BIOINFORMATICS DISCOVERY NOTE

Vol. 30 no. 3 2014, pages 301–304
doi:10.1093/bioinformatics/btt688

Gene expression

Advance Access publication December 6, 2013

RNA-seq differential expression studies: more sequence or more replication?

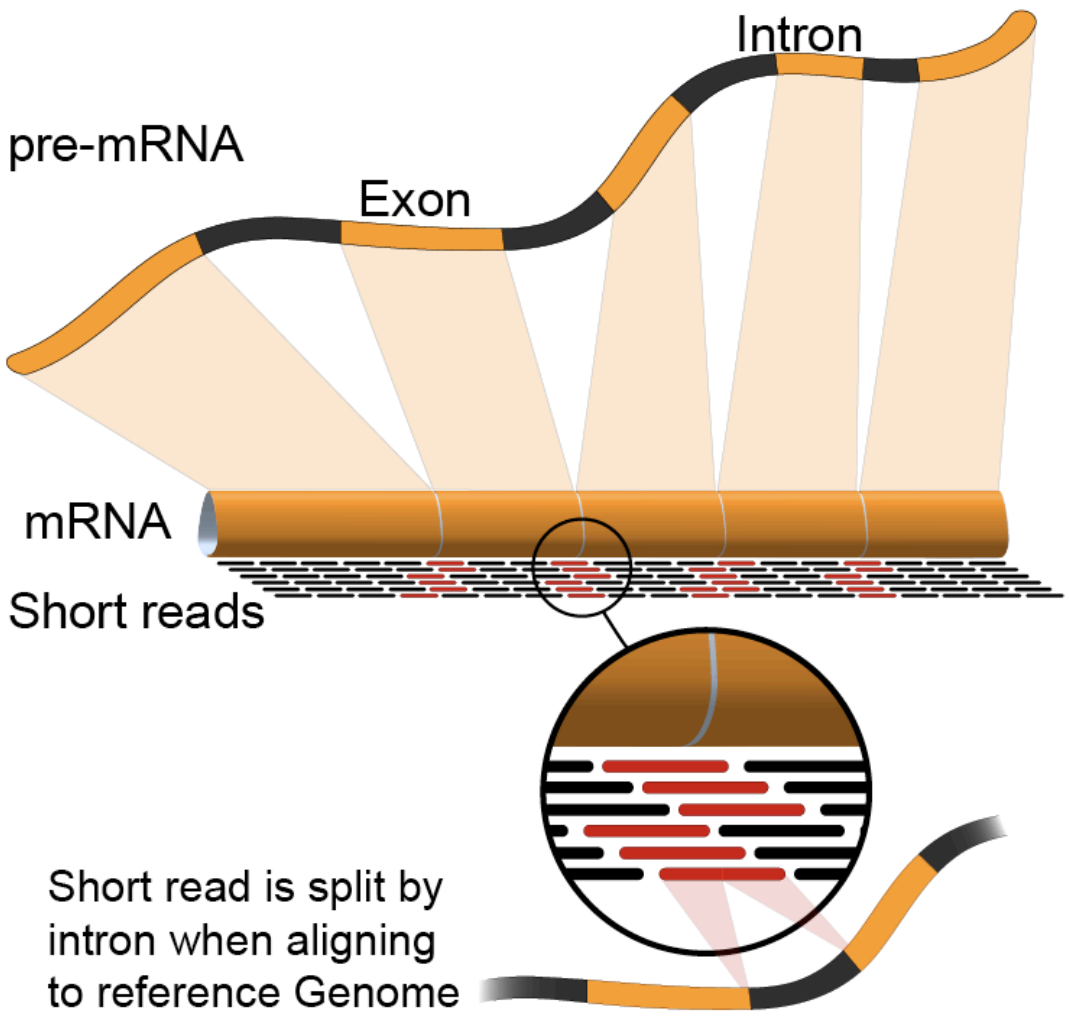
Yuwen Liu[1,2], Jie Zhou[1,3] and Kevin P. White[1,2,3,*]

[1]Institute of Genomics and Systems Biology, [2]Committee on Development, Regeneration, and Stem Cell Biology and [3]Department of Human Genetics, University of Chicago, Chicago, IL 60637, USA

Associate Editor: Janet Kelso

J Club 14.02.2014

RNA-seq



http://upload.wikimedia.org/wikipedia/en/0/01/RNA-Seq-alignment.png

- Whole Transcriptome Shotgun Sequencing
- Used widely for differential expression (DE) studies
- Higher gene coverage than microarrays
- High technical reproducibility

Motivation

- Most large-scale RNA-seq studies favor low level biological replication with deep sequencing
- Inefficient designs of RNA-seq studies can lead suboptimal power and waste of resources
- High-sequencing depth generates more informational reads, but there is a saturation level
- To achieve maximum power to detect DE genes within a budget, a compromise must be made

Data

- MCF7 breast cancer cells
- Seven samples were treated with 10nM 17β-estradiol (E2) for 24h, other seven samples were as controls
- RNA-seq libraries (14 in total) were constructed using the Illumina TruSeq RNA sample preparation protocol
- A total of 50 bp single end reads were generated
- More than 30 million reads per library

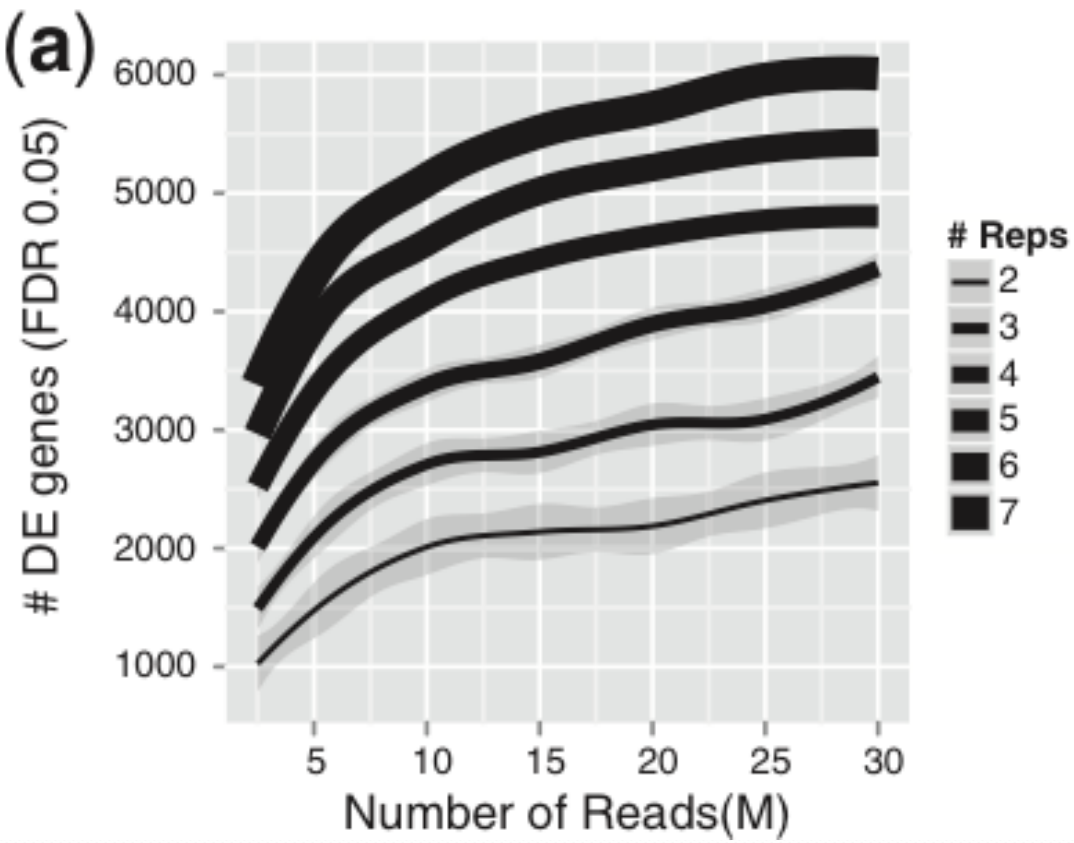
Analysis (1)

- Sequences were aligned with TopHat to the hg18
- Reads were randomly downsampled to generate seven datasets: 2.5, 5, 10, 15, 20, 25, and 30 M; only aligned reads
- Raw counts of number of tags on each gene were created
- edgeR and DESeq packages were used to detect significantly DE genes between control and E2-treated samples (FDR cutoff 0.05)
- Genes with < 5 reads were removed from calculation (standard cutoff for most analyzing programs)

Analysis (2)

- Each sequencing depth and treatment sample were simulated 100 times (data randomly picked)
- For power calculation (ROC curves) a list of 3292 "true positive" genes was used <– *DE genes detected by edgeR from 7 replicas, sequencing depth 30M and FDR 0.001*
- LogFC CV fold change coefficient of variation was computed for top 100 DE genes (lowest FDR in 7 replicas) of 100 simulations
- LogCPM CV (logarithm of counts per million reads) for estimating expression level
- Cost per DE gene:
  - Illumina sequencing per lane $1,200 (reagents, personnel, equipment, contracts); 150M read per lane; multiplexing for lane is 24x
  - Library preparation for each sample $250 (reagents and personnel)

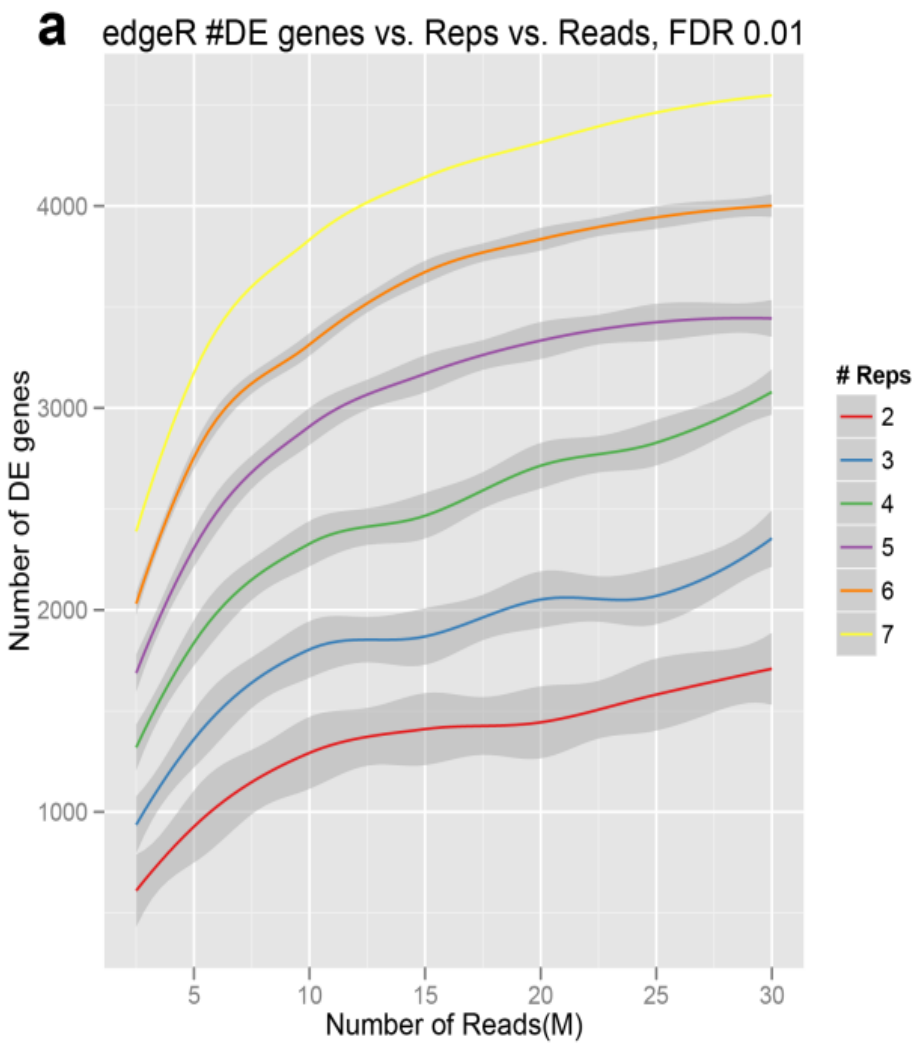
Number of DE genes detected (1)

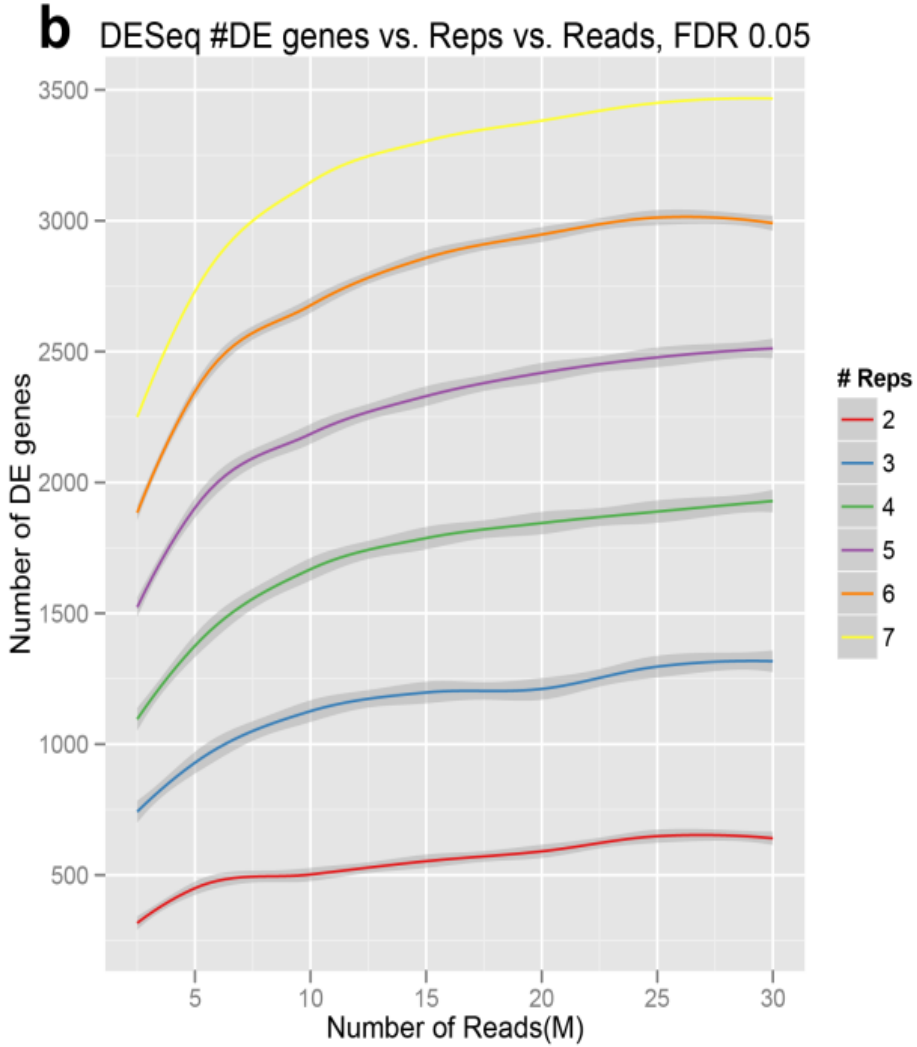


Line thickness indicates depth of replication, with 2 replicates the darkest and 7 replicates the lightest. The lines are smoothed averages for each replication level, with the shaded regions corresponding to the 95% confidence intervals.

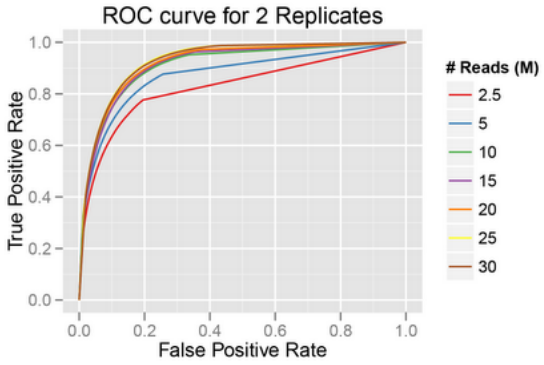
- Increase in biological replication significantly increases the number of DE genes identified
- Number of sequencing reads have a diminishing returns after 10 M reads
- Similar results were observed using different significance cutoffs or software (DESeq)

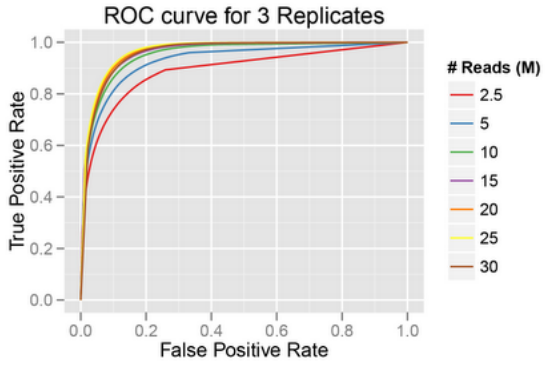
Number of DE genes detected (2)

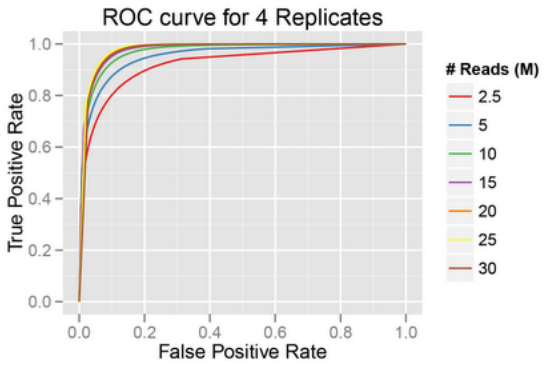


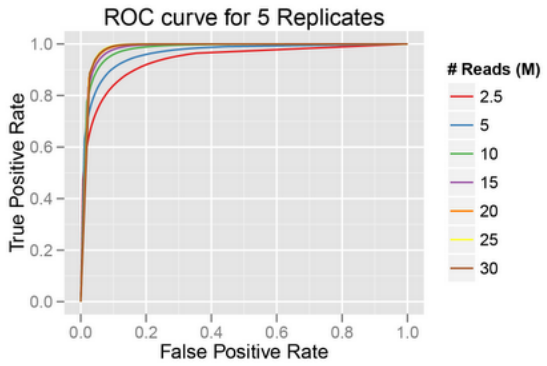


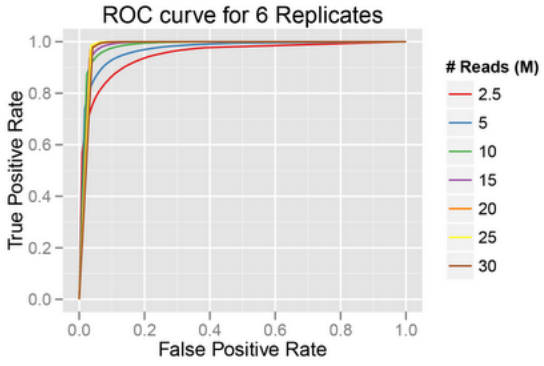
Power of detecting DE genes



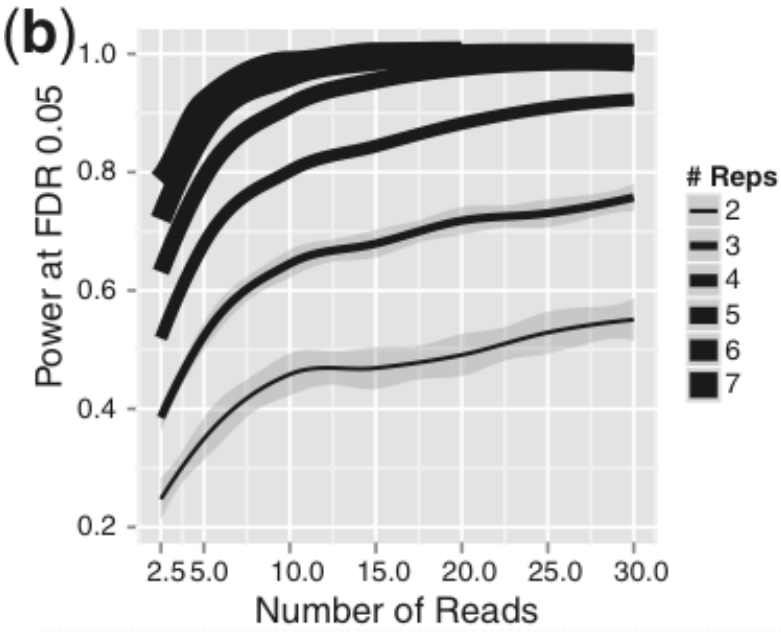




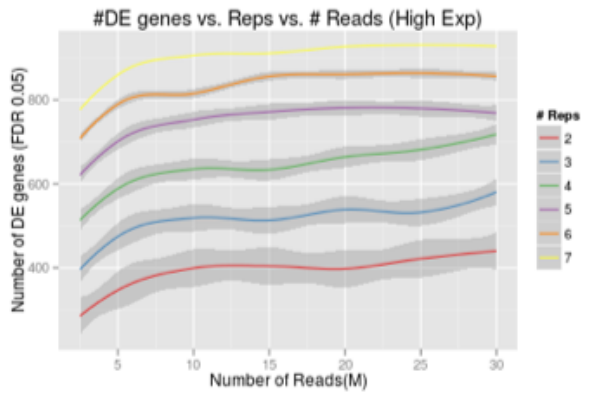


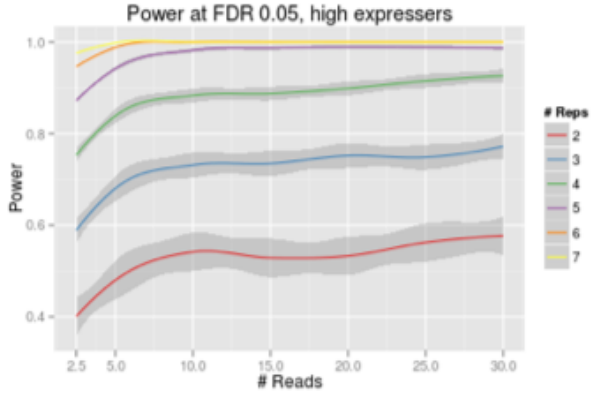


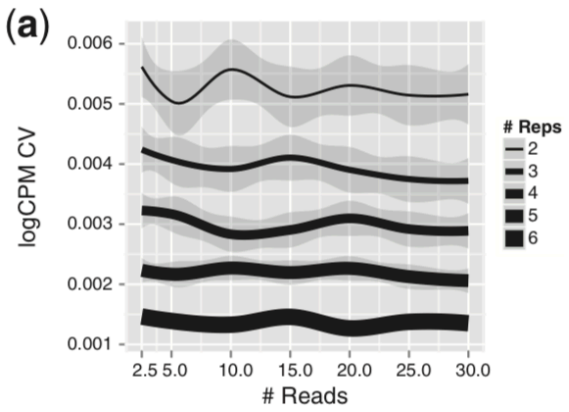
- Sequencing depth over 10 M reads barely leads to any significant improvement in either precision or power

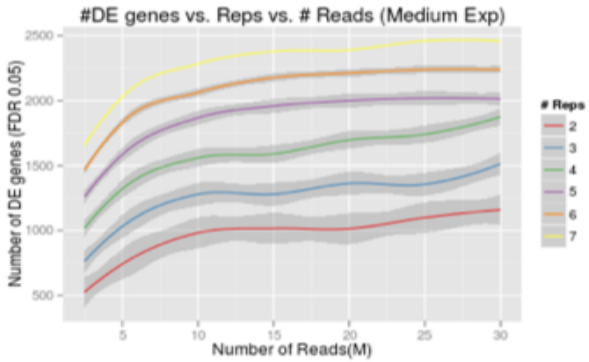


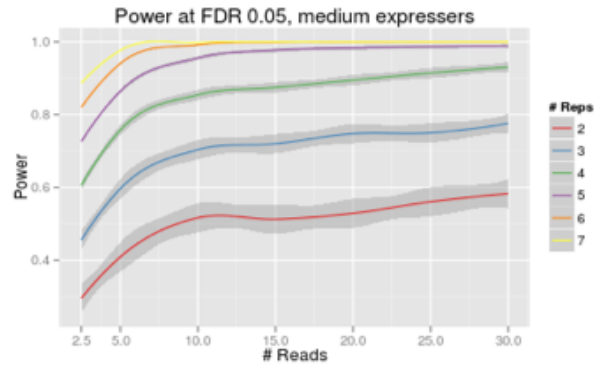
DE genes by different expression levels

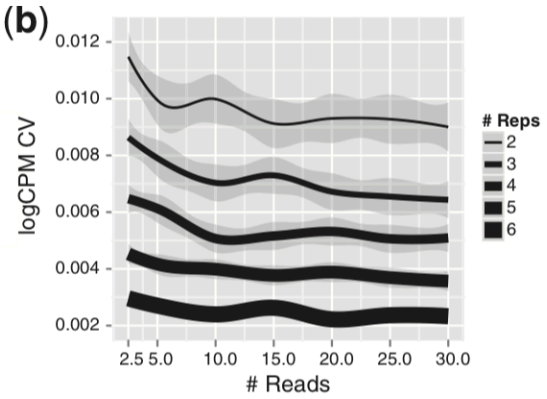


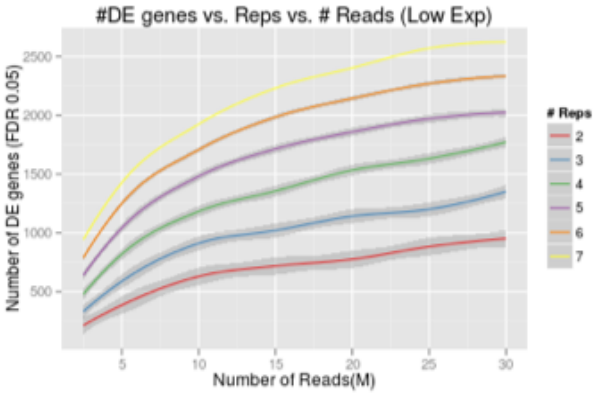


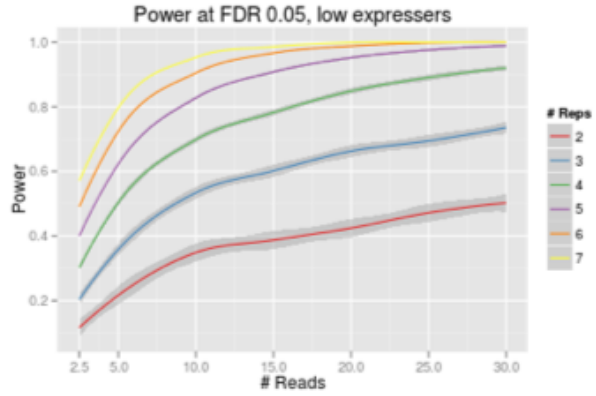


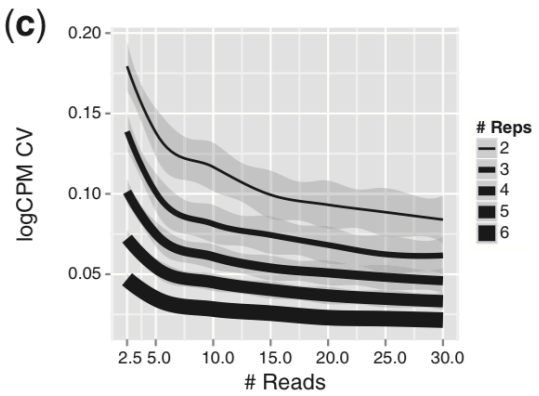




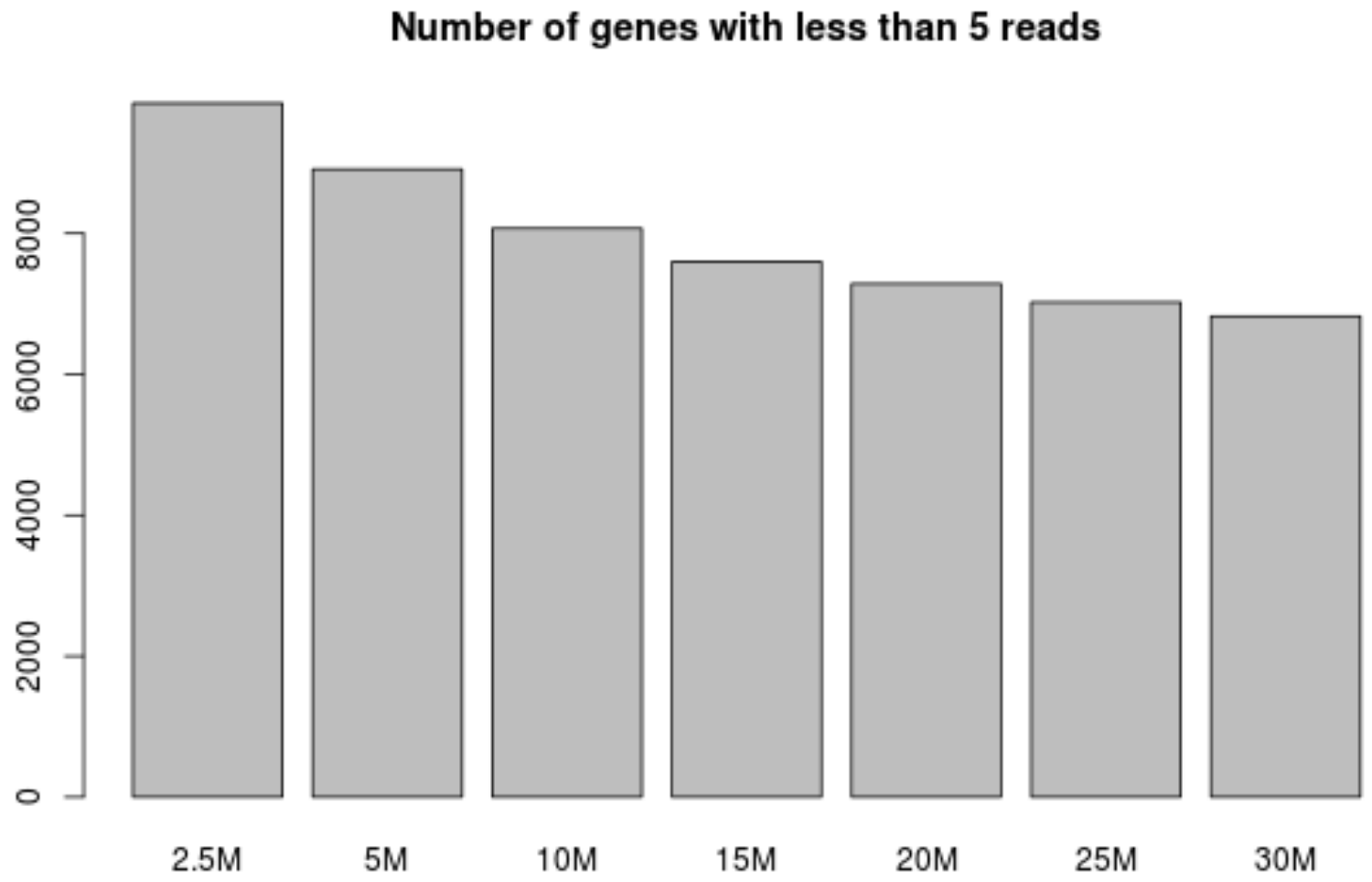








Number of of genes with <5 reads



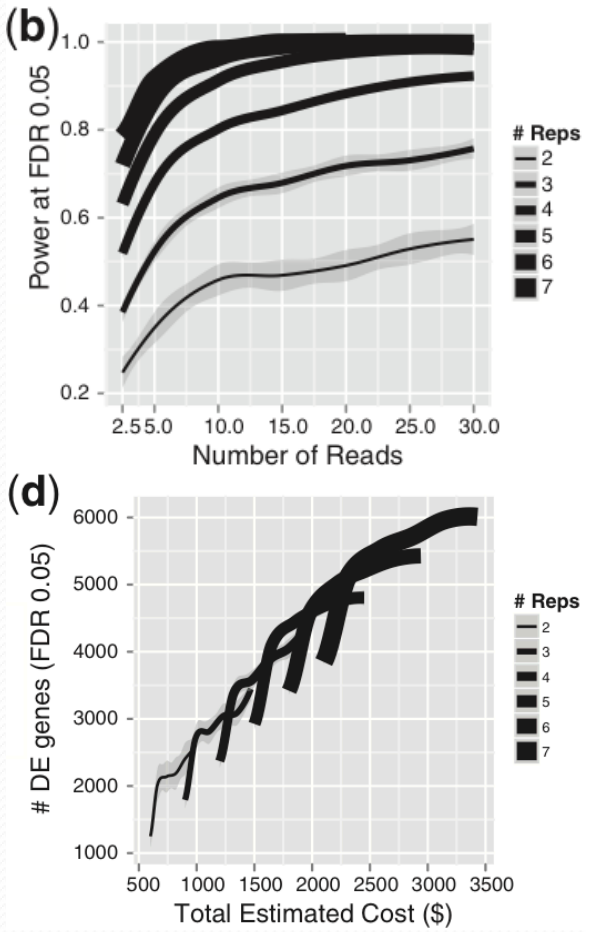
Cost-effectiveness

$$Cost = \frac{(\text{fixed costs per sample} * \text{number of samples} + \text{sequencing costs})}{\text{power}}$$

Table 1. Cost efficiency for power to detect DE genes (cost per 1% power given each experimental design where the variables are)

Relative cost	2.5 M	5 M	10 M	15 M	20 M	25 M	30 M
2 replicates	24.2	17.2	14.4[a]	15.8	16.7	17.0	17.8
3 replicates	23.4	17.2	15.3[a]	16.3	17.1	18.5	19.4
4 replicates	23.1	17.7	16.5[a]	17.5	18.6	19.8	21.2
5 replicates	23.8	19.0	18.1[a]	19.4	21.0	22.8	24.9
6 replicates	25.0	20.7	20.6[a]	22.4	24.6	27.0	29.4
7 replicates	26.8	23.0[a]	23.5	26.0	28.7	31.5	34.3

Note: Assumptions made during calculations are described in Section 2.
[a]Lowest cost per 1% power in each replication level. Units are in dollars.



Conclusions

- In typical DE study using RNA-Seq, increasing sequencing depth beyond a certain level does not result detecting more genes
- Power, estimation accuracy for fold change and absolute expression levels greatly improve with more replicas, not much with increased sequencing
- For MCF7 cell samples their cost metric suggests that sequencing depth 10M reads is enough
- For other species and samples, the exact depth will be different, but overall guidelines should still remain
- However, DE of exons for example a deep sequencing is advantageous, because informative region are much shorter