Introduction to RNA-Seq

Laura Saba, PhD
Assistant Professor
Department of Pharmaceutical Sciences
Skaggs School of Pharmacy and Pharmaceutical Sciences
University of Colorado Anschutz Medical Campus
Laura.Saba@ucdenver.edu
http://www.ucdenver.edu/pharmacy/SystemsGeneticsandBioinformatics

Overview

- Why RNA-Seq?
- Technical Overview
- Experimental Design
- Transcriptome Profiling
- Differential Expression
- Functional Enrichment
- In-class example

https://github.com/LauraSaba/IntroToRNASeq

WHY RNA-SEQ?

Why Study the RNA Dimension?

Transcriptome links DNA and complex traits/diseases

A. One of the first quantitative links between DNA sequence and phenotype

B. First step where DNA sequence and environment interact

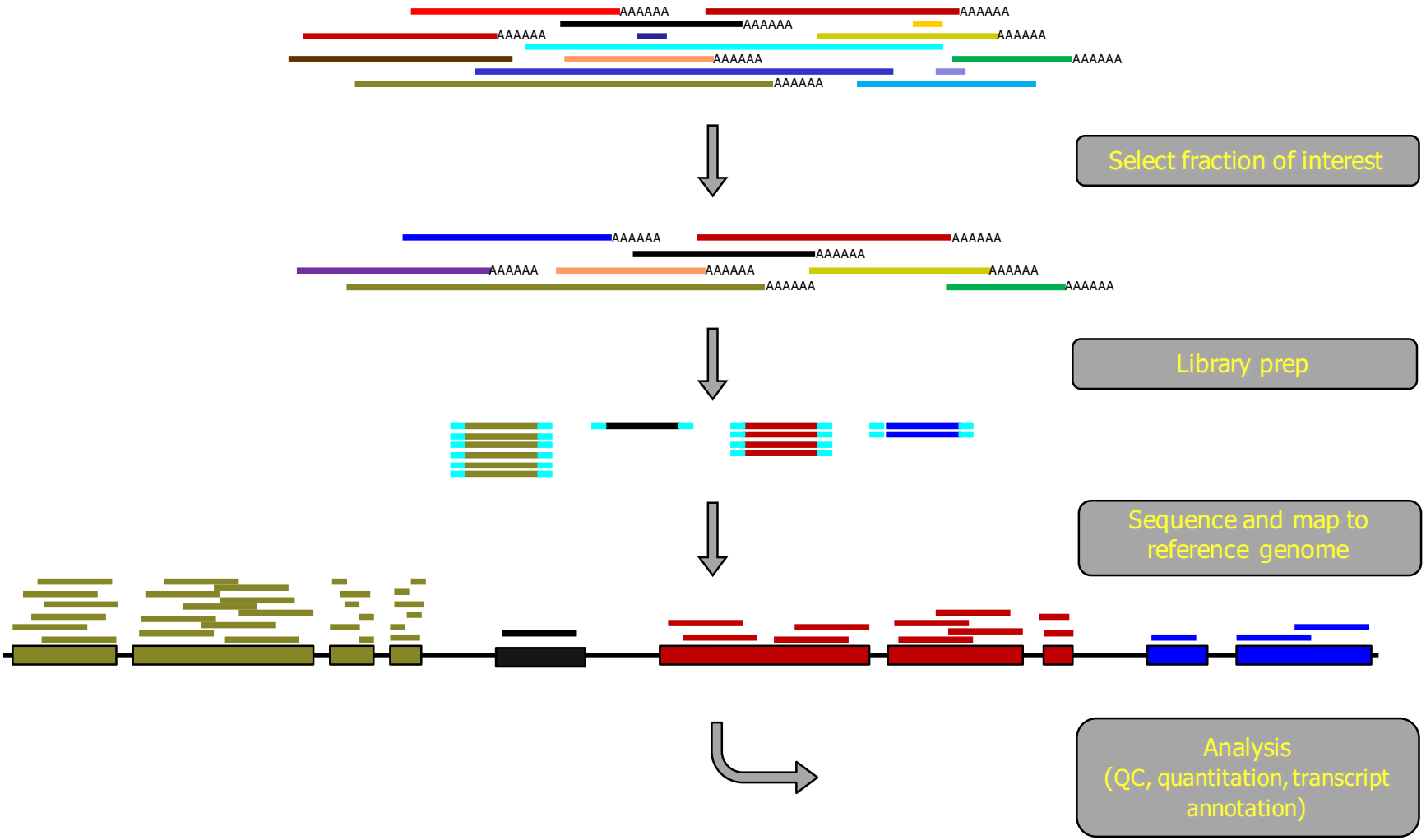
C. Implementation of graph theory at the transcript level provides insight into genetic/environmental interactions that are the basis for susceptibility to complex diseases.

Why RNA-Seq?

- Allows for discovery of:
  - novel protein-coding genes
  - novel splice variants of known 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 of:
  - Single nucleotide variants
  - RNA editing

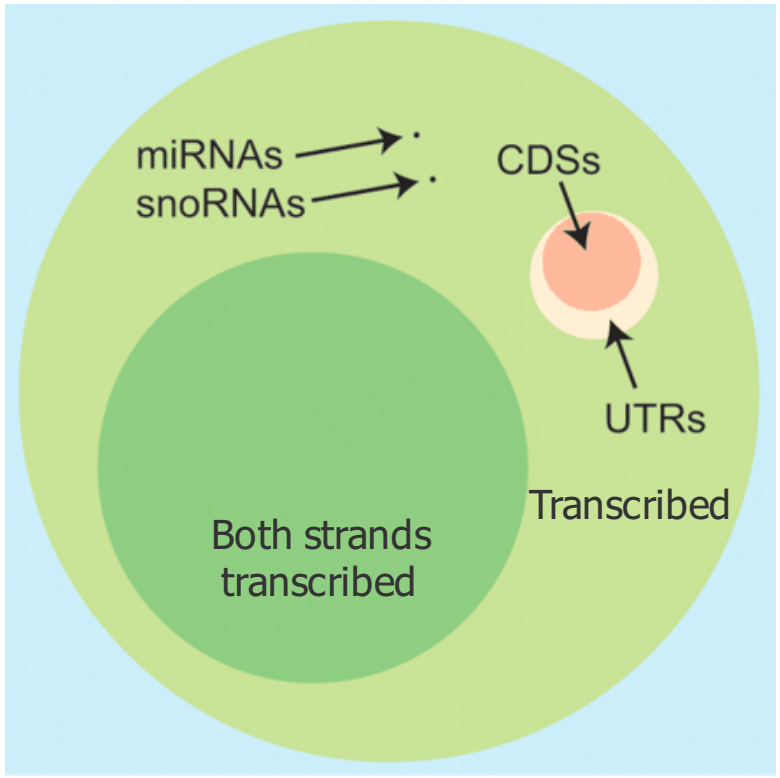
TECHNICAL OVERVIEW

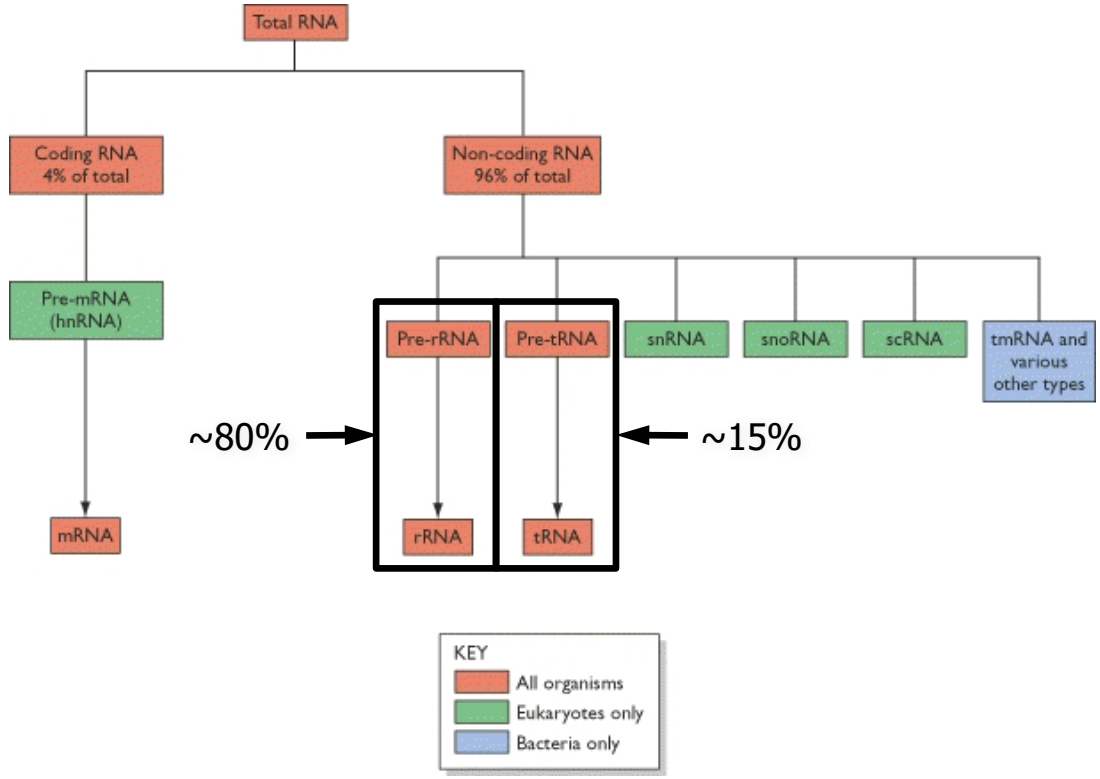
RNA-Seq Overview



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

RNA Fraction





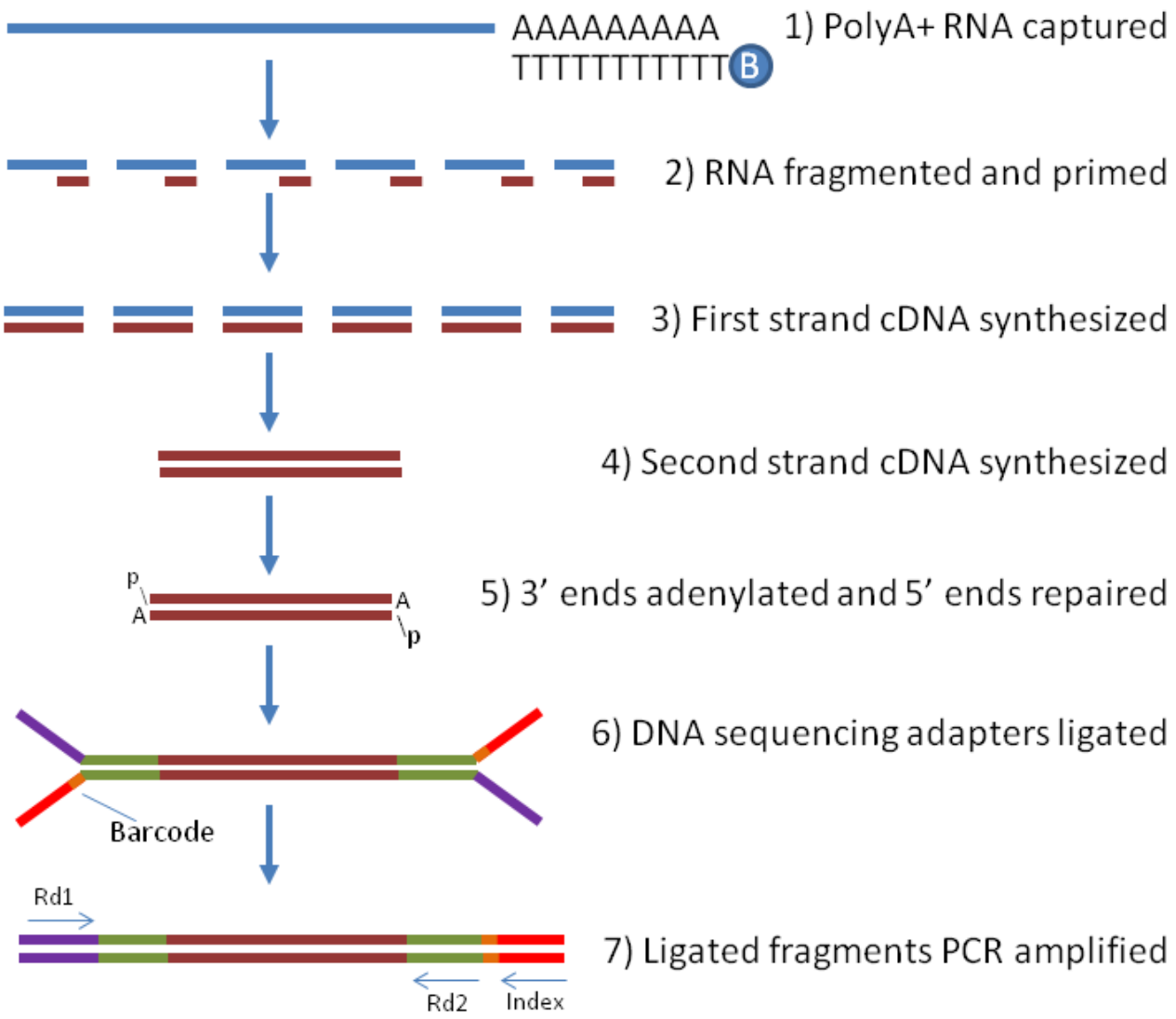
Genomic Distribution

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

Total RNA Distribution

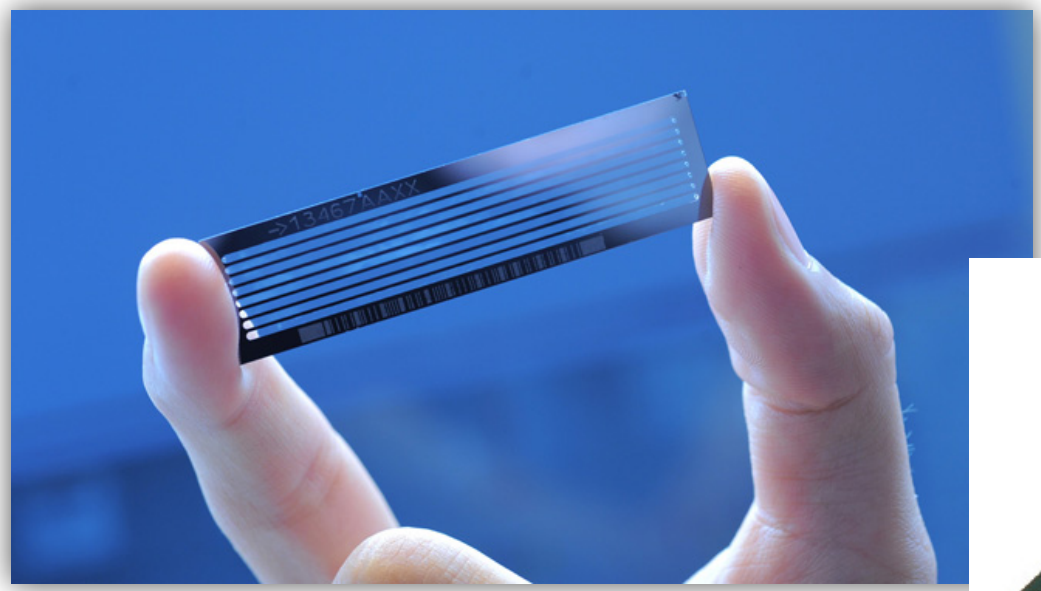
Genomes (2002), 2nd Edition, Chapter 3

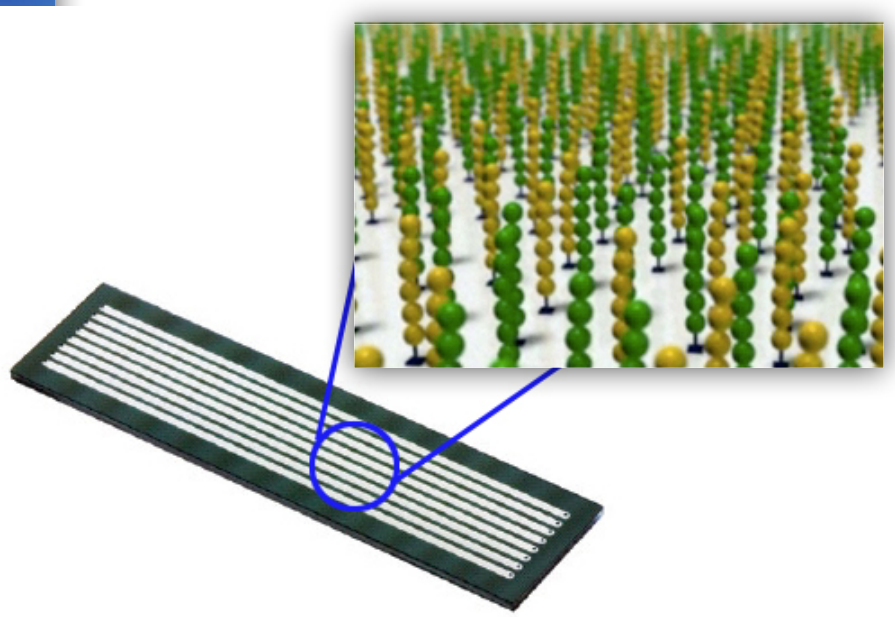
Library Preparation



Corney (2013) Mater Methods 3:203

Illumina System for Sequencing





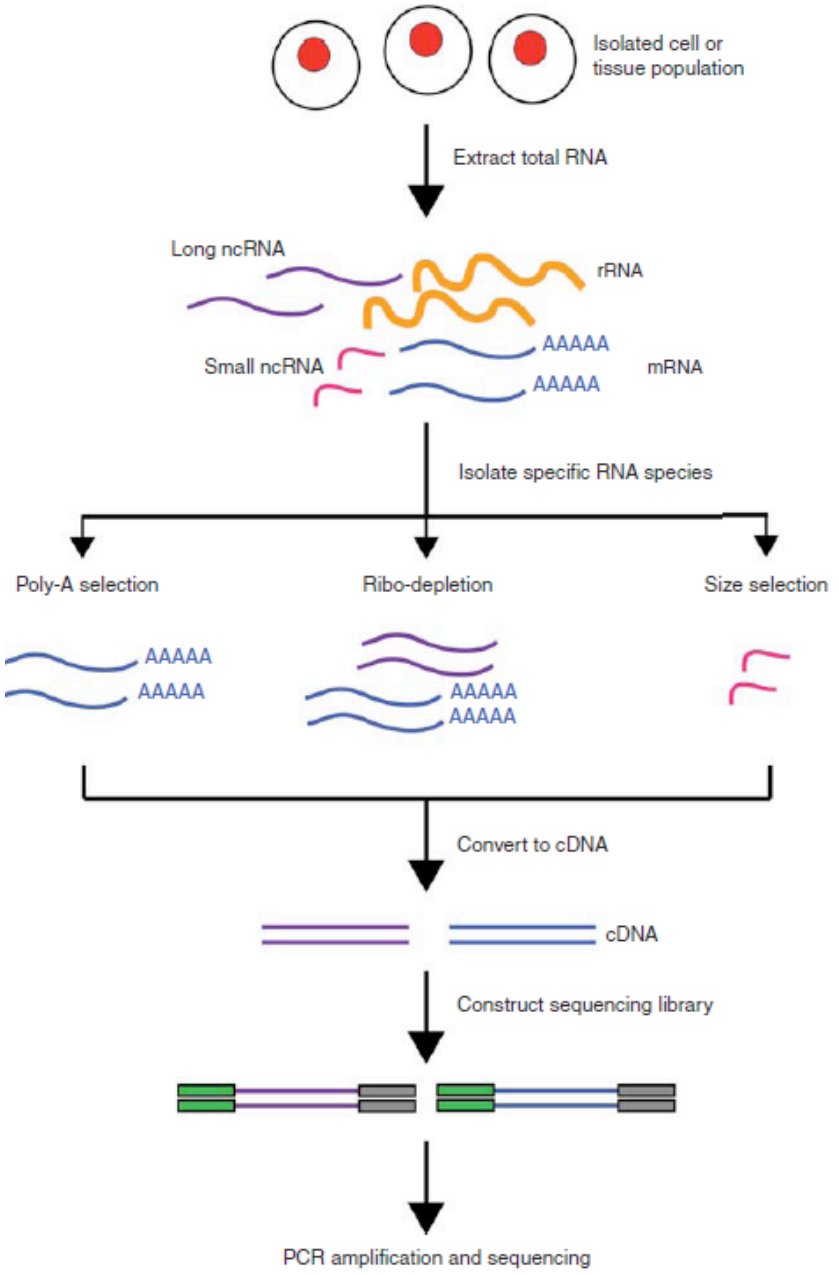
Experimental Design

Experimental Design

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

RNA Fraction

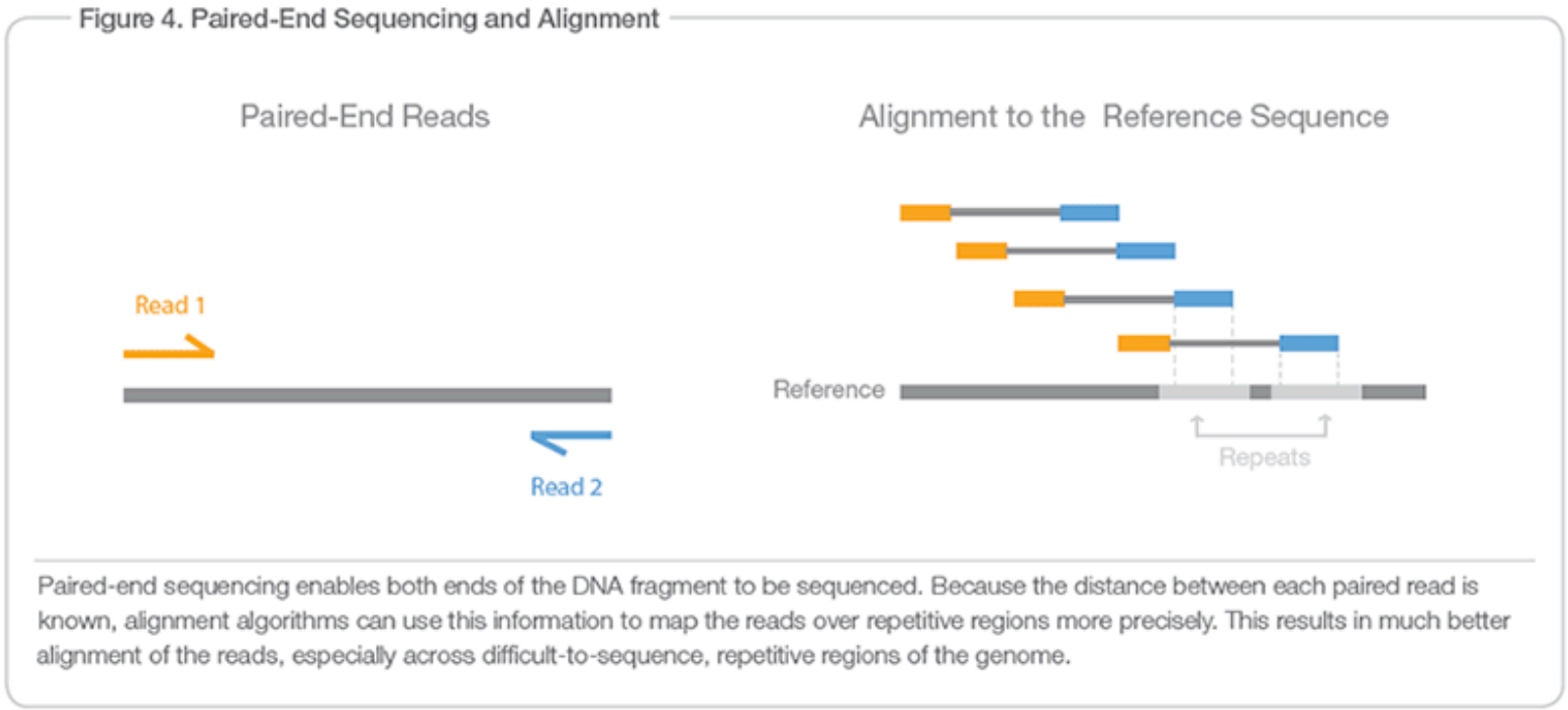
- **PolyA selection** - Oligo-dT beads capture polyA tails
  - ➡ targets mRNAs
- **Ribosomal RNA depletion** - standard kit removes cytoplasmic (nuclear-encoded) rRNAs; 'gold' kit removes both mitochondrial and cytoplasmic rRNAs; special 'blood' kit also removes global mRNA
  - ➡ targets mRNA and long 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

- **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



https://www.illumina.com/technology/next-generation-sequencing/paired-end-sequencing_assay.html

Sequencing Length

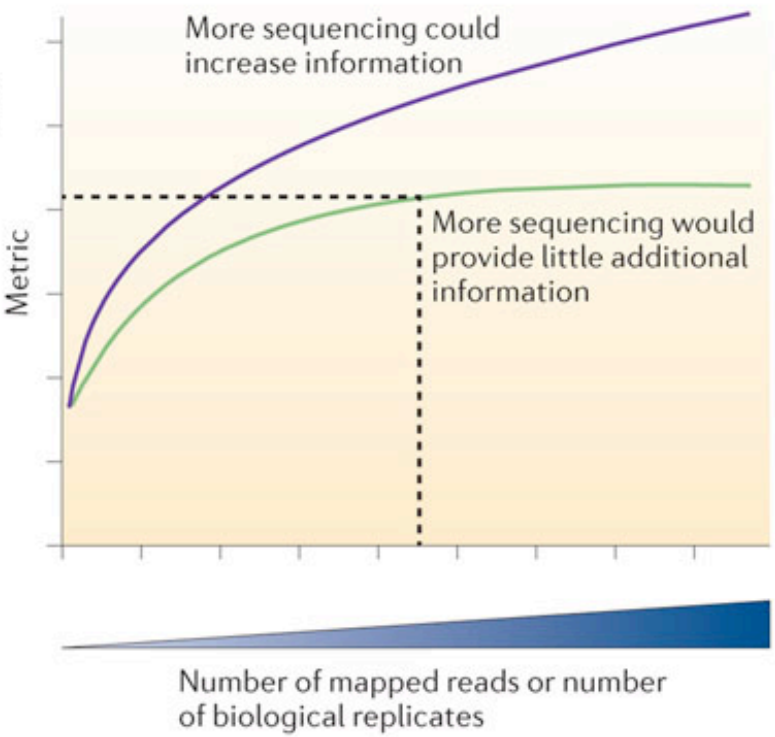
- Shorter reads - less than 100 bp
    - Cheaper
- Longer reads - greater than 100 bp
    - More expensive
    - Improved mappability
    - Improved accuracy of transcript identification

Sequencing Depth

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

Possible metrics:

- General transcriptome coverage: percentage of genes covered over 90% at a given expression level
- Differential expression: number of differentially expressed genes
- Alternative isoform detection: percentage of split reads (that is, junction that spans reads)
- ChIP–seq peak detection: number of enriched loci



Nature Reviews | Genetics

"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. http://doi.org/10.1038/nrg3642

Number of Replicates

Number of replicates needed are dependent on the following that vary from gene to gene:

- Within group variability
- Read coverage
- Desired 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. http://doi.org/10.1186/s13059-016-0881-8

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

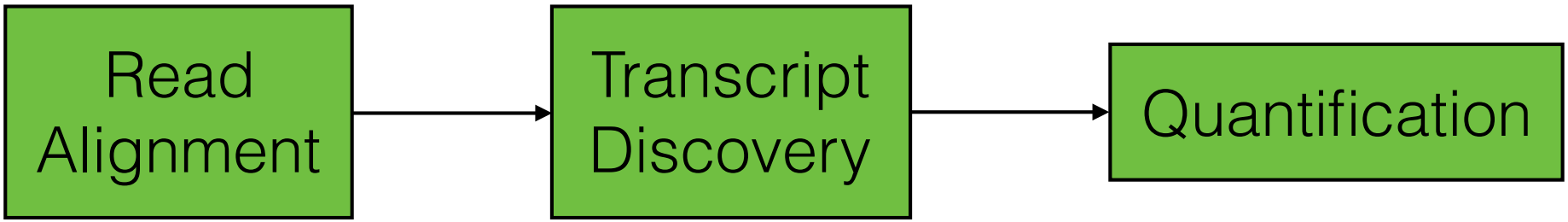
- Library preparation tends to generate the greatest batch effects
- Randomization at library preparation (i.e., spreading samples within a group across multiple batches) helps to:
  - avoid confounding between technical effects (e.g., batch effects) and effects of interest
  - minimizes effects of a ‘bad batch’

Randomization At Sequencing Run

- Actual sequencing 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

TRANSCRIPTOME PROFILING

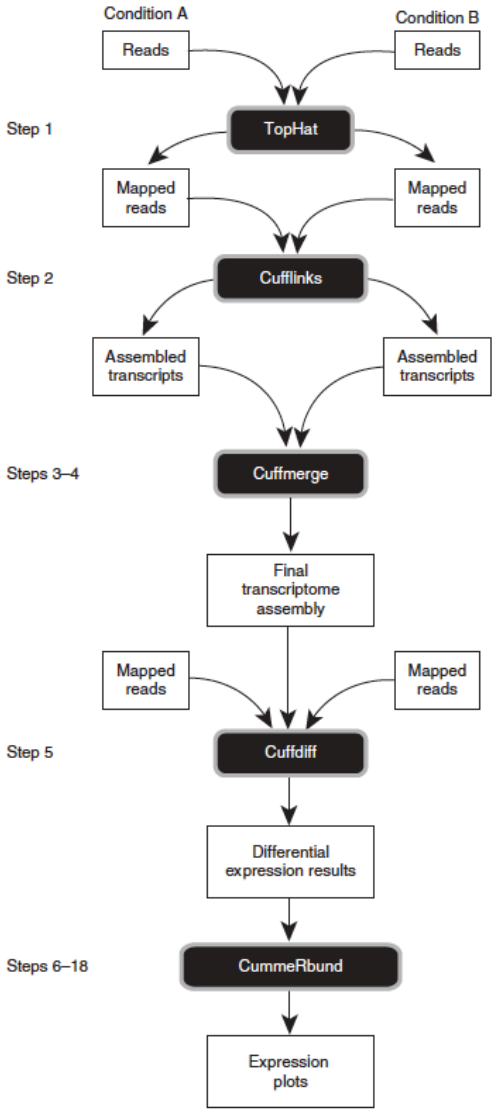
Transcriptome Profiling



Popular Strategies/Pipelines

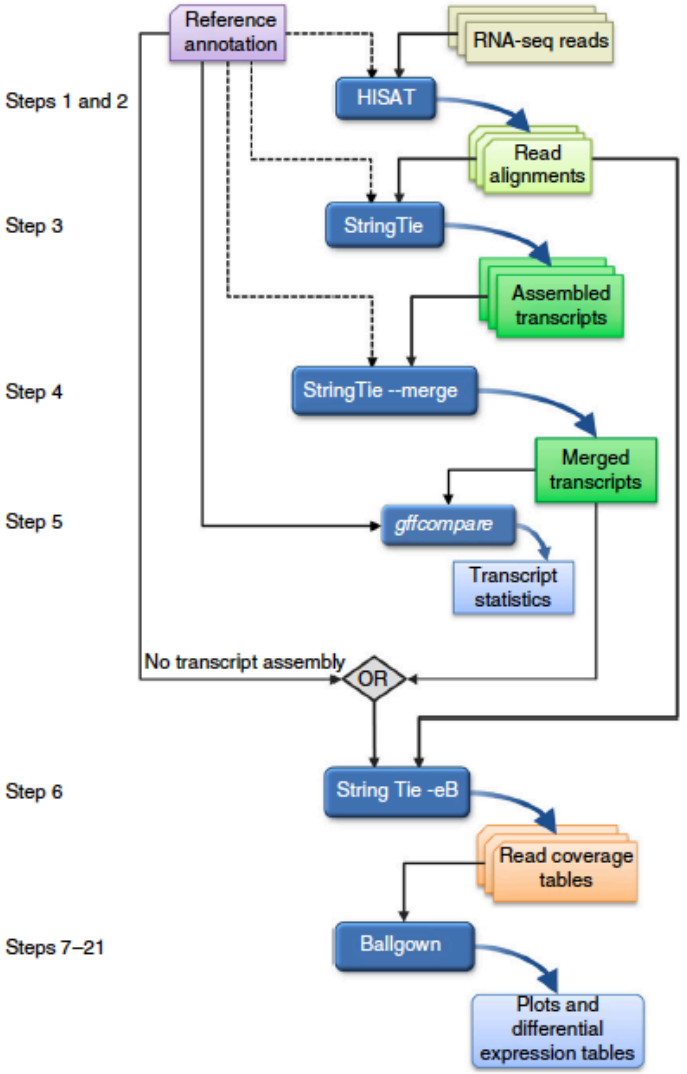
Most Popular Tuxedo Pipeline

Trapnell, C., et al. (2012). Nature Protocols, 7(3), 562–578.



Most Recent Tuxedo Pipeline

Pertea, M., et al (2016). Nature Protocols, 11(9), 1650–1667.



TRANSCRIPTOME PROFILING

READ ALIGNMENT

Raw Reads

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

```
Riken-M-WT-5-brain-total-RNA-cDNA_CGGCTATG_L003_R1_001.fastq
Riken-M-WT-5-brain-total-RNA-cDNA_CGGCTATG_L003_R2_001.fastq
```

fastq files have 4 rows per read:

Row 1: Header - includes information to uniquely identify read
Row 2: Sequence - actual nucleotide sequence read
Row 3: Spacer - ??
Row 4: Quality - quality metric for each individual base call

```
@GWZHISEQ02:301:C97NKANXX:3:1101:1486:1895 1:N:0:CGGCTATG
NTCCGGTCTGAACTCAGATCACGTAGGACTTTAATCGTTGAACAAACGAACCATTAATAGCNTCTGCACCATTGGGATGTCCTGATCCAACAGATCGGAAG
+
#<<BB/BFFFFFFFFFFFFFFBFFFFFF<BBBF/FB<BFFFFFFFFFFFFF/<<FF<<<F#/<<<FFFFFF/FF<FBF/</F/FBFFFFFFBF/F/7BF/
@GWZHISEQ02:301:C97NKANXX:3:1101:1705:1878 1:N:0:CGGCTATG
NACGCTTCTTCTCGCTCTTTCATTTTTAGCTTCTCTTTGAGTTTCTCTGCTTTCAACTGCTGCGTGCAGTCATCGAAGAGTTTTTGGTTCATCTCCATGAA
+
#<</<FFF<FFFF/<<FF/FFF<BF//<BBF/</<B//<F/<BFF/B<<</////<FB<///<<B/7///</<<//</F<<<B/FB/<<FFFFFFFF//<F
@GWZHISEQ02:301:C97NKANXX:3:1101:1704:1906 1:N:0:CGGCTATG
NTTAGATGAATCCCAGAGTATCATTCTTGATAAGACTCCCGGACTGACAGGTTTATTTTGGCCTCTAAGAGCAGGAAATCCAAGGTTAAATATTTCCCTGA
+
#<<BBFBFFFFFFFFFFFFFFFFFFFFF<FFFFFFFFFFFFFFFFFFFFFFBFFFFFFFFFFFF<FFFFFFFFFFFBFFFFFFFFFF<FFF<FFFF
```

Initial Quality Control

- Count number of raw reads and the average length of each read in each file:
    - Needed later for total number of sequenced reads
    - Double checked that paired files have the same number of reads
    - Double check that average length is the requested read length
- Can use programs like FASTQC to examine quality of reads

Trimming

- cutadapt - command line program (based in Python) 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)

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

Read Alignment

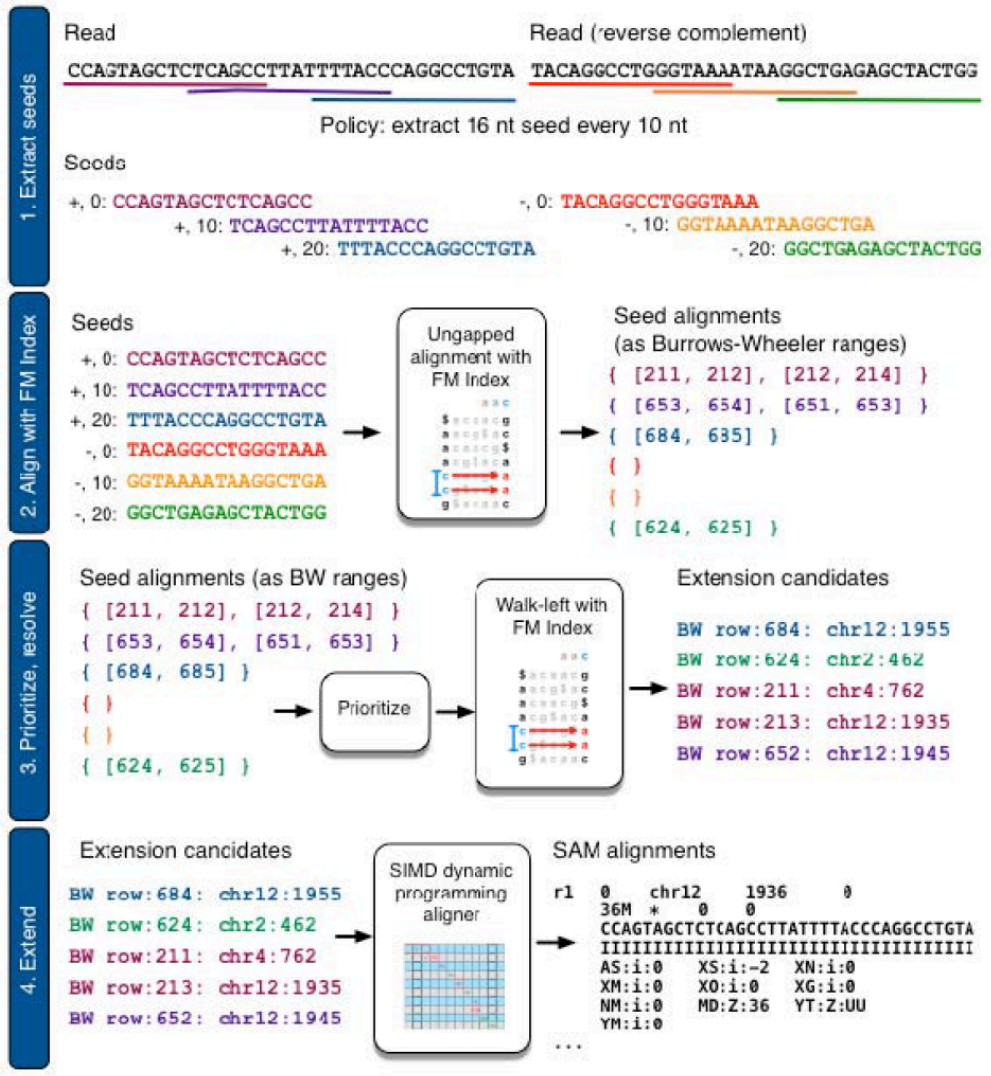
Initial Choices

- 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

Bowtie2

- Bowtie2 - short read alignment tool (command line based)
- Allows for mismatches, gaps, and soft clipping but doesn't identify exon junctions
- Used for alignment to transcriptome and for short RNA-Seq
- Included in other programs such as TopHat2 and RSEM



http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods, 9(4), 357–359.

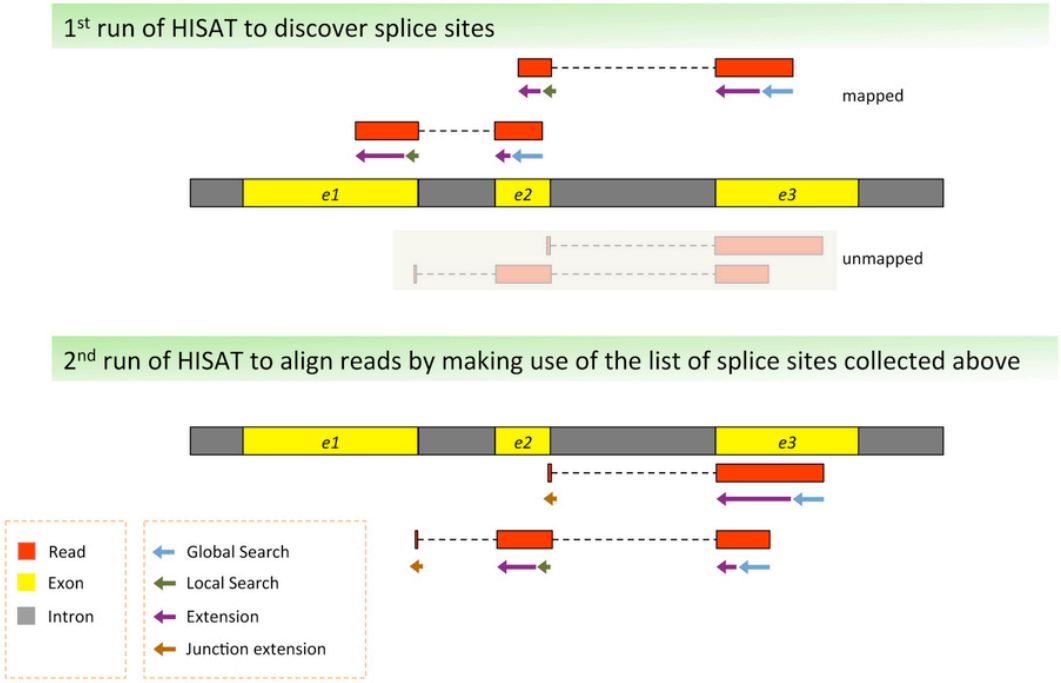
Read Alignment

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

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

SAM and BAM Files

- Output of most alignment 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

```
GWZHISEQ02:301:C97NKANXX:3:1101:1629:190683  6   1414042 60  100M1S  =   1414027 -116
AAAAGAAGAAAAGCAACAGCAAGCGAATCTGAAAAATAGGCAGAAGAGCCTGAAGGAGGAAGAACGGGAAAGGCGTGACATTGGGCTGAAGAATGCGCTGN
   FFFBFFFB/FFBFFFF//BF<FF<FFBBFFFFF/FBB<FFFBBFFFFFFBFFBFFFFFFFF</
BFFFFFBFFFFFFFF<FFFFFFFFFFBFFFFFFB<<<#   AS:i:-1 XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:100
YS:i:0   YT:Z:CP  XS:A:+   NH:i:1
GWZHISEQ02:301:C97NKANXX:3:1101:1629:1906163 6   1414027 60  101M    =   1414042 116
CCGAGAAGCCCGGCGAAAAGAAGAAAAGCAACAGCAAGCGAATCTGAAAAATAGGCAGAAGAGCCTGAAGGAGGAAGAACGGGAAAGGCGTGACATTGGGC

BBBBBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFBFBFFFFFFFFFFF<FFFFFFFFFFBFFFFFFFFFFFFFFFFFFFFFFFFFBFF
   AS:i:0   XN:i:0   XM:i:0   XO:i:0   XG:i:0  NM:i:0  MD:Z:101    YS:i:-1 YT:Z:CP XS:A:+  NH:i:1
```

TRANSCRIPTOME PROFILING

TRANSCRIPT DISCOVERY

Transcript Discovery

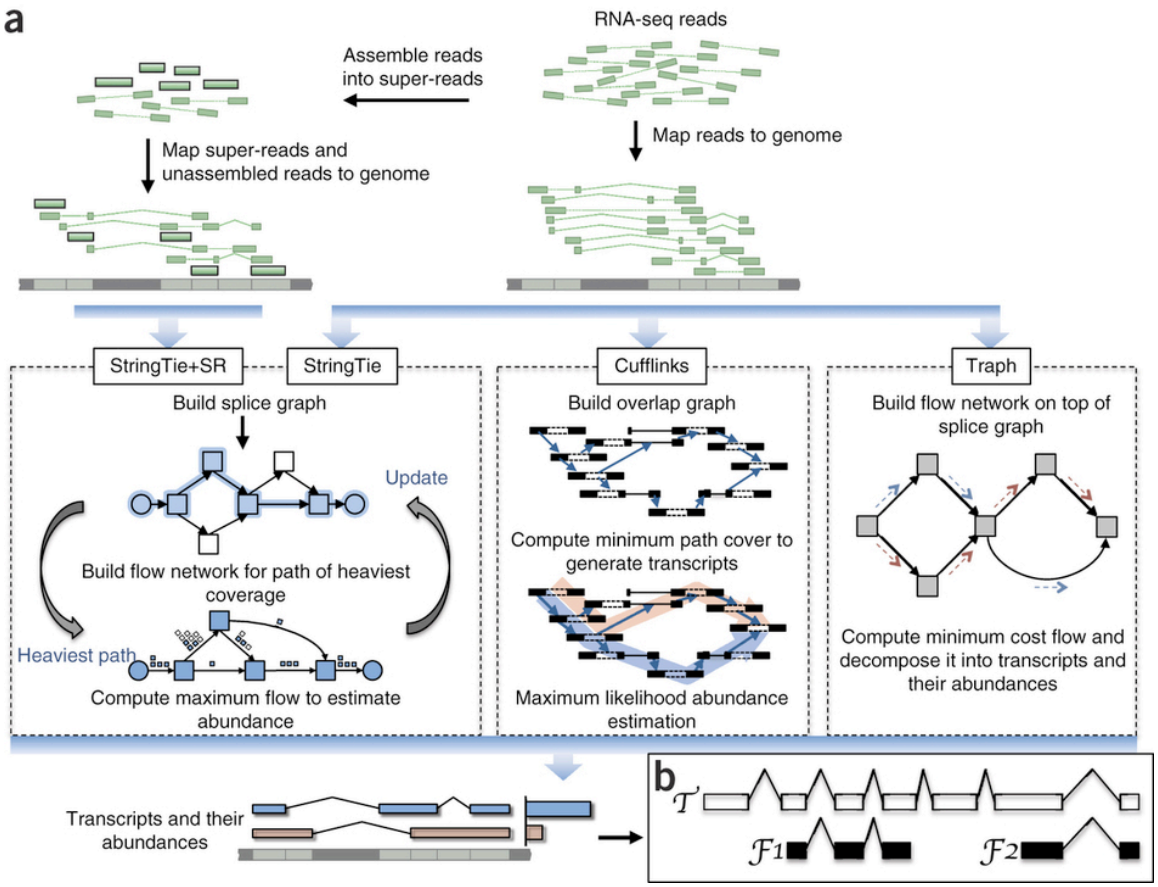
When and why do we need this?

- **In humans**
  - 20,441 coding genes
  - 22,219 non-coding genes
  - 198,002 gene transcripts
- **In rats**
  - 22,263 coding genes
  - 8,879 non-coding genes
  - 40,459 gene transcripts

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.

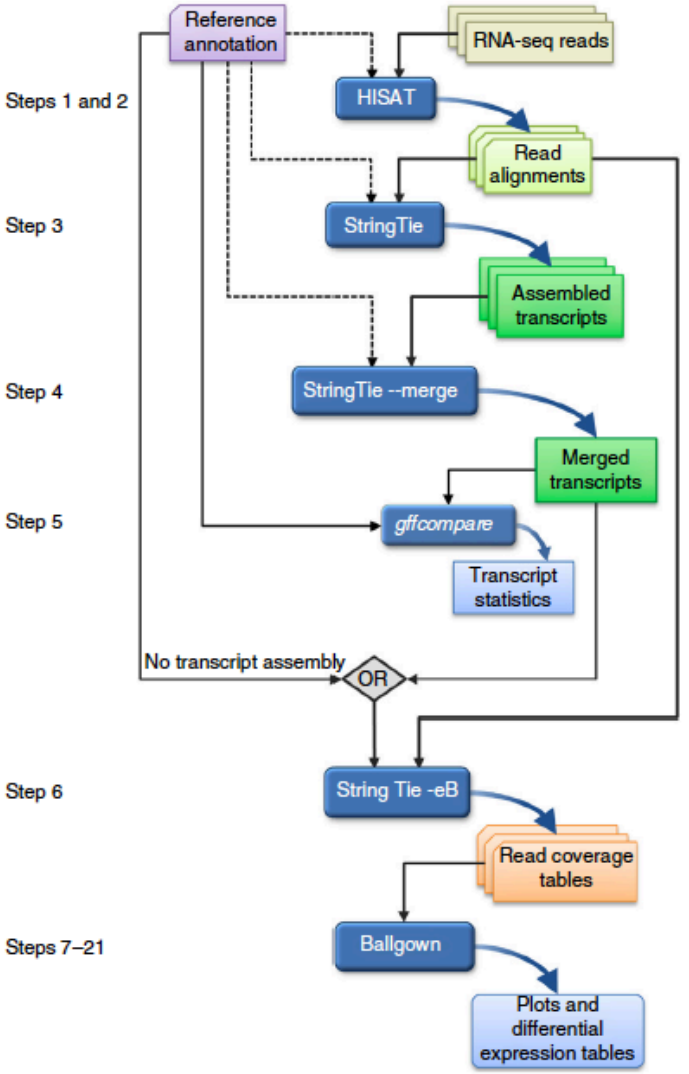
Transcript Discovery

StringTie Pipeline

1. Create sample-specific transcriptome assemblies

2. Merge sample-specific transcriptome assemblies

3. Quantify transcripts in merged transcriptome assembly



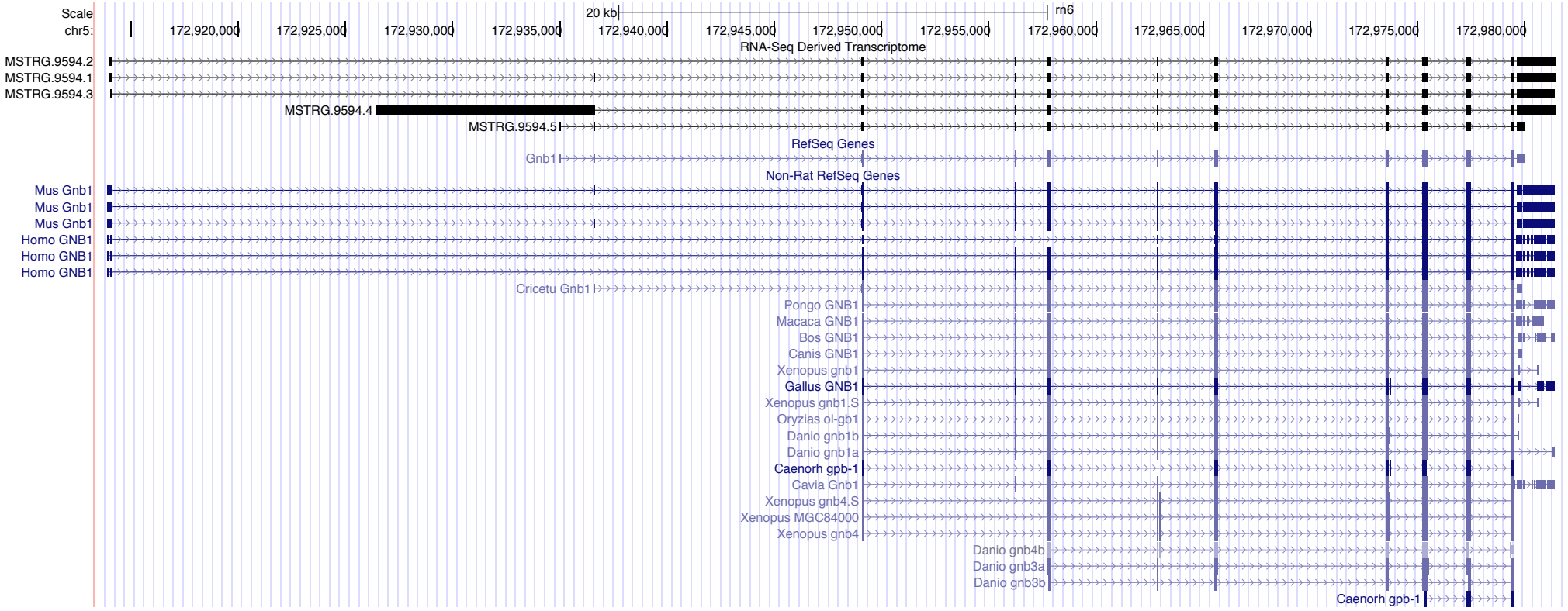
Transcript Discovery
Quality Control

- Distribution of novel vs. annotated transcripts
- Top expressed transcripts
- Non-coding vs protein-coding transcripts

Transcript 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.



Transcript 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

TRANSCRIPTOME PROFILING

QUANTIFICATION

Quantification
Initial Options

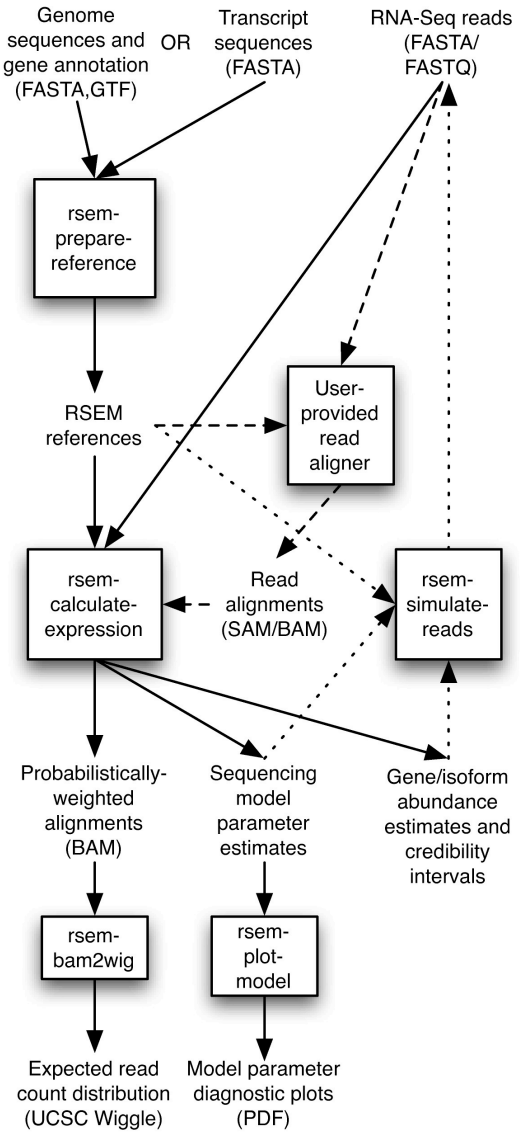
- **Gene-level analysis** - reads are counted that align to any isoform of the gene
    - Used with a shallow read depth per sample
    - Used when interest is in identifying key pathways, not necessarily key players in the pathway
- **Isoform-level analysis** - read counts for individual isoforms (i.e., 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 (Teng, M., et al. *Genome Biology*, *17*(1), 74.).



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

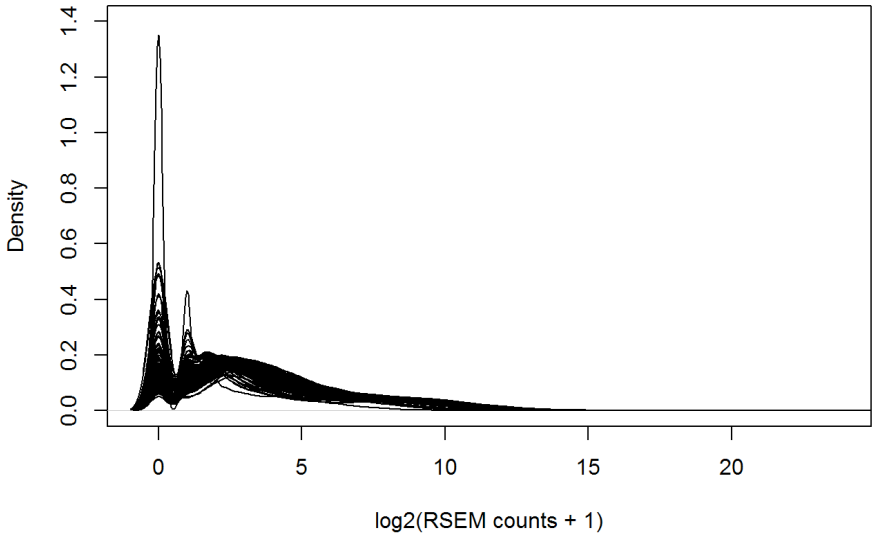
Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, *12*(1), 323.

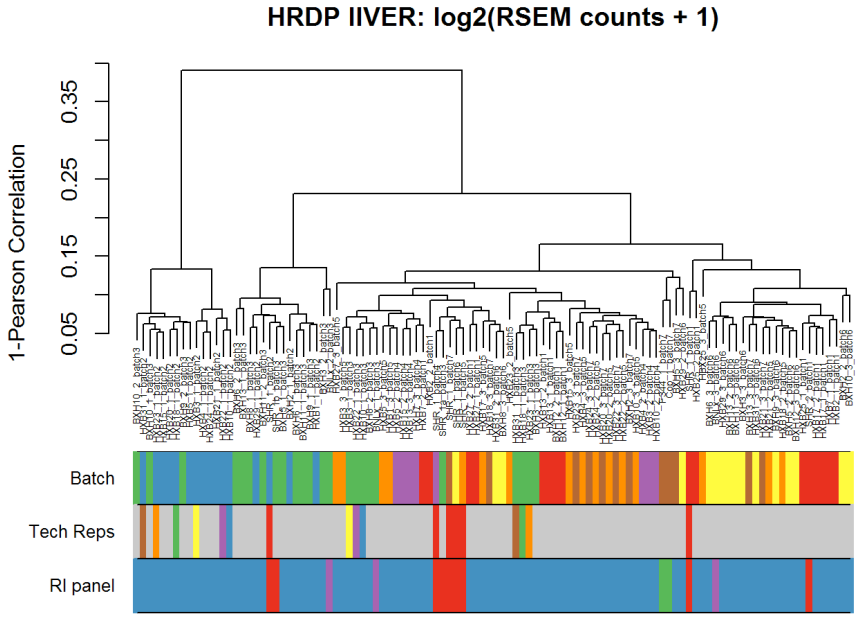
Quantification
Units

- ***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 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





Differential Expression

Differential Expression

- Preprocessing
- Differential expression
- 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

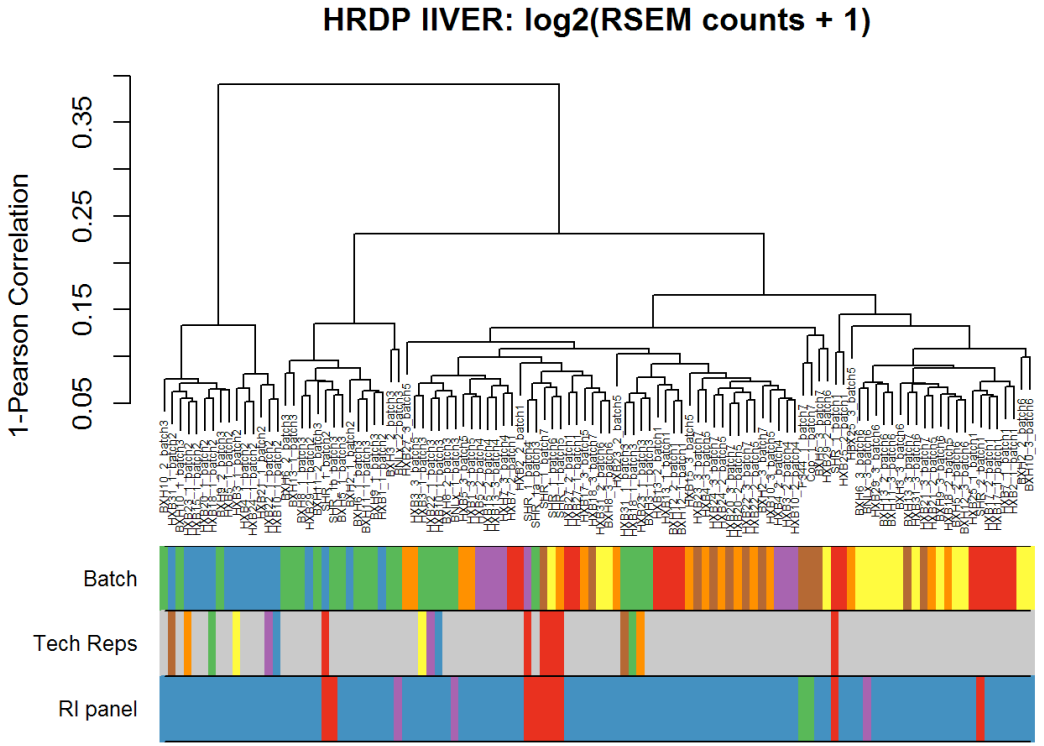
- Often, low counts genes/isoforms are removed prior to differential expression analysis with the assumption that their expression estimates are ‘below background’.

- 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.

Differential Expression

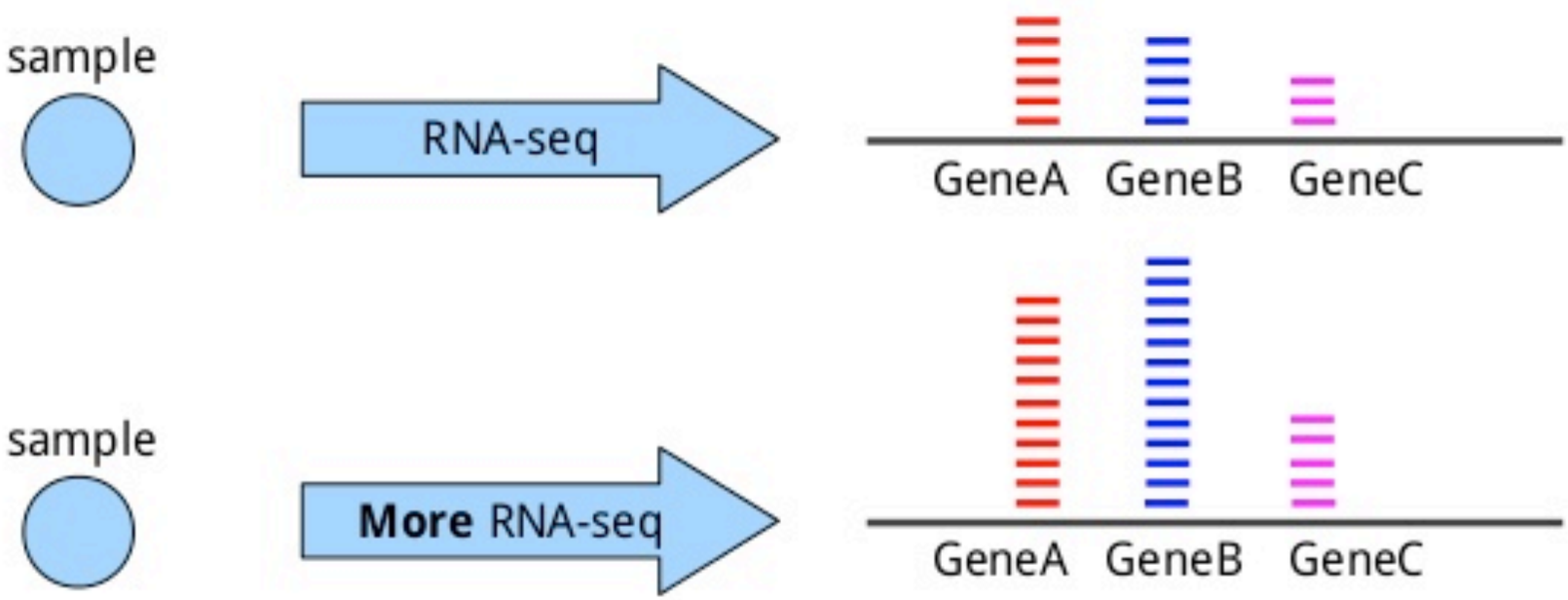
Preprocessing - Visualizations

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



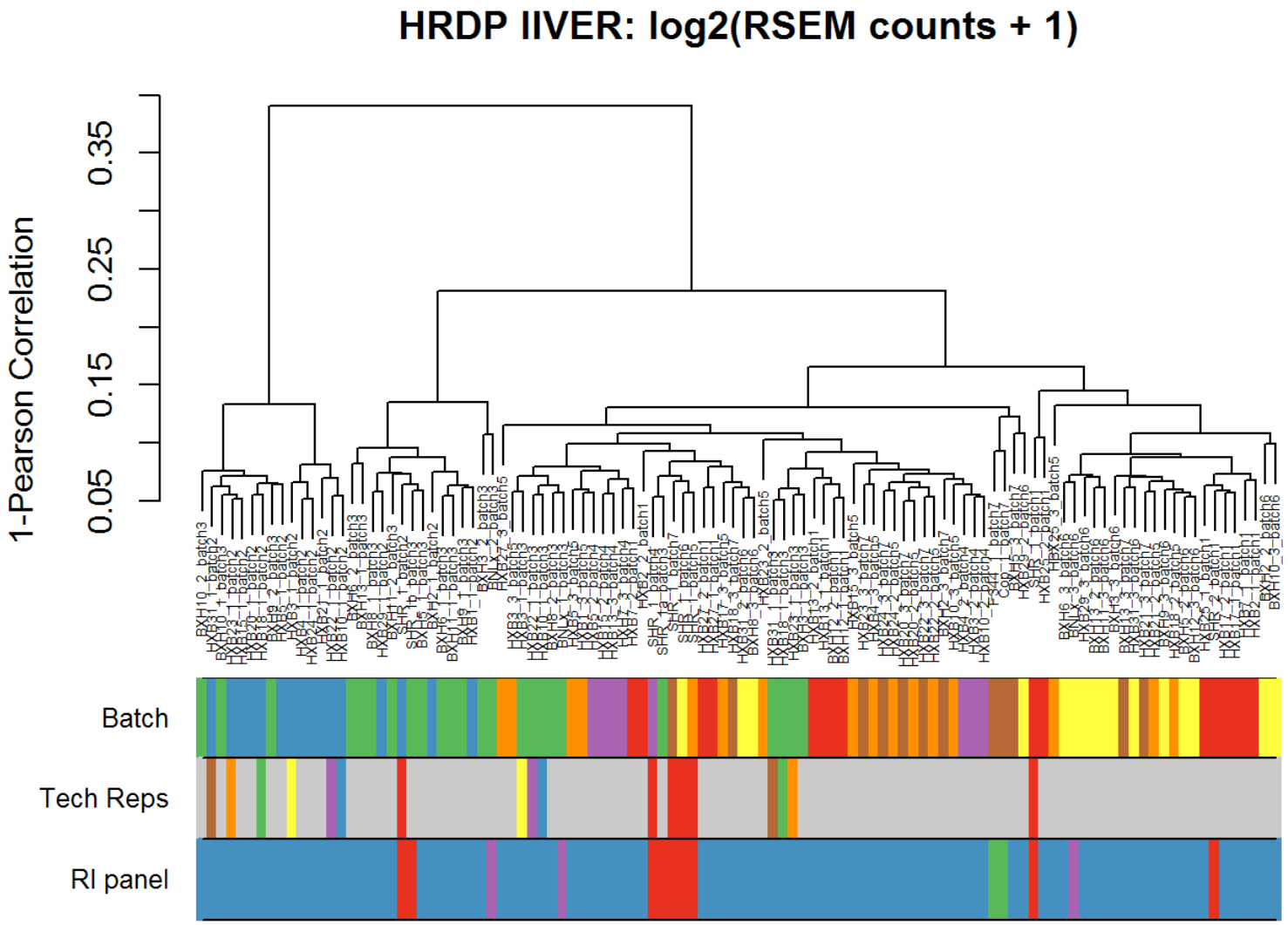
Differential Expression

Preprocessing - Library Size Bias



Differential Expression

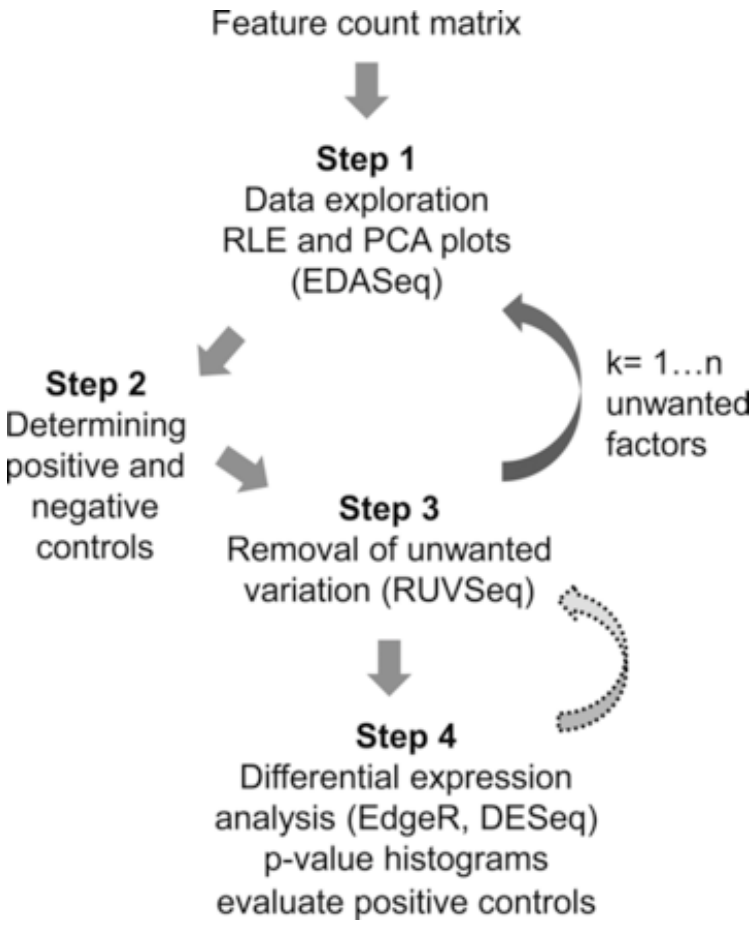
Preprocessing - Batch Effects



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
- 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 Distribution

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

- Probability of having seen Office Space is 20%

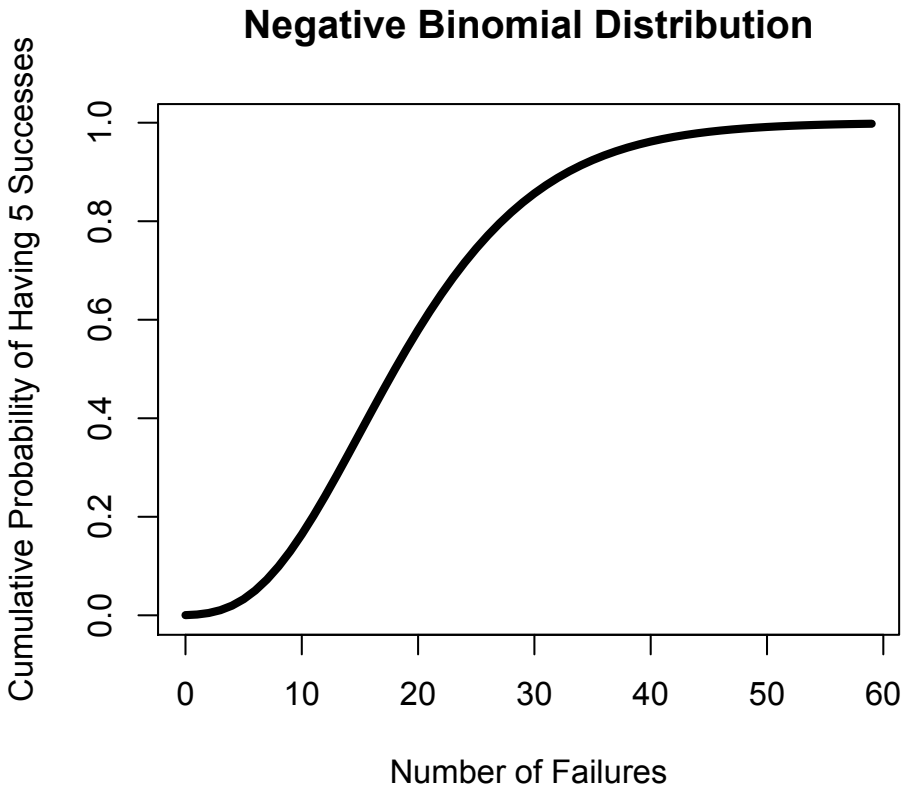
How many people will I have to ask before I find 5 people?

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 expressed in different "proportions" in to populations



Differential Expression
Comparison of Software

- Many software/methods available
- No clear winner
- DESeq2 is one of the most popular

Differential Expression
DESeq2

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

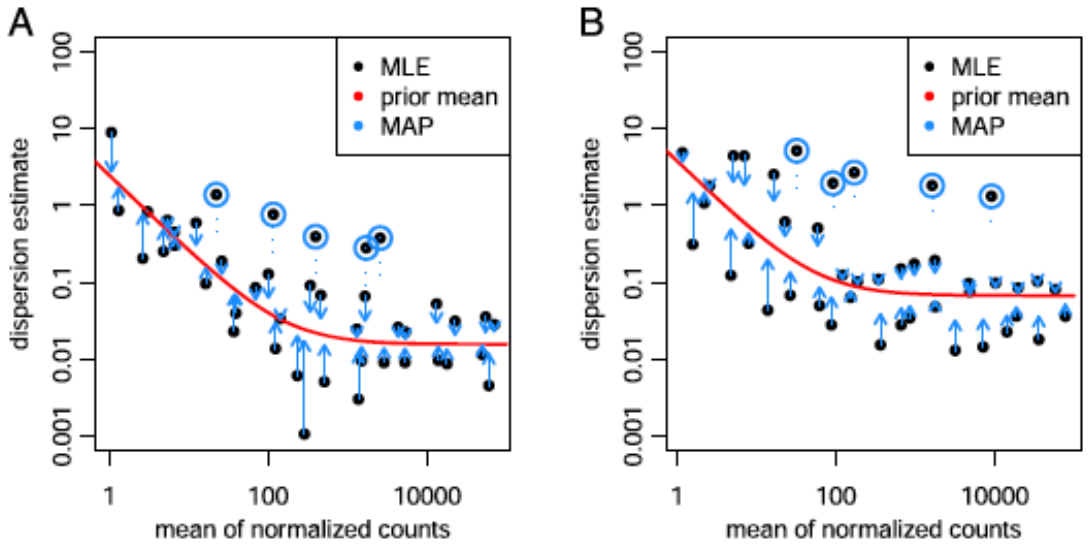


Figure 1 Shrinkage estimation of dispersion. Plot of dispersion estimates over the average expression strength **(A)** for the Bottomly *et al.* [16] dataset with six samples across two groups and **(B)** for five samples from the Pickrell *et al.* [17] dataset, fitting only an intercept term. First, gene-wise MLEs are obtained using only the respective gene's data (black dots). Then, a curve (red) is fit to the MLEs to capture the overall trend of dispersion-mean dependence. This fit is used as a prior mean for a second estimation round, which results in the final MAP estimates of dispersion (arrow heads). This can be understood as a shrinkage (along the blue arrows) of the noisy gene-wise estimates toward the consensus represented by the red line. The black points circled in blue are detected as dispersion outliers and not shrunk toward the prior (shrinkage would follow the dotted line). For clarity, only a subset of genes is shown, which is enriched for dispersion outliers. Additional file 1: Figure S1 displays the same data but with dispersions of all genes shown. MAP, maximum *a posteriori*; MLE, maximum-likelihood estimate.

Differential Expression
DESeq2

- 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 Logs

- 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 potentially be used in mixed models

Differential Expression

False Discovery Rate

- Many times for genetic studies, we use a false discovery rate (FDR) rather than a traditional p-value to help account for multiple comparisons.
- FDR is the proportion of “significant” tests that are false positives.
- 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).
- When we use a 5% FDR threshold for significance, we are estimating that 5% of the significant genes 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

Functional Enrichment

Functional Enrichment

- Overrepresentation of Pathways/Ontologies
- Functions of unannotated transcripts

Overrepresentation of Functions

Background

- Differential expression analyses often results in a long list of ‘candidate’ genes
- Finding common functions/pathways among these genes often helps interpret the observed biological process
- Using known functional annotation, we can identify pathways/functions *enriched* among genes that are differentially expressed in our analysis

Overrepresentation of Functions

EnrichR

- EnrichR is a web-based tool that examines enrichment in *many* databases (e.g., KEGG pathways, GO terms, OMIM disease genes)

- It is very simple to use, but does not allow the user to define a background set of genes.

    - Some pathways are enriched due to tissue type rather than due to treatment factor

- Does not take into account biases in RNA-Seq (e.g., gene length)

Overrepresentation of Functions

GOSeq

- goseq is an R package that also takes into account both transcript length and sequencing depth

Young *et al. Genome Biology* 2010, **11**:R14
http://genomebiology.com/2010/11/2/R14

METHOD Open Access

Gene ontology analysis for RNA-seq: accounting for selection bias

Matthew D Young, Matthew J Wakefield, Gordon K Smyth and Alicia Oshlack*

Unannotated Transcripts
Blast2GO

- Blast2Go is a software that links sequences of unannotated transcripts to Gene Ontology (GO) terms.
- Blast2GO Pro also includes software for Gene Set Enrichment Analysis