Reconstructing the past: deep learning for population genetics

Flora Jay and Guillaume Charpiat
LRI (Bioinfo and TAO)

CDS Pitching Day

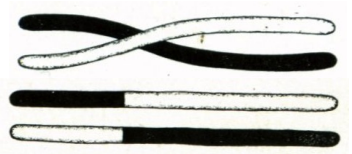
Overview

1. Introduction: past demography and genetics
2. Extracting information from genomic data
3. State of the art: summary statistics (hand designed) + ABC
4. Deep learning
    4a. Challenges: variable input size
    4b. Desired properties: invariances
    4c. Plan
5. Summary

Introduction

Population genetic and demography

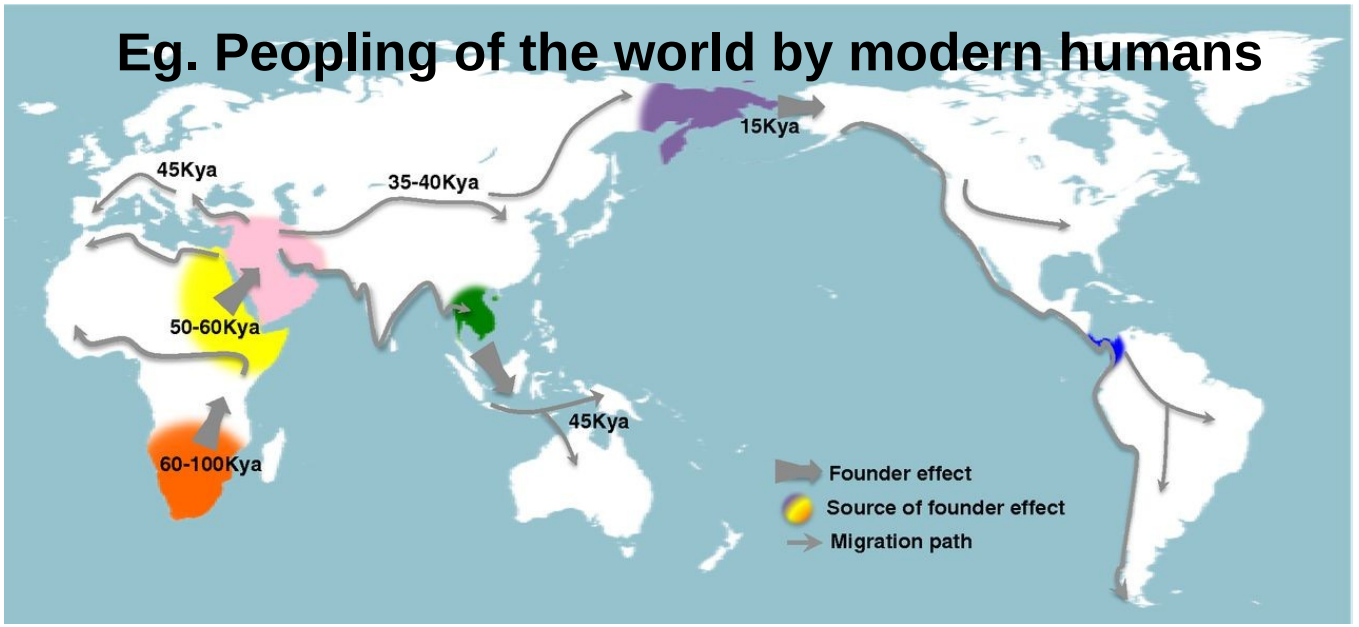
MUTATION
RECOMBINATION



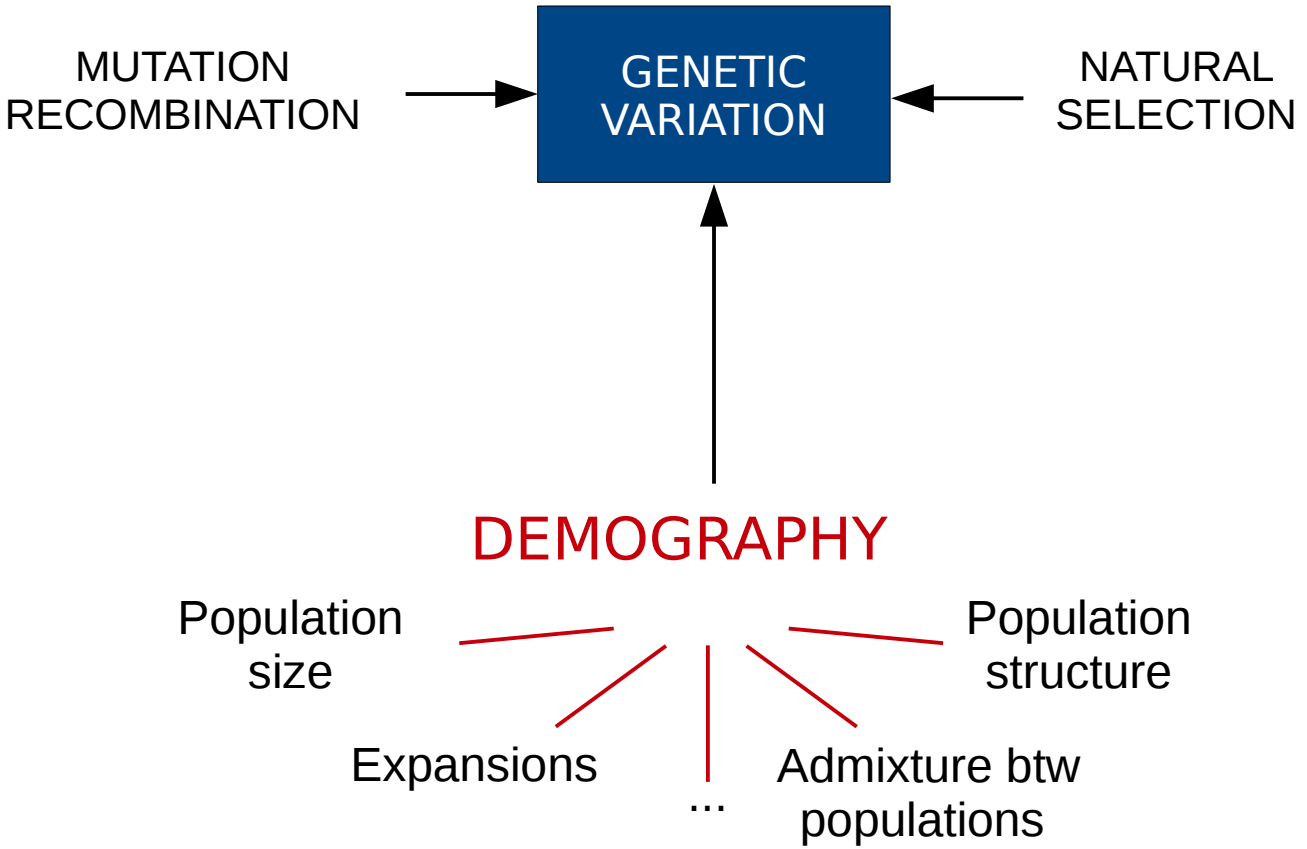
GENETIC VARIATION

NATURAL SELECTION

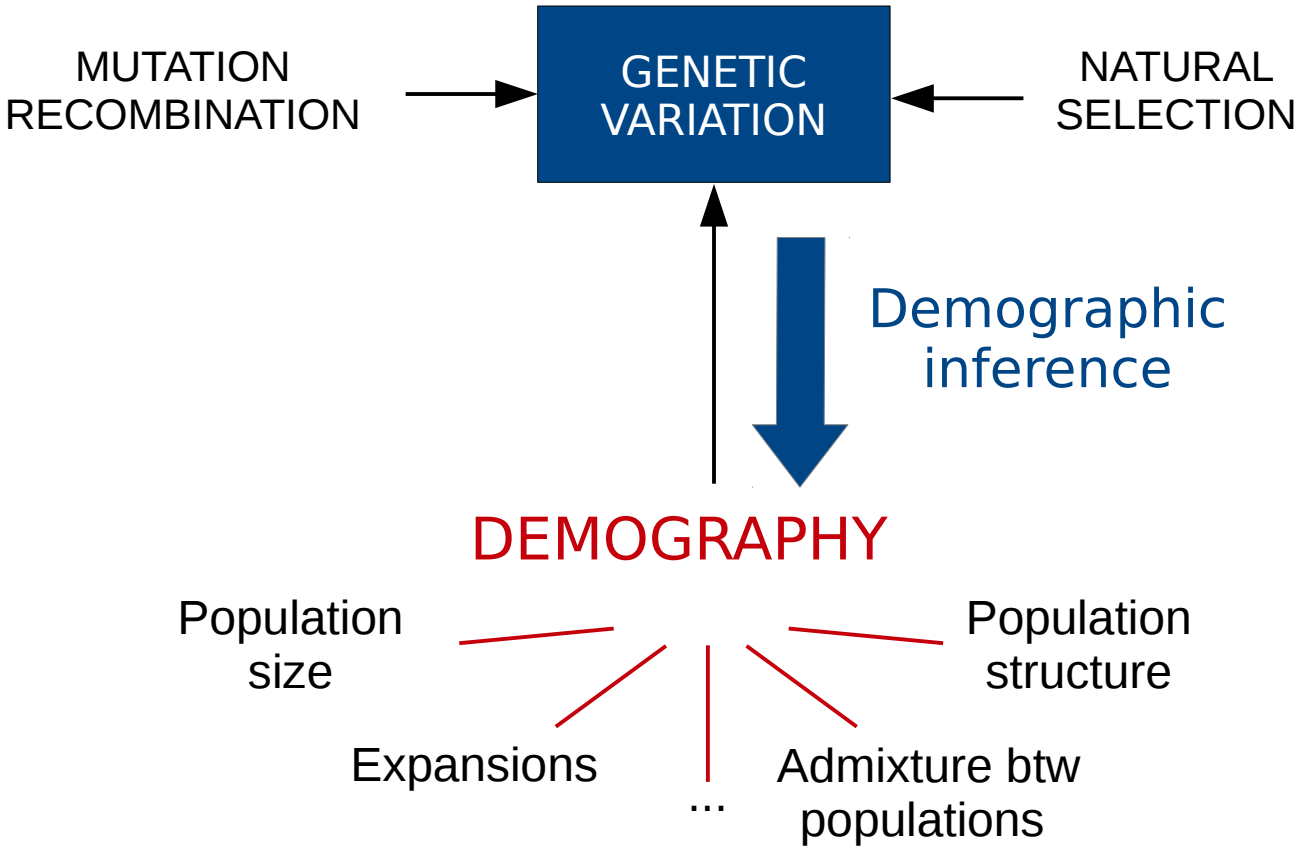
DEMOGRAPHY



Population genetic and demography



Population genetic and demography



Methods and applications

Identify events, date them, estimate their strength, ...

Why inferring past demography ?

If you love at least one of the following...

- History — *eg. did we mingle with Neandertals 50k year ago while peopling earth?*
- Medicine — *eg. is there a mutation increasing the risk of getting breast cancer?*
- Evolution — *eg. are Tibetan adapted to altitude and why? Are plant populations adapted to their environment and what could be the impact of climate change?*

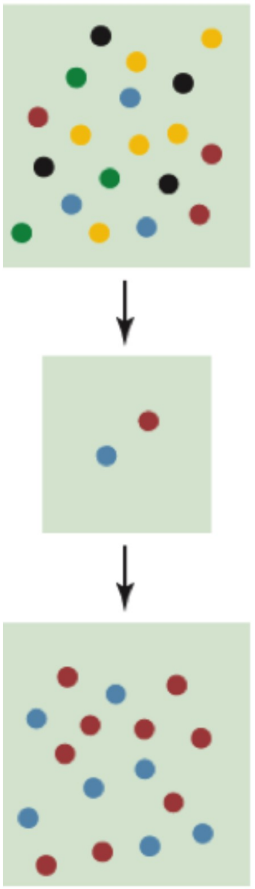
then you already have a good reason for trying to infer demography!

→ *gives a null model to test non-neutral hypotheses*
eg. observed signal at a gene due to [demography] versus [demography+selection] ?

Where is the information?

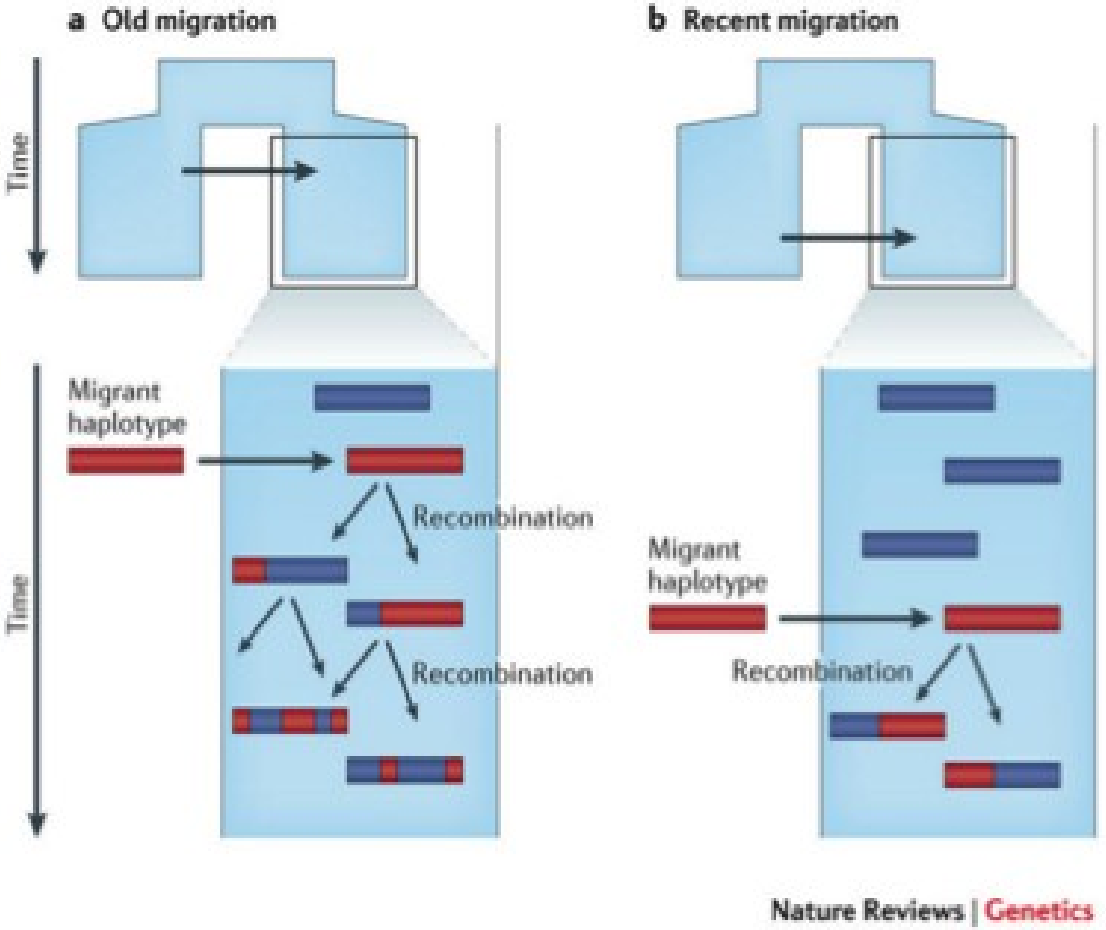
Past demography leaves signatures in genetic data

- Population sizes



© Slatkin

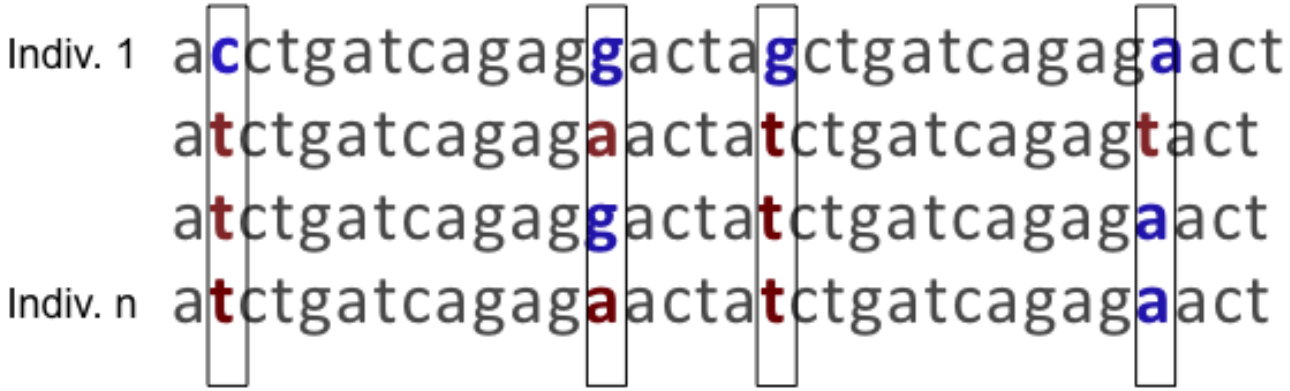
- Migration / admixture between populations



Sousa & Hey 2013

Project:
Inferring demography using whole genomes

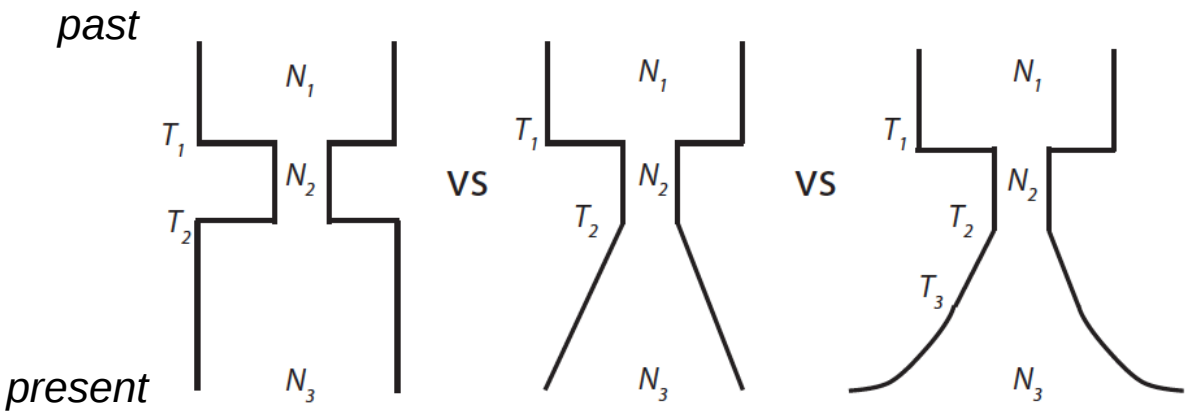
- Develop a method using sequence data...

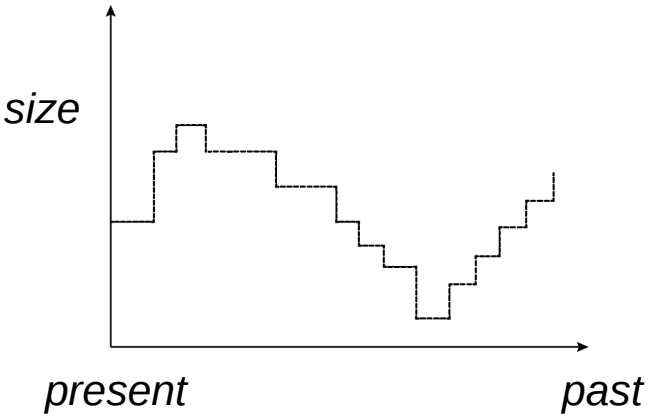


Dim ~ n x 3Gb

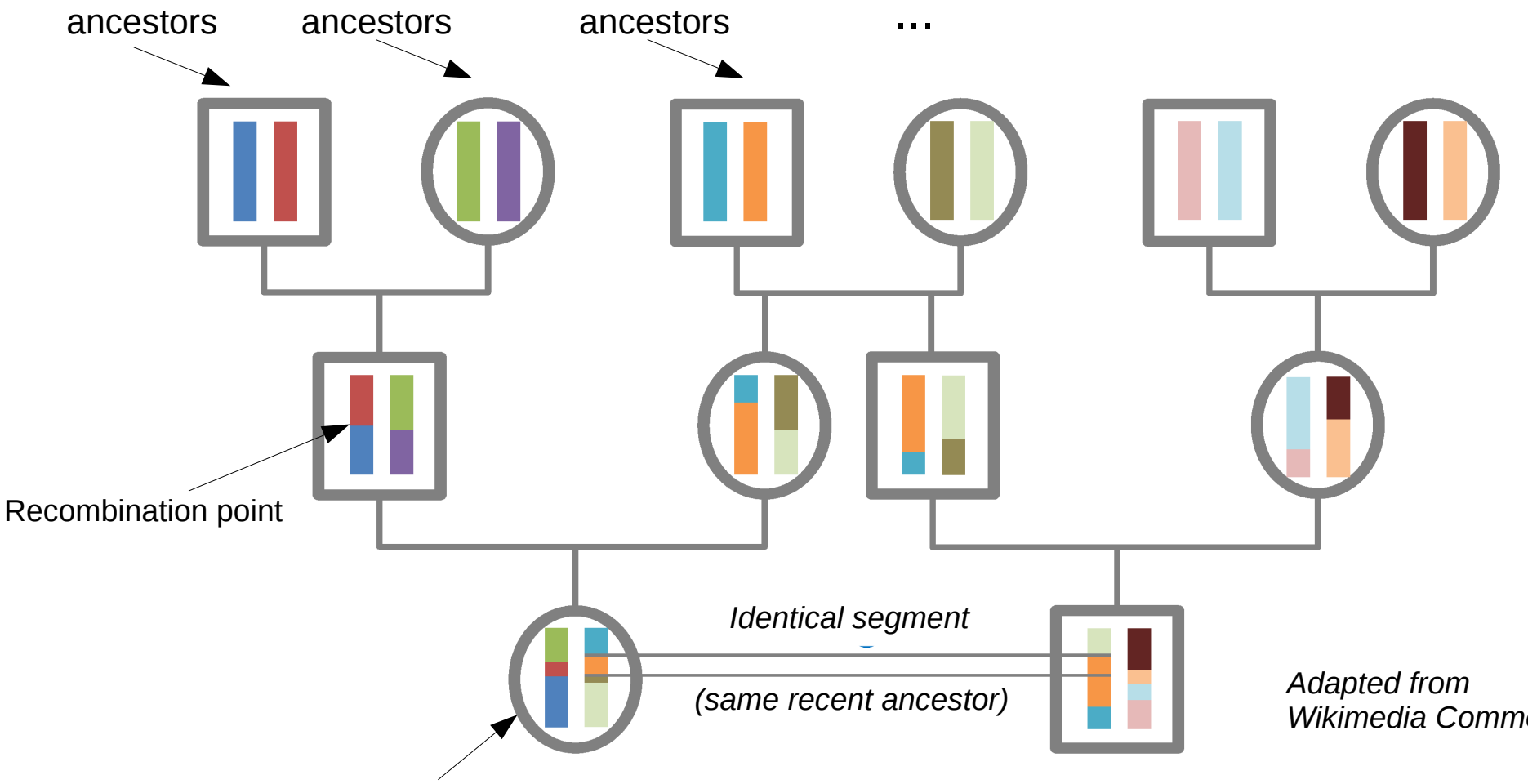
Mutation
→ polymorphism (SNP)

- ...to identify complex demographic histories.



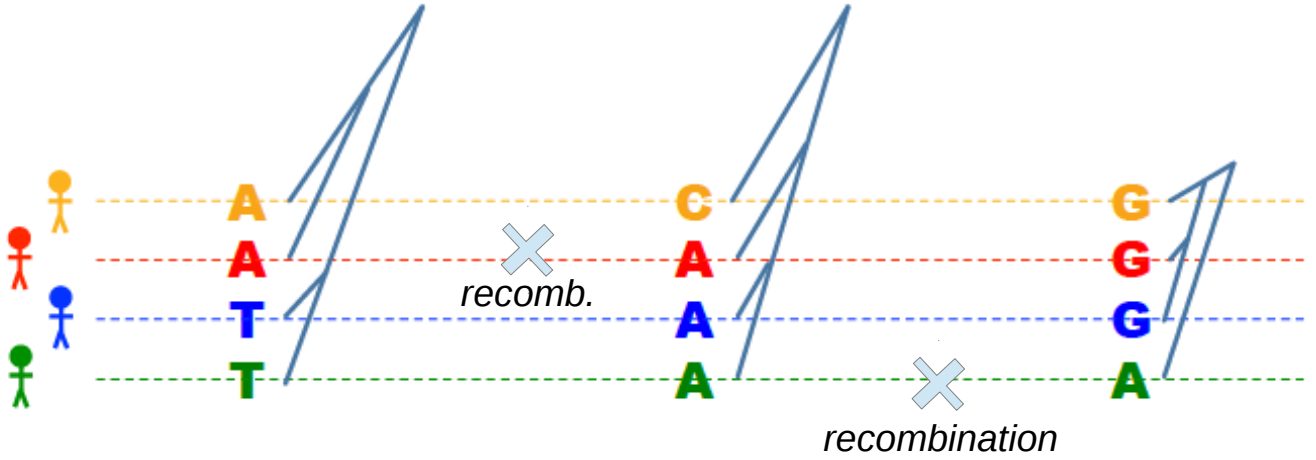


Genetic process



Genetic process

- Recombination along the DNA sequence at each generation
  → not a single tree for the whole genome BUT <u>multiple genealogies</u>

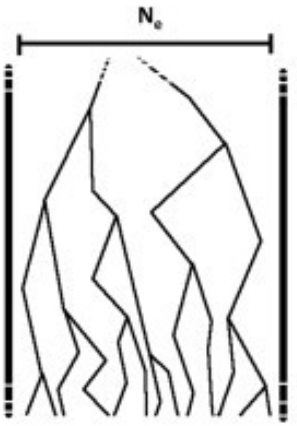


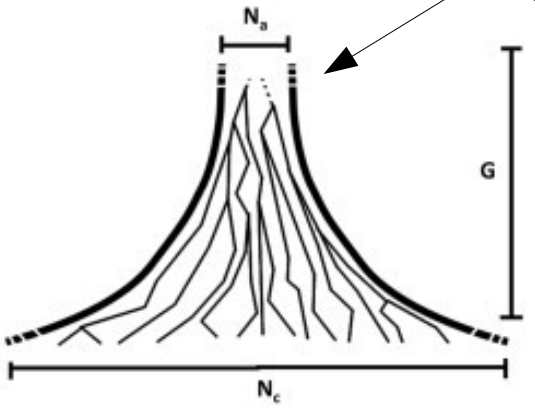
Not observed

© Sheehan

- Genealogies are influenced by <u>past demography</u>

More coalescence when population is small





Palamara et al. 2012

Inferring demographic history

The approach: Combining different type of information
to learn about hidden genealogies and thus demography

(1) from “expert statistics” (past and current research)
(site-frequency spectrum, linkage disequilibrium, diversity, ...)
Using Approximate Bayesian Computation

(2) by learning interesting features from raw data (PROJECT)
Deep learning: build a deep neural network tuned for
population genetic data

State of the art: summary statistics + ABC

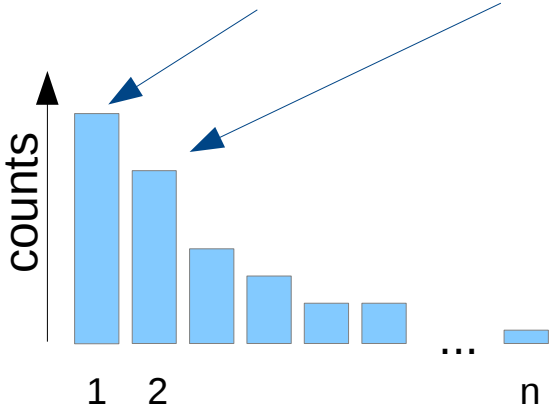
Understanding the data - Summary statistics

Goal: extract information from genomes about hidden genealogies

DNA is spatially structured

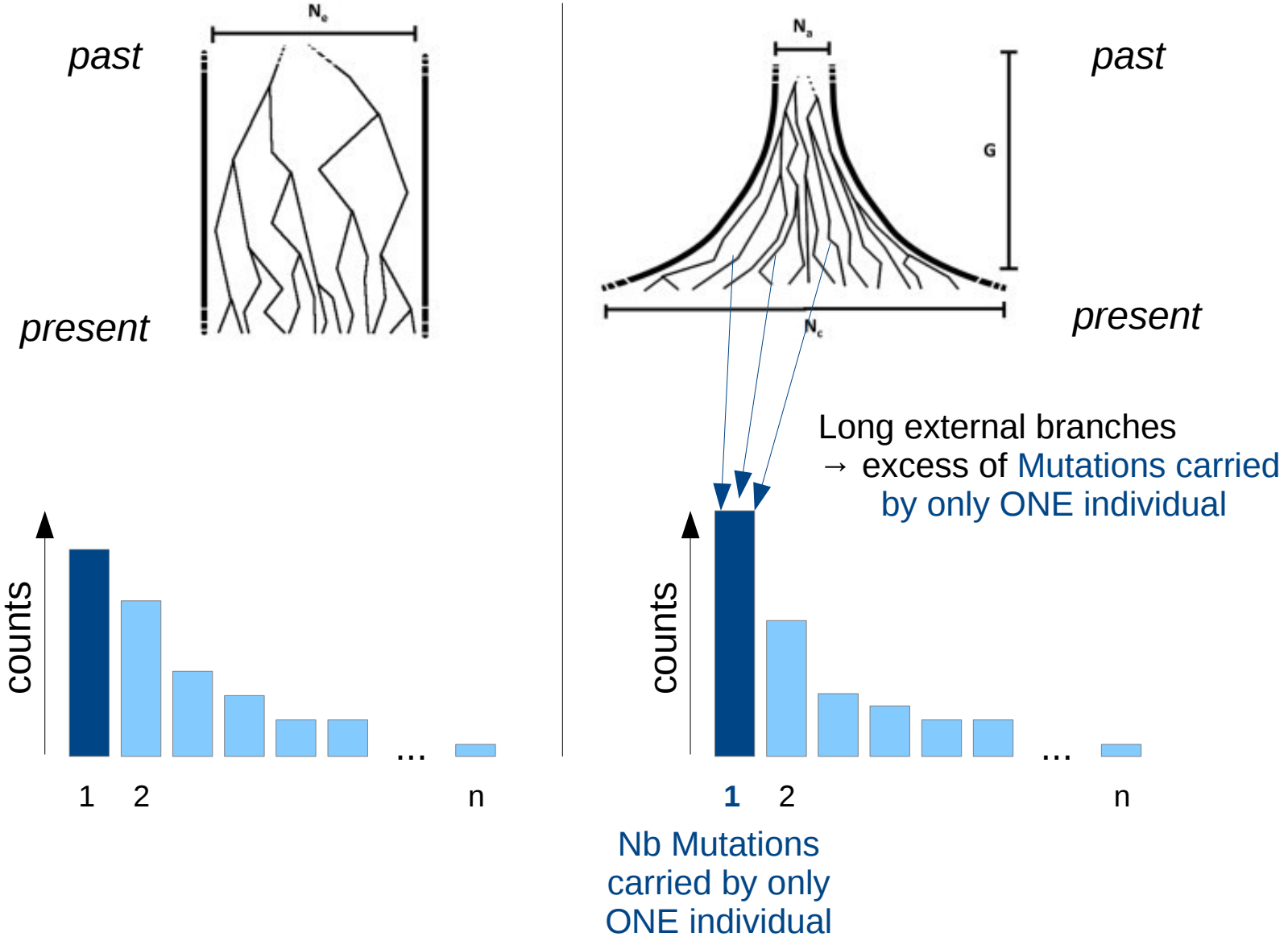
1. Distant sites: **distant** histories (different evolutionary trees)
   → site frequency spectrum = histogram of allele counts

Counts histogram: 2 singletons, 1 doubleton, ...



Summary statistics

1. histogram of allele counts at all segregating sites



Summary statistics

DNA is spatially structured

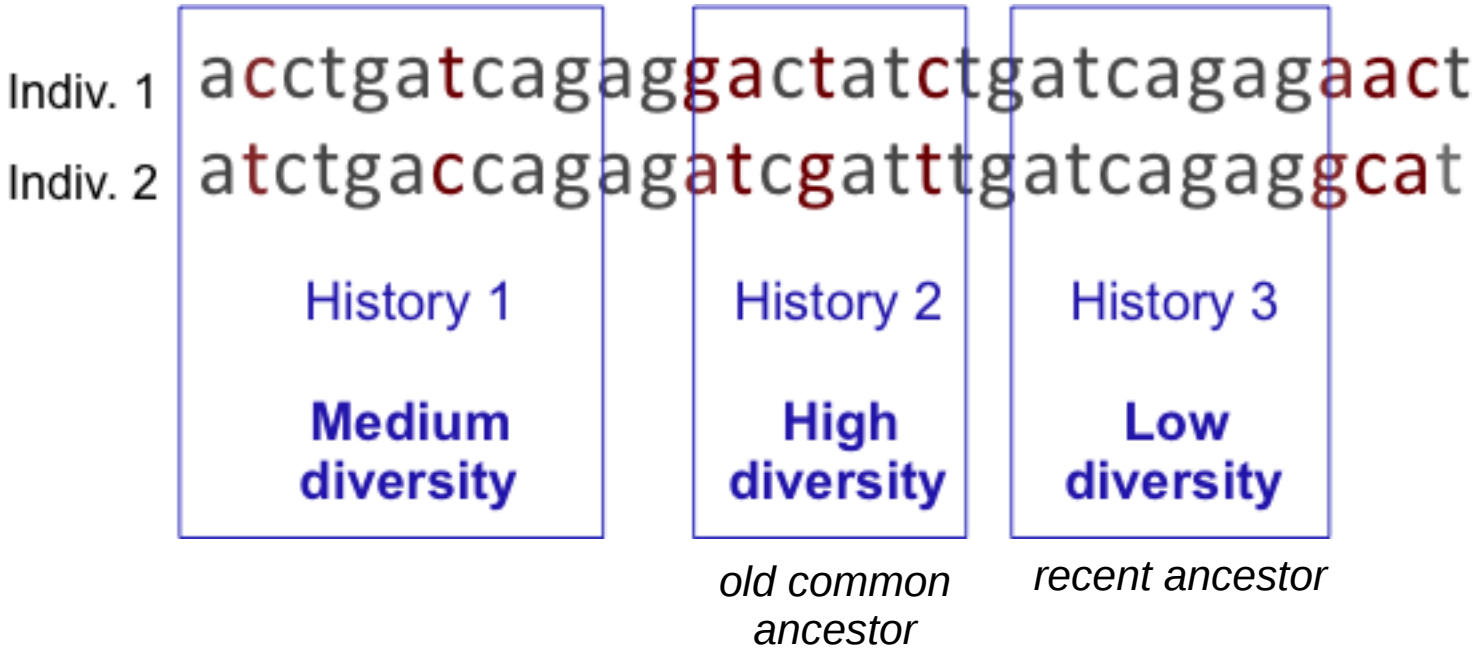
1. Distant sites: **distant** histories (different evolutionary trees)

2. Less distant sites: **related** histories
   → linkage disequilibrium (correlation between SNPs)

Summary statistics

DNA is spatially structured

1. Distant sites: **distant** histories (different evolutionary trees)
2. Less distant sites: **related** histories
3. Adjacent sites share the **same** history
   → Diversity per regions



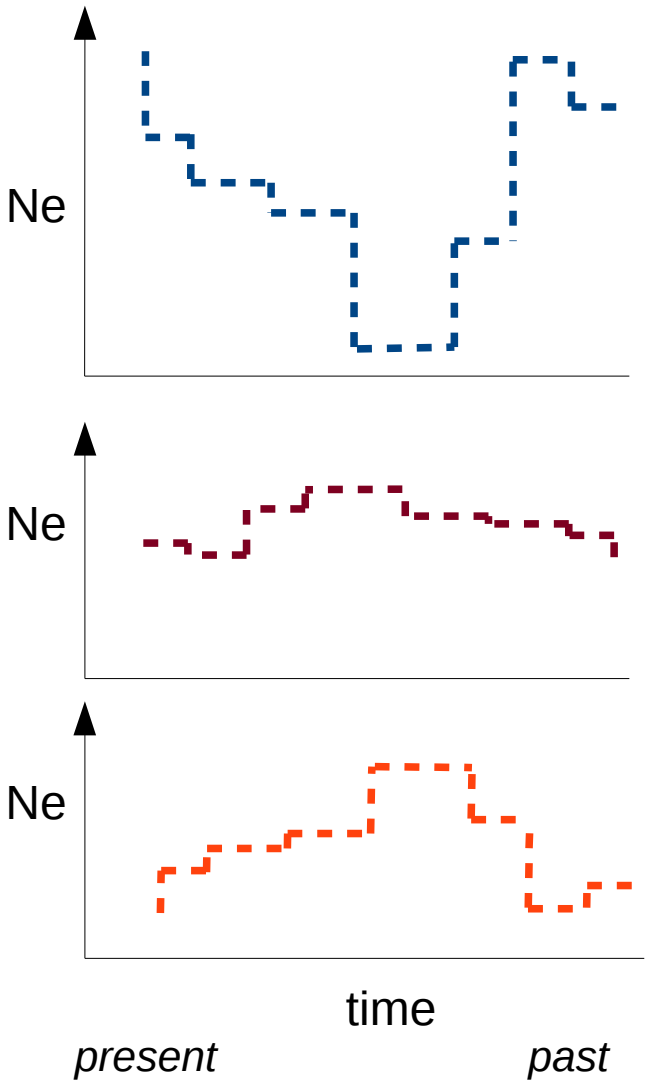
Summary statistics

1. Distant sites: **distant** histories (different evolutionary trees)
2. Less distant sites: **related** histories
3. Adjacent sites share the **same** history
   → Diversity per regions

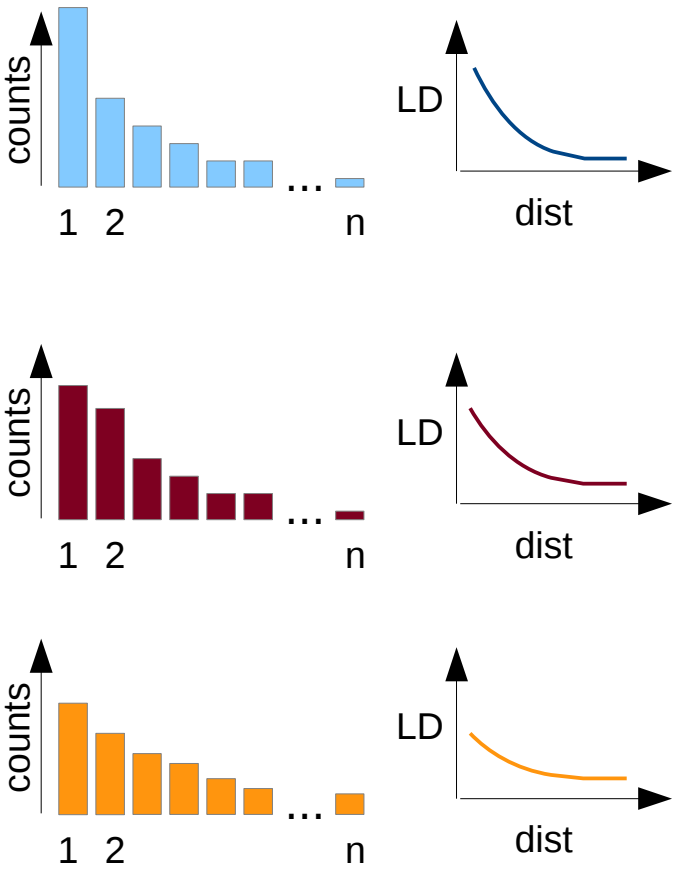
And so on...

Approximate Bayesian Computation (ABC)

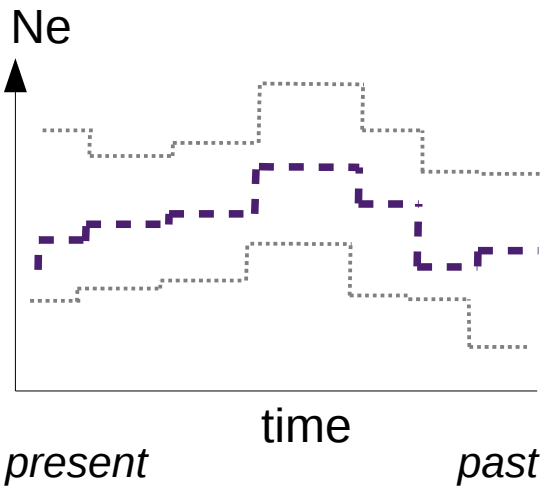
Generate randomly thousands of histories



Compute summary statistics



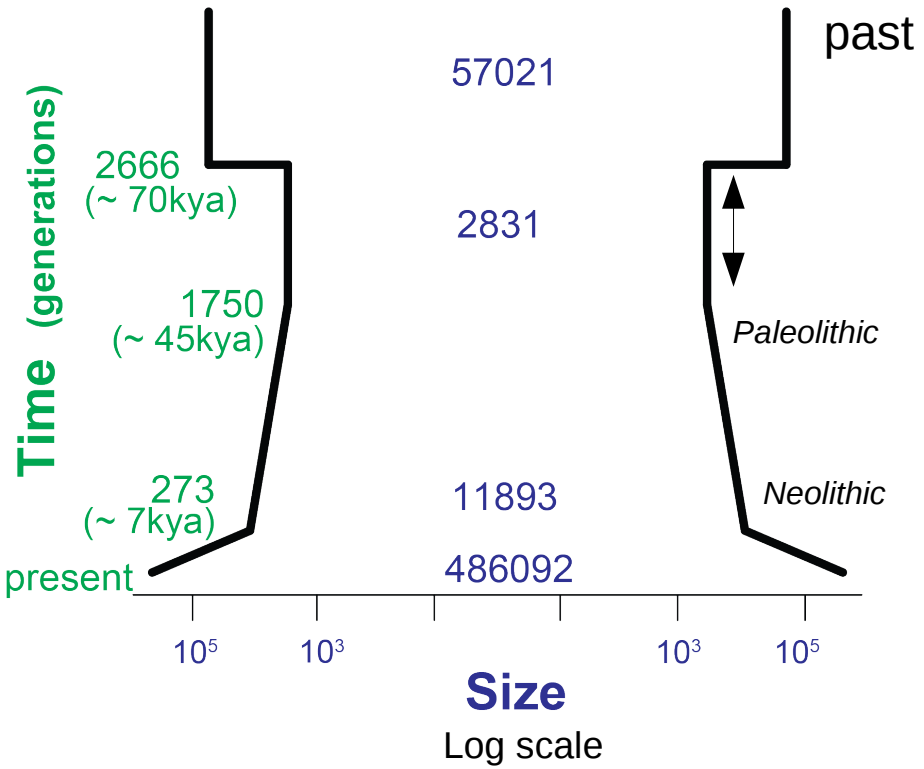
Keep histories that produce sum. stat. similar to real ones
Infer history



Application to real data

Human data

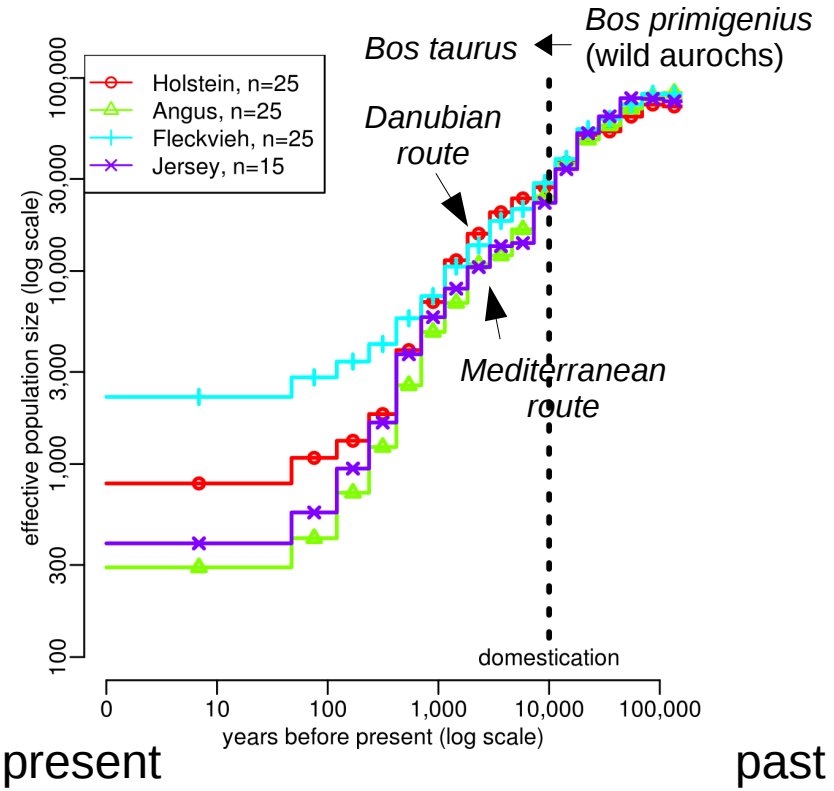
Bottleneck + Paleolithic and neolithic expansions in European data



Jay et al (in prep)

Cattle breed data

population decline



PopSizeABC
Boitard, Rodríguez, Jay et al.
(PloS Genet. 2016)

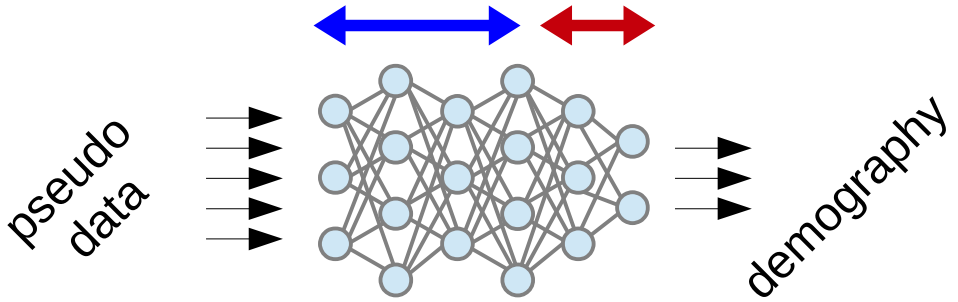
Project: Deep Learning instead

Project

Learn a **global relationship** between the **data** and the demographic **parameters** with a **deep** neural network

Learning

- data features
- functions predicting demographic model/parameters/...



Multi-layer neural networks

Project

Learn a **global relationship** between the **data** and the demographic **parameters** with a **deep** neural network

Motivation from previous work in population genetics:

- Improvement when using a simple neural network inside ABC to learn locally the relationship between the summary statistics S(.) and the demographic parameters
  (Blum&François 2010, Boitard et al 2016, Jay et al in prep, …)
- A global relationship can be learned between S(.) and popgen. parameters using deep neural networks (fully connected) (Sheehan&Song (2016))
  → they get rid of the rejection step
  (Method tested on coarse demographic models only)
- **Natural next step: learn automatically the features from raw data**
  → get rid of S(), gain information?

Project - Challenges

Learn a **global relationship** between the **data** and the demographic **parameters** with a **deep** neural network

Challenges due to input data:

- Raw genetic data are large (larger than images)
- Number of sequenced individuals vary
- Length of sequences vary

→ Need for flexibility & generalization w.r.t. input size

→ e.g.: if knowing how to predict past demography for sets of 10 sequences of length 100 000 000, want not to start from scratch for a new set of 9 sequences of length 70 000 000.

→ Recurrent networks can somehow deal with (1D) variable length, but not necessarily suited here (2D, more information by contemplating a whole column at once...)

Project – Desired properties

Learn a **global relationship** between the **data** and the demographic **parameters** with a **deep** neural network

Incorporate coalescence knowledge in the architecture:

- Invariance by translation along the genome
- Invariance by permutation of the individuals*
- Correlation decreases with distance (but rate depends on the demography)
- ...

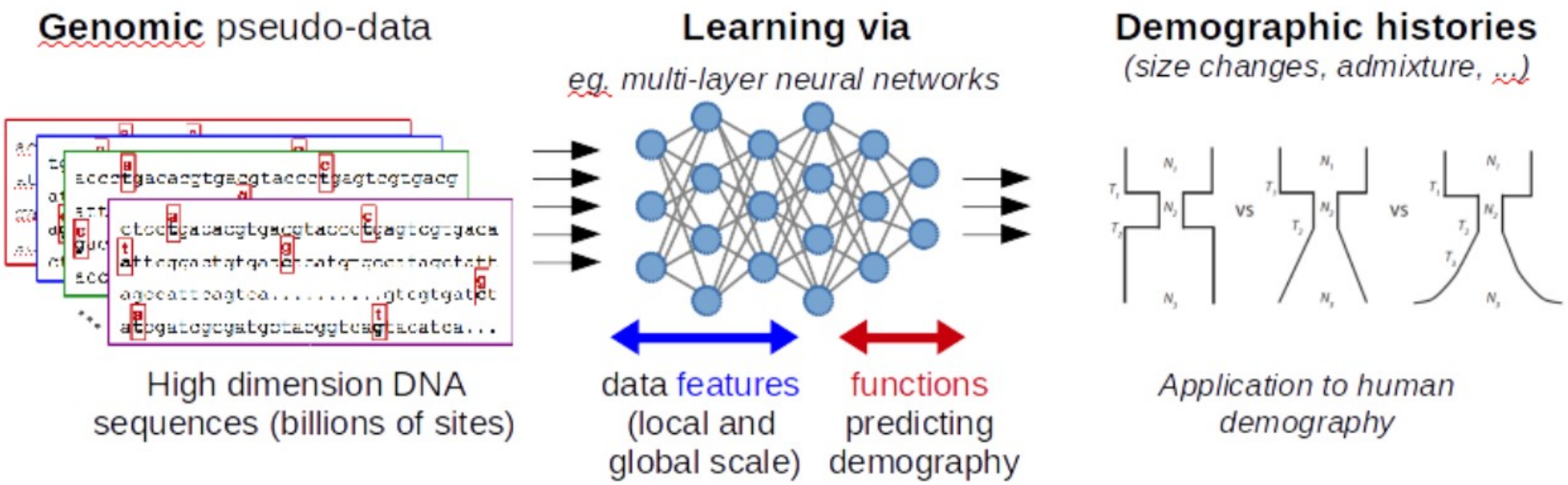
* but maybe not on the permutation of the haplotypes

Project - Plan

Learn a **global relationship** between the **data** and the demographic **parameters** with a **deep** neural network

- Step 1: describe all **properties specific** to genetic data AND population genetic data to be used in the DNN
- Step 2: practical test with common DNN layers such as recurrent and convolution layers
- Step 3: study flexible architectures that scale to different input size
  - → naturally **scalable node functions** (e.g.: max, average, variance...)
  - → training a ***family*** of neural nets, by combining pre-defined families of layers (indexed by input size)
  - → **meta DNN** generating architectures (take as input the data size and outputs a neural network)

Summary



Asked funding: Grant for a Master internship = 3000 euros

References:

- Boitard S, Rodriguez W, Jay F, et al. Inferring population size history from large samples of genomewide molecular dataan approximate Bayesian computation approach. PLoS Genet. 2016 12(3):e1005877.
- Sheehan S, Song YS. Deep learning for population genetic inference. PLoS Comput Biol. 2016 12(3):e1004845.
- Stanley, Kenneth O., David B. D'Ambrosio, and Jason Gauci. A hypercubebased encoding for evolving largescale neural networks. Artificial life 15.2 (2009): 185-212.