So you want to do a RNAseq Experiment Differential Expression Analysis

Matt Settles

Director, Bioinformatics Core

Treating Bioinformatics as a Data Science

Seven stages to data science

1. Define the question of interest
2. Get the data
3. Clean the data
4. Explore the data
5. Fit statistical models
6. Communicate the results
7. Make your analysis reproducible

Data science done well looks easy and that's a big problem for data scientists

simplystatistics.org
March 3, 2015 by Jeff Leek

What is Differential Expression

Differential expression analysis means taking the *normalised* sequencing fragment count data and performing statistical analysis to discover *quantitative* changes in expression levels between experimental groups.

For example, we use statistical testing to decide whether, for a given gene, an observed difference in fragment counts is significant, that is, whether it is greater than what would be expected just due to natural random variation.

Designing Experiments

Beginning with the question of interest (and working backwards)

- The final step of a DE analysis is the application of a linear model to each gene in your dataset.

  Traditional statistical considerations and basic principals of statistical design of experiments apply.

  - **Control** for effects of outside variables, avoid/consider possible biases, avoid confounding variables in sample preparation.
  - **Randomization** of samples, plots, etc.
  - **Replication** is essential (triplicates are THE minimum)

- You should know your final (DE) model and comparison contrasts before beginning your experiment.

Three outcomes
Goldilocks and the three bears

- Technical and/or biological variation exceeds that of experimental variation, results in 0 differentially expressed genes
- Experiment induces a significant phenotype with cascading effects and/or little to no biological variation between replicates (ala cell lines), results in 1000s of DE genes. Some of which are directly due to experiment; however, most due to cascading effects.
- Technical artifacts are controlled. Biological variation is induced in the experiment, and cascading effects are controlled, or accounted for, results in 100s of DE genes directly applicable to the question of interest.

General rules for preparing samples

- Prepare more samples then you are going to need, i.e. expect some will be of poor quality, or fail
- Preparation stages should occur across all samples at the same time (or as close as possible) and by the same person
- Spend time practicing a new technique to produce the highest quality product you can, reliably
- Quality should be established using Fragment analysis traces (pseudo-gel images, RNA RIN > 7.0)
- DNA/RNA should not be degraded
  - 260/280 ratios for RNA should be approximately 2.0 and 260/230 should be between 2.0 and 2.2. Values over 1.8 are acceptable
- Quantity should be determined with a Fluorometer, such as a Qubit.

Generating RNA-seq libraries

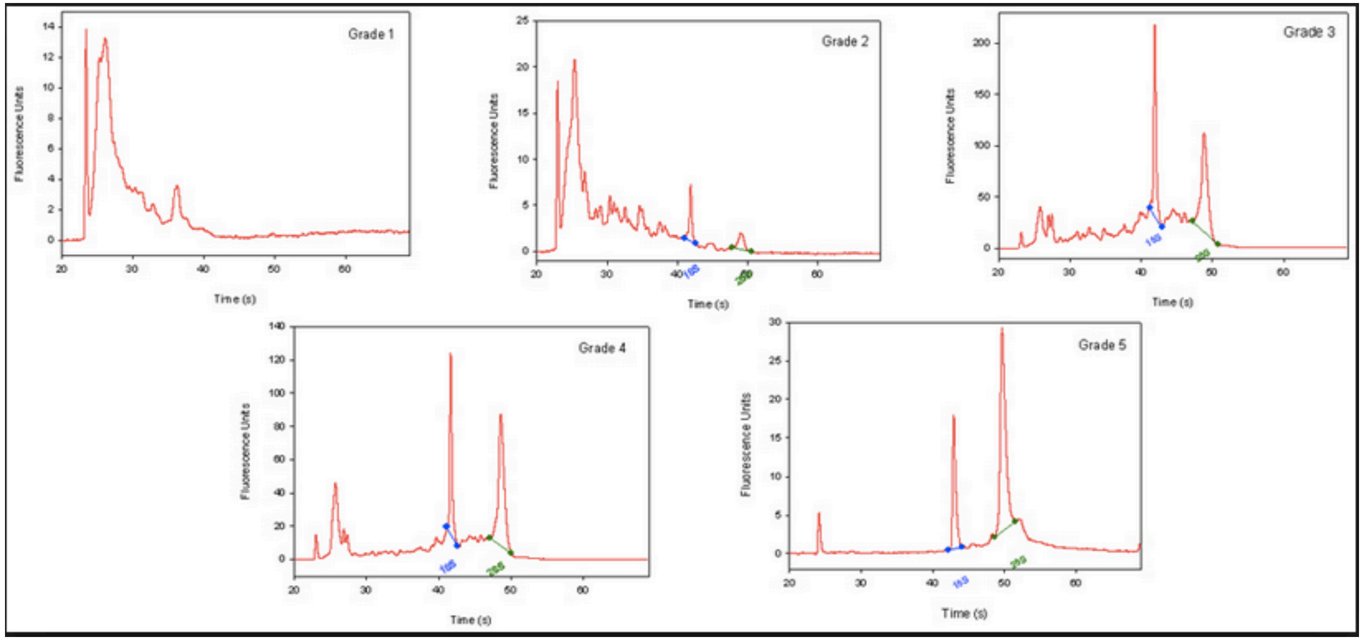
Considerations

- QA/QC of RNA samples
- What is the RNA of interest
- Library Preparation
  - Stranded Vs. Unstranded
- Size Selection/Cleanup
  - Final QA

QA/QC of RNA samples

RNA Quality and RIN (RQN on AATI Fragment Analyzer)

- RNA sequencing begins with high-quality total RNA, only an Agilant BioAnalyzer (or equivalent) can adequately determine the quality of total RNA samples. RIN values between 7 and 10 are desirable.



BE CONSISTANT!!!

RNA of interest

- From “total RNA” we extract “RNA of interest”. Primary goal is to NOT sequence 90% (or more) ribosomal RNAs, which are the most abundant RNAs in the typical sample. there are two main strategies for enriching your sample for “RNA of interest”.
  - polyA selection. Enrich mRNA (those with polyA tails) from the sample by oligo dT affinity.
  - rRNA depletion. rRNA knockdown using RiboZero (or Ribominus) is mainly used when your experiment calls for sequencing non-polyA RNA transcripts and non-coding RNA (ncRNA) populations. This method is also usually more costly.

rRNA depletion will result in a much larger proportion of reads align to intergenic and intronic regions of the genome.

Library Preparation

- Some library prep methods first require you to generate cDNA, in order to ligate on the Illumina barcodes and adapters.
  - cDNA generation using oligo dT (3’ biased transcripts)
  - cDNA generation using random hexomers (less biased)
  - full-length cDNAs using SMART cDNA synthesis method
- Also, can generate strand specific libraries, which means you only sequence the strand that was transcribed.
  - This is most commonly performed using dUDP rather than dNTPs in cDNA generation and digesting the “rna” strand.
  - Can also use a RNA ligase to attach adapters and then PCR the second strand and remainder of adapters.

Size Selection/Cleanup/qA

Final insert size optimal for DE are ~ 150bp

- Very important to be consistent across all samples in an experiment on how you size select your final libraries. You can size select by:
  - Fragmenting your RNA, prior to cDNA generation.
    - Chemically heat w/magnesium
    - Mechanically (ex. ultra-sonicator)
- Cleanup/Size select after library generation using SPRI beads or (gel cut)
- QA the samples using an electrophoretic method (Bioanalyzer) and quantify with qPCR.

Most important thing is to be consistent!!!

[SUMMARY] Generating RNA-seq libraries

Considerations

- QA/QC of RNA samples [Consistency across samples is most important.]
- What is the RNA of interest [polyA extraction is recommended.]
- Library Preparation
  - Stranded Vs. Unstranded [Standard stranded library kits]
- Size Selection/Cleanup [Target mean 150bp or kit recommendation]
  - Final QA [Consistency across samples is most important.]

Sequencing Depth

- Coverage is determined differently for ”Counting” based experiments (RNAseq, amplicons, etc.) where an expected number of reads per sample is typically more suitable.
- The first and most basic question is how many reads per sample will I get Factors to consider are (per lane):
    1. Number of reads being sequenced
    2. Number of samples being sequenced
    3. Expected percentage of usable data

$$\frac{reads}{sample} = \frac{reads.sequenced \ * 0.8}{samples.pooled}$$

- Read length, or SE vs PE, does not factor into sequencing depth.

Sequencing

Characterization of transcripts, or differential gene expression

Factors to consider are:

- Read length needed depends on likelihood of mapping uniqueness, but generally longer is better and paired-end is better than single-end. (2 x >75bp is best)
- Interest in measuring genes expressed at low levels (<< level, the >> the depth and necessary complexity of library)
- The fold change you want to be able to detect (< fold change more replicates, more depth)
- Detection of novel transcripts, or quantification of isoforms requires >> sequencing depth

The amount of sequencing needed for a given sample/experiment is determined by the goals of the experiment and the nature of the RNA sample.

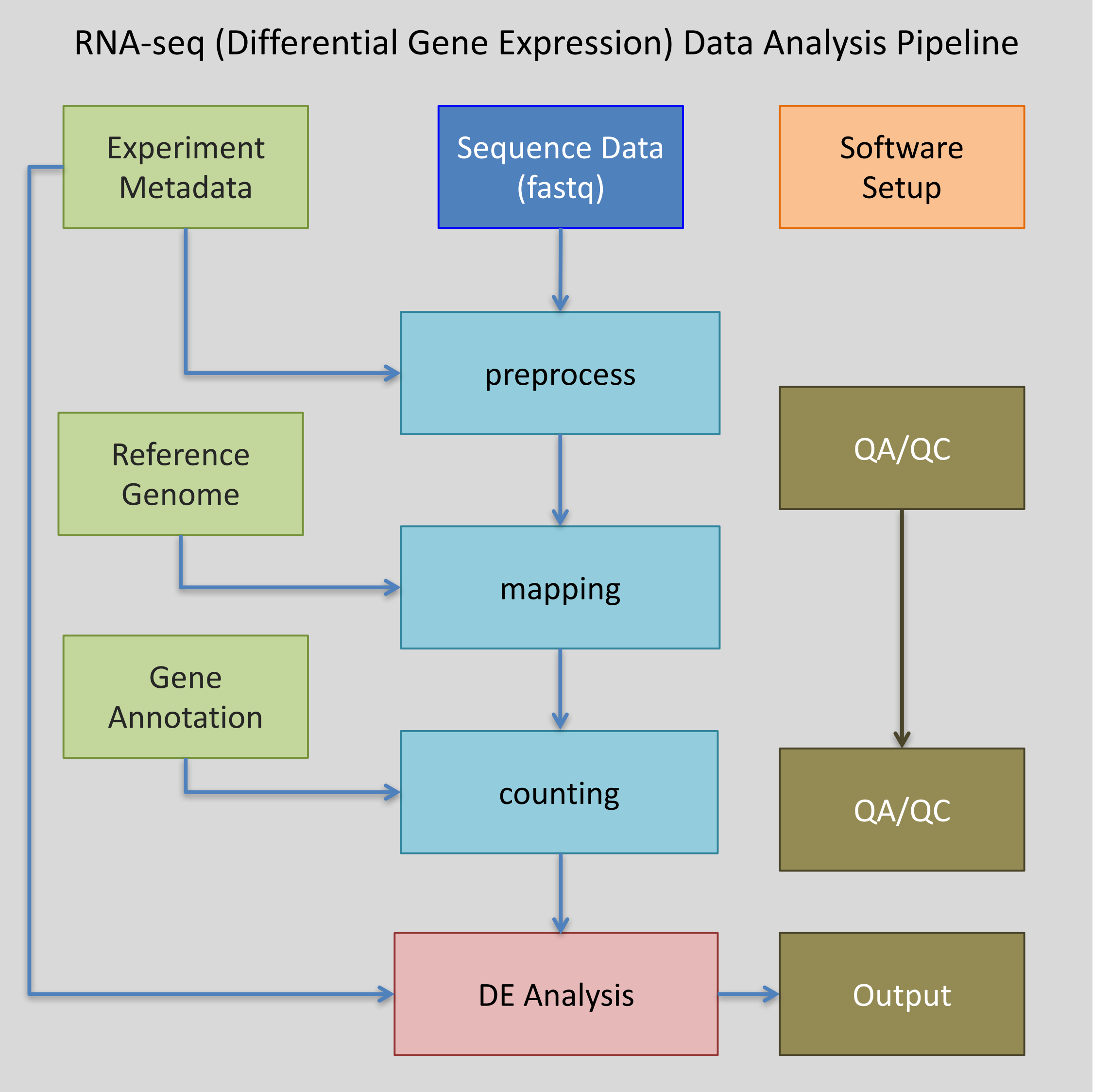
Overview of RNA-SEQ data analysis

Prerequisites

- Access to a multi-core (24 cpu or greater), ‘high’ memory 64Gb or greater Linux server.
- Familiarity with the ’command line’ and at least one programming language.
- Basic knowledge of how to install software
- Basic knowledge of R (or equivalent) and statistical programming
- Basic knowledge of Statistics and model building

UCDAVIS Bioinformatics Core

RNA-seq pipeline overview



Sequence Preprocessing

Why Preprocess reads

- We have found that aggressively “cleaning” and processing reads can make a large difference to the **speed** and **quality** of assembly and mapping results. Cleaning your reads means, removing reads/bases that are:
  - other unwanted sequence (polyA tails in RNA-seq data)
  - artificially added onto sequence of primary interest (vectors, adapters, primers)
  - join short overlapping paired-end reads
  - low quality bases
  - originate from PCR duplication
  - not of primary interest (contamination)
- Preprocessing also produces a number of statistics that are technical in nature that should be used to evaluate “experimental consistancy”

QA/QC

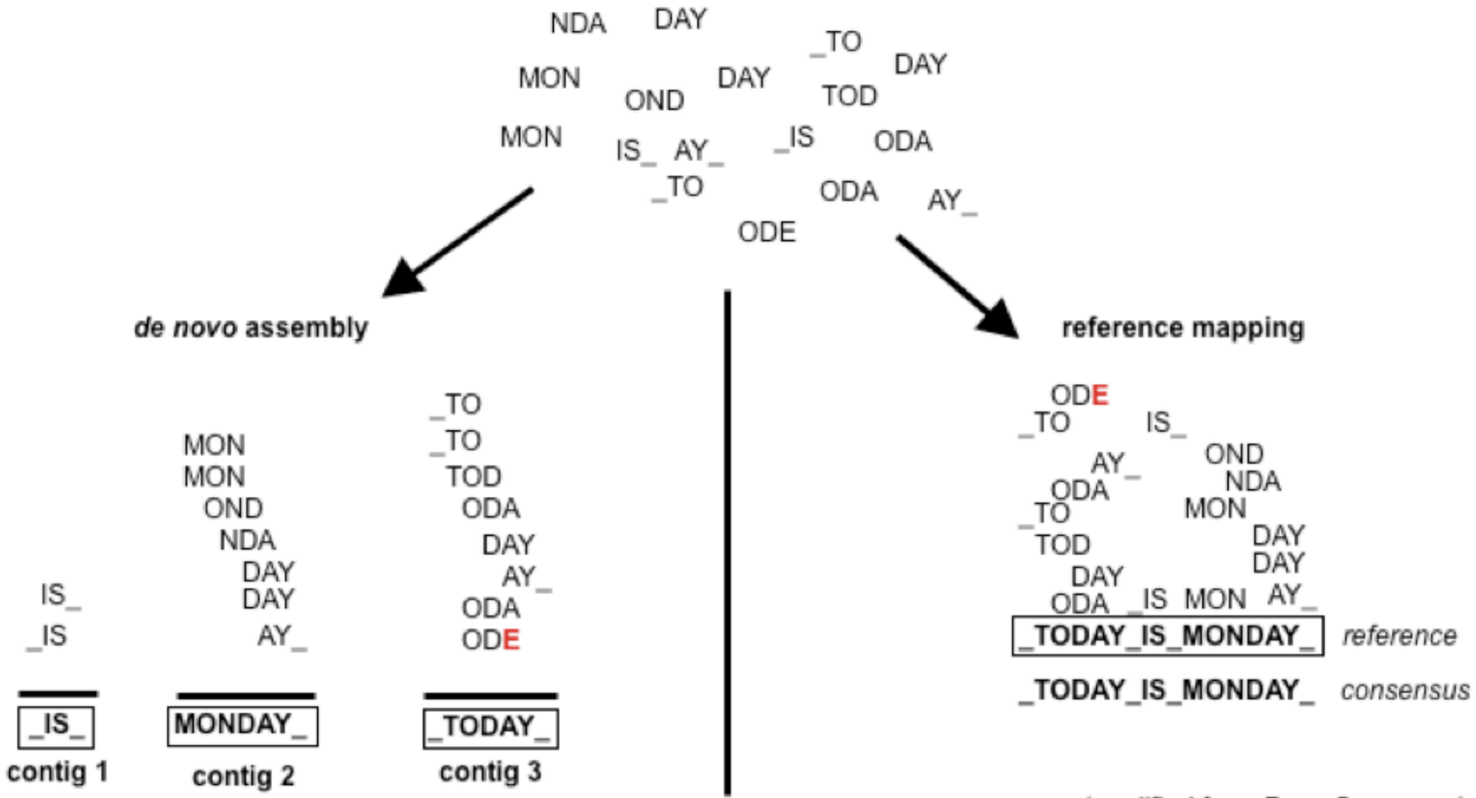
- Beyond generating better data for downstream analysis, cleaning statistics also give you an idea as to the quality of the sample, library generation, and sequencing quality used to generate the data.
- This can help inform you of what you might do in the future.
- I’ve found it best to perform QA/QC on both the run as a whole (poor samples can affect other samples) and on the samples themselves as they compare to other samples (REMEMBER, BE CONSISTANT).
  - Reports such as Basespace for Illumina, are great ways to evaluate the runs as a whole.
  - PCA/MDS plots of the preprocessing summary are a great way to look for technical bias across your experiment

Sequence Mapping

Mapping vs Assembly

- Given sequence data,
  - Assembly seeks to put together the puzzle without knowing what the picture is
  - Mapping tries to put together the puzzle pieces directly onto an image of the picture
- In mapping the question is more, given a small chunk of sequence, where in the genome did this piece most likely come from.
- The goal then is to find the match(es) with either the “best” edit distance (smallest), or all matches with edit distance less than max edit dist. Main issues are:
  - Large search space
  - Regions of similarity (aka repeats)
  - Gaps (INDELS)
  - Complexity (RNA, transcripts)

Example



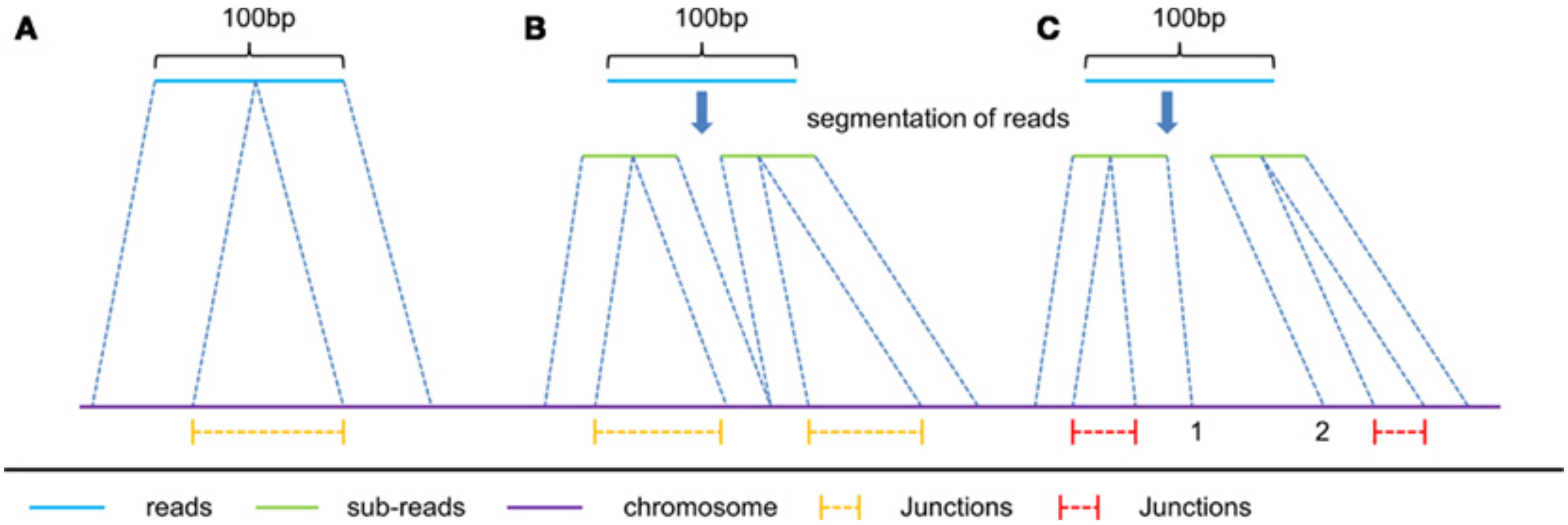
(modified from Panu Somervuo)

Consideration

- Placing reads in regions that do not exist in the reference genome (reads extend off the end) [mitochondrial, plasmids, structural variants, etc.].
- Sequencing errors and variations: alignment between read and true source in genome may have more differences than alignment with some other copy of repeat.
- What if the closest fully sequenced genome is too divergent? (3% is a common assumed alignment capability)
- Placing reads in repetitive regions: Some algorithms only return 1 mapping; If multiple: map quality = 0
- Algorithms that use paired-end information => might prefer correct distance over correct alignment.

Intron/exon junctions

- In RNA-seq data, you must also consider splice junctions, reads may span an intron



Some Aligners

- Spliced Aligners
  - Tophat (Bowtie2)
  - GSNAP
  - SOAPsplice
  - MapSplice
  - TrueSite
  - star
- Aligners that can ’clip’
  - Bowtie2 in local mode
  - bwa-mem

https://en.wikipedia.org/wiki/List_of_sequence_alignment_software

Genome vs Transcriptome Reference

- May seem intuitive to map RNAseq data to transcriptome, but its not that simple.
  - Transcriptomes are rarely complete,
  - Which transcript, canonical transcript? Shouldn’t map to all splice variants as these would show up as multi-mappers
- More so, a mapper will do its best to map every read, somewhere, provided the result meets its minimum requirements.
  - Need to provide a mapper with all possible places the read could arisen from, which is best represented by the genome. Otherwise you get mismapping because its close enough.

Genome and Annotation

- Genome fasta files and Annotation files go together! Should be identified before beginning any analysis
  - Genome fasta files should include all primary chromosomes, unplaced sequences and un-localized sequences, as well as any organelles. Should not contain any contigs that represent patches or alternative haplotypes.
  - Annotation file should be GTF (preferred), and should be the most comprehensive you can find.
    - Chromosome names in the GTF should match those in the fasta, they don’t always do.
    - Star recommends the Gencode annotations for mouse/human

Preparing a sam file for counting and stats

- Samtools is used to manipulate mapping files for counting, common steps include:
  - samtools view [to convert from sam to bam]
  - samtools sort [possibly by read and not by position, htseq-count requirement]
  - samtools index
  - samtools idxstats
  - samtools flagstat
  - samtools stats
- Check with the counting application as to its input requirements.

QA/QC

- Mapper produce summary statistics, view the summary report (in a text editor) and compare across samples.
  - Other additional summary statistics can be produced with:

  ```
  samtools flagstat
  samtools idxstats
  samtools stats
  ```

- Produce a multi-dimensional scaling (MDS) plots of the summary files, the purpose is to look for patterns in the plot that are non-random, and may be influenced by technical artifacts

Estimate known genes and transcripts expression – Counting

Counting as a measure of expression

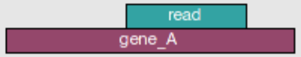
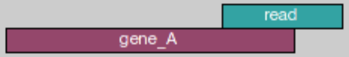
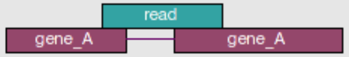
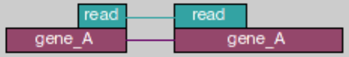
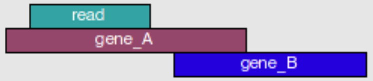
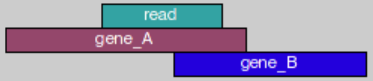
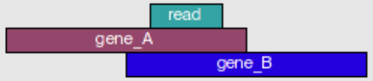
- The more you can count (and HTS sequencing systems can count a lot) the better the measure of copy number for even rare transcripts in a population.
  - Most RNA-seq techniques deal with count data. Reads are mapped to a reference genome, transcripts are detected, and the number of reads that map to a transcript (or gene) are counted.
  - Read counts for a transcript are roughly proportional to the gene’s length and transcript abundance.
- technical artifacts should be considered during counting
  - mapping quality
  - mapability (uniqueness), the read is not ambiguous

Read Counting with HTSEQ-COUNT

Problem:

- Given a sam/bam file with aligned sequence reads and a list of genomic feature (genes locations), we wish to count the number of reads (fragments) than overlap each feature.
  - Features are defined by intervals, they have a start and stop position on a chromosome.
  - For this workshop and analysis, features are genes which are the union of all its exons. You could consider each exon as a feature, for alternative splicing.
- Htseq-count has three overlapping modes
  - union:
  - intersection-strict
  - intersection-nonempty

Htseq-count

	union	intersection _strict	intersection _nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

Counting genes -- STAR

- Counts coincide with Htseq-counts under default parameters. Need to specify GTF file at genome generation step or during mapping.
- Output, 4 columns
  - GeneID
  - Counts for unstranded
  - Counts for first read strand
  - Counts for the second read strand
- Chose the columns that makes sense and generate a matrix table, columns are sample, rows are genes.

QA/QC

- View summary report (in a text editor)
- Produce a multi-dimensional scaling (MDS) plots of the summary files, the purpose is to look for patterns in the plot that are non-random, and may be influenced by technical means.
- Statistics such as:
  - % Multimapped reads
  - % Uniquely mapped reads
  - Splice sites
  - Unmapped
  - Chimeric
  - Etc.

Differential Expression Analysis using edgeR/Limma Voom

Differential Expression Analysis

- Differential Expression between conditions is determined from count data, which is modeled by a distribution (ie. Negative Binomial Distribution, Poisson, etc.)
- Generally speaking differential expression analysis is performed in a very similar manner to DNA microarrays, once and normalization have been performed.
- A lot of RNA-seq analysis has been done in R and so there are many packages available to analyze and view this data. Two of the 'best' are:
  - DESeq, developed by Simon Anders (also created htseq) in Wolfgang Huber's group at EMBL
  - edgeR/Voom (extension to Limma [microarrays] for RNA-seq), developed out of Gordon Smyth's group from the Walter and Eliza Hall Institute of Medical Research in Australia
  - http://bioconductor.org/packages/release/BiocViews.html#___RNASeq

Basic steps procedure – edgeR/limma voom

1. Read the count data in
2. Filter genes(uninteresting genes, e.g. unexpressed)
3. Calculate normalizing factors (sample-specific adjustment)
4. Calculate dispersion (gene-gene variance-stabilizing transformation) [edgeR]
5. Fit a model of your experiment
6. Perform likelihood ratio tests on comparisons of interest (using contrasts)
7. Adjust for multiple testing, Benjamini-Hochberg (BH) is the defaults.
8. Check results for confidence
9. Attach annotation if available and write tables

Filtering genes

- Most common filter is to ***remove*** genes that are less then X reads counts across a certain number of samples. EX.
  - rowSums(cpms <= 1) < 3 , require at least 1 cpm in at least 3 samples to keep

- A second less used filter to is minimum variance across all samples, so if a gene isn’t changing (constant expression) its not interesting no need to test.

NORMALIZATION

- In differential expression analysis, only sample-specific effects need to be normalized, NOT concerned with comparisons and quantification of absolute expression.
  - Sequence depth – is a sample specific effect and needs to be adjusted for.
  - RNA composition - finding a set of scaling factors for the library sizes that minimize the log-fold changes between the samples for most genes (uses a trimmed mean of M-values between each pair of sample)
  - GC content – is NOT sample-specific (except when it is)
  - Gene Length – is NOT sample-specific (except when it is)

- Normalization in edgeR/Voom is model-based, you calculate normalizing factors using the function calcNormFactors function which by default uses TMM (trimmed means of M values).

RPKM vs FPKM vs CPM vs model based

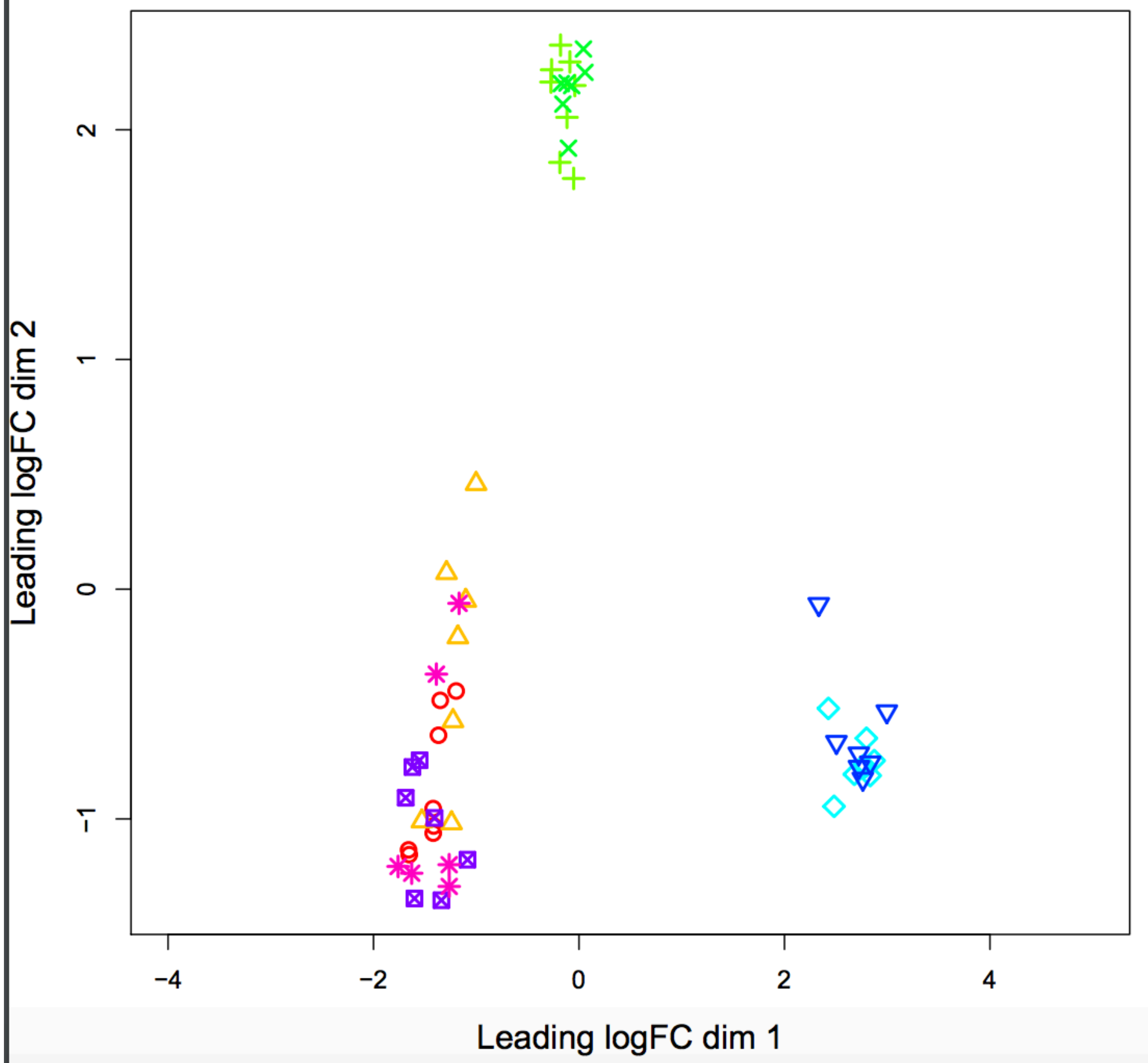
- RPKM - Reads per kilobase per million mapped reads
- FPKM - Fragments per kilobase per million mapped reads
- CPM/TPM – Counts per million [after logging good for producing MDS plots, estimation of normalized values in model based]
- Model based - original read counts are not themselves transformed, but rather correction factors are used in the DE model itself.

Transformation

- Transformation turn the gene count data into a distribution more suitable for statistical analysis (more ”normal” like).
- Most common
  - Limma-trended transformation, logCPM, best when total counts per sample are relatively close to each other (~3-fold)
  - Voom, is best used when library sizes are quite variable across samples

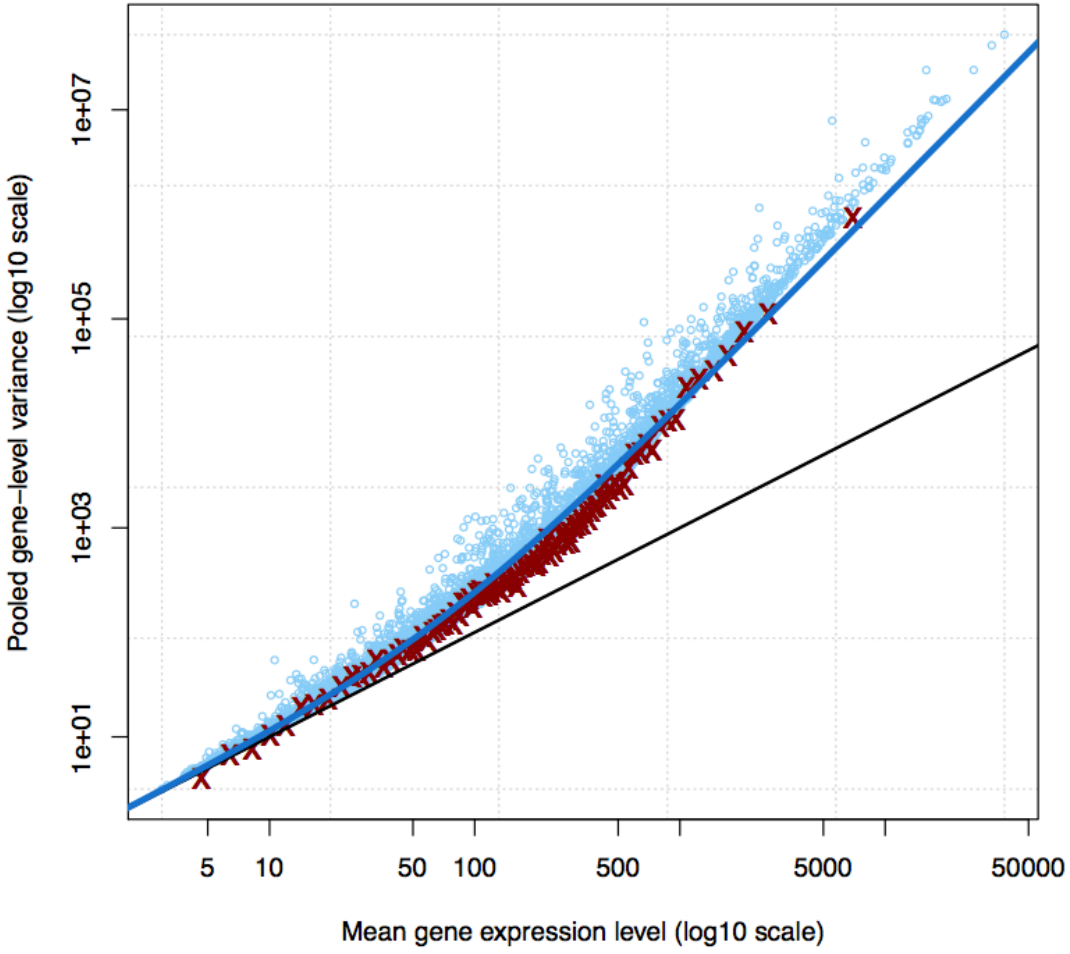
QA/QC

- MDS plots of logCPM



Variance Stabilization - eBayes

- The variance characteristics of low expressed genes are different from high expressed genes, if treated the same, the effect is to over represent low expressed genes in the DE list.



Multiple testing correction

- Simply a must! Best choices are
    - FDR (false discovery rate)
    - qvalue
- The FDR (or qvalue) is a statement about the list and no longer about the gene. So a FDR 0.05, says you expect 5% false positives in the list of genes with an FDR of 0.05 and less.
- The statement “Statistically significant” **means** FDR of 0.05 or less.
    - My opinion is these genes do not require further validation (eg qrtPCR)
    - You can dip below FDR 0.05, but in my opinion you then need to validate those genes.

EdgeR/Limma Manual

- Both edgeR and limma voom have VERY comprehensive user manuals
  - Limma voom https://bioconductor.org/packages/release/bioc/vignettes/limma/inst/doc/usersguide.pdf
  - edgeR http://bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf

Summarization and Visualization

The Top Table

- The basic table
  - Gene_ID: The Gene Id from the GTF file
  - logFC: log fold change, positive values indicate up-regulation, negative numbers indicate down-regulation
  - logCPM: log counts per million, average ‘expression’ value of the gene
  - LR: log ratio of the test (ignore)
  - Pvalue: raw p-value for that gene (best to sort on)
  - FDR: false discover rate for that gene
- Annotation is added in additional columns (must first uncomment the line to do so in the R script

Visualization and Next step tools

Visualization

1. Integrated Genome Viewer (https://www.broadinstitute.org/igv/)

Further Annotation of Genes

1. DAVID (http://david.abcc.ncifcrf.gov/tools.jsp)
2. ConsensusPathdb (http://cpdb.molgen.mpg.de/)
3. NetGestalt (http://www.netgestalt.org/)
4. Molecular Signatures Database (http://www.netgestalt.org/)
5. PANTHER (http://www.pantherdb.org/)
6. Cognoscente (http://vanburenlab.medicine.tamhsc.edu/cognoscente.shtml)
7. Pathway Commons (http://www.pathwaycommons.org/)
8. Readctome (http://www.reactome.org/)
9. PathVisio (http://www.pathvisio.org/)
10. Moksiskaan (http://csbi.ltdk.helsinki.fi/moksiskaan/)
11. Weighed Gene Co-Expression Network Analysis (WGCNA)s
12. More tools in R Bioconductor

Gene Set enrichment analysis (GSEA) And GO/Pathway Enrichment

Gene set enrichment analysis

- A computational method that determines whether an a priori set of genes (e.g. gene ontology group, or pathway) shows statistically significant, concordant differences between two biological states (e.g. phenotypes)

Gene Ontology/Pathways enrichment analysis

- Given a set of genes that are up-regulated, which gene ontologies or pathways are over-represented (or under-represented) using annotations for that gene set.

Software

Preprocessing:

- Python 2.7
  - Modules: argparse, optparse, distutils
- bowtie2 - contaminant screening
  - http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
- Super-Deduper – Identify and remove PCR duplicates
  - https://github.com/dstreett/Super-Deduper
- Sickle – Trim low quality regions
  - https://github.com/dstreett/sickle
- Scythe – Identify and remove adapters in SE reads
  - https://github.com/ucdavis-bioinformatics/scythe
- FLASH2 – Join overlapping reads, identify and remove adapter in PE reads
  - https://github.com/dstreett/FLASH2

Software

Mapping:

- Bwa mem – map reads to a reference
  - http://sourceforge.net/projects/bio-bwa/files/
- samtools – processing of sam/bam file
  - http://www.htslib.org/

Read Counting:

- samtools – processing of sam/bam file
  - http://www.htslib.org/
- HTeq-0.6.1 htseq_count – count reads occurrences within genes
  - http://www-huber.embl.de/users/anders/HTSeq/

OR simultaneous read mapping and counting:

- Star
  - https://github.com/alexdobin/STAR [performs both alignment and counting]

Software

Analysis of differential expression:

- R http://www.r-project.org/
  - R Packages: EdgeR, limma from bioconductor – differential expression analysis
    - http://bioconductor.org/packages/release/bioc/html/edgeR.html
    - http://bioconductor.org/packages/release/bioc/html/limma.html
    - https://genomebiology.biomedcentral.com/articles/10.1186/gb-2014-15-2-r29
- RStudio
  - https://www.rstudio.com/