HIGH-PERFORMANCE BIOLOGICAL COMPUTING
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

RNA-Seq and Transcriptome Analysis

Jessica Holmes

High Performance Biological Computing (HPCBio)

Roy J. Carver Biotechnology Center

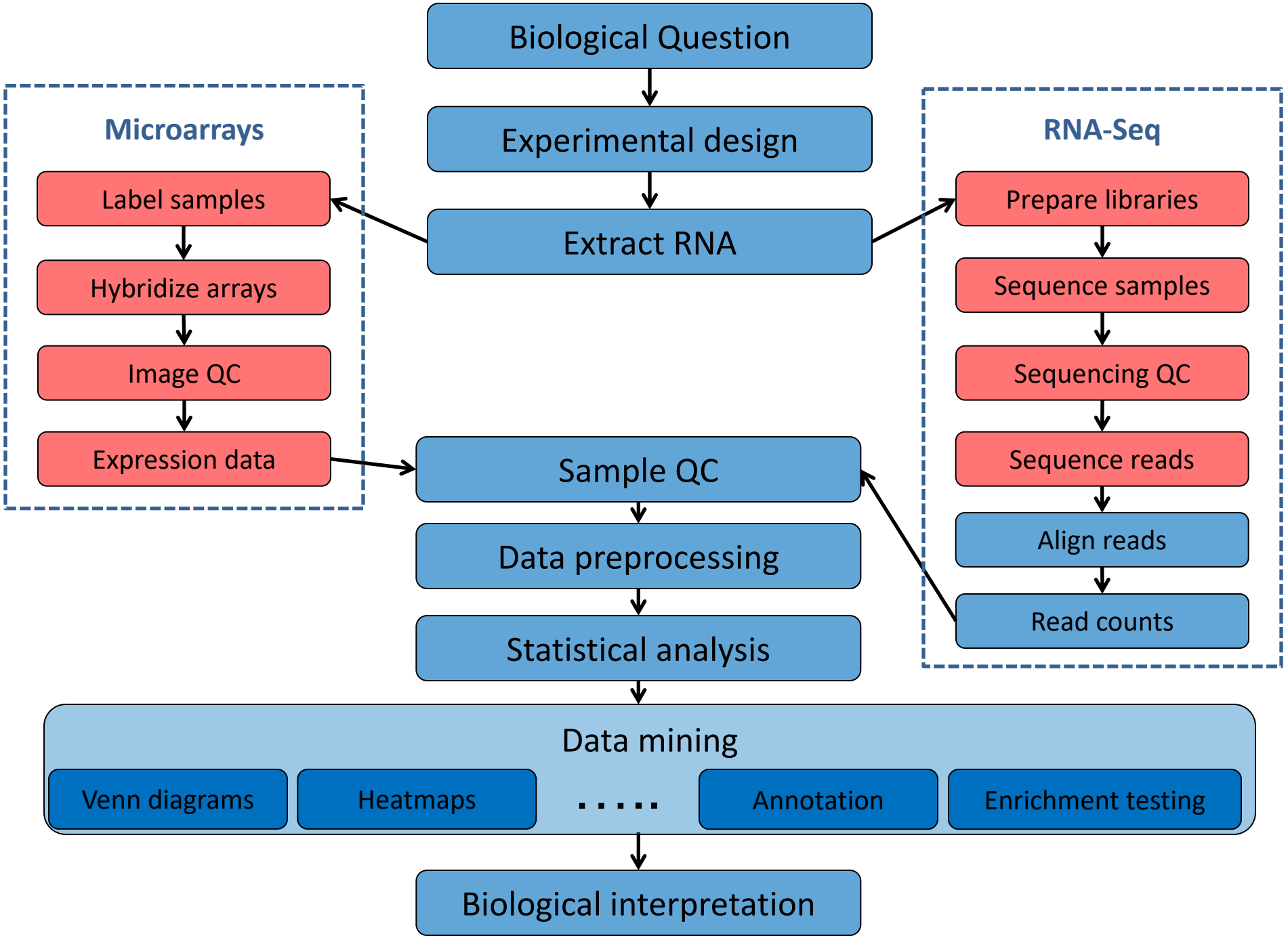
General Outline

1. Getting the RNA-Seq data: from RNA -> Sequence 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)

Why sequence RNA?

- **Differential Gene Expression**
  - Quantitative evaluation and 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

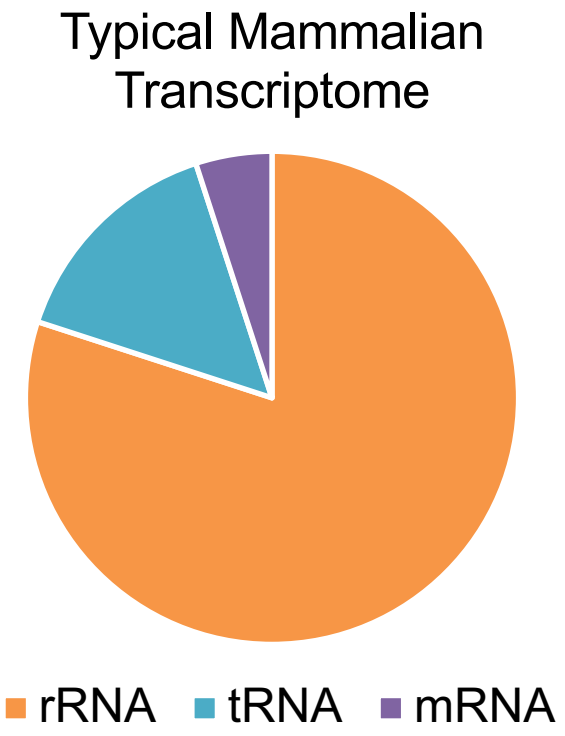
Types of RNA

- 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.)

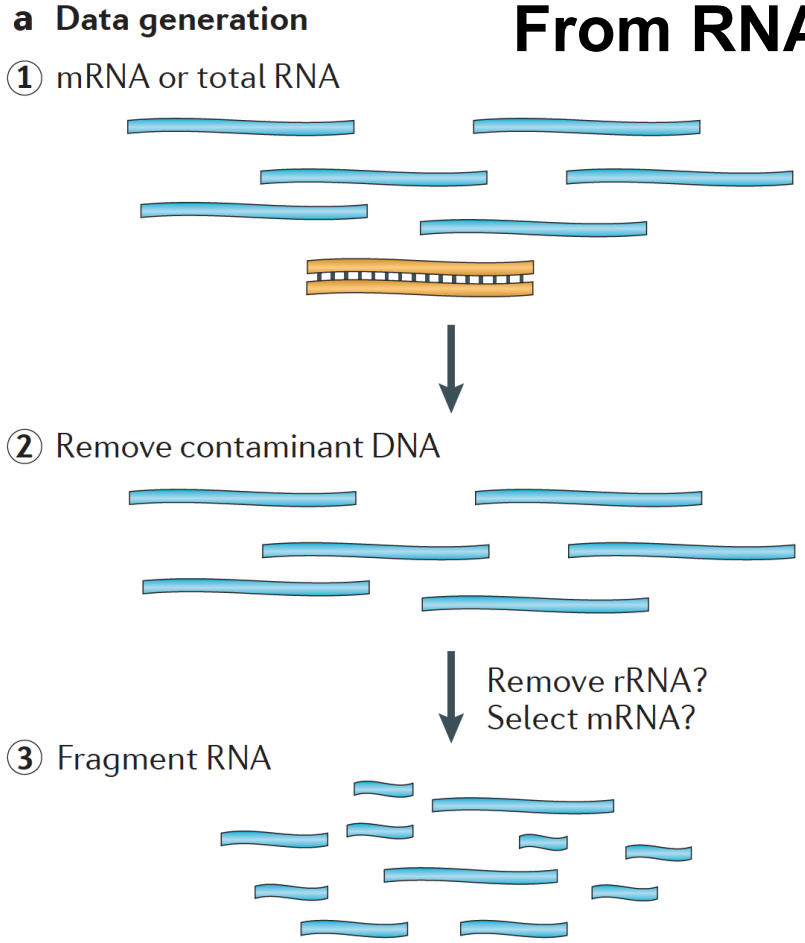
Removal of rRNA is almost always recommended

Removal Methods:

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

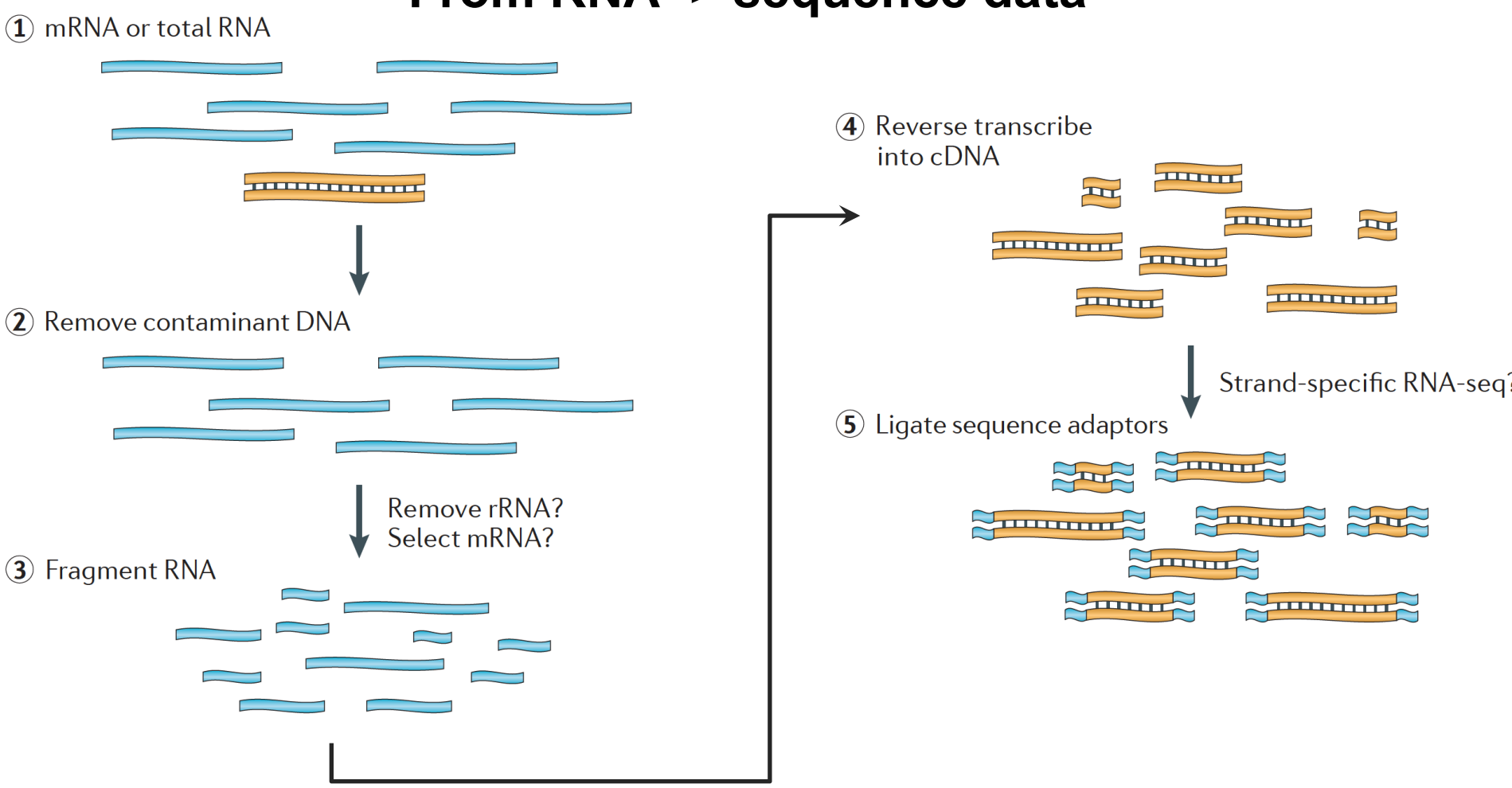


From RNA -> sequence data



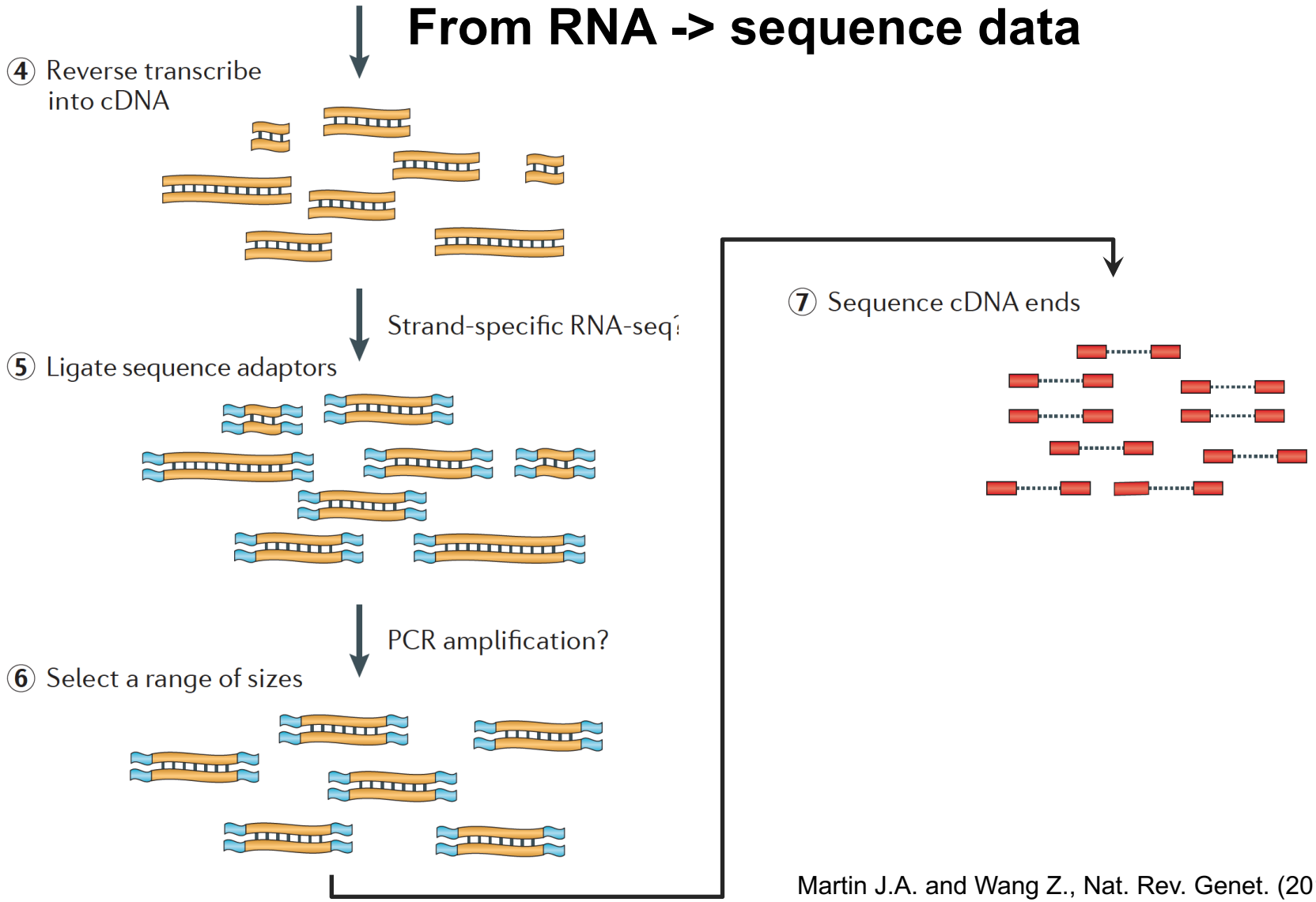
Martin J.A. and Wang Z., Nat. Rev. Genet. (2011) 12:671–682

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):

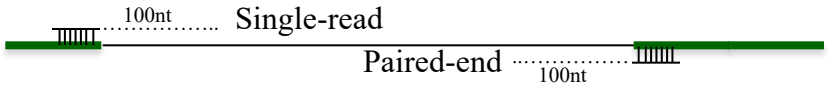
- **454** - pyrosequencing
- **SOLiD** – sequencing by ligation
- **Illumina** – sequencing by synthesis
- **Ion Torrent** – ion semiconductor
- **Pac Bio** – Single Molecule Real-Time sequencing, 1000 bp

3rd generation (2015)

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

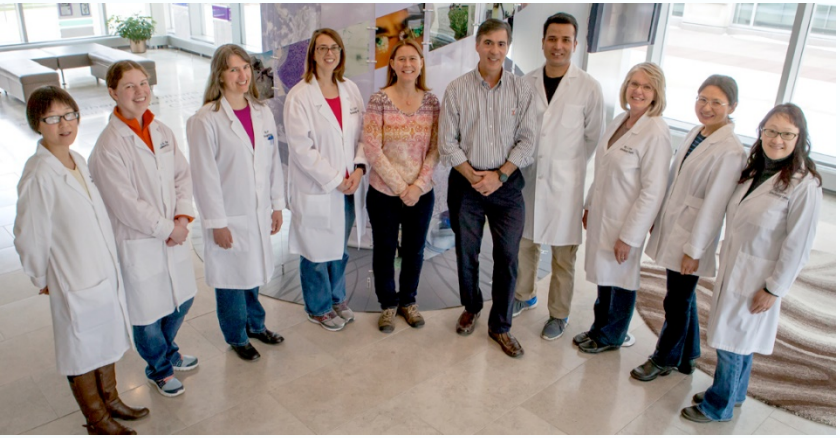
Illumina – "short read" sequencing

- 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**

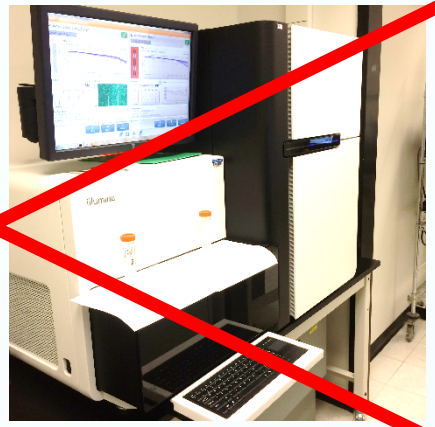


- **Accuracy**: 99.99% **Biases**: yes
- Most common platform for transcriptome sequencing

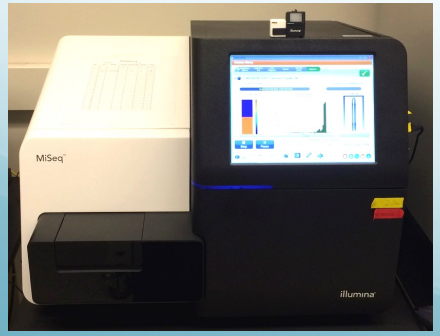
Library Construction and Sequencing Personnel and Equipment



2 Illumina HiSeq 4000 and two 2500

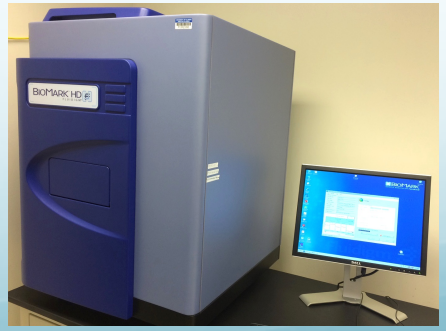


3 MiSeq



2 EpMotion

Fluidigm (FG)



1.5 PB archive

NovaSeq 6000

Any Genome. Any Method. Any Scale.

PE 150 | Q30 ≥ 75%

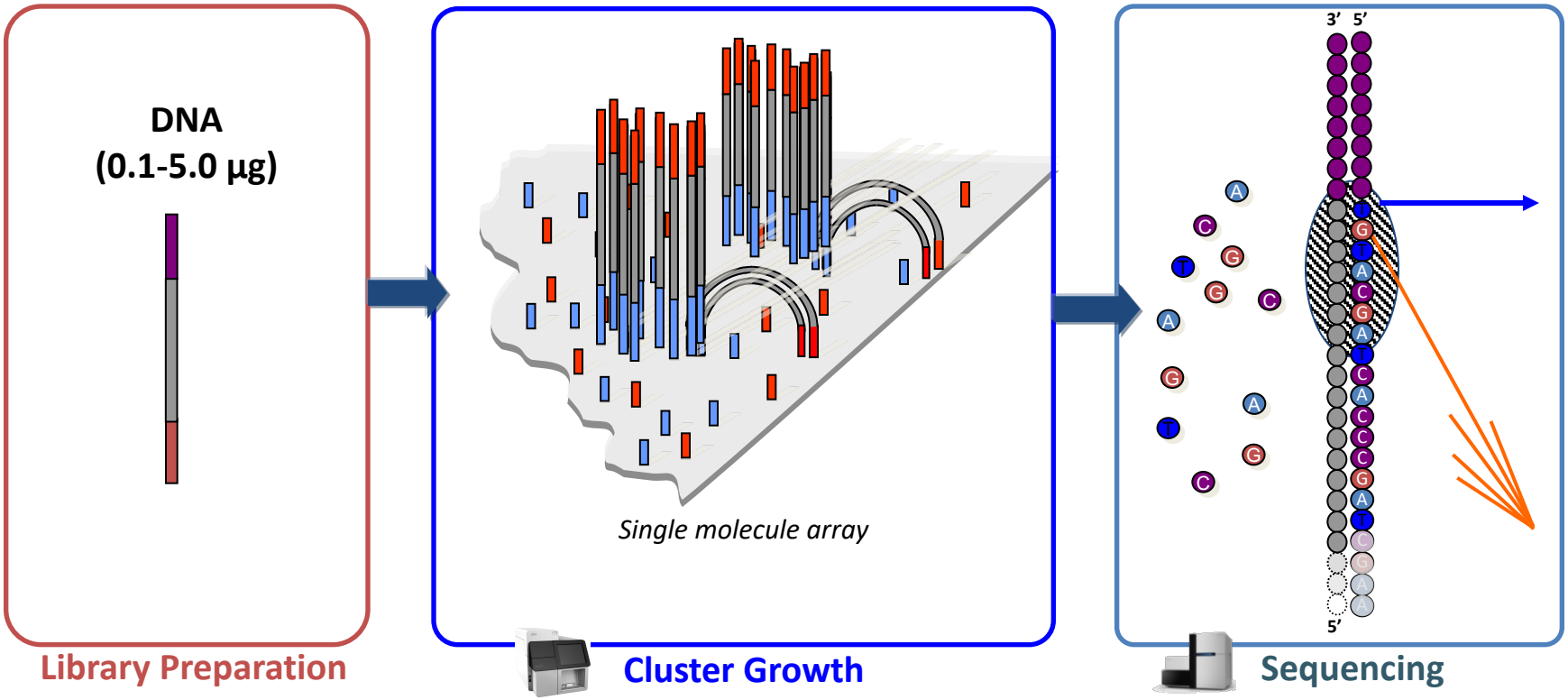
OUTPUT	167 – 3000 Gb
SINGLE READS	1.6 – 10B
RUN TIME	Fastest (40 Hr. for 2T Run)
Flow Cells	Scalable Flow Cell Format

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

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. Getting the RNA-Seq data: from RNA -> Sequence 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...

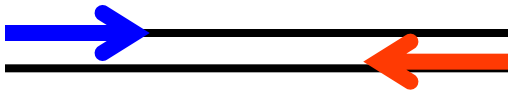
Differential Gene Expression

- 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

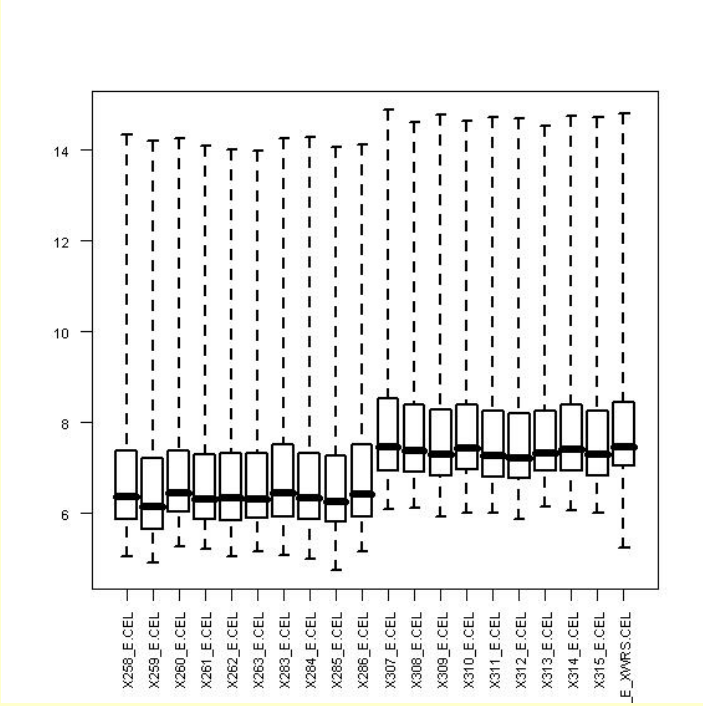
- Collect RNA from many various sources for a robust transcriptome
  - These can be pooled before or after sequencing (but before assembly)
- Poly-A enrichment is optional depending on your focus
- Remove ribosomal RNA (rRNA)
  - Unless you're interested in rRNA!
- Paired-end (PE) is recommended. The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

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.

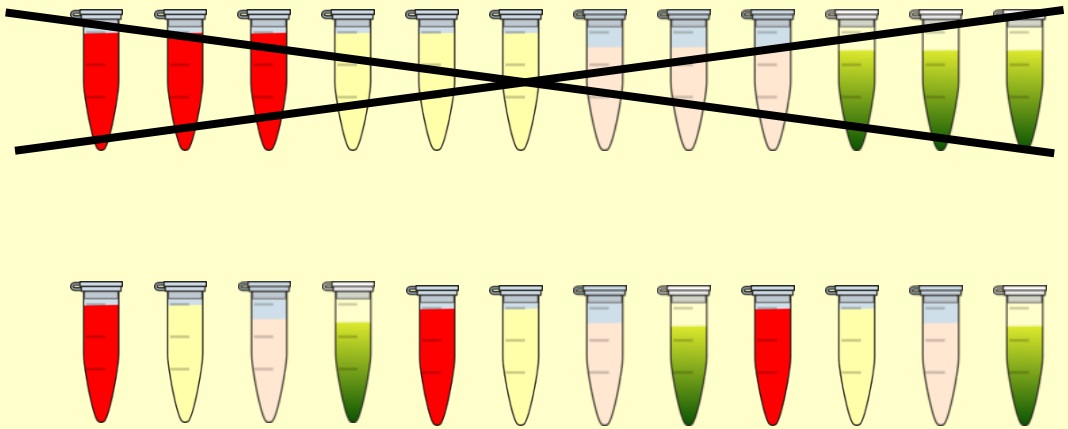


If batch effects are spread evenly over factor levels, they can be accounted for statistically

Slide courtesy of Jenny Drnevich

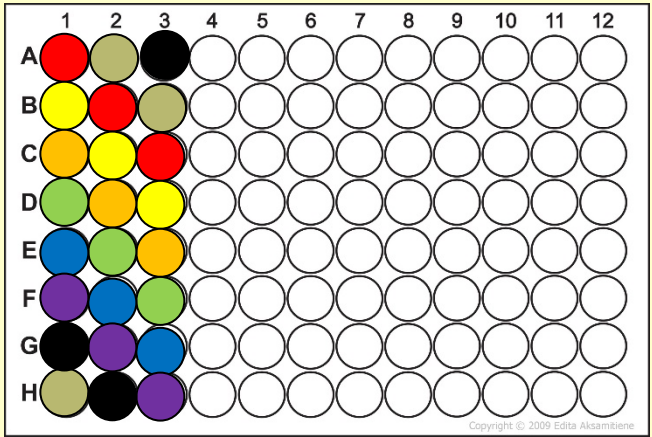
Beware systematic biases!

- 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!



Slide courtesy of Jenny Drnevich

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

General Outline

1. Getting the RNA-Seq data: from RNA -> Sequence 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

A brief note

Sequence formats

- FASTA
- FASTQ

Feature formats

- GFF
- GTF

Alignment formats

- *SAM*
- *BAM*

Formats: **FASTA**

```
>unique_sequence_ID My sequence is pretty cool
ATTCATTAAAGCAGTTTATTGGCTTAATGTACATCAGTGAAATCATAAATGCTAAAAA
```

- Deceptively simple format (e.g. there is no standard)
- However in general:
  - Header line, starts with '>'
  - followed **directly** by an ID
  - … and an optional description (separated by a space)
- Files can be fairly large (whole genomes)
- Any residue type (DNA, RNA, protein), but simple alphabet

Formats: **FASTA**

E.g. a read

```
>unique_sequence_ID
ATTCATTAAAGCAGTTTATTGGCTTAATGTACATCAGTGAAATCATAAATGCTAAAAATTTATGATAAAA
```

E.g. a chromosome

```
>Group10 gi|323388978|ref|NC_007079.3| Amel_4.5, whole genome shotgun
sequence
TAATTTATATATCTATTTTTTTTTATTAAAAAATTTATATTTTTGTTAAAATTTTATTTGATTAGAAATAT
TTTTACTATTGTTCATTAATCGTTAATTAAAGATAGCACAGCACATGTAAGAATTCTAGGTCATGCGAAA
TTAAAAATTAAAAATATTCATATTTCTATAATAATTAAATTATTGTTTTAATTTAAGTAAAAAAATTTCT
AAGAAATCAAAAATTTGTTGTAATATTGAAACAAAATTTTGTTGTCTGCTTTTTATAGTAACTAATAAAT
ATTTAATAAAAAATTACTTTATTTAATATTTTATAATAAATCAAATTGTCCAATTTGAAATTTATTTTAT
CACTAAAAATATCTTTATTATAGTCAATATTTTTTGTTAGGTTTAAATAATTGTTAAAATTAGAAAATGA
TCGATATTTTCAAATAGTACGTTTAACTAATACTTAAGTGAAAGGTAAAGCGGTTATTTAAAATATTGAT
TTATAATATTCGTGACATAATATATTTATAAATAGATTATATATATATATATACATCAAAATATTATACG
AGAACTAGAAAATATTACAGATGCAAAATAAATTAAATTTTGTAAATGTTACAGAATTAAAAATCGAAGT
```

Formats: **FASTQ**

- **FASTQ – FASTA with quality**

```
@unique_sequence_ID
ATTCATTAAAGCAGTTTATTGGCTTAATGTACATCAGTGAAATCATAAATGCTAAAAATTTATGATAAAA
+
=-(DD--DDD/DD5:*1B3&)-B6+8@+1(DDB:DD07/DB&3((+:?=8*D+DDD+B)*)B.8CDBDD4
```

- DNA sequence with quality metadata
- The header line, starts with '@',followed directly by an ID and an optional description (separated by a space)
- May be 'raw' data (straight from sequencing) or processed (trimmed)
- Variations: Sanger, Illumina, Solexa (Sanger is most common)
- Can hold 100's of millions of records
- **Files can be very large - 100's of GB apiece**

"Phred" quality (Q) scores

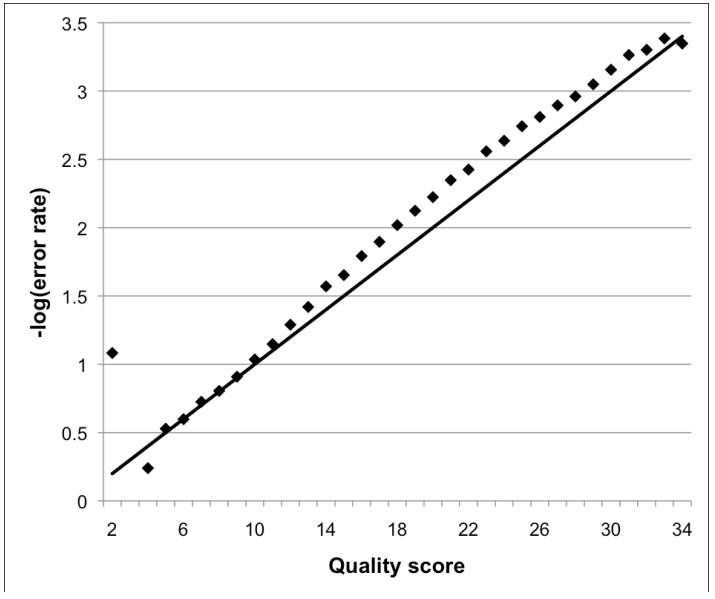
Historically developed for the phred program, an open source base caller for Sanger sequencing

$$Q = -10 * \log10\ (P)$$

Where P is the probability that a base call is erroneous

Q score	Prob. of wrong call	Accuracy
10	1 in 10 (0.1)	90%
20	1 in 100 (0.01)	99%
30	1 in 1000 (0.001)	99.9%
40	1 in 10000 (0.0001)	99.99%

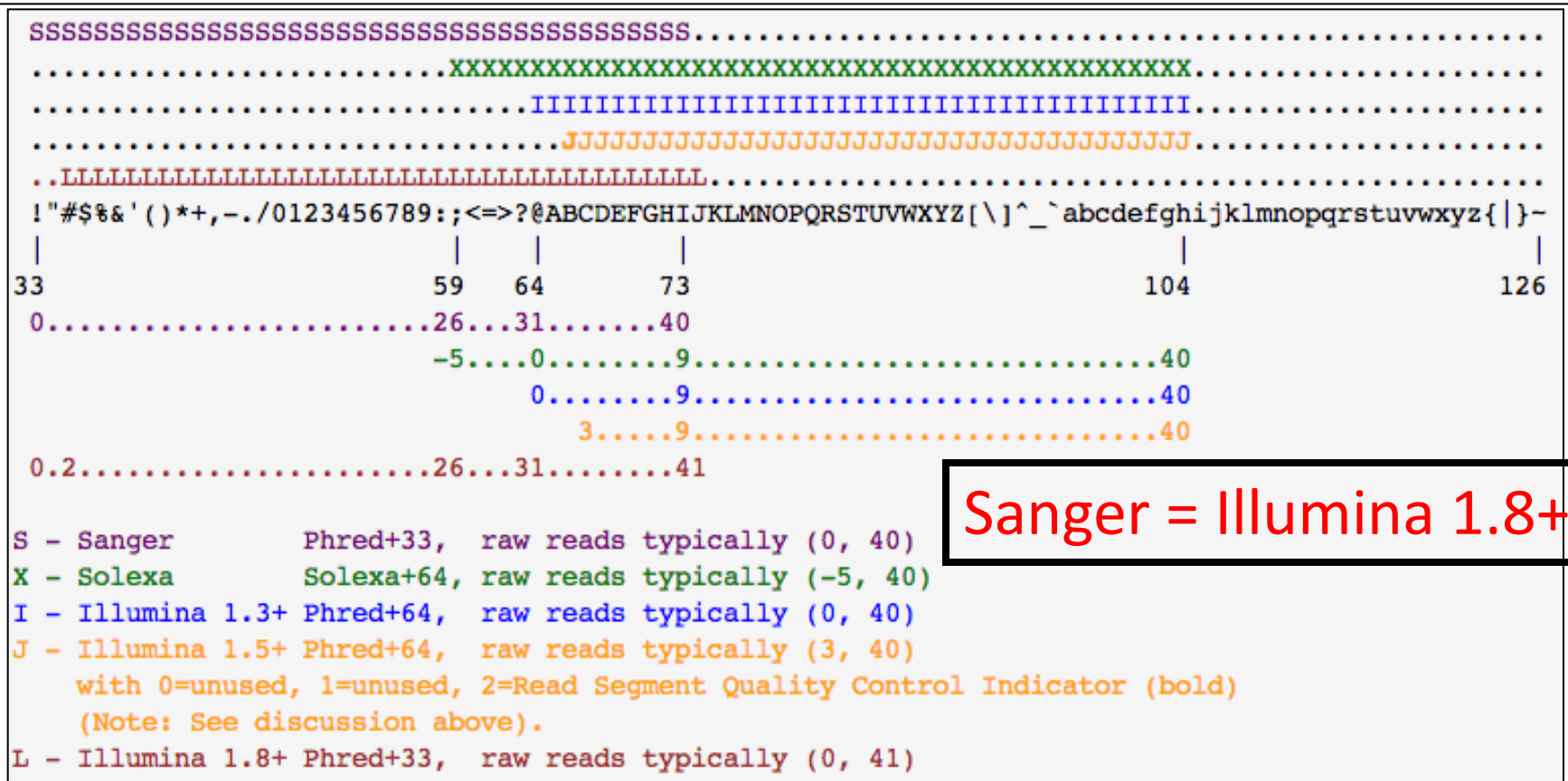
Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing
Akihiro Fujimoto, Hidewaki Nakagawa, Naoya Hosono, Kaoru Nakano, Tetsuo Abe, Keith A Boroevich, Masao Nagasaki, Rui Yamaguchi, Tetsuo Shibuya, Michiaki Kubo, Satoru Miyano, Yusuke Nakamura & Tatsuhiko Tsunoda
Nature Genetics, 2010



Formats: **FASTQ**

- **FASTQ – FASTA with quality**

```
@unique_sequence_ID
ATTCATTAAAGCAGTTTATTGGCTTAATGTACATCAGTGAAATCATAAATGCTAAAAATTTATGATAAAA
+unique_sequence_ID
=-(DD--DDD/DD5:*1B3&)-B6+8@+1(DDB:DD07/DB&3((+:?=8*D+DDD+B)*)B.8CDBDD4
```



Sanger = Illumina 1.8+

http://en.wikipedia.org/wiki/FASTQ_format

Feature formats

- ✧ GTF/GFF3
- ✧ SAM/BAM
- ✧ UCSC formats (BED, WIG, etc.)

Feature formats

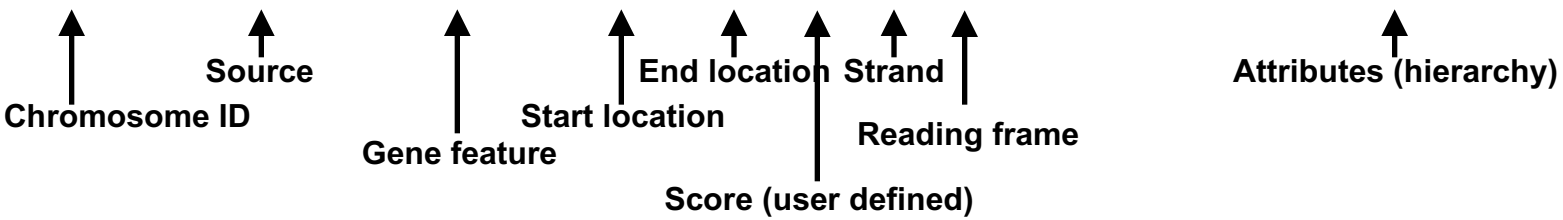
- Used for mapping features against a particular sequence or genome assembly
- May or may not include sequence data
- **The reference sequence must match** the names from a related file (possibly FASTA)
- **These are version (assembly)-dependent** - they are tied to a specific version (assembly/release) 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";
```

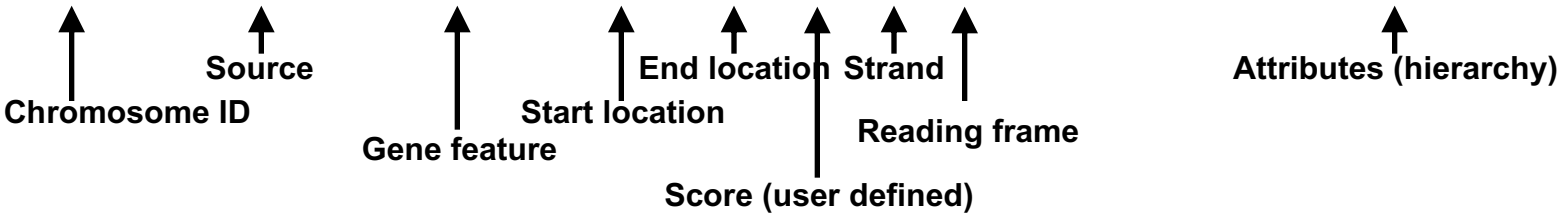


Feature formats : **GTF**

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";
```

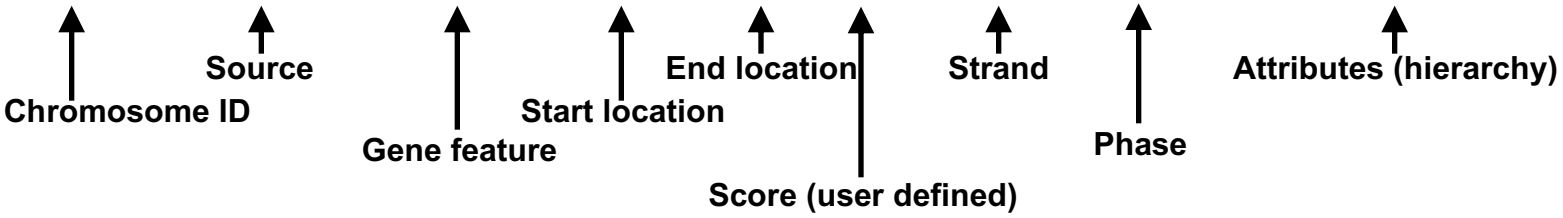


Feature formats : **GFF3**

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
```

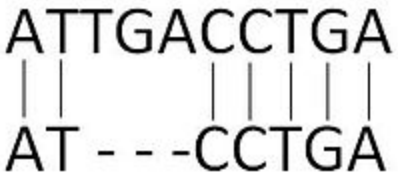
✧ GTF – Gene transfer format

```
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

What is an alignment?

- Wikipedia - “a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences”



- How can we store this information about millions of reads that align to our reference genome?

Formats : **SAM**

- **SAM – Sequence Alignment/Map format**
  - SAM file format stores alignment information
- **Plain text**
- **Specification**: http://samtools.sourceforge.net/SAM1.pdf
- Contains quality information, meta data, alignment information, sequence etc.
- **Files can be very large:** Many 100's of GB or more
- Normally converted into **BAM** to save space (and text format is mostly useless for downstream analyses)

@HD *[format version]*
@SQ SN:chr_1 LN:12345678
@PG *[information about program that made this]*
HWI-D00758:59:C7U2JANXX:1:1101:1398:2079 0 chr_1 130447256 255 1S9M * 0 0 NAGCTCTTTA #/<<BFBBFF NH:i:1 HI:i:1 AS:i:93 nM:i:2

Formats : **BAM**

- **BAM – BGZF compressed SAM format**
  - Compressed/binary version of SAM and is **not human readable.** Uses a specialized compression algorithm optimized for indexing and record retrieval (bgzip)
  - Makes the alignment information easily accessible to downstream applications (large genome file not necessary)
  - Unsorted, sorted by sequence name, **sorted by genome coordinates**
  - May be accompanied by an index file (.bai) (only if coordinate sorted)

- **Files are typically very large:** ~ 1/5 of SAM, but still very large

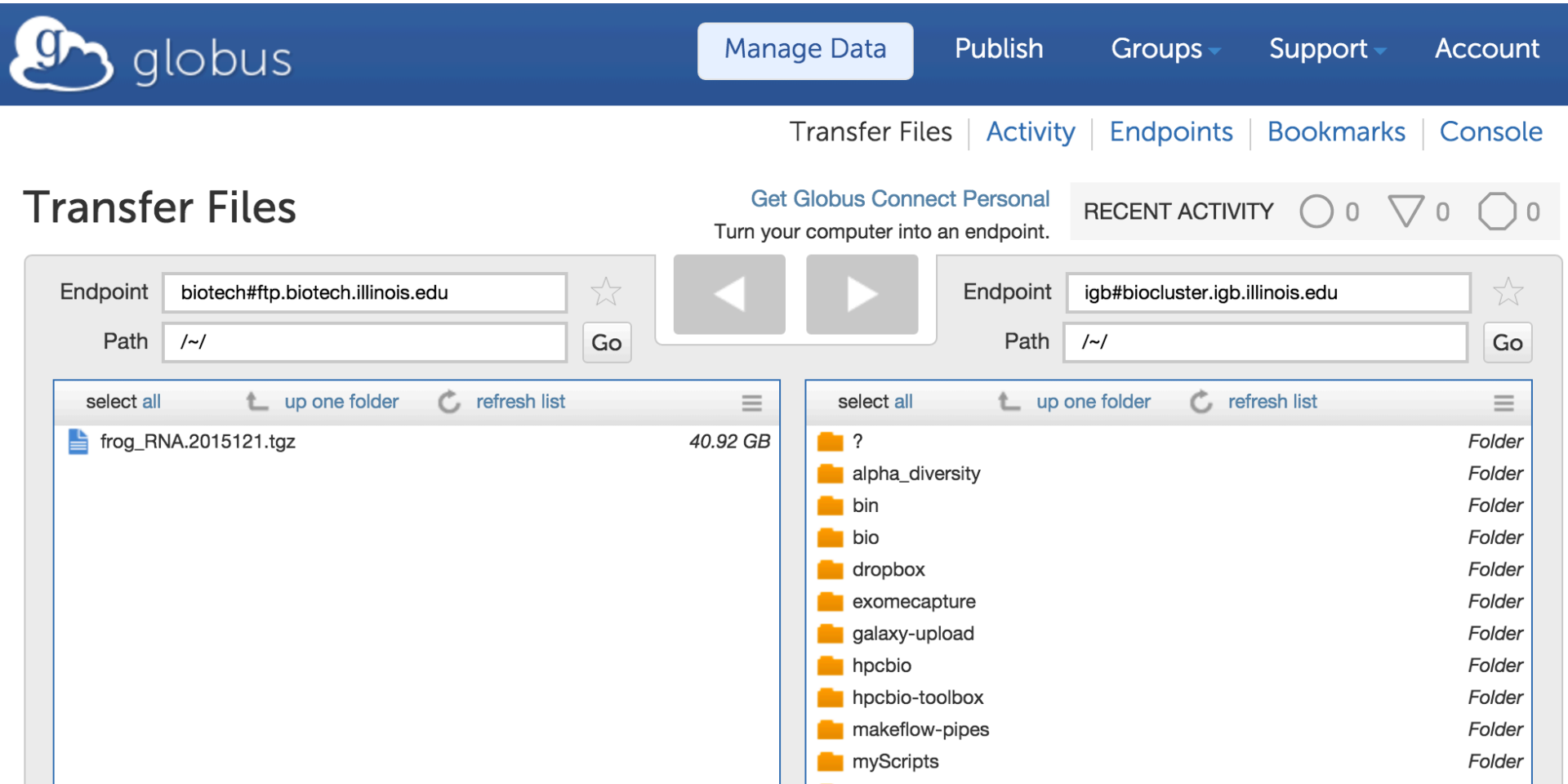
General Outline

4. Transcriptomic analysis methods and tools
    a. Transcriptome Analysis; aspects 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

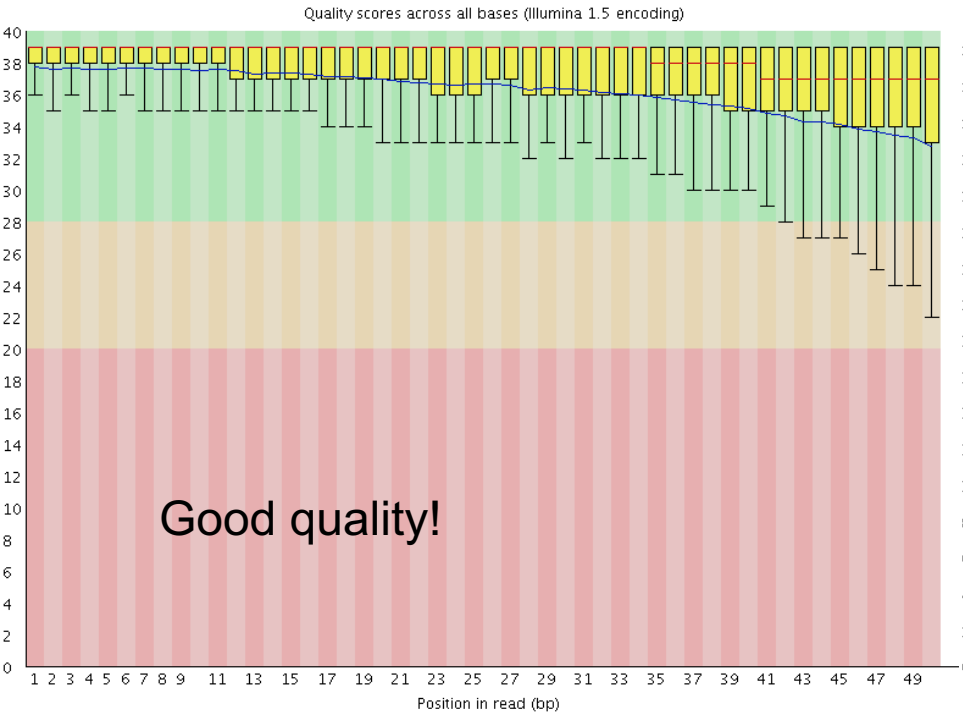

globus
Manage Data
Publish
Groups
Support
Account
Transfer Files
Activity
Endpoints
Bookmarks
Console
Transfer Files
Get Globus Connect Personal
Turn your computer into an endpoint.
RECENT ACTIVITY 0 0 0
Endpoint
biotech#ftp.biotech.illinois.edu
Path
/~/
Go
select all
up one folder
refresh list
frog_RNA.2015121.tgz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

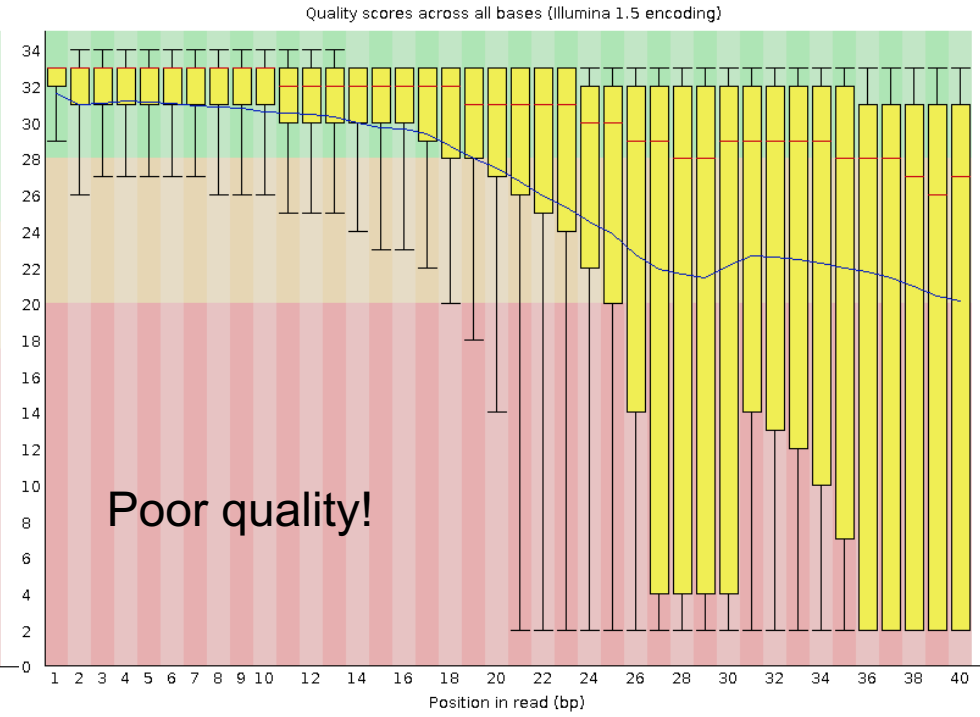
So how can we check the quality of our raw sequences?

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

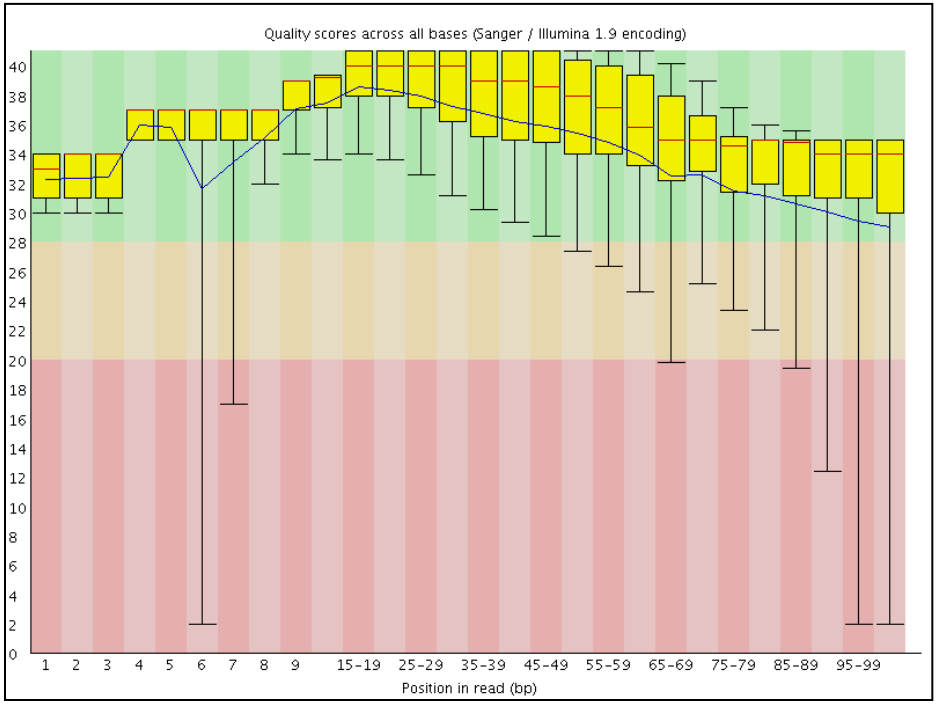
What do I do when FastQC calls my data poor?

- 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

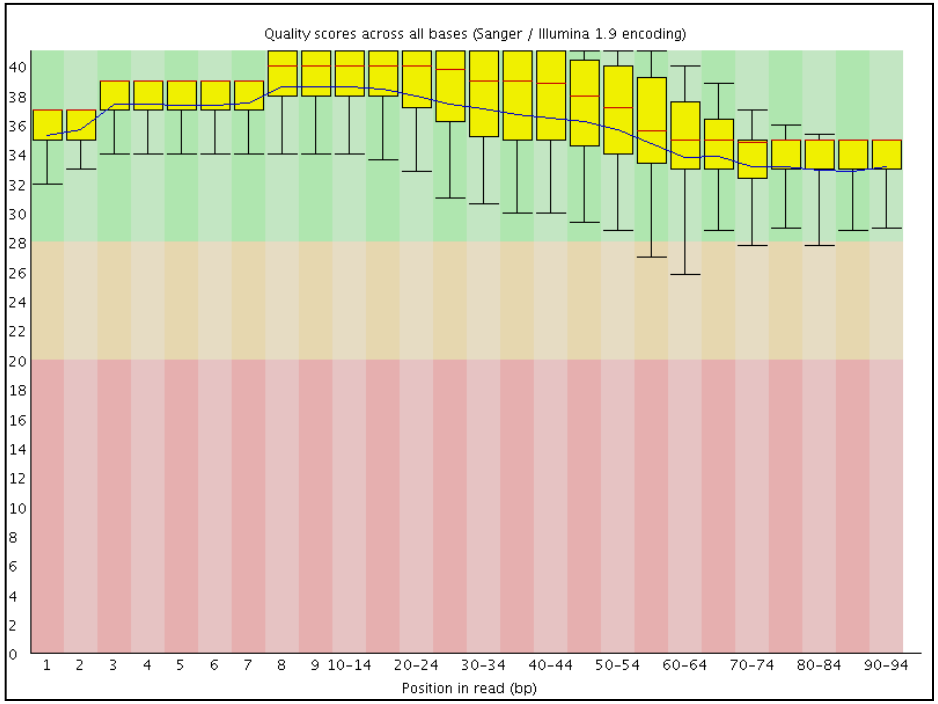
Transcriptome Analysis

Quality Checks

Before quality trimming



After quality trimming

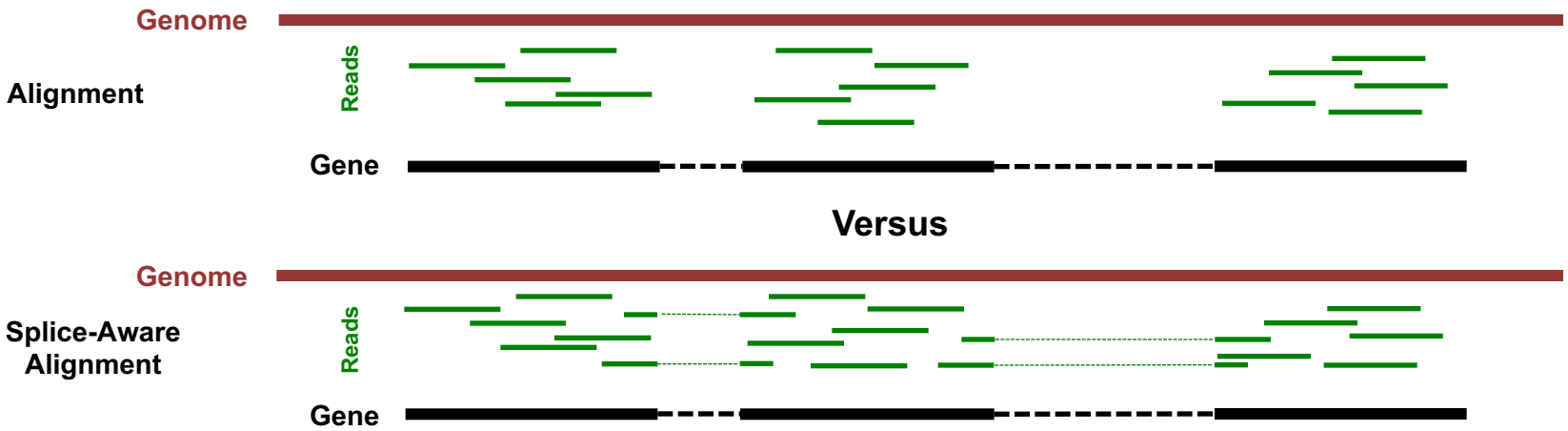


Transcriptome Analysis

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



Transcriptome Analysis

Data Alignment

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

- ✧ If aligning RNA-Seq data to the genome, always pick a splice-aware aligner (unless it's a bacterial genome!)

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

- ✧ There are excellent aligners available that are offer non-splice-aware alignment. This is ideal for bacterial genomes.

BWA, Novoalign (not free), Bowtie2, HiSat2

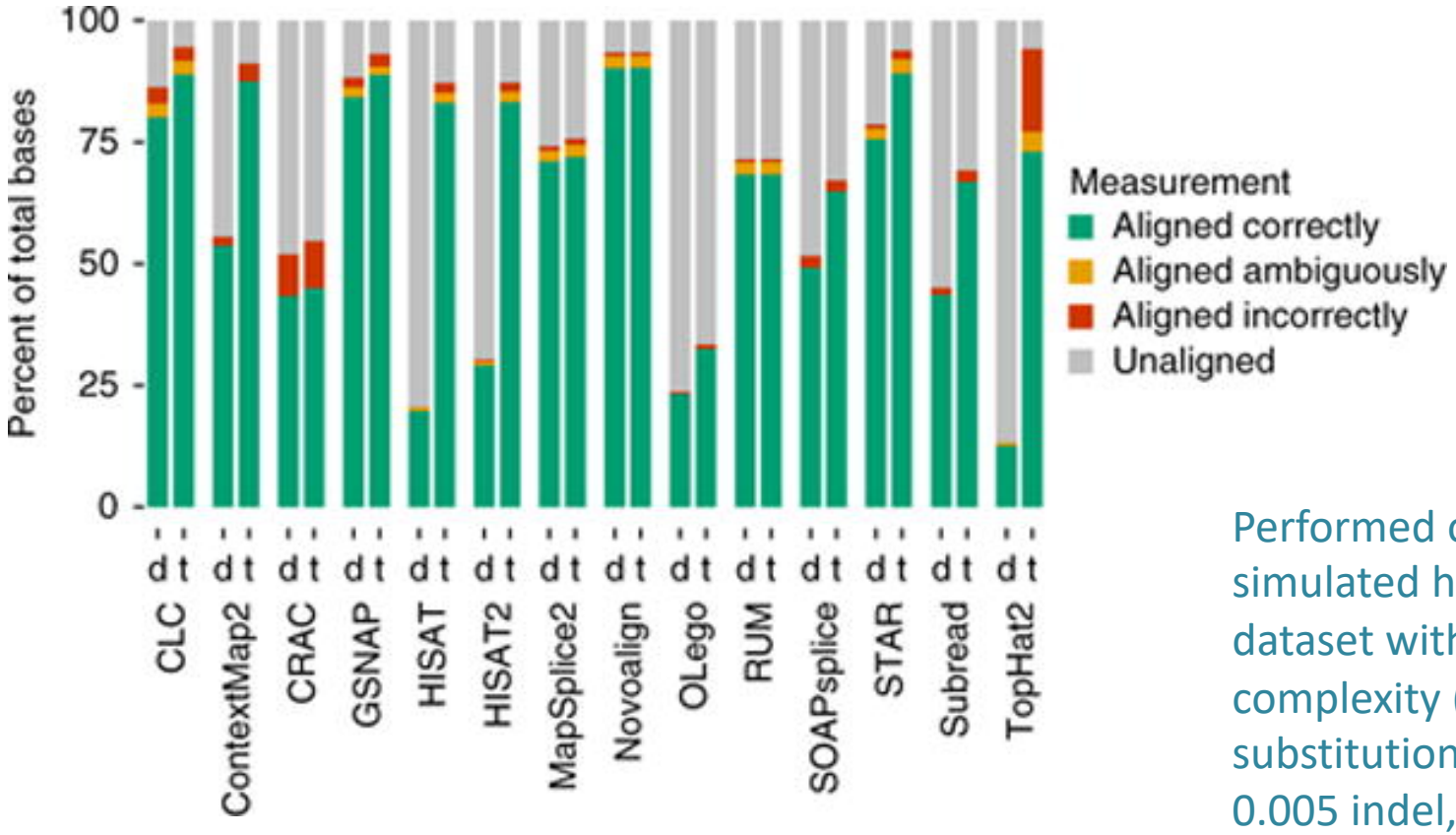
Transcriptome Analysis

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

Transcriptome Analysis

Alignment Visualization

IGV is the visualization tool used for this snapshot

General Outline

4. Transcriptomic analysis methods and tools
    a. Transcriptome Analysis; aspects 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

Two main types of assembly

a. Reference-based assembly

b. A *de novo* assembly

Transcriptome Assembly

Reference-based assembly

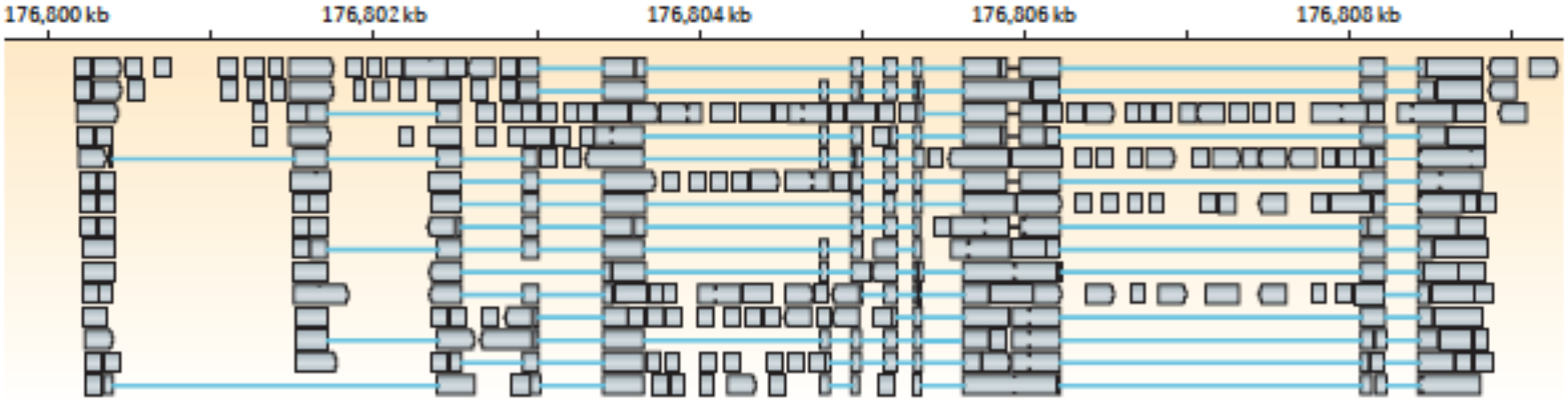
Used when the genome reference sequence is known, and:

- ✧ 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

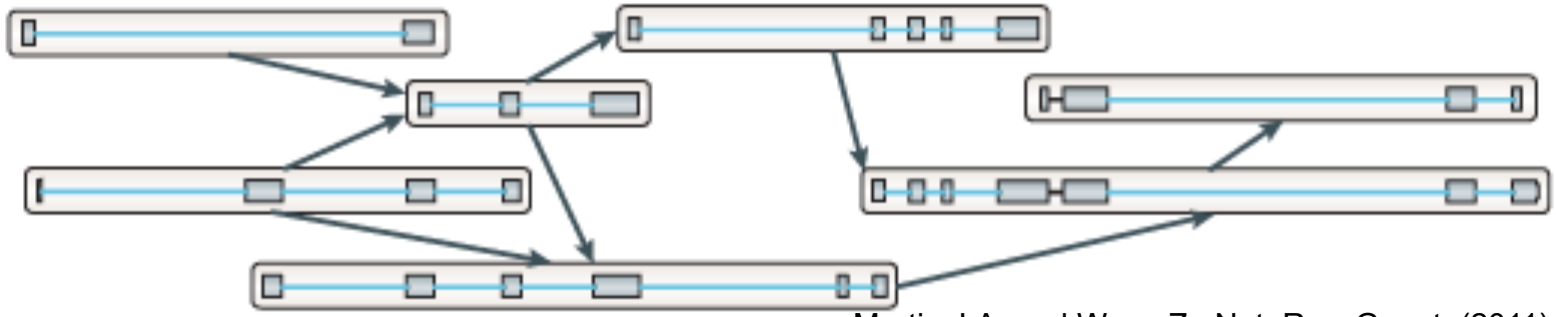
Transcriptome Assembly

a. Splice align reads to genome

Reference-based assembly



b. Build graph representing alternative splicing events

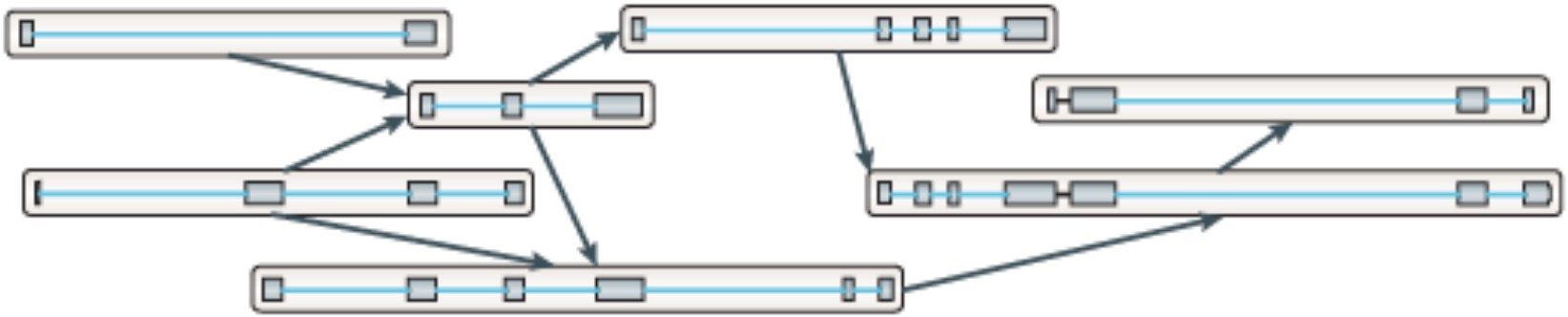


Martin J.A. and Wang Z., Nat. Rev. Genet. (2011) 12:671–682

Transcriptome Assembly

Reference-based assembly

b. Build graph representing alternative splicing events

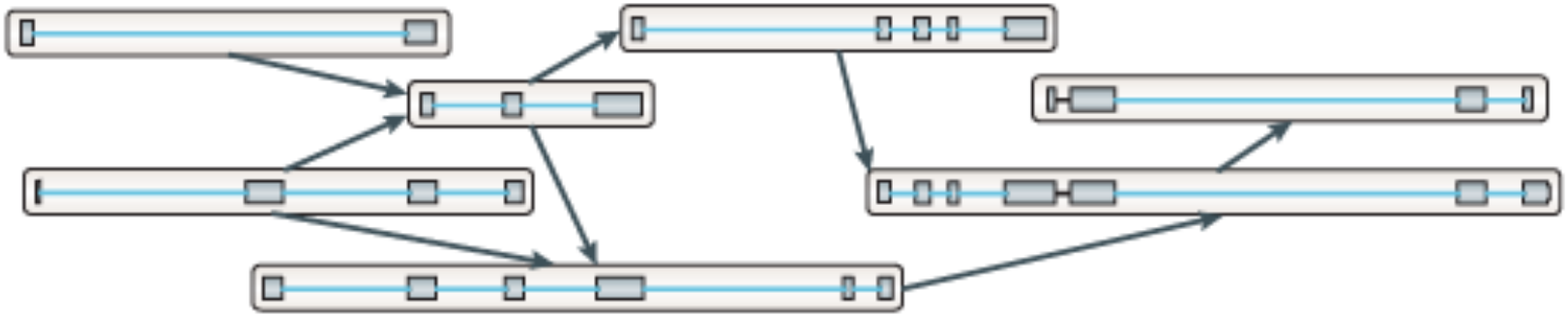


Martin J.A. and Wang Z., Nat. Rev. Genet. (2011) 12:671–682

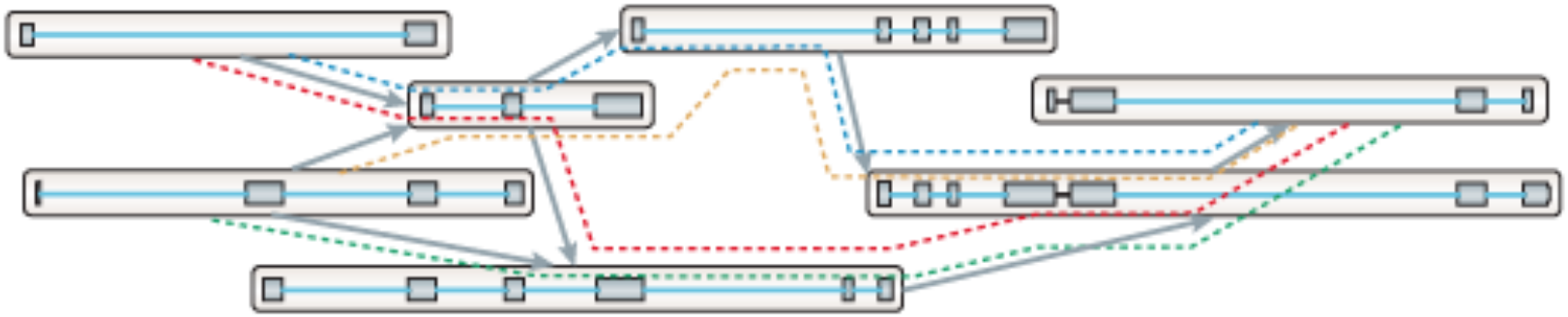
Transcriptome Assembly

Reference-based assembly

b. Build graph representing alternative splicing events



c. Traverse the graph to assemble variants

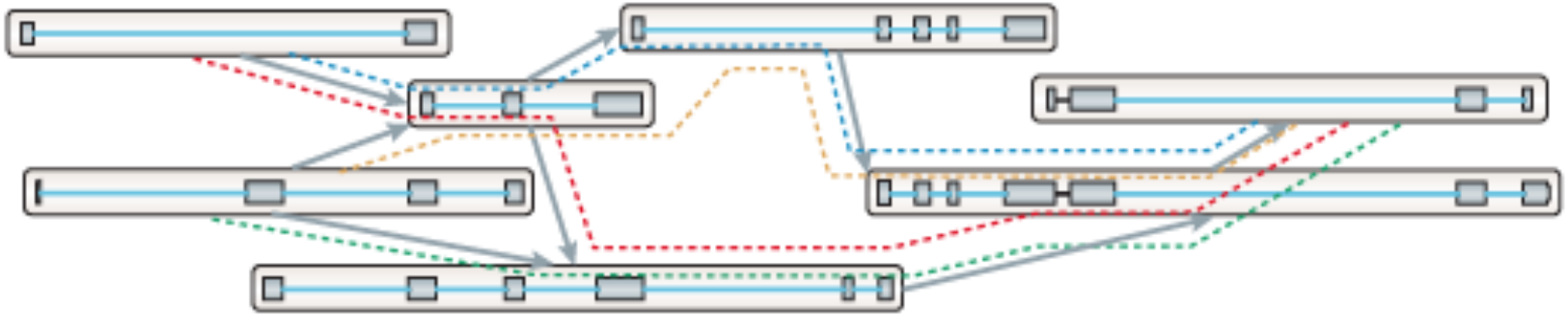


Martin J.A. and Wang Z., Nat. Rev. Genet. (2011) 12:671–682

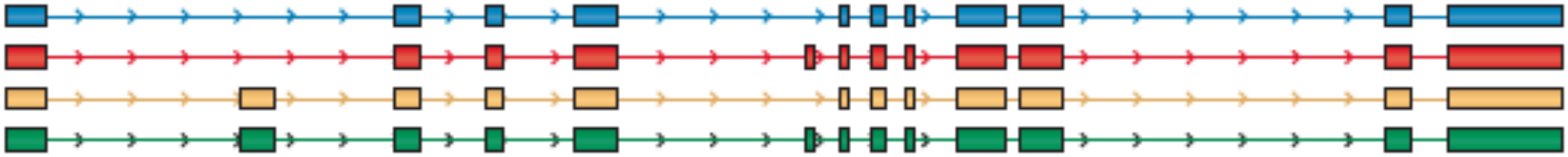
Transcriptome Assembly

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

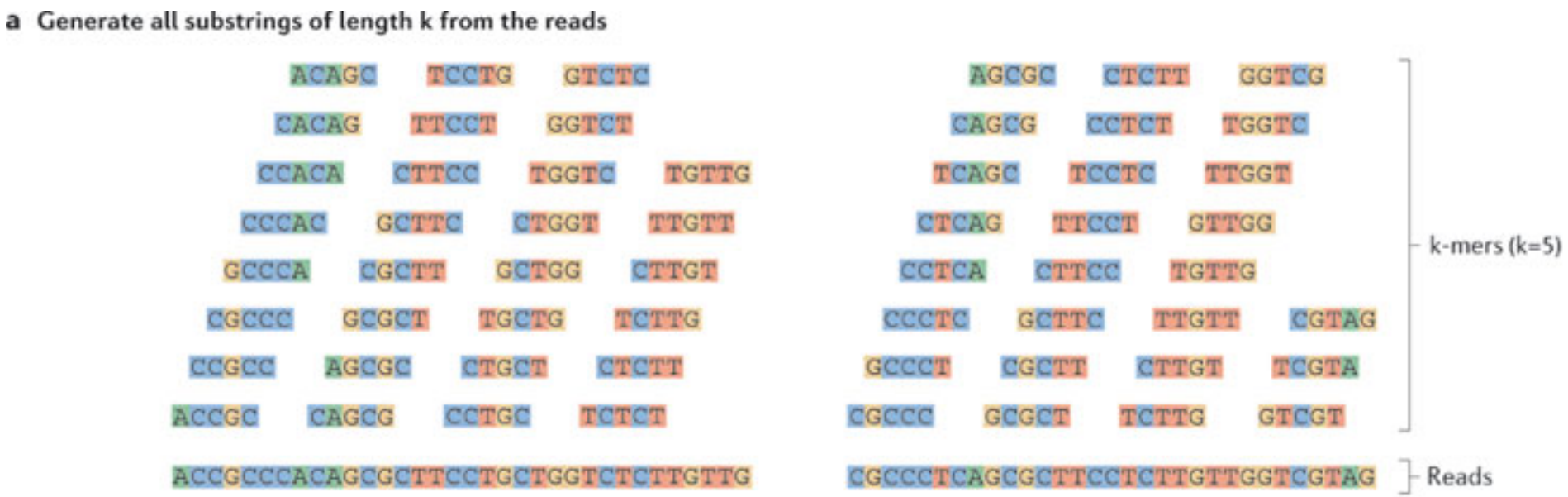
De novo assembly

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

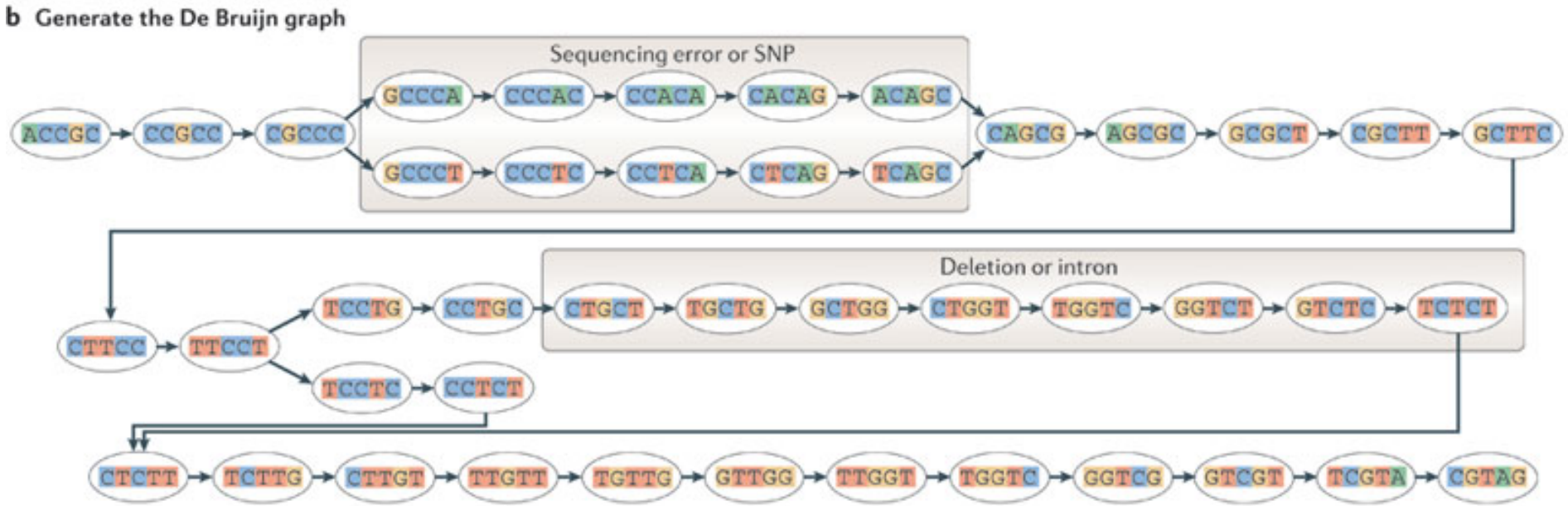
De novo assembly (De Bruijn graph construction)



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

Transcriptome Assembly

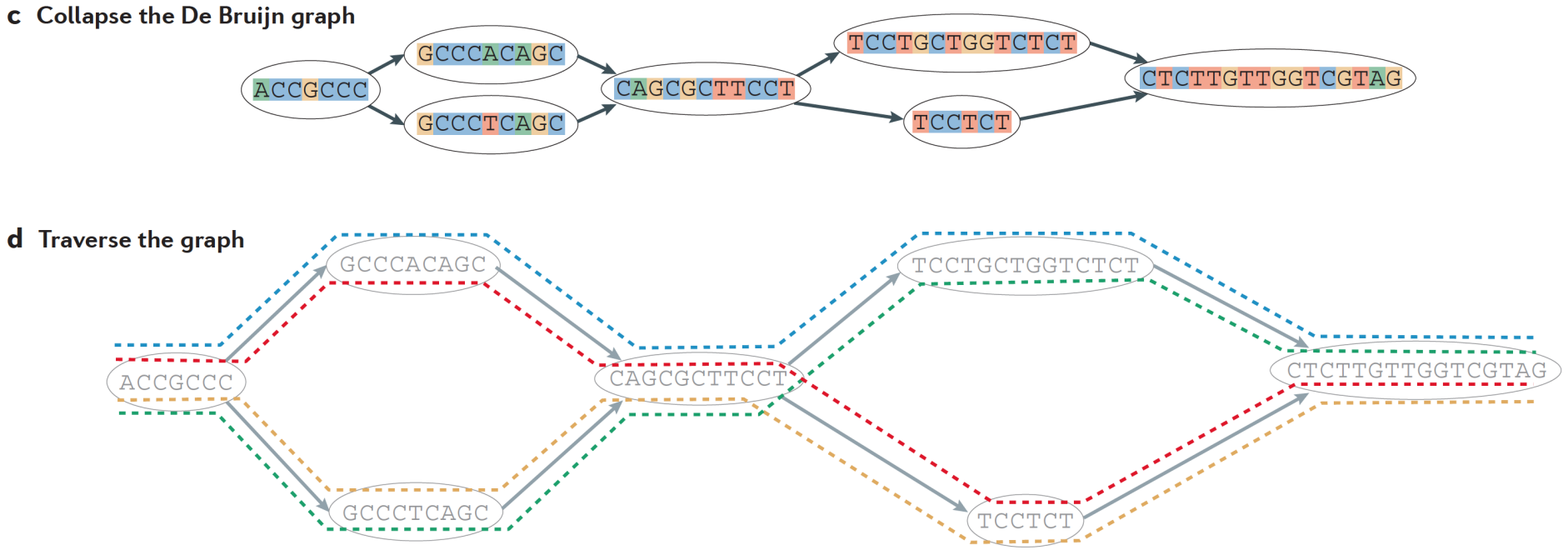
De novo assembly (De Bruijn graph construction)



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

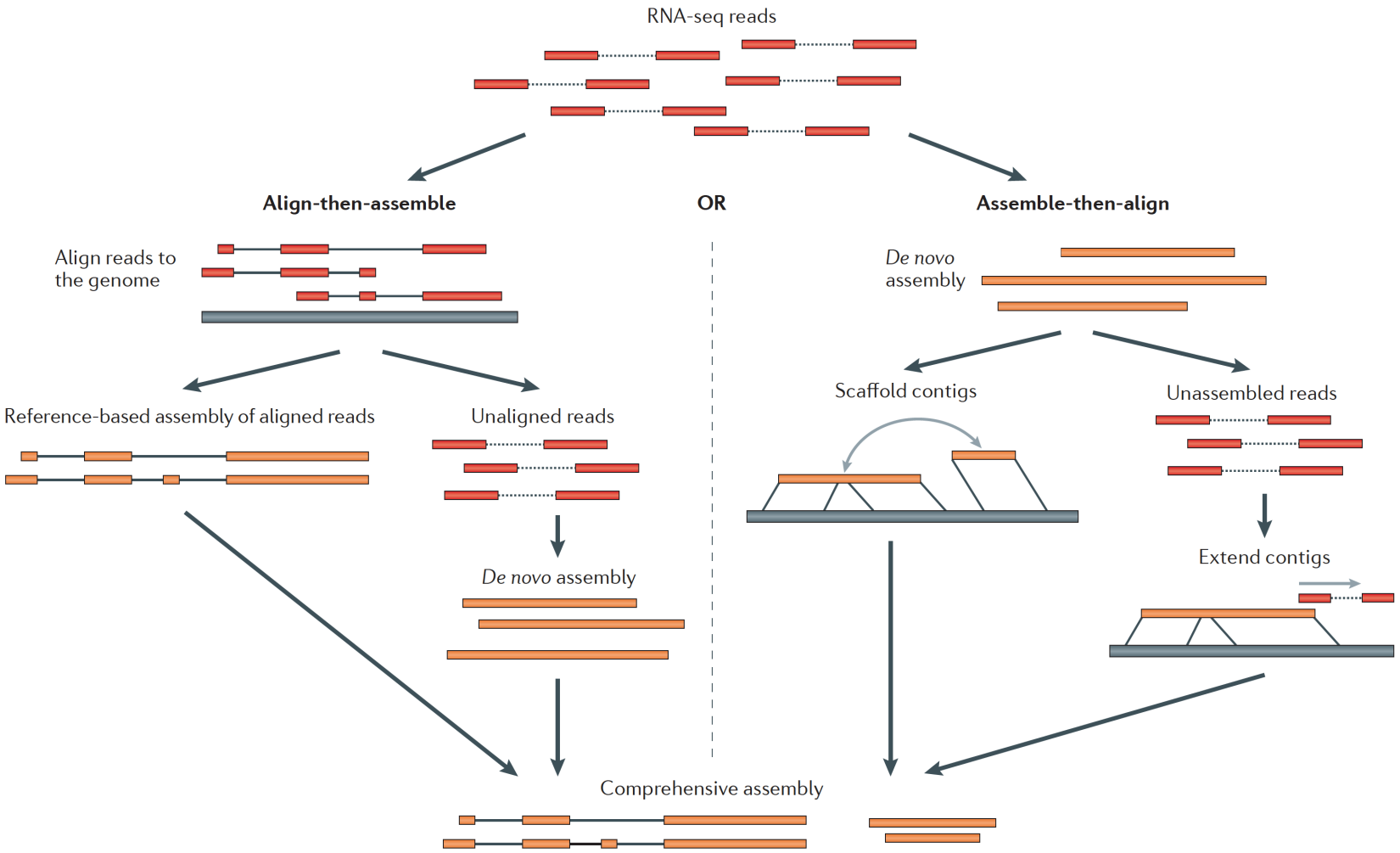
Transcriptome Assembly

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.

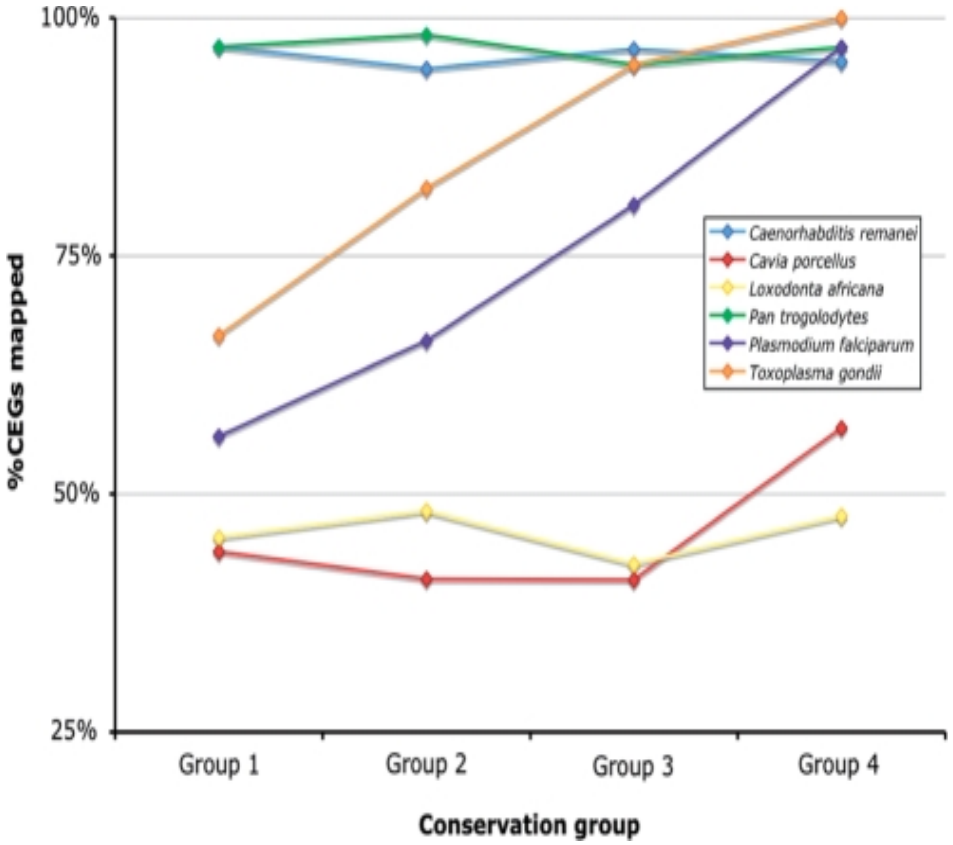
- Can generate a 'cleaned-up' or optimized assembly based on metrics

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



Outline

3. Transcriptomic analysis methods and tools

a. Transcriptome Analysis; aspects common to both assembly and differential gene expression
   - ✧ Quality check
   - ✧ Data alignment

b. Assembly

c. Differential Gene Expression

d. Choosing a method, the considerations…

e. Final thoughts and observations

Differential Gene Expression Overview

① Obtain/download sequence data

② Check quality of data and

③ Trim low quality bases, and remove adapter sequence

④ Align trimmed reads to genome of interest

a. Pick alignment tool

b. Index genome file

c. Run alignment after choosing the relevant parameters

Check every parameter and confirm that the aligner makes the correct assumptions for your genome! Otherwise, change them

Differential Gene Expression overview

④ Set up to do differential gene expression (DGE)

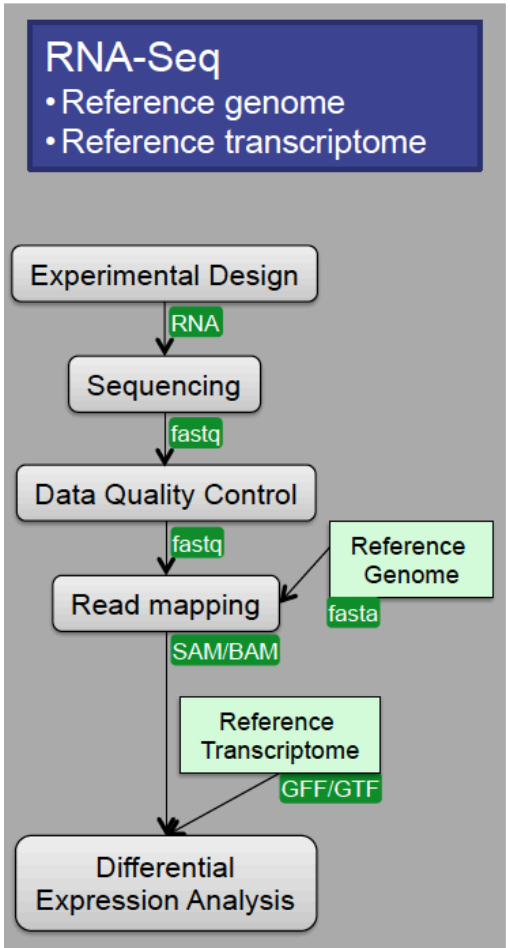
Identify read counts associated with genes

a. Do 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

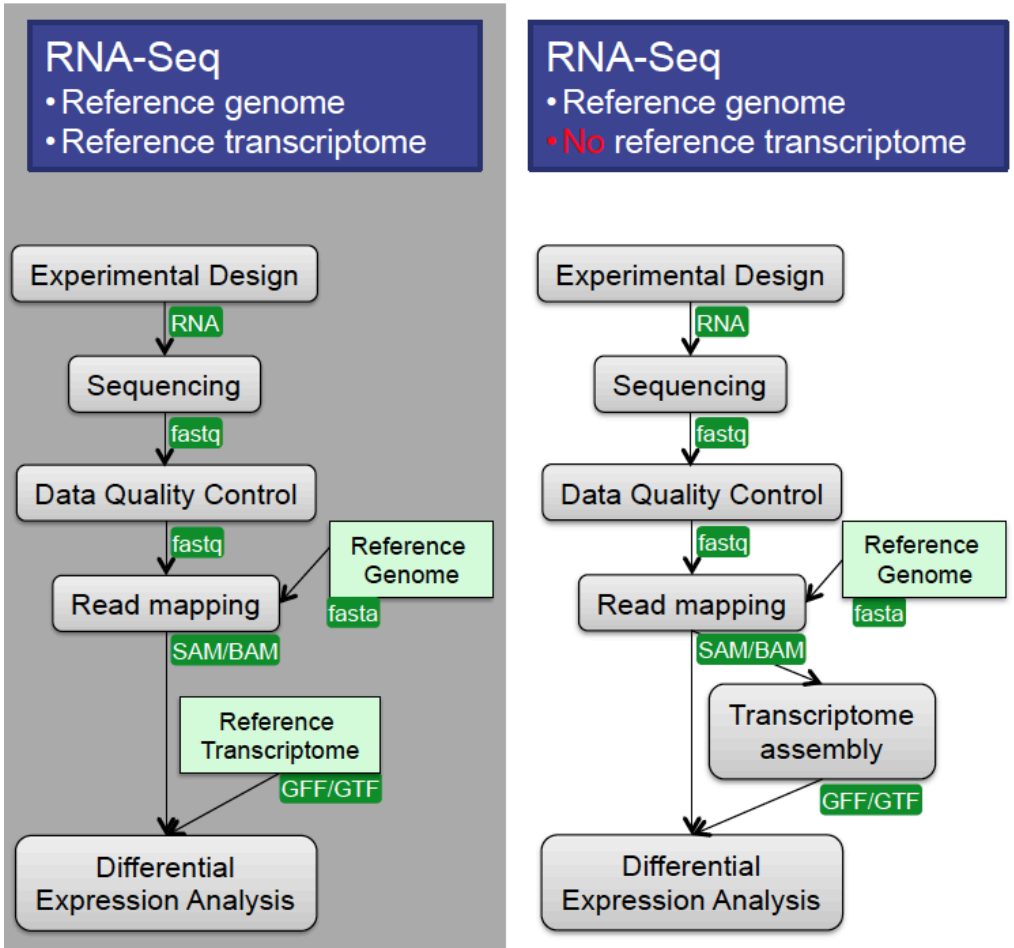
Differential Gene Expression

Options for DGE analysis



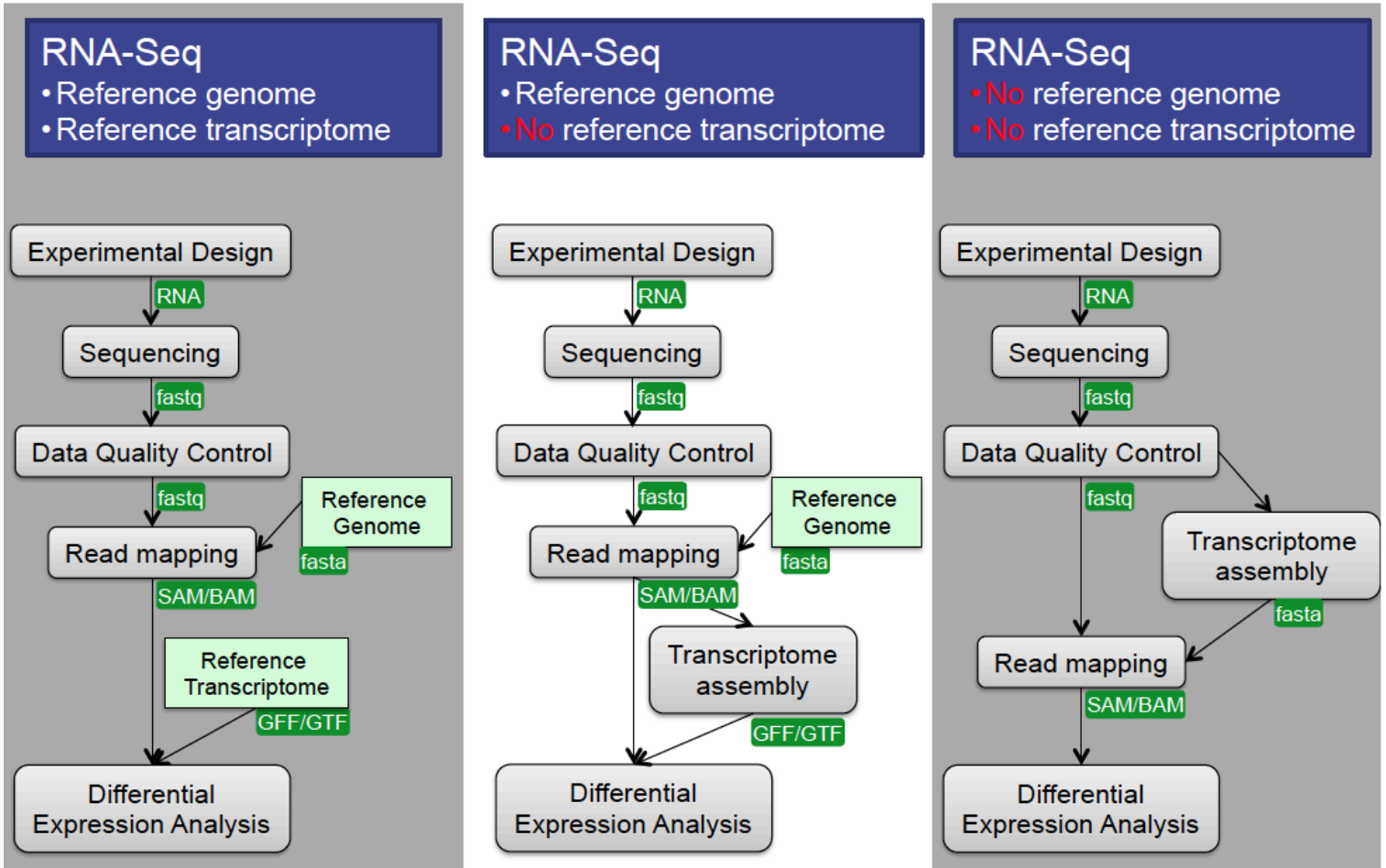
Differential Gene Expression

Options for DGE analysis



Differential Gene Expression

Options for DGE analysis



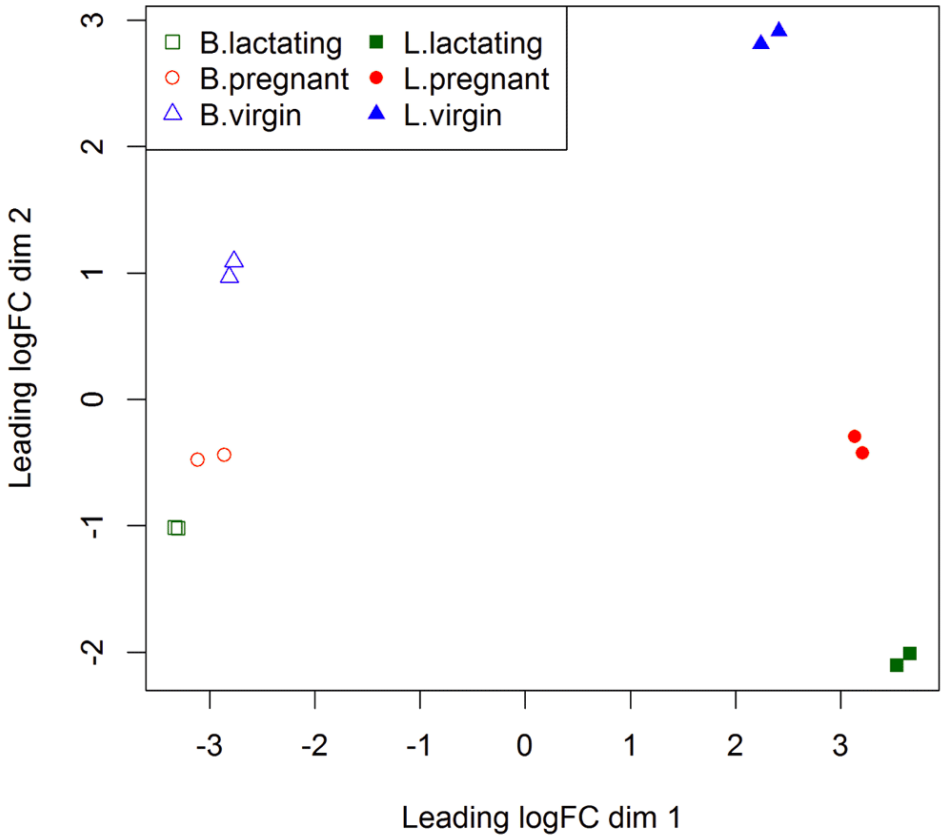
DGE Statistical Analyses

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

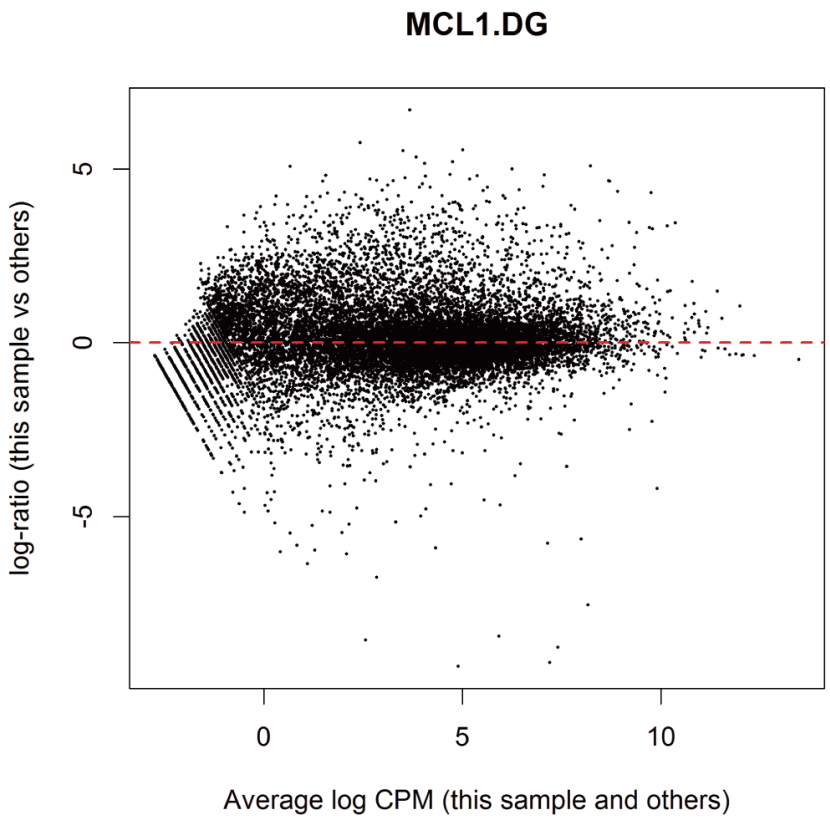
- A list of significantly differentially expressed genes
- Heatmaps, Venn Diagrams, and more
- Annotation
- WGCNA
- ... and more!

EdgeR: MDS Plot



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

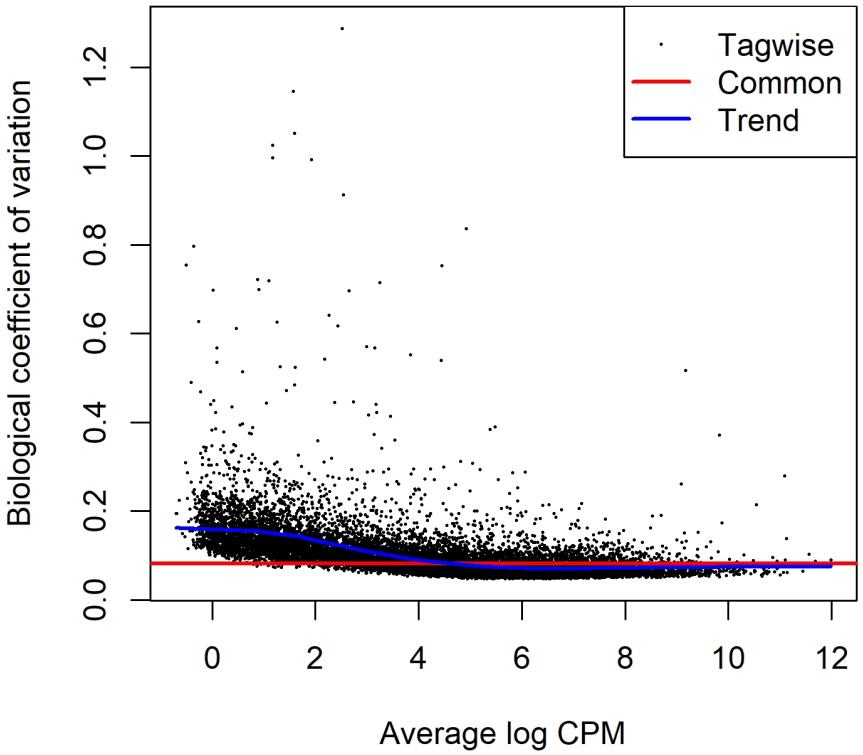
EdgeR: MD Plot



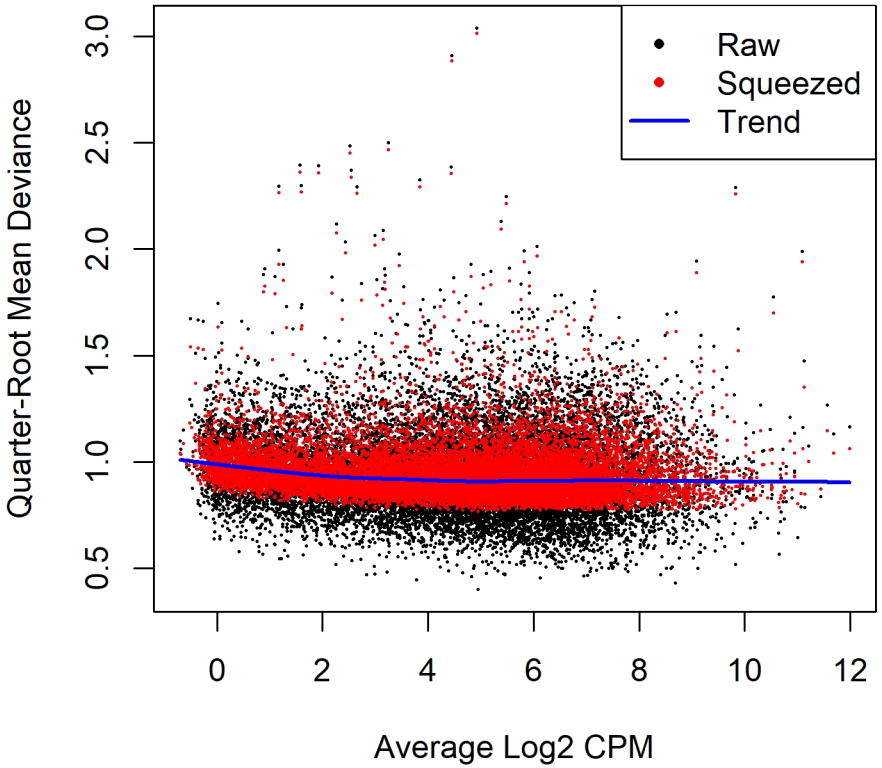
https://f1000research.com/articles/5-1438 (doi: 10.12688/f1000research.8987.2)

EdgeR Results: Dispersion Estimation

BCV Plot



QL Plot

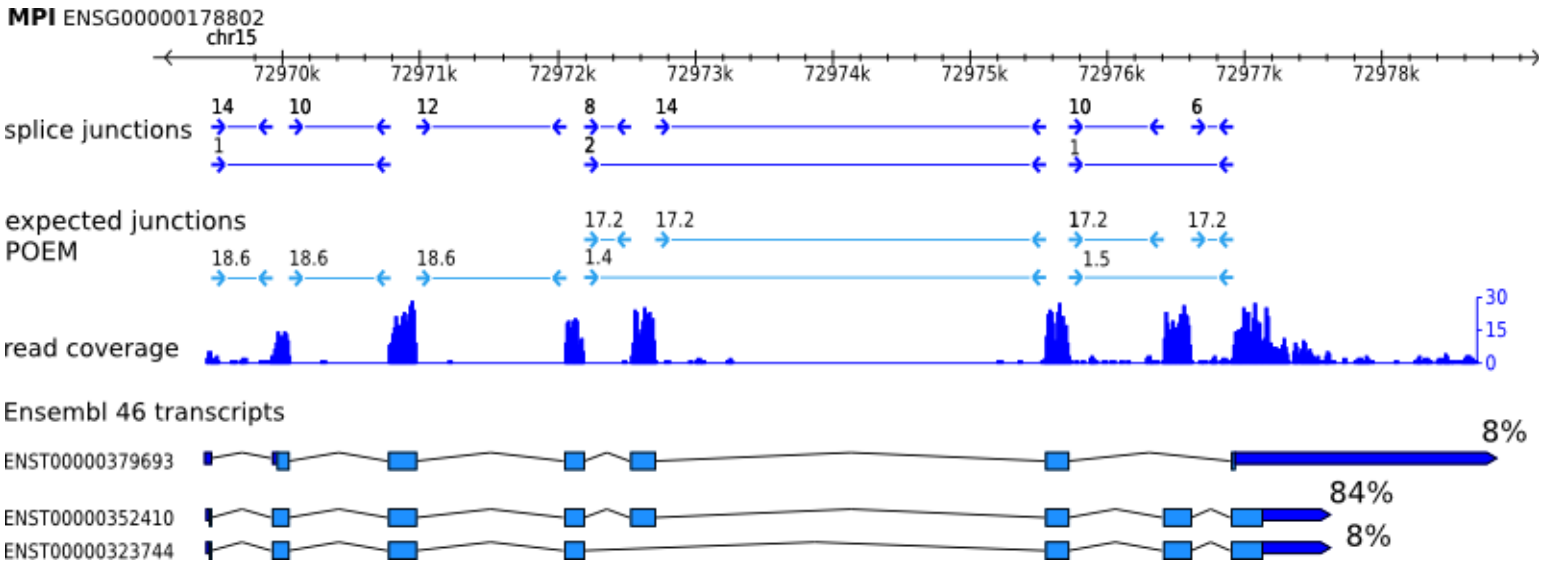


https://f1000research.com/articles/5-1438

TRANSCRIPT COUNTING METHODS

Can't use STAR/featureCounts at transcript level

- If want to count at transcript level, many more reads now ambiguous and will be discarded

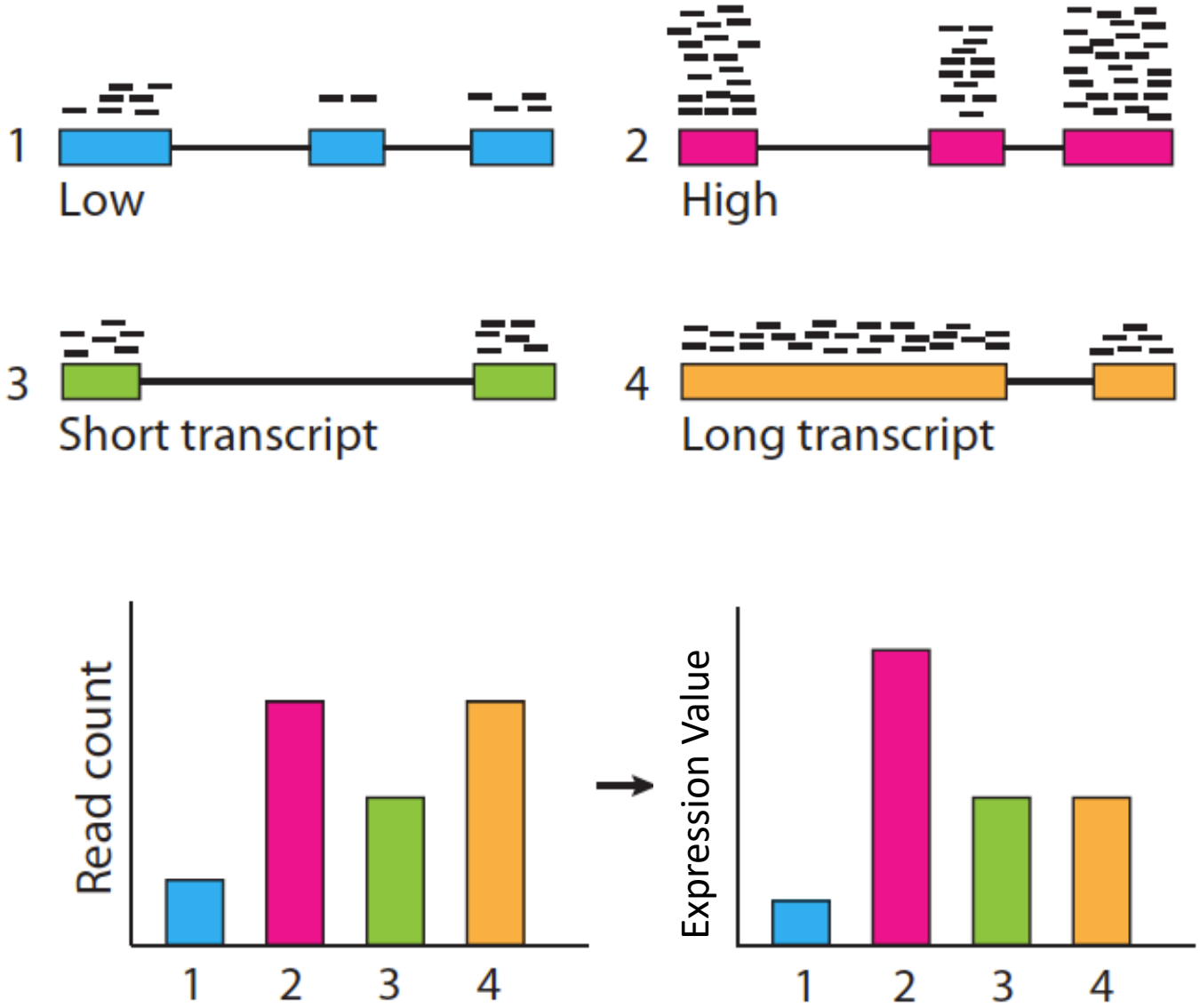


http://www.cs.cmu.edu/~maschulz/projects.html

Problems with STAR/featureCounts at gene level:

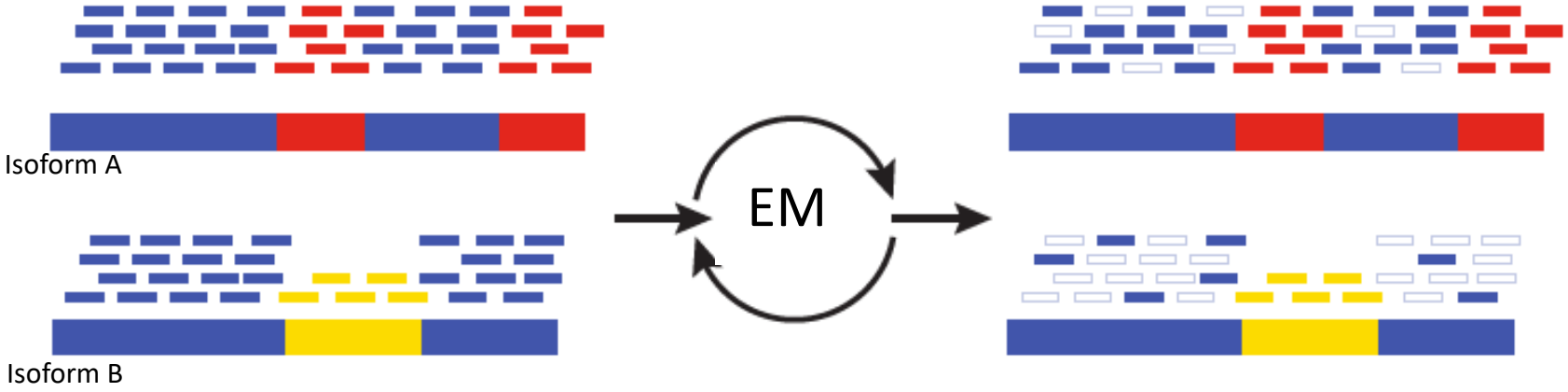
1. Multimapping reads not used, leading to underestimation of gene abundances, particularly for genes with more shared sequence
2. A small percentage of genes may not ever be quantifiable using this method.
3. Genes that change relative isoform usage can have erroneous results due to changes in isoform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

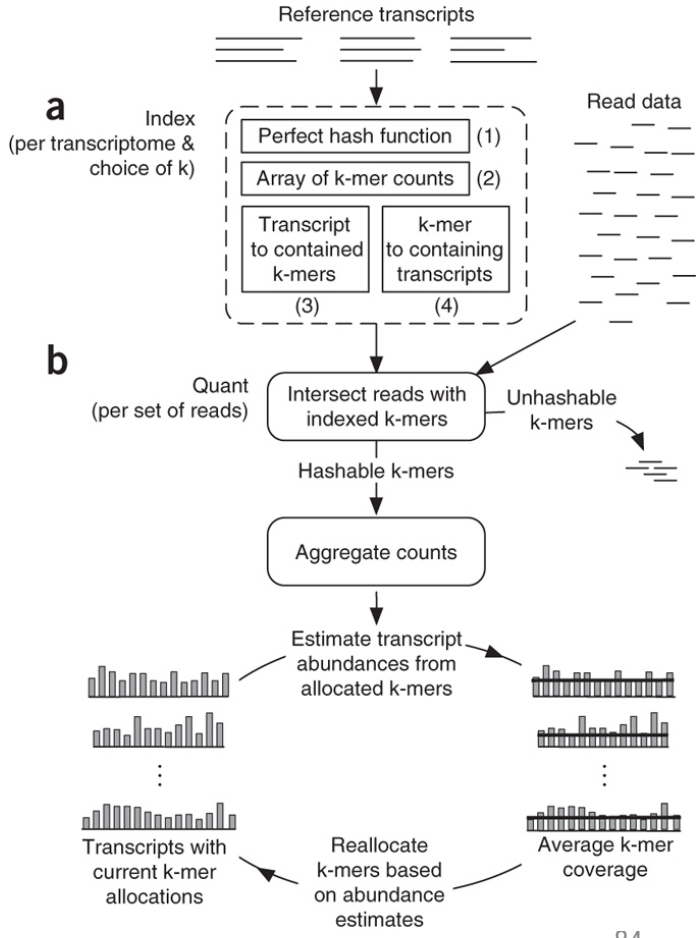
slide from Brian Haas; title modified

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

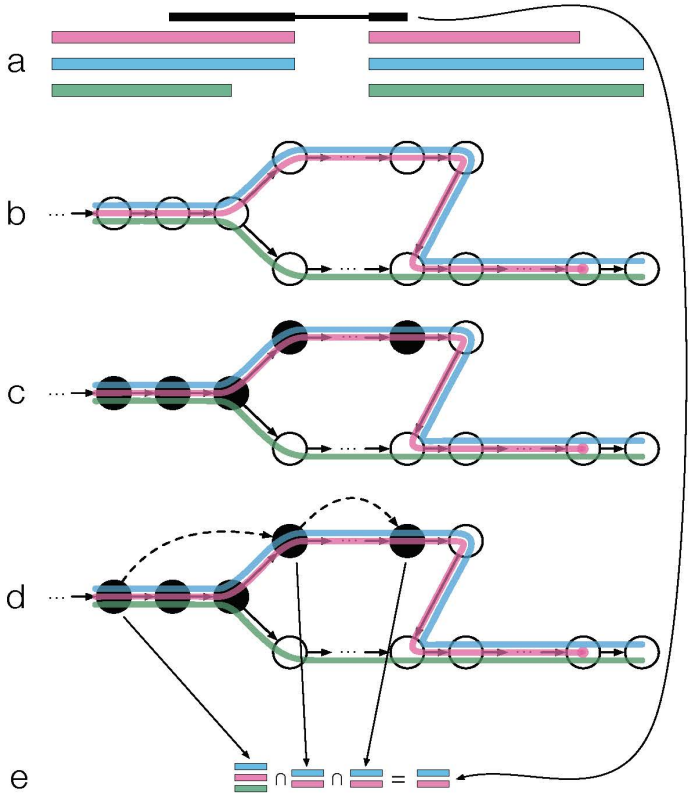
Radically new transcript counting programs have recently come out...

- 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!



Radically new 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.

Outline

3. Transcriptomic analysis methods and tools

a. Transcriptome Analysis; aspects 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**

De novo transcriptome assembly - **Trinity**

How do I learn more about these steps?

- Your lab will go through some of these steps on a very 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

Outline

3. Transcriptomic analysis methods and tools

a. Transcriptome Analysis; aspects 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

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 assemblies can be done, but are usually challenging

5.R is an excellent language to learn, if you are interested in performing in-depth statistical analyses for differential gene expression analysis

- Not within the scope of this lecture/lab section
- We do offer a long RNA-Seq workshop that covers the “HPCBio” RNA-Seq pipeline: http://hpcbio.illinois.edu/hpcbio-workshops

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!