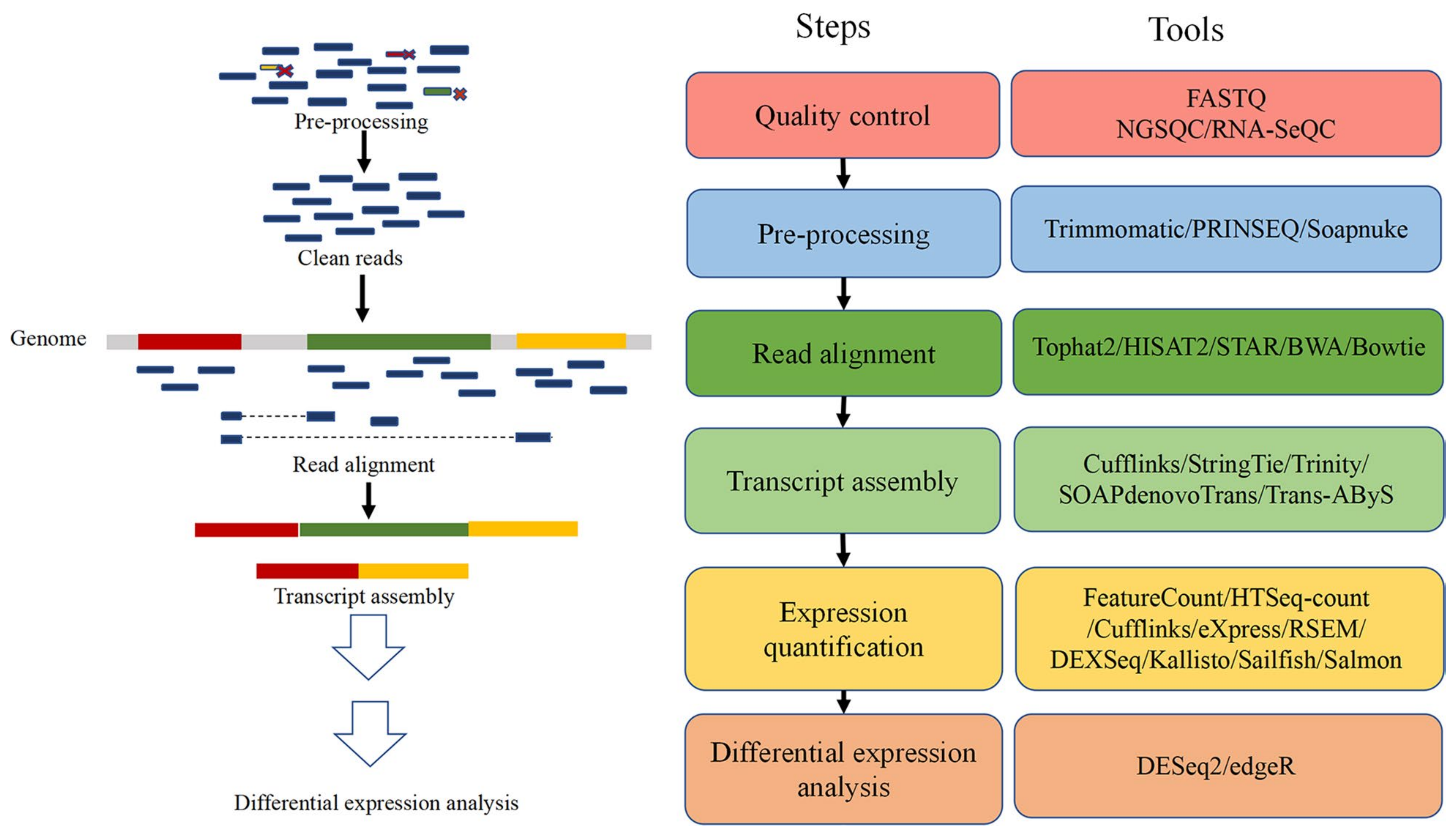
Reference-based

Genes (IDs)

Classification-based

Transcripts (isoforms)

Pipeline: from reads to candidate genes



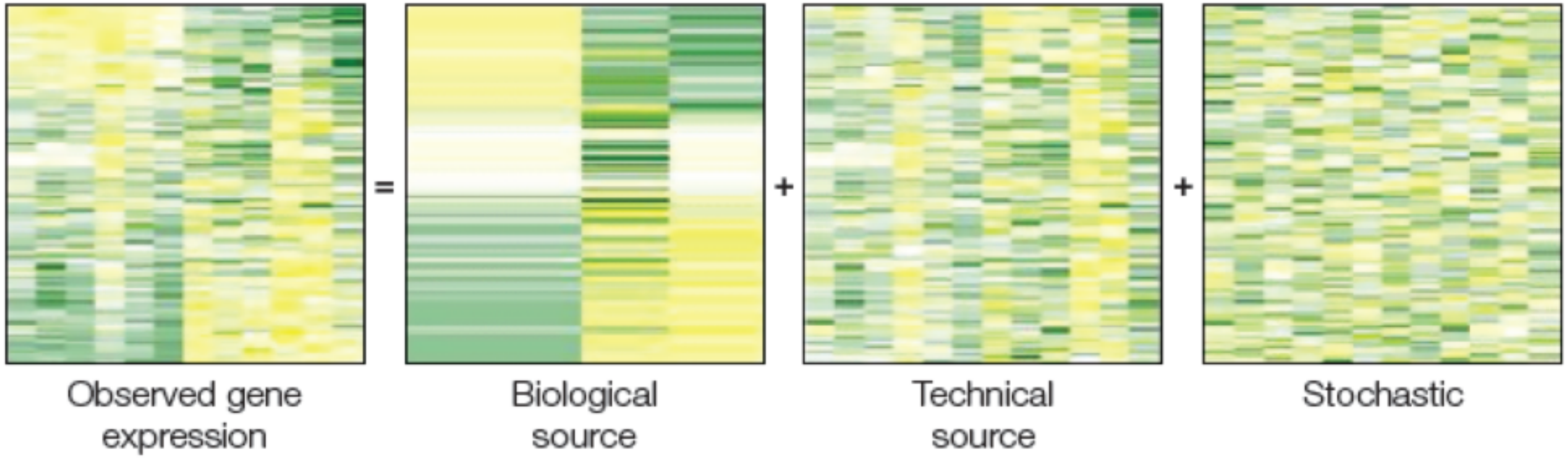
Steps	Tools
Quality control	FASTQ NGSQC/RNA-SeQC
Pre-processing	Trimmomatic/PRINSEQ/Soapnuke
Read alignment	Tophat2/HISAT2/STAR/BWA/Bowtie
Transcript assembly	Cufflinks/StringTie/Trinity/ SOAPdenovoTrans/Trans-AByS
Expression quantification	FeatureCount/HTSeq-count /Cufflinks/eXpress/RSEM/ DEXSeq/Kallisto/Sailfish/Salmon
Differential expression analysis	DESeq2/edgeR

Fig. 4 Bioinformatics tools commonly used in RNA-seq data analysis. These tools are primarily used in the four main processes of RNA-seq data analysis, including quality control, read alignment and transcript assembly, expression quantification and differential expression analysis

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

Pieces you’ll need to put together

Normalization



Approaches to Normalization

Mean or Median transform, simply centers the distribution

- Something like this is essential to control for overall distributional effects (eg RNA concentration)

Variance transforms, such as standardization or inter-quartile range

- Depends on whether you think the overall distributions should have similar variance

Quantile normalization

- Transforms the ranks to the average expression value for each rank

Gene-level model fitting

- Remove technical or biological effects before model fitting on the residuals

Supervised normalization

- Optimally estimate the biological effect while fitting technical factors across the entire experiment

DEG analysis

name	control	shock
Gene A	100	200
Gene B	80	60
Gene C	120	180
...		

How do we get this?

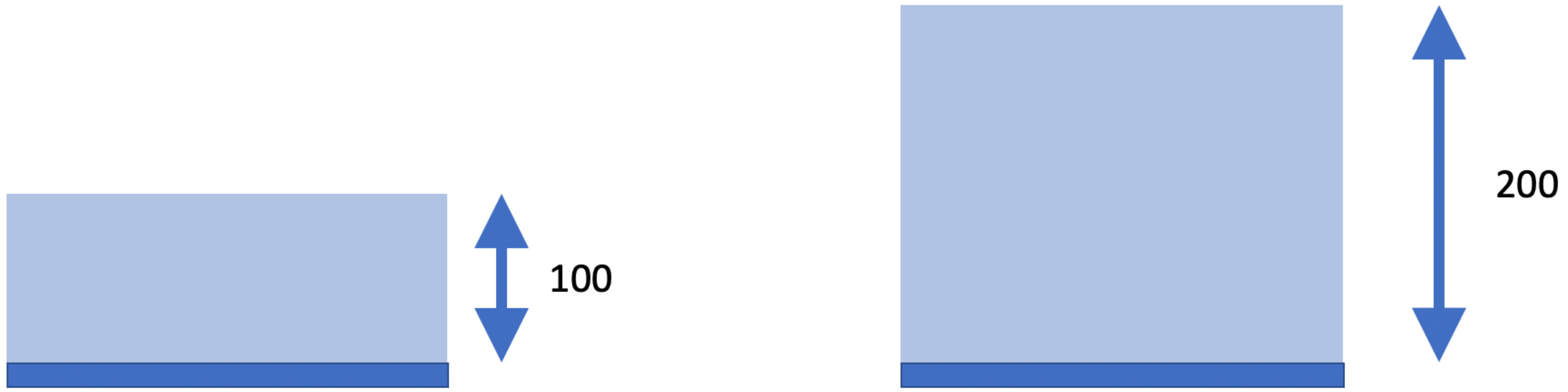
fold_change	pvalue
2.0	0.000035
0.75	0.234
1.5	0.013

How about this part?

RNAseq in its essence

Condition 1

Condition 2



Gene A
control

Gene A
cold shock

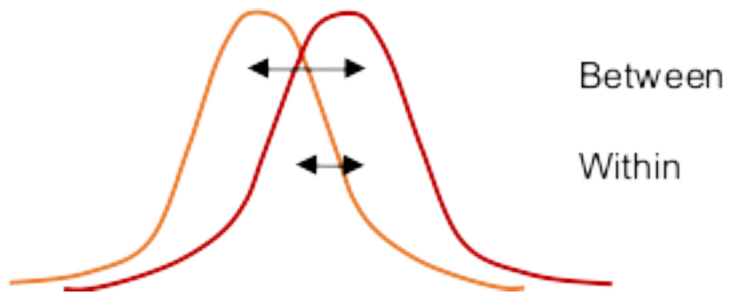
Pieces you'll need to put together

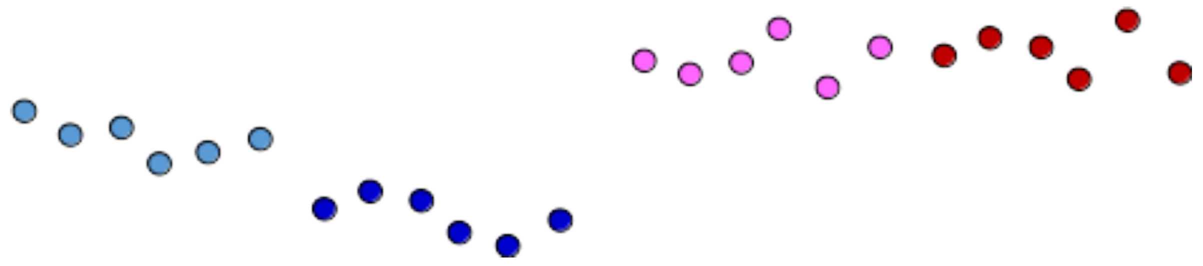
Statistical analysis and hypothesis testing

Two-conditions:
T-test

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{2/n}}$$

$$s_p = \sqrt{\frac{s^2_{X_1} + s^2_{X_2}}{2}}$$





Any questions?

Multiple testing and FDR?

FDR-controlling procedures are designed to control the FDR, which is the expected proportion of "discoveries" (rejected null hypotheses) that are false (incorrect rejections of the null). Equivalently, the FDR is the expected ratio of the number of false positive classifications (false discoveries) to the total number of positive classifications (rejections of the null). The total number of rejections of the null include both the number of false positives (FP) and true positives (TP).

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

Primary category	Tool name	Notes
Splice-aware read alignment	GEM	Filtration-based approach to approximate string matching for alignment
	GSNAP	Based on seed and extend alignment algorithm aware of complex variants
	MapSplice	Based on Burrows-Wheeler Transform (BWT) algorithm
	RUM	Integrates alignment tools Blat and Bowtie to increase accuracy
	STAR	Based on seed searching in an uncompressed suffix arrays followed by seed clustering and stitching procedure; fast but memory-intensive
	TopHat	Uses Bowtie, based on BWT, to align reads; resolves spliced reads using exons by split read mapping
Transcript assembly and quantification	Cufflinks	Assembles transcripts to reference annotations or de novo and quantifies abundance
	FluxCapacitor	Quantifies transcripts using reference annotations
	iReckon	Models novel isoforms and estimates their abundance
Differential expression (DE)	BaySeq	Count-based approach using empirical Bayesian method to estimate posterior likelihoods
	Cuffdiff2	Isoform-based approach based on beta negative binomial distribution
	DESeq	Exon-based approach using the negative binomial model
	DEGSeq	Isoform-based approach using the Poisson model
	EdgeR	Count-based approach using empirical Bayes method based on the negative binomial model
	MISO	Isoform-based model using Bayes factors to estimate posterior probabilities
Other tools	HCP	Normalizes expression data by inferring known and hidden factors with prior knowledge
	PEER	Normalizes expression data by inferring known and hidden factors using a probabilistic estimation based on the Bayesian framework
	Matrix eQTL	Fast eQTL detection tool that uses linear models (linear regression or ANOVA)