Identifying high-confidence de novo mutations with GEMINI

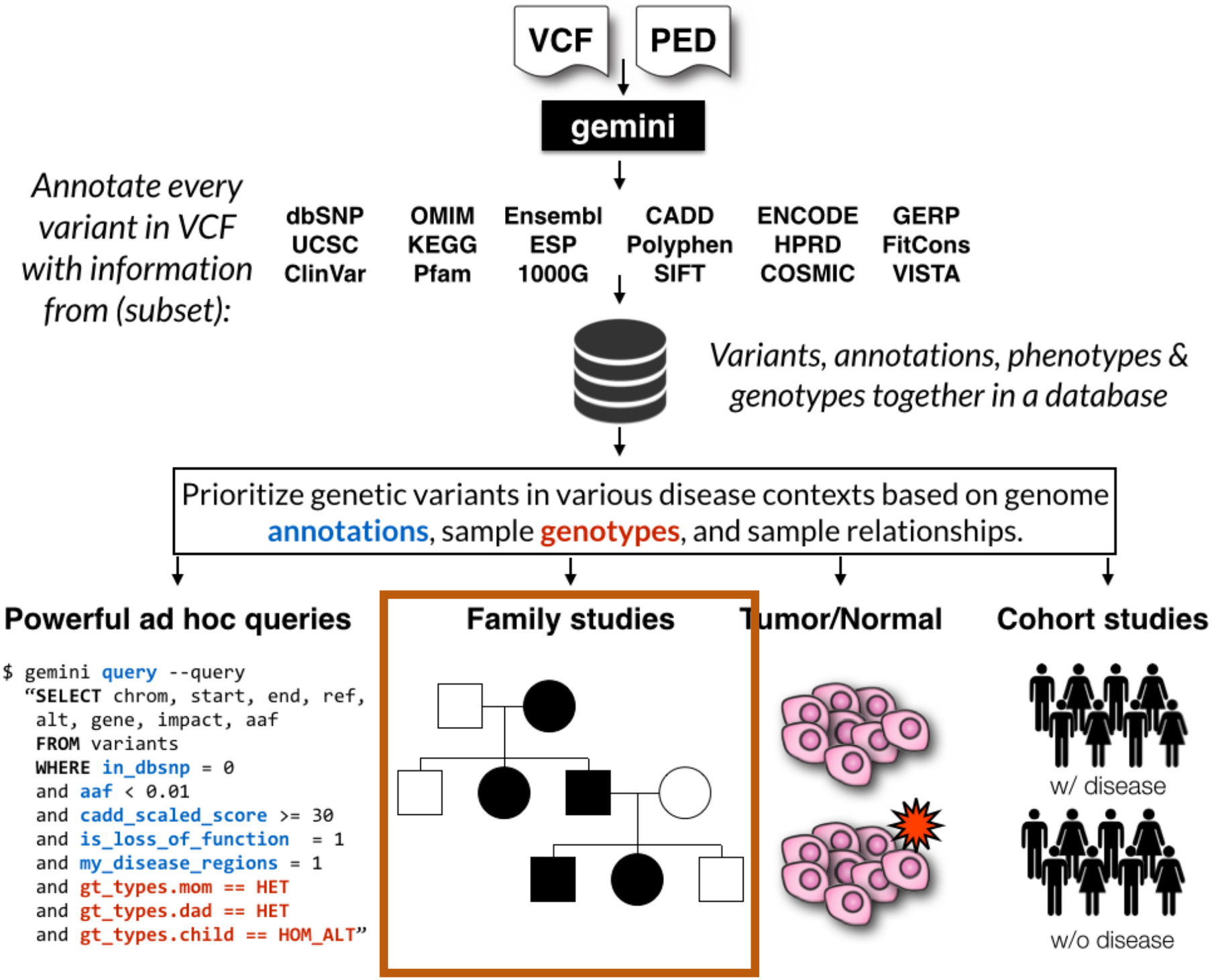
Aaron Quinlan
University of Utah

quinlanlab.org

Commands for this session:

https://gist.github.com/arq5x/9e1928638397ba45da2e#file-denovo-sh

Automated tools for disease inheritance models



Automated tools for disease inheritance models

- 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.
  - **`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
  - **`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

Common options for disease model tools.

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)

Unfortunately, spurious inherited variants can often appear due to insufficient sequence coverage One simple way to filter such artifacts is to enforce a minimum sequence depth (default: 0) for each sample. We can do that with this flag.

`--allow-unaffected`

By default, candidates that also appear in unaffected samples are not reported if this flag is specified, such variants will be reported.

`--lenient`

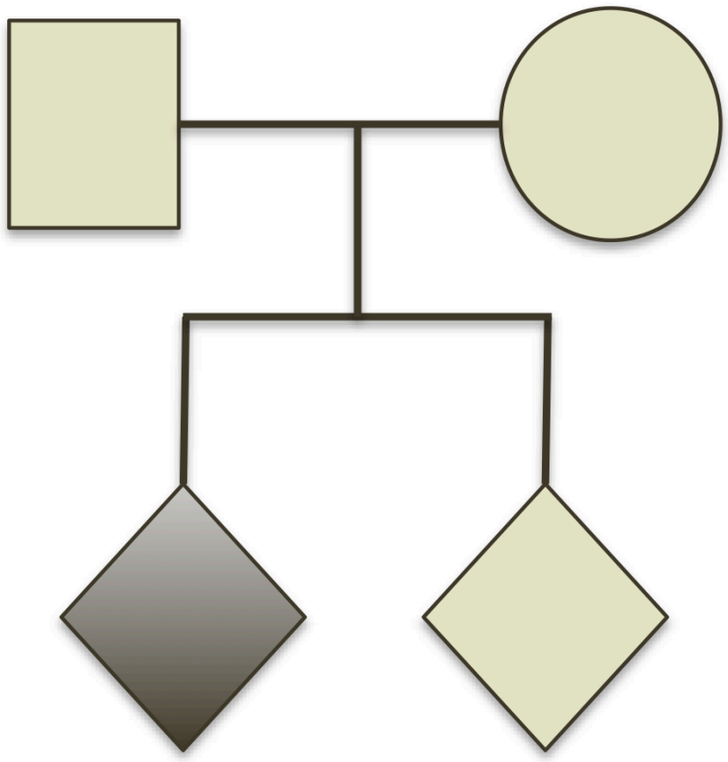
Loosen the restrictions on family structure. This will allow, for example, finding compound_hets in unaffected samples.

`--gt-pl-max`

In order to eliminate less confident genotypes, it is possible to enforce a maximum PL value for each sample. On this scale, lower values indicate more confidence that the called genotype is correct. 10 is a reasonable value:

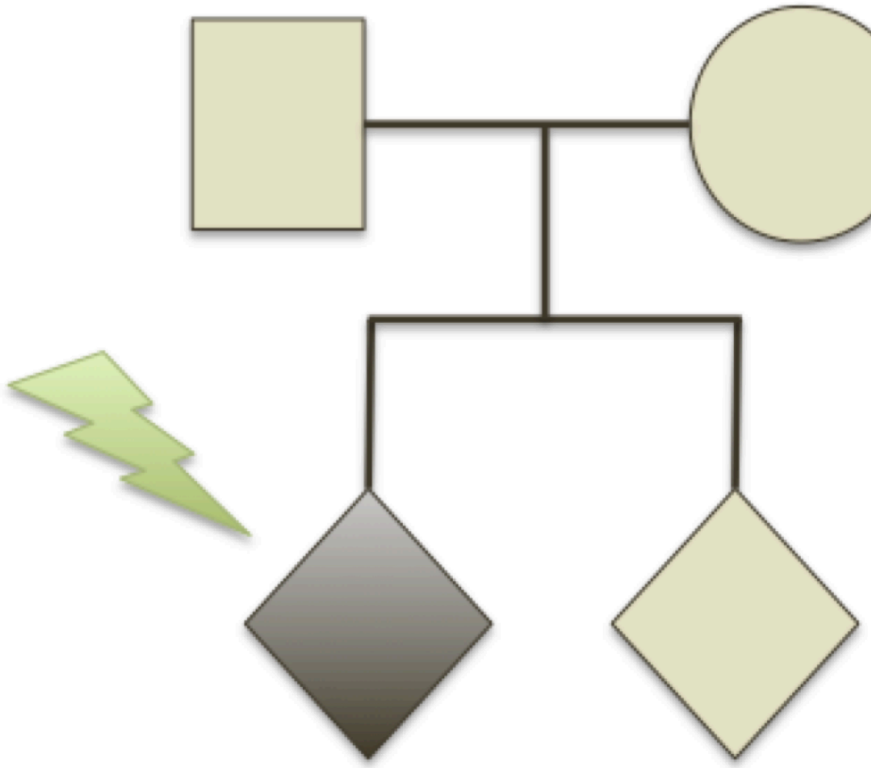
Why search for de novo mutations?

- Unsolved disorder
- Sporadic-Little to no family recurrence
- Appears genetic
- Severe (likely to reduce fitness)
- Heterogeneity?

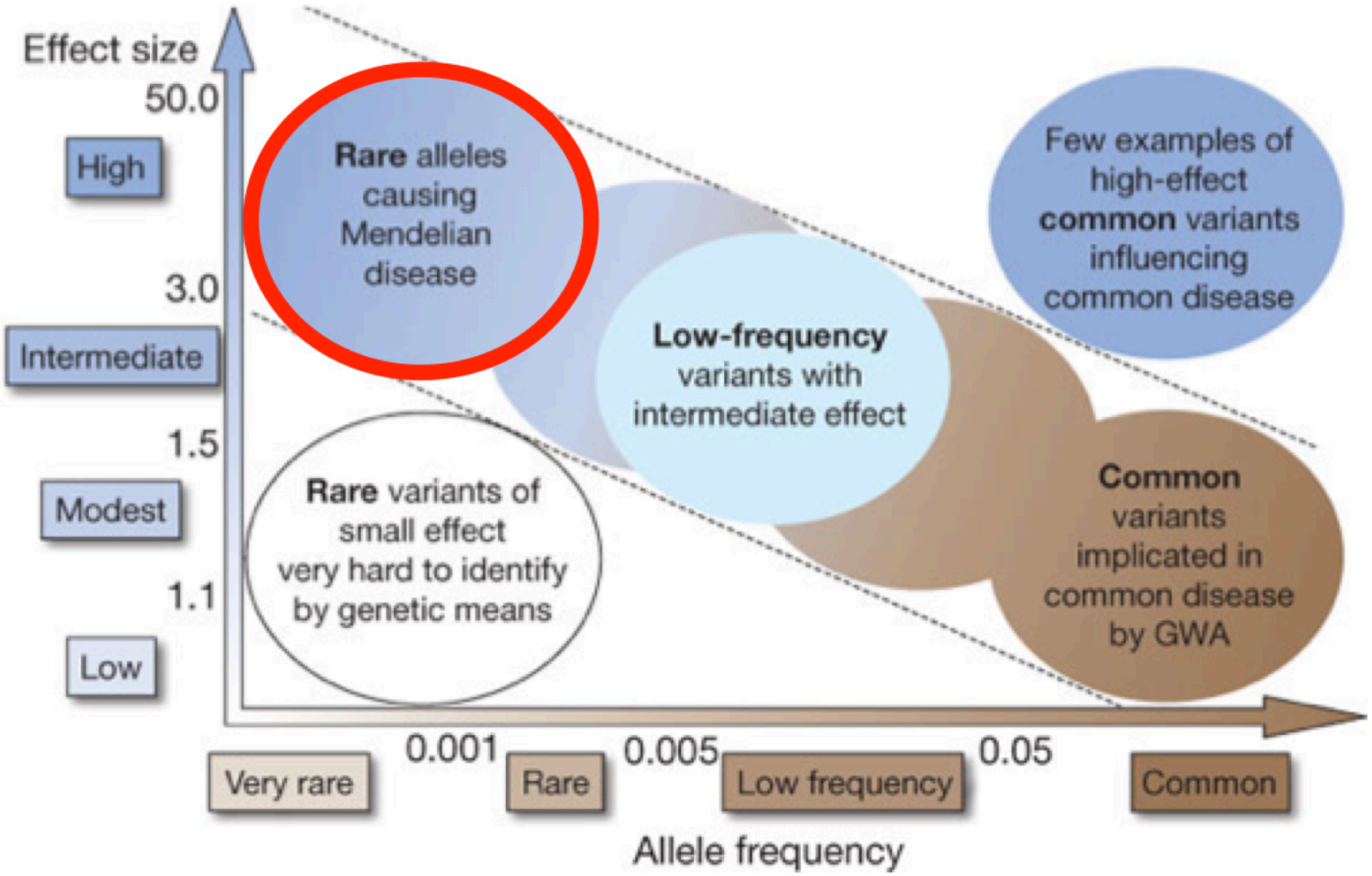


Brian O'Roak

High impact variants



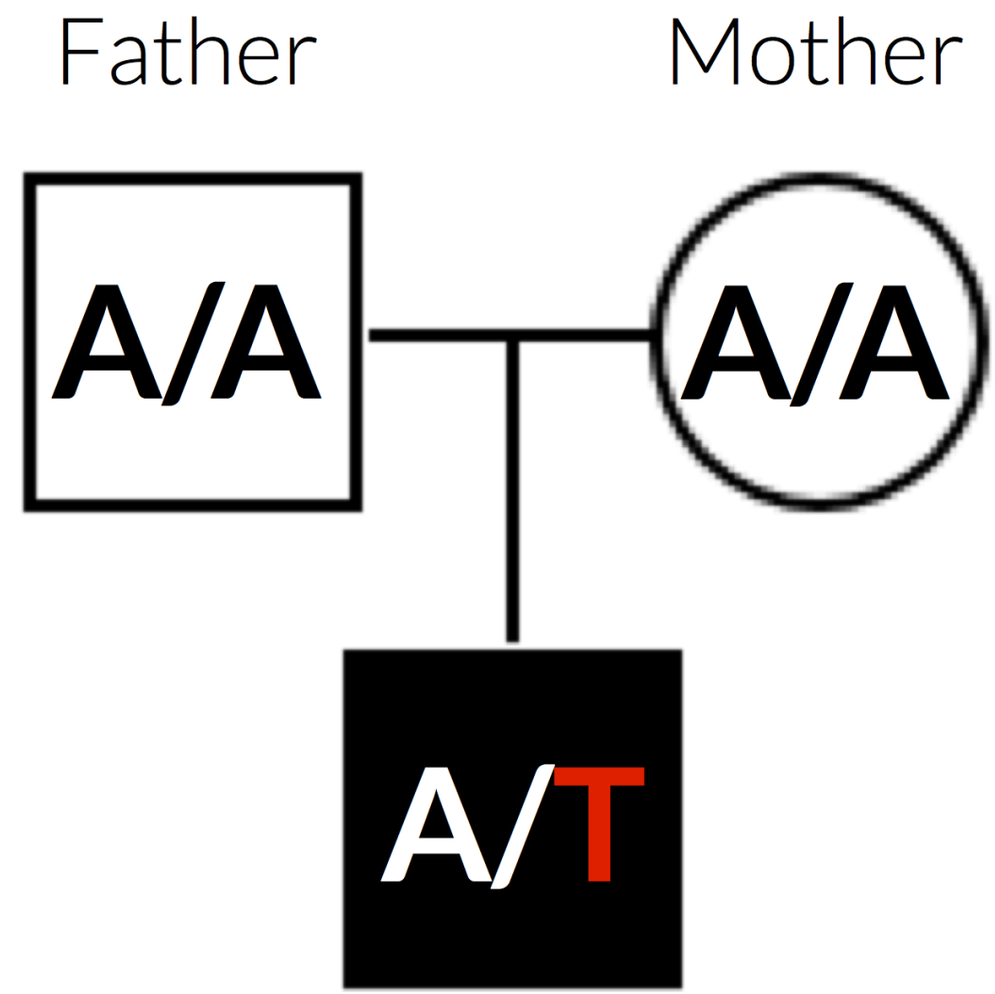
de novo
“new”
point mutations



Manolio *et al.*, 2009 nature

Brian O’Roak

De novo mutations



Affected proband

How many de novo mutations should we expect?

De novo (spontaneous) 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**

De novo (spontaneous) 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**

In practice, it's not so simple.

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.

Brian O'Roak

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

Why are there so many artifacts?

- **Prior probabilities** - the more interesting something is, the less likely it is to be real

- **If something can go wrong, it will.**
  - Incorrect genotype assignment
  - Low coverage in one or more of the individuals in the family (especially the parents…**why**?)
  - Mismapping
  - Misalignment
  - Paralogy
  - Systematic artifacts
  - Somatic events

Brian O'Roak

Detective work with GEMINI

The **de_novo** tool in GEMINI

de_novo: Identifying potential de novo mutations.

Note

1. This tool requires that you identify familial relationships via a PED file when loading your VCF into gemini via:

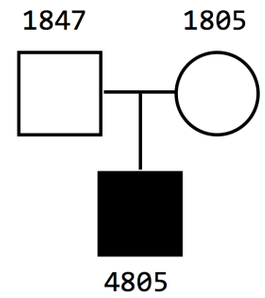
```
gemini load -v my.vcf -p my.ped my.db
```

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




```
$ gemini de_novo --columns "chrom,start,end" test.de_novo.db
chrom       start   end     variant_id      family_id       family_members  family_genotypes        samples family_count
chr10       1142207 1142208 1       1       1_dad(dad;unaffected),1_mom(mom;unaffected),1_kid(child;affected)       T/T,T/T,T/C     1_kid   1
chr10       48003991        48003992        2       2       2_dad(dad;unaffected),2_mom(mom;unaffected),2_kid(child;affected)       C/C,C/C,C/T     2_kid   1
chr10       48004991        48004992        3       3       3_dad(dad;unaffected),3_mom(mom;unaffected),3_kid(child;affected)       C/C,C/C,C/T     3_kid   1
chr10       135336655       135336656       4       4       1_dad(dad;unaffected),1_mom(mom;unaffected),1_kid(child;affected)       G/G,G/G,G/A     1_kid   2
chr10       135336655       135336656       4       4       2_dad(dad;unaffected),2_mom(mom;unaffected),2_kid(child;affected)       G/G,G/G,G/A     2_kid   2
chr10       135369531       135369532       5       5       1_dad(dad;unaffected),1_mom(mom;unaffected),1_kid(child;affected)       T/T,T/T,T/C     1_kid   3
chr10       135369531       135369532       5       5       2_dad(dad;unaffected),2_mom(mom;unaffected),2_kid(child;affected)       T/T,T/T,T/C     2_kid   3
chr10       135369531       135369532       5       5       3_dad(dad;unaffected),3_mom(mom;unaffected),3_kid(child;affected)       T/T,T/T,T/C     3_kid   3
```

Note

The output will always start with the the requested columns followed by the 5 columns enumerated at the start of this document.

http://gemini.readthedocs.org/en/latest/content/tools.html#de-novo-identifying-potential-de-novo-mutations

Login

```
$ ssh -YC USERNAME@cmghead.gs.washington.edu
$ qrsh
$ cd wed/mydata
```

Create a GEMINI database from a VCF

Notes:

1. The VCF has been normalized and decomposed with VT
2. The VCF has been annotated with VEP.

```
curl https://s3.amazonaws.com/gemini-tutorials/trio.trim.vep.vcf.gz > trio.trim.vep.vcf.gz
curl https://s3.amazonaws.com/gemini-tutorials/denovo.ped > denovo.ped
gemini load --cores 2 \
            -v  trio.trim.vep.vcf.gz \
            -t VEP \
            --tempdir . \
            --skip-gene-tables --skip-cadd --skip-gerp-bp \
            -p denovo.ped \
       trio.trim.vep.denovo.db
```

Note: copy and paste the full command from the Github Gist to avoid errors

~4 minutes

While we are waiting - what are normalization and decomposition?

Variant normalization

Reference and alternative alleles of a multi nucleotide polymorphism (MNP)

```
REF    GGGCATGGG
ALT    GGGTGCGGG
```

Genome Reference		Variant Call Format			
	GGGGCATGGGG	POS	REF	ALT	
REF	GCAT	4	GCAT	GTGC	Not left trimmed
ALT	GTGC				
REF	CATG	5	CATG	TGCG	Not right trimmed
ALT	TGCG				
REF	GCATG	4	GCATG	GTGCG	Not left and right trimmed
ALT	GTGCG				
REF	CAT	5	CAT	TGC	Normalized
ALT	TGC				

Alleles represented against the human genome reference. Allele pairs are colored the same, all are representations of the same variant.

Alleles represented in Variant Call Format, all are representations of the same variant.

http://genome.sph.umich.edu/wiki/File:Normalization_mnp.png

Why is this important?

Variant decomposition

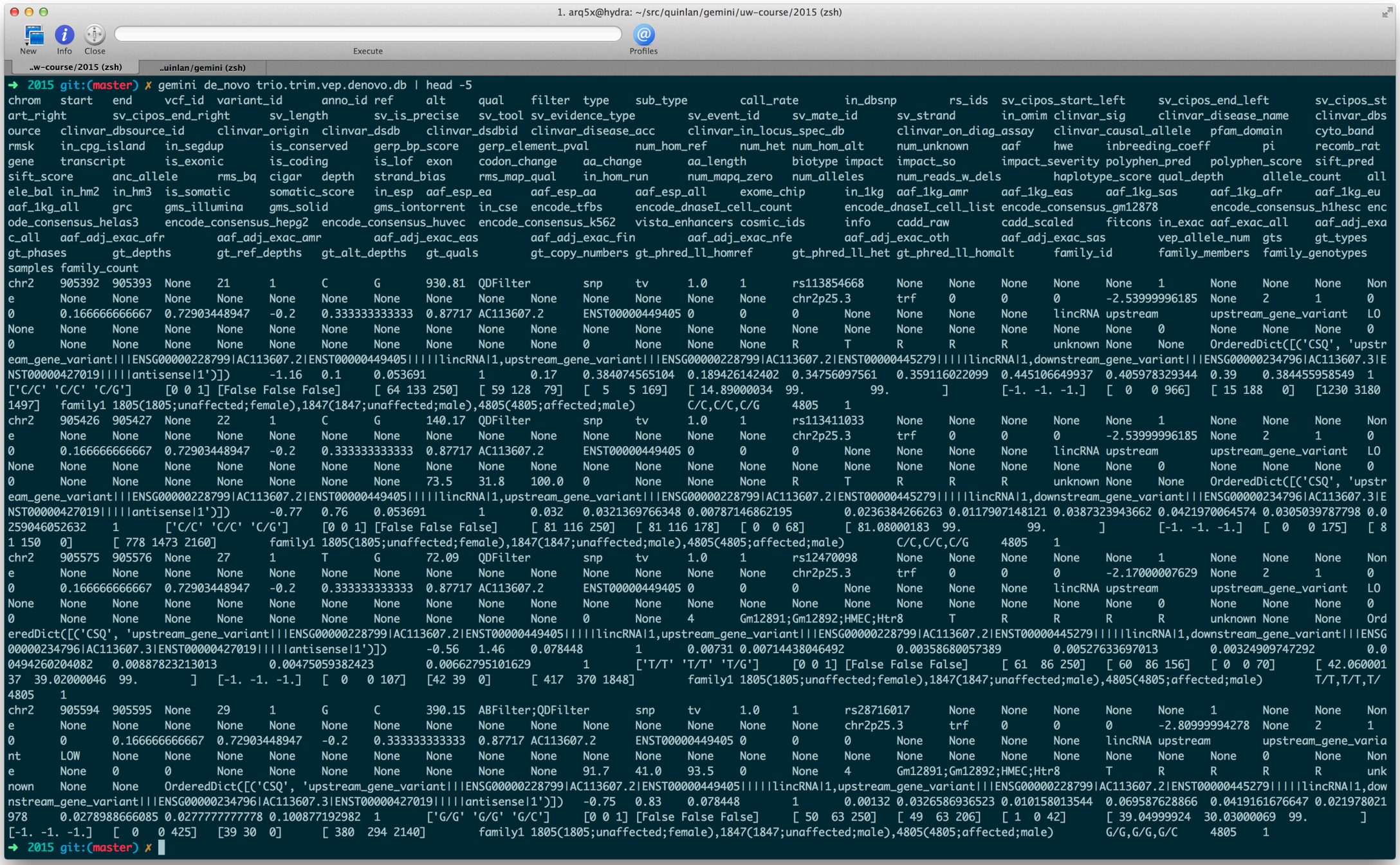
```
#decomposes multiallelic variants into biallelic variants and write out to gatk.decomposed.vcf
vt decompose gatk.vcf -o gatk.decomposed.vcf

#before decomposition
#CHROM  POS      ID   REF     ALT         QUAL   FILTER  INFO                  FORMAT    S1                                  S2
1       3759889  .    TA      TAA,TAAA,T  .      PASS    AF=0.342,0.173,0.037  GT:DP:PL  1/2:81:281,5,9,58,0,115,338,46,116,809 0/0:86:0,30,323,31,365,483,38,291,325,567

#after decomposition
#CHROM  POS      ID   REF     ALT         QUAL   FILTER  INFO                                    FORMAT   S1              S2
1       3759889  .    TA      TAA         .      PASS    OLD_MULTIALLELIC=1:3759889:TA/TAA/TAAA/T  GT:PL    1/.:281,5,9     0/0:0,30,323
1       3759889  .    TA      TAAA        .      .       OLD_MULTIALLELIC=1:3759889:TA/TAA/TAAA/T  GT:PL    ./1:281,58,115  0/0:0,31,483
1       3759889  .    TA      T           .      .       OLD_MULTIALLELIC=1:3759889:TA/TAA/TAAA/T  GT:PL    ./.:281,338,809 0/0:0,38,567

One might want to post process the partial genotypes like 1/. to the best guess genotype based on the PL values.
```

http://genome.sph.umich.edu/wiki/Vt#Decompose

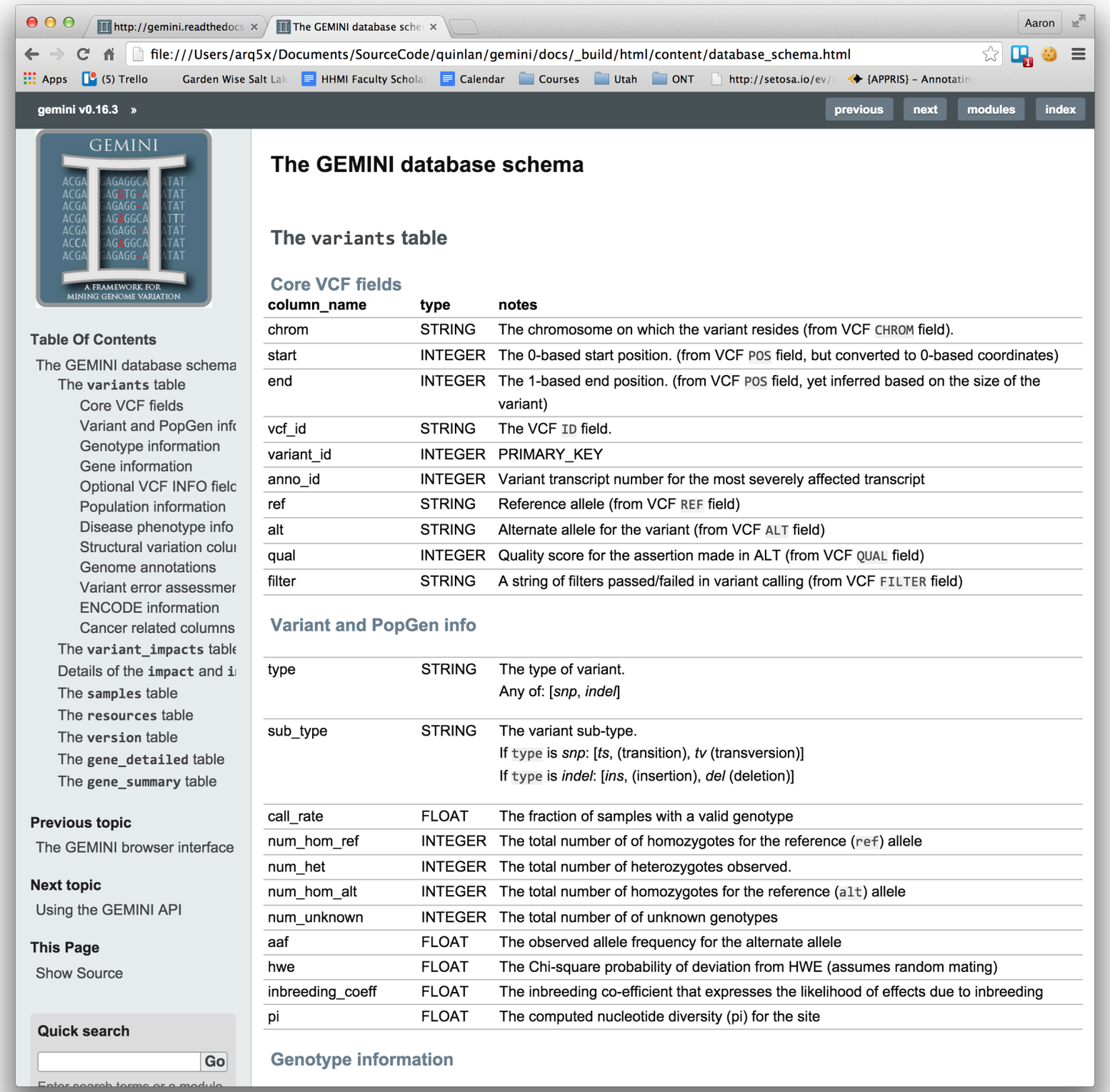
Running the de_novo tool

```
$ gemini de_novo trio.trim.vep.denovo.db
```

Note: copy and paste the full command from the Github Gist



Information overload



There are currently 115 columns in the **`variants`** table.

Perhaps a bit of overkill for a typical analysis

http://gemini.readthedocs.org/en/latest/content/database_schema.html#the-variants-table

Limit the attributes returned w/ the `--columns` option.

`--columns`

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

```
$ gemini de_novo -d 50 --columns "chrom, start, end, ref, alt" my.db

family_id   family_members  family_genotypes   family_genotype_depths       chrom   start   er
1   238(father; unknown),239(mother; unknown),173(child; affected)  A/A,A/A,A/G     152,214,25
1   238(father; unknown),239(mother; unknown),173(child; affected)  A/A,A/A,A/G     189,250,25
1   238(father; unknown),239(mother; unknown),173(child; affected)  G/G,G/G,G/A     197,247,25
1   238(father; unknown),239(mother; unknown),173(child; affected)  T/T,T/T,T/A     195,250,24
...
```

Note

The output will always start with the family ID, the family members, the observed genotypes, and the observed aligned sequencing depths for the family members.

Note: copy and paste the full command from the Github Gist

```
$ gemini de_novo \
    --columns "chrom, start, end, ref, alt, \
                    filter, qual, gene, impact" \
    trio.trim.vep.denovo.db
```

Limit the attributes returned w/ the `--columns` option.

`--columns`

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

```
$ gemini de_novo -d 50 --columns "chrom, start, end, ref, alt" my.db

family_id   family_members  family_genotypes   family_genotype_depths       chrom   start   er
1   238(father; unknown),239(mother; unknown),173(child; affected)  A/A,A/A,A/G     152,214,25
1   238(father; unknown),239(mother; unknown),173(child; affected)  A/A,A/A,A/G     189,250,25
1   238(father; unknown),239(mother; unknown),173(child; affected)  G/G,G/G,G/A     197,247,25
1   238(father; unknown),239(mother; unknown),173(child; affected)  T/T,T/T,T/A     195,250,24
...
```

Note

The output will always start with the family ID, the family members, the observed genotypes, and the observed aligned sequencing depths for the family members.

Note: copy and paste the full command from the Github Gist

```
$ gemini de_novo \
   --columns "chrom, start, end, ref, alt, \
                  filter, qual, gene, impact" \
   trio.trim.vep.denovo.db
```

1. arq5x@hydra: ~/src/quinlan/gemini/uw-course/2015 (zsh)

New Info Close Execute Profiles

..w-course/2015 (zsh) ..uinlan/gemini (zsh) ..uinlan/gemini (zsh)

```
➜ 2015 git:(master) ✗ gemini de_novo \
   --columns "chrom, start, end, ref, alt, \
              filter, qual, gene, impact" \
   trio.trim.vep.denovo.db | head
```

chrom	start	end	ref	alt	filter	qual	gene	impact	variant_id	family_id	family_members	family_genotypes	samples	family_count
chr2	905392	905393	C	G	QDFilter	930.81	AC113607.2	upstream	21	family1	1805(1805;unaffected;female),1847(1847;unaffected;male),4805(4805;affected;male)	C/C,C/C,C/G	4805	1
chr2	905426	905427	C	G	QDFilter	140.17	AC113607.2	upstream	22	family1	1805(1805;unaffected;female),1847(1847;unaffected;male),4805(4805;affected;male)	C/C,C/C,C/G	4805	1
chr2	905575	905576	T	G	QDFilter	72.09	AC113607.2	upstream	27	family1	1805(1805;unaffected;female),1847(1847;unaffected;male),4805(4805;affected;male)	T/T,T/T,T/G	4805	1
chr2	905594	905595	G	C	ABFilter;QDFilter	390.15	AC113607.2	upstream	29	family1	1805(1805;unaffected;female),1847(1847;unaffected;male),4805(4805;affected;male)	G/G,G/G,G/C	4805	1
chr2	905633	905634	A	C	QDFilter	393.91	AC113607.2	upstream	30	family1	1805(1805;unaffected;female),1847(1847;unaffected;male),4805(4805;affected;male)	A/A,A/A,A/C	4805	1
chr2	905762	905763	G	C	ABFilter;LowQual;QDFilter;QUALFilter;SBFilter	15.12	AC113607.2	upstream	33	family1	1805(1805;unaffected;female),1847(1847;unaffected;male),4805(4805;affected;male)	G/G,G/G,G/C	4805	1
chr2	905804	905805	G	C	QDFilter	443.57	AC113607.2	upstream	35	family1	1805(1805;unaffected;female),1847(1847;unaffected;male),4805(4805;affected;male)	G/G,G/G,G/C	4805	1
chr2	905818	905819	T	G	QDFilter;SBFilter	689.37	AC113607.2	upstream	36	family1	1805(1805;unaffected;female),1847(1847;unaffected;male),4805(4805;affected;male)	T/T,T/T,T/G	4805	1
chr2	905819	905820	T	A	ABFilter;QDFilter;SBFilter	296.27	AC113607.2	upstream	37	family1	1805(1805;unaffected;female),1847(1847;unaffected;male),4805(4805;affected;male)	T/T,T/T,T/A	4805	1

```
➜ 2015 git:(master) ✗
```

As of version 0.14, the tools have a standardized output that is different from previous versions. Requested *–columns* will come first followed by a standard set of columns:

- variant_id - unique id from the databse
- family_id - family id for this row
- family_members - which family members were tested
- family_genotypes - genotypes of this family
- samples - samples contributing to this row appearing in the results
- family_count - number of families with this effect

http://gemini.readthedocs.org/en/latest/content/tools.html#common-args-common-arguments

Better, but there are still so many (likely false) candidates.

```
$ gemini de_novo \
    --columns "chrom, start, end, ref, alt, \
               filter, qual, gene, impact" \
    trio.trim.vep.denovo.db | wc -l
```

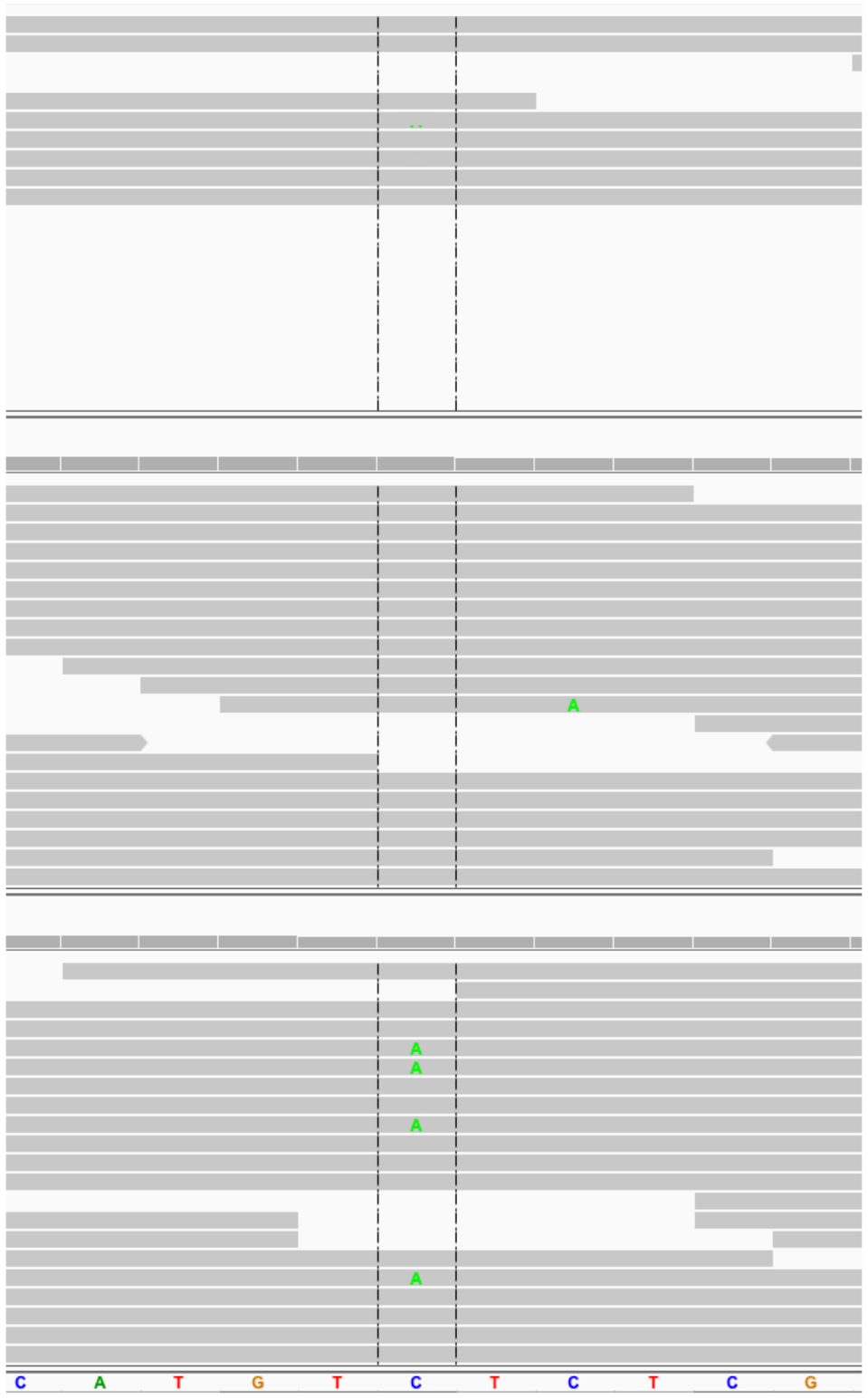
771 candidates!

Causes of erroneous genotype predictions: **lack of depth**

Father

Mother

Proband



Let's enforce a minimum sequence depth for each subject: **-d**

-d [0] (depth)

Unfortunately, spurious inherited variants can often appear due to insufficient sequence coverage One simple way to filter such artifacts is to enforce a minimum sequence depth (default: 0) for each sample. We can do that with this flag.

```
$ gemini de_novo \
    --columns "chrom, start, end, ref, alt, \
               filter, qual, gene, impact" \
    -d 6 \
    trio.trim.vep.denovo.db | wc -l
```

727 candidates

Also use a minimum genotype quality for each subject: **--min-gq**

--min-gq [0]

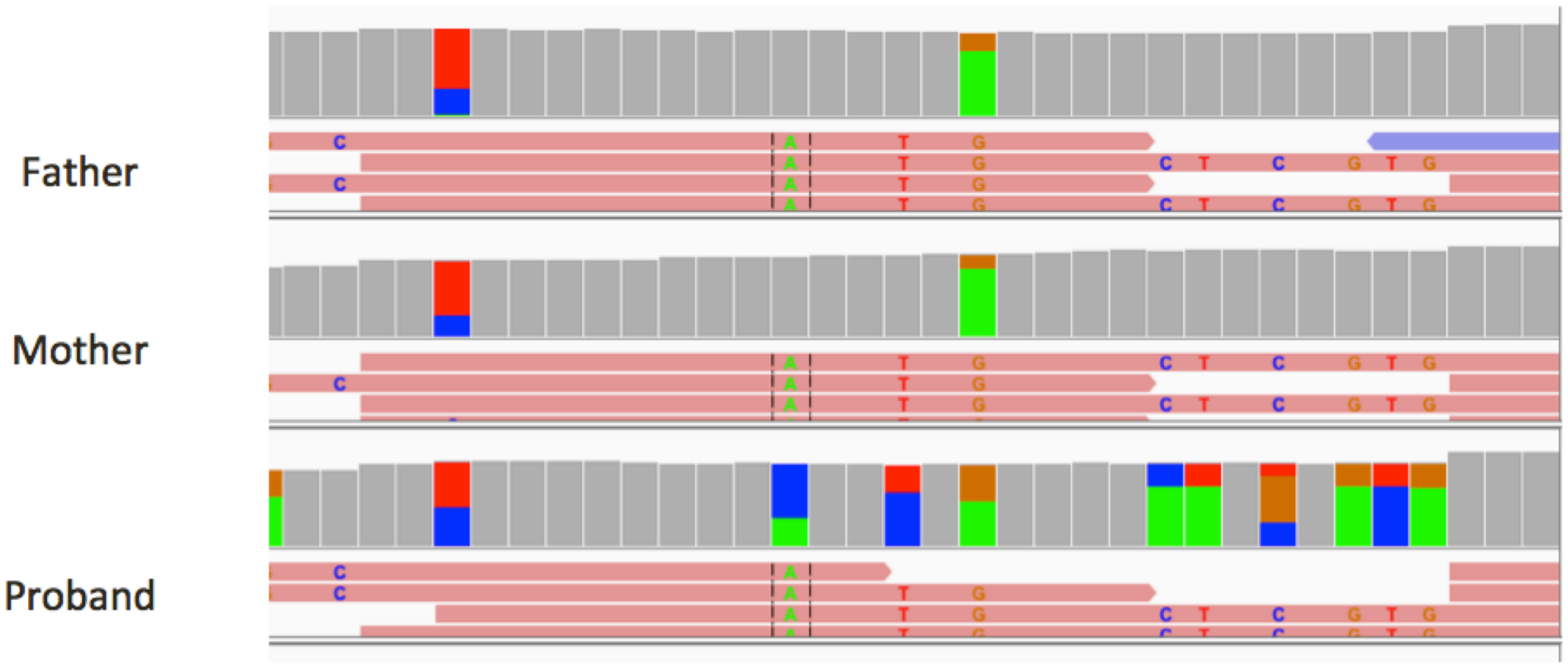
Filter variants that do not have at least this genotype quality for each sample in a family. Default is 0. Higher values are more stringent.

Genotype qualities (GQ) use a Phred scaling system…

```
$ gemini de_novo \
    --columns "chrom, start, end, ref, alt, \
               filter, qual, gene, impact" \
  -d 6 \
  --min-gq 20 \
  trio.trim.vep.denovo.db | wc -l
```

617 candidates

Causes of erroneous genotype predictions: **low quality variants**



Require that the mutation passes GATK QC with `--filter`

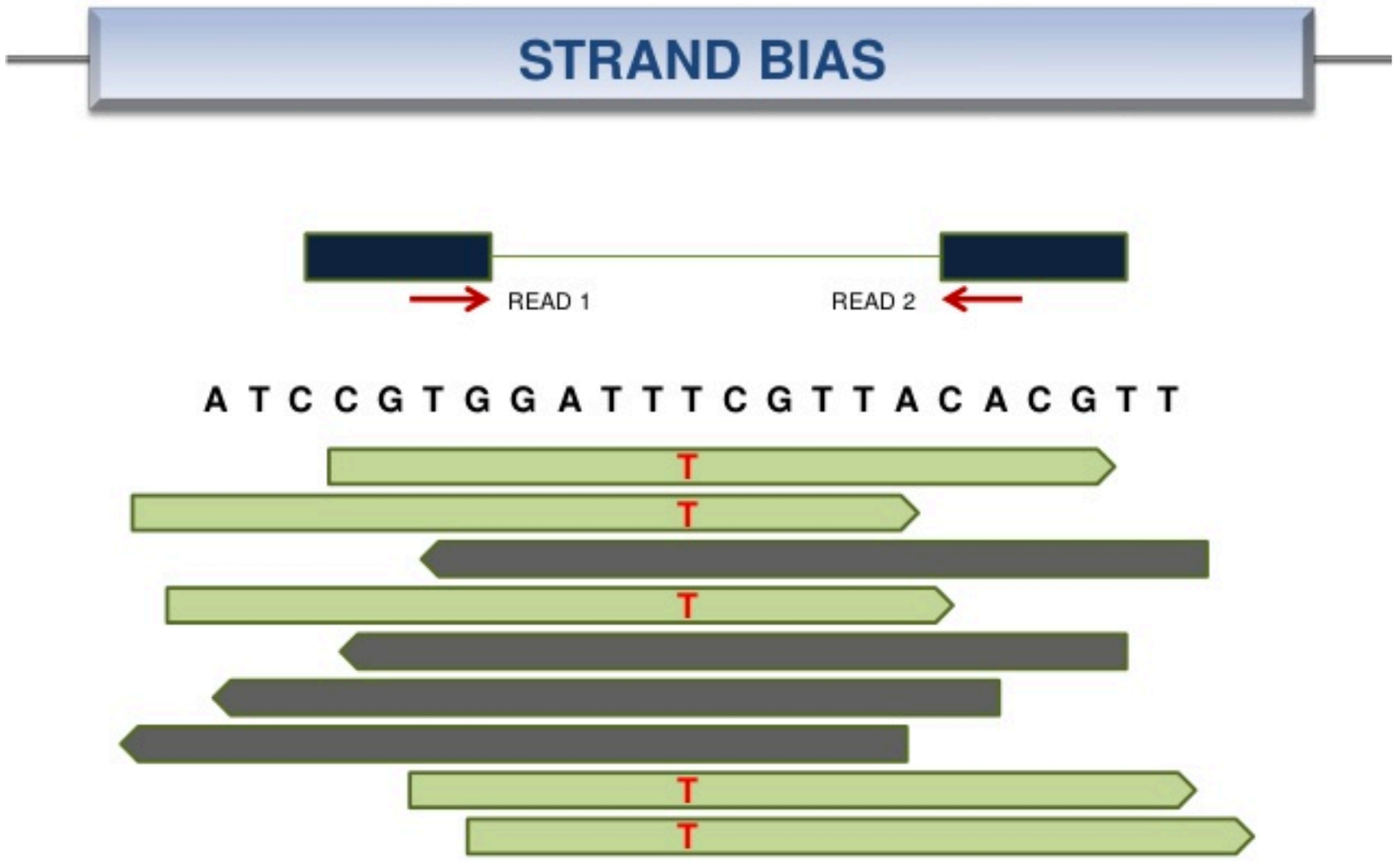
`--filter`

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.

```
$ gemini de_novo \
    --columns "chrom, start, end, ref, alt, \
               filter, qual, gene, impact" \
  -d 6 \
  --min-gq 20 \
  --filter "filter is NULL" \
  trio.trim.vep.denovo.db | wc -l
```

52 candidates

But we also want to be a bit tolerant of strand bias…



*See also: Allhoff et al. BMC Bioinformatics 2013 **14** (Suppl 5):S1 doi:10.1186/1471-2105-14-S5-S1*

as higher strand bias is expected in exome sequencing owing to different hybridization efficiency among probes

Require that the mutation passes GATK QC with **`--filter`**

`--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.

```
$ gemini de_novo \
   --columns "chrom, start, end, ref, alt, \
              filter, qual, gene, impact" \
   -d 6 \
   –min-gq 20 \
   --filter "(filter is NULL or filter=='SBFilter')" \
   trio.trim.vep.denovo.db | wc -l
```

53 candidates

Require that the mutation is likely to have functional consequence

Details of the `impact` and `impact_severity` columns

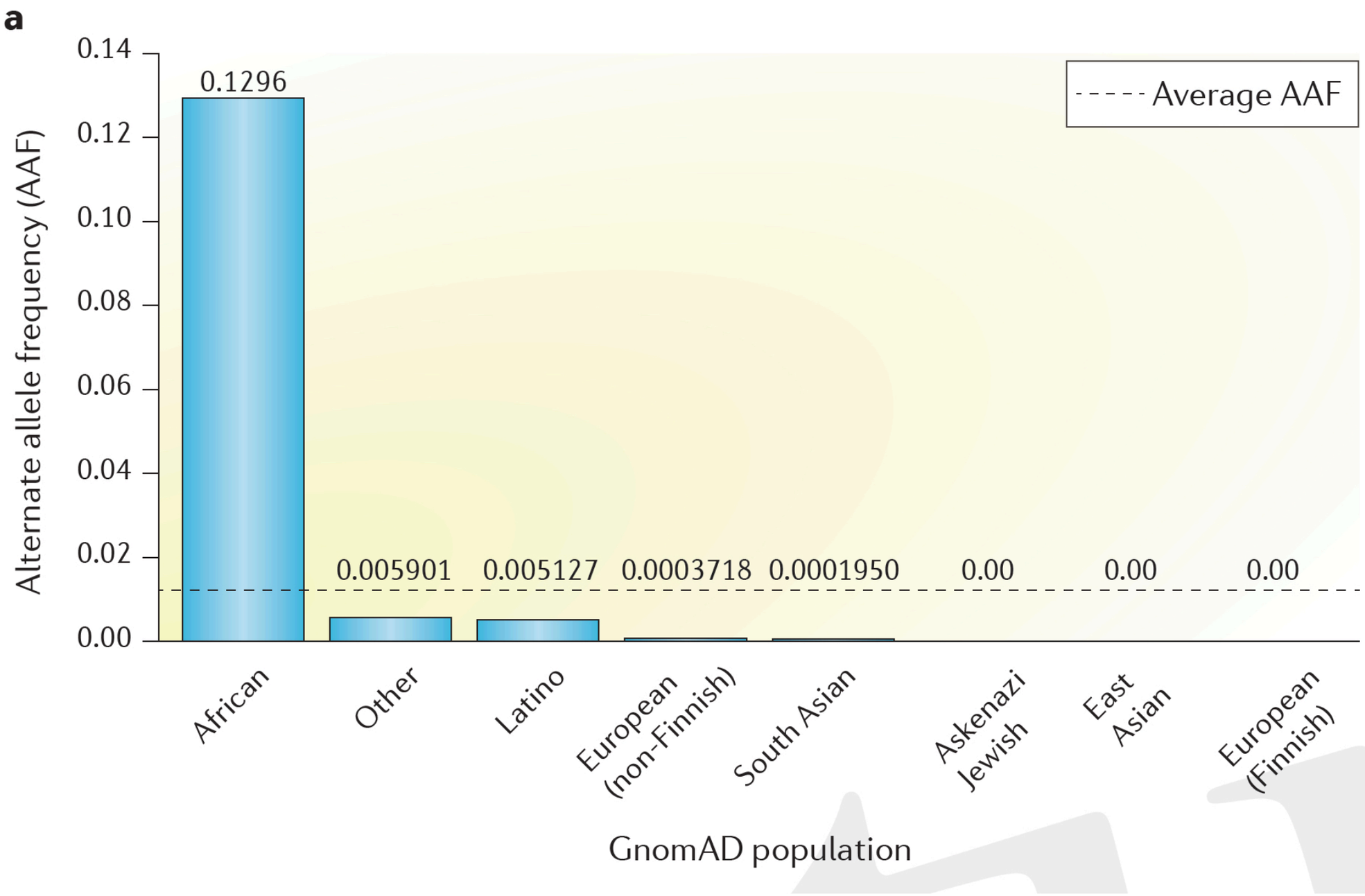
impact severity	impacts	SO_impacts
HIGH	• exon_deleted • frame_shift • splice_acceptor • splice_donor • start_loss • stop_gain • stop_loss • non_synonymous_start • transcript_codon_change • rare_amino_acid • chrom_large_del	• exon_loss_variant • frameshift_variant • splice_acceptor_variant • splice_donor_variant • 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 • CDS • upstream • downstream • intergenic • intragenic • gene • transcript • exon • start_gain • synonymous_start • intron_conserved • nc_transcript • NMD_transcript • incomplete_terminal_codon • nc_exon • transcript_ablation • transcript_amplification • feature elongation • feature truncation	• stop_retained_variant • synonymous_variant • 5_prime_UTR_variant • 3_prime_UTR_variant • intron_variant • coding_sequence_variant • upstream_gene_variant • downstream_gene_variant • intergenic_variant • intragenic_variant • gene_variant • transcript_variant • exon_variant • 5_prime_UTR_premature_start_codon_gain_variant • start_retained_variant • conserved_intron_variant • nc_transcript_variant • NMD_transcript_variant • incomplete_terminal_codon_variant • non_coding_exon_variant • transcript_ablation • transcript_amplification • feature_elongation • feature_truncation

Require that the mutation is likely to have functional consequence

```
$ gemini de_novo \
  --columns "chrom, start, end, ref, alt, \
             filter, qual, gene, impact" \
  -d 6 \
  --min-gq 20 \
 --filter "(filter is NULL or filter=='SBFilter') \
           and impact_severity != 'LOW'" \
 trio.trim.vep.denovo.db | wc -l
```

14 candidates

Require that the mutation is not likely to be a known polymorphism



Eilbeck, Quinlan, Yandell (in press)

Filter based upon maximum alternate allele freq (**max_aaf_all**)

(more on this later)

ESP	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
1000G	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)
ExAC	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_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, 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.

Require that the mutation is not likely to be a known polymorphism

```
$ gemini de_novo \
  --columns "chrom, start, end, ref, alt, \
             filter, qual, gene, impact" \
  -d 6 \
  --min-gq 20 \
  --filter "(filter is NULL or filter=='SBFilter') \
            and impact_severity != 'LOW' and max_aaf_all < 0.005" \
  trio.trim.vep.denovo.db | wc -l
```

5 candidates!

5 candidates. Which is causal? Requires manual inspection…

```
$ gemini de_novo \
  --columns "chrom, start, end, ref, alt, \
             filter, qual, gene, impact" \
  -d 6 \
  --min-gq 20 \
  --filter "(filter is NULL or filter=='SBFilter') \
            and impact_severity != 'LOW' and max_aaf_all < 0.005" \
  trio.trim.vep.denovo.db | wc -l
```

chrom	start	end	ref	alt	filter	qual	gene	impact	family_genotypes	samples	family_count
chr17	76471351	76471352	G	A	None	1245	DNAH17	splice_region_variant	G/G,G/G,G/A	4805	1
chr22	43027436	43027437	C	T	None	1320	CYB5R3	missense_variant	C/C,C/C,C/T	4805	1
chr15	41229630	41229631	T	G	None	2116	DLL4	missense_variant	T/T,T/T,T/G	4805	1
chr22	23040690	23040691	G	A	None	2574	IGLV2-23	missense_variant	G/G,G/G,G/A	4805	1
chr22	23040728	23040729	G	C	None	2087	IGLV2-23	missense_variant	G/G,G/G,G/C	4805	1

Phenotype: blue skin disease

Which gene can we rule out at a glance?

Load the following files into IGV

BAM alignment files (don't forget to use qsh to run igv):

```
wed/data/1805.workshop2018.mini.sorted.bam
wed/data/1847.workshop2018.mini.sorted.bam
wed/data/4805.workshop2018.mini.sorted.bam
```

VCF variant file:

```
trio.trim.vep.vcf.gz
```