Variant Detection & Interpretation in a diagnostic context

Christian Gilissen

c.gilissen@gen.umcn.nl

28-05-2013
So far…

Sequencing

Mapping

Variant calling

Interpretation

Johan den Dunnen
Marja Jakobs
Ewart de Bruijn

Victor Guryev

Laurent Francioli
What to interpret?

- Variants → SNVs and small indels
- 3 million SNVs per individual genome
- 20,000 to 50,000 variants per individual exome

How to identify variants that are involved in a patient’s disease?
Variant interpretation

1. Annotation of variants
2. Strategies for prioritization
3. Computational prediction of pathogenicity
Part I - Interpretation of exome data

• An initial approach:

~150-500 private non-synonymous variants
Annotation

• Publicly available sources
  • SeattleSeq, Annovar, Vaast, Ensembl AP, SNPEff, dbNSFP

• Commercial packages:
  • CLC Bio
  • NextGene
  • Cartagenia
  • Ingenuity VA

• Home-made software

• All tools:
  • Effect of variant on protein coding gene
  • Overlap with databases of polymorphisms
What can you get?

- **SeattleSeq** ([http://snp.gs.washington.edu/SeattleSeqAnnotation/](http://snp.gs.washington.edu/SeattleSeqAnnotation/))
  - Conservation scores, Polyphen predictions, on-line
  - No indels, input format is very specific

- **Annovar** ([http://www.openbioinformatics.org/annovar/](http://www.openbioinformatics.org/annovar/)):
  - Pro: Sift (old) and polyphen predictions
  - Con: local install required → web interface now available: **wAnnovar**

- **Vaast** ([http://www.yandell-lab.org/software/vaast.html](http://www.yandell-lab.org/software/vaast.html)):
  - Pro: statistic framework for candidate gene selection
  - Con: local install required, no indels (yet)

- **Ensembl API** ([http://www.ensembl.org/info/docs/api/variation/index.html](http://www.ensembl.org/info/docs/api/variation/index.html))
  - Pro: flexible
  - Con: requires installation and programming, not all data available

  - Pro: fast, indels, multiple species
  - Con: local install, only does effect on protein
Variant frequency sources

• **dbSNP**: largest dataset, but polluted

• **1000 genomes**: frequencies available but from cell-lines

• **ESP database**: no indels, patients, no validation

• **Published studies**: GONL, Complete genomics genomes

• **In house databases / DVD**: population/sequencing specific variants
ESP6500 variants for ASXL1

<table>
<thead>
<tr>
<th>Variant Pos</th>
<th>rs ID</th>
<th>Alleles</th>
<th>EA Allele #</th>
<th>AA Allele #</th>
<th>All Allele #</th>
<th>Avg. Sample Read Depth</th>
<th>Genes</th>
<th>mRNA Accession #</th>
<th>GVS Function</th>
<th>Amino Acid</th>
</tr>
</thead>
<tbody>
<tr>
<td>20:3121163</td>
<td>rs145699348</td>
<td>A/G</td>
<td>A=1/G=8599</td>
<td>A=0/G=4406</td>
<td>A=1/G=13005</td>
<td>81</td>
<td>ASXL1</td>
<td>NM_015338.5</td>
<td>missense</td>
<td>ILE, VAL</td>
</tr>
<tr>
<td>20:3121190</td>
<td>unknown</td>
<td>T/C</td>
<td>T=1/C=8599</td>
<td>T=0/C=4406</td>
<td>T=1/C=13005</td>
<td>90</td>
<td>ASXL1</td>
<td>NM_015338.5</td>
<td>missense</td>
<td>CYS, ARG</td>
</tr>
<tr>
<td>20:3121211</td>
<td>unknown</td>
<td>T/C</td>
<td>T=0/C=8600</td>
<td>T=2/C=4404</td>
<td>T=2/C=13004</td>
<td>92</td>
<td>ASXL1</td>
<td>NM_015338.5</td>
<td>stop-gained</td>
<td>stop, ARG</td>
</tr>
<tr>
<td>20:3121232</td>
<td>rs14864681</td>
<td>T/C</td>
<td>T=1/C=8599</td>
<td>T=0/C=4406</td>
<td>T=1/C=13005</td>
<td>90</td>
<td>ASXL1</td>
<td>NM_015338.5</td>
<td>missense</td>
<td>CYS, ARG</td>
</tr>
<tr>
<td>20:3121233</td>
<td>rs143719307</td>
<td>A/G</td>
<td>A=0/G=8600</td>
<td>A=1/G=4405</td>
<td>A=1/G=13005</td>
<td>90</td>
<td>ASXL1</td>
<td>NM_015338.5</td>
<td>missense</td>
<td>HIS, ARG</td>
</tr>
<tr>
<td>20:3121250</td>
<td>unknown</td>
<td>T/C</td>
<td>T=1/C=8599</td>
<td>T=0/C=4406</td>
<td>T=1/C=13005</td>
<td>87</td>
<td>ASXL1</td>
<td>NM_015338.5</td>
<td>stop-gained</td>
<td>stop, ARG</td>
</tr>
<tr>
<td>20:3121324</td>
<td>unknown</td>
<td>C/T</td>
<td>C=0/T=8600</td>
<td>C=1/T=4405</td>
<td>C=1/T=13005</td>
<td>98</td>
<td>ASXL1</td>
<td>NM_015338.5</td>
<td>coding-synonymous</td>
<td>none</td>
</tr>
<tr>
<td>20:3121332</td>
<td>unknown</td>
<td>G/C</td>
<td>G=1/C=8599</td>
<td>G=0/C=4405</td>
<td>G=1/C=13005</td>
<td>97</td>
<td>ASXL1</td>
<td>NM_015338.5</td>
<td>stop-gained</td>
<td>stop, SER</td>
</tr>
<tr>
<td>20:3121337</td>
<td>unknown</td>
<td>A/G</td>
<td>A=0/G=8600</td>
<td>A=0/G=4405</td>
<td>A=1/G=13005</td>
<td>97</td>
<td>ASXL1</td>
<td>NM_015338.5</td>
<td>missense</td>
<td>ILE, VAL</td>
</tr>
<tr>
<td>20:3121364</td>
<td>unknown</td>
<td>A/G</td>
<td>A=2/G=8596</td>
<td>A=0/G=4406</td>
<td>A=2/G=13004</td>
<td>102</td>
<td>ASXL1</td>
<td>NM_015338.5</td>
<td>coding-synonymous</td>
<td>none</td>
</tr>
<tr>
<td>20:3121369</td>
<td>unknown</td>
<td>A/G</td>
<td>A=2/G=8596</td>
<td>A=1/G=4405</td>
<td>A=2/G=13004</td>
<td>104</td>
<td>ASXL1</td>
<td>NM_015338.5</td>
<td>missense</td>
<td>ASN, SER</td>
</tr>
<tr>
<td>20:3121430</td>
<td>rs141346625</td>
<td>C/G</td>
<td>C=2/G=8598</td>
<td>C=17/G=4369</td>
<td>C=19/G=12987</td>
<td>107</td>
<td>ASXL1</td>
<td>NM_015338.5</td>
<td>missense</td>
<td>GLN, GLU</td>
</tr>
<tr>
<td>20:3121466</td>
<td>rs142172134</td>
<td>G/C</td>
<td>G=0/C=8600</td>
<td>G=10/C=4396</td>
<td>G=10/C=12996</td>
<td>106</td>
<td>ASXL1</td>
<td>NM_015338.5</td>
<td>missense</td>
<td>GLY, ARG</td>
</tr>
<tr>
<td>20:3121475</td>
<td>rs14591372</td>
<td>C/G</td>
<td>C=0/G=8600</td>
<td>C=2/G=4404</td>
<td>C=2/G=13004</td>
<td>109</td>
<td>ASXL1</td>
<td>NM_015338.5</td>
<td>missense</td>
<td>PRO, ALA</td>
</tr>
<tr>
<td>20:3121527</td>
<td>rs138971201</td>
<td>A/T</td>
<td>A=0/T=8600</td>
<td>A=14/T=4392</td>
<td>A=14/T=12992</td>
<td>131</td>
<td>ASXL1</td>
<td>NM_015338.5</td>
<td>missense</td>
<td>ASN, VAL</td>
</tr>
<tr>
<td>20:3121544</td>
<td>unknown</td>
<td>A/G</td>
<td>A=0/G=8600</td>
<td>A=1/G=4405</td>
<td>A=1/G=13005</td>
<td>140</td>
<td>ASXL1</td>
<td>NM_015338.5</td>
<td>missense</td>
<td>MET, VAL</td>
</tr>
</tbody>
</table>

Bohring-Opitz syndrome is often fatal in early childhood.

http://evs.gs.washington.edu/EVS/
Not just interpretation: also QC

% called variants in dbSNP
QC from annotation: Tr/Ti

Transitions/Transversions
QC from annotation: stop mutations

# stop mutations

![Bar chart showing the distribution of stop mutations across different samples or categories. The x-axis represents different categories or samples, and the y-axis represents the number of stop mutations. The chart shows a variation in the number of stop mutations across the categories.]
Other (common) annotations:

- **Variant based:**
  - Grantham / substitution scores
  - HGMD
  - Protein domains
  - Protein level conservation
  - Repeat

- **Gene based**
  - OMIM (disease gene),
  - MGI: Mouse knock-out phenotypes / zebrafish knock-out
  - Kegg pathways and GO biological processes
  - Loss off function gene
Protein-protein interactions

How to use?
• Simulate 100 exomes with a “spiked-in” mutation in a deafness gene
• Raking of variants using PPI and conservation compared to only on conservation
Interpretation of non-coding variants

• Many more variants, much less information

• What can you use?
  • Evolutionary conservation
  • Overlap with regulator regions (Encode)
  • Proximity to known genes

• Similar ways of reducing the candidates as exome analysis: *de novo* variants, family analysis
Part II – Strategies to prioritize variants from exome studies

Linkage based strategy
Homozygosity based strategy
Double-hit based strategy

Overlap based strategy
De novo based strategy
Candidate based strategy
Linkage strategy

- Select variants that segregate with the disease or lie within a region that segregates with the disease
- Applies to both dominant and recessive disorders

1. Overlap / exclude variants from family members
   Two affected siblings, reducing the number of candidates to 9 genes.\(^1\)

2. Determine regions of Identity By Descent
   Three affected siblings, reducing the number from 14 to 2 genes.\(^2\)

\(^1\)Ng et al. Exome sequencing identifies the cause of a mendelian disorder. Nat Genet. 2010
Homozygosity strategy

• Select variants that lie within a large homozygous region of the patient

Reduced the number of homozygous candidate variants from 17 to 3.³

Double hit strategy

- Select variants that are homozygous or compound-heterozygous in the patient

- Applies only to recessive disorders (with no consanguinity)

- A single exome can be sufficient, \(^4,^5\) reducing the number of candidates from 139 and 158 to 3 and 4 respectively.

\(^4\)Pierce et al. Mutations in the DBP-deficiency protein HSD17B4 cause ovarian dysgenesis, hearing loss, and ataxia of Perrault Syndrome. Am J Hum Genet. 2010

Overlap strategy

- Select unrelated patients and determine variants in multiple patients in the same gene\textsuperscript{6,7}
- Used for rare sporadic dominant disorders
- Depends crucially on good phenotyping
- Disorder must be monogenic
- Three individuals can be enough to pinpoint a single gene.\textsuperscript{8}

\textsuperscript{6}Hoischen \textit{et al.} De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. Nat Genet. 2010
\textsuperscript{7}Ng \textit{et al.} Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nat Genet. 2010
\textsuperscript{8}Hoischen \textit{et al.} De novo nonsense mutations in ASXL1 cause Bohring-Opitz syndrome. Nat Genet. 2011
**De novo strategy**

- Exome sequencing an affected patient and his unaffected parents and select variants that are not inherited.\(^9\),\(^10\),\(^11\)

- Applies to sporadic disorders with large genetic heterogeneity

- Methods for detecting *de novo* mutations enrich for sequencing and analysis errors.

\(^9\)Vissers *et al.* A *de novo* paradigm for mental retardation. Nat Genet. 2010
\(^10\)O’roak *et al.* Exome sequencing in sporadic autism spectrum disorders identifies severe *de novo* mutations. Nat Genet. 2011
\(^11\)Xu et al. Exome sequencing supports a *de novo* mutational paradigm for schizophrenia. Nat Genet. 2011
Prioritization of candidate de novo variants

- 38 not validated in proband
  → Median variant reads: 5

- 13 validated: 9 de novo!!!
  → Median variant reads: 17

Systematic validation using Sanger sequencing

<table>
<thead>
<tr>
<th>MR trio</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>average</th>
</tr>
</thead>
<tbody>
<tr>
<td>High confidence variant calls</td>
<td>20,810</td>
<td>21,658</td>
<td>21,338</td>
<td>22,647</td>
<td>17,694</td>
<td>22,333</td>
<td>21,369</td>
<td>22,658</td>
<td>24,085</td>
<td>22,962</td>
<td>21,755</td>
</tr>
</tbody>
</table>

n=51
Candidate strategy

• Selection of variants based on variant and gene interpretation

• Traditional gene prioritization techniques\(^\text{12}\)

• Variant interpretation: Polyphen, SIFT, Mutpred, etc.

• Evolutionary conservation

\(^\text{12}\)Erlich et al. Exome sequencing and disease-network analysis of a single family implicate a mutation in KIF1A in hereditary spastic paraparesis. Genome Res. 2011
Evolutionary conservation for variant prioritization
Part III – Computational Predictions

- **Polyphen2**: Bayesian classification based on sequence/structure attributes and MSA ([http://genetics.bwh.harvard.edu/ppyh2/](http://genetics.bwh.harvard.edu/ppyh2/))

- **Mutpred**: Random forest classification on protein structure attributes and evolutionary attributes. ([http://mutpred.mutdb.org/](http://mutpred.mutdb.org/))

- **SIFT**: probability of substitution tolerance based on MSA ([http://sift.jcvi.org/](http://sift.jcvi.org/))

- **Mutation taster**: Naïve bayes classifier, sequence distribution and protein domains ([http://www.mutationtaster.org/](http://www.mutationtaster.org/))
Performance comparison of prediction programs

Thusberg et al/ Hum mut. 2011
Prediction on 57 blindness variants

Neveling et al. *Hum mut.* 2012
Conclusions

- Open source annotation tools available for variant annotation

- Think about your method of prioritization before starting any experiments. Most successful studies:
  - Clear Mendelian disorders
  - Good control dataset
  - Family members available for follow up
  - Cohort available for finding recurrence

- Pathogenicity prediction can help but should be used with care.
All families & clinicians involved!