The Elizabeth H.
and James S. McDonnell III
McDonnell Genome Institute
at Washington University

You've got the data…now what?

Karyn Meltz Steinberg

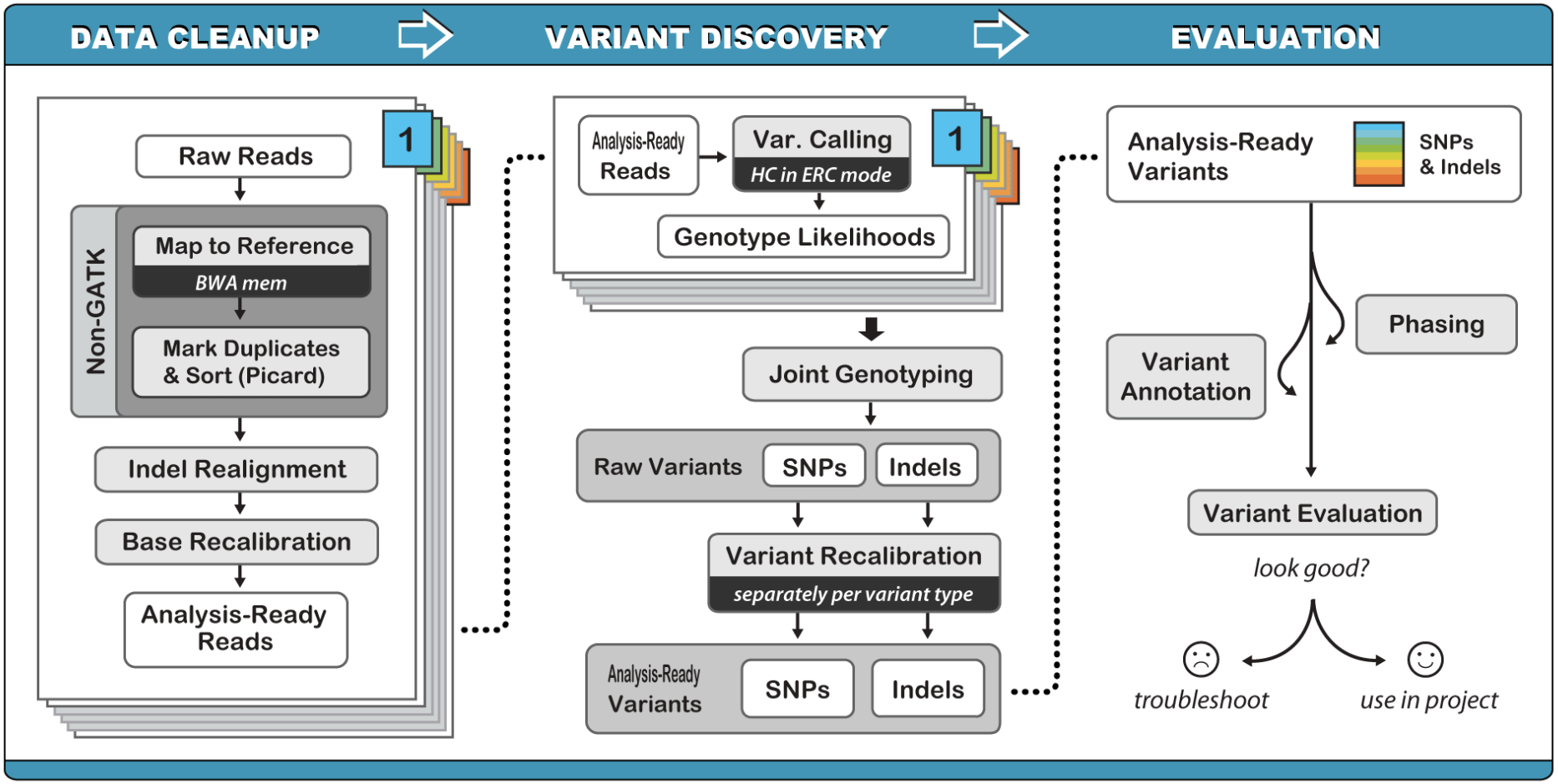
November 17, 2016

HLA course

Outline

- Variant calling and filtering
- Variant Evaluation
- Annotation and GEMINI
- Identifying rare variants in large cohorts

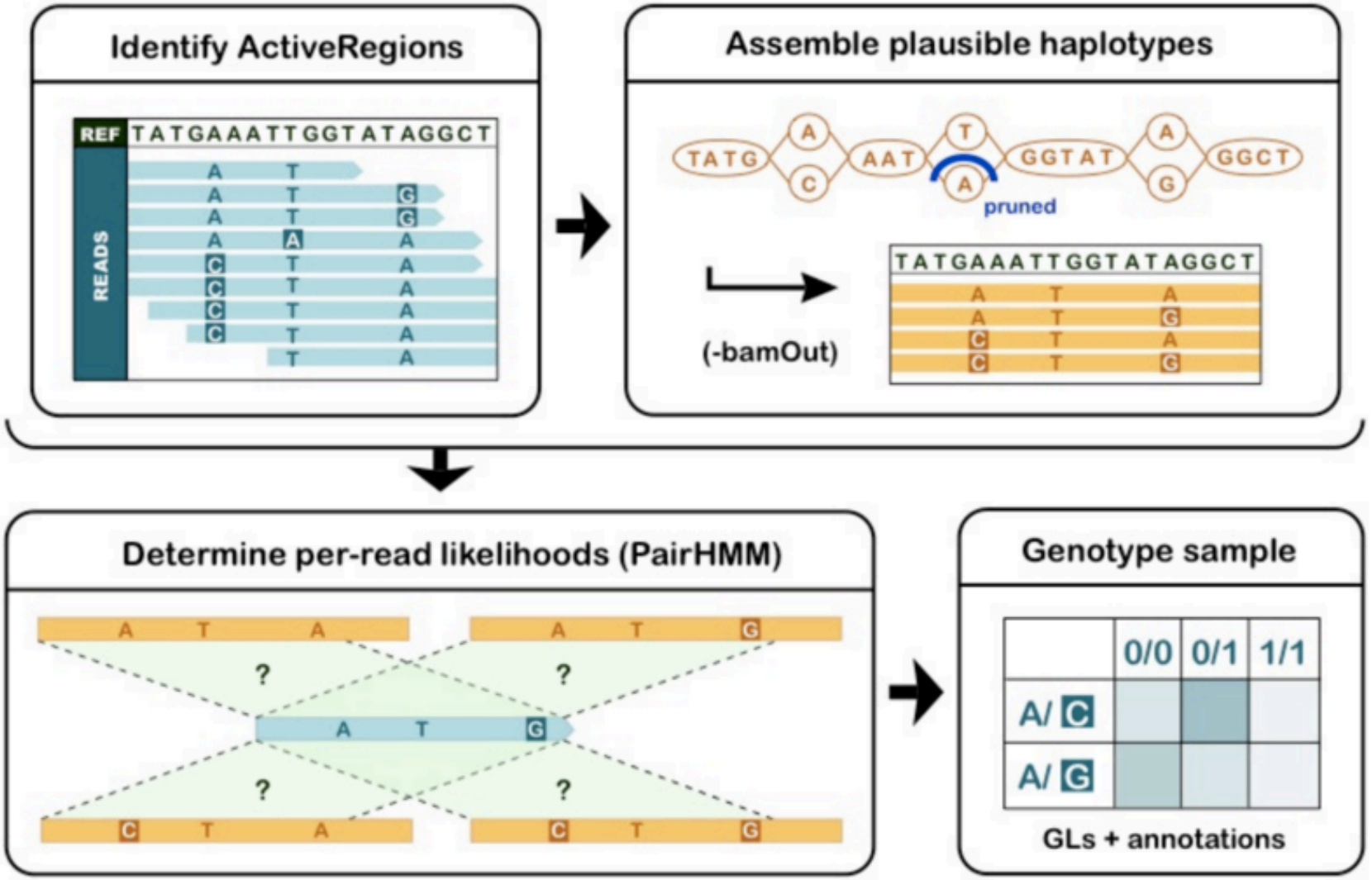
Variant calling and filtering


DATA CLEANUP
VARIANT DISCOVERY
EVALUATION
1
Raw Reads
Non-GATK
Map to Reference
BWA mem
Mark Duplicates & Sort (Picard)
Indel Realignment
Base Recalibration
Analysis-Ready Reads
1
Analysis-Ready Reads
Var. Calling
HC in ERC mode
Genotype Likelihoods
Joint Genotyping
Raw Variants
SNPs
Indels
Variant Recalibration
separately per variant type
Analysis-Ready Variants
SNPs
Indels
Analysis-Ready Variants
SNPs & Indels
Phasing
Variant Annotation
Variant Evaluation
look good?
troubleshoot
use in project

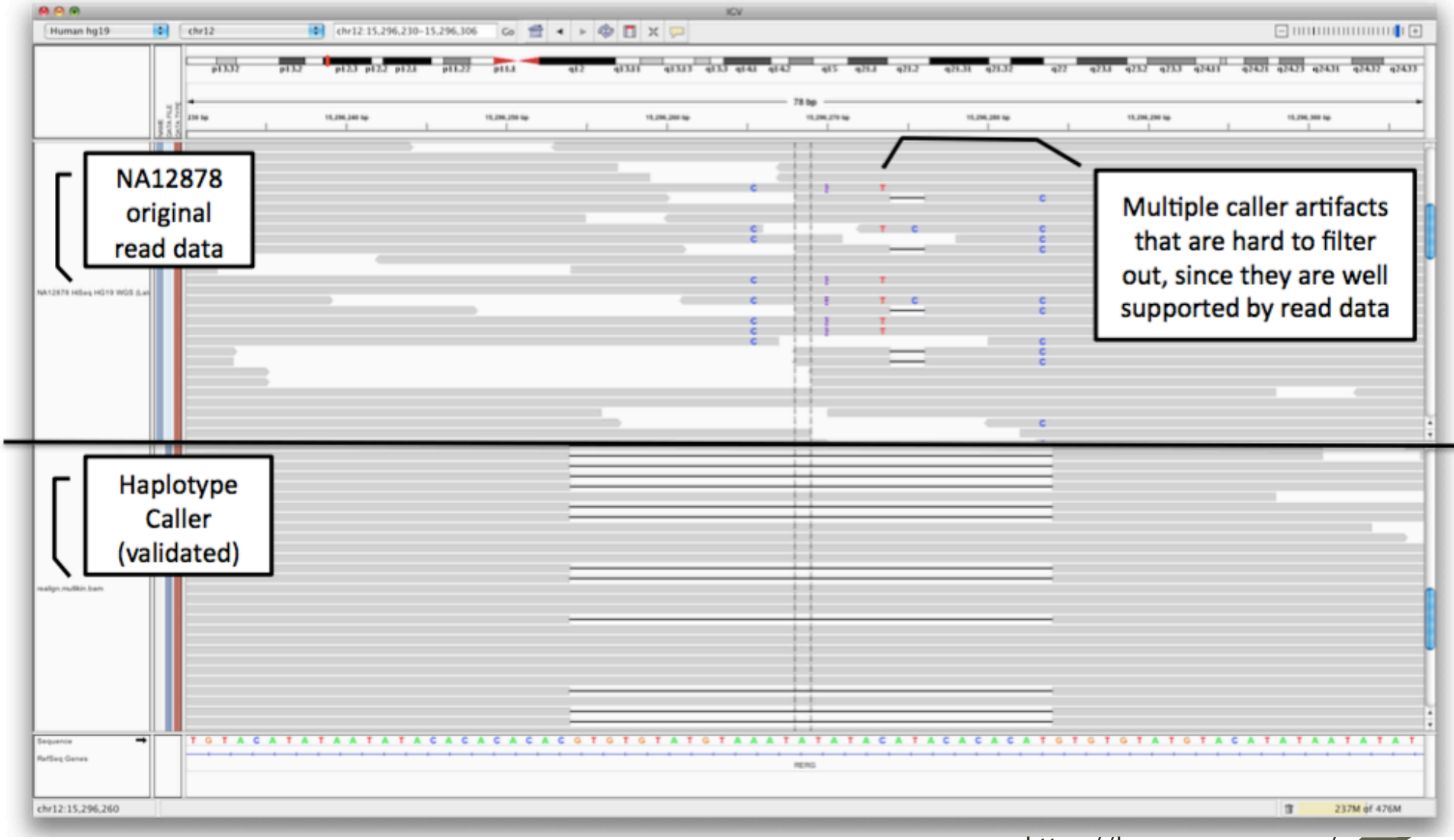
Different methods for calling variants

- Call SNVs and indels separately by considering each variant locus
  - Very fast
  - Assumes bases are independent
- Call SNVs and indels simultaneously via Bayesian genotype likelihood model
  - More computationally intensive
  - GATK UnifiedGenotyper
- Call SNVs, indels and SVs simultaneously by performing a local de novo assembly
  - More computationally intensive
  - More accurate—gets rid of many false positives especially indels
  - GATK HaplotypeCaller

HC method illustrated



Why all the steps?

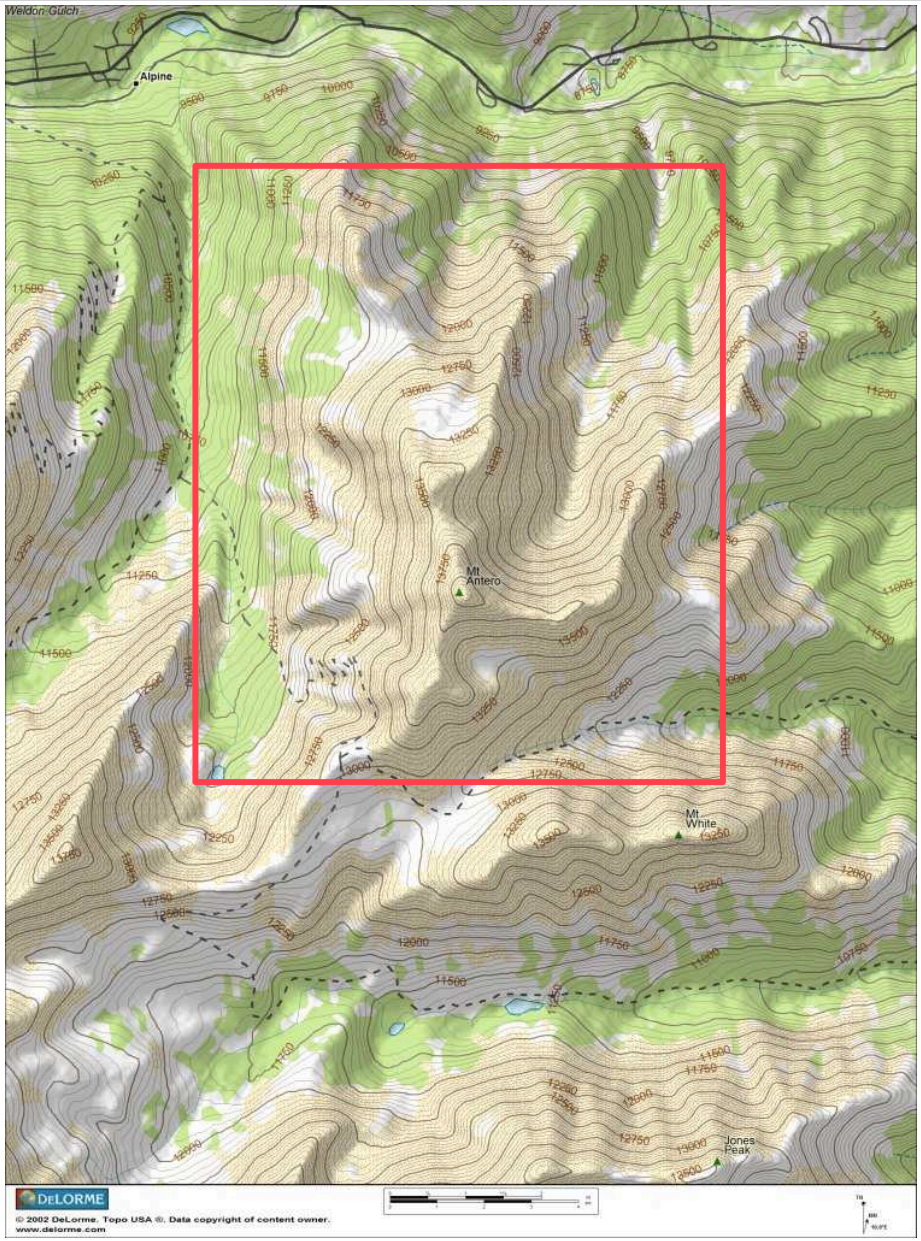


https://kccg.garvan.org.au/

GATK recommended filters

- SNPs
  - QD < 2.0 (variant quality/depth of non-ref samples)
  - MQ < 40.0 (Mapping quality)
  - FS > 60.0 (Phred score Fisher's test pvalue for strand bias)
  - SOR > 3.0 (Strand odds ratio, aims to evaluate whether there is strand bias in the data—updated form of FET)
  - MQRankSum < -12.5 (mapping quality of reference reads vs alt reads)
  - ReadPosRankSum < -8.0 (distance of alt reads from end of the read)
- Indels
  - QD < 2.0
  - ReadPosRankSum < -20.0
  - InbreedingCoeff < -0.8
  - FS > 200.0
  - SOR > 10.0

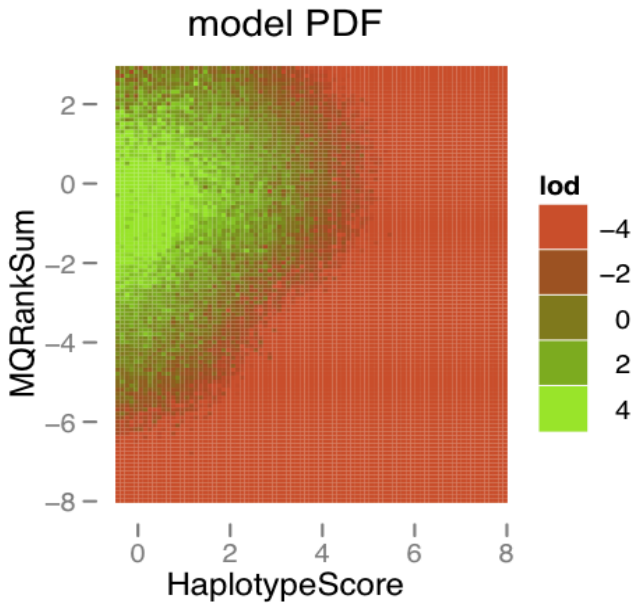
http://gatkforums.broadinstitute.org/discussion/3225/how-can-i-filter-my-callset-if-i-cannot-use-vqsrrecalibrate-variants

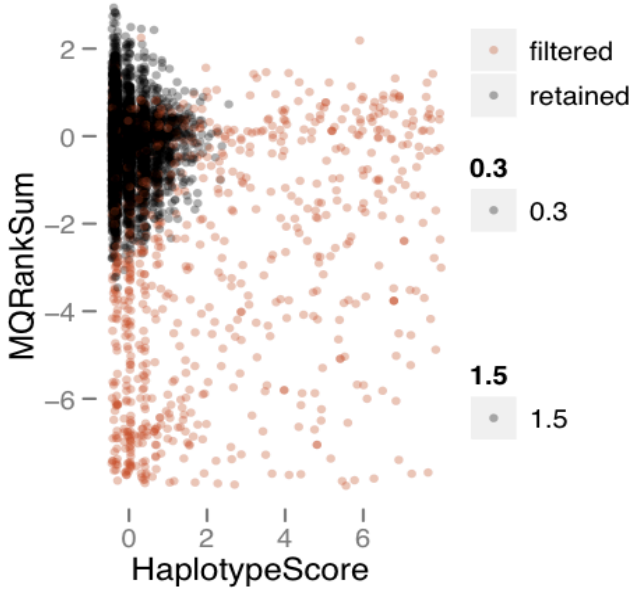

Weldon Gulch
Alpine
Mt Antero
Mt White
Jones Peak
DeLORME
© 2002 DeLorme. Topo USA ®. Data copyright of content owner.
www.delorme.com

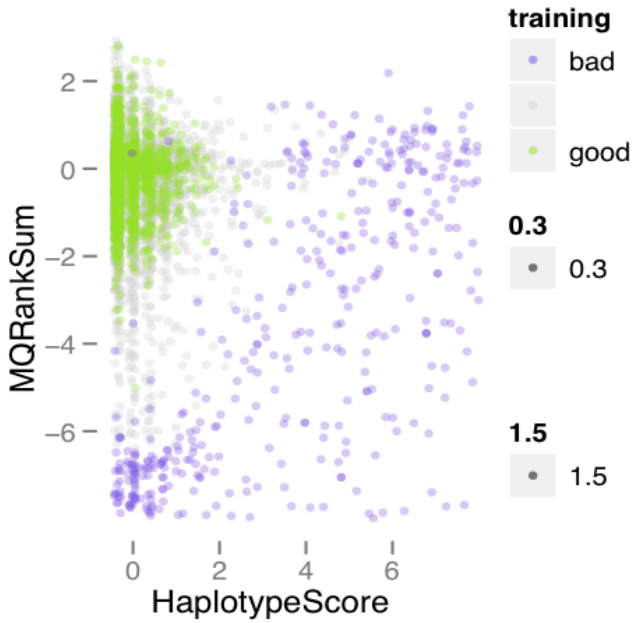
VQSR

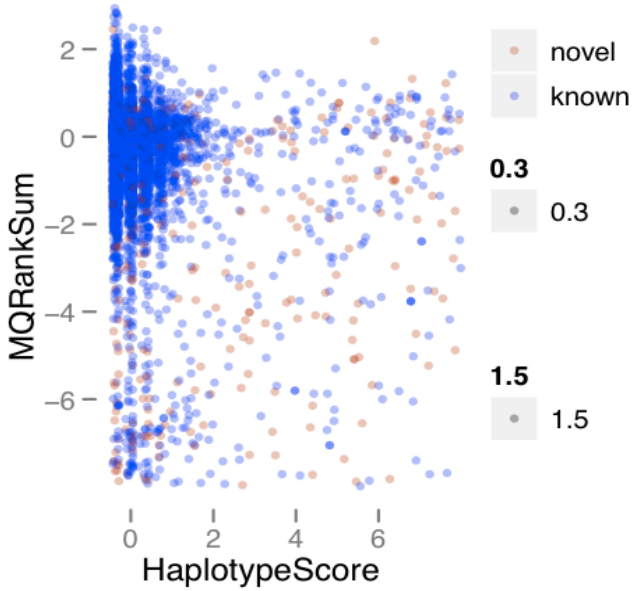
- VariantRecalibrator
  - Gaussian mixture model by looking at the annotation values over a high quality subset of input call set and then evaluate all variants
- ApplyRecalibration
  - Apply model parameters to each variant producing a recalibrated VCF file

- Run these separately for SNPs and indels
- Need to have some resource files
  - resource:hapmap,known=false,training=true,truth=true,prior=15.0 HAPMAP
  - resource:omni,known=false,training=true,truth=true,prior=12.0 OMNI
  - resource:1000G,known=false,training=true,truth=false,prior=10.0 G1000
  - resource:dbsnp,known=true,training=false,truth=false,prior=2.0 DBSNP

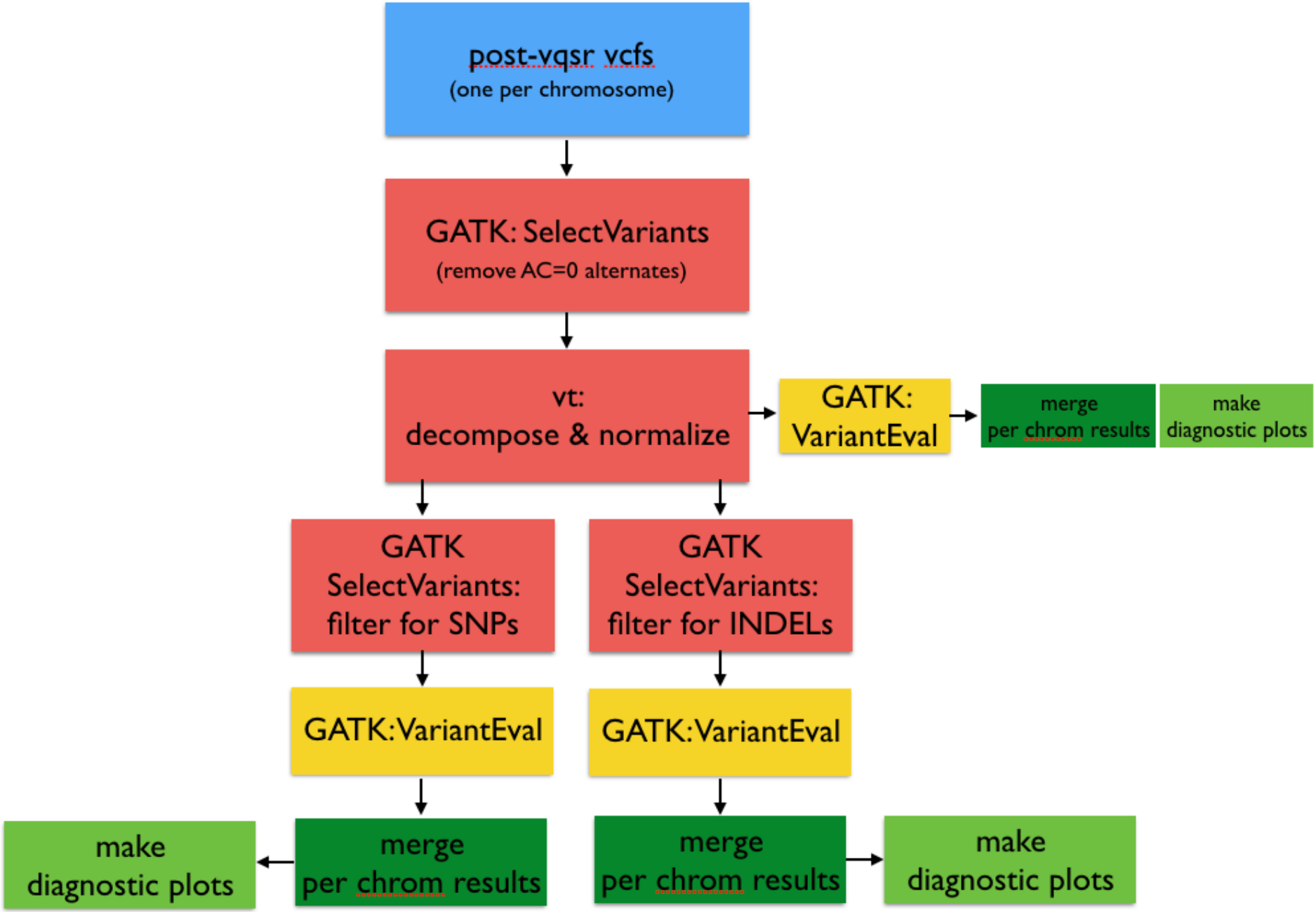

model PDF
MQRankSum
HaplotypeScore
lod
−4
−2
0
2
4


MQRankSum
HaplotypeScore
filtered
retained
0.3
0.3
1.5
1.5

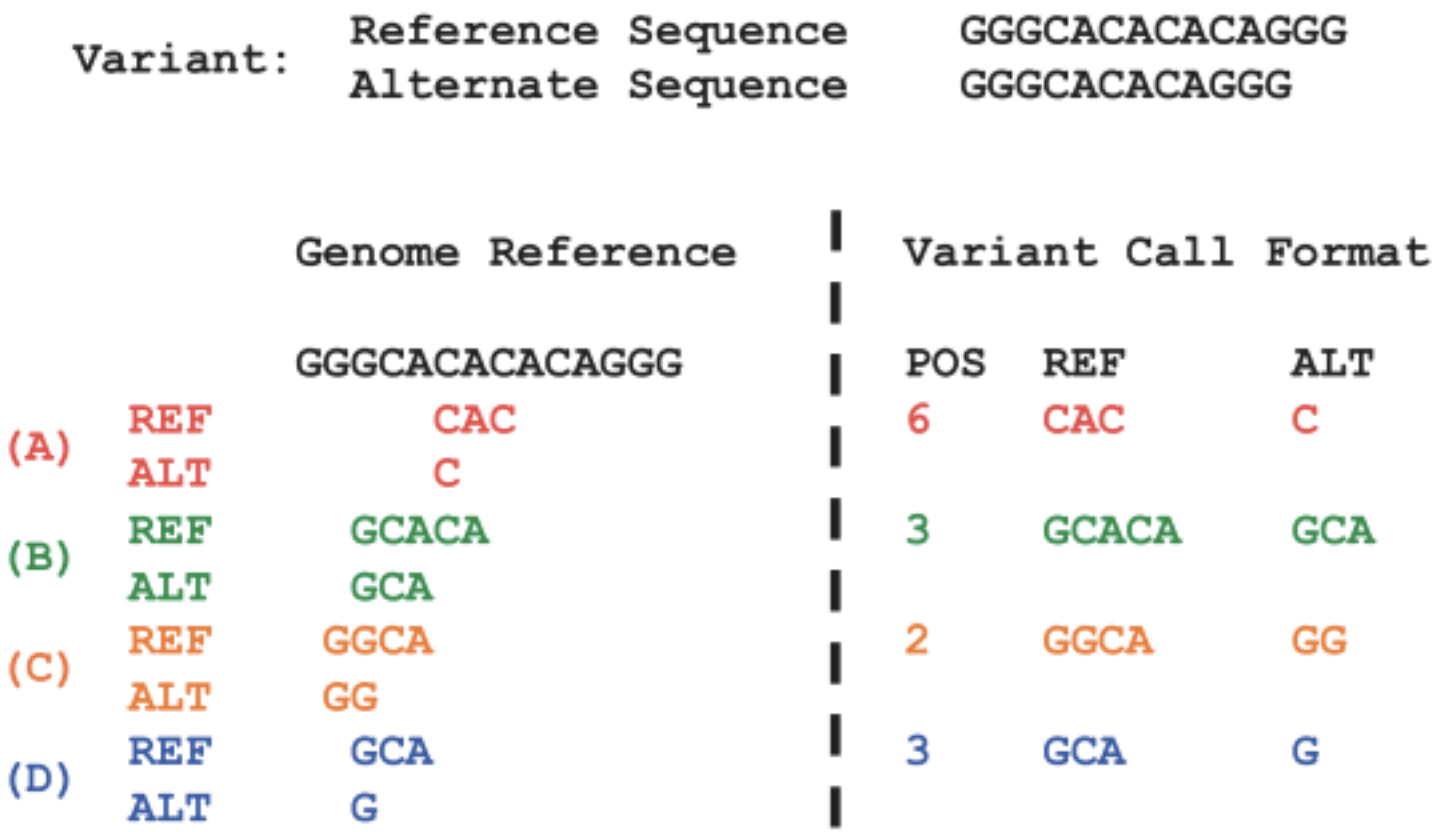

MQRankSum
HaplotypeScore
training
bad
good
0.3
0.3
1.5
1.5


MQRankSum
HaplotypeScore
novel
known
0.3
0.3
1.5
1.5

Variant Evaluation


post-vqsr vcfs
(one per chromosome)
GATK: SelectVariants
(remove AC=0 alternates)
vt:
decompose & normalize
GATK:
VariantEval
merge
per chrom results
make
diagnostic plots
GATK
SelectVariants:
filter for SNPs
GATK
SelectVariants:
filter for INDELs
GATK:VariantEval
GATK:VariantEval
make
diagnostic plots
merge
per chrom results
merge
per chrom results
make
diagnostic plots

Decomposing and normalizing indels



doi: 10.1093/bioinformatics/btv112

Normalize indels

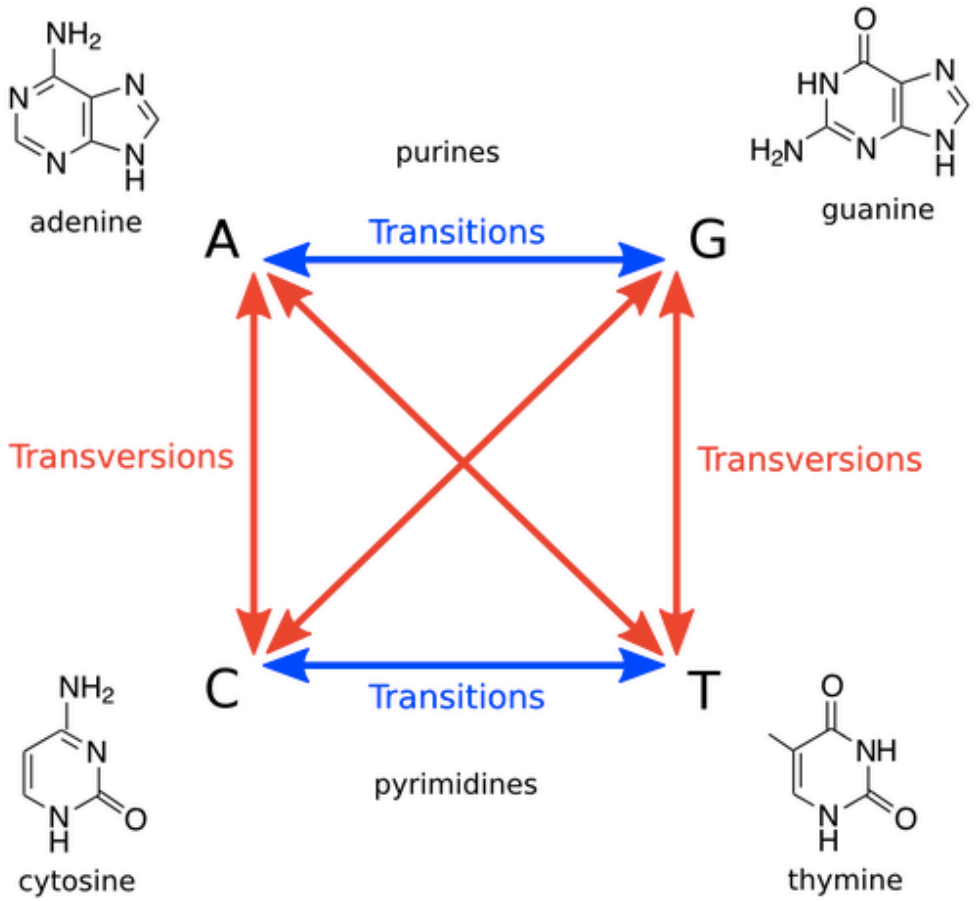
Algorithm 1 Normalize a VCF entry

Input: A VCF entry and the reference genome sequence.
Output: A normalized VCF entry

```
1 : do
2 : if all alleles end with same nucleotide then
3 : truncate the rightmost nucleotide of each allele
4 : if any allele is length zero then
5 : extend all alleles by 1 nucleotide to the left
6 : while changes made in the VCF entry in the loop
7 : while all alleles start with same nucleotide and length ≥2
      do
8 : truncate the leftmost nucleotide of each allele
9 : end while
10 : return the VCF entry
```

doi: 10.1093/bioinformatics/btv112

Transition/transversion ratio (Ti/Tv)



Random = 0.5

WGS = 2.0-2.1

Exome = 3-3.5

Hardy Weinberg Equilibrium

$$(p + q)^2 = p^2 + 2pq + q^2 = 1$$

Where:

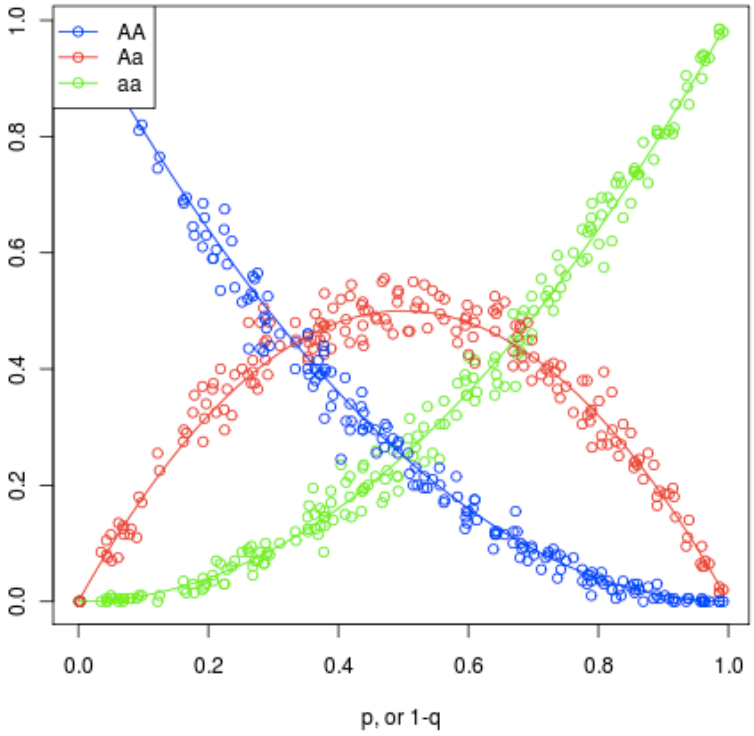
p = the frequency of allele A

q = the frequency of allele a

p^2 = the frequency of individual AA

q^2 = the frequency of individual aa

2pq = the frequency of individual Aa

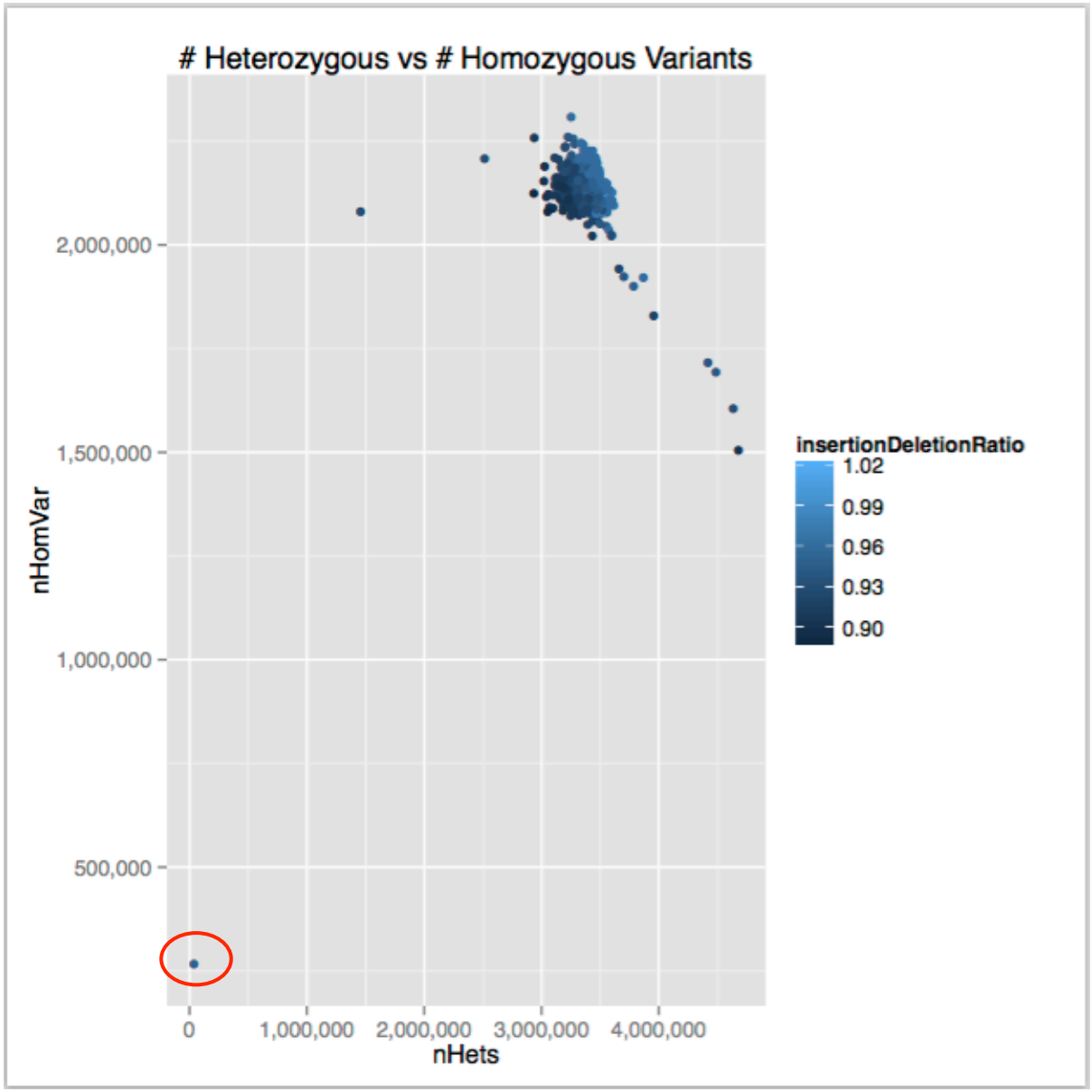


https://bayesianbiologist.com/2011/06/13/using-simulation-to-demonstrate-theory-hardy-weinberg-equilibrium/

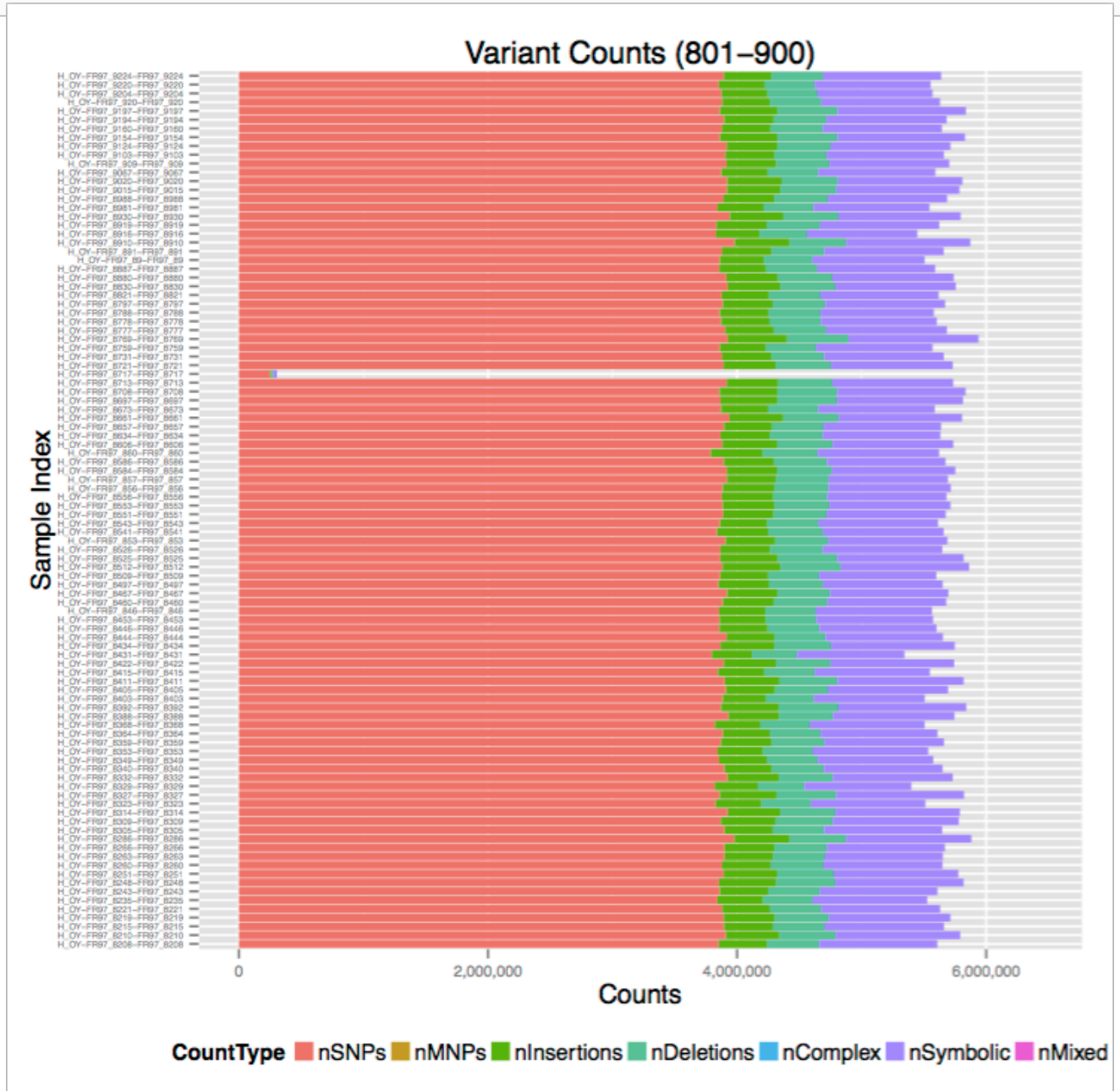
Assumptions of Hardy Weinberg Equilibrium

- organisms are diploid
- only sexual reproduction occurs
- generations are non overlapping
- mating is random
- population size is infinitely large
- allele frequencies are equal in the sexes
- there is no migration, mutation or selection

Heterozygous and homozygous variants



Missingness per sample and variant site



Other things to evaluate

- Sex check
- Novel vs Known sites
  - dbSNP/1KG concordance (should be >98%)
- Number of singletons per sample
  - excess singletons may indicate a bad sample
- PCA
  - Is there significant population stratification?
    - Do you expect this?
    - What could be causing? Batch effects?

Annotation and GEMINI

Annotation

- Based on conservation
  - SIFT
  - PolyPhen
  - GERP
  - Grantham
  - phyloP
  - CADD
  - …
- Based on protein function
  - SNPEff
  - VEP
  - …

SIFT

- Sorting Intolerant From Tolerant
- SIFT predicts whether an amino acid substitution affects protein function
- Given a protein sequence, SIFT chooses related proteins and obtains an alignment of these proteins with the query. Based on the amino acids appearing at each position in the alignment, SIFT calculates the probability that an amino acid at a position is tolerated conditional on the most frequent amino acid being tolerated. If this normalized value is less than a cutoff, the substitution is predicted to be deleterious
  - Scores range from 0-1
  - <0.05 → deleterious

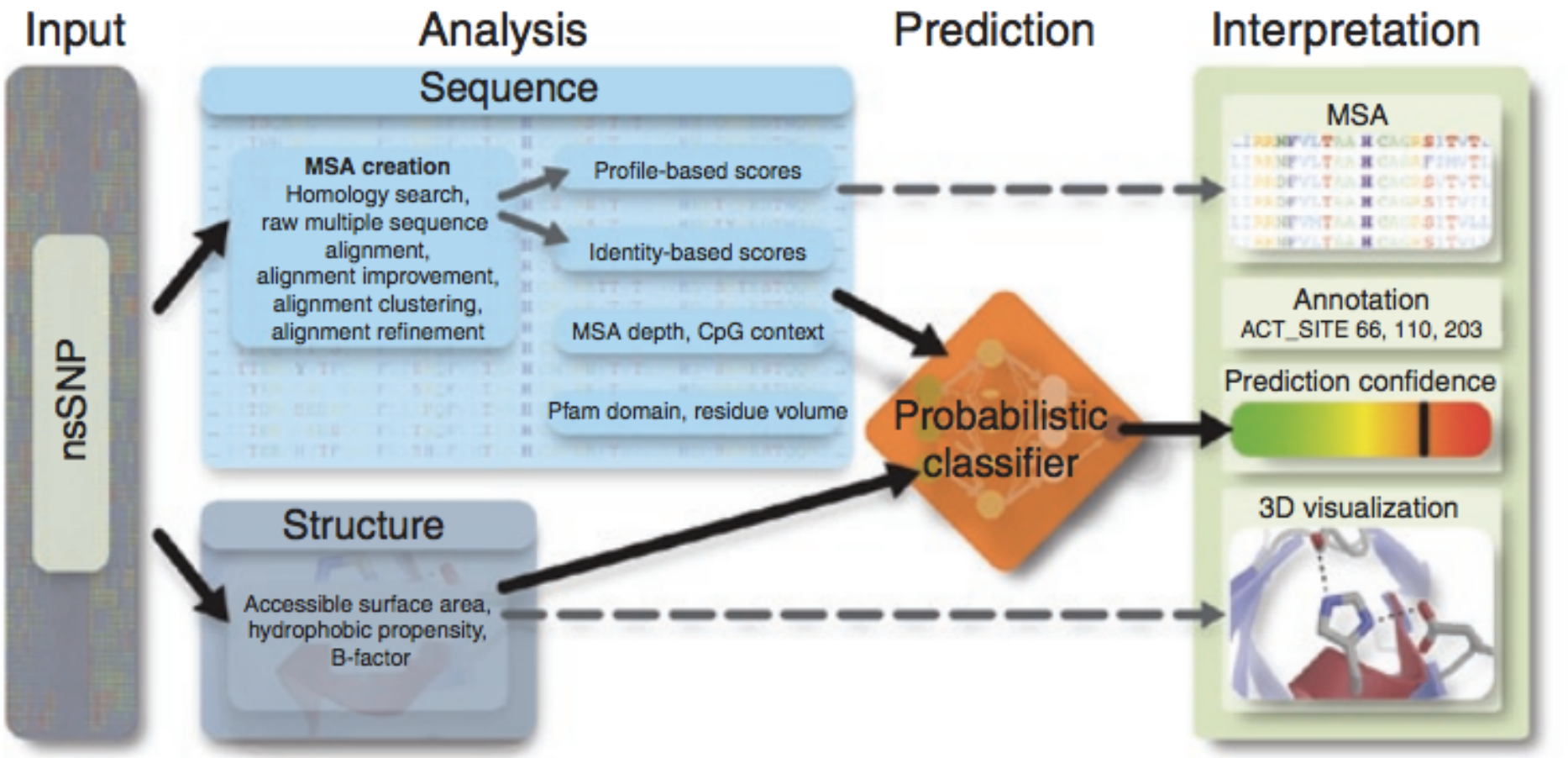
http://sift.jcvi.org/

PolyPhen

- **Poly**morphism **Phen**otyping
  - Now on v2
- Predicts possible impact of amino acid substitution on structure and function of human protein using physical and comparative considerations
  - 8 sequence based predictive features
    - Multiple sequence alignment
    - Comparison of ancestral and mutant allele
  - 3 structure based predictive features

http://genetics.bwh.harvard.edu/pph2/

PolyPhen

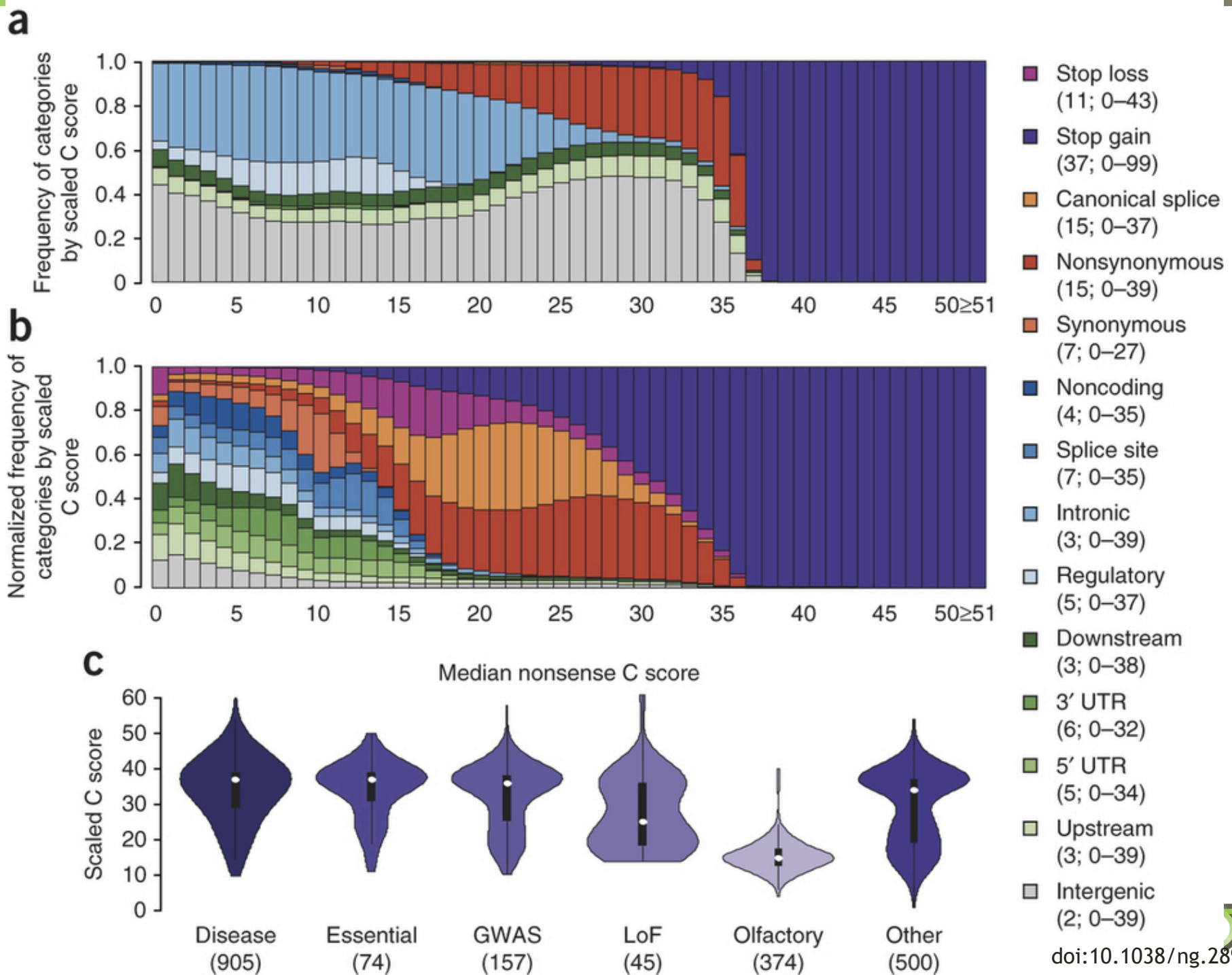


Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. Nat Methods 7(4):248-249 (2010).

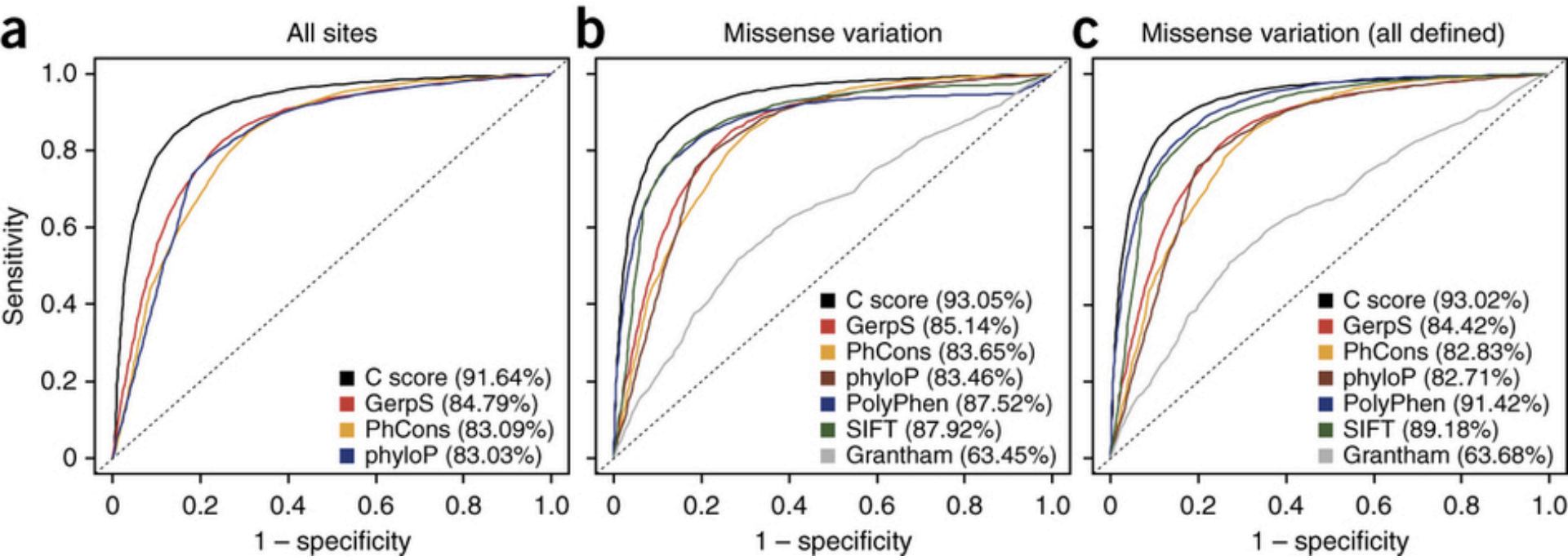
CADD

- Combined Annotation Dependent Depletion
- Score deleteriousness of single nucleotide variants as well as indels in human genome
- Integrates multiple annotations into one metric by contrasting variants that survived natural selection with simulated mutations
- C-scores strongly correlate with allelic diversity, pathogenicity of coding and non-coding variants
- Ranks causal variants within individual genome sequences

http://cadd.gs.washington.edu/

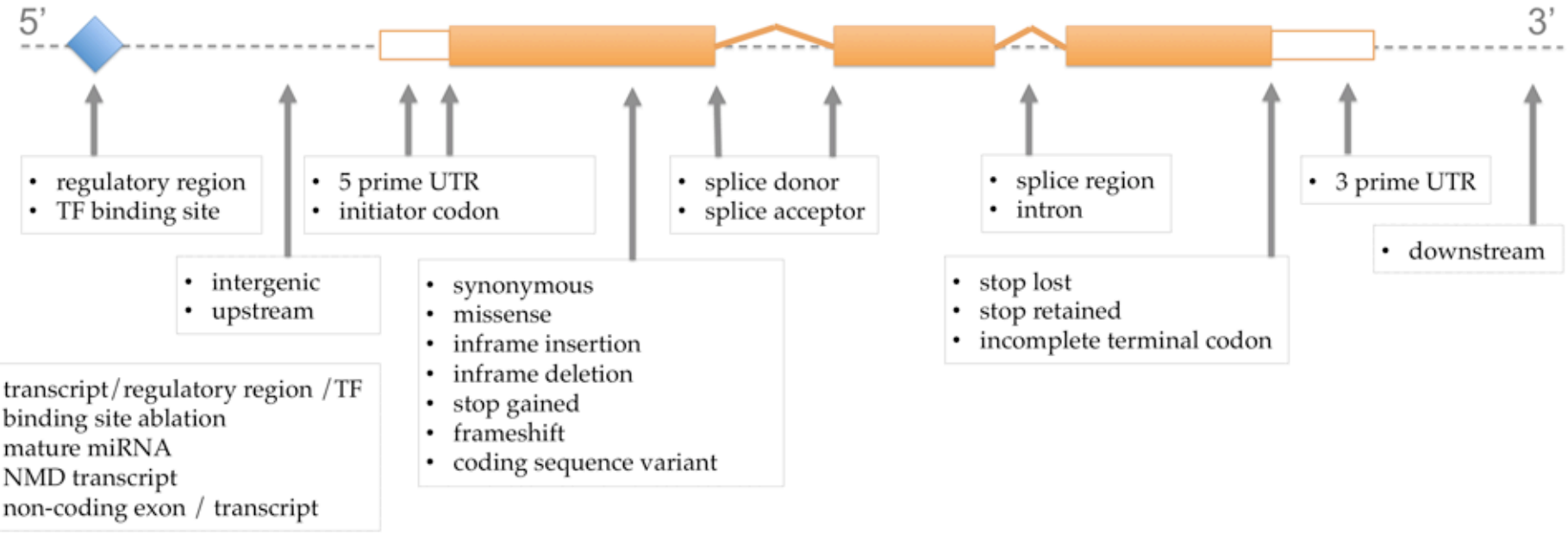


doi:10.1038/ng.2892



doi:10.1038/ng.2892

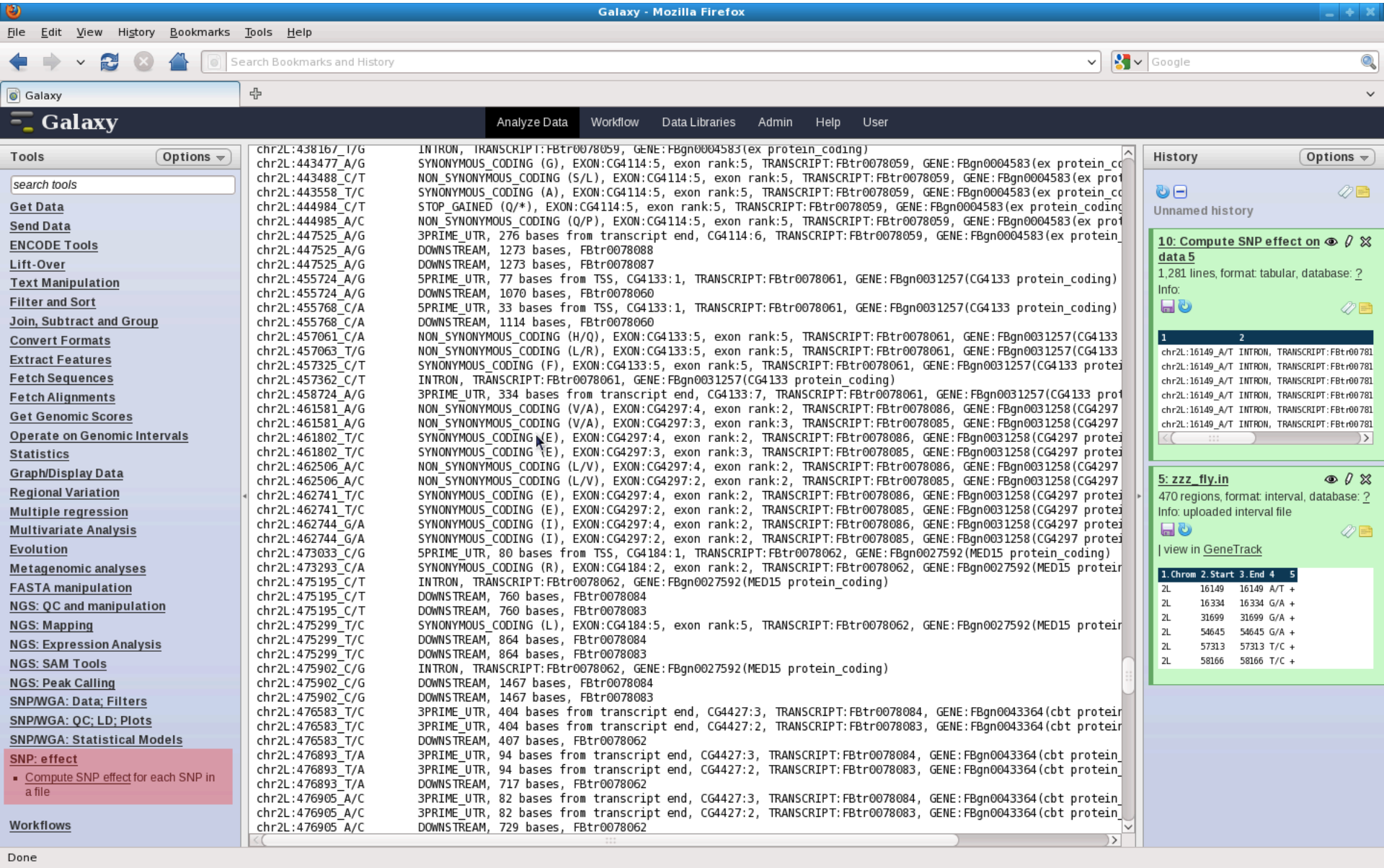
Functional Categorization



SNPEff

- Categorizes effects of single nucleotide polymorphisms, multiple nucleotide polymorphisms and indels based on effects on annotated genes
  - Synonymous or non-synonymous
  - 3’ or 5’UTR
  - Intronic
  - Upstream, downstream of genes
  - Intergenic regions
- Will output effects for every transcript
- Integrated into Galaxy, GATK and other sequence data analysis pipelines

http://snpeff.sourceforge.net/SnpEff.html


Galaxy - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Search Bookmarks and History
Google
Galaxy
Galaxy
Analyze Data
Workflow
Data Libraries
Admin
Help
User
Tools
Options
search tools
Get Data
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Metagenomic analyses
FASTA manipulation
NGS: QC and manipulation
NGS: Mapping
NGS: Expression Analysis
NGS: SAM Tools
NGS: Peak Calling
SNP/WGA: Data; Filters
SNP/WGA: QC; LD; Plots
SNP/WGA: Statistical Models
SNP: effect
Compute SNP effect for each SNP in a file
Workflows
chr2L:438167_T/G INTRON, TRANSCRIPT:FBtr0078059, GENE:FBgn0004583(ex protein_coding)
chr2L:443477_A/G SYNONYMOUS_CODING (G), EXON:CG4114:5, exon rank:5, TRANSCRIPT:FBtr0078059, GENE:FBgn0004583(ex protein_co
chr2L:443488_C/T NON_SYNONYMOUS_CODING (S/L), EXON:CG4114:5, exon rank:5, TRANSCRIPT:FBtr0078059, GENE:FBgn0004583(ex prot
chr2L:443558_T/C SYNONYMOUS_CODING (A), EXON:CG4114:5, exon rank:5, TRANSCRIPT:FBtr0078059, GENE:FBgn0004583(ex protein_co
chr2L:444984_C/T STOP_GAINED (Q/*), EXON:CG4114:5, exon rank:5, TRANSCRIPT:FBtr0078059, GENE:FBgn0004583(ex protein_coding
chr2L:444985_A/C NON_SYNONYMOUS_CODING (Q/P), EXON:CG4114:5, exon rank:5, TRANSCRIPT:FBtr0078059, GENE:FBgn0004583(ex prot
chr2L:447525_A/G 3PRIME_UTR, 276 bases from transcript end, CG4114:6, TRANSCRIPT:FBtr0078059, GENE:FBgn0004583(ex protein_
chr2L:447525_A/G DOWNSTREAM, 1273 bases, FBtr0078088
chr2L:447525_A/G DOWNSTREAM, 1273 bases, FBtr0078087
chr2L:455724_A/G 5PRIME_UTR, 77 bases from TSS, CG4133:1, TRANSCRIPT:FBtr0078061, GENE:FBgn0031257(CG4133 protein_coding)
chr2L:455724_A/G DOWNSTREAM, 1070 bases, FBtr0078060
chr2L:455768_C/A 5PRIME_UTR, 33 bases from TSS, CG4133:1, TRANSCRIPT:FBtr0078061, GENE:FBgn0031257(CG4133 protein_coding)
chr2L:455768_C/A DOWNSTREAM, 1114 bases, FBtr0078060
chr2L:457061_C/A NON_SYNONYMOUS_CODING (H/Q), EXON:CG4133:5, exon rank:5, TRANSCRIPT:FBtr0078061, GENE:FBgn0031257(CG4133
chr2L:457063_T/G NON_SYNONYMOUS_CODING (L/R), EXON:CG4133:5, exon rank:5, TRANSCRIPT:FBtr0078061, GENE:FBgn0031257(CG4133
chr2L:457325_C/T SYNONYMOUS_CODING (F), EXON:CG4133:5, exon rank:5, TRANSCRIPT:FBtr0078061, GENE:FBgn0031257(CG4133 protei
chr2L:457362_C/T INTRON, TRANSCRIPT:FBtr0078061, GENE:FBgn0031257(CG4133 protein_coding)
chr2L:458724_A/G 3PRIME_UTR, 334 bases from transcript end, CG4133:7, TRANSCRIPT:FBtr0078061, GENE:FBgn0031257(CG4133 prot
chr2L:461581_A/G NON_SYNONYMOUS_CODING (V/A), EXON:CG4297:4, exon rank:2, TRANSCRIPT:FBtr0078086, GENE:FBgn0031258(CG4297
chr2L:461581_A/G NON_SYNONYMOUS_CODING (V/A), EXON:CG4297:3, exon rank:3, TRANSCRIPT:FBtr0078085, GENE:FBgn0031258(CG4297
chr2L:461802_T/C SYNONYMOUS_CODING (E), EXON:CG4297:4, exon rank:2, TRANSCRIPT:FBtr0078086, GENE:FBgn0031258(CG4297 protei
chr2L:461802_T/C SYNONYMOUS_CODING (E), EXON:CG4297:3, exon rank:3, TRANSCRIPT:FBtr0078085, GENE:FBgn0031258(CG4297 protei
chr2L:462506_A/C NON_SYNONYMOUS_CODING (L/V), EXON:CG4297:4, exon rank:2, TRANSCRIPT:FBtr0078086, GENE:FBgn0031258(CG4297
chr2L:462506_A/C NON_SYNONYMOUS_CODING (L/V), EXON:CG4297:2, exon rank:2, TRANSCRIPT:FBtr0078085, GENE:FBgn0031258(CG4297
chr2L:462741_T/C SYNONYMOUS_CODING (E), EXON:CG4297:4, exon rank:2, TRANSCRIPT:FBtr0078086, GENE:FBgn0031258(CG4297 protei
chr2L:462741_T/C SYNONYMOUS_CODING (E), EXON:CG4297:2, exon rank:2, TRANSCRIPT:FBtr0078085, GENE:FBgn0031258(CG4297 protei
chr2L:462744_G/A SYNONYMOUS_CODING (I), EXON:CG4297:4, exon rank:2, TRANSCRIPT:FBtr0078086, GENE:FBgn0031258(CG4297 protei
chr2L:462744_G/A SYNONYMOUS_CODING (I), EXON:CG4297:2, exon rank:2, TRANSCRIPT:FBtr0078085, GENE:FBgn0031258(CG4297 protei
chr2L:473033_C/G 5PRIME_UTR, 80 bases from TSS, CG4184:1, TRANSCRIPT:FBtr0078062, GENE:FBgn0027592(MED15 protein_coding)
chr2L:473293_C/A SYNONYMOUS_CODING (R), EXON:CG4184:2, exon rank:2, TRANSCRIPT:FBtr0078062, GENE:FBgn0027592(MED15 protein
chr2L:475195_C/T INTRON, TRANSCRIPT:FBtr0078062, GENE:FBgn0027592(MED15 protein_coding)
chr2L:475195_C/T DOWNSTREAM, 760 bases, FBtr0078084
chr2L:475195_C/T DOWNSTREAM, 760 bases, FBtr0078083
chr2L:475299_T/C SYNONYMOUS_CODING (L), EXON:CG4184:5, exon rank:5, TRANSCRIPT:FBtr0078062, GENE:FBgn0027592(MED15 protein
chr2L:475299_T/C DOWNSTREAM, 864 bases, FBtr0078084
chr2L:475299_T/C DOWNSTREAM, 864 bases, FBtr0078083
chr2L:475902_C/G INTRON, TRANSCRIPT:FBtr0078062, GENE:FBgn0027592(MED15 protein_coding)
chr2L:475902_C/G DOWNSTREAM, 1467 bases, FBtr0078084
chr2L:475902_C/G DOWNSTREAM, 1467 bases, FBtr0078083
chr2L:476583_T/C 3PRIME_UTR, 404 bases from transcript end, CG4427:3, TRANSCRIPT:FBtr0078084, GENE:FBgn0043364(cbt protein
chr2L:476583_T/C 3PRIME_UTR, 404 bases from transcript end, CG4427:2, TRANSCRIPT:FBtr0078083, GENE:FBgn0043364(cbt protein
chr2L:476583_T/C DOWNSTREAM, 407 bases, FBtr0078062
chr2L:476893_T/A 3PRIME_UTR, 94 bases from transcript end, CG4427:3, TRANSCRIPT:FBtr0078084, GENE:FBgn0043364(cbt protein_
chr2L:476893_T/A 3PRIME_UTR, 94 bases from transcript end, CG4427:2, TRANSCRIPT:FBtr0078083, GENE:FBgn0043364(cbt protein_
chr2L:476893_T/A DOWNSTREAM, 717 bases, FBtr0078062
chr2L:476905_A/C 3PRIME_UTR, 82 bases from transcript end, CG4427:3, TRANSCRIPT:FBtr0078084, GENE:FBgn0043364(cbt protein_
chr2L:476905_A/C 3PRIME_UTR, 82 bases from transcript end, CG4427:2, TRANSCRIPT:FBtr0078083, GENE:FBgn0043364(cbt protein_
chr2L:476905_A/C DOWNSTREAM, 729 bases, FBtr0078062
History
Options
Unnamed history
10: Compute SNP effect on data 5
1,281 lines, format: tabular, database: ?
Info:
1 2
chr2L:16149_A/T INTRON, TRANSCRIPT:FBtr00781
chr2L:16149_A/T INTRON, TRANSCRIPT:FBtr00781
chr2L:16149_A/T INTRON, TRANSCRIPT:FBtr00781
chr2L:16149_A/T INTRON, TRANSCRIPT:FBtr00781
chr2L:16149_A/T INTRON, TRANSCRIPT:FBtr00781
chr2L:16149_A/T INTRON, TRANSCRIPT:FBtr00781
5: zzz_fly.in
470 regions, format: interval, database: ?
Info: uploaded interval file
| view in GeneTrack
1.Chrom 2.Start 3.End 4 5
2L 16149 16149 A/T +
2L 16334 16334 G/A +
2L 31699 31699 G/A +
2L 54645 54645 G/A +
2L 57313 57313 T/C +
2L 58166 58166 T/C +
Done

VEP

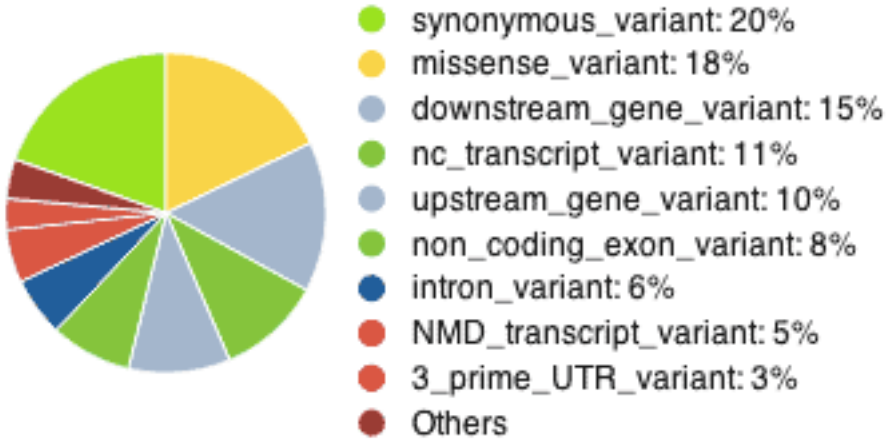
- Variant Effect Predictor
- Standalone perl script or Web interface
- Output includes
  - Genes and transcripts affected by variants
  - Location of variants (e.g. upstream of transcript, in coding sequence, in non-coding RNA)
  - Consequence of variants on protein sequence (e.g. stop gained, missense, etc)
  - Known variants that match yours and associated minor allele frequencies from 1000 genomes project
  - SIFT and PolyPhen scores

http://useast.ensembl.org/info/docs/tools/vep/index.html

VEP output

Category	Count
Variants processed	498955
Variants remaining after filtering	498955
Novel / known variants	-
Overlapped genes	825
Overlapped transcripts	2888
Overlapped regulatory features	7309

Consequences (all)



VEP output

Show/hide columns

Uploaded variation	Location	Feature	Feature type	Consequence	CDS position	Protein position	Amino acids
rs116383664	1:1115461	ENSR00000528923	RegulatoryFeature	regulatory_region_variant	-	-	-
rs116383664	1:1115461	ENST00000379317	Transcript	upstream_gene_variant	-	-	-
rs116383664	1:1115461	ENST00000486379	Transcript	upstream_gene_variant	-	-	-
rs116383664	1:1115461	ENST00000379289	Transcript	missense_variant	247	83	R/W
rs116383664	1:1115461	ENST00000460998	Transcript	upstream_gene_variant	-	-	-
rs116383664	1:1115461	ENST00000514695	Transcript	upstream_gene_variant	-	-	-
rs116383664	1:1115461	ENST00000379290	Transcript	missense_variant	247	83	R/W
rs116383664	1:1115461	ENST00000379288	Transcript	missense_variant	28	10	R/W

Download

All VCF VEP TXT

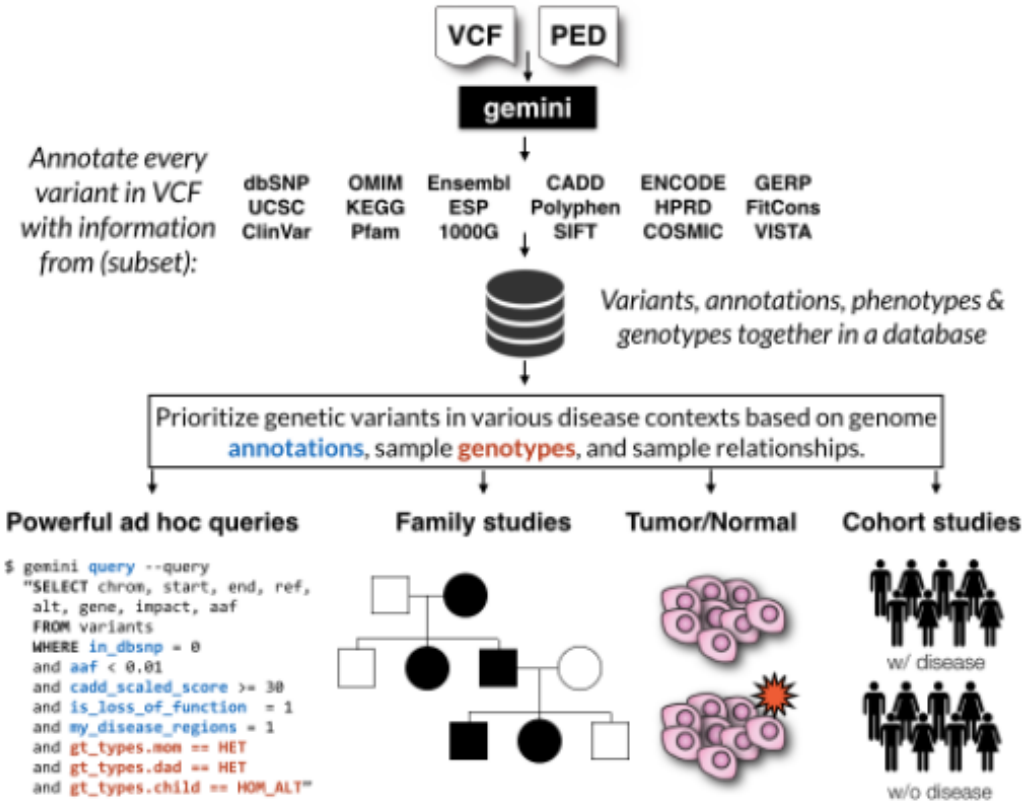
Filtered VCF VEP TXT

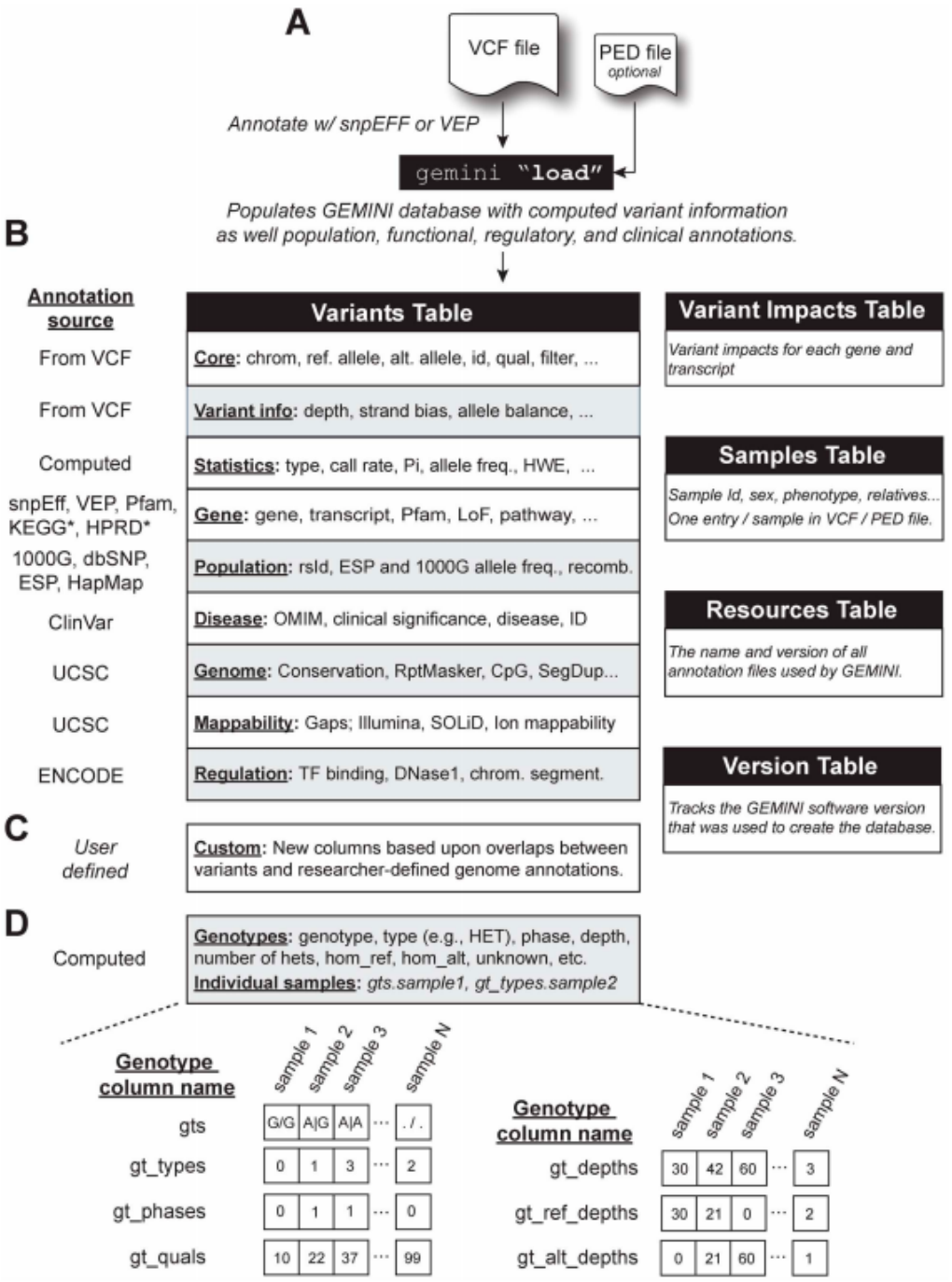
Download filtered results in TXT format (best for Excel)

GEMINI

http://gemini.readthedocs.io/en/latest/index.html

GEMINI: *a flexible framework for exploring genome variation*




A
VCF file
PED file
optional
Annotate w/ snpEFF or VEP
gemini "load"
Populates GEMINI database with computed variant information
as well population, functional, regulatory, and clinical annotations.
B
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, Pfam, KEGG*, HPRD*
Gene: gene, transcript, Pfam, LoF, 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.
Variant Impacts Table
Variant impacts for each gene and 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.
C
User defined
Custom: New columns based upon overlaps between variants and researcher-defined genome annotations.
D
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
Genotype 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_quals
10
22
37
99
Genotype column name
sample 1
sample 2
sample 3
sample N
gt_depths
30
42
60
3
gt_ref_depths
30
21
0
2
gt_alt_depths
0
21
60
1

GEMINI

- Decompose original VCF so that variants with multiple alleles are expanded into distinct variant records (one record for each REF/ALT combination)
- Normalize the decomposed VCF so that variants are left aligned and represented using the most parsimonious alleles
- Annotate with VEP or snpEff
- bgzip and tabix
- Load VCF
  - If family based study be sure to use a ped file!

GEMINI

1. Extract all transitions with a call rate > 95%

```
$ gemini query -q "select * from variants \
                   where sub_type = 'ts' \
                   and call_rate >= 0.95" my.db
```

2. Extract all loss-of-function variants with an alternate allele frequency < 1%:

```
$ gemini query -q "select * from variants \
                   where is_lof = 1 \
                   and aaf >= 0.01" my.db
```

3. Extract the nucleotide diversity for each variant:

```
$ gemini query -q "select chrom, start, end, pi from variants" my.db
```

4. Combine `GEMINI` with `bedtools` to compute nucleotide diversity estimates across 100kb windows:

```
$ gemini query -q "select chrom, start, end, pi from variants \
                   order by chrom, start, end" my.db | \
  bedtools map -a hg19.windows.bed -b - -c 4 -o mean
```

GEMINI

- Built in analysis tools
  - de_novo
  - comp_hets
  - autosomal_recessive
  - autosomal_dominant
  - x_linked_recessive
  - x_linked_dominant
  - x_linked_de_novo

GEMINI

- Limit to confidently called genotypes
  - gt-pl-max 10
- Limit to those predicted to disrupt the protein
  - --filter “impact_severity = ‘HIGH’”
- Limit to rare alleles
  - --filter “max_aaf_all < 0.005”
- Limit to high genotype quality
  - --min-gq 20

GEMINI

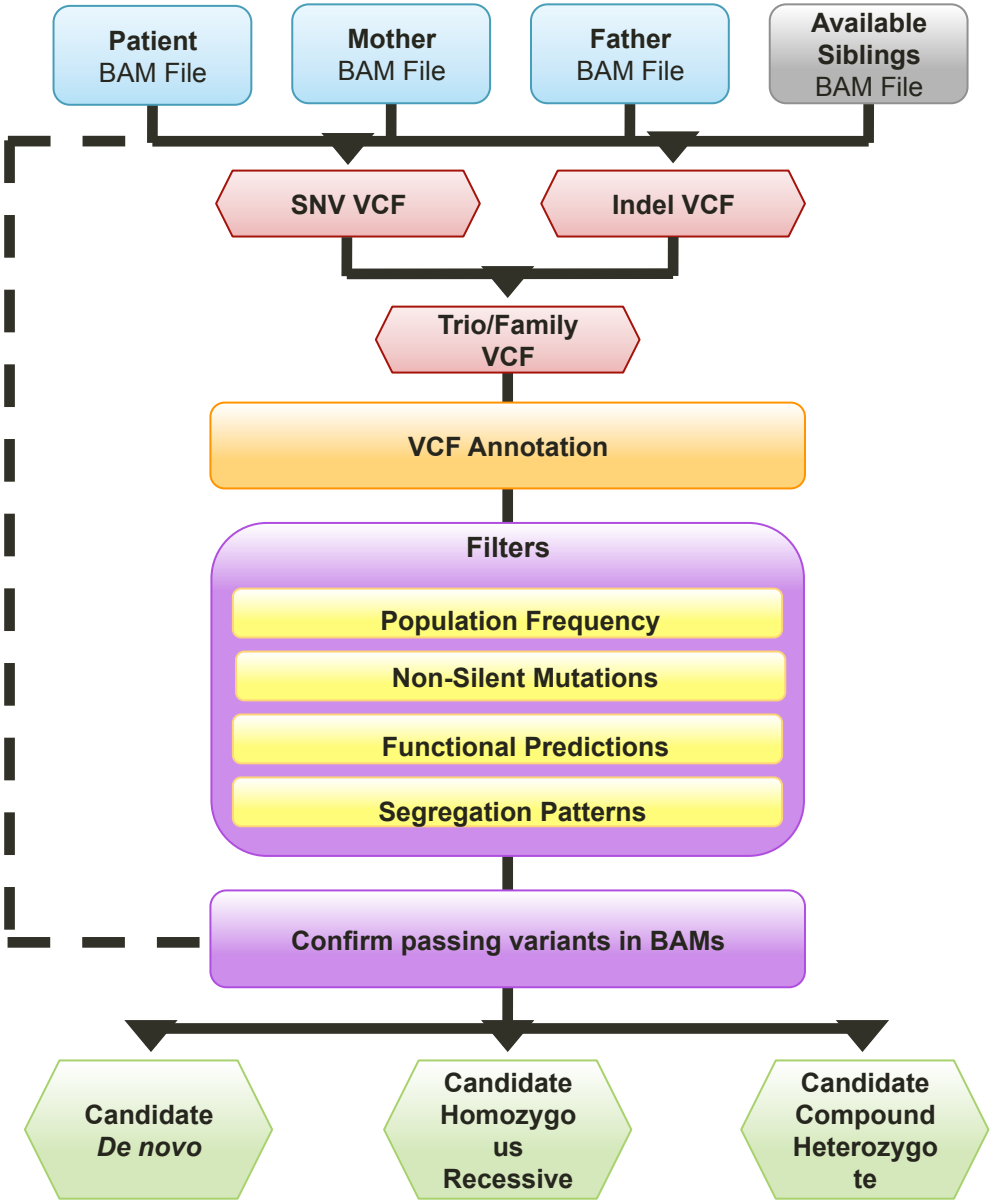
```
$ gemini de_novo \
      --columns "chrom, start, end, ref, alt" \
      --filter "impact_severity = 'HIGH'" \
      --min-kindreds 2 \
      test.de_novo.db
```

Restrict candidate genes to those genes with a de novo variant in at least 2 families

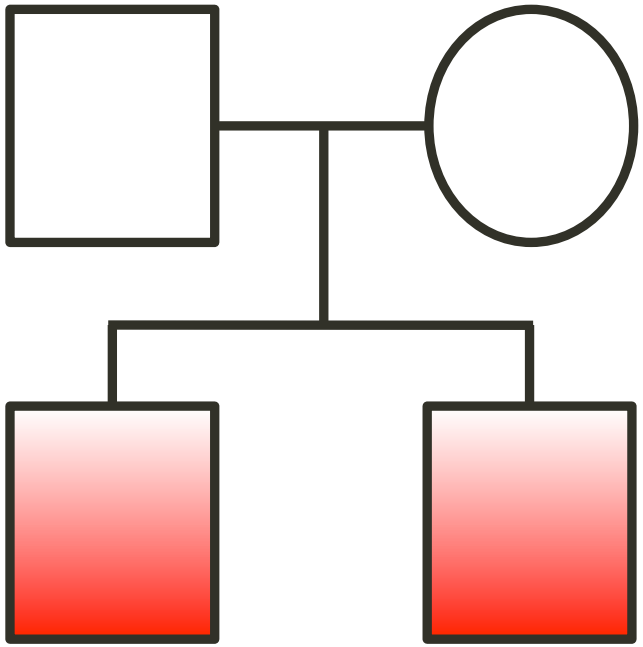
Practical Example: Genomics Pediatric Board

Washington University

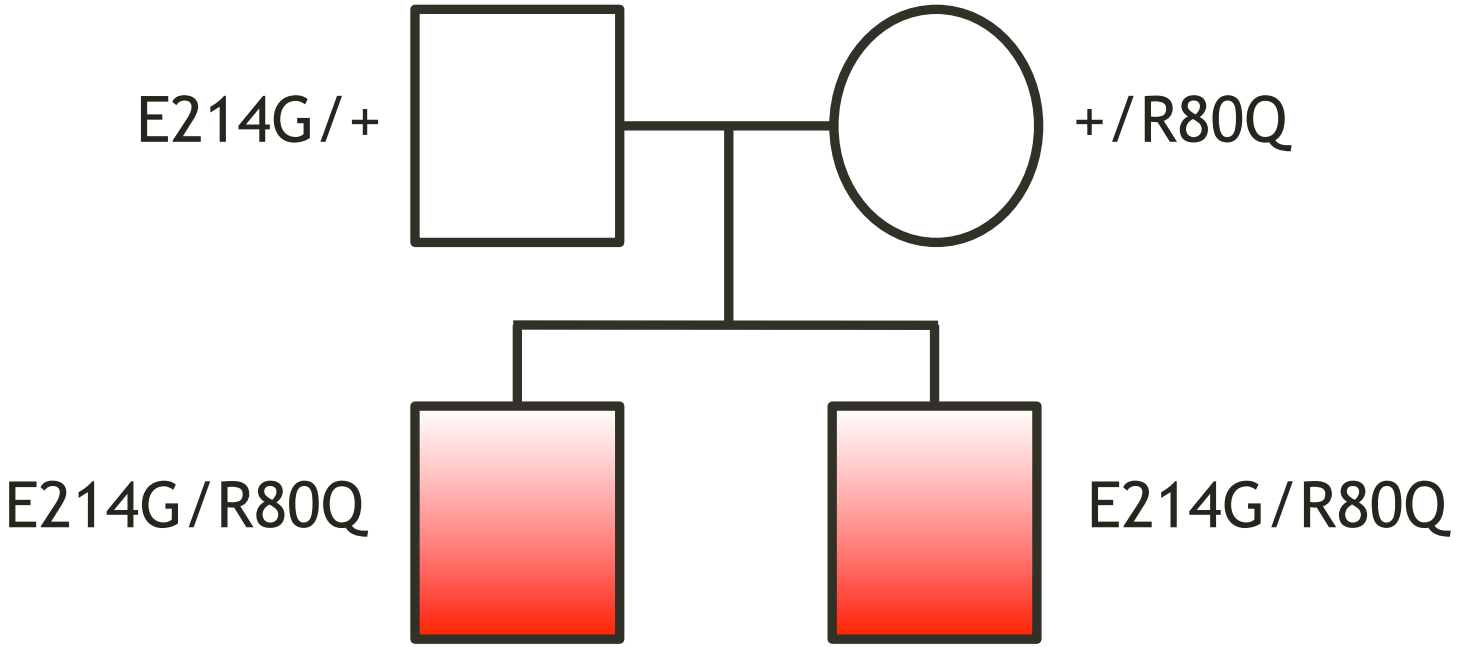
Daniel Wegner, Jennifer Waumbach, Thomas Nicholas, F. Sessions Cole

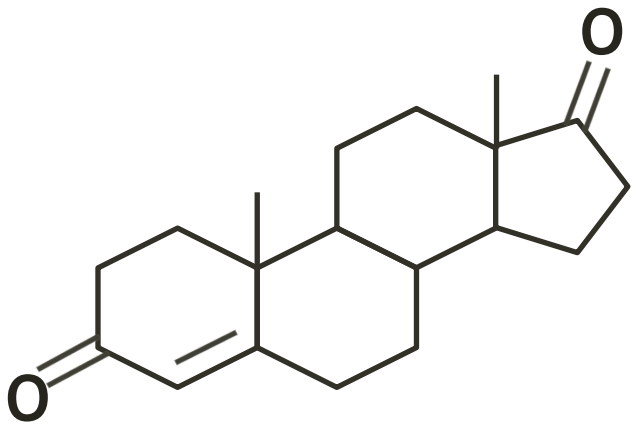

Patient
BAM File
Mother
BAM File
Father
BAM File
Available
Siblings
BAM File
SNV VCF
Indel VCF
Trio/Family
VCF
VCF Annotation
Filters
Population Frequency
Non-Silent Mutations
Functional Predictions
Segregation Patterns
Confirm passing variants in BAMs
Candidate
De novo
Candidate
Homozygo
us
Recessive
Candidate
Compound
Heterozygo
te

GBD-012 presented 2 male siblings with ambiguous genitalia



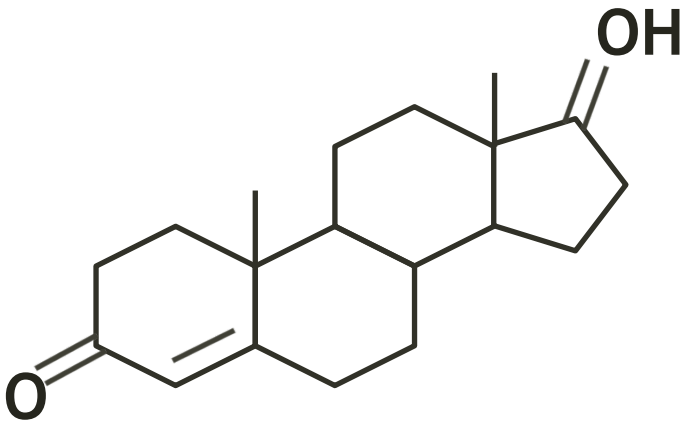
Each sibling is compound heterozygous for *HSD17B3*





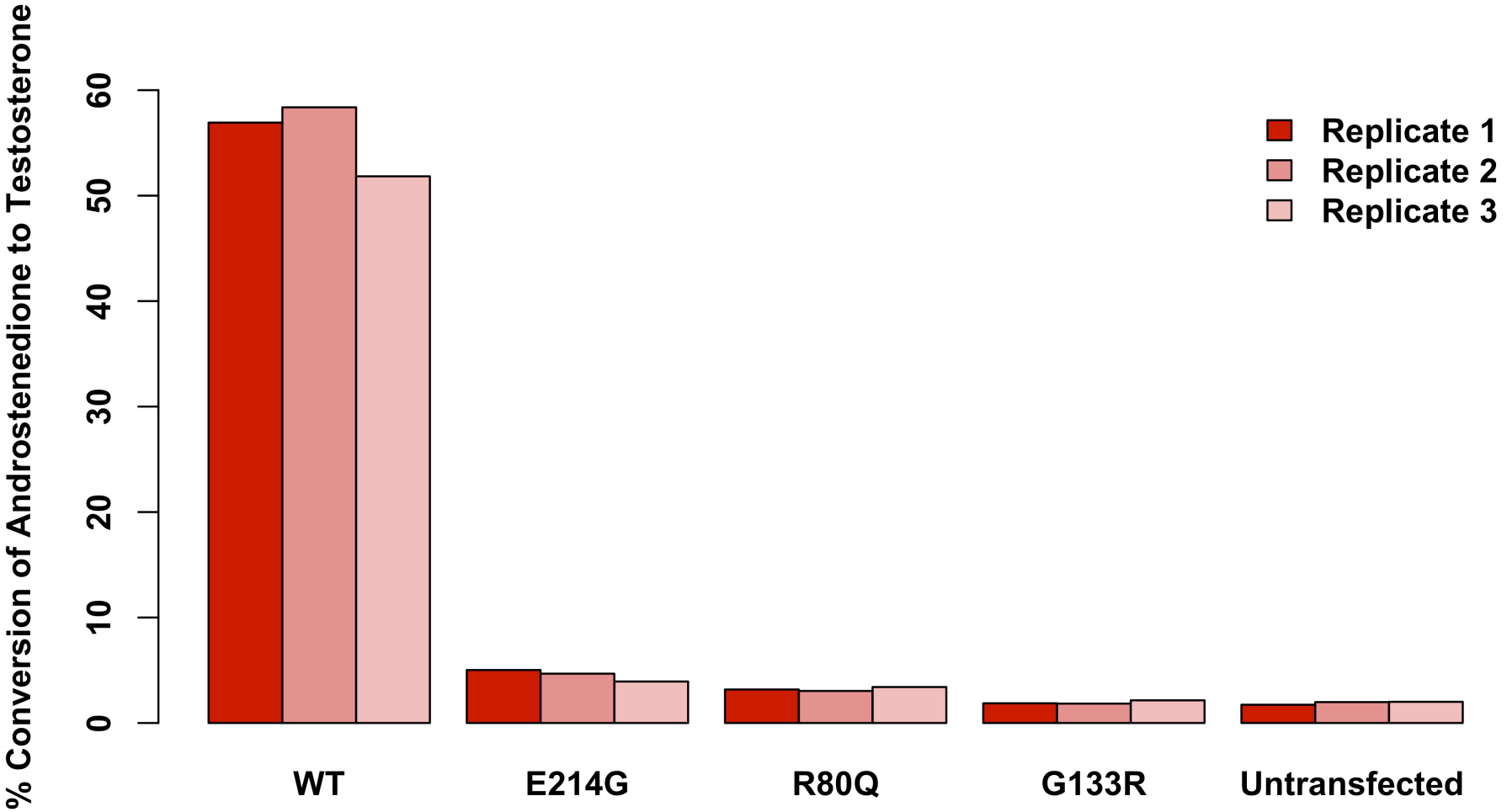
Androstenedione

HSD17B3



Testosterone

Decreases in androstenedione conversion in cells with mutations



Identifying rare variants in large cohorts

Why study rare variants?

- Identify the complete genetic architecture of traits
- Rare variants should identify new susceptibility loci at known trait loci

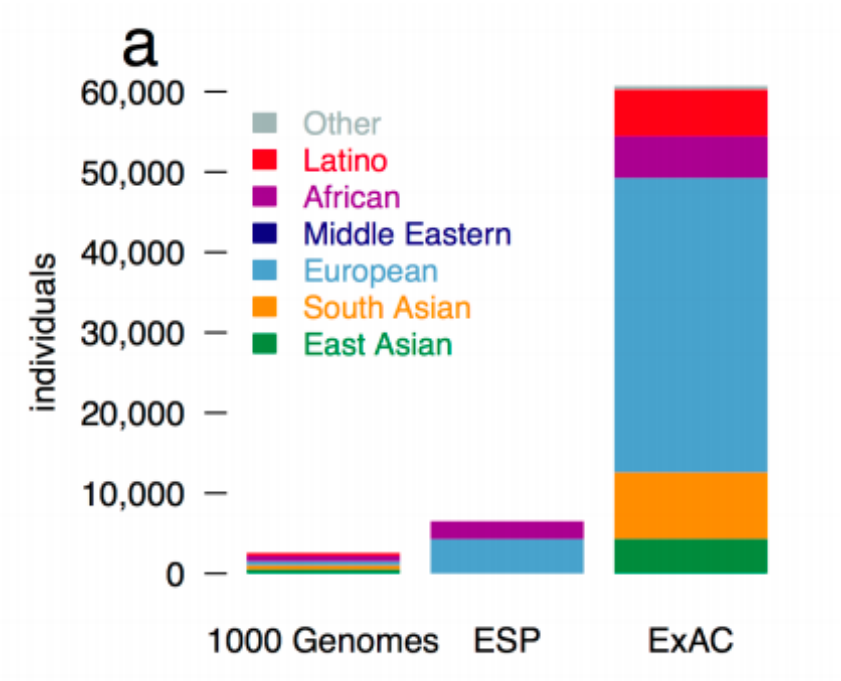
- May have higher penetrance
- May also have higher effect size
  - Most variants that impact function are expected to be deleterious and removed by natural selection

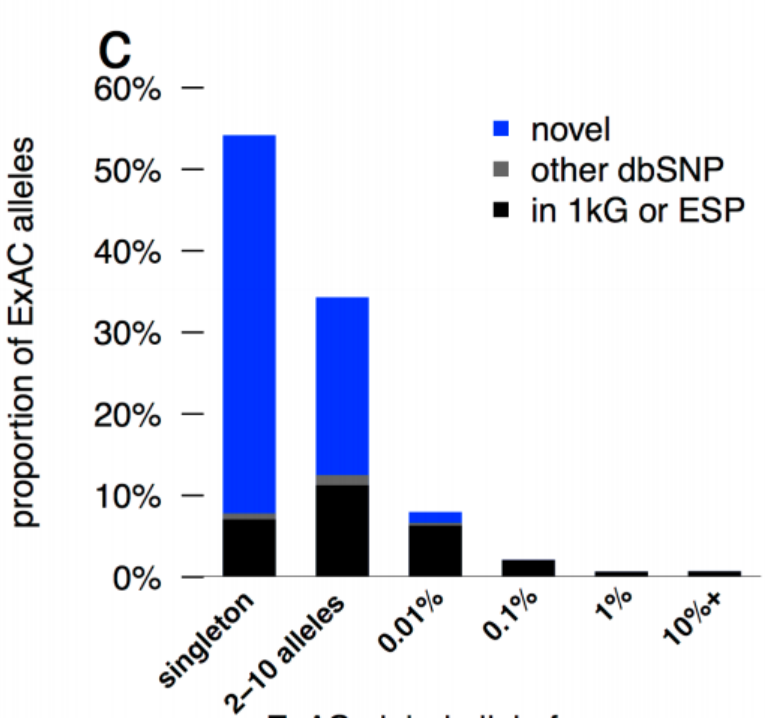
Many rare, likely functional coding variants

	# SNVs	Singletons	Doubletons	Tripletons	> 3 occurrences
Synonymous	270,263	128,319 (47%)	29,340 (11%)	13,129 (5%)	99,475 (37%)
Nonsynonymous	410,956	234,633 (57%)	46.740 (11%)	19,274 (5%)	110,309 (27%)
Nonsense	8,913	6,196 (70%)	926 (10%)	326 (4%)	1,465 (16%)
Ratio (NS/S)		1.8 to 1	1.6 to 1	1.4 to 1	1.1 to 1

NHLBI Exome Sequencing Project

Increasing sample size to 60K identifies even more rare variants





MOST VARIATION IS RARE

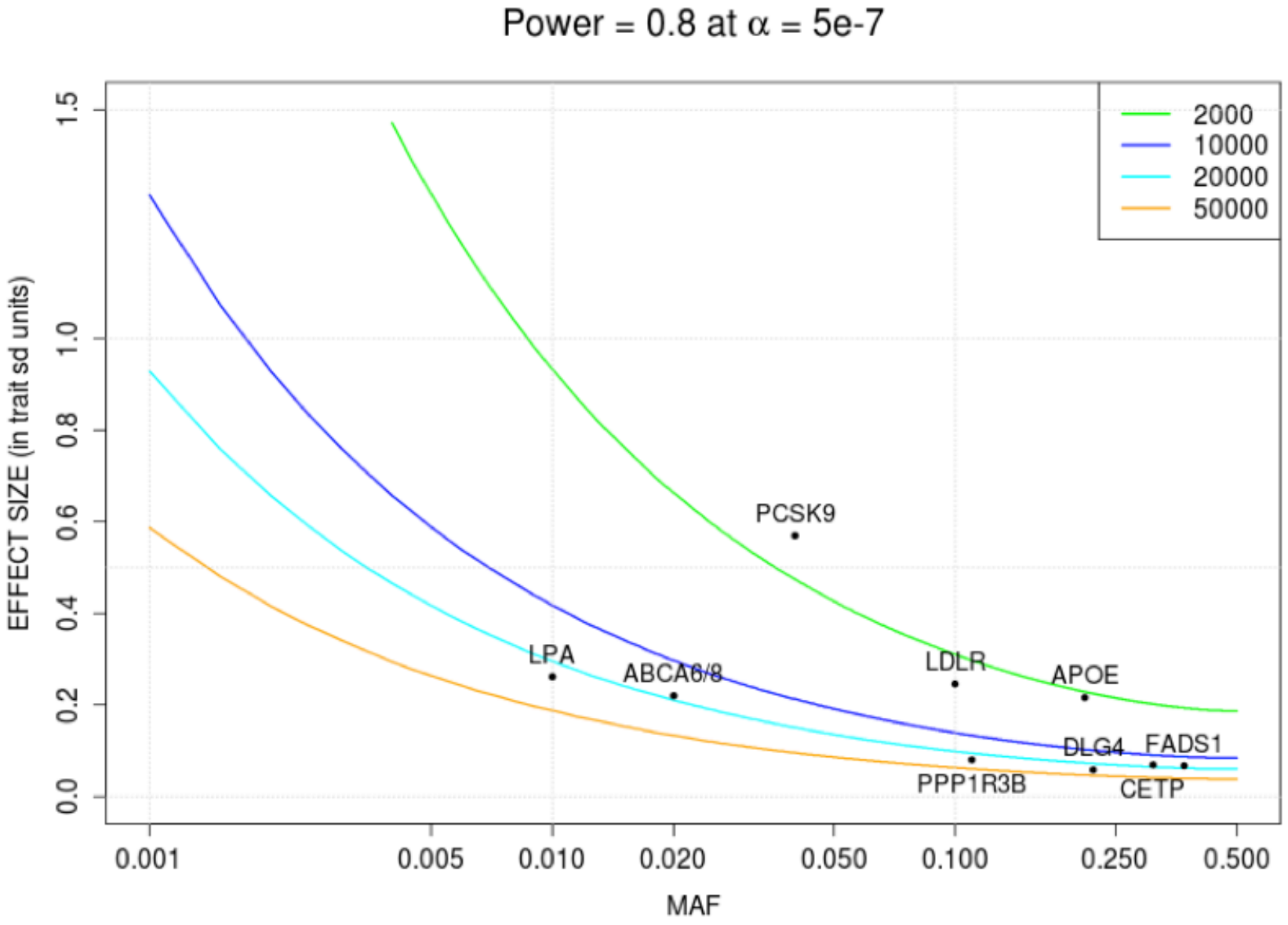
Lek et al, 2015 BiorXiv

Single Variant Association Testing

Sample sizes (cases and controls) required for 80% power to detect Effect Size=2.0 at $\alpha=5x10^{-8}$

MAF	Sample size
0.05	2,500
0.01	12,000
0.001	117,000

With increased sample sizes we have greater power to detect associations at lower effect sizes and MAF



Slide courtesy of Matti Pirinen

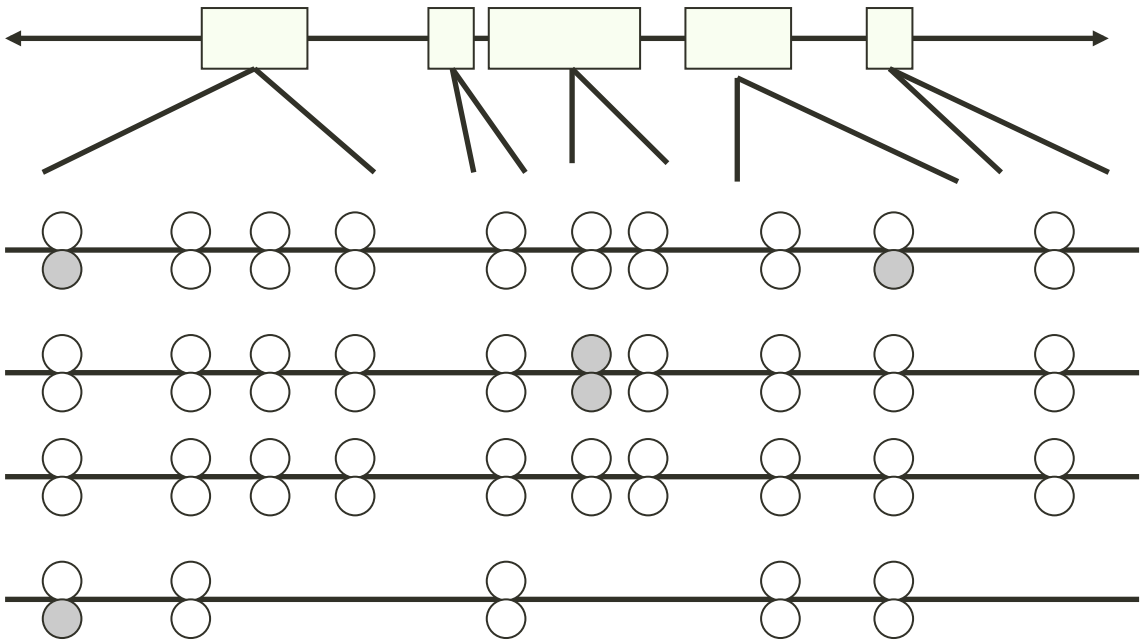
Power depends on:

Frequency

Effect Size

Even with large effects, rare variants can only be detected in large samples

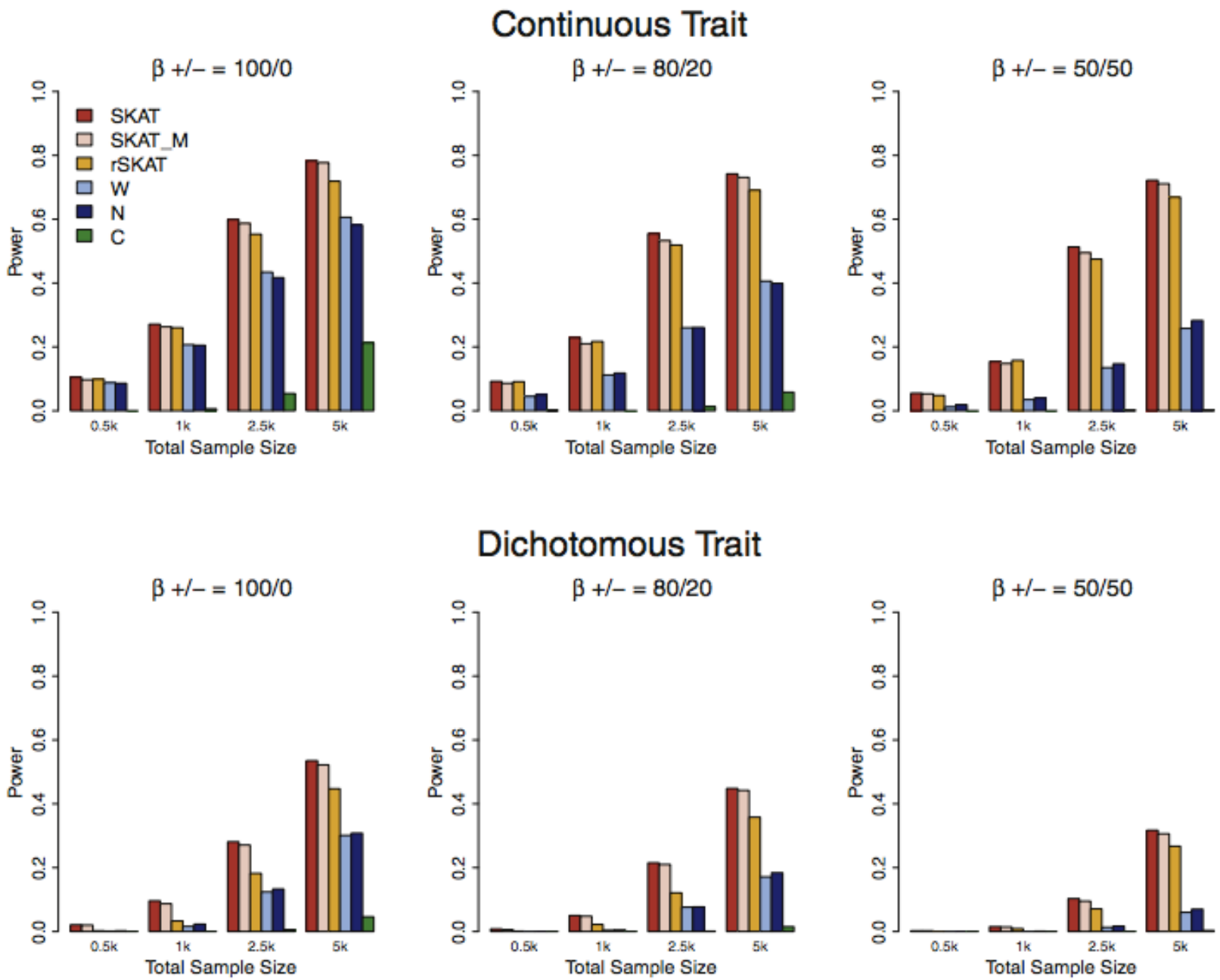
What if we could group variants together?



SKAT

- Sequence kernel association test
  - Regression approach
    - Directly performs multiple regression of a phenotype on genotypes for all variants
  - Tests for association between common and rare variants in a region
  - Dichotomous or continuous phenotype
  - Adjusts for covariates like population stratification
  - Incorporates flexible weight functions to boost power
    - e.g. increase weight of variants with lower MAF

Wu et al, 2011

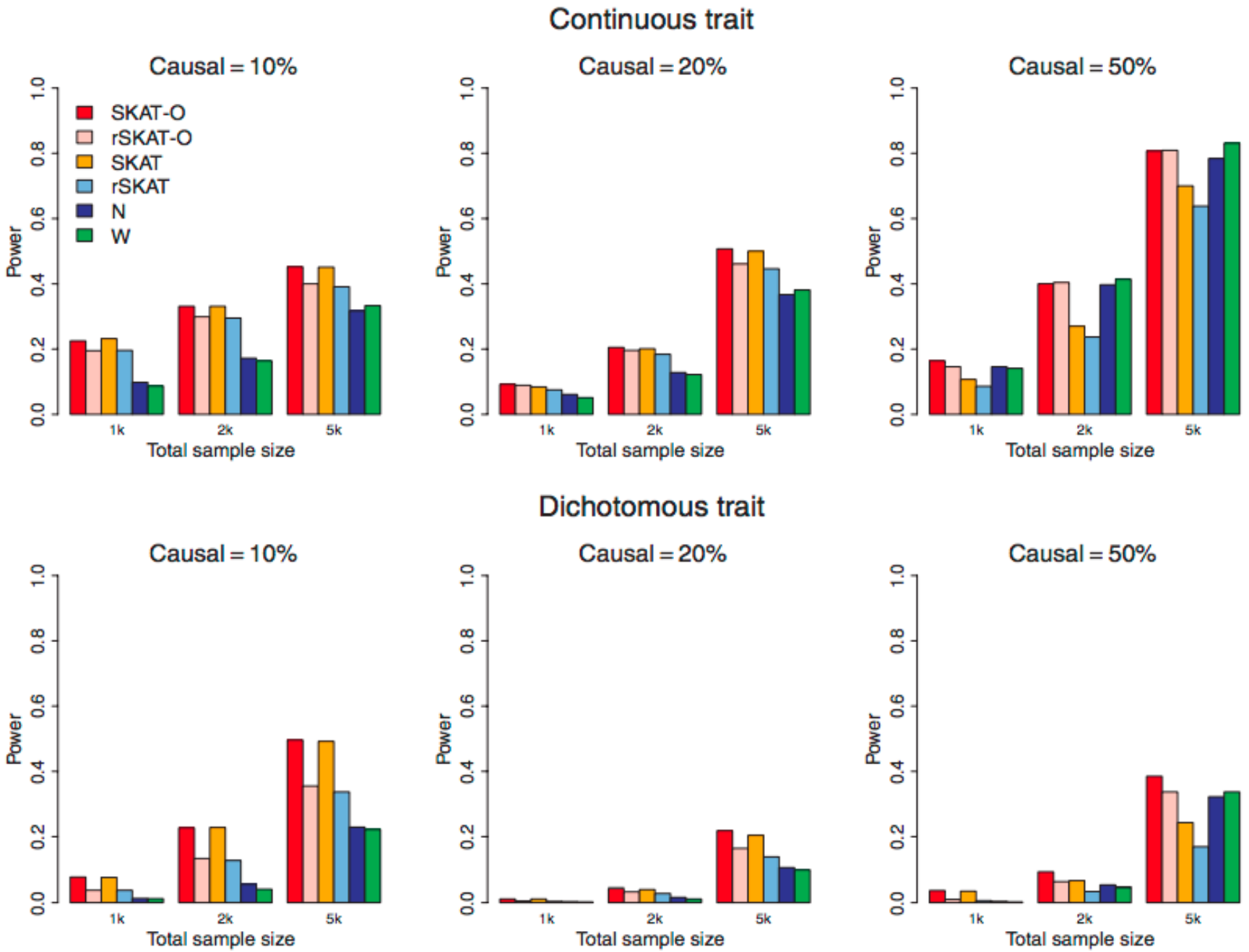

Continuous Trait
β +/− = 100/0
β +/− = 80/20
β +/− = 50/50
SKAT
SKAT_M
rSKAT
W
N
C
Power
0.0
0.2
0.4
0.6
0.8
1.0
0.5k
1k
2.5k
5k
Total Sample Size
Dichotomous Trait
β +/− = 100/0
β +/− = 80/20
β +/− = 50/50
Power
Total Sample Size

Wu et al, 2011

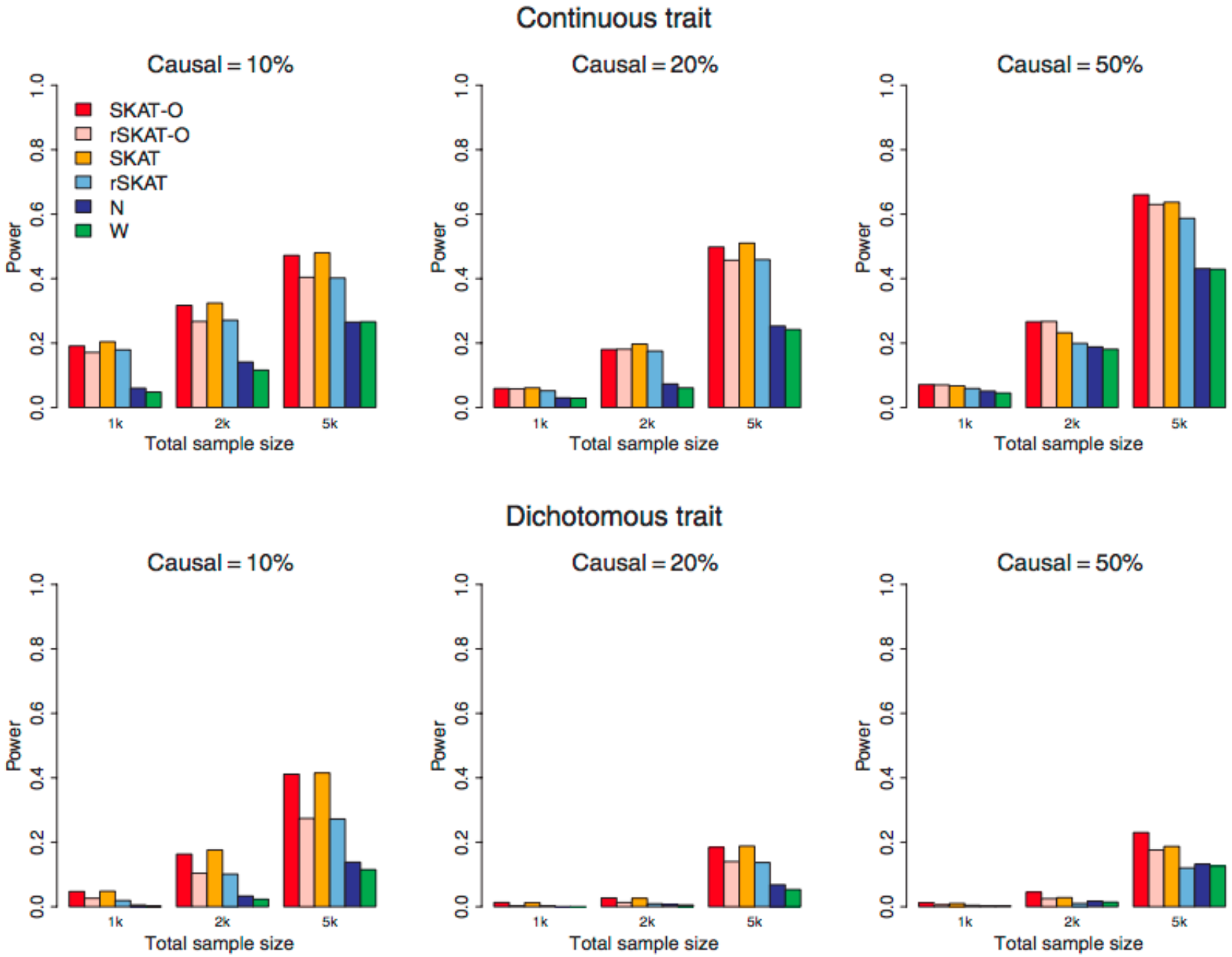
SKAT-O

- Use both burden-based tests and SKAT
- Incorporate a correlation structure of variant effects through family of kernels
  - When effects of variants are perfectly correlated, it is essentially just a burden test
- Optimal test derived by estimating correlation parameter in kernel matrix to maximize power

Lee et al, 2012



Lee et al, 2012



Lee et al, 2012

EMMAX

- Statistical test for large scale association studies accounting for sample structure (e.g. relatedness)
- Avoids repetitive variance component estimation
  - Each loci only explains a small fraction of complex traits
- Very computationally efficient
- http://genome.sph.umich.edu/wiki/EMMAX

Kang et al, 2010

Maximize the power

- Choose threshold for defining rare carefully
- Focus on loss of function (LOF) variants only
- Focus on individuals with extreme trait values for continuous phenotypes
- Focus on individuals with strong family history of disease for dichotomous traits
- Population isolates

Practical Example: FinMetSeq Project

Washington University, University of Michigan, UCLA, FIMM, University of Kuopio

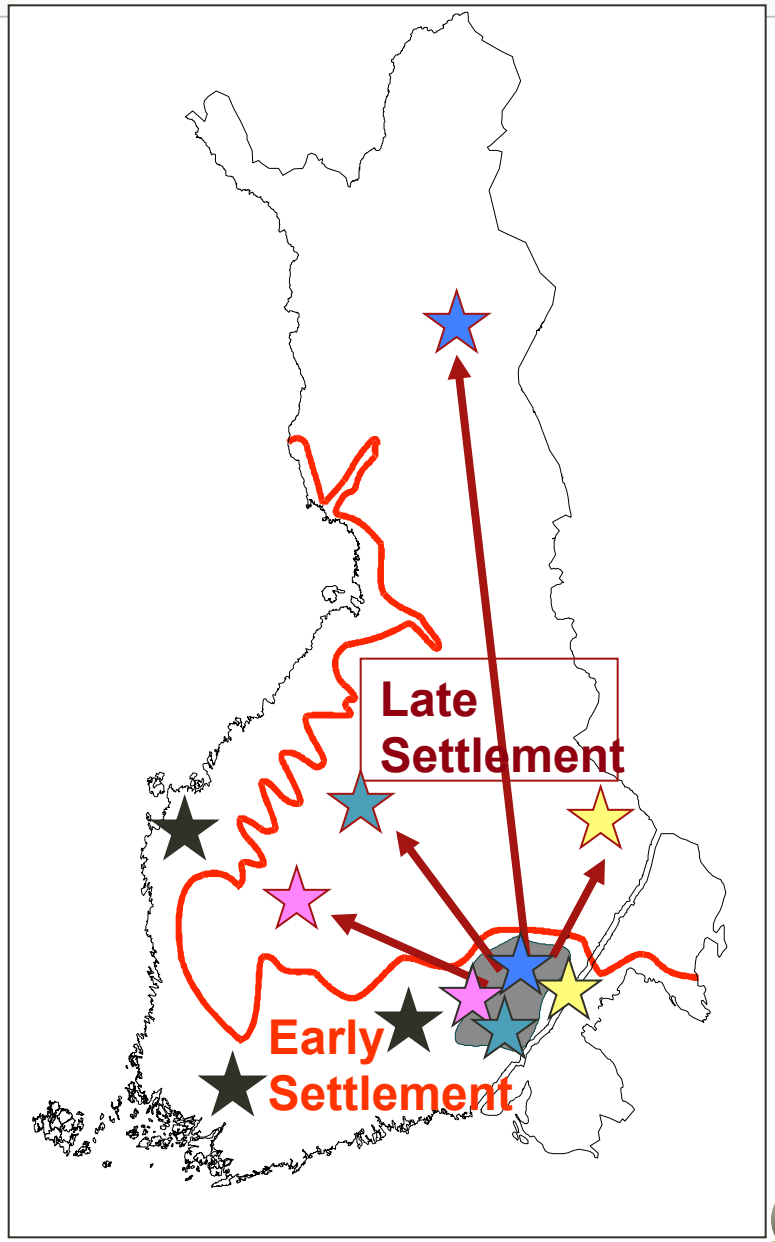
Karyn Meltz Steinberg and Adam Locke

Finland Population History

Early Settlement
- 2,000-10,000 years ago
- South and Coast

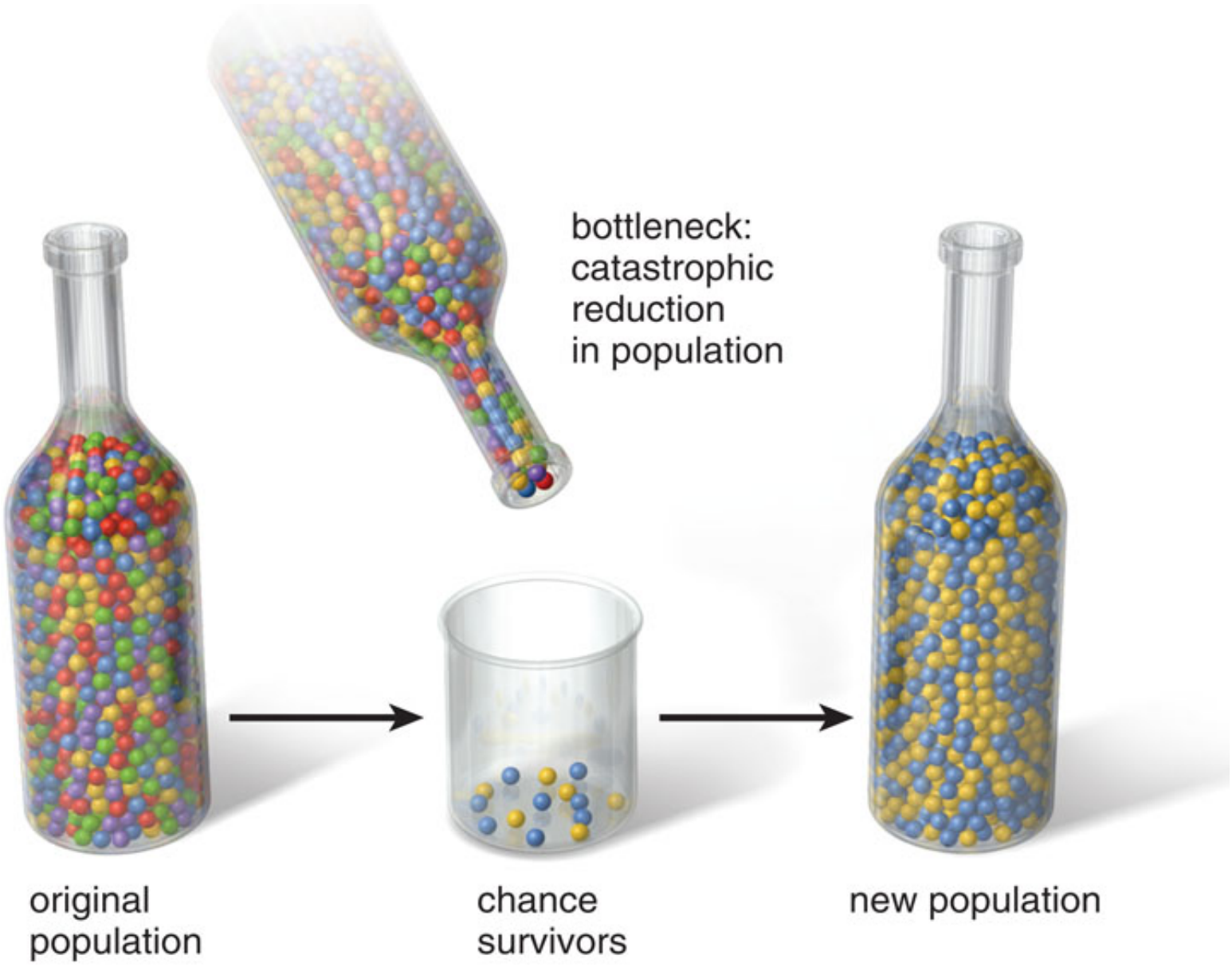
Late Settlement
- 16th century
- Multiple bottlenecks

Expansion
- 18th century (pop 250K)
- Today (pop 5.3M)



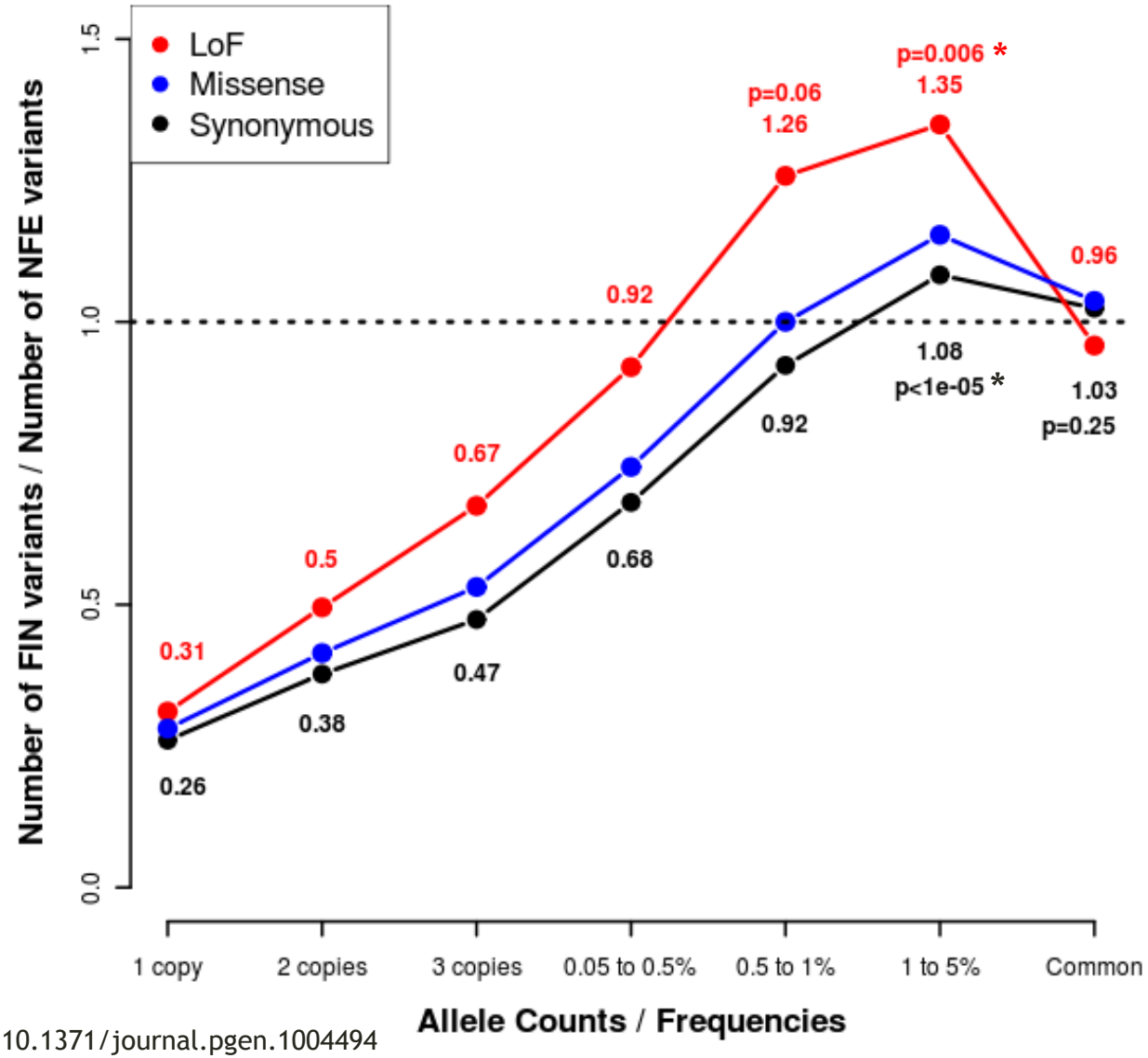
Slide courtesy of A. Palotie

Population isolates: Bottleneck effect and genetic drift



Slide courtesy of A. Palotie

There are proportionally more LoF variants in Finnish individuals compared to Non-Finnish Europeans



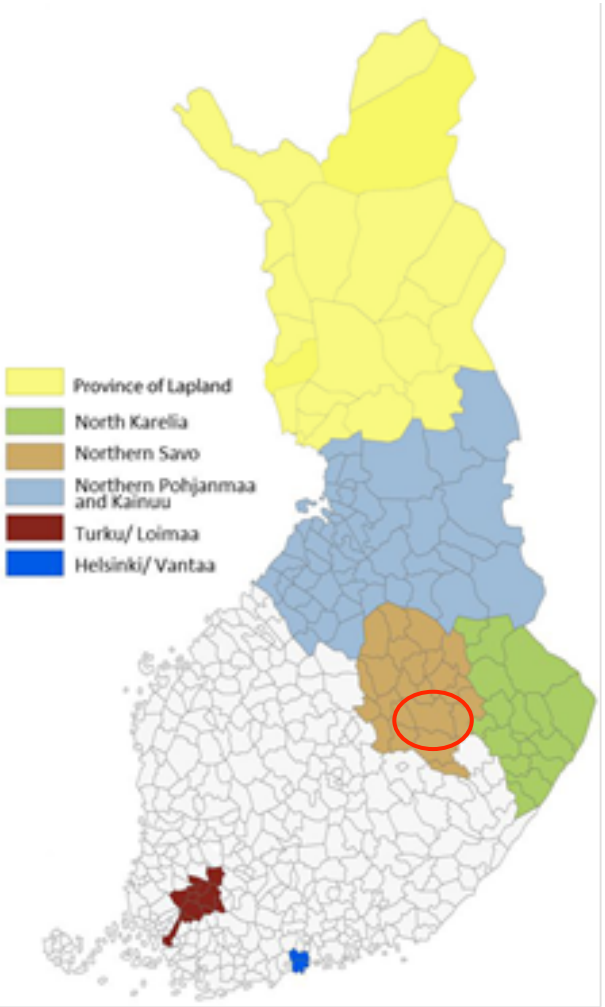
Lim et al, 2015, DOI: 10.1371/journal.pgen.1004494

Major Questions

- Do currently unknown infrequent/rare variants contribute substantially to common disease risk and/or important quantitative traits relevant to cardiometabolic disorders?
  - Novel loci
  - Fine mapping known loci and identifying additional signals
- Can we relate sequence variants that we identify to specific clinical outcomes?
  - Finnish disease heritage

Cohorts for exome sequencing

- METSIM
  - Population based study of 10,197 men from Kuopio
  - Ages 45-70 years at baseline
  - Baseline study conducted in 2005-2010
- FINRISK
  - Health examination surveys every five years since 1972
  - Sample sizes 6,000-12,000 individuals
  - Ages 25-74
  - Stratified by 10 year ago group, gender and study area
  - Included all participants from eastern and northern areas except those that overlap METSIM



Association analysis methodology

- Generate phenotype residuals and transform separately for METSIM and FINRISK
- Inverse normalize phenotype residuals for all quantitative traits (except height)
- Use EMMAX to account for relatedness within and across studies

- Five major trait groups
  - Lipids
  - Glucose and insulin
  - Anthropometrics
  - Blood pressure
  - Other (inflammatory markers, etc)

Overview

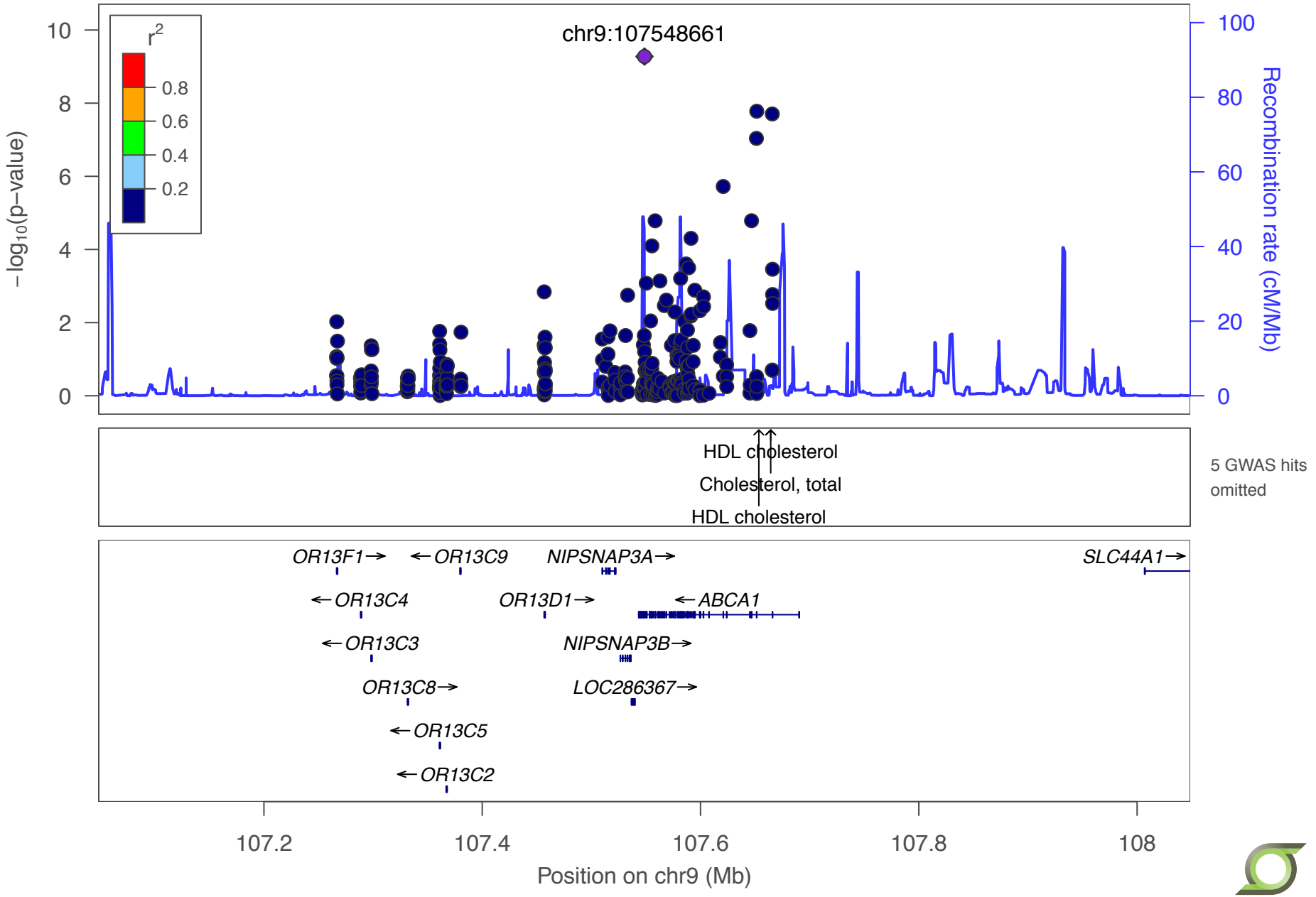
- Protein truncating variants + Non-synonymous variants MAF<0.1, deleterious by all 5 algorithms in Purcell et al 2015
- 12,880 genes
- Bonferroni corrected p-value 3.88e-6

ABCA1 and ApoA1 (p=3.26e-06)

Variants in gene based test	MAC	MAF	ExAC_AF_FIN	ExAC_AF_NFE	P-value	Beta	CADD	CSQ
9:107548661_A/G (Lead SNP)	7	2.32E-04	4.50E-04	0.00E+00	4.42E-05	-1.531	32	Missense
9:107550254_G/A	3	7.72E-05	1.50E-04	1.50E-05	4.14E-03	-1.644	34	Missense
9:107550798_T/G	1	2.57E-05	NA	NA	NA	NA	31	Missense
9:107558416_T/C	15	8.24E-04	NA	NA	7.90E-02	-0.4499	32	Missense
9:107560784_C/T	6	3.35E-04	9.00E-04	3.00E-05	1.20E-01	-0.6298	35	Missense
9:107568659_G/A	1	2.57E-05	0.00E+00	0.00E+00	NA	NA	39	Stop gain
9:107589238_C/G	7	3.60E-04	1.50E-04	3.30E-03	3.12E-01	-0.3791	26.4	Missense
9:107593948_C/T	1	2.57E-05	NA	NA	NA	NA	21.1	Missense
9:107599789_A/G	1	2.57E-05	NA	NA	NA	NA	29.6	Missense

ABCA1 and ApoA1 (p=3.26e-06)

- The membrane-associated protein encoded by this gene is a member of the superfamily of ATP-binding cassette (ABC) transporters. ABC proteins transport various molecules across extra- and intracellular membranes.
- With cholesterol as its substrate, this protein functions as a cholesterol efflux pump in the cellular lipid removal pathway. Mutations in this gene have been associated with Tangier's disease and familial high-density lipoprotein deficiency. [provided by RefSeq, Jul 2008]
- Gene previously associated with Cholesterol (Willer et al, 2013), Lipid metabolism phenotypes (Kettunen et al, 2012), Metabolite levels (Inouye et al, 2012), Metabolic syndrome (Kristiansson et al, 2012), Coronary heart disease (Lettre et al, 2011), HDL cholesterol (Waterworth et al, 2010; Willer et al, 2008)

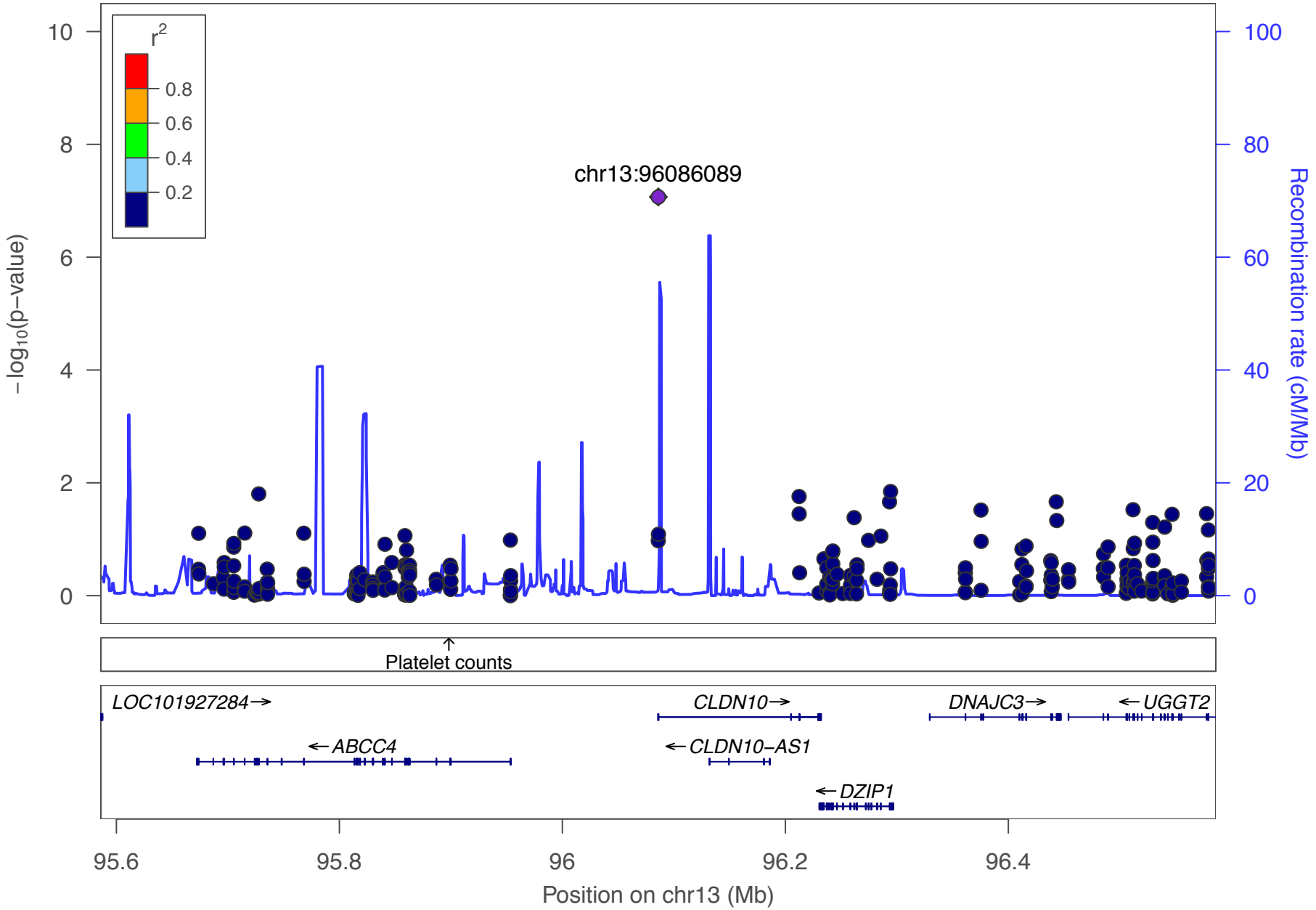

chr9:107548661
r2
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
HDL cholesterol
Cholesterol, total
HDL cholesterol
5 GWAS hits omitted
OR13F1
OR13C9
NIPSNAP3A
SLC44A1
OR13C4
OR13D1
ABCA1
OR13C3
NIPSNAP3B
OR13C8
LOC286367
OR13C5
OR13C2
Position on chr9 (Mb)

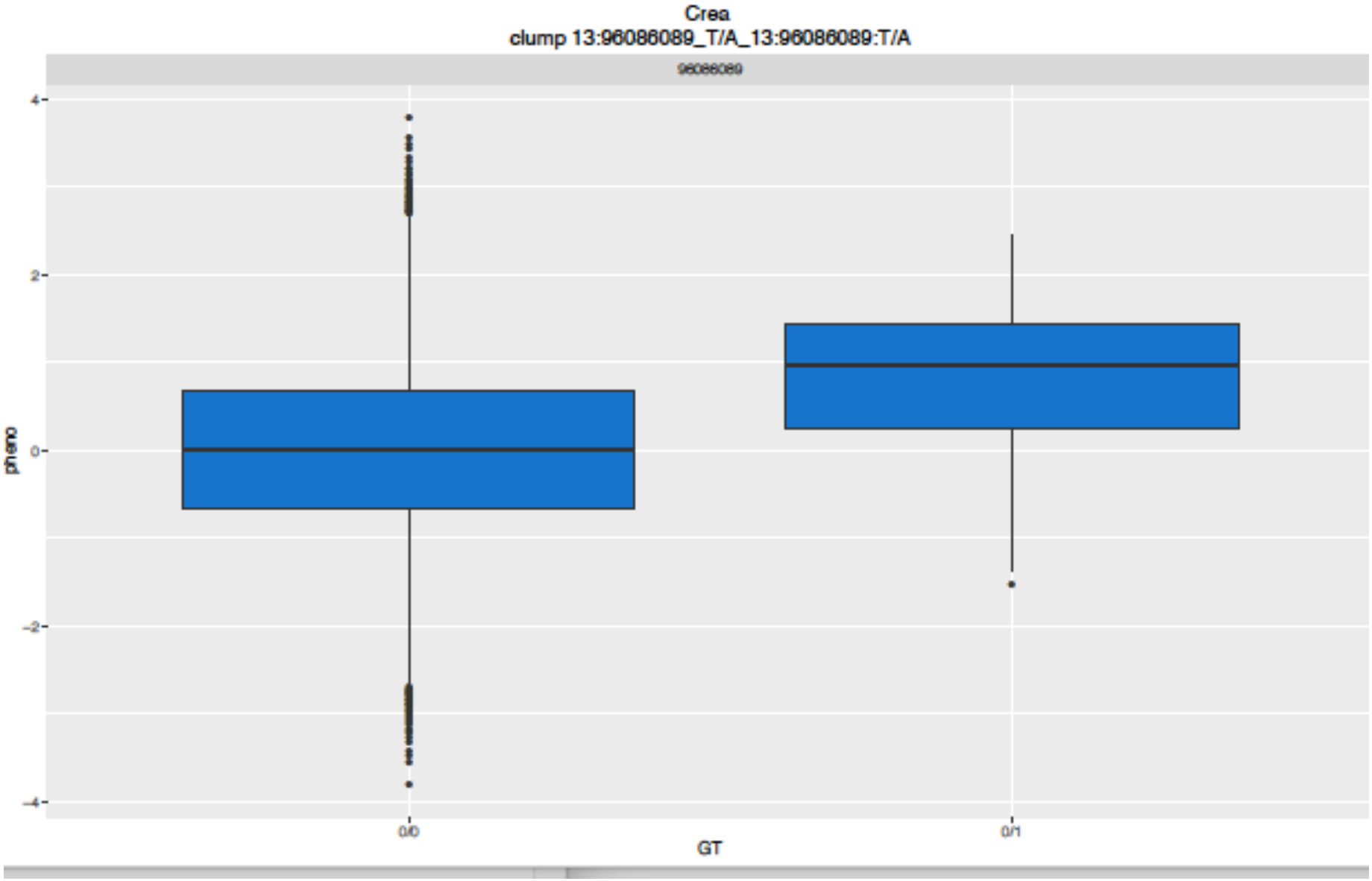
CLDN10 and Creatinine levels (p=2.90E-08)

Variants in gene based test	MAC	MAF	ExAC_AF_FIN	ExAC_AF_NFE	P-value	Beta	CADD	CSQ
13:96086089:T/A (Lead SNP)	40	1.80E-03	1.50E-03	1.50E-05	8.55E-08	0.8487	24.2	Start loss
13:96212394:C/T	4	2.00E-04	1.50E-04	0.00E+00	3.55E-02	1.046	40	Stop gain
13:96229547:G/T	1	5.00E-05	0.00E+00	1.50E-05	NA	NA	22.1	missense

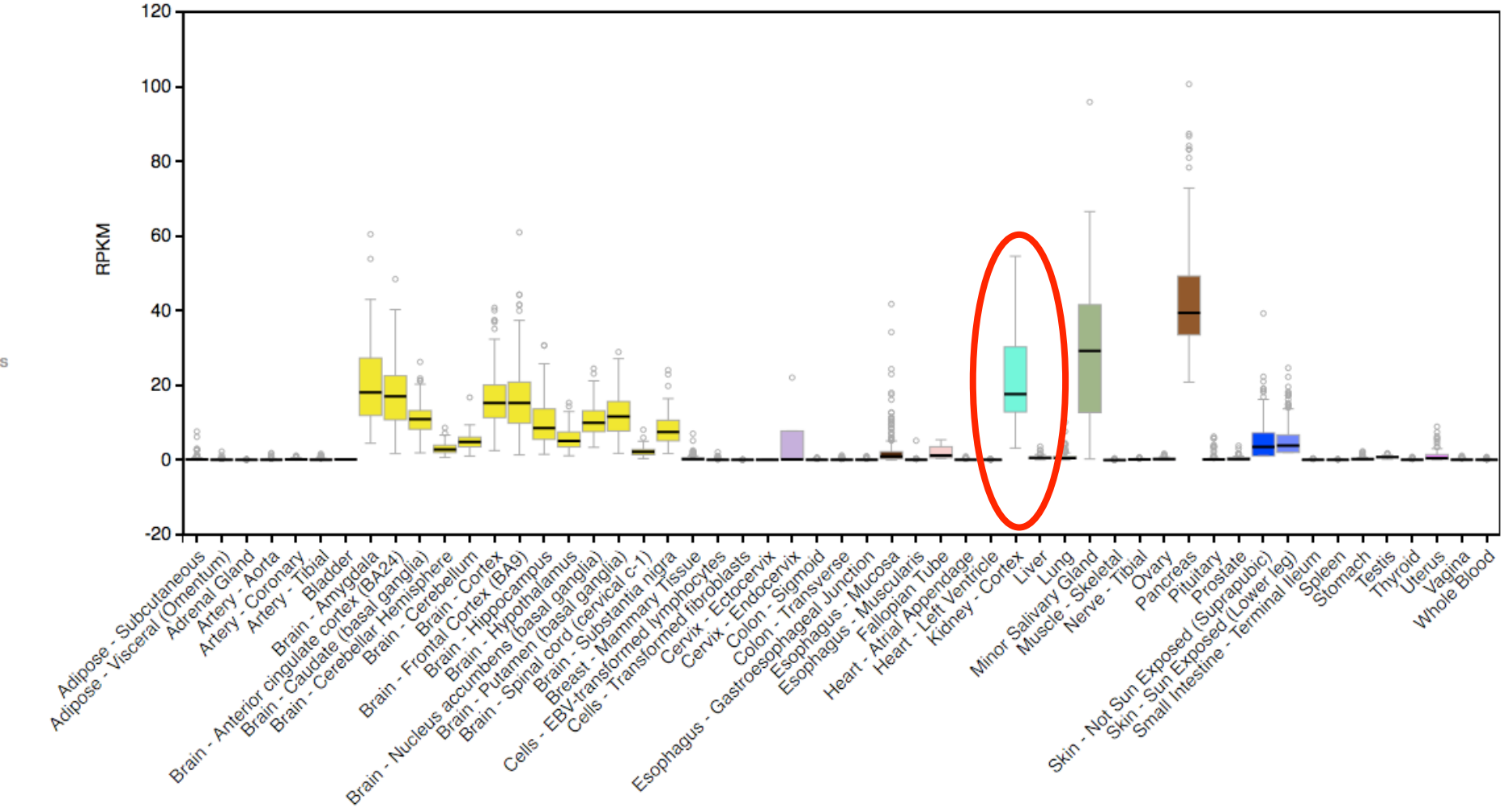
Claudins

- integral membrane proteins and components of tight junction strands
- Plays a major role in tight junction-specific obliteration of the intercellular space, through calcium-independent cell-adhesion activity
- May form permselective paracellular pores
  - isoform 1 appears to create pores preferentially permeable to cations
  - isoform 2 for anions.
- Plays a key role in controlling cation selectivity and transport in the thick ascending limb (TAL) of Henles loop in **kidney**.


r2
0.8
0.6
0.4
0.2
chr13:96086089
−log10(p−value)
Recombination rate (cM/Mb)
Platelet counts
LOC101927284→
←ABCC4
CLDN10→
←CLDN10−AS1
←DZIP1
DNAJC3→
←UGGT2
Position on chr13 (Mb)


Crea
clump 13:96086089_T/A_13:96086089:T/A
96086089
4
2
0
−2
−4
pheno
0/0
0/1
GT

Gene expression from GTEx



Summary

- Call variants using local de novo assembly (GATK HaplotypeCaller)
- Filter variants using hard filters or VQSR
- Decompose and normalize indels
- Filter variants based on many metrics
  - Ti/Tv, HWE, het/hom, missingness
- Annotate variants
- For small, family based studies GEMINI can be used to analyze inheritance patterns of rare variants
- For large, case-control cohort studies rare variants can be identified using gene based association tests