SFS inference from NGS data to detect recent adaptive selection

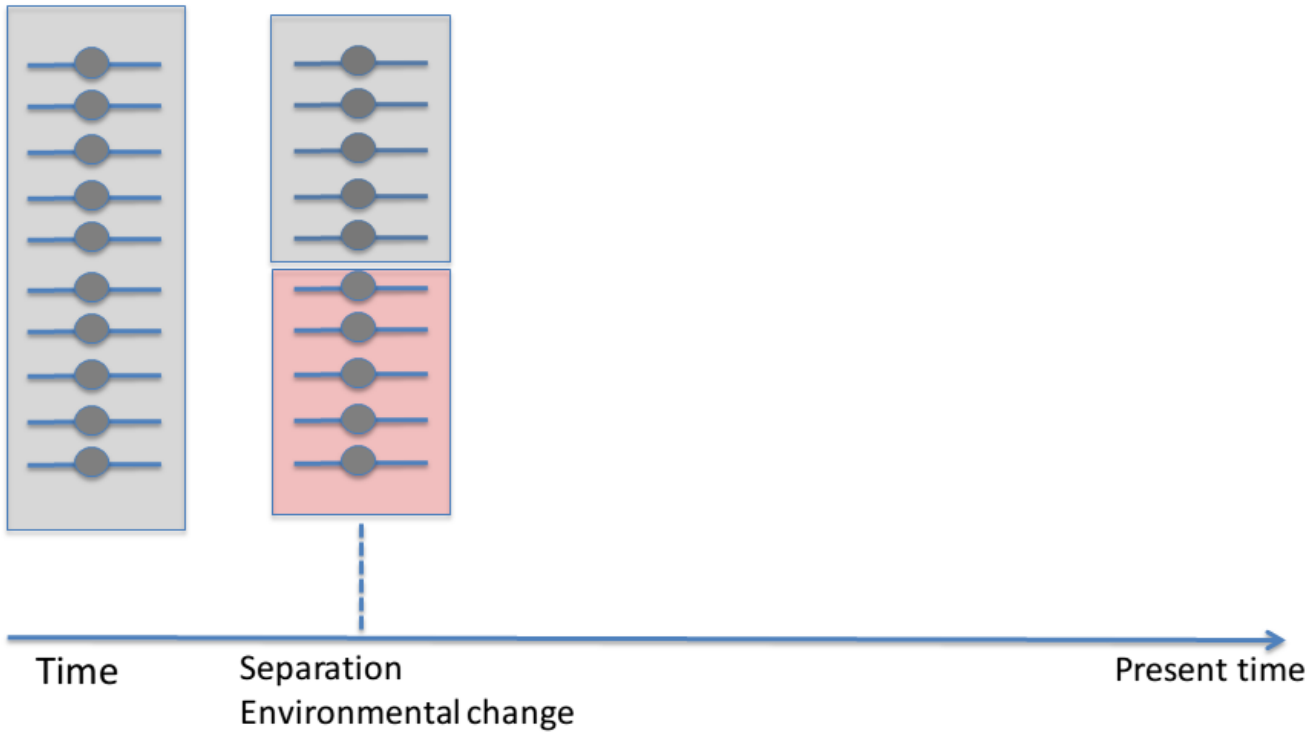
Anders Albrechtsen
The bioinformatic Centre, Copenhagen University

Outline

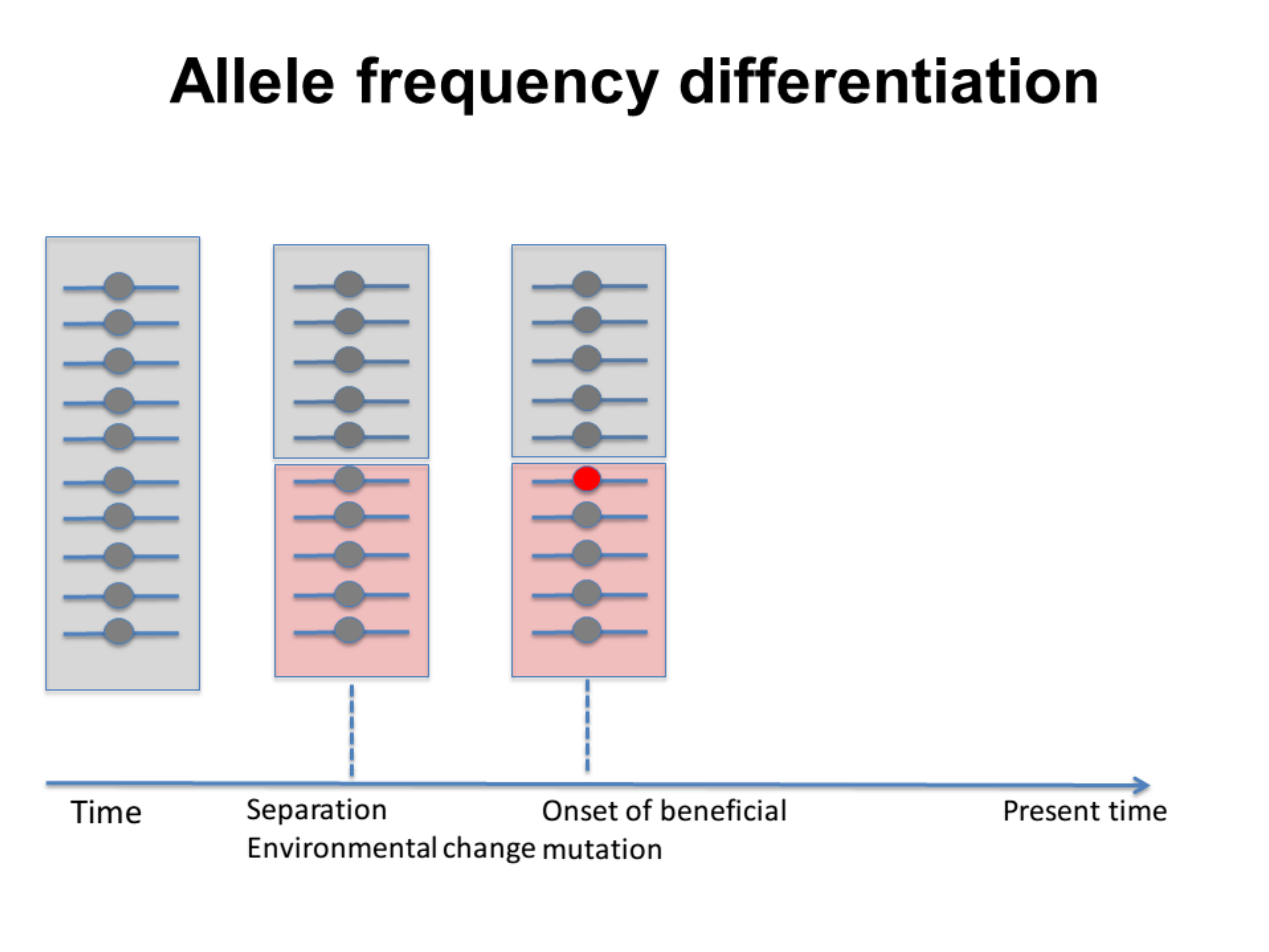
1. Allele frequency differentiation and selection
2. Tibet
   - background and hypothesis
3. Greenland
   - Background and hypothesis
4. SFS for NGS data
   - Bias for low/medium depth sequencing data
   - Genotype likelihood based SFS

Allele frequency differentiation

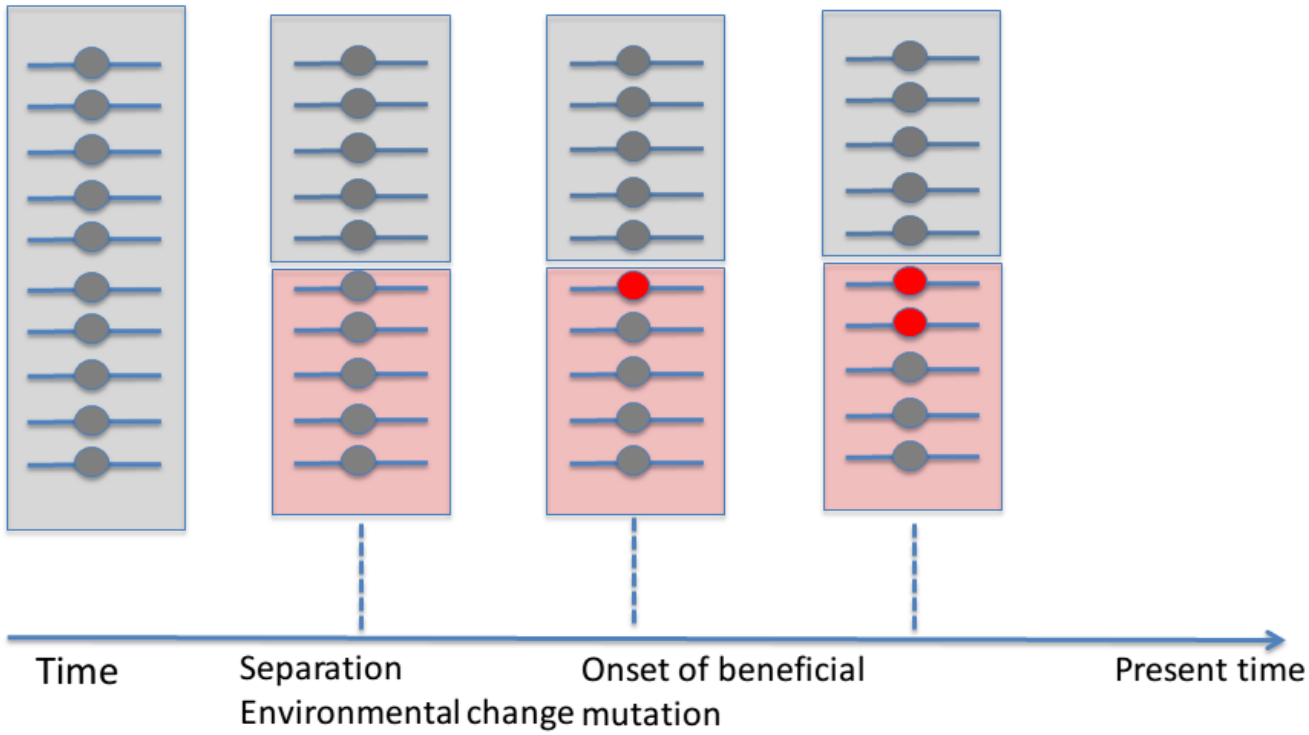
Allele frequency differentiation



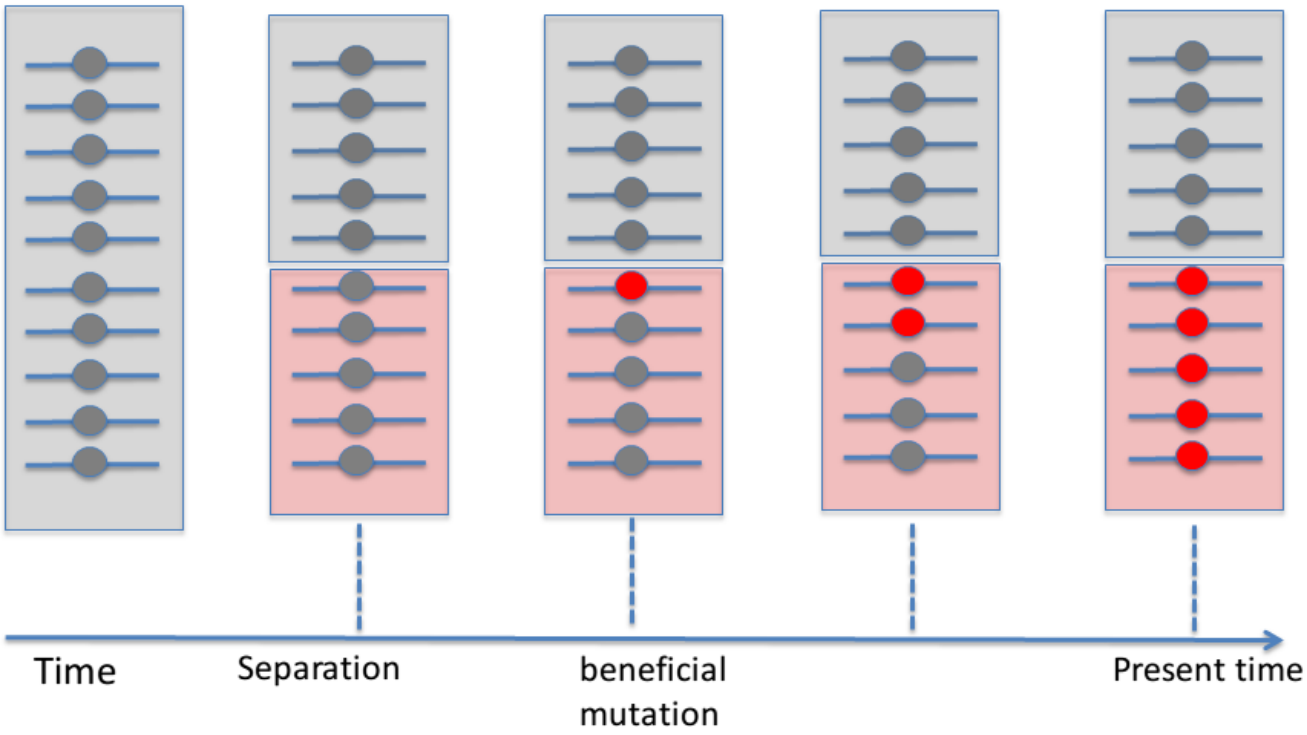
Probability of fixation



Allele frequency differentiation



Allele frequency differentiation

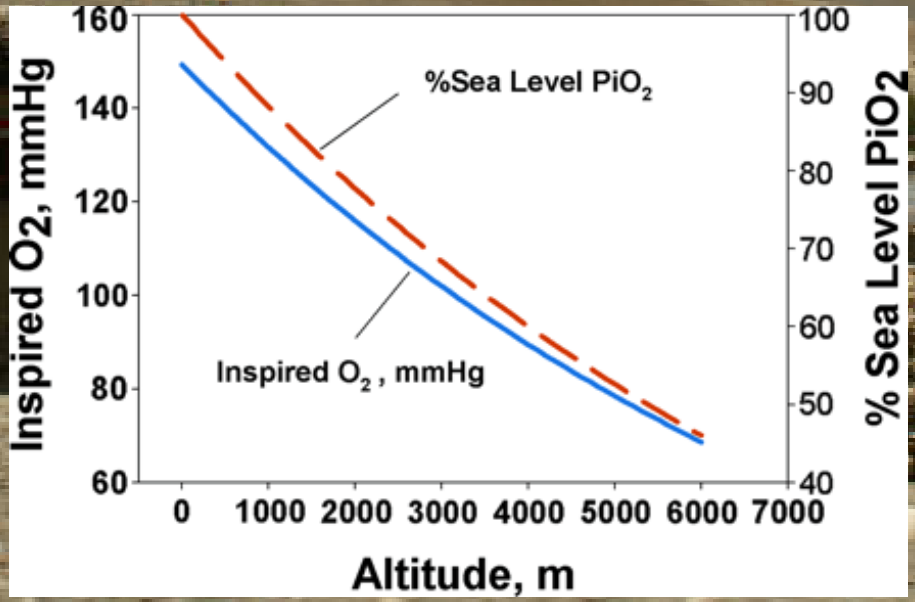


Altitude adaption in Tibet

Altitude adaption in Tibet

Yi et al. 2010

- Low oxygen has a large effect on fitness
- People living in high altitude are at greater risk of problematic births

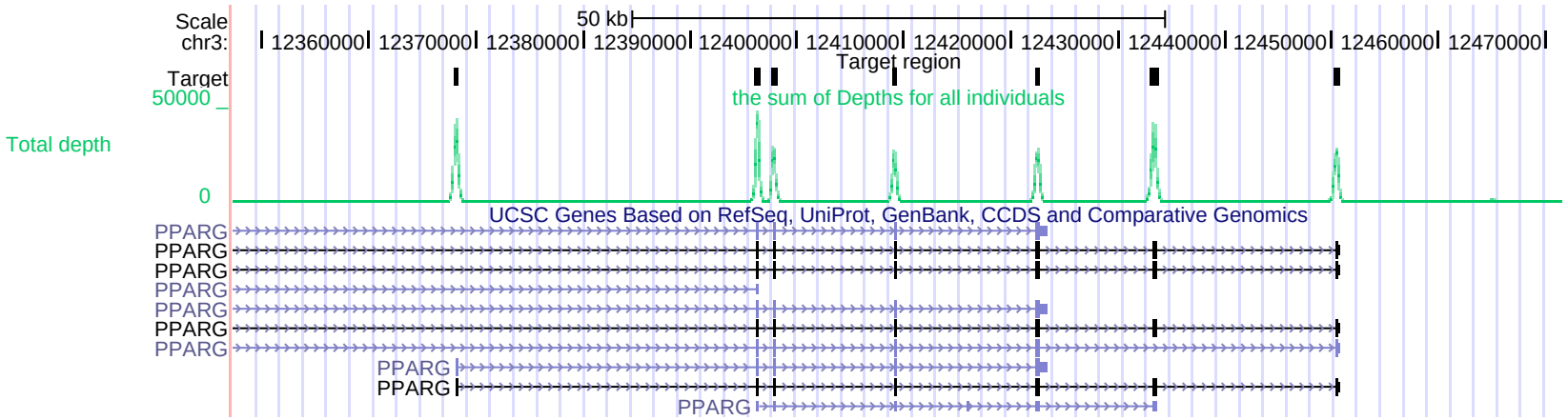


Altitude adaption in Tibet

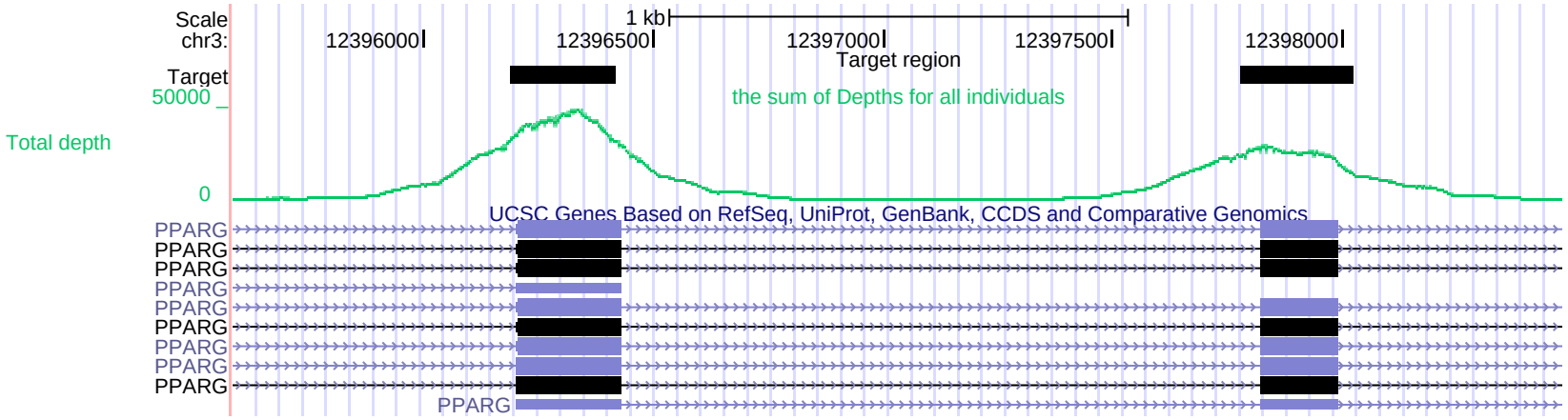
Yi et al. 2010

- The exomes of 50 Tibetan individuals at an average coverage of 18X.
- Compared to 40 Han Chinese individuals sequenced at an average of 6X (1000G).
- and 200 Danish exome sequenced individuals (8X)
- Estimated joint allele frequencies for each SNP using Bayesian approach.

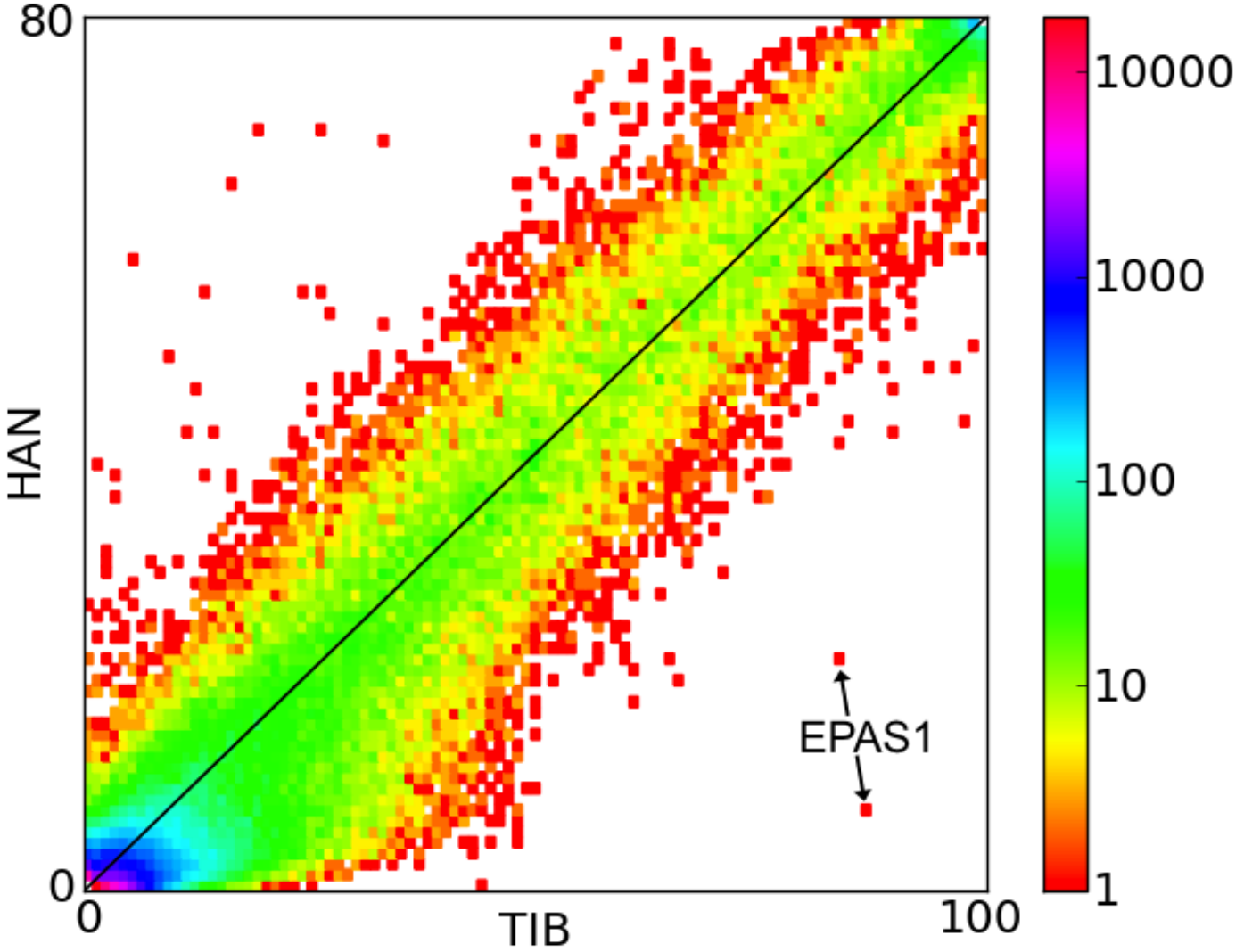
PPARG



PPARG - zoom

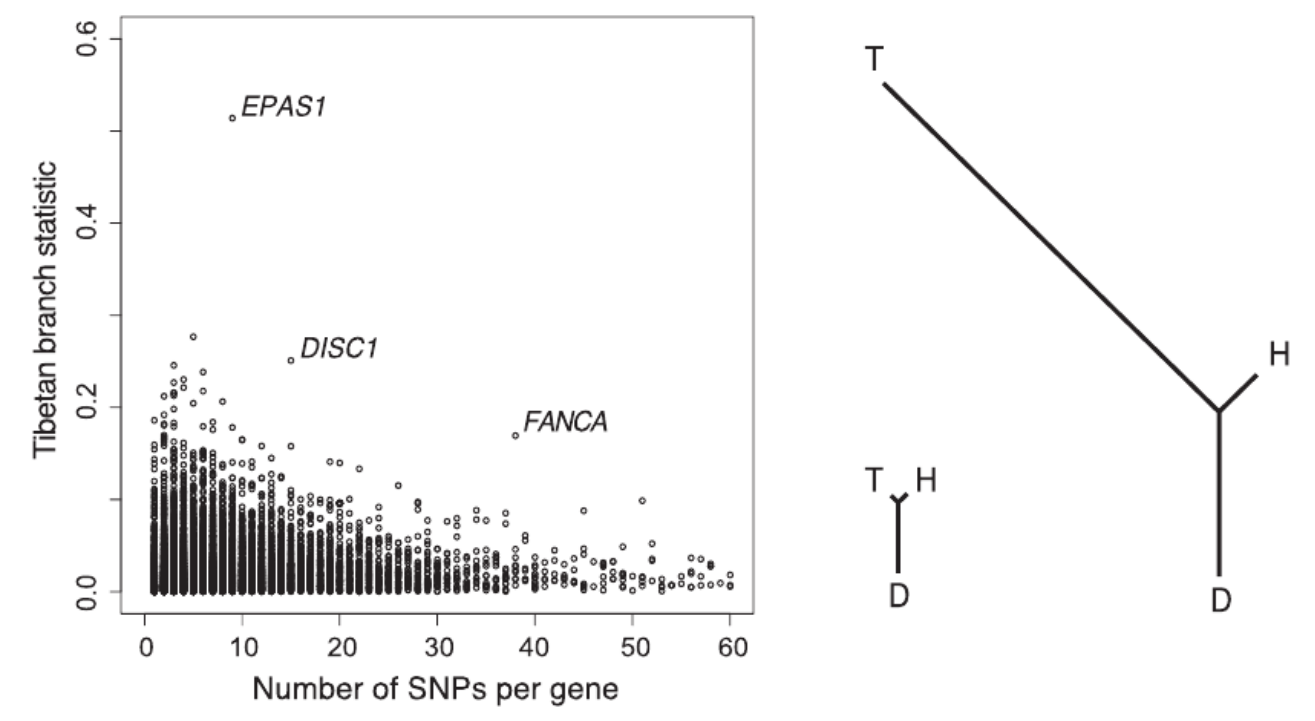


2D site frequency spectrum



Population Branch Statistic (PBS)

$$PBS = TBS = (T^{TH} + T^{TD} - T^{HD})/2, \qquad T^{AB} = log(1 - F_{st}^{AB})$$



Population frequencies

EPAS1 SNP allele frequencies

Allele	Tibetan	Han	Danish
C	0.13	0.9125	1
G	0.87	0.0875	0

EPAS1

- type of hypoxia-inducible factors
- active under low oxygen
- variant of gene confers increased athletic performance - called the "super athlete gene".

Genotyping in 366 individuals

Independent genotyping

- 366 Tibetans
- Genotyped for the EPAS1 SNP
- Phenotypes availeble

Associations within the Tibetan population

	CC	CG	GG	p-value
N	10	84	272	
Hemoglobin concentration	178	178.9	167.5	0.0013
erythrocyte counts	5.3	5.6	5.2	0.0015

Is this extreme compared to populations

Africans
1 Bantu
2 Mandenka
3 Yoruba
4 San
5 Mbuti pygmy
6 Biaka
7 Mozabite

Europeans
8 Orcadian
9 Adygei
10 Russian
11 Basque
12 French
13 North Italian
14 Sardinian
15 Tuscan

Western Asians
16 Bedouin
17 Druze
18 Palestinian

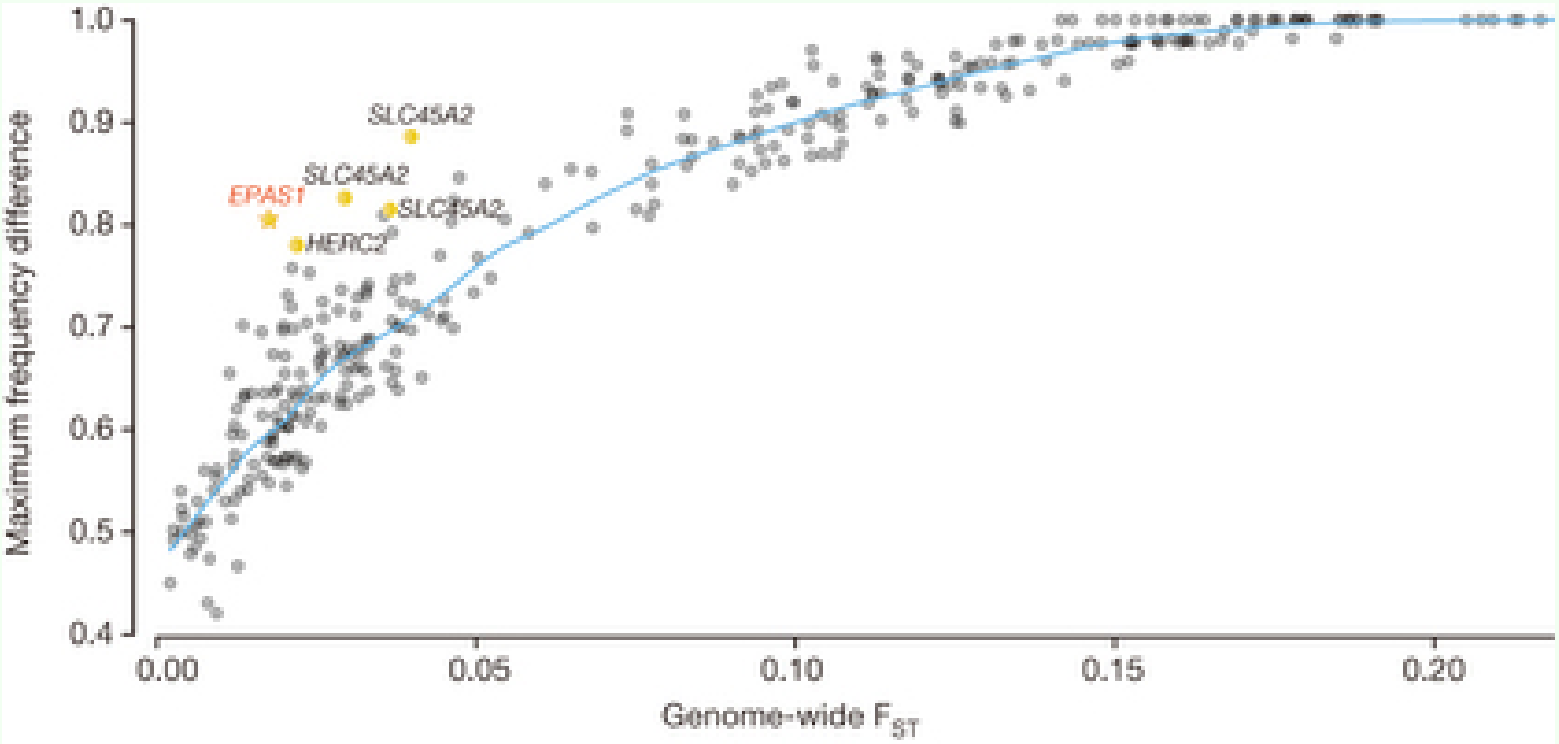
Central and Southern Asians
19 Balochi
20 Brahui
21 Makrani
22 Sindhi
23 Pathan
24 Burusho
25 Hazara
26 Uygur

Eastern Asians
28 Han (S. China)
29 Han (N. China)
30 Dai
31 Daur
32 Hezhen
33 Lahu
34 Miao
35 Oroqen
36 She
37 Tujia
38 Tu
39 Xibo
40 Yi
41 Mongola
42 Naxi

Oceanians
46 Melanesian
47 Papuan

Native Americans
48 Karitiana
49 Surui
50 Colombian
51 Maya
52 Pima

Other genes with large FST



conclusion

- Tibetans have adapted to life in high altitude
- A loci EPAS1 was found that has undergone strong adaptive selection
- The loci associated with hemoglobin concentrations and erythrocyte counts
- Followup study (Huerta-Snchez *et al* 2014) showed that
  - The mutations were introduced by Denisovan introgression
  - Example of adaptive introgression in human

Human adaption to arctic environment

Brief overview of Greenland's history

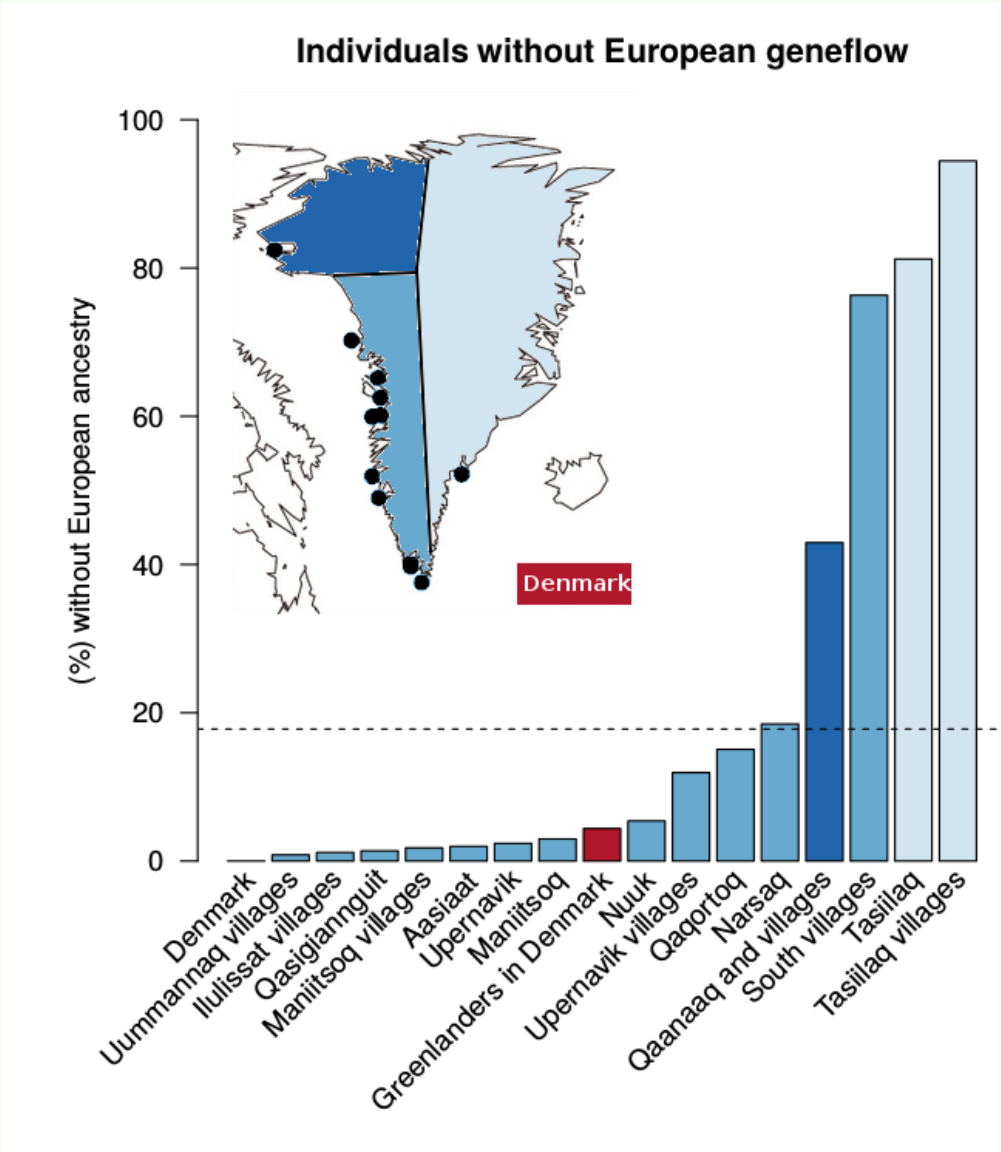
- Inhabited on and off by different Arctic cultures for ∼4500 years:
- 
- Visited by Vikings, Danish colony from 1814, now autonomous country
- 

The modern Greenlandic population

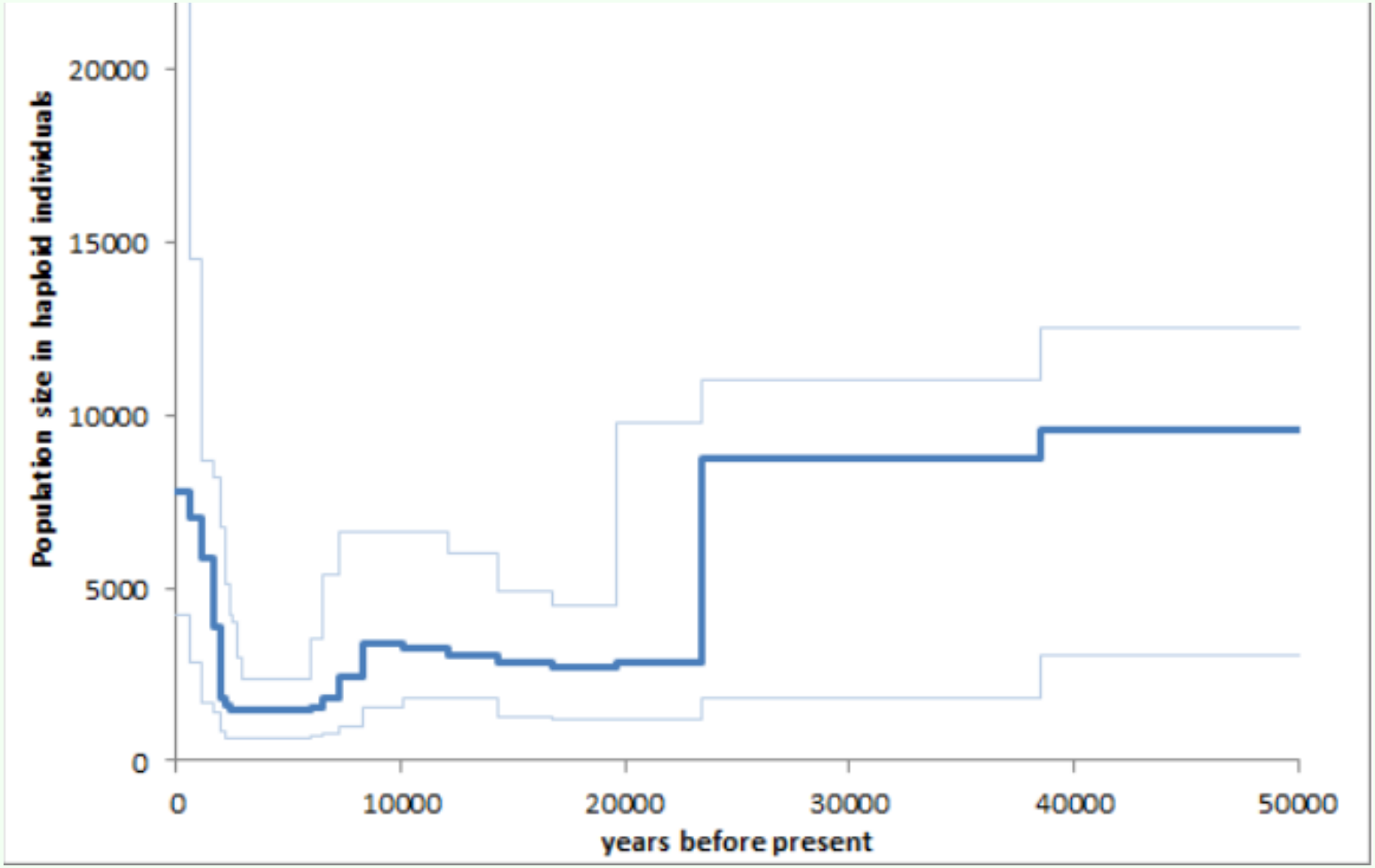
- Small: N≃57,000
- Live in coastal towns
- Descendents of Inuit

- But most also have European ancestry
- On average $\sim$ 25%

From Moltke et al. 2014

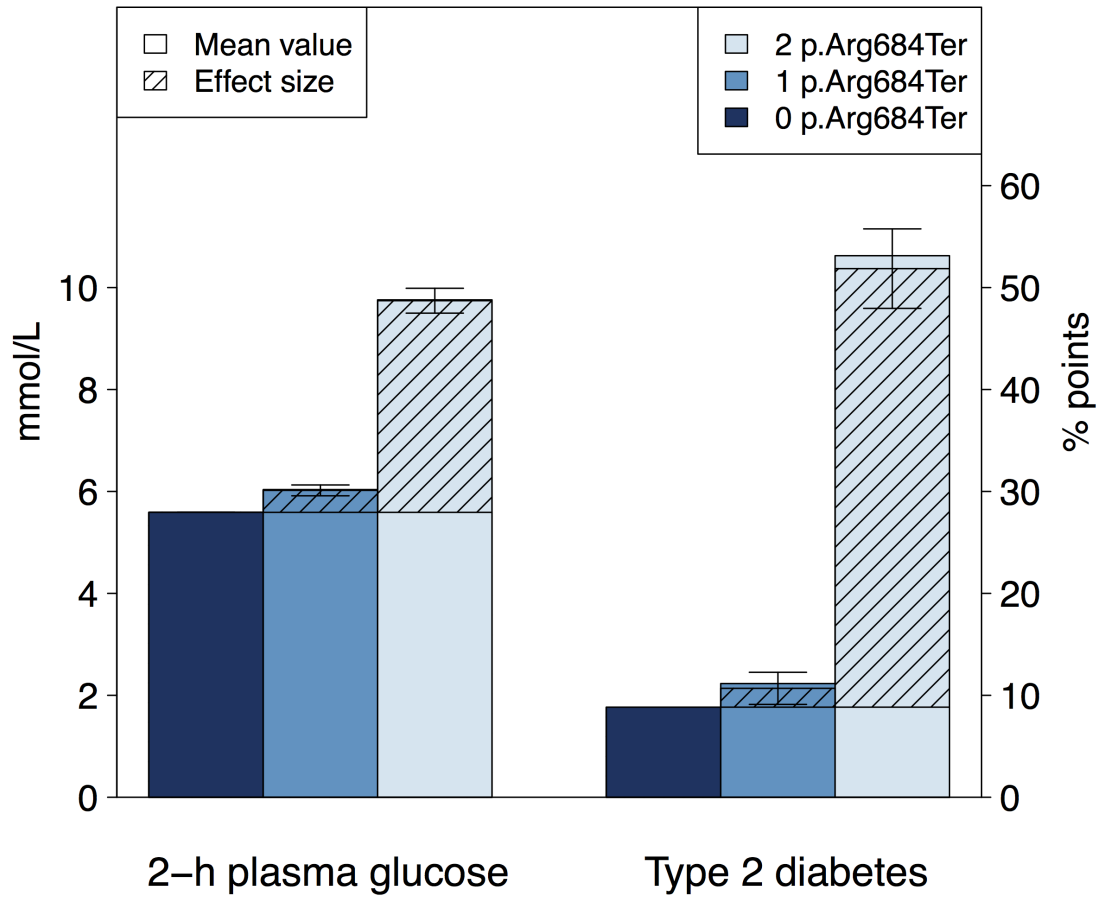


Recent changes in population size



A mutation causes 15% of type 2 diabetes in Greenland[1]

Very large almost recessive effect



Rec model

2-h Glucose:3.8mmol/l
T2D: OR=10.3

heredibility

The variation explain 15% of all T2D in Greenland

[1] Moltke *et al.* 2014

Life in the Arctic is extreme: cold temperatures & fat-rich diet

Questions we recently tried to answer

Long term history

Who are the ancestors of the Inuit and Greenlanders?

Recent history

How do modern Greenlanders relate to each other and Europe

Disease and selective pressure

Effect of being a small population - can we identify the genetic basis

Adaptation

How did the Inuit adapt to the extreme environment

Effect of being a small and isolated population

Allele frequencies

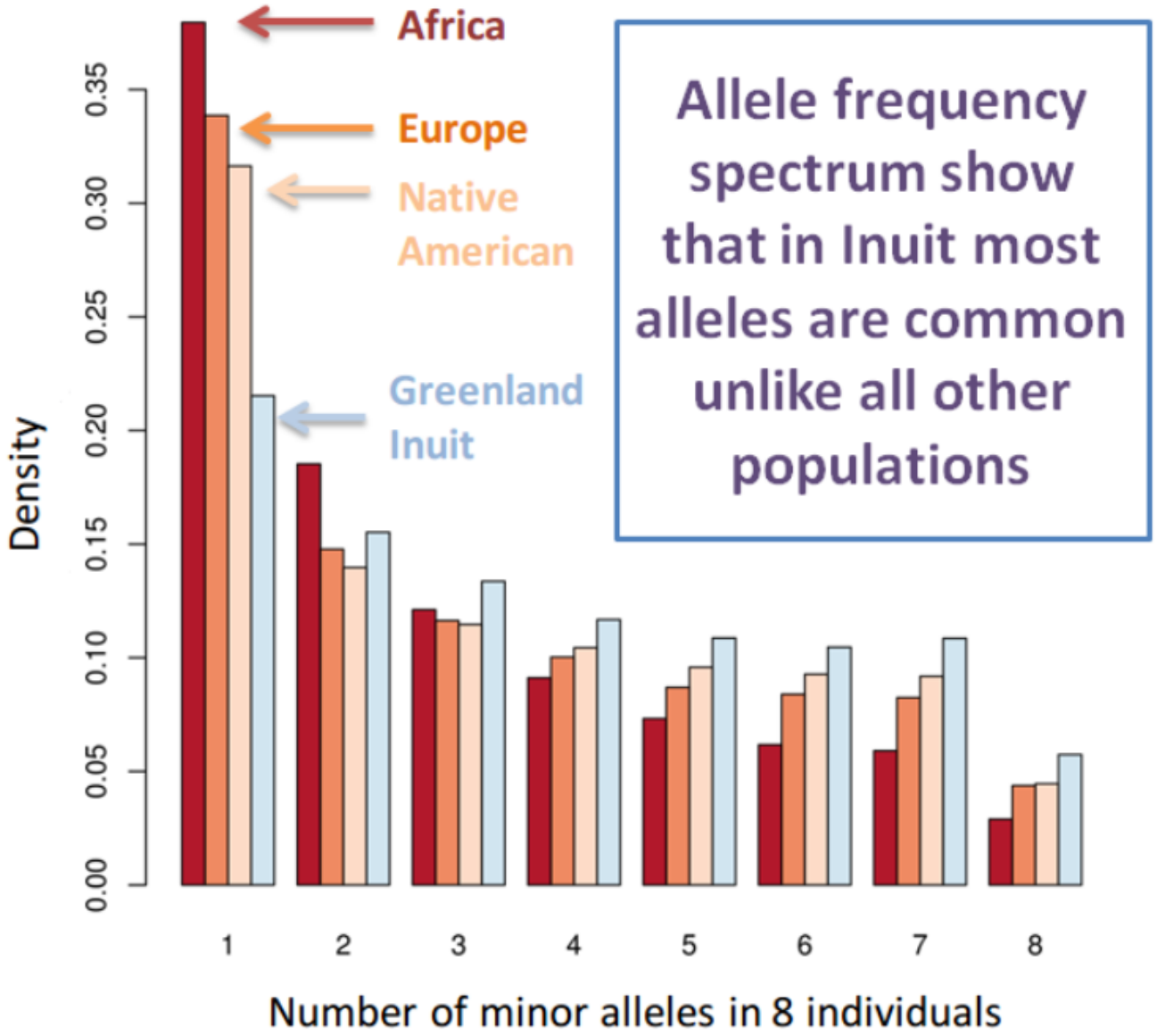
drift
- By far the most important factor
- Stronger effect in small populations

selection
- Important for alleles with phenotypic effect
- For small populations only alleles under very strong selection will be significantly affected

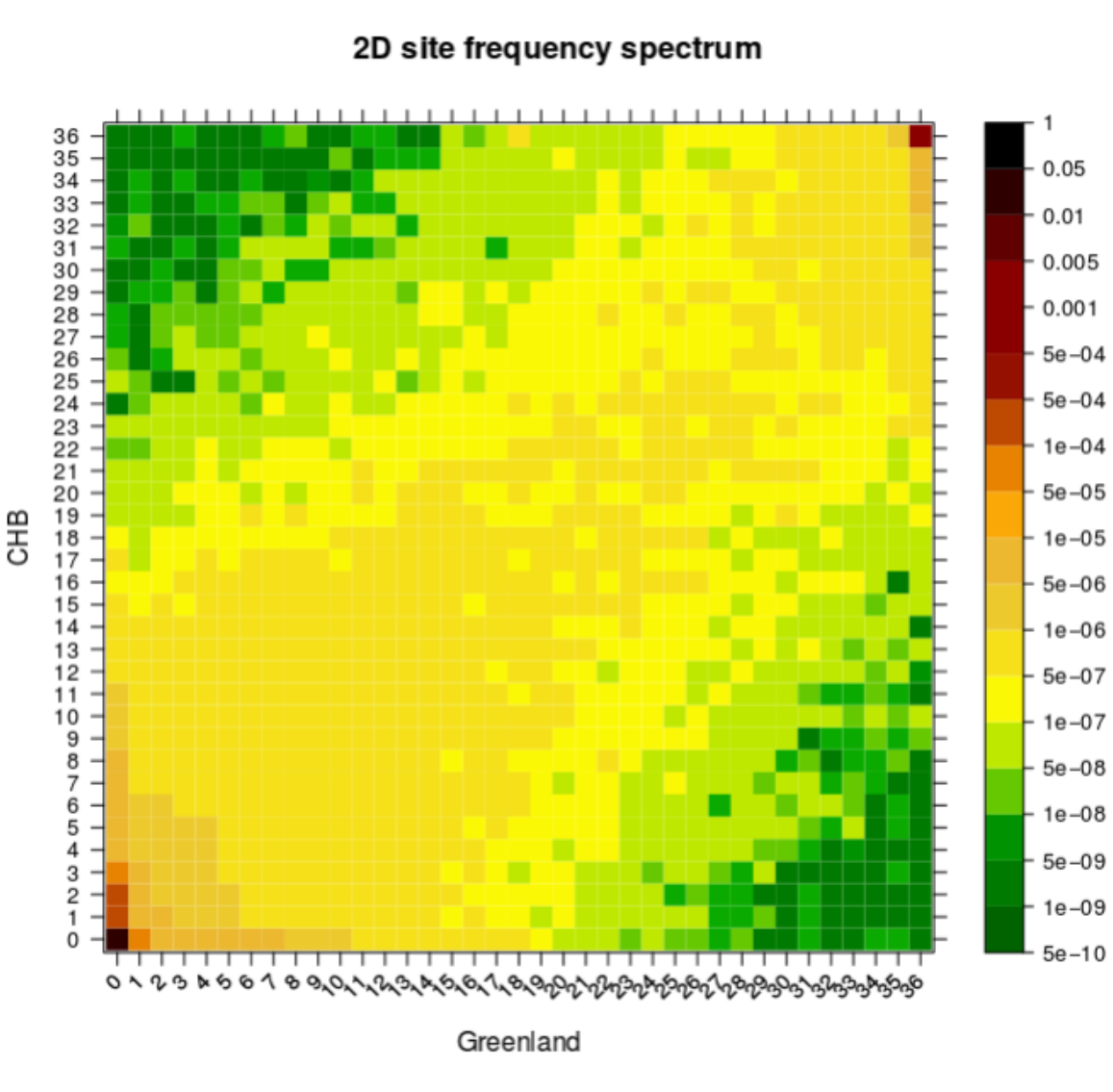
causal loci
- loci with a strong effect will be at very low frequency in large populations
- loci with a strong effect can have a large frequency in small populations

all loci Allele frequencies will differ from all other populations

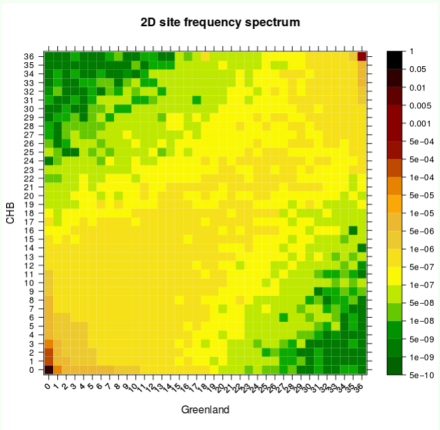
Frequency spectrum of Inuit



2D SFS between GL and Han

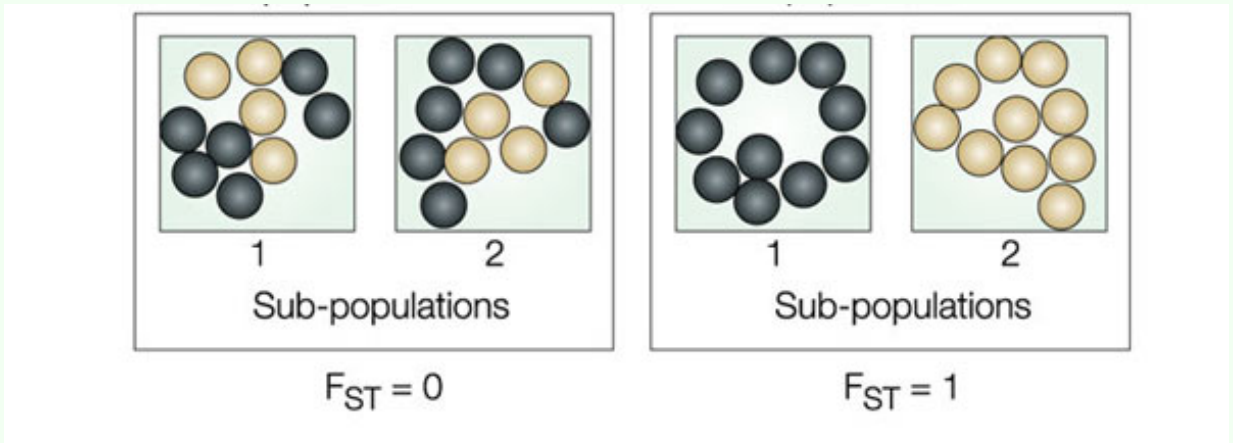


2D SFS and Fst

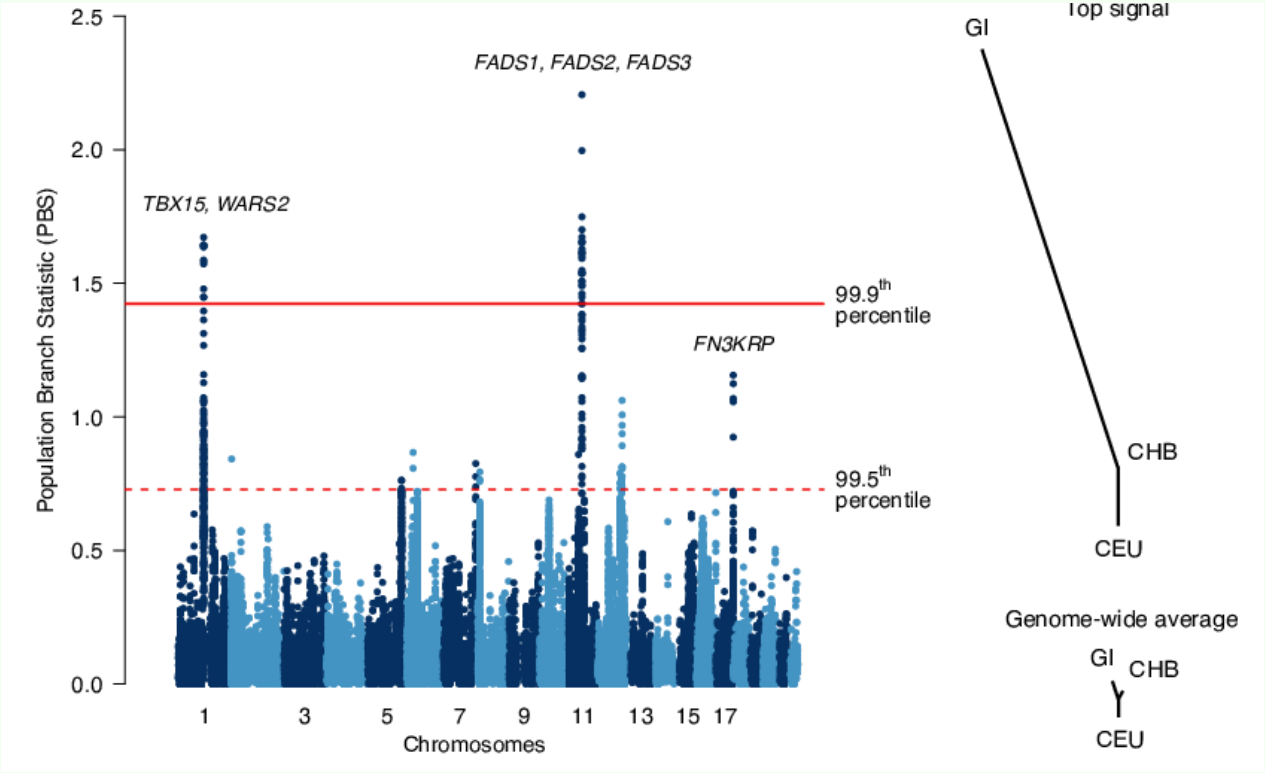


Fst from heterozygosity

$$F_{st} = \frac{\sigma_B}{\sigma_T} = \frac{H_{total} - H_{subpolulations}}{H_{total}}$$



Selection scan using PBS - ((HAN, GR) CEU)



Top loci

FADS

fatty acid desaturase.

TBX15

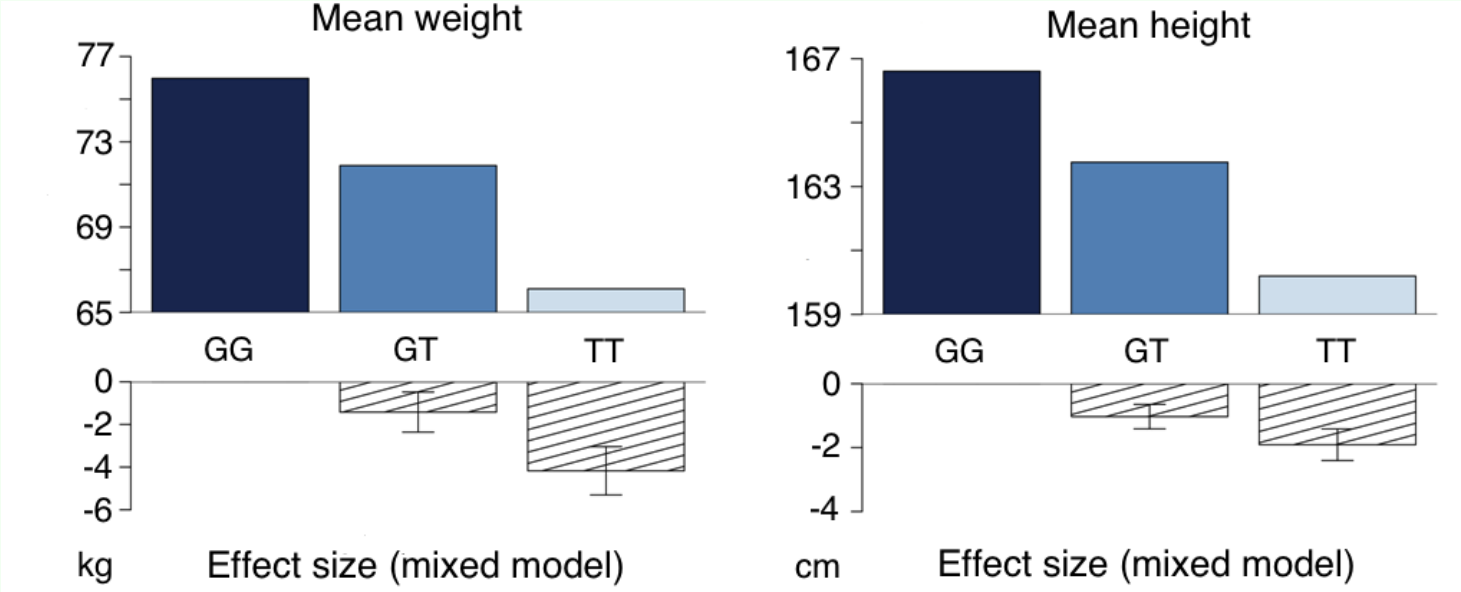
- TBX15 plays an important role in differentiation of brown (subcutaneous) adipocytes.
- Upon stimulation by cold exposure can produces heat by lipid oxidation.

FN3KRP

- an enzyme that catalyzes fructosamines, psicosamines and ribulosamines that protects against nonenzymatic glycation.
- FN3KRP can act to counteract the negative fitness caused by a PUFA rich diet.

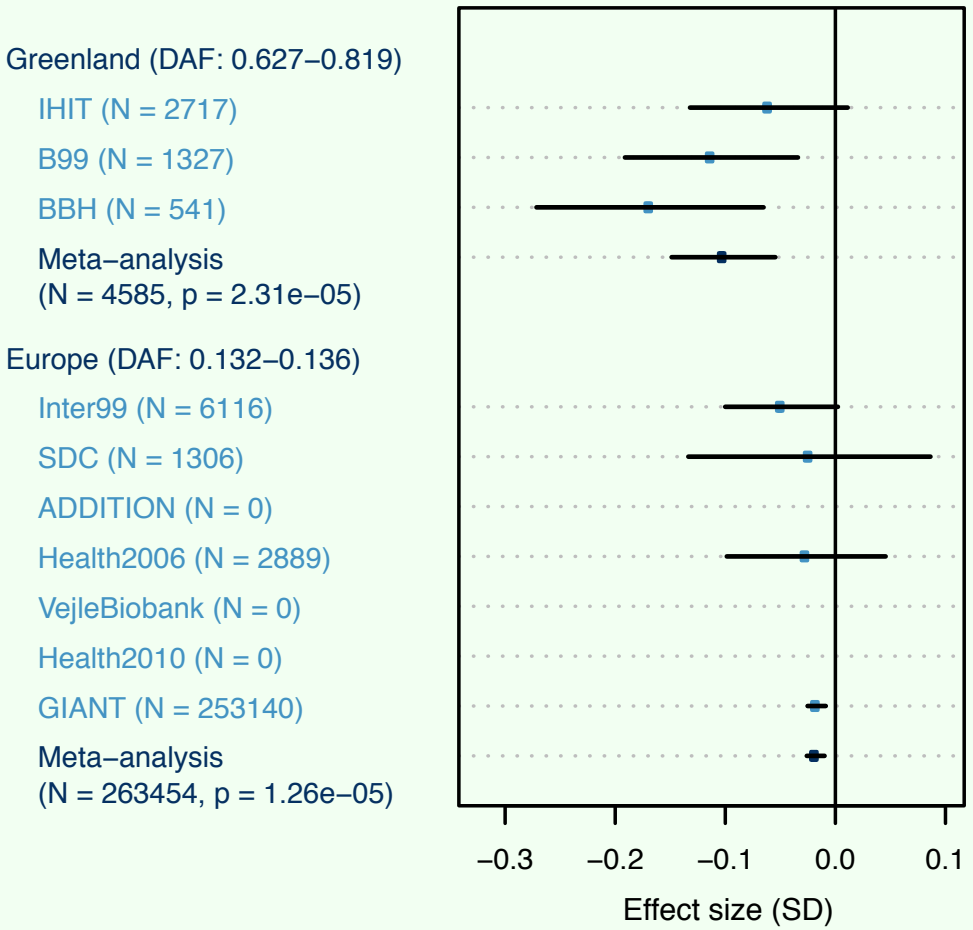
Why selection?

- Tested for association between top SNPs and metabolic traits
- Marginally significant associations with multiple traits, including LDL
- Selected alleles associated with decreased weight and height:



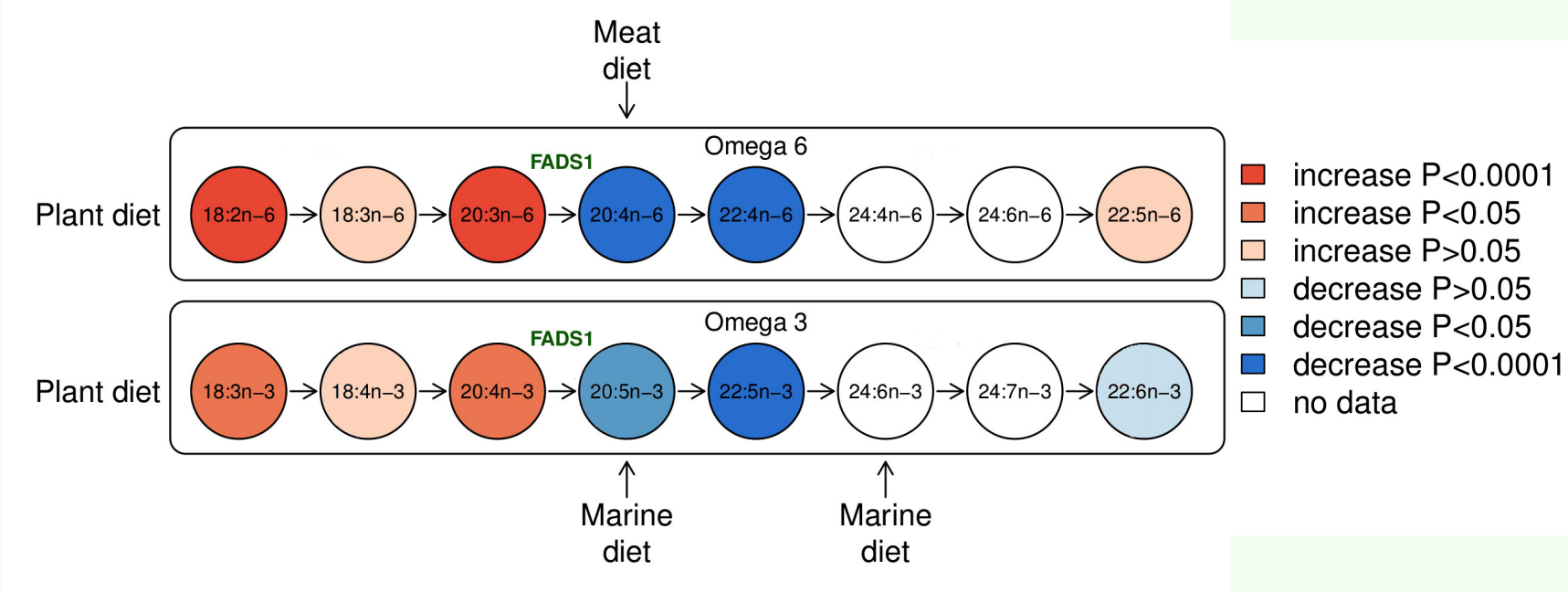
Why selection?

- The association with height replicates in Europe:



Why selection? Take 2

- Testing for association w. red blood cell membrane fatty acid composition:



- Mutation seems to compensate for high-fat diet
- Height due to effect of fatty acid composition on growth hormone levels?
- Either way, the results suggest that selection in this region is a new example of human adaptation where we know the genetic basis

Conclusion

- We find multiple interesting loci which some evidence of recent adaptation to life in the arctic
- As expected the genes are involved in poly unsaturated fatty acid metabolism and cold adaption
- Surprisingly the loci also affects high and weight
- variants also have an effect in on height in Europe

How are the SFS estimated?

With high depth sequencing

simple counts of derived alleles

Can we construct the SFS using low/medium sequencing

Yes - maybe - use genotype likelihoods and be careful

When can calling SNPs and genotypes be a problem?

low/medium depth data

- Capture data
- low depth sequencing due to price
- ancient DNA (only a finite amount of DNA)

What depth is high enough?

Depends on the analysis. e.g.

- SFS is extremely sensitive to both genotype and SNP calling
- admixture proportions are sensitive to genotype calling
- ABBA-BABA (D-stats) can be used regardless of depth

Estimating SFS while taking uncertainty of data into account

Likelihood of SFS for a single site:[a]

[a] fast calculations with dynamic programming (Nielsen et al. 2012)

$$P(X^s \mid \eta) = \sum_{j=0}^{2N} p(X^s \mid J = j) p(J = j \mid \eta)$$

SFS for a region

$$P(X \mid \eta) = \prod_{s=1}^{r} P(X^s \mid \eta)$$

Estimating SFS while taking uncertainty of data into account

Likelihood of SFS for a single site:[a]

[a] fast calculations with dynamic programming (Nielsen et al. 2012)

$$P(X^s \mid \eta) = \sum_{j=0}^{2N} p(X^s \mid J = j) p(J = j \mid \eta)$$

$$\propto \sum_{j=0}^{2N} \eta_j \sum_{g \in \{0,1,2\}^N} p(G = g \mid J = j) \prod_{i=1}^{N} P(X_i^s \mid G_i = g_i),$$

$p(G = g \mid J = j)$

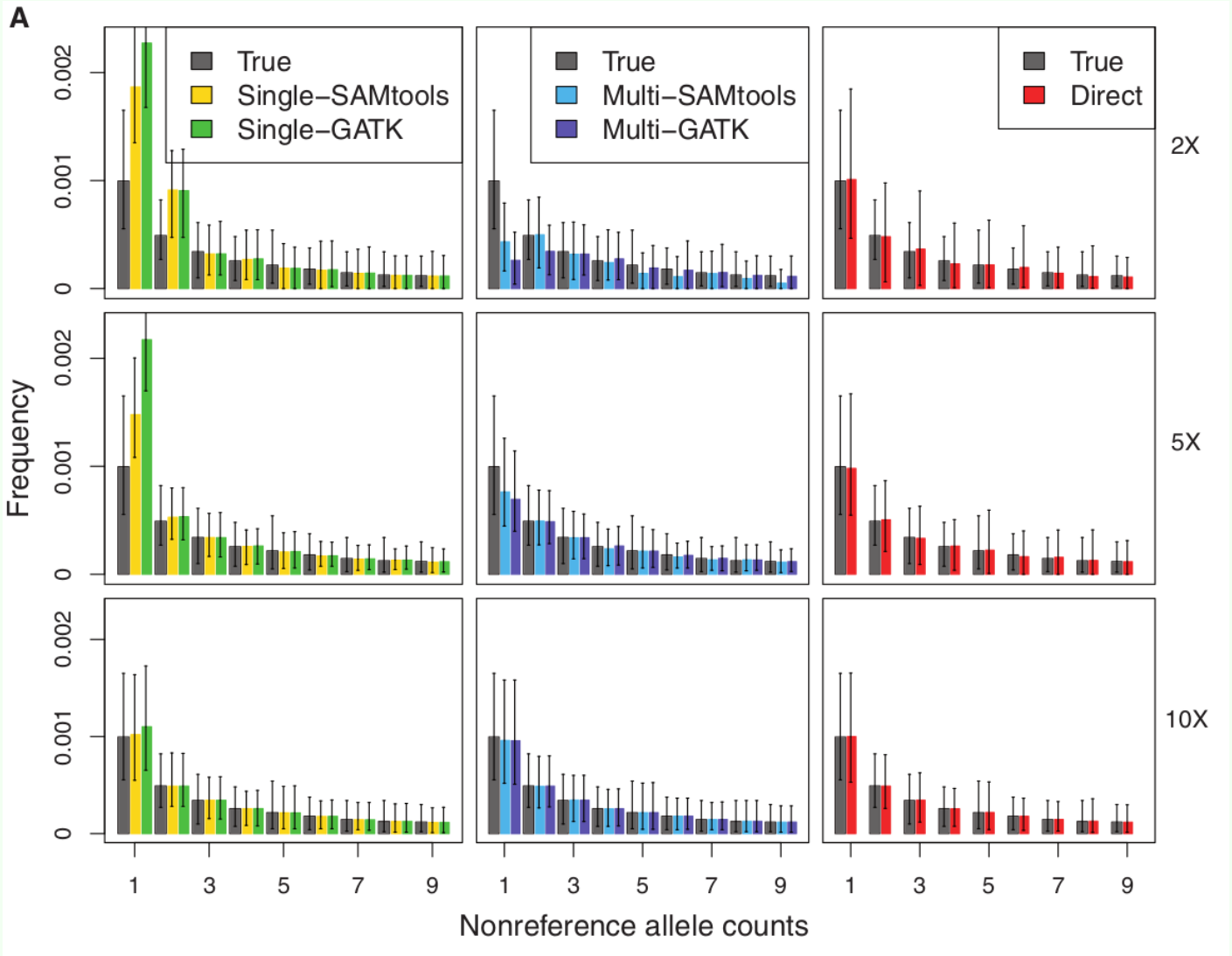
$p(G = g \mid J = j) = \binom{2N}{j} 2^{\sum_i^N I_1(g_i)}$

when $\sum_{i=1}^{2N} g_i = j$, else 0

SFS for a region

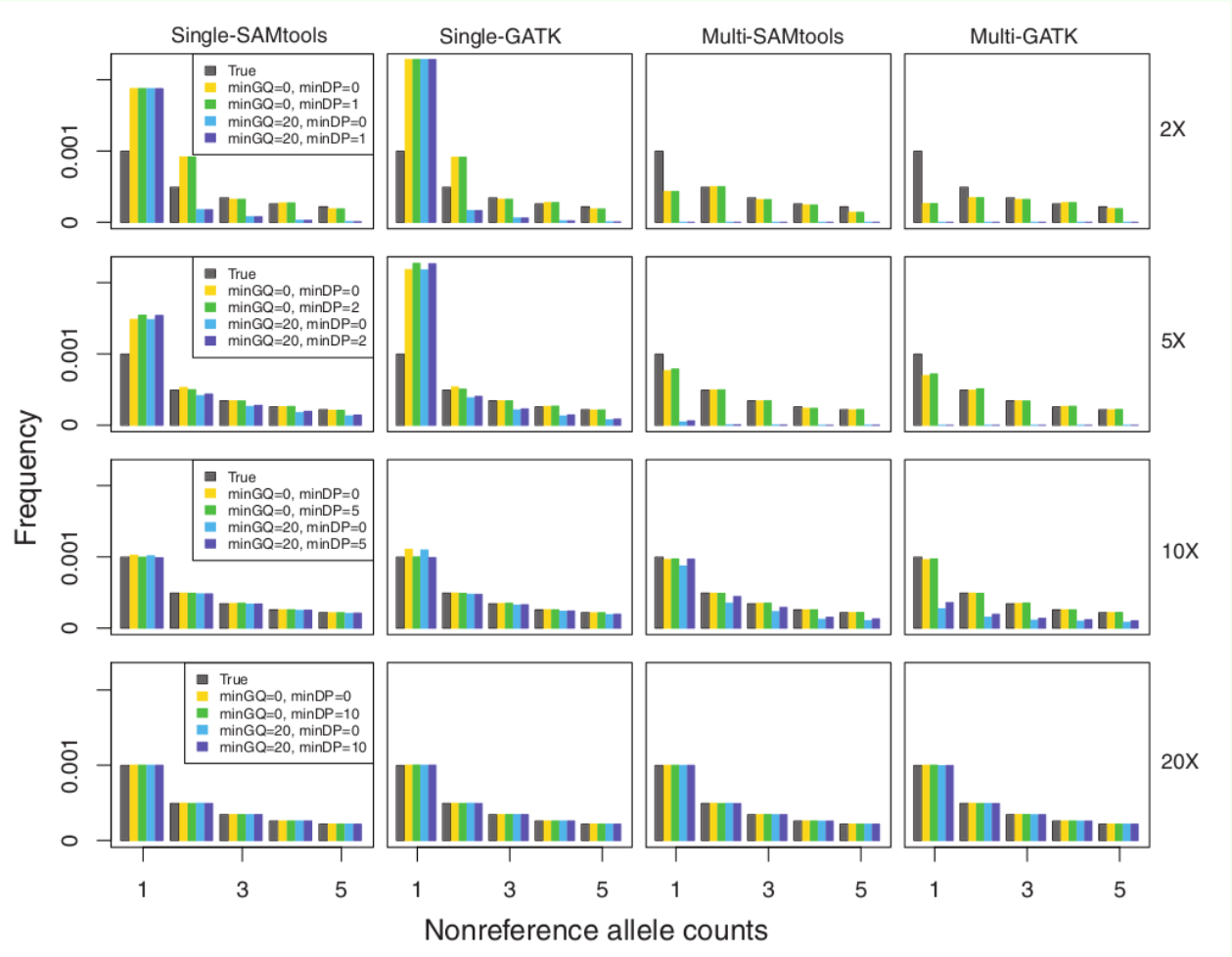
$P(X \mid \eta) = \prod_{s=1}^{r} P(X^s \mid \eta)$

Site frequency spectrum for low/medium depth data[2]



[2]E Han et al 2013

Filters do not solve the problem



3

[3] E Han et al 2013

Conclusion on SFS based on genotype likelihoods

- can be estimated even with low(ish) depth e.g. 2 X
- We use genotype likelihoods unless depth is high (>10X) unless you have other information
- Can be done in multiple dimension
  - 1D thetas e.g. Tajimas pi, Tajimas D, Population sizes
  - 2D f_{st} and PBS
  - XD usefull for Demography inference

Thank you for listening

Questions?