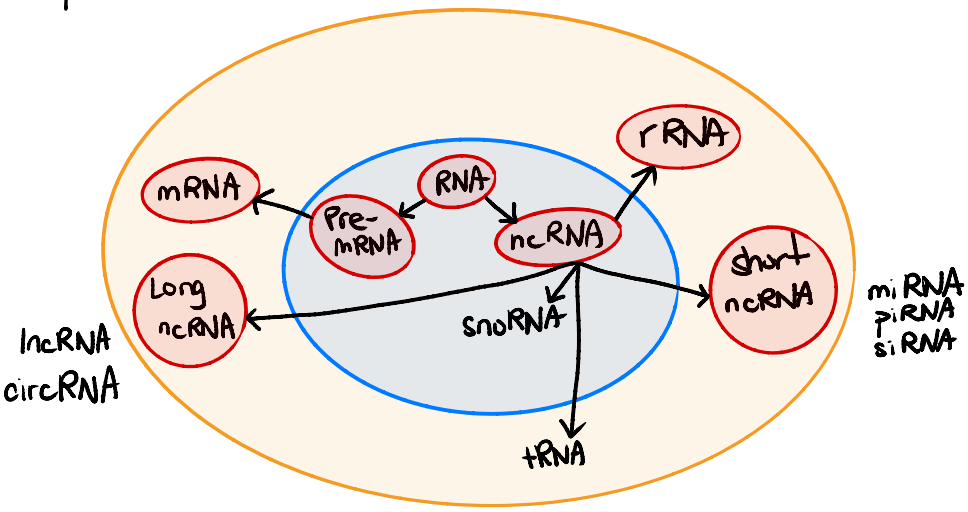
NOTES

Lecture 6: RNA-seq

Transcriptome

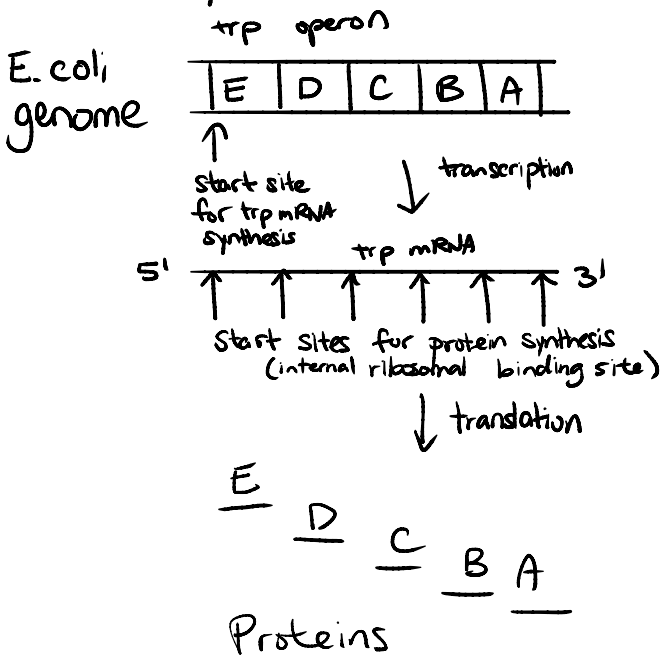
- All RNA molecules, including mRNA, rRNA, tRNA and other non-coding RNA produced in one or a population of cells

Types of RNA in the cell



RNA Transcription Principles

Prokaryotes



*tend not to have intronic structures

Eukaryotes

Double stranded genomic DNA template

↓ transcription and polyadenylation

Single-stranded pre-mRNA

5' CAP —— 3'SS —— 5'SS —AAA

↓ RNA processing (splicing)

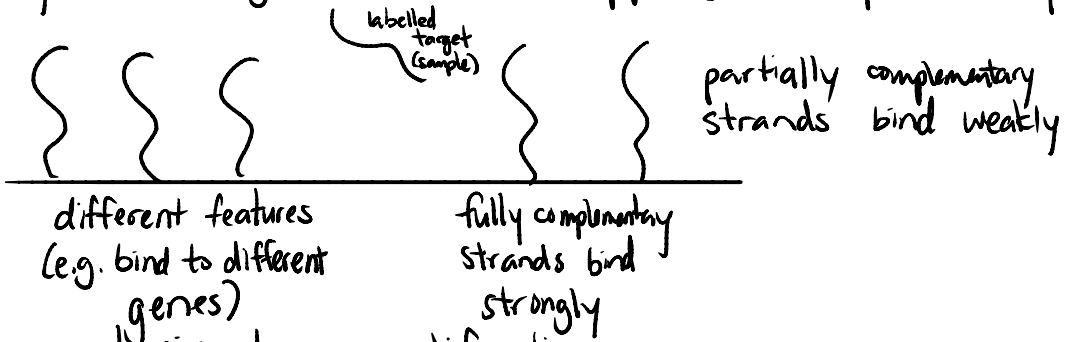
Mature RNA

↓ export and translation

protein

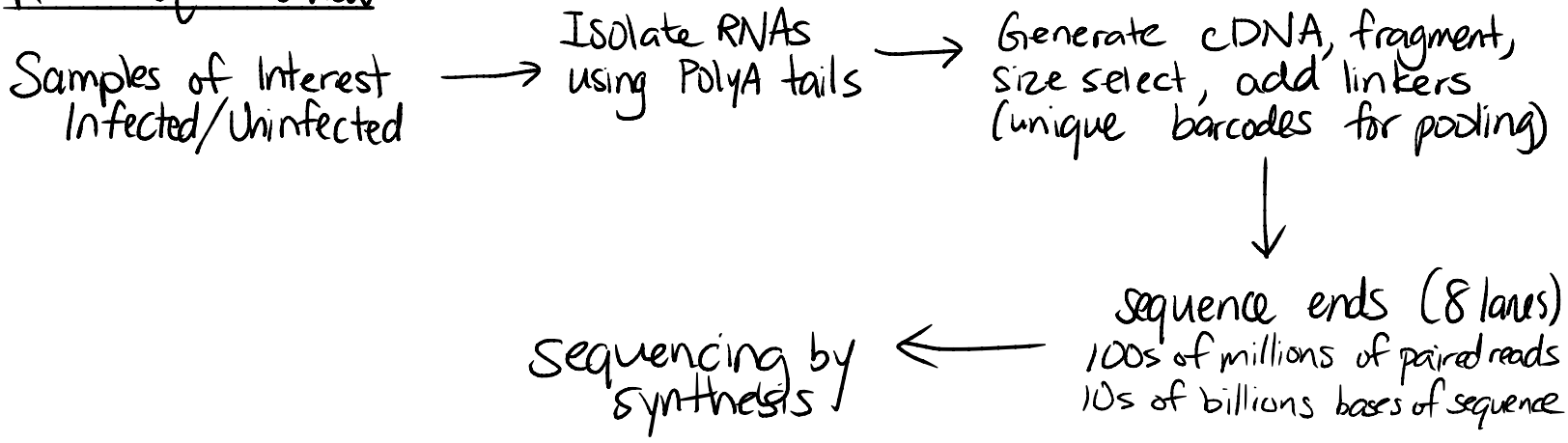
Techniques for Measuring Gene Expression

1. Quantitative PCR
   - low throughput (1-10 genes)
     - can do more, but noisy/expensive
   - design qPCR primer pairs
   - PCR from RNA or cDNA template, with fluorescent dye incorporated
   - the amount of intercalating dye fluorescence is proportional to number of DNA molecules
   - PCR cycle at which fluorescence enters exponential phase (i.e. the cycle threshold (Ct)), is correlated to the number of template DNA molecules
     - on fluorescence vs cycle plot, have housekeeping genes and gene of interest
2. Microarrays
   - 100s to 1000s of genes
   - DNA probes (for exons or gene regions) printed on a solid array
   - cDNA or RNA is fluorescently labeled and hybridized to the array
   - Array is imaged, a mask applied, and probes quantified

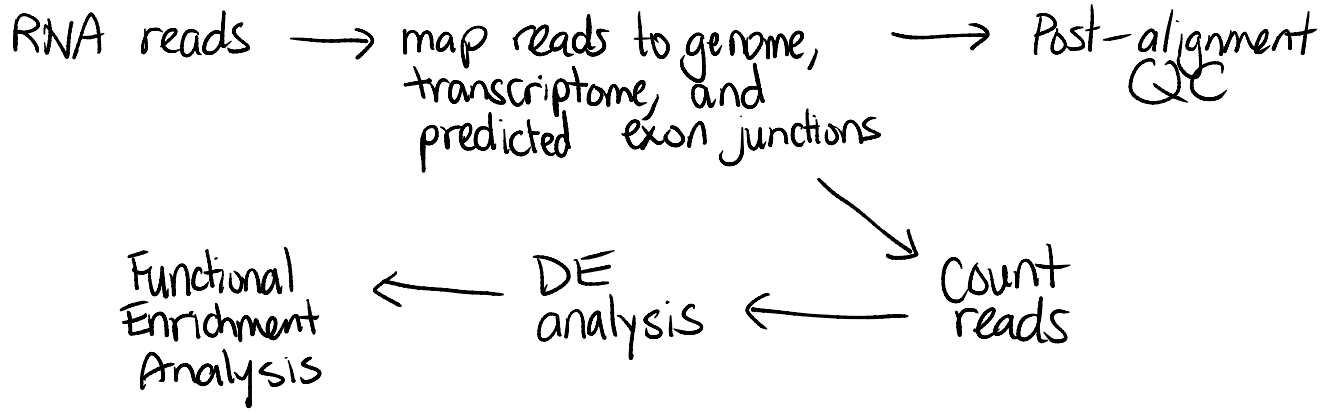


   - fluorescent signal = quantification
   - each DNA spot contains picomoles (10-12 pmol) of a specific DNA sequence
     - red: control
     - green: test
     - yellow: overall
3. Serial Analysis of Gene Expression (SAGE)
   - 100s to 1000s of genes
   - any transcript with a known sequence can be identified by a short signature sequence or "tag" (uniquely identified)
   - SAGE tags are then cloned into cloning vector before Sanger sequencing
4. RNA-Seq

RNA-seq Overview



RNA-Seq Analysis Workflow



Why Sequence RNA (versus DNA)?

- Functional studies
  - understand gene expression changes in response to an experimental condition (whereas genome may be constant)
- some molecular features can only be observed at the RNA level
  - alternative isoforms
  - fusion transcripts
  - RNA editing
- Predicting transcript sequence from genome sequence is difficult
  - alternative splicing
  - RNA editing
- interpret mutations that do not have an obvious effect on protein sequence
  - "regulatory" mutations that affect what mRNA isoform is expressed and how much

- prioritize protein coding somatic mutations (often heterozygous)
  - if the gene is not expressed, a mutation in that gene would be less interesting
  - if the gene is only expressed from the wild type allele, this might suggest loss-of-function (haploinsufficiency)
  - if the mutant allele is expressed, this could be candidate drug target

RNA-Seq Experimental Design

- number of replicates
  - always set up experiment with more replicates than you think you need (since not every sample will pass QC)
  - minimum number of 3 biological replicates for statistical analysis
- library type
  - what kind of RNA?
- sequencing depth
- ribosomal RNA or globin (for blood samples) removal method
- ways to minimize batch/confounding effects

Good Experimental Design

1. Replication
   - measurements are usually subject to variation and uncertainty
   - replication allows us to better estimate the true effects of treatments, to further strengthen the experiment's reliability and validity
2. Randomization
   - assign individuals at random to groups or to different groups in an experiment to reduce bias
3. Blocking
   - Blocking reduces known but irrelevant sources of variation between treatments

EXAMPLE: Many different plants (replicates) are assigned randomly to one of two treatments (randomization) then grouped and placed in each field (blocking)

Sources of Variation

1. Biological variation
   - intrinsic to all organisms
   - may be influenced by genetic or environmental factors
2. Technical variation
   - variability in measurements (ie. the uncertainty in the abundance of each gene in each sample that is estimated by the sequencing technology)
   - rRNA / library prep method / extraction
   - variations between flow cells / lanes within the same flow cell

Batch Effects

- Sources of variation that are "unrelated" to the biological or scientific variables in a study
- technical variabilities that potentially contribute to batch effects:
   - different personnel / lab
   - different experimental / sample processing dates
   - different sample processing methods/reagents/equipment

Experimental Design Summary

- replication, randomization and blocking are essential
- the best way to ensure reproducibility and accuracy of results is to include independent biological replicates (technical replicates are not appropriate substitutes)
- differential expression results from unreplicated data cannot be generalized beyond the one sample tested
- recognize potential confounding factors (e.g. lane, batch, and flow cell effects) in the design

Challenges

- sample
   - purity
   - quantity (abundance can vary widely)
   - quality (Good RNA vs Bad RNA)

- RNAs consist of small exons that may be separated by large introns
- Relative abundance of RNAs vary wildly
- RNAs come in a wide range of sizes
- RNA is fragile compared to DNA (and easily degraded)

Good vs. Bad RNA

- RNA quality assessed via bioanalyzer (uses ratio of rRNA band intensities
- output = RIN (RNA Integrity Number)
- Good RIN → 10
- Bad RIN → 0

Separating mRNA from other RNA

- in mammalian cells, ~80% of total RNA is rRNA
- rRNA is similarly abundant in bacteria
- need to separate mRNA from other types of RNA before library construction
- Eukaryotic mRNA has poly-A tail on 3' end, therefore poly-A selection is performed
- Prokaryotic mRNA doesn't have a poly-A tail, so ribosomal depletion is performed

mRNA RNA-Seq Library Construction

1. PolyA+ RNA captured (using polydT bead)
2. RNA fragmented and primed
3. First strand cDNA synthesized
4. Second strand cDNA synthesized
5. 3' ends adenylated and 5' ends repaired
6. DNA sequencing adapters ligated (e.g. index)
7. Ligated fragments PCR amplified (5–10 cycles)

Stranded vs Unstranded Library Prep

- Unstranded Protocol
  - Synthesis of randomly primed double-stranded cDNA followed by the addition of adapters for NGS
  - loss of information on which strand the original mRNA template is coming from
  - cannot accurately determine gene expression from overlapping genes (i.e. genes with partially overlapping genomic coordinates, but are transcribed from opposite strand
- Stranded Protocol
  - many different methods
  - dUTP second-strand marking is recommended
  - uses dUTPs instead of dTTPs during the synthesis of the second strand in cDNA synthesis of sequencing library step
  - prior to PCR, the second strand with uracils is degraded using uracil-N-glycosylase

RNA-Seq Analysis

- gene expression and differential expression
- transcript discovery or annotation
- allele specific expression (and in relation to SNPs or mutations)
- mutation discovery
- fusion detection
- RNA editing

PCRdup Removal

- Concerns
  - duplicates may correspond to biased PCR amplification of particular fragments
  - For highly expressed, short genes, duplicates are expected even if there is no amplification bias
  - removing them may reduce the dynamic range of expression estimates
- if removed, assess duplicates at the level of paired-end reads (fragments) and not single end reads

RNA-Seq Library Depth

- factors
  - what is your question? (gene expression changes, alternative splicing? Mutation calling?)
    - rare transcripts requires more depth
  - tissue type, RNA prep, RNA quality, library construction method (how well it went, etc.)
  - sequencing type: read length, paired/unpaired, etc.
  - computational approach with similar goals
- identify publications with similar goals
- pilot experiment

RNA-seq Library Quality

- compute metrics of quality and compare across libraries
- validate expression values (or other measure) on alternate platform

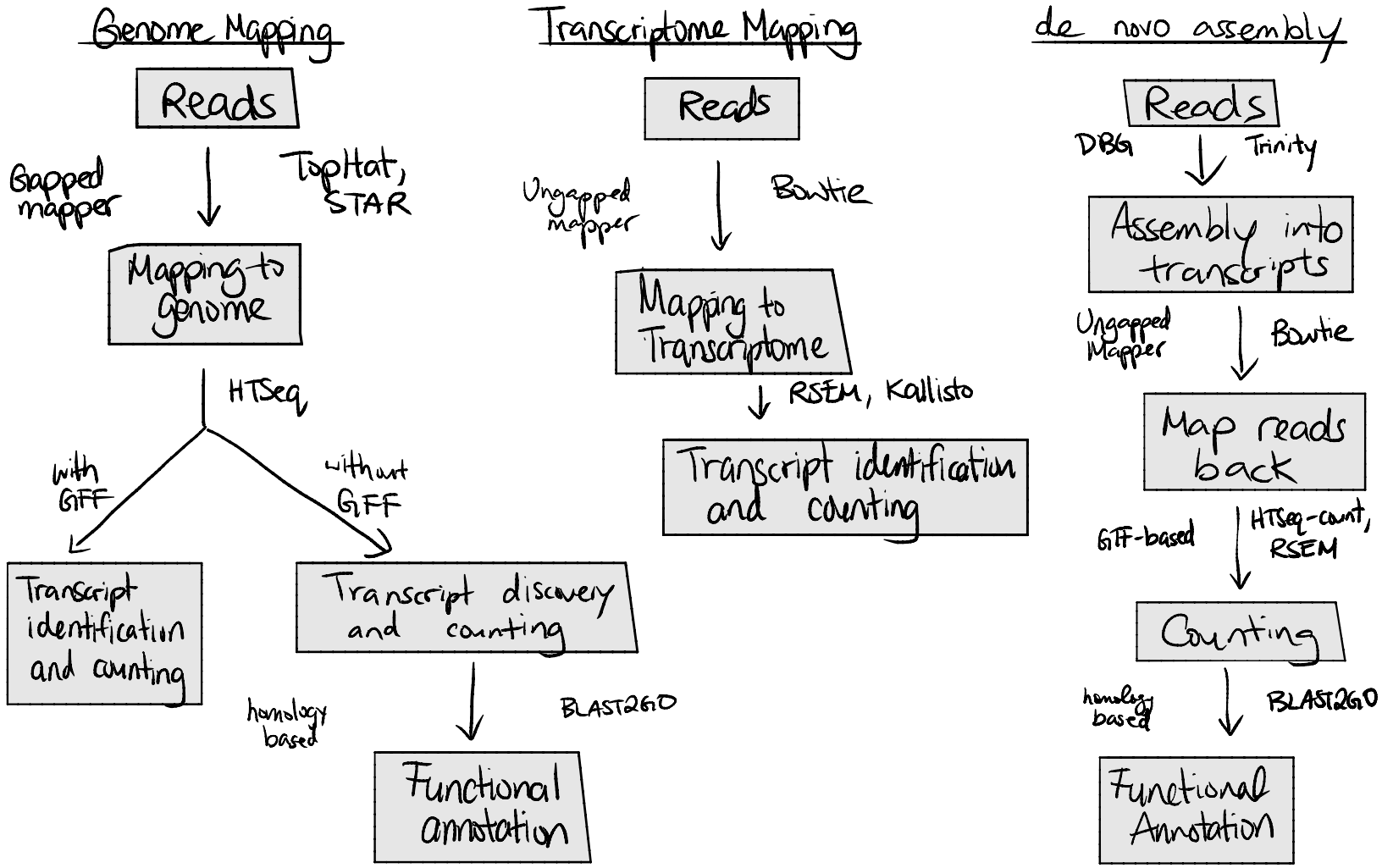
RNA-Seq vs Microarray

- Microarray uses fluorescence (detection limit) background noise at low levels, saturation at high levels
  - decent correlation for genes with medium level expression
  - poor correlation for genes with high/low levels of expression
- very different distributions of expression levels
  - RNA-seq normalized counts → Poisson distribution
    - no prior assumptions made
    - count what you see
  - microarray
    - assumed normal distribution

Lecture 7: RNA-Seq Alignment

What to Do With RNAseq Reads

- when reference genome is available:
  1. Mapping to reference genome
     - reads are aligned to reference genome with a gapped mapper
     - novel transcript discovery and quantification can proceed with or without an annotation file
  2. Mapping to reference transcriptome
     - reads aligned to the reference transcriptome using an ungapped aligner
     - transcript identification and quantification can occur simultaneously
- when a reference genome is not available
  - reads need to be assembled first into contigs or transcripts
  - for quantification, reads are mapped back to the novel reference transcriptome for further analysis or annotation



Approaches to Spliced Mapping

1. Exon-first approach (e.g. TOPHat)
   - exon read mapping
     - align reads with complete alignment
   - spliced read mapping
     - spliced reads chopped to bits and then realigned
2. Seed-extend approach
   - seed matching
   - seed extend

STAR (Spliced Transcript Alignment to a Reference) Aligner

- high accuracy
- outperforms other aligners by >50X in mapping speed, but is memory intensive

STAR Algorithm

1. Seed searching
   - for each read, STAR will search for the longest sequence that matches exactly one or more locations on the reference genome / (Maximal Mappable Prefixes — MMPs)
   - different parts of the reads that are mapped separately are "seeds"
   - STAR will search again for ONLY the unmapped portion of the read to find the next longest sequence that exactly matches the reference genome ("seed2")
   - sequential searching of only the unmapped portions of reads is why STAR is an efficient aligner
     - tophat has to align multiple times to find the splice site
   - STAR extends the previous MMPs / seeds to accomodate for mismatches
     - penalizing gaps end up selecting for pseudogenes (introns unspliced)
   - if extension does not give a good alignment, then poor quality or adapter sequence will be soft clipped

2. Clustering, Stitching, and Scoring
   - the separate seeds are stitched together to create a complete read by:
     - clustering the seeds together based on how close they are to a set of "anchor" seeds (uniquely mapped seeds)
     - seeds are stitched together based on the best alignment for the stitched read (using scores based on mismatches, indels, gaps, etc.

Running STAR

1. Create a genome index
   - genomes from ENSEMBL (better annotations) and NCBI
   - GENCODE annotations are recommended for human and mouse
   - Reference genomes
     - haploid representation of a species genome
     - human genome is a haploid mosaic derived from 13 volunteer donors from Buffalo, NY (maintained by the Reference Genome Consortium)
     - not perfect, contain gaps and patches
     - toplevel fasta file
       - contains all sequence regions flagged as toplevel in an Ensembl schema
       - includes chromosomes, regions NOT assembled into chromosomes, and N padded haplotype/patch regions
     - primary assembly fasta file
       - contains toplevel regions excluding haplotypes and patches
       - best used for performing sequence similarity searches where patch and haplotype sequences would confuse analysis
       - if primary assembly not present, there are no haplotype/patch regions, where toplevel is equivalent
   - for small (i.e. bacterial) genomes, use
     genomeSAindexNbases = $\frac{\log_2(\text{genome length})}{2} - 1$

2. Map reads to genome
    - output files
        - BAM file
        - log.out
            - main log file
            - detailed information
            - used for troubleshooting
        - log.final.out
            - summary mapping statistics (visualized in MultiQC)
            - for QC (should see 60-90% uniquely mapped reads in human/mouse)
        - log.progress.out
            - job progress per minute
        - SJ.out.tab
            - highly confident collapsed splice junctions

Alignment QC Assessment

- 3' and 5' bias (in coverage)
- nucleotide content
- base/read quality
- sequencing depth
- base distribution
- insert size distribution

Quality of RNA-Seq Toolset (QoRTs)

- GC content
- 5'-3' bias (gene body coverage)
    - a 3'-bias may be due to RNA degradation
    - or stem from polyA-enrichment
- "right" stranded protocol?

HTSeq

- given BAM file and list of gene locations (annotation), counts how many reads map to each gene
    - a gene = union of all its exons
- locations supplied in GTF file
    - GTF and BAM file must use same chromosome naming

- multimapping reads and ambiguous reads are not counted
- 3 modes to handle reads which overlap several genes
  1. union (default)
  2. intersection-strict
  3. intersection-nonempty
- stranded / unstranded library

Stranded vs Unstranded RNAseq Data

Library Prep Method	HISAT2/Cuffdiff	HTSeq
Un-stranded	fr-unstrand	--stranded no
First strand	fr-firststrand	--stranded reverse
Second strand	fr-secondstrand	--stranded yes

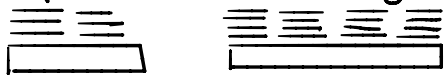
Sequence Alignment to Expression Values

- need to normalize for varying sequencing "depth" and gene length
  - sequencing runs with more sequencing depth will have more reads mapping to each gene
  - longer genes will have more reads mapping to them
- both can bias our analyses
- common normalization strategies
  - RPKM: "Reads per Kilobase Mapped per Million Sequence Reads" (for single end RNA-Seq) $\frac{\text{reads on gene}}{\text{10M reads}} \div \text{length of gene in Kb} = \text{RPKM}$
  - FPKM: "Fragments per Kilobase Mapped per Million Sequence Reads (~ RPKM/2 for paired-end reads) [Fragment = 2 reads]
  - TPM: "Transcripts Per Million"

Lecture 8: Differential Expression

Library Normalization

- need to be able to compare expression between and (less importantly) within samples
  - within-sample comparison: at the same expression level, longer transcripts have more read counts



  - between-sample comparison: higher counts at higher sequencing depth

Problem #1: Adjusting for differences in library sizes
- sequencing depth differences (technical)

Problem #2: Adjusting for differences in library composition
- library composition differences due to expression differences in tissue types, genetic differences (biological)

Gene	Sample 1	Sample 2
A	30	60
B	24	48
C	0	0
D	563	2126
E	5	10
F	13	26
Library Size	635	1270

Gene	Sample 3	Sample 4
A	30	235
B	24	188
C	0	0
D	563	0
E	5	39
F	13	102
Library Size	635	635

Why Do DE Analysis?

- need to be able to account for differences in library size and composition
- count data – unique distribution, typically follows a negative binomial distribution
- large data
- small number of biological replicates
- variance of the measured data is dependent on the mean → "heteroscedasticity"

How Does DESeq2 Scale the Different Samples?

Gene	Sample 1	Sample 2	Sample 3
A	0	10	4
B	2	6	12
C	33	55	200

- goal is to calculate a scaling factor for each sample
- scaling factor needs to take read depth and library composition into account
- low or high count genes → high variance
- median count genes → good correlation
- DESeq2 uses T-tests

① Step 1: Take the natural log of all the values

Gene	ln(Sample 1)	ln(Sample 2)	ln(Sample 3)
A	-inf	2.3	1.4
B	0.7	1.8	2.5
C	3.5	4.0	5.3

② Average each row
- average of log values (Geometric Means) are not easily swayed by outliers

Gene	Average of ln
A	-inf
B	1.7
C	4.3

③ Filter out genes with infinity
- Filter out genes with zero read counts in one or more samples
- This focuses the scaling factors on the housekeeping genes (i.e. genes transcribed at similar levels regardless of tissue types)

Gene	Average of ln
B	1.7
C	4.3

④ Subtract the average log value from the ln(counts)
- this step lets us check out the ratio of the reads in each sample compared to the average across samples

$$\ln(\text{reads for gene X}) - \ln(\text{average for gene X}) = \ln\left(\frac{\text{reads for gene X}}{\text{average for gene X}}\right)$$

Gene	ln(Sample 1)	ln(Sample 2)	ln(Sample 3)
B	-1.0	0.1	0.5
C	-0.8	-0.3	1.3

⑤ Calculate the median of the ratios for each sample
- using median prevents extreme genes from swaying the value too much
- this will focus the scaling factor on the expression of housekeeping genes

Gene	ln(Sample 1)	ln(Sample 2)	ln(Sample 3)
B	-1.0	0.1	0.5
C	-0.8	-0.3	1.3
Median	-0.9	-0.1	0.9

⑥ Convert the medians to "normal numbers" to get the final scaling factors for each sample

Gene	ln(Sample 1)	ln(Sample 2)	ln(Sample 3)
Scaling Factor	$e^{-0.9} = 0.4$	$e^{-0.1} = 0.9$	$e^{0.9} = 2.5$

⑦ Divide the original read counts by the scaling factors

Gene	Sample 1	Sample 2	Sample 3
A	0	10	4
B	2	6	12
C	33	55	200

← Original Read Counts

Gene	Sample 1	Sample 2	Sample 3
A	0	11	2
B	5	7	5
C	83	61	80

← Scaled Read Counts

Normalization with DESeq2

1. Take the natural log of all the values
2. Average each row (=gene)
3. Filter out genes with infinity
4. Subtract the average ln value from ln(counts)
5. Calculate the median of the ratios for each sample
6. Convert the medians to get final scaling factors for each sample
7. Divide the original read counts by the scaling factor

DESeq2's Library Size Scaling Factor

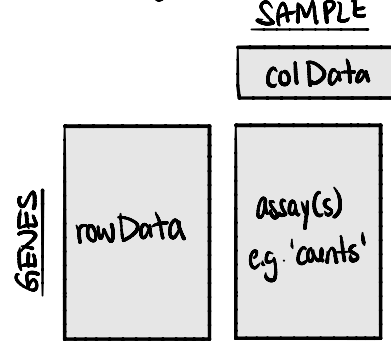
- eliminate genes that are only transcribed in one sample type
- smooth over outlier read counts (via Geometric Mean)
- median further downplays genes that soak up a lot of reads, putting emphasis on moderately expressed genes

DESeq2 Hypothesis Testing

- since we have limited number of replicates per treatment group, DESeq2 will "pool" information across genes by assuming that genes of similar expression strength will have similar variance
- DESeq2 will perform a "hypothesis test" for each gene to see if there's sufficient evidence against the null hypothesis
- Null Hypothesis: logarithmic fold change between treatment and control for a gene's expression is exactly 0
  - the gene is not affected by the treatment

How to Run DESeq2

- DESeq2 expects a count matrix of "un-normalized" counts (i.e. raw counts)
  - this is important for the DESeq2's statistical model to work
  - DESeq2 model internally corrects for library size



① Build DESeqDataSet from HTSeq count data:
- "design" indicates how you want to compare the samples (also needs to be in the metadata)

② Choosing what your "reference level" is for the comparison of interest (using ref=)
- by default, R will choose a reference level for you, based on alphabetical order

③ Perform the differential expression analysis using the DESeq function
- 3 steps wrapped into this one function
  - estimating size factor
  - estimating dispersion
  - perform hypothesis testing

④ Generate results tables (using results)

How Do We Assess Sample Relationships?

- What can we ask with sample distances:
  - which samples are similar/different to each other?
  - does this fit with our expectation from experimental design

① Can calculate sample distance and visualize using hierarchal clustering and heatmap

② Perform principal components analysis (PCA) by projecting data points onto 2D plane so that they spread out in the 2 directions that explain most of the differences

DESeq2 Results

average of normalized count values

expression changes comparing treated vs untreated (log2)

adjusted pvalue

Gene	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj

Multiple Test Correction

- when p-value < 0.05 (significant), this means that 5% of the time, we may report a false positive

Functional Enrichment

- how do you get mechanistic insights into the underlying biology from DE genes?

Pathway Analysis

- simplify analysis by grouping long lists of individual genes into smaller sets of "related functions" reduces the complexity of analysis
- use pre-existing knowledge base to help with this task
  - describe biological processes, components or structures in which individual genes are known to be involved in
  - how and where gene products interact
- analysis at the functional level is useful:
  1. Grouping genes by pathways they are involved in reduces the complexity from thousands of genes to just hundreds of pathways
  2. Finding a pathway that differ between two conditions can provide (testable) biological explanation than just a gene list

"Pathway" Databases: Gene Ontology (GO)

- "An ontology is a formal representation of a body of knowledge within a given domain"
- Ontologies use a set of terms/concepts and highlight the relationships between them
- Gene Ontology uses 3 classes to describe biological functions:
  - Molecular Function
  - Biological Process
  - Cellular Components

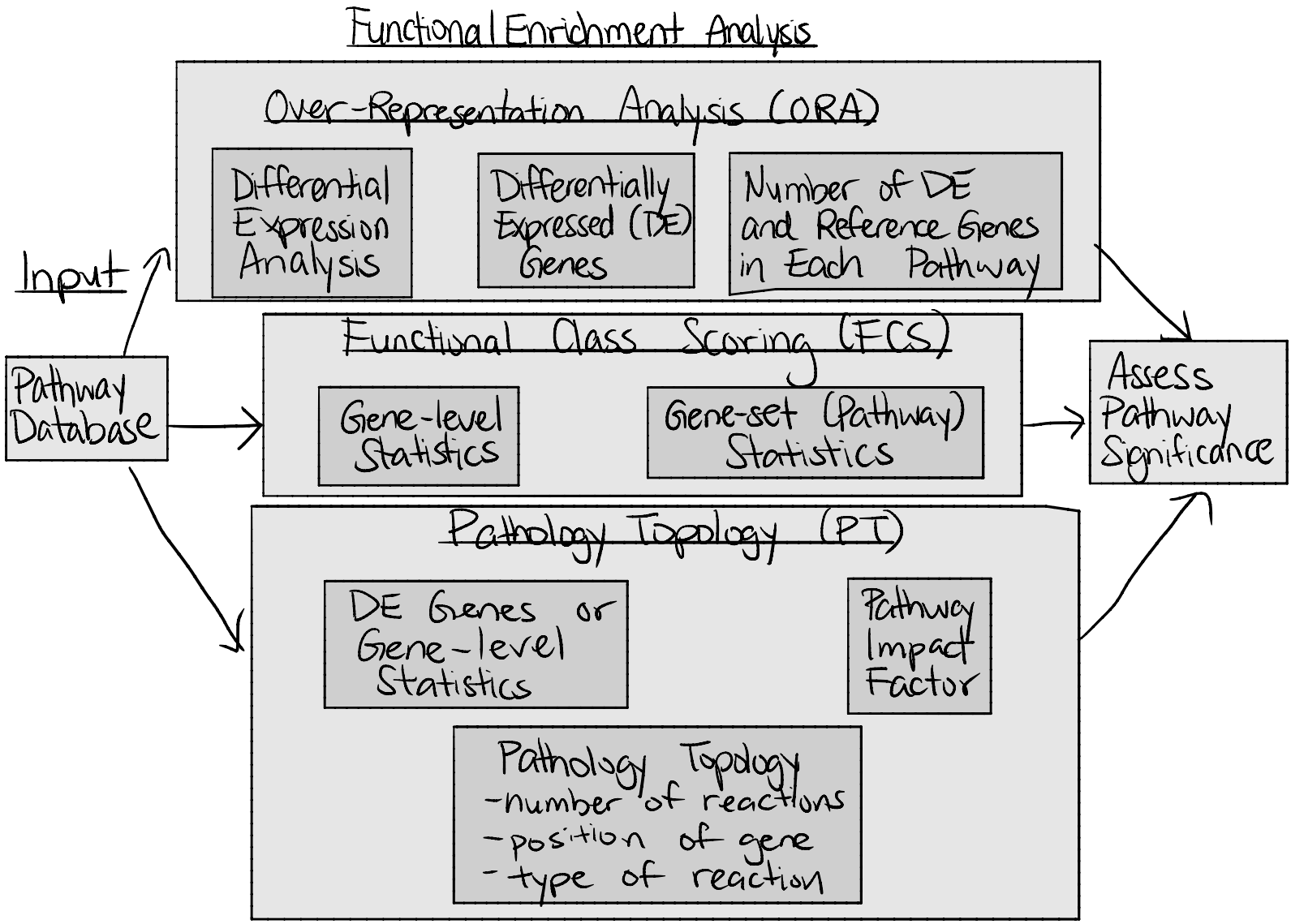
Pathway Databases: Kyoto Encyclopedia of Genes and Genomes (KEGG)

- KEGG is a computer representation of the biological system
  - molecular building blocks of genes and proteins
  - chemical metabolites
  - integrated in a molecular wiring diagrams of interaction and reaction networks
  - contains information for both prokaryotes and eukaryotes

Pathway Databases: Reactome

- a database of signalling and metabolic molecules and how they are organized into biological pathways and processes
- can visualize the relationships using graphical map of known biological processes and pathways
- gives information for how genes/proteins interact with each other
- mostly focused on mammalian systems

Functional Enrichment Analysis



Pathway Analysis Approaches

- can be roughly divided into three generations of pathway analysis approaches
  1. Over-Representation Analysis (ORA) Approaches
     - input is a list of differentially expressed genes
  2. Functional Class Scoring (FCS) Approaches
     - input is the entire data matrix
  3. Pathway Topology (PT) – Based Approaches
     - use the number and type of interactions between gene products

Over-Representation Analysis (ORA)

- statistically evaluates the fraction of genes in a particular pathway found among the set of genes showing changes in expression
- also known as the "2x2 table method"

	Differential Expression	NO Differential Expression	Total
IN Transcription Elongation	12	3	15
NOT IN Transcription Elongation	3	12	15
Total	15	15	30

1. Create an input list using a threshold
2. For each pathway, input genes that are part of the pathway are counted
3. Repeat this for the "background" list of genes (i.e., all the genes with a count in RNASeq)
4. Each pathway is tested for over-representation in the list of input genes with statistical test such as hypergeometric test

Limitations of ORA Methods

- ORA considers the number of genes alone but ignore any values (e.g. fold changes) associated with them
- ORA uses the most significant genes and discards the others, which can result in information loss
- ORA assumes each gene is independent of the other genes
  - not true since there is a complex web of interactions between genes
- ORA assumes each pathway is independent of other pathways