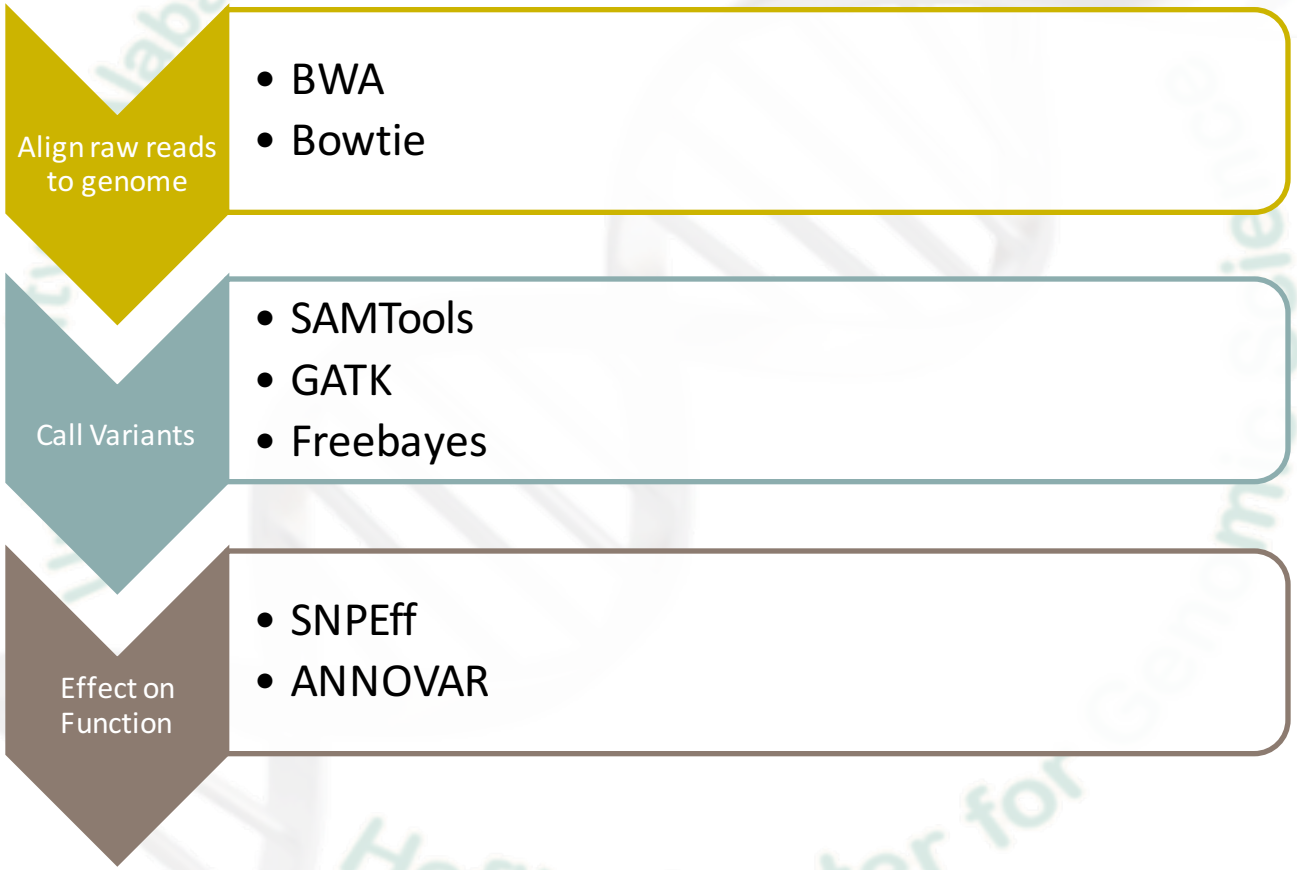
Variant Analysis of Exome/Genome Sequencing Data

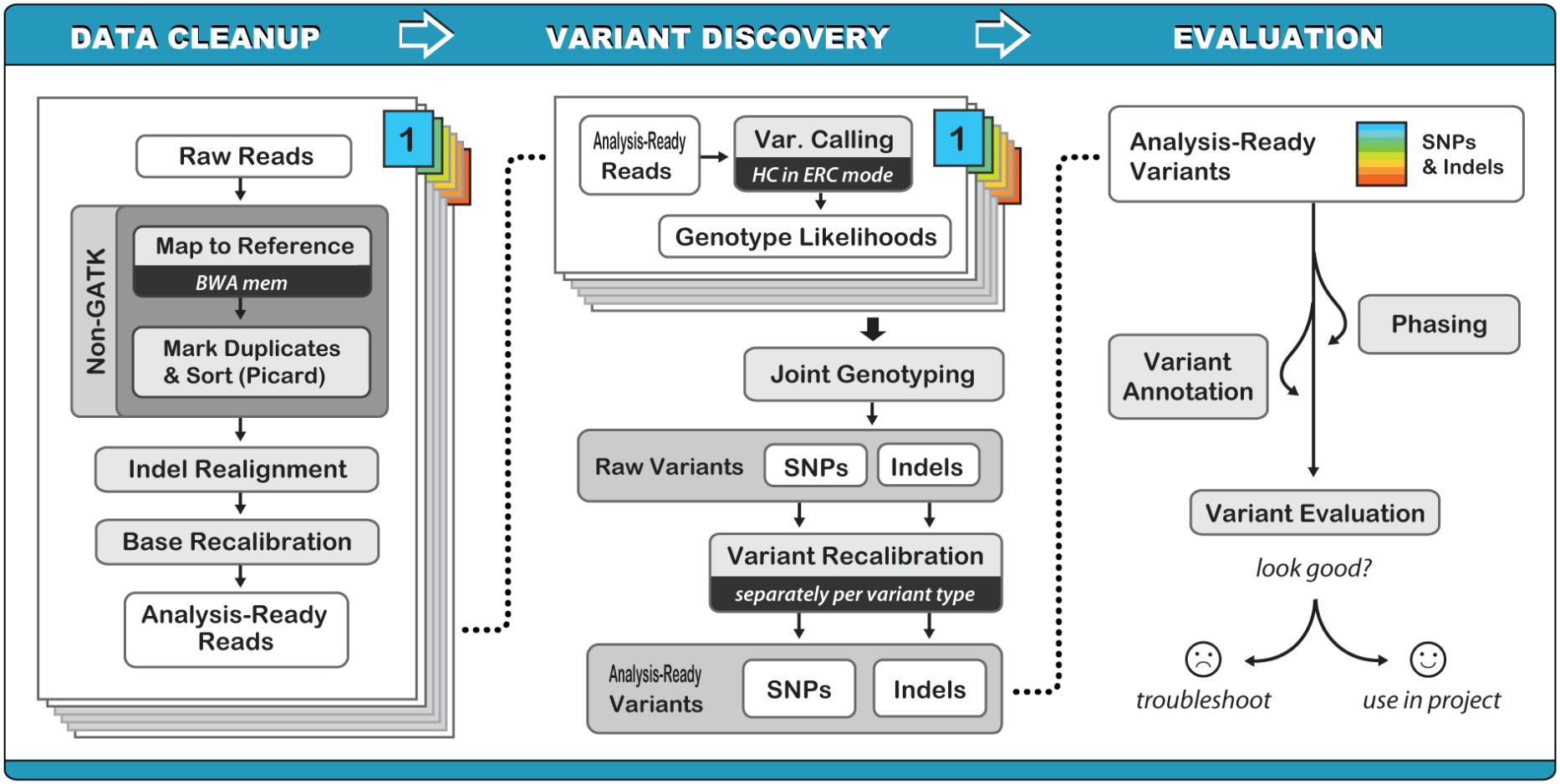
David Crossman, Ph.D.
UAB Heflin Center for Genomic Science

Immersion Course

Whole Genome/Exome (DNA-Seq) analysis pipeline



GATK pipeline



GATK Best Practices

(http://www.broadinstitute.org/gatk/)

Best Practice Variant Detection with the GATK v4, for release 2.0

There are 18 comments on this article. To see them or add your own, read this post on the forum →

Introduction

1. The basic workflow

Our current best practice for making SNP and indel calls is divided into four sequential steps: initial mapping, refinement of the initial reads, multi-sample indel and SNP calling, and finally variant quality score recalibration. These steps are the same for targeted resequencing, whole exomes, deep whole genomes, and low-pass whole genomes. Example commands for each tool are available on the individual tool's wiki entry. There is also a list of which resource files to use with which tool.

Note that due to the specific attributes of a project the specific values used in each of the commands may need to be selected/modified by the analyst. Care should be taken by the analyst running our tools to understand what each parameter does and to evaluate which value best fits the data and project design.

2. Lane, Library, Sample, Cohort

There are four major organizational units for next-generation DNA sequencing processes that used throughout this documentation:

- Lane: The basic machine unit for sequencing. The lane reflects the basic independent run of an NGS machine. For Illumina machines, this is the physical sequencing lane.
- Library: A unit of DNA preparation that at some point is physically pooled together. Multiple lanes can be run from aliquots from the same library. The DNA library and its preparation is the natural unit that is being sequenced. For example, if the library has limited complexity, then many sequences are duplicated and will result in a high duplication rate across lanes.
- Sample: A single individual, such as human CEPH NA12878. Multiple libraries with different properties can be constructed from the original sample DNA source. Here we treat samples as independent individuals whose genome sequence we are attempting to determine. From this perspective, tumor / normal samples are different despite coming from the same individual.
- Cohort: A collection of samples being analyzed together. This organizational unit is the most subjective and depends intimately on the design goals of the sequencing project. For population discovery projects like the 1000 Genomes, the analysis cohort is the ~100 individual in each population. For exome projects with many samples (e.g., ESP with 800 EOMI samples) deeply sequenced we divide up the complete set of samples into cohorts of ~50 individuals for multi-sample analyses.

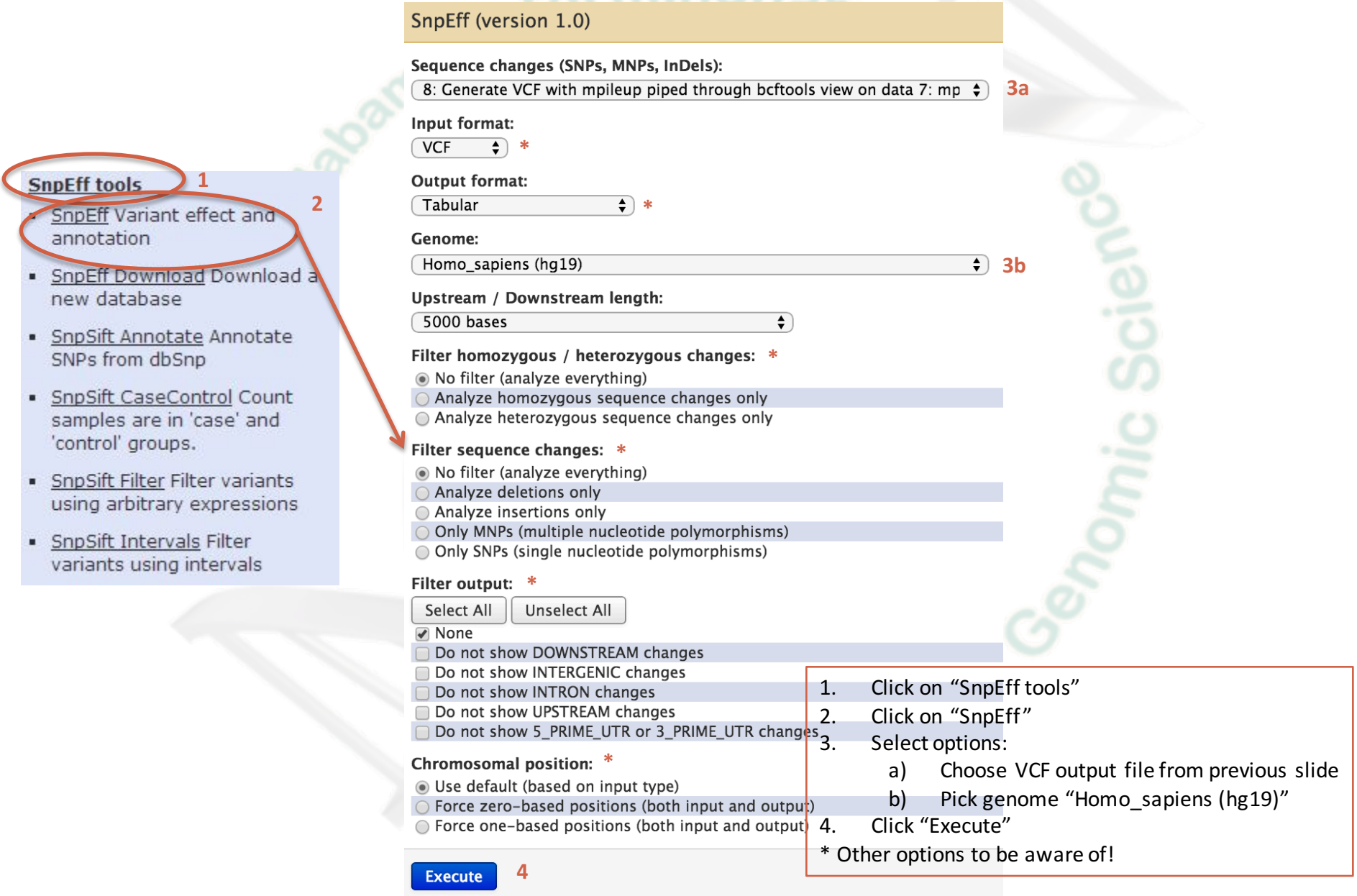
This document describes how to call variation within a single analysis cohort, comprised for one or many samples, each of one or many libraries that were sequenced on at least one lane of an NGS machine.

Note that many GATK commands can be run at the lane level, but will give better results seeing all of the data for a single sample, or even all of

VCF file

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	Gen011
chr1	3311261	rs6702992	C	T	663.71	PASS	AC=2;AF=1.00;AN=2;DB;DP=3;FS=0.000;MQ=60.00;MQ0=0;POSITIVE_TRAIN_SITE;QD=31.56;VQSLOD=4.22;culprit=FS;SF=0	GT:AD:DP:GQ:PL	1/1:0,3:3:9:109,9,0
chr1	5937168	rs3747992	G	A	1245.09	PASS	AC=1;AF=0.500;AN=2;BaseQRankSum=1.23;ClippingRankSum=-7.620e-01;DB;DP=5;FS=0.000;MQ=60.00;MQ0=0;MQRankSum=1.18;POSITIVE_TRAIN_SITE;QD=19.16;ReadPosRankSum=0.067;VQSLOD=4.25;culprit=FS;SF=0	GT:AD:DP:GQ:PL	0/1:3,2:5:51:58,0,51

SNPEff (effect on function of variants)



1. Click on "SnpEff tools"
2. Click on "SnpEff"
3. Select options:
   a) Choose VCF output file from previous slide
   b) Pick genome "Homo_sapiens (hg19)"
4. Click "Execute"

* Other options to be aware of!

SNPEff output (cropped)

Chr	Pos	REF	ALT	Change_type	Homozygous	Quality	Cov.	Gene name	Trancript_ID	Effect	old_AA/new_AA	Old_codon/New_codon	Codon_Num(CDS)	Codon_Degeneracy	CDS_size
21	10910311	T	G	SNP	Hom	99	737	TPTE	NM_199259	NON_SYNONYMOUS_CODING	Y/S	tAt/tCt	464	1	1602
21	10970008	C	T	SNP	Hom	4.13	286	TPTE	NM_199259	SPLICE_SITE_DONOR					1602
21	11058226	G	C	SNP	Hom	21	882	BAGE2	NM_182482	NON_SYNONYMOUS_CODING	P/A	Cct/Gct	72	1	330
21	15481365	G	T	SNP	Hom	99	171	LIPI	NM_198996	NON_SYNONYMOUS_CODING	D/E	gaC/gaA	465	2	1446
21	15596772	T	G	SNP	Hom	99	63	RBM11	NM_144770	NON_SYNONYMOUS_CODING	L/V	Ttg/Gtg	116	2	846
21	15954528	G	A	SNP	Hom	99	405	SAMSN1	NM_001256370	NON_SYNONYMOUS_CODING	H/Y	Cac/Tac	64	1	1326
21	30339120	C	A	SNP	Hom	99	111	LTN1	NM_015565	NON_SYNONYMOUS_CODING	G/C	Ggc/Tgc	611	1	5439
21	31744127	A	T	SNP	Hom	99	110	KRTAP13-2	NM_181621	STOP_GAINED	C/*	tgT/tgA	135	2	528
21	34948686	*	+A	INS	Hom	99	74	SON	NM_138927	FRAME_SHIFT	-/?	-/A	2413		7281

Tool	Link
CADD	http://cadd.gs.washington.edu/
Broad ExAC	http://exac.broadinstitute.org/
SeattleSeq	http://snp.gs.washington.edu/SeattleSeqAnnotation138/
Ensembl VEP	http://useast.ensembl.org/info/docs/tools/vep/index.html
MutationAssessor	http://mutationassessor.org/
MutationTaster	http://www.mutationtaster.org/
OMIM	http://www.ncbi.nlm.nih.gov/omim
ClinVar	http://www.ncbi.nlm.nih.gov/clinvar/
GeneCards	http://www.genecards.org/
Wellcome Trust Sanger Institute Mouse Genomes Project	http://www.sanger.ac.uk/resources/mouse/genomes/

Demo

CADD

CADD

- **C**ombined **A**nnotation **D**ependent **D**epletion (**CADD**)
- A tool that scores deleteriousness of SNVs as well as INDELs in the human genome
- Integrates multiple annotations into one metric by contrasting variants that survived natural selection with simulated mutations.
- How to interpret scaled CADD scores:
  - 10^{th}-% of CADD scores are assigned to CADD-10
  - Top 1% to CADD-20
  - Top 0.1% to CADD-30

Prepare SNPEff output files for downstream analysis (i.e. CADD)

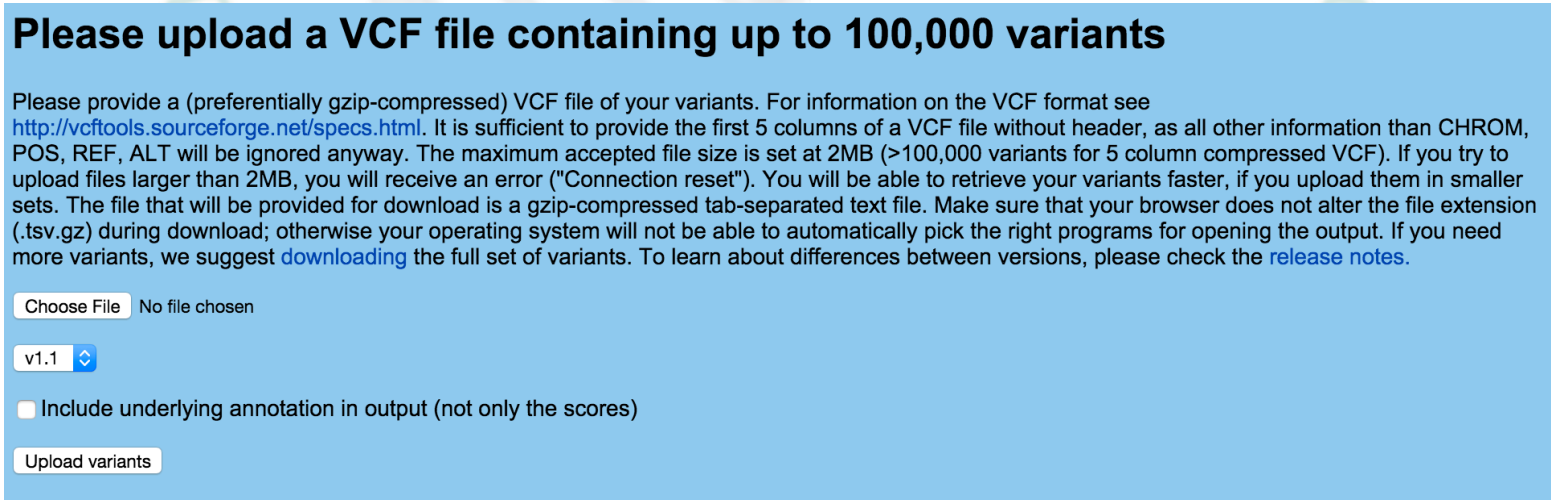
1. Download SNPEff output file
2. Open in Excel
3. Keep first four (4) columns; remove all header columns
4. Create a 3rd column called ID and place a "." for each variant. Final output in this format:

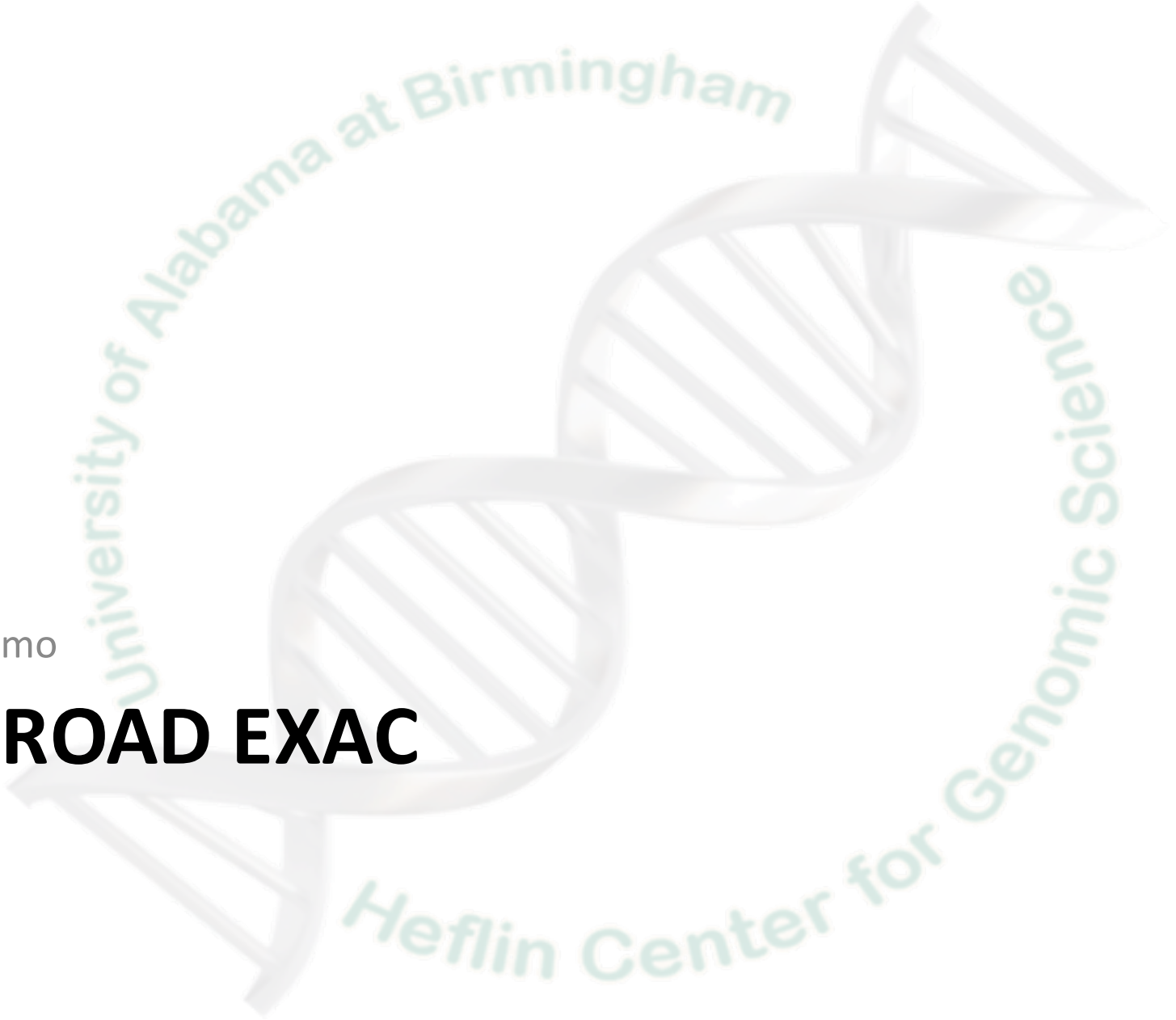
# CHROM	POS	ID	REF	ALT
21	1097008	.	C	T

1. Save as a Windows or Unix text file
2. Manually change .txt to .vcf

How to use CADD (on a small set of SNVs)

- Go to: http://cadd.gs.washington.edu/score
- Upload a gzip-compressed vcf file of variants
- Click on “Upload variants”
- Once analysis completes, download the tab-delimited file and open in Excel.





Demo

BROAD EXAC

Broad's ExAC Browser

- **Ex**ome **A**ggregation **C**onsortium (**ExAC**)
- Consists of 60,706 unrelated individuals that were exome sequenced
- Various disease-specific and population genetic studies
- Search for any human gene or variant or region

How to use ExAC

- Go to: http://exac.broadinstitute.org/
- Input either gene or variant or region

Gene: PRDM1

PRDM1 PR domain containing 1, with ZNF domain
Number of variants 509 (Including filtered: 555)
UCSC Browser 6:106534195-106557814
GeneCards PRDM1
OMIM PRDM1
Other External References

Transcripts

Gene summary

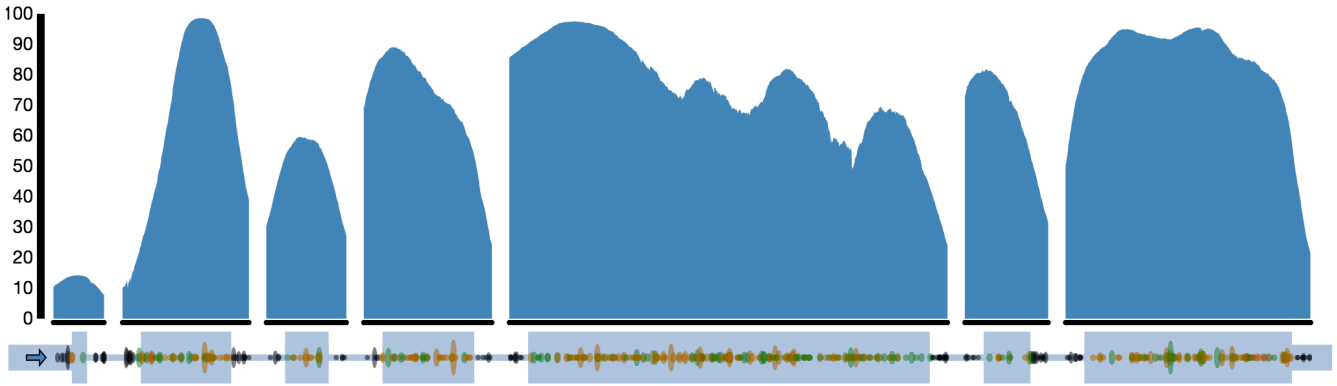
(Coverage shown for canonical transcript: ENST00000369096)

Mean coverage 70.23

Display: Overview | Detail **Include UTRs in plot**

Coverage metric: Average | Individuals over X

Metric: mean



Save coverage plot | Save exon image

All | Missense + LoF | LoF **Include filtered (non-PASS) variants**

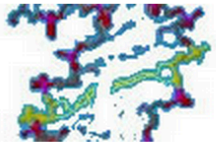
Invert (highlight rare variants)

Export table to CSV

Variant	Chrom	Position	Protein Consequence	Filter	Annotation	Allele Count	Allele Number	Number of Homozygotes	Allele Frequency
6:106534391 A / G	6	106534391		PASS	5' UTR	1	23570	0	0.00004243
6:106534402 G / C	6	106534402		PASS	5' UTR	1	23606	0	0.00004236
6:106534408 G / A	6	106534408		PASS	5' UTR	1	23640	0	0.00004230
6:106534419 C / G	6	106534419		PASS	5' UTR	162	23696	1	0.006837
6:106534419 CT / C	6	106534419		PASS	5' UTR	3	23696	0	0.0001266
6:106534430 T / C	6	106534430	p.Met1?	PASS	initiator codon	1	23752	0	0.00004210
6:106534431 G / T	6	106534431	p.Met1?	PASS	initiator codon	1	23756	0	0.00004209

dbSNP

dbSNP
Short Genetic Variations



ar GaP PubMed Nucleotide Protein
ll variations in dbSNP or large structural variations in dbVar
for Go

Reference SNP (refSNP) Cluster Report: rs2822432

RefSNP	
Organism:	human (Homo sapiens)
Molecule Type:	Genomic
Created/Updated in build:	100/142
Map to Genome Build:	106/Weight
Validation Status:	

Allele	
Variation Class:	SNV: single nucleotide variation
RefSNP Alleles:	C/T (FWD)
Allele Origin:	
Ancestral Allele:	C
Variation Viewer:	unknown
Clinical Significance:	NA
MAF/MinorAlleleCount:	T=0.3173/1589
MAF Source:	1000 Genomes

HGVS Names
NC_000021.8:g.15516948C>T
NC_000021.9:g.14144627C>T
NG_021434.1:g.67307G>A
NM_198996.2:c.1354G>A
NP_945347.1:p.Glu452Lys
XM_005260920.1:c.1378G>A
XM_005260920.2:c.1378G>A
XM_005260921.1:c.1360G>A
XM_005260921.2:c.1360G>A
XM_005260922.1:c.1288G>A
XM_005260922.2:c.1288G>A

...

SNP Details are organized in the following sections:
GeneView | Map | Submission | Fasta | Resource | Diversity | Validation

Integrated Maps (Hint: click on 'Chr Pos' to see variant in the new NCBI variation viewer)

Assembly	Annotation Release	Chr	Chr Pos	Contig	Contig Pos	SNP to Chr	Contig allele	Contig to Chr	Neighbor SNP
GRCh38	106	21	14144627	NT_011512.12	1178819	Fwd	C	Fwd	view
GRCh37.p13	105	21	15516948	NT_011512.11	1178819	Fwd	C	Fwd	view

GeneView

OMIM

*609252

LIPASE I; LIPI

Alternative titles; symbols
LPD LIPASE; LPDL
PRED5

HGNC Approved Gene Symbol: LIPI

Cytogenetic location: 21q11.2 ***Genomic coordinates (GRCh37): 21:15,480,783-15,583,360*** (from NCBI)

Gene-Phenotype Relationships

Location	Phenotype	Phenotype MIM number	Phenotype mapping key
21q11.2	{Hypertriglyceridemia, susceptibility to}	145750	3

TEXT

Cloning and Expression

By screening a testis cDNA library using mouse Lpdl as probe, Wen et al. (2003) cloned human LIPI, which they called LPDL. The deduced 460-amino acid protein contains a hydrophobic leader sequence with a putative cleavage site after amino acid 15, a central lipase consensus sequence (GxSxG), a central 12-amino acid lipase lid sequence,

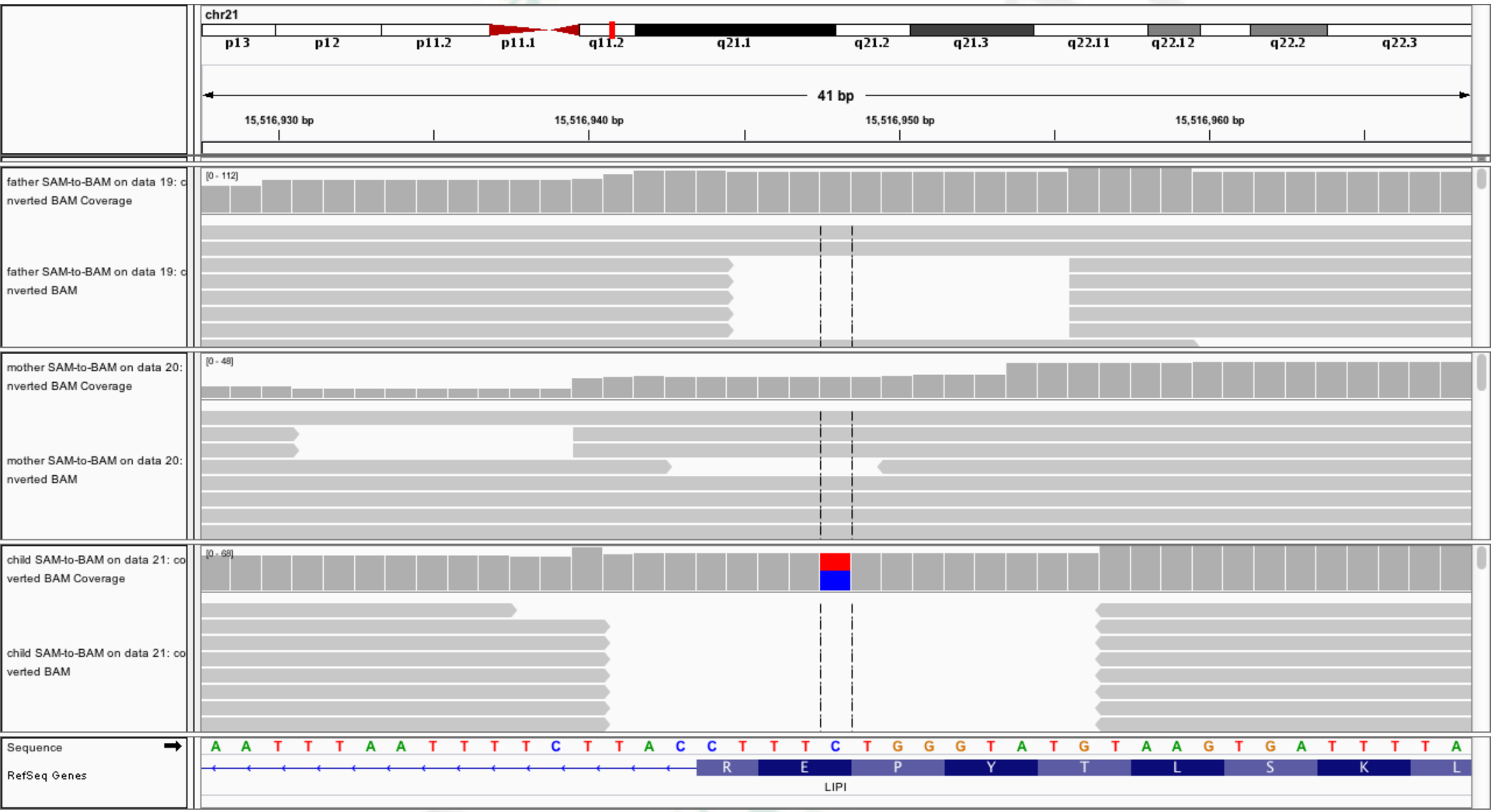
ClinVar

Results: 1 to 20 of 49

<< First < Prev Page 1 of 3 Next > Last >>

	Gene(s)	Condition(s)	Frequency	Clinical significance (Last reviewed)	Review status	C
1. NM_198996.2(**LIPI**):c.1358+1266T>C	**LIPI**	Lung cancer		Uncertain significance	classified by single submitter	2
2. NM_198996.3(**LIPI**):c.322G>A (p.Gly108Ser)	**LIPI**	Malignant melanoma		not provided	not classified by submitter	2
3. NM_198996.3(**LIPI**):c.691C>T (p.His231Tyr)	**LIPI**	Malignant melanoma		not provided	not classified by submitter	2
4. NM_198996.3(**LIPI**):c.749C>T (p.Pro250Leu)	**LIPI**	Malignant melanoma		not provided	not classified by submitter	2
5. NM_198996.3(**LIPI**):c.1158G>A (p.Met386Ile)	**LIPI**	Malignant melanoma		not provided	not classified by submitter	2
NM_198996.3(**LIPI**):c.227G>A	**LIPI**	Hypertriglyceridemia	GO-ESP:0.00969(T)	risk factor	classified	2

IGV


chr21
p13
p12
p11.2
p11.1
q11.2
q21.1
q21.2
q21.3
q22.11
q22.12
q22.2
q22.3
41 bp
15,516,930 bp
15,516,940 bp
15,516,950 bp
15,516,960 bp
father SAM-to-BAM on data 19: c
nverted BAM Coverage
[0 - 112]
father SAM-to-BAM on data 19: c
nverted BAM
mother SAM-to-BAM on data 20:
nverted BAM Coverage
[0 - 48]
mother SAM-to-BAM on data 20:
nverted BAM
child SAM-to-BAM on data 21: co
verted BAM Coverage
[0 - 68]
child SAM-to-BAM on data 21: co
verted BAM
Sequence
A A T T T A A T T T T C T T A C C T T T C T G G G T A T G T A A G T G A T T T T A
RefSeq Genes
R E P Y T L S K L
LIPI

References and web links

- Galaxy
  - PSU Public website: https://usegalaxy.org/
  - UAB: https://www.uab.edu/galaxy
- Bowtie
- GATK
- SAMTools
- Picard Tools
- FreeBayes
- IGV
- SNPEff
- CADD
- Broad Exac
- dbSNP
- OMIM
- ClinVar

Thanks! Questions?

Contact info:

David K. Crossman, Ph.D.

Bioinformatics Director

Heflin Center for Genomic Science

University of Alabama at Birmingham

http://www.heflingenetics.uab.edu

dkcrossm@uab.edu