Finding Selection in All the Right Places
TA Notes and Key
Lab 9

Objectives:

1. Use published genome data to look for evidence of selection in individual genes.
2. Understand the need for DNA sequence alignment before sequence analysis.
3. Understand the use of the McDonald-Kreitman test.
4. Form hypotheses for the observed presence or absence of selection on individual genes.

Order of activities and answers to questions:

1. Quiz
2. have them work through the alignment exercise
3. while students work on alignment exercise, transfer genes to their flash drives (5 per pair) – works best if you move the files to a subfolder (called "used" or similar) simultaneously, so you don't give the same sequence to multiple pairs
4. discuss answers to alignment exercise
5. best to work through one gene with them (I'll give you extra so you can go through a new one yourself, plus have replacements for the few that end up being pseudogenes or otherwise problematic), go all the way through both MK tests and the data entry. Be sure to use the N-count.xls spreadsheet so the students see how to use it.
6. turn them loose to finish their other 4 genes on their own
7. it's best if they do all 4 alignments in Mega first, then do all 4 sets of MK tests (because they need your help the most for the alignment portion, and, since the MK tests are just a website, it's very easy to complete that part on their own time)
8. they can do MK test and data entry simultaneously, but they still need the gene pages for the first information, since they'll be finishing through the alignment for all 4 genes before moving on

Key to questions with known answers:

Now try it for these sequence reads:

CTATCTCCCACGAGGATACT
CTAAAGGACAAAAATATTCT
ATGGCTCTAGCTATCTCCCA
CAGATTTTGCTAAAGGACAA
CGAGGATACTCAGATTTTGC

** What is the consensus sequence you built from these reads?

ATGGCTCTAGCTATCTCCCACGAGGATACTCAGATTTTGCTAAAGGACAAAAATATTCT

** Consider the 5 sequences below. How many polymorphic sites (base pair positions that are not the same nucleotide for all 5 samples) are there? Hint: make a mark underneath each polymorphic site and then count your marks.

A	A	A	A	A	G	G	A	G	T	C	A	A	C	T	A	T	T	G	G	A	A	A	T	T	C	T	G	C	A	T							
A	A	G	A	G	T	C	A	A	C	T	A	T	T	G	G	A	A	A	T	T	C	T	G	C	A	T											
A	T	A	A	A	G	G	A	G	T	C	A	A	C	T	A	T	T	G	G	A	A	A	T	T	C	T	G	C	A	T							
A	A	A	A	A	G	G	A	G	T	C	A	A	C	T	A	T	T	G	G	A	A	A	T	T	G	G	A	A	A	T	T	C	T	G	C	A	T
A	A	A	A	A	G	G	A	G	T	C	A	A	G	T	A	T	T	G	G	A	A	A	T	T	C	T	G	C	A	T							

This can be easier to see if each nucleotide is highlighted with a different color (which you can do with the colored highlighters on your lab bench).

A	A	A	A	A	G	G	A	G	T	C	A	A	C	T	A	T	T	G	G	A	A	A	T	T	C	T	G	C	A	T							
A	A	G	A	G	T	C	A	A	C	T	A	T	T	G	G	A	A	A	T	T	C	T	G	C	A	T											
A	T	A	A	A	G	G	A	G	T	C	A	A	C	T	A	T	T	G	G	A	A	A	T	T	C	T	G	C	A	T							
A	A	A	A	A	G	G	A	G	T	C	A	A	C	T	A	T	T	G	G	A	A	A	T	T	G	G	A	A	A	T	T	C	T	G	C	A	T
A	A	A	A	A	G	G	A	G	T	C	A	A	G	T	A	T	T	G	G	A	A	A	T	T	C	T	G	C	A	T							
	*	*		*	*	*		*	*	*		*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*

=34

** Look at the color-coded sequences more carefully, paying close attention to the patterns made by the colors. Pay particular attention to the sequences that are different lengths than the others. Do you notice anything about the color pattern? Could you shift the sequences in any way to make the pattern better? What do you think caused these changes?

If sequence 2 was shifted to the right 4 bases and either 7 bases were removed from the center of sequence 4 or all the other sequences were shifted over 7 bases, then the patterns would match up much better. The change in sequence 2 was probably the result of a deletion mutation and the change in sequence 4 was probably the result of an insertion mutation.

A	A	A	A	A	G	G	A	G	T	C	A	A	C	T	A	T	T	G	G	A	A	A	T	T	C	T	G	C	A	T							
A	A	G	A	G	T	C	A	A	C	T	A	T	T	G	G	A	A	A	T	T	C	T	G	C	A	T											
A	T	A	A	A	G	G	A	G	T	C	A	A	C	T	A	T	T	G	G	A	A	A	T	T	C	T	G	C	A	T							
A	A	A	A	A	G	G	A	G	T	C	A	A	C	T	A	T	T	G	G	A	A	A	T	T	G	G	A	A	A	T	T	C	T	G	C	A	T
A	A	A	A	A	G	G	A	G	T	C	A	A	G	T	A	T	T	G	G	A	A	A	T	T	C	T	G	C	A	T							

** Redraw the 5 sequences in the table below making the changes you suggested above. Put one nucleotide in each box of the table. Don't worry if you don't fill the table completely, at this point there's more than one way to answer the question.

A	A	A	A	A	G	G	A	G	T	C	A	A	C	T	A	T	T	G	G	A	A	A	T	T	-	-	-	-	-	-	-	C	T	G	C	A	T
A	A	-	-	-	-	G	A	G	T	C	A	A	C	T	A	T	T	G	G	A	A	A	T	T	-	-	-	-	-	-	-	C	T	G	C	A	T
A	T	A	A	A	G	G	A	G	T	C	A	A	C	T	A	T	T	G	G	A	A	A	T	T	-	-	-	-	-	-	-	C	T	G	C	A	T
A	A	A	A	A	G	G	A	G	T	C	A	A	C	T	A	T	T	G	G	A	A	A	T	T	G	G	A	A	A	T	T	C	T	G	C	A	T
A	A	A	A	A	G	G	A	G	T	C	A	A	G	T	A	T	T	G	G	A	A	A	T	T	-	-	-	-	-	-	-	C	T	G	C	A	T
	*	*	*	*	*								*												*	*	*	*	*	*	*						13

Also OK if they just deleted the insertion in sequence 4, but then point out that they lost that polymorphism and ask which way (deleting insertion or adding spaces in other sequences) gives a more accurate representation of the differences between the sequences.

** Highlight the different bases in the table above different colors using the highlighters provided. It doesn't matter which color you use for which nucleotide. How many polymorphic sites are there now?

13

** Of the two counts of polymorphic sites that you made, which gives the more realistic representation of the differences between these sequences?

The second is probably more accurate because the chance that the overall pattern would be recreated by independent base changes is much lower than the chance of the one insertion and one deletion.

- **Familiarize yourself with the Mega window.**

** What do the asterices along the top row indicate?

Base positions in which all the strains have the same base (conserved bases)

** What does a deletion in only one strain look like?

A deletion in only one strain will be a "–" in a white space while the other strains have nucleotides in colored spaces.

** How about an insertion in only one strain?

An insertion in only one strain will be a nucleotide in that one strain while all the other strains have dashes in white spaces.

** Arginine is one of the amino acids with the most codons encoding it, while tryptophan is one of the amino acids with the least. Using a codon table, determine how many single base pair mutations to the arginine codon CGA (written in RNA, but the answer would be the same for mutations to the DNA, right?) would be synonymous and how many would be nonsynonymous. Do the same for the tryptophan codon UGG. So what can you say about the possible number of synonymous and nonsynonymous substitutions for all codons?

Arginine: 4 synonymous, 5 nonsynonymous

Tryptophan: 0 synonymous, 9 nonsynonymous

There will always be more possible nonsynonymous than synonymous substitution mutations

** There are three possible relationships that could be seen in the Contingency Table (think about comparing the nonsynonymous/synonymous ratios of between and within species). What are they and what does each one mean?

Ratios equal between and within species and more synonymous than nonsynonymous -> exactly what is predicted by the neutral theory

Ratios not equal, between species > within species -> evidence that positive selection is driving the between species differences

Ratios not equal, between species < within species -> either purifying selection so the strains have low frequency within species polymorphisms that were never fixed between species, or balancing selection is maintaining polymorphism within one or both species

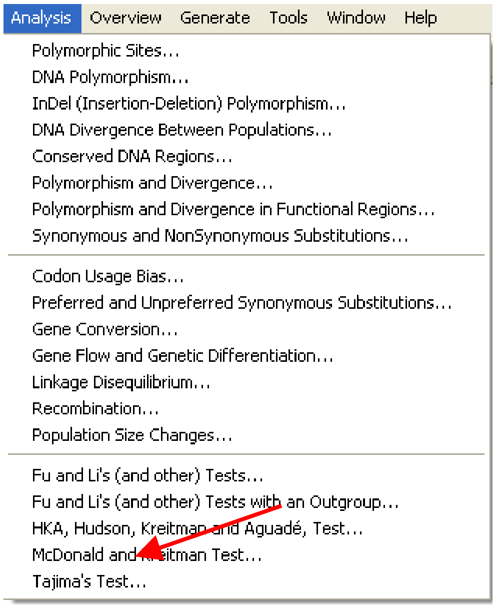
Alternate McDonald-Kreitman Test instructions, in case http://mkt.uab.es is down

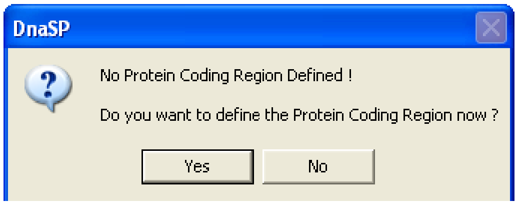
NOTE: requires Windows and the free software DnaSP (http://www.ub.edu/dnasp/)

- **start the DnaSP program** by clicking on the icon in the Taskbar.
- **Open your aligned .fas file** by clicking the Open button and navigating to your file.

  DnaSP should see it, but if it doesn't, make sure that "Files of type" at the bottom is set to "All Data Files" or "FASTA files."
- You can **click Close** on the Data Information window that pops up.

- **Go to Analysis -> McDonald and Kreitman Test** and a popup will appear asking if you want to define a protein coding region – **click Yes**.


Analysis
Overview
Generate
Tools
Window
Help
Polymorphic Sites...
DNA Polymorphism...
InDel (Insertion-Deletion) Polymorphism...
DNA Divergence Between Populations...
Conserved DNA Regions...
Polymorphism and Divergence...
Polymorphism and Divergence in Functional Regions...
Synonymous and NonSynonymous Substitutions...
Codon Usage Bias...
Preferred and Unpreferred Synonymous Substitutions...
Gene Conversion...
Gene Flow and Genetic Differentiation...
Linkage Disequilibrium...
Recombination...
Population Size Changes...
Fu and Li's (and other) Tests...
Fu and Li's (and other) Tests with an Outgroup...
HKA, Hudson, Kreitman and Aguadé, Test...
McDonald and Kreitman Test...
Tajima's Test...



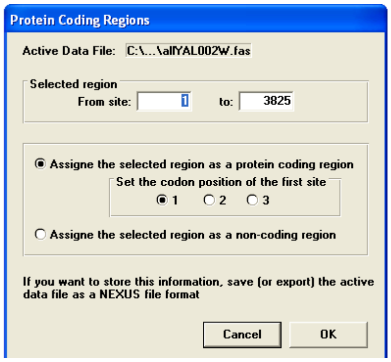
No Protein Coding Region Defined !

Do you want to define the Protein Coding Region now ?

- You can **either click OK** for the default values that appear (because the default is the entire sequence with the reading frame starting with the first base, which is true of your data)

  or you can subtract 3 from the ending site because DnaSP does not consider STOP codons to be coding sequence.

  If you don't subtract the last 3 bases, then DnaSP will ask if you want it to do that for you.



Protein Coding Regions

Active Data File: C:\...\allYAL002W.fas

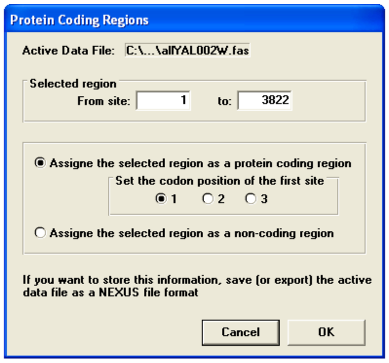
Selected region
From site: 1 to: 3825

● Assigne the selected region as a protein coding region
Set the codon position of the first site
● 1 ○ 2 ○ 3

○ Assigne the selected region as a non-coding region

If you want to store this information, save (or export) the active data file as a NEXUS file format

OR



Protein Coding Regions

Active Data File: C:\...\allYAL002W.fas

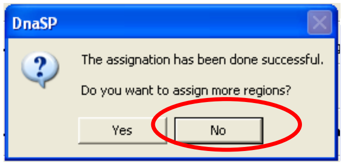
Selected region
From site: 1 to: 3822

● Assigne the selected region as a protein coding region
Set the codon position of the first site
● 1 ○ 2 ○ 3

○ Assigne the selected region as a non-coding region

If you want to store this information, save (or export) the active data file as a NEXUS file format

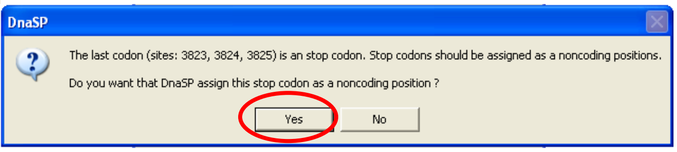
- A popup will ask if you want to define more coding regions – **click No**.
- If you didn't subtract 3 from the ending site, you will now be asked if you want DnaSP to assign the STOP codon as a noncoding position, **click Yes**.



DnaSP

The assignation has been done successful.

Do you want to assign more regions?



DnaSP

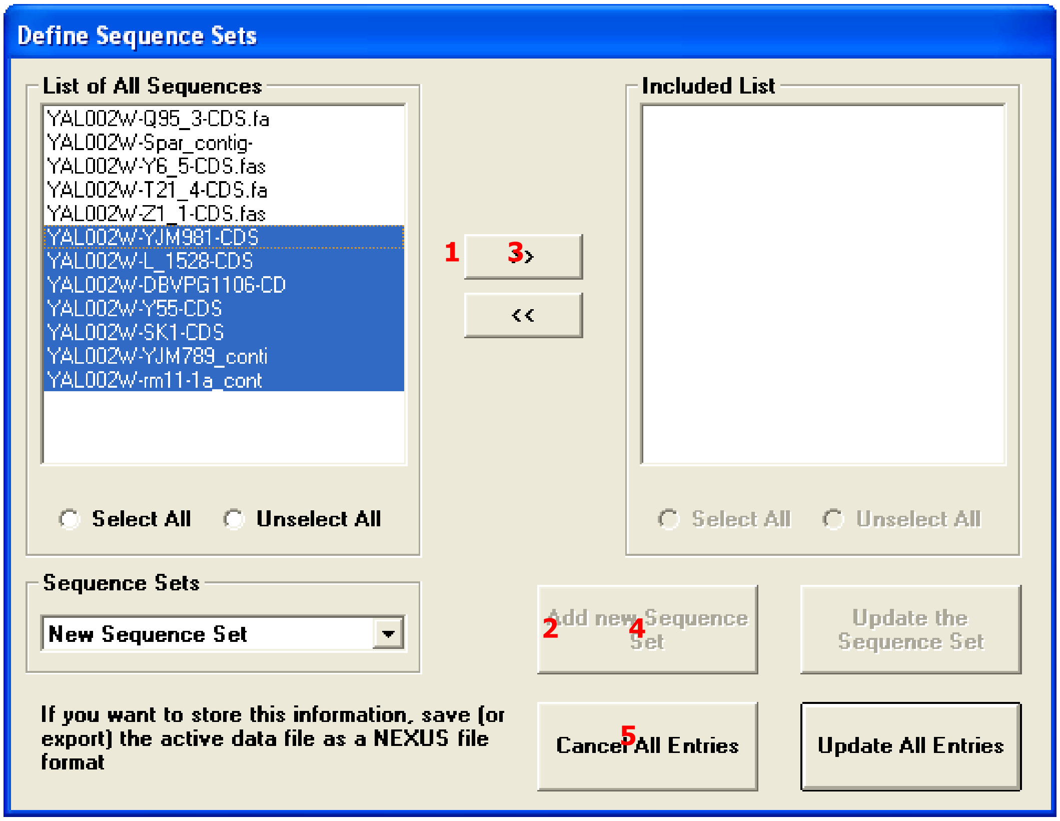
The last codon (sites: 3823, 3824, 3825) is an stop codon. Stop codons should be assigned as a noncoding positions.

Do you want that DnaSP assign this stop codon as a noncoding position ?

3- **Define sequence sets.**

For the following instructions, the number in parentheses refers to the red number in the image.

- A popup will then prompt you to define sequence sets (sequences for each species) – **click Yes**.
- Now you need to know which strains belonged to each species. **Select all of the *D. pseudoobscura* strains and click on >> (1)**.
- **Click the "Add new Sequence Set" button (2) and name that set Dpseudoobscura.**



Define Sequence Sets

List of All Sequences

YAL002W-Q95_3-CDS.fa
YAL002W-Spar_contig-
YAL002W-Y6_5-CDS.fas
YAL002W-T21_4-CDS.fa
YAL002W-Z1_1-CDS.fas
YAL002W-YJM981-CDS
YAL002W-L_1528-CDS
YAL002W-DBVPG1106-CD
YAL002W-Y55-CDS
YAL002W-SK1-CDS
YAL002W-YJM789_conti
YAL002W-rm11-1a_cont

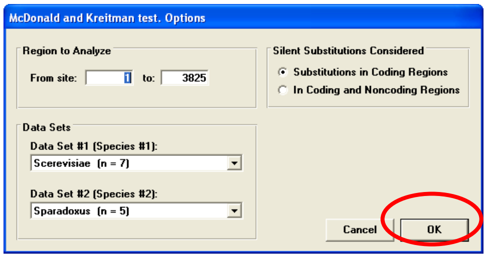
Included List

Sequence Sets

New Sequence Set

If you want to store this information, save (or export) the active data file as a NEXUS file format

- Then **select all the *D. miranda* strains and click >> (3).**
- **Click "Add new Sequence Set" (4) again, and name that set Dmiranda**.
- **Click the "Update All Entries" button (5).** Make sure everything in the popup looks right, then **click OK**.



McDonald and Kreitman test. Options

Region to Analyze

From site: 1 to: 3825

Silent Substitutions Considered

● Substitutions in Coding Regions
○ In Coding and Noncoding Regions

Data Sets

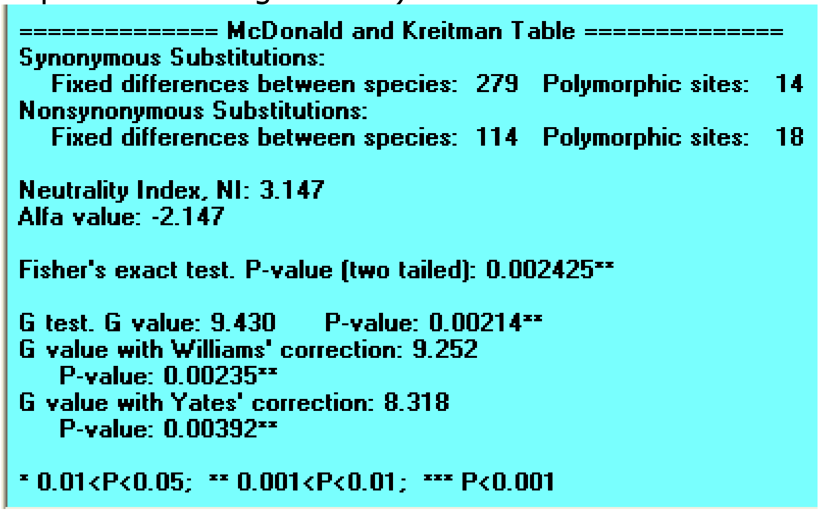
Data Set #1 (Species #1): Scerevisiae (n = 7)

Data Set #2 (Species #2): Sparadoxus (n = 5)

4- **Record your McDonald-Kreitman results.**
The most important information is at the bottom of the output screen. An example is shown below.

The McDonald and Kreitman Table summarizes the values used in the McDonald-Kreitman test. Fixed differences between species are sites at which all sequences in one species contain nucleotide variants that are not in the second species. Polymorphic sites are those that are variable within a single species.

- **Use the results of the Fisher's exact test** to determine whether the MK result is significant (DnaSP will tell you explicitly if it is not significant, it uses P<0.1 as the requirement for significance).



```
============== McDonald and Kreitman Table ==============
Synonymous Substitutions:
   Fixed differences between species:  279   Polymorphic sites:   14
Nonsynonymous Substitutions:
   Fixed differences between species:  114   Polymorphic sites:   18

Neutrality Index, NI: 3.147
Alfa value: -2.147

Fisher's exact test. P-value (two tailed): 0.002425**

G test. G value: 9.430      P-value: 0.00214**
G value with Williams' correction: 9.252
    P-value: 0.00235**
G value with Yates' correction: 8.318
    P-value: 0.00392**

* 0.01<P<0.05;  ** 0.001<P<0.01;  *** P<0.001
```