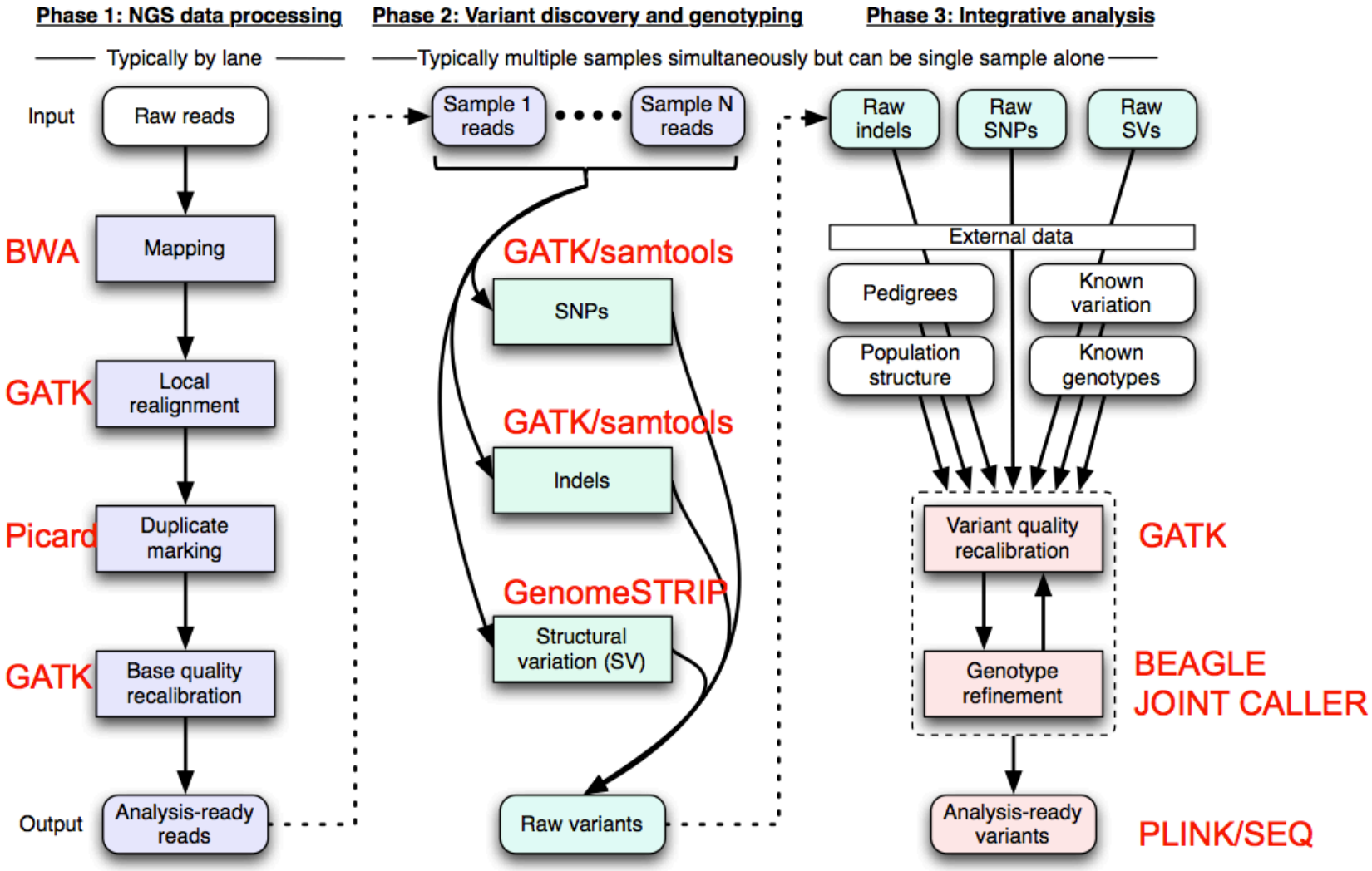
Analysis of next-generation sequencing data: PSEQ practical

Shaun Purcell

shaun.purcell@mssm.edu

Boulder Course, CO
March 2013

The Picard/GATK NGS analysis pipeline



example.vcf

```
##fileformat=VCF4.0
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total read depth">
##INFO=<ID=HM,Number=0,Type=Flag,Description="Seen in HapMap 2 or 3">
##FILTER=<ID=q10,Description="Quality below 10">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype quality">
#CHROM   POS    ID      REF  ALT  QUAL      FILTER   INFO      FORMAT   S001      S002
1        1001   .       A    G    55        PASS     DP=82     GT:GQ    0/0:56    0/1:80
1        8888   rs123   T    A    5         q10      DP=8;HM   GT:GQ    ./.:.     0|1:58
```

Numeric allele encoding & multi-allelic sites:

If REF is A, ALT is T 0/0 is A/A 0/1 is A/T

If REF is A, ALT is T,C 0/2 is A/C 2/1 is C/T

Missing genotype ./.

Representing haplotype phase

0/1 unphased heterozygote

1|0 phased with respect to previous site (implies 01/10 haplotypes)

(Encoding of haplotypic information will be more explicit in future VCF specs.)

- What does this genotype mean?

```
GT   : AD       : DP : GQ : PL
0/0:366,11:200:99:0,600,5980
```

- GT hard genotype call
- AD and DP read-depth information
- GQ quality score
- PL is (phred-scaled) *genotype-likelihoods* (soft-calls)

- VCF likely to be primary out of imputation packages in the future, as it can represent hard-calls (most likely genotype) but also the expected dosage and/or posterior probabilities (and also R^2 in the INFO field, etc)

Phred-scale

Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90 %
20	1 in 100	99 %
30	1 in 1000	99.9 %
40	1 in 10000	99.99 %
50	1 in 100000	99.999 %

$$Q = -10 \log_{10} P$$

$$P = 10^{\frac{-Q}{10}}$$

GT:AD:DP:GQ:PL 0/1:28,35:62:99:1151,0,889

Heterozygote genotype; 28 reference reads, 35 alternate; 62 of 63 reads used in calling; genotype quality score of 99; Phred-scaled genotyped likelihoods {1151,0,889}

Genotype	VCF	PL Phred-scaled genotype likelihood	GL Genotype likelihoods P(read data | true genotype)	P(genotype | read data) assuming flat priors
Homozygous reference	0/0	1151	0.000	0.000
Heterozygote	0/1	0	1.000	1.000
Homozygous alternate	1/1	889	0.000	0.000

GT:AD:DP:GQ:PL 0/0:174,11:1:3:0,3,25

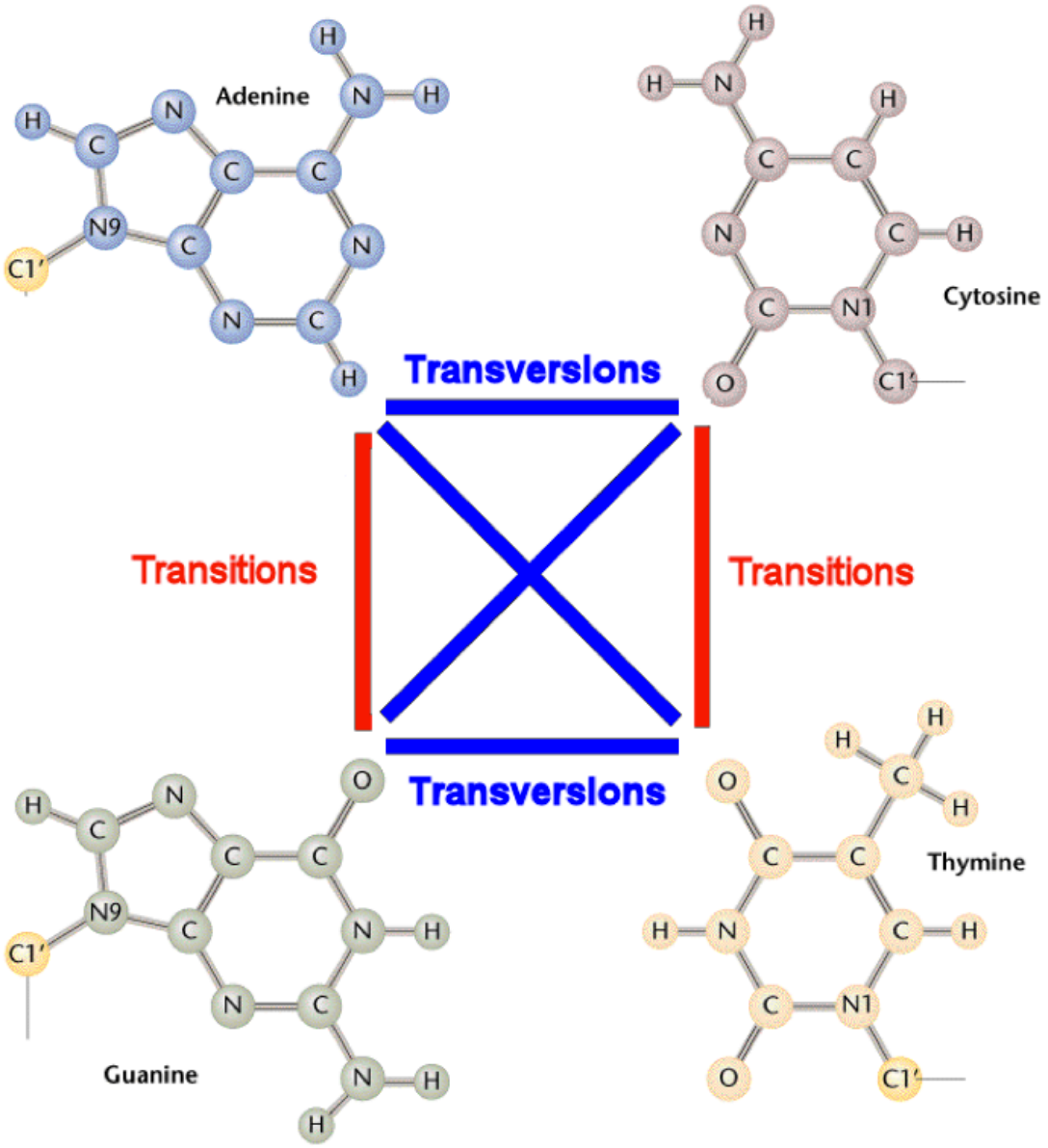
Homozygous reference genotype; 174 reference reads, 11 alternate; but only 1 read used for calling; genotype quality score of 3 (0-99); Phred-scaled genotyped likelihoods {0,3,25}

Genotype	VCF	PL Phred-scaled genotype likelihood	GL Genotype likelihoods P(read data | true genotype)	P(genotype | read data) assuming flat priors
Homozygous reference	0/0	0	1.000	0.66
Heterozygote	0/1	3	0.501	0.33
Homozygous alternate	1/1	25	0.003	0.00

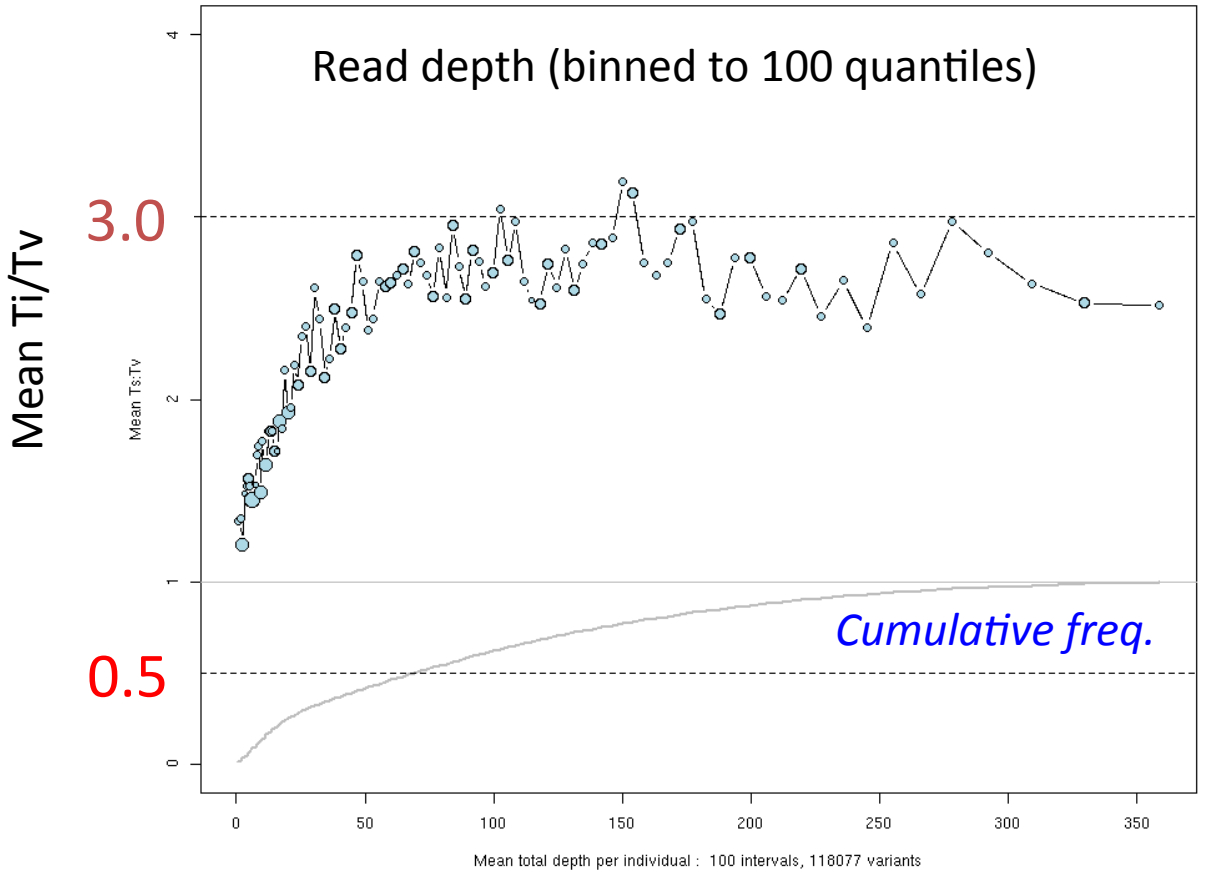
Variant (and genotype) annotations

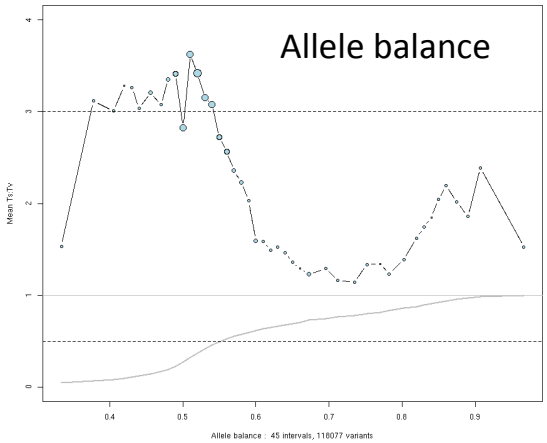
- **Technical**
  - Read depth, allele balance, mean mapping quality, etc
  - QC Filters (PASS, or reasons for exclusion)
  - Hardy-Weinberg disequilibrium
  - (Distribution of) individual genotype likelihoods
- **Population**
  - **Novelty (presence in dbSNP and/or 1000Genomes)**
  - Population frequencies
  - Linkage disequilibrium/phase information on individual genotypes
  - Prior disease/functional associations
- **Genomic**
  - **Gene and coding status**
  - **Transition/transversion**
  - Ancestral/derived allele status

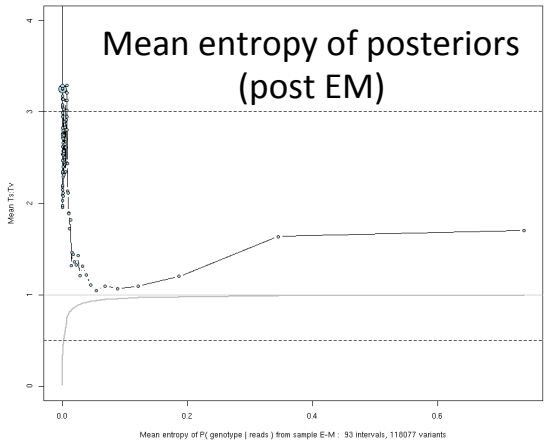
two-ring **purines** one-ring **pyrimidines**

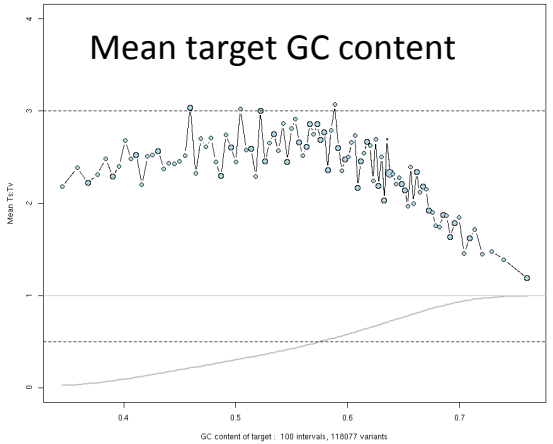


http://www.mun.ca/biology/scarr/Transitions_vs_Transversions.html


Read depth (binned to 100 quantiles)
Mean Ti/Tv
3.0
0.5
Cumulative freq.
Mean Ts:Tv
Mean total depth per individual : 100 intervals, 118077 variants

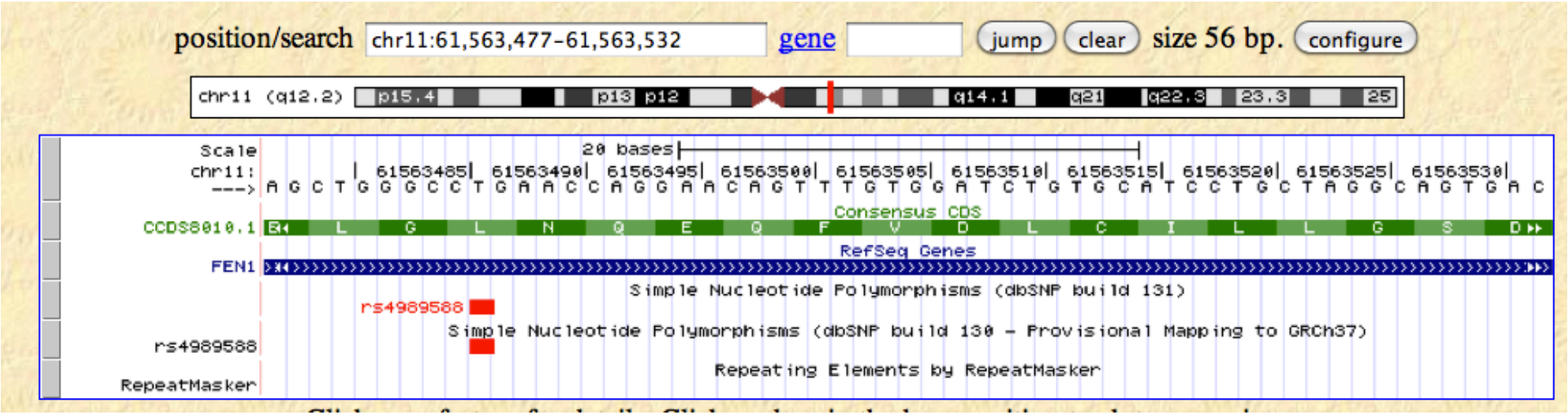

Allele balance
Allele balance : 45 intervals, 118077 variants


Mean entropy of posteriors (post EM)
Mean entropy of P(genotype | reads) from sample E-M : 93 intervals, 118077 variants


Mean target GC content
GC content of target : 100 intervals, 118077 variants

AMINO ACID TABLE

First Position		Second Position: U				C				A				G				Third Position
		code	Amio Acid			code	Amio Acid			code	Amio Acid			code	Amio Acid			
U		UUU	phe			UCU	ser			UAU	tyr			UGU	cys			U
		UUC				UCC				UAC				UGC				C
		UUA	leu			UCA				UAA	STOP			UGA	STOP			A
		UUG				UCG				UAG	STOP			UGG	trp			G
C		CUU	leu			CCU	pro			CAU	his			CGU	arg			U
		CUC				CCC				CAC				CGC				C
		CUA				CCA				CAA	gln			CGA				A
		CUG				CCG				CAG				CGG				G
A		AUU	ile			ACU	thr			AAU	asn			AGU	ser			U
		AUC				ACC				AAC				AGC				C
		AUA				ACA				AAA	lys			AGA	arg			A
		AUG	met			ACG				AAG				AGG				G
G		GUU	val			GCU	ala			GAU	asp			GGU	gly			U
		GUC				GCC				GAC				GGC				C
		GUA				GCA				GAA	glu			GGA				A
		GUG				GCG				GAG				GGG				G

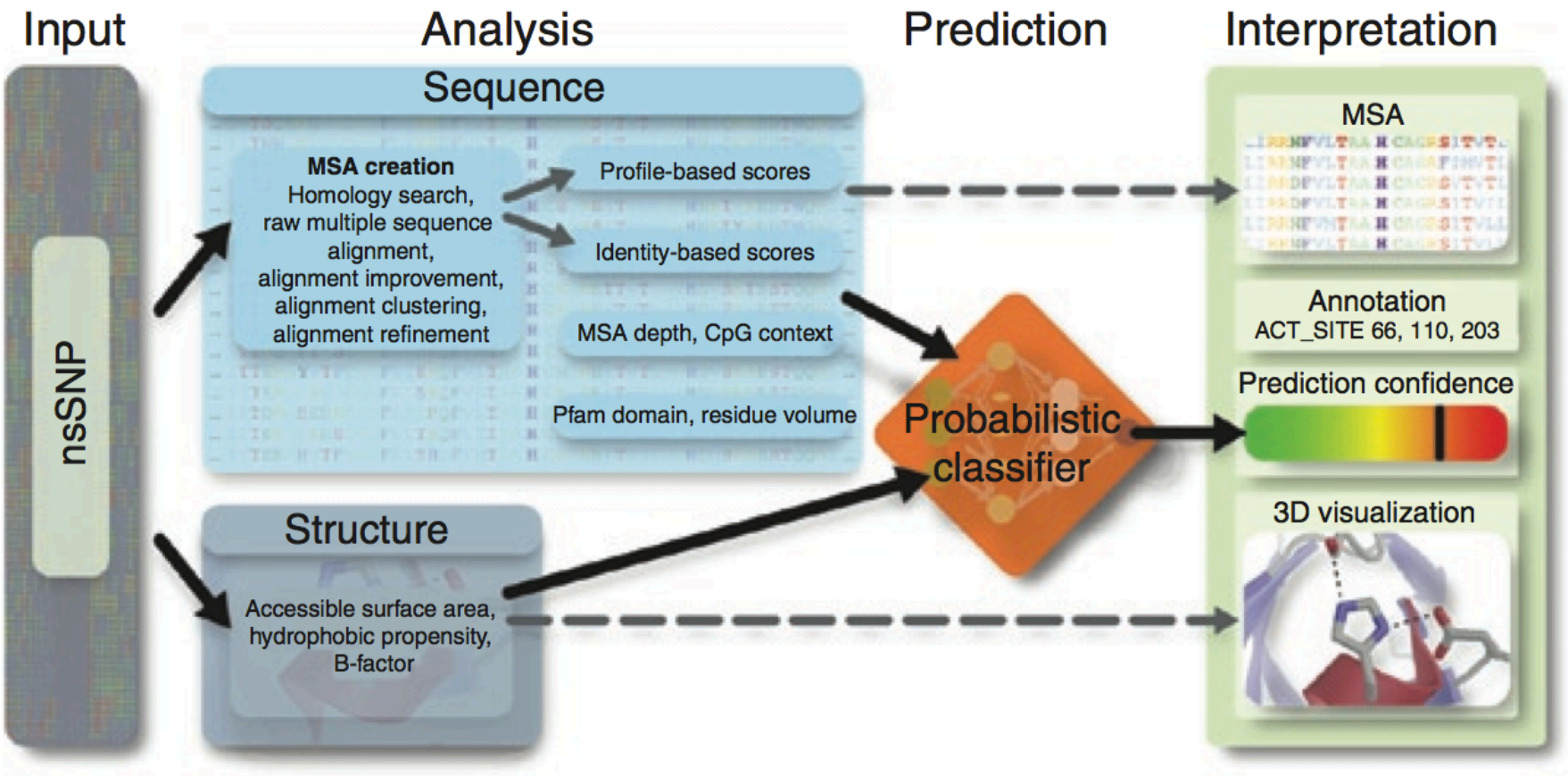


Gene Model(s)							
Function	mRNA				Protein		
	SNP to mRNA	Accession	Position	Allele change	Accession	Position	Residue change
missense	+	NM_004111.4	1025	CTG ⇒ CAG	NP_004102.1	218	L [Leu] ⇒ Q [Gln]

In addition to assigning missense/nonsense coding status, other complications include:

- Splice-site, UTR, regions
- Frameshift mutations (indels)
- Multi-nucleotide polymorphisms
- Full haplotype-based per-individual annotation & compound heterozygosity
- Nonsense-medicated decay
- Ranking of missense variants
- Genomic annotation for non-coding variants

PolyPhen2: predicting the damaging effects of missense mutations



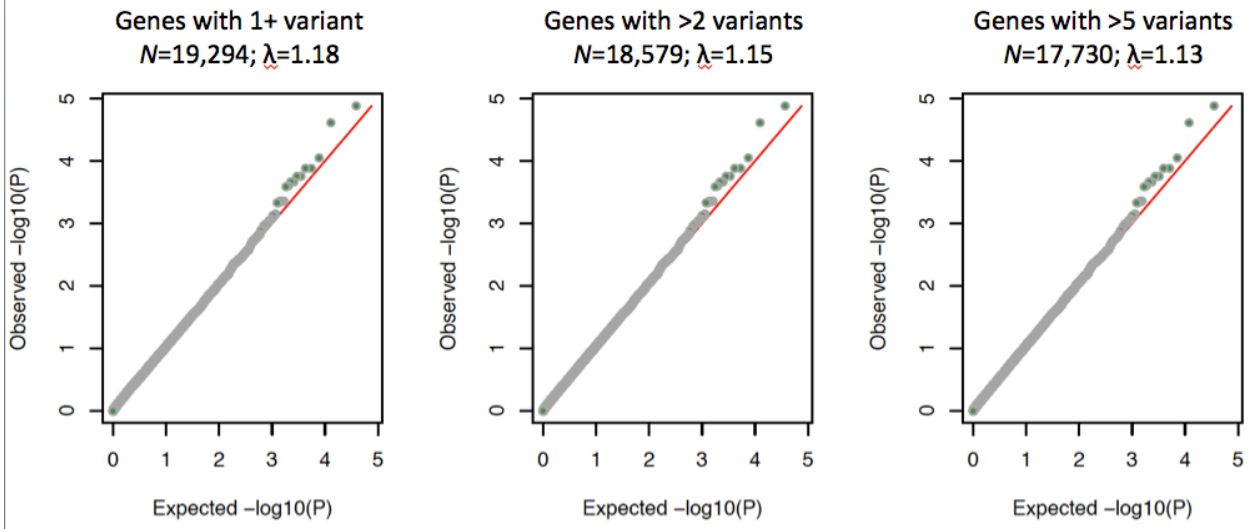
Adzhubei *et al*. Nature Methods (2010).

Functional annotation of variants

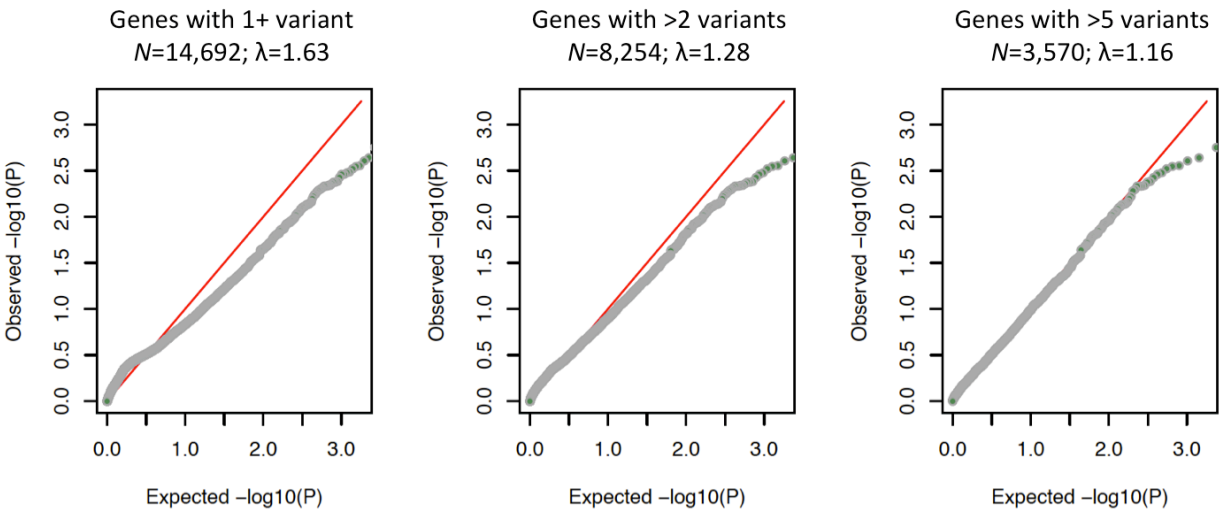
Type	Number	MAF	% singletons
Intronic	101,799	0.066	25
Silent	89,732	0.047	40
Missense	136,847	0.025	51
Benign	60,626	0.038	45
Possibly damaging	22,676	0.020	52
Probably damaging	44,592	0.010	59
Essential splice site	5,020	0.016	54
Nonsense	2,902	0.010	66

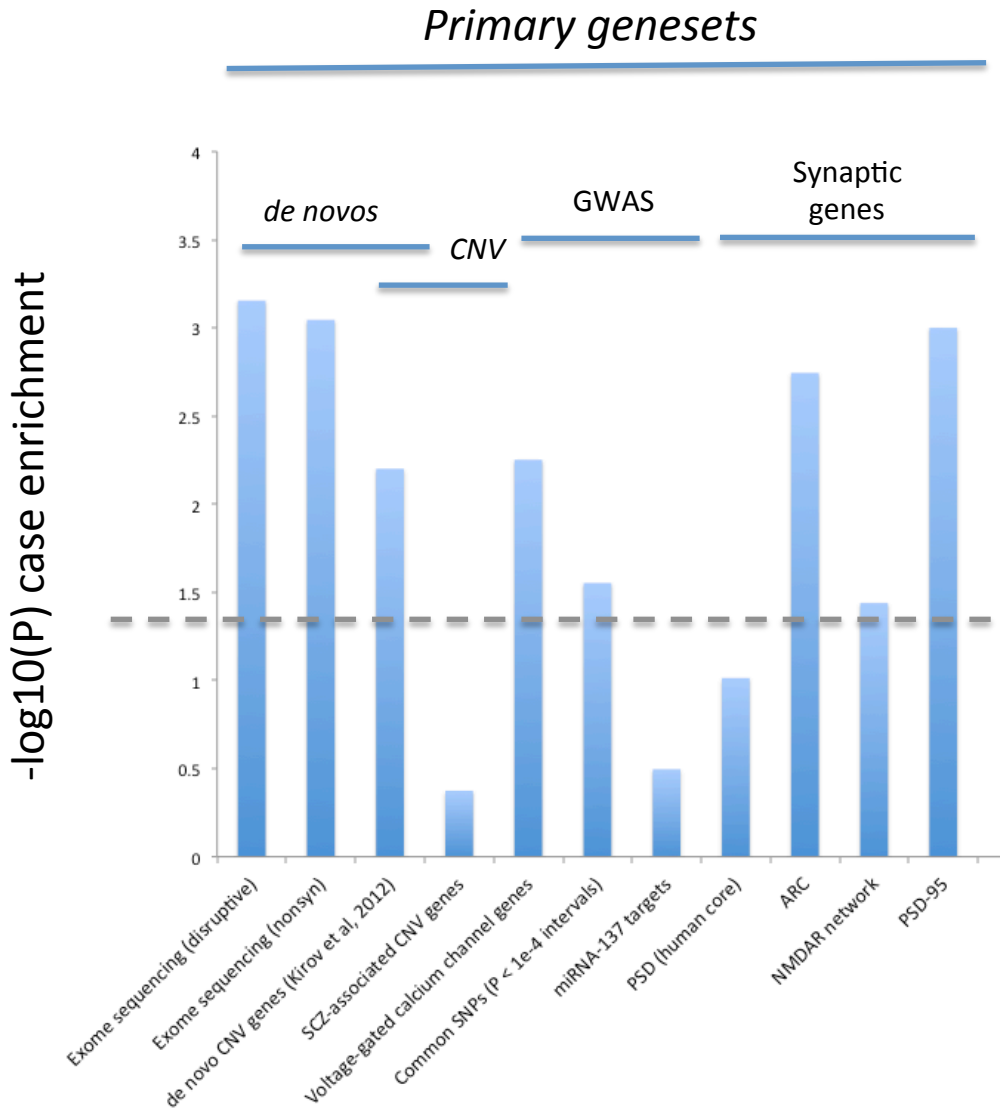
(based on RefSeq transcripts and hg19; missense ranking w/ PolyPhen2)

SKAT | all variants | MDS covariates



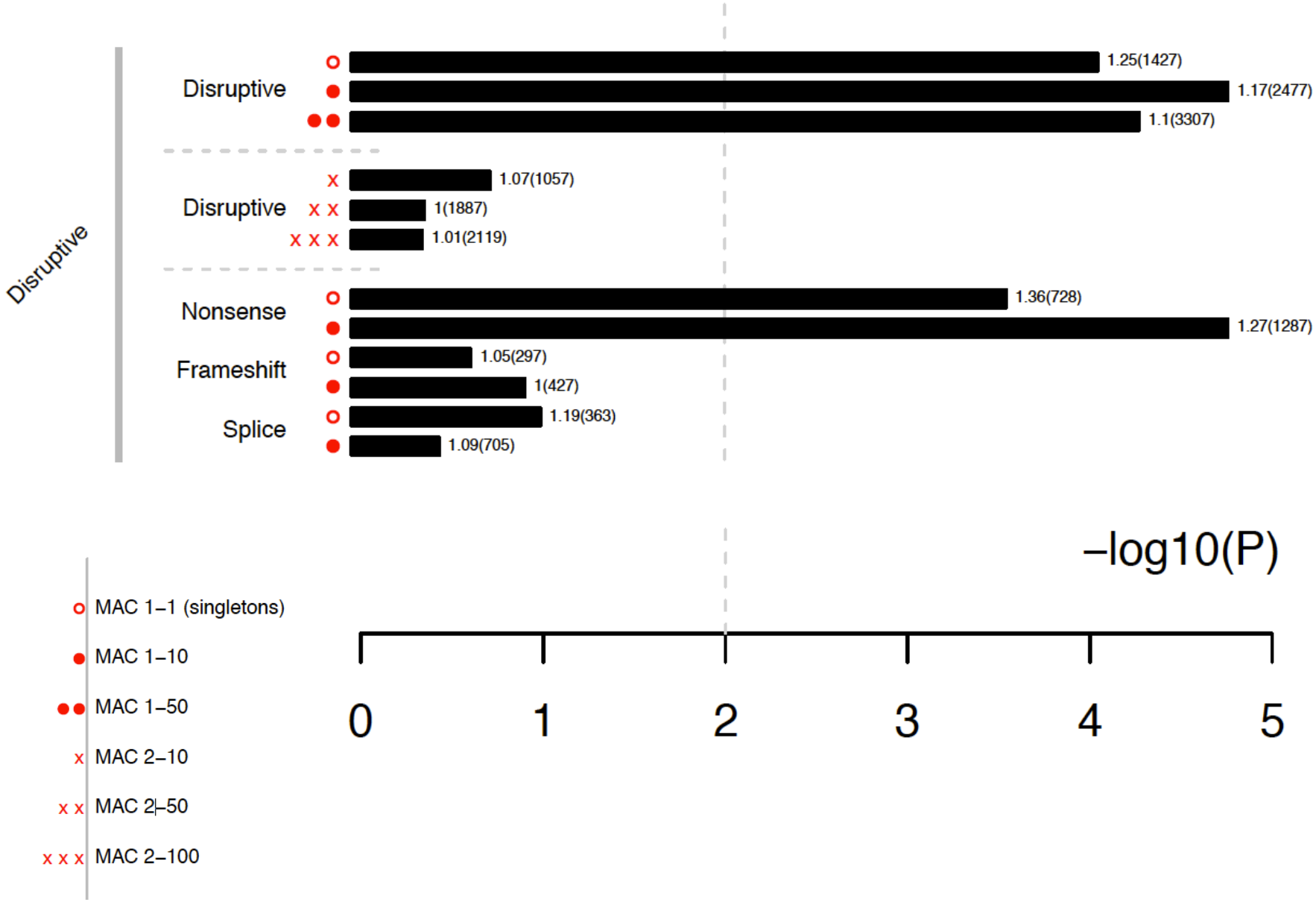
SKAT | nonsyn. (strict) variants| MDS covariates

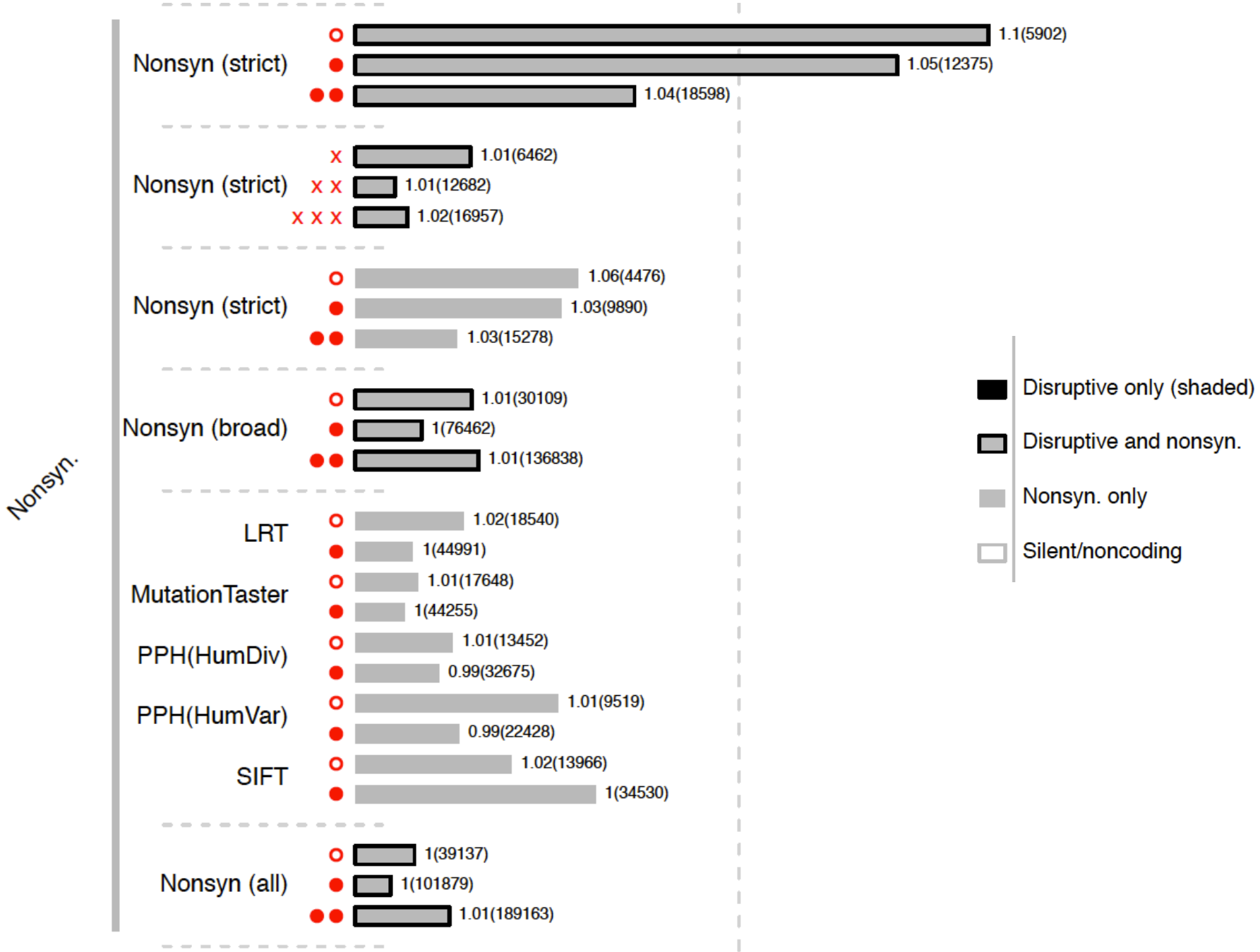


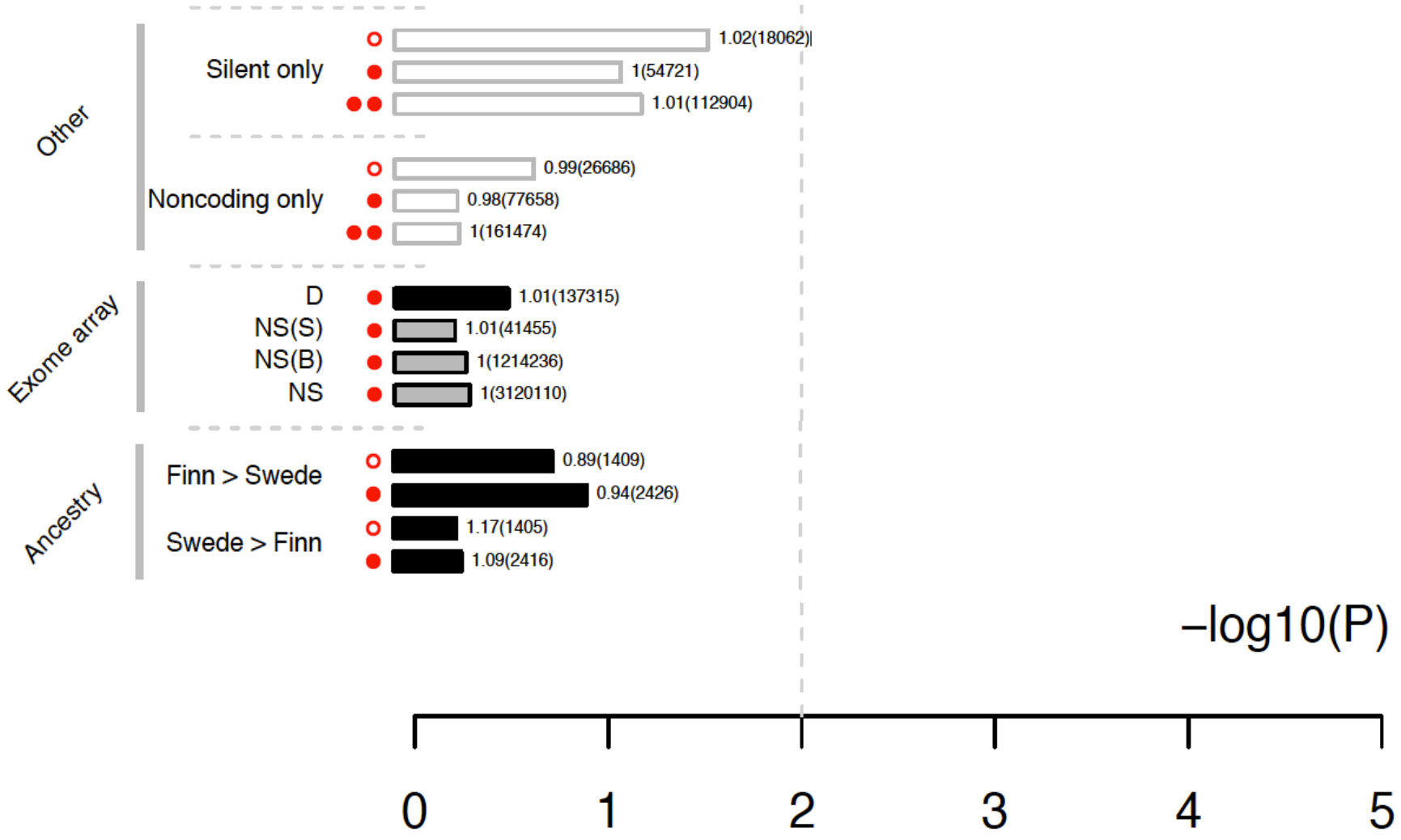


Summary of case burden geneset enrichment results for disruptive mutations (MAF<0.1%)

Characterizing this increased burden: a “core” gene-set of 1858 genes




Nonsyn.
Nonsyn (strict)
1.1(5902)
1.05(12375)
1.04(18598)
Nonsyn (strict)
1.01(6462)
1.01(12682)
1.02(16957)
Nonsyn (strict)
1.06(4476)
1.03(9890)
1.03(15278)
Nonsyn (broad)
1.01(30109)
1(76462)
1.01(136838)
LRT
1.02(18540)
1(44991)
MutationTaster
1.01(17648)
1(44255)
PPH(HumDiv)
1.01(13452)
0.99(32675)
PPH(HumVar)
1.01(9519)
0.99(22428)
SIFT
1.02(13966)
1(34530)
Nonsyn (all)
1(39137)
1(101879)
1.01(189163)
Disruptive only (shaded)
Disruptive and nonsyn.
Nonsyn. only
Silent/noncoding


Other
Silent only
1.02(18062)
1(54721)
1.01(112904)
Noncoding only
0.99(26686)
0.98(77658)
1(161474)
Exome array
D
1.01(137315)
NS(S)
1.01(41455)
NS(B)
1(1214236)
NS
1(3120110)
Ancestry
Finn > Swede
0.89(1409)
0.94(2426)
Swede > Finn
1.17(1405)
1.09(2416)
−log10(P)
0
1
2
3
4
5

Working with NGS data / VCFs

- "How do I…"

  … find variants in a specific gene in my dataset?

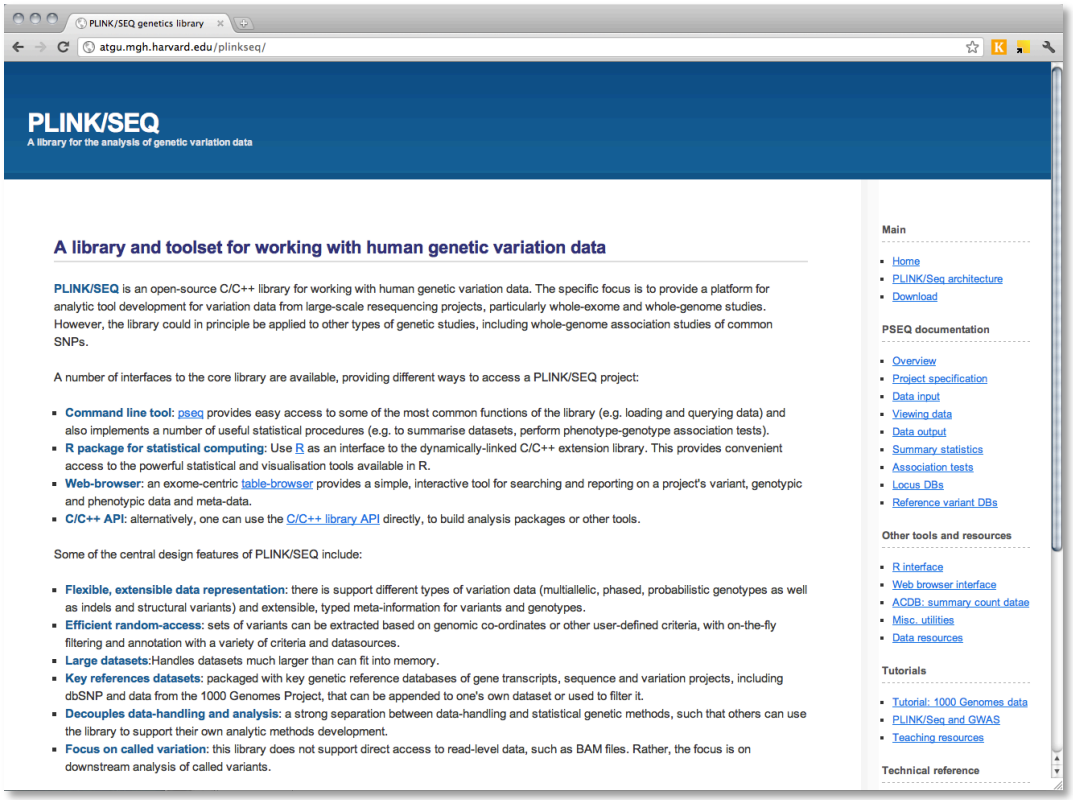
  … lookup a list of variants?

  … annotate a list of sites?

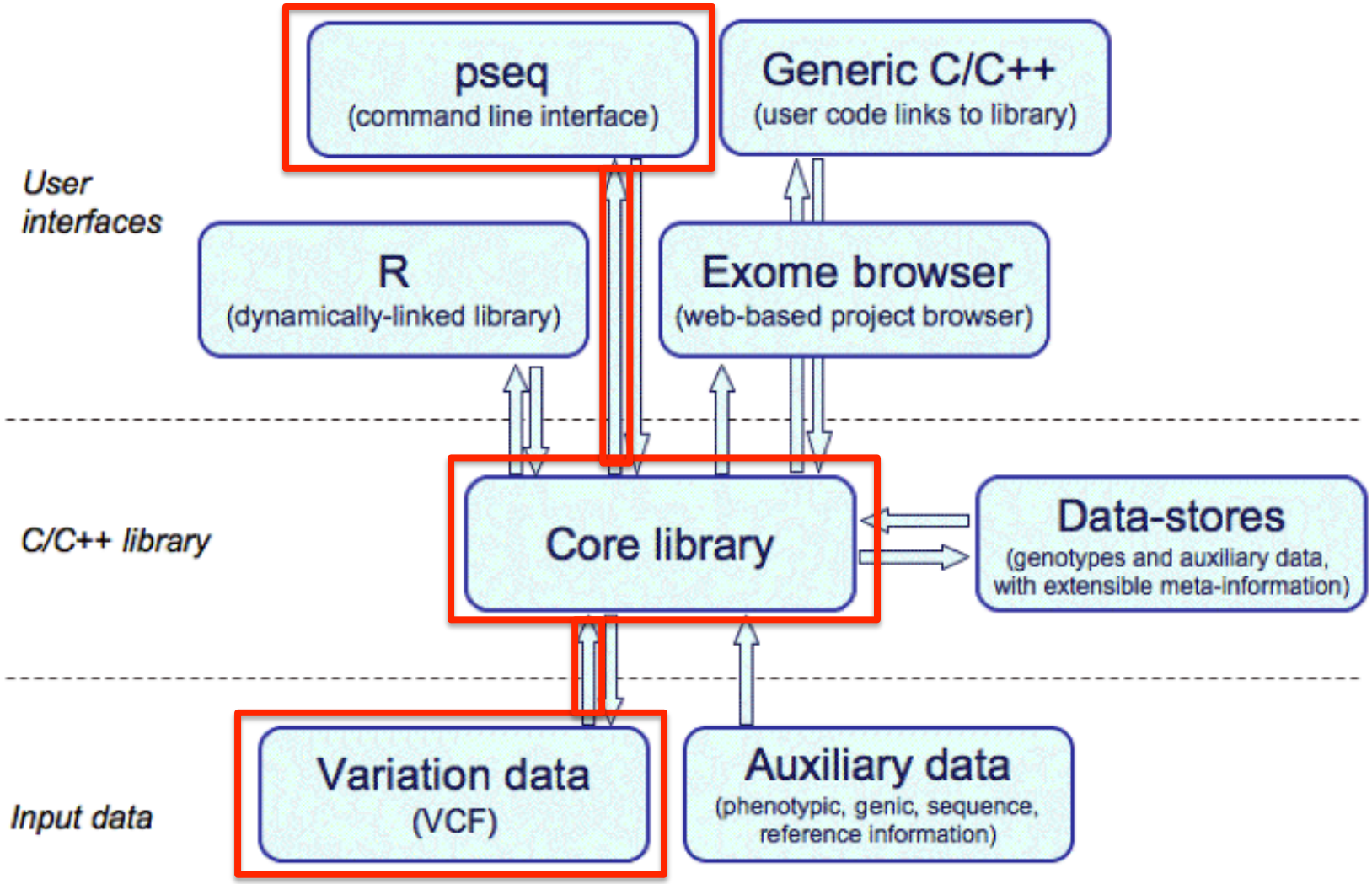
  … get summary QC metrics for regions, samples?

PLINK/Seq : a toolset for NGS variation datasets

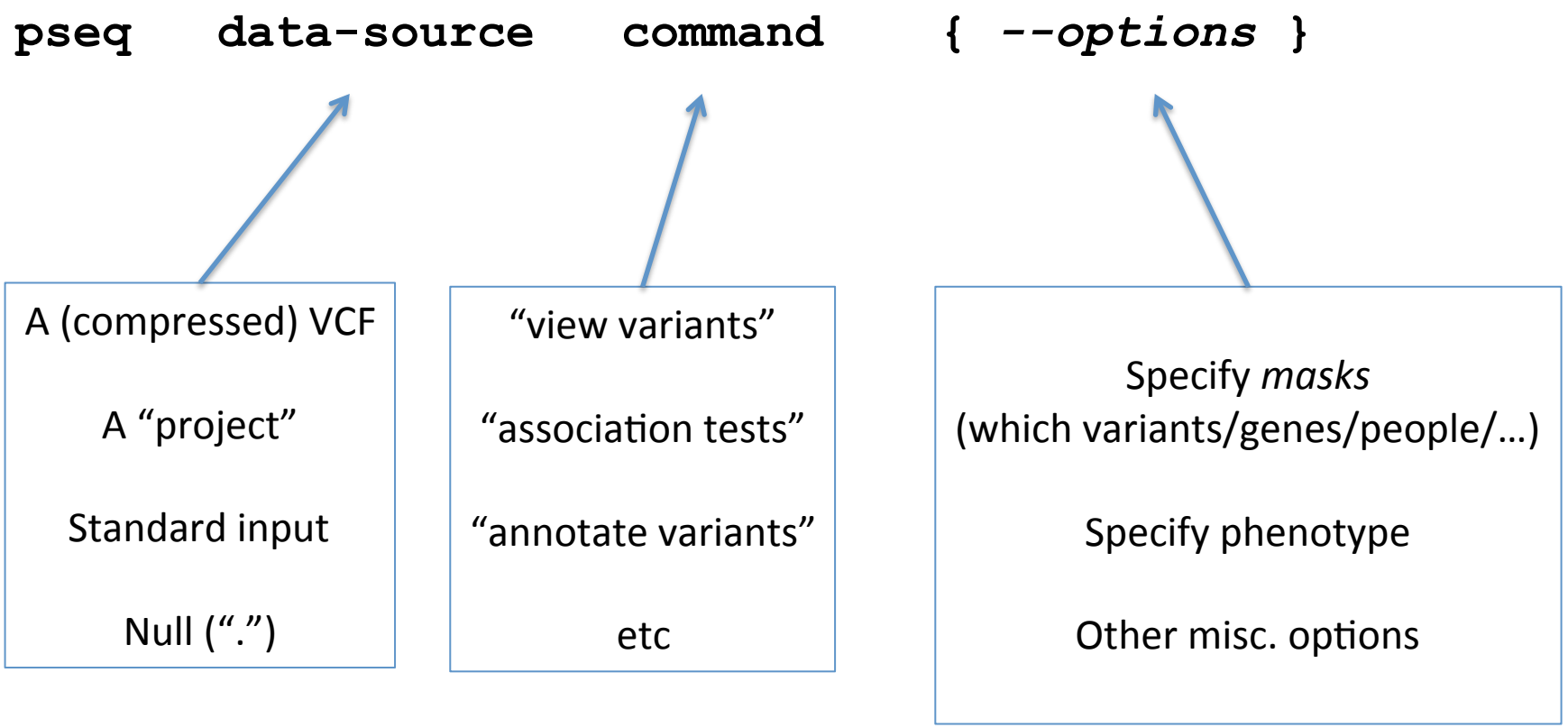
http://atgu.mgh.harvard.edu/plinkseq/



- VCF as primary input
- Focus on analysis of rare variation
- Extensible meta-information on locus, genotypes, individuals
- Bundled with key reference databases that can be directly intersected with one's own project
- Command-line and R library; web-based GUI under-development


pseq
(command line interface)
Generic C/C++
(user code links to library)
User interfaces
R
(dynamically-linked library)
Exome browser
(web-based project browser)
C/C++ library
Core library
Data-stores
(genotypes and auxiliary data, with extensible meta-information)
Input data
Variation data
(VCF)
Auxiliary data
(phenotypic, genic, sequence, reference information)

Basic structure of *PSEQ* commands



Note: other types of genotypic data can be incorporated into existing projects:

PLINK files (BED/BIM/FAM format)
"dosage" files, post imputation
BCF2 files

Three main modes of operation :
given various inclusion/exclusion masks, to iterate over all …

1) Variants
   - e.g. viewing/filtering a VCF
   - calculating various summary statistics

2) Groups of variants (e.g. genes, all nonsense variants)
   - primarily gene-based (or set-based) association tests

3) Individuals
   - per-individual statistics, burden of rare variants, etc

Core project databases

VARDB	Variant information
INDDB	Individual phenotype/covariate information
LOCDB	Gene/transcript information (e.g. RefSeq, CCDS)
SEQDB	Human genome sequence (e.g. hg19)
REFDB	(Annotated) known variants (e.g. dbSNP, HGMD)
NETDB	Network information (e.g. PPI)
PROTDB	Protein domain/motif information (e.g. Pfam)
IBDDB	Pairwise identity-by-descent (IBD) information
SEGDB	Known variants (exclude novel)

Examples of common *masks* (following `--mask`)

`reg=chr2`	Variants on chromosome 2
`reg=chr2:123456`	A specific variant(s) by position
`reg=chr2:1000000..20000000`	Variants by range
`reg=chr2,chr7`	List of ranges/positions
`reg=@pos.txt`	List of regions in file `pos.txt`
`id=rs123456`	Specific variant by ID (from VCF)
`novel`	Variants with no known ID
`novel.ex`	Known variants (exclude novel)
`var=coding`	Variants by *variant set "coding"* from VARDB
`loc=refseq`	Variants in *locus group* "refseq" from LOCDB

Exome data used in this practical

Deep-coverage, whole-exome sequence data from the 1000 Genomes Project, Phase 1

GBR Great British, N=63

TSI Toscans in Italy, N=60

LWK Luhya in Webuye, Kenya, N=27

Overview of practical

Look at 1000 Genomes VCF files directly
Filter and output a new VCF
Look at resource files
Create a PLINK/SEQ “project”
Variant level summary statistics
Individual level summary statistics
Extracting meta-information
Experimenting with filters/masks
Gene-based summaries
Annotation
Single-site metrics
Comparing two VCFs
Ancestry inference and relatedness
Single-site association
Gene-based association
Using pbrowse and Rplinkseq interfaces
Geneset enrichment analyses

- Main script to follow:

- ~pshaun/2013/pseq/extra/commands.txt