MCB5472
Computer methods in molecular evolution

Lecture 4/7/2014

Old Assignment

Write a script that takes all phylip formated aligned multiple sequence files present in a directory, and performs a bootstrap analyses using maximum parsimony.

Files you might want to use are A.fa, B.fa, alpha.fa, beta.fa from last week's assignment, and atp_all.phy. BUT you first have to **align** them and convert them to **phylip format*** AND you should replace gaps with "?"

(In the end you would be able to answer the question "does the resolution increase if a more related subgroup is analyzed independent from an outgroup?)

- clustalw2 is one program frequently used to convert formats
- `system("clustalw -infile=$file.fa –convert -output=PHYLIP");`

Phylip

written and distributed by Joe Felsenstein and collaborators (some of the following is copied from the PHYLIP homepage)

PHYLIP (the *PHYL*ogeny *I*nference *P*ackage) is a package of programs for inferring phylogenies (evolutionary trees).

PHYLIP is the most widely-distributed phylogeny package, and competes with PAUP* to be the one responsible for the largest number of published trees. PHYLIP has been in distribution since 1980, and has over 15,000 registered users.

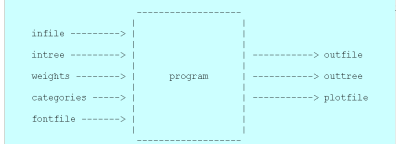
Output is written onto special files with names like "outfile" and "outtree". Trees written onto "outtree" are in the Newick format, an informal standard agreed to in 1986 by authors of a number of major phylogeny packages.

Input is either provided via a file called "infile" or in response to a prompt.

input and output

Input and output files

For most of the PHYLIP programs, information comes from a series of input files, and ends up in a series of output files:



The programs interact with the user by presenting a menu. Aside from the user's choices from the menu, they read all other input from files. These files have default names. The program will try to find a file of that name - if it does not, it will ask the user to supply the name of that file. Input data such as DNA sequences comes from a file whose default name is `infile`. If the user supplies a tree, this is in a file whose default name is `intree`. Values of weights for the characters are in `weights`, and the tree plotting program need some digitized fonts which are supplied in `fontfile` (all these are default names).

What's in PHYLIP

Programs in PHYLIP allow to do parsimony, distance matrix, and likelihood methods, including bootstrapping and consensus trees. Data types that can be handled include molecular sequences, gene frequencies, restriction sites and fragments, distance matrices, and discrete characters.

Phylip works well with protein and nucleotide sequences
Many other programs mimic the style of PHYLIP programs. (e.g. TREEPUZZLE, phyml, protml)

Many other packages use PHYIP programs in their inner workings (e.g., SEAVIEW)

PHYLIP runs under all operating systems

Web interfaces are available

Programs in PHYLIP are Modular

For example:

SEQBOOT take one set of aligned sequences and writes out a file containing bootstrap samples.

PROTDIST takes a aligned sequences (one or many sets) and calculates distance matices (one or many)

FITCH (or NEIGHBOR) calculate best fitting or neighbor joining trees from one or many distance matrices

CONSENSE takes many trees and returns a consensus tree

…. modules are available to draw trees as well, but often people use figtree or njplot

The Phylip Manual is an excellent source of information.

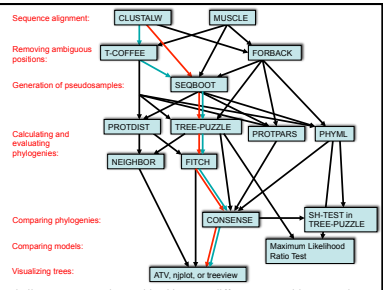
Brief one line descriptions of the programs are here

The easiest way to run PHYLIP programs is via a command line menu (similar to clustalw). The program is invoked through clicking on an icon, or by typing the program name at the command line.

```
> seqboot
> protpars
> fitch
```

If there is no file called infile the program responds with:

```
[gogarten@carrot gogarten]$ seqboot
seqboot: can't find input file "infile"
Please enter a new file name>
```



Phylip programs can be combined in many different ways with one another and with programs that use the same file formats.

Example 1 Protpars

example: seqboot, protpars, consense on atp_all.phy

NOTE the bootstrap majority consensus tree does not necessarily have the same topology as the "best tree" from the original data!

threshold parsimony,
gap symbols - versus ?
(in vi you could use `:%s/-/?/g` to replace all – ?)
outfile
outtree compare to distance matrix analysis

create *.phy files

the easiest (probably) is to run clustalw with the phylip option:
For example (here):

```
#!/usr/bin/perl -w
print "# This program aligns all multiple sequence files with names *.fa \n
# found in its directory using clustalw,  and saves them in phyip format.\n";
while(defined($file=glob("*.fa*"))){
        @parts=split(/\./,$file);
        $file=$parts[0];
        system("clustalw -infile=$file.fa -align -output=PHYLIP");
        #if you only want to convert files use
        #system("clustalw -infile=$file.fa -convert -output=PHYLIP");
        };
# cleanup:
system ("rm *.dnd");
exit;
```

Alternative for entering the commands for the menu:

```
#!/usr/bin/perl -w
        system ("cp A.phy infile");
        system ("echo -e 'y\n9\n'|seqboot");
exit;
```

`echo` returns the string in ` `, i.e., `y\n9\n`.
The `-e` options allows the use of `\n`
The `|` symbol pipes the output from echo to seqboot

New Assignment

•Given a multiple fasta sequence file*, write a script that for each sequence extract the gi number and the species name, and then rewrites the file so that the annotation line starts with the gi number, followed by the species/strain name, followed by a space. (The gi number and the species name should not be separated by or contain any spaces – replace them by _. This is useful, because many programs will recognize the number and name as handle for the sequence (e.g., clustalw2 and phyml)

Assume that the annotation line follows the NCBI convention and begins with the
> followed by the gi number, and ends with the species and strain designation given in []
Example:

```
>gi|229240723|ref|ZP_04365119.1| primary replicative DNA helicase; intein [Cellulomonas flavigena DSM 20109]
```

*An example multiple sequence file in the unofficial NCBI formatted annotation line is here.

Bayes' Theorem

Reverend Thomas Bayes (1702-1761)

$$P(model|data, I) = P(model, I)\ \frac{P(data|model, I)}{P(data,I)}$$

Likelihood: describes how well the model predicts the data

Posterior Probability: represents the degree to which we believe a given **model** accurately describes the situation given the available **data** and all of our prior information **I**

Prior Probability: describes the degree to which we believe the model accurately describes reality based on all of our prior information.

Normalizing constant

Illustration of a biased random walk

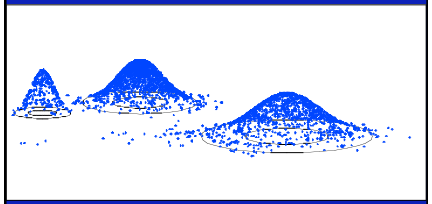


Figure generated using MCRobot program (Paul Lewis, 2001)

Alternative Approaches to Estimate Posterior Probabilities

Bayesian Posterior Probability Mapping with MrBayes
(Huelsenbeck and Ronquist, 2001)

Problem:

Strimmer's formula $p_i = \frac{L_i}{L_1+L_2+L_3}$ only considers 3 trees (those that maximize the likelihood for the three topologies)

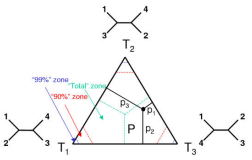
Solution:

Exploration of the tree space by sampling trees using a biased random walk (Implemented in MrBayes program)

Trees with higher likelihoods will be sampled more often

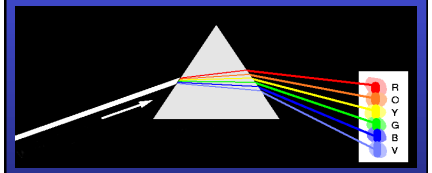
$p_i = \frac{N_i}{N_{total}}$,where N_i - number of sampled trees of topology i, i=1,2,3
N_{total} – total number of sampled trees (has to be large)

ml mapping



From: Olga Zhaxybayeva and J Peter Gogarten *BMC Genomics* 2002, 3:4

Decomposition of Phylogenetic Data



Phylogenetic information present in genomes

Break information into small quanta of information (bipartitions or embedded quartets)

Analyze spectra to detect transferred genes and plurality consensus.

TOOLS TO ANALYZE PHYLOGENETIC INFORMATION FROM MULTIPLE GENES IN GENOMES:

Bipartition Spectra (Lento Plots)

BIPARTITION OF A PHYLOGENETIC TREE

Bipartition (or split) – a division of a phylogenetic tree into two parts that are connected by a single branch.
It divides a dataset into two groups, but it does not consider the relationships within each of the two groups.

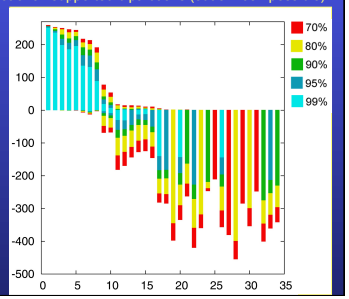
Yellow *vs* Rest
compatible to illustrated bipartition

Orange vs Rest
incompatible to illustrated bipartition

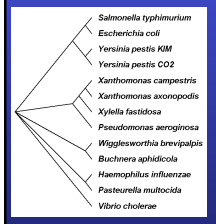
"Lento"-plot of 34 supported bipartitions (out of 4082 possible)

13 gamma-proteobacterial genomes (258 putative orthologs):
- E.coli
- Buchnera
- Haemophilus
- Pasteurella
- Salmonella
- Yersinia pestis (2 strains)
- Vibrio
- Xanthomonas (2 sp.)
- Pseudomonas
- Wigglesworthia

There are 13,749,310,575 possible unrooted tree topologies for 13 genomes

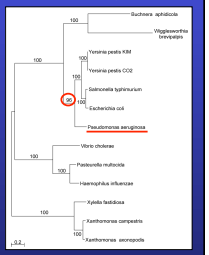


Consensus clusters of eight significantly supported bipartitions



only 258 genes analyzed

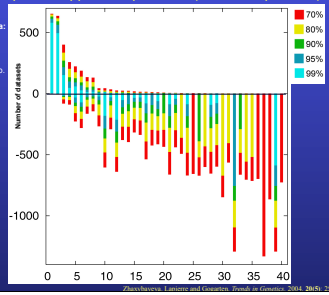
Phylogeny of putatively transferred gene (virulence factor homologs (mviN))



"Lento"-plot of supported bipartitions (out of 501 possible)

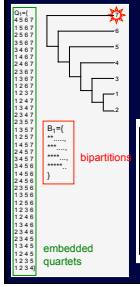
10 cyanobacteria:
- *Anabaena*
- *Trichodesmium*
- *Synechocystis* sp.
- *Prochlorococcus marinus* (3 strains)
- *Marine Synechococcus*
- *Thermo-synechococcus elongatus*
- *Gloeobacter*
- *Nostoc punctiforme*

Based on 678 sets of orthologous genes



Zhaxybayeva, Lapierre and Gogarten. *Trends in Genetics*. 2004. **20**(5): 254-260

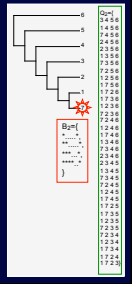
PROBLEMS WITH BIPARTITIONS (A)

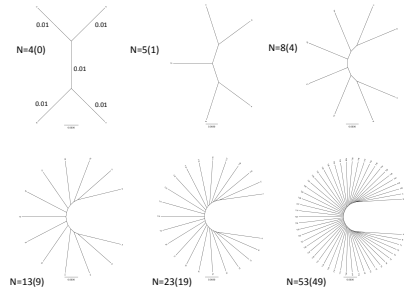


A single rogue sequence that moves from one end of a Hennigian comb to the other changes all bipartition

supported quartets
Q1 ∩ Q2 =
{3 4 5 6, 1 4 5 6, 2 4 5 6, 2 3 5 6, 1 3 5 6, 1 2 5 6, 1 2 3 6, 1 2 4 6, 1 3 4 6, 2 3 4 6, 2 3 4 5, 1 3 4 5, 1 2 4 5, 1 2 3 5, 1 2 3 4}

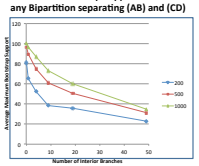
supported bipartitions:
B1 ∩ B2 = ∅



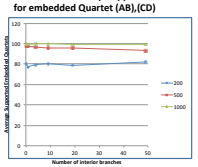


Results :

Maximum Bootstrap Support value for any Bipartition separating (AB) and (CD)



Maximum Bootstrap Support value for embedded Quartet (AB),(CD)



Bipartition Paradox:

- The more sequences are added, the lower the support for bipartitions that include all sequences. The more data one uses, the lower the bootstrap support values become.
- This paradox disappears when only embedded splits for 4 sequences are considered.

Bootstrap support values for embedded quartets

: tree calculated from one pseudo-sample generated by bootstraping from an alignment of one gene family present in 11 genomes

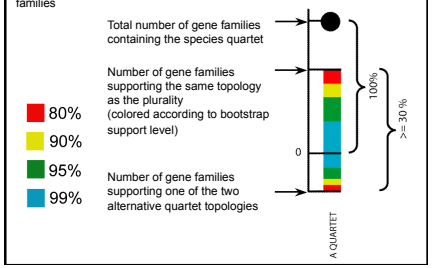
: embedded quartet for genomes 1, 4, 9, and 10 .
This bootstrap sample supports the topology ((1,4),9,10).

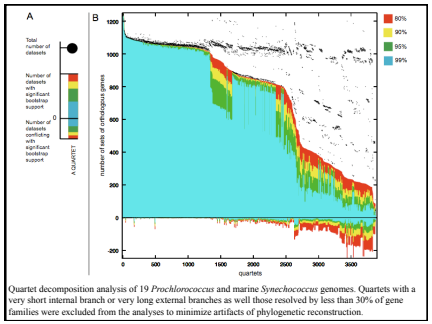
Quartet spectral analyses of genomes iterates over three loops:
- ➢Repeat for all bootstrap samples.
- ➢Repeat for all possible embedded quartets.
- ➢Repeat for all gene families.

Zhaxybayeva et al. 2006. Genome Research. 16(6): 1586-96

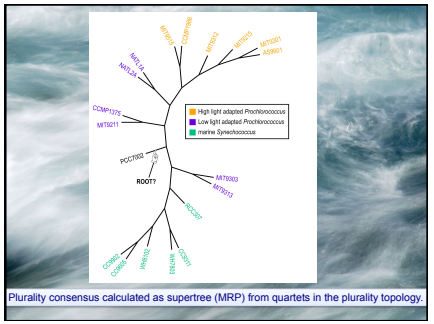
Illustration of one component of a quartet spectral analyses

Summary of phylogenetic information for one genome quartet for all gene families

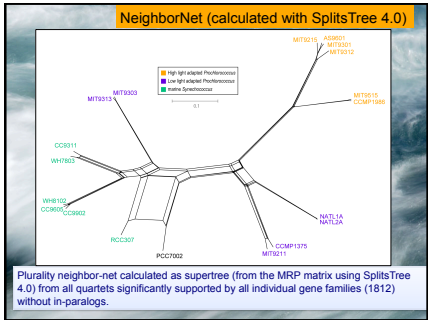




Quartet decomposition analysis of 19 *Prochlorococcus* and marine *Synechococcus* genomes. Quartets with a very short internal branch or very long external branches as well those resolved by less than 30% of gene families were excluded from the analyses to minimize artifacts of phylogenetic reconstruction.



Plurality consensus calculated as supertree (MRP) from quartets in the plurality topology.



Plurality neighbor-net calculated as supertree (from the MRP matrix using SplitsTree 4.0) from all quartets significantly supported by all individual gene families (1812) without in-paralogs.

Neutral theory:

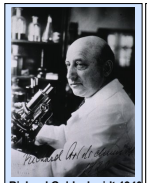
The vast majority of observed sequence differences between members of a population are neutral (or close to neutral). These differences can be fixed in the population through random genetic drift. Some mutations are strongly counter selected (this is why there are patterns of conserved residues). Only very seldom is a mutation under positive selection.

The neutral theory does not say that all evolution is neutral and everything is only due to to genetic drift.

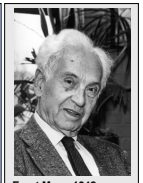
Nearly Neutral theory:

Even synonymous mutations do not lead to random composition but to codon bias. Small negative selection might be sufficient to produce the observed codon usage bias.

How do you define evolution?



Richard Goldschmidt 1940
hopeful monsters
Mutationism HGT/WGD!
Punctuated Equilibrium
Few genes / large effect
Vilified by Mayr, celebrated 1977 Gould & Evo-devo

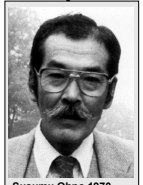


Ernst Mayr 1942
NeoDarwinian Synthesis
Natural Selection
Gradualism
Many genes/small effect
Dario – "Fisher right"

Motoo Kimura 1968
Neutral Theory
Genetic Drift is main force for changing allele frequencies

Slide from Chris Pires

Duplications and Evolution



Susumu Ohno 1970
Evolution by gene duplication
1R and 2R hypothesis

"Junk DNA" 1972

Ohno postulated that gene duplication plays a major role in evolution

Small scale duplications (SSD)
Whole genome duplications (WGD)

- **Polyploid: nucleus contains three or more copies of each chromosome**
- **Autopolyploid: formed within a single species**
  Diploids AA and A'A' ⇨ Polyploid AAA'A'
- **Allopolyploid: formed from more than one species**
  Diploids AA and BB ⇨ Polyploid AABB

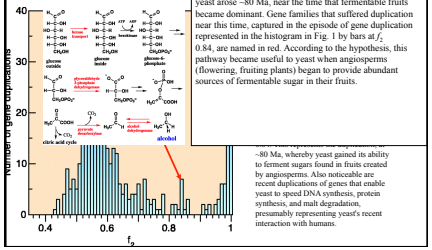
Slide from Chris Pires

What is it good for?

Gene duplication events can provide an outgroup that allows rooting a molecular phylogeny.
Most famously this principle was applied in case of the tree of life – the only outgroup available in this case are ancient paralogs (see http://gogarten.uconn.edu/cvs/Publ_Pres.htm for more info).
However, the same principle also is applicable to any group of organisms, where a duplication preceded the radiation (example).
Lineage specific duplications also provide insights into which traits were important during evolution of a lineage.

e.g. gene duplications in yeast

from Benner et al., 2002



The chemical pathway that converts glucose to alcohol in yeast arose ~80 Ma, near the time that fermentable fruits became dominant. Gene families that suffered duplication near this time, captured in the episode of gene duplication represented in the histogram in Fig. 1 by bars at f_g 0.84, are named in red. According to the hypothesis, this pathway became useful to yeast when angiosperms (flowering, fruiting plants) began to provide abundant sources of fermentable sugar in their fruits.

~80 Ma, whereby yeast gained its ability to ferment sugars found in fruits created by angiosperms. Also noticeable are recent duplications of genes that enable yeast to speed DNA synthesis, protein synthesis, and malt degradation, presumably representing yeast's recent interaction with humans.

the gradualist point of view

Evolution occurs within populations where the fittest organisms have a selective advantage. Over time the advantages genes become fixed in a population and the population gradually changes.

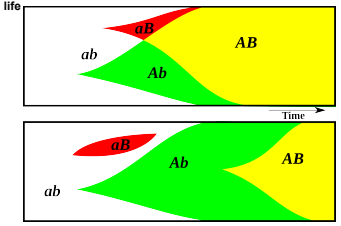
Note: this is not in contradiction to the the theory of neutral evolution. (which says what ?)

Processes that MIGHT go beyond inheritance with variation and selection?

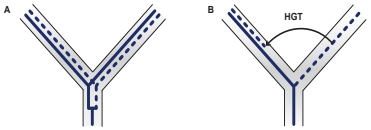
- Horizontal gene transfer and recombination
- Polyploidization (botany, vertebrate evolution) see here or here
- Fusion and cooperation of organisms (Kefir, lichen, also the eukaryotic cell)
- Targeted mutations (?), genetic memory (?) (see Foster's and Hall's reviews on directed/adaptive mutations; see here for a counterpoint)
- Random genetic drift
- Gratuitous complexity
- Selfish genes (who/what is the subject of evolution??)
- Parasitism, altruism, Morons
- Evolutionary capacitors
- Hopeless monsters (in analogy to Goldschmidt's hopeful monsters)

Gene Transfer, Sex, and Recombination:

- **Inventions do not need to be made sequentially**
- **Gene transfer, followed by homologous or non-homologous recombination, allows inventions to be shared across the tree of life**



Aside: Gene and genome duplication versus Horizontal Gene Transfer



Autochtonous gene/genome duplication are rare in prokaryotes

Gene family expansion through horizontal gene transfer – the most common process in prokaryotes

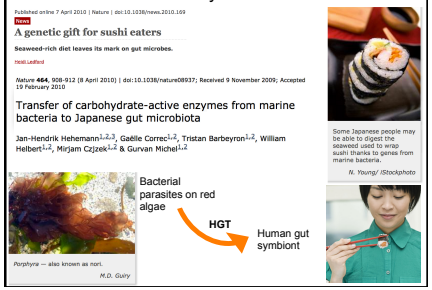
Horizontal Gene Transfer (HGT) and the Acquisition of New Capabilities

- Most important process to adapt microorganisms to new environments. E.g.: Antibiotic and heavy metal resistance, pathways that allow acquisition and breakdown of new substrates.
- Creation of new metabolic pathways.
- HGT not autochthonous gene duplication is the main process of gene family expansion in prokaryotes.
- Also important in the recent evolution of multicellular eukaryotes (HGT between fish species and between grasses).

Selection acts on the Holobiont (= Host + Symbionts)

- To adapt to new conditions, new symbionts can be acquired, or existing symbionts can acquire new genes through HGT.

Gene Transfer in Eukaryotes



Published online 7 April 2010 | Nature | doi:10.1038/news.2010.169

News

A genetic gift for sushi eaters

Seaweed-rich diet leaves its mark on gut microbes.

Heidi Ledford

Nature **464**, 908-912 (8 April 2010) | doi:10.1038/nature08937; Received 9 November 2009; Accepted 19 February 2010

Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota

Jan-Hendrik Hehemann[1,2,3], Gaëlle Correc[1,2], Tristan Barbeyron[1,2], William Helbert[1,2], Mirjam Czjzek[1,2] & Gurvan Michel[1,2]

Some Japanese people may be able to digest the seaweed used to wrap sushi thanks to genes from marine bacteria.

N. Young/ iStockphoto

Porphyra — also known as nori.

M.D. Guiry

Bacterial parasites on red algae

HGT

Human gut symbiont

Gene Transfer in Eukaryotes – Example 2

Highlights

- Key genes for C_4 photosynthesis were transmitted between distantly related grasses
- These genes contributed to the adaptation of the primary metabolism
- Their transmission was independent from most of the genome

Curr Biol. 2012 Mar 6;22(5):445-9. Epub 2012 Feb 16.

Adaptive Evolution of C(4) Photosynthesis through Recurrent Lateral Gene Transfer.

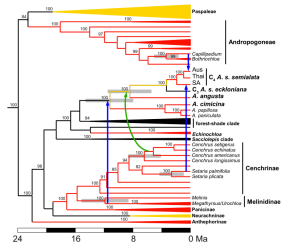
Christin PA, Edwards EJ, Besnard G, Boxall SF, Gregory R, Kellogg EA, Hartwell J, Osborne CP.

C_4 Photosynthesis: Need a Gene? Borrow One!

Eric H. Roalson Current Biology Vol 22 No 5 R162

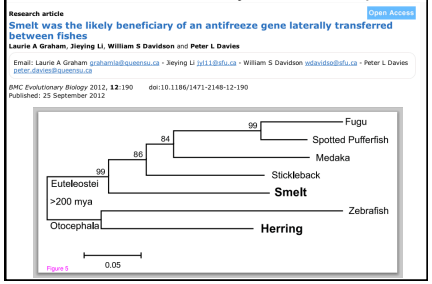
Horizontal gene transfer has been increasingly documented between eukaryotes, but a new study suggests a much larger role for horizontal gene transfer in physiological adaption through the transfer of photosynthetic pathway genes.

Gene Transfer in Eukaryotes – Example 2



From: Christin PA, Edwards EJ, Besnard G, Boxall SF, Gregory R, Kellogg EA, Hartwell J, Osborne CP. Adaptive Evolution of C(4) Photosynthesis through Recurrent Lateral Gene Transfer. Curr Biol. 2012 Mar 6;22(5):445-9. Epub 2012 Feb 16.

Gene Transfer in Eukaryotes – Example 3



Research article

Open Access

Smelt was the likely beneficiary of an antifreeze gene laterally transferred between fishes

Laurie A Graham, Jieying Li, William S Davidson and Peter L Davies

Email: Laurie A Graham grahamla@queensu.ca - Jieying Li jyl11@sfu.ca - William S Davidson wdavidso@sfu.ca - Peter L Davies peter.davies@queensu.ca

BMC Evolutionary Biology 2012, **12**:190 doi:10.1186/1471-2148-12-190

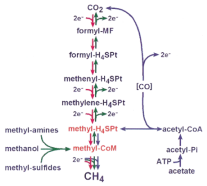
Published: 25 September 2012

Fugu
Spotted Pufferfish
Medaka
Stickleback
Smelt
Zebrafish
Herring
Osmeriformes
Euteleostei
Otocephala
>200 mya
99
84
86
99
0.05
Figure 5

HGT as a force creating new pathways

HGT as a force creating new pathways – Example I

Acetoclastic Methanogenesis

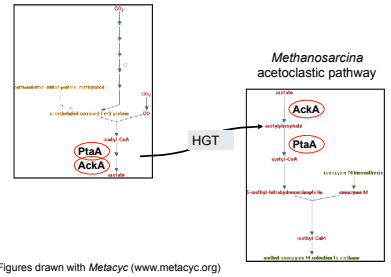


- Unique to subset of Archaea
- Energy production via reduction of multiple carbon substrates to CH_4
- 900 Million metric tons of biogenic methane produced annually.
- Over 66% of biogenic methane is produced from acetate, mostly by *Methanosarcina* genera.

From: Galagan et al., 2002

Fournier and Gogarten (2008) Evolution of Acetoclastic Methanogenesis in Methanosarcina via Horizontal Gene Transfer from Cellulolytic Clostridia. *J. Bacteriol.* 190(3): 1124-7

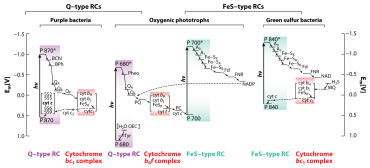
Clostridia acetigenic pathway



Figures drawn with *Metacyc* (www.metacyc.org)

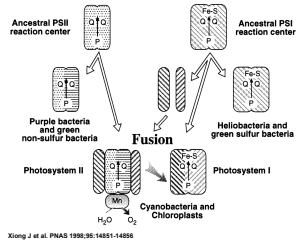
HGT as a force creating new pathways – Example 2

Oxygen producing photosynthesis



Hohmann-Marriott MF, Blankenship RE. 2011. Annu. Rev. Plant Biol. 62:515–48

A heterologous fusion model for the evolution of oxygenic photosynthesis based on phylogenetic analysis.



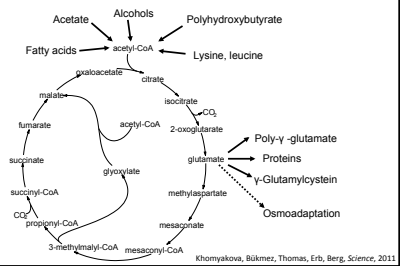
Xiong J et al. PNAS 1998;95:14851-14856

©1998 by National Academy of Sciences

PNAS

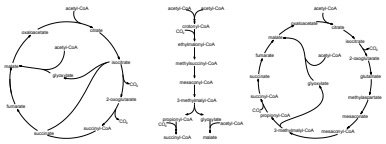
HGT as a force creating new pathways – Example 3

Acetyl-CoA Assimilation: Methylaspartate Cycle



Khomyakova, Bükmez, Thomas, Erb, Berg, *Science*, 2011

Comparison of different anaplerotic pathways



Citric acid cycle and Glyoxylate cycle

Bacteria, Eukarya and some Archaea

Ethylmalonyl-CoA pathway

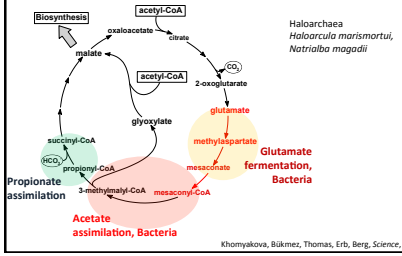
α-Proteobacteria, streptomycetes

Methylaspartate cycle

haloarchaea

HGT as a force creating new pathways – Example 3

Acetyl-CoA Assimilation: methylaspartate cycle

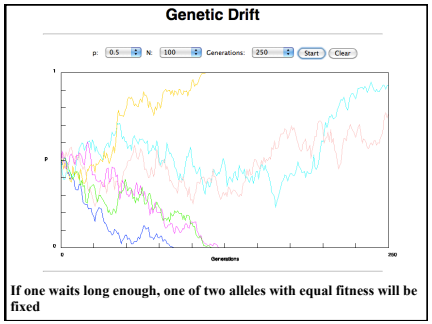


Haloarchaea
Haloarcula marismortui, Natrialba magadii

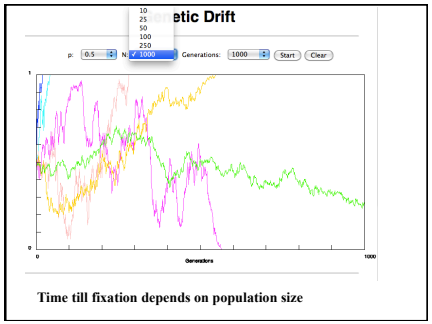
Khomyakova, Bükmez, Thomas, Erb, Berg, *Science*, 2011

selection versus drift

The larger the population the longer it takes for an allele to become fixed.
Note: Even though an allele conveys a strong selective advantage of 10%, the allele has a rather large chance to go extinct.
Note#2: Fixation is faster under selection than under drift.



If one waits long enough, one of two alleles with equal fitness will be fixed



Time till fixation depends on population size

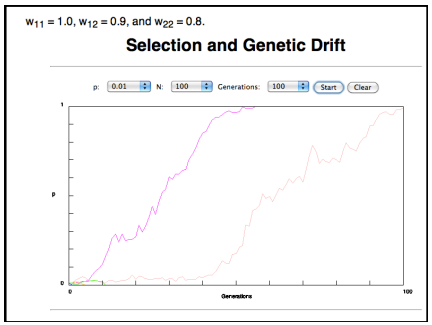
s=0

Probability of fixation, P, is equal to frequency of allele in population.
Mutation rate (per gene/per unit of time) = u ;
freq. with which allele is generated in diploid population size N =u*2N
Probability of fixation for each allele = 1/(2N)

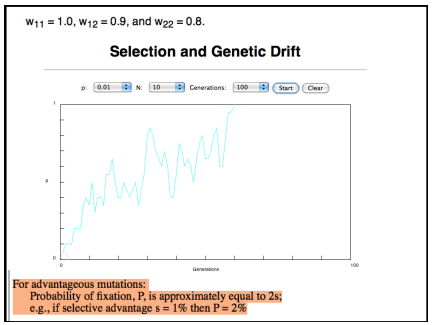
Substitution rate =
frequency with which new alleles are generated * Probability of fixation=
u*2N *1/(2N) = u = Mutation rate
Therefore:
If f s=0, the substitution rate is independent of population size, and equal to the mutation rate !!!! (NOTE: Mutation unequal Substitution!)
This is the reason that there is hope that the molecular clock might sometimes work.

Fixation time due to drift alone:
t_{av}=4*N_e generations
(N_e=effective population size; For n discrete generations
N_e= n/(1/N_1+1/N_2+.....1/N_n)

w_{11} = 1.0, w_{12} = 0.9, and w_{22} = 0.8.



w_{11} = 1.0, w_{12} = 0.9, and w_{22} = 0.8.

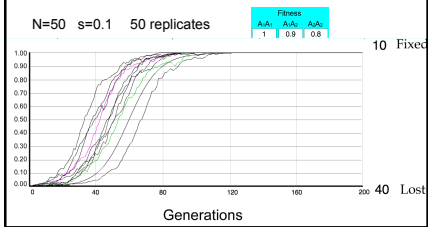


For advantageous mutations:
Probability of fixation, P, is approximately equal to 2s:
e.g., if selective advantage s = 1% then P = 2%

How does the survival of multiple alleles in a population depend on the population size.

Drift and Selection (interesting setting: P=0.01, N=50)

Note: Even though the allele conveys a strong selective advantage of 10%, the allele has a rather large chance to go extinct quickly.



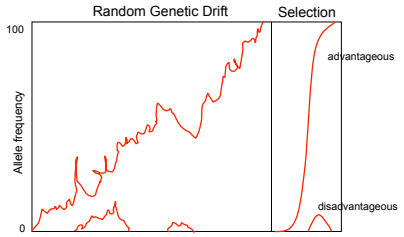
s>0

Time till fixation on average:
t_{av}= (2/s) ln (2N) generations
(also true for mutations with negative "s" ! discuss among yourselves)

E.g.: N=10^6,
s=0: average time to fixation: 4*10^6 generations
s=0.01: average time to fixation: 2900 generations

N=10^4,
s=0: average time to fixation: 40.000 generations
s=0.01: average time to fixation: 1.900 generations

=> substitution rate of mutation under positive selection is larger than the rate wite which neutral mutations are fixed.



Modified from from www.tcd.ie/Genetics/staff/Aoife/GE3026/GE3026_1+2.ppt

Positive selection (s>0)

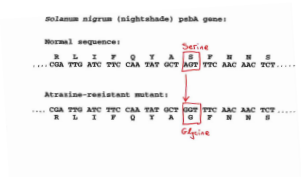
- A new allele (mutant) confers some increase in the **fitness** of the organism
- Selection acts to favour this allele
- Also called adaptive selection or Darwinian selection.

NOTE: **Fitness** = ability to survive and reproduce

Modified from from www.tcd.ie/Genetics/staff/Aoife/GE3026/GE3026_1+2.ppt

Advantageous allele

Herbicide resistance gene in nightshade plant



Modified from from www.tcd.ie/Genetics/staff/Aoife/GE3026/GE3026_1+2.ppt

Negative selection (s<0)

- A new allele (mutant) confers some decrease in the fitness of the organism
- Selection acts to remove this allele
- Also called purifying selection

Modified from from www.tcd.ie/Genetics/staff/Aoife/GE3026/GE3026_1+2.ppt

Neutral mutations

- Neither advantageous nor disadvantageous
- Invisible to selection (no selection)
- Frequency subject to 'drift' in the population
- **Random drift** – random changes in small populations

Types of Mutation-Substitution

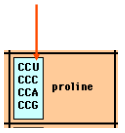
- Replacement of one nucleotide by another
- Synonymous (Doesn't change amino acid)
  - Rate sometimes indicated by Ks
  - Rate sometimes indicated by d_s
- Non-Synonymous (Changes Amino Acid)
  - Rate sometimes indicated by Ka
  - Rate sometimes indicated by d_n

(this and the following 4 slides are from mentor.lscf.ucsb.edu/course/ spring/eemb102/lecture/Lecture7.ppt)

Genetic Code – Note degeneracy of 1st vs 2nd vs 3rd position sites

UUU UUC	phenyl alanine	UCU UCC UCA UCG	serine	UAU UAC	tyrosine	UGU UGC	cysteine
UUA UUG	leucine			UAA UAG	stop	UGA	stop
						UGG	tryptophan
CUU CUC CUA CUG	leucine	CCU CCC CCA CCG	proline	CAU CAC	histidine	CGU CGC CGA CGG	arginine
				CAA CAG	glutamine		
AUU AUC AUA	isoleucine	ACU ACC ACA ACG	threonine	AAU AAC	asparagine	AGU AGC	serine
AUG	methionine			AAA AAG	lysine	AGA AGG	arginine
GUU GUC GUA GUG	valine	GCU GCC GCA GCG	alanine	GAU GAC	aspartic acid	GGU GGC GGA GGG	glycine
				GAA GAG	glutamic acid		

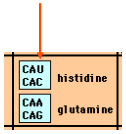
Genetic Code



Four-fold degenerate site – Any substitution is synonymous

From: mentor.lscf.ucsb.edu/course/spring/eemb102/lecture/Lecture7.ppt

Genetic Code



Two-fold degenerate site – Some substitutions synonymous, some non-synonymous

From: mentor.lscf.ucsb.edu/course/spring/eemb102/lecture/Lecture7.ppt

Genetic Code

Degeneracy of 1st vs 2nd vs 3rd position sites results in 25.5% synonymous changes and 74.5% non synonymous changes (Yang&Nielsen,1998).

First Position	Second Position U		Second Position C		Second Position A		Second Position G		Third Position
	code	Amino Acid	code	Amino Acid	code	Amino Acid	code	Amino Acid	
U	UUU	phe	UCU	ser	UAU	tyr	UGU	cys	U
	UUC		UCC		UAC		UGC		C
	UUA	leu	UCA		UAA	STOP	UGA	STOP	A
	UUG		UCG		UAG	STOP	UGG	trp	G
C	CUU	leu	CCU	pro	CAU	his	CGU	arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	ile	ACU	thr	AAU	asn	AGU	ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	lys	AGA	arg	A
	AUG	met	ACG		AAG		AGG		G
G	GUU	val	GCU	ala	GAU	asp	GGU	gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	glu	GGA		A
	GUG		GCG		GAG		GGG		G

Measuring Selection on Genes

- Null hypothesis = neutral evolution
- Under neutral evolution, synonymous changes should accumulate at a rate equal to mutation rate
- Under neutral evolution, amino acid substitutions should also accumulate at a rate equal to the mutation rate

From: mentor.lscf.ucsb.edu/course/spring/eemb102/lecture/Lecture7.ppt

Counting #s/#a

	Ser	Ser	Ser	Ser	Ser
Species1	TGA	TGC	TGT	TGT	TGT
	Ser	Ser	Ser	Ser	Ala
Species2	TGT	TGT	TGT	TGT	GGT

#s = 2 sites
#a = 1 site

#a/#s=0.5

To assess selection pressures one needs to calculate the rates (Ka, Ks), i.e. the occurring substitutions as a fraction of the possible syn. and nonsyn. substitutions.

Things get more complicated, if one wants to take transition transversion ratios and codon bias into account. See chapter 4 in Nei and Kumar, Molecular Evolution and Phylogenetics.

Modified from: mentor.lscf.ucsb.edu/course/spring/eemb102/lecture/Lecture7.ppt

Testing for selection using dN/dS ratio

dN/dS ratio (aka **Ka/Ks** or **ω** (omega) ratio) where

dN = number of non-synonymous substitutions / number of possible non-synonymous substitutions

dS =number of synonymous substitutions / number of possible non-synonymous substitutions

dN/dS >1 positive, Darwinian selection

dN/dS =1 neutral evolution

dN/dS <1 negative, purifying selection

PAML (codeml) the basic model

$$q_{ij} = \begin{cases} 0, & \text{if the two codons differ at more than one position,} \\ \pi_j, & \text{for synonymous transversion,} \\ \kappa\pi_j, & \text{for synonymous transition,} \\ \omega\pi_j, & \text{for nonsynonymous transversion,} \\ \omega\kappa\pi_j, & \text{for nonsynonymous transition,} \end{cases}$$

The equilibrium frequency of codon j (π_j) can be considered a free parameter, but can also be calculated from the nucleotide frequencies at the three codon positions (control variable `CodonFreq`). Under this model, the relationship holds that $\omega = d_N/d_S$, the ratio of nonsynonymous/synonymous substitution rates. This basic model is fitted by specifying model = 0 NSsites = 0, in the control file codeml.ctl. It forms the basis for more sophisticated models implemented in codeml.

dambe

Three programs worked well for me to align nucleotide sequences based on the amino acid alignment,

One is DAMBE (works well for windows). This is a handy program for a lot of things, including reading a lot of different formats, calculating phylogenies, it even runs codeml (from PAML) for you.

The procedure is not straight forward, but is well described on the help pages. After installing DAMBE go to HELP -> general HELP -> sequences -> align nucleotide sequences based on …-->

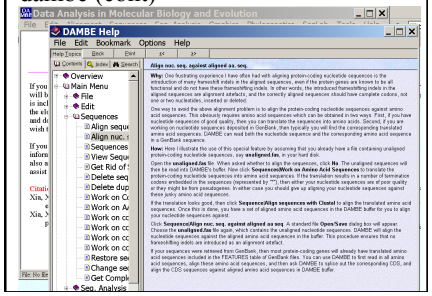
If you follow the instructions to the letter, it works fine.

DAMBE also calculates Ka and Ks distances from codon based aligned sequences.

Alternatives are

- tranalign from the EMBOSS package, and
- Seaview (see below)

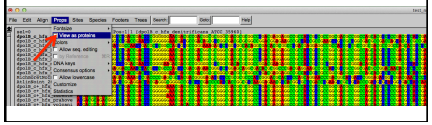
dambe (cont)



Codon based alignments in Seaview

Load nucleotide sequences (**no gaps in sequences, sequence starts with nucleotide corresponding to 1st codon position**)

Select view as proteins



Codon based alignments in Seaview

With the protein sequences displayed, align sequences

Select view as nucleotides

PAML (codeml) the basic model

$$q_{ij} = \begin{cases} 0, & \text{if the two codons differ at more than one position,} \\ \pi_j, & \text{for synonymous transversion,} \\ \kappa\pi_j, & \text{for synonymous transition,} \\ \omega\pi_j, & \text{for nonsynonymous transversion,} \\ \omega\kappa\pi_j, & \text{for nonsynonymous transition,} \end{cases}$$

The equilibrium frequency of codon j (π_j) can be considered a free parameter, but can also be calculated from the nucleotide frequencies at the three codon positions (control variable `CodonFreq`). Under this model, the relationship holds that $\omega = d_N/d_S$, the ratio of nonsynonymous/synonymous substitution rates. This basic model is fitted by specifying model = 0 NSsites = 0, in the control file codeml.ctl. It forms the basis for more sophisticated models implemented in codeml.

sites versus branches

You can determine omega for the whole dataset; however, usually not all sites in a sequence are under selection all the time.

PAML (and other programs) allow to either determine omega for each site over the whole tree, *Branch Models* **, or determine omega for each branch for the whole sequence,** *Site Models* **.**

It would be great to do both, i.e., conclude codon 176 in the vacuolar ATPases was under positive selection during the evolution of modern humans – alas, a single site does not provide much statistics ….

Sites model(s)

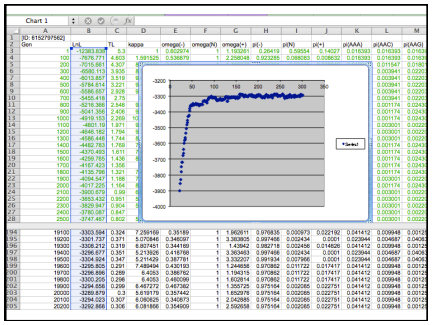
work great have been shown to work great in few instances. The most celebrated case is the influenza virus HA gene.

A talk by Walter Fitch (slides and sound) on the evolution of this molecule is here .
This article by Yang et al. 2000 gives more background on ml aproaches to measure omega. The dataset used by Yang et al is here: flu_data.paup .

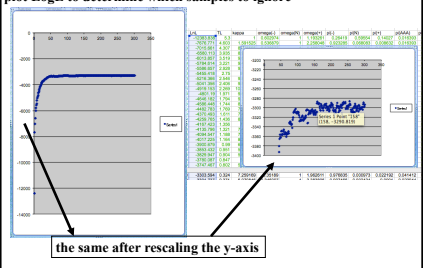
sites model in MrBayes

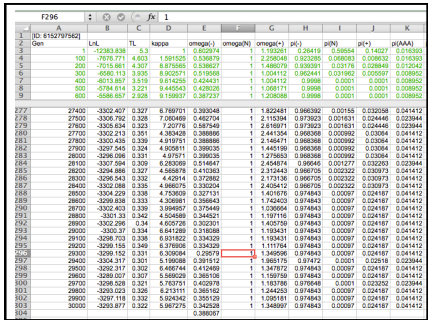
The MrBayes block in a nexus file might look something like this:

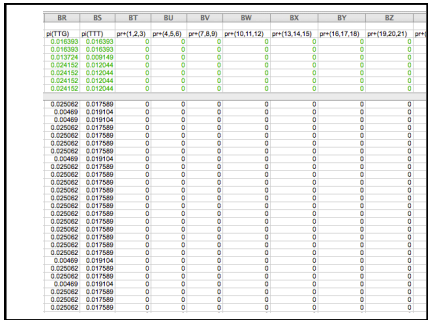
```
begin mrbayes;
set autoclose=yes;
lset nst=2 rates=gamma nucmodel=codon omegavar=Ny98;
mcmcp samplefreq=500 printfreq=500;
mcmc ngen=500000;
sump burnin=50;
sumt burnin=50;
end;
```



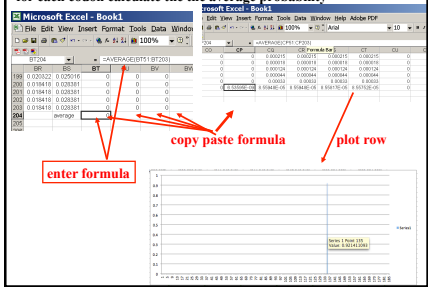
plot LogL to determine which samples to ignore



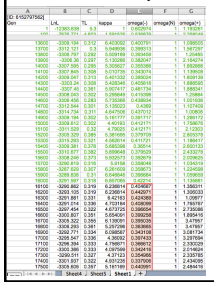




for each codon calculate the the average probability

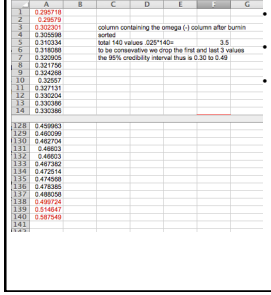


To determine credibility interval for a parameter (here omega<1):



Select values for the parameter, sampled after the burning.

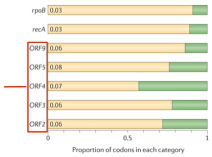
Copy paste to a new spreadsheet,



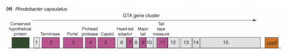
- **Sort values according to size,**
- **Discard top and bottom 2.5%**
- **Remainder gives 95% credibility interval.**

Purifying selection in GTA genes

dN/dS <1 for GTA genes has been used to infer selection for function



Lang AS, Zhaxybayeva O, Beatty JT. Nat Rev Microbiol. 2012 Jun 11;10(7):472-82



Lang, A.S. & Beatty, J.T. Trends in Microbiology , Vol.15, No.2 , 2006

Purifying selection in *E.coli* ORFans

dN-dS < 0 for some ORFan *E. coli* clusters seems to suggest they are functional genes.

Gene groups	Number	dN-dS>0	dN-dS<0	dN-dS=0
E. coli ORFan clusters	3773	944 (25%)	1953 (52%)	876 (23%)
Clusters of *E.coli* sequences found in *Salmonella sp., Citrobacter sp.*	610	104 (17%)	423(69%)	83 (14%)
Clusters of *E.coli* sequences found in some *Enterobacteriaceae* only	373	8 (2%)	365 (98%)	0 (0%)

Adapted after Yu, G. and Stoltzfus, A. Genome Biol Evol (2012) Vol. 4 1176-1187

Vincent Daubin and Howard Ochman: Bacterial Genomes as New Gene Homes: The Genealogy of ORFans in *E. coli*. *Genome Research 14:1036-1042, 2004*

The ratio of non-synonymous to synonymous substitutions for genes found only in the E.coli - Salmonella clade is lower than 1, but larger than for more widely distributed genes.

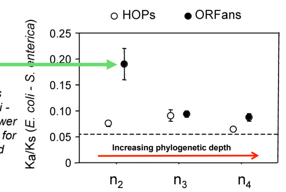


Fig. 3 from Vincent Daubin and Howard Ochman, *Genome Research 14:1036-1042, 2004*

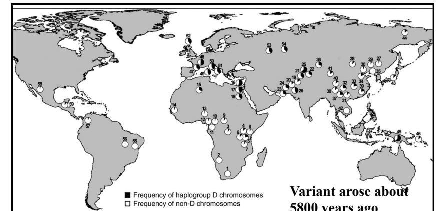
Trunk-of-my-car analogy: Hardly anything in there is the is the result of providing a selective advantage. Some items are removed quickly (purifying selection), some are useful under some conditions, but most things do not alter the fitness.

Could some of the inferred purifying selection be due to the acquisition of novel detrimental characteristics (e.g., protein toxicity, HOPELESS MONSTERS)?

Other ways to detect positive selection

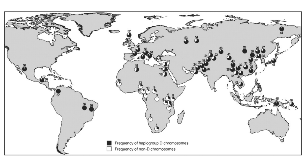
Selective sweeps -> fewer alleles present in population (see contributions from archaic Humans for example)

Repeated episodes of positive selection -> high dN



Each population's geographic origin, number of individuals, and frequency of haplogroup D chromosomes are given in parentheses as follows:

Fig. 3. Global frequencies of Microcephalin haplogroup D chromosomes (defined as having the derived C allele at the G37995C diagnostic SNP) in a panel of 1184 individuals



P. D. Evans et al., Science 309, 1717 -1720 (2005)

The age of haplogroup D was found to be ~37,000 years

Science AAAS

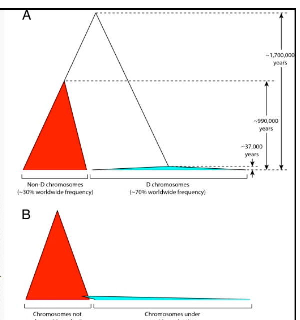
Published by AAAS

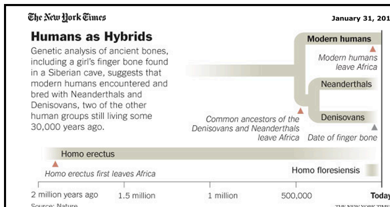
Proc Natl Acad Sci U S A. 2006 November 28; 103(48): 18178–18183. PMCID: PMC1635020
Published online 2006 November 7. doi: 10.1073/pnas.0606966103.

Evolution

Evidence that the adaptive allele of the brain size gene *microcephalin* introgressed into *Homo sapiens* from an archaic *Homo* lineage

Patrick D. Evans,*†‡ Nitzan Mekel-Bobrov,*†‡ Eric J. Vallender,*†‡ Richard R. Hudson,§ and Bruce T. Lahn*†¶

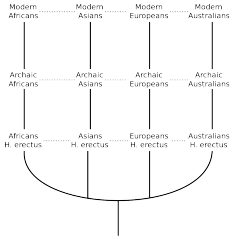




Close Window

From:
http://www.nytimes.com/2012/01/31/science/gains-in-dna-are-speeding-research-into-human-origins.html?_r=1

The multiregional hypothesis



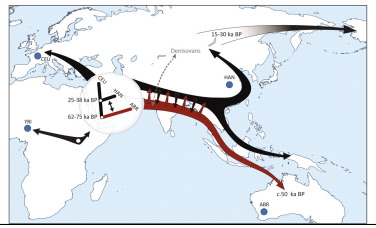
From http://en.wikipedia.org/wiki/Multiregional_Evolution

Published Online September 22 2011
Science 7 October 2011:
Vol. 334 no. 6052 pp. 94–98
DOI: 10.1126/science.1211177

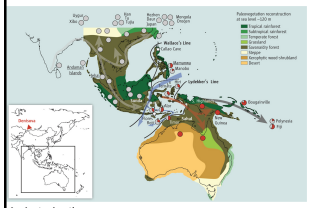
REPORT

An Aboriginal Australian Genome Reveals Separate Human Dispersals into Asia

Morten Rasmussen[1,2,3], Xiaosen Guo[1,2,3], Yong Wang[2,3], Kirk E. Lohmueller[2,3], Simon Rasmussen[3],

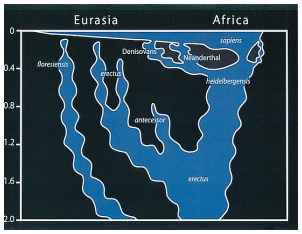


Did the Denisovans Cross Wallace's Line?
Science 18 October 2013:
vol. 342 no. 6156 321-323



Ancient migrations.
The proportions of Denisovan DNA in modern human populations are shown as red in pie charts, **relative to New Guinea and Australian Aborigines** (3). Wallace's Line (8) is formed by the powerful Indonesian flow-through current (blue arrows) and marks the limit of the Sunda shelf and Eurasian placental mammals.

Archaic human admixture with modern Homo sapiens



From: http://en.wikipedia.org/wiki/Archaic_human_admixture_with_modern_Homo_sapiens

For more discussion on archaic and early humans see:
http://en.wikipedia.org/wiki/Denisova_hominin

http://www.nytimes.com/2012/01/31/science/gains-in-dna-are-speeding-research-into-human-origins.html

http://www.sciencedirect.com/science/article/pii/S0002929711003958
http://www.abc.net.au/science/articles/2012/08/31/3580500.htm

http://www.sciencemag.org/content/334/6052/94.full
http://www.sciencemag.org/content/334/6052/94/F2.expansion.html

http://haplogroup-a.com/Ancient-Root-AJHG2013.pdf