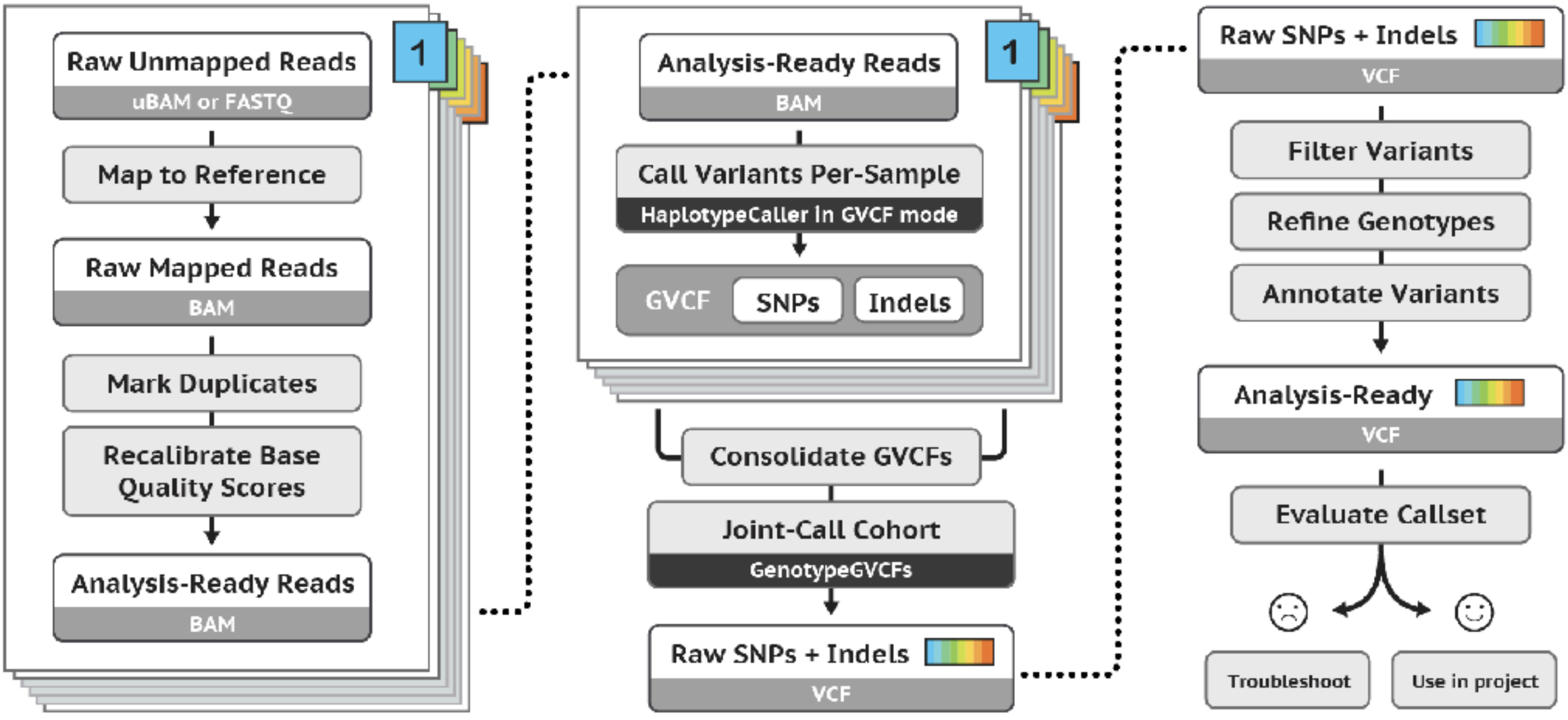
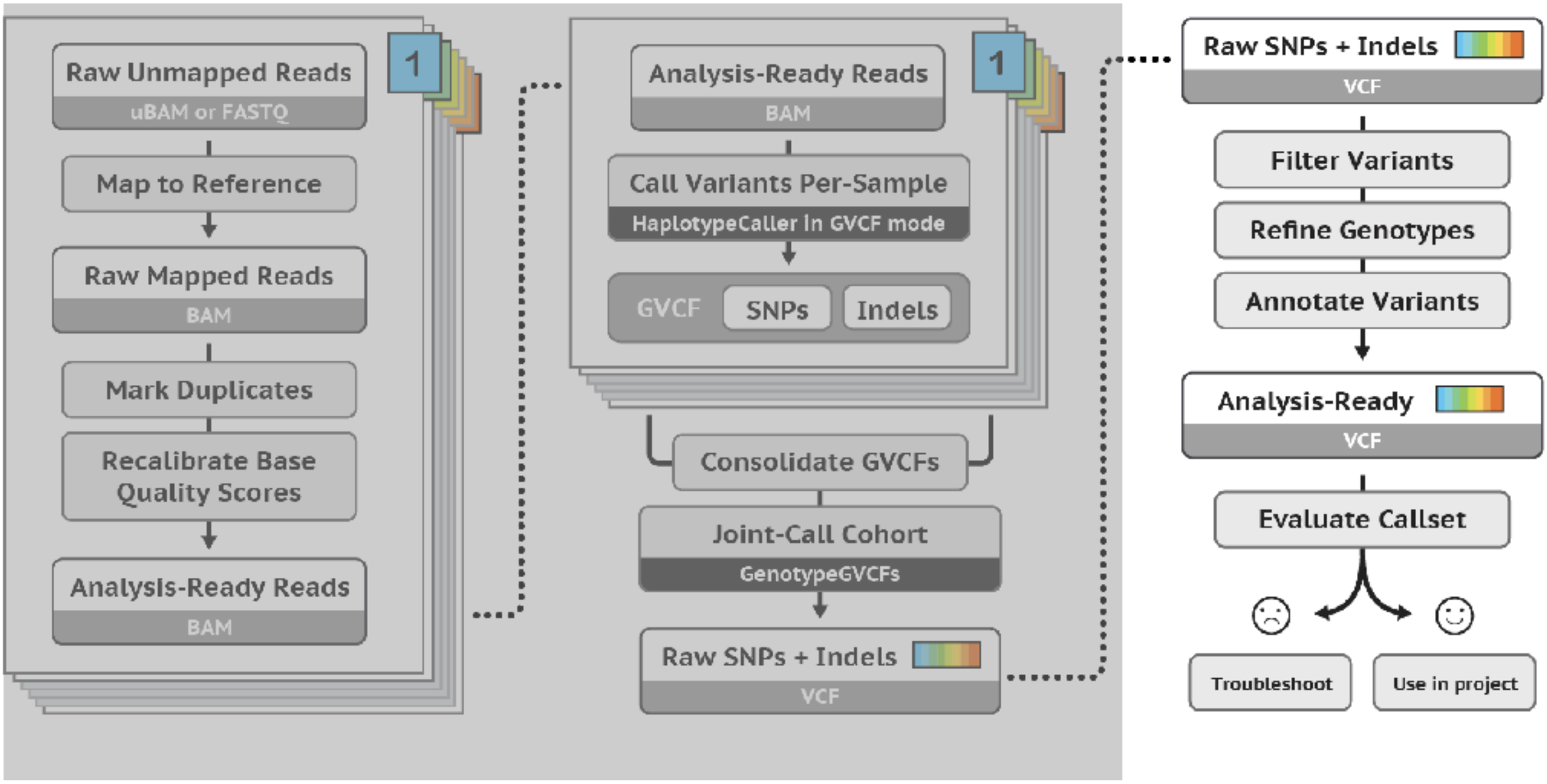
Topic 9: SNP Filtering

and Analysis

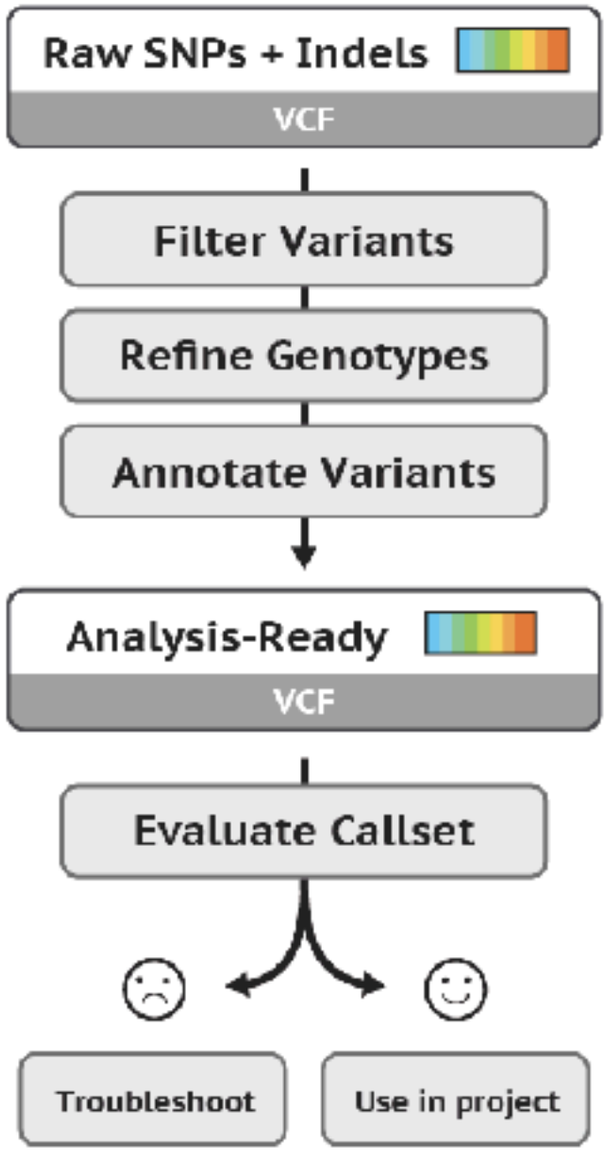
Overview


Raw Unmapped Reads
uBAM or FASTQ
Map to Reference
Raw Mapped Reads
BAM
Mark Duplicates
Recalibrate Base Quality Scores
Analysis-Ready Reads
BAM
1
Analysis-Ready Reads
BAM
Call Variants Per-Sample
HaplotypeCaller in GVCF mode
GVCF
SNPs
Indels
1
Consolidate GVCFs
Joint-Call Cohort
GenotypeGVCFs
Raw SNPs + Indels
VCF
Raw SNPs + Indels
VCF
Filter Variants
Refine Genotypes
Annotate Variants
Analysis-Ready
VCF
Evaluate Callset
Troubleshoot
Use in project

Overview


Raw Unmapped Reads
uBAM or FASTQ
Map to Reference
Raw Mapped Reads
BAM
Mark Duplicates
Recalibrate Base Quality Scores
Analysis-Ready Reads
BAM
1
Analysis-Ready Reads
BAM
Call Variants Per-Sample
HaplotypeCaller in GVCF mode
GVCF
SNPs
Indels
1
Consolidate GVCFs
Joint-Call Cohort
GenotypeGVCFs
Raw SNPs + Indels
VCF
Raw SNPs + Indels
VCF
Filter Variants
Refine Genotypes
Annotate Variants
Analysis-Ready
VCF
Evaluate Callset
Troubleshoot
Use in project

Overview



Today we'll focus on SNP filtering, annotation and one analysis in particular

There are MANY different ways analyze population genomic data

Too many to cover them all in one lecture

Why filter data?

Are all variants equally reliable?

Review: VCF Files

```
##INFO=<ID=MLEAC,Number=A,Type=Integer,Description="Maximum likelihood expectation (MLE) for the allele counts (not necessarily t
##INFO=<ID=MLEAF,Number=A,Type=Float,Description="Maximum likelihood expectation (MLE) for the allele frequency (not necessarily
##INFO=<ID=MQ,Number=1,Type=Float,Description="RMS Mapping Quality">
##INFO=<ID=MQRankSum,Number=1,Type=Float,Description="Z-score From Wilcoxon rank sum test of Alt vs. Ref read mapping qualities">
##INFO=<ID=QD,Number=1,Type=Float,Description="Variant Confidence/Quality by Depth">
##INFO=<ID=RAW_MQandDP,Number=2,Type=Integer,Description="Raw data (sum of squared MQ and total depth) for improved RMS Mapping Q
##INFO=<ID=ReadPosRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test of Alt vs. Ref read position bias"
##INFO=<ID=SOR,Number=1,Type=Float,Description="Symmetric Odds Ratio of 2x2 contingency table to detect strand bias">
##contig=<ID=chr_1,length=5000000>
##contig=<ID=chr_2,length=5000000>
##source=GenomicsDBImport
##source=GenotypeGVCFs
##source=HaplotypeCaller
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  Chinook.p1.i1.r400000   Chinook.p1.i2.r400000   Chinook.
chr_1   163     .       T       C       179.59  .       AC=3;AF=0.375;AN=8;BaseQRankSum=-4.310e-01;DP=16;ExcessHet=1.0474;FS=0.0
chr_1   196     .       T       C       686.01  .       AC=8;AF=1.00;AN=8;DP=17;ExcessHet=3.0103;FS=0.000;MLEAC=7;MLEAF=0.875;MQ
chr_1   296     .       T       A       514.38  .       AC=6;AF=1.00;AN=6;DP=14;ExcessHet=3.0103;FS=0.000;MLEAC=6;MLEAF=1.00;MQ=
chr_1   726     .       A       C       714.91  .       AC=6;AF=1.00;AN=6;DP=20;ExcessHet=3.0103;FS=0.000;MLEAC=6;MLEAF=1.00;MQ=
chr_1   755     .       T       A       987.52  .       AC=6;AF=1.00;AN=6;DP=29;ExcessHet=3.0103;FS=0.000;MLEAC=7;MLEAF=1.00;MQ=
chr_1   804     .       T       C       173.03  .       AC=1;AF=0.125;AN=8;BaseQRankSum=-1.097e+00;DP=28;ExcessHet=3.0103;FS=2.63
chr_1   1052    .       G       T       1106.76 .       AC=8;AF=1.00;AN=8;DP=29;ExcessHet=3.0103;FS=0.000;MLEAC=8;MLEAF=1.00;MQ=
chr_1   1420    .       G       A       1181.88 .       AC=8;AF=1.00;AN=8;DP=30;ExcessHet=3.0103;FS=0.000;MLEAC=8;MLEAF=1.00;MQ=
chr_1   1492    .       C       G       645.47  .       AC=6;AF=0.750;AN=8;DP=26;ExcessHet=0.3218;FS=0.000;MLEAC=6;MLEAF=0.750;MQ
chr_1   1886    .       A       G       475.50  .       AC=4;AF=0.500;AN=8;BaseQRankSum=-4.310e-01;DP=22;ExcessHet=2.4304;FS=0.00
chr_1   1939    .       A       T       1122.43 .       AC=8;AF=1.00;AN=8;DP=29;ExcessHet=3.0103;FS=0.000;MLEAC=8;MLEAF=1.00;MQ=
chr_1   3434    .       A       G       691.97  .       AC=6;AF=1.00;AN=6;DP=18;ExcessHet=3.0103;FS=0.000;MLEAC=6;MLEAF=1.00;MQ=
chr_1   3462    .       A       C       543.54  .       AC=6;AF=1.00;AN=6;DP=14;ExcessHet=3.0103;FS=0.000;MLEAC=6;MLEAF=1.00;MQ=
chr_1   3851    .       T       C       504.65  .       AC=4;AF=0.500;AN=8;DP=20;ExcessHet=0.1902;FS=0.000;MLEAC=4;MLEAF=0.500;MQ
chr_1   4139    .       A       T       1007.38 .       AC=8;AF=1.00;AN=8;DP=26;ExcessHet=3.0103;FS=0.000;MLEAC=8;MLEAF=1.00;MQ=
chr_1   4267    .       A       G       303.58  .       AC=3;AF=0.375;AN=8;BaseQRankSum=-1.036e+00;DP=25;ExcessHet=1.0474;FS=0.00
chr_1   4455    .       G       C       187.46  .       AC=2;AF=0.250;AN=8;DP=20;ExcessHet=0.3218;FS=0.000;MLEAC=2;MLEAF=0.250;MQ
chr_1   4750    .       G       A       443.30  .       AC=2;AF=0.250;AN=8;DP=31;ExcessHet=0.3218;FS=0.000;MLEAC=2;MLEAF=0.250;MQ
chr_1   4780    .       G       A       144.69  .       AC=2;AF=0.250;AN=8;BaseQRankSum=1.28;DP=32;ExcessHet=0.3218;FS=0.000;MLEA
chr_1   5139    .       G       T       1078.75 .       AC=8;AF=1.00;AN=8;DP=28;ExcessHet=3.0103;FS=0.000;MLEAC=8;MLEAF=1.00;MQ=
chr_1   5354    .       G       C       327.28  .       AC=3;AF=0.375;AN=8;BaseQRankSum=1.53;DP=26;ExcessHet=1.0474;FS=2.059;MLE
```

Headers

Variants

VCF: Header

```
##fileformat=VCFv4.2
##FILTER=<ID=PASS,Description="All filters passed">
##ALT=<ID=NON_REF,Description="Represents any possible alternative allele not already represented at this
##FILTER=<ID=LowQual,Description="Low quality">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (reads with MQ=255 or with bad m
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=MIN_DP,Number=1,Type=Integer,Description="Minimum DP observed within the GVCF block">
##FORMAT=<ID=PGT,Number=1,Type=String,Description="Physical phasing haplotype information, describing how
phased in relation to one another; will always be heterozygous and is not intended to describe called alle
##FORMAT=<ID=PID,Number=1,Type=String,Description="Physical phasing ID information, where each unique ID w
not across samples) connects records within a phasing group">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods for genotypes as d
specification">
##FORMAT=<ID=PS,Number=1,Type=Integer,Description="Phasing set (typically the position of the first variar
##FORMAT=<ID=RGQ,Number=1,Type=Integer,Description="Unconditional reference genotype confidence, encoded a
p(genotype call is wrong)">
##FORMAT=<ID=SB,Number=4,Type=Integer,Description="Per-sample component statistics which comprise the Fish
strand bias.">
```

Contains detailed information on what each column contains, the file version, commands used to generate file etc.

Lines starting with ##

VCF: Records

```
##fileformat=VCFv4.2
##FILTER=<ID=PASS,Description="All filters passed">
##ALT=<ID=NON_REF,Description="Represents any possible alternative allele not already represented at this location by REF and ALT">
##FILTER=<ID=LowQual,Description="Low quality">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (reads with MQ=255 or with bad mates are filtered)">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=MIN_DP,Number=1,Type=Integer,Description="Minimum DP observed within the GVCF block">
##FORMAT=<ID=PGT,Number=1,Type=String,Description="Physical phasing haplotype information, describing how the alternate alleles are phased in
be heterozygous and is not intended to describe called alleles">
##FORMAT=<ID=PID,Number=1,Type=String,Description="Physical phasing ID information, where each unique ID within a given sample (but not acros
phasing group">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification">
##FORMAT=<ID=PS,Number=1,Type=Integer,Description="Phasing set (typically the position of the first variant in the set)">
##FORMAT=<ID=RGQ,Number=1,Type=Integer,Description="Unconditional reference genotype confidence, encoded as a phred quality -10*log10 p(genot
##FORMAT=<ID=SB,Number=4,Type=Integer,Description="Per-sample component statistics which comprise the Fisher's Exact Test to detect strand bi
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  -e Chinook.p1.i0
```

Contains detailed information on what each column contains, the file version, commands used to generate file etc.

VCF: Records

```
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  -e Chinook.p1.i0        -e Chinook.p1.i1
chr_1   102     .       C       T       173.69  .
AC=4;AF=0.026;AN=156;BaseQRankSum=0.524;DP=209;ExcessHet=0.0860;FS=0.000;InbreedingCoeff=0.2702;MLEAC=5;MLEAF=0.032;MQ=60.00;MQRankSum=0.0
0;QD=14.47;ReadPosRankSum=0.00;SOR=0.693      GT:AD:DP:GQ:PL  0/0:3,0:3:9:0,9,102     0/0:2,0:2:6

chr_1   163     .       T       C       2919.39 .
AC=38;AF=0.271;AN=140;BaseQRankSum=-1.800e-01;DP=235;ExcessHet=0.0000;FS=5.509;InbreedingCoeff=0.3039;MLEAC=56;MLEAF=0.400;MQ=60.00;MQRank
Sum=0.00;QD=27.54;ReadPosRankSum=0.00;SOR=3.587       GT:AD:DP:GQ:PL  0/0:4,0:4:12:0,12,144
```

Records for two SNPs

How do we know that they are SNPs?

VCF: Records - INFO

```
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  -e Chinook.p1.i0        -e Chinook.p1.i1
chr_1   102     .       C       T       173.69  .
AC=4;AF=0.026;AN=156;BaseQRankSum=0.524;DP=209;ExcessHet=0.0860;FS=0.000;InbreedingCoeff=0.2
702;MLEAC=5;MLEAF=0.032;MQ=60.00;MQRankSum=0.00;QD=14.47;ReadPosRankSum=0.00;SOR=0.693
GT:AD:DP:GQ:PL  0/0:3,0:3:9:0,9,102     0/0:2,0:2:6
```

```
AC=4
AF=0.026
AN=156
BaseQRankSum=0.524
DP=209
ExcessHet=0.0860
FS=0.000
InbreedingCoeff=0.2702
MLEAC=5
MLEAF=0.032
MQ=60.00
MQRankSum=0.00
QD=14.47
ReadPosRankSum=0.00
SOR=0.693
```

Semi-colon separated data held the INFO field

VCF: Records - INFO

```
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  -e Chinook.p1.i0        -e Chinook.p1.i1
chr_1   102     .       C       T       173.69  .
AC=4;AF=0.026;AN=156;BaseQRankSum=0.524;DP=209;ExcessHet=0.0860;FS=0.000;InbreedingCoeff=0.2
702;MLEAC=5;MLEAF=0.032;MQ=60.00;MQRankSum=0.00;QD=14.47;ReadPosRankSum=0.00;SOR=0.693
GT:AD:DP:GQ:PL  0/0:3,0:3:9:0,9,102     0/0:2,0:2:6
```

```
AC=4
AF=0.026
AN=156
BaseQRankSum=0.524
DP=209
ExcessHet=0.0860
FS=0.000
InbreedingCoeff=0.2702
MLEAC=5
MLEAF=0.032
MQ=60.00
MQRankSum=0.00
QD=14.47
ReadPosRankSum=0.00
SOR=0.693
```

Semi-colon separated data held the INFO field

```
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in g
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency, f
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of a
##INFO=<ID=BaseQRankSum,Number=1,Type=Float,Description="Z-score f
##INFO=<ID=DP,Number=1,Type=Integer,Description="Approximate read
##INFO=<ID=END,Number=1,Type=Integer,Description="Stop position of
##INFO=<ID=ExcessHet,Number=1,Type=Float,Description="Phred-scaled
##INFO=<ID=FS,Number=1,Type=Float,Description="Phred-scaled p-valu
##INFO=<ID=InbreedingCoeff,Number=1,Type=Float,Description="Inbree
per-sample when compared against the Hardy-Weinberg expectation">
##INFO=<ID=MLEAC,Number=A,Type=Integer,Description="Maximum likeli
the same as the AC), for each ALT allele, in the same order as lis
##INFO=<ID=MLEAF,Number=A,Type=Float,Description="Maximum likeliho
the same as the AF), for each ALT allele, in the same order as lis
##INFO=<ID=MQ,Number=1,Type=Float,Description="RMS Mapping Quality
##INFO=<ID=MQRankSum,Number=1,Type=Float,Description="Z-score From
##INFO=<ID=QD,Number=1,Type=Float,Description="Variant Confidence/
##INFO=<ID=RAW_MQandDP,Number=2,Type=Integer,Description="Raw data
Quality calculation. Incompatible with deprecated RAW_MQ formulati
##INFO=<ID=ReadPosRankSum,Number=1,Type=Float,Description="Z-score
bias">
##INFO=<ID=SOR,Number=1,Type=Float,Description="Symmetric Odds Rat
```

The Key to the INFO Field is in the header

VCF: Records - FORMAT

```
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  -e Chinook.p1.i0        -e Chinook.p1.i1
chr_1   102     .       C       T       173.69  .
AC=4;AF=0.026;AN=156;BaseQRankSum=0.524;DP=209;ExcessHet=0.0860;FS=0.000;InbreedingCoeff=0.2702;MLEAC=5;MLEAF=0.032;MQ=60.00;MQ
RankSum=0.00;QD=14.47;ReadPosRankSum=0.00;SOR=0.693     GT:AD:DP:GQ:PL  0/0:3,0:3:9:0,9,102     0/0:2,0:2:6
```

`GT:AD:DP:GQ:PL`

Colon separated key to the data in the column for each sample

`0/0:3,0:3:9:0,9,102`

Colon separated data for sample “Chinook.p1.i0”

VCF: Records - FORMAT

```
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  -e Chinook.p1.i0        -e Chinook.p1.i1
chr_1   102     .       C       T       173.69  .
AC=4;AF=0.026;AN=156;BaseQRankSum=0.524;DP=209;ExcessHet=0.0860;FS=0.000;InbreedingCoeff=0.2702;MLEAC=5;MLEAF=0.032;MQ=60.00;MQ
RankSum=0.00;QD=14.47;ReadPosRankSum=0.00;SOR=0.693     GT:AD:DP:GQ:PL  0/0:3,0:3:9:0,9,102     0/0:2,0:2:6
```

`GT:AD:DP:GQ:PL`

`0/0:3,0:3:9:0,9,102`

Colon separated key to the data in the column for each sample

Colon separated data for sample “Chinook.p1.i0”

The Key to abbreviations in the FORMAT field is in the header

Why filter data?

By default, GATK is very permissive
(It will output false positive sites!)

Two approaches to filtering:

Hard filtering

Variant recalibration

Intersect of diff. program SNP call sets

Hard filtering

Define fixed cutoff thresholds for various summary statistics

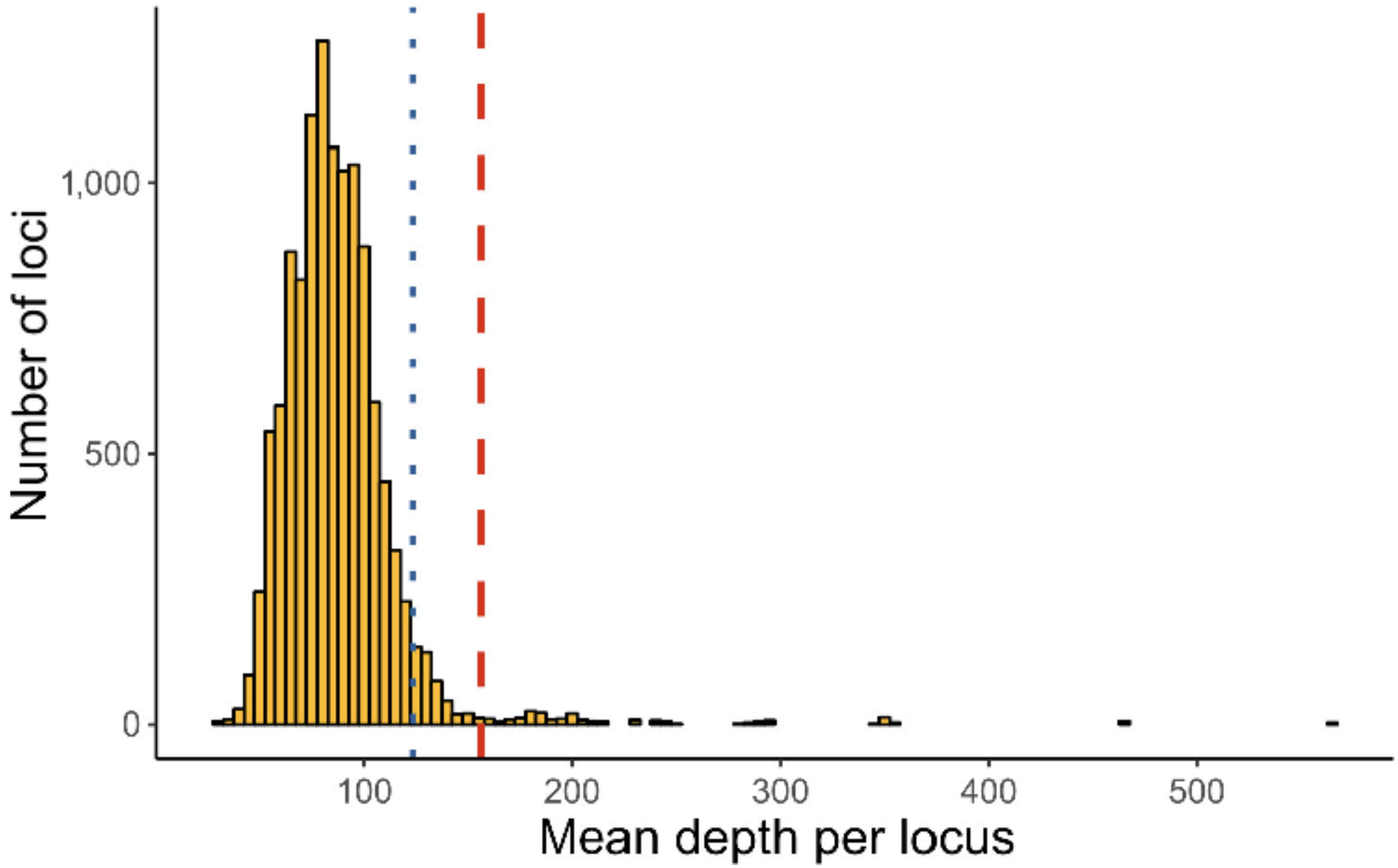
Bioinformatic filters:

- Genotype quality
- Individual depth
- Heterozygosity
- Strand bias

Statistical and population genetic filters:

- Allele frequency
- HWE Deviations
- Missing data
- Linkage disequilibrium

Hard filtering



Read: O'Leary et al 2018 - Molecular Ecology

Hard filtering

Low Confidence SNP Calls	**minDP > 5**	Minimum Depth/Coverage (for individual)
	Qual > 20	SNP Quality Score
	meanDP > 15	Mean Depth/Coverage (across individuals)
	mac < 3	Biallelic SNPs
Missing Data	**geno > 50%**	At least 50% of individuals have genotype
	imiss < 90%	Each individual has at least 90% of all SNP

Why filter data?

By default, GATK is very permissive
(It will output false positive sites!)

Two approaches to filtering:

- Hard filtering
- **Variant recalibration**

Intersection of outputs from different SNP calling programs

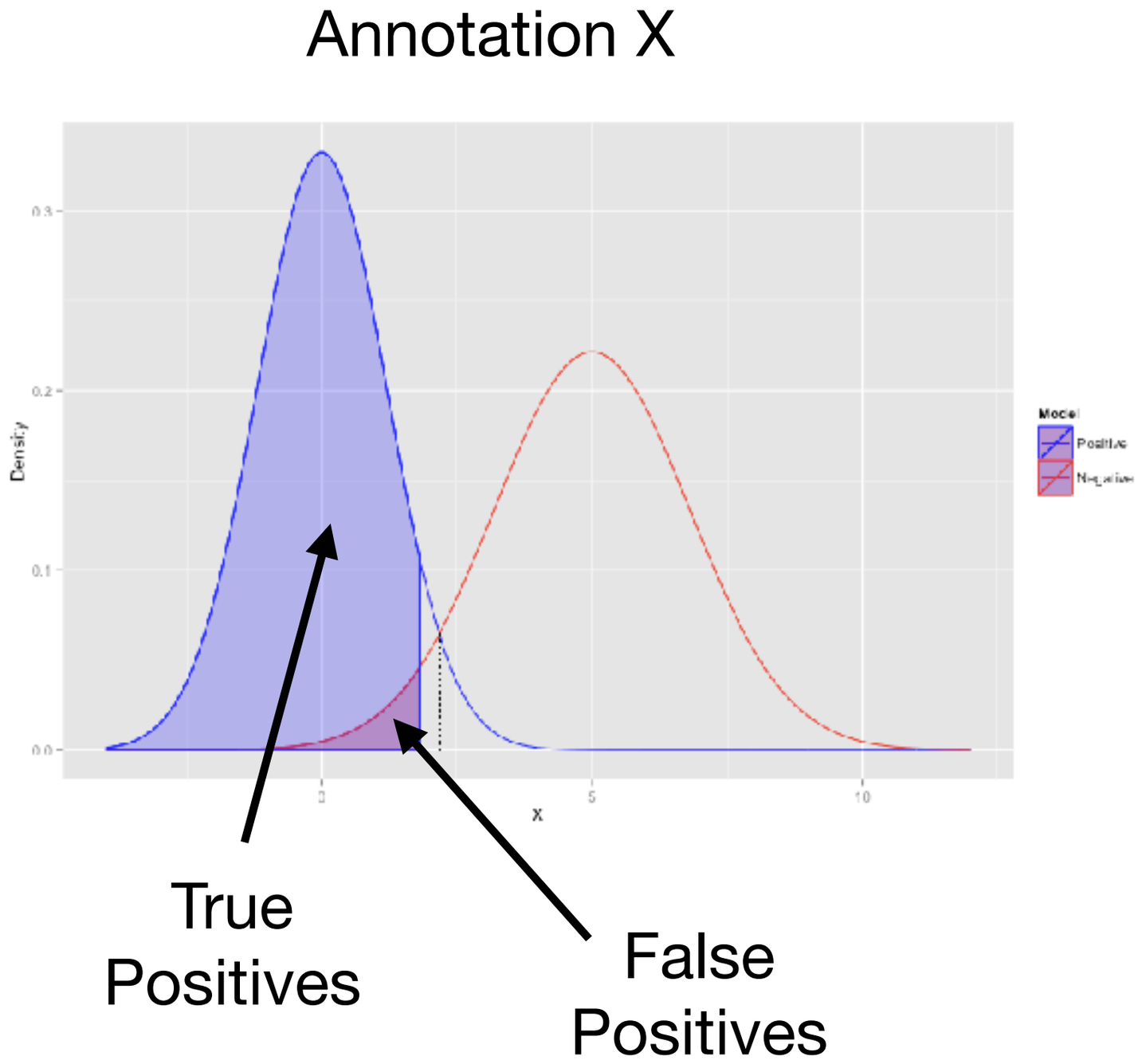
Variant Recalibration

GATK Method - Train a statistical model to learn what a “good” variant looks like

Assumes groups of annotations/statistics form Gaussian clusters

Build Gaussian mixture models from annotations of known variants from the dataset

Use this variation to score all variants



Variant Recalibration

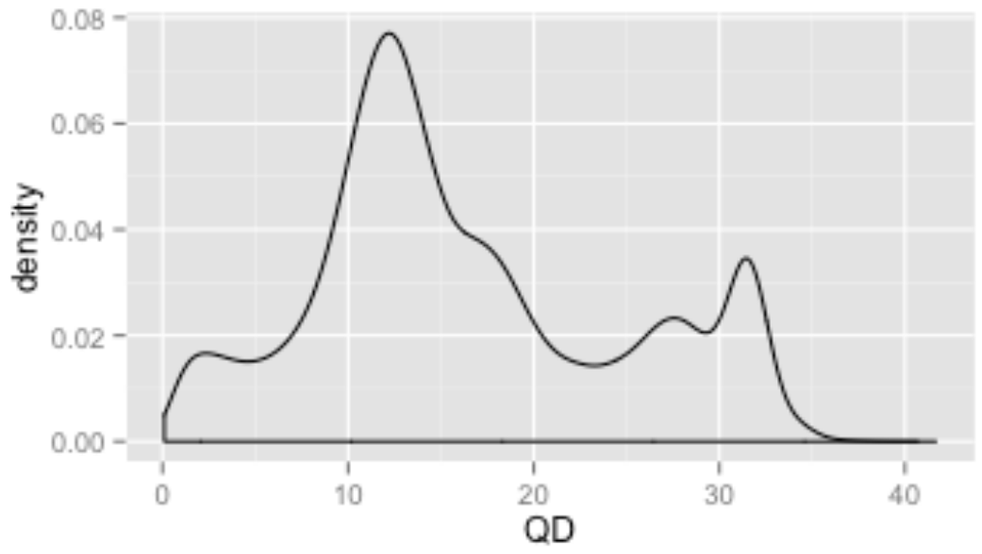
GATK Method - Train a statistical model to learn what a “good” variant looks like

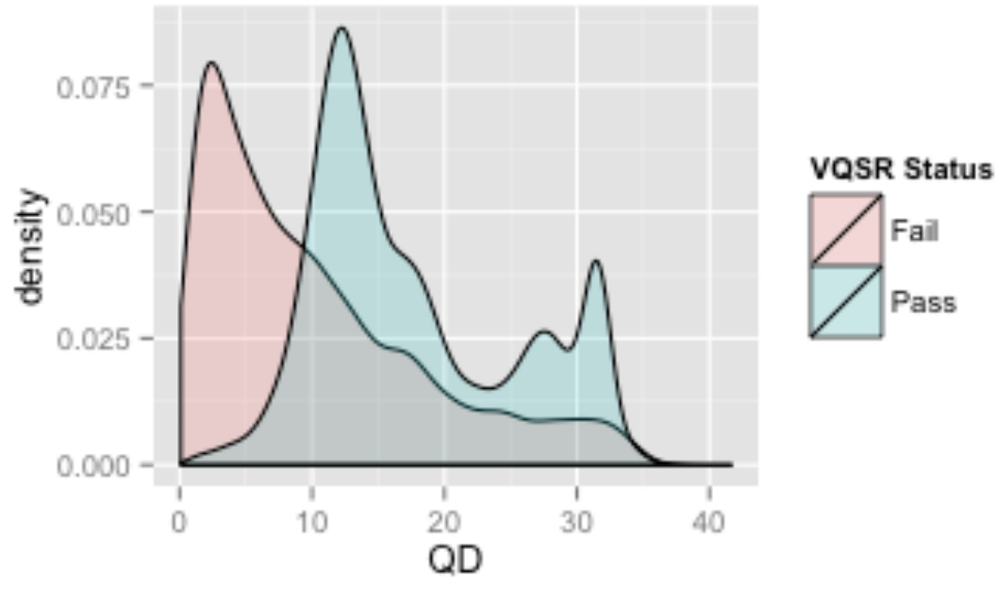
Assumes groups of annotations/statistics form Gaussian clusters

Build Gaussian mixture models from annotations of known variants from the dataset

Use this variation to score all variants

Quality By Depth





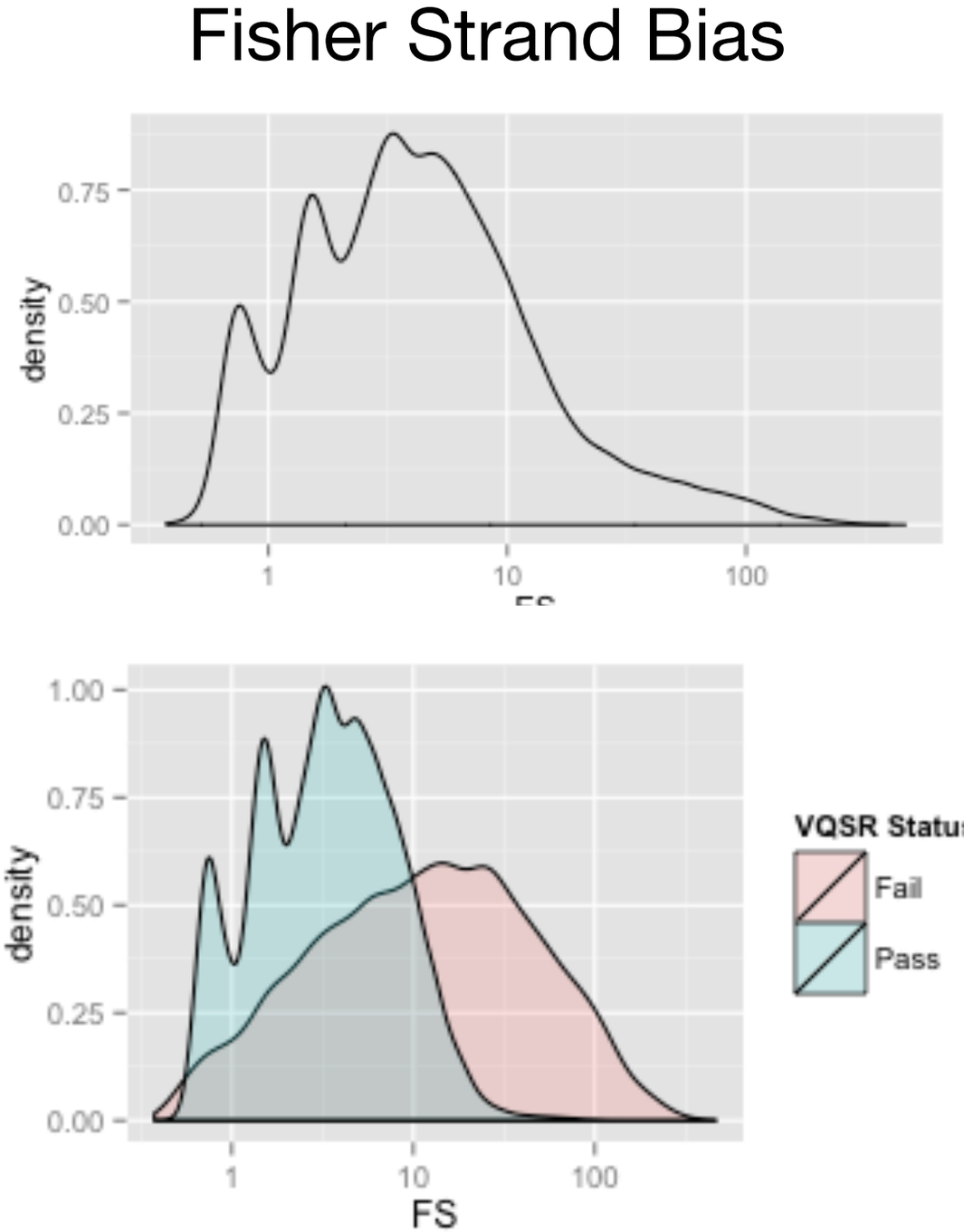
Variant Recalibration

GATK Method - Train a statistical model to learn what a “good” variant looks like

Assumes groups of annotations/statistics form Gaussian clusters

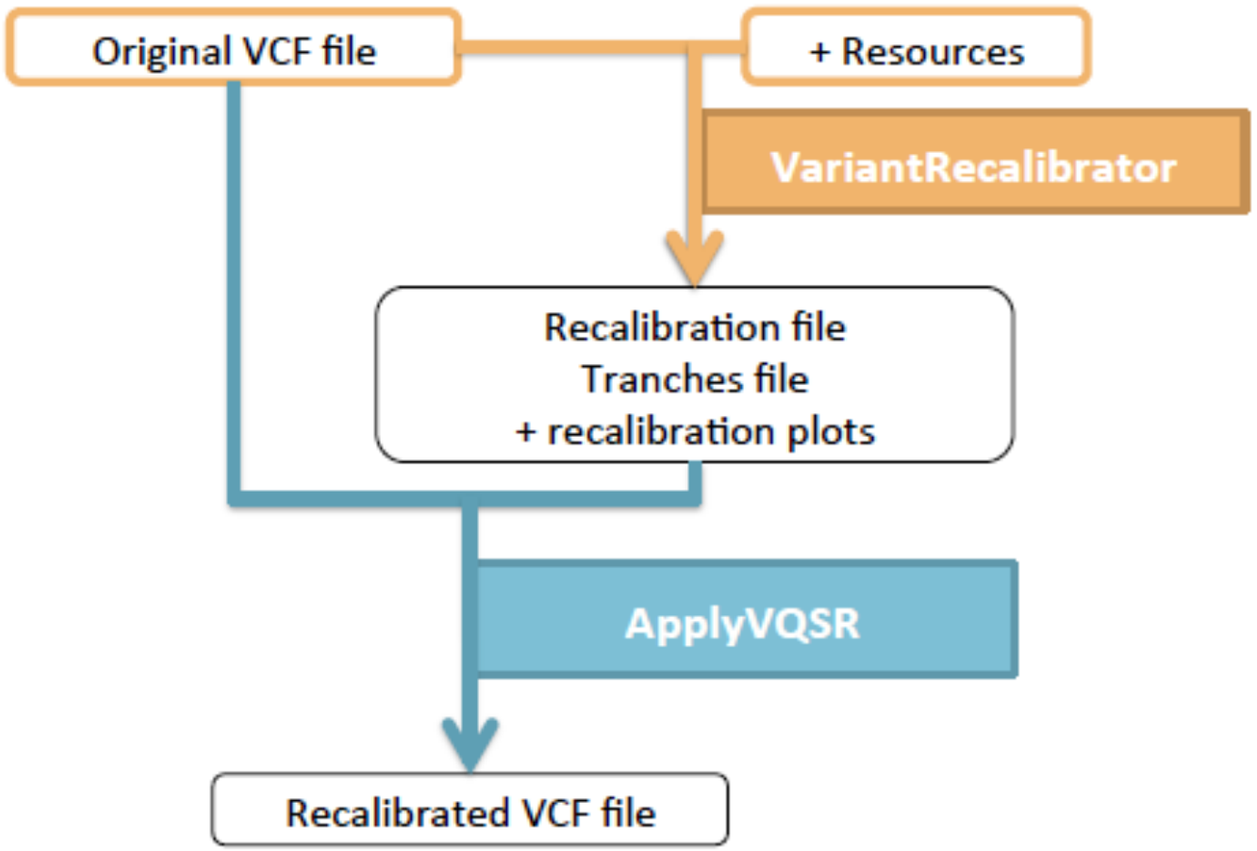
Build Gaussian mixture models from annotations of known variants from the dataset

Use this variation to score all variants



Variant Recalibration

Part of the GATK pipeline



But, where do we get the truth set?

For some organisms these exist (humans, Drosophila, Arabidopsis etc.)

SNP chip data, or other existing datasets

Alternatively, use stringent hard filters and use SNPs that pass as the truth set

Why filter data?

By default, GATK is very permissive
(It will output false positive sites!)

Two approaches to filtering:

- Hard filtering
- Variant recalibration

Intersection of outputs from different SNP calling programs

What kinds of analyses are you planning?

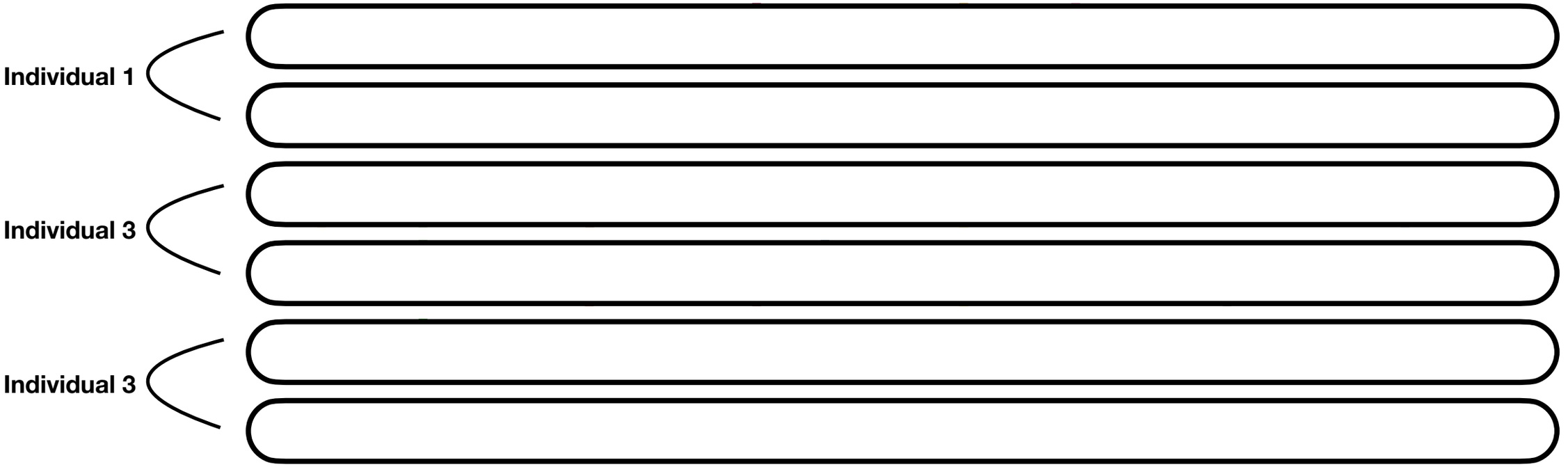
What is a genome scan

A statistical test applied to an entire genome’s worth of data

What is a genome scan

A statistical test applied to an entire genome's worth of data

Sample genomes from a population (or populations) of interest

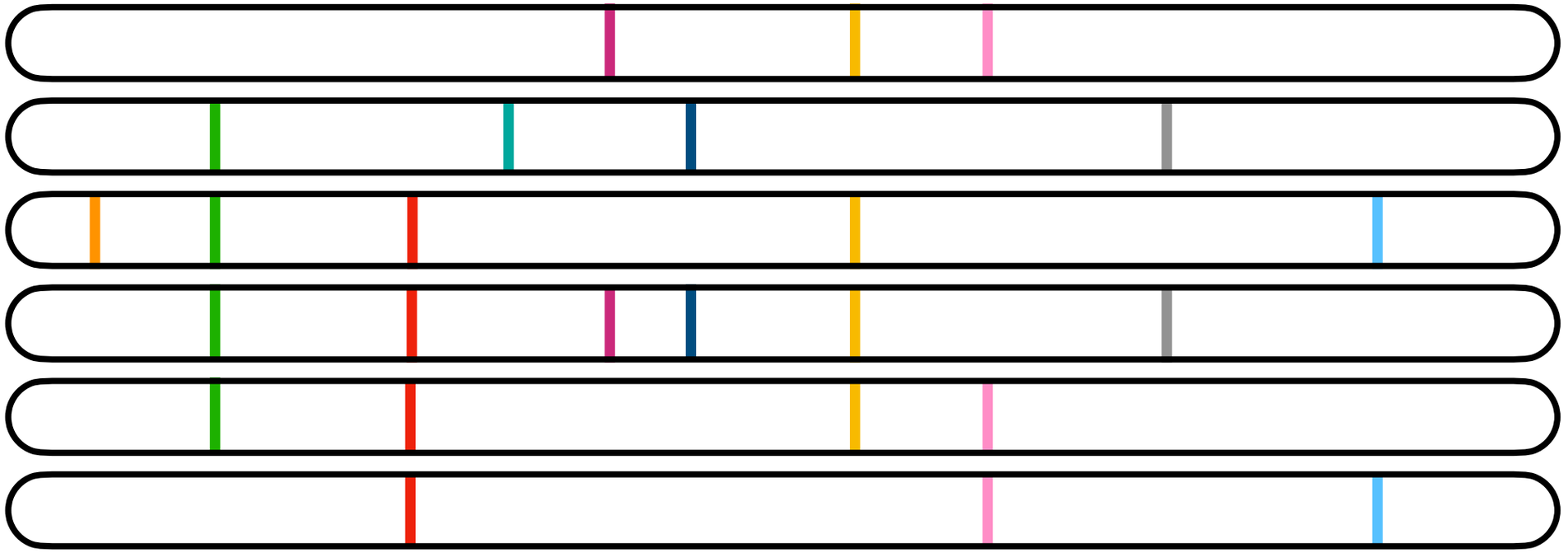


A cartoon of a chromosome

What is a genome scan

A statistical test applied to an entire genome's worth of data

Genotype polymorphisms across the genome

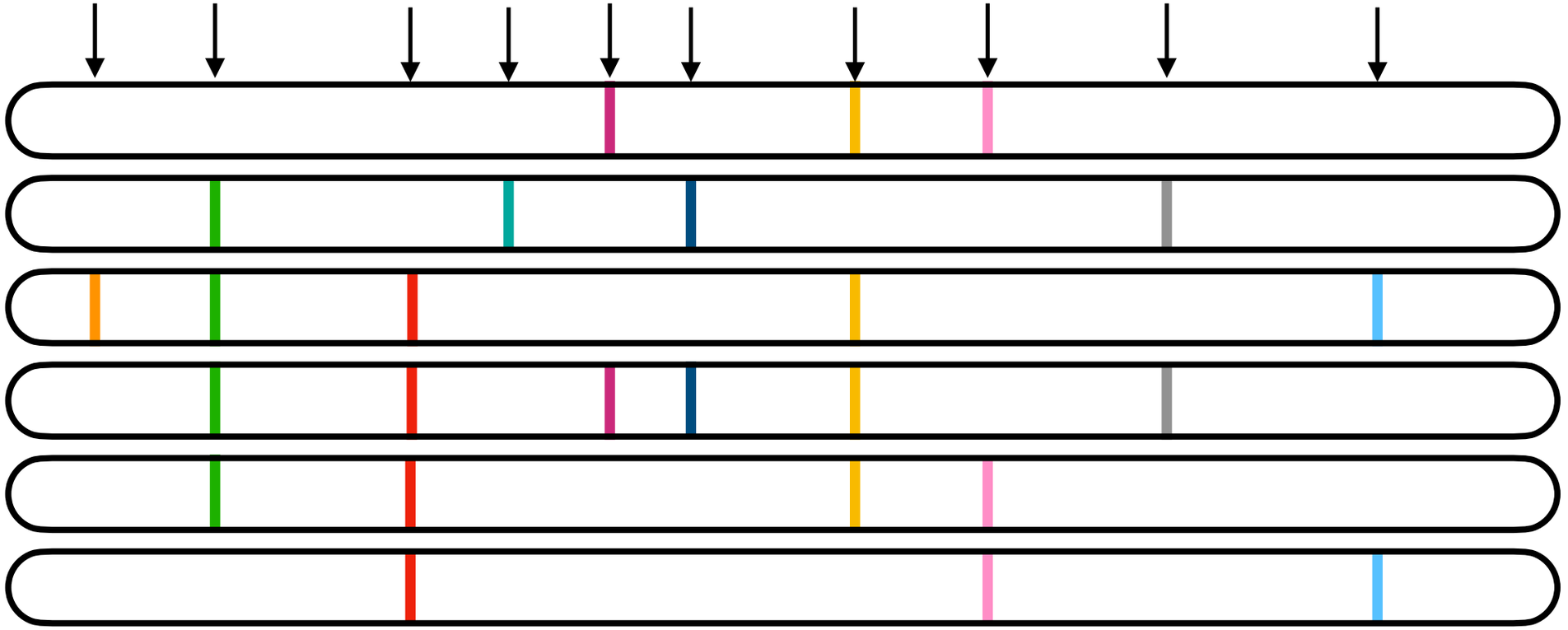


A cartoon of a chromosome

What is a genome scan

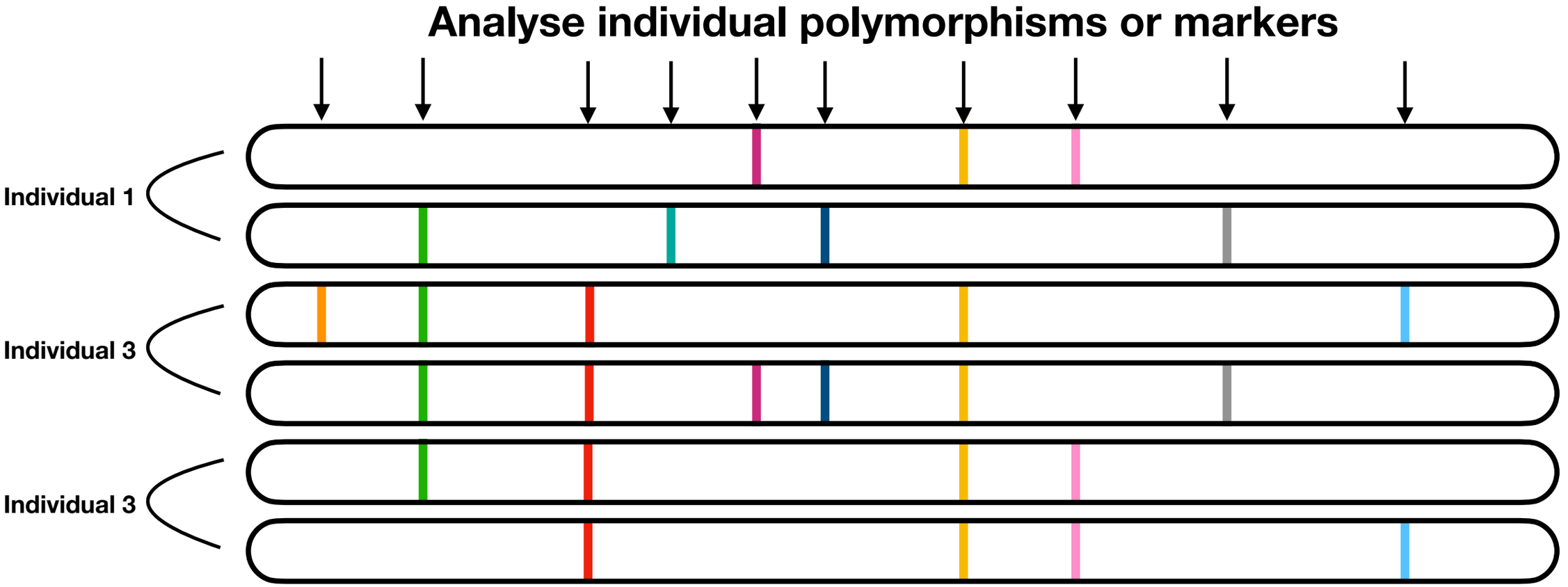
A statistical test applied to an entire genome's worth of data

Analyse individual polymorphisms or markers



What is a genome scan

A statistical test applied to an entire genome's worth of data

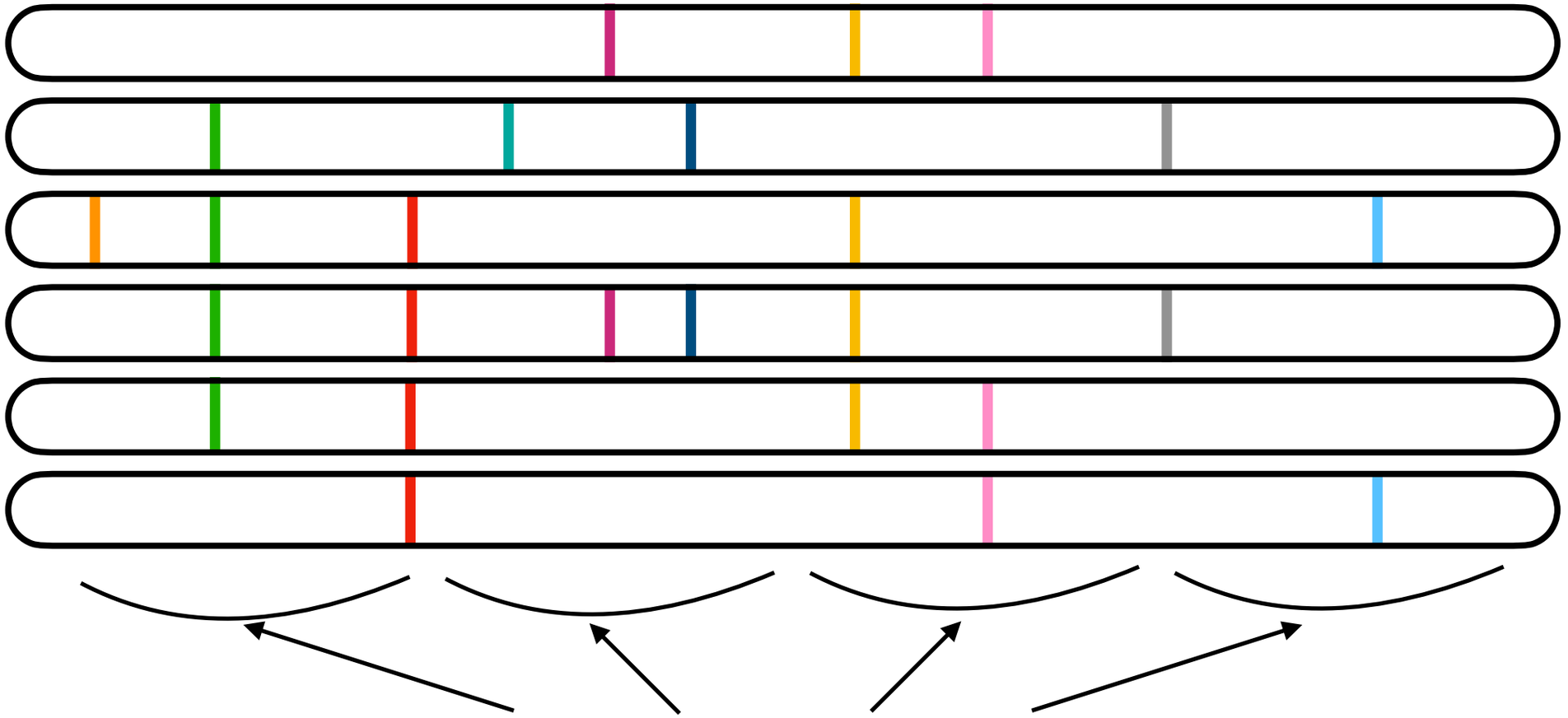


Allele Frequency	1/6	4/6	4/6	1/6	2/6	2/6	5/6	3/6	2/6	2/6
Hobserved	1/3	2/3	2/3	1/3	2/3	2/3	1/3	1/3	2/3	2/3

For example, look for excess heterozygosity

What is a genome scan

A statistical test applied to an entire genome's worth of data

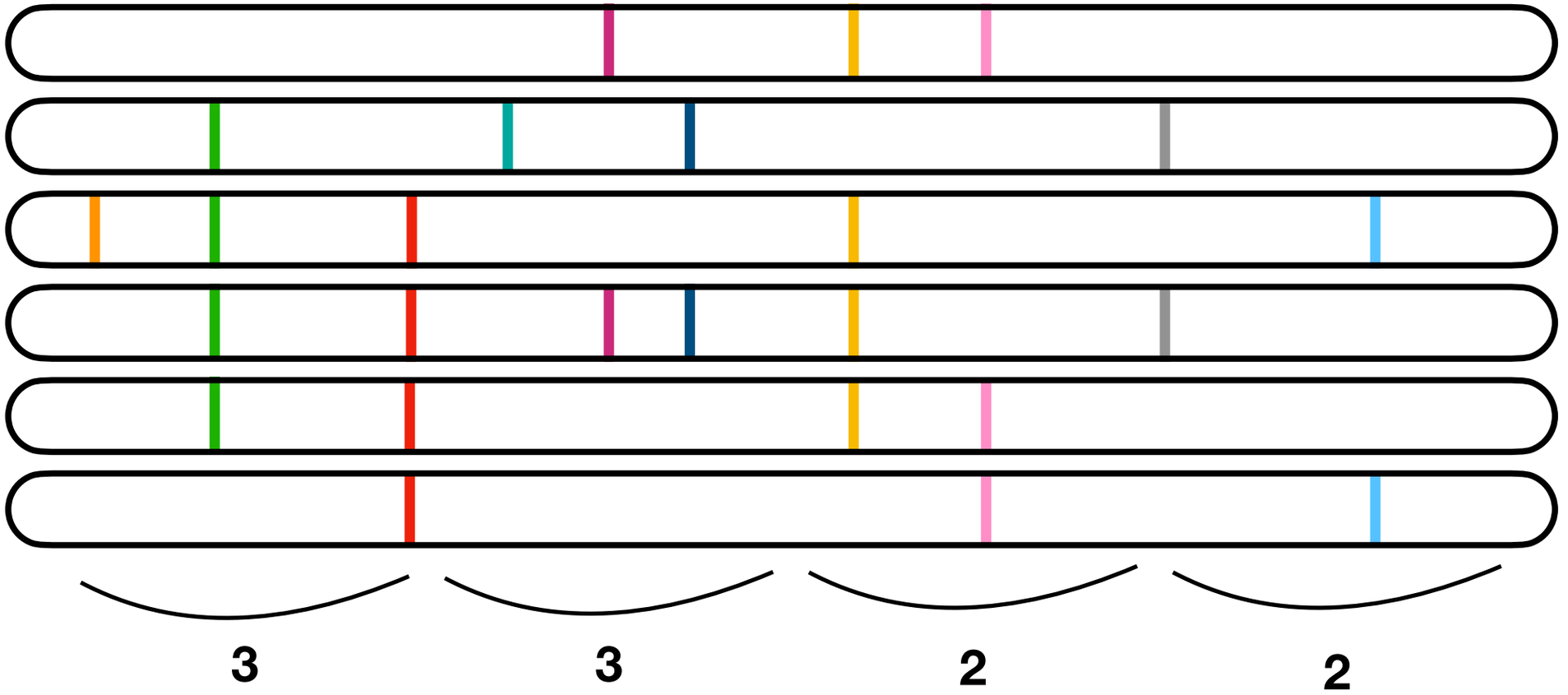


…or groups of markers

These are called analysis windows

What is a genome scan

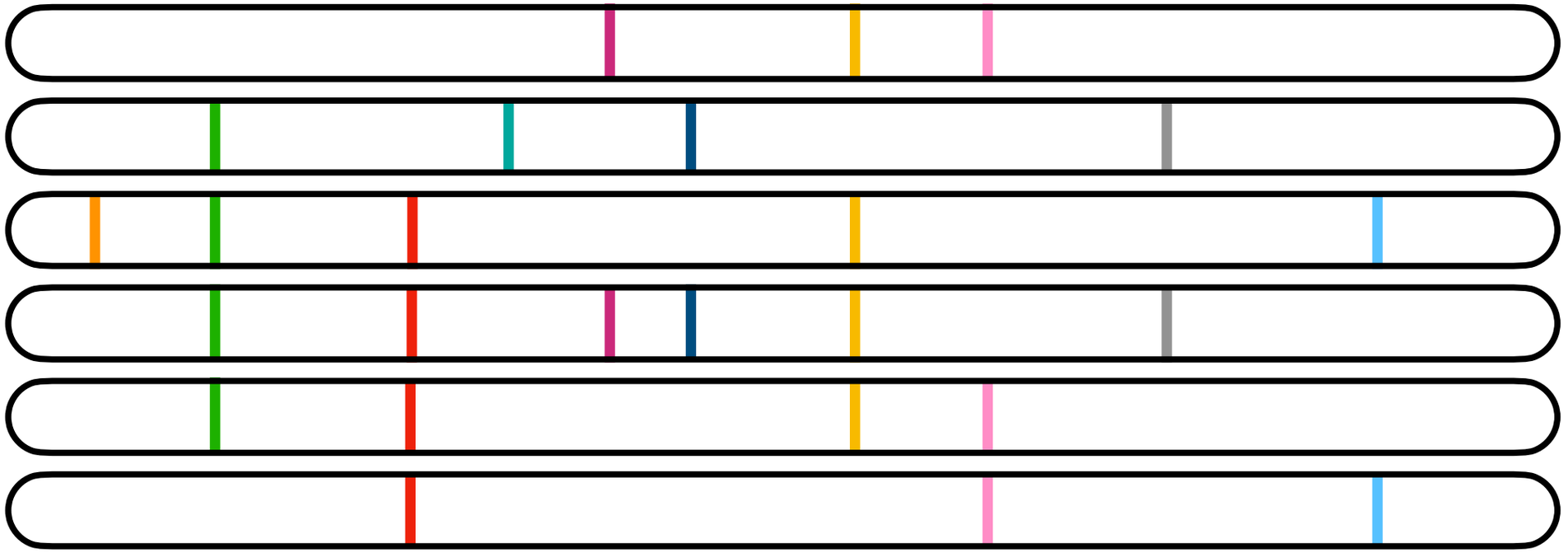
A statistical test applied to an entire genome's worth of data



The number of segregating sites in a window (proportional to the local effective population size)

What is a genome scan

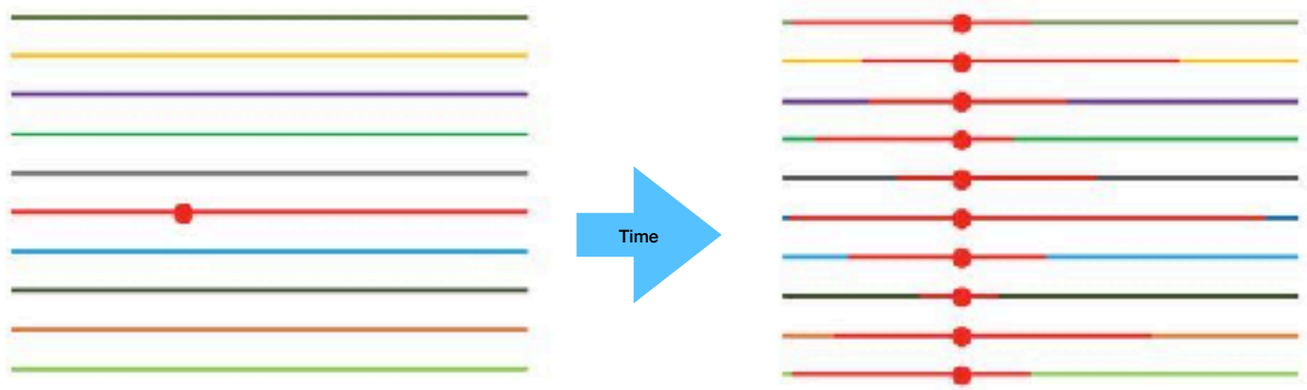
Examining the distribution of genetic variation across the genome can lead to insights into evolutionary processes



What is a genome scan

Genome scans may look for evidence of selective sweeps

E.g. hard selective sweeps

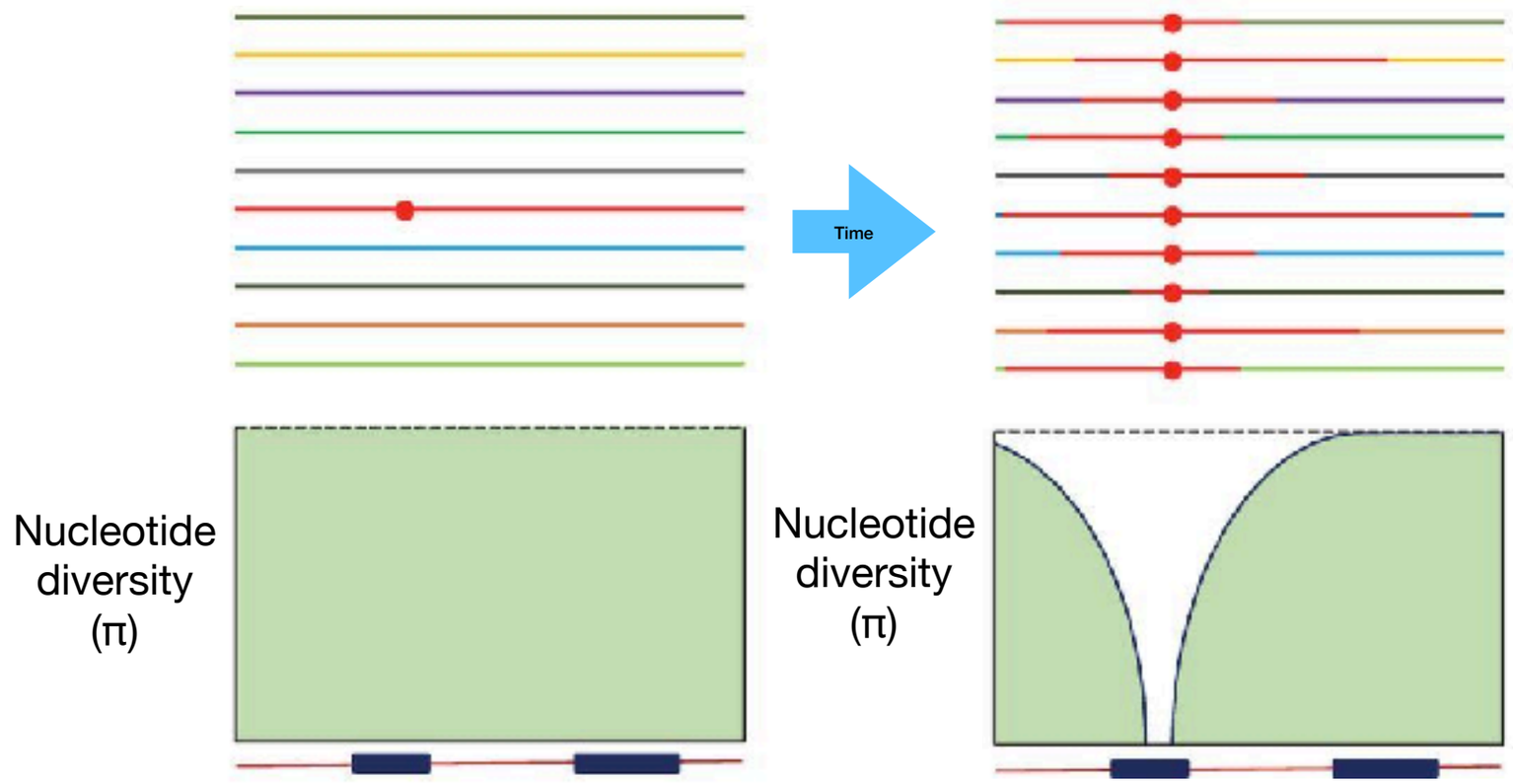


Modified from Comeron 2017 - Proc B.

What is a genome scan

Genome scans may look for evidence of selective sweeps

E.g. hard selective sweeps

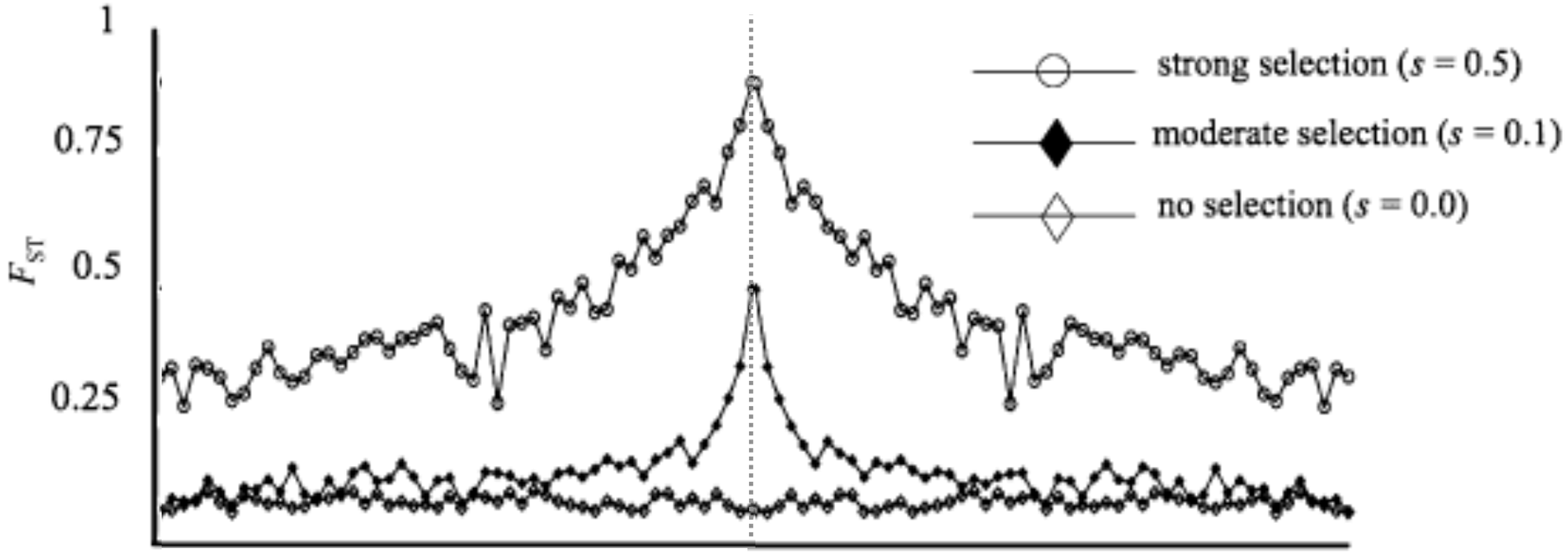


Modified from Comeron 2017 - Proc B.

What is a genome scan

Genome scans may look for evidence of divergent selection

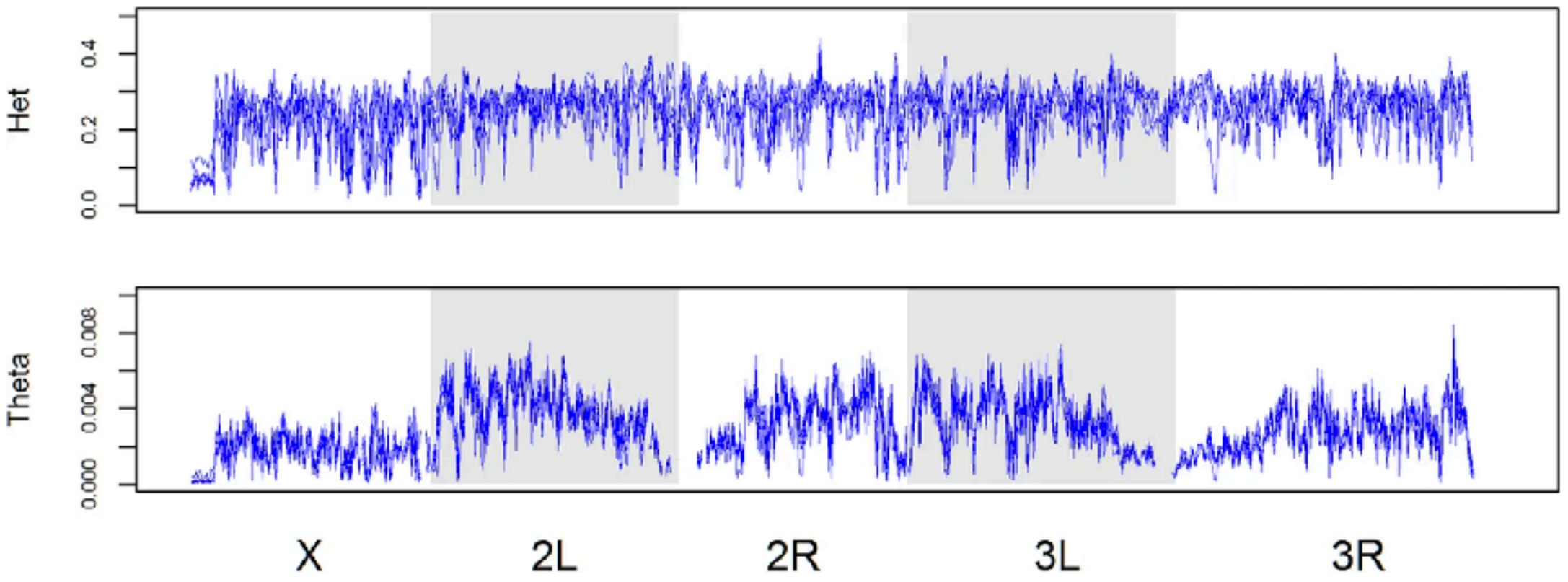
E.g. local adaptation



Modified from Nosil et al 2009 - Molecular Ecology

What is a genome scan

The distribution of heterozygosity and the number of segregating sites across the *D. melanogaster* genome



100,000bp Windows

Phillips et al 2016 - Scientific Reports

What is a genome scan

There are LOTS of population genetic summary statistics

Table 1 The arsenal of parameters for population genetics/genomics analysis: measures of nucleotide diversity, LD, and tests of selection

Measure/test	Description	Reference
Nucleotide diversity measures (un-dimensional measures)		
S, s	Number of segregating sites (per DNA sequence or per site, respectively)	Nei (1987)
H, η	Minimum number of mutations (per DNA sequence or per site, respectively)	Tajima (1996)
k	Average number of nucleotide differences (per DNA sequence) between any two sequences	Tajima (1983)
π	Nucleotide diversity: average number of nucleotide differences per site between any two sequences	Jukes and Cantor (1969); Nei and Gojobori (1986); Nei (1987)
θ, θ_{π}	Nucleotide polymorphism: proportion of nucleotide sites that are expected to be polymorphic in any suitable sample	Watterson (1975); Tajima (1983, 1996)
SFS	Site-allele frequency spectrum: distribution of allele frequencies at a given set of loci in a population or sample	Ronen et al. (2013)
LD (multi-dimensional association among variable sites) and recombination		
D	Coefficient of LD whose range depends of the allele frequencies	Lewontin and Kojima (1960)
D'	Normalized D, independent of allele frequencies	Lewontin (1964)
r, r^2	Statistical correlation between pairs of sites	Hill and Robertson (1968)
Z_{nS}	Average of r^2 over all pairwise comparisons	Kelly (1997)
Z_A/Z_Z	Z_A is the average of r^2 only between adjacent polymorphic sites. Z_Z is Z_A minus Z_{nS}, which is an estimate of the recombination parameter γ	Rozas et al. (2001)
Four-gamete test	Measure of historical recombination under the infinite-sites model	Hudson and Kaplan (1985)
ρ	Population-scaled recombination rate $\rho = 4N_e r$ [computed, e.g., by LDhat (Auton and McVean 2007) and LDhelmet (Chan et al. 2012)]	Hudson (1987)
Selection tests based on the allele frequency spectrum and/or levels of variability		
Tajima's D	Number of nucleotide polymorphisms with the mean pairwise difference between sequences	Tajima (1989)
Fu and Li's D, D^*	Number of derived nucleotide variants observed only once in a sample with the total number of derived nucleotide variants	Fu and Li (1993)
Fu and Li's F, F^*	Number of derived nucleotide variants observed only once in a sample with the mean pairwise difference between sequences	Fu and Li (1993)
Fay and Wu's H	Number of derived nucleotide variants at low and high frequencies with the number of variants at intermediate frequencies	Fay and Wu (2000)
Zeng's E, θ_L, DH	Difference between θ_L and θ_W the first is sensitive to changes in high-frequency variants. DH is a joint test including Tajima's D and Fay and Wu's H	Zeng et al. (2006)
Achaz's Y	Unified framework for θ estimators on the basis of the allele frequency spectrum	Achaz (2009)
Fu's F_S	Test based on the allele frequency spectrum	Fu (1997)
Ramos-Onsins' and Rozas' R_2, R_3, R_4, R_{2E}, R_{3E}, R_{4E}	Tests based on the difference between the number of singleton mutations and the average number of nucleotide differences	Ramos-Onsins and Rozas (2002)
CL, CLR	Genome scan for candidate regions of selective sweeps based on aberrant allele frequency spectrum	Nielsen et al. (2005)
Selection tests based on comparisons of polymorphism and/or divergence between different classes of mutation		
d_N/d_S, K_a/K_s	Ratio of nonsynonymous to synonymous nucleotide divergence/polymorphism (ω)	Li et al. (1985); Nei and Gojobori (1986)
HKA	Degree of polymorphism within and between species at two or more loci	Hudson et al. (1987)
MK	Ratios of synonymous and nonsynonymous nucleotide divergence and polymorphism	McDonald and Kreitman (1991)
Estimators derived from extensions of the MK test or the DFE		
NI	Neutrality index that summarizes the four values in an MK test table as a ratio of ratios	Rand and Kann (1996)
DoS	Direction of selection: difference between the proportion of nonsynonymous divergence and nonsynonymous polymorphism	Stoletzki and Eyre-Walker (2011)

(continued)

Table 1, continued

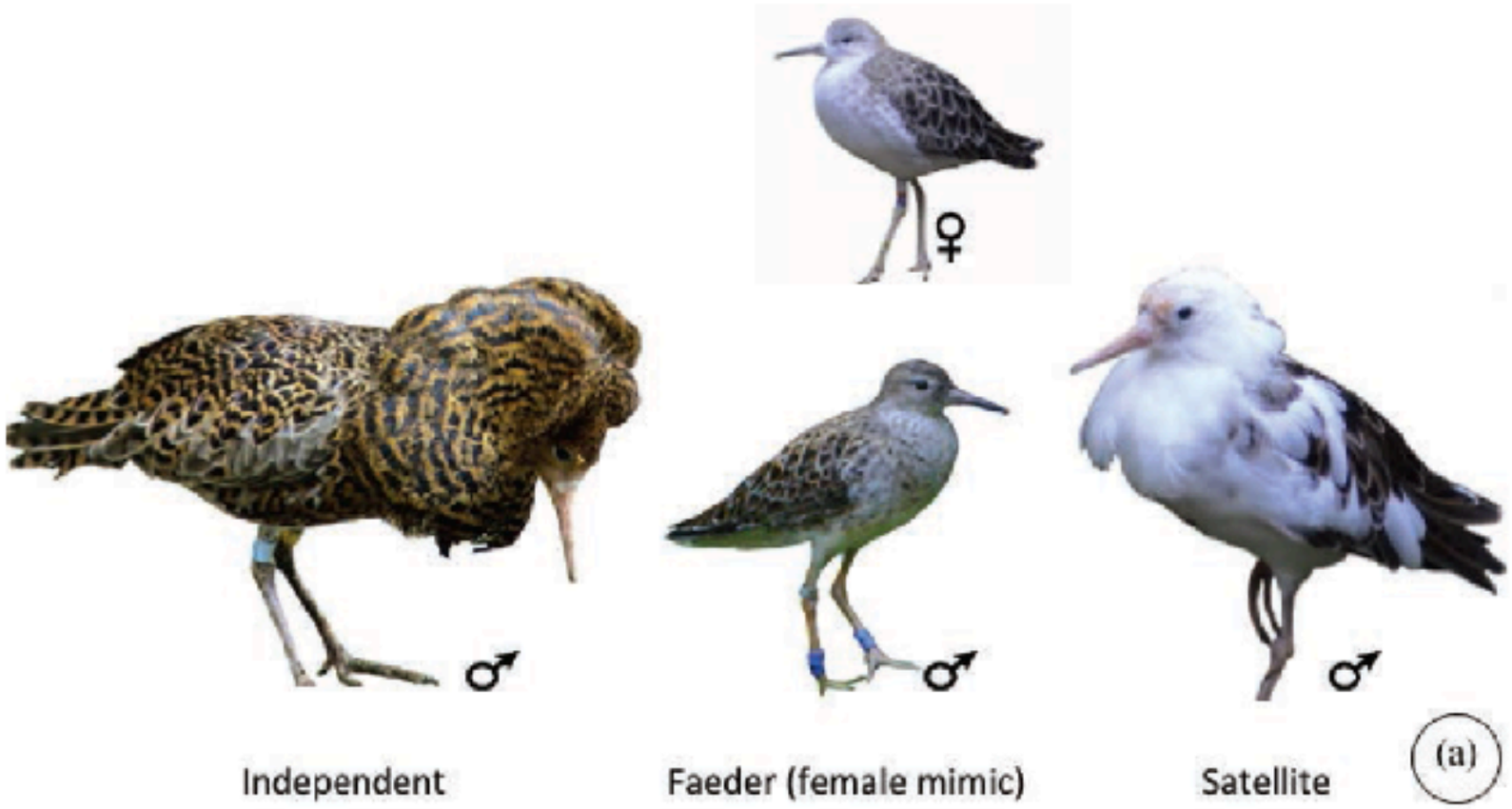
Measure/test	Description	Reference
α	Proportion of substitutions that are adaptive	Charlesworth (1994); Smith and Eyre-Walker (2002)
DFE-α	Fraction of adaptive nonsynonymous substitutions, robust to low recombination	Eyre-Walker and Keightley (2009)
ω_A	Rate of adaptive evolution relative to the mutation rate	Gossmann et al. (2010)
K_{aa}	Rate of adaptive amino acid substitution ($K_{aa} = \omega_A \theta_a$)	Gossmann et al. (2010); Castellano et al. (2016)
$\hat{a}$, $\hat{b}$, $\hat{f}$, $\hat{\gamma}$, $\hat{\alpha}$	Fractions of five different selection regimes derived from an extension of the MK test: $\hat{a}$, fraction of new mutations that are strongly deleterious and do not segregate in the population; $\hat{b}$, fraction of new mutations that are slightly deleterious and segregate at minor allele frequency (MAF) <5%; $\hat{f}_0$, fraction of new mutations that are neutral, calculated after removing the excess of sites at MAF <5% due to slightly deleterious mutations; $\hat{\gamma}$, subset of $\hat{f}_0$ corresponding to recently neutral sites; $\hat{\alpha}$, fraction of new mutations that are adaptive, calculated after removing slightly deleterious mutations	Macay et al. (2012)
α_{MK}	Proportion of adaptive substitutions lost due to HRI	Castellano et al. (2016)
L_{min}	Optimal baseline recombination, above which the genome is free of the HRI and thus $L_{min} = 0$	Macay et al. (2012); Castellano et al. (2016)
Selection tests based on LD		
Hudson's haplotype test	Detection of derived and ancestral alleles on unusually long haplotypes	Hudson et al. (1994)
BVQ	Based on LD between adjacent pairs of segregating sites, under the coalescent model with recombination	Wall (1999)
iHS	Integrated haplotype score, based on the frequency of alleles in regions of high LD	Voight et al. (2006)
LRH	Long-range haplotype test, based on the frequency of alleles in regions of long-range LD	Sabeti et al. (2002)
HS	Haplosimilarity score: long-range haplotype similarity	Hanchard et al. (2006)
EHH	Extended haplotype homozygosity: measurement of the decay of LD between loci with distance	Sabeti et al. (2002)
LDD	LD decay: expected decay of adjacent SNP LD at recently selected alleles	Wang et al. (2006)
SGS	Shared genomic segment analysis: detection of shared regions across individuals within populations	Cai et al. (2011)
GIBDLD	Detection of genomic loci with excess of identity-by-descent sharing in unrelated individuals as signature of recent selection	Han and Abney (2013)
XP-EHH	Long-range haplotype method to detect recent selective sweeps	Sabeti et al. (2007)
H12, H2/H1	Haplotype homozygosity	Garud et al. (2015)
Population differentiation and associated selection tests		
Gst	Analysis of gene diversity (heterozygosity) within and between subpopulations	Nei (1975)
F_{ST}	Average levels of gene flow based on allele frequencies, under the infinite-sites model	Hudson et al. (1992b)
Bayesian F_{ST}	Probability that a locus is subject to selection based on locus-specific population differentiation, using a Bayesian method	Foll and Gaggiotti (2008)
Gst, Hst, Kst	Different test statistics based on haplotype frequencies and/or the number of nucleotide differences between sequences	Hudson et al. (1992a)
S_{nn}	Genetic differentiation of subpopulations based on haplotypic data	Hudson (2000)
Rho_{ST}	Correlation of haplotypic diversity at different levels of hierarchical subdivision	Excoffier et al. (1992)
Strobeck's S	Measure of population structure based on the comparison of the observed number of alleles in a sample to that expected when this is estimated from the average number of nucleotide differences	Strobeck (1987)
XP-CLR	Cross-population composite likelihood ratio test, based on allele frequency differentiation across populations	Chen et al. (2010)
TLK, TF-LK	Original Lewontin–Krakauer test (TLK) and an extension (TF-LK), aimed at detecting selection based on the variance of F_{ST} across loci	Lewontin and Krakauer (1973); Bonhomme et al. (2010)
LSBL	Locus-specific branch length, based on pairwise F_{ST} distances	Shriver et al. (2004)
HapFLK	Detecting of selection based on differences in haplotype frequencies among populations with a hierarchical structure	Fariello et al. (2013)

Not an exhaustive list!

Table 1 from Casillas and Barbadilla 2017 Genetics

What is a genome scan

Ruffs have a very interesting mating system



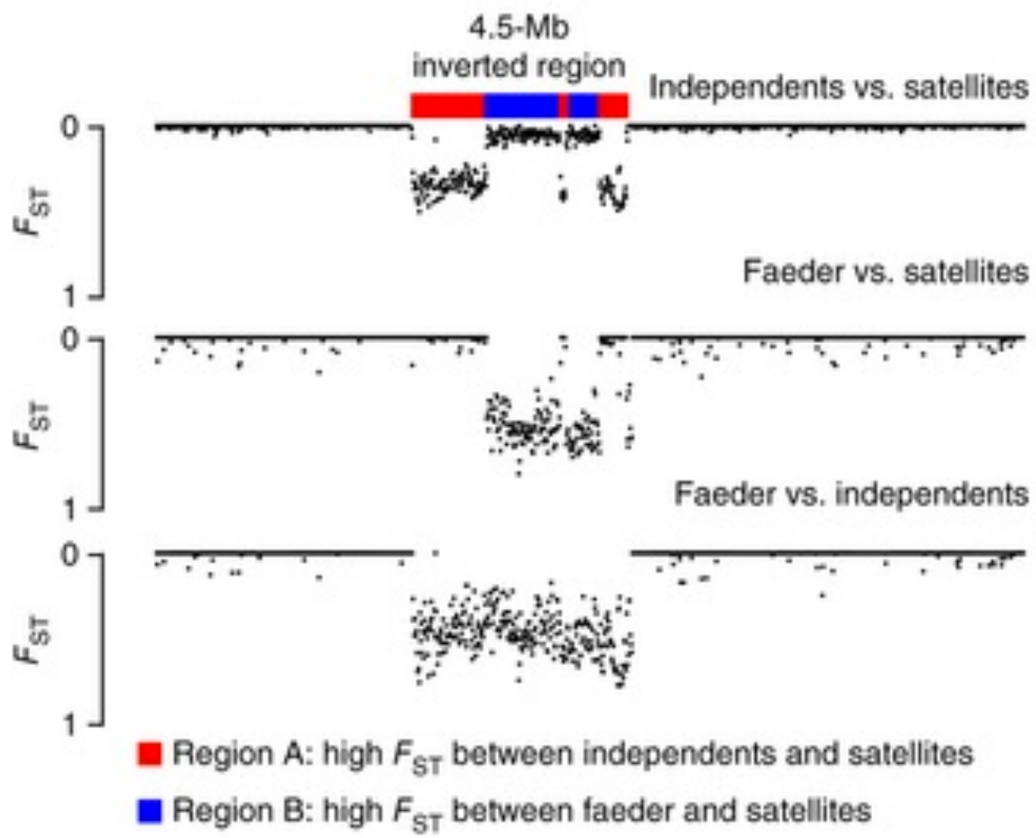
https://www.molecularecologist.com/2016/03/11/chromosomal-inversion-determines-male-morphs-in-the-ruff/

What is a genome scan

Male plumage is controlled by a “supergene” maintained by an inversion

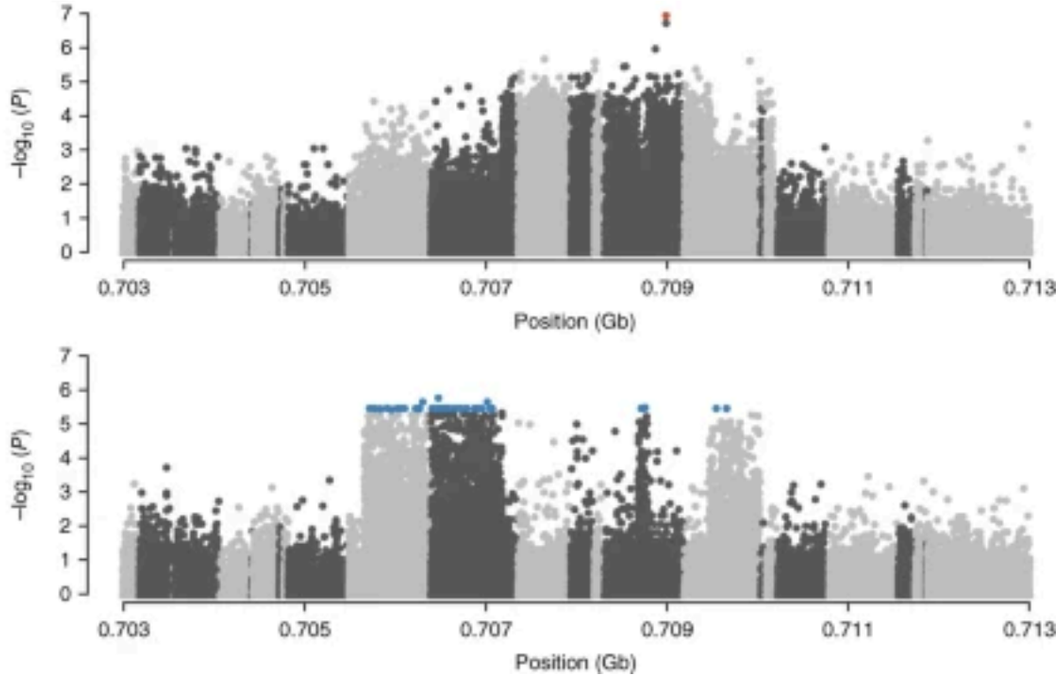
The inversion is under negative frequency dependant selection

***F_{ST}* scan**



Lamichhaney et al 2016 - Nature Genetics

GWAS on plumage



Küpper et al 2016 - Nature Genetics

https://www.molecularecologist.com/2016/03/11/chromosomal-inversion-determines-male-morphs-in-the-ruff/

What is a genome scan

Applying genome scans to understand the genetics of male plumage in the ruff led to insights into the genetic and evolutionary mechanisms that maintain complex traits

Küpper et al 2016 - Nature Genetics

Lamichhaney et al 2016 - Nature Genetics

https://www.newscientist.com/article/dn28493-ruff-bird-orgies-have-four-sexes-thanks-to-a-supergene-flip/

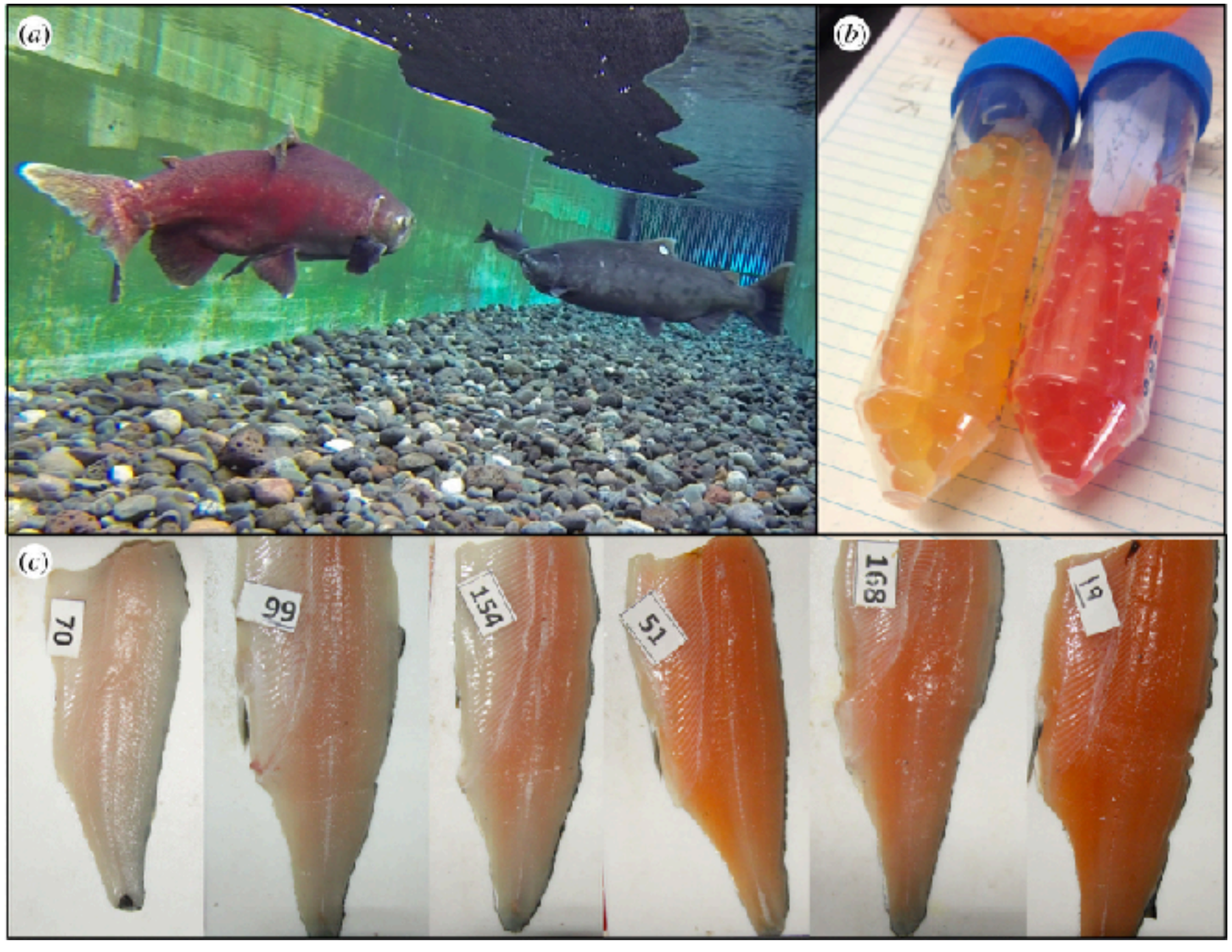
Assignment:

The genetic basis of a flesh colour polymorphism in Chinook Salmon

Chinook

Fish pic from Wikipedia

Chinook: Flesh colour polymorphism



Lehnert et al 2019 - *Proc B.*

Chinook: Flesh colour polymorphism

TABLE 1. Proportions of white-fleshed chinook salmon recovered in Southeast Alaska commercial and sport fisheries from 21 localities in western North America. The localities are arranged from north (1) to south (21); their geographic locations are shown in Fig. 1. Localities with the same letter preceeding them belong to the same maximal acceptable subset. White-fleshed scores are relative (+, −) to the overall mean.

Locality			White-fleshed chinook salmon recovered			
No.		Name	No.	%	Score	Total
1	A	Chilkat River, AK	174	37.1	+	468
2	B	Taku River, AK	183	12.0	+	1528[a]
3	C	Stikine River, AK	23	2.6	−	893
4	D	Unuk River, AK	505	16.3	+	3094
5	D	Chickamin River, AK	29	17.4	+	167
6	E	Upper Skeena River, B.C.	0	0.0	−	61
7	F	Lower Skeena River, B.C.	134	41.2	+	325
8	E	Bella Coola River, B.C.	6	8.1	−	74
9	E	Upper Fraser River, B.C.	1	4.2	−	24
10	E,H	Southern Coastal B.C.	1	4.2	−	24
11	G	Lower Fraser River, B.C.	35	53.8	+	65
12	E,H	East Vancouver Island, B.C.	20	2.1	−	936
13	E	West Vancouver Island, B.C.	7	0.5	−	1509
14	H	Northwestern Washington	8	3.6	−	220
15	H	Priest Rapids, Columbia River	4	1.5	−	262
16	H	Snake River, ID	2	4.8	−	42
17	H	Lower Columbia River	2	0.8	−	245
18	H	Mid-Columbia River	3	0.6	−	494
19	H	Northern Coastal Oregon	0	0.0	−	196
20	H	Willamette River, OR	3	0.5	−	613
21	H	Southern Coastal OR	2	1.5	−	133
		Total	1142	$\bar{X} = 10.1$		11 373[b]

Hard et al 1989 - *Can. Journ. Fish. Aq. Sci*

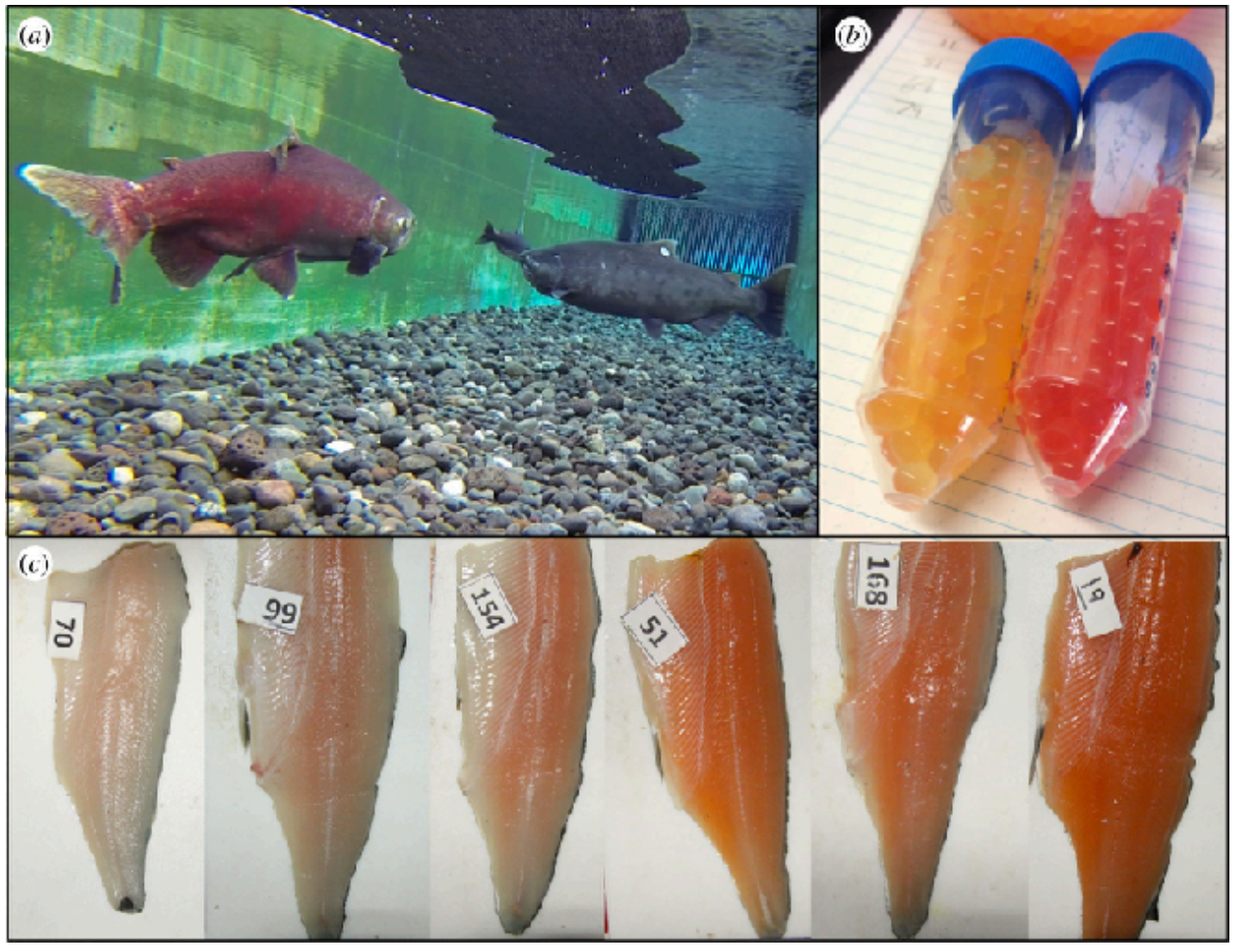
Chinook: Flesh colour polymorphism

TABLE 1. Proportions of white-fleshed chinook salmon recovered in Southeast Alaska commercial and sport fisheries from 21 localities in western North America. The localities are arranged from north (1) to south (21); their geographic locations are shown in Fig. 1. Localities with the same letter preceeding them belong to the same maximal acceptable subset. White-fleshed scores are relative (+, −) to the overall mean.

Locality			White-fleshed chinook salmon recovered			
No.		Name	No.	%	Score	Total
1	A	Chilkat River, AK	174	37.1	+	468
2	B	Taku River, AK	183	12.0	+	1528[a]
3	C	Stikine River, AK	23	2.6	−	893
4	D	Unuk River, AK	505	16.3	+	3094
5	D	Chickamin River, AK	29	17.4	+	167
6	E	Upper Skeena River, B.C.	0	0.0	−	61
7	F	Lower Skeena River, B.C.	134	41.2	+	325
8	E	Bella Coola River, B.C.	6	8.1	−	74
9	E	Upper Fraser River, B.C.	1	4.2	−	24
10	E,H	Southern Coastal B.C.	1	4.2	−	24
11	**G**	**Lower Fraser River, B.C.**	**35**	**53.8**	**+**	**65**
12	E,H	East Vancouver Island, B.C.	20	2.1	−	936
13	E	West Vancouver Island, B.C.	7	0.5	−	1509
14	H	Northwestern Washington	8	3.6	−	220
15	H	Priest Rapids, Columbia River	4	1.5	−	262
16	H	Snake River, ID	2	4.8	−	42
17	H	Lower Columbia River	2	0.8	−	245
18	H	Mid-Columbia River	3	0.6	−	494
19	H	Northern Coastal Oregon	0	0.0	−	196
20	H	Willamette River, OR	3	0.5	−	613
21	H	Southern Coastal OR	2	1.5	−	133
		Total	1142	$\bar{X} = 10.1$		11 373[b]

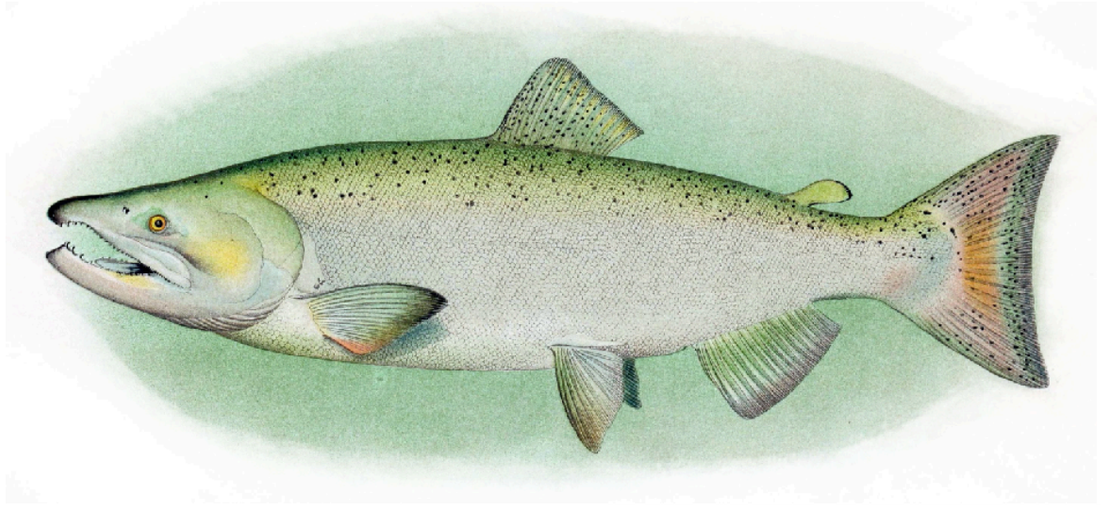
Hard et al 1989 - *Can. Journ. Fish. Aq. Sci*

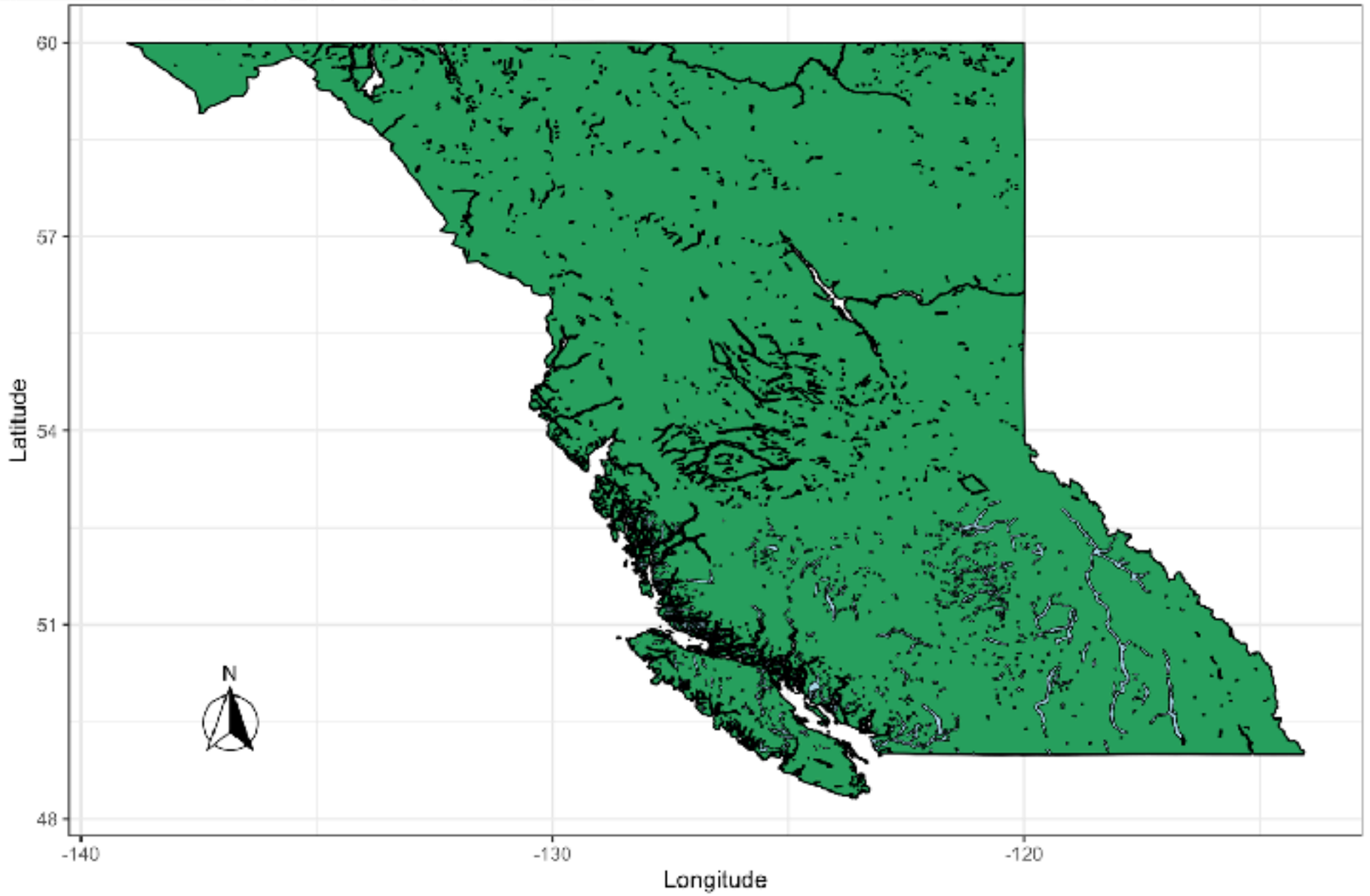
Chinook: Flesh colour polymorphism

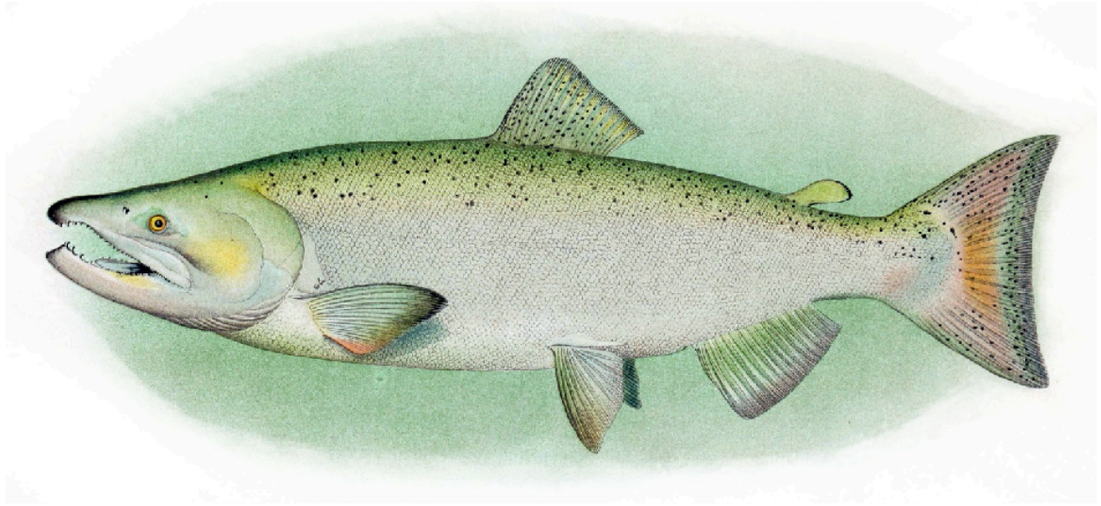


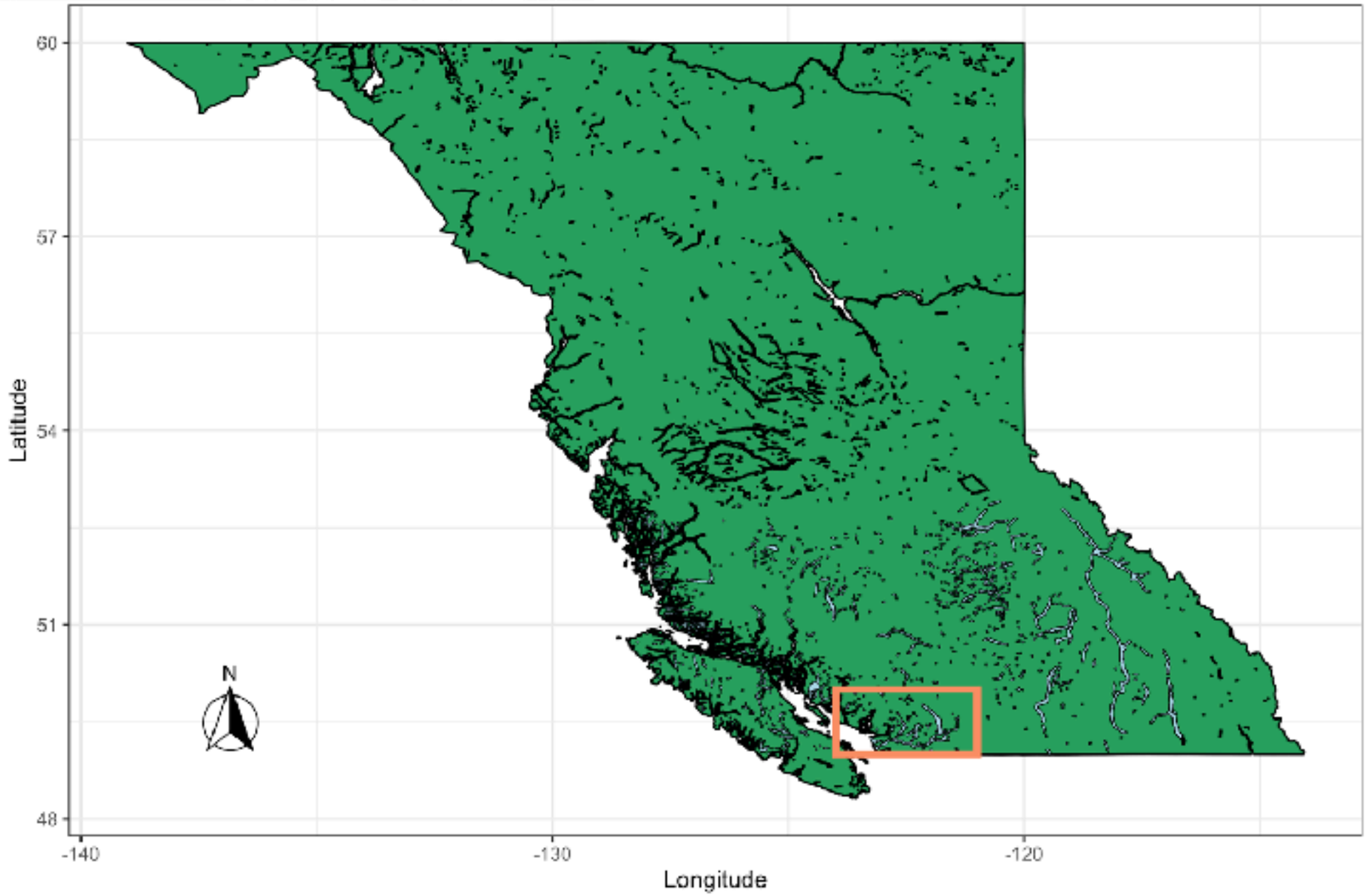
- About 5-10% of Chinook Salmon have white flesh
- This is due to differences in the metabolism of dietary caretenoids
- There is evidence that egg colour is related to predation
- Controlled crosses indicate the polymorphism has a fairly simple genetic basis

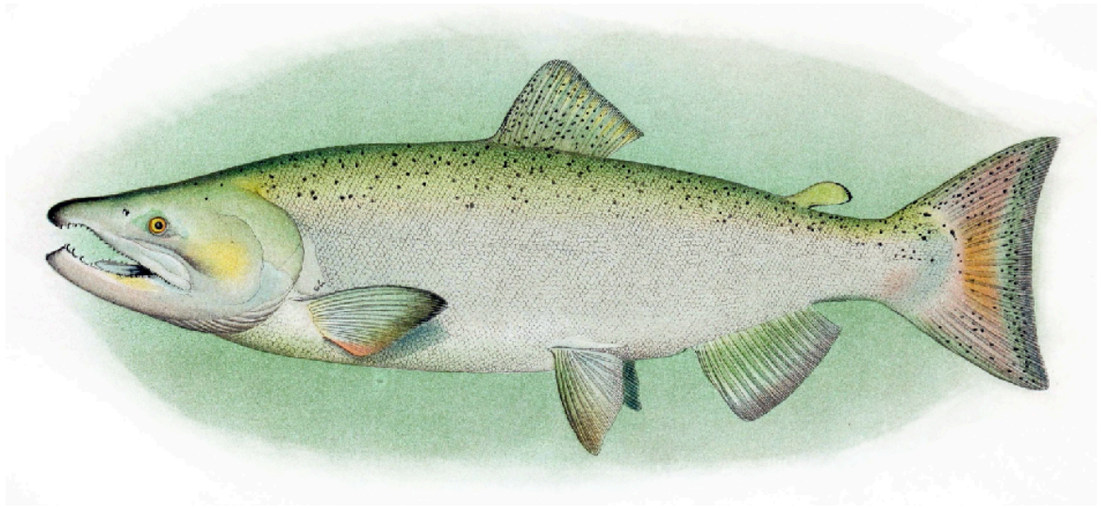
Lehnert et al 2019 - *Proc B.*

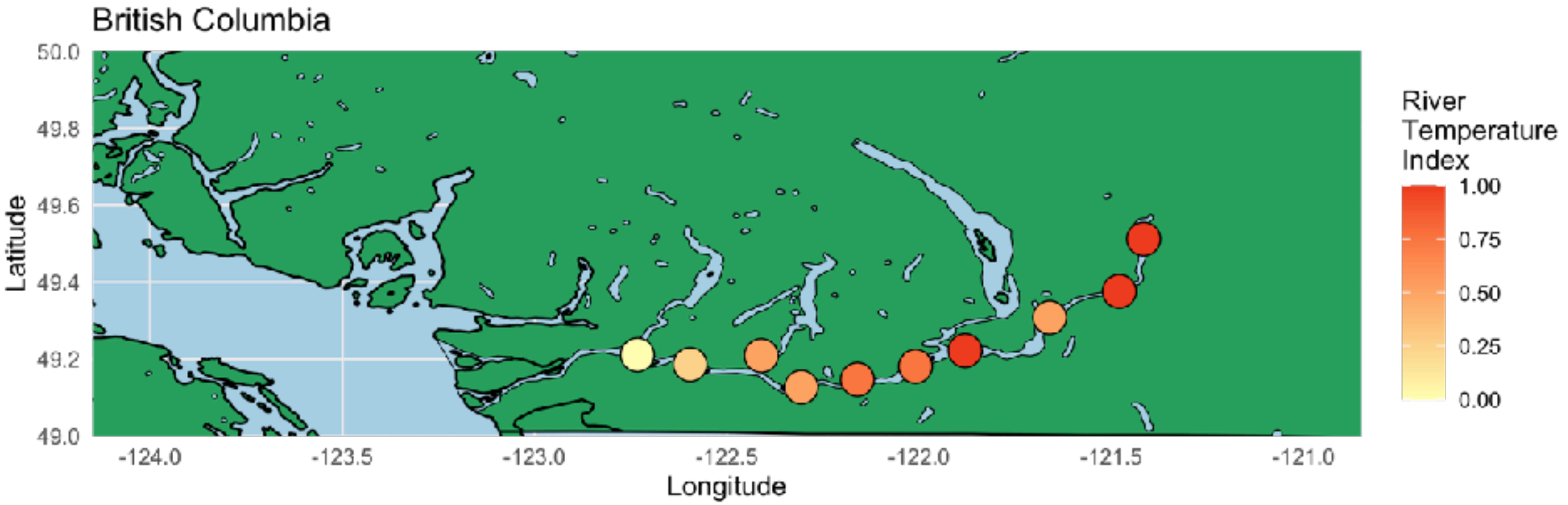



60
57
54
51
48
Latitude
-140
-130
-120
Longitude
N



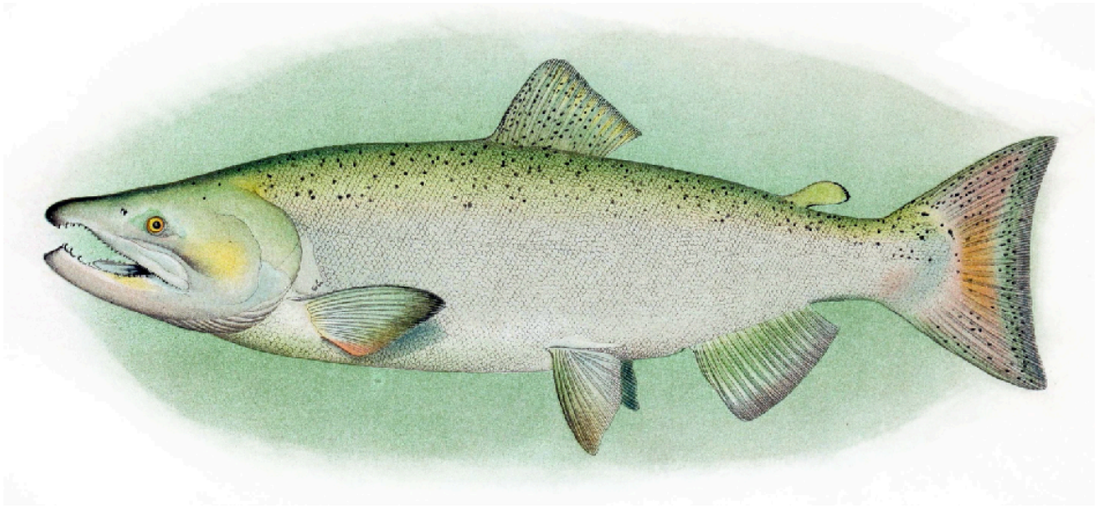

60
57
54
51
48
Latitude
-140
-130
-120
Longitude
N

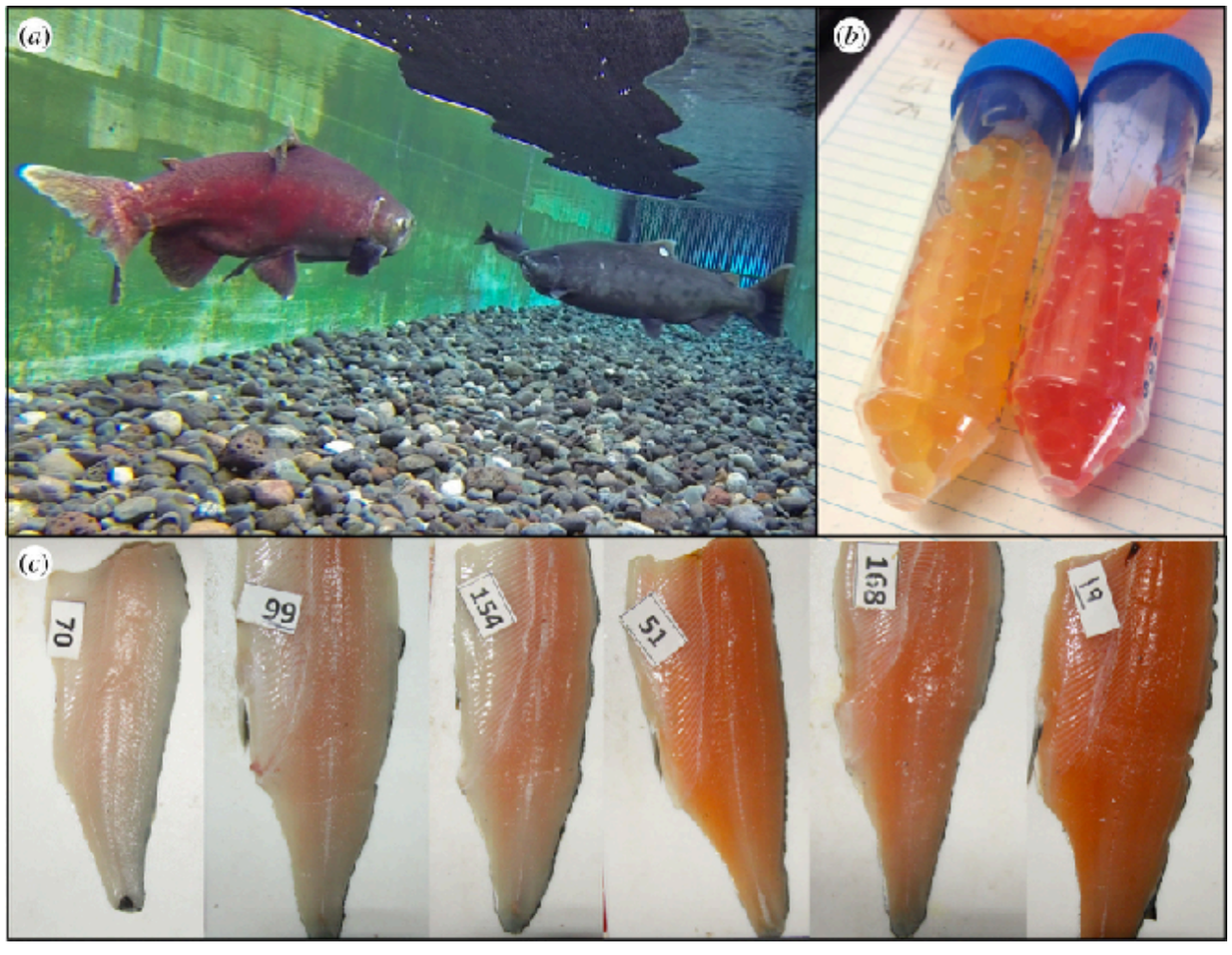




In our simulated population, white fleshed fish make up ~30% of the meta-population

What factors maintain polymorphism in a population?

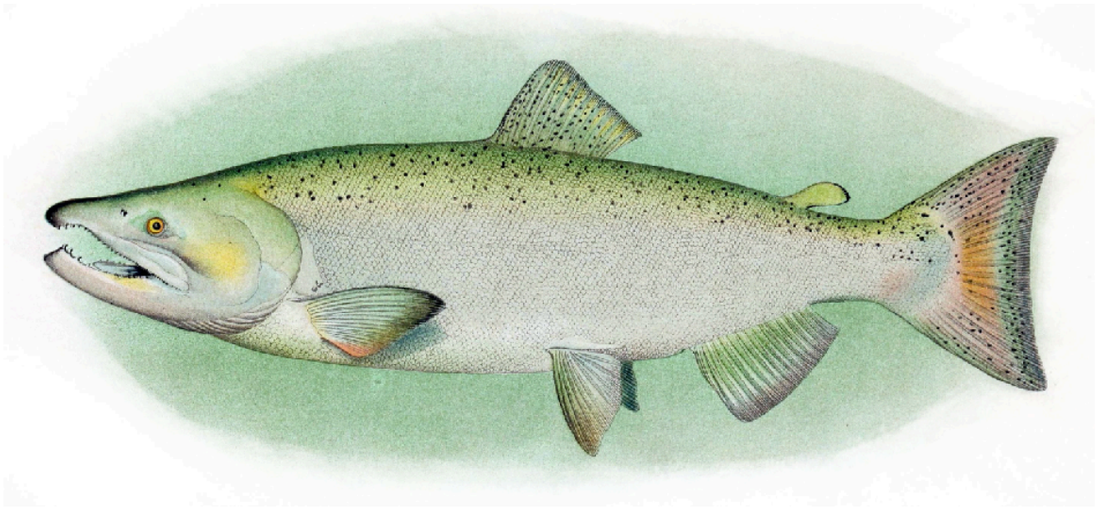


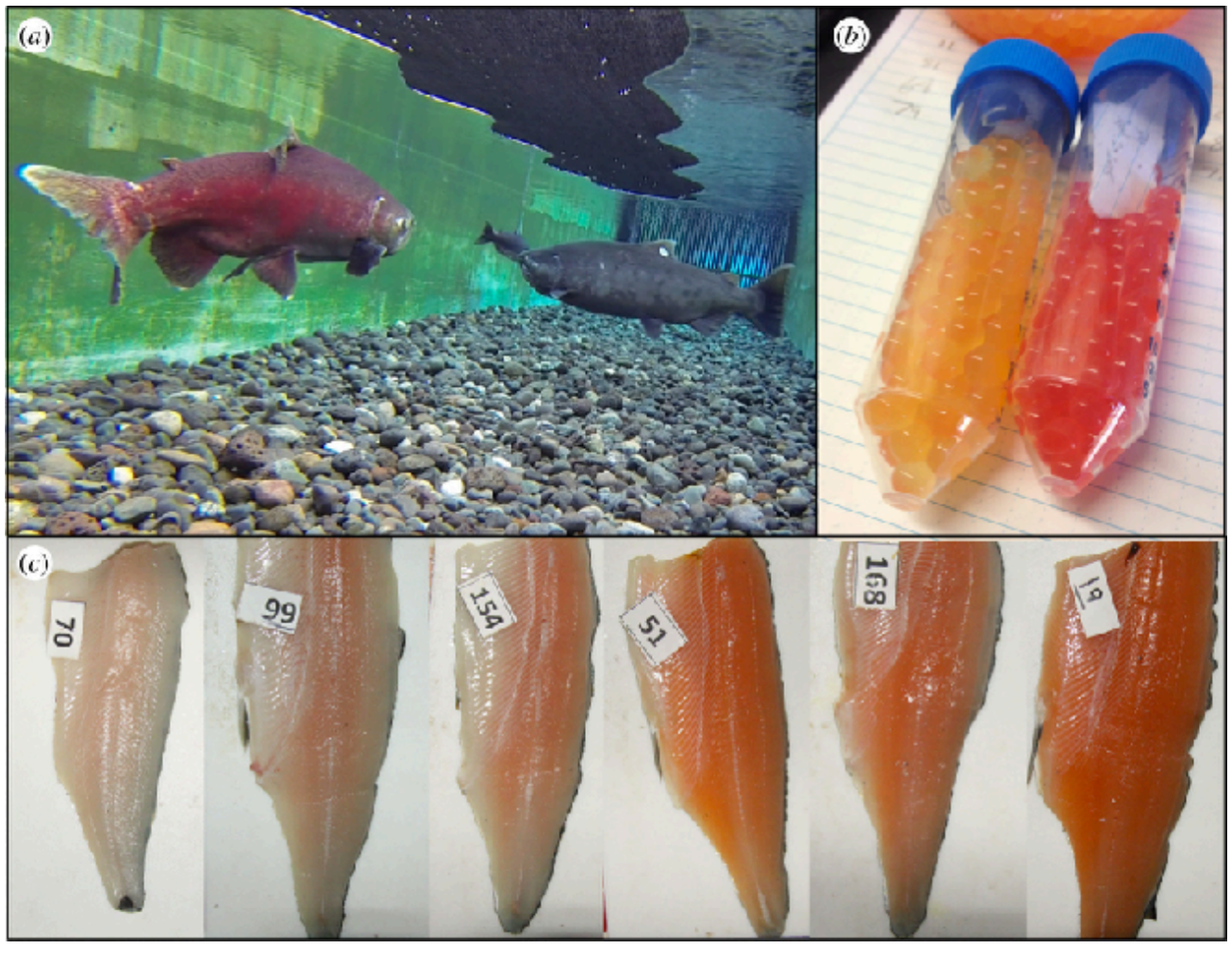


We modelled the white flesh polymorphism as an instance of heterozygote advantage

This polymorphism has been maintained in our simulation for 40,000 years

What approaches could be used to identify a long-term balanced polymorphism?





Using a genome sequences of white and red fleshed salmon, can you identify the locus that gives rise to the polymorphism?

We recommend using an Fst genome scan - but you are free to experiment!

Data

You've used these already…

- Reference Genome (**SalmonReference.fasta**)
- Gene Annotations (**SalmonAnnotations.gff**)

You also get this:

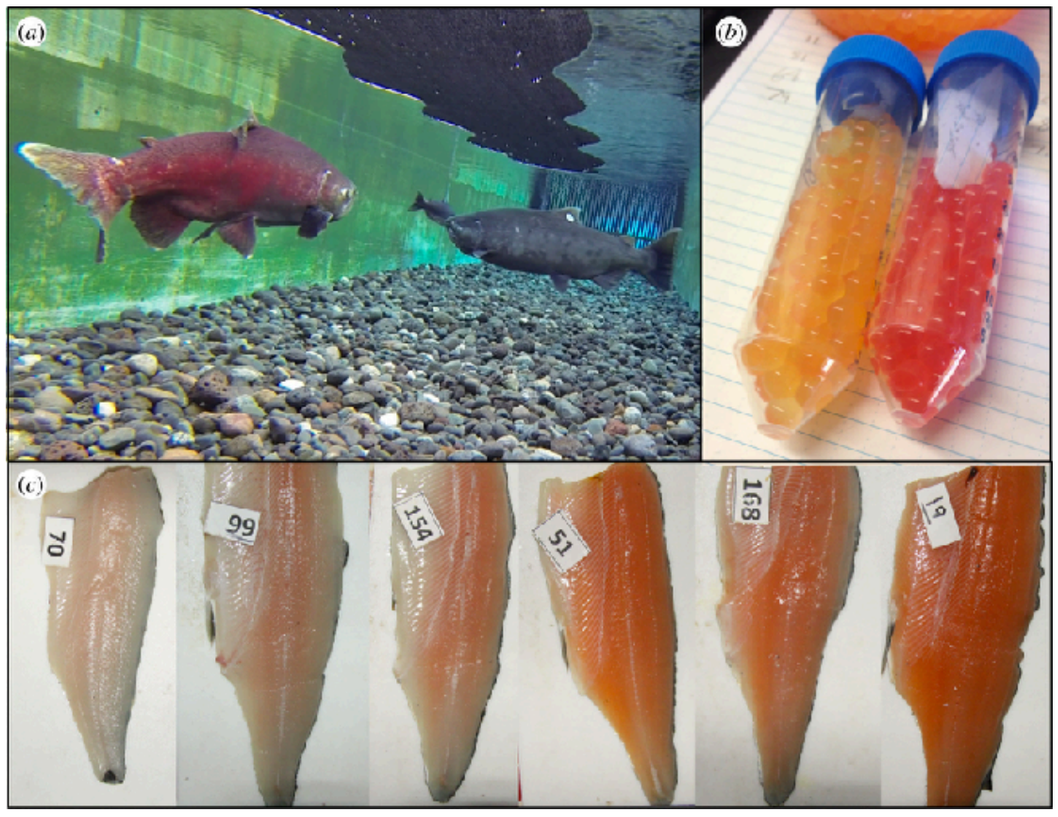
Whole genome sequences of 22 Salmon
11 with red flesh
11 with white flesh

Each individual has been sequenced to approximately 10x coverage using Illumina HiSeq paired-end reads

The FASTQs have been trimmed and screened for quality

```
/mnt/data/Assignment/fastq/
```

Assignment

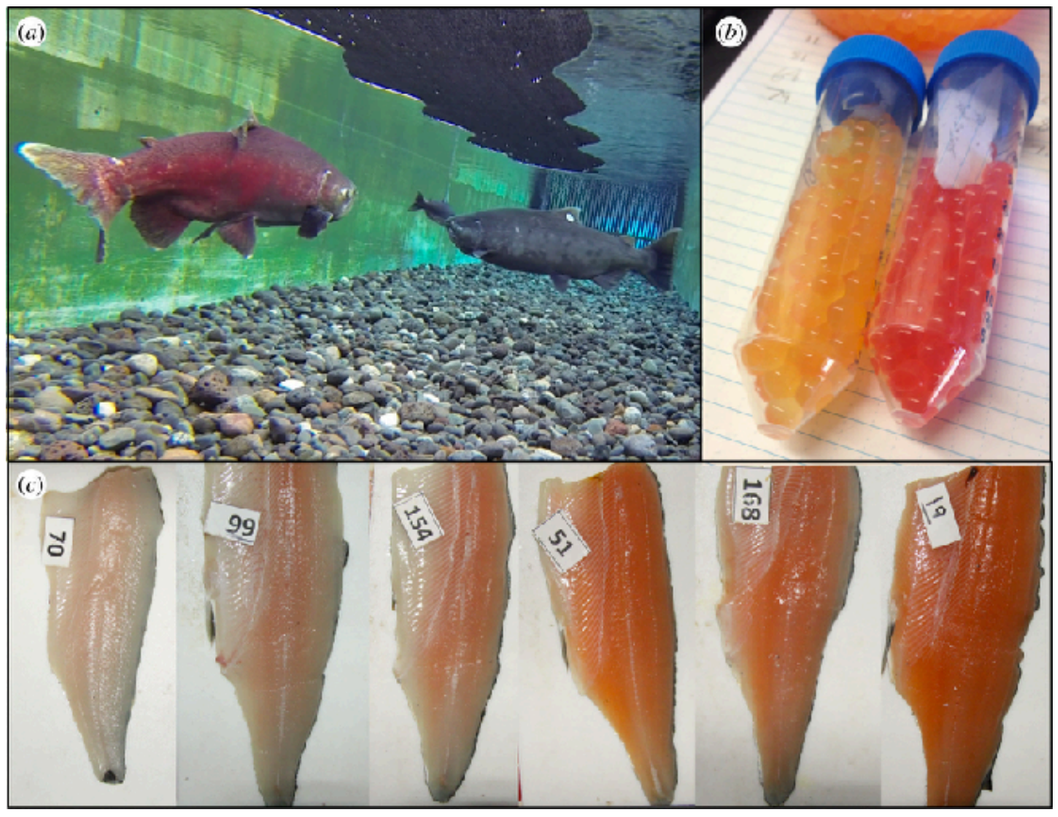


Build a pipeline to use the sequence data to identify the causal gene

- 75% of your mark comes from building the pipeline - using comments to show us you understand the different steps

- You'll get the remaining 25% of the mark if your pipeline works on the server you're working on and tell us the causal gene

Assignment



Submit your shell script to me two weeks after the last class