Probabilistic approaches for detecting and locating whole genome duplications

Charles-Elie Rabier

joint work with

Cécile Ané and Tram Ta

INRA MIAT / Statistics Department UW Madison

March 2015

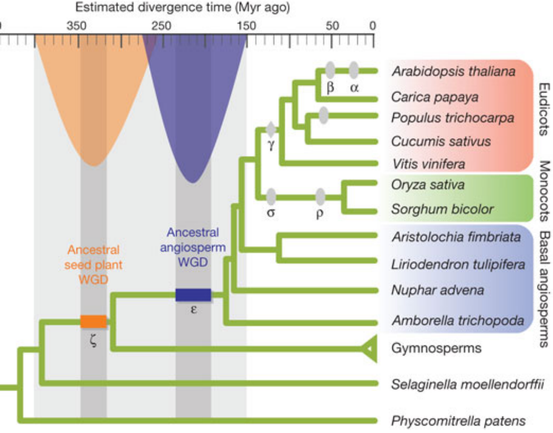
Whole Genome Duplication (WGD)

"Ancestral polyploidy in seed plants and angiosperms", Jiao et al. (Nature 2009)

"Whole-genome duplication followed by gene loss and diploidization has long been recognized as an important evolutionary force in animals, fungi and other organisms, especially plants"

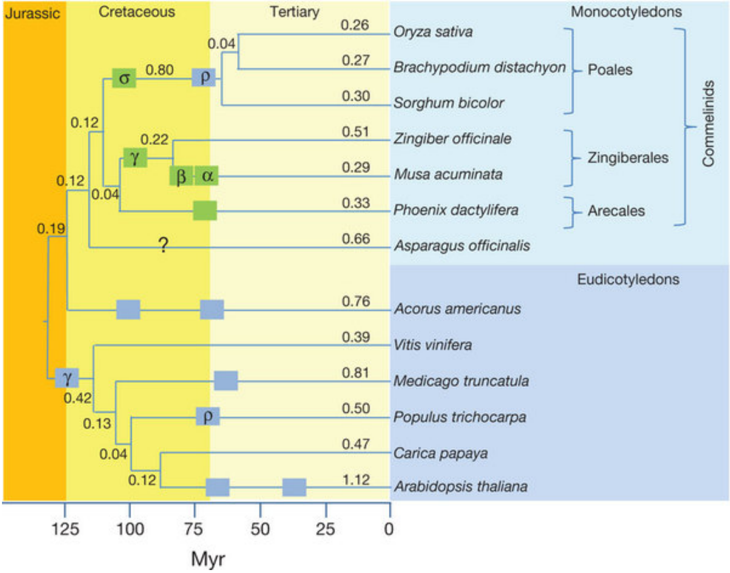
WGD in seed plants and angiosperms

Jiao et al. (Nature 2009)



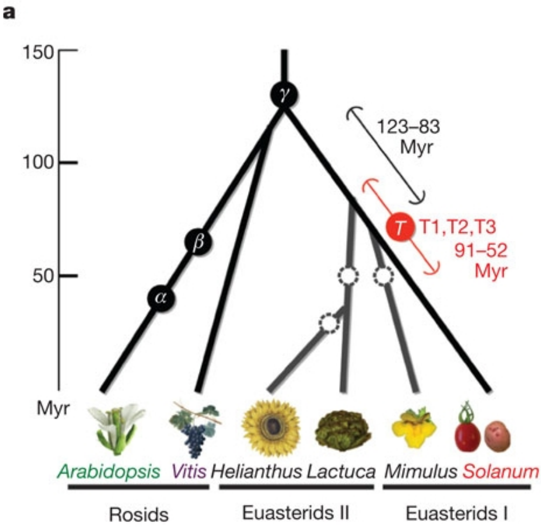
WGD in bananas

D'Hont et al. (Nature 2012)



WGD in tomatoes

Sato et al. (Nature, 2012)

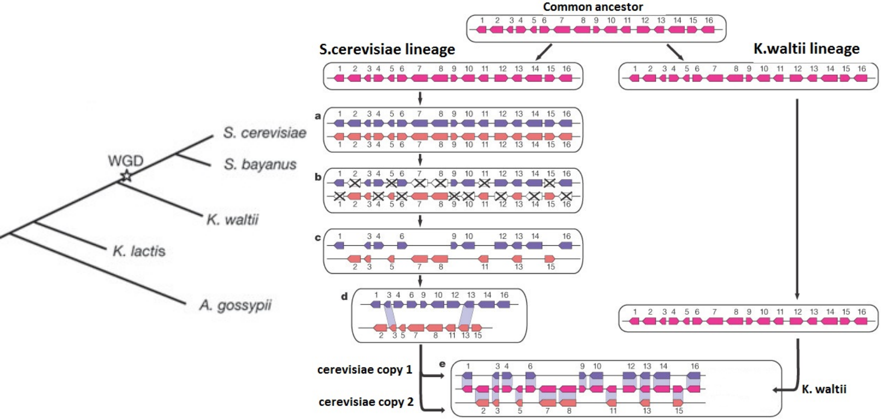


Traditional methods to detect WGDs

- Synteny-based method : search for synteny gene blocks in and between different genomes
- Age distribution-based method : infer the age of the different duplications (do not require positional informations on the paralogs)

Synteny-based methods (e.g. in yeast S.cerevisiae)

Kellis et al. (Nature, 2004) : 2 :1 mapping of syntenic blocks from *Saccharomyces cerevisiae* to *Kluyveromyces waltii*



Method sensitive to genome rearrangements and gene loss

Synteny-based methods (e.g. in yeast S.cerevisiae)

Kellis et al. (Nature, 2004)

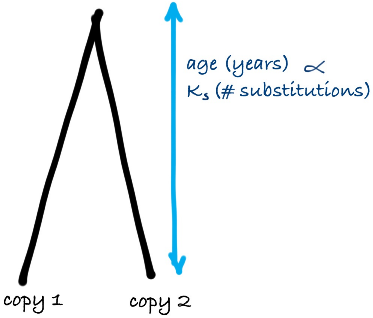
“S. cerevisiae genome is only 13% larger than K. waltii”

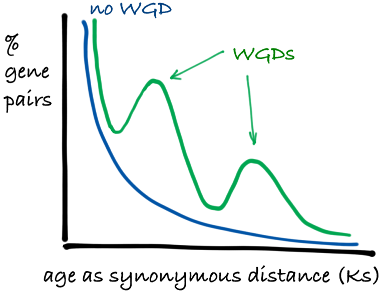
“We can infer that 12% of the paralogous genes pairs were retained in each DCS block, and the remaining 88% of paralogous genes were lost”

K_s-based methods

Duplication ages measured by synonymous distance
K_s : number of synomymous substitutions per synonymous site.

Using all pairs of paralogous genes, one genome :



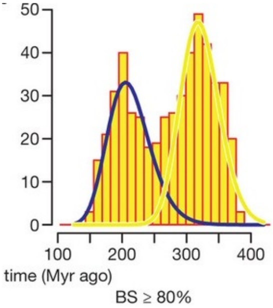


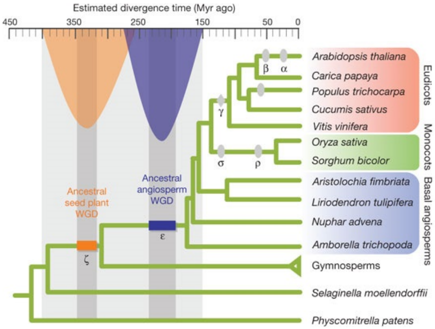
Limitation : K_s saturation for old duplicates

Age-based method on a phylogeny

Jiao et al. (Nature, 2011) :

- genes clustered into families (“gene family” = a set of genes with common or similar function)
- retained families with particular trees, with duplication prior to monocot-eudicot split
- mixture model





Probabilistic model for gene family evolution

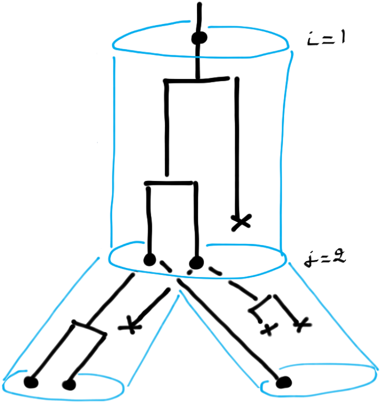
- phylogenetic framework : multiple species
- probabilistic model to avoid ad-hoc filtering of families or nodes
- requires : genes clustered into families. No synteny.

Birth-death model for small-scale events, and
WGD model for large-scale events.

$$\text{likelihood} = \prod_{\text{families } f} \text{likelihood}(f)$$

Birth - death process for small scale events

Data
(1 family)

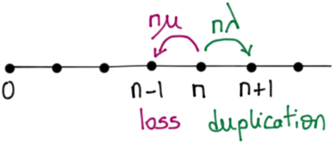


gene counts: X=2 X=1

sequences: a1=AACC b1=AATT

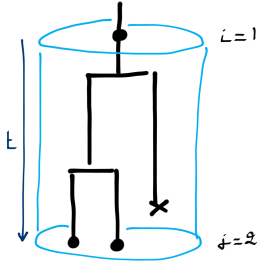
a2=AACT

Birth rate λ, death rate μ



Likelihood of gene counts, birth - death process

λ, μ : birth & death rates



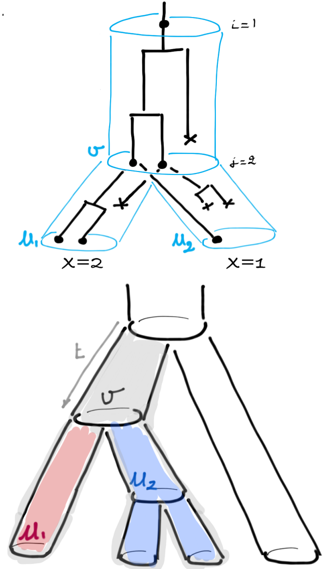
$$p_t(i,j) = \mathbb{P}\left(X_t = j | X_0 = i\right)$$

$$p_t(1,0) = \gamma_t = \frac{\mu\left(e^{(\lambda-\mu)t} - 1\right)}{\lambda e^{(\lambda-\mu)t} - \mu},$$
$$p_t(1,1) = (1-\gamma_t)(1-\psi_t) \text{ with } \psi_t = \frac{\lambda}{\mu}\gamma_t$$

$$p_t(i,j) = \sum_{k=0}^{i \wedge j} \binom{i}{k} \binom{i+j-k-1}{i-1} \gamma_t^{i-k} \psi_t^{j-k} (1-\gamma_t-\psi_t)^k$$

Bailey (1964)

Likelihood of gene counts, birth - death process



Conditional likelihood $L_v(i)$ at node v : probability of gene count data below v given $X = i$ at parent of v, calculated recursively :

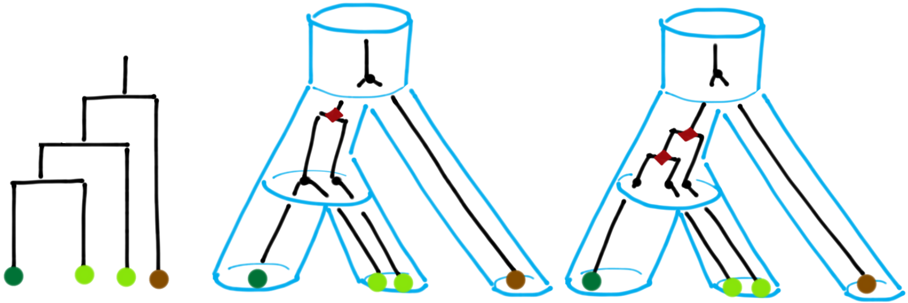
$$L_v(i) = \sum_j p_t(i,j) L_{u_1}(j) L_{u_2}(j)$$

Geometric prior π for # at the root :

$$\text{likelihood} = \sum_j \pi(j) L_{u_1}(j) L_{u_2}(j)$$

or Csűrös & Miklós (2009)

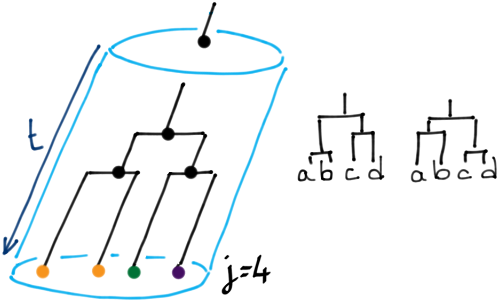
Likelihood of gene tree reconciliations, BD process



Problem 1 : each gene tree has many ”reconciliations” : to map gene tree inside species tree.

Likelihood of gene tree reconciliations, BD process

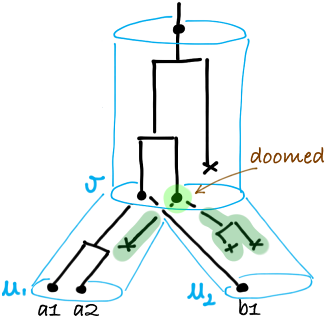
Problem 2 : labels

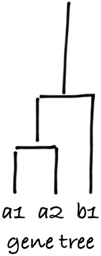


For a reconciled subtree within a ‘slice’, *j* tips, 3 colors

Arvestad et al. (2009), Rasmussen & Kellis (2011)

Likelihood of gene trees reconciliations, BD process





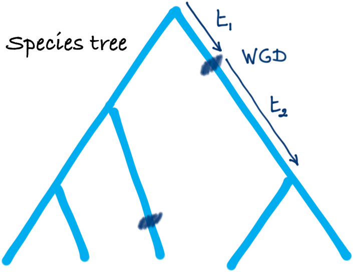
gene tree

Problem 3 : gene trees lack doomed lineages

d_v : probability that a lineage starting at node v leaves no descendent (or : is doomed). Recursively :

$$d_v = \left(\sum_j p_{t_1}(1,j) d^j_{u_1}\right)\left(\sum_j p_{t_2}(1,j) d^j_{u_2}\right)$$

WGD model for large-scale events



At the WGD :

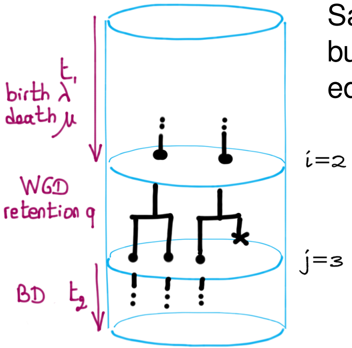
- each gene is duplicated
- second copy lost immediately with probability $1 - q$.

Each WGD has its own retention rate q, to explain :

- Large-scale events
- fragmentation : tendency to lose the extra copy,
  increased background loss rate
  shortly after WGD
- extension to whole genome triplications

Rabier, Ta, Ané (2014)

Likelihood : birth-death + WGD model

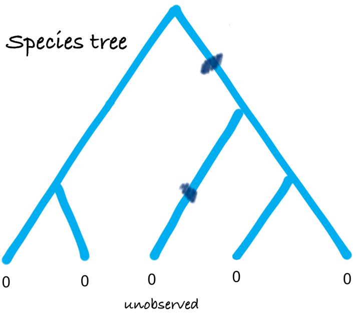


Same recursive algorithm through the tree, but new transition probabilities along WGD edges :

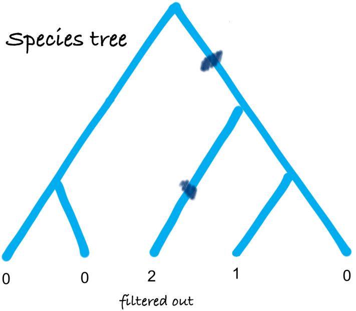
$$p_{\text{WGD}}(i,j) = \binom{i}{j-i} q^{j-i}(1-q)^{2i-j}$$

$$(i \leq j \leq 2i)$$

Conditioning on data collection process



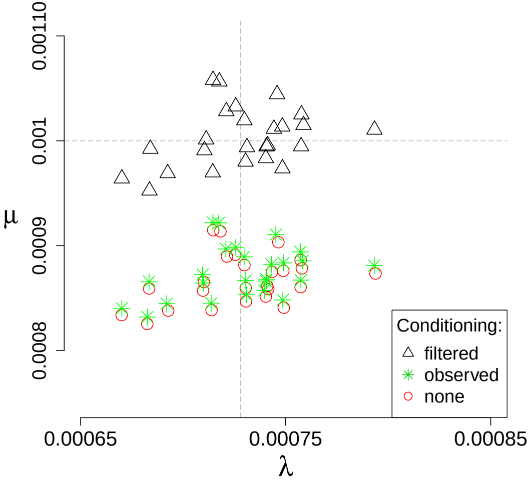
extinct families are unobservable



families with no gene in outgroup or ingroup species may be excluded (*de novo* or transferred genes)

Importance of conditioning

Simulated sets of 1000 gene families on 16-species yeast tree,
Families with 0 genes in ingroup or outgroup clades : excluded.
Birth & death rates (λ, μ) estimated from gene counts :



Two methods to detect WGDs

Using gene counts only :

- **fast** ($<$ 1s)
- exact likelihood
- optimize λ, μ and separate *q*'s at each WGD
- but : **limited** information

R package `WGDgc`

Two methods to detect WGDs

Using full sequences :

- **rich** information and model, but
- **slow** (e.g. 1h/family) : integrate over tree, reconciliation, branch lengths (gene-specific and species-specific rates).
- approximate likelihood
    - search over gene trees, but most parsimonious reconciliation.
    - new algorithm to find MP reconciliation with WGDs
- fixed $\hat{\lambda}, \hat{\mu}$

C++ program `spimapWGD`, based on `SPIMAP` (Rasmussen & Kellis 2011)

If you are interested in the gene tree ...

Some notations

- S : species tree
- D : data (ie. alignment data)
- T : gene tree topology
- ℓ : branch length
- R : reconciliation

Bayesian framework

- $\mathbb{P}(T, R|S)$: topology prior
- $\mathbb{P}(\ell|T, R, S)$: branch length prior
- $\mathbb{P}(T, R, \ell|D, S)$: posterior

$\Rightarrow$ Markov Chain Monte Carlo (Hasting Metropolis) to estimate posterior distribution $\mathbb{P}(T, R, \ell|D, S)$

Approximate versus exact likelihood

Exact Likelihood

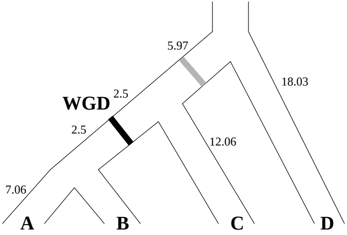
$$\begin{aligned}\mathbb{P}(D|S) &= \sum_{T,R} \int_{l} \mathbb{P}(D, l, T, R|S) \\ &= \sum_{T,R} \int_{l} \mathbb{P}(D|l, T, S)\, \mathbb{P}(l|T, R, S)\, \mathbb{P}(T, R|S)\end{aligned}$$

Approximate Likelihood

$$\mathbb{P}(D|S) \approx \mathbb{P}(D, \hat{\ell}, \hat{T}, \hat{R}|S)$$

with $\hat{\ell}$, $\hat{T}$, $\hat{R}$ maximum a posteriori estimators of ℓ, T, R given the data

Performance on simulated data



20,000 families per replicate
$\lambda = .02$, $\mu = .03$
500-bp sequences

- using gene counts : R package `WGDgc`
- using full sequences : C++ program `spimapWGD`, based on `SPIMAP` (Rasmussen & Kellis 2011)

Our simulation framework for the reconciliation method

- Equal base frequencies (Jukes-Cantor)
- Data simulated either under no WGD, or with WGD (true retention rate $q = 0.2$, 0.5 or 0.9)
- 20000 gene families
- Each gene family analyzed 11 times ($q = 0, q = 0.1, ..., q = 1$), in order to try the different retention rates

⇒ 220000 jobs = 75 years completed in 2 days using the high throughput computing ressources with Condor, Open Science Grid.

The Condor team

Where are my Condor jobs running ?

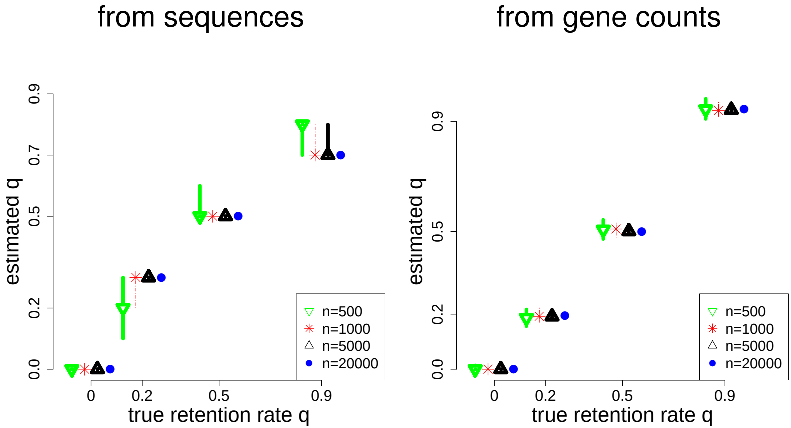
```
>condor q -run rabier

10505346.0 rabier glidein10012@ iut2-c086.iu.edu
10505347.0 rabier glidein4215@ compute-2-1.nys1
10505348.0 rabier glidein2561@ iut2-c048.iu.edu
10505349.0 rabier slot1@ wid-exec-1.chtc.wisc.edu
10505353.0 rabier glidein15691@hansen-a003.rcac.purdue.edu
10505354.0 rabier glidein25903@node254.red.hcc.unl.edu
10505355.0 rabier glidein11128@ acas0584.usatlas.bnl.gov
10505356.0 rabier glidein9966@ node198.red.hcc.unl.edu
```

Indiana university, Cornell university, University of Wisconsin, Purdue university, university Nebraska-Lincoln, Brookhaven national lab ....

Some areas of application of Condor

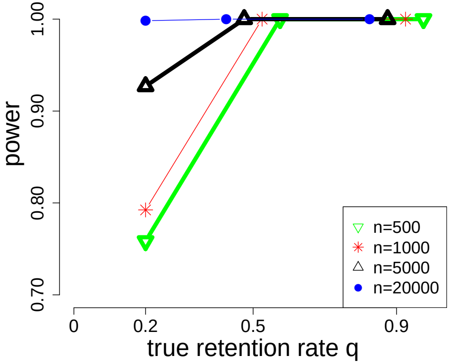
Peter Higgs

Estimation of retention rate *q*



Power to detect the WGD

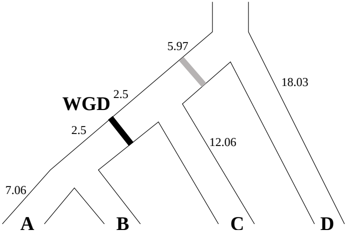
from sequences



from gene counts

100%
from $q \geq 0.2$ and
$\geq$500-gene families

WGD location



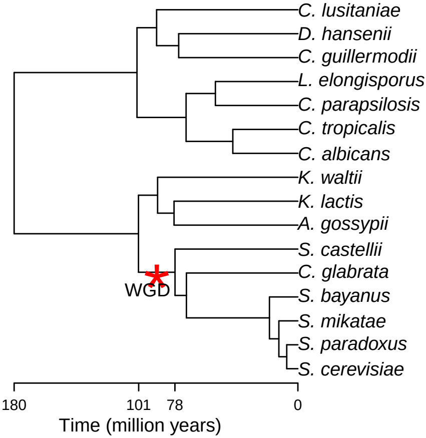
With uncertain location of WGD : likelihood maximized over two hypothesized edges.

When detected, the WGD location was correctly estimated.

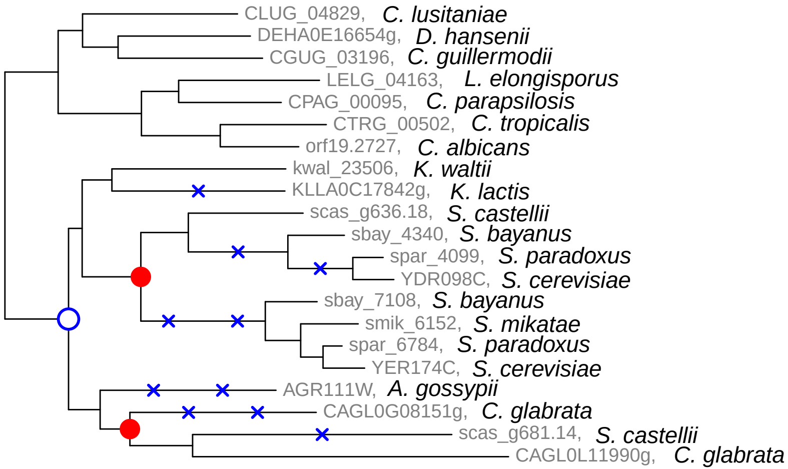
Yeast genome evolution

Kellis et al. (2004), from synteny on *Kluyveromyces waltii* and *S. cerevisiae* : "12% of the paralogous gene pairs were retained in each doubly conserved synteny block"

- 9209 gene families (Butler et al 2009)
- filter : 3932 families with ≥ 1 gene in both *Candida* and *Saccharomyces* subclades

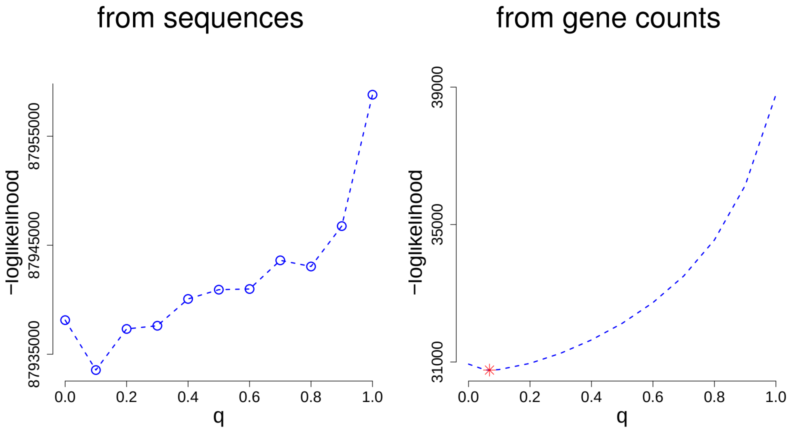


A phylogenetic tree of gene family 1306


CLUG_04829, C. lusitaniae
DEHA0E16654g, D. hansenii
CGUG_03196, C. guillermodii
LELG_04163, L. elongisporus
CPAG_00095, C. parapsilosis
CTRG_00502, C. tropicalis
orf19.2727, C. albicans
kwal_23506, K. waltii
KLLA0C17842g, K. lactis
scas_g636.18, S. castellii
sbay_4340, S. bayanus
spar_4099, S. paradoxus
YDR098C, S. cerevisiae
sbay_7108, S. bayanus
smik_6152, S. mikatae
spar_6784, S. paradoxus
YER174C, S. cerevisiae
AGR111W, A. gossypii
CAGL0G08151g, C. glabrata
scas_g681.14, S. castellii
CAGL0L11990g, C. glabrata

2 duplications at the WGD (red circles), 0 loss at the WGD
1 duplication, 10 losses (blue crosses)

Testing the Yeast WGD



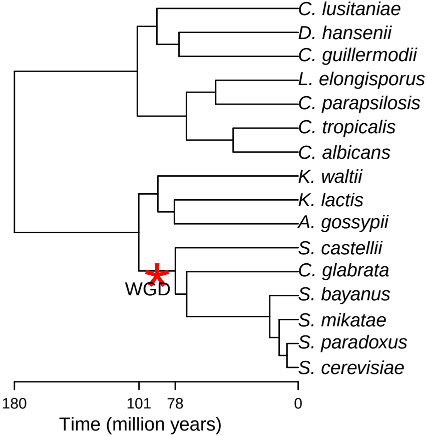
LRT : 9159.5

LRT : 348.1

retention rate : $\hat{q} = 6.81\%$, in $[0.058, 0.079]$ with 95% confidence

Yeast WGD timing

$\hat{t} = 0$: immediately before speciation,
$\hat{t} \leq 5.04$ My with 95% confidence.



Caveats

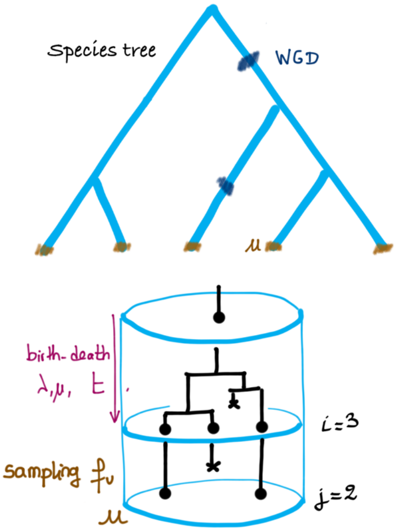
- variation in background duplication/loss rates across **families**
- errors in species tree branch **lengths**
- errors in gene count **data**, e.g. from low-coverage genomes or transcriptomes

Extension : error model for gene counts

Incompletely sampled genomes : sampling frequency f_u for species u. transition probability, extra edge at u :

$$p_u^{\text{sampling}}(i,j) = \binom{i}{j} f_u^j (1-f_u)^{i-j}$$

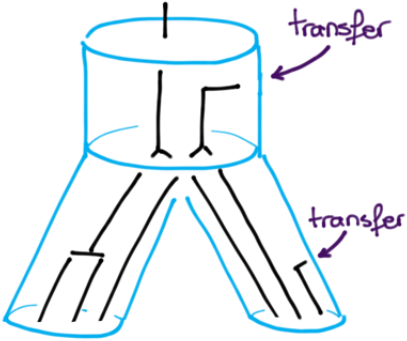
Error models for assembly and clustering errors : Han et al. (2013)



Extension : gene transfers

Include gene transfers : duplication-loss-**gain** process, or duplication-**transfer**-loss.

Csűrös & Miklós (2009) : rates λ, μ and κ.



Thanks to

Cécile Ané
Tram Ta

Matt Rasmussen
Bill Taylor

NSF

DEB-0949121

HTCondor
High Throughput Computing