14 March, 2016: Introduction to course

529053 Evolutionary Genomics

Topics and aims of this course

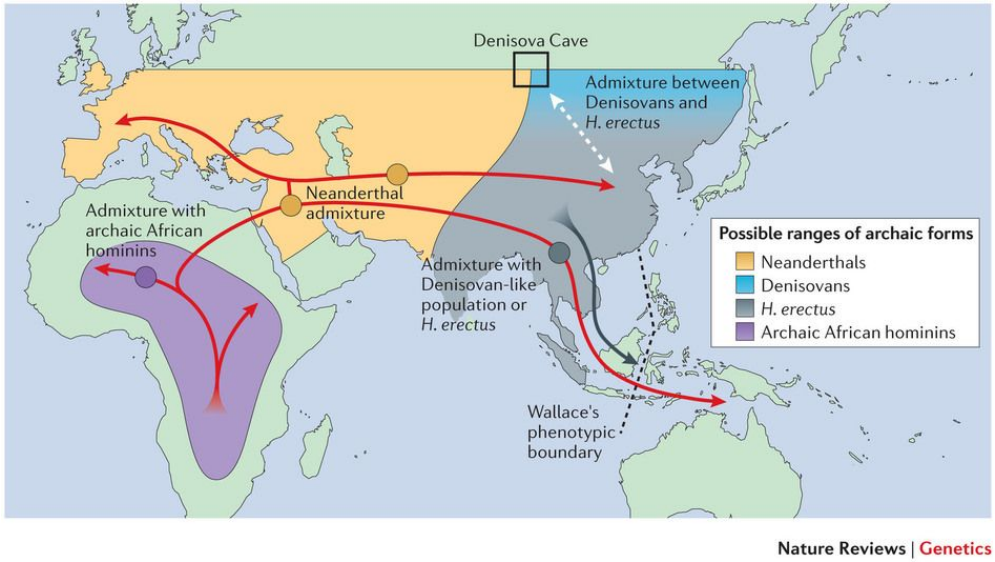
- evolutionary analysis of genome-scale data
  - resequencing data, genomic sequence data, homology data
  - population genomics, comparative genomics
- efficient computational analysis of genome-scale data sets
  - Linux command-line work, use of command-line programs
  - scripting with bash, awk, perl, python etc
  - understanding computational constraints
- theoretical basis of central concepts and methods
  - population genetics
  - coalescent theory
  - sequence evolution

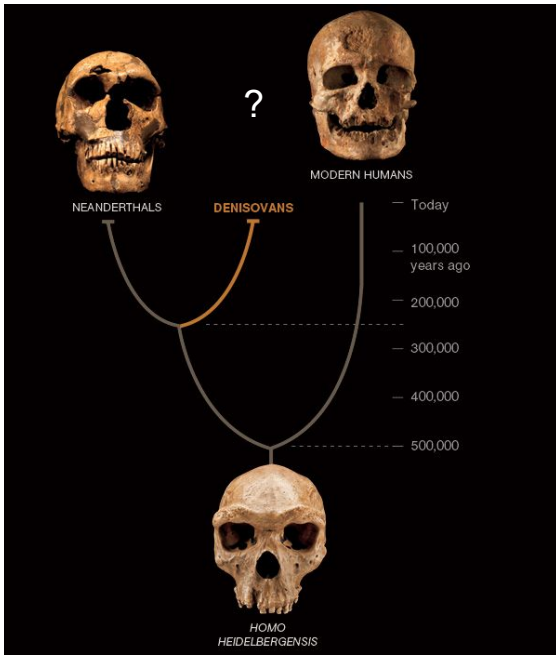
Why this course?

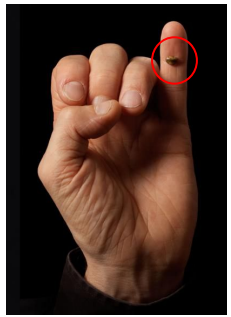
- DNA sequencing is getting cheaper and cheaper
  - important to understand the data and the population genomics theory behind it
- biology is getting computational, “genomics” is the forefront of this change
  - every biologist doesn’t need to script; those doing sequence analysis **have** to!
- U. Helsinki has several interesting genomics projects ongoing
  - lack of competent students, possibilities to get into high-class research
- Evolutionary genomics is a hot topic!
  - it touches us, humans, as a species
  - revolution in analysis methods, applied to human first
  - research on other organisms gains from this development

Why this course?

- recent insights about human evolution
  - human history, introgression







ngm.nationalgeographic.com

Why this course?

- recent insights about human evolution
  - adaptation through introgression

High altitudes in Tibet

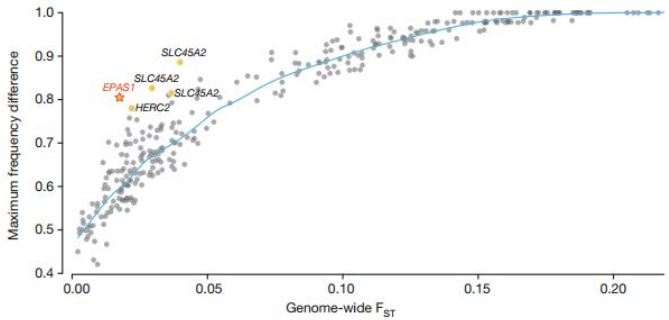


Figure 1 | Genome-wide F_{ST} versus maximal allele frequency difference.

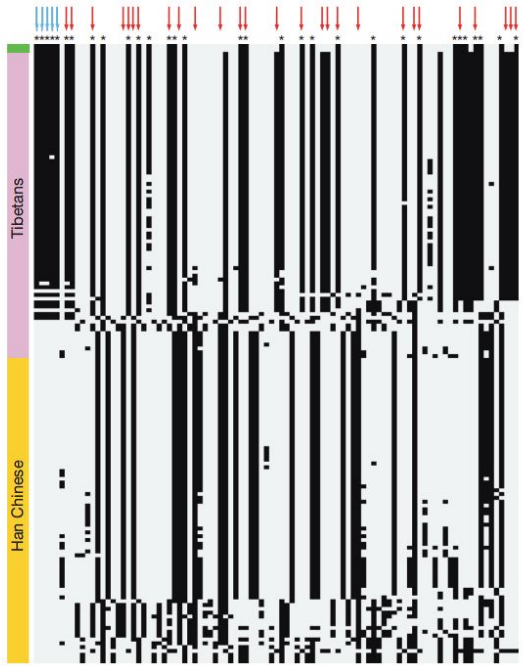
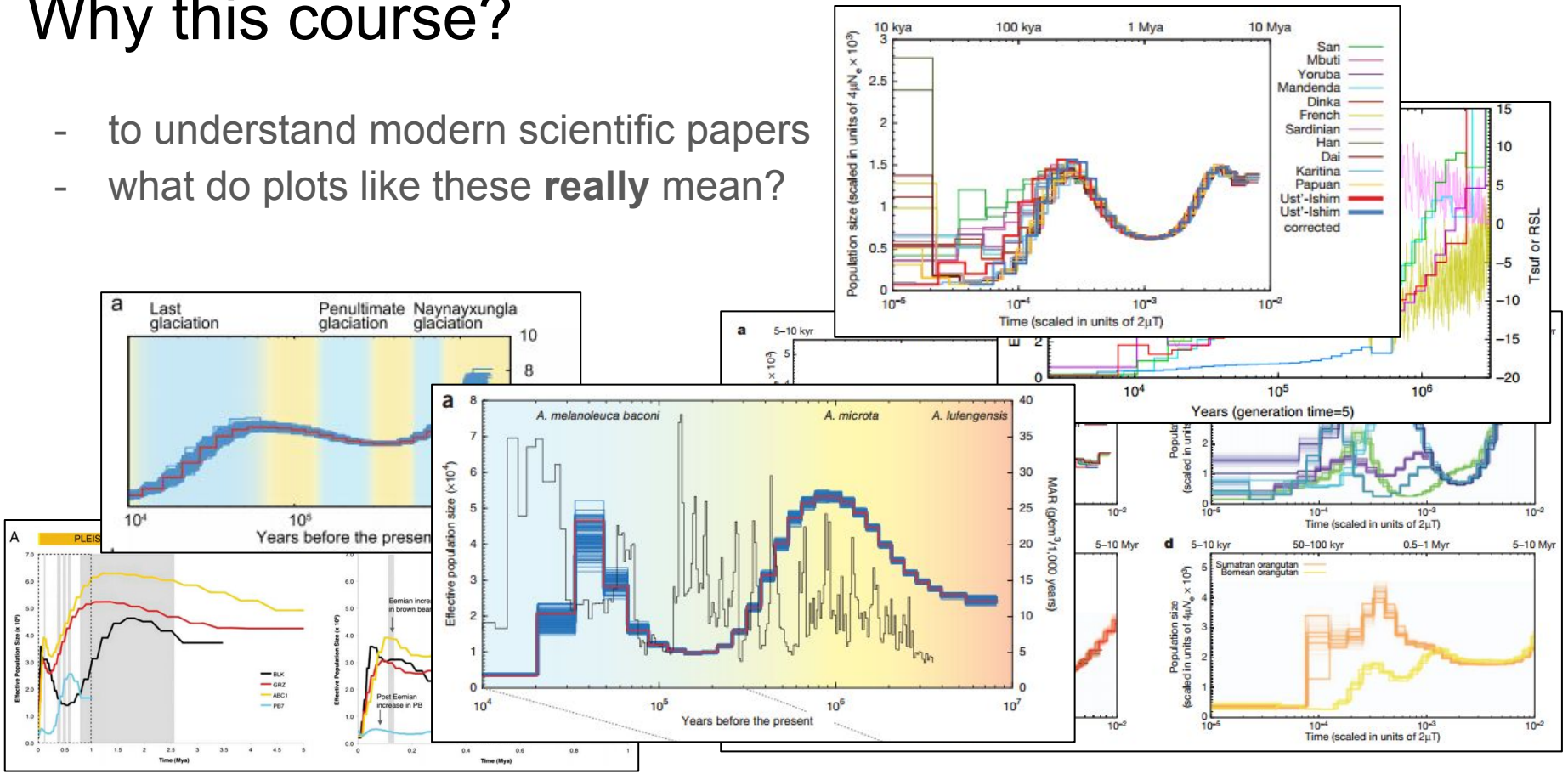


Figure 2 | Haplotype pattern in a region defined by SNPs that are at high frequency in Tibetans and at low frequency in Han Chinese. Each column i

Huerta-Sanchez et al, Nature, 2014

Why this course?

- to understand modern scientific papers
- what do plots like these **really** mean?



Why Linux?

No meaningful alternatives

- serious bioinformatic analysis on Windows is a bad joke
- Macs were Unix computers, now becoming iPads with a keyboard
- many software packages only work on Linux

If you want to do genomics data analysis, you have to learn Linux

- If you don't want, this course may not be for you

U. Helsinki doesn't provide Linux computers

- one hast to use Linux remotely or use own computers
- we use Linux Virtual Machine, compatible with most modern laptops

Course format

- Mondays, Tuesdays and Thursdays at 14-16, break around Eastern
  - see WebOodi / Moodle
- no separation between lectures and computer exercises
  - computer exercises with own laptops
  - (possibly) using CSC: apply for an account if you don’t have one
  - many/most analyses can be done with Linux Virtual Machine (VM)
  - VM uses VirtualBox, runs on Windows, Mac, Linux
  - most software pre-installed on VM
- home exercises with brief reports
  - extra points for the final exam
  - Brownie points for my own records
  - *opportunity to practise what is learned, to become fluent with the methods used*
- final exam on-line: data analysis using methods learned, brief essays

Course data

Mostly using nine-spined stickleback data from Juha Merilä

- 500 Mbp genome, we use 10 Mbp subset
- good-quality 10X/5X for non-model organism
- lots of samples, many populations
- closely-related to three-spined stickleback, a model organism

Using Google and Wikipedia is allowed!

I try to explain logic behind concepts and connections between them

- I won’t spend time explaining things that you can easily find in the net

First home exercise:

- what is “A” in awk, “B” in bash and “R” in R ?