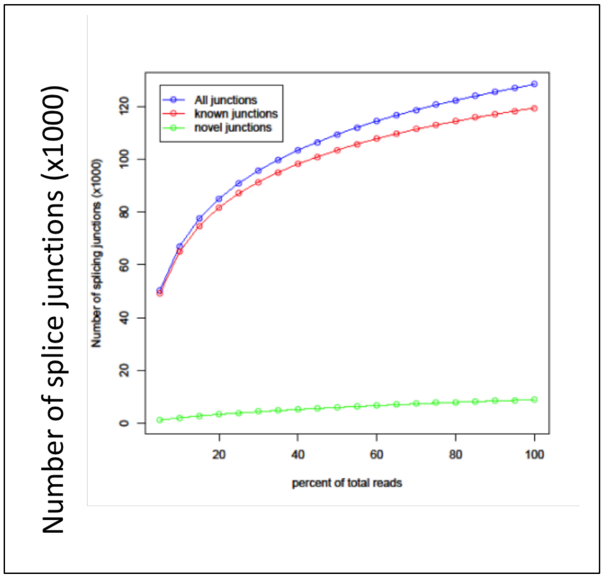
QC of RNAseq data from cell/tissue populations (not single cells)

Jain Lab and Eric Tycksen (GTAC, Washington University)

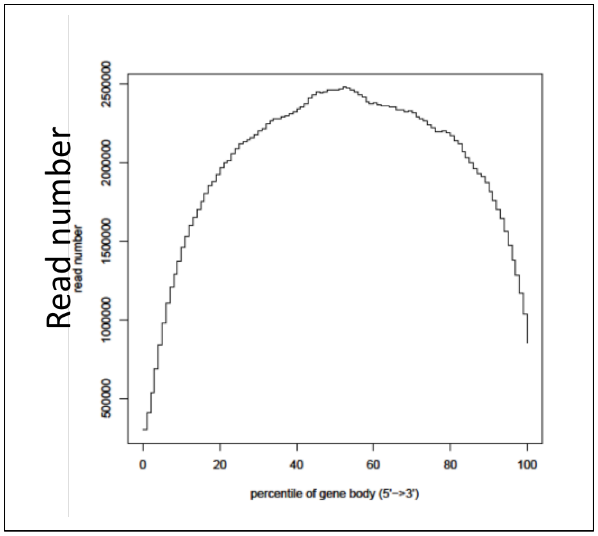
1) **Biological replicates:** at least 3 (preferably each from pool of different mice or litters).
2) **Total reads:** > 20 million single end, if planning isoform analysis then >20 million paired-end reads. Create junction saturation plots with RSeQC junction_saturation.py to plot of the coverage over known and novel splicing junctions. When the known/novel junctions plateau, the plot indicates that the sensitivity for known/novel junctions has been maximized and continued sequencing would only increase coverage rather than sensitivity. If the plot shows no signs of plateauing for the known/novel junctions, the plot indicates that low and low-medium expressors are likely under-sampled and the abundances measured will not reflect the true distribution of expressed genes or transcripts. Continued sequencing will increase sensitivity and would likely prove beneficial for all downstream analysis such as gene or transcript level differential expression.



3) **Ribosomal fraction:** < 1% (unless goal is to study ribosomal transcripts)

4) **End bias:** Use RSeQC geneBody_coverarge.py to plot the 3 prime and 5 prime bias averaged across all known genes for a given sample normalized to 100bp gene bodies. Ideally, the plot should show high and consistent coverage from the 3 prime to the 5 prime ends that appears as a large plateau over the normalized 100bp gene body. High coverage at one end and low coverage at the opposite end indicates end bias during sample and library preparation. Inconsistent spikes in coverage across the gene body

with low coverage in between spikes indicates sample degradation or library preparation artifacts that are typical with non-strand displacing random priming library preparation such as the Sigma kit for low input or degraded FFPE samples.



5) **Quality of replicates:** Spearman correlation among biological replicates > 0.9.

6) **Batch effects:** Principal component analysis or Multidimensional scaling can be used to ensure that the samples don't cluster based on the library prep or NGS runs. It is advisable to spread out the control and experimental samples if more than one batch would be run at different times.

Our pipeline summer 2015

The analysis methods are continuously evolving, some notes are provided to give brief pointers on methods that were the standard several years ago and their shortfalls. For analysis, all samples should be subjected to the same processing pipeline using the raw fastq files. The following are methods that are now the current accepted standard according to the ABRF and RNASeQC consortium studies [Li et al, Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. *Nat Biotechnol*. 2014 Sep;32(9):915-25; SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat Biotechnol.* 2014 Sep;32(9):903-14. doi: 10.1038/nbt.2957. Epub 2014 Aug 24.
PMID: 25150838].

1) Accept only raw demultiplexed fastq files with notes for what adapters were used (needed for trimming).

2) Align all fastq files with STAR to the latest version of Ensembl (GRCm38.p2 assembly)

Note: Because new methods come out every year, always keep the fastq files available since that will allow the database to be updated and the Ensembl ID's can always be kept current with biomart (Ensembl ID's are conserved between assemblies and dropped entirely and never reused again if the gene or transcript is dropped by the database).

Note: STAR has the advantage of performing local re-alignments, alignment of segmented reads without using a fixed segment size (as is the casefor Tophat), local re-alignment of the aligned segments, and automatic soft-clipping of low quality bases and adapters, and the added ability of aligning to chimeric genes. Moreover, we replaced HTSeq with Subread:featureCounts that is much faster and has the ability to enumerate any chimeric read found by Star.

3) Enumerate all known genes with raw counts to the matching gtf file from the same reference build from Ensembl with Subread:featureCounts using the Ensembl gene ID as the key.

4) Use the Sailfish or Salmon package to estimate known transcript/isoform expression if desired.

5) Use the counts listed in the output in the Subread:featureCounts files to generate Countsper-Million (CPM) normalized values for heat map generation.
Note: RPKM values can also be created if desired, but they are very susceptible to bias gene length bias and ribosomal contamination.

6) Load all raw counts and CPM values along with pertinent sample and experiment information into a SQL database that can then be queried online or through an R/Bioconductor interface to create and download matrices reads for differential expression analysis or plotting using a R Shiny server with built in Ensembl annotations available through the R/Bioconductor Biomart package.

The raw counts will be used for differential expression analysis, the CPM values will be used for plotting, and the fastq files can be maintained in a secure FTP archive.

Note 1: Differential expression can be done using EdgeR to calculate library size scaling factors and normalize the counts between libraries using the TMM method and then calculate differential expression using a negative-binomial generalized linear model.

Note 2: Differential Expression analysis currently preferred is using the Limma-Voom method that is essentially a conglomeration of EdgeR for RNA-seq and Limma for microarrays that were both published by Gordon Smyth et. Al. The results are nearly identical to our current EdgeR method, but it accommodates more complex experimental designs.

TOOLS:

STAR: https://github.com/alexdobin/STAR/releases

Subread: http://sourceforge.net/projects/subread/

RSeQC: http://rseqc.sourceforge.net

Sailfish/Salmon: https://github.com/kingsfordgroup/sailfish