Experimental Design

BCB 504: Applied Bioinformatics

Matt Settles

University of Idaho
Bioinformatics and Computational Biology Program

February 13, 2013

Outline

Sample Preparation

In high throughput biological work (Microarrays, Sequencing, HT Genotyping, etc.), what may seem like small technical artifacts introduced during sample extraction/preparation can lead to large changes, or bias, in the data. Not to say this doesn't occur with smaller scale analysis such as Sanger sequencing or rtPCR, but they do become more appart and may cause significant issues during analysis.

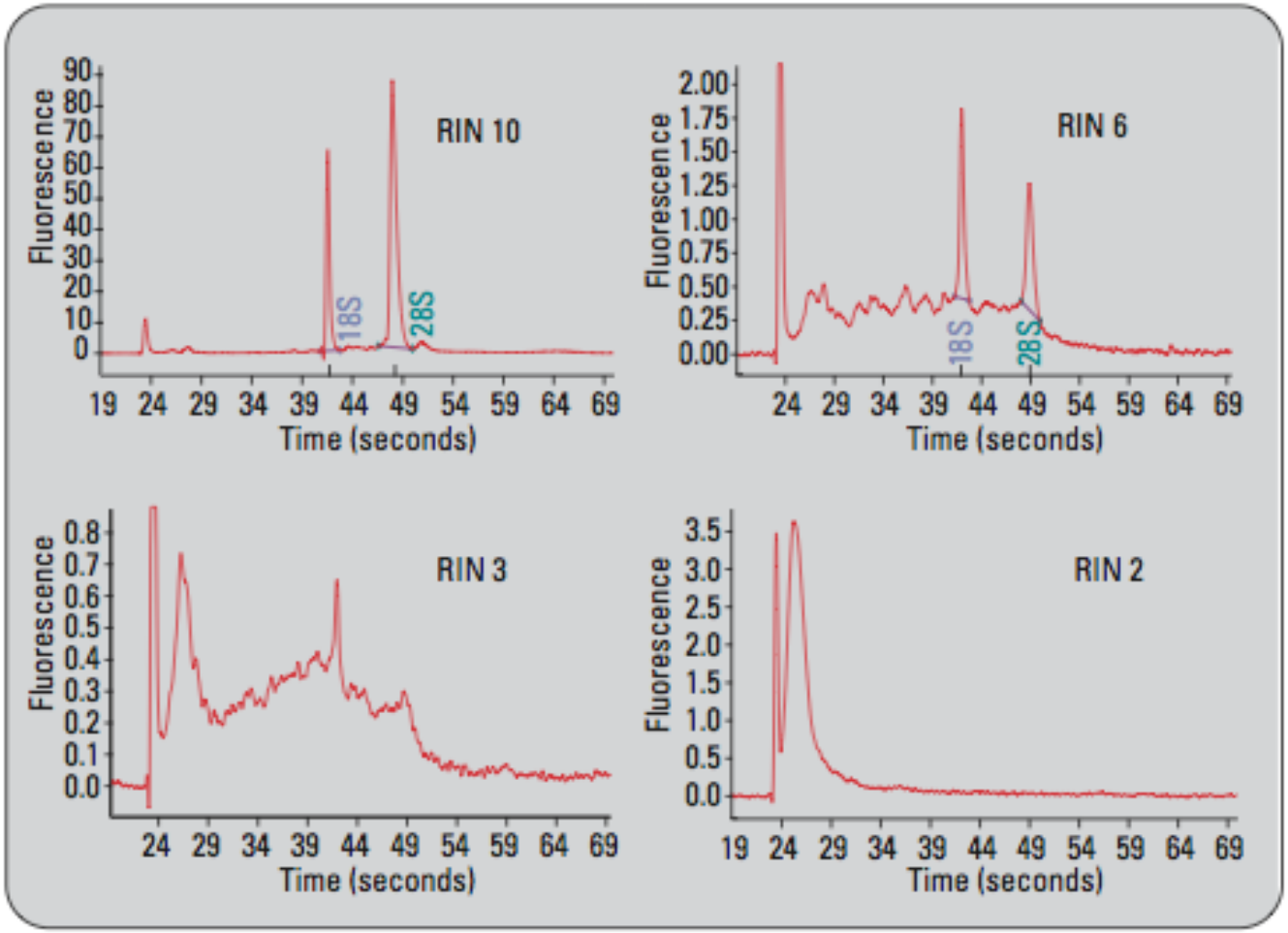
Sample Preparation

Some suggestions

1. Prepare more samples then you are going to need, ie expect some will be of poor quality.
2. Preparation stages should occur across all samples at the same time (or as close to) and by the same person.
3. Spend time practicing to produce the highest quality product you can.
4. Quality and quantity should be established using Bioanalyzer traces (pseudo-gel images), 260/280 & 260/230 $>$ 1.8, and quantified by fluourametry.
5. RNA should not be degrated, Bioanalyzer give RIN numbers to measure relative degradation.

Sample Preparation

The GRC likes to see a minimum RIN value of XX.



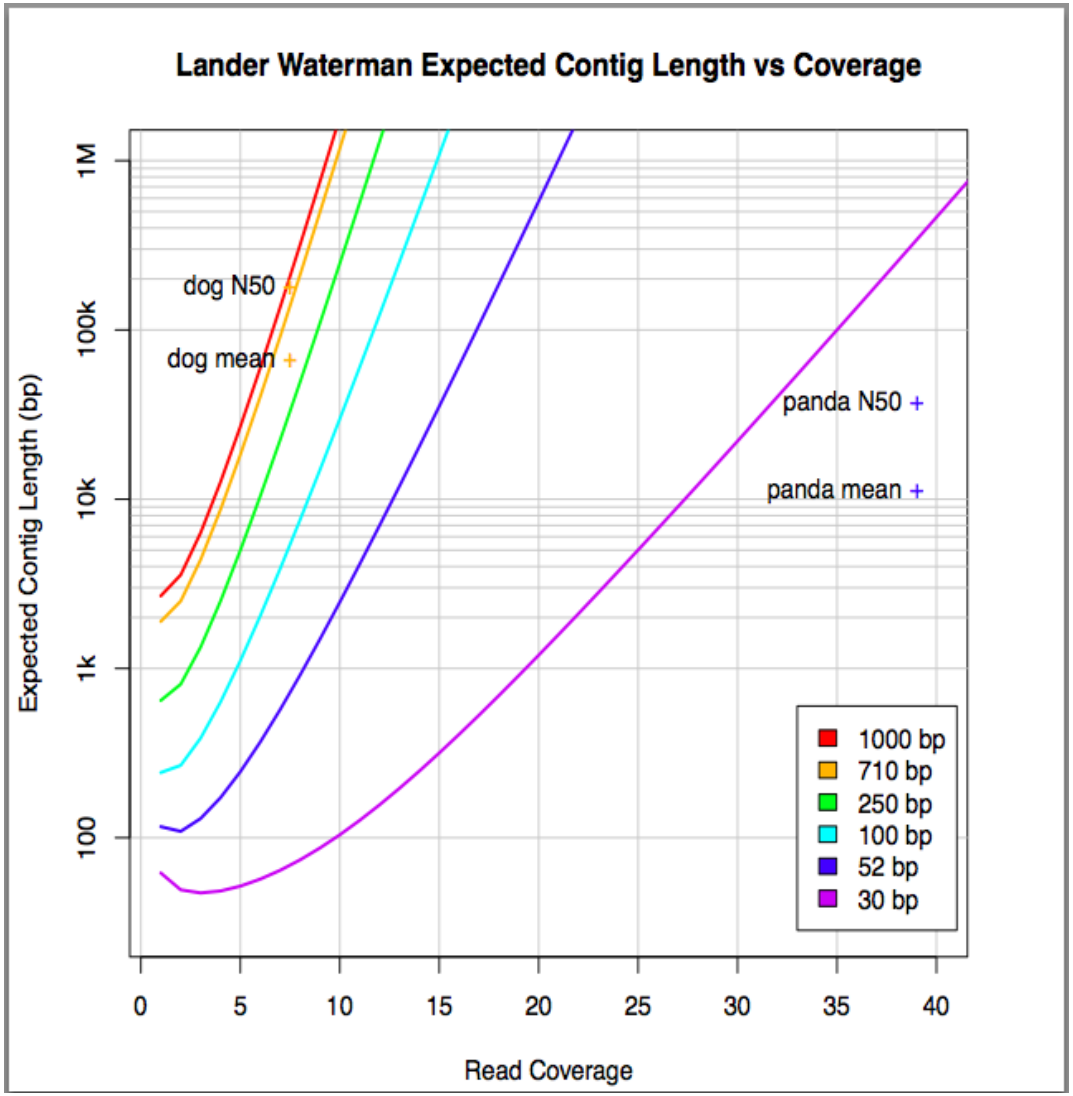
How much coverage do you need (assembly)

Idealized Lander-Waterman model (1988)

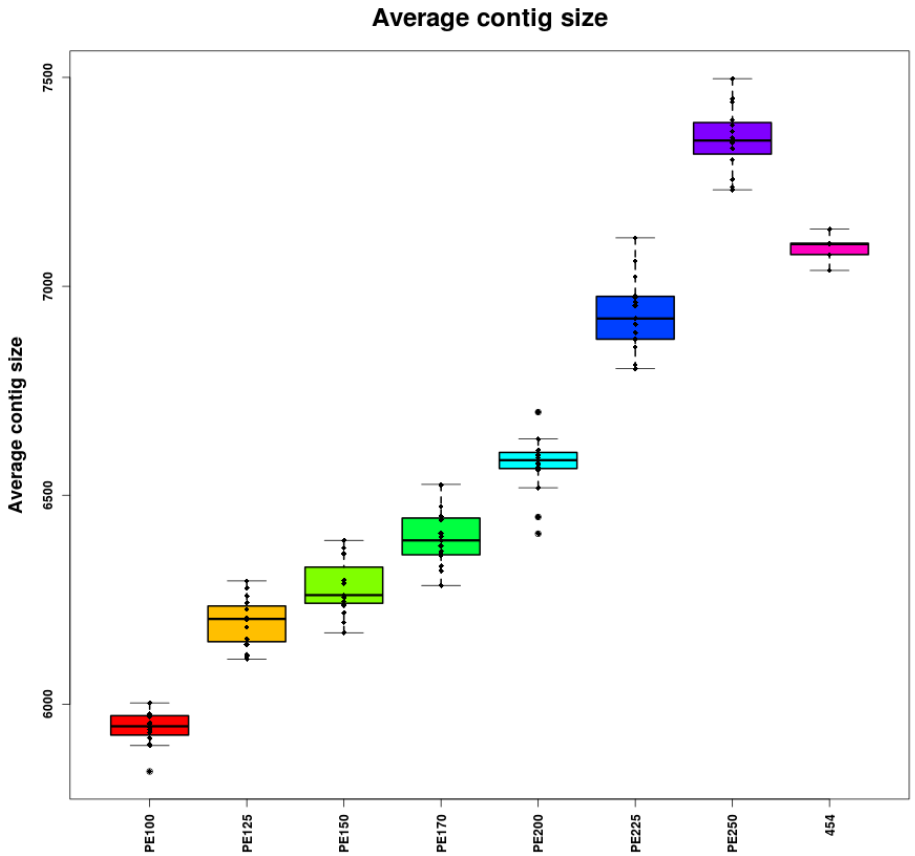
- Reads start at perfectly random position
- Poisson distribution in coverage
- Contig length is a function of coverage and read length

the probability a base is not sequenced is given by: $P_0 = e^{-c}$
where $e = 2.718$, $c =$ fold sequence coverage
$c = LN/G$
$L =$ read length, $N =$ number of reads, $G =$ genome size.
Formula used for the original Human Genome project.
Does not consider read error rate, or genome complexity. Only probability of sequencing a base.
Roach 1995 published an expanded theory.

Lander-Waterman model



Practical Size Considerations



	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
PE100	5839	5926	5947	5942	5972	6003
PE125	6108	6150	6204	6198	6235	6295
PE150	6171	6242	6261	6281	6328	6392
PE170	6284	6358	6392	6401	6446	6526
PE200	6408	6564	6584	6573	6602	6699
PE225	6803	6874	6923	6934	6976	7116
PE250	7231	7316	7349	7353	7392	7497
454	7038	7076	7101	7091	7103	7137

Example Genome Assembly Approach

Approach published Jan. 2012, crocodilian genome, expected completion Jun. 2012. (estimated genome size 2.75Gb.)

- 50x coverage from an overlapping, Illumina, short-insert library.
- 20x coverage from an Illumina 2kb mate-pair library.
- 50x coverage from a non-overlapping 2x100bp, Illumina, short-insert library.
- 1x coverage from a 700bp, 454 library
- 2x coverage from a 3kb and 6kb, 454 library
- BAC end sequencing
- Finally, FISH mapping the BACs to assign scaffolds to chromosomes.

No updates on progress since then, though data is being updated

Coverage by sequencing run

Determine run data (Illumina specs). Get read (or bp) per lane =

Dual Flow Cell / 16
Single Flow Sell / 8
Est. Coverage then is:
$(readsLane * readLength)/(genomeSize * numberOfSamples)$

I usually use low estimate for reads (or bp) and then multiple Est. Coverage by 0.8 for quality.

Also know actual equal proportions between multiplexed samples is rare and can be multiple fold off.

Bottom Line

Bottom Line:
Spend the time (and money) producing good quality, accurate and sufficient data for your experiment. You will spend much more time (and more dollars) trying to pull out biological significant and results in bioinformatic analysis.