Transcriptomics Part 1: RNA-seq Common Files and Pipelines

Bridging the Bench-Machine Learning Gap

Dr. Emily A. Beck

Dr. Jake Searcy

Learning Objectives

Understand basic structures of RNA-seq file types not covered in DNA-seq

Learn basic pipelines for getting from raw sequencing files to common desired outcomes (annotations and differential expression analyses)

Learning Objectives

Understand basic structures of RNA-seq file types not covered in DNA-seq

Learn basic pipelines for getting from raw sequencing files to common desired outcomes (annotations and differential expression analyses)

Main reasons people use RNA-Seq:

1. Improve genomic knowledge for a non-model organism

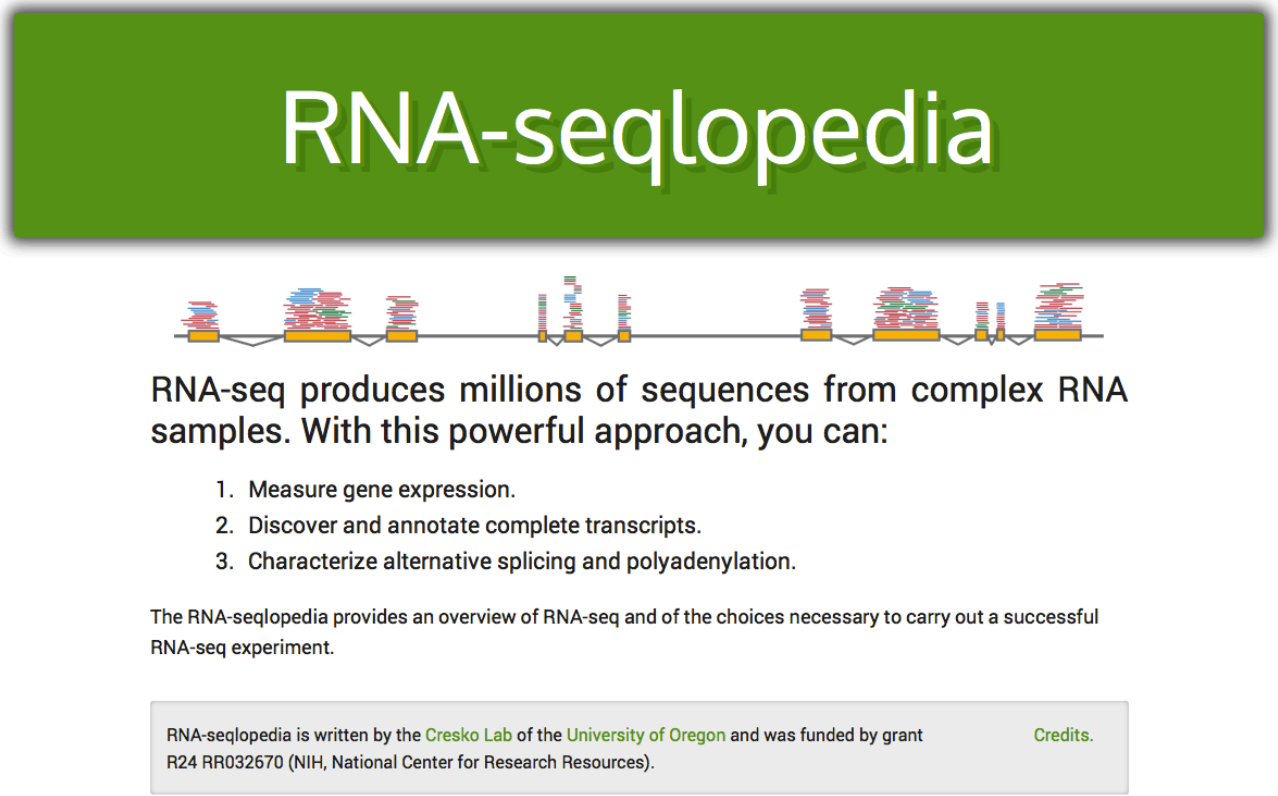
2. Measure and compare transcript abundance among groups
(e.g. exp. treatments, genetic backgrounds, species, cells, etc.)

3. Characterize a transcriptome at some finer scale
(e.g. discover novel splice variants or polyadenylation variants,
document pre-processing of RNAs, etc.)

“Annotation”

“Differential Expression”

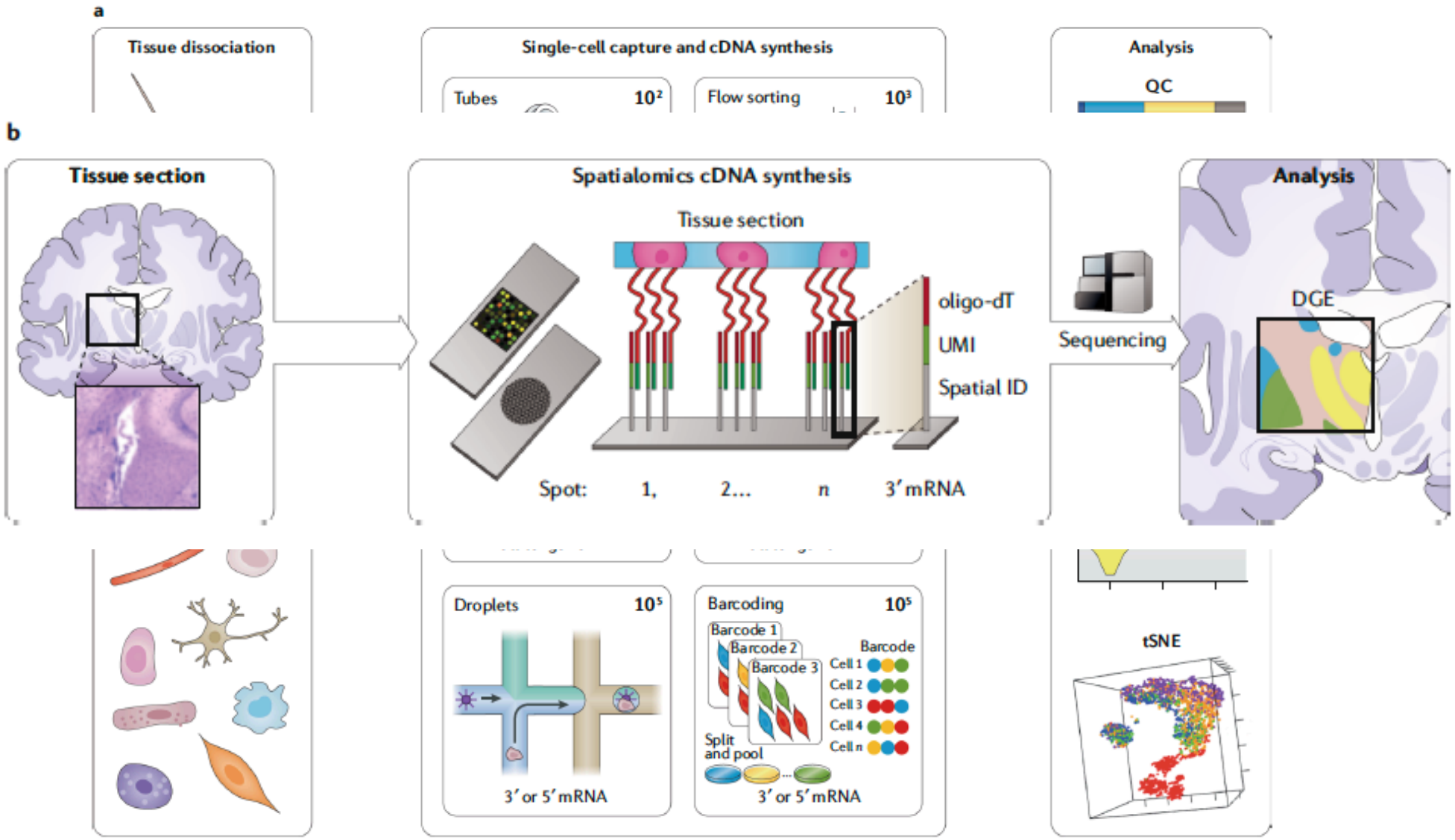
RNA-seqlopedia is an excellent resource for understanding the basics of RNA-seq



http://rnaseq.uoregon.edu/

Read through as a basic resource

Newer applications of RNA-Seq (single-cell, “slice-seq”, “spatialomics)



Stark et al., Nat. Rev. Genet. 2019

Generic RNA-Seq workflow

Just like with DNA-seq how you prep your starting material matters!

-High-quality total RNA

degraded RNA will result in 3’ biases

-Try to maximize RNA diversity/evenness!

- broad range of tissue types (if using adult organism)
- early stages in development

-Minimize polymorphism

-use inbred population or an individual

-avoid pooling many individuals

Researchers often focus on mRNAs encoding proteins

DNA transcribed into **messenger RNA (mRNA)**; mRNA is translated into protein on the surface of a ribosome

tRNAs and rRNAs

- **tRNA** matches amino acids to codons during protein production
- **rRNA** can be separated in a centrifuge and is named after the *band* that it separates at:
  18S, 28S: Forms structural and functional components of ribosomes

Small nuclear RNA (snRNA):

- Localized to the nucleus; used in RNA splicing
- Five snRNAs (along with many proteins) make up the *spliceosome*, which removes introns from pre-mRNA

MicroRNA (miRNA):

- ~22 nucleotide RNAs
- Down-regulates protein abundance by inhibiting translation from mRNA
- Found in intergenic DNA or oriented antisense to neighboring genes; often found in introns or exons and transcribed along with host gene

snoRNA, siRNA, lncRNA, etc…

Depending on the focus of your project you may need to alter your methods for RNA-seq library prep! Not all transcripts are polyadenylated, different preps are able to capture different classes of RNA

Target Enrichment upstream in the pipeline allows you to focus on different types of RNAs

polyA+ : messenger RNA enrichment

*most common, emphasis on coding RNAs

rRNA- : ribosomal RNA removal

*leaves mature and immature mRNAs, need quality total RNA, utility in divergent non-model species a concern

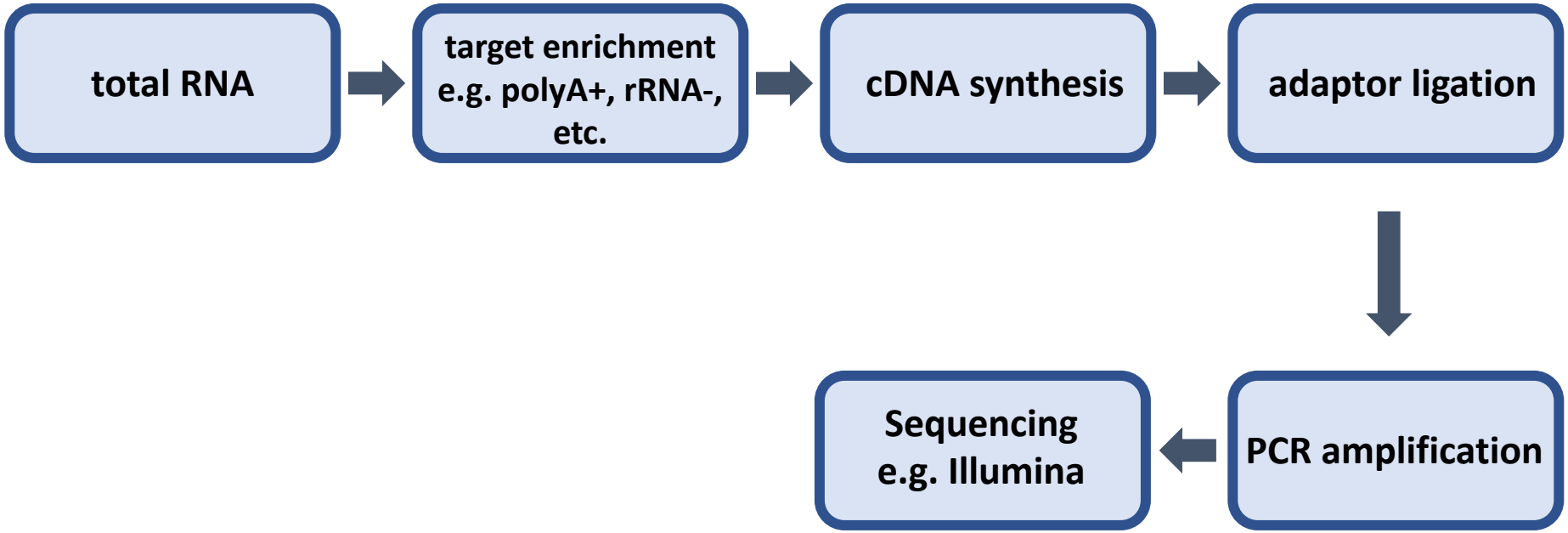
short ncRNA+ : enrichment for snoRNAs, microRNAs, siRNAs, snRNAs, exRNAs, piRNAs, etc.

*size selection-dependent, special adapters

DSN (Duplex-specific normalization)

*reduces abundant RNAs, used for annotation studies

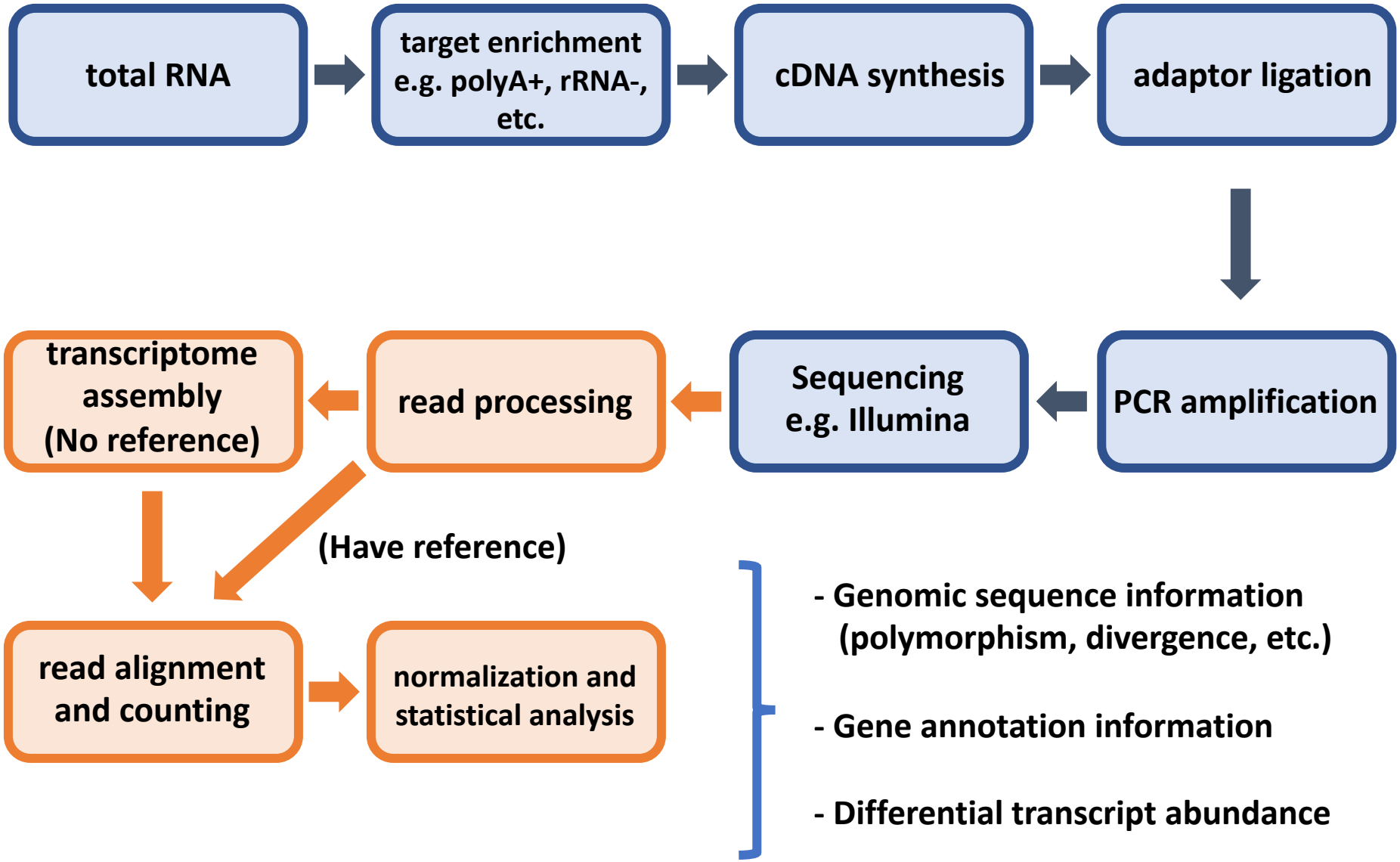
Generic RNA-Seq workflow



We are going to talk through the specifics of differential expression analyses later

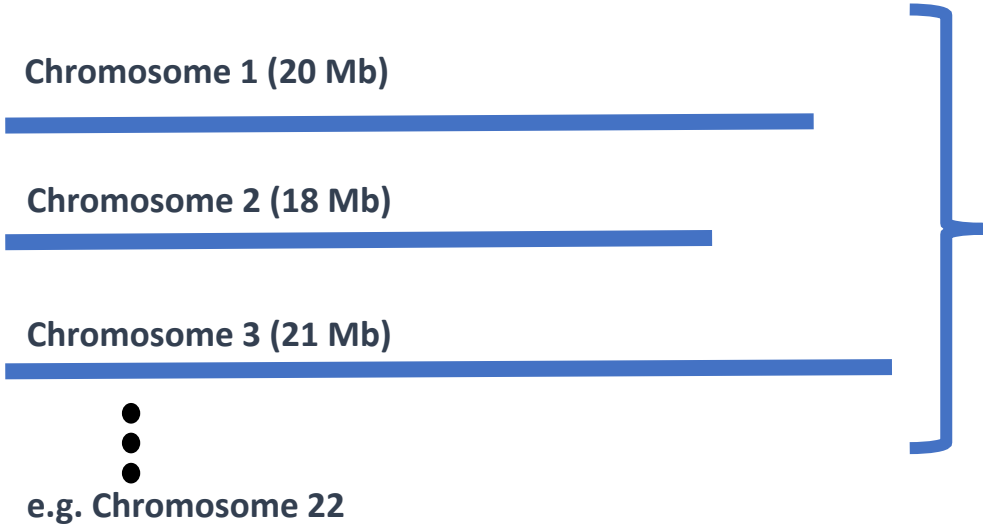
For now we can focus on getting from clean reads to a count file needed to assess differential expression

Generic RNA-Seq workflow



There are different goals and challenges in transcriptomics compared to genomics

Genome Assembly

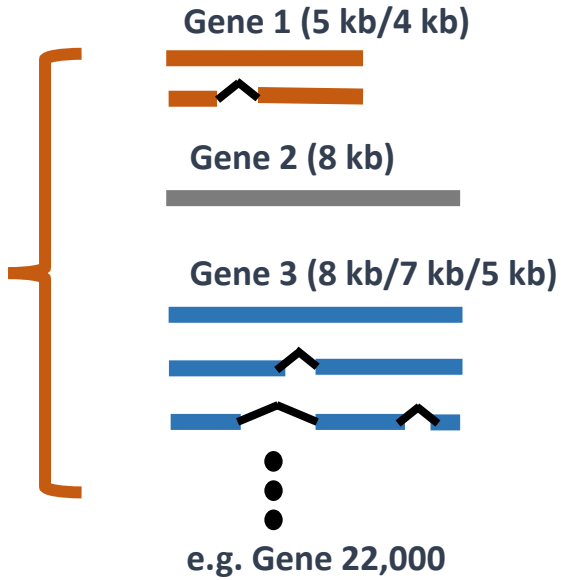


-Goal is to assemble long, continuous chromosomes

-Relatively even coverage

-Minimize # of scaffolds to recapitulate a few to dozens of chromosomes

Transcriptome Assembly



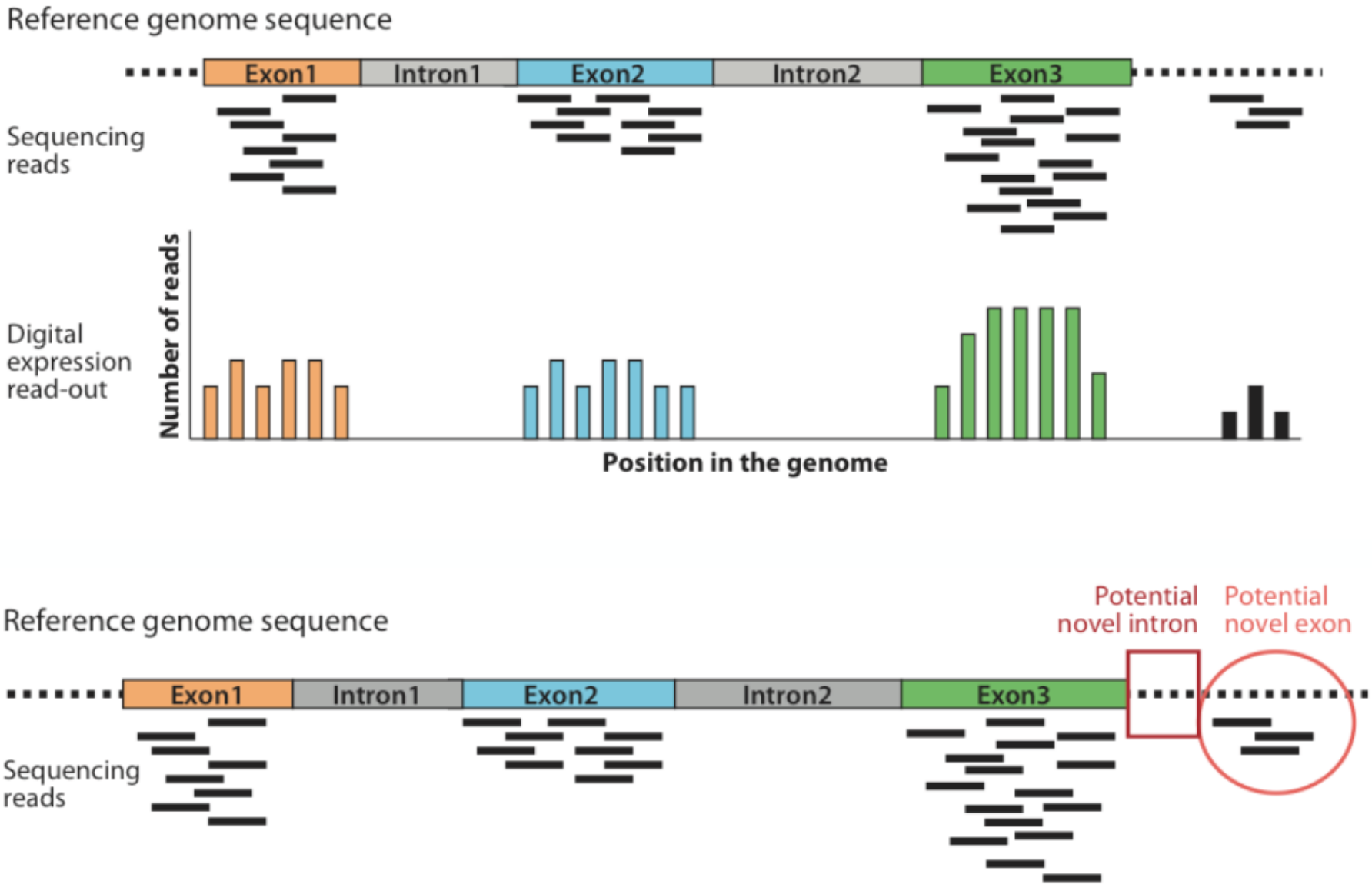
-Trying to assemble many separate, short transcripts

-Drastically uneven coverage!

-Multiple splice variants per gene

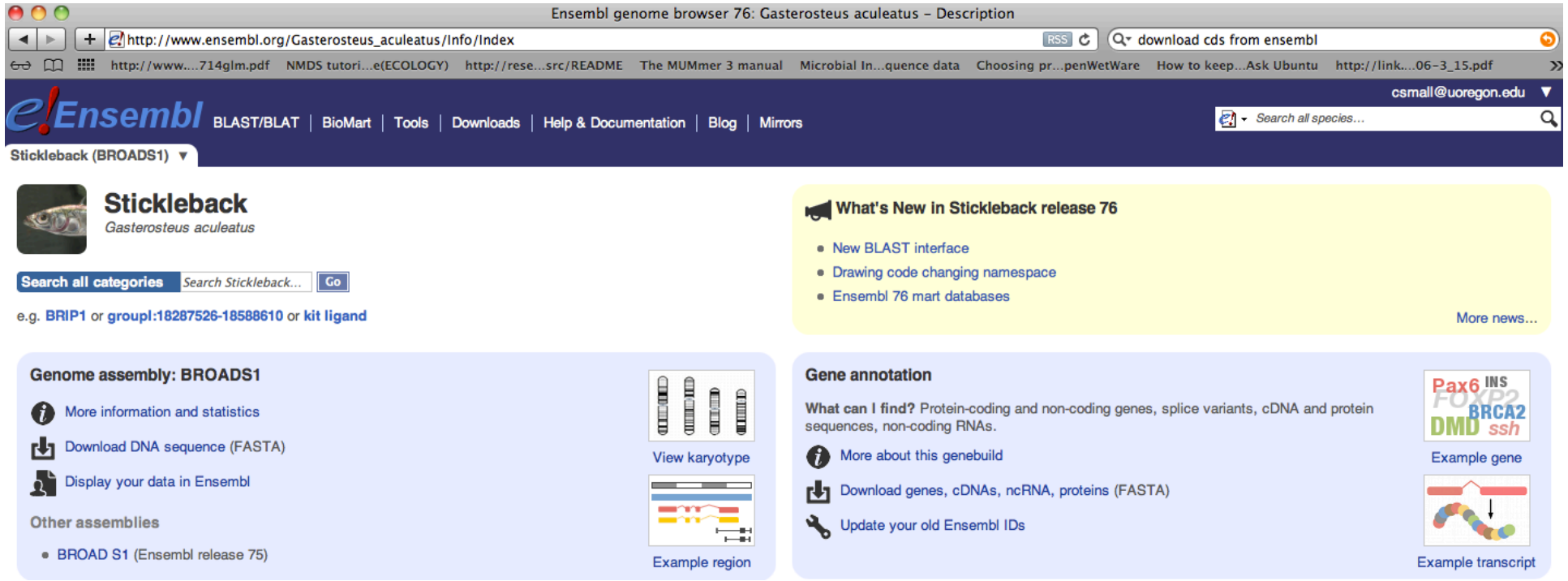
-Goal is to reconstruct “true” gene models

Gene Expression Quantification



Morozova, et al. 2009

Need a reference genome

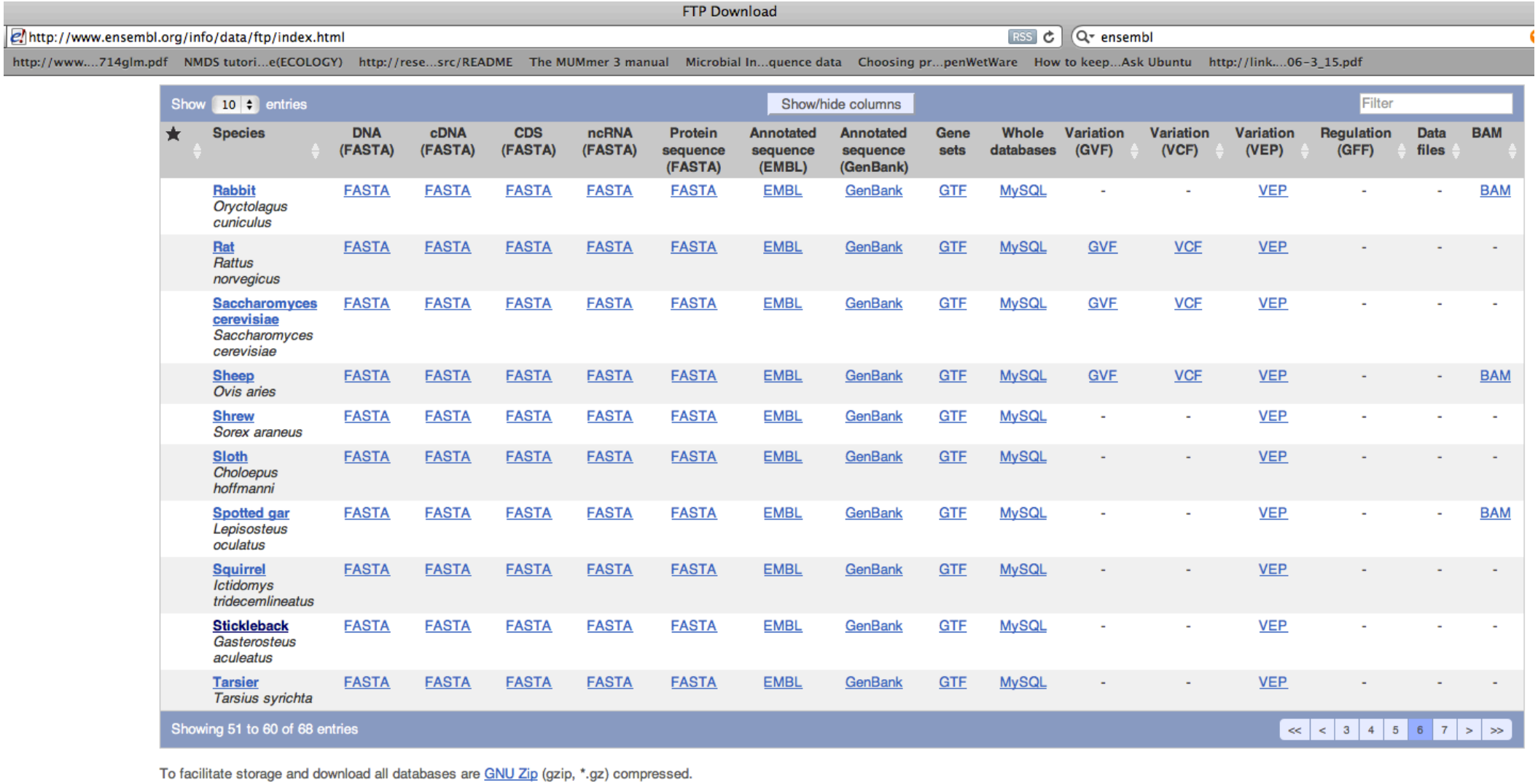


- Raw .fasta file

- Each sequence corresponds to a chromosome or scaffold

Need an annotation file (GTF)

Make sure the annotation matches the assembly version!



Species	DNA (FASTA)	cDNA (FASTA)	CDS (FASTA)	ncRNA (FASTA)	Protein sequence (FASTA)	Annotated sequence (EMBL)	Annotated sequence (GenBank)	Gene sets	Whole databases	Variation (GVF)	Variation (VCF)	Variation (VEP)	Regulation (GFF)	Data files	BAM
Rabbit *Oryctolagus cuniculus*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF	MySQL	-	-	VEP	-	-	BAM
Rat *Rattus norvegicus*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF	MySQL	GVF	VCF	VEP	-	-	-
Saccharomyces cerevisiae *Saccharomyces cerevisiae*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF	MySQL	GVF	VCF	VEP	-	-	-
Sheep *Ovis aries*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF	MySQL	GVF	VCF	VEP	-	-	BAM
Shrew *Sorex araneus*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF	MySQL	-	-	VEP	-	-	-
Sloth *Choloepus hoffmanni*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF	MySQL	-	-	VEP	-	-	-
Spotted gar *Lepisosteus oculatus*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF	MySQL	-	-	VEP	-	-	BAM
Squirrel *Ictidomys tridecemlineatus*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF	MySQL	-	-	VEP	-	-	-
Stickleback *Gasterosteus aculeatus*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF	MySQL	-	-	VEP	-	-	-
Tarsier *Tarsius syrichta*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF	MySQL	-	-	VEP	-	-	-

- This file contains the genomic coordinates of attributes

The .gtf file lists each annotated “feature” by genomic coordinate (contains 9 fields)

1. Chromosome or scaffold

2. Gene type (e.g., coding, rRNA, etc.)

3. Gene feature type

4.&5. Start & End Coordinates

6. Score (for predicted features)

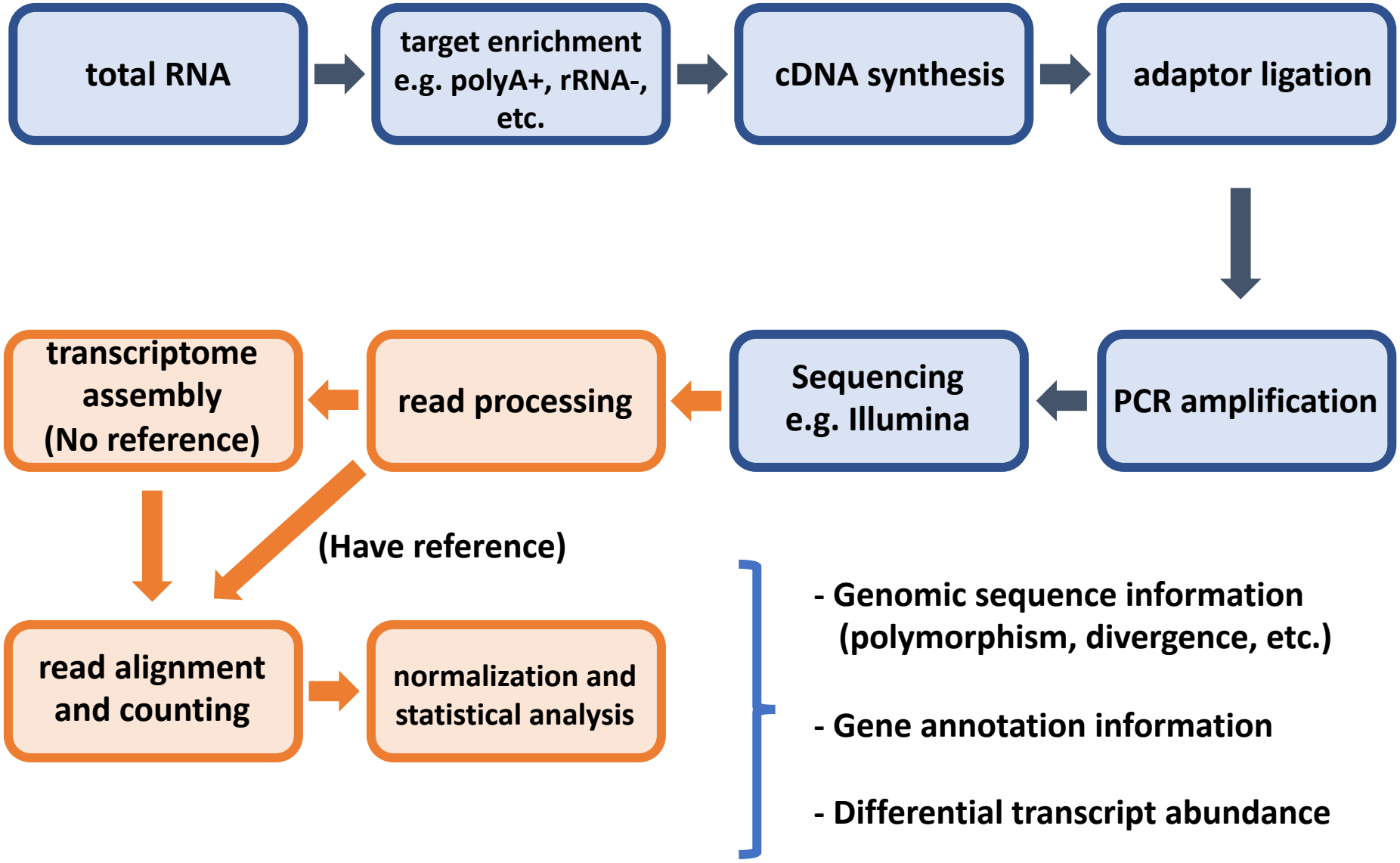
7. Strand (- opposite; + same)

8. Reading Frame

9. Attribute String

groupIV	protein_coding	gene	3396	9380	.	-	.	gene_id "ENSGACG00000016217"; gene_name "im:7142942";...
groupIV	protein_coding	transcript	3396	9380	.	-	.	gene_id "ENSGACG00000016217"; transcript_id"ENSGACT00000...
groupIV	protein_coding	exon	9231	9380	.	-	.	gene_id "ENSGACG00000016217"; transcript_id "ENSGACT0000...
groupIV	protein_coding	exon	7580	7601	.	-	.	gene_id "ENSGACG00000016217"; transcript_id "ENSGACT0000...
groupIV	protein_coding	CDS	7580	7591	.	-	0	gene_id "ENSGACG00000016217"; transcript_id "ENSGACT0000...
groupIV	protein_coding	start_codon	7589	7591	.	-	0	gene_id "ENSGACG00000016217"; transcript_id "ENSGACT0000...

You have followed our pipeline to create a Sam file which has the mapping coordinates of your reads



We can use our mapping file (Sam) to count the number of reads aligned to each feature in the gtf file

Parse .sam file for library 1, count reads mapping only between 3396 & 9380 of groupIV

Parse .sam file for library 2, …

Parse .sam file for library 3, …

Iterate over all features of interest and format all read counts in one tabular file

groupIV	protein_coding	gene	3396	9380	.	-	.	gene_id "ENSGACG00000016217"; gene_name "im:7142942";...
groupIV	protein_coding	transcript	3396	9380	.	-	.	gene_id "ENSGACG00000016217"; transcript_id"ENSGACT00000...
groupIV	protein_coding	exon	9231	9380	.	-	.	gene_id "ENSGACG00000016217"; transcript_id "ENSGACT0000...
groupIV	protein_coding	exon	7580	7601	.	-	.	gene_id "ENSGACG00000016217"; transcript_id "ENSGACT0000...
groupIV	protein_coding	CDS	7580	7591	.	-	0	gene_id "ENSGACG00000016217"; transcript_id "ENSGACT0000...
groupIV	protein_coding	start_codon	7589	7591	.	-	0	gene_id "ENSGACG00000016217"; transcript_id "ENSGACT0000...

Fortunately there are toold to help you

htseq-count (python script from HTSeq package)

(Anders et al. 2014, *bioRxiv)*

End with a count file (tsv)

Gene	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	Sample7	Sample8	Sample 9	SampleX
ENSGAC00002	3	6	10	3	3	400	243	654	345	244
ENSGAC00003	0	0	0	0	0	2	8	5	9	7
ENSGAC00004	1000	3456	6444	4564	5677	4525	5422	6553	3456	2333
ENSGAC00005	0	0	0	0	0	0	0	0	0	0
ENSGAC00006	0	0	0	0	0	0	0	0	0	0
ENSGAC00007	12	13	54	32	67	32	56	76	54	54
ENSGAC00008	54	56	34	54	65	54	4	34	55	34
ENSGAC00009	3435	5466	4356	6544	5467	7765	5444	5567	5444	5644
ENSGAC00010	0	20	30	3	35	9	7	65	7	7
ENSGAC00011	67	89	78	99	87	0	0	0	0	0

etc.

After the break we will work through an example using count files and metadata to look for patterns of differential expression