CSC 334: TOPICS IN COMPUTATIONAL BIOLOGY

"Algorithms for Genomic Data"

Fall 2015

Smith College

Instructor: Prof. Sara Sheehan

Outline: 10/19

- Review Sankoff’s Algorithm
- Midterm and Project discussion
- HW 5
- Next half of the semester overview

Talk Today!

- Colleen Lewis from Harvey Mudd College
- “*Increasing Diversity in Computer Science*”
- 4:10pm in Ford 241 (here)

Midterm and Final Project

- Midterm:
  - Take-home, mix of programming and on paper
  - Out around Wednesday Oct. 28
  - Covers everything up through Sankoff's Algorithm
  - We will do some in class review

Midterm and Final Project

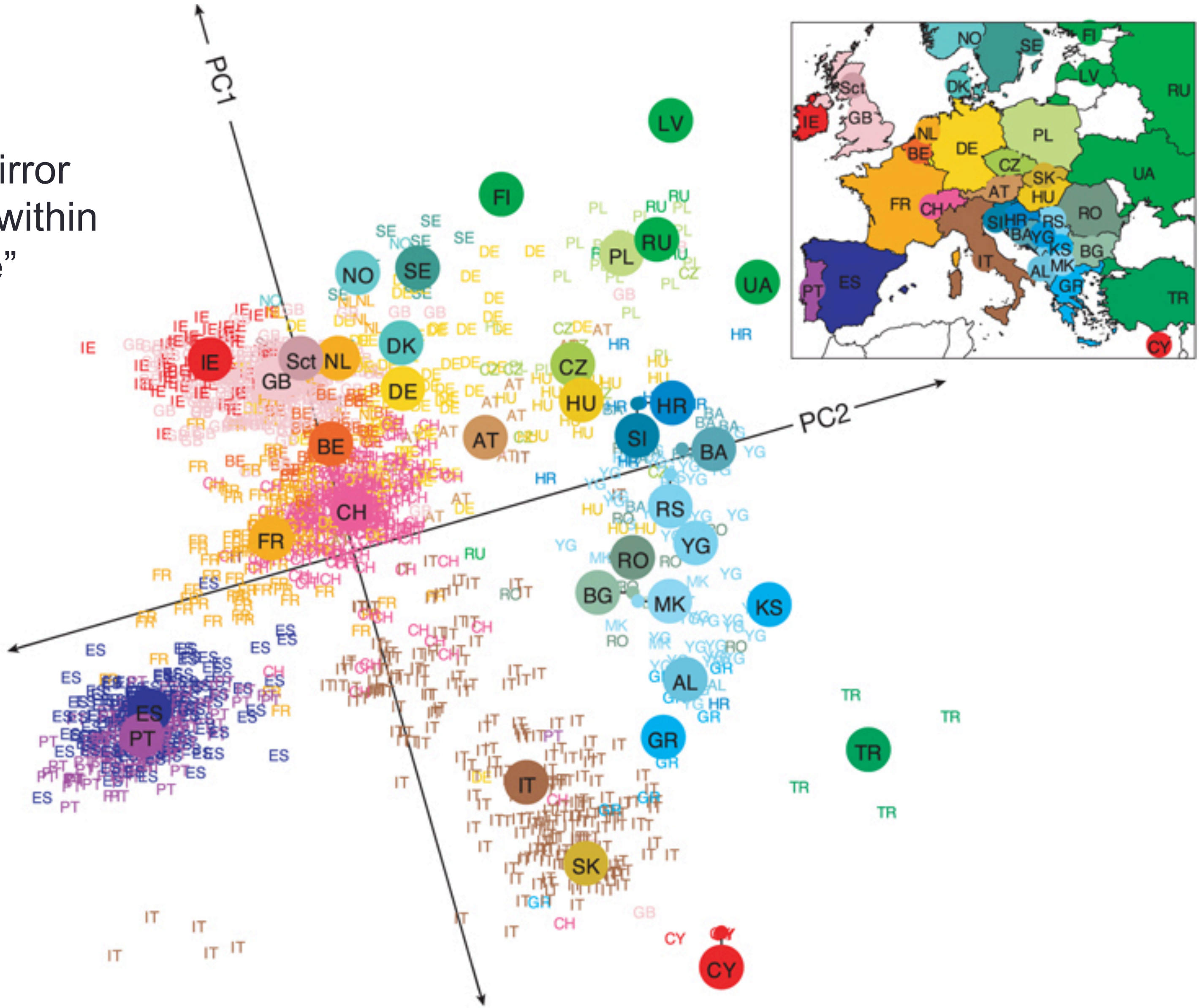
- Midterm:
  - Take-home, mix of programming and on paper
  - Out around Wednesday Oct. 28
  - Covers everything up through Sankoff’s Algorithm
  - We will do some in class review

- Final Project:
  - HW 5 is a small paper presentation
  - Designed to help you find a project topic
  - List of papers up soon, or you can find your own
  - Involve algorithms somehow
  - DNA, RNA, or protein

Population history using genomics

- “Genes mirror geography within Europe”
- Application of PCA (Principal Components Analysis) to genomics
  - Classical population genetics, could be applied to any dataset

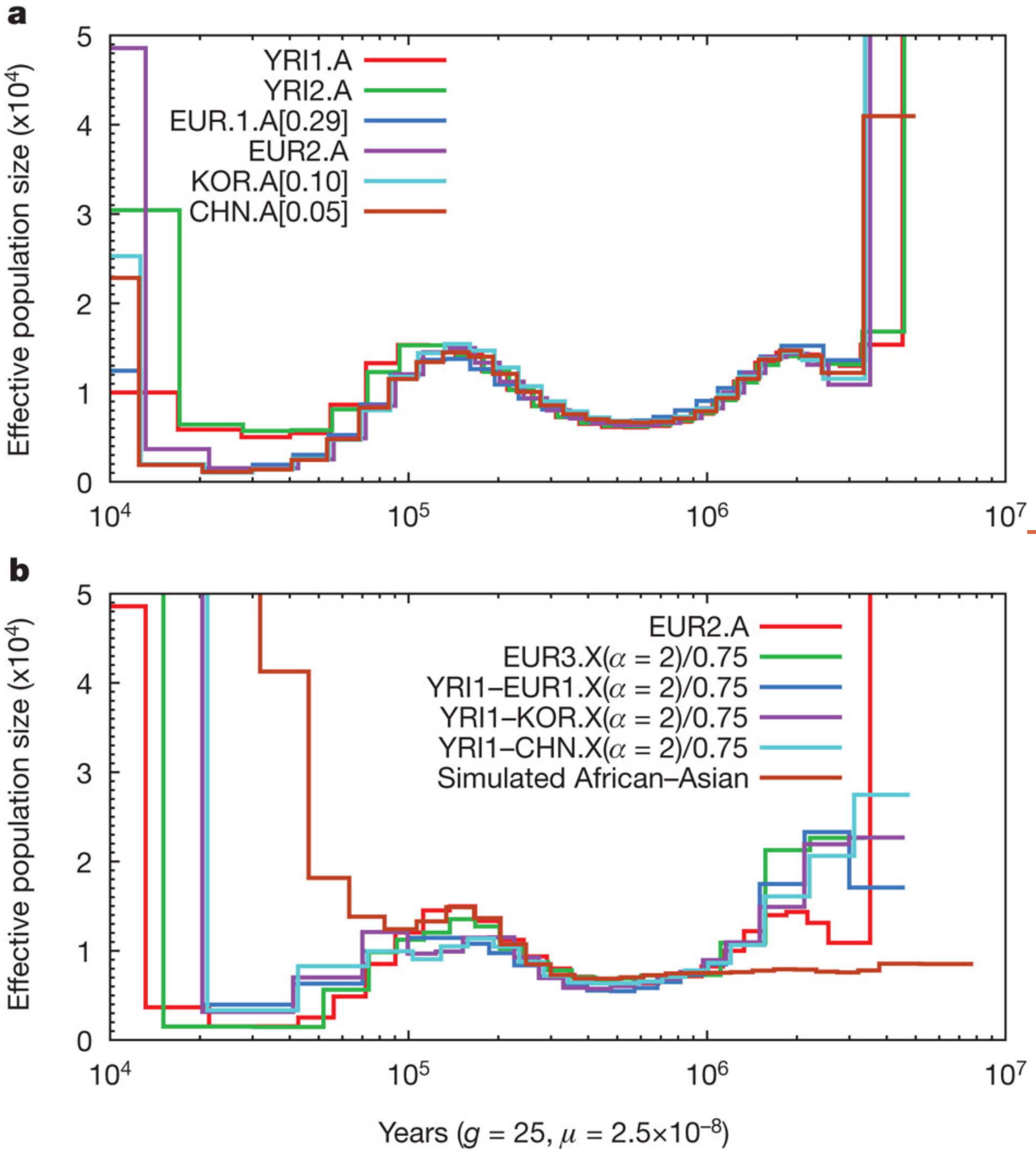
"Genes mirror geography within Europe"



Population size estimation

- “Inference of human population history from individual whole-genome sequences”

PSMC estimate on real data.



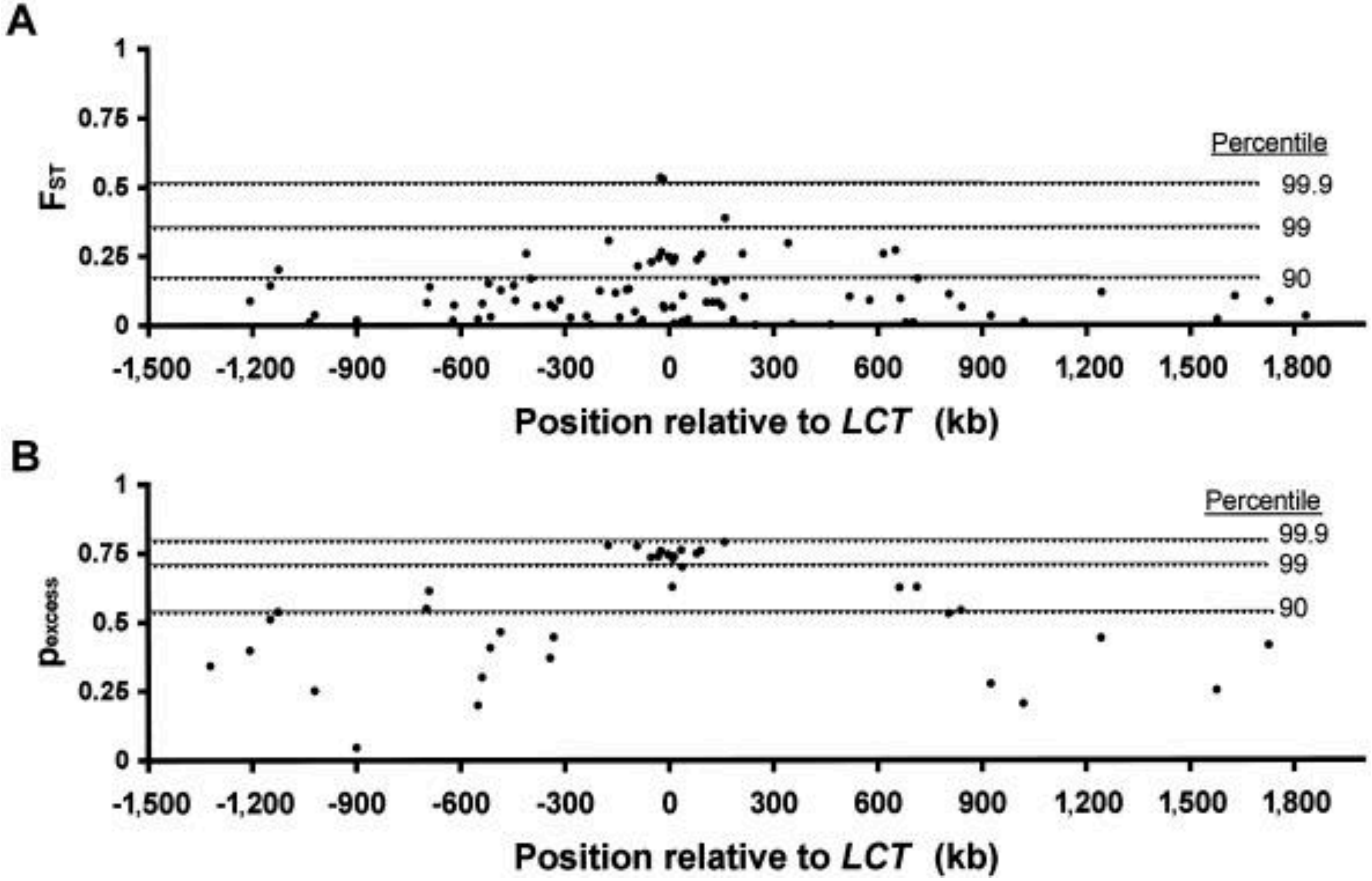
H Li & R Durbin *Nature* **000**, 1-4 (2011) doi:10.1038/nature10231

nature

Algorithms for Natural Selection

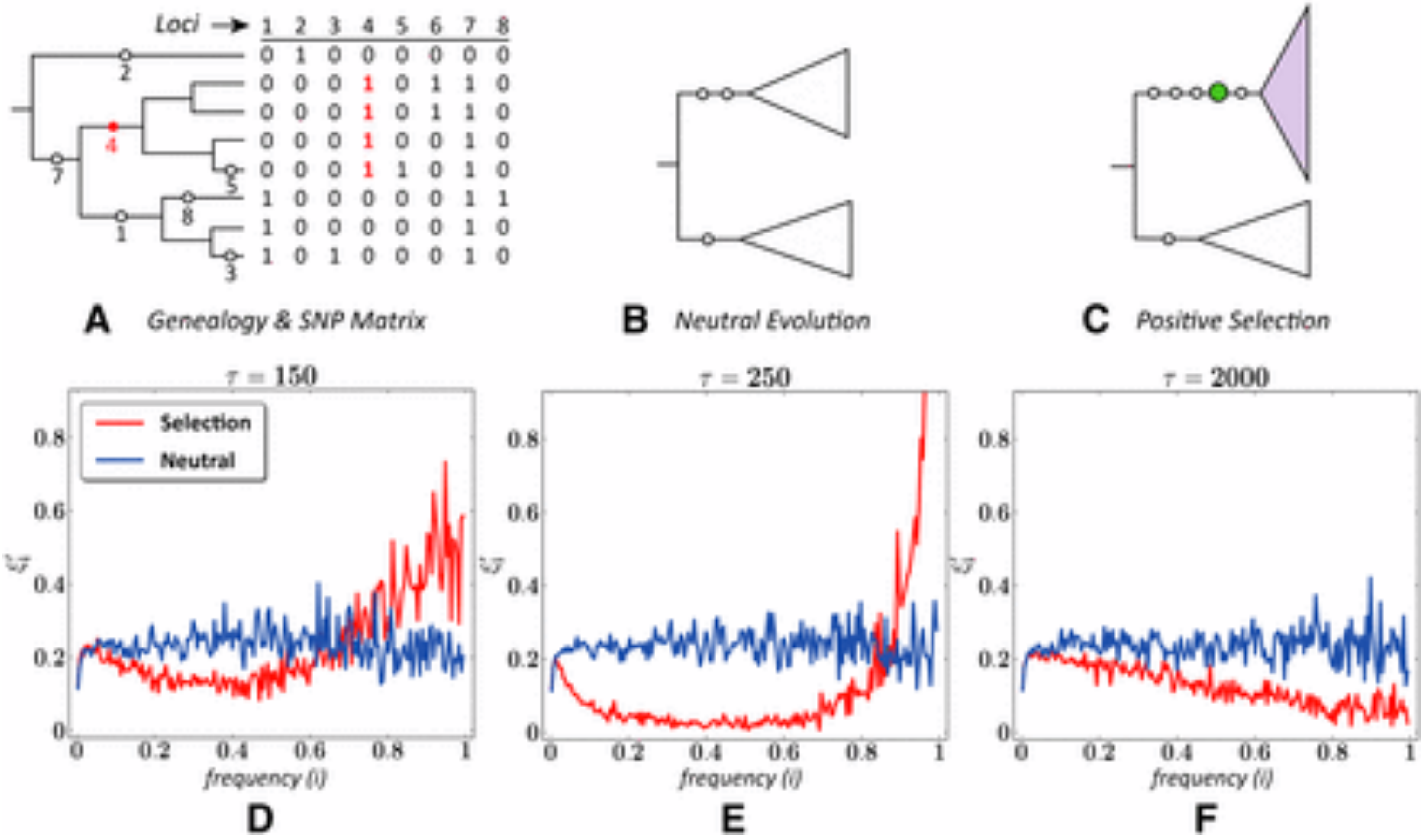
- “Genetic Signatures of Strong Recent Positive Selection at the Lactase Gene”
- “Learning Natural Selection from the Site Frequency Spectrum”

Positive selection on the lactase gene



“Genetic Signatures of Strong Recent Positive Selection at the Lactase Gene”

Machine learning for selection



Loci →	1	2	3	4	5	6	7	8
	0	1	0	0	0	0	0	0
	0	0	0	1	0	1	1	0
	0	0	0	1	0	1	1	0
	0	0	0	1	0	0	1	0
	0	0	0	1	1	0	1	0
	1	0	0	0	0	0	1	1
	1	0	0	0	0	0	1	0
	1	0	1	0	0	0	1	0

A *Genealogy & SNP Matrix*

B *Neutral Evolution*

C *Positive Selection*

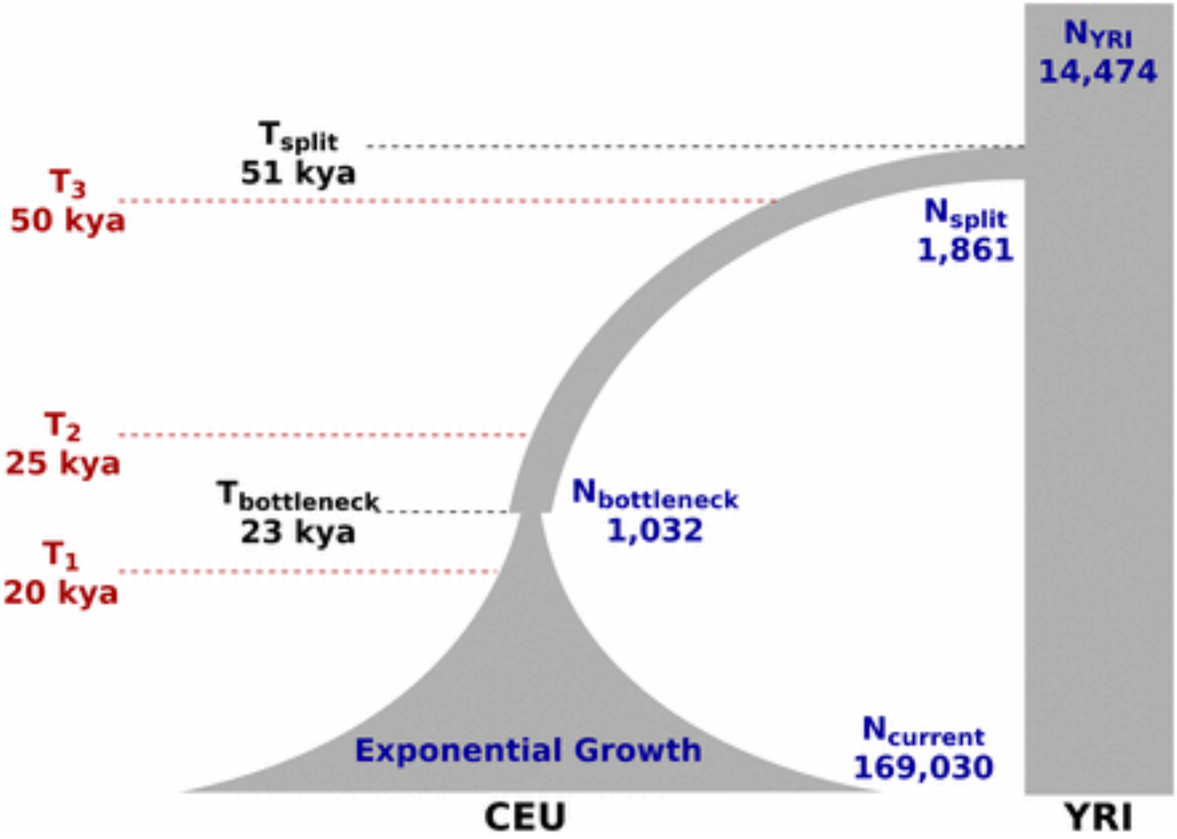
D

E

F

“Learning Natural Selection from the Site Frequency Spectrum”

Demographic Models

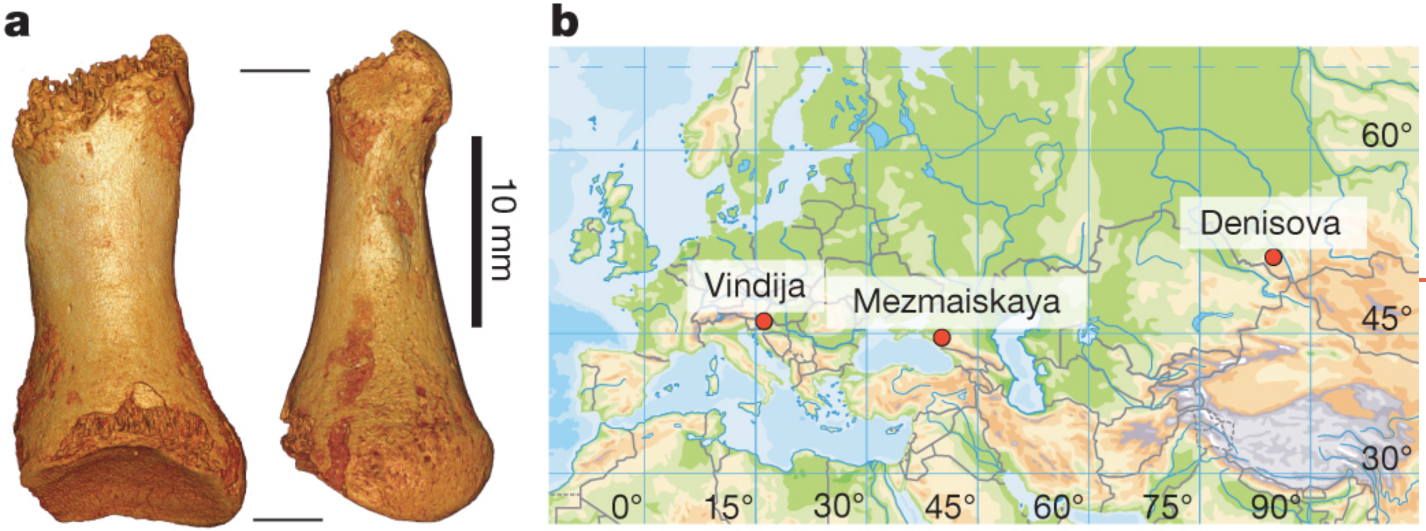


"Learning Natural Selection from the Site Frequency Spectrum"

Ancient Genomics

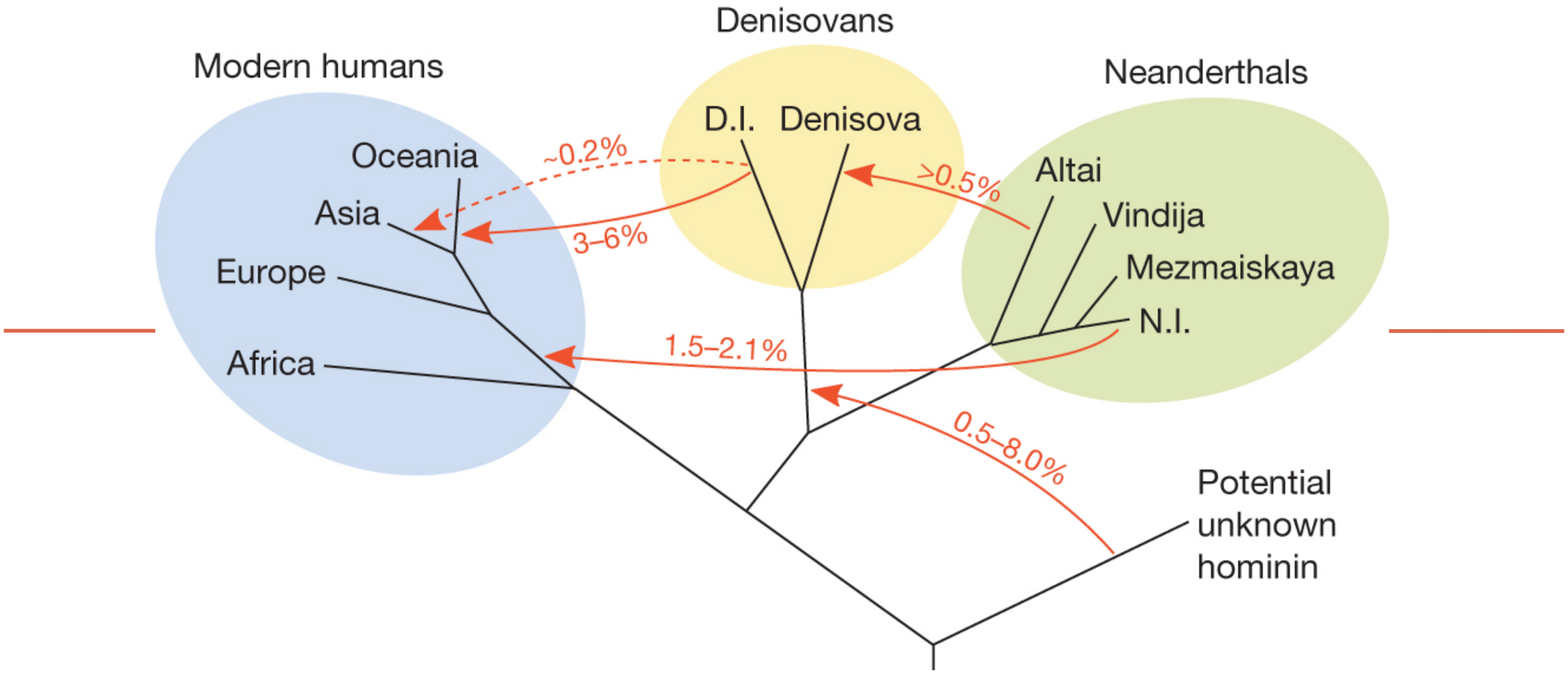
- “The complete genome sequence of a Neanderthal from the Altai Mountains”

Toe phalanx and location of Neanderthal samples for which genome-wide data are available.



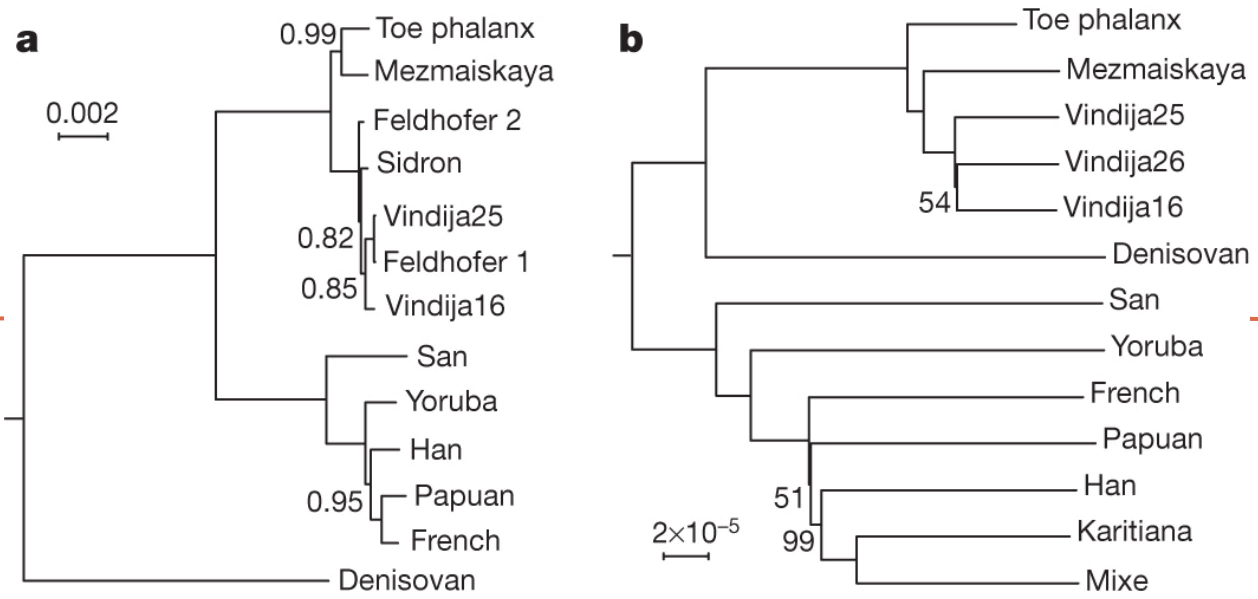
K Prüfer *et al. Nature* **000**, 1-7 (2013) doi:10.1038/nature12886

A possible model of gene flow events in the Late Pleistocene.



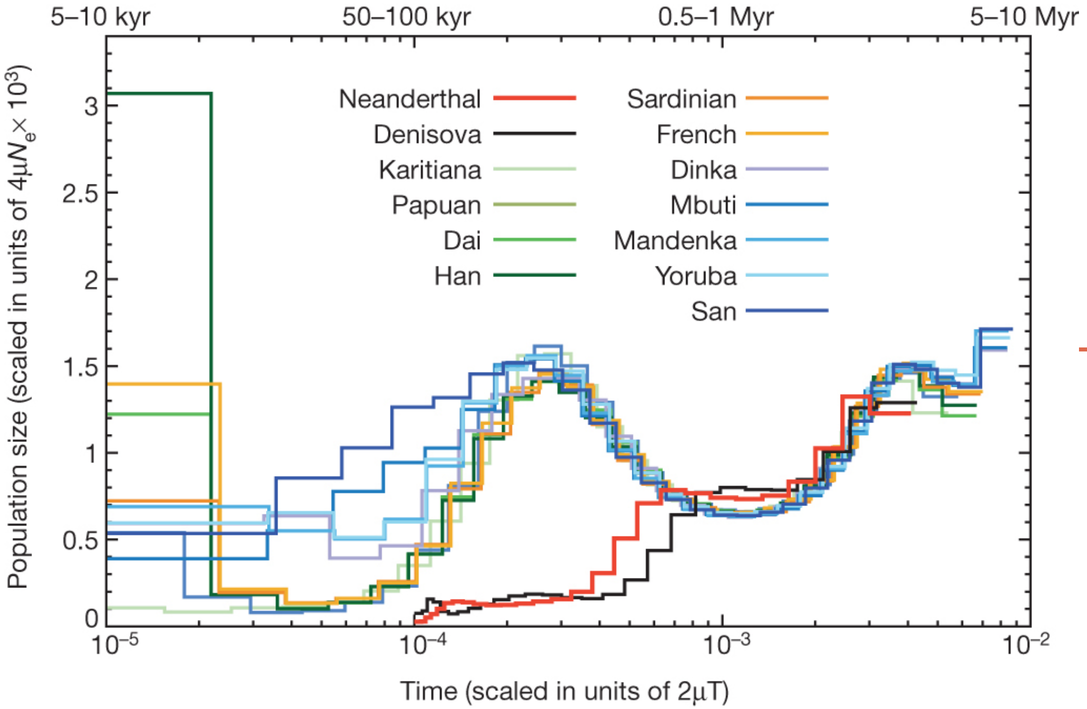
K Prüfer *et al. Nature* **000**, 1-7 (2013) doi:10.1038/nature12886

Phylogenetic relationships of the Altai Neanderthal.



K Prüfer *et al. Nature* **000**, 1-7 (2013) doi:10.1038/nature12886

Inference of population size change over time.



K Prüfer *et al. Nature* **000**, 1-7 (2013) doi:10.1038/nature12886

Cancer Genomics

- “Next generation sequencing in cancer research and clinical application”

“Next generation sequencing in cancer research and clinical application”

Alignment

Mutation calling

Structural Variation (SV) detection

Functional effect of mutation

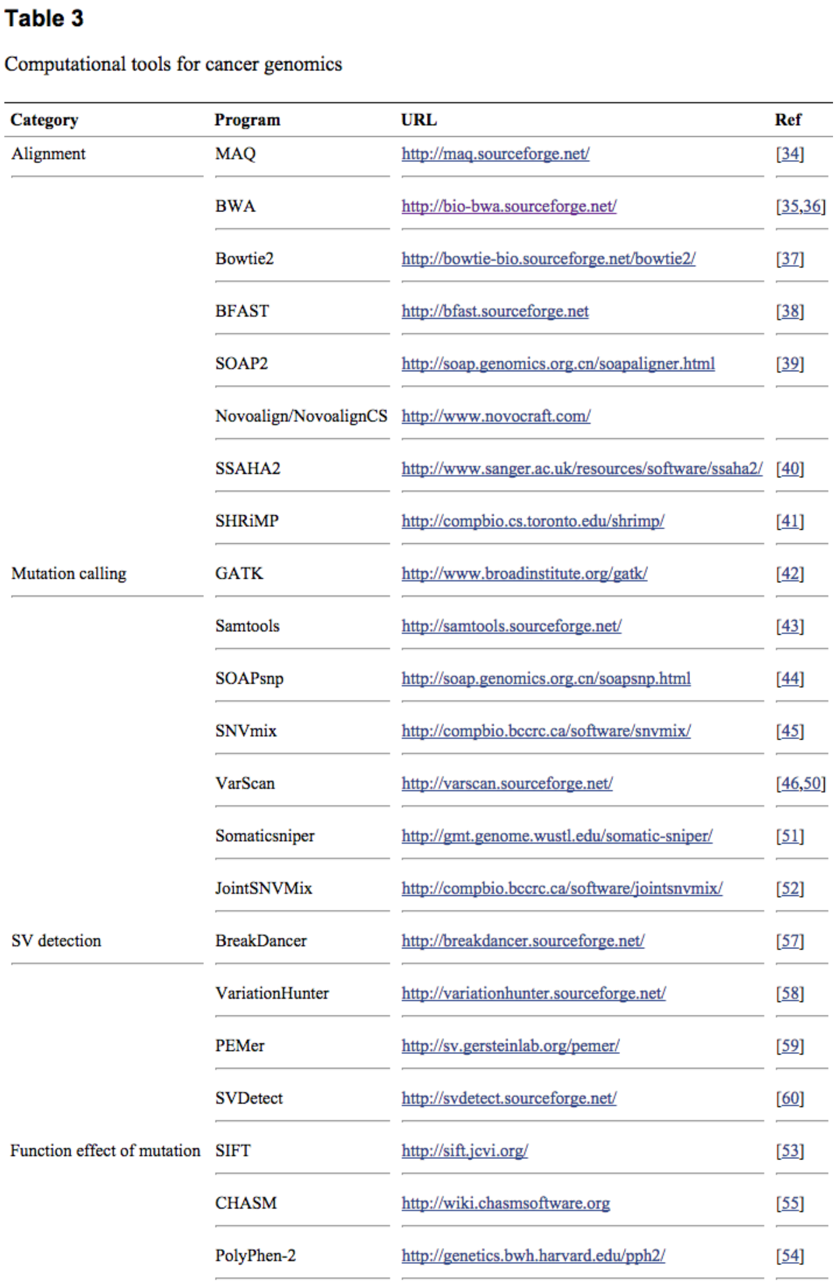


Table 3

Computational tools for cancer genomics

Category	Program	URL	Ref
Alignment	MAQ	http://maq.sourceforge.net/	[34]
	BWA	http://bio-bwa.sourceforge.net/	[35,36]
	Bowtie2	http://bowtie-bio.sourceforge.net/bowtie2/	[37]
	BFAST	http://bfast.sourceforge.net	[38]
	SOAP2	http://soap.genomics.org.cn/soapaligner.html	[39]
	Novoalign/NovoalignCS	http://www.novocraft.com/	
	SSAHA2	http://www.sanger.ac.uk/resources/software/ssaha2/	[40]
	SHRiMP	http://compbio.cs.toronto.edu/shrimp/	[41]
Mutation calling	GATK	http://www.broadinstitute.org/gatk/	[42]
	Samtools	http://samtools.sourceforge.net/	[43]
	SOAPsnp	http://soap.genomics.org.cn/soapsnp.html	[44]
	SNVmix	http://compbio.bccrc.ca/software/snvmix/	[45]
	VarScan	http://varscan.sourceforge.net/	[46,50]
	Somaticsniper	http://gmt.genome.wustl.edu/somatic-sniper/	[51]
	JointSNVMix	http://compbio.bccrc.ca/software/jointsnvmix/	[52]
SV detection	BreakDancer	http://breakdancer.sourceforge.net/	[57]
	VariationHunter	http://variationhunter.sourceforge.net/	[58]
	PEMer	http://sv.gersteinlab.org/pemer/	[59]
	SVDetect	http://svdetect.sourceforge.net/	[60]
Function effect of mutation	SIFT	http://sift.jcvi.org/	[53]
	CHASM	http://wiki.chasmsoftware.org	[55]
	PolyPhen-2	http://genetics.bwh.harvard.edu/pph2/	[54]

Other topics

- RNA secondary structure (very nice dynamic programming algorithm!)
- GWAS: genome-wide association studies
    - Human disease applications
- Pre-natal diagnostics
    - Uses string alignment, next-generation sequencing, etc
- Literature review?
    - Come talk to me