Exome Capture
Why Exome (WES, WXS)

• Exome Sequencing attempts to cover all coding regions of the genome. In Humans, the exome is estimated to comprise approximately 1–2% of the genome (30Mb – 60Mb), yet contains ~85% of recognized disease-causing mutations.

• The goal of this approach is to identify genetic variation that is responsible for both Mendelian and common diseases without the high costs associated with whole genome sequencing.

• Lower costs of sequencing per individual means more individuals can be sequenced.

• Capture techniques can only target known sequence
Data Reduction Techniques

Genomic reduction allows for greater coverage and multiplexing of samples.

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

Whole Genome

Reduction Techniques
- RADseq
- GBS
- Fluidigm Access Array
- Amplicons
- Few or Single Amplicons

Capture Techniques

Greater Multiplexing

1X Depth of Coverage

1KB

Whole Genome

Read 1 (50-300bp)

Read 2 (50-300bp)

Barcode (8bp)

Barcode Read primer

Single Multiplexing

Read 2 primer
Capture Libraries

- Capture libraries are a subset of a typical library
Exome-seq, sequencing considerations

Target is even coverage across all exons in a sample

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, particularly for mapping into repetitive regions.

• Depth of sequencing (reads per sample) is primarily dependent on capture content, 60Mb capture requires more reads per sample than 40Mb capture
Sequencing Coverage

• Once you have the number of base pairs per sample you can then determine expected coverage
Factors to consider are:
  1. Total Length of the targeted region (in bp) [total genomic content]
  2. Any extra-genomic sequence, or contamination
  3. Desired Coverage

\[ bpPerSample = totalGenomicContent \times expectedCoverage \]

• Coverage is determined differently for ”Counting” based experiments (RNAseq, amplicons) where an expected number of reads per sample is typically more suitable.
Samples per lane

• The first and most basic question is how many base pairs of sequence data do I need, after you factors to consider are:
  1. Number of reads being sequenced
  2. Read length (if paired consider them as individuals, e.g. 2x150 = 300bp read length)
  3. Basepairs needed per sample
  4. Expected percentage of usable data

\[
\text{Sample Count} = \frac{\text{readLength} \times \text{readCount}}{\text{bpPerSamle}} \times 0.8
\]

• 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.
Illumina sequencing


<table>
<thead>
<tr>
<th></th>
<th>HISEQ 3000 SYSTEM</th>
<th>HISEQ 4000 SYSTEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of Flow Cells per Run</td>
<td>1</td>
<td>1 or 2</td>
</tr>
<tr>
<td>Data Yield:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 x 150 bp</td>
<td>650-750 Gb</td>
<td>1300-1500 Gb</td>
</tr>
<tr>
<td>2 x 75 bp</td>
<td>325-375 Gb</td>
<td>650-750 Gb</td>
</tr>
<tr>
<td>1 x 50 bp</td>
<td>105-125 Gb</td>
<td>210-250 Gb</td>
</tr>
<tr>
<td>Clusters Passing Filter (Single Reads) (8 lanes per flow cell)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.1-2.5 billion</td>
<td>4.3-5 billion</td>
</tr>
<tr>
<td>Quality Scores:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 x 50 bp</td>
<td>≥ 85% bases above Q30</td>
<td>≥ 85% bases above Q30</td>
</tr>
<tr>
<td>2 x 75 bp</td>
<td>≥ 80% bases above Q30</td>
<td>≥ 80% bases above Q30</td>
</tr>
<tr>
<td>2 x 150 bp</td>
<td>≥ 75% bases above Q30</td>
<td>≥ 75% bases above Q30</td>
</tr>
<tr>
<td>Daily Throughput</td>
<td>&gt; 200 Gb</td>
<td>&gt; 400 Gb</td>
</tr>
<tr>
<td>Run Time</td>
<td>&lt; 1-3.5 days</td>
<td>&lt; 1-3.5 days</td>
</tr>
<tr>
<td>Human Genomes per Run*</td>
<td>up to 6</td>
<td>up to 12</td>
</tr>
<tr>
<td>Exomes per Run**</td>
<td>up to 48</td>
<td>up to 96</td>
</tr>
<tr>
<td>Transcriptomes per Run***</td>
<td>up to 50</td>
<td>up to 100</td>
</tr>
</tbody>
</table>

2500 MiSeq
Depth Recommendation

• Depth to target as the average across all exomes in the sample, 100X
• Local depth (individual SNP)
  • 3x Homozygous call
  • 13x Heterozygous gall
• However, the greater the depth the more confident in a call you will be


**Quantifying single nucleotide variant detection sensitivity in exome sequencing.**

Meynert AM¹, Bicknell LS, Hurles ME, Jackson AP, Taylor MS.
Generating Exome-seq libraries

Considerations

• QA/QC of DNA samples
• Which Exome Capture provider (Blocking oligos for Library prep)
• Library Preparation
  • QA
• Exome Capture Protocol
  • How many samples per capture
• Amplification
  • Final QA
QA/QC of DNA samples

Assess DNA Quality on a Bioanalyzer or AATI Fragment Analyzer (best), gel images are ok

BE CONSISTANT!!!
Microarray Platforms are now capture platforms

- Roche Nimblegen
- Agilent
- Illumina
- Mycroarray


http://www.nature.com/nbt/journal/v29/n10/full/nbt.1975.html
Library Preparation

- Fragmentation size should be driven by desired sequencing length, i.e. desire longer fragments for 2x150bp than for 2x100bp
- Library preparation kit choice should be one that does not introduce bias (Nextera has a bias)
Size Selection/Cleanup/QA

• Very important to be consistent across all samples in an experiment on how you size select your final libraries.
• Cleanup/Size select after library generation using SPRI beads or (gel cut)
• QA all samples using an electrophoretic method (Bioanalyzer) and quantify with qPCR.

Most important thing is to be consistent!!!
Exome Capture

• Finite number of capture probes per tube
  • More samples you combine in a tube >> the chance of missing an exon

• The laboratory should also determine the number of samples that can be pooled per sequencing run to achieve the desired coverage level. Coverage here means coverage of the genome.
  • Use recommendation from manufacturer, others who have used the kit and publications.
BE CONSISTANT!

• 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.
Bioinformatics

- Evaluation of Capture Efficiency
  - % of total targeted bases at X coverage

- Analysis of SNP variants, that are ‘causative mutants’

- Regulatory variants in non-coding regions that may effect expression

- Analysis of CNV

- INDEL analysis is not recommended for WES
Ex. Exome-seq Data Analysis Pipeline

- Experiment Metadata
- Sequence Data (fastq)
- Software Setup
- Reference Genome
- Gene/Exon Annotation
- preprocess
- mapping
- variant analysis
- filters
- QA/QC
- Output
Filters

- Preprocessing/Mapping
  - PCR duplicates
  - Multiple mapping (important to align to genome)

- Variant filtering depends on choice of caller and experimental expectations, there will be many
  - Filter on quality of the variant
  - Type of SNP, ex. heterozygous/homozygous
  - SNP is synonymous, nonsynonymous
  - Within a particular gene set

- Domain knowledge and expectations play a key role in final filtering