Frederick National Laboratory
for Cancer Research
sponsored by the National Cancer Institute

Annotation Visualization and Impact Analysis AVIA

Hue Vuong Reardon
Advanced Biomedical Computational Science
April 23, 2020

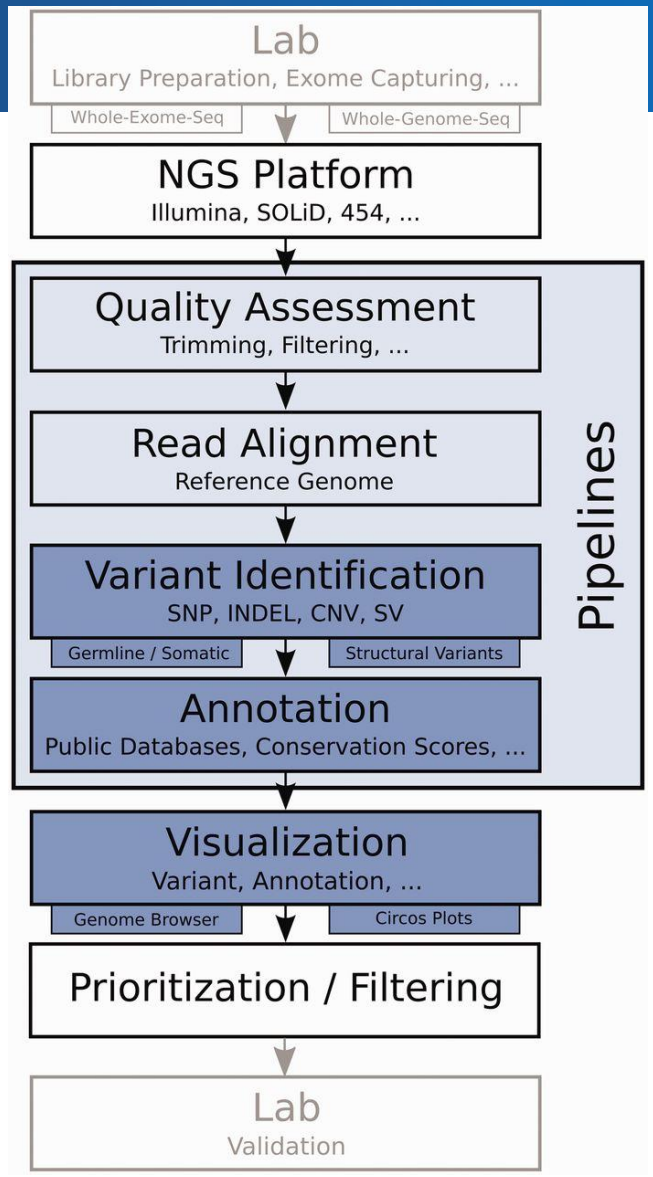
DEPARTMENT OF HEALTH AND HUMAN SERVICES • National Institutes of Health • National Cancer Institute

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

Overview

- Background
  - Sequencing and variants
  - Variant annotations
  - Impact analysis
- Demo
  - Submitting to AVIA and retrieving results
  - Single sample example
  - Multi sample example
  - Registered users
    - Project management
    - Cohort annotations

NGS Workflow



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

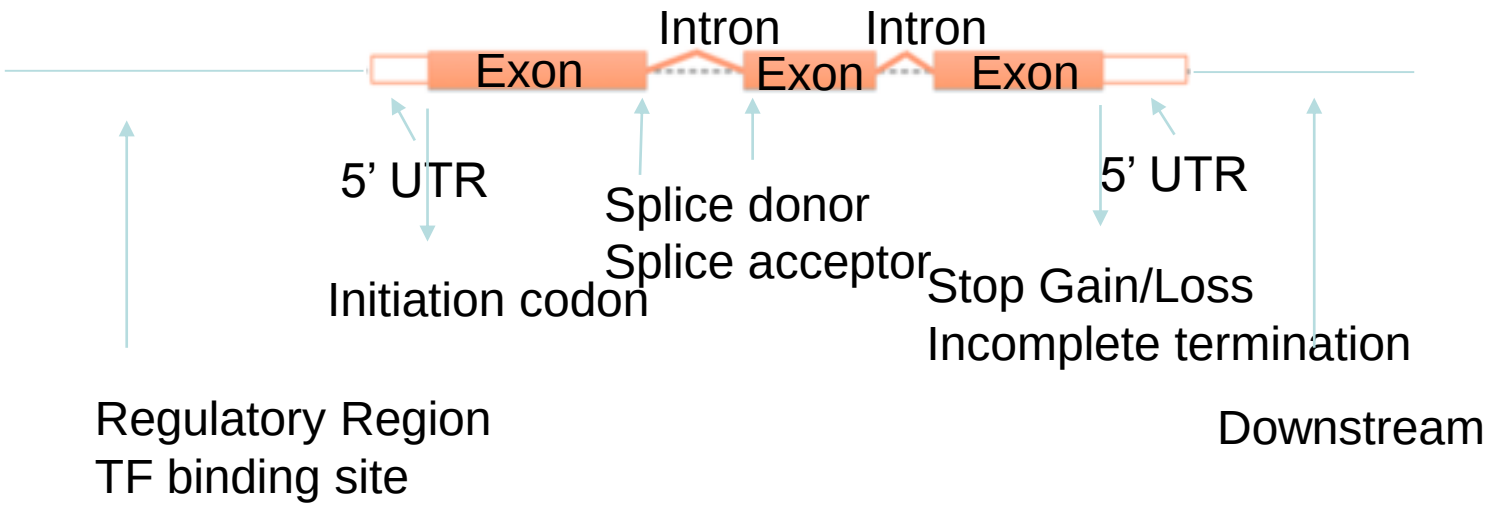
What is Annotation and Impact Analysis?

- Annotation: To process of identifying locations of genes and all of the coding regions in a genome.
  - Identifying other associated data for a variant at a given position

- Impact Analysis: To determine the **unexpected, negative** effects of a specific nucleotide change, or a set of nucleotide changes to the DNA of individuals
  - How do the changes affect individuals or populations?
  - What are the functions of the genes affected?

How do we sift through the hundreds/thousands of variants to find those of interest?

Variant location - Gene



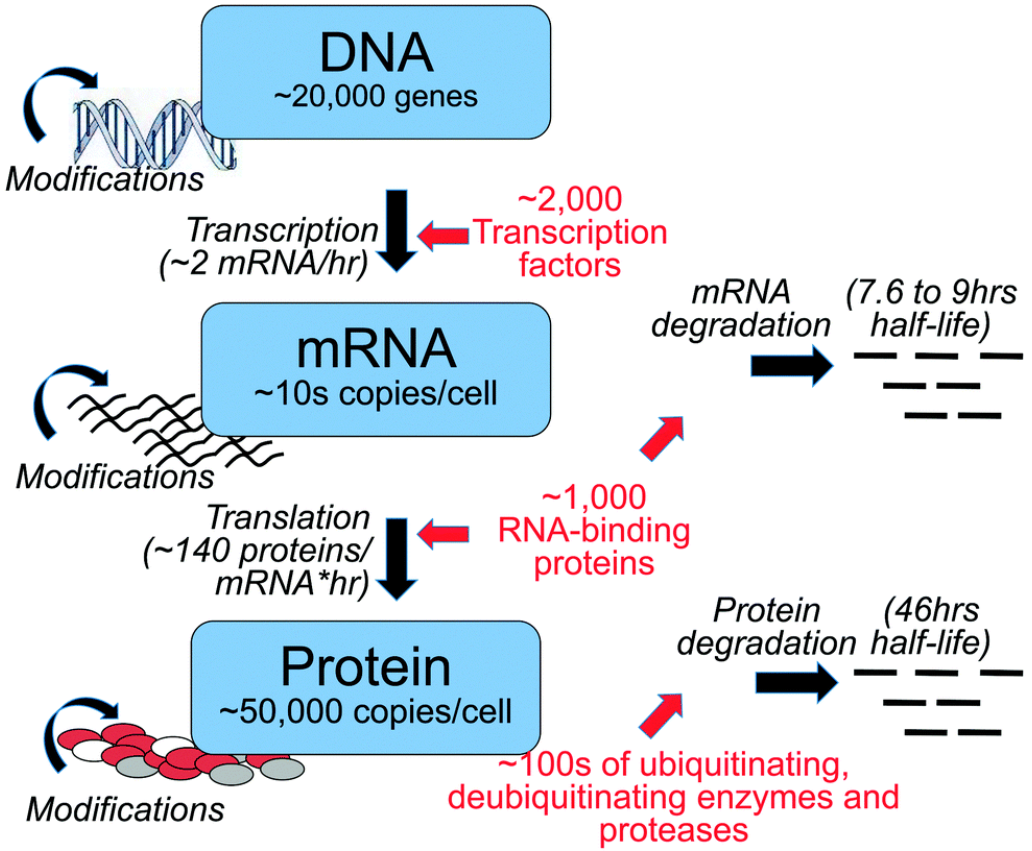
Can also be intergenic (between genes) or in a non-coding gene (ncRNA)

Gene Regulation

Methylation
Promoters
Tf Binding Sites

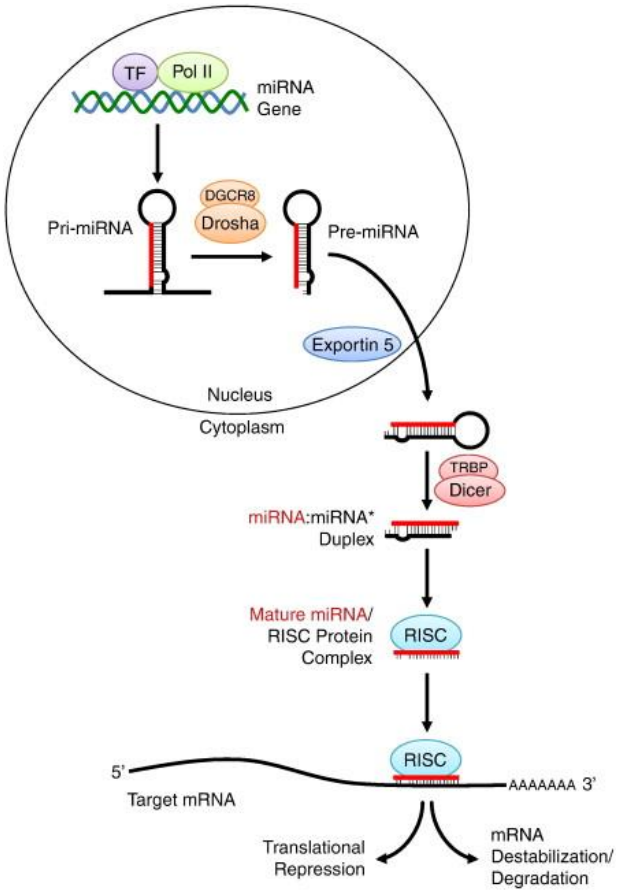
Splicing
miRNA

PTM
Protein Binding
Binding Sites
Domains, etc



http://pubs.rsc.org/services/images/RSCpubs.ePlatform.Service.FreeContent.ImageService.svc/ImageService/Articleimage/2015/MB/c5mb00310e/c5mb00310e-f1_hi-res.gif

miRNA Annotation Databases in AVIA



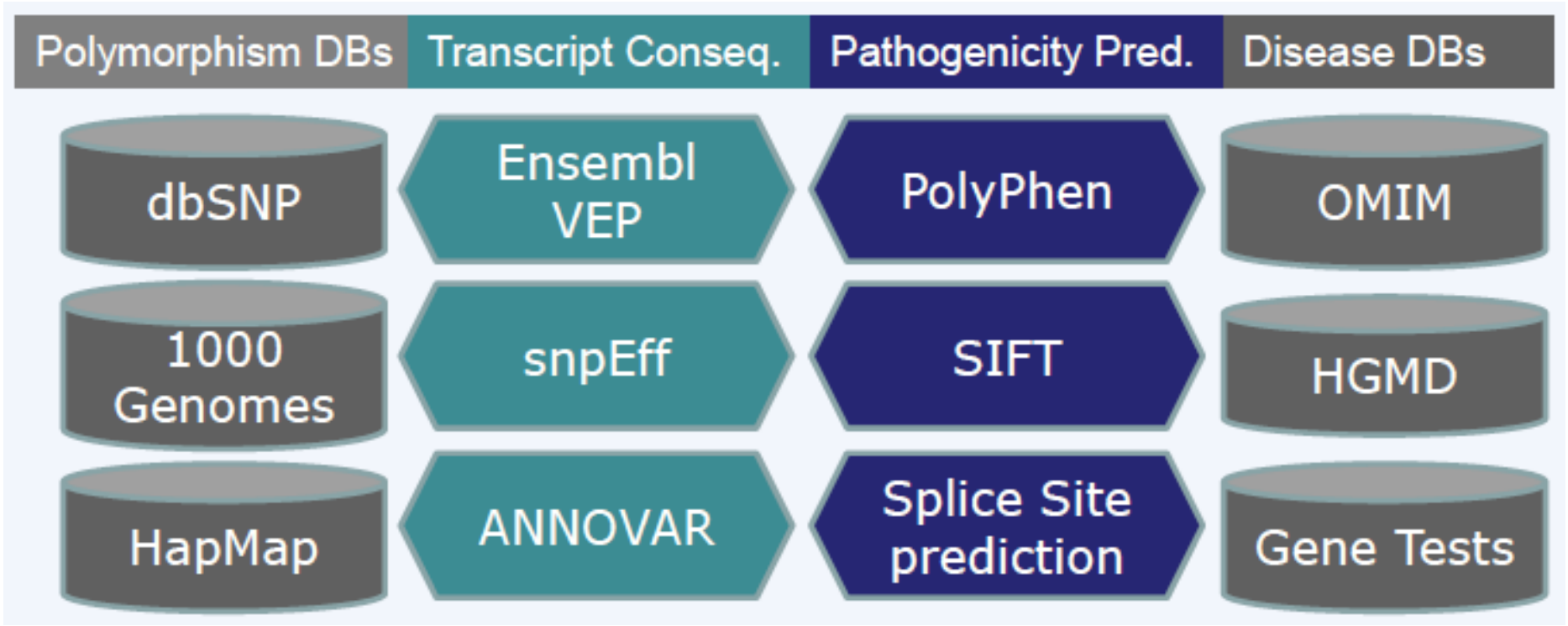
HMDD
snoRNA
miRNA

TargetScan
SomamiR
microPIR

Source : Curtan, A and Phillip Sharp. The Role of miRNAs in Regulating Gene Expression Networks. *J. of Mol. Biol. (2013) 425(19):3582-3600.*

Annotation and Functional Prediction

- Now let's take a quick look at some ways of predicting and visualising the effect of variation on protein structure and function.



Annotation and Impact Analysis

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

- Impact Analysis: Assessing the impact of that change
  - gene/protein/pathway
  - Pathogenicity predictions

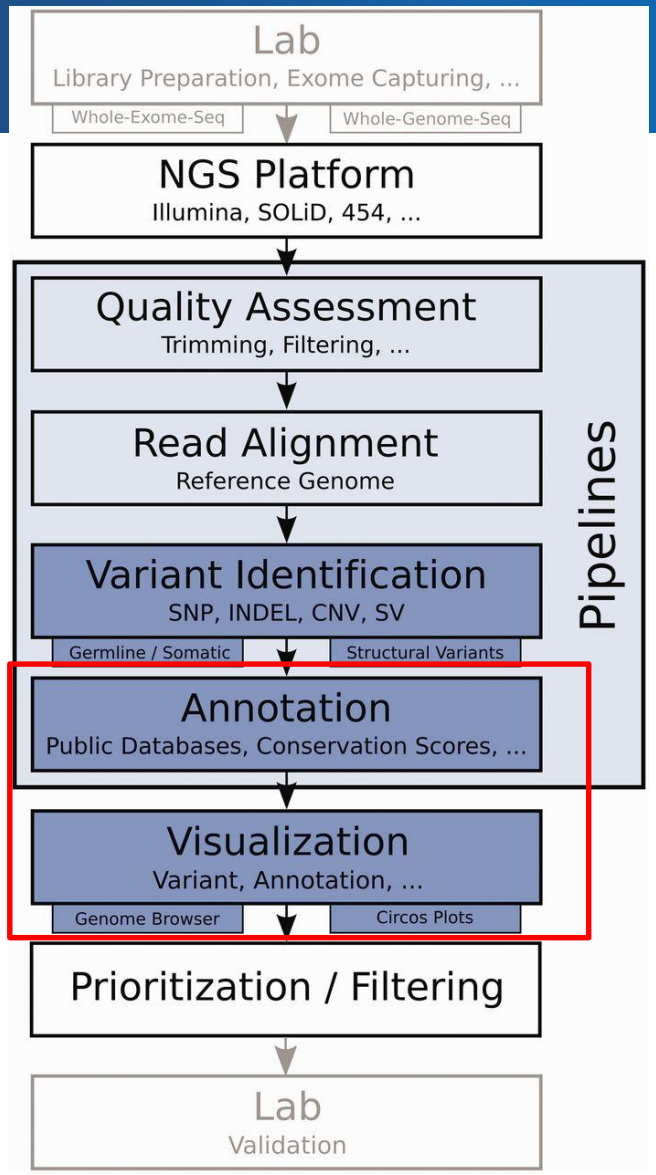
How do we prioritize the hundreds/thousands of variants?

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

NGS Workflow

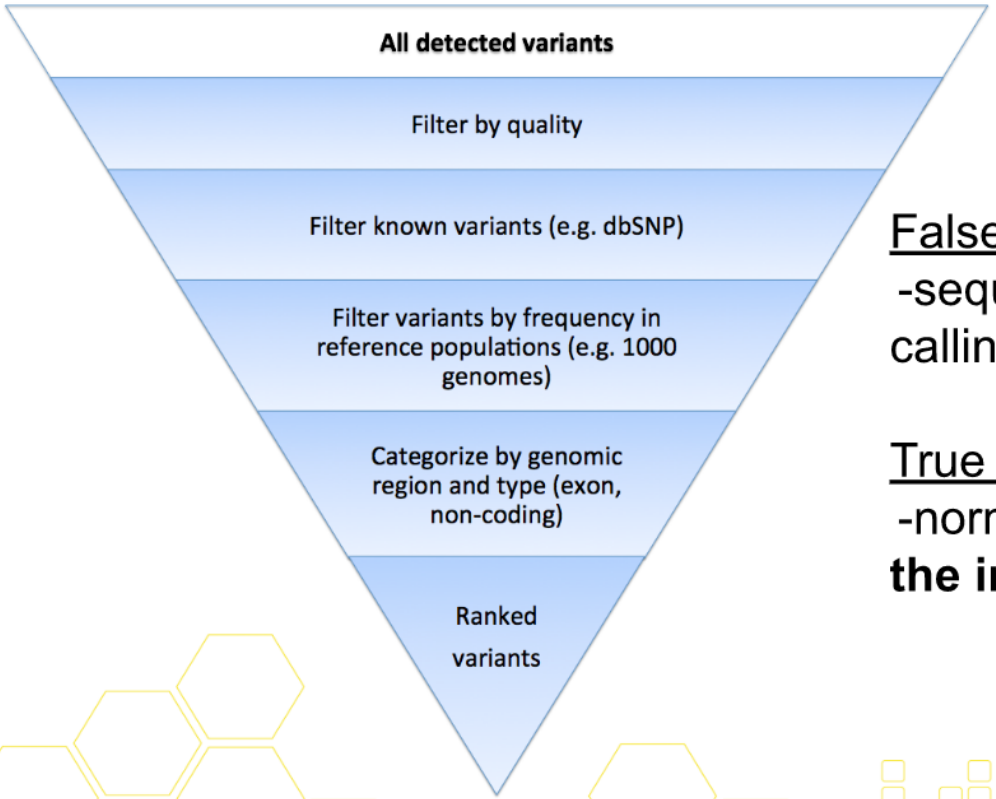


AVIA

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

Standard Variant Annotation Workflow

All variants detected in sample(s)



Variants of interest

False positive variants
-sequencer, alignment, variant calling

True variants
-normal variation, benign variants, **the interesting ones**

AVIA Allele Frequency Designations

- We adopted these terms from Marth's Lab gene.iobio. The bins are as follows and annotations display the group with the highest allele frequency:
  - common - AF >5%
  - uncommon - AF 1-5%
  - rare - AF = <1%
  - superrare - <0.1%
  - uberrare AF <0.01%

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

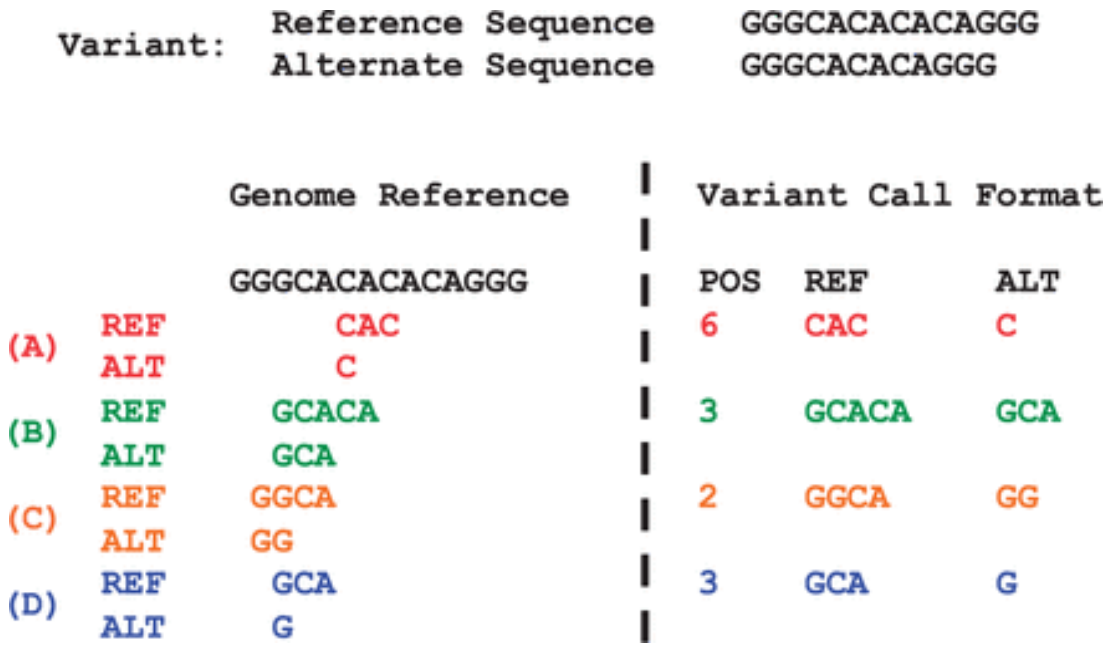


Fig. 1. Example of VCF entries representing the same variant. Left panel aligns each allele to the reference genome, and the right panel represents the variant in VCF. (**A**) is not left-aligned (**B**) is neither left-aligned nor parsimonious, (**C**) is not parsimonious and (**D**) is normalized

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

6	CAC	C
3	GCACA	GCA
2	GGCA	GG

3	GCA	G

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
- Submit variant list to AVIA
- Data Retrieval
- Walk through of visualization and data
- Advanced Features
- Registration and Additional Tools
  - Custom Annotations
  - Project Management
    - Data Sharing
    - Saving and sharing dashboards
    - Building cohorts
    - Reannotating

Basic Navigation

Navigation

NIH NATIONAL CANCER INSTITUTE NCIatFrederick

AVIA Annotation, Visualization, and Impact Analysis

Annotation, Visualization, and Impact Analysis

Home Analysis Examples About Sign in

AVIA v3.0 Features

AVIAv3 Annotation Databases for Human

CSV Excel

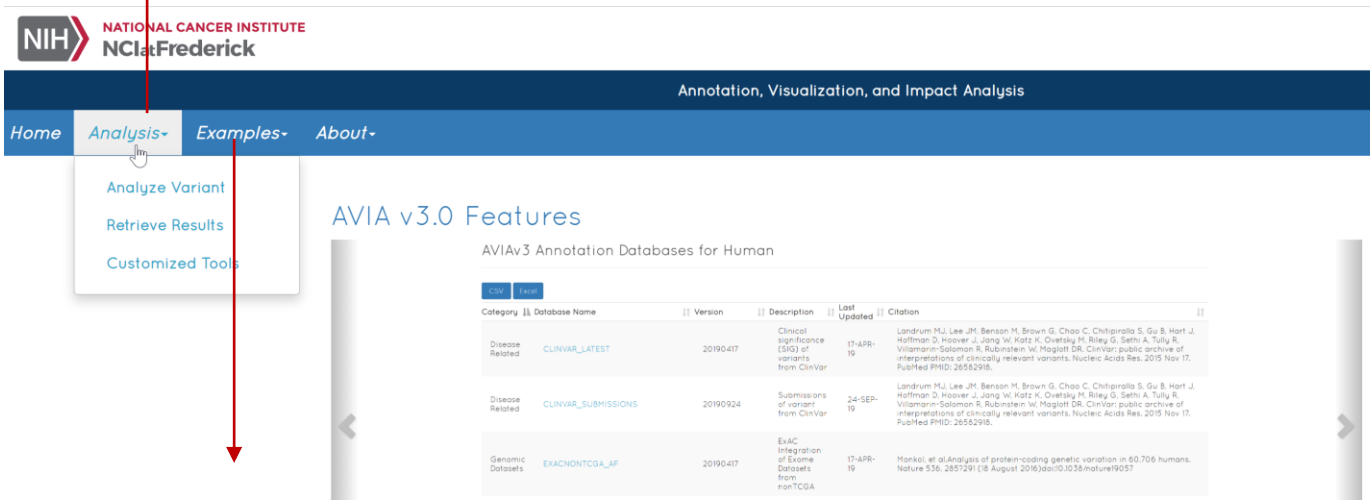
Category	Database Name	Version	Description	Last Updated	Citation
Disease Related	CLINVAR_LATEST	20190417	Clinical significance (SIG) of variants from ClinVar	17-APR-19	Landrum M.J, Lee J.M, Benson M, Brown G, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Hoover J, Jang W, Katz K, Ovetsky M, Riley G, Sethi A, Tully R, Villamarin-Salomon R, Rubinstein W, Maglott D.R. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. 2015 Nov 17. PubMed PMID: 26582918.
Disease Related	CLINVAR_SUBMISSIONS	20190924	Submissions of variant from ClinVar	24-SEP-19	Landrum M.J, Lee J.M, Benson M, Brown G, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Hoover J, Jang W, Katz K, Ovetsky M, Riley G, Sethi A, Tully R, Villamarin-Salomon R, Rubinstein W, Maglott D.R. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. 2015 Nov 17. PubMed PMID: 26582918.
Genomic Datasets	EXACNONTCGA_AF	20190417	ExAC Integration of Exome Datasets from nonTCGA	17-APR-19	Monkol, et al Analysis of protein-coding genetic variation in 60,706 humans. Nature 536, 285?291 (18 August 2016)doi:10.1038/nature19057
Genomic	EXACNONTCGA_THS	20190417	ExAC Integration of Exome	17-APR-	Monkol, et al Analysis of protein-coding genetic variation in 60,706 humans.

Over 70 databases updated automatically

URL: https://avia-abcc.ncifcrf.gov

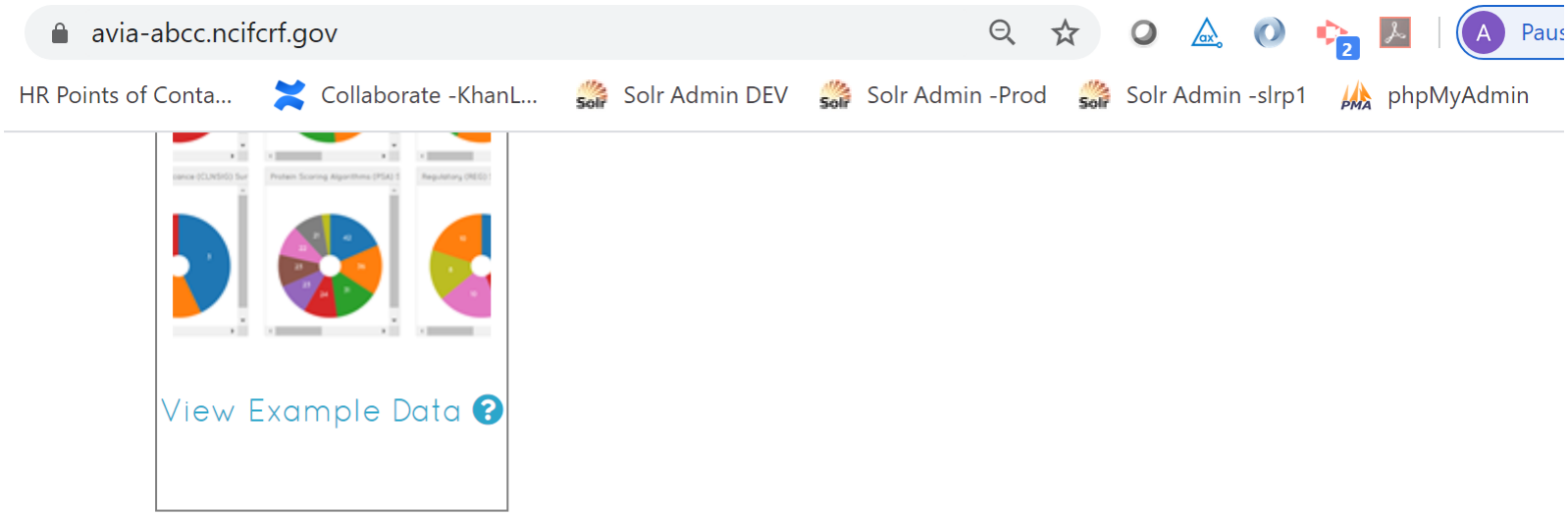
Navigation

Use any tool or retrieve results



View a sample analysis

Contact Us



Email: NCIFrederickAVIA@mail.nih.gov

Web form to ask questions or request services

Navigation

Information about AVIA

Accepted Input Formats

Understanding VCF format

- **VCF – variant call format**
- **standard file format** for storing variation data
- used by large scale variant mapping projects
- the standard output of variant calling software
- can be compressed and indexed

VCF is a preferred format because it is **unambiguous, scalable and flexible**, allowing extra information to be added to the info field. Many millions of variants can be stored in a single VCF file.

Header in VCF files

```
##fileformat=VCFv4.3
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#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
```

https://en.wikipedia.org/wiki/Variant_Call_Format

VCF (ctd)

VCF files are tab-delimited text files

#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

Sample Columns in VCF file

```
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL  FILTER   INFO
```

FORMAT	NA00001	NA00002	NA00003
GT:GQ:DP:HQ	0|0:48:1:51,51	1|0:48:8:51,51	
GT:GQ:DP:HQ	0|0:49:3:58,50	0|1:3:5:65,3	0/0:41:3
GT:GQ:DP:HQ	1|2:21:6:23,27	2|1:2:0:18,2	2/2:35:4
GT:GQ:DP:HQ	0|0:54:7:56,60	0|0:48:4:51,51	0/0:61:2
GT:GQ:DP	0/1:35:4	0/2:17:2	1/1:40:3

/ : genotype unphased

| : genotype phased

BED – like (ANNOVAR format)

The first five space or tab delimited fields are Chromosome ("chr" prefix is optional), Start, End, Reference Allelel, Alternative Allele. The rest of the columns are completely optional.

```
-> cat ex1.avinput
1       948921  948921  T       C       comments: rs15842, a SNP in 5' UTR of ISG15
1       1404001 1404001 G       T       comments: rs149123833, a SNP in 3' UTR of ATAD3C
1       5935162 5935162 A       T       comments: rs1287637, a splice site variant in NPHP4
1       162736463       162736463       C       T       comments: rs1000050, a SNP in Illumina SNP arrays
1       84875173        84875173        C       T       comments: rs6576700 or SNP_A-1780419, a SNP in Affymetrix SNP arrays
1       13211293        13211294        TC      -       comments: rs59770105, a 2-bp deletion
1       11403596        11403596        -       AT      comments: rs35561142, a 2-bp insertion
1       105492231       105492231       A       ATAAA   comments: rs10552169, a block substitution
1       67705958        67705958        G       A       comments: rs11209026 (R381Q), a SNP in IL23R associated with Crohn's disease
2       234183368       234183368       A       G       comments: rs2241880 (T300A), a SNP in the ATG16L1 associated with Crohn's disease
16      50745926        50745926        C       T       comments: rs2066844 (R702W), a non-synonymous SNP in NOD2
16      50756540        50756540        G       C       comments: rs2066845 (G908R), a non-synonymous SNP in NOD2
16      50763778        50763778        -       C       comments: rs2066847 (c.3016_3017insC), a frameshift SNP in NOD2
13      20763686        20763686        G       -       comments: rs1801002 (del35G), a frameshift mutation in GJB2, associated with hearing loss
13      20797176        21105944        0       -       comments: a 342kb deletion encompassing GJB6, associated with hearing loss
8       8887543 8887543 A       T       comments: a mutation that abolishes stop codon
8       8887539 8887539 A       T       comments: a mutation that results in premature stop codon
8       8887536 8887537 AG      GATT    comments: a mutation that creates a stop codon 2 amino acids downstream
8       8887540 8887540 G       GGAA    comments: a mutation that results in insertion of a new amino acid
5       1295288 1295288 G       A       comments: a variant upstream of transcriptional start site
chr14   95602958        95602958        A       C       comments: a variant that affects splicing of UTR regions
->
```

Indels cannot be normalized using this input format. VCF format is the preferred method.

https://doc-openbio.readthedocs.io/projects/annovar/en/latest/user-guide/input/

BED file discrepancies

```
1       105492231       105492231       A       ATAAA   comments: rs10552169, a block substitution
```

If this was a VCF file, ANNOVAR would have converted this to:

1 105492231 105492231 - TAAA

Refer to ANNOVAR about normalization and variant highjacking:
https://doc-openbio.readthedocs.io/projects/annovar/en/latest/articles/VCF/

Sample Mining for Single Sample (ss VCF, BED, or dbSNP ids)

- AVIA does not mine BED files for sample information
- When no sample information is presented, a dummy sample id is used instead

Submit to AVIA

Input to AVIA

Accepted Input Types

- VCF (single and multisample)
- Bed-like (Chr – start – stop – ref – alt)
- dbSNP identifiers

AVIA Annotation and Visualization Request

Section I. Describe your data

A field with an asterisk (*) before it is a required field.

Name your submission (optional) — Enter name (optional)

Input Format* — --Select a format--

Organism*

E-mail address*

Confirm e-mail address*

--Select a format--
VCF4 formatted file
ANNOVAR format input (BED)
dbSNP identifiers

Section II. Upload your data (Required)

Enter your data here using a space delimited format (one variant per line)

-- OR --

Select a file

Examples -- Select an Input --

Notification of Results

AVIA
Annotation, Visualization, and Impact Analysis

Your results for vizssbed can be found at AVIA. You can get your results one of two ways: 1) If you are not a registered user, click on the 'Retrieve Results' under the 'Analysis' menu on the link above. Then enter your AVIA id vizssbed and email address. 2) If you are a registered user, please log into your AVIA Account by selecting "Sign in" and follow directions for CILogon. After authentication, select your AVIA id from your personal dashboard. Thank you for using AVIA. If the link does not work, please use https://avia-abcc.ncifcrf.gov.

1)

avia-abcc.ncifcrf.gov

NIH NATIONAL CANCER INSTITUTE NCIatFrederick

AVIA Annotation, Visualization, and Impact Analysis

Annotation, Visualization, and Impact Analysis

2)

Analysis Examples About Sign in

AVIA v3.0 Features

AVIAv3 Annotation Databases for Human

3)

Analysis Examples

Analyze Variant

Retrieve Results

Customized Tools

Retrieve AVIA Results

Please enter the AVIA Identifier provided when you submitted your request and the email address. You will be required to log in to see your results if you are already registered user. If you are not registered, you can still see your results.

5)

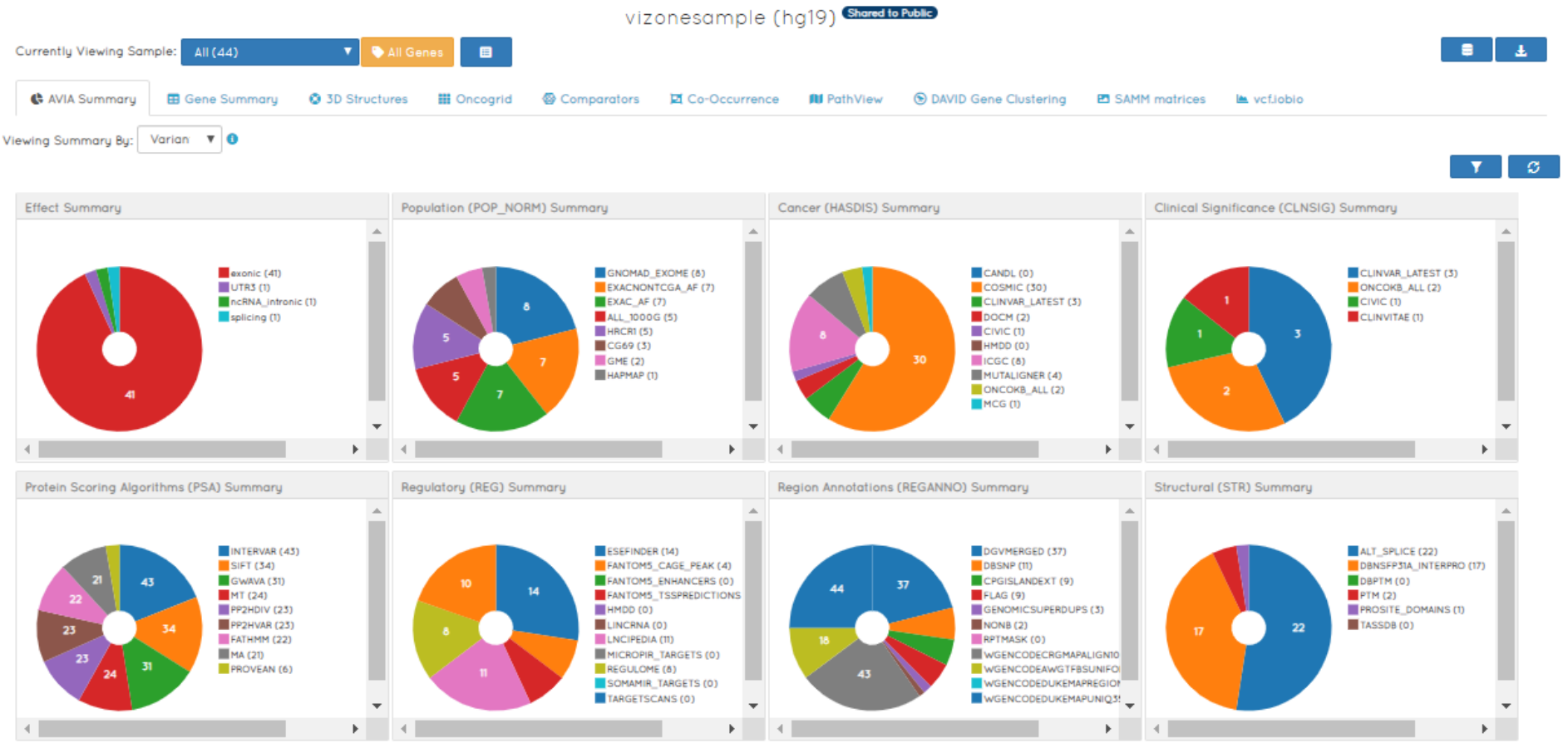
AVIA ID: vizssbed

User Name/Email: huetogo@hotmail.com

6)

Submit Reset

Results Navigation Landing Page



Demo

Questions?

Contact Us:

NCI-FrederickAVIA@mail.nih.gov

Or https://avia-abcc.ncifcrf.gov