Variant Prioritization in Cancer and in Trios

So you have the vcf and bam files, what's next?

- Variant annotation
- Variant manipulation
- Sample/Project level Quality Check
- Variant filtering
- Variant visualization (covered by Brandi)
- Result interpretation

Variant annotation

- What annotation does and does not do
- Types of annotation
- Annotators: VEP, SnpEff, ANNOVAR …

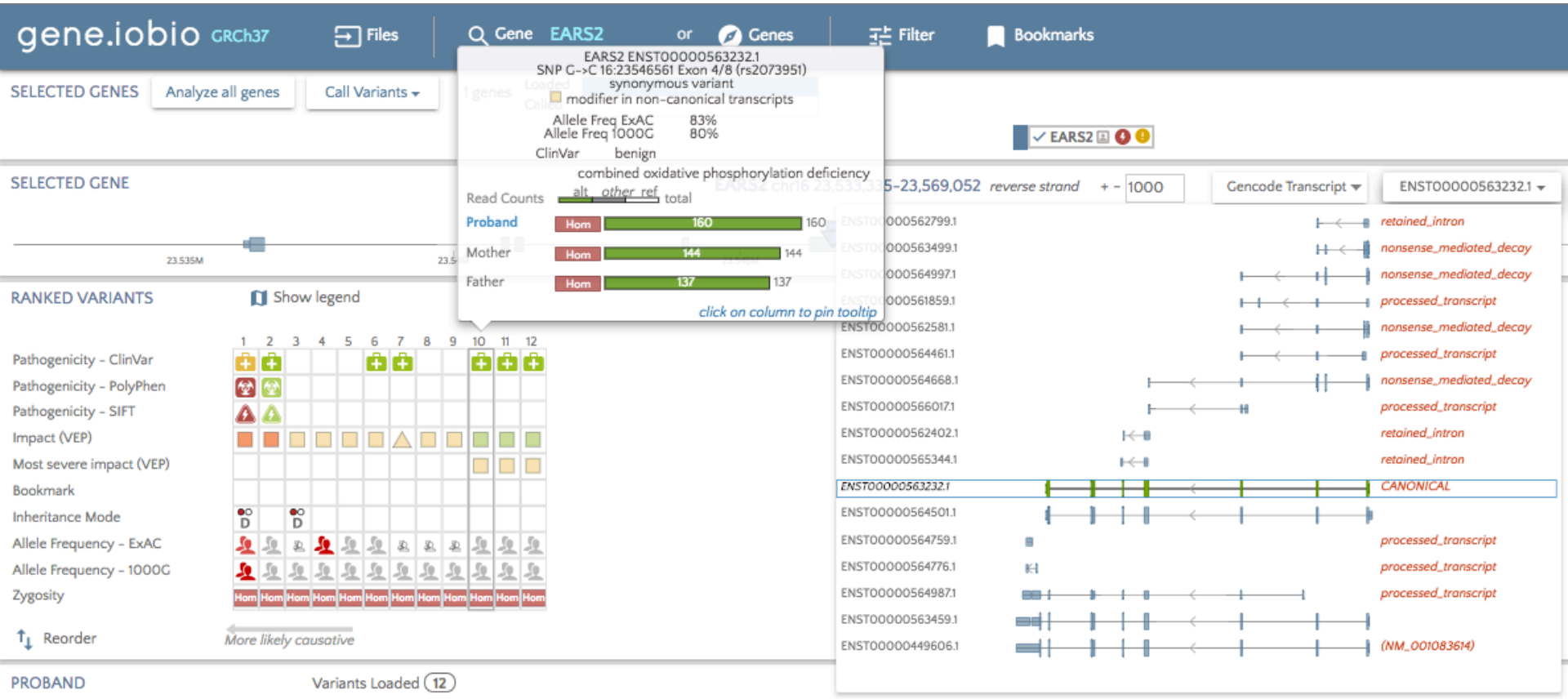
Variant annotation: what annotation does and does not do

- What annotation does
  - Is the variant in a coding exon, and if so does it change an amino acid of the translated protein? If the variant changes the protein, is that likely to be deleterious? Is it in or near a splice site? Does it disturb gene regulation?
  - Has the reference allele been conserved in evolution?
  - Has the variant been seen before? What is its frequency in a population of interest?
  - Does the variant appear in a disease database?
- What annotation does not do
  - The various annotations enable you to prioritize variants for further investigation, but do not by themselves identify causality.

Variant annotation: Types of annotation and sources of data for each

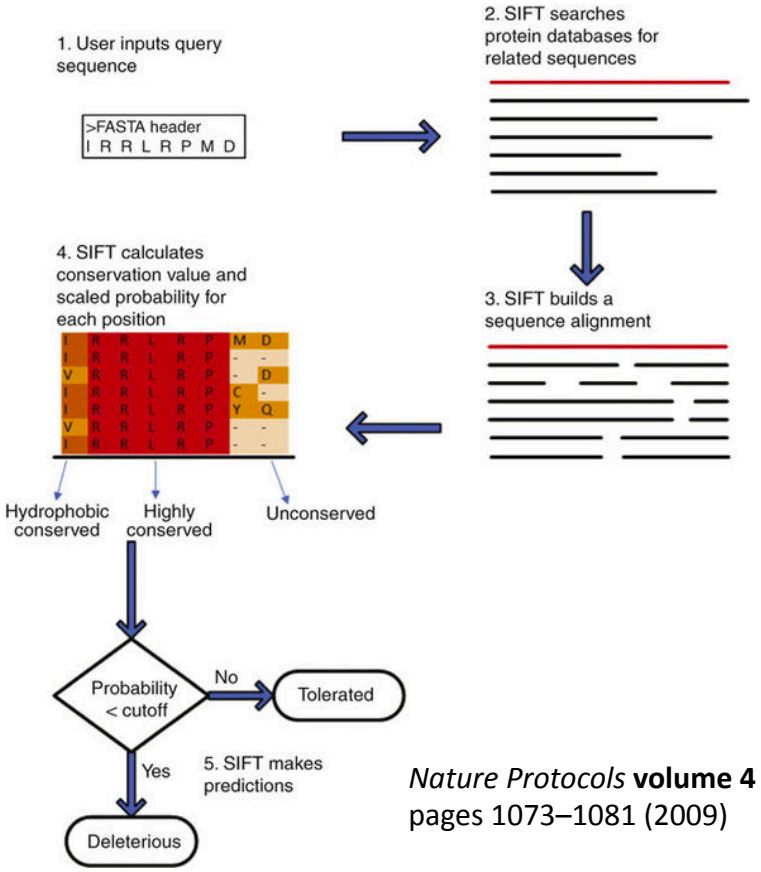
- Genes and other sequence features
- Mutation effect prediction
- Variant frequency in populations
- Clinical association

Variant annotation: Genes and other sequence features



The same mutation has different effects in different transcripts

Variant annotation: SIFT (protein sequence conservation)



Nature Protocols **volume 4**
pages 1073–1081 (2009)

SIFT: sorting intolerant from tolerant

- For single amino acid substitution
- Based on protein sequence
- Estimated by conservation
- Structure information is not used
- Score from 0 to 1, at or below 0.05 is damaging

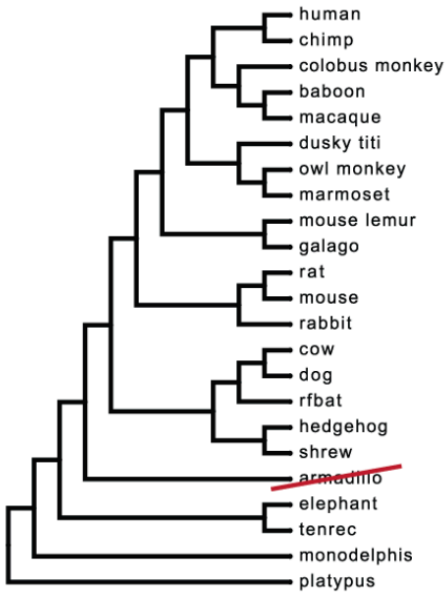
Variant annotation: GERP (DNA sequence conservation)

GERP: Genomic Evolutionary Rate Profiling

- Genome-wide, single-base resolution
- Range of -12.3 to 6.17, with 6.17 being the most conserved

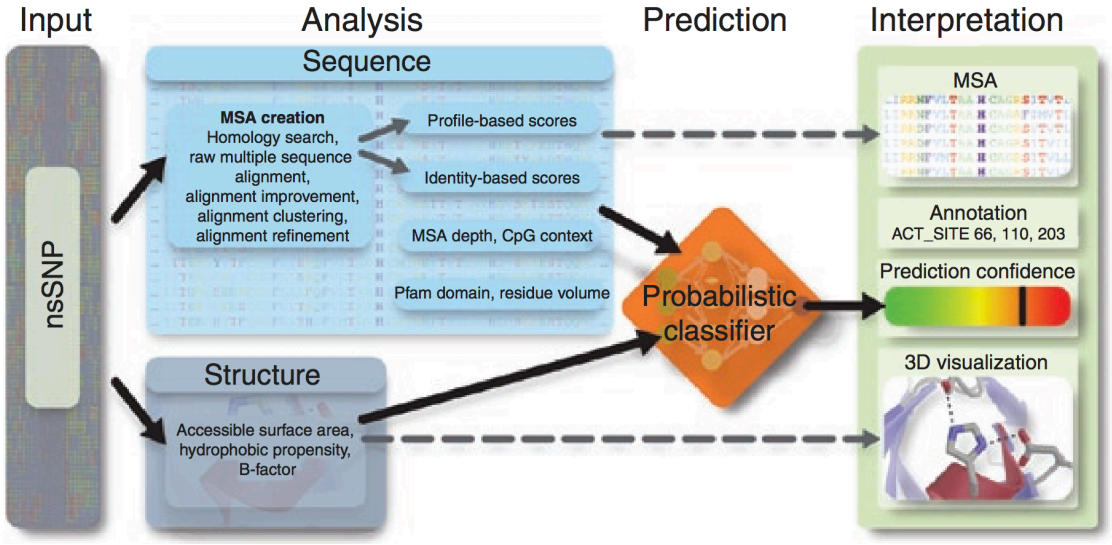
Multiple Sequence Alignment

human	AATACGG	A	ACTTCATTCATT
chimp	AATATGG	A	ACTTCATTCATT
colobus monkey	AGTATGG	A	ACTTCATTCATT
baboon	AGTATGG	A	ACTTCATTCATT
macaque	AGTATGG	A	ACTTCATTCATT
dusky titi	AGTATGG	A	ACTTCATTCATT
owl monkey	AGTATGG	A	ACTTCATTCATT
marmoset	AGTATGG	A	ACTTCATTCATT
mouse lemur	AGTACGG	A	ACTTCATTCATT
galago	AGTACGG	A	ACTTCATTCATT
rat	AGTATGG	A	ACATCGTTCATT
mouse	AGTATGG	A	ACATCTTTCATT
rabbit	AGTATGG	A	ACATCATTCATT
cow	AGTATGG	A	ACATCATTCATT
dog	AGTACGG	A	ACATCATTCATT
rfbat	AGTATGG	A	ACATCGTTCATT
hedgehog	AGTATGG	A	ACATCATTCATT
shrew	AGTATGG	G	ACATCCTTCATT
armadillo	-------	-	------------
elephant	AGTATGG	A	ACATCGTTCATT
tenrec	AGTATGG	A	ACATCGTTCATT
monodelphis	AGTATGG	G	ACATCTTTCATT
platypus	AGTATGG	A	ACGTCATTCATT



PLoS Comput Biol. 2010 Dec 2;6(12)

Variant annotation: Consensus Methods



PolyPhen: Polymorphism Phenotyping

- 8 protein sequence features
- 3 protein structure features
- Between 0 and 1, larger being more deleterious

Nat Methods 7(4):248-249 (2010)

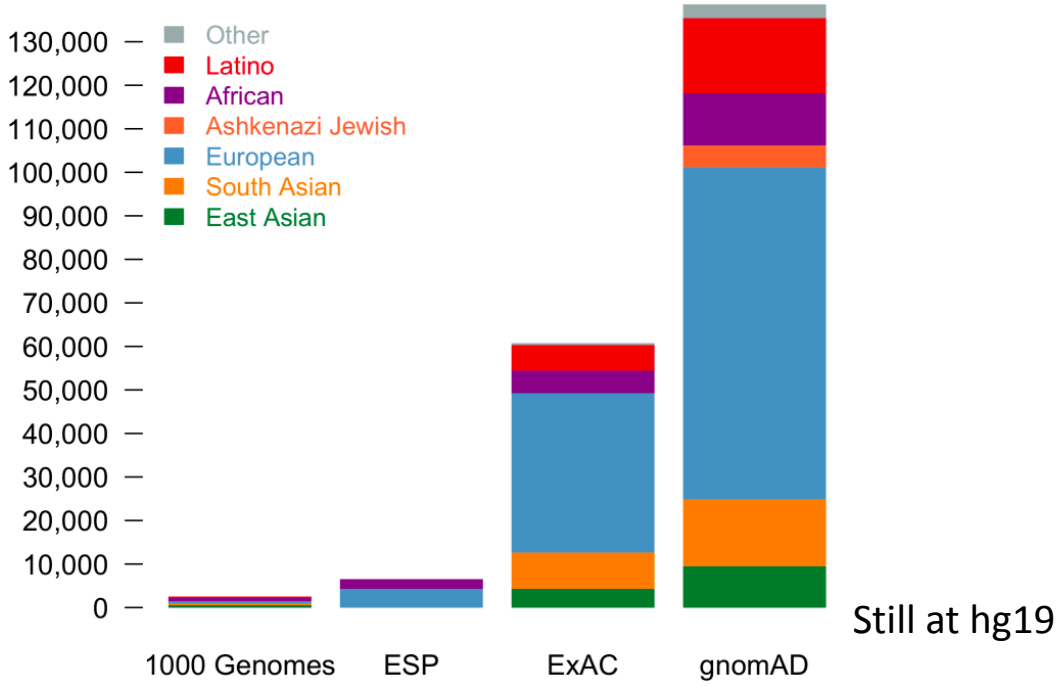
CADD: Combined Annotation Dependent Depletion

- Generate 63 distinct annotations, including scores from PhastCons, GERP, PhyloP, SIFT and PolyPhen, etc., ENCODE data (summarized at various levels), gene body annotations…
- Build a support vector machine (SVM) that estimates, for a given variant, whether it is likely to be observed or simulated, based on its combined annotation profile
- The C score ranges from 1 to 99, with a higher score indicating greater deleteriousness. Values >= 10 are predicted to be the 10% most deleterious substitutions, >= 20 indicate the 1% most deleterious.
- Score all possible ~8.6 billion possible SNVs of hg19

Nature Genetics 46, 310 (2014) and http://cadd.gs.washington.edu

Variant annotation: Variant frequency in populations

gnomAD



Still at hg19

https://macarthurlab.org/2017/02/27/the-genome-aggregation-database-gnomad/
http://gnomad.broadinstitute.org/

Variant	Source	Consequence	Annotation	Flags	Allele Count	Allele Number	Number of Homozygotes	Allele Frequency
16:23568591 CGCCGTCCTACGGGGCGGCC... / C (rs771097223)	E	c.26_63+10delTGCAGCG... †	splice donor	LC LoF	2	144656	0	1.383e-5
16:23568602 G / T (rs113410455)	E G	c.63C>A†	splice region		12	168482	0	7.122e-5

Contributing projects

1000 Genomes
1958 Birth Cohort
ALSGEN
Alzheimer's Disease Sequencing Project (ADSP)
Atrial Fibrillation Genetics Consortium (AFGen)
Estonian Genome Center, University of Tartu (EGCUT)
Bulgarian Trios
Finland-United States Investigation of NIDDM Genetics (FUSION)
Finnish Twin Cohort Study
FINN-ADGEN
FINRISK
Framingham Heart Study
Génome Québec - Genizon Biobank
Genomic Psychiatry Cohort
GoT2D
Genotype-Tissue Expression Project (GTEx)
Health2000
Inflammatory Bowel Disease:
- Helsinki University Hospital Finland
- NIDDK IBD Genetics Consortium
- Quebec IBD Genetics Consortium

Jackson Heart Study
Kuopio Alzheimer Study
LifeLines Cohort
MESTA
METabolic Syndrome In Men (METSIM)
Finnish Migraine Study
Myocardial Infarction Genetics Consortium (MIGen):
- Leicester Exome Seq
- North German MI Study
- Ottawa Genomics Heart Study
- Pakistan Risk of Myocardial Infarction Study (PROMIS)
- Precocious Coronary Artery Disease Study (PROCARDIS)
- Registre Gironi del COR (REGICOR)
- South German MI Study
- Variation in Recovery: Role of Gender on Outcomes of Young AMI Patients (VIRGO)

National Institute of Mental Health (NIMH) Controls
NHLBI-GO Exome Sequencing Project (ESP)
NHLBI TOPMed
Schizophrenia Trios from Taiwan
Sequencing Initiative Suomi (SiSu)
SIGMA-T2D
Swedish Schizophrenia & Bipolar Studies
T2D-GENES
- GoDARTS

T2D-SEARCH
The Cancer Genome Atlas (TCGA)

Variant annotation: Clinical association

Cancer

https://cancer.sanger.ac.uk/cosmic

Mendelian Disease

https://www.ncbi.nlm.nih.gov/clinvar/

Variant annotation: a practical example

gene	rs_ids	impact_ severity	impact	clinvar_sig	clinvar_disease_ name	max_aaf_all	gnomad_ num_het	gnomad_nu m_hom_alt	cadd_ scaled	gerp_bp _score
PAH	rs5030855	LOW	intron_variant	pathogenic	Phenylketonuria|not_provided	0.003	60	0	5.65	3.73
PAH	rs5030858	MED	missense_variant	pathogenic	Phenylketonuria|not_provided	0.001744186	191	0	19.02	1.44

Variant annotation: Annotators

- VEP
- SnpEff
- ANNOVAR

Web based services:
http://wannovar.wglab.org/
http://uswest.ensembl.org/Tools/VEP
http://snp.gs.washington.edu/SeattleSeqAnnotation150/

An example of a SnpEff annotated variant

```
##INFO=<ID=ANN,Number=.,Type=String,Description="Functional annotations: 'Allele | Annotation | Annotation_Impact | Gene_Name | Gene_ID | Feature_Type | Featur
e_ID | Transcript_BioType | Rank | HGVS.c | HGVS.p | cDNA.pos / cDNA.length | CDS.pos / CDS.length | AA.pos / AA.length | Distance | ERRORS / WARNINGS / INFO'"
```

```
16      23546218        .       C       A       234.965 .       AB=0.25;ABP=5.18177;AC=3;AF=0.5;AN=6;AO=12;CIGAR=1X;DP=21;DPB=21;DPRA=1.25;EPP=9.52472;EPPR=3.2
5157;GTI=0;LEN=1;MEANALT=1;MQM=60;MQMR=60;NS=3;NUMALT=1;ODDS=2.17108;PAIRED=1;PAIREDR=1;PAO=0;PQA=0;PQR=0;PRO=0;QA=391;QR=204;RO=9;RPL=1;RPP=21.1059;RPPR=3.251
57;RPR=11;RUN=1;SAF=10;SAP=14.5915;SAR=2;SRF=8;SRP=14.8328;SRR=1;technology.illumina=1;TYPE=snp;CallSet=platypus,sam,ssvar;ANN=A|missense_variant|MODERATE|EARS
2|ENSG00000103356|transcript|ENST00000563232|protein_coding|4/8|c.949G>T|p.Gly317Cys|950/1709|949/1605|317/534||,A|missense_variant|MODERATE|EARS2|ENSG00000103
356|transcript|ENST00000449606|protein_coding|4/9|c.949G>T|p.Gly317Cys|981/3961|949/1572|317/523||,A|missense_variant|MODERATE|EARS2|ENSG00000103356|transcript
|ENST00000563459|protein_coding|4/10|c.949G>T|p.Gly317Cys|956/3840|949/1572|317/523||,A|missense_variant|MODERATE|EARS2|ENSG00000103356|transcript|ENST00000564
501|protein_coding|4/9|c.949G>T|p.Gly317Cys|1337/2027|949/1521|317/506||,A|upstream_gene_variant|MODIFIER|SUB1P4|ENSG00000260247|transcript|ENST00000566062|pro
cessed_pseudogene||n.-2164C>A|||||2164|,A|downstream_gene_variant|MODIFIER|EARS2|ENSG00000103356|transcript|ENST00000566017|processed_transcript||n.*377G>T||||
|377|,A|downstream_gene_variant|MODIFIER|EARS2|ENSG00000103356|transcript|ENST00000564668|nonsense_mediated_decay||c.*17327G>T|||||423|,A|non_coding_transcript
_exon_variant|MODIFIER|EARS2|ENSG00000103356|transcript|ENST00000564987|processed_transcript|3/9|n.573G>T||||||,A|non_coding_transcript_exon_variant|MODIFIER|E
ARS2|ENSG00000103356|transcript|ENST00000565344|retained_intron|1/2|n.322G>T||||||,A|non_coding_transcript_exon_variant|MODIFIER|EARS2|ENSG00000103356|transcri
pt|ENST00000562402|retained_intron|1/2|n.553G>T||||||      GT:DP:AD:AO:RO  1/1:11:0,11:11:0        0/1:4:3,1:1:3   0/0:6:6,0:0:6
```

Variant manipulation

- What's in a VCF?
- What to do with a VCF?
  - Indexing
  - Extract regions
  - Stats
  - Filtering
  - Merging
  - Intersecting
  - …..
- Tools
  - Tabix
  - VCFtools
  - BCFtools

Variant manipulation: what's in a vcf?

Every VCF file has three parts in the following order:

- Meta-information lines (lines beginning with "##").
- One header line (line beginning with "#CHROM").
- Data lines contain marker and genotype data (one variant per line). A data line is called a VCF record.

Example VCF file

```
##fileformat=VCFv4.2
##FORMAT=<ID=GT,Number=1,Type=Integer,Description="Genotype">
##FORMAT=<ID=GP,Number=G,Type=Float,Description="Genotype Probabilities">
##FORMAT=<ID=PL,Number=G,Type=Float,Description="Phred-scaled Genotype Likelihoods">
#CHROM  POS      ID       REF     ALT     QUAL    FILTER  INFO    FORMAT  SAMP001 SAMP002
20      1291018  rs11449  G       A       .       PASS    .       GT      0/0     0/1
20      2300608  rs84825  C       T       .       PASS    .       GT:GP   0/1:.   0/1:0.03,0.97,0
20      2301308  rs84823  T       G       .       PASS    .       GT:PL   ./.:.   1/1:10,5,0
```

https://faculty.washington.edu/browning/intro-to-vcf.html

Variant manipulation tools: Tabix

- A tool for working with sorted block-wise compressed tab-separated files containing coordinates (positions or ranges)
- Input file must be sorted and compressed by bgzip
- Documentation: http://www.htslib.org/doc/tabix.html

```
Usage:   tabix [OPTIONS] [FILE] [REGION [...]]

Indexing Options:
   -0, --zero-based           coordinates are zero-based
   -b, --begin INT            column number for region start [4]
   -c, --comment CHAR         skip comment lines starting with CHAR [null]
   -C, --csi                  generate CSI index for VCF (default is TBI)
   -e, --end INT              column number for region end (if no end, set INT to -b) [5]
   -f, --force                overwrite existing index without asking
   -m, --min-shift INT        set minimal interval size for CSI indices to 2^INT [14]
   -p, --preset STR           gff, bed, sam, vcf
   -s, --sequence INT         column number for sequence names (suppressed by -p) [1]
   -S, --skip-lines INT       skip first INT lines [0]

Querying and other options:
   -h, --print-header         print also the header lines
   -H, --only-header          print only the header lines
   -l, --list-chroms          list chromosome names
   -r, --reheader FILE        replace the header with the content of FILE
   -R, --regions FILE         restrict to regions listed in the file
   -T, --targets FILE         similar to -R but streams rather than index-jumps
```

Variant manipulation tools: Tabix

Tabix needs a sorted VCF, sort it if it's not already sorted:

```
(zgrep "^#" variants.vcf; zgrep -v "^#" variants.vcf.gz | sort -k1,1 -k2,2n )| bgzip  > variants.sorted.vcf.gz
```

Variant manipulation tools: BCFtools

More practical examples in the second session

- A toolset for manipulating data stored in VCF: http://www.htslib.org/doc/bcftools.html

```
-- Indexing
   index        index VCF/BCF files

-- VCF/BCF manipulation
   annotate     annotate and edit VCF/BCF files
   concat       concatenate VCF/BCF files from the same set of samples
   convert      convert VCF/BCF files to different formats and back
   isec         intersections of VCF/BCF files
   merge        merge VCF/BCF files files from non-overlapping sample sets
   norm         left-align and normalize indels
   plugin       user-defined plugins
   query        transform VCF/BCF into user-defined formats
   reheader     modify VCF/BCF header, change sample names
   sort         sort VCF/BCF file
   view         VCF/BCF conversion, view, subset and filter VCF/BCF files

-- VCF/BCF analysis
   call         SNP/indel calling
   consensus    create consensus sequence by applying VCF variants
   cnv          HMM CNV calling
   csq          call variation consequences
   filter       filter VCF/BCF files using fixed thresholds
   gtcheck      check sample concordance, detect sample swaps and contamination
   mpileup      multi-way pileup producing genotype likelihoods
   roh          identify runs of autozygosity (HMM)
   stats        produce VCF/BCF stats
```

Sample/Project level Quality Check: Motivation

There are many reasons why the genotyped samples may not represent the intended relationships with each other

- Sample swaps
- Duplicate samples
- Sample contamination
- Cryptic relatedness
- Pedigree errors/False paternity

Sample/Project level Quality Check: Tools

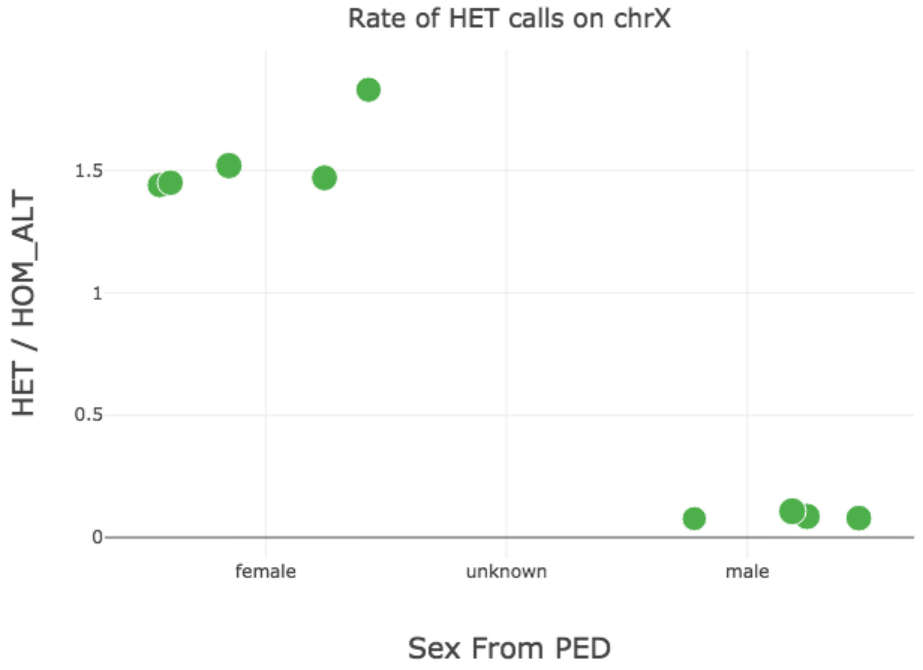
- PLINK
- KING
- Peddy

Peddy seeks discrepancies between relationships and sexes as indicated in a .*PED* file with those inferred from genotypes.

- Works directly on the VCF and PED
- It samples ~23,000 well-behaved exome sites in the genome
- Generates interactive plots for QC
- Runs on a trio in a few seconds.
- Runs on 2K samples in ~10 minutes.

Sample/Project level Quality Check: Sex Check

- Sex check -- rate of HET on chrX
  - Samples indicated as Male should have ~ 0 HET calls in non-PAR regions of X chromosome

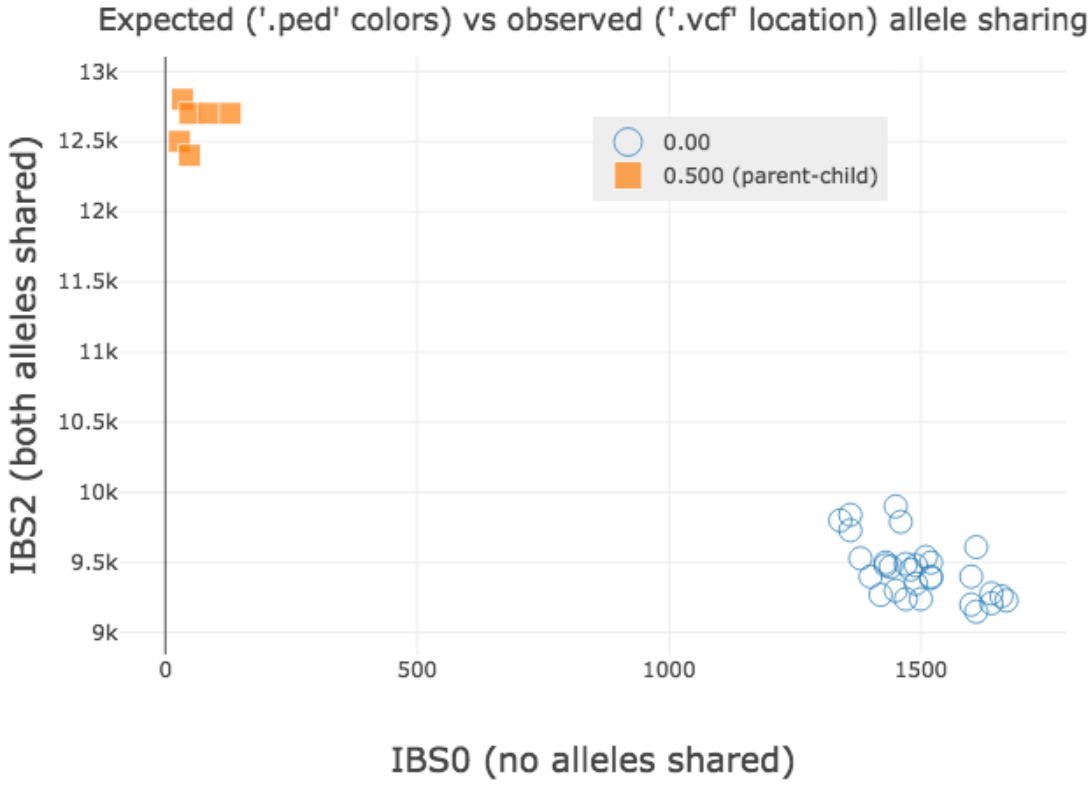


Sample/Project level Quality Check: Ancestry Check

- PCA / Ancestry
  - matrix of genotypes and project onto principal components of the 2504 1KG samples
  - Infer ancestry using SVM trained on 1KG samples

Sample/Project level Quality Check: Pedigree Check

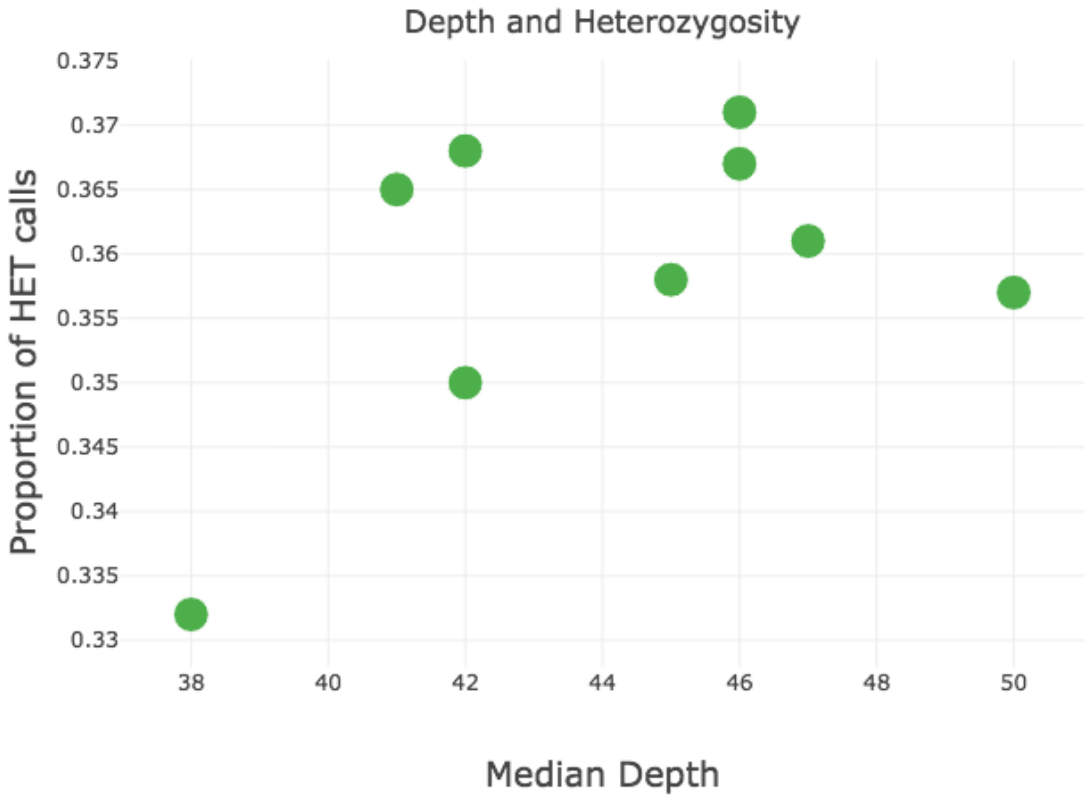
- Ped check (23K sampled sites) on all pairwise combinations of samples
  - IBS0 (0 alleles are shared) -- should be rare for parent-kid pairs
  - IBS2 (both HOM_REF, both HET, both HOM_ALT, i.e. sharing the same genotype)



Parent-child pairs should have ~ 0 sites where they share no alleles (e.g. Mom: A/A, Kid: T/T). This is called IBS0 (identity-by-state)

Sample/Project level Quality Check: Het Check

- Het check (23K sampled sites across autosome)
  - Rate of HET calls (high can indicate contamination, low can indicate consanguinity)
  - Median depth
  - Call-rate



Sample/Project level Quality Check: Table View

family_id	sample_id	paternal_id	maternal_id	sex	phenotype	het_call_rate	het_ratio	het_mean_depth	het_iqr_baf	ancestry-prediction	sex_het_ratio	sex_error
GMDP_3_0047	GMDP_3_0047_1	GMDP_3_0047_2	GMDP_3_0047_3	female	affected	0.9985	0.3647	60.15	0.2232	EUR	1.831	False
GMDP_3_0047	GMDP_3_0047_2	-9	-9	male	unaffected	0.9993	0.3707	80.43	0.1987	EUR	0.1066	False
GMDP_3_0047	GMDP_3_0047_3	-9	-9	female	unaffected	0.9987	0.3685	65.92	0.2154	EUR	1.472	False
GMDP_3_0046	GMDP_3_0046_1	GMDP_3_0046_2	GMDP_3_0046_3	female	affected	0.9997	0.361	91.62	0.1849	AMR	1.441	False
GMDP_3_0046	GMDP_3_0046_2	-9	-9	male	unaffected	0.9995	0.3675	80.83	0.1965	AMR	0.07947	False
GMDP_3_0046	GMDP_3_0046_3	-9	-9	female	unaffected	0.9993	0.3578	80.56	0.1943	AMR	1.451	False
GMDP_3_0045	GMDP_3_0045_1	GMDP_3_0045_2	GMDP_3_0045_3	male	affected	0.9925	0.3319	44.88	0.2534	AMR	0.07692	False
GMDP_3_0045	GMDP_3_0045_2	-9	-9	male	unaffected	0.9969	0.3501	61.59	0.2195	AMR	0.08621	False
GMDP_3_0045	GMDP_3_0045_3	-9	-9	female	unaffected	0.9998	0.3572	91.54	0.1835	AMR	1.521	False

family_id sample_ic paternal_id maternal_id sex phenotyp het_call_rate het_ratio het_mean_dep het_iqr_baf ancestry-prediction sex_het_ratio sex_error

peddy reports additional metrics that can be plotted to visually diagnose common sample problems.

Sample/Project level Quality Check: Running peddy

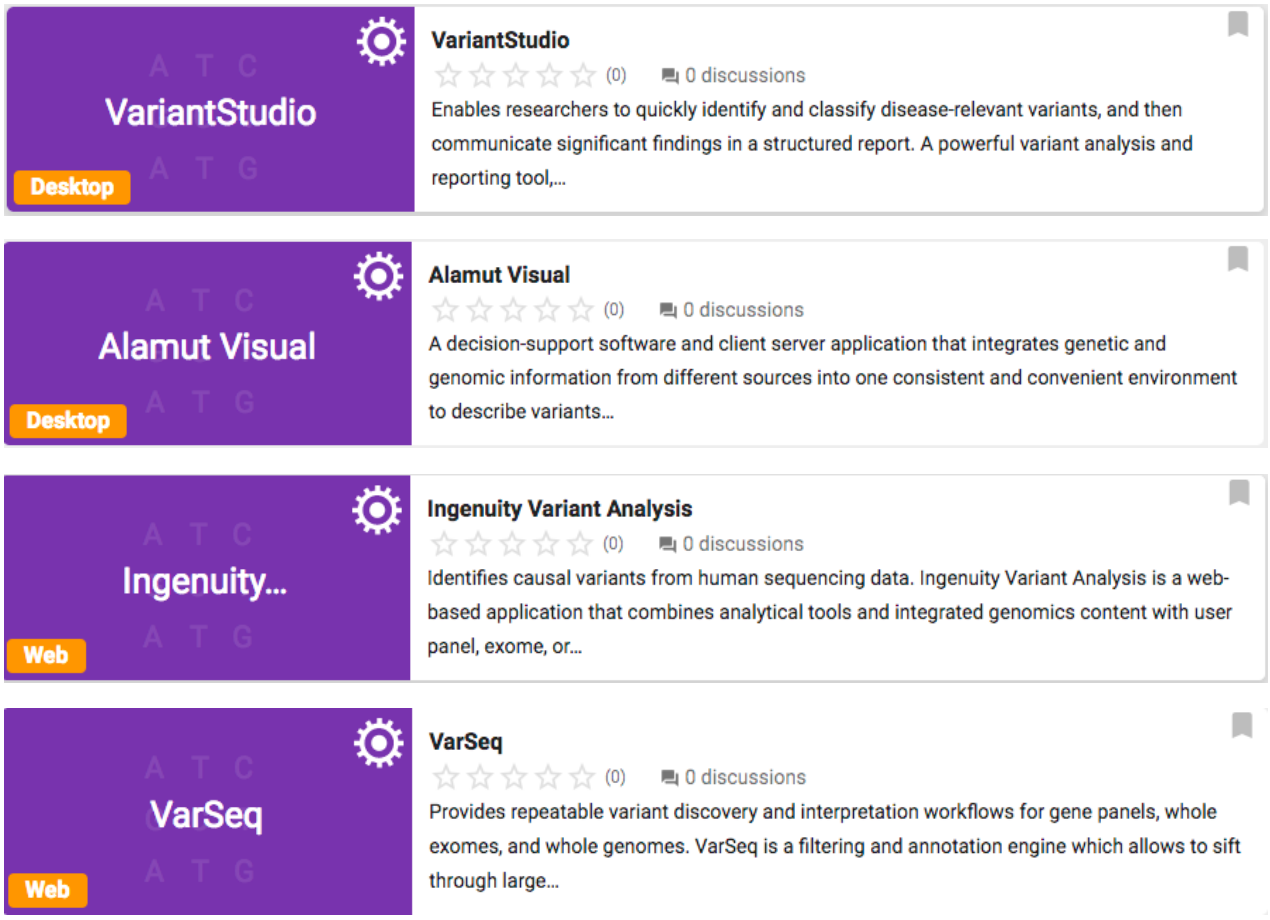
GMDP_3_0045	GMDP_3_0045_1	GMDP_3_0045_2	GMDP_3_0045_3	1	2
GMDP_3_0045	GMDP_3_0045_2	0	0	1	1
GMDP_3_0045	GMDP_3_0045_3	0	0	2	1
GMDP_3_0046	GMDP_3_0046_1	GMDP_3_0046_2	GMDP_3_0046_3	2	2
GMDP_3_0046	GMDP_3_0046_2	0	0	1	1
GMDP_3_0046	GMDP_3_0046_3	0	0	2	1
GMDP_3_0047	GMDP_3_0047_1	GMDP_3_0047_2	GMDP_3_0047_3	2	2
GMDP_3_0047	GMDP_3_0047_2	0	0	1	1
GMDP_3_0047	GMDP_3_0047_3	0	0	2	1

Columns in a .ped file
Family_id
Sample_id
Father_id
Mother_id
Gender
Affected_status

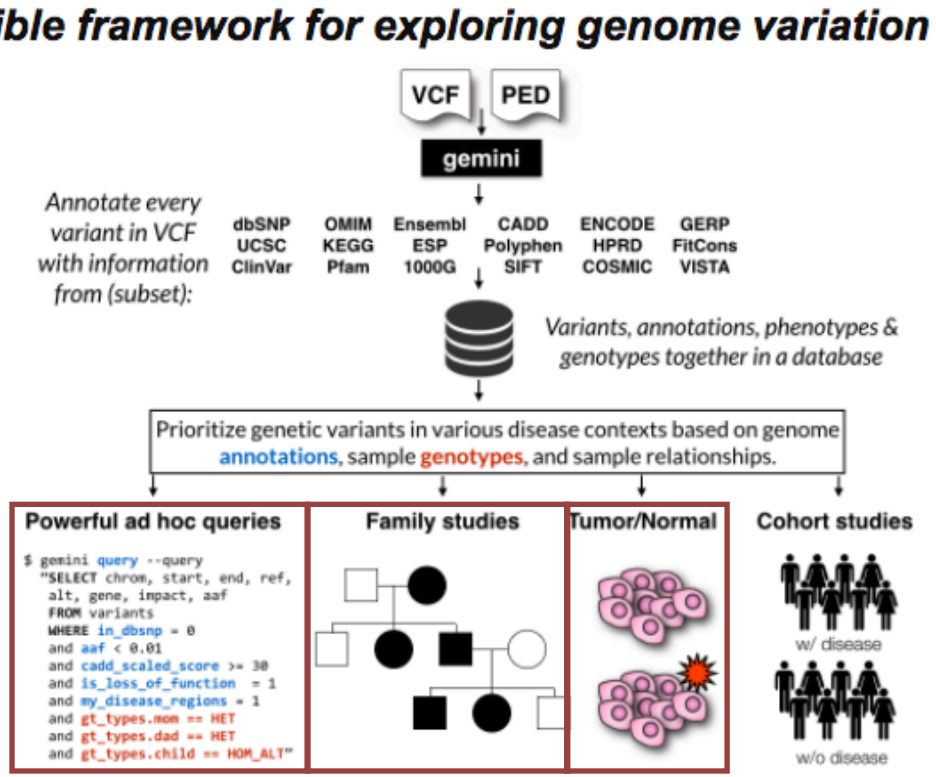
```
python -m peddy vcf_file ped_file
```

Results will be generated in your working directory

Variant Filtering: commercial GUI tools



Variant Filtering: GEMINI



What can you do with Gemini?

- Load a VCF into an "easy to use" database
- Query (fetch data) from database based on annotations or subject genotypes
- Analyze simple genetic models

https://gemini.readthedocs.io/en/latest/#

Gemini Documentation:
https://gemini.readthedocs.org
Annotations and information available for Gemini:
https://gemini.readthedocs.org/en/latest/content/database_schema.html

Extensive built-in annotation for Hg19 but not Hg38 :(
Use vcf2db to import vcf if using Hg38

Variant Filtering: GEMINI built-in analysis tools

- Built-in analysis tools
  - **common_args**: common arguments
  - **comp_hets**: Identifying potential compound heterozygotes
  - **mendelian_error**: Identify non-mendelian transmission.
  - **de_novo**: Identifying potential de novo mutations.
  - **autosomal_recessive**: Find variants meeting an autosomal recessive model.
  - **autosomal_dominant**: Find variants meeting an autosomal dominant model.
  - **x_linked_recessive**: x-linked recessive inheritance
  - **x_linked_dominant**: x-linked dominant inheritance
  - **x_linked_de_novo**: x-linked de novo
  - **gene_wise**: Custom genotype filtering by gene.
  - **pathways**: Map genes and variants to KEGG pathways.
  - **interactions**: Find genes among variants that are interacting partners.
  - **lof_sieve**: Filter LoF variants by transcript position and type
  - **amend**: updating / changing the sample information
  - **annotate**: adding your own custom annotations
  - **region**: Extracting variants from specific regions or genes
  - **windower**: Conducting analyses on genome “windows”.
  - **stats**: Compute useful variant statistics.
  - **burden**: perform sample-wise gene-level burden calculations
  - **ROH**: Identifying runs of homozygosity
  - **set_somatic**: Flag somatic variants
  - **actionable_mutations**: Report actionable somatic mutations and drug-gene interactions
  - **fusions**: Report putative gene fusions
  - **db_info**: List the gemini database tables and columns

Variant Filtering: GEMINI database schema

- The GEMINI database schema
  - The **`variants`** table
  - The **`variant_impacts`** table
  - Details of the **`impact`** and **`impact_severity`** columns
  - The **`samples`** table
  - The **`resources`** table
  - The **`version`** table
  - The **`gene_detailed`** table
  - The **`gene_summary`** table

https://gemini.readthedocs.io/en/latest/content/database_schema.html

Variant Filtering: Core VCF fields in the variants table

Core VCF fields

column_name	type	notes
chrom	STRING	The chromosome on which the variant resides (from VCF `CHROM` field).
start	INTEGER	The 0-based start position. (from VCF `POS` field, but converted to 0-based coordinates)
end	INTEGER	The 1-based end position. (from VCF `POS` field, yet inferred based on the size of the variant)
vcf_id	STRING	The VCF `ID` field.
variant_id	INTEGER	PRIMARY_KEY
anno_id	INTEGER	Variant transcript number for the most severely affected transcript
ref	STRING	Reference allele (from VCF `REF` field)
alt	STRING	Alternate allele for the variant (from VCF `ALT` field)
qual	INTEGER	Quality score for the assertion made in ALT (from VCF `QUAL` field)
filter	STRING	A string of filters passed/failed in variant calling (from VCF `FILTER` field)

Variant Filtering: Genotype fields in the variants table

Genotype information

gts	BLOB	A compressed binary vector of sample genotypes (e.g., "A/A", "A|G", "G/G") - Extracted from the VCF `GT` genotype tag.
gt_types	BLOB	A compressed binary vector of numeric genotype "types" (e.g., 0, 1, 2) - Inferred from the VCF `GT` genotype tag.
gt_phases	BLOB	A compressed binary vector of sample genotype phases (e.g., False, True, False) - Extracted from the VCF `GT` genotype tag's allele delimiter e.g., `A/G` means an unphased genotype. Value is **FALSE**. e.g., `A|G` means a phased genotype. Value is **TRUE**.
gt_depths	BLOB	A compressed binary vector of the depth of aligned sequence observed for each sample - Extracted from the VCF `DP` genotype tag.
gt_ref_depths	BLOB	A compressed binary vector of the depth of reference alleles observed for each sample - Extracted from the VCF `AD` genotype tag.
gt_alt_depths	BLOB	A compressed binary vector of the depth of alternate alleles observed for each sample - Extracted from the VCF `AD` genotype tag.
gt_alt_freqs	BLOB	A compressed binary (float) vector of the frequency of alternate alleles observed for each sample - equivalent to gt_alt_depths / (gt_alt_depths + gt_ref_depths)
gt_quals	BLOB	A compressed binary vector of the genotype quality (PHRED scale) estimates for each sample - Extracted from the VCF `GQ` genotype tag.
gt_phred_ll_homref	BLOB	A compressed binary vector of the phred-scaled genotype likelihood of the 0/0 genotype estimates for each sample - Extracted from the VCF `GL` or `PL` tag. - New in version 0.13.0
gt_phred_ll_het	BLOB	A compressed binary vector of the phred-scaled genotype likelihood of the 0/1 genotype estimates for each sample - Extracted from the VCF `GL` or `PL` tag. - New in version 0.13.0
gt_phred_ll_homalt	BLOB	A compressed binary vector of the phred-scaled genotype likelihood of the 1/1 genotype estimates for each sample - Extracted from the VCF `GL` or `PL` tag. - New in version 0.13.0

Variant Filtering: Gene fields in the variants table

Gene information

gene	STRING	Corresponding gene name of the highly affected transcript
transcript	STRING	The variant transcript that was most severely affected (for two equally affected transcripts, the protein_coding biotype is prioritized (SnpEff/VEP)
is_exonic	BOOL	Does the variant affect an exon for >= 1 transcript?
is_coding	BOOL	Does the variant fall in a coding region (excl. 3' & 5' UTRs) for >= 1 transcript?
is_lof	BOOL	Based on the value of the impact col, is the variant LOF for >= transcript?
is_splicing	BOOL	Does the variant affect a canonical or possible splice site? That is, set to TRUE if the SO term is any of `splice_acceptor_variant`, `splice_donor_variant`, or `splice_region_variant`.
exon	STRING	Exon information for the severely affected transcript
codon_change	STRING	What is the codon change?
aa_change	STRING	What is the amino acid change (for a snp)?
aa_length	STRING	Has the format pos/len when biotype=protein_coding, is empty otherwise. len=protein length. pos = position of the amino acid change when is_coding=1 and is_exonic=1, '-' otherwise.
biotype	STRING	The 'type' of the severely affected transcript (e.g., protein-coding, pseudogene, rRNA etc.) (`only SnpEff`)
impact	STRING	The consequence of the most severely affected transcript
impact_so	STRING	The Sequence ontology term for the most severe consequence
impact_severity	STRING	Severity of the highest order observed for the variant
polyphen_pred	STRING	Polyphen predictions for the snps for the severely affected transcript (`only VEP`)
polyphen_score	FLOAT	Polyphen scores for the severely affected transcript (`only VEP`)
sift_pred	STRING	SIFT predictions for the snp's for the most severely affected transcript (`only VEP`)
sift_score	FLOAT	SIFT scores for the predictions (`only VEP`)
pfam_domain	STRING	Pfam protein domain that the variant affects

Variant Filtering: Pop frequency fields in the variants table

Population information

in_dbsnp	BOOL	Is this variant found in dbSNP? 0 : Absence of the variant in dbsnp 1 : Presence of the variant in dbsnp
rs_ids	STRING	A comma-separated list of rs ids for variants present in dbSNP
in_hm2	BOOL	Whether the variant was part of HapMap2.
in_hm3	BOOL	Whether the variant was part of HapMap3.
in_esp	BOOL	Presence/absence of the variant in the ESP project data
in_1kg	BOOL	Presence/absence of the variant in the 1000 genome project data (phase 3)
aaf_esp_ea	FLOAT	Minor Allele Frequency of the variant for European Americans in the ESP project
aaf_esp_aa	FLOAT	Minor Allele Frequency of the variant for African Americans in the ESP project
aaf_esp_all	FLOAT	Minor Allele Frequency of the variant w.r.t both groups in the ESP project
aaf_1kg_amr	FLOAT	Allele frequency of the variant in AMR population based on AC/AN (1000g project, phase 3)
aaf_1kg_eas	FLOAT	Allele frequency of the variant in EAS population based on AC/AN (1000g project, phase 3)
aaf_1kg_sas	FLOAT	Allele frequency of the variant in SAS population based on AC/AN (1000g project, phase 3)
aaf_1kg_afr	FLOAT	Allele frequency of the variant in AFR population based on AC/AN (1000g project, phase 3)
aaf_1kg_eur	FLOAT	Allele frequency of the variant in EUR population based on AC/AN (1000g project, phase 3)
aaf_1kg_all	FLOAT	Global allele frequency (based on AC/AN) (1000g project - phase 3)
in_exac	BOOL	Presence/absence of the variant in ExAC (Exome Aggregation Consortium) data (Broad)
aaf_exac_all	FLOAT	Raw allele frequency (population independent) of the variant based on ExAC exomes (AF)
aaf_adj_exac_all	FLOAT	Adjusted allele frequency (population independent) of the variant based on ExAC (Adj_AC/Adj_AN)
aaf_adj_exac_afr	FLOAT	Adjusted allele frequency of the variant for AFR population in ExAC (AC_AFR/AN_AFR)
aaf_adj_exac_amr	FLOAT	Adjusted allele frequency of the variant for AMR population in ExAC (AC_AMR/AN_AMR)
aaf_adj_exac_eas	FLOAT	Adjusted allele frequency of the variant for EAS population in ExAC (AC_EAS/AN_EAS)
aaf_adj_exac_fin	FLOAT	Adjusted allele frequency of the variant for FIN population in ExAC (AC_FIN/AN_FIN)
aaf_adj_exac_nfe	FLOAT	Adjusted allele frequency of the variant for NFE population in ExAC (AC_NFE/AN_NFE)
aaf_adj_exac_oth	FLOAT	Adjusted allele frequency of the variant for OTH population in ExAC (AC_OTH/AN_OTH)
aaf_adj_exac_sas	FLOAT	Adjusted allele frequency of the variant for SAS population in ExAC (AC_SAS/AN_SAS)
max_aaf_all	FLOAT	the maximum of aaf_gnomad{afr,amr,eas,nfe,sas},aaf_esp_ea, aaf_esp_aa, aaf_1kg_amr, aaf_1kg_eas,aaf_1kg_sas,aaf_1kg_afr,aaf_1kg_eur,aaf_adj_exac_afr,aaf_adj_exac_amr,aaf_adj_exac_eas,aaf_adj_exac_nfe,aaf_adj_exac_sas. and -1 if none of those databases/populations contain the variant.
exac_num_het	INTEGER	The number of heterozygote genotypes observed in ExAC. Pulled from the ExAC `AC_Het` INFO field.
exac_num_hom_alt	INTEGER	The number of homozygous alt. genotypes observed in ExAC. Pulled from the ExAC `AC_Het` INFO field.
exac_num_chroms	INTEGER	The number of chromosomes underlying the ExAC variant call. Pulled from the ExAC `AN_Adj` INFO field.
aaf_gnomad_all	FLOAT	Allele frequency (population independent) of the variant in gnomad,
aaf_gnomad_afr	FLOAT	Allele frequency (AFR population) of the variant in gnomad
aaf_gnomad_amr	FLOAT	Allele frequency (AMR population) of the variant in gnomad
aaf_gnomad_asj	FLOAT	Allele frequency (ASJ population) of the variant in gnomad
aaf_gnomad_eas	FLOAT	Allele frequency (EAS population) of the variant in gnomad
aaf_gnomad_fin	FLOAT	Allele frequency (FIN population) of the variant in gnomad
aaf_gnomad_nfe	FLOAT	Allele frequency (NFE population) of the variant in gnomad
aaf_gnomad_oth	FLOAT	Allele frequency (OTH population) of the variant in gnomad
aaf_gnomad_sas	FLOAT	Allele frequency (SAS population) of the variant in gnomad
gnomad_num_het	INTEGER	Number of het genotypes observed in gnomad
gnomad_num_hom_alt	INTEGER	Number of hom_alt genotypes observed in gnomad
gnomad_num_chroms	INTEGER	Number of chromosomes genotyped in gnomad

Variant Filtering: disease phenotype fields in the variants table

Disease phenotype info (from ClinVar).

in_omim	BOOL	0 : Absence of the variant in OMIM database 1 : Presence of the variant in OMIM database
clinvar_causal_allele	STRING	The allele(s) that are associated or causal for the disease.
clinvar_sig	STRING	The clinical significance scores for each of the variant according to ClinVar: *unknown*, *untested*, *non-pathogenic* *probable-non-pathogenic*, *probable-pathogenic* *pathogenic*, *drug-response*, *histocompatibility* *other*
clinvar_disease_name	STRING	The name of the disease to which the variant is relevant
clinvar_dbsource	STRING	Variant Clinical Channel IDs
clinvar_dbsource_id	STRING	The record id in the above database
clinvar_origin	STRING	The type of variant. Any of: *unknown*, *germline*, *somatic*, *inherited*, *paternal*, *maternal*, *de-novo*, *biparental*, *uniparental*, *not-tested*, *tested-inconclusive*, *other*
clinvar_dsdb	STRING	Variant disease database name
clinvar_dsdbid	STRING	Variant disease database ID
clinvar_disease_acc	STRING	Variant Accession and Versions
clinvar_in_locus_spec_db	BOOL	Submitted from a locus-specific database?
clinvar_on_diag_assay	BOOL	Variation is interrogated in a clinical diagnostic assay?
clinvar_gene_phenotype	STRING	'|' delimited list of phenotypes associated with this gene (includes any variant in the same gene in clinvar not just the current variant).
geno2mp_hpo_ct	INTEGER	Value from geno2mp indicating count of HPO profiles. Set to -1 if missing

Variant Filtering: other categories of fields in the variants table

- Variant and PopGen info
- Optional VCF INFO fields
- Structural variation columns
- Genome annotations
- Variant error assessment
- ENCODE information
- Cancer related columns

Variant Filtering: impact severity explained

Details of the `impact` and `impact_severity` columns

impact severity	impacts	SO_impacts
HIGH	• exon_deleted • frame_shift • splice_acceptor • splice_donor • start_loss • stop_gain • stop_loss • non_synonymous_start • transcript_codon_change • rare_amino_acid • chrom_large_del	• exon_loss_variant • frameshift_variant • splice_acceptor_variant • splice_donor_variant • start_lost • stop_gained • stop_lost • initiator_codon_variant • initiator_codon_variant • rare_amino_acid_variant • chromosomal_deletion
MED	• non_syn_coding • inframe_codon_gain • inframe_codon_loss • inframe_codon_change • codon_change_del • codon_change_ins • UTR_5_del • UTR_3_del • splice_region • mature_miRNA • regulatory_region • TF_binding_site • regulatory_region_ablation • regulatory_region_amplification • TFBS_ablation • TFBS_amplification	• missense_variant • inframe_insertion • inframe_deletion • coding_sequence_variant • disruptive_inframe_deletion • disruptive_inframe_insertion • 5_prime_UTR_truncation + exon_loss_variant • 3_prime_UTR_truncation + exon_loss_variant • splice_region_variant • mature_miRNA_variant • regulatory_region_variant • TF_binding_site_variant • regulatory_region_ablation • regulatory_region_amplification • TFBS_ablation • TFBS_amplification
LOW	• synonymous_stop • synonymous_coding • UTR_5_prime • UTR_3_prime • intron	• stop_retained_variant • synonymous_variant • 5_prime_UTR_variant • 3_prime_UTR_variant • intron_variant

Variant Filtering: loading vcf and ped file into GEMINI

```
gemini load -v my.vcf -p my.ped my.db
```

```
usage: gemini load [-h] [-v VCF] [-t {snpEff,VEP,BCFT,all}] [-p PED_FILE]
                   [--skip-gerp-bp] [--skip-cadd] [--skip-gene-tables]
                   [--save-info-string] [--no-load-genotypes] [--no-genotypes]
                   [--cores CORES] [--scheduler {lsf,sge,slurm,torque}]
                   [--queue QUEUE] [--tempdir TEMPDIR] [--passonly]
                   [--test-mode] [--skip-pls]
                   db
```

`common_args`: common arguments

The inheritance tools share a common set of arguments. We will describe them here and refer to them in the corresponding sections:

`--columns`

This flag is followed by a comma-delimited list of columns the user is requestin in the output.

`--min-kindreds 1`

This is the number of families required to have a variant in the same gene in order for it to be reported. For example, we may only be interested in candidates where at least 4 families have a variant in that gene.

`--families`

By default, candidate variants are reported for all families in the database. One can restrict the analysis to variants in specific familes with the `--families` option. Families should be provided as a comma-separated list

`--filter`

By default, each tool will report all variants regardless of their putative functional impact. In order to apply additional constraints on the variants returned, one can use the `--filter` option. Using SQL syntax, conditions applied with the `--filter` options become WHERE clauses in the query issued to the GEMINI database.

`-d [0]` (depth)

Filter variants that do not have at least this depth for all members in a a family. Default is 0.

Variant Filtering: de novo mutations

Human mutation rate
~ 1.1e-8 / bp / generation

X

Haploid human genome
is 3.1e9 nucleotides

~ 31 de novo mutations
per haploid **genome**

↓

But the exome is only ~ 2% of genome

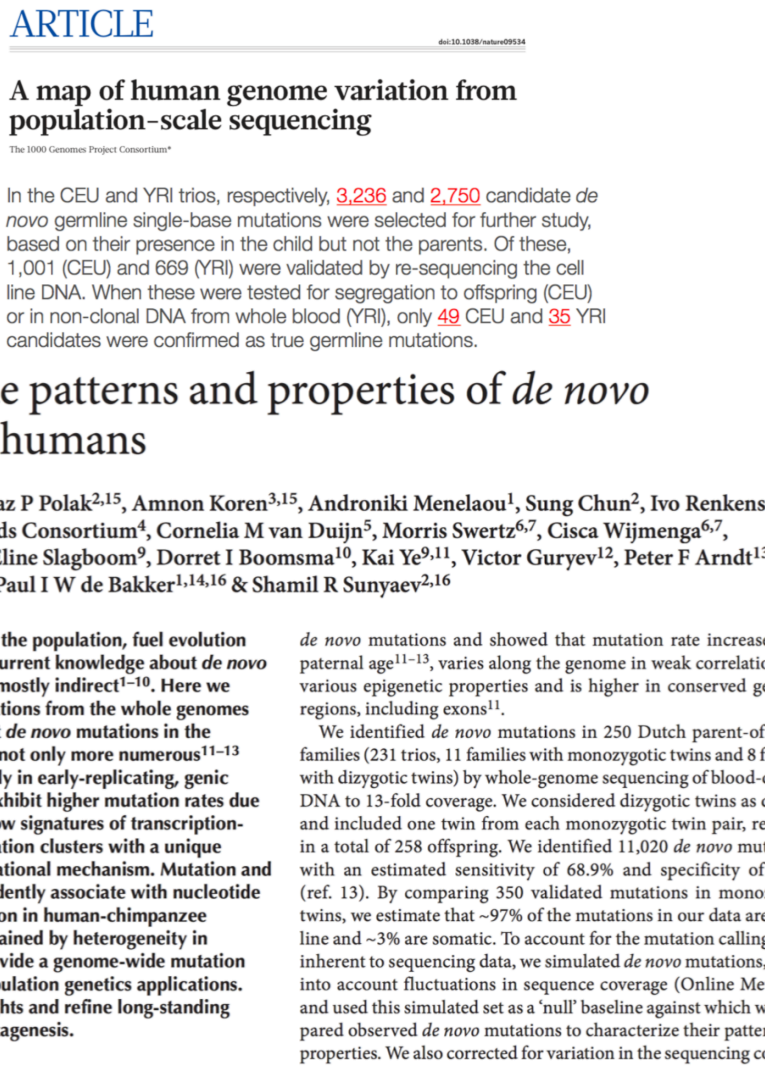
↓

So, ~ 0.62 mutations per haploid exome,
or **0 - 2 mutations per exome**

Why de novo mutations have a high False Positive Rate (FPR)?

- Incorrect genotype assignment
- Low coverage in one or more of the individuals in the family
- Mismapping
- Misalignment
- Paralogy
- Systematic artifacts
- Somatic events

Aaron Quinlan



ARTICLE

doi:10.1038/nature09534

A map of human genome variation from population-scale sequencing

The 1000 Genomes Project Consortium*

In the CEU and YRI trios, respectively, 3,236 and 2,750 candidate *de novo* germline single-base mutations were selected for further study, based on their presence in the child but not the parents. Of these, 1,001 (CEU) and 669 (YRI) were validated by re-sequencing the cell line DNA. When these were tested for segregation to offspring (CEU) or in non-clonal DNA from whole blood (YRI), only 49 CEU and 35 YRI candidates were confirmed as true germline mutations.

Genome-wide patterns and properties of *de novo* mutations in humans

Laurent C Francioli[1,15], Paz P Polak[2,15], Amnon Koren[3,15], Androniki Menelaou[1], Sung Chun[2], Ivo Renkens[1], Genome of the Netherlands Consortium[4], Cornelia M van Duijn[5], Morris Swertz[6,7], Cisca Wijmenga[6,7], Gertjan van Ommen[8], P Eline Slagboom[9], Dorret I Boomsma[10], Kai Ye[9,11], Victor Guryev[12], Peter F Arndt[13], Wigard P Kloosterman[1], Paul I W de Bakker[1,14,16] & Shamil R Sunyaev[2,16]

Mutations create variation in the population, fuel evolution and cause genetic diseases. Current knowledge about *de novo* mutations is incomplete and mostly indirect[1–10]. Here we analyze 11,020 *de novo* mutations from the whole genomes of 250 families. We show that *de novo* mutations in the offspring of older fathers are not only more numerous[11–13] but also occur more frequently in early-replicating, genic regions. Functional regions exhibit higher mutation rates due to CpG dinucleotides and show signatures of transcription-coupled repair, whereas mutation clusters with a unique signature point to a new mutational mechanism. Mutation and recombination rates independently associate with nucleotide diversity, and regional variation in human-chimpanzee divergence is only partly explained by heterogeneity in mutation rate. Finally, we provide a genome-wide mutation rate map for medical and population genetics applications. Our results provide new insights and refine long-standing hypotheses about human mutagenesis.

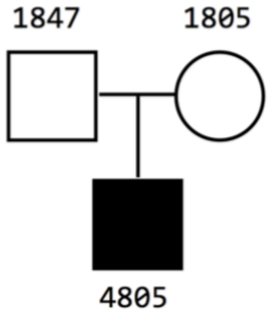
de novo mutations and showed that mutation rate increases with paternal age[11–13], varies along the genome in weak correlation with various epigenetic properties and is higher in conserved genomic regions, including exons[11].

We identified *de novo* mutations in 250 Dutch parent-offspring families (231 trios, 11 families with monozygotic twins and 8 families with dizygotic twins) by whole-genome sequencing of blood-derived DNA to 13-fold coverage. We considered dizygotic twins as distinct and included one twin from each monozygotic twin pair, resulting in a total of 258 offspring. We identified 11,020 *de novo* mutations, with an estimated sensitivity of 68.9% and specificity of 94.6% (ref. 13). By comparing 350 validated mutations in monozygotic twins, we estimate that ~97% of the mutations in our data are germline and ~3% are somatic. To account for the mutation calling biases inherent to sequencing data, we simulated *de novo* mutations, taking into account fluctuations in sequence coverage (Online Methods), and used this simulated set as a 'null' baseline against which we compared observed *de novo* mutations to characterize their patterns and properties. We also corrected for variation in the sequencing coverage

Variant Filtering: de_novo mode of GEMINI

Genotype Requirements

- all affecteds must be het
- [affected] all unaffected must be homref or homalt
- at least 1 affected kid must have unaffected parents
- [strict] if an affected has affected parents, it's not de_novo
- [strict] all affected kids must have unaffected (or no) parents
- [strict] warning if none of the affected samples have parents.



Requires a PED file

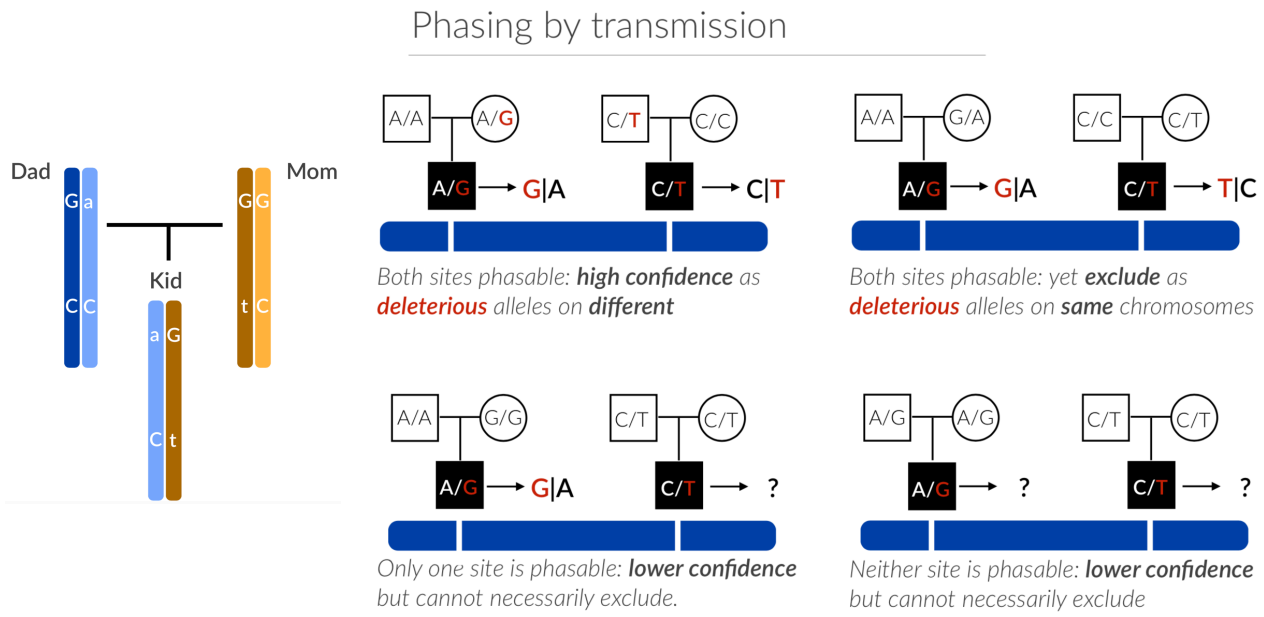
#family_id	sample_id	paternal_id	maternal_id	sex	phenotype
family1	1805	-9	-9	2	1
family1	1847	-9	-9	1	1
family1	4805	1847	1805	1	2

example

if we wanted to restrict candidate variants to solely those with a HIGH predicted functional consequence, we could use the following:

```
$ gemini de_novo \
      --columns "chrom, start, end, ref, alt" \
      --filter "impact_severity = 'HIGH'" \
      test.de_novo.db
chrom       start   end     ref     alt     variant_id      family_id       family_members  fa
chr10       1142207 1142208 T       C       1       1       1_dad(dad;unaffected),1_mom(mom;un
```

Variant Filtering: compound heterozygous mutations



Aaron Quinlan

* Convention for phased genotype is maternal allele first

mom	dad	kid	phaseable	priority	notes
R-H	H-R	H-H	both	1	both sites phaseable and alts on opposite chroms
n/a	n/a	H-H	NO	2	singleton (unphaseable) HETs have priority 2.
R-H	H-H	H-H	one	3	should be a rare occurrence
H-H	H-H	H-H	NO	3	should be a rare occurrence
H-H	UNK	H-H	NO	3	missing parent and all hets.
A-R	H-H	H-H	both	NA	exclude hom-alts from un-affecteds
R-R	H-H	H-H	both	NA	phaseable, but alts are on the same chroms.

Variant Filtering: comp_hets mode of GEMINI

Genotype Requirements

- All affected individuals must be heterozygous at both sites.
- No unaffected can be homozygous alternate at either site.
- Neither parent of an affected sample can be homozygous reference at both sites.
- If any unphased-unaffected is het at both sites, the site will be give lower priority
- No phased-unaffected can be heterozygous at both sites.
  a. *`--allow-unaffected`* keeps sites where a phased unaffected shares the het-pair
  b. unphased, unaffected that share the het pair are counted and reported for each candidate pair.
- Remove candidates where an affected from the same family does NOT share the same het pair.
- Sites are automatically phased by transmission when parents are present in order to remove false positive candidates.
  a. If data from one or both parents are unavailable and the child's data was not phased prior to loading into GEMINI, all comp_het variant pairs will automatically be given at most priority == 2. If there's only a single parent and both the parent and the affected are HET at both sites, the candidate will have priority 3.
  b. *`--max-priority x`* can be used to set the maximum allowed priority level at which candidate pairs are included in the output

```
$ gemini comp_hets -d 50 \
      --columns "chrom, start, end, ref, alt" \
      --filter "impact_severity = 'HIGH'" \
      --min-kindreds 2 \
      my.db
```

Variant Filtering: autosomal_recessive and x_linked_recessive modes of GEMINI

autosomal_recessive

Genotype Requirements

- all affecteds must be hom_alt
- [affected] no unaffected can be hom_alt (can be unknown)
- [strict] if parents exist they must be unaffected and het for all affected kids
- [strict] if there are no affecteds that have a parent, a warning is issued.

x_linked_recessive

Genotype Requirements

- Affected females must be HOM_ALT
- Unaffected females are HET or HOM_REF
- Affected males are not HOM_REF
- Unaffected males are HOM_REF

Variant Filtering: set_somatic mode of GEMINI

`set_somatic`: Flag somatic variants

Somatic mutations in a tumor-normal pair are variants that are present in the tumor but not in the normal sample.

Note

1. This tool requires that you specify the sample layout via a PED file when loading your VCF into GEMINI via:

```
gemini load -v my.vcf -p my.ped my.db
```

Example PED file format for GEMINI

```
#Family_ID      Individual_ID   Paternal_ID     Maternal_ID     Sex     Phenotype       Ethnic
1       Normal  -9      -9      0       1       -9
1       Tumor   -9      -9      0       2       -9
```

default behavior

By default, `set_somatic` simply marks variants that are genotyped as homozygous reference in the normal sample and non-reference in the tumor. More stringent somatic filtering criteria are available through tunable command line parameters.

```
$ gemini set_somatic \
    --min-depth 30 \
    --min-qual 20 \
    --min-somatic-score 18 \
    --min-tumor-depth 10 \
    --min-norm-depth 10 \
    tumor_normal.db
tum_name        tum_gt  tum_alt_freq    tum_alt_depth   tum_depth       nrm_name        nrm_gt
tumor   GAAAAAAAAAAAAAGGTGAAAATT/GAAAAAAAAAAAAAGGTGAAAATT       0.217391304348  5       23
tumor   CTGCTATTTTG/CG  0.22    11      50      normal  CTGCTATTTTG/CTGCTATTTTG 0.0     0
tumor   C/A     0.555555555556  10      18      normal  C/C     0.0     0       17      chr17
tumor   C/T     0.1875  12      64      normal  C/C     0.0     0       30      chr2    128046
Identified and set 4 somatic mutations
```

Variant Filtering: set_somatic mode options

`--min-depth [None]`

The minimum required combined depth for tumor and normal samples.

`--min-qual [None]`

The minimum required variant quality score.

`--min-somatic-score [None]`

The minimum required somatic score (SSC). This score is produced by various somatic variant detection algorithms including SpeedSeq, SomaticSniper, and VarScan 2.

`--max-norm-alt-freq [None]`

The maximum frequency of the alternate allele allowed in the normal sample.

`--max-norm-alt-count [None]`

The maximum count of the alternate allele allowed in the normal sample.

`--min-norm-depth [None]`

The minimum depth required in the normal sample.

`--min-tumor-alt-freq [None]`

The minimum frequency of the alternate allele required in the tumor sample.

`--min-tumor-alt-count [None]`

The minimum count of the alternate allele required in the tumor sample.

`--min-tumor-depth [None]`

The minimum depth required in the tumor sample.

Result interpretation

- Supporting evidence
- False positives
- False negatives
- Complications

Result interpretation: Supporting evidence - ACMG guidelines

ACMG STANDARDS AND GUIDELINES

	Benign		Pathogenic			
	Strong	**Supporting**	**Supporting**	**Moderate**	**Strong**	**Very strong**
Population data	MAF is too high for disorder BA1/BS1 **OR** observation in controls inconsistent with disease penetrance BS2			Absent in population databases PM2	Prevalence in affecteds statistically increased over controls PS4	
Computational and predictive data		Multiple lines of computational evidence suggest no impact on gene /gene product BP4 Missense in gene where only truncating cause disease BP1 Silent variant with non predicted splice impact BP7 In-frame indels in repeat w/out known function BP3	Multiple lines of computational evidence support a deleterious effect on the gene /gene product PP3	Novel missense change at an amino acid residue where a different pathogenic missense change has been seen before PM5 Protein length changing variant PM4	Same amino acid change as an established pathogenic variant PS1	Predicted null variant in a gene where LOF is a known mechanism of disease PVS1
Functional data	Well-established functional studies show no deleterious effect BS3		Missense in gene with low rate of benign missense variants and path. missenses common PP2	Mutational hot spot or well-studied functional domain without benign variation PM1	Well-established functional studies show a deleterious effect PS3	
Segregation data	Nonsegregation with disease BS4		Cosegregation with disease in multiple affected family members PP1	Increased segregation data →		
De novo data				De novo (without paternity & maternity confirmed) PM6	De novo (paternity and maternity confirmed) PS2	
Allelic data		Observed in *trans* with a dominant variant BP2 Observed in *cis* with a pathogenic variant BP2		For recessive disorders, detected in trans with a pathogenic variant PM3		
Other database		Reputable source w/out shared data = benign BP6	Reputable source = pathogenic PP5			
Other data		Found in case with an alternate cause BP5	Patient's phenotype or FH highly specific for gene PP4			

Figure 1 Evidence framework. This chart organizes each of the criteria by the type of evidence as well as the strength of the criteria for a benign (left side) or pathogenic (right side) assertion. Evidence code descriptions can be found in **Tables 3** and **4**. BS, benign strong; BP, benign supporting; FH, family history; LOF, loss of function; MAF, minor allele frequency; path., pathogenic; PM, pathogenic moderate; PP, pathogenic supporting; PS, pathogenic strong; PVS, pathogenic very strong.

https://www.acmg.net/docs/standards_guidelines_for_the_interpretation_of_sequence_variants.pdf

Result interpretation: Supporting evidence - ACMG guidelines

ACMG STANDARDS AND GUIDELINES

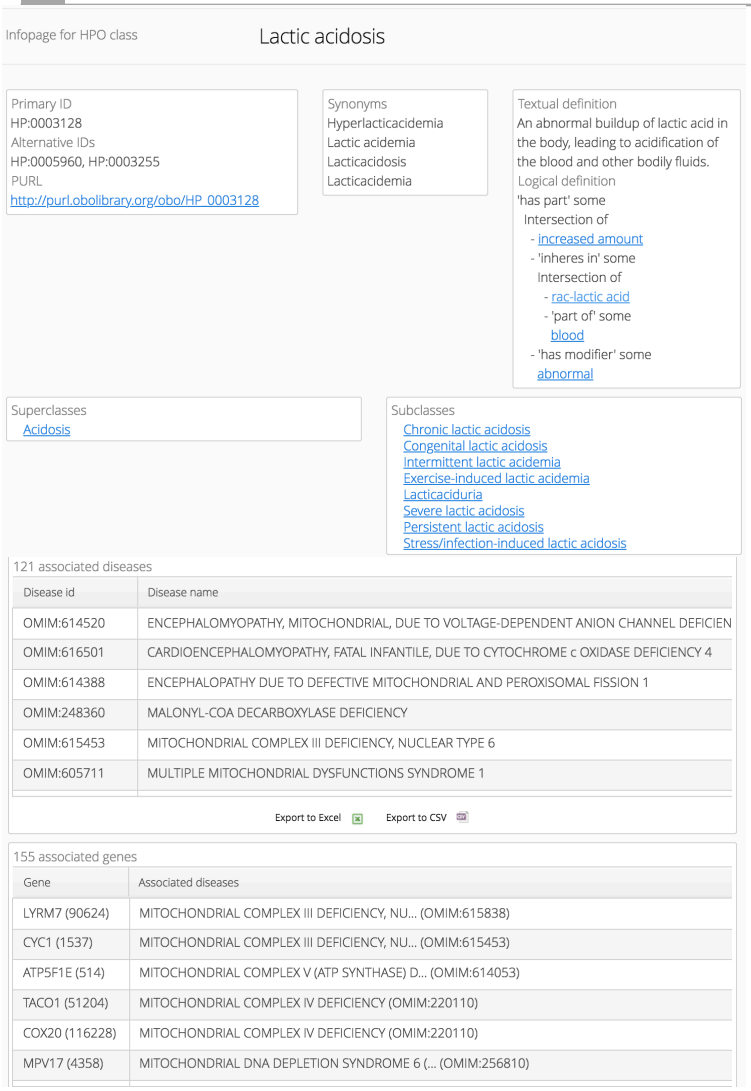
Table 3 Criteria for classifying pathogenic variants

Evidence of pathogenicity	Category
Very strong	PVS1 null variant (nonsense, frameshift, canonical ±1 or 2 splice sites, initiation codon, single or multiexon deletion) in a gene where LOF is a known mechanism of disease Caveats: • Beware of genes where LOF is not a known disease mechanism (e.g., *GFAP*, *MYH7*) • Use caution interpreting LOF variants at the extreme 3′ end of a gene • Use caution with splice variants that are predicted to lead to exon skipping but leave the remainder of the protein intact • Use caution in the presence of multiple transcripts
Strong	PS1 Same amino acid change as a previously established pathogenic variant regardless of nucleotide change Example: Val→Leu caused by either G>C or G>T in the same codon Caveat: Beware of changes that impact splicing rather than at the amino acid/protein level PS2 De novo (both maternity and paternity confirmed) in a patient with the disease and no family history Note: Confirmation of paternity only is insufficient. Egg donation, surrogate motherhood, errors in embryo transfer, and so on, can contribute to nonmaternity. PS3 Well-established in vitro or in vivo functional studies supportive of a damaging effect on the gene or gene product Note: Functional studies that have been validated and shown to be reproducible and robust in a clinical diagnostic laboratory setting are considered the most well established. PS4 The prevalence of the variant in affected individuals is significantly increased compared with the prevalence in controls Note 1: Relative risk or OR, as obtained from case–control studies, is >5.0, and the confidence interval around the estimate of relative risk or OR does not include 1.0. See the article for detailed guidance. Note 2: In instances of very rare variants where case–control studies may not reach statistical significance, the prior observation of the variant in multiple unrelated patients with the same phenotype, and its absence in controls, may be used as moderate level of evidence.
Moderate	PM1 Located in a mutational hot spot and/or critical and well-established functional domain (e.g., active site of an enzyme) without benign variation PM2 Absent from controls (or at extremely low frequency if recessive) (**Table 6**) in Exome Sequencing Project, 1000 Genomes Project, or Exome Aggregation Consortium

Table 4 Criteria for classifying benign variants

Evidence of benign impact	Category
Stand-alone	BA1 Allele frequency is >5% in Exome Sequencing Project, 1000 Genomes Project, or Exome Aggregation Consortium
Strong	BS1 Allele frequency is greater than expected for disorder (see **Table 6**) BS2 Observed in a healthy adult individual for a recessive (homozygous), dominant (heterozygous), or X-linked (hemizygous) disorder, with full penetrance expected at an early age BS3 Well-established in vitro or in vivo functional studies show no damaging effect on protein function or splicing BS4 Lack of segregation in affected members of a family Caveat: The presence of phenocopies for common phenotypes (i.e., cancer, epilepsy) can mimic lack of segregation among affected individuals. Also, families may have more than one pathogenic variant contributing to an autosomal dominant disorder, further confounding an apparent lack of segregation.
Supporting	BP1 Missense variant in a gene for which primarily truncating variants are known to cause disease BP2 Observed in *trans* with a pathogenic variant for a fully penetrant dominant gene/disorder or observed in *cis* with a pathogenic variant in any inheritance pattern BP3 In-frame deletions/insertions in a repetitive region without a known function BP4 Multiple lines of computational evidence suggest no impact on gene or gene product (conservation, evolutionary, splicing impact, etc.) Caveat: Because many in silico algorithms use the same or very similar input for their predictions, each algorithm cannot be counted as an independent criterion. BP4 can be used only once in any evaluation of a variant. BP5 Variant found in a case with an alternate molecular basis for disease BP6 Reputable source recently reports variant as benign, but the evidence is not available to the laboratory to perform an independent evaluation BP7 A synonymous (silent) variant for which splicing prediction algorithms predict no impact to the splice consensus sequence nor the creation of a new splice site AND the nucleotide is not highly conserved

Result interpretation: gene-phenotype association



Infopage for HPO class

Lactic acidosis

Primary ID
HP:0003128
Alternative IDs
HP:0005960, HP:0003255
PURL
http://purl.obolibrary.org/obo/HP_0003128

Synonyms
Hyperlacticacidemia
Lactic acidemia
Lacticacidosis
Lacticacidemia

Textual definition
An abnormal buildup of lactic acid in the body, leading to acidification of the blood and other bodily fluids.
Logical definition
'has part' some
Intersection of
- increased amount
- 'inheres in' some
Intersection of
- rac-lactic acid
- 'part of' some
blood
- 'has modifier' some
abnormal

Superclasses
Acidosis

Subclasses
Chronic lactic acidosis
Congenital lactic acidosis
Intermittent lactic acidemia
Exercise-induced lactic acidemia
Lacticaciduria
Severe lactic acidosis
Persistent lactic acidosis
Stress/infection-induced lactic acidosis

121 associated diseases

Disease id	Disease name
OMIM:614520	ENCEPHALOMYOPATHY, MITOCHONDRIAL, DUE TO VOLTAGE-DEPENDENT ANION CHANNEL DEFICIEN
OMIM:616501	CARDIOENCEPHALOMYOPATHY, FATAL INFANTILE, DUE TO CYTOCHROME c OXIDASE DEFICIENCY 4
OMIM:614388	ENCEPHALOPATHY DUE TO DEFECTIVE MITOCHONDRIAL AND PEROXISOMAL FISSION 1
OMIM:248360	MALONYL-COA DECARBOXYLASE DEFICIENCY
OMIM:615453	MITOCHONDRIAL COMPLEX III DEFICIENCY, NUCLEAR TYPE 6
OMIM:605711	MULTIPLE MITOCHONDRIAL DYSFUNCTIONS SYNDROME 1

Export to Excel Export to CSV

155 associated genes

Gene	Associated diseases
LYRM7 (90624)	MITOCHONDRIAL COMPLEX III DEFICIENCY, NU... (OMIM:615838)
CYC1 (1537)	MITOCHONDRIAL COMPLEX III DEFICIENCY, NU... (OMIM:615453)
ATP5F1E (514)	MITOCHONDRIAL COMPLEX V (ATP SYNTHASE) D... (OMIM:614053)
TACO1 (51204)	MITOCHONDRIAL COMPLEX IV DEFICIENCY (OMIM:220110)
COX20 (116228)	MITOCHONDRIAL COMPLEX IV DEFICIENCY (OMIM:220110)
MPV17 (4358)	MITOCHONDRIAL DNA DEPLETION SYNDROME 6 (... (OMIM:256810)

Class Hierarchy

Thing
+ Phenotype
+ abnormal phenotype
+ Phenotypic abnormality
+ Abnormality of metabolism/homeostasis
+ Abnormality of acid-base homeostasis
+ Acidosis
+ Metabolic acidosis
+ Renal tubular acidosis
+ Ketoacidosis
+ Hyperchloremic acidosis
+ Increased serum lactate
- Methylmalonic acidemia
- Oroticaciduria
- Glutaric acidemia
- Propionicacidemia
- Phenylpyruvic acidemia
+ Respiratory acidosis
+ Chronic acidosis
+ Dicarboxylic acidemia
- Lactic acidosis
- Lacticaciduria
- Stress/infection-induced lactic acidosis
- Persistent lactic acidosis
- Severe lactic acidosis
- Exercise-induced lactic acidemia
- Congenital lactic acidosis
- Intermittent lactic acidemia
- Chronic lactic acidosis

Human phenotype ontology (HPO)

HPO browser from http://human-phenotype-ontology.github.io/

Result interpretation: Other Supporting evidence

- Fits into observed inheritance model
- Rare in any population
- Significant pathogenicity predictions
- Phenotypes observed in model organisms

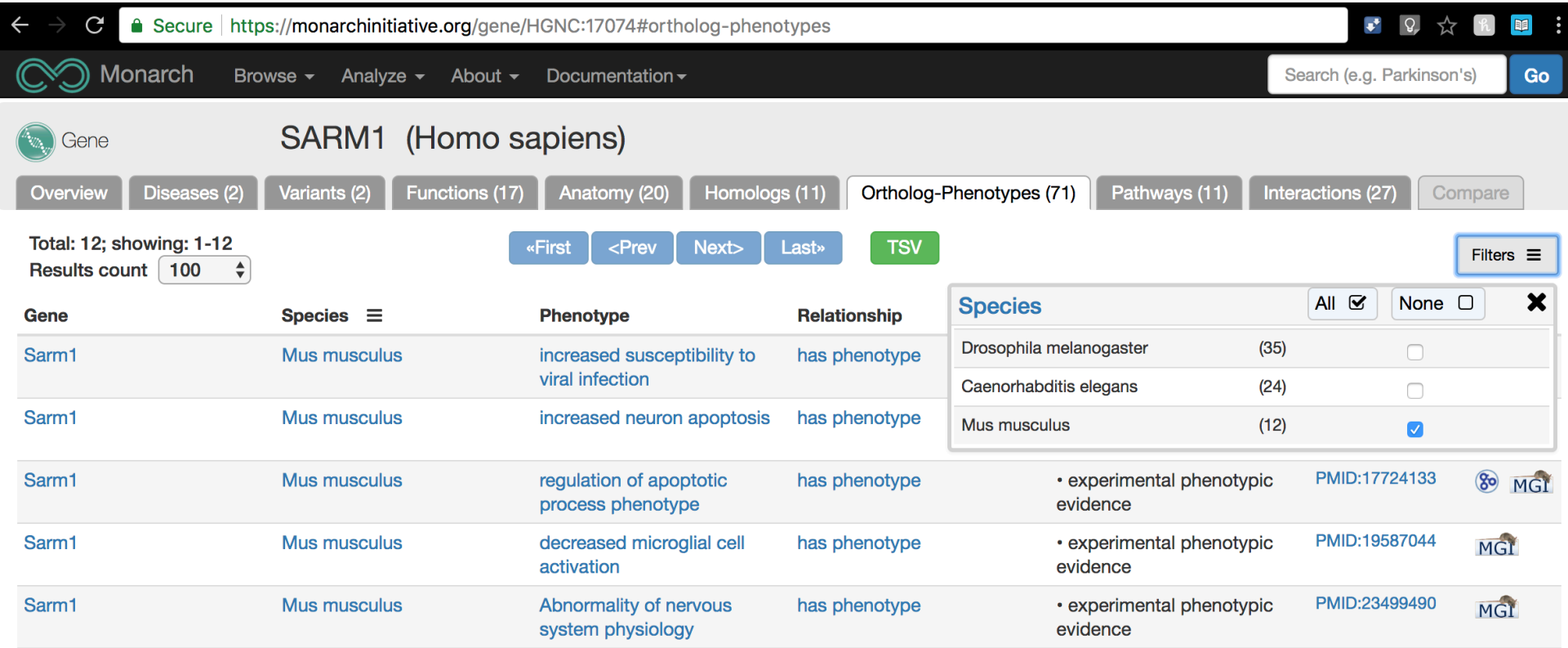
(Monarch Initiative - http://monarchinitiative.org)

Gene-based predictions

- ClinGen haploinsufficency score https://www.ncbi.nlm.nih.gov/projects/dbvar/clingen/

- ExAC Probability of loss of function intolerance (pLI)

Monarch Initiative - http://monarchinitiative.org



Result interpretation: False Positives

- Large genes
  - TTN
  - USH2A
  - MUC16
  - FLG
- Lots of paralogs/part of gene family

Don't rule out if phenotype makes sense!

- *TTN* mutations cause dilated cardiomyopathy and muscular dystrophy
- *MUC1* mutations cause medullary cystic kidney disease
- *KRT** gene mutations cause ichthyosis, keratoderma, keratosis

Jessica Chong

Count	String	Description
91	LOC	LOC genes
22	ENS	Ensembl genes
21	FAM	FAM proteins
15	GOL	Golgi-like GOLGA8E
13	PRA	PRAMEF genes
9	NBP	Nuclear breakpoint family
7	POT	POTE ankyrin domain family
6	DEF	defensins
5	OR2	Olfactory receptor
5	MUC	Mucins
5	KRT	Keratins
4	WAS	WAS protein family homolog
4	ANK	ankyrins
3	TRI	tri-partite motif containing
3	OR1	Olfactory receptor
3	FRG	FSHD region gene

Pseudogene database: http://pseudofam.pseudogene.org
http://massgenomics.org/2013/06/ngs-false-positives.html

Result interpretation: False Negatives

Loosen the cut-off to salvage those that were called but filtered out

- Population frequency
- Mutation impact
- Predicted pathogenicity scores
- Variant caller filters, e.g. GATK VQSR
- Read depth
- Genotypes

Improve method to pick up those that were not detected

- A different caller
- Deeper coverage
- WGS instead of WES

Result interpretation: Complications

- Same apparent condition, different underlying genes
- Same condition, same gene, different inheritance patterns
- Different conditions, same gene, same inheritance pattern
- Different conditions, same gene, different inheritance patterns
- Compound inheritance:
  - SNV+CNV
  - SNV+indel
- Mosaicism:
  - Tissue from the somatic mosaic parent contains the mutation
  - Tissue from the somatic mosaic proband does not contain mutation
- Non-coding
  - Synonymous mutations that actually affect splicing
  - Intronic mutations that activates a pseudoexon
  - Mutations affecting lncRNA, enhancers, etc
- Complex Events, e.g. repeat expansions
- Polygenic conditions
- Uniparental disomy
- Imprinting
- Environmental rather than genetic

Jessica Chong

Acknowledgement

Some of the slides in this presentation are adapted from UW CMG workshop in 2016
http://uwcmg.org/#/analysisWorkshop