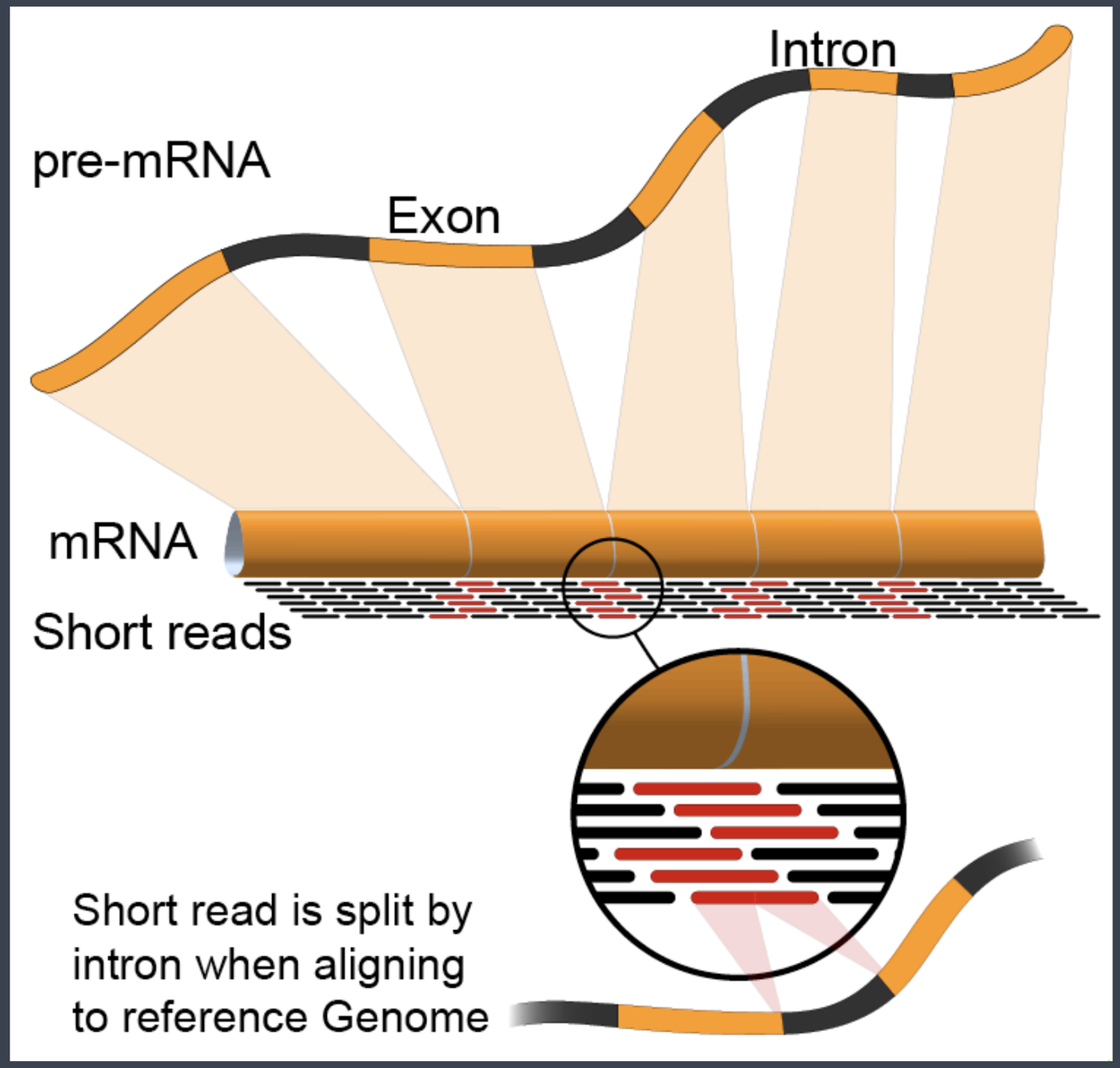
RNA-Seq: experimental design



http://upload.wikimedia.org/wikipedia/commons/0/01/RNA-Seq-alignment.png

Transcriptomics (RNA-Seq)

- The process of sequencing the “transcriptome”
- Uses include –
  - Differential Gene Expression

    Quantitative evaluation and comparison of transcript levels
  - Transcriptome assembly

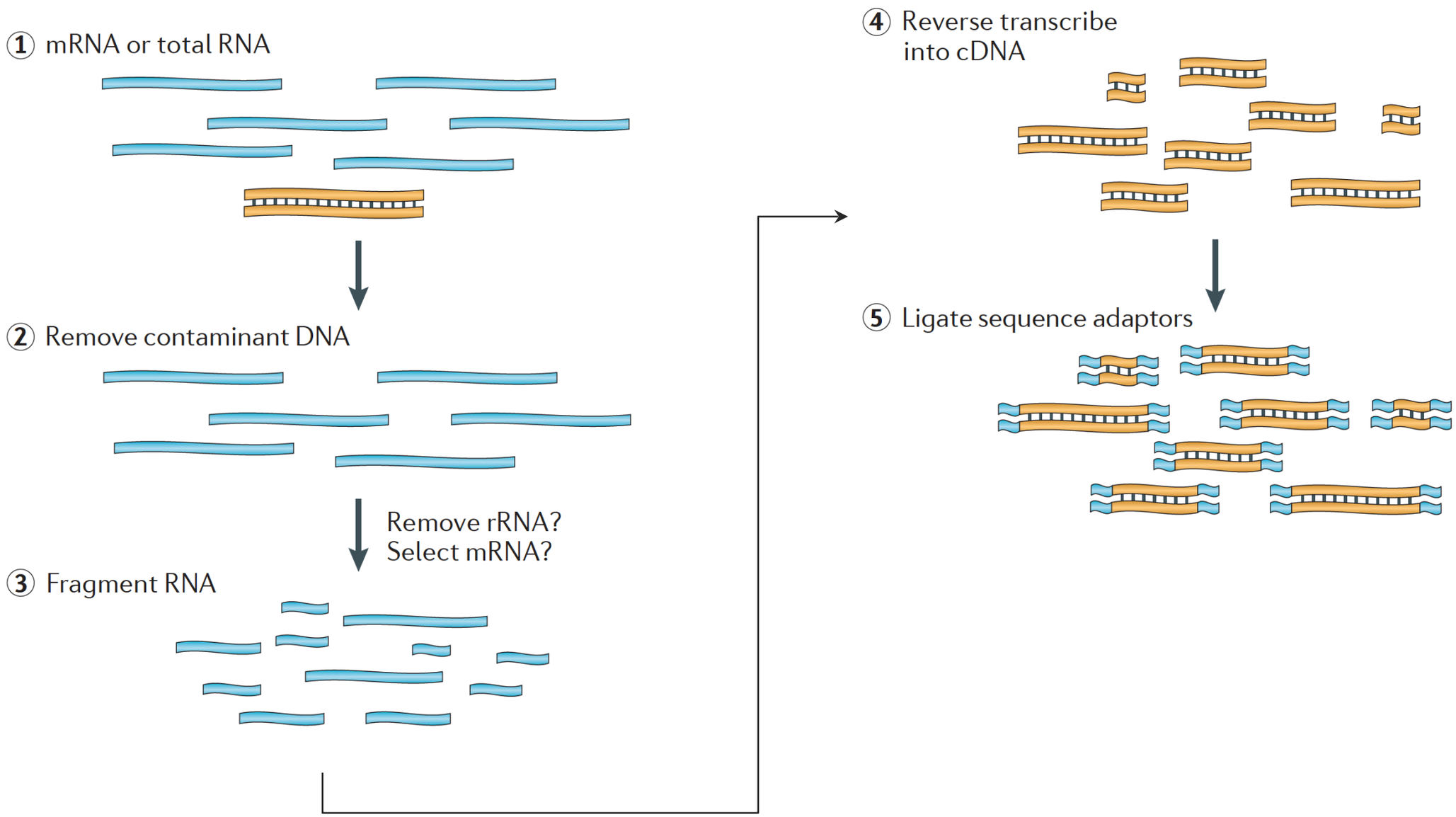
    Building the profile of transcribed regions of the genome, a qualitative evaluation.
  - Can be used to help build better gene models, and verify them using the assembly
  - Metatranscriptomics or community transcriptome analysis

Outline

- Library preparation
- Experimental and Practical Considerations
- Analysis workflow and options
- *Commonly used file formats*

Outline

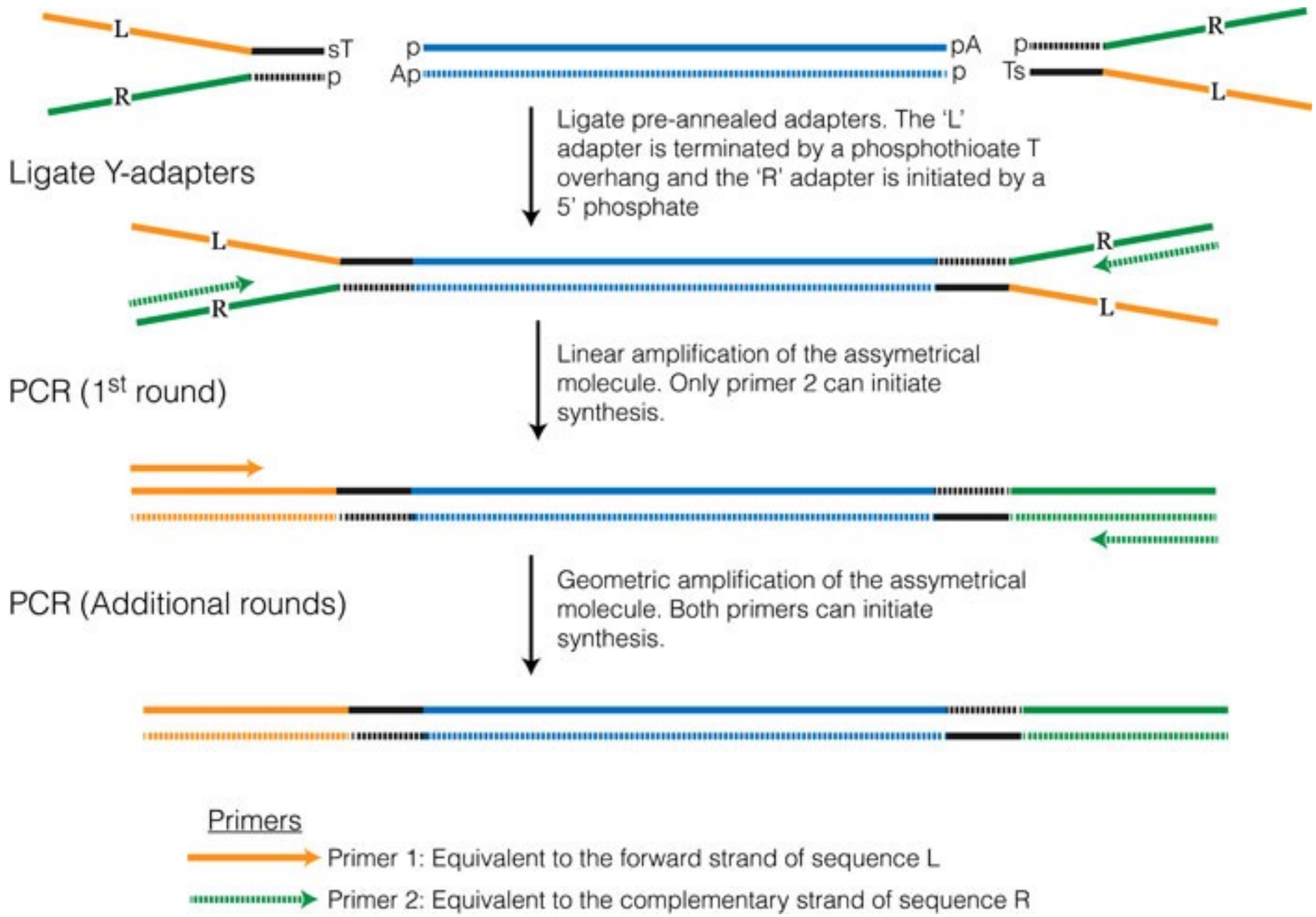
- Library preparation
- Experimental and Practical Considerations
- Analysis workflow and options
- *Commonly used file formats*



RNA-Seq library prep

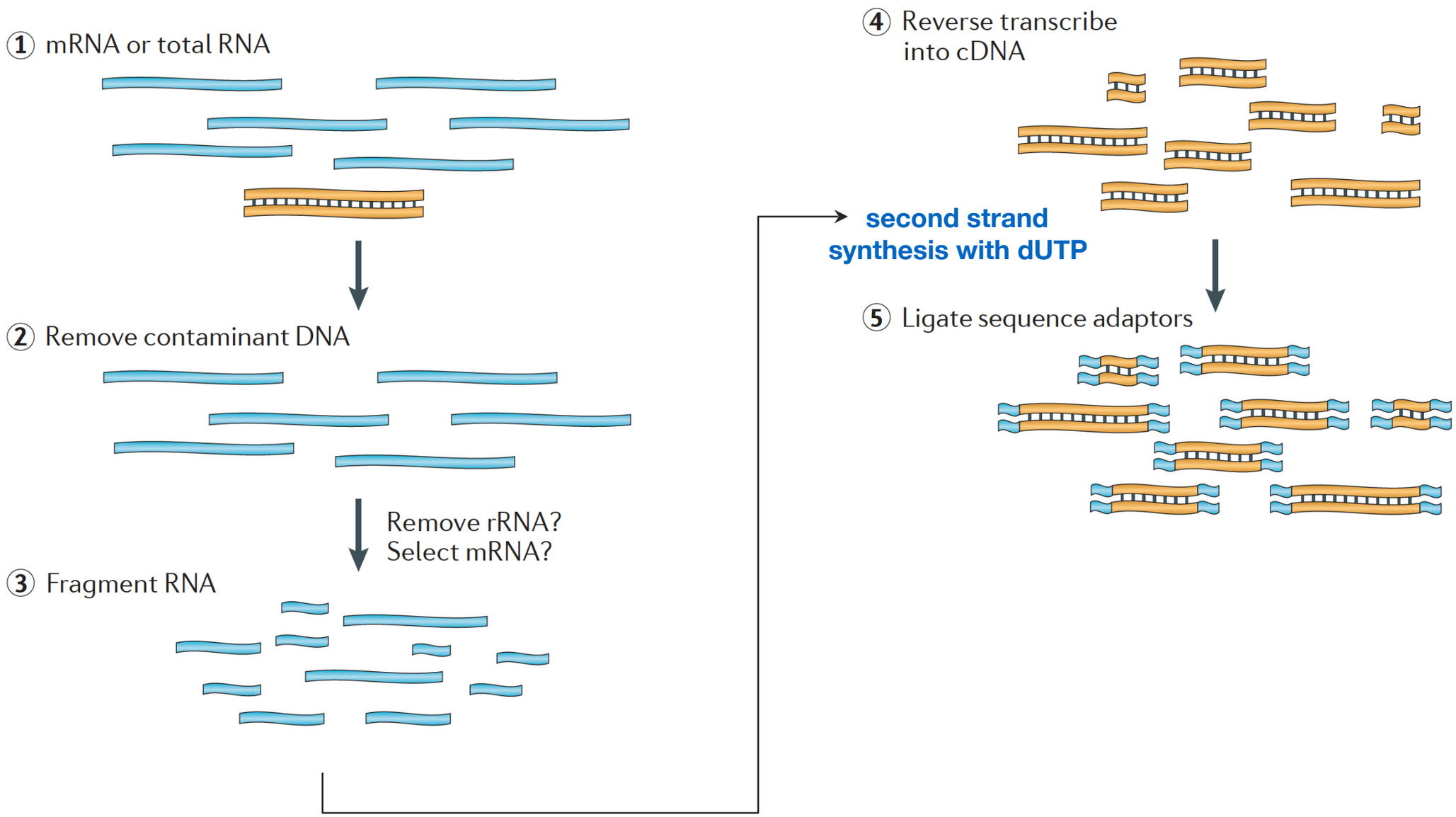
http://rnaseq.uoregon.edu/#rna-prep

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



Y-adapters :: RNA-Seq library prep

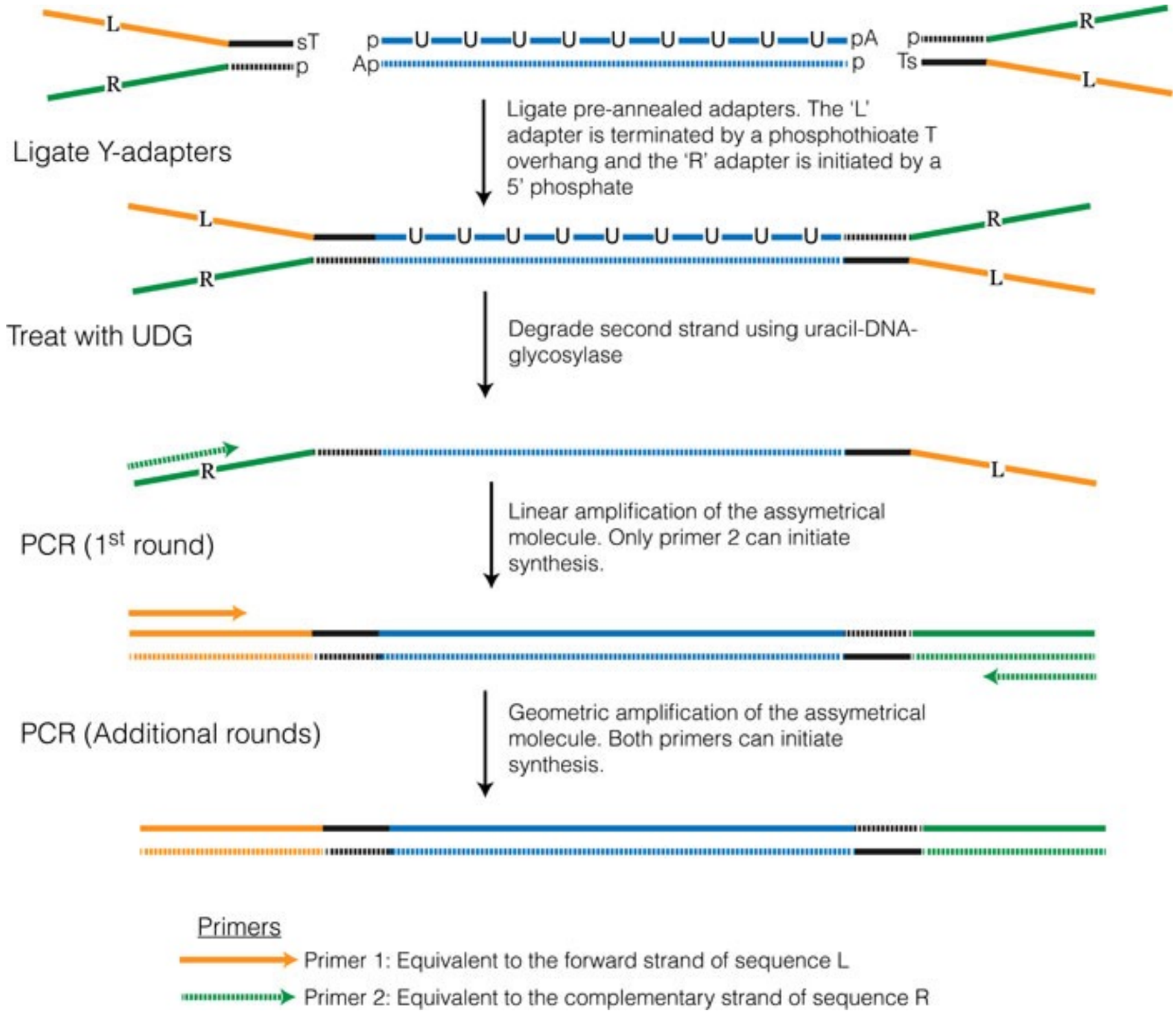
http://rnaseq.uoregon.edu/#library-prep



Stranded library prep (dUTP method)

http://rnaseq.uoregon.edu/#rna-prep

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



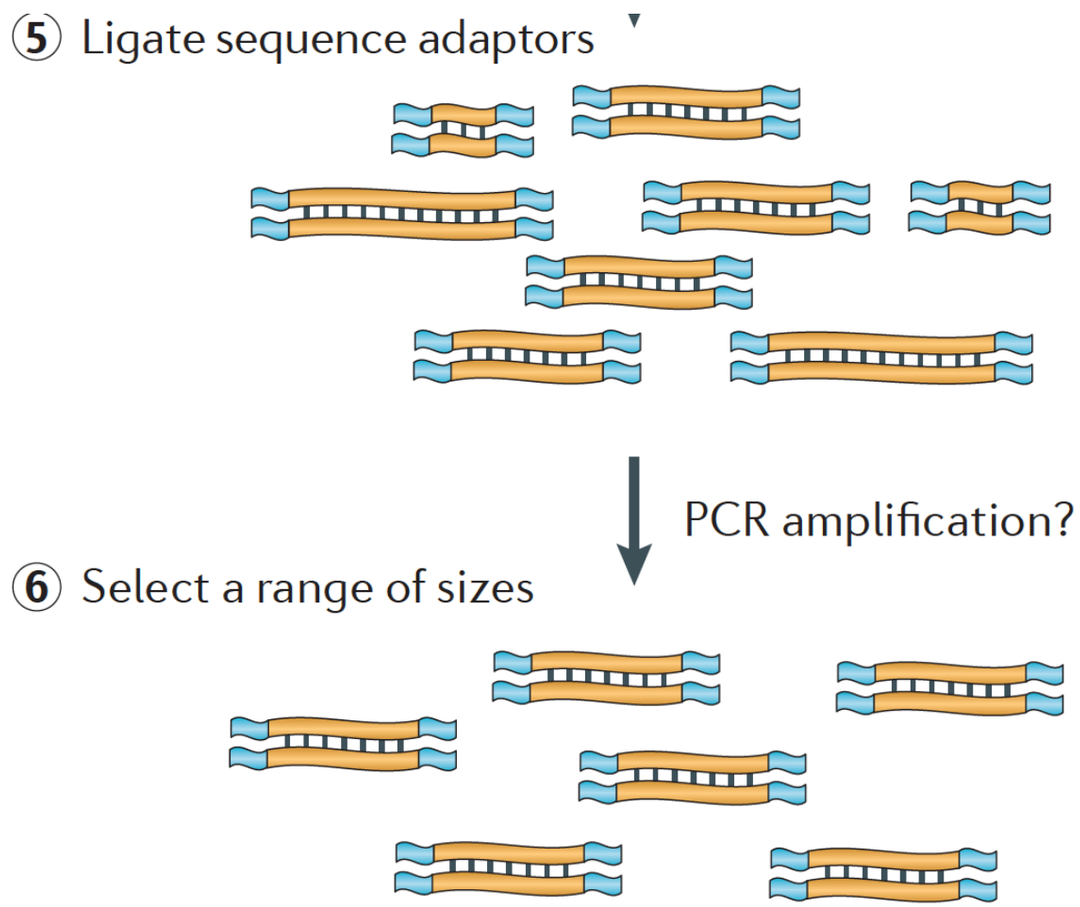
Stranded library prep (dUTP method)

http://rnaseq.uoregon.edu/#library-prep

- ✓ **reverse** or **fr-firststrand** => reverse complement of coding strand sequenced almost exclusively
  - dUTP (Illumina Truseq Stranded), NSR, NNSR
- ✓ **forward** or **fr-secondstrand** => coding strand sequenced almost exclusively
- ✓ **unstranded** => roughly equal amounts of both coding and it's reverse complement

Stranded library preps

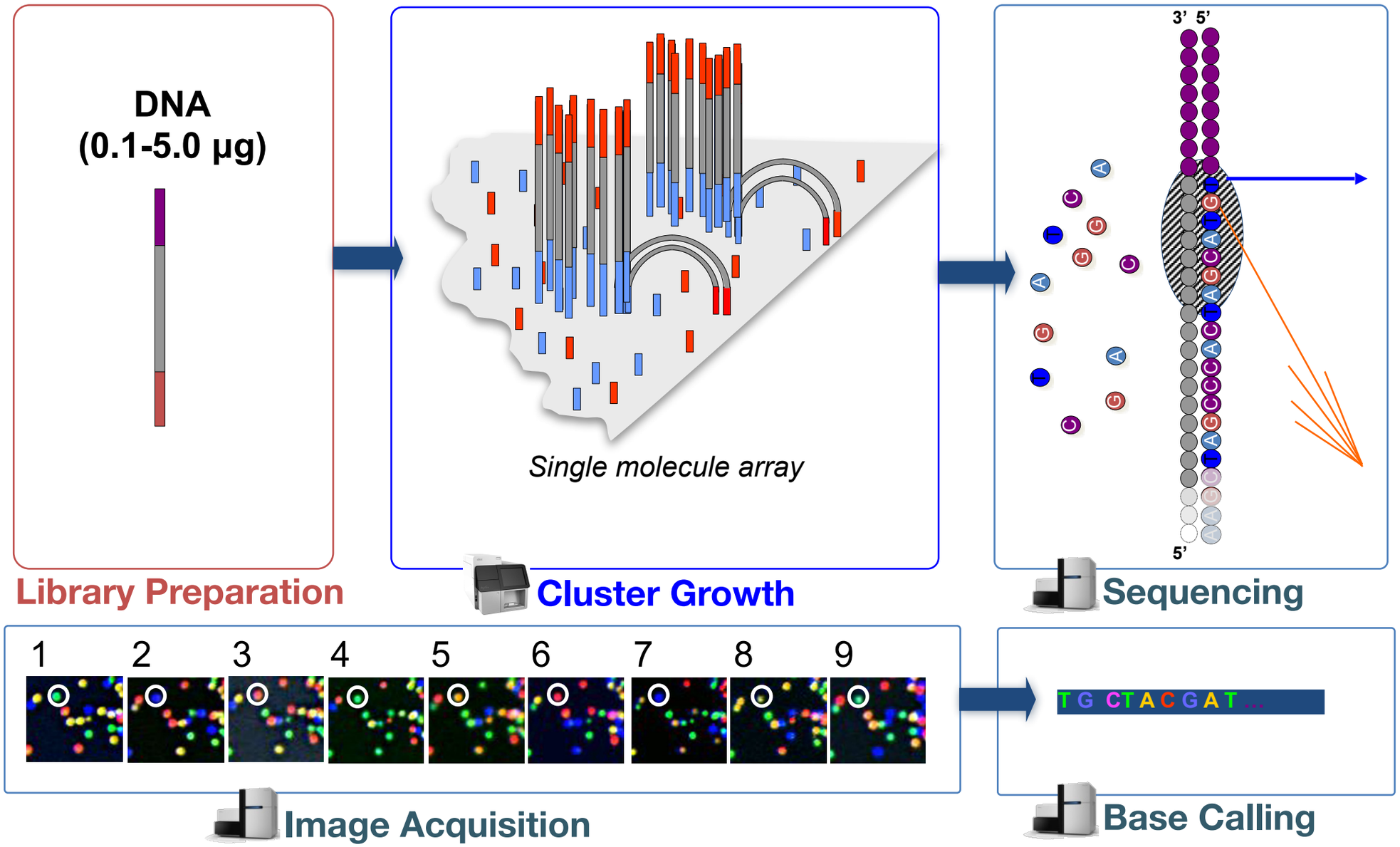
http://onetipperday.sterding.com/2012/07/how-to-tell-which-library-type-to-use.html



RNA-Seq library prep

http://rnaseq.uoregon.edu/#rna-prep

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



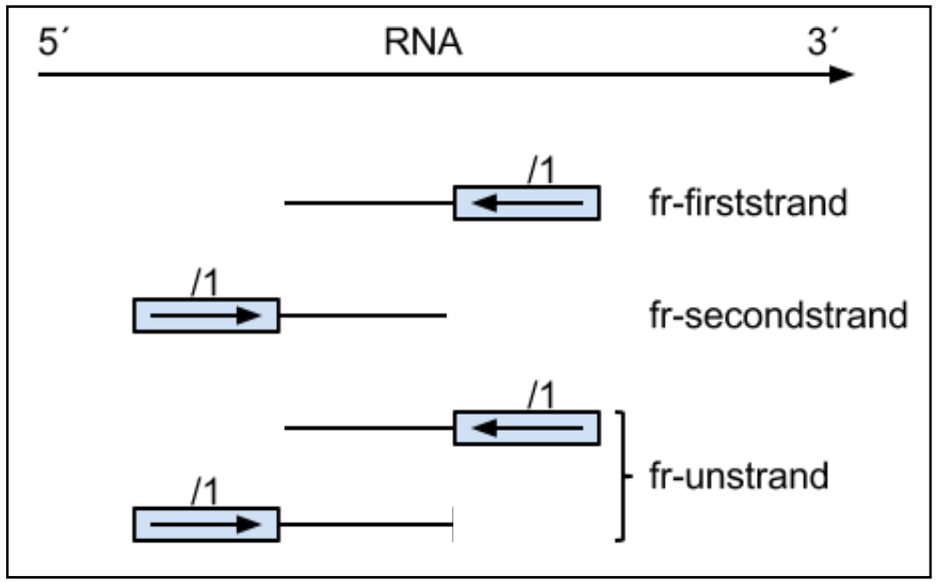
Illumina: Sequencing Workflow

- ✓ SE - Single end dataset => Only Read1
- ✓ PE - Paired-end dataset => Read1 + Read2
  - can be 2 separate FastQ files or just one with interleaved pairs
  - <u>insert</u> refers to the DNA fragment** flanked by the adapters

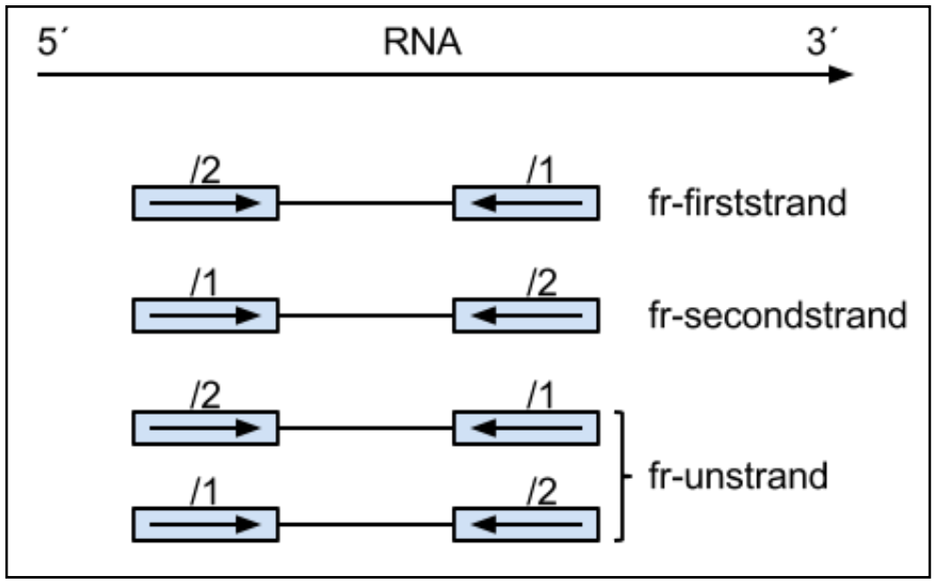
*** "fragment" during library prep (Illumina) refers to the whole piece of DNA (insert + adapters). But, during downstream processing steps "fragment" can sometime refer to only the insert.*

Options for sequencing

Single-end read



Paired-end reads



Strandedness in the context of SE or PE

http://onetipperday.sterding.com/2012/07/how-to-tell-which-library-type-to-use.html

- ✓ SE - Single end dataset => Only Read1
- ✓ PE - Paired-end dataset => Read1 + Read2
  - can be 2 separate FastQ files or just one with interleaved pairs
  - insert refers to the DNA fragment** flanked by the adapters
- ✓ Read length - 50bp - 250bp, depends on the sequencer

Options for sequencing

http://tucf-genomics.tufts.edu/home/faq

				
MiniSeq System	**MiSeq Series**	**NextSeq Series**	**HiSeq Series**	**HiSeq X Series***
Amplicon, targeted RNA, small RNA, and targeted gene panel sequencing.	Small genome, amplicon, and targeted gene panel sequencing.	Everyday exome, transcriptome, and targeted resequencing.	Production-scale genome, exome, transcriptome sequencing, and more.	Population- and production-scale whole-genome sequencing.

A poster from Illumina documenting library preps for various applications:
http://www.illumina.com/content/dam/illumina-marketing/documents/applications/ngs-library-prep/ForAllYouSeqMethods.pdf

Illumina's sequencing systems

http://www.illumina.com/systems/sequencing.html

Outline

- Library preparation
- Experimental and Practical Considerations
- Analysis workflow and options
- *Commonly used file formats*

Experimental and Practical considerations

1. Experimental Design
2. Poly(A) enrichment or ribosomal RNA depletion?
3. Single-end or Paired-end data?
4. Stranded libraries?
5. How much sequencing data to collect?
6. Multiplexing

Experimental and Practical considerations

1. Experimental design

- Technical replicates: Illumina has low technical variation unlike microarrays, hence technical replicates are unnecessary.
- Batch effects are still a problem. Be consistent!
- Biological replicates, are absolutely essential. Have at least 3!
- For differential gene expression, pooling RNA from multiple biological replicates is usually not advisable; do so only if you have multiple pools from each experimental condition.

Experimental and Practical considerations

2. Poly(A) enrichment or ribosomal RNA depletion?

Depends on which RNA entities you are interested in…

- 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, then do a ribosomal RNA depletion.

Experimental and Practical considerations

3. Single-end or Paired-end data?

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

Experimental and Practical considerations

3. Single-end or Paired-end data?

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

- For differential gene expression, which one you pick depends on-
  - If you are specifically interested in isoform-level differences
  - The abundance of paralogous genes in your system of interest
  - Your budget, paired-end data is usually 2x more expensive

Experimental and Practical considerations

4. Stranded libraries?

Stranded libraries are now standard with Illumina's TruSeq stranded RNA-Seq kits. This means that with a great amount of certainty you can identify which strand of DNA the RNA was transcribed from.

3 types of libraries –

- Reverse (firststrand)– reads resemble the complementary sequence (TruSeq)
- Unstranded
- Forward (secondstrand) – reads resemble the gene sequence

Experimental and Practical considerations

5. How much sequencing data to collect?

Depends heavily on the size of the transcriptome of interest:

- 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.

Experimental and Practical considerations

5. How much sequencing data to collect?

Depends heavily on the size of the transcriptome of interest:

- 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.
- Only ~2% of the human genome transcribes protein-coding RNA.
- In general, some mRNAs will be much more abundant than others, and some genes are much longer than others, so the coverage metric breaks down.
- For human samples ~30-50 million reads/sample is recommended (ENCODE guidelines).
- More replicates >> More reads (for standard DGE).
- Your budget

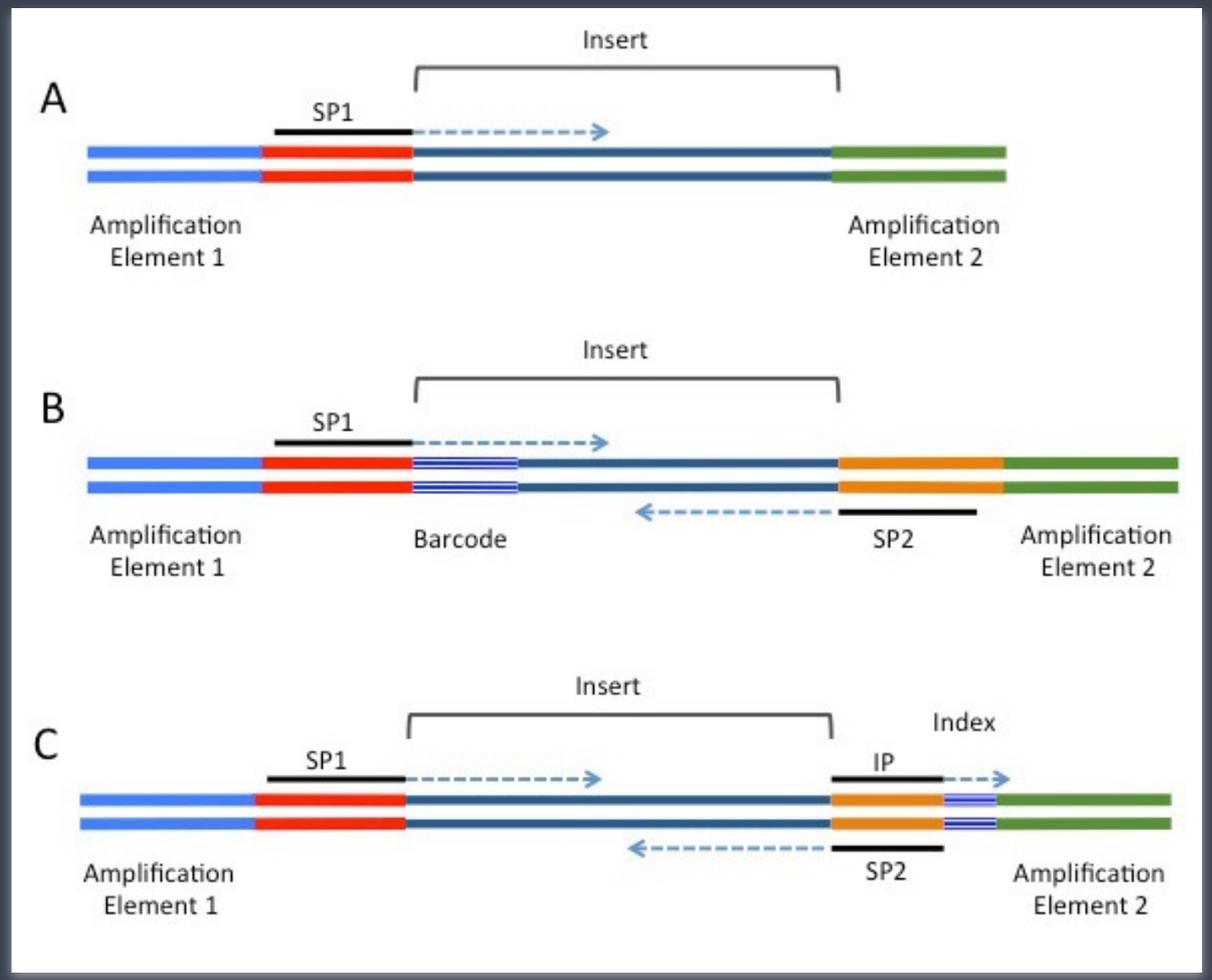
Experimental and Practical considerations

6. Multiplexing (with barcodes and indices)

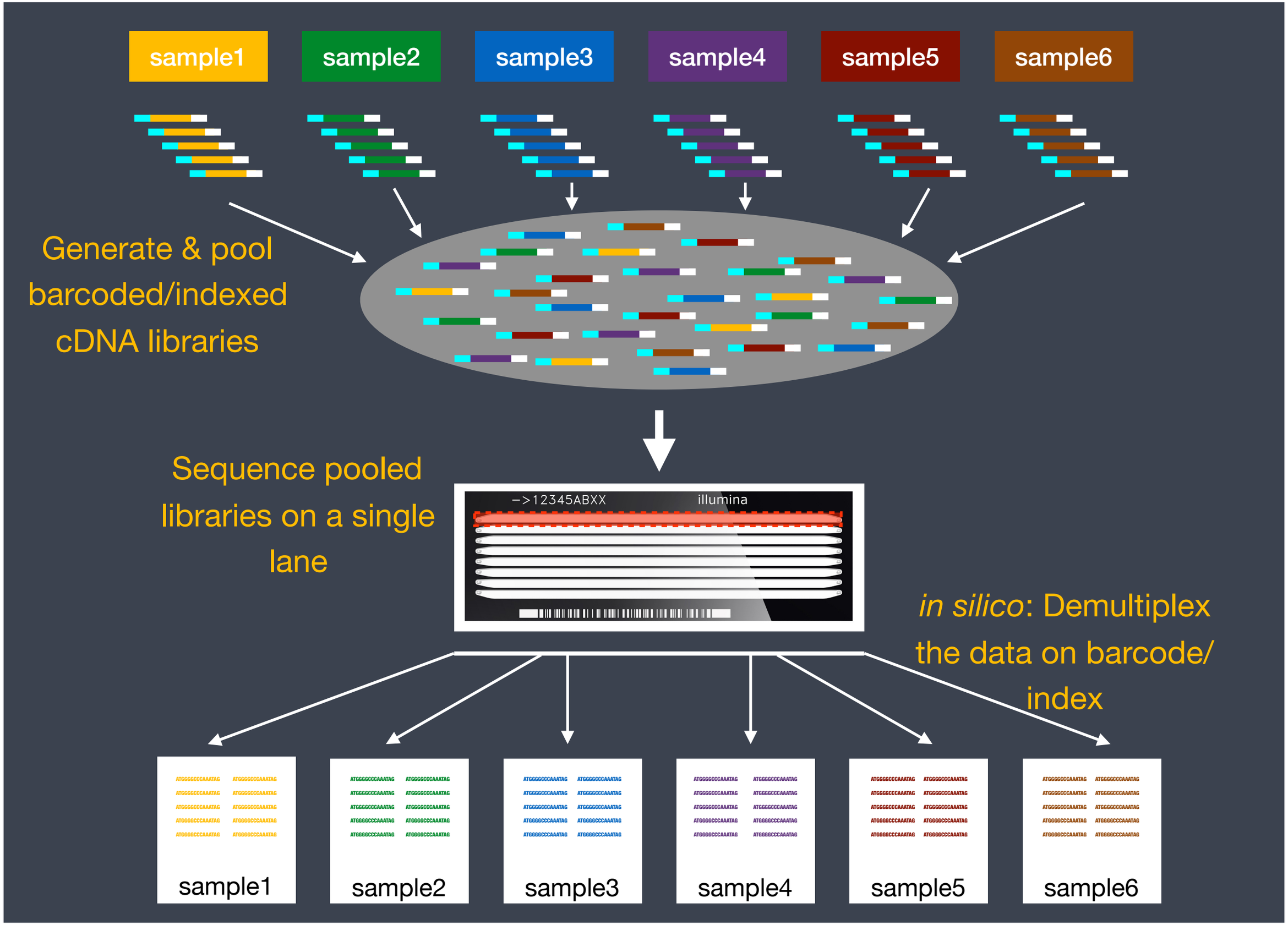
- Charges for sequencing are usually per lane of the flow cell
- Each lane generates ~150 million reads
- For RNA-Seq, the required data per sample is much lower than that
- Sequencing of multiple samples per lane possible with addition of barcodes and special indices to adapters or directly to each cDNA prep

Experimental and Practical considerations

6. Multiplexing (with barcodes and indices)



http://rnaseq.uoregon.edu/

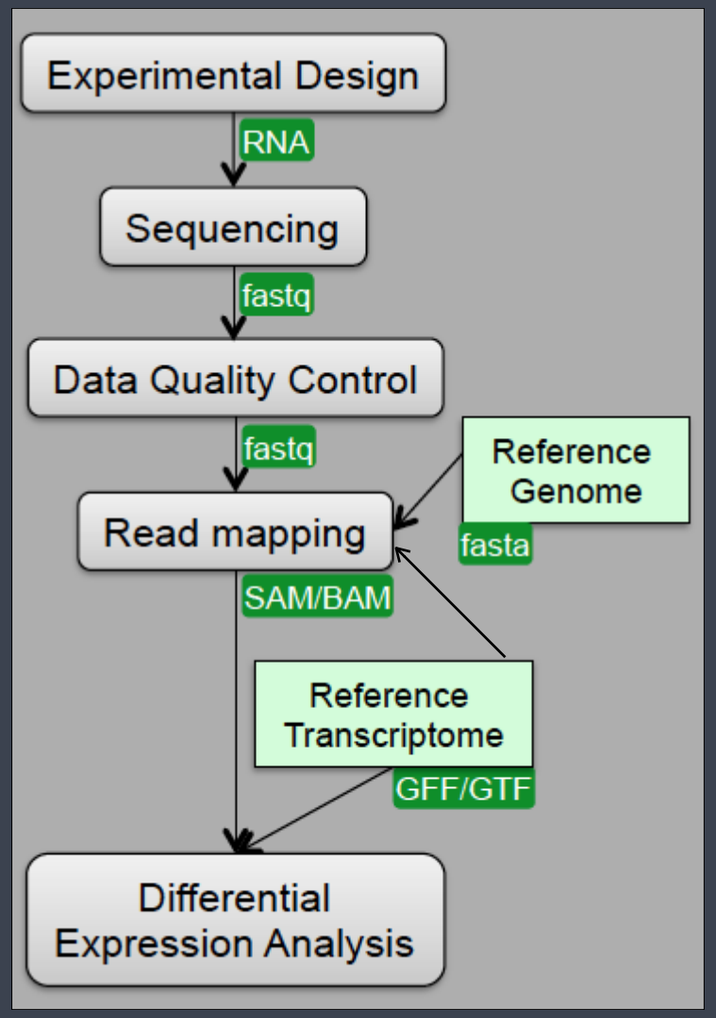


http://www.illumina.com/

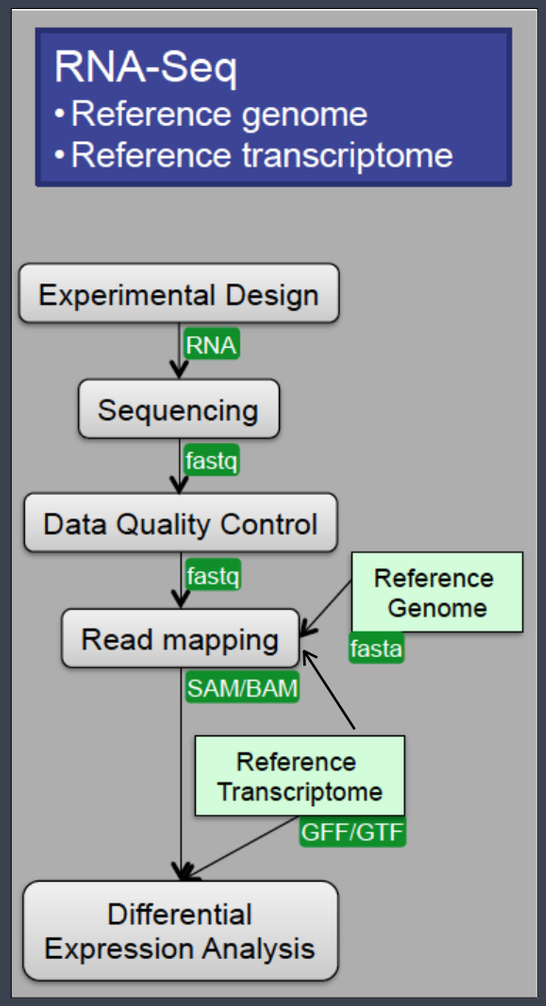
Outline

- Library preparation
- Experimental and Practical Considerations
- Analysis workflow and options
- *Commonly used file formats*

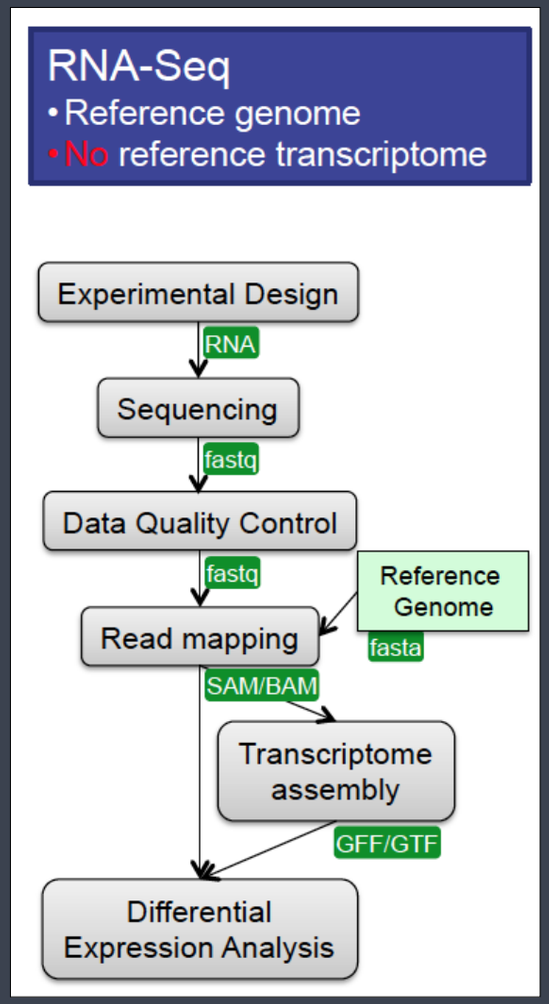
RNA-Seq analysis workflow and options



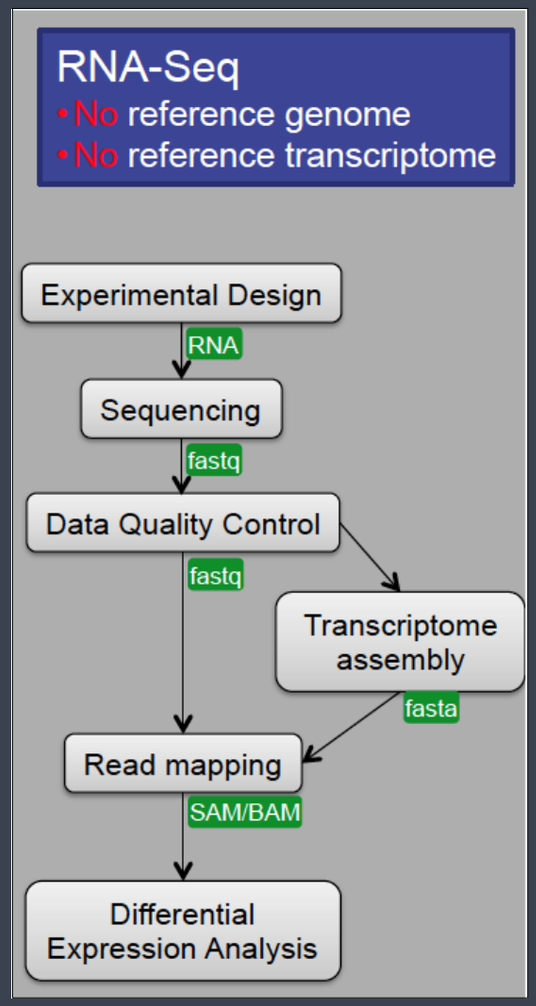
RNA-Seq analysis workflow and options



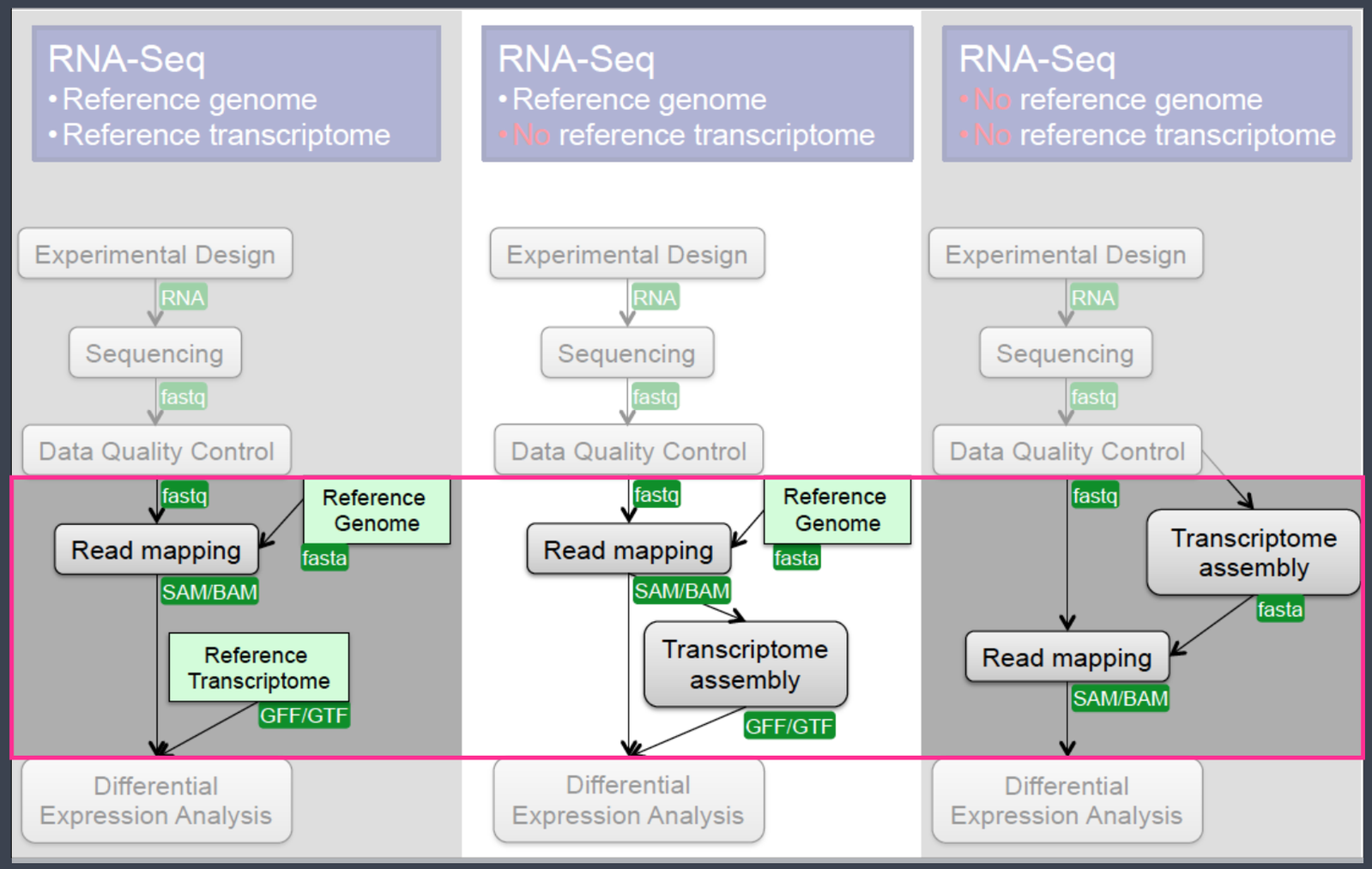
RNA-Seq analysis workflow and options



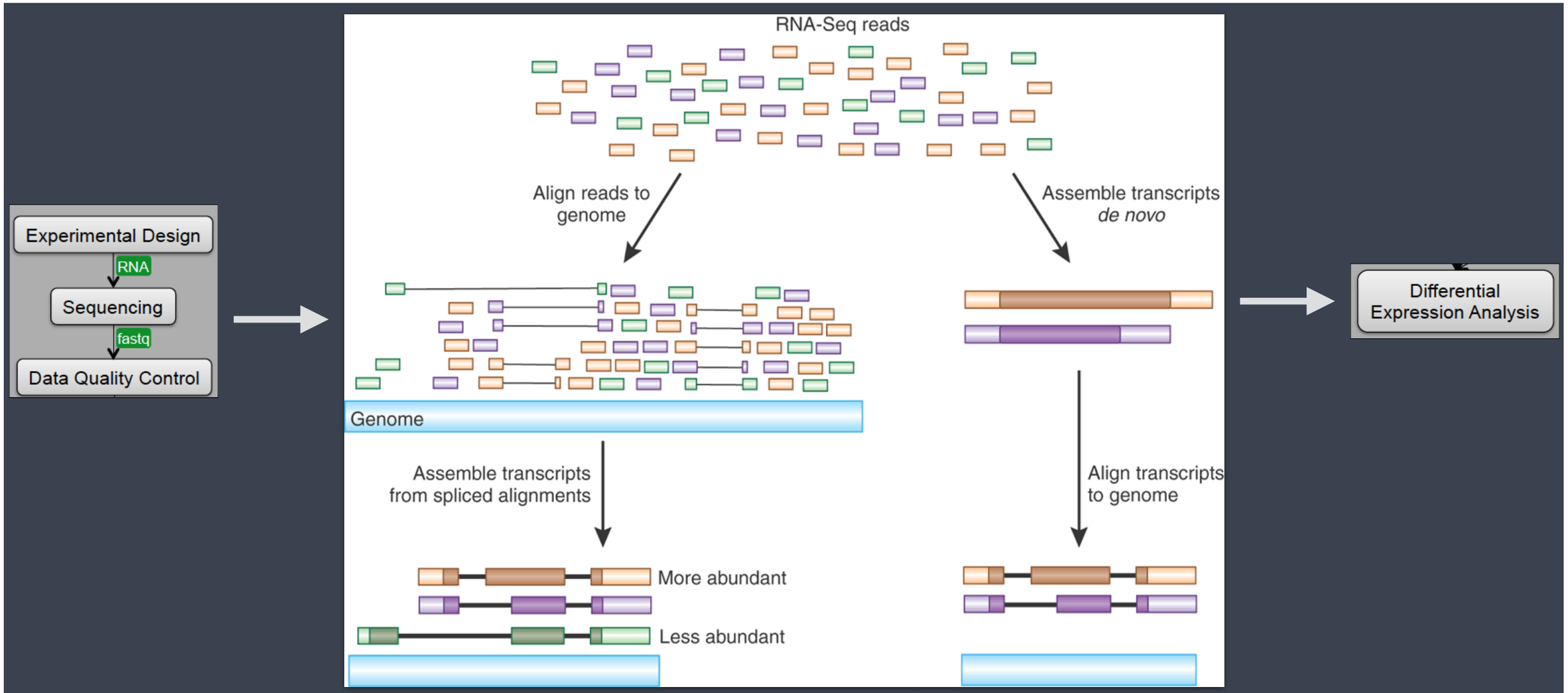
RNA-Seq analysis workflow and options



RNA-Seq analysis workflow and options



RNA-Seq analysis workflow and options



Haas & Zody, Nature Biotechnology 28, 421–423 (2010)

Outline

- Library preparation
- Experimental and Practical Considerations
- Analysis workflow and options
- *Commonly used file formats*

Common data types and file formats

- You will encounter 3 major types of data, with several associated file formats:
  - Sequence data
  - Genome feature data
  - Alignment data
- File formats represent these data types in a structured manner, and can combine multiple data types in one file.
- Some file formats are not human-readable (binary).
- Many are human readable, but extremely large; never use Word or Excel to open these!

Simple sequence formats

- FASTA
- FASTQ

Feature formats

- GTF/GFF (GTF v2, and GFF v3)
- SAM/BAM
- UCSC formats (BED, WIG, etc.)

Feature formats

- Tab-delimited
- Contain specific information about genome (or assembly) coordinates
- May or may not include sequence data
- The chromosome (or contig) names MUST match the reference sequence name
  - 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'** versus **Ensembl/NCBI – '1'**
  - Best practice: get these from the same source as the reference genome

Feature formats: GTF (Gene Transfer Format)

- Evolved from Sanger Centre GFF (gene feature format) originally, but repeatedly modified
- Differences in representation of information make it distinct from GFF
- **1-based coordinates**

```
chr1 unknown  exon         113217048  113217252  .   +    .  gene_id "MOV10";p_id "P5535";transcript_id "NM_001130079"
chr1 unknown  exon         113217048  113217351  .   +    .  gene_id "MOV10";p_id "P5535";transcript_id "NM_020963"
chr1 unknown  exon         113217470  113217671  .   +    .  gene_id "MOV10";p_id "P5535";transcript_id "NM_001130079"
chr1 unknown  CDS          113217535  113217671  .   +    0  gene_id "MOV10";p_id "P5535";transcript_id "NM_001130079"
chr1 unknown  start_codon  113217535  113217537  .   +    .  gene_id "MOV10";p_id "P5535";transcript_id "NM_001130079"
```

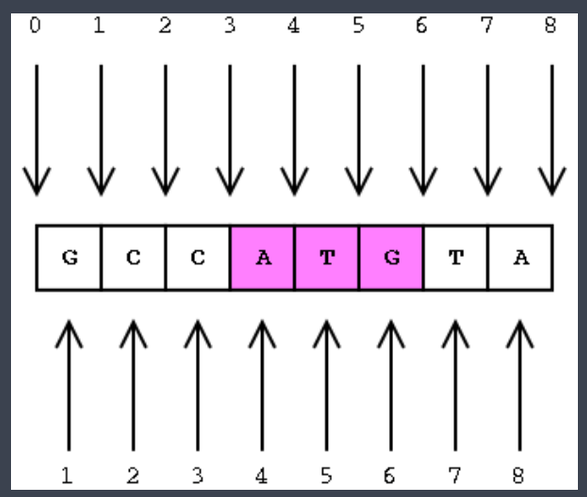
Chromosome ID
Source
Gene feature
Start location
End location
Score (user defined)
Strand
Reading frame
Attributes

Genomic coordinates can be represented in 2 ways

Where is 1 and where is 8?

Genomic coordinates can be represented in 2 ways

Coords	***Where is ATG?***	***Length***
0-based (half-open) preferred by programmers	`( 3, 6 ]`	Len = end - start



1-based (closed) preferred by biologists	`[ 4, 6 ]`	Len = end – start + 1

Feature formats: GTF (Gene Transfer Format)

- Evolved from Sanger Centre GFF (gene feature format) originally, but repeatedly modified
- Differences in representation of information make it distinct from GFF
- **1-based coordinates**
- Source of the GTF is important, subtle differences between an Ensembl version and a UCSC version can cause issues.

```
chr1 unknown  exon         113217048  113217252  .  +  .  gene_id "MOV10";p_id "P5535";transcript_id "NM_001130079"
chr1 unknown  exon         113217048  113217351  .  +  .  gene_id "MOV10";p_id "P5535";transcript_id "NM_020963"
chr1 unknown  exon         113217470  113217671  .  +  .  gene_id "MOV10";p_id "P5535";transcript_id "NM_001130079"
chr1 unknown  CDS          113217535  113217671  .  +  0  gene_id "MOV10";p_id "P5535";transcript_id "NM_001130079"
chr1 unknown  start_codon  113217535  113217537  .  +  .  gene_id "MOV10";p_id "P5535";transcript_id "NM_001130079"
```

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

Feature formats: GFF3 (Gene Feature Format)

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

Feature formats: GFF3 versus GTF

GFF3 – Gene feature format

chr1	ensembl_havana	transcript	112674487	112700739	.	+	.	ID=transcript:ENST00000369645;Parent=gene:ENSG00000155363;Name=MOV10-006;biotype=protein_coding;ccdsid=CCDS853.1;havana_transcript=OTTHUMT00000032911;havana_version=1;tag=basic;transcript_id=ENST00000369645;transcript_support_level=5 (assigned to previous version 4);version=5
chr1	havana	exon	112674487	112674729	.	+	.	Parent=transcript:ENST00000369645;Name=ENSE00001450533;constitutive=0;ensembl_end_phase=-1;ensembl_phase=-1;exon_id=ENSE00001450533;rank=1;version=1
chr1	havana	five_prime_UTR	112674487	112674729	.	+	.	Parent=transcript:ENST00000369645
chr1	havana	five_prime_UTR	112674848	112674912	.	+	.	Parent=transcript:ENST00000369645
chr1	havana	exon	112674848	112675049	.	+	.	Parent=transcript:ENST00000369645;Name=ENSE00003676444;constitutive=0;ensembl_end_phase=2;ensembl_phase=-1;exon_id=ENSE00003676444;rank=2;version=1

GTF – Gene transfer format

chr1	havana	transcript	112674487	112700739	.	+	.	gene_id "ENSG00000155363"; gene_version "18"; transcript_id "ENST00000369645"; transcript_version "5"; gene_name "MOV10"; gene_source "ensembl_havana"; gene_biotype "protein_coding"; havana_gene "OTTHUMG00000011906"; havana_gene_version "1"; transcript_name "MOV10-006"; transcript_source "havana"; transcript_biotype "protein_coding"; tag "CCDS"; ccds_id "CCDS853"; havana_transcript "OTTHUMT00000032911"; havana_transcript_version "1"; tag "basic"; transcript_support_level "5 (assigned to previous version 4)";
chr1	havana	exon	112674487	112674729	.	+	.	gene_id "ENSG00000155363"; gene_version "18"; transcript_id "ENST00000369645"; transcript_version "5"; exon_number "1"; gene_name "MOV10"; gene_source "ensembl_havana"; gene_biotype "protein_coding"; havana_gene "OTTHUMG00000011906"; havana_gene_version "1"; transcript_name "MOV10-006"; transcript_source "havana"; transcript_biotype "protein_coding"; tag "CCDS"; ccds_id "CCDS853"; havana_transcript "OTTHUMT00000032911"; havana_transcript_version "1"; exon_id "ENSE00001450533"; exon_version "1"; tag "basic"; transcript_support_level "5 (assigned to previous version 4)";
chr1	havana	five_prime_utr	112674487	112674729	.	+	.	gene_id "ENSG00000155363"; gene_version "18"; transcript_id "ENST00000369645"; transcript_version "5"; gene_name "MOV10"; gene_source "ensembl_havana"; gene_biotype "protein_coding"; havana_gene "OTTHUMG00000011906"; havana_gene_version "1"; transcript_name "MOV10-006"; transcript_source "havana"; transcript_biotype "protein_coding"; tag "CCDS"; ccds_id "CCDS853"; havana_transcript "OTTHUMT00000032911"; havana_transcript_version "1"; tag "basic"; transcript_support_level "5 (assigned to previous version 4)";
chr1	havana	five_prime_utr	112674848	112674912	.	+	.	gene_id "ENSG00000155363"; gene_version "18"; transcript_id "ENST00000369645"; transcript_version "5"; gene_name "MOV10"; gene_source "ensembl_havana"; gene_biotype "protein_coding"; havana_gene "OTTHUMG00000011906"; havana_gene_version "1"; transcript_name "MOV10-006"; transcript_source "havana"; transcript_biotype "protein_coding"; tag "CCDS"; ccds_id "CCDS853"; havana_transcript "OTTHUMT00000032911"; havana_transcript_version "1"; tag "basic"; transcript_support_level "5 (assigned to previous version 4)";
chr1	havana	exon	112674848	112675049	.	+	.	gene_id "ENSG00000155363"; gene_version "18"; transcript_id "ENST00000369645"; transcript_version "5"; exon_number "2"; gene_name "MOV10"; gene_source "ensembl_havana"; gene_biotype "protein_coding"; havana_gene "OTTHUMG00000011906"; havana_gene_version "1"; transcript_name "MOV10-006"; transcript_source "havana"; transcript_biotype "protein_coding"; tag "CCDS"; ccds_id "CCDS853"; havana_transcript "OTTHUMT00000032911"; havana_transcript_version "1"; exon_id "ENSE00003676444"; exon_version "1"; tag "basic"; transcript_support_level "5 (assigned to previous version 4)";

Always check which of the two formats is accepted by the application you're using

Alignment formats: SAM

- SAM – Sequence Alignment/Map format
- SAM file format stores alignment information
- Plain text
- **1-based coordinates**
- 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)

Alignment 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 (bgzip)
- **0-based coordinates**
- Makes the alignment information easily accessible to downstream applications
- Files are typically very large: ~ 1/5 of SAM, but still very large

Commonly used file formats

- FASTA
- FASTQ – Fasta with quality
- GFF3 – Gene feature format (genome interval ++)
- GTF – Gene transfer format (genome interval ++)
- SAM – Sequence Alignment/Map format
- BAM – Binary Sequence Alignment/Map format
- *Bed – Basic genome interval (0-based coordinates)*
- *Wiggle (wig, bigwig) – tab-limited format to represent values, usually associated with a set of genomic coordinates (0-based coordinates)*

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

- 3 replicates from each sample group (transfected HEK293F cell lines)
- stranded libraries (dUTP method)
- single-end, 100 nt long reads on Illumina HiSeq-2500
- ~40 million reads/sample (we will be using a tiny subset from chromosome 1, ~150,000 - 300,000 reads)

Control

Replicate1 Replicate2

Replicate3

Mov10 oe (overexpression)

Replicate1 Replicate2

Replicate3

Mov10 kd (knockdown)

Replicate1 Replicate2

These materials have been developed by members of the teaching team at the Harvard Chan Bioinformatics Core (HBC). These are open access materials distributed under the terms of the Creative Commons Attribution license (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.