Experimental Design

Dr. Matthew L. Settles

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

Designing Experiments

Beginning with the question of interest (and work backwards)

- In many cases the final step is the application of a model 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 statistical 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 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.

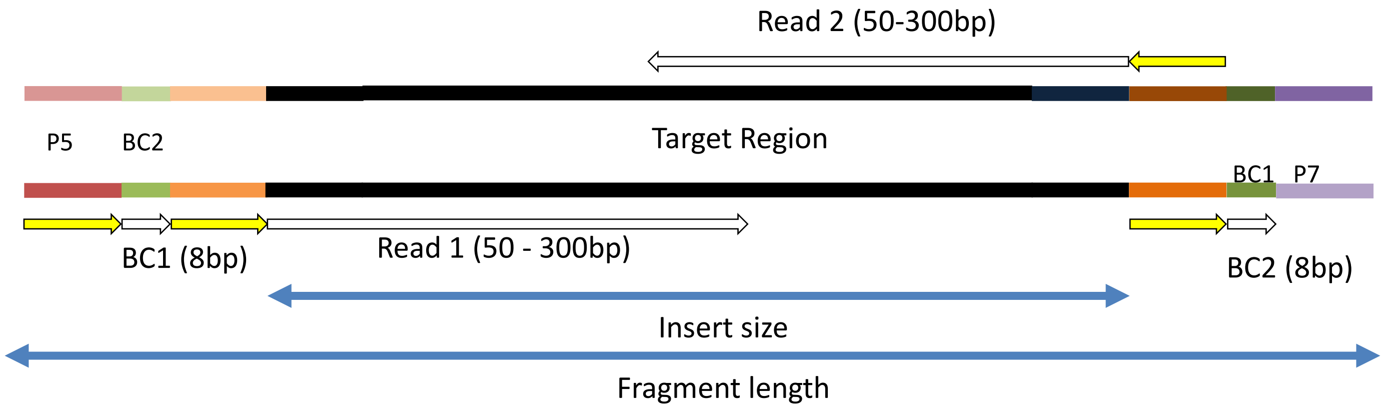
Not to say this doesn’t occur with smaller scale analysis such as Sanger sequencing or qRT-PCR, but they do become more apparent (seen on a global scale) and may cause significant issues during analysis.

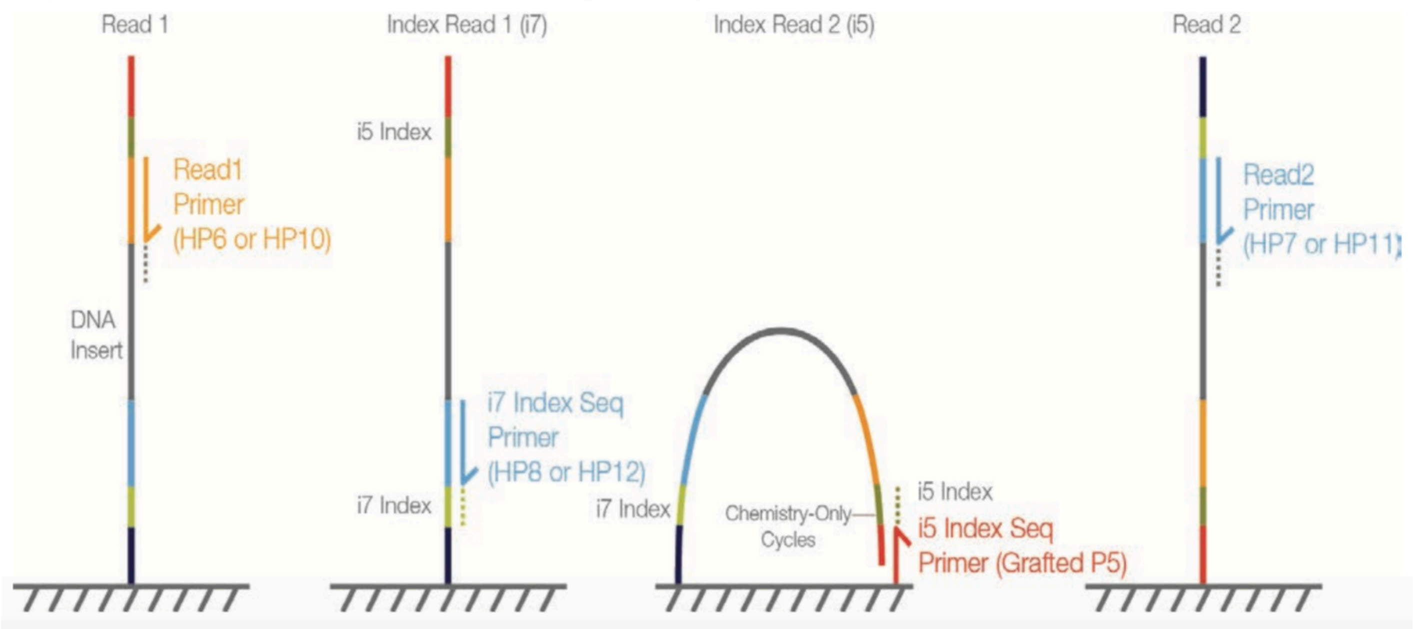
Be Consistent

BE CONSISTENT ACROSS ALL SAMPLES!!!

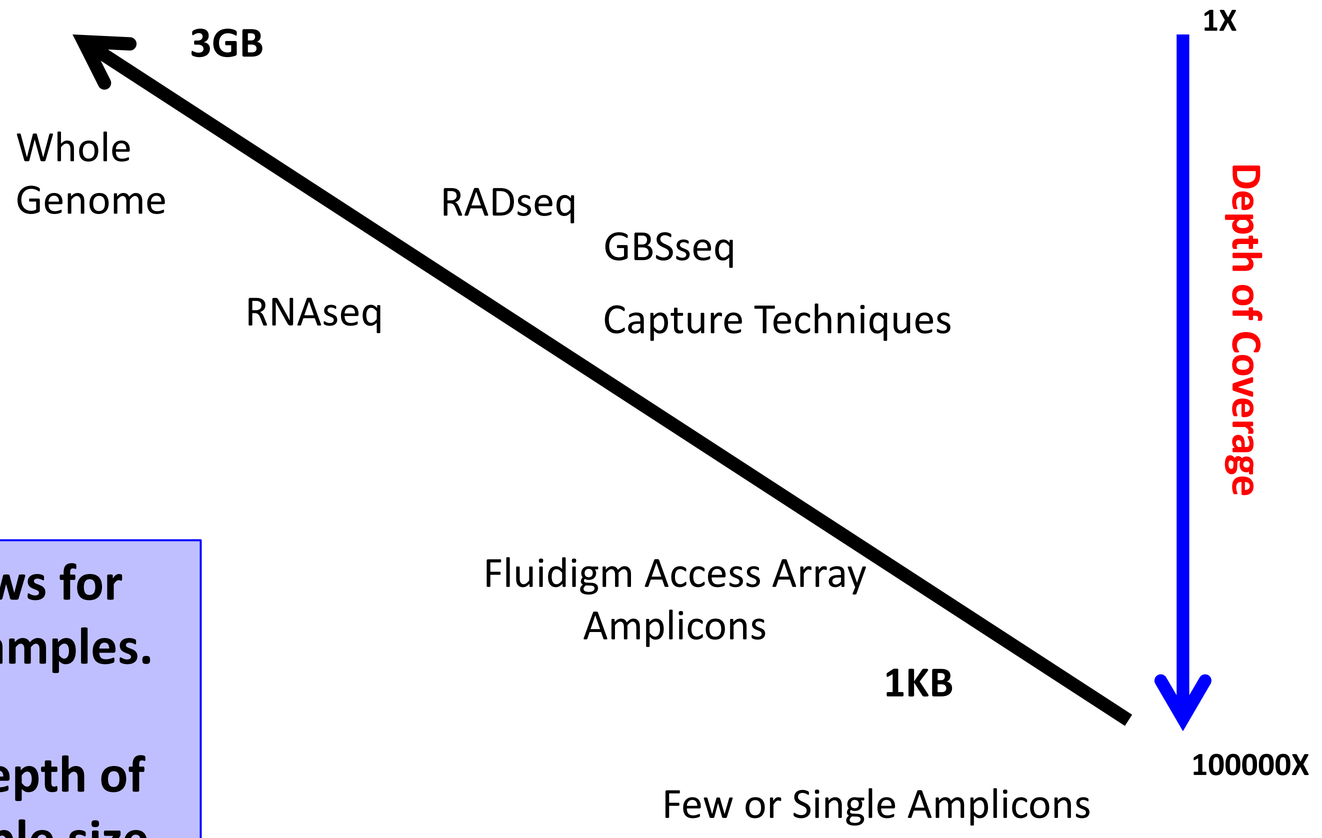
Illumina sequencing

- Illumina SBS




Read 1
Index Read 1 (i7)
Index Read 2 (i5)
Read 2
i5 Index
Read1 Primer (HP6 or HP10)
DNA Insert
i7 Index Seq Primer (HP8 or HP12)
i7 Index
i7 Index
Chemistry-Only Cycles
i5 Index
i5 Index Seq Primer (Grafted P5)
Read2 Primer (HP7 or HP11)

Genomic Reduction



Genomic reduction allows for greater multiplexing of samples.

You can fine tune your depth of coverage needs and sample size with the reduction technique

Sequenced Basepairs per samples per lane

The first and most basic question is how many base pairs of sequence will I get

Factors to consider then are:

1. Number of reads being sequenced
2. Read length (if reads are paired, consider them as individuals for this calculations)
3. Number of samples being sequenced
4. Expected percentage of good bases/reads

$$\frac{bp}{sample} = \frac{readLength * (\# reads)}{\# samples} * 0.8$$

The number of reads and read length data are best obtained from the manufacturer’s website (search for specifications) and always use the lower end of the estimate.

Genomic Coverage

Once you have the number of base pairs per sample you can then determine expected coverage

Factors to consider then are:

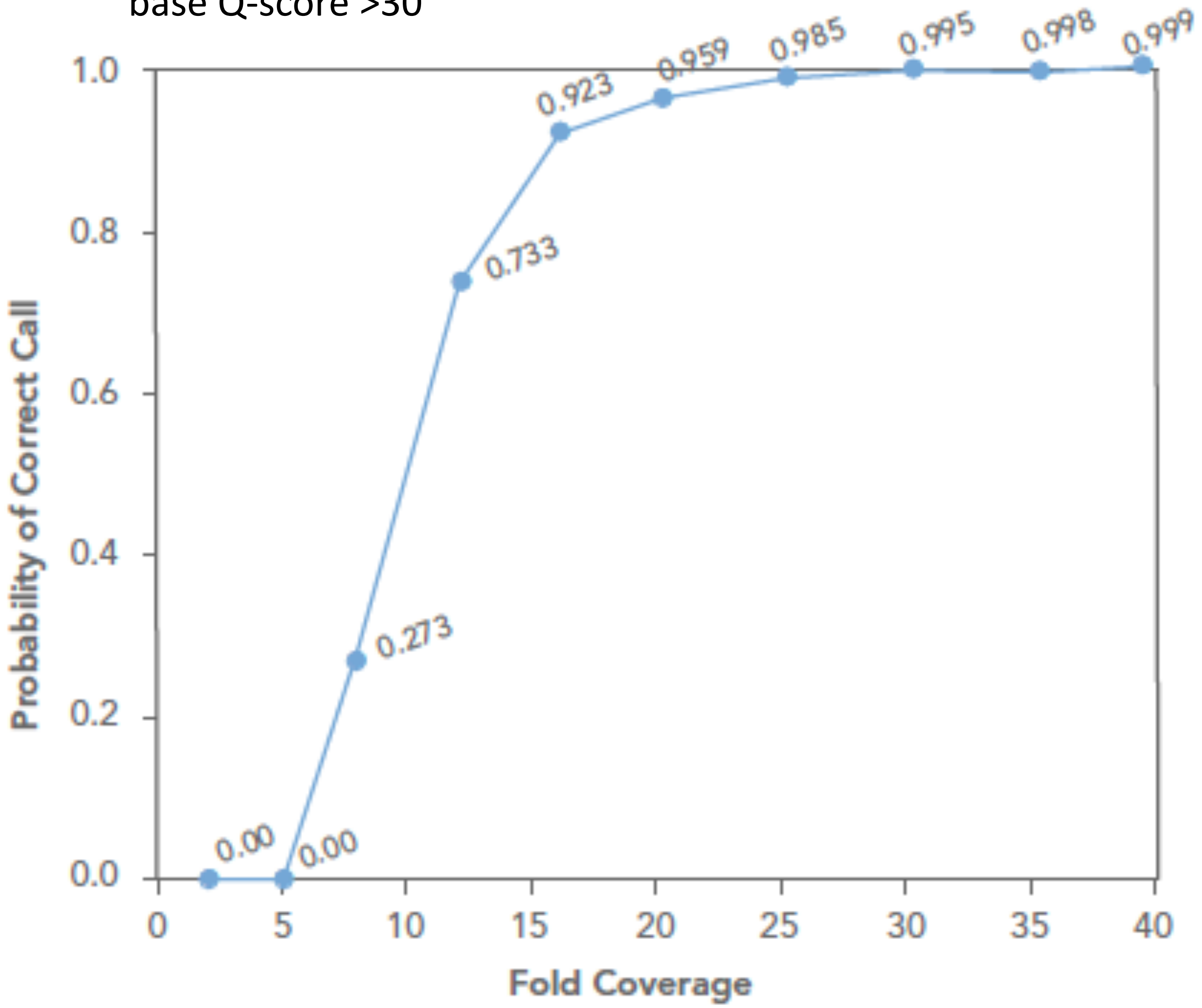
1. Length of the genome
2. Any extra-genomic sequence (ie mitochondria, virus, plasmids, etc.). For bacteria in particular, these can become a significant percentage

$$\frac{ExpectedCoverage}{sample} = \frac{\frac{(readLength\ * numReads) * 0.8}{numSamples} * \text{num.lanes}}{TotalGenomicContent}$$

Variant Analysis

- Read length contributes to uniqueness of mapping
- Paired reads are required to identify structure changes
- For a single individual we target > 30x coverage.
- In population studies, the greater the number of samples less coverage per samples that is required. (ex. with 1000 samples 2x coverage per sample may be sufficient)

Probability of a correct SNP call, assuming base Q-score >30



Sequencing Depth – Counting based experiments

- 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

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

- Read length, or SE vs PE, does not factor into sequencing depth.

Amplicon Sequencing (Communities, genotyping)

Considerations

- Number of reads being sequenced
- Proportion that is diversity sample (e.g. PhiX)
- Number of samples being pooled in the run

The back of the envelope calculation

$$\frac{reads}{sample} = \frac{reads\_sequenced * (1 \ - diversity\_sample)}{num\_samples}$$

example

$$\frac{102{,}000}{sample} = \frac{18e6 \ * (1 - 0.15)}{150}$$

Recommendations

- Illumina ‘recommends’ 100K per sample
- I’ve used 30K per sample historically, others are fine with 3K per sample
- Really should have as many reads as your experiment needs

Metagenomics Sequencing

Considerations (when a literature search turns up nothing)

- Proportion that is host (non-microbial genomic content)
- Proportion that is microbial (genomic content of interest)
- Number of species
- Genome size of each species
- Relative abundance of each species

The back of the envelope calculation

$$\frac{numReads}{sample} = \frac{Coverage * (AverageGenomeSize)}{ReadLen\ * DilutionFactor\ * (1 - hostProportion)} * \frac{1}{0.8}$$

ReadLen = 200

Coverage = 30

hostProportion = 0.5

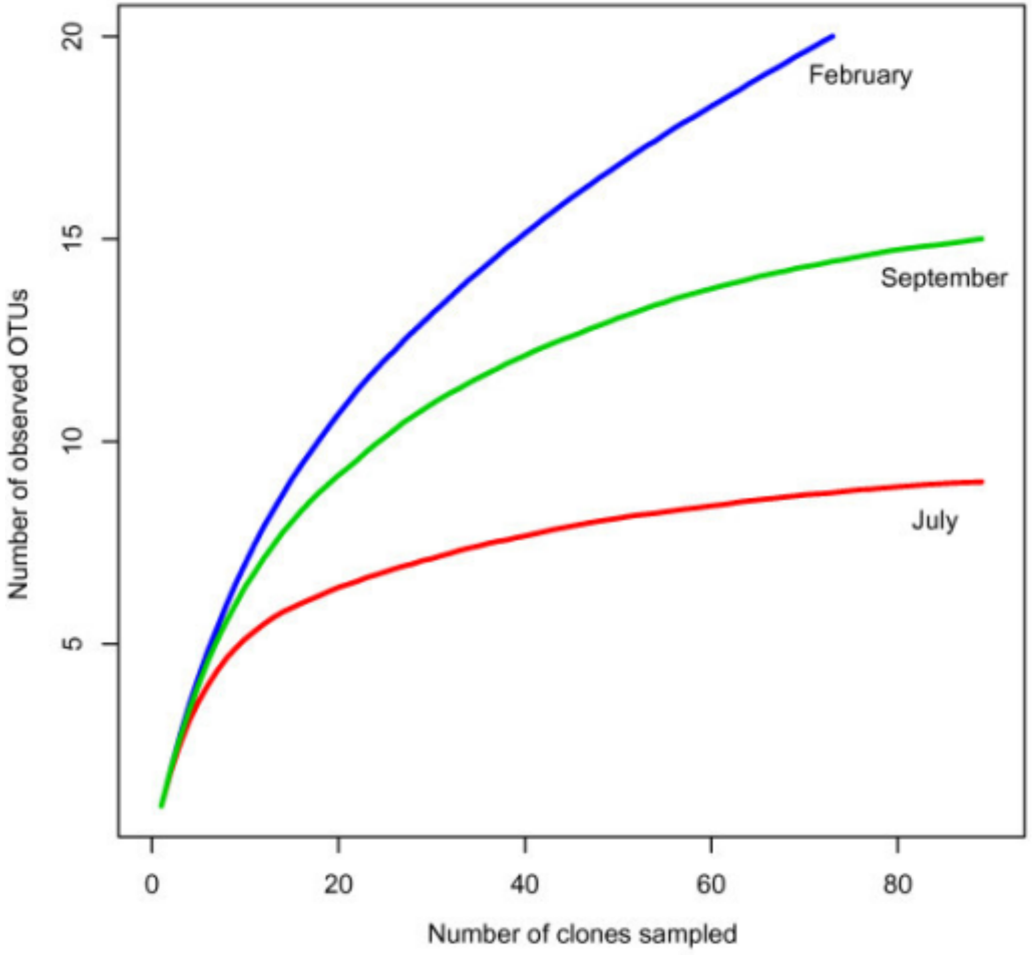
DillutionFactor = 0.01

AverageGenomeSize = 5Mb

Community Rarefaction curves

- ’Deep’ sequence a number of test samples
  amplicons: ~ 1M+ reads.
  metagenomics: 1 full HiSeq lane

- Plot rarefactions curves of organism identification, to determine if saturation is achieved



Take Homes

- Experience and/or literature searches (other peoples experiences) will provide the best justification for estimates on needed depth.
- ‘Longer’ reads are better than short reads.
- Paired-end reads are more useful than single-end reads
- Libraries can be sequenced again, so do a pilot, perform a preliminary analysis, then sequence more accordingly.

Cost Estimation

- Extractions from tissue (DNA/RNA): cost per sample
- Sample quality assurance. Including quantification and sample degradation: cost per sample
- PCR reactions: Cost per sample
- Library generation and quantification: cost per sample
- Pooling and quantification of libraries: cost per group
- Sequencing (type if sequencing PE/SE, length of reads, number of lanes / runs): cost per lane/run
- Bioinformatics, general rule is to estimate double your budget)

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

Bioinformatics Costs

Bioinformatics includes:

1. Storage of data
2. Access and use of computational resources and software
3. System Administration time
4. Bioinformatics Data Analysis time
5. Back and forth consultation/analysis to extract biological meaning

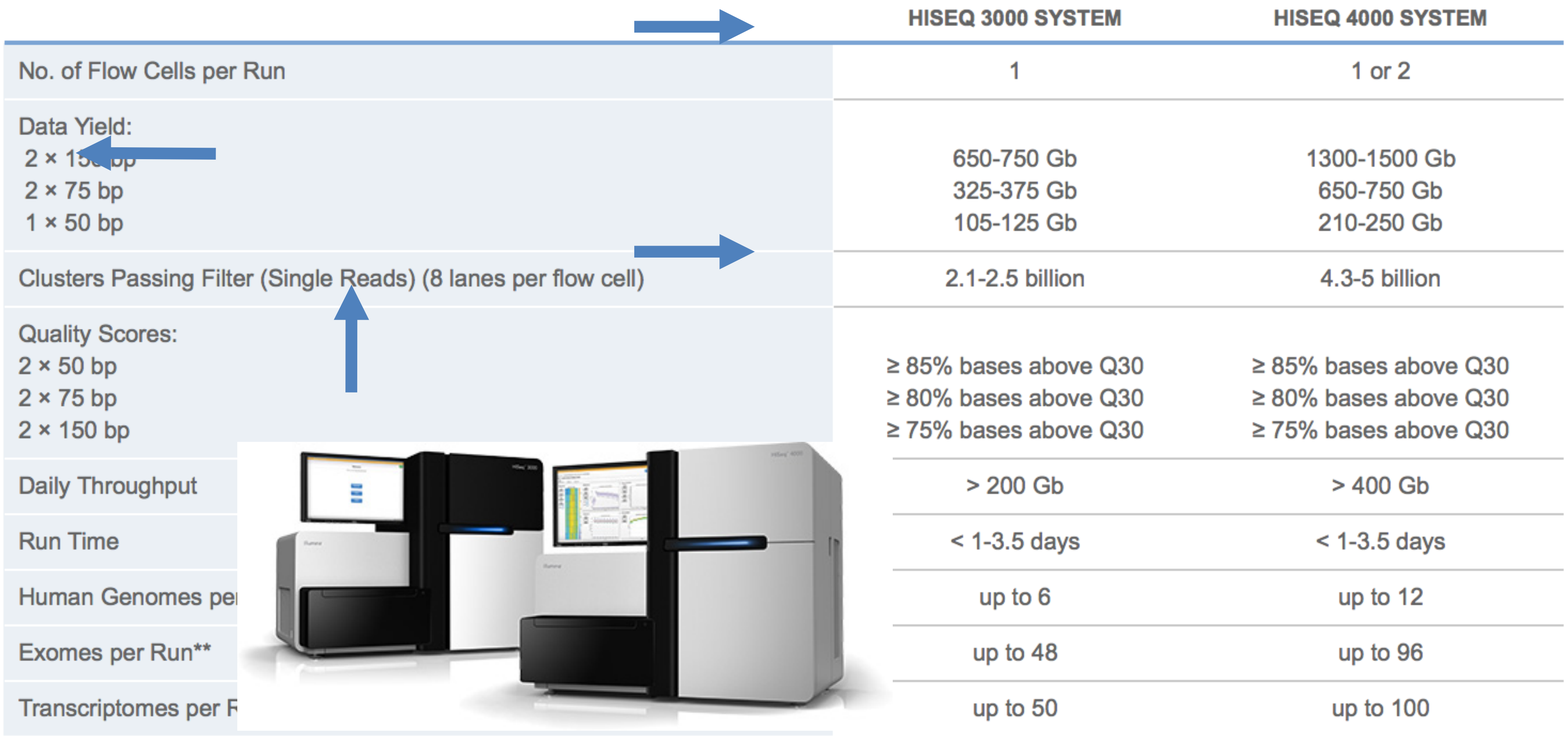
Rule of thumb:
Bioinformatics can and should cost as much (sometimes more) as the cost of data generation.

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 DNA technologies 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

Illumina HISEQ sequencing

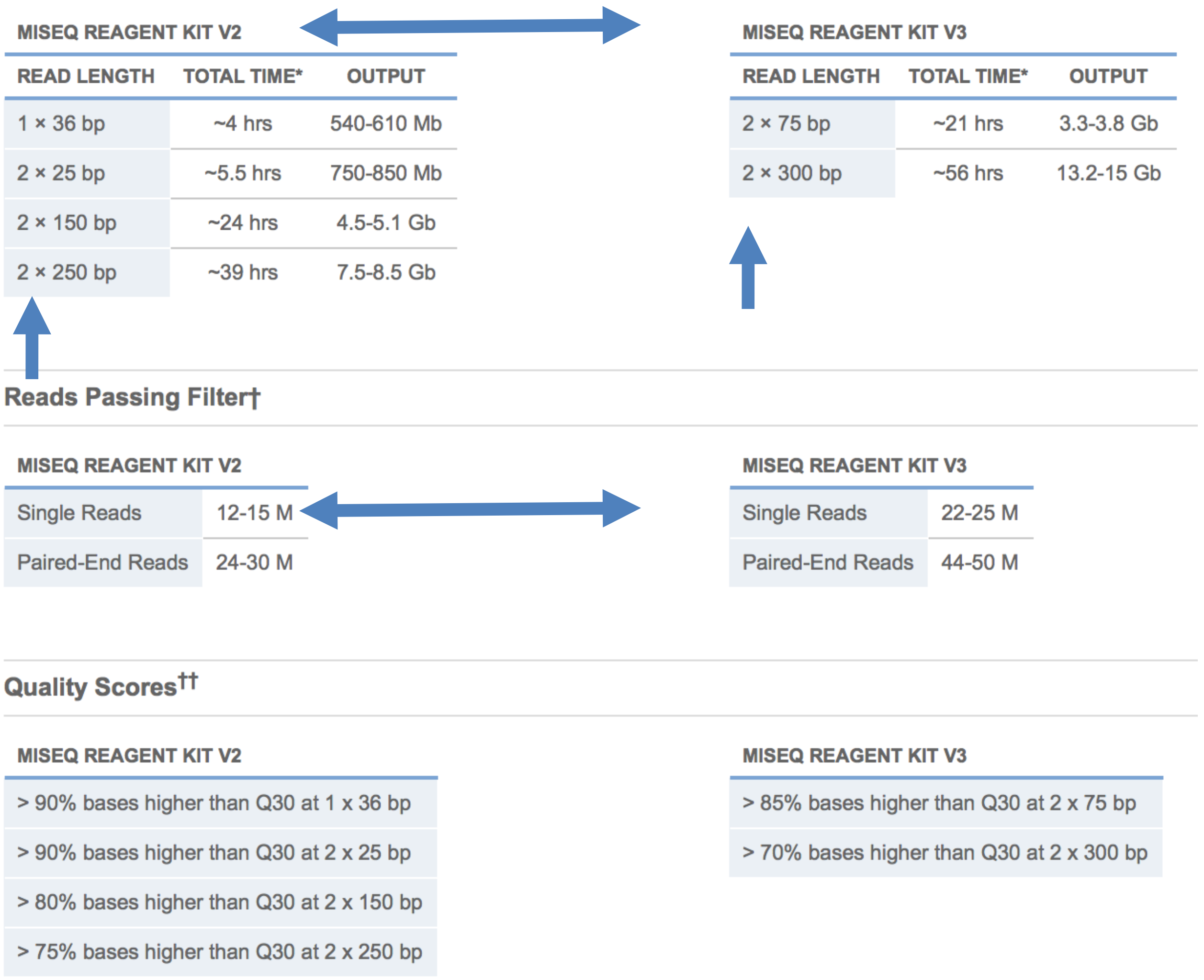
- http://www.illumina.com/systems/hiseq-3000-4000/specifications.html



	HISEQ 3000 SYSTEM	HISEQ 4000 SYSTEM
No. of Flow Cells per Run	1	1 or 2
Data Yield: 2 × 150 bp 2 × 75 bp 1 × 50 bp	 650-750 Gb 325-375 Gb 105-125 Gb	 1300-1500 Gb 650-750 Gb 210-250 Gb
Clusters Passing Filter (Single Reads) (8 lanes per flow cell)	2.1-2.5 billion	4.3-5 billion
Quality Scores: 2 × 50 bp 2 × 75 bp 2 × 150 bp	 ≥ 85% bases above Q30 ≥ 80% bases above Q30 ≥ 75% bases above Q30	 ≥ 85% bases above Q30 ≥ 80% 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 pe	up to 6	up to 12
Exomes per Run**	up to 48	up to 96
Transcriptomes per R	up to 50	up to 100

Illumina MISEQ SEQUENCING

MiSeq



MISEQ REAGENT KIT V2

READ LENGTH	TOTAL TIME*	OUTPUT
1 × 36 bp	~4 hrs	540-610 Mb
2 × 25 bp	~5.5 hrs	750-850 Mb
2 × 150 bp	~24 hrs	4.5-5.1 Gb
2 × 250 bp	~39 hrs	7.5-8.5 Gb

MISEQ REAGENT KIT V3

READ LENGTH	TOTAL TIME*	OUTPUT
2 × 75 bp	~21 hrs	3.3-3.8 Gb
2 × 300 bp	~56 hrs	13.2-15 Gb

Reads Passing Filter†

MISEQ REAGENT KIT V2

Single Reads	12-15 M
Paired-End Reads	24-30 M

MISEQ REAGENT KIT V3

Single Reads	22-25 M
Paired-End Reads	44-50 M

Quality Scores††

MISEQ REAGENT KIT V2

- > 90% bases higher than Q30 at 1 x 36 bp
- > 90% bases higher than Q30 at 2 x 25 bp
- > 80% bases higher than Q30 at 2 x 150 bp
- > 75% bases higher than Q30 at 2 x 250 bp

MISEQ REAGENT KIT V3

- > 85% bases higher than Q30 at 2 x 75 bp
- > 70% bases higher than Q30 at 2 x 300 bp

Cost Estimation

- DNA/RNA extraction and QA/QC (Per sample)
- library preparation (Per sample)
  - Library QA/QC (Bioanalyzer and Qubit)
- 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/

Example: RNA - 12 samples, ribo-depletion libraries, target 30M reads per sample, Hiseq 3000 (2x100).

Cost Estimation

- 12 Samples
  - QA Bioanalyzer = $98 for all 12 samples
  - Library Preparation (ribo-depletion) = $383/sample = $4,596
- Sequencing = $2,346 per lane PE100
  - 2.1 - 2.5 Billion reads per run / 8 lanes = Approximately 300M reads per lane
  - Multiplied by a 0.8 buffer equals 240M expected good reads
  - Divided by 12 samples in the lane = 20M reads per sample per lane.
  - Target 30M reads means 2 lanes of sequencing $2346 x 2 = $4692
- Bioinformatics, simple comparison design, DE only $2000
  - This is the most basic analysis, for in depth collaborative analysis double sequencing budget.

Total = $98 + $4596 + $4692 + $2000 = $11,386

Approximately $950 per sample @ 40M reads per sample