NamesforLife Semantic Resolution Services for the Life Sciences

Charles T. Parker[1], Dorothea Rohlfs[1], Kara Mannor[1], Sarah Wigley[1], Nicole Osier[1], Catherine Lyons[1], George M. Garrity[1,2]*(garrity@namesforlife.com)

[1]NamesforLife, LLC, East Lansing, MI and [2]Michigan State University, East Lansing, MI

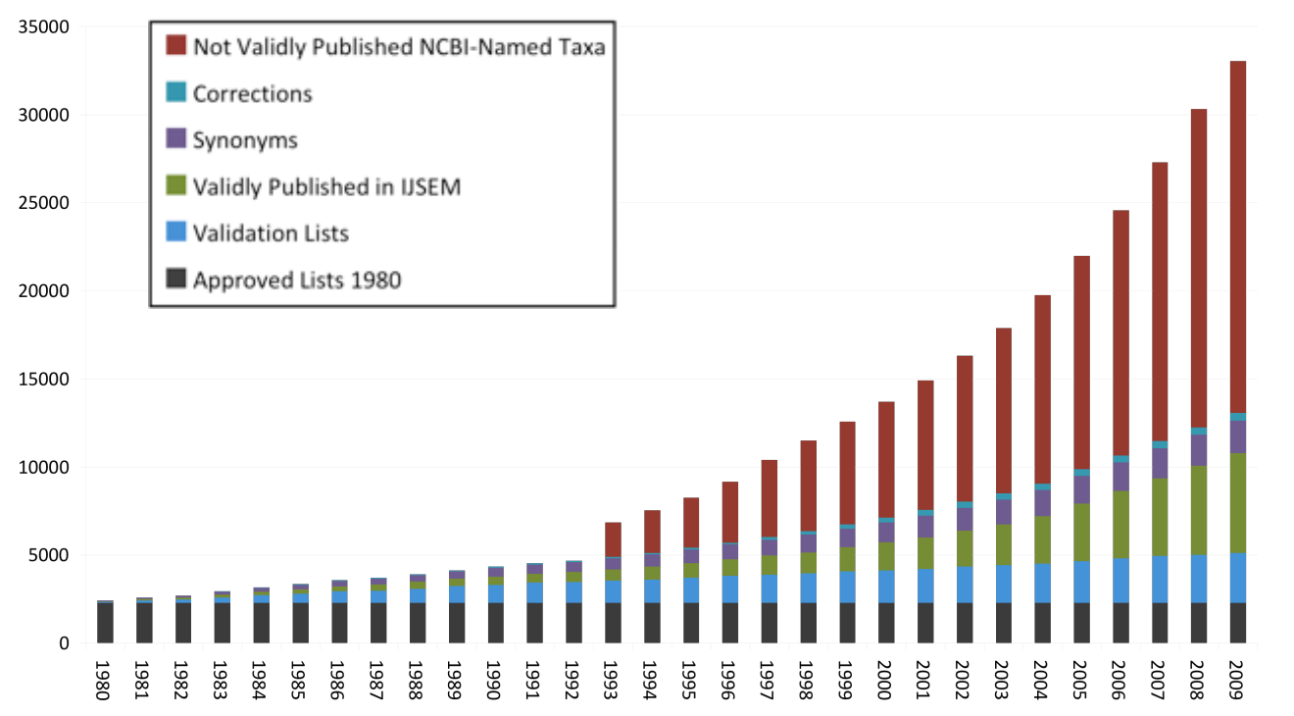
Abstract

Within the Genomes-to-Life Roadmap, the DOE states that a significant barrier to effective communication in the life sciences is a lack of standardized semantics that accurately describe data objects and persistently express knowledge change over time. As research methods and biological concepts evolve, certainty about correct interpretation of prior data and published results decreases because both become overloaded with synonymous and polysemous terms. Ambiguity in rapidly evolving terminology is a common and chronic problem in science and technology. N4L is a novel technology designed to solve this problem.

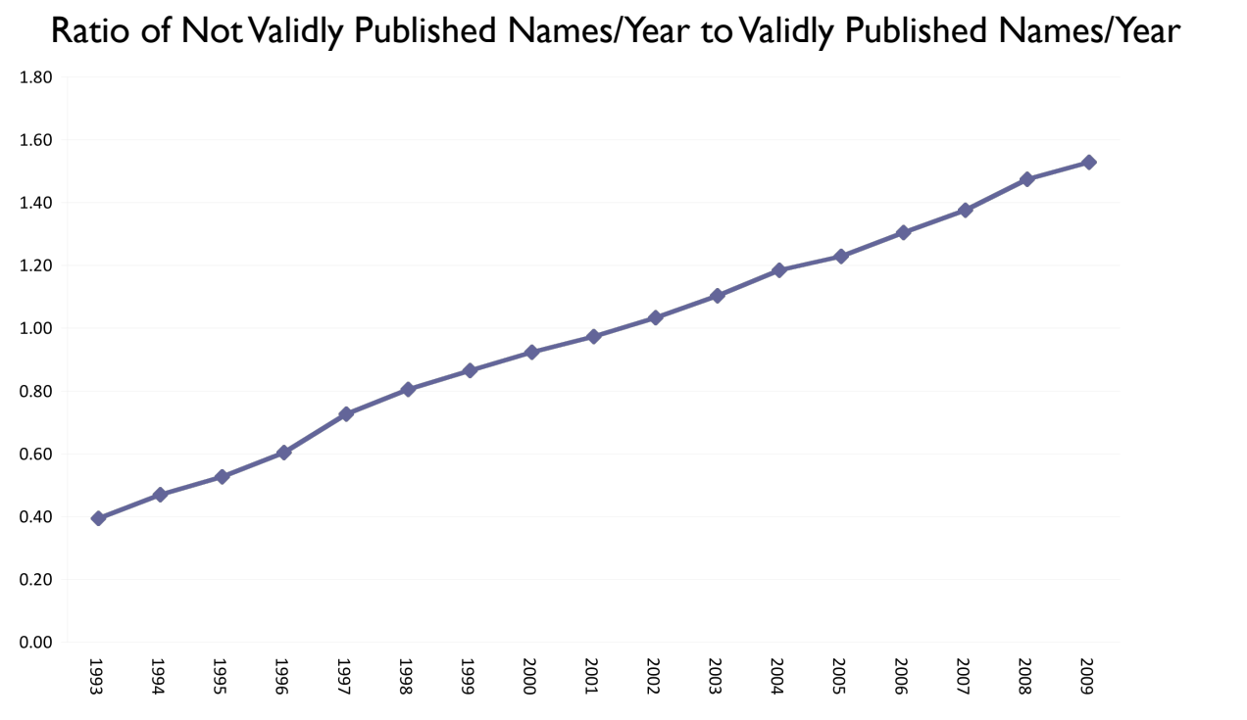
Background

Understanding the correct meaning of a biological name in the appropriate context is important. It is not a trivial task, and the number of individuals with expertise in biological nomenclature is limited. Such knowledge can, however, be accurately modeled and delivered through a networked semantic resolution service that provides end-users of biological nomenclature or other dynamic terminologies with the appropriate information, in proper context, on demand. Such a service can also be used by database owners, publishers, or other information providers to semantically enable their offerings, making them more readily discoverable, even when the definition of a name or term has changed.

The figure below shows the cumulative growth in the number of validly published names since 1980. At the beginning of 1Q 2010, there were 13,069 validly published names (subspecies through classes) and 1,854 synonyms for which there were publicly available type strains. As seen in the chart below, the rate at which new validly published names appear in the literature plateaued in 2006, at around 800 names per year (excluding taxonomic rearrangements and emendations). GenBank began cataloging taxonomic information for sequences derived from Bacteria and Archaea in 1993, and introduced the practice of tagging data with trivial or invalid names. By 2002, the number of invalid names associated with GenBank records exceeded the number of validly published names. At present, invalidly published names outnumber validly published names (3:2). For both tracking and identification purposes, this adds significantly to the burden of those who must rely on this information for various purposes by adding an additional 4.8 names/day, none of which are governed by any rules or principles for assignment, usage, or uniqueness.



Ratio of Not Validly Published Names/Year to Validly Published Names/Year

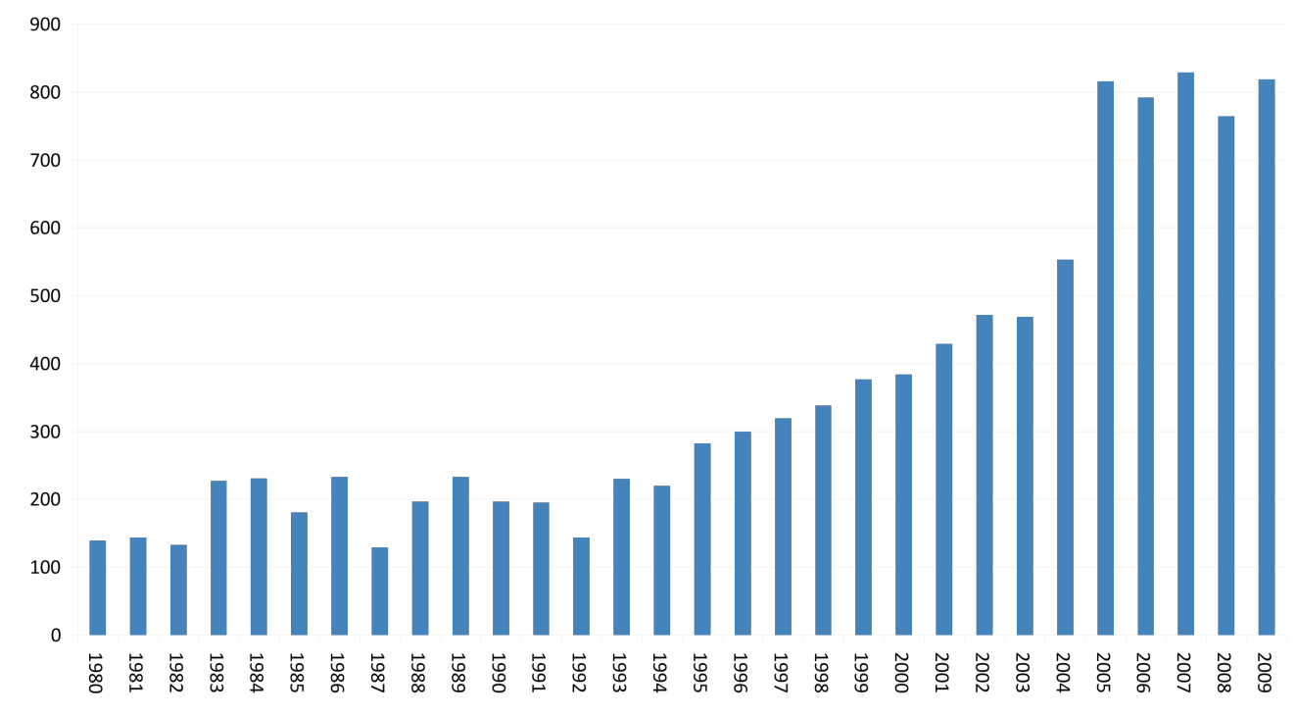


The chart above shows that the ratio of not validly published names to validly published names has been increasing at a linear rate since 1993.

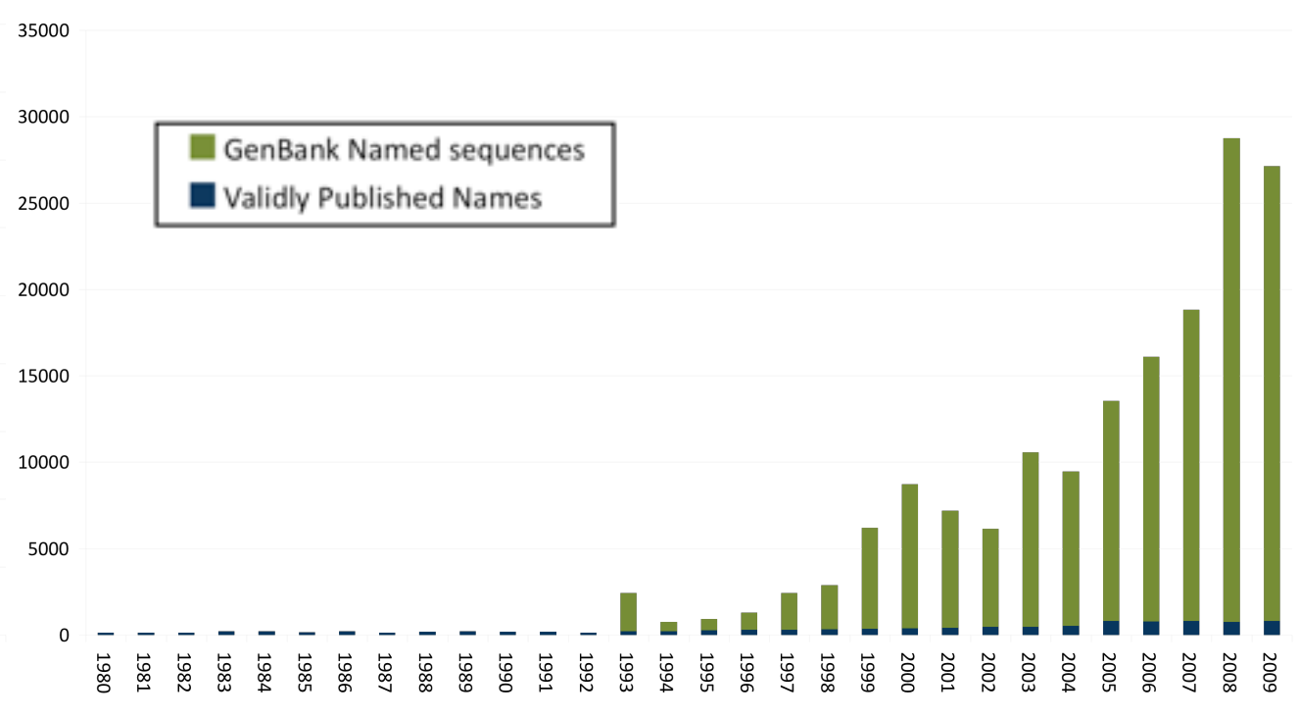
A looming problem

The adoption of DNA sequencing as the preferred method of rapidly characterizing *Bacteria* and *Archaea* has tremendously accelerated during the past five years, with the expected consequences. At present, the rate at which "named" sequences are added to the GenBank taxonomy exceeds the rate at which validly published names appear in the taxonomic record by a factor of approximately 35. This confounds the retrieval of related information from various databases and the scientific, technical and medical literature as many of these invalidly named species can not be readily tracked over time, nor can relationships be inferred to those species for which at least one genome sequence is available. This disconnect between the knowledge contained in the literature and the accumulated genomic data is likely to grow as faster and cheaper sequencing methods come into the market place.

Validly Published Names



Validly Published Names and GenBank "named" Sequences



Database Statistics

The NamesforLife Nomenclature Database (N4LDB) contains 13,534 distinct names, 13,117 of which are validly published, 111 Candidatus, and 306 that are illegitimate but relevant to Bacterial Nomenclature. N4LDB also contains 10,293 exemplars, 9,005 of which represent distinct type strains for 10,558 taxa and 10,942 names. The remaining 2,592 names are associated with higher taxa that do not have exemplars. The breakdown is shown in Table 1 to the right.

Table 2 shows a breakdown of the nomenclatural events by event type since the publication of the Approved Lists in 1980. Some of the less common events (Judicial Opinions, etc) are not shown here.

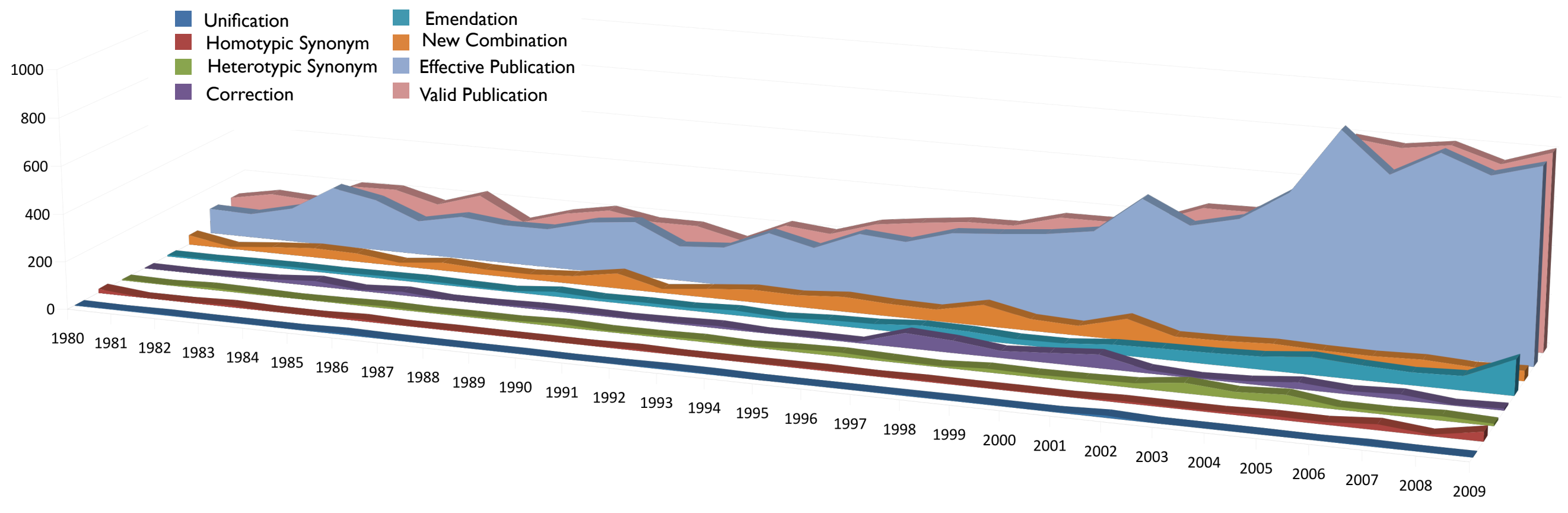
Table 1: N4LDB Records by Rank

Rank	Taxa	Names
Domain	2	2
Phylum	34	35
Class	62	63
Subclass	5	5
Order	119	123
Suborder	20	21
Family	317	322
Subfamily	1	1
Genus	1965	1996
Subgenus	5	5
Species	10058	10403
Subspecies	519	558
Total	**13107**	**13534**

Table 2: Nomenclatural Events Recorded in N4LDB

Event	Count
Corrections	431
New Combinations	1243
Heterotypic Synonyms	343
Homotypic Synonyms	171
Unifications	46
Automatically created names via rule 40d	65
Emendations	1104
Validation List events	2841
Valid Publication	10372

The graph below shows nomenclatural activity from the Approved Lists through the end of 2009. Clearly visible is the slight lag between effective and valid publication, as well as periods of higher activity for taxonomic rearrangements and emendations. The total number of references in the N4L database describing the events is 9,812.



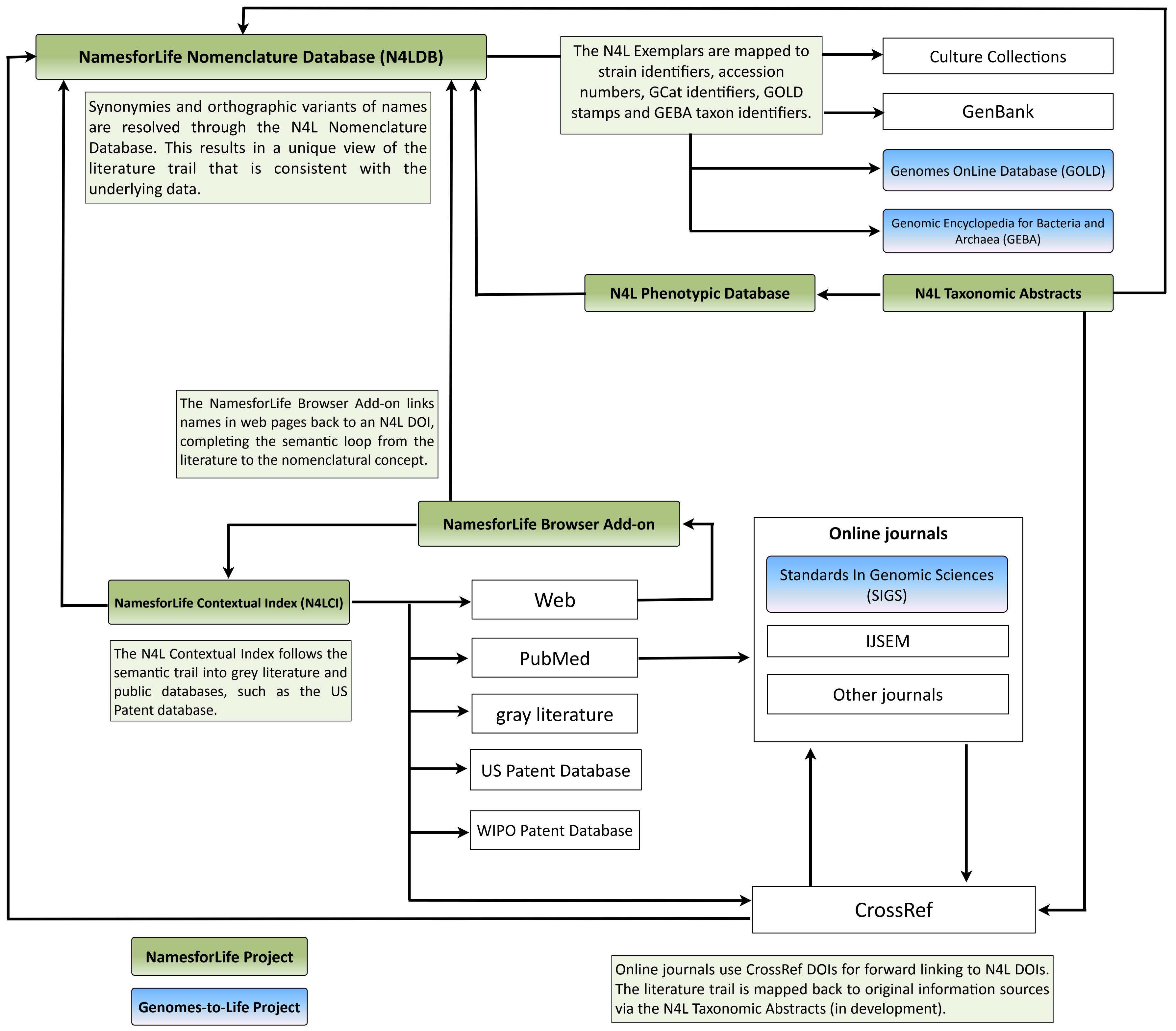
N4L Technology

A full-text literature search can quickly reach a dead end if the terminology has changed. N4L, the technology developed by NamesforLife, resolves changes in Bacterial Nomenclature via a fully referenced model of the synonymies and corrections (see nomenclature example below).

The power of the N4L model is that it does not aggregate data, but rather metadata; it links existing unique identifiers of literature and data together through the terminology itself (a bacterial name or strain), via a single persistent identifier: the N4L Digital Object Identifier (DOI). As shown in the figure below, several once-disparate information sources are now linked together through terminology in a way that is traversable.

Now that the Bacterial Nomenclature database is complete and updated in synchrony with the valid publication of nomenclatural changes, NamesforLife is in the process of linking together Bacterial Nomenclature, technical literature, and the various projects of the Genomes-to-Life program. In N4L, each individual organism is represented by a metadata object (an N4L Exemplar), which is identified by a DOI.

An N4L Exemplar aggregates what is known about an individual organism. The Genomes OnLine Database (GOLD), Standards in Genomic Sciences (SIGS), Genomic Encyclopedia for Bacteria and Archaea (GEBA) and Genomes and Metagenomes Catalogue (GEM) all use unique identifiers that link to each other in some way; via the GCat identifier, GOLD stamp, and GEBA Taxon Identifier. However, there is no single common link to the literature. NamesforLife is closing this gap by tying these disparate sources of information together via N4L Exemplars, which are integrated with the N4L Nomenclature Database and N4L Contextual Index.



NamesforLife Nomenclature Database (N4LDB)

Synonymies and orthographic variants of names are resolved through the N4L Nomenclature Database. This results in a unique view of the literature trail that is consistent with the underlying data.

The N4L Exemplars are mapped to strain identifiers, accession numbers, GCat identifiers, GOLD stamps and GEBA taxon identifiers.

Culture Collections

GenBank

Genomes OnLine Database (GOLD)

Genomic Encyclopedia for Bacteria and Archaea (GEBA)

N4L Phenotypic Database

N4L Taxonomic Abstracts

The NamesforLife Browser Add-on links names in web pages back to an N4L DOI, completing the semantic loop from the literature to the nomenclatural concept.

NamesforLife Browser Add-on

NamesforLife Contextual Index (N4LCI)

The N4L Contextual Index follows the semantic trail into grey literature and public databases, such as the US Patent database.

Web

PubMed

gray literature

US Patent Database

WIPO Patent Database

Online journals

Standards In Genomic Sciences (SIGS)

IJSEM

Other journals

CrossRef

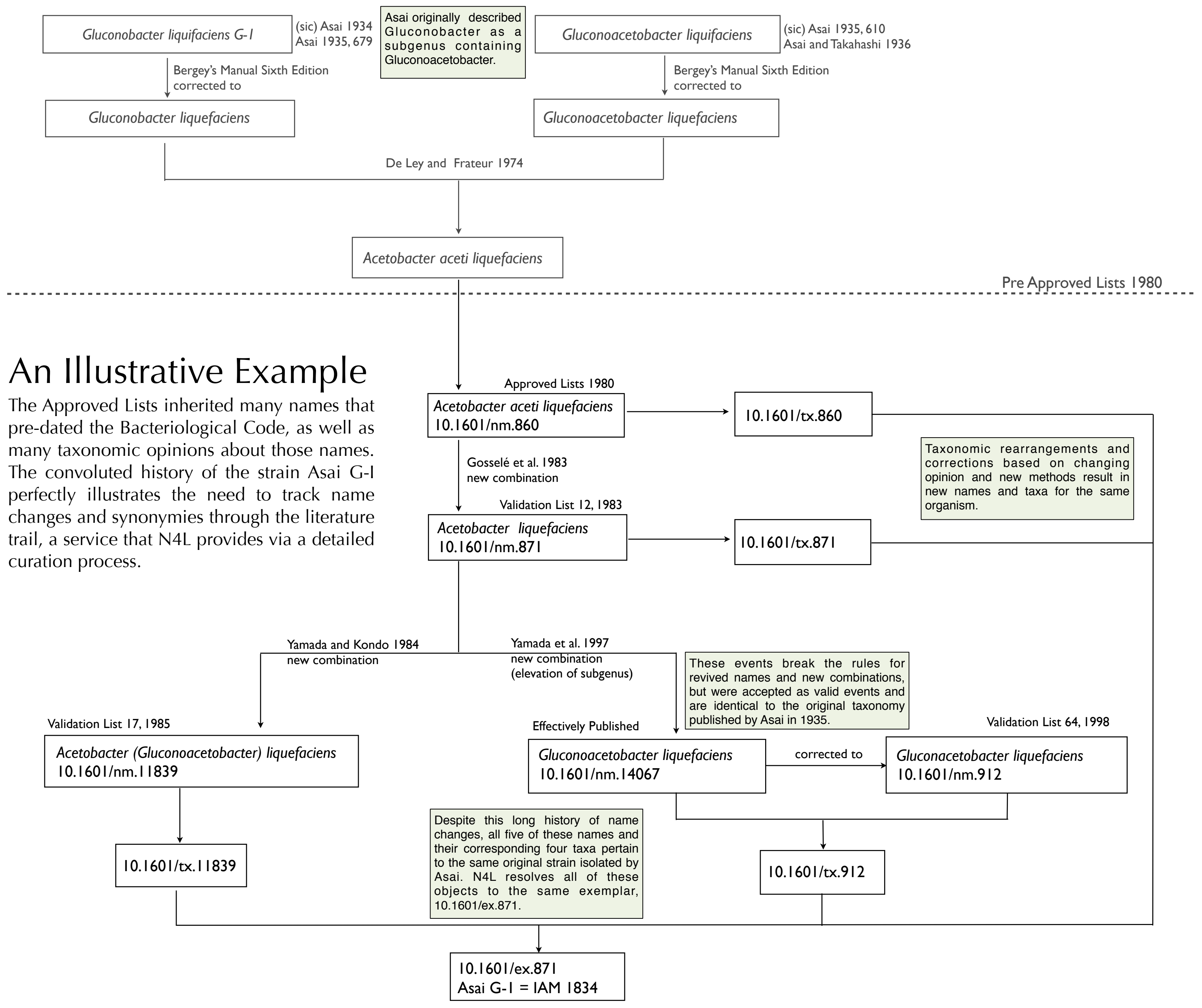
NamesforLife Project

Genomes-to-Life Project

Online journals use CrossRef DOIs for forward linking to N4L DOIs. The literature trail is mapped back to original information sources via the N4L Taxonomic Abstracts (in development).

An Illustrative Example

The Approved Lists inherited many names that pre-dated the Bacteriological Code, as well as many taxonomic opinions about those names. The convoluted history of the strain Asai G-1 perfectly illustrates the need to track name changes and synonymies through the literature trail, a service that N4L provides via a detailed curation process.



Gluconobacter liquifaciens G-1 (sic) Asai 1934 Asai 1935, 679

Asai originally described Gluconobacter as a subgenus containing Gluconoacetobacter.

Gluconoacetobacter liquifaciens (sic) Asai 1935, 610 Asai and Takahashi 1936

Bergey's Manual Sixth Edition corrected to

Gluconobacter liquefaciens

Bergey's Manual Sixth Edition corrected to

Gluconoacetobacter liquefaciens

De Ley and Frateur 1974

Acetobacter aceti liquefaciens

Pre Approved Lists 1980

Approved Lists 1980

Acetobacter aceti liquefaciens 10.1601/nm.860

10.1601/tx.860

Gosselé et al. 1983 new combination

Validation List 12, 1983

Acetobacter liquefaciens 10.1601/nm.871

10.1601/tx.871

Taxonomic rearrangements and corrections based on changing opinion and new methods result in new names and taxa for the same organism.

Yamada and Kondo 1984 new combination

Yamada et al. 1997 new combination (elevation of subgenus)

These events break the rules for revived names and new combinations, but were accepted as valid events and are identical to the original taxonomy published by Asai in 1935.

Validation List 17, 1985

Acetobacter (Gluconoacetobacter) liquefaciens 10.1601/nm.11839

Effectively Published

Gluconoacetobacter liquefaciens 10.1601/nm.14067

corrected to

Validation List 64, 1998

Gluconacetobacter liquefaciens 10.1601/nm.912

10.1601/tx.11839

Despite this long history of name changes, all five of these names and their corresponding four taxa pertain to the same original strain isolated by Asai. N4L resolves all of these objects to the same exemplar, 10.1601/ex.871.

10.1601/tx.912

10.1601/ex.871
Asai G-1 = IAM 1834

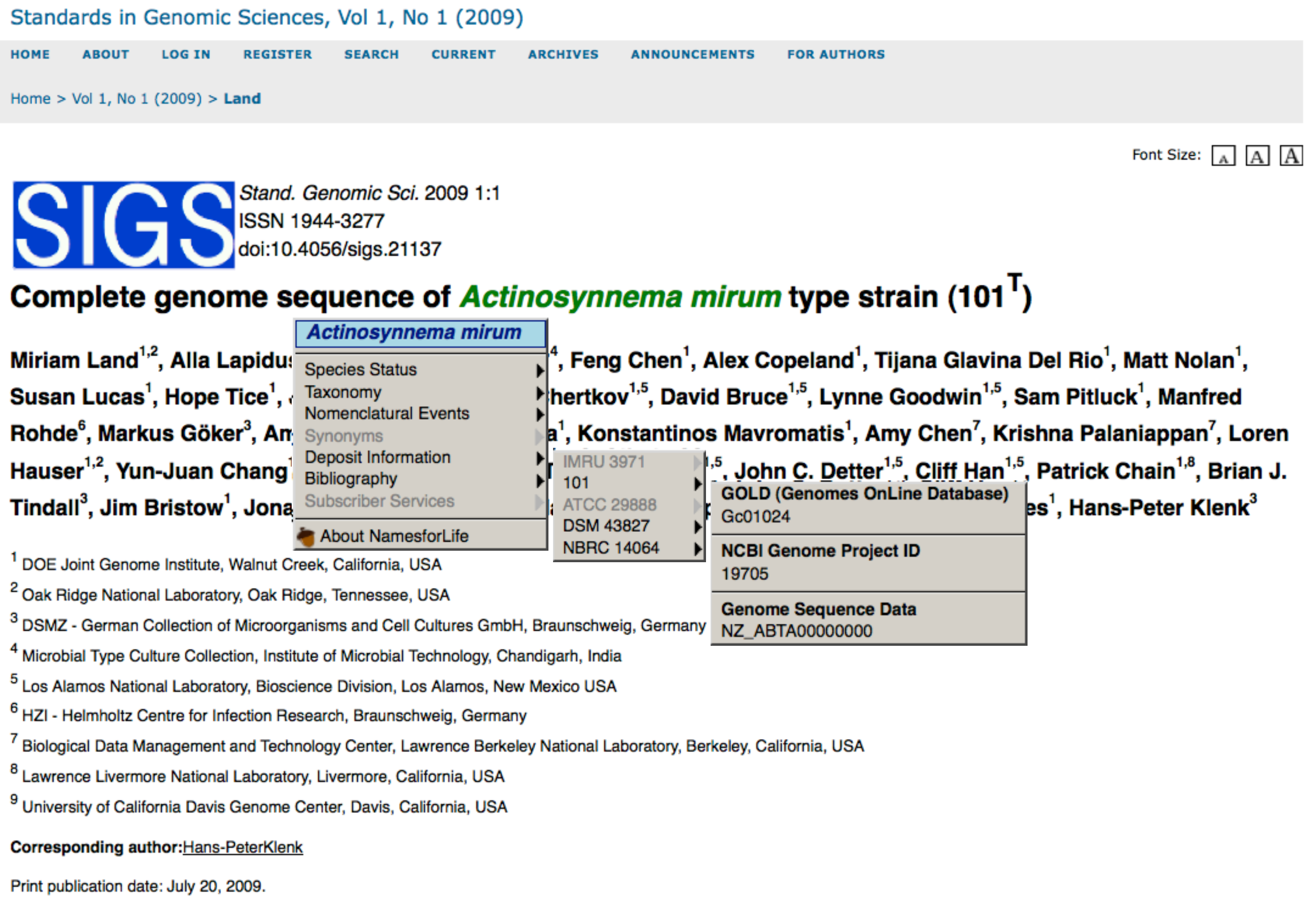
Current Work

The NamesforLife curators are currently reviewing references for the last groups of type exemplars and finishing the links into the Genomes-to-Life projects, including non-type genomes. We expect this work to be completed by April of this year. The N4L Nomenclature Database will continue to be updated as nomenclatural events occur.

The development of the N4L Contextual Index is nearing completion. The first component of this system will come online this spring and will create a semantic path into US Patents and Patent Applications. Additional resources will become available via the Contextual Index throughout the year.

The N4L Taxonomic Abstracts are currently in development. These will provide a snapshot of Bacterial Nomenclature in the form of a citable micro-publication, and will serve to link existing literature to nomenclature via CrossRef.

The Beta release of the N4L Browser Add-on is officially scheduled to coincide with the Society for General Microbiology conference at the end of March 2010, but it is already available for early testing. Instructions on installation and use can be found at the NamesforLife services website, https://services.namesforlife.com. This Firefox Add-on detects and links bacterial names to the N4LDB, providing up-to-date nomenclature, strain and genome information, and a full bibliography. The screenshots below demonstrate the use of this tool on a Standards in Genomic Sciences article.



Standards in Genomic Sciences, Vol 1, No 1 (2009)

HOME ABOUT LOG IN REGISTER SEARCH CURRENT ARCHIVES ANNOUNCEMENTS FOR AUTHORS
Home > Vol 1, No 1 (2009) > Land
Font Size: A A A
SIGS

Stand. Genomic Sci. 2009 1:1
ISSN 1944-3277
doi:10.4056/sigs.21137

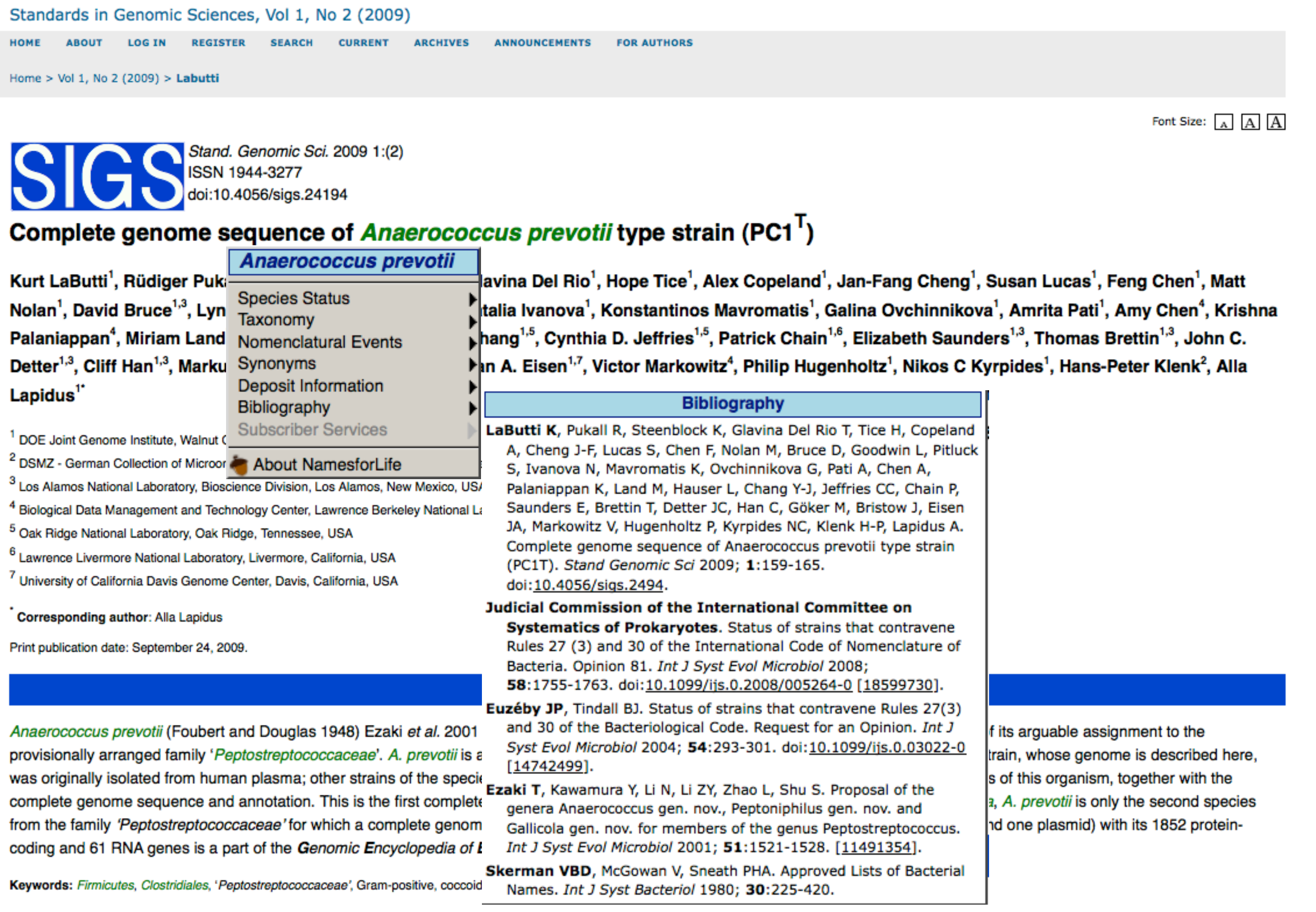
Complete genome sequence of *Actinosynnema mirum* type strain (101[T])

Miriam Land[1,2], Alla Lapidu… [4], Feng Chen[1], Alex Copeland[1], Tijana Glavina Del Rio[1], Matt Nolan[1], Susan Lucas[1], Hope Tice[1], … hertkov[1,5], David Bruce[1,5], Lynne Goodwin[1,5], Sam Pitluck[1], Manfred Rohde[6], Markus Göker[3], Am… a[1], Konstantinos Mavromatis[1], Amy Chen[7], Krishna Palaniappan[7], Loren Hauser[1,2], Yun-Juan Chang … [1,5], John C. Detter[1,5], Cliff Han[1,5], Patrick Chain[1,8], Brian J. Tindall[3], Jim Bristow[1], Jona… es[1], Hans-Peter Klenk[3]

[1] DOE Joint Genome Institute, Walnut Creek, California, USA
[2] Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA
[3] DSMZ - German Collection of Microorganisms and Cell Cultures GmbH, Braunschweig, Germany
[4] Microbial Type Culture Collection, Institute of Microbial Technology, Chandigarh, India
[5] Los Alamos National Laboratory, Bioscience Division, Los Alamos, New Mexico USA
[6] HZI - Helmholtz Centre for Infection Research, Braunschweig, Germany
[7] Biological Data Management and Technology Center, Lawrence Berkeley National Laboratory, Berkeley, California, USA
[8] Lawrence Livermore National Laboratory, Livermore, California, USA
[9] University of California Davis Genome Center, Davis, California, USA

Corresponding author: Hans-PeterKlenk

Print publication date: July 20, 2009.



Standards in Genomic Sciences, Vol 1, No 2 (2009)

HOME ABOUT LOG IN REGISTER SEARCH CURRENT ARCHIVES ANNOUNCEMENTS FOR AUTHORS
Home > Vol 1, No 2 (2009) > Labutti
Font Size: A A A
SIGS

Stand. Genomic Sci. 2009 1:(2)
ISSN 1944-3277
doi:10.4056/sigs.24194

Complete genome sequence of *Anaerococcus prevotii* type strain (PC1[T])

Kurt LaButti[1], Rüdiger Puk… avina Del Rio[1], Hope Tice[1], Alex Copeland[1], Jan-Fang Cheng[1], Susan Lucas[1], Feng Chen[1], Matt Nolan[1], David Bruce[1,3], Lyn… talia Ivanova[1], Konstantinos Mavromatis[1], Galina Ovchinnikova[1], Amrita Pati[1], Amy Chen[4], Krishna Palaniappan[4], Miriam Land … hang[1,5], Cynthia D. Jeffries[1,5], Patrick Chain[1,6], Elizabeth Saunders[1,3], Thomas Brettin[1,3], John C. Detter[1,3], Cliff Han[1,3], Marku… n A. Eisen[1,7], Victor Markowitz[4], Philip Hugenholtz[1], Nikos C Kyrpides[1], Hans-Peter Klenk[2], Alla Lapidus[1*]

[1] DOE Joint Genome Institute, Walnut C…
[2] DSMZ - German Collection of Microor…
[3] Los Alamos National Laboratory, Bioscience Division, Los Alamos, New Mexico, US…
[4] Biological Data Management and Technology Center, Lawrence Berkeley National L…
[5] Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA
[6] Lawrence Livermore National Laboratory, Livermore, California, USA
[7] University of California Davis Genome Center, Davis, California, USA

[*] **Corresponding author**: Alla Lapidus

Print publication date: September 24, 2009.

LaButti K, Pukall R, Steenblock K, Glavina Del Rio T, Tice H, Copeland A, Cheng J-F, Lucas S, Chen F, Nolan M, Bruce D, Goodwin L, Pitluck S, Ivanova N, Mavromatis K, Ovchinnikova G, Pati A, Chen A, Palaniappan K, Land M, Hauser L, Chang Y-J, Jeffries CC, Chain P, Saunders E, Brettin T, Detter JC, Han C, Göker M, Bristow J, Eisen JA, Markowitz V, Hugenholtz P, Kyrpides NC, Klenk H-P, Lapidus A. Complete genome sequence of Anaerococcus prevotii type strain (PC1T). *Stand Genomic Sci* 2009; **1**:159-165. doi:10.4056/sigs.2494.

Judicial Commission of the International Committee on Systematics of Prokaryotes. Status of strains that contravene Rules 27 (3) and 30 of the International Code of Nomenclature of Bacteria. Opinion 81. *Int J Syst Evol Microbiol* 2008; **58**:1755-1763. doi:10.1099/ijs.0.2008/005264-0 [18599730].

Euzéby JP, Tindall BJ. Status of strains that contravene Rules 27(3) and 30 of the Bacteriological Code. Request for an Opinion. *Int J Syst Evol Microbiol* 2004; **54**:293-301. doi:10.1099/ijs.0.03022-0 [14742499].

Ezaki T, Kawamura Y, Li N, Li ZY, Zhao L, Shu S. Proposal of the genera Anaerococcus gen. nov., Peptoniphilus gen. nov. and Gallicola gen. nov. for members of the genus Peptostreptococcus. *Int J Syst Evol Microbiol* 2001; **51**:1521-1528. [11491354].

Skerman VBD, McGowan V, Sneath PHA. Approved Lists of Bacterial Names. *Int J Syst Bacteriol* 1980; **30**:225-420.

Anaerococcus prevotii (Foubert and Douglas 1948) Ezaki *et al.* 2001 … f its arguable assignment to the provisionally arranged family '*Peptostreptococcaceae*'. *A. prevotii* is a… train, whose genome is described here, was originally isolated from human plasma; other strains of the specie… s of this organism, together with the complete genome sequence and annotation. This is the first complete… a, *A. prevotii* is only the second species from the family *'Peptostreptococcaceae'* for which a complete genom… nd one plasmid) with its 1852 protein-coding and 61 RNA genes is a part of the ***Genomic Encyclopedia of B…***

Keywords: *Firmicutes*, *Clostridiales*, '*Peptostreptococcaceae*', Gram-positive, coccoid

Future Work

The Taxonomic Outline of Bacteria and Archaea (TOBA) is an open publication that has served the microbiology community for several years. NamesforLife will continue to support its ongoing publication. Release 8.0 of TOBA is planned following the completion of the current curatorial efforts.

In tandem with the new release of TOBA, NamesforLife will begin to publish the N4L Taxonomic Abstracts, described above. These will leverage the existing citation system developed by CrossRef to complete the semantic link between existing publications as shown in the diagram to the left.

An N4L Phenotypic Database is currently in the planning stages, and is intended to fill a gap in the publication of phenotypic data. A prototype of this database will be available in the summer of 2010 and will be incorporated into the N4L Taxonomic Abstracts. The goal is to provide a source of fully-referenced MIGS-compliant data for bacterial strains based on historical literature.

Additional tools and delivery methods for N4L content are also planned, including authoring and research tools, and alert services. New services will be announced as they become available.

Acknowledgments

We wish to thank B.J. Tindall (DSMZ, Braunschweig) and J. Euzéby (École Nationale Vétérinaire de Toulouse) for their helpful discussions regarding problematic nomenclature issues, R. Dunford and R. Fraser (SGM, Reading) for their assistance in making IJSEM content available for this study and their time, energy and willingness to make this project a reality. We would also like to thank members of the International Commission on Prokaryotic Nomenclature for their support of these efforts, and Matt Winters, Denise Searles, Austin Kuo, Julia Bell, Judy Leventhal and Sheena Tapo for their assistance in curating the underlying taxonomic and nomenclatural information used in our models. This research is supported by the Office of Biological and Environmental Research of the U.S. Department of Energy under Phase II STTR Award DE-FG02-07ER86321 A001.