Experimental Design

BCB 511: Applied Bioinformatics

Matt Settles

University of Idaho

January 27, 2015

Outline

Sample Preparation

Sequencing Needs

Library Preparation

Sequence Data

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 qRT-PCR, but they do become more apparent and may cause significant issues during analysis.

Sample Suggestions

- 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
- Quality and quantity should be established using Fragment analysis traces (pseudo-gel images).
- DNA/RNA should not be degraded,
- BE CONSISTENT ACROSS ALL SAMPLES

Sequencing Depth

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

Factors to consider are:

1. Number of reads being sequenced
2. Read length (if paired consider then as individuals)
3. Number of samples being sequenced
4. Expected percentage of usable data

$$bpPerSample = \frac{readLength * readCount}{sampleCount} * 0.8$$

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.

Coverage

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

Factors to consider are:

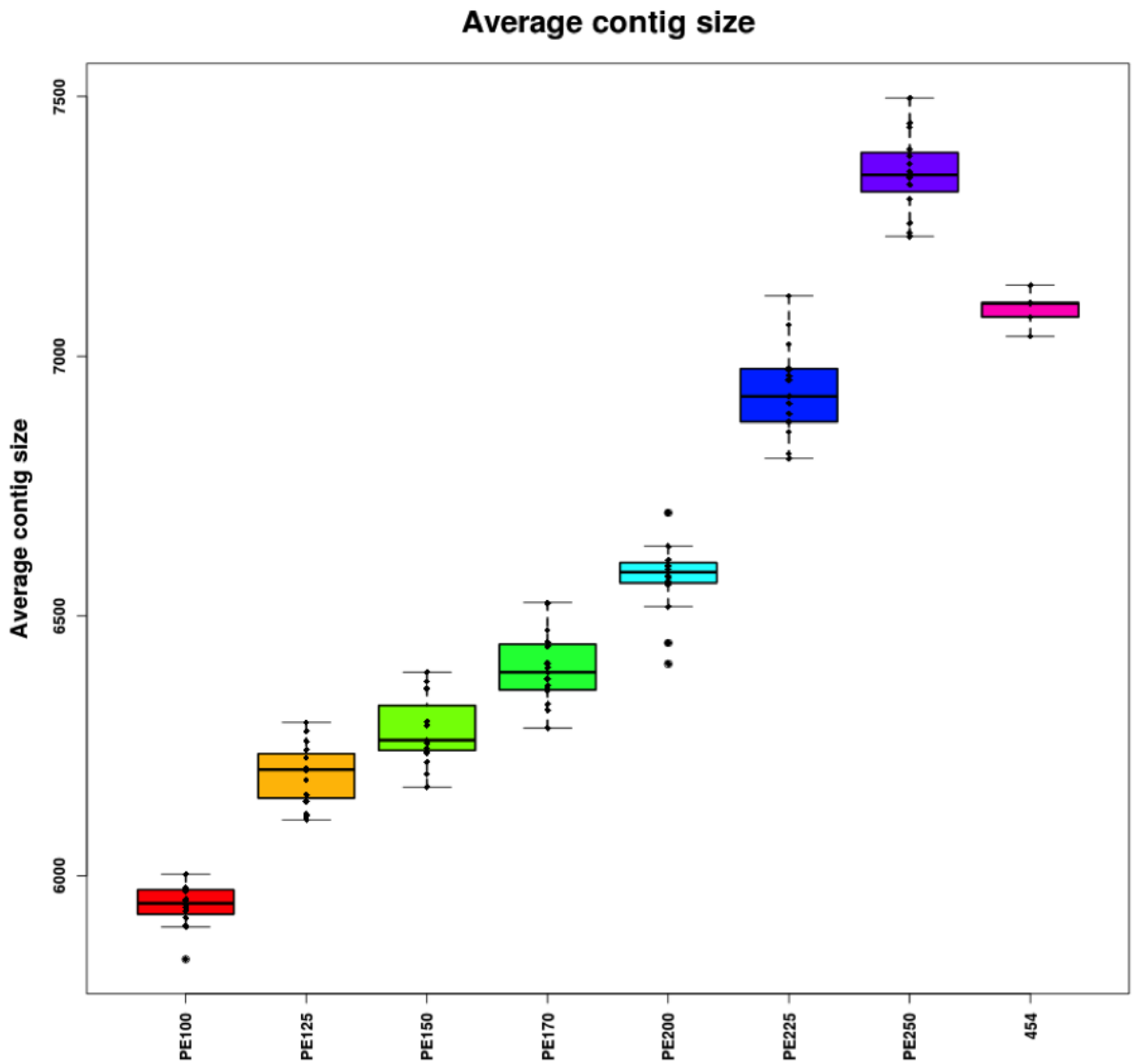
1. Length of the genome (in bp)
2. Any extra-genomic sequence, or contamination

$$expectedCoverage = \frac{bpPerSample}{totalGenomicContent}$$

Coverage is counted differently for "Counting" based experiments (RNAseq, amplicons) where an expected reads per sample is more suitable.

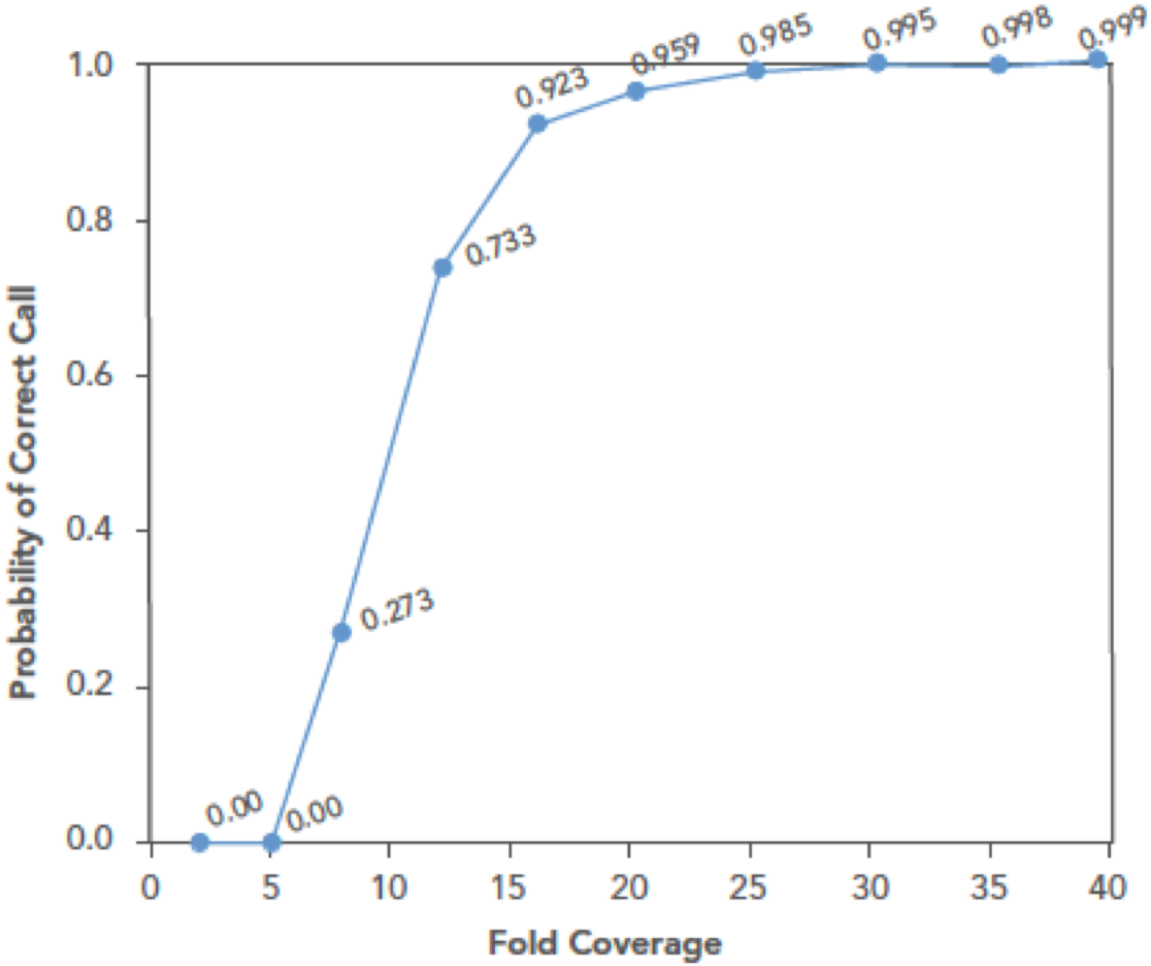
Read length and assembly

Read length matters



Coverage and mapping

Coverage in Variant calling matters



Further, read length contributes to uniqueness of mapping

Metagenomics Coverage

First determine the dilution factor of the rarest species you want to sequence completely. For example if you wish to assemble a species that is present at 1 percent of the community, then your dilution factor is 1 in 100.

$$bpNeeded = \frac{DilutionFactor * AverageGenomeSize * Cov)}{0.8}$$

1. Average microbial genome size estimates
   $2.5x10^6$ lower bound to $5.0x10^6$ mean estimate
2. Coverage, 30x for Illumina

Also consider the expected percentage of "contamination" in your calculation

RNAseq Coverages

Characterization of transcripts or Differential Gene Expression

1. Read length needed depends on likelihood of mapping uniqueness, but generally longer is better and paired-end is better than single-end.
2. Interest in measuring genes expressed at low levels
3. The fold change you want to be able to detect

Uses the same equation as metagenomics (typical average gene sizes is 1500bp), average coverage should be greater

Library Preparation types

- Shotgun - randomly fragmented DNA (100bp - 1kb)
- RNA - Random nanomers or 3' bias (stranded or unstranded)
- Amplicons
- Selected (Capture, RADseq, etc.)
- Paired end / Mate pair
- Synthetic Long Reads

fasta,qual and fastq files

- fasta files

  >sequence1
  ACCCATGATTTGCGA
- qual files

  >sequence1
  40 40 39 39 40 39 40 40 40 40 20 20 36 39 39
- fastq files

  @sequence1
  ACCCATGATTTGCGA
  +
  IIHHIHIIII55EHH

phred scores

$$Q = -10 log_{10} P$$

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%

phred score conversion

$Q_{sanger} = -10 log_{10} P$ - based on probability (aka phred)
$Q_{solexa} = -10 log_{10} \frac{P}{1-P}$ - based on odds

S - Sanger	Phred+33,	raw reads typically (0, 40)
X - Solexa	Solexa+64,	raw reads typically (-5, 40)
I - Illumina 1.3+	Phred+64,	raw reads typically (0, 40)
J - Illumina 1.5+	Phred+64,	raw reads typically (3, 40)
L - Illumina 1.8+	Phred+33,	raw reads typically (0, 41)