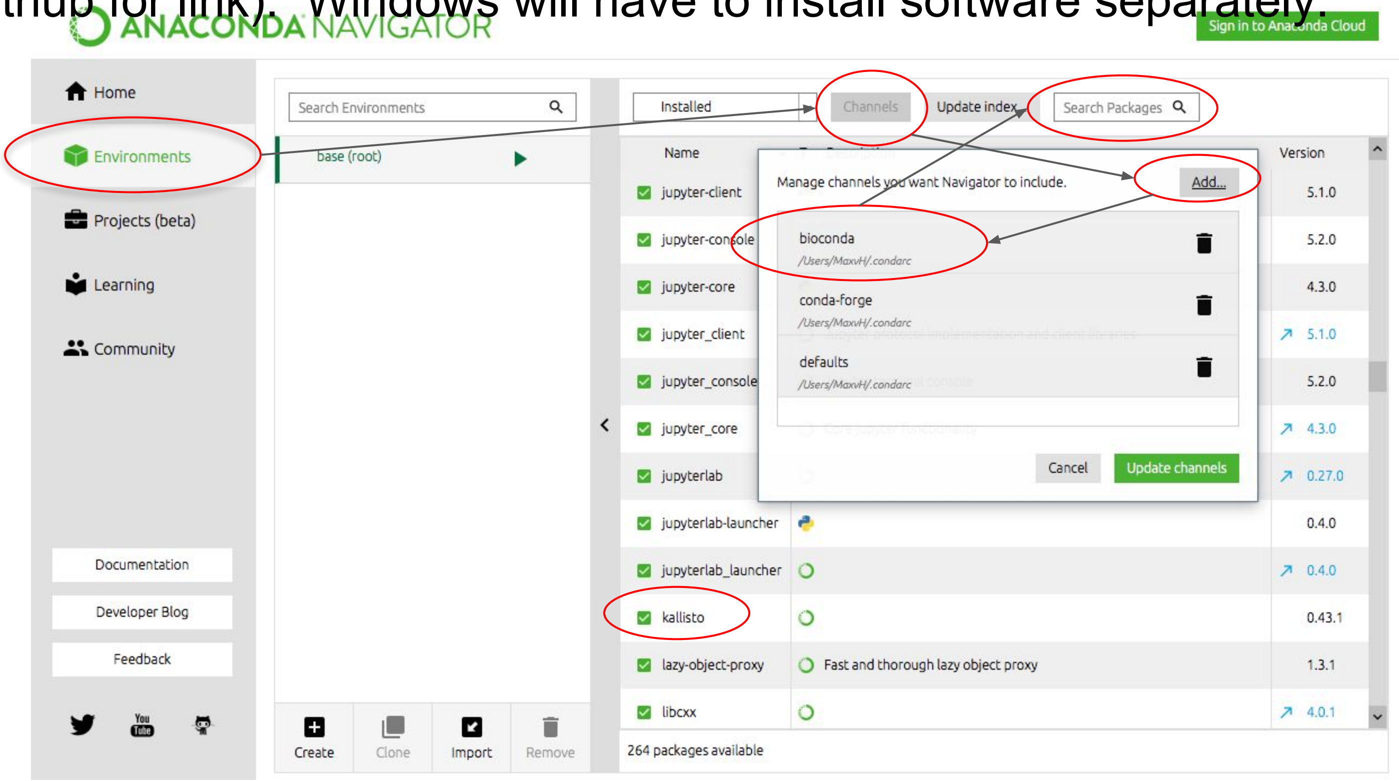
RNAseq Analysis

If you are on either Mac or Linux, start with installing Anaconda navigator (see github for link). Windows will have to install software separately.



Manage channels you want Navigator to include.

bioconda
/Users/MaxvH/.condarc

conda-forge
/Users/MaxvH/.condarc

defaults
/Users/MaxvH/.condarc

Name	Description	Version
☑ jupyter-client		5.1.0
☑ jupyter-console		5.2.0
☑ jupyter-core		4.3.0
☑ jupyter_client		↗ 5.1.0
☑ jupyter_console		5.2.0
☑ jupyter_core		↗ 4.3.0
☑ jupyterlab		↗ 0.27.0
☑ jupyterlab-launcher		0.4.0
☑ jupyterlab_launcher	○	↗ 0.4.0
☑ kallisto	○	0.43.1
☑ lazy-object-proxy	○ Fast and thorough lazy object proxy	1.3.1
☑ libcxx	○	↗ 4.0.1

264 packages available

link to github lesson: **https://tinyurl.com/y9msejcf** **OR**
https://github.com/ScienceParkStudyGroup/studyGroup/tree/gh-pages/lessons/20180220_RNA-Seq_Max_and_Fred

Experimental design

-What do you want to test?
-Will your samples give you your answers?
-cell type, genes/repeats/both?

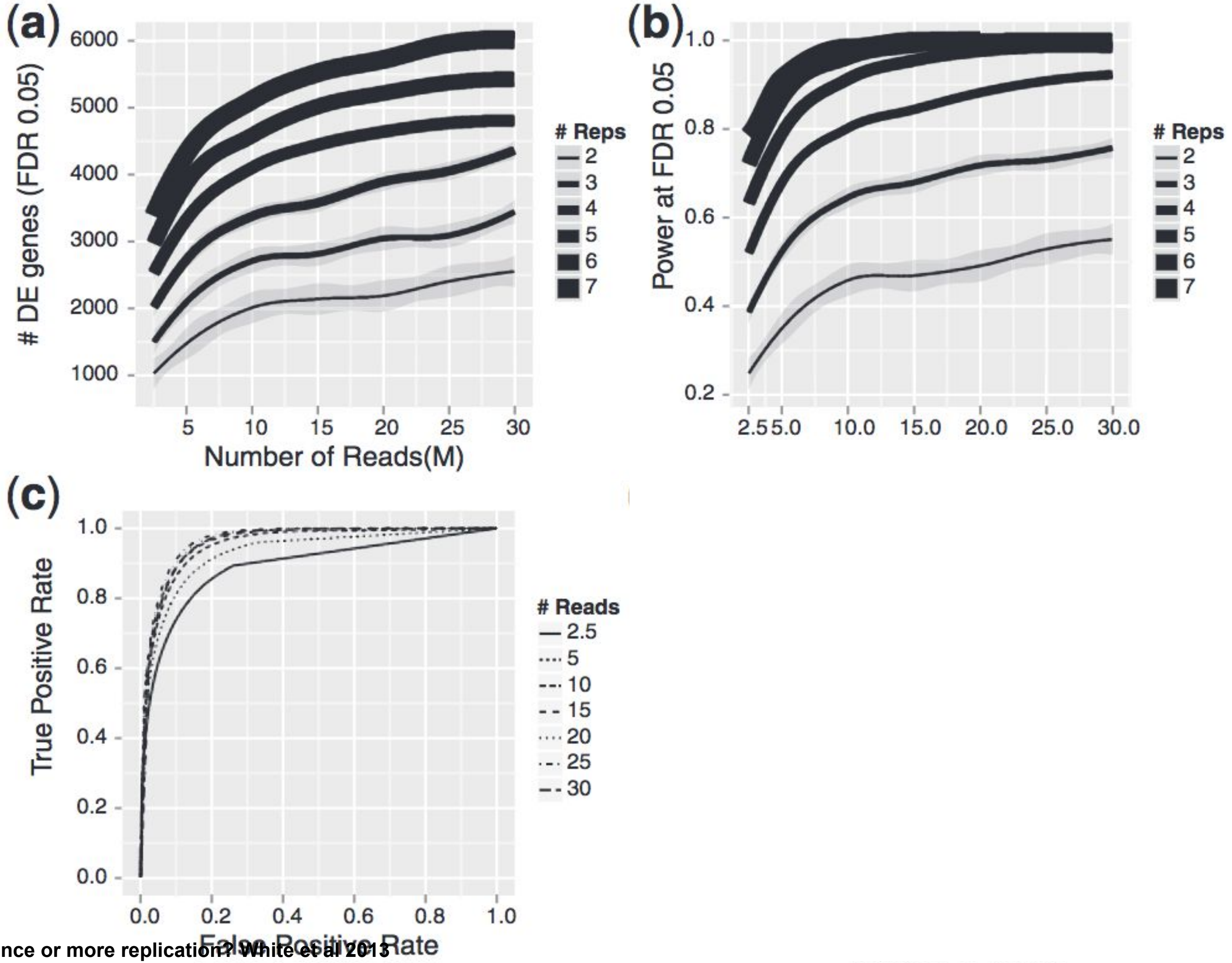
How many samples can you afford

Sequencing platform

<u>short reads</u>
-Illumina
-SOliD

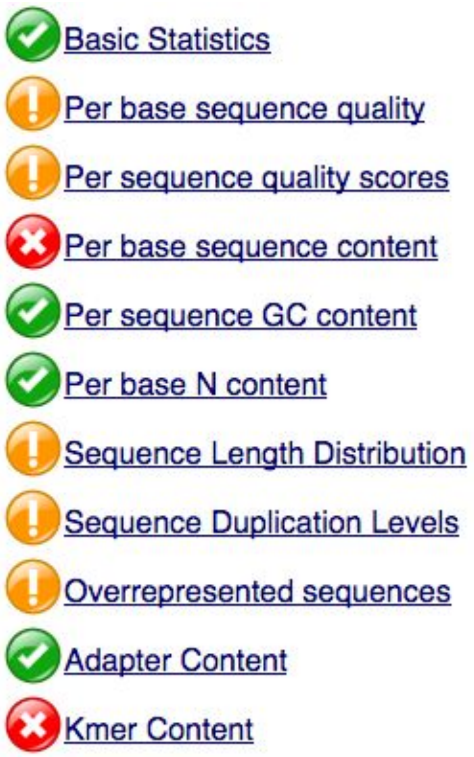
<u>Long reads</u>
-Ion torrent
-Pac bio
-roche 454

# samples matter more than # of reads

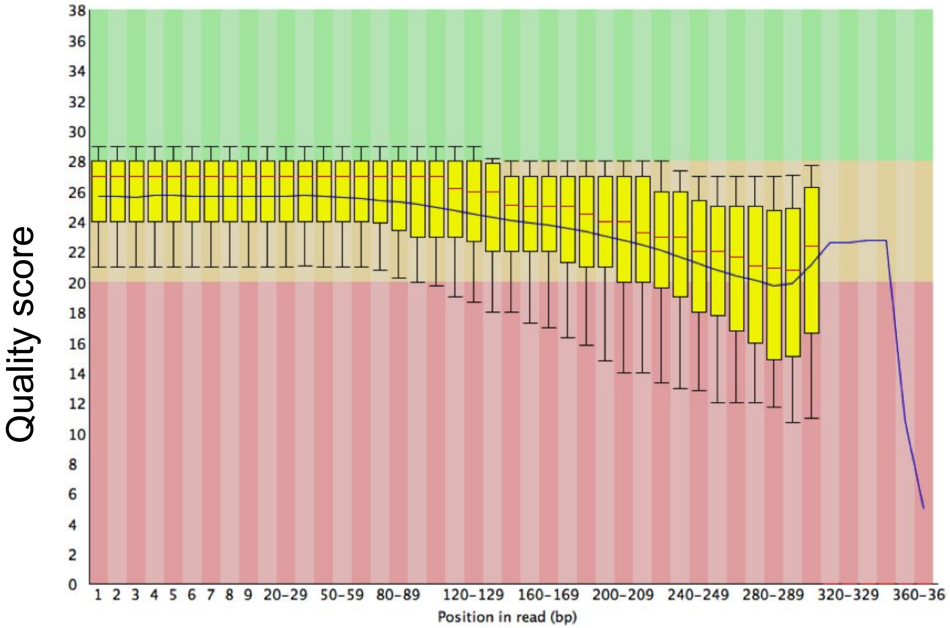


RNA-seq differential expression studies: more sequence or more replication? White et al 2013

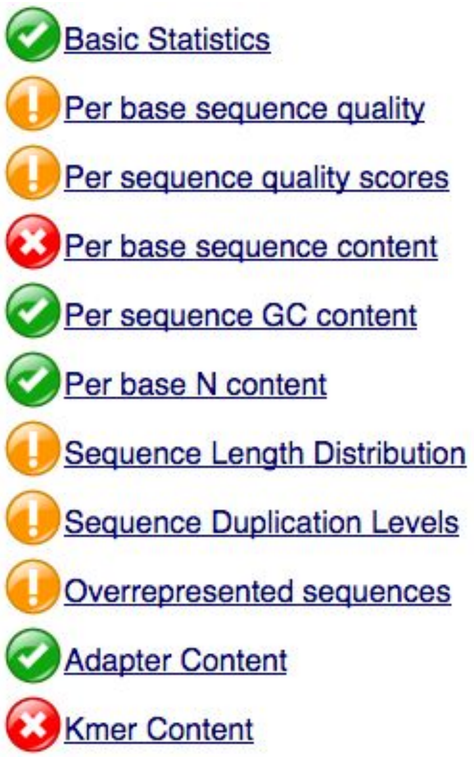
FastQC



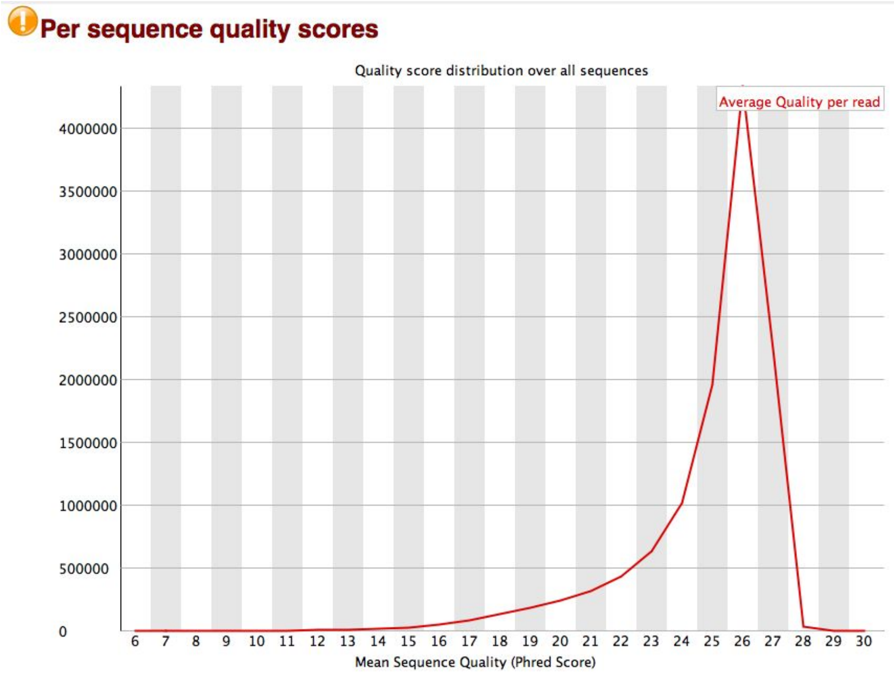
- ✓ Basic Statistics
- ! Per base sequence quality
- ! Per sequence quality scores
- ✗ Per base sequence content
- ✓ Per sequence GC content
- ✓ Per base N content
- ! Sequence Length Distribution
- ! Sequence Duplication Levels
- ! Overrepresented sequences
- ✓ Adapter Content
- ✗ Kmer Content



FastQC

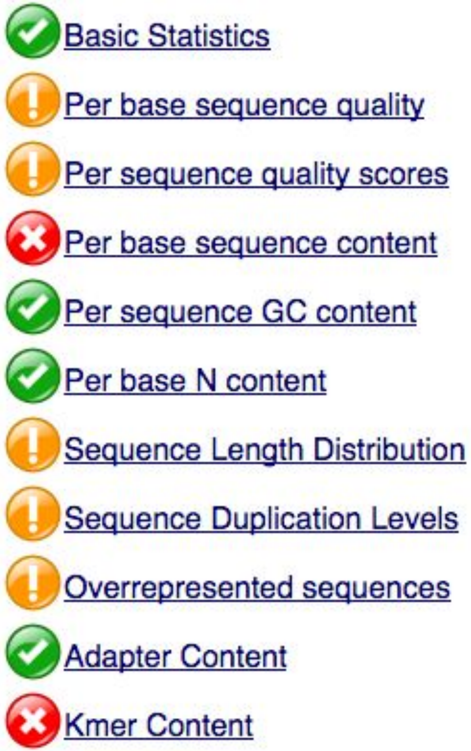


- ✓ Basic Statistics
- ! Per base sequence quality
- ! Per sequence quality scores
- ✗ Per base sequence content
- ✓ Per sequence GC content
- ✓ Per base N content
- ! Sequence Length Distribution
- ! Sequence Duplication Levels
- ! Overrepresented sequences
- ✓ Adapter Content
- ✗ Kmer Content

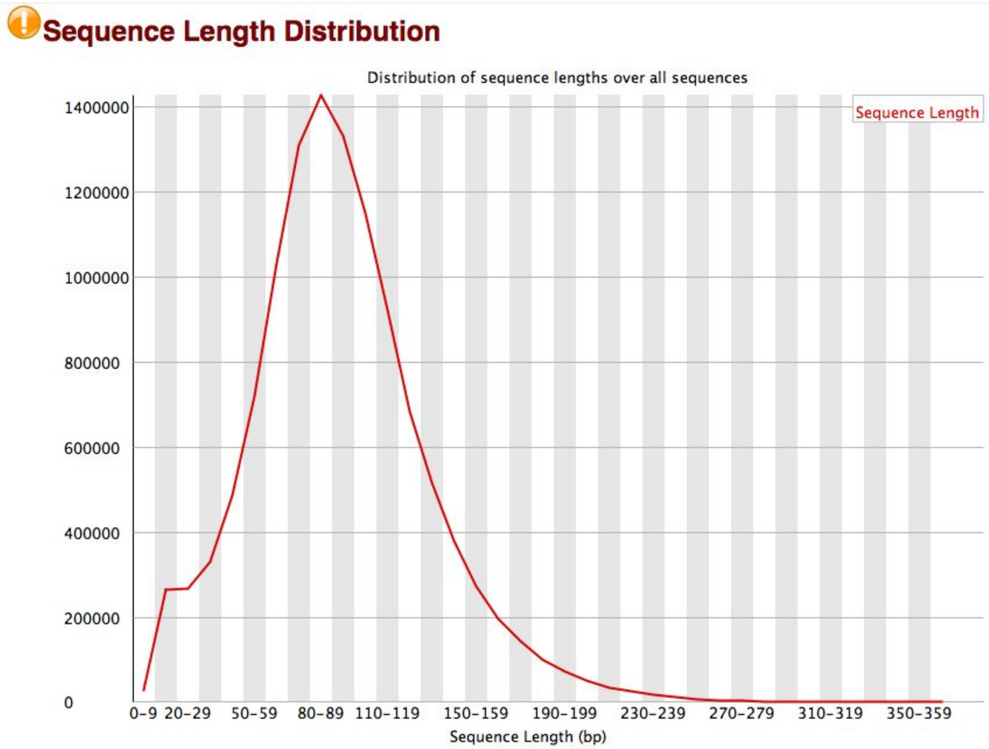


! Per sequence quality scores

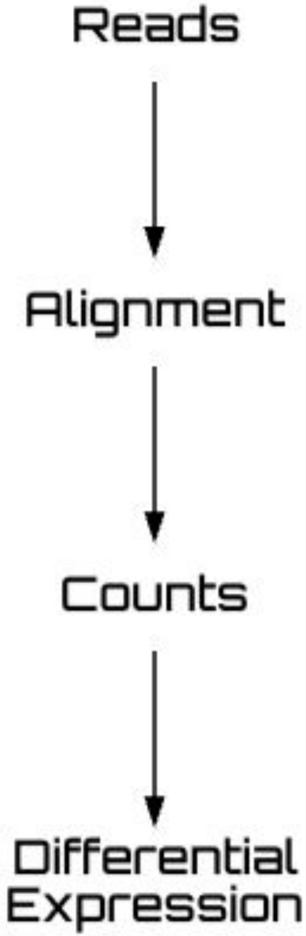
FastQC



- ✓ Basic Statistics
- ! Per base sequence quality
- ! Per sequence quality scores
- ✗ Per base sequence content
- ✓ Per sequence GC content
- ✓ Per base N content
- ! Sequence Length Distribution
- ! Sequence Duplication Levels
- ! Overrepresented sequences
- ✓ Adapter Content
- ✗ Kmer Content



! Sequence Length Distribution



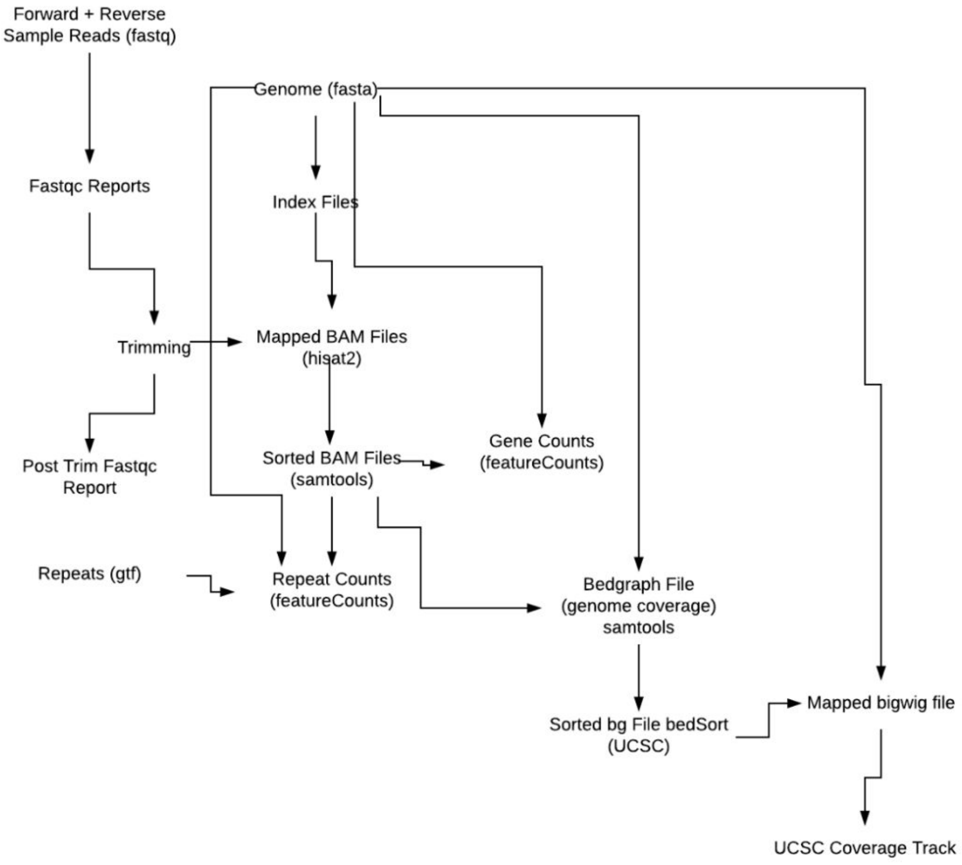
Reads

Alignment

Counts

Differential Expression

Workflow Model



Forward + Reverse Sample Reads (fastq)

Fastqc Reports

Trimming

Post Trim Fastqc Report

Genome (fasta)

Index Files

Mapped BAM Files (hisat2)

Sorted BAM Files (samtools)

Gene Counts (featureCounts)

Repeats (gtf)

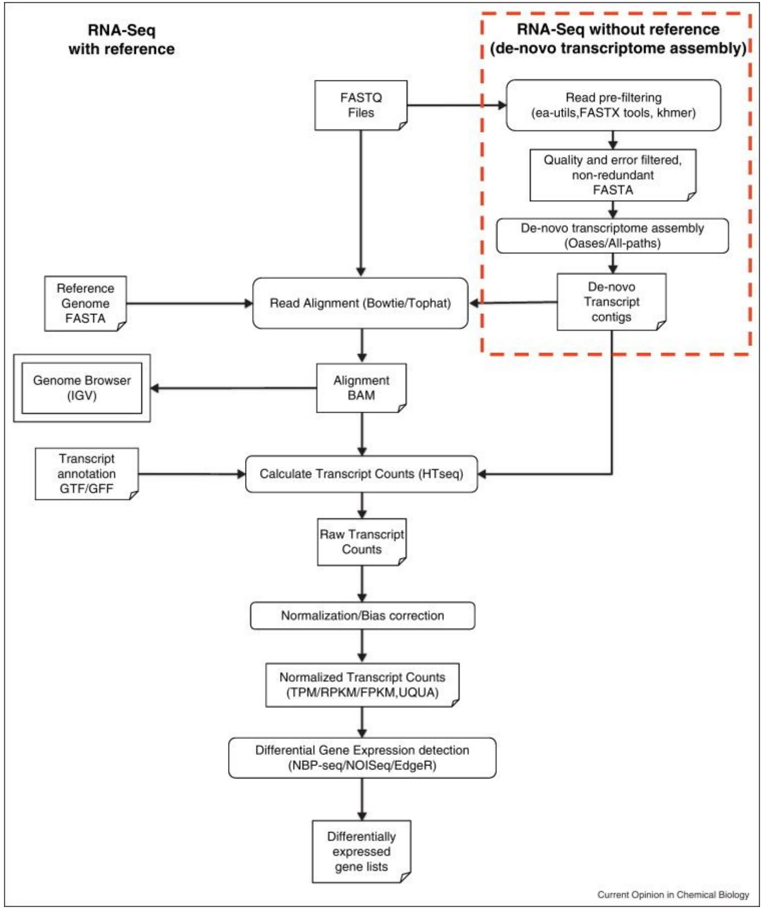
Repeat Counts (featureCounts)

Bedgraph File (genome coverage) samtools

Sorted bg File bedSort (UCSC)

Mapped bigwig file

UCSC Coverage Track



RNA-Seq with reference
RNA-Seq without reference (de-novo transcriptome assembly)
FASTQ Files
Read pre-filtering (ea-utils,FASTX tools, khmer)
Quality and error filtered, non-redundant FASTA
De-novo transcriptome assembly (Oases/All-paths)
De-novo Transcript contigs
Reference Genome FASTA
Read Alignment (Bowtie/Tophat)
Genome Browser (IGV)
Alignment BAM
Transcript annotation GTF/GFF
Calculate Transcript Counts (HTseq)
Raw Transcript Counts
Normalization/Bias correction
Normalized Transcript Counts (TPM/RPKM/FPKM,UQUA)
Differential Gene Expression detection (NBP-seq/NOISeq/EdgeR)
Differentially expressed gene lists

Current Opinion in Chemical Biology

Metrics to consider when choosing the aligner

- On-target hits
- False positives
- User time
- Memory Consumption
- Splice variant considerations
- Analysing for genes or transposable elements

Mapper	**Pros**	**Cons**
Bowtie/Tophat	Widely used, Developed for splice variants, Output Cufflinks compatible	Time, Memory Consumption
STAR	Widely used, Used by ENCODE, Clear documentation	Time, Memory Consumption
Hisat	Widely used, Works well with Cufflinks	Conservative hit rate, Time, Memory Consumption
Kallisto	Speed, Accuracy, Ease of use, Specifically developed to circumvent multimapping problem (i.e. splicing and repeats)	Small user base

Differential Expression	Pros	Cons
EdgeR	Can be used without replicates (with caution)	Works with CPM
Cuffdiff	Widely used, Combined with CummeRbund (R package)	Blackbox
DESeq2	Can be used without replicates (with caution)	Input must be integer
Sleuth	Ease of use, Optimised for small no. of replicates,	Small user base

Aligners or pseudo-aligners

- Genome alignments are time consuming and do not always result in optimised mapping, especially for repeats.
- Advances in bioinformatic algorithms have attempted to circumvent this problem by pseudo-mapping to a transcriptome to quantify transcript abundance directly from the raw sequence reads (fastq files)
- Kallisto: a recent update in RNAseq analysis

Pseudo aligners - Kallisto

- Kallisto/Sleuth analysis has a similar workflow
- BUT alignment to a transcriptome fasta file
- Speedier analysis as not forced to properly align to the entire genome

How does Kallisto Work?

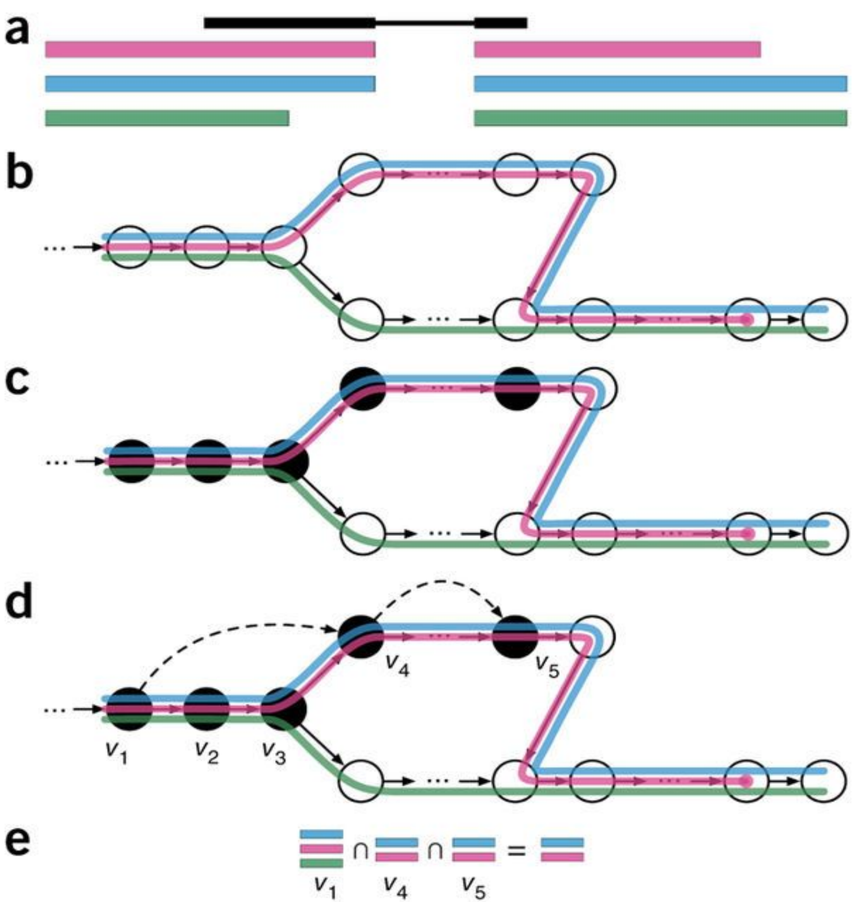


Figure 2 *Kallisto Schematic*
(Bray, Pimentel, Melsted & Pachter, 2016)

Example Kallisto Run

(a) Read example, with three overlapping potential transcripts to map to.

(b) **I**ndex made through T-DBG, nodes are k-mers**,** coloured paths = transcripts. The paths create a k-compatibility class over the nodes per k-mer.

(c) K-mers are hashed in order to find this k-compatibility.

(d) If a k-mer is redundant i.e. has the same k-compatibility to each node then the hashing skips to the next k-mer of the sequence (redundancy is marked as dotted lines).

(e) K-compatibility class of each read is determined by the intersection of the compatibility classes of its constituent k-mers.

References

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nature biotechnology, 34(5), 525.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., ... & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. Genome biology, 17(1), 13.

Otto C, Stadler PF, Hoffmann S: 'Lacking alignments? The next generation sequencing mapper segemehl revisited', Bioinformatics. 2014 Jul 1;30(13):1837-43 (2014)

Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome biology, 15(2), R29.

Love, M.I., Huber, W., Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15:550. 10.1186/s13059-014-0550-8

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nature methods, 14(4), 417.

Pimentel, H., Bray, N. L., Puente, S., Melsted, P., & Pachter, L. (2017). Differential analysis of RNA-seq incorporating quantification uncertainty. Nature methods, 14(7), 687.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic acids research, 43(7), e47-e47.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26(1), 139-140.

Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics, 25(9), 1105-1111.

Useful Link

Kallisto/Sleuth support group:

https://groups.google.com/forum/#!forum/kallisto-sleuth-users