UNIVERSITÀ di VERONA | Dipartimento di BIOTECNOLOGIE

Bioinformatic tools for bacterial identification and characterization

Giovanna Felis

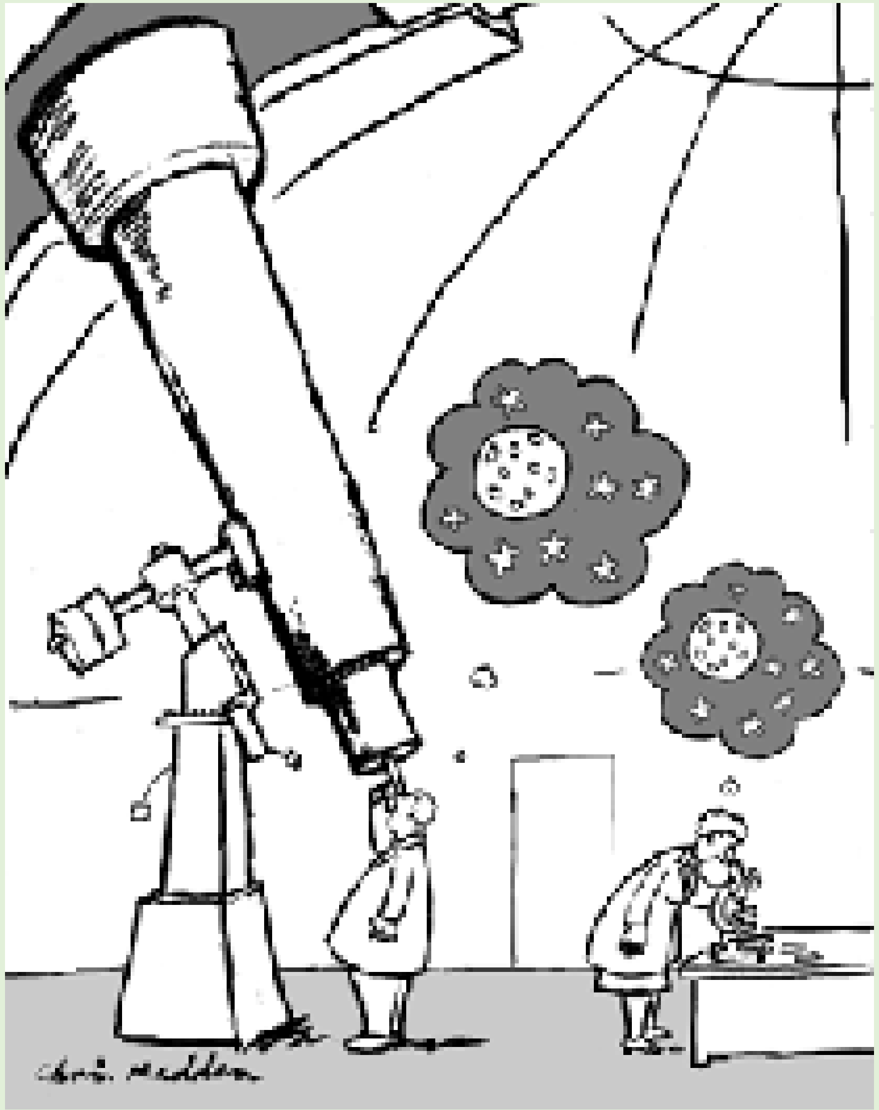
University of Verona – Dept. Biotechnology

giovanna.felis@univr.it

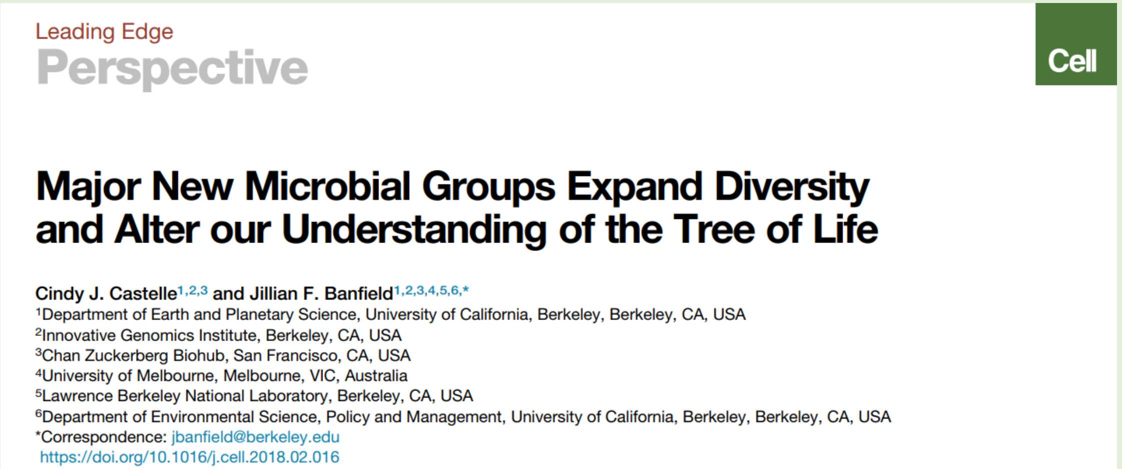
@FelisGiovanna

"Where the telescope ends the microscope begins,
and who can say which has the wider vision?"

- Victor Hugo (?) -



http://www.microbial-systems-ecology.de/links_taxonomy.html

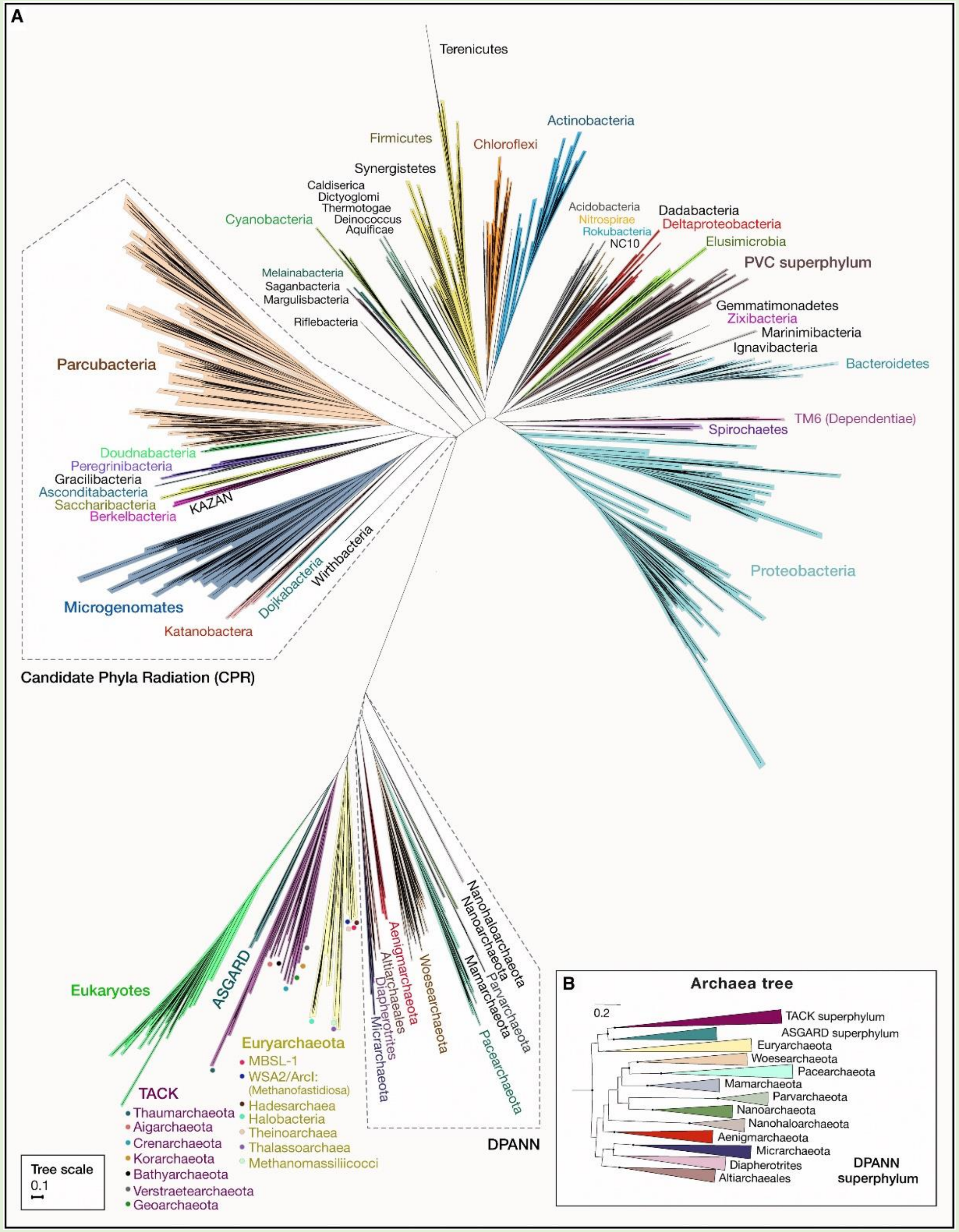


Leading Edge
Perspective
Cell

Major New Microbial Groups Expand Diversity and Alter our Understanding of the Tree of Life

Cindy J. Castelle[1,2,3] and Jillian F. Banfield[1,2,3,4,5,6,*]
[1]Department of Earth and Planetary Science, University of California, Berkeley, Berkeley, CA, USA
[2]Innovative Genomics Institute, Berkeley, CA, USA
[3]Chan Zuckerberg Biohub, San Francisco, CA, USA
[4]University of Melbourne, Melbourne, VIC, Australia
[5]Lawrence Berkeley National Laboratory, Berkeley, CA, USA
[6]Department of Environmental Science, Policy and Management, University of California, Berkeley, Berkeley, CA, USA
*Correspondence: jbanfield@berkeley.edu

https://doi.org/10.1016/j.cell.2018.02.016



UNIVERSITÀ di VERONA
Dipartimento di BIOTECNOLOGIE

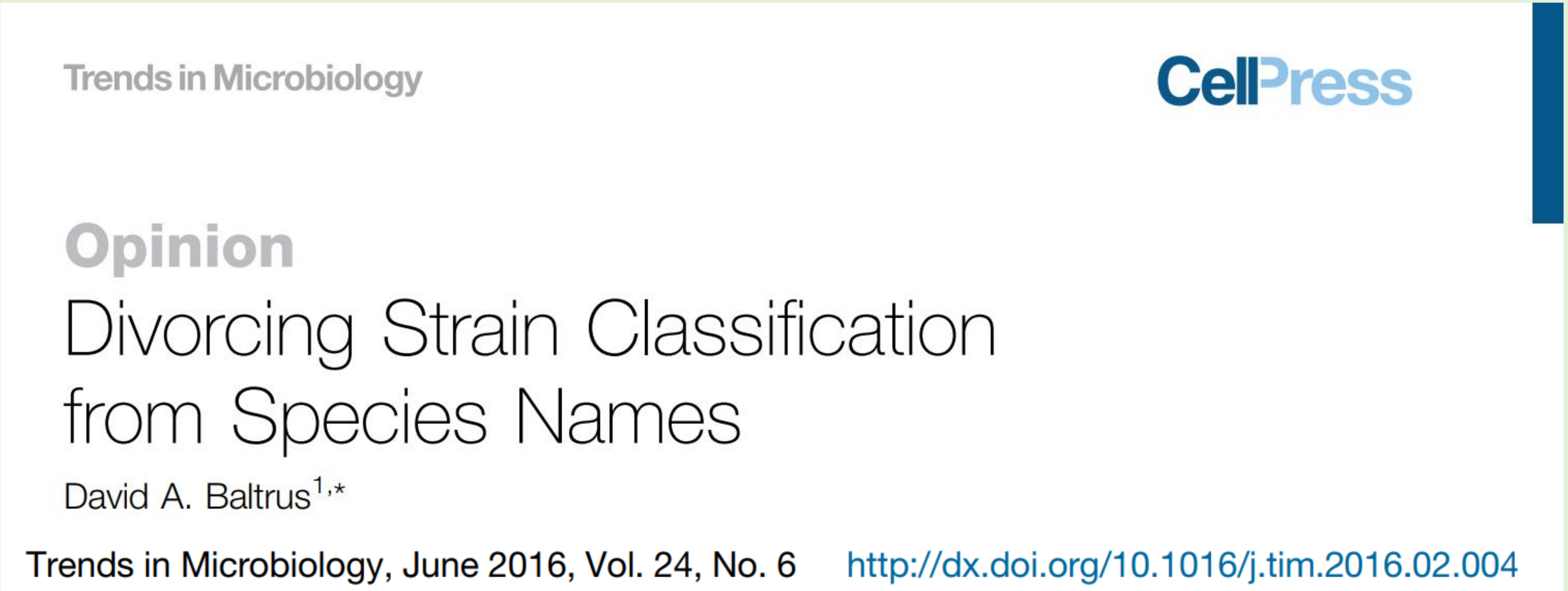
Today

Topics

- The need for **names** in an applied context (food labelling, risk groups of microorganisms, search and discovery in biotechnology)
- Names are the result of **taxonomic studies**
    - **What is a species? How do we circumscribe species?**
    - Identification, classification and nomenclature
    - Procedures and resources
- Evolution in taxonomy: **phylogenetic trees** as tools for inferring relationships among genes and organisms

Be interactive!

The strain is everything



Trends in Microbiology

CellPress

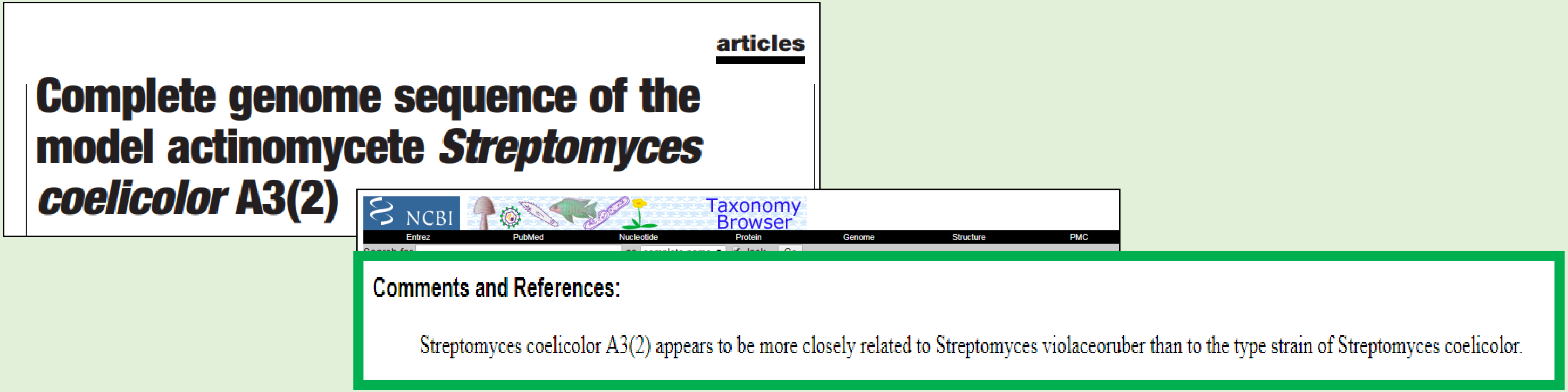
Opinion

Divorcing Strain Classification from Species Names

David A. Baltrus[1,*]

Trends in Microbiology, June 2016, Vol. 24, No. 6 http://dx.doi.org/10.1016/j.tim.2016.02.004

The strain is everything



articles

Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2)

Comments and References:

Streptomyces coelicolor A3(2) appears to be more closely related to Streptomyces violaceoruber than to the type strain of Streptomyces coelicolor.

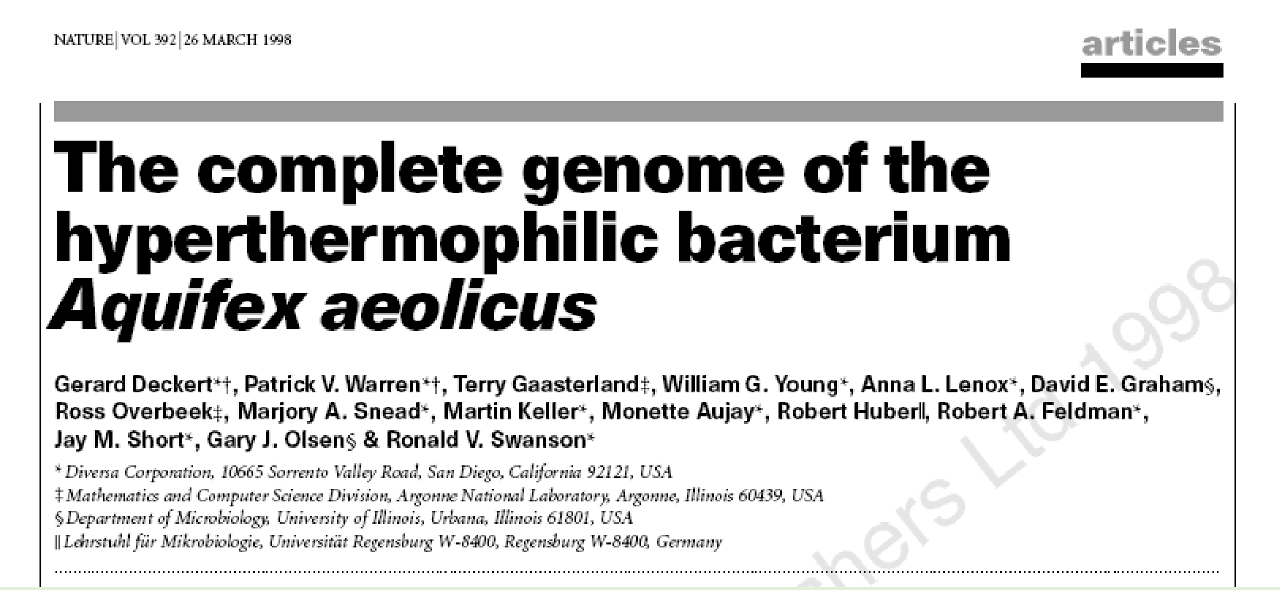
The strain is everything

Aquifex aeolicus VF5 (Nature, 1998)

April 2018:

2670 papers referring to *Aquifex aeolicus* in PubMed Central

(519 in PubMed)



NATURE | VOL 392 | 26 MARCH 1998

articles

The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*

Gerard Deckert*†, Patrick V. Warren*†, Terry Gaasterland‡, William G. Young*, Anna L. Lenox*, David E. Graham§, Ross Overbeek‡, Marjory A. Snead*, Martin Keller*, Monette Aujay*, Robert Huber‖, Robert A. Feldman*, Jay M. Short*, Gary J. Olsen§ & Ronald V. Swanson*

* *Diversa Corporation, 10665 Sorrento Valley Road, San Diego, California 92121, USA*

‡ *Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois 60439, USA*

§ *Department of Microbiology, University of Illinois, Urbana, Illinois 61801, USA*

‖ *Lehrstuhl für Mikrobiologie, Universität Regensburg W-8400, Regensburg W-8400, Germany*

"Aquifex aeolicus" is not a validly published **name**

The strain is everything, but…

“What's in a name? that which we call a rose *by any other name* would smell as sweet...”

The need for **names**

“What's in a name? that which we call a rose by any other name would smell as sweet...”

- **Scientific importance**: conventional way for referring to organisms

Names provide

- a unique framework for scientific communication

- the definition of a “structured knowledge”

The need for **names**

- **Scientific importance**: conventional way for referring to organisms
- **What if we deal with**
  - Pro-technological organisms?
  - Pathogens?
  - Microbiome data?

  Are names important?

  Baltrus (2016) suggested that classification should be independent on nomenclature, based on numerical non-Linnean classification system… we'll see what happens in the future

The need for **names**

- **Scientific importance**: conventional way for referring to organisms
- **Applied** importance:
  - food labelling
  - risk groups of microorganisms
  - search and discovery in biotechnology

The need for **names**

Safety rules and regulations (national and international, public health, environmental laws, intellectual property rights etc.)

- **Risk groups**
- **QPS status**

are LISTS OF NAMES

Links

- https://www.efsa.europa.eu/en/topics/topic/qualified-presumption-safety-qps (https://efsa.onlinelibrary.wiley.com/doi/epdf/10.2903/j.efsa.2018.5131)

- **GRAS** (generally regarded as safe) status (FDA, www.fda.gov/ EFFCA, www.effca.org)
- ABSA: American Biological Safety Association https://my.absa.org/Riskgroups

The need for **names**

- **Scientific importance**: conventional way for referring to organisms
- **Applied** importance:
  - **food labelling**
  - risk groups of microorganisms
  - search and discovery in biotechnology

The need for **names**

- **Scientific names and/or commercial names?**

The need for **names**

- **Scientific importance**: conventional way for referring to organisms
- **Applied** importance:
  - food labelling
  - risk groups of microorganisms
  - **search and discovery in biotechnology**

- **Microbiome data**
- **Colturomic analyses**

Could reveal *novel* organisms… How do I know if this is *NOVEL* or *ALREADY KNOWN?*

The need for **names**

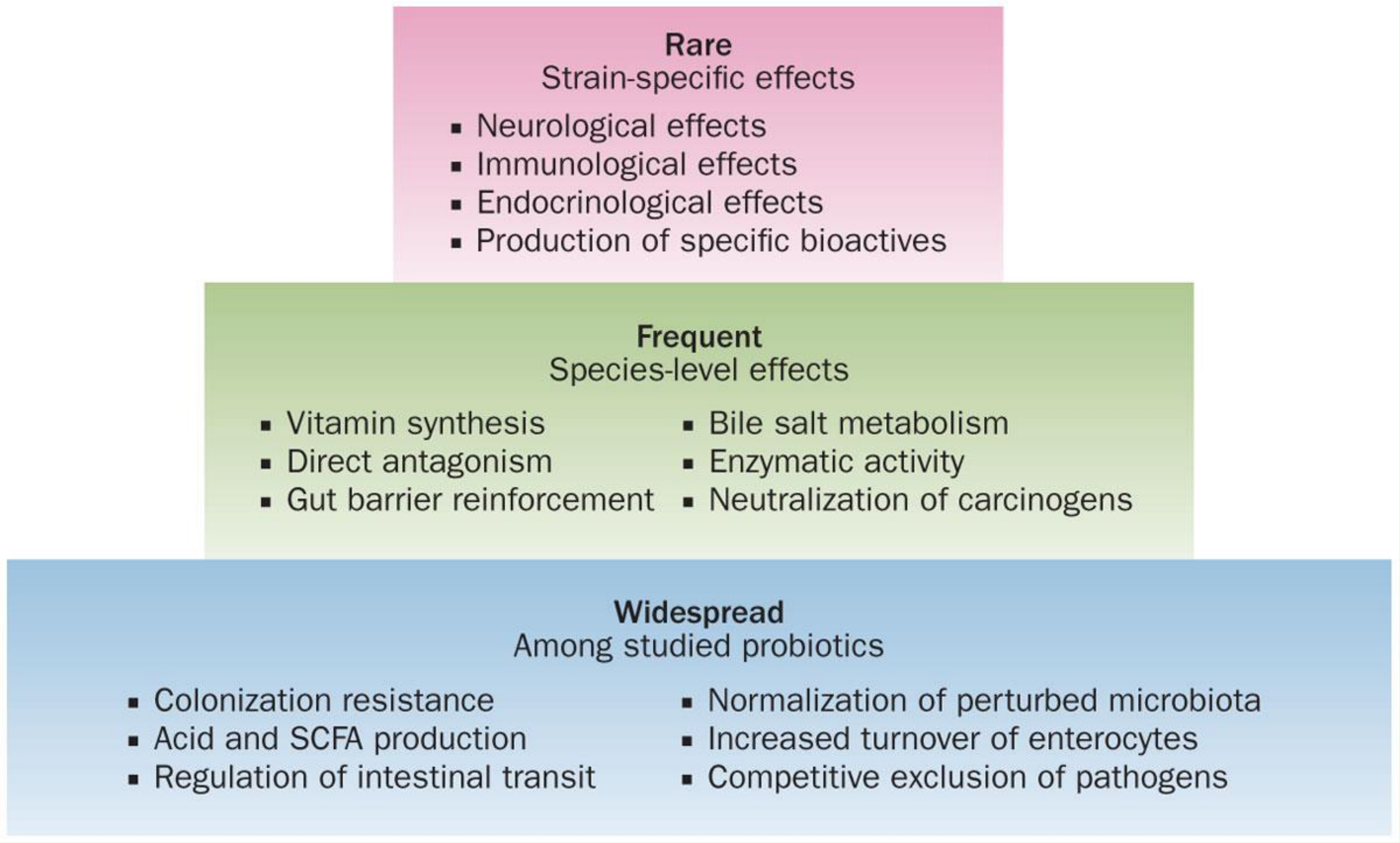
- **Scientific importance**: conventional way for referring to organisms
- **Applied** importance:
  - food labelling
  - risk groups of microorganisms
  - **search and discovery in biotechnology**

- **Microbiome data**
- **Colturomic analyses**

Could reveal *novel* organisms… How do I know if this is *NOVEL* or *ALREADY KNOWN?* → **NAMES and species descriptions!**

A focus on probiotics

Possible distribution of mechanisms



Rare
Strain-specific effects

- Neurological effects
- Immunological effects
- Endocrinological effects
- Production of specific bioactives

Frequent
Species-level effects

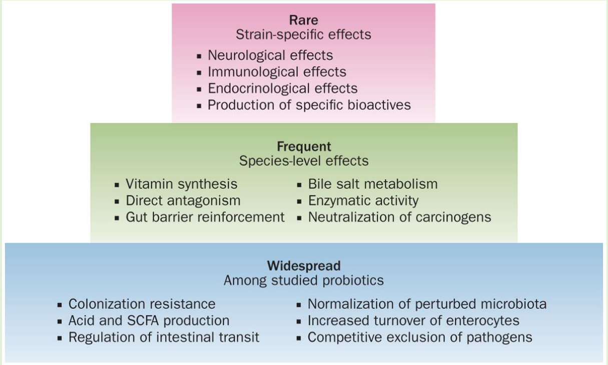
- Vitamin synthesis
- Direct antagonism
- Gut barrier reinforcement
- Bile salt metabolism
- Enzymatic activity
- Neutralization of carcinogens

Widespread
Among studied probiotics

- Colonization resistance
- Acid and SCFA production
- Regulation of intestinal transit
- Normalization of perturbed microbiota
- Increased turnover of enterocytes
- Competitive exclusion of pathogens

The International Scientific Association for Probiotics and Prebiotics consensus statement on the scope and appropriate use of the term probiotic Hill *et al.* 2014 *Nat. Rev. Gastroenterol. Hepatol.* doi:10.1038/nrgastro.2014.66

A focus on probiotics



Rare
Strain-specific effects

- Neurological effects
- Immunological effects
- Endocrinological effects
- Production of specific bioactives

Frequent
Species-level effects

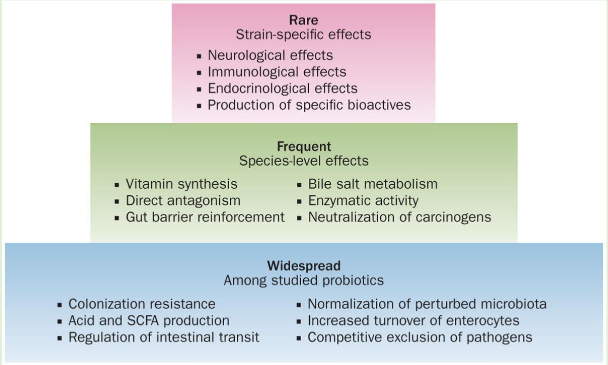
- Vitamin synthesis
- Direct antagonism
- Gut barrier reinforcement
- Bile salt metabolism
- Enzymatic activity
- Neutralization of carcinogens

Widespread
Among studied probiotics

- Colonization resistance
- Acid and SCFA production
- Regulation of intestinal transit
- Normalization of perturbed microbiota
- Increased turnover of enterocytes
- Competitive exclusion of pathogens

- Probiotic effects are generally considered **strain-specific**
- **Strain identity** is important to:
  - link a strain to a specific health effect
  - enable accurate surveillance and epidemiological studies
  - possible **exception** → *S. thermophilus* and *L. delbrueckii* subsp. *bulgaricus* to enhance lactose digestion in lactose intolerant individuals → where there is suitable scientific substantiation of health benefits that are not strain specific, individual strain identity is not critical
- **Speciation** of the bacteria must be established using the **most current, valid methodology**, combination of phenotypic and genetic tests be used.

Genus/species/strain



Rare
Strain-specific effects

- Neurological effects
- Immunological effects
- Endocrinological effects
- Production of specific bioactives

Frequent
Species-level effects

- Vitamin synthesis
- Direct antagonism
- Gut barrier reinforcement
- Bile salt metabolism
- Enzymatic activity
- Neutralization of carcinogens

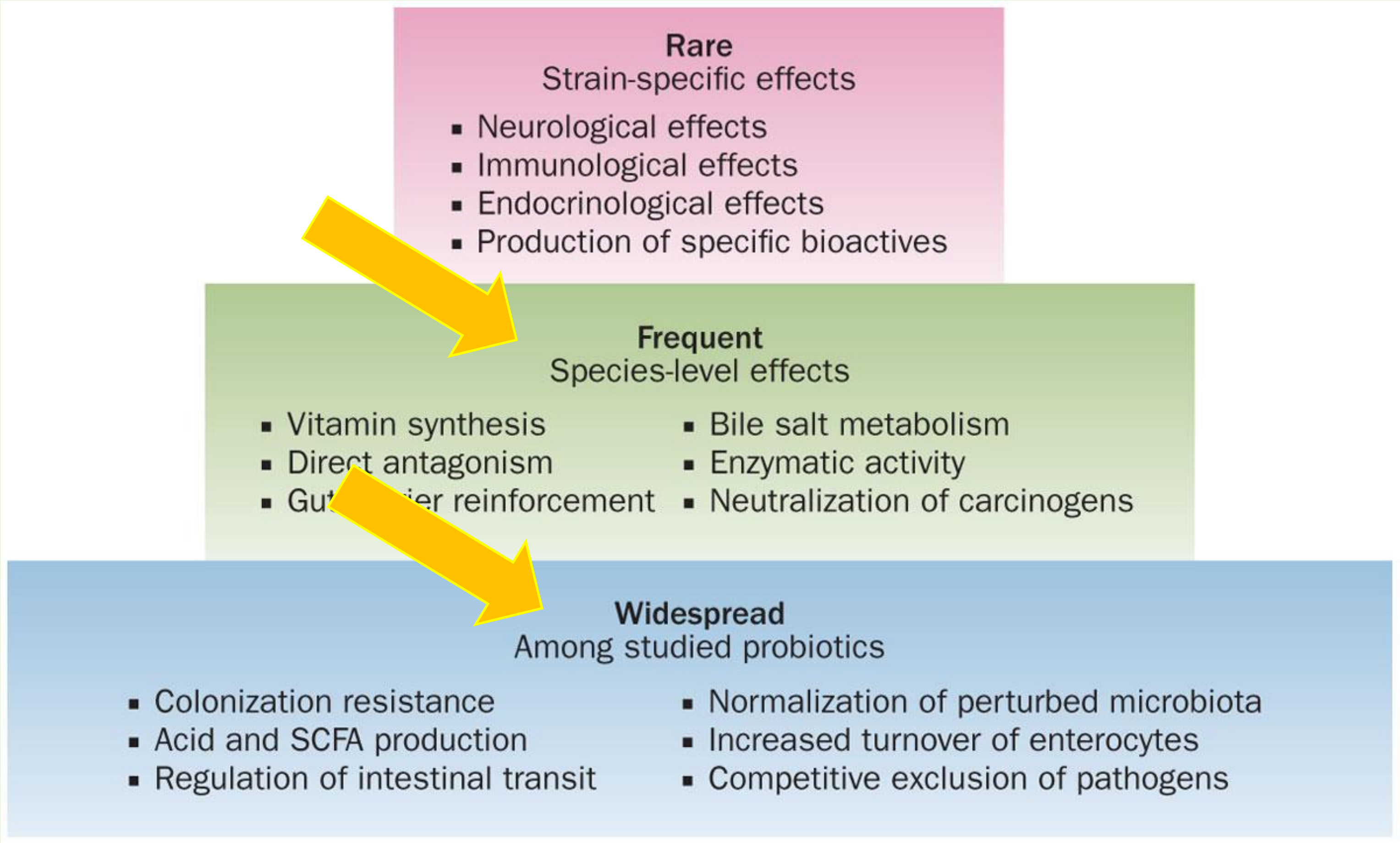
Widespread
Among studied probiotics

- Colonization resistance
- Acid and SCFA production
- Regulation of intestinal transit
- Normalization of perturbed microbiota
- Increased turnover of enterocytes
- Competitive exclusion of pathogens

- **Nomenclature** of the bacteria must conform to the current, scientifically recognized names.
- Protracted use of older or misleading nomenclature is not acceptable on product **labels**
- The use of **incorrect names**
  - does not properly identify the probiotic bacterium in the product
  - forces consumers and regulatory agencies to make assumptions about the identity of the real bacterium being sold.

Probiotics, mechanisms and taxonomic levels

Speciation of the bacteria must be established using the **most current, valid methodology**, combination of phenotypic and genetic tests be used.



Rare
Strain-specific effects

- Neurological effects
- Immunological effects
- Endocrinological effects
- Production of specific bioactives

Frequent
Species-level effects

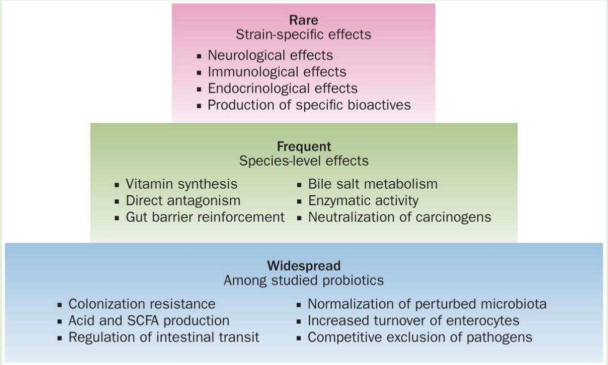
- Vitamin synthesis
- Direct antagonism
- Gut barrier reinforcement
- Bile salt metabolism
- Enzymatic activity
- Neutralization of carcinogens

Widespread
Among studied probiotics

- Colonization resistance
- Acid and SCFA production
- Regulation of intestinal transit
- Normalization of perturbed microbiota
- Increased turnover of enterocytes
- Competitive exclusion of pathogens

Hill *et al.* 2014 *Nat. Rev. Gastroenterol. Hepatol.* doi:10.1038/nrgastro.2014.66

Techniques for identification



Rare
Strain-specific effects

- Neurological effects
- Immunological effects
- Endocrinological effects
- Production of specific bioactives

Frequent
Species-level effects

- Vitamin synthesis
- Direct antagonism
- Gut barrier reinforcement
- Bile salt metabolism
- Enzymatic activity
- Neutralization of carcinogens

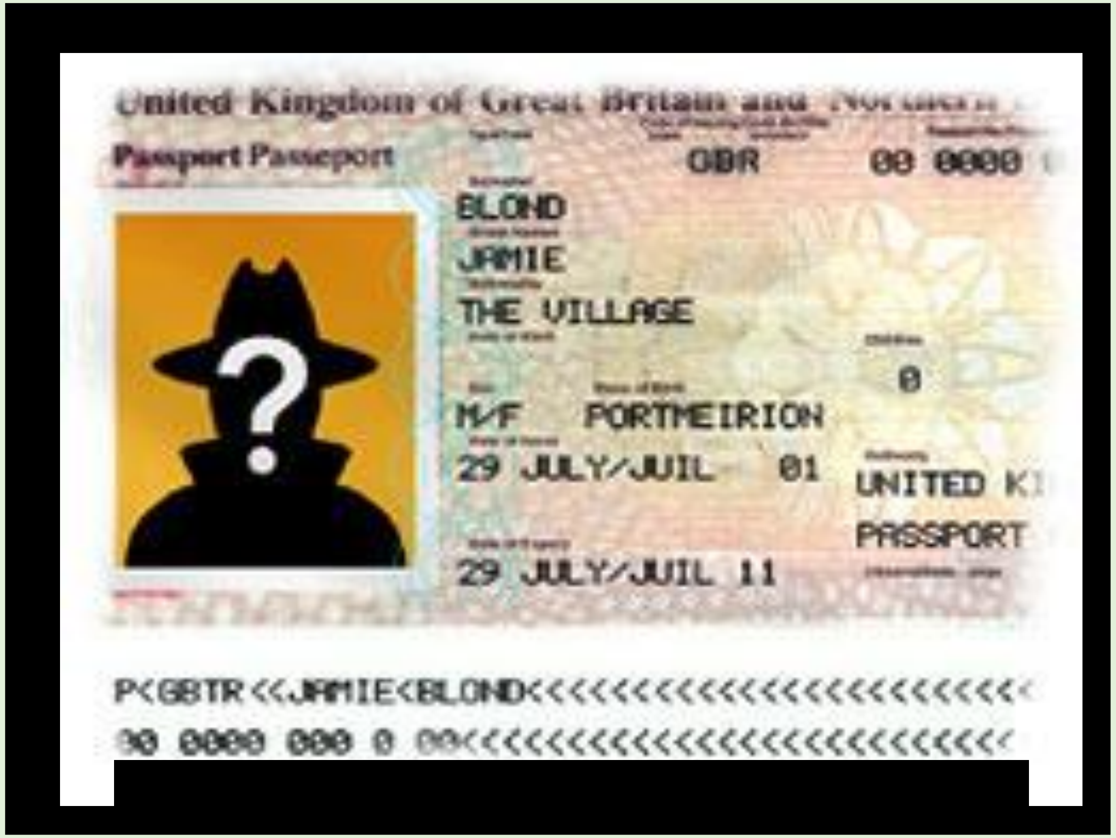
Widespread
Among studied probiotics

- Colonization resistance
- Acid and SCFA production
- Regulation of intestinal transit
- Normalization of perturbed microbiota
- Increased turnover of enterocytes
- Competitive exclusion of pathogens

- **DNA-DNA hybridization**
- **16S rRNA sequencing**, it is recommended that this genotypic technique be combined with **phenotypic** tests for confirmation.
- Patterns generated from the fermentation of a range of sugars and final fermentation products obtained from glucose utilization are **key phenotypes** that should be investigated for identification purposes.
- **Strain typing**
  - Pulsed Field Gel Electrophoresis (PFGE) is the gold standard.
  - Randomly Amplified Polymorphic DNA (RAPD) can also be used, but is less reproducible.
  - Determination of the presence of extrachromosomal genetic elements, such as plasmids can contribute to strain typing and characterization.
- It is recommended that all strains be deposited in an internationally recognized culture collection.

- **Today: genome sequencing, DDH and ANI values calculation**

Taxonomy: what’s in a name?



United Kingdom of Great Britain and Northern

Passport Passeport GBR 00 0000

BLOND

JAMIE

THE VILLAGE

0

M/F PORTMEIRION

29 JULY/JUIL 01

UNITED KI

PASSPORT

29 JULY/JUIL 11

P<GBTR<<JAMIE<BLOND<<<<<<<<<<<<<<<<<<<<<<<<<

00 0000 000 0 00<<<<<<<<<<<<<<<<<<<<<<<<<<<<

Taxonomy

grouping and NAMING of organisms on the basis of SIMILARITY

diversity (ecological concept) **exists**

names (artificial delineation of diversity) are **needed**

Names indicate **species,**
the species is an artificial and pragmatic unit

Keywords

- taxonomy/systematics: 3 inter-related but different sub-disciplines
  - **classification**: involves the recognition of similarities and relationships as a basis for the arrangement of the bacteria into taxonomic groups or taxa. The basic unit is the species
  - **identification**: the recognition of an organism as a member of one of the established taxa, by the comparison of a number of characters with those in the description
  - **nomenclature**: attribution of univocal names to taxa classified and identified

Key points

- classification / identification:
  - dependent on technical advancements
  - intrinsic characteristics of the analysed organisms
  - vary in time
- nomenclature:
  - given a classification scheme, rules are fixed and standard among scientists
  - names could change according to classification
- the species...

What is a (bacterial) species?

Species Concept

idea and theoretical framework that explain

what the unit *species* can be

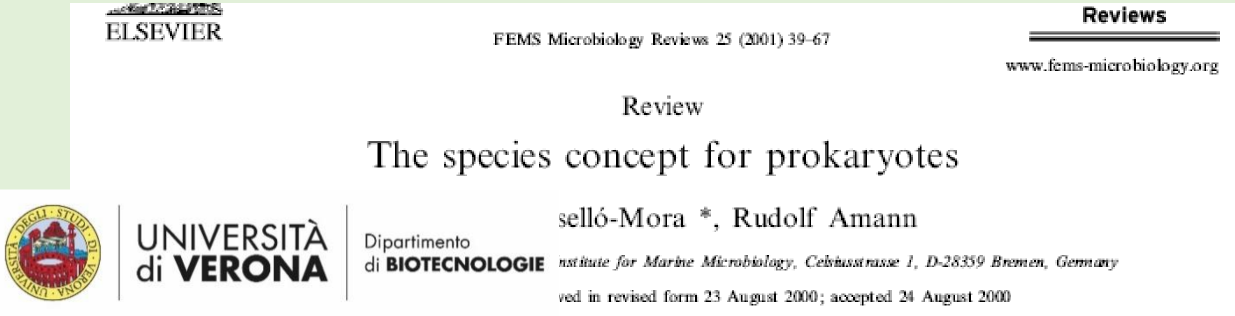
Different interpretations by taxonomists, ecologists, evolutionary biologists!!

Evolving concept

The species concept for prokaryotes

2001

“a **monophyletic** and **genomically coherent** cluster of individual organisms that show a **high degree of overall similarity** in many independent characteristics, and is diagnosable by a **discriminative phenotypic** property”



ELSEVIER

FEMS Microbiology Reviews 25 (2001) 39–67

Reviews

www.fems-microbiology.org

Review

The species concept for prokaryotes

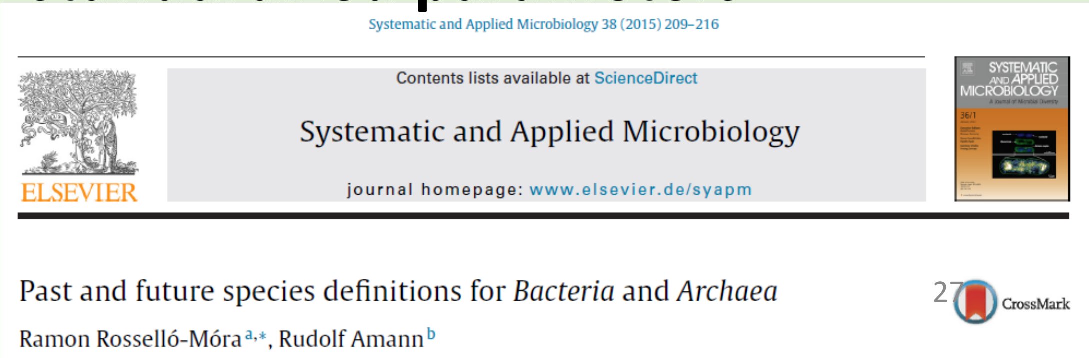
selló-Mora *, Rudolf Amann

nstitute for Marine Microbiology, Celsiusstrasse 1, D-28359 Bremen, Germany

ved in revised form 23 August 2000; accepted 24 August 2000

2015

“a category that circumscribes **monophyletic**, and **genomically** and **phenotypically** coherent populations of individuals that can be clearly discriminated from other such entities by means of **standardized parameters**”



Systematic and Applied Microbiology 38 (2015) 209–216

Contents lists available at ScienceDirect

Systematic and Applied Microbiology

journal homepage: www.elsevier.de/syapm

ELSEVIER

Past and future species definitions for *Bacteria* and *Archaea*

Ramon Rosselló-Móra[a,*], Rudolf Amann[b]

The “species problem”

- **phylosophical** aspect: species **concept**
  - a category or an evolving population?
  - defined by the characteristics that biologists use to identify it? or an evolving entity existing in nature?

- **practical** aspect: species **delineation/definition**
  - how is a species recognized and described?

Species concept-delineation

- Linnean taxonomic scheme is based on species
- higher organisms:
  - the species consists of populations of organisms that can reproduce with one another and that are reproductively isolated from other such populations (Ernst Mayr, Biological Species Concept, 1942)

definition of “organisms” and “sex” for bacteria?

Bacterial organisms and sex

- bacterial “organisms” are the strains:
    - groups of cells (cultures) descending from the division of one cell
    - cells evolve...
- bacterial sex: conjugation, natural competence, HGT, mobile elements, plasmids...

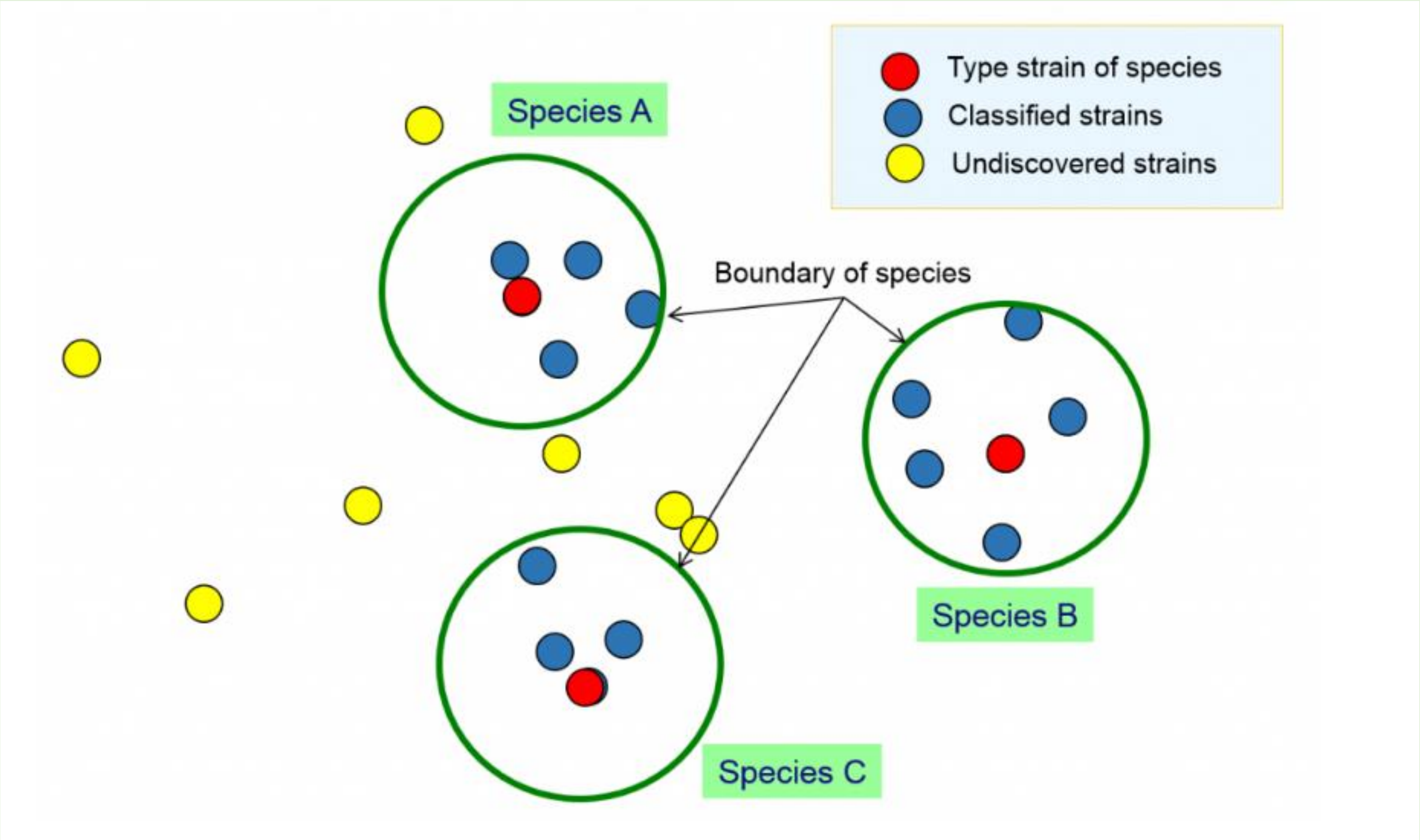
how can we define and delimitate a microbial species?

Species Definition

the way we circumscribe the unit, i.e. compilation of different parameters that allows unequivocal identification

We need a reference point, link between existing diversity and the (artificial) taxonomic scheme

→**type strain**



http://help.bioiplug.com/bacterial-species-concept-explained/

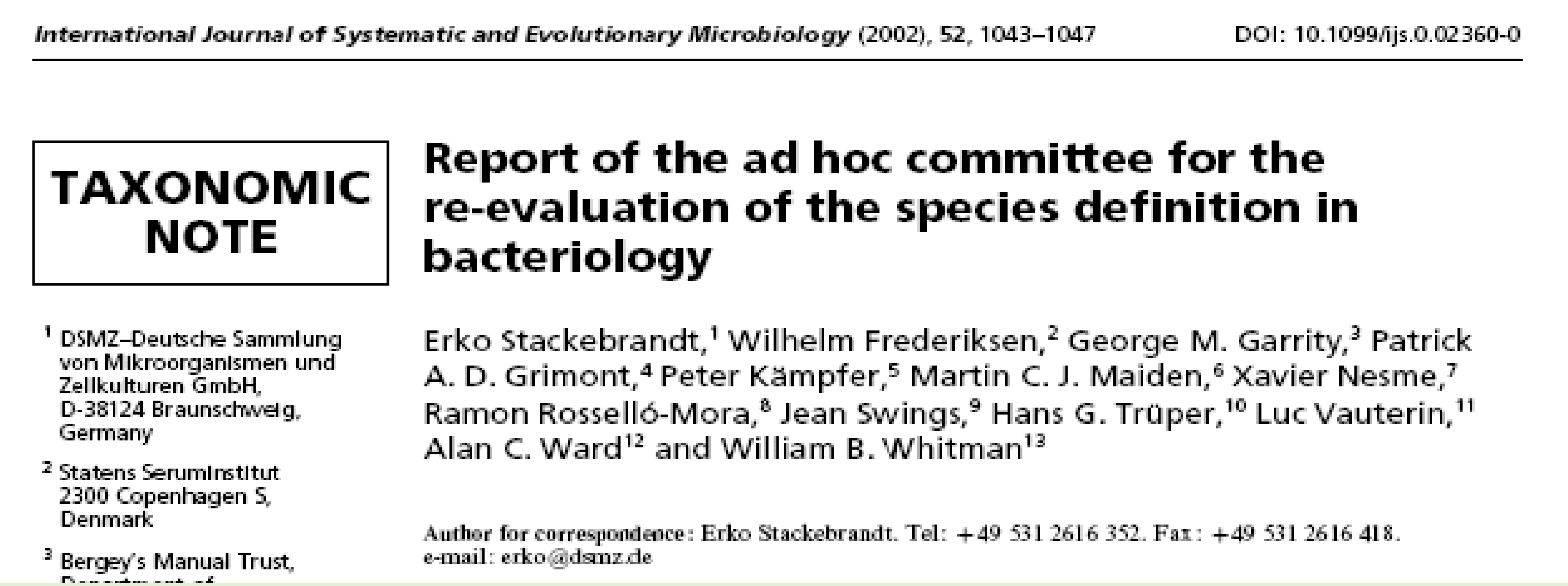
What’s the type strain

- strain to which the **name** of the taxon is permanently attached, definition of the reference point, the link between existing diversity and the artificial taxonomic scheme
- type strain must to be available to the scientific community (deposit in at least TWO culture collections)
- Publication must be on
  - Int J of Systematic and Evolutionary Microbiology (IJSEM)
  - Other journals **+ Validation Lists on IJSEM**

Species is a pragmatic unit

- **DNA-DNA hybridization (DDH)**
- **16S rRNA gene sequence analysis** → useful also for phylogeny

- **Type strain**: strain to which the name of the taxon is permanently attached
- **Techniques** used for species delineation determine similarity → **cut-off values** for identification



International Journal of Systematic and Evolutionary Microbiology (2002), 52, 1043–1047 DOI: 10.1099/ijs.0.02360-0

TAXONOMIC NOTE

Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology

Erko Stackebrandt,[1] Wilhelm Frederiksen,[2] George M. Garrity,[3] Patrick A. D. Grimont,[4] Peter Kämpfer,[5] Martin C. J. Maiden,[6] Xavier Nesme,[7] Ramon Rosselló-Mora,[8] Jean Swings,[9] Hans G. Trüper,[10] Luc Vauterin,[11] Alan C. Ward[12] and William B. Whitman[13]

[1] DSMZ–Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH, D-38124 Braunschweig, Germany

[2] Statens Seruminstitut 2300 Copenhagen S, Denmark

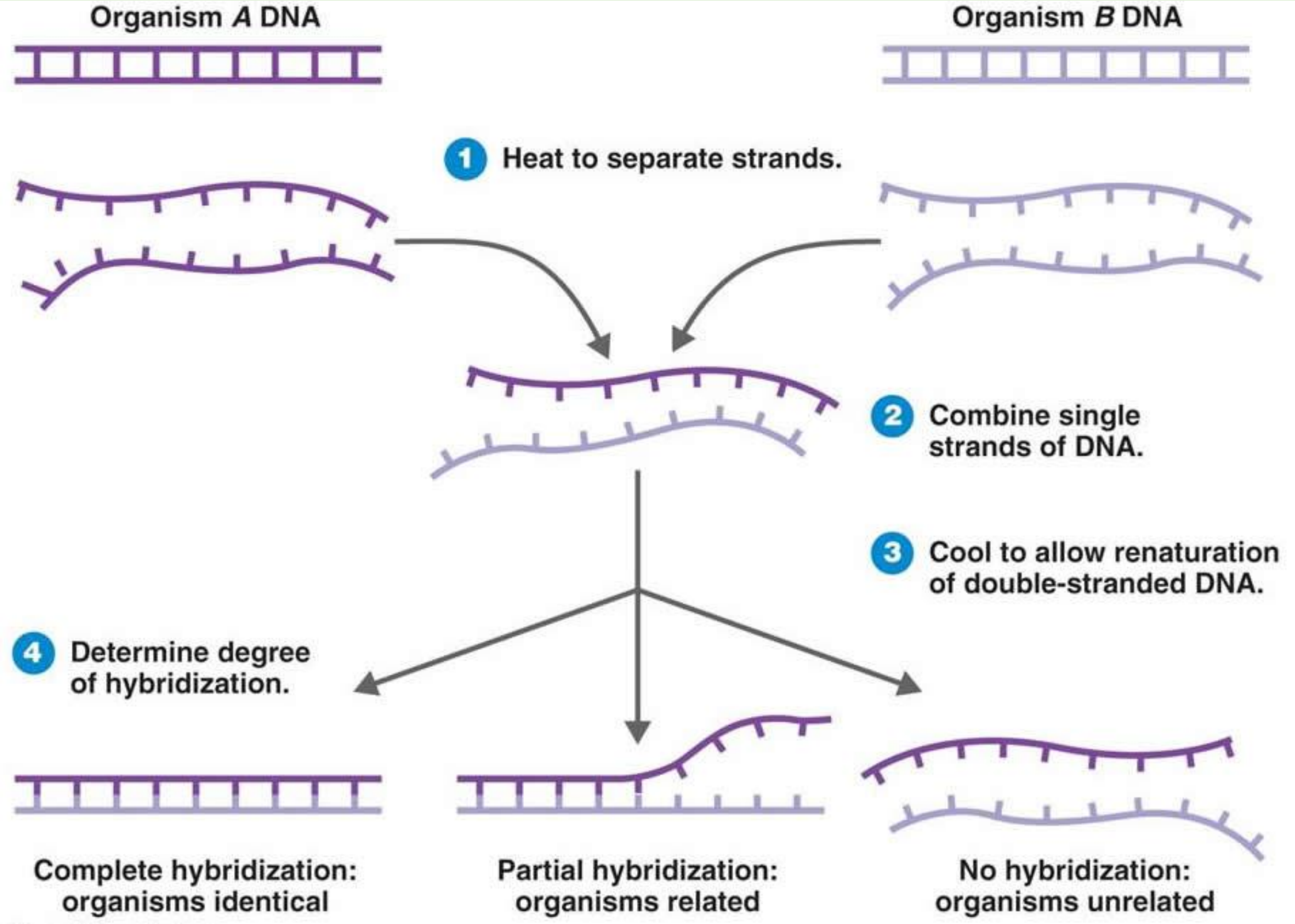
[3] Bergey's Manual Trust, Department of

Author for correspondence: Erko Stackebrandt. Tel: +49 531 2616 352. Fax: +49 531 2616 418. e-mail: erko@dsmz.de

PHYLO-PHENETIC delineation of the bacterial species :

1. phylogeny: 16S rRNA gene sequence analysis
2. overall similarity (>70% DNA-DNA hybridization)
3. distinctive phenotype

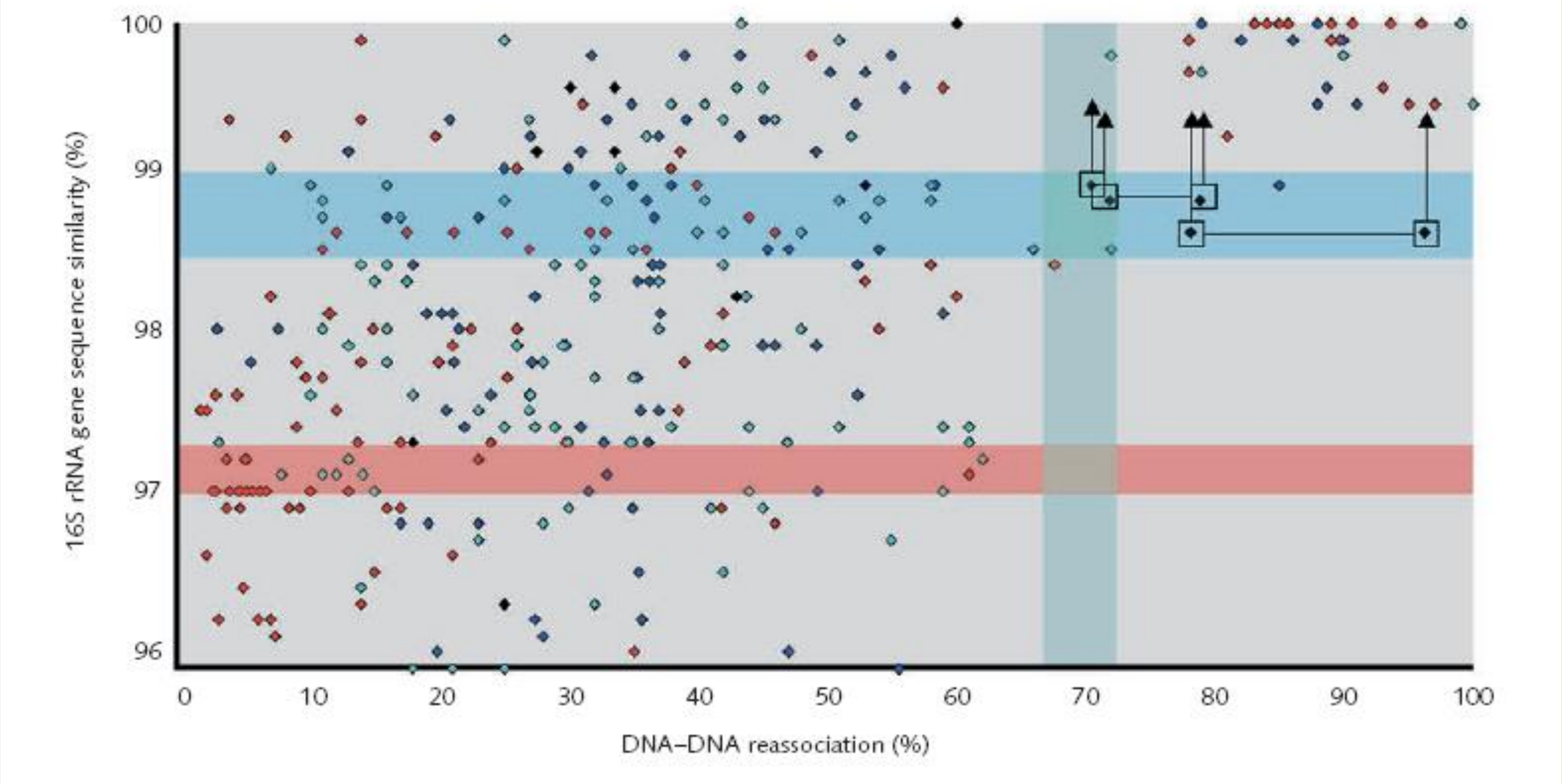
What is DDH?


Organism A DNA
Organism B DNA
1 Heat to separate strands.
2 Combine single strands of DNA.
3 Cool to allow renaturation of double-stranded DNA.
4 Determine degree of hybridization.
Complete hybridization: organisms identical
Partial hybridization: organisms related
No hybridization: organisms unrelated

https://isntsciencewonderful.wordpress.com/2016/04/10/dna-hybridisation-as-cool-as-it-sounds/

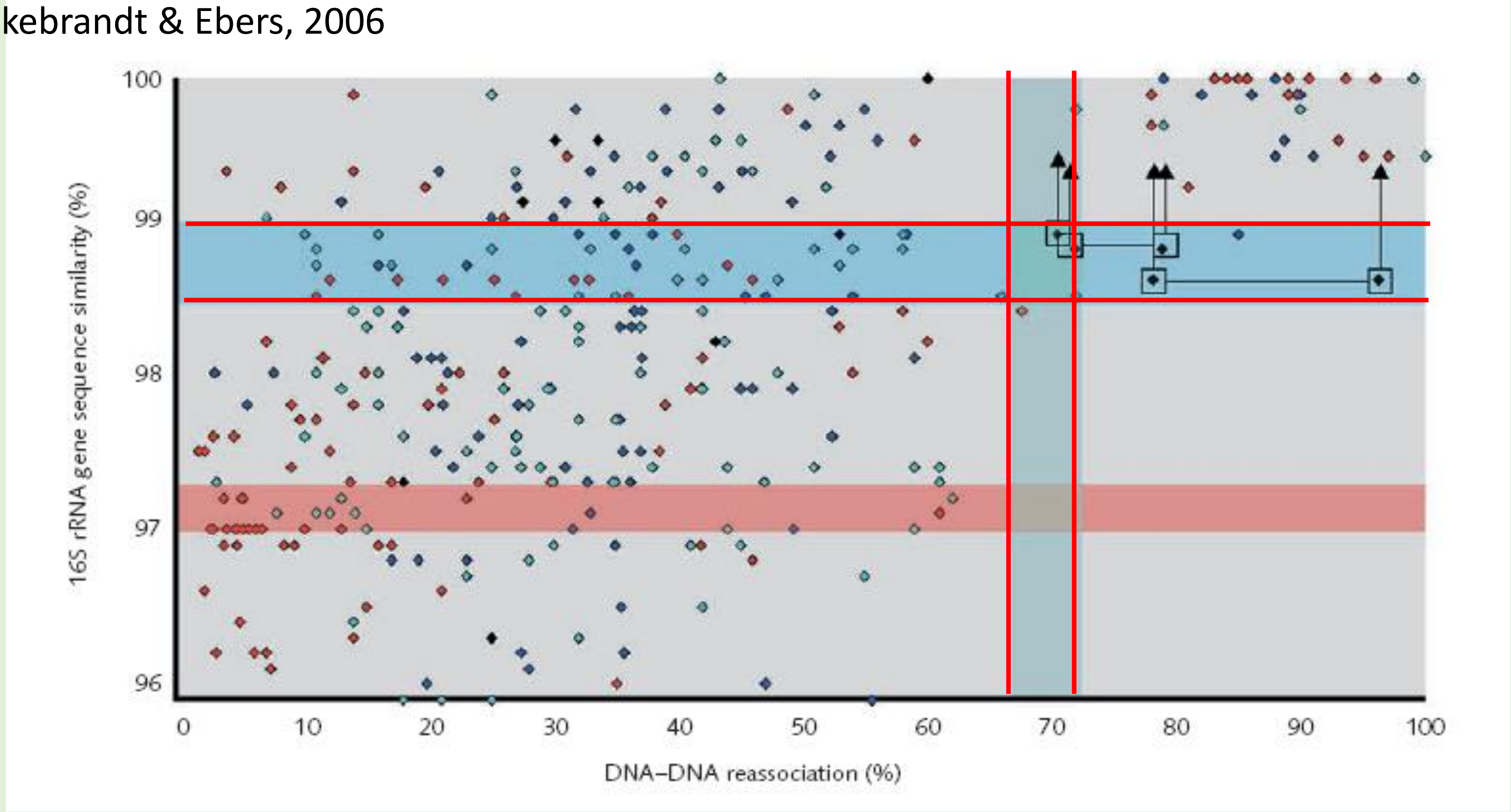
DDH and 16S rRNA gene similarity

Stackebrandt & Ebers, 2006



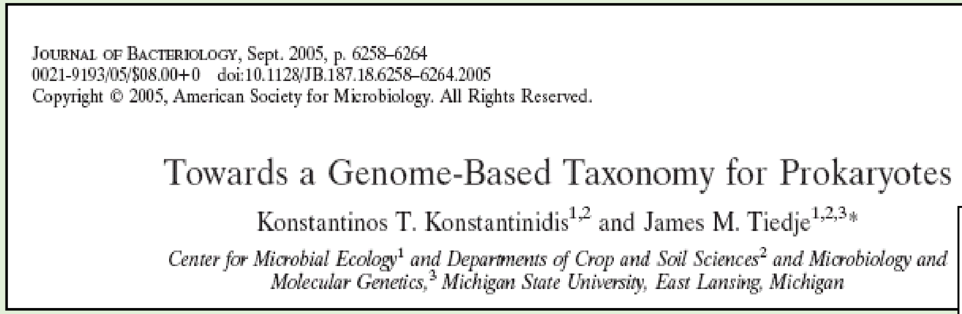
DDH and 16S rRNA gene similarity

Stackebrandt & Ebers, 2006



16S rRNA gene sequence similarity < 98.8%: strains belong to different species

Improvements in genome sequencing

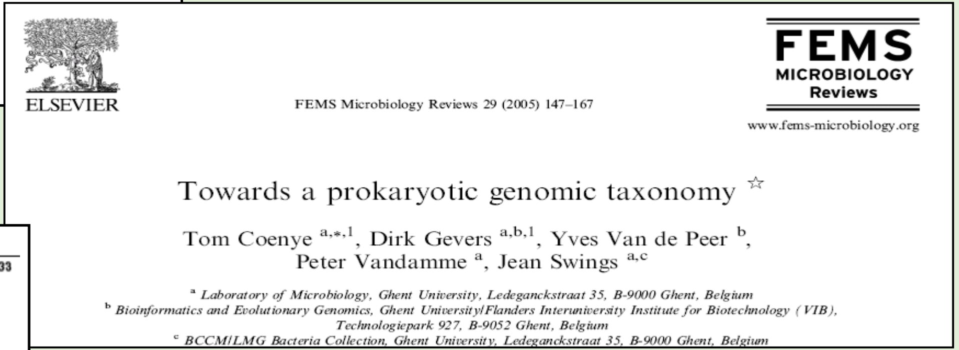


JOURNAL OF BACTERIOLOGY, Sept. 2005, p. 6258–6264 Vol. 187, No. 18
0021-9193/05/$08.00+0 doi:10.1128/JB.187.18.6258–6264.2005
Copyright © 2005, American Society for Microbiology. All Rights Reserved.

Towards a Genome-Based Taxonomy for Prokaryotes

Konstantinos T. Konstantinidis[1,2] and James M. Tiedje[1,2,3]*

Center for Microbial Ecology[1] and Departments of Crop and Soil Sciences[2] and Microbiology and Molecular Genetics,[3] Michigan State University, East Lansing, Michigan



ELSEVIER

FEMS Microbiology Reviews 29 (2005) 147–167

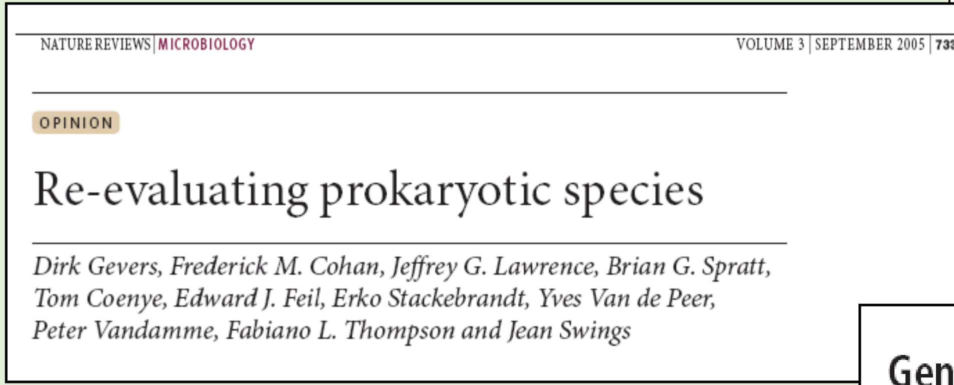
FEMS MICROBIOLOGY Reviews

www.fems-microbiology.org

Towards a prokaryotic genomic taxonomy ☆

Tom Coenye [a,*,1], Dirk Gevers [a,b,1], Yves Van de Peer [b], Peter Vandamme [a], Jean Swings [a,c]

[a] *Laboratory of Microbiology, Ghent University, Ledeganckstraat 35, B-9000 Ghent, Belgium*
[b] *Bioinformatics and Evolutionary Genomics, Ghent University/Flanders Interuniversity Institute for Biotechnology (VIB), Technologiepark 927, B-9052 Ghent, Belgium*
[c] *BCCM/LMG Bacteria Collection, Ghent University, Ledeganckstraat 35, B-9000 Ghent, Belgium*

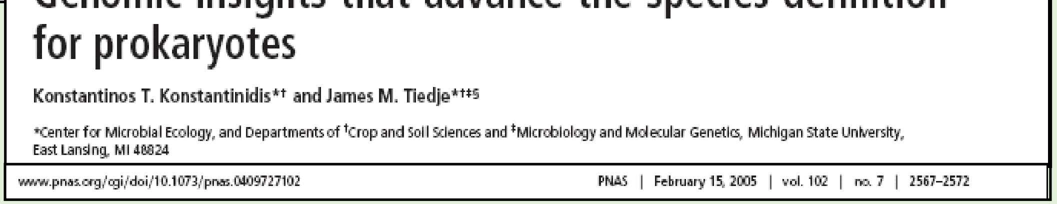


NATURE REVIEWS | MICROBIOLOGY VOLUME 3 | SEPTEMBER 2005 | 733

OPINION

Re-evaluating prokaryotic species

Dirk Gevers, Frederick M. Cohan, Jeffrey G. Lawrence, Brian G. Spratt, Tom Coenye, Edward J. Feil, Erko Stackebrandt, Yves Van de Peer, Peter Vandamme, Fabiano L. Thompson and Jean Swings



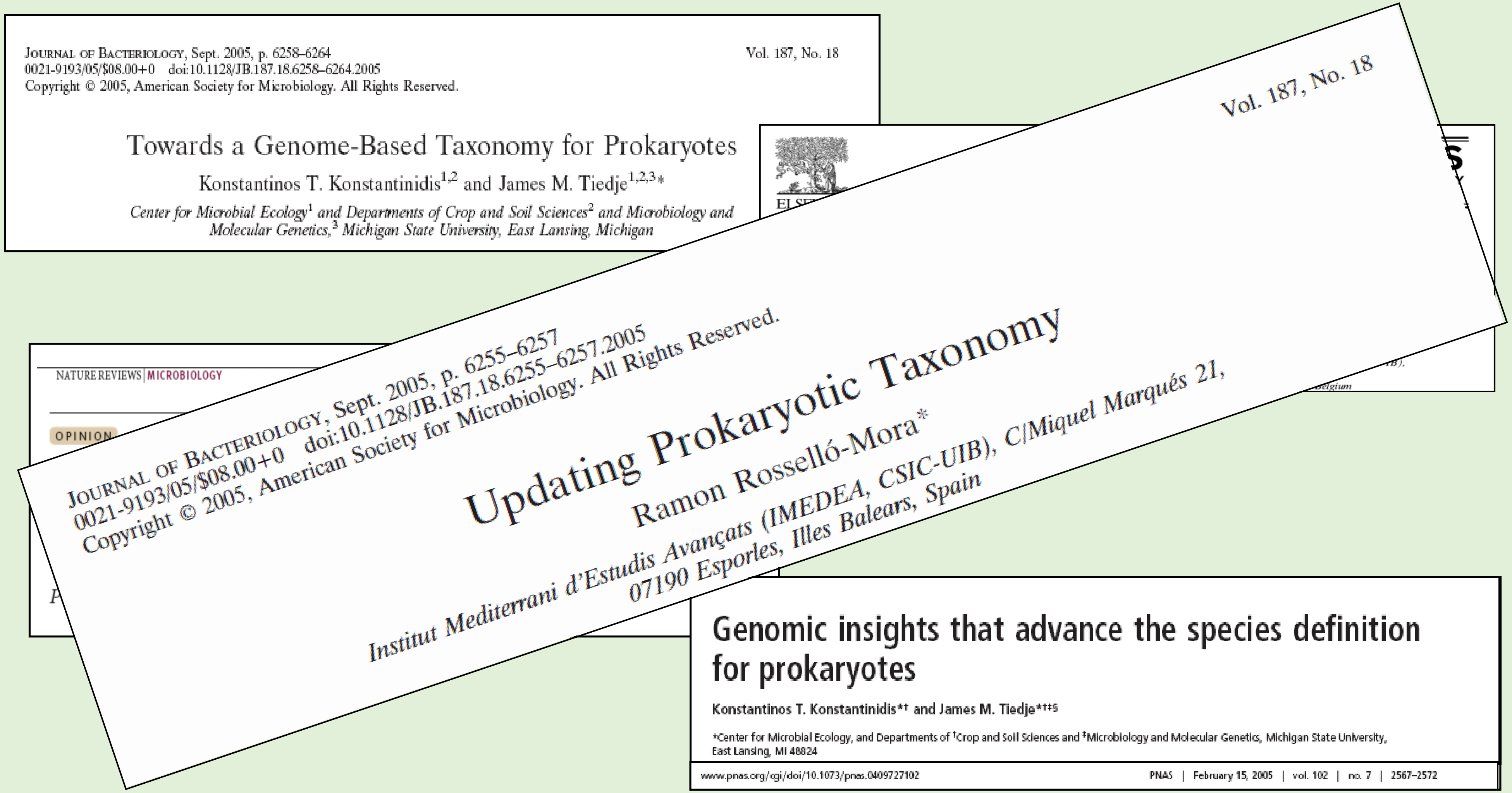
Genomic insights that advance the species definition for prokaryotes

Konstantinos T. Konstantinidis*† and James M. Tiedje*†‡§

*Center for Microbial Ecology, and Departments of †Crop and Soil Sciences and ‡Microbiology and Molecular Genetics, Michigan State University, East Lansing, MI 48824

www.pnas.org/cgi/doi/10.1073/pnas.0409727102 PNAS | February 15, 2005 | vol. 102 | no. 7 | 2567–2572

Improvements in genome sequencing



JOURNAL OF BACTERIOLOGY, Sept. 2005, p. 6258–6264 Vol. 187, No. 18
0021-9193/05/$08.00+0 doi:10.1128/JB.187.18.6258–6264.2005
Copyright © 2005, American Society for Microbiology. All Rights Reserved.

Towards a Genome-Based Taxonomy for Prokaryotes

Konstantinos T. Konstantinidis[1,2] and James M. Tiedje[1,2,3]*

Center for Microbial Ecology[1] and Departments of Crop and Soil Sciences[2] and Microbiology and Molecular Genetics,[3] Michigan State University, East Lansing, Michigan

JOURNAL OF BACTERIOLOGY, Sept. 2005, p. 6255–6257 Vol. 187, No. 18
0021-9193/05/$08.00+0 doi:10.1128/JB.187.18.6255–6257.2005
Copyright © 2005, American Society for Microbiology. All Rights Reserved.

Updating Prokaryotic Taxonomy

Ramon Rosselló-Mora*

Institut Mediterrani d'Estudis Avançats (IMEDEA, CSIC-UIB), C/Miquel Marqués 21, 07190 Esporles, Illes Balears, Spain

NATURE REVIEWS | MICROBIOLOGY

OPINION

Genomic insights that advance the species definition for prokaryotes

Konstantinos T. Konstantinidis*† and James M. Tiedje*‡§

*Center for Microbial Ecology, and Departments of †Crop and Soil Sciences and ‡Microbiology and Molecular Genetics, Michigan State University, East Lansing, MI 48824

www.pnas.org/cgi/doi/10.1073/pnas.0409727102 PNAS | February 15, 2005 | vol. 102 | no. 7 | 2567–2572

2014-2015

Bergey’s International Society for Microbial Systematics (BISMiS) April 7-10th, 2014 Edinburgh, Scotland

BISMiS

BISMiS-2014

Edinburgh, Scotland

Defining Microbial Diversity in the Genomic Era

Int J Syst Evol Microbiol Volume 64, Issue 2, February 2014

Special Collection: **Genomics for Next-Generation Taxonomy and Phylogenetics of Micro-Organisms**

Syst Appl Microbiol Volume 38, Issue 4, June 2015

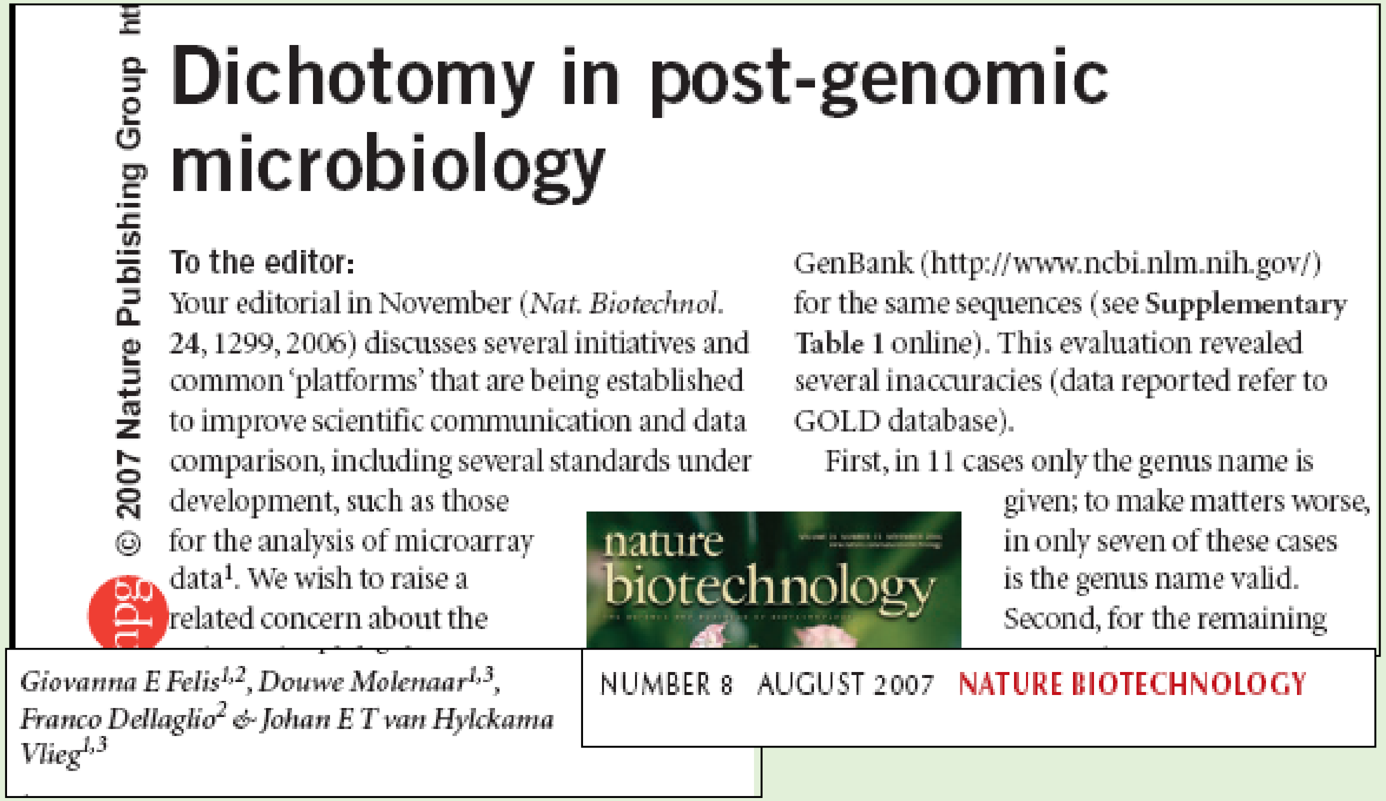
Special issue: **Taxonomy in the age of genomics**

Standardized parameters…

Overall Genome Relatedness Indices:

- **Average Nucleotide Identity (ANI)** (Konstantidinis & Tedje 2005, Goris et al. 2007, Richter & Rossello-Mora 2009)
- **digital DNA-DNA hybridization (dDDH)** (Meier-Kolthoff et al. 2014)
- Maximal Unique Matches (MUM) (Deloger et al. 2009)
- Tetranucleotide signature regression (TETRA) (Richter & Rossello-Mora 2009)
- Average Aminoacid Identity (AAI) (Rodrigues & Konstantinidis 2014)
- Percentage of conserved proteins (POCP) (Qin et al. 2014)

However…



Dichotomy in post-genomic microbiology

To the editor:
Your editorial in November (*Nat. Biotechnol.* 24, 1299, 2006) discusses several initiatives and common 'platforms' that are being established to improve scientific communication and data comparison, including several standards under development, such as those for the analysis of microarray data[1]. We wish to raise a related concern about the

GenBank (http://www.ncbi.nlm.nih.gov/) for the same sequences (see **Supplementary Table 1** online). This evaluation revealed several inaccuracies (data reported refer to GOLD database).

First, in 11 cases only the genus name is given; to make matters worse, in only seven of these cases is the genus name valid. Second, for the remaining

Giovanna E Felis[1,2], *Douwe Molenaar*[1,3], *Franco Dellaglio*[2] *& Johan E T van Hylckama Vlieg*[1,3]

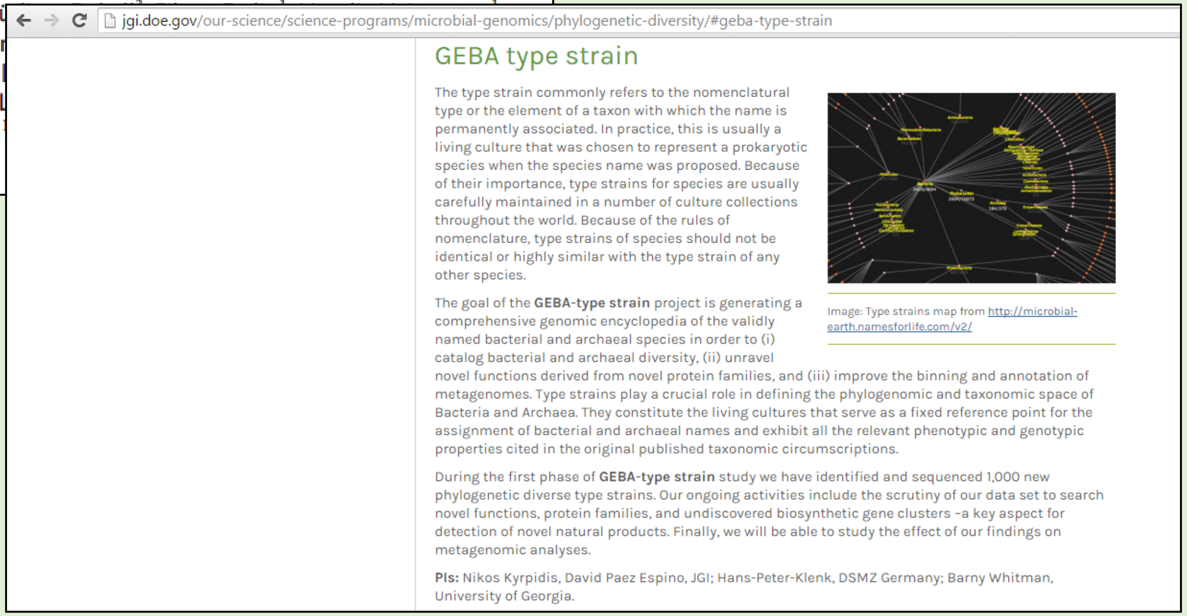
Genome sequencing initiative for the type strains

nature Vol 462|24/31 December 2009|**doi:10.1038/nature08656**

LETTERS

A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea

Dongying Wu[1,2], Philip Hugenholtz[1], Konstantinos Mavromatis[1], Rü
Victor Kunin[1], Lynne Goodwin[4], Martin Wu[5], Brian J. Tindall[3], Sear
Stefan Spring[3], Iain J. Anderson[1], Patrik D'haeseleer[1,6], Adam Zeml
Alex Copeland[1], Cliff Han[4], Feng Chen[1], Jan-Fang Cheng[1], Susan L
Sabine Gronow[3], Patrick Chain[1,4], David Bruce[4], Edward M. Rubin
& Jonathan A. Eisen[1,2]



jgi.doe.gov/our-science/science-programs/microbial-genomics/phylogenetic-diversity/#geba-type-strain

GEBA type strain

The type strain commonly refers to the nomenclatural type or the element of a taxon with which the name is permanently associated. In practice, this is usually a living culture that was chosen to represent a prokaryotic species when the species name was proposed. Because of their importance, type strains for species are usually carefully maintained in a number of culture collections throughout the world. Because of the rules of nomenclature, type strains of species should not be identical or highly similar with the type strain of any other species.

Image: Type strains map from http://microbial-earth.namesforlife.com/v2/

The goal of the **GEBA-type strain** project is generating a comprehensive genomic encyclopedia of the validly named bacterial and archaeal species in order to (i) catalog bacterial and archaeal diversity, (ii) unravel novel functions derived from novel protein families, and (iii) improve the binning and annotation of metagenomes. Type strains play a crucial role in defining the phylogenomic and taxonomic space of Bacteria and Archaea. They constitute the living cultures that serve as a fixed reference point for the assignment of bacterial and archaeal names and exhibit all the relevant phenotypic and genotypic properties cited in the original published taxonomic circumscriptions.

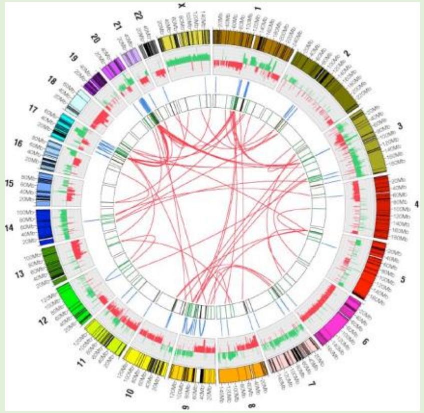
During the first phase of **GEBA-type strain** study we have identified and sequenced 1,000 new phylogenetic diverse type strains. Our ongoing activities include the scrutiny of our data set to search novel functions, protein families, and undiscovered biosynthetic gene clusters –a key aspect for detection of novel natural products. Finally, we will be able to study the effect of our findings on metagenomic analyses.

PIs: Nikos Kyrpidis, David Paez Espino, JGI; Hans-Peter-Klenk, DSMZ Germany; Barny Whitman, University of Georgia.

However less than 50 % of species with validly published names are represented by genome sequences of their type strains “as of the time of writing” (Chun et al., 2018)

UNIVERSITÀ di VERONA Dipartimento di BIOTECNOLOGIE

Species delineation



- Phylo-phenetic approach:
  - phylogeny: **16S** rRNA gene sequence analysis
  - overall similarity (>70% **DNA-DNA hybridization**)
  - distinctive phenotype

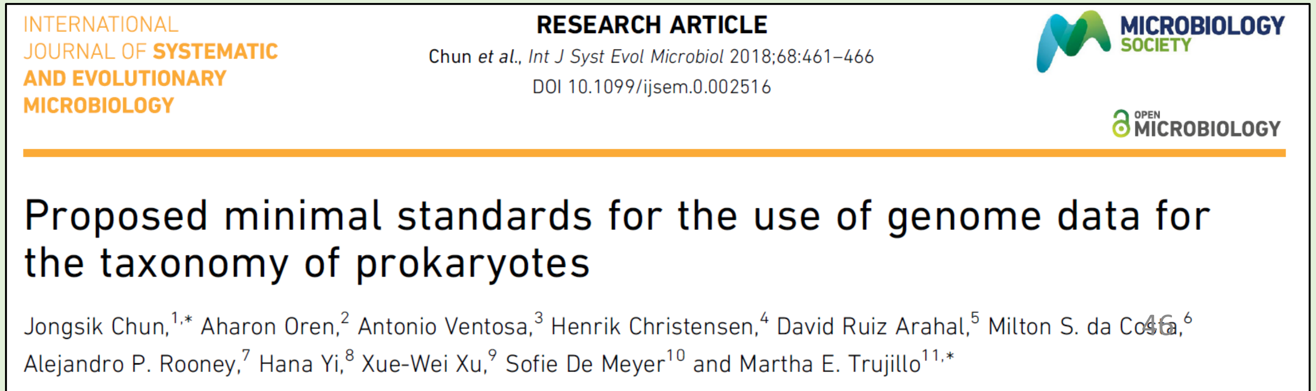
Overall **G**enome **R**elatedness **I**ndices (**OGRI**):

- **Average Nucleotide Identity (ANI)** (Konstantidinis & Tedje 2005, Goris et al. 2007, Richter & Rossello-Mora 2009)
- **digital DNA-DNA hybridization (dDDH)** (Meier-Kolthoff et al. 2014)

Phenotypic characterization

Overall genome related index (OGRI)

- values analogous to DDH values; similarity or distance
- OGRIs can be used to check if a strain belongs to a known species by calculating the relatedness between genome sequences of the strains and type strain of a species
- generally accepted species boundaries
  - for ANI, 95~96%
  - dDDH 70 %



INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY

RESEARCH ARTICLE

Chun *et al*., *Int J Syst Evol Microbiol* 2018;68:461–466

DOI 10.1099/ijsem.0.002516

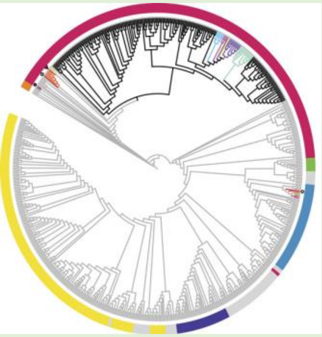
MICROBIOLOGY SOCIETY

OPEN MICROBIOLOGY

Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes

Jongsik Chun,[1,*] Aharon Oren,[2] Antonio Ventosa,[3] Henrik Christensen,[4] David Ruiz Arahal,[5] Milton S. da Costa,[6] Alejandro P. Rooney,[7] Hana Yi,[8] Xue-Wei Xu,[9] Sofie De Meyer[10] and Martha E. Trujillo[11,*]

ANI- Average Nucleotide Identity



Measure of nucleotide-level genomic similarity between the coding regions of two genomes

Important elements

→Sequence identity

→Coverage

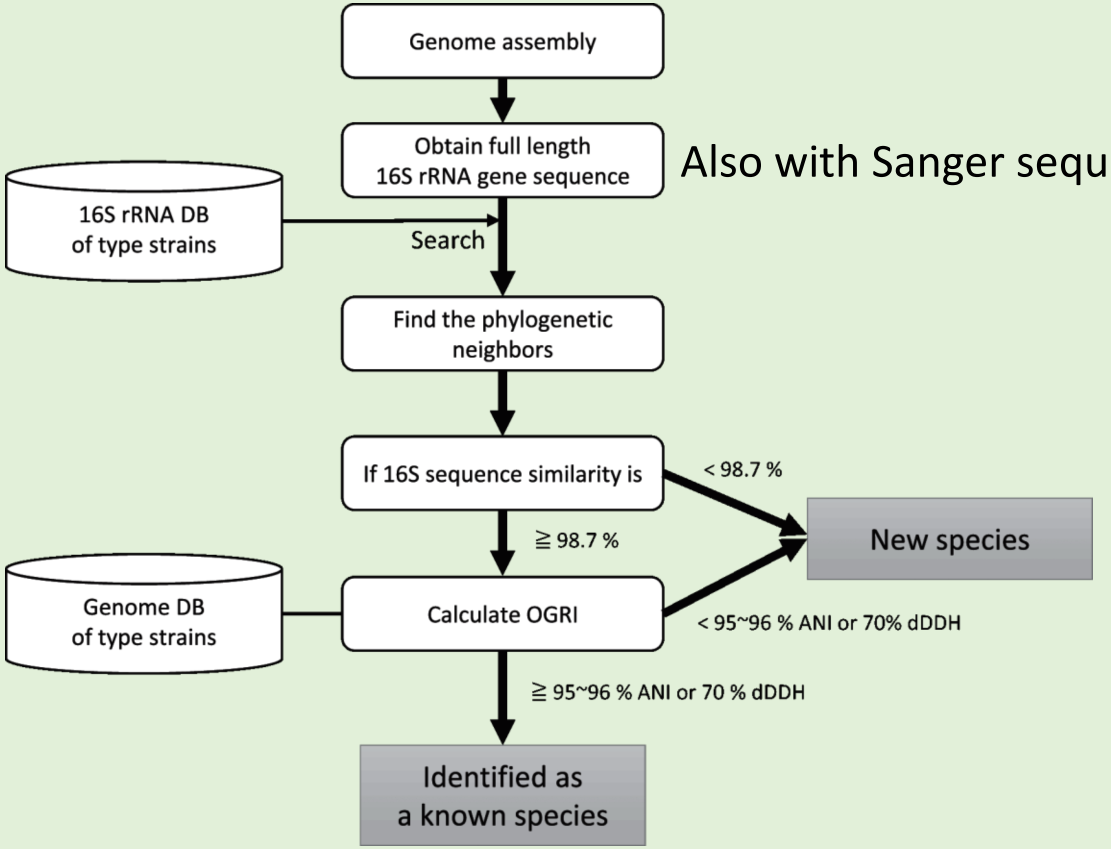
- Completeness of the genomes

However less than 50% of genomes of the type strains of validly described species is available (almost complete database of 16S rRNA gene sequences of the type strains)

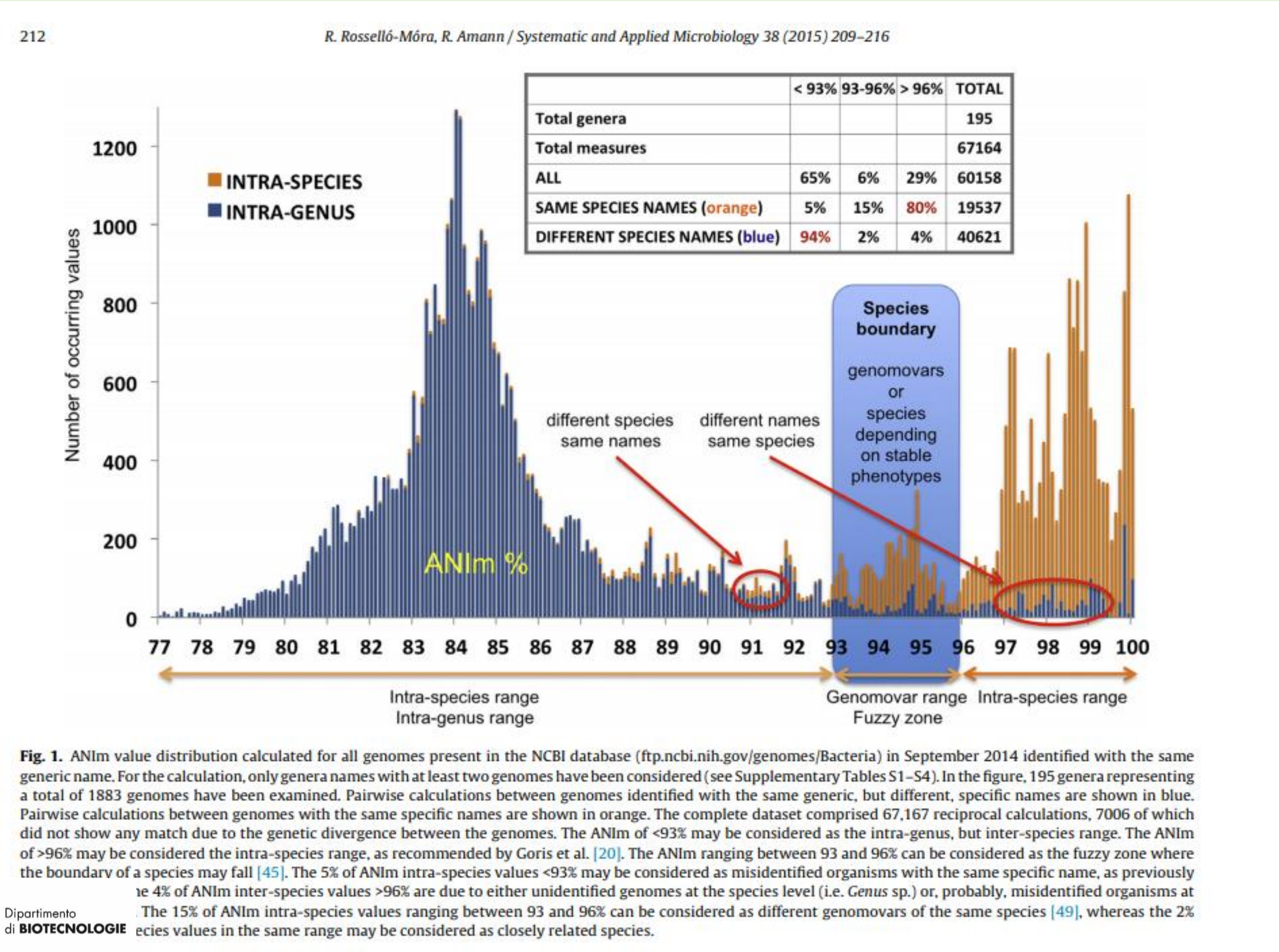
Identification in the genomic era (Chun et al., 2018)

→combination of 16S similarity and OGRI can be used

- Use of **98.7 %** as cutoff (assurance in the quality of 16S sequences)
  - if genome sequence data of the type strains of the hit species are not available, it is recommended to obtain it


Genome assembly
Obtain full length 16S rRNA gene sequence
16S rRNA DB of type strains
Search
Find the phylogenetic neighbors
If 16S sequence similarity is
< 98.7 %
≧ 98.7 %
New species
Genome DB of type strains
Calculate OGRI
< 95~96 % ANI or 70% dDDH
≧ 95~96 % ANI or 70 % dDDH
Identified as a known species

Also with Sanger sequencing



212 R. Rosselló-Móra, R. Amann / Systematic and Applied Microbiology 38 (2015) 209–216

	< 93%	93-96%	> 96%	TOTAL
Total genera				195
Total measures				67164
ALL	65%	6%	29%	60158
SAME SPECIES NAMES (orange)	5%	15%	80%	19537
DIFFERENT SPECIES NAMES (blue)	94%	2%	4%	40621

Fig. 1. ANIm value distribution calculated for all genomes present in the NCBI database (ftp.ncbi.nih.gov/genomes/Bacteria) in September 2014 identified with the same generic name. For the calculation, only genera names with at least two genomes have been considered (see Supplementary Tables S1–S4). In the figure, 195 genera representing a total of 1883 genomes have been examined. Pairwise calculations between genomes identified with the same generic, but different, specific names are shown in blue. Pairwise calculations between genomes with the same specific names are shown in orange. The complete dataset comprised 67,167 reciprocal calculations, 7006 of which did not show any match due to the genetic divergence between the genomes. The ANIm of <93% may be considered as the intra-genus, but inter-species range. The ANIm of >96% may be considered the intra-species range, as recommended by Goris et al. [20]. The ANIm ranging between 93 and 96% can be considered as the fuzzy zone where the boundary of a species may fall [45]. The 5% of ANIm intra-species values <93% may be considered as misidentified organisms with the same specific name, as previously ne 4% of ANIm inter-species values >96% are due to either unidentified genomes at the species level (i.e. *Genus* sp.) or, probably, misidentified organisms at . The 15% of ANIm intra-species values ranging between 93 and 96% can be considered as different genomovars of the same species [49], whereas the 2% ecies values in the same range may be considered as closely related species.

Nomenclature

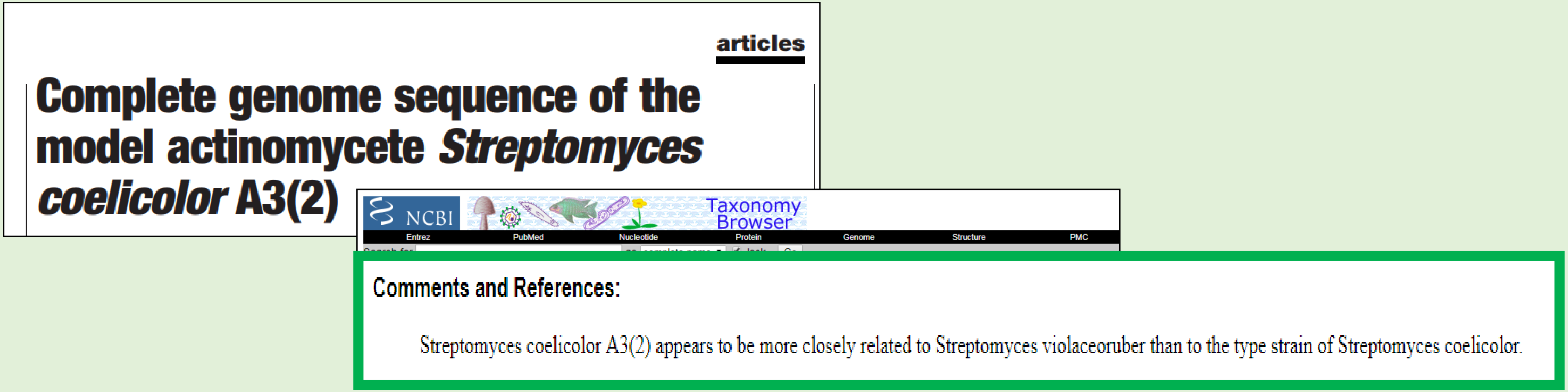
Classification is hierachical

Taxonomic rank/Suffix		Example
• **Phylum**		
• **Class**		
• **Order**	-ales	Pseudomonadales
Suborder	-ineae	Pseudomonadineae
Family	-aceae	Pseudomonadaceae
Subfamily	-oideae	Pseudomonadoideae
Tribe	-eae	Pseudomonadeae
Subtribe	-inae	Pseudomonadinae
Genus	—	*Pseudomonas*
(Subgenus)	—	(not for *Pseudomonas)*
Species	—	*Pseudomonas fluorescens*
Subspecies	—	*Pseudomonas pseudoalcaligenes* subsp. *citrulli*
Biovar	—	*Pseudomonas fluorescens* biovar I
Pathovar	—	*Pseudomonas syringae* pathovar *tabaci*

Bergey’ s Manual of Systematic Bacteriology

Taxonomic outlines are available online

The strain is everything



articles

Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2)

Comments and References:

Streptomyces coelicolor A3(2) appears to be more closely related to Streptomyces violaceoruber than to the type strain of Streptomyces coelicolor.

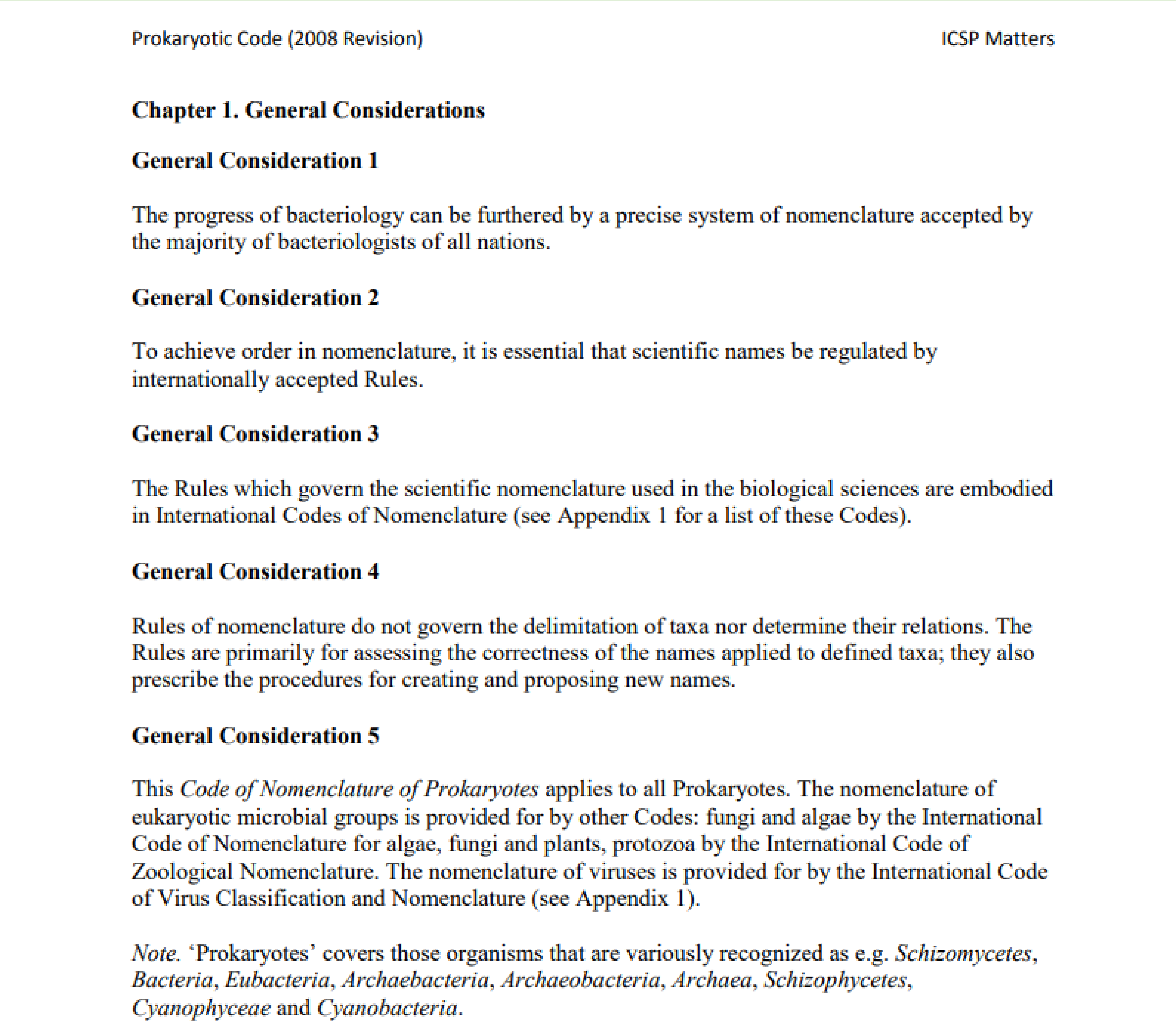
Nomenclature

- "is one step in an information management system, the scope of which is only limited by the bounds of the methods available for studying the organisms themselves and our ability to interpret and comprehend that information"- preface to the Prokaryotic Code (2008 Revision)
- "The International Code of Nomenclature of Prokaryotes is an instrument of scientific communication. Names have meaning only in the context in which they were formed and used" – general recommendation 8

International Code of Nomenclature of Prokaryotes

Cited as the **“Prokaryotic Code (2008 Revision)“**

Applied from the date of publication (2016).



Chapter 1. General Considerations

General Consideration 1

The progress of bacteriology can be furthered by a precise system of nomenclature accepted by the majority of bacteriologists of all nations.

General Consideration 2

To achieve order in nomenclature, it is essential that scientific names be regulated by internationally accepted Rules.

General Consideration 3

The Rules which govern the scientific nomenclature used in the biological sciences are embodied in International Codes of Nomenclature (see Appendix 1 for a list of these Codes).

General Consideration 4

Rules of nomenclature do not govern the delimitation of taxa nor determine their relations. The Rules are primarily for assessing the correctness of the names applied to defined taxa; they also prescribe the procedures for creating and proposing new names.

General Consideration 5

This *Code of Nomenclature of Prokaryotes* applies to all Prokaryotes. The nomenclature of eukaryotic microbial groups is provided for by other Codes: fungi and algae by the International Code of Nomenclature for algae, fungi and plants, protozoa by the International Code of Zoological Nomenclature. The nomenclature of viruses is provided for by the International Code of Virus Classification and Nomenclature (see Appendix 1).

Note. ‘Prokaryotes’ covers those organisms that are variously recognized as e.g. *Schizomycetes*, *Bacteria*, *Eubacteria*, *Archaebacteria*, *Archaeobacteria*, *Archaea*, *Schizophycetes*, *Cyanophyceae* and *Cyanobacteria*.

General Consideration 6

Code is divided into

- Principles
- Rules
- Recommendations

General Consideration 6

Code is divided into

- Principles
- Rules
- Recommendations

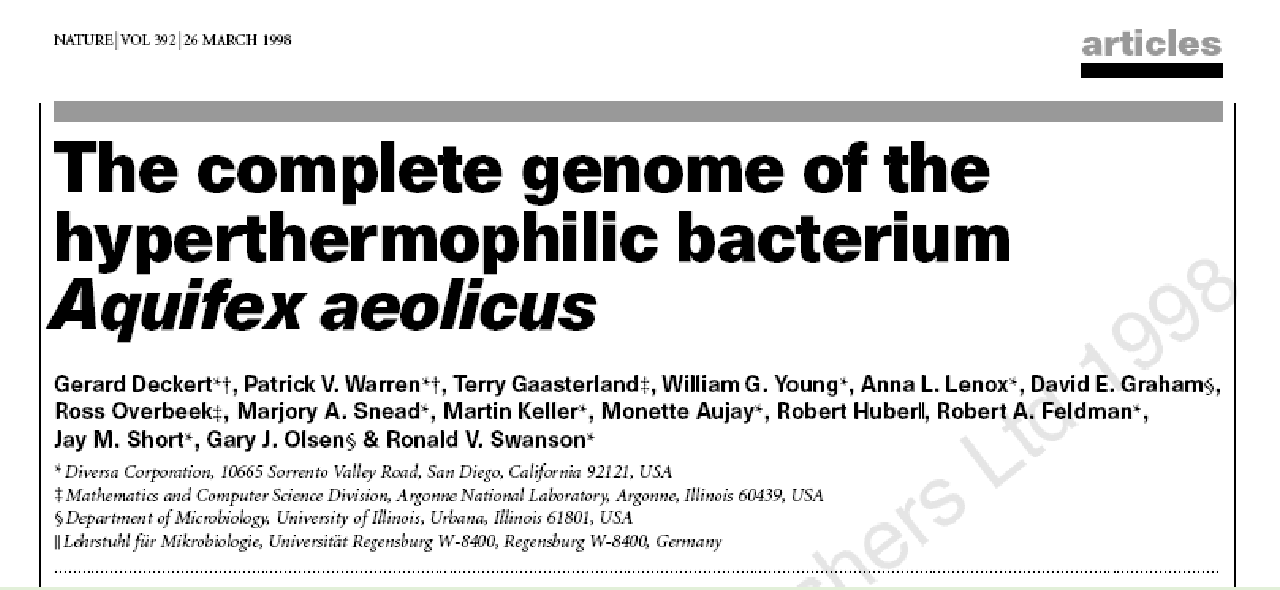
1. **Principles** (Chapter 2) form the **basis of the Code**, and the Rules and Recommendations are derived from them.
2. **Rules** (Chapter 3) are
   - designed to make effective the Principles,
   - to put the nomenclature of the past in order, and
   - to provide for the nomenclature of the future.
3. **Recommendations** (Chapter 3) deal with subsidiary points and are appended to the Rules which they supplement. Recommendations do not have the force of Rules, intended to be guides to desirable practice in the future

The strain is everything

Aquifex aeolicus VF5 (Nature, 1998)

April 2018:

2670 papers referring to *Aquifex aeolicus* in PubMed Central (519 in PubMed)



articles

The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*

Gerard Deckert*†, Patrick V. Warren*†, Terry Gaasterland‡, William G. Young*, Anna L. Lenox*, David E. Graham§, Ross Overbeek‡, Marjory A. Snead*, Martin Keller*, Monette Aujay*, Robert Huber‖, Robert A. Feldman*, Jay M. Short*, Gary J. Olsen§ & Ronald V. Swanson*

* Diversa Corporation, 10665 Sorrento Valley Road, San Diego, California 92121, USA
‡ Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois 60439, USA
§ Department of Microbiology, University of Illinois, Urbana, Illinois 61801, USA
‖ Lehrstuhl für Mikrobiologie, Universität Regensburg W-8400, Regensburg W-8400, Germany

***"Aquifex aeolicus"* is not a validly published name**

Valid publication of new names: fulfillment of requirements (rules 27, 30 and others)

among others:

- list of the strains included in the species
- characteristics of each strain, traits essential of the species, diagnostic characteristics
- designation of the **type strain** for that species

Subspecies

A species may be divided into **subspecies,**

- minor but consistent phenotypic variations within the species or
- genetically determined clusters of strains within the species

Variety is a synonym of subspecies; its use is not encouraged as it leads to confusion

Taxa below the rank of subspecies (**infrasubspecific subdivisions**) are **not covered** by the Rules of the Code

Where to find updated names?

List of Prokaryotic names with standing in Nomenclature

- LPSN http://www.bacterio.net/

Reference for classification *Bergey's Manual of Systematics of Archaea and Bacteria (BMSAB)* - Bergey’s manual Taxonomic Outline

- https://wol-prod-cdn.literatumonline.com/pb-assets/assets/9781118960608/Taxonomic_Outline_October_2017-1507044705000.pdf

- SILVA- living Tree
  - https://www.arb-silva.de/fileadmin/silva_databases/living_tree/LTP_release_123/LSU_release_02_2017/LTPs123_LSU_tree.pdf

Bioinformatics for taxonomic purposes

Bioinformatics for taxonomic purposes (Chun et al., 2018)

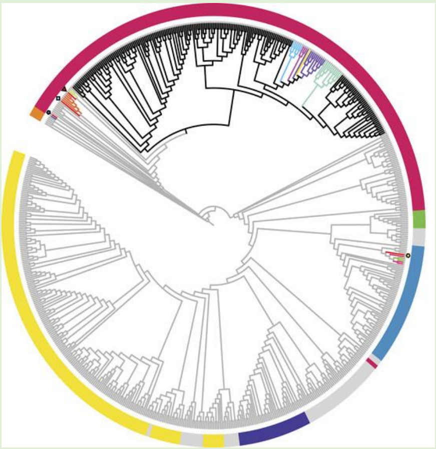
1. ***OGRI***

- any measurements indicating how similar two genome sequences are
- direct descendant of DDH (still gold standard)
- taxonomic resolution limited to differentiate only closely related species
- not suitable for phylogenetic inference, especially at the suprageneric rank level
- average nucleotide identity (ANI) most widely used
- an alternative to ANI is digital DDH (Genome-to-Genome Distance Calculator; GGDC)
- authors who propose new species should provide OGRI values between the type strain of proposed species and type strains of related species that show ≥98.7 % 16S sequence similarity

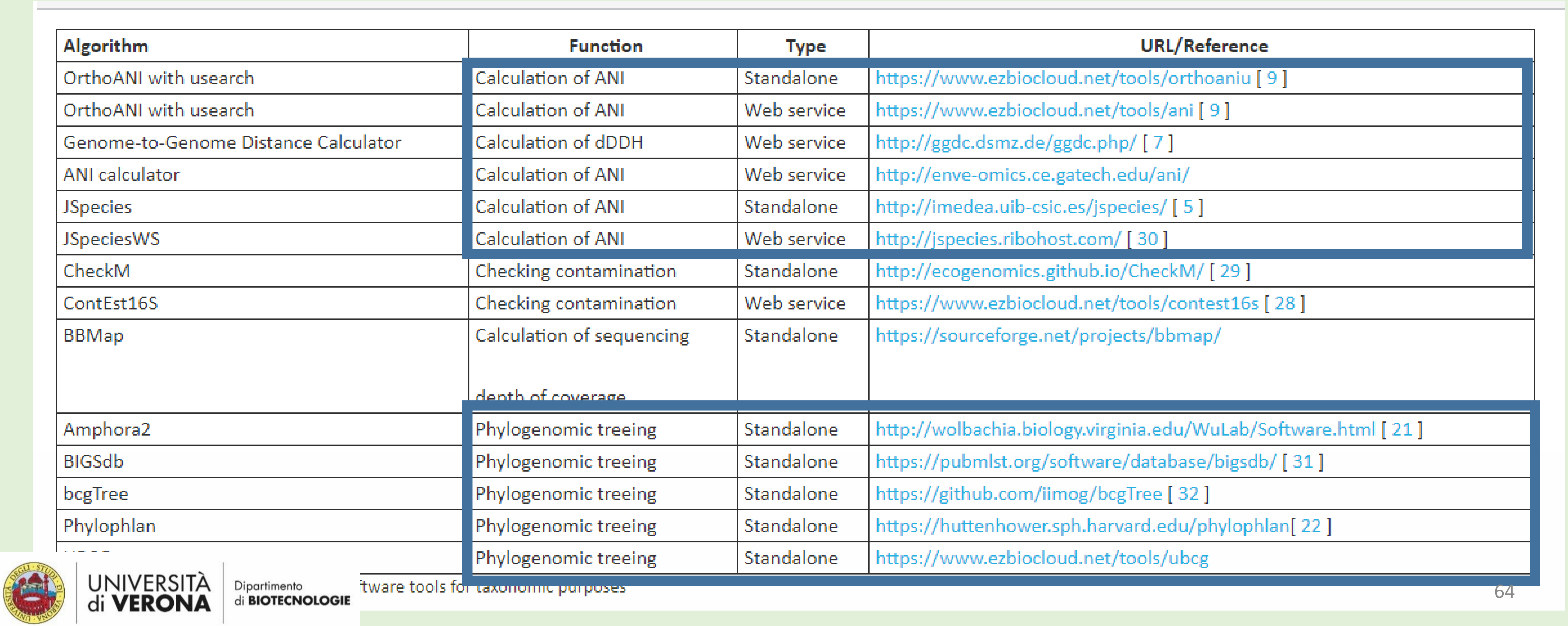
Bioinformatics for taxonomic purposes (Chun et al., 2018)

2. ***Phylogenomic treeing*** (use of genome data to phylogenetic analysis)

- to explore the phylogenetic relationship at various taxonomic levels
- Inference of phylogenetic trees on the basis of multiple genes, instead of a single gene such as 16S
- active area of research with different scientific views
- Recommendation of using at least **30** genes, which is higher than that used in the traditional multilocus sequence analysis (MLSA)



Software tools available (web-services and standalone)



Algorithm	Function	Type	URL/Reference
OrthoANI with usearch	Calculation of ANI	Standalone	https://www.ezbiocloud.net/tools/orthoaniu [9]
OrthoANI with usearch	Calculation of ANI	Web service	https://www.ezbiocloud.net/tools/ani [9]
Genome-to-Genome Distance Calculator	Calculation of dDDH	Web service	http://ggdc.dsmz.de/ggdc.php/ [7]
ANI calculator	Calculation of ANI	Web service	http://enve-omics.ce.gatech.edu/ani/
JSpecies	Calculation of ANI	Standalone	http://imedea.uib-csic.es/jspecies/ [5]
JSpeciesWS	Calculation of ANI	Web service	http://jspecies.ribohost.com/ [30]
CheckM	Checking contamination	Standalone	http://ecogenomics.github.io/CheckM/ [29]
ContEst16S	Checking contamination	Web service	https://www.ezbiocloud.net/tools/contest16s [28]
BBMap	Calculation of sequencing depth of coverage	Standalone	https://sourceforge.net/projects/bbmap/
Amphora2	Phylogenomic treeing	Standalone	http://wolbachia.biology.virginia.edu/WuLab/Software.html [21]
BIGSdb	Phylogenomic treeing	Standalone	https://pubmlst.org/software/database/bigsdb/ [31]
bcgTree	Phylogenomic treeing	Standalone	https://github.com/iimog/bcgTree [32]
Phylophlan	Phylogenomic treeing	Standalone	https://huttenhower.sph.harvard.edu/phylophlan[22]
UBCG	Phylogenomic treeing	Standalone	https://www.ezbiocloud.net/tools/ubcg

ftware tools for taxonomic purposes

Important aspects (Chun et al., 2018)

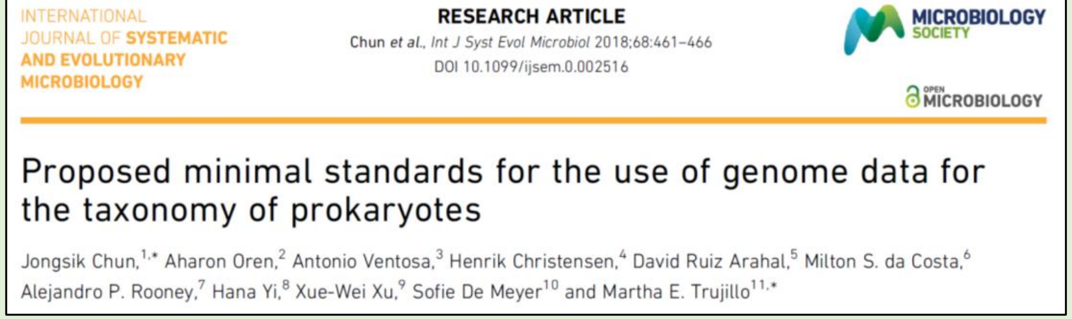
- **Choice of reference genome data from the public domain**
  - multiple genome sequences can be available for the same type strains

    →authentic genome sequences of the best quality are chosen for OGRI and phylogenomic treeing

    →recommended criterion: N50 statistic* rather than the number of contigs

    →sequencing depth of coverage can also be useful, but usually not available

Other relevant elements



INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY

RESEARCH ARTICLE

Chun *et al.*, *Int J Syst Evol Microbiol* 2018;68:461–466

DOI 10.1099/ijsem.0.002516

MICROBIOLOGY SOCIETY

OPEN MICROBIOLOGY

Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes

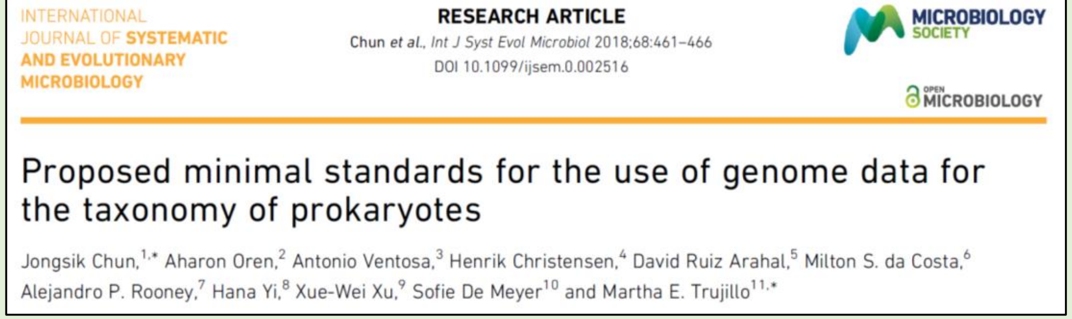
Jongsik Chun,[1,*] Aharon Oren,[2] Antonio Ventosa,[3] Henrik Christensen,[4] David Ruiz Arahal,[5] Milton S. da Costa,[6] Alejandro P. Rooney,[7] Hana Yi,[8] Xue-Wei Xu,[9] Sofie De Meyer[10] and Martha E. Trujillo[11,*]

- **DNA sequencing platforms**
  - Illumina (USA),
  - Ion Torrent (Thermo Fisher Scientific, USA)
  - Pacific Biosciences (USA)

  generate DNA sequence data that meet the general standards, if used with adequate experimental protocols

  “Any other NGS platform that will be available in the future should be subject to rigorous evaluation before it can be used in prokaryotic taxonomic studies”

Other relevant elements



INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY

RESEARCH ARTICLE

Chun *et al.*, *Int J Syst Evol Microbiol* 2018;68:461–466

DOI 10.1099/ijsem.0.002516

MICROBIOLOGY SOCIETY

OPEN MICROBIOLOGY

Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes

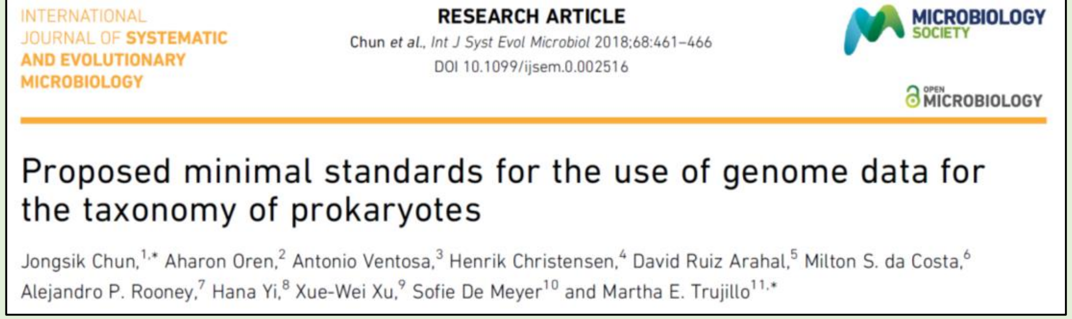
Jongsik Chun,[1,*] Aharon Oren,[2] Antonio Ventosa,[3] Henrik Christensen,[4] David Ruiz Arahal,[5] Milton S. da Costa,[6] Alejandro P. Rooney,[7] Hana Yi,[8] Xue-Wei Xu,[9] Sofie De Meyer[10] and Martha E. Trujillo[11,*]

- **Quality of raw NGS data and assembled genome sequences**
  - the important statistic is the quality of the final assembly, not that of the raw data
  - various software tools can be used to assemble the filtered raw reads into contigs
  - Full genomes are better than contigs, but fragmented assemblies could be sufficient if redundancy is sufficient:
    - *Genome size*. defined as the length sum of all contigs
    - *The number of contigs* and *N50*
    - *Sequencing depth of coverage* ≥50X is recommended (measured for all DNA sequencing platforms with adequate genome assembler software)

N50 statistic

- defines assembly quality
- Given a set of contigs, each with its own length, the *N50* length is defined as the shortest sequence length at 50% of the genome
  - example consider 9 contigs with the lengths 3,5,7,9,11,13,15,17,and 19
    - sum = 99
    - half of the sum = 49,5
    - 50% of this assembly would be 19 + 17 + 15 = 51 (about half the length of the sequence)
    - N50= 15 –> size of the contig which, along with the larger contigs, contain half of sequence of a particular genome
- *L50* count: smallest number of contigs whose length sum produces N50 (L50=3)

Other relevant elements



INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY

RESEARCH ARTICLE

Chun *et al.*, *Int J Syst Evol Microbiol* 2018;68:461–466
DOI 10.1099/ijsem.0.002516

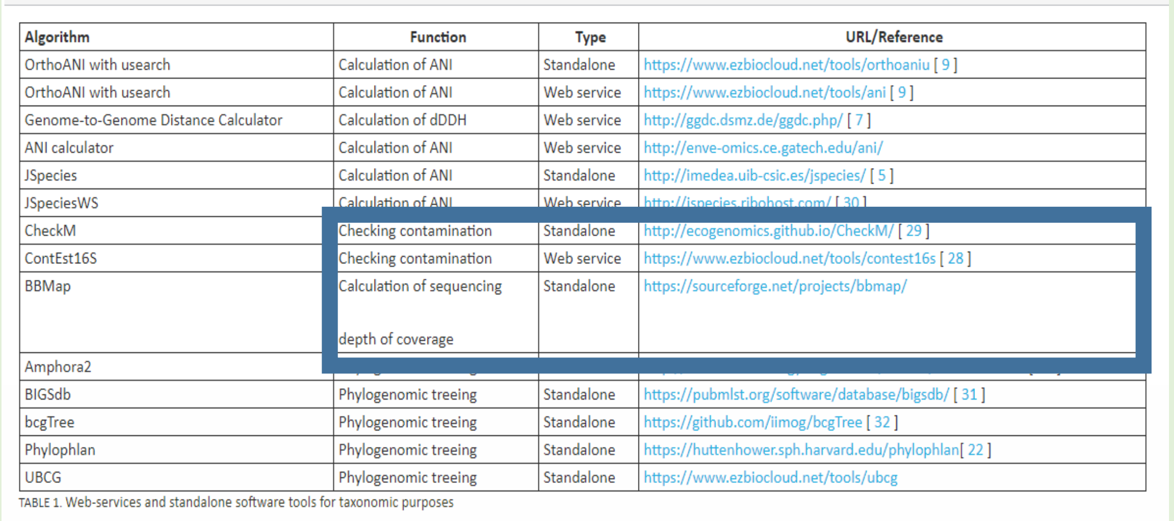
MICROBIOLOGY SOCIETY

OPEN MICROBIOLOGY

Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes

Jongsik Chun,[1,*] Aharon Oren,[2] Antonio Ventosa,[3] Henrik Christensen,[4] David Ruiz Arahal,[5] Milton S. da Costa,[6] Alejandro P. Rooney,[7] Hana Yi,[8] Xue-Wei Xu,[9] Sofie De Meyer[10] and Martha E. Trujillo[11,*]

- **Contamination in the genome assembly**
  - contaminating DNA sequences, even in a minor amount, can be incorporated into the genome assembly, in both culturing and DNA sequencing steps
  - at present, only a few bioinformatic tools for detecting potential contaminations are available using 16S and protein-coding genes
  - Be careful: HGT could be confusing



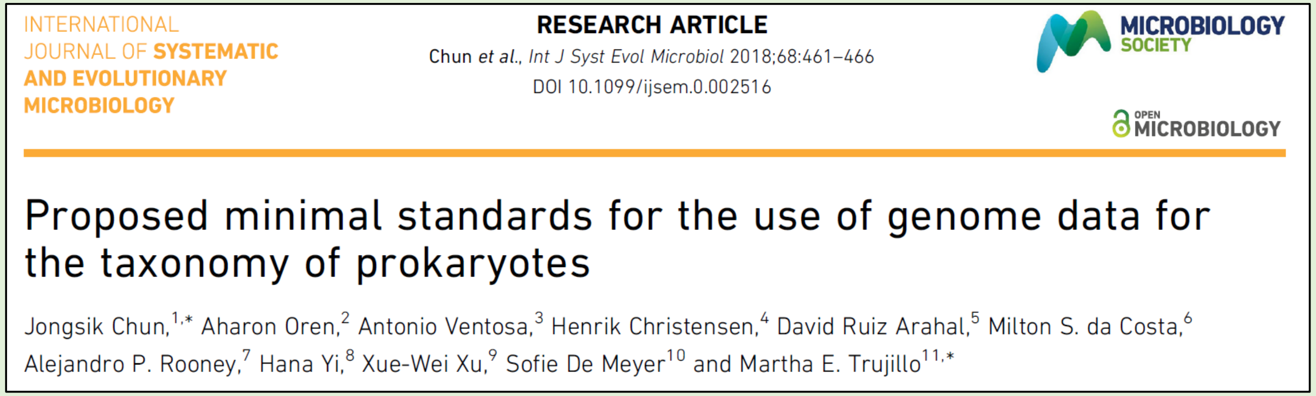
Algorithm	Function	Type	URL/Reference
OrthoANI with usearch	Calculation of ANI	Standalone	https://www.ezbiocloud.net/tools/orthoaniu [9]
OrthoANI with usearch	Calculation of ANI	Web service	https://www.ezbiocloud.net/tools/ani [9]
Genome-to-Genome Distance Calculator	Calculation of dDDH	Web service	http://ggdc.dsmz.de/ggdc.php/ [7]
ANI calculator	Calculation of ANI	Web service	http://enve-omics.ce.gatech.edu/ani/
JSpecies	Calculation of ANI	Standalone	http://imedea.uib-csic.es/jspecies/ [5]
JSpeciesWS	Calculation of ANI	Web service	http://jspecies.ribohost.com/ [30]
CheckM	Checking contamination	Standalone	http://ecogenomics.github.io/CheckM/ [29]
ContEst16S	Checking contamination	Web service	https://www.ezbiocloud.net/tools/contest16s [28]
BBMap	Calculation of sequencing depth of coverage	Standalone	https://sourceforge.net/projects/bbmap/
Amphora2	[illegible]	[illegible]	[illegible]
BIGSdb	Phylogenomic treeing	Standalone	https://pubmlst.org/software/database/bigsdb/ [31]
bcgTree	Phylogenomic treeing	Standalone	https://github.com/iimog/bcgTree [32]
Phylophlan	Phylogenomic treeing	Standalone	https://huttenhower.sph.harvard.edu/phylophlan[22]
UBCG	Phylogenomic treeing	Standalone	https://www.ezbiocloud.net/tools/ubcg

TABLE 1. Web-services and standalone software tools for taxonomic purposes

Classification of genera and higher taxa

- ❖ OGRI: no taxonomic resolution above the species level
- ❖ **multigene-based phylogenomic treeing approach** for defining genera or higher taxa
- “The combination of phylogenomic treeing and highly conserved phenotypes, including chemotaxonomic markers, should play a significant role in the classification of genera and higher taxa”

- We’ll have a look at phylogeny afterwards



INTERNATIONAL JOURNAL OF **SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY**

RESEARCH ARTICLE

Chun *et al.*, *Int J Syst Evol Microbiol* 2018;68:461–466

DOI 10.1099/ijsem.0.002516

MICROBIOLOGY SOCIETY

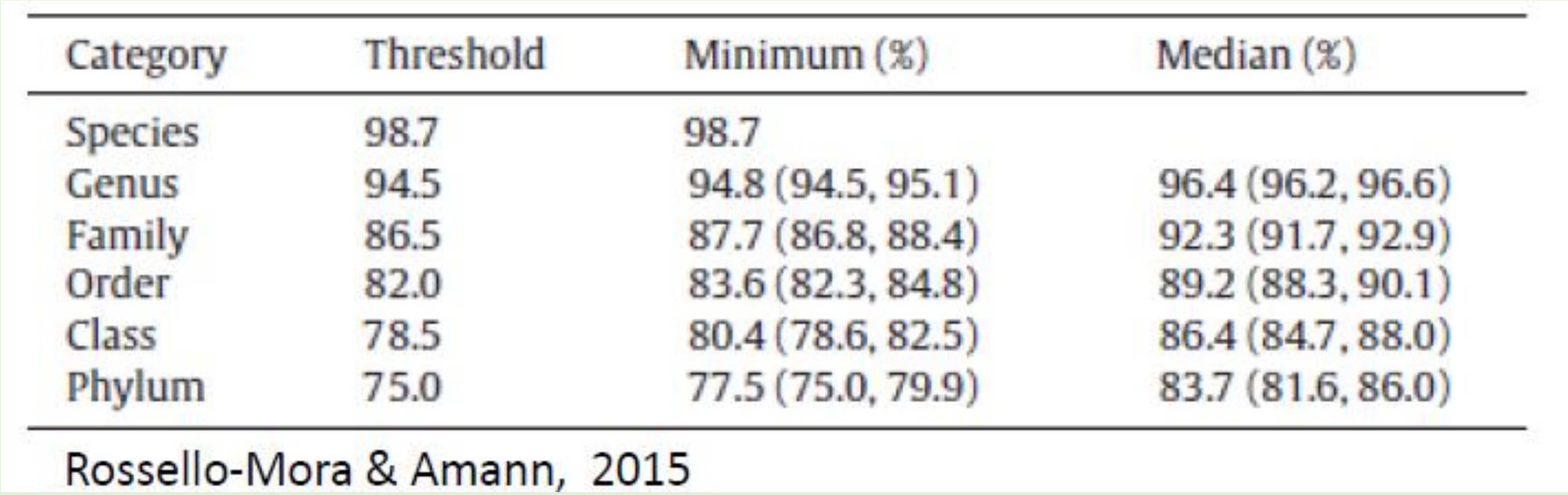
OPEN MICROBIOLOGY

Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes

Jongsik Chun,[1,*] Aharon Oren,[2] Antonio Ventosa,[3] Henrik Christensen,[4] David Ruiz Arahal,[5] Milton S. da Costa,[6] Alejandro P. Rooney,[7] Hana Yi,[8] Xue-Wei Xu,[9] Sofie De Meyer[10] and Martha E. Trujillo[11,*]

Classification of genera and higher taxa

- 16S similarity



Category	Threshold	Minimum (%)	Median (%)
Species	98.7	98.7	
Genus	94.5	94.8 (94.5, 95.1)	96.4 (96.2, 96.6)
Family	86.5	87.7 (86.8, 88.4)	92.3 (91.7, 92.9)
Order	82.0	83.6 (82.3, 84.8)	89.2 (88.3, 90.1)
Class	78.5	80.4 (78.6, 82.5)	86.4 (84.7, 88.0)
Phylum	75.0	77.5 (75.0, 79.9)	83.7 (81.6, 86.0)

Rossello-Mora & Amann, 2015

Infra-specific ranks

Taxonomic rank/Suffix		Example
• **Phylum**		
• **Class**		
• **Order**	-ales	Pseudomonadales
Suborder	-ineae	Pseudomonadineae
Family	-aceae	Pseudomonadaceae
Subfamily	-oideae	Pseudomonadoideae
Tribe	-eae	Pseudomonadeae
Subtribe	-inae	Pseudomonadinae
Genus	—	*Pseudomonas*
(Subgenus)	—	(not for *Pseudomonas)*
Species	—	*Pseudomonas fluorescens*
Subspecies	—	*Pseudomonas pseudoalcaligenes* subsp. *citrulli*
Biovar	—	*Pseudomonas fluorescens* biovar I
Pathovar	—	*Pseudomonas syringae* pathovar *tabaci*

Bergey' s Manual of Systematic Bacteriology

Taxonomic outlines are available online

Genome data in subspecies recognition

- No general guideline at the moment
- a good practice should include that (among others)
  i) OGRIs between subspecies and other species should be lower than the species-level cutoff value
  ii) OGRIs between subspecies should be higher than the species-level cutoff,
  iii) strains belonging to different subspecies should be genomically coherent and form distinguishable clades by OGRIs and phylogenomic treeing

INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY

RESEARCH ARTICLE

Chun *et al.*, *Int J Syst Evol Microbiol* 2018;68:461–466

DOI 10.1099/ijsem.0.002516

MICROBIOLOGY SOCIETY

OPEN MICROBIOLOGY

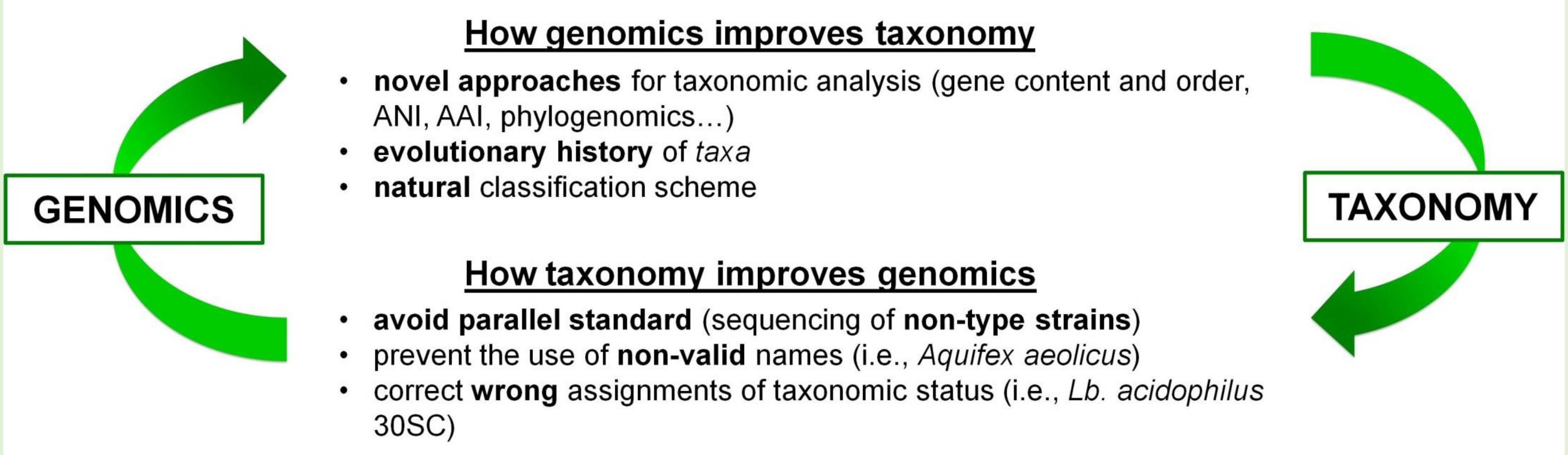
Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes

Jongsik Chun,[1,*] Aharon Oren,[2] Antonio Ventosa,[3] Henrik Christensen,[4] David Ruiz Arahal,[5] Milton S. da Costa,[6] Alejandro P. Rooney,[7] Hana Yi,[8] Xue-Wei Xu,[9] Sofie De Meyer[10] and Martha E. Trujillo[11,*]

Useful resources

- Classification
  - Bergey’s Manual of Systematic Bacteriology, now 2nd ed. (2001), reference book for classification
  - IJSEM
  - International Committee on Systematics of Prokaryotes (ICSP) (www.the-icsp.org) and subcommittees
- Nomenclature
  - Prokariotic Code available online
  - IJSEM
  - SAM and Ant van Leew
  - Approved Lists of Bacterial Names (Int. J. Syst. Bacteriol, 1980,30:225-420) also available in http://www.bacterio.cict.fr/
  - Validation Lists, published in the International Journal of Systematic and Evolutionary Microbiology (or International Journal of Systematic Bacteriology, prior to 2000), available online at www.bacterio.cict.fr
- Culture collections
  - e.g. ATCC, LMG, DSMZ, JCM

Genome-based taxonomy & taxonomy-based genomics



How genomics improves taxonomy

- **novel approaches** for taxonomic analysis (gene content and order, ANI, AAI, phylogenomics…)
- **evolutionary history** of *taxa*
- **natural** classification scheme

How taxonomy improves genomics

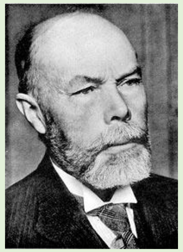
- **avoid parallel standard** (sequencing of **non-type strains**)
- prevent the use of **non-valid** names (i.e., *Aquifex aeolicus*)
- correct **wrong** assignments of taxonomic status (i.e., *Lb. acidophilus* 30SC)

Examples in the genus *Lactobacillus*

The genus level

Lactobacillus

First description by **Beijerinck** in **1901,** Type species: *L. delbrueckii*



1909 “The Lactic Acid Bacteria**”** by **Orla Jensen**

- **184** species, **220** validly published names since 1980

QPS List (EFSA)	GRAS notice (FDA)	EFFCA Inventory	Patents (ESPACENET)
36 species	12 species	86 species	22 species

Salvetti & O’Toole 2017

the beginning

Beijerinck, M.W. 1901. Archives Neerlandaises des Sciences Exactes et Naturelles (Section 2) **6:**212–243.

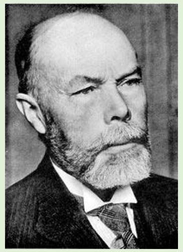
Thermobacterium, ***Streptobacterium*** and *Streptococcus*: mainly lactic acid besides traces of other by-products

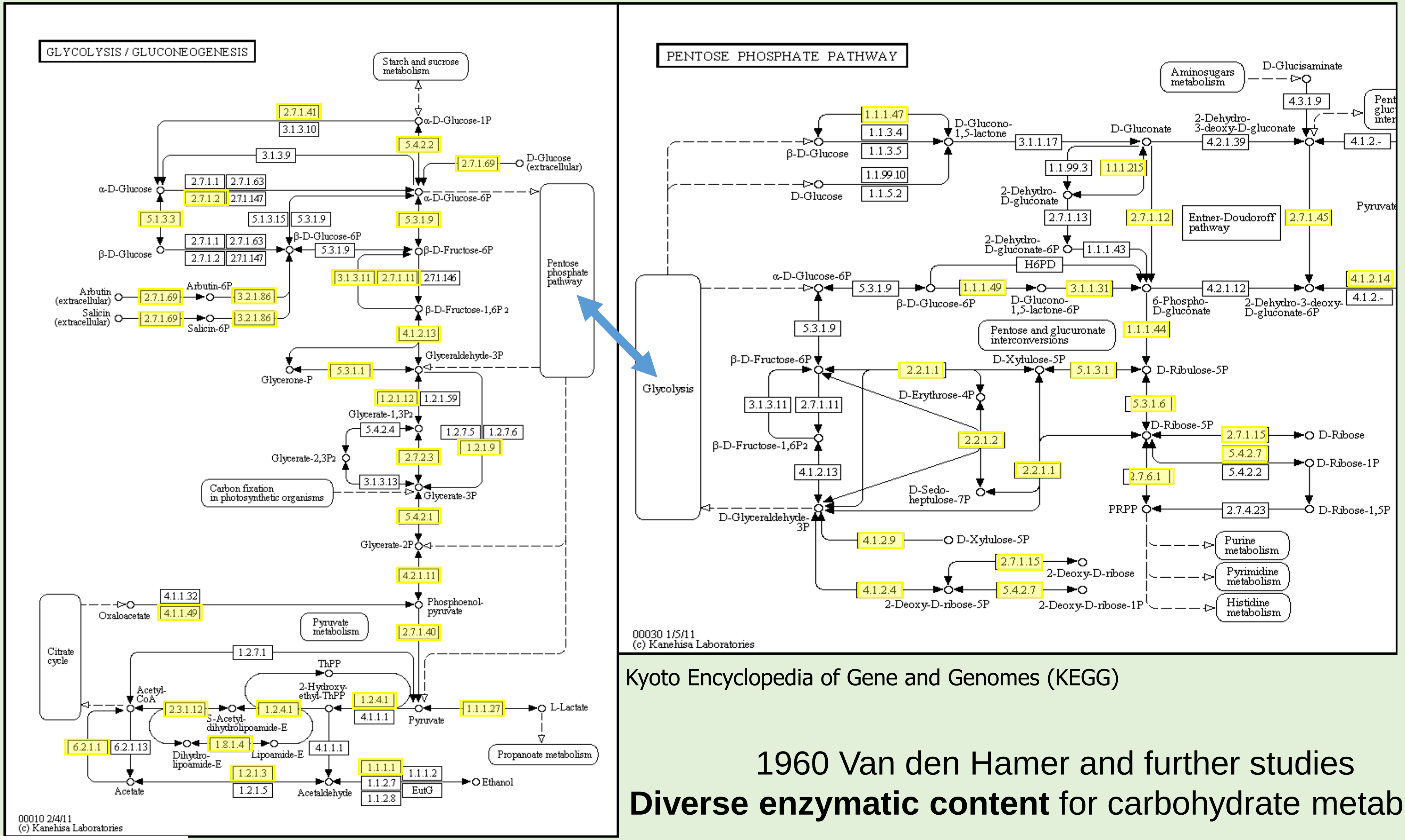
Betabacterium and *Betacoccus:* detectable amounts of gas and other by-products

Three subgenera of *Lactobacillus*

1919 Orla Jensen:

morphology, nutritional characteristics, temperature range for growth and agglutination effects

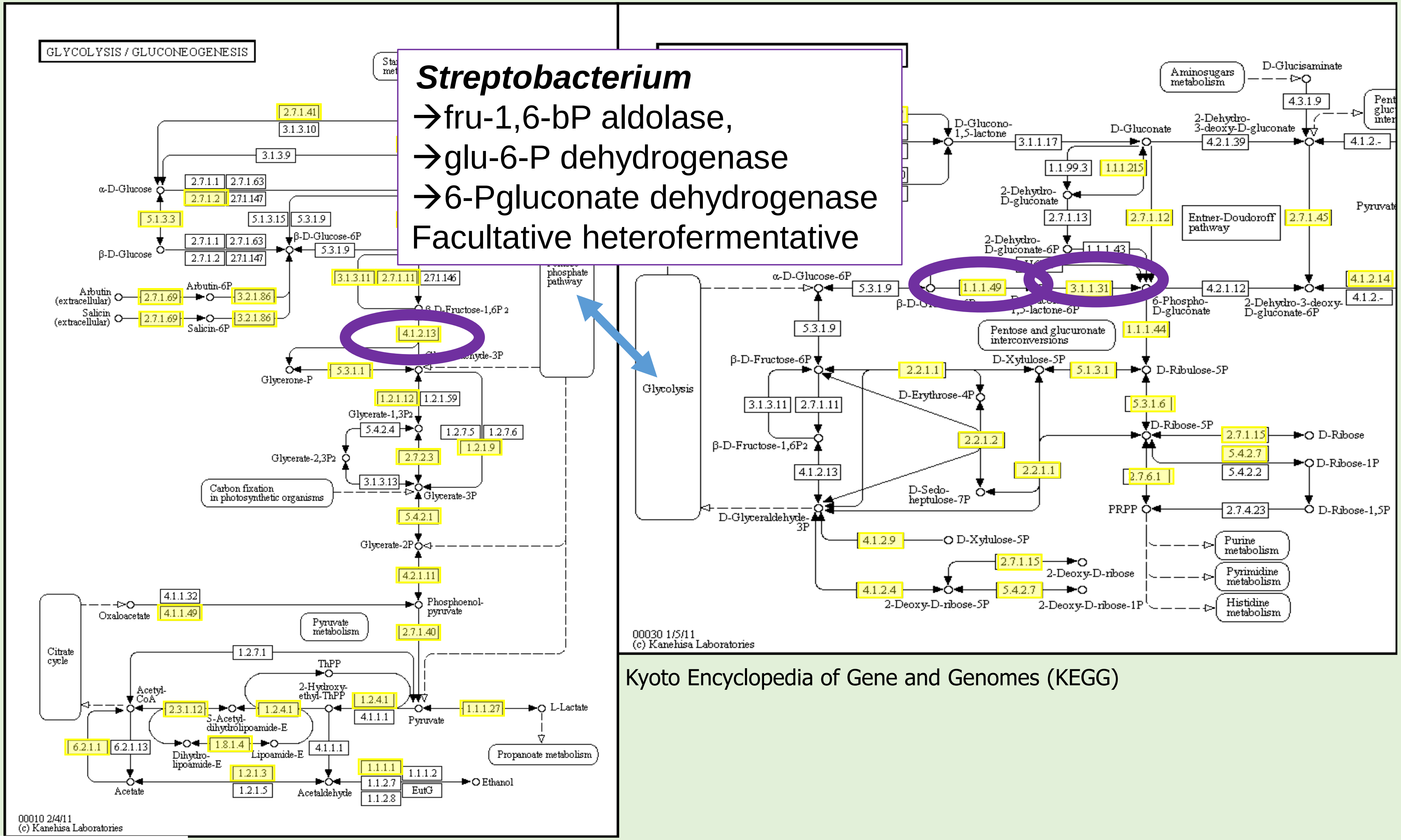




Kyoto Encyclopedia of Gene and Genomes (KEGG)

1960 Van den Hamer and further studies

Diverse enzymatic content for carbohydrate metabolism



Streptobacterium

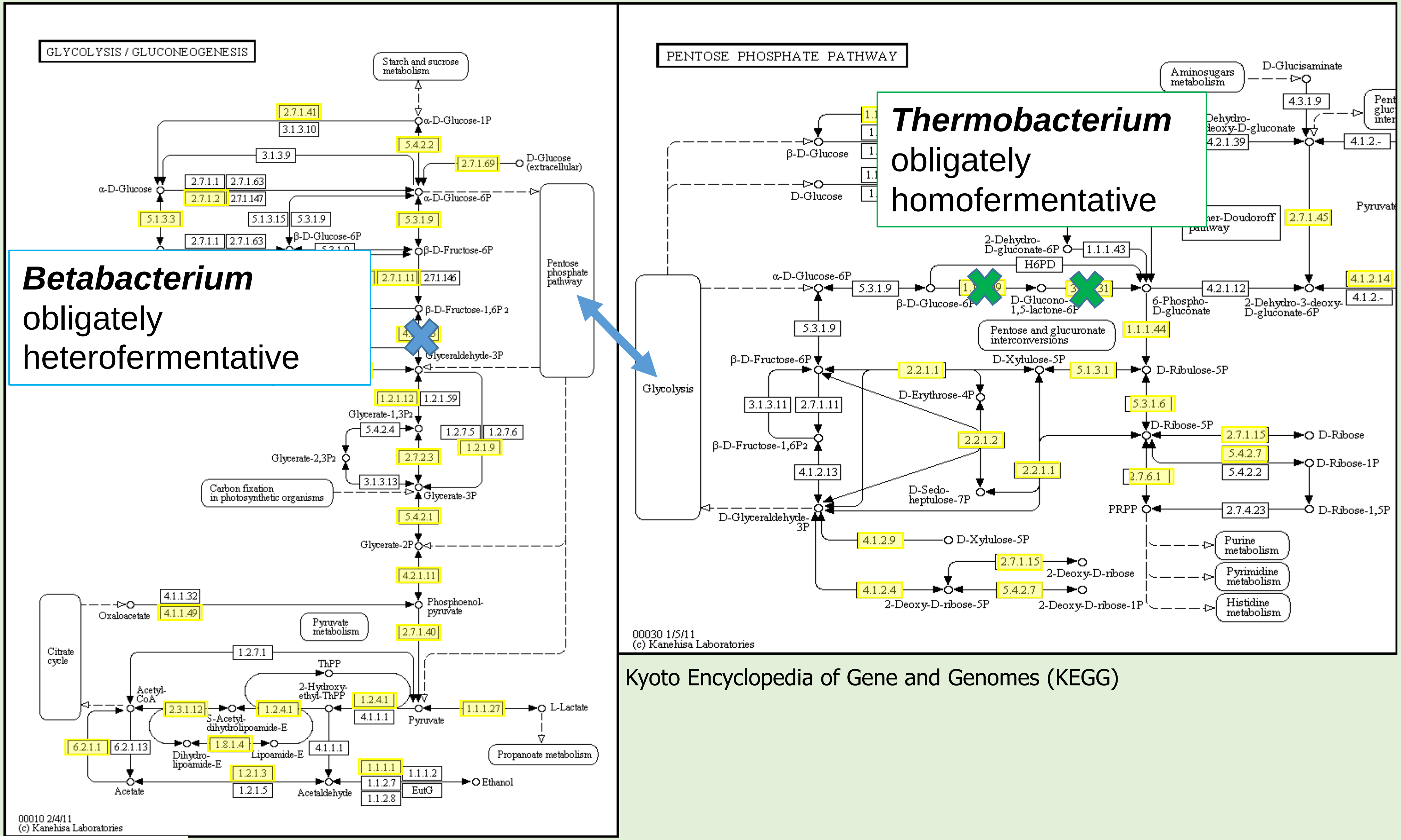
→fru-1,6-bP aldolase,

→glu-6-P dehydrogenase

→6-Pgluconate dehydrogenase

Facultative heterofermentative

Kyoto Encyclopedia of Gene and Genomes (KEGG)



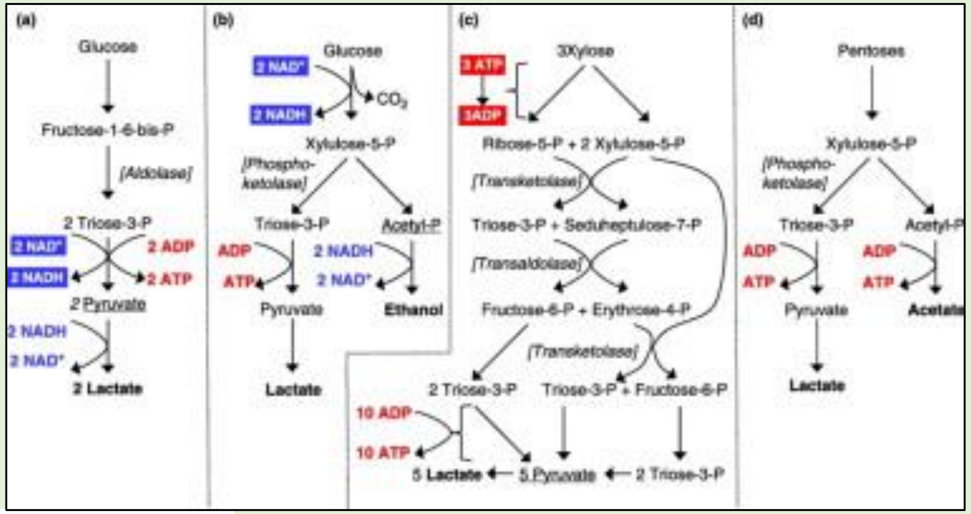
Betabacterium
obligately
heterofermentative

Thermobacterium
obligately
homofermentative

Kyoto Encyclopedia of Gene and Genomes (KEGG)

Metabolic characteristics of LAB

- Oxygen tolerant, growth 2-53°C
- Capacity for respiration, fermentative metabolism
- Multiple auxotrophies for aminoacids, nucleotides and vitamins (nutrient-rich environment)

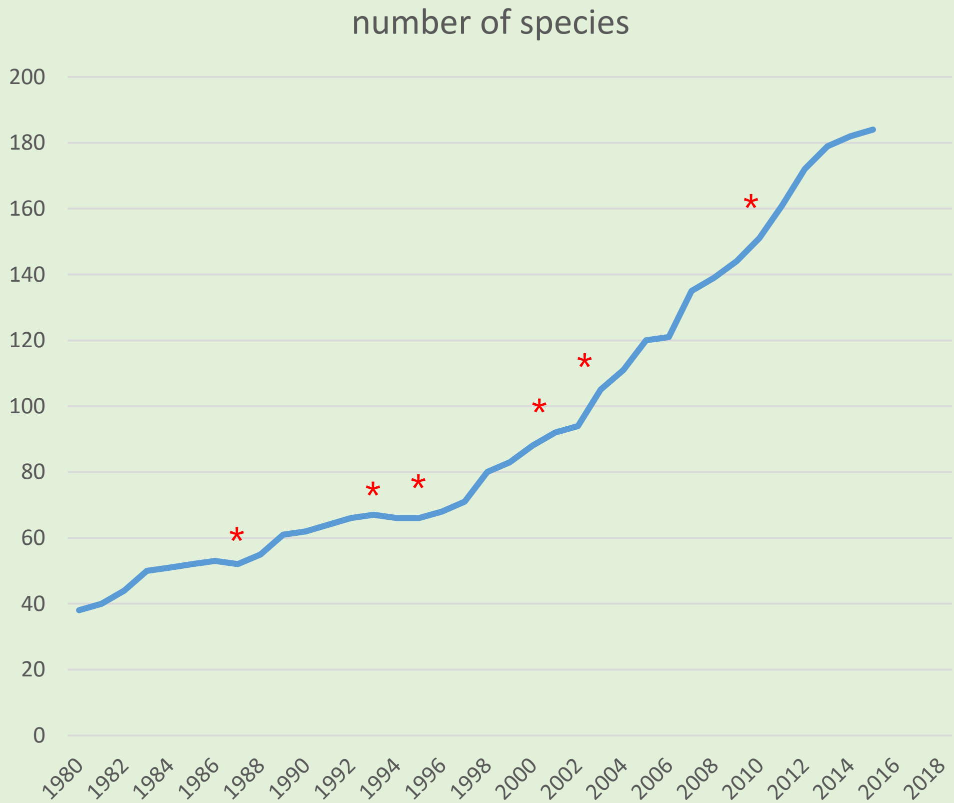


Two major metabolic groups:

1. Homofermentative:
Hexoses via EMP pathway
2. Heterofermentative:
Hexoses via phosphoketolase pathway
Pentoses and hexoses utilised simultaneously

(Gänzle 2015, Duar et al., 2017)

Never-ending species description



- 1980 Approved List of bacterial names **35** valid **species** of *Lactobacillus*

* 1987 *Carnobacterium*

* 1993 *Atopobium*

* 1994 *Weissella*

* 2001 *Olsenella*

* 2002 *Leuconostoc*

* 2011 *Eggerthia* and *Kandleria*

* 2000-2011 *Paralactobacillus*

UNIVERSITÀ di VERONA Dipartimento di BIOTECNOLOGIE

Phylogenetic framework at *order* level

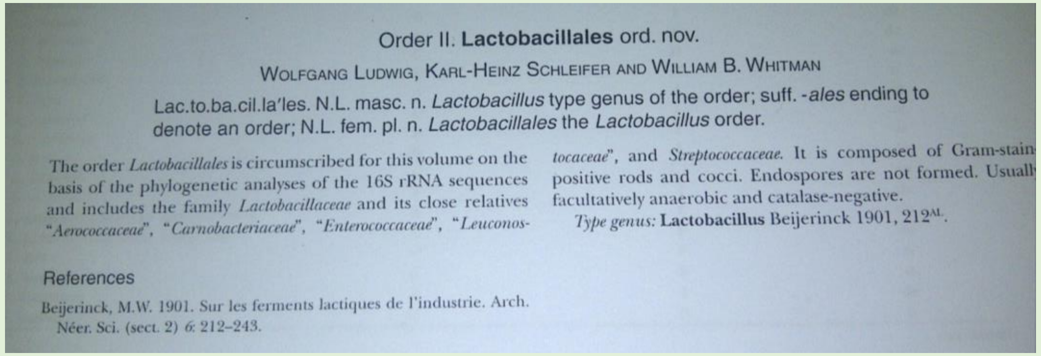
- ***Lactobacillus*** (‘*Paralactobacillus*’)
- *Pediococcus*
- *Enterococcus*
- *Leuconostoc*
- *Oenococcus*
- *Lactococcus*
- *Streptococcus*

Family

Lactobacillaceae

Order *Lactobacillales*

Main genera of



Order II. **Lactobacillales** ord. nov.

Wolfgang Ludwig, Karl-Heinz Schleifer and William B. Whitman

Lac.to.ba.cil.la'les. N.L. masc. n. *Lactobacillus* type genus of the order; suff. *-ales* ending to denote an order; N.L. fem. pl. n. *Lactobacillales* the *Lactobacillus* order.

The order *Lactobacillales* is circumscribed for this volume on the basis of the phylogenetic analyses of the 16S rRNA sequences and includes the family *Lactobacillaceae* and its close relatives “*Aerococcaceae*”, “*Carnobacteriaceae*”, “*Enterococcaceae*”, “*Leuconostocaceae*”, and *Streptococcaceae*. It is composed of Gram-stain positive rods and cocci. Endospores are not formed. Usuall facultatively anaerobic and catalase-negative.

Type genus: **Lactobacillus** Beijerinck 1901, 212^{AL}.

References

Beijerinck, M.W. 1901. Sur les ferments lactiques de l'industrie. Arch. Néer. Sci. (sect. 2) *6*: 212–243.

Bergey’s Manual of Systematic Bacteriology

Domain, Phylum, Class, Order, Family, Genus, Species

Curr. Issues Intestinal Microbiol. 8: 44–61. Onlir

Taxonomy of Lactobacilli and Bifidobacteria

Giovanna E. Felis and Franco Dellaglio*†

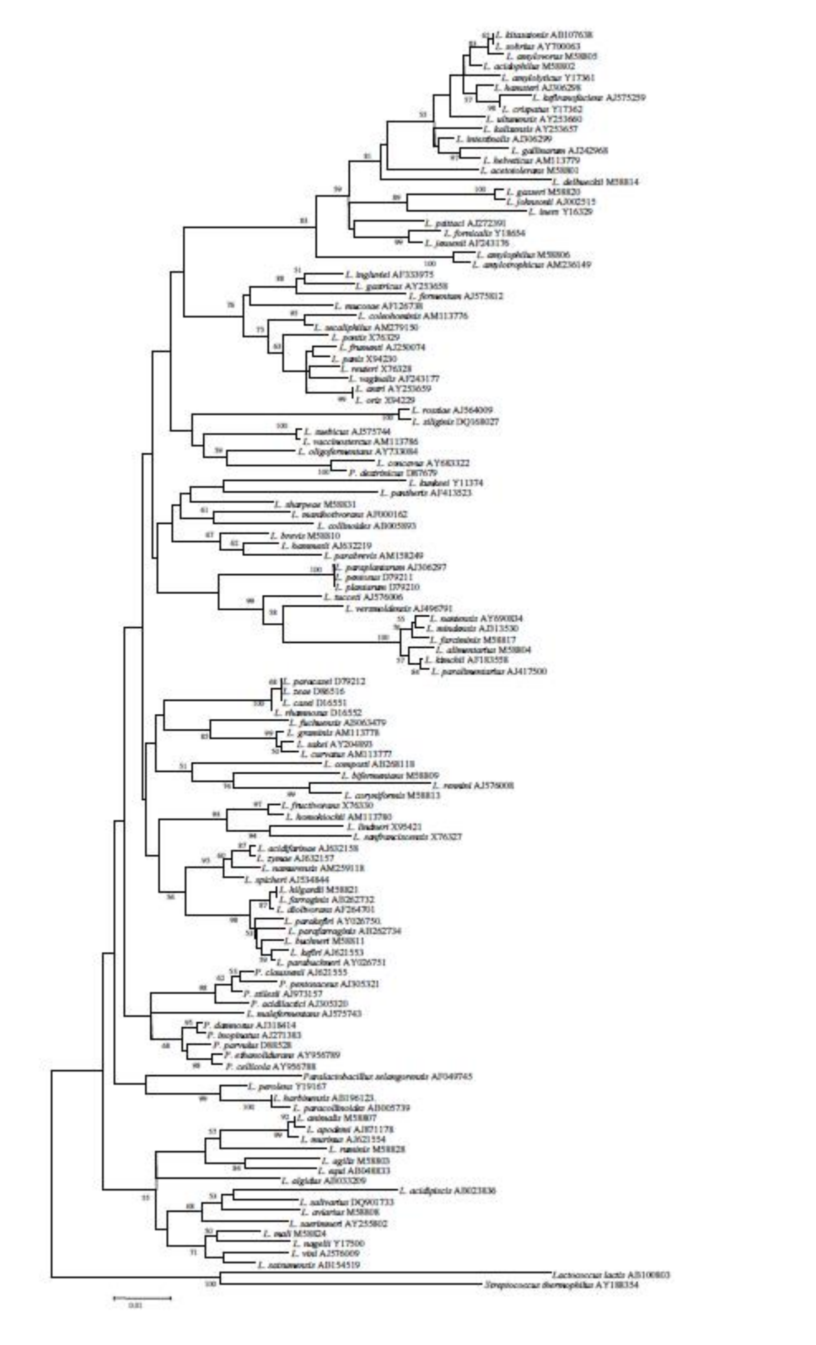
2007

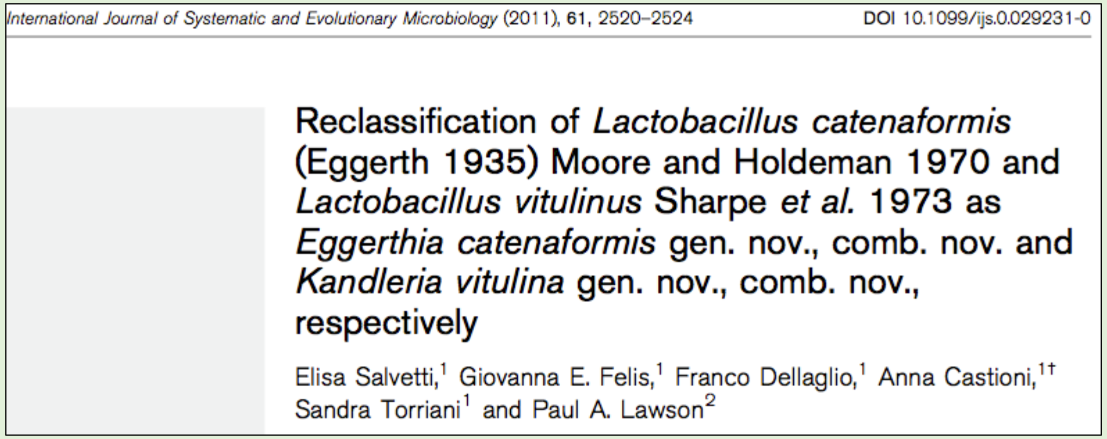
108 species

16S rRNA gene sequence analysis

- 13 groups (≥3 species)
- 3 couples
- 5 single lines of descent

intermixed with *Pediococcus* (**1** group)





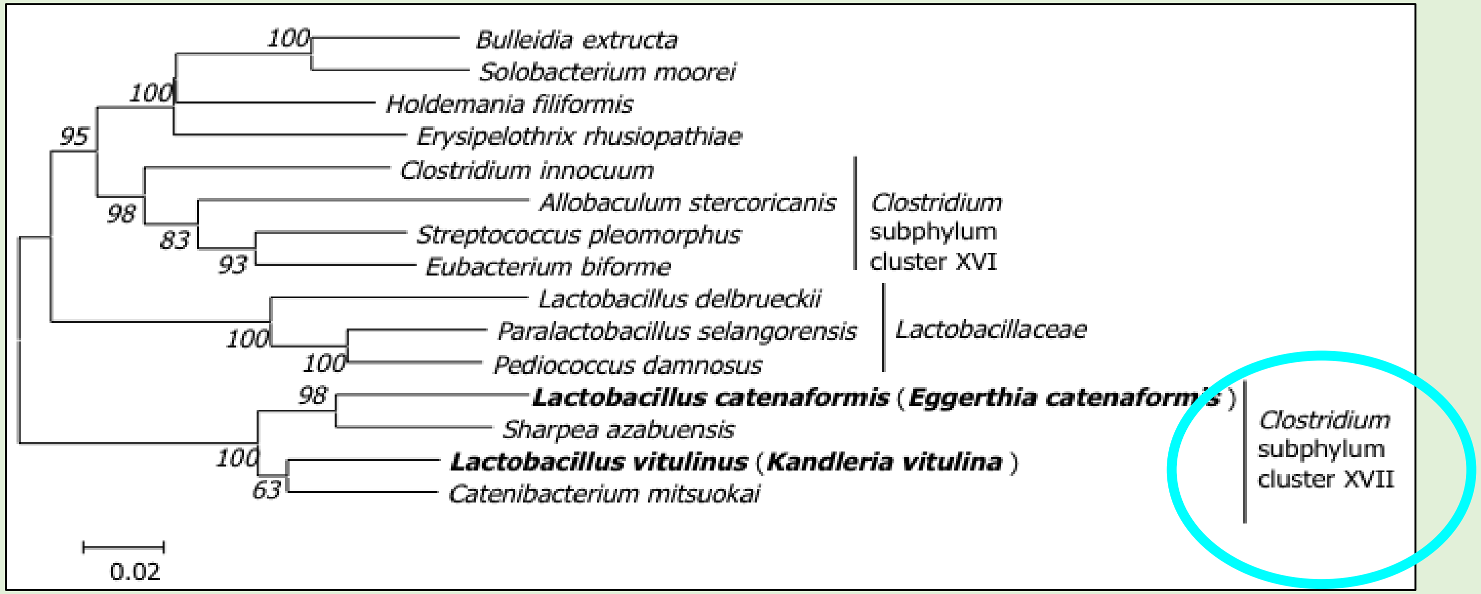
International Journal of Systematic and Evolutionary Microbiology (2011), 61, 2520–2524 DOI 10.1099/ijs.0.029231-0

Reclassification of *Lactobacillus catenaformis* (Eggerth 1935) Moore and Holdeman 1970 and *Lactobacillus vitulinus* Sharpe *et al.* 1973 as *Eggerthia catenaformis* gen. nov., comb. nov. and *Kandleria vitulina* gen. nov., comb. nov., respectively

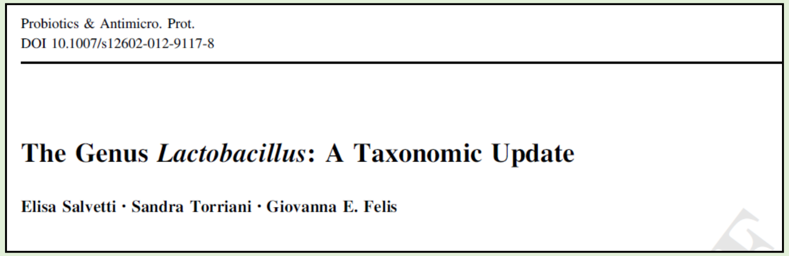
Elisa Salvetti,[1] Giovanna E. Felis,[1] Franco Dellaglio,[1] Anna Castioni,[1†] Sandra Torriani[1] and Paul A. Lawson[2]

- *L. catenaformis* and *L. vitulinus* 16S rRNA gene sequence comparison and phylogenetic analysis

- phenotypic data



Bulleidia extructa
Solobacterium moorei
Holdemania filiformis
Erysipelothrix rhusiopathiae
Clostridium innocuum
Allobaculum stercoricanis
Streptococcus pleomorphus
Eubacterium biforme
Clostridium subphylum cluster XVI
Lactobacillus delbrueckii
Paralactobacillus selangorensis
Pediococcus damnosus
Lactobacillaceae
Lactobacillus catenaformis (Eggerthia catenaformis)
Sharpea azabuensis
Lactobacillus vitulinus (Kandleria vitulina)
Catenibacterium mitsuokai
Clostridium subphylum cluster XVII
100
100
95
98
83
93
100
100
98
100
63
0.02



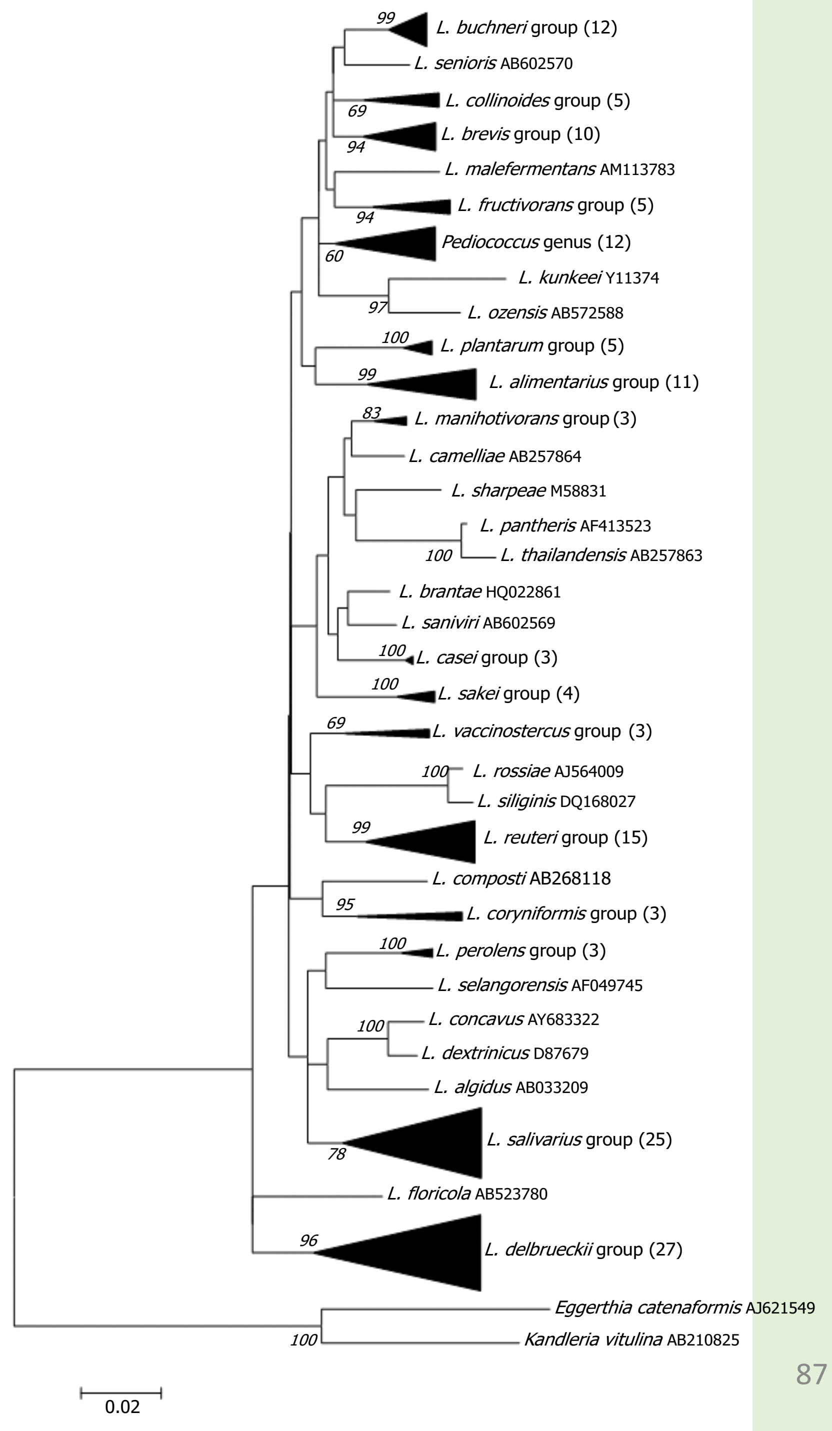
Probiotics & Antimicro. Prot.
DOI 10.1007/s12602-012-9117-8

The Genus *Lactobacillus*: A Taxonomic Update

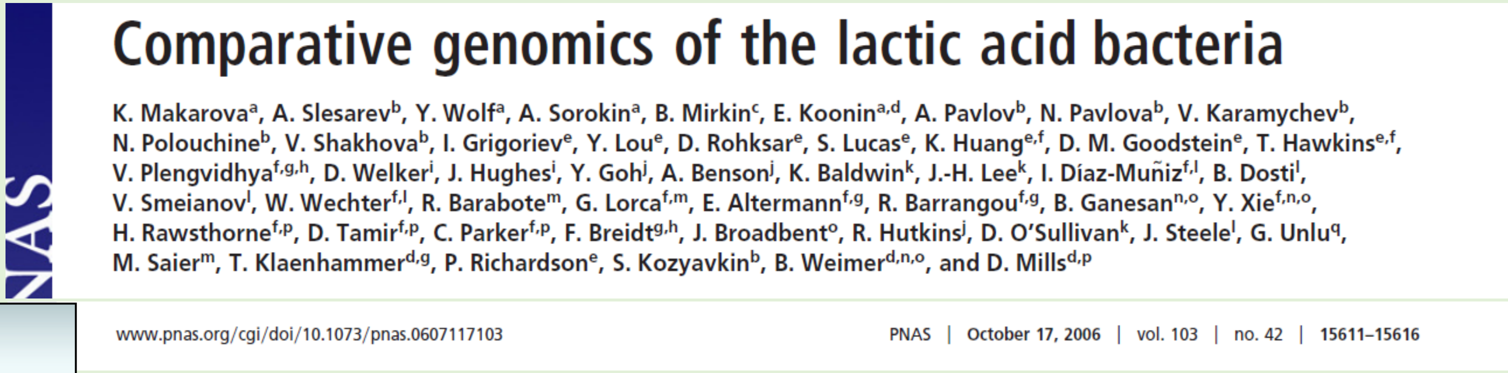
Elisa Salvetti · Sandra Torriani · Giovanna E. Felis

- **2012**
- **152** validly described species
- **16S rRNA gene** sequence analysis

- **14** groups (≥3 species)
- **4** couples
- **10** single lines of descent
- intermixed with *Pediococcus* (**1** group)



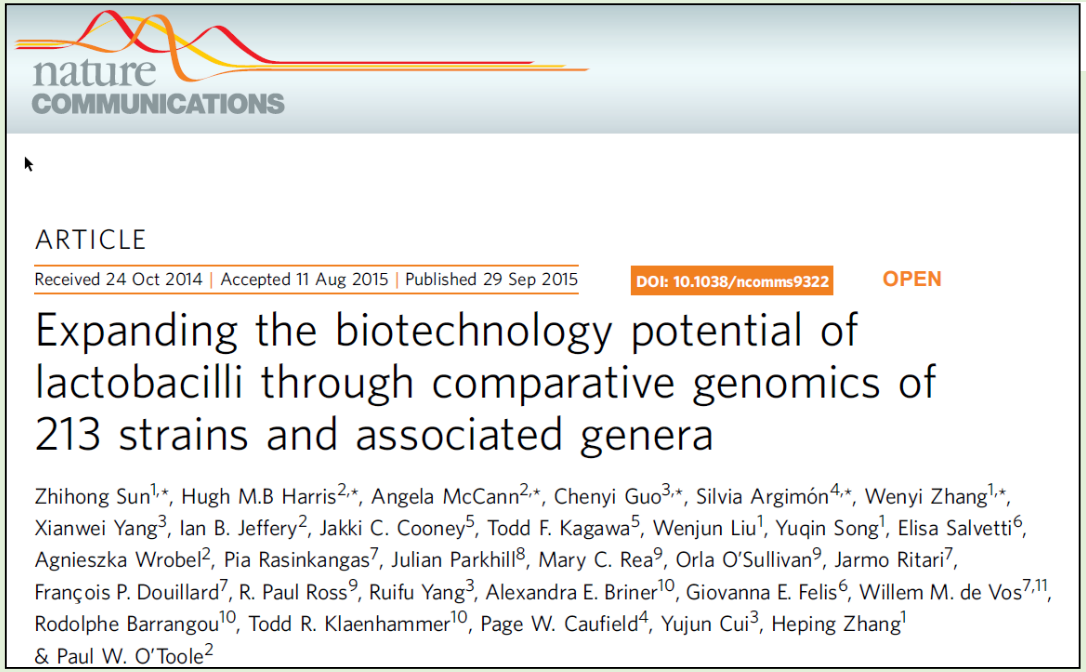
Genome data



Comparative genomics of the lactic acid bacteria

K. Makarova[a], A. Slesarev[b], Y. Wolf[a], A. Sorokin[a], B. Mirkin[c], E. Koonin[a,d], A. Pavlov[b], N. Pavlova[b], V. Karamychev[b], N. Polouchine[b], V. Shakhova[b], I. Grigoriev[e], Y. Lou[e], D. Rohksar[e], S. Lucas[e], K. Huang[e,f], D. M. Goodstein[e], T. Hawkins[e,f], V. Plengvidhya[f,g,h], D. Welker[i], J. Hughes[i], Y. Goh[j], A. Benson[j], K. Baldwin[k], J.-H. Lee[k], I. Díaz-Muñiz[f,l], B. Dosti[l], V. Smeianov[l], W. Wechter[f,l], R. Barabote[m], G. Lorca[f,m], E. Altermann[f,g], R. Barrangou[f,g], B. Ganesan[n,o], Y. Xie[f,n,o], H. Rawsthorne[f,p], D. Tamir[f,p], C. Parker[f,p], F. Breidt[g,h], J. Broadbent[o], R. Hutkins[j], D. O'Sullivan[k], J. Steele[l], G. Unlu[q], M. Saier[m], T. Klaenhammer[d,g], P. Richardson[e], S. Kozyavkin[b], B. Weimer[d,n,o], and D. Mills[d,p]

www.pnas.org/cgi/doi/10.1073/pnas.0607117103 PNAS | October 17, 2006 | vol. 103 | no. 42 | 15611–15616



ARTICLE

Received 24 Oct 2014 | Accepted 11 Aug 2015 | Published 29 Sep 2015 DOI: 10.1038/ncomms9322 OPEN

Expanding the biotechnology potential of lactobacilli through comparative genomics of 213 strains and associated genera

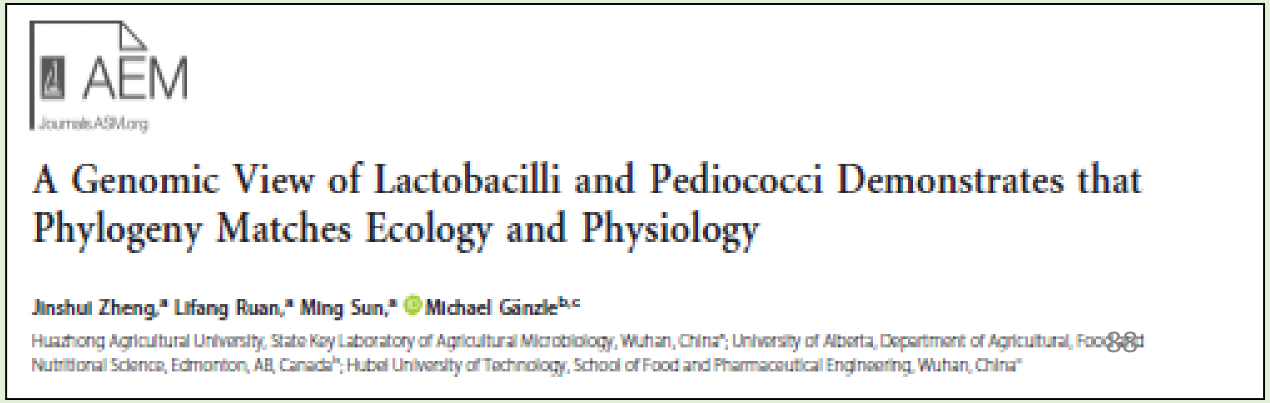
Zhihong Sun[1,*], Hugh M.B Harris[2,*], Angela McCann[2,*], Chenyi Guo[3,*], Silvia Argimón[4,*], Wenyi Zhang[1,*], Xianwei Yang[3], Ian B. Jeffery[2], Jakki C. Cooney[5], Todd F. Kagawa[5], Wenjun Liu[1], Yuqin Song[1], Elisa Salvetti[6], Agnieszka Wrobel[2], Pia Rasinkangas[7], Julian Parkhill[8], Mary C. Rea[9], Orla O'Sullivan[9], Jarmo Ritari[7], François P. Douillard[7], R. Paul Ross[9], Ruifu Yang[3], Alexandra E. Briner[10], Giovanna E. Felis[6], Willem M. de Vos[7,11], Rodolphe Barrangou[10], Todd R. Klaenhammer[10], Page W. Caufield[4], Yujun Cui[3], Heping Zhang[1] & Paul W. O'Toole[2]

- *Lactobacillus*
- *Pediococcus*
- *Leuconostoc*
- *Oenococcus*
- *Weissella*
- *Fructobacillus*

Type strains sequenced: taxonomic value

Heterofermentative species lack Phosphofructokinase gene

24 phylogenetic groups described (Duar *et al*., 2017)



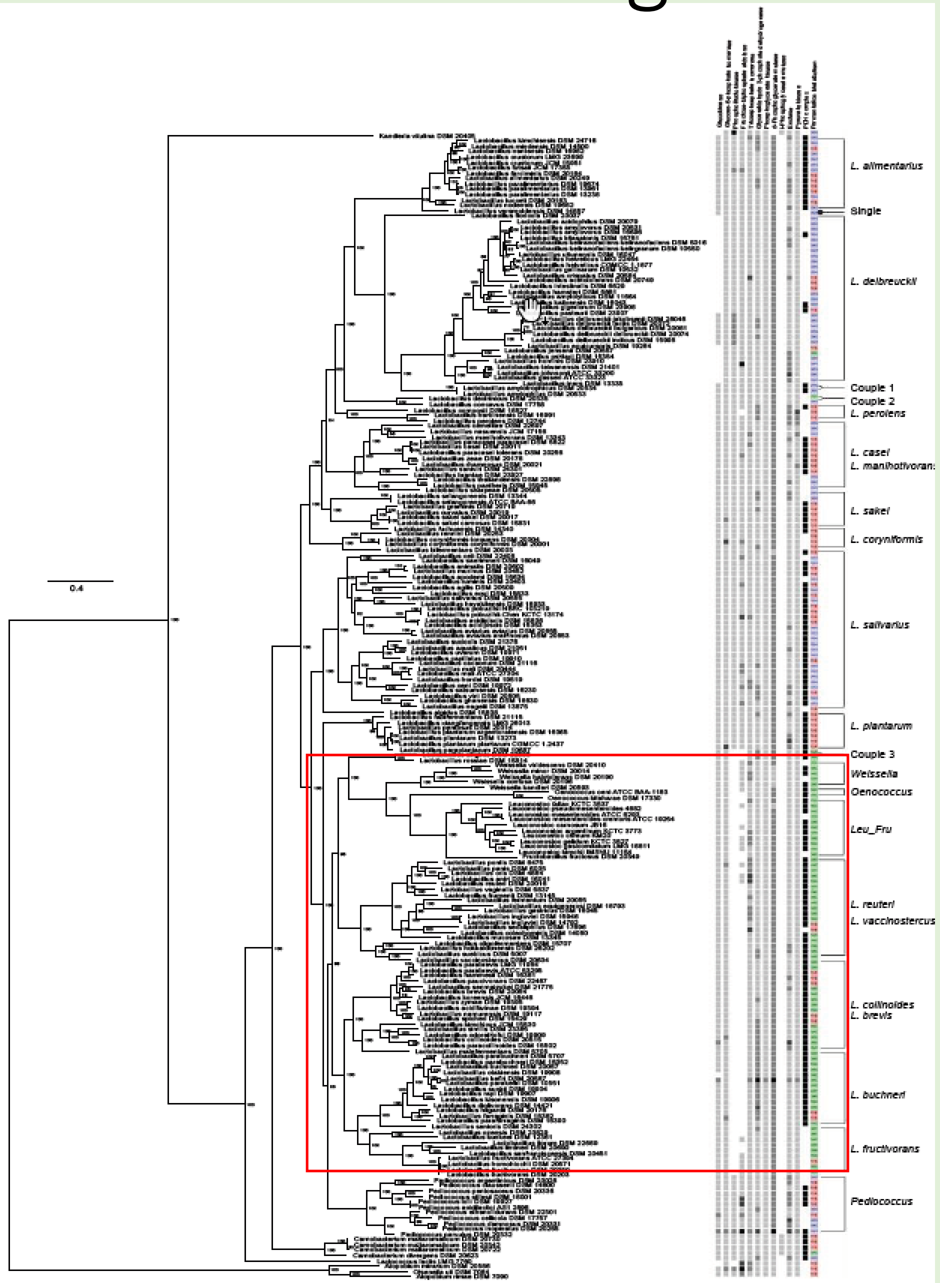
AEM Journals.ASM.org

A Genomic View of Lactobacilli and Pediococci Demonstrates that Phylogeny Matches Ecology and Physiology

Jinshui Zheng,[a] Lifang Ruan,[a] Ming Sun,[a] Michael Gänzle[b,c]

Huazhong Agricultural University, State Key Laboratory of Agricultural Microbiology, Wuhan, China[a]; University of Alberta, Department of Agricultural, Food and Nutritional Science, Edmonton, AB, Canada[b]; Hubei University of Technology, School of Food and Pharmaceutical Engineering, Wuhan, China[c]

Lactobacillus genomics – metabolic potential



- Robust correlation between absence of glycolytic **Phosphofructokinase** and **heterofermentative** species

✓ *L. hilgardii*

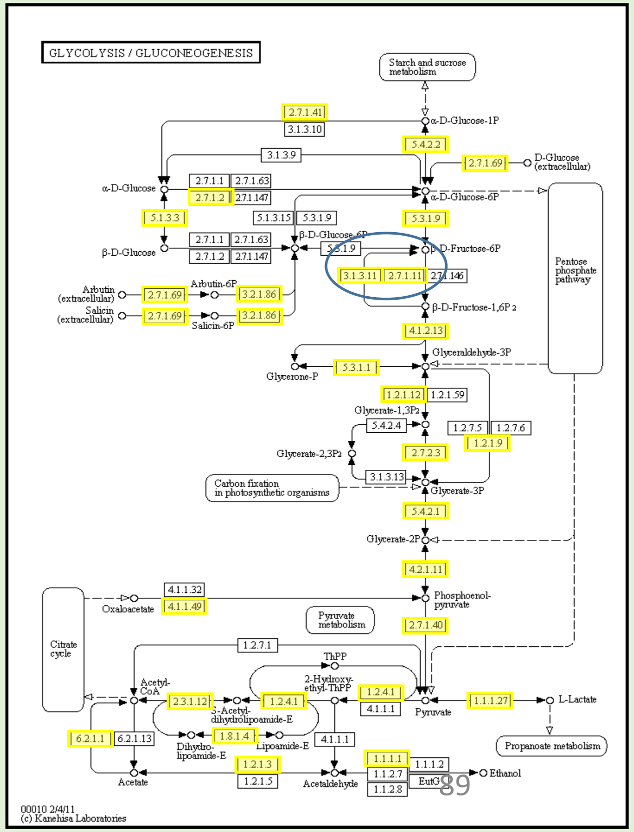
✓ *L. buchneri*

✓ *L. brevis*

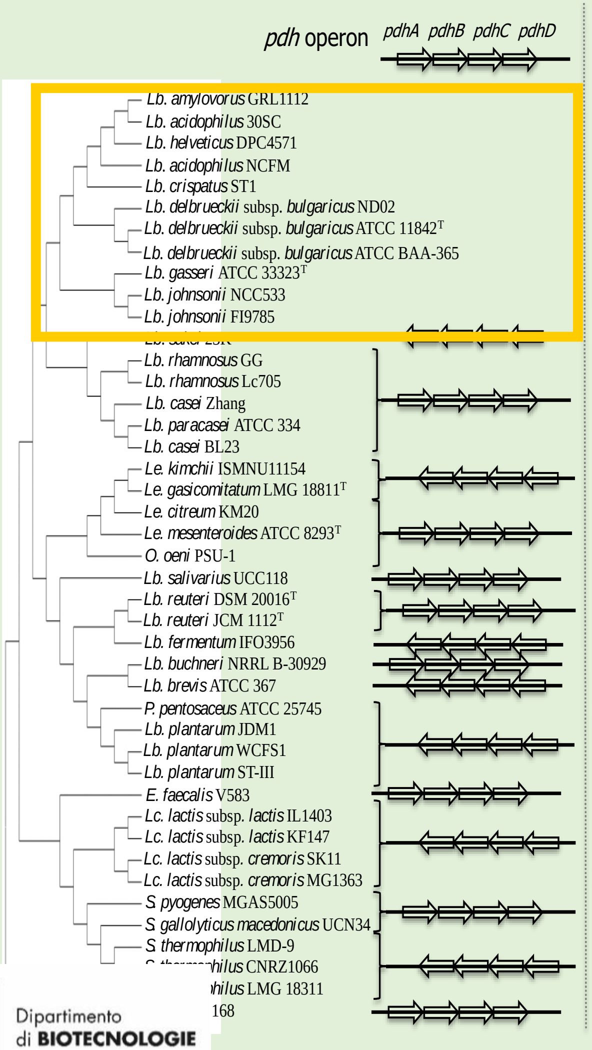
✓ *O. oeni*

✓ *Leuconostoc* spp.

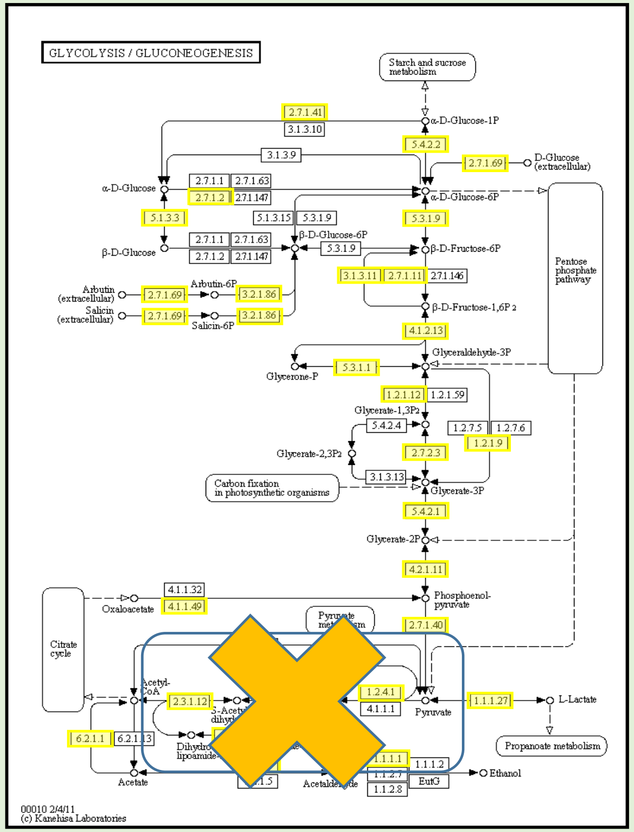
Sun *et al.*, 2015



Pdh operon in *L. delbrueckii group*



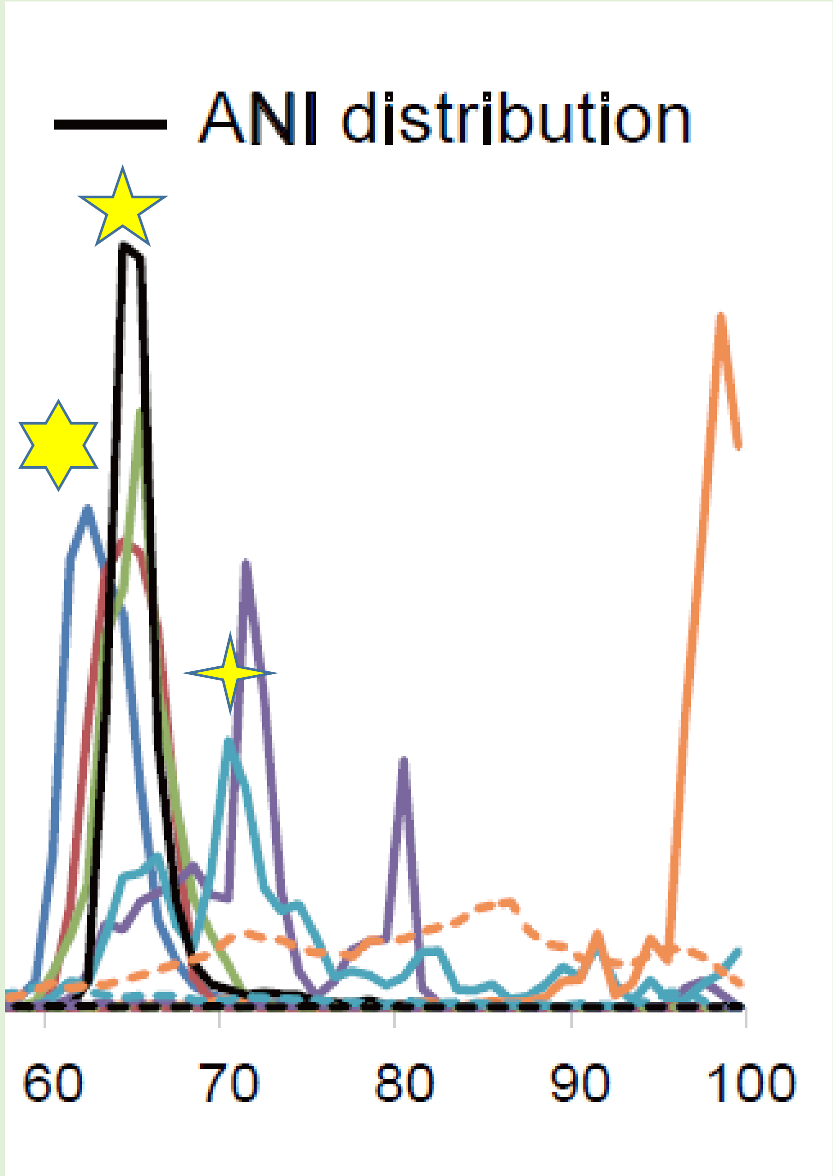
- *L. delbrueckii* group **lack** *pdh* operon
- Some homologs of specific genes are present (HGT)



Salvetti et al., 2013 and unpublished

As diverse as a order or even as a class?

- 🟦 same phylum,diff class
- 🟥 same class,diff order
- 🟩 same order,diff family
- 🟪 same family,diff genus
- 🟦 same genus,diff species
- 🟧 same species
- ⬛ *Lactobacillus*



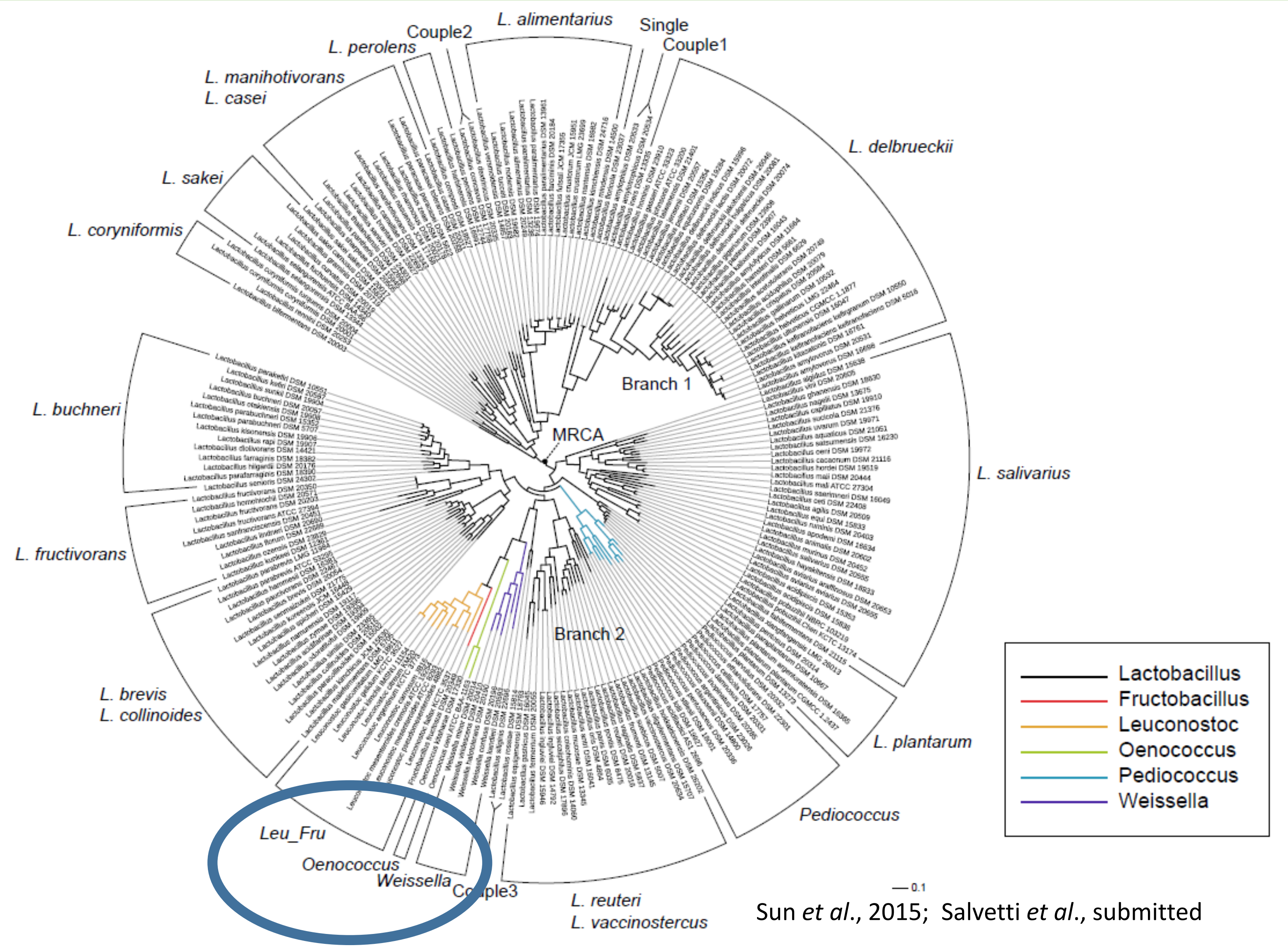
Sun, Harris, McCann, Guo, Argimón, Zhang *et al.* (2015)

Lactobacillus phylogenomics

genomics
sequence-based data:

cgMLST – 73 proteins
rMLST – 29 proteins
MLST – 12 markers

- Genus *Lactobacillus* is **polyphyletic**, **intermixed** with members of other genera
- **complex** evolutionary history



Sun *et al*., 2015; Salvetti *et al*., submitted

towards a new classification/1

- presence of
  - about 10 consistent groups which can be considered the nuclei for new genera - supported by combination of sequence-based and distance-based methods (Average Aminoacid Identity and Percentage of conserved proteins)
  - few couples and single lines of descent
- Back to the past: Lactobacillaceae and Leuconostocaceae appear to be intermixed: a revised classification beyond the genus level (family/order)?

towards a new classification/2

- Principle 1 of Prokaryotic Code (2008 Revision)
  - names should aim at stability, and
  - useless creation of names should be avoided

- careful revaluation of phenotypic characteristics and geno-pheno matching, discussed among experts (Subcommittee on the taxonomy of LAB)

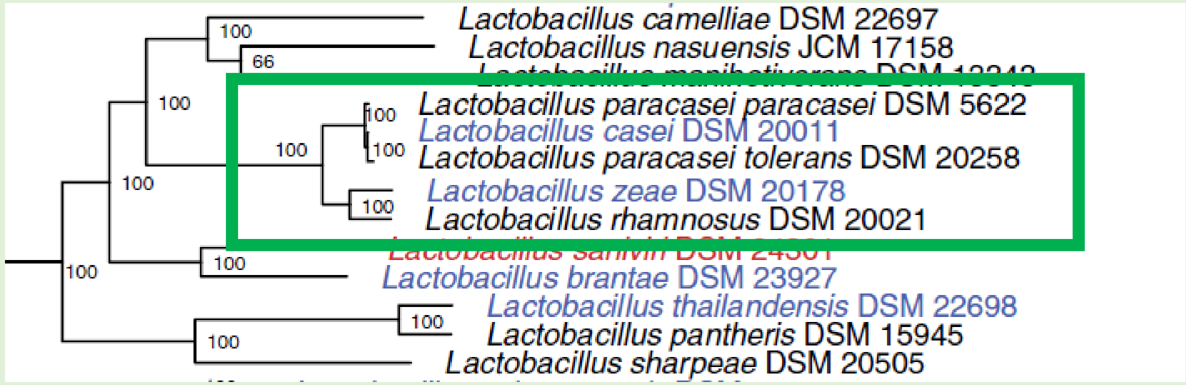
Examples in the genus *Lactobacillus*

The species level

Lactobacillus casei

L. casei group includes 3 species:

- *L. casei*
- *L. paracasei*
- *L. rhamnosus*


Lactobacillus camelliae DSM 22697
Lactobacillus nasuensis JCM 17158
Lactobacillus manihotivorans DSM 13343
Lactobacillus paracasei paracasei DSM 5622
Lactobacillus casei DSM 20011
Lactobacillus paracasei tolerans DSM 20258
Lactobacillus zeae DSM 20178
Lactobacillus rhamnosus DSM 20021
Lactobacillus saniviri DSM 24301
Lactobacillus brantae DSM 23927
Lactobacillus thailandensis DSM 22698
Lactobacillus pantheris DSM 15945
Lactobacillus sharpeae DSM 20505
100
66
100
100
100
100
100
100
100
100
100

former “*L. zeae*” synonym of *L. casei*

Lactobacillus casei - group

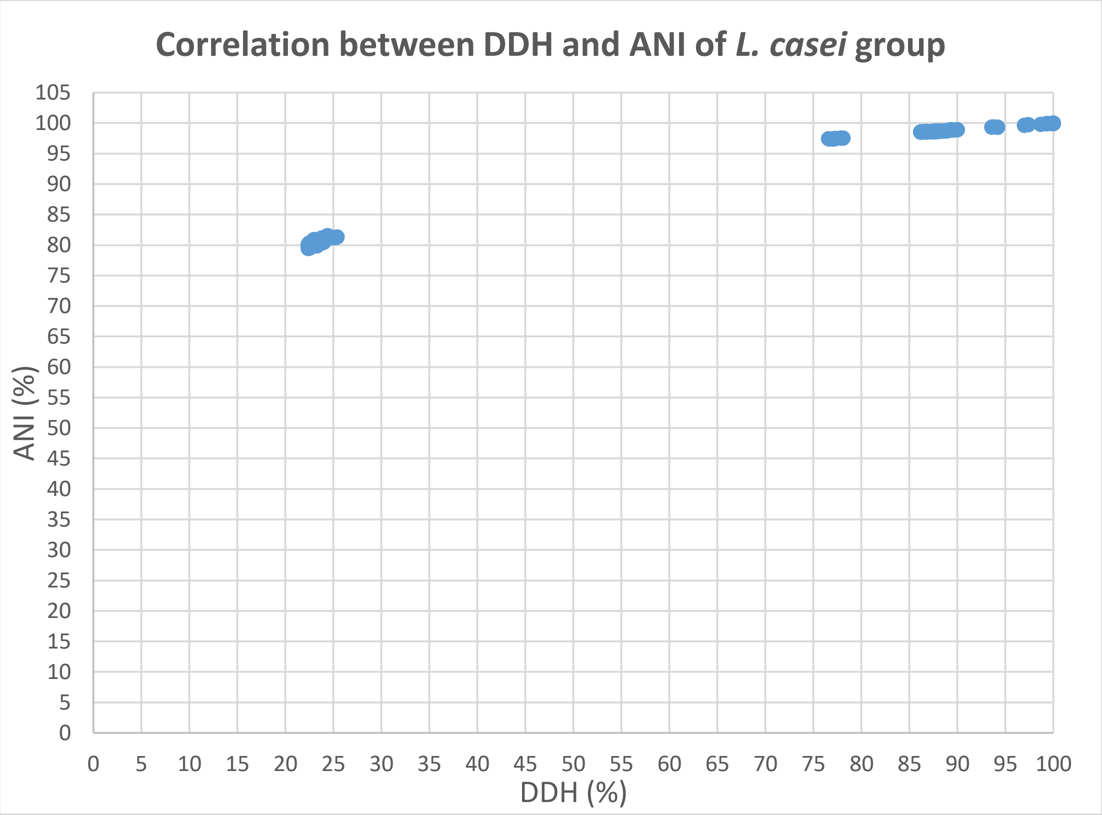
20 complete and public genome sequences (GenBank)

- dDDH

http://ggdc.dsmz.de/distcalc2.php

- ANI

http://enve-omics.ce.gatech.edu /ani/



Unpublished results

Lactobacillus casei - group

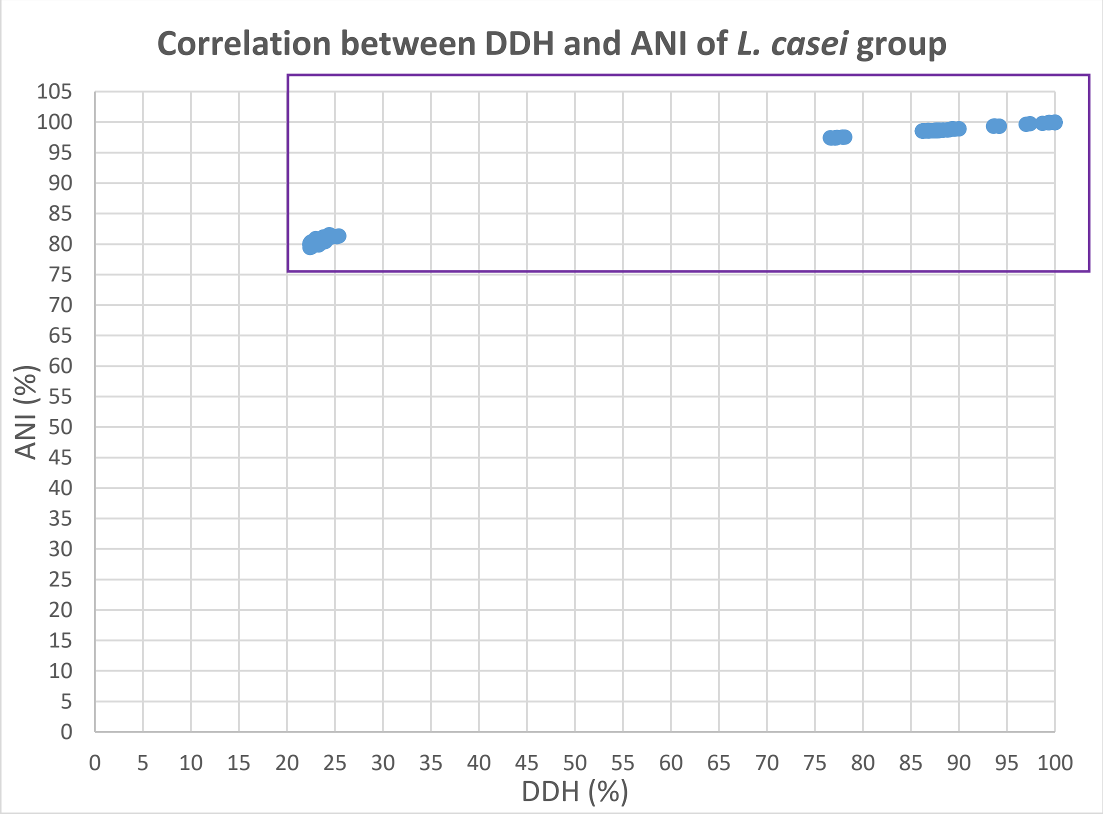
20 complete and public genome sequences (GenBank)

- dDDH

http://ggdc.dsmz.de/distcalc2.php

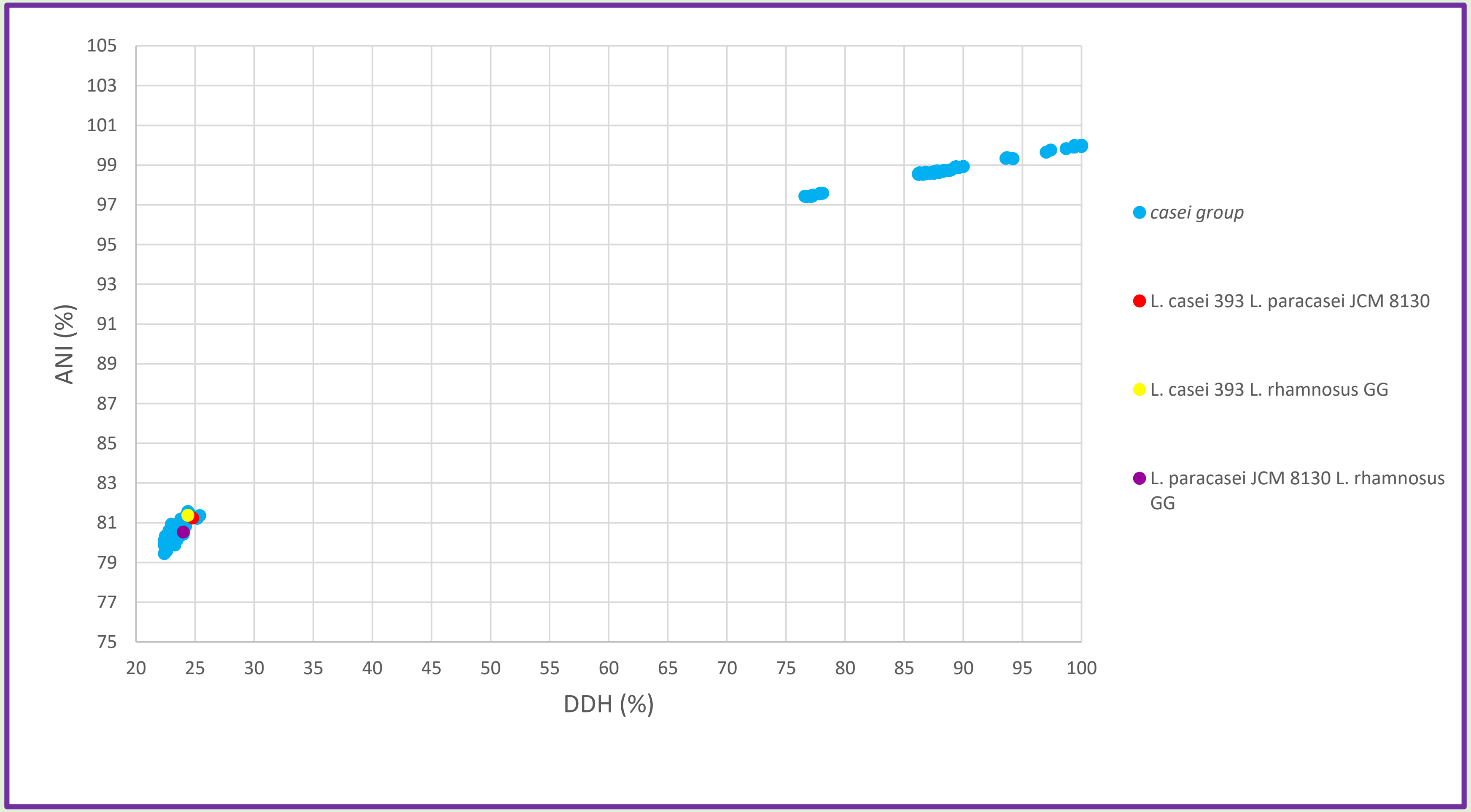
- ANI

http://enve-omics.ce.gatech.edu /ani/

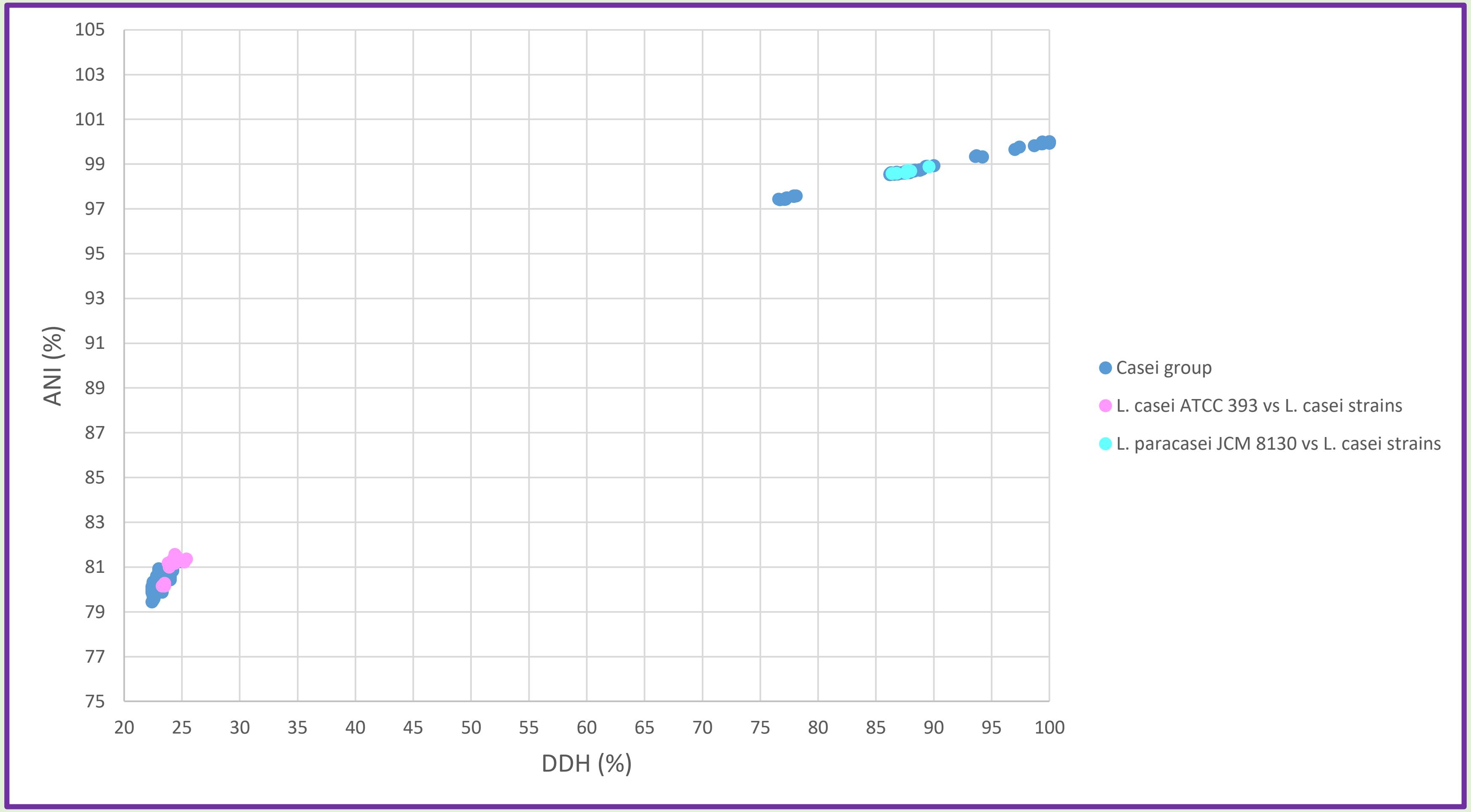


Unpublished results

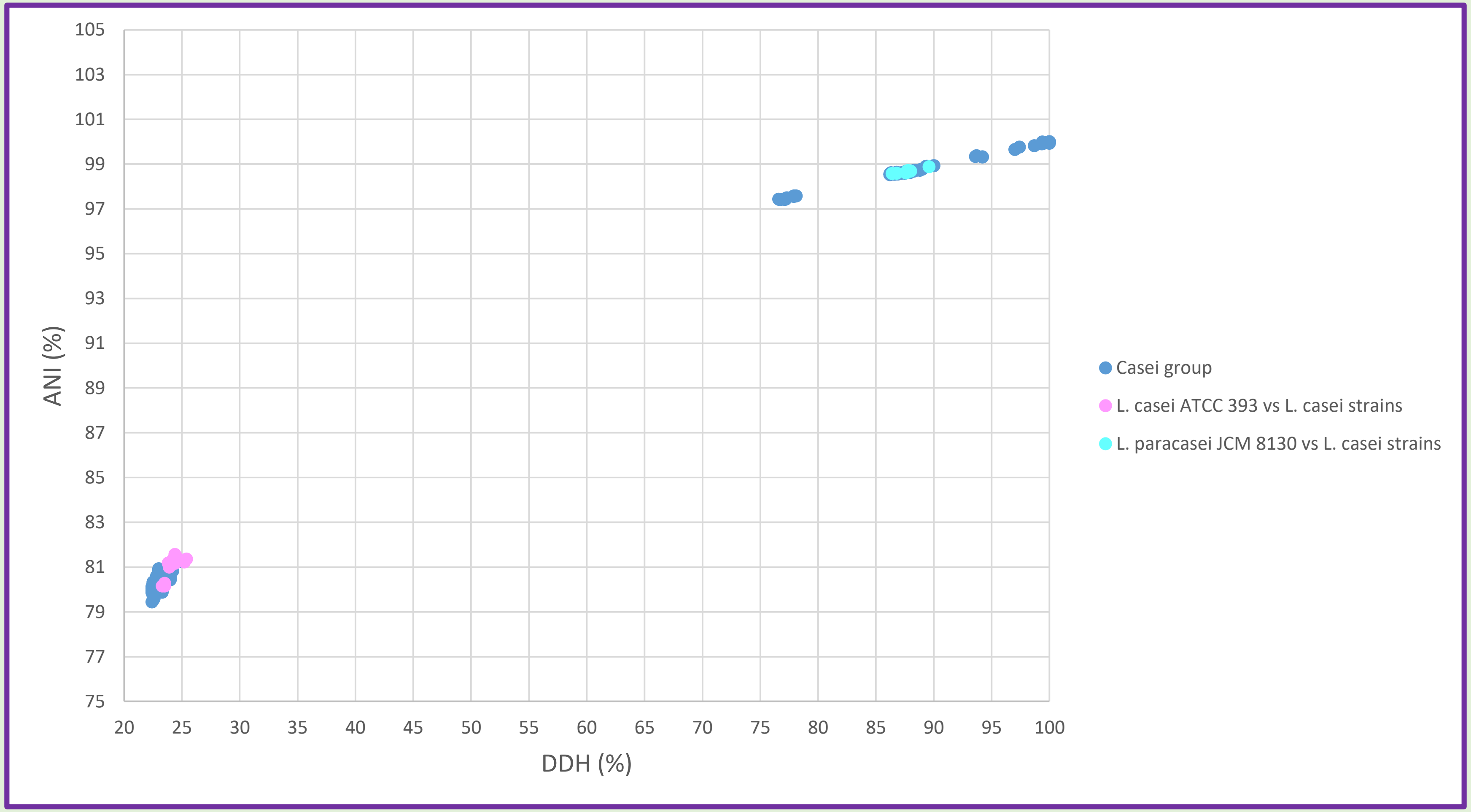
Lactobacillus casei - group



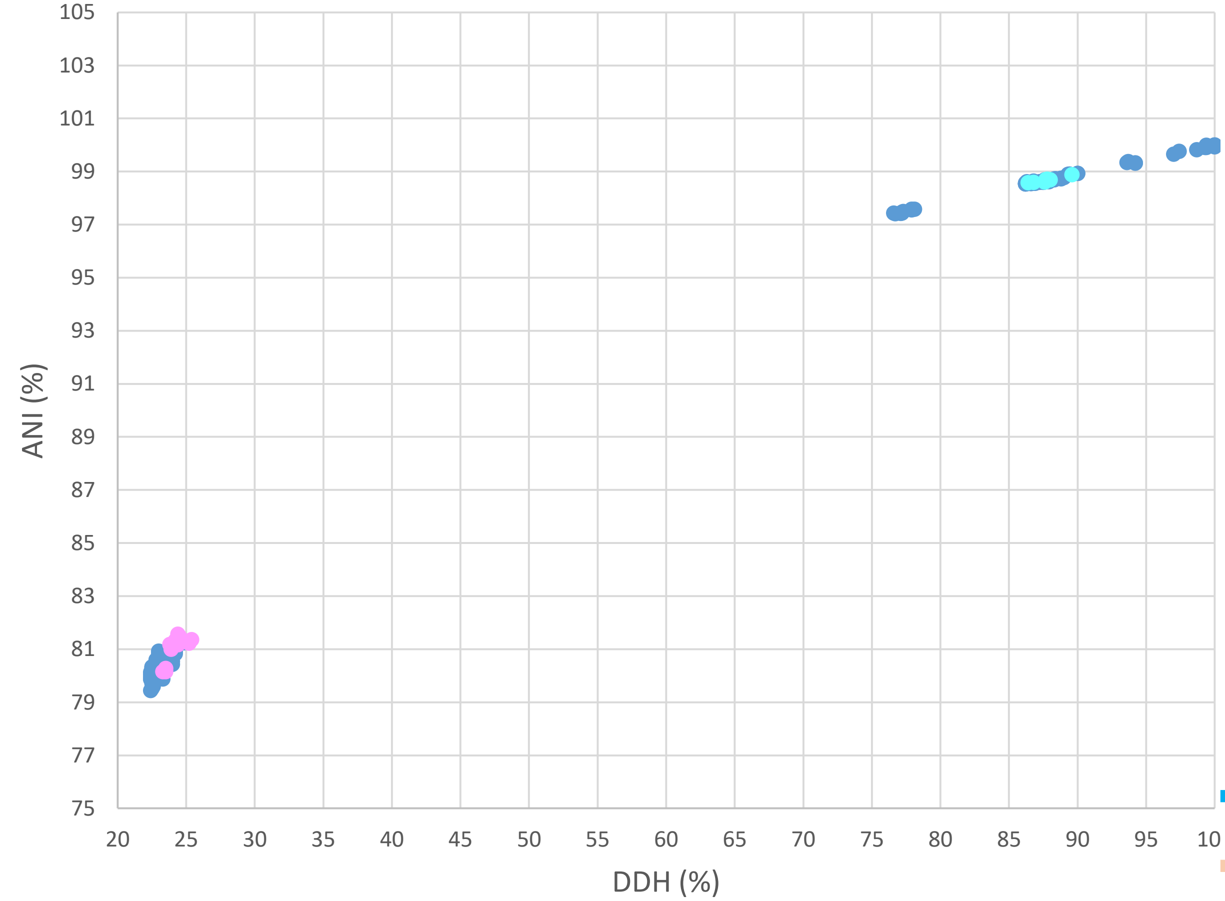
Lactobacillus casei - group



Lactobacillus casei - group



Lactobacillus casei - group



***L. casei* 12A**
***L. casei* ATCC334**
***L. casei* BD-II**
***L. casei* BL23**
***L. casei* LC2W**
***L. casei* LcA**
***L. casei* LcY**
***L. casei* LOCK919**
***L. casei* str. Zhang**
***L. casei* W56**
are more related to *L. paracasei* than to *L. casei*

Lactobacillus casei - group

- All the strains, except 12A and ATCC 334, are reported as probiotics

- Use of name L. casei could determine ambiguities and difficulty in communication and analysis of species-level properties

***L. casei* 12A**
***L. casei* ATCC334**
***L. casei* BD-II**
***L. casei* BL23**
***L. casei* LC2W**
***L. casei* LcA**
***L. casei* LcY**
***L. casei* LOCK919**
***L. casei* str. Zhang**
***L. casei* W56**
are more related to *L. paracasei* than to *L. casei*

Strain characterization

Safety evaluation

Other strain characteristics

Safety

European Food Safety Authority (EFSA) guidelines (EFSA 2013) requires the absence of the genetic make-up for

  - virulence factor (VF),
  - transmissible antibiotic resistance (AR) and
  - other deleterious characteristics

- safety assessments including complete genome sequences
  - *Bifidobacterium* strains (Bennedsen et al. 2011),
  - *Lactobacillus plantarum* JDM1 (Zhang et al. 2012)
  - *Bifidobacterium longum* JDM301 (Wei et al. 2012),
  - *Streptococcus salivarius* strains NU10 and YU10 (Barbour and Philip 2014),
  - *Enterococcus faecium* NRRL B-2354 (Kopit et al. 2014)
  - *Butyricicoccus pulicaecorum* 25-3T (Steppe et al. 2014)
  - *Lactobacillus helveticus* MTCC 5463 (Senan et al. 2015)
  - *Bacillus coagulans* GBI-30, 6086 (Salvetti et al., 2016)
  - *Lactobacillus helveticus* KLDS1.8701 (Li et al., 2017)

Bacillus coagulans GBI-30, 6086 as a case study

- sporeforming lactic acid-producing bacterium,
  - resists the harsh conditions of GIT
  - displays good stability during shelf life (Hyronimus et al. 2000; Maathuis et al. 2010).
  - Commercial name: GanedenBC30™ (BC30), deposited in the American Type Culture Collection as *B. coagulans* PTA-6086.
- probiotic properties :
  - improves gastrointestinal quality of life in adults with postprandial intestinal gas-related symptoms (Kalman et al. 2009);
  - aid in protein, lactose and fructose digestion (Maathuis et al. 2010);
  - antimicrobial activity in distal regions of the GI tract (Honda et al. 2011) and
  - Improvement of some parameters of *Clostridium difficile*-induced colitis in mice and limitation of recurrence (Fitzpatrick et al. 2011; Fitzpatrick et al. 2012).
- Other aspects include
  - studies assessing its immunomodulatory properties (Jensen et al. 2010; Benson et al. 2012) and
  - stimulating effects on other beneficial genera of bacteria, organic acid production in the elderly (Nyangale et al. 2014).

Preliminary indications on safety

- **Safe history of use** supported by
  - a toxicological safety assessment (Endres et al. 2009)
  - a 1-year chronic oral toxicity study (Endres et al. 2011).
- Notice of Ganeden Biotech, Inc. to US FDA (Food and Drug Administration) reported **unpublished PCR protocols** that demonstrated that the strain does not contain genes homologous to those encoding known **protein toxins and haemolysin** (Ganeden Biotech, Inc. 2011) → Generally Recognized As Safe (**GRAS**) status in 2012 from the FDA.
- *B. coagulans* is in the Qualified Presumption of Safety (**QPS**) list by EFSA as feed additive since 2007 (EFSA 2007) thanks to the certified absence of toxigenic potential.

Antibiotic resistance - phenotype

Phenotypic tests were performed, and results were compared to MIC cut-off values **for *Bacillus* species**

GBI-30, 6086 was

- resistant to kanamycin and streptomycin

- MIC values > 1500 mg/L
- MIC cut-off values for *Bacillus* species 8 mg/L or 64 mg/L according to a previous EU document

- susceptible to ampicillin (0.125 mg/L), chloramphenicol (0.25 mg/L), ciprofloxacin (0.03 mg/L), clindamycin (0.125 mg/L), erythromycin (0.125 mg/L), gentamycin (0.031 mg/L), linezolid (0.06 mg/L), neomycin (2 mg/L), rifampicin (0.016 mg/L), tetracycline (0.25 mg/L), trimethoprim (0.063 mg/L), vancomycin (0.063 mg/L) and virginiamycin (0.016 mg/L).

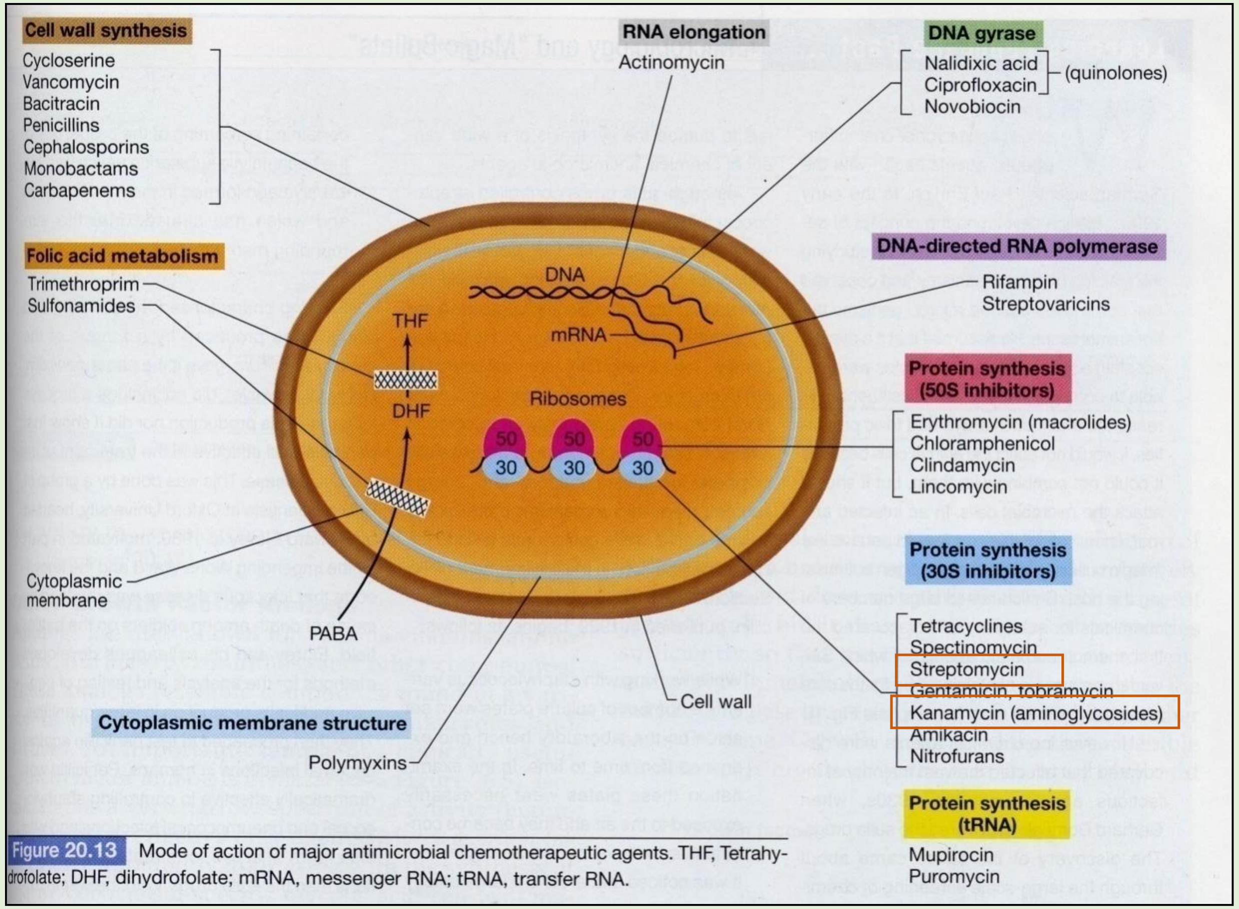


Figure 20.13 Mode of action of major antimicrobial chemotherapeutic agents. THF, Tetrahydrofolate; DHF, dihydrofolate; mRNA, messenger RNA; tRNA, transfer RNA.

Antibiotic resistance - genotype

- **Comprehensive Antibiotic Resistance Database (CARD)** (AR-related genes (E < 1e-2, coverage > 70 % and similarity > 30 %).
- Identification of **109 putative AR genes**:
    - transporters (57),
    - genes modulating the antibiotic efflux (9),
    - genes associated with resistance to daptomycin (6), polymyxin (1), streptothricin (1), penicillin (5), vancomycin (13), elfamycin (1), rifampin (2), sulphonamide (1), macrolides (as erythromycin, streptogramin and chloramphenicol) (2), fluoroquinolone (2), aminocoumarin (2) trimethoprim (1),
    - other genes related to a non-specified antibiotic resistance (4) and aminoglycosides (2).

Antibiotic resistance

- The two identified **aminoglycoside** resistance genes
  1. IE89_07115 → ribosomal protein S12 of subunit 30S
     - the ribosome alteration is one of the main aminoglycoside resistance mechanisms that can be mediated by 16S rRNA methylases and methyltransferases or intrinsic mechanisms as chromosomal mutations
     - No other rRNAmethylases or methyltransferases were detected → it can be assumed that *B. coagulans* GBI-30, 6086 underwent **events of mutation in IE89_07115**, thus, becoming **intrinsically resistant**.
     - The **absence of mobile elements** in the surrounding regions suggests the **low risk of gene transfer**

Antibiotic resistance

- The two identified **aminoglycoside** resistance genes
  2. IE89_03650 → aminoglycoside 3-Nacetyltransferase.
     - Gene similar (e-value: 3e-41; similarity: 31, 36 %, query coverage 98 %) to the gene encoding for an aminoglycoside 3-N-acetyltransferase from a *Micromonospora chalcea* isolate.
     - analysis of the **flanking regions**:
       - the gene is co-localized on the chromosome with a gene encoding for a multidrug transporter MatE (IE89_03645), and this organization is detectable in all available *B. coagulans* genomes in NCBI
       - no mobile elements as transposases and insertion sequences in the flanking regions of the gene →very low risk of HGT

Antibiotic resistance/5

- The phenotypic and genomic analysis of AR in *B. coagulans* GBI-30, 6086 showed:
  - phenotypic resistance to streptomycin and kanamycin.
  - probable determinants for this resistance appear to be not easily transferrable to other bacteria
  - → support to the safety of this strain with respect to antibiotic resistance.
- **no other AR phenotypes despite the genes highlighted**

Biogenic amine production: pheno-geno

- HPLC analyses → tyramine, histamine, putrescine, cadaverine and phenyletilamine, and the polyamines, spermine and spermidine, were **not produced** by *B. coagulans* GBI-30, 6086 in the conditions used
- **genes** for BA production were **generally absent, except** entire metabolic pathway
  - from arginine to putrescine
  - from putrescine to spermidine
  - carboxyspermidine dehydrogenase/carboxyspermidine decarboxylase (CASDH/CASDC) system

  → Could those compounds be produced in gut-like conditions?

Putative virulence factors/VFDB

- BLAST analysis against the **Virulence Factor Database (VFDB)**(Chen et al. 2012)
- Identification of **200 genes putatively related to virulence** (E < 1e-2, coverage > 70 % and similarity > 30%)
    - **eight genes were classified as related to defense mechanisms,** annotated as:
        - Multidrug transporters and resistance proteins (also previously detected by CARD),
        - a peroxidase
        - an alkyl hydroperoxide reductase, essential to adapt in response to redox changes (Zuo et al. 2014).
    - several putative VFs: the majority related to **extracellular structures**

    → could represent **essential probiotic traits for the adhesion** to the host cells, or for the **sporulation** mechanism!

Putative virulence factors/2

- According to Clusters of Orthologous Groups (COG) database (http://www.ncbi.nlm.nih.gov/COG/), most of these genes were defensive or non-classical virulence factors, such as determinants related to:
  - transcription, translation, post-translational modifications,
  - ribosomal structure and biogenesis,
  - replication, recombination and repair,
  - cell motility,
  - signal transduction mechanisms,
  - intra- and extracellular transportation,
  - metabolism and transport of lipids, coenzymes, amino acids and carbohydrates,
  - signal transduction mechanisms,
  - cell cycle control,
  - cell division and chromosome partitioning,
  - protein turnover and chaperones,
  - energy production and conversion and
  - membrane biogenesis.

Putatively adverse metabolites/1

- BLASTX analysis showed that *B. coagulans* GBI-30, 6086 **does not carry:**
  - **any known enterotoxin genes**
  - genes encoding for surfactins, cyclic lipopeptides (create damages to the host epithelial and sperm cells) produced by all haemolytic *Bacillus* strains
  - genes encoding for other lipopeptides with toxin activity as the fengycin and the lychenisin (EFSA 2011)
  - **genes** encoding for the haemolysin BL, the non-haemolytic enterotoxin (Nhe, mostly associated with diarrhoeal outbreaks), the enterotoxins K and T and the emetic toxin (cereulide) (EFSA 2011)

→ confirming the toxicological analysis previously performed (Endres et al. 2009)

Stability of the genome/1

- presence of proteins annotated as **transposases**
  - 9 complete transposase-encoding genes were identified, but **none of their flanking genes were associated with AR or other putatively adverse genes**.
- ProphageFinder:
  - presence of 2 prophage-like elements:
    - **no gene was found for the tail tape measure protein**, one of the phage essential proteins, **no attL and attR sites** in both the prophage regions → defective and non-functional phages

Proposed modus operandi

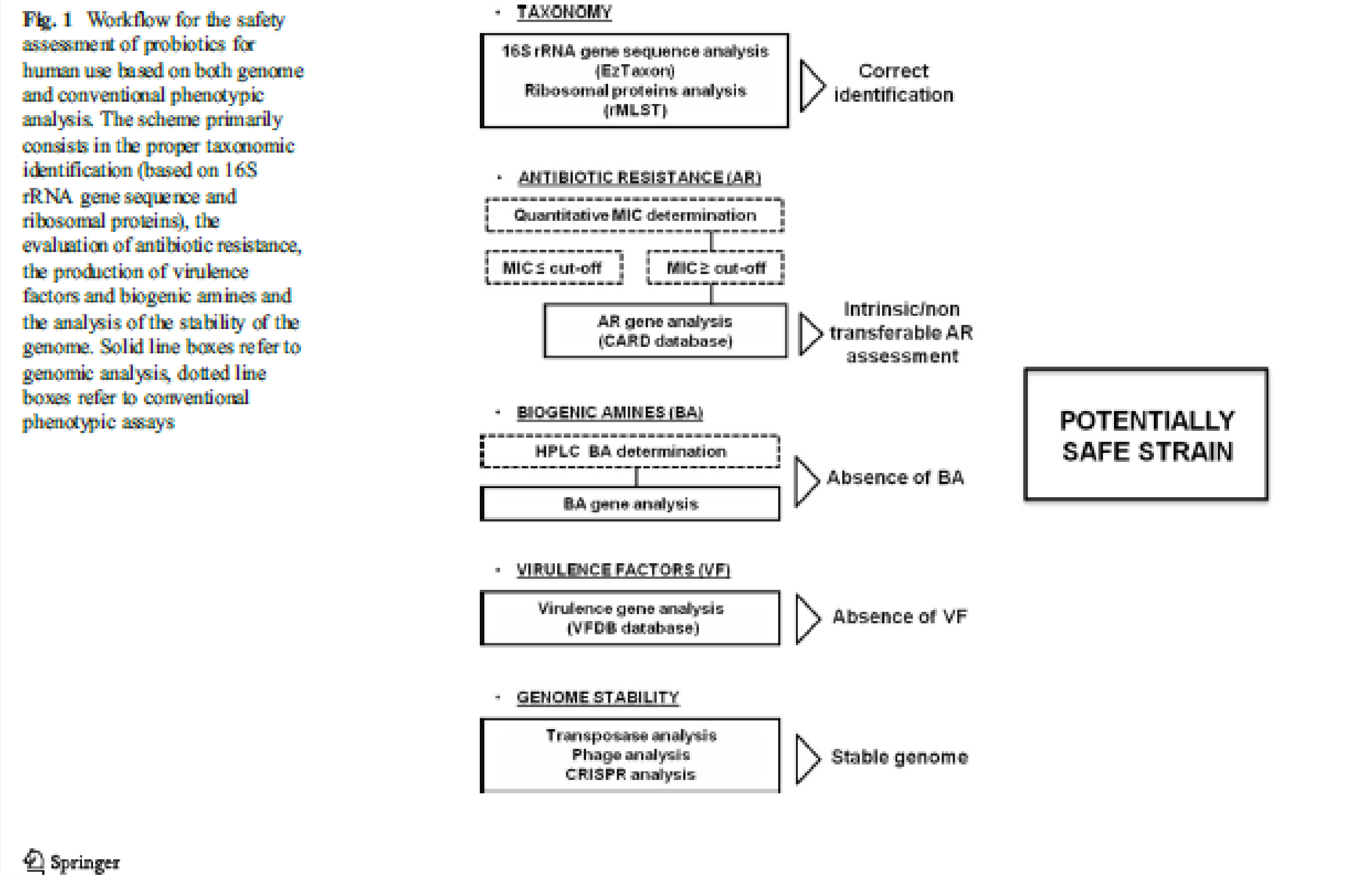


Fig. 1 Workflow for the safety assessment of probiotics for human use based on both genome and conventional phenotypic analysis. The scheme primarily consists in the proper taxonomic identification (based on 16S rRNA gene sequence and ribosomal proteins), the evaluation of antibiotic resistance, the production of virulence factors and biogenic amines and the analysis of the stability of the genome. Solid line boxes refer to genomic analysis, dotted line boxes refer to conventional phenotypic assays

- TAXONOMY

16S rRNA gene sequence analysis (EzTaxon)
Ribosomal proteins analysis (rMLST)

Correct identification

- ANTIBIOTIC RESISTANCE (AR)

Quantitative MIC determination

MIC ≤ cut-off

MIC ≥ cut-off

AR gene analysis (CARD database)

Intrinsic/non transferable AR assessment

- BIOGENIC AMINES (BA)

HPLC BA determination

BA gene analysis

Absence of BA

- VIRULENCE FACTORS (VF)

Virulence gene analysis (VFDB database)

Absence of VF

- GENOME STABILITY

Transposase analysis
Phage analysis
CRISPR analysis

Stable genome

POTENTIALLY SAFE STRAIN

Springer

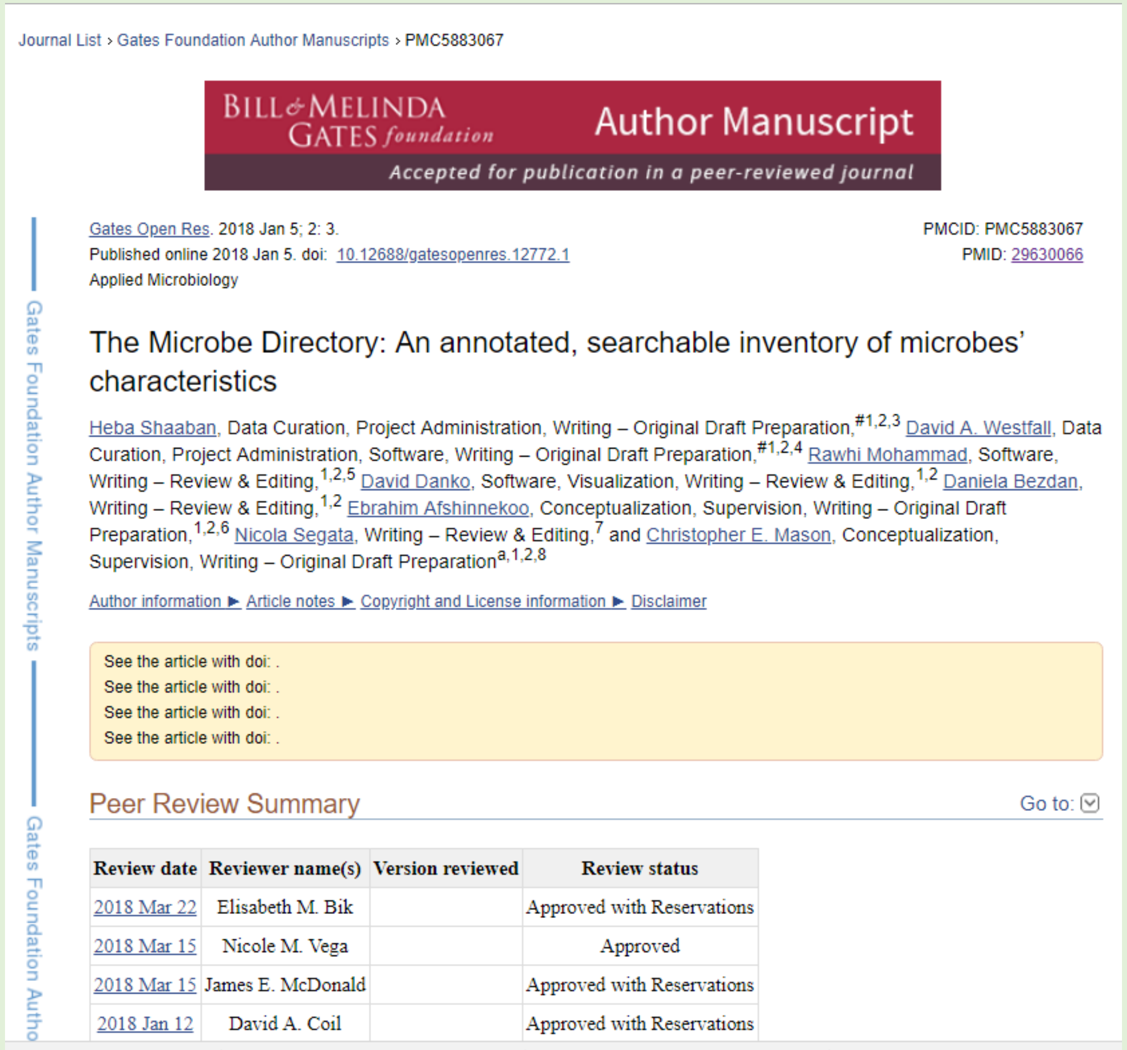
Other interesting databases

- CARD

for function identification

- Kyoto Encyclopaedia of Genes and Genomes (KEGG)
- Carbohydrate-Active enZYmes Database **http://www.cazy.org/**
- database of Clusters of Orthologous Groups of proteins (COG) http://www.ncbi.nlm.nih.gov/COG/ ftp://ftp.ncbi.nih.gov/pub/COG/COG2014/static/lists/listCOGs.html

Eventually…



Journal List › Gates Foundation Author Manuscripts › PMC5883067

Gates Open Res. 2018 Jan 5; 2: 3.
Published online 2018 Jan 5. doi: 10.12688/gatesopenres.12772.1
Applied Microbiology

PMCID: PMC5883067
PMID: 29630066

The Microbe Directory: An annotated, searchable inventory of microbes' characteristics

Heba Shaaban, Data Curation, Project Administration, Writing – Original Draft Preparation,[#1,2,3] David A. Westfall, Data Curation, Project Administration, Software, Writing – Original Draft Preparation,[#1,2,4] Rawhi Mohammad, Software, Writing – Review & Editing,[1,2,5] David Danko, Software, Visualization, Writing – Review & Editing,[1,2] Daniela Bezdan, Writing – Review & Editing,[1,2] Ebrahim Afshinnekoo, Conceptualization, Supervision, Writing – Original Draft Preparation,[1,2,6] Nicola Segata, Writing – Review & Editing,[7] and Christopher E. Mason, Conceptualization, Supervision, Writing – Original Draft Preparation[a,1,2,8]

Author information ► Article notes ► Copyright and License information ► Disclaimer

See the article with doi: .
See the article with doi: .
See the article with doi: .
See the article with doi: .

Peer Review Summary

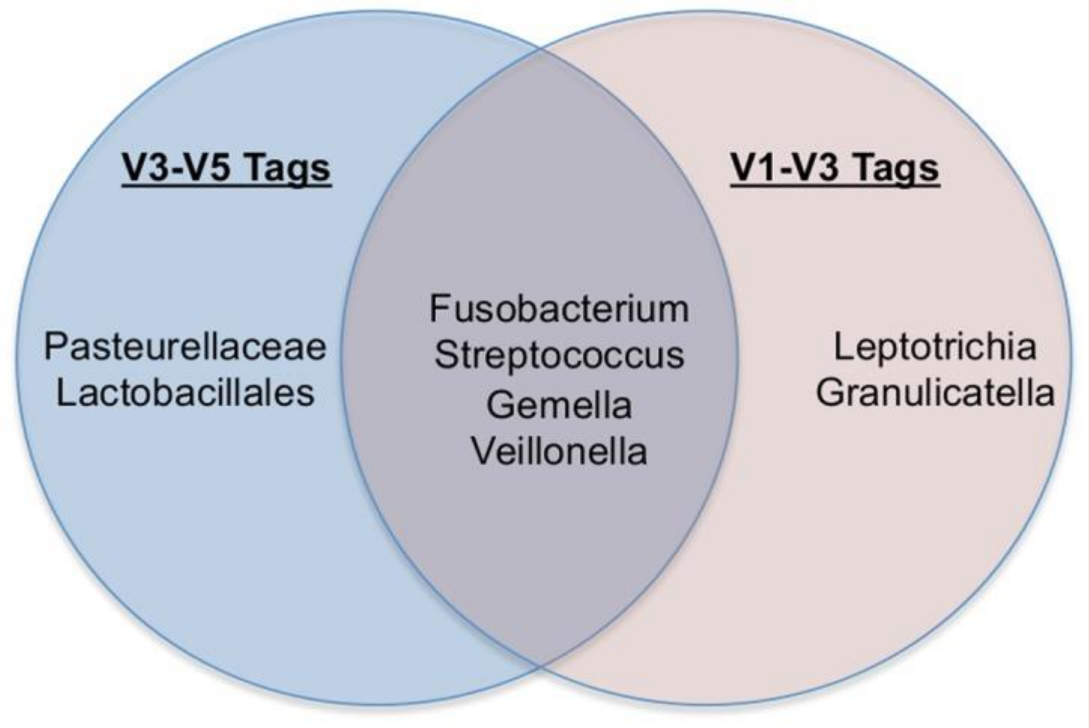
Go to:

Review date	Reviewer name(s)	Version reviewed	Review status
2018 Mar 22	Elisabeth M. Bik		Approved with Reservations
2018 Mar 15	Nicole M. Vega		Approved
2018 Mar 15	James E. McDonald		Approved with Reservations
2018 Jan 12	David A. Coil		Approved with Reservations

Taxonomy and metagenomic data

Getting sense from metagenomic data

- Amplicon sequencing (16S partial sequencing) and OTU assignments
  - 97% as threshold for OTU assignment
  - SNPs (DADA2 R package)
  - Tax4Fun, R Package (http://tax4fun.gobics.de/) to infer metabolic capabilities
- WMS
  - Strain level? (Segata mSystems. 2018 Mar 13;3(2). pii: e00190-17. doi: 10.1128/mSystems.00190-17)



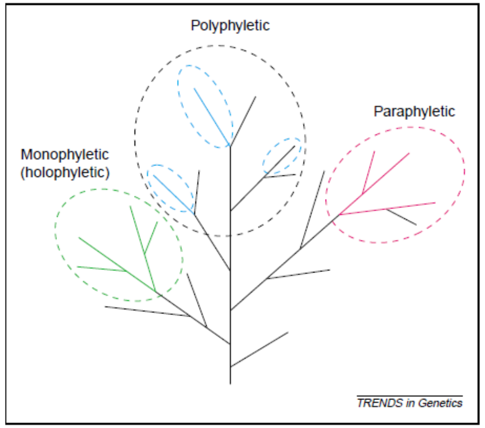
Susan M. Huse, et al. PLoS One. 2012;7(6):e34242.

Phylogenetic analysis

From Baldauf 2003

Phylogenetic analysis

- is a powerful tool for sorting and interpreting molecular data.
- With a very basic understanding of general principles and conventions it is possible to glean valuable information from a phylogenetic tree, e.g., on the origin, evolution and possible function of genes and the proteins they might encode



TRENDS in Genetics

Fig. 1. Trees are about groups: monophyletic (holophyletic), paraphyletic and 'polyphyletic'.

Terminology

- A phylogenetic tree is a graph, composed of **branches** (edges) and **nodes**
- Branches connect nodes
- A node is the point at which two (or more) branches diverge.
- Branches and nodes can be internal or external (terminal).
  - internal node → hypothetical last common ancestor (LCA) of everything arising from it
  - Terminal nodes → sequences from which the tree was derived (also referred to as operational taxonomic units or 'OTUs').
- **Trees can be made up of multigene families (gene trees) or a single gene from many taxa (species trees, at least theoretically) or a combination of the two. In the first case, the internal nodes correspond to gene duplication events, in the second to speciation events.**

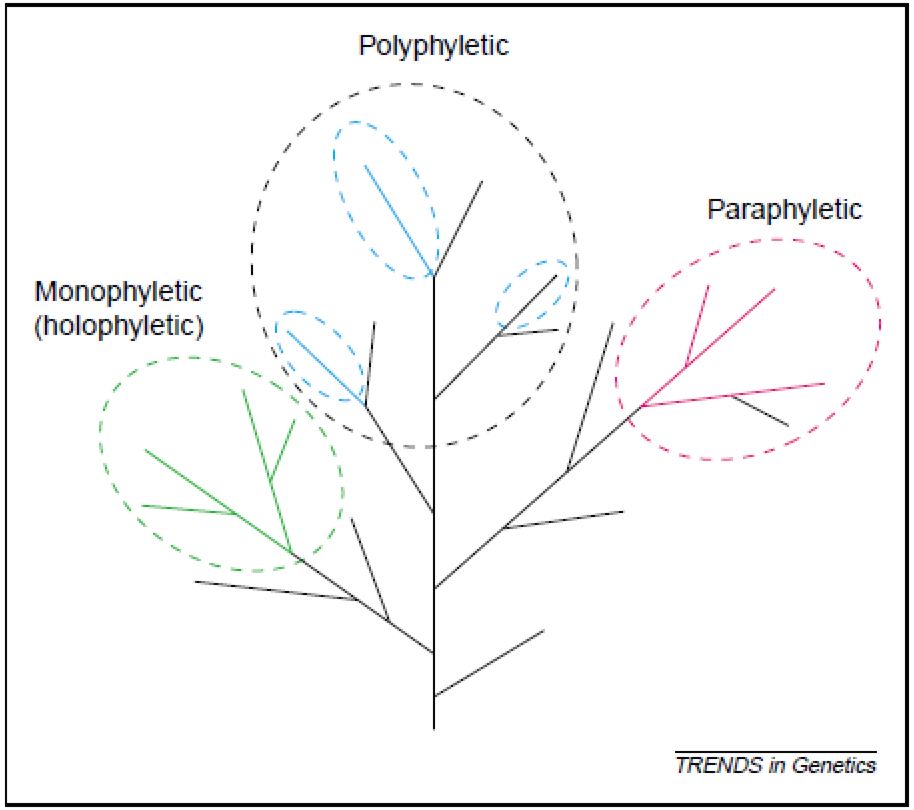


Fig. 1. Trees are about groups: monophyletic (holophyletic), paraphyletic and 'polyphyletic'.

Groups

- Trees are about groupings
- A node and everything arising from it is a ‘clade’ or a ‘monophyletic group’.
- A **monophyletic** group is a **natural** group; all members are derived from a unique common ancestor (with respect to the rest of the tree) and have inherited a set of unique common traits (characters) from it.
- A group excluding some of its descendents is a **paraphyletic** group

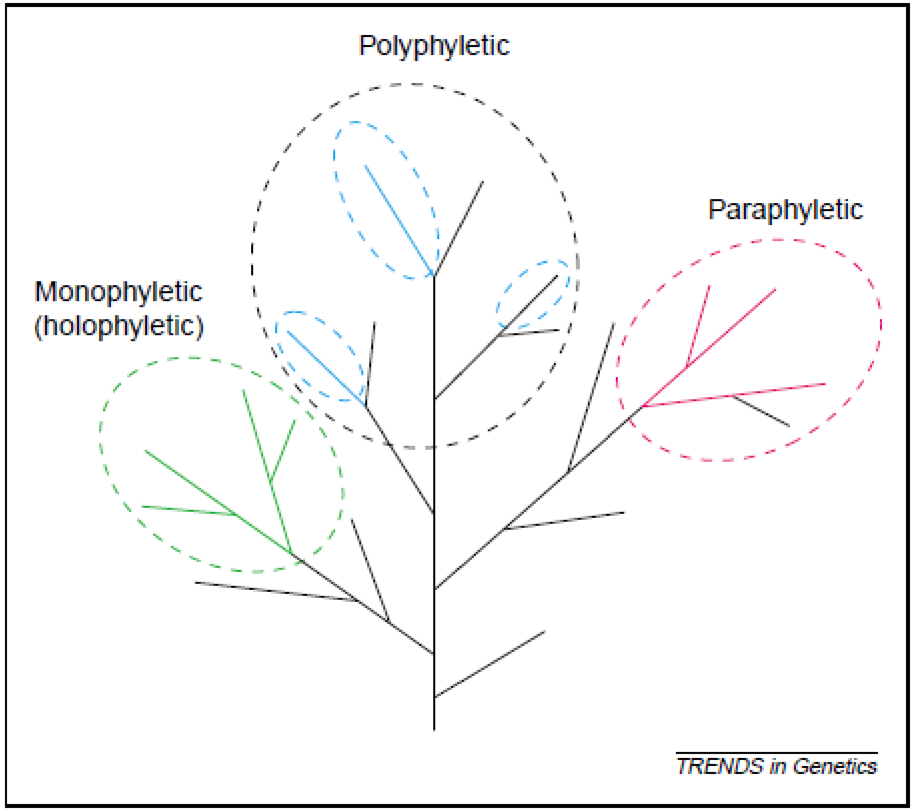


Fig. 1. Trees are about groups: monophyletic (holophyletic), paraphyletic and ‘polyphyletic’.

Trees

- Intuitively we draw trees from the ground up (Fig. a).
- To make large tree more readable, we can expand the nodes (Fig. b) and turn the tree on its side (Fig. c).
- → tree grows left to right, and all the labels are horizontal
  - easier to read and to annotate
  - widths of the nodes have no meaning
- all branches can rotate freely about the plane of their nodes, so all trees in Fig. are identical (except tree F, unrooted)

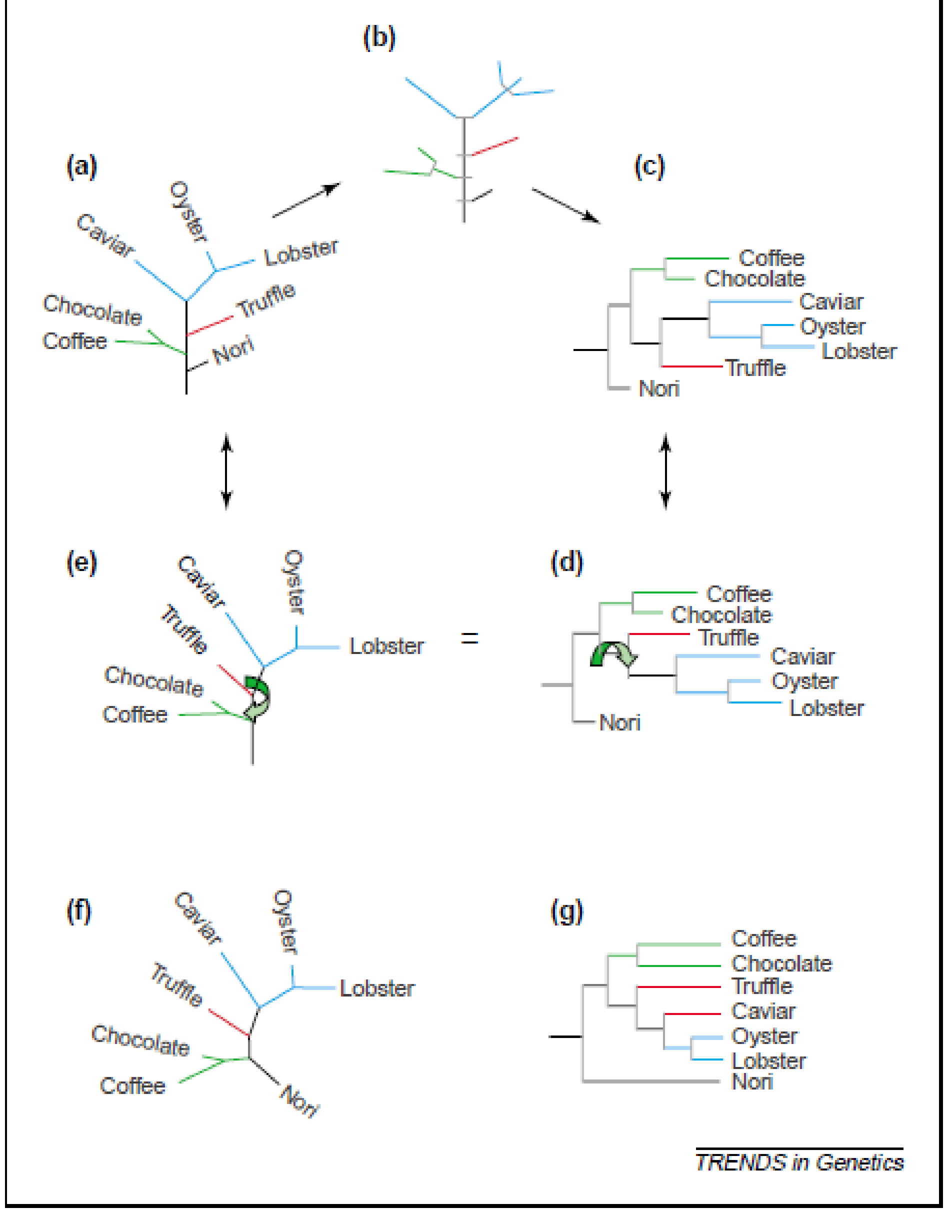


Fig. 2. Phylogenetic tree styles. All these trees have identical branching patterns. The only differences are (f), which is unrooted. (g) is a cladogram, so the branch lengths are right justified and not drawn to scale (i.e. they are not proportional to estimated evolutionary difference).

Trees

- trees are usually drawn with proportional branch lengths → the lengths of the branches correspond to the amount of evolution (roughly, % seq divergence) between the two nodes they connect (Fig. a–f)
- the longer the branches the more relatively divergent (highly evolved) are the sequences attached to them
- Alternatively, trees can be drawn to display branching patterns only ('cladograms')→ **lengths of the branches have no meaning** (Fig. g), (rarely done with molecular sequence trees)

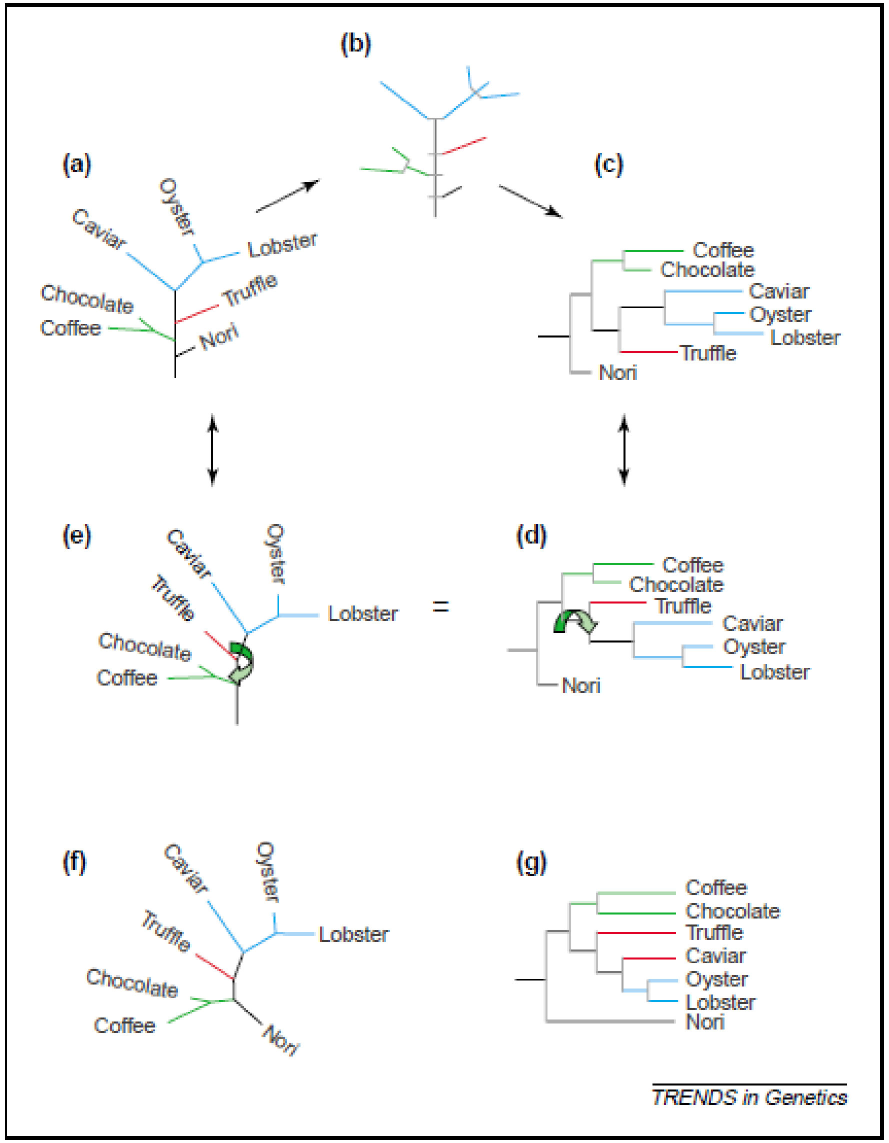


Fig. 2. Phylogenetic tree styles. All these trees have identical branching patterns. The only differences are (f), which is unrooted. (g) is a cladogram, so the branch lengths are right justified and not drawn to scale (i.e. they are not proportional to estimated evolutionary difference).

Root and outgroup

- The root is the base of a phylogenetic tree
- It is the oldest point in the tree → it implies the order of branching in the rest of the tree
- Branching order → who shares a more recent common ancestor with whom.
- **The only way to root a tree is with an 'outgroup', an external point of reference. An outgroup is anything that is not a natural member of the group of interest (i.e. the 'ingroup')**
- In the absence of a certain outgroup, place the root in the middle of the tree (at its midpoint), or don't root the tree (Fig. f)

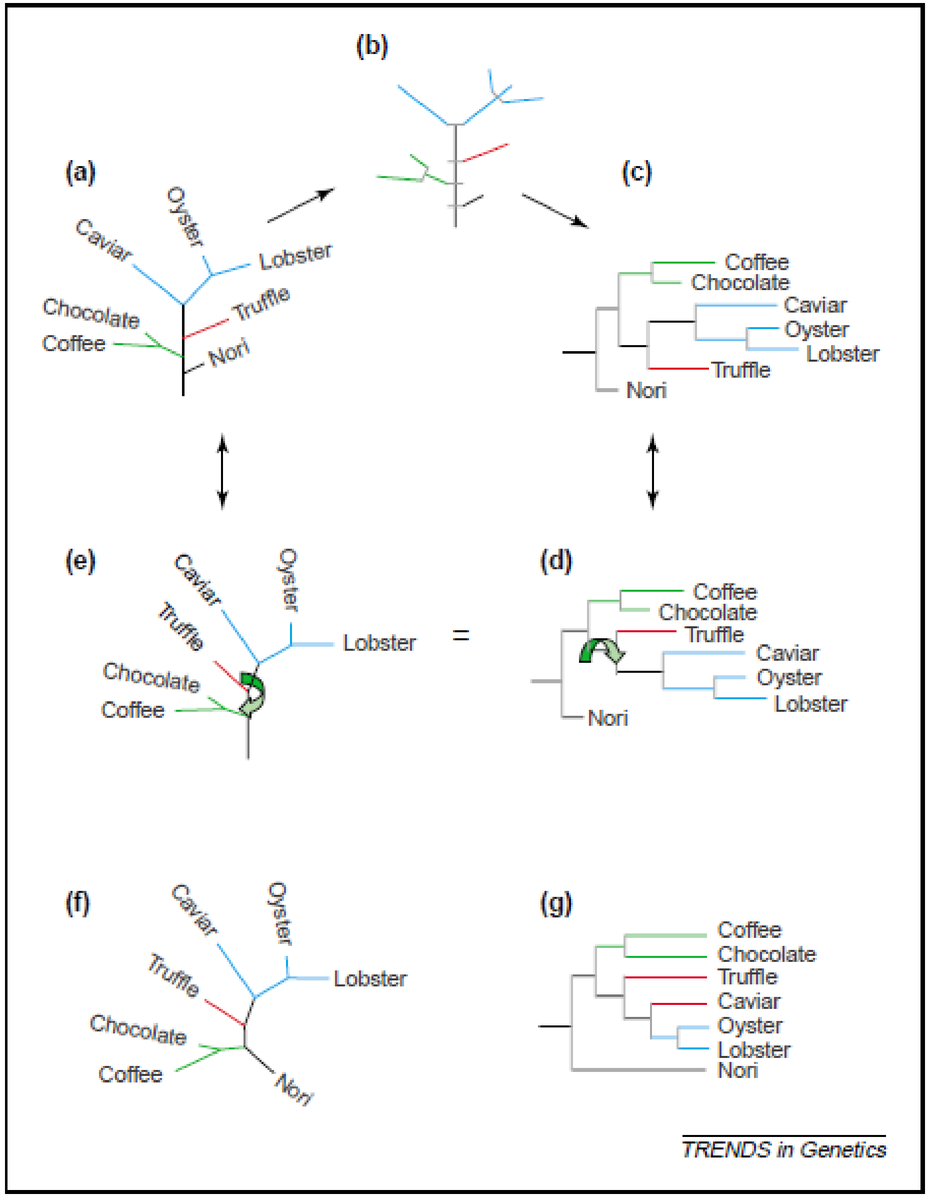
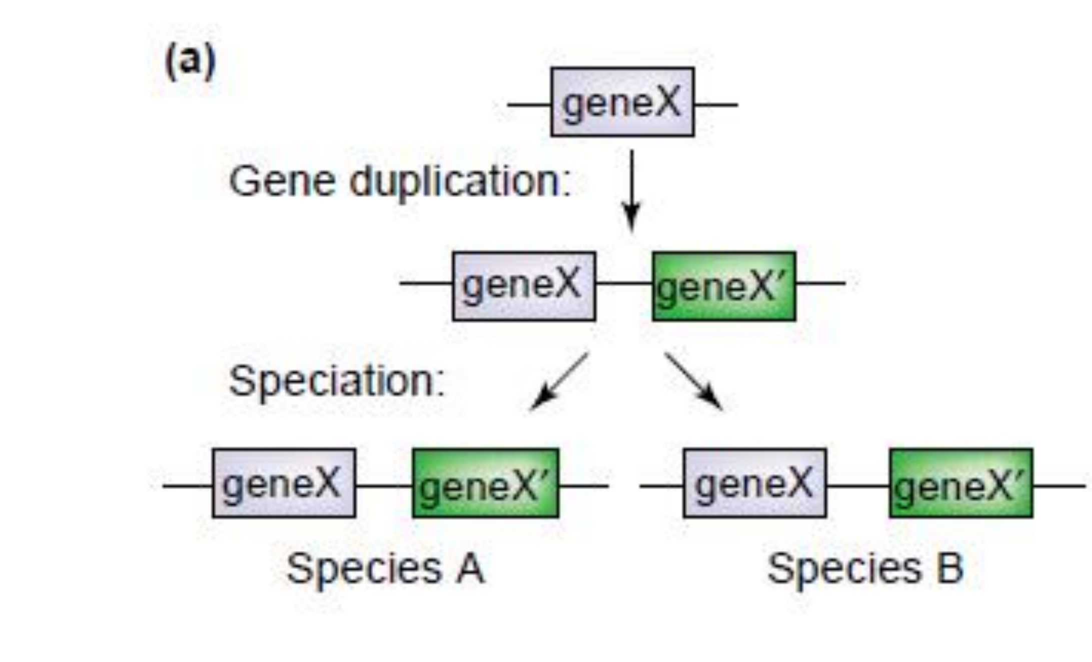


Fig. 2. Phylogenetic tree styles. All these trees have identical branching patterns. The only differences are (f), which is unrooted. (g) is a cladogram, so the branch lengths are right justified and not drawn to scale (i.e. they are not proportional to estimated evolutionary difference).

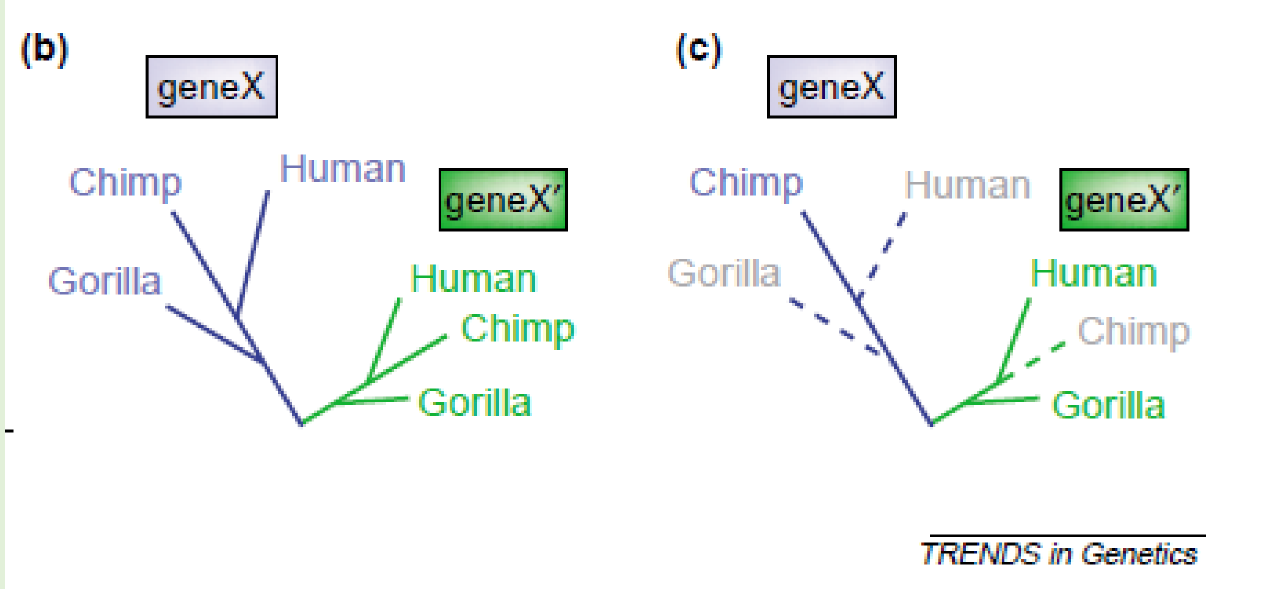
Homology

- Evolution is about homology → similarity due to common ancestry
- Homologues can be
  - **Orthologues**: only duplicate when their host divides, strictly vertically transmitted → their phylogeny traces that of their host lineage
  - **Paralogues:** come from gene duplications, member of a multigenic family



The problem with paralogues

- Inference of species relationships with paralogues can lead to troubles
  - if all copies of two paralogues are in the tree, OK (Fig b), also, there are two mirror phylogenies and paralogues can serve as each other's natural outgroup
  - if some of the copies are missing, phylogeny is misleading (Fig. c)



TRENDS in Genetics

Building trees

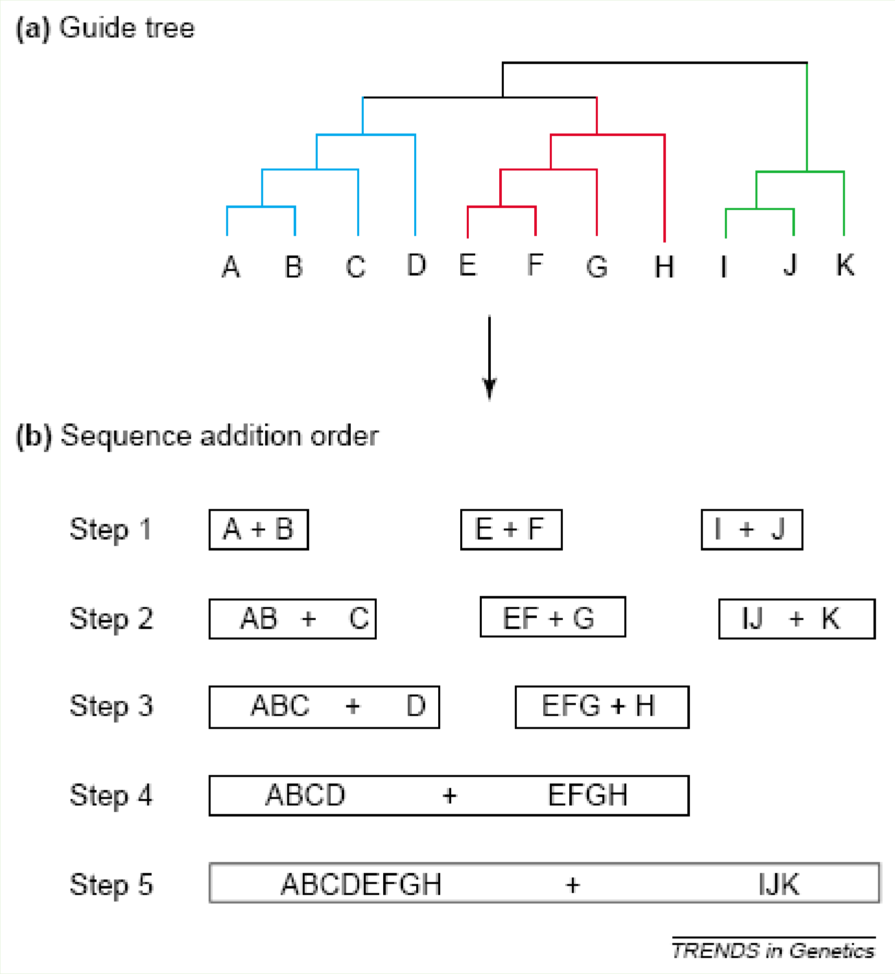
Five steps

1. Assembling a dataset
2. Multiple sequence alignment
3. Trees
4. Tests
5. Data presentation

Step 1. Assembling a dataset

- Finding and retrieving sequences from the public domain (GenBank, EMBL, DDBJ)
- Avoid text search, prefer sequence similarity search (Blast)

Step 2. Multiple sequence alignment



(a) Guide tree

(b) Sequence addition order

Step 1 A + B E + F I + J

Step 2 AB + C EF + G IJ + K

Step 3 ABC + D EFG + H

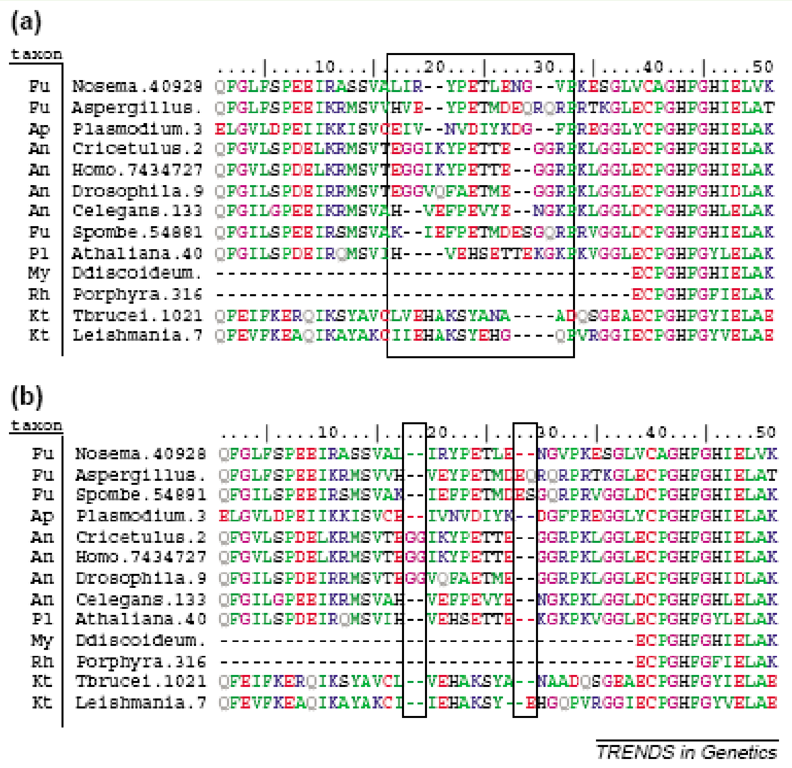
Step 4 ABCD + EFGH

Step 5 ABCDEFGH + IJK

TRENDS in Genetics

- Steps in progressive sequence alignment
  - guide tree which determines the order in which sequences are added to the growing alignment
  - Refinement of the alignment

Step 2. Multiple sequence alignment



(a)

```
taxon
                          ....|....10...|....20...|....30...|....40...|....50
 Fu   Nosema.40929        QFGLFSPEEIRASSVALIR--YPETLENG--VPKESGLVCAGHFGHIELVK
 Fu   Aspergillus.        QFGLFSPEEIKRMSVVHVE--YPETMDEQRQRPRTKGLECPGHFGHIELAT
 Ap   Plasmodium.3        ELGVLDPEIIKKISVCEIV--NVDIYKDG--FPREGGLYCPGHFGHIELAK
 An   Cricetulus.2        QFGVLSPDELKRMSVTEGGIKYPETTE--GGRPKLGGLECPGHFGHIELAK
 An   Homo.7434727        QFGVLSPDELKRMSVTEGGIKYPETTE--GGRPKLGGLECPGHFGHIELAK
 An   Drosophila.9        QFGILSPDEIRRMSVTEGGVQFAETME--GGRPKLGGLECPGHFGHIDLAK
 An   Celegans.133        QFGILGPEEIKRMSVAH--VEFPEVYE--NGKPKLGGLDCPGHFGHLELAK
 Fu   Spombe.54881        QFGILSPEEIRSMSVAK--IEFPETMDESGQRPRVGGLDCPGHFGHIELAK
 Pl   Athaliana.40        QFGILSPDEIRQMSVIH----VEHSETTEKGKPKVGGLECPGHFGYLELAK
 My   Ddiscoideum.        --------------------------------------ECPGHFGHIELAK
 Rh   Porphyra.316        --------------------------------------ECPGHFGFIELAK
 Kt   Tbrucei.1021        QFEIFKERQIKSYAVCLVEHAKSYANA----ADQSGEAECPGHFGYIELAE
 Kt   Leishmania.7        QFEVFKEAQIKAYAKCIIEHAKSYEHG----QPVRGGIECPGHFGYVELAE
```

(b)

```
taxon
                          ....|....10...|....20...|....30...|....40...|....50
 Fu   Nosema.40928        QFGLFSPEEIRASSVAL--IRYPETLE--NGVPKESGLVCAGHFGHIELVK
 Fu   Aspergillus.        QFGLFSPEEIKRMSVVH--VEYPETMDEQRQRPRTKGLECPGHFGHIELAT
 Fu   Spombe.54891        QFGILSPEEIRSMSVAK--IEFPETMDESGQRPRVGGLDCPGHFGHIELAK
 Ap   Plasmodium.3        ELGVLDPEIIKKISVCE--IVNVDIYK--DGFPREGGLYCPGHFGHIELAK
 An   Cricetulus.2        QFGVLSPDELKRMSVTEGGIKYPETTE--GGRPKLGGLECPGHFGHIELAK
 An   Homo.7434727        QFGVLSPDELKRMSVTEGGIKYPETTE--GGRPKLGGLECPGHFGHIELAK
 An   Drosophila.9        QFGILSPDEIRRMSVTEGGVQFAETME--GGRPKLGGLECPGHFGHIDLAK
 An   Celegans.133        QFGILGPEEIKRMSVAH--VEFPEVYE--NGKPKLGGLDCPGHFGHLELAK
 Pl   Athaliana.40        QFGILSPDEIRQMSVIH--VEHSETTE--KGKPKVGGLECPGHFGYLELAK
 My   Ddiscoideum.        --------------------------------------ECPGHFGHIELAK
 Rh   Porphyra.316        --------------------------------------ECPGHFGFIELAK
 Kt   Tbrucei.1021        QFEIFKERQIKSYAVCL--VEHAKSYA--NAADQSGEAECPGHFGYIELAE
 Kt   Leishmania.7        QFEVFKEAQIKAYAKCI--IEHAKSY--EHGQPVRGGIECPGHFGYVELAE
```

TRENDS in Genetics

- inspect alignment carefully
- decide what should and should not be included in the analysis
- General rule: delete all positions with gaps plus any adjacent, ambiguously aligned positions (i.e. columns in the alignment)
- In case of protein-encoding gene: analysis of DNA or protein?
  - Protein for more distant relationships

Step 3. Trees

- Methods, **two** general categories:
  - **distance-matrix** methods, also known as clustering or algorithmic methods (e.g. UPGMA, neighbour-joining, Fitch–Margoliash);
    - transformation of all sequence information into a distance matrix, which is then analyzed using an algorithm for clustering the taxa. Building a tree with this method is fast but all sequence information is lost in the process
  - **discrete data** methods, also known as tree searching methods (e.g., maximum parsimony (MP), maximum likelihood (ML), or Bayesian).

→distance methods are much faster than discrete data methods

→Discrete data methods are time-consuming because all the sequence information is used for the evaluation of the best phylogenetic tree

Step 3. Trees

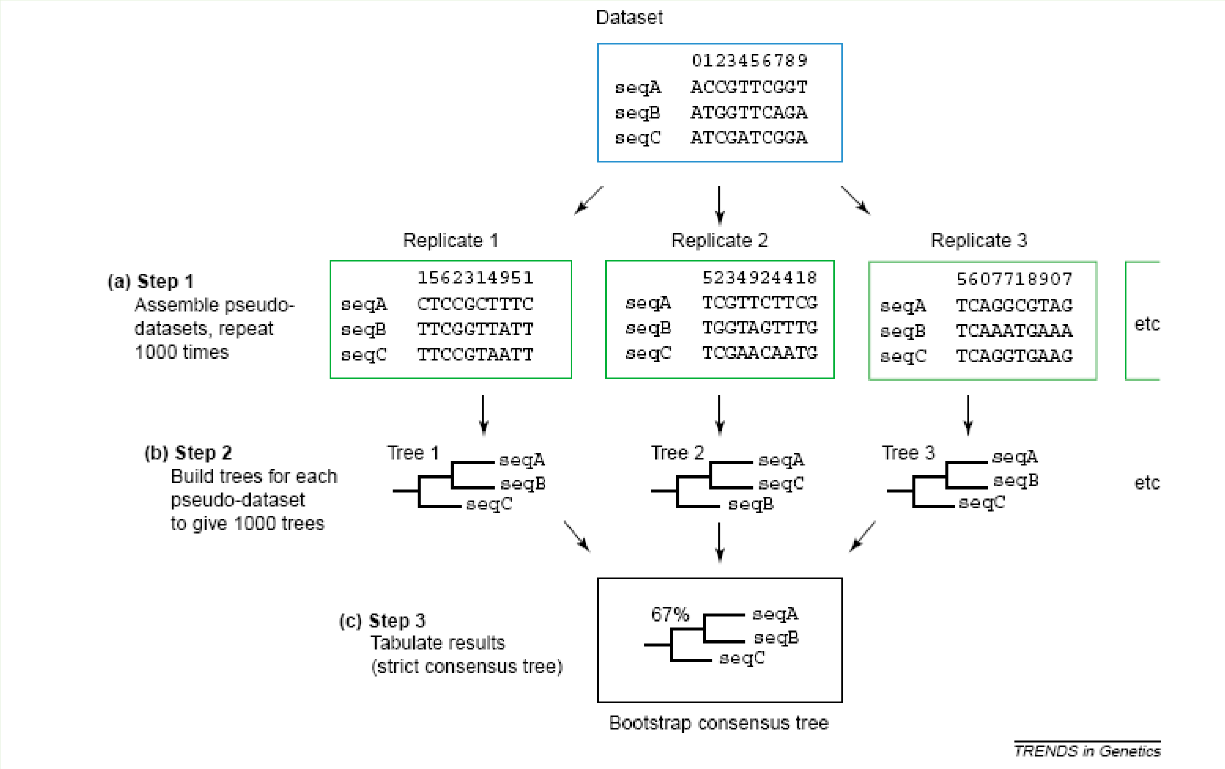
- Distance: relatively simple and straightforward
  - a **single statistic**, the distance (roughly, the percent sequence difference), is calculated for all pairwise combinations of OTUs, and then the distances are assembled into a tree
- Discrete data methods examine each column of the alignment separately and look for the tree that best accommodates all of this information
  - Discrete data analysesare information rich; there is an hypothesis for every column in the alignment, so you can trace the evolution at specific sites in the molecule (e.g. catalytic sites or regulatory regions)
- Models are many and complex either
- Packages (inexpensive or free) for phylogenetic analysis are PHYLIP, Mega and PAUP*, implementing a variety of models and methods
- MrBayes, PhyloBayes and BEAST for Bayesian phylogeny

Step 4. Tests – the bootstrap

- Bootstrapping: so how good is the tree?
- The simplest test of phylogenetic accuracy is the **bootstrap**
- Bootstrapping tests whether your whole dataset is supporting your tree, or if the tree is just a marginal winner among many nearly equal alternatives

Boostrap analysis

1. The dataset is randomly sampled with replacement to create multiple pseudo-datasets of the same size as the original
2. Individual trees are constructed from each of the pseudo-datasets
3. Each of the pseudo-dataset trees are scored for which nodes (groupings) appear and how often



Dataset

```
      0123456789
seqA  ACCGTTCGGT
seqB  ATGGTTCAGA
seqC  ATCGATCGGA
```

(a) **Step 1**
Assemble pseudo-datasets, repeat 1000 times

Replicate 1

```
      1562314951
seqA  CTCCGCTTTC
seqB  TTCGGTTATT
seqC  TTCCGTAATT
```

Replicate 2

```
      5234924418
seqA  TCGTTCTTCG
seqB  TGGTAGTTTG
seqC  TCGAACAATG
```

Replicate 3

```
      5607718907
seqA  TCAGGCGTAG
seqB  TCAAATGAAA
seqC  TCAGGTGAAG
```

etc

(b) **Step 2**
Build trees for each pseudo-dataset to give 1000 trees

(c) **Step 3**
Tabulate results (strict consensus tree)

Bootstrap consensus tree

TRENDS in Genetics

In this case, a node uniting seqA plus seqB is found in two of the three replicate trees, this gives a bootstrap support for this grouping of 2/3 or 67%

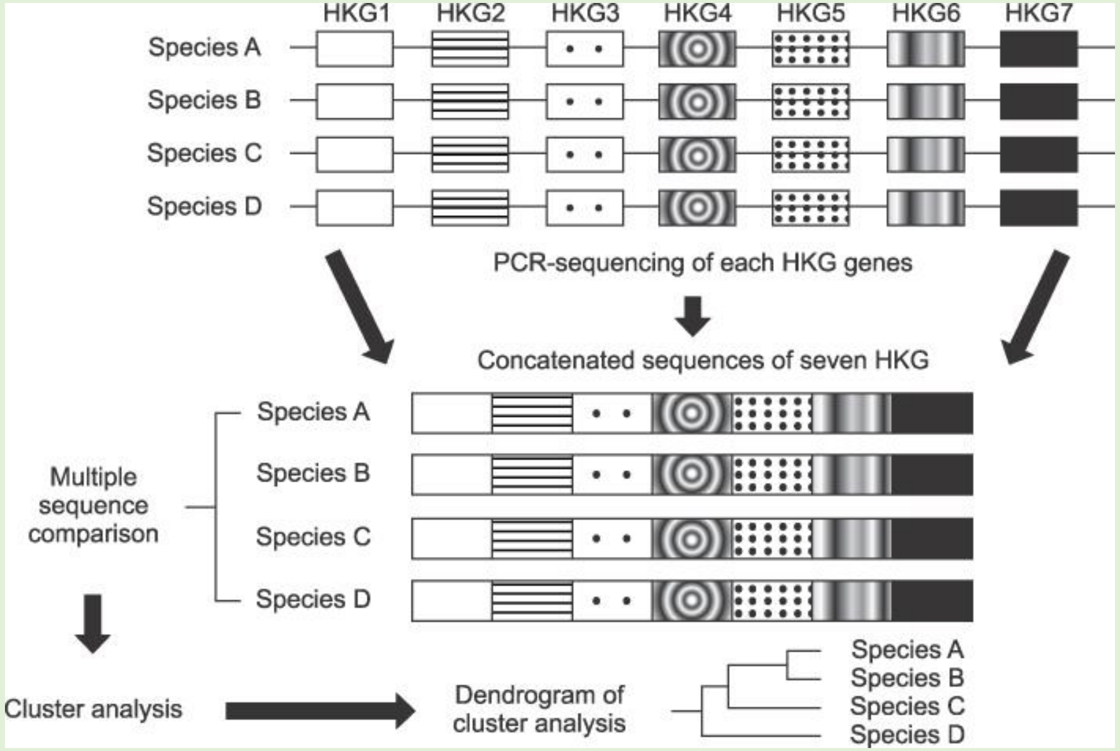
Step 5. Data presentation

- Branch lengths are almost always drawn to scale: that is, proportional to the amount of evolution estimated to have occurred along them.
- Lengths still give a good general impression of relative rates of change across a tree.
- Bootstrap values should be displayed as percentages, not raw values: this makes the tree easier to read and to compare with other trees.
- By convention, only bootstrap values of 50% or higher are reported; lower values mean that the node in question was found in less than half of the bootstrap replicates.

Issues

- Long branches
  - The most problematic and pervasive problem in molecular phylogeny
  - the ‘**long branch attraction**’ is the tendency of highly divergent sequences (i.e. those with long terminal branches) to group together in a tree *regardless of their true relationships*
- Sampling/over- or under-representation of some taxa, might impact of tree reconstruction

Multi Locus Sequence Analysis



Kim & Jang, https://doi.org/10.5145/KJCM.2012.15.3.79

- Be careful with alignments and sequence frames!
- Usually 5-7 genes
- At least 30 genes for genome-level comparison (Chun et al., 2018)