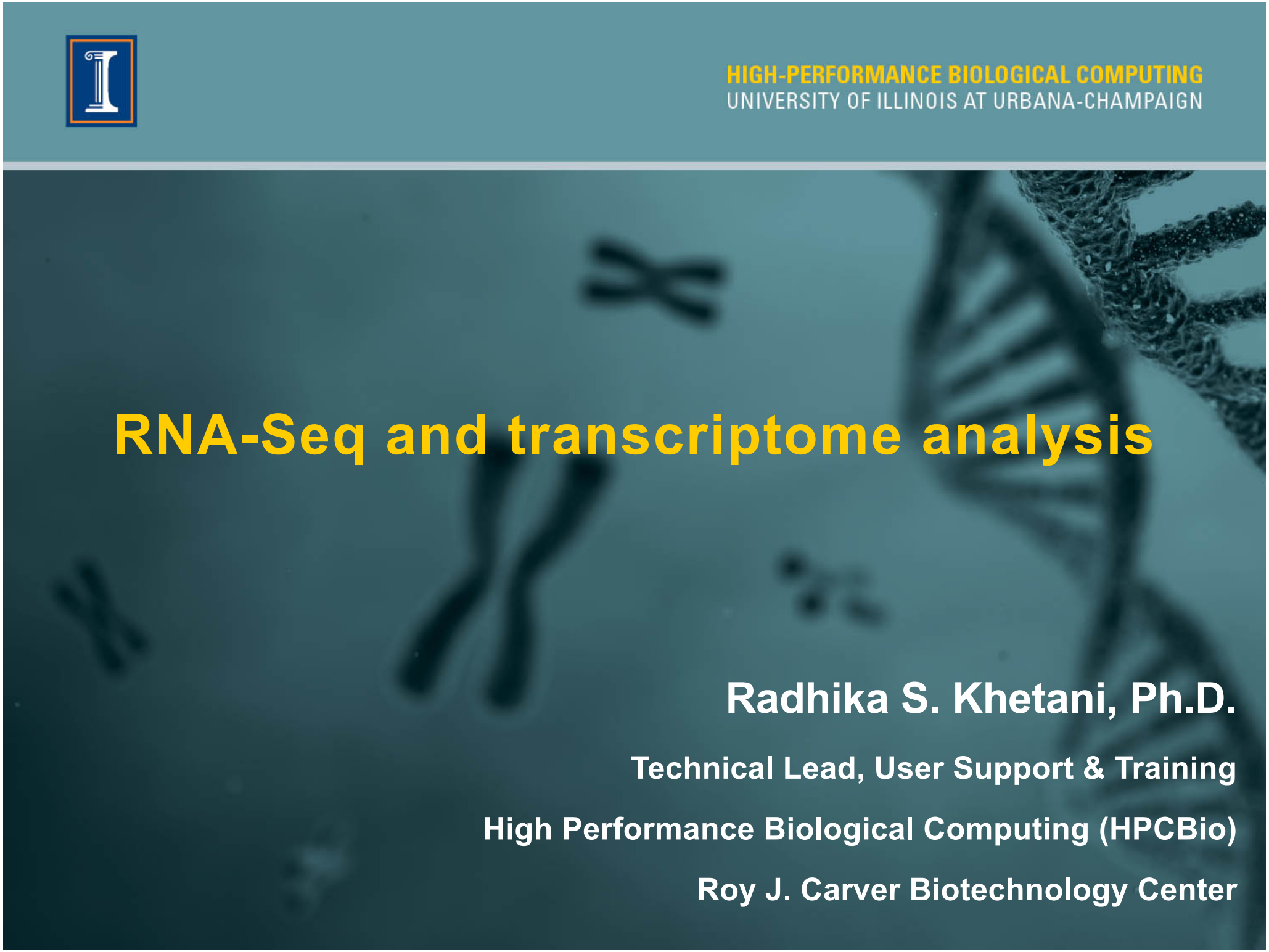


HIGH-PERFORMANCE BIOLOGICAL COMPUTING
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

RNA-Seq and transcriptome analysis

Radhika S. Khetani, Ph.D.

Technical Lead, User Support & Training

High Performance Biological Computing (HPCBio)

Roy J. Carver Biotechnology Center

RNA-Seq or Transcriptome Sequencing

RNA-Seq

- It is the process of sequencing the transcriptome
- Its uses include –
  - Differential Gene Expression
    - ✧ Quantitative evaluation and comparison of transcript levels
  - Transcriptome assembly
    - ✧ Building the profile of transcribed regions of the genome, a qualitative evaluation.
  - Can be used to help build better gene models, and verify them using the assembly
  - Metatranscriptomics or community transcriptome analysis
  - Small RNA analysis

RNA-Seq or Transcriptome Sequencing

RNA-Seq

- It is the process of sequencing the transcriptome
- Its uses include –
  - **Differential Gene Expression**
    - ✧ Quantitative evaluation and comparison of transcript levels
  - **Transcriptome assembly**
    - ✧ Building the profile of transcribed regions of the genome, a qualitative evaluation.
  - Can be used to help build better gene models, and verify them using the assembly
  - Metatranscriptomics or community transcriptome analysis
  - Small RNA analysis

RNA-Seq or Transcriptome Sequencing

Sequencing technologies applicable to RNA-Seq

High throughput

- Illumina HiSeq 2500
- Illumina Next-Seq 500
- Illumina MiSeq
- Illumina X Ten

Illumina…

“Lower” throughput

- Roche 454

Low throughput

- Sanger

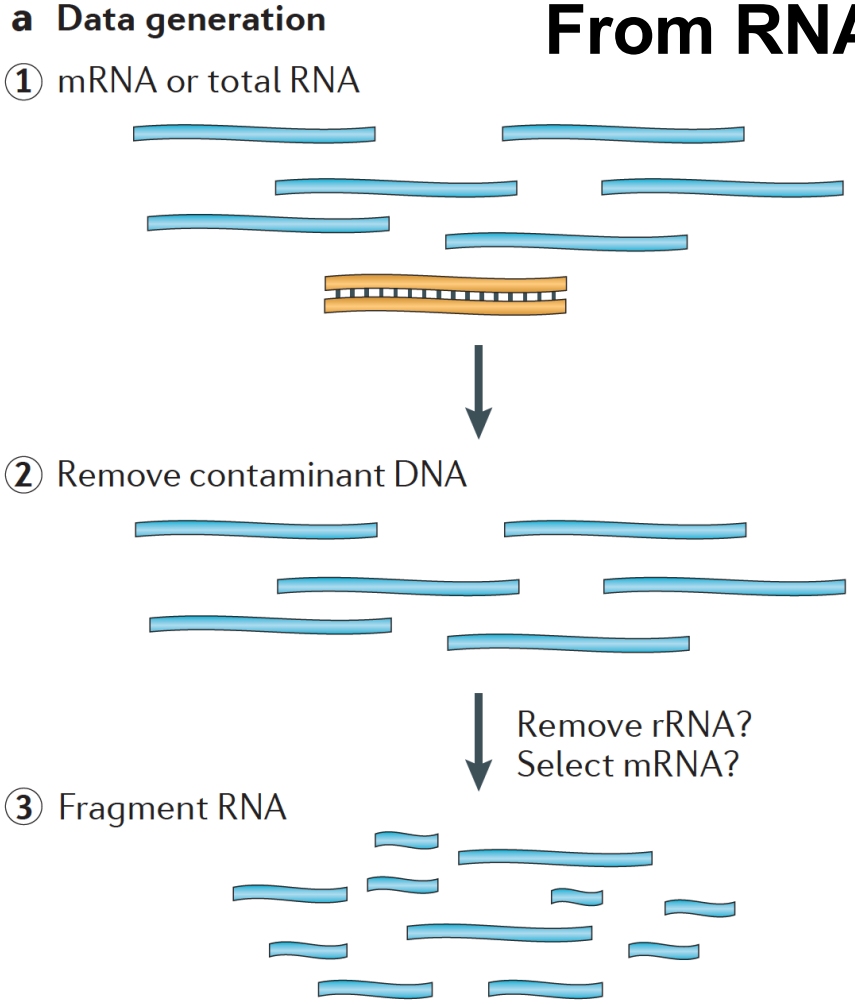
Outline

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

Outline

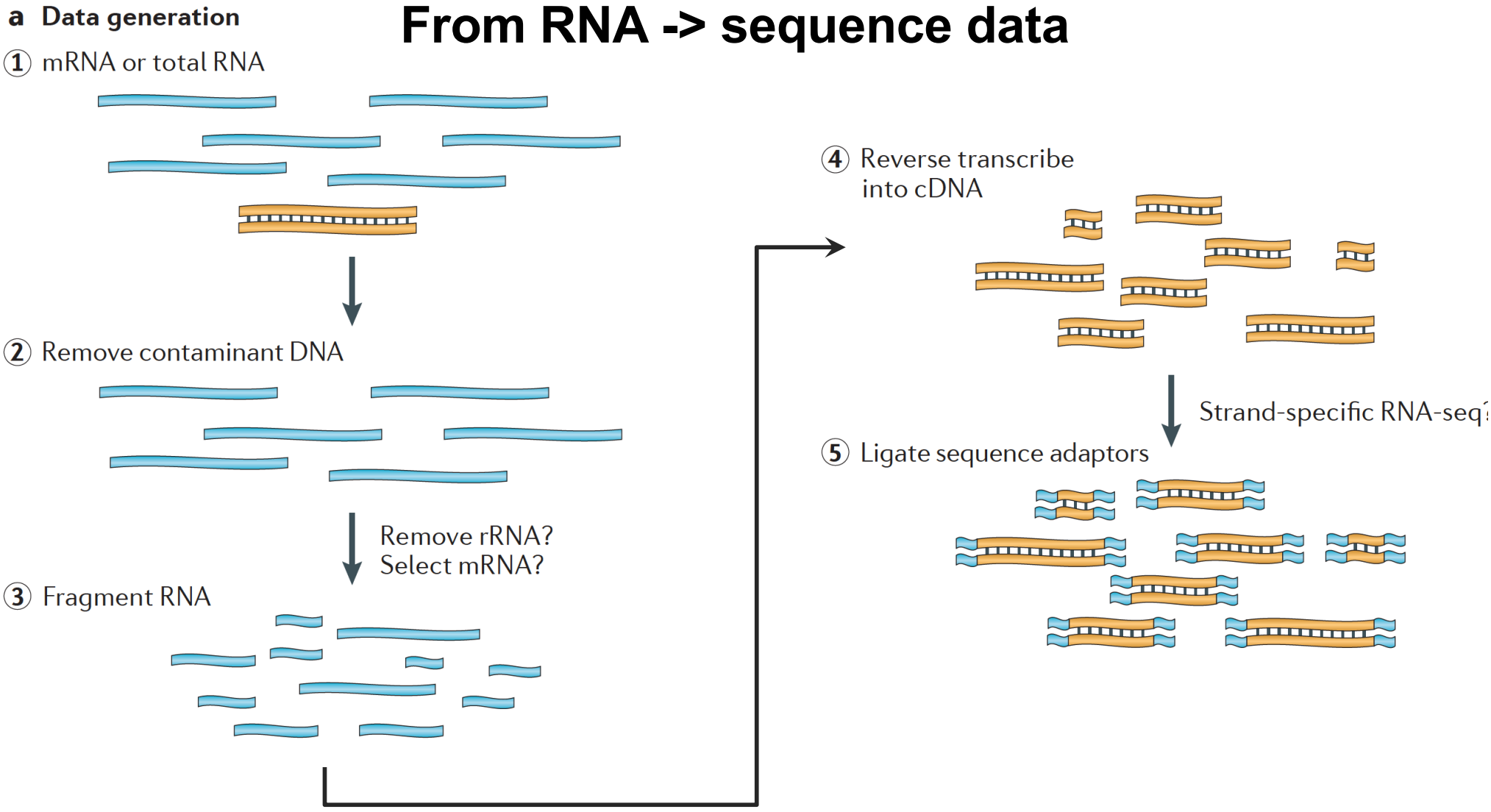
1. Getting the RNA-Seq data: from RNA -> Sequence data
2. Experimental and Practical considerations
3. Transcriptomic analysis methods and tools
    a. Assemblies
    b. Differential Gene expression

From RNA -> sequence data



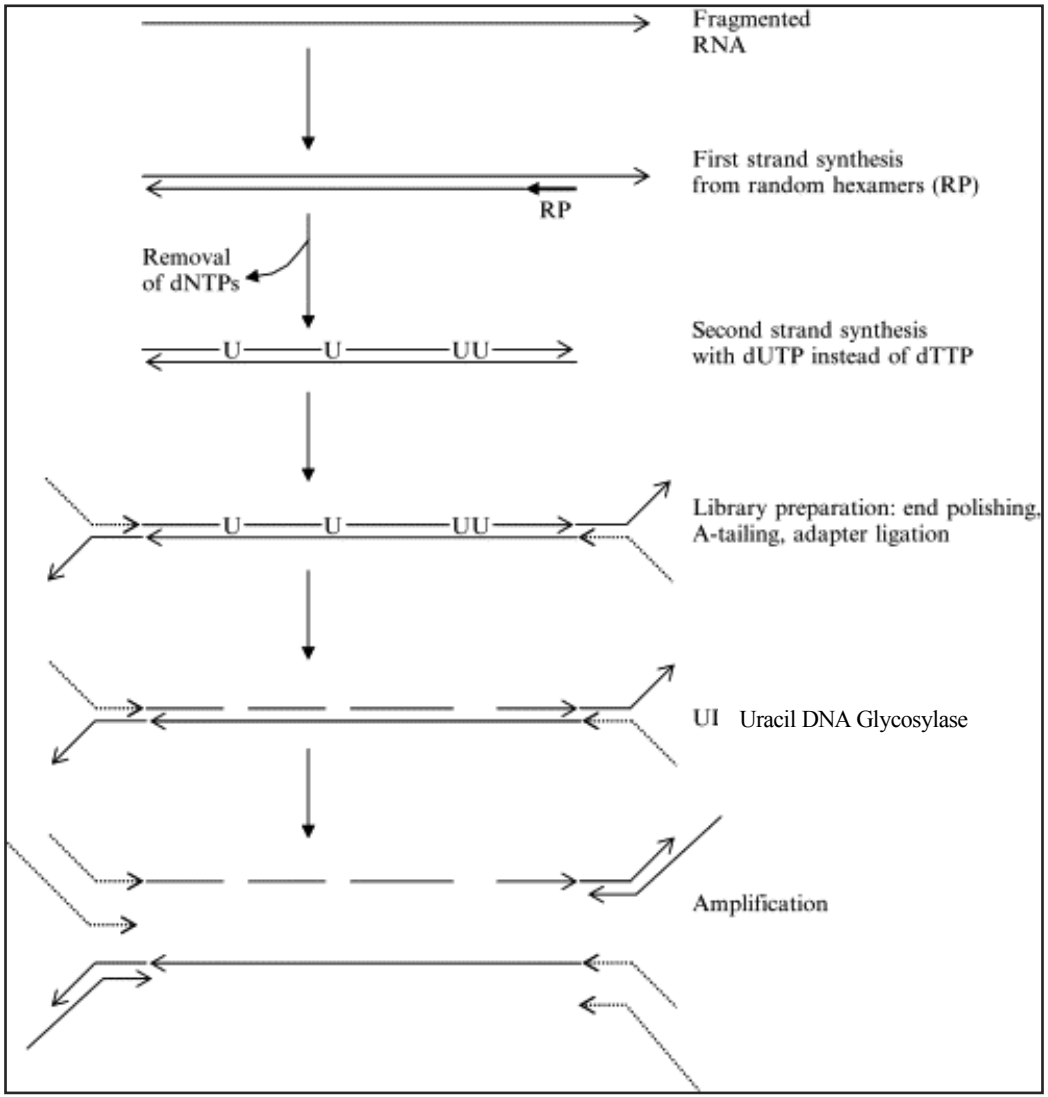
Martin J.A. and Wang Z., Nat. Rev. Genet. (2011) 12:671–682

From RNA -> sequence data



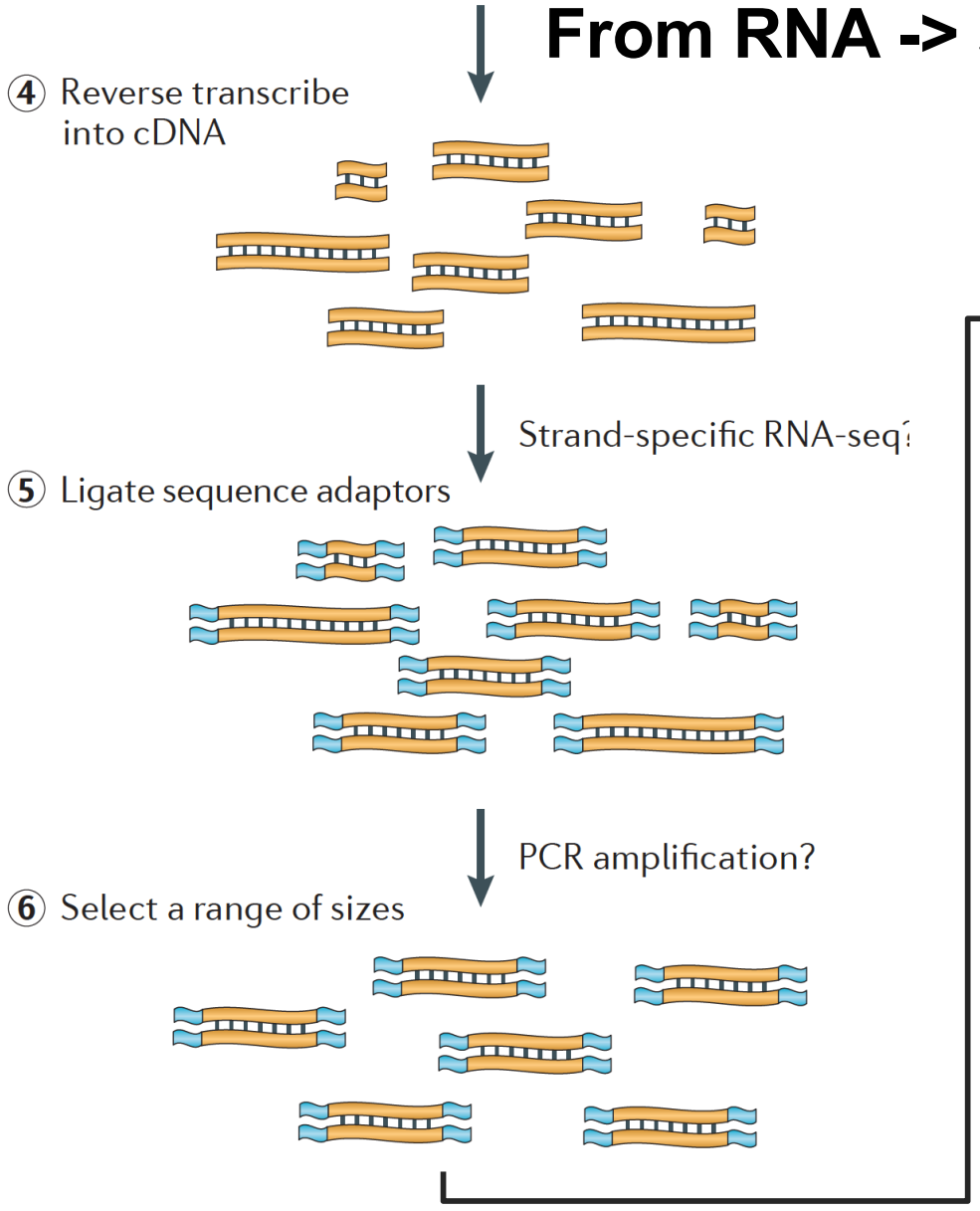
Martin J.A. and Wang Z., Nat. Rev. Genet. (2011) 12:671–682

From RNA -> sequence data



Borodina T., Methods in Enzymology (2011) 500:79–98

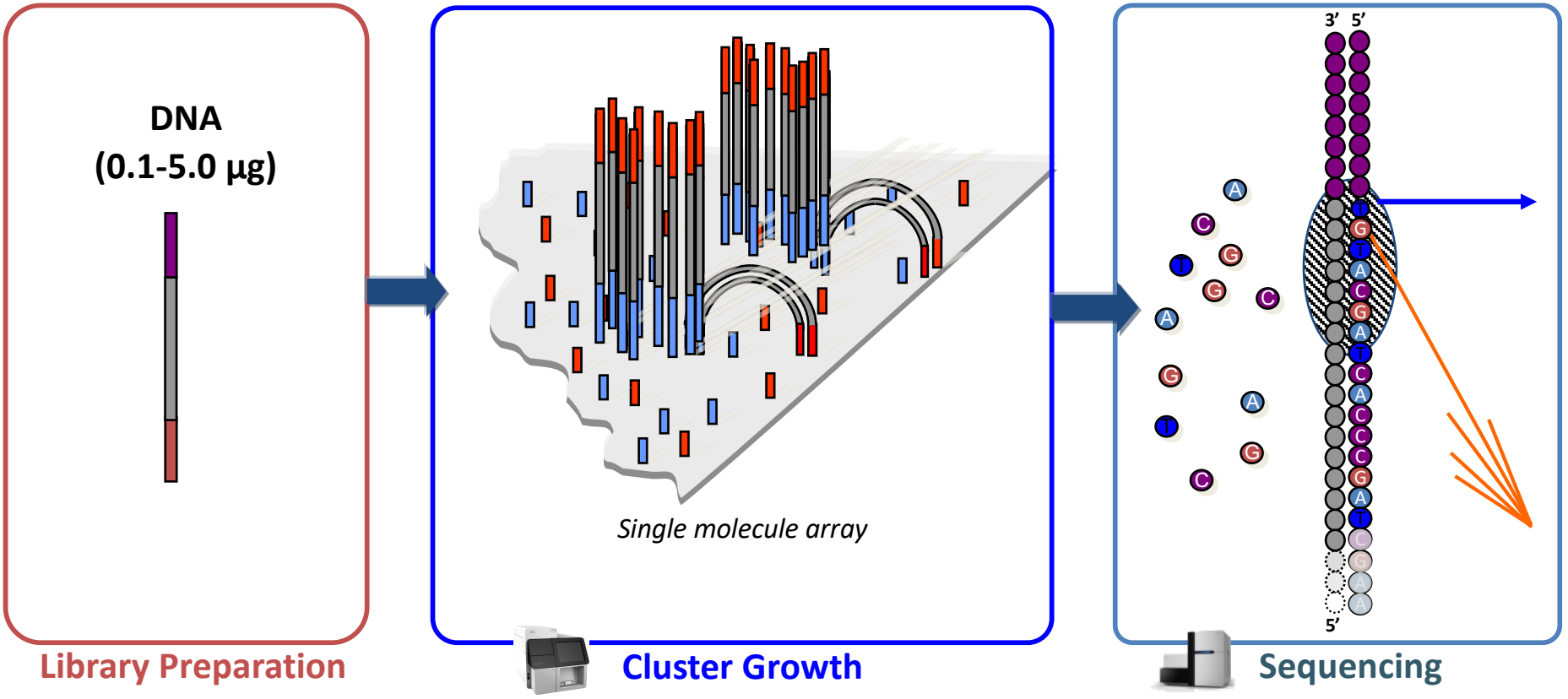
From RNA -> sequence data



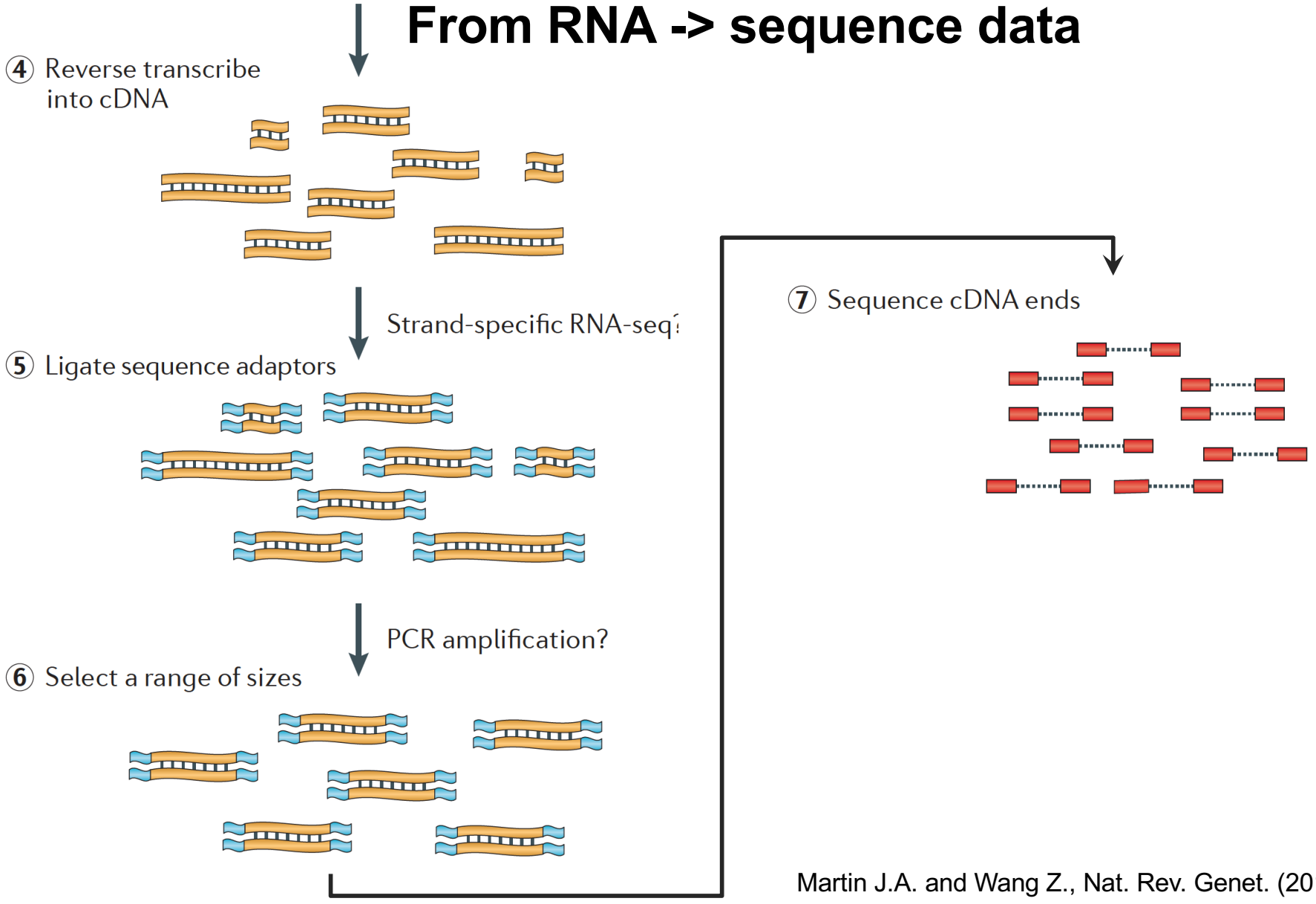
Ready for sequencing

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

Illumina Sequencing Technology Workflow



From RNA -> sequence data



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

Outline

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

Outline

2. Experimental and Practical considerations
    a. Experimental Design
    b. Poly(A) enrichment or ribosomal RNA depletion?
    c. Single-end or Paired end?
    d. Insert size for paired-end data?
    e. Stranded or not?
    f. How much sequencing data to collect?

RNA-Seq

Experimental and Practical considerations

Experimental design

- Technical replicates: Illumina has low technical variation unlike microarrays, hence technical replicates are unnecessary.
- Batch effects are still a problem, try and sequence everything for a given experiment at the same time (different flow cells are usually okay). If you are preparing the libraries, try to be consistent and make them simultaneously
- Biological replicates, are absolutely essential for your experiment to have any statistical power. Have at least 3.

RNA-Seq

Experimental and Practical considerations

Experimental design

- For transcriptome assembly, RNA can be pooled from various sources to ensure the most robust transcriptome. Pooling can also be done after sequencing, prior to entering the data into an assembler.
- For differential gene expression, pooling RNA from multiple biological replicates is usually not advisable; only do so if you have multiple pools from each experimental condition.

RNA-Seq

Experimental and Practical considerations

Poly(A) enrichment or ribosomal RNA depletion?

Depends on which RNA entities you are interested in…

- For transcriptome assembly, it is best to remove all ribosomal RNA (and maybe enrich for only polyA+ transcripts)
- For differential gene expression, it is best to enrich for Poly(A)
  - *EXCEPTION* – If you are aiming to obtain information about long non-coding RNAs
- For metatranscriptomics, e.g. gut microbiome, it is best to remove all the host materials. Remove most of the rRNA by molecular methods prior to sequencing, and remove host mRNA by computational methods post-sequencing

RNA-Seq

Experimental and Practical considerations

Single-end or Paired end?

Depends on what your goals are, paired-end reads are thought to be better for reads that map to multiple locations, for assemblies and for isoform differentiation.

Single-end read

Read1

ATGTTCCATAAGC…

Paired-end reads

Read1

ATGTTCCATAAGC…

Read2

CCGTAATGGCATG…

RNA-Seq

Experimental and Practical considerations

Single-end or Paired end?

Depends on what your goals are, paired-end reads are thought to be better for reads that map to multiple locations, for assemblies and for isoform differentiation.

- For transcriptome assembly, paired-end is the best way to go.
- For differential gene expression, single-end and paired-end are both okay, which one you pick depends on-
  - The abundance of paralogous genes in your system of interest
  - How you will be doing your analysis, and if your downstream methods are able to take advantage of the extra data you are collecting
  - Your budget, paired-end data is usually 2x more expensive
- For metatranscriptomics, paired-end is better to allow you to differentiate between orthologous genes from different species.

RNA-Seq

Experimental and Practical considerations

Stranded?

Most kits for RNA-Seq library preparation have moved to producing stranded libraries. This means that with some amount of certainty you can identify which strand of DNA the RNA was transcribed from. Strandedness is advisable for all applications.

3 types of libraries –

- **Unstranded** – you have no idea which strand of DNA was used to transcribe the reads, the information is lost during the cDNA library prep stage.
- **Reverse** – reads were transcribed from the strand with complementary sequence. dUTP incorporation during second-strand synthesis is a commonly used library prep method that yields “reverse” data.
- **Forward** – reads were transcribed from the strand that has a sequence identical to the reads.

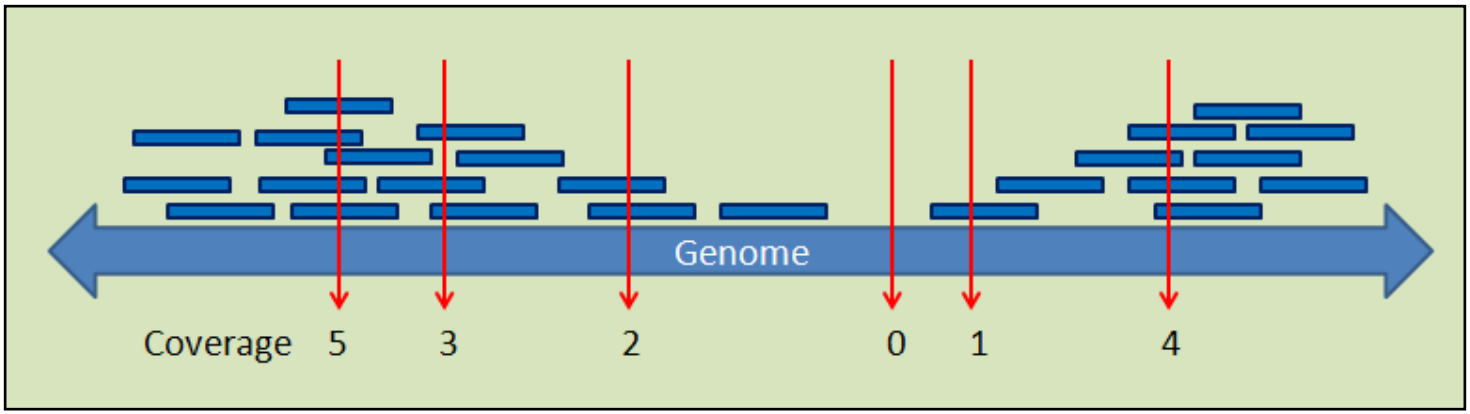
RNA-Seq

Experimental and Practical considerations

How much sequencing data to collect?

It depends heavily on the size of the transcriptome of interest, and in the case of metatranscriptomics, the diversity you expect in the community you are sequencing.

- The factor used to estimate the depth of sequencing for genomes is coverage - how many times do the total nucleotides you sequenced “cover” the genome.



RNA-Seq

Experimental and Practical considerations

How much sequencing data to collect?

It depends heavily on the size of the transcriptome of interest, and in the case of metatranscriptomics, the diversity you expect in the community you are sequencing.

- The factor used to estimate the depth of sequencing for genomes is coverage - how many times do the total nucleotides you sequenced "cover" the genome.
- But, this is not a good measure for RNA-Seq, since transcription does not occur from the whole genome (it's controversial what % is transcribed), and only ~2% of the human genome transcribes protein-coding RNA.
- You can use a rough estimate of nucleotide coverage if you only consider the protein-coding areas (depending upon exactly what you chose to sequence). But this is only a very crude, inaccurate measure, since some mRNAs will be much more abundant than others, and some genes are much longer than others!
- For human samples ~30 – 50 million reads per sample is recommended.

RNA-Seq

Experimental and Practical considerations

How much sequencing data to collect?

It depends heavily on the size of the transcriptome of interest, and in the case of metatranscriptomics, the diversity you expect in the community you are sequencing.

- The ENCODE project has some very in-depth guidelines on how to make this choice for different types of projects at http://encodeproject.org/ENCODE/experiment_guidelines.html

File formats

A brief note

Sequence formats

- FASTA
- FASTQ

Alignment formats

- SAM/BAM

Feature formats

- GFF
- GTF

File formats

FASTA

```
>unique_sequence_ID
ATTCATTAAAGCAGTTTATTGGCTTAATGTACATCAGTGAAATCATAAATGCTAAAAATTTATGATAAAAGAATAC
```

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

File formats

FASTQ

```
@unique_sequence_ID
ATTCATTAAAGCAGTTTATTGGCTTAATGTACATCAGTGAAATCATAAATGCTAAAAATTTATGATAAAAGAATAC
+
=-(DD--DDD/DD5:*1B3&)-B6+8@+1(DDB:DD07/DB&3((+:?=8*D+DDD+B)*)B.8CDBDD4DDD@@D
```

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

File formats

GFF3

- 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
```

Chromosome ID
Source
Gene feature
Start location
End location
Score (user defined)
Strand
Phase
Attributes (hierarchy)

File formats

GTF

- Evolved from Sanger Centre GFF (gene feature format) originally, but repeatedly modified
- 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";
```

Chromosome ID
Source
Gene feature
Start location
End location
Score (user defined)
Strand
Reading frame
Attributes (some hierarchy)

File formats

GTF vs GFF3

GFF3 – Gene 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

File formats

SAM

- **SAM – Sequence Alignment/Map format**
  - SAM file format stores alignment information
- **Plain text**
- **Specification**: http://samtools.sourceforge.net/SAM1.pdf
- Contains FASTQ reads, quality information, meta data, alignment information, 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)

File formats

BAM

BAM – BGZF compressed SAM format

- Compressed/binary version of SAM and is not human readable. Uses a specialize compression algorithm optimized for indexing and record retrieval
- Makes the alignment information easily accessible to downstream applications (large genome file not necessary)
- Relatively simple format makes it easy to extract specific features, e.g. genomic locations

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

Outline

3. Transcriptomic analysis methods and tools
    a. Transcriptome Analysis; aspects common to both assembly and differential gene expression
        - ✧ Quality checks
        - ✧ Data alignment
    b. Assembly
    c. Differential Gene Expression
    d. Choosing a method, the considerations…
    e. Final thoughts and observations

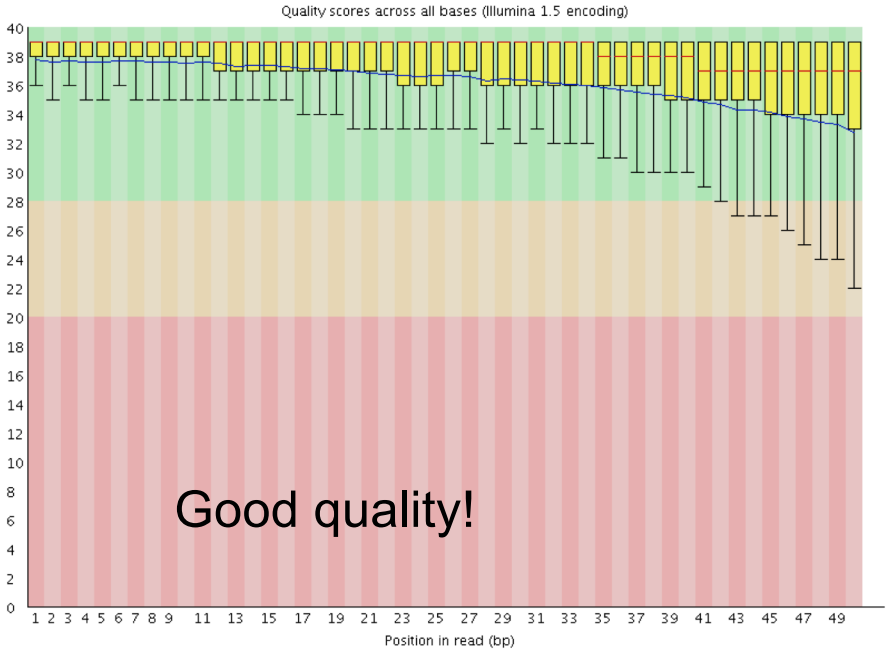
Transcriptome Analysis

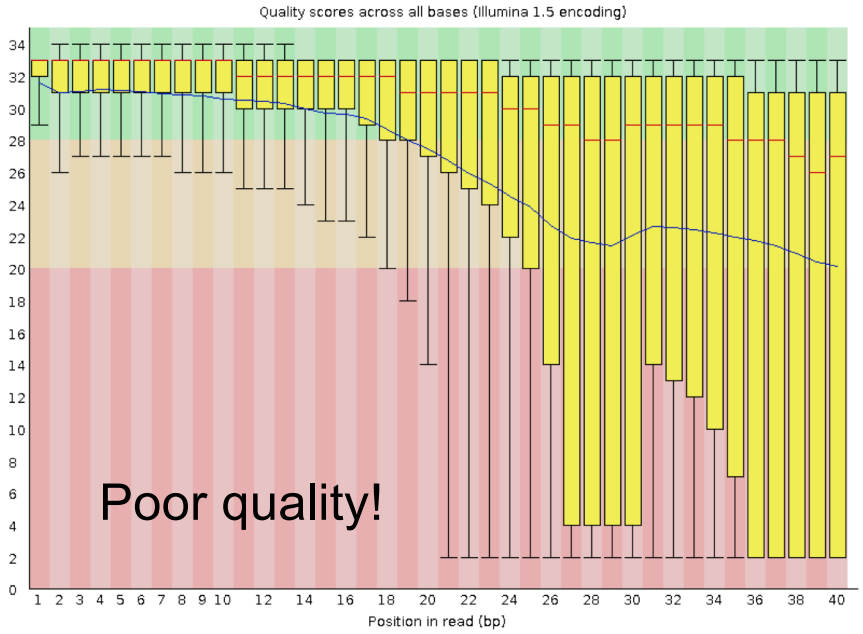
Methods and Tools

Quality checks

How does my newly obtained data look?

- ✧ Check for overall data quality. FastQC is a great tool that enables the quality assessment.





Transcriptome Analysis

Methods and Tools

Quality checks

How does my newly obtained data look?

- ✧ Check for overall data quality. FastQC is a great tool that enables the quality assessment.
- ✧ In addition to the quality of each sequenced base, it will give you an idea of
  - Presence of, and abundance of contaminating sequences.
  - Average read length
  - GC content
- ✧ *NOTE* – FastQC is good, but it is very strict and will not hesitate to call your dataset bad on one of the many metrics it tests the raw data for. Use logic and read the explanation for why and if it is acceptable.

Transcriptome Analysis

Methods and Tools

Quality checks

What do I do when FastQC calls my data poor?

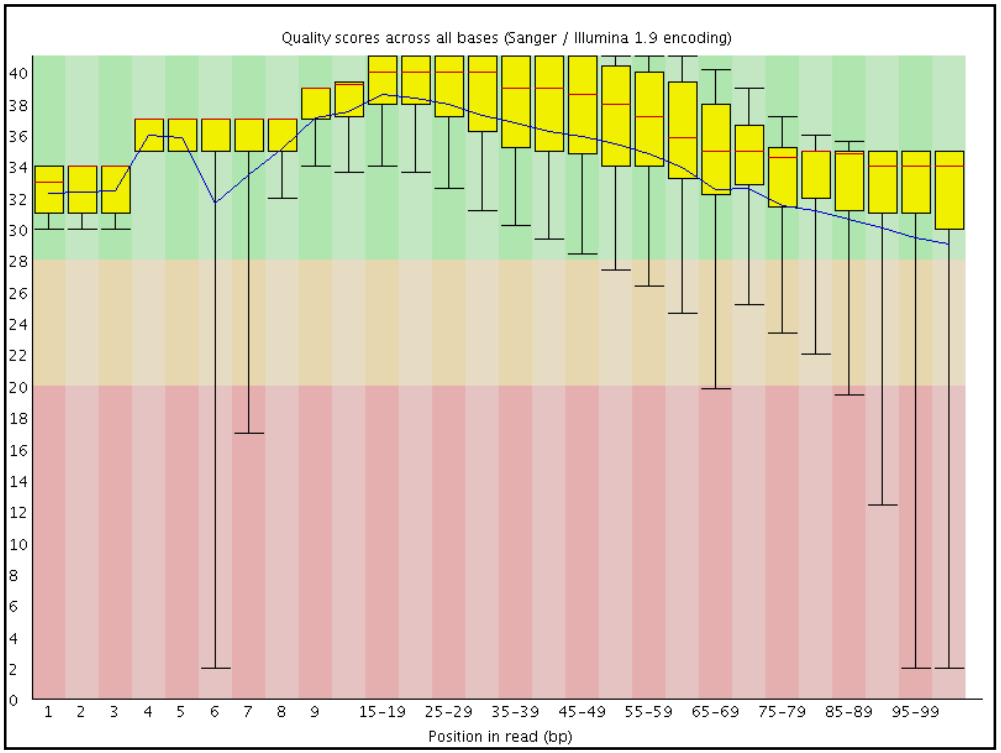
- ✧ Poor quality at the ends can be remedied by using “quality trimmers” like trimmomatic, fastx-toolkit, etc.
- ✧ Left-over adapter sequences in the reads can be remedied by using “adapter trimmers” like trimmomatic. Always trim adapters as a matter of routine (trimmomatic does both types of trimming at once).
- ✧ We need to take care of these 2 types of issues so we get the best possible alignment, since with short reads only a few mismatches are allowed.
- ✧ Once the trimmers have been used, it is best to rerun the data through FastQC to check the resulting data.

Transcriptome Analysis

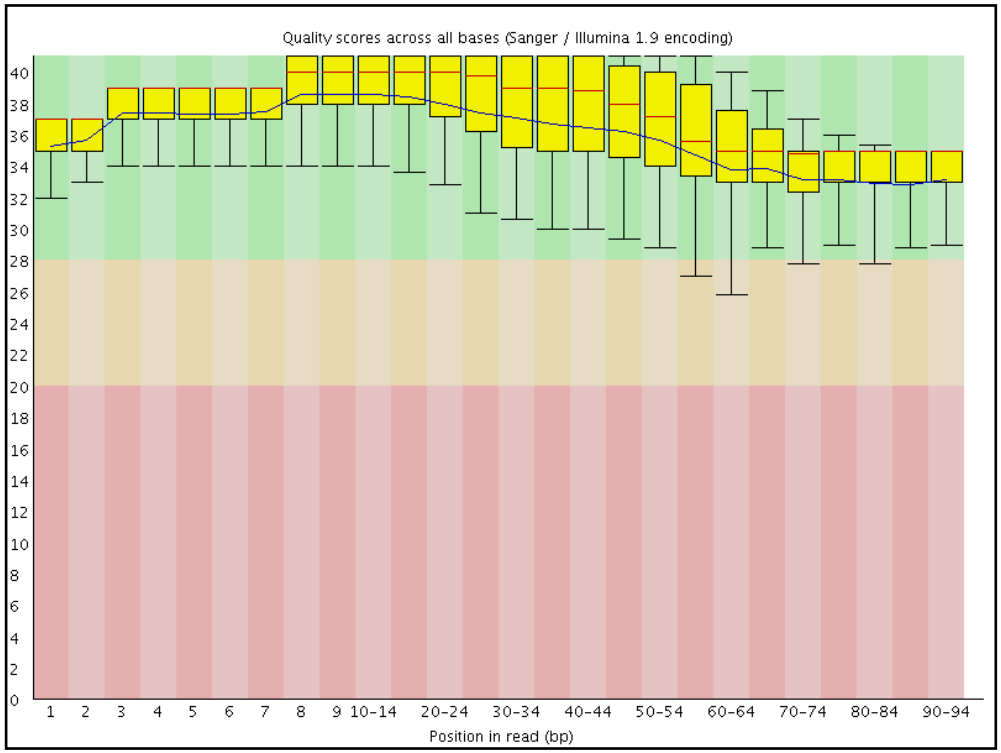
Methods and Tools

Quality checks

Before quality trimming



After quality trimming



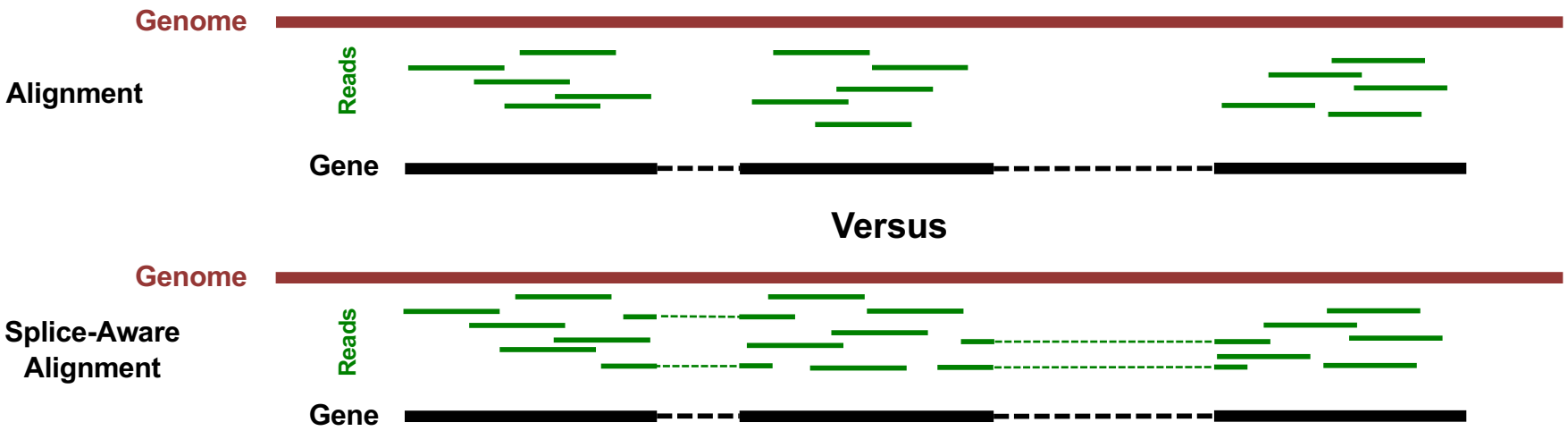
Transcriptome Analysis

Methods and Tools

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 slice-aware aligner



Transcriptome Analysis

Methods and Tools

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 slice-aware aligner

TopHat2, MapSplice, SOAPSplice, Passion, SpliceMap, RUM, ABMapper, CRAC, GSNAP, HMMSplicer, Olego, BLAT

- ✧ There are excellent aligners available that are not splice-aware. These are useful for aligning directly to an already available transcriptome (gene models, so you are not worrying about introns). However, be aware that you will lose isoform information.

Bowtie2, BWA, Novoalign (not free), SOAPaligner

Transcriptome Analysis

Methods and Tools

Data alignment

What other considerations do you have to make 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?

Transcriptome Analysis

Methods and Tools

Data alignment

How does one pick from all the tools available?

- Tophat is the most commonly used splice-aware aligner, and is part of a suite of software that make up the Tuxedo pipeline/suite. It is reliable.
- Some of the listed tools are a little better than the others at doing specific things; e.g. better speed or memory usage, available options for reads that have multiple hits, and so on.

Transcriptome Analysis

Methods and Tools

Data alignment

IGV is the visualization tool used for this snapshot

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

Transcriptome Assembly overview

Methods and Tools

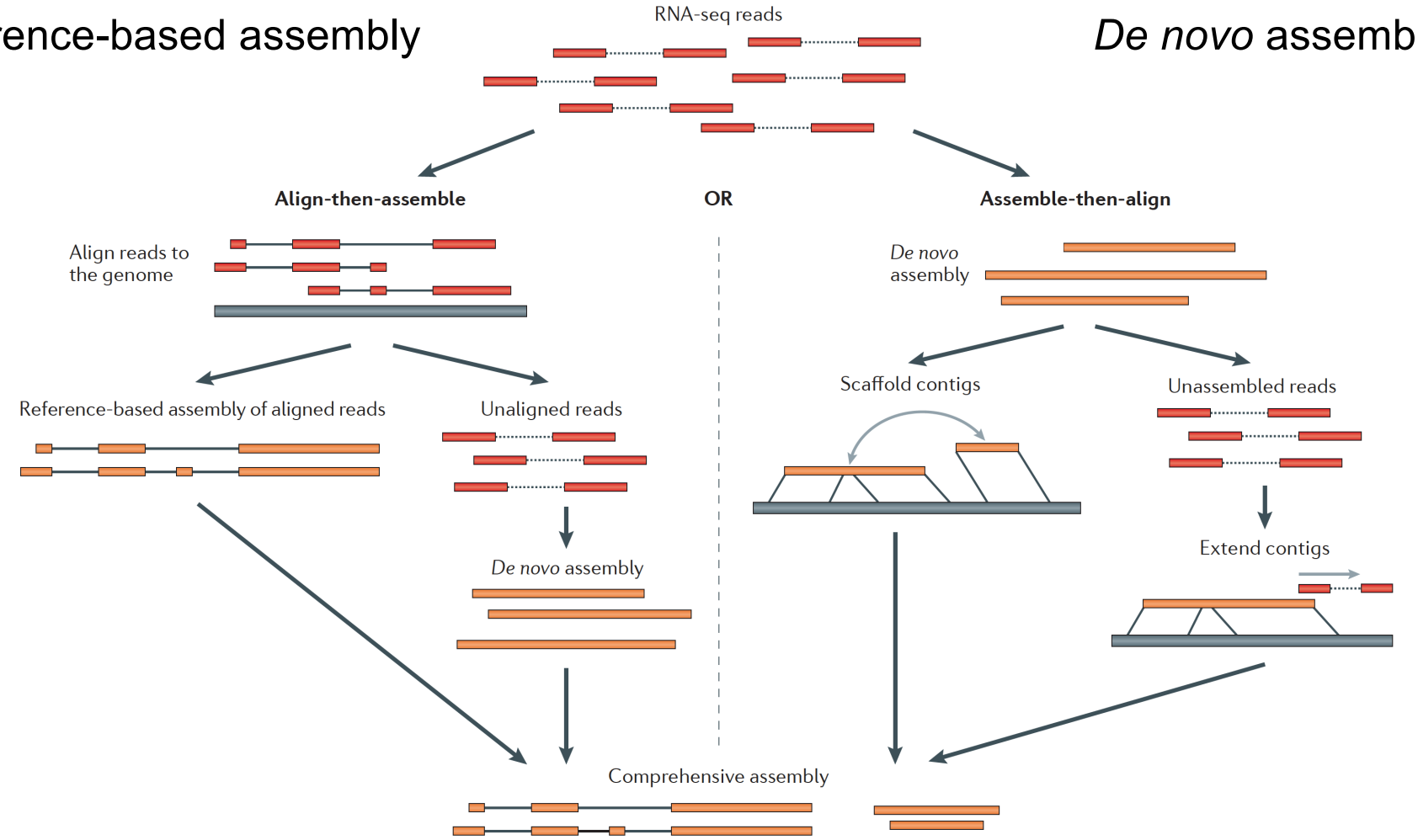
1) Obtain/download sequence data from sequencing center
2) Check quality of data and trim low quality bases from ends
3) Pick your method of choice for assembly
    a. Reference-based assembly?
        *(*Align to reference and assemble)
    b. A *de novo* assembly?

Transcriptome Assembly

Methods and Tools

Reference-based assembly

De novo assembly



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

Transcriptome Assembly

Methods and Tools

Reference-based assembly

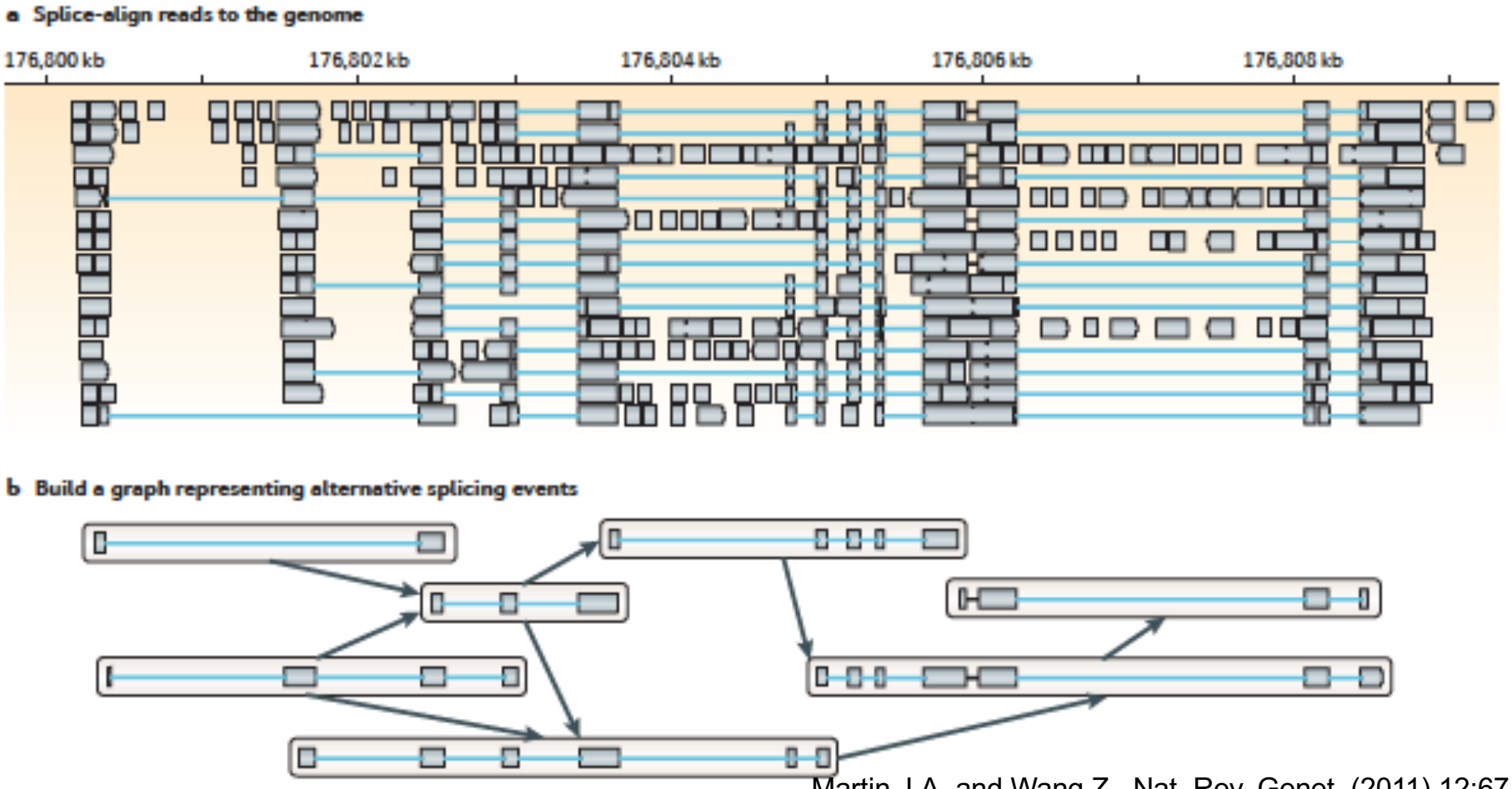
This type of assembly is used when the genome sequence is known.

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

Transcriptome Assembly

Methods and Tools

Reference-based assembly

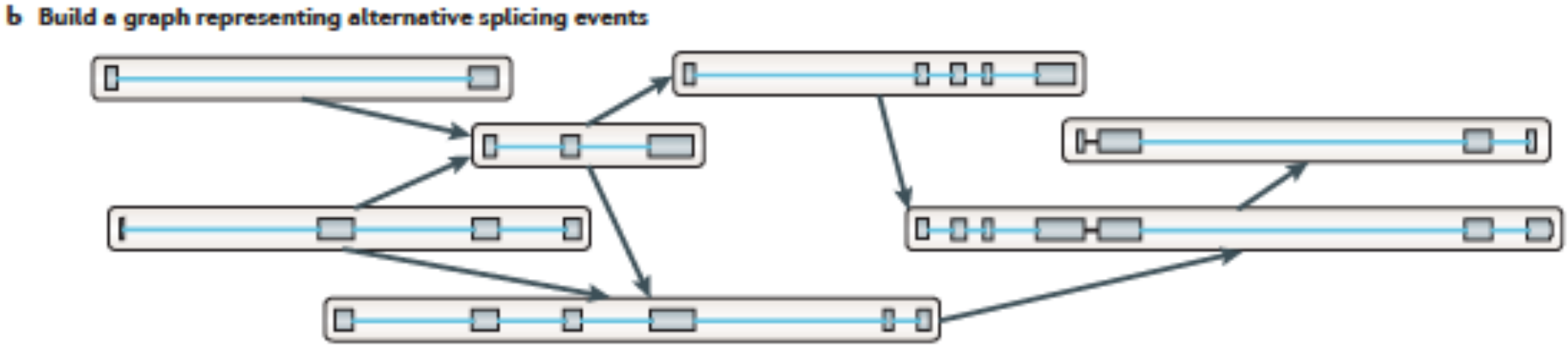


Martin J.A. and Wang Z., Nat. Rev. Genet. (2011) 12:671–682

Transcriptome Assembly

Methods and Tools

Reference-based assembly

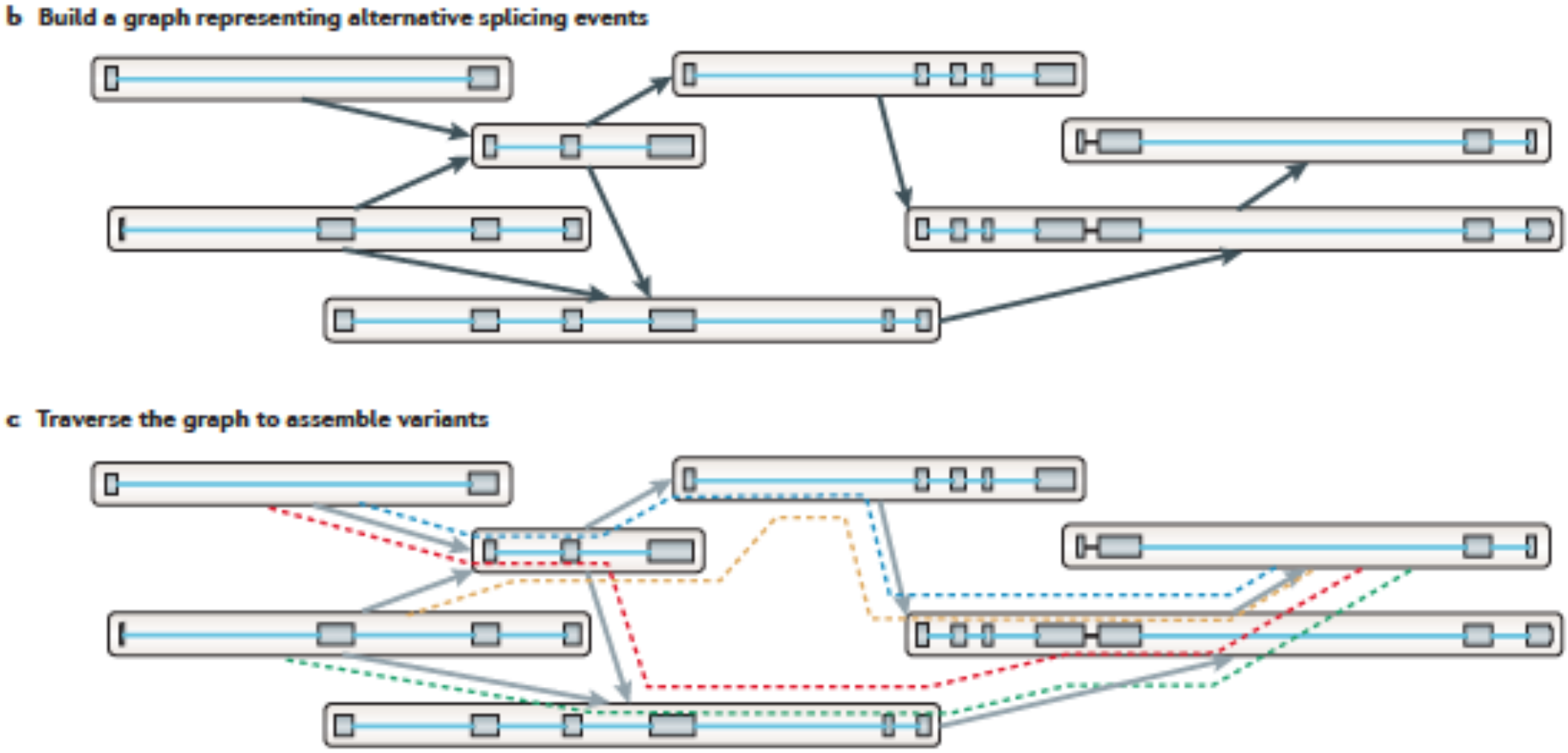


Martin J.A. and Wang Z., Nat. Rev. Genet. (2011) 12:671–682

Transcriptome Assembly

Methods and Tools

Reference-based assembly

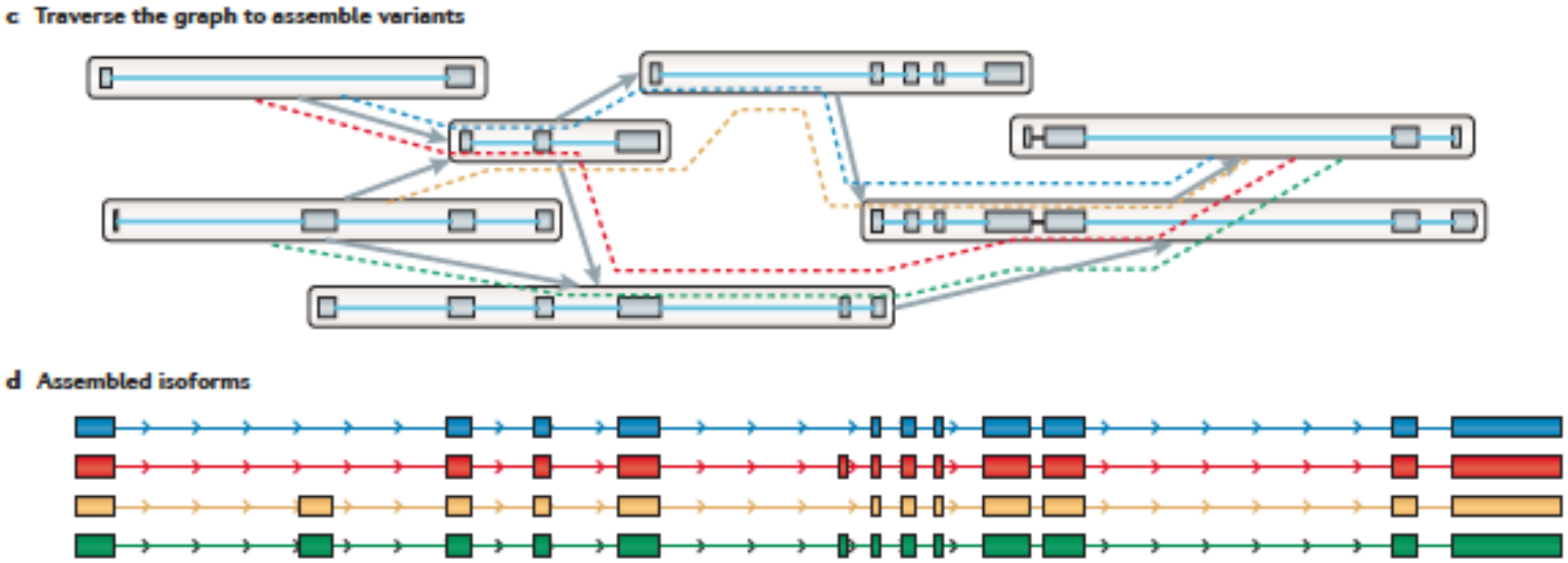


Martin J.A. and Wang Z., Nat. Rev. Genet. (2011) 12:671–682

Transcriptome Assembly

Methods and Tools

Reference-based assembly



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

Transcriptome Assembly

Methods and Tools

De novo assembly

This type of assembly is used when very little information is available for the genome

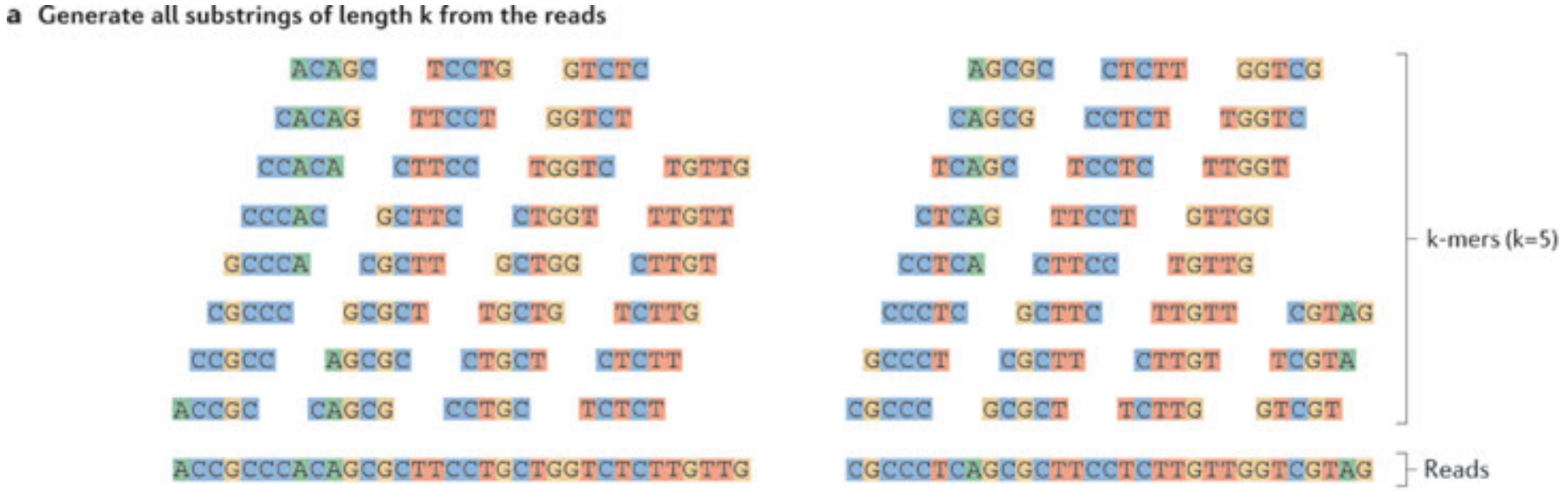
- An assembly of this type is often the first step in putting together information about an unknown genome
- Amount of data needed for a good *de novo* assembly is higher than what is needed for a reference-based assembly
- Assemblies of this sort can be used for genome annotation, once the genome is assembled
- Oases, TransABySS, Trinity are examples of well-regarded transcriptome assemblers, especially Trinity

It is not uncommon to used both methods and compare and combine the assemblies, even when a genome sequence is known, especially for a new genome.

Transcriptome Assembly

Methods and Tools

De novo assembly (De Bruijn graph construction)

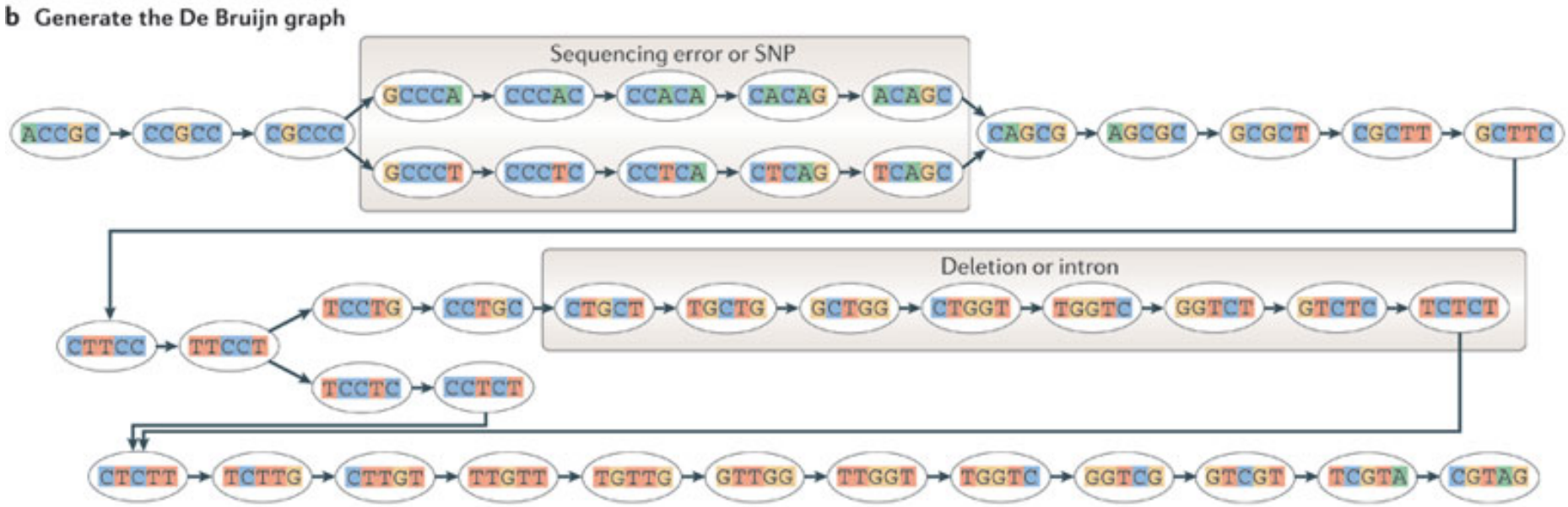


Martin J.A. and Wang Z., Nat. Rev. Genet. (2011) 12:671–682

Transcriptome Assembly

Methods and Tools

De novo assembly (De Bruijn graph construction)

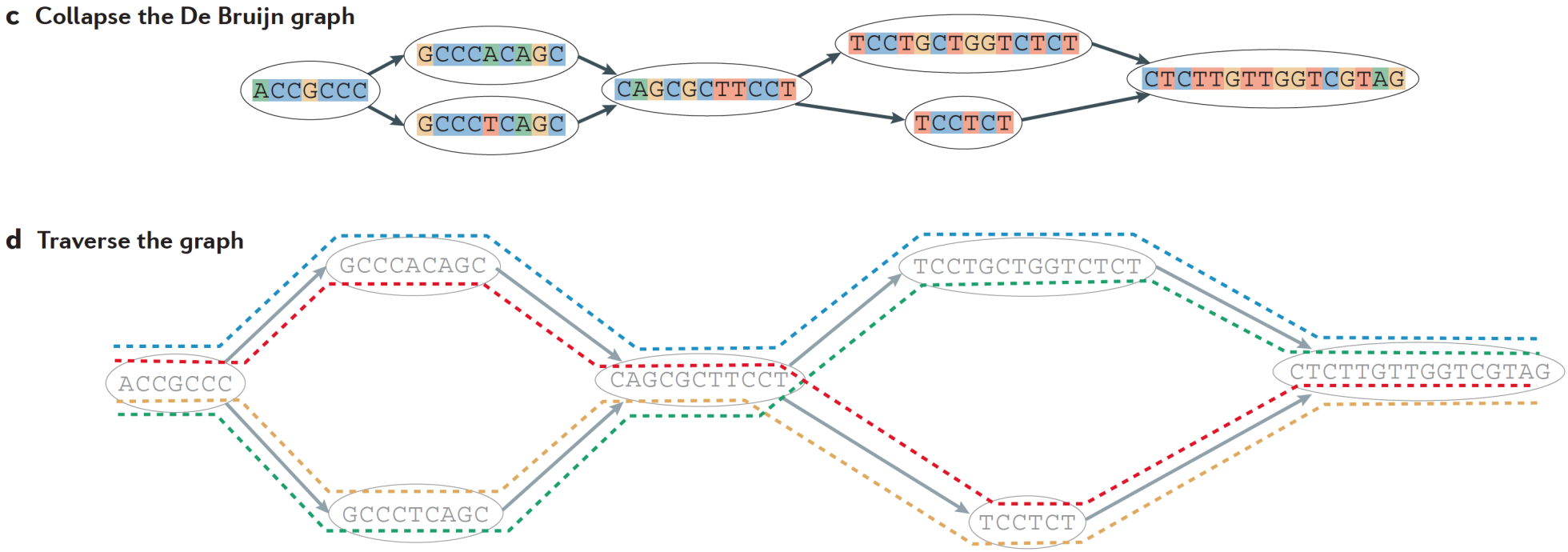


Martin J.A. and Wang Z., Nat. Rev. Genet. (2011) 12:671–682

Transcriptome Assembly

Methods and Tools

De novo assembly (De Bruijn graph construction)



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

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

Methods and Tools

① Obtain/download sequence data from sequencing center

② Check quality of data and trim low quality bases from ends

③ Align trimmed reads to genome of interest

a. Pick alignment tool, splice-aware or not? (map to gene set?)

b. Index genome file according to instructions for that tool

c. Run alignment after choosing the relevant parameters, like how many mismatches to allow between reads and genome? what is to be done with reads that map to multiple locations?

Differential Gene Expression overview

Methods and Tools

④ Set up to do differential gene expression

Identify read counts associated with genes using the gene annotation file

a. Make sure that your genome information and gene annotation information match (release numbers and chromosome names)

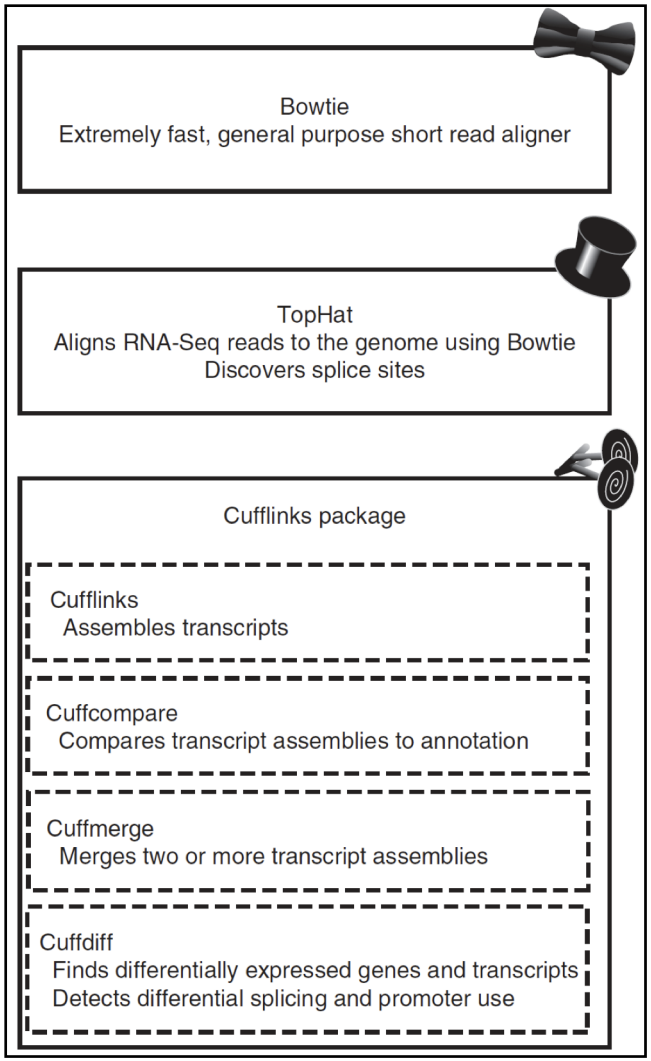
b. Do you want to obtain raw read counts or normalized read counts? This will depend on the statistical analysis you wish to perform downstream.

- ✧ htseq will take an alignment file and a gene annotation file to give you read counts associated with each gene
- ✧ Cufflinks will take the same information as htseq and give you FPKM normalized counts for each gene

Differential Gene Expression

Methods and Tools

Options for DGE analysis (tuxedo suite)



Bowtie and Bowtie use Burrows-Wheeler indexing for aligning reads. With bowtie2 there is no upper limit on the read length

Tophat uses either Bowtie or Bowtie2 to align reads in a splice-aware manner and aids the discovery of new splice junctions

The Cufflinks package has 4 components, the 2 major ones are listed below -

Cufflinks does **reference-based transcriptome assembly**

Cuffdiff does statistical analysis and identifies differentially expressed transcripts in a simple pairwise comparison, and a series of pairwise comparisons in a time-course experiment

Trapnell *et al.*, Nature Protocols, March 2012

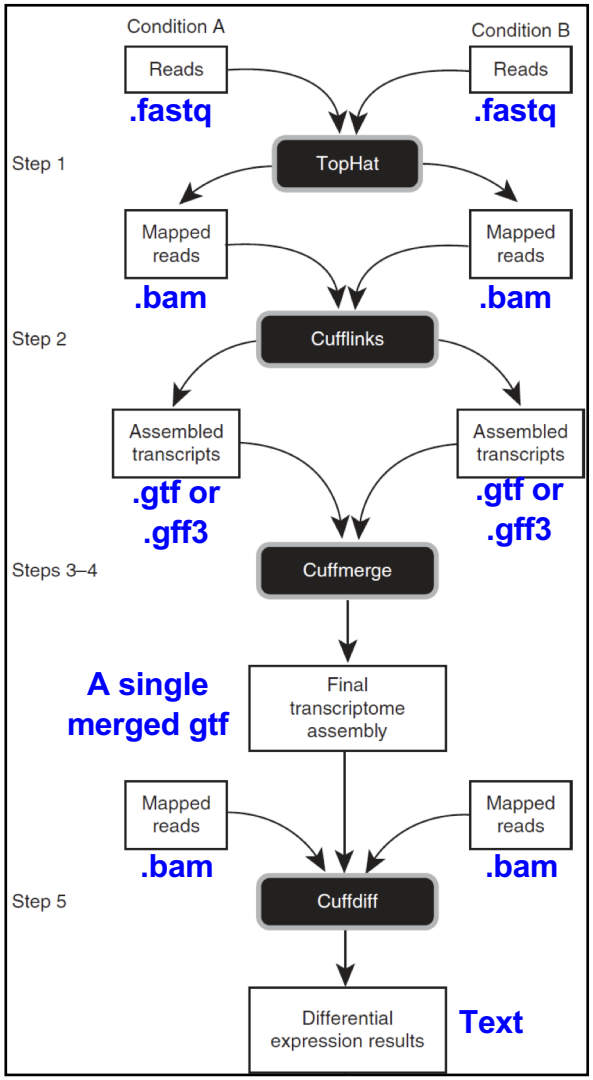
Differential Gene Expression

Methods and Tools

Options for DGE analysis (tuxedo suite)

Want to learn more about the formats?

https://genome.ucsc.edu/FAQ/FAQformat.html



Raw sequence data file

Alignment file

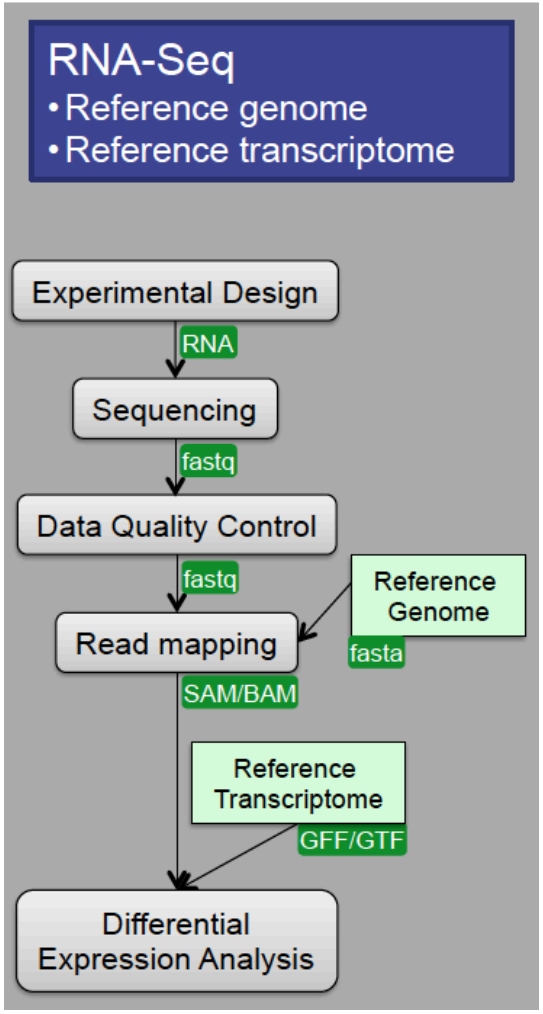
Gene annotation file

Trapnell *et al.*, Nature Protocols, March 2012

Differential Gene Expression

Methods and Tools

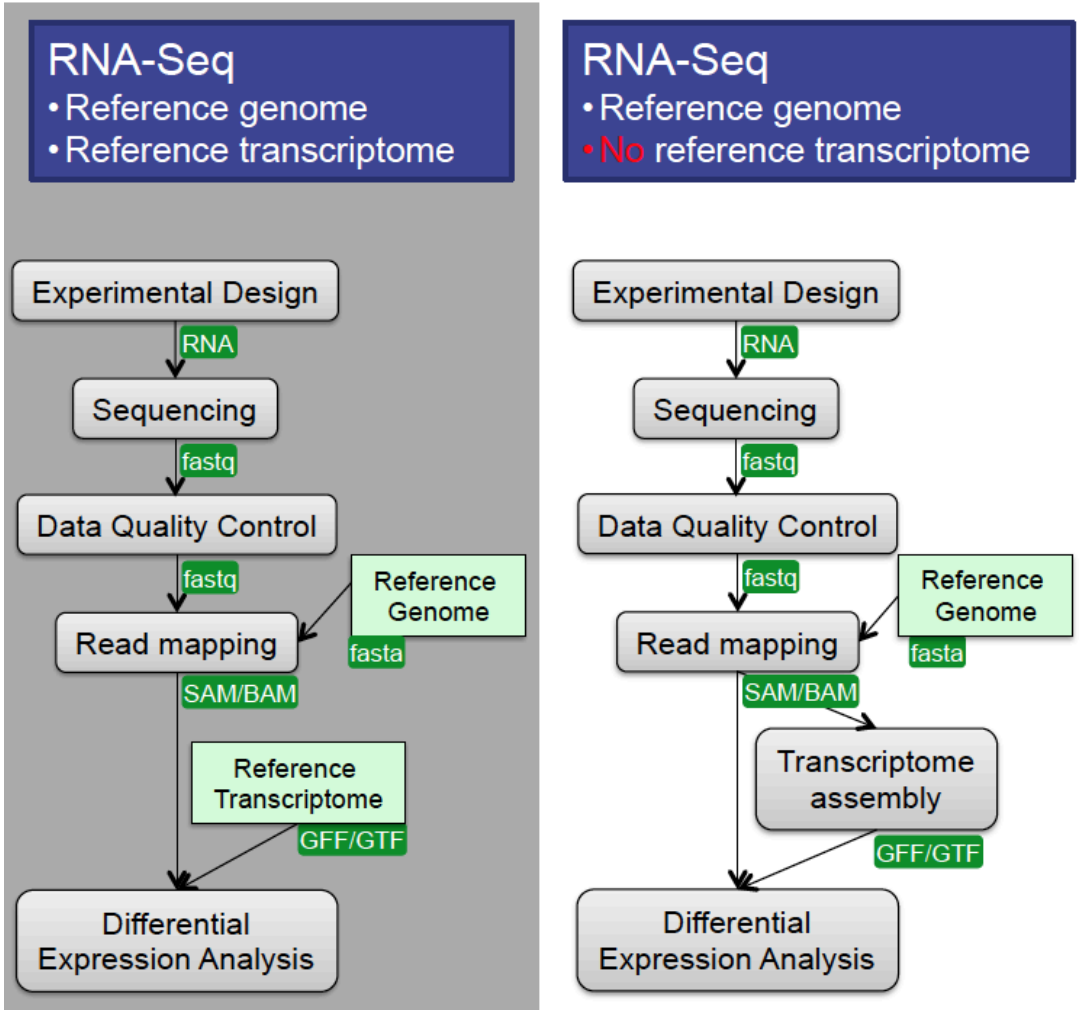
Options for DGE analysis



Differential Gene Expression

Methods and Tools

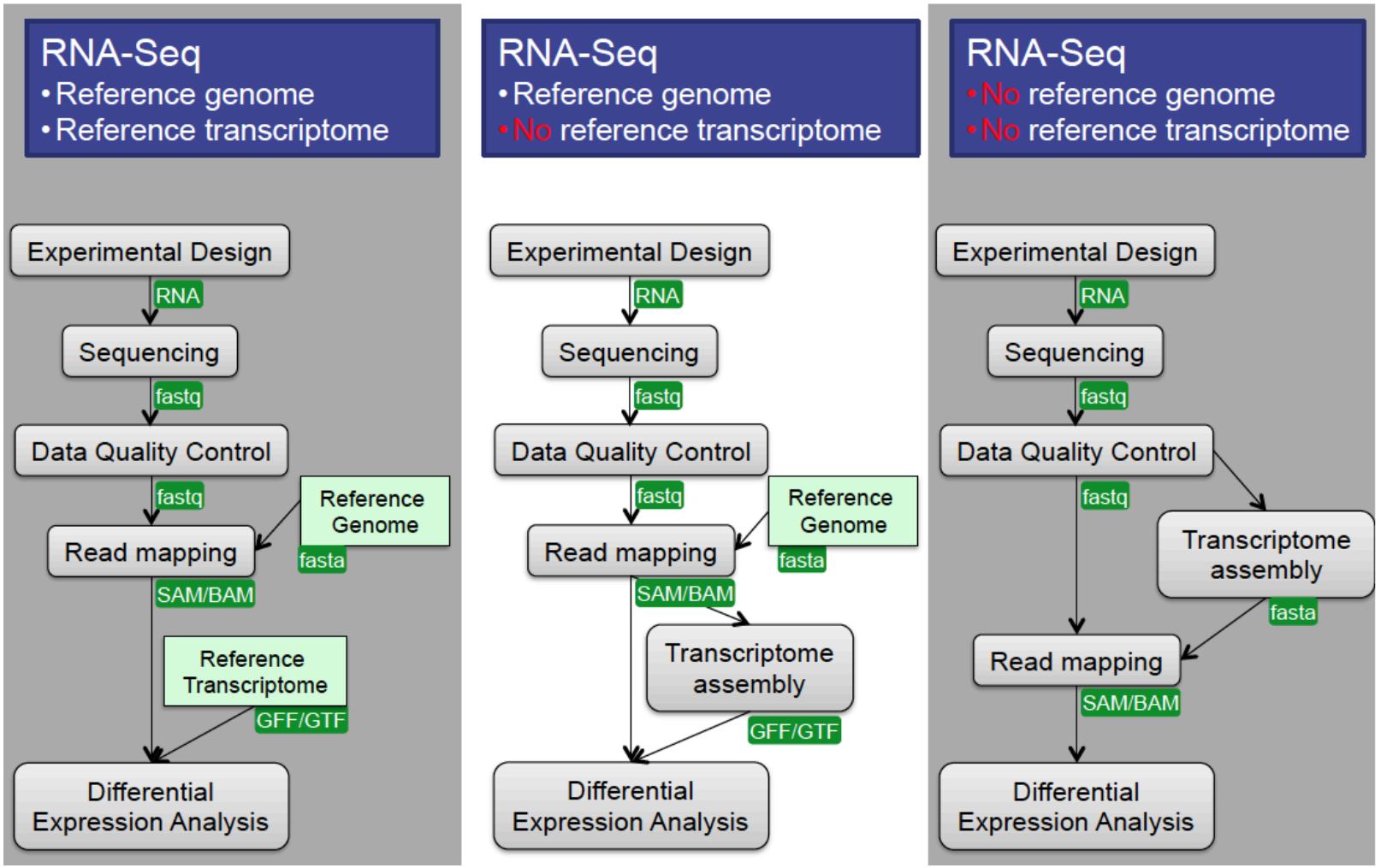
Options for DGE analysis



Differential Gene Expression

Methods and Tools

Options for DGE analysis



Differential Gene Expression

Methods and Tools

Differential Gene Expression

What genes are being differentially expression in the various test conditions

- The first step is proper normalization of the data, several methods exist, and often the statistical package you use (see below) will have a normalization method that it prefers and uses exclusively. E.g. Voom, FPKM, scaling (used by EdgeR)
- Is your experiment a pairwise comparison? Tools -> Cuffdiff, EdgeR, DESeq
- Is it a more complex design? Tools -> EdgeR, DESeq, other R/Bioconductor packages
- In general, RNA-Seq data do not follow a normal (Poisson) distribution, but follow a negative binomial distribution. Use a statistical program that makes the correct assumptions about the data distribution.

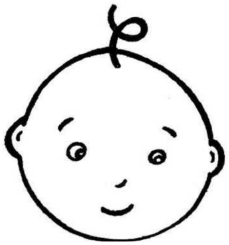
Outline

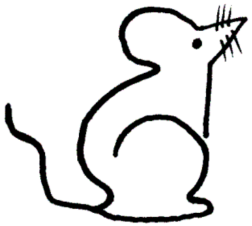
3. Transcriptomic analysis methods and tools
    a. Transcriptome Analysis; aspects common to both assembly and differential gene expression
        - ✧ Quality check
        - ✧ Data alignment
    b. Assembly
    c. Differential Gene Expression
    d. Choosing a method, the considerations…
    e. Final thoughts and observations

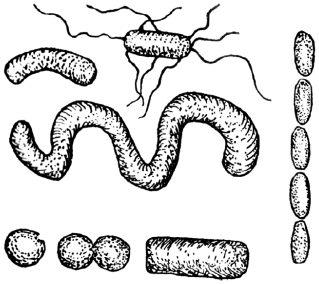
Transcriptome Analysis

Methods and Tools

How does one pick the right tool?







Step 1: Quality Control

Step 2: Data prepping

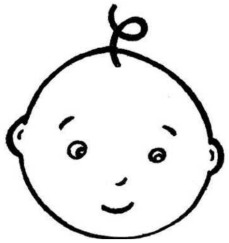
Step 3: Map Reads to Reference Genome/Transcriptome

Step 4: Assemble Transcriptome

Discovery mode

Other applications:
De novo Assembly
Refine gene models

Identify Differentially Expressed Genes



Step 1: Quality Control — FastQC

Step 2: Data prepping — Filter/Trimmer/Converter

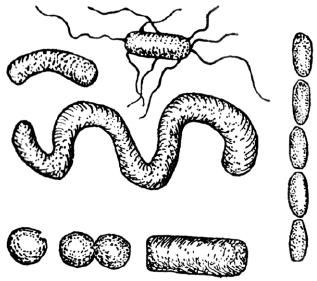
Step 3: Map Reads to Reference Genome/Transcriptome — TopHat, GSNAP

Step 4: Assemble Transcriptome — Cufflinks, Cuffmerge

IGV

Other applications: Refine gene models

Identify Differentially Expressed Genes — Cuffdiff



Step 1: Quality Control — FastQC

Step 2: Data prepping — Filter/Trimmer/Converter

Step 3: Map Reads to Reference Genome/Transcriptome — **Bowtie, BWA**

Step 4:

"We don't recommend assembling bacteria transcripts using Cufflinks at first. If you are working on a new bacteria genome, consider a computational gene finding application such as Glimmer." – Cufflinks developer

Glimmer
Artemis
Trinity, Velvet

Other applications:
De novo Assembly
Refine gene models

Identify Differentially Expressed Genes — Cuffdiff

Step 1: Quality Control — FastQC

Step 2: Data prepping — Filter/Trimmer/Converter

Step 3: Map Reads to Reference Genome/Transcriptome — **TopHat, GSNAP**

Step 4: Assemble Transcriptome — Cufflinks, Cuffmerge

Identify Differentially Expressed Genes — Cuffdiff

Other applications: De novo Assembly, Refine gene models — **GeneMark, FGeneSH, Trinity, TransABySS, BLAT**

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

Topics covered today

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

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)

   These sites cover a variety of topics, and questions from people with a variety of expertise. If you know what you are looking for, it is very likely that someone has already asked the question. If not, it is good forum to ask it yourself.

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)

   These sites cover a variety of topics, and questions from people with a variety of expertise. If you know what you are looking for, it is very likely that someone has already asked the question. If not, it is good forum to ask it yourself.
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 not other types of sequence data. Usually you are going to encounter short-read Illumina data for these types of analyses, but it is not uncommon for people to use 454 data as well. Hybrid assemblies can be done, but are challenging and no straightforward method exists.

Final thoughts and stray observations

4. Today we covered how to deal with Illumina data, but not other types of sequence data. Usually you are going to encounter short-read Illumina data for these types of analyses, but it is not uncommon for people to use 454 data as well. Hybrid assemblies can be done, but are challenging and no straightforward method exists.
5. For evaluating *de novo* transcriptome assemblies, you can compare the new genes to closely related species or evolutionarily conserved genes and check for representation (CEGMA, BUSCO).

Final thoughts and stray observations

4. Today we covered how to deal with Illumina data, but not other types of sequence data. Usually you are going to encounter short-read Illumina data for these types of analyses, but it is not uncommon for people to use 454 data as well. Hybrid assemblies can be done, but are challenging and no straightforward method exists.
5. For evaluating *de novo* transcriptome assemblies, you can compare the new genes to closely related species or evolutionarily conserved genes and check for representation (CEGMA, BUSCO).
6. 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.)

Documentation and Support

Online resources for RNA-Seq analysis questions –

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

Contact us at:

hpcbiohelp@illinois.edu

hpcbiotraining@igb.illinois.edu

rkhetani@illinois.edu

Thank you for your attention!

For this presentation, figures and slides came from publications, web pages and presentations, and I am grateful for all the help.