Frederick National Laboratory
for Cancer Research

sponsored by the National Cancer Institute

Annotation Visualization and Impact Analysis AVIA

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

DEPARTMENT OF HEALTH AND HUMAN SERVICES • National Institutes of Health • National Cancer Institute

Frederick National Laboratory is a Federally Funded Research and Development Center operated by Leidos Biomedical Research, Inc., for the National Cancer Institute

Advanced Biomedical Computational Science

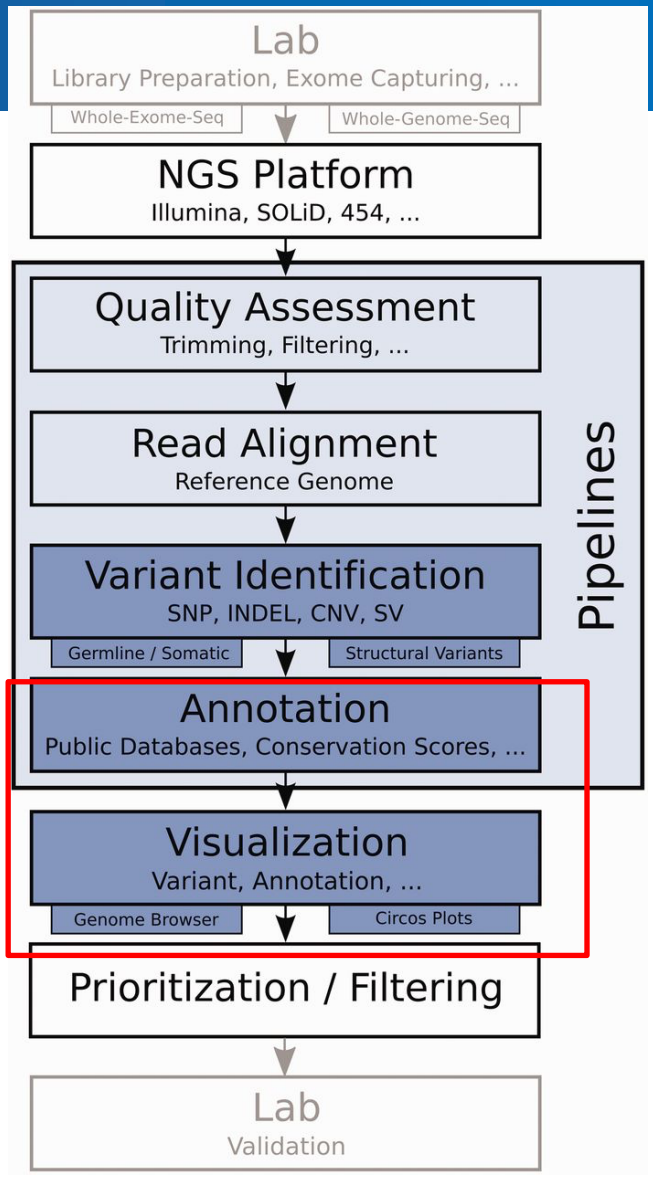
- Bioinformatics - CCBR, SF, NCBR
- Data mining, integration
- Infrastructure - biomedical databases, software
- Scientific web programming
- Imaging and Visualization
- Structural biology
- Computational chemistry
- Statistical analysis

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



AVIA

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

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

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA00001	NA00002	NA00003
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0|0:48:1:51,51	1|0:48:8:51,51	
1/1:43:5:.,.											
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017	GT:GQ:DP:HQ	0|0:49:3:58,50	0|1:3:5:65,3	0/0:41:3
20	1110696	rs6040355	A	G,T	67	PASS	NS=2;DP=10;AF=0.333,0.667;AA=T;DB	GT:GQ:DP:HQ	1|2:21:6:23,27	2|1:2:0:18,2	2/2:35:4
20	1230237	.	T	.	47	PASS	NS=3;DP=13;AA=T	GT:GQ:DP:HQ	0|0:54:7:56,60	0|0:48:4:51,51	0/0:61:2
20	1234567	microsat1	GTC	G,GTCT	50	PASS	NS=3;DP=9;AA=G	GT:GQ:DP	0/1:35:4	0/2:17:2	1/1:40:3

position and alt allele

Contains info about variant
e.g. counts, allele frequencies (AF), depth, etc

Sample Information
1 – many samples in a single VCF file

Format and Sample go hand in hand

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

Indel Representation

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

OXFORD UNIVERSITY PRESS

AVIA Indel Normalization

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

(D)	REF	GCA	3	GCA	G
	ALT	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

NIH NATIONAL CANCER INSTITUTE NCIatFrederick

AVIA Annotation, Visualization, and Impact Analysis

Annotation, Visualization, and Impact Analysis

Home | Analysis | Examples | About | Projects | Hello huetogo | Sign out

- What's new
- FAQs
- AVIA Database Sources

AVIAv3 Annotation Dat ...n(hg19)

Human GRCh38/hg38 | Mouse (mm10)

CSV | Excel

Search:

Category	Database Name	Version	Description	Last Updated	Citation
Alternative Splicing	ALT_SPLICE	NA	Ensembl Splice events	21-FEB-17	Koscielny G, Le Texier V, Gopalakrishnan C, Kumanduri V, Riethoven JJ, Nardone F, Stanley E, Fallsehr C, Hofmann O, Kull M, Harrington E, Boue S, Eyras E, Plass M, Lopez F, Ritchie W, Moucadel V, Ara T, Pospisil H, Herrmann A, G Reich J, Guigo R, Bork P, Doeberitz MK, Vilo J, Hide W, Apweiler R, Thanaraj TA, Gautheret D.
Disease Related	CANDL	20161222	Cancer Driver Log (CanDL): Catalog of Potentially	21-FEB-17	

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.

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

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