What else can we do with lcWGS data?

This is not an exhaustive list….

- Genotype-phenotype association
- Genotype-environment associations
- Introgression, local ancestry or gene flow analyses
- Selection scans
- Relatedness analyses
- Species/Population assignment
- ….

Types of input formats for downstream analyses:

1. Many programs use allele counts or allele frequency tables as input:

   A. Minor allele frequency tables can be generated in R by joining xxx.maf.gz files for each population to generate a table in the following format:

Chr	Pos	MAF Pop1	MAF Pop2
1	10	0.4	0.1
1	20	0.3	0.2
2	50	0.1	0.7
2	80	0.8	0.4

Types of input formats for downstream analyses:

1. Many programs use allele counts or allele frequency tables as input:

   A. Minor allele frequency tables can be generated in R by joining xxx.maf.gz files for each population to generate a table in the following format.

   B. Allele counts per site and population can be generated in multiple ways. In R, one can multiply the allele frequency (knownEM) column in the xxx.maf.gz file with the number of genotyped individuals for that site (nInd column).

```
#Estimate minor allele count (x) and major allele count (y) for each population
JIGA.maf$x=round((JIGA.maf$knownEM)*(2*(JIGA.maf$nInd)), digits = 0)
JIGA.maf$y=(2*(JIGA.maf$nInd))-JIGA.maf$x
```

   One can join them again into a single table with 1 column per population.

Types of input formats for downstream analyses:

1. Many programs use allele counts or allele frequency tables as input:

   A. Minor allele frequency tables can be generated in R by joining xxx.maf.gz files for each population to generate a table in the following format.

   B. Allele counts per site and population can be generated in multiple ways. In R, one can multiply the allele frequency (knownEM) column in the xxx.maf.gz file with the number of genotyped individuals for that site (nInd column).

   DOWNSIDE: Genotype uncertainty is not carried over. But with sufficient sample sizes, inferred allele frequencies from lcWGS data are generally very accurate.

Approaches that can use these data:

Approaches that can use these data:

1. Population-level GWAS and Genotype-Environment associations with BayPass:

Allele counts as input

```
--- file begins here ---
81 19 86 14 2 98 8 92 32 68 23 77
89 11 81 19 9 91 1 99 27 73 27 73
89 11 91 9 0 0 15 85 77 23 80 20

[...97 more lines...]

--- file ends here ---
```

Approaches that can use these data:

1. Population-level GWAS and Genotype-Environment associations with BayPass:

Allele counts as input

Pop1 Pop2

```
--- file begins here ---
snp1 81 19 86 14 2 98 8 92 32 68 23 77
snp2 89 11 81 19 9 91 1 99 27 73 27 73
snp3 89 11 91 9 0 0 15 85 77 23 80 20

[...97 more lines...]

--- file ends here ---
```

Approaches that can use these data:

1. Population-level GWAS and Genotype-Environment associations with BayPass.

2. Selective sweep analyses within species based on minor or derived allele counts.

position	x	n	folded
460000	9	100	0
460010	100	100	0
460210	30	78	1
463000	0	94	0
…	…	…	…

x = allele count
n = number of samples
folded = polarized or not? (1 = not polarized)

1. Estimate SFS for the entire genome

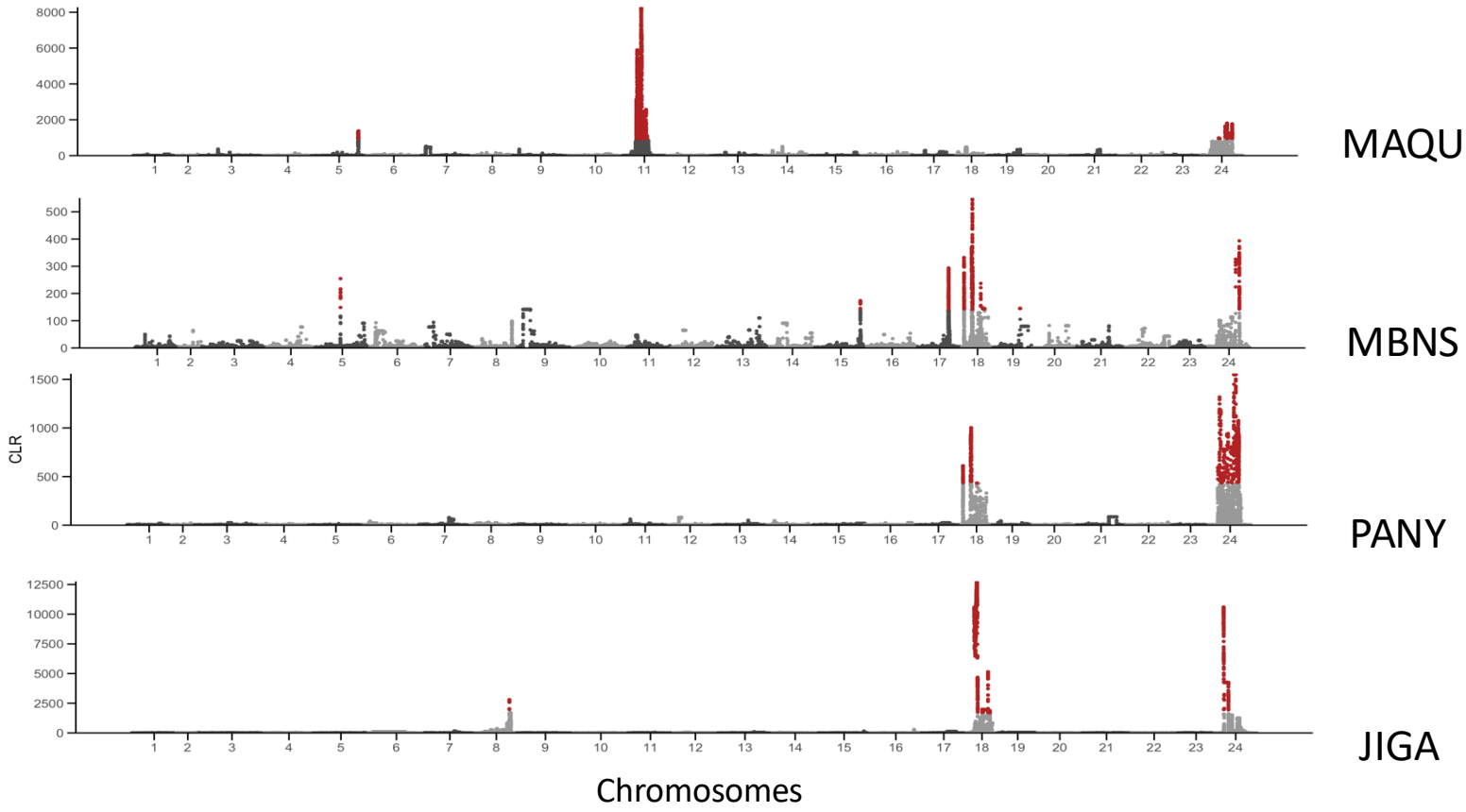
2. Run Sweepfinder2 or sweeD by chromosome for the entire genome

Approaches that can use these data:

Sweepfinder2

position	x	n	folded
460000	9	100	0
460010	100	100	0
460210	30	78	1
463000	0	94	0
…	…	…	…

x = allele count
n = number of samples
folded = polarized or not? (1 = not polarized)



1. Estimate SFS for the entire genome

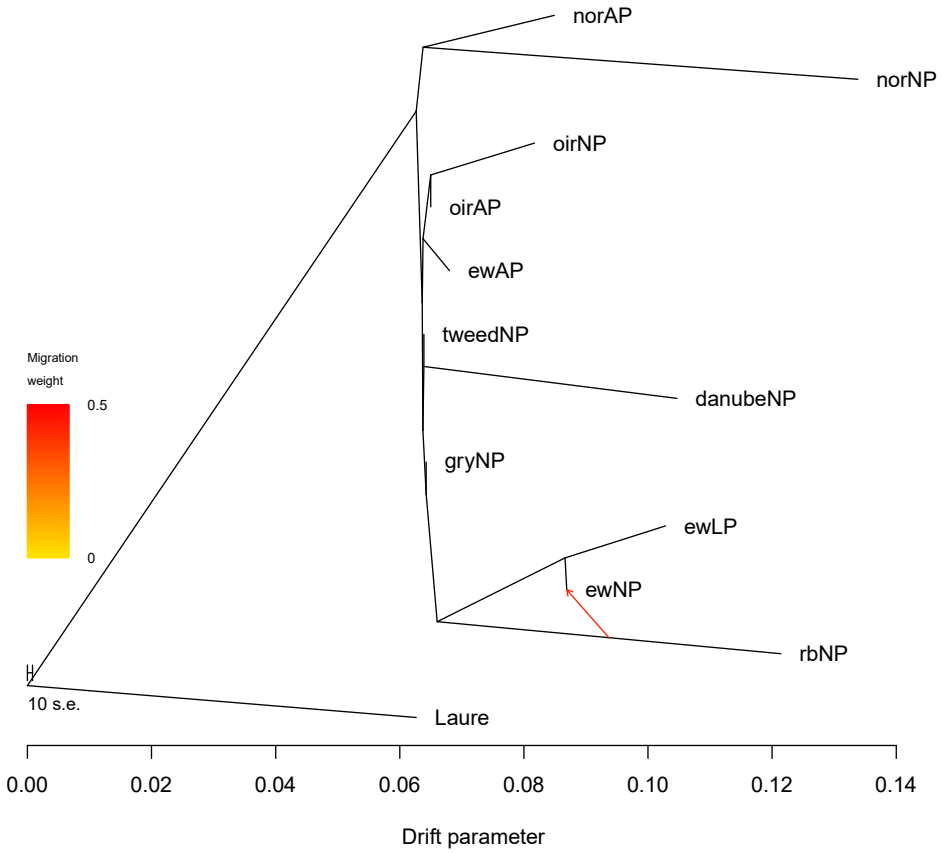
2. Run Sweepfinder2 by chromosome for the entire genome

Phylogenetic relationship and gene flow

TreeMix analyses

TreeMix uses a simple allele counts table as input:

```
pop1 pop2 pop3 pop4
5,1 1,1 4,0 0,4
3,3 0,2 2,2 0,4
1,5 0,2 2,2 1,3
```



Pickrell & Pritchard. 2012 PloS Genetics

Genome-wide association studies

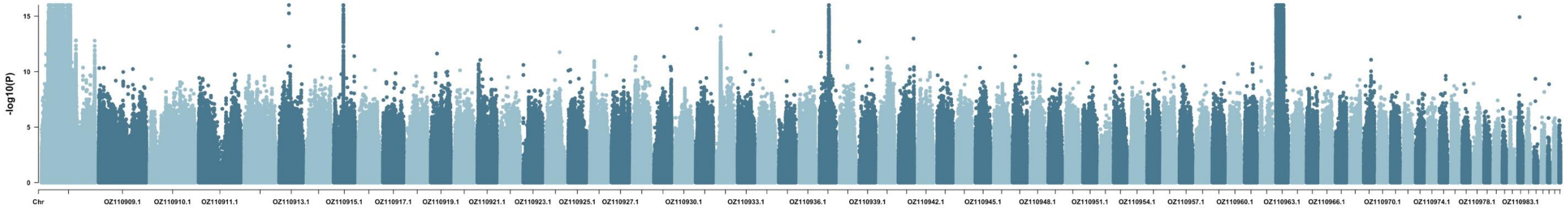
Genome-wide association studies

GWAS in ANGSD:

```
./angsd -doAsso
abcAsso.cpp:
        -doAsso 0
        1: Frequency Test (Known Major and Minor)
        2: Score Test
        4: Latent genotype model
        5: Score Test with latent genotype model - hybrid test
        6: Dosage regression
        7: Latent genotype model (wald test) - NOT PROPERLY TESTED YET!
  Frequency Test Options:
        -yBin           (null)  (File containing disease status)

  Score, Latent, Hybrid and Dosage Test Options:
        -yBin           (null)  (File containing disease status)
        -yCount         (null)  (File containing count phenotypes)
        -yQuant         (null)  (File containing phenotypes)
        -cov            (null)  (File containing additional covariates)
        -sampleFile             (null)  (.sample File containing phenotypes and covariates)
        -whichPhe       (null)  Select which phenotypes to analyse, write phenos comma seperated ('phe1,phe2,...'), only works with a .sample
file
        -whichCov       (null)  Select which covariates to include, write covs comma seperated ('cov1,cov2,...'), only works with a .sample
file
        -model  1
        1: Additive/Log-Additive (Default)
        2: Dominant
        3: Recessive

        -minHigh        10      (Require atleast minHigh number of high credible genotypes)
        -minCount       10      (Require this number of minor alleles, estimated from MAF)
        -assoThres      0.000001        Threshold for logistic regression
        -assoIter       100     Number of iterations for logistic regression
        -emThres        0.000100        Threshold for convergence of EM algorithm in doAsso 4 and 5
        -emIter 40      Number of max iterations for EM algorithm in doAsso 4 and 5

        -doPriming      1       Prime EM algorithm with dosage derived coefficients (0: no, 1: yes - default)

        -Pvalue 0       Prints a P-value instead of a likelihood ratio (0: no - default, 1: yes)

  Hybrid Test Options:
        -hybridThres            0.050000        (p-value value threshold for when to perform latent genotype model)
```

Genotype-phenotype/environment association

GWAS for lamprey for life-history using angsd with population structure correction using PC1 to 3.
Model: Score test with latent genotype model.



Genome-wide association studies

However, some more sophisticated GWAS software can also use ‘mean genotypes’ as input that still contain some of the uncertainty:

1. What is a mean genotype (or sometimes called allele dosage)?

Mean genotypes are a composite score of genotype likelihoods (from beagle file) and range from 0 to 2.

0 = homozygous for major/major
1 = heterozygous (major/minor)
2 = homozygous for minor/minor

However, mean genotypes can also have all possible intermediate values, e.g. 0.5, 0.75, 1.2, 1.8

This captures the uncertainty of the genotype.

Genome-wide association studies

However, some more sophisticated GWAS software can also use 'mean genotypes' as input that still contain some of the uncertainty:

1. What is a mean genotype (or sometimes called allele dosage)?

2. How do we estimate the mean genotype:

   We can treat GLs as genotype probabilities and estimate the **expected genotype** (dosage) as 0·P(AA) + 1·P(AB) +2·P (BB) = **P(AB) + 2P(BB).**

Marker	Allele 1	Allele 2	Ind0 (AA)	Ind0 (AB)	Ind0 (BB)	Ind1	Ind1	Ind1
1_14000023	1	0	0.941177	0.058822	0.000001	0.799685	0.199918	0.000397
1_14000113	0	2	0.855993	0.106996	0.037010	0.333333	0.333333	0.333333
1_14000202	2	0	0.835380	0.104420	0.060201	0.799685	0.199918	0.000397

1. If sites are 0.3333 set to NA

2. Allele dosage = Major/Minor + 2*Minor/Minor

For example: 0.058822 + (2*0.00001) = 0.058842

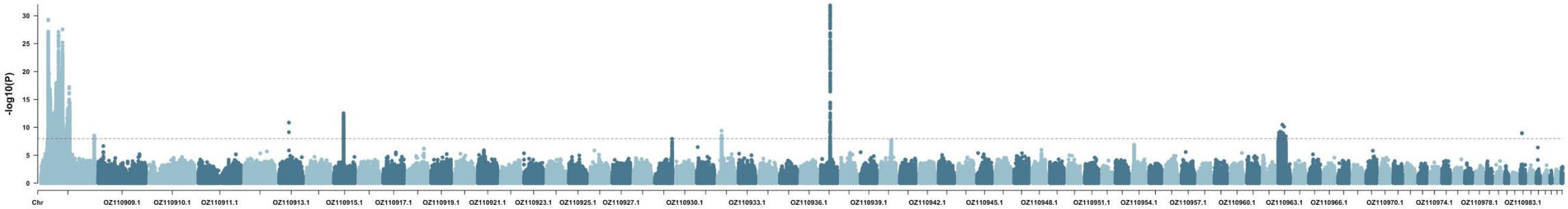
Note: We also polarize by a random reference individual to always use the same minor allele

Genome-wide association studies

However, some more sophisticated GWAS software can also use ‘mean genotypes’ as input that still contain some of the uncertainty:

Mean genotype file can e.g. be used for linear mixed models in GEMMA with correction using a genetic covariance matrix and principal components.

GWAS with GEMMA



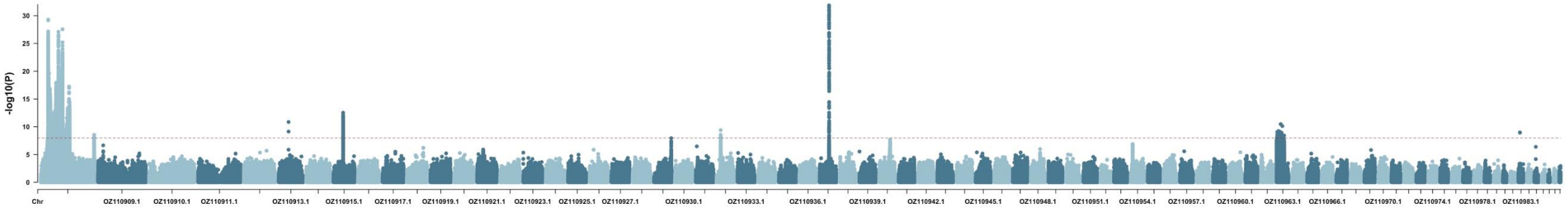
GEMMA: Zhou & Stephens 2012 Nature Genetics

Genome-wide association studies

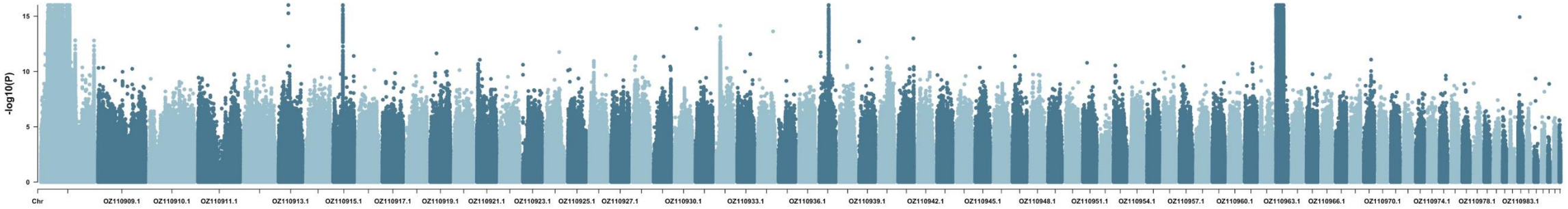
However, some more sophisticated GWAS software can also use ‘mean genotypes’ as input that still contain some of the uncertainty:

Mean genotype file can e.g. be used for linear mixed models in GEMMA with correction using a genetic covariance matrix and principal components.

GWAS with GEMMA



GWAS with ANGSD



Demographic inference

Demographic inference

Option 1: Use the 1D and 2D (or 3D …) SFS produced in angsd or winSFS with demographic inference software (e.g. ðaði, GADMA, fastsimcoal2, stairwayplot2)

Important: SFS can be problematic with very low coverage data (< 2x), which might lead to mis-inferences. But new approaches such as winSFS might help with that.
NOTE: Don’t use a minor allele frequency filter!

Option 2: Calculate summary statistics from lcWGS data and use them for demographic inference using ABC.

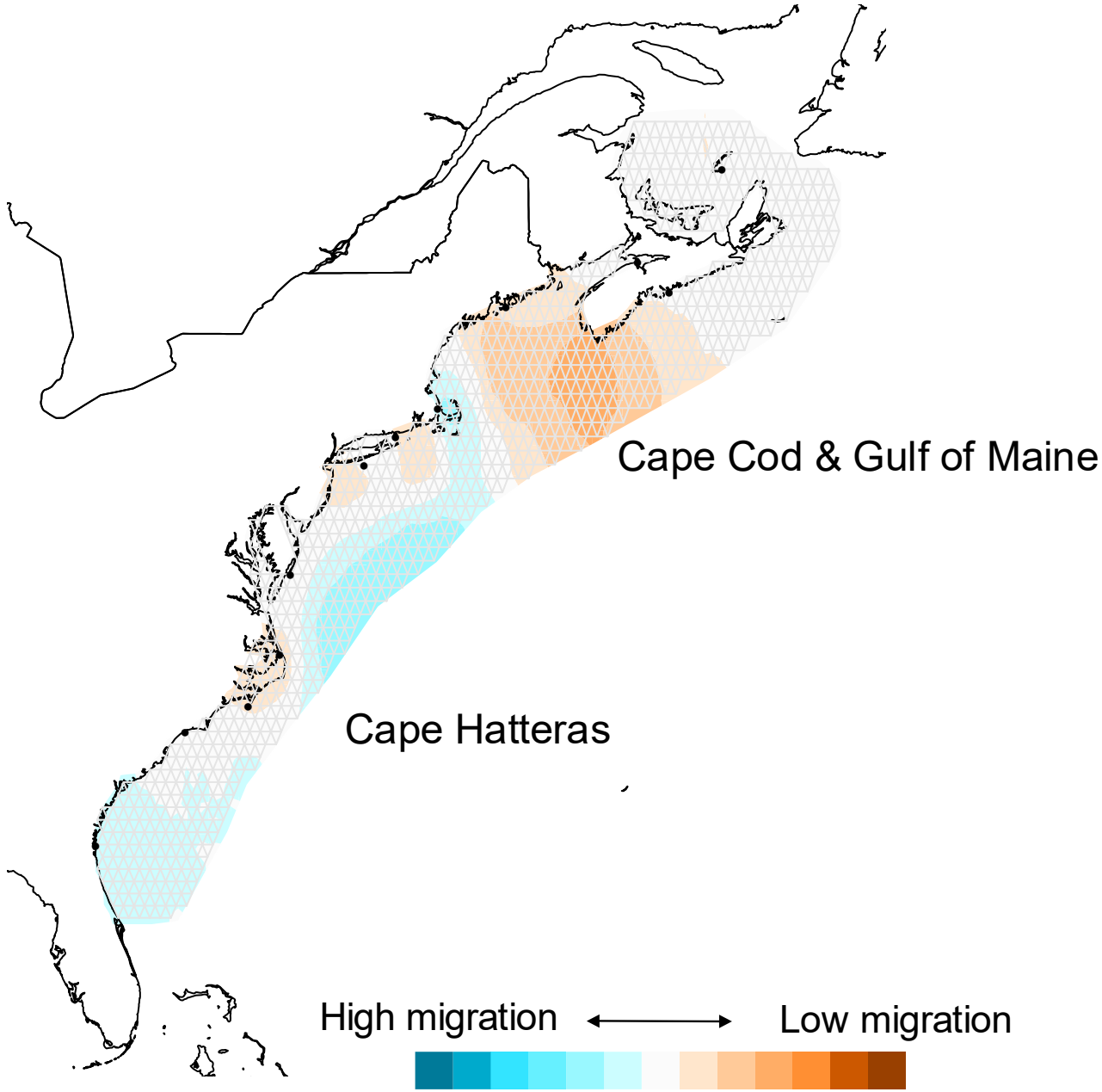
Disclaimer: Have not tested this and one would need to implement custom approaches.

Effective migration surfaces

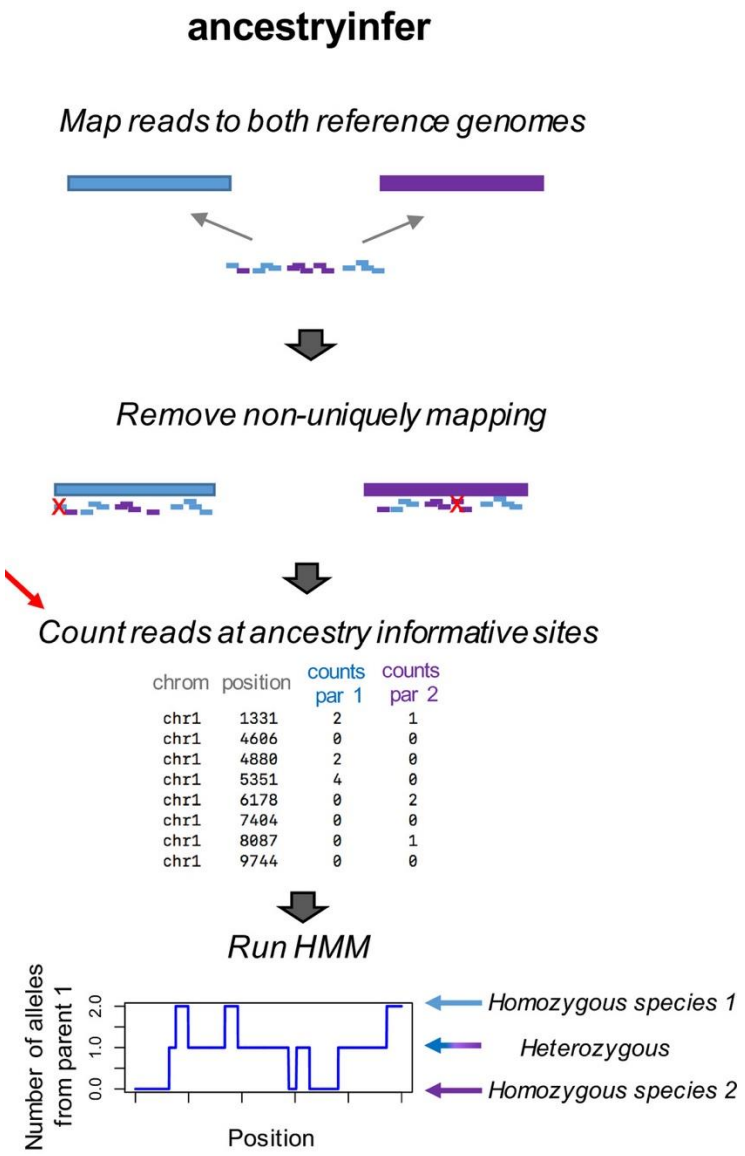
EEMS, Petkova et al. 2016

Input:

- **Pairwise Fst matrix**
- **IBS matrix** from angsd



Local ancestry



Input: References genomes for each species/population and fastq files for individuals.

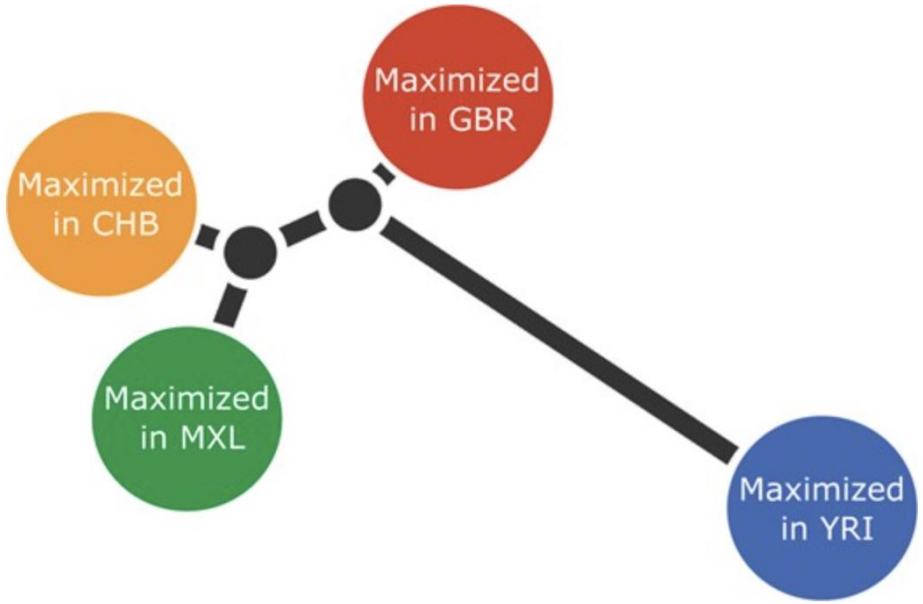
Caveats: One needs ancestry informative SNPs.

ancestryinfer, Schumer et al. 2020 MER

Selective sweep analysis

Ohana

Fig. 4.



JOURNAL ARTICLE

Detecting Selection in Multiple Populations by Modeling Ancestral Admixture Components

Jade Yu Cheng ✉, Aaron J Stern, Fernando Racimo, Rasmus Nielsen

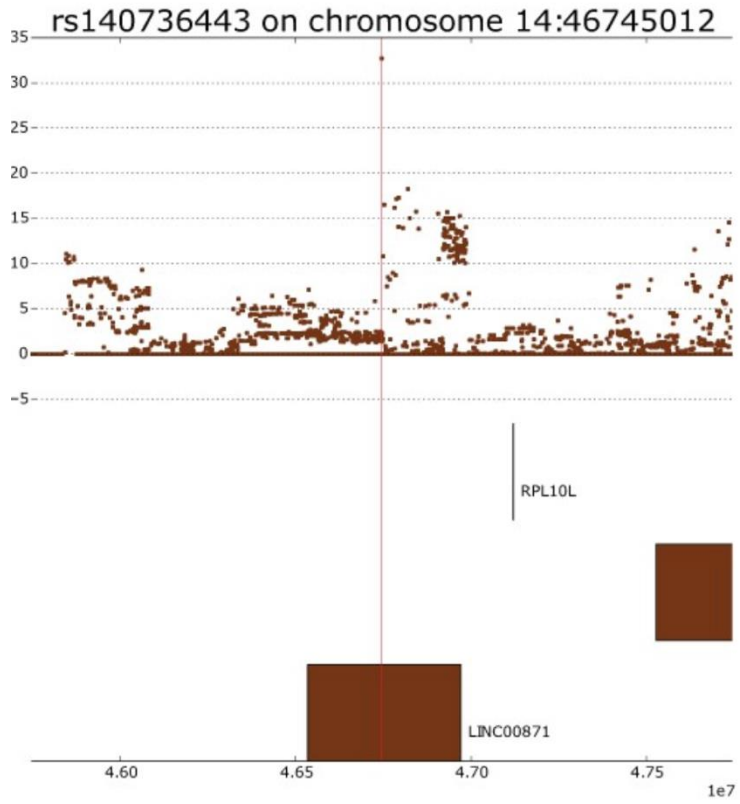
Molecular Biology and Evolution, Volume 39, Issue 1, January 2022, msab294,
https://doi.org/10.1093/molbev/msab294
Published: 09 October 2021

Input: Beagle genotype likelihood file

1. Reconstruct population structure
2. Correct allele frequencies using inferred structure
3. Scan for signatures of selection in specific ancestry clusters

Selective sweep analysis

Ohana



JOURNAL ARTICLE

Detecting Selection in Multiple Populations by Modeling Ancestral Admixture Components

Jade Yu Cheng ✉, Aaron J Stern, Fernando Racimo, Rasmus Nielsen

Molecular Biology and Evolution, Volume 39, Issue 1, January 2022, msab294,
https://doi.org/10.1093/molbev/msab294
Published: 09 October 2021

Input: Beagle genotype likelihood file

1. Reconstruct population structure
2. Correct allele frequencies using inferred structure
3. Scan for signatures of selection in specific ancestry clusters

Population assignment

One can use super low-coverage WGS data (< 0.5x) to assign individuals to source species or populations

--> WGSassign

Received: 19 September 2023 | Accepted: 15 December 2023

DOI: 10.1111/2041-210X.14286

Methods in Ecology and Evolution

RESEARCH ARTICLE

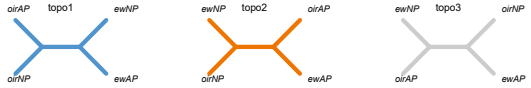
Population assignment from genotype likelihoods for low-coverage whole-genome sequencing data

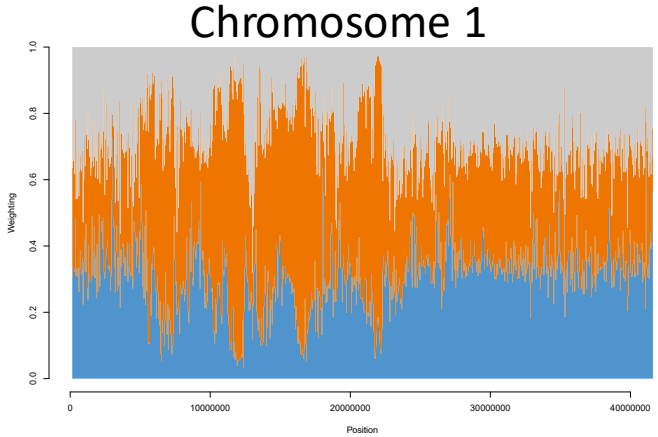
Matthew G. DeSaix[1] | Marina D. Rodriguez[1] | Kristen C. Ruegg[1] | Eric C. Anderson[1,2,3]

Phylogenetic relationship across the genome

Topological weighting analyses across the genome using TWISST based on IBS-based NJ trees

Or Which tree topology is more common in specific genomic regions?





1. Estimate IBS matrix in windows across the genome (iterate over windows) → e.g. window size can be determined based on LD decay
2. Generate neighbour-joining tree from each IBS matrix using bioNJ algorithm e.g. using ape-package in R
3. Use TWISST software to perform topological weighting analyses based on all input trees across the genome

Martin & Belleghem 2017 Genetics

List of some other approaches possible with lcWGS data:

Ancestry relationships/Gene flow:

- D-stats/ABBA-BABA using angsd: -doAbbababa
- TreeMix
- Effective migration surfaces (EEMS)
- TWISST
- WGSassign

Relatedness:

- ngsRelate
- LowKi
- NGSremix

Selection:

- PCAngsd (e.g. pcadapt)
- Sweepfinder2
- Ohana
- BayPass
- Parallel allele frequency changes (AFVaper)
- Allele frequency differentiation

Linkage Mapping:

- LepMap3

Genomic/geographic clines:

- BGC-HM (based on GL)
- HZAR (e.g. based on mean genotypes)