Impact of Next Generation Sequencing Capabilities on Genomics Based Research

Anup Madan, Mollie McWhorter, Tanja Smith, Anne Ho, Sally Dow, Xiaojun Di, Mark Parrish and Sergey Stepaniants
Covance Genomics Laboratory, Seattle, Washington

Presented at the
Merck Technology Symposium
Edison, New Jersey
5–6 October 2010
Impact of Next Generation Sequencing Capabilities on Genomics Based Research
Anup Madan, Mollie McWhorter, Tanja Smith, Anne Ho, Sally Dow, Xiaojun Di, Mark Parrish and Sergey Stepaniants; Covance Genomic s Laboratory, Seattle, Washington

Abstract
The advent of Next-Generation sequencing technology has fundamentally altered the approach to biological and clinical questions. This technological revolution has revolutionized genomics and opened up new avenues to expand research and development. Covance Genomics Laboratory is a Illumina Certified Service Provider and offers complete sequencing solutions for a range of applications such as genome sequencing, transcriptomics, gene regulation and epigenomics, as well as metagenomics. This poster will present the possible value to a next generation sequencing approach in comparison to traditional genomics research.

Before commencing the project, detailed project information is provided, including guidelines for project set-up, QC parameters for sample preparation and a precise description of the service workflow. We have established pipeline for QC/QA analysis of sequence data, identify SNPs, perform gene expression analysis and evaluate ChIP sequencing data. This bioinformatic analysis of NextGen sequence data plays a key role in realizing the value of next generation sequencing and paving the way to new insights and scientific breakthroughs.

Advantages of NextGen Sequencing

<table>
<thead>
<tr>
<th>Microarray</th>
<th>Sequencing</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNA</td>
<td></td>
</tr>
<tr>
<td>Analag signal</td>
<td>Digital signal</td>
</tr>
<tr>
<td>Restricted to known probes</td>
<td>Greater dynamic range</td>
</tr>
<tr>
<td>Limited structural information</td>
<td>No prior sequence knowledge necessary</td>
</tr>
<tr>
<td>No SNP information</td>
<td>All structural information available</td>
</tr>
<tr>
<td>SNPs in parallel</td>
<td>Potential to measure exon SNPs in parallel</td>
</tr>
</tbody>
</table>

DNA

- Analag signal
- Restricted to known SNPs
- Today max = 1 Million SNPs
- Digital signal
- No prior sequence knowledge necessary
- All data available
- Human = 3 Billion potential SNPs

Sequencing Capabilities at CGL
- 2 Illumina GAIIx Sequencers and 1 SOLID 4 System
- Consensus accuracy >99.99%
- 70-85% of base calls score Q >= 30
- Up to 2 x 100 bp runs yield with greater 70% reads mapping to targeted region

Quality of Sequence Data at CGL

- Majority of bases are sequenced uniformly at same coverage in different library preps showing high reproducibility of data generated at CGL
- 75% of data mapped to targeted regions at various chromosomes suggesting randomness of libraries made at CGL
- Majority of the exons are covered uniformly in targeted enrichment process
- Consistent SNP calls (~ 85 %, with minimum 20X coverage) from different library preparations of same tissue samples demonstrate CGLs ability to produce good quality sequence data.

Quality of Library Made at CGL

Quality of libraries prepared for sequencing was measured by
a) determining number of duplicate inserts in libraries,
b) determining randomness of sequence data,
c) evaluating coverage of various targeted regions, and
d) comparing SNP calls between different sequence libraries from same tissue samples.