RNA-Seq Analyses

Jessica Holmes

High Performance Biological Computing (HPCBio)
Roy J. Carver Biotechnology Center

CARL R. WOESE INSTITUTE FOR GENOMIC BIOLOGY
WHERE SCIENCE MEETS SOCIETY

ILLINOIS

General Outline

1. From RNA to sequencing data
2. Experimental and practical considerations
3. Commonly encountered file formats
4. Transcriptomic analysis methods and tools
    a. Transcriptome assembly
    b. Differential gene expression

Transcriptome Sequencing (aka RNA-Seq)

Differential Gene Expression

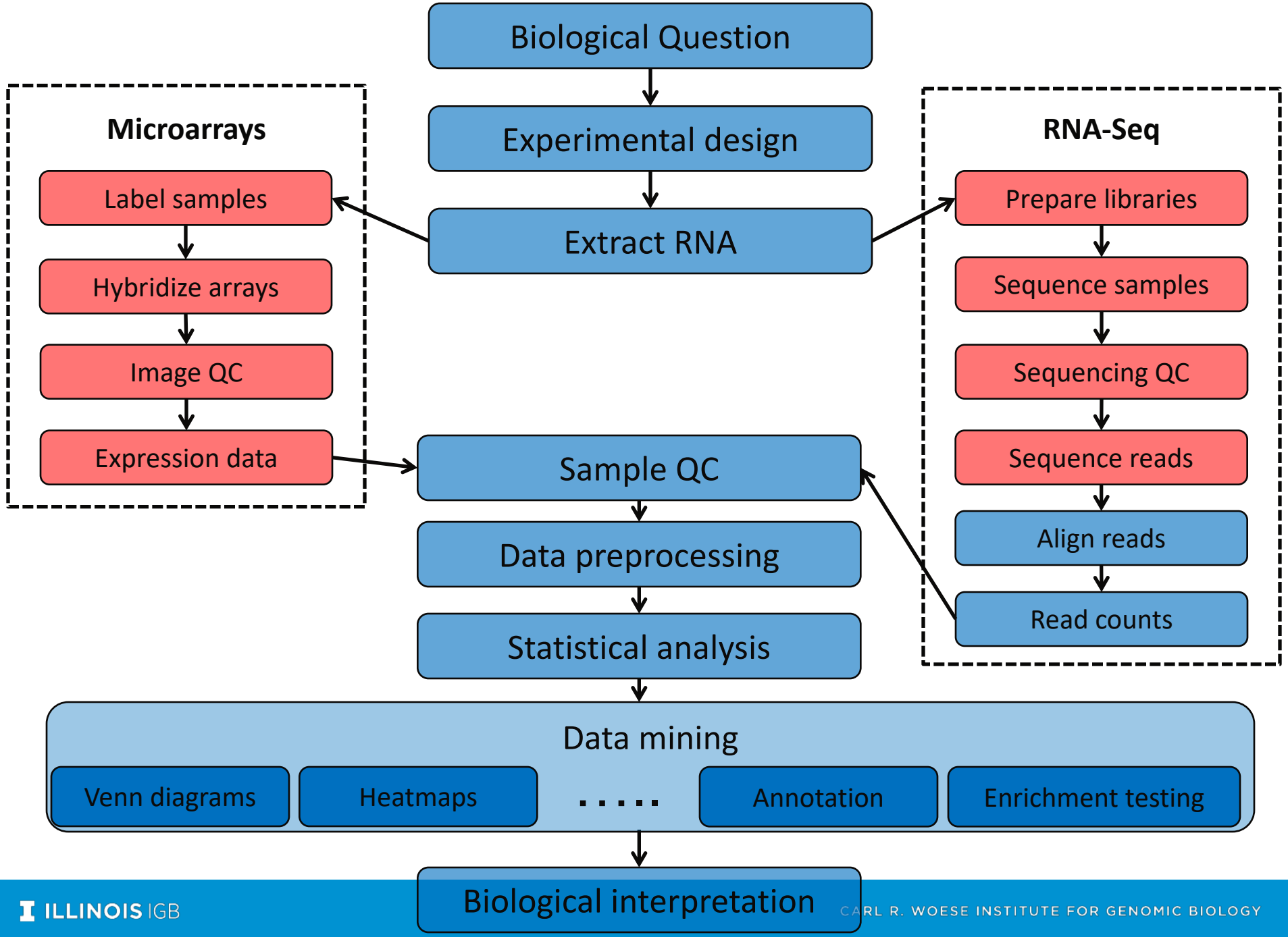
- Quantitative evaluation
- Comparison of transcript levels, usually between different groups
- Vast majority of RNA-Seq is for DGE

Transcriptome Assembly

- Build new or improved profile of transcribed regions (“gene models”) of the genome
- Can then be used for DGE

Metatranscriptomics

- Transcriptome analysis of a community of different species (e.g., gut bacteria, hot springs, soil)
- Gain insights on the functioning and activity rather than just who is present


Biological Question
Experimental design
Extract RNA
Microarrays
Label samples
Hybridize arrays
Image QC
Expression data
RNA-Seq
Prepare libraries
Sequence samples
Sequencing QC
Sequence reads
Align reads
Read counts
Sample QC
Data preprocessing
Statistical analysis
Data mining
Venn diagrams
Heatmaps
. . . . .
Annotation
Enrichment testing
Biological interpretation

Types of RNA

Ribosomal (rRNA)

- Responsible for protein synthesis
- up to 95% of total RNA in a cell

Messenger (mRNA)

- Translated into protein in ribosome
- 3-4% of total RNA in a cell
- have poly-A tails in eukaryotes

Micro (miRNA)

- short (22 bp) non-coding RNA involved in expression regulation

Transfer (tRNA)

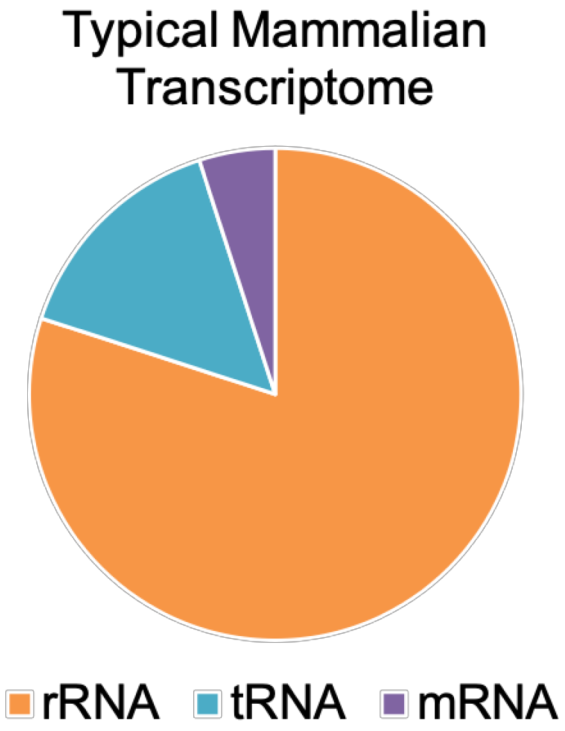
- Bring specific amino acids for protein synthesis

Others (lncRNA, shRNA, siRNA, snoRNA, etc.)

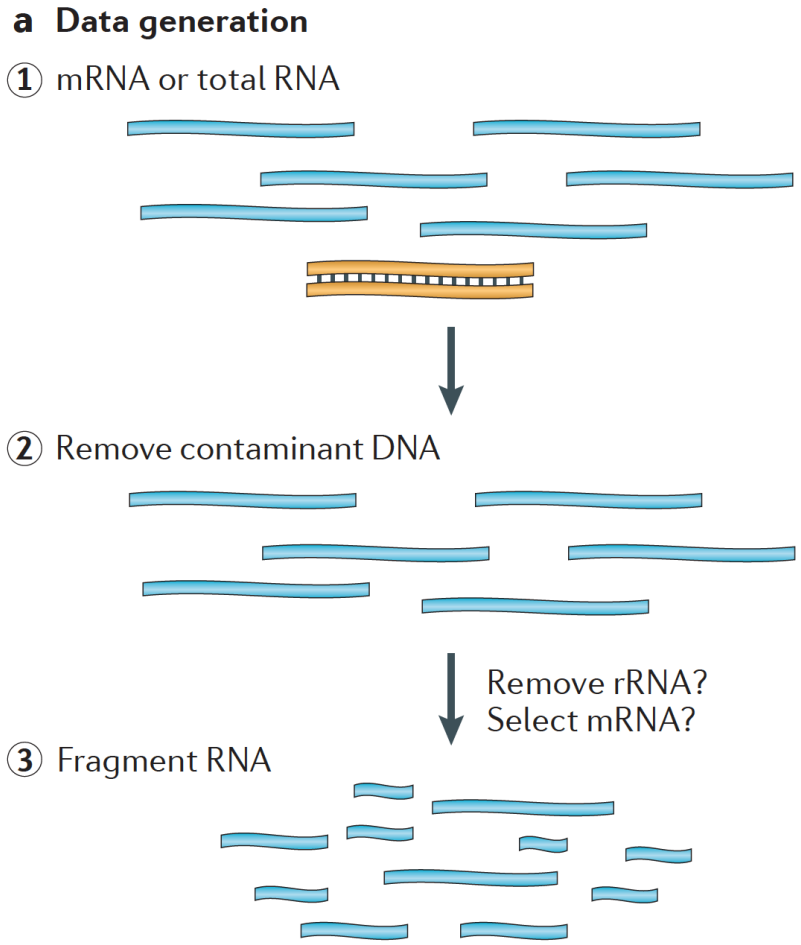
Removal of rRNA is almost always recommended

Removal Methods:

- poly-A selection (eukaryotes only)
- ribosomal depletion
- Size selection

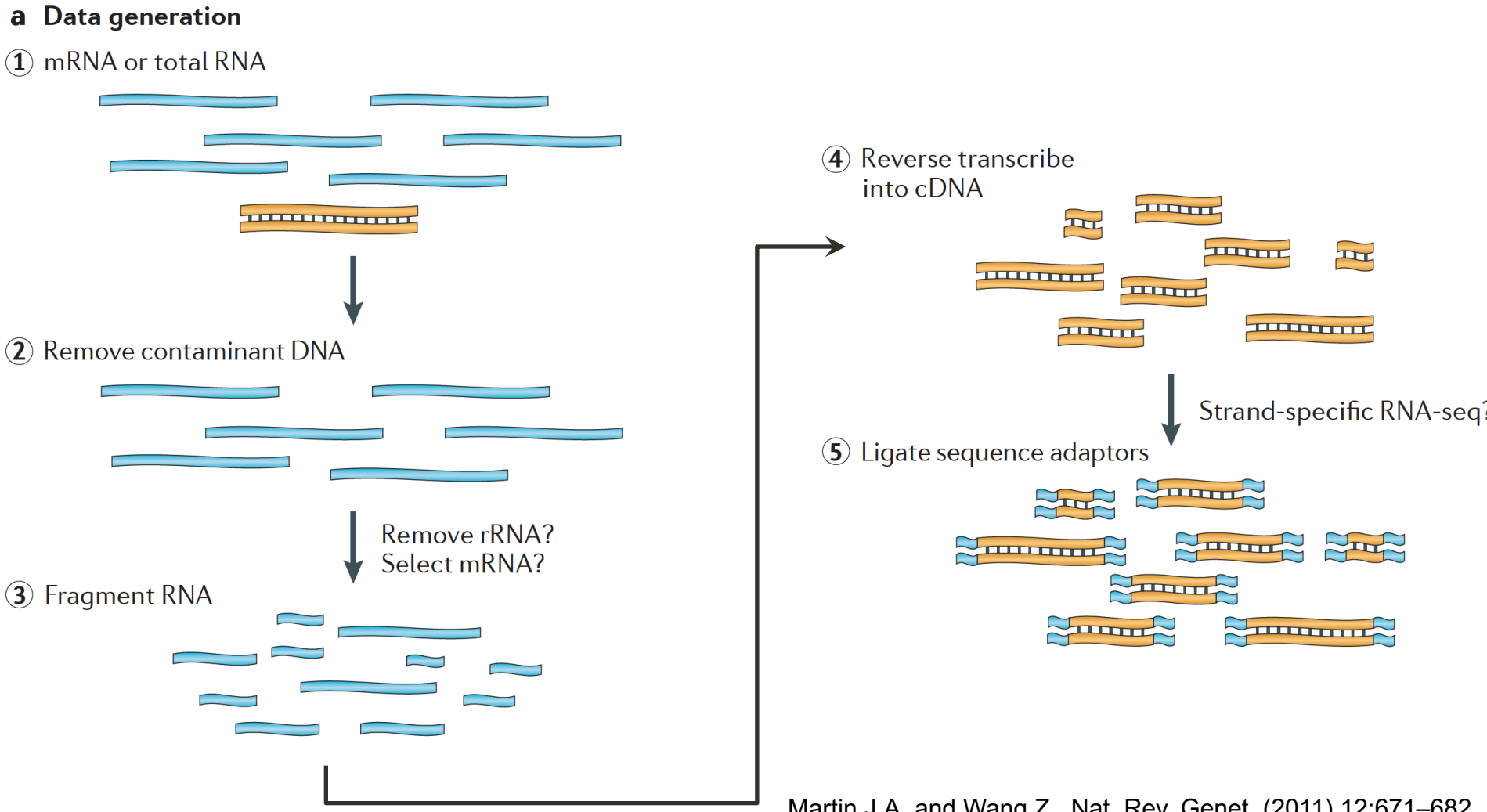


From RNA -> sequence data



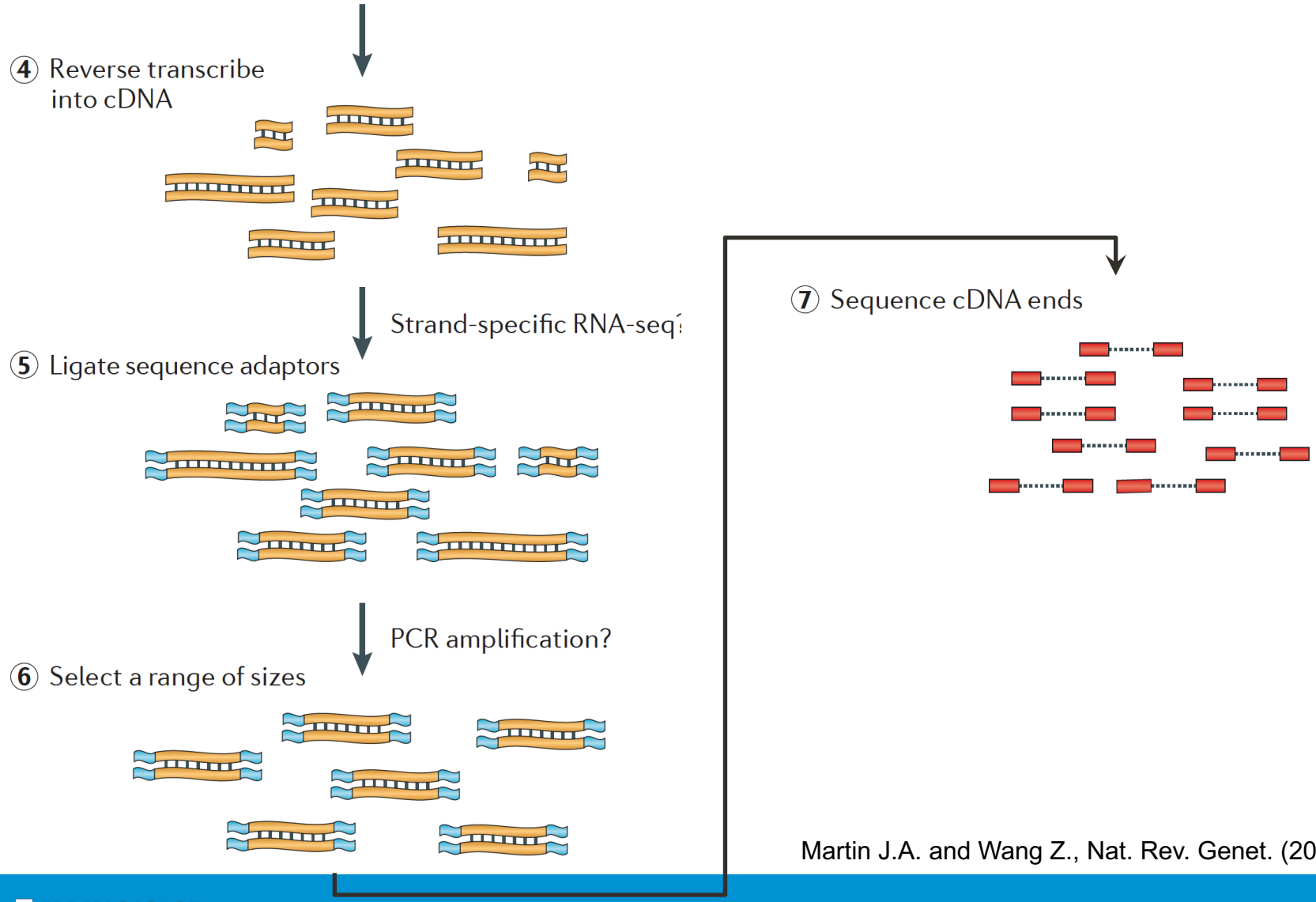
Martin J.A. and Wang Z., Nat. Rev. Genet. (2011) 12:671–682

From RNA -> sequence data



Martin J.A. and Wang Z., Nat. Rev. Genet. (2011) 12:671–682

From RNA -> sequence data



Martin J.A. and Wang Z., Nat. Rev. Genet. (2011) 12:671–682

How do we sequence DNA?

1st generation: **Sanger** method (1987)

2nd generation ("next generation"; 2005):

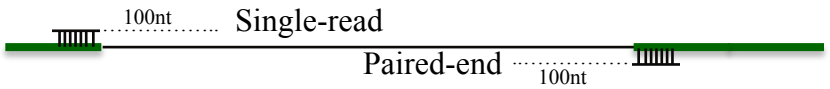
- **454** - pyrosequencing
- **SOLiD** – sequencing by ligation
- **Illumina** – sequencing by synthesis
- **Ion Torrent** – ion semiconductor
- **Pac Bio** – Single Molecule Real-Time sequencing, 1000 bp

3rd generation (2015)

- **Pac Bio** – SMRT, Sequel system, 20,000 bp
- **Nanopore** – ion current detection
- **10X Genomics** – novel library prep for Illumina

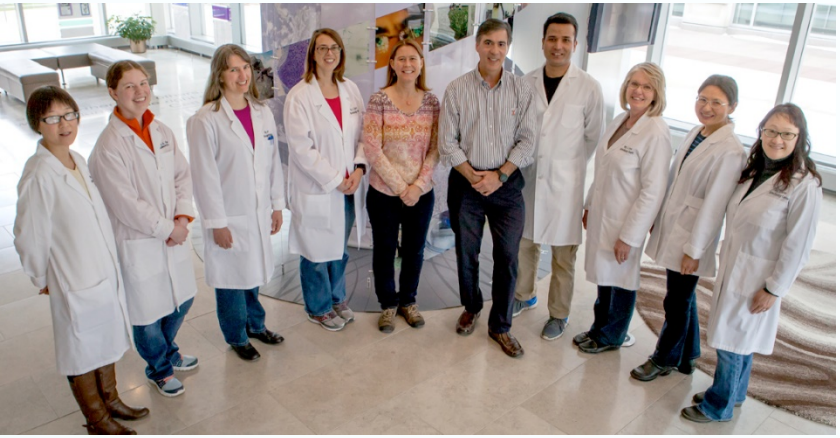
Illumina – “short read” sequencing

- Rapid improvements over the years from 36 bp to **300 bp**; highest throughput at 100/150 bp; many different types of sequencers for various applications.
- Can also “flip” a longer DNA strand and sequence from the other end to get **paired-end reads**

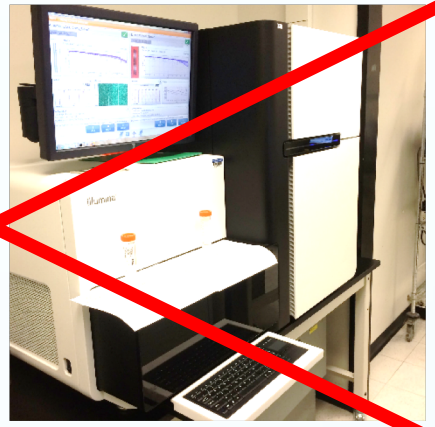


- **Accuracy**: 99.99% **Biases**: yes
- Most common platform for transcriptome sequencing

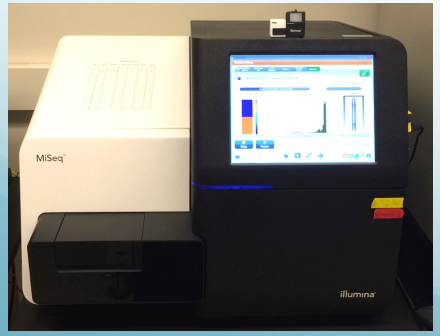
Library Construction and Sequencing Personnel and Equipment



2 Illumina HiSeq 4000 and two 2500

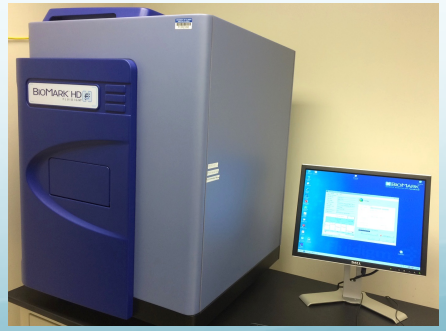


3 MiSeq



2 EpMotion

Fluidigm (FG)



1.5 PB archive

NovaSeq 6000

Any Genome. Any Method. Any Scale.

PE 150 | Q30 ≥ 75%

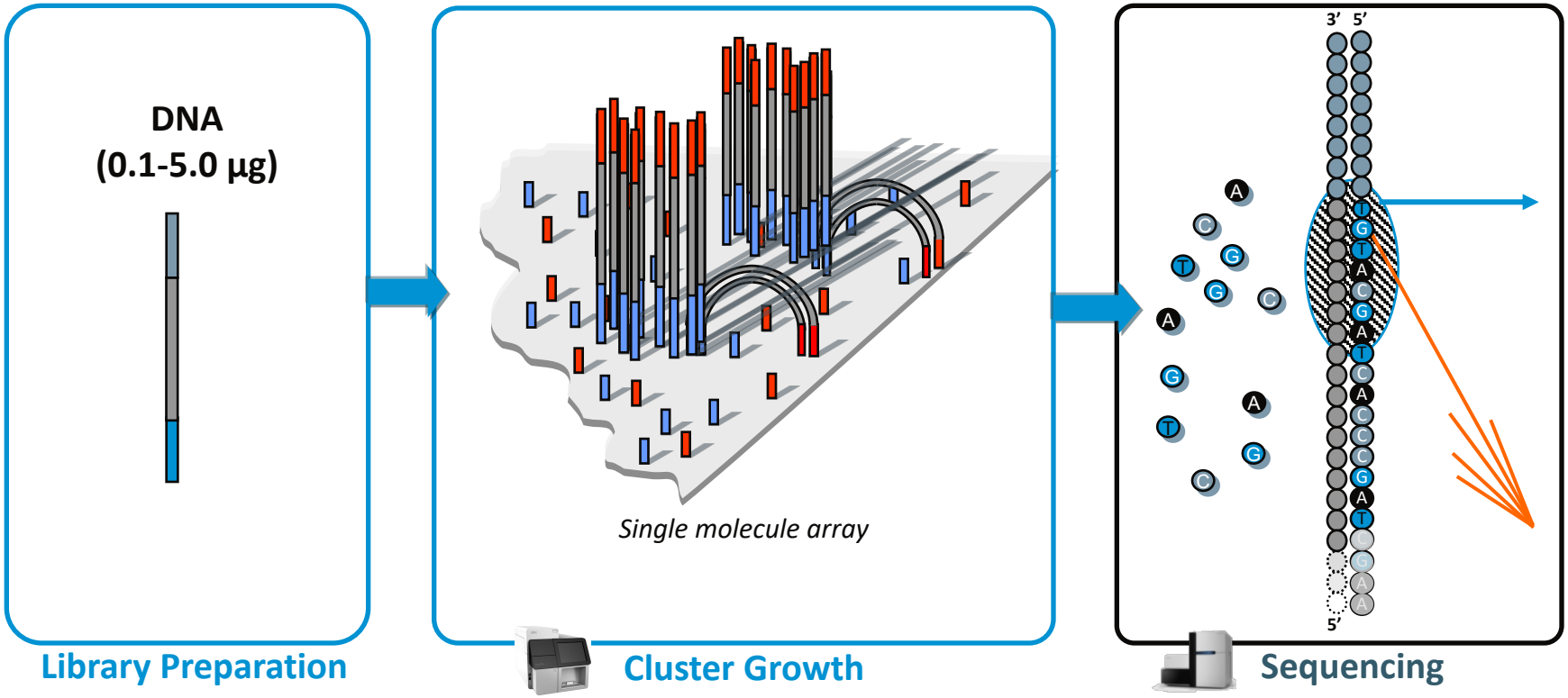
OUTPUT	167 – 3000 Gb
SINGLE READS	1.6 – 10B
RUN TIME	Fastest (40 Hr. for 2T Run)
Flow Cells	Scalable Flow Cell Format

Output and Read Metrics are per flow cell

illumina®

Output range shown based on currently unreleased flow cells

Illumina Sequencing Technology Workflow

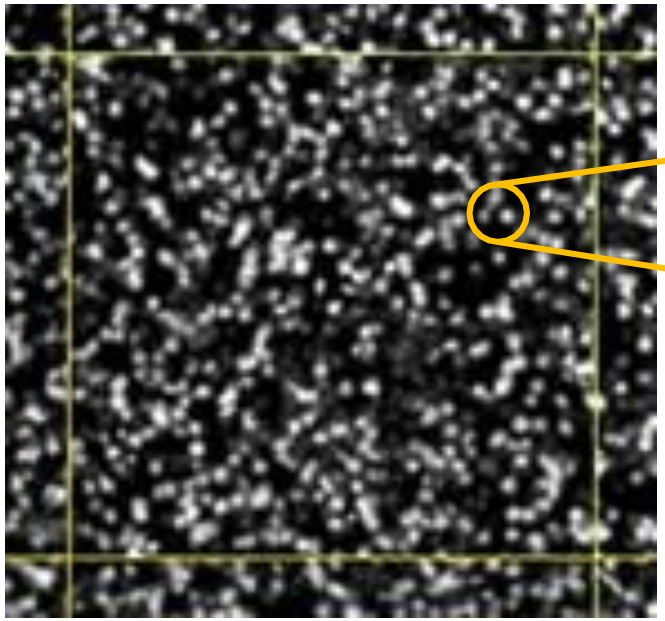


Illumina Sequencing Video

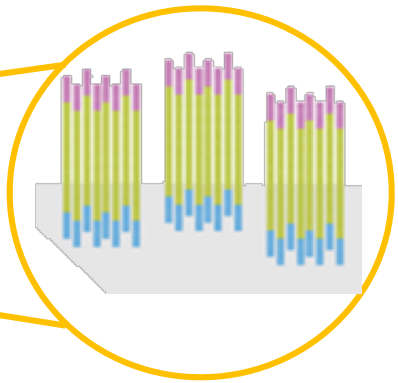
- Introduction to Sequencing by Synthesis

What is a Cluster?

Clusters are bright spots on an image



Each cluster represents thousands of copies of the same DNA strand in a 1–2 micron spot



illumina®

Quality Scoring

Quality Scores	• Estimate the probability of an error in base calling based on a quality model
Quality model	• Includes quality predictors of single bases, neighboring bases and reads
Reported	• After clusters passing filter calculation

ASCII Quality Score	Probability of Incorrect Based Call	Base Call Accuracy	Q-score
+	1 in 10	90%	Q10
5	1 in 100	99%	Q20
?	1 in 1000	99.9%	Q30
I	1 in 10000	99.99%	Q40

illumina®

General Outline

1. From RNA to sequencing data
2. **Experimental and practical considerations**
3. Commonly encountered file formats
4. Transcriptomic analysis methods and tools
    a. Transcriptome Assembly
    b. Differential Gene expression

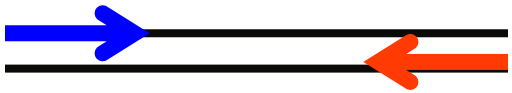
Considerations for…
Differential Gene Expression

- Keep biological replicates separate
- Poly-A enrichment is generally recommended
  - Unless you're interested in non-coding RNA!
- Remove ribosomal RNA (rRNA)
  - Unless you're interested in rRNA!
- Usually single-end (SE) is enough
  - Paired-end (PE) may be recommended for more complex genomes

Single-end read

Read1

ATGTTCCATAAGC…



Paired-end reads

Read1

ATGTTCCATAAGC…

Read2

CCGTAATGGCATG…

Considerations for…
Transcriptome Assembly

- Collect RNA from many various sources for a robust transcriptome
  - These can be pooled before or after sequencing (but before assembly)
- Poly-A enrichment is optional depending on your focus
- Remove ribosomal RNA (rRNA)
  - Unless you're interested in rRNA!
- Paired-end (PE) is recommended. The more sequence, the better.
  - Even better if you use long-read technology in addition

Considerations for…
Metatranscriptomics

- Keep biological replicates separate
- Poly-A enrichment is optional depending on your focus
- Remove ribosomal RNA (rRNA)
- Paired-end (PE) reads will help you separate out orthologous genes
- May need to remove host mRNA computationally downstream
    - e.g. removing human mRNA from gut samples

Experimental Design Issues

(or Why you need to think about how you will analyze the data ***before*** you do the experiment)

- Poorly designed experiments (especially with confounding factors) can lead to lower power to detect differences, ambiguous results, or even a waste of time and money!
- What to consider:
  - How many factors do you have?
  - How many levels per factor?
  - How many independent replicates should you do? (3 minimum, 5 is better, and put 5 more in the -70 if you can)
- The more complex the experiment, the more difficult the statistical analysis will be.

How many independent replicates (N)?

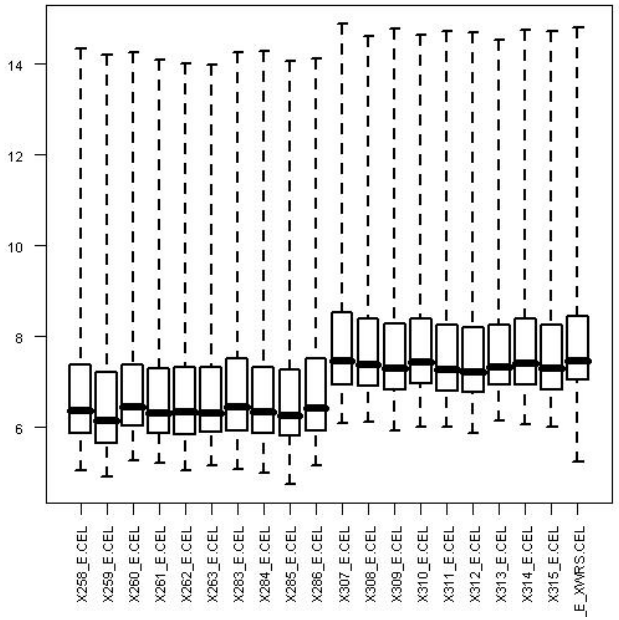
Realistically, the most-used formula is:

$$N = \frac{(\$ \text{ you have})}{(\$ \text{ / measurement})}$$

Inspiration and graphic from Jeff Leek's Statistics for Genomic Data Science course on Coursera.org

Beware confounding factors! (aka batch effects)

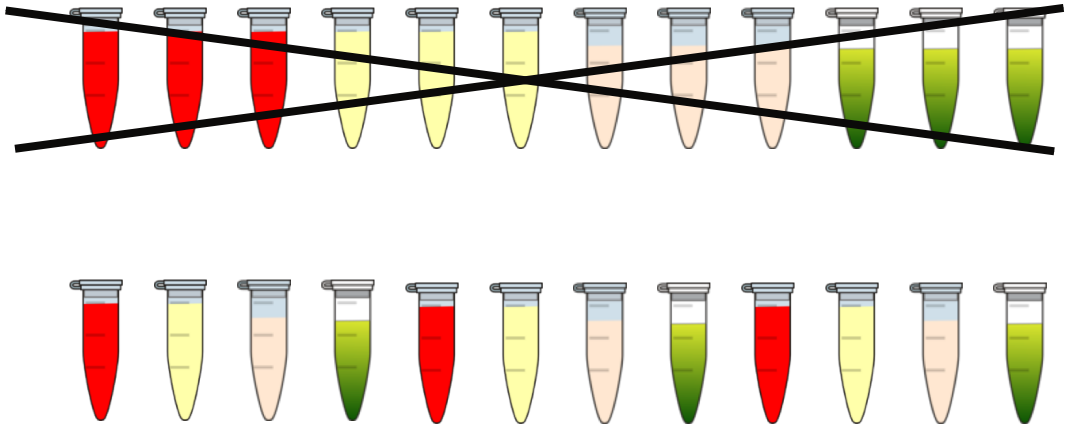
- In good experimental design, you compare two groups that **only differ in one factor**.
- Batch effect can occur when subsets of the replicates are handled separately at any stage of the process; handling group becomes in effect another factor. **Avoid processing all or most of one factor level together** if you can't do all the samples at once.



If batch effects are spread evenly over factor levels, they can be accounted for statistically

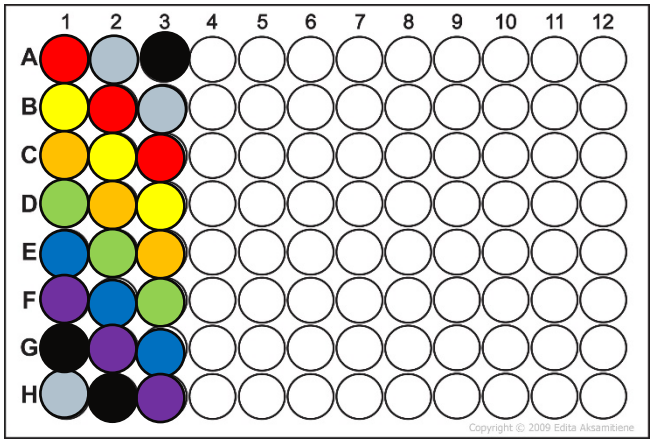
Beware systematic biases!

- Avoid systematic biases in the arrangement of replicates
  - **Don't** do all of one factor level first (circadian rhythms, experimenter experience, time-on-ice effects)
  - **Don't** send samples to the Keck Center in order



http://www.clker.com/clipart-eppendorf-tube-closed.html

Have one rep in each row and each column!



http://www.cellsignet.com/media/templ.html

A word on technical replication…

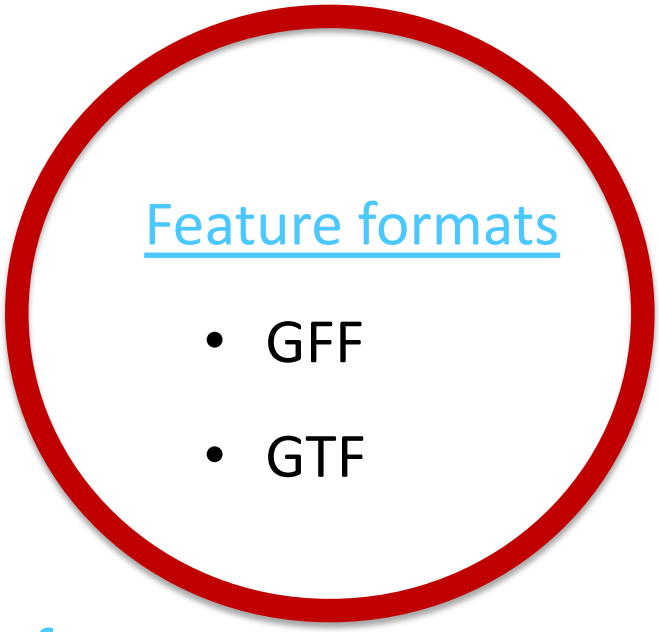
Technical replication is seen by many statisticians as a waste of time and resources because they do not substantially increase your power to detect differences… **biological replicates do**!

If you cannot increase the number of biological replicates but want to get extra certainty for the samples you do have, then you could do technical replicates if you have the $$ to spend.

General Outline

1. From RNA to sequencing data
2. Experimental and practical considerations
3. **Commonly encountered file formats**
4. Transcriptomic analysis methods and tools
    a. Transcriptome Assembly
    b. Differential Gene expression

File Formats



Sequence formats

- FASTA
- FASTQ

Feature formats

- GFF
- GTF

Alignment formats

- *SAM*
- *BAM*

Feature formats

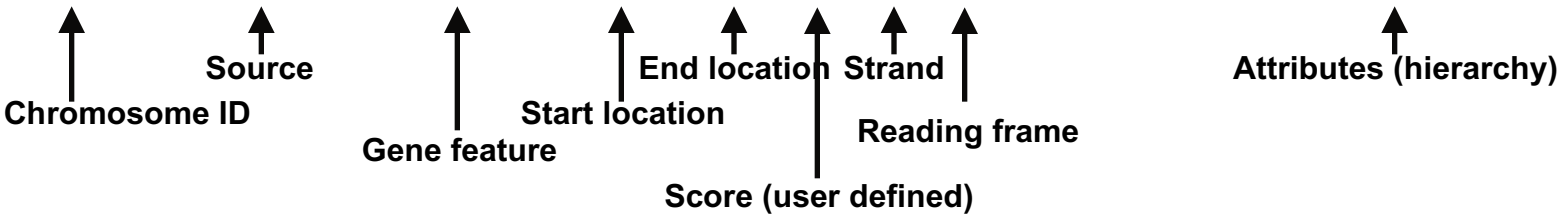
- Used for mapping features against a particular sequence or genome assembly
- May or may not include sequence data
- **The reference sequence must match** the names from a related file (possibly FASTA)
- **These are version (assembly)-dependent** - they are tied to a specific version (assembly/release) of a reference genome
- Not all reference genomes are the represented the same! E.g. human chromosome 1
  - UCSC – '**chr1**'
  - Ensembl – '**1**'
  - NCBI – '**NC_000001.11**'
- **Best practice:** get these from the same source as the reference

Feature formats : **GTF**

Gene transfer format

✧ Differences in representation of information make it distinct from GFF

```
AB000381 Twinscan  CDS          380   401   .   +   0   gene_id "001"; transcript_id "001.1";
AB000381 Twinscan  CDS          501   650   .   +   2   gene_id "001"; transcript_id "001.1";
AB000381 Twinscan  CDS          700   707   .   +   2   gene_id "001"; transcript_id "001.1";
AB000381 Twinscan  start_codon  380   382   .   +   0   gene_id "001"; transcript_id "001.1";
AB000381 Twinscan  stop_codon   708   710   .   +   0   gene_id "001"; transcript_id "001.1";
```

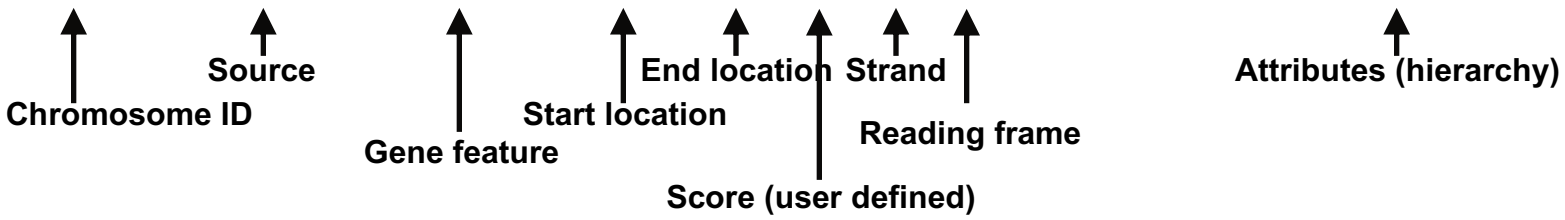


Feature formats : **GTF**

Gene transfer format

- Differences in representation of information make it distinct from GFF
- **Source of GTF is important** – Ensembl GTF is not quite the same as UCSC GTF

```
AB000381 Twinscan  CDS          380   401   .   +   0   gene_id "001"; transcript_id "001.1";
AB000381 Twinscan  CDS          501   650   .   +   2   gene_id "001"; transcript_id "001.1";
AB000381 Twinscan  CDS          700   707   .   +   2   gene_id "001"; transcript_id "001.1";
AB000381 Twinscan  start_codon  380   382   .   +   0   gene_id "001"; transcript_id "001.1";
AB000381 Twinscan  stop_codon   708   710   .   +   0   gene_id "001"; transcript_id "001.1";
```

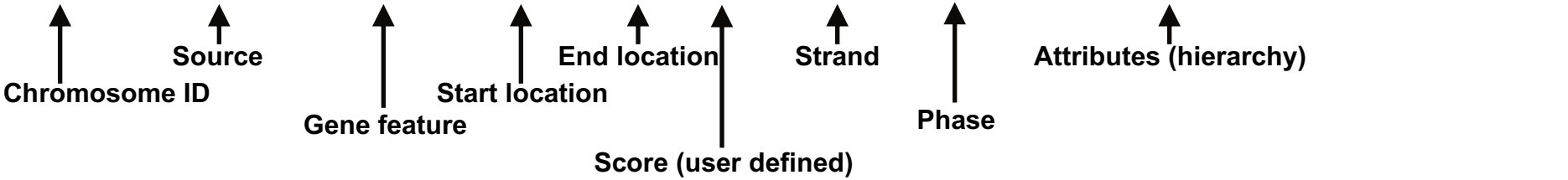


Feature formats : **GFF3**

General feature format (v3)

- Tab-delimited file to store genomic features, e.g. genomic intervals of genes and gene structure
- Meant to be unified replacement for GFF/GTF (includes specification)
- All but UCSC have started using this (UCSC prefers their own internal formats)

```
Chr1  amel_OGSv3.1  gene   204921  223005  .  +  .  ID=GB42165
Chr1  amel_OGSv3.1  mRNA   204921  223005  .  +  .  ID=GB42165-RA;Parent=GB42165
Chr1  amel_OGSv3.1  3'UTR  222859  223005  .  +  .  Parent=GB42165-RA
Chr1  amel_OGSv3.1  exon   204921  205070  .  +  .  Parent=GB42165-RA
Chr1  amel_OGSv3.1  exon   222772  223005  .  +  .  Parent=GB42165-RA
```



Feature formats: **GFF3 vs. GTF**

✧ GFF3 – General feature format

```
Chr1  amel_OGSv3.1   gene     204921  223005  .  +  .  ID=GB42165
Chr1  amel_OGSv3.1   mRNA     204921  223005  .  +  .  ID=GB42165-RA;Parent=GB42165
Chr1  amel_OGSv3.1   3'UTR    222859  223005  .  +  .  Parent=GB42165-RA
Chr1  amel_OGSv3.1   exon     204921  205070  .  +  .  Parent=GB42165-RA
Chr1  amel_OGSv3.1   exon     222772  223005  .  +  .  Parent=GB42165-RA
```

✧ GTF – Gene transfer format

```
AB000381 Twinscan  CDS          380   401   .  +  0  gene_id "001"; transcript_id "001.1";
AB000381 Twinscan  CDS          501   650   .  +  2  gene_id "001"; transcript_id "001.1";
AB000381 Twinscan  CDS          700   707   .  +  2  gene_id "001"; transcript_id "001.1";
AB000381 Twinscan  start_codon  380   382   .  +  0  gene_id "001"; transcript_id "001.1";
AB000381 Twinscan  stop_codon   708   710   .  +  0  gene_id "001"; transcript_id "001.1";
```

Always check which of the two formats is accepted by your application of choice, sometimes they cannot be swapped

General Outline

1. From RNA to sequencing data
2. Experimental and practical considerations
3. Commonly encountered file formats
4. **Transcriptomic analysis methods and tools**
    a. Transcriptome assembly
    b. Differential gene expression

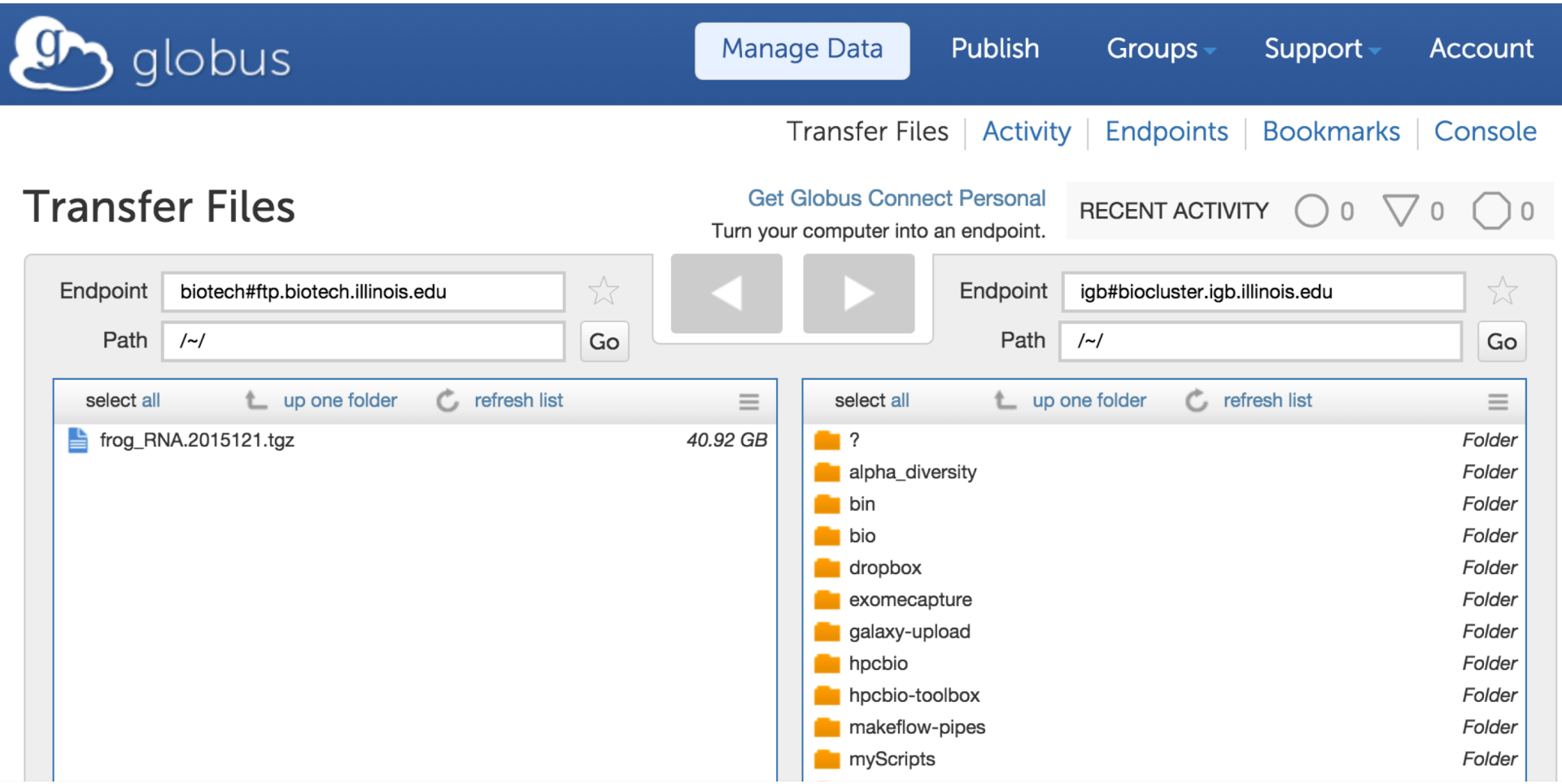
Detailed Outline

4. **Transcriptomic analysis methods and tools**
    a. **Steps common to both assembly and differential gene expression**
        - ✧ **Download data**
        - ✧ **Quality check**
        - ✧ **Data alignment**
    b. Assembly
    c. Differential Gene Expression
    d. Choosing a method, the considerations…
    e. Final thoughts and observations

Obtain sequence data

1. If you are using the R.J.C. Biotechnology Center and the Biocluster
   - ✧ Globus is most direct route
     - ✧ CNRG instructions
2. Download data to a computer and upload to Biocluster using an SFTP client
   - ✧ Cyberduck, WinSCP…
3. Can also use linux commands such as:
   - ✧ scp, rsync, wget, …

Globus

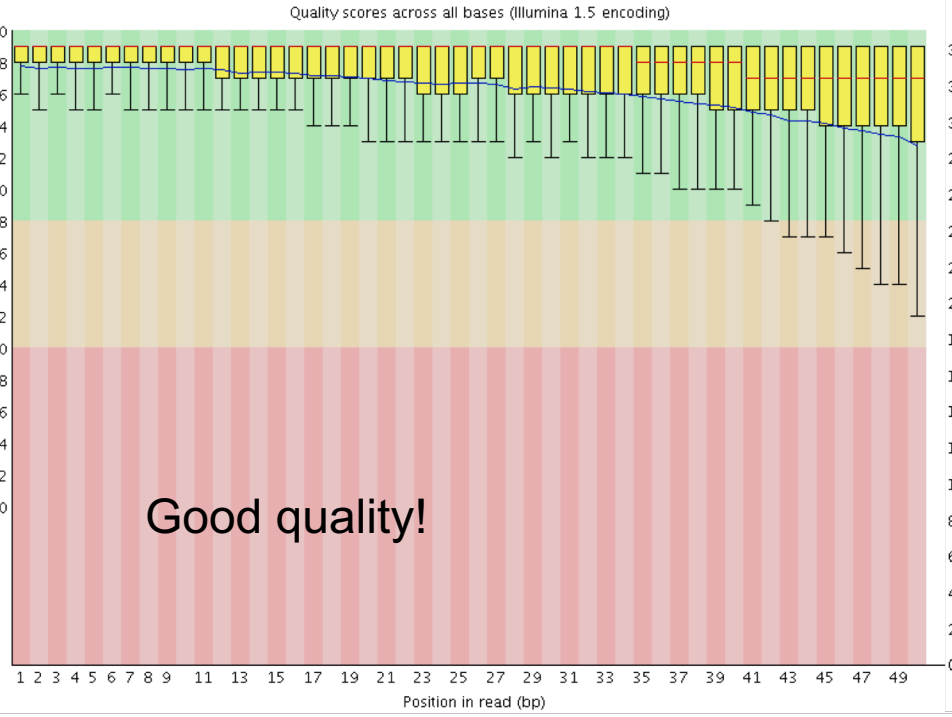

globus
Manage Data
Publish
Groups
Support
Account
Transfer Files
Activity
Endpoints
Bookmarks
Console
Transfer Files
Get Globus Connect Personal
Turn your computer into an endpoint.
RECENT ACTIVITY 0 0 0
Endpoint
biotech#ftp.biotech.illinois.edu
Path
/~/
Go
Endpoint
igb#biocluster.igb.illinois.edu
Path
/~/
Go
select all
up one folder
refresh list
frog_RNA.2015121.tgz
40.92 GB
select all
up one folder
refresh list
? Folder
alpha_diversity Folder
bin Folder
bio Folder
dropbox Folder
exomecapture Folder
galaxy-upload Folder
hpcbio Folder
hpcbio-toolbox Folder
makeflow-pipes Folder
myScripts Folder

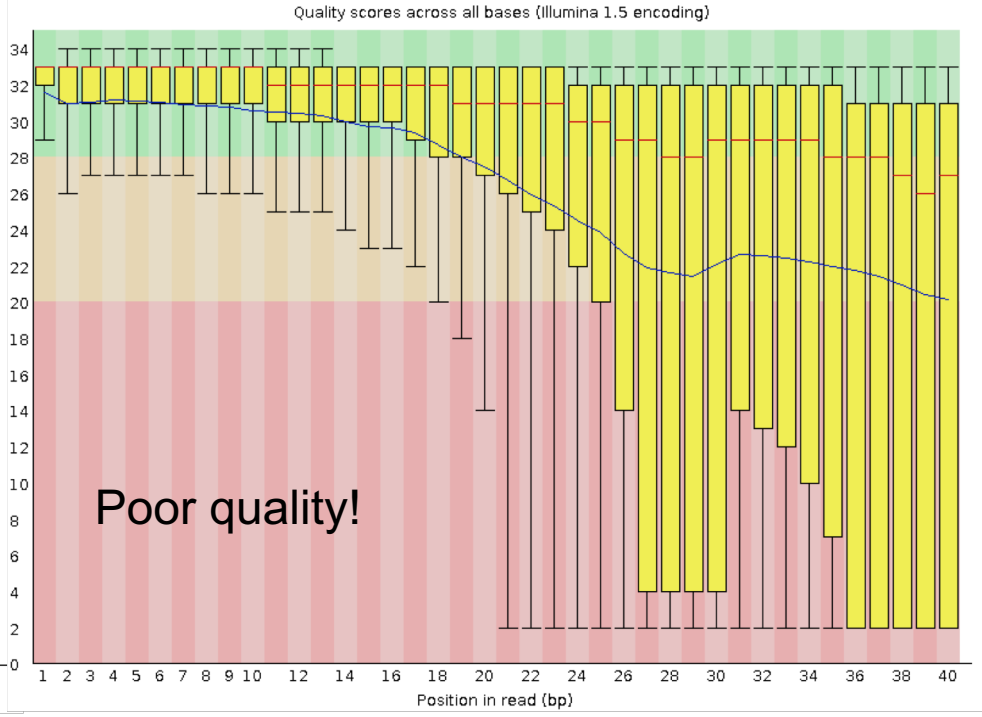
So how can we check the quality of our raw sequences?

Software called **FASTQC**

- Name is a play on FASTQ format and QC (Quality Control)
- Checks quality by several metrics, and creates a visual report

FASTQC: Quality Scores





FASTQC cont…

Additional metrics

- Presence of, and abundance of contaminating sequences
- Average read length
- GC content
- And more!

Assumes that your data is:

- WGS (i.e. evenish sampling of the whole genome)
- Derived from DNA
- Derived from one species

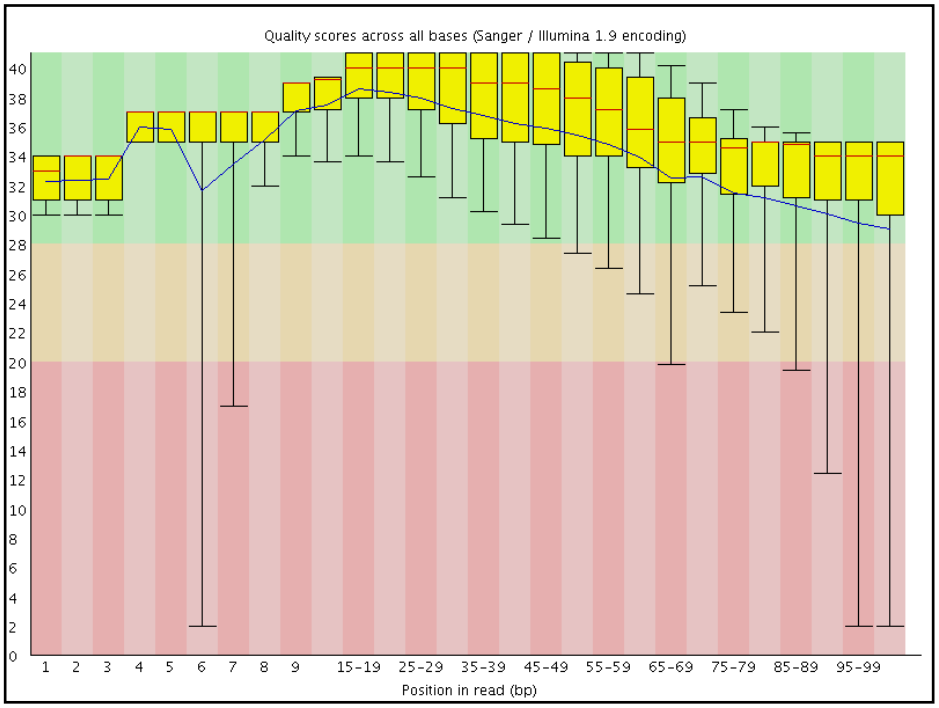
So keep this in mind when interpreting results

What do I do when FastQC calls my data poor?

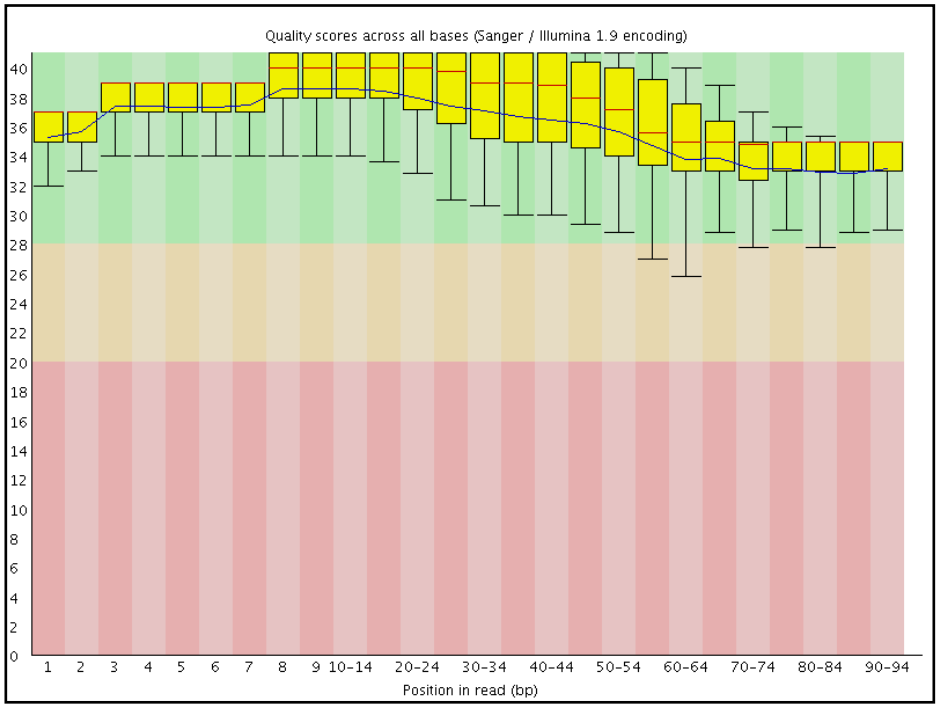
- Poor quality at the ends can be remedied
- Left-over adapter sequences in the reads can be removed
  - Always trim adapters as a matter of routine
- We need to amend these issues so we get the best possible alignment
- After trimming, it is best to rerun the data through FastQC to check the resulting data

Quality Before & After

Before quality trimming



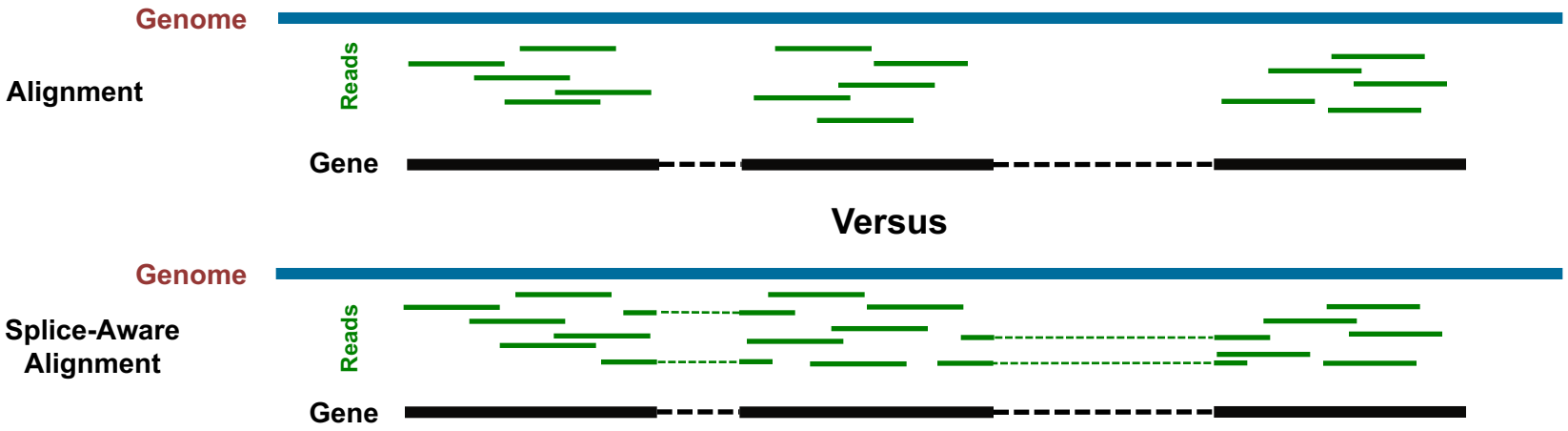
After quality trimming



Data Alignment

We need to align the sequence data to our genome of interest

- If aligning RNASeq data to the genome, almost always pick a splice-aware aligner



Data Alignment

Splice-aware aligners: recommended for most applications

STAR, HiSat2, Novoalign (not free), MapSplice2, GSNAP, ContextMap2 …

Non-splice aware aligners: ideal for bacterial genomes

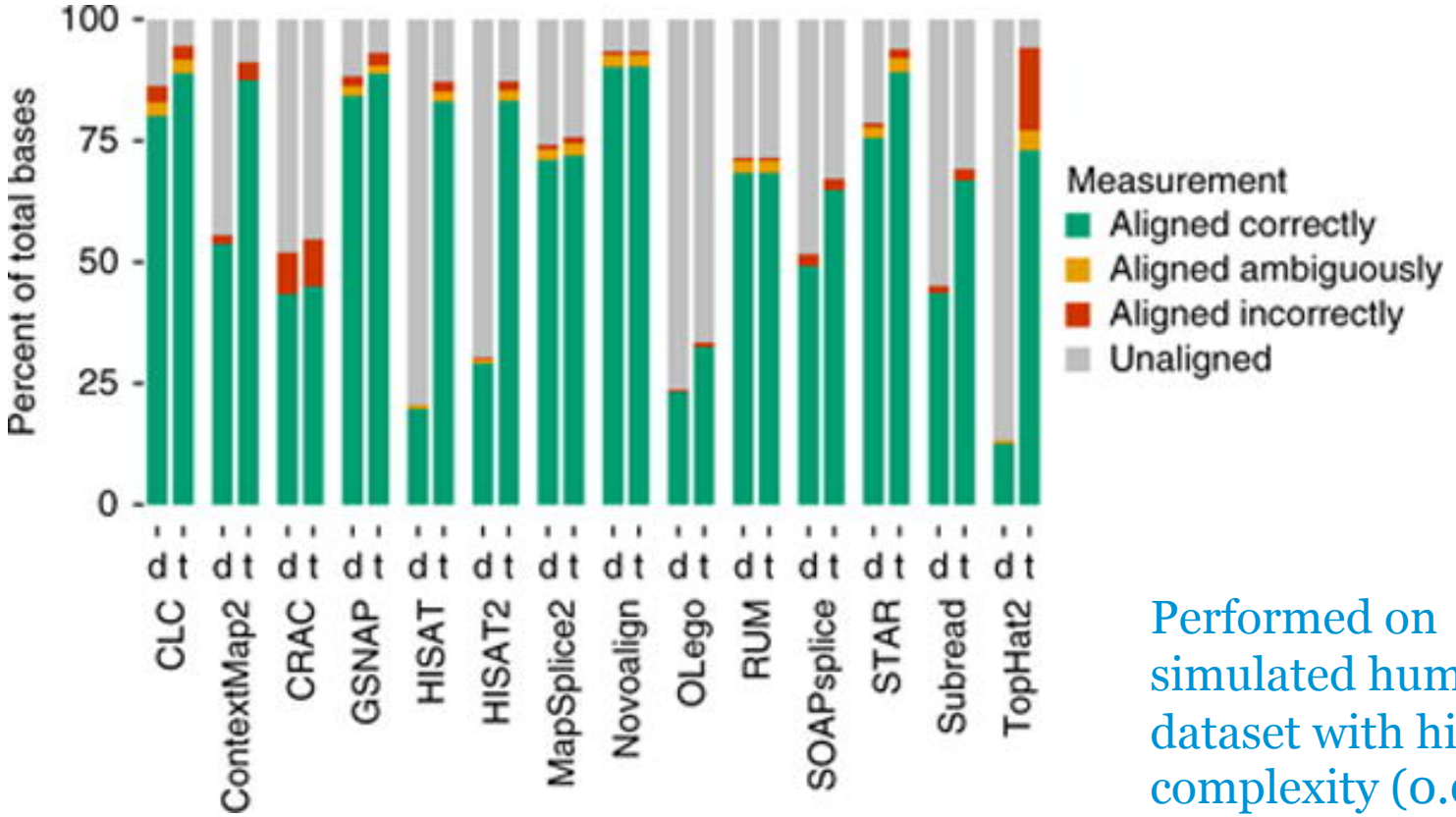
BWA, Novoalign (not free), Bowtie2, HiSat2

Data Alignment

Other considerations when choosing an aligner:

- How does it deal with reads that map to **multiple locations**?
- How does it deal with **paired-end versus single-end** data?
- How many **mismatches** will it allow between the genome and the reads?
- What **assumptions** does it make about my genome, and can I change these assumptions?

Always check the default settings of any software you use!!!



Performed on simulated human dataset with high complexity (0.03 substitution, 0.005 indel, 0.02 error)

Baruzzo et. al, 2017, doi: 10.1038/nmeth.4106

Alignment Visualization

IGV is the visualization tool used for this snapshot

Detailed Outline

Transcriptomic analysis methods and tools

a. Steps common to both assembly and differential gene expression
   - ✧ Download data
   - ✧ Quality check
   - ✧ Data alignment

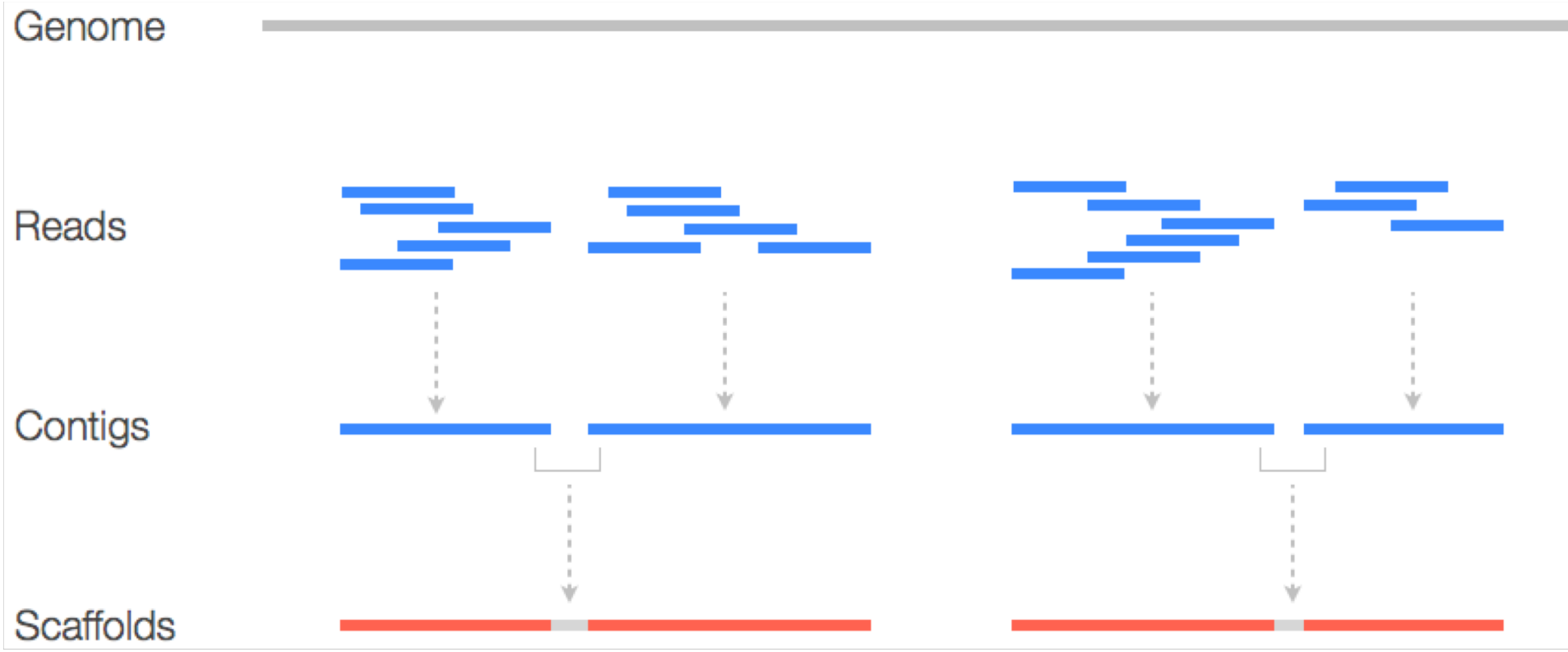
b. Assembly

c. Differential Gene Expression

d. Choosing a method, the considerations…

e. Final thoughts and observations

Transcriptome Assembly Overview



End goal: A complete and comprehensive set of gene-models

Transcriptome Assembly Overview

Two main types of assemblies

- Reference-based assembly
- *de novo* assembly

Transcriptome Assembly

Reference-based assembly

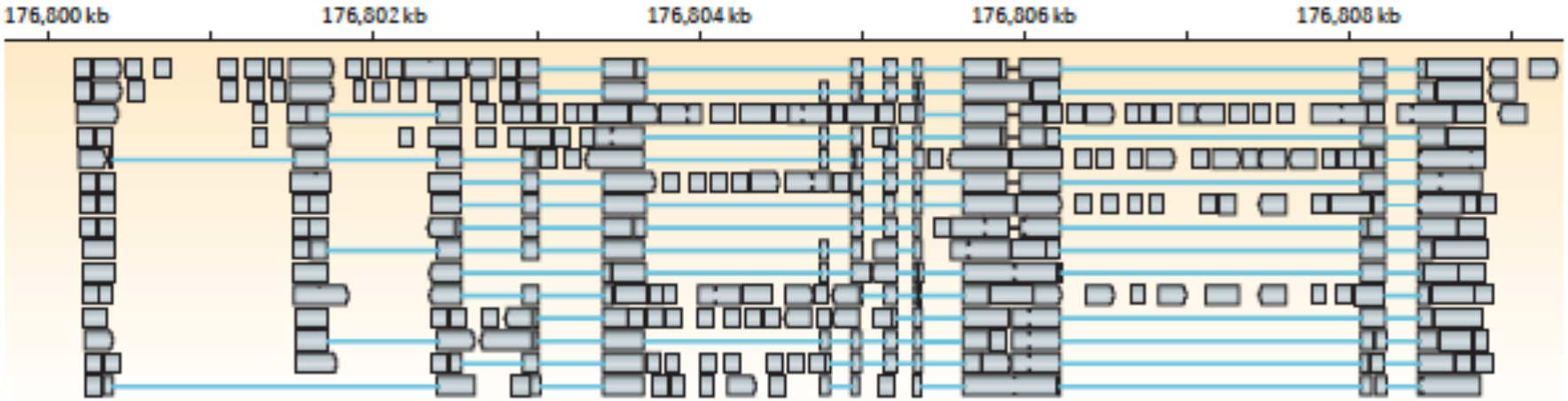
Used when the genome reference sequence is known, and:

- Transcriptome data is not available
- Transcriptome data is available but not good enough,
  - i.e. missing isoforms of genes, or unknown non-coding regions
- The existing transcriptome information is for a different tissue type
- Stringtie, and Scripture are some reference-based transcriptome assemblers

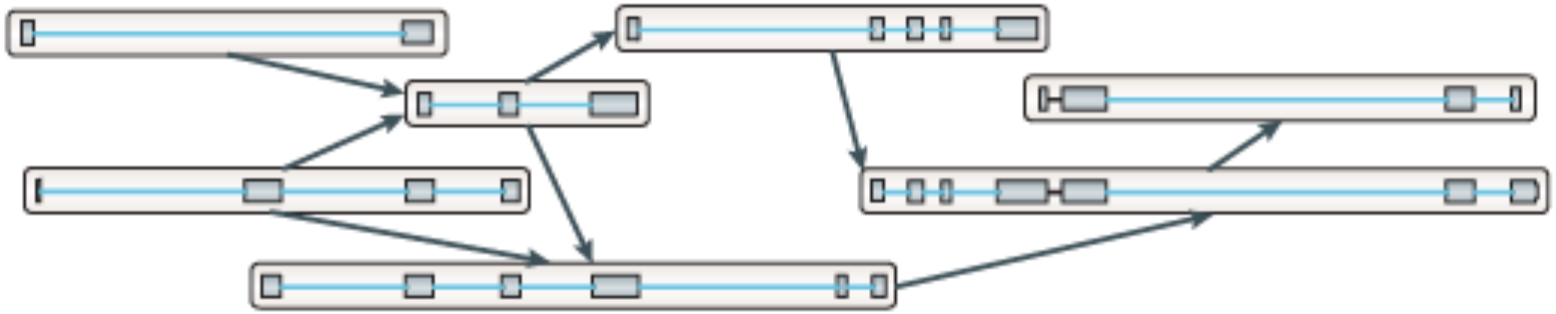
Transcriptome Assembly

Reference-based assembly

a. Splice align reads to genome



b. Build graph representing alternative splicing events

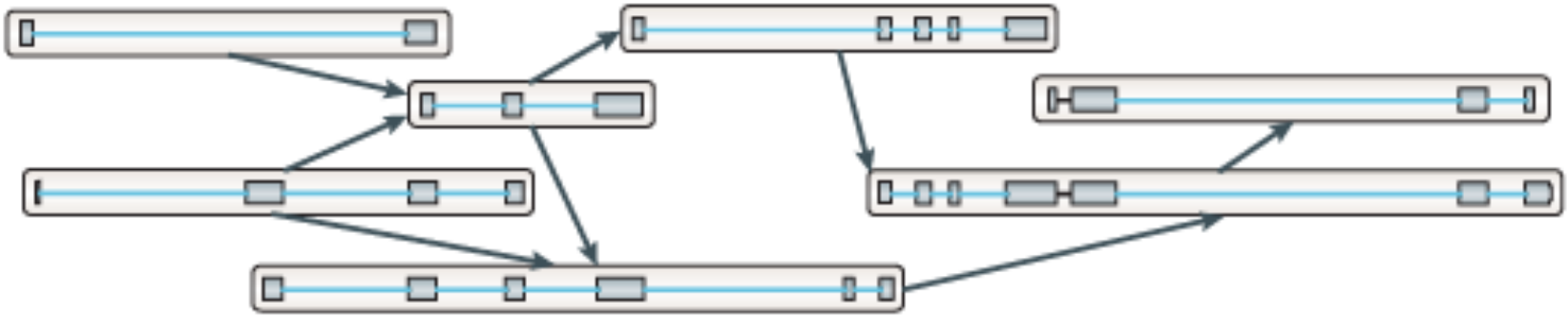


Martin J.A. and Wang Z., Nat. Rev. Genet. (2011) 12:671–682

Transcriptome Assembly

Reference-based assembly

b. Build graph representing alternative splicing events

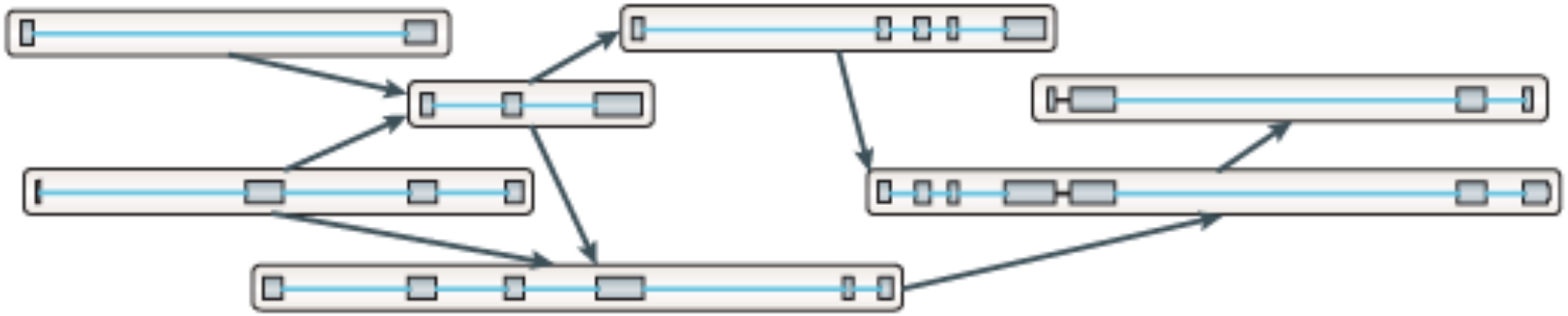


Martin J.A. and Wang Z., Nat. Rev. Genet. (2011) 12:671–682

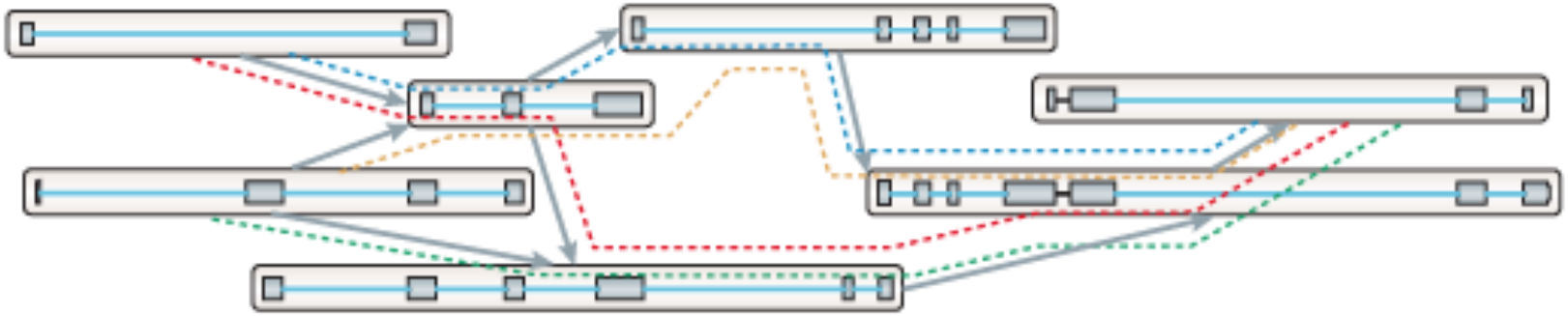
Transcriptome Assembly

Reference-based assembly

b. Build graph representing alternative splicing events



c. Traverse the graph to assemble variants

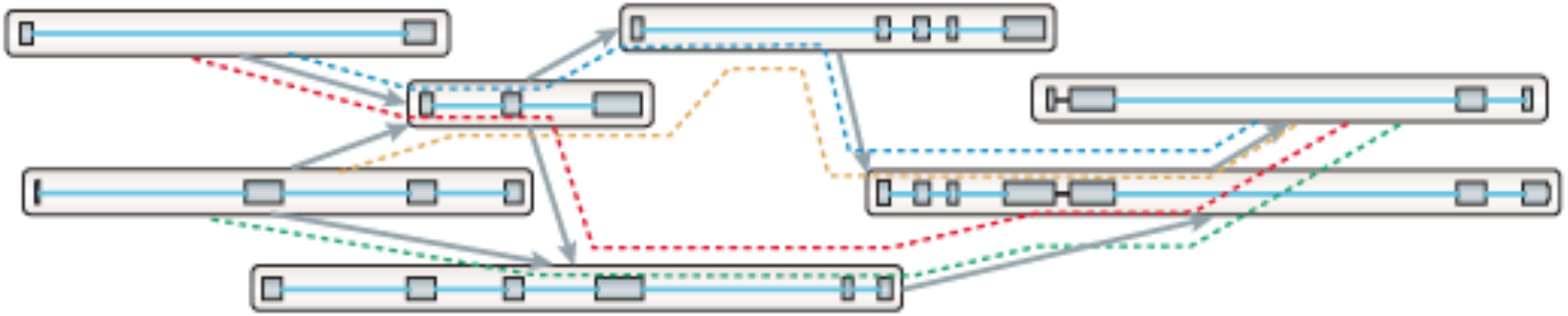


Martin J.A. and Wang Z., Nat. Rev. Genet. (2011) 12:671–682

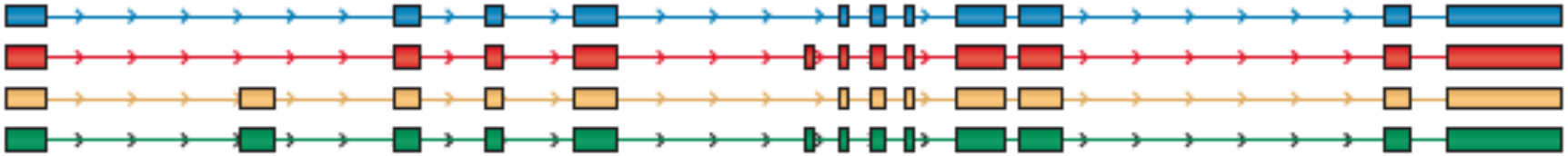
Transcriptome Assembly

Reference-based assembly

c. Traverse the graph to assemble variants



d. Assembled isoforms



Martin J.A. and Wang Z., Nat. Rev. Genet. (2011) 12:671–682

Transcriptome Assembly

De novo assembly

Used when very little information is available for the genome

- ✧ Often the first step in putting together information about an unknown genome
- ✧ Amount of data needed for a good *de novo* assembly is higher than what is needed for a reference-based assembly
- ✧ Can be used for genome annotation, once the genome is assembled
- ✧ Trinity, SPAdes, and TransABySS, are examples of well-regarded transcriptome assemblers

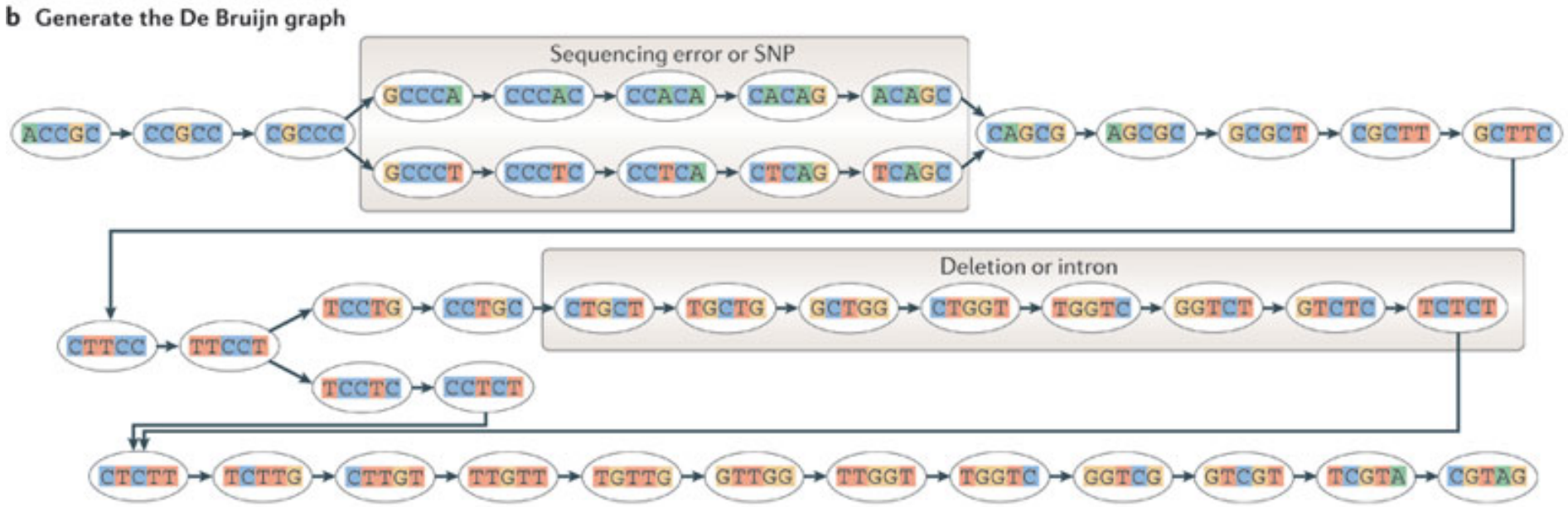
Transcriptome Assembly

De novo assembly (De Bruijn graph construction)

Martin J.A. and Wang Z., Nat. Rev. Genet. (2011) 12:671–682

Transcriptome Assembly

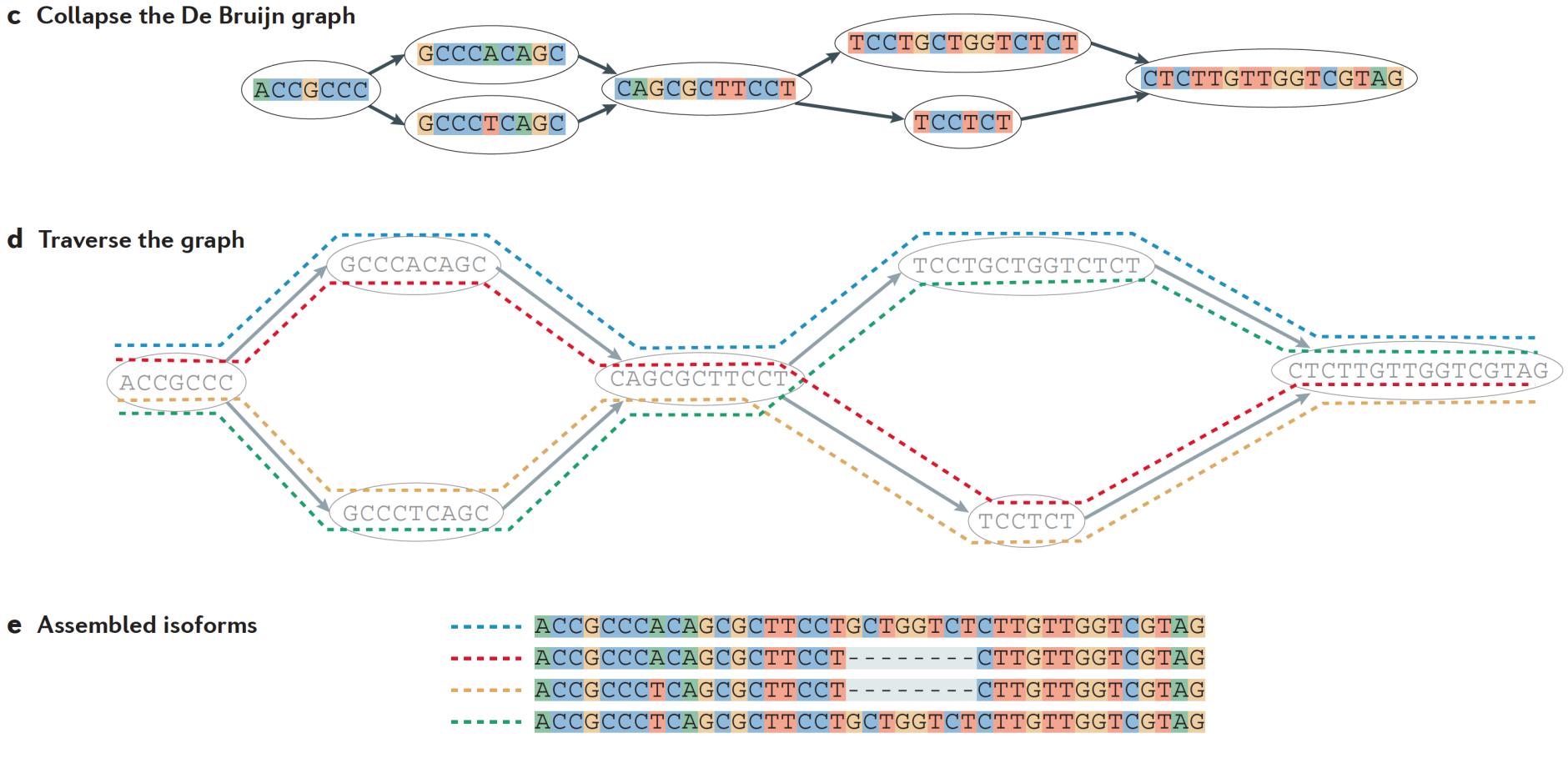
De novo assembly (De Bruijn graph construction)



Martin J.A. and Wang Z., Nat. Rev. Genet. (2011) 12:671–682

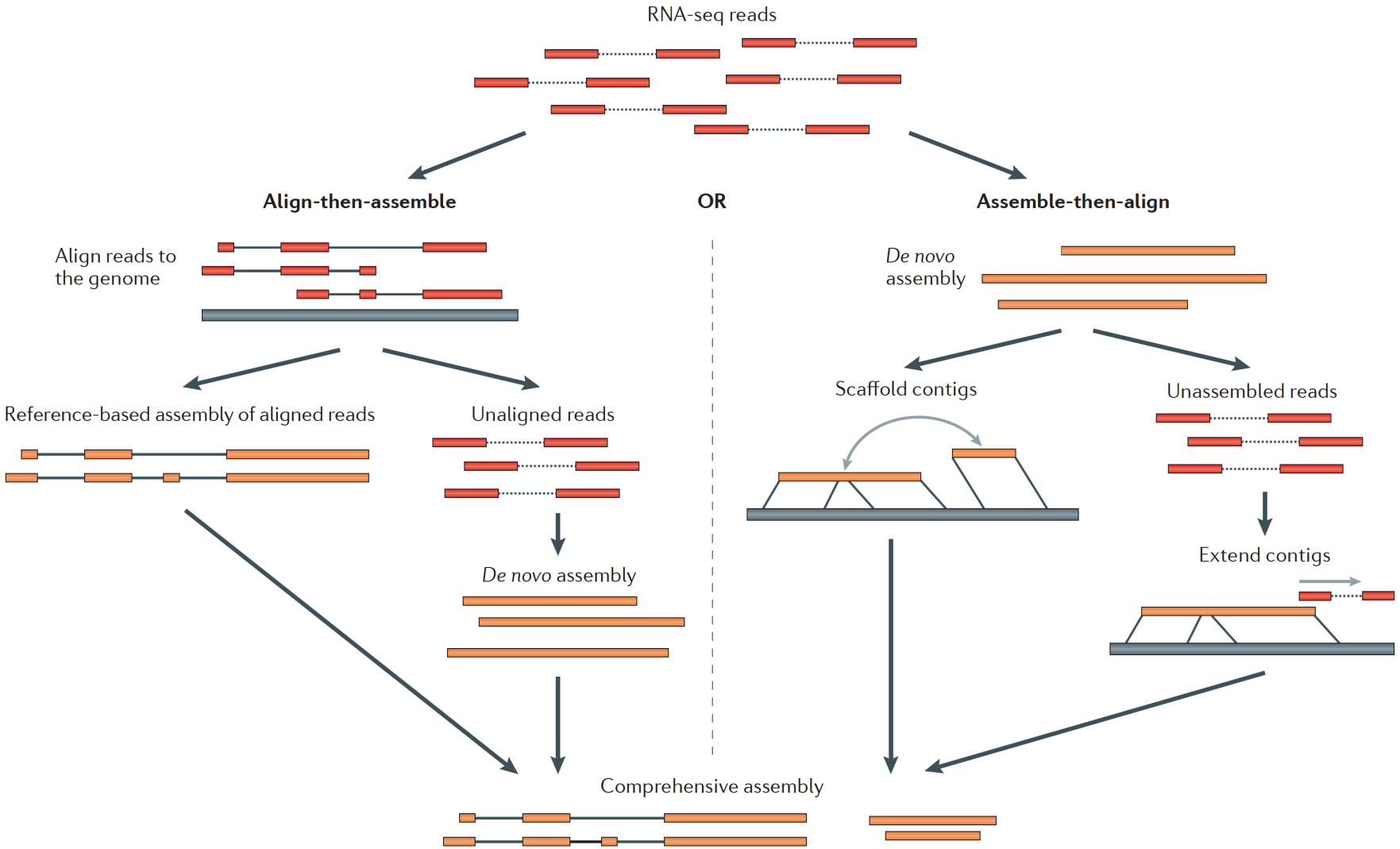
Transcriptome Assembly

De novo assembly (De Bruijn graph construction)



Martin J.A. and Wang Z., Nat. Rev. Genet. (2011) 12:671–682

Combined Transcriptome Assembly



Martin J.A. and Wang Z., Nat. Rev. Genet. (2011) 12:671–682

How good is my assembly?

- Are all the genes I expected in the assembly?
- Do I have complete genes?
- Are the contigs assembled correctly?
- How does it look compared to a close reference?

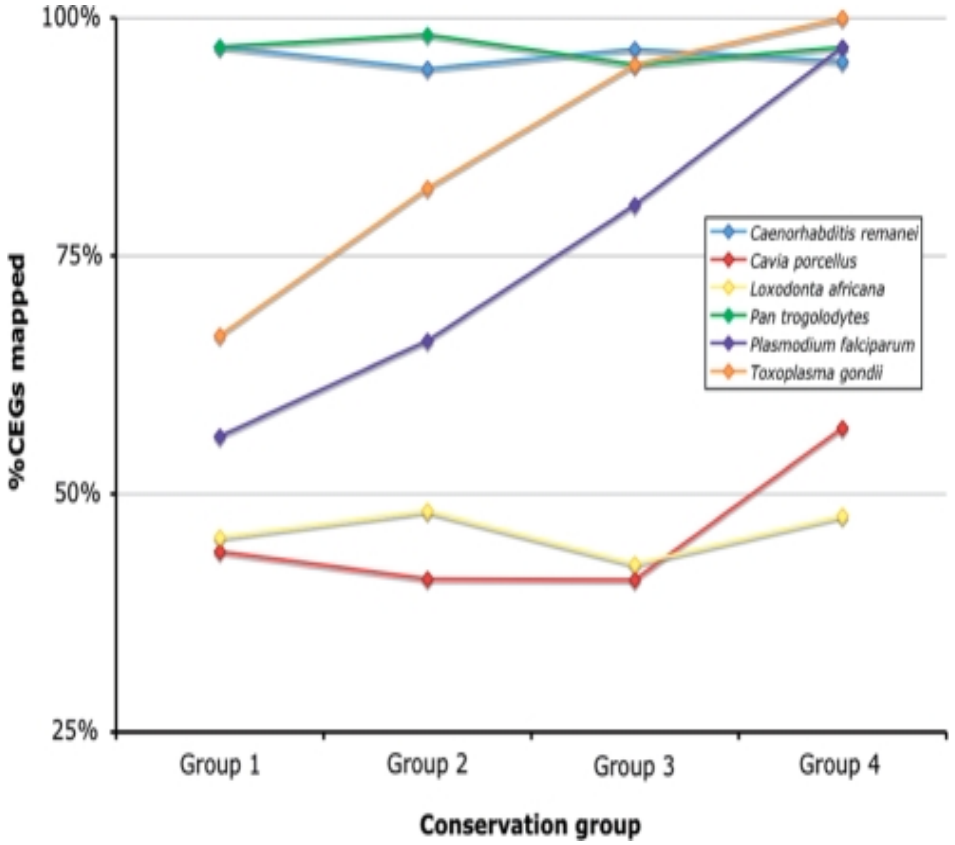
Tools for Evaluating Assembly: *using the information you have*

- TransRate – evaluates assembly using reads, paired end information, reference genome, protein data, etc.
    - Can generate a 'cleaned-up' or optimized assembly based on metrics
- DETONATE – evaluates assembly based on read mapping and/or reference information

Tools for Evaluating Assembly: *conserved gene sets*

BUSCO: From Evgeny Zdobnov's group, University of Geneva

Coverage is indicative of quality and completeness of assembly



Detailed Outline

Transcriptomic analysis methods and tools

a. Steps common to both assembly and differential gene expression
   - ✧ Download data
   - ✧ Quality check
   - ✧ Data alignment

b. Assembly

c. Differential Gene Expression

d. Choosing a method, the considerations…

e. Final thoughts and observations

Differential Gene Expression Steps

1. Quality control steps

2. Align reads to a reference genome with splice aware software (unless bacterial)

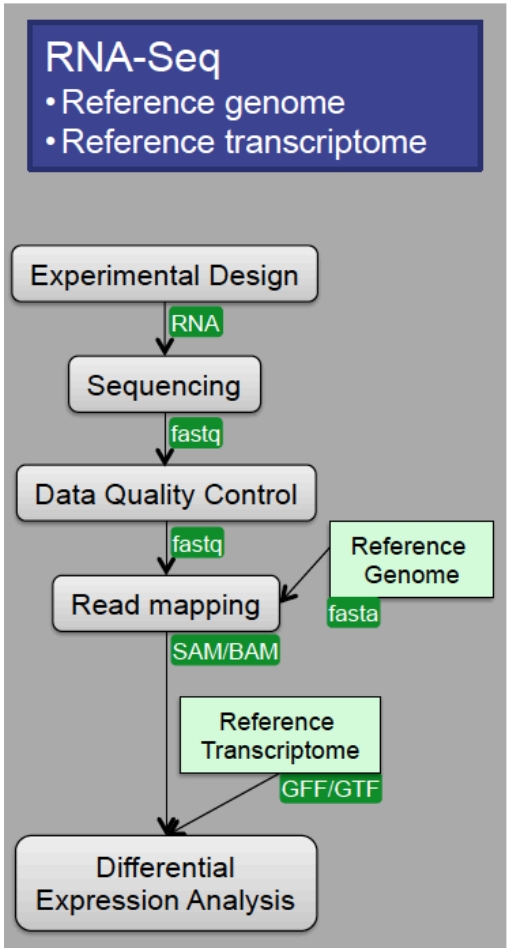
3a. Use a gene counting software to obtain the number of read counts per known gene.

3b. Alternatively, use a transcript counting software

4. Run statistical software on the gene/transcript counts

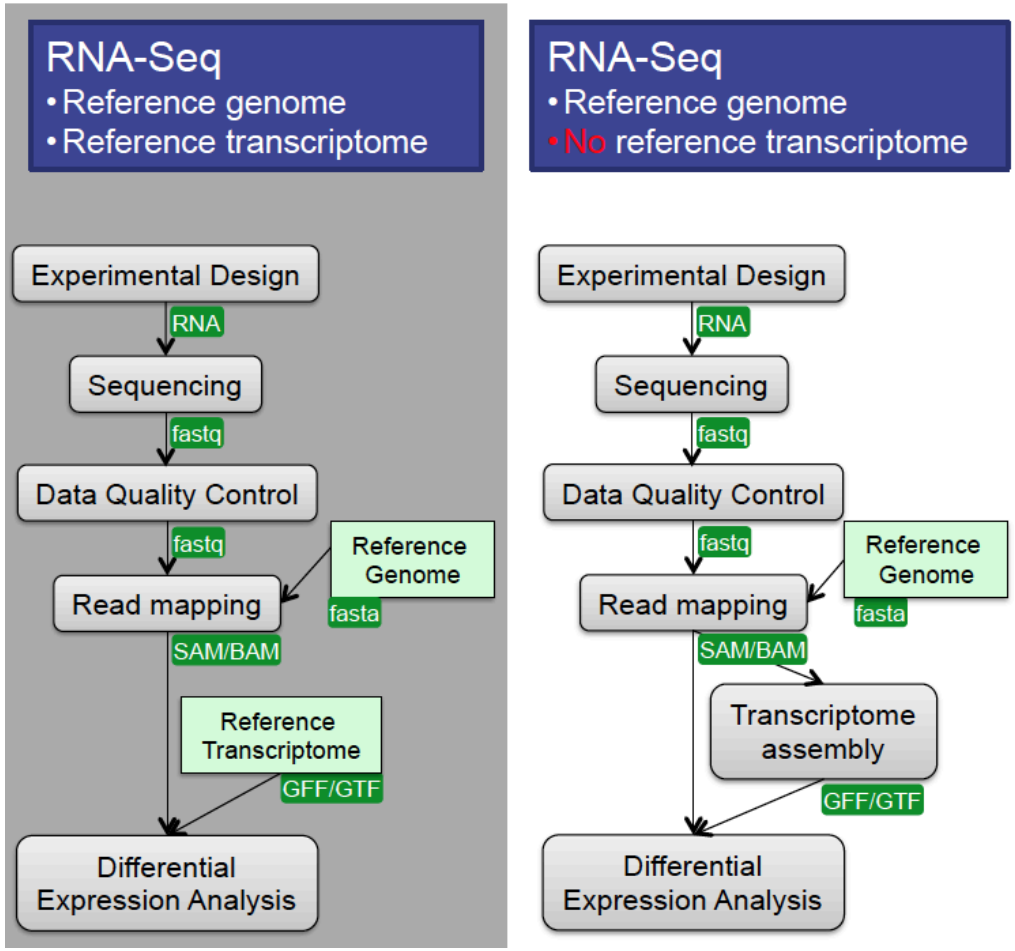
Differential Gene Expression

different options



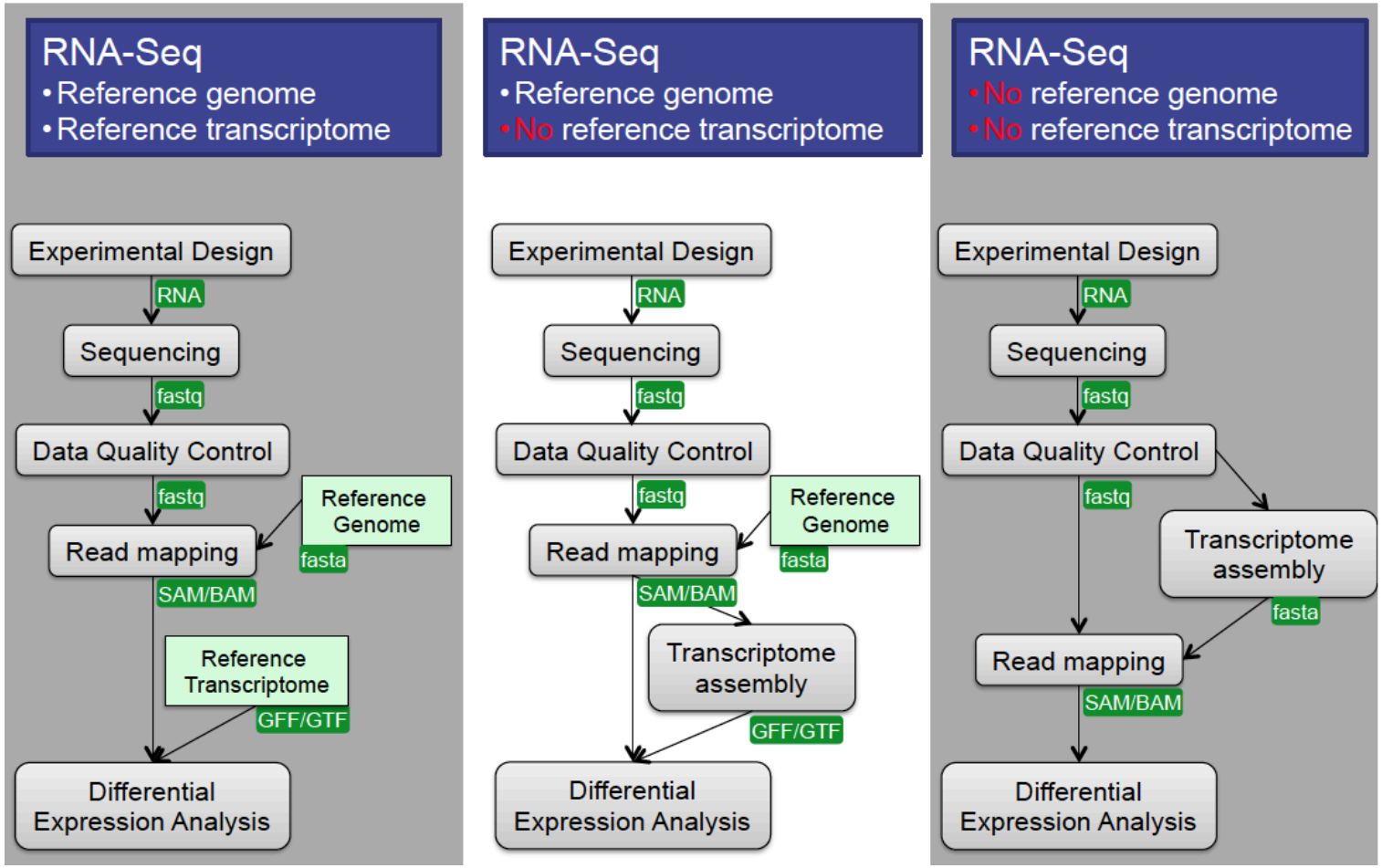
Differential Gene Expression

different options



Differential Gene Expression

different options



Gene Counting

When selecting software consider whether you want to obtain raw read counts or normalized read counts? This will depend on the statistical analysis you wish to perform downstream

- htseq & feature-counts return raw read counts
  - Required for R programs like DESeq & EdgeR
- StringTie returns FPKM normalized counts for each gene

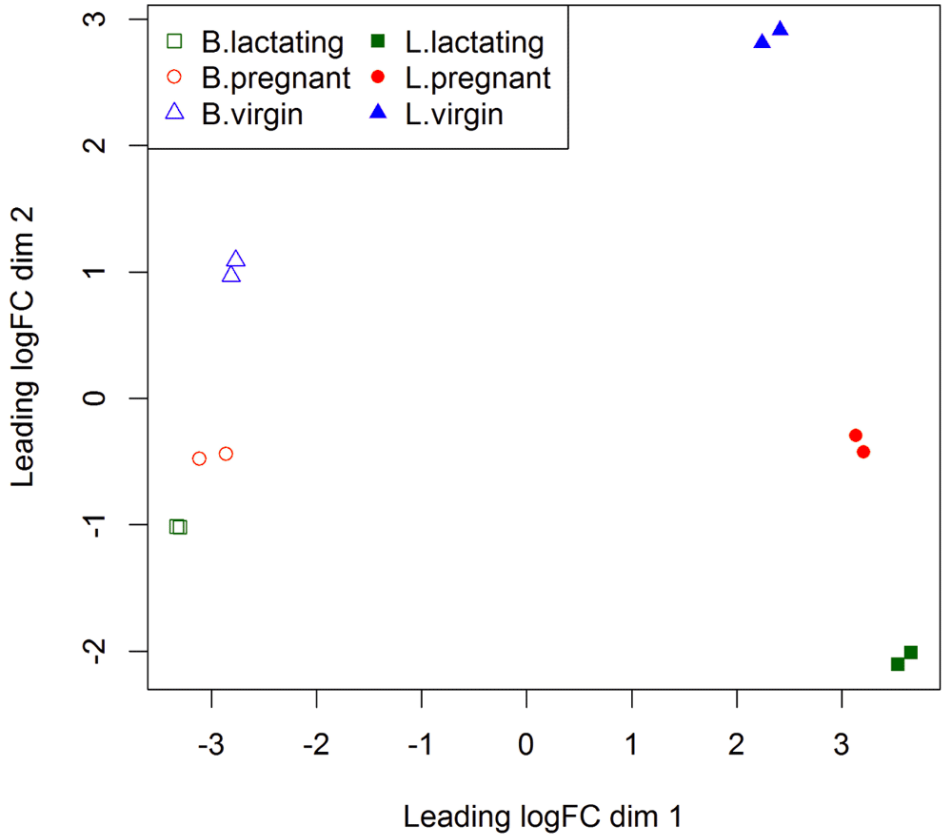
DGE Statistical Analyses

1. The first step is proper normalization of the data
   - ✧ Often the statistical package you use will have a normalization method that it prefers and uses exclusively (e.g. Voom, FPKM, TMM (used by EdgeR))
2. Is your experiment a pairwise comparison?
   - ✧ Ballgown, EdgeR, DESeq
3. Is it a more complex design?
   - ✧EdgeR, DESeq, other R/Bioconductor packages

Statistical Results

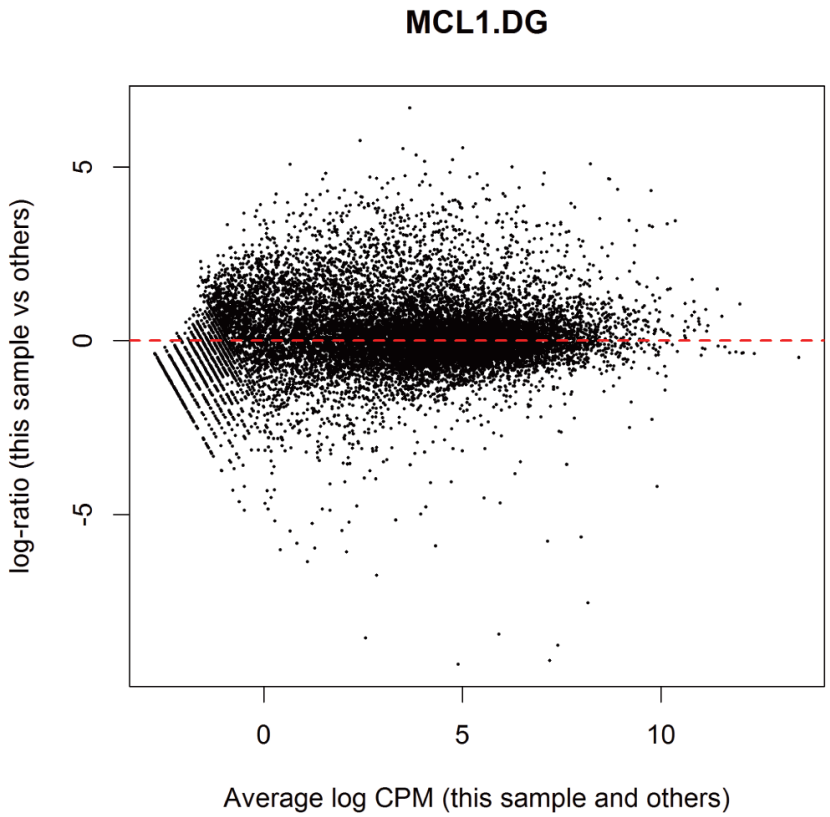
- A list of significantly differentially expressed genes
- Heatmaps, Venn Diagrams, and more
- Annotation
- WGCNA
- ... and more!

EdgeR: MDS Plot



https://f1000research.com/articles/5-1438 (doi: 10.12688/f1000research.8987.2)

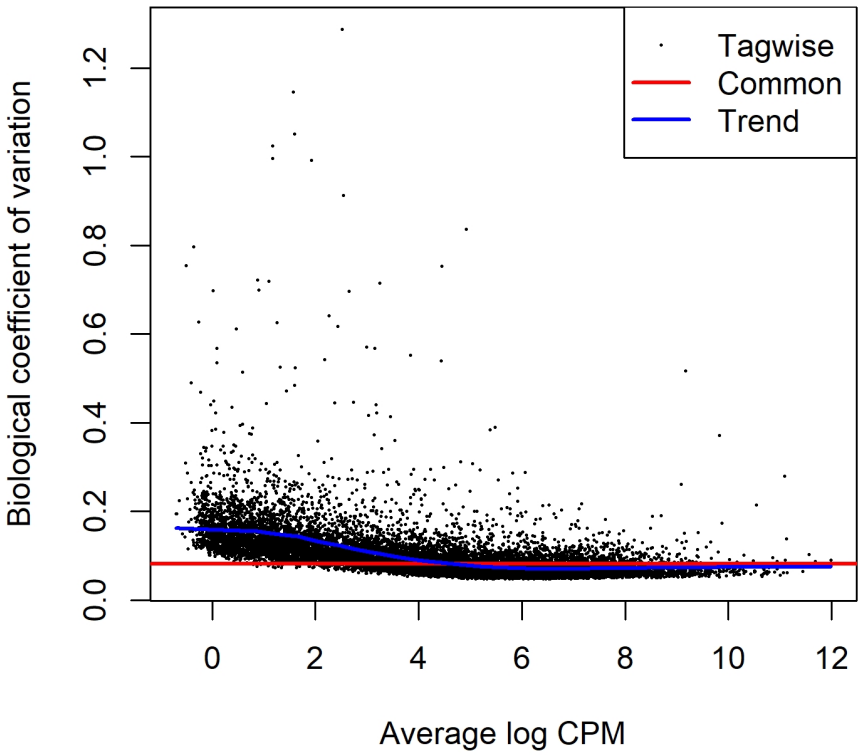
EdgeR: MD Plot



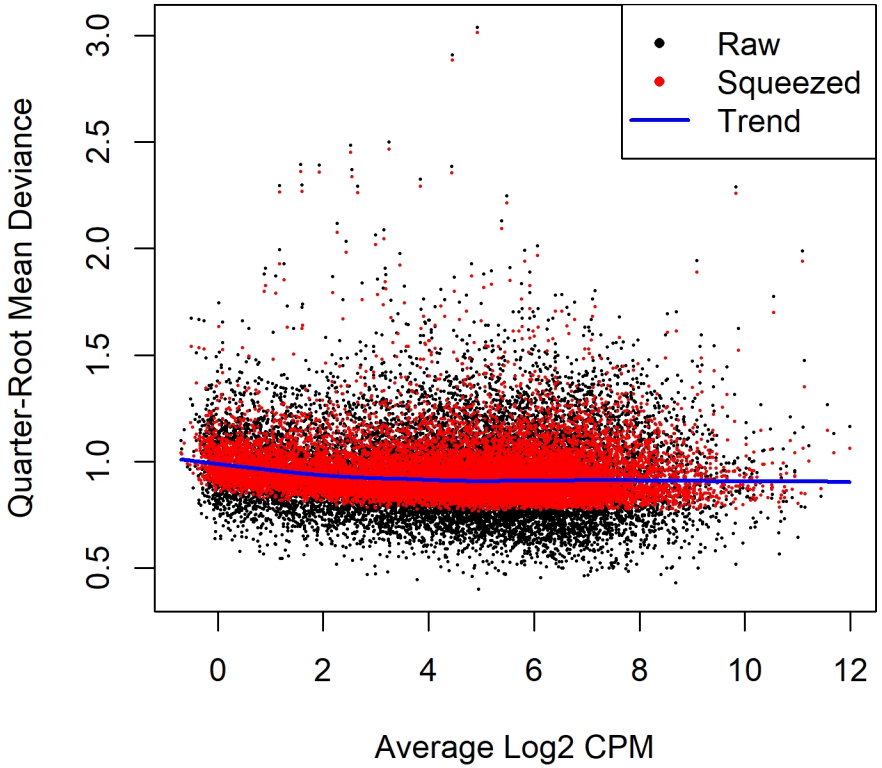
https://f1000research.com/articles/5-1438 (doi: 10.12688/f1000research.8987.2)

EdgeR Results: Dispersion Estimation

BCV Plot



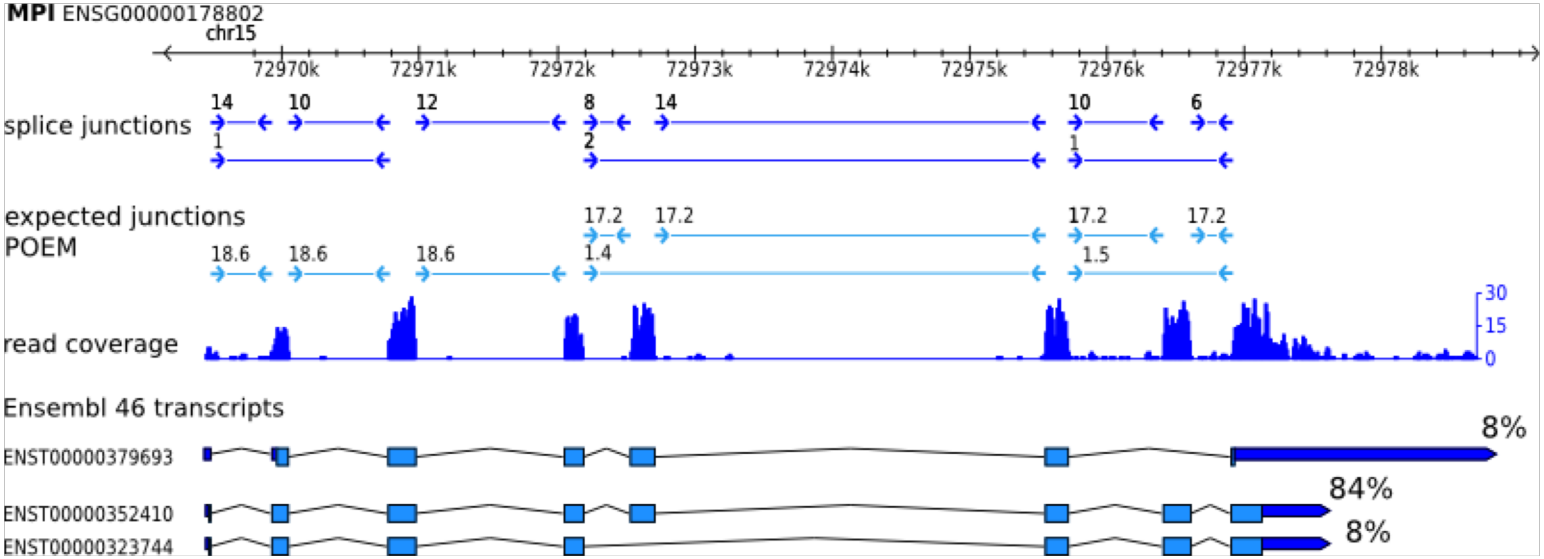
QL Plot



https://f1000research.com/articles/5-1438

Transcript Counting Methods

- Can't use STAR/featureCounts at transcript level
- If you try, many more reads will be ambiguous and will be discarded

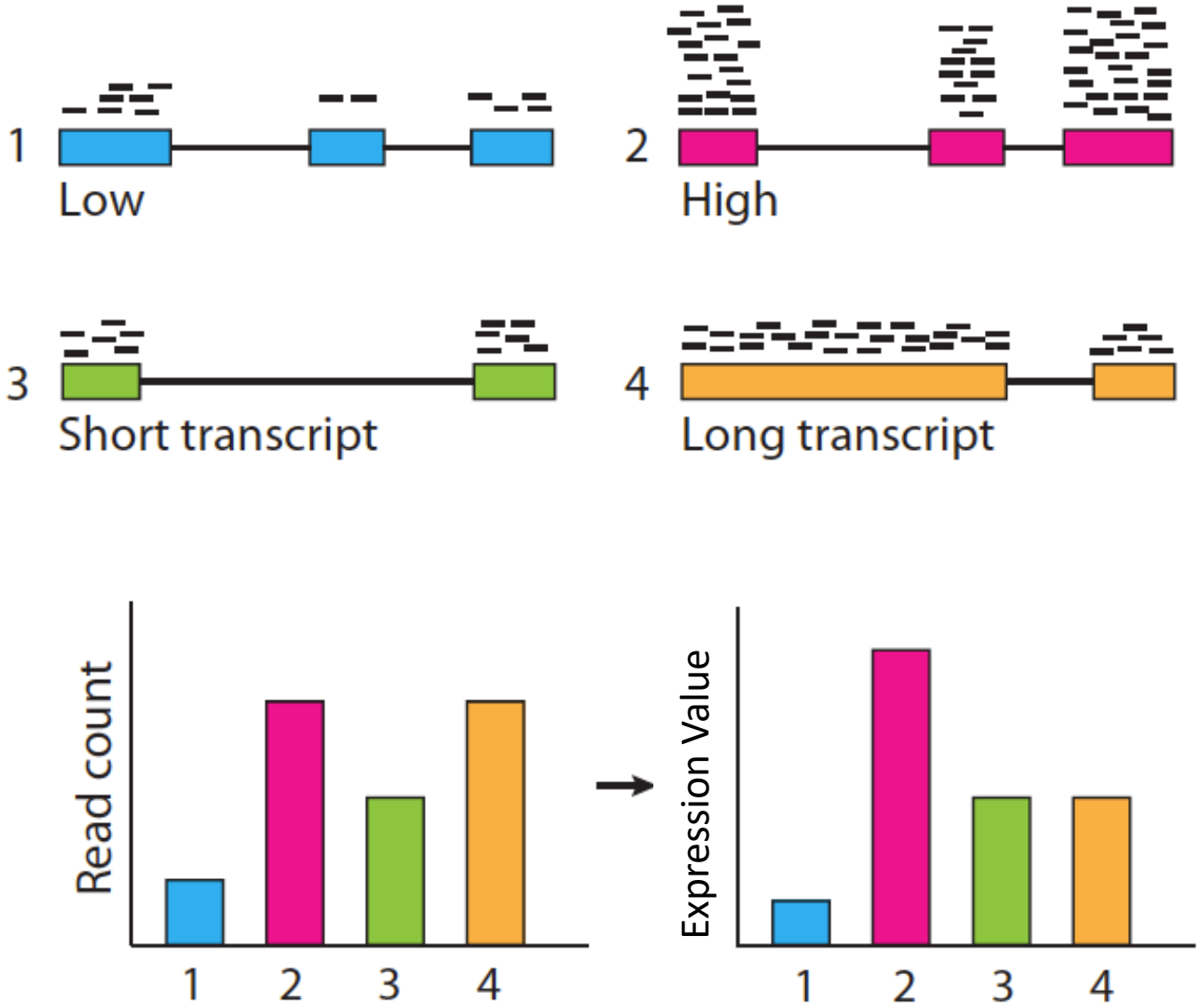


http://www.cs.cmu.edu/~maschulz/projects.html

Problems with STAR/featureCounts at gene level:

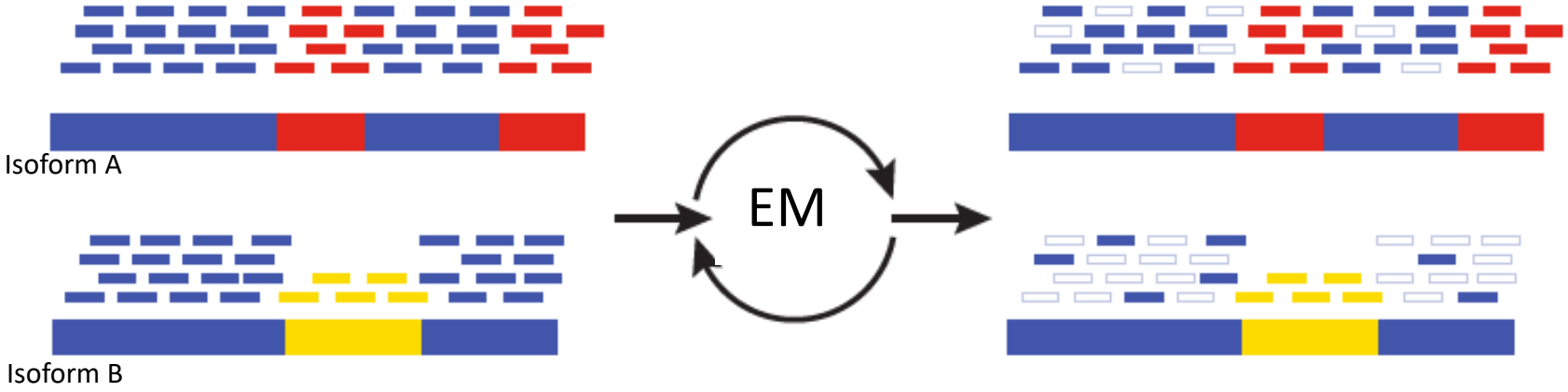
1. Multimapping reads not used, leading to underestimation of gene abundances, particularly for genes with more shared sequence
2. A small percentage of genes may not ever be quantifiable using this method.
3. Genes that change relative isoform usage can have erroneous results due to changes in isoform length

Calculating expression of genes and transcripts



Slide courtesy of Cole Trapnell

Solution: Expectation Maximization algorithms



Blue = multiply-mapped reads
Red, Yellow = uniquely-mapped reads

Use Expectation Maximization (EM) to find the most likely assignment of reads to transcripts.

Performed by:

- Cufflinks and Cuffdiff (Tuxedo)
- RSEM
- eXpress
- Salmon/kallisto

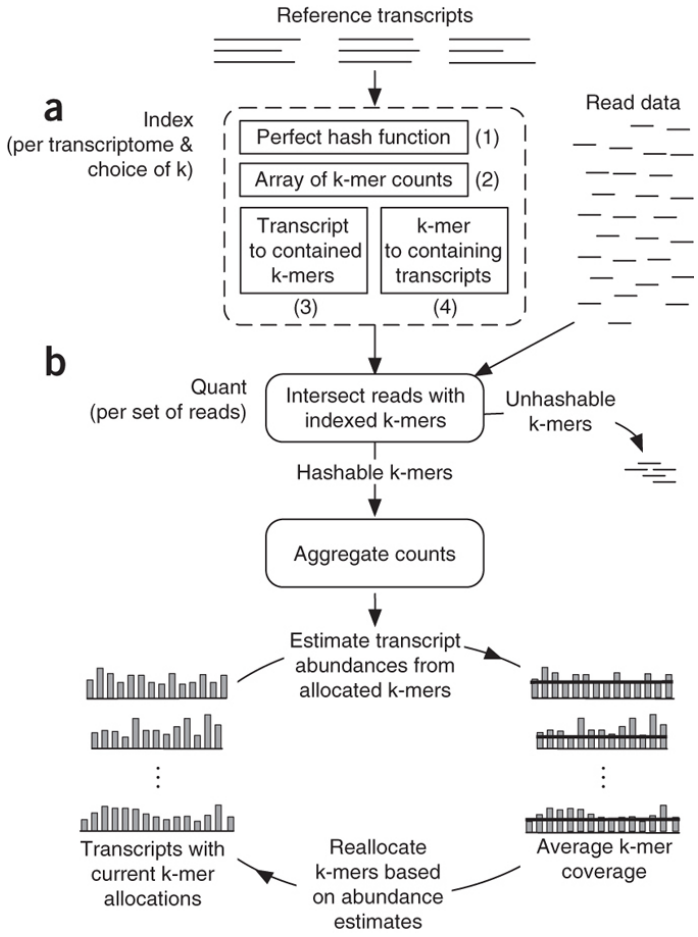
slide from Brian Haas; title modified

Traditional transcript counting programs

- Cufflinks (Trapnell et al. 2010)
  - Part of Tuxedo suite (Bowtie, Tophat)
  - Also reference-based transcriptome assembler - find new splice junctions, isoforms and genes
  - Takes ~2-4 hrs, including alignment
- RSEM
  - Typically run after Trinity, a de-novo transcriptome assembler
  - Uses Bowtie to align reads to transcriptome
  - Takes ~6 hrs, including alignment

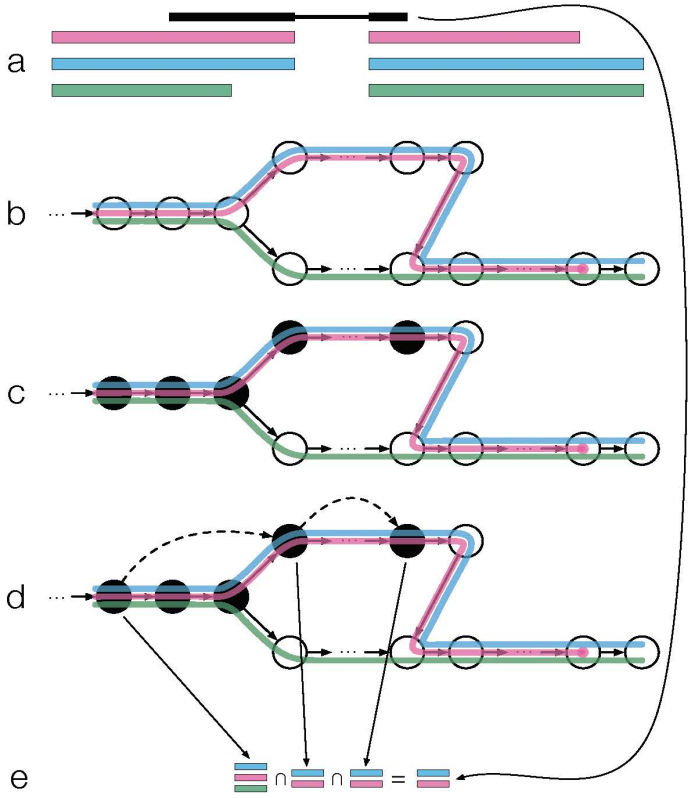
Modern transcript counting programs

- Sailfish (Patro et al. 2014)
  - estimates transcript coverage by k-mer counting approach
  - Takes 5-20 minutes
  - Cannot find new splice junctions/isoforms
- Salmon (Patro et al. 2017)
  - More accurate than Sailfish
  - Even faster: 3-5 min!



Modern transcript counting program based on pseudo-alignments

- Kallisto (Bray et al. 2016)
  - First creates a De Bruijn graph of the transcripts
  - Defines relationships between a read and possible transcripts
  - less than 5 min on laptop computer!!



When to use transcript-counting methods

- Genome duplications
- Many gene families
- When you have a large percentage (>15%) of multi-mapped reads

•**Note:** After counting at the transcript-level, you can then group by gene-level, which is more accurate.

Detailed Outline

4. **Transcriptomic analysis methods and tools**
    a. Steps common to both assembly and differential gene expression
        - ✧ Download data
        - ✧ Quality check
        - ✧ Data alignment
    b. Assembly
    c. Differential Gene Expression
    d. Choosing a method, the considerations…
    e. Final thoughts and observations

Transcriptome Analysis

How does one pick the right tools?

What does HPCBio use?

1. Quality Check - **FASTQC**
2. Trimming - **Trimmomatic**
3. Splice-aware alignment - **STAR**

   Bacterial alignment - **BWA** or **Novoalign**
4. Counting reads per gene - **featureCounts**

   Counting reads per isoform **- Salmon**
5. DGE Analysis - **edgeR** or **limma**

- Alignment visualization - **IGV**
- De novo transcriptome assembly – **Trinity**
- Reference-based transcriptome assembly – **StringTie**

How do I learn more about these steps?

- Your lab will go through some of these steps on a small dataset: **alignment, gene-counting, DGE analysis, and alignment visualization**
- We do offer a longer and very detailed workshop on these methods during Spring semester every year
- Check http://hpcbio.illinois.edu/hpcbio-workshops at the beginning of the year for updates

Detailed Outline

4. **Transcriptomic analysis methods and tools**
    a. Steps common to both assembly and differential gene expression
        - ✧ Download data
        - ✧ Quality check
        - ✧ Data alignment
    b. Assembly
    c. Differential Gene Expression
    d. Choosing a method, the considerations…
    e. Final thoughts and observations

Final thoughts and stray observations

1. Think carefully about what your experimental goals are before designing your experiment and choosing your bioinformatics tools

Final thoughts and stray observations

1. Think carefully about what your experimental goals are before designing your experiment and choosing your bioinformatics tools
2. When in doubt “Google it” and ask questions.
   - http://www.biostars.org/ - Biostar (Bioinformatics explained)
   - http://seqanswers.com/ - SEQanswers (the next generation sequencing community)

Final thoughts and stray observations

1. Think carefully about what your experimental goals are before designing your experiment and choosing your bioinformatics tools
2. When in doubt "Google it" and ask questions.
   - http://www.biostars.org/ - Biostar (Bioinformatics explained)
   - http://seqanswers.com/ - SEQanswers (the next generation sequencing community)
3. Another good resource if you are not ready to use the command line routinely is Galaxy. It is a web-based bioinformatics portal that can be locally installed, if you have the necessary computational infrastructure.

Final thoughts and stray observations

4. Today we covered how to deal with Illumina data, but you may also encounter long-read data as well
   - Hybrid transcriptome assemblies can be done, but are usually challenging
   - Using modern long read data on its own is usually sufficient

Documentation and Support

Online resources for RNA-Seq analysis questions –

- Software manuals
- http://www.biostars.org/ - Biostar (Bioinformatics explained)
- http://seqanswers.com/ - SEQanswers (the next generation sequencing community)
- Most tools have a dedicated lists/forums

Contact us at:

hpcbiohelp@illinois.edu

hpcbiotraining@igb.illinois.edu

jholmes5@illinois.edu

See website for upcoming workshops & services:

http://hpcbio.illinois.edu/

Thank you for your attention!