A beginner's guide to bulk RNA-Seq analysis

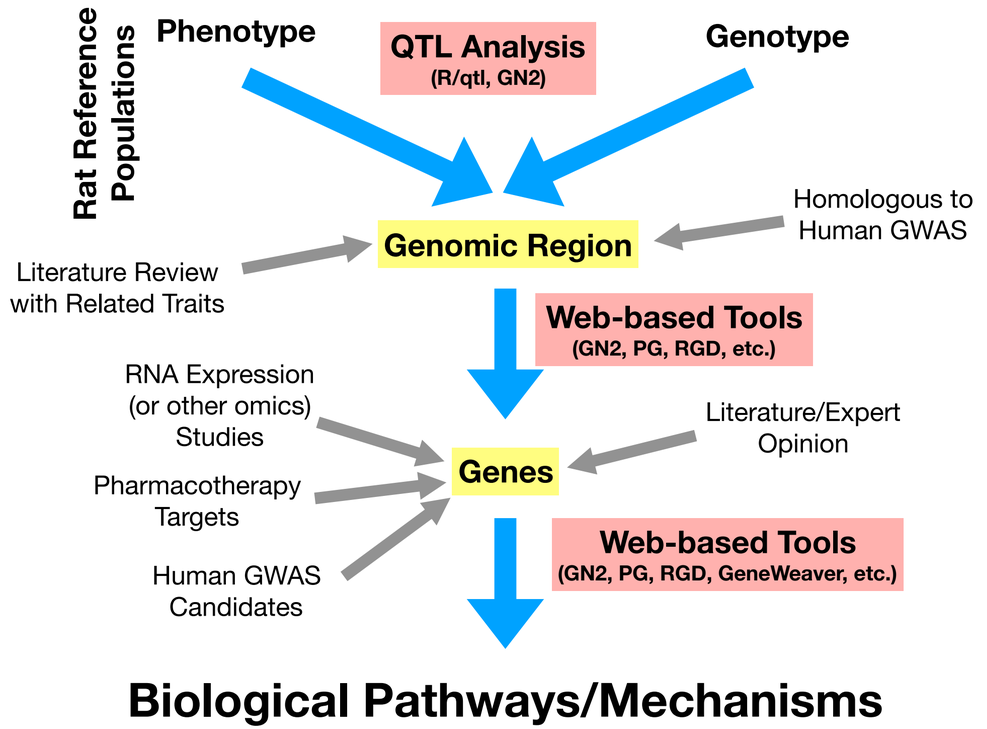
11th Webinar for Quantitative Genetics Tools for Mapping Trait Variation to Mechanisms, Therapeutics, and Interventions

Laura Saba, PhD
Associate Professor
University of Colorado Anschutz Medical Campus
NIDA Center of Excellence in Omics, Systems Genetics and the Addictome

Quantitative Genetics Tools for Mapping Trait Variation to Mechanisms, Therapeutics, and Interventions Webinar Series

Goal of the Series:
Transverse the path from trait variance to QTL to gene variant to molecular networks to mechanisms to therapeutic and interventions

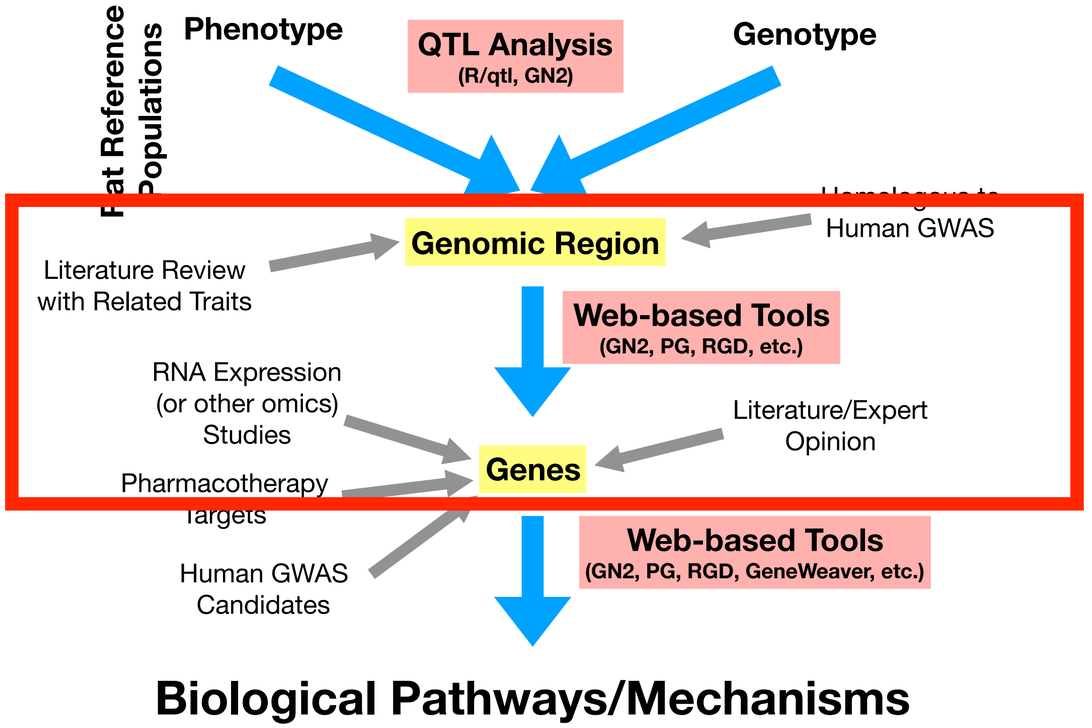
Forward Genetics in Model Organisms



Outline

- Why RNA-Seq?
- Technical Overview
- Experimental Design
- Transcriptome Profiling
- Differential Expression

Forward Genetics in Model Organisms



Why RNA-Seq?

Why study the RNA dimension?

Transcriptome links DNA and complex traits/diseases

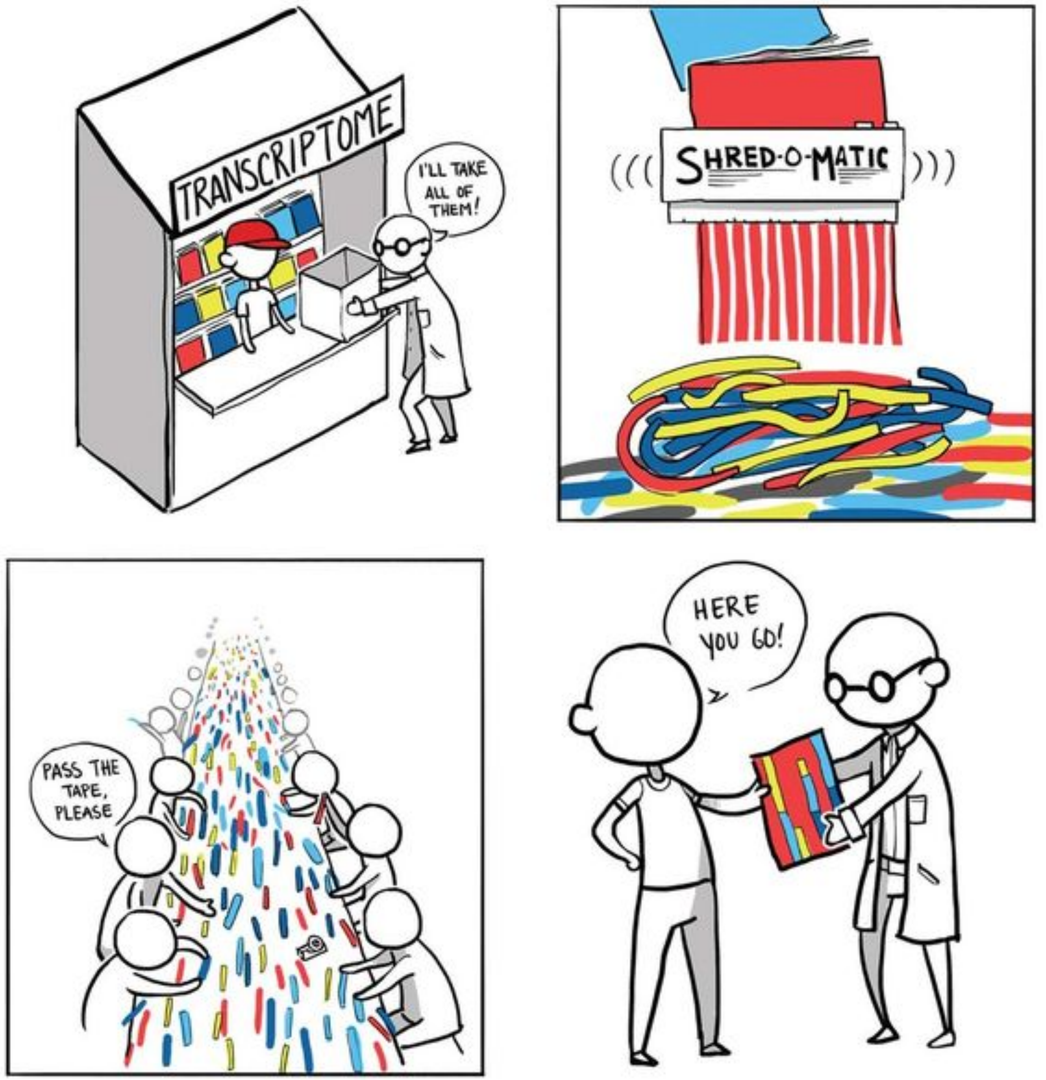
1. RNA represents one the the first **quantitative** links between DNA sequence and phenotype

2. Transcription of RNA is the **first step** where DNA sequence and environment interact

3. Implementation of graph theory at the transcript level provides insight into genetic/environmental **interactions** that are the basis for susceptibility to complex diseases

Why RNA-Seq?

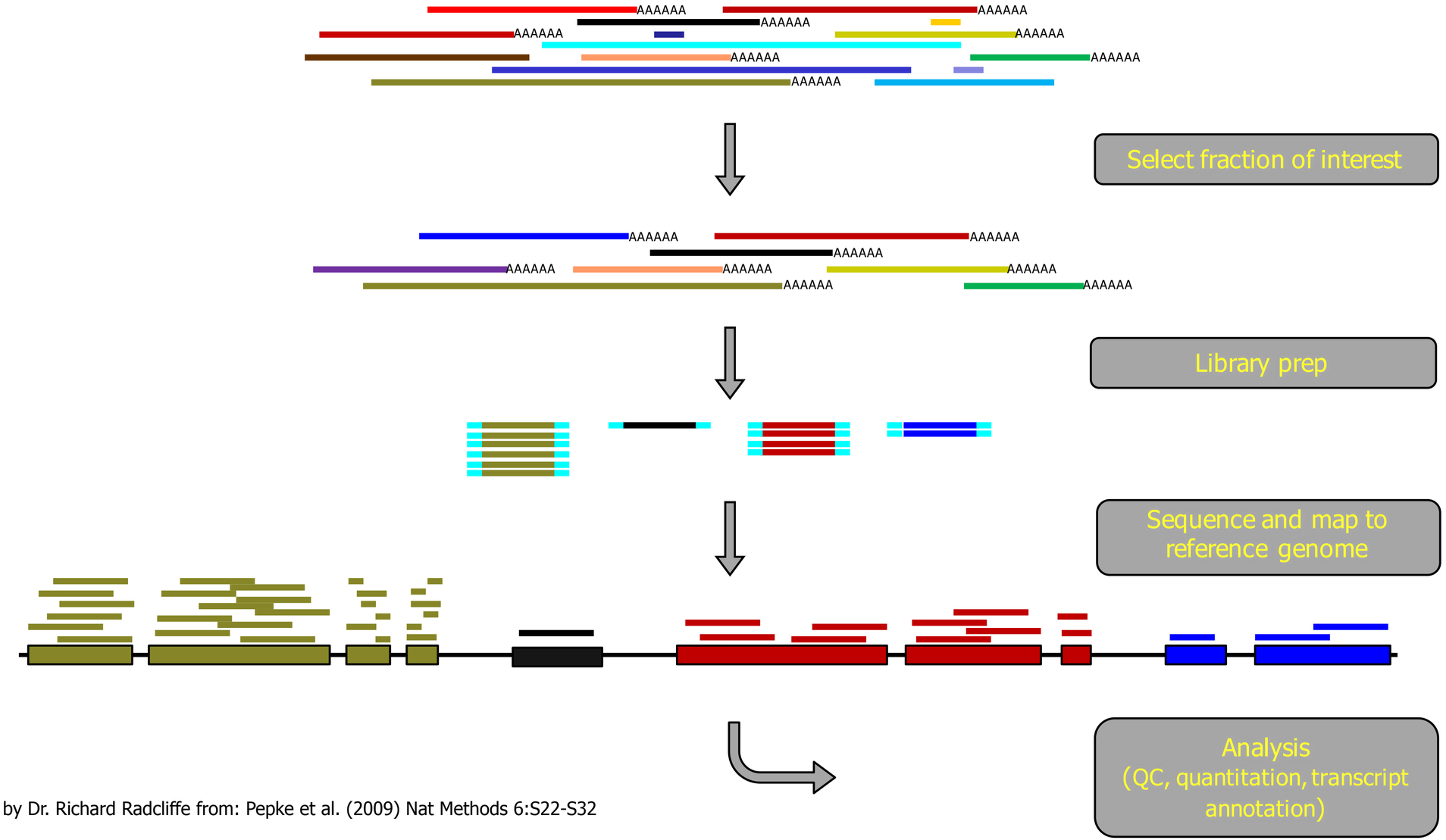
- Allows for discovery of:
  - Novel protein-coding genes
  - Novel splice variants of annotated protein-coding genes
  - Novel non-coding transcripts
  - Novel types of non-coding transcripts
- Allows for the quantitation of:
  - Protein-coding and non-coding genes
  - Alternative splicing
  - Alternative UTR usage
- Allows for the identification:
  - Single nucleotide variants
  - RNA editing



Korf, I. Genomics: the state of the art in RNA-seq analysis. *Nat Methods* **10,** 1165–1166 (2013).

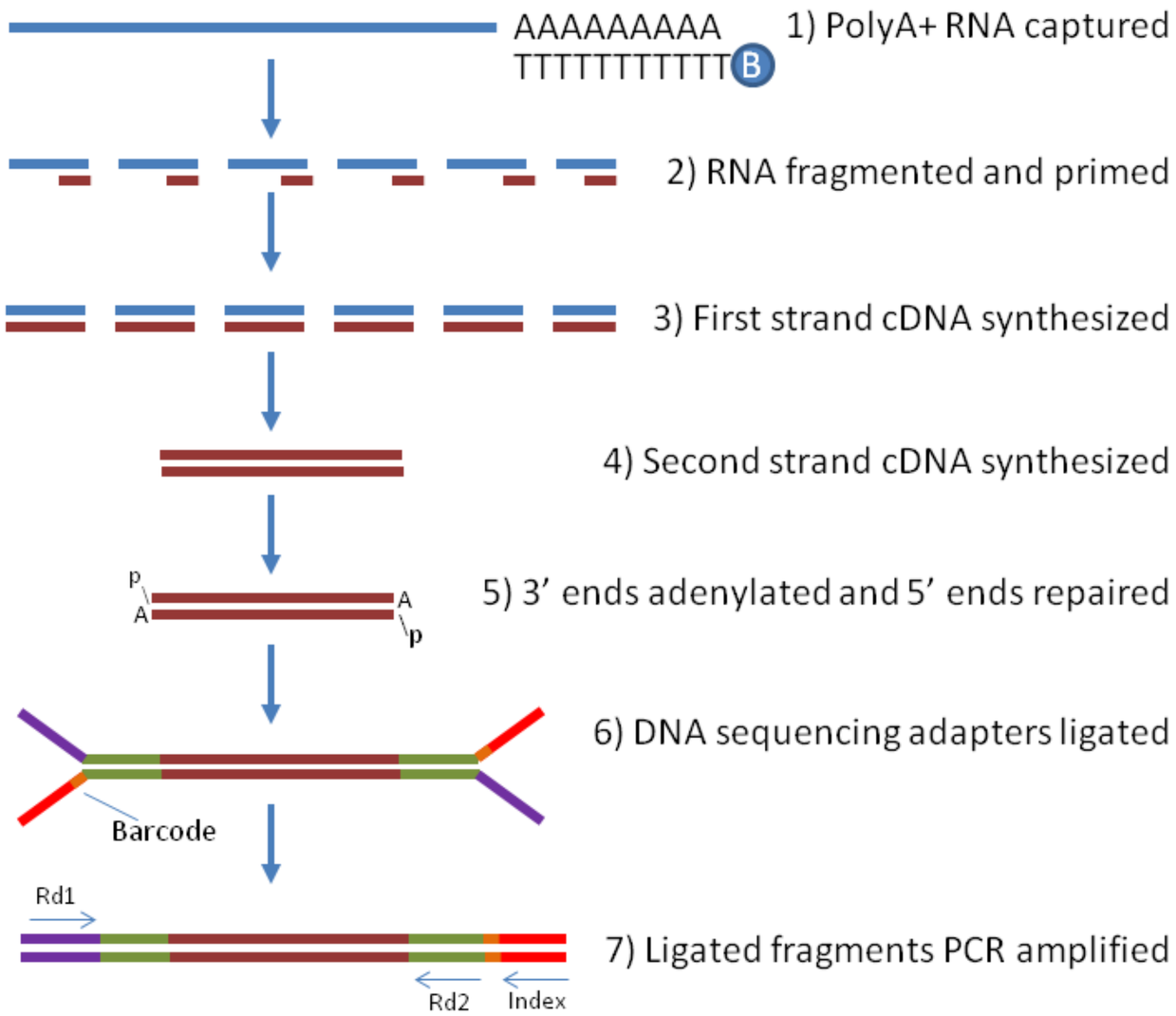
Technical Overview

RNA-Seq Overview



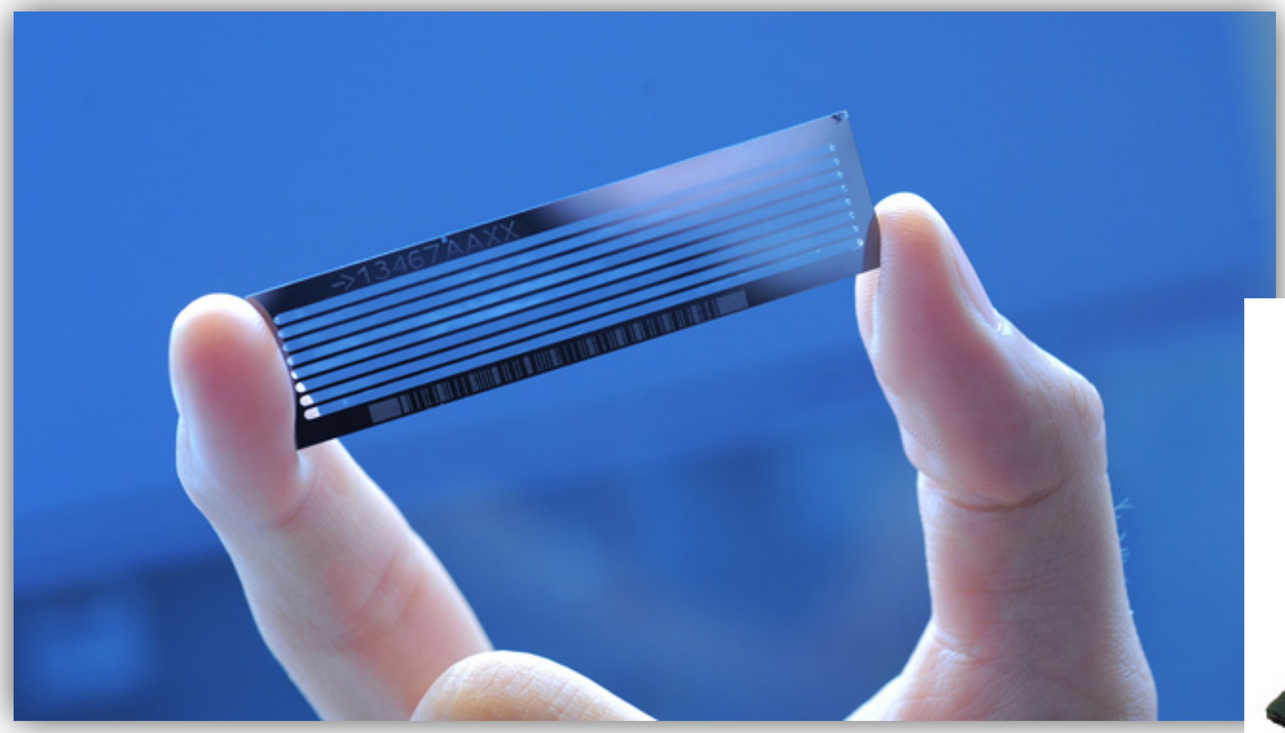
Adapted by Dr. Richard Radcliffe from: Pepke et al. (2009) Nat Methods 6:S22-S32

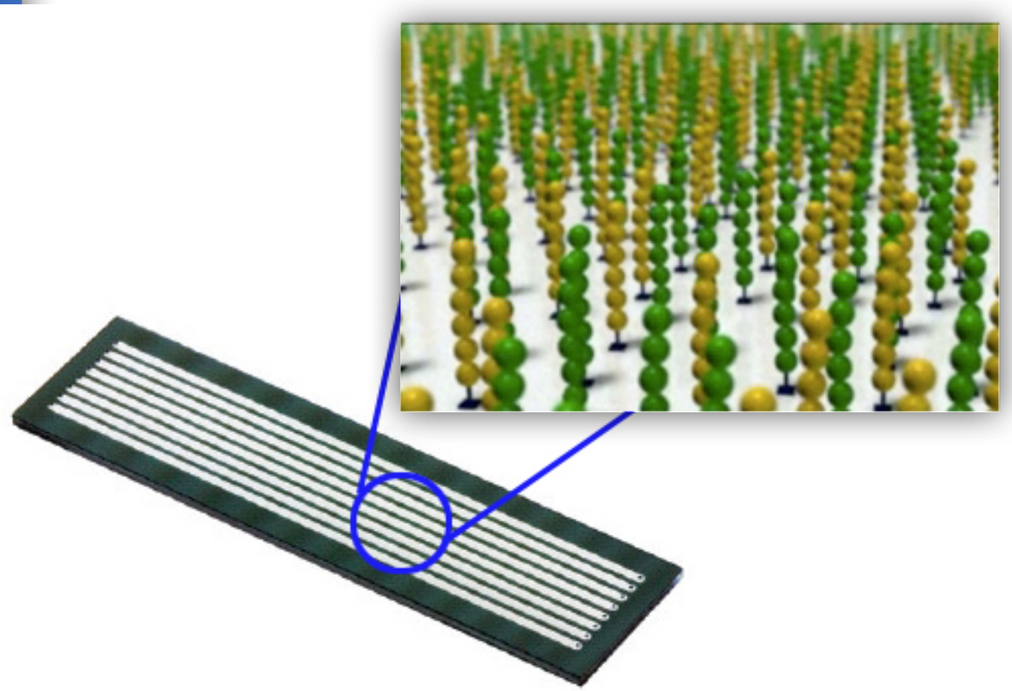
Library Preparation



Corney (2013) Mater Methods 3:203

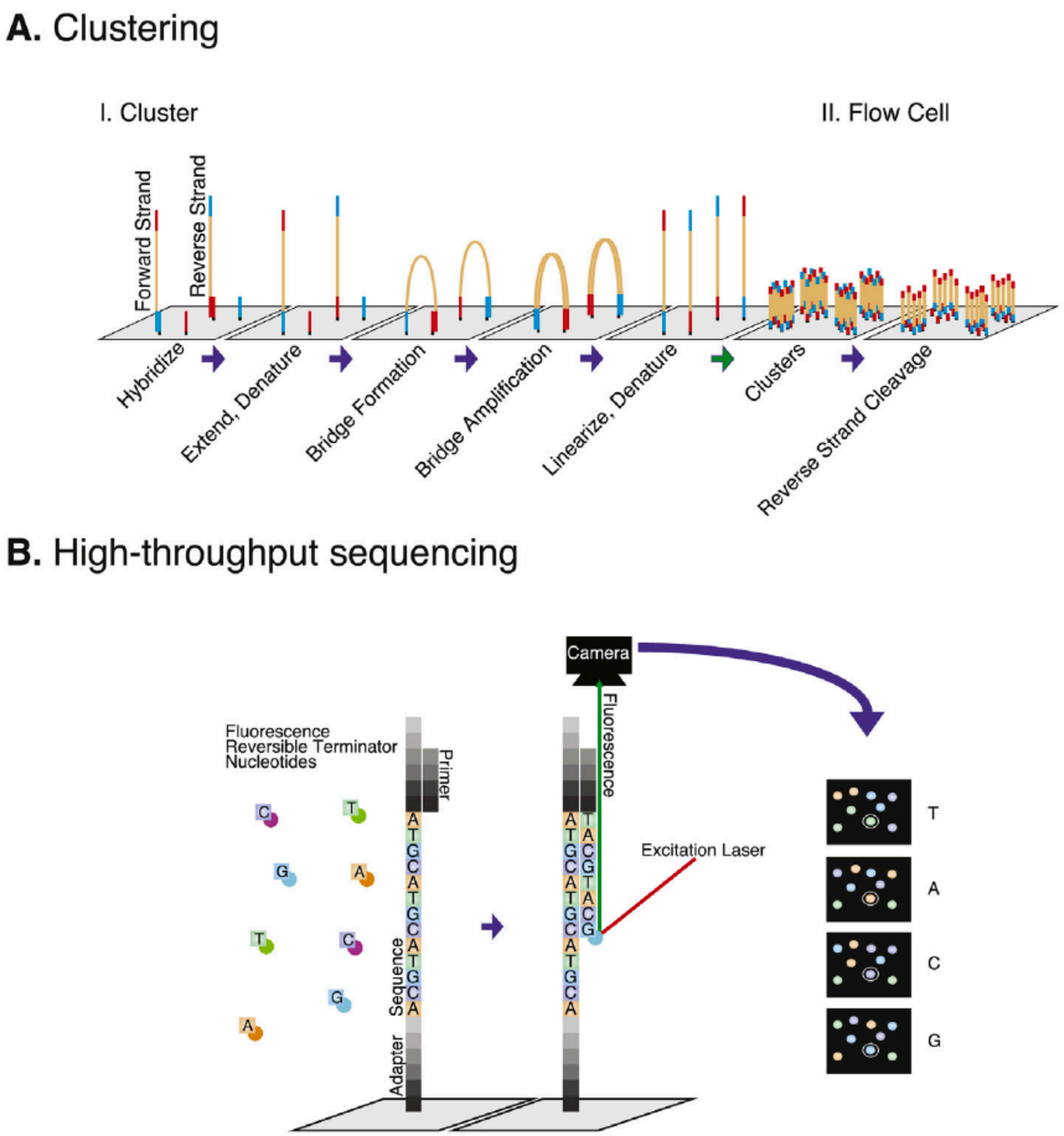
Illumina System for Sequencing





Sequencing by Synthesis

YouTube Video - https://www.youtube.com/watch?app=desktop&v=fCd6B5HRaZ8



Chaitankar V, Karakülah G, Ratnapriya R, Giuste FO, Brooks MJ, Swaroop A. Next generation sequencing technology and genomewide data analysis: Perspectives for retinal research. Prog Retin Eye Res. 2016 Nov;55:1-31.

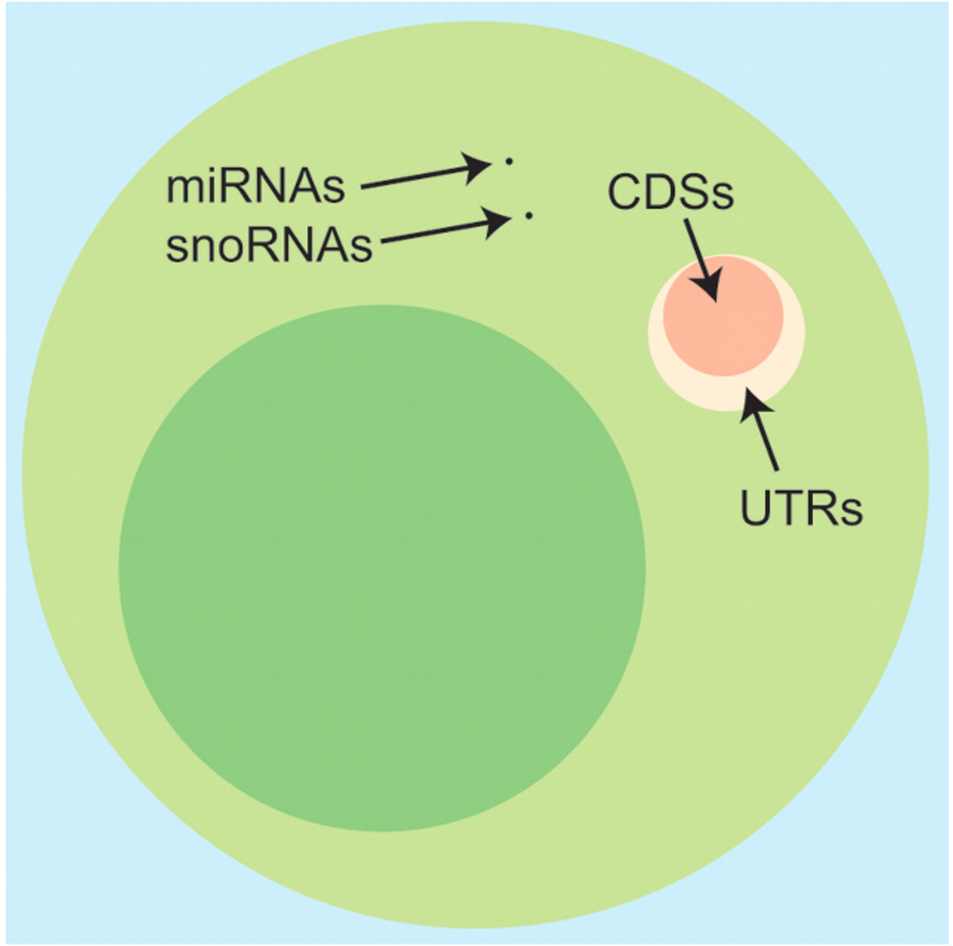
Experimental Design

Main components of experimental design

1. RNA fraction
2. Library type
3. Sequencing length
4. Sequencing depth
5. Number of replicates

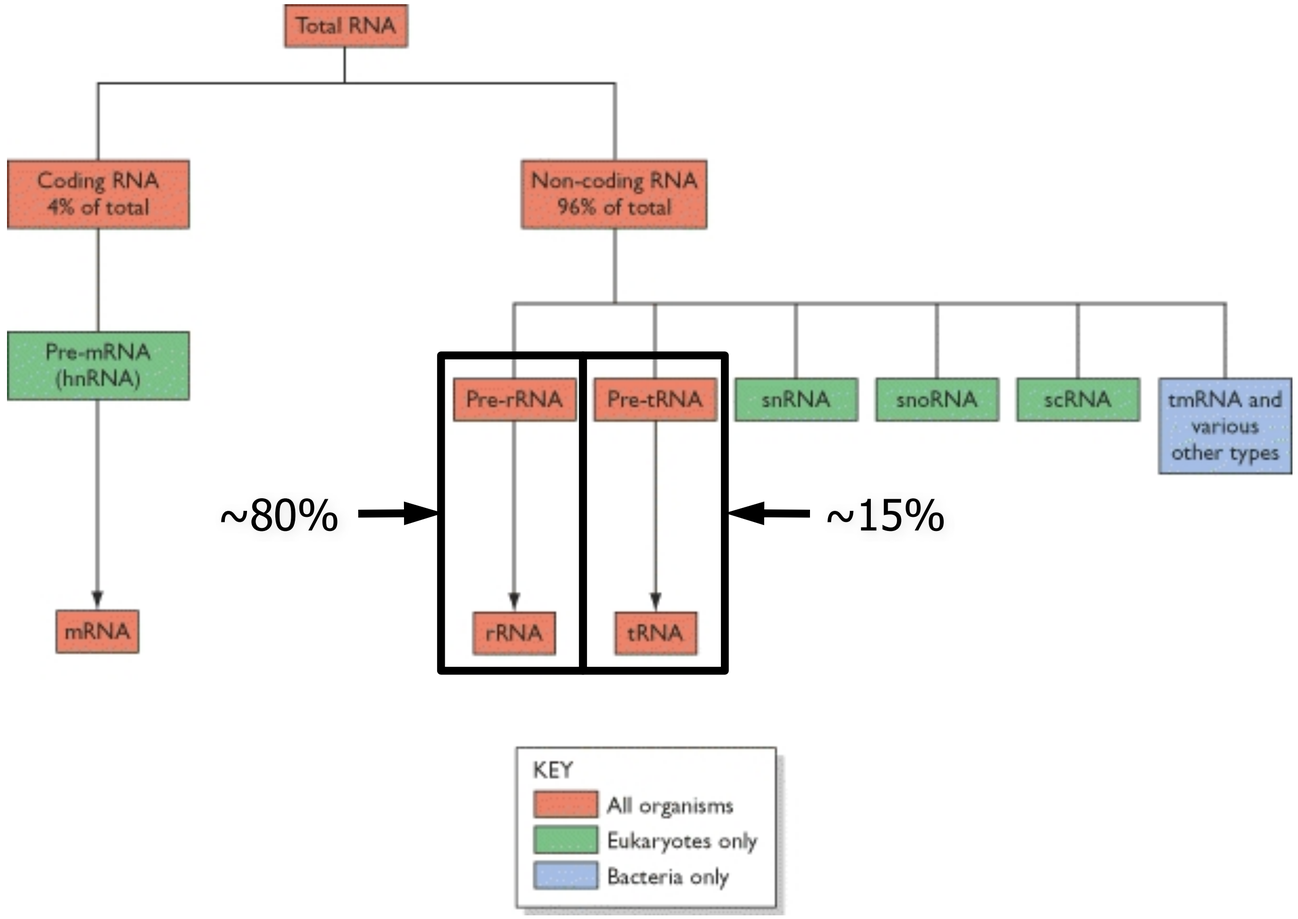
RNA Fraction

Proportion of genome that is transcribed



Mattick & Makunin (2006) Hum Mol Genet 1:R17-29

Total RNA Distribution



Genomes (2002), 2nd Edition, Chapter 3

RNA Fraction

- **PolyA selection** - Oligo-dT beads capture polyA tails

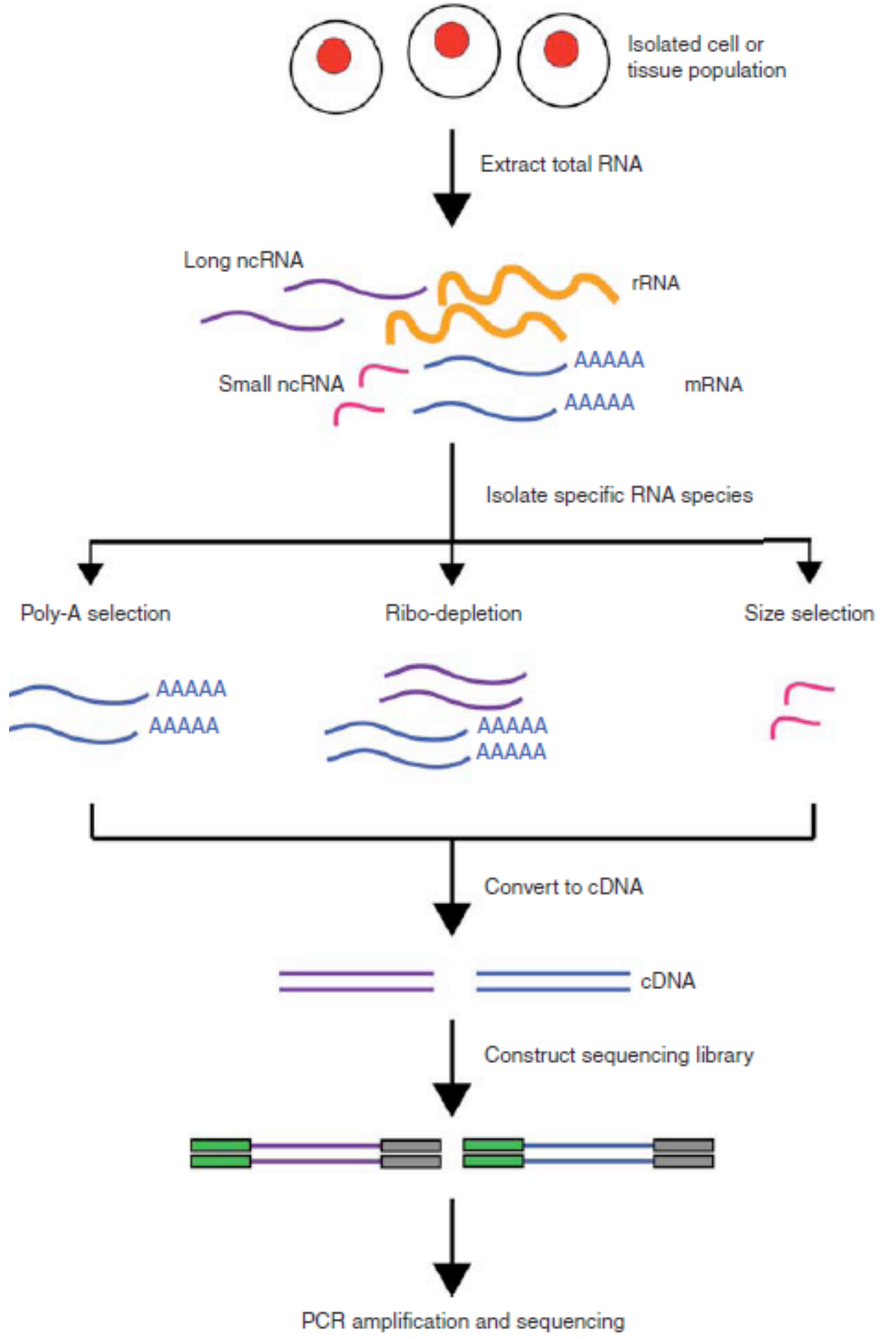
  ➡targets mRNAs

- **Ribosomal RNA depletion** - standard kit removes cytoplasmic (nuclear-encoded) rRNAs; 'gold' kit removes both mitochondrial and cytoplasmic rRNAs; special 'blood' kit also removes global mRNA

  ➡targets mRNA and long ncRNA

- **Size selection** - targets a specific size of transcript

  ➡targets smRNA and microRNA

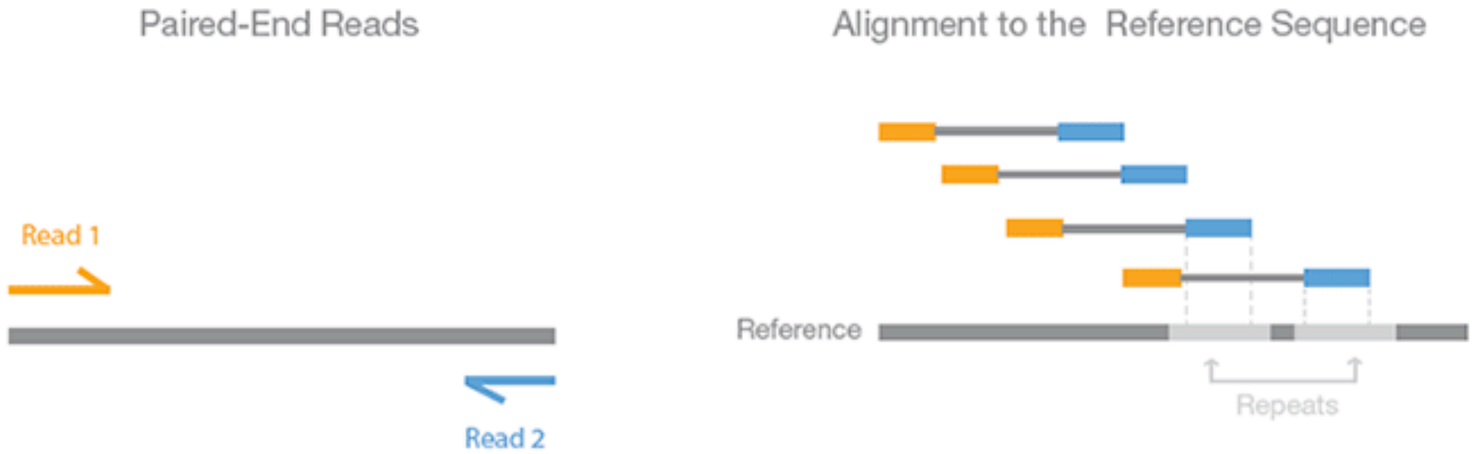


http://cshprotocols.cshlp.org/content/early/2015/04/11/pdb.top084970.full.pdf+html

Library Type

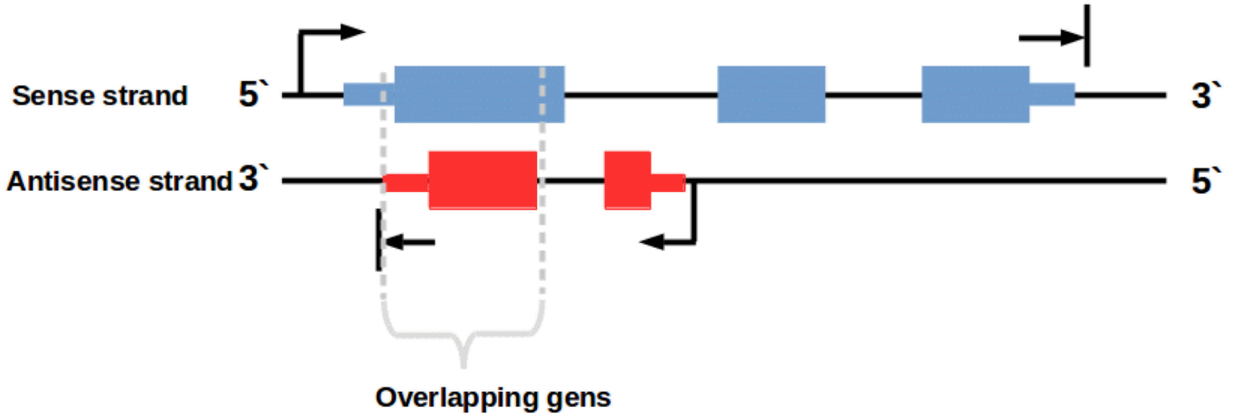
paired-end vs. single end

- **Single-end reads** - sequences only one end of the RNA fragment
  - cheaper
  - ‘good enough’ in many situations
- **Paired-end reads** - sequences both ends of the RNA fragment
  - more expensive
  - allows for the unambiguous alignment of more reads
  - better for *de novo* transcript discovery and isoform-level expression



Library Type

stranded vs. unstranded



- **unstranded reads** - sequences both the original strand and the complement; cannot distinguish between the two
  - Sometimes cheaper
  - 'good enough' in many situations
- **stranded reads** - sequences only one strand
  - Can be more expensive
  - allows for the alignment of reads to a specific strand
  - better for *de novo* transcript discovery - especially one-exon transcripts (often ncRNA)

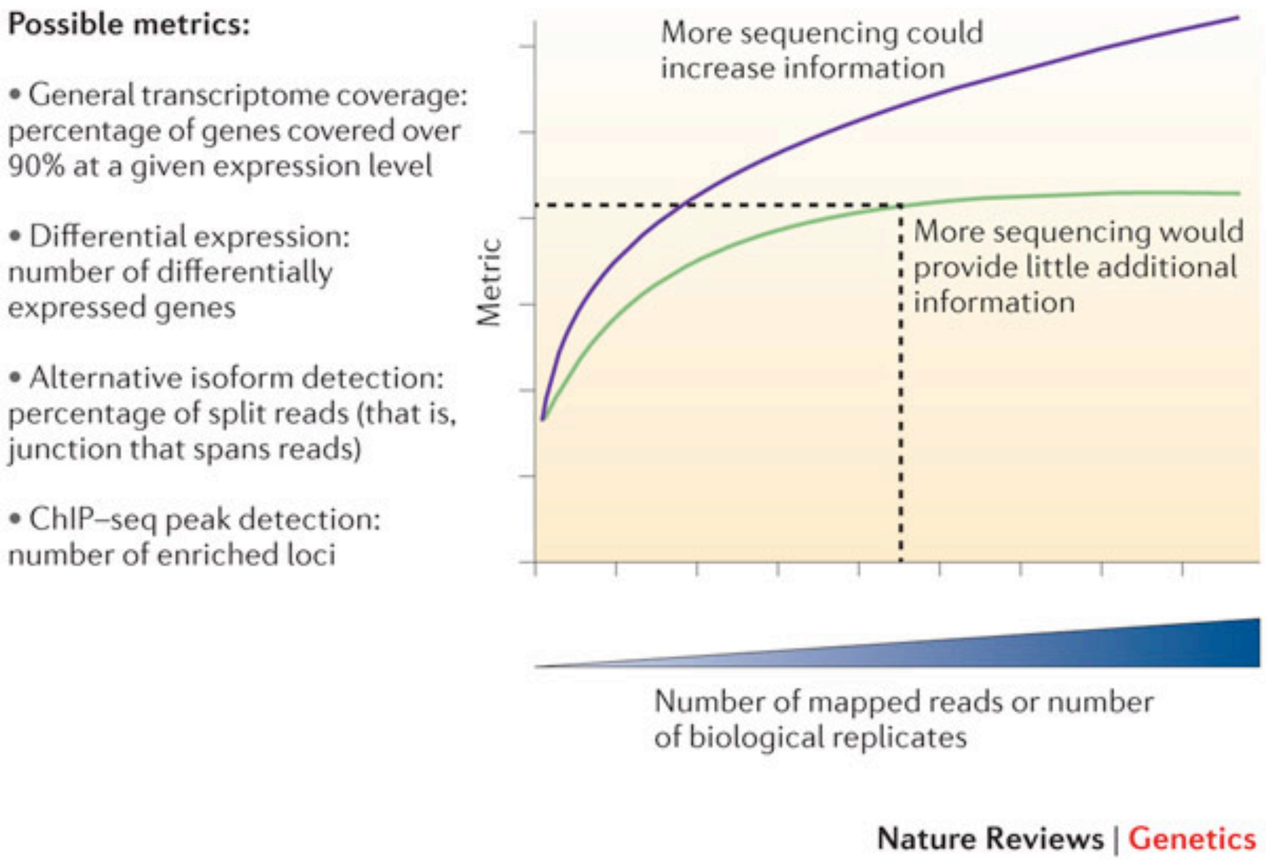
https://www.ecseq.com/support/ngs/how-do-strand-specific-sequencing-protocols-work

Sequencing Length

- **Shorter reads**
  - Cheaper
  - Longer reads are not necessary for small RNA fraction sequencing
- **Longer reads**
  - More expensive
  - Improved mappability
  - Improved accuracy of transcript identification

Sequencing Depth

- Gene-level analysis with known transcriptome from polyA-selected RNA in liver
  - 20-30 million reads/sample
- Isoform-level analysis with *de novo* transcript discovery from rRNA-depleted total RNA in brain
  - 70-80 million reads/sample



"the number of reads that is required in an experiment is determined by the least abundant RNA species of interest — a variable that is not known before sequencing"

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., & Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. Nature Reviews. Genetics, 15(2), 121–132.

Number of Replicates

Number of replicates needed is dependent on the following that varies from gene to gene:

- Within group variance
- Read coverage
- Desired detectable effect size

Table 1 Statistical power to detect differential expression varies with effect size, sequencing depth and number of replicates

	Replicates per group		
	3	5	10
Effect size (fold change)			
1.25	17 %	25 %	44 %
1.5	43 %	64 %	91 %
2	87 %	98 %	100 %
Sequencing depth (millions of reads)			
3	19 %	29 %	52 %
10	33 %	51 %	80 %
15	38 %	57 %	85 %

Example of calculations for the probability of detecting differential expression in a single test at a significance level of 5 %, for a two-group comparison using a Negative Binomial model, as computed by the RNASeqPower package of Hart et al. [190]. For a fixed within-group variance (package default value), the statistical power increases with the difference between the two groups (effect size), the sequencing depth, and the number of replicates per group. This table shows the statistical power for a gene with 70 aligned reads, which was the median coverage for a protein-coding gene for one whole-blood RNA-seq sample with 30 million aligned reads from the GTEx Project [214]

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., et al. (2016). A survey of best practices for RNA-seq data analysis. Genome Biology, 17(1), 13.

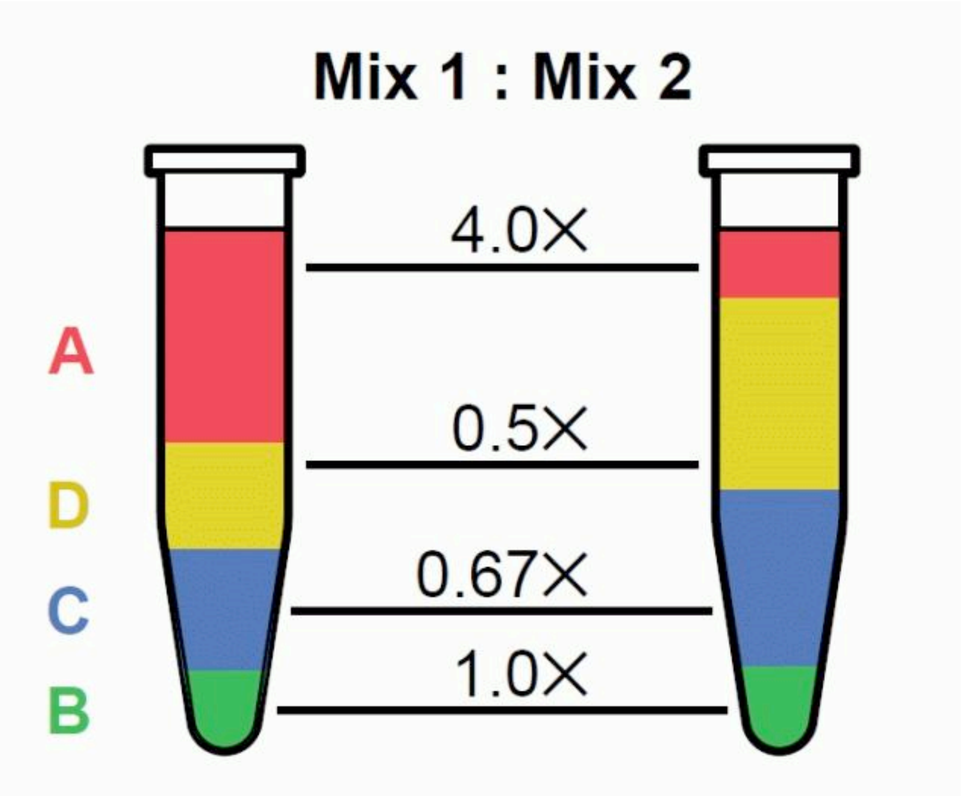
Sequencing Design

1. Synthetic spike-ins
2. Randomization at library prep
3. Randomization at sequencing run

Synthetic Spike-Ins

External RNA Controls Consortium (ERCC) Synthetic Spike Ins

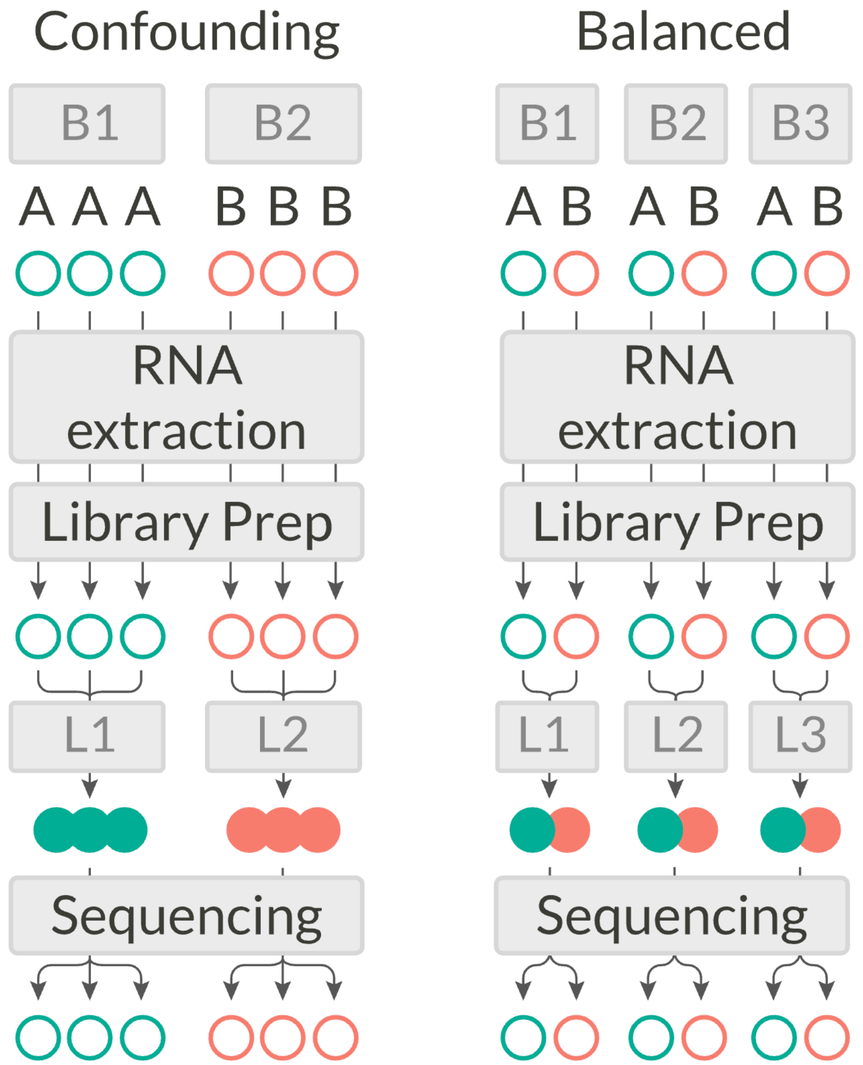
- aid normalization
- generate detection limits for both quantitation and differential expression



Transcript molar ratios in ERCC Spike-In Mixes. The transcripts in Spike-In Mix 1 and Spike-In Mix 2 are present at defined Mix 1:Mix 2 molar concentration ratios, described by 4 subgroups. Each subgroup contains 23 transcripts spanning a 106-fold concentration range, with approximately the same transcript size distribution and GC content.

Randomization at Library Preparation

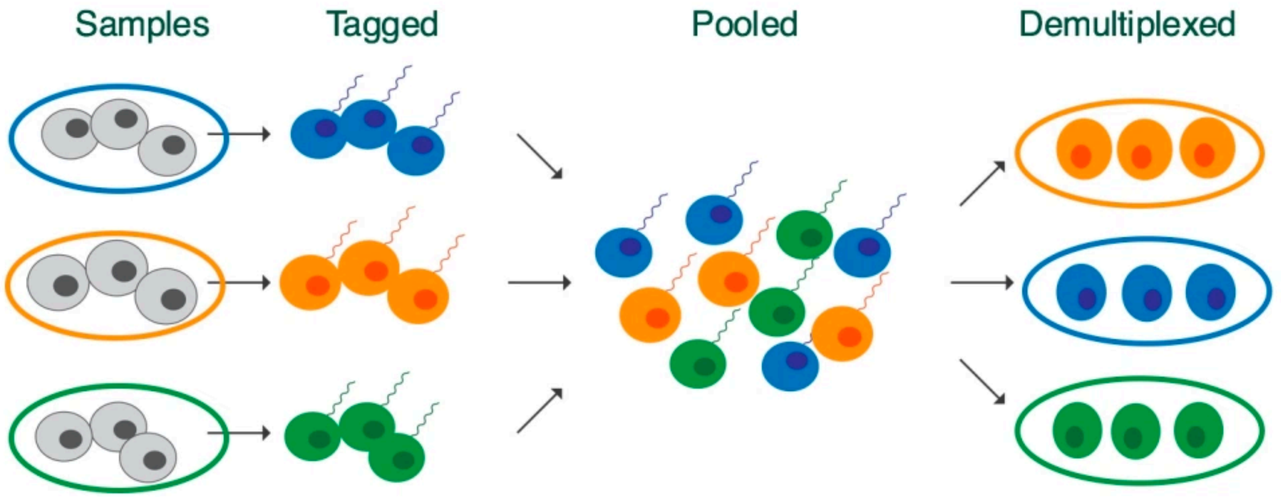
- Library preparation tends to generate the most profound batch effects
- Randomization at library preparation (i.e., spreading samples within a group across multiple batches) helps to:
  - Avoid confounding between technical effects (e.g., batch effects) and effects of interest
  - Minimizes the loss of information from a ‘bad batch’



https://nbisweden.github.io/workshop-ngsintro/2001/slide_rnaseq.html#6

Randomization at Sequencing Run

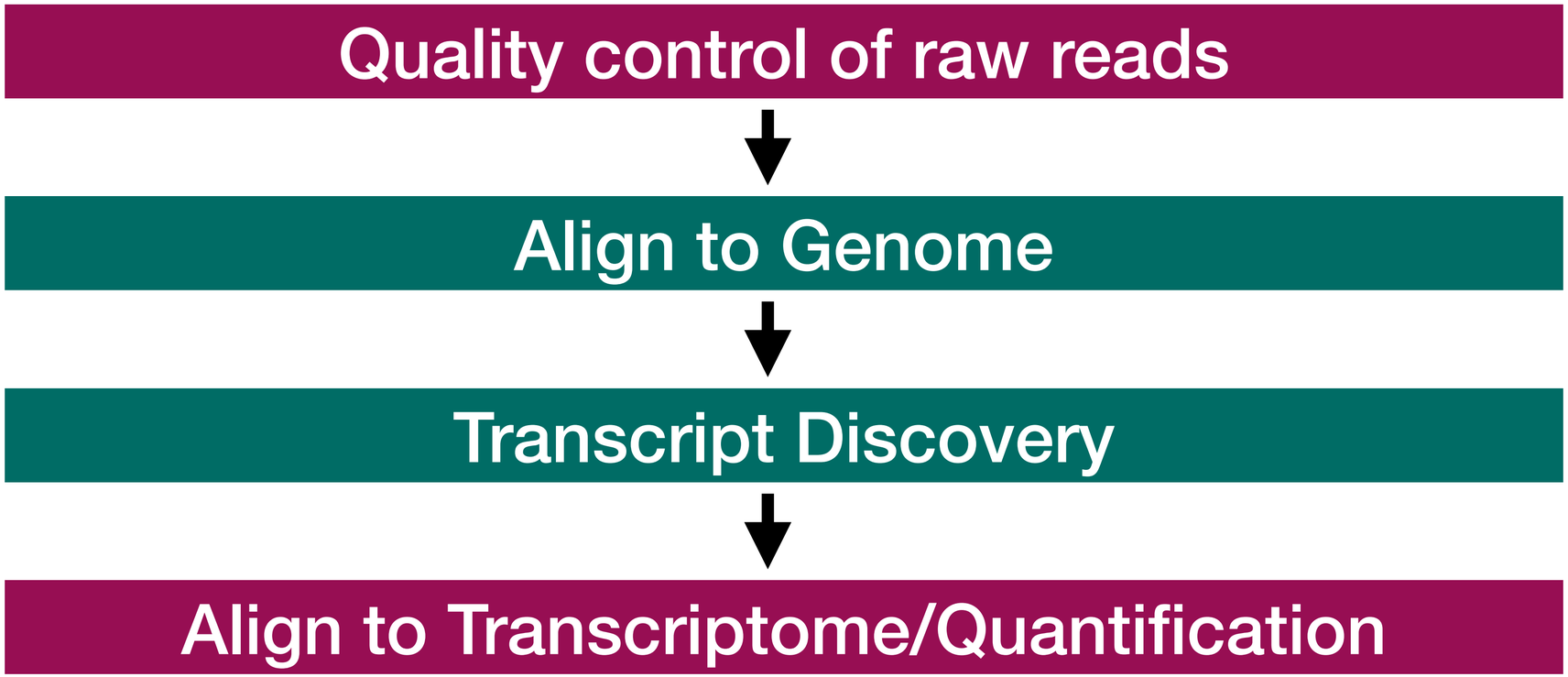
- Sequencing batch tends to contribute little to batch effects
- Multiplexing - RNA fragments from a sample are given a unique nucleotide sequence in the adapter that is sequenced along with the RNA fragment
  - Multiple samples can be combined and put into the same lane
  - This ‘mixture’ of samples can be used in all lanes of a flow cell
  - Not recommended if there are big differences in RNA quality between samples



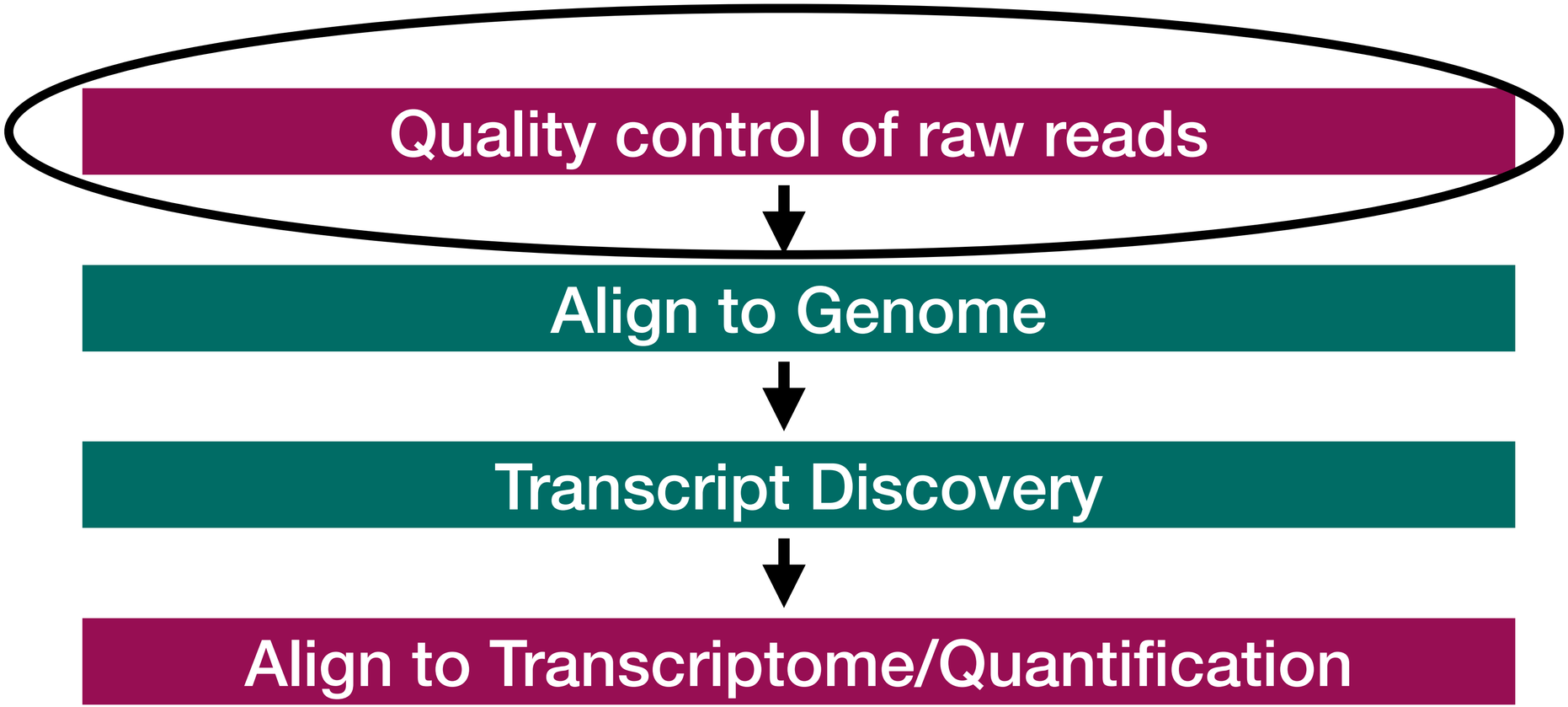
https://liorpachter.wordpress.com/2018/05/30/the-benefits-of-multiplexing/

Transcriptome Profiling

Overview of Transcriptome Profiling



Overview of Transcriptome Profiling



Example data set

GEO Series GSE155709

- CRISPR/Cas9 Knockout of Lrap (a long non-coding RNA associated with alcohol consumption HXB/BXH recombinant inbred panel)
- 3 groups - wild type, heterozygotes, and homozygotes
- 3 male biological replicates per group
- Whole brain samples
- Ribosomal RNA depleted Total RNA
- Paired-end reads

Raw Reads

For paired-end reads, 2 fastq files are generated:

```
Riken-M-KO-1-naive-brain-total-RNA-cDNA_TCTCGCGC_L003_R1_001.fastq.gz
Riken-M-KO-1-naive-brain-total-RNA-cDNA_TCTCGCGC_L003_R2_001.fastq.gz
```

Fastq files have 4 rows per read:

Row 1: Header - includes information to uniquely identify read

Row 2: Sequence - actual nucleotide sequence of read

Row 3: Spacer?

Row 4: Quality - quality metric for each individual base call

```
@GWZHISEQ02:301:C97NKANXX:3:1101:1633:1832 1:N:0:TCTCGCGC
NAGGGAACTCATCAAGTTGCCAAGGTCAGCAGCATGGAGTTGTGTCTCTGTCCCCTCACTGCTGAGTTGTGCTGGCATTTGAGAAGCGGCTGTCATCGAGA
+
#<<<B<BFFFFFFFF<<BFFFFBF<FFFFFFFFFFFFFFBBF<FFFFB/FFFFFFBBFFFFF<<<FFFFB<FF/FFFBFFFFFFBFFFFF/<FB<///<<<
@GWZHISEQ02:301:C97NKANXX:3:1101:1623:1850 1:N:0:TCTCGCGC
NTTGTTGCCATGGTAATCCTGCTCAGTACGAGAGGAACCGCAGGTTCAGACATTTGGTGTATGTGCTTGGCTGAGGAGCCAATGGGGCGAAGCTACCATCT
+
#<<BBBFFFFFFFFBFFFFFFFFFFFFFFFFFFFBFFFFFFB7F<FFFFFBB/<<BF<FF/FFFFFFBFFFFFFFFFFFFFBFFFFFFFFFFFFFFF<</
```

Initial Quality Control

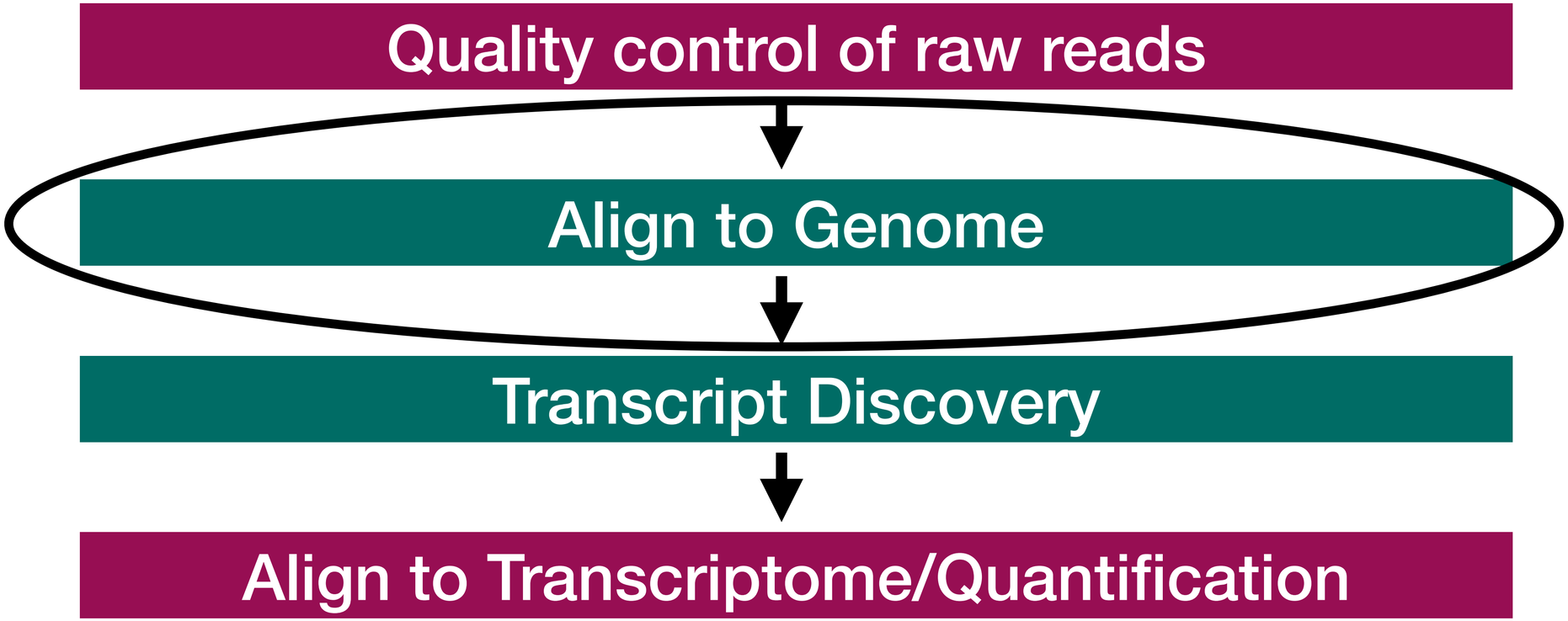
- Count number of raw reads and the average length of each read in each file:
  - Needed later to determine alignment rates
  - Check that paired files have the same number of reads
  - Check that average length is the requested read length
  - Compare read depth across samples (extremely large differences across samples and especially between groups could present confounding issues later on or could represent differences in RNA quality)
- Can use programs like FASTQC to examine quality of reads

Trimming

- cutadapt - command line program (Python based) for removing adaptors and low quality bases from RNA-Seq reads (http://cutadapt.readthedocs.io/en/stable/index.html)
- Quality-based trimming is controversial (Williams, C. R., et al (2016) BMC Bioinformatics, 17(1), 103)
- General summaries about trimming can flag problems with RNA degradation (e.g., trimmed reads are significantly smaller than raw reads)

https://www.med.unc.edu/pharm/calabreselab/files/tufts-sequencing-primer

Overview of Transcriptome Profiling



Read Alignment

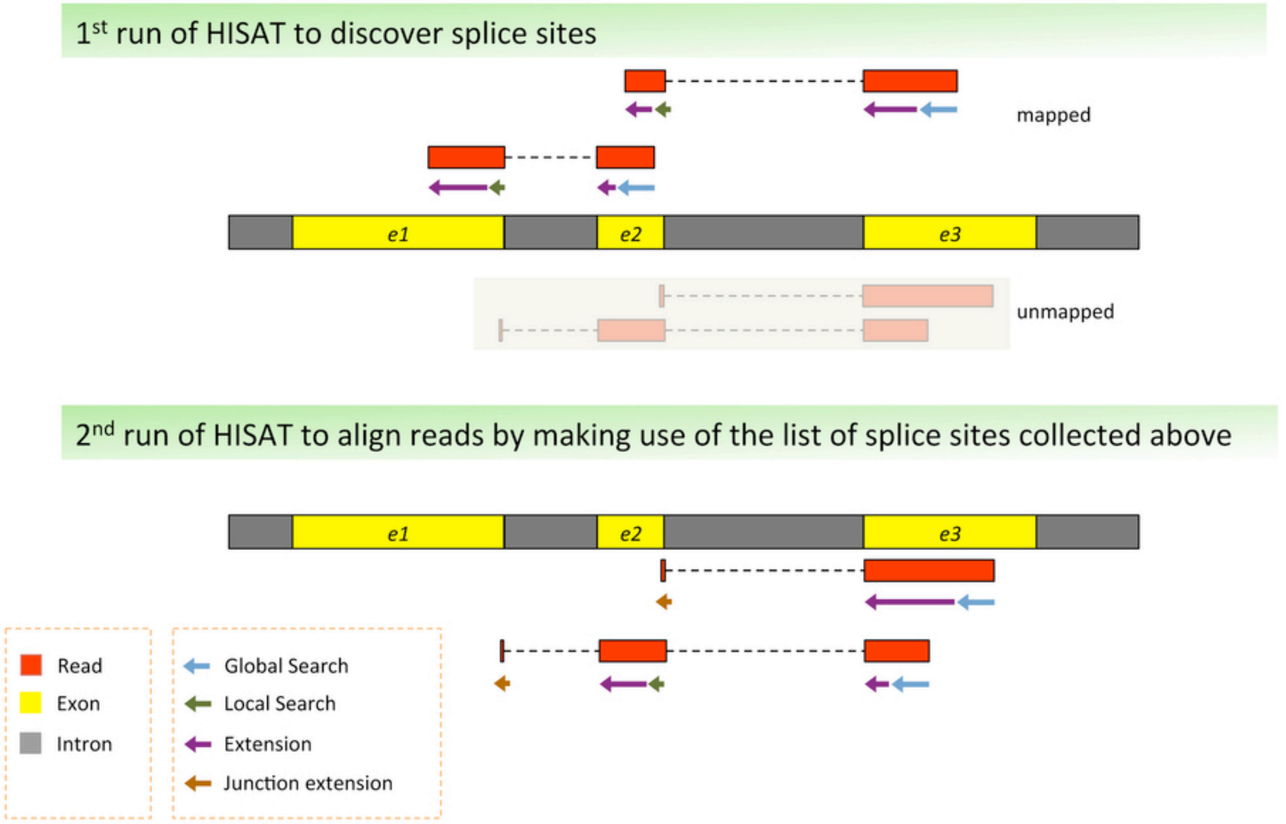
Initial Choice

- Genome vs. Transcriptome
  - **Genome** - when doing de novo transcript identification and also for visualization
  - **Transcriptome** - when using known transcriptome assembly
- Subject-specific vs Reference
  - **Reference genome/transcriptomes** are available from Ensembl among others
  - **Subject-specific genomes** can be used if available to alleviate alignment issues due to variants

Read Alignment to Genome

HISAT2

Hierarchical Indexing for Spliced Alignment of Transcripts (HISAT) - command line tool for spliced alignment



http://ccb.jhu.edu/software/hisat2/index.shtml

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, *12*(4), 357–360.

Read Alignment to Genome

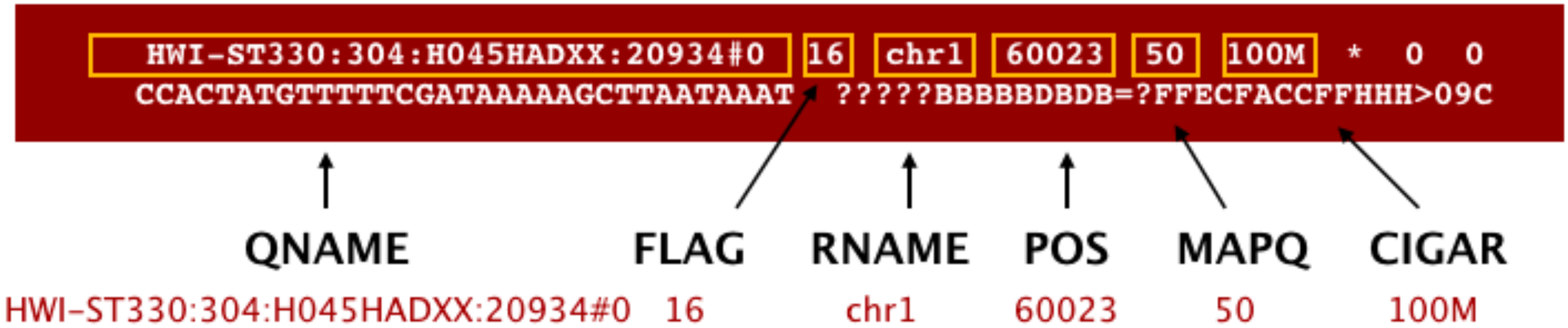
Quality Control

```
10000 reads; of these:
  10000 (100.00%) were paired; of these:
    650 (6.50%) aligned concordantly 0 times
    8823 (88.23%) aligned concordantly exactly 1 time
    527 (5.27%) aligned concordantly >1 times
    ----
    650 pairs aligned concordantly 0 times; of these:
      34 (5.23%) aligned discordantly 1 time
    ----
    616 pairs aligned 0 times concordantly or discordantly; of these:
      1232 mates make up the pairs; of these:
        660 (53.57%) aligned 0 times
        571 (46.35%) aligned exactly 1 time
        1 (0.08%) aligned >1 times
96.70% overall alignment rate
```

Read Alignment to Genome

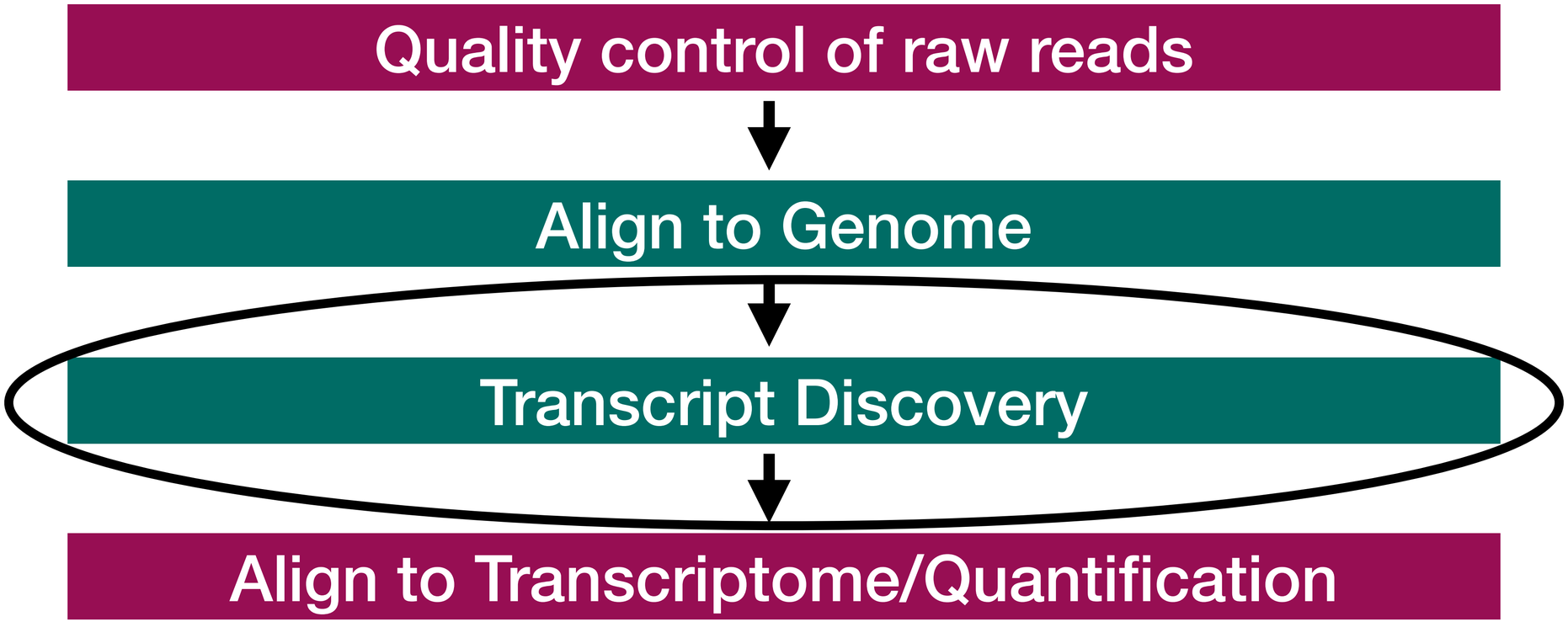
SAM and BAM Files

- Output of most programs is either:
  - SAM file - sequence alignment map
  - BAM file - binary sequence alignment map
- *samtools* - a suite of tools that allows users to manipulate these types of files



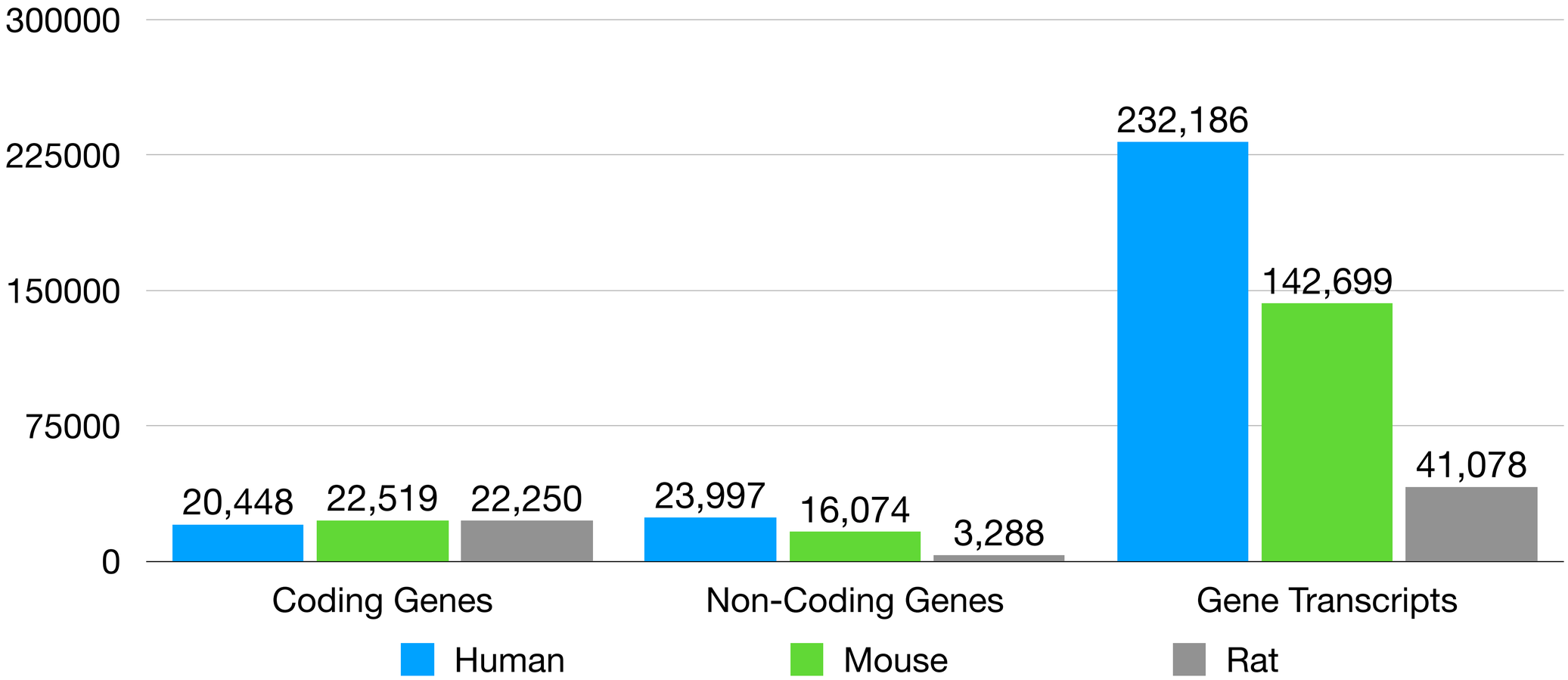
https://hbctraining.github.io/Intro-to-rnaseq-hpc-O2/lessons/04_alignment_quality.html

Overview of Transcriptome Profiling



Transcript Discovery

When and why do we need this?



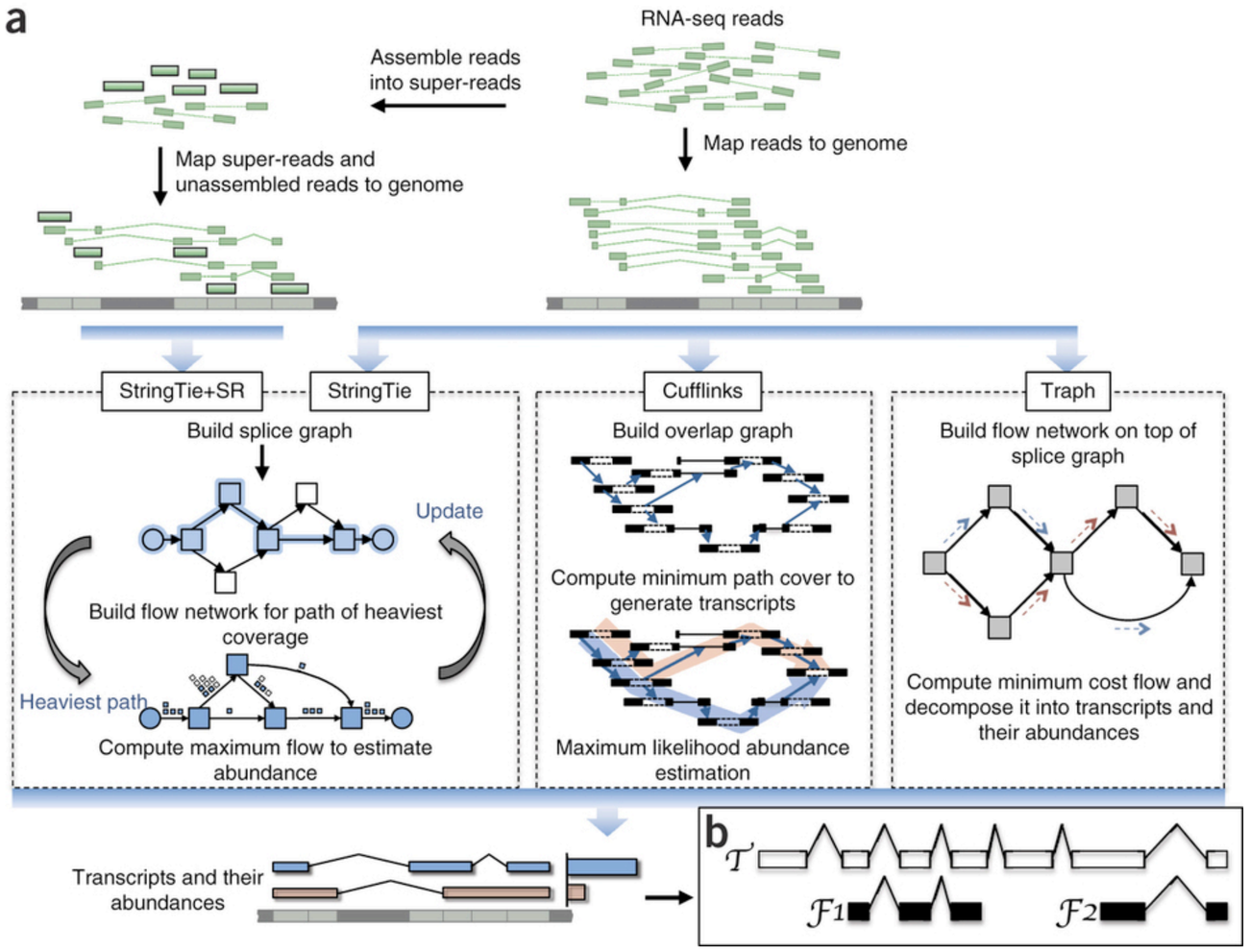
Transcript Discovery

StringTie Algorithm

StringTie - command line tool for *de novo* transcriptome assembly from genome-aligned reads

https://ccb.jhu.edu/software/stringtie/

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, *33*(3), 290–295.

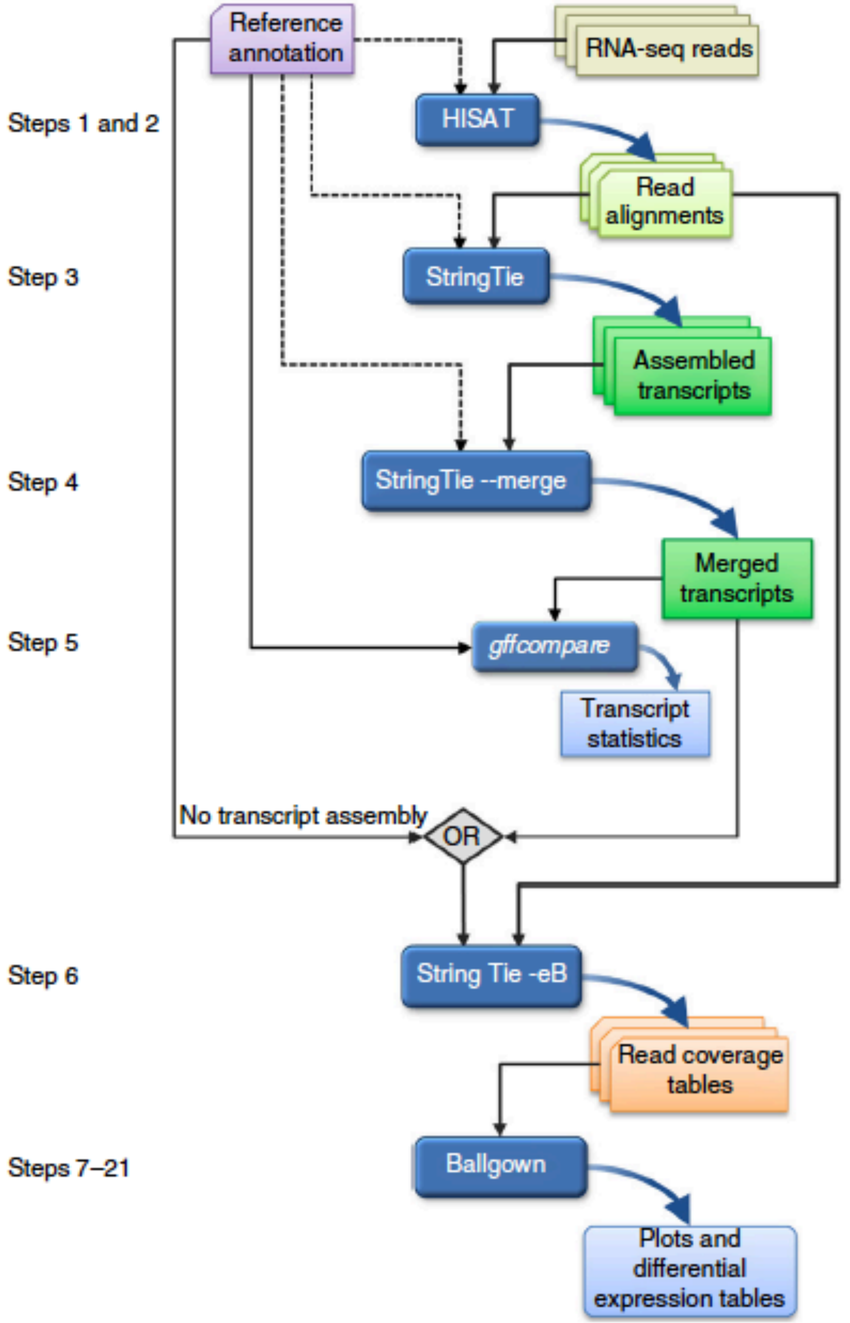


Transcript Discovery

StringTie Pipeline

1. Create sample-specific transcriptome assemblies
2. Merge sample-specific transcriptome assemblies
3. Quantify transcripts in merged transcriptome assembly

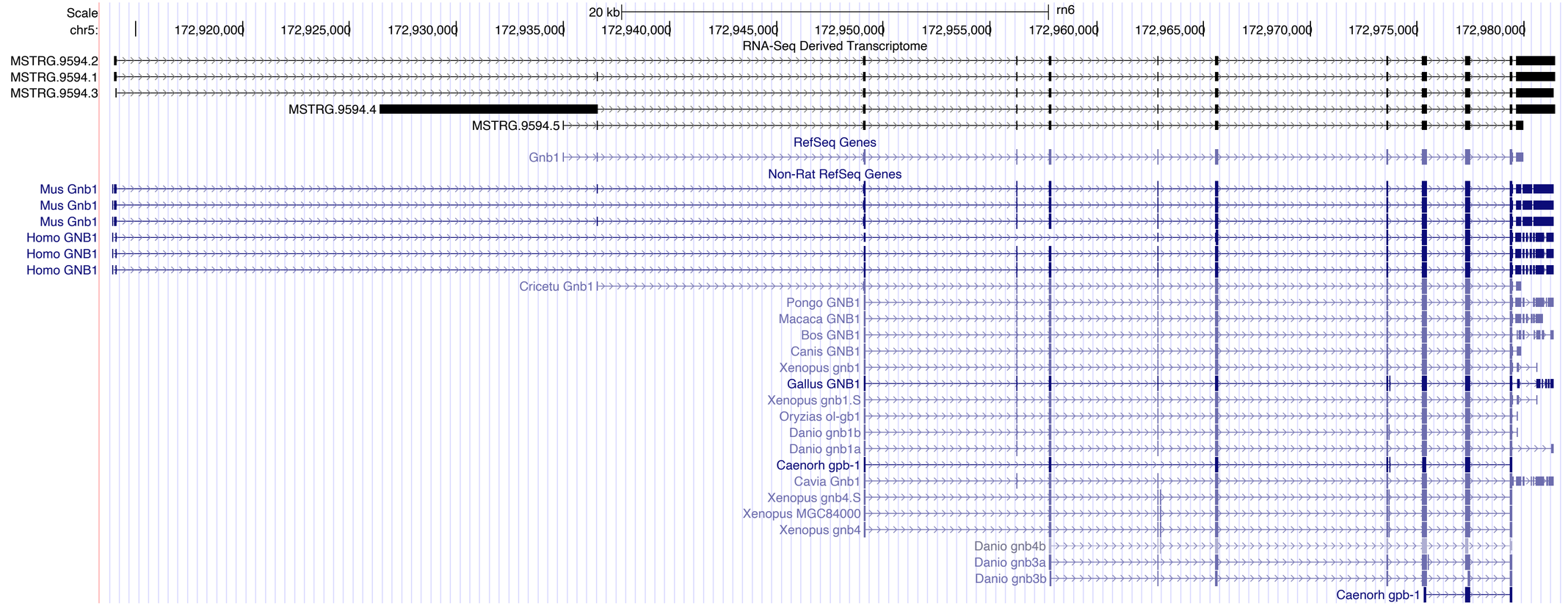
Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat Protoc. 2016 Sep;11(9):1650-67.



Transcriptome Discovery

Visualizations

GTF formatted files generated by StringTie can be easily visualized using the UC Santa Cruz Genome Browser (https://genome.ucsc.edu/) by uploading it as a custom track.

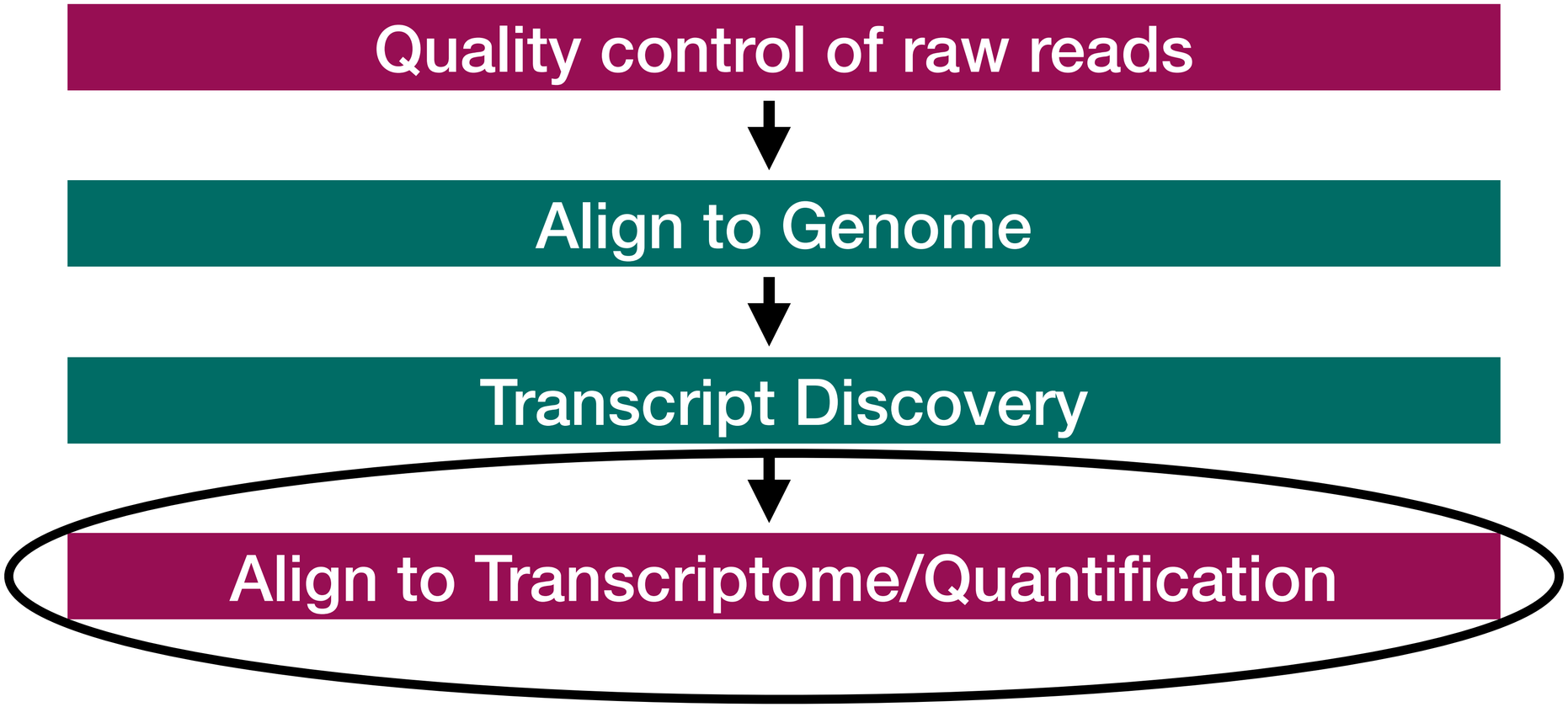


Transcriptome Discovery

Caveats and Potential Pitfalls

- Accuracy of transcript reconstruction is still not optimal
- Transcription start sites and transcription stop sites are notoriously bad (not what algorithm was designed for)
- Uncertainty in structure is not propagated into differential expression analysis

Overview of Transcriptome Profiling



Quantification

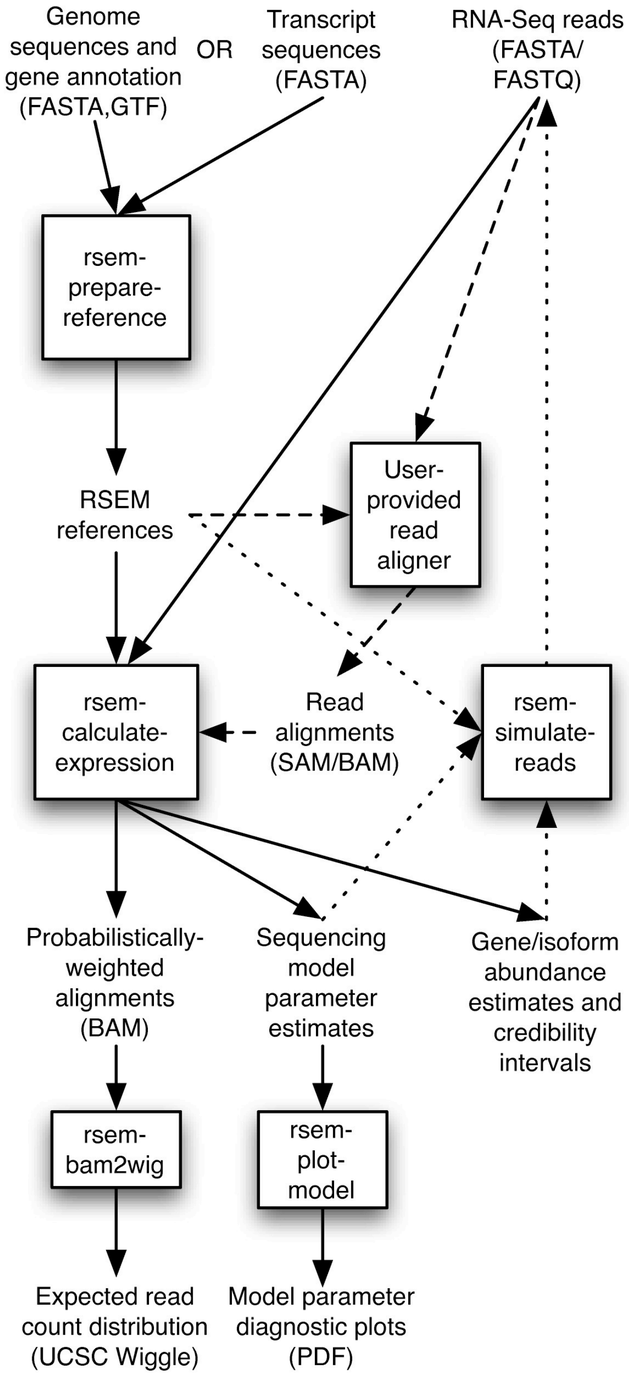
Initial Options

- **Gene-level analysis** - reads are counted that align to any isoform of the gene
  - Used with a shallow read depth per sample
  - Used when interest is in identifying key pathways, not necessarily key players in the pathway
- **Isoform-level analysis** - read counts for individual isoforms (i.e., transcripts in Ensembl) are estimated; counts for reads that align to multiple isoforms are probabilistically split between isoforms
  - Used with a deeper read depth per sample and longer paired end reads
  - Used when interested in individual transcripts and when alternative splicing is suspected.

Quantification

Gene and Isoform Counts - RSEM

- RNA-Seq by Expectation Maximization (RSEM) calculates the expected read counts for individual isoforms and genes
- When compared to other quantitation methods, RSEM has a better ROC-based performance in both simulated and real data sets (Teng M, et al. A benchmark for RNA-seq quantification pipelines. Genome Biol. 2016 Apr 23;17:74).



https://deweylab.github.io/RSEM/

Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics, 12(1), 323.

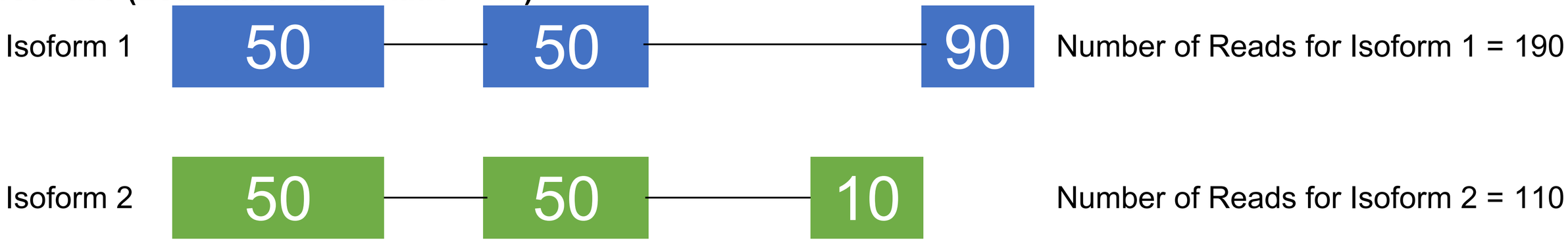
Quantification

General Concept Behind RSEM

Total Number of Reads Aligned to Region

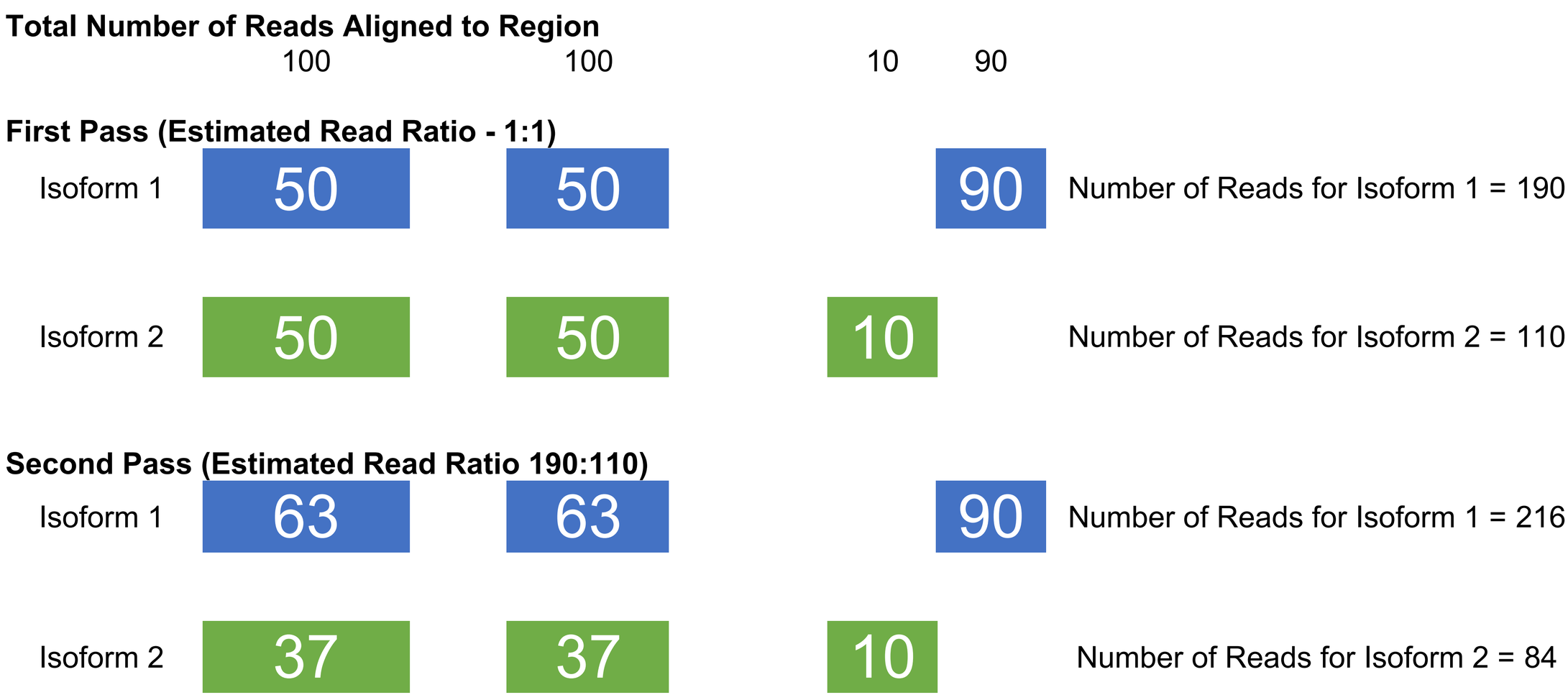
100 100 10 90

First Pass (Estimated Read Ratio - 1:1)



Quantification

General Concept Behind RSEM



Quantification

General Concept Behind RSEM

Quantification

Units

- ***estimated read counts*** - estimated based on model because of reads that align to multiple isoforms/genes
  - used for differential expression analysis
  - not adjusted for library size
- ***FPKM*** - Fragments Per Kilobase of transcript per Million mapped reads (FKPM for paired end reads/RPKM for single end reads)
  - popular, but even the people who first coined this term no longer recommend its use
- ***TPM*** - Transcripts Per Million transcripts
  - recommended over FPKM or RPKM when accounting for library size

Quantification

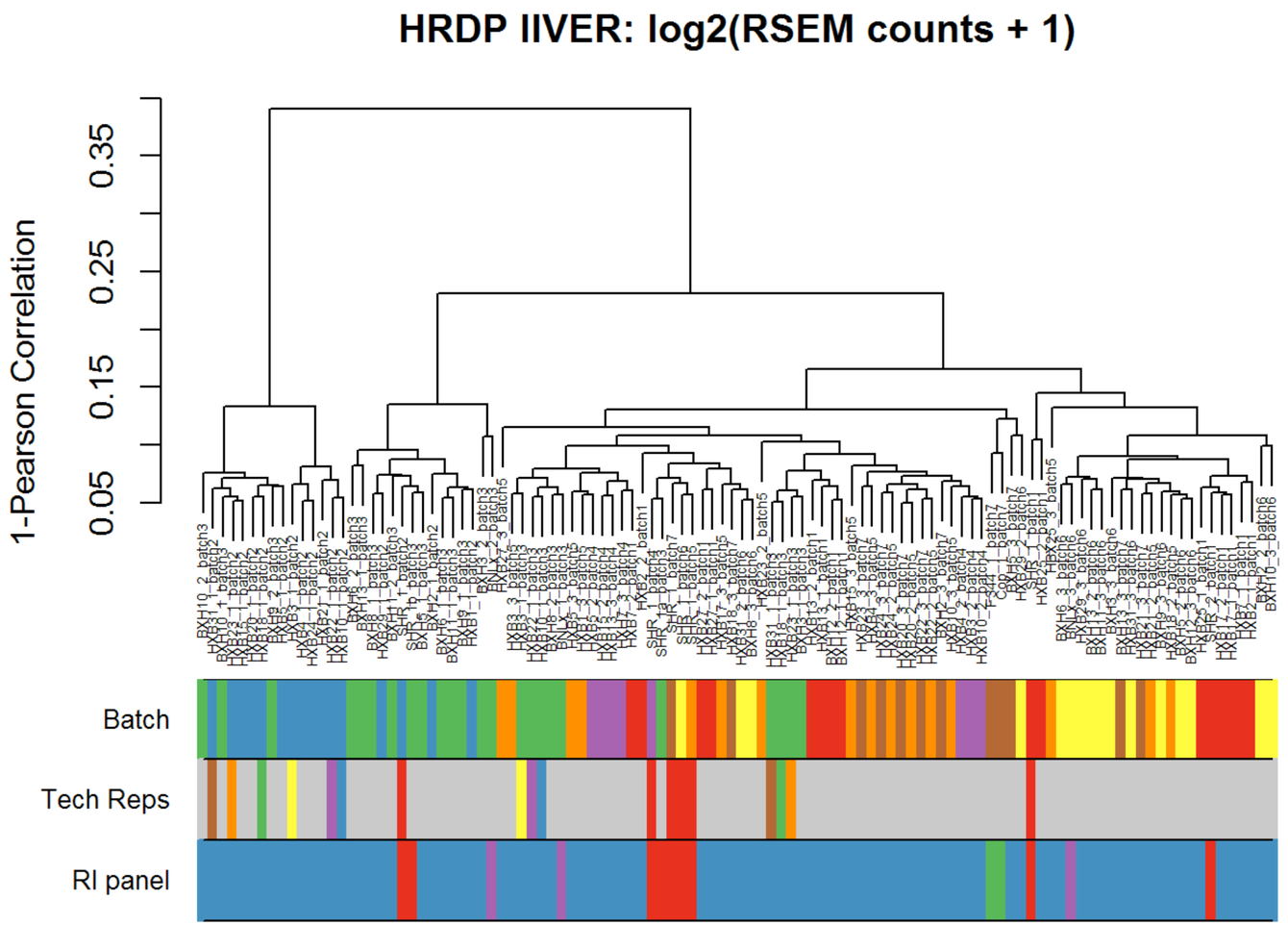
Quality Control

- Exploratory Data Analysis - i.e., visualization (principle component analysis, hierarchical clustering, relative log expression)
  - **Key observations** - do the samples ‘look similar’? How strong is the correlation among samples from the same group at a global scale?
  - **Possible consequences** - removal of outlier/bad samples, further processing to remove technical artifacts (e.g., batch effects)
- Other metrics useful to compare among samples:
  - Proportion of all reads aligned that are attributed to the top 10 transcripts/genes
  - Proportion of transcripts/genes with a zero read count

Quantification

Using a hierarchical clustering for QC

- Dendrogram indicate the ‘natural’ clustering of samples.
- Goal - more clustering based on biological factors, less clustering based on technical factors



Differential Expression

Differential Expression

1. Preprocessing
2. Differential expression of genes and/or isoforms
3. Other types of differential expression

Differential Expression

Preprocessing - Low Counts

- In general, when counts are low, differential expression analyses lose power and are more likely to produce spurious results
- Often, low counts genes/isoforms are removed prior to differential expression analysis with the assumption that their expression estimates are ‘below background’.
- No commonly accepted level for ‘background’ expression. Often use the total number of counts across all samples to include genes/isoforms that are expressed in only one group or the proportion of samples with read counts above a given value.
- DESeq2 threshold - valuable if differential expression is your only analysis but may not be as satisfying as a concrete threshold

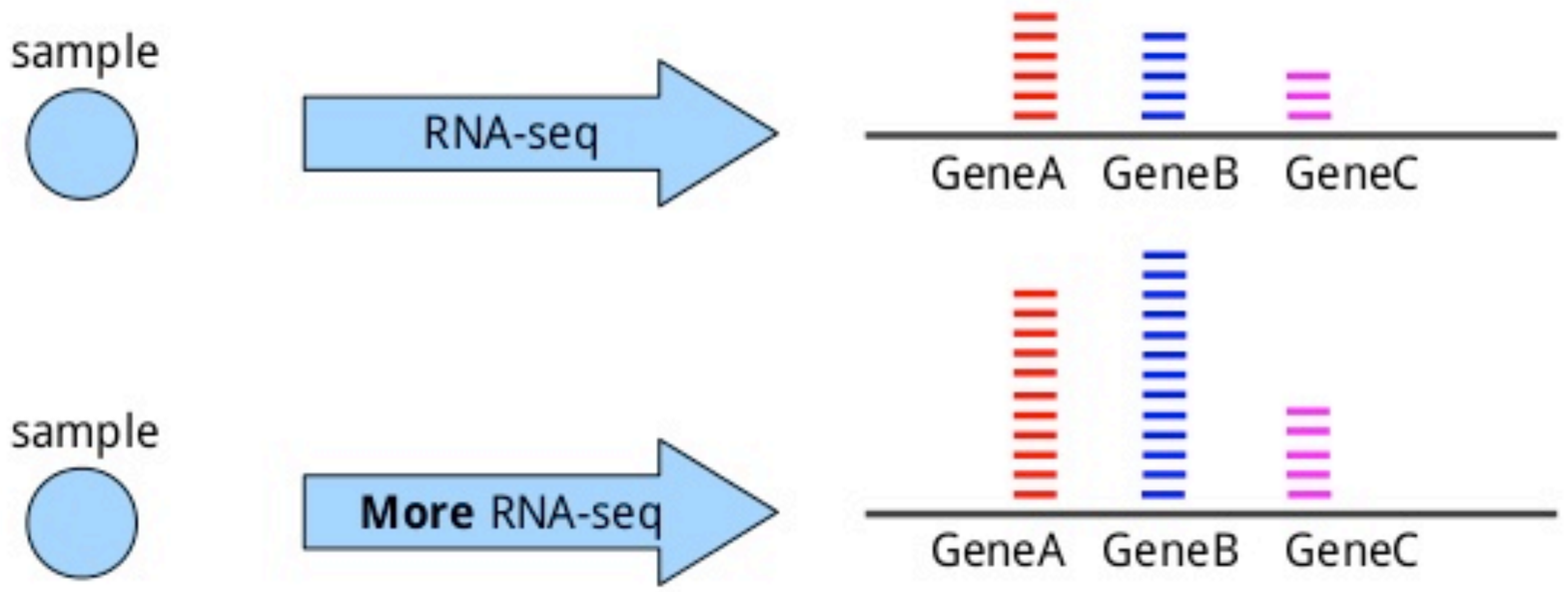
General Philosophy - Never perform a statistical test that you don’t intend on interpreting

Differential Expression

Preprocessing - Library Size Bias

This bias is often taken care of in one of the following ways:

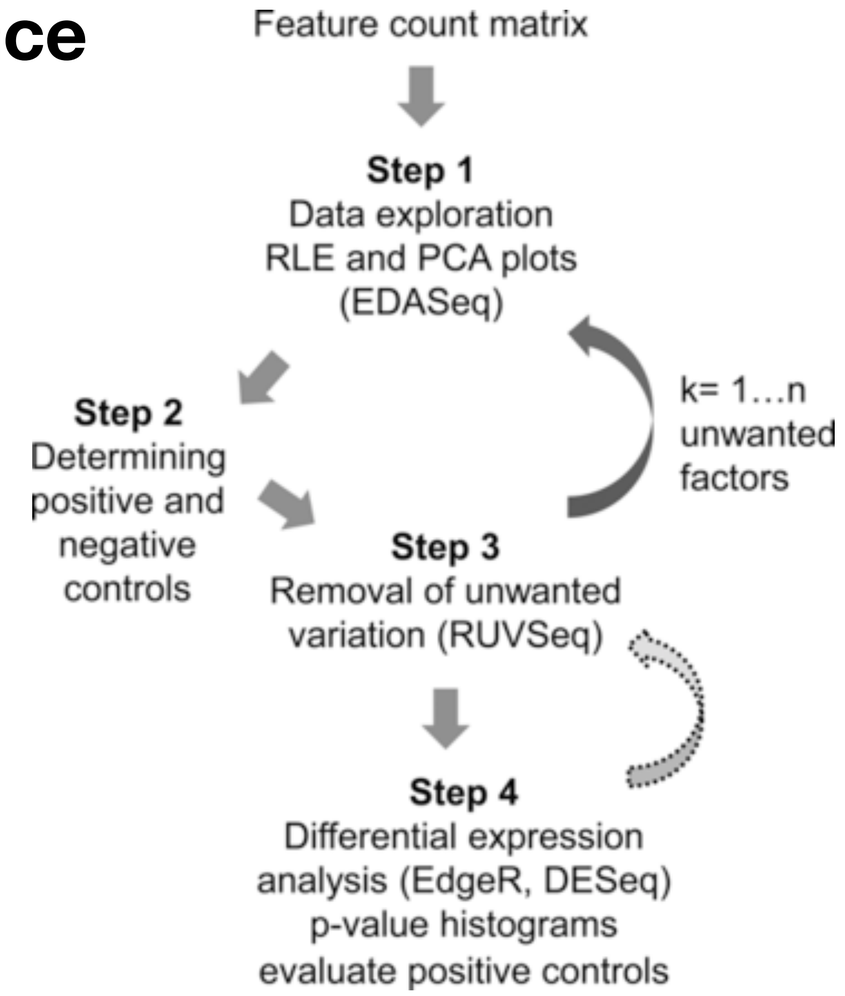
- TPM/CPM/FPKM measures
- Negative binomial model that includes library size
- Scaling - e.g., upper quantile



Differential Expression

Preprocessing - Remove Unwanted Variance

- Determines latent factors (must specify number) that account for unwanted variance using either:
  - control samples
  - negative control genes
  - empirically-derived negative control genes (performed best in our hands even when the other two are available)
- Factors can be included as covariates in differential expression analyses or counts can be directly ‘adjusted’.
- Implemented in RUVSeq package in R (Risso, D., et al. (2014). Normalization of RNA-seq data using factor analysis of control genes or samples. Nature Biotechnology, 32(9), 896–902)



Differential Expression

Comparing Read Counts

Sample A	Sample B
1	2
100	200
1100	1200
0	2
0	2000

Differential Expression

Negative Binomial Model

Goal – Find 5 people that have seen the movie *Office Space*

How many people will I have to ask before I find 5 people?

Differential Expression

Negative Binomial Model

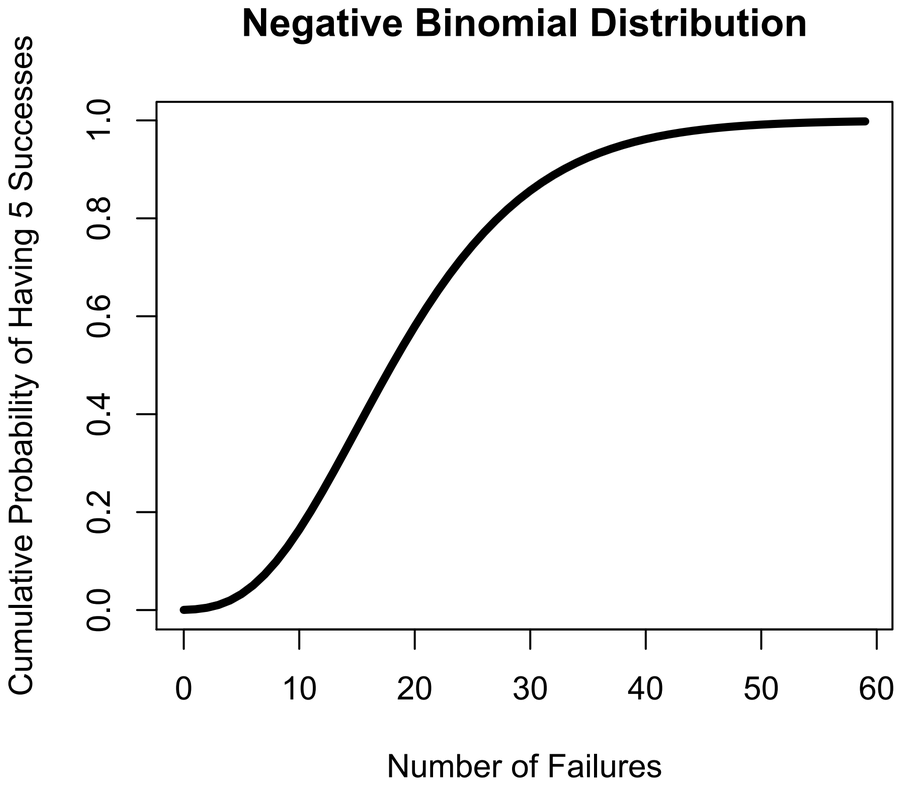
- Observation: I had to ask 20 people before encountered 5 people that had seen it
  - My instincts tell me:
    - Estimate: the proportion of people that have seen Office Space is 5/20 = 25%
    - I would be surprised if the **true** proportion wasn't between 10% to 40%
- Observation: I had to ask 200 people before encountered 50 people that had seen it
  - My instincts tell me:
    - Estimate: the proportion of people that have seen Office Space is 50/200 = 25%
    - I would be surprised if the **true** proportion wasn't between 20% to 30%

Differential Expression

Negative Binomial Distribution

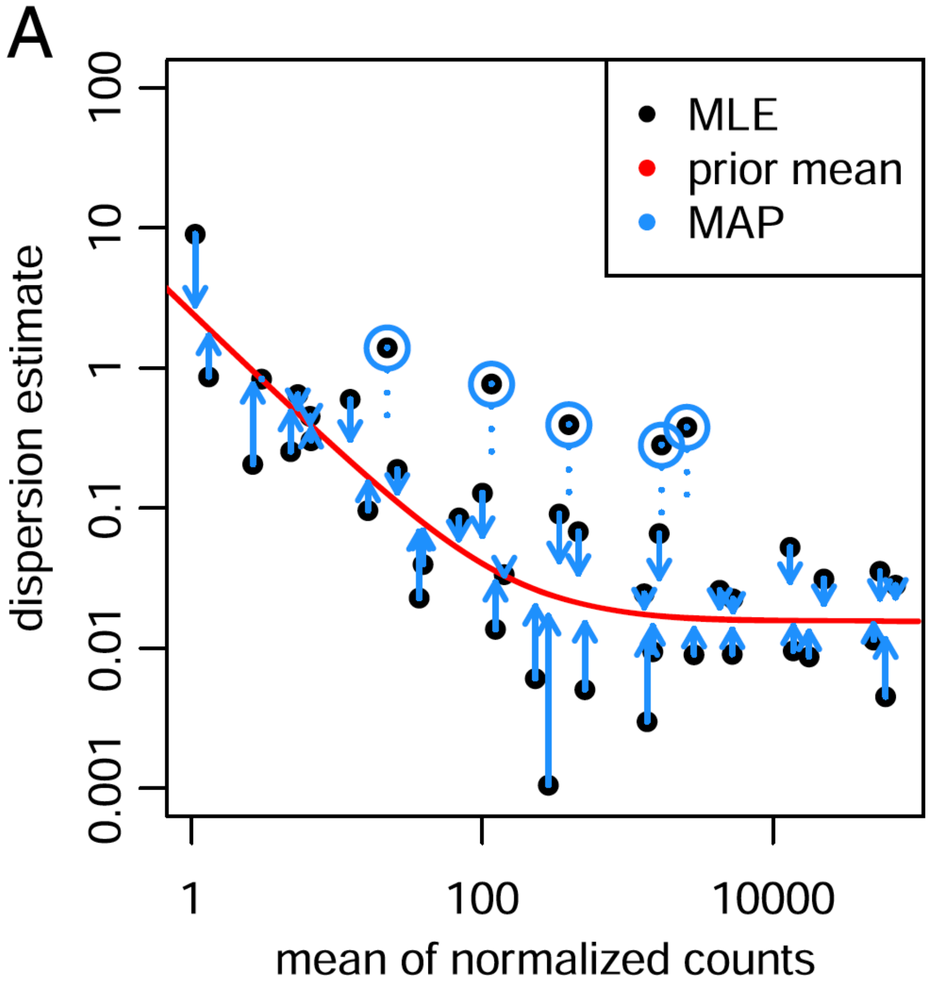
Can we estimate the “proportion of people who have seen Office Space” from how many failures before we have 5 successes?

Goal in RNA-Seq – identify genes that expressed in different “proportions” across groups



Differential Expression

DESeq2



DESeq2, implemented in R, uses a negative binomial model that ‘shrinks’ both dispersion and log fold differences to stabilize estimates

Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.

Differential Expression

DESeq2 - Pros and Cons

- Easy to implement in R
- Can handle multiple covariates
- Can test for differences between models (e.g., omnibus test for group effect in the presence of more than two groups)
- Cannot incorporate random effects (e.g., longitudinal studies with multiple observations from the same sample)

Differential Expression

Regularized Log Transformation/Variance Stabilizing Transformation

- Regularized log transformation implemented in DESeq2 transforms counts into a value that is approximately normally distributed with homoskedastic variances
- These values can be used for correlation style analyses
- Could be used in repeated measures models or more complex models.

Differential Expression

False Discovery Rate

- Many times for RNA-Seq studies, a false discovery rate (FDR) rather than a traditional p-value is reported to account for multiple comparisons.
- FDR is the estimated proportion of “significant” tests that are false positives at a particular threshold.
- An FDR value is calculated for each test (e.g., gene), but it is dependent on the distribution of the other test results (e.g., other genes).
- Although an FDR value is calculated for each test, it is more appropriate to report an FDR threshold rather than reporting the FDR for an individual gene.
- When we use a 10% FDR threshold for identification of ‘positive’ results, we are estimating that 10% of the ‘positive’ results are false positives.

Differential Expression

Other Types of Differential Expression

- Alternative splicing
  - e.g., mixture-of-isoforms (MISO) model, a statistical model that estimates expression of alternatively spliced exons and isoforms and assesses confidence in these estimates
- Alternative polyadenylation
  - e.g., Dynamitic analysis of Alternative PolyAdenylation from RNA-seq (DaPars), an algorithm that identifies alternative polyadenylation (APA) sites and dynamic APA usages between two conditions

Conclusions

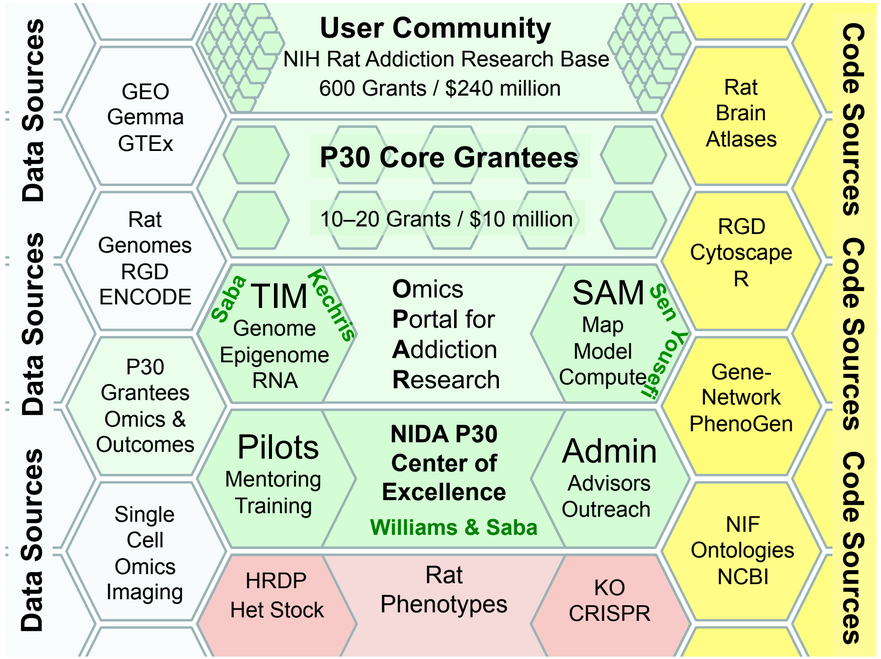
Conclusions

- **Why RNA-Seq?**
  - RNA is one of the first quantitive links between DNA and phenotypes
  - RNA-Seq allows for both identification AND quantization of expression
- **Experimental Design**
  - There are several choices for experimental design that depend on type of experiment and the research question of interest.
  - Because each gene/transcript will have different read coverage and different within group variances, power analysis related to number of samples per group and number of reads per sample are done for a ’typical’ gene rather than all genes.
- **Transcriptome Profiling**
  - Methods and tools have matured and make it relatively simple to move from unmapped reads to quantitated genes/transcripts.
  - Quality control should be explored at multiple steps throughout the procedure to ensure integrity of the data set.
- **Differential Expression**
  - Differential expression is typically measured using a negative binomial model with an FDR correction because of the nature of count data and the multiple testing burden, respectfully.

NIDA Core Center of Excellence in Omics, System Genetics, and the Addictome

Co-Directors: Rob Williams (UTHSC) and Laura Saba (CU-AMC)

The **purpose** of the NIDA P30 Core Center of Excellence in Omics, Systems Genetics, and the Addictome is to empower and train researchers supported by NIH, NIDA, NIAAA, and other federal and state institutions to use more quantitative and testable ways to analyze genetic, epigenetic, and the environmental factors that influence drug abuse risk and treatment.



Our Approach:

- Omics Portal for Addiction Research (OPAR)
- Study design and RNA-Seq analysis services
- Training in Systems Genetics, RNA-Seq, and OPAR usage
- Funding for pilot grants

Acknowledgements

- Saba Lab:
  - Current: Ryan Lusk, Cheyret Wood, Samuel Rosean, and Angela Yoder
  - Former: Lauren Vanderlinden, Harry Smith, and Sean Hickey
- Boris Tabakoff, Paula Hoffman, and their lab
  - Spencer Mahaffey and Jenny Mahaffey

- Financial Support:
  - NIDA Core “Center of Excellence" in Omics, Systems Genetics and the Addictome (NIDA - P30DA044223; MPIs - Williams, Saba)
  - The heritable transcriptome and alcoholism (NIAAA - R24AA013162; MPIs - Tabakoff, Hoffman, Saba