Guideline for RNAseq analysis (Eckart Stolle, ZFMK Bonn)

How many / which reads do I need?

Depends on your goal:

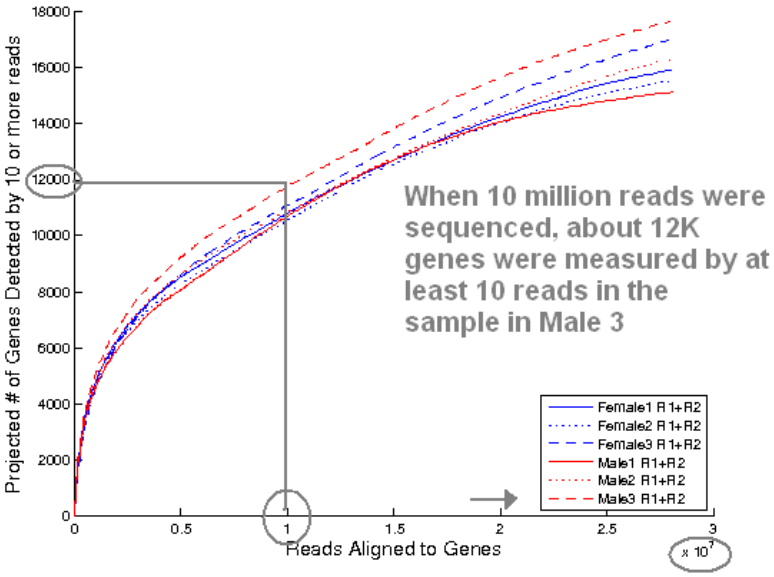
- Gene expression profiling / quick snapshot of highly expressed genes may only need 5–25 million reads per sample (source: Illumina).
- Experiments looking for a more global view of gene expression, and some information on alternative splicing, typically require 30–60 million reads per sample, for an in-depth view even 100–200 million reads
- miRNA-Seq or small RNA Analysis experiments may require fewer reads than whole transcriptome sequencing
- **Transcriptome assembly (>50M reads recommended):** Trinity performs well (see: Martin Hölzer, Manja Marz (2019): De novo transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers. GigaScience, Volume 8, Issue 5, May 2019, giz039, https://doi.org/10.1093/gigascience/giz039)
- Single end and shorter read length (e.g. 50, 75 or 100 bp) may suffice to obtain read counts, at the disadvantage of reliable mapping (especially important with very similar paralogs), hence the general recommendation: paired end, 150 bp read. (exception: small RNAs: a single 50bp read typically covers the entire sequence).

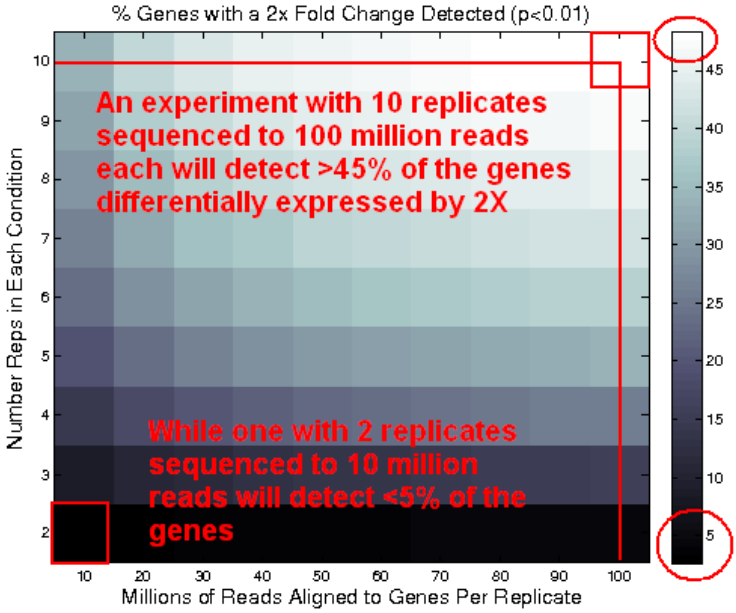
Depth vs Replicates?

- 7 replicates or more give good power to detect differential expression
- 10M reads may be insufficient except when aiming at highly expressed genes or larger expression differences. Otherwise 20M or more is recommended.

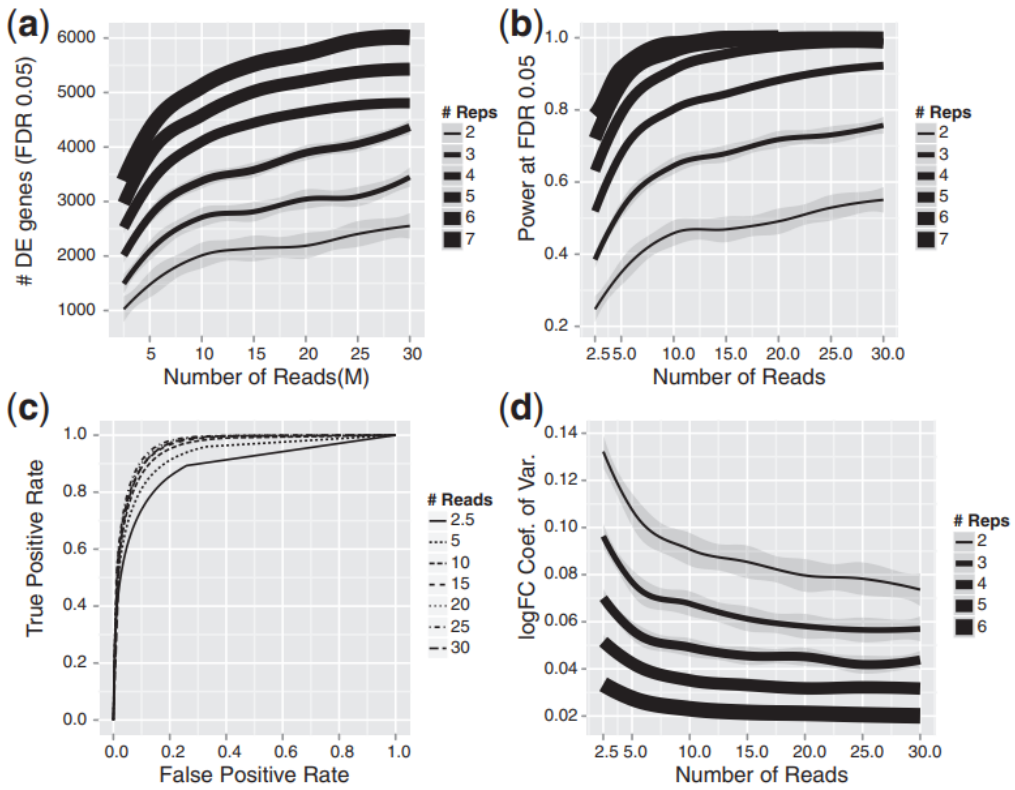
Consider these:

- Scotty: A Web Tool For Designing RNA-Seq Experiments to Measure Differential Gene Expression. M.A. Busby; C. Stewart; C. Miller; K. Grzeda; G. Marth Bioinformatics 2013; doi: 10.1093/bioinformatics/btt015 http://scotty.genetics.utah.edu/help.html





- Schurch et al. (2016): How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? RNA. 2016 Jun; 22(6): 839–851. doi: 10.1261/rna.053959.115 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4878611
- Baccarella et al. (2018): Empirical assessment of the impact of sample number and read depth on RNA-Seq analysis workflow performance. BMC Bioinformatics volume 19, Article number: 423. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-018-2445-2
- Lui et al. (2014): RNA-seq differential expression studies: more sequence or more replication? Bioinformatics, Volume 30, Issue 3, 1 February 2014, Pages 301–304, https://doi.org/10.1093/bioinformatics/btt688 https://academic.oup.com/bioinformatics/article/30/3/301/228651


(a)
DE genes (FDR 0.05)
6000
5000
4000
3000
2000
1000
5
10
15
20
25
30
Number of Reads(M)
Reps
2
3
4
5
6
7
(b)
Power at FDR 0.05
1.0
0.8
0.6
0.4
0.2
2.5
5.0
10.0
15.0
20.0
25.0
30.0
Number of Reads
Reps
2
3
4
5
6
7
(c)
True Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
Reads
2.5
5
10
15
20
25
30
(d)
logFC Coef. of Var.
0.14
0.12
0.10
0.08
0.06
0.04
0.02
2.5
5.0
10.0
15.0
20.0
25.0
30.0
Number of Reads
Reps
2
3
4
5
6

polyA enrichment vs ribo-depletion?

- polyA enrichment (mRNAs) is well established/cheap, but omits non-polyA-tailed RNAs (e.g. some long non coding RNAs and many TEs)
- depletion of ribosomal RNA from total RNA ideally needs species-specific probes. For some insects (e.g. bumblebees) it works well to use the Illumina RiboZero Kit for Human/Mouse/Rat. There are also homemade solutions, e.g. Kraus et al. 2019: https://www.nature.com/articles/s41598-019-48692-2 or https://journals.asm.org/doi/10.1128/mBio.00010-20 .
- post-library prep depletion of rRNA and other custom / abundant RNAs: NuGen AnyDeplete https://www.nugen.com/anydeplete-rrna-depletion