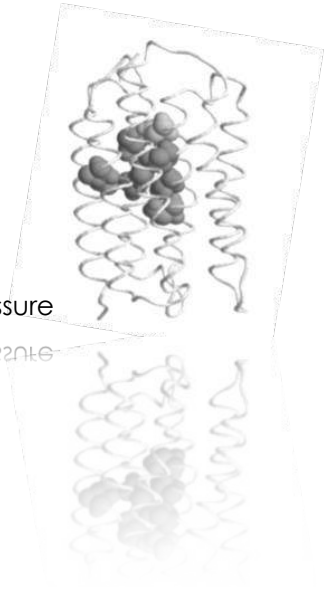
part 3: analysis of natural selection pressure



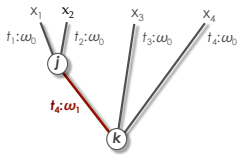
markov models are good

phenomenological codon models do have many benefits:

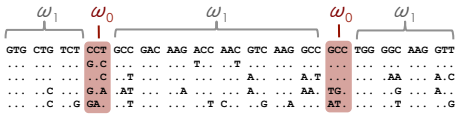
- principled framework for statistical inference
- avoiding *ad hoc* corrections of "counting" methods
- computation of transition probabilities *
- explicit use of phylogeny
- model ω variation among sites
- model ω variation among branches
- many other kinds of models for ω

* Computation of transition probabilities accomplishes, in just one step, (1) a proper correction for multiple substitutions, (2) weighting for alternative pathways between codons and (3) is the basis for estimating the values of the model parameters from the data in hand.

two basic types of models

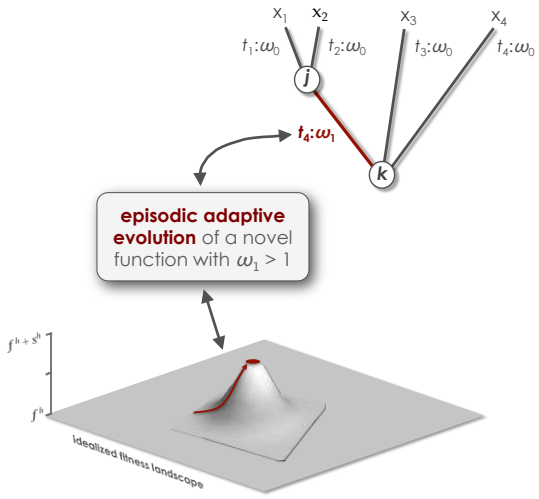


branch models
(*ω* varies among branches)

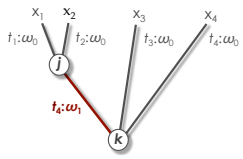


site models
(*ω* varies among sites)

interpretation of a branch model



branch models*



variation (ω) among branches:	approach
Yang, 1998	fixed effects
Bielawski and Yang, 2003	fixed effects
Seo et al. 2004	auto-correlated rates
Kosakovsky Pond and Frost, 2005	genetic algorithm
Dutheil et al. 2012	clustering algorithm

* *these methods can be useful when selection pressure is strongly* ***episodic***

site models*

```
GTG CTG TCT CCT GCC GAC AAG ACC AAC GTC AAG GCC GCC TGG GGC AAG GTT GGC GCG CAC
... ... ... G.C ... ... ... T.. ..T ... ... ... ... ... ... ... ... ... .GC A..
... ... ... ..C ..T ... ... ... ... A.. ... A.T ... ... ... .AA ... A.C ... AGC ...
... ..C ... G.A .AT ... ..A ... ... A.. ... AA. TG. ... ..G ... A.. ..T .GC ..T
... ..C ..G GA. ..T ... ... ..T C.. ..G ..A ... AT. ... ..T ... ..G ..A .GC ...
```

variation (ω) among sites:	approach
Yang and Swanson, 2002	fixed effects (ML)
Bao, Gu and Bielawski, 2006	fixed effects (ML)
Massingham and Goldman, 2005	site wise (LRT)
Kosakovsky Pond and Frost, 2005	site wise (LRT)
Nielsen and Yang, 1998	mixture model (ML)
Kosakovsky Pond, Frost and Muse, 2005	mixture model (ML)
Huelsenbeck and Dyer, 2004; Huelsenbeck et al. 2006	mixture (Bayesian)
Rubenstein et al. 2011	mixture model (ML)
Bao, Gu, Dunn and Bielawski 2008 & 2011	mixture (LiBaC/MBC)
Murell et al. 2013	mixture (Bayesian)

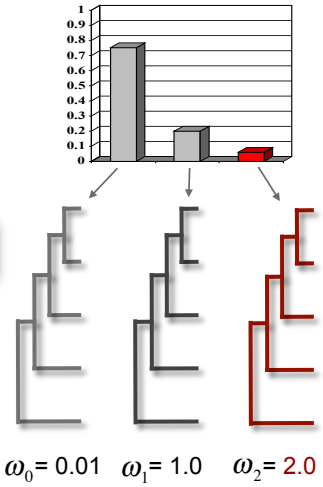
- *useful when at some sites evolve under* ***diversifying selection*** *pressure over long periods of time*
- *this is not a comprehensive list*

site models: discrete model (M3)

MIXTURE-MODEL LIKELIHOOD

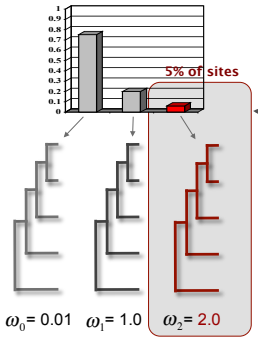
$$P(\mathbf{x}_h) = \sum_{i=0}^{K-1} p_i P(\mathbf{x}_h \mid \omega_i)$$

conditional likelihood calculation (see part 1)



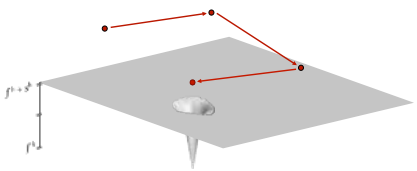
ω_0 = 0.01 ω_1 = 1.0 ω_2 = 2.0

interpretation of a sites-model

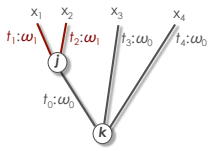


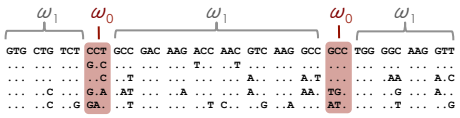
ω_0 = 0.01 ω_1 = 1.0 ω_2 = 2.0

diversifying selection (frequency dependent) at 5% of sites with $\omega_2 = 2$



models for variation among branches & sites





branch models
(ω varies among branches)

site models
(ω varies among sites)

branch-site models
(combines the features of above models)

models for variation among branches & sites

variation (ω) among branches & sites:	approach
Yang and Nielsen, 2002	fixed+mixture (ML)
Forsberg and Christiansen, 2003	fixed+mixture (ML)
Bielawski and Yang, 2004	fixed+mixture (ML)
Giundon et al., 2004	switching (ML)
Zhang et al. 2005	fixed+mixture (ML)
Kosakovsky Pond et al. 2011, 2012	full mixture (ML)

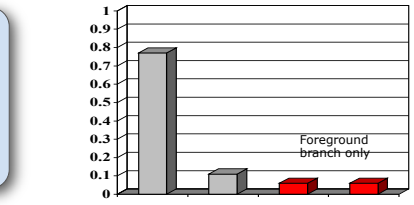
* *these methods can be useful when selection* ***pressures change over time at just a fraction of sites***

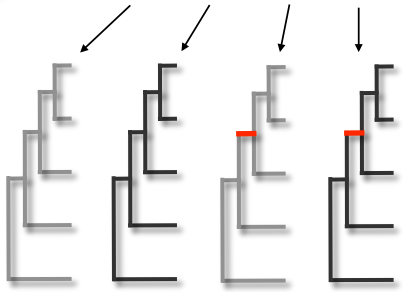
* *it can be a challenge to apply these methods properly (****more about this later****)*

branch-site "Model B"

MIXTURE-MODEL LIKELIHOOD

$$P(\mathbf{x}_h) = \sum_{i=0}^{K-1} p_i P(\mathbf{x}_h \mid \omega_i)$$

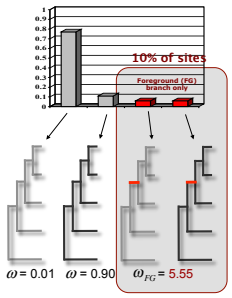




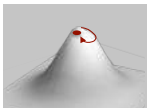
$\omega = 0.01$ $\omega = 0.90$ $\omega = 5.55$

ω for background branches are from site-classes 1 and 2 (0.01 or 0.90)

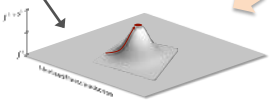
two scenarios can yield branch-sites with dN/dS > 1



10% of sites have **shifting balance** on a fixed peak (**same function**)



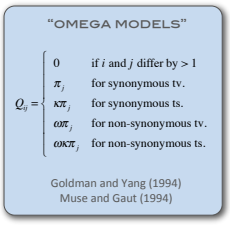
episodic adaptive evolution at **10% of sites** for **novel function**



branch-site codon models cannot tell which scenario is correct without external information!

Jones et al (2016) MBE

model-based inference



model based inference

3 analytical tasks

task 1. parameter estimation (e.g., ω)

task 2. hypothesis testing

task 3. make predictions (e.g., sites having $\omega > 1$)

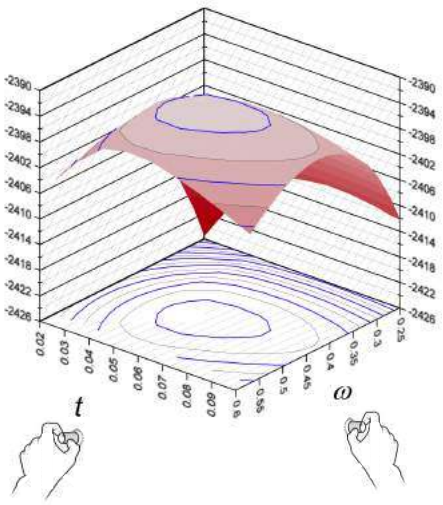
task 1: parameter estimation

t, κ, ω = unknown constants estimated by ML

π's = empirical [GY: F3×4 or F61 in Lab]

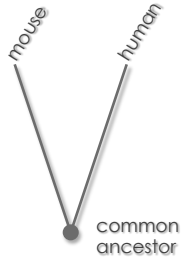
use a numerical hill-climbing algorithm to maximize the likelihood function

task 1: parameter estimation



Parameters: t and ω

Gene: acetylcholine α receptor



lnL = -2399

task 2: statistical significance

task 1. parameter estimation (e.g., ω) ✔

task 2. hypothesis testing ⬅ LRT

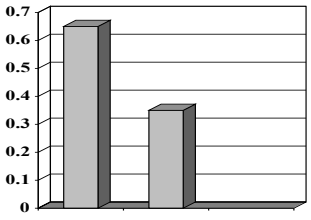
task 3. prediction / site identification

task 2: likelihood ratio test for positive selection

H_0: variable selective pressure but NO positive selection (M1)
H_1: variable selective pressure with positive selection (M2)

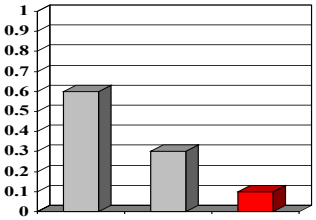
Compare **$2\Delta l$** = $2(l_1 - l_0)$ with a χ^2 distribution

Model 1a



$\hat{\omega}$ = 0.5 (ω = 1)

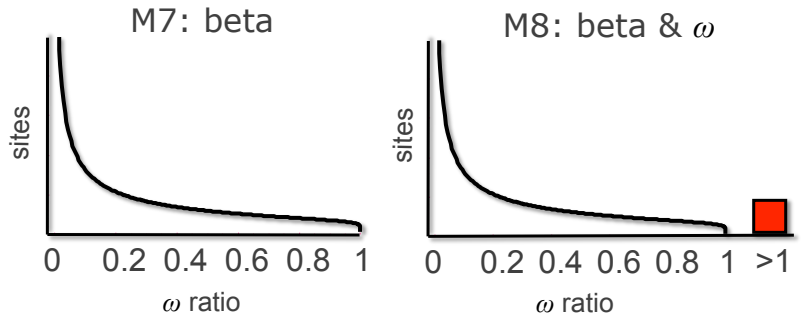
Model 2a



$\hat{\omega}$ = 0.5 (ω = 1) $\hat{\omega}$ = 3.25

task 2: likelihood ratio test for positive selection

H_0: Beta distributed variable selective pressure (M7)
H_1: Beta plus positive selection (M8)
Compare $2\Delta l = 2(l_1 - l_0)$ with a χ^2 distribution



task 3: identify the selected sites

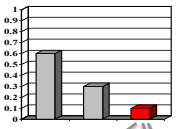
task 1. parameter estimation (e.g., ω) ✔

task 2. hypothesis testing ✔

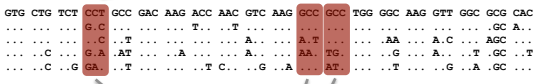
task 3. prediction / site identification ← Bayes' rule

task 3: which sites have *dN*/*dS* > 1

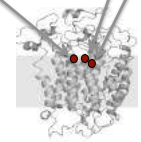
model:
9% have $\omega > 1$



Bayes' rule:
site 4, 12 & 13



structure:
sites are in contact



Bayes' rule: yet another (silly) example

Suppose that a population consists of 60% males and 40% females, and a disease occurs at the rate 1% in males and 0.1% in females.

Q_1: What is the probability that any individual carries the disease?

A_1: $0.6 \times 0.01 + 0.4 \times 0.001 = 0.0064$

$$P(\mathrm{D}) = P(\mathrm{M})P(\mathrm{D}|\mathrm{M}) + P(\mathrm{F})P(\mathrm{D}|\mathrm{F})$$

See Yang and Bielawski (2000) TREE 15:496-503 for a detailed presentation of this example

Bayes' rule: yet another (silly) example

Q_2: Given that an individual carries the disease, what is the probability that it is a male?

A_2: 0.6 × 0.01/0.0064 = 0.94

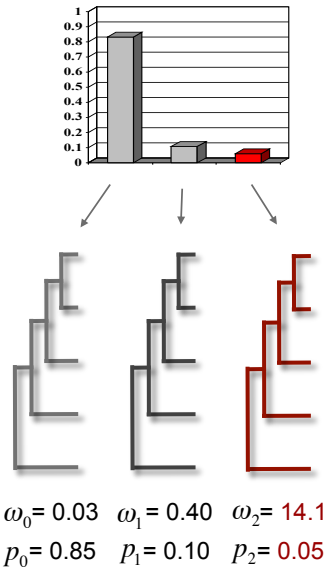
$$P(\text{M}|\text{D}) = \frac{P(\text{M})\,P(\text{D}|\text{M})}{P(\text{D})}$$

See Yang and Bielawski (2000) TREE 15:496-503 for a detailed presentation of this example

review the mixture likelihood

$$P(\mathbf{x}_h) = \sum_{i=0}^{K-1} p(\omega_i) P(\mathbf{x}_h \mid \omega_i)$$

Total probability | Prior | Likelihood



Bayes' rule for identifying selected sites

Site class 0: ω_0 = .03, 85% of codon sites
Site class 1: ω_1 = .40, 10% of codon sites
Site class 2: ω_2 = 14, 05% of codon sites

?

Prior probability of hypothesis (ω_2)

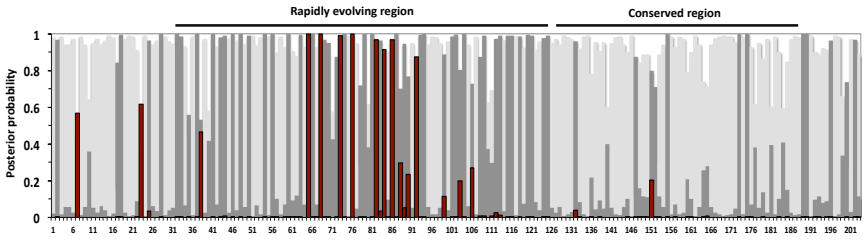
Likelihood of hypothesis (ω_2)

$$P(\omega_2 \mid x_h) = \frac{P(\omega_2)P(x_h \mid \omega_2)}{\sum_{i=0}^{K-1} P(\omega_i)P(x_h \mid \omega_i)}$$

Posterior probability of hypothesis (ω_2)

Marginal probability (Total probability) of the data

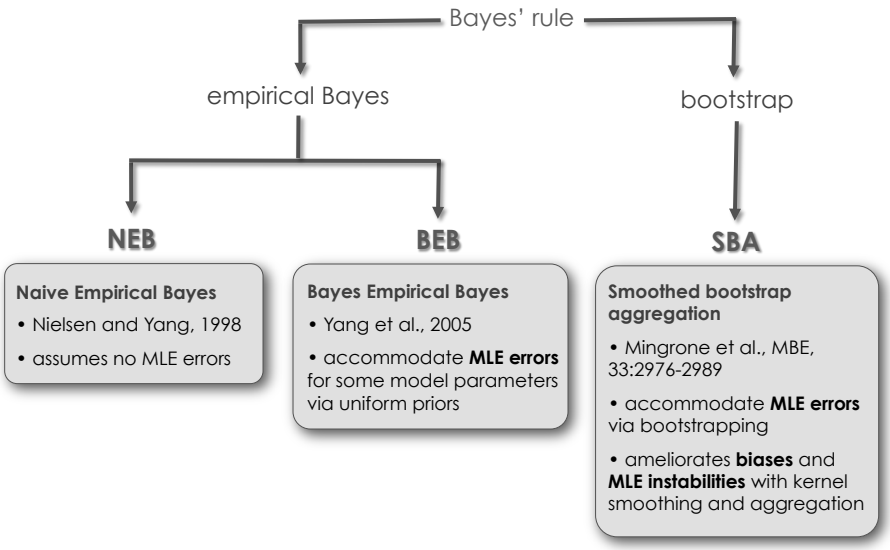
task 3: Bayes rule for which sites have dN/dS > 1



Site class 0: ω_0 = .03 (strong purifying selection)
Site class 1: ω_1 = .40 (weak purifying selection)
Site class 2: ω_2 = 14 (positive selection)

NOTE: The posterior probability should NOT be interpreted as a "*P*-value"; it can be interpreted as a measure of relative support, although there is rarely any attempt at "calibration".

task 3: Bayes rule for which sites have *dN*/*dS* > 1

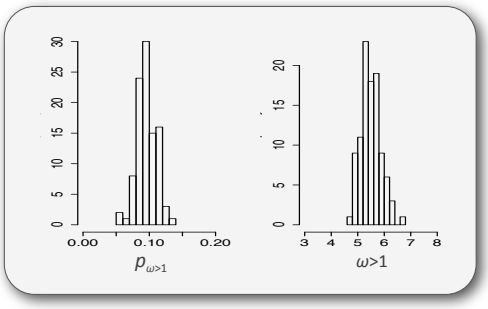


critical question:

Have the requirements for maximum likelihood inference been met?

(rarely addressed in real data analyses)

regularity conditions have been met

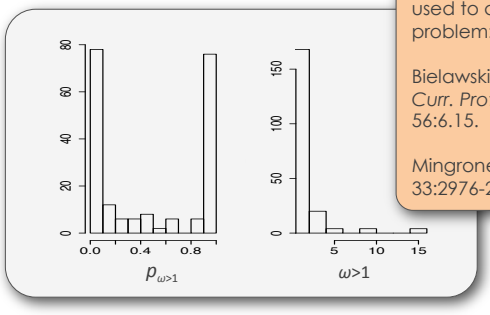


Normal MLE uncertainty (M2a)

- large sample size with regularity conditions
- MLEs approximately unbiased and minimum variance

$$\hat{\theta} \sim N\left(\theta, I\left(\hat{\theta}\right)^{-1}\right)$$

regularity conditions have **NOT** been met



bootstrapping can be used to diagnose this problem:

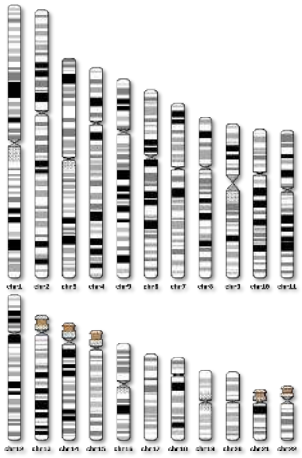
Bielawski et al. (2016) *Curr. Protoc. Bioinf.* 56:6.15.

Mingrone et al., *MBE*, 33:2976-2989

MLE instabilities (M2a)

- small sample sizes and θ on boundary
- continuous θ has been discretized (*e.g.*, M2a)
- non-Gaussian, over-dispersed, divergence among datasets

multi-gene evolutionary survey
(*& experimental validation*)

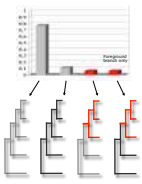


evolutionary survey

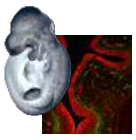
"Identification and assessment of *NR2C1* (*TR2*) as a modulator of pluripotentiality during hominid evolution"

Baker et al. (2016) *Genetics*, 203:905-922

Sample, process, and align 48 nuclear receptors



(1) evolutionary survey
(2) rank candidates

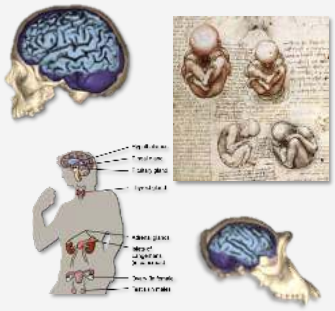


Experimental assessment of evolutionary change in pluripotentiality

evolutionary survey

nuclear receptor (NR) family

reproduction, development, and endocrine signaling



hormones & behavior

Intracellular transcription factors regulate gene expression

evolutionary survey

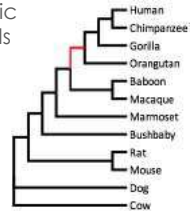
analysis of 48 primate NRs

Phase 1:

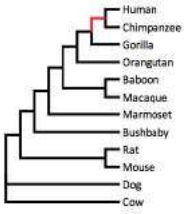
1. processing and Q.C.
2. structurally-aware modelling of codon evolution
3. test each gene for 5 different evolutionary scenarios
4. use "FDR control" to identify families of genes associated with different evolutionary scenarios

evolutionary survey

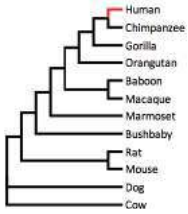
episodic models



1A) H_1: $\omega_M \neq \omega_{GA} = \omega_{HC} = \omega_H$

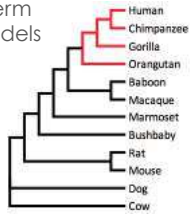


1B) H_2: $\omega_M = \omega_{GA} \neq \omega_{HC} = \omega_H$

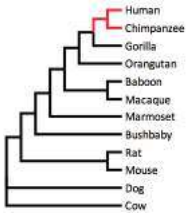


1C) H_3: $\omega_M = \omega_{GA} = \omega_{HC} \neq \omega_H$

long-term shift models



1D) H_4: $\omega_M \neq \omega_{GA} = \omega_H$



1E) H_5: $\omega_M = \omega_{GA} \neq \omega_{HC}$

KEY

ω_M : ω_{Mammal}

ω_{GA} : $\omega_{GreatApe}$

ω_{HC} : $\omega_{Human\text{-}Chimpanzee}$

ω_H : ω_{Human}

evolutionary survey

phase 1: analysis of 48 NRs

evolutionary scenario	**LRT: $p < 0.05$**	**FDR: $q < 0.05$**
episodic models		
H_1: great apes	1 gene	0 genes
H_2: human-chimp	0 genes	0 genes
H_3: humans	4 genes	1 genes
long-term shift		
H_4: great apes	10 genes	5 genes
H_5: human-chimp	8 genes	5 genes

number of unique candidate genes: 9

evolutionary survey

analysis of 9 NRs

phase 2: reliability and robustness assessment

1. alignment (independent evaluations)
2. recombination
3. MG94 style codon model
4. alternative tree topologies
5. robustness to variation in baseline DNA/RNA rates
6. bootstrapping

nuclear receptor *NR1D1*: positive selection along human lineage at 1% of sites

1. alignment assessment ✔
2. no evidence of recombination ✔
3. alternate framework (MG94): *LRTs and MLEs robust*

brief detour

how to model codon frequencies?

	substitution rates are proportional to empirical frequency of:
Goldman and Yang 1994 (GY):	target codon
Muse and Gaut 1994 (MG):	target nucleotide

See Rodrique et al. (2008) for a comparison of GY and MG style codon models that suggests the MG style, combined with parameters for codon preferences, might be the most desirable core-model for future development.

The MutSel process (part 1) is inherently a process whereby the transition probability depends on the target nucleotide (MG).

how to model codon frequencies?

example: A → C

AAA → CAA

AAA → ACA

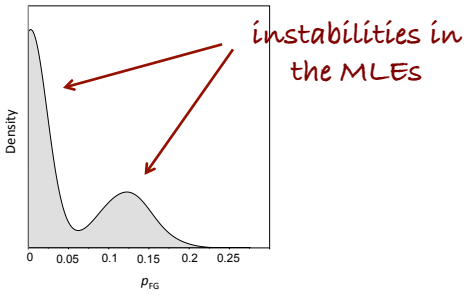
AAA → AAC

	Δ at codon position		
	1st	2nd	3rd
GY	π_{CAA}	π_{ACA}	π_{AAC}
MG	π_c^1	π_c^2	π_c^3

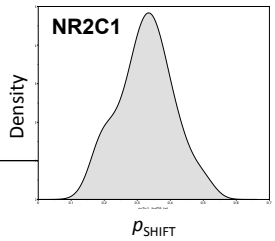
nuclear receptor *NR1D1*: positive selection along human lineage at 1% of sites

1. alignment assessment ✔
2. no evidence of recombination ✔
3. alternate framework (MG94): *LRTs and MLEs robust* ✔
4. alternate tree topology: *LRTs and MLEs robust* ✔
5. variation in baseline DNA/RNA rate of evolution: ✔
6. bootstrap distribution: ✔

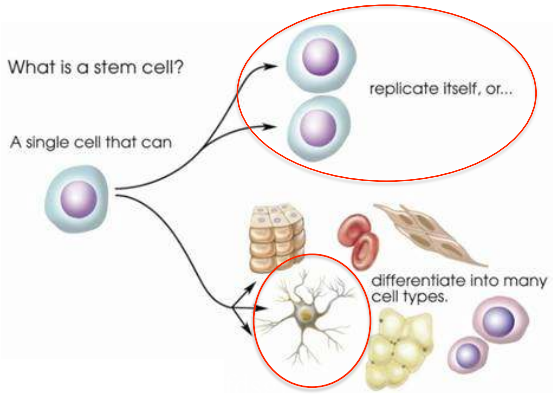
FULL STOP



gene	status	note
ESSRA	**excluded**	MLE instabilities
ESSRB	**excluded**	MLE instabilities
NR1D1	**excluded**	MLE instabilities
NR2E3	**excluded**	recombination
RARG	**excluded**	MLE instabilities
ESSRB	ranked 4th	
PGR	ranked 3rd	
RORA	ranked 2nd	
NR2C1	**ranked 1st**	



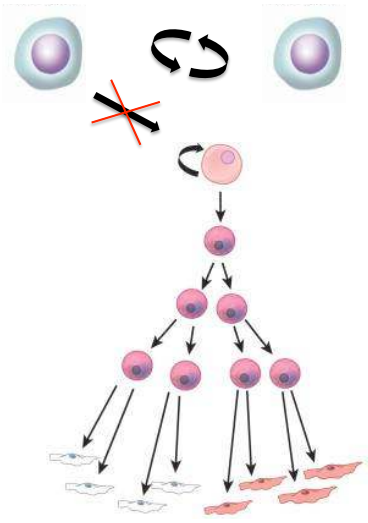
What is the function of *NR2C1?*



- Prefrontal cortex
- Embryonic stem cell maintenance
- Neuronal differentiation

image: C. Twomey

Oct4 & Nanog

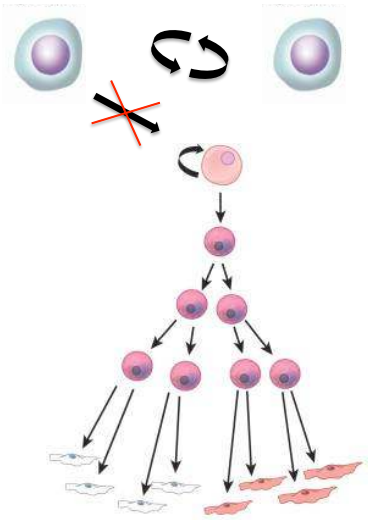


→ *OCT4 & Nanog*

OCT4 & Nanog are essential for maintaining the self-renewing undifferentiated state of embryonic stem cells

image: White 2015

Oct4 & Nanog



→ *OCT4 & Nanog*

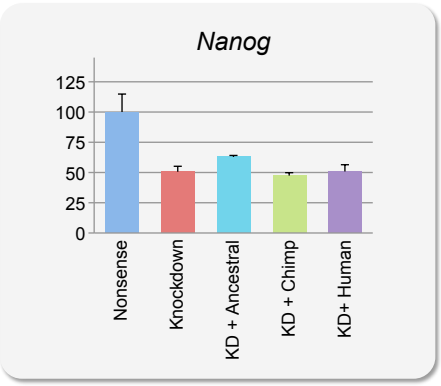
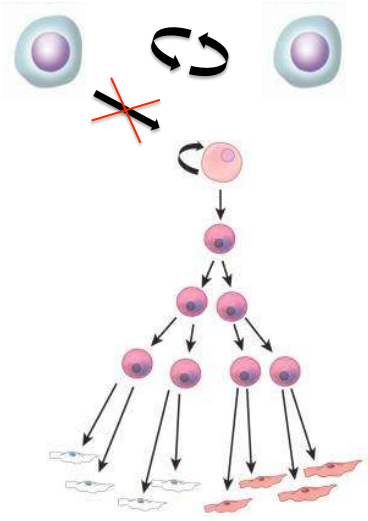


image: White 2015

I. Introduction II. Comparative genomics III. Gene expression IV. Ancestral gene reconstruction V. Future directions VI. Conclusions

Oct4 & Nanog



→ *OCT4 & Nanog*

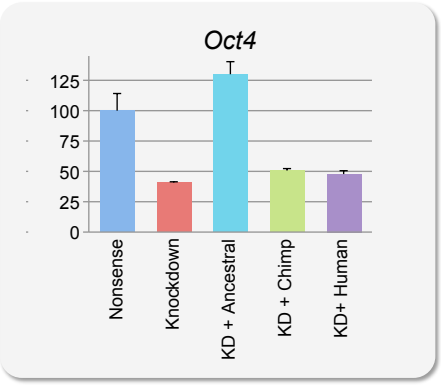


image: White 2015

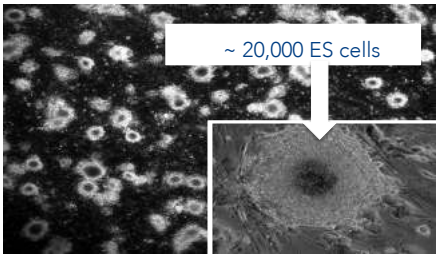
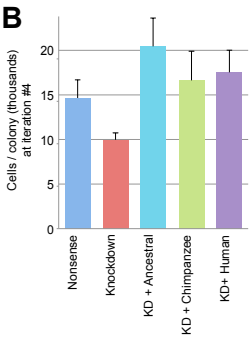
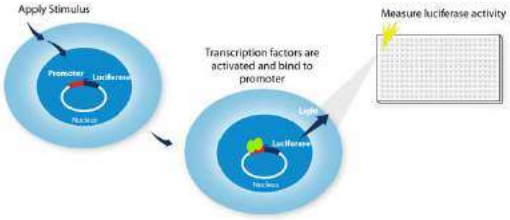
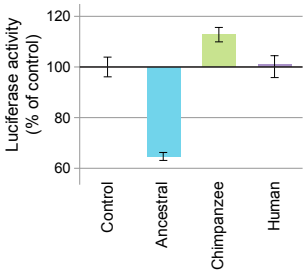
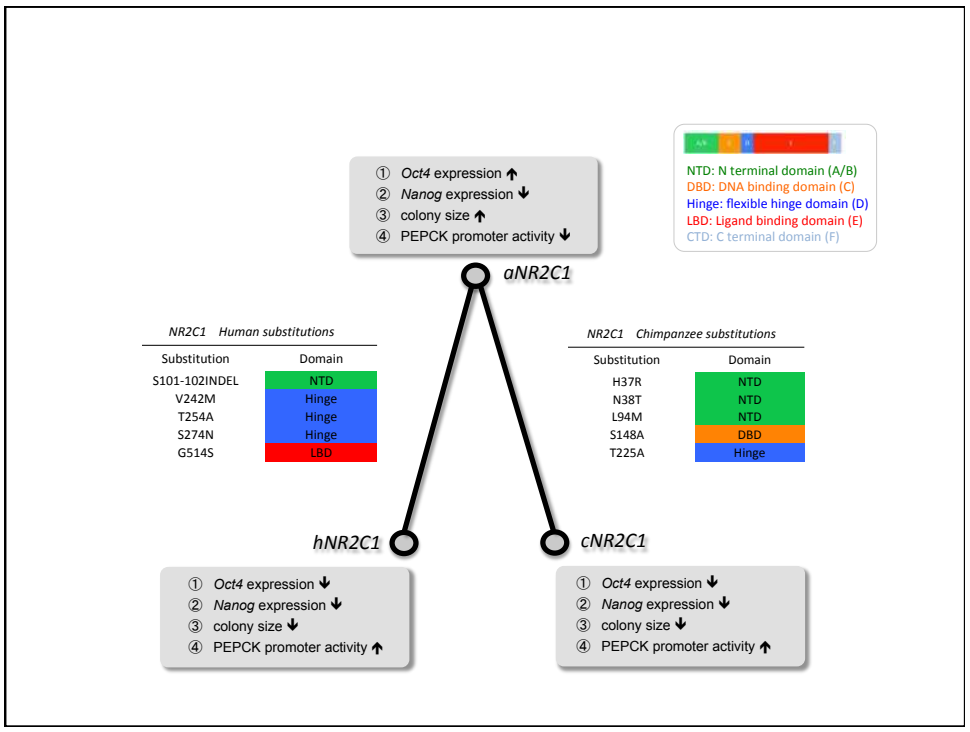


Photo: Tom Maynard

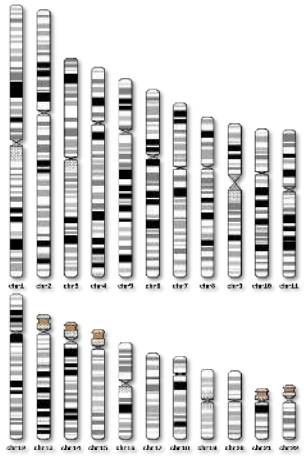








best practices in large-scale (genomic) surveys



best practices for "genome scans"

Phase 1: Large scale survey

1. Do as much quality control and cross validation in the pipeline as possible. Remove as many poor candidates as possible.
2. alignment validation
3. plan your analysis carefully; cover a variety of evolutionary questions; execute that plan
4. control the false discovery rate: identify "families" of genes which are significantly associated with specific evolutionary scenarios

best practices for "genome scans"

Phase 2: Robustness analysis

1. independent alignment validation (possible re-analysis)
2. test for recombination
3. re-run the analytical plan using a different formulation of the codon model (e.g., MG vs, GY).
4. re-run the analytical plan using alternative tree topologies.
5. bootstrap to assess MLE distributions for instabilities.
6. smoothed aggregated bootstrap for site-wise inference

best practices for "genome scans"

for further discussion of best practices see …

- Baker et al. (2016) *Genetics*, 203:905-922
- Bielawski et al. (2016) *Curr. Protoc. Bioinf.*, 56: Unit 6.15
- Mingrone et al. (2016) *MBE*, 33:2976-2989

THE END.