UCDAVIS Bioinformatics Core

RNAseq: Differential Gene Expression Experimental Design

Dr. Matthew L. Settles

Genome Center
University of California, Davis
settles@ucdavis.edu

What is Differential Expression

Differential expression analysis means taking *normalized* 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 between group A and group B is significant, that is, whether it is greater than what would be expected just due to natural random variation.

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

Designing Experiments

Beginning with the question of interest (and work backwards)

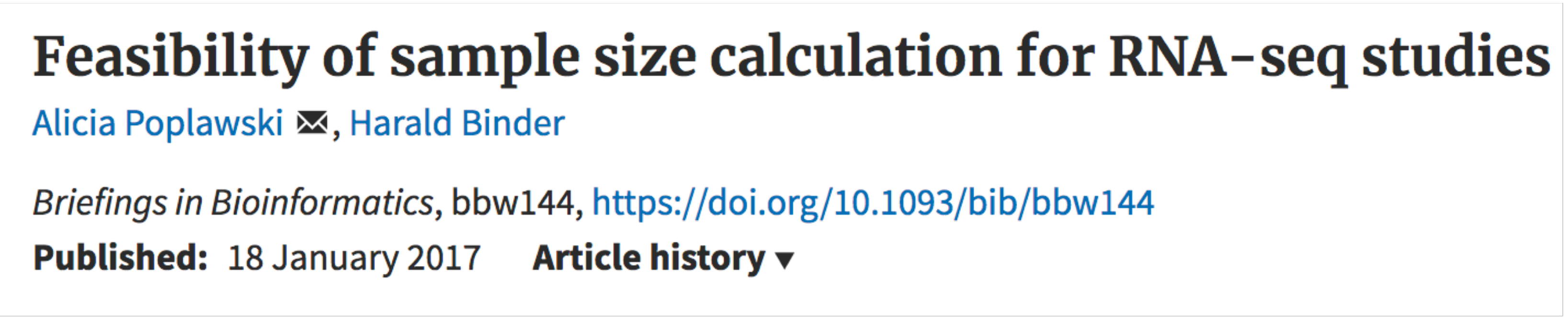
- The final step of a DE analysis is the application of a statistical model to each gene in your dataset.

  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.

Power Analysis



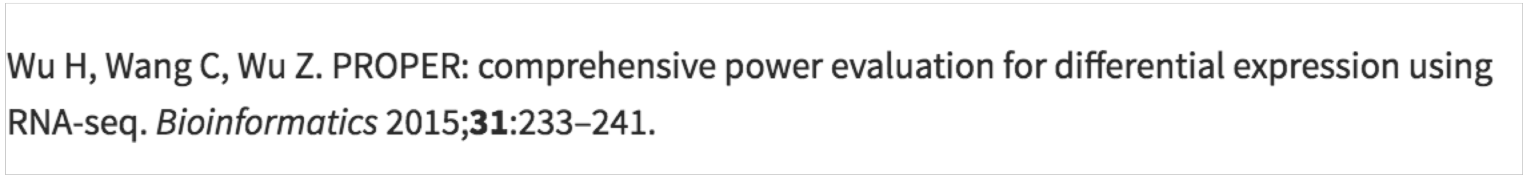
Feasibility of sample size calculation for RNA-seq studies

Alicia Poplawski ✉, Harald Binder

Briefings in Bioinformatics, bbw144, https://doi.org/10.1093/bib/bbw144

Published: 18 January 2017 **Article history ▾**

- A systematic search resulted in six open source tools for sample size calculation for RNA-seq differential expression analysis. As of Jan 2017
- Exemplary sample size estimation performed by the remaining six tools using real mouse and human data as input files led to widely different results.
- Tool evaluation using simulations showed that most tools estimate the sample sizes incorrect in particular for low true effects and large dispersion.
- The use of pilot data is recommended.



Wu H, Wang C, Wu Z. PROPER: comprehensive power evaluation for differential expression using RNA-seq. *Bioinformatics* 2015;**31**:233–241.

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 and experiment/ 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.

Sample preparation

In high throughput biological work (Microarrays, Sequencing, HT Genotyping, etc.), what may seem like small technical details introduced during sample extraction/preparation can lead to large changes, or technical bias, in the data.

Not to say this doesn’t occur with smaller scale analysis such as Sanger sequencing or qRT-PCR, but they do become more apparent (seen on a global scale) and may cause significant issues during analysis.

Be Consistent

BE CONSISTENT ACROSS ALL SAMPLES!!!

Generating RNA-seq libraries

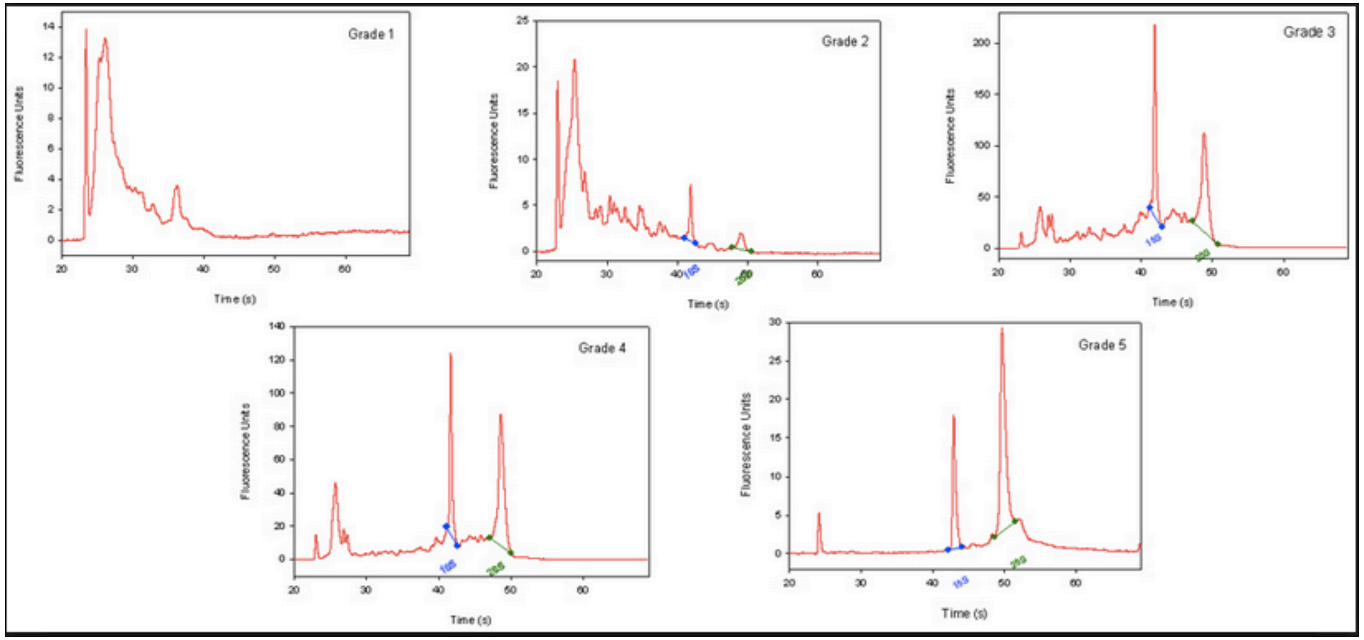
Considerations

- QA/QC of RNA samples
- What is the RNA of interest
- Library Preparation
  - Stranded Vs. Unstranded
  - Whole transcript Vs. 3-prime biased
- 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

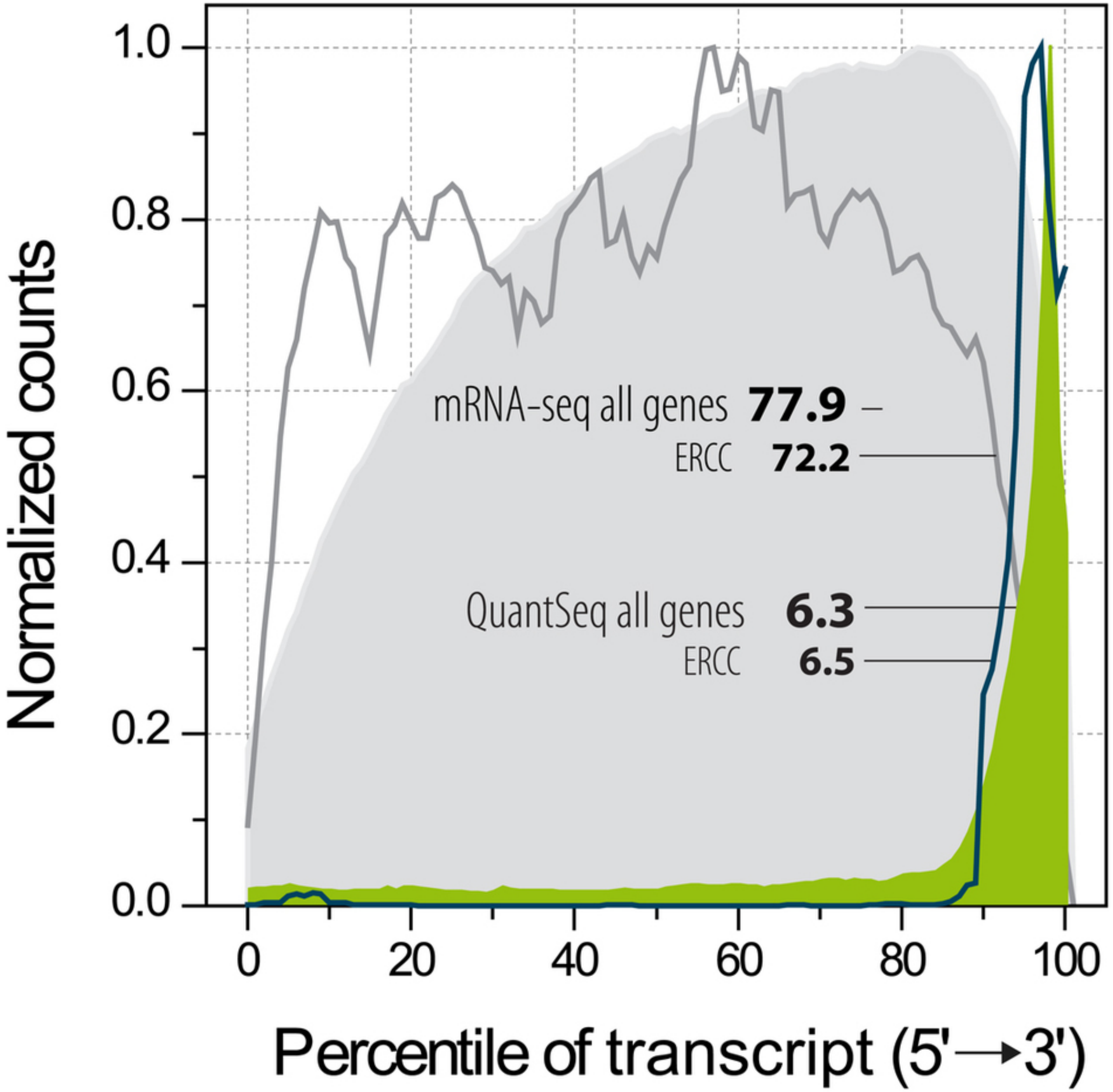
- From “total RNA” we extract “RNA of interest”. Primary goal is to NOT sequence 90% (or more) ribosomal RNAs, which are the most abundant RNAs in the typical sample. there are two main strategies for enriching your sample for “RNA of interest”.
  - polyA selection. Enrich mRNA (those with polyA tails) 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 which align to intergenic and intronic regions of the genome.

Library Preparation

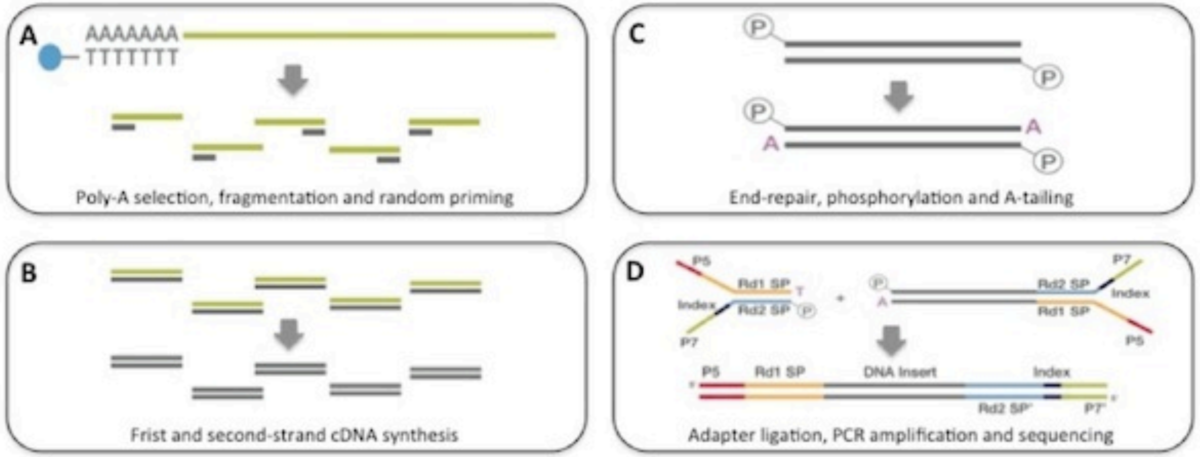
- Some library prep methods first require you to generate cDNA, in order to ligate on the Illumina barcodes and adapters.
  - cDNA generation using oligo dT (3’ biased transcripts)
  - cDNA generation using random hexomers (less biased)
  - full-length cDNAs using SMART cDNA synthesis method
- Also, 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.

Library Preparation – whole transcript/3’ biased



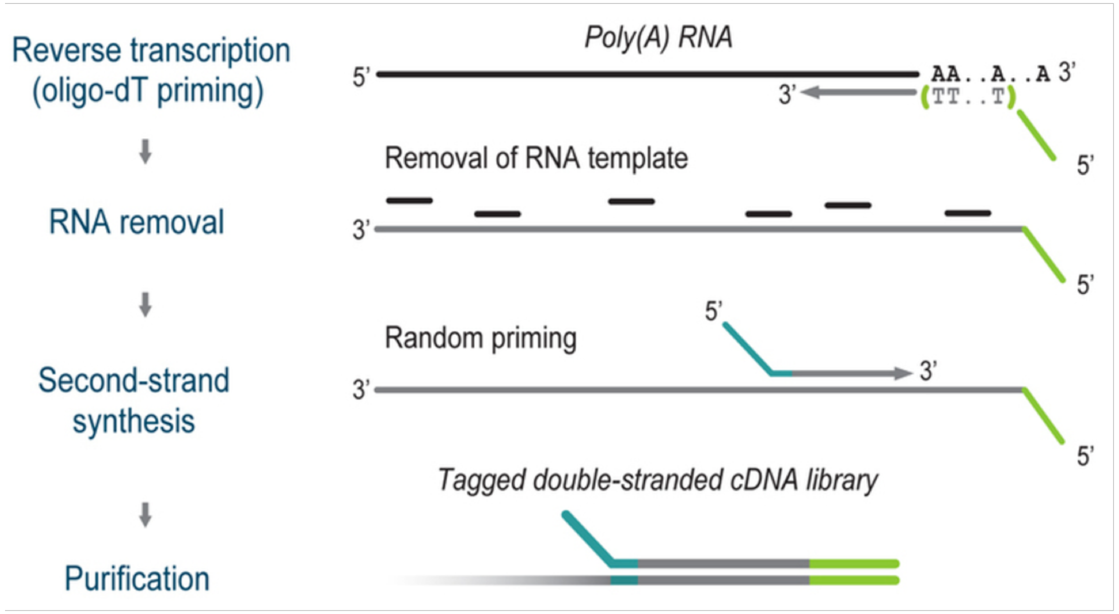
Protocol – whole transcript vs 3-prime biased

Illumina Tru-Seq RNA-seq protocol


A
AAAAAAA
TTTTTTT
Poly-A selection, fragmentation and random priming
C
P
A
End-repair, phosphorylation and A-tailing
B
Frist and second-strand cDNA synthesis
D
P5
Rd1 SP
Index
Rd2 SP
P7
DNA Insert
Adapter ligation, PCR amplification and sequencing

Library prep begins from 100ng-1ug of Total RNA which is poly-A selected (A) with magnetic beads. Double-stranded cDNA (B) is phosphorylated and A-tailed (C) ready for adapter ligation. The library is PCR amplified (D) ready for clustering and sequencing.

Lexogen Quantseq protocol



3-prime differences from whole transcript

- Low input and low quality samples (say as low as 100ng), BUT protocol is a (a bit) more sensitive to chemical contaminants (spin column cleaned RNA samples are recommended)
- Significant cost saving through faster library preparation protocol and multiplexing [need less sequencing per sample] comparable to microarrays.
- Best suited for gene counting as data does not contain transcript splicing information on the whole transcript.
- Requires (sort of) a reference genome with good annotation (plus known UTRs)
- Only applicable to Eukaryotic samples (requires polyA)
- Strand specific
- Single read sequencing is sufficient

Size Selection/Cleanup/QA

Final insert size optimal for DE are ~ 150bp (or kit suggestion)

- 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!!!

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
4. Number of lanes being sequenced

$$\frac{reads}{sample} = \frac{reads.sequenced * 0.8}{samples.pooled} \text{ X num.lanes}$$

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.

Barcodes and Pooling samples for sequencing

- Best to have as many barcodes as there are samples
  - Can purchase barcodes from vendor, generate them yourself and purchase from IDTdna (example), or consult with the sequencing core.
- Best to pool all samples into one large pool, then sequence multiple lanes
- IF you cannot generate enough barcodes, or pool into one large pool, RANDOMIZE samples into pools.
  - Bioinformatics core can produce a randomization scheme for you.
  - This must be considered/determined PRIOR to library preparation

[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]
  - Whole transcript vs 3-prime biased [IF DE only, 3-prime biased]
- Size Selection/Cleanup [Target mean 150bp or kit recommendation]
  - Final QA [Consistency across samples is most important.]

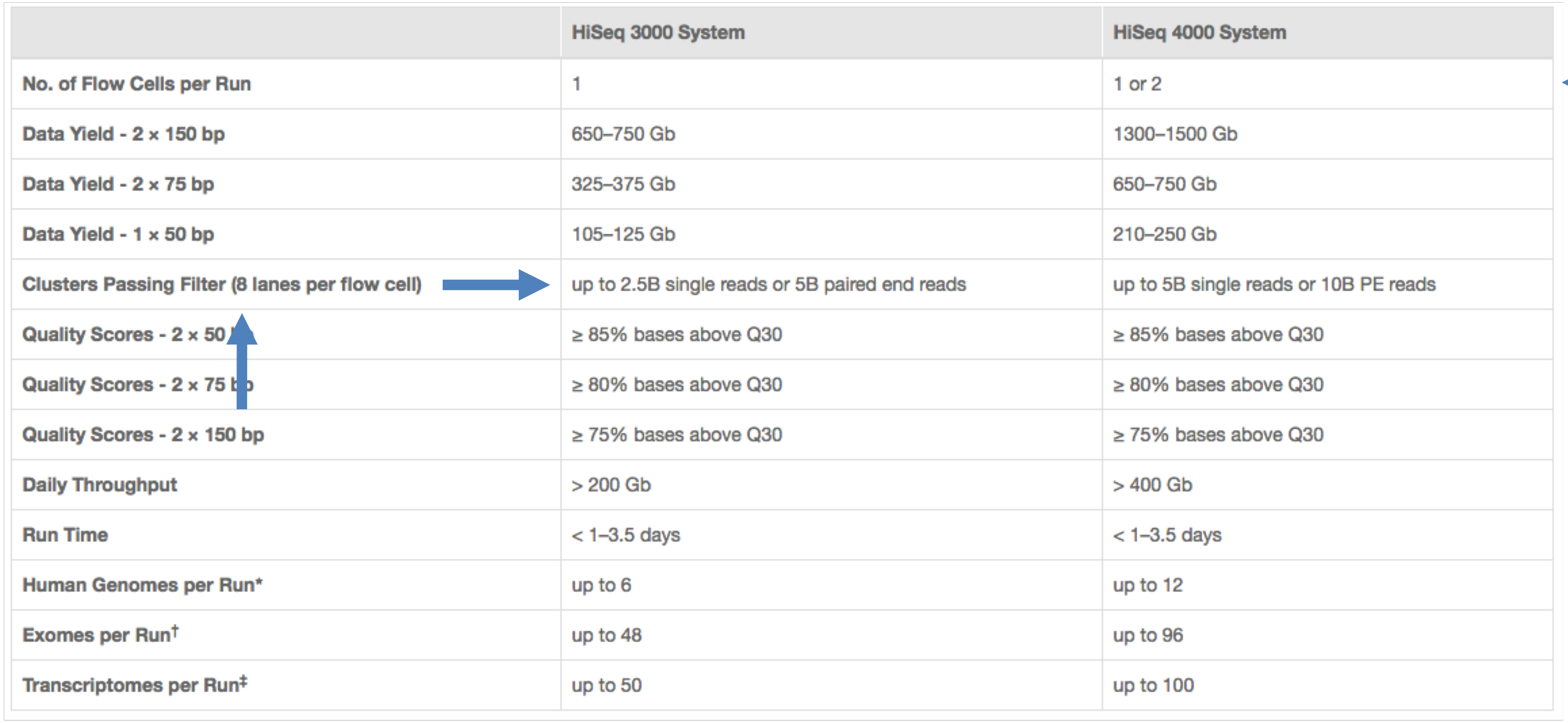
Cost Estimation

- RNA extraction and QA/QC (Per sample)
- Enrichment of RNA of interest + library preparation (Per sample)
  - Library QA/QC (Bioanalyzer and Qubit)
  - Pooling [If you generate your own libraries]
- Sequencing (Number of lanes)
- Bioinformatics (General rule is to estimate the same amount as data generation, i.e. double your budget)

http://dnatech.genomecenter.ucdavis.edu/prices/

Illumina sequencing costs

I use 350M fragments per lane



	HiSeq 3000 System	HiSeq 4000 System
No. of Flow Cells per Run	1	1 or 2
Data Yield - 2 × 150 bp	650–750 Gb	1300–1500 Gb
Data Yield - 2 × 75 bp	325–375 Gb	650–750 Gb
Data Yield - 1 × 50 bp	105–125 Gb	210–250 Gb
Clusters Passing Filter (8 lanes per flow cell)	up to 2.5B single reads or 5B paired end reads	up to 5B single reads or 10B PE reads
Quality Scores - 2 × 50 bp	≥ 85% bases above Q30	≥ 85% bases above Q30
Quality Scores - 2 × 75 bp	≥ 80% bases above Q30	≥ 80% bases above Q30
Quality Scores - 2 × 150 bp	≥ 75% bases above Q30	≥ 75% bases above Q30
Daily Throughput	> 200 Gb	> 400 Gb
Run Time	< 1–3.5 days	< 1–3.5 days
Human Genomes per Run*	up to 6	up to 12
Exomes per Run†	up to 48	up to 96
Transcriptomes per Run‡	up to 50	up to 100

http://www.illumina.com/systems/hiseq-3000-4000/specifications.html

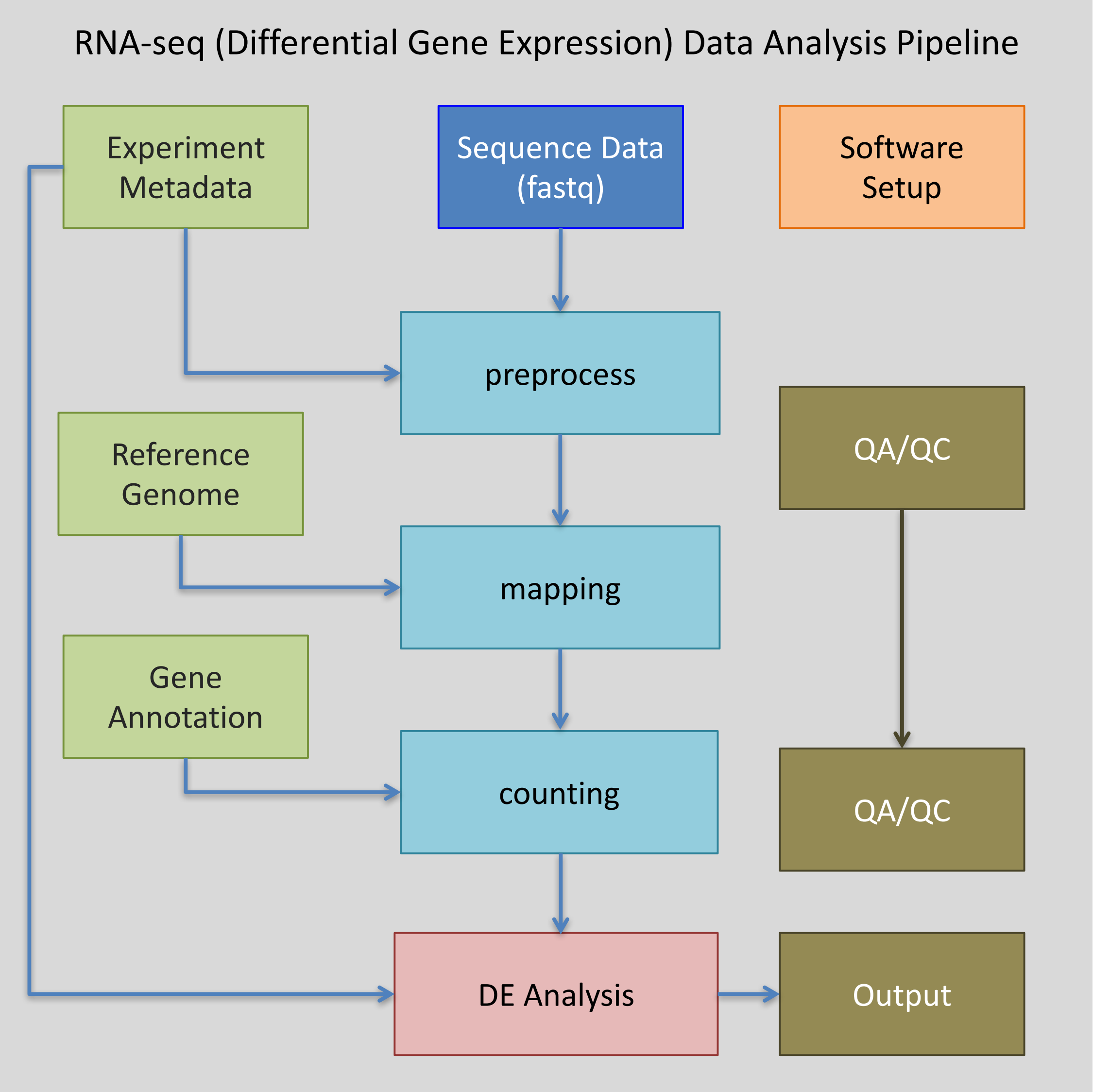
Cost Estimation

- 48 Samples, QA/QC and Library Prep (Poly-A)
  - QA Bioanalyzer = 4*$98 for all 48 samples, 4 chips
  - Library Preparation (Poly-A Enrichment) = $100/sample = 48*$100 = $4,800
- Sequencing, targeting 10M reads per sample (Illumina HiSeq 4000)
  - 2.1 - 2.5 Billion reads per run / 8 lanes = Approximately 350M reads per lane
  - Multiplied by a 0.8 buffer equals 280M expected good reads
  - Divided by 48 samples in the lane = 5.8M reads per sample per lane.
  - Target 10M reads means 2 lanes of sequencing 2*$2,346 = $4,692
- Bioinformatics
  - Double your budget

Total = $392 + $4,800 + $4,692 = $9,884, w/Bioinformatics $19,768

Approximately $412 per sample @ 10M reads per sample w/bioinformatics

RNA-seq pipeline overview



RNA-seq (Differential Gene Expression) Data Analysis Pipeline