Practical RNA-seq analysis

Prat Thiru
Bioinformatics and Research Computing (BaRC)

http://barc.wi.mit.edu/hot_topics/

Whitehead Institute

Why do RNA-seq?

- RNA-Seq includes experiments to
  - Profile abundance of mRNA and other RNAs
  - Identify alternated spliced transcript isoforms
  - Assemble transcriptome
  - Identify "differentially expressed" genes
  - Identify variants in transcribed regions of genome

Outline

- Experimental design*
- Quality control
- Sequence preparation*
- Mapping spliced reads
- Counting gene levels
- Normalization and identifying "differentially expressed" genes
- Creating figures and summaries*

*not included in the hands-on exercises

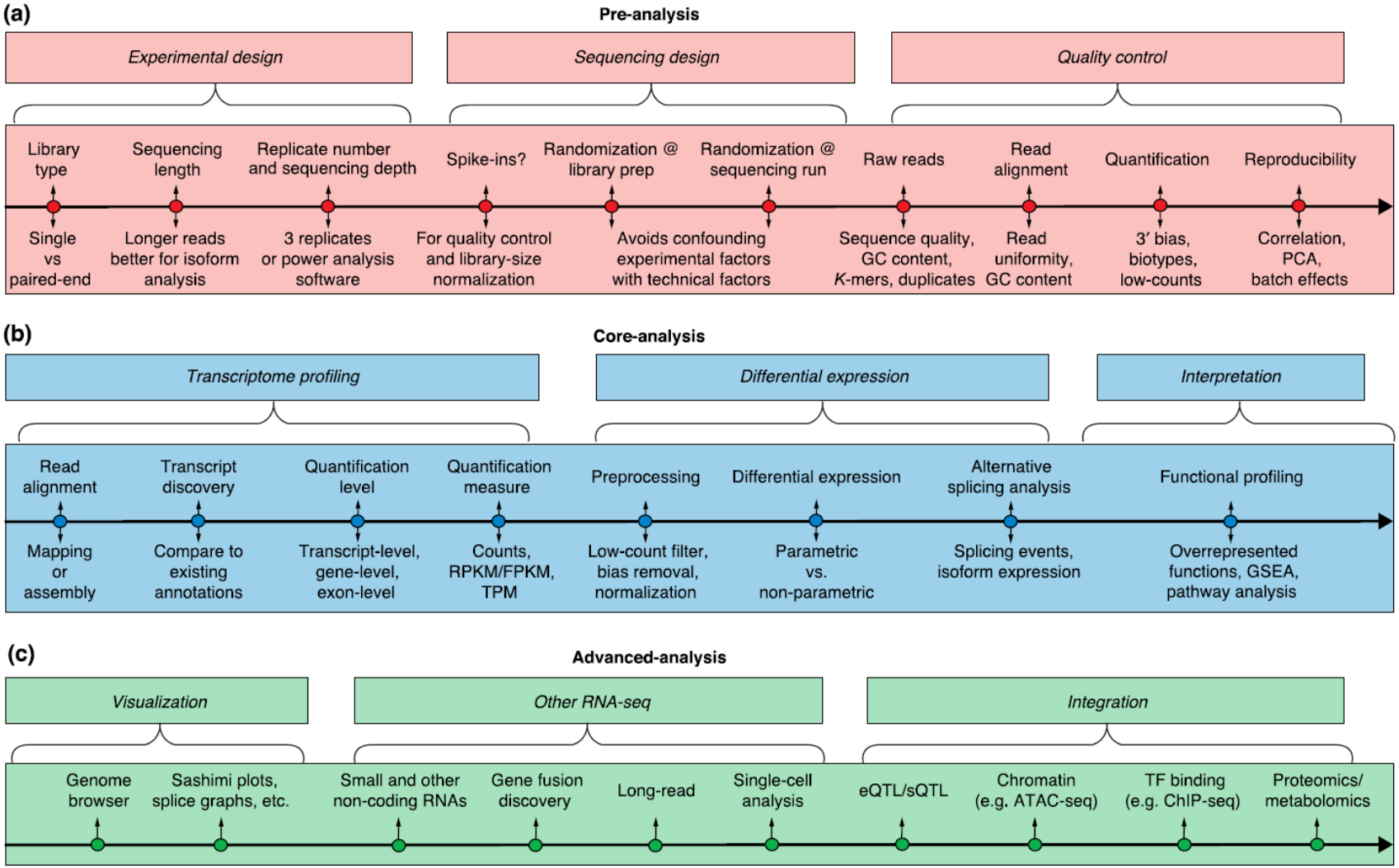
Hands-on exercises

- All data is in
  /nfs/BaRC_Public/Hot_Topics/RNAseq_Apr2018
- Create directories on tak
- Link to data files on BaRC_Public

- See handout for series of commands (step 0)

- Commands can be copied from file
  RNA-seq_Feb_2018.commands.txt

RNA-seq Analysis Overview



Conesa, A., et al. *A survey of best practices for RNA-seq data analysis* Genome Biology (2016)

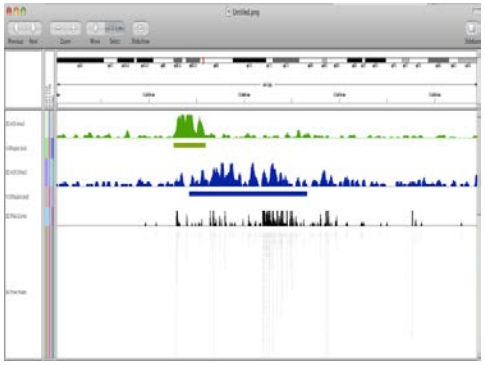
Whitehead Institute
BaRC

Interpreting RNA-seq

RNA-Seq

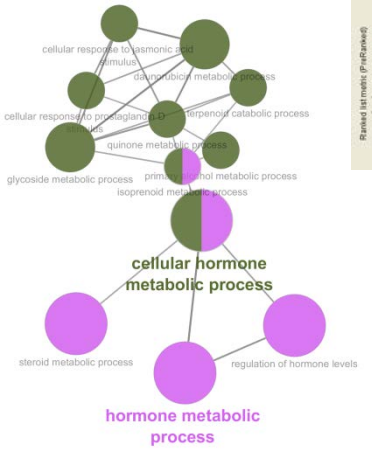
Gene	Log Ratio	p-value
Abcg1	-2.09614	4.72E-07
Adamts5	2.483321	1.33E-07
Alox12b	-2.41347	3.59E-07
Arg1	-2.27214	3.06E-07
AU018091	2.048711	4.62E-07
Bex1	2.591349	4.08E-07
Degs2	-2.46253	1.54E-07
Klk7	-2.18902	3.77E-07
Krt78	-2.89916	2.18E-07
Ly6c1	3.085592	9.41E-08
Ly6g6c	-2.55108	3.62E-07
Sdr16c6	-2.16277	4.05E-07
Sdr9c7	-2.25984	2.63E-07
Sept5	-2.08797	6.31E-07
Kprp	-2.34542	6.77E-07
Ly6a	2.839925	6.04E-07
Slc2a3	2.199118	6.52E-07
Sprr2i	-2.22872	5.67E-07
Mxd1	-1.77522	9.66E-07
Cidea	-1.93749	1.20E-06
Krt16	-1.91642	1.24E-06
Krt8	2.057569	1.22E-06
Trex2	-1.71243	1.29E-06
Aldh3b2	-1.7556	2.63E-06
Asprv1	-1.56796	2.35E-06

Gene List

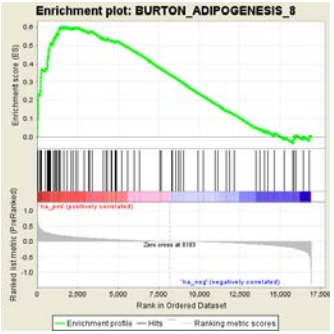


ChIP-Seq or Epigenomics

?



GO Enrichment (ClueGO)



Gene Set Enrichment Analysis (GSEA)

Public RNASeq Datasets

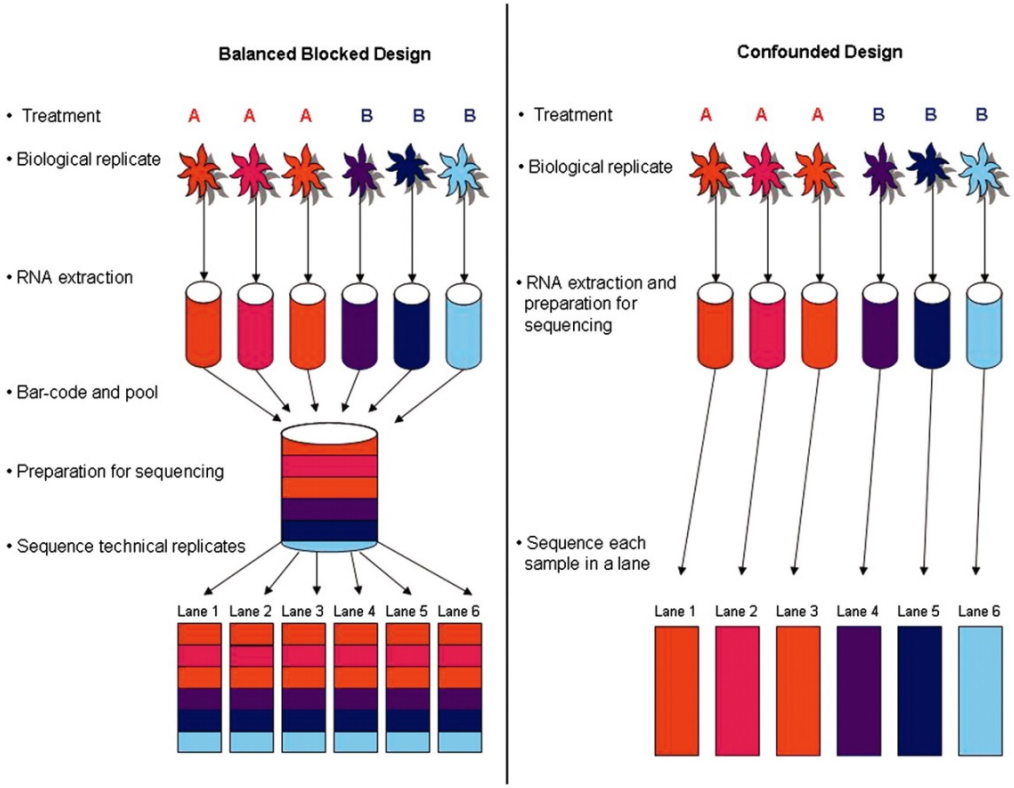
- NCBI GEO
- EBI ArrayExpress
- Broad (e.g. CCLE, GTEx)
- Recount2
- ENCODE
- TCGA
- /nfs/BaRC_datasets

Experimental Design

- Replication is essential if results with confidence are desired.
- With the combination of high numbers of reads per sample and multiplexing, the number of Illumina lanes can be much fewer than number of samples.
- Lots of details to think about:
  - Has someone already done an experiment like this?
  - Total RNA or poly(A) RNA or …
  - Number of samples?
  - Read length?
  - Paired or unpaired reads?
  - Number of reads?
  - What reference genome to use?
  - Stranded or unstranded?
  - What reference transcriptome to use?

Experimental Design

- Lots of data typically cannot make up for a poor experimental design.
- Look out for bias and confounding.
- Short-read sequencing requires an effectively designed experiment.
- See BaRC about reducing batch effects



Auer, PL and Doerge, RW *Statistical design and analysis of RNA sequencing data* Genetics (2010)

Design of Sample Experiment

- How does "gene expression" (really: transcript levels) differ between European and Africans?
- European samples (n=2) Montgomery et al., 2010
  - 37nt paired-end reads
- African samples (n=2) Pickrell et al., 2010
  - 46nt single-end reads
- Sample data contains about 10% of total reads
- Is this a good design?

QC Before Alignment

- FastQC, use mulitQC to view
- Check quality of file of raw reads (fastqc_report.html)
- Respond to QC analysis:
  - Filter poor-quality reads
  - Trim poor-quality positions
  - Trim adapter and/or other vector
- Check quality of file of modified reads
- See previous Hot Topic:

  NGS: Quality Control and Mapping Reads (Feb 2014)

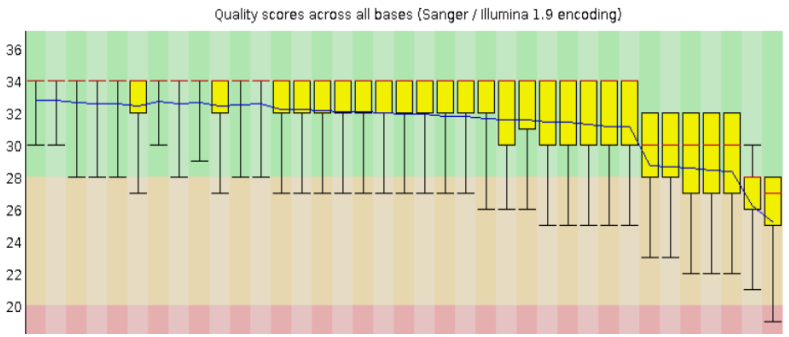
- See handout for fastqc command (step 1)

```
@WIGTC-HISEQ3:1:1212:18183:1997#TTAGGC/1
NCCACAACAGGGCACGGTGCGGAATAGAGAACTATCCCTT
+WIGTC-HISEQ3:1:1212:18183:1997#TTAGGC/1
B[[[[aeaeeaeeeeeeeeeeeeeeeeeeeeeeeeeeee
```

Basic Statistics

Measure	Value
Filename	NA07357_CEU_1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	7940176
Filtered Sequences	0
Sequence length	37
%GC	56

Per base sequence quality



Responding to Quality Issues

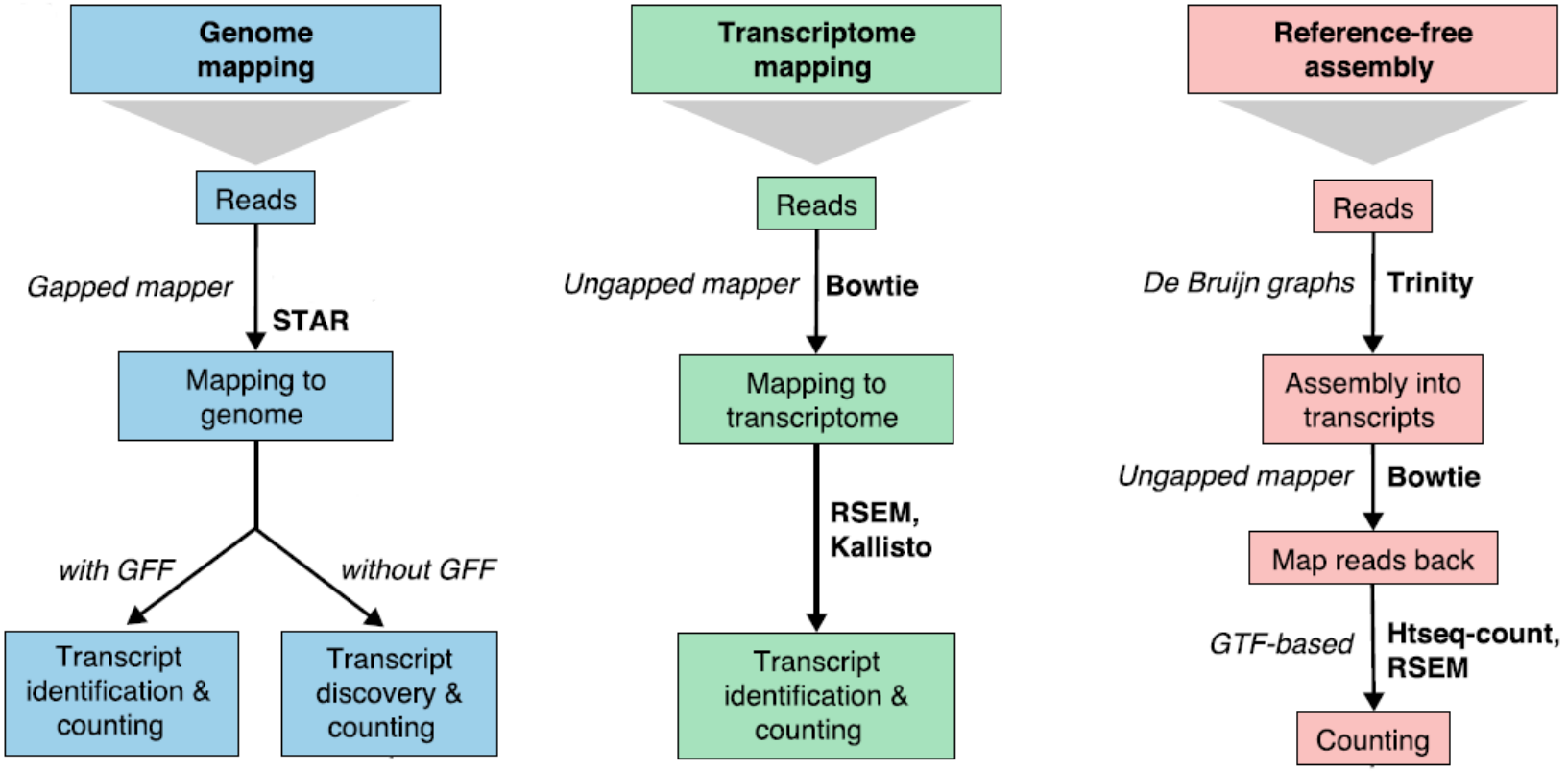
- Method 1:
  - Keep all reads as is
  - Map as many as possible
- Method 2:
  - Drop all poor-quality reads
  - Trim poor-quality bases
  - Map only good-quality bases
- Which makes more sense for your experiment?

RNA-Seq Genome Mapping

- Reads can be mapped with a **splice-aware alignment tool** such as STAR (recommended), or many others (e.g. HISAT2*)
- The ideal tool should map to best hit(s), whether to continuous or spliced genome segments
- Look at alignment stats and mapped reads in a browser (and re-map if needed).

*Successor of TopHat (note: TopHat no longer recommended!)

Mapping Considerations



Adapted from Conesa, A., et al. *A survey of best practices for RNA-seq data analysis* Genome Biology (2016)

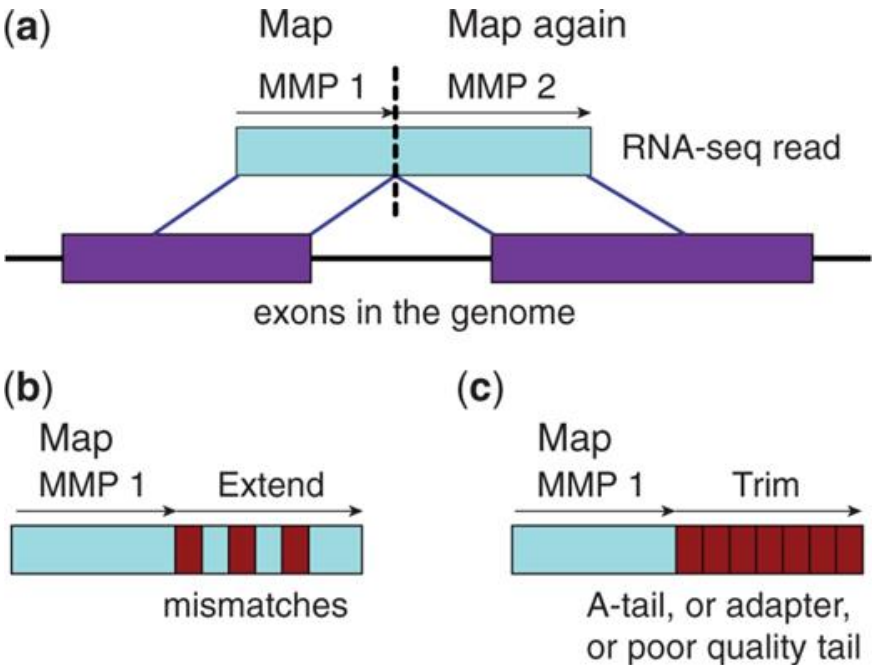
Mapping considerations

- Type of quality score encoding?
- Use all or just canonical chromosomes?
- Include known splice junctions (in GTF file)?
- Look for novel splice junctions?
- How short of a sub-read should map to an exon boundary?
- How long are your introns?

STAR Aligner

I. Sequential search for Maximal Mappable Prefix (MMP)

II. Stitch together all the seeds that were aligned to the genome from I)



Dobin, A., et al. *STAR: ultrafast universal RNA-seq aligner* Bioinformatics (2012)

Alignment with STAR

- Create genome index using *genomeGenerate,* also see /nfs/genomes

- Run alignment, e.g.

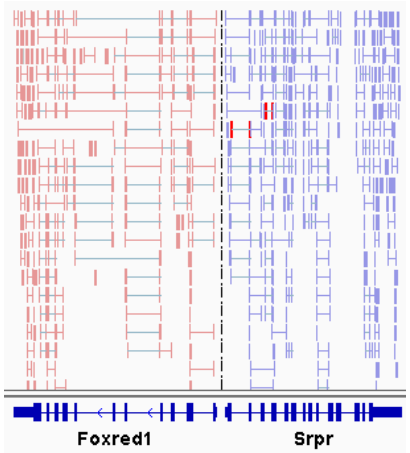
STAR --genomeDir /path/to/GenomeDir --readFilesIn /path/to/read1.fq.gz /path/to/read2.fq.gz --sjdbScore 2 --outFileNamePrefix whateverPrefix --runThreadN 8 --readFilesCommand zcat --outSAMtype BAM SortedByCoordinate

--runMode <alignReads, genomeGenerate>	"alignReads" does the actual mapping. "genomeGenerate" generates the genomeDir required for mapping (default = alignReads).
--genomeDir </path/to/GenomeDir>	Specifies the path to the directory used for storing the genome information created in the genomeGenerate step.
--genomeFastaFiles <genome FASTA files>	Specifies genome FASTA files to be used.
--readFilesIn <read1.fastq read2.fastq>	Specifies the fastq files containing the reads, can be single-end or paired-end.
--sjdbScore <n>	Provides extra alignment score for alignments that cross database junctions (default = 2).
--runThreadN <n>	Specifies the number of threads to use.
--readFilesCommand <cmd>	Specifies the command to uncompress compressed fastq files. For gzipped files (*.gz) use --readFilesCommand zcat.
--outSAMtype <BAM sortingMode>	Specifies the type of BAM file to create. Options: 'BAM Unsorted', 'BAM SortedByCoordinate', 'BAM Unsorted SortedByCoordinate' (to create both unsorted and sorted BAMs)

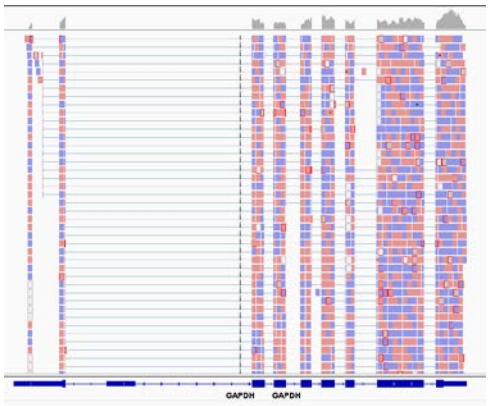
See handout for STAR command (step 2)

QC after Alignment

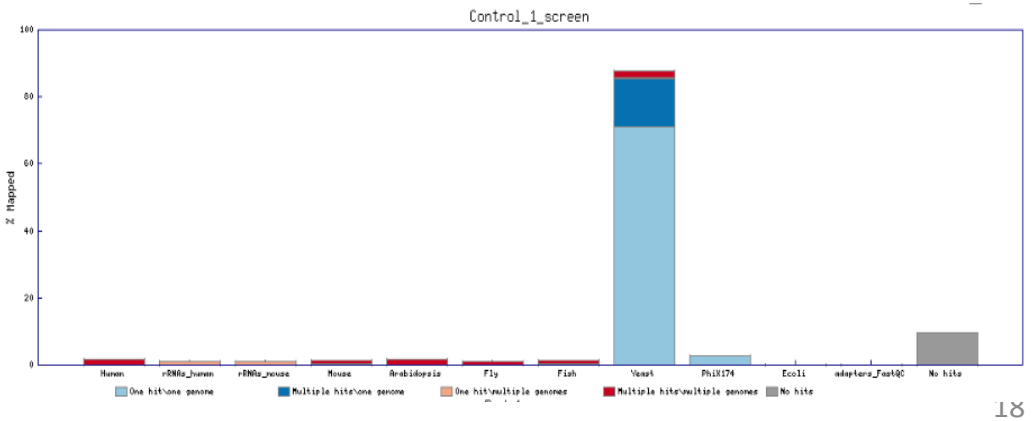
- Confirm that reads are stranded or unstranded
  - Run infer_experiment.py (from **RseQC** package)
  - Look at BAM reads in genome browser
- Contamination?
  - **FastQ Screen**



stranded

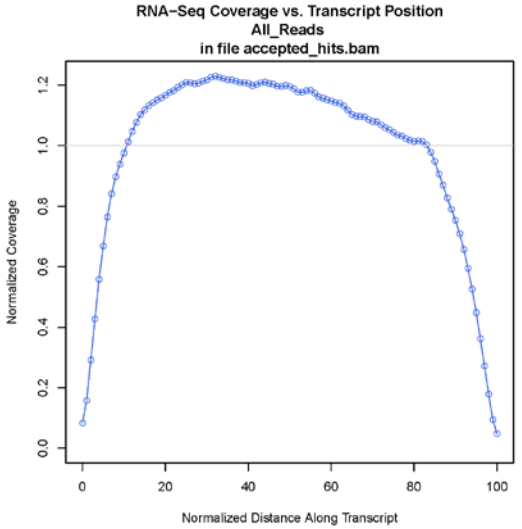


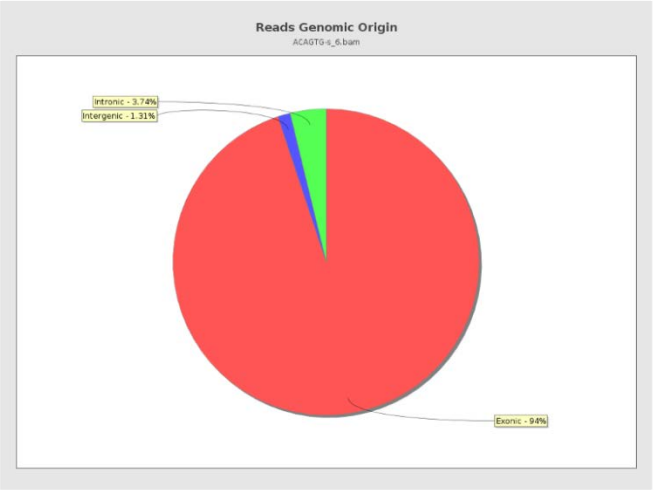
unstranded



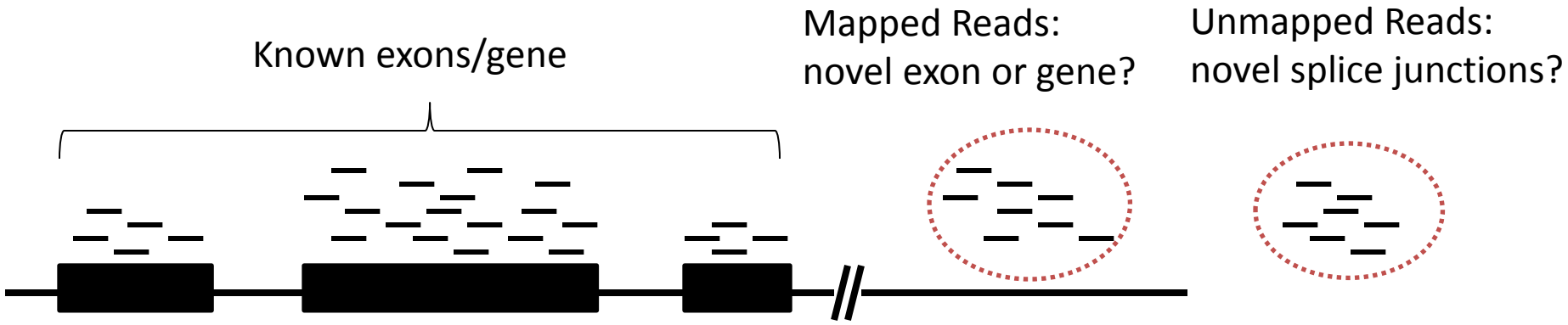
QC after Alignment

- Do reads cover the length of a typical transcript, or is there 3' or 5' bias?
  - Run **Picard** tool: CollectRnaSeqMetrics
- What fraction of reads map to annotated exons?
  - Run '**qualimap** rnaseq'
- See BaRC SOPs for commands





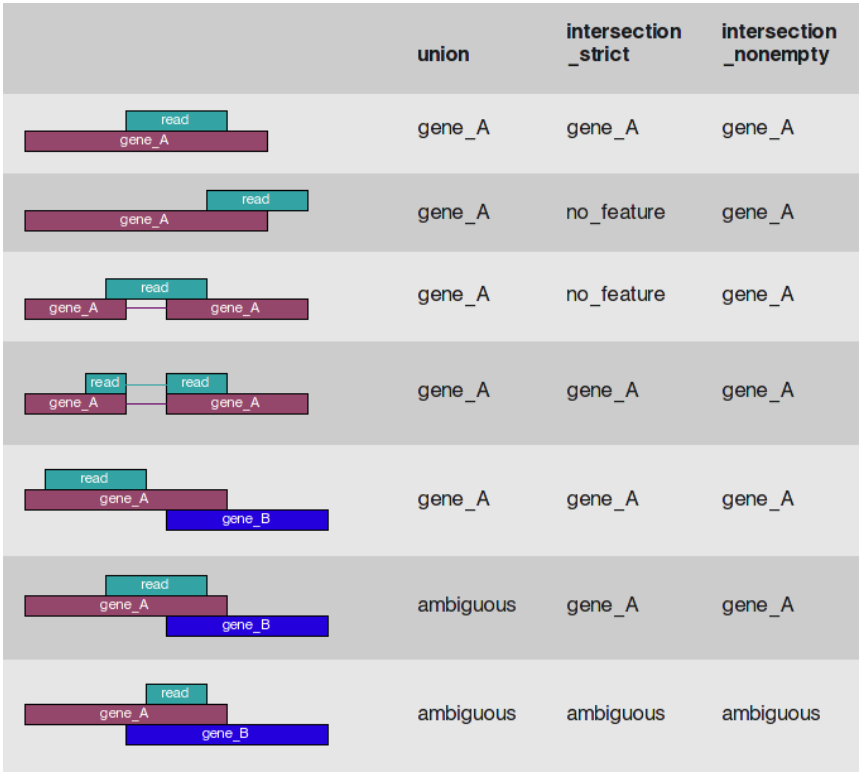
RNA-Seq Mapping



Counting RNA-Seq Features

- What features are of interest?
- Gene, transcript, and/or exon counts?
- What happens if a read maps to multiple places?
- What happens if a read maps to multiple features?
- Does the direction of a read need to agree with the direction of the feature?
- What are the limitations of your experiment and analysis details?

htseq-count "modes"



Counting Methods

- htseq-count

  htseq.readthedocs.io/en/master/count.html
  - Output is raw counts

- featureCounts (recommended)

  bioinf.wehi.edu.au/featureCounts/
  - Output is raw counts

- Cufflinks

  cole-trapnell-lab.github.io/cufflinks/
  - Output is FPKM and related statistics

- Bedtools (intersectBed; coverageBed)

  bedtools.readthedocs.io/
  Output is raw counts (but may need post-processing)

Running featureCounts

Count reads mapping to the specified gene models:

- Usage:

```
featureCounts [options] -a <annotation_file> -o
<output_file> input_file1 [input_file2] ...
```

- Example:

```
#single-end reads (unstranded)
 featureCounts -a gene_anotations.gtf -o
 MySample.featureCounts.txt MySample.bam
#paired-end reads (forward stranded)
 featureCounts –p -s 1 -a gene_anotations.gtf –o
 MySample.featureCounts.txt MySample.sorted.bam
```

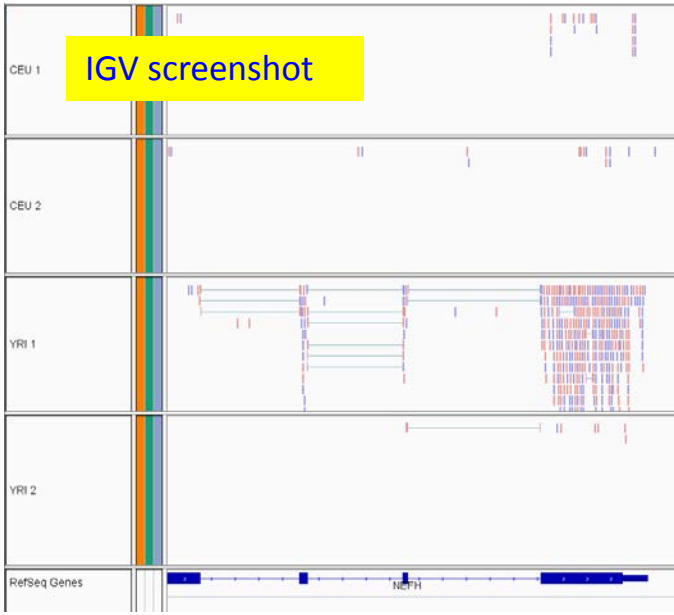
Running featureCounts: Options

Option	Description
--minOverlap	Minimum number of overlapping bases in a read that is required for read assignment. 1 by default.
--fracOverlap	Minimum fraction of overlapping bases in a read that is required for read assignment.
-M	Multi-mapping reads will also be counted.
-s	Perform strand-specific read counting. Acceptable values: 0 (unstranded), 1 (stranded) and 2 (reversely stranded). 0 by default.

See handout for featureCounts commands (step 3)

Differential Expression Statistics

- RNA-Seq RNA levels can be written as a continuous value (FPKM) or as integer counts
- Statistics of these are inherently different
- All statistics are dependent on coverage (number of mapped reads)
- All statistics require sample replication to quantify within-sample variability
- Lack of replication greatly reduces the strength of one's conclusions.



Whitehead Institute
BaRC

Normalization

- Raw counts cannot be compared directly
- Correct for sequencing depth (i.e. library size) and RNA composition bias
  - CPM: counts per million
  - FPKM*: fragments per kilobase per million mapped reads
  - TPM*: transcripts per million
  - DESeq: Relative Log Expression (RLE)
  - edgeR: Trimmed Mean M-values (TMM)

*can be used to compare across genes or transcripts

Differential Expression Methods

- Count-based methods (recommended)
  - Input is matrix of raw counts
  - DESeq2 (R package) -- recommended
  - edgeR (R package)
  - Typically used to compare gene counts
- Cufflinks (cuffdiff)
  - Inputs are mapped reads (SAM format)
  - Typically used to compare gene and transcript counts
- See handout for DESeq2 commands (step 4)

Interpreting DESeq2 output

Gene ID (from GTF file)	Mean norm counts	Log2 (fold change)	logFC std error	Wald statistic	Raw p-value	FDR p-value	Raw counts				Normalized counts = raw / (size factor)			
Feature.ID	baseMean	log2(YRI/CEU)	lfcSE	stat	pvalue	padj	CEU_NA07357	CEU_NA11881	YRI_NA18502	YRI_NA19200	CEU_NA07357.norm	CEU_NA11881.norm	YRI_NA18502.norm	YRI_NA19200.norm
ENSG00000251705	114.15	-3.48	0.46	-7.60	2.90E-14	4.48E-11	197	250	8	14	172.84	257.56	5.77	20.42
ENSG00000236552	66.39	-3.86	0.52	-7.47	8.29E-14	1.07E-10	180	96	2	5	157.93	98.9	1.44	7.29
ENSG00000226958	2073.59	-2.45	0.38	-6.52	6.88E-11	7.58E-08	3324	4148	407	556	2916.42	4273.49	293.32	811.14
ENSG00000064886	54.20	3.31	0.54	6.09	1.11E-09	1.05E-06	9	1	84	101	7.9	1.03	60.54	147.35
ENSG00000198786	2006.17	-1.95	0.32	-6.08	1.23E-09	1.05E-06	4601	2391	834	633	4036.84	2463.33	601.05	923.47
ENSG00000100292	58.87	-2.86	0.47	-6.04	1.54E-09	1.19E-06	131	98	13	7	114.94	100.96	9.37	10.21

```
sizeFactors (from DESeq2):

CEU_NA07357 CEU_NA11881 YRI_NA18502 YRI_NA19200
  1.1397535   0.9706359   1.3875763   0.6854579
```


Whitehead Institute
BaRC

Differential Expression Issues

- Given that statistics are
  - based on complex models
  - influenced by even more complex biology

  The p-values may not be accurate but can be very effective at ranking genes

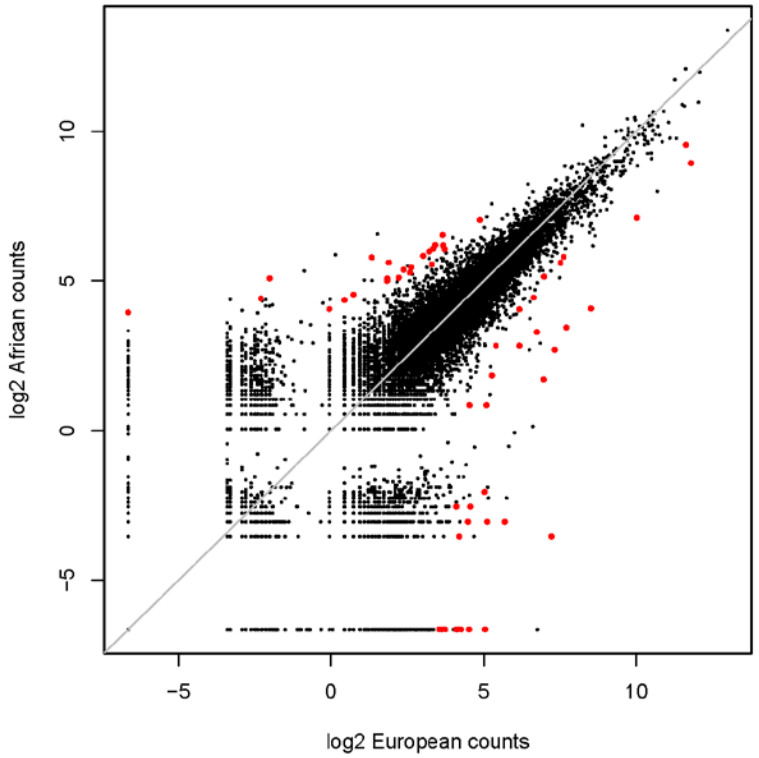
- Statistics don't work very well when one sample has no counts.

- You have to choose appropriate thresholds.

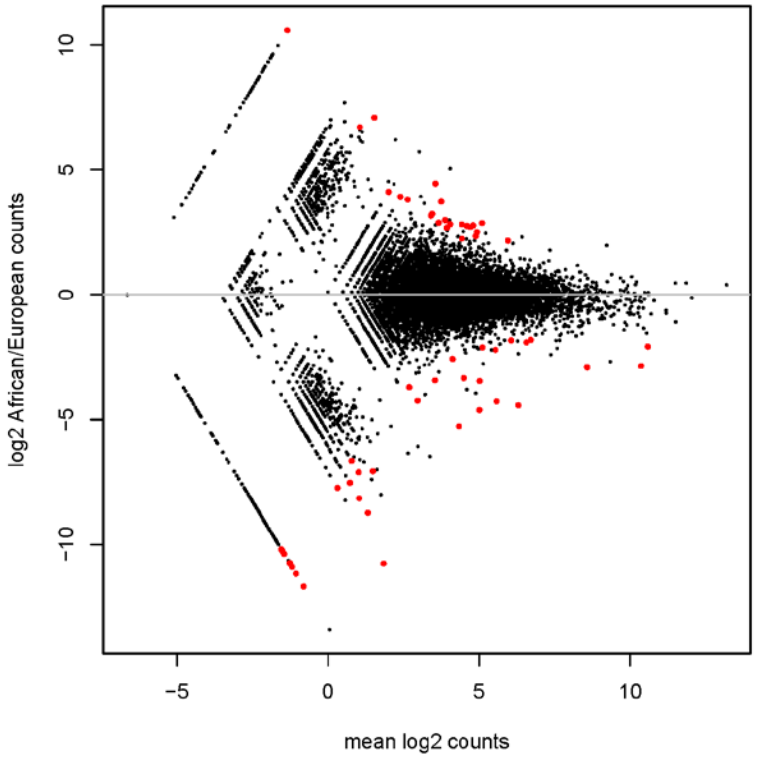
Presenting Results

- What do you want to show?
- All expression-type figures are possible
- All-gene scatterplots can be helpful to
  - See level and fold-change ranges
  - Identify sensible thresholds
  - Hint at data or analysis problems
- Heatmaps are useful if many conditions are being compared but only for gene subsets
- Output normalized read counts with same method used for DE statistics
- Whenever one gene is especially important, look at the mapped reads in a genome browser

Scatterplots

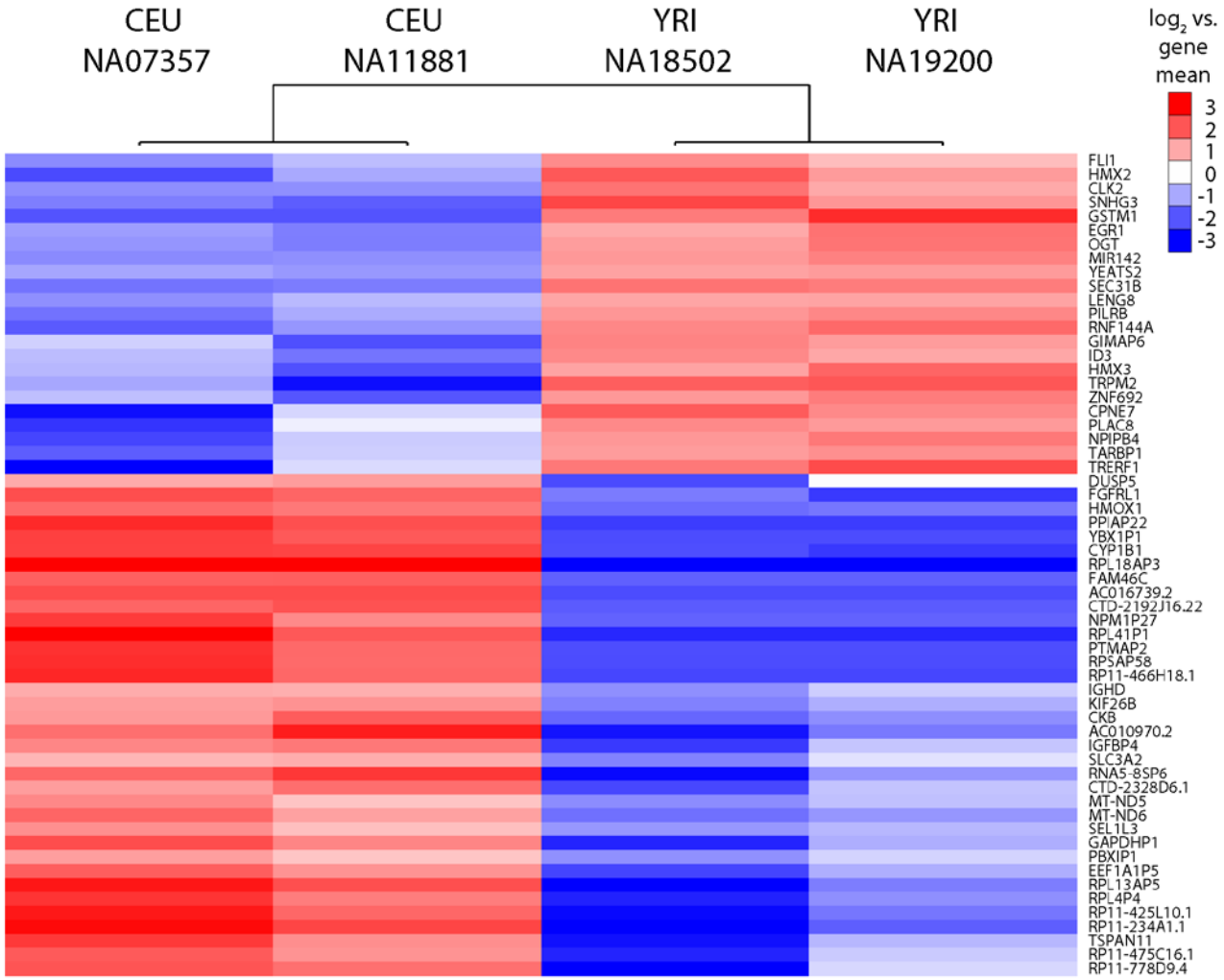


Standard scatterplot



MA (ratio-intensity) plot

Heatmap example: genes with FDR < 0.2



Excel:
- Add pseudocounts

Cluster 3.0:
- Log-transform
- Mean center
- Cluster

Java TreeView:
- Visualize
- Export

Illustrator
- Assemble

Summary

- Experimental design
- Quality control (fastqc)
- Sequence preparation
- Mapping spliced reads (STAR)
- Counting gene levels (featureCounts)
- Normalization and identifying "differentially expressed" genes (DESeq2 R package)
- Creating figures and summaries
- Save your commands!

Resources

- Previous Hot Topics (http://jura.wi.mit.edu/bio/education/hot_topics/)
- An introduction to R and Bioconductor: A BaRC Short Course
- BaRC SOPs (http://barcwiki.wi.mit.edu/wiki/SOPs)
- Online software manuals
  - STAR, featureCounts, DESeq2, etc.
- Conesa *et al.* A survey of best practices for RNA-seq data analysis. Genome Biol. 2016 Jan 26;17:13.
- Various datasets: **`/nfs/BaRC_datasets`**
- Genome index and GTF files are in **`/nfs/genomes`**