C3G Analysis Workshop: RNA-Seq

Part III: Introduction to RNA-seq

January 22-23, 2019

Canadian Centre for Computational Genomics

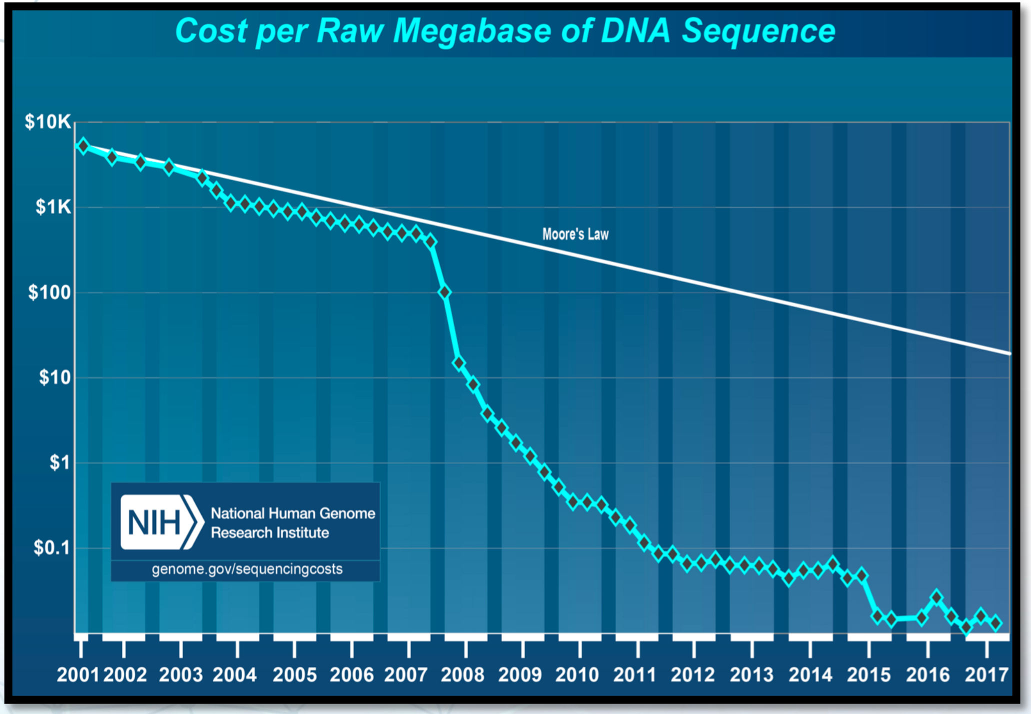
Learning objectives

1. Understand the technical principles behind NGS
2. Understand the biological principles behind RNA-seq
3. Understand the standard steps of RNA-seq analyses
4. Introduce the GenPipes RNA-seq pipeline

Part 1:
Principles of NGS

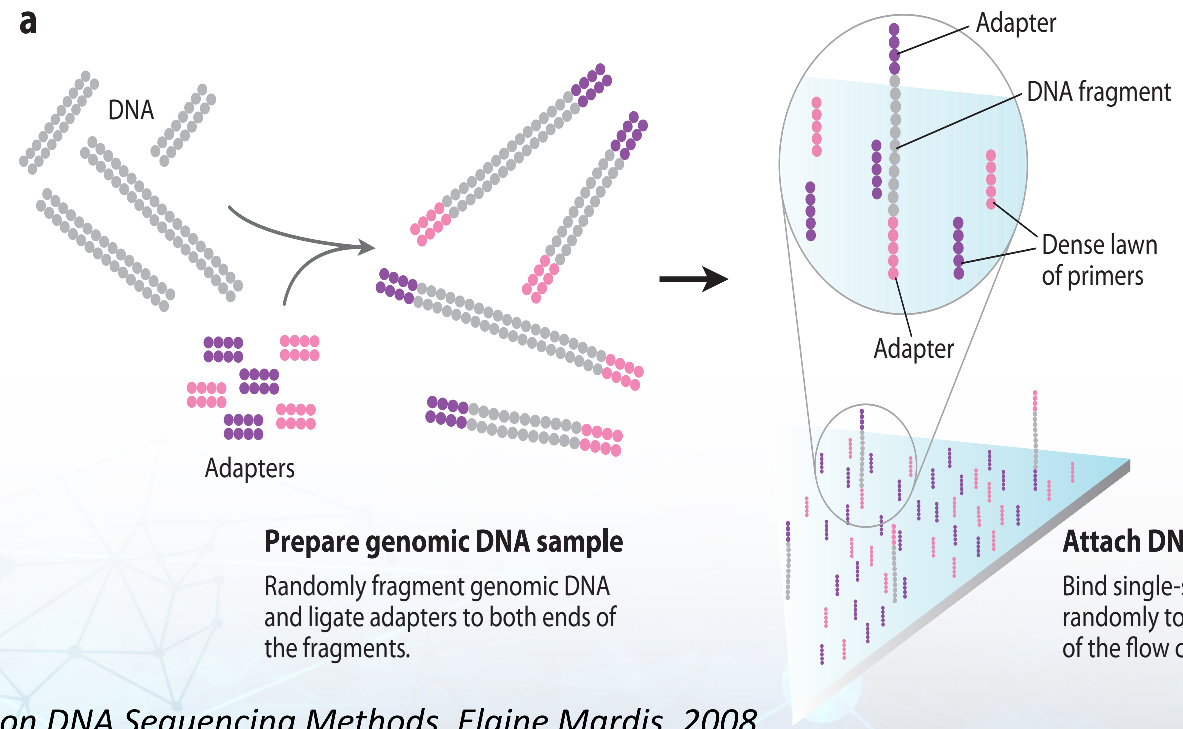
“Next Generation” Sequencing (NGS) has Revolutionized Genomics

Canadian Centre for Computational Genomics



- Sequencing costs have dropped dramatically.
- The processing time has also been greatly reduced.

Illumina NGS Technology is Based on Sequencing-by-Synthesis



Prepare genomic DNA sample

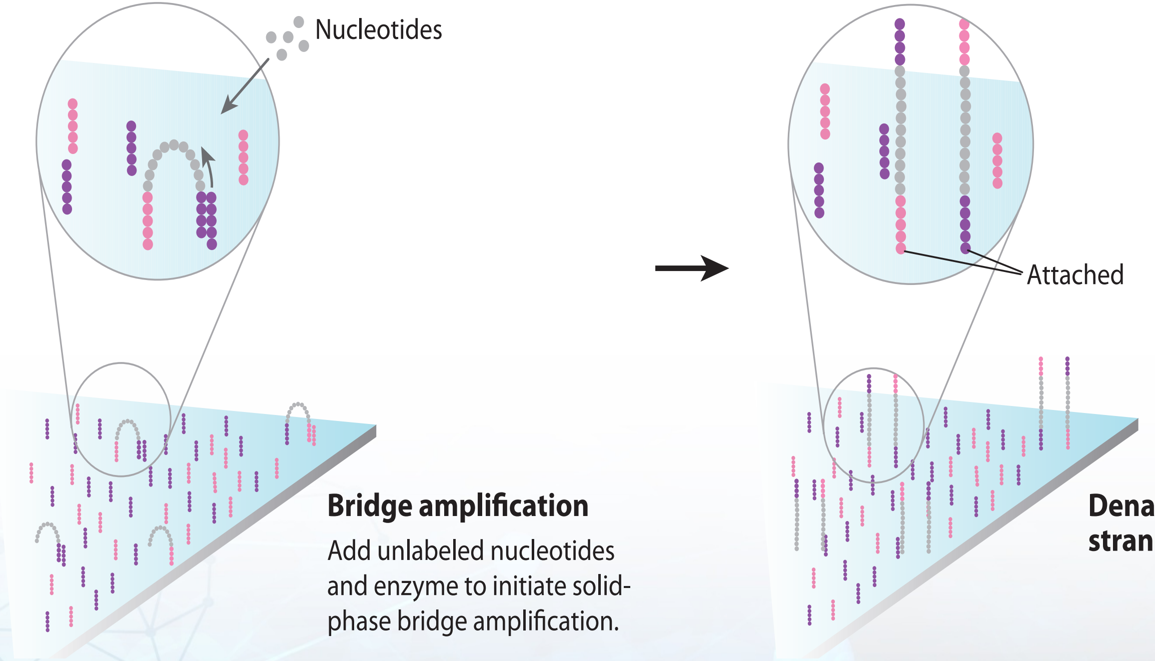
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

Attach DNA to surface

Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

Next-Generation DNA Sequencing Methods, Elaine Mardis, 2008

Illumina NGS Technology is Based on Sequencing-by-Synthesis



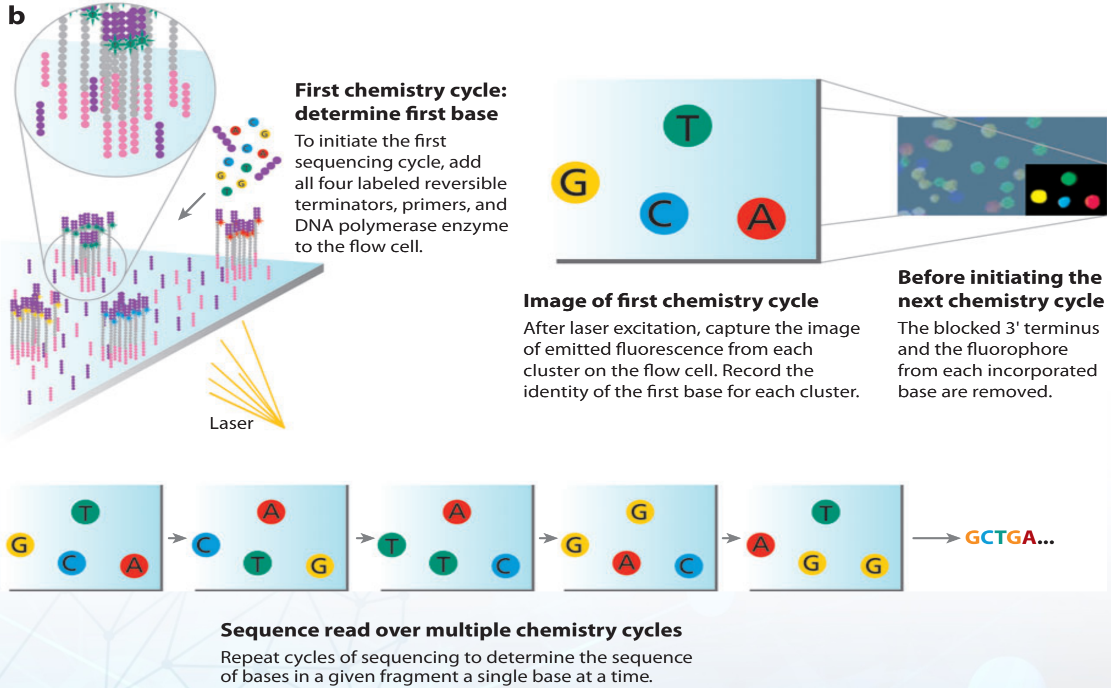
Bridge amplification

Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

Denature the double stranded molecules

Next-Generation DNA Sequencing Methods, Elaine Mardis, 2008

Illumina NGS Technology is Based on Sequencing-by-Synthesis



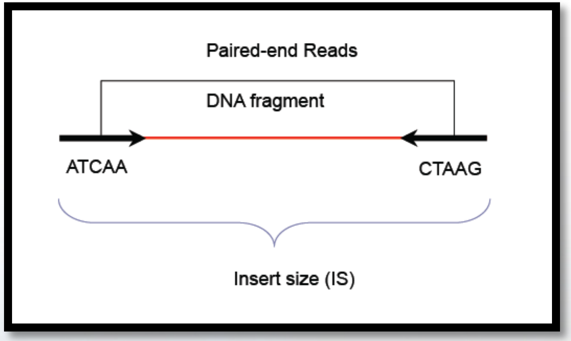
First chemistry cycle: determine first base
To initiate the first sequencing cycle, add all four labeled reversible terminators, primers, and DNA polymerase enzyme to the flow cell.

Image of first chemistry cycle
After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.

Before initiating the next chemistry cycle
The blocked 3' terminus and the fluorophore from each incorporated base are removed.

Sequence read over multiple chemistry cycles
Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at a time.

In paired-end libraries, one pair is read first, then the second one, not both at once.



Next-Generation DNA Sequencing Methods, Elaine Mardis, 2008

Sequencing-by-Synthesis offers Many Advantages

- **Low cost and time:** Sequencing-by-Synthesis (*Illumina)* is usually the cheapest sequencing option with the shortest turnaround time
- **Versatility:** many different types of analyses and libraries can be sequenced using this kind of sequencer
  - Including new libraries that allow for single-cell resolution
- **Support:** because it is the most common type of sequencing, it is supported by most providers and software packages

Sequencing-by-Synthesis also has Important Drawbacks

- **Relatively short reads:** *Illumina* can provide up to 250-300bp reads, but for now 100-150bp is still the standard
- **Sequencing errors:** although quite low compared to other alternatives (approx. 0.1%)

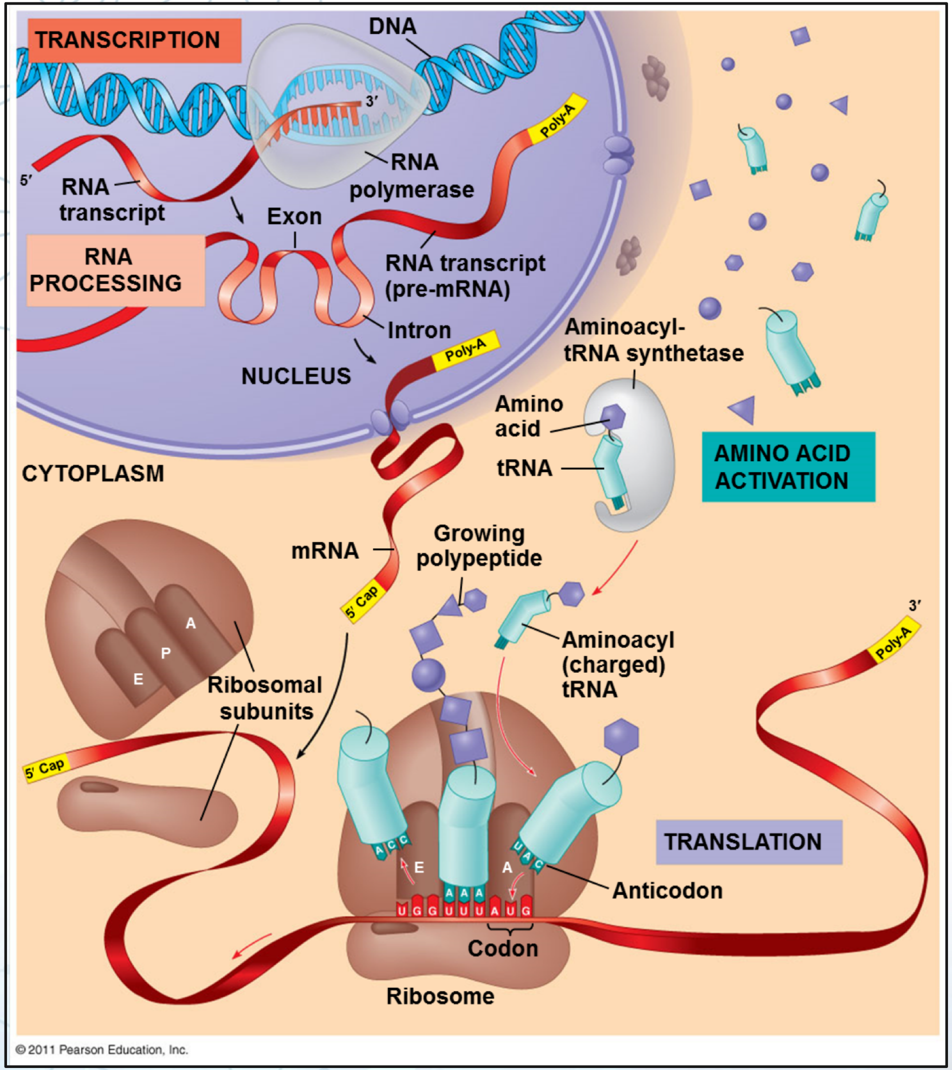
Phred Quality Scores

Indicate the probability of each base call being correct

(higher score = higher quality)

Phred Score	Prob. of Incorrect Base Call	Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	**1 in 1000**	**99.9%**
40	1 in 10,000	99.99%

Part 2:
Principles of RNA-seq



In brief:

RNA-seq is focused on analyzing and comparing a collection of **RNA molecules** (library) *from one or more samples.*

RNA-Seq can help answer several types of biological questions

What genes are being expressed?

- Transcriptome profiling

Is there a difference in gene expression between two conditions?

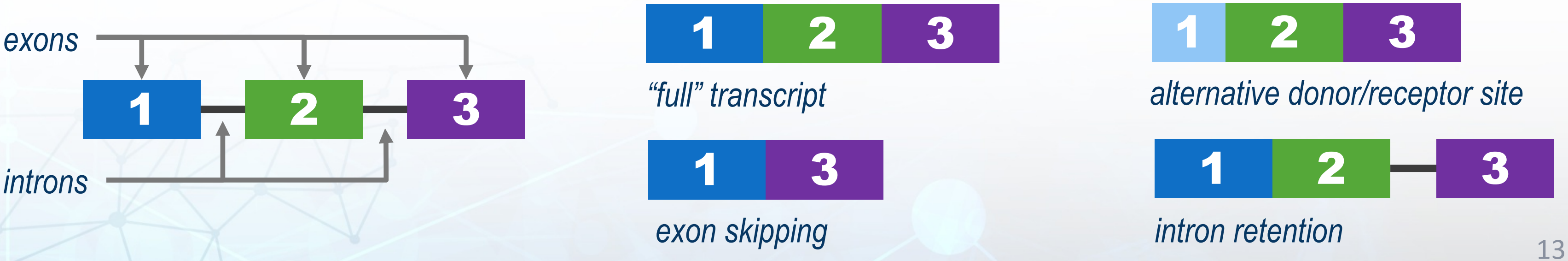
- Differential expression analysis

Are there novel genes/transcripts being expressed?

- Alternative splicing, gene fusions, etc.
- *De novo* assembly

There are several complications to RNA-Seq analyses

- RNA is a less stable molecule than DNA
- RNA usually has small exons separated by introns
- **Very** large variation in abundances
- RNA molecules have very different sizes
- Gene splicing complicates assigning reads to transcripts



"full" transcript

alternative donor/receptor site

exon skipping

intron retention

It is very important to consider the library preparation method for RNA-Seq analyses

What is the library preparation strategy?

- **Total RNA:** Abundant RNA's dominate, high amounts of unprocessed RNA, rRNA and genomic DNA.
- **rRNA reduction:** Abundant rRNA's de-emphasized, still high amounts of unprocessed RNA and genomic DNA.
- **PolyA selection:** Limited transcript representation, low unprocessed RNA and genomic DNA.
- **cDNA capture:** Targeted transcript representation (using cDNA), all other RNA molecules de-emphasized.

Experimental design should consider the hypotheses and factors affecting RNA-Seq

How many replicates do you need?

- **Technical replicates:** Sequences derived from the *same sample* (lanes, flow cells, etc.)
  - More **technical replicates** are recommended if higher coverage is required
- **Biological replicates:** Sequences derived from *different samples*, but with the same phenotype/genotype or experimental condition
  - Recommended: *minimum* of 3 biological replicates per experimental group
  - More replicates are recommended if samples are expected to have high variation

Experimental design should consider the hypotheses and factors affecting RNA-Seq

Canadian Centre for Computational Genomics

How much coverage do you need?

Depends on the purpose (*examples for human*):

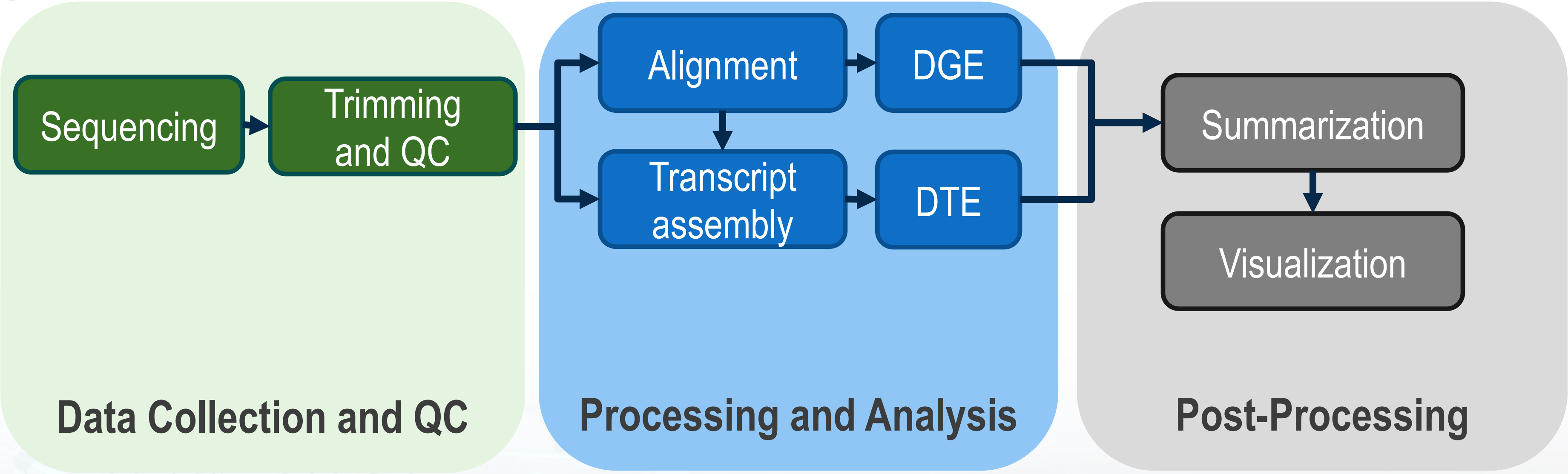
Type of experiment	No. of mapped reads (*per sample*)	Length of reads
Gene expression profiling	10-25 million	50-75 bp
Differential analysis and alternative splicing	40-60 million	75 bp
Transcriptome assembly	> 100 million	> 75 bp
miRNA and sRNA analysis	1-5 million (targeted)	50 bp (single-end)

Adjust for smaller/larger genomes

Check ***illumina*** *website for updated guidelines and costs*

Part 3:
RNA-seq Standard Analysis

Most RNA-Seq analyses follow similar steps

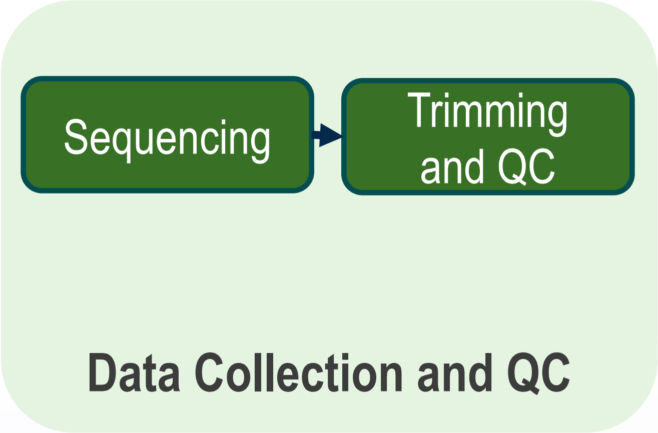


The first steps ensure the quality of the sequencing data

Review of raw data:

Using data provided by sequencing provider.

- Samples are complete and properly named
- Initial library sizes are similar
- No large technical issues
  - No sudden drops in quality
  - Read length is appropriate
- Reads mostly align to organism of interest
  - **Check using BLAST**



Raw sequences are usually reported in FASTQ format

There are two main formats for raw sequencing data:

- **FASTA:** sequence data
- **FASTQ:** sequence data + quality

These are text files (not binary), which means:

- They have **several possible extensions**:
  .fasta, .fa, .fastq, .fq
- They **can be very large** in size
  - Often compressed with gzip (extension .gz)

Data Collection and QC

The FASTA format is the basic sequence data format

FASTA format characteristics:

- **FASTA record start:** **>** symbol
- **Header:** text after **>**
- **Sequence:** subsequent line(s) after header
  - Lines should not be too long
  - Lines should have same width

The FASTA format is loosely defined, so there may be variations based on source!

example.fa

```
> sequence1
ATGCATGCATGCATGCATGC
ATGCATGCATGATGCATGCA
TGCATGCA
> sequence2
GCATTGCATCATGCATGCAT
TGCATCAATGTGCATGCCAT
ATG
```

The FASTQ format is similar to FASTA with the addition of Phred scores

Canadian Centre for Computational Genomics

FASTQ format characteristics:

- **FASTQ record start:** **@** symbol
- **Header:** text after **@**
- **Sequence:** *single* line after header
- **Section separator:** **+** symbol (optional header)
- **Quality:** line with encoded **Phred score**
  - <u>Same length as sequence</u>

The FASTQ format is loosely defined, so there may be variations based on source!

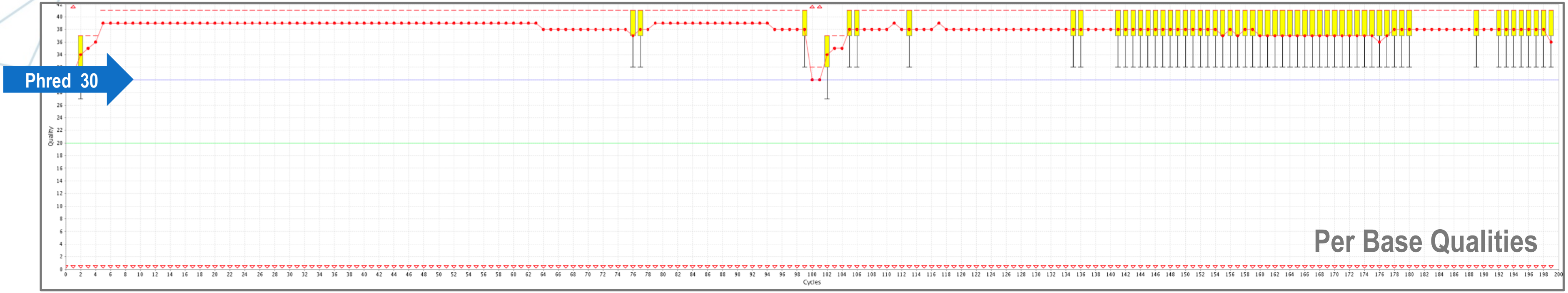
example.fq

```
@ sequence1
ATGCATGCATGCATGCATGC
+ sequence1
!''*((((***+))%%%++%
@ sequence2
GCATGCATATGCATGCATGC
+ sequence2
(((***+))%!''*(%%++%
```

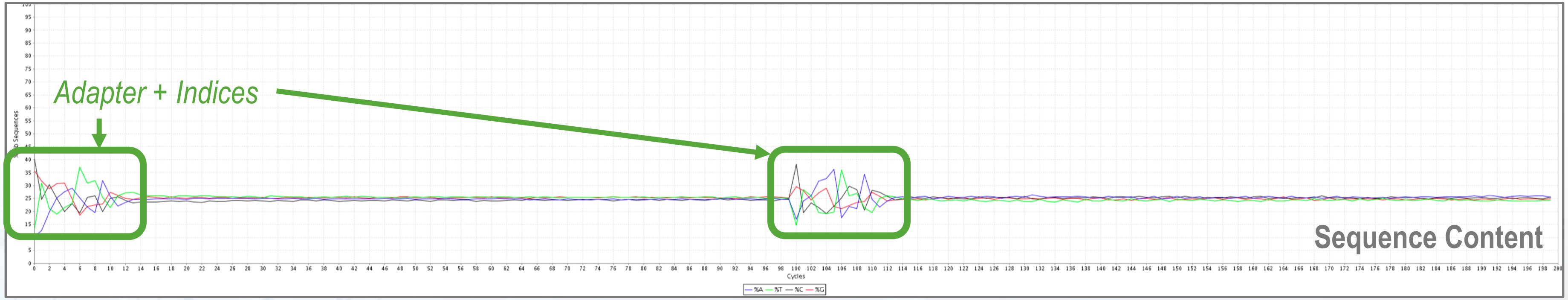
The first steps ensure the quality of the sequencing data

Canadian Centre for Computational Genomics

FastQC



Per Base Qualities



Adapter + Indices

Sequence Content

Data Collection and QC

Trimming removes adapter sequences and low quality reads

Provide adapter sequences to trimming software

Set thresholds for quality and read length

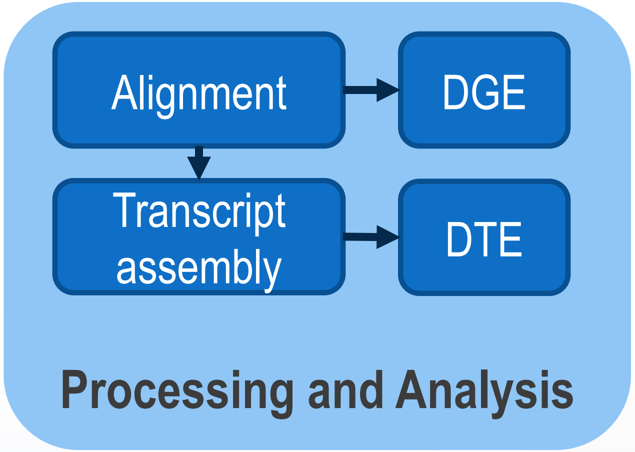
- Minimum quality (phred score) should be 30
- Minimum length of reads should be around 60% of original length

Software:

- `Trimmomatic`
- `FastQC`
- `FASTX-Toolkit`

The key to RNA-Seq analysis is how reads are assigned and counted

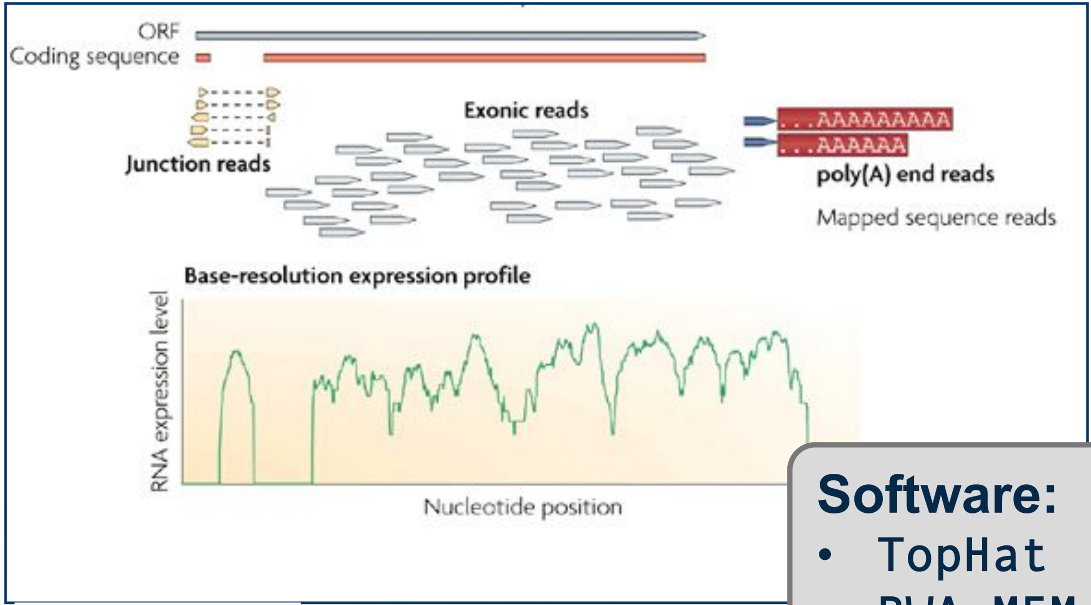
1. **Assign** reads to genes or transcripts
   - Alignment (genome/transcriptome)
   - Assembly (*de novo*/guided)
   - *Pseudo-Alignment**
2. **Estimate** abundances
3. **Compare** abundances
   - Normalization



**Not covered in this course*

Mapping RNA reads requires an adequate alignment strategy

Reads may span large introns, so using **splice-aware** aligners is key.



Nature Reviews | Genetics

Software:

- TopHat
- BWA-MEM
- STAR
- HISAT2

Mapping data is usually reported in the SAM/BAM format

SAM: Sequence Alignment Format

- BAM files are just binary SAM files
- Usually sorted and indexed (.bai)

Composed of two sections:

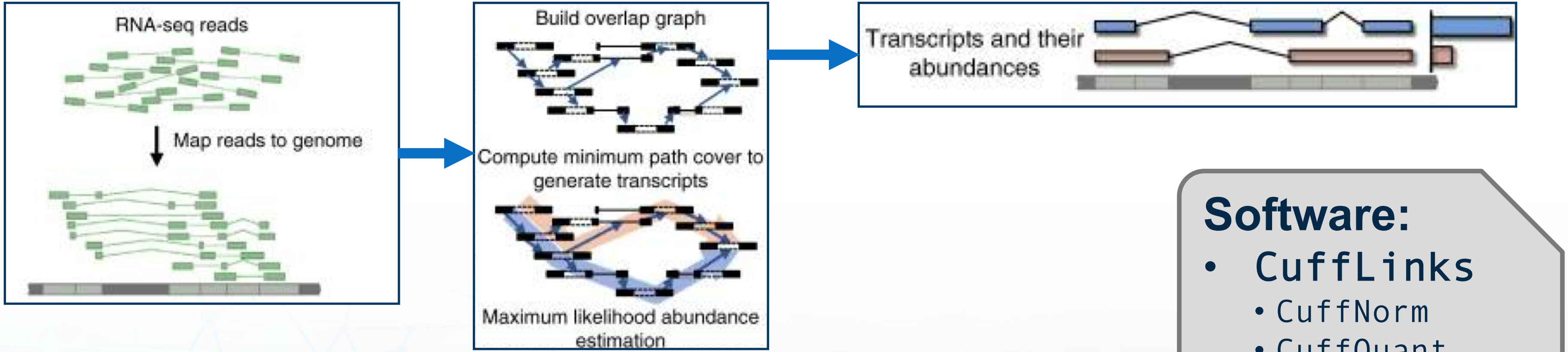
- **Header section:** information about reference, aligner and flags (lines begin with @)
- **Alignment section:** each row represents a query sequence, and includes its name, position in reference, flags, mapping quality, etc.

Processing and Analysis

Assembling transcripts can help answer additional biological questions

- Discovery of novel splice variants
- Differential transcript analysis

CuffLinks

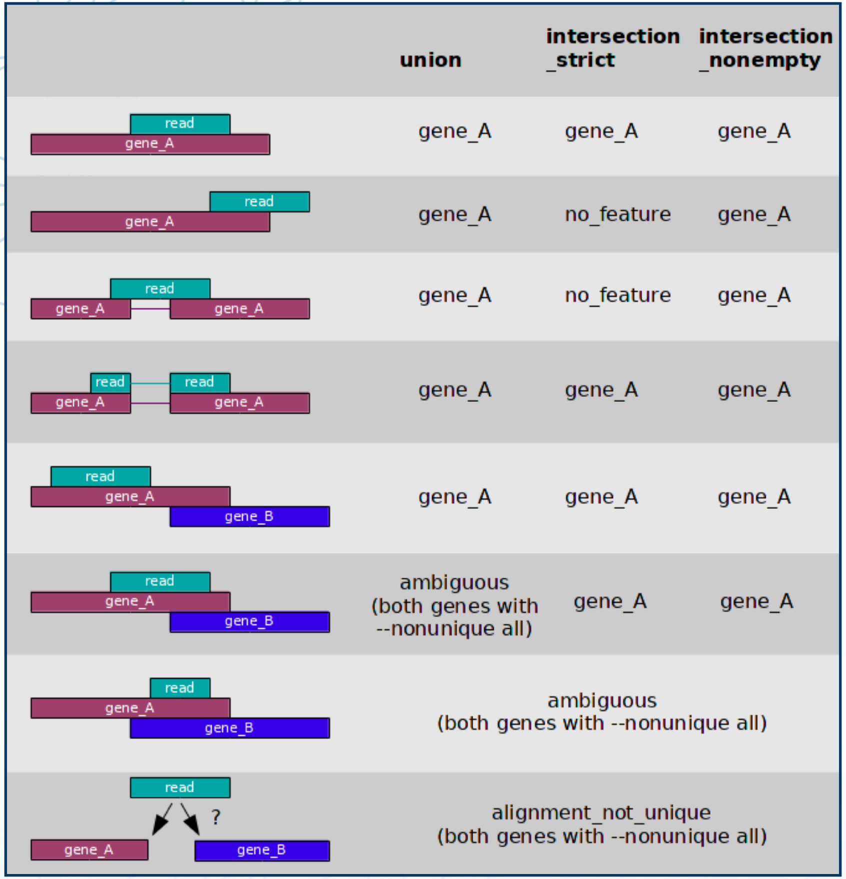

RNA-seq reads
Map reads to genome
Build overlap graph
Compute minimum path cover to generate transcripts
Maximum likelihood abundance estimation
Transcripts and their abundances

Processing and Analysis

Software:

- CuffLinks
  - CuffNorm
  - CuffQuant
  - CuffDiff
- StringTie
  - Ballgown

Alignment data is used to estimate gene expression



	union	intersection_strict	intersection_nonempty
read gene_A	gene_A	gene_A	gene_A
read gene_A	gene_A	no_feature	gene_A
read gene_A gene_A	gene_A	no_feature	gene_A
read read gene_A gene_A	gene_A	gene_A	gene_A
read gene_A gene_B	gene_A	gene_A	gene_A
read gene_A gene_B	ambiguous (both genes with --nonunique all)	gene_A	gene_A
read gene_A gene_B	ambiguous (both genes with --nonunique all)		
read ? gene_A gene_B	alignment_not_unique (both genes with --nonunique all)		

It is important to think how reads that align to more than one gene are counted.

Software:

- `HTSeq-Count`
- `featureCount`

Processing and Analysis

Comparing expression requires normalization and statistical tests

Counts are normalized to account for:

- Library size
- Effective feature length

It is important to know if and how your "counts" have been normalized.

FPKM/RPKM are normalized units!

Software:

- DESeq2
- EdgeR
- Ballgown
- CuffDiff
- Sleuth

Processing and Analysis

Comparing expression requires normalization and statistical tests

The most simple statistical test is a pairwise comparison

Hypothesis: gene/transcript expression changed between two conditions
Null hypothesis: gene/transcript expression did not change…

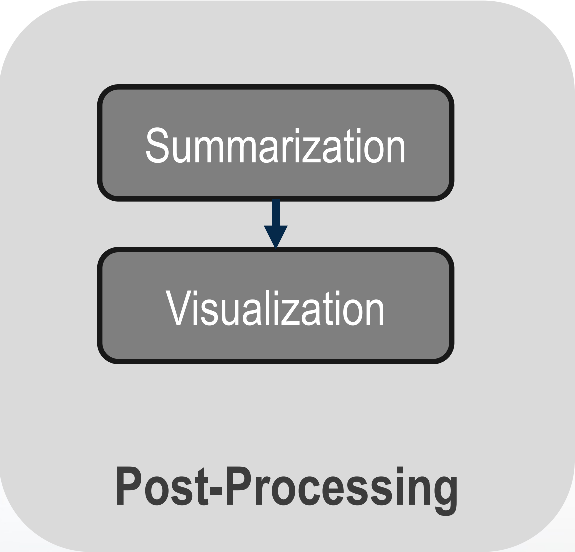
P-value: probability of the ***null*** hypothesis

- Lower P-values indicate a lower probability that the effect is due to random chance
- Smaller P-values do not always indicate “stronger” or “better” results
- Use P-values as a *cutoff* to select values for further analysis, but not to “rank” them

Summarizing and interpreting results is key to gaining knowledge

Once statistical tests have been performed, results should be contextualized and validated

•Different approaches depending on the purpose of the experiment

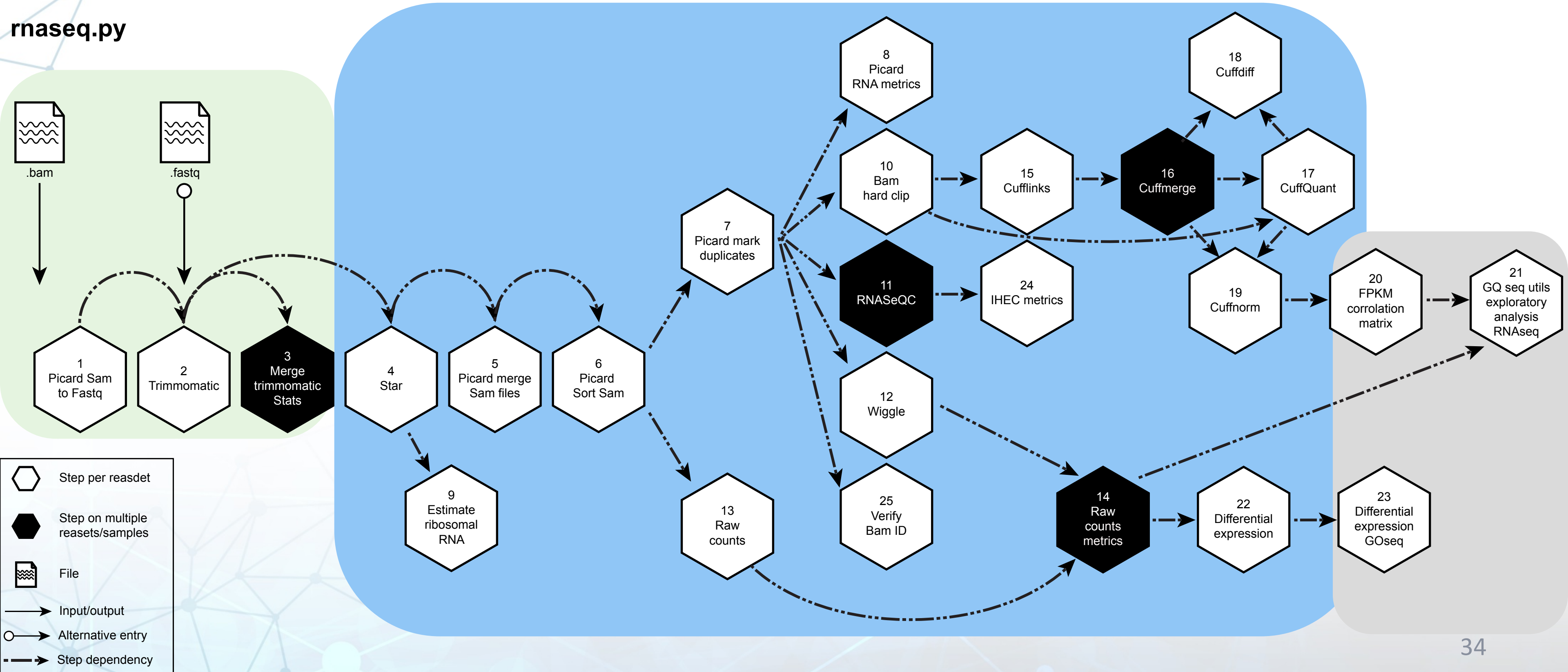


Part 4:
RNA-seq with GenPipes

The GenPipes RNA-Seq workflow

Canadian Centre for Computational Genomics

rnaseq.py



.bam

.fastq

1 Picard Sam to Fastq

2 Trimmomatic

3 Merge trimmomatic Stats

4 Star

5 Picard merge Sam files

6 Picard Sort Sam

7 Picard mark duplicates

8 Picard RNA metrics

9 Estimate ribosomal RNA

10 Bam hard clip

11 RNASeQC

12 Wiggle

13 Raw counts

14 Raw counts metrics

15 Cufflinks

16 Cuffmerge

17 CuffQuant

18 Cuffdiff

19 Cuffnorm

20 FPKM corrolation matrix

21 GQ seq utils exploratory analysis RNAseq

22 Differential expression

23 Differential expression GOseq

24 IHEC metrics

25 Verify Bam ID

⬡ Step per reasdet

⬢ Step on multiple reasets/samples

File

→ Input/output

○→ Alternative entry

- - → Step dependency

There are two types of files that can be used as input for the pipeline

Starting from BAM files (step 1)

- The BAM files will be converted back to FASTQ files, and aligned again with appropriate parameters
- Make sure that BAM files include **unaligned reads**

Starting from FASTQ files (skips step 1)

- Don’t skip the trimming step

The STAR two-pass alignment method increases novel junction discovery

Two-step alignment method:

1. **First pass mapping**
   - Using regular parameters
   - Detect novel junctions
2. **Merge novel junctions** discovered in first alignment
   - Create new genome indices with all junctions (**SJ.out.tab**)
3. **Second pass mapping**
   - Using new genome index

Differential Analysis for both genes and transcripts

Differential Gene Analysis:

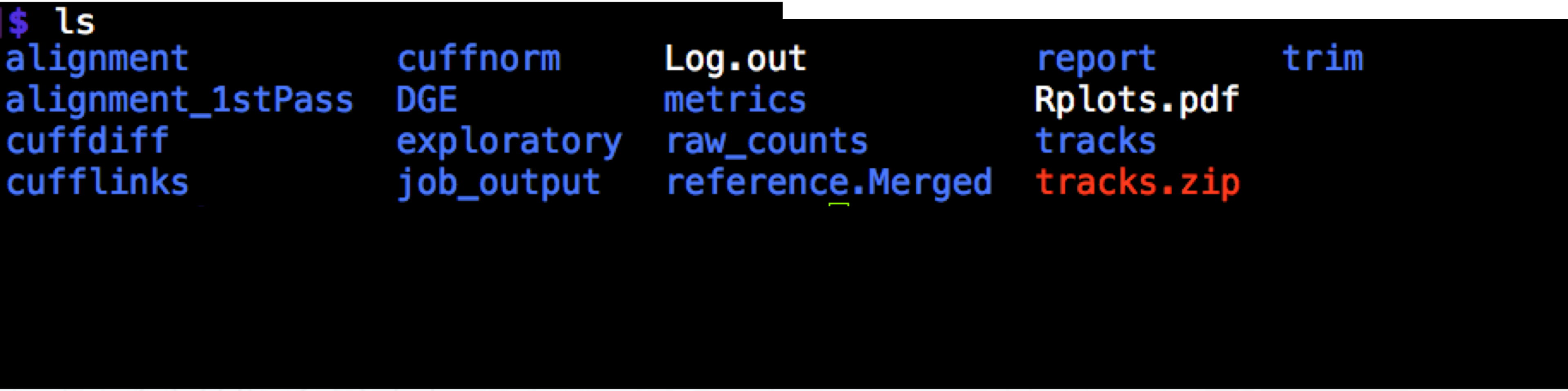
- Raw counts with `HTSeq-count`
- Differential analysis using both `DESeq2` and `EdgeR`
- Differential GO analysis using `GOSeq`

Differential Transcript Analysis:

- Transcript assembly with `CuffLinks`
- Raw counts with `CuffMerge`, `CuffCount`
- Differential analysis with `CuffDiff`

Outputs will be saved in different appropriately labeled directories

GenPipes output structure:

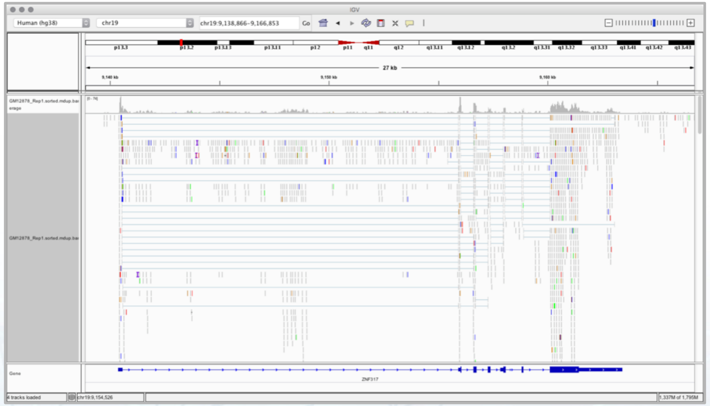


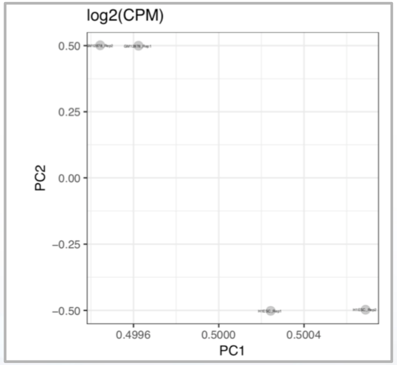
```
$ ls
alignment          cuffnorm     Log.out           report      trim
alignment_1stPass  DGE          metrics           Rplots.pdf
cuffdiff           exploratory  raw_counts        tracks
cufflinks          job_output   reference.Merged  tracks.zip
```

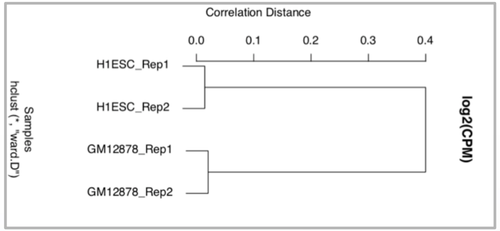
Post-Processing

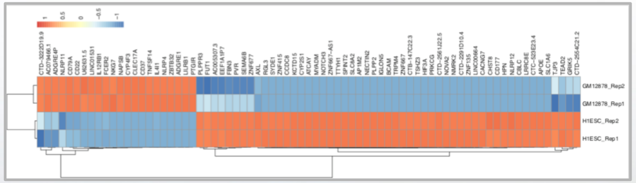
GenPipes generates a report with summary and visualization

- HTML report with links to plots, tables and data
- Alignment files can be explored with genome browsers
- Use R or spreadsheets for additional data exploration


IGV
Human (hg38)
chr19
chr19:9,138,866-9,166,853
Go
27 kb
GM12878_Rep1.sorted.mdup.bam Coverage
GM12878_Rep1.sorted.mdup.bam
Gene
ZNF317
4 tracks loaded
chr19:9,154,526







Post-Processing

Part 5: Review

Conclusions

1. There are many biological and technical factors that can affect the results of an RNA-seq experiment
2. Most RNA-seq analysis follow similar steps, but there are variations in the methods and assumptions
3. The GenPipes RNA-seq pipeline is a tool that allows for a simple, reproducible way to perform RNA-seq analyses

Acknowledgement

McGill

MiCM McGill initiative in Computational Medicine

compute canada | calcul canada

GenomeCanada

GenomeQuébec

Ontario Genomics

SickKids®

bioinformatics.ca

canarie

IHEC

Calcul Québec

GenAP

Terry Fox PROFYLE

HPC4Health

This presentation includes material prepared by Dr. Mathieu Bourgey