Analysis of Natural Selection

Michael G. Campana

Overview

- Many methods of inferring selection from genomic data
- All require additional verification/experiments to ensure selective effects

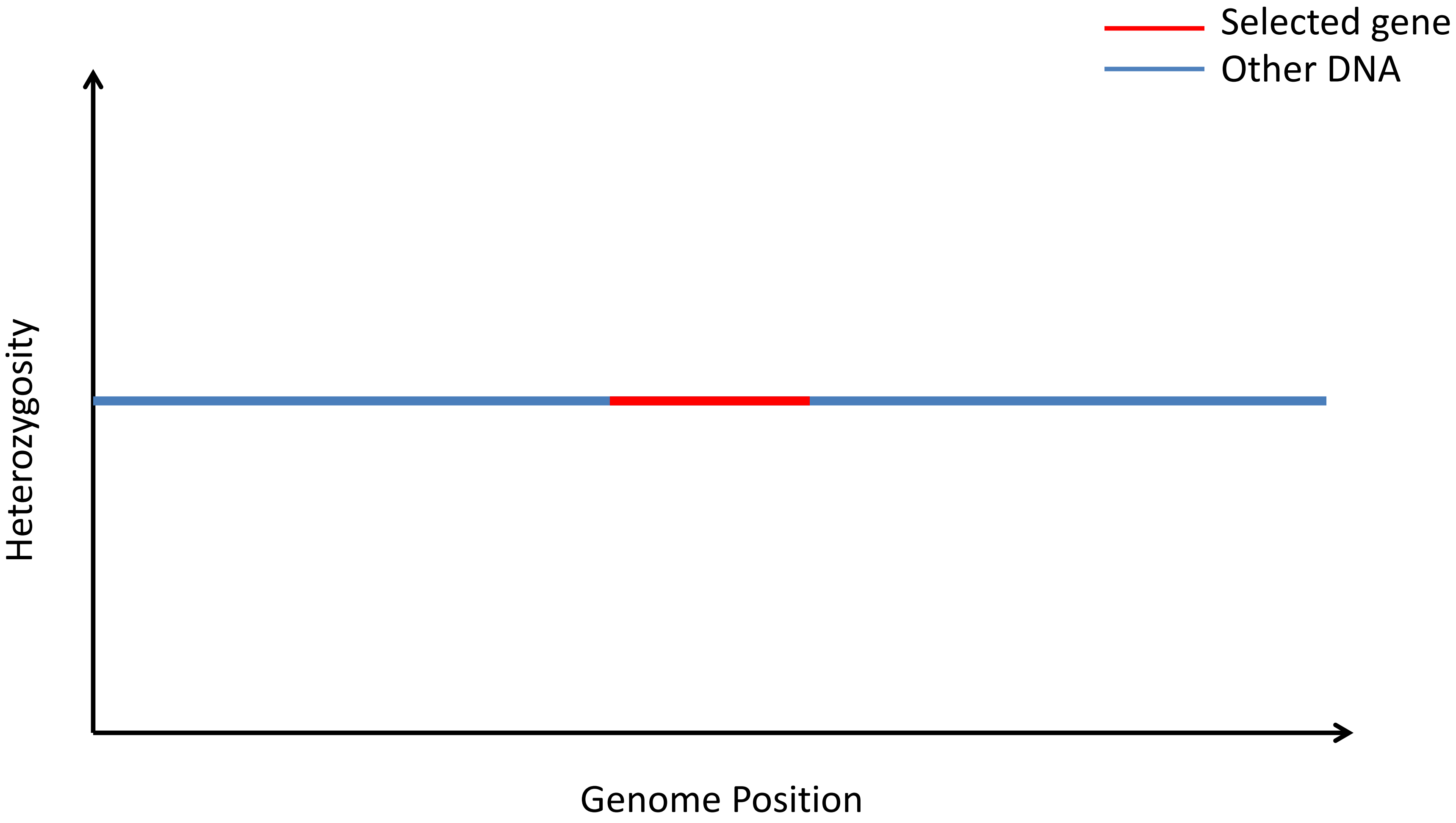
Overview

- Purifying Selection using ROH
- Prediction of SNP effects on genes
- GWAS

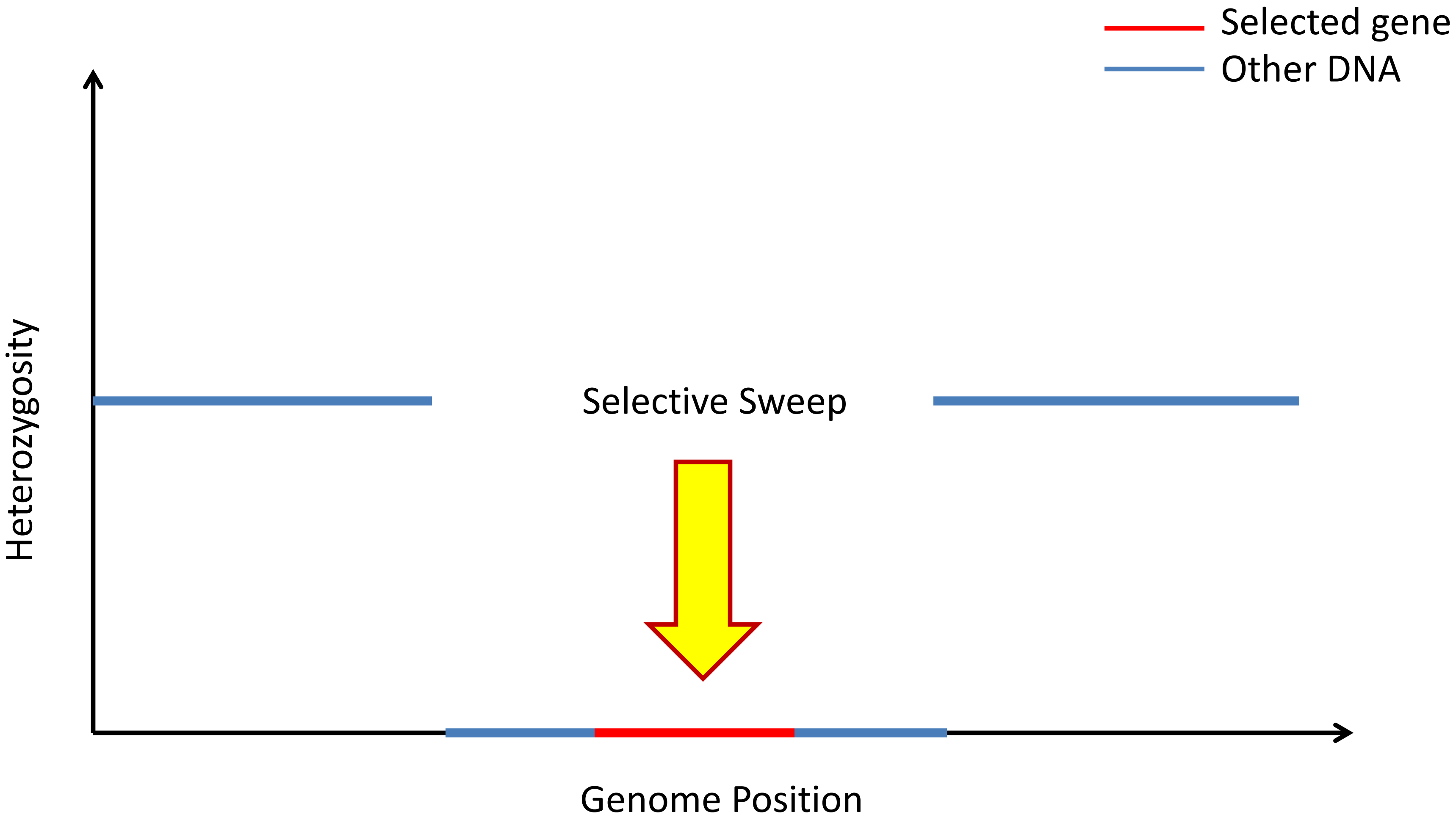
Inferring Purifying Selection

- Purifying selection causes runs-of-homozygosity (ROH) around selected gene
- Length of ROH inversely proportional to time since selective event
- Subsequent recombination reduces ROH length
- ROH algorithms vary greatly in terms of models, output, sensitivity, etc.

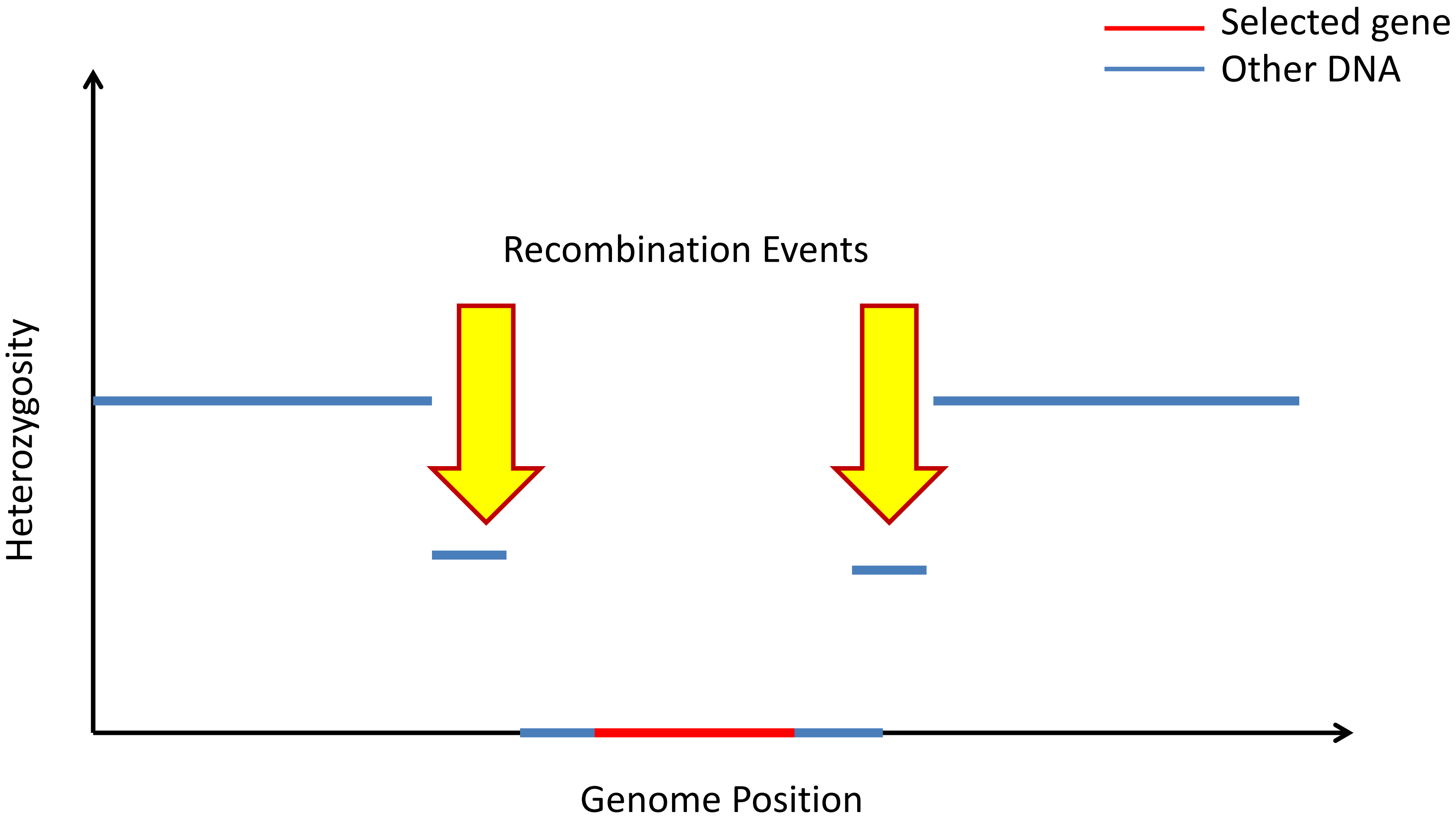
Inferring Purifying Selection

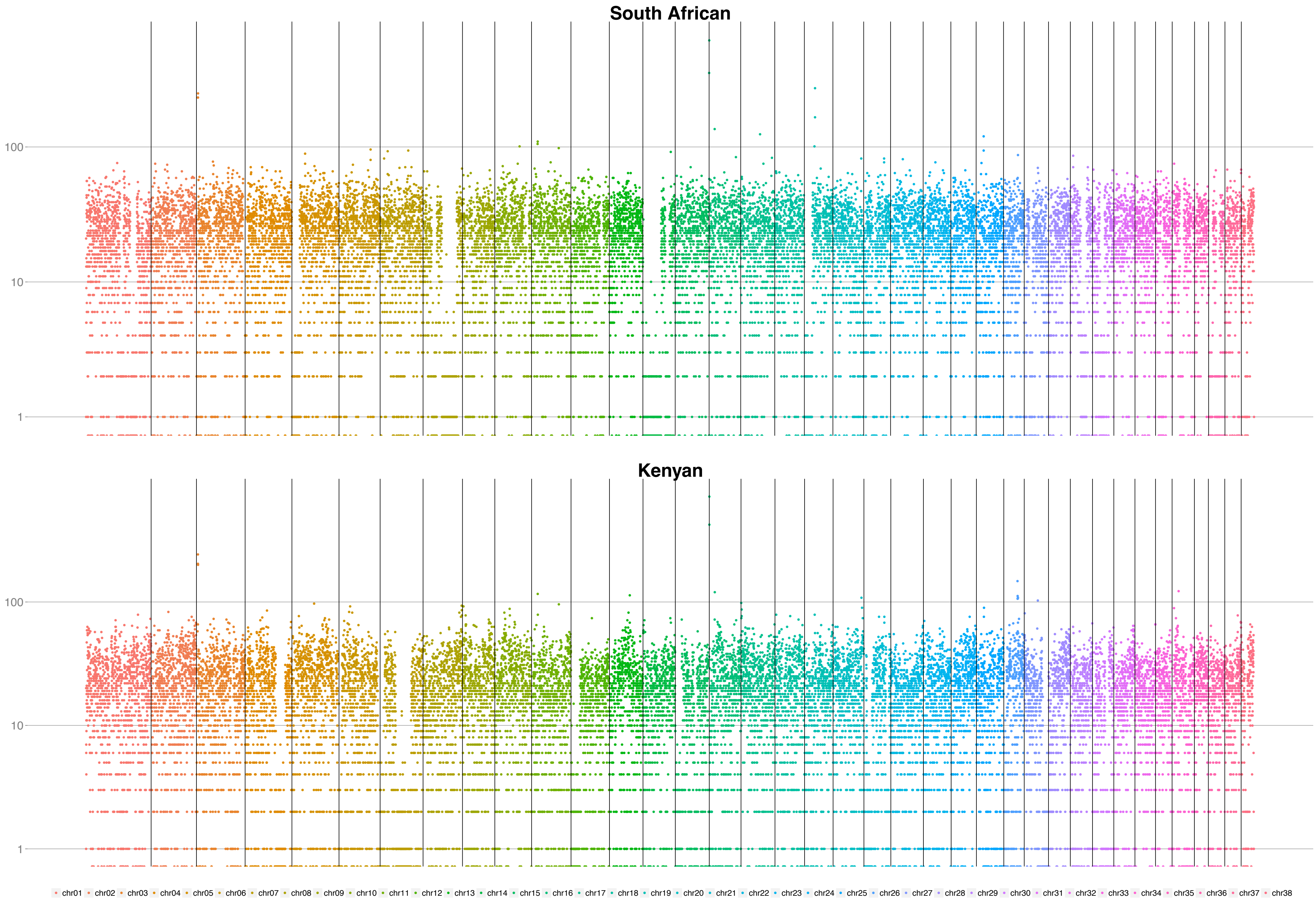


Inferring Purifying Selection



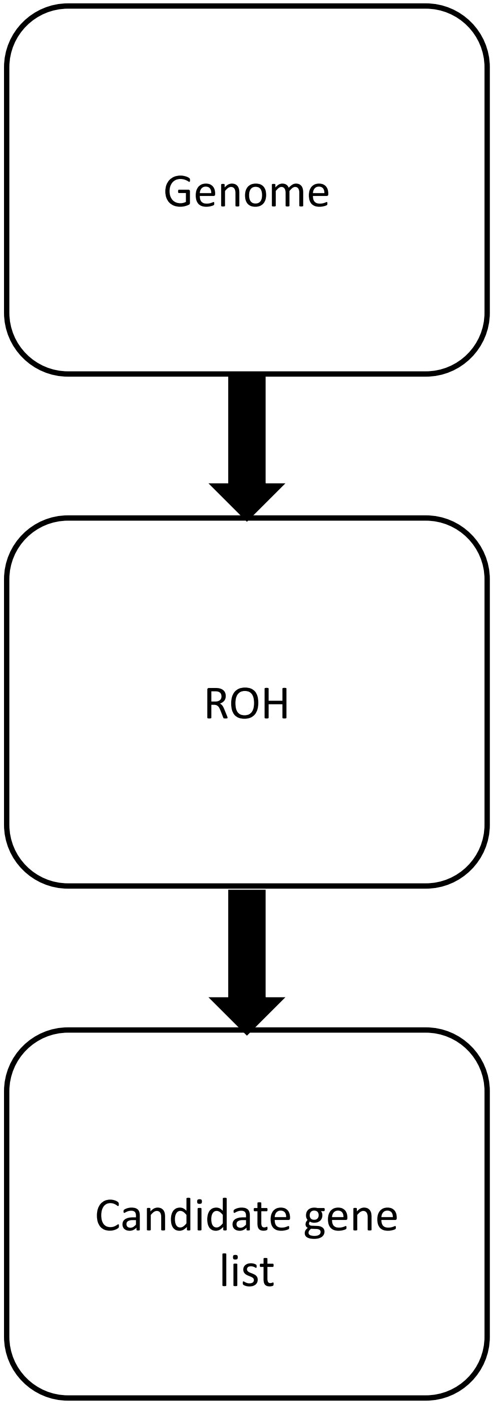
Inferring Purifying Selection




South African
100
10
1
Kenyan
100
10
1
chr01
chr02
chr03
chr04
chr05
chr06
chr07
chr08
chr09
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chr23
chr24
chr25
chr26
chr27
chr28
chr29
chr30
chr31
chr32
chr33
chr34
chr35
chr36
chr37
chr38

Interpreting ROH

- ROH identifies regions under selection, not genes selected nor cause/direction of selection
- Needs further investigation to link ROH to selection
- Inbreeding also causes ROH
- ROH need to be defined for target species due to variation in effective pop sizes, heterozygosity, etc.



Genome

ROH

Candidate gene list

So how do we narrow the gene list?

Variant Effect Prediction

- Some variants are more likely to be functional than others
- E.g. variants in introns are less likely to have an effect
- Non-synonymous substitutions/nonsense substitutions more likely to have functional significance than synonymous substitutions

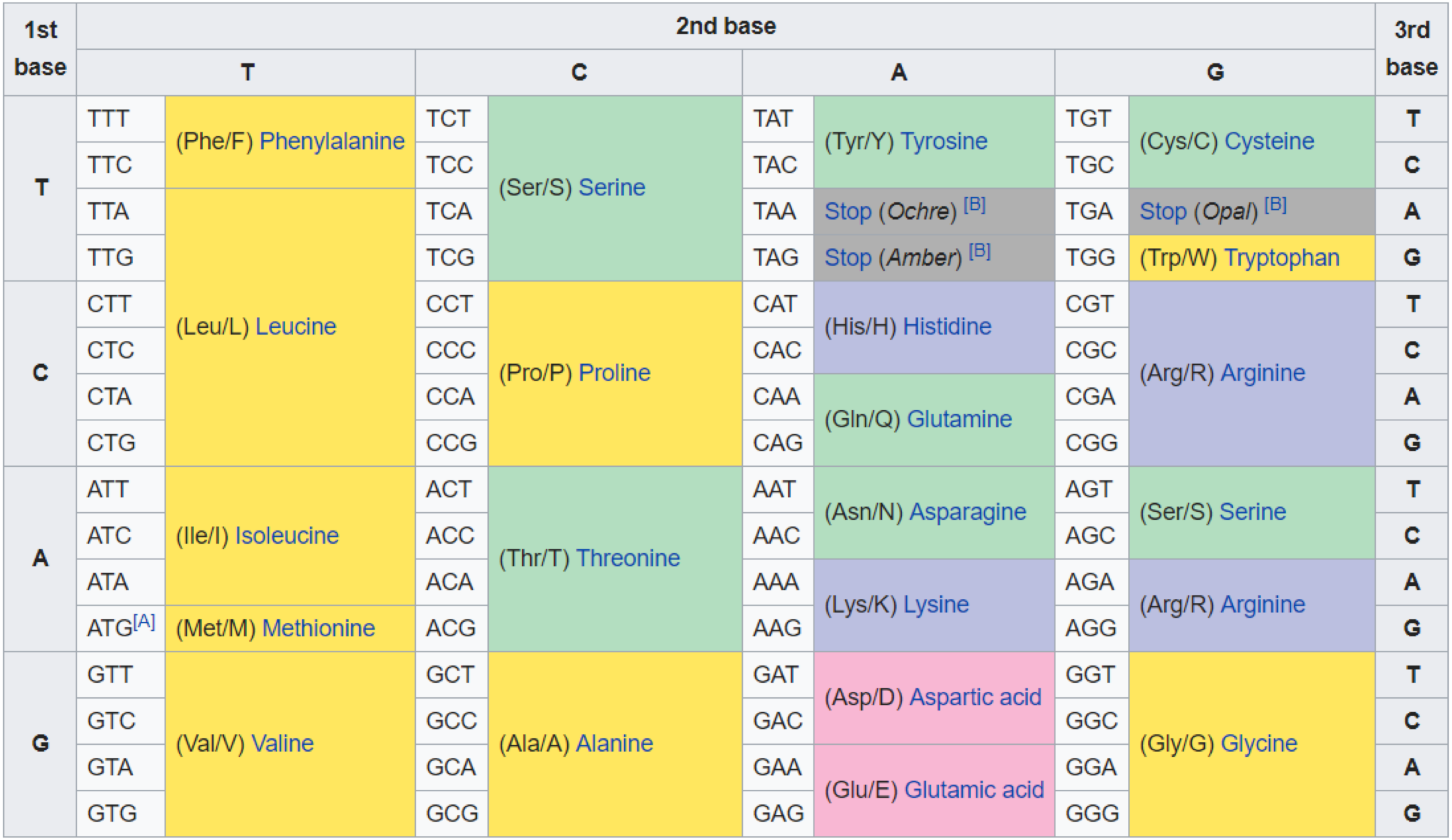
N/S ratio

- Ratio of non-synonymous to synonymous substitutions
- Genes under positive selection >1
- Most genes <1 since most nonsynonymous subs are deleterious (purifying selection)
- HOWEVER – interpreting significant deviations dependent on background *N/S* ratio

dN/dS

- Non-synonymous substitution per non-synonymous site/synonymous substitution per synonymous site
- Interpretation same as *N/S*
- *dN/dS* accounts for codon redundancy (BETTER)

Standard genetic code



1st base	2nd base								3rd base
	T		C		A		G		
T	TTT	(Phe/F) Phenylalanine	TCT	(Ser/S) Serine	TAT	(Tyr/Y) Tyrosine	TGT	(Cys/C) Cysteine	T
	TTC		TCC		TAC		TGC		C
	TTA	(Leu/L) Leucine	TCA		TAA	Stop (*Ochre*) [B]	TGA	Stop (*Opal*) [B]	A
	TTG		TCG		TAG	Stop (*Amber*) [B]	TGG	(Trp/W) Tryptophan	G
C	CTT		CCT	(Pro/P) Proline	CAT	(His/H) Histidine	CGT	(Arg/R) Arginine	T
	CTC		CCC		CAC		CGC		C
	CTA		CCA		CAA	(Gln/Q) Glutamine	CGA		A
	CTG		CCG		CAG		CGG		G
A	ATT	(Ile/I) Isoleucine	ACT	(Thr/T) Threonine	AAT	(Asn/N) Asparagine	AGT	(Ser/S) Serine	T
	ATC		ACC		AAC		AGC		C
	ATA		ACA		AAA	(Lys/K) Lysine	AGA	(Arg/R) Arginine	A
	ATG[A]	(Met/M) Methionine	ACG		AAG		AGG		G
G	GTT	(Val/V) Valine	GCT	(Ala/A) Alanine	GAT	(Asp/D) Aspartic acid	GGT	(Gly/G) Glycine	T
	GTC		GCC		GAC		GGC		C
	GTA		GCA		GAA	(Glu/E) Glutamic acid	GGA		A
	GTG		GCG		GAG		GGG		G

Codon

- Each base in codon can mutate 3 ways (9 total possibilities)
- Not all amino acids have same ratios of nonsynonymous possibilities

Example

- Substitution of TGG → TGA (Trp → Stop)
- N = 1, nonsyn subs for Trp: 9

- Substitution of AGA → AGC (Arg → Ser)
- N = 1, nonsyn subs for Arg: 7

Sites calculation

- Sum of sequence length multiplied by nonsyn/syn proportion at each site

- TGG-AGA (Trp-Arg)
- nonsyn sites = 3 * (9/9) + 3 * (7/9) = 5.333
- syn sites = 3 * (0/9) + 3 * (2/9) = 0.667

Example

- TGG-AGA-AGA → TGA-AGC-AGG
- N = 1, S = 1
- N/S = 1
- nonsyn sites = 7.667
- syn sites = 1.333
- dN = 1/7.667
- dS = 1/1.333
- dN/dS = 0.174

Caveat

- Simple count estimates of *N/S* and *dN/dS* are underestimates
- Back mutations and multiple mutations at a site not counted
- More complex algorithms use ML to estimate these rates to improve accuracy

SnpEff

- Software that models the effects of variants in a VCF by comparison against annotations in a GFF file
- Genes with splice site mutations, nonsense, missense mutations more likely to be under selection

Functional Modelling

- Missense/nonsense mutations/etc are not necessarily selected
- Protein folding algorithms can predict whether mutations have selective effects
- Only functional/transgenic experiments can truly verify function of these mutations

SnpSift

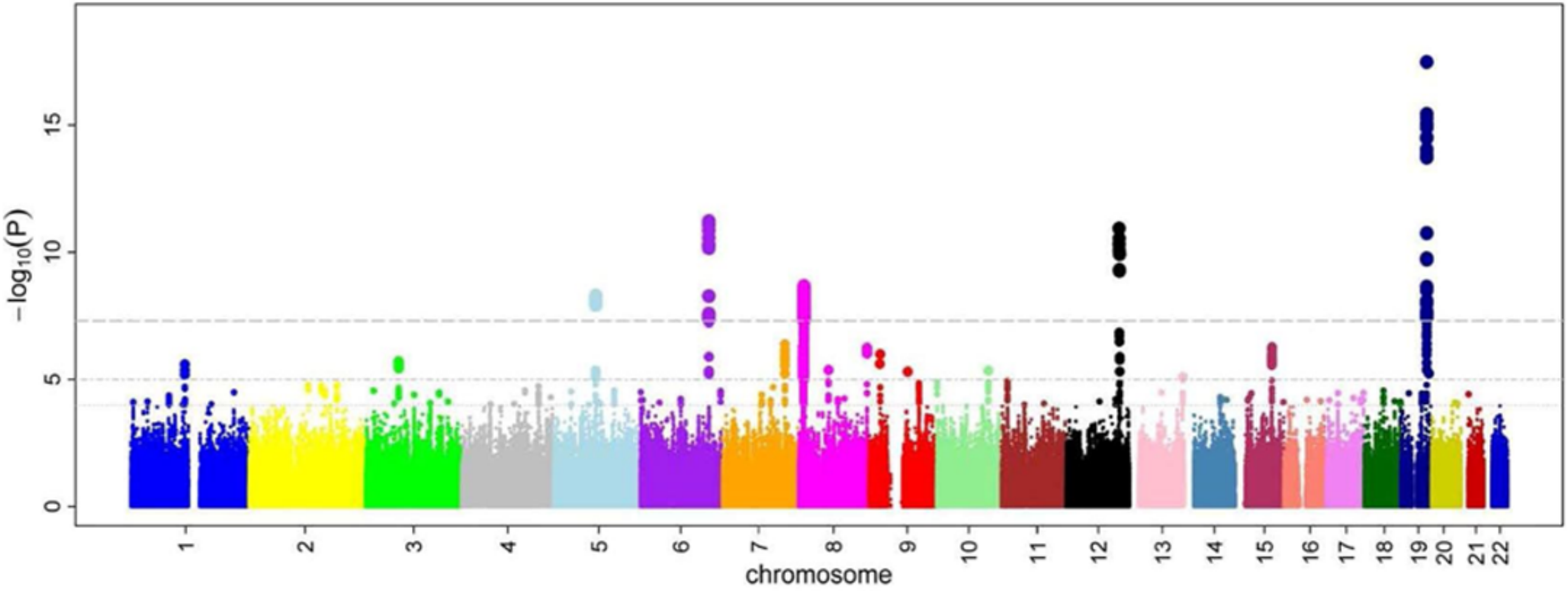
- Extracts genes with classes of mutations from SnpEff models
- I.e. can look for all genes with nonsense mutations

Pathway Annotation

- Known gene functions are categorized using Gene Ontology (GO) terms
- Various other databases (Kyoto Encylopedia of Gene and Genomes [KEGG], PANTHER)
- Pathway selection can be determined by finding statistically overly frequent GO terms in a likely-selected gene list (e.g. DAVID)

GWAS

- Genome Wide Association Study
- Capitalizes on Linkage Disequilibrium
- Dense of panel of known SNPs equally spaced across genome
- Genes under selection will be in disequilibrium with SNPs that are physically close


-log10(P)
0
5
10
15
chromosome
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22

GWAS

- Experiment is divided into (large) sets of cases and controls
- F_{ST} and statistical association between allele frequencies and trait calculated for each SNP
- Very large number of tests requires stringent correction for multiple testing (e.g. Bonferroni correction)
- Individuals need to be corrected for background kinship

GWAS

- Only gives REGION of selection
- Region needs to be investigated using other methods (gene annotations, functional experiments)
- Many GWAS experiments find only ambiguous linkages to certain traits, especially traits with large numbers of additive small-effects