An integrated annotation system to support whole-exome sequencing experiments design and data management

Ivan Limongelli, Angelo Nuzzo, Annalisa Vetro, Erika Della Mina, Roberto Ciccone, Orsetta Zuffardi, Riccardo Bellazzi

IRCCS C. Mondino, Pavia
Dipartimento di Genetica Medica
Dipartimento di Informatica e Sistemistica
Exome Sequencing – Analysis Workflow

Nucleotide sequences (short reads)

@HWUSI-EAS703_0001:1:1:1009:17571#0/2
AAGCAAAAGATCCATAGCCACGCAACAGAATGTGAATCTCTGGCATAATTGGTGTG
+
SDDDDDDD:DAD:DDDDDDDDDDDDDD:AD=D;
@HWUSI-EAS703_0001:1:1:1009:12086#0/2
TCCACAGCTTGTCATCTATTTCTTTCCACTCTAGGGGCACCGAGTGTG
+
@HWUSI-EAS703_0001:1:1:1010:6169#0/2
GGTCAGTCAAGGTTGACTCTTCAGGATTTACGTAACGTACGTACGTTAGT
+

Primary Analysis

Secondary Analysis
Exome Sequencing – Secondary Analysis

- Variants Detection

SNVs

Short In-dels
Secondary Analysis – Variants Annotation

**Public bio-databases**

*chr1, 153170600, A>G, NM_015383, NBPF14, I>R, ...*

**Does this variant affect the product (protein structure and function) coded in this region?**
Secondary Analysis – Variants Prediction

- Some open tools addressing this aim:
  - *Polyphen2*, *Mutation Taster*, *SIFT*, *Annovar*, *Sequence variant Analyzer*
- Suitable for loss-of-function mutations
- Score assignment to each mutation corresponding to its probability to damage protein
- Principally based on type of amino acid substitution, conservation across species, polymorphisms databases
Secondary Analysis – Data management

• Some considerations
  ▪ Illumina GAIIx platform: max 8 whole-exome samples sequenced in each experiment
  ▪ 13-18K variants per sample (SNVs/Indels) are detected, but about 95% of them are common variants

• Needs
  ▪ **I would like** to easily perform cross-samples and cross-experiments analysis
  ▪ **I would NOT like** to annotate and predict changes again for previously (already annotated) identified variants
Secondary Analysis – Data management

Output by software device manufacturer

Pipeline 1
Pipeline 2
Pipeline N

MiddelWare
(data uploads)

Applications Module

DB

-omics DB

Synchronising

Functional prediction Tools
Secondary Analysis – Data management

Step 1: create experiment
Secondary Analysis – Data management

Step 2: choose experiment to add a sample
Secondary Analysis – Data management

NGSDB Web Site

**EXPLORE**
- Experiments
- Samples
- Mutations

**NEW**
- Experiment
- Sample
- Mutations

**ANALYSIS**
- Cases Vs Controls

**MANAGE**
- Update
  - UCSC
  - DBSnp

DbSnp FTP
Software
Theses

**Sample - Creation**
- **Sex**
  - M

- **Code**
  - 1234-11

- **Add Sample**

**Upload VCF file**
- **Vcf File**

- **Reference build**
  - hg19

- **Upload**

- **Step 3: create sample**
- **Step 4: upload mutation data for that sample**
Secondary Analysis – Data management

Step 5: select cases/controls
Secondary Analysis – Data management

Step 5: filtering parameters setup
### Secondary Analysis – Data management

<table>
<thead>
<tr>
<th>SampleDescription</th>
<th>Reference</th>
<th>Chromosome</th>
<th>Position</th>
<th>refUCSC</th>
<th>Observed</th>
<th>Genotype</th>
<th>Quality</th>
<th>Depth Of Coverage Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>GM11-126</td>
<td>hg19</td>
<td>1</td>
<td>566048</td>
<td>G</td>
<td>A</td>
<td>hom</td>
<td>42.2</td>
<td>0</td>
</tr>
<tr>
<td>GM11-126</td>
<td>hg19</td>
<td>1</td>
<td>567579</td>
<td>C</td>
<td>T</td>
<td>hom</td>
<td>49.3</td>
<td>0</td>
</tr>
<tr>
<td>GM11-126</td>
<td>hg19</td>
<td>1</td>
<td>808631</td>
<td>G</td>
<td>A</td>
<td>hom</td>
<td>36.65</td>
<td>1</td>
</tr>
<tr>
<td>GM11-126</td>
<td>hg19</td>
<td>1</td>
<td>909238</td>
<td>G</td>
<td>C</td>
<td>hom</td>
<td>43.95</td>
<td>0</td>
</tr>
<tr>
<td>GM11-126</td>
<td>hg19</td>
<td>1</td>
<td>909238</td>
<td>G</td>
<td>C</td>
<td>hom</td>
<td>43.95</td>
<td>0</td>
</tr>
<tr>
<td>GM11-126</td>
<td>hg19</td>
<td>1</td>
<td>909238</td>
<td>G</td>
<td>C</td>
<td>hom</td>
<td>43.95</td>
<td>0</td>
</tr>
<tr>
<td>GM11-126</td>
<td>hg19</td>
<td>1</td>
<td>1956399</td>
<td>C</td>
<td>T</td>
<td>het</td>
<td>1089.48</td>
<td>48</td>
</tr>
<tr>
<td>GM11-126</td>
<td>hg19</td>
<td>1</td>
<td>1956579</td>
<td>G</td>
<td>A</td>
<td>het</td>
<td>464.87</td>
<td>20</td>
</tr>
<tr>
<td>GM11-126</td>
<td>hg19</td>
<td>1</td>
<td>1957037</td>
<td>T</td>
<td>C</td>
<td>het</td>
<td>2970.61</td>
<td>111</td>
</tr>
<tr>
<td>GM11-126</td>
<td>hg19</td>
<td>1</td>
<td>154541971</td>
<td>T</td>
<td>G</td>
<td>het</td>
<td>142.39</td>
<td>56</td>
</tr>
<tr>
<td>GM11-126</td>
<td>hg19</td>
<td>1</td>
<td>154744807</td>
<td>C</td>
<td>G</td>
<td>het</td>
<td>6490.54</td>
<td>243</td>
</tr>
<tr>
<td>GM11-126</td>
<td>hg19</td>
<td>1</td>
<td>154744807</td>
<td>C</td>
<td>G</td>
<td>het</td>
<td>6490.54</td>
<td>243</td>
</tr>
<tr>
<td>GM11-126</td>
<td>hg19</td>
<td>1</td>
<td>154744807</td>
<td>C</td>
<td>G</td>
<td>het</td>
<td>6490.54</td>
<td>243</td>
</tr>
<tr>
<td>GM11-126</td>
<td>hg19</td>
<td>1</td>
<td>154744852</td>
<td>A</td>
<td>G</td>
<td>het</td>
<td>6271.88</td>
<td>202</td>
</tr>
<tr>
<td>GM11-126</td>
<td>hg19</td>
<td>1</td>
<td>154744852</td>
<td>A</td>
<td>G</td>
<td>het</td>
<td>6271.88</td>
<td>202</td>
</tr>
<tr>
<td>GM11-126</td>
<td>hg19</td>
<td>1</td>
<td>154744937</td>
<td>C</td>
<td>T</td>
<td>het</td>
<td>2371.62</td>
<td>121</td>
</tr>
<tr>
<td>GM11-126</td>
<td>hg19</td>
<td>1</td>
<td>154744937</td>
<td>C</td>
<td>T</td>
<td>het</td>
<td>2371.62</td>
<td>121</td>
</tr>
<tr>
<td>GM11-126</td>
<td>hg19</td>
<td>1</td>
<td>154745031</td>
<td>T</td>
<td>G</td>
<td>het</td>
<td>1249.55</td>
<td>46</td>
</tr>
<tr>
<td>GM11-126</td>
<td>hg19</td>
<td>1</td>
<td>154745031</td>
<td>T</td>
<td>G</td>
<td>het</td>
<td>1249.55</td>
<td>46</td>
</tr>
<tr>
<td>GM11-126</td>
<td>hg19</td>
<td>1</td>
<td>154745031</td>
<td>T</td>
<td>G</td>
<td>het</td>
<td>1249.55</td>
<td>46</td>
</tr>
</tbody>
</table>
# Secondary Analysis – Data management

<table>
<thead>
<tr>
<th>Depth of coverage/mutation</th>
<th>UCSC transcript</th>
<th>UCSC gene symbol</th>
<th>UniProt Protein</th>
<th>IdDbSnip</th>
<th>Region</th>
<th>change_type</th>
<th>AA_change</th>
<th>Polyphen Prediction</th>
<th>pph_prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>unknown</td>
<td>null</td>
<td>null</td>
<td>rs6421780</td>
<td>genomic</td>
<td>na</td>
<td>na</td>
<td>null</td>
<td>null</td>
</tr>
<tr>
<td>2</td>
<td>uc001aaz.2</td>
<td>BC018860</td>
<td>null</td>
<td>rs112232512</td>
<td>utr</td>
<td>na</td>
<td>na</td>
<td>null</td>
<td>null</td>
</tr>
<tr>
<td>2</td>
<td>uc001abt.3</td>
<td>FAM41C</td>
<td>null</td>
<td>rs112040799</td>
<td>intron</td>
<td>na</td>
<td>na</td>
<td>null</td>
<td>null</td>
</tr>
<tr>
<td>2</td>
<td>uc001acd.2</td>
<td>PLEKH1</td>
<td>Q494U1-2</td>
<td>rs3829740</td>
<td>exon</td>
<td>missense</td>
<td>R&gt;P</td>
<td>benign</td>
<td>0.0</td>
</tr>
<tr>
<td>2</td>
<td>uc001ace.2</td>
<td>PLEKH1</td>
<td>Q494U1</td>
<td>rs3829740</td>
<td>exon</td>
<td>missense</td>
<td>R&gt;P</td>
<td>benign</td>
<td>0.0</td>
</tr>
<tr>
<td>2</td>
<td>uc001acf.2</td>
<td>PLEKH1</td>
<td>Q494U1-3</td>
<td>rs3829740</td>
<td>exon</td>
<td>missense</td>
<td>R&gt;P</td>
<td>benign</td>
<td>0.0</td>
</tr>
<tr>
<td>43</td>
<td>uc001aip.2</td>
<td>GABRD</td>
<td>Q14764</td>
<td>rs79685811</td>
<td>exon</td>
<td>synonymous</td>
<td>G&gt;G</td>
<td>null</td>
<td>null</td>
</tr>
<tr>
<td>21</td>
<td>uc001aip.2</td>
<td>GABRD</td>
<td>Q14764</td>
<td>null</td>
<td>exon</td>
<td>synonymous</td>
<td>G&gt;G</td>
<td>null</td>
<td>null</td>
</tr>
<tr>
<td>127</td>
<td>uc001aip.2</td>
<td>GABRD</td>
<td>Q14764</td>
<td>rs2229110</td>
<td>exon</td>
<td>synonymous</td>
<td>G&gt;G</td>
<td>null</td>
<td>null</td>
</tr>
<tr>
<td>68</td>
<td>uc001aff.2</td>
<td>CHRNB2</td>
<td>Q5SXY3</td>
<td>null</td>
<td>exon</td>
<td>missense</td>
<td>V&gt;G</td>
<td>probably damaging</td>
<td>0.96</td>
</tr>
<tr>
<td>197</td>
<td>uc001aff.2</td>
<td>KCNN3</td>
<td>Q8WXG7</td>
<td>rs1051614</td>
<td>exon</td>
<td>synonymous</td>
<td>L&gt;L</td>
<td>null</td>
<td>null</td>
</tr>
<tr>
<td>197</td>
<td>uc001aff.2</td>
<td>KCNN3</td>
<td>Q6JXY2</td>
<td>rs1051614</td>
<td>exon</td>
<td>synonymous</td>
<td>L&gt;L</td>
<td>null</td>
<td>null</td>
</tr>
<tr>
<td>197</td>
<td>uc009wox.1</td>
<td>KCNN3</td>
<td>Q6JXY2</td>
<td>rs1051614</td>
<td>exon</td>
<td>synonymous</td>
<td>L&gt;L</td>
<td>null</td>
<td>null</td>
</tr>
<tr>
<td>208</td>
<td>uc001aff.2</td>
<td>KCNN3</td>
<td>Q8WXG7</td>
<td>rs1131820</td>
<td>exon</td>
<td>synonymous</td>
<td>N&gt;N</td>
<td>null</td>
<td>null</td>
</tr>
<tr>
<td>208</td>
<td>uc001aff.2</td>
<td>KCNN3</td>
<td>Q6JXY2</td>
<td>rs1131820</td>
<td>exon</td>
<td>synonymous</td>
<td>N&gt;N</td>
<td>null</td>
<td>null</td>
</tr>
<tr>
<td>208</td>
<td>uc009wox.1</td>
<td>KCNN3</td>
<td>Q6JXY2</td>
<td>rs1131820</td>
<td>exon</td>
<td>synonymous</td>
<td>N&gt;N</td>
<td>null</td>
<td>null</td>
</tr>
<tr>
<td>96</td>
<td>uc001aff.2</td>
<td>KCNN3</td>
<td>Q8WXG7</td>
<td>rs11589471</td>
<td>intron</td>
<td>na</td>
<td>na</td>
<td>null</td>
<td>null</td>
</tr>
<tr>
<td>96</td>
<td>uc001aff.2</td>
<td>KCNN3</td>
<td>Q6JXY2</td>
<td>rs11589471</td>
<td>intron</td>
<td>na</td>
<td>na</td>
<td>null</td>
<td>null</td>
</tr>
<tr>
<td>96</td>
<td>uc009wox.1</td>
<td>KCNN3</td>
<td>Q6JXY2</td>
<td>rs11589471</td>
<td>intron</td>
<td>na</td>
<td>na</td>
<td>null</td>
<td>null</td>
</tr>
<tr>
<td>44</td>
<td>uc001aff.2</td>
<td>KCNN3</td>
<td>Q8WXG7</td>
<td>rs11584403</td>
<td>intron</td>
<td>na</td>
<td>na</td>
<td>null</td>
<td>null</td>
</tr>
<tr>
<td>44</td>
<td>uc001aff.2</td>
<td>KCNN3</td>
<td>Q6JXY2</td>
<td>rs11584403</td>
<td>intron</td>
<td>na</td>
<td>na</td>
<td>null</td>
<td>null</td>
</tr>
<tr>
<td>44</td>
<td>uc009wox.1</td>
<td>KCNN3</td>
<td>Q6JXY2</td>
<td>rs11584403</td>
<td>intron</td>
<td>na</td>
<td>na</td>
<td>null</td>
<td>null</td>
</tr>
</tbody>
</table>

**NGSDB Web Site**

**EXPLORE**
- Experiments
- Samples
- Mutations

**NEW**
- Experiment
- Sample
- Mutations

**ANALYSIS**
- Cases Vs Controls

**MANAGE**
- Update
- UCSC
- DBSnp

**DbSnip FTP**

**Software**

**Theses**
Secondary Analysis – Data management

- Current prototype implementation allows to:
  - Store experiments and samples data
  - Store identified variants (SNVs/Indels) and their reliability parameters (VCF 4.0 currently supported)
  - Annotate variants
  - Predict their probability to damage protein and store results (Polyphen2, Mutation Taster, SIFT)
  - Control-case studies modelling
The end..

Thank you for your attention!