Variant interpretation and prioritization with GEMINI

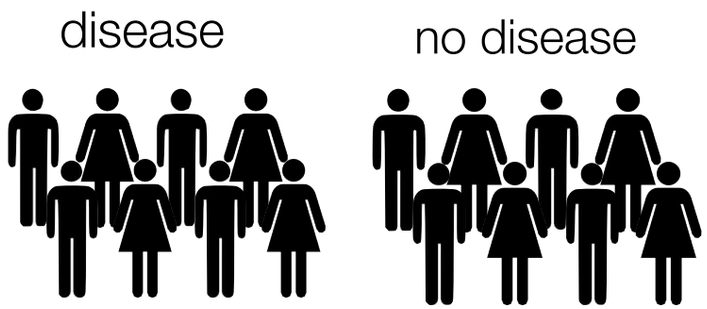
Aaron Quinlan | University of Virginia | Oct. 3, 2013

Beyond the Genome, 2013 | San Francisco, CA

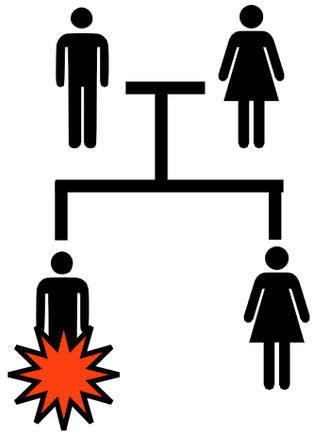
github.com/arq5x/gemini

quinlanlab.org

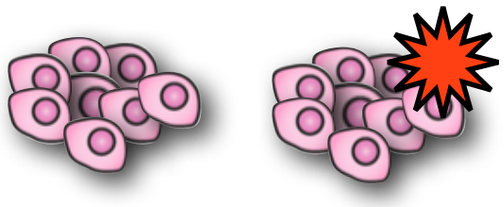
Typical genetics study designs



"case" v. "control"

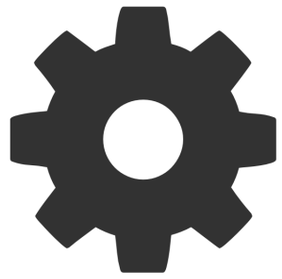


family studies



cancer genomics

Sequence DNA

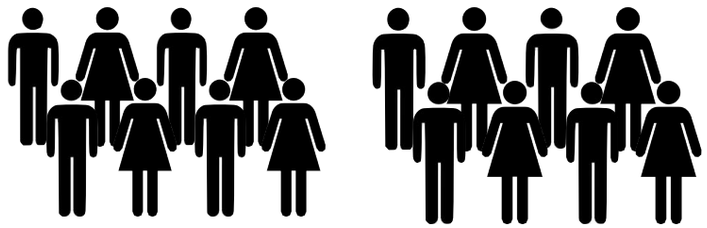


Align.
Remove artifacts.
Etc., etc.
Find genetic variants.

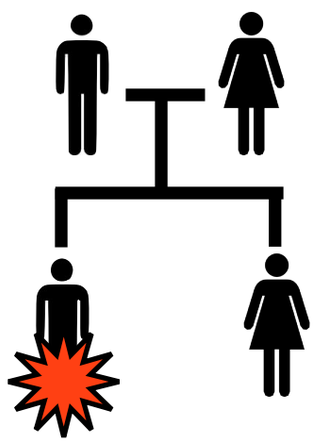
Typical genetics study designs

disease

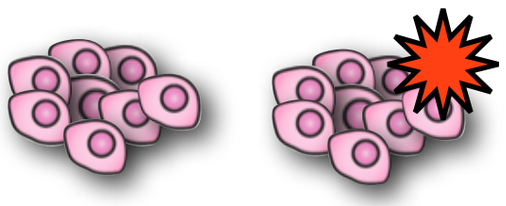
no disease



“case” v. “control”



family studies

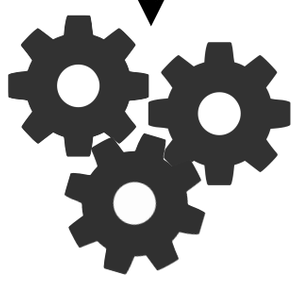


normal tumor

cancer genomics

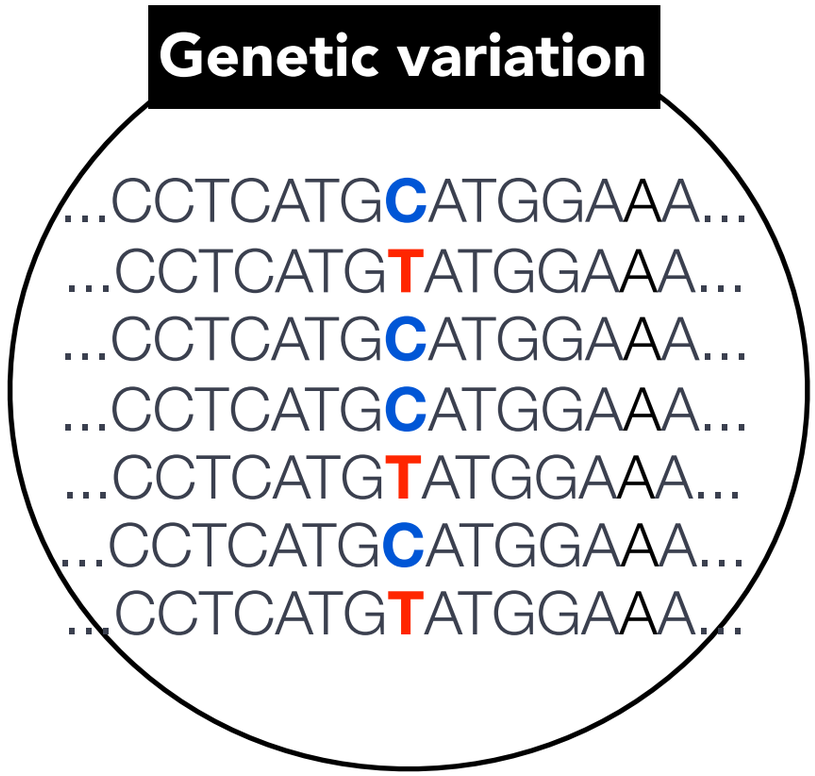
Sequence DNA

Align.
Remove artifacts.
Etc., etc.
Find genetic variants.

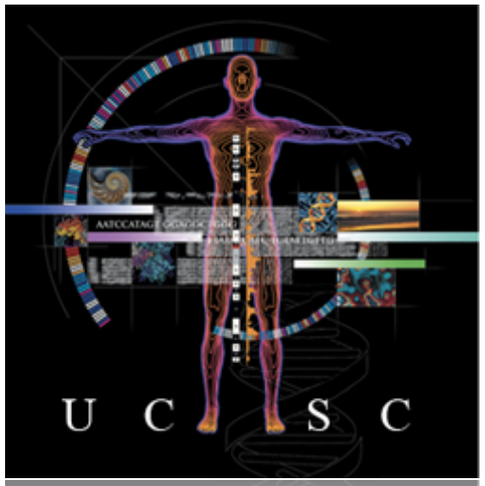


Interpret.
Prioritize.
Repeat.

Analytical challenges: data integration

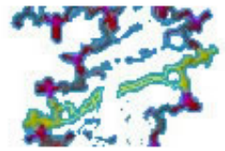


Analytical challenges: data integration



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

dbSNP
Short Genetic Variations



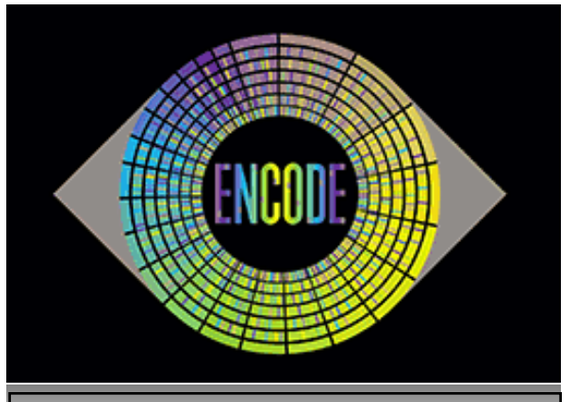
ClinVar

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

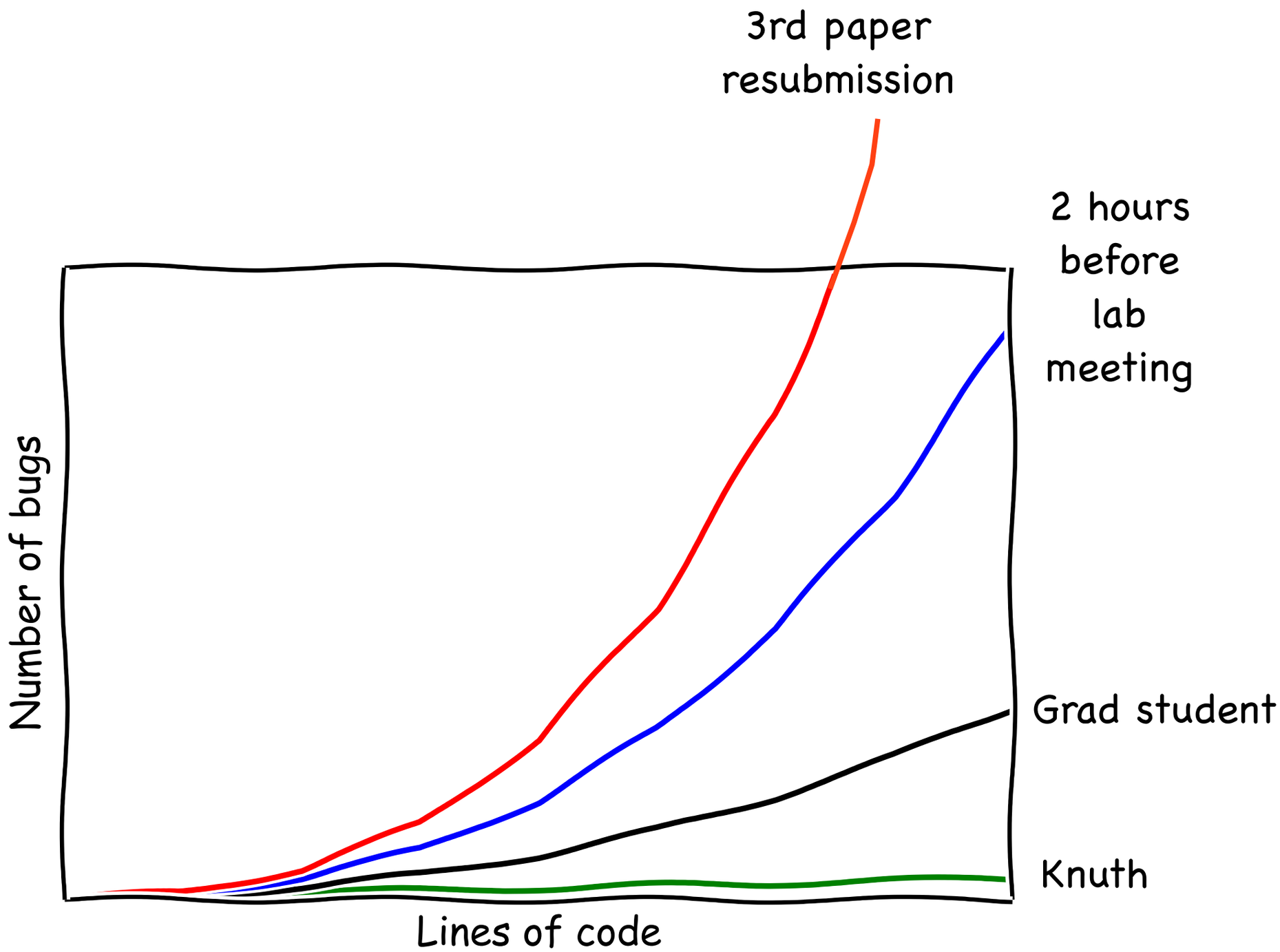
KEGG Kyoto Encyclopedia of Genes and Genomes

HGMD®

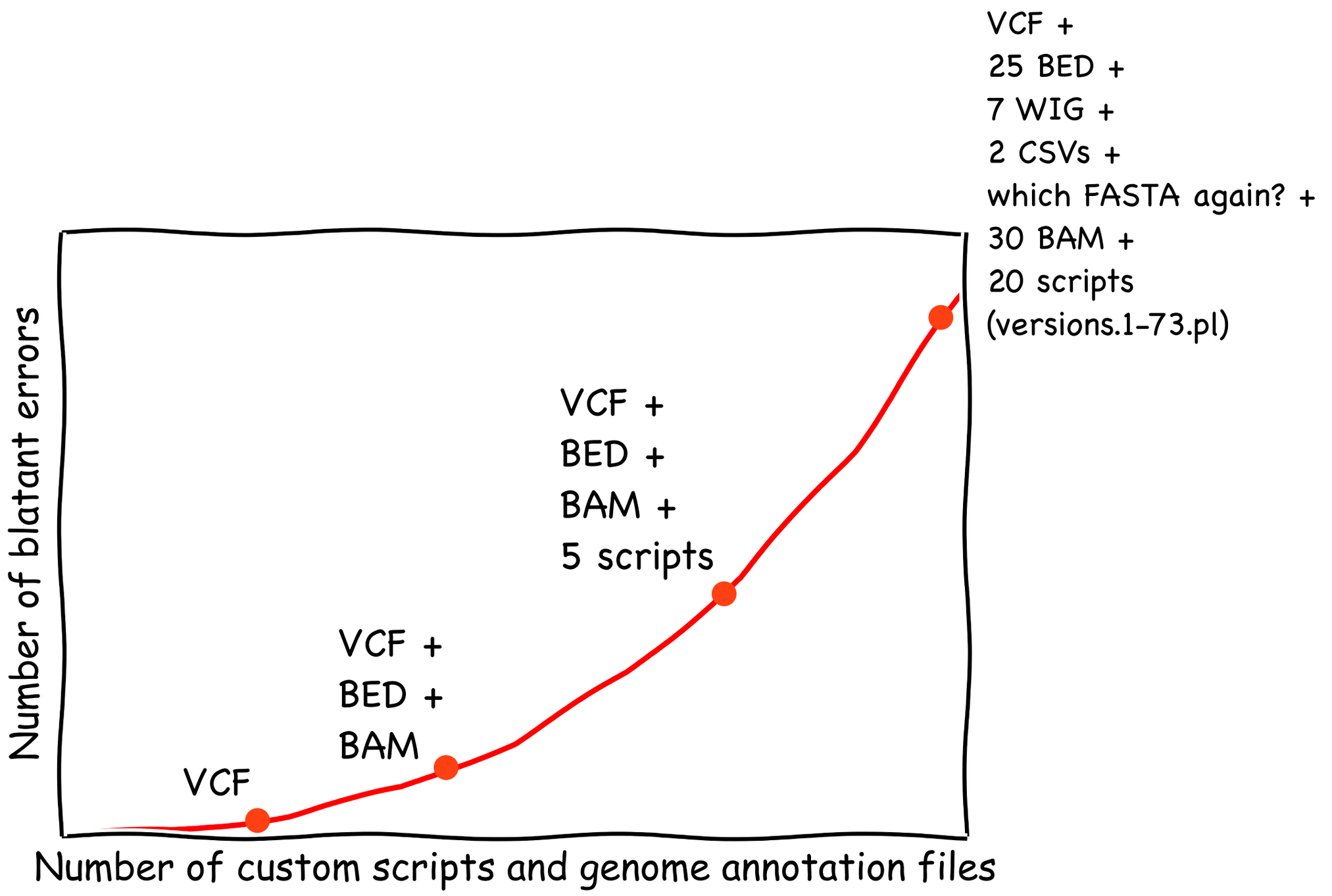
Human Protein Reference Database

More code, more problems

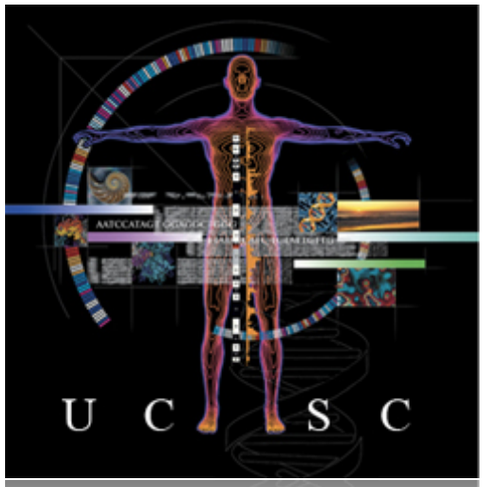
More code, more problems



A genomics corollary

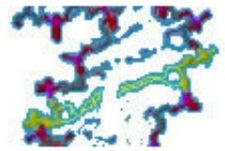


Directly Integrate variants with annotations



Conservation
Repeat elements
Genome Gaps
Cytobands
Gene annotations
“Mappability”
DeCIPHER
ISGA

dbSNP
Short Genetic Variations



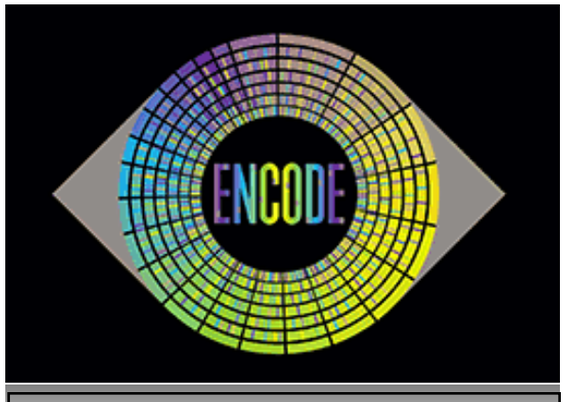
ClinVar

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

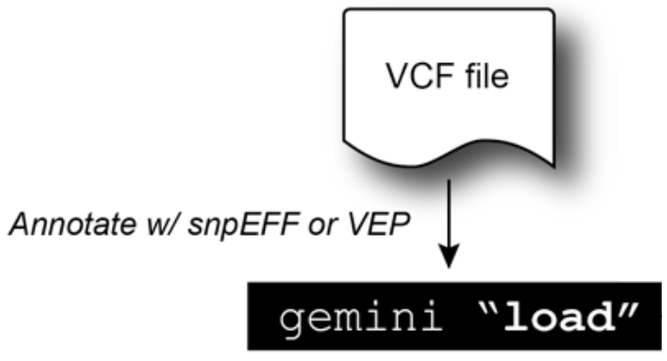
Pfam

Kyoto Encyclopedia of Genes and Genomes
KEGG

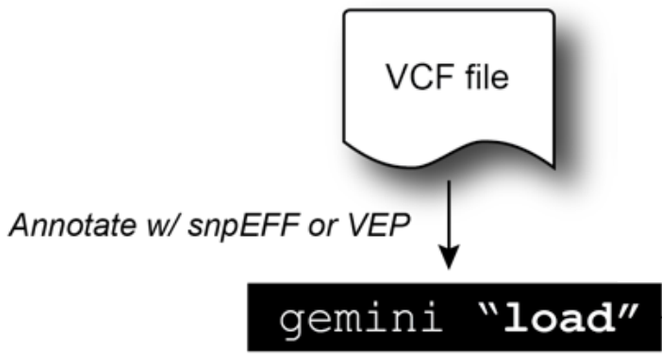
HGMD®

Human Protein Reference Database

The GEMINI framework



The GEMINI framework



Parallelized loading on SGE, Torque, LFS

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.

The GEMINI framework

VCF file

Annotate w/ snpEFF or VEP

`gemini "load"`

Parallelized loading on SGE, Torque, LFS

User-defined annotations

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.

The GEMINI framework

VCF file

Annotate w/ snpEFF or VEP

`gemini "load"`

Parallelized loading on SGE, Torque, LFS

User-defined annotations

Column name	sample 1	sample 2	sample 3	⋯	sample N
gts	G/G	A|G	A|A	⋯	./.
gt_types	0	1	3	⋯	2
gt_phases	0	1	1	⋯	0
gt_depths	73	91	53	⋯	4

Access to (and filtering upon) 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

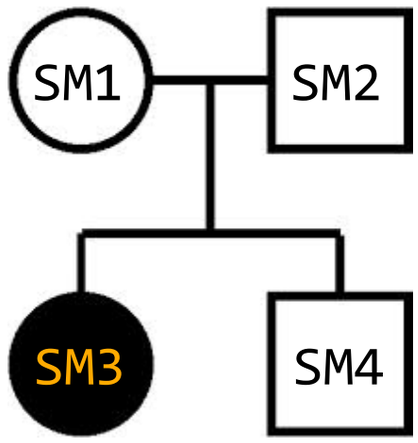
Tracks the GEMINI software version that was used to create the database.

We can exploit sample *relationships* (not in VCF)

				SM1	SM2	SM3	SM4
CHR1	123	A	G	**A/G**	**A/G**	**G/G**	**A/A**

We can exploit sample *relationships* (not in VCF)

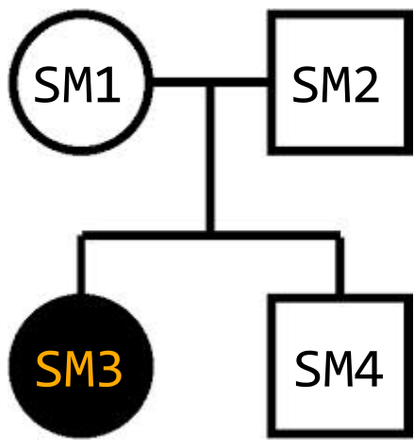
				SM1	SM2	SM3	SM4
CHR1	123	A	G	**A/G**	**A/G**	**G/G**	**A/A**



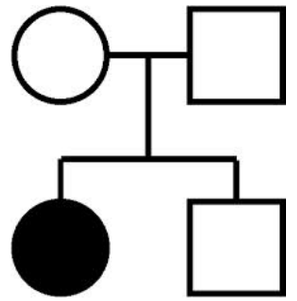
PED file

We can exploit sample *relationships* (not in VCF)


```
                    SM1   SM2   SM3   SM4
CHR1 123   A   G    A/G   A/G   G/G   A/A
```



PED file



PED file

Disease models

Inheritance patterns

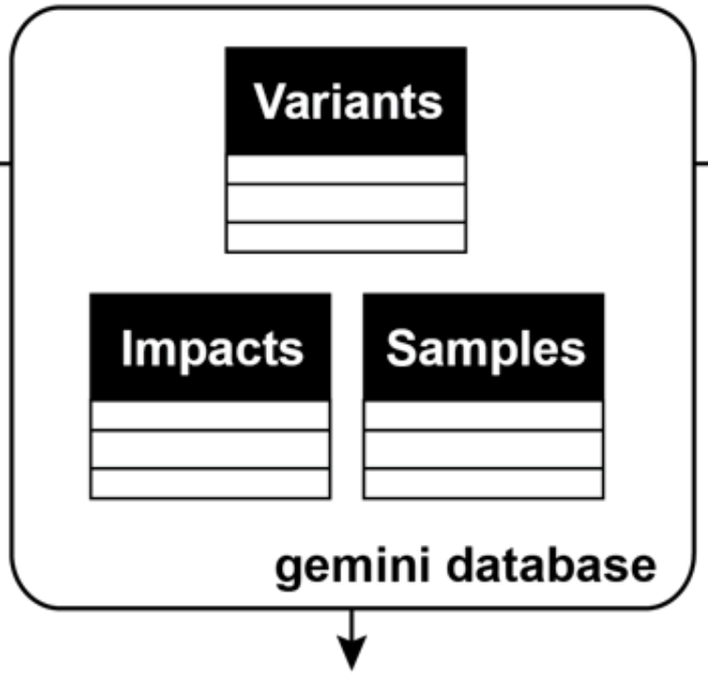
Transmission tests

Phasing

Mining variation with GEMINI

ad hoc data exploration

```
gemini query
--query
"select chrom, start, end,
        ref, alt, gene,
        impact, aaf, gts.kid
 from variants
 where in_dbsnp = 0
 and aaf < 0.01
 and is_lof  = 1
 and my_disease_regions = 1"
--gt-filter
"gt_types.mom == HET
 and
 gt_types.dad == HET
 and
 gt_types.proband == HOM_ALT"
```



Framework for new tools

- Burden tests
- Population genetics
- Pedigree studies
- Haplotype analysis
- *Custom* tools and new methods

Built-in tools and analyses

Tool	*Description*
`region`	*extract variants spec. genomic intervals or genes*
`stats`	*compute variant statisics (SFS, Ts/Tv, counts, etc.)*
`annotate`	*add new columns based on custom annotations*
`windower`	*compute variant statistics across genome "windows"*
`comp_hets`	*identify candidate compund heterozygotes*
`pathways`	*maps genes and variants to KEGG pathways*
`lof_sieve`	*prioritize candidate loss-of-function variants*
`interact`	*find protein interactions for genes/variants/samples*
`auto_rec`	*identify variants meeting an autosomal recessive model*
`auto_dom`	*identify variants meeting an autosomal dominant model*
`de_novo`	*identify candidate de novo mutations*
`browser`	*launch the interactive gemini web browser interface*

Example analyses: *ad hoc* exploration

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

```
$ gemini query -q “select * from variants
                   where is_lof = 1
                   and aaf <= 0.01)”
                   my.gemini.db
```

Example analyses: *ad hoc* exploration

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

```
$ gemini query -q “select * from variants
                   where is_lof = 1
                   and aaf <= 0.01)”
                   my.gemini.db
```

2. Find all LoF variants and show me genotypes for all *affected samples having blue eyes*

```
$ gemini query -q “select * from variants
                     where is_lof = 1”
                --sample-filter "phenotype=1 and eye_color='blue'"
                     my.gemini.db
```

Automatically built from custom sample annotations in PED file

Example analyses: *ad hoc* exploration

3. Find variants with a known clinical phenotype

```
$ gemini query -q “select * from variants
                   where (in_omim = 1
                          or clinvar_disease_name is not NULL)”
```

Example analyses: *ad hoc* exploration

3. Find variants with a known clinical phenotype

```
$ gemini query -q “select * from variants
                   where (in_omim = 1
                          or clinvar_disease_name is not NULL)”
                  --gt-filter “gt_types.proband == HET
                  and
                  gt_types.mommy == HOM_REF
                  and
                  gt_types.daddy == HOM_REF”
```

Example analyses: *ad hoc* exploration

3. Find variants with a known clinical phenotype

```
$ gemini query -q “select * from variants
                   where (in_omim = 1
                          or clinvar_disease_name is not NULL)”
                  --gt-filter “gt_types.proband == HET
                              and
                              gt_types.mommy == HOM_REF
                              and
                              gt_types.daddy == HOM_REF”
                  --min-kindreds 3
```

Example analyses: *ad hoc* exploration

3. Find variants with a known clinical phenotype

```
$ gemini query -q “select * from variants
                   where (in_omim = 1
                          or clinvar_disease_name is not NULL)”
                  --gt-filter “gt_types.proband == HET
                              and
                              gt_types.mommy == HOM_REF
                              and
                              gt_types.daddy == HOM_REF”
                  --min-kindreds 3
                  --region chr1:149000000-153000000
```

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
```

Example: autosomal recessive candidates

```
gemini auto_recessive
        --columns “chrom, start, end, ref,
                   alt, gene, impact”
        --filter  “impact_severity != ‘LOW’”
        --min-kindreds 2
        --depth 40
        disease.gemini.db
```

request specific columns from database

restrict to higher impact variants

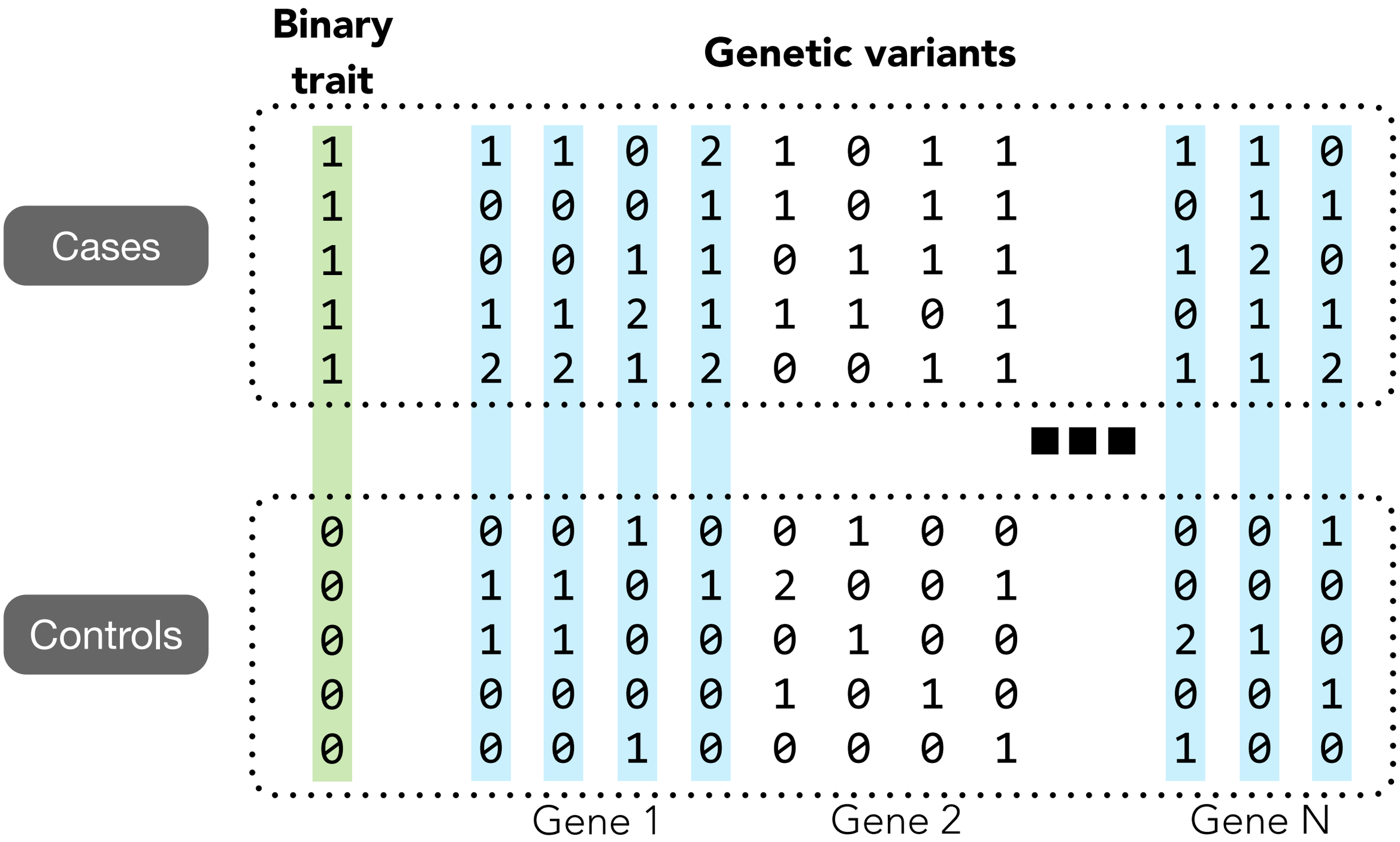
found in >= 2 families

each sample must have >= 40 reads

fam	genotypes (F,M,C)	depths	gene	chrom	start	end	ref	alt	impact
1	T/C,T/C,C/C	69,48,70	WDR37	chr10	1142207	1142208	T	C	stop_loss
2	T/C,T/C,C/C	59,49,82	WDR37	chr10	1142207	1142208	T	C	stop_loss

Ongoing research.

Burden tests




```
> gemini burden --calpha my.db --min-aaf 0.0 --max-aaf 0.01
```

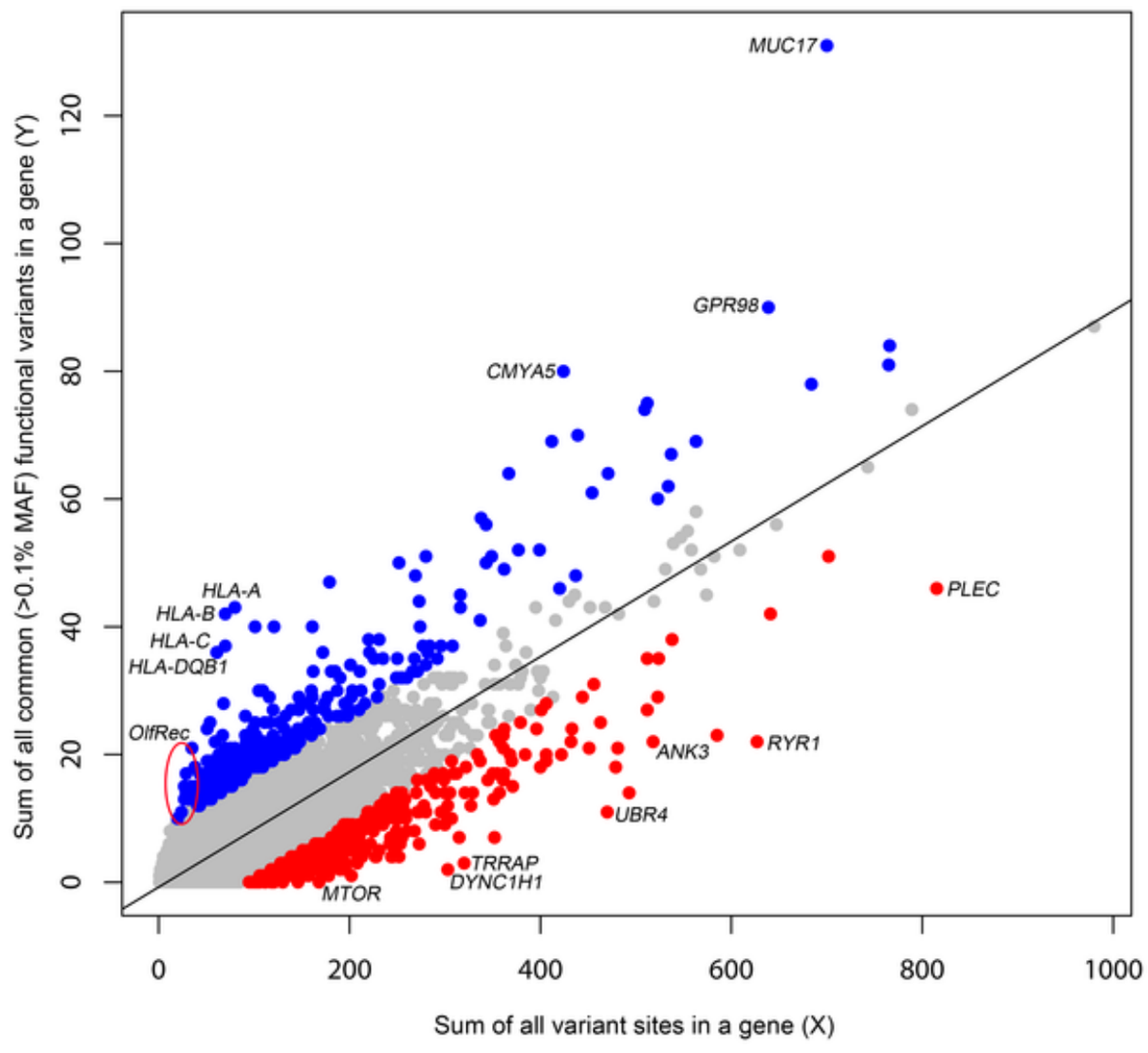
Gene prioritization

OPEN ACCESS Freely available online

PLOS | GENETICS

Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes

Slavé Petrovski[1,2]*, Quanli Wang[1], Erin L. Heinzen[1,3], Andrew S. Allen[1,4], David B. Goldstein[1]*



An API for data viz. and tool dev.

query

↓ JSON

JBrowse

Cytoscape

Galaxy/Trackser

IOBIO

Summary

- Flexible framework for mining genetic variation.
- Integrates important genome annotations.
- Query access to individual genotypes
- Extensible for new analyses and tool dev.
- Free. Open source. github.com/arq5x/gemini
- Well documented. gemini.readthedocs.org
- Extensible, portable, & reproducible

OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

GEMINI: Integrative Exploration of Genetic Variation and Genome Annotations

Umadevi Paila[1], Brad A. Chapman[2], Rory Kirchner[2], Aaron R. Quinlan[1]*

Acknowledgements

Uma Paila*

Postdoctoral Fellow

github.com/udp3f

Brad Chapman

Oliver Hofmann

Rory Kirchner

National Human Genome Research Institute

R01HG006693-01