RNA-Seq

https://dbsloan.github.io/TS2022/

Regulation of Gene Expression

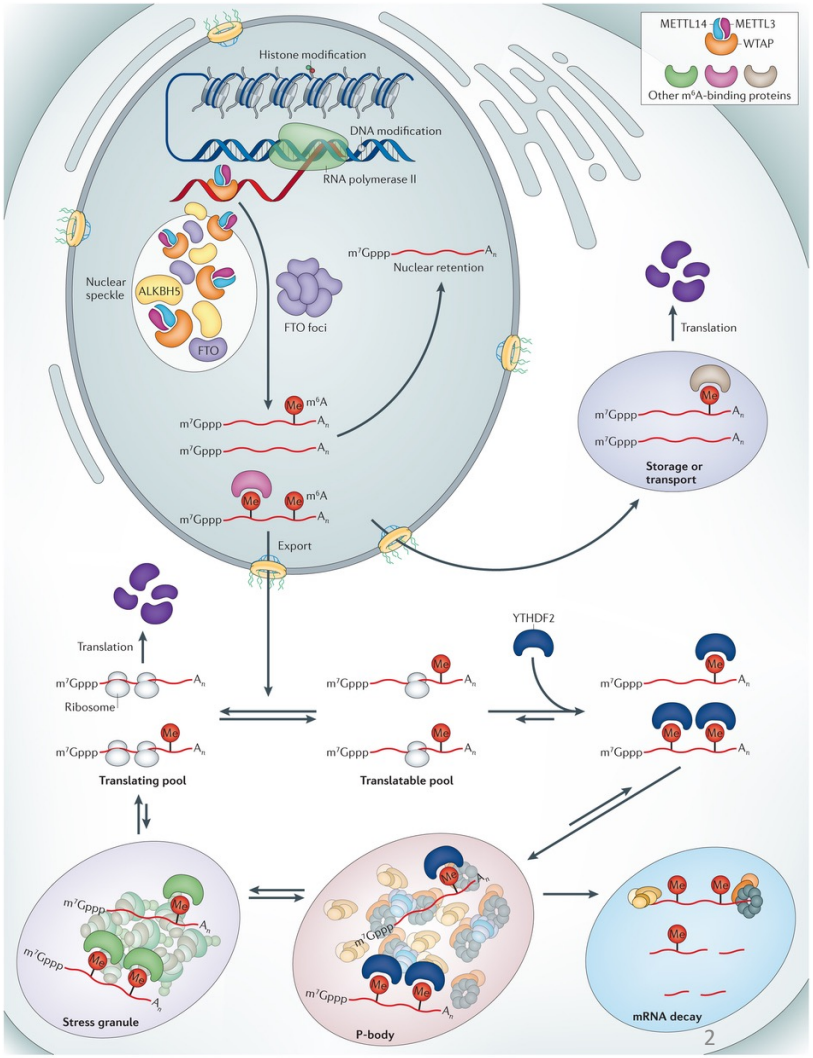
Regulation of transcription:
- Transcription factors
- Histone modifications
- DNA methylation

Regulation of RNA processing:
- Polyadenylation
- Splicing
- Capping
- RNA export

Regulation of translation:
- mRNA decay
- Translational repression
- Sequestration

RNA-seq measures steady state mRNA levels

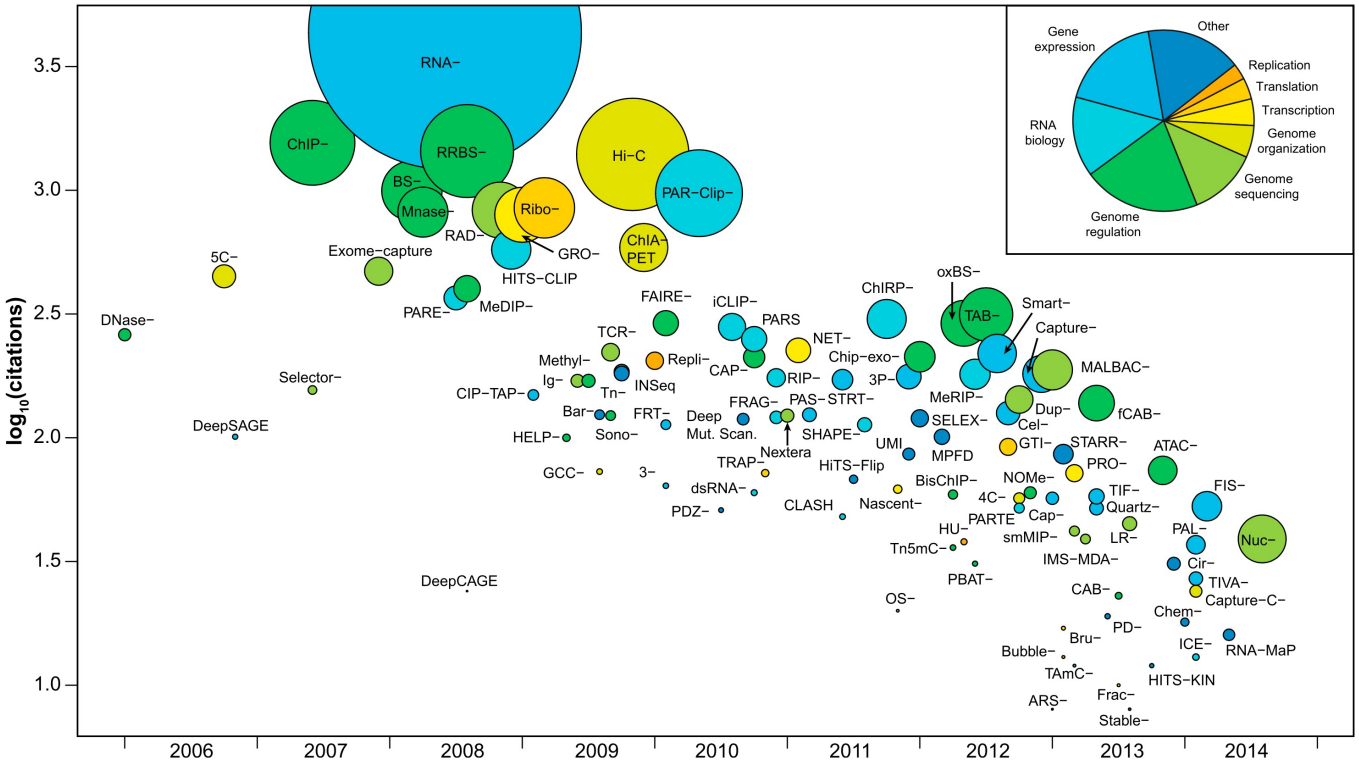
Posttranslational regulation:
- Chemical modifications (e.g. phosphorylation)
- Protein turnover (proteolysis)



Fu et al. (2014)

RNA-seq: a very common application of HTS

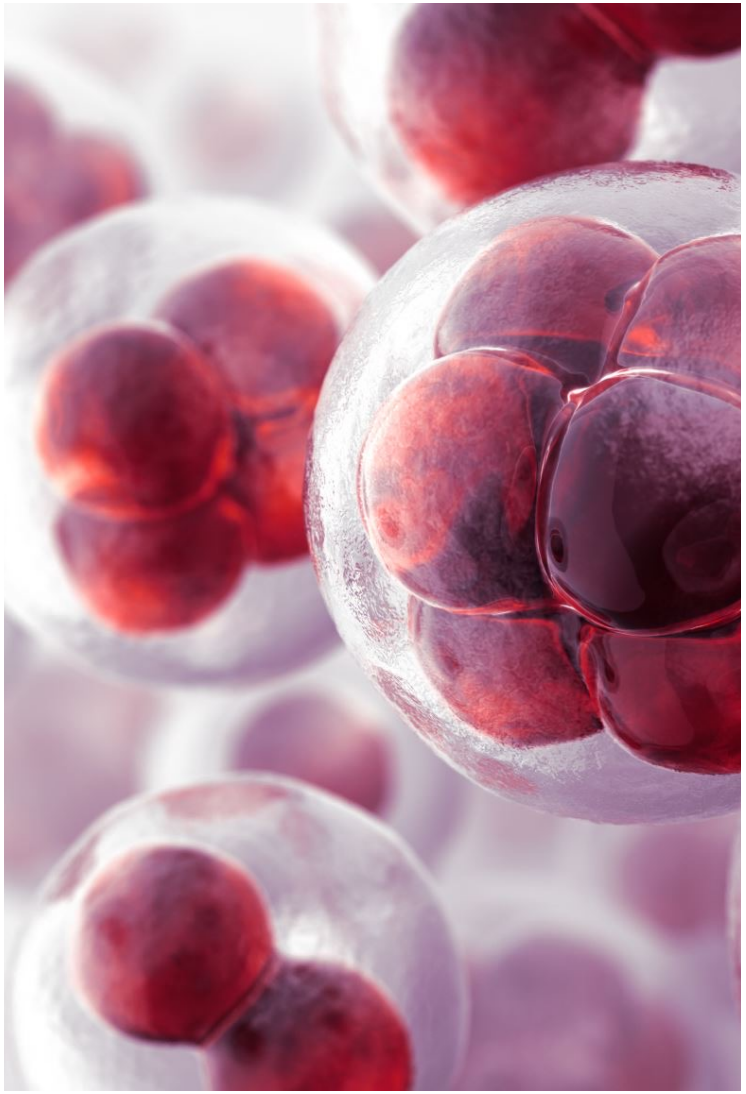
RNA-seq is most commonly used to explore the transcriptome and compare gene expression across different conditions.



Reuter et al 2015

Comparative Studies

Want to compare differences in gene expression? RNA-seq!



Important considerations for RNA-seq experiments

Experiment design:

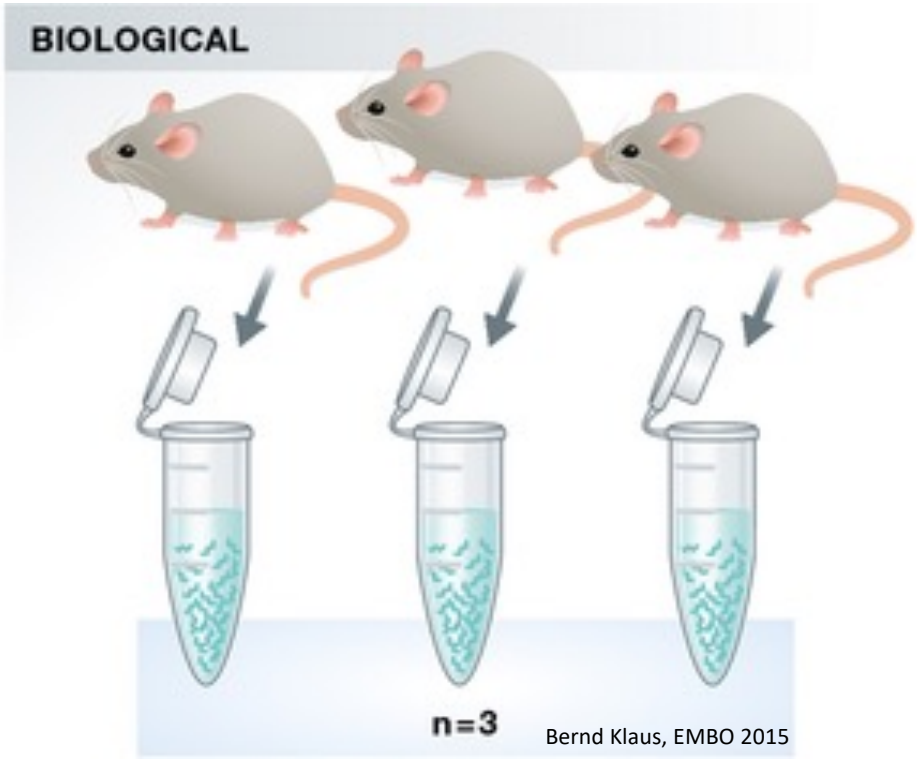
- Carefully plan experiment to minimize confounding variables
  - Stage match cells/organisms
  - Grow cells/organisms in parallel under same conditions
  - Collect the same composition of cells/tissue and at the same time
- Include biological replicates
  - At least 2, typically 3-4

Sample preparation:

- Treat RNA samples identically
- Check the integrity of the RNA before preparing libraries
- Prepare libraries in parallel

Replication

Not needed for RNA-seq



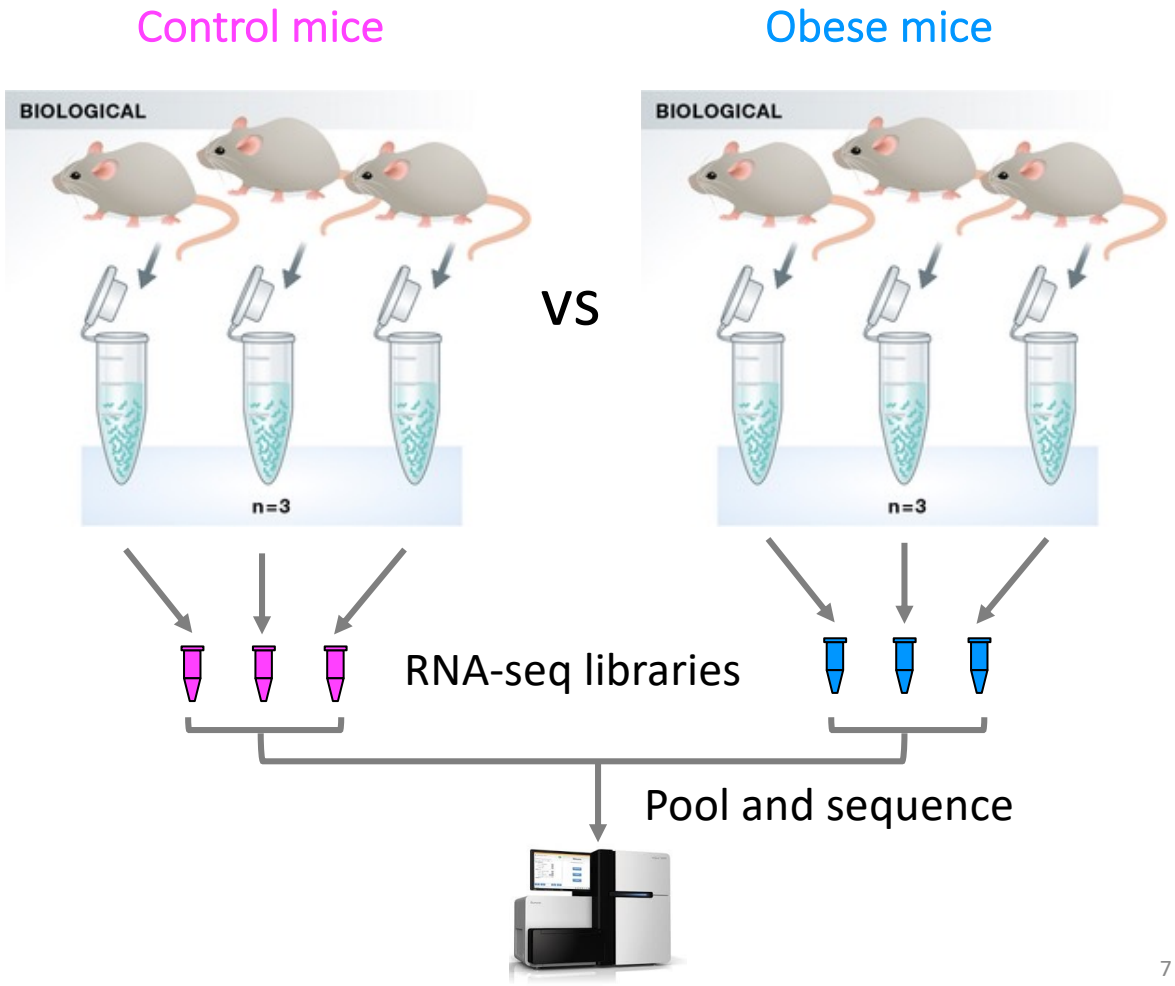
Nearly essential for RNA-seq

Grow mice in parallel

Collect tissue in parallel

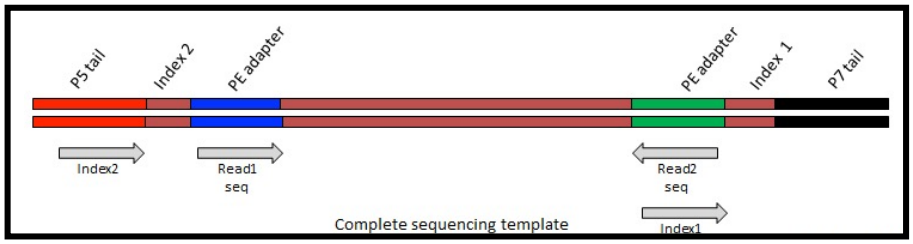
Extract RNA in parallel

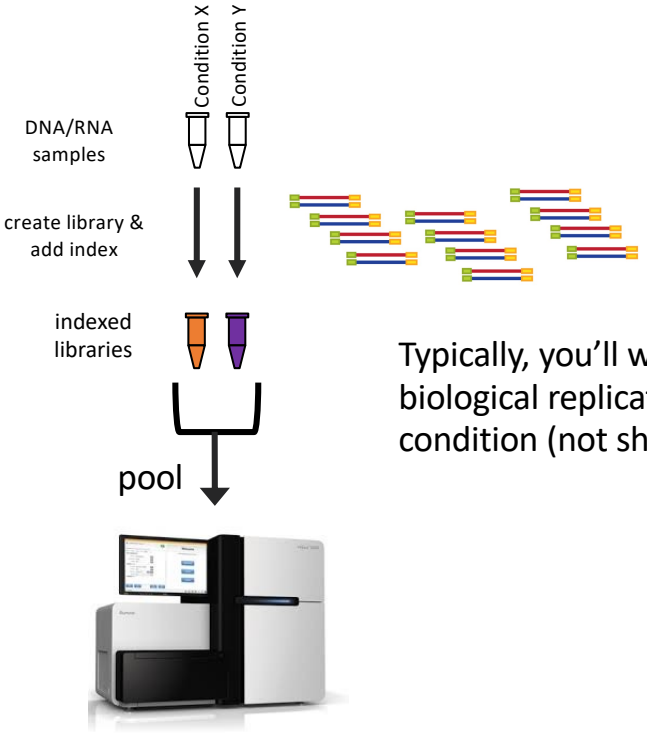
Prepare RNA-seq libraries in parallel



Sequencing

Libraries

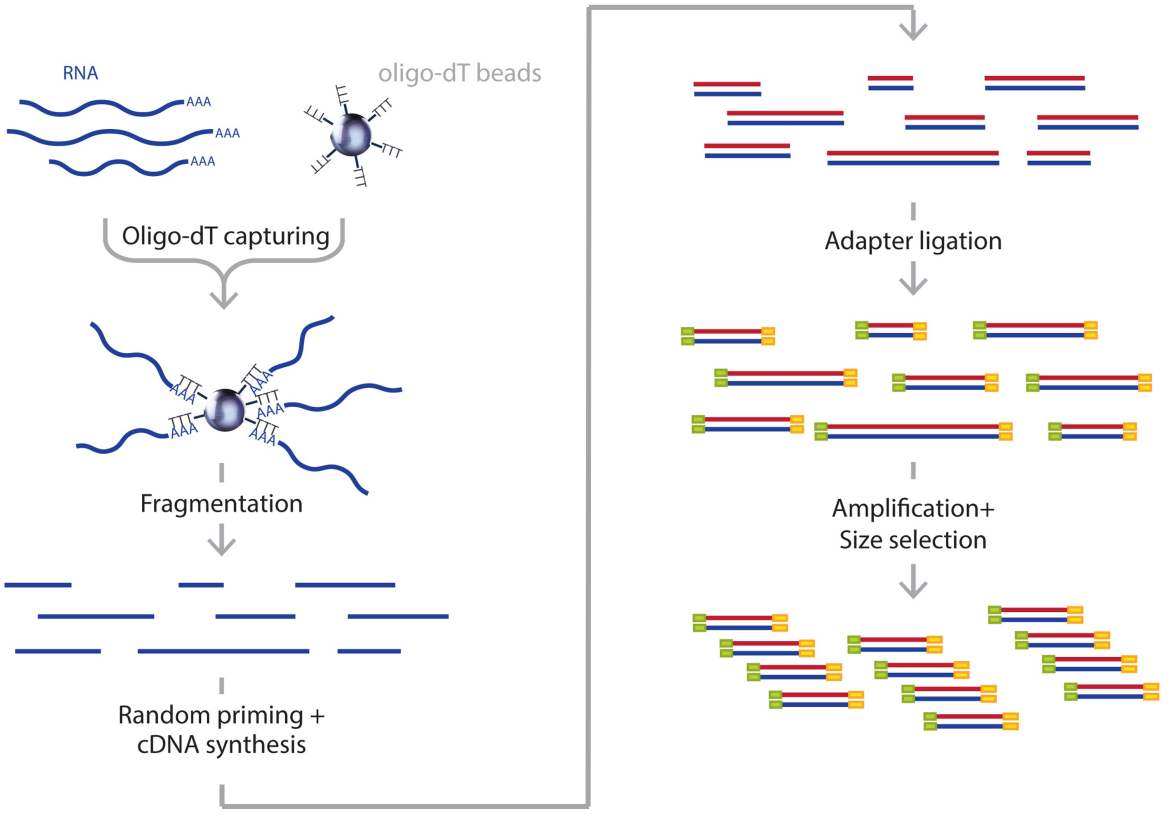




Typically, you'll want 3 biological replicates per condition (not shown here).

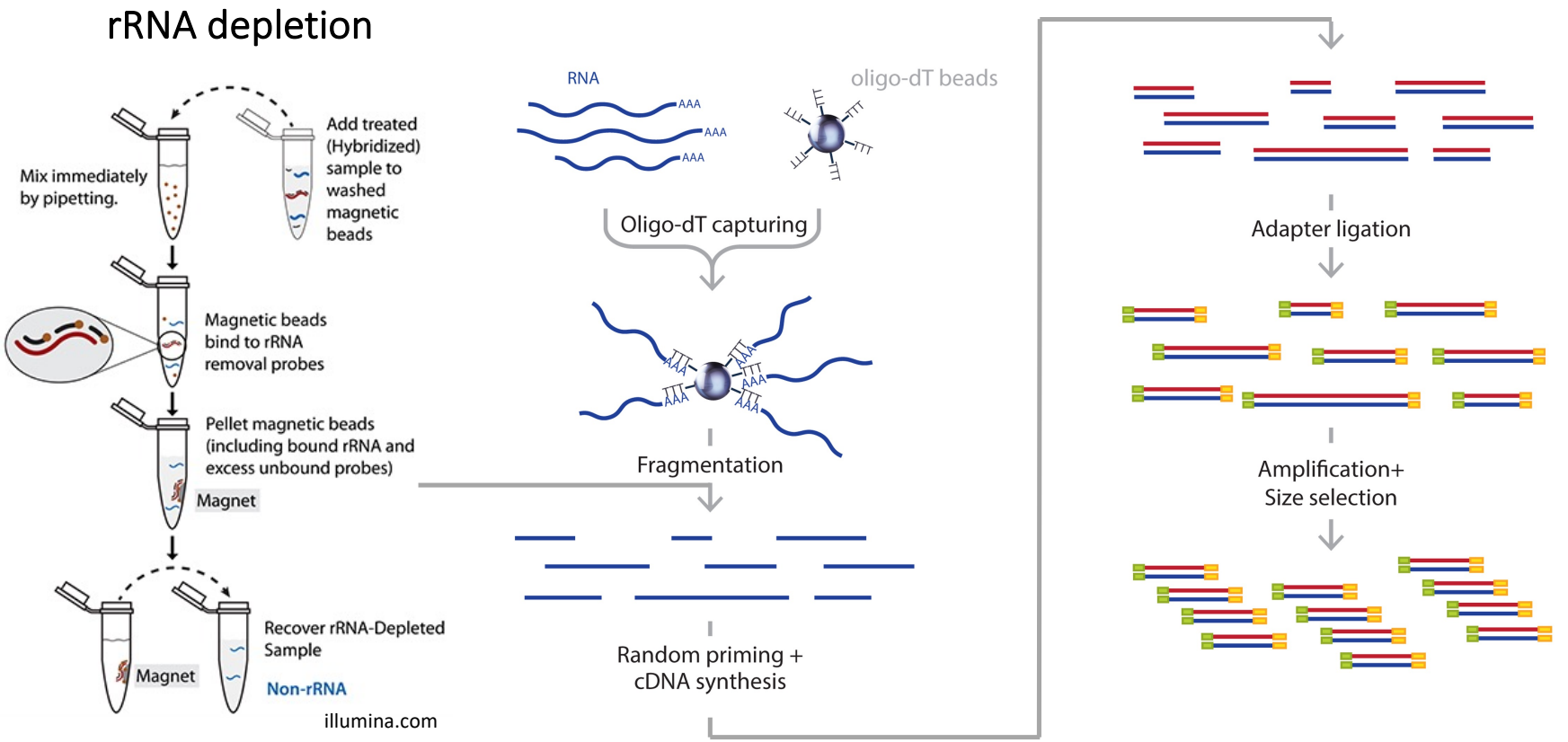
Aim for ~30 million reads per library

RNA-seq library prep



Zhernakova et al. (2009)

RNA-seq library prep



Zhernakova et al. (2009)

Sample preparation

- Starting RNA
  - Typically 1-5 ug of high-quality total RNA is ideal.
- Sequencing depth
  - Typically you want 20-30 million high quality reads/library.
- Considerations
  - Strand specific (default is yes)
  - Single-end or paired-end (single-end is typically sufficient)
  - Long reads vs short reads (short Illumina reads, 50-150 nt, are often sufficient)
  - rRNA depletion or oligo-dT
  - Low quantity/single cell

Files Needed for RNA-seq

High-throughput sequencing data (2 files per library if paired-end data)
- fastq format

Genome sequence, if available
- fasta format

mRNA coordinates, if available
- gff or gtf format

Alternatively, transcript sequences
- fasta format

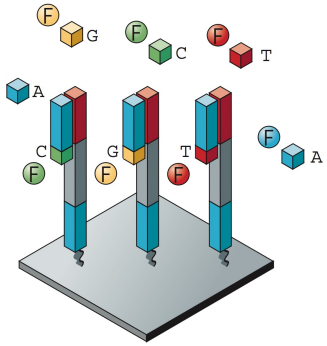
FASTQ format commonly used for HTS data

Index sequence

Read 1
```
1 @D64TDFP1:248:C50DMACXX:5:1101:1241:2095 1:N:0:ATCACG
2 CACCGCCCGTCGCTATCCGGGACTGGAATTCTCGGGTGCCAAGGAACTCCA
3 +
4 CCCFFFFFHHHHHJIJGHJJJJIJJJJJGGGFFFFEABDHHHFHFF@@DD>
```

Read 2
```
1 @D64TDFP1:248:C50DMACXX:5:1101:1371:2154 1:N:0:ATCACG
2 TCAATATTTGCATAGGGTATCTGGAATTCTCGGGTGCCAAGGAACTCCAGT
3 +
4 CCCFFFFFHHHHHJJJJGFHIJJJJJJJJJJJJJFHHIIJJHGHJFGHJJI
```

Read 3
```
1 @D64TDFP1:248:C50DMACXX:5:1101:1461:2205 1:N:0:ATCACG
2 GAAAGACGTCTTCCTAGATTATGGAATTCTCGGGTGCCAAGGAACTCCAGT
3 +
4 CCCFFFFFHHHHHJJJJJJJJJJJIJJJJJJJJJHIJJJJJGIIJFGIJJJ
```



Metzker, M.L. (2010) NRG

Line 1: sequence ID, description, and index; begins with @

Line 2: sequence; contains only A, C, T, G, and N

Line 3: optional sequence ID; begins with +

Line 4: signal quality of each base, cryptic code, phred 33 or 64

Genome sequences are often in FASTA format

FASTA: DNA sequence alignment software. The software gave rise to the fasta format, now ubiquitous sequence file format.

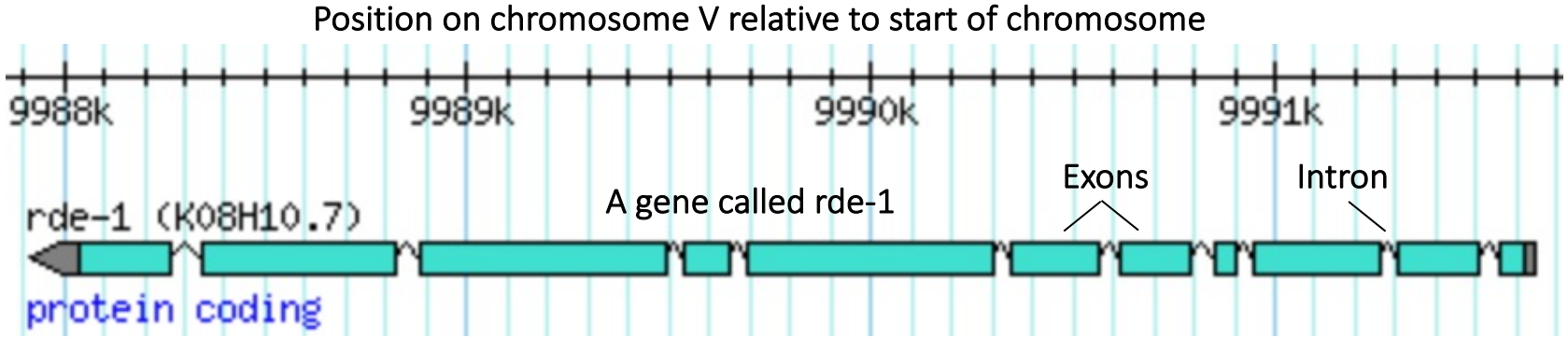
```
>sequence1_description
AGCTAGCATCGACTAGCTACGATCGATCGATCACGAGCTACGACGTAGGCATGGGGGCTTACGATGCTA
CGGCGGAGCTACGGCGACTGCGATCTACGGCGATCGACGGACGGACGTCAGGCGACGATCTATCATCTA
TCGAGCGAGCTACTTACTCTTCTCTATCTACTTATCCCCTTCTTAGGGGTTGATTAGTCTAGCTGGTAC
GATCGAGCGATCTAGAGCGATCGACGAGCTGACGGACGTACTTACTATCGTAGCGACTACTTC

>sequence2_description
CTCTAGCATCGACTAGCTACGATCGATCGATCACGAGCTACGACGTAGGCATGGGGGCTTACGATGCTA
CCCCGGAGCTACGGCGACTGCGATCTACGGCGATCGACGGACGGACGTCAGGCGACGATCTATCATCTA
TCAAGCGAGCTACTTACTCTTCTCTATCTACTTATCCCCTTCTTAGGGGTTGATTAGTCTAGCTGGTAC
GATCTTTCTAGCGAGCGATCTAGAGCGATCGACGAGCTGACGGACGTACTTACTATCGTAGCGACTACT
TC
```

*DNA, RNA, or amino acid sequence

For more details, see https://en.wikipedia.org/wiki/FASTA_format

GFF/GTF format for feature annotations

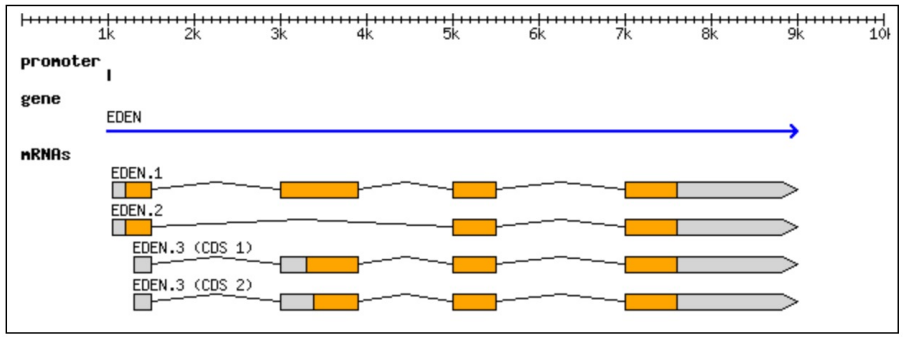
Positional/annotation information about sequence features, such as genes, is often stored as a gff or gtf formatted file.




```
#Typical gff file format for gene annotations
CHROMOSOME    SOURCE    FEATURE    START    END      SCORE    STRAND    FRAME    NOTE
V             WormBase  gene       9987196  9991642  .        -          .       Name=rde-1
```

Generic Feature Format (gff3): 9 tab-delimited columns

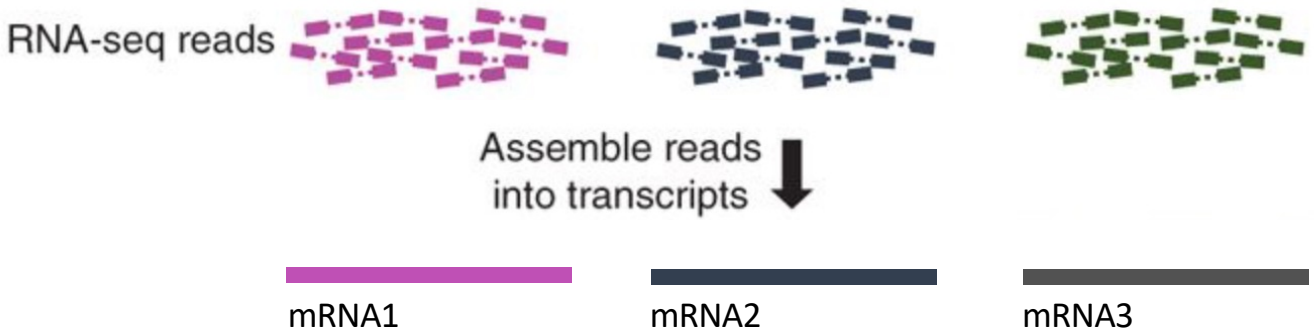
GFF/GTF format for feature annotations




```
ctg123 . exon            1300  1500  .  +  .  Parent=mRNA00003
ctg123 . exon            1050  1500  .  +  .  Parent=mRNA00001,mRNA00002
ctg123 . exon            3000  3902  .  +  .  Parent=mRNA00001,mRNA00003
ctg123 . exon            5000  5500  .  +  .  Parent=mRNA00001,mRNA00002,mRNA00003
ctg123 . exon            7000  9000  .  +  .  Parent=mRNA00001,mRNA00002,mRNA00003
ctg123 . CDS             1201  1500  .  +  0  ID=cds00001;Parent=mRNA00001
ctg123 . CDS             3000  3902  .  +  0  ID=cds00001;Parent=mRNA00001
ctg123 . CDS             5000  5500  .  +  0  ID=cds00001;Parent=mRNA00001
ctg123 . CDS             7000  7600  .  +  0  ID=cds00001;Parent=mRNA00001
ctg123 . CDS             1201  1500  .  +  0  ID=cds00002;Parent=mRNA00002
ctg123 . CDS             5000  5500  .  +  0  ID=cds00002;Parent=mRNA00002
ctg123 . CDS             7000  7600  .  +  0  ID=cds00002;Parent=mRNA00002
ctg123 . CDS             3301  3902  .  +  0  ID=cds00003;Parent=mRNA00003
ctg123 . CDS             5000  5500  .  +  1  ID=cds00003;Parent=mRNA00003
ctg123 . CDS             7000  7600  .  +  1  ID=cds00003;Parent=mRNA00003
ctg123 . CDS             3391  3902  .  +  0  ID=cds00004;Parent=mRNA00003
ctg123 . CDS             5000  5500  .  +  1  ID=cds00004;Parent=mRNA00003
ctg123 . CDS             7000  7600  .  +  1  ID=cds00004;Parent=mRNA00003
```

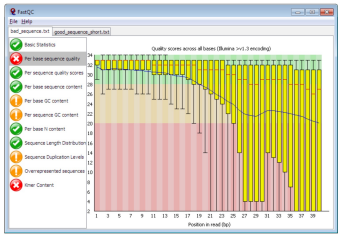
http://www.sequenceontology.org/gff3.shtml

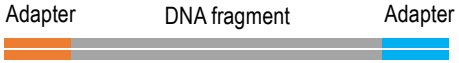
Assembling mRNAs if working with an uncharacterized species



RNA-seq data analysis workflow

```
@D64TDFP1:248:C50DMACXX:5:1101:1241:2095 1:N:0:ATCACG
CACCGCCCGTCGCTATCCGGGACTGGAATTCTCGGGTGCCAAGGAACTCCA
+
CCCFFFFFHHHHHJIJGHJJJJIJJJJJGGGFFFFEABDHHHFHFF@@DD>
@D64TDFP1:248:C50DMACXX:5:1101:1371:2154 1:N:0:ATCACG
TCAATATTTGCATAGGGTATCTGGAATTCTCGGGTGCCAAGGAACTCCAGT
+
CCCFFFFFHHHHHJJJJGFHIJJJJJJJJJJJJJFHHIIJJHGHJFGHJJI
@D64TDFP1:248:C50DMACXX:5:1101:1461:2205 1:N:0:ATCACG
GAAAGACGTCTTCCTAGATTATGGAATTCTCGGGTGCCAAGGAACTCCAGT
+
CCCFFFFFHHHHHJJJJJJJJJJJIJJJJJJJJJHIJJJJJGIIJFGIJJJ
```

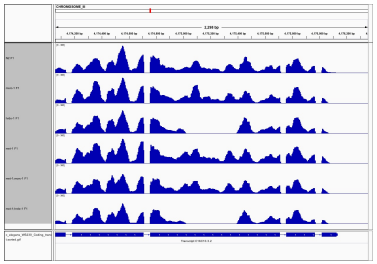


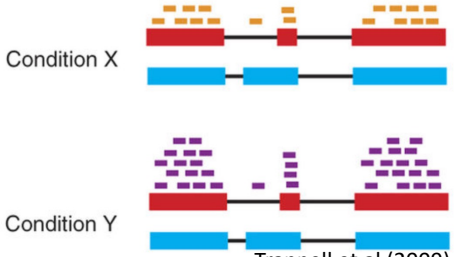


Obtain fastq files, genome or mRNA sequences, and feature annotations

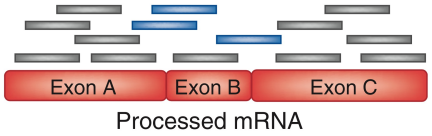
Quality control

Adapter trimming, quality filtering





Trapnell et al (2009)



Trapnell et al (2009)

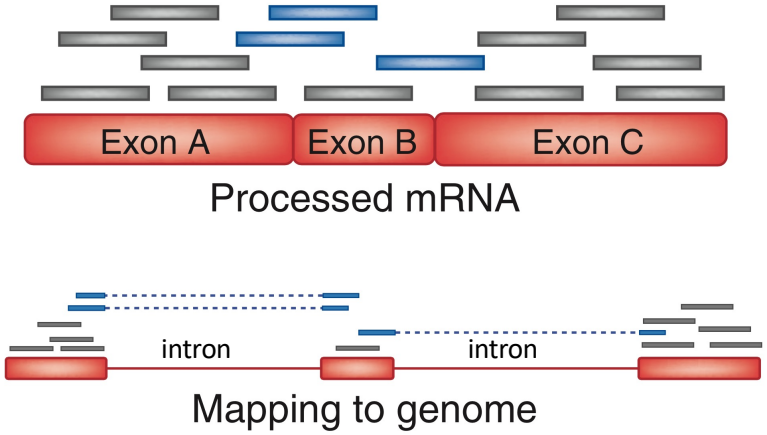
Data visualization and presentation

Differential gene expression analysis

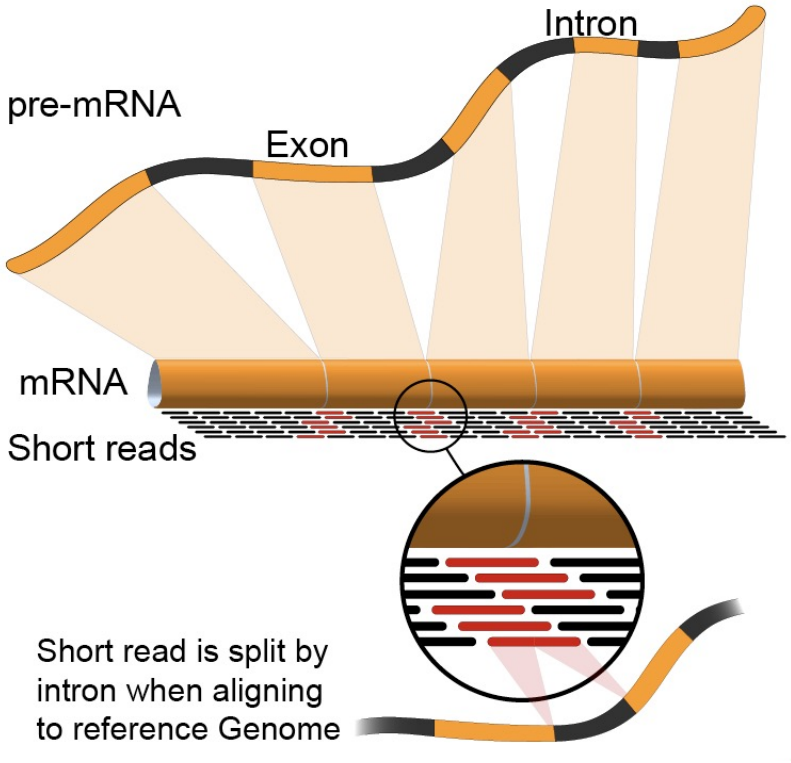
Map reads to transcriptome
Transcript quantification

Aligning reads across splice junctions

Splicing: removal of introns and joining of exons during mRNA maturation.

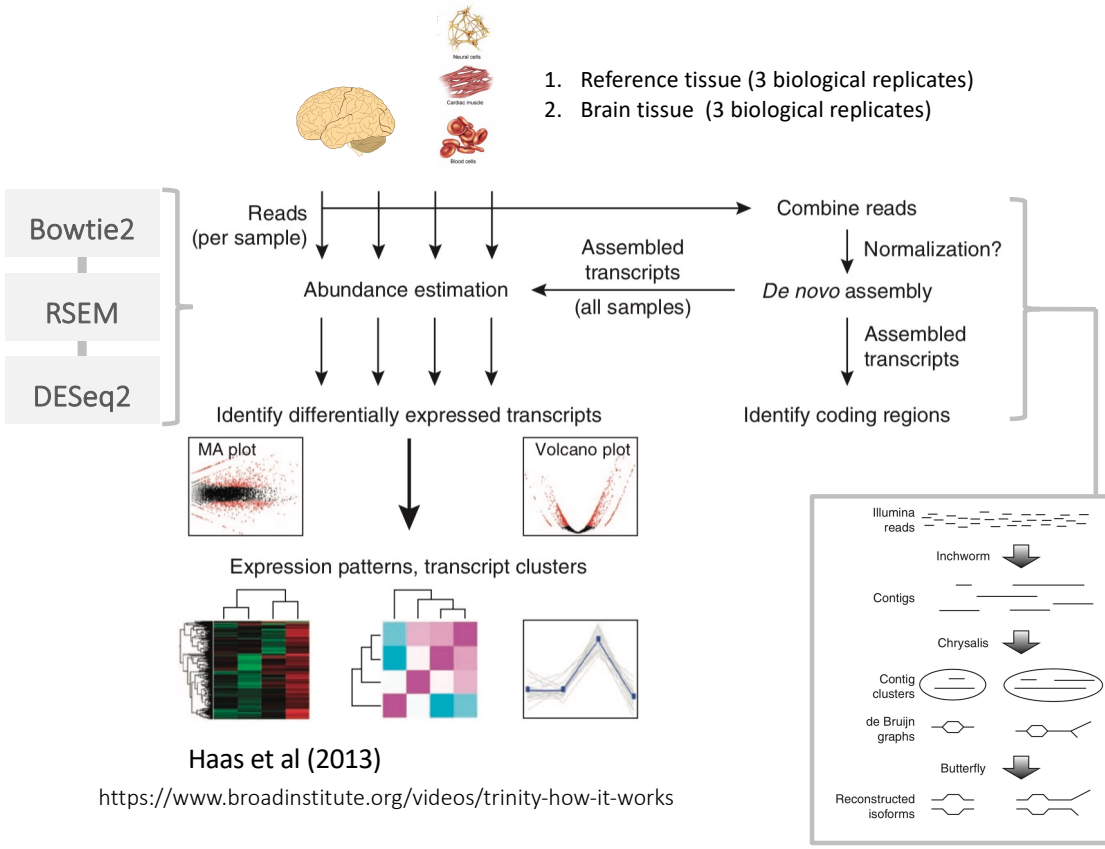


Trapnell et al (2009)



Rgocs, https://commons.wikimedia.org/w/index.php?curid=27664181

Trinity workflow

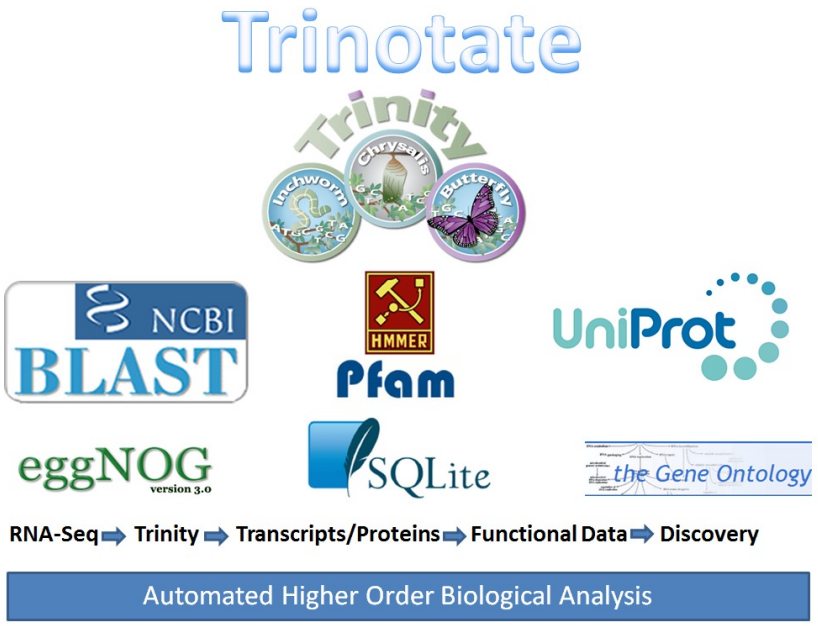


Haas et al (2013)

https://www.broadinstitute.org/videos/trinity-how-it-works

Functional annotation of newly identified transcripts

We won't cover functional analysis of our transcripts in this workshop but Trinotate can be used to automate homology searches, domain discovery, and functional analysis of your RNA-seq data.

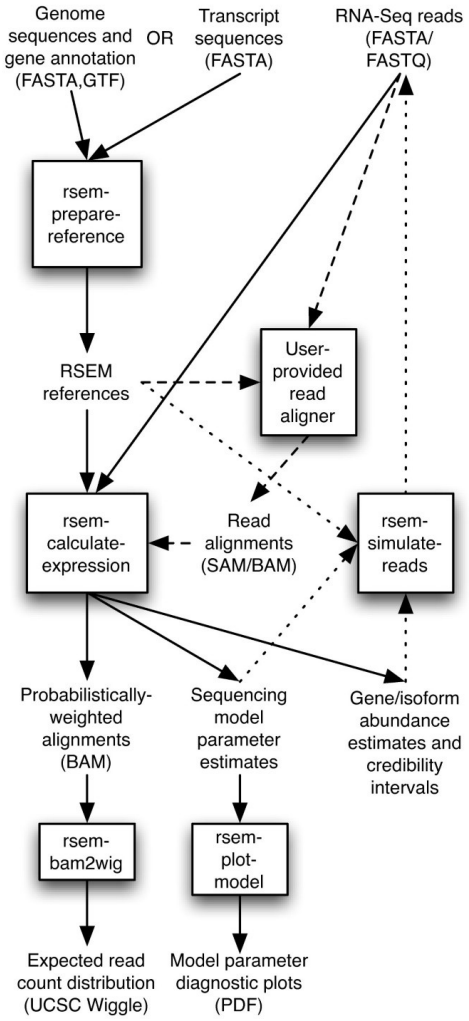


http://trinotate.github.io

Transcript quantification with RSEM

Two steps:

1. rsem-prepare-reference
   - Prepares transcript or genome for mapping reads to.
2. rsem-calculate-expression:
   - Assign single-mapping reads.
   - Assign multi-mapping reads based on probability of alignment one locus vs another.
     - If many reads map to only one of the possible loci, reads all reads are more likely to originate from that locus.
     - If all reads map to multiple loci, split reads between loci.



Output from STAR and RSEM

Counts tables.
Read mapping information (coordinates, etc in sam/bam format).

RSEM counts table

genes.results

gene_id	transcript_id(s)	length	effective_length	expected_count	TPM	FPKM
2L52.1	2L52.1	1284	1209.84	181	3.87	7.08
2RSSE.1	2RSSE.1	1032	957.84	111	3	5.48
2RSSE.2	2RSSE.2	666	591.84	25	1.09	2
3R5.1	3R5.1	648	573.84	177	7.98	14.6
4R79.1	4R79.1a,4R79.1b	947.78	873.62	16	0.47	0.87
4R79.2	4R79.2a,4R79.2b	1162.85	1088.68	26	0.62	1.13
6R55.2	6R55.2	258	183.84	1	0.14	0.26
AC3.1	AC3.1	1068	993.84	3	0.08	0.14
AC3.10	AC3.10	1056	981.84	19	0.5	0.92
AC3.12	AC3.12	285	210.84	1	0.12	0.22
AC3.2	AC3.2	1578	1503.84	473	8.14	14.88
AC3.3	AC3.3	1278	1203.84	1	0.02	0.04

Read Counting Metrics

FPKM (fragment per kilobase of transcript per million reads):

1. Divide total fragments in a sample by the total # million fragments (=RPM)
2. Divide RPM by length of the transcript in kilobases (=FPKM).

* One read/fragment for single-end data, two reads/fragment for paired-end data.

TPM (transcripts per kilobase million, preferred approach):

1. Divide fragments by length of transcript in kilobases (RPK).
2. Divide RPK by sum of RPKs for all genes in millions.

Subtract mean fragment length from transcript length + 1

genes.results

gene_id	transcript_id(s)	length	effective_length	expected_count	TPM	FPKM
2L52.1	2L52.1	1284	1209.84	181	3.87	7.08
2RSSE.1	2RSSE.1	1032	957.84	111	3	5.48
2RSSE.2	2RSSE.2	666	591.84	25	1.09	2
3R5.1	3R5.1	648	573.84	177	7.98	14.6

Read Counting Metrics

FPKM (fragment per kilobase of transcript per million reads):

1. Divide total fragments in a sample by the total # million fragments (=RPM)
2. Divide RPM by length of the transcript in kilobases (=FPKM).

* One read/fragment for single-end data, two reads/fragment for paired-end data.

TPM (transcripts per kilobase million, preferred approach):

1. Divide fragments by length of transcript in kilobases (RPK).
2. Divide RPK by sum of RPKs for all genes in millions.

Raw # fragments computed to align to a gene.

genes.results

gene_id	transcript_id(s)	length	effective_length	expected_count	TPM	FPKM
2L52.1	2L52.1	1284	1209.84	181	3.87	7.08
2RSSE.1	2RSSE.1	1032	957.84	111	3	5.48
2RSSE.2	2RSSE.2	666	591.84	25	1.09	2
3R5.1	3R5.1	648	573.84	177	7.98	14.6

Read Counting Metrics

FPKM example (total reads = 21130730.98):
FPKM for 2L52.1 = (181/21.13073098)/1.20984 => 7.08

TPM example (total RPK = 38.66207856):
TPM for 2L52.1 = (181/1.20984)/38.66207856 => 3.87

genes.results

gene_id	transcript_id(s)	length	effective_length	expected_count	TPM	FPKM
2L52.1	2L52.1	1284	1209.84	181	3.87	7.08
2RSSE.1	2RSSE.1	1032	957.84	111	3	5.48
2RSSE.2	2RSSE.2	666	591.84	25	1.09	2
3R5.1	3R5.1	648	573.84	177	7.98	14.6

SAM/BAM Format

SAMtools: a software package for mining NextGen sequencing data after alignment.

SAM (Sequence Alignment Map): A widely use format for storing alignment data for high-throughput sequencing reads.

BAM (binary SAM): Compressed SAM (binary format).

The file is broken down into two sections:

1. **Header section (optional):** contains general information about the data such as alignment software used, reference genome aligned againsts, etc. Header lines start with @.
2. **Alignment section:** contains much of the same information as a fastq file, such as sequence and base quality scores, as well as information about alignment to reference sequence.

For a more complete description, see https://genome.sph.umich.edu/wiki/SAM

SAM/BAM Format

11 fields + optional 12th TAGs field (not shown)

Col	Field	Type	Brief Description
1	QNAME	String	Query template NAME
2	FLAG	Int	bitwise FLAG
3	RNAME	String	References sequence NAME
4	POS	Int	1- based leftmost mapping POSition
5	MAPQ	Int	MAPping Quality
6	CIGAR	String	CIGAR String
7	RNEXT	String	Ref. name of the mate/next read
8	PNEXT	Int	Position of the mate/next read
9	TLEN	Int	observed Template LENgth
10	SEQ	String	segment SEQuence
11	QUAL	String	ASCII of Phred-scaled base QUALity+33