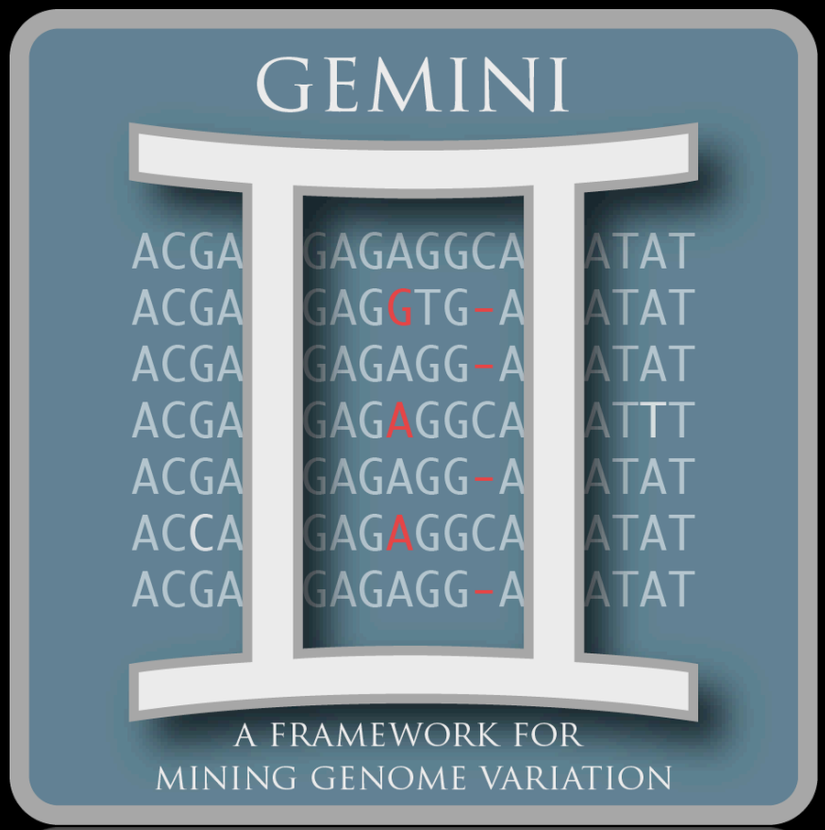


Exploring genetic variation with a tour guide.

Aaron Quinlan
Public Health Sciences
Center for Public Health Genomics
quinlanlab.org

UNIVERSITY of VIRGINIA

CENTER for PUBLIC HEALTH GENOMICS

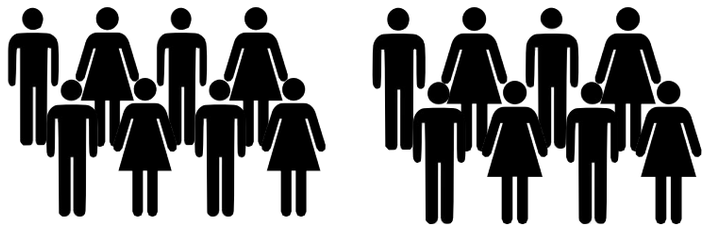
Exploring human genomes: outline

- Technical challenges (many)
- Analytical challenges (more still?)
- Data integration (large; diverse)
- Our solution: GEMINI

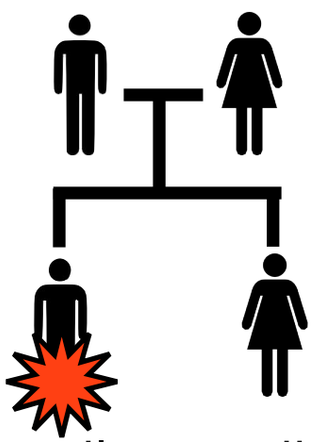
Typical study designs

phenotype

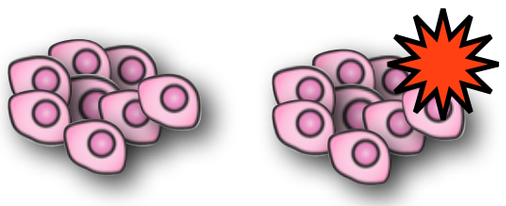
no phenotype



case / control



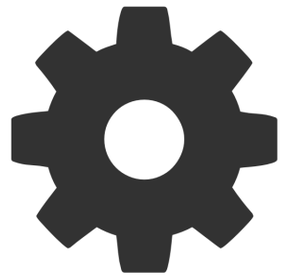
family studies



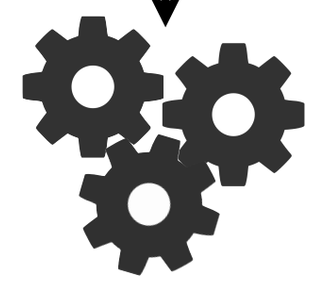
normal tumor

cancer genomics

Genome or exome sequencing

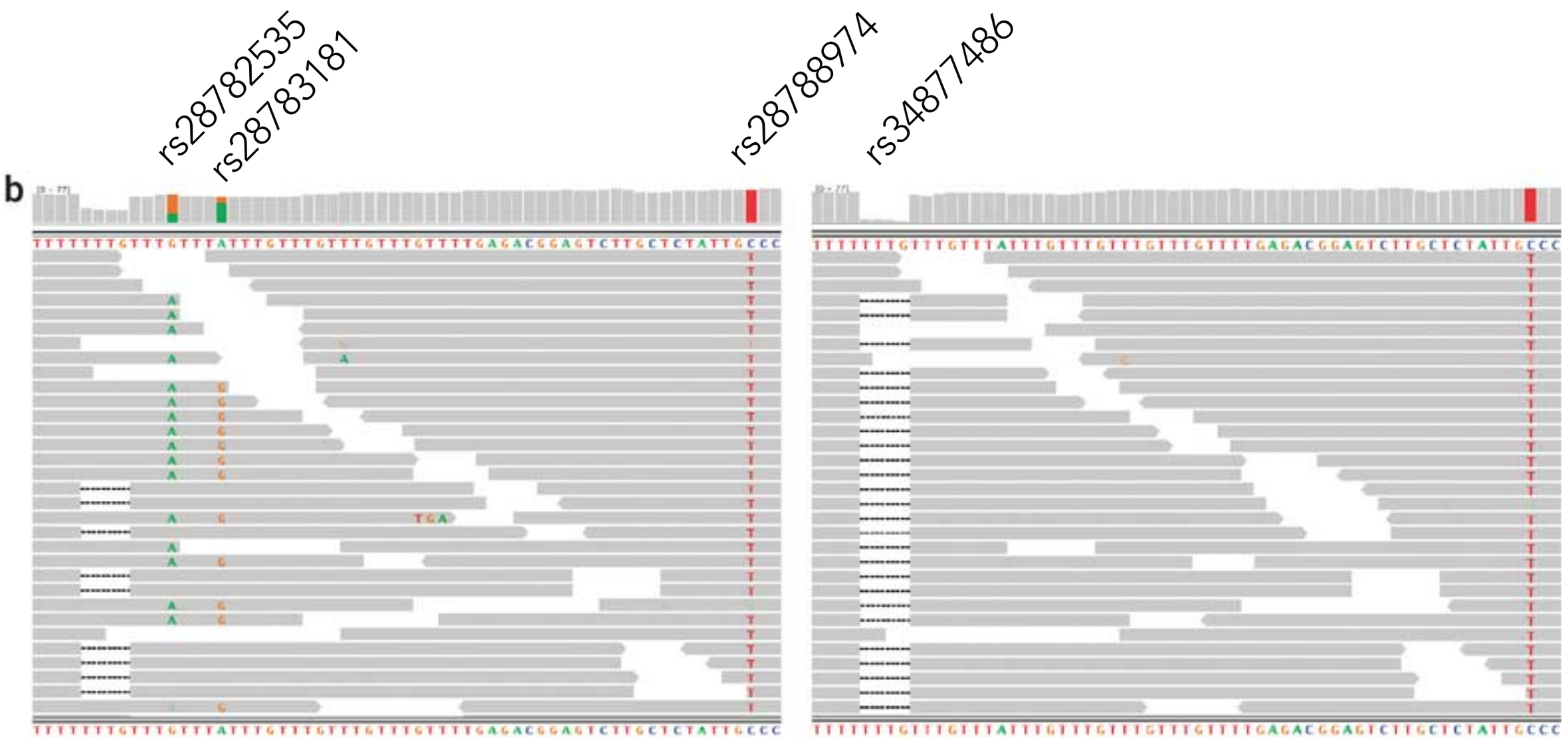


Align sequences
Scrub data
Call variants
(e.g., GATK, etc.)



Interpret.
Prioritize.
Repeat.

Technical Challenges: alignment issues

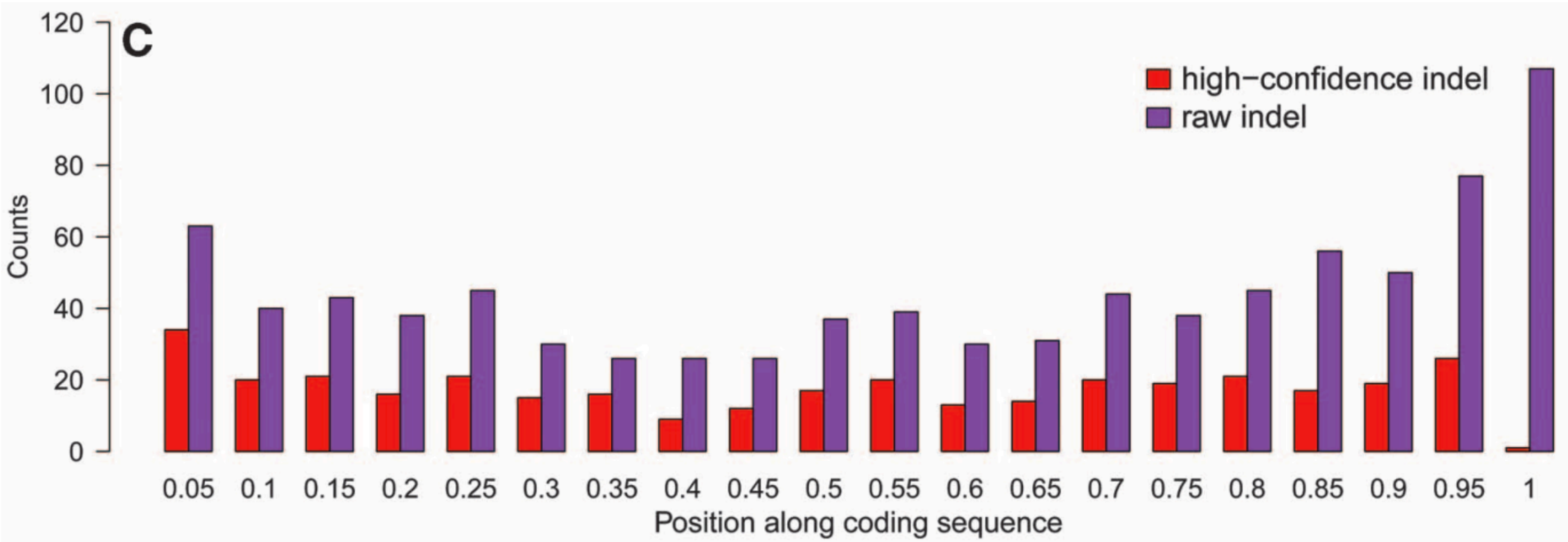


"raw" alignments

after MSA correction

http://www.nature.com/ng/journal/v43/n5/fig_tab/ng.806_F2.html

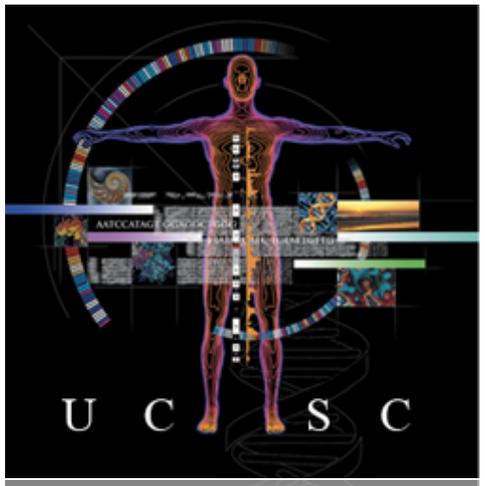
Technical Challenges: LoF variants



Variant type	Before filtering					After filtering				
	Total	1000G low-coverage average per individual			NA12878	Total	1000G low-coverage average per individual			NA12878
		CEU	CHB+JPT	YRI			CEU	CHB+JPT	YRI	
Stop	1111	85.7 (21.8)	113.4 (26.7)	109.1 (23.7)	115 (25)	565	26.2 (5.2)	27.4 (6.9)	37.2 (6.3)	23 (2)
Splice	658	80.5 (29.5)	98.1 (35.6)	89.0 (30.4)	95 (32)	267	11.2 (1.9)	13.2 (2.5)	13.7 (1.9)	12 (1)
Frameshift indel	1040	217.8 (112.1)	225.5 (121.7)	247.2 (118.7)	348 (159)	337	38.2 (9.2)	36.2 (9.0)	44.0 (8.0)	38 (11)
Large deletion	142	32.4 (12.2)	31.2 (11.8)	31.4 (9.7)	31 (5)	116	28.3 (6.2)	26.7 (5.9)	26.6 (5.5)	24 (4)
Total	2951	416.4 (175.6)	468.2 (195.8)	476.7 (316.0)	654 (286)	1285	103.9 (22.5)	103.5 (24.3)	121.5 (21.7)	97 (18)

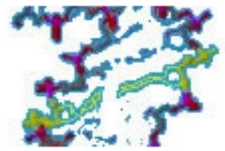
MacArthur et al. Science. 2012

Analytical challenges: data integration



Conservation
Repeat elements
Genome Gaps
Cytobands
Gene annotations
"Mappability"
DeCIPHER
ISGA

dbSNP
Short Genetic Variations



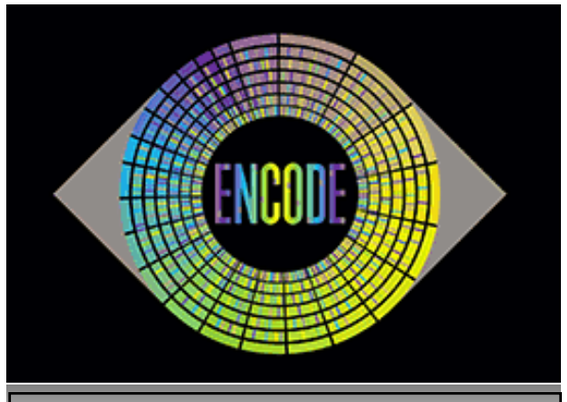
ClinVar

Genetic variation

...CCTCATG**C**ATGGAAA...
...CCTCATG**T**ATGGAAA...
...CCTCATG**C**ATGGAAA...
...CCTCATG**C**ATGGAAA...
...CCTCATG**T**ATGGAAA...
...CCTCATG**C**ATGGAAA...
...CCTCATG**T**ATGGAAA...

OMIM
Online Mendelian Inheritance in Man

1000 Genomes
A Deep Catalog of Human Genetic Variation



Chromatin marks
DNA methylation
RNA expression
TF binding

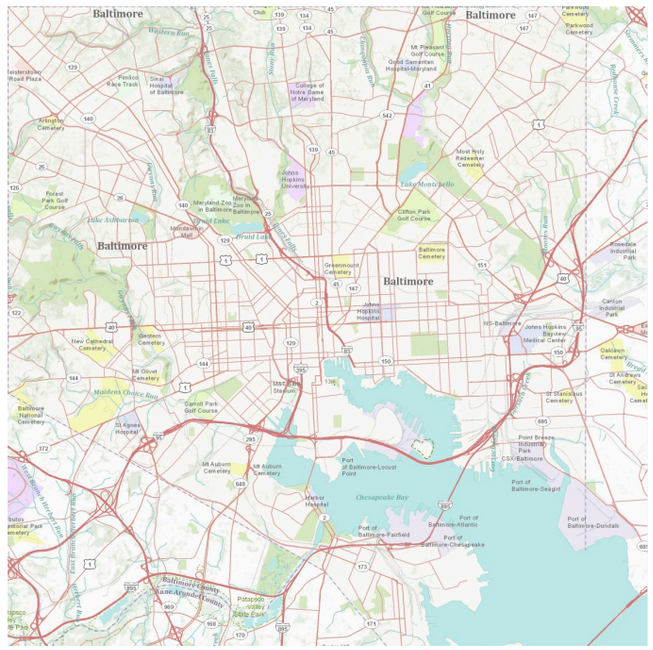
Pfam

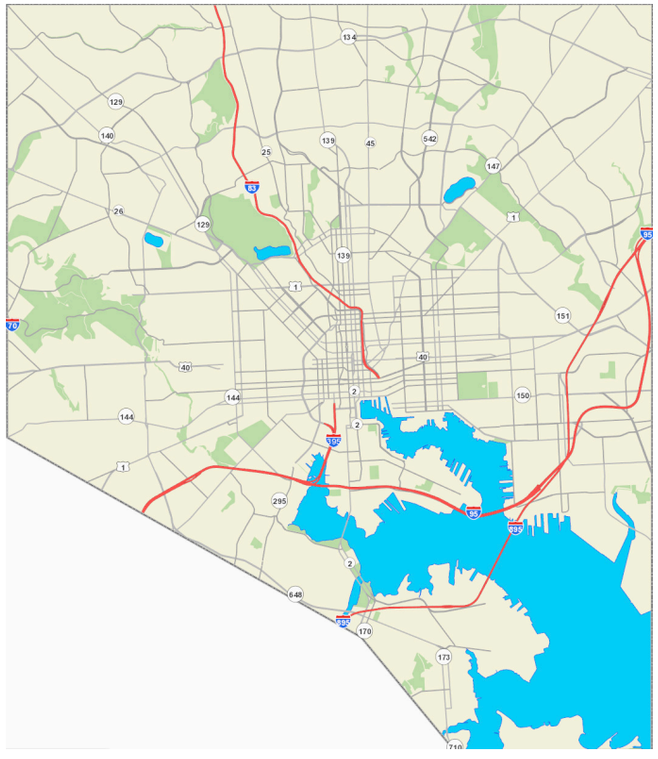
KEGG Kyoto Encyclopedia of Genes and Genomes

HGMD®

Human Protein Reference Database

Much like map “overlays”





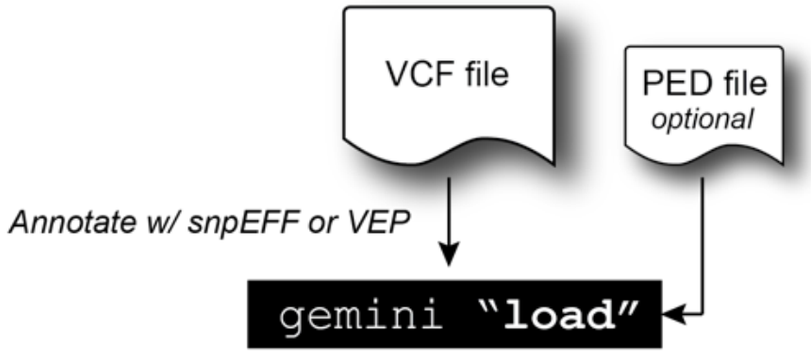
Each layer provides distinct information in different contexts.

Our solution: GEnome MINIng

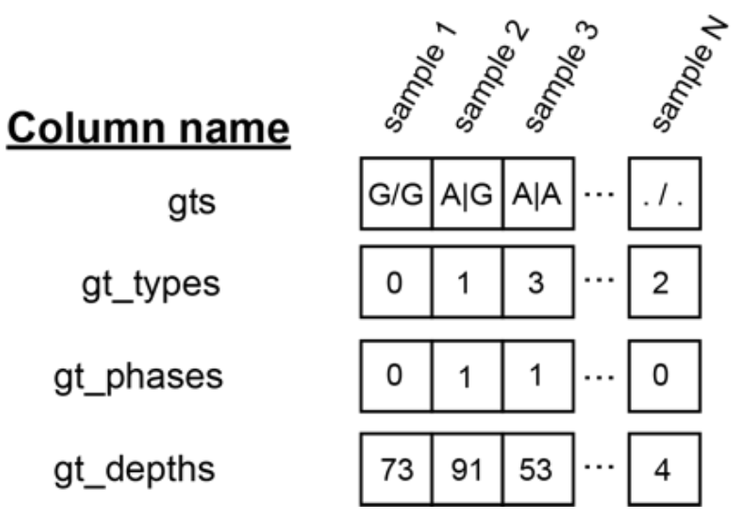
- Powerful
- Flexible
- Comprehensive
- Scalable
- Portable
- Reproducible

Framework for personal, medical, and population genomics

The GEMINI framework



User-defined annotations



Access to and filtering on individual genotypes

Annotation source	Variants Table
From VCF	**Core**: chrom, ref. allele, alt. allele, id, qual, filter, ...
From VCF	**Variant info**: depth, strand bias, allele balance, ...
Computed	**Statistics**: type, call rate, Pi, allele freq., HWE, ...
snpEff, VEP, KEGG*, HPRD*	**Gene**: gene, transcript, impact, LoF, SIFT, pathway, ...
1000G, dbSNP, ESP, HapMap	**Population**: rsId, ESP and 1000G allele freq., recomb.
ClinVar	**Disease**: OMIM, clinical significance, disease, ID
UCSC	**Genome**: Conservation, RptMasker, CpG, SegDup...
UCSC	**Mappability**: gaps; Illumina, SOLiD, Ion mappability
ENCODE	**Regulation**: TF binding, DNase1, chrom. segment.
User defined	**Custom**: New columns based upon overlaps between variants and researcher-defined genome annotations.
Computed	**Genotypes**: genotype, type (e.g., HET), phase, depth, number of hets, hom_ref, hom_alt, unknown, etc. **Individual samples**: *gts.sample1, gt_types.sample2*

Variant Impacts Table

Variant impacts for each gene/transcript

Samples Table

Sample Id, sex, phenotype, relatives...
One entry / sample in VCF / PED file.

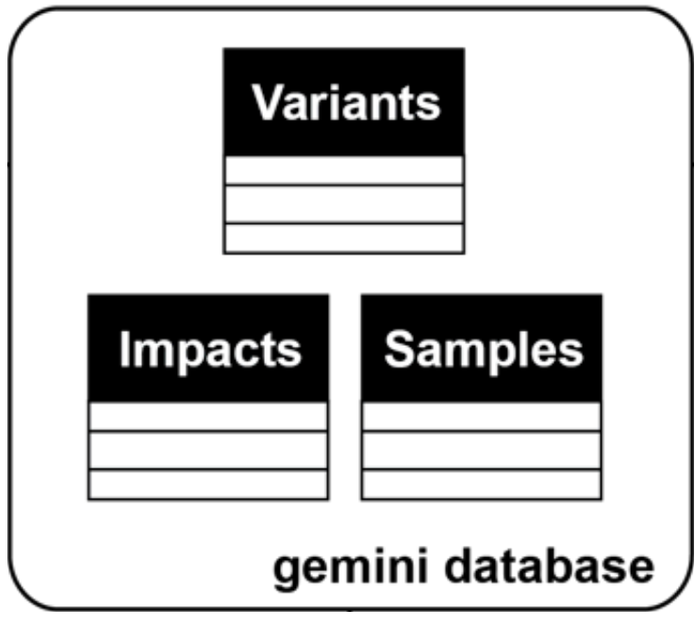
Resources Table

The name and version of all annotation files used by GEMINI.

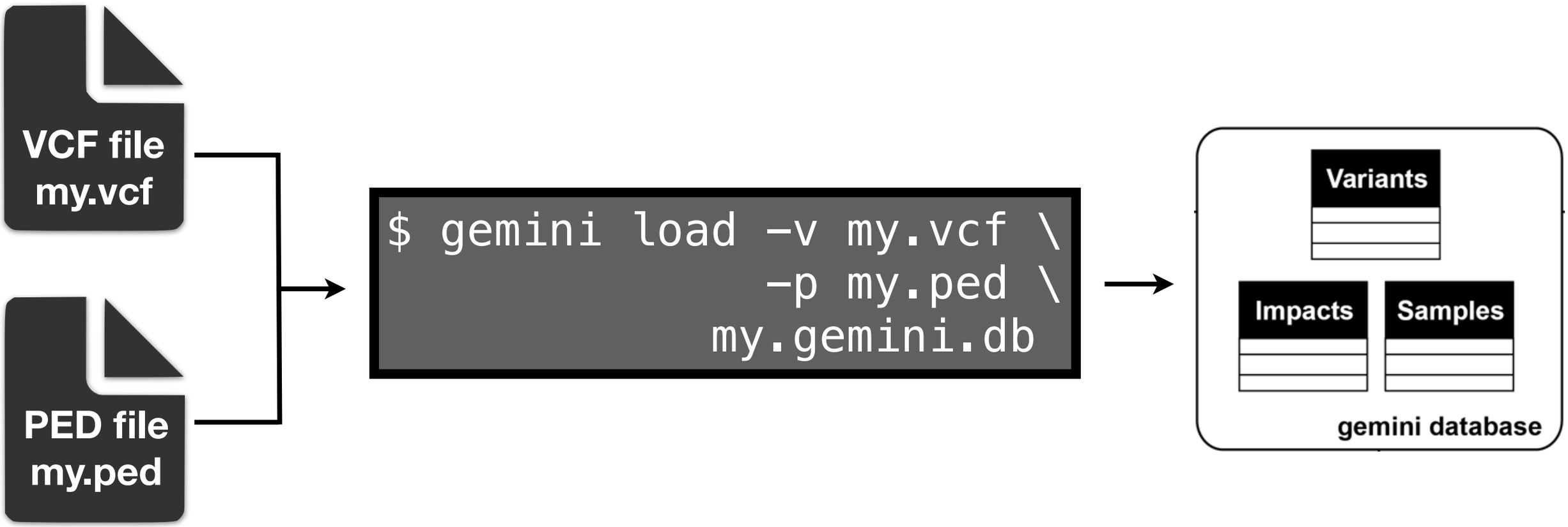
Version Table

Tracks the GEMINI software version that was used to create the database.

Mining variation with GEMINI



Example workflow



Example analyses: *ad hoc* exploration

1. Find all novel or rare (<1%), LoF variants

```
$ gemini query -q "select * from variants
                   where is_lof = 1
                   and (in_dbsnp = 0 or
                        aaf <= 0.01)"
                   my.gemini.db
```

2. Find all variants with a known clinical phenotype

```
$ gemini query -q "select * from variants
                   where in_omim = 1
                   or clinvar_disease_name is not NULL"
                   my.gemini.db
```

3. Find LoF variants where a sample has a specific genotype

```
$ gemini query -q "select * from variants
                   where is_lof = 1"
                 --gt_filter "gts.sampleXYZ == 'HET'"
                   my.gemini.db
```

Example analyses: *built-in analyses*

1. Find *de novo* mutations

```
$ gemini de_novo my.gemini.db
```

2. Find compound heterozygotes

```
$ gemini comp_hets my.gemini.db
```

3. Find variants meeting an auto. recessive pattern

```
$ gemini auto_recessive my.gemini.db
```

4. Find variants meeting an auto. dominant pattern

```
$ gemini auto_dominant my.gemini.db
```

5. Find interacting genes each with LoF variants.

```
$ gemini lof_interactions my.gemini.db
```

Example analyses: *tool / method dev.*

```
from gemini import GeminiQuery

query = GeminiQuery(“my.db”)
query.run("select * from variants")
for row in query:
    # print specific columns
    print row[‘chrom’], row[‘rsid’]

# extract sample genotypes
genotype_types = row.gt_types

# association test
if assoc_test(genotype_types) < 1E-8:
    print row

# your groundbreaking idea!
if whizbang_test(genotype_types) < 1E-8:
    print row
```

Example datasets

- Illumina Platinum Trio (CEU trio)
  - ftp://ftp.platinumgenomes.org/trio/trio.vcf.gz
  - 6.4 million variants; 3 genotypes / variant
  - VCF is 319 Megabytes
- 1000 genomes VCF from 1046 samples
  - http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/integrated_call_sets/
  - 39.7 million variants; 1046 genotypes / variant
  - VCF is 144 Gigabytes

Performance: *database loading and size*

Loading is slowest step, but can be parallelized on clusters or multi-CPU machines.

- Illumina Platinum Trio (CEU trio)
  - 84 minutes with 4 CPUs on a Macbook Pro
  - Database is 1.2 Gigabytes
- 1000 genomes VCF from 1046 samples
  - 28 hours with 30 CPUs on a multi-CPU machine.
  - 2 hours on a cluster using 250 CPUs.
  - Database is 78 Gb versus 144 Gb for original VCF

Performance: *querying the database*

Experiment	Platinum Trio	1046 samples
Return all novel variants `select * from variants where in_dbsnp = 0`	24 sec (N=345,028)	11 sec (N=87,939)
Return all loss-of-func. variants `select * from variants where is_lof = 1`	2 sec (N=1,126)	177 sec (N=13,049)

Return all rare, loss-of-func. variants
`select * from variants`
`where is_lof = 1 and aaf < 0.01`
	2 sec
	(N=112)
		152 sec
		(N=12,683)

Filtering variants based on sample genotype criteria
 	
`select * from variants where is_lof = 1”`
`--gt-filter "gt_types.NA12878 == HET"`
	2 sec
	(N=487)
		194 sec
		(N=384)

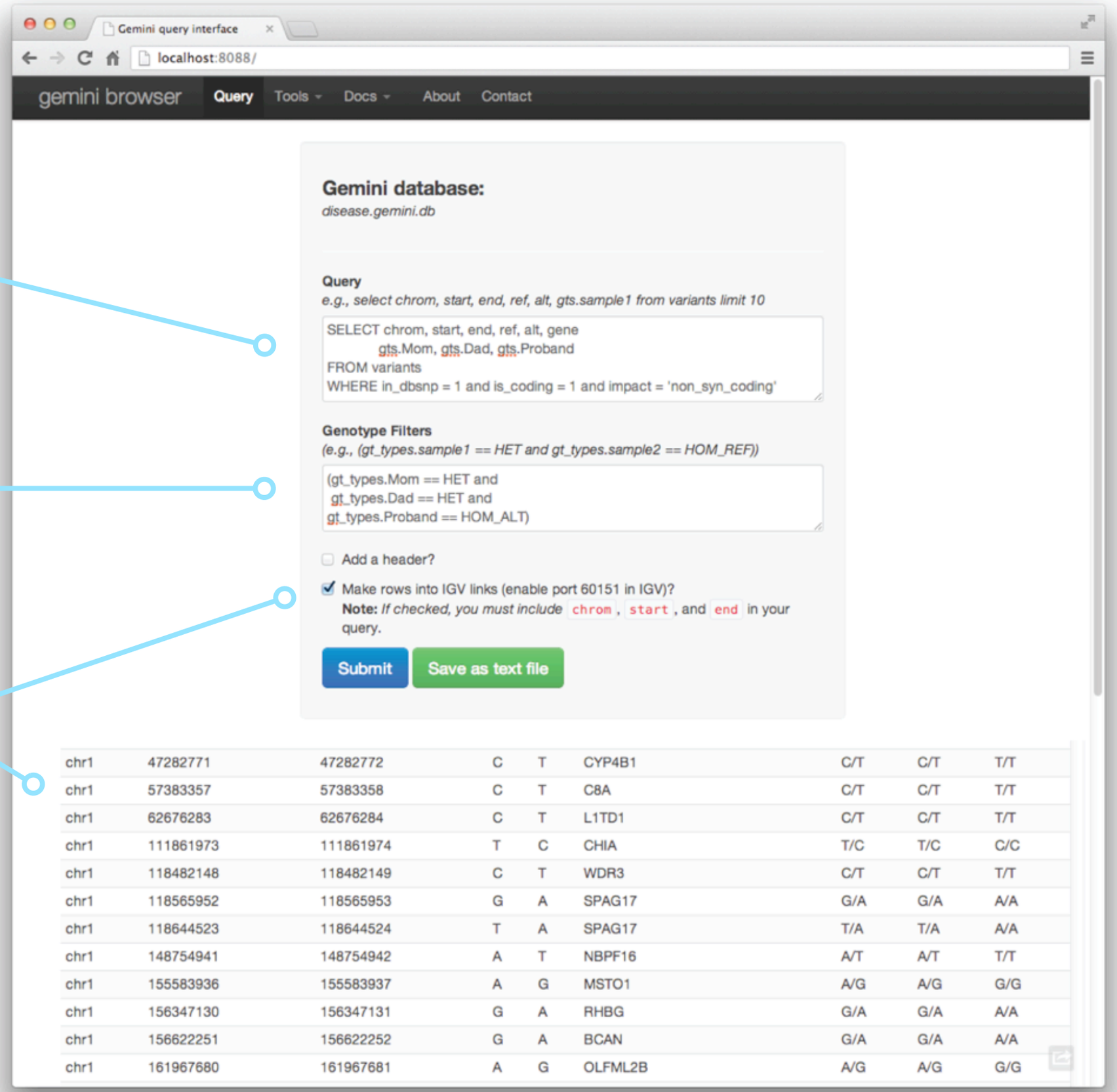
Filtering variants by specific genotypes requires decompression of sample genotype arrays. Slower, but thankfully only marginally so. Hope for scalability.

GEMINI browser

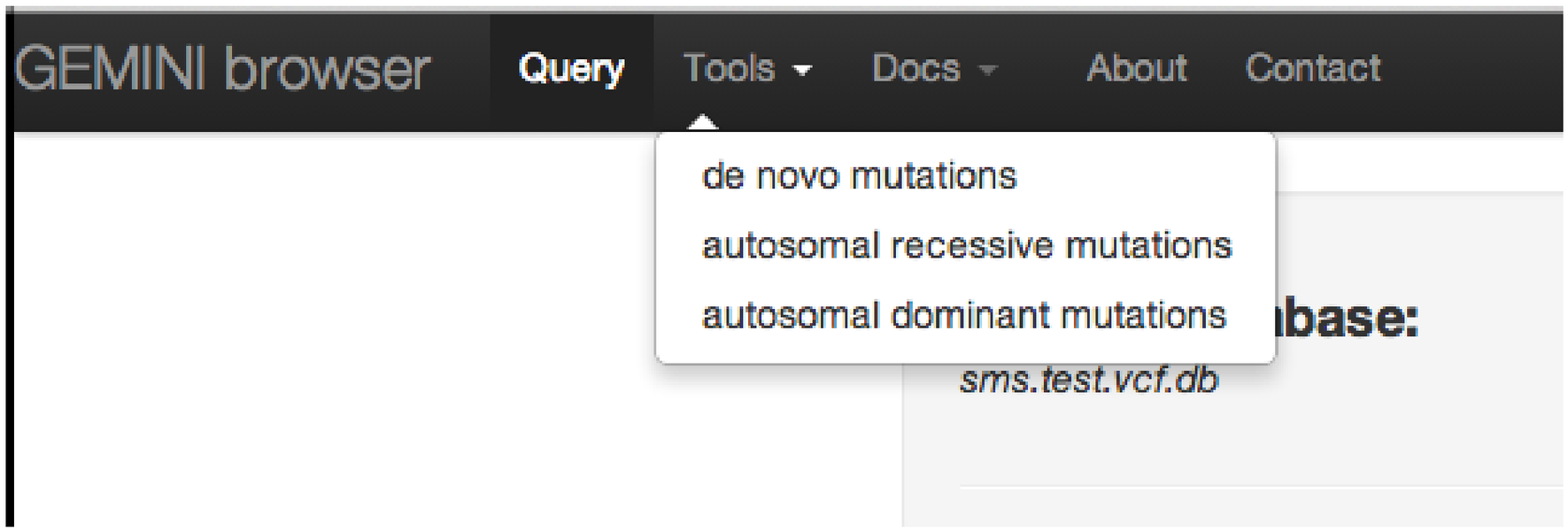
SQL query

genotype filters (e.g., inheritance patterns)

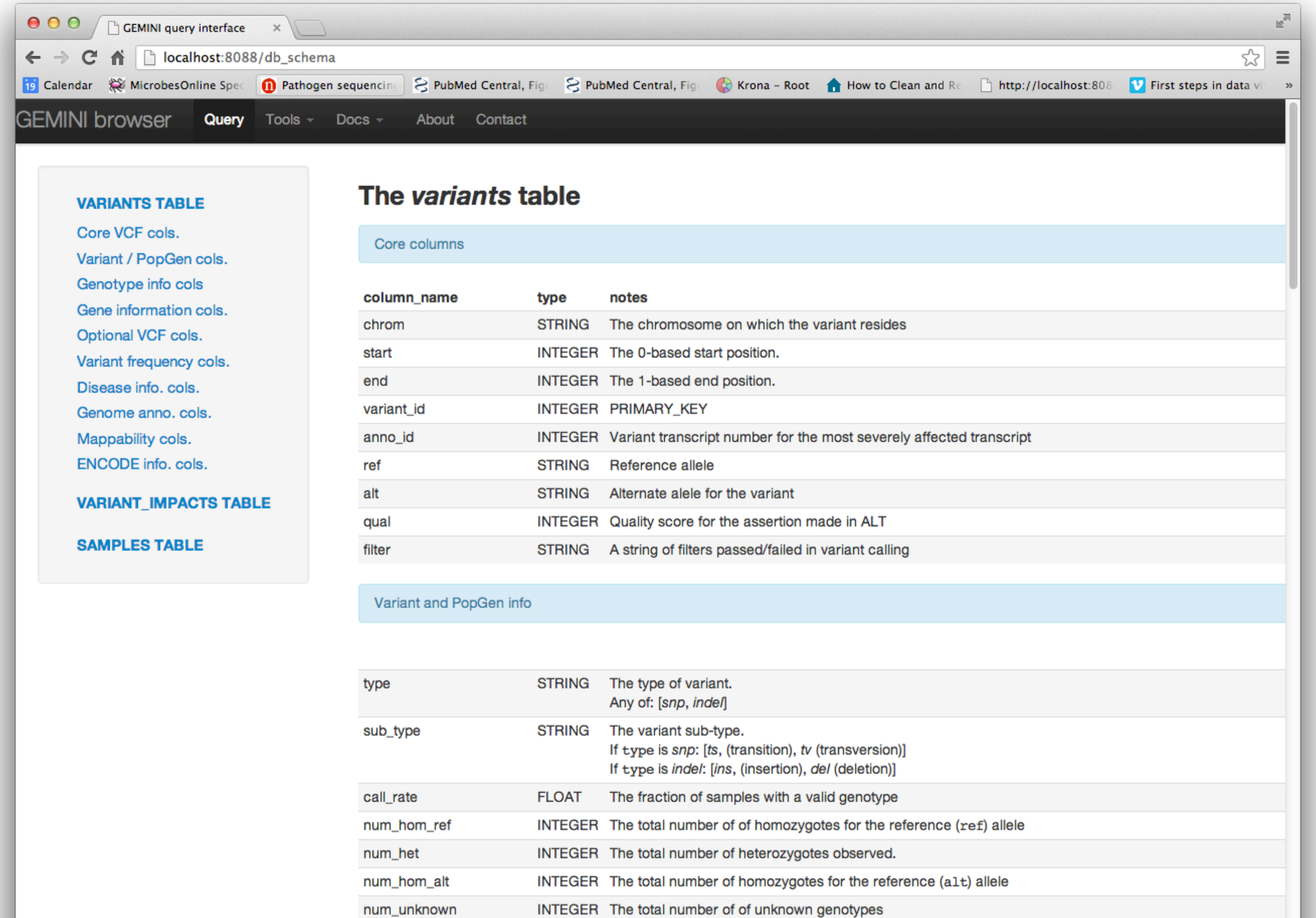
built-in links to IGV



Built-in tools



Built-in documentation



Our goals.

- Free. Open source.
  - github.com/arq5x/gemini
- Well tested and documented.
  - gemini.readthedocs.org
- Extensible, portable, & reproducible

Acknowledgements

Quinlan Lab

Uma Paila*

Postdoctoral Fellow

github.com/udp3f

Ryan Layer

Graduate Student

Co-mentored with Ira Hall

github.com/ryanlayer

Brad Chapman

Oliver Hofmann

Rory Kirchner

National Human Genome Research Institute

R01HG006693-01