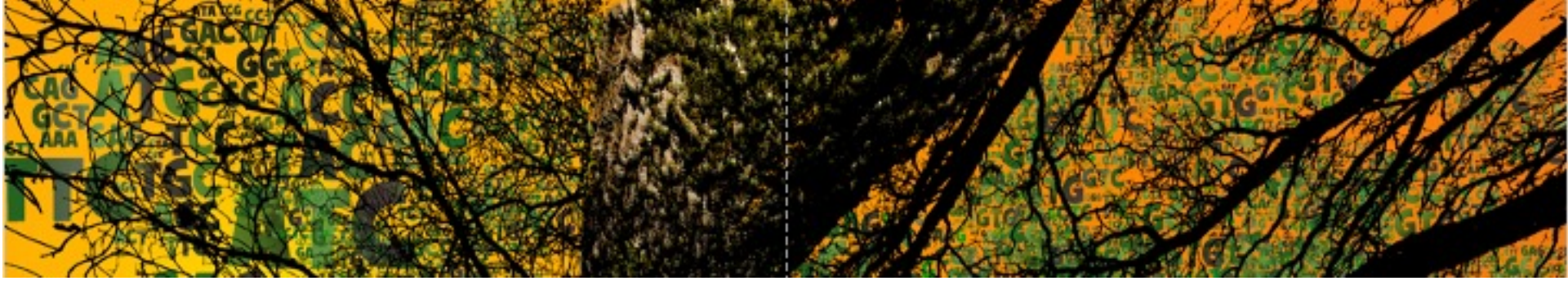


Large-scale selection analyses

Maria Anisimova

Institute of Computational Life Sciences
Zurich University of Applied Sciences - ZHAW

12.05.2022, Lund

Life Sciences and Facility Management
Institute of Computational Life Sciences

Swiss Institute of Bioinformatics

Two decades of large-scale selection scans

Inferring Nonneutral Evolution from Human-Chimp-Mouse Orthologous Gene Trios

Andrew G. Clark,[1] Stephen Glanowski,[3] Rasmus
Paul D. Thomas,[4] Anish Kejariwal,[4] Melissa A.
David M. Tanenbaum,[5] Daniel Civello,[6] Fu Lu,[5] Bri
Steve Ferriera,[3] Gary Wang,[3] Xianqgun Zh
Thomas J. White,[6] John J. Sninsky,[6] Mark D. A
Michele Cargill[6]†

2003, Science

Open access, freely available online PLoS BIOLOGY

A Scan for Positively Selected Genes in the Genomes of Humans and Chimpanzees

Rasmus Nielsen[1,2*], Carlos Bustamante[1], Andrew G. Clark[3], Stephen Glanowski[4], Timothy B. Sackton[3], Melissa J. Hubisz[1], Adi Fledel-Alon[1], David M. Tanenbaum[5], Daniel Civello[6], Thomas J. White[6], John J. Sninsky[6], Mark D. Adams[5¤], Michele Cargill[6]

2005

OPEN ACCESS Freely available online PLoS GENETICS

Patterns of Positive Selection in Six Mammalian Genomes

Carolin Kosiol[1], Tomáš Vinař[1], Rute R. da Fonseca[2], Melissa J. Hubisz[3], Carlos D. Bustamante[1], Rasmus Nielsen[2], Adam Siepel[1*]

2008

Research article

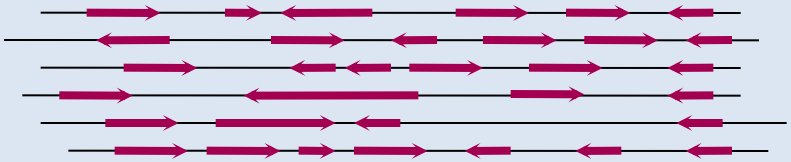
A systematic search for positive selection in higher plants (Embryophytes)

Christian Roth[1,2,3] and David A Liberles*[1,3]

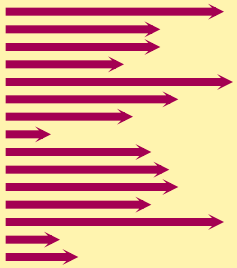
2006

Large-scale selection scans step-by-step

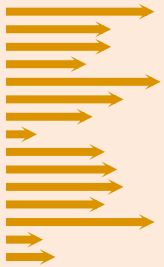
CDs from complete genomes /databases



DNA CDs library

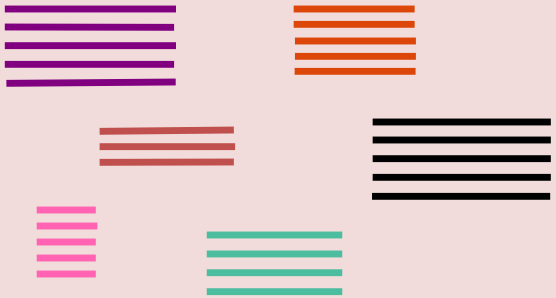


AA library



Homologous gene clusters
(unaligned AA data)

MSAs (both AA and CDs)

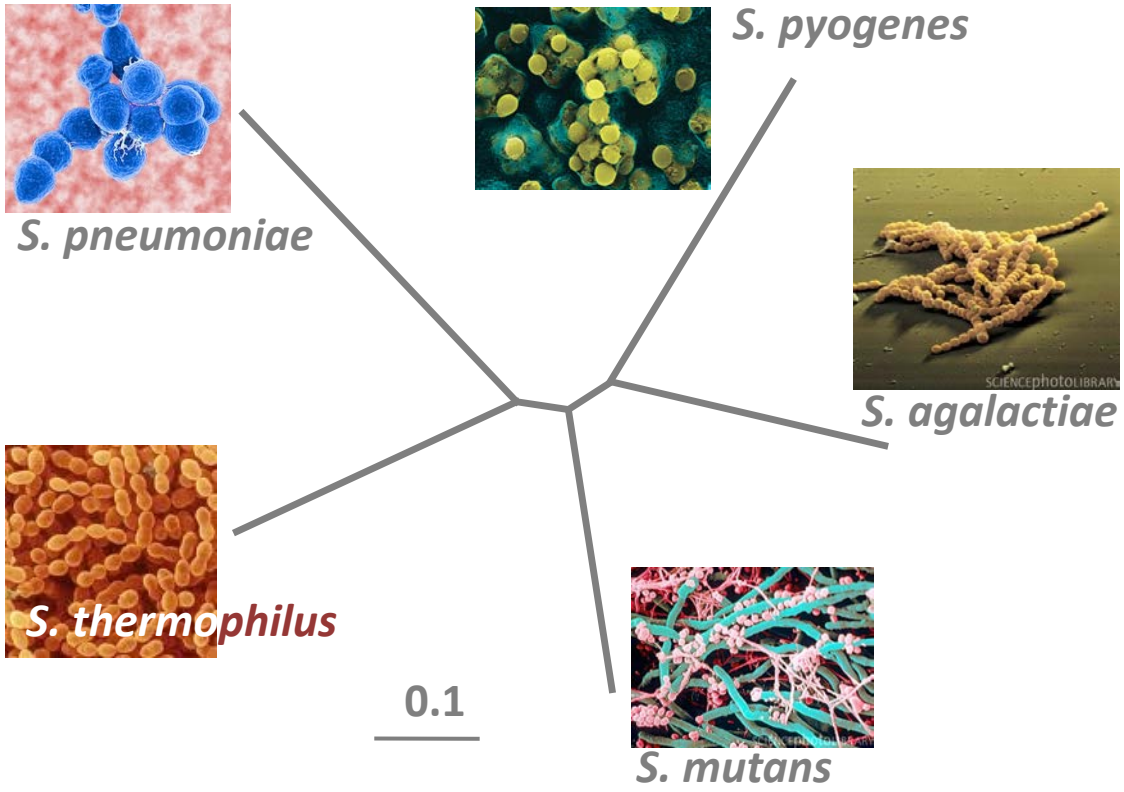


Statistical analysis

0.236251 9929 1.1542 0.00012 1.1542 0.03029 1.1542 0.03029
0.6767251 9945 1.1542 0.03029 1.1542 0.03029
0.1111251 9981 1.1542 0.01998 0.1111251 9981
0.230004 9139 1.1542 0.00912 0.1111251 9981 1.1542 ..

What do we do with all these numbers?

Natural selection in Streptococcus



Anisimova et al 2007 BMC Evol Biol

12 complete genomes

Positive selection in 136 genes:

29% connected to virulence

10% no ascribable function

7% essential to *S. pneumoniae*

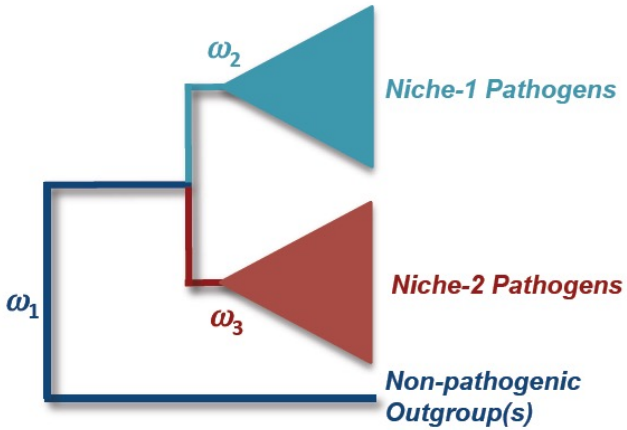
19% with body-site specific patterns of gene expression during invasive disease in *S. pyogenes* (infected blood, cerebrospinal fluid, epithelial cell contact)

Positive selection affects both core and accessory genes, most likely due to the antagonistic interaction between host and parasite.

Products of both core and auxiliary genes participate in complex networks that comprise the molecular basis of virulence.

Listeria phylogenomics
Mapping selection to phenotype

A: Gene-level data analysis



Null model (1 parameter):

$H_0: \omega_1 = \omega_2 = \omega_3$

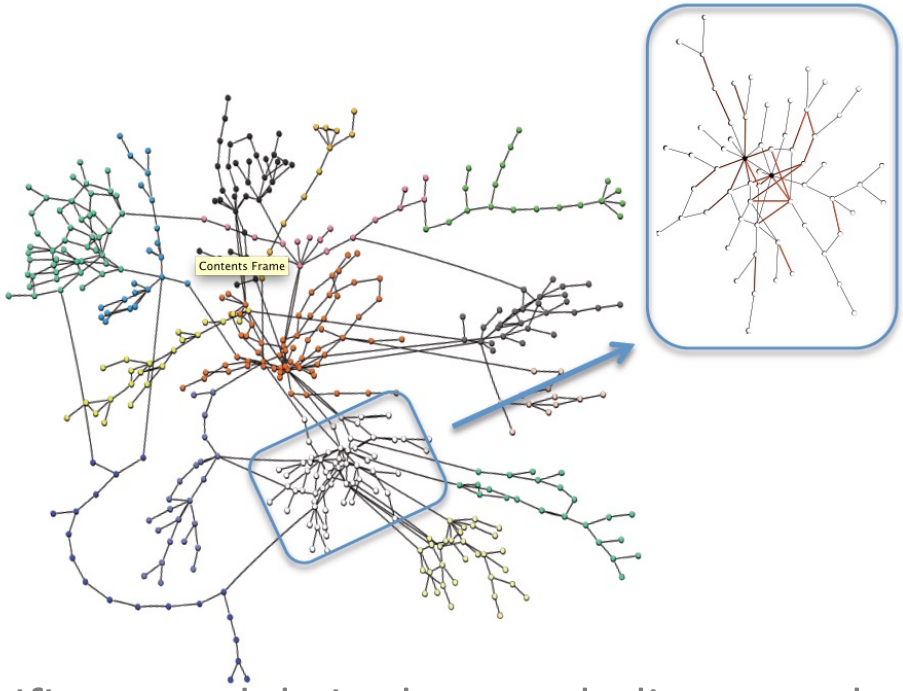
Alternatives (2 parameters):

$H_1: \omega_1 \neq \omega_2 = \omega_3$

$H_2: \omega_1 = \omega_2 \neq \omega_3$

$H_3: \omega_1 = \omega_3 \neq \omega_2$

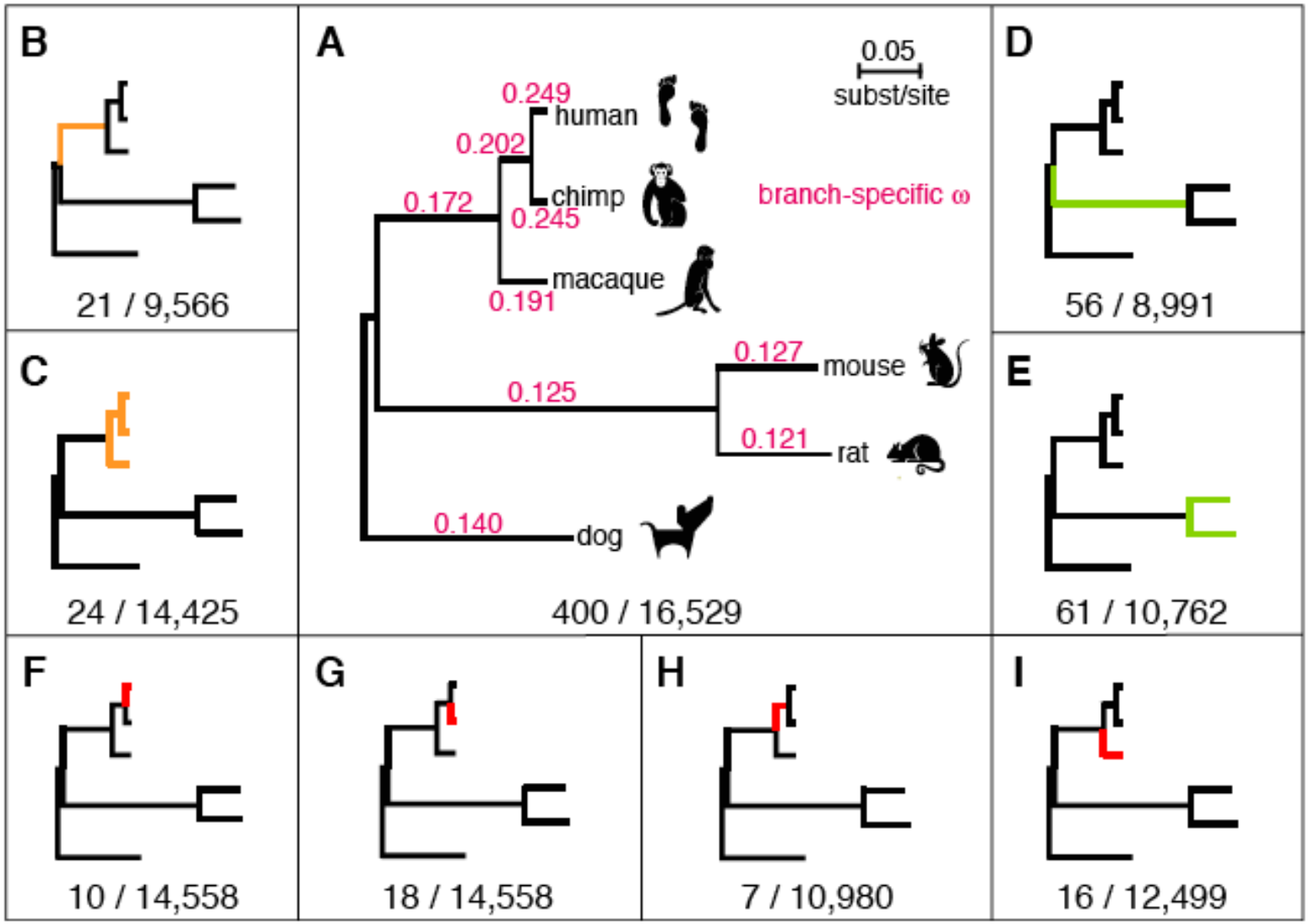
B: Phenotype-level data analysis



Blue box identifies a module in the metabolic network. Red links in the expanded view of this module indicate a significant cluster of genes subject to niche specific selection in "lineage I" *of L. monocytogenes.*

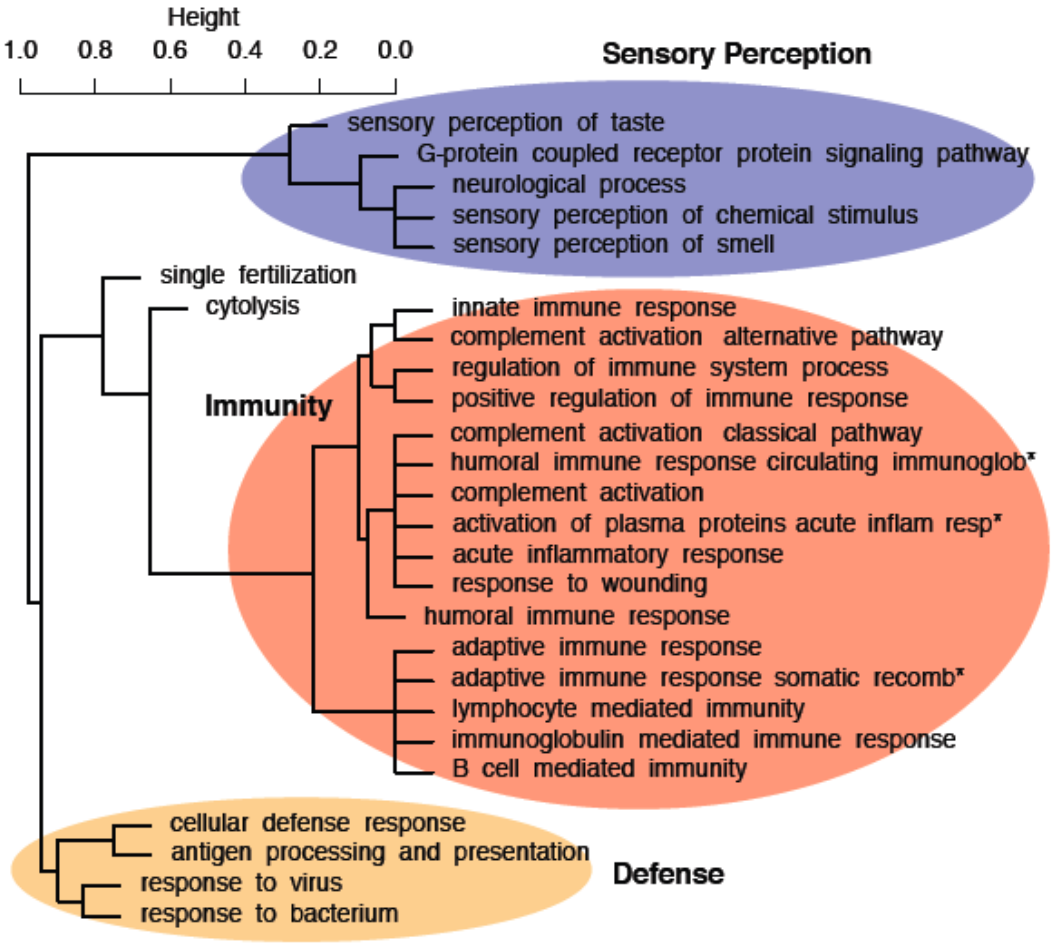
From Dunn et al 2009, MBE

Multiple LRTs: scan of mammalian genomes



Kosiol et al (2008)

Multiple LRTs: scan of mammalian genomes



Kosiol et al (2008)

Dissimilarity measure:

$$d_{AB} = 1 - \frac{|N(A) \cap N(B)|}{\min\{|N(A)|, |N(B)|\}}$$

Hierarchical clustering of GO categories (biological process) over-represented with genes under positive selection

Which proteins are under positive selection?

- Host proteins involved in defence or immunity against viral, bacterial, fungal or parasite attacks (MHC, immunoglobulin VH, class 1 chitinas).
- Viral or pathogen proteins involved in evading host defence (HIV env, nef, gap, pol, etc., capsid in FMD virus, flu virus hemagglutinin gene).
- Proteins or pheromones involved in reproduction (abalone sperm lysin, sea urchin bindin, proteins in mammals)
- Proteins that acquired new functions after gene duplication.
- Miscellaneous (diet, globins,etc.)


Conclusions

Detecting positive selection

- Pairwise methods – very low power
- Branch models allow variation over time but assume one ω for all sites - low power
- Site models allow variation among sites but assume selection pressure does not change over time – have higher power if positive selection is long term
- Branch-site models may be more successful at detecting episodic selection but are more difficult to fit, require more data and often have multiple sub-optimal peaks (caution with genome scans!)

Testing for positive selection

- LRT is accurate even for small datasets
- Power of LRT is better for larger datasets
- Watch out for recombination
- Accurate parameter estimation is more difficult, depends on model assumptions
- Bayesian site prediction is even more difficult than LRTs and parameter estimation
- There is an optimal window of sequence divergence (sequences should be not too similar and not saturated)
- Robustness of results: Use several models & tests
- Check for local optima, especially for complex models

Weaknesses of methods based on codon models

- Model assumptions may be unrealistic (but some assumptions matter more than others)
- The method detects positive selection only if it generates excessive nonsynonymous substitutions. It may lack power in detecting one-off directional selection or when the sequences are highly similar or highly divergent. Little power with population data.
- Do not work for noncoding DNA (but see Wong & Nielsen 2003 Genetics)
- Sensitive to sequence and alignment errors (Fletcher & Yang 2010 Mol Biol Evol 27; Privman et al. 2011 Mol Biol Evol 29; Jordan & Goldman 2012 Mol Biol Evol 29)

Criticisms on codon models

by M. Nei, Y. Suzuki, & A.L. Hughes

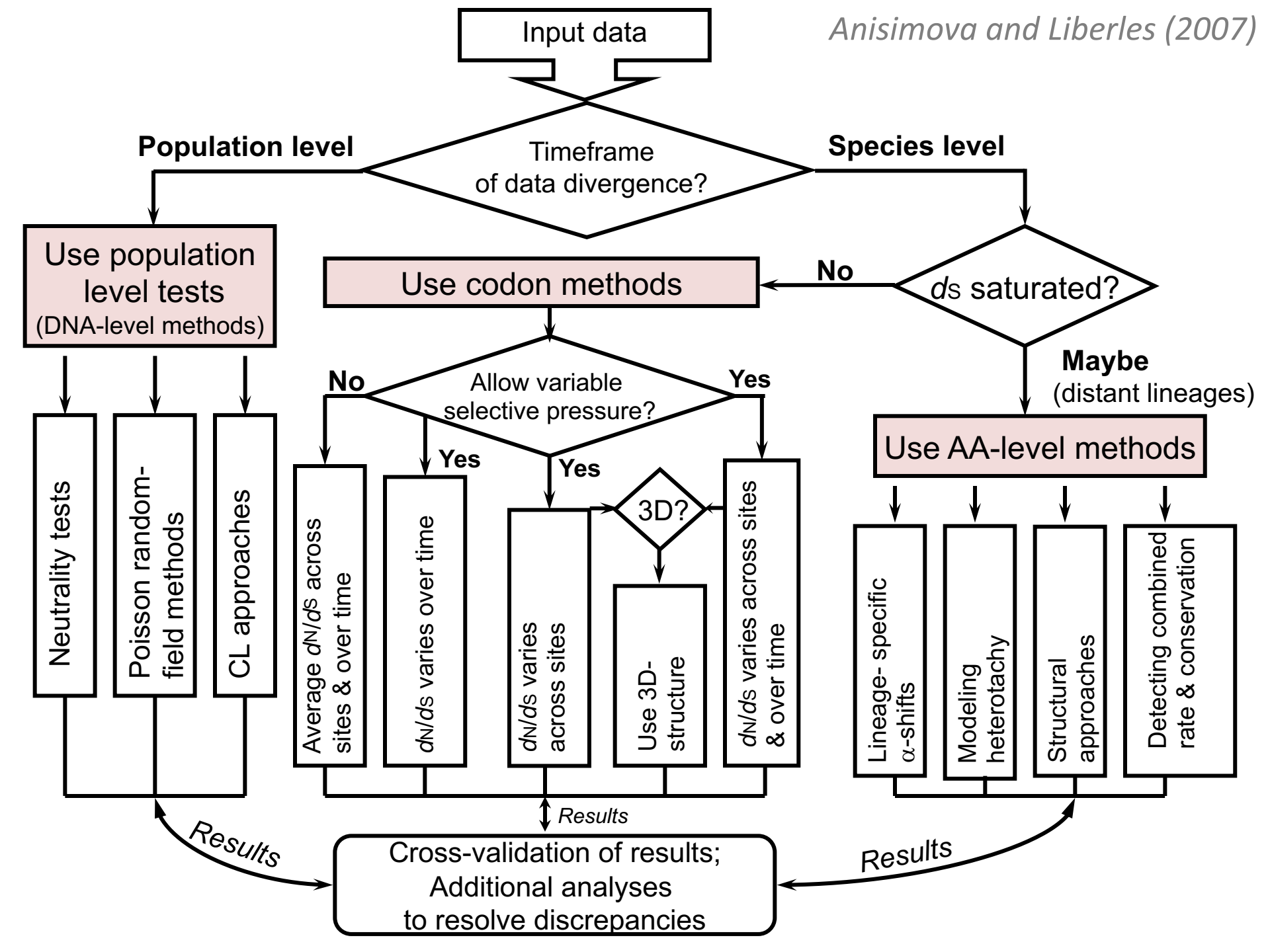
Hughes AL. 2007. Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. *Heredity* 99

Nozawa, Suzuki & Nei. 2009. *PNAS* 106

Yang Z, dos Reis M. 2011. Statistical properties of the branch-site test of positive selection. *Mol Biol Evol* 28

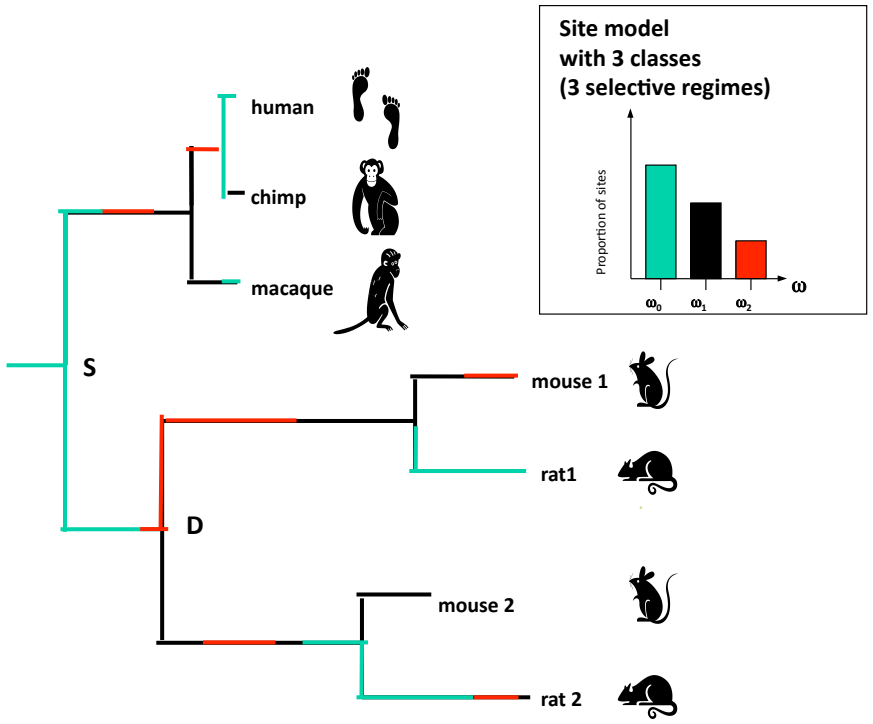
Zhai W, Nielsen R, Goldman N, Yang Z. 2012. Looking for Darwin in genomic sequences - validity and success of statistical methods. *Mol Biol Evol* 29

MacCallum, C. & Hill, E. 2006 Being positive about selection. *PLoS Biol* **4**, e87


Anisimova and Liberles (2007)
Input data
Timeframe of data divergence?
Population level
Species level
Use population level tests (DNA-level methods)
Use codon methods
No
dS saturated?
Maybe (distant lineages)
No
Allow variable selective pressure?
Yes
Yes
Yes
3D?
Use AA-level methods
Neutrality tests
Poisson random-field methods
CL approaches
Average dN/dS across sites & over time
dN/dS varies over time
dN/dS varies across sites
Use 3D-structure
dN/dS varies across sites & over time
Lineage- specific α-shifts
Modeling heterotachy
Structural approaches
Detecting combined rate & conservation
Results
Results
Results
Cross-validation of results; Additional analyses to resolve discrepancies

The many faces of codon models

- Detecting selection
- Studying codon bias
- Inferring phylogenies
- Dating speciation events
- Ancestral reconstruction
- Changes in time & space
- Predicting coding regions
- Improved alignment
- Inferring gene features (phyloHMM, netHMM)
- Simulation of data



Markov modulated model:
Guindon et al. 2004

Reviews of codon models:
Kosiol and Anisimova 2012
Anisimova and Kosiol 2009