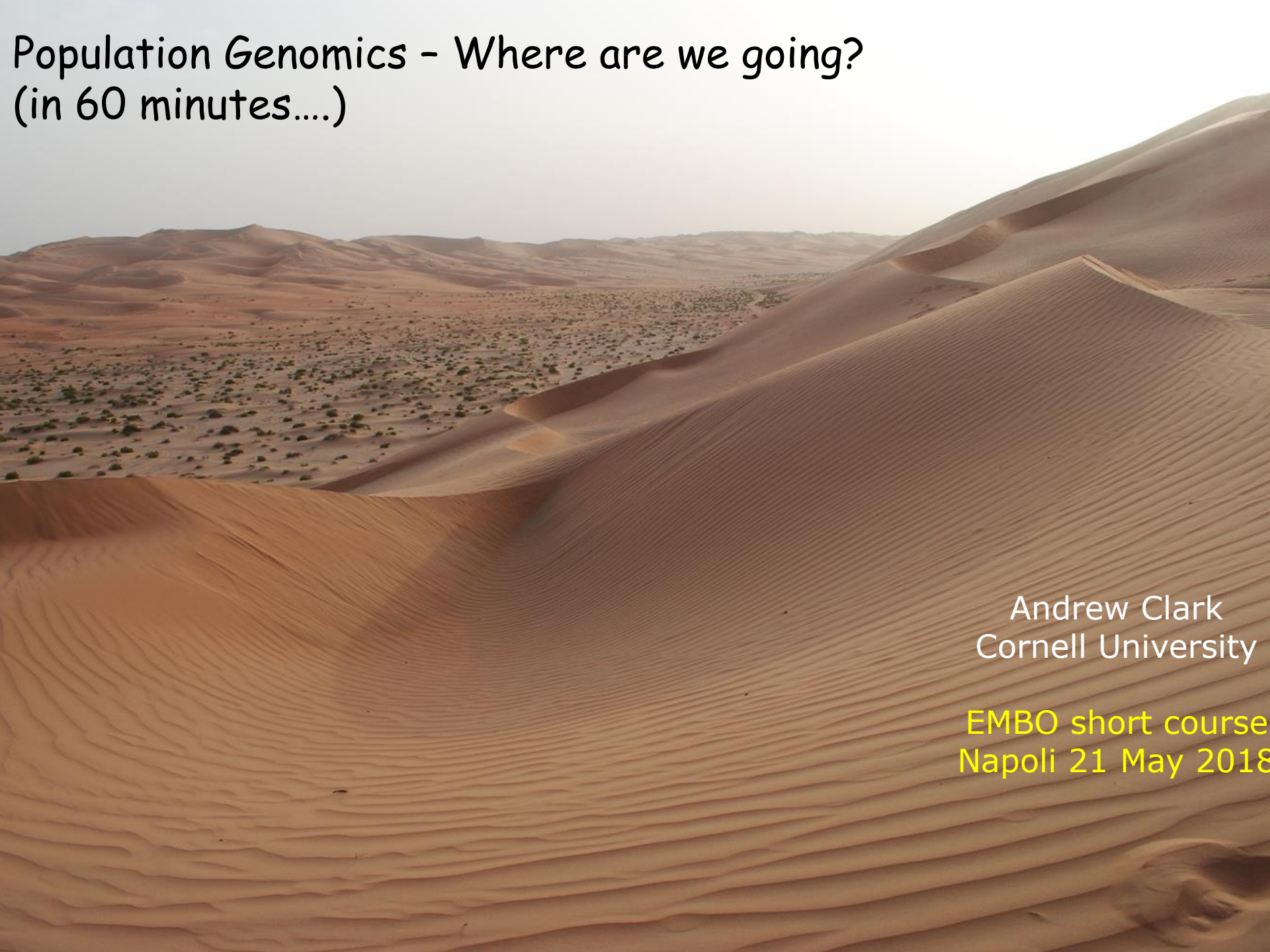
# Population Genomics – Where are we going? (in 60 minutes….)



Andrew Clark
Cornell University

EMBO short course
Napoli 21 May 2018

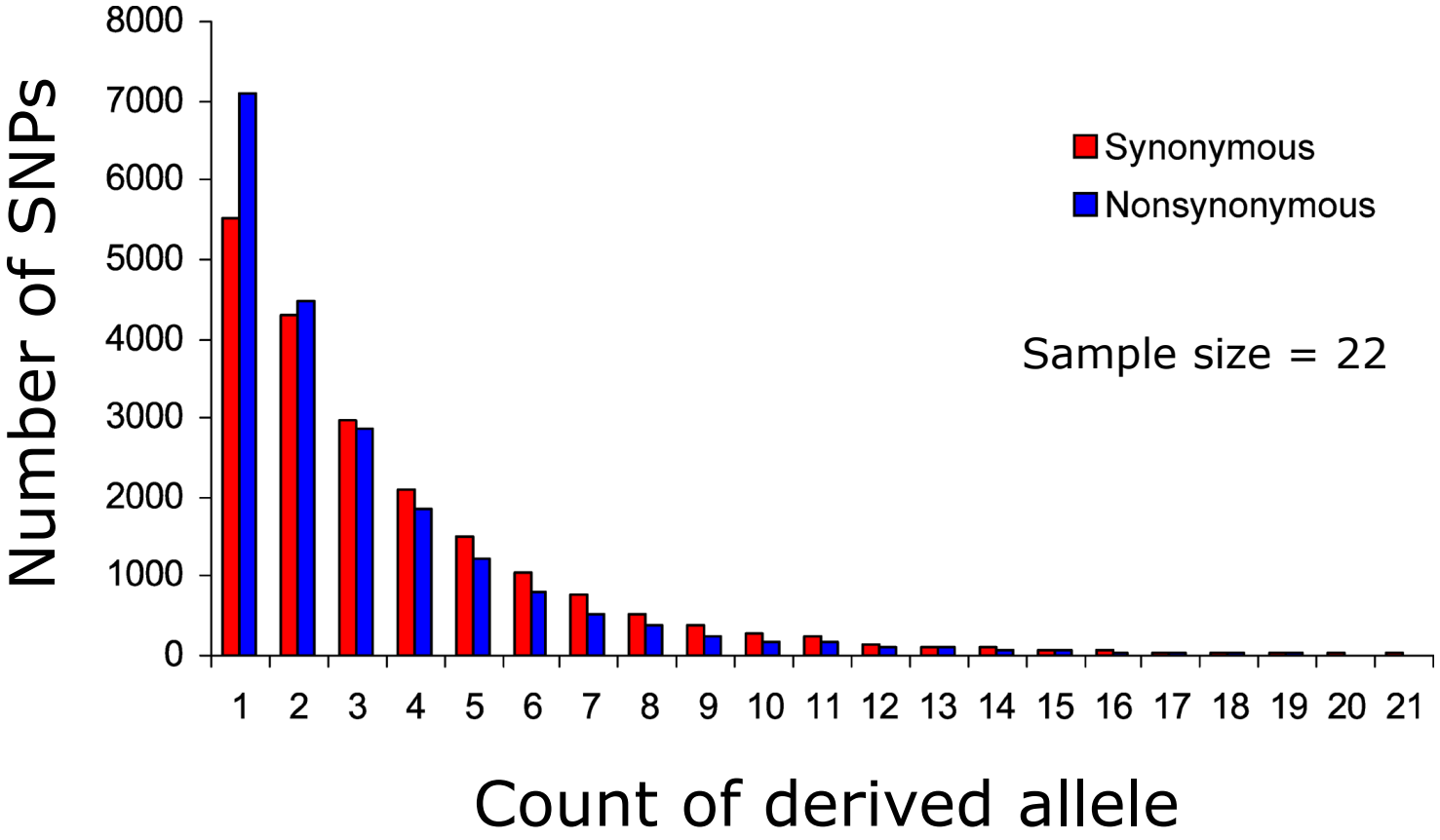
# Outline

- Demographic inference
- Population structure and history
- Admixture / Introgression
- Random genetic drift
- Natural selection
- Mutation spectrum
- Disease association
- Genome function

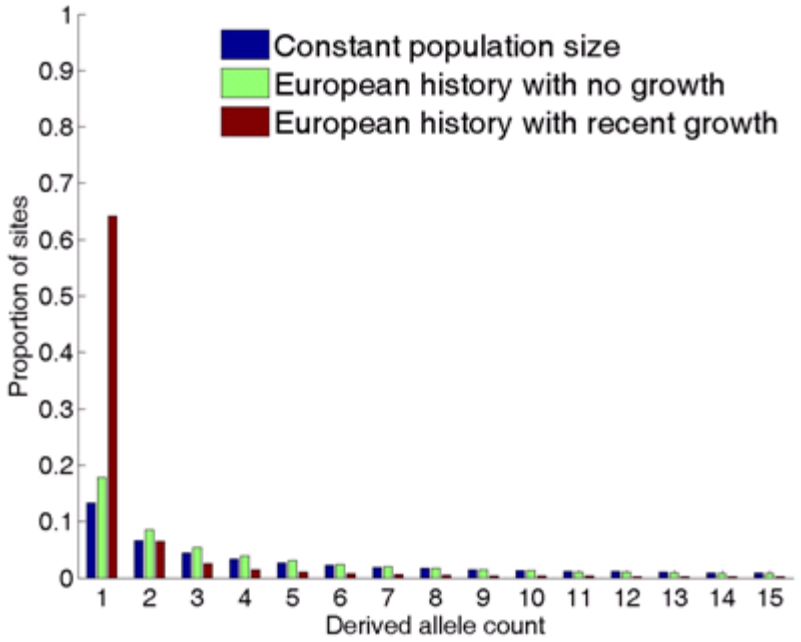
# DEMOGRAPHY

How can we infer past changes in population size, bottlenecks, etc.?

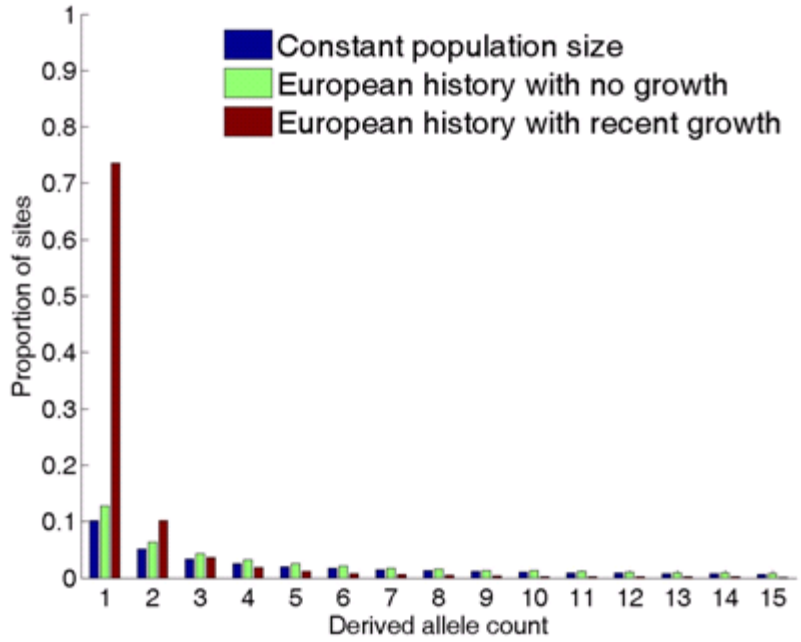
# The Site Frequency Spectrum



# Large samples are needed to see extent of skew to SFS



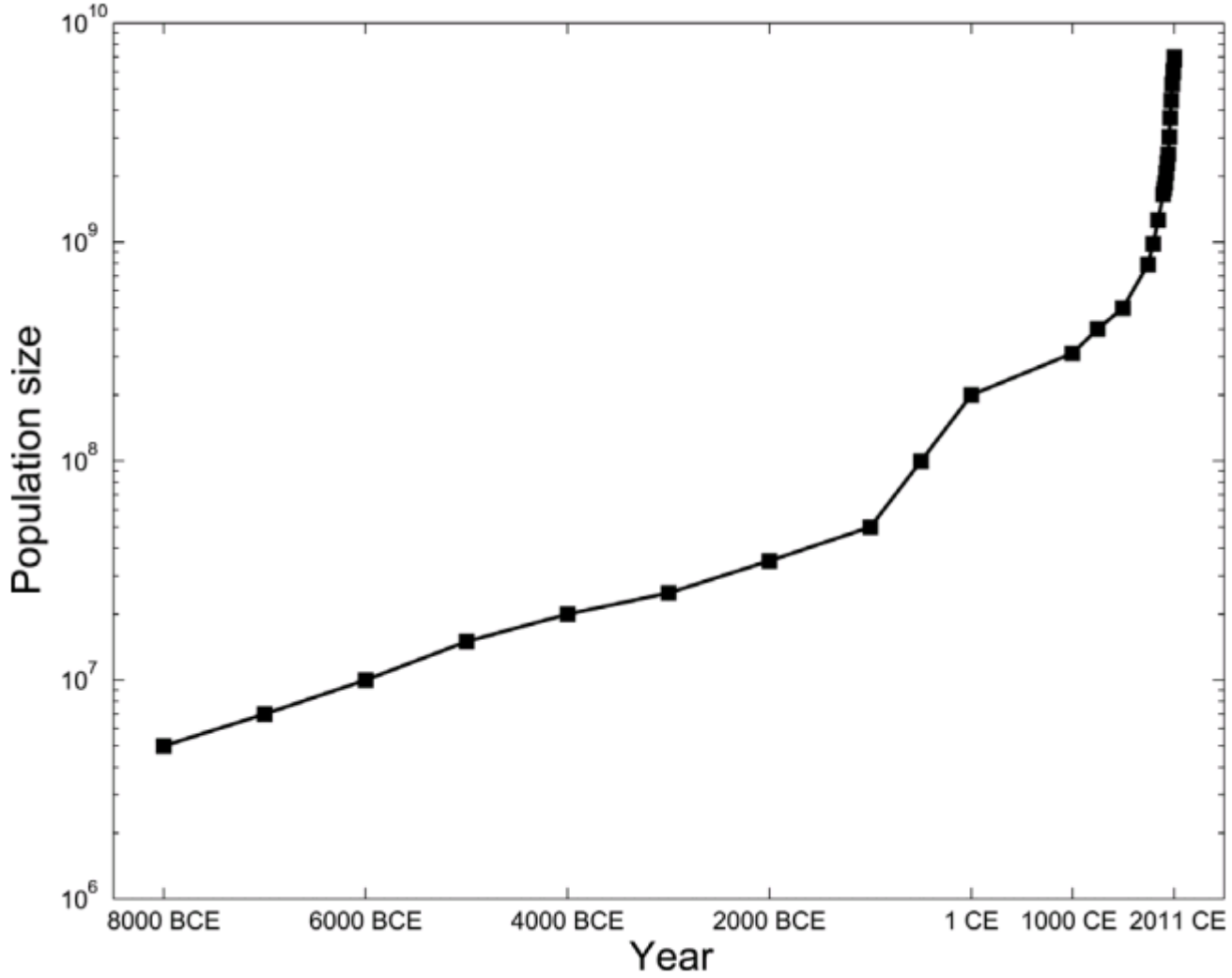

*n* = 1000




*n* = 10,000

Keinan & Clark 2012 *Science*

# The human population has grown super-exponentially




Cohen J 1995 *How Many People can the Earth Support*?

Science
AAAS

**An Abundance of Rare Functional Variants in 202 Drug Target Genes Sequenced in 14,002 People**

Matthew R. Nelson *et al.*

*Science* **337**, 100 (2012);
DOI: 10.1126/science.1217876

# An Abundance of Rare Functional Variants in 202 Drug Target Ge
# Sequenced in 14,002 Peo

Matthew R. Nelson,[1*†] Daniel Wegmann,[2*] M
Pamela St. Jean,[1] Claudio Verzilli,[3] Judo
Dana Fraser,[1] Liling Warren,[1] Jenni
Yong Zhang,[6] Jun Li,[7] Yun Li
Keith Nangle,[1] Jun Wan
John C. Whittaker

nature
International weekly journal of science
doi:10.1038/nature11690

LETTER

## Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants

Wenqing Fu[1], Timothy D. O'Connor[1], Goo Jun[2], Hyun Min Kang[2], Goncalo Abecasis[2], Suzanne M. Leal[3], Stacey Gabriel[4], David Altshuler[4], Jay Shendure[1], Deborah A. Nickerson[1], Michael J. Bamshad[1,5], NHLBI Exome Sequencing Project*
& Joshua M. Akey[1]

Sci

# Evolut
# Rare C
# Sequenc

Jacob A. Tennessen,[1*] Al
imothy D. O'Connor,[1*] Wenqing Fu,[1]
Eimear E. Kenny,[3] Simon
icGee,[1] Ron Do,[4,5] Xiaoming Liu,[6] Goo Jun,[7]
Hyun Min Kang,[7] Daniel Jo
uzanne M. Leal,[9] Stacey Gabriel,[4] Mark J. Rieder,[1]
Goncalo Abecasis,[7] David Altshuler,[4] Deborah A. Nickerson,[1] Eric Boerwinkle,[6,10]
Shamil Sunyaev,[4,8] Carlos D. Bustamante,[3] Michael J. Bamshad,[1,2]‡ Joshua M. Akey,[1]‡
Broad GO, Seattle GO, on behalf of the NHLBI Exome Sequencing Project

# Implications of recent growth for complex disease studies

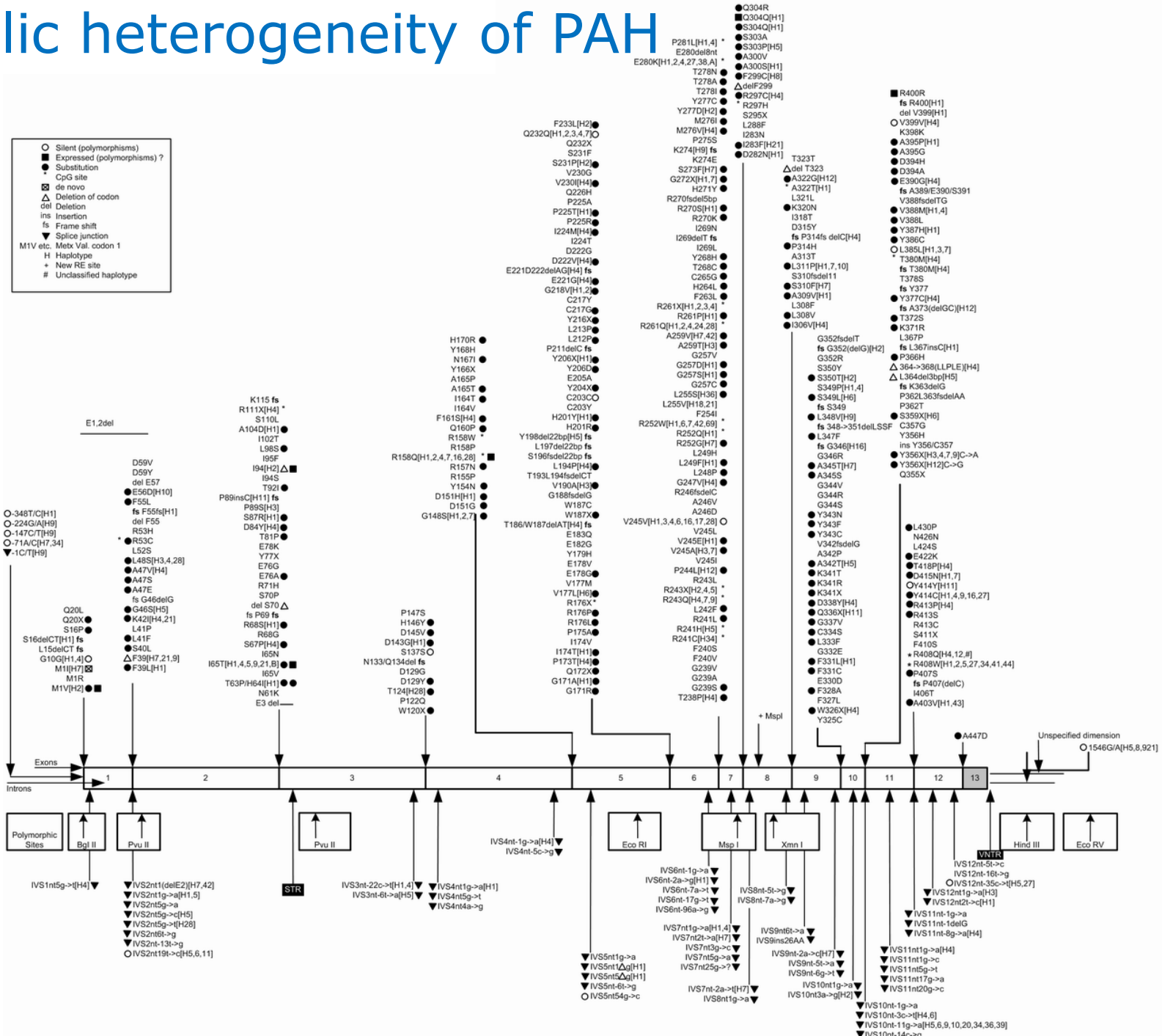
- Many more variants of all classes.
- Massive excess of rare variants.
- Rare variants are more likely to have deleterious effects.

# Implications of recent growth for complex disease studies

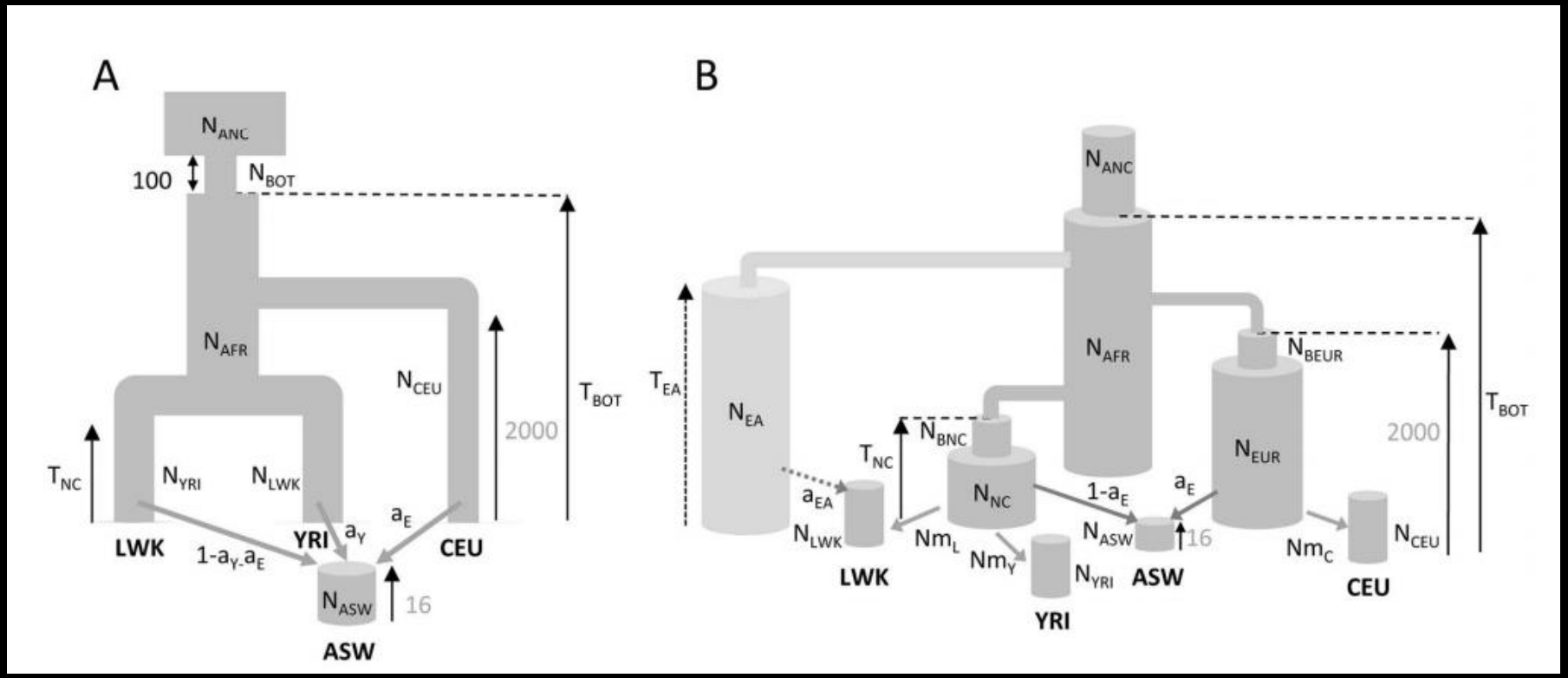
- Many more variants of [illegible]asses.
- Frequency shifted [illegible]er variants.
- Rare variant[illegible] more likely to have deleteri[illegible]ects.

**Genetic heterogeneity**

# Allelic heterogeneity of PAH

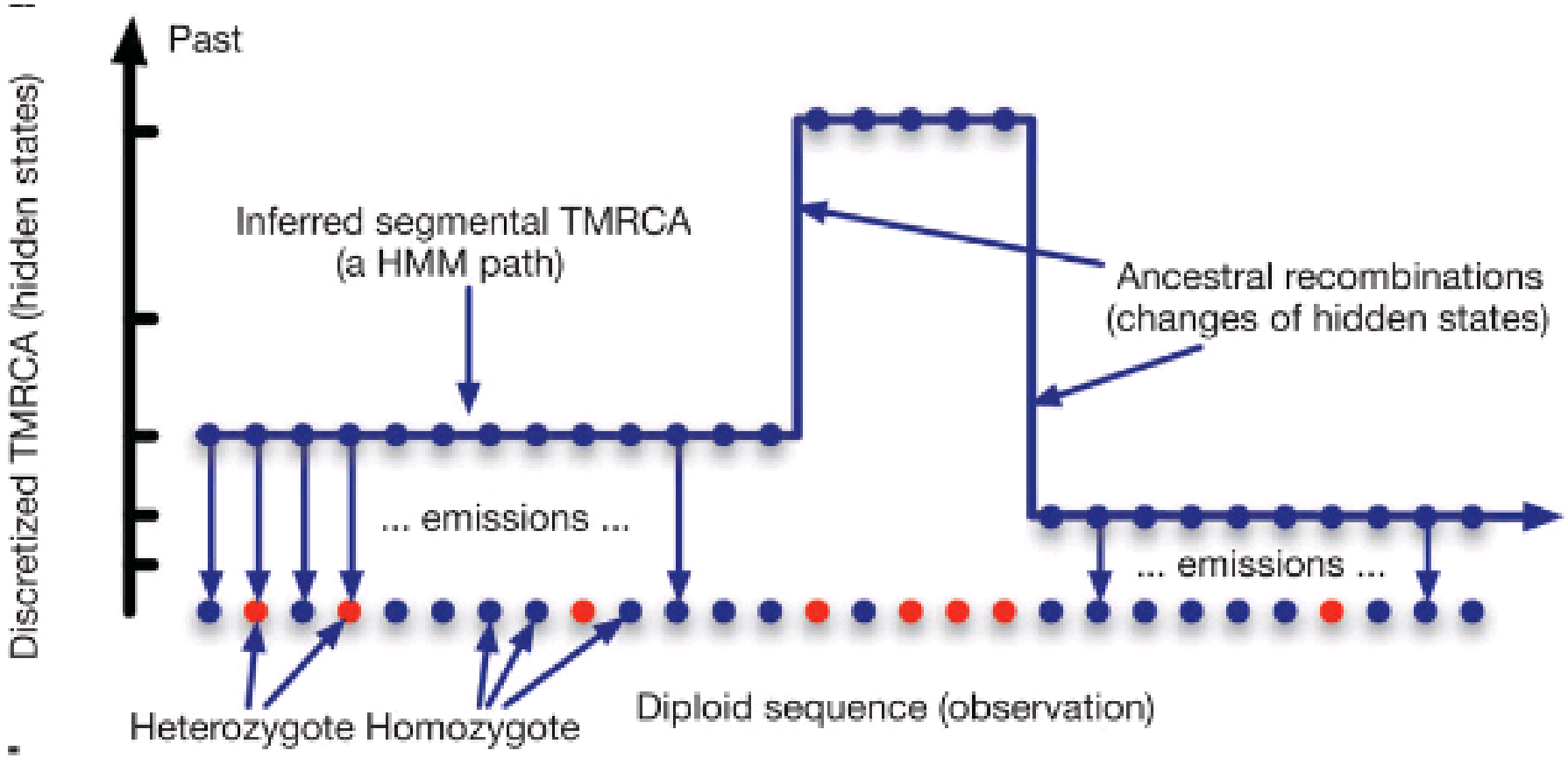


# Many methods for inference of demography: IM, dadi, fastsimcoal



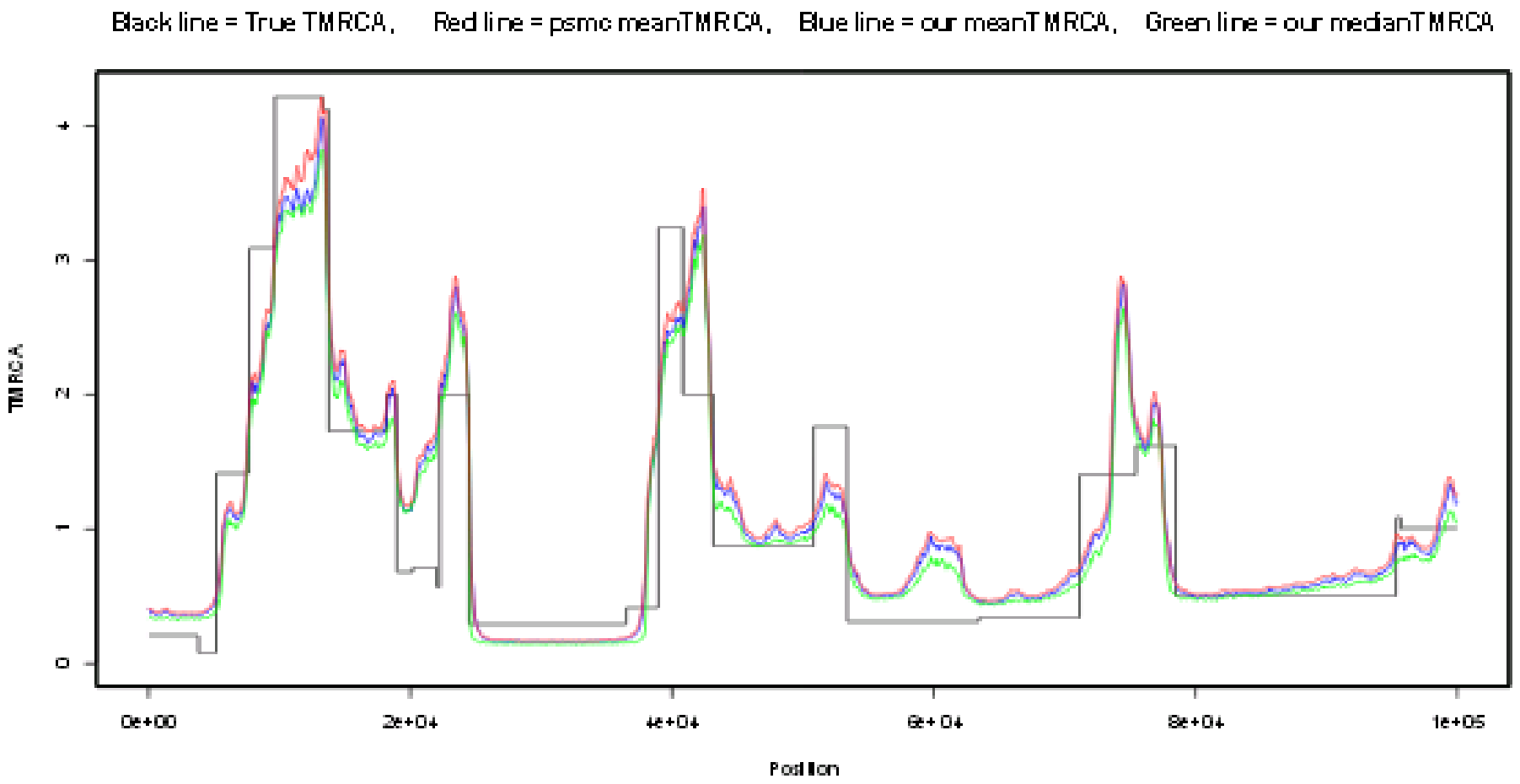

Excoffier *et al*. 2013 *PLoS Genetics*

# Inferring demography from a single individual: Pairwise Sequentially Markovian Coalescent



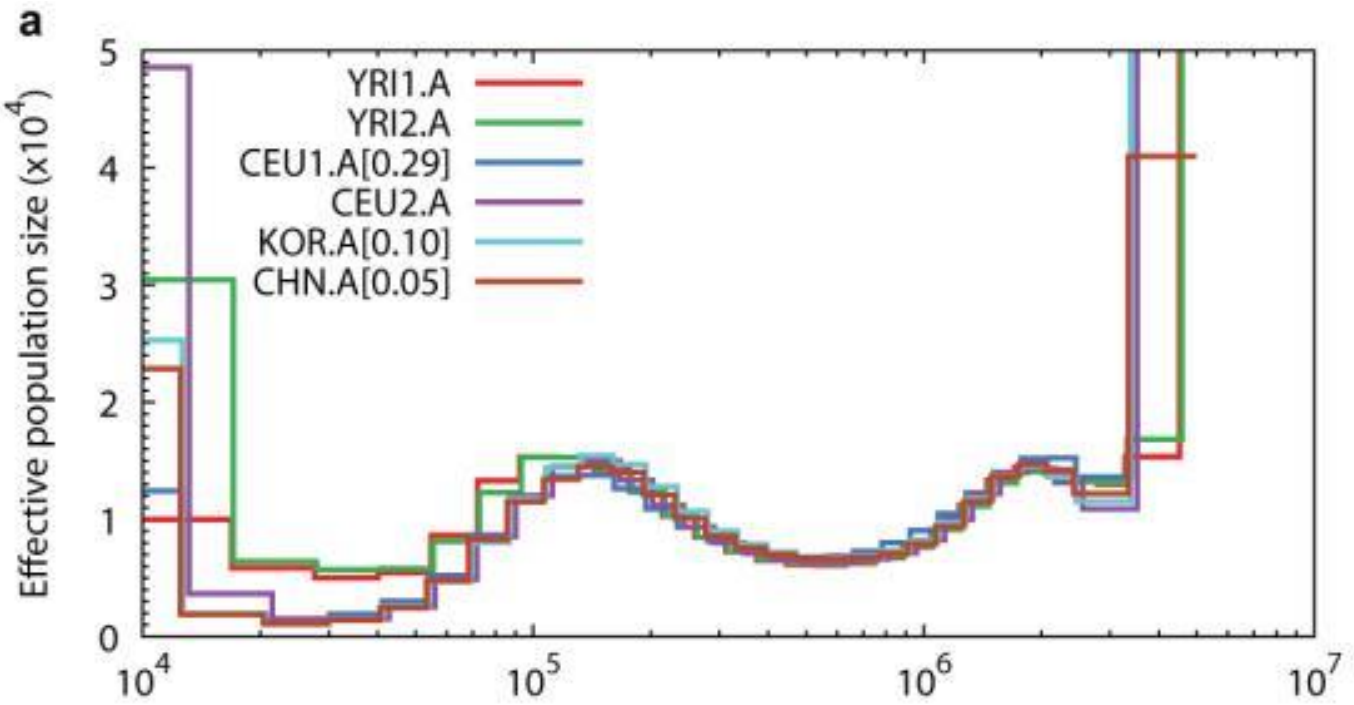

Li & Durbin 2011 *Nature*

# Simulations of past population demography shows reasonably good accuracy of PSMC, with an important caveat …..



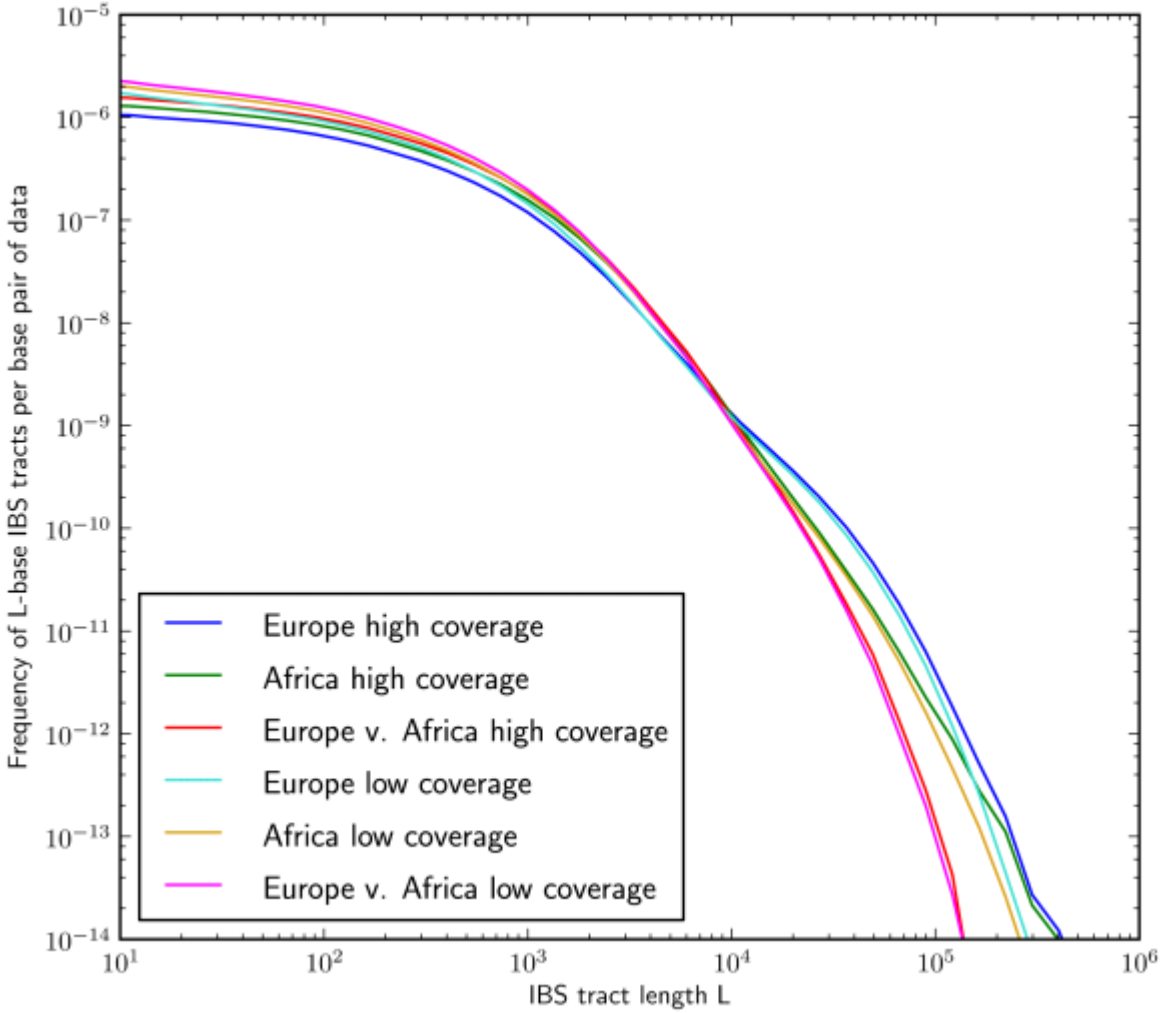

Li & Durbin 2011 *Nature*

# If the population has deme structure, and we sample from just one deme, we can spuriously infer a bottleneck !




Li & Durbin 2011 *Nature*

# Identity-by-Descent tracts for inference of demography




Harris & Nielsen 2013 *PLoS Genetics*

## DEMOGRAPHY (Population collapse)

What happens to genetic variation in genomes of populations that are crashing?

# Florida Scrub-Jay

(*Aphelocoma coerulescens*)

Cooperative breeder

Federally Threatened

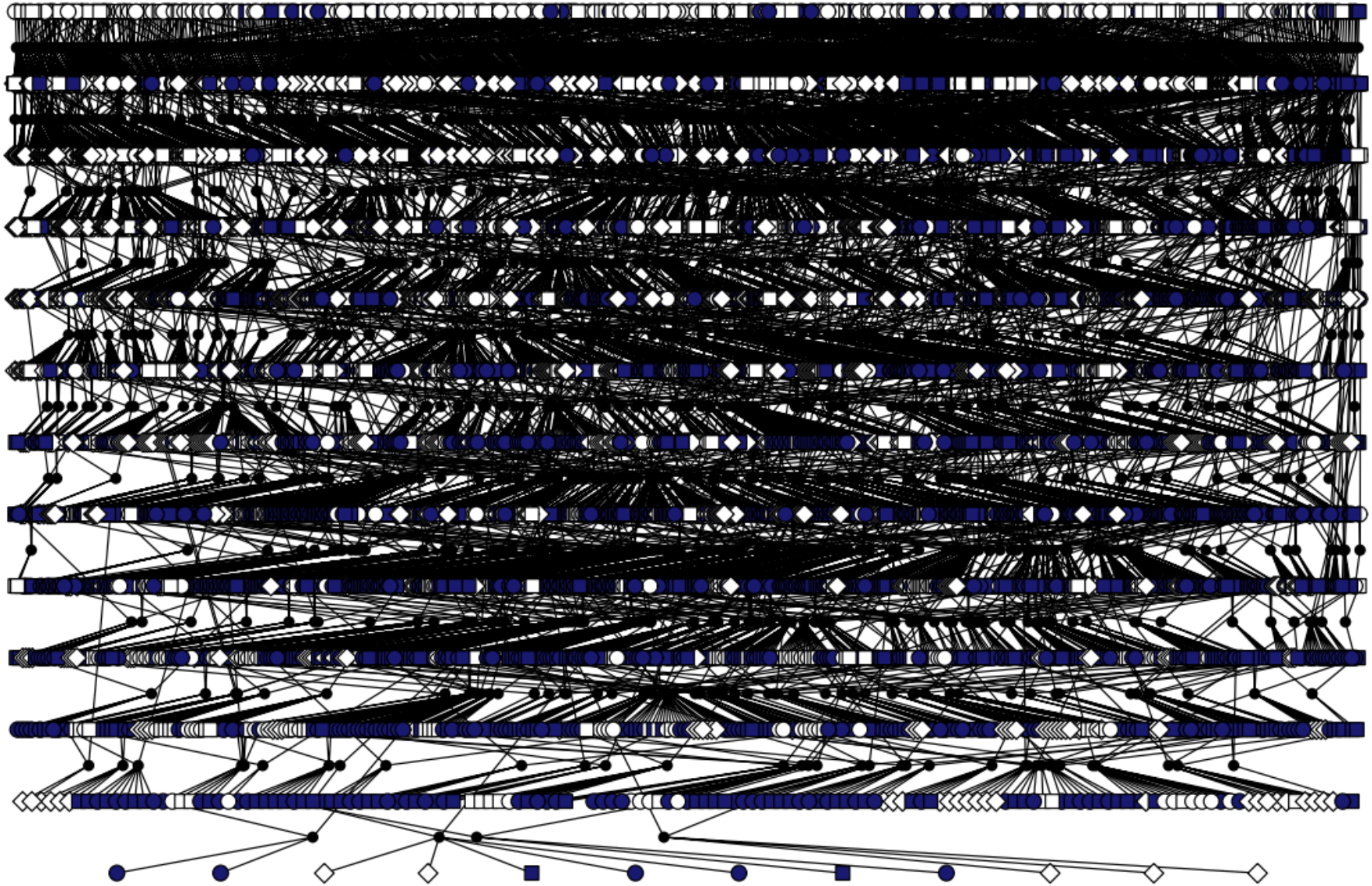
Non-migratory

Highly territorial & philopatric

Socially & (mostly) genetically monogamous

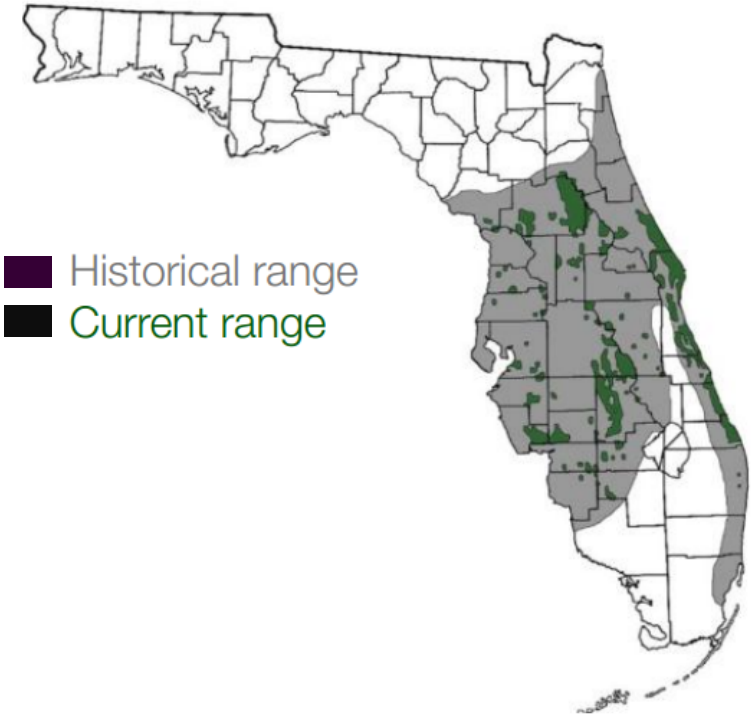
All individuals banded and tracked throughout life

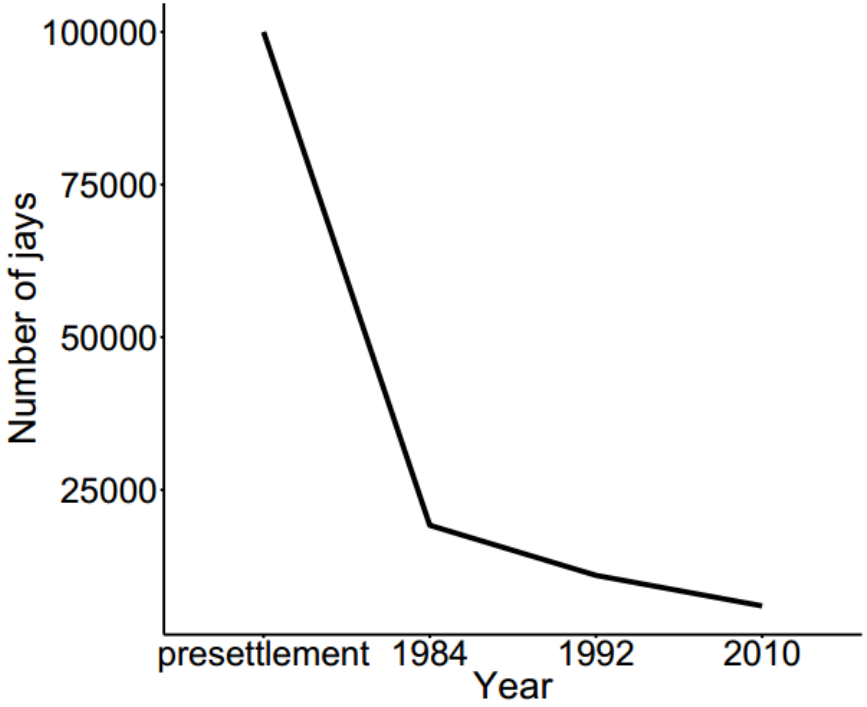

Quinn *et al.* 1999; Townsend *et al.* 2011; Woolfenden & Fitzpatrick 1984, 1996



We have blood samples as well as life history & morphological data for >4,000 pedigreed individuals

# Florida Scrub-Jay populations have drastically declined due to habitat loss




97% decline in past century. 50% decline in past 20 years

map by Monica E. McGarrit

# Florida Scrub-Jay genomic resources

Genome-wide SNPs

Genome assembly

Transcriptome assembly

Genome annotation

Linkage map construction

# Florida Scrub-Jay genomic resources

Genome-wide SNPs

Genome assembly

Transcriptome assembly

Genome annotation

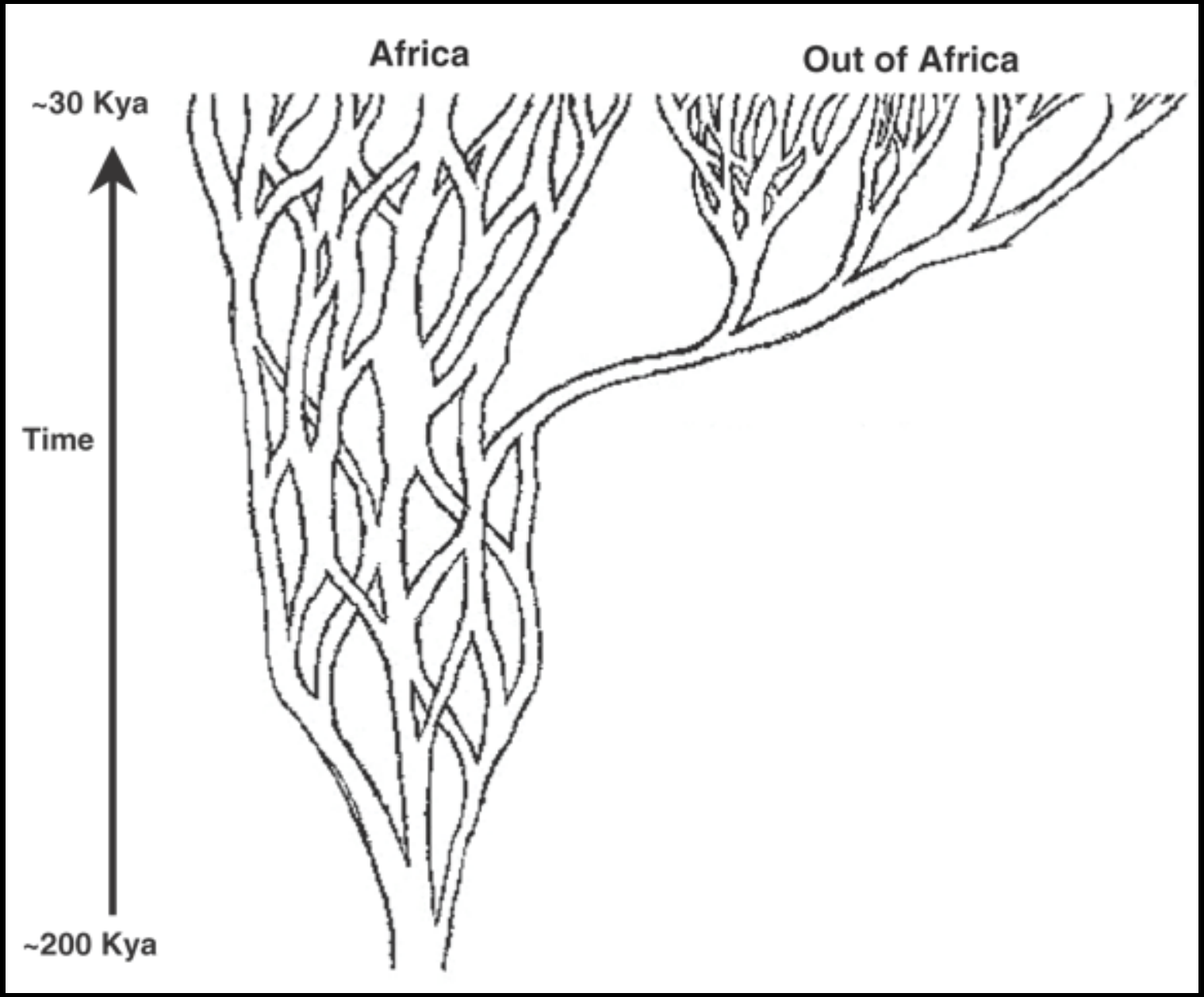
Linkage map construction

How does the recent population crash manifest itself in the structure of genetic variation?

# Population Structure

How can we infer past patterns of migration from genome sequence?

# Human population history



# Principal Components Analysis for population structure

OPEN ACCESS Freely available online

PLoS GENETICS

## Population Structure and Eigenanalysis

**Nick Patterson[1*], Alkes L. Price[1,2], David Reich[1,2]**

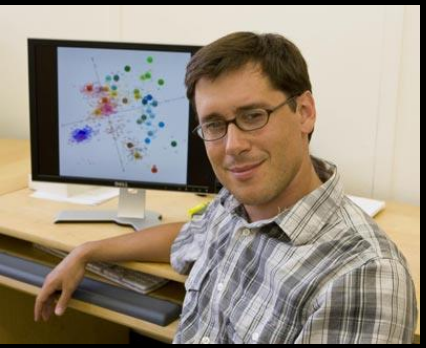
1 Broad Institute of Harvard and MIT, Cambridge, Massachusetts, United States of America, 2 Department of Genetics, Harvard Medical School, Boston, Massachusetts, United States of America

**Current methods for inferring population structure from genetic data do not provide formal significance tests for population differentiation. We discuss an approach to studying population structure (principal components analysis) that was first applied to genetic data by Cavalli-Sforza and colleagues. We place the method on a solid statistical footing, using results from modern statistics to develop formal significance tests. We also uncover a general "phase change" phenomenon about the ability to detect structure in genetic data, which emerges from the statistical theory we use, and has an important implication for the ability to discover structure in genetic data: for a fixed but large dataset size, divergence between two populations (as measured, for example, by a statistic like $F_{ST}$) below a threshold is essentially undetectable, but a little above threshold, detection will be easy. This means that we can predict the dataset size needed to detect structure.**

Citation: Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLoS Genet 2(12): e190. doi:10.1371/journal.pgen.0020190

Patterson N, Price AL, Reich D (2006) *PLoS Genet* 2(12): e190.

# European genetic differentiation using PCA

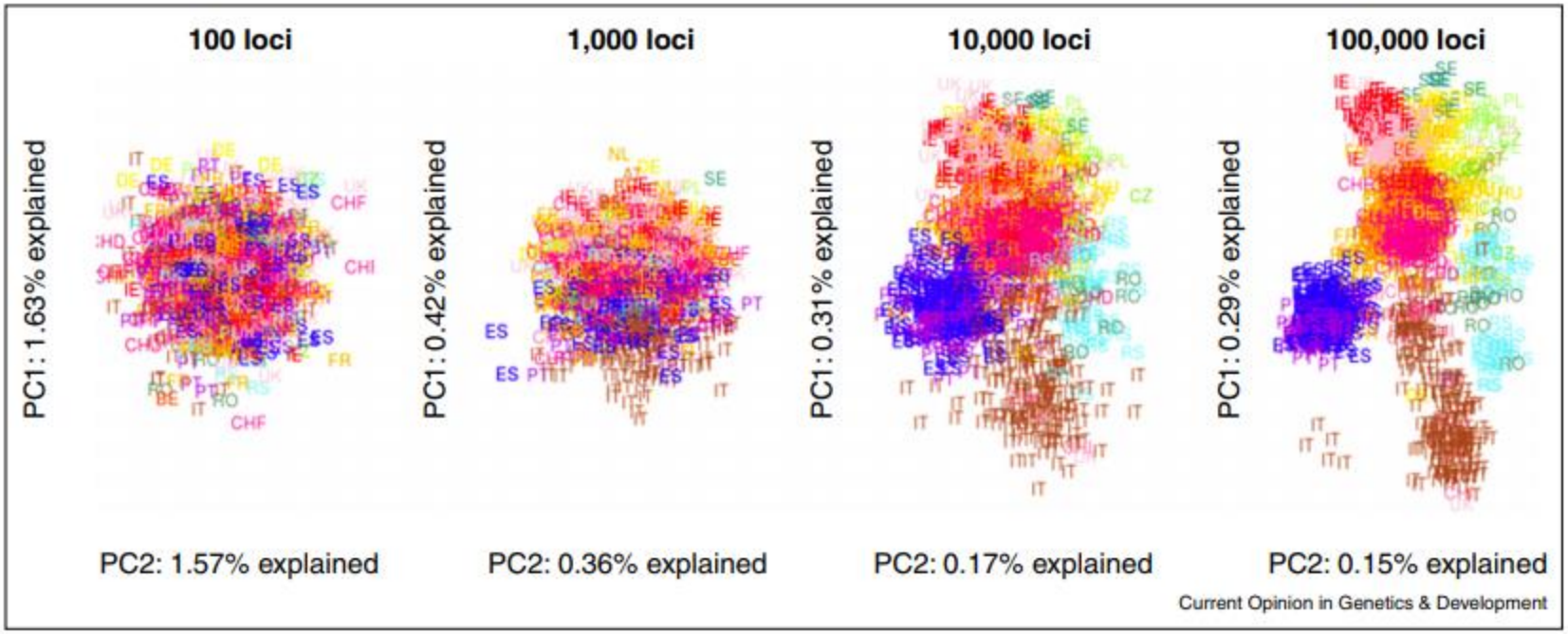


John Novembre

Novembre *et al*. 2008 *Nature* 456:274.

# Accurate fine structure inference from PCA requires a huge number of SNPs

**Figure 2**




Novembre & Peter 2016 *Cur. Op. Genes & Dev.*

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW. 2002 Genetic structure of human populations. Science. 298:2381–2385.

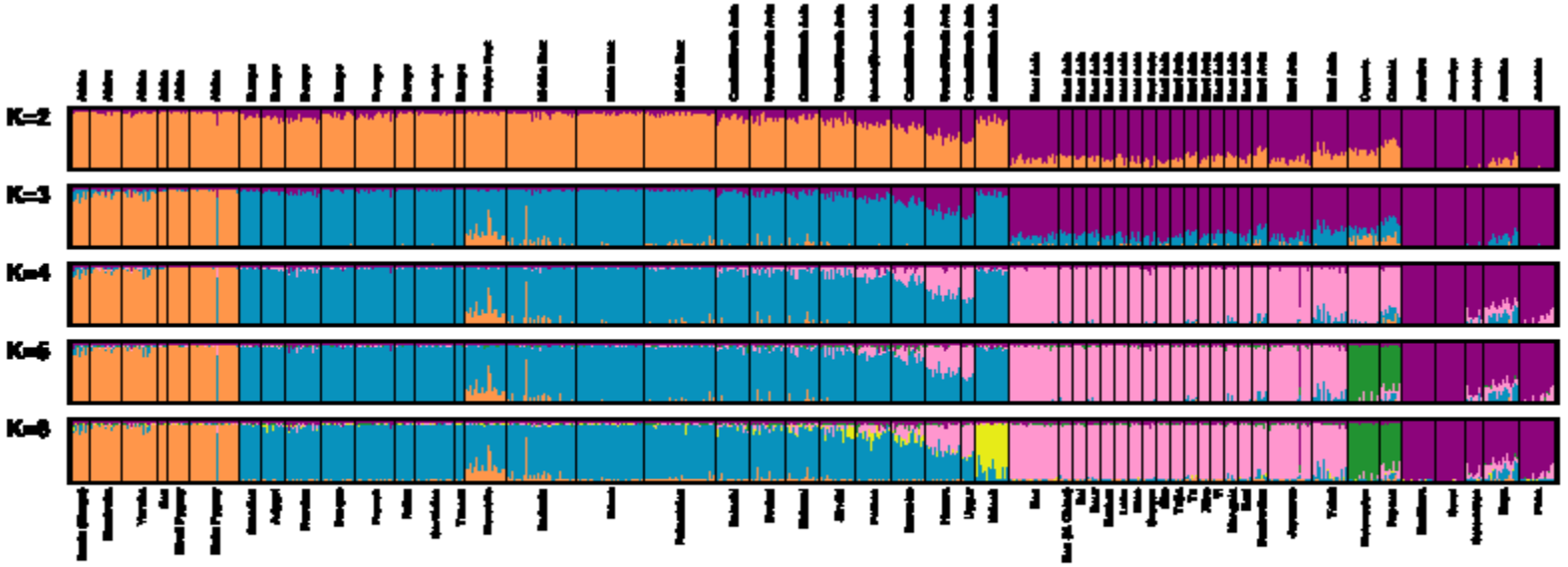
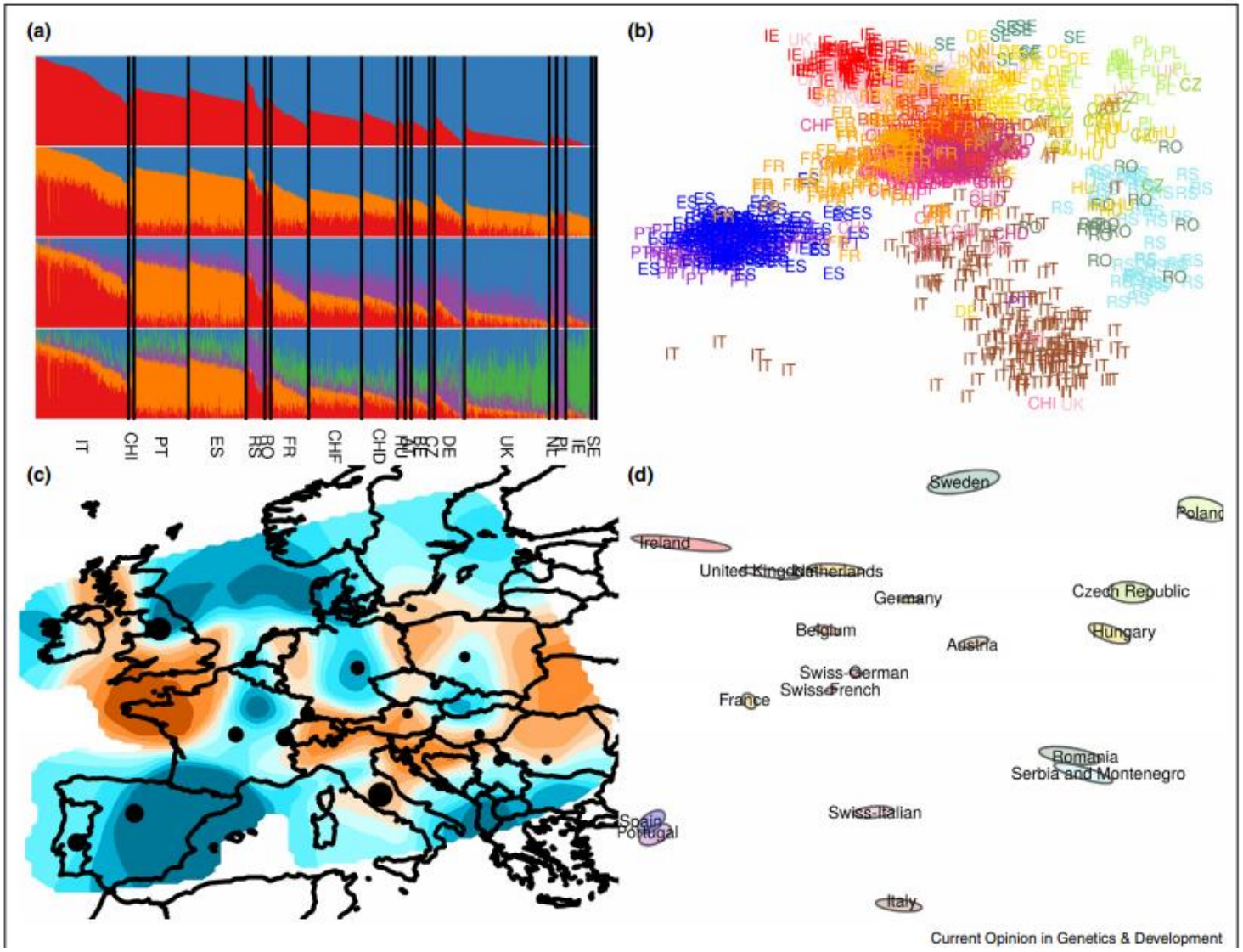



**Fig. 1.** Estimated population structure. Each individual is represented by a thin vertical line, which is partitioned into *K* colored segments that represent the individual's estimated membership fractions in *K* clusters. Black lines separate individuals of different populations. Populations are labeled below the figure, with their regional affiliations above it. Ten *structure* runs at each *K* produced nearly identical individual membership coefficients, having pairwise similarity coefficients above 0.97, with the exceptions of comparisons involving four runs at $K = 3$ that separated East Asia instead of Eurasia, and one run at $K = 6$ that separated Karitiana instead of Kalash. The figure shown for a given *K* is based on the highest probability run at that *K*.

ADMIXTURE

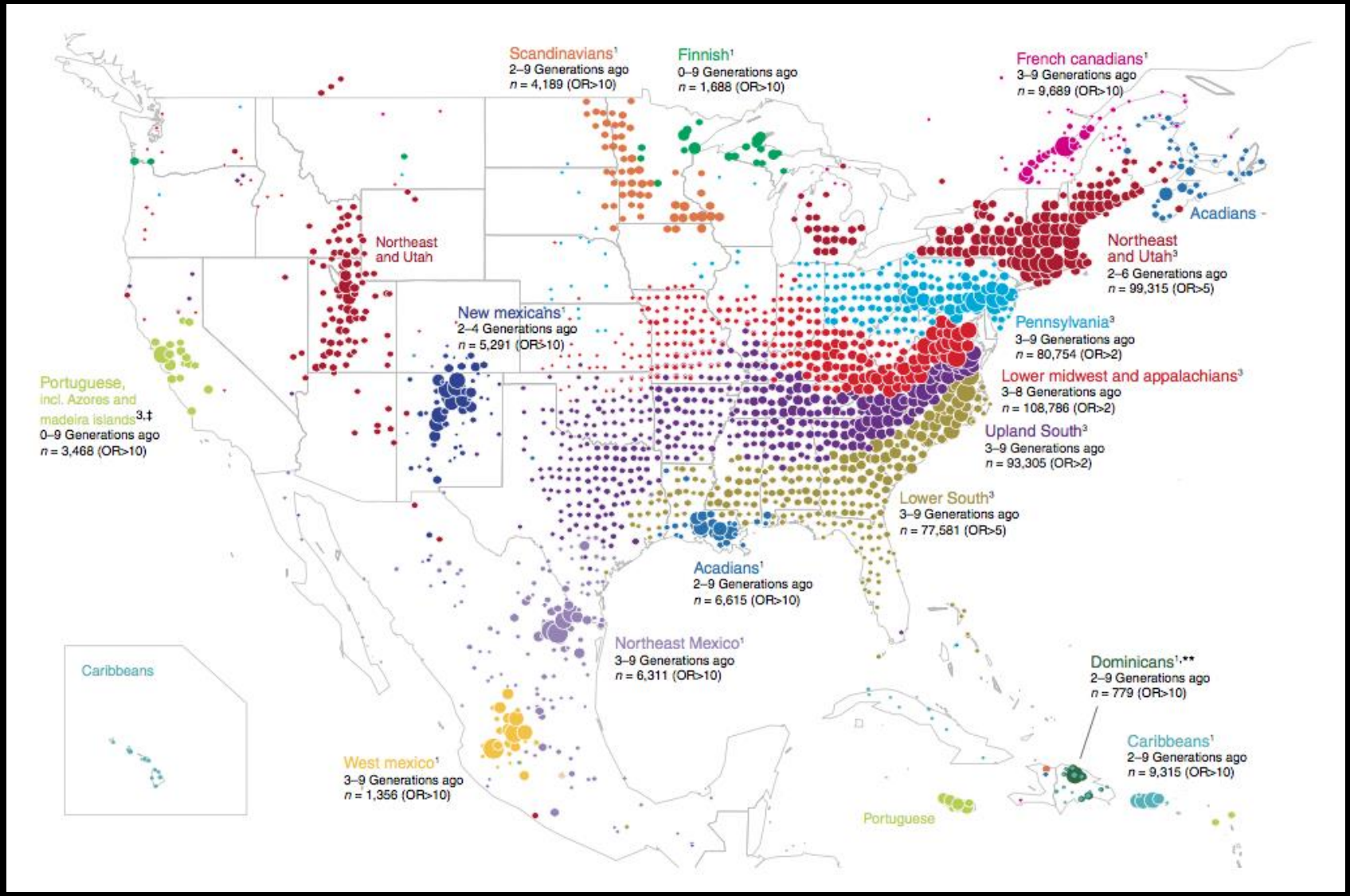
PCA




EEMS

SpaceMix

# Inference of population fine-structure from IBD

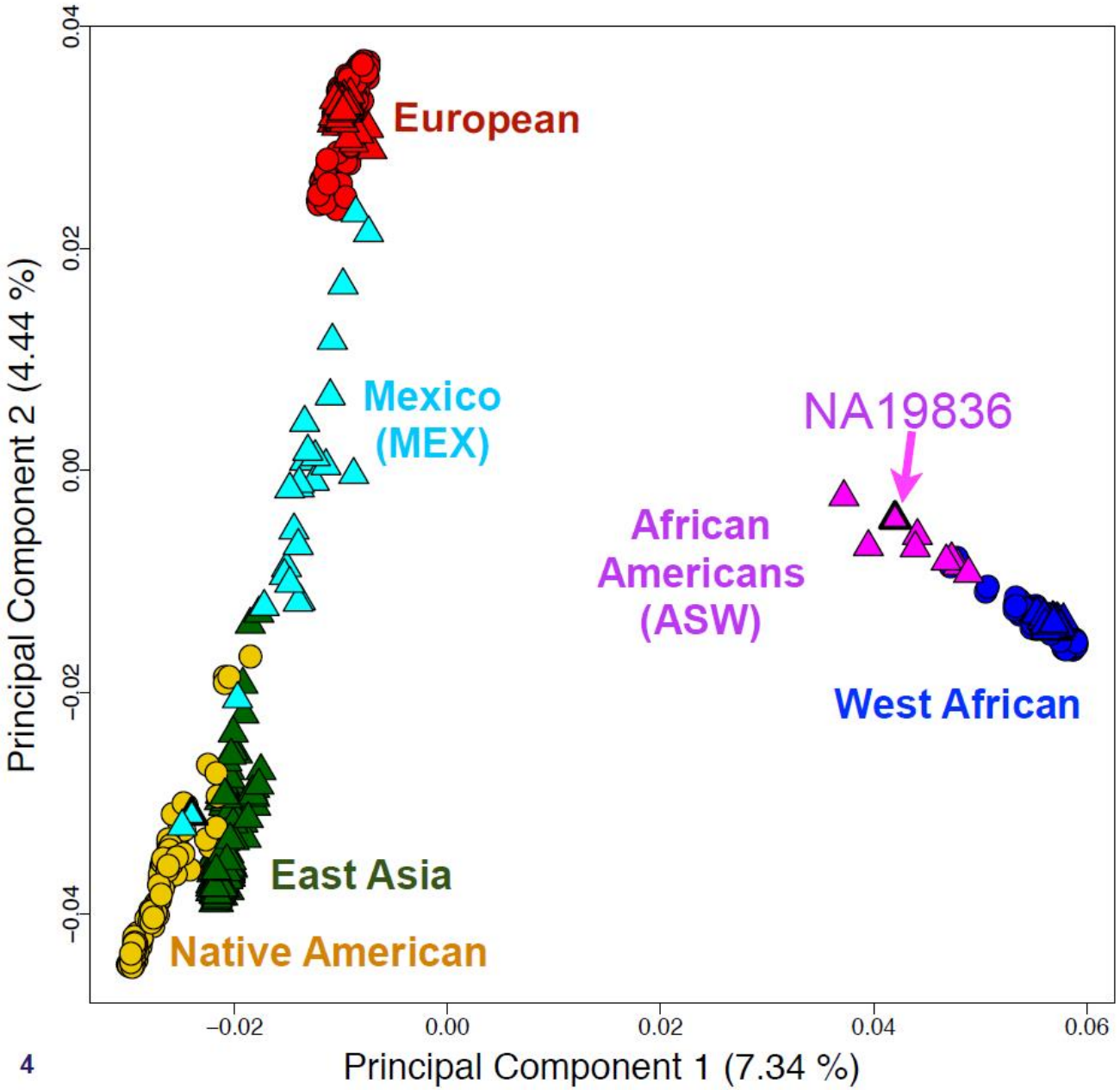



Eimear Kenny

# ADMIXTURE

How can we infer consanguinity and admixture from genetic data?

# Using PCA to infer admixed individuals




Africa
African American
Europe
Mexico
Native American
East Asia

HGDP
HapMap

Kasia Bryc

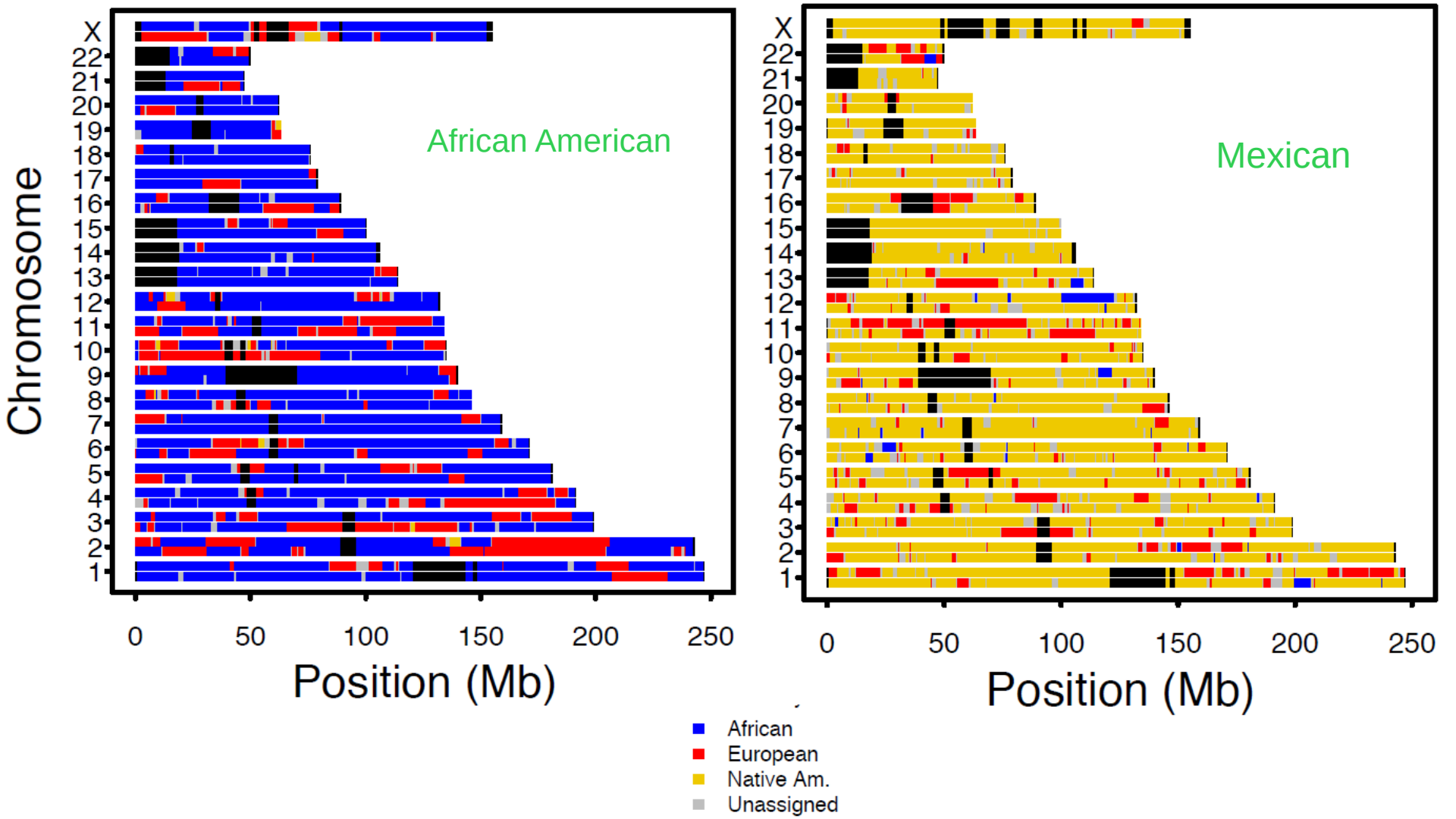
Adam Auton

# Local ancestry inference

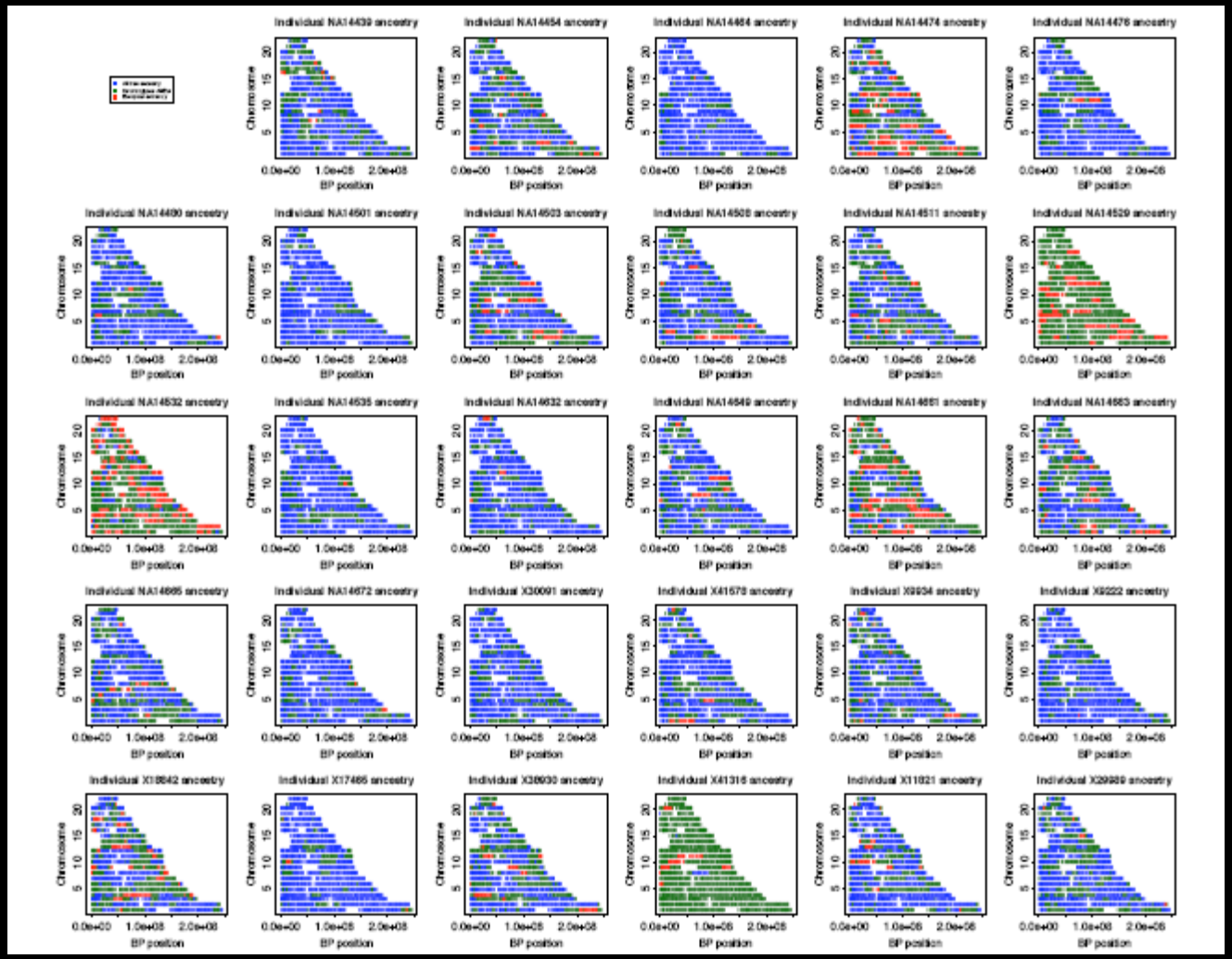
- Run PCA on African, European, and African American samples.
- From PC loadings of 15 SNP windows, infer 0, 1, or 2 copies of African ancestry.
- Slide along the genome.
- RESULT:  Call of 0, 1, 2 copies of African ancestry for each chunk of the genome in each individual.

Kasia Bryc

# PCAdmix can identify the population-of-origin of segments of the genome

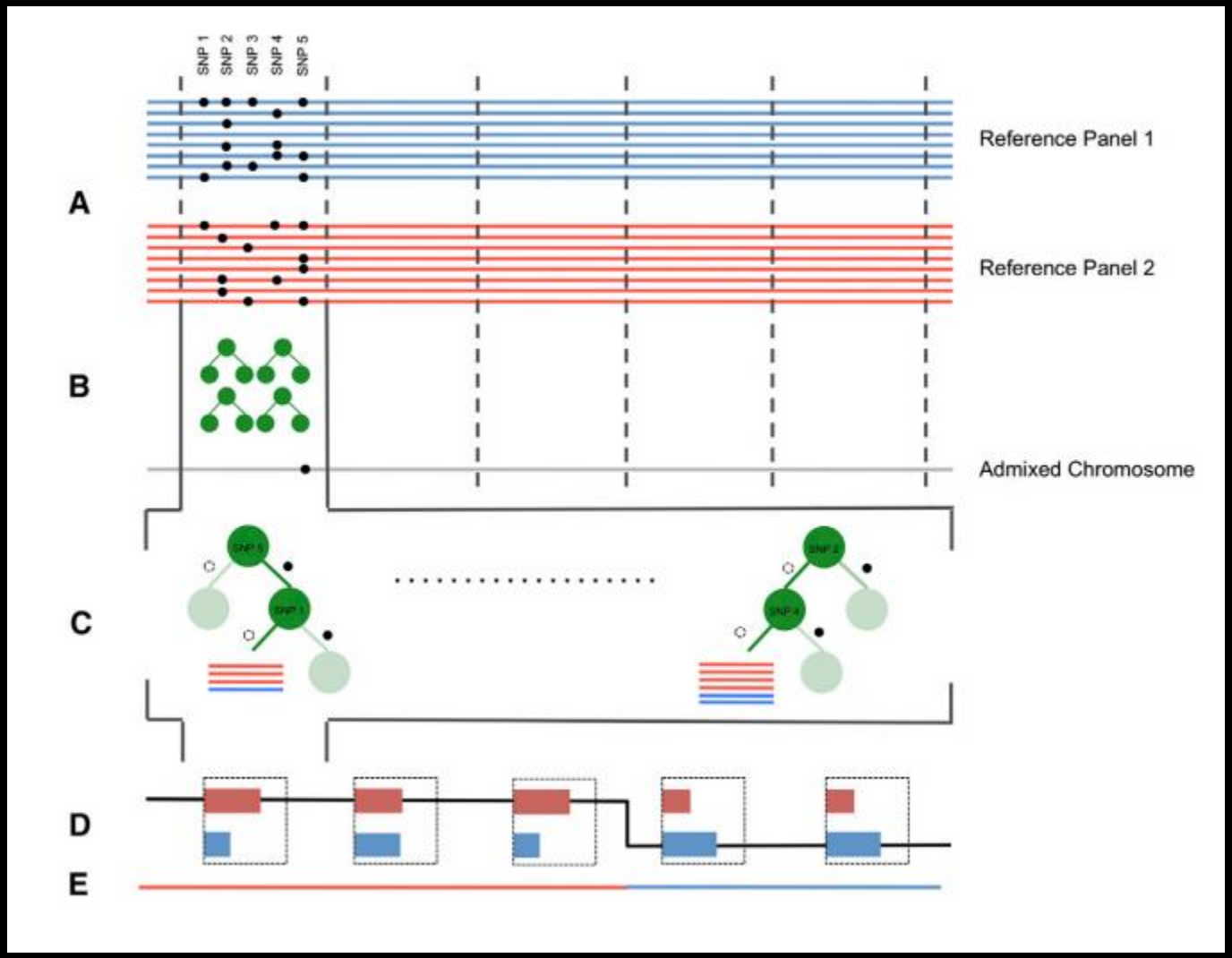


# High variability among individuals in admixture patterns




Kasia Bryc

# RFMix uses sequence from ancestral populations and a window-based random forest approach to map ancestry blocks



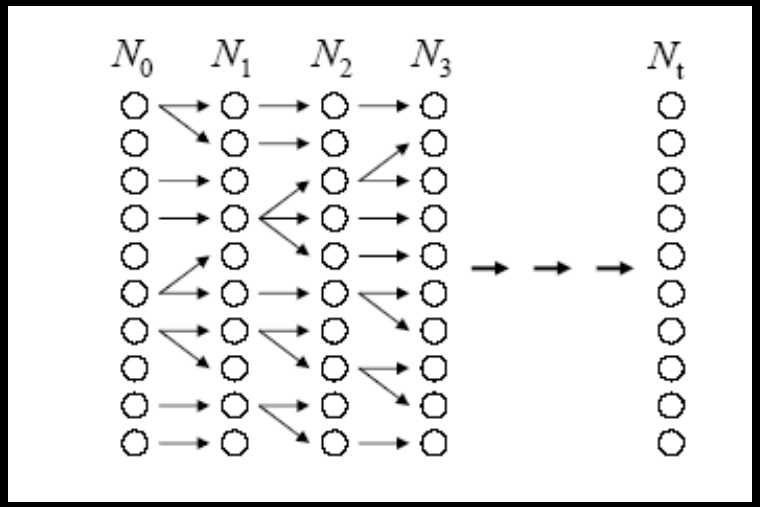

Maples *et al*. 2013 *AJHG*

# Random Genetic Drift

What can we infer from allele frequency dynamics of every nucleotide in the genome?

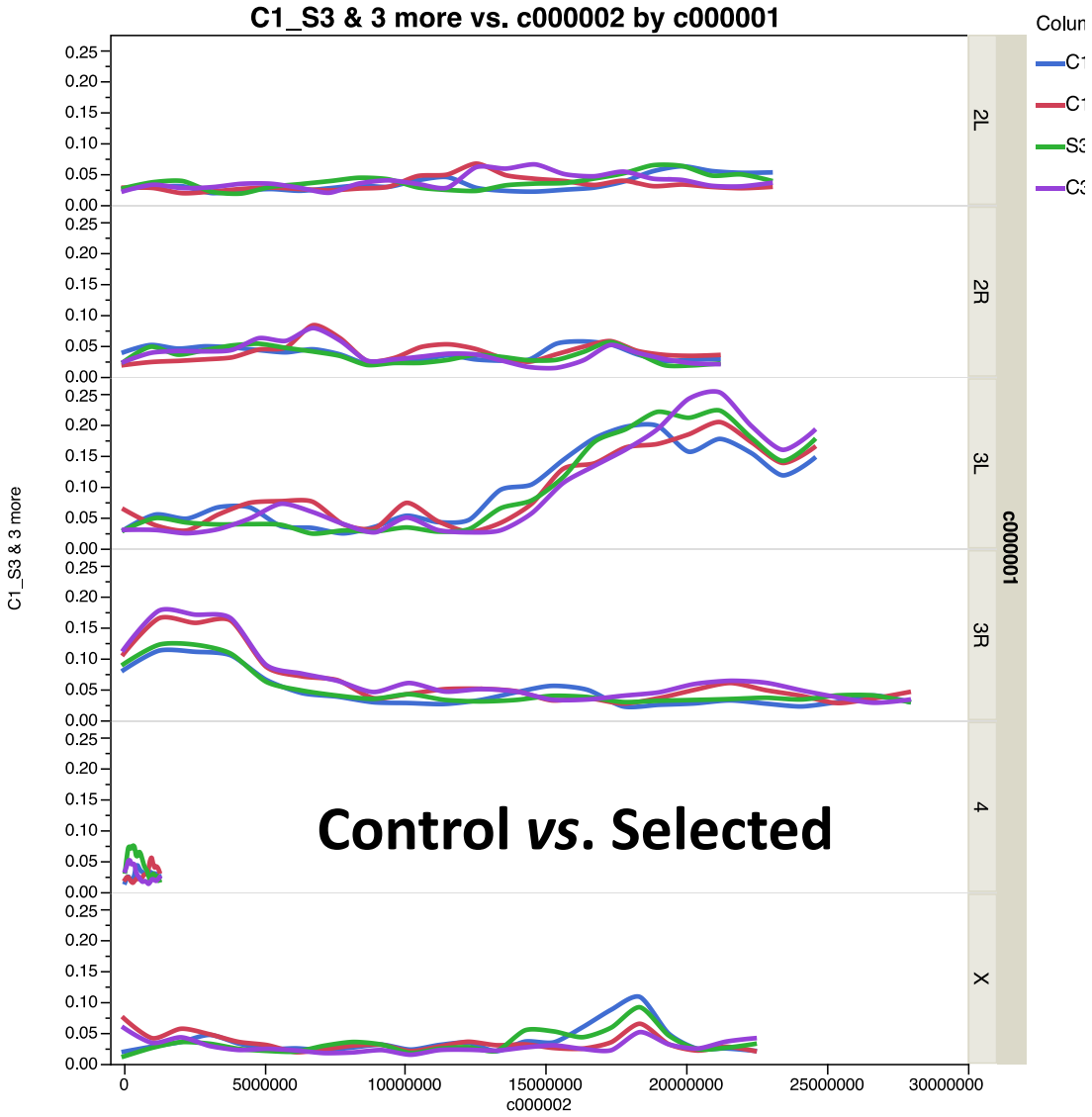
# The Wright-Fisher drift model:

## the Null model for Evolve-and-Resequence




- Selfing allowed
- Random mating
- Non-overlapping generations
- Constant population size
- No migration
- No selection

# Genomic regions show consistent changes in allele frequency



*Bioinformatics*, 31(11), 2015, 1762–1770
doi: 10.1093/bioinformatics/btv014
Advance Access Publication Date: 21 January 2015
Original paper

OXFORD

Genetics and population analysis

# Gaussian process test for high-throughput sequencing time series: application to experimental evolution

**Hande Topa[1,*,†], Ágnes Jónás[2,3,*,†], Robert Kofler[2], Carolin Kosiol[2,*] Antti Honkela[4,*]**

[1]Helsinki Institute for Information Technology (HIIT), Department of Information and Computer Science, Aalto University, Espoo, Finland, [2]Institut für Populationsgenetik, Vetmeduni Vienna, 1210 Wien, Austria, [3]Vienna Graduate School of Population Genetics, Wien, Austria and [4]Helsinki Institute for Information Technology (HIIT), Department of Computer Science, University of Helsinki, Helsinki, Finland

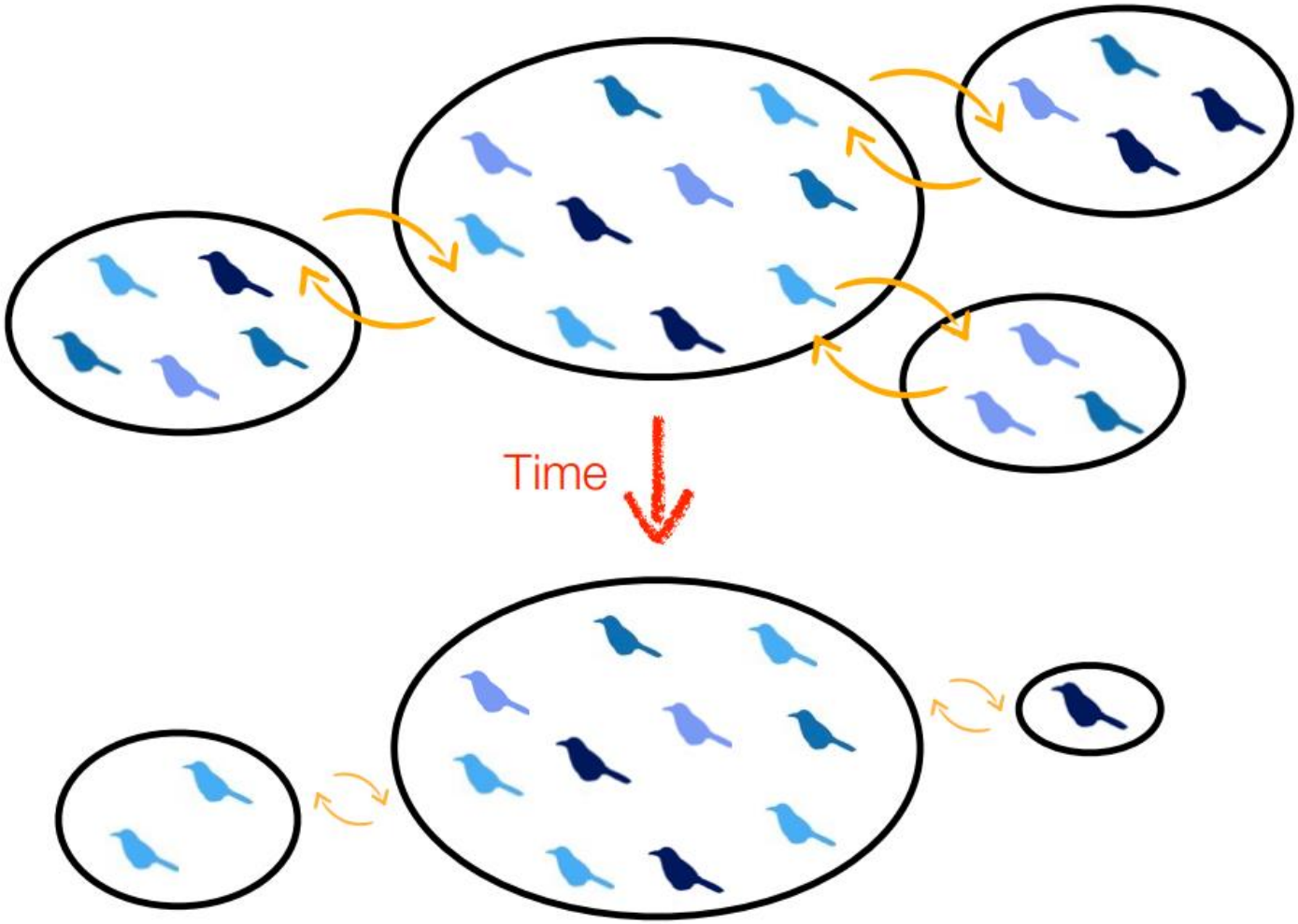
# Fitting pool-seq data by Gaussian process

- Sampling from one generation to the next each generation is Binomial($p_i$,$n_j$)
- The $p_i$ are drawn from a Beta distribution.
- The sampling is correlated along the genome, with covariance depending on linkage and LD.
- This model is fitted by Gaussian Process regression (least square).

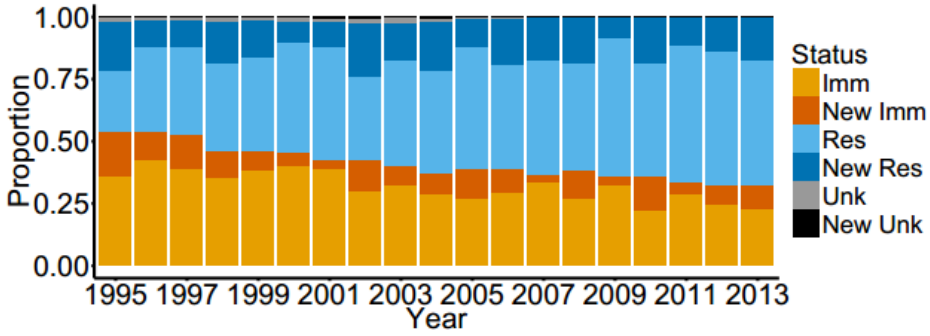
# Migration

How does spatial movement of individuals interact with other forces?

# Consequences of habitat fragmentation

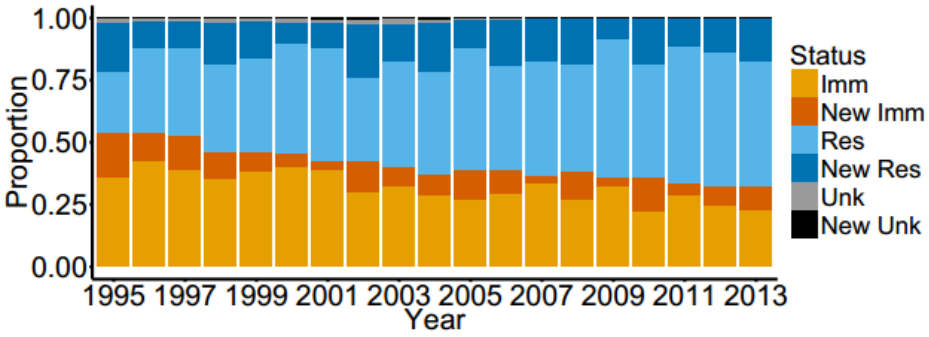


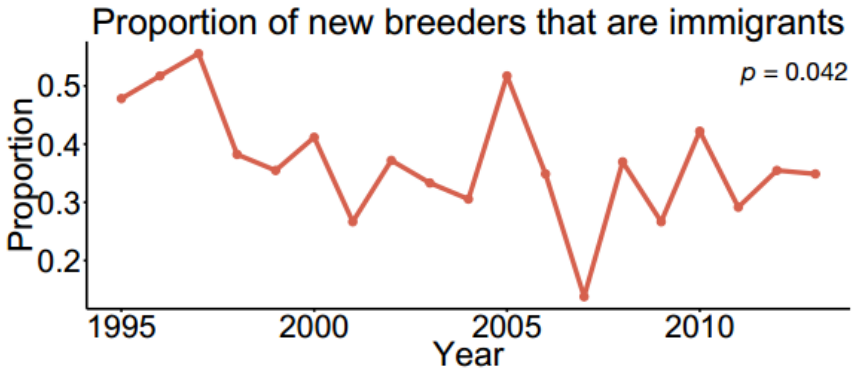
# Immigration rate decreased over time




Chen *et al.* 2016 *Current Biology*

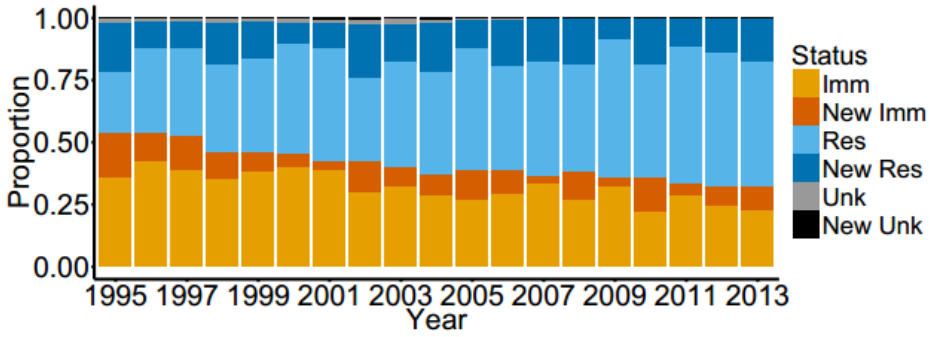
# Immigration rate decreased over time

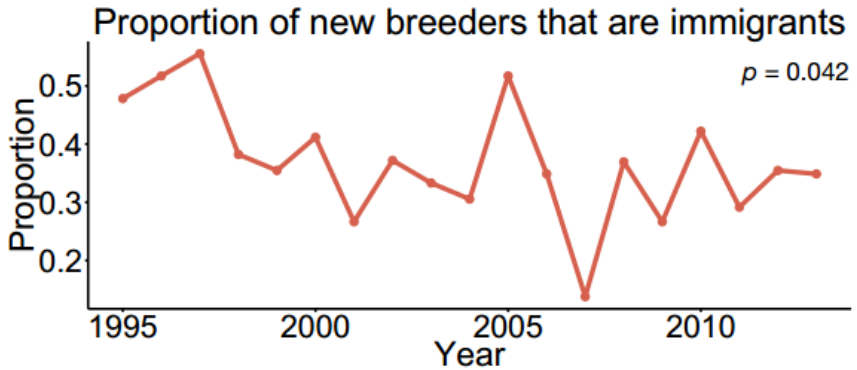


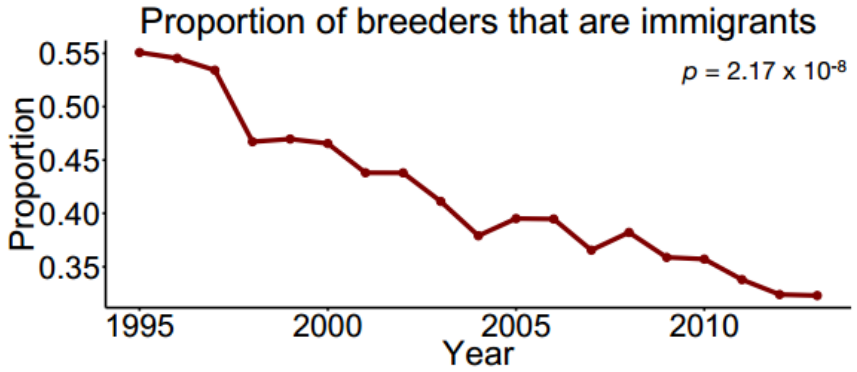




Chen *et al.* 2016 *Current Biology*

# Immigration rate decreased over time

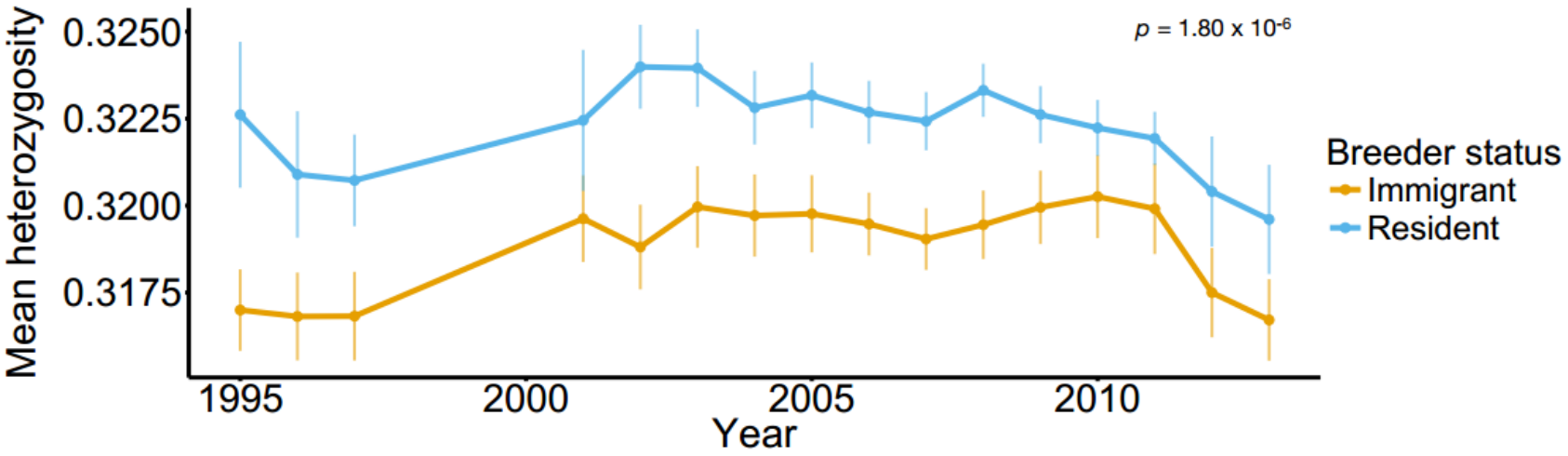


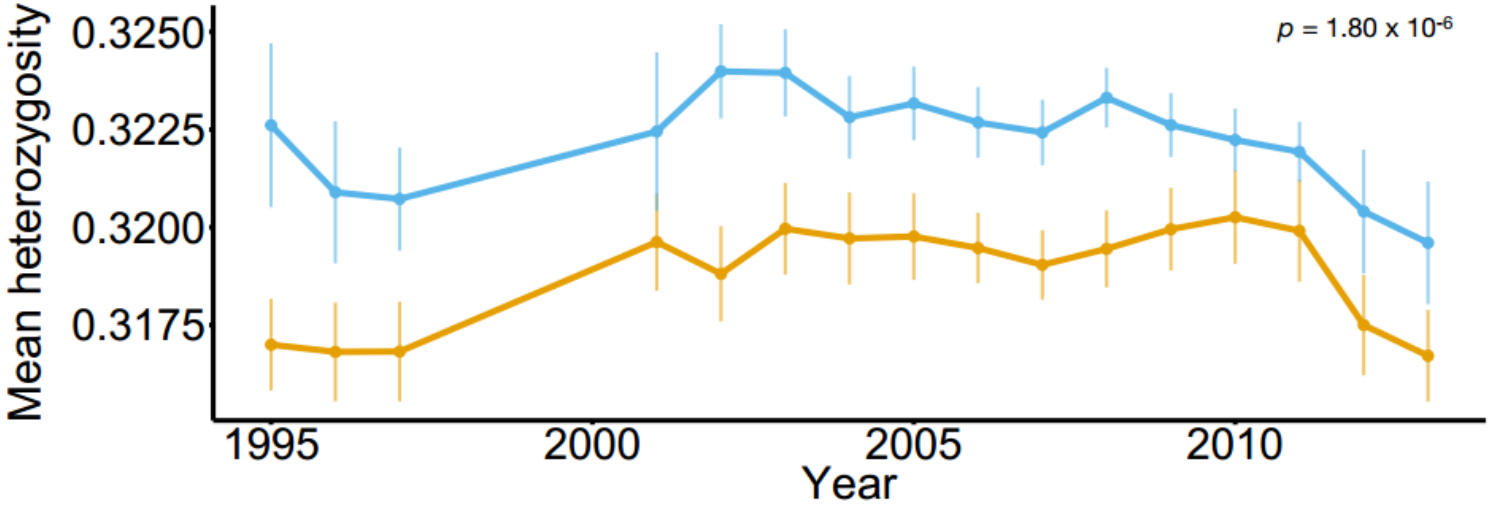

Chen *et al.* 2016 *Current Biology*

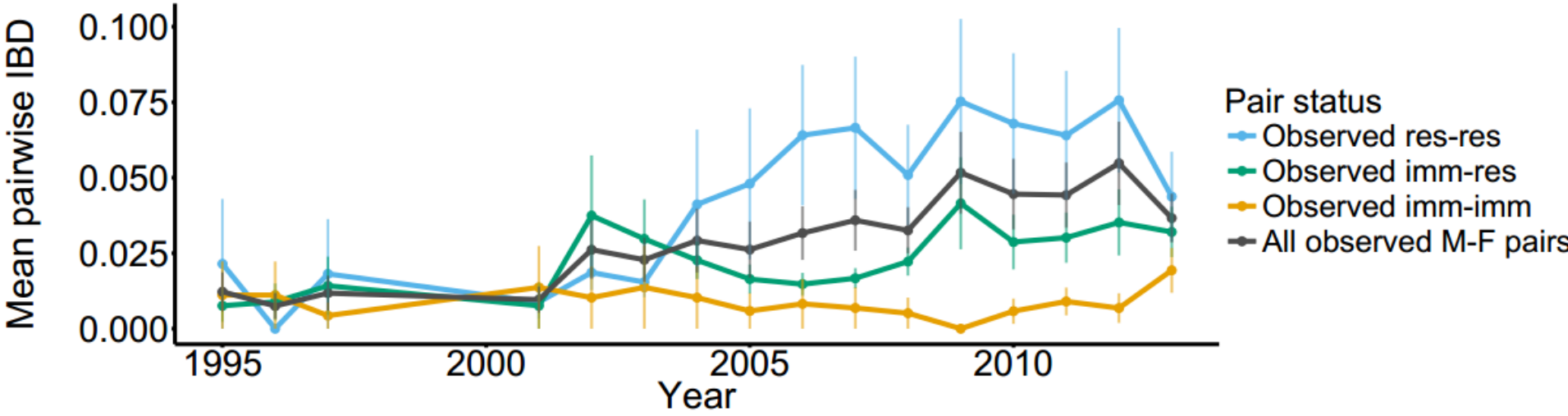
# Immigrants were less heterozygous than residents



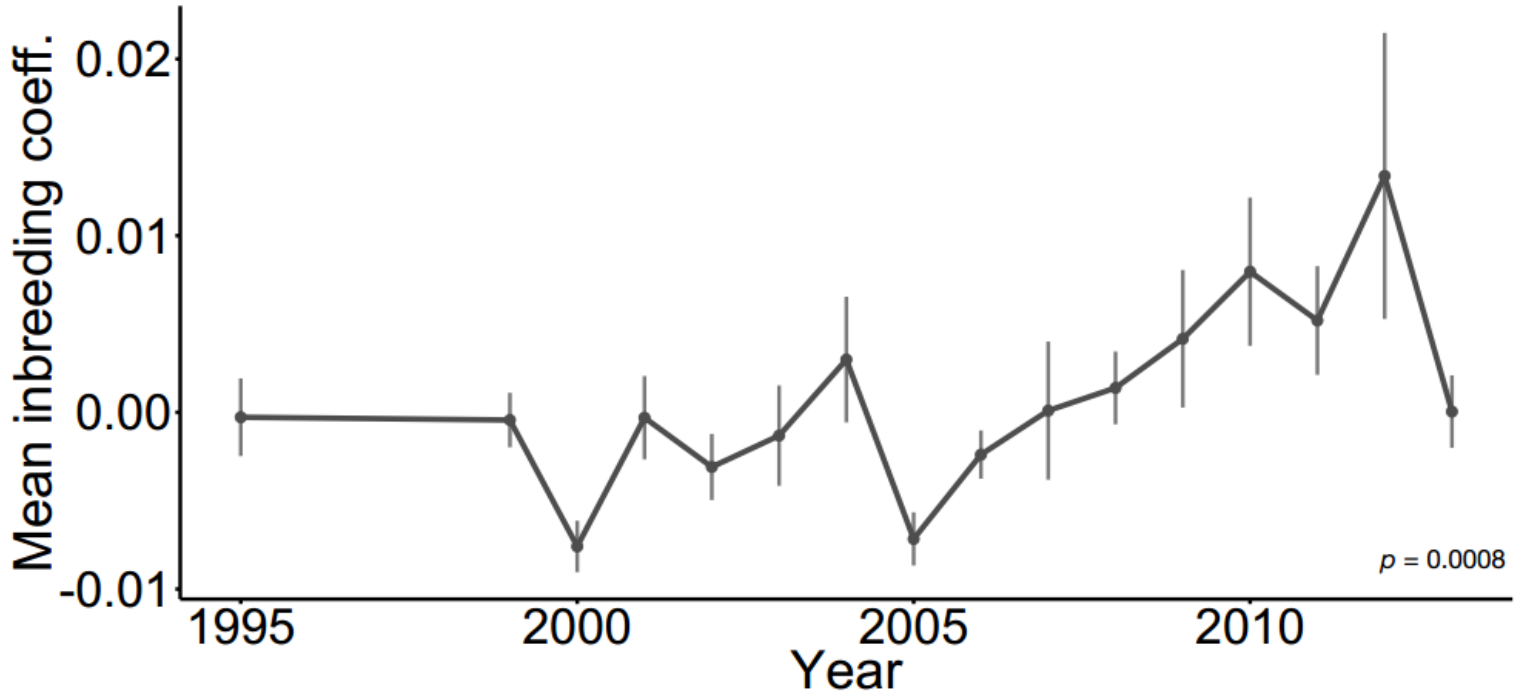
# Immigrants were less heterozygous than residents

## but still contributed genetic variation



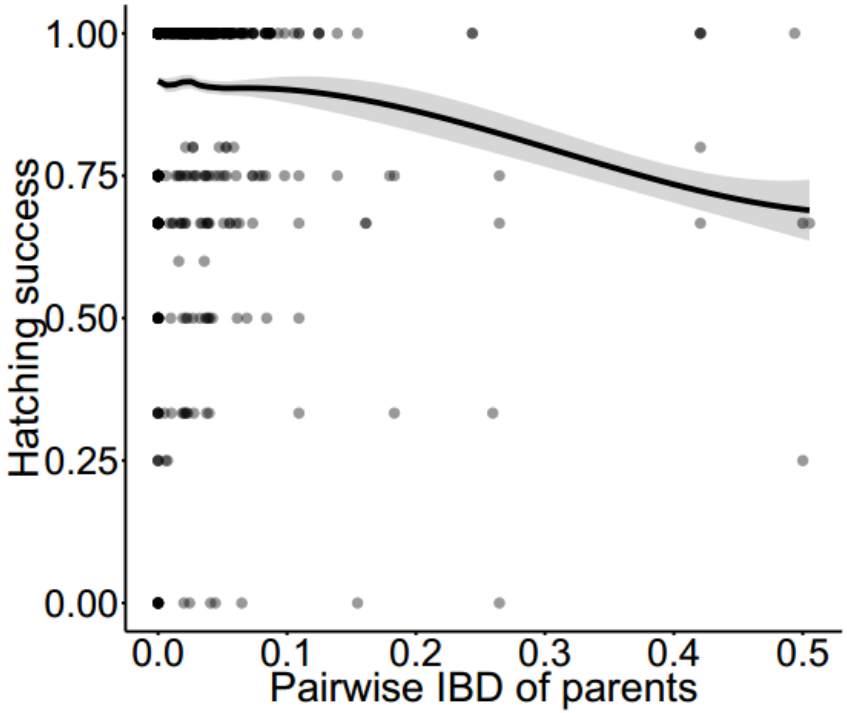



# Mean inbreeding coefficient of the birth cohort increased over time

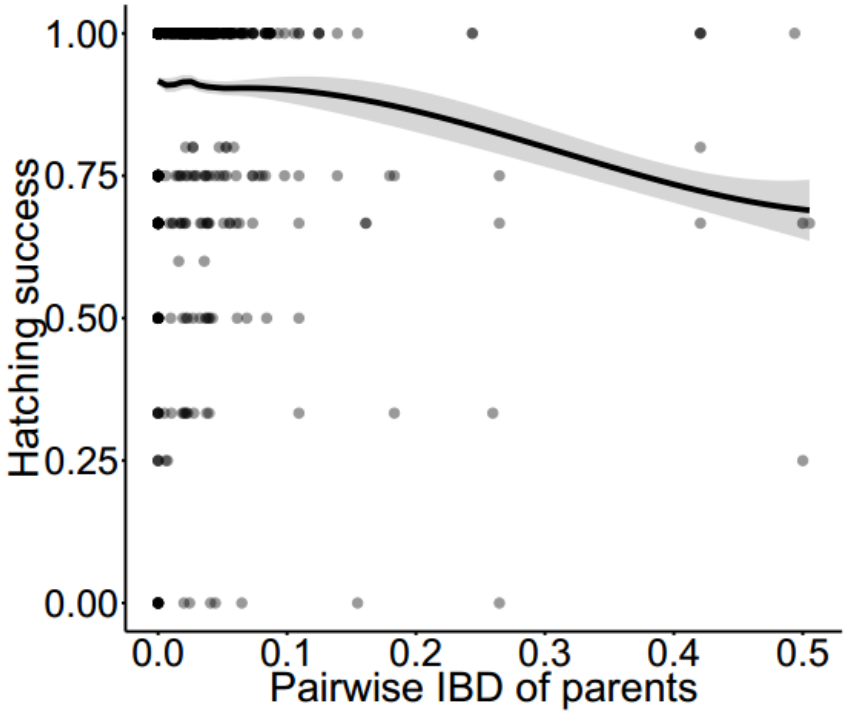


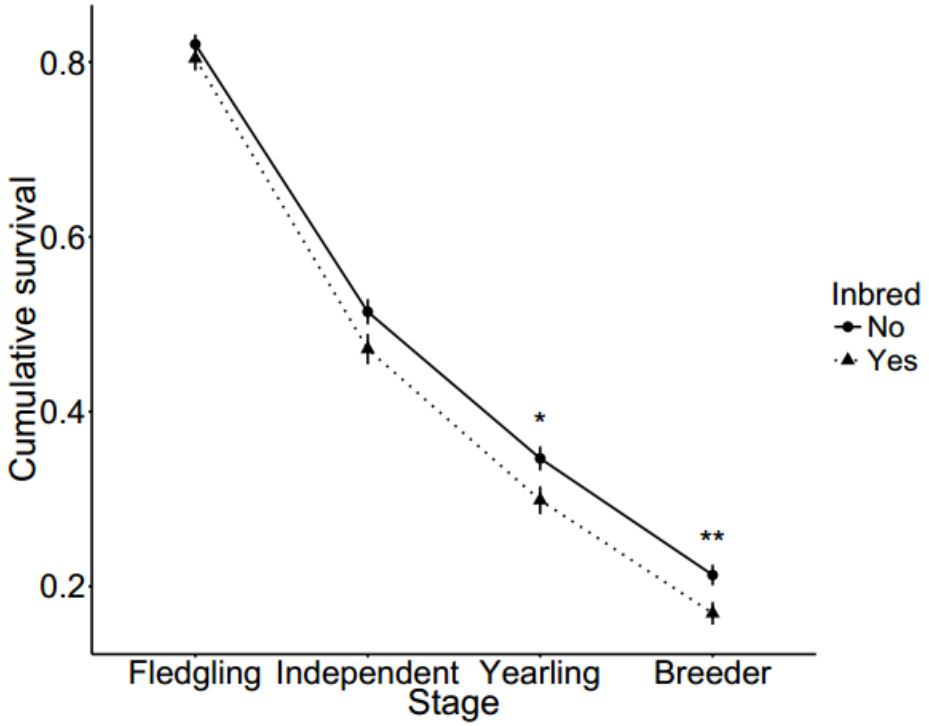

Chen *et al.* 2016 *Current Biology*

# Inbreeding depression in multiple life-history stages

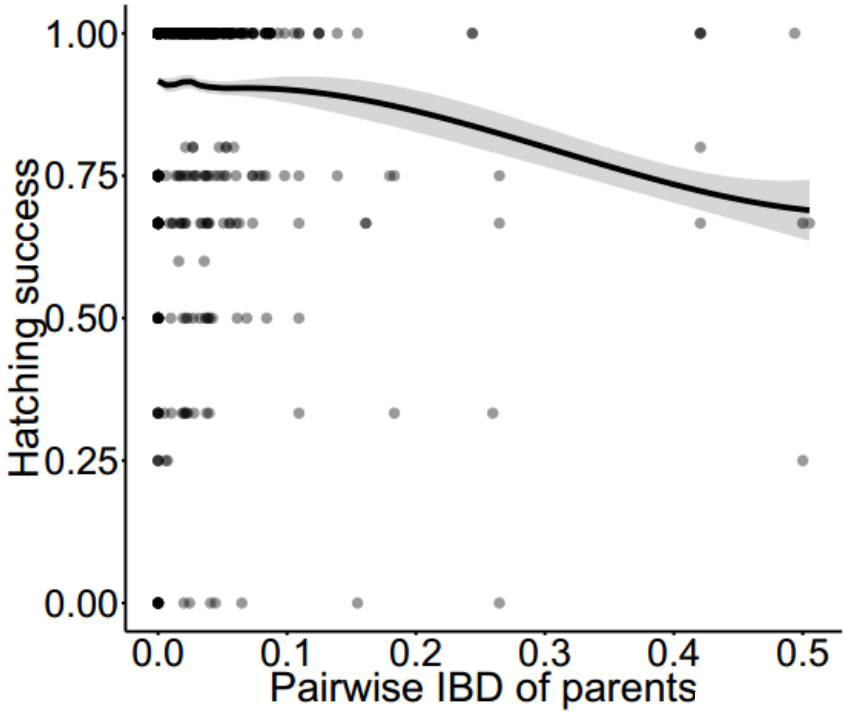


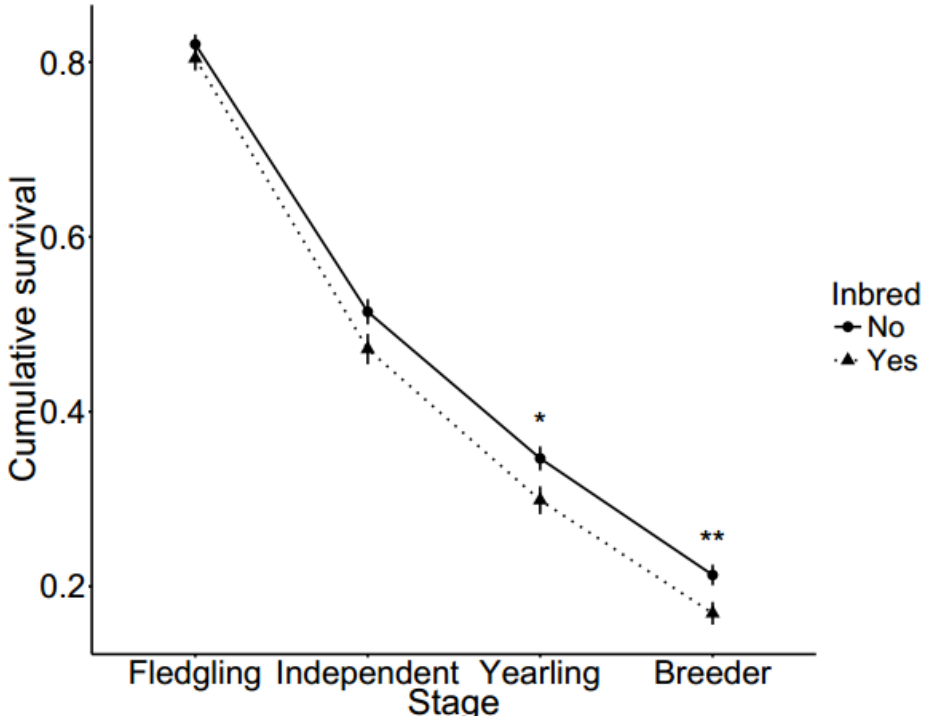
# Inbreeding depression in multiple life-history stages



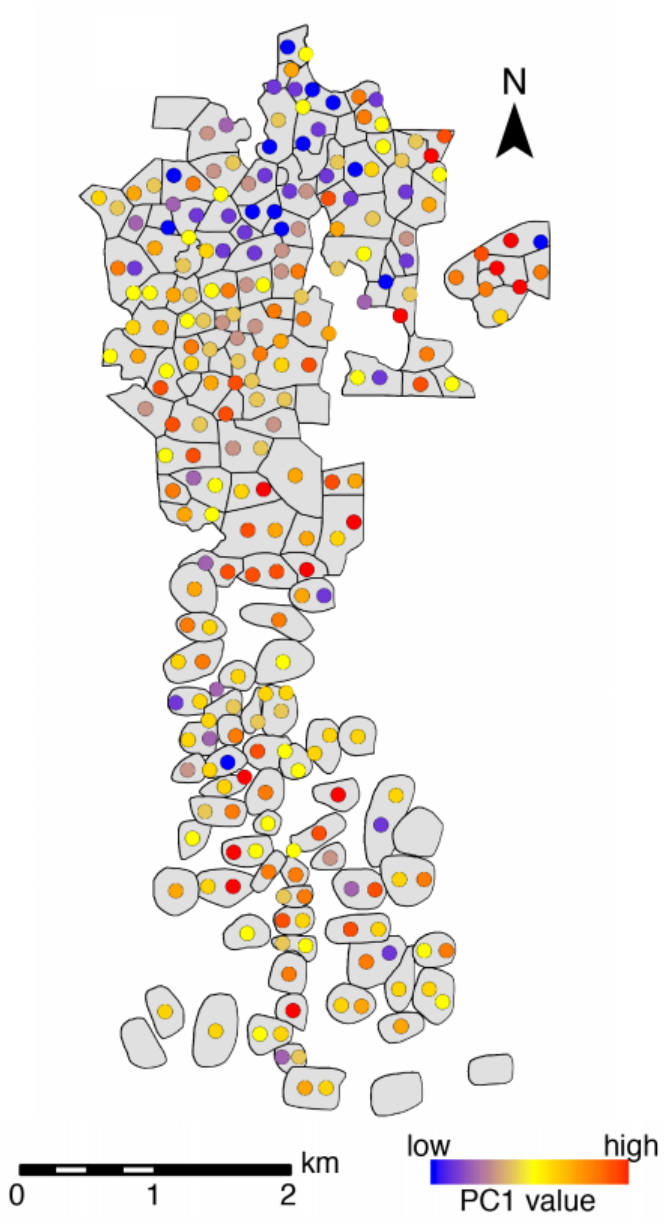
# Inbreeding depression in multiple life-history stages

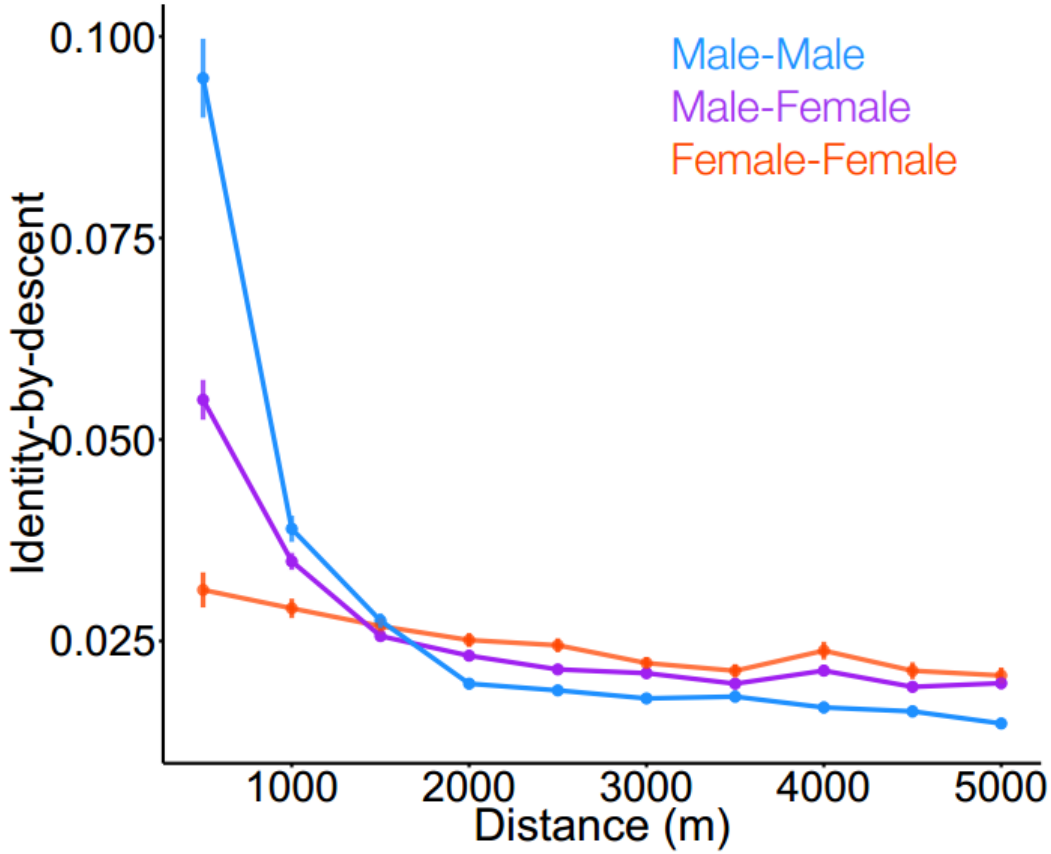



- ✓ Hatching success
- ✓ Nestling weight
- ✓ Juvenile survival
- ✓ Breeder lifespan
- ✓ Lifetime reproductive success

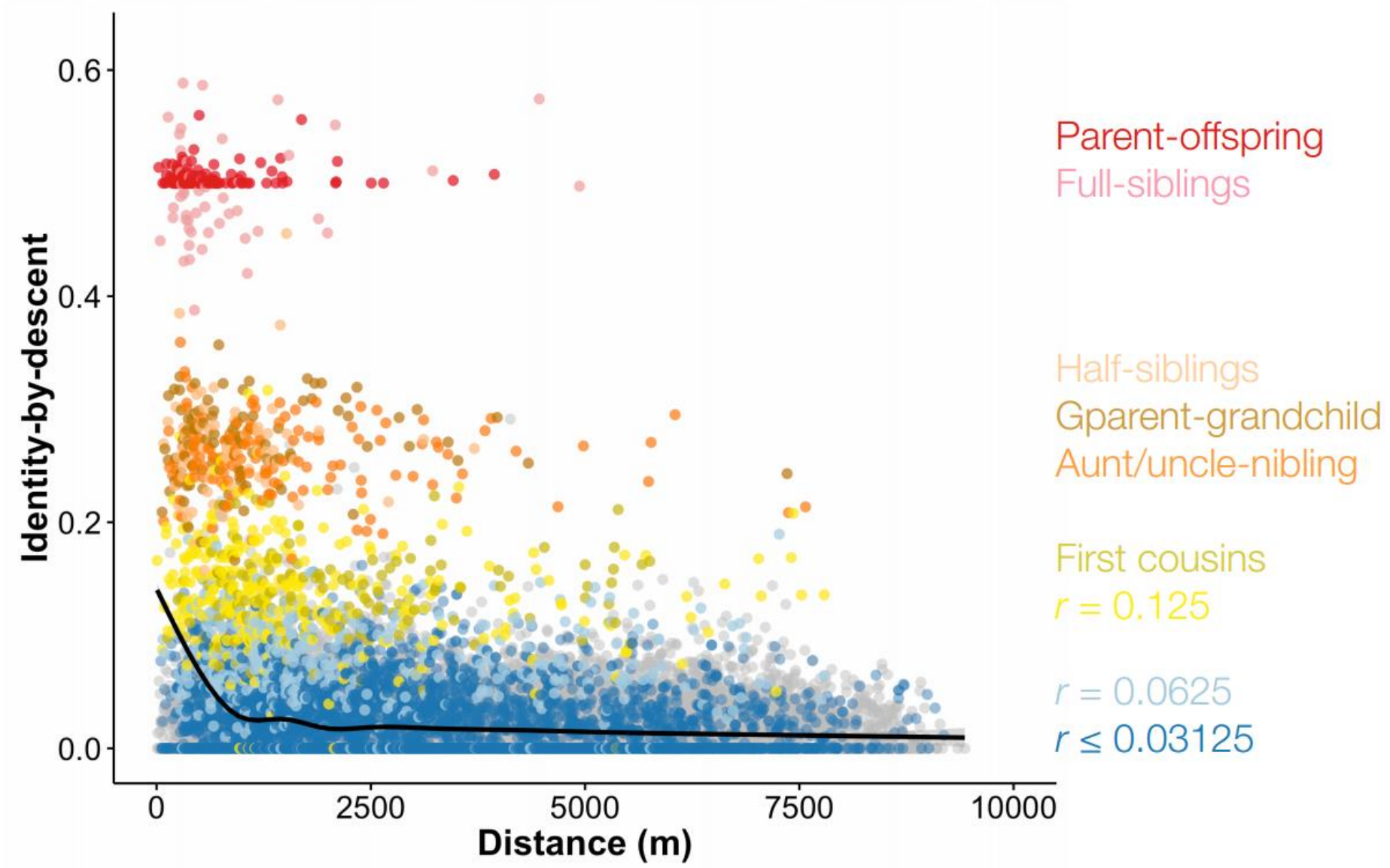
# Limited dispersal leads to isolation-by-distance




Aguillon, …, Chen *In review*

# A closer look at the isolation-by-distance pattern

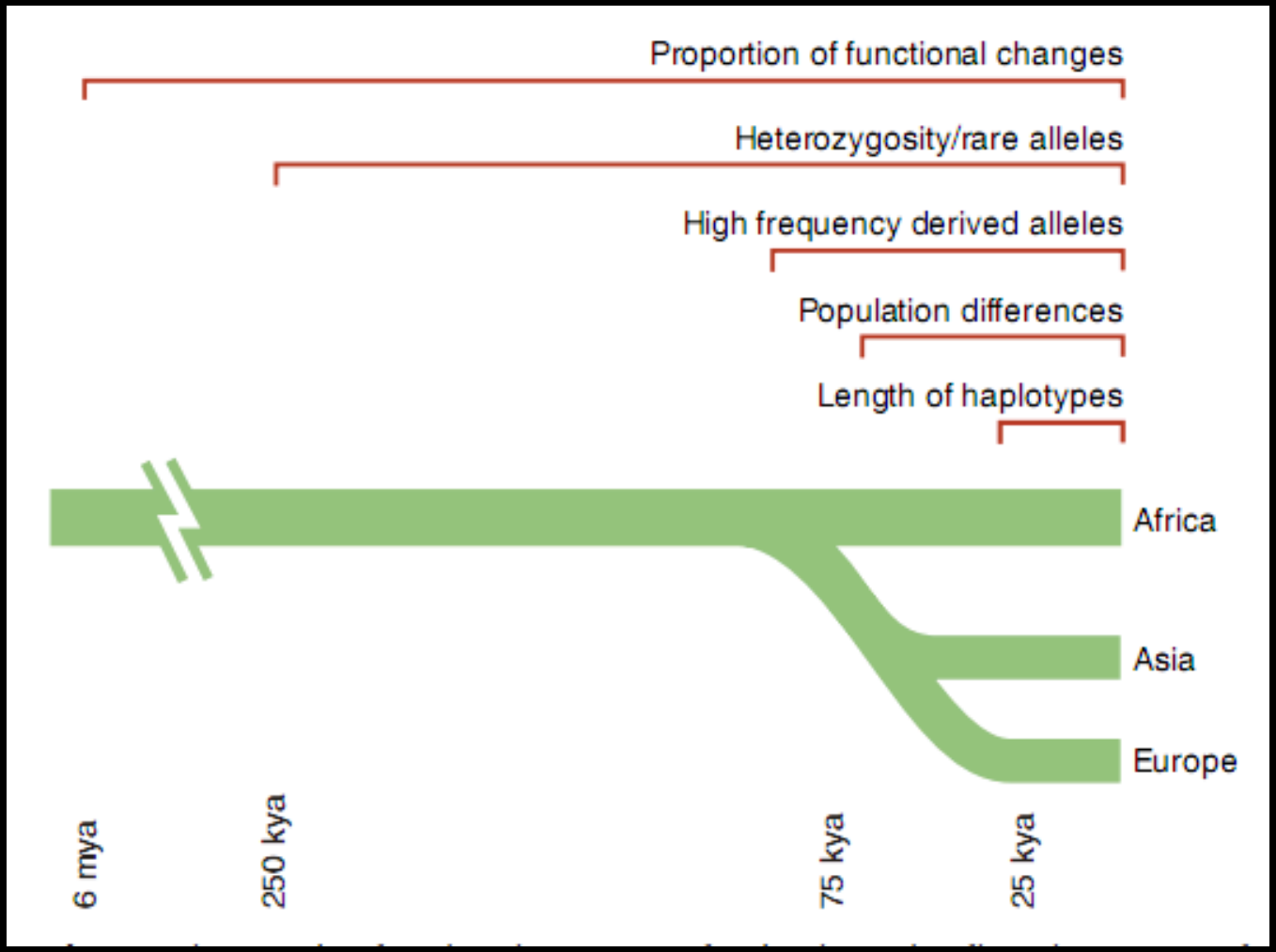


# Natural Selection

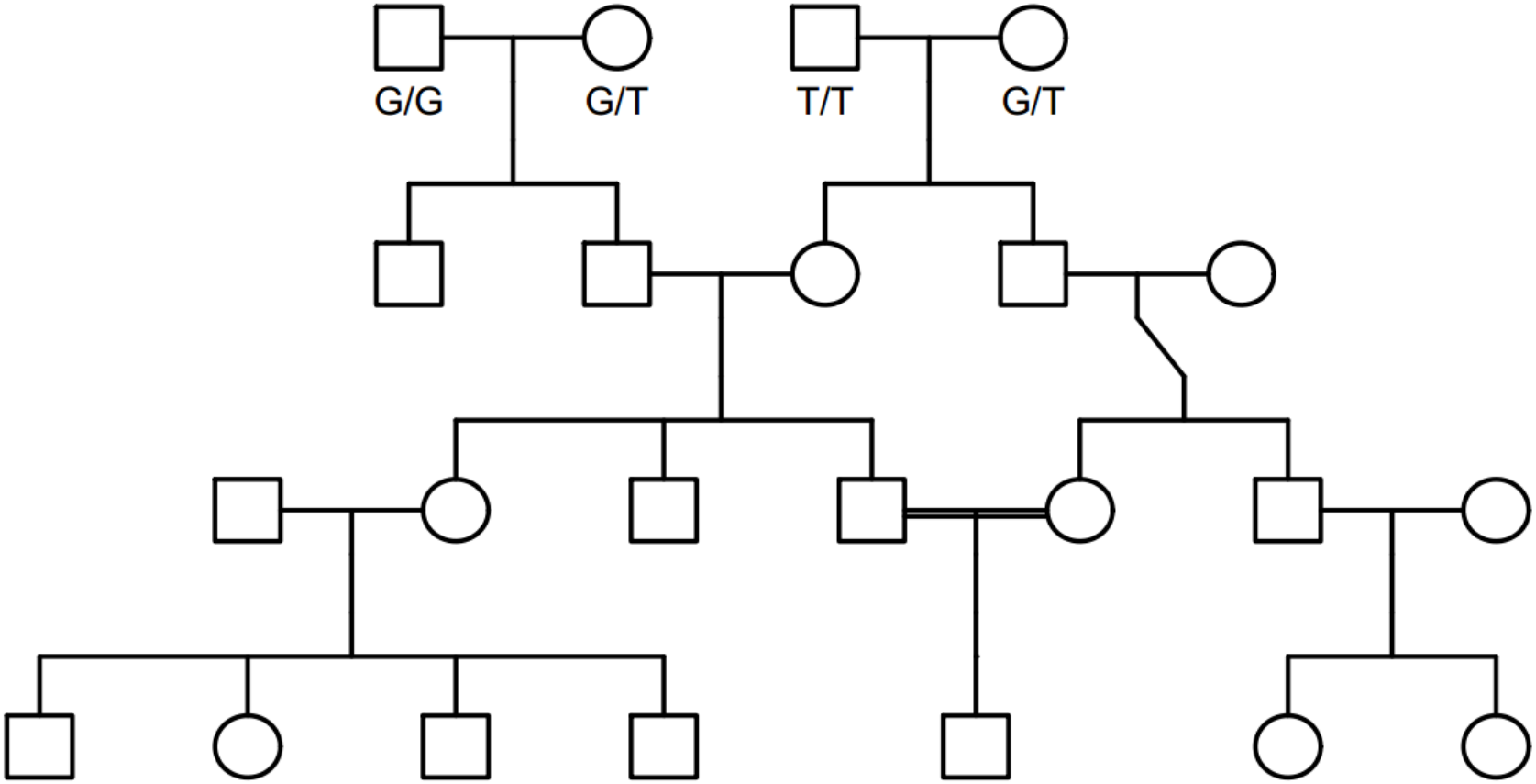
How best to do inference of selection genome-wide?

# Scans for selection

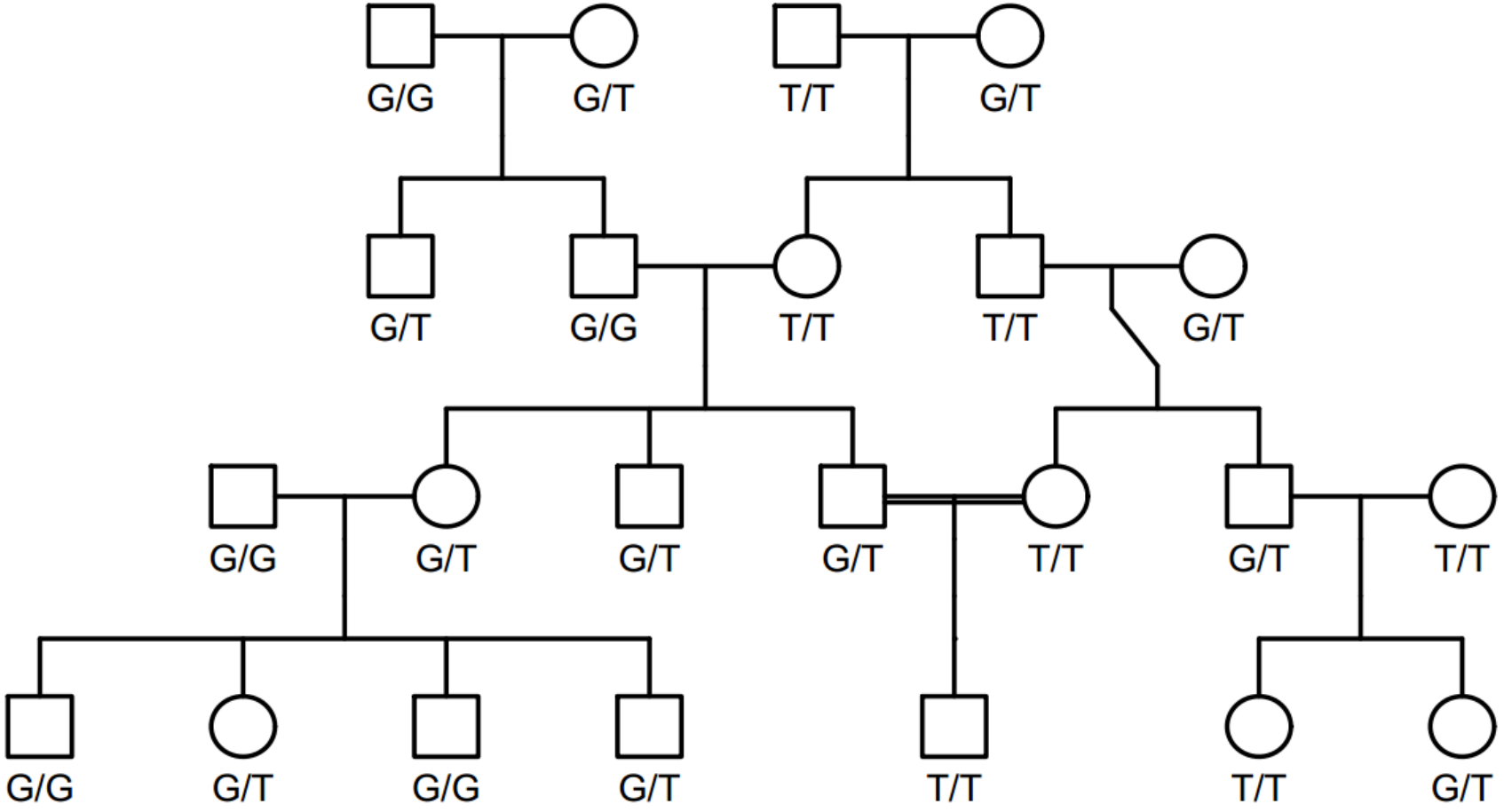
## Different signals at different time depths



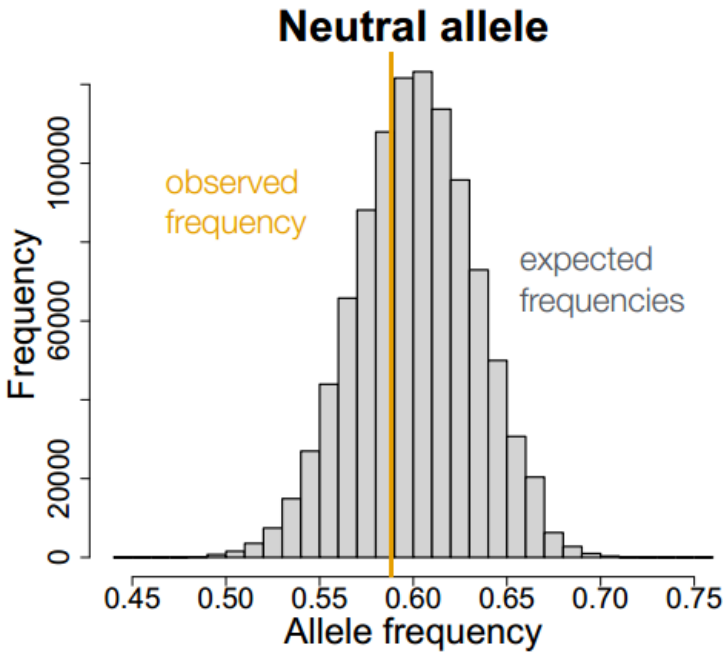
# Testing for selection using pedigree information



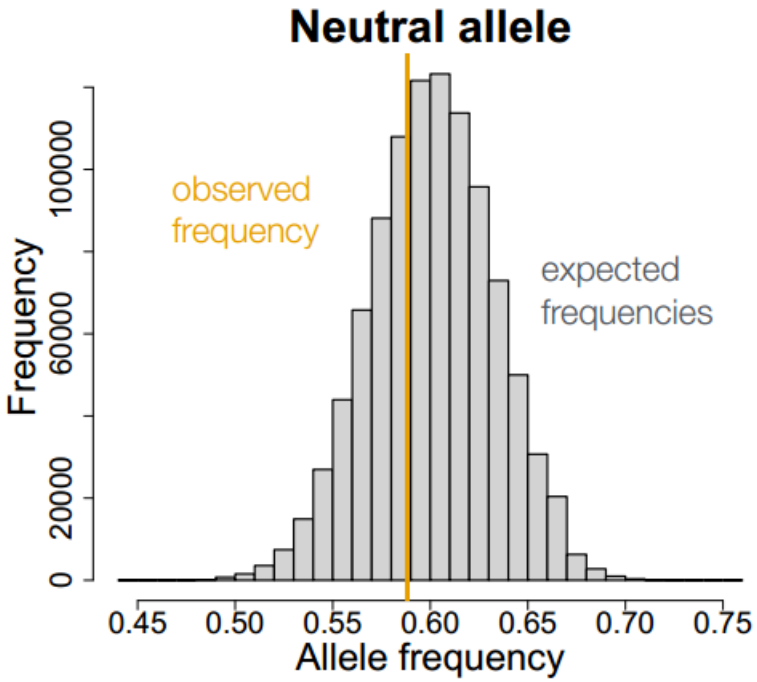
# Gene-dropping to generate neutral expectations

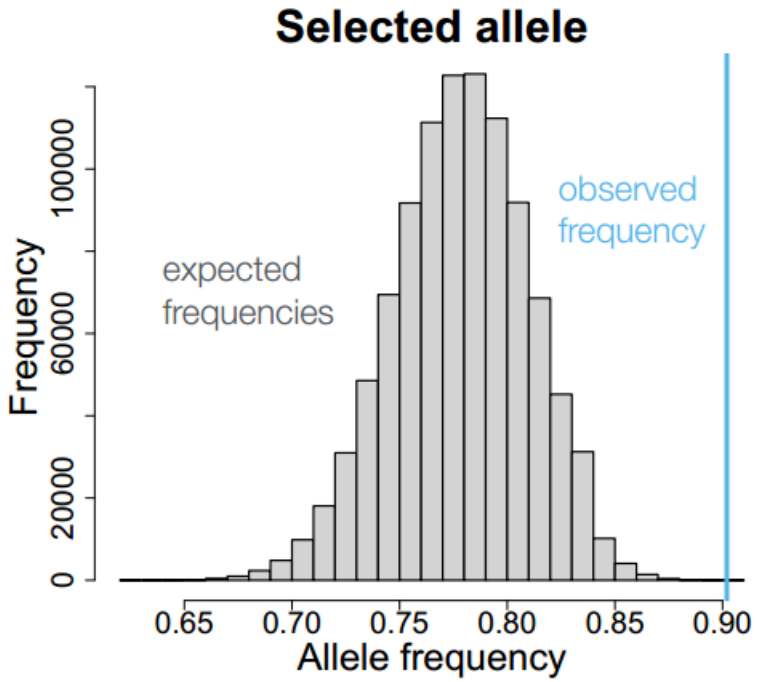


# Most SNPs conform to the null hypothesis

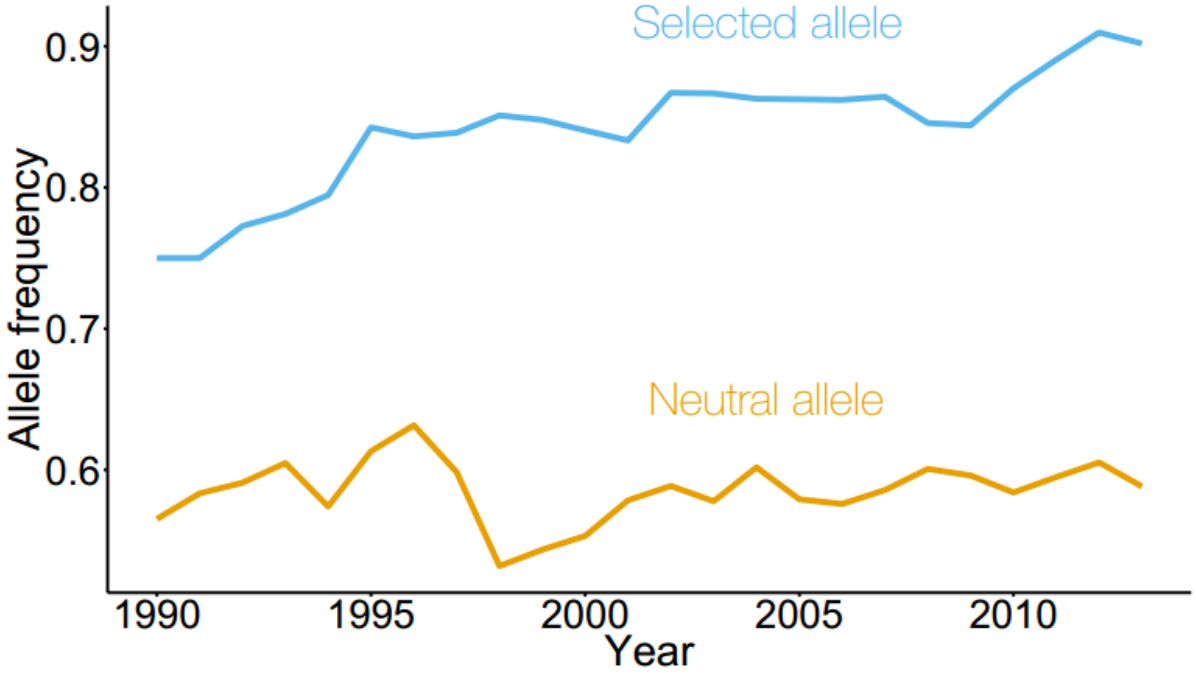


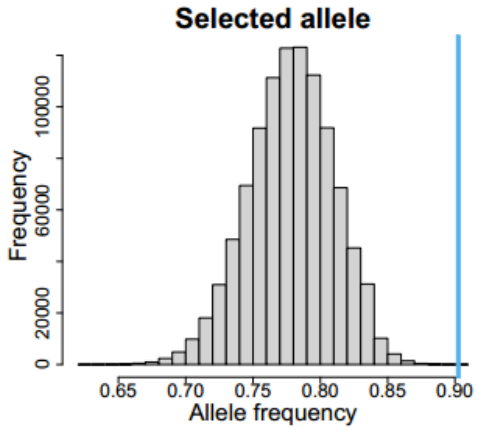
# … but 67 SNPs display a significant departure.

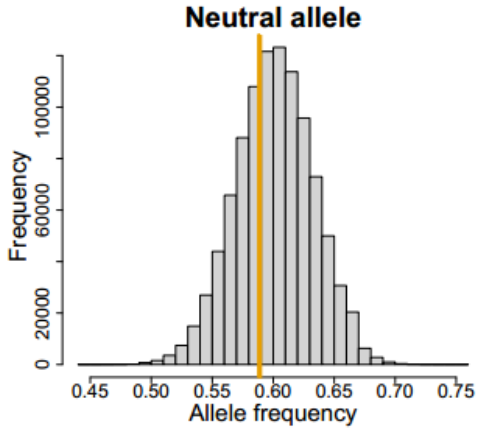


# Evidence of selection at 67 SNPs



# Selection can act at different stages of the life cycle

zygotes

viability selection

adults

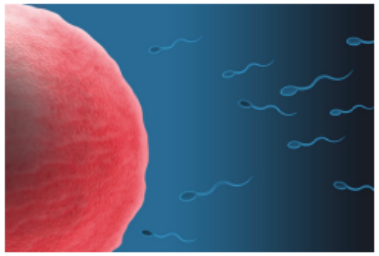
sexual selection

parents

fecundity selection

gametes

gametic selection



# Disease Association

How can variation in genome sequence be used to infer disease risk?

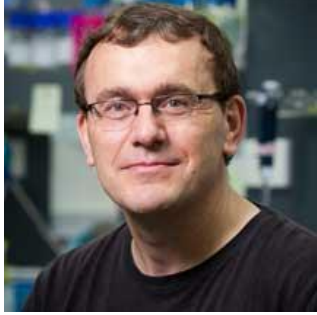
# Population genetics and disease-causing variants

GWAS is fundamentally a population genetic idea – common variants that are causal to a phenotype are likely to show direct or indirect statistical association

Early onset disease-risk enhancing alleles ought to have lower allele frequency.

Rare variants can also be found by association, but require huge samples.

# Variance explained



Peter Visscher

Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index

Jian Yang[1,2,24], Andrew Bakshi[1], Zhihong Zhu[1], Gibran Hemani[1,3], Anna A E Vinkhuyzen[1], Sang Hong Lee[1,4], Matthew R Robinson[1], John R B Perry[5], Ilja M Nolte[6], Jana V van Vliet-Ostaptchouk[6,7], Harold Snieder[6], The LifeLines Cohort Study[8], Tonu Esko[9–12], Lili Milani[9], Reedik Mägi[9], Andres Metspalu[9,13], Anders Hamsten[14], Patrik K E Magnusson[15], Nancy L Pedersen[15], Erik Ingelsson[16,17], Nicole Soranzo[18,19], Matthew C Keller[20,21], Naomi R Wray[1], Michael E Goddard[22,23] & Peter M Visscher[1,2,24]

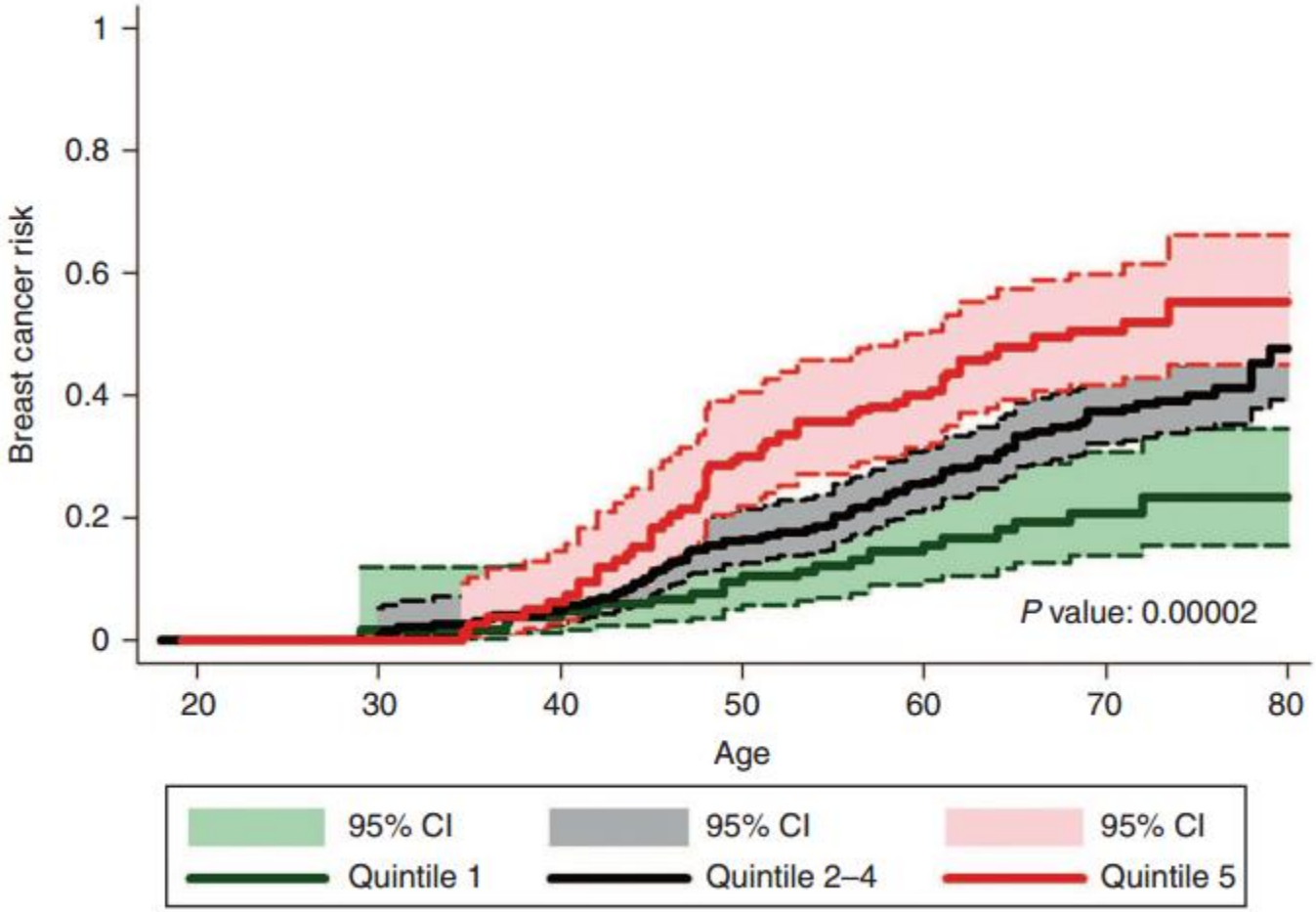
## GCTA

The sum of infinitesimal effects of 17 M imputed SNPs yields 56% of variance.

# GLM and prediction – polygenic risk score

- Purcell *et al*. (*Nature* 2009)
  - Polygenic Risk Score (schizophrenia $R^2$ = 3%)
- Chen et al. (*Genet. Epi*. 2015)
  - Big improvement with ancestry correction
  - hair color $R^2$ = 4-7%, tanning ability $R^2$ = 1-3%, basal cell carcinoma $R^2$ = 1-2%
- Vilhalmsson *et al.* (2015 *AJHG*)
  - Direct modeling LD increases accuracy of PRS

# Polygenic Risk Score and Breast Cancer Risk



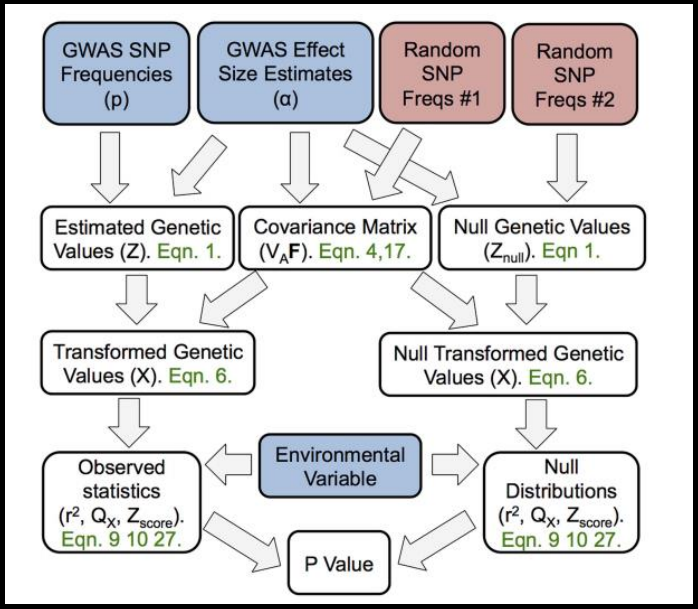

Li *et al*. 2016 *Genetics in Medicine*

# Polygenic selection

Estimate effect size from GWAS

Contrast population from different places or times

Ask if allele frequencies change in a correlated way




Berg & Coop 2014 *PLoS Genetics*

# Evidence for selection on height in humans

GWAS on height finds many SNPs.

Many of these SNPs show a consistent cline from North to South in Europe.

There is a significant tendency of “short” alleles to be found in the south.

Turchin *et al*. 2012 *Nature Genetics*

# Caveats about Polygenic Selection

1. Population stratification in GWAS – errors in effect sizes.

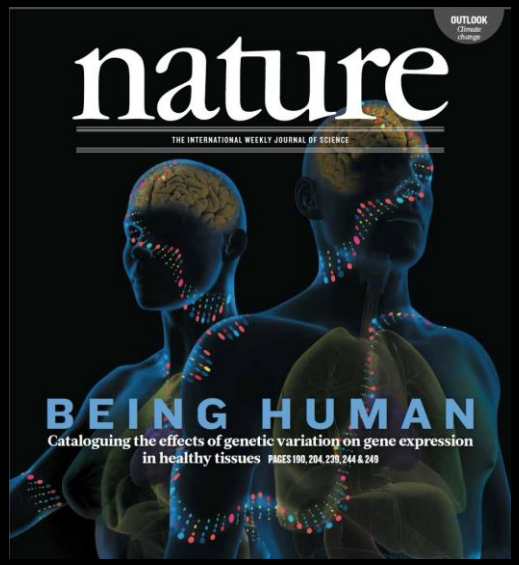
2. Assumes constant effect sizes over space and time.

3. SNPs that impact fitness have inflated ascertainment, producing a bias

Novembre & Barton 2018 *Genetics*

# Genome functional variation

How can analysis of variation at the sequence and gene-expression levels inform us about genome function?

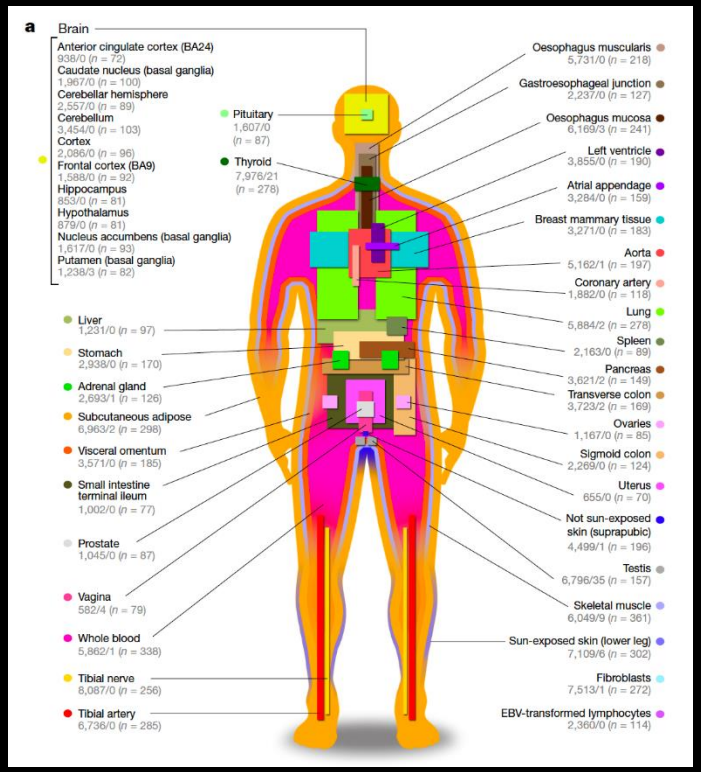



# GTEx (Genotype-Tissue-Expression) Project

- Pilot (2015): 1658 samples from 175 donors
- Mid-phase (2017) : 7,051 samples from 450 donors
- Final (ongoing analysis): 17,382 from 838 donors

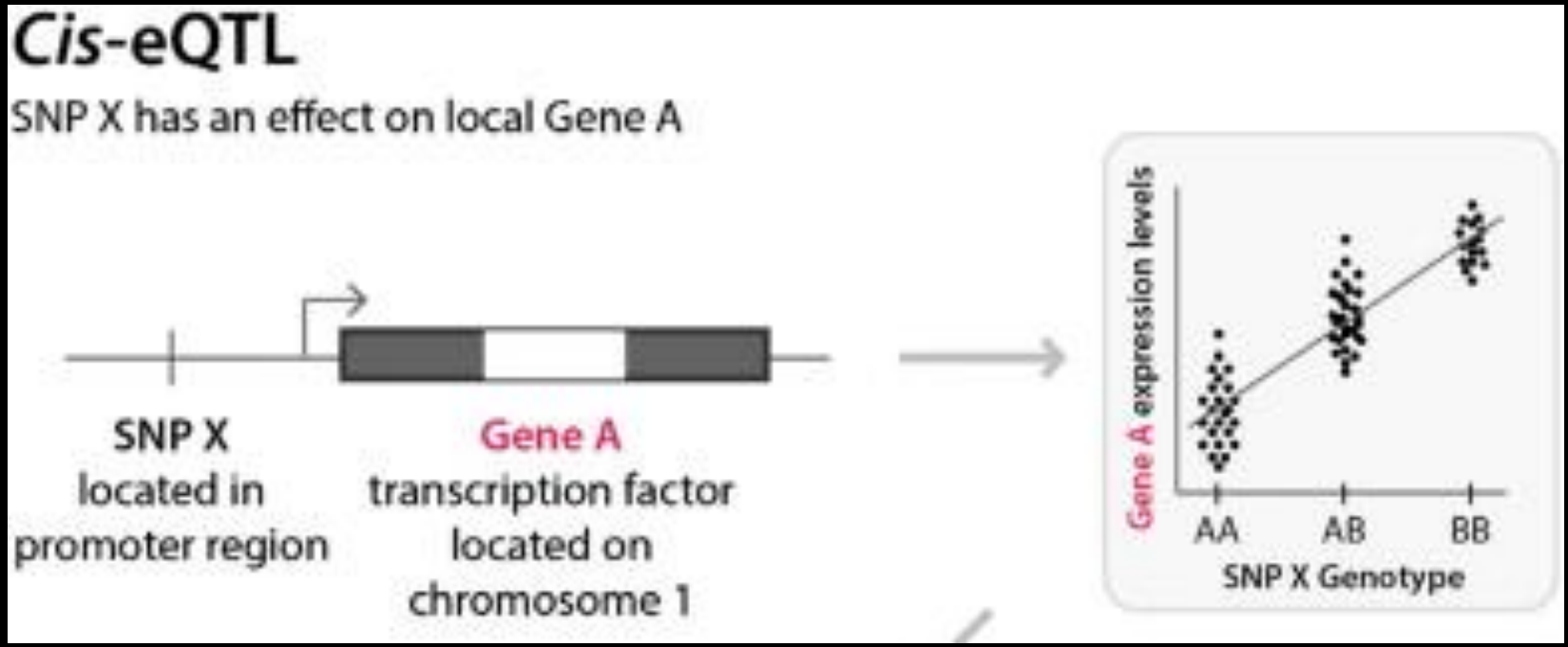
- Transcriptome: mRNA-seq
- Genotyping: < 2017: arrays,WES, WGS

- Banked biospecimens & procurement protocols
- Raw data (dbGap) + summary data (GTEx Portal) + browsers (Portal, UCSC, Ensembl)

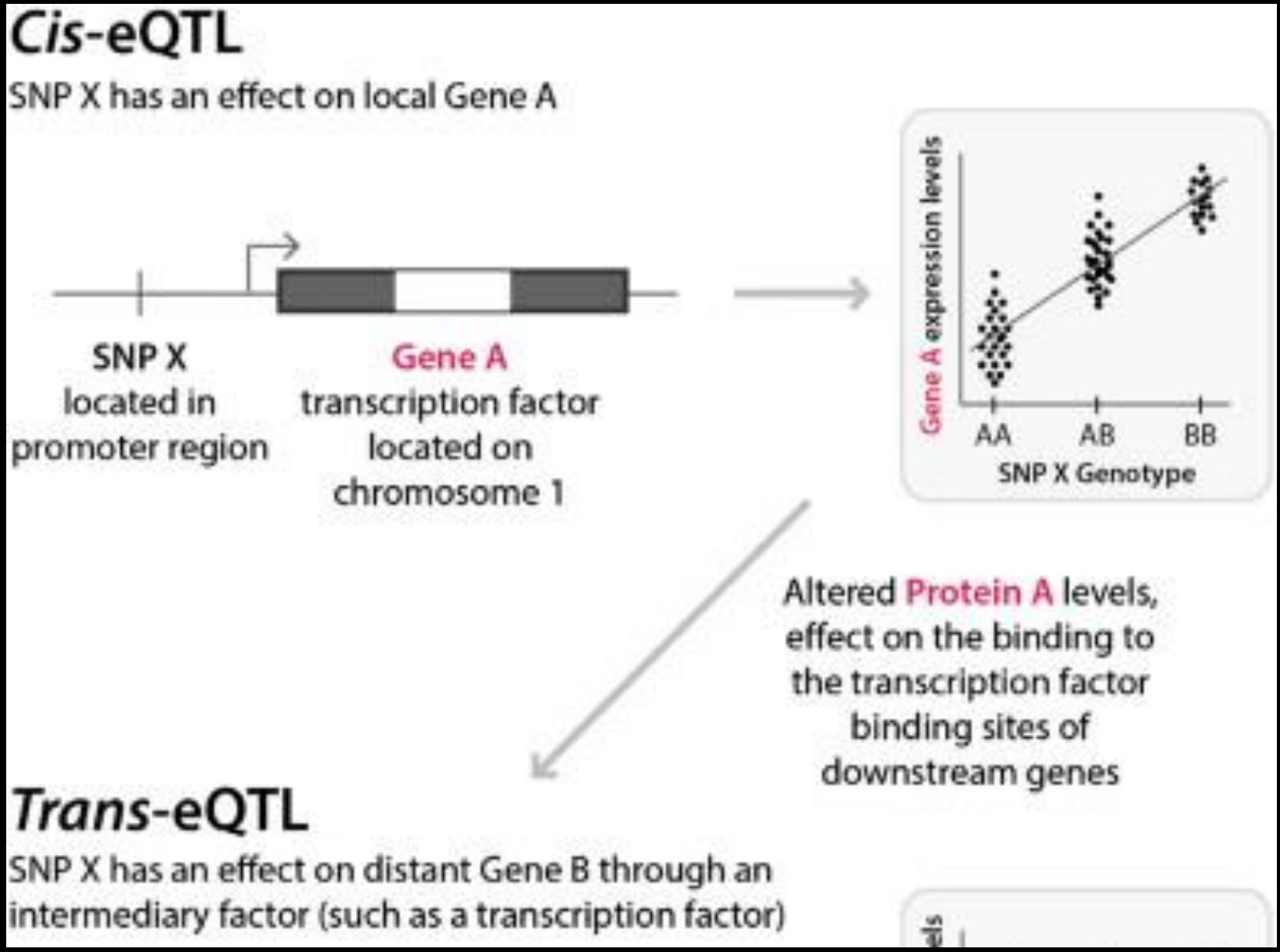



Courtesy of Tuuli Lappalainen

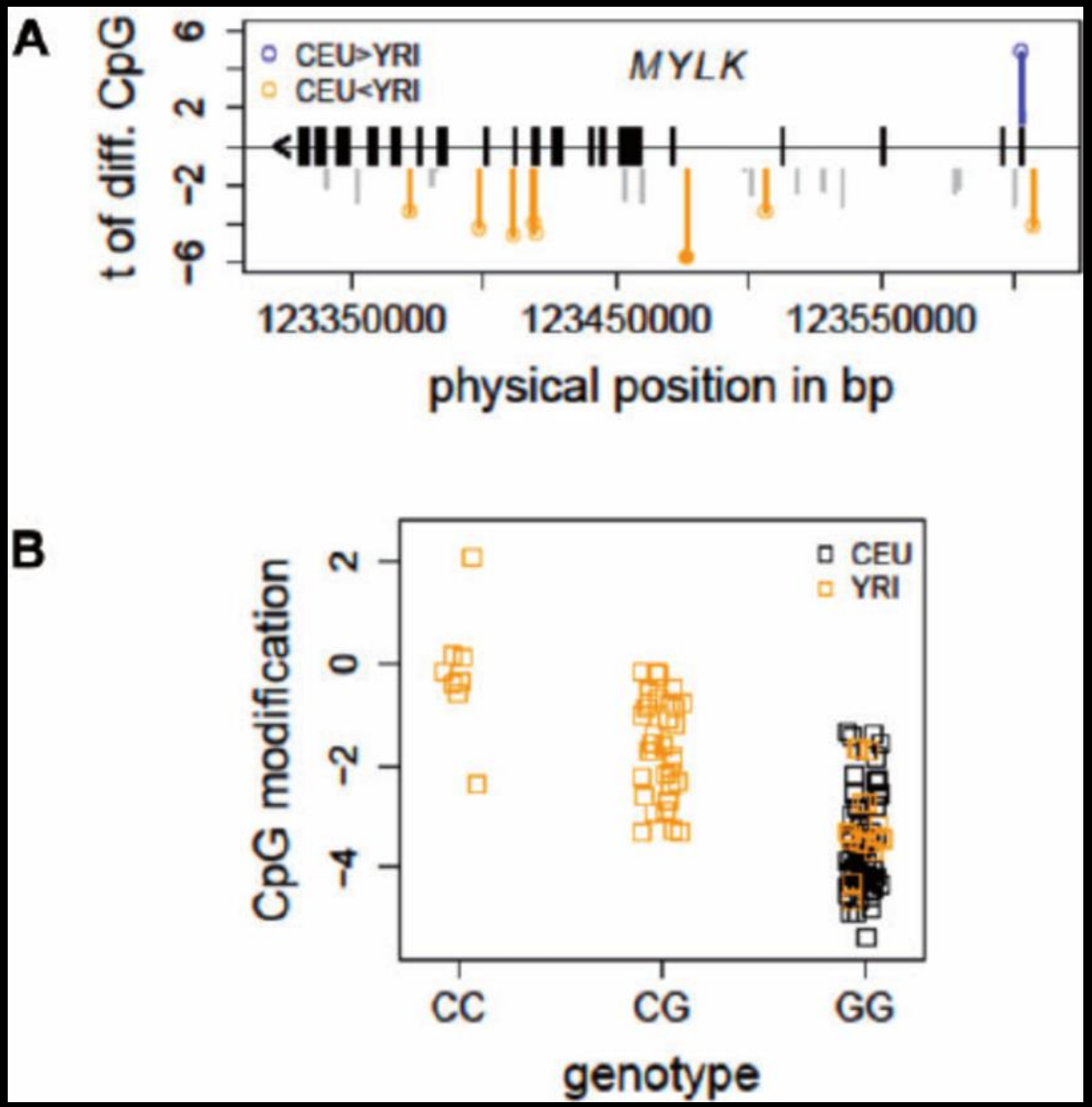
# Expression QTL have revealed regulatory variation



# Expression QTL have revealed regulatory variation

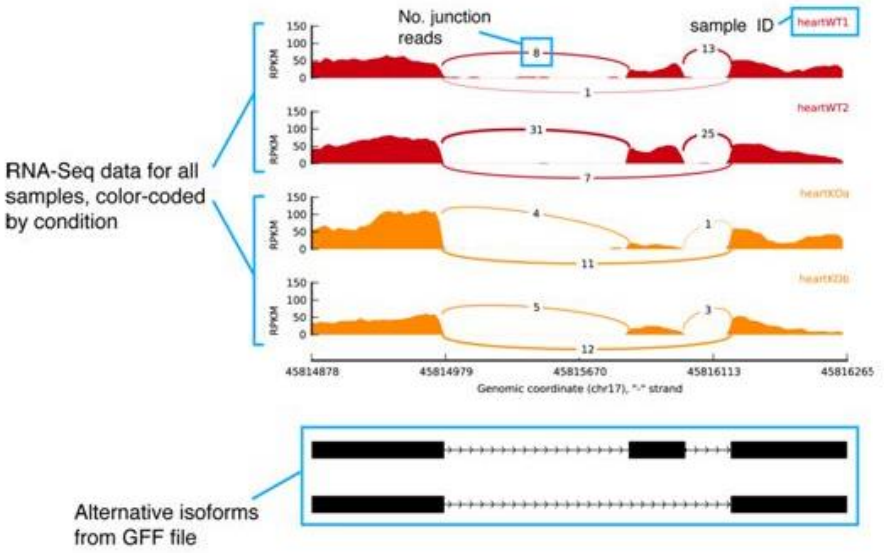


# Methylation QTL (mQTL)



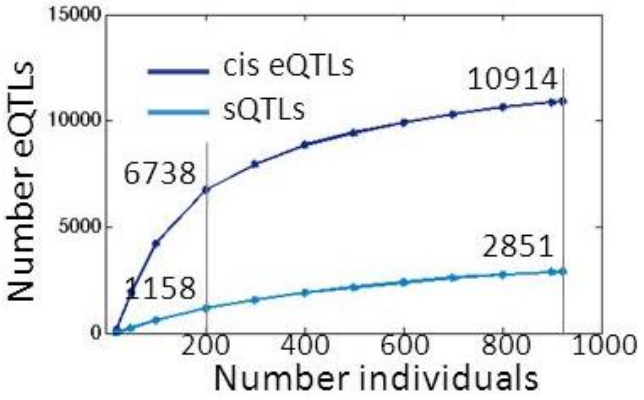
# Splicing QTL – polymorphism for splicing

Can investigate relative transcript ratios or reads across junctions.



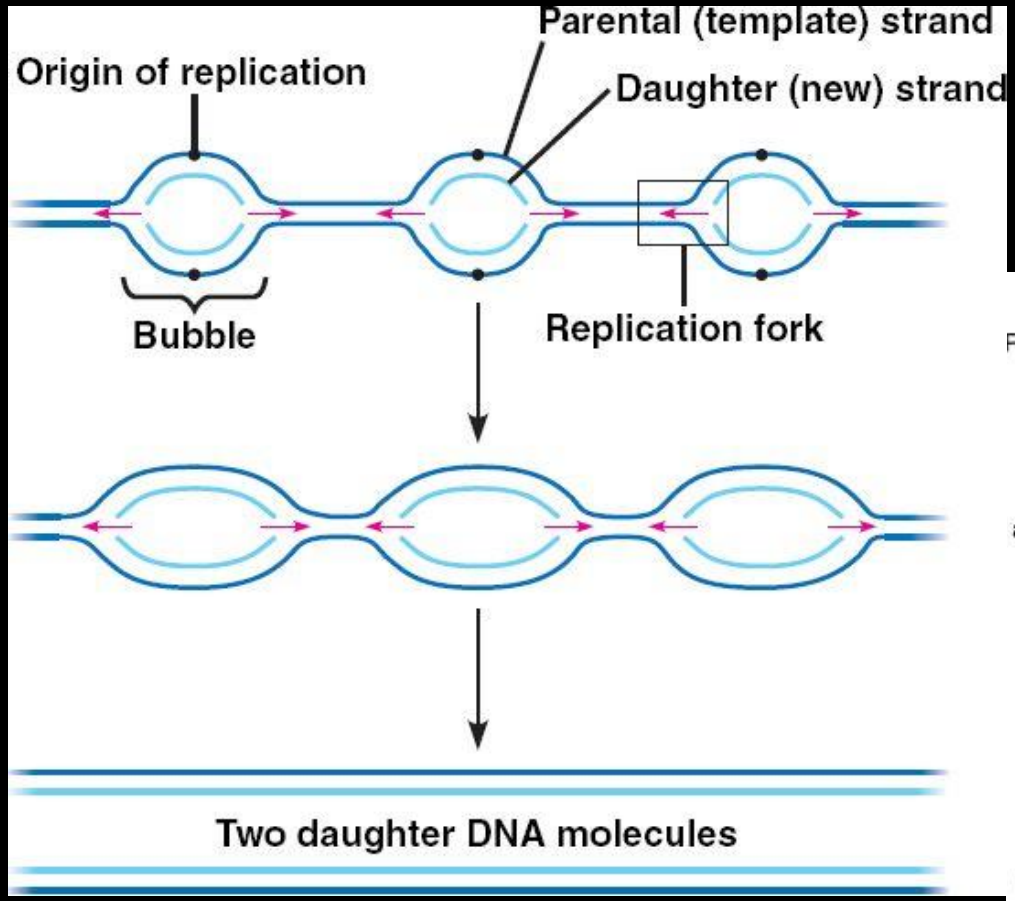

Katz et al, Nature Methods, 2010

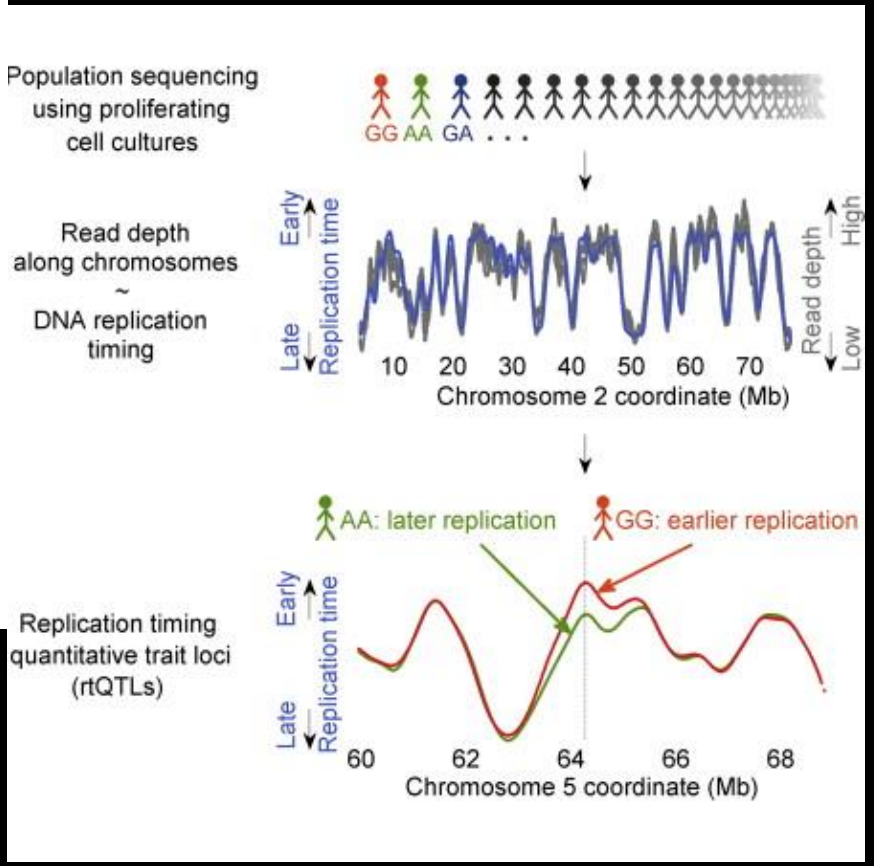
- Splicing also affected for many genes




Battle et al, Genome Research, 2014

# Replication timing QTL – variation in DNA copying!




Koren *et al*. 2014 *CELL*

# Outline

- Demographic inference
- Population structure and history
- Admixture/ Introgression
- Random genetic drift
- Natural selection
- Mutation spectrum
- Disease association
- Genome function