Computational Approaches For IDing Signatures of Local Adaptation

J. Roman Arguello

November 8, 2019

Dept. of Ecology & Evolution
University of Lausanne

what is local adaptation?

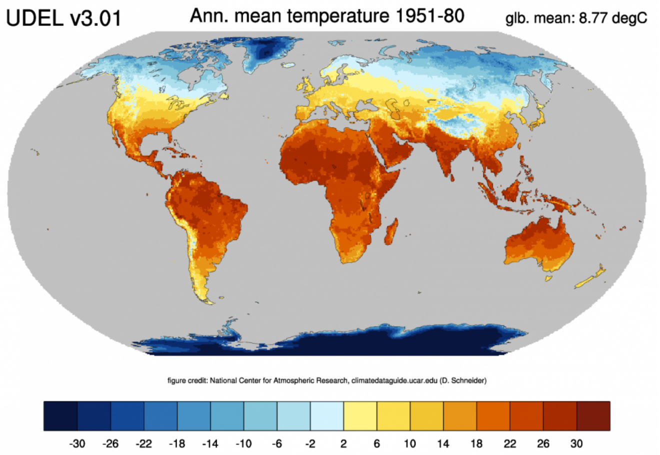
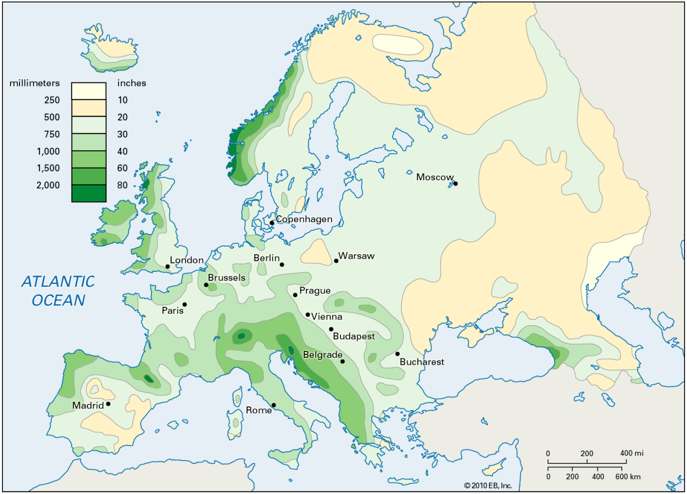
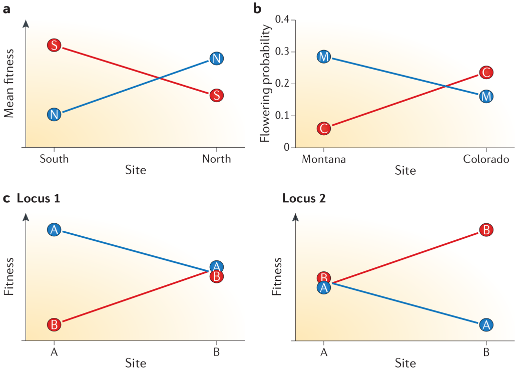


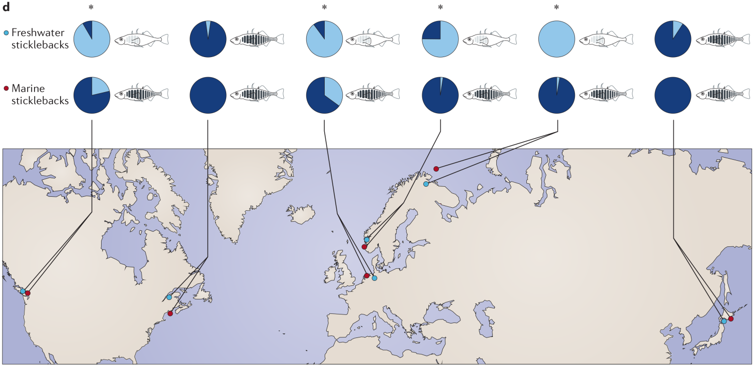
figure credit: National Center for Atmospheric Research, climatedataguide.ucar.edu (D. Schneider)


millimeters
inches
250 10
500 20
750 30
1,000 40
1,500 60
2,000 80
ATLANTIC OCEAN
London
Brussels
Paris
Madrid
Berlin
Copenhagen
Warsaw
Prague
Vienna
Budapest
Belgrade
Bucharest
Rome
Moscow
0 200 400 mi
0 200 400 600 km


boredpanda.com
© Monterey Bay Aquarium



Savolainen et al. Nat. Rev. Genetics, 2013, modified from Anderson et al. Evol. 2011 and Lowry, New Phytol., 2012



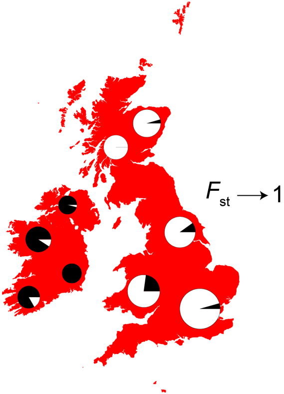
Savolainen et al. Nat. Rev. Genetics, 2013, modified from DeFaveri et al. Evol. 2011

What are approaches used to detect local adaptation?

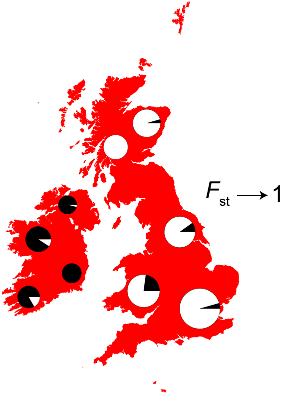
- population genomic approaches
- association mapping
- quantitative trait loci mapping (QTL)

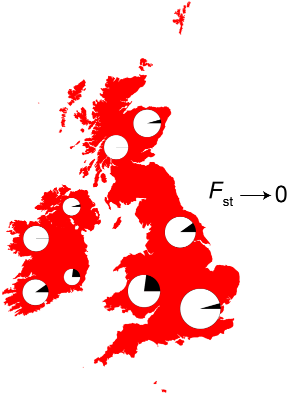
Population genetic approaches: Fst and empirical distributions

Scanning for recent selection using F_{st}



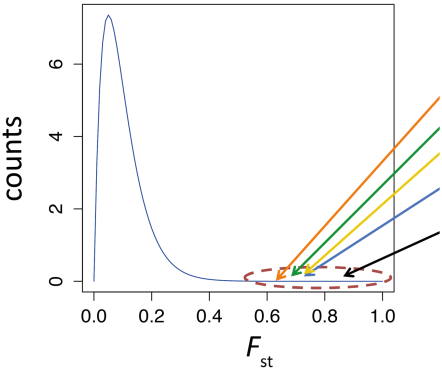
Scanning for recent selection using F$_{st}$





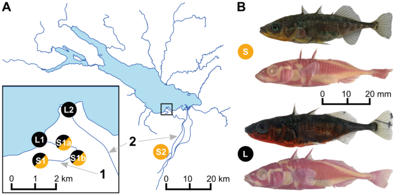
- formulas for Fst: see notes (and links)

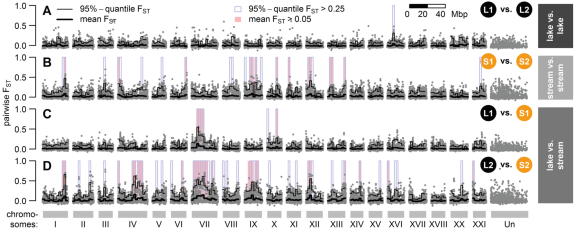
Scanning for recent selection using F$_{st}$



- pros/cons

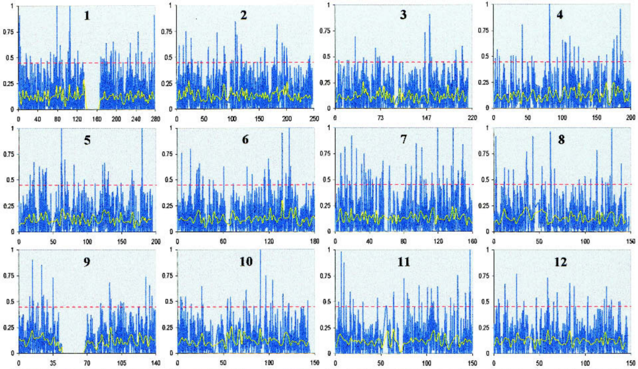
Scanning for recent selection using F_{st}





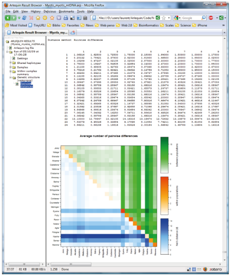
Marques et al. PLoS Genetics, 2016

Scanning for recent selection using F$_{st}$



Akey et al. Genome Res, 2002

software

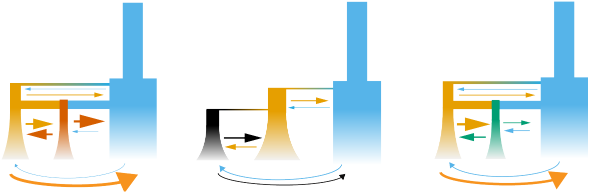


Julio Rozas, Albert Ferrer-Mata, Juan Carlos Sánchez-DelBarrio, Sara Guirao-Rico, Pablo Librado, Sebastián Ramos-Onsins and Alejandro Sánchez-Gracia
Universitat de Barcelona
Current Beta Version: **6.12.03** (February 26, 2019)

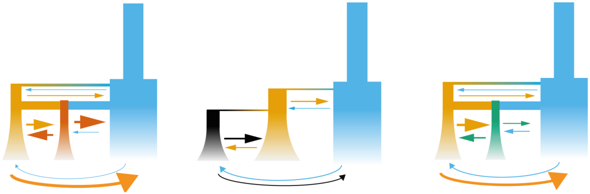
- custom scripts
- arlequin
- DnaSP
- fsthet

- But can we have a better expectation of what an outlier would be?

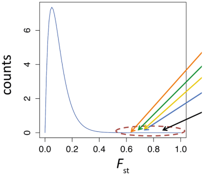
“raw” F_{st} usage



“raw” F_{st} usage

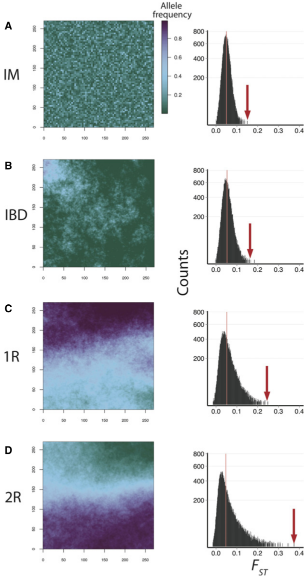


simulante data from the best fitting model(s)



- pros/cons

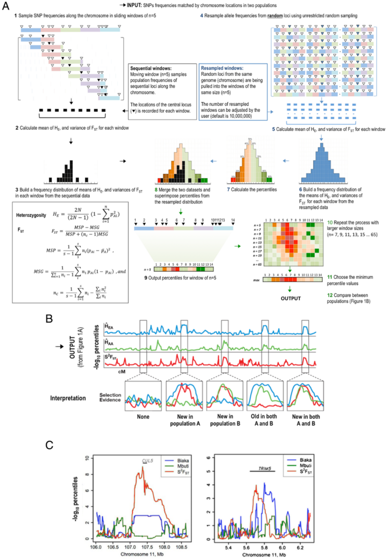
“raw” F_{st} usage



Latterhos, Whitlock, Mol Eco, 2014

software

- **demographic simulators:** ms, fastsimcoal, primems, cosi, ...
- **GppFst:** "will compute the probability of observing an empirical proportion of loci within a given FST range conditioned on the particular coalescent model of population divergence"
  - only works for 2 populations
- **SmileFinder:** "builds a resampled distribution (the baseline) of random multi-locus sets matched to the sizes of sliding windows.."
- **OutFLANK:** a likelihood method that uses a trimmed distribution of F_{st} values as a way to infer distribution of F_{st} for neutral variants



INPUT: SNPs frequencies matched by chromosome locations in two populations

1 Sample SNP frequencies along the chromosome in sliding windows of n=5

4 Resample allele frequencies from **random** loci using unrestricted random sampling

Sequential windows:
Moving window (n=5) samples population frequencies of sequential loci along the chromosome.

The locations of the central locus (▼) is recorded for each window.

Resampled windows:
Random loci from the same genome (chromosome) are being pulled into the windows of the same size (n=5)

The number of resampled windows can be adjusted by the user (default is 10,000,000)

2 Calculate mean of H_E, and variance of F_{ST} for each window

5 Calculate mean of H_E, and variance of F_{ST} for each window

3 Build a frequency distribution of means of H_E, and variances of F_{ST} in each window from the sequential data

8 Merge the two datasets and superimpose percentiles from the resampled distribution

7 Calculate the percentiles

6 Build a frequency distribution of the means of H_E, and variances of F_{ST} for each window from the resampled data

Heterozygosity
$$H_E = \frac{2N}{(2N-1)}\left(1 - \sum_{i=1}^{n} p_{Ai}^2\right)$$

F_{ST}
$$F_{ST} = \frac{MSP - MSG}{MSP + (n_c - 1)MSG}$$

$$MSP = \frac{1}{s-1}\sum_{1}^{s} n_i (p_{Ai} - \bar{p}_A)^2 \ ,$$

$$MSG = \frac{1}{\sum_{i=1}^{s} n_i - 1}\sum_{i}^{s} n_1\, p_{Ai}(1 - p_{Ai}) \ , and$$

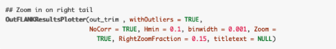
$$n_c = \frac{1}{s-1}\sum_{i=1}^{s} n_i - \frac{\sum_i n_i^2}{\sum_i n_i}$$

9 Output percentiles for window of n=5

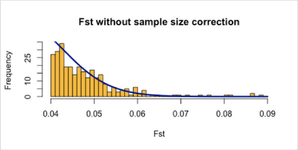
10 Repeat the process with larger window sizes (n= 7, 9, 11, 13, 15 … 65)

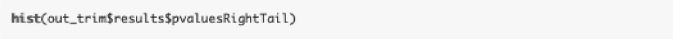
11 Choose the minimum percentile values

12 Compare between populations (Figure 1B)

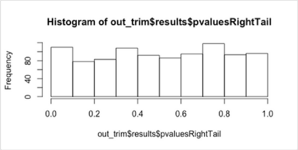


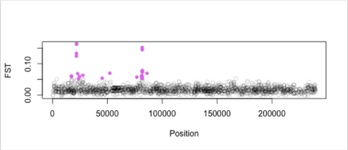
```
## Zoom in on right tail
OutFLANKResultsPlotter(out_trim , withOutliers = TRUE,
                       NoCorr = TRUE, Hmin = 0.1, binwidth = 0.001, Zoom =
                         TRUE, RightZoomFraction = 0.15, titletext = NULL)
```





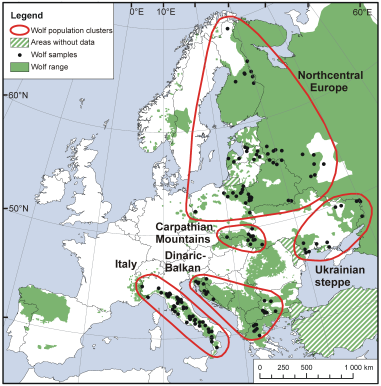
```
hist(out_trim$results$pvaluesRightTail)
```



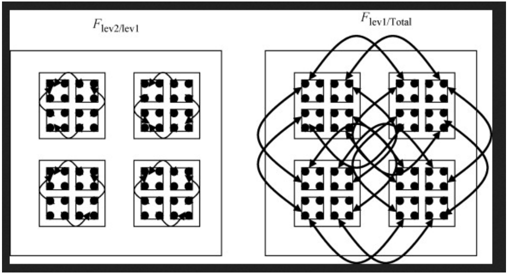


OutFLANK, Latterhos, Whitlock, Mol Eco, 2014, Whitlock and Latterhos, Am Nat. 2015

- Populations and F_{st} can be thought of hierarchically

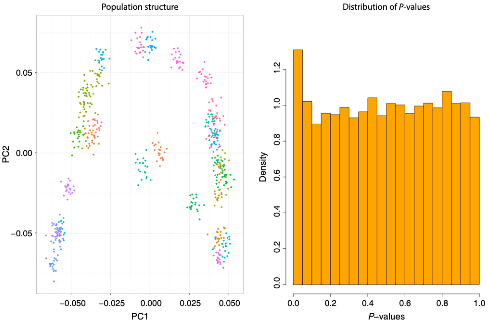


Stronen et. al, PLoS One, 2013



- **software:** HierFstat
    - while some software allows a limited number of hierarchical levels, HierFstat allows an arbitrary number of levels

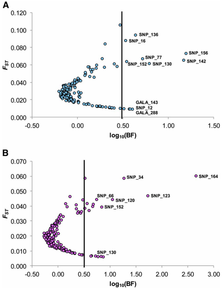
Goudet, Molecular Ecology Notes, 2015 implementation of Yang, Evol.1998



- **software:** pcadapt
  - PCA-based approach for identifying outliers

Luu et al. Mol. Eco. Res. 2017

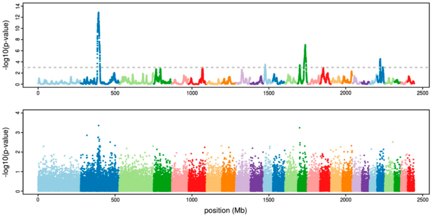
More complicated model-based methods



- **software:** Bayescan
  - subpopulations are correlated to each other through a common migrant gene pool (poulation structure)
    - each subpopulation's allele frequency difference is measured via an F_{st} coefficient
    - the F_{st} coefficient has a "shared" β component and a locus specific α component
    - selection is inferred when the α is necessary to explain allele frequency differences

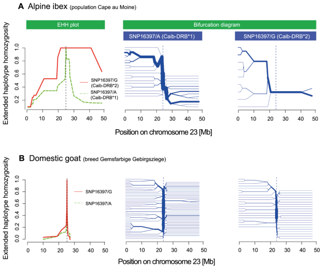
- **software:** BlockFeST
    - extension on Bayescan that allows the grouping of SNPs into predefined blocks

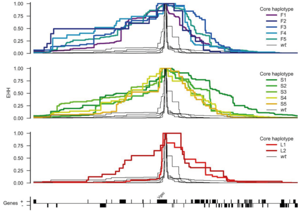
using haplotype data



- **software:** hapFLK
  - extension of F_{st} scan based on haplotype data

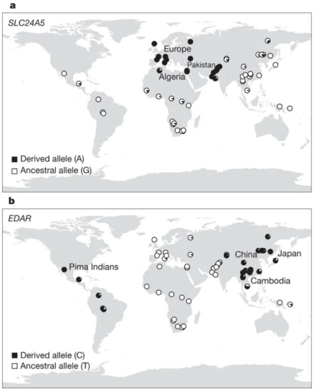
Fariello et al. Genetics, 2013

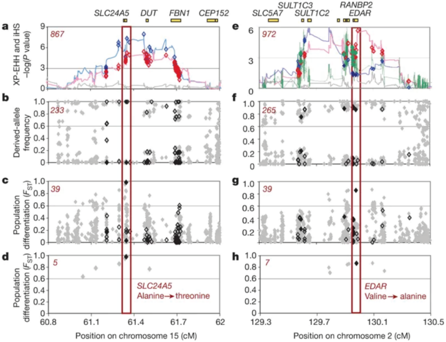




- **tests:** iHS, EHH, XP-EHH
- **software:** rehh, fastPHASE, hapbin, selscan,...
    - tests based on the size of haplotypes
    - can be used among populations to identify population-specific signals (XP-EHH)

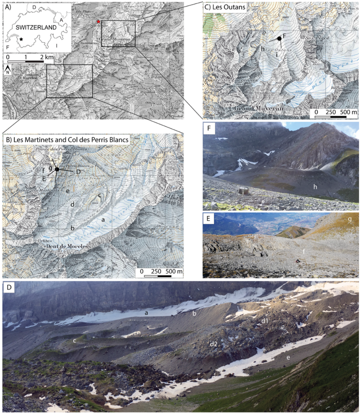
Grossen et al. PLoS Genetics, 2015; Clarkson et al. bioRiv, 2018

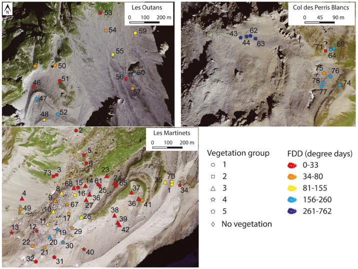




Sabeti et al. Nature, 2007

incorporating population genetic data with phenotypes





- in addition to genotypes, one might also have environmental/ecological variables that vary among populations
    - temp, sun/shade, humidity, soil types, longitude,...

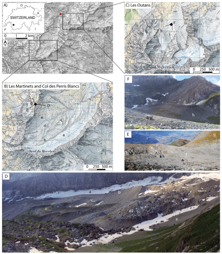
Elisa et al., Influence of microclimate and geomorphological factors on alpine vegetation in the Western Swiss Alps, 2019

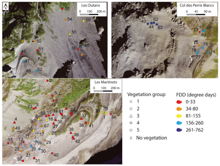
BayeScanEnv

- uses much of the same machinery as Bayescan, but adds additional capacity to incorporate environmental variables (environmental differentiation, E)
  - : **M1:** Neutral model
  - : **M2:** Local adaptation model with environmental differentiation E
  - : **M3:** Locus-specific model
- M2 is the primary model, while M3 can be considered as a "nuisance" model that helps reduce false positive rates by helping to explain increase variation in F_{st} by locus-specific effects that do not correlate with E

Villemereuil & Gaggiotti, Methods in Ecology & Evolution, 2015

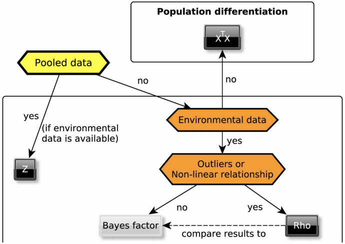
Bayenv 1.0





- **software:** Bayenv 1.0
  - Aims to address genetic relatedness by inferring an allele covariance matrix using a reference SNP set (null model)
  - Tests the null model against models where the environmental variable has linear effects on (transformed) allele frequencies
  - MCMC is used to sample the posterior distribution of the covariance matrix
  - First introduced in 2008 “informally” and then formally in 2010

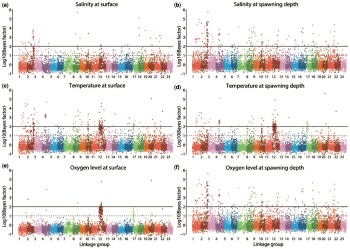
Bayenv 2.0



- **software:** Bayenv 2.0
  - uses the same Bayesian framework as Bayenv1.0
  - Aims to address unequal population sampling (standardize allele frequencies)
  - Results in an F_{st} analog that accounts for population structure
  - Aims to address pool-seq data
  - Bayenv1.0 updated to 2.0 in 2013

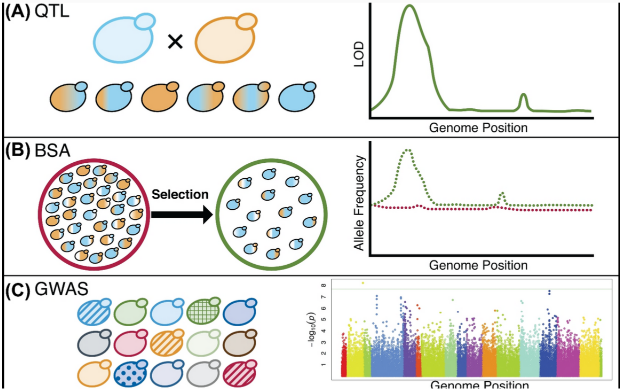
local adaptation to open water conditions

Fig. 1.—Sampling locations of Atlantic cod specimens. Samples were obtained between 2002 and 2008 using trawl. Only mature specimens were selected for genetic analysis. See table 1 for a detailed description of the samples. Average surface salinity is denoted in ‰.



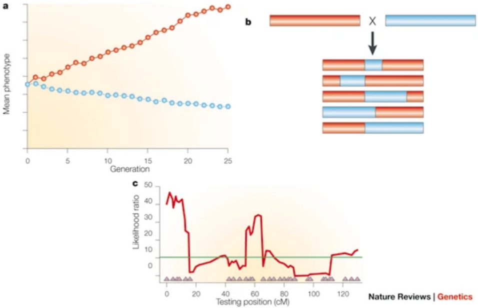
Berg et al. GBE, 2015

other genetic mapping approaches



Sardi & Gasch, FEMS Yeast Res. 2017

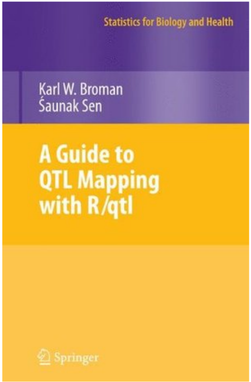
QTL: quantitative trait loci mapping



- many methods for correlating genotype data with a mapping population

Mackay, Nat. Rev. Gen., 2001

QTL: quantitative trait loci mapping



Statistics for Biology and Health

Karl W. Broman
Śaunak Sen

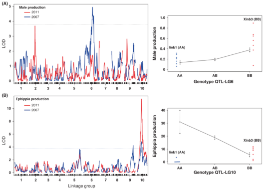
A Guide to
QTL Mapping
with R/qtl

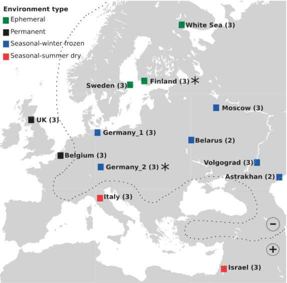
Springer

- **software:**
  - R/qtl

Broman et al. Bioinformatics, 2010

QTL: quantitative trait loci mapping

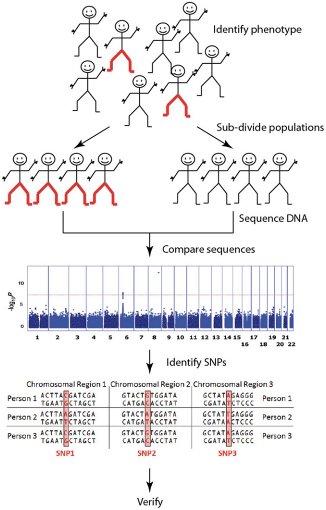




Roulin et al. Molecular Ecology, 2013

- association studies
- human elevation
- human diving?

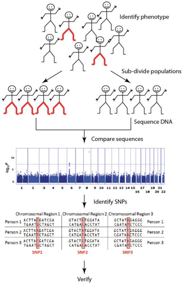
GWAS: genome-wide association studies



	Chromosomal Region 1	Chromosomal Region 2	Chromosomal Region 3	
Person 1	ACTTACGATCGA TGAATGCTAGCT	GTACTGTGGATA CATGACACCTAT	GCTATAGAGGG CGATATCTCCC	Person 1
Person 2	ACTTAAGATCGA TGAATTCTAGCT	GTACTATGGATA CATGATACCTAT	GCTATTGAGGG CGATAACTCCC	Person 2
Person 3	ACTTACGATCGA TGAATGCTAGCT	GTACTGTGGATA CATGACACCTAT	GCTATAGAGGG CGATATCTCCC	Person 3
	SNP1	SNP2	SNP3	

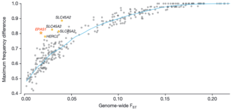
Schierding et al. Frontiers in Genetics, 2014

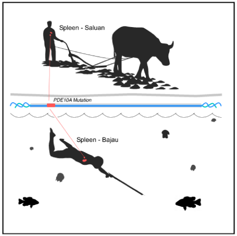
GWAS: genome-wide association studies



- **software:**
    - GWASTools, Plink, rrBLUP, GWASpoly, Hail, BGENIE

GWAS: genome-wide association studies





Melissa A. Ilardo, Ida Moltke,
Thorfinn S. Korneliussen, ...,
Suhartini Salingkat, Rasmus Nielsen,
Eske Willerslev

Correspondence
rasmus_nielsen@berkeley.edu (R.N.),
ewillerslev@snm.ku.dk (E.W.)

In Brief
Genetic and physiological adaptations enable the remarkable breath-holding ability of marine nomads.