Universidade de São Paulo
Instituto de Química

Análise filogenética para dados moleculares – aula 2

João Carlos Setubal

2021

Programas para inferência filogenética

- Pacotes
  - Oferecem vários diferentes programas
  - Diferentes métodos para o mesmo objetivo
  - Podem incluir programas auxiliares
- Programas individuais
  - São especializados num método

Pacotes

- PHYLIP
  - Joe Felsenstein
  - http://evolution.genetics.washington.edu/phylip.html
- PAUP
  - David Swofford
  - http://paup.csit.fsu.edu/
- MEGA
  - Sudhir Kumar, Koichiro Tamura & Masatoshi Nei
  - http://www.megasoftware.net/
  - Atualmente na versão X

Programas que implementam métodos não-probabilísticos

- **Distância**
  - Pacotes mencionados no slide anterior
    - Neighbor-joining
    - UPGMA
- **Parcimônia**
  - pacotes mencionados no slide anterior

Máxima verossimilhança

- RaXML
  - A. Stamatakis
  - http://www.exelixis-lab.org/
- phyML
  - O. Gascuel et al. Systematic Biology, 59(3):307-21, 2010
  - http://www.atgc-montpellier.fr/phyml/
- fastTree
  - Morgan N. Price in Adam Arkin's group
  - http://www.microbesonline.org/fasttree/
  - "FastTree can handle alignments with up to a million of sequences in a reasonable amount of time and memory"
- IQ-Tree
  - http://www.iqtree.org/
  - A fast and effective stochastic algorithm to infer phylogenetic trees by maximum likelihood. *IQ-TREE compares favorably to RAxML and PhyML* in terms of likelihoods with similar computing time (Nguyen et al., 2015)

Um resultado de desempenho pontual

- Criação de uma árvore ML para 500 sequências de proteínas com aprox. 300 aa
- Computador desktop “normal” (4 GB de RAM)
- RAxML or PHYml levaram aprox. 10 horas
- Fasttree levou menos do que 1 hora

Inferência bayesiana

- MrBayes
- Ronquist and Huelsenbeck. Bioinformatics. 2003 19(12):1572-4.
- http://mrbayes.sourceforge.net/
- Mais lento comparado a RAxML e phyML
- Resultados não são conclusivamente melhores do que ML

O problema da caixa preta

- Idealmente: todo usuário de um método e respectivo programa deveria entender os princípios do método
- No caso de métodos de filogenia
  - Estatística não trivial

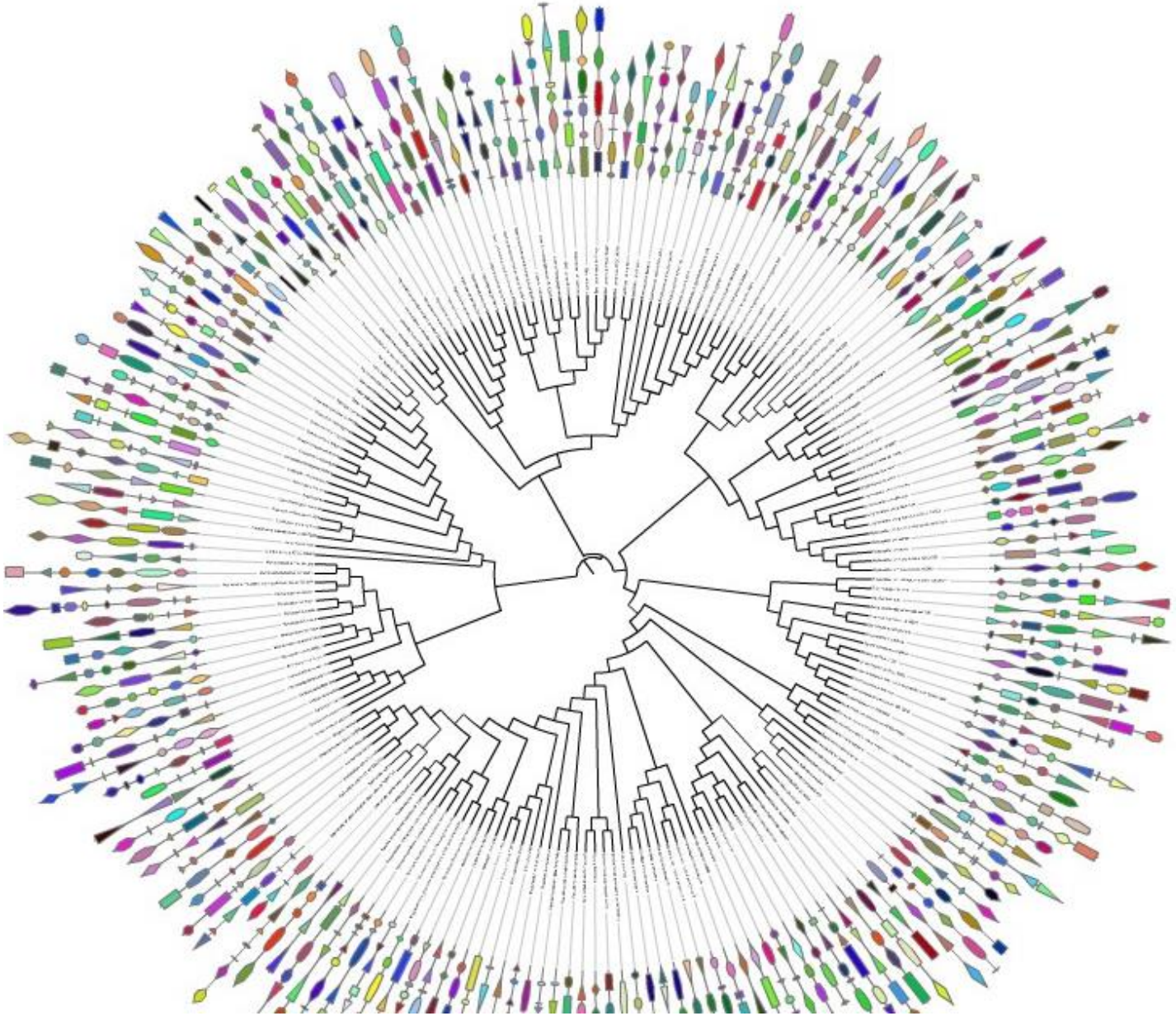
Visualização de árvores: formatos

- **Newick, NEXUS**
- (((erHomoC:0.28006,erCaelC:0.22089):0.40998,(erHomoA:0.32304,(erpCaelC:0.58815,((erHomoB:0.5807,erCaelB:0.23569):0.03586,erCaelA:0.38272):0.06516):0.03492):0.14265):0.63594,(TRXHomo:0.65866,TRXSacch:0.38791):0.32147,TRXEcoli:0.57336);
- **http://molecularevolution.org/resources/treeformats**

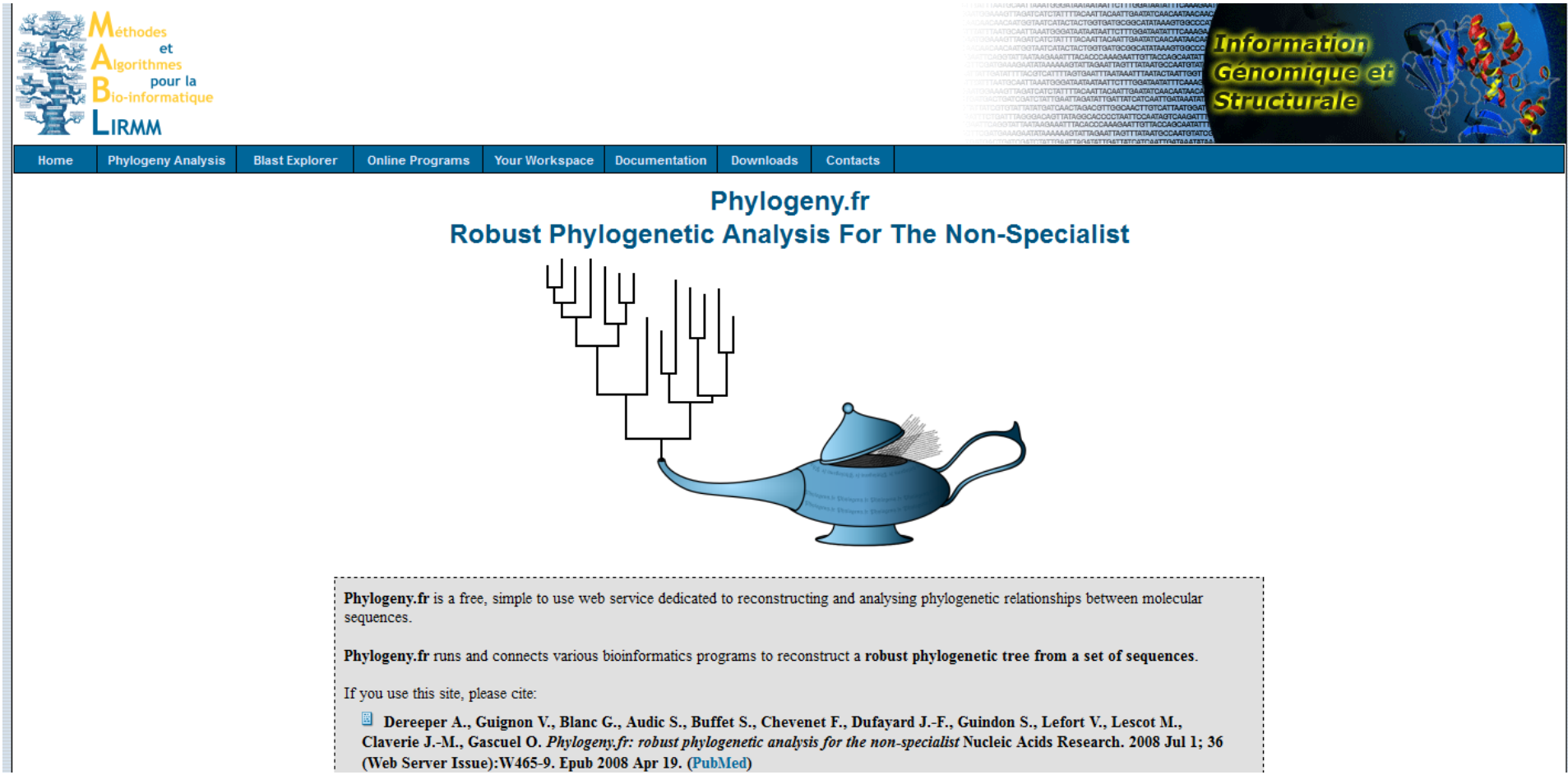
Visualização de árvores

- Interactive Tree of Life http://itol.embl.de

- **http://en.wikipedia.org/wiki/List_of_phylogenetic_tree_visualization_software**



All-in-one: http://phylogeny.fr



Phylogeny.fr (2)

Phylogeny analysis

"One Click"
Paste your set of sequences and let the software make decisions on your behalf (Each step is optimized for your data).

"Advanced"
Manually set parameters for the various steps.

"A la Carte"
Create your own phylogeny workflow using more programs available.

Explore your sequence neighbors

Paste your single sequence, run Blast and explore its homologous sequences.

Online phylogeny programs

Direct access to the individual tools available on this server.

Multiple Alignment:	Phylogeny:	Tree viewers:	Utilities:
MUSCLE	PhyML	TreeDyn	Gblocks
T-Coffee / 3D-Coffee	TNT	Drawgram	Jalview
ClustalW	BioNJ	Drawtree	Readseq
ProbCons	MrBayes	ATV	Format converter

This project is funded by the Réseau National des Génopoles (RNG).
This project is managed in a GForge project, which aims to help collaboration and development management (using Subversion).

RSS Feed Mailing-list Mentions légales

Marseille - Nice genopole®

cnrs Centre National de la Recherche Scientifique
dépasser les frontières

Building your tree locally: SeaView

SeaView

Version 4.3.5

NEW: seaview computes and draws parsimony, distance and PhyML phylogenetic trees.
NEW: seaview prints trees and outputs them in scalable vector graphics (SVG) format.
NEW: seaview drives the Gblocks program to select blocks of conserved sites.

SeaView is a multiplatform, graphical user interface for multiple sequence alignment and molecular phylogeny.

- SeaView reads and writes various file formats (NEXUS, MSF, CLUSTAL, FASTA, PHYLIP, MASE, Newick) of DNA and protein sequences and of phylogenetic trees.
- SeaView drives programs muscle or clustalw for multiple sequence alignment, and also allows to use any external alignment algorithm able to read and write FASTA-formatted files
- Seaview drives the Gblocks program to select blocks of evolutionarily conserved sites.
- SeaView computes phylogenetic trees by
  - parsimony, using PHYLIP's dnapars/protpars algorithm,
  - distance, with NJ or BioNJ algorithms on a variety of evolutionary distances,
  - maximum likelihood, driving program PhyML 3.0.
- SeaView prints and draws phylogenetic trees on screen, SVG, PDF or PostScript files.
- SeaView allows to download sequences from EMBL/GenBank/UniProt using the Internet.

Screen shots of the main alignment and tree windows. On-line help document. Old seaview version 3.2

Download SeaView

Mac Universal **MacOS X**	**Linux on PC** **64-bit Linux on x86**
MS Windows **self-extractible archive**	**Solaris on SPARC**

Interpretação

- Árvores são apenas hipóteses
- GIGO: garbage in, garbage out
- Os métodos em geral (menos distância) fornecem uma árvore com nota (score)
  - Parcimônia: número mínimo de mutações
  - ML: valor da verossimilhança logarítmica
  - Bayesiano: probabilidade posterior
- A árvore de melhor nota pode não ser a árvore “verdadeira”
- Para avaliar a qualidade da árvore
  - Confiabilidade de sua topologia

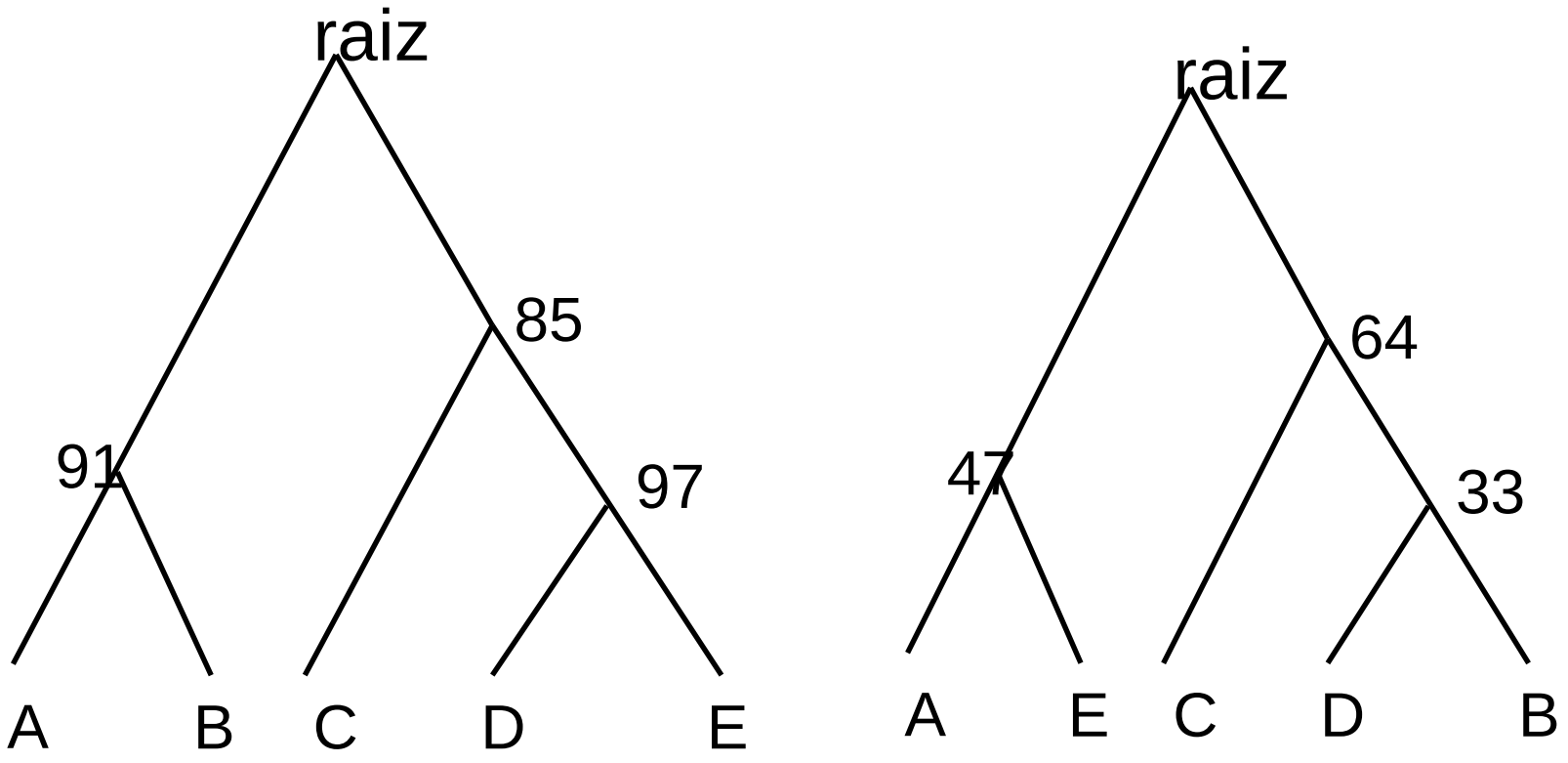
Topologia das árvores

- Uma árvore é uma forma de relacionar as folhas (nós externos) e os nós internos
- Usamos o termo *topologia* para nos referirmos à estrutura da árvore (quais são os grupos, subgrupos, supergrupos de nós), ignorando os comprimentos dos ramos
- É comum que o termo *clado* seja usado para designar um subgrupo na árvore
- Obtida uma árvore, como saber se sua topologia é confiável?

Confiabilidade da topologia

- Valores de bootstrap
- Colunas do AM são amostradas aleatoriamente em várias corridas (replicatas; geralmente entre 100 e 1000)
- Árvores resultantes são comparadas entre si
- Concordâncias nos clados são calculadas, resultando em número de vezes (ou %) que clados se repetem nas replicatas
- Valores bons são considerados aqueles maiores do que 0.7 (70%)
- Custosos para calcular
- PhyML fornece valores aproximados de bootstrap (ALRT) muito mais rapidamente

Exemplo de árvores com bootstrap



Como lidar com todas essas incertezas?

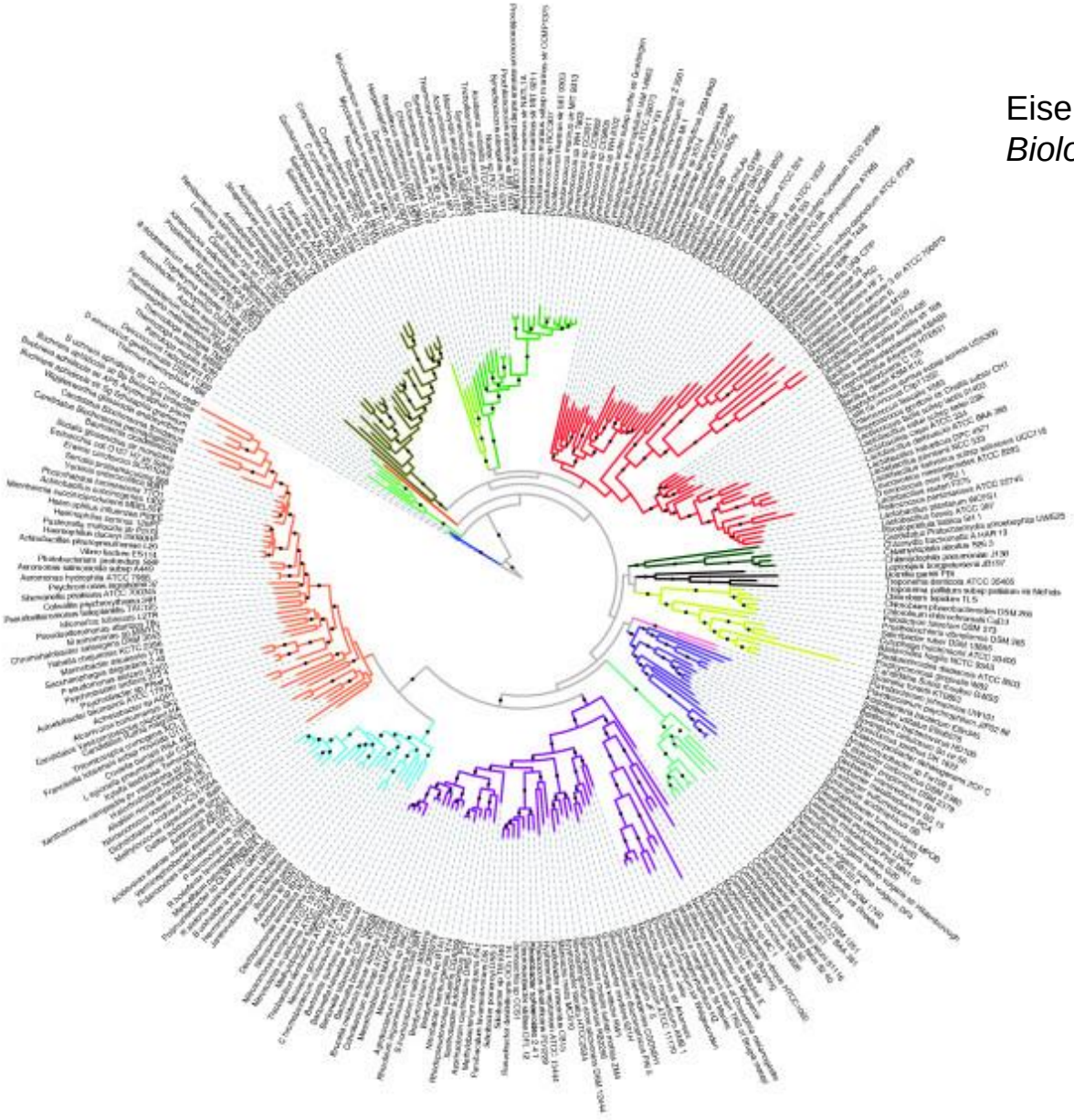
- Aprenda mais sobre evolução e inferência filogenética
- Se a filogenia é crucial para seus resultados
  - Use mais de um método!

Supermatrizes

- Método bom para obter árvores robustas de espécies quando genomas completos ou quase completos estão disponíveis
- Determinar famílias de proteínas para os genomas de interesse
- Determinar quais famílias tem exatamente um representante de cada genoma
- AM para cada família
- Concatenar todos os AMs (“a supermatriz”)
- Construir árvore com base no AM concatenado
- Foi usando esse método que foi construída a árvore da vida do próximo slide

Ciccarelli et al, *Science*, 2006

No próximo slide há um outro exemplo de uso da técnica de supermatrizes, neste caso para construir uma árvore apenas de bactérias



Eisen & Wu, *Genome Biology*, 2008

Problemas de supermatrizes

- Diferentes genes tem diferentes taxas de evolução
- Long branch attraction
  - Ramos longos (muitas mutações) tendem a ficar artificialmente próximos uns dos outros (e próximos da raiz)
  - Topologia errada
- HGT insuspeita sempre vai introduzir erros

Próximo slide: exemplo recente de árvore de bactérias e arquéias

Exercício: compare essa árvore com a árvore de Wu e Eisen do slide anterior

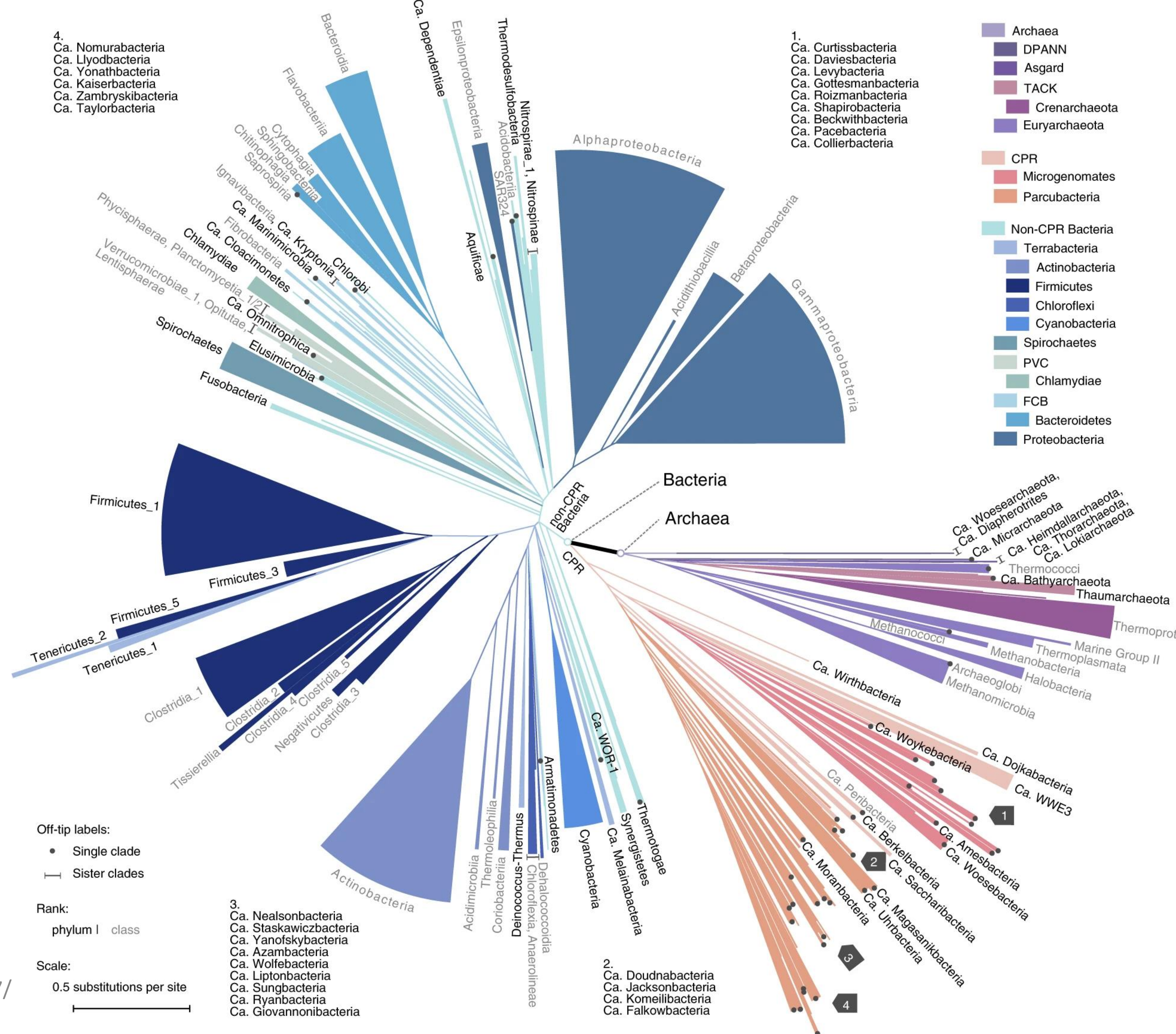
nature communications

Article | Open Access | Published: 02 December 2019

Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea

Qiyun Zhu, Uyen Mai, [...] Rob Knight ✉

Nature Communications **10**, Article number: 5477 (2019) | Cite this article

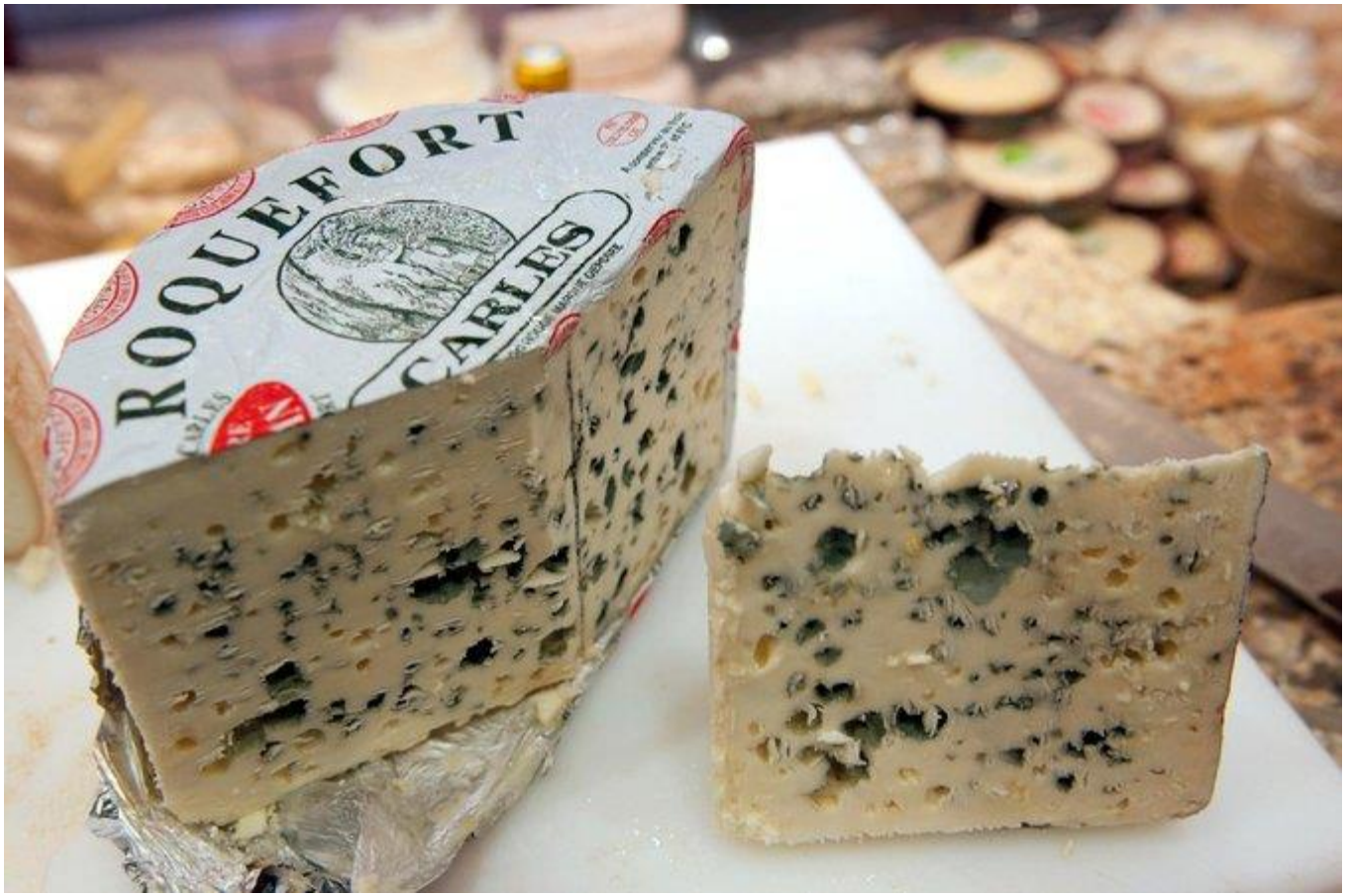


Bacteria
Archaea
non-CPR Bacteria
CPR
1.
Ca. Curtissbacteria
Ca. Daviesbacteria
Ca. Levybacteria
Ca. Gottesmanbacteria
Ca. Roizmanbacteria
Ca. Shapirobacteria
Ca. Beckwithbacteria
Ca. Pacebacteria
Ca. Collierbacteria
2.
Ca. Doudnabacteria
Ca. Jacksonbacteria
Ca. Komeilibacteria
Ca. Falkowbacteria
3.
Ca. Nealsonbacteria
Ca. Staskawiczbacteria
Ca. Yanofskybacteria
Ca. Azambacteria
Ca. Wolfebacteria
Ca. Liptonbacteria
Ca. Sungbacteria
Ca. Ryanbacteria
Ca. Giovannonibacteria
4.
Ca. Nomurabacteria
Ca. Llyodbacteria
Ca. Yonathbacteria
Ca. Kaiserbacteria
Ca. Zambryskibacteria
Ca. Taylorbacteria
Archaea
DPANN
Asgard
TACK
Crenarchaeota
Euryarchaeota
CPR
Microgenomates
Parcubacteria
Non-CPR Bacteria
Terrabacteria
Actinobacteria
Firmicutes
Chloroflexi
Cyanobacteria
Spirochaetes
PVC
Chlamydiae
FCB
Bacteroidetes
Proteobacteria
Off-tip labels:
Single clade
Sister clades
Rank:
phylum | class
Scale:
0.5 substitutions per site
Alphaproteobacteria
Acidithiobacillia
Betaproteobacteria
Gammaproteobacteria
Ca. Dependentiae
Epsilonproteobacteria
Thermodesulfobacteria
Acidobacteria
SAR324
Nitrospirae_1, Nitrospinae
Aquificae
Bacteroidia
Flavobacteriia
Cytophagia
Sphingobacteriia
Chitinophagia
Saprospiria
Ignavibacteria, Ca. Kryptonia
Ca. Marinimicrobia
Chlorobi
Fibrobacteria
Ca. Cloacimonetes
Phycisphaerae, Planctomycetia_1/2
Chlamydiae
Verrucomicrobiae_1, Opitutae
Lentisphaerae
Ca. Omnitrophica
Spirochaetes
Elusimicrobia
Fusobacteria
Firmicutes_1
Firmicutes_3
Firmicutes_5
Tenericutes_2
Tenericutes_1
Clostridia_1
Clostridia_2
Clostridia_4
Clostridia_5
Negativicutes
Clostridia_3
Tissierellia
Actinobacteria
Acidimicrobiia
Thermoleophilia
Coriobacteriia
Deinococcus-Thermus
Chloroflexia, Anaerolineae
Dehalococcoidia
Armatimonadetes
Cyanobacteria
Ca. WOR-1
Ca. Melainabacteria
Synergistetes
Thermotogae
Ca. Woesearchaeota,
Ca. Diapherotrites
Ca. Micrarchaeota
Ca. Heimdallarchaeota,
Ca. Thorarchaeota,
Ca. Lokiarchaeota
Thermococci
Ca. Bathyarchaeota
Thaumarchaeota
Thermoprotei
Marine Group II
Thermoplasmata
Methanococci
Methanobacteria
Halobacteria
Archaeoglobi
Methanomicrobia
Ca. Wirthbacteria
Ca. Woykebacteria
Ca. Dojkabacteria
Ca. WWE3
Ca. Peribacteria
Ca. Berkelbacteria
Ca. Saccharibacteria
Ca. Amesbacteria
Ca. Woesebacteria
Ca. Moranbacteria
Ca. Uhrbacteria
Ca. Magasanikbacteria
1
2
3
4

Transferência Horizontal de Genes

- Material genético é passado de uma célula (doadora) para outra (receptora)
- O doador pode ser completamente diferente do receptor
- Exemplo: humanos e bactérias

Exemplo de HGT



Fonte: NY Times, 24-set-2015

Fungos e queijos

- Fabricação de queijos depende da ação de fungos
- Roquefort
  - *Penicillium roqueforti*
- Camembert
  - *P. camemberti*
- Esses fungos vem sendo selecionados e cultivados há séculos

Resultado recente

- Ao comparar diferentes espécies de fungos usados em queijos, descobriu-se
  - *Multiple Recent Horizontal Gene Transfers between Distant Penicillium Species, Flanked by Specific Retrotransposons*

Please cite this article in press as: Ropars et al., Adaptive Horizontal Gene Transfers between Multiple Cheese-Associated Fungi, Current Biology (2015), http://dx.doi.org/10.1016/j.cub.2015.08.025

Current Biology
Report

CellPress

Adaptive Horizontal Gene Transfers between Multiple Cheese-Associated Fungi

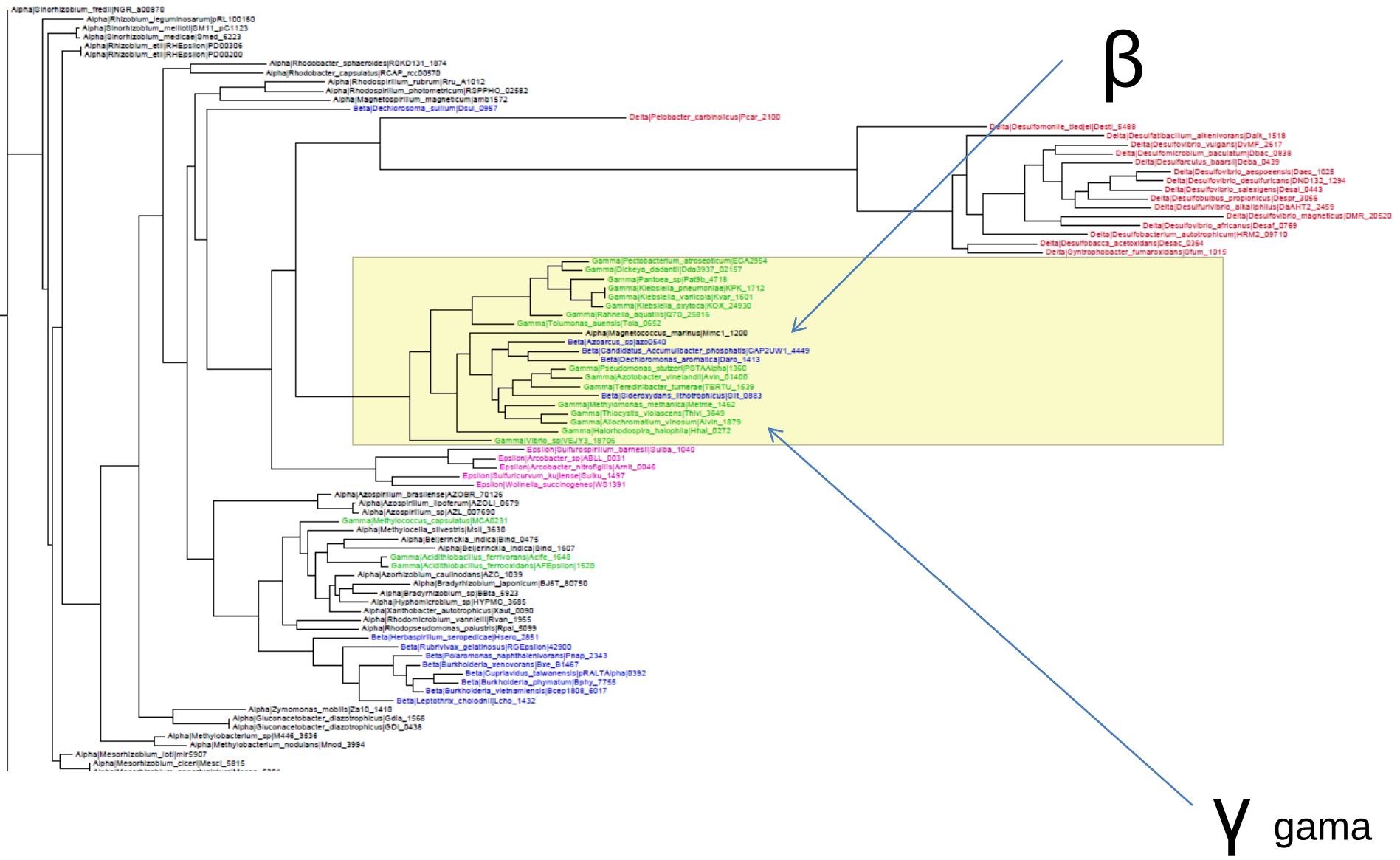
Jeanne Ropars,[1,2,8] Ricardo C. Rodríguez de la Vega,[1,2,8] Manuela López-Villavicencio,[3] Jérôme Gouzy,[4,5] Erika Sallet,[4,5] Émilie Dumas,[1,2] Sandrine Lacoste,[3] Robert Debuchy,[6,7] Joëlle Dupont,[3] Antoine Branca,[1,2,9,*] and Tatiana Giraud[1,2,9,*]
[1]Ecologie, Systématique et Evolution, UMR8079, Univ. Paris-Sud, 91405 Orsay, France
[2]Ecologie, Systématique et Evolution, UMR8079, CNRS, 91405 Orsay, France
[3]Institut de Systématique, Evolution, Biodiversité, UMR 7205 CNRS-MNHN-UPMC-EPHE, Muséum national d'Histoire naturelle, Sorbonne Université, CP39, 57 Rue Cuvier, 75231 Paris Cedex 05, France
[4]Laboratoire des Interactions Plantes-Microorganismes (LIPM), UMR441, INRA, Castanet-Tolosan 31326, France
[5]Laboratoire des Interactions Plantes-Microorganismes (LIPM), UMR2594, CNRS, Castanet-Tolosan 31326, France
[6]Institut de Génétique et Microbiologie, UMR8621, Univ. Paris-Sud, 91405 Orsay, France
[7]Institut de Génétique et Microbiologie, UMR8621, CNRS, 91405 Orsay, France
[8]Co-first author
[9]Co-senior author
*Correspondence: antoine.branca@u-psud.fr (A.B.), tatiana.giraud@u-psud.fr (T.G.)
http://dx.doi.org/10.1016/j.cub.2015.08.025

Transferência Horizontal de Genes

- Atrapalha a construção de árvores de espécies
- Como detectar?
- THG antiga
- THG recente

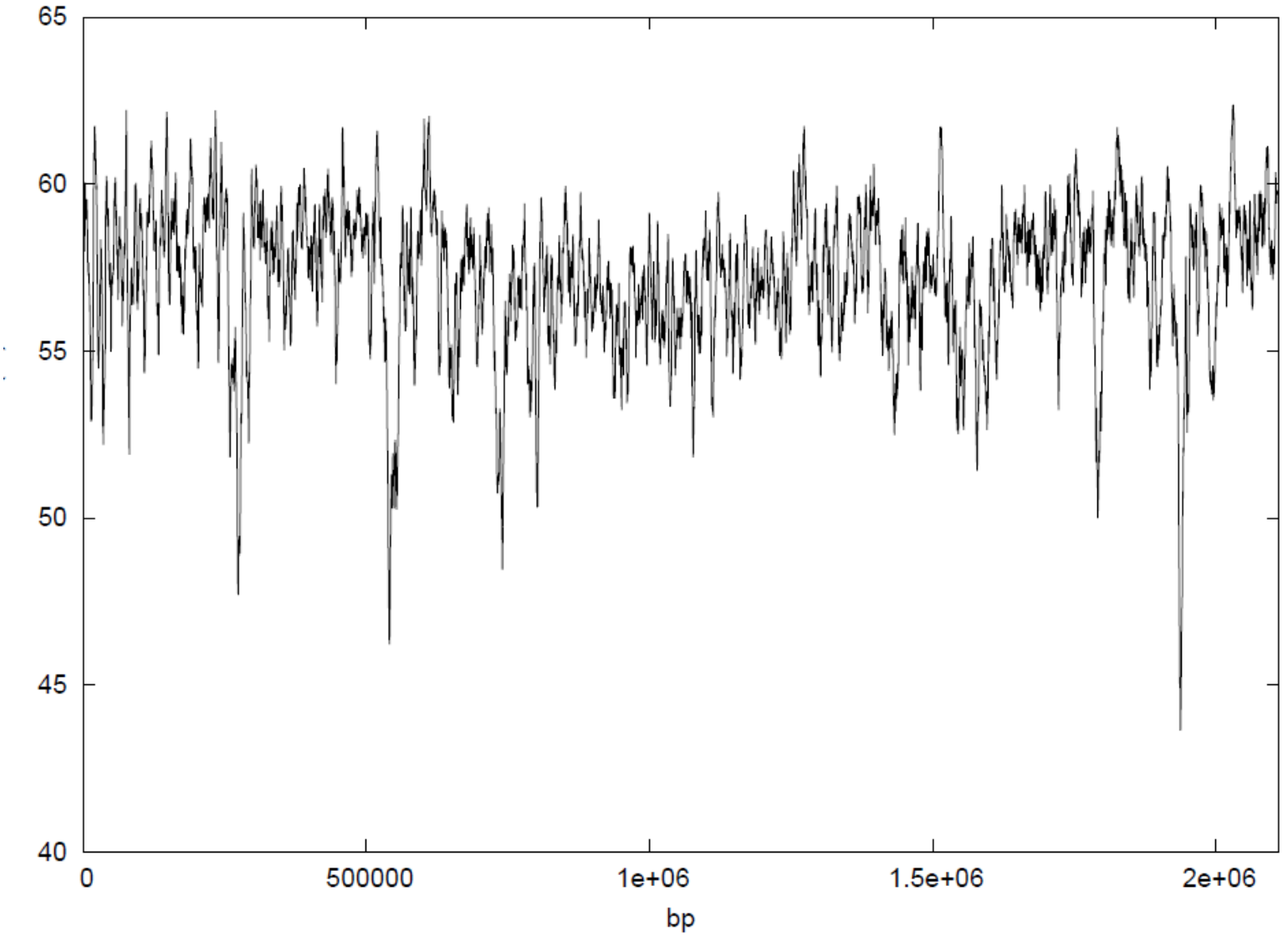
THG antiga

- Incongruência de árvores
  - Quando a árvore de um gene difere da árvore (robusta) de espécies

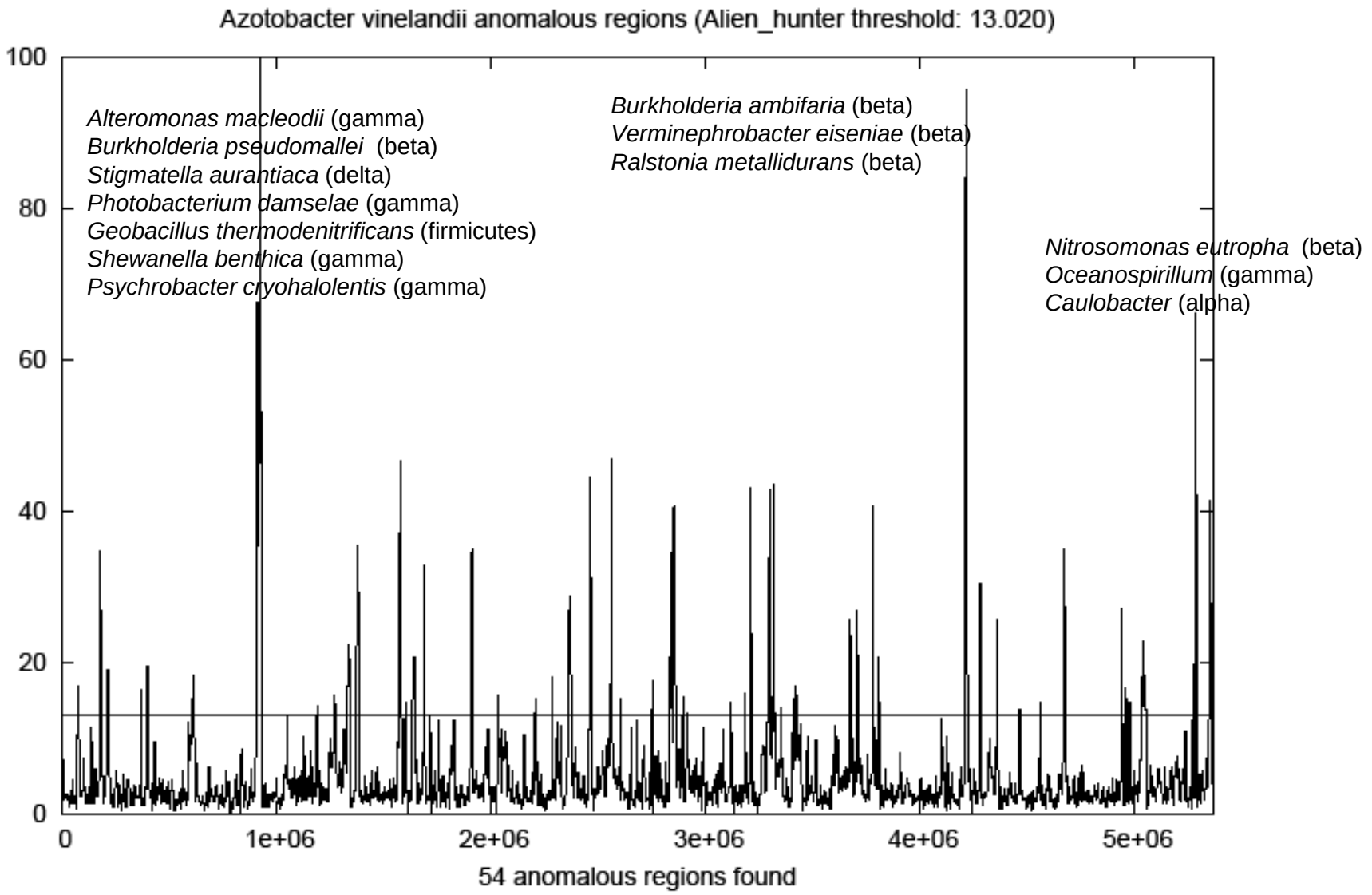

β
γ gama

THG recente

- Incongruência de árvores
- Outros métodos
  - Desvios na composição (%GC, dinucleotídeos, uso de codons) da sequência
  - Ilhas genômicas

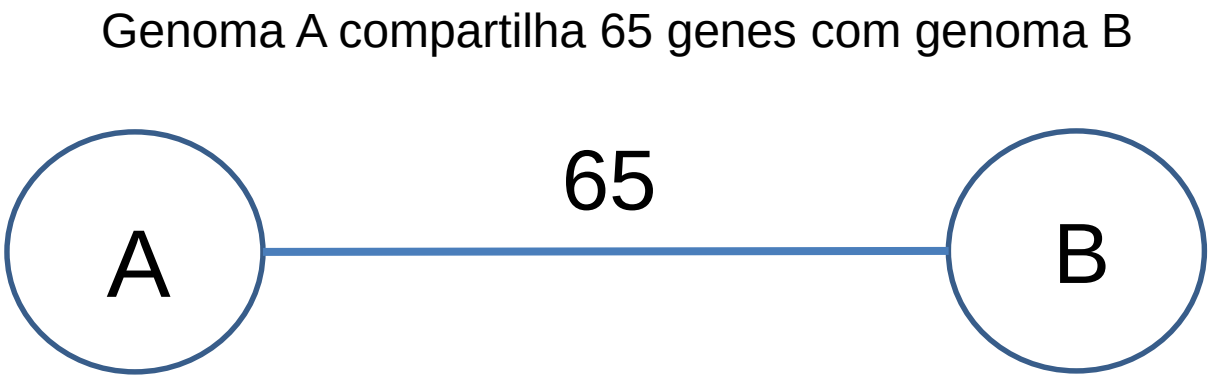


Variação do %GC no cromossomo principal de *Brucella ovis* ATCC25840


Azotobacter vinelandii anomalous regions (Alien_hunter threshold: 13.020)
100
80
60
40
20
0
0
1e+06
2e+06
3e+06
4e+06
5e+06
54 anomalous regions found
Alteromonas macleodii (gamma)
Burkholderia pseudomallei (beta)
Stigmatella aurantiaca (delta)
Photobacterium damselae (gamma)
Geobacillus thermodenitrificans (firmicutes)
Shewanella benthica (gamma)
Psychrobacter cryohalolentis (gamma)
Burkholderia ambifaria (beta)
Verminephrobacter eiseniae (beta)
Ralstonia metallidurans (beta)
Nitrosomonas eutropha (beta)
Oceanospirillum (gamma)
Caulobacter (alpha)

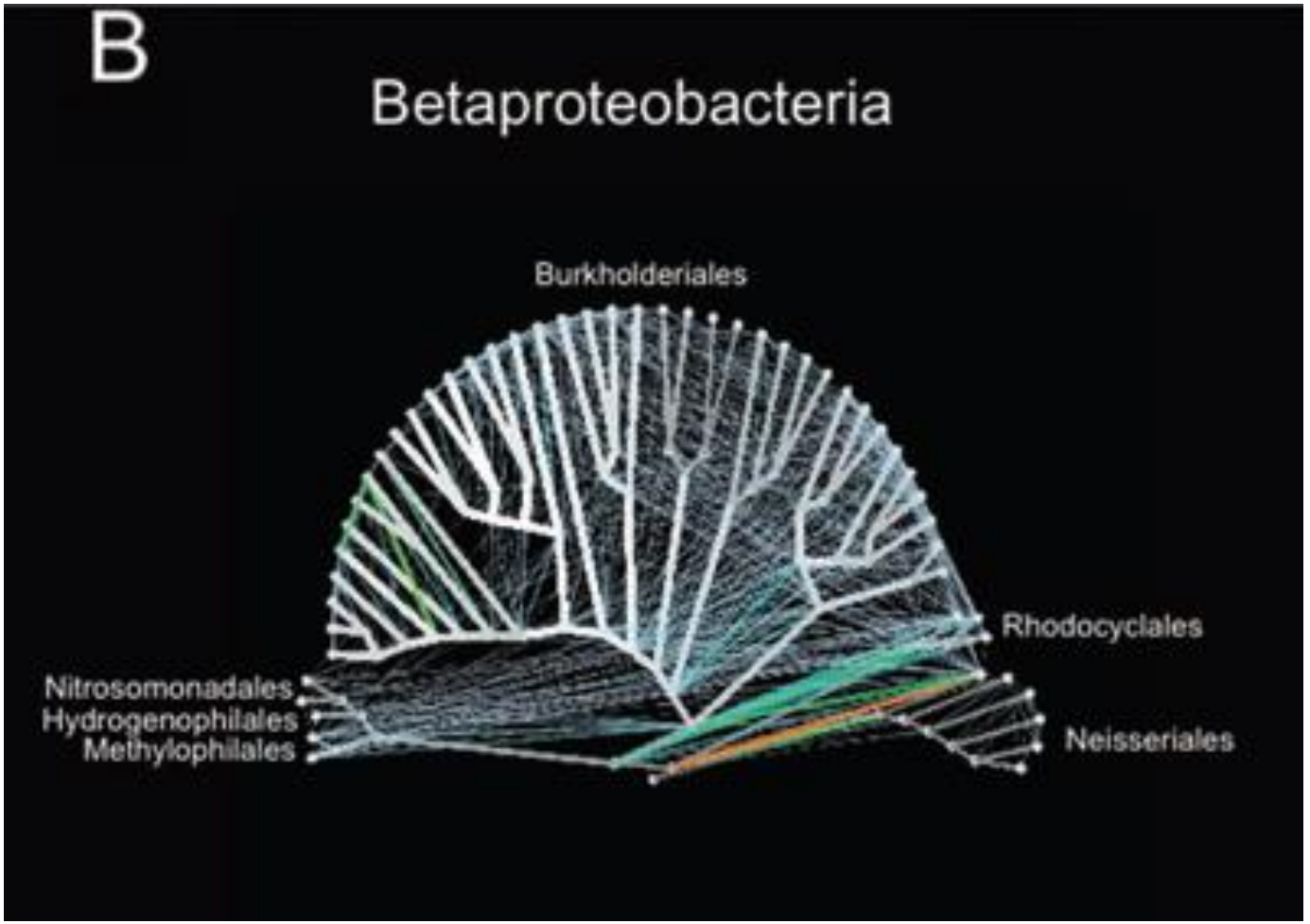
Redes filogenômicas

- Redes que mostram compartilhamento de genes

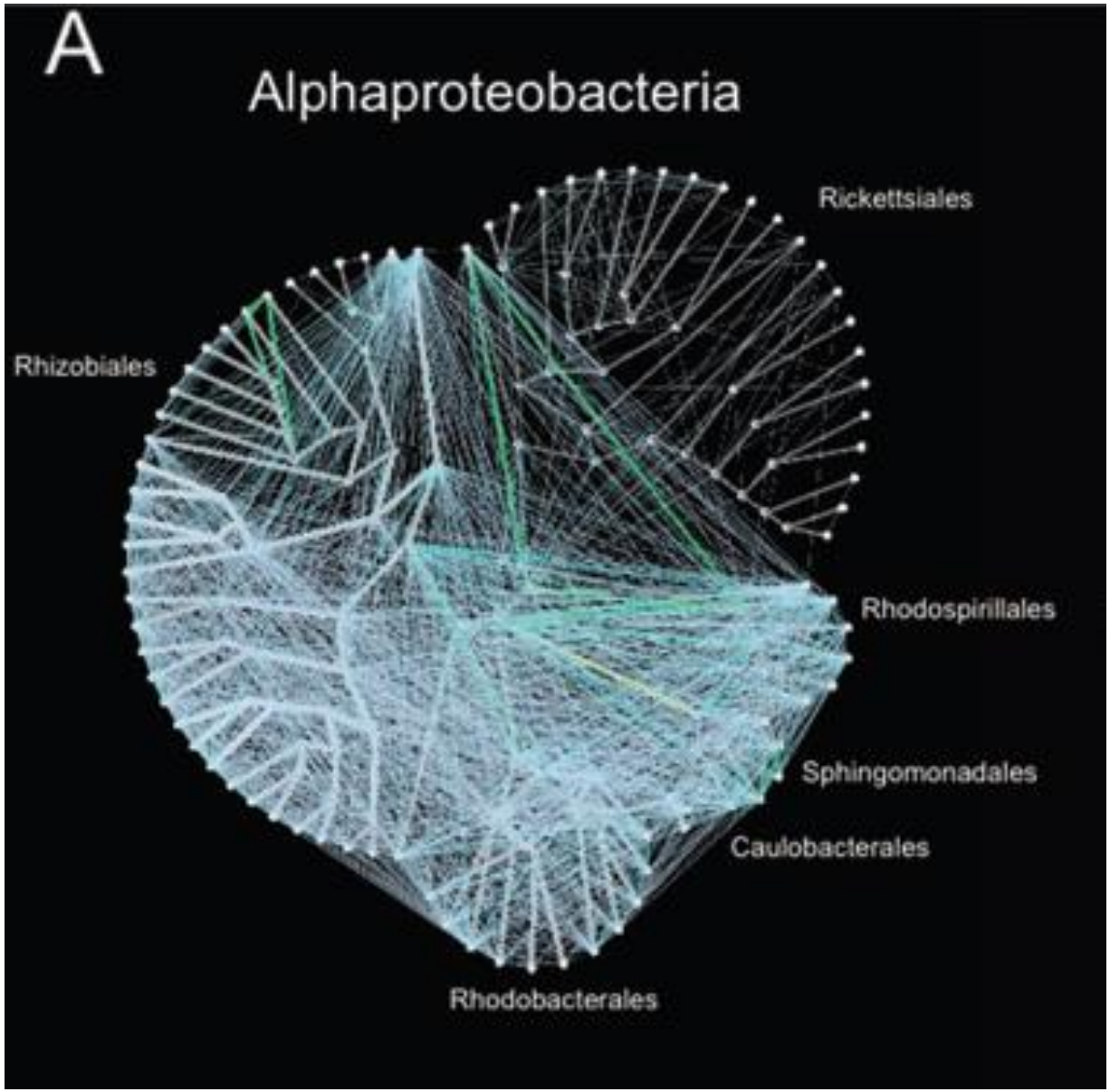


A superposição de uma árvore de espécies numa tal rede mostra possíveis eventos de transferência horizontal

Uma rede filogenômica

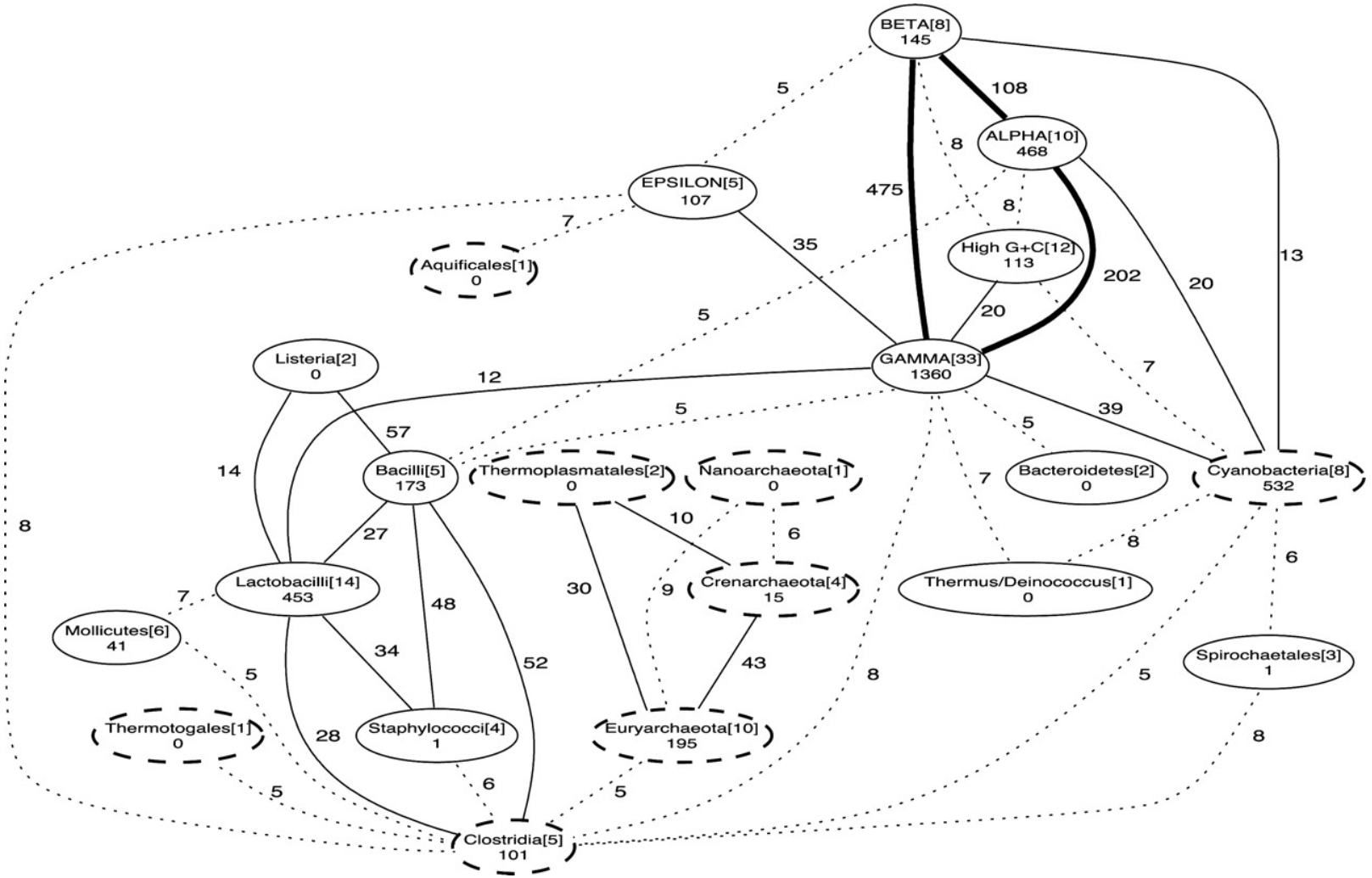


Kloesges et al, *Molecular Biology and Evolution*, 2011



Kloesges et al, *Molecular Biology and Evolution*, 2011

Highways of obligate gene transfer within and among phyla and divisions of prokaryotes, based on analysis of the 22,348 protein trees for which a minimal edit path could be resolved.



Beiko R G et al. PNAS 2005;102:14332-14337

Substituições sinônimas e não-sinônimas

- Código genético é degenerado
- Glicina: GGA, GGC, GGG, GGU
- Mutação na terceira base não altera o aminoácido
  - Sinônima (silenciosa)
- Mutação na primeira base altera o aminoácido
  - Não-sinônima

Razão Ka/Ks

- Ka/Ks ou dN/dS
- Razão entre o número de subs. não-sinônimas (Ka) e o número de subs. sinônimas (Ks)
- Usado para inferir a direção e magnitude de seleção natural agindo em genes codificadores de proteínas
- Ka/Ks > 1: seleção positiva ou Darwiniana
- Ka/Ks < 1: seleção purificadora ou estabilizadora
- Ka/Ks = 1: não há seleção (neutra)

Para calcular Ka/Ks

- Hurst, L. (2002). "The Ka/Ks ratio: diagnosing the form of sequence evolution". *Trends in Genetics* **18**: 486–489
- http://services.cbu.uib.no/tools/kaks

Para saber mais

- Yang e Rannala. Molecular phylogenetics: principles and practice. *Nature Reviews Genetics,* 13:303-314, 2012
- ***Bioinformatics***. Baxevanis and Ouellette (Eds.) Wiley-Interscience, 2005 (3rd edition), **ch. 14**
- D. Mount. ***Bioinformatics***. CSHL Press, 2004 (2nd edition), **ch. 7**
- ***The phylogenetic handbook***. Lemey, Salemi and Vandamme (Eds.) Cambridge University Press, 2009 (2nd edition)

NEWS FEATURE

THE TOP 100 PAPERS

Nature **explores the most-cited research of all time.**

BY RICHARD VAN NOORDEN, BRENDAN MAHER AND REGINA NUZZO

The discovery of high-temperature superconductors, the determination of DNA's double-helix structure, the first observations that the expansion of the Universe is accelerating — all of these breakthroughs won Nobel prizes and international acclaim. Yet none of the papers that announced them comes anywhere close to ranking among the 100 most highly cited papers of all time.

Citations, in which one paper refers to ear-

of carbon nanotubes (number 36) are indeed classic discoveries. But the vast majority describe experimental methods or software that have become essential in their fields.

The most cited work in history, for example, is a 1951 paper[2] describing an assay to determine the amount of protein in a solution. It has now gathered more than 305,000 citations — a recognition that always puzzled its lead author, the late US biochemist Oliver Lowry. "Although I really know it is not a

to other scientists what kind of wo
doing". Another common practice i
ensures that truly foundational d
ies — Einstein's special theory of r
for instance — get fewer citations t
might deserve: they are so important
quickly enter the textbooks or are inc
into the main text of papers as terms
so familiar that they do not need a ci

Citation counts are riddled with o
founding factors. The volume of cita
increased, for example — yet older pa
had more time to accrue citations. E
tend to cite one another's work more f
than, say, physicists. And not all fields
the same number of publications. Mo
liometricians therefore recoil from m
crude as simply counting citations w
want to measure a paper's value: inst
prefer to compare counts for papers
age, and in comparable fields.

Nor is Thomson Reuters' list the o
ing system available. Google Scholar
its own top-100 list for *Nature*. It is
many more citations because the sear
culls references from a much greater (
poorly characterized) literature base, i
from a large range of books. In that l
able at www.nature.com/top100, ec
papers have more prominence. Goog
ar's list also features books, which T
Reuters did not analyse. But among th
papers, many of the same titles show

Yet even with all the caveats, the
ioned hall of fame still has value. If
else, it serves as a reminder of the natu
entific knowledge. To make exciting
researchers rely on relatively unsung
describe experimental methods, datal
software.

Here *Nature* tours some of the key
that tens of thousands of citations hav
to the top of science's Kilimanjaro —
but rarely thrust into the limelight.

Nature, 30/10/2014

Papers de bioinformática: número de citações (2014)

- 10) clustalW: 40289
- 12) blast1: 38380
- 14) blast2: 36410
- 20) NJ: 30176
- 28) clustalX: 24098
- 41) bootstrap: 21373
- 45) MEGA: 18286
- 76) modelTest: 14099
- 100) MrBayes: 12209