Annotation Visualization and Impact Analysis
AVIA

Hue Vuong, Anney Che, Uma Mudunuri
Advanced Biomedical Computational Science
Feb 27, 2020
Advanced Biomedical Computational Science

- Bioinformatics - CCBR, SF, NCBR
- Data mining, integration
- Infrastructure - biomedical databases, software
- Scientific web programming
- Imaging and Visualization
- Structural biology
- Computational chemistry
- Statistical analysis

https://ncifrederick.cancer.gov/dsitp/abcc/abcc-groups/
Overview

• Background
  – Sequencing and variants
  – Variant annotations
  – Impact analysis

• Demo
  – Single sample example
  – Multi sample example
  – Registered users
    • Project management
    • Cohort annotations
NGS Workflow

Lab
Library Preparation, Exome Capturing, ...

Whole-Exome-Seq
Whole-Genome-Seq

NGS Platform
Illumina, SOLiD, 454, ...

Quality Assessment
Trimming, Filtering, ...

Read Alignment
Reference Genome

Variant Identification
SNP, INDEL, CNV, SV

Germline / Somatic
Structural Variants

Annotation
Public Databases, Conservation Scores, ...

Visualization
Variant, Annotation, ...

Genome Browser
Circos Plots

Prioritization / Filtering

Lab
Validation

Stephan Pabinger et al. Brief Bioinform 2013;bib.bbs086

Frederick National Laboratory for Cancer Research
Reference Genomes in AVIA

• Human
  – UCSC hg19 – NCBI GRCh37 (current)
  – UCSC hg38 – NCBI GRCh38

• Mouse
  – UCSC mm10 – GRCm38
Variant Uploads

Variant Call Format (VCF) is the preferred format

- **Position and alt allele**
- Contains info about variant (e.g., counts, allele frequencies (AF), depth, etc)

Format and Sample go hand in hand

Sample Information:
1 – many samples in a single VCF file.
Variant Types

- **Single base-pair substitution**
  - Single nucleotide polymorphisms (SNPs)

- **Multiple nucleotide substitution**
  - Substitutions where length > 1

- **Insertion or deletion, also known as ‘indel’**
  - Insertion or deletion of a DNA sequence
  - 2 to 100’s of base-pairs in length
  - For AVIA, limited to small indels < 50

- **Structural variation**
  - larger DNA sequence
  - copy number variation
  - chromosomal rearrangement events
**Fig. 1.** Example of VCF entries representing the same variant. Left panel aligns each allele to the reference genome, and the right panel represents the variant in VCF. (A) is not left-aligned (B) is neither left-aligned nor parsimonious, (C) is not parsimonious and (D) is normalized.
AVIA Indel Normalization

- All indels are normalized using U. Michigan’s VT package

<table>
<thead>
<tr>
<th>(D)</th>
<th>REF</th>
<th>GCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALT</td>
<td>G</td>
<td></td>
</tr>
</tbody>
</table>

3 GCA G

- Annotations against normalized indels
- Indel alias table
  - Maintain all aliases
Annotation and Impact Analysis

• Annotation: Identifying other associated data at a variant’s genomic location
  – Presence of gene or regulatory regions
  – Uniqueness and repeat regions
  – Presence in other samples or studies

• Impact Analysis: Assessing the impact of that change
  – gene/protein/pathway
  – Pathogenecity predictions

How do we prioritize the hundreds/thousands of variants?
Annotations

- Gene - RefSeq, (Ensembl)
- Regulatory regions - TargetScan, HMDD,
- Population databases - dbSNP, gnomAD, 1000 genomes
- Disease associated variants - COSMIC, ClinVar, TCGA
- Genomic Features - Genomicsuperdups, nonb, ENCODE
- Protein Features - Prosite_domain, dbptm
- Protein scoring algorithms - SIFT, polyphen, CADD

- 88 annotations in current version
- Regular updates through automated downloads
AVIA Full Annotations List

**AVIAv3 Annotation Database**

<table>
<thead>
<tr>
<th>Category</th>
<th>Database Name</th>
<th>Version</th>
<th>Description</th>
<th>Last Updated</th>
<th>Citation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease Related</td>
<td>CANDL</td>
<td>20161222</td>
<td>Cancer Driver Log (CanDL); Catalog of Potentially</td>
<td>21-FEB-17</td>
<td>Frederick National Laboratory for Cancer Research</td>
</tr>
</tbody>
</table>

https://avia-abcc.ncifcrf.gov
Impact Assessment

• Variant – overview, analytics
• Gene – gene.iobio
• Protein – ProtVista, MolArt
• Gene Functional clustering - DAVID
• Pathway – PathView
• Tissue – SAMM

• Literature references
• Comparisons - between and within annotations, samples
AVIA Demo Overview

• Basic Navigation
• Walk through of features
• Submit variant list
• Data Retrieval
• Registration and Additional Tools
  – Custom Annotations
  – Project Management
    • Data Sharing
    • Saving and sharing dashboards
    • Building cohorts
    • Reannotating
Demo
# Tabs Displayed by View

Below is a table of tabs displayed for each project/sample combination.

<table>
<thead>
<tr>
<th>User Project (viz*)</th>
<th>All Samples</th>
<th>Sample Selected</th>
<th>Cohorts (coh*)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVIA Summary</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Gene Summary</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>vcf.iobio</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gene.iobio (Human Only)</td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>PathView</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Oncogrid</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Co-occurrence</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>DAVID Gene Clustering</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>SAMM matrices (Human Only)</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Comparators (VENN)</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>molArt (3D structures)</td>
<td></td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>

Also available in the FAQ section under “Navigation”
Other Tools available through AVIA

- **VCF4 validator**
  - Validates VCF files using vcf-validator

- **Liftover / Converters**
  - Converts between builds of the same genome (e.g. hg38 to hg19)
  - Converts by protein positions (shows only the genomic location of the 3 codon positions)

- **Application Programming Interface**
  - Allows for programmatic submission to AVIA (bypasses web interface)
  - Requires an API key to access tied to a specific user
  - Contact us at NCI-FrederickAVIA@mail.nih.gov for more information or to register

- **Single variant annotation**
  - Allows to view annotations for a single variant
Questions?
NCI-FrederickAVIA@mail.nih.gov