UNIVERSITY OF TORONTO

1

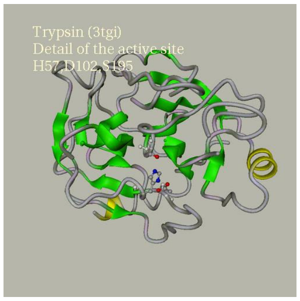
Selection Analysis

- attempts to identify if natural selection is acting on DNA sequences
- if natural selection is indeed acting, is the selection “positive” or “negative”, i.e. is there selection for or against a site?
- addresses how sequences are changing because of natural selection

2

Natural selection

- Positive ("Darwinian") selection occurs when a beneficial mutation arises in a population and increases in frequency (e.g. antibiotic resistance)
- Negative (purifying) selection is the obverse of positive selection. It occurs when a detrimental mutation is selected out of a population.
- Balancing or diversifying selection is selection that favours the maintenance of genetic variation at a locus → multiple environments select for different allelic forms of a protein.



from the Protein Picture Generator at http://bioserv.rpbs.jussieu.fr

Measuring selection

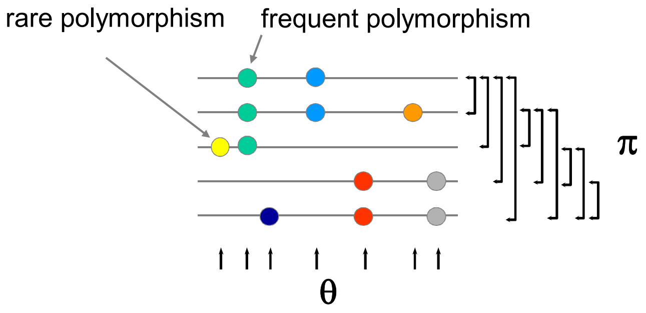
There are two commonly used measures: **Tajima's *D*** and the **dN/dS** test

There are many ways to quantify genetic variation. Population geneticists primarily focus on two metrics:

- *Theta* (θ) is based on the number of segregating / variable sites in the sample.
- Pi (π) is based on the average number of differences between all pairwise combinations of sequences → pairwise nucleotide diversity.

Both θ and π measure variation, but π is much more sensitive to the frequency of genetic variants. θ considers all variants equal, regardless of whether they are found in only a single sequence, or half of all sequences, whereas π takes these differences into account.

Measuring selection



Number of segregating sites = θ

Nucleotide diversity = π

Tajima's *D*

- Natural selection acts directly on specific mutations that change fitness, but *genetic regions that surround the selected site are also influenced*
- Genetic regions surrounding a site get "dragged" along due to a process called "genetic hitchhiking"
- The length of the genetic region is dependent on the rate of recombination
- Most of the genetic variation in surrounding regions is neutral (e.g. 3rd position substitutions that don't change the coding sequence), nevertheless, it can still show signs of selection due to its linkage to a selected polymorphism
- Tajima's *D* looks at θ and π and asks which one predominates...

How θ and π are used in calculating Tajima's *D* to assess selection

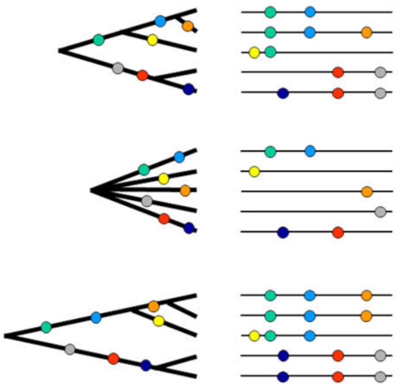
- In the case of a neutral gene, where no selection is acting,
  θ in this case is roughly equal to π.

- Positive selection effectively sweeps genetic variation from a population – a process called a selective sweep. Very little genetic variation will remain in a population immediately after the spread of a beneficial mutation through a population, but then new genetic variation will begin to accumulate. Since all this new variation is effectively accumulating in a genetically homogeneous background, most mutations will be at low frequency.
  $\therefore$ θ in this case is greater than π.

- Balancing selection retains alleles (genetic variants) in the population longer than would be expected. Since alleles are kept around longer than expected, they will rise to intermediate frequencies.
  $\therefore$ θ in this case is less than π.

7

How θ and π are used in calculating Tajima's *D* to assess selection



$\theta \approx \pi$
neutral

$D \approx 0$

$\theta > \pi$
positive selection
(selective sweep)

$D < 0$

$\theta < \pi$
balancing selection

$D > 0$

8

Tajima’s *D*: caveats

- Tajima’s *D* is a powerful statistic for detecting selection, but it can be fooled by other factors
- Two of the main factors are if the population you are looking at has gone through a “bottleneck” – in this case positive selection at a locus would be masked – or if the locus is in a region of low recombination
- You should not use this test without critically examining your data, e.g. looking at population substructure with other tools...

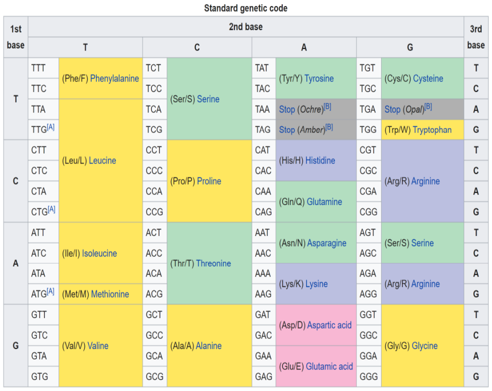
9

dN/dS ratio test for selection

- The dN/dS (also known as Ka/Ks) Ratio Test is perhaps the most widely used method for detecting the pattern of natural selection from nucleotide sequence data
- This test is particularly useful because it can infer selection acting all the way down to the level of the codon → are there specific sites that are being selected for?

10

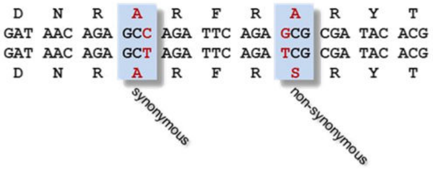
Non-synonymous versus synonymous nucleotide changes



Standard genetic code

1st base	2nd base								3rd base
	T		C		A		G		
T	TTT	(Phe/F) Phenylalanine	TCT	(Ser/S) Serine	TAT	(Tyr/Y) Tyrosine	TGT	(Cys/C) Cysteine	T
	TTC		TCC		TAC		TGC		C
	TTA	(Leu/L) Leucine	TCA		TAA	Stop (*Ochre*)[B]	TGA	Stop (*Opal*)[B]	A
	TTG[A]		TCG		TAG	Stop (*Amber*)[B]	TGG	(Trp/W) Tryptophan	G
C	CTT		CCT	(Pro/P) Proline	CAT	(His/H) Histidine	CGT	(Arg/R) Arginine	T
	CTC		CCC		CAC		CGC		C
	CTA		CCA		CAA	(Gln/Q) Glutamine	CGA		A
	CTG[A]		CCG		CAG		CGG		G
A	ATT	(Ile/I) Isoleucine	ACT	(Thr/T) Threonine	AAT	(Asn/N) Asparagine	AGT	(Ser/S) Serine	T
	ATC		ACC		AAC		AGC		C
	ATA		ACA		AAA	(Lys/K) Lysine	AGA	(Arg/R) Arginine	A
	ATG[A]	(Met/M) Methionine	ACG		AAG		AGG		G
G	GTT	(Val/V) Valine	GCT	(Ala/A) Alanine	GAT	(Asp/D) Aspartic acid	GGT	(Gly/G) Glycine	T
	GTC		GCC		GAC		GGC		C
	GTA		GCA		GAA	(Glu/E) Glutamic acid	GGA		A
	GTG		GCG		GAG		GGG		G

From https://en.wikipedia.org/wiki/DNA_codon_table



```
 D   N   R   A   R   F   R   A   R   Y   T
GAT AAC AGA GCC AGA TTC AGA GCG CGA TAC ACG
GAT AAC AGA GCT AGA TTC AGA TCG CGA TAC ACG
 D   N   R   A   R   F   R   S   R   Y   T
```

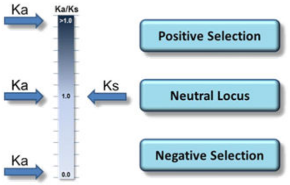
- Non-synonymous substitutions result in a change in the protein sequence
- Synonymous substitutions change the DNA sequence, but not the protein sequence due to the degeneracy of the genetic code

The dN/dS ratio test

- calculates the ratio of the rate of non-synonymous substitutions (dN, the number of ***non-synonymous*** substitutions per non-synonymous site) to the rate of synonymous substitutions (dS, the number of ***synonymous*** substitutions per synonymous site).
- Synonymous substitutions are not exposed to (strong) selective pressures since they don't result in a change to the protein sequence ∴ they tend to accumulate at roughly a constant rate → baseline to compare the rate of substitutions that change the protein sequence, i.e. non-synonymous substitutions.
- *What is critical to have for this kind of analysis?*

Interpretation of dN/dS ratio

- In the case of a completely neutral sequence (one that is free to change with no constraints), you would expect dN to be the same as dS, or dN/dS = 1
- When there are selective constraints on a sequence (negative selection), you would expect fewer substitutions that change the protein sequence, or a lower dN ∴ dN/dS < 1
- In the case of positive selection, you would expect to see a higher proportion of amino acid substitutions in your population (because they are being increased by positive selection), so a higher dN ∴ dN/dS > 1
- Most functional genes are under some level of selection constraint ∴ dN/dS ratios are typically well below 1.0 for most coding regions. Certain <u>sites</u> *may* be under positive selection, however.



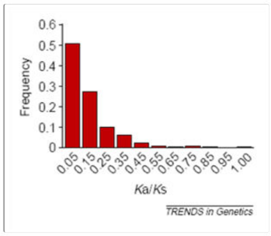
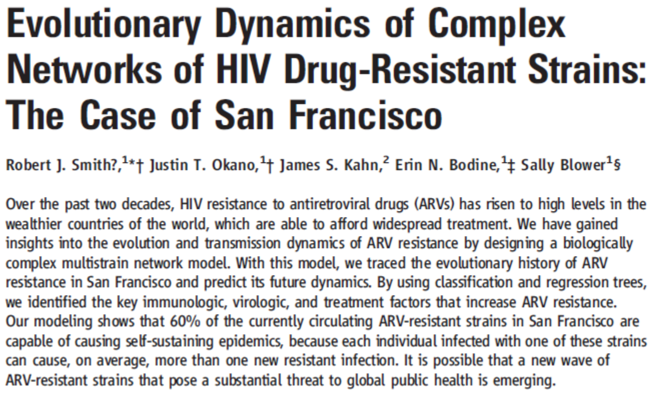


Fig. 1. The frequency of different values of *Ka*/*Ks* for 835 mouse–rat orthologous genes. Figures on the *x* axis represent the middle figure of each bin; that is, the 0.05 bin collects data from 0 to 0.1

from Hurst (2002) TIG 18:466

13

Evolutionary arms race…



Evolutionary Dynamics of Complex Networks of HIV Drug-Resistant Strains: The Case of San Francisco

Robert J. Smith?,[1*†] Justin T. Okano,[1†] James S. Kahn,[2] Erin N. Bodine,[1‡] Sally Blower[1§]

Over the past two decades, HIV resistance to antiretroviral drugs (ARVs) has risen to high levels in the wealthier countries of the world, which are able to afford widespread treatment. We have gained insights into the evolution and transmission dynamics of ARV resistance by designing a biologically complex multistrain network model. With this model, we traced the evolutionary history of ARV resistance in San Francisco and predict its future dynamics. By using classification and regression trees, we identified the key immunologic, virologic, and treatment factors that increase ARV resistance. Our modeling shows that 60% of the currently circulating ARV-resistant strains in San Francisco are capable of causing self-sustaining epidemics, because each individual infected with one of these strains can cause, on average, more than one new resistant infection. It is possible that a new wave of ARV-resistant strains that pose a substantial threat to global public health is emerging.

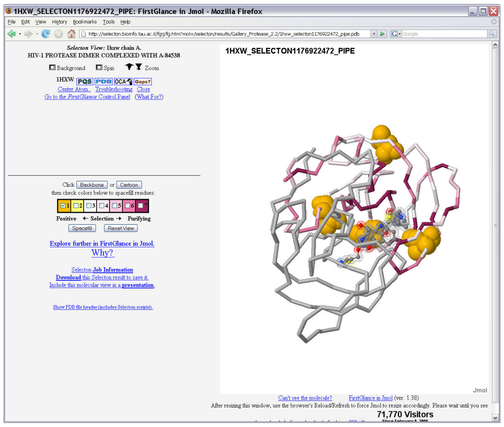
Smith et al. (2010) Science 327:697-701; DOI:10.1126/science.1180556

14

Site-specific dN/dS analysis

SELECTON Server
http://selecton.tau.ac.il/

- Exampe with HIV-1 protease
- HIV-1 protease is an essential enzyme for viral replication and has been a target for drug development, e.g. Ritonavir
- Seventy HIV-1 protease gene sequences from Ritonavir-treated HIV-1 patients were extracted from the Stanford HIV drug resistance Database (hivdb.stanford.edu/).
- These sequences, together with the PDB structure of the protease dimer bound to Ritonavir, were fed into the SELECTON server


1HXW_SELECTON1176922472_PIPE: FirstGlance in Jmol - Mozilla Firefox

http://selecton.bioinfo.tau.ac.il/fgij/fg.htm?mol=/selecton/results/Gallery_Protease_2.2/1hxw_selecton1176922472_pipe.pdb

Selecton View: 1hxw chain A.
HIV-1 PROTEASE DIMER COMPLEXED WITH A-84538

Click Backbone or Cartoon then check colors below to spacefill residues:

After resizing this window, use the browser's Reload/Refresh to force Jmol to resize accordingly. Please wait until you see

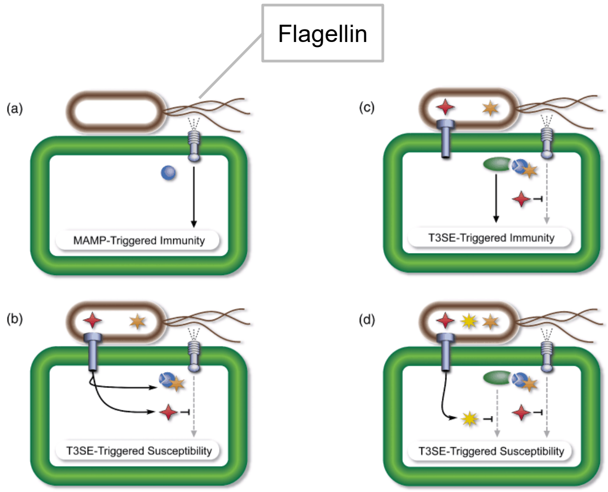
71,770 Visitors

File
Edit
View
History
Bookmarks
Tools
Help
Google
□ Background
□ Spin
Zoom
1HXW
PQS
PDB
OCA
Gaps?
Center Atom
Troubleshooting
Close
Go to the FirstGlance Control Panel
(What For?)
☑1 □2 □3 □4 □5 □6 □7
Positive ← Selection → Purifying
Spacefill
Reset View
Explore further in FirstGlance in Jmol.
Why?
Selecton Job Information
Download this Selecton result to save it.
Include this molecular view in a presentation.
Show PDB file header (includes Selecton scripts).
1HXW_SELECTON1176922472_PIPE
Jmol
Can't see the molecule?
FirstGlance in Jmol (ver. 1.38)

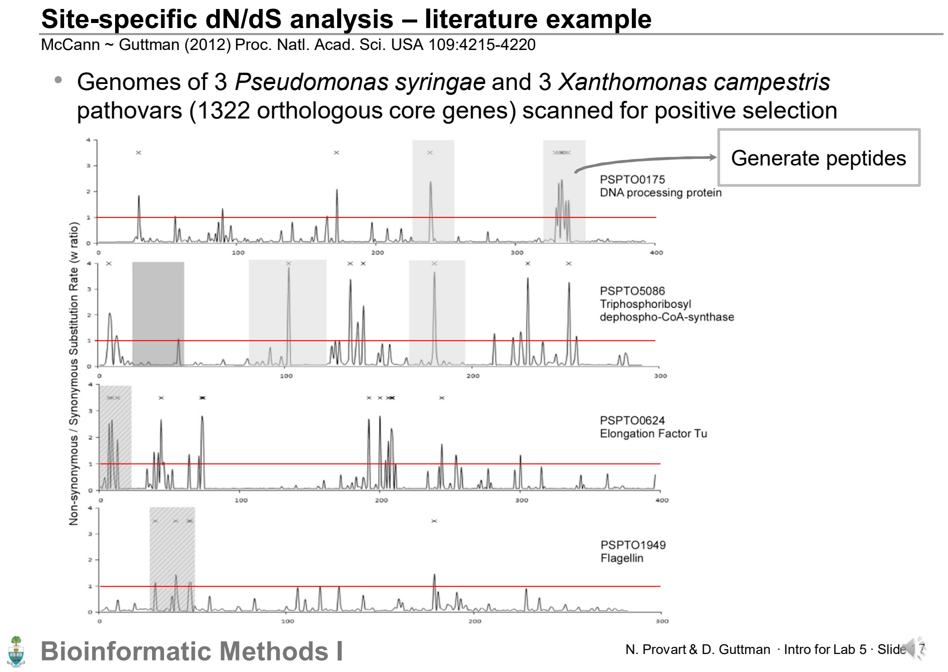
Stern, A., Doron-Faigenboim, A., Erez, E., Martz, E., Bacharach, E., Pupko, T. Selecton 2007: advanced models for detecting positive and purifying selection using a Bayesian inference approach. Nucleic Acids Research. 35: W506-W511.

Site-specific dN/dS analysis – literature example

Evolutionary arms race…



from McCann and Guttman (2007) New Phytologist 177:33



Site-specific dN/dS analysis – literature example

McCann ~ Guttman (2012) Proc. Natl. Acad. Sci. USA 109:4215-4220

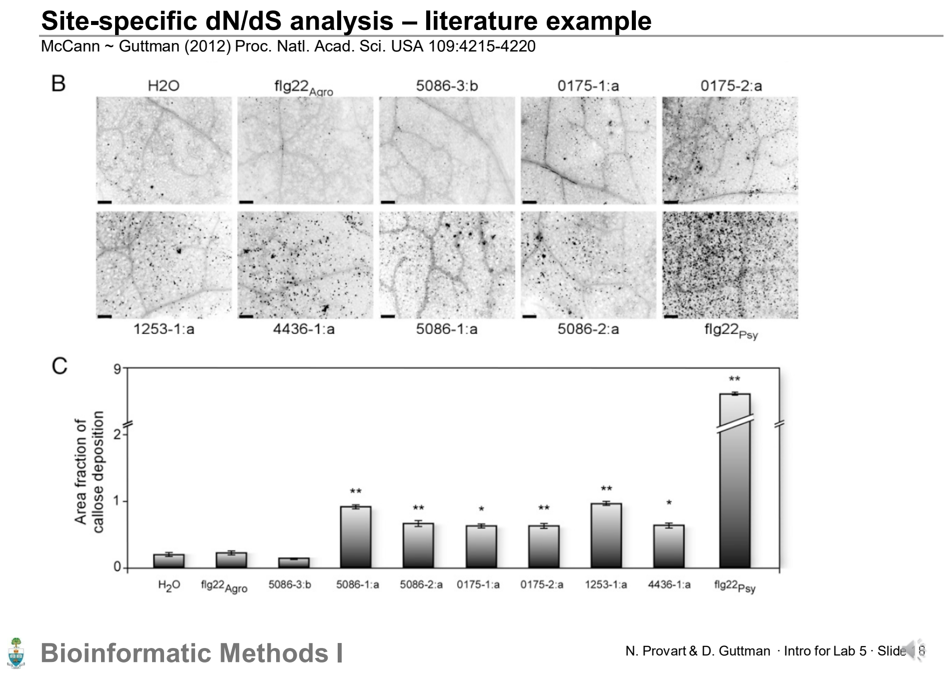
- Genomes of 3 *Pseudomonas syringae* and 3 *Xanthomonas campestris* pathovars (1322 orthologous core genes) scanned for positive selection

Generate peptides

Bioinformatic Methods I

N. Provart & D. Guttman · Intro for Lab 5 · Slide 17

17



Site-specific dN/dS analysis – literature example

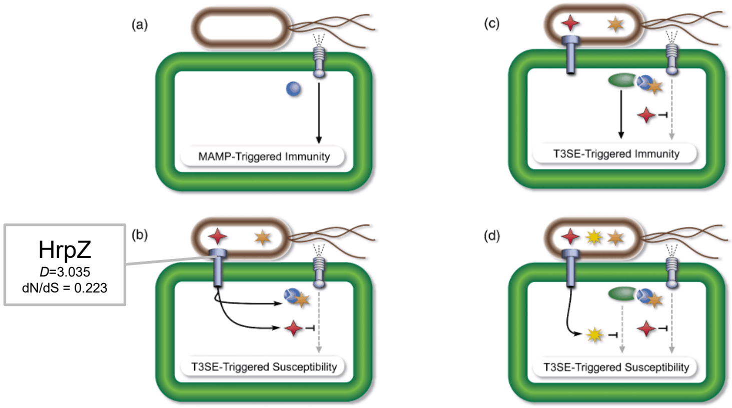
McCann ~ Guttman (2012) Proc. Natl. Acad. Sci. USA 109:4215-4220

Bioinformatic Methods I

N. Provart & D. Guttman · Intro for Lab 5 · Slide 18

18

Evolutionary arms race…



HrpZ
D=3.035
dN/dS = 0.223

from McCann and Guttman (2007) New Phytologist 177:33