Great Interest in Exome Interpretation

Clinical application of exome sequencing in undiagnosed genetic conditions

Anna C Need, Vandana Shashi, Yuki Hitomi, Kelly Schoch, Kevin V Shianna, Marie T McDonald, Miriam H Meisler, David B Goldstein

Bioinformatics for personal genome interpretation

Emidio Coprriotti*, Nathan L. Nehrt*, Maricel G. Kann* and Yana Bromberg*

Submitted: 19th September 2011; Received (in revised form): 8th November 2011

An integrative variant analysis suite for whole exome next-generation sequencing data

Denny Challis†, Jin Yu†, Uday S Evani†, Andrew R Jackson‡, Sameer Paithankar, Cristian Coarfa‡, Aleksandar Milosavljevic‡, Richard A Gibbs* and Fuli Yu†

Whole-Exome Sequencing Identifies Compound Heterozygous Mutations in WDR62 in Siblings With Recurrent Polymicrogyria

David R. Murdock, Gary D. Clark, Matthew N. Bainbridge, Irene Newsham, Yua Donna M. Muzny, Sau Wai Cheung, Richard A. Gibbs, and Melissa B. Ramocki

Whole-Genome Sequencing in a Patient with Charcot–Marie–Tooth Neuropathy

Rational for Automated Pipeline

- Filtering of data to biologically significant targets
- Novel/rare variant detection
- HGSC produces 16-18 Tbases/month
  - WGL can sequence up to 60 Exomes/month
- Quality assurance/control
- Complexity of pipeline
- Manpower
BCM Exome Sequencing Overview

• HGSC Exome Pipeline
  ▪ Now in place at the WGL and HGSC
• Prep Sample DNA
• Enrich for exon sequence – VCrome2.1
• Sequence sample “exome” – HiSeq2000
• Identify variants
• Annotate variants for biological significance
• Provide clinical interpretation through the WGL
Primary Sequence Data Production

...Amplify to form Clusters

Sequencing Instrument

Four-channel Images

Image Processing & Basecalling

Sequence Reads & Qualities (Basecall Confidence)

@HWI-ST115_0142:1:1:9815:1977#0/1
ATTCGAGTCACCTTGACGGCCTGACTGACAAAAAGACAGAGGTAGTGTTGGAT
+HWI-ST115_0142:1:1:9815:1977#0/1
BIIQNPMMOMYZZY[]\Y][][_`^`^`Z`^`^_UVVMURTPPP_ZZZ
Obligatory Cost Per Megabase
Pipeline Design Requirements

- **Modularity**
  - Allow components to be substituted at will

- **LIMS compatibility**
  - Need to communicate with LIMS

- **Automated**
  - As little human intervention as possible

- **Hardened**
  - Graceful error catching with alerts

- **Code reusability - Ruby**
Sequence Data Analysis

~10Gbp of Reads

Align to Reference Genome

Data QA/QC

Call Variants and Estimate Quality

Production & Data QA/QC

Annotate Variants

Cassandra

TGP

dbSNP

HGMD

REPORT
Mercury Pipeline

Data from Illumina Instrument

MCP
- .bcl
- .fastq

CASAVA
- .bam

BAM
- finishing
- final.bam

BWA

Atlas-SNP
- snp.vcf
- indel.vcf

Cassandra SNP
- “final.vcf”
- indel.vcf

SNParray data
- eGenoType concordance

QC

LIMS
Variant Calls & Annotation

- **Variant Calls with Atlas Suite (Fuli Yu)**
  - Logistic regression & Bayesian framework for variant quality
  - Adjustable (very light in WGL) heuristic filtering
  - Genotyping
  - Available on HGSC website & Genboree workbench

- **Annotation with Cassandra (Matthew Bainbridge)**
  - Queries databases of known variation
  - Provides ‘Pileup String’, Atlas QC measures
  - Identifies effect on genes & key regions
  - Includes gene function information to aid interpretation
## Databases Used in Cassandra

<table>
<thead>
<tr>
<th>Database</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SwissProt</td>
<td>SwissProt data, functional description, Expression disease association</td>
</tr>
<tr>
<td>Phase1MAF EXOME</td>
<td>MAF data from Thousand genomes exome</td>
</tr>
<tr>
<td>Phase1MAF WGS</td>
<td>MAF data from Thousand genomes low cov whole genome</td>
</tr>
<tr>
<td>UW MAF</td>
<td>MAF data from the UW Variant Exome Server</td>
</tr>
<tr>
<td>CG MAF</td>
<td>MAF data from the complete genomics 75 whole genomes</td>
</tr>
<tr>
<td>dbNSFP</td>
<td>deleteriousness data from dbNSFP</td>
</tr>
<tr>
<td>Mappability</td>
<td>The mapability of the genome from UCSC</td>
</tr>
<tr>
<td>HGMD snp</td>
<td>SNV data from HGMD</td>
</tr>
<tr>
<td>dbSNP clinical</td>
<td>Clinically implicated variants from dbSNP</td>
</tr>
<tr>
<td>dbSNP</td>
<td>dbSNP variants</td>
</tr>
<tr>
<td>ESE indel</td>
<td>Indels from the HGSC database ESE</td>
</tr>
<tr>
<td>HGMD indel</td>
<td>Indels from HGMD</td>
</tr>
<tr>
<td>dbSNP indel</td>
<td>Indels from dbSNP</td>
</tr>
<tr>
<td>TG indel exome</td>
<td>Indels from Thousand Genomes exome project</td>
</tr>
</tbody>
</table>
# Technical Replication

<table>
<thead>
<tr>
<th>Sample #</th>
<th>Test1</th>
<th>Test1 (%)</th>
<th>Test1 and 2</th>
<th>Test1 and 2 (%)</th>
<th>Test2</th>
<th>Test2 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HS-1011</td>
<td>133</td>
<td>0.561%</td>
<td>23,320</td>
<td>98.409%</td>
<td>244</td>
<td>1.029%</td>
</tr>
<tr>
<td>HS-1015</td>
<td>155</td>
<td>0.542%</td>
<td>28,312</td>
<td>98.910%</td>
<td>157</td>
<td>0.548%</td>
</tr>
<tr>
<td>HS-1016</td>
<td>155</td>
<td>0.644%</td>
<td>23,752</td>
<td>98.732%</td>
<td>150</td>
<td>0.624%</td>
</tr>
<tr>
<td>HS-1017</td>
<td>105</td>
<td>0.456%</td>
<td>22,767</td>
<td>98.768%</td>
<td>179</td>
<td>0.777%</td>
</tr>
<tr>
<td>HS-1018</td>
<td>165</td>
<td>0.693%</td>
<td>23,531</td>
<td>98.795%</td>
<td>122</td>
<td>0.512%</td>
</tr>
<tr>
<td>HS-1019</td>
<td>162</td>
<td>0.682%</td>
<td>23,441</td>
<td>98.686%</td>
<td>150</td>
<td>0.631%</td>
</tr>
<tr>
<td>HS-1020</td>
<td>493</td>
<td>2.041%</td>
<td>23,518</td>
<td>97.355%</td>
<td>146</td>
<td>0.604%</td>
</tr>
<tr>
<td>HS-1021</td>
<td>161</td>
<td>0.681%</td>
<td>23,188</td>
<td>98.125%</td>
<td>282</td>
<td>1.193%</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>191</strong></td>
<td><strong>0.718%</strong></td>
<td><strong>23,979</strong></td>
<td><strong>98.473%</strong></td>
<td><strong>179</strong></td>
<td><strong>0.613%</strong></td>
</tr>
</tbody>
</table>
Possible Sources of Error

- ~1% of sequenced bases are errors
- Repeats, pseudogenes, etc. lead to misplaced reads (mismapping)
- Uneven sampling (statistical sampling errors) lead to undercalled alleles
- Annotation databases imperfect & incomplete
- Imperfect phenotyping can be misleading
# Computational Requirements

<table>
<thead>
<tr>
<th></th>
<th>Casava</th>
<th>BWA</th>
<th>GATK/Picard/SAMtools</th>
<th>Atlas</th>
<th>Cassandra</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>bcl to fastq</td>
<td>final fastq</td>
<td>BWA align</td>
<td>build BAM</td>
<td>Merge</td>
</tr>
<tr>
<td>nodes</td>
<td>1</td>
<td>0.25</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>RAM(GB)</td>
<td>28</td>
<td>8</td>
<td>28</td>
<td>28</td>
<td>28</td>
</tr>
<tr>
<td>hours</td>
<td>8</td>
<td>1</td>
<td>12</td>
<td>24</td>
<td>3</td>
</tr>
<tr>
<td>node*hrs</td>
<td>8</td>
<td>0.25</td>
<td>24</td>
<td>24</td>
<td>3</td>
</tr>
</tbody>
</table>

HGSC Compute Cluster

- 370 Nodes
- 3040 processors
- Moab/Torque
- 4PB Storage
We Need a Bridge

What we do well now:

Primary
Casava, etc.

Secondary
Mapping, BAMS

Tertiary
Annotation, filtering

Hand
Curation

Diagnosis
Example Report

### Table 1. Deleterious Mutations in Disease Genes Related to Clinical Phenotype

<table>
<thead>
<tr>
<th>Disease</th>
<th>Inheritance Pattern</th>
<th>Gene</th>
<th>Isoform</th>
<th>Nucleotide</th>
<th>Amino Acid</th>
<th>Zygosity</th>
<th>References/Comments</th>
<th>Relevance of the Variant to Patient's Current Symptoms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Congenital Variant of Reit Syndrome</td>
<td>Autosomal Dominant</td>
<td>FOGX1</td>
<td>MN9999.9</td>
<td>c.111G&gt;T</td>
<td>p.G37X</td>
<td>Heterozygous</td>
<td>PMID:488888</td>
<td>Likely to be relevant</td>
</tr>
</tbody>
</table>

### Table 2. Variants of Unknown Clinical Significance in Disease Genes Related to Clinical Phenotype

<table>
<thead>
<tr>
<th>Disease</th>
<th>Inheritance Pattern</th>
<th>Gene</th>
<th>Isoform</th>
<th>Nucleotide</th>
<th>Amino Acid</th>
<th>Zygosity</th>
<th>References/Comments</th>
<th>Relevance of the Variant to Patient's Current Symptoms</th>
</tr>
</thead>
<tbody>
<tr>
<td>None Detected</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 3. Medically Actionable Deleterious Mutations in Disease Genes Unrelated to Clinical Phenotype

<table>
<thead>
<tr>
<th>Disease</th>
<th>Inheritance Pattern</th>
<th>Gene</th>
<th>Isoform</th>
<th>Nucleotide</th>
<th>Amino Acid</th>
<th>Zygosity</th>
<th>References/Comments</th>
<th>Variant Classification/Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neurofibromatosis Type 1</td>
<td>Autosomal Recessive</td>
<td>NF1</td>
<td>NM_000492.3</td>
<td>c.3846G&gt;A</td>
<td>p.W334X</td>
<td>Heterozygous</td>
<td>PMID:2236053</td>
<td>Deleterious</td>
</tr>
<tr>
<td>Lynch Syndrome</td>
<td>Autosomal Dominant</td>
<td>PMS2</td>
<td>NM_000535.5</td>
<td>c.990G&gt;T</td>
<td>p.R33X</td>
<td>Heterozygous</td>
<td>PMID:4444444</td>
<td>Deleterious</td>
</tr>
</tbody>
</table>

### Table 4. Carrier Status for Recessive Mendelian Disorders

<table>
<thead>
<tr>
<th>Disease</th>
<th>Inheritance Pattern</th>
<th>Gene</th>
<th>Isoform</th>
<th>Location</th>
<th>Nucleotide</th>
<th>Amino Acid</th>
<th>Zygosity</th>
<th>References/Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cystic Fibrosis, Congenital Absence of the Vas Deferens, CFTR-Related Hereditary Pancreatitis</td>
<td>Autosomal Recessive</td>
<td>CFTR</td>
<td>NM_000492.3</td>
<td>Exon 3</td>
<td>c.3846G&gt;A</td>
<td>p.W1282X</td>
<td>Heterozygous</td>
<td>PMID:2236053</td>
</tr>
</tbody>
</table>

### Table 5. Pharmacogenetic Profile Variant Alleles (optional)

<table>
<thead>
<tr>
<th>Drug</th>
<th>Allele</th>
<th>Zygosity</th>
<th>References/Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plavix</td>
<td>CYP2C19*17</td>
<td>Heterozygous</td>
<td>PMID:16413245; 17625515; 20083681</td>
</tr>
</tbody>
</table>

**Note:** An expanded report of the Whole Exome Sequencing Test is available. The Expanded Report will give additional information on mutations and variants in genes which cause disease unrelated to the indication for testing and predicted deleterious mutations in genes with no known current association with disease. If this information is requested by the physician, please complete Requisition Form and Patient Consent for Expanded Report.
Future Directions

- **Cancer Pipeline**
  - Tumor/Normal, Tumor/Normal/Normal
- **More Instrument Integration**
- **Variant calls for list of VIPs**
- **Addition of More Annotation Databases**
- **More Filtering Tools**
- **Automation of Physician Assisted Diagnosis**
  - NLP, Heuristics
Acknowledgements

- Code will be open source.

- Art Beaudet & the whole WGL team
- Jeff Reid, Richard Gibbs, & Donna Muzny
- Eric Boerwinkle, Fuli Yu, & Matthew Bainbridge
- Peter Pham & Mark Wang
- Alicia Hawes, Ziad Khan, & Mike Dahdouli