Announcement

Feb 3 Notify the selected paper(s)

From Jan 29, bring your laptop to the class

The paper(s) should be published in 2022 and after, on the journals I listed, and use high throughput data.

RNA-Seq

Modified from Jessica Holmes

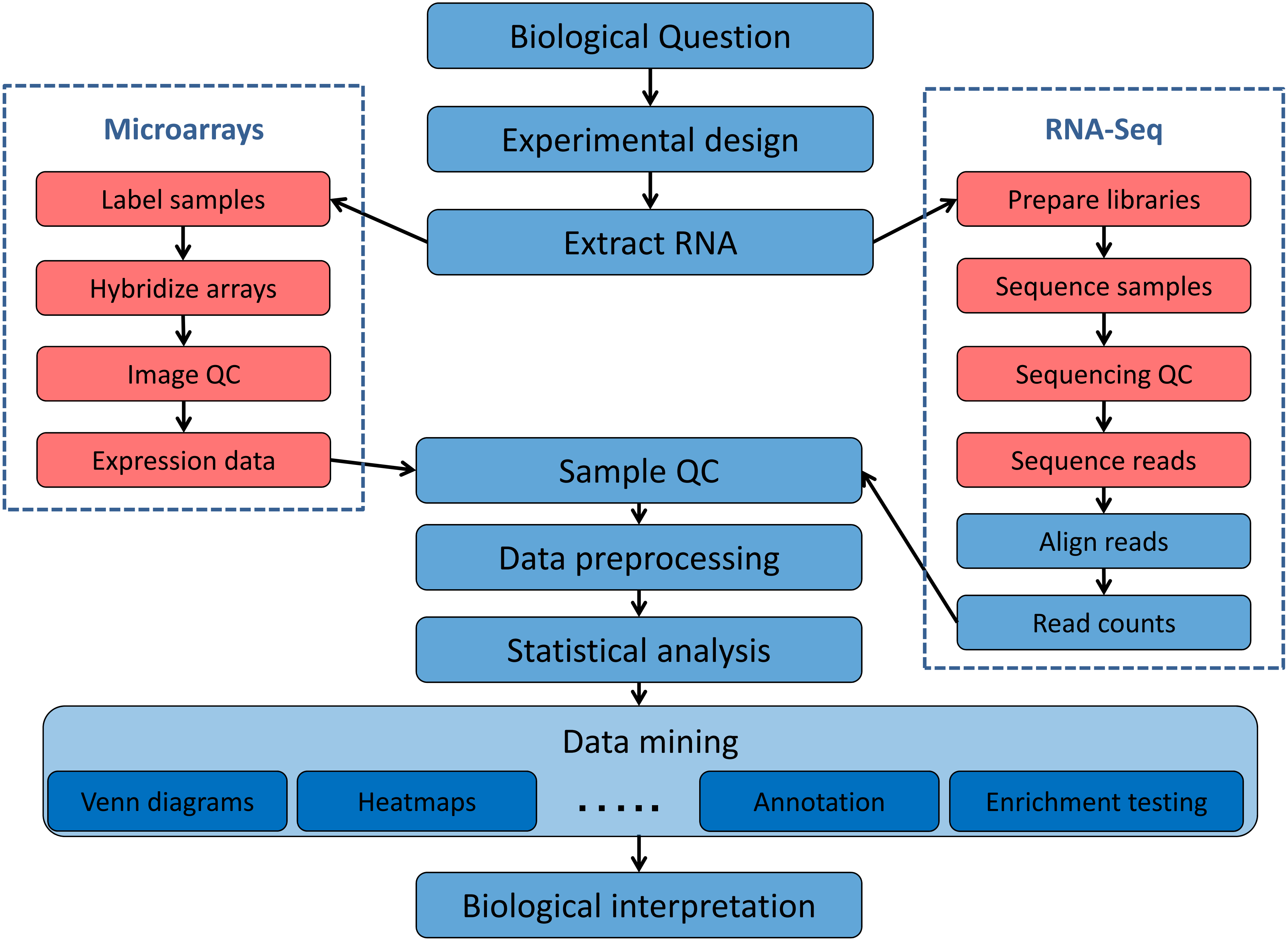
https://wiki.illinois.edu/wiki/display/HPCBio/RNA-Seq+Analysis+-+Spring+2020

Outline

1. Getting the RNA-Seq data: from RNA -> Sequence data
2. Experimental and practical considerations
3. Transcriptomic analysis methods and tools
    a. Transcriptome Assembly
    b. Differential Gene expression

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

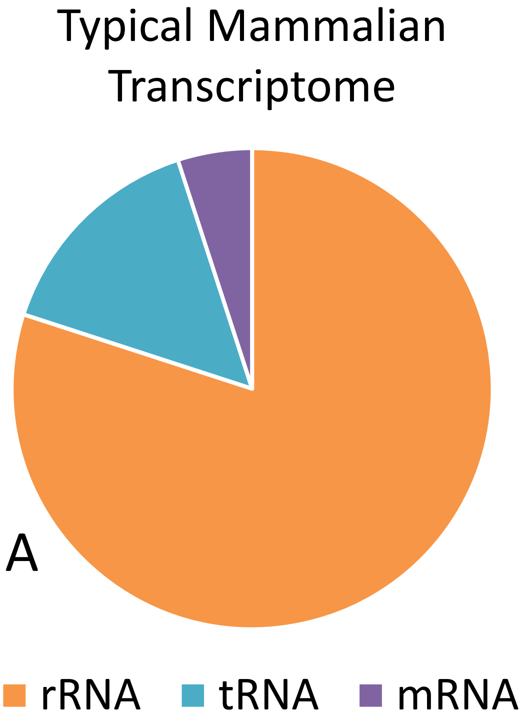

Biological Question
Experimental design
Extract RNA
Microarrays
Label samples
Hybridize arrays
Image QC
Expression data
RNA-Seq
Prepare libraries
Sequence samples
Sequencing QC
Sequence reads
Align reads
Read counts
Sample QC
Data preprocessing
Statistical analysis
Data mining
Venn diagrams
Heatmaps
. . . . .
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, siRNA, snoRNA, etc.)

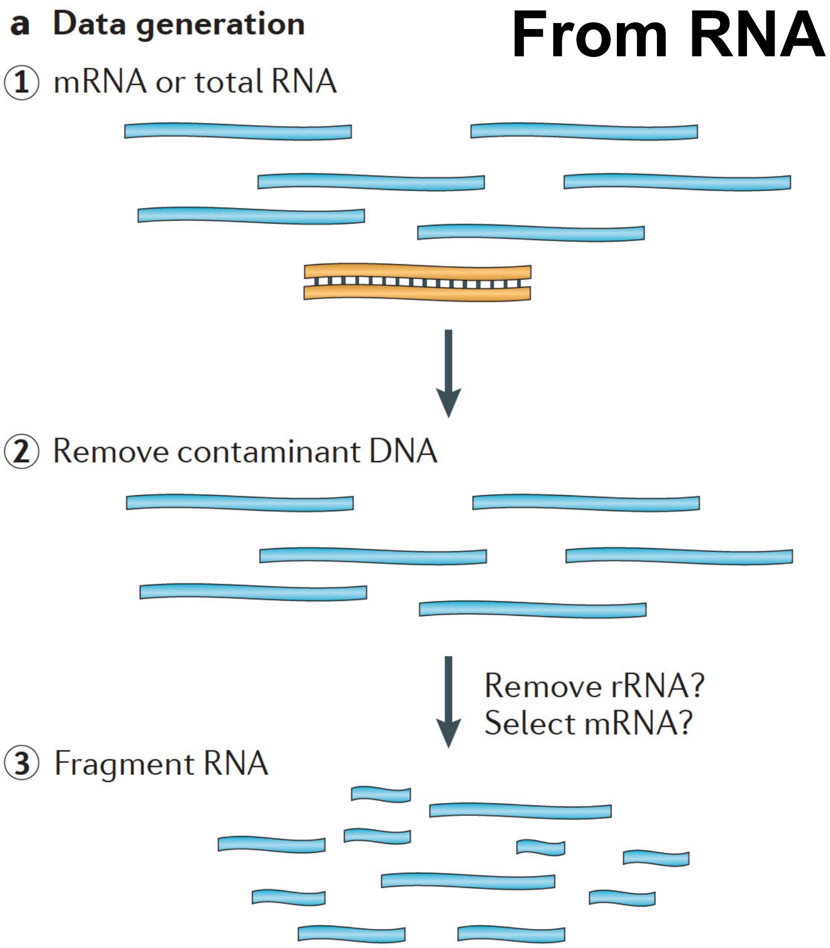
Removal of rRNA is almost always recommended

- Removal Methods:
  - poly-A selection (eukaryotes only)
  - rRNA depletion



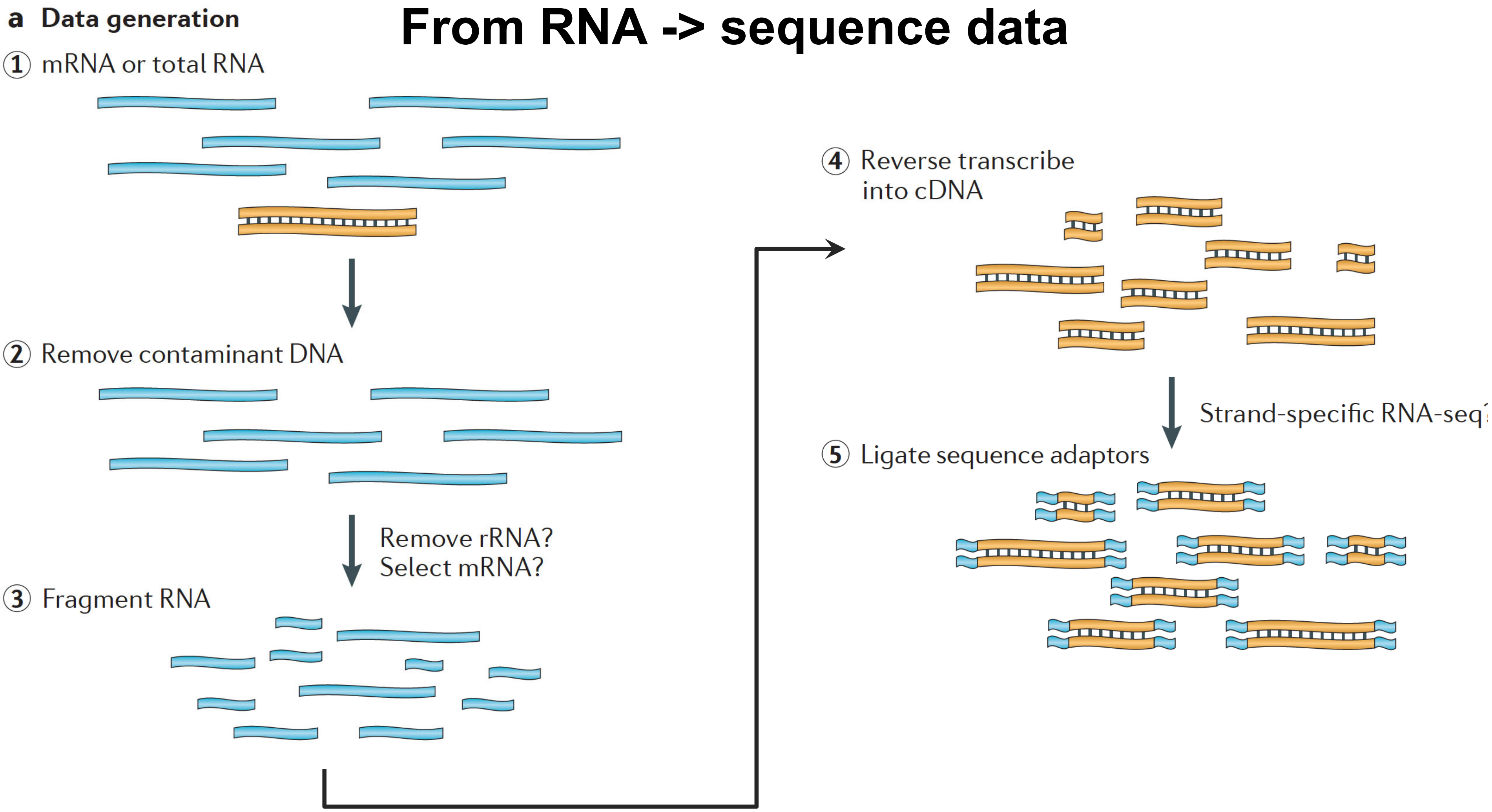
rRNA depletion captured more unique transcriptome features, whereas polyA+ selection outperformed rRNA depletion with higher exonic coverage and better accuracy of gene quantification (https://www.nature.com/articles/s41598-018-23226-4)

From RNA -> sequence data



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

From RNA -> sequence data



a Data generation

1 mRNA or total RNA

2 Remove contaminant DNA

Remove rRNA?
Select mRNA?

3 Fragment RNA

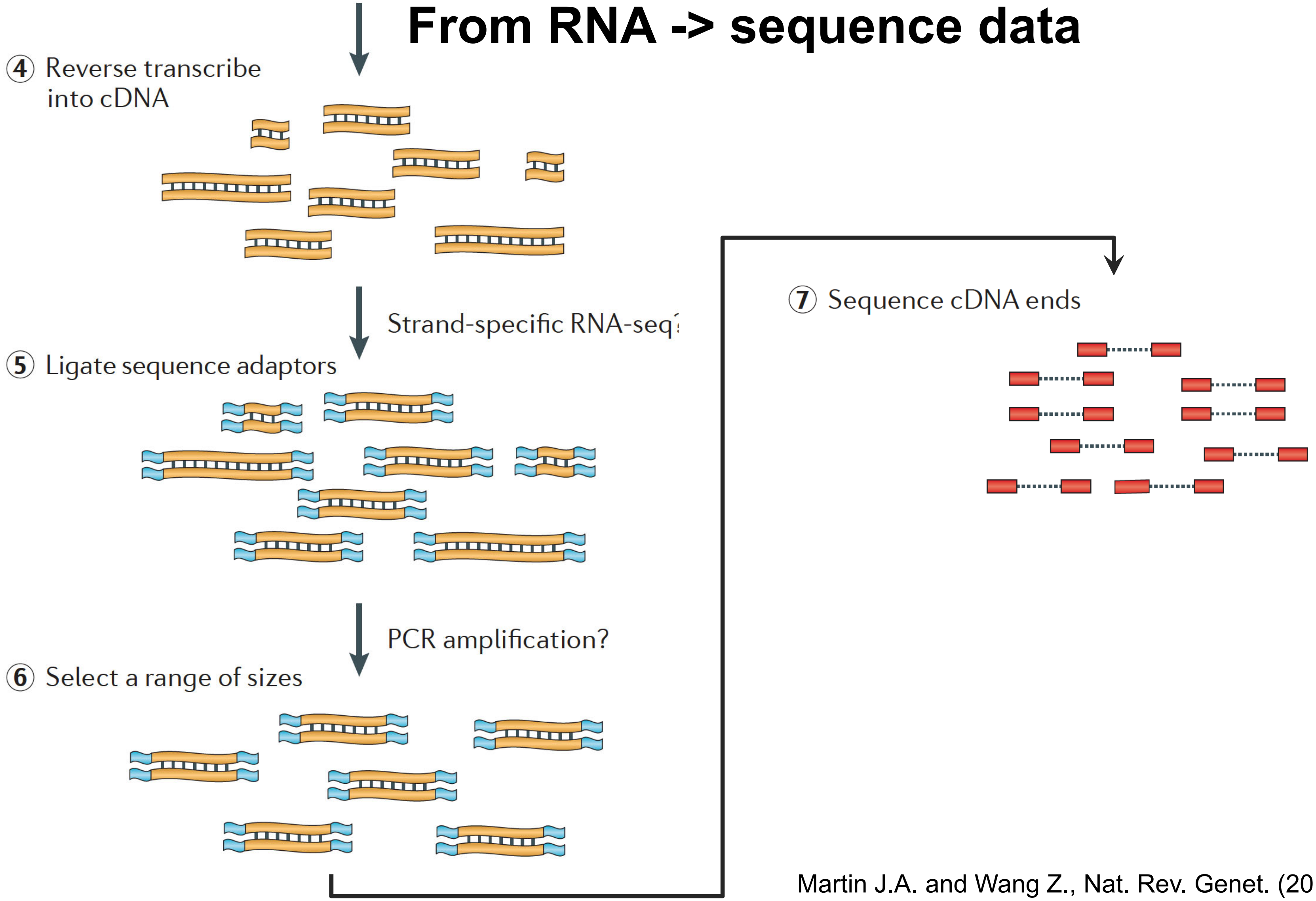
4 Reverse transcribe into cDNA

Strand-specific RNA-seq?

5 Ligate sequence adaptors

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

From RNA -> sequence data



④ Reverse transcribe into cDNA

Strand-specific RNA-seq?

⑤ Ligate sequence adaptors

PCR amplification?

⑥ Select a range of sizes

⑦ Sequence cDNA ends

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):

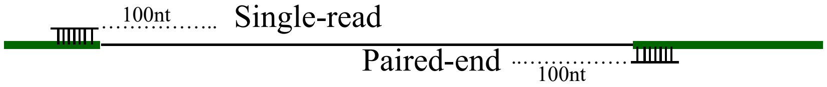
- **454** - pyrosequencing
- **SOLiD** – sequencing by ligation
- **Illumina** – sequencing by synthesis
- **Ion Torrent** – ion semiconductor
- **Pac Bio** – Single Molecule Real-Time sequencing, 1000 bp

3rd generation (2015)

- **Pac Bio** – SMRT, Sequel system, 20,000 bp
- **Nanopore** – ion current detection

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

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

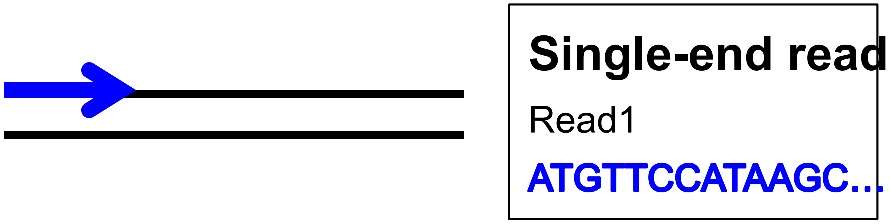
General Outline

1. Getting the RNA-Seq data: from RNA -> Sequence data
2. **Experimental and Practical considerations**
3. Transcriptomic analysis methods and tools
    a. Transcriptome Assembly
    b. Differential Gene expression

Considerations for...

Differential Gene Expression

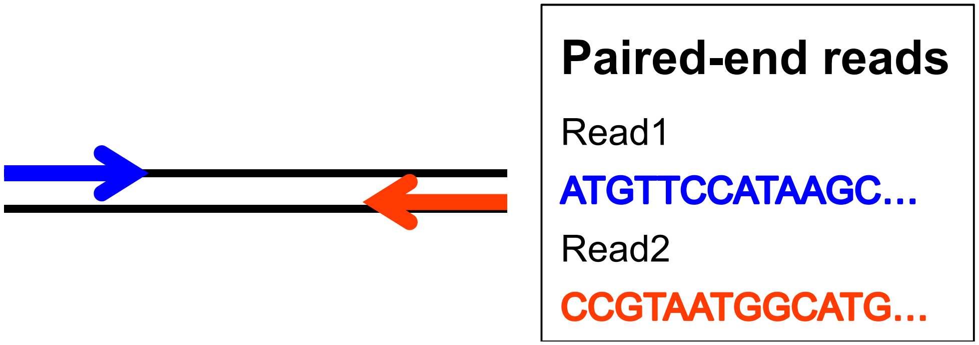
- Keep biological replicates separate
- Poly-A enrichment is generally recommended
  - Unless you’re interested in non-coding RNA!
- Remove ribosomal RNA (rRNA)
  - Unless you’re interested in rRNA!
- Usually single-end (SE) is enough
  - Paired-end (PE) may be recommended for more complex genomes



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

General Outline

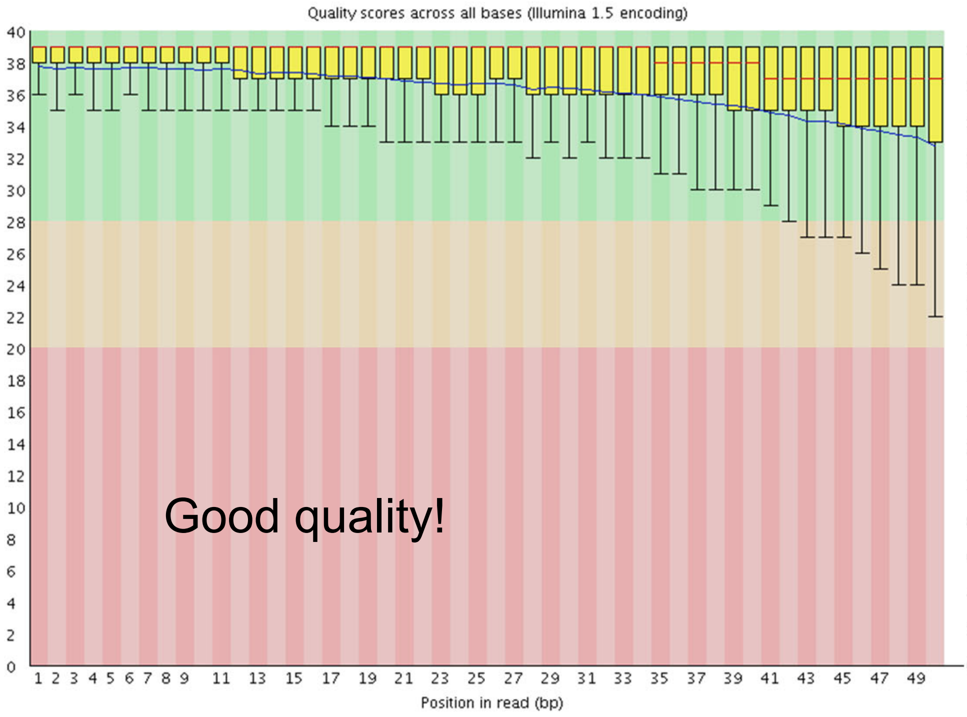
1. Getting the RNA-Seq data: from RNA -> Sequence data
2. Experimental and Practical considerations
3. Transcriptomic analysis methods and tools
    a. Transcriptome Assembly
    b. Differential Gene expression

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

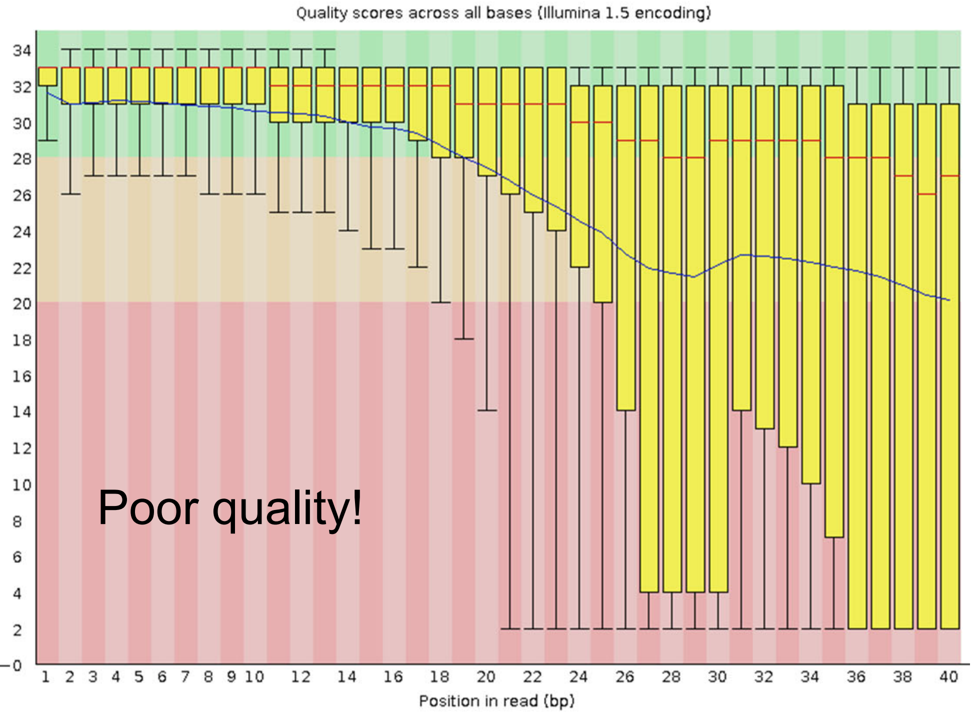
Software called **FASTQC**

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

FASTQC: Quality Scores



Good quality!



Poor quality!

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

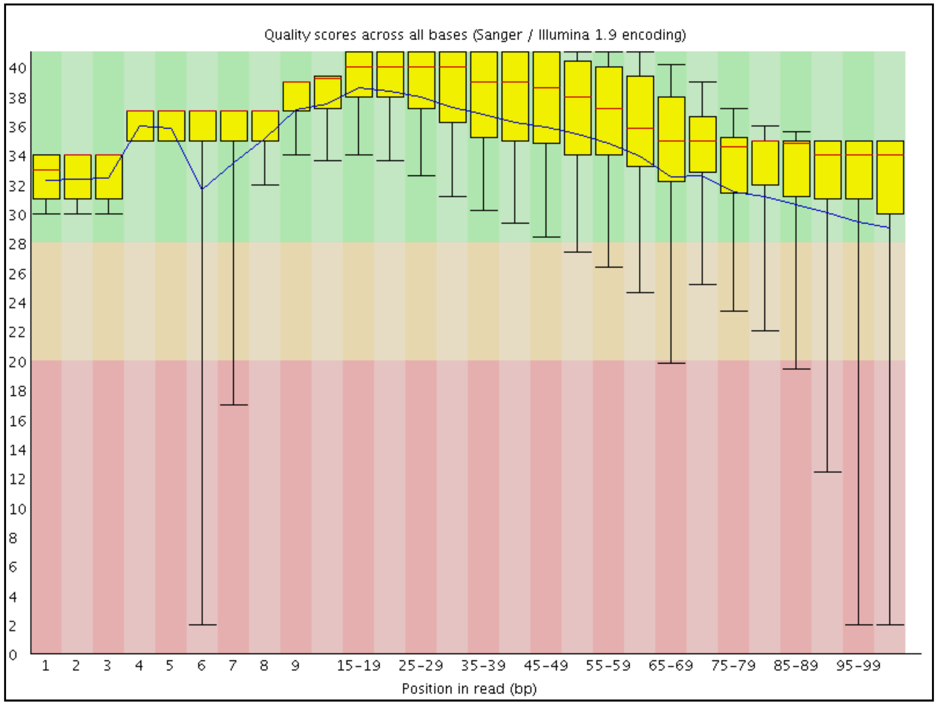
What do I do when FastQC calls my data poor?

- ✧ 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

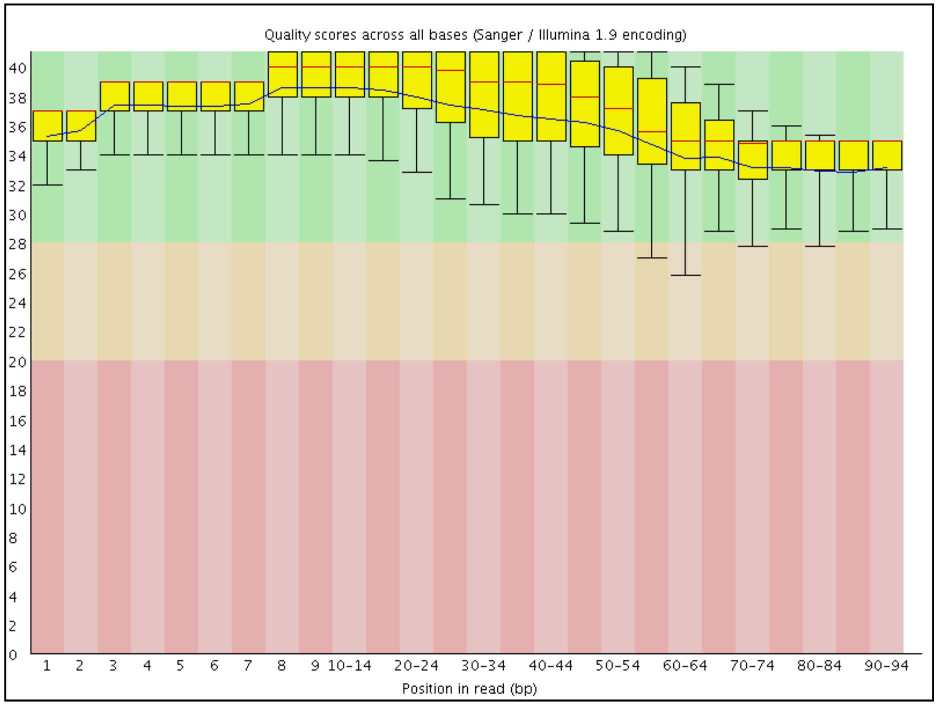
Transcriptome Analysis

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 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

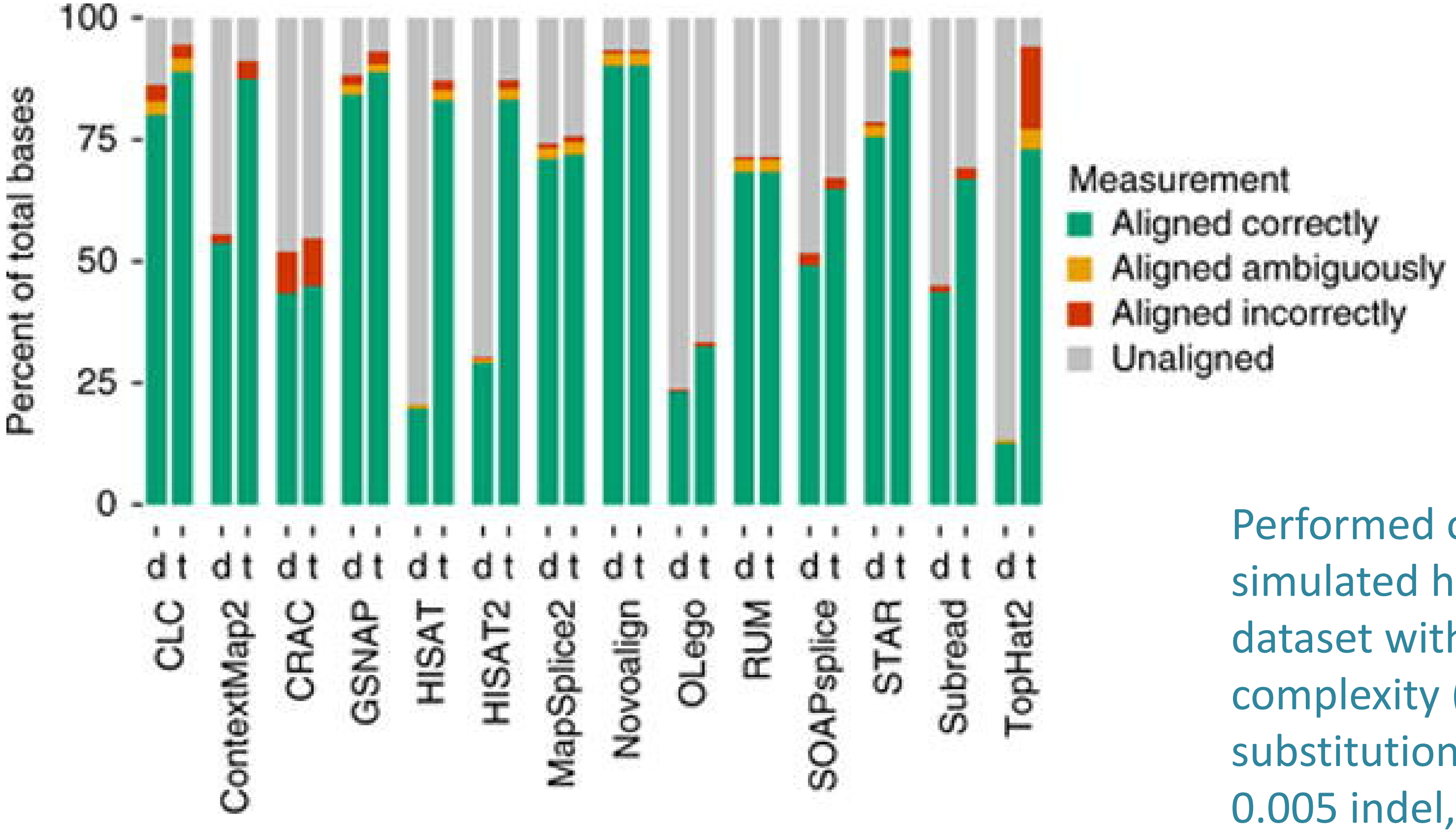
Transcriptome Analysis

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!!!

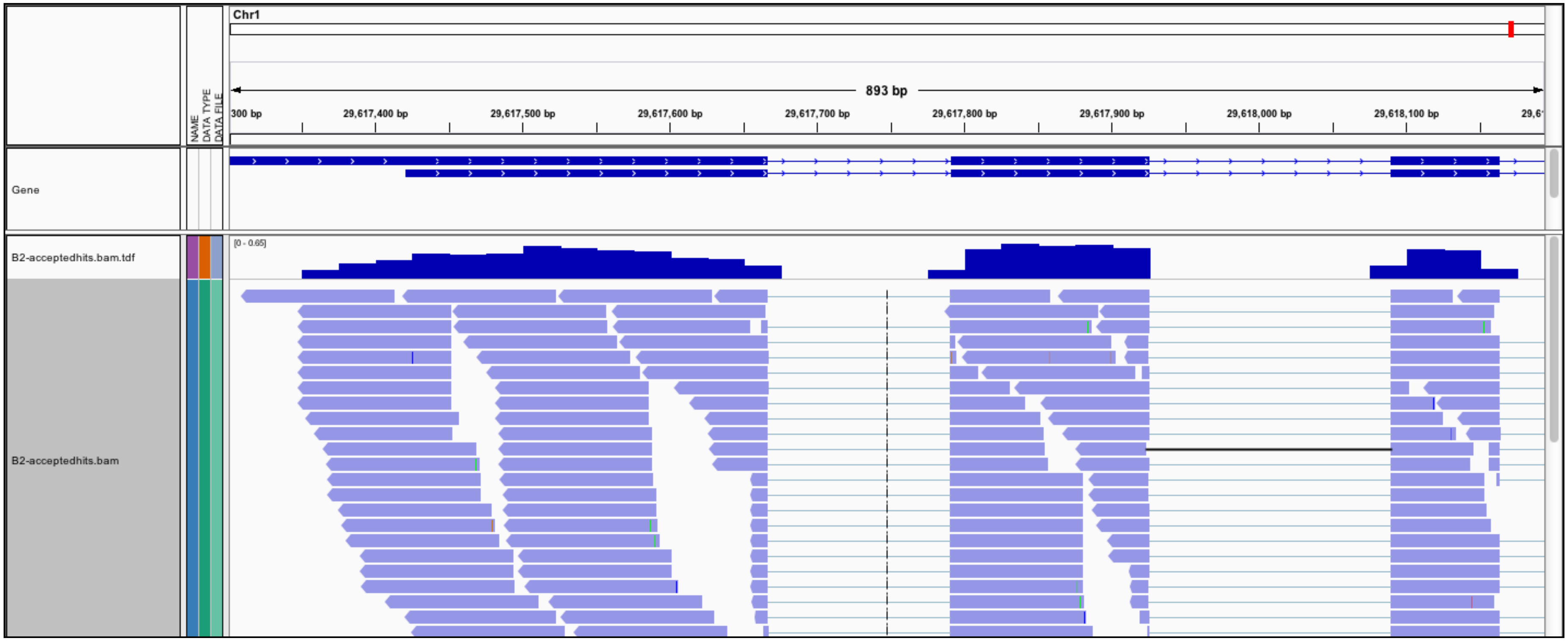


Performed on simulated human dataset with high complexity (0.03 substitution, 0.005 indel, 0.02 error)

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

      ✧ 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

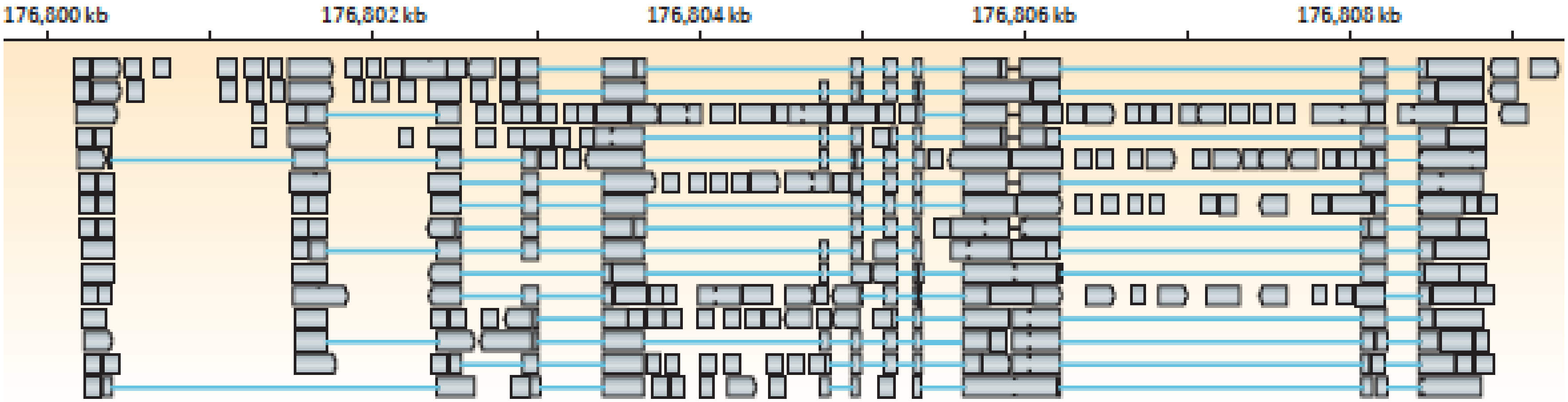
Used when the genome reference sequence is known, and:

- 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

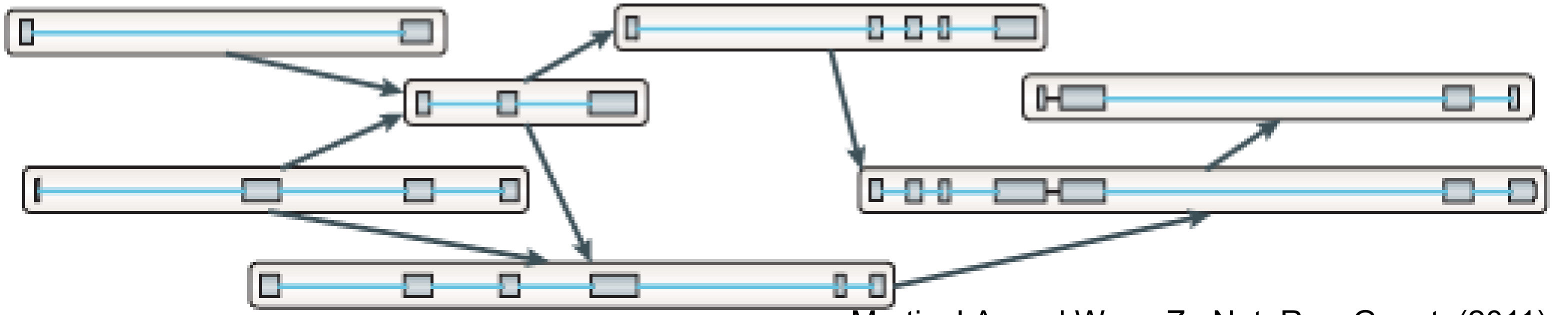
Transcriptome Assembly

a. Splice align reads to genome

Reference-based assembly



b. Build graph representing alternative splicing events

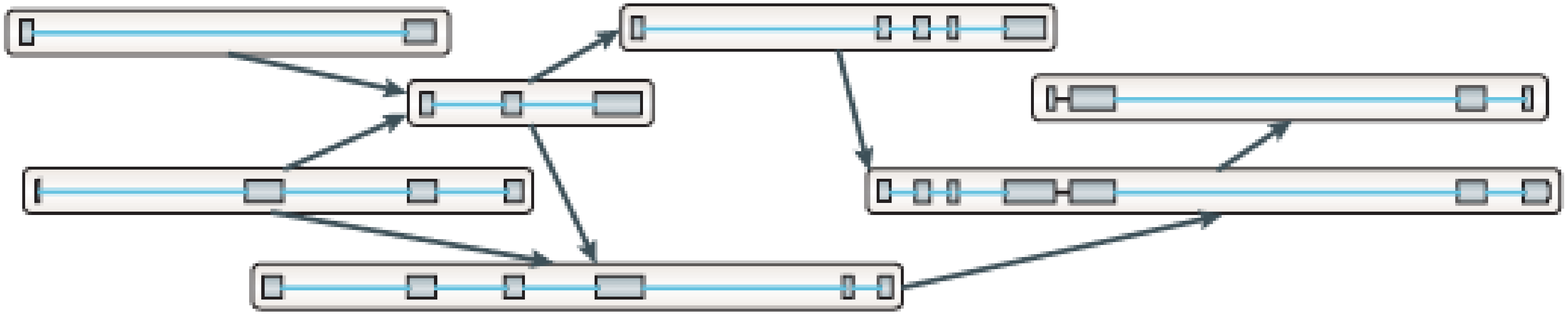


Martin J.A. and Wang Z., Nat. Rev. Genet. (2011) 12:671–682

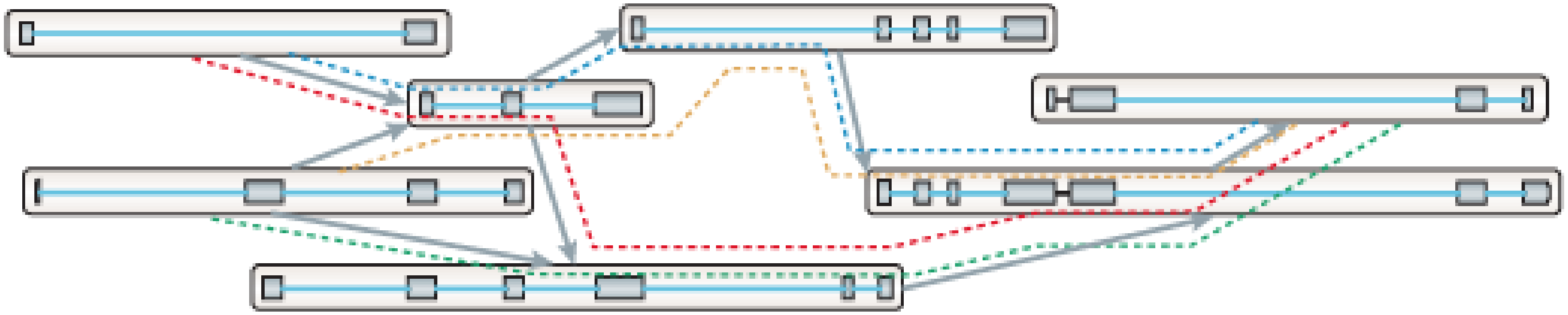
Transcriptome Assembly

Reference-based assembly

b. Build graph representing alternative splicing events



c. Traverse the graph to assemble variants

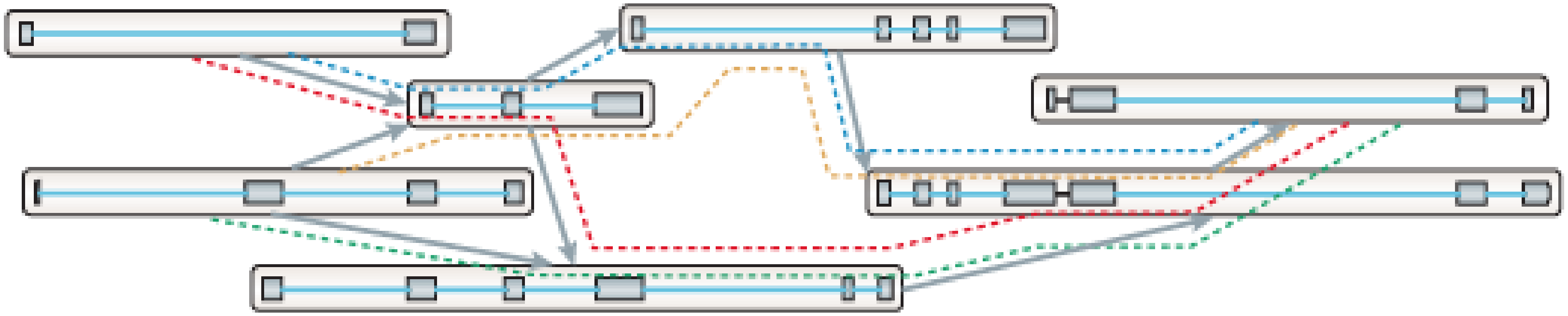


Martin J.A. and Wang Z., Nat. Rev. Genet. (2011) 12:671–682

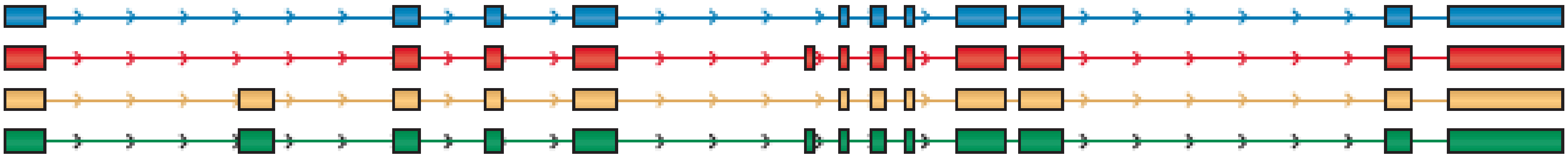
Transcriptome Assembly

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

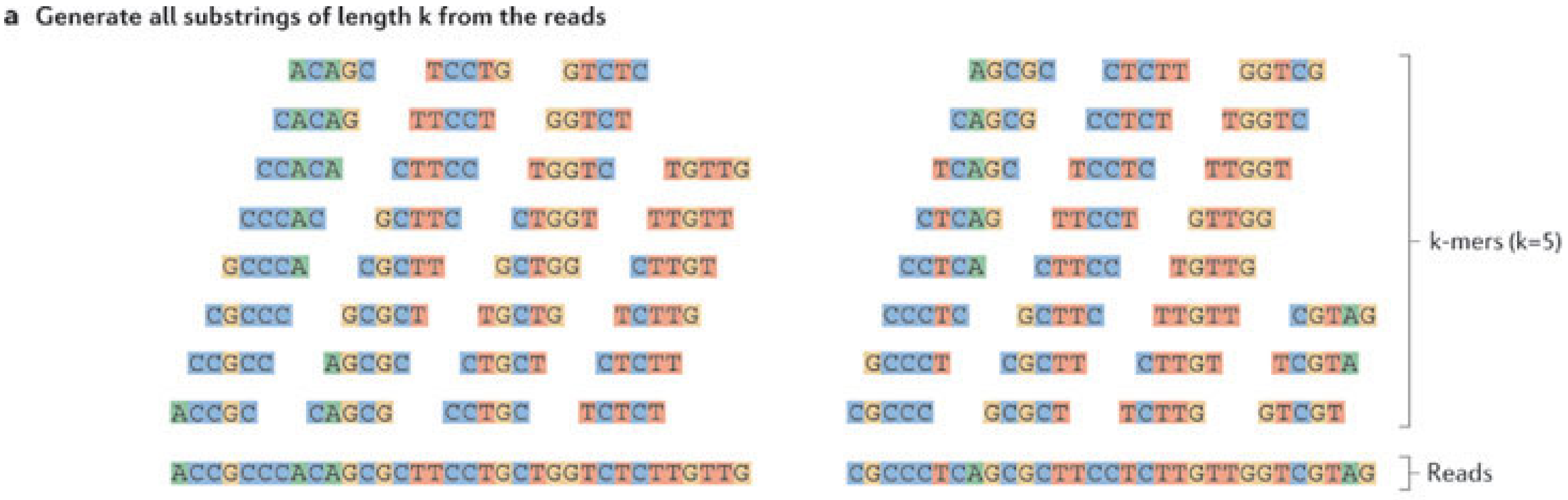
De novo assembly

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

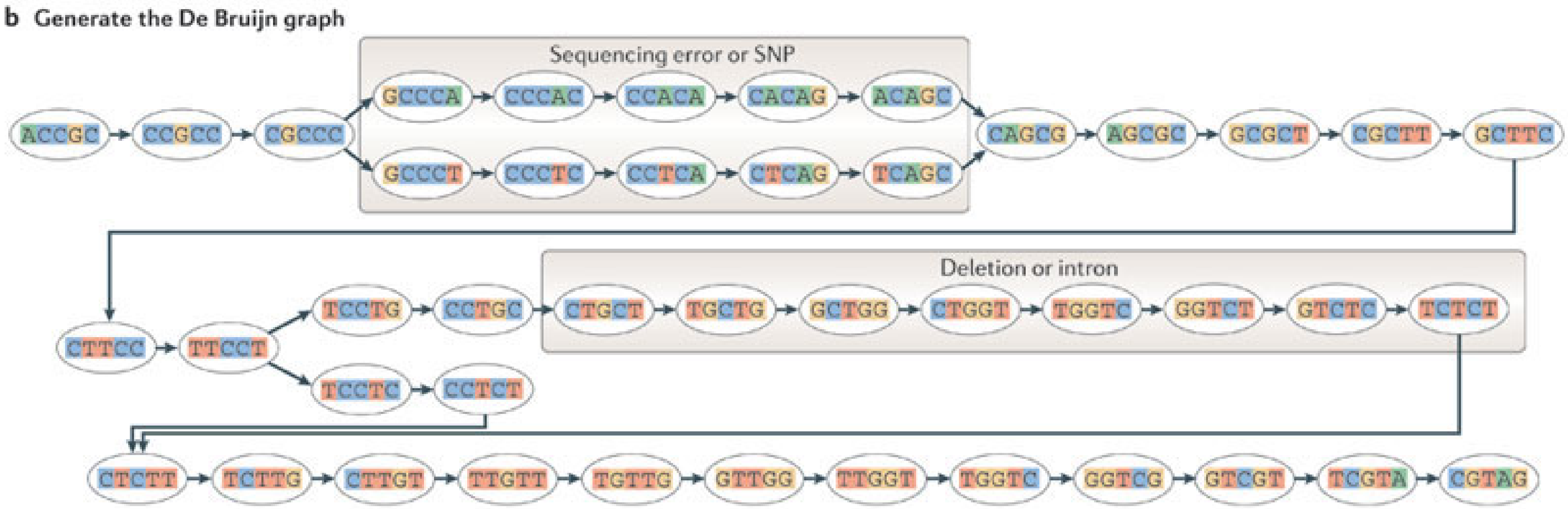
De novo assembly (De Bruijn graph construction)



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

Transcriptome Assembly

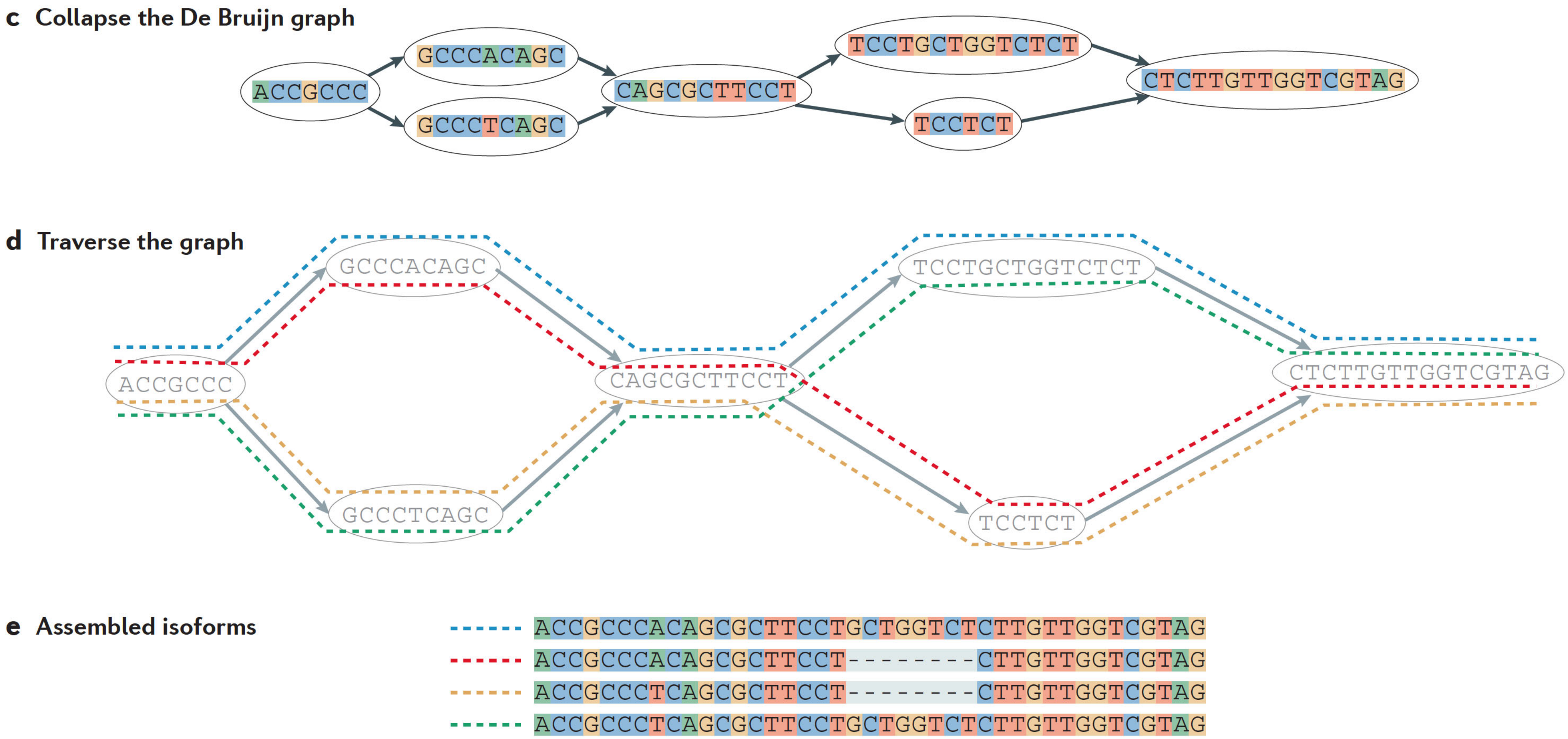
De novo assembly (De Bruijn graph construction)



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

Transcriptome Assembly

De novo assembly (De Bruijn graph construction)



c Collapse the De Bruijn graph

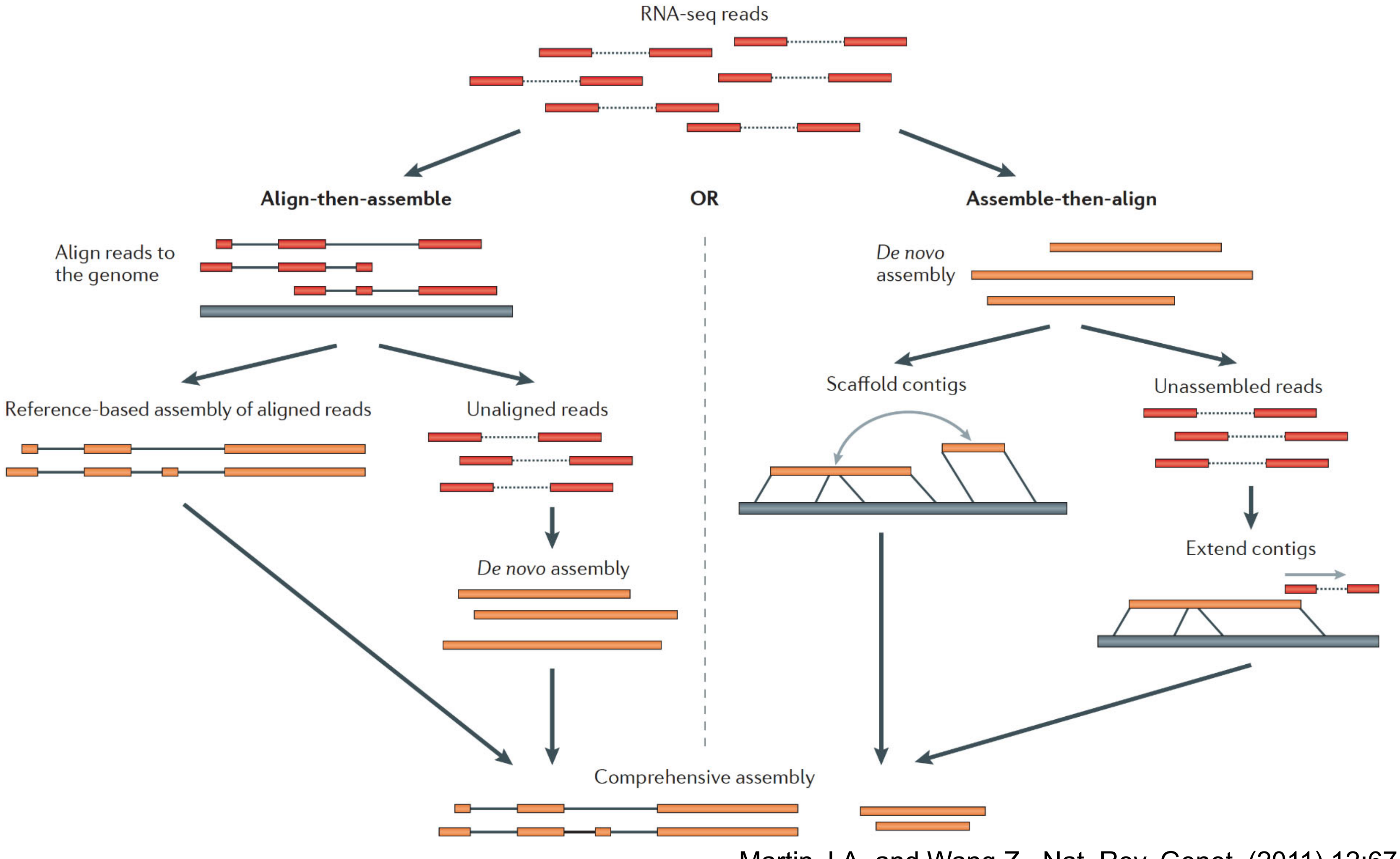
d Traverse the graph

e Assembled isoforms

ACCGCCCACAGCGCTTCCTGCTGGTCTCTTGTTGGTCGTAG
ACCGCCCACAGCGCTTCCT--------CTTGTTGGTCGTAG
ACCGCCCTCAGCGCTTCCT--------CTTGTTGGTCGTAG
ACCGCCCTCAGCGCTTCCTGCTGGTCTCTTGTTGGTCGTAG

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?

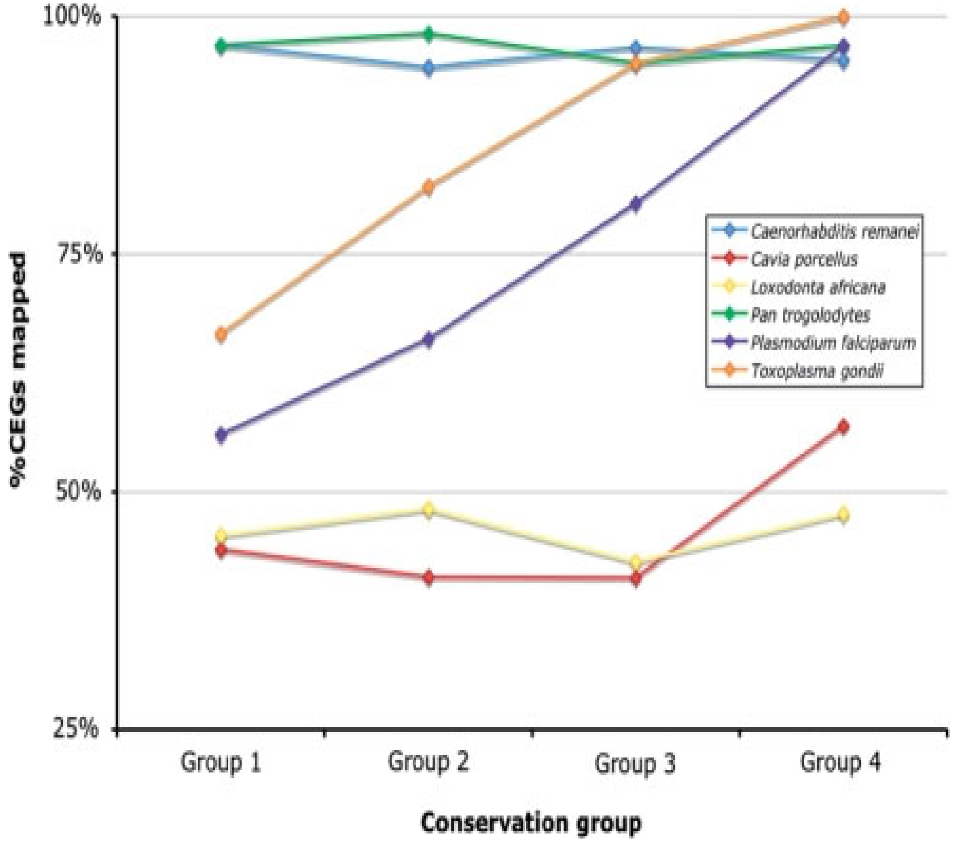
Tools for Evaluating Assembly: *using the information you have*

- 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

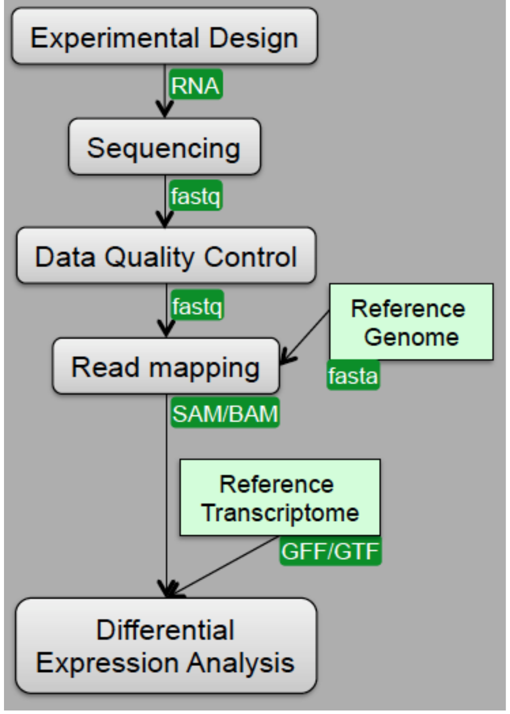
   - ✧ htseq & feature-counts return raw read counts
     - ✧ Required for R programs like DESeq & EdgeR
   - ✧ StringTie returns FPKM normalized counts for each gene

Differential Gene Expression

Options for DGE analysis

RNA-Seq

- Reference genome
- Reference transcriptome



Experimental Design

Sequencing

Data Quality Control

Read mapping

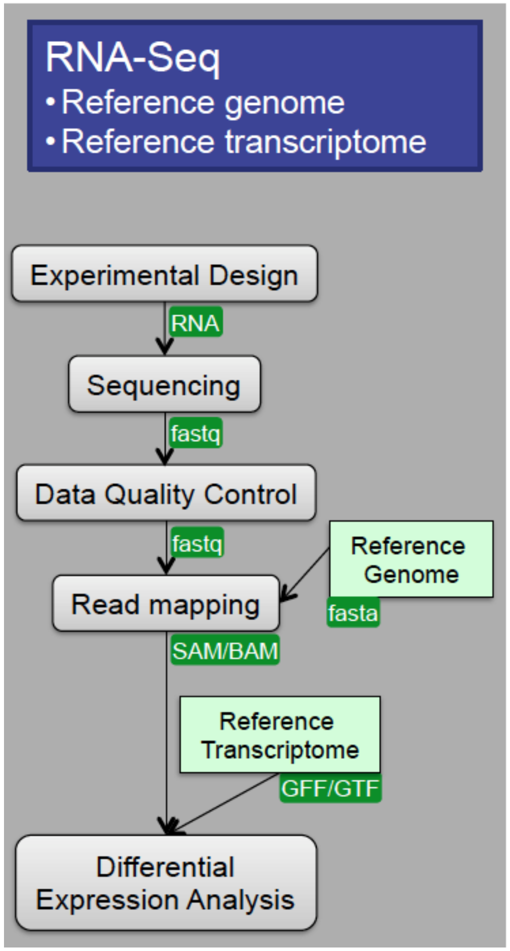
Reference Genome

Reference Transcriptome

Differential Expression Analysis

Differential Gene Expression

Options for DGE analysis



RNA-Seq
- Reference genome
- Reference transcriptome

Experimental Design

RNA

Sequencing

fastq

Data Quality Control

fastq

Reference Genome

fasta

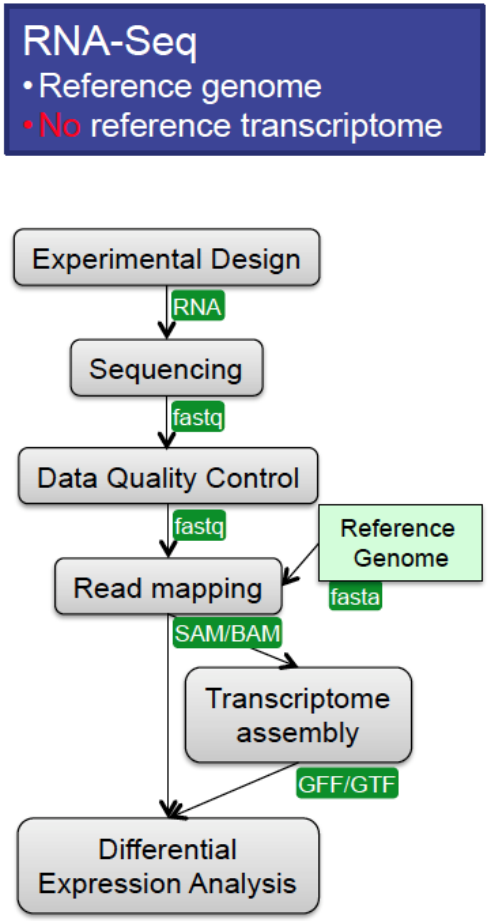
Read mapping

SAM/BAM

Reference Transcriptome

GFF/GTF

Differential Expression Analysis



RNA-Seq
- Reference genome
- No reference transcriptome

Experimental Design

RNA

Sequencing

fastq

Data Quality Control

fastq

Reference Genome

fasta

Read mapping

SAM/BAM

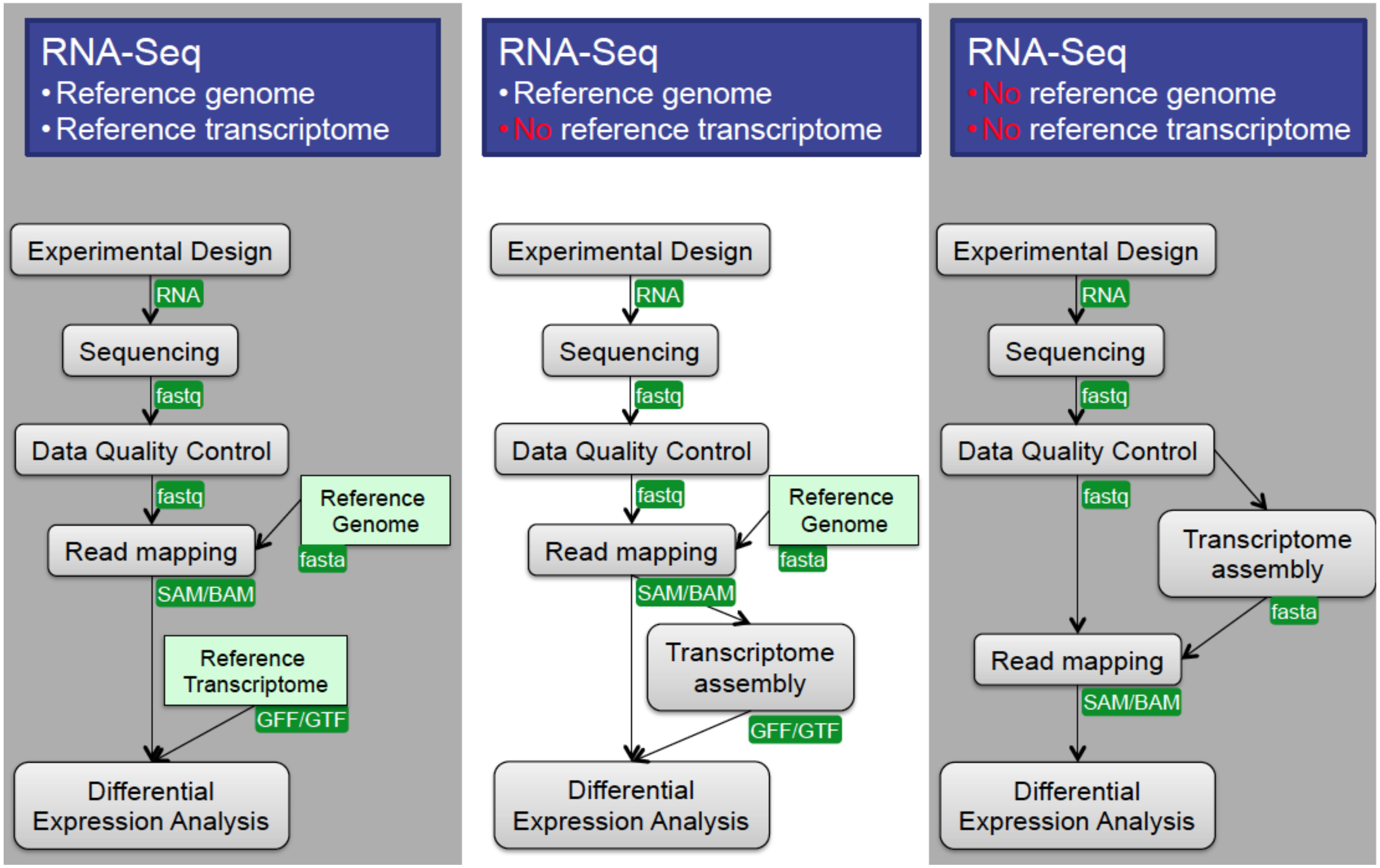
Transcriptome assembly

GFF/GTF

Differential Expression 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
- ... and more!

How does one pick the right tools?

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**

TIPs

1. When in doubt “Google it” and ask questions.

- http://www.biostars.org/ - Biostar (Bioinformatics explained)
- http://seqanswers.com/ - SEQanswers (the next generation sequencing community)

2. 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.

3. http://hpcbio.illinois.edu/hpcbio-workshops

2nd In-Class question

What are the main steps to analyze ChIP-seq data? Please list at least three steps and the tools that you can use.

Due 6pm today. Submit your answers at webcourses.