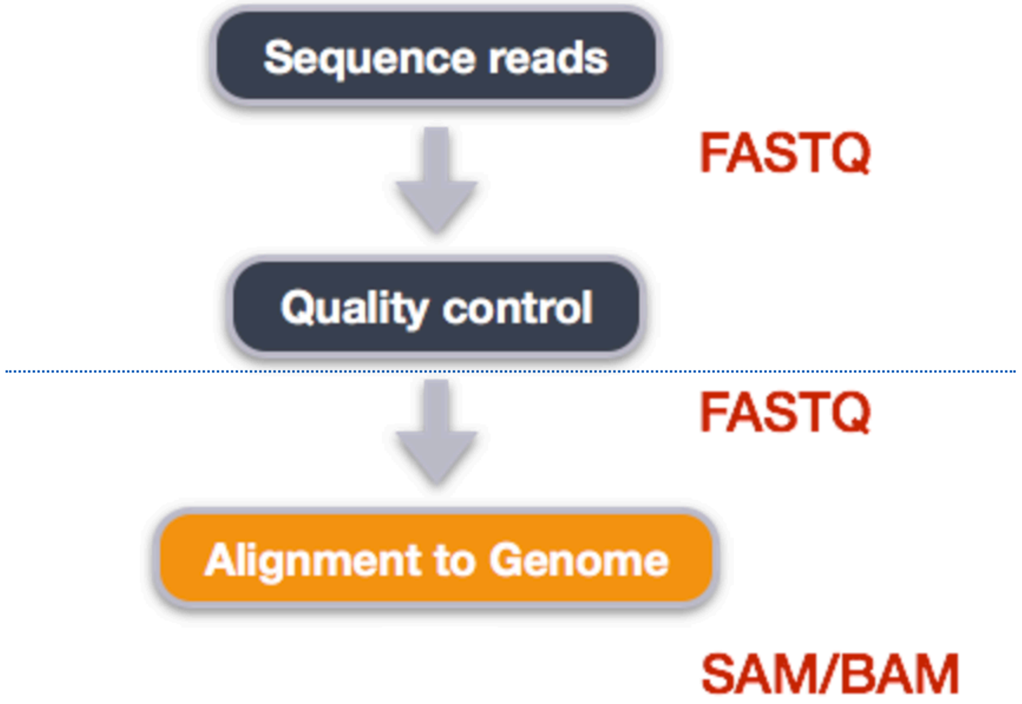
Today

- SAM, BAM, vcf files
- SNP calling
- natural selection from SNPs

FULL Tutorial

- https://training.galaxyproject.org/training-material/topics/data-science/tutorials/bash-variant-calling/tutorial.html

Alignment to a reference genome



```
$ conda create -n name_of_your_env bwa samtools bcftools
$ conda activate name_of_your_env
```

Software	Version	Manual	Available for	Description
BWA	0.7.17	BWA Manual	Linux, MacOS	Mapping DNA sequences against reference genome.
SAMtools	1.15.1	SAMtools Manual	Linux, MacOS	Utilities for manipulating alignments in the SAM format.
BCFtools	1.15.1	BCFtools manual	Linux, MacOS	Utilities for variant calling and manipulating VCFs and BCFs.
IGV	IGV Download	IGV User Guide	Linux, MacOS, Windows	Visualization and interactive exploration of large genomics datasets.

SAM/BAM format

The SAM file, is a tab-delimited text file that contains information for each individual read and its alignment to the genome. While we do not have time to go into detail about the features of the SAM format, the paper by Heng Li et al. provides a lot more detail on the specification.

- SAM = Sequence Alignment Map
- BAM = Binary Alignment Map

SAM/BAM format

The file begins with a **header**, which is optional.

The header is used to describe the source of data, reference sequence, method of alignment, etc., this will change depending on the aligner being used.

Following the header is the **alignment section**. Each line that follows corresponds to alignment information for a single read.

Each alignment line has **11 mandatory fields** for essential mapping information and a variable number of other fields for aligner specific information.

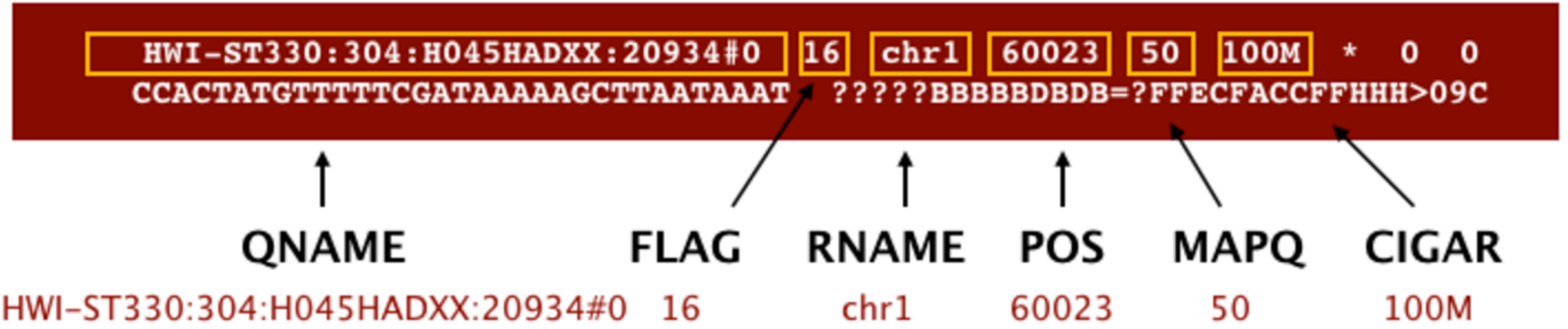
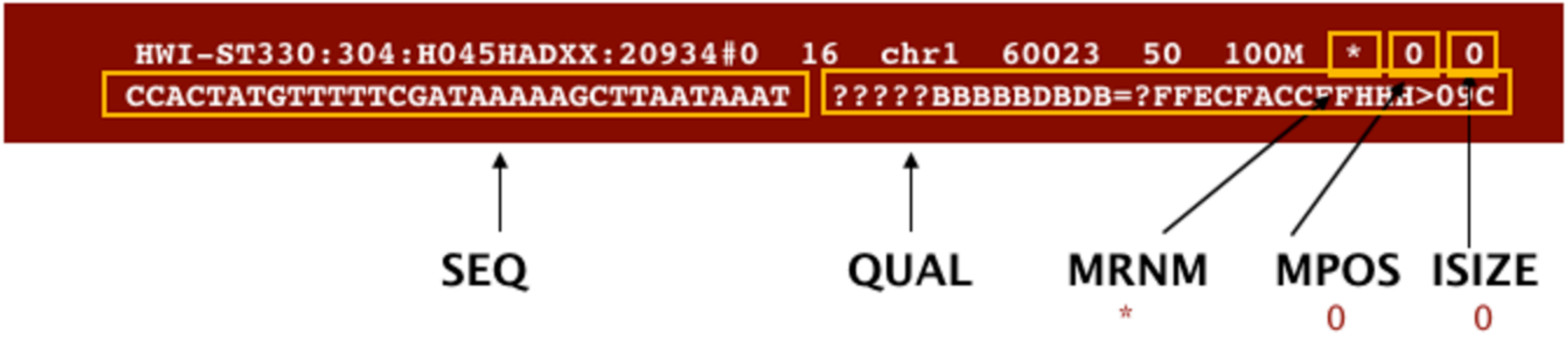


Image from Data Wrangling and Processing for Genomics



SAM

SAM/BAM files can be sorted in multiple ways, e.g. by location of alignment on the chromosome, by read name, etc.

different alignment tools will output differently sorted SAM/BAM, and different downstream tools require differently sorted alignment files as input.

samtools can be used to learn more about this bam file as well.

SAMTOOLS

Samtools is a set of utilities that manipulate alignments in the SAM (Sequence Alignment/Map), BAM, and CRAM formats.

convert between the formats, sorting, merging and indexing, and can retrieve reads in any regions swiftly.

Samtools is designed to work on a stream. It regards an input file `-' as the standard input (stdin) and an output file `-' as the standard output (stdout). Several commands can thus be combined with Unix pipes.

Samtools always output warning and error messages to the standard error output (stderr).

SAMTOOLS

Samtools is a set of utilities that manipulate alignments in the SAM (Sequence Alignment/Map), BAM, and CRAM formats.

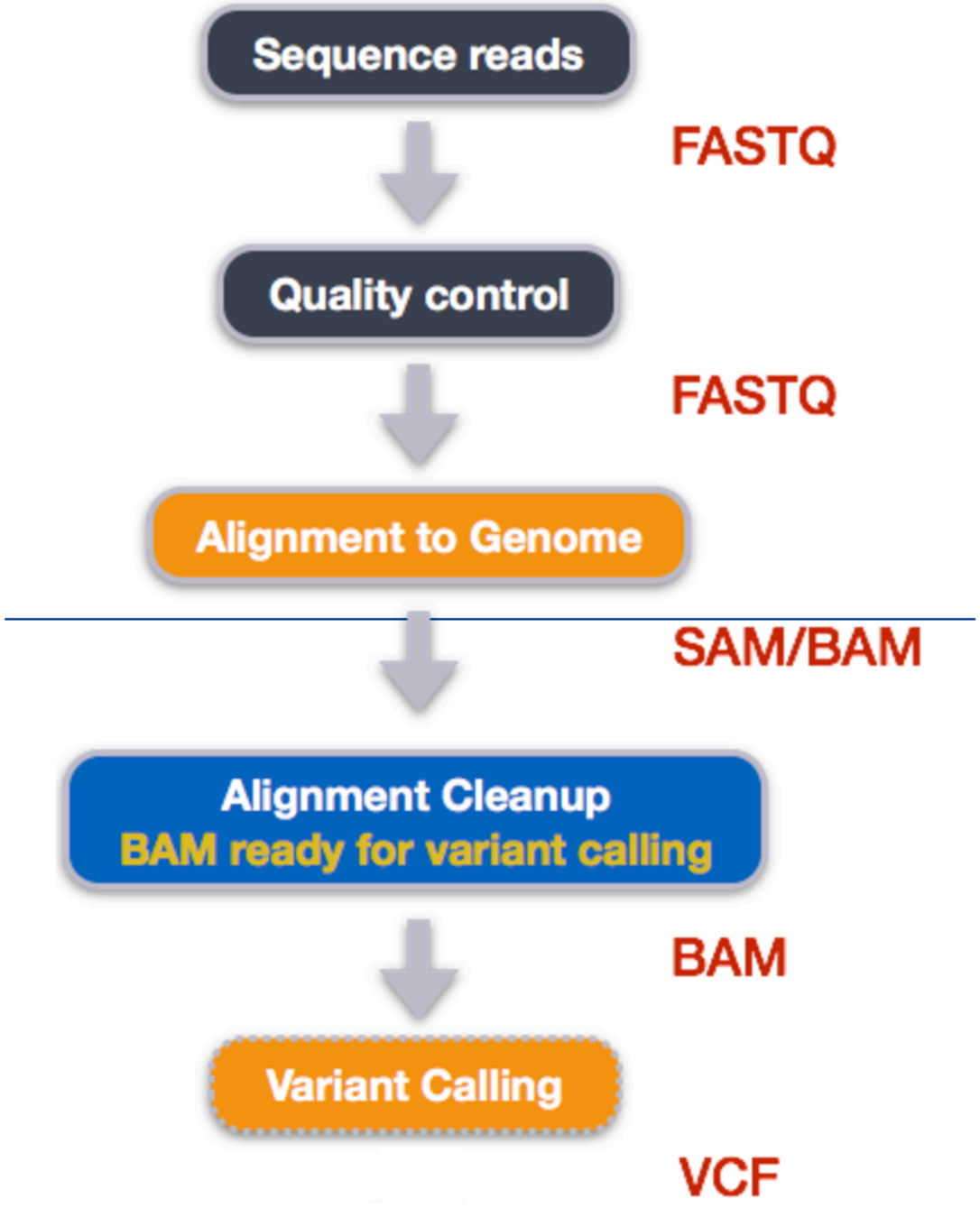
Samtools is also able to open files on remote FTP or HTTP(S) servers if the file name starts with `ftp://', `http://', etc.

Samtools checks the current working directory for the index file and will download the index upon absence.

Samtools does not retrieve the entire alignment file unless it is asked to do so.

BAM ?

The compressed binary version of SAM is called a BAM file. Use this file format to reduce size and to allow for *indexing*, which enables efficient random access of the data contained within the file.


Sequence reads
FASTQ
Quality control
FASTQ
Alignment to Genome
SAM/BAM
Alignment Cleanup
BAM ready for variant calling
BAM
Variant Calling
VCF

Variant calling

- https://samtools.github.io/bcftools/bcftools.html

- bcftools

STEPS

Step 1: Calculate the read coverage of positions in the genome

Step 2: Detect the single nucleotide variants (SNVs)

Step 3: Filter and report the SNV variants in variant calling format (VCF)

Output = .vcf

```
##fileformat=VCFv4.2
##FILTER<ID=PASS,Description="All filters passed">
##bcftoolsVersion=1.8+htslib-1.8
##bcftoolsCommand=mpileup -O b -o results/bcf/SRR2584866_raw.bcf -f data/ref_genome/ecoli_rel606.fast
##reference=file://data/ref_genome/ecoli_rel606.fasta
##contig<ID=CP000819.1,length=4629812>
##ALT<ID=*,Description="Represents allele(s) other than observed.">
##INFO<ID=INDEL,Number=0,Type=Flag,Description="Indicates that the variant is an INDEL.">
##INFO<ID=IDV,Number=1,Type=Integer,Description="Maximum number of reads supporting an indel">
##INFO<ID=IMF,Number=1,Type=Float,Description="Maximum fraction of reads supporting an indel">
##INFO<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
##INFO<ID=VDB,Number=1,Type=Float,Description="Variant Distance Bias for filtering splice-site artefa
##INFO<ID=RPB,Number=1,Type=Float,Description="Mann-Whitney U test of Read Position Bias (bigger is b
##INFO<ID=MQB,Number=1,Type=Float,Description="Mann-Whitney U test of Mapping Quality Bias (bigger is
##INFO<ID=BQB,Number=1,Type=Float,Description="Mann-Whitney U test of Base Quality Bias (bigger is be
##INFO<ID=MQSB,Number=1,Type=Float,Description="Mann-Whitney U test of Mapping Quality vs Strand Bias
##INFO<ID=SGB,Number=1,Type=Float,Description="Segregation based metric.">
##INFO<ID=MQ0F,Number=1,Type=Float,Description="Fraction of MQ0 reads (smaller is better)">
##FORMAT<ID=PL,Number=G,Type=Integer,Description="List of Phred-scaled genotype likelihoods">
##FORMAT<ID=GT,Number=1,Type=String,Description="Genotype">
##INFO<ID=ICB,Number=1,Type=Float,Description="Inbreeding Coefficient Binomial test (bigger is better
##INFO<ID=HOB,Number=1,Type=Float,Description="Bias in the number of HOMs number (smaller is better)"
##INFO<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes for each ALT allele, in the
##INFO<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO<ID=DP4,Number=4,Type=Integer,Description="Number of high-quality ref-forward , ref-reverse, al
##INFO<ID=MQ,Number=1,Type=Integer,Description="Average mapping quality">
##bcftools_callVersion=1.8+htslib-1.8
##bcftools_callCommand=call --ploidy 1 -m -v -o results/bcf/SRR2584866_variants.vcf results/bcf/SRR25
```

```
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  results/bam/SRR2584866.aligned.sorted.bam
CP000819.1      1521    .       C       T       207     .       DP=9;VDB=0.993024;SGB=-0.662043;MQSB=0.974597;MQ0F=0;AC=1;AN=1;DP4=0,0,4,5;MQ=60
CP000819.1      1612    .       A       G       225     .       DP=13;VDB=0.52194;SGB=-0.676189;MQSB=0.950952;MQ0F=0;AC=1;AN=1;DP4=0,0,6,5;MQ=60
CP000819.1      9092    .       A       G       225     .       DP=14;VDB=0.717543;SGB=-0.670168;MQSB=0.916482;MQ0F=0;AC=1;AN=1;DP4=0,0,7,3;MQ=60
CP000819.1      9972    .       T       G       214     .       DP=10;VDB=0.022095;SGB=-0.670168;MQSB=1;MQ0F=0;AC=1;AN=1;DP4=0,0,2,8;MQ=60      GT:PL
CP000819.1      10563   .       G       A       225     .       DP=11;VDB=0.958658;SGB=-0.670168;MQSB=0.952347;MQ0F=0;AC=1;AN=1;DP4=0,0,5,5;MQ=60
CP000819.1      22257   .       C       T       127     .       DP=5;VDB=0.0765947;SGB=-0.590765;MQSB=1;MQ0F=0;AC=1;AN=1;DP4=0,0,2,3;MQ=60      GT:PL
CP000819.1      38971   .       A       G       225     .       DP=14;VDB=0.872139;SGB=-0.680642;MQSB=1;MQ0F=0;AC=1;AN=1;DP4=0,0,4,8;MQ=60      GT:PL
CP000819.1      42306   .       A       G       225     .       DP=15;VDB=0.969686;SGB=-0.686358;MQSB=1;MQ0F=0;AC=1;AN=1;DP4=0,0,5,9;MQ=60      GT:PL
CP000819.1      45277   .       A       G       225     .       DP=15;VDB=0.470998;SGB=-0.680642;MQSB=0.95494;MQ0F=0;AC=1;AN=1;DP4=0,0,7,5;MQ=60
CP000819.1      56613   .       C       G       183     .       DP=12;VDB=0.879703;SGB=-0.676189;MQSB=1;MQ0F=0;AC=1;AN=1;DP4=0,0,8,3;MQ=60      GT:PL
CP000819.1      62118   .       A       G       225     .       DP=19;VDB=0.414981;SGB=-0.691153;MQSB=0.906029;MQ0F=0;AC=1;AN=1;DP4=0,0,8,10;MQ=59
CP000819.1      64042   .       G       A       225     .       DP=18;VDB=0.451328;SGB=-0.689466;MQSB=1;MQ0F=0;AC=1;AN=1;DP4=0,0,7,9;MQ=60      GT:PL
•
```

```
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  results/bam/SRR2584866.aligned.sorted.bam
```

This is a lot of information, in this output.

The first few columns represent the information we have about a predicted variation.

column	**info**
CHROM.	contig location where the variation occurs
POS	position within the contig where the variation occurs
ID.	a . until we add annotation information
REF.	reference genotype (forward strand)
ALT	sample genotype (forward strand)
QUAL.	Phred-scaled probability that the observed variant exists at this site (higher is better)
FILTER	if no quality filters have been applied, PASS if a filter is passed, or the name of the filters this variant failed

The last two columns contain the genotypes and can be tricky to decode.

•

.vcf

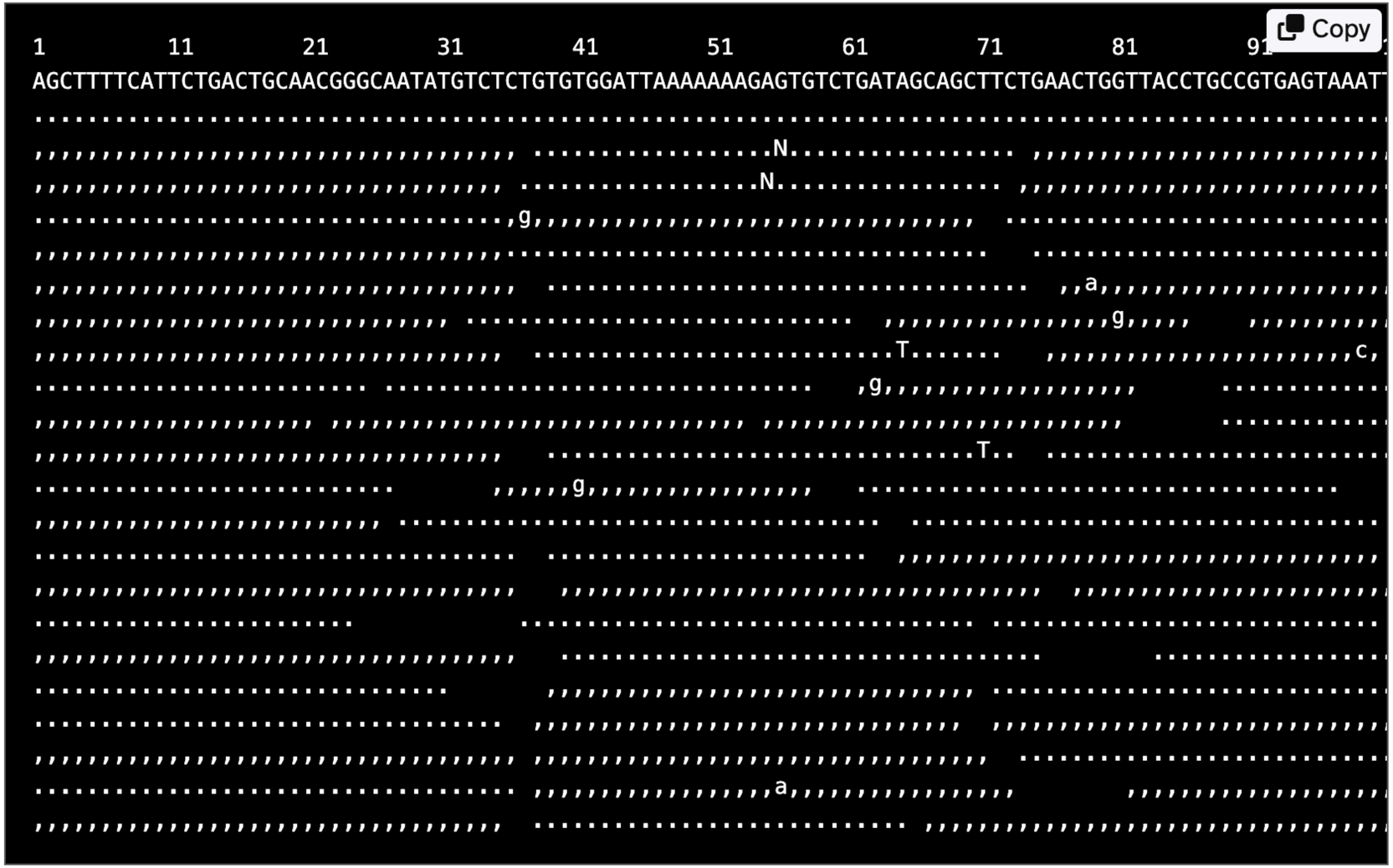
column
info
FORMAT: lists in order the metrics presented in the final column
results: lists the values associated with those metrics in order

metric	**definition**
AD, DP.	the depth per allele by sample and coverage
GT.	the genotype for the sample at this loci.
PL.	the likelihoods of the given genotypes
GQ.	the Phred-scaled confidence for the genotype

RC	POS	fere	a_ge	ea_	Pinky_	Pink	N_gene_na	llel	Annotation	notation_Im	Gene_Name	Gene_ID	eature_Typ	Feature_ID	script_Bio	Rank	HGVS.c	HGVS.p	pos_cDNA.	pos_CDS.le	pos_AA.ler	Distance	_WARNINGS_INFO
1	956227	C	./.	.	C/G	4	AGRN	G	intron_var	MODIFIER	AGRN	ENSG0000	transcript	ENST0000	protein_co	1/35	c.201+474C>G						
1	956374	A	T/T	34	./.	.	AGRN	T	intron_var	MODIFIER	AGRN	ENSG0000	transcript	ENST0000	protein_co	1/35	c.201+621A>T						
1	956622	C	T/T	47	./.	.	AGRN	T	intron_var	MODIFIER	AGRN	ENSG0000	transcript	ENST0000	protein_co	1/35	c.201+869C>T						
1	957105	A	G/G	32	./.	.	AGRN	G	intron_var	MODIFIER	AGRN	ENSG0000	transcript	ENST0000	protein_co	1/35	c.202-476A>G						
1	957568	A	G/G	51	G/G	22	AGRN	G	intron_var	MODIFIER	AGRN	ENSG0000	transcript	ENST0000	protein_co	1/35	c.202-13A>G						
1	957640	C	T/T	47	T/T	24	AGRN	T	synonymou	LOW	AGRN	ENSG0000	transcript	ENST0000	protein_co	2/36	c.261C>T	p.Asp87As	311/7323	261/6138	87/2045		
1	957820	G	A/A	44	./.	.	AGRN	A	synonymou	LOW	AGRN	ENSG0000	transcript	ENST0000	protein_co	2/36	c.441G>A	p.Glu147G	491/7323	441/6138	147/2045		
1	957898	G	T/T	35	T/T	18	AGRN	T	intron_var	MODIFIER	AGRN	ENSG0000	transcript	ENST0000	protein_co	2/35	c.463+56G>T						
1	957933	T	C/C	28	./.	.	AGRN	C	intron_var	MODIFIER	AGRN	ENSG0000	transcript	ENST0000	protein_co	2/35	c.463+91T>C						
1	957965	T	T/G	11	T/A	8	AGRN	A	intron_var	MODIFIER	AGRN	ENSG0000	transcript	ENST0000	protein_co	2/35	c.463+123T>A						
1	957965							G	intron_var	MODIFIER	AGRN	ENSG0000	transcript	ENST0000	protein_co	2/35	c.463+123T>G						
1	957967	T	T/C	9	./.	.	AGRN	C	intron_var	MODIFIER	AGRN	ENSG0000	transcript	ENST0000	protein_co	2/35	c.463+125T>C						
1	957971	G	G/A	13	G/A	8	AGRN	A	intron_var	MODIFIER	AGRN	ENSG0000	transcript	ENST0000	protein_co	2/35	c.463+129G>A						
1	957975	C	./.	.	C/T	6	AGRN	T	intron_var	MODIFIER	AGRN	ENSG0000	transcript	ENST0000	protein_co	2/35	c.463+133C>T						
1	957976	A	./.	.	A/G	7	AGRN	G	intron_var	MODIFIER	AGRN	ENSG0000	transcript	ENST0000	protein_co	2/35	c.463+134A>G						
1	958017	A	G/G	36	G/G	17	AGRN	G	intron_var	MODIFIER	AGRN	ENSG0000	transcript	ENST0000	protein_co	2/35	c.463+175A>G						
1	958216	C	./.	.	C/T	24	AGRN	T	intron_var	MODIFIER	AGRN	ENSG0000	transcript	ENST0000	protein_co	2/35	c.463+374C>T						
1	958248	A	G/G	37	G/G	13	AGRN	G	intron_var	MODIFIER	AGRN	ENSG0000	transcript	ENST0000	protein_co	2/35	c.463+406A>G						
1	958427	G	C/C	29	./.	.	AGRN	C	intron_var	MODIFIER	AGRN	ENSG0000	transcript	ENST0000	protein_co	2/35	c.463+585G>C						
1	958624	G	./.	.	C/C	24	AGRN	C	intron_var	MODIFIER	AGRN	ENSG0000	transcript	ENST0000	protein_co	2/35	c.463+782G>C						
1	958731	A	./.	.	G/G	22	AGRN	G	intron_var	MODIFIER	AGRN	ENSG0000	transcript	ENST0000	protein_co	2/35	c.463+889A>G						
1	958905	A	./.	.	G/G	24	AGRN	G	intron_var	MODIFIER	AGRN	ENSG0000	transcript	ENST0000	protein_co	2/35	c.463+1063A>G						
1	958953	A	./.	.	G/G	22	AGRN	G	intron_var	MODIFIER	AGRN	ENSG0000	transcript	ENST0000	protein_co	2/35	c.463+1111A>G						
1	959026	G	./.	.	A/A	18	AGRN	A	intron_var	MODIFIER	AGRN	ENSG0000	transcript	ENST0000	protein_co	2/35	c.463+1184G>A						
1	959169	G	G/C	23	./.	.	AGRN	C	intron_var	MODIFIER	AGRN	ENSG0000	transcript	ENST0000	protein_co	2/35	c.463+1327G>C						
1	959842	C	./.	.	T/T	26	AGRN	T	intron_var	MODIFIER	AGRN	ENSG0000	transcript	ENST0000	protein_co	2/35	c.463+2000C>T						
1	961294	G	./.	.	A/A	17	AGRN	A	intron_var	MODIFIER	AGRN	ENSG0000	transcript	ENST0000	protein_co	2/35	c.463+3452G>A						
1	961370	G	T/T	49	./.	.	AGRN	T	intron_var	MODIFIER	AGRN	ENSG0000	transcript	ENST0000	protein_co	2/35	c.463+3528G>T						
1	961512	C	./.	.	T/T	24	AGRN	T	intron_var	MODIFIER	AGRN	ENSG0000	transcript	ENST0000	protein_co	2/35	c.463+3670C>T						
1	961827	G	A/A	35	A/A	13	AGRN	A	intron_var	MODIFIER	AGRN	ENSG0000	transcript	ENST0000	protein_co	2/35	c.463+3985G>A						
1	962892	C	./.	.	T/T	16	AGRN	T	intron_var	MODIFIER	AGRN	ENSG0000	transcript	ENST0000	protein_co	2/35	c.463+5050C>T						
1	963706	C	./.	.	A/A	9	AGRN	A	intron_var	MODIFIER	AGRN	ENSG0000	transcript	ENST0000	protein_co	2/35	c.463+5864C>A						
1	963712	C	./.	.	G/G	4	AGRN	G	intron_var	MODIFIER	AGRN	ENSG0000	transcript	ENST0000	protein_co	2/35	c.463+5870C>G						
1	963721	T	C/C	6	T/C	3	AGRN	C	intron_var	MODIFIER	AGRN	ENSG0000	transcript	ENST0000	protein_co	2/35	c.463+5879T>C						
1	963723	C	./.	.	C/G	3	AGRN	G	intron_var	MODIFIER	AGRN	ENSG0000	transcript	ENST0000	protein_co	2/35	c.463+5881C>G						
1	963871	C	./.	.	G/G	9	AGRN	G	intron_var	MODIFIER	AGRN	ENSG0000	transcript	ENST0000	protein_co	2/35	c.463+6029C>G						
1	964046	C	C/T	89	./.	.	AGRN	T	intron_var	MODIFIER	AGRN	ENSG0000	transcript	ENST0000	protein_co	2/35	c.463+6204C>T						
1	964079	C	C/T	59	./.	.	AGRN	T	intron_var	MODIFIER	AGRN	ENSG0000	transcript	ENST0000	protein_co	2/35	c.463+6237C>T						
1	964389	C	T/T	55	./.	.	AGRN	T	intron_var	MODIFIER	AGRN	ENSG0000	transcript	ENST0000	protein_co	2/35	c.464-6268C>T						
1	964437	C	./.	.	C/G	2	AGRN	G	intron_var	MODIFIER	AGRN	ENSG0000	transcript	ENST0000	protein_co	2/35	c.464-6220C>G						
1	964473	C	T/T	38	./.	.	AGRN	T	intron_var	MODIFIER	AGRN	ENSG0000	transcript	ENST0000	protein_co	2/35	c.464-6184C>T						
1	964485	C	./.	.	C/T	5	AGRN	T	intron_var	MODIFIER	AGRN	ENSG0000	transcript	ENST0000	protein_co	2/35	c.464-6172C>T						
1	964494	T	C/C	32	./.	.	AGRN	C	intron_var	MODIFIER	AGRN	ENSG0000	transcript	ENST0000	protein_co	2/35	c.464-6163T>C						
1	964498	G	C/C	32	./.	.	AGRN	C	intron_var	MODIFIER	AGRN	ENSG0000	transcript	ENST0000	protein_co	2/35	c.464-6159G>C						
1	964506	C	T/T	36	./.	.	AGRN	T	intron_var	MODIFIER	AGRN	ENSG0000	transcript	ENST0000	protein_co	2/35	c.464-6151C>T						
1	964582	C	A/A	44	./.	.	AGRN	A	intron_var	MODIFIER	AGRN	ENSG0000	transcript	ENST0000	protein_co	2/35	c.464-6075C>A						
1	964766	G	G/C	25	./.	.	AGRN	C	intron_var	MODIFIER	AGRN	ENSG0000	transcript	ENST0000	protein_co	2/35	c.464-5891G>C						
1	964797	C	./.	.	C/T	13	AGRN	T	intron_var	MODIFIER	AGRN	ENSG0000	transcript	ENST0000	protein_co	2/35	c.464-5860C>T						

Viewing with `tview`

SAMTOOLS



Viewing with IGV

IGV is a stand-alone browser, which has the advantage of being installed locally and providing fast access.

Web-based genome browsers, like Ensembl or the UCSC browser, are slower, but provide more functionality.

Viewing with IGV

They not only allow for more polished and flexible visualization, but also provide easy access to a wealth of annotations and external data sources.

This makes it straightforward to relate your data with information about repeat regions, known genes, epigenetic features or areas of cross-species conservation, to name just a few.

Viewing with IGV

There should be two tracks: one coresponding to our BAM file and the other for our VCF file.

In the VCF track, each bar across the top of the plot shows the allele fraction for a single locus. The second bar shows the genotypes for each locus in each sample. We only have one sample called here, so we only see a single line. Dark blue = heterozygous, Cyan = homozygous variant, Grey = reference. Filtered entries are transparent.

Zoom in to inspect variants you see in your filtered VCF file to become more familiar with IGV. See how quality information corresponds to alignment information at those loci. Use this website and the links therein to understand how IGV colors the alignments.

- GALAXY

https://usegalaxy.org/

ATTGTGTA**T**ATTGTATGTA
ATTGTGTA**G**ATTGTATGTA

SNP

Sidebar 18.1 Parameters incorporated into modern population genetics models: After Marjoram and Tavare

Mutation and recombination rates: among the first parameters to be approached by these new techniques.

Demographic parameters: population size, population substructure and mating patterns, and migration.

Selection: regions of the genome under selective pressure. In particular, the HapMap project has benefited from this approach in identifying episodic selection at the genome level. In addition, a phenomenon called selective sweep can be examined via this process.

Ancestral inference: time to the most recent common ancestor (TMRCA). TMRCA identifies the divergence time of a population and is also used to infer the age of certain mutations.

Genome-level phenomena: genomic changes within a population, such as identification of recombination hotspots, reconstruction of haplotypes, and linkage disequilibrium.

Human disease association studies: identification and mapping of disease genes mostly used in human population genetics. Genome-wide association studies (GWAS) are used extensively in disease association studies and are reliant on population genetic modeling.

General population description: estimates of genetic diversity, genetic distances between populations, estimates of population subdivision, tests of selective neutrality within populations, estimates of gametic phase, allelic richness, and kinship and relatedness at the individual level. Many of these parameters and estimates can be obtained from statistical packages and programs that we will discuss in the next few chapters. These kinds of analyses have become an important part of conservation genetic research.

Individual-centered analysis: estimates of recent migration rates, assignment of individuals to populations, detection of genetic structure among individuals in populations, and hybridization in populations, as well as detection of hybrids in populations.

OLD TOOLS NEW CONTEXT

Tajima's D distinguishes between sequences evolving neutrally and those evolving non-neutrally using allele frequencies

F statistics measure the degree of isolation of entities

OLD TOOLS NEW CONTEXT

FIT which is the correlation between gametes within an individual relative to the entire population; the FIS, which is the correlation between gametes within an individual relative to the subpopulation to which that individual belongs; and the FST, which is the correlation between gametes chosen randomly from within the same subpopulation relative to the entire population.

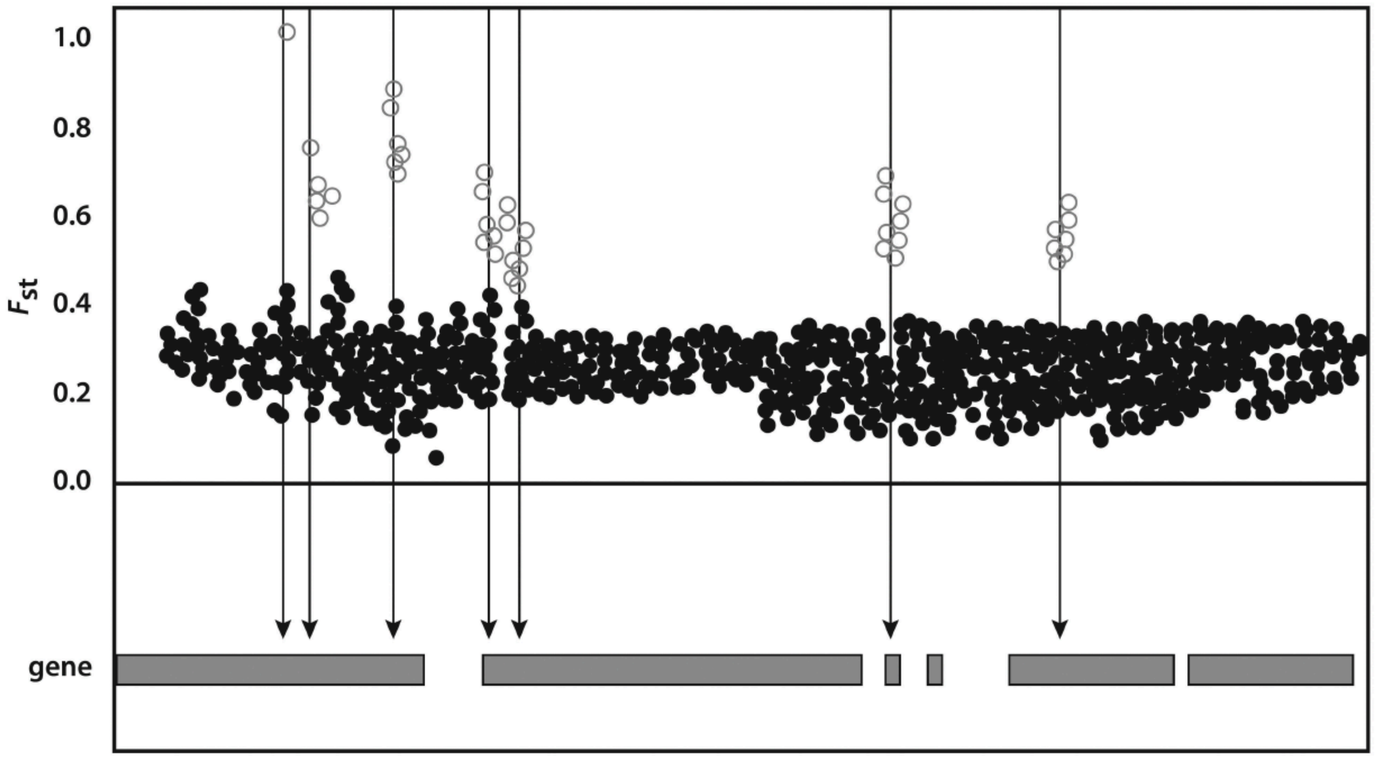
FST and related measures have four major uses in evolutionary biology

Estimating migration rates.

Inferring demographic history.

Identifying genomic regions under selection.

Forensic science and association mapping.



of FST values over a genomic region with more than six gene regions and over a large number of single nucleotide polymorphisms. In this hypothetical example, the majority of single nucleotide polymorphisms (SNPs) were inferred to be not under selection (closed circles), while approximately 50 SNPs (open circles) showed signs of selection as inferred from the F_{ST} estimates calculated with Bayesian approaches. Seven of the SNP regions showed extremely large F_{ST} and hence strong degrees of statistically significant selection. The solid rectangles at the bottom of the figure refer to gene regions on a chromosome. (Adapted from K.E Holsinger & B.S. Weir. *Nature Reviews Genetics* 10:639–650, 2009. Courtesy of Nature Publishing Group.)

STRUCTURE analysis reveals substructure and genetic cross talk

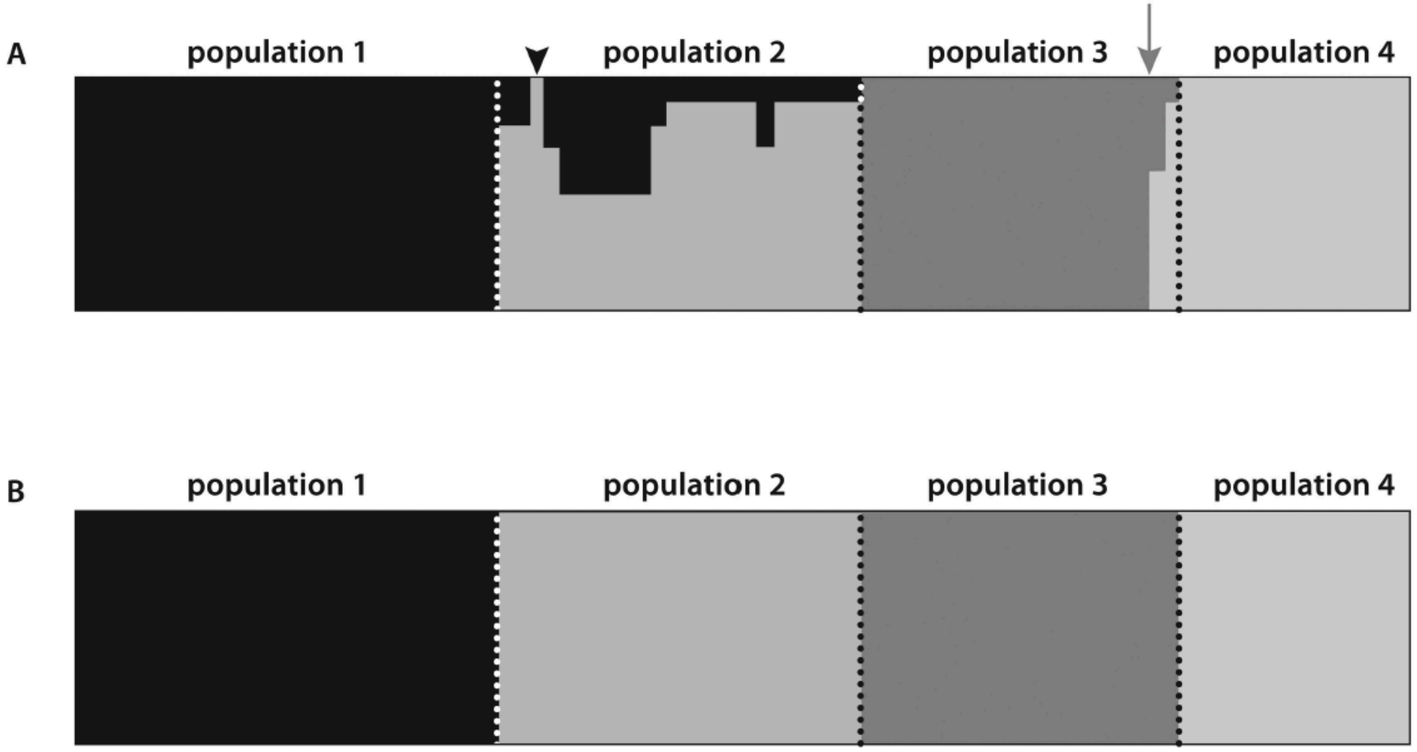
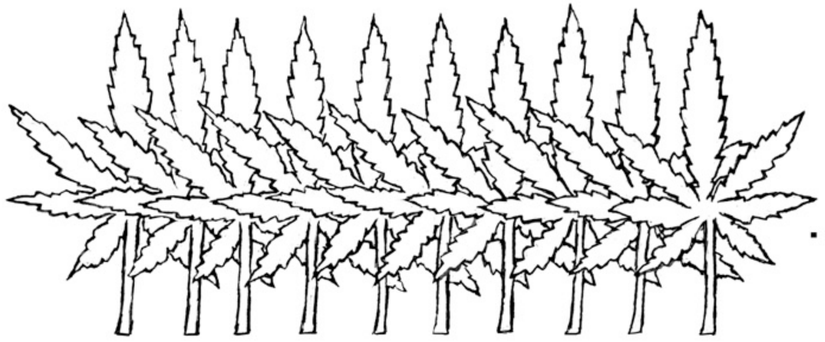


Figure 18.7 STRUCTURE diagrams. A: STRUCTURE analysis where the assignments to populations are "imperfect." The different shades (black, gray, dark green, and light green) represent assignment to a particular "population." Dotted lines indicate boundaries of the populations. Each column represents an individual in the study. The green arrow points to an individual that was assigned to at least two populations, in this case to populations 3 and 4. The black arrow points to an individual that was assigned to a single population (gray), in this case to population 2. B: STRUCTURE analysis where the assignments to the four populations are "one-to-one." Dotted lines indicate boundaries of the populations.

CASES



SNP1 = AT/TT *SNP2* = GC/CC *SNP....*

CASES
Count of T/T
100 of 400

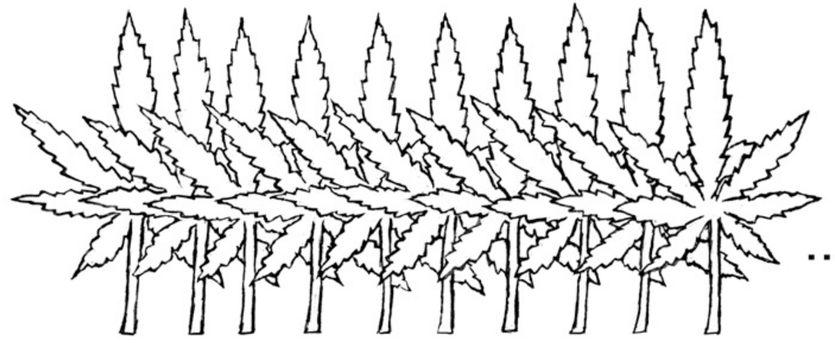
Frequency of T:
25%

CASES
Count of G/T
210 of 400

Frequency of G:
52%

Repeat for all SNPs

CONTROLS



CONTROLS
Count of T/T
300 of 600

Frequency of T:
50%

CONTROLS
Count of G/C
300 of 600

Frequency of T:
50%

P-value:
1.537459e-12

P-value:
0.4795

Case-Control GWAS

cases

*T-**A**-C-T-G-T*

*T-**A**-T-A-G-A*

*G-**A**-T-A-G-T*

*T-**A**-C-A-T-A*

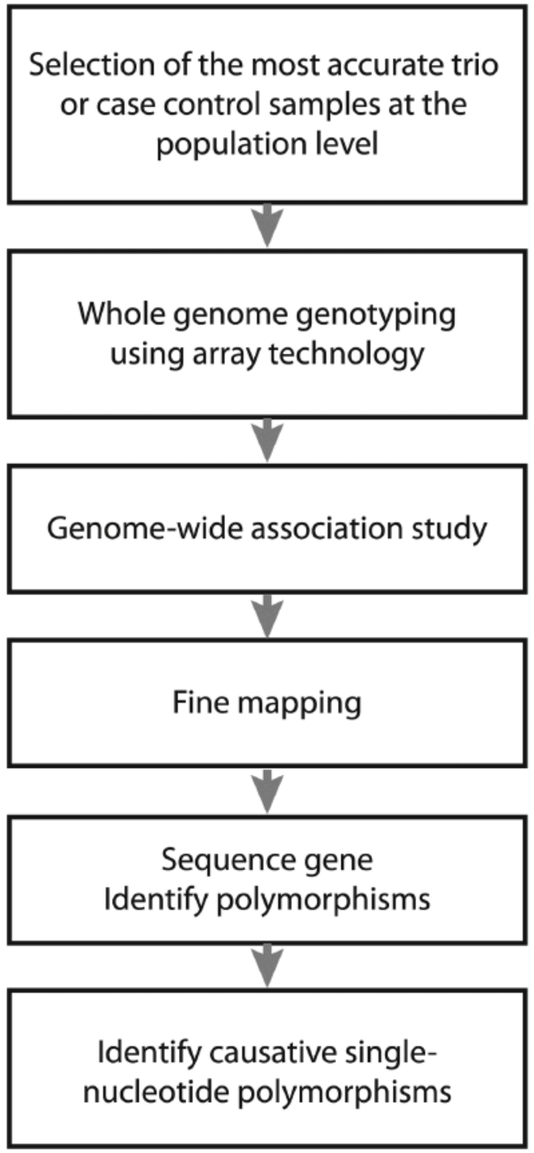
controls

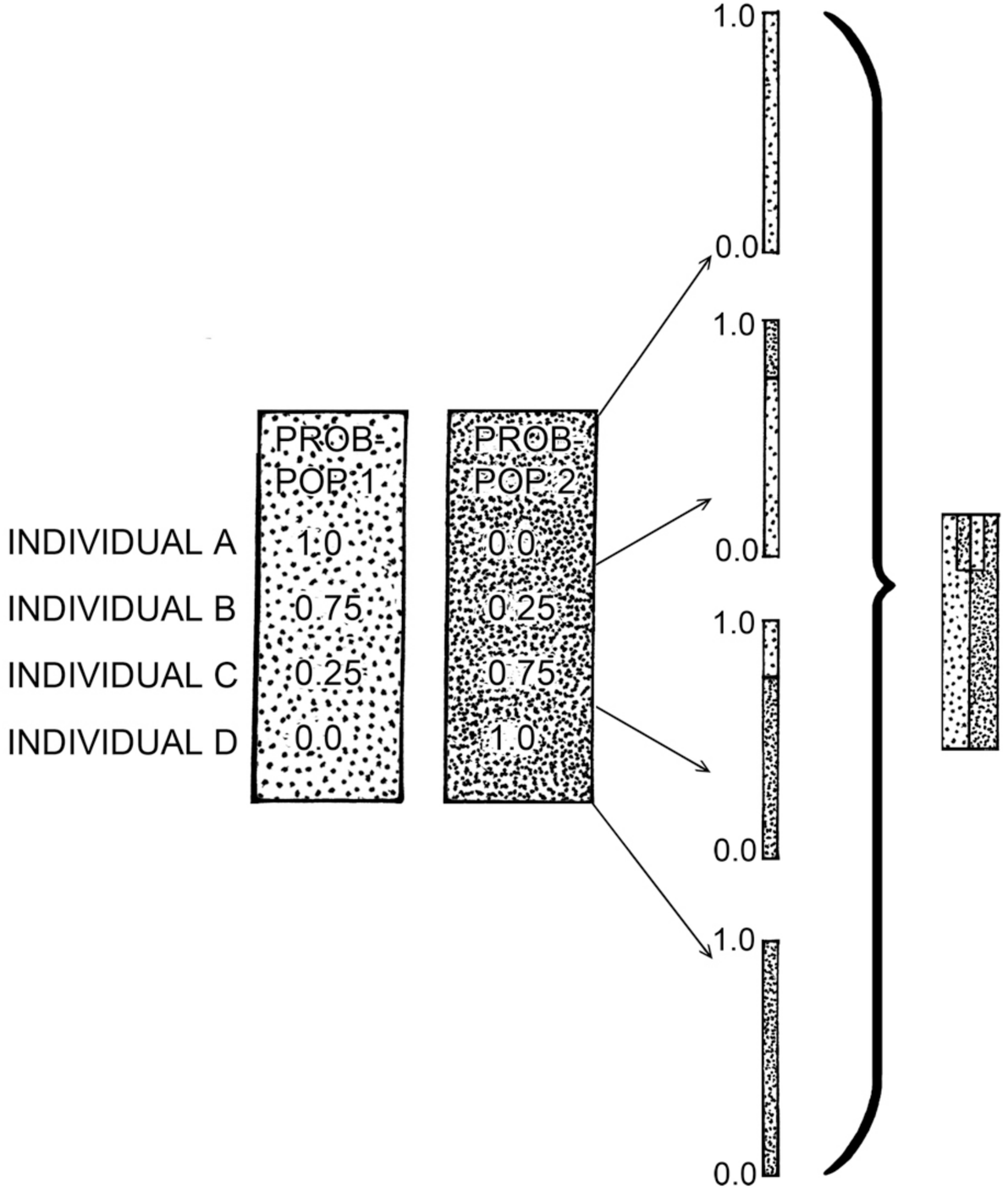
*T-**T**-C-T-T-A*

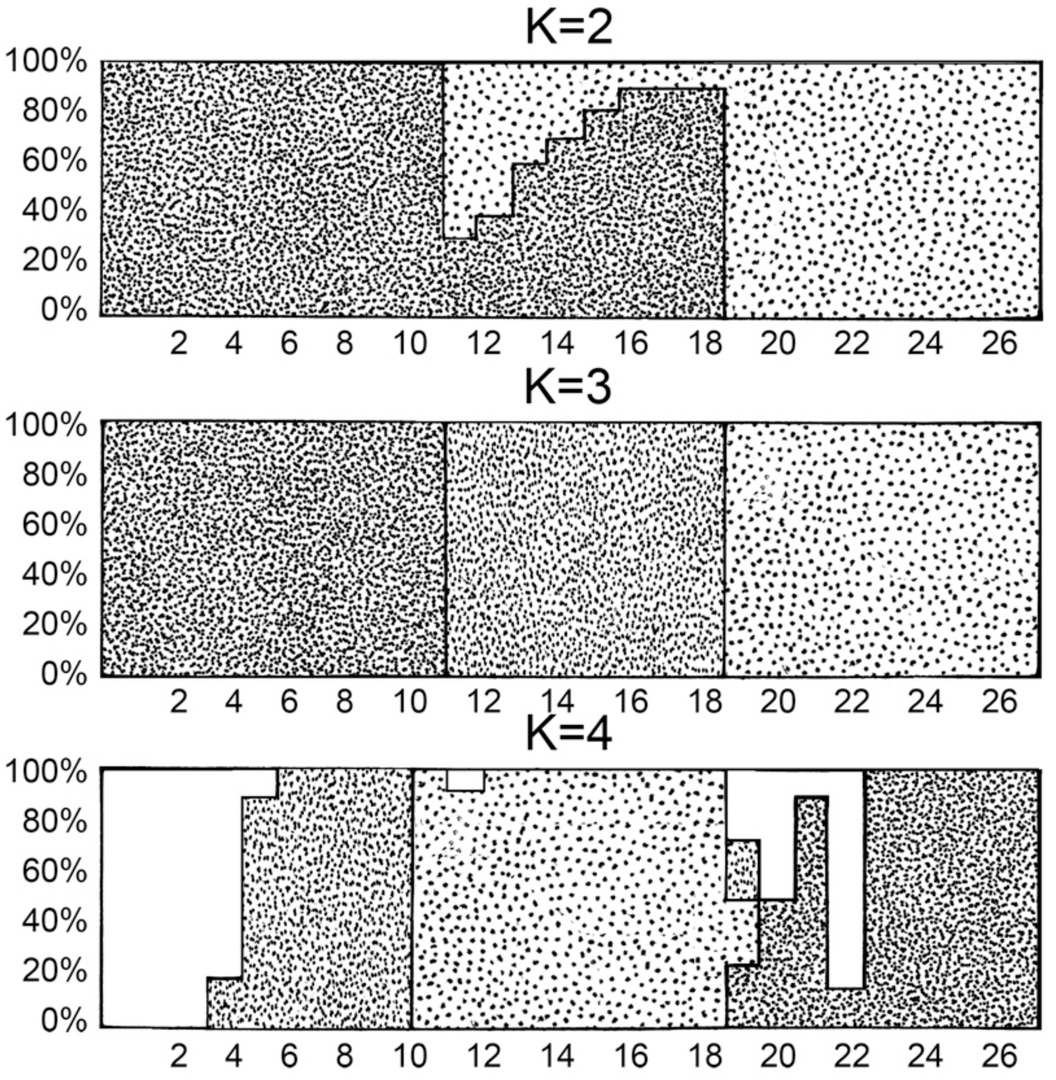
*T-**T**-T-A-G-T*

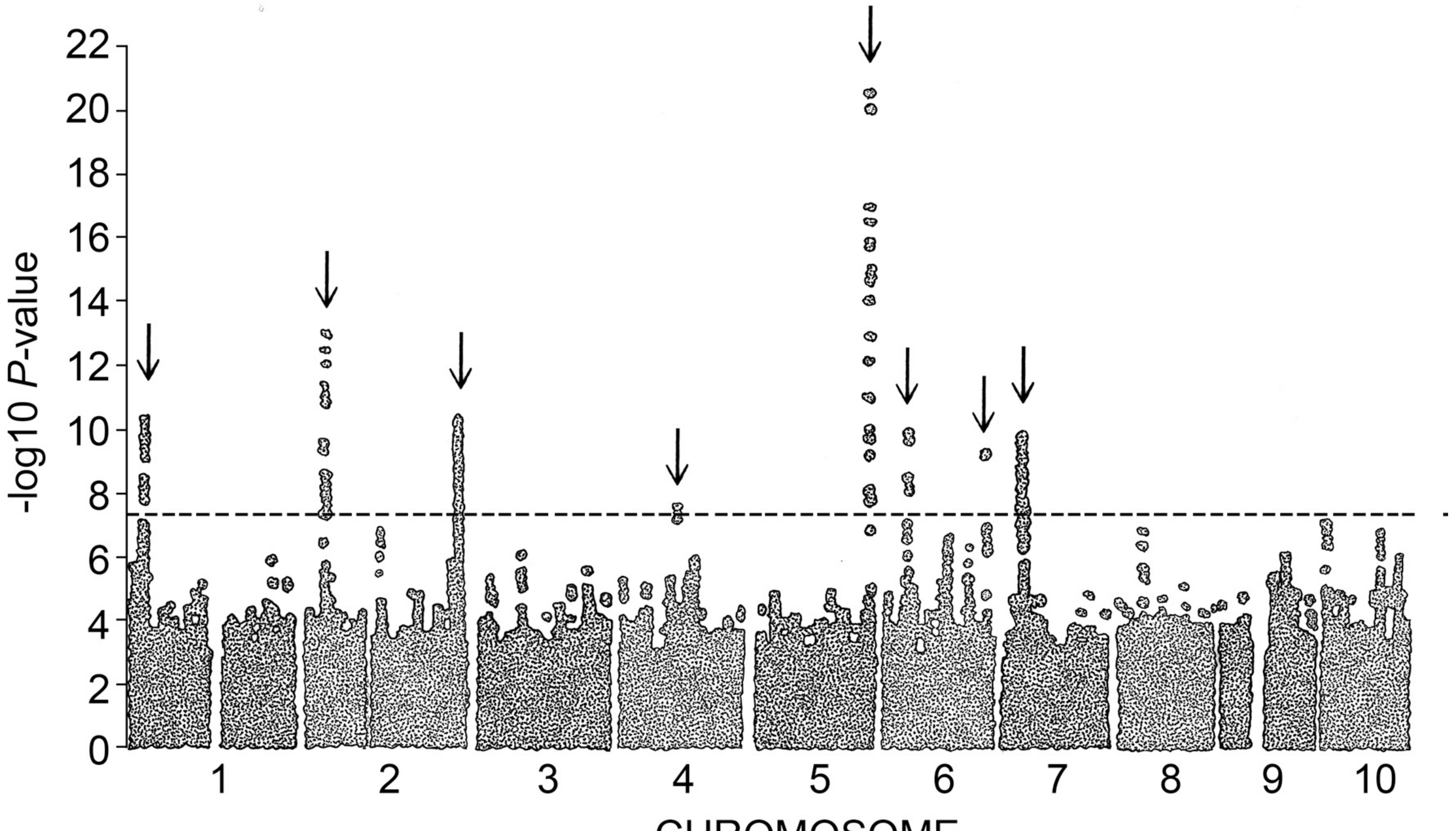
*G-**T**-C-T-G-T*

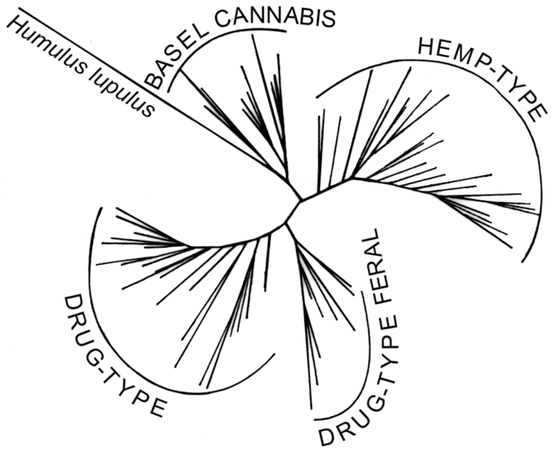
*G-**T**-T-A-G-A*

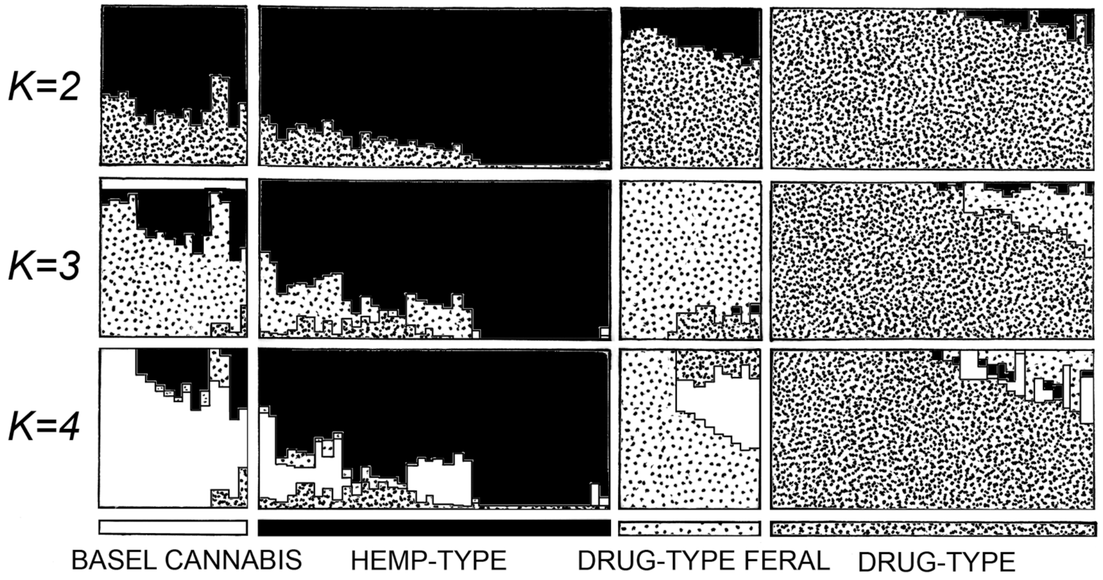


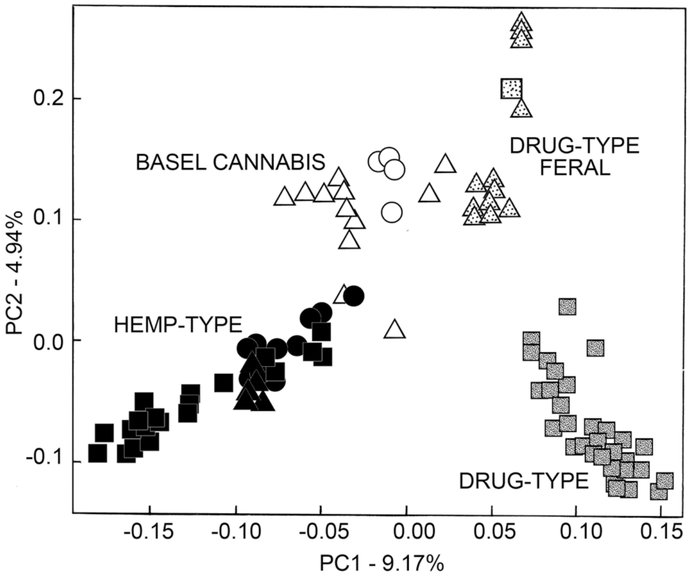

PROB-
POP 1
PROB-
POP 2
INDIVIDUAL A
1 0
0 0
INDIVIDUAL B
0 75
0 25
INDIVIDUAL C
0 25
0 75
INDIVIDUAL D
0 0
1 0
1.0
0.0
1.0
0.0
1.0
0.0
1.0
0.0


K=2
K=3
K=4
100%
80%
60%
40%
20%
0%
2 4 6 8 10 12 14 16 18 20 22 24 26


-log10 P-value
0
2
4
6
8
10
12
14
16
18
20
22
1
2
3
4
5
6
7
8
9
10
CHROMOSOME


Humulus lupulus
BASEL CANNABIS
HEMP-TYPE
DRUG-TYPE
DRUG-TYPE FERAL


K=2
K=3
K=4
BASEL CANNABIS
HEMP-TYPE
DRUG-TYPE FERAL
DRUG-TYPE


BASEL CANNABIS
DRUG-TYPE FERAL
HEMP-TYPE
DRUG-TYPE
PC2 - 4.94%
PC1 - 9.17%
0.2
0.1
0.0
-0.1
-0.15
-0.10
-0.05
0.00
0.05
0.10
0.15

Genetic Hitchhiking and Selective Sweeps

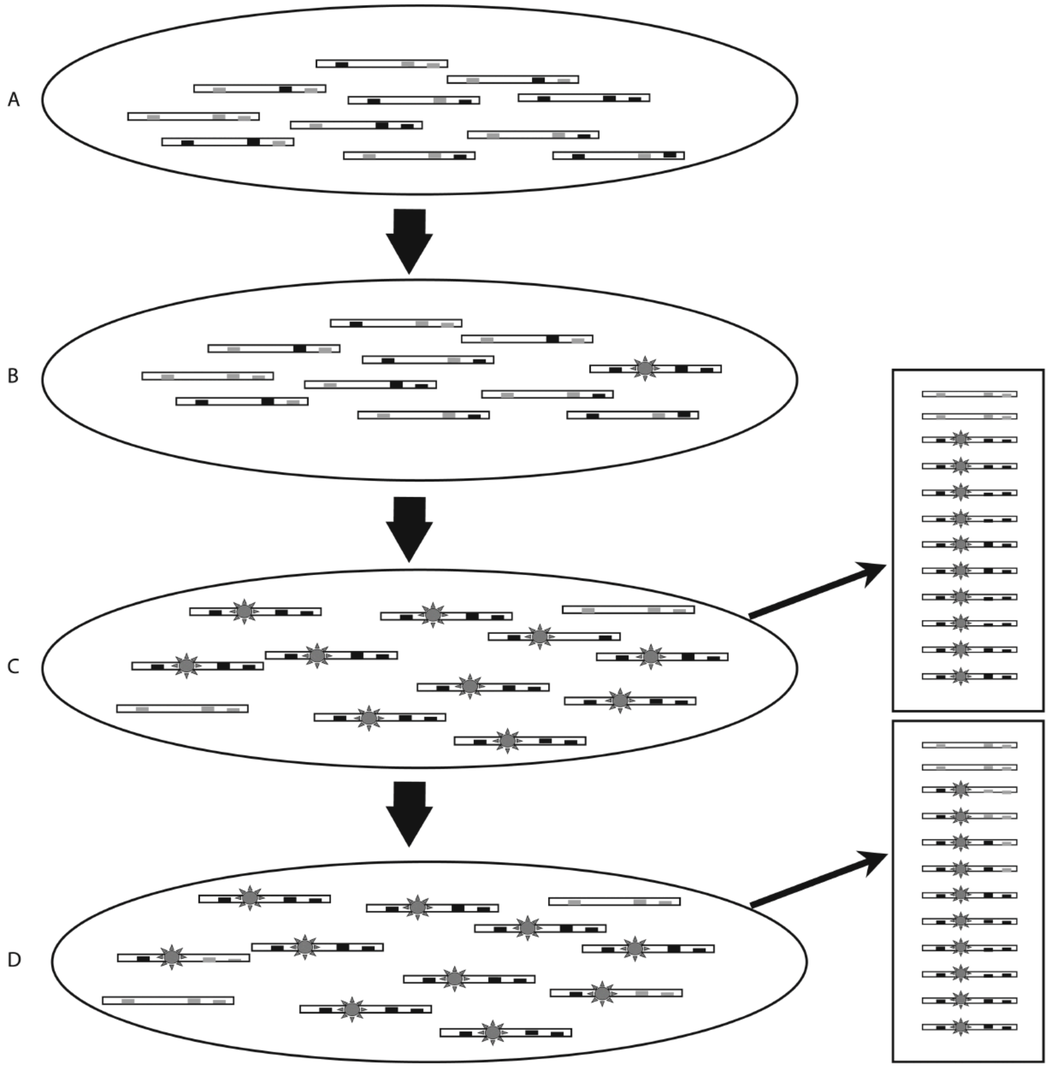


Figure 19.2 Processes involved in a selective sweep. In the four breeding populations shown, different single-nucleotide polymorphisms are represented by gray and black boxes on the chromosomes. The arrows between the populations indicate breeding between generations. A: the initial breeding population has a certain amount of genetic variation on a chromosome. B: a beneficial mutation arises in the population (single starburst) and, due to strong selection, nearly all chromosomes in the population are selected against and eliminated. C: natural selection drives the increase of the favorable mutation (shown by the starbursts) and alleles linked to it. This stage of a selective sweep is most easily recognized via phylogenomics. D: as recombination occurs in subsequent generations of the population, the chromosomal markers start to lose their initial signal of a selective sweep for markers further away from the locus under selection. The chromosomes from panels C and D are summarized in the adjacent boxes and lined up to show how the clarity of the result of a selective sweep erodes with recombination.

Genetic Hitchhiking and Selective Sweeps

Selective sweeps are detected in four basic ways

through the extended haplotype approach, which uses linkage disequilibrium methods. In this method haplotypes are used to examine a population of individuals for extended linkage disequilibrium and homozygosity.

The second approach is called the site frequency spectrum (SFS) approach. This approach quantifies the excess or surfeit of genetic variation across regions of the genome. For instance, a recently derived allele in a large population, if neutral, should be found in low frequency,

recent strong positive selection will result in the subdivisioning of large populations. Classical measures of subdivisioning in populations can be used to examine various parts of the genome, and regions that show strong subdivisioning are assumed to be under a selective sweep.

selective sweeps are also expected to reduce variability in regions of the genome where they have been an influence. In this approach, the genome is then scanned for degree of variability, and any region that appears to depart from background levels of variability is assumed to be under a selective sweep.

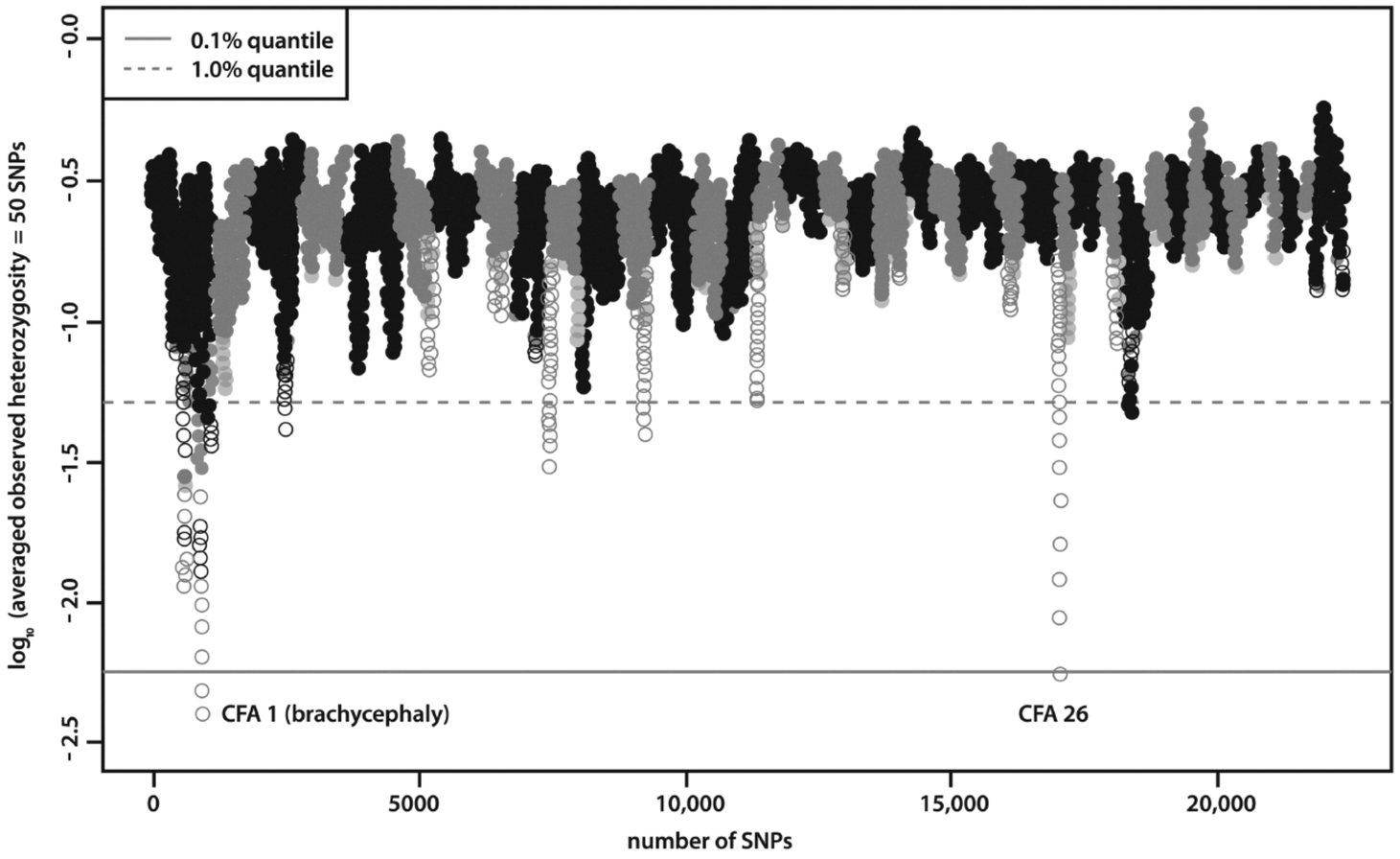
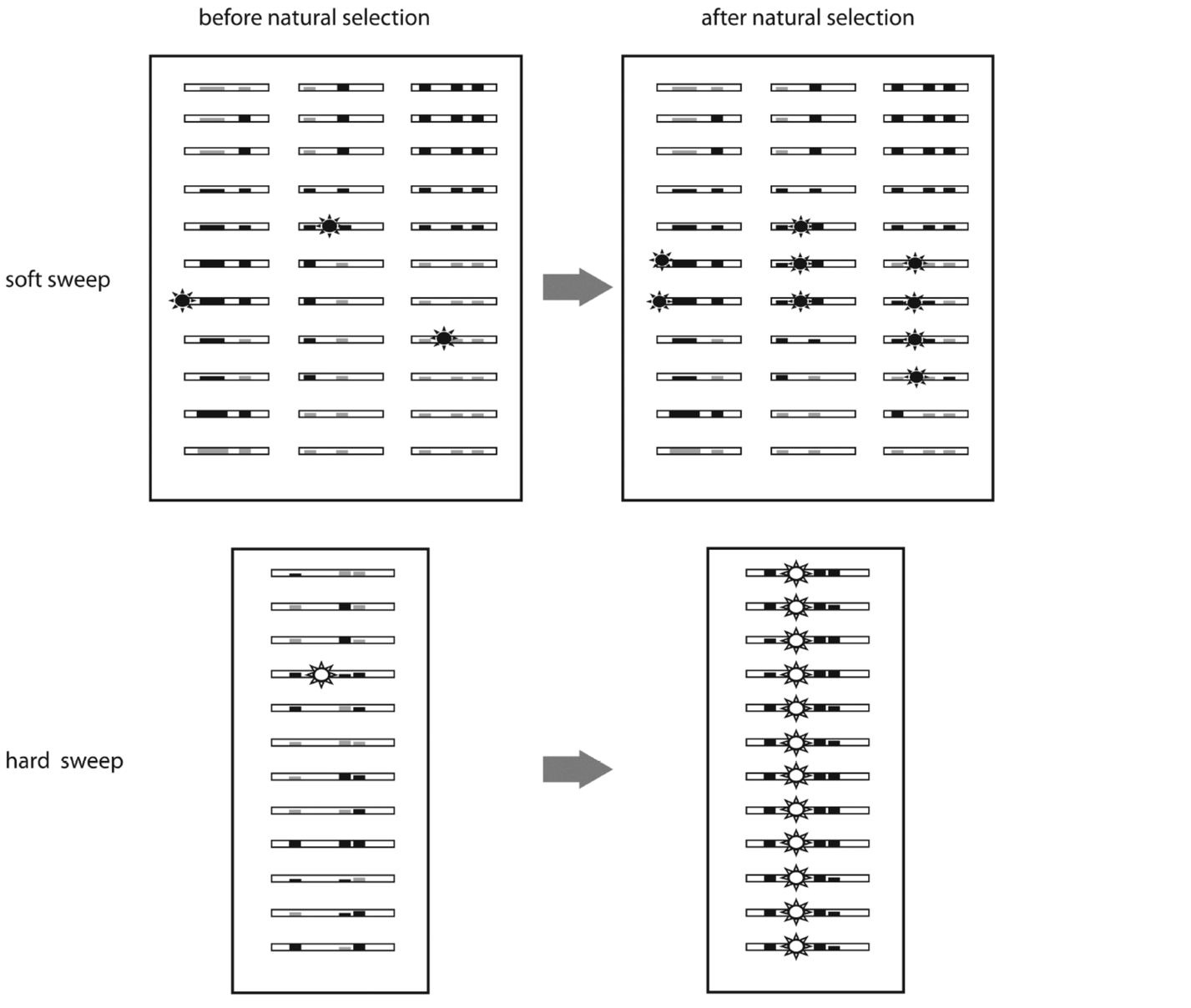


Figure 19.3 Sliding window scans of the boxer genome, showing departures from normal levels of heterozygosity. Two regions of the boxer genome that show higher than usual levels of homozygosity (a signal of a selective sweep) are identified as CFA 1 and CFA 26. CFA 1 represents a region on the first chromosome that is associated with brachycephaly, a disorder that results in a malformed head. CFA 26 refers to the novel region identified in the study that is associated with heart disorders. The numbers in these regions refer to the chromosomes in the boxer genome. The black and green symbols in the figure refer to alternating chromosomes. SNPs, single-nucleotide polymorphisms. (Adapted from J. Quilez, A.D. Short, & V. Martinez. *BMC Genomics* 12:339, 2011. Courtesy of BioMed Central.)

Table 19.4 Selective Sweep-Based Programs

Program	Website	Reference
SweepFinder	https://omictools.com/sweepfinder-tool	Nielsen et al., 2005
SweeD	http://pop-gen.eu/wordpress/software/sweed	Pavlidis et al., 2013
RAiSD	https://github.com/alachins/raisd	Alachiotis and Pavlidis, 2018
Sweep Dynamics (SD) plots	https://github.com/hzi-bifo/SDplots	Mooren et al., 2018
SAFE	https://github.com/alek0991/iSAFE	Akbari et al., 2018
LASSI	http://personal.psu.edu/mxd60/LASSI.html	Harris and DeGiorgio, 2019
selscan	https://github.com/szpiech/selscan	Szpiech and Hernandez, 2014
OmegaPlus	https://github.com/alachins/omegaplushttp://pop-gen.eu/wordpress/software	Alachiotis and Pavlidis, 2016

Hard and soft sweeps produce different effects in the genome



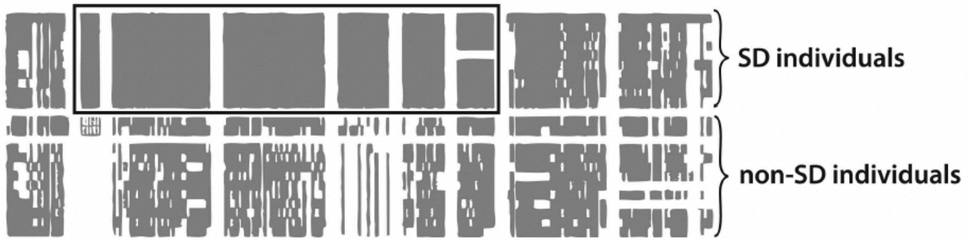


Figure 19.4 Bird's-eye view plots of sequence variation in the chromosomal regions near the segregation distorter locus in *Drosophila*. Green regions indicate identity of sequence to a reference genome with a segregation distorter (SD) chromosome. Hence all SD individuals in the figure (upper panel) show extreme similarity to the reference SD chromosome and, more importantly, high levels of homozygosity. The lower panel shows non-SD individuals; note the spottiness of the distribution of green. This pattern indicates a high degree of heterozygosity in the non-SD population. The rectangle shows the region of the chromosome that has undergone an extreme selective sweep in the SD individuals. (Adapted from D. Presgraves, PR. Gerard, A. Cherukuri, & T.W Lyttle. *PLoS Genetics* 5:e1000463, 2009. Courtesy of the Public Library of Science.)

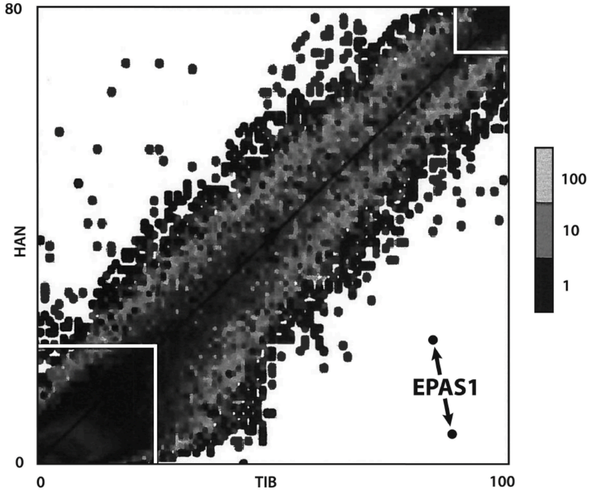


Figure 19.5 Plot of expected versus observed heterozygosity for single-nucleotide polymorphisms in Tibetan (TIB) and Chinese (HAN) populations. The diagonal in the diagram would be complete agreement of expected with observed heterozygosity. The points in the outlined boxes are less than 1000 and do not correspond to the scale bar on the right. For the TIB and HAN populations, two single-nucleotide polymorphisms (SNPs) stand out as departing significantly from the expected heterozygosity. Both of these SNPs reside in the EPAS1 gene. (Adapted from X. Yi, Y. Liang, E. Huerta- Sanchez, et al. *Science* 329:75–78, 2010.)

Regions of the human genome experience accelerated evolution

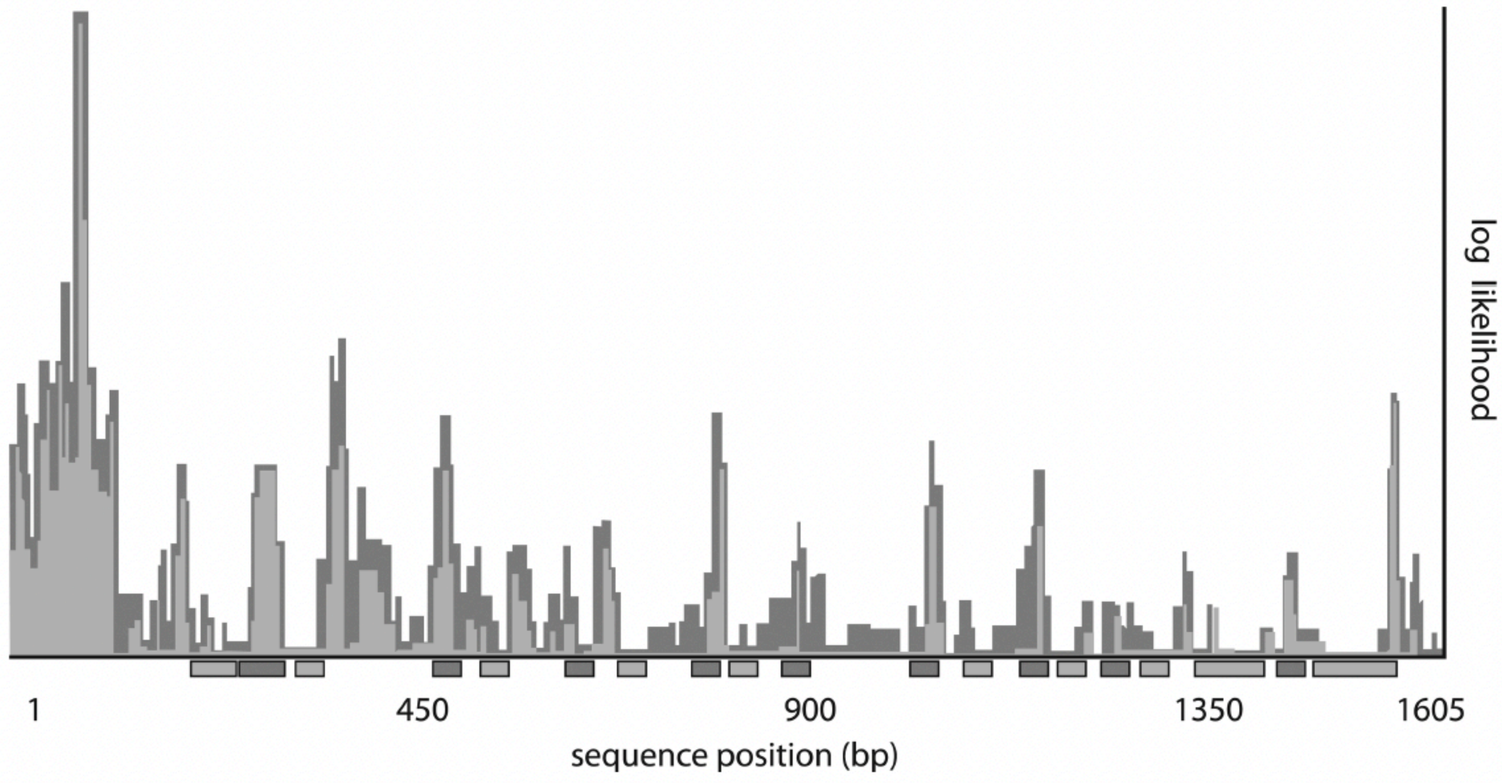
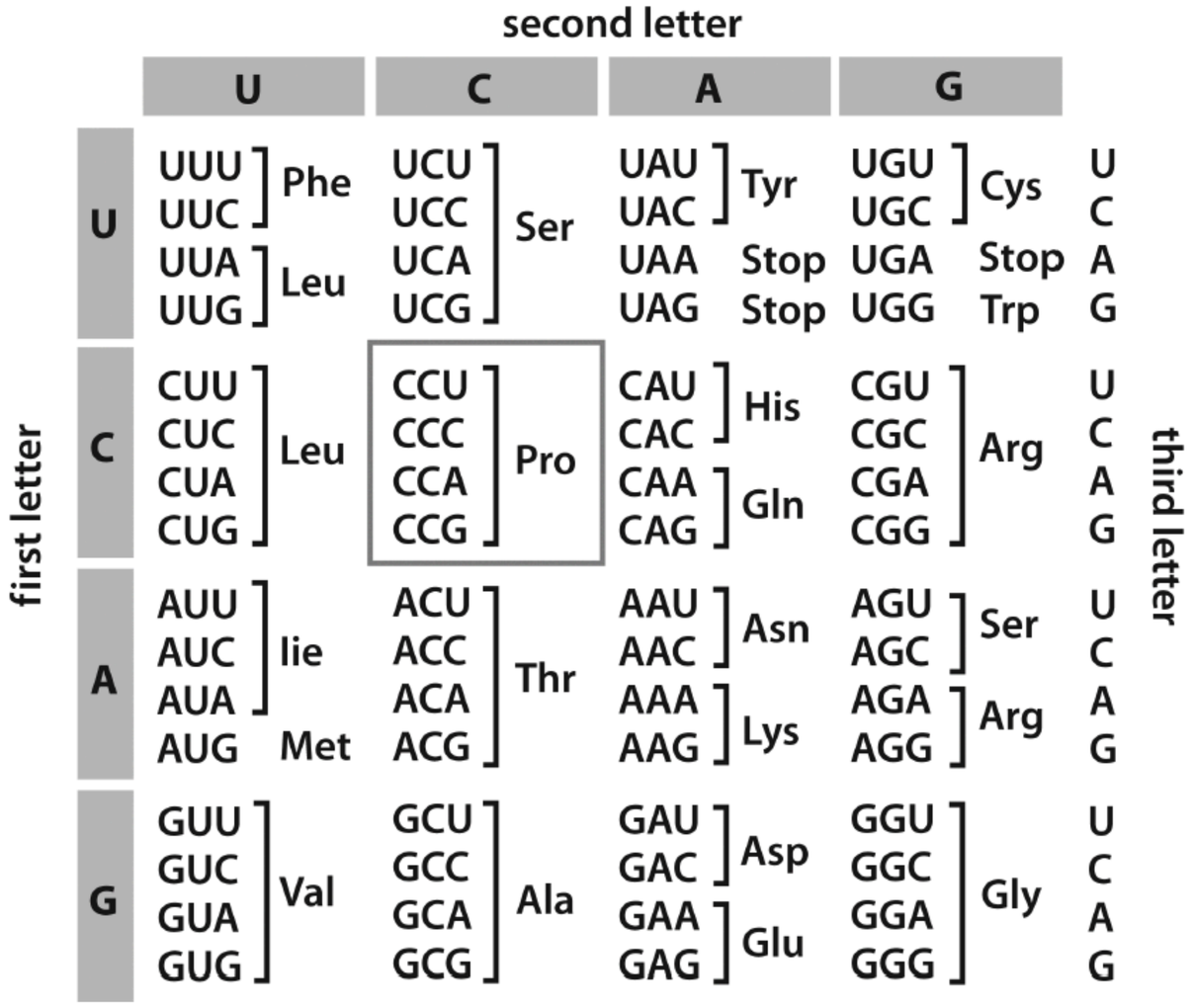


Figure 19.7 Phylogenetic shadowing comparison of the Apo(a) protein over several mammalian taxa. The graph shows where in the length of the gene strong conservation occurs across the large number of mammal species (represented by bars under the *x*-axis). The differences in length of the gray and green bars refer to the differences in length of the signal detected. The position in the gene is given in base pairs along the *x*-axis, and the degree of conservation is given as log likelihood on the *y*-axis. bp, base pair position on the chromosome. (Adapted from I. Ovcharenko, D. Boffelli, and G. Loots. *Genome Research* 14:1191–1198, 2004.)

Analyzing DNA Sequences for Natural Selection



Several variables affect the detection of natural selection at the genomic level

accurately calculating the dN and dS rates is not trivial.

natural selection on a protein is
not necessarily constant during the divergence of that protein. Different lineages are subject to different selection pressures during phylogenetic divergence.

natural selection may be different in different parts of a protein, a situation created by the modular structure of proteins in which domains with specific functions are interspersed with regions where other functions are implemented.

Sidebar 20.1 Definitions and steps for approximate calculation of dN/dS

The first step is to define dN and dS and related terms:

dN	total number of nonsynonymous substitutions observed in two sequences
dS	total number of synonymous substitutions observed in two sequences
S	number of potential synonymous substitutions [average for two compared sequences, $(S_1 + S_2)/2$]
N	number of potential nonsynonymous substitutions [average for two compared sequences, $(N_1 + N_2)/2$]
pS	proportion of observed to potential synonymous substitutions
pN	proportion of observed to potential nonsynonymous substitutions

To calculate dN/dS via the approximate method, follow the steps given:

Step 1. Count the number of potential synonymous (S) or silent sites in sequence 1.

Step 2. Count the number of potential synonymous or silent sites in sequence 2.

Step 3. Count the number of potential nonsynonymous (N) or replacement sites in sequence 1.

Step 4. Count the number of potential nonsynonymous or replacement sites in sequence 2.

Step 5. Obtain the average for S and N from the values for the two sequences being compared.

Step 6. Count the actual number of synonymous and nonsynonymous differences between sequences 1 and 2.

Step 7. Correct for codon pathways and multiple substitutions at the same site.

						S							S					S	
	M	V	L	S	D	A	E	W	Q	L	V	L	N	I	W	A	K	V	E	A
						*							*						*	
whale	atg	gtg	ctc	agc	gac	gca	gaa	tgg	cag	ttg	gtg	ctg	aac	atc	tgg	gcg	aag	gtg	gaa	gct
sheep	atg	ggg	ctc	agc	gac	ggg	gaa	tgg	cag	ttg	gtg	ctg	aat	gcc	tgg	ggg	aag	gtg	gag	gct
		#				#								##		#				
	M	G	L	S	D	G	E	W	Q	L	V	L	N	A	W	G	K	V	E	A
		R				R								R		R				
														R						

On the basis of these results, would it be possible to calculate dN/dS as 5/3 = 1.67?

NOPE

Scaling for redundancy and getting the number of potential substitutions is necessary for determining dN/dS

Sidebar 20.2 Potential number of silent and replacement events for the 20 amino acids

For codons that are one-, two-, three-, and fourfold degenerate, we have listed the potential events by category. For the three amino acids that are sixfold degenerate (that is, L, R, and S have six codons each), the silent and replacement events vary for each codon, so we have listed these counts per individual codon.

Degeneracy	Amino Acids	Silent	Replacement
Onefold	M, W	0	3
Twofold	F, Y, H, Q, N, K, D, E, C	1/3	8/3
Threefold	I	2/3	7/3
Fourfold	V P, T, A, G	1	2
Sixfold	L, R, S	Variable	Variable
Codons	**Silent**	**Replacement**	
L			
TTA	2/3	7/3	
TTG	2/3	7/3	
CTG	4/3	5/3	
CTA	4/3	5/3	
CTT	1	2	
CTC	1	2	
R			
AGA	2/3	7/3	
AGG	2/3	7/3	
CGA	4/3	5/3	
CGG	4/3	5/3	
CGT	1	2	
CGC	1	2	
S			
AGT	2/3	7/3	
AGC	2/3	7/3	
TCT	4/3	5/3	
TCC	4/3	5/3	
TCG	1	2	
TCA	1	2	

	M	V	L	S	D	A	E	W	Q	L	V	L	N	I	W	A	K	V	E	A
whale	atg	gtg	ctc	agc	gac	gca	gaa	tgg	cag	ttg	gtg	ctg	aac	atc	tgg	gcg	aag	gtg	gaa	gct
	0	1	1	2/3	1/3	1	1/3	0	1/3	2/3	1	4/3	1/3	2/3	0	1	1/3	1	1/3	1

$$S_{whale} = 3(0) + 7(1) + 1(4/2) + 3(2/2) + 6(1/2) = 12.33 \quad (20.2)$$

sheep	atg	ggg	ctc	agc	gac	ggg	gaa	tgg	cag	ttg	gtg	ctg	aat	gcc	tgg	ggg	aag	gtg	gag	gct
	M	G	L	S	D	G	E	W	Q	L	V	L	N	A	W	G	K	V	E	A
	0	1	1	2/3	1/3	1	1/3	0	1/3	2/3	1	4/3	1/3	1	0	1	1/3	1	1/3	1

$$S_{sheep} = 3(0) + 8(1) + 1(4/3) + 2(2/3) + 6(1/3) = 12.67 \quad (20.3)$$

To obtain the number of potential synonymous sites (S), the average for the two sequences is calculated as follows:

$$S = \left(S_{whale} + S_{sheep}\right)/2 = (12.33 + 12.67)/2 = 12.5 \quad (20.4)$$

	M	V	L	S	D	A	E	W	Q	L	V	L	N	I	W	A	K	V	E	A
whale	atg	gtg	ctc	agc	gac	gca	gaa	tgg	cag	ttg	gtg	ctg	aac	atc	tgg	gcg	aag	gtg	gaa	gct
	3	2	2	7/3	8/3	2	8/3	3	8/3	7/3	2	5/3	8/3	7/3	3	2	8/3	2	8/3	2

$$N_{whale} = 3(3) + 7(2) + 1\left(\tfrac{5}{3}\right) + 3\left(\tfrac{7}{3}\right) + 6\left(\tfrac{8}{3}\right) = 47.67 \tag{20.5}$$

sheep	atg	ggg	ctc	agc	gac	ggg	gaa	tgg	cag	ttg	gtg	ctg	aat	gcc	tgg	ggg	aag	gtg	gag	gct
	M	G	L	S	D	G	E	W	Q	L	V	L	N	A	W	G	K	V	E	A
	3	2	2	7/3	8/3	2	8/3	3	8/3	7/3	2	5/3	8/3	2	3	2	8/3	2	8/3	2

$$N_{sheep} = 3(3) + 8(2) + 1\left(\tfrac{5}{3}\right) + 2\left(\tfrac{7}{3}\right) + 6\left(\tfrac{8}{3}\right) = 47.33 \tag{20.6}$$

To obtain the number of potential nonsynonymous sites (N), the average for the two sequences is calculated as follows:

$$\mathrm{N} = \left(N_{whale} + N_{sheep}\right)/2 = \left(47.67 + 47.33\right)/2 = 47.5 \tag{20.7}$$

We can now calculate the proportion of observed synonymous (pS) and non-synonymous (pN) substitutions (where proportion = actual number observed/potential number calculated):

$$pS = 3 / 12.5 = 0.24 \quad (20.8)$$

$$pN = 5 / 47.5 = 0.105 \quad (20.9)$$

The ratio pN/pS, which we will take as a proxy for dN/dS, would therefore be 0.105/0.24 = 0.438. Since this ratio is less than 1.0, it would imply that these sequences are evolving as a result of purifying selection, which is a different inference than if we had simply calculated the raw synonymous and non-synonymous changes that have occurred since the divergence of these two species.

Does that mean we are done??

NOPE

Pathways of codon change are an important element in calculating dN/dS

Although the above corrections for redundancy yield a partial solution for a proxy for dN/dS, we need to take into account the different pathways that can occur in analyzing each codon before we can settle on the final dN/dS. For each codon in an alignment, there are three positions that need to be considered. For a given pair of codons, there can either be zero changes (the two codons are identical at the DNA sequence level), one change (only one position in the codon is different between the two sequences), two changes (two positions are different), or three changes (all three nucleotide states are different). In terms of the possible pathways that could be followed to arrive at these differences, the following observations apply.

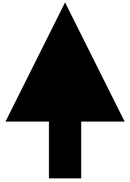
- If there is one difference between the sequences, there is one pathway.
- If there are two differences between sequences, there are two pathways.
- If there are three differences between the sequences, there are six pathways.

The best way to account for these pathways is to apply a maximum likelihood (ML) approach to correct for the potential occurrence of multiple pathways. In the ML approach, certain pathways are weighted differently. To demonstrate this approach, we examine the problem using a simpler counting method developed by Nei and Gojobori that does not weight the kinds of changes; this is called an unweighted counting method.

Look at the sixth position in the whale and sheep sequences:

	A
	*
whale	gca
sheep	ggg
	#
	G

	M	V	L	S	D	A	E	W	Q	L	V	L	N	I	W	A	K	V	E	A
						*							*						*	
whale	atg	gtg	ctc	agc	gac	gca	gaa	tgg	cag	ttg	gtg	ctg	aac	atc	tgg	gcg	aag	gtg	gaa	gct
sheep	atg	ggg	ctc	agc	gac	ggg	gaa	tgg	cag	ttg	gtg	ctg	aat	gcc	tgg	ggg	aag	gtg	gag	gct



In order for the two species to possess their codon states, one needs to consider the identity of their common ancestor. The nucleotide state of the common ancestor of whale and sheep is inferred through character reconstruction as explained below. In the codon shown above, there are differences in two positions (marked by * and #). Therefore, two pathways need to be examined.

	A
	*
whale	gca
sheep	ggg
	#
	G

A	G	G
gca →	gga →	ggg
A	A	G
gca →	gcg →	ggg

Each pathway requires one nonsynonymous (N) change and one synonymous (S) change:

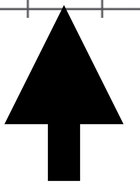
			S	N
A	G	G		
gca →	gga →	ggg	1	1
A	A	G		
gca →	gcg →	ggg	1	1

In the comparison of sheep and whale myoglobin, there are two potential synonymous replacements and two potential nonsynonymous replacements for the two pathways. The adjustment for this codon is simple. The proportion of potential changes that are synonymous is $2/4$ or $1/2$, and likewise $2/4$ or $1/2$ of the potential changes are nonsynonymous. If these proportions are applied to the two actual changes, $1/2(2)=1$ should be synonymous and $1/2(2)=1$ should be nonsynonymous. Because this matches the proportion of observed changes, no adjustment is needed by the multiple pathways correction.

The other codon with 2 or more changes is the 14th amino acid in our sequence:

	I
whale	atc
sheep	gcc
	##
	A

	M	V	L	S	D	A	E	W	Q	L	V	L	N	I	W	A	K	V	E	A
						*							*						*	
whale	atg	gtg	ctc	agc	gac	gca	gaa	tgg	cag	ttg	gtg	ctg	aac	atc	tgg	gcg	aag	gtg	gaa	gct
sheep	atg	ggg	ctc	agc	gac	ggg	gaa	tgg	cag	ttg	gtg	ctg	aat	gcc	tgg	ggg	aag	gtg	gag	gct



Again, there are two differences and therefore two pathways of change between these two codon states:

I	T	A
atc →	acc →	gcc
I	V	A
atc →	gtc →	gcc

In this case, each pathway leads to two nonsynonymous (N) changes and zero synonymous changes (S):

			S	N
I	T	A		
atc →	acc →	gcc	0	2
I	V	A		
atc →	gtc →	gcc	0	2

Codon change pathways can be used to account for redundancy

To demonstrate how the approximate approaches to estimating natural selection are accomplished, we will use the carboxy terminus of the myoglobin gene of sheep and whale as an example. This sequence is shown below.

	R	H	P	G	D	F	G	A	D	A	Q	A	A	M	N	K	A
				*		*		*	*	*		*					
whale	agg	cat	cct	ggg	gac	ttt	ggt	gcc	gac	gcc	cag	gca	gcc	atg	aac	aag	gcc
sheep	aag	cat	cct	tca	gac	ttc	ggt	gct	gat	gca	cag	ggc	gcc	atg	agc	aag	gcc
	#			##								#			#		
	K	H	P	S	D	F	G	A	D	A	Q	G	A	M	S	K	A

Scoring individual positions in each codon for their degeneracy in the whale sequence yields

	R	H	P	G	D	F	G	A	D	A	Q	A	A	M	N	K	A
whale	agg	cat	cct	ggg	gac	ttt	ggt	gcc	gac	gcc	cag	gca	gcc	atg	aac	aag	gcc
	214	112	114	114	112	112	114	114	112	114	112	114	114	111	112	112	114

For the sheep sequence, scoring individual positions for their degeneracy results in the following:

sheep	aag	cat	cct	tca	gac	ttc	ggt	gct	gat	gca	cag	ggc	gcc	atg	agc	aag	gcc
	K	H	P	S	D	F	G	A	D	A	Q	G	A	M	S	K	A
	112	112	114	214	112	112	114	114	112	114	112	114	114	111	214	112	114

The whale sequence has 34 positions of onefold degeneracy, 8 positions of twofold degeneracy, and 9 positions of fourfold degeneracy.

$$S_{whale} = 34(0) + 8\left(1/3\right) + 9(1) = 11.67 \tag{20.10}$$

$$N_{whale} = 34(1) + 8\left(2/3\right) + 9(0) = 39.33 \tag{20.11}$$

The sheep sequence has 33 positions of onefold degeneracy, 9 positions of twofold degeneracy, and 9 positions of fourfold degeneracy.

$$S_{sheep} = 33(0) + 9\left(1/3\right) + 9(1) = 12 \tag{20.12}$$

$$N_{sheep} = 33(1) + 9\left(2/3\right) + 9(0) = 39 \tag{20.13}$$

The number of potential synonymous sites is therefore [(11.67 + 12)/2] = 11.835, and the number of potential nonsynonymous sites is [(39.33 + 39)/2] = 39.165. Since there are six synonymous differences (marked by *) between the two sequences, dS is 6/11.835 = 0.507; dN is 5/39.165 = 0.128 (there are five nonsynonymous differences, marked by #); and dN/dS is 0.128/0.507, or 0.252. These results are clearly in the purifying selection zone. However, there is one codon in the comparison that requires further examination because of the pathway problem:

G
*
ggg
tca
##
S

There are three changes in this codon: two are nonsynonymous (marked by #) and one is synonymous (marked by *). These three changes could have arisen by six possible pathways, which requires that we recalculate the potential synonymous (S) and nonsynonymous (N) changes that can occur in this codon.

The six pathways are as follows:

					S	N
(1)	G	G	A	S	1	2
	ggg →	gga →	gca →	tca		
(2)	G	A	A	S		
	ggg →	gcg →	gca →	tca	1	2
(3)	G	W	stp	S	–	–
	ggg →	tgg →	tga →	tca		
(4)	G	W	stp	S	–	–
	ggg –	gga –	tga –	tca		
(5)	G	A	S	S	1	2
	ggg –	gcg –	tcg –	tca		
(6)	G	W	S	S	1	2
	ggg –	tgg –	tcg –	tca		

Pathways 3 and 4 are ignored because they require stop codons as intermediates in the reconstructions, and such stop codons would disrupt the protein and be highly disadvantageous. Since this evolutionary scenario would be highly unlikely compared to other pathways, these pathways can be ignored.

So does this mean we are done??

NOPE

This analysis results in four pathways of three changes each, for a total of 12 potential

changes, with 4 of the 12 being synonymous $\left(\frac{4}{12}=\frac{1}{3}\right)$ and 8 of the 12 being nonsynonymous $\left(\frac{8}{12}=\frac{2}{3}\right)$. If these proportions are applied to the three actual changes, $\frac{1}{3}(3)=1$ should be synonymous and $\frac{2}{3}(3)=2$ should be nonsynonymous. This matches the results found, so no corrections are needed in this case. However, in some cases not all codon pathway changes can be accommodated by simple counting as we will see in the next chapter.

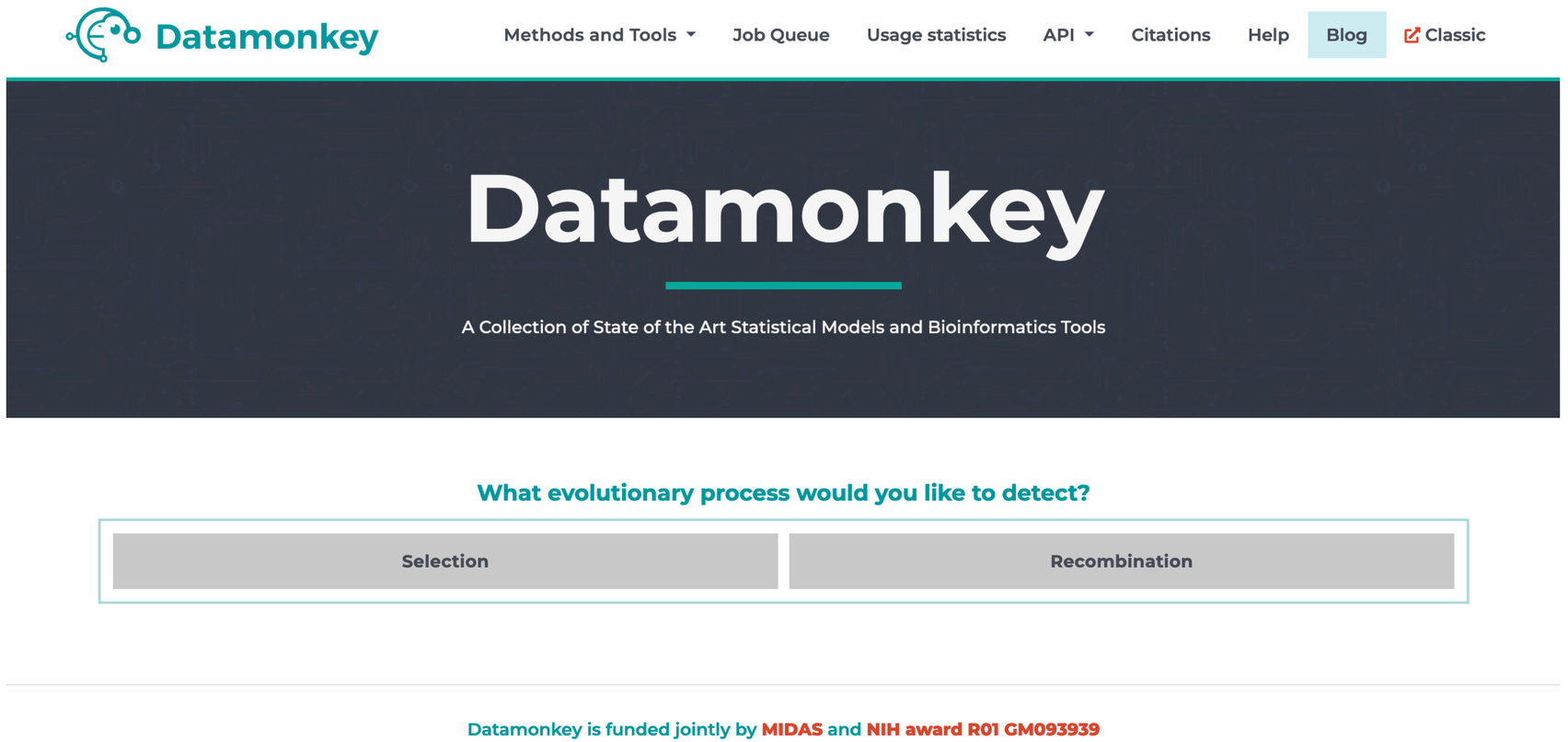
Accounting for Multiple Hits in DNA Sequences for dN/dS Measures

Estimating natural selection requires adjusting the calculation of sequence changes

In the examples provided above and in the previous chapter, we have moved from simple counting of nucleotide changes to adjusting the method for calculating synonymous and nonsynonymous changes by the following protocol:

- Align sequences.
- Count the number of "actual" synonymous and nonsynonymous sites in each sequence.
- Obtain the total of synonymous and nonsynonymous changes by averaging for sequences in the comparison.
- Count the number of "potential" changes by summing all possible pathways between each pair of codons. For the majority of changes in closely related species and for slowly evolving sequences, the correction for pathways will be trivial.
- Correct for the substitution model.

DATA MONKEY



https://www.datamonkey.org/