UCDAVIS Bioinformatics Core

So you want to do a: Peak based study

Matthew L. Settles

Genome Center Bioinformatics Core

University of California, Davis

settles@ucdavis.edu; bioinformatics.core@ucdavis.edu

Disclaimer

- This talk/workshop is full of opinion, there are as many different way to perform analysis as there are Bioinformaticians.
- My opinion is based on over a decade of experience and spending a considerable amount of time to understand data, how its created and how it relates to the biological questions of interest.
- Each experiment is unique, this lecture is a starting place and should be adapted to the specific characteristics of your experiment.

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.

General rules for preparing and experiment/ samples

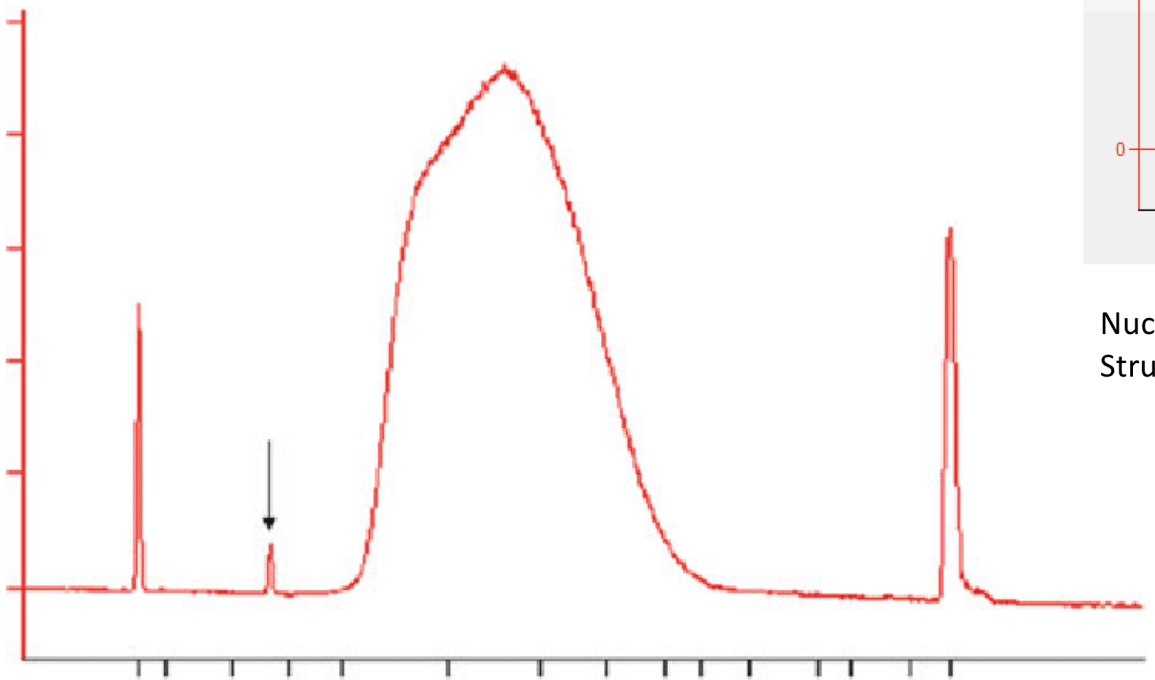
- Prepare more samples then you are going to need, i.e. expect some will be of poor quality, or fail
- Preparation stages should occur across all samples at the same time (or as close as possible) and by the same person
- Spend time practicing a new technique to produce the highest quality product you 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 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.

Be Consistent

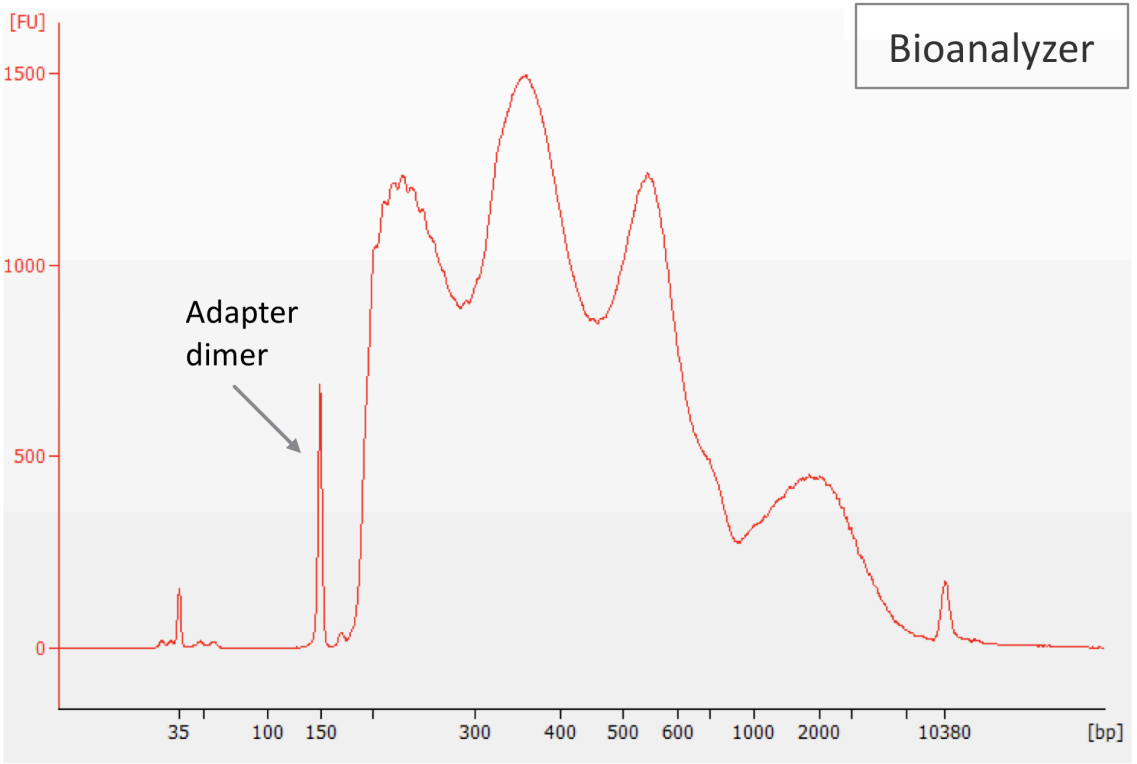
BE CONSISTENT ACROSS ALL SAMPLES!!!

UCDAVIS Bioinformatics Core

QA/QC traces



Bioanalyzer



Nucleosome Structure

BE CONSISTANT!!!

Sequencing Depth

Coverage is determined differently for ”Counting” based experiments (RNAseq, ChIPseq, 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 peaks

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 peaks expressed at low levels (<< level, the >> the depth and necessary complexity of library)
- The change you want to be able to detect (< fold change more replicates, more depth)
- Paired end allows you to better define the start and end region of the peak.

The amount of sequencing needed for a given sample/experiment is determined by the goals of the experiment and the nature of the sample.

UCDAVIS Bioinformatics Core

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

Cost Estimation

- QA/QC (Per sample)
- Enrichment of DNA 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/

https://bioinformatics.sf.ucdavis.edu/rates

UCDAVIS Bioinformatics Core

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