Variants prioritization: annotation and filtration steps

Variant Effect Prediction Training Course
Heraklion, Greece, 2016
Jean-Pierre Desvignes & Christophe Béroud,
Aix-Marseille University
INSERM UMR_S910 Medical Genetics and Functional Genomics
Context

- Next generation Sequencing has facilitated the discovery of new genes and genetic variants in a multitude of human disorders
- 1st Whole-Exome Sequencing (WES) done by Ng et al 2009 in Miller Syndrome
- In 2013, >150 Mendelian disorders were studied by WES (Rabbani et al, J Hum Genet, 2014)
- IRDiRC and Orphanet ➔ 3500 genes involved in RD, >1000 identified between 2010 and 2014 (IRDiRC website)
...But...

- Despite all these encouraging figures
  - Only 23-26% of WES are successful (higher rate if several individuals from the same family are sequenced 34-37% for a trio) [Farwell et al. 2015, Sawyer et al. 2016]
    - Technical factors (homopolymers, GC reach regions, poor quality at read ends …)
    - Type of disease causing mutations (not captured, triplet repeat expansions, CNVs, pseudogenes …)
    - Bioinformatics pipeline to generate VCF (same sequencing technology, not same VCF)
    - Wrong annotations/filtrations
Variant annotation

- Part of the data analysis process
  - Mandatory for prioritization and filtration of variants
- Two objectives
  - Help to refine our estimate of how likely a variant is to be true, genotype, quality …
  - Provide functional annotations to determine the links between a genetic variation and a disease

Adapted from Pabinger et al. Briefings in Bioinformatics 2014
Variant annotation

- It is performed at various levels
Variant annotation

Quality data
- Quality score, quality filter,
- Genotype, read depth, ...

Mutation localization
- Exonic, intronic, splice site, 3'UTR, 5'UTR, ...
- Mutation type and HGVS nomenclature (g., c., p.)

Mutation frequency
- EVS, dbSNP, 1000G, ExAC, ...

Pathogenicity predictions
- SIFT, Polyphen, CADD, Mutation taster,
- UMD-Predictor...
Variant annotation

**Variant level**

**Gene level**

- **Expression profile**
  - GTEx, Expression Atlas, RNA-seq data, ...

- **Gene Ontology**
  - Molecular function, cellular component, biological process

- **Biological Pathways**
  - KEGG, BioCarta, WikiPathways, Reactome
Variant annotation

Involvement in Human diseases
OMIM, UNIPRO, LSDB (LOVD, UMD, ...), Clinvar,
COSMIC
Animal models (mutants, knockdowns, transgenics)
MGI, Zfin, FlyBase, ...

Phenotype level
Gene level
Variant level
Variant annotation
### Variant annotation systems

<table>
<thead>
<tr>
<th>Availability</th>
<th>Annovar</th>
<th>SNPeff</th>
<th>Ensembl VEP</th>
<th>SeattleSeq</th>
<th>AnnTools</th>
<th>Oncotator</th>
<th>Vanno</th>
<th>Variant Annotation Tools</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Command line</td>
<td>Command line</td>
<td>Command line</td>
<td>web</td>
<td>command line</td>
<td>Command line</td>
<td>Web</td>
<td>Command line</td>
</tr>
<tr>
<td>Variant quality</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>-</td>
<td>Yes</td>
<td>-</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Variant localization</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Gene/transcript annotation</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Genotype</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>-</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Population frequency</td>
<td>Yes</td>
<td>-</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Impact at the RNA level</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Impact at the protein level</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Conservation</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>-</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Reported impact</td>
<td>-</td>
<td>-</td>
<td>Yes</td>
<td>Yes</td>
<td>-</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Predicted pathogenicity</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>-</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Gene ontology</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Yes</td>
<td>Yes</td>
<td>-</td>
</tr>
<tr>
<td>Pathways</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Yes</td>
<td>-</td>
<td>-</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Tissue expression</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Various types of annotation software are available (command line/web)
No system is providing annotations at all levels ➞ need to be combined
Variant filtration

Identify disease causing/private mutations

Adapted from Pabinger et al. Briefings in Bioinformatics 2014
# Filtration tools

## Automatic systems

- **Disease causing genes based on pedigree and phenotypic data**

<table>
<thead>
<tr>
<th>Software name</th>
<th>Availability</th>
<th>Mode of inheritance</th>
<th>Custom analysis</th>
<th>Mutation localization</th>
<th>Mutation type</th>
<th>Mutation frequency</th>
<th>Pathogenicity predictions</th>
<th>Functional evidences</th>
<th>Clinical report</th>
<th>Prioritization score</th>
</tr>
</thead>
<tbody>
<tr>
<td>ExomeWalker</td>
<td>Web App</td>
<td>Yes</td>
<td>-</td>
<td>-</td>
<td>Yes</td>
<td>No but provided</td>
<td>No but provided</td>
<td>Yes</td>
<td>Yes</td>
<td><strong>Yes</strong></td>
</tr>
<tr>
<td>Exomiser</td>
<td>Command line</td>
<td>Yes</td>
<td>-</td>
<td>-</td>
<td>Yes</td>
<td>Yes</td>
<td>-</td>
<td>Yes</td>
<td>Yes</td>
<td><strong>Yes</strong></td>
</tr>
<tr>
<td>eXtasy</td>
<td>Command line</td>
<td>Yes</td>
<td>-</td>
<td>-</td>
<td>Yes</td>
<td>No but provided</td>
<td>No but provided</td>
<td>-</td>
<td>Yes</td>
<td><strong>Yes</strong></td>
</tr>
<tr>
<td></td>
<td>Web App</td>
<td></td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>No - But provided</strong></td>
<td></td>
</tr>
<tr>
<td>MirTRIOS</td>
<td>Web App</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>-</td>
<td>No but provided</td>
<td>Yes</td>
<td><strong>Yes</strong></td>
</tr>
<tr>
<td>OMIM Explorer</td>
<td>Web App</td>
<td>Yes*</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Yes</td>
<td>Yes</td>
<td><strong>Yes</strong></td>
</tr>
<tr>
<td>OVA</td>
<td>Web App</td>
<td>Yes</td>
<td>-</td>
<td>Yes</td>
<td>Yes (2)</td>
<td>-</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td><strong>Yes</strong></td>
</tr>
<tr>
<td>wKGGSeq</td>
<td>Web App</td>
<td>Yes</td>
<td>-</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td><strong>Yes</strong></td>
</tr>
</tbody>
</table>

- Fast and accurate method for known genes/diseases
- Automatically gather additional information
- Work only with known genes/diseases with annotations
- Limited flexibility

* Does not combine multiple samples
### Filtration tools

- **Semi automatic/manual systems**
  - Users can select candidate mutations by applying various set of filters

<table>
<thead>
<tr>
<th>Software name</th>
<th>Availability</th>
<th>Mode of inheritance</th>
<th>Custom analysis</th>
<th>Mutation localization</th>
<th>Mutation type</th>
<th>Mutation frequency</th>
<th>Pathogenicity predictions</th>
<th>Functional evidences</th>
<th>Clinical report</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANNOVAR</td>
<td>Command line</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>-</td>
<td>Yes</td>
</tr>
<tr>
<td>BIERapp</td>
<td>Web App</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>FITUS</td>
<td>GUI</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes - if provided</td>
<td>Yes - if provided</td>
<td>Yes - if provided</td>
<td>Yes - if provided</td>
<td>Yes - if provided</td>
<td>Yes - if provided</td>
</tr>
<tr>
<td>FMFilter</td>
<td>GUI</td>
<td>Yes</td>
<td>-</td>
<td>Yes - if provided</td>
<td>Yes - if provided</td>
<td>Yes - if provided</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Gemini</td>
<td>Command line</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Web App</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes - if provided</td>
<td>Yes - if provided</td>
<td>Yes - if provided</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Vanno</td>
<td>Web App</td>
<td>-</td>
<td>-</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No but provided</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>VarAFT</td>
<td>GUI</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No - But provided</td>
</tr>
<tr>
<td>VarSifter</td>
<td>Command line</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes - if provided</td>
<td>Yes - if provided</td>
<td>Yes - if provided</td>
<td>Yes - if provided</td>
<td>Yes - if provided</td>
<td>Yes - if provided</td>
</tr>
<tr>
<td></td>
<td>Web App</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes - if provided</td>
<td>Yes - if provided</td>
<td>Yes - if provided</td>
<td>Yes - if provided</td>
<td>Yes - if provided</td>
<td>Yes - if provided</td>
</tr>
<tr>
<td>VCF-MINER</td>
<td>Local Web App</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes - if provided</td>
<td>Yes - if provided</td>
<td>Yes - if provided</td>
<td>Yes - if provided</td>
<td>Yes - if provided</td>
<td>Yes - if provided</td>
</tr>
</tbody>
</table>

- Provide a good flexibility and traceability
- Can be tedious and difficult
- No filtration at all annotation levels
For more information

_Hum Mutat._ 2016 Sep 7. doi: 10.1002/humu.23110. [Epub ahead of print]

How to Identify Pathogenic Mutations among All Those Variations: Variant Annotation and Filtration in the Genome Sequencing Era.

Salgado D¹, Bellgard M², Desvignes JP⁴, Béroud C⁴,⁵.
Challenges

Variant annotation and filtration are not a simple and solved problem

Variant Annotation

- Require the management of multiple and large data sources (local)
- Incorrect or incomplete annotations can cause researchers to overlook and dilute interesting variants in a pool of false positives
- From Davis McCarthy (Genome Medicine 2014)
  - Choice of transcript set and software can have a large effect on the variant annotation
    - Matching annotation for ANNOVAR on Ensembl or RefSeq genesets is only 83% for all exonic variants (lof, missenses, synonymous …)
    - Comparison between ANNOVAR and VEP on Ensembl transcripts 87% of all exonic variants were in agreement
    - Even more discrepancies with splicing variants
- Where a specific tissue of interest is known, restrict annotation to transcript known to be expressed in that tissue (GENCODE)
Challenges

- **Variant filtration/prioritization**
  - Good *phenotypic description* of patients
  - *Mode of inheritance* should be known (identify the right model, hypothesis on the penetrance …)
  - **No Gold standard** but frequently used filters
    - Frequency in the population
    - Genotype
    - Mutation type / Pathogenicity
    - Need to be done interactively (add/remove filters)
  - **Discrepancies between pathogenicity predictors** (may introduce false +/-)
  - **Population frequency** (not all databases gather healthy individuals, ethnicity) e.g. presence in dbSNP does not mean “polymorphism”
  - **Privacy issue** may arise when using “online” system
Our systems

To improve and facilitate disease causing mutation identification

Pathogenicity prediction systems

variant annotation

Variant annotation and filtration
Pathogenicity prediction system for any Human cDNA substitutions

- Precomputed all possible substitutions for all nucleotides of any human transcripts (280,315,899 substitutions)
- Combined multiple features in a unique score (0-100)
  - AA change – substitution and biochemical matrices (BLOSUM/Yu)
  - Exonic splicing signal (HSF - Acceptors and Donors splice sites)
  - Protein key residues (UNIPROT HCD)
  - Conserved and functional domains -100 species protein alignments (Phastcons) + Grantham
  - Allele frequency – dbSNP

Available through a web application and webservices
UMD-Predictor Evaluation

- **4 datasets** (more than 140,000 mutations) - Varibench + dbSNP, Uniprot, ClinVar and PredictSNP
- **7 references pathogenicity predictors**

Varibench + dbSNP = 17,329 variations

<table>
<thead>
<tr>
<th></th>
<th>SIFT</th>
<th>PPH2</th>
<th>Provean</th>
<th>Mutation Assessor</th>
<th>CONDEL</th>
<th>Mutation Taster</th>
<th>CADD</th>
<th>UMD-Predictor</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP</td>
<td>9596</td>
<td>10290</td>
<td>9638</td>
<td>9775</td>
<td>8797</td>
<td><strong>11174</strong></td>
<td>10182</td>
<td>10727</td>
</tr>
<tr>
<td>TN</td>
<td>2805</td>
<td>3045</td>
<td>3068</td>
<td>3162</td>
<td>3287</td>
<td>2937</td>
<td>3214</td>
<td><strong>4024</strong></td>
</tr>
<tr>
<td>FP</td>
<td>1229</td>
<td>1189</td>
<td>1147</td>
<td>1073</td>
<td>948</td>
<td><strong>1298</strong></td>
<td>1021</td>
<td>211</td>
</tr>
<tr>
<td>FN</td>
<td>3498</td>
<td>2803</td>
<td>3456</td>
<td>3319</td>
<td>4297</td>
<td><strong>1929</strong></td>
<td>2912</td>
<td>2367</td>
</tr>
</tbody>
</table>

- **Sensitivity** 0.73, 0.79, 0.74, 0.75, 0.67, 0.85, 0.78, 0.82
- **Specificity** 0.70, 0.72, 0.73, 0.75, 0.78, 0.69, 0.76, 0.95
- **DOR** 6.3, 9.7, 7.7, 9.0, 7.2, 12.6, 11.2, 86.6
- **log(DOR)** 1.84, 2.27, 2.04, 2.20, 1.97, 2.53, 2.42, 4.46

Similar results with other datasets (Cf. Salgado, Desvignes, et al. Human Mutation 2016)

DOR : Measure the effectiveness of a diagnostic test
Trade-off between sensitivity and specificity
UMD-Predictor Evaluation

Comparison using 3 WES performed in a clinical diagnostic context

<table>
<thead>
<tr>
<th></th>
<th>SIFT\textsuperscript{A}</th>
<th>PPH2\textsuperscript{A}</th>
<th>Provean\textsuperscript{A}</th>
<th>Mutation Assessor\textsuperscript{A,D}</th>
<th>CONDEL\textsuperscript{A,B}</th>
<th>Mutation Taster</th>
<th>CADD\textsuperscript{A}</th>
<th>UMD-Predictor</th>
</tr>
</thead>
<tbody>
<tr>
<td>NV1</td>
<td>1958</td>
<td>2881</td>
<td>1540</td>
<td>1339</td>
<td>1376</td>
<td>2677</td>
<td>3241</td>
<td>871</td>
</tr>
<tr>
<td>NV2</td>
<td>1341</td>
<td>2350</td>
<td>1332</td>
<td>1049</td>
<td>1111</td>
<td>2437</td>
<td>2555</td>
<td>540</td>
</tr>
<tr>
<td>NV3</td>
<td>1842</td>
<td>2781</td>
<td>1542</td>
<td>1350</td>
<td>1376</td>
<td>3401</td>
<td>3098</td>
<td>807</td>
</tr>
</tbody>
</table>

Shortest list of potential pathogenic mutations

Time required to process VCF files

<table>
<thead>
<tr>
<th></th>
<th>SIFT\textsuperscript{A}</th>
<th>PPH2\textsuperscript{A}</th>
<th>Provean\textsuperscript{A}</th>
<th>Mutation Assessor\textsuperscript{A,D}</th>
<th>CONDEL\textsuperscript{A,B}</th>
<th>Mutation Taster</th>
<th>CADD\textsuperscript{A}</th>
<th>UMD-Predictor</th>
</tr>
</thead>
<tbody>
<tr>
<td>PT1 (s)</td>
<td>1200</td>
<td>420</td>
<td>3240</td>
<td>540</td>
<td>3000</td>
<td>2100</td>
<td>8700</td>
<td>93</td>
</tr>
<tr>
<td>PT2 (s)</td>
<td>240</td>
<td>420</td>
<td>8100</td>
<td>960</td>
<td>1500</td>
<td>2340</td>
<td>9360</td>
<td>206</td>
</tr>
<tr>
<td>PT3 (s)</td>
<td>540</td>
<td>420</td>
<td>4140</td>
<td>600</td>
<td>1500</td>
<td>2340</td>
<td>11160</td>
<td>240</td>
</tr>
</tbody>
</table>

Even faster by using webservices

Fastest system to process variations from VCF files
Human Splicing Finder

- Pathogenicity prediction system for any mutations on splicing signals
Human Splicing Finder

Source: Splicing regulation: From a parts list of regulatory elements to an integrated splicing code, Wang, Burge - RNA, 2008
Human Splicing Finder

- Pathogenicity prediction system for **any mutations on splicing signals**
- **Reference system** ( >611 citations since 2009, Web of Science)
- A “One stop-Shop” system
  - Splicing signals
  - Branch points
  - Auxiliary signals (ESE, ESS, ISE, ISS)
- Combine various predictive systems, matrices and specific algorithms
- **Expert system for data interpretation** (establishment of rules to provide a conclusion) e.g. **MSH2 c.274_276del**

- Compliant with **NGS technologies:** webservices
Human Splicing Finder
Evaluation

HSF

Sroogle

MaxEnt

SpliceView

SplicePredictor

SplicePort

GeneSplicer

NNSplice
VARiant Annotation and Filtration Tool

http://varaft.eu

- An user-friendly variant annotation and filtration tool
- Standalone multiplatform application
- Evaluation of the data coverage for WGS/WES/panel
- One click annotation (based on ANNOVAR) and other sources
- Only system to provide UMD-Predictor and HSF annotations
- Interactive filtration at all levels
  - Combine multiple samples
  - Automatic selection of variants (mode of inheritance)
  - Genetic population studies
  - Cancers
- Optimal selection of candidate pathogenic mutations
Variant Annotation and Filtration Tool
Conclusions

- No ideal annotation/filtration systems
  - Many of them are built to be used by bioinformaticians (command line systems)
- To maximize chances of identify causing mutations
  - be aware of challenges posed by each steps of the data analysis pipeline
  - use family members (when possible)
  - collect exact and complete phenotypic information
- Many challenges remain to be solved for both annotation and filtration systems - benchmarking initiatives
- Most of the current available systems were created for WES and need to be adapted to WGS
- Many more issues arise with WGS
  - Annotation for non-coding regions but also for non-protein coding genes
  - Need to develop and improve pathogenicity prediction system for non-exonic mutations