Other Types of Sequence Analyses

James Pease
Indiana University
14 July 2014

Heterozygosity

"π"

```
TATCGCGAAGGCTAGAACACAAGGTAAAATATCGCGA
TATCGCGAAGGCTAGAACGCAAGGTAAAATATCGCGA
TATCGCGAAGACTAGAACACGAGGCAAAATATCGCGA
TATCGCGAAGACTAGAACACGAGGCAAAATATCGCGA
TATCGCGAAGACTAGAACACGAGGCAAAATATCGCGA
```

Heterozygosity

```
TATCGCGAAGGCTAGAACACAAGGTAAAATATCGCGA
TATCGCGAAGGCTAGAACGCAAGGTAAAATATCGCGA
TATCGCGAAGGCTAGAACACAAGGTAAAATA-CGCGA
TATCGCGAAGACTAGAACACGAGGCAAAATA-CGCGA
TATCGCGAAGACTAGAACACGAGGCAAAATAACGCGA
TATCGCGAAGACTAGAACACGAGGCAAAATATCGCGA
```

$$\pi = \text{mean}\left\{ \frac{3}{6} \ldots \quad \frac{1}{6} \ldots \quad \frac{1}{4} \right\}$$

Heterozygosity

	Heterozygosity is high	Heterozygosity is low
Genetic diversity	higher	lower
population size	larger	smaller
inbreeding	less	more
possible ancestral population size…	expansion	reduction “bottleneck”

Population Structure

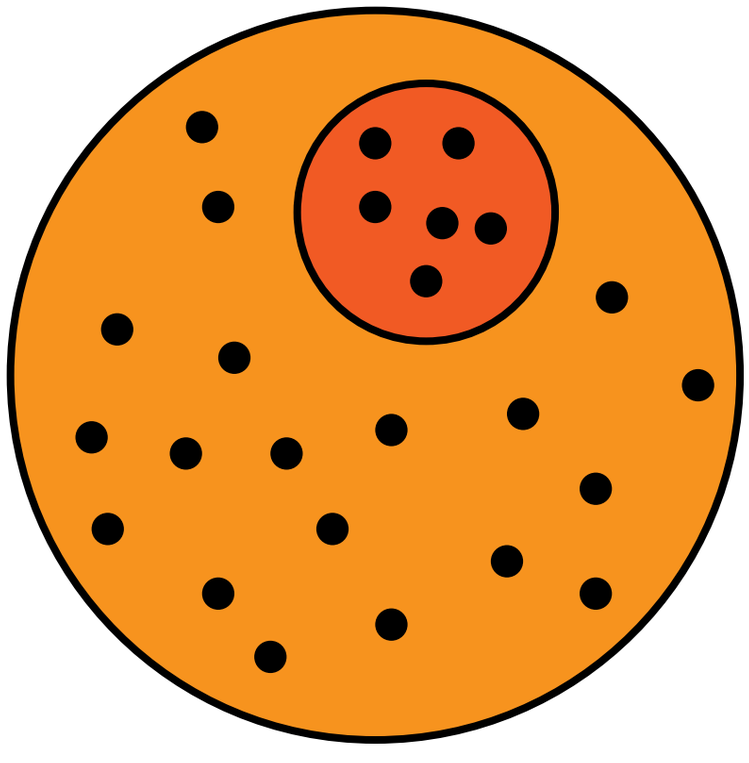
Is there genetic structure in the population

$$F_{ST} = \frac{\pi_{\text{total}} - \pi_{\text{subpopulation}}}{\pi_{\text{total}}}$$

How different is the proportion of alleles in each subpopulation compared to the population as a whole?

Population Structure

If heterozygosity in the subpopulation is higher than in the total population, then there is evidence of non-random mating that creates population structure.



Protein analysis

Now you have the genome sequence...

GCCCCTTTCGCTGCAGTTGACGGGGGATGTGCGTGATTACTCACGTCTAGCCAGGTAAATAGTCCAAGATCTTCGGTCCGTACTTATTTCGAAAACACCG
AGTCACTAATCAGCGGACAAGTCTGTGGCCTTAACGGCTGGGTTTAACCAAAACACGAGCTGACCGCGTGCAGTGCCCCAAACGAGCATCCTATAACGAT
CAACAAGGAGGGAGGCGGTCTGCCCTTCGGTGACCGCCGTCTTGCCTAGCGAAGGCGGATGTACAATGCTGGTCTTCTACTGGGGGACATCTTTCTAATA
TTTTACAATCATGCGCATAATCCCGTAGTACCACGTCGAGTTAATGTTGTACGCCCAGAAGTCTCACACTCTCATGTTGGTGAACACCATTCTCCACTCC
TTCTACTACTTACTCAAGCACAGGCGTTCGGTACGAATTGGGATACGCAGCTCTTACTTCGTTACATAACACGTATGCGCCCCTCAAGTGAAAGCCGGTG
AACCTATGCATTGGCTATACCATGCCTAAGTACCTGTGCTTCCGCGACCACCATAAATATTCGTTCGGTGGGGCGTCTTGGAATGAGGGCATGTACTGTT
ACGGTAGCAGAAATAGTTTTGGACTTCACCGCGGGAATATTATTTCTCCTTACCCTGACTAGAATCACGTTACGAAACTTCCCATGTCTCTTAAATGTAT
AAACTTGACGAATAACAGGAAGTCACGTACTGCACCTGTCTCTTTGGTGCACAAAGCACCATAACTTCAGCATACGCAAGATCTAAACCCCGCGGGCGGT
ACGCAAAGTTTTACCGAAAGTGCAGTCCTGGGCGTCCTTCGTCCTACGCCTTATCTCAGTTCAACTTGGGTCGGACCCTTTTGTGAAGGATGAAATGCCA
GTTCCCCGGTCAGACCGGTGGATCGTACTGCAGTTTCTTGCGGGTTGTGCTATGTTGCTCACGAGCTTCCCCTGTTGTCACAAATAGGATGGTATATGAC
ATCAAGTCTTAATATTGAAGGTCGACAGTCTATCCCACCGTTCTAGACCTAATACCCCCCGCGCGCACTACGAATCTTTACGTGCCACCTGGCATTTCGG
TACATTAAAAGTATTGTTCGGGCCTCTTGCGCGTGTACATTTGTTACCTTTGCTGACCGTCACCCAGCTAGTTACGAACGGATATCTGGCGCCACAGTAC
AACTGACTGAAGAATCACTCCCAGTAATAATAAAAGACGCTGGAAACACTCACAAAGACCGAGTTTTAAATCTTGATCGAGATCTGTCAGACAAGATACC
AGATGGAGATAGGGGACCATGTCGAAGGCTGTGCGAGGGGGAAGCTTATGTAAACACAGACCTGGCACGGCAACTATCGCCGGTAAAGTGCTGGTAGCTT
TGGGCGTCCAGATTATGTATTACCTCACTATGAGCACGGACGCCCGACAAATCGCAGACGGATCACGAATCAATCCGTTGGCTTCGGGGACATTAATGCG

So where are the genes?

Software for computational gene prediction in the genome:

GLIMMER
GENSCAN
SNAP

Protein analysis

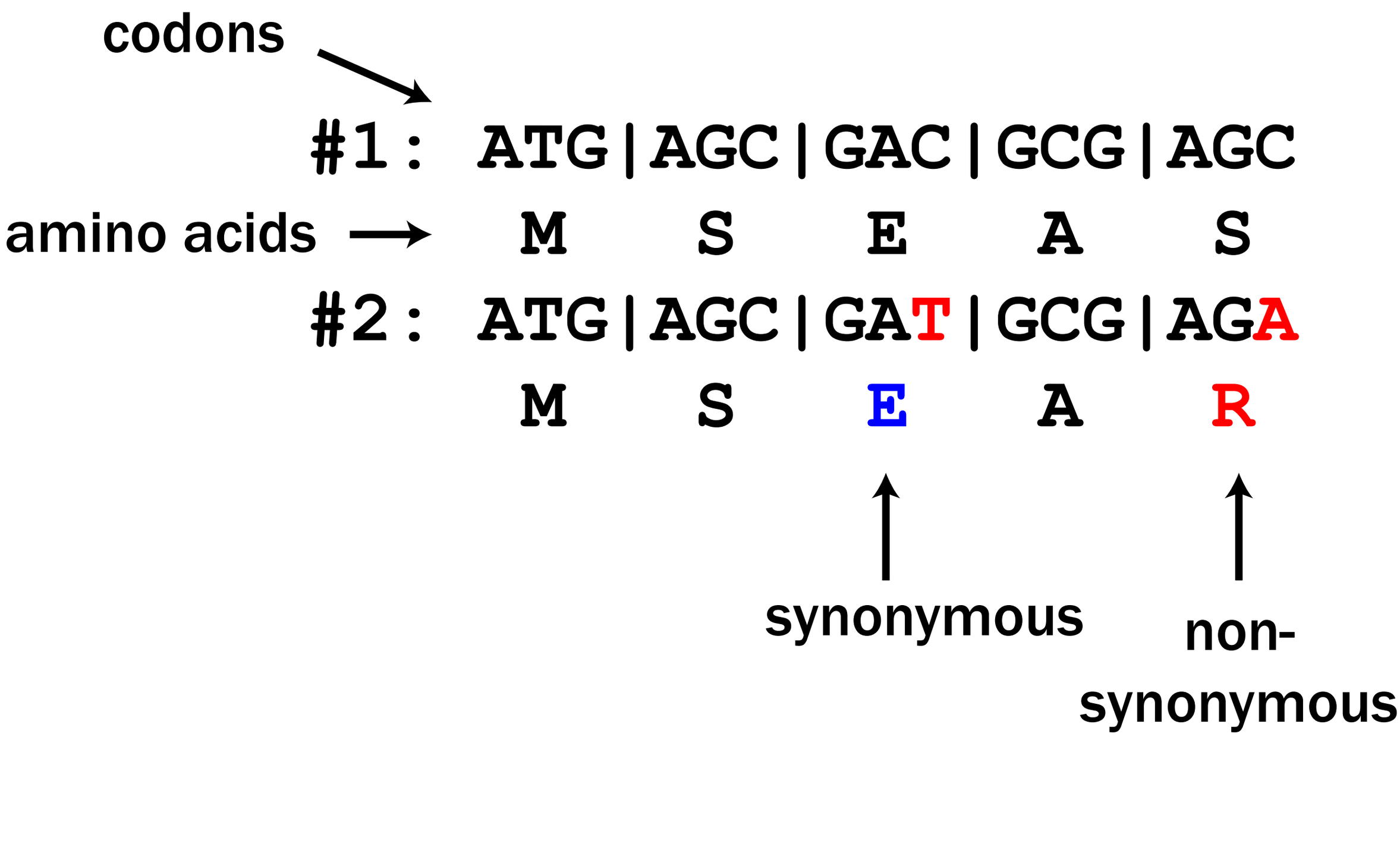
for the transcriptome:

“TransDecoder”

(comes with Trinity)

http://trinityrnaseq.sourceforge.net/

Analysis of Natural Selection



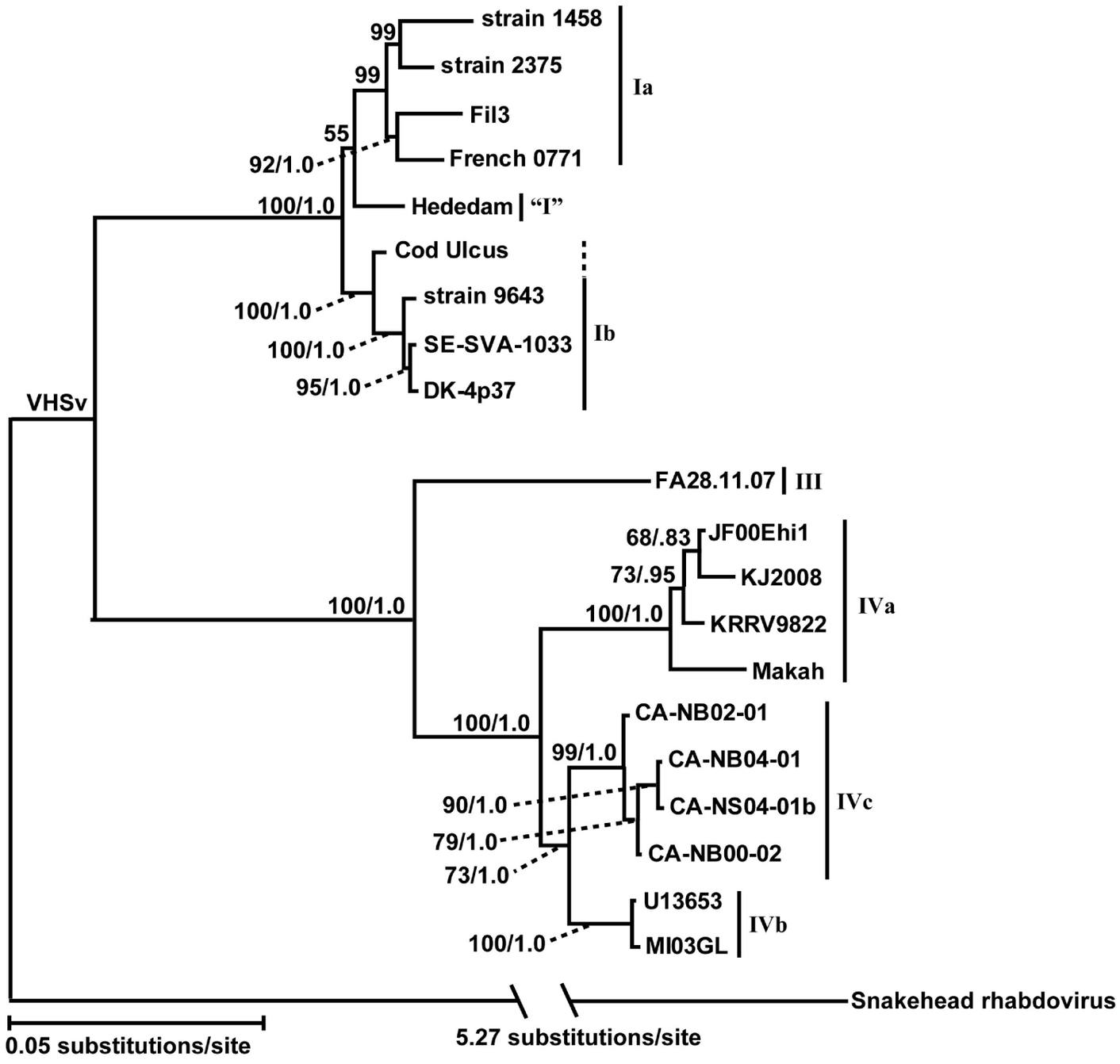
Analysis of Natural Selection

$$dN/dS = \frac{\#\ \text{non-synonymous}}{\#\ \text{synonymous}}$$

dN/dS > 1: Positive selection (adaptation)

dN/dS < 1: Negative selection (conservation)

E


strain 1458
strain 2375
Fil3
French 0771
Hededam
"I"
Ia
Cod Ulcus
strain 9643
SE-SVA-1033
DK-4p37
Ib
99
99
55
92/1.0
100/1.0
100/1.0
100/1.0
95/1.0
VHSv
FA28.11.07
III
68/.83
JF00Ehi1
73/.95
KJ2008
100/1.0
KRRV9822
Makah
IVa
100/1.0
CA-NB02-01
99/1.0
CA-NB04-01
CA-NS04-01b
CA-NB00-02
IVc
90/1.0
79/1.0
73/1.0
U13653
MI03GL
IVb
100/1.0
Snakehead rhabdovirus
0.05 substitutions/site
5.27 substitutions/site

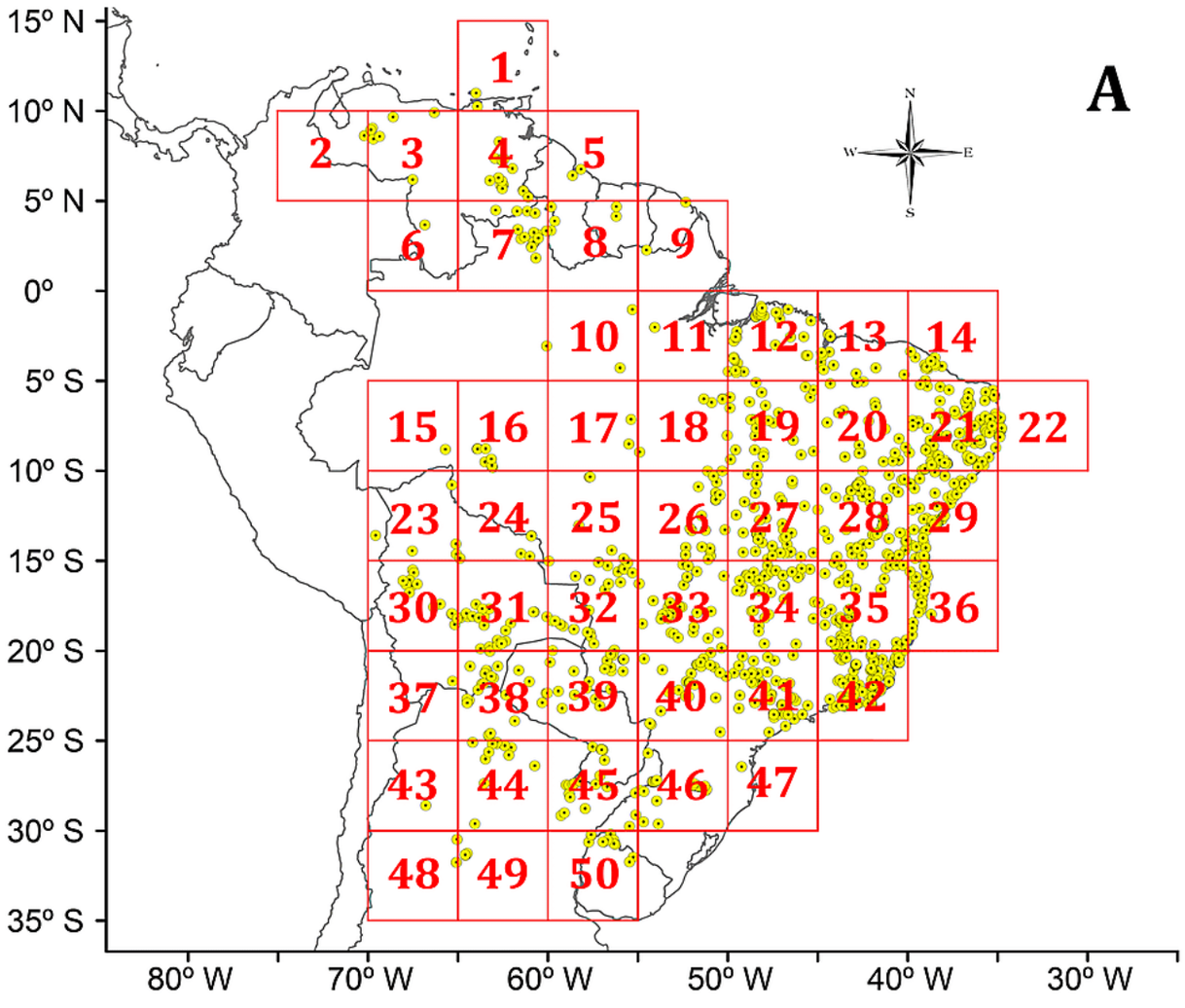
Analysis of Natural Selection

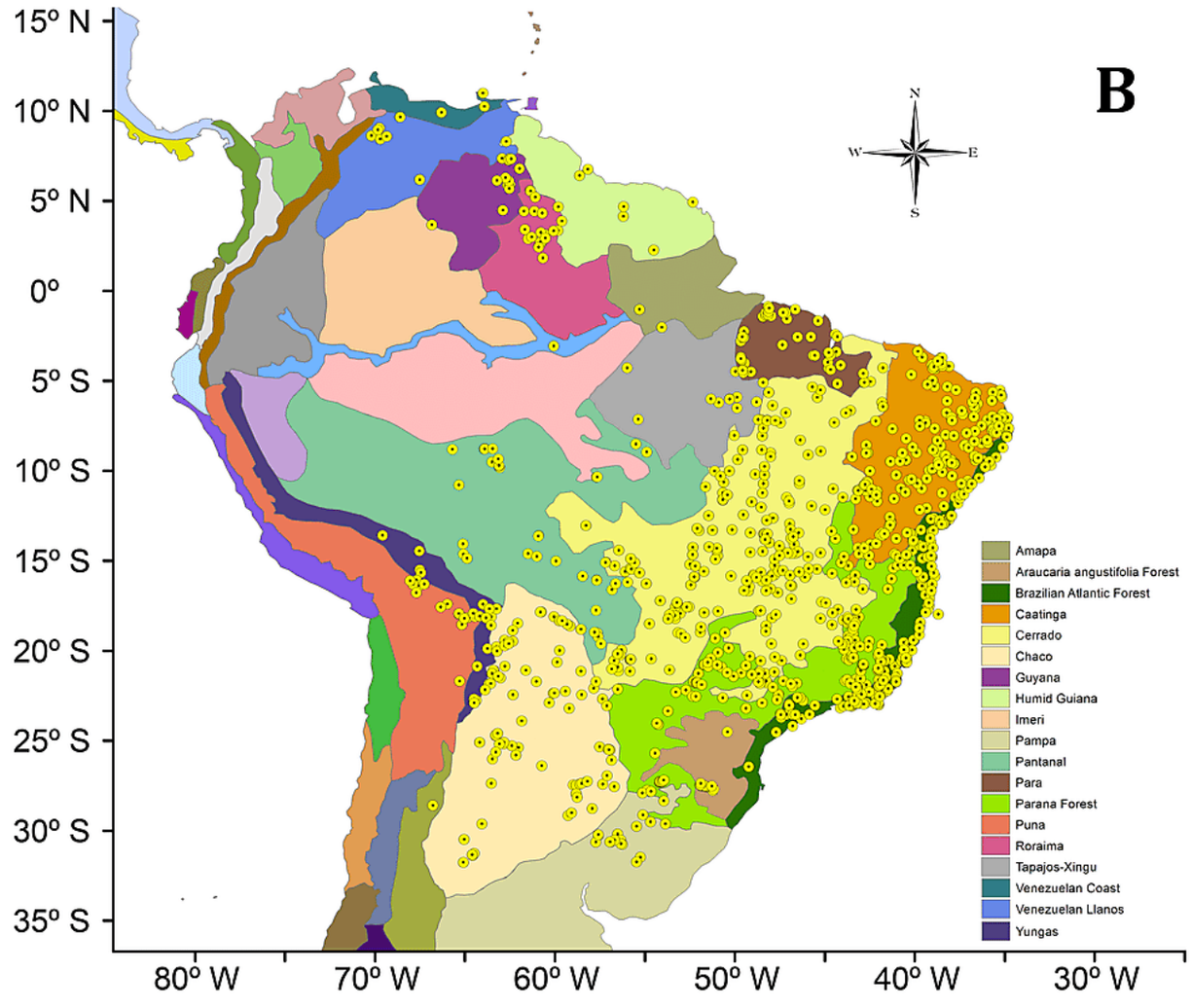
Software: PAML

Phylogenetic Analysis by Maximum Likelihood

http://abacus.gene.ucl.ac.uk/software/paml.html

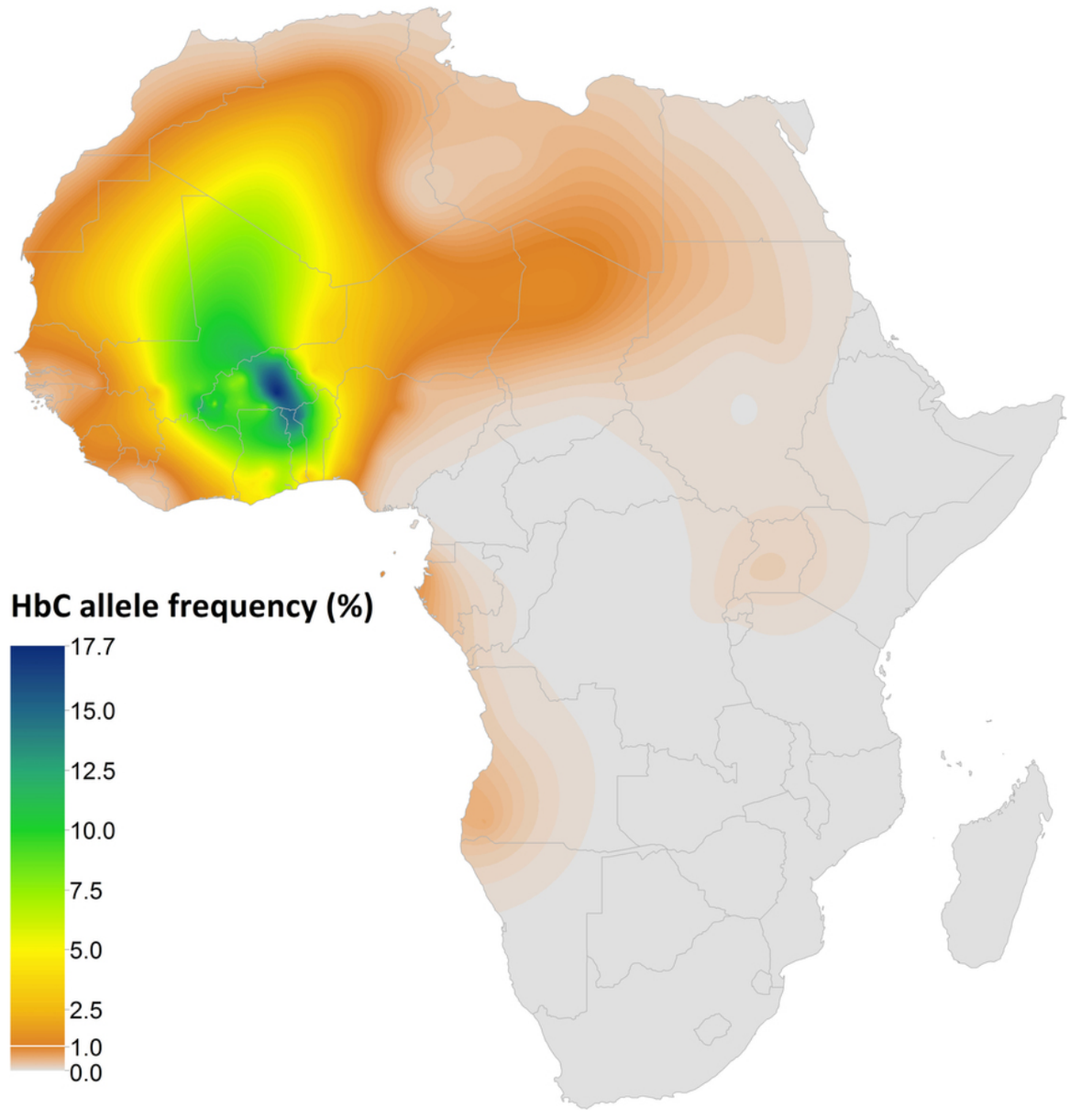
Biogeography





Biogeography

correlate allele frequencies with GPS coordinates



Biogeography

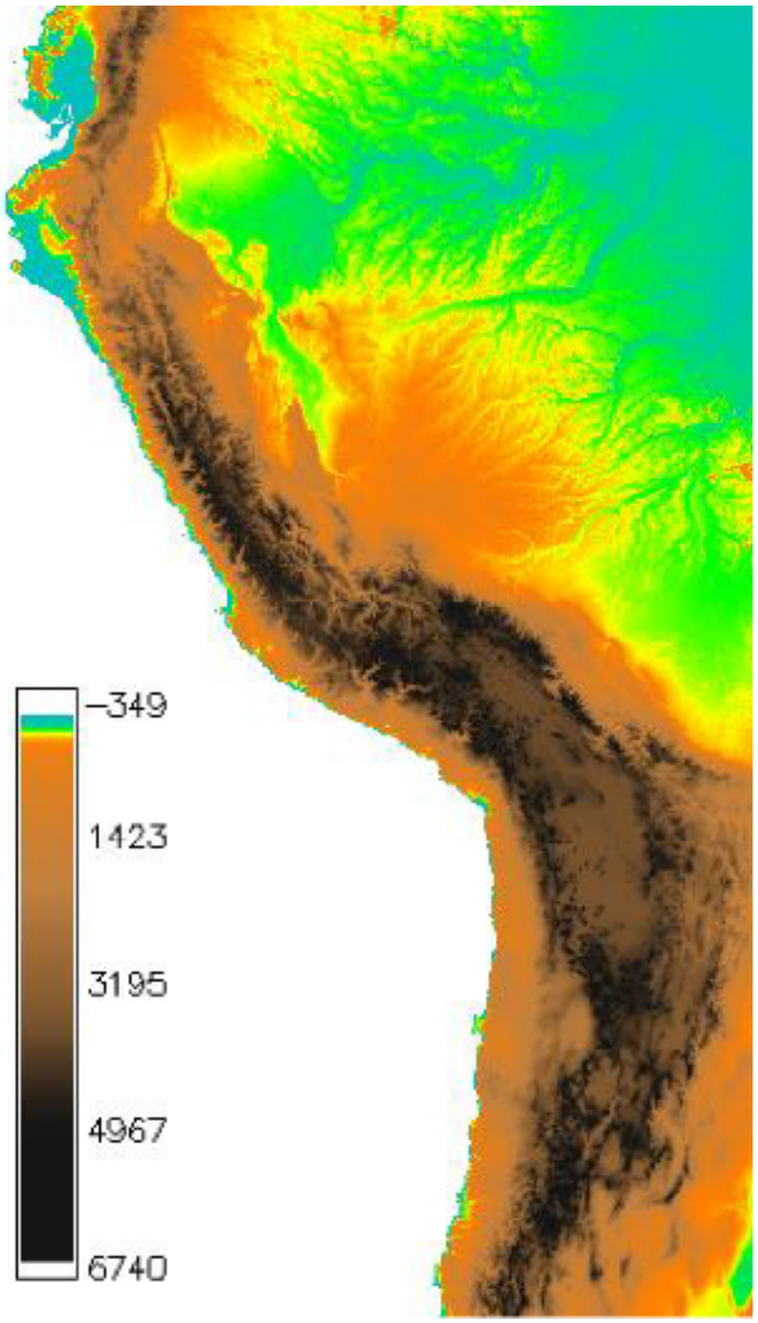
Correlation of allele frequency or SNPs with

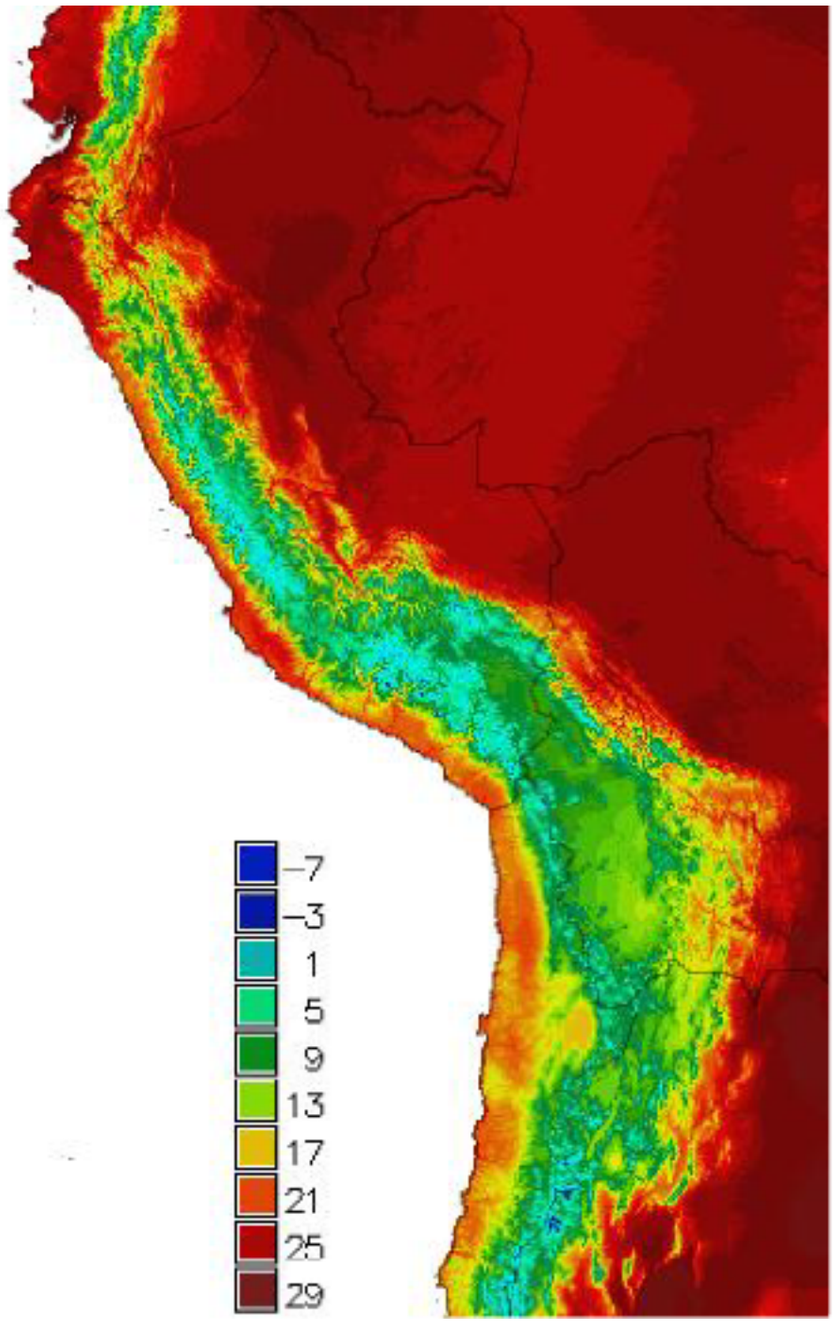
- Altitude

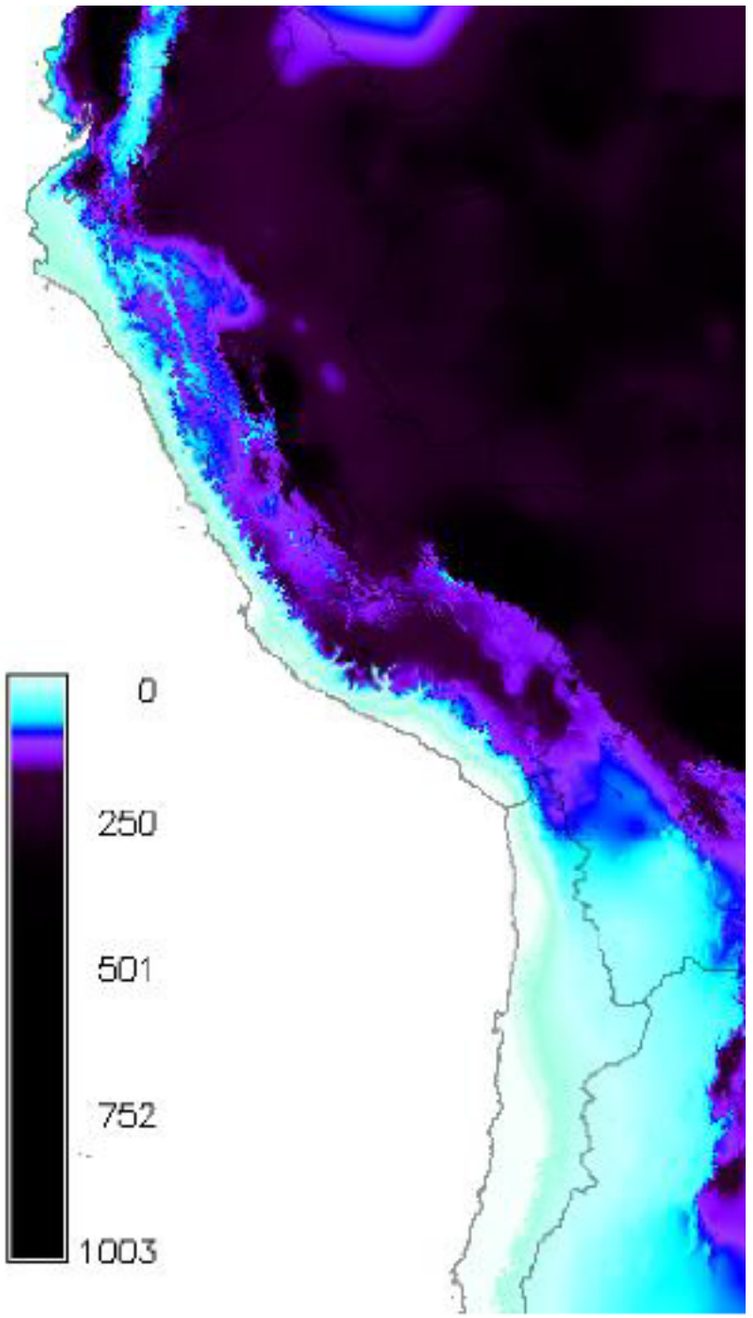
- Precipitation

- Temperature

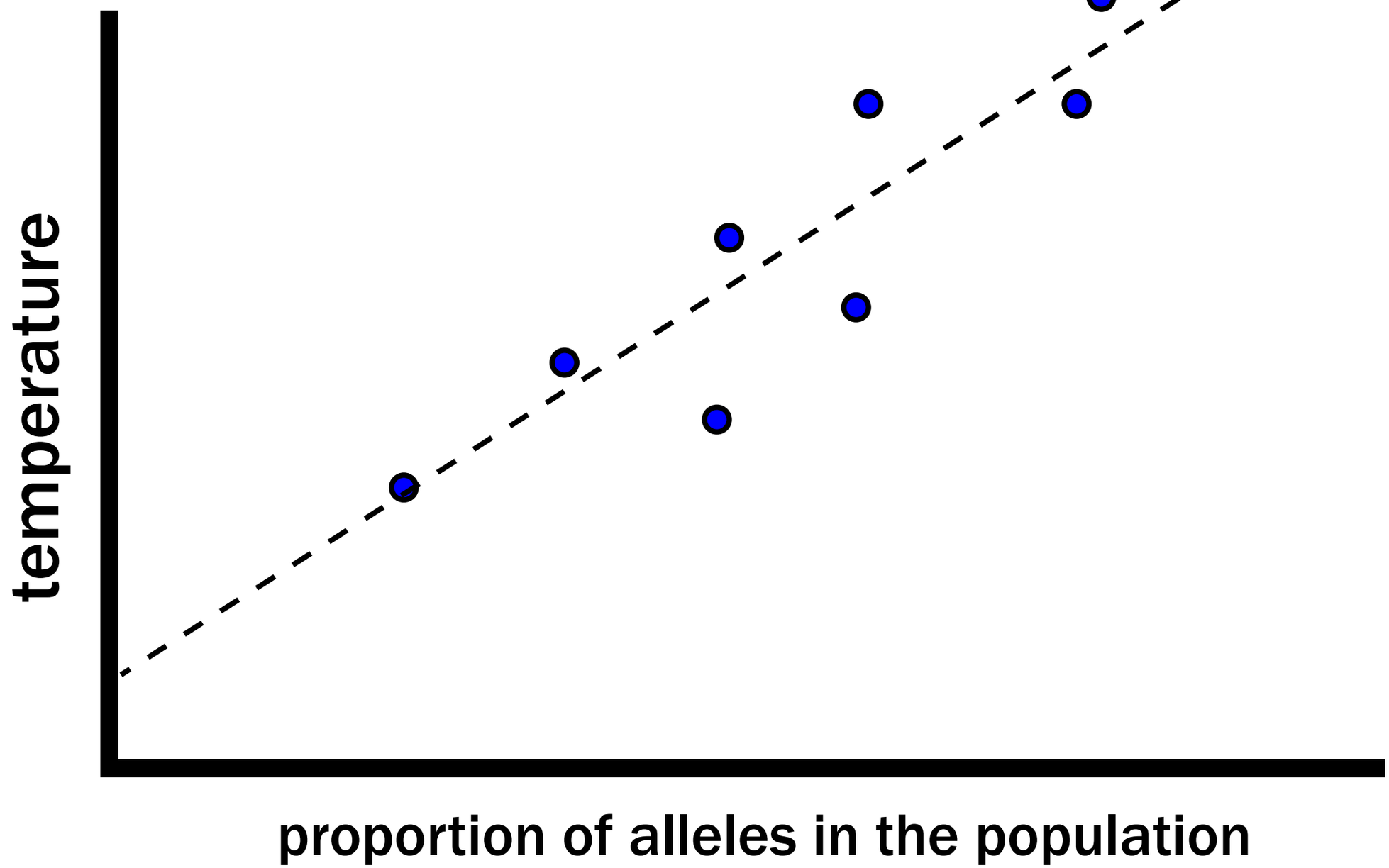
Biogeography







Biogeography



Biogeography

http://worldclim.org/

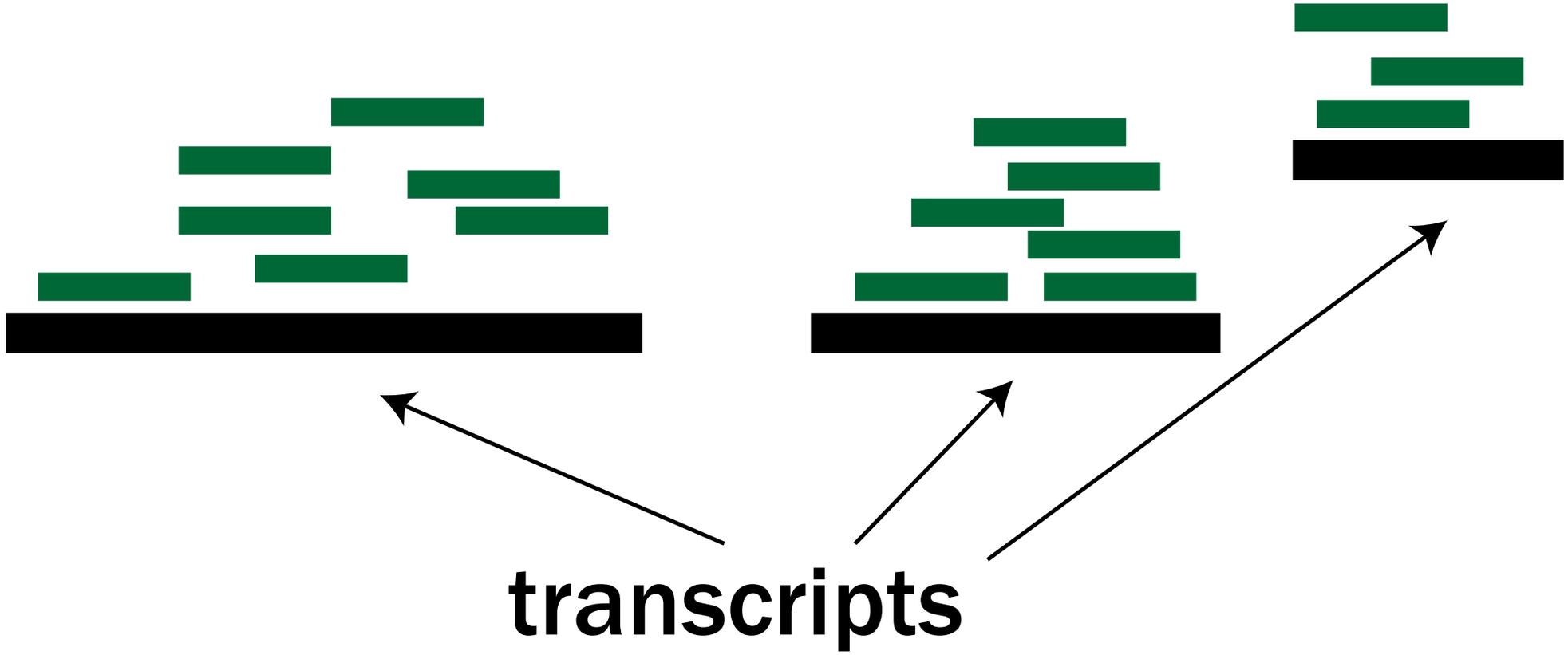
GRASS GIS

http://grass.osgeo.org/

GIS: Geographic Information System

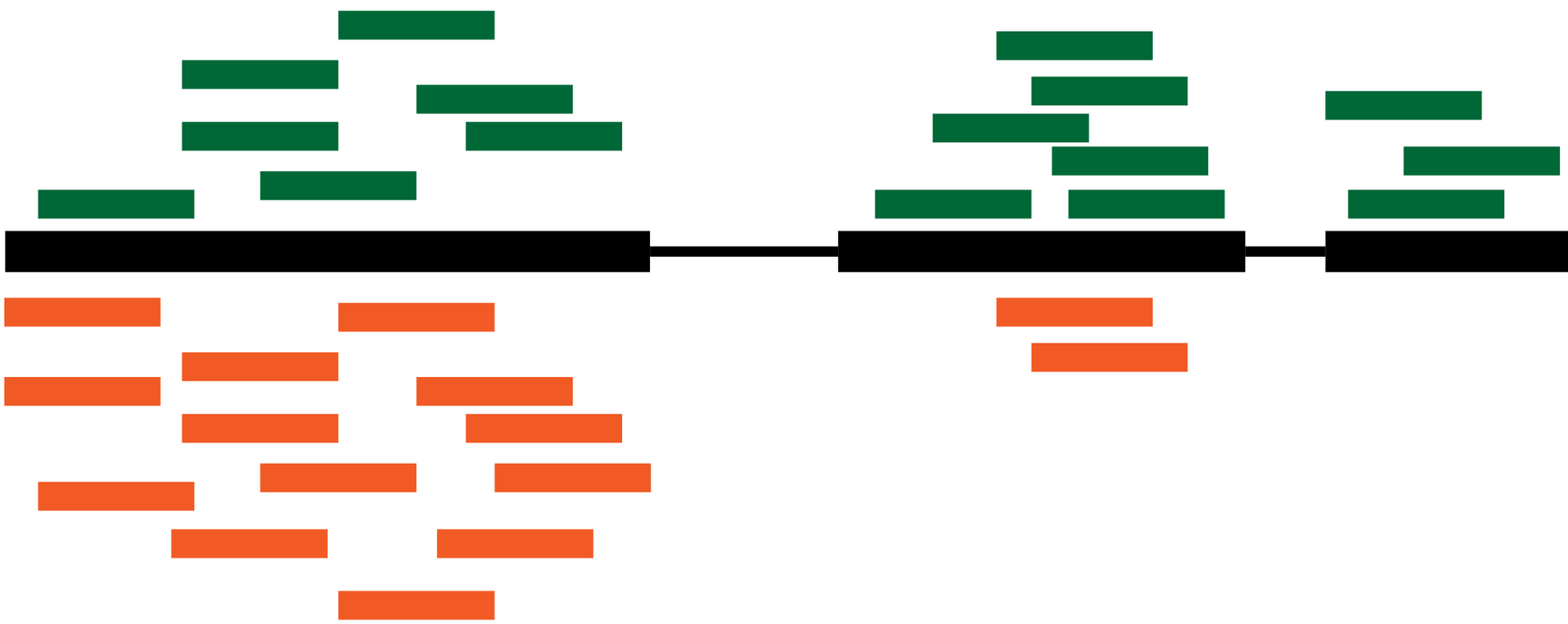
Differental Expression of Genes

Map reads



Differental Expression of Genes

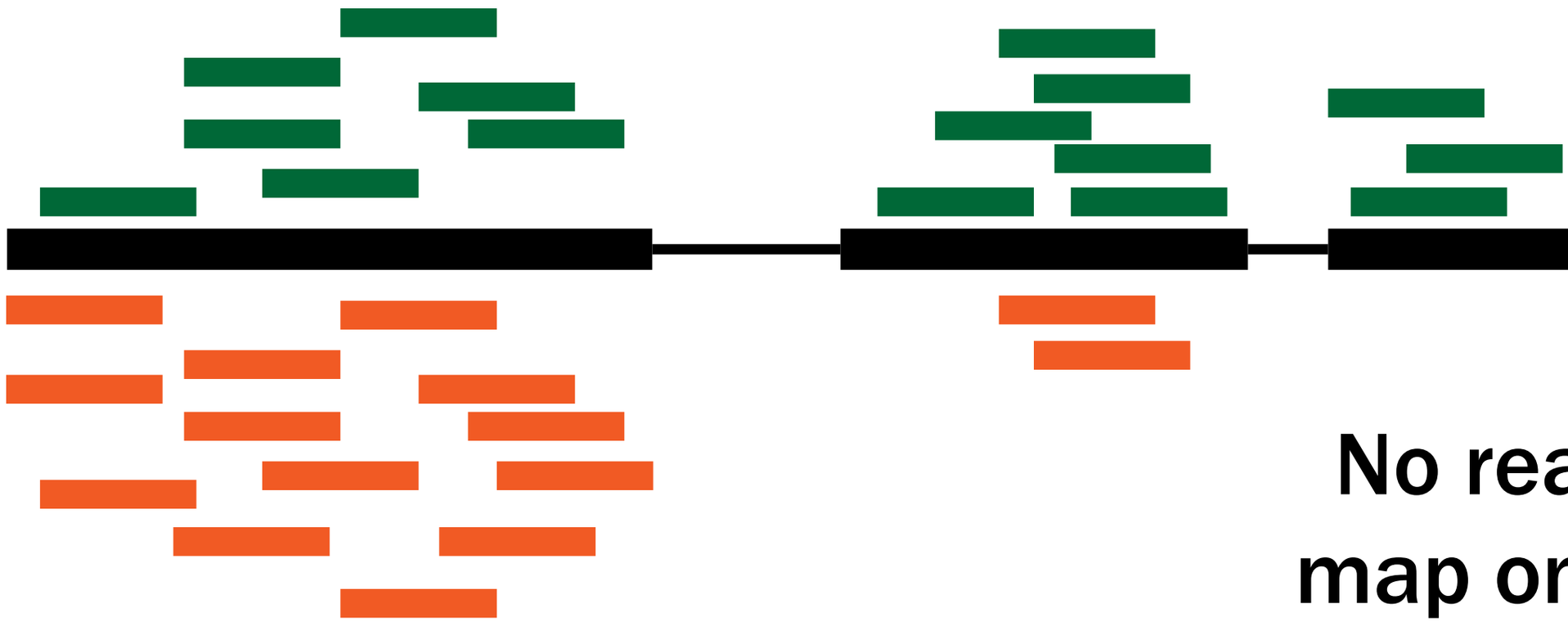
Sample #1



Sample #2

Differental Expression of Genes

More reads
map on #1



No reads
map on #2

More reads
map on #2

Differental Expression of Genes

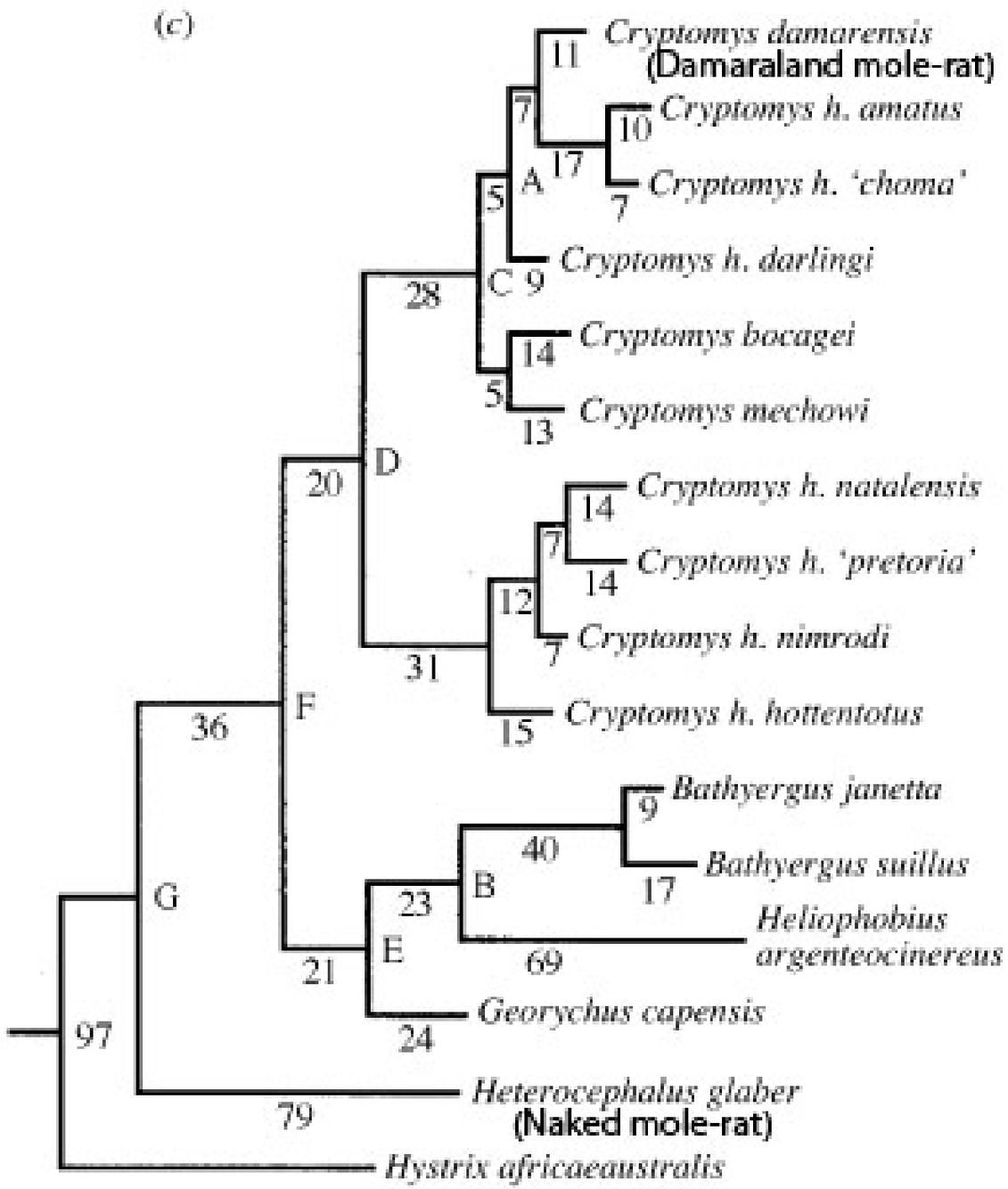
rank	gene name	t-statistic	P-val	fdr	log.fold change
1	L071601g000000	1128.8141060138	0	0	-3.58966974473853
2	L071608g006320	1000.22820252227	0	0	-3.75859437550754
3	L071603g078450	985.588268639028	0	0	-4.08715503043163
4	L071604g033420	737.336266145832	0	0	-3.74280522087657
5	L0716xxg003484	540.946677455803	0	0	-2.6152892502199
6	L0716xxg020555	535.93209804871	0	0	-3.95851648463624
7	L071612g055560	496.863848818756	0	0	-3.42695937641938
8	L071609g015670	478.278599761731	0	0	-3.50901279993839
9	L0716xxg024559	441.393366092215	0	0	-3.76356530270462
10	L0716xxg009114	408.424885129679	0	0	-2.88531810525456
11	L0716xxg005805	391.992960502799	0	0	-3.38198656632194
12	L071610g009310	372.565495161956	0	0	-3.49789184672633
13	L0716xxg005214	369.978745933731	0	0	-2.79756200226048
14	L0716xxg030016	368.841990445128	0	0	-1.97515466753591
15	L0716xxg007307	361.968975178021	0	0	-3.11137460910863
16	L0716xxg009467	361.285578767151	0	0	-2.43401469419127
17	L071603g113380	355.228562896043	0	0	-1.80252285029543
18	L0716xxg018667	341.904804510281	0	0	-1.85807794138901
19	L0716xxg003234	322.437357593479	0	0	-2.12621214896404
20	L071605g055310	304.645724744926	0	0	-1.61652185282518
21	L0716xxg007333	284.955310652926	0	0	-3.28216779375703
22	L0716xxg006541	272.94331844079	0	0	-3.90293184556053

Differental Expression of Genes

Software: DEseq, PoissonSeq, Cuffdiff, et al.

Check for a statistical difference between the number of reads between samples

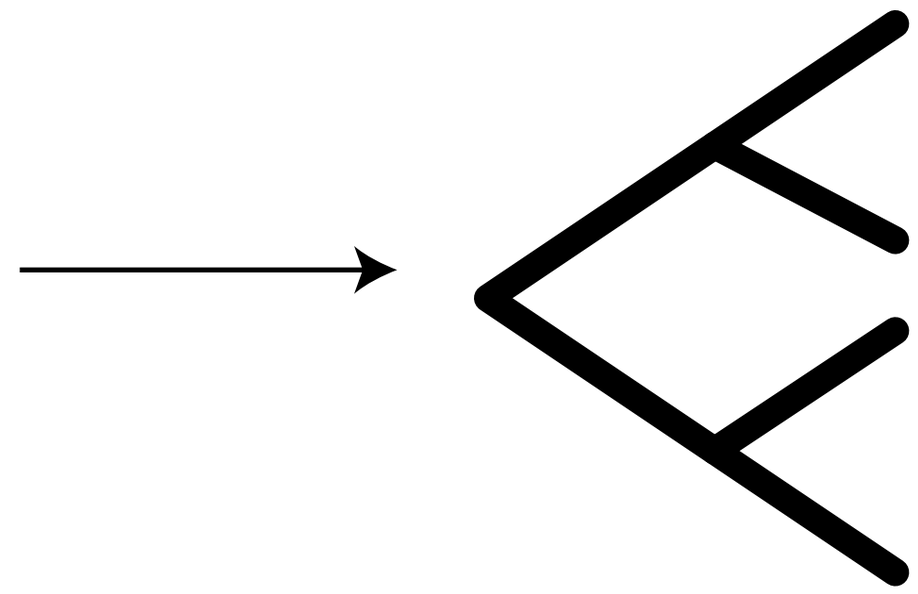
Phylogenetics



Phylogenetics

alignment to phylogeny

```
TATCGCGAAGGCTAGAACACAAGGTAAAATATCGCGA
TATCGCGAAGGCTAGAACGCAAGGTAAAATATCGCGA
TATCGCGAAGGCTAGAACACAAGGTAAAATATCGCGA
TATCGCGAAGACTAGAACACGAGGCAAAATATCGCGA
```



“Maximum Likelihood”

“Bayesian”

Phylogenetics

RAxML
http://sco.h-its.org/exelixis/software.html
(online)
http://embnet.vital-it.ch/raxml-bb/index.php
(and in Galaxy)

PHYML (online)
http://www.atgc-montpellier.fr/phyml/

Phylogenetics

alignments from many genes

```
TATCGCGAAGGCTAGAACA
TATCGCGAAGGCTAGAACA
TATCGCGAAGGCTAGAACA
TATCGCGAAGACTAGAACA
```

```
CAAGGTAAAATATCGCGA
CAAGGTAAAATATCGCGA
CAAGGTAAAATATCGCGA
```

```
AAGGCTAGAATCGCG
AAGGCTAGAATCGCG
AAGGCTAGAATCGCG
```

two primary methods for inferring trees

Phylogenetics

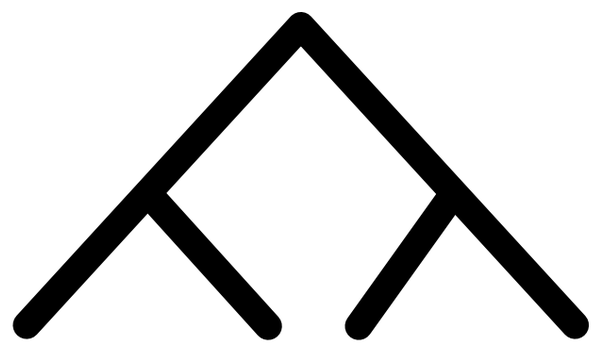
concatenation

TATCGCGAAGGCTAGAACA
TATCGCGAAGGCTAGAACA
TATCGCGAAGGCTAGAACA
TATCGCGAAGACTAGAACA

CAAGGTAAAATATCGCGA
CAAGGTAAAATATCGCGA
CAAGGTAAAATATCGCGA
CAAGGTAAAATATCGCGA

AAGGCTAGAATCGCG
AAGGCTAGAATCGCG
AAGGCTAGAATCGCG
AAGGCTAGAATCGCG

TATCGCGAAGGCTAGAACACAAGGTAAAATATCGCGAAAGGCTAGAATCGCG
TATCGCGAAGGCTAGAACACAAGGTAAAATATCGCGAAAGGCTAGAATCGCG
TATCGCGAAGGCTAGAACACAAGGTAAAATATCGCGAAAGGCTAGAATCGCG
TATCGCGAAGACTAGAACACAAGGTAAAATATCGCGAAAGGCTAGAATCGCG



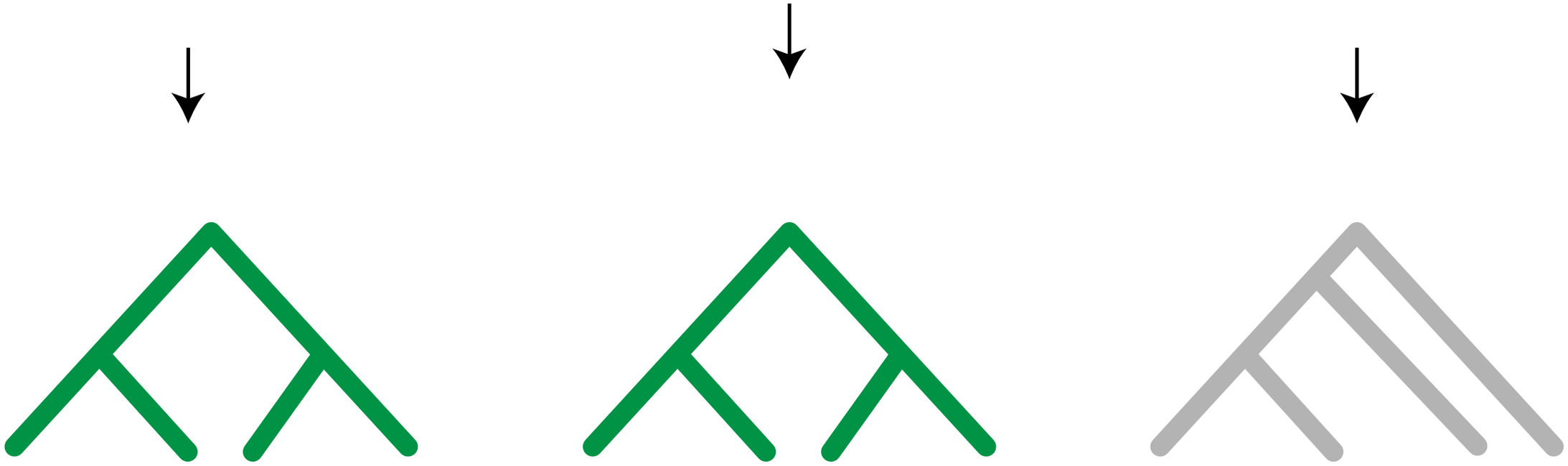
Phylogenetics

consensus

```
TATCGCGAAGGCTAGAACA
TATCGCGAAGGCTAGAACA
TATCGCGAAGGCTAGAACA
TATCGCGAAGACTAGAACA
```

```
CAAGGTAAAATATCGCGA
CAAGGTAAAATATCGCGA
CAAGGTAAAATATCGCGA
CAAGGTAAAATATCGCGA
```

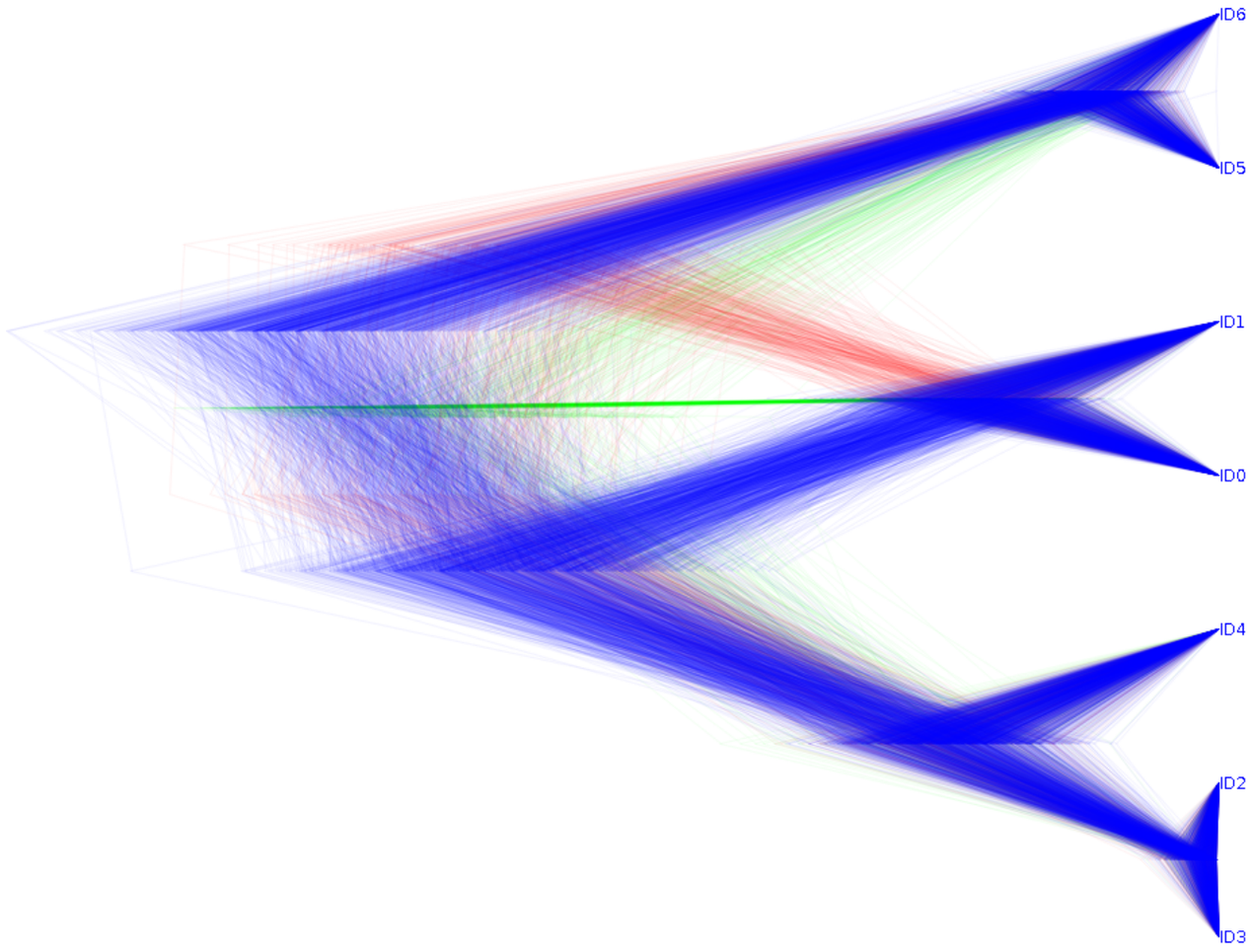
```
AAGGCTAGAATCGCG
AAGGCTAGAATCGCG
AAGGCTAGAATCGCG
AAGGCTAGAATCGCG
```



Phylogenetics

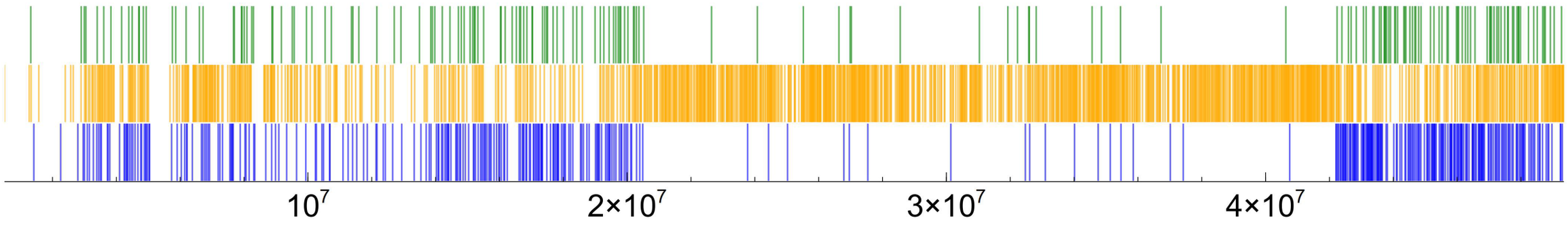
“DensiTree”

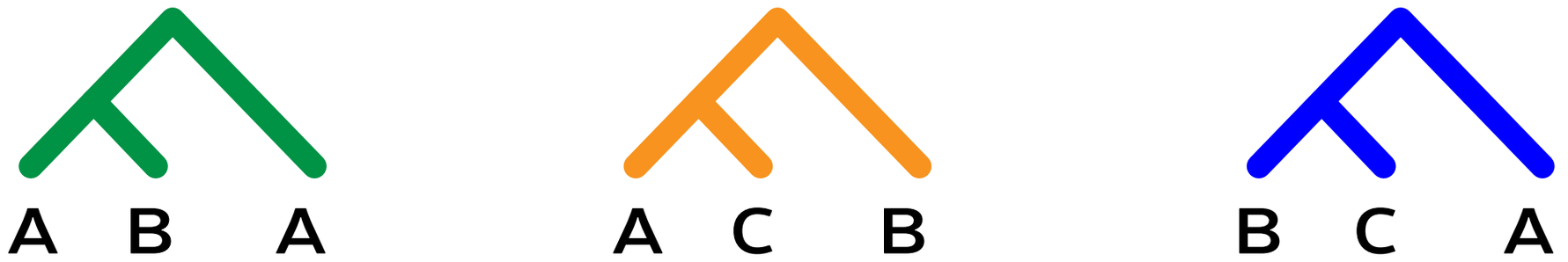
all gene trees overlaid in one picture



Phylogenetics

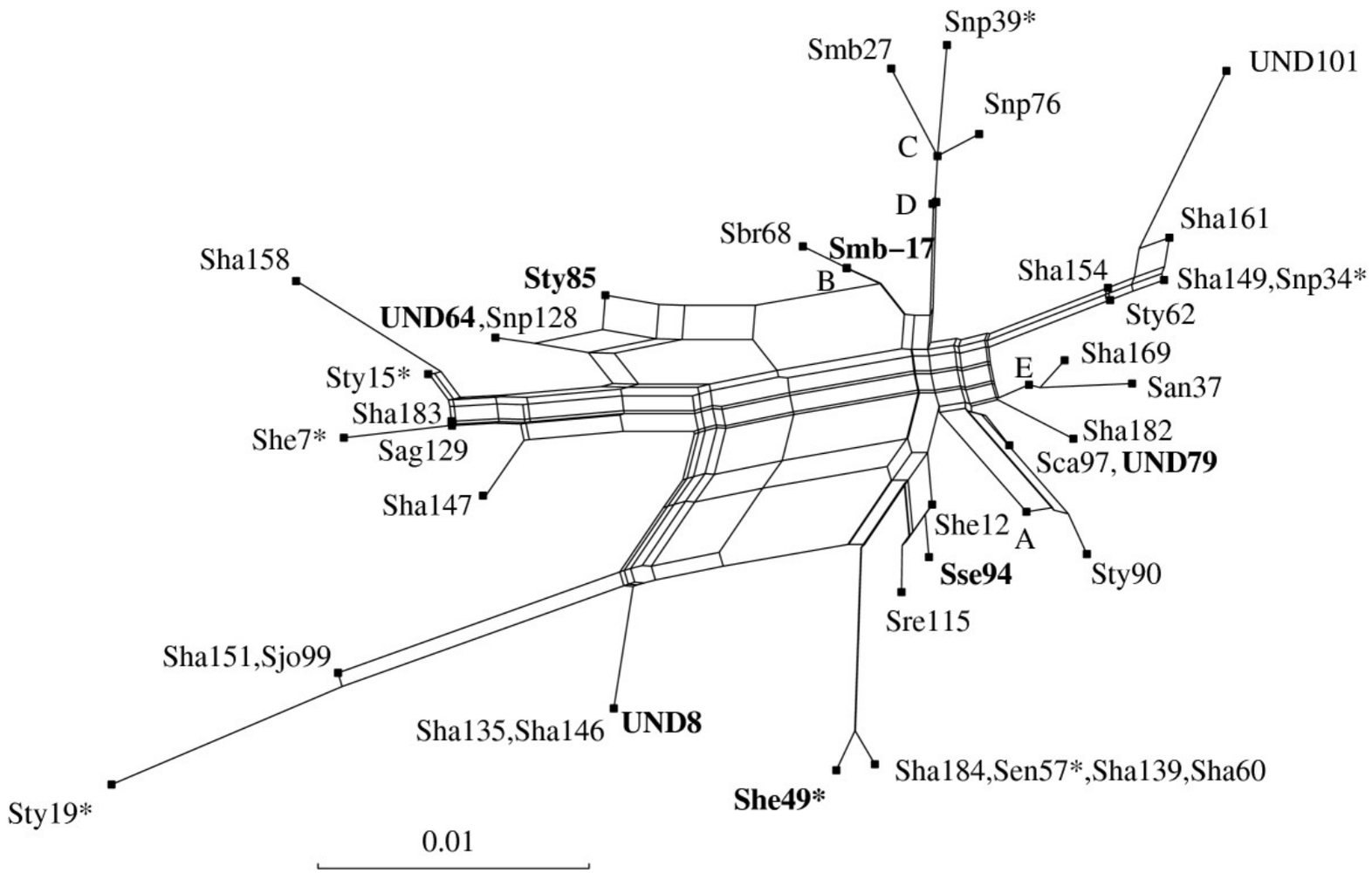
a phylogeny from each region
along the chromosome





Phylogenetics

NetworkNet



http://ab.inf.uni-tuebingen.de/software/splitstree4/welcome.html