Reviews in
Computational Biology

Inferring Lateral Gene Transfer

Christophe Dessimoz

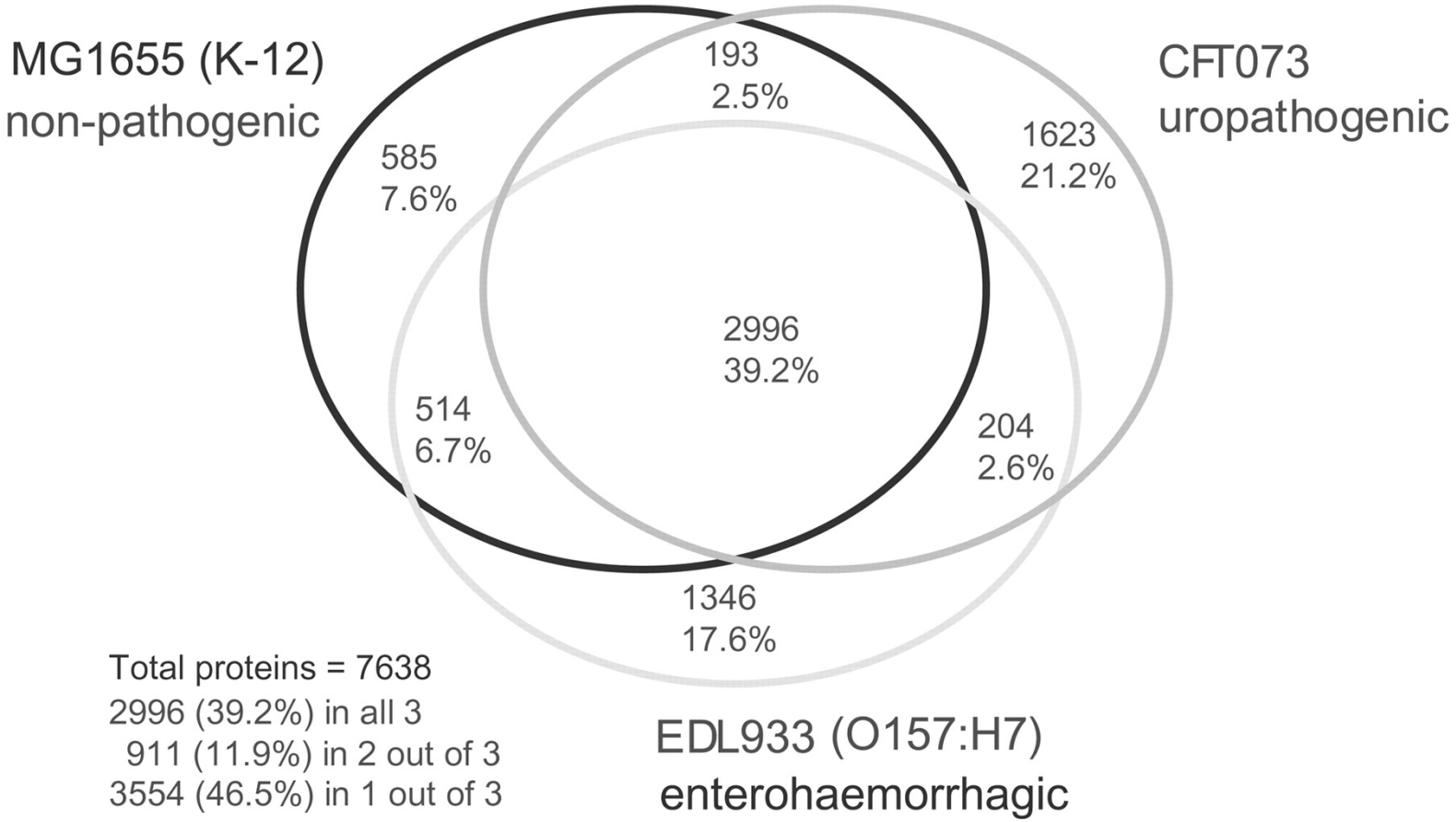
March 28th, 2011

ETH Zürich inf | Informatik
Computer Science

Structure

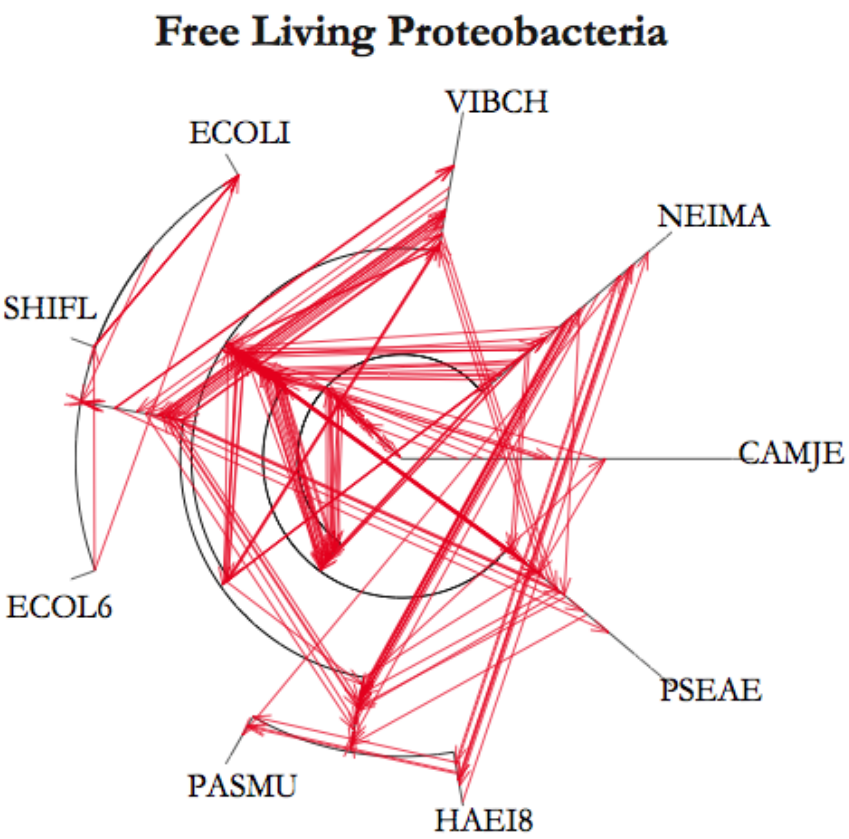
- **Introduction**
- **Methods to Detect Lateral Gene Transfer**
  - Parametric
  - Phylogenetic
    - Explicit
    - Implicit
- **Outlook**

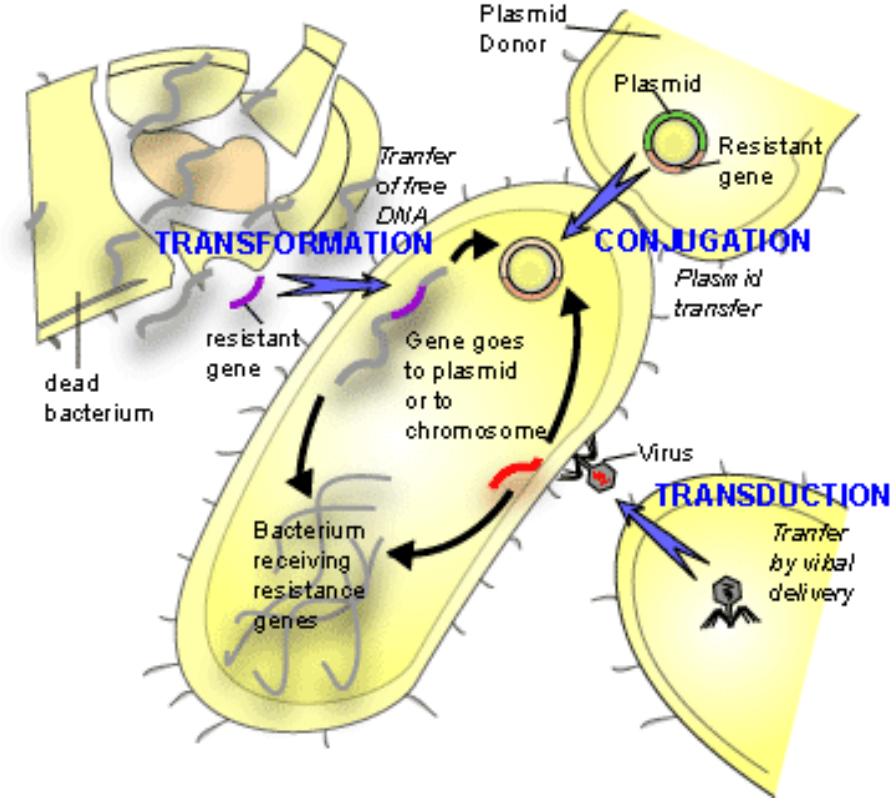
Shared E. coli proteins.



Welch R A et al. PNAS 2002;99:17020-17024

Lateral Gene Transfer (LGT)



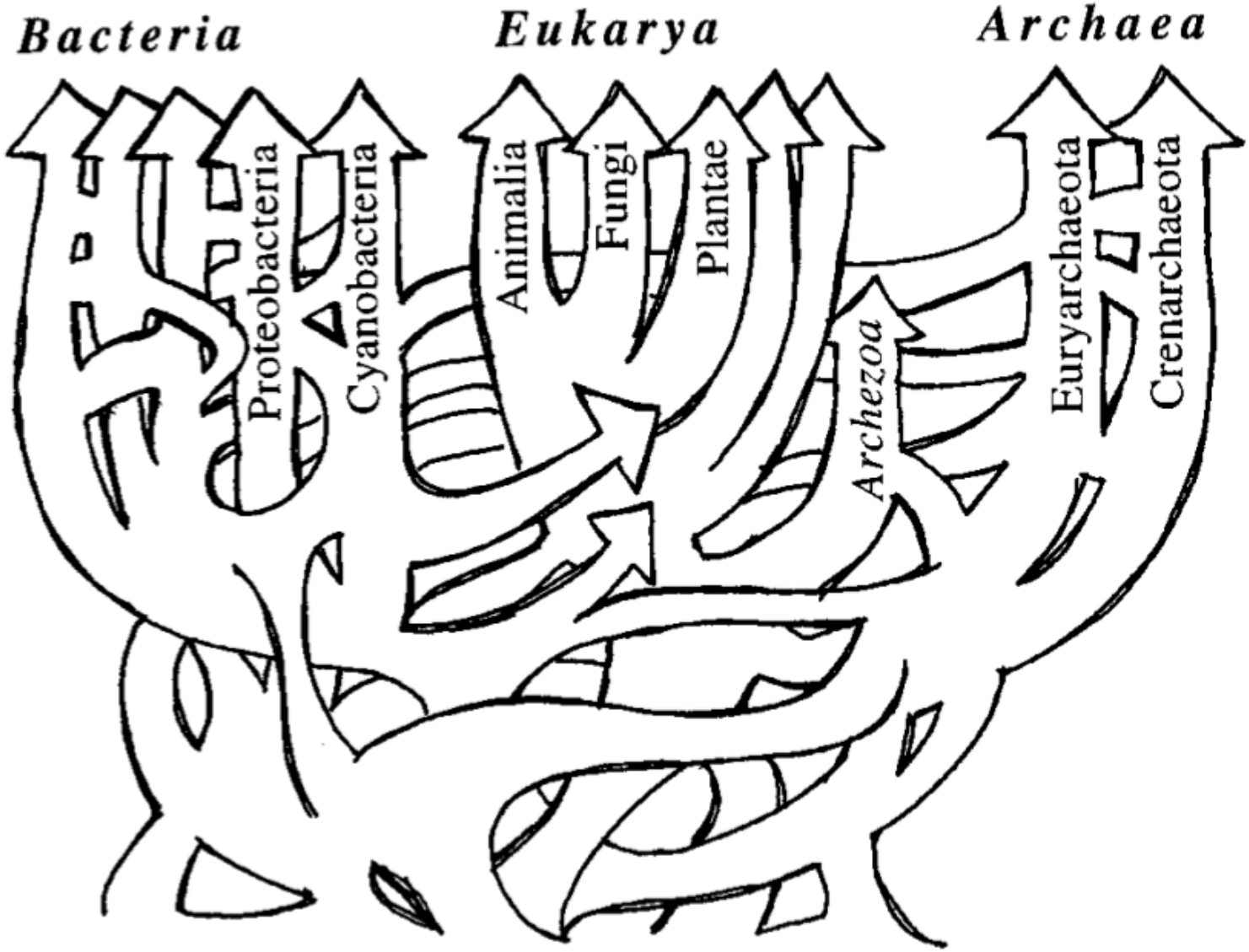


http://www.scq.ubc.ca/attack-of-the-superbugs-antibiotic-resistance/

Types of LGTs

- novel gene acquisition
- orthologous gene replacement

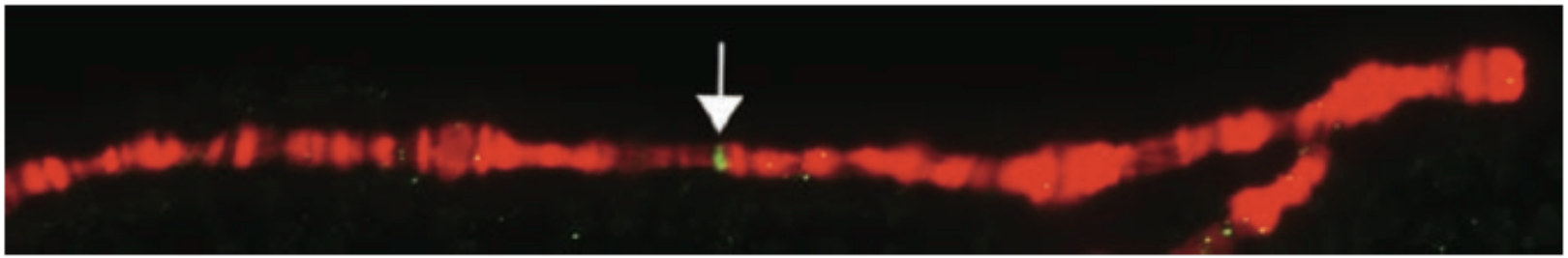
Phylogenetic Classification and the Universal Tree
W. Ford Doolittle, *et al.*
Science **284**, 2124 (1999);



Science VOL 317 21 SEPTEMBER 2007

Widespread Lateral Gene Transfer from Intracellular Bacteria to Multicellular Eukaryotes

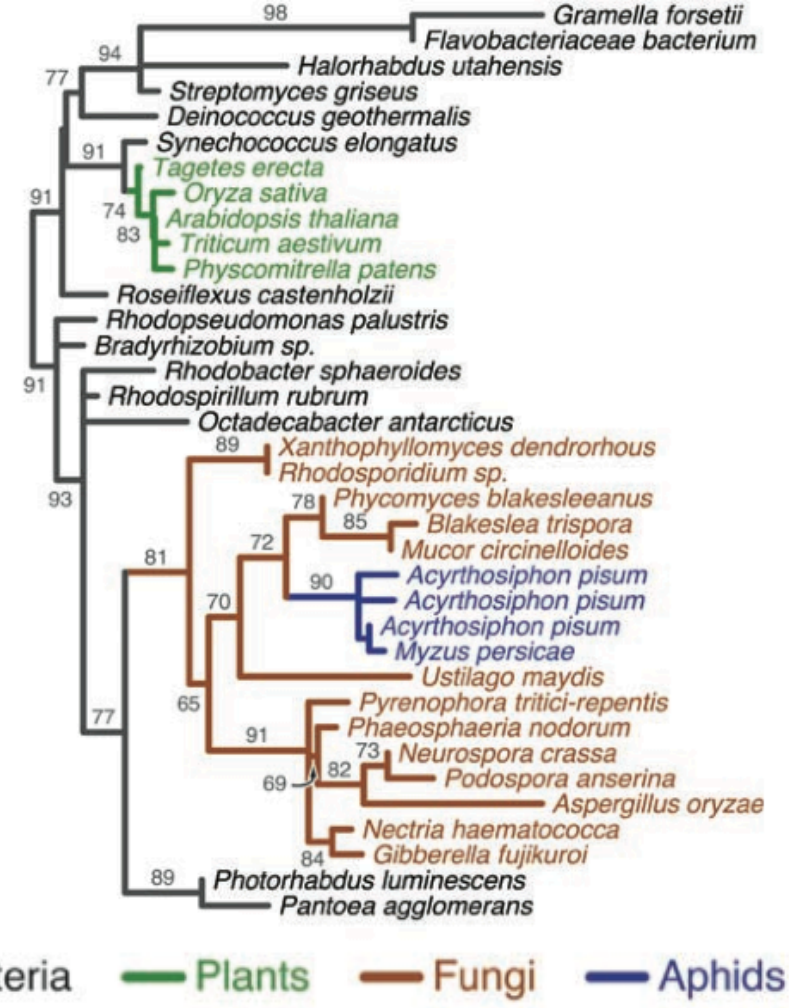
Julie C. Dunning Hotopp,[1]*†‡ Michael E. Clark,[2]* Deodoro C. S. G. Oliveira,[2] Jeremy M. Foster,[3] Peter Fischer,[4] Mónica C. Muñoz Torres,[5] Jonathan D. Giebel,[2] Nikhil Kumar,[1]‡ Nadeeza Ishmael,[1]‡ Shiliang Wang,[1] Jessica Ingram,[3] Rahul V. Nene,[1]§ Jessica Shepard,[1]‖ Jeffrey Tomkins,[5] Stephen Richards,[6] David J. Spiro,[1] Elodie Ghedin,[1,7] Barton E. Slatko,[3] Hervé Tettelin,[1]‡¶ John H. Werren[2]¶



Science
AAAS
30 APRIL 2010 VOL 328

Lateral Transfer of Genes from Fungi Underlies Carotenoid Production in Aphids

Nancy A. Moran[1*] and Tyler Jarvik[2]



Overview

Parametric methods (based on genome signatures)

- CG-content
- Codon bias
- k-nucleotide frequencies

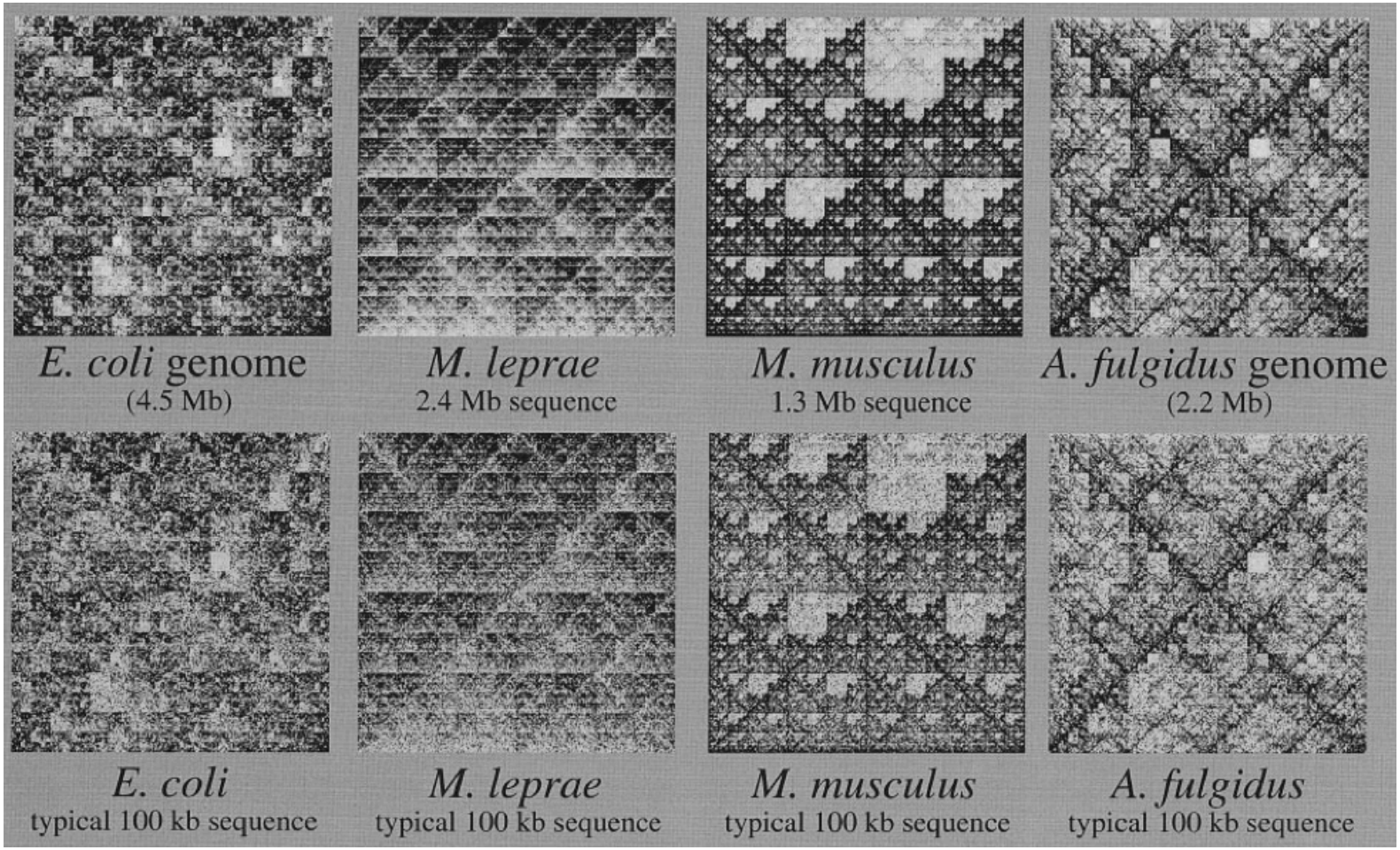
Phylogenetic methods

- Explicit: tree reconciliation
- Implicit: based on underlying tree, but no tree inference/reconciliation

Parametric Methods

Genome signatures

Representation of 7-nucleotide frequencies



Mol. Biol. Evol. 16(10):1391–1399. 1999

Deschavanne et al.

nature genetics VOLUME 36 | NUMBER 7 | JULY 2004

Biased biological functions of horizontally transferred genes in prokaryotic genomes

Yoji Nakamura[1,5], Takeshi Itoh[2,3], Hideo Matsuda[4] & Takashi Gojobori[1,2]

$$P(COD_1|F) = \frac{P(F|\,COD_1)\,P(COD_1)}{\sum_{m=1}^{6} P(F|\,COD_m)\,P(COD_m) + P(F|\,NON)\,P(NON)}$$

1/12 (numerator $P(COD_1)$); 1/12 (denominator $P(COD_m)$); 1/2 (denominator $P(NON)$)

$(m = 1, 2, 3, 4, 5, 6)$.

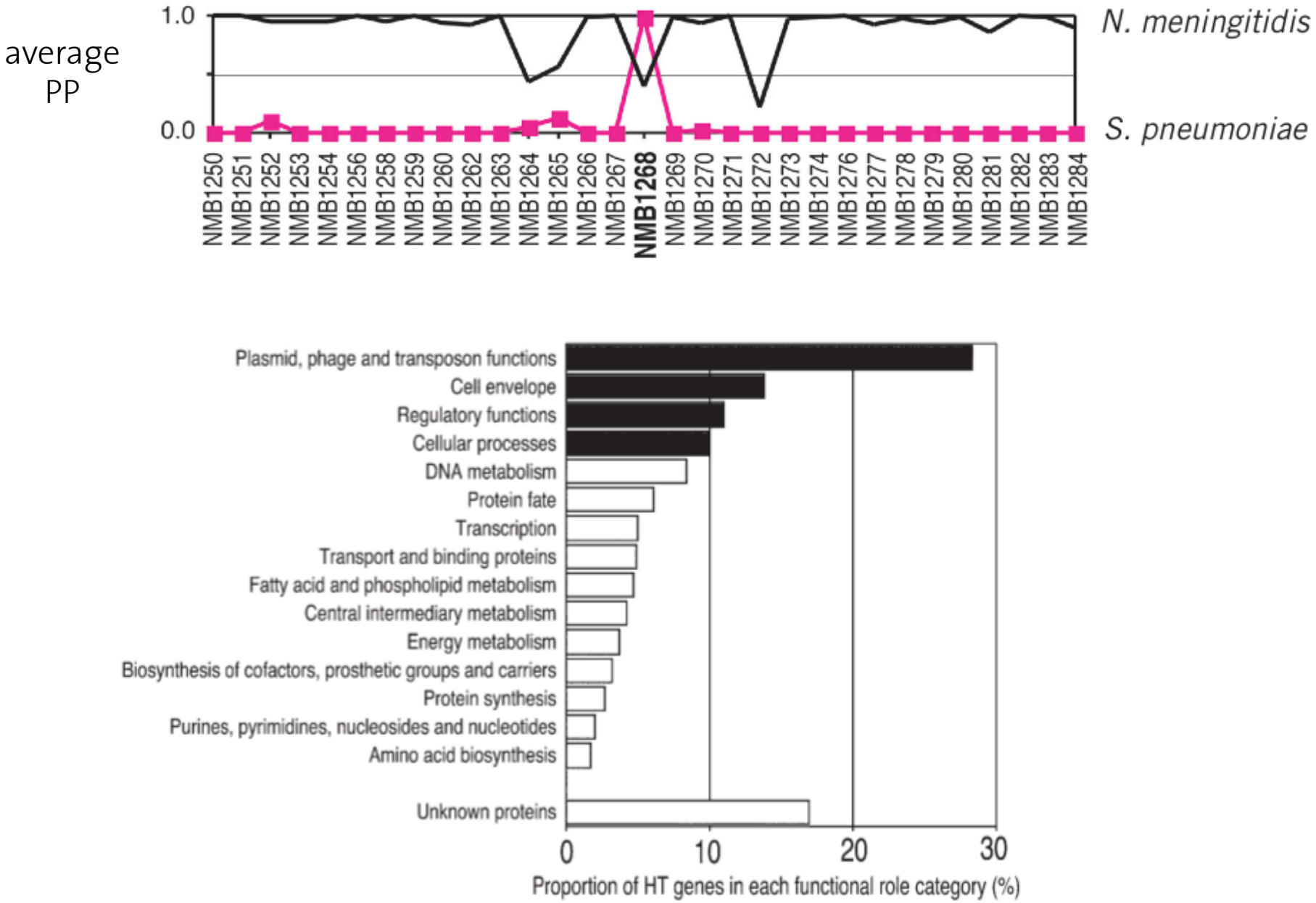
reading frames

F: hexamer

COD$_i$: coding in frame i

NON: non-coding

The Horizontal Transfer Index is the average $P(COD_1|F)$ value over a window of size 96 bp, slid on the gene sequence by a step of 12 bp



Nakamura et al. Nat Gen 2004

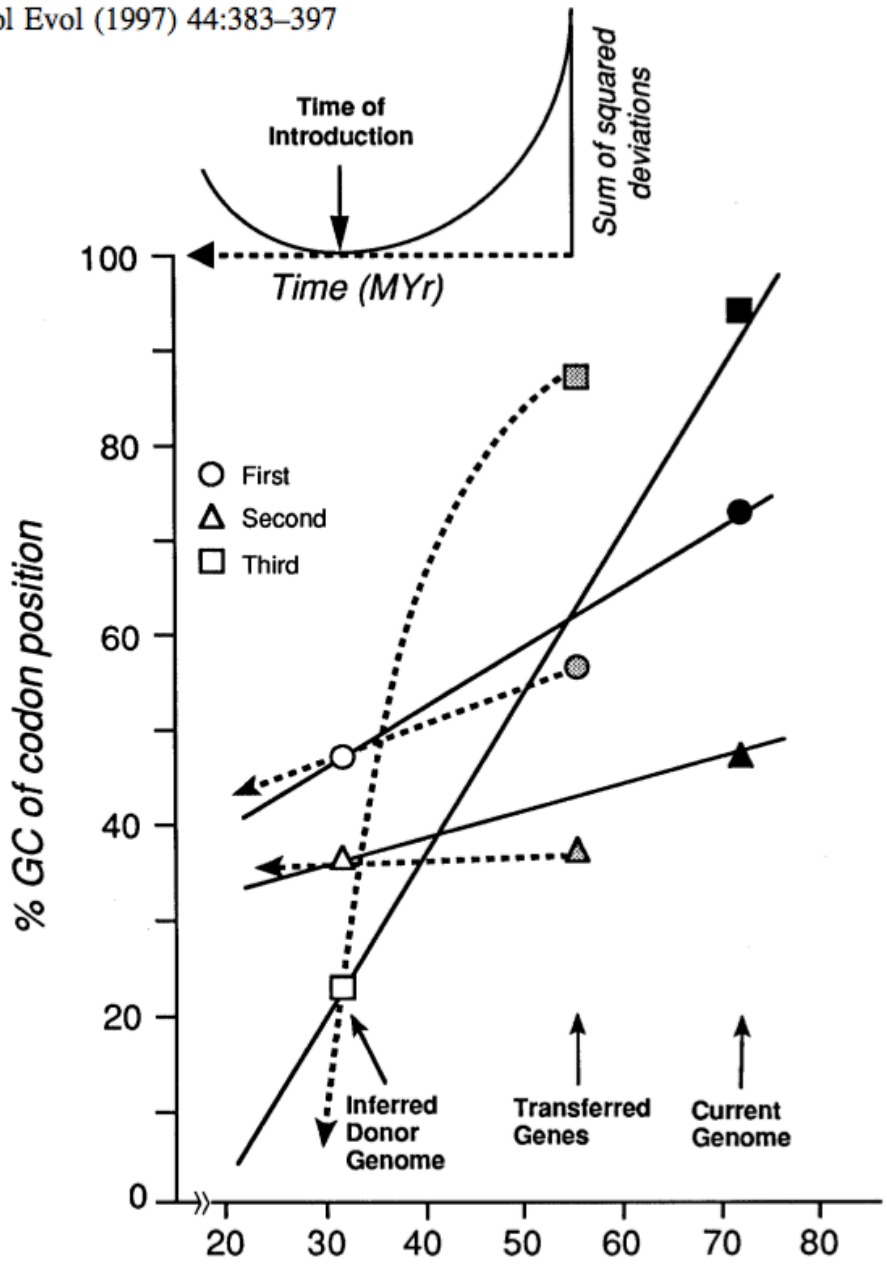
Parametric Methods: Limitations

- "Amelioration": adjustment of laterally transferred gene to the nucleotide composition of its new host.
  → restricted to relatively recent transfers
- Donor species are difficult to identify
- Only works for genes transferred from species with significantly different parameters

Amelioration of Bacterial Genomes: Rates of Change and Exchange

Jeffrey G. Lawrence,[1]* Howard Ochman[2]

J Mol Evol (1997) 44:383–397



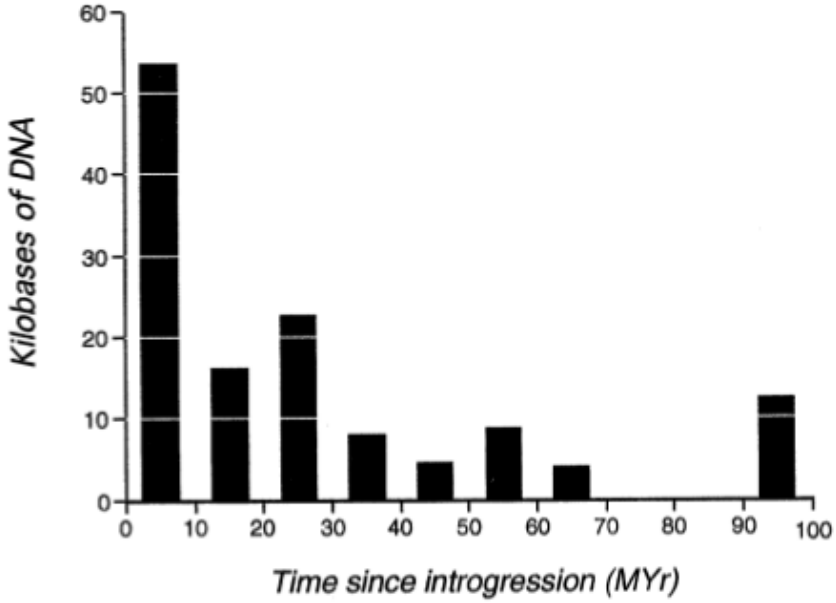


Fig. 7. Times of introgression of DNA fragments into the *E. coli* chromosome. The amount of DNA reflects cumulative pooled sequences (see Table 3).

Phylogenetic Methods

Phylogenetic Methods:
Overview

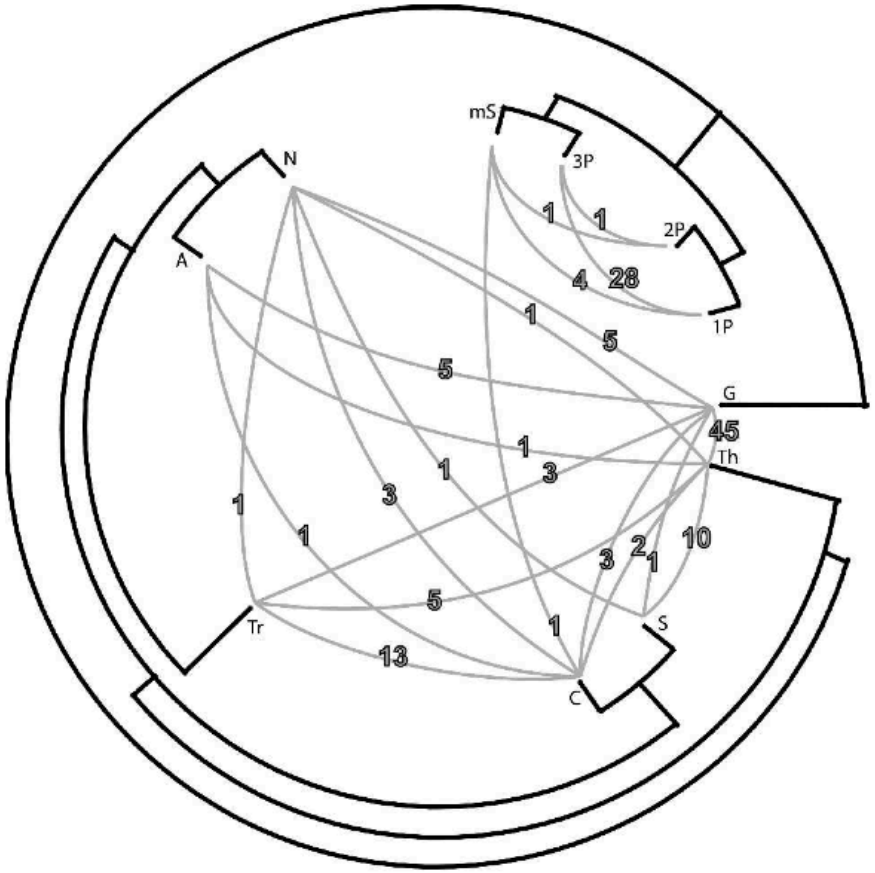
- **Explicit (reconstruct trees)**
  - Discordance
  - Model-based
- **Implicit (rely indirectly on tree)**

Discordance

CSH PRESS GENOME RESEARCH

Phylogenetic analyses of cyanobacterial genomes: Quantification of horizontal gene transfer events

Olga Zhaxybayeva, J. Peter Gogarten, Robert L. Charlebois, et al.



(plurality signal = "species tree")

But (5-30% FP on simulation w.o. LGT)

are very probable candidates for HGT. (At the same time, our simulation study shows how frequently false positives arise, and casts a shadow on the reliability of phylogenetic reconstruction in general.)

Simulations

We performed simulations of genome evolution using EvolSimulator (http://bioinformatics.org.au/evolsim/). Seven hundred genes were simulated for 10,000 generations in a dynamic population of genomes, with speciation balancing extinction (at a nominal rate of 0.015 events per generation) to maintain ~50 extant lineages at any given time following a brief initial phase of population growth. Disabling paralogous duplication as well as gene loss, each genome maintained exactly 700 genes whose orthology could thus be perfectly tracked. Parameters affecting se-

14332–14337 | PNAS | October 4, 2005 | vol. 102 | no. 40

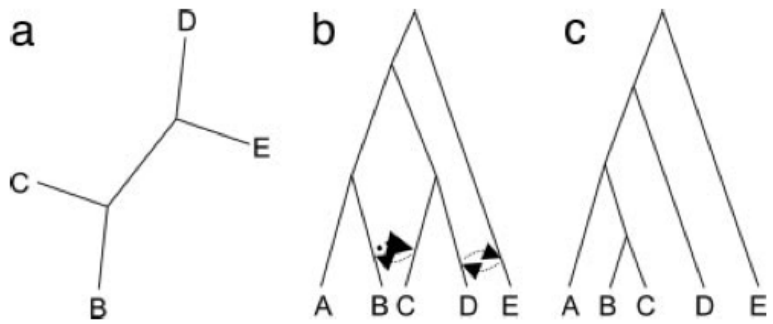
Highways of gene sharing in prokaryotes

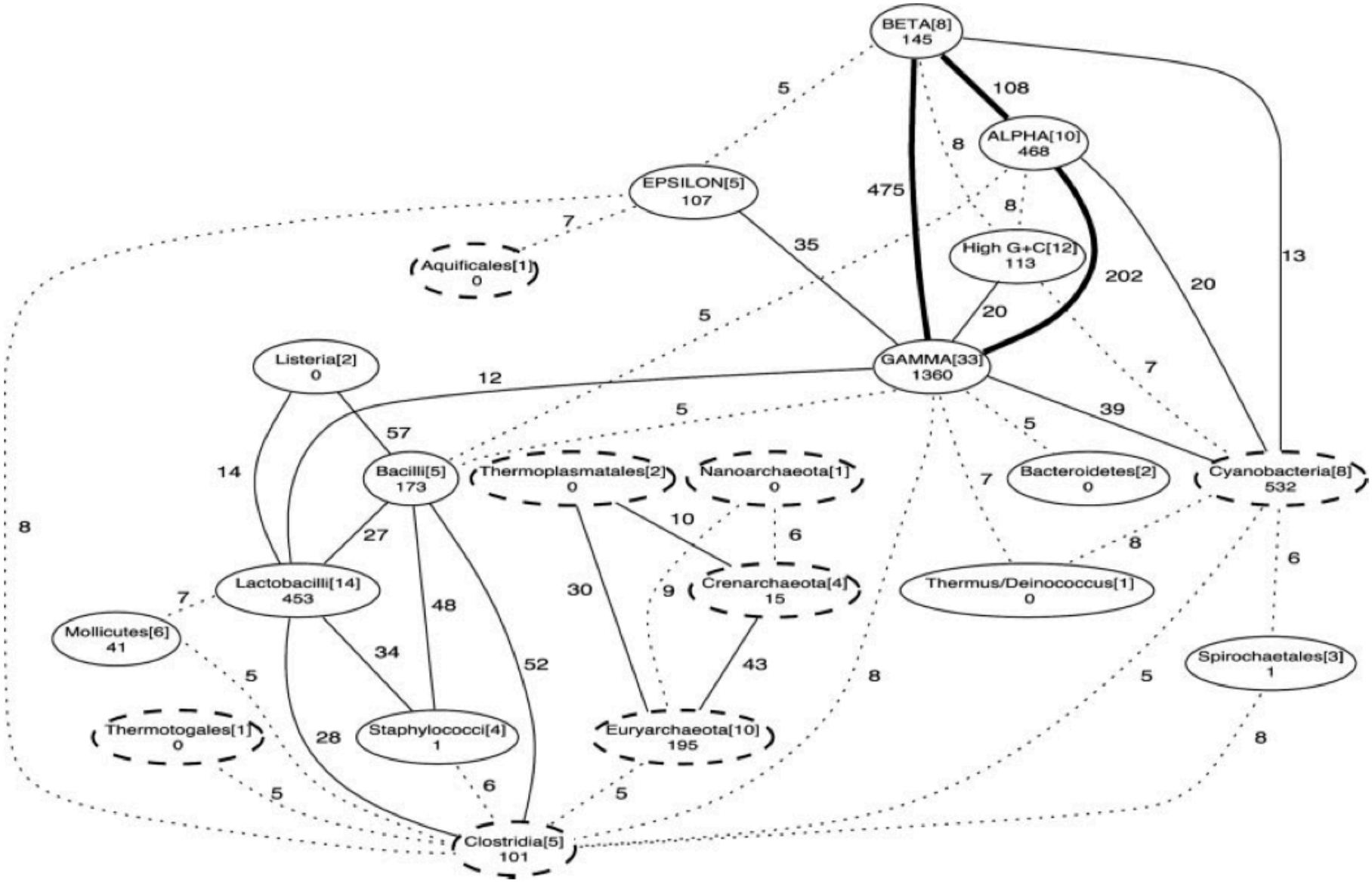
Robert G. Beiko, Timothy J. Harlow, and Mark A. Ragan*

- Considered 144 prokaryotes
- Built Bayesian trees for 22,437 families
- Combined them into a supertree
- Incongruence between bipartitions with high posterior probability and supertree suggests LGT:

> support a regime of vertical inheritance at that node. Protein trees that are strongly incongruent with a supertree node are discordant, and provide *prima facie* evidence of LGT. Of the

- Heuristic to find SPR-moves to explain incongruences





Beiko et al. Highways of gene sharing in prokaryotes. PNAS (2005)

Explicit, Model-Based

- **Usually based on subtree prunning and regraft (SPR)**
- **If only interested in the "prunning" -> Maximum Agreement Forest Problem** (smallest number of edges to cut in 2 trees to yield two identical forests of rooted subtrees)
- **SPR distance between two unrooted binary trees is NP-hard**
  → practical algorithms are approximations

Simultaneous Identification of Duplications and Lateral Transfers

[Extended Abstract]

Mike Hallett
McGill Centre for Bioinformatics
3775 University Ave.
Montréal, Québec, Canada
hallett@mcb.mcgill.ca

Jens Lagergr
Stockholm Bioinfor
Centre
Dept. of Numerical A
and Computer Scien
Stockholm, Swe
jensl@nada.kt

Integrating Sequence and Topology for Efficient and Accurate Detection of Horizontal Gene Transfer

Cuong Than, Guohua Jin, and Luay Nakhleh*

Inferring and Validating Horizontal Gene Transfer Events Using Bipartition Dissimilarity

ALIX BOC[1], HERVÉ PHILIPPE[2], AND VLADIMIR MAKARENKOV[1,*]

Detecting Highways of Horizontal Gene Transfer

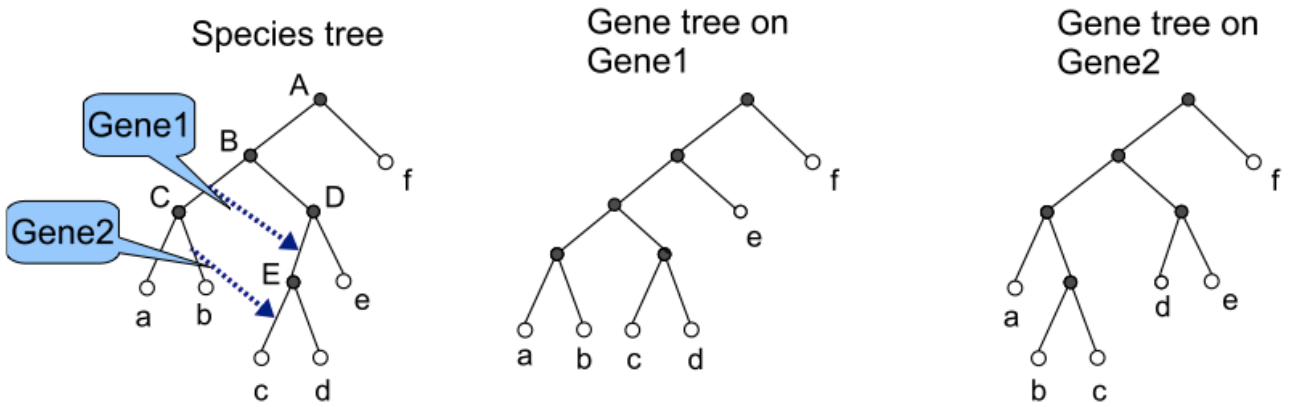
Mukul S. Bansal[1], J. Peter Gogarten[2], and Ron Shamir[1]

Detecting lateral gene transfers by statistical reconciliation of phylogenetic forests

Sophie S Abby, Eric Tannier, Manolo Gouy and Vincent Daubin*

Detecting Highways of Horizontal Gene Transfer

Mukul S. Bansal[1], J. Peter Gogarten[2], and Ron Shamir[1] (2011)



Step 1: Decompose each input gene tree T into its constituent set of $\binom{|Le(T)|}{4}$ quartet trees, and combine the quartet trees from the different gene trees into a single weighted set, Φ, of quartet trees.

Step 2: Remove from Φ all those quartet trees that are consistent with S.

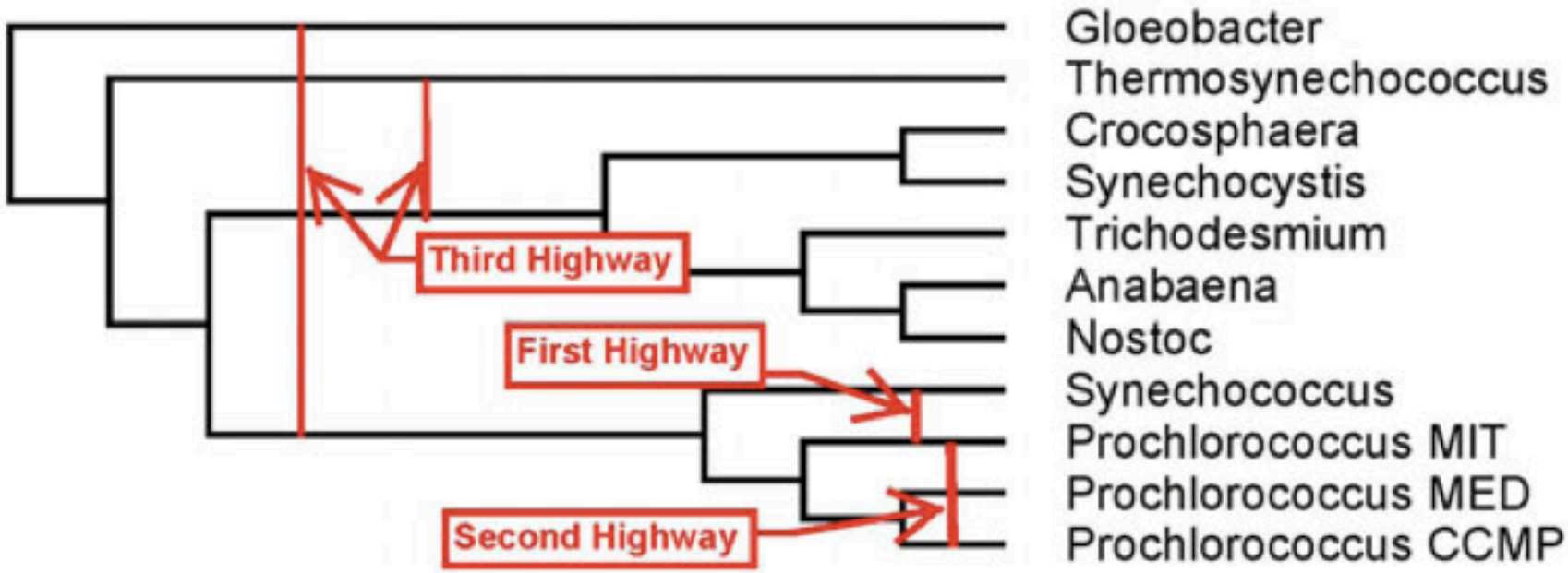
Step 3: Compute the HGT score of each edge in $H(S)$. This HGT score for an edge is computed based on Φ, and is explained in detail below.

Step 4: Select the highest scoring horizontal edge as a highway.

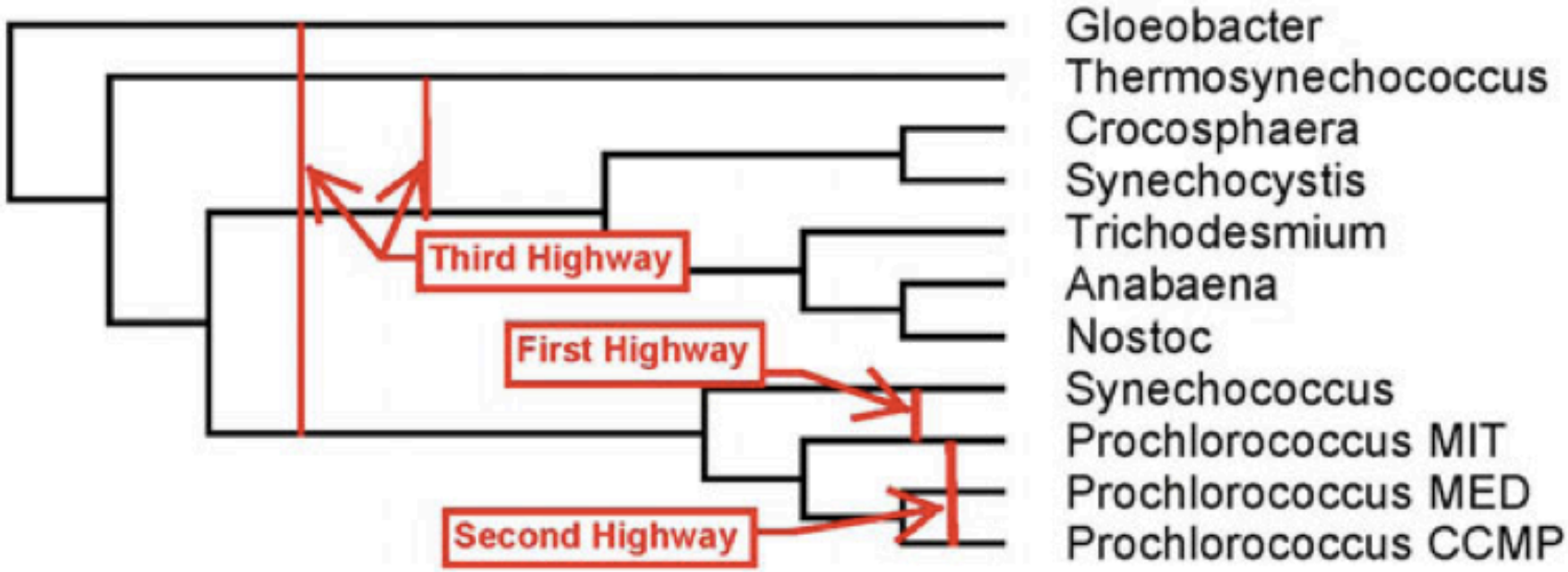
Step 5: Remove from Φ all those quartet trees that are explained by the proposed highway, and go to Step 3 to start the next iteration.

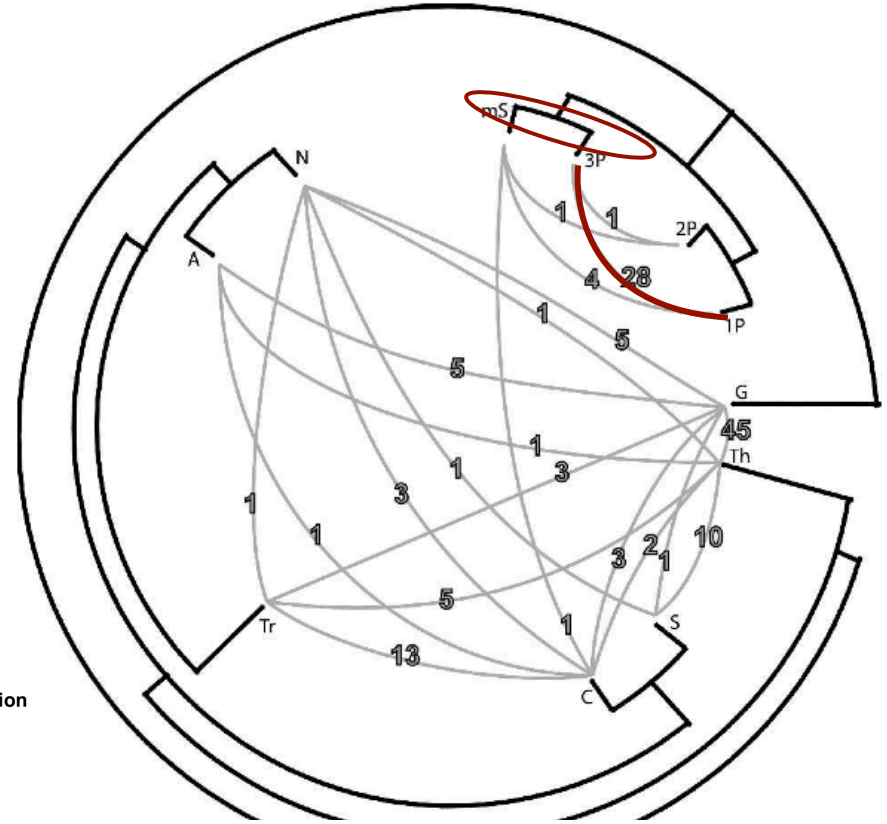
Detecting Highways of Horizontal Gene Transfer

Mukul S. Bansal[1], J. Peter Gogarten[2], and Ron Shamir[1]



(data from Zhaxybayeva et al. 2006)





GENOME RESEARCH

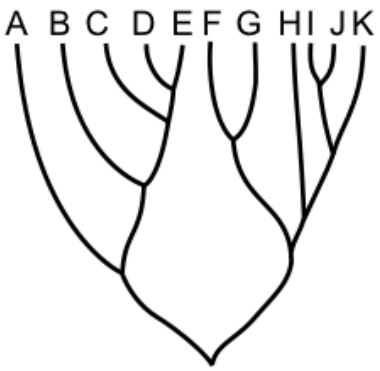
Phylogenetic analyses of cyanobacterial genomes: Quantification of horizontal gene transfer events

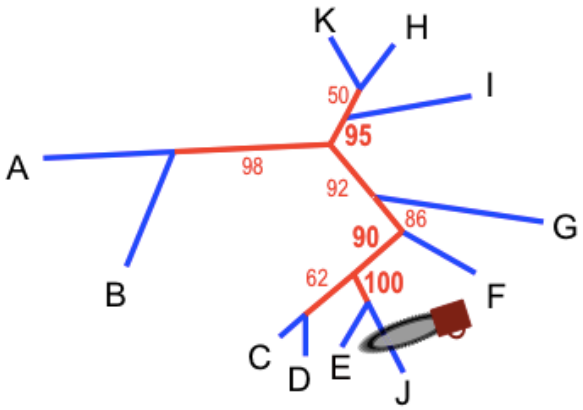
Olga Zhaxybayeva, J. Peter Gogarten, Robert L. Charlebois, et al.

Detecting lateral gene transfers by statistical reconciliation of phylogenetic forests

Sophie S Abby, Eric Tannier, Manolo Gouy and Vincent Daubin* (2010)

Species tree S ***Gene tree G***





Prunier 1st round

Agree(*TS*, *TG*) => NO

Sort candidate transfers (common edges):

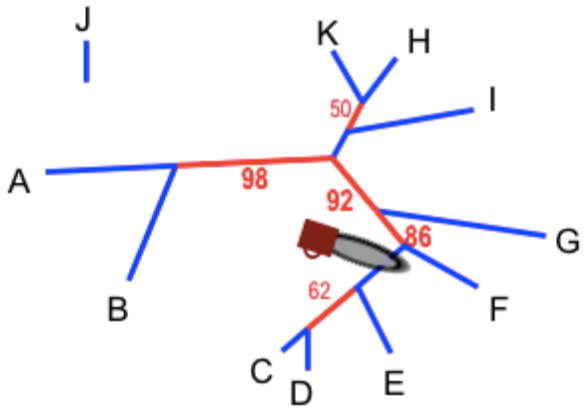
1) Rm{T(J)} discards: **100** ; **95** ; **90**
2) Rm{T(E)} discards: 100
3) Rm{*T(A)*} or {*T(B)*} discards: 98
4) Rm{*T(C)*} or {*T(D)*} discards: 62
5) Rm{*T(K)*} or {*T(H)*} discards: 50

Agree{*TS(J)*, *TG(J)*} => YES
Prune the highest scoring candidate: *T(J)*

Detecting lateral gene transfers by statistical reconciliation of phylogenetic forests

Sophie S Abby, Eric Tannier, Manolo Gouy and Vincent Daubin*

Species tree S ***Gene tree G***



Prunier 2nd round

Agree(*TS*, *TG*) => NO

Sort candidate transfers (common edges):

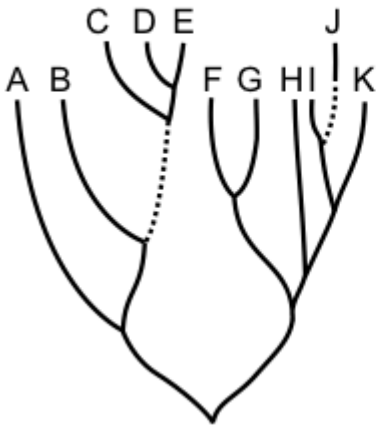
1) Rm{*T(CDE)*} discards: **98** ; **92** ; **86**
2) Rm{*T(A)*} or {*T(B)*} discards: 98
3) Rm{*T(C)*} or {*T(D)*} or {*T(E)*} discards: 62
4) Rm{*T(K)*} or {*T(H)*} or {*T(I)*} discards: 50

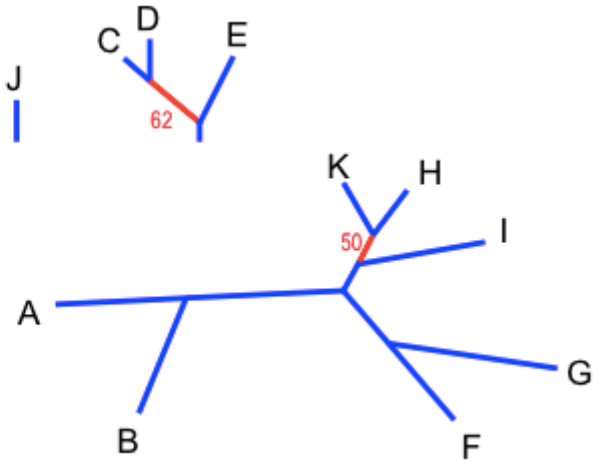
Agree{*TS(CDE)*, *TG(CDE)*} => YES
Prune the highest scoring candidate: *T(CDE)*

Detecting lateral gene transfers by statistical reconciliation of phylogenetic forests

Sophie S Abby, Eric Tannier, Manolo Gouy and Vincent Daubin*

Species tree S ***Gene tree G***





Prunier 3rd round

Agree(*TS*, *TG*) => **YES**

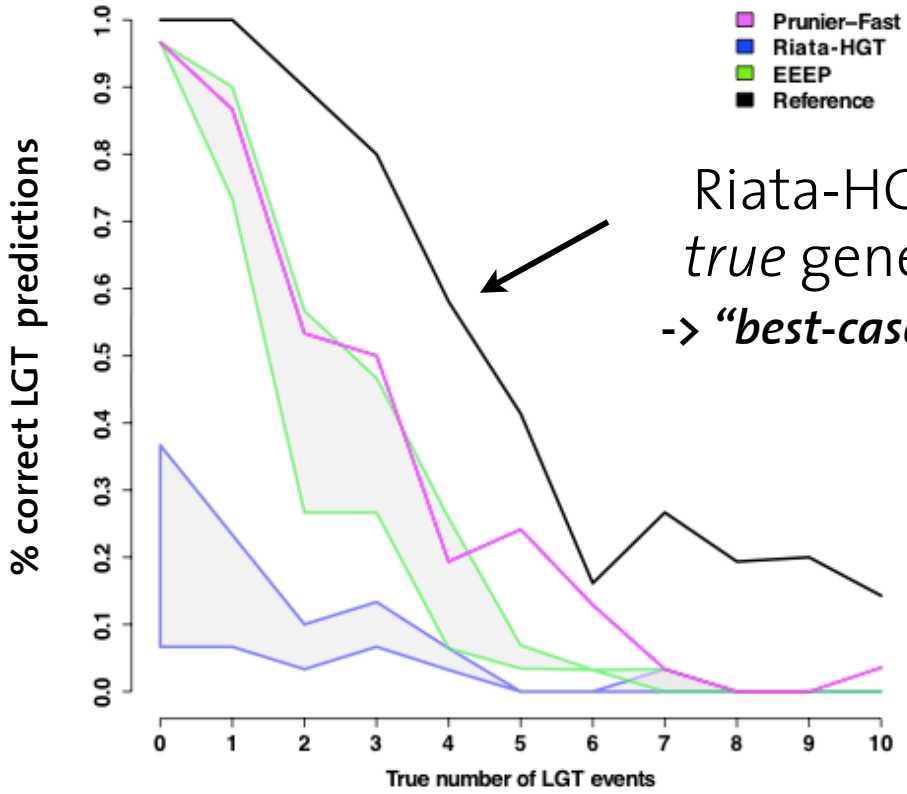
Output: a MSAF decomposition

A **reconciled backbone tree:** *T(ABFGHIK)*

Two **reconciled LGT subtrees:** *T(J)* and *T(CDE)*

Detecting lateral gene transfers by statistical reconciliation of phylogenetic forests

Sophie S Abby, Eric Tannier, Manolo Gouy and Vincent Daubin*



Riata-HGT on
true gene tree
-> ***"best-case" scenario***

Limitations of Explicit Phylogenetic Methods

- Factors other than LGT can lead to discordance.
- Cannot identify LGT between sister taxa.
- Relatively computationally expensive.
- *Discordance*: does not characterize the LGT: number of events, donor/recipient species, etc.
- *Model-based*: heuristic-based due to computational complexity

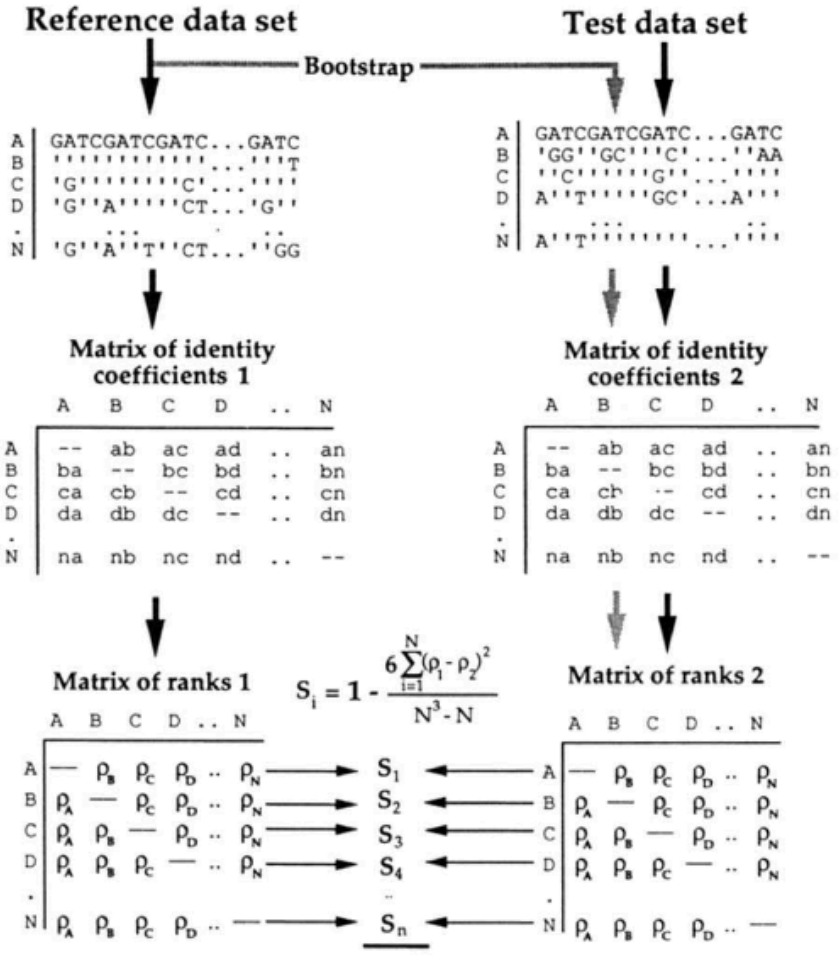
Implicit Phylogenetic Methods

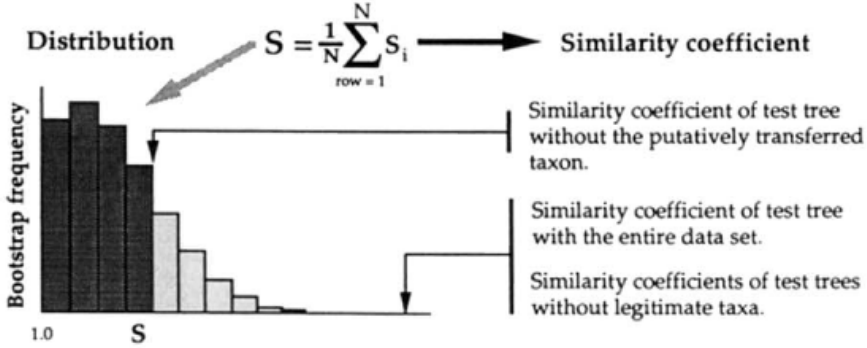
Lawrence & Hartl 1992
Clarke et al. 2002
Novichkov et al. 2004
Choi & Kim 2007
Podell & Gaasterland 2007
Dessimoz et al. 2008
Kanhere & Vingron 2009

Genetics 131: 753–760 (July, 1992)

Inference of Horizontal Genetic Transfer From Molecular Data: An Approach Using the Bootstrap

Jeffrey G. Lawrence[1] and Daniel L. Hartl

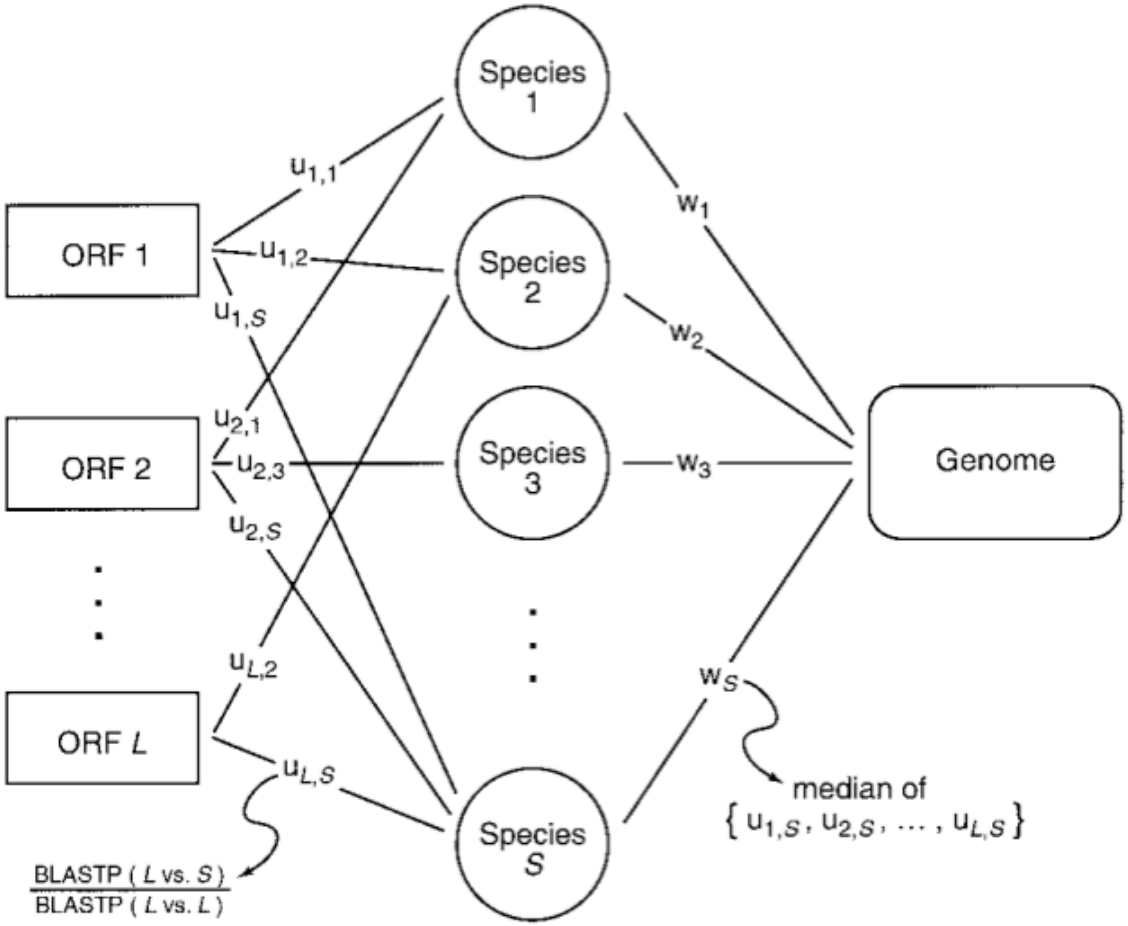




JOURNAL OF BACTERIOLOGY, Apr. 2002, p. 2072–2080 Vol. 184, No. 8
0021-9193/02/$04.00+0 DOI: 10.1128/JB.184.8.2072–2080.2002

Inferring Genome Trees by Using a Filter To Eliminate Phylogenetically Discordant Sequences and a Distance Matrix Based on Mean Normalized BLASTP Scores

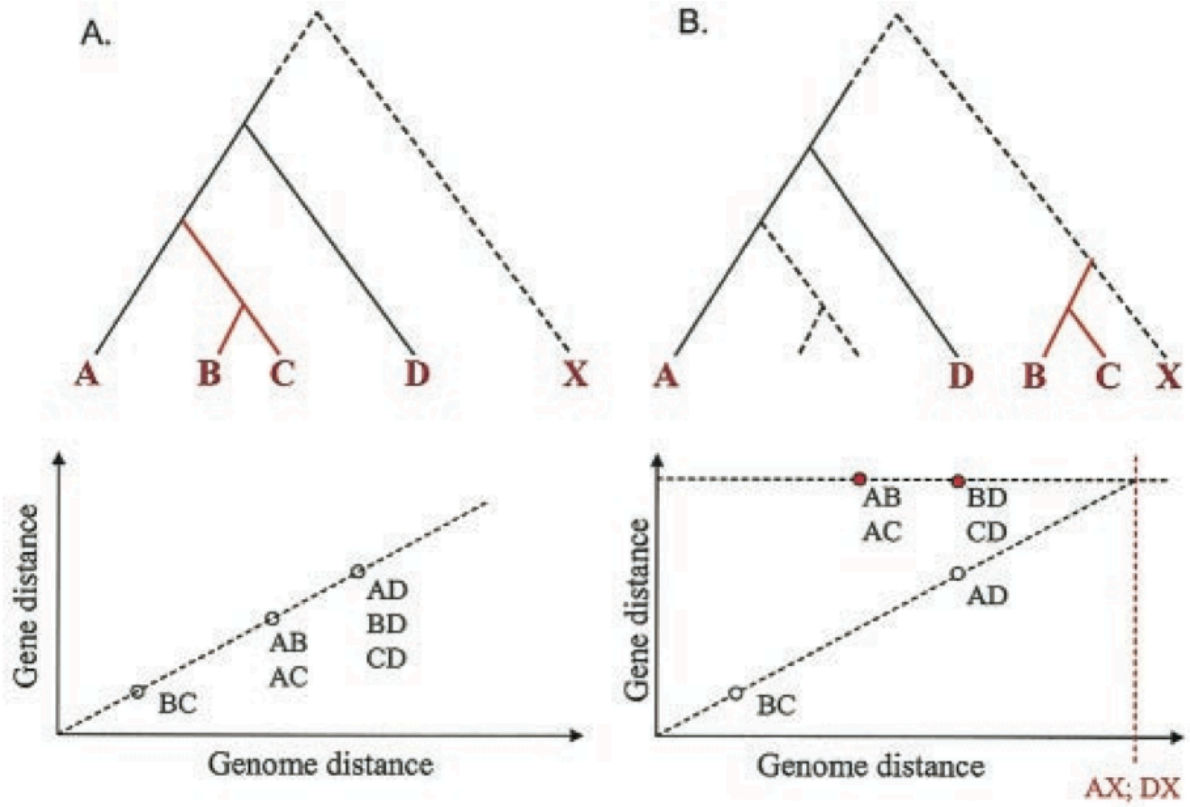
G. D. Paul Clarke,[1] Robert G. Beiko,[2] Mark A. Ragan,[3,4]* and Robert L. Charlebois[1,2,4]



JOURNAL OF BACTERIOLOGY, Oct. 2004, p. 6575–6585 Vol. 186, No. 19
0021-9193/04/$08.00+0 DOI: 10.1128/JB.186.19.6575–6585.2004

Genome-Wide Molecular Clock and Horizontal Gene Transfer in Bacterial Evolution

Pavel S. Novichkov,[1] Marina V. Omelchenko,[2,3] Mikhail S. Gelfand,[4,5] Andrei A. Mironov,[1] Yuri I. Wolf,[3] and Eugene V. Koonin[3]*

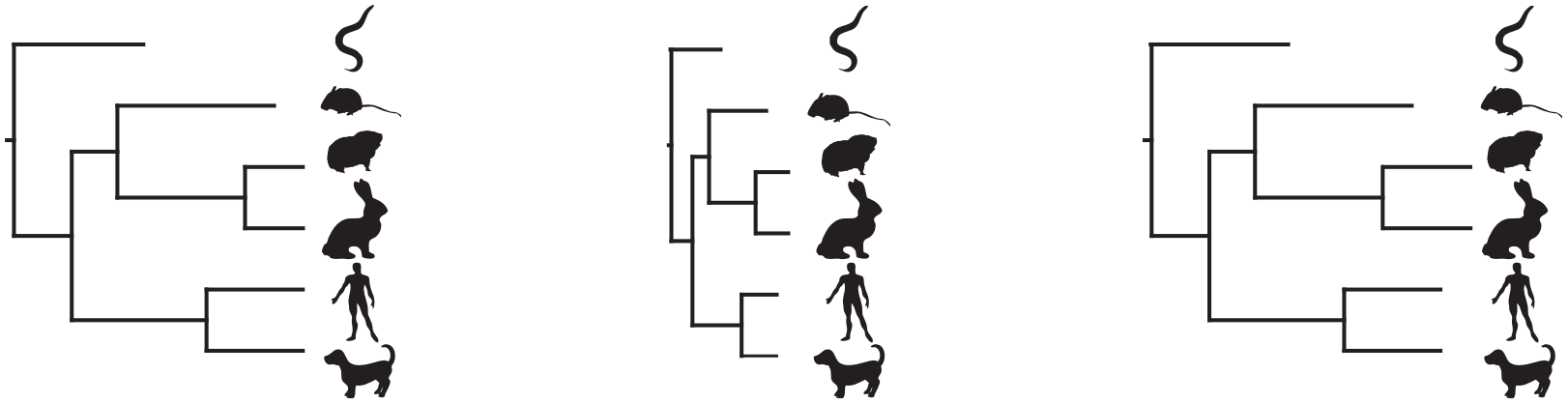


RECOMB 2008, LNBI 4955, pp. 315–330, 2008.

DLIGHT – Lateral Gene Transfer Detection Using Pairwise Evolutionary Distances in a Statistical Framework

Christophe Dessimoz*, Daniel Margadant, and Gaston H. Gonnet

- gene tree = $\text{rate}_f \times$ species tree



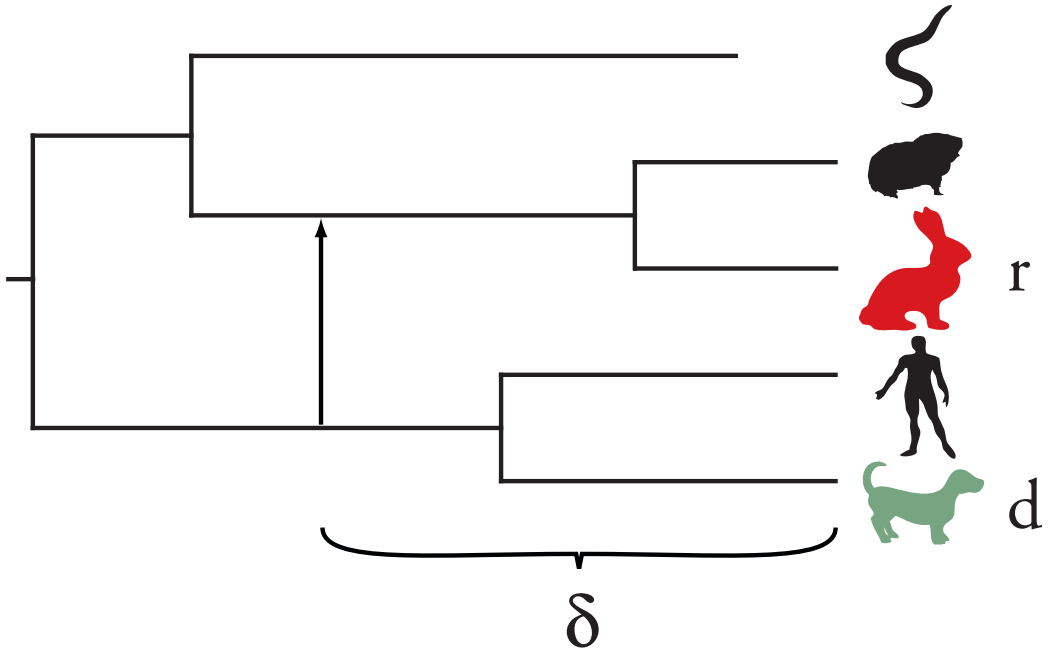
Algorithm

for all orthologous families f **do**

for all pairs d, r with a seq. in f **do**

if $2 \ln \frac{l(f,d,r,\delta_{ML})}{l(f,d,r,\delta=\infty)} > \chi^2(\alpha, 1)$ **then**

the triplet (f,d,r) is a LGT transfer



Likelihood computation: the idea

- Observed distances

$$\mathbf{x} = (x_1, x_2, \ldots, x_n)^{\mathrm{T}} \sim \mathcal{N}_n(\mathbf{T}, \Sigma_x)$$

Likelihood computation: the idea

- Observed distances

  $\mathbf{x} = (x_1, x_2, \ldots, x_n)^{\mathrm{T}} \sim \mathcal{N}_n(\mathbf{T}, \Sigma_x)$

- Modeled distances

  $\mathbf{y} = (y_1, y_2, \ldots, y_n)^{\mathrm{T}} \sim \mathcal{N}_n(\mathbf{T}, \Sigma_y)$

  where $y_i = g(\mathrm{seqs}_i, d, r, \delta)$

Likelihood computation: the idea

- Observed distances

  $\mathbf{x} = (x_1, x_2, \ldots, x_n)^{\mathrm{T}} \sim \mathcal{N}_n(\mathbf{T}, \Sigma_x)$

- Modeled distances

  $\mathbf{y} = (y_1, y_2, \ldots, y_n)^{\mathrm{T}} \sim \mathcal{N}_n(\mathbf{T}, \Sigma_y)$

  where $y_i = g(\mathrm{seqs}_i, d, r, \delta)$

- Thus, assuming that $\mathbf{x}, \mathbf{y}$ are independent,

$$l(f, d, r, \delta) = \mathrm{pdf}(\mathbf{x} - \mathbf{y}) = \frac{exp(-\frac{1}{2}(\mathbf{x}-\mathbf{y})^{\mathrm{T}}(\Sigma_x+\Sigma_y)^{-1}(\mathbf{x}-\mathbf{y}))}{\sqrt{(2\pi)^n|\Sigma_x+\Sigma_y|}}$$

Limitations of Implicit Phylogenetic Methods

- Cannot identify LGT between sister taxa
- *Most methods* do not characterize the LGTs identified: number of events, donor/recipient species, time of events, etc.

Outlook

- **Traditionally parametric vs phylogenetic, but latter is gaining momentum because profits more from more genomes**
- **Somewhat fragmented community:**
  - Microbiologists: parametric methods
  - Theoreticians: model-based phylo.
  - Pragmatic bioinform: discordance
- **I see potential for (1) comparing methods; (2) providing large-scale results into DB**