Chapter 8

Dealing with the Data Deluge – New Strategies in Prokaryotic Genome Analysis

Leonid Zaslavsky, Stacy Ciufo, Boris Fedorov, Boris Kiryutin, Igor Tolstoy and Tatiana Tatusova

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/62125

Abstract

Recent technological innovations have ignited an explosion in microbial genome sequencing that has fundamentally changed our understanding of biology of microbes and profoundly impacted public health policy. This huge increase in DNA sequence data presents new challenges for the annotation, analysis, and visualization bioinformatics tools. New strategies have been designed to bring an order to this genome sequence shockwave and improve the usability of associated data. Genomes are organized in a hierarchical distance tree using single-copy ribosomal protein marker distances for distance calculation. Protein distance measures dissimilarity between markers of the same type and the subsequent genomic distance averages over the majority of marker-distances, ignoring the outliers. More than 30,000 genomes from public archives have been organized in a marker distance tree resulting in 6,438 species-level clades representing 7,597 taxonomic species. This computational infrastructure provides a foundation for prokaryotic gene and genome analysis, allowing easy access to pre-calculated genome groups at various distance levels. One of the most challenging problems in the current data deluge is the presentation of the relevant data at an appropriate resolution for each application, eliminating data redundancy but keeping biologically interesting variations.

Keywords: Genome analysis, clusters, proteins, bacteria, prokaryotes

1. Introduction

Prokaryotes are probably the largest and the most diverse group of cellular organisms.

The number of described species is now about 12,000, and the number of species on earth is estimated in the millions [1]. Recent rapid advances in sequencing technologies provided a

relatively cheap and fast way of studying the diversity of microbial species by discovering representatives of novel divisions or even phyla [2] and analyzing the variation within the species by sequencing closely related genomes from the ecological microbial populations or clinical studies of pathogenic bacteria.

Historically, prokaryotic organisms were organized by classical taxonomic ranking system (species, genus, family, order, and phylum). Delineation of prokaryotic species was originally based on phenotypic information, pathogenicity, and environmental observations. Due to the high mutation level, fast replication rate, and efficient DNA exchange mechanisms, microbial organisms can easily adapt to their habitats. Genomic studies have shown that different species living in similar ecological environments show similarity at the genomic level (e.g., congruent evolution of water-living bacteria from various taxonomic origins [3]) while same pathogenic species (or symbionts) rapidly adapting to the new hosts become quite different at genomic level (e.g., *Buchnera aphidicola* [4], *Serratia symbiotica* [5]).

Next-generation sequencing technologies provide new insights into the life of microbes and their interactions with the host, but they do not classify the organisms in a traditional way. Many novel species are described as "candidatus" or "<genus> sp."

The genomes of uncharacterized isolates of the Candidatus Arthromitus, host-specific intestinal symbionts, comprise a distinct clade within the Clostridiaceae [6].

http://www.ncbi.nlm.nih.gov/genome/13597

The number of uncharacterized species is rapidly growing in public genome collections. As of November 2014, almost half of bacterial and archaeal species in NCBI Refseq data set remain uncharacterized. (Bacteria: 3,559 uncharacterized, 7,597 total; Archaea: 162 uncharacterized, 399 total.)

The need for different approaches to the identification of microbial species that can take into account the advantages of the growing massive volume of genomic sequence data is being actively discussed in the research community.

Scientists from different disciplines (taxonomists, ecologists, and evolutionary biologists) have different interpretations of species defined by the framework of their needs and the tools they use for identification. A recent review [7] describes the history and present state of various methods of description of prokaryotic species. The authors suggest the concept of species as "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."

Comparative analysis requires a target: a coherent group of isolates with some degree of similarity defined by the goal of the study (the analysis of pathogen outbreak performed at the species level or below, while biodiversity studies use broader group such as families or phyla). Several groups have attempted to delineate the taxonomy of Archaea and Bacteria using the methods based on single-copy universally conserved markers [8-13]. Other methods are discussed in recent reviews [14].

http://dx.doi.org/10.5772/62125

Different species vary dramatically in terms of the sampling density and data quality. Clinical and epidemiological studies produce large data sets of closely related (clonal) genomes (Table 1), while other species are sampled very coarsely. Genomic and proteomic structure of a densely sampled group of related strains is commonly described by the concept of pan-genome [15] species.

The complexity of the data is challenging to the analysis, representation, and visualization of the data sets. One of the challenges is the amount of the resources required for a brute-force processing approach (e.g., BLAST all-to-all of 35 million proteins will take five days on 1,000 processors). Another big problem is data heterogeneity and redundancy: the closest-neighbor results will often contain long list of nearly identical objects, making it difficult to identify more distant neighbors.

Here we describe a combined approach that provides a robust, fast, and scalable method of defining the sequence similarity genome groups that can be used for comparative genome analysis and resolve some known issues with the delineation of species in traditional taxonomy.

2. Materials and methods

The genomes are organized in hierarchical groups calculated with different methods. The universal ribosomal markers approach is used to build a distance tree and to define species and superspecies-level clades (genome groups). The species-level clades are further refined by using whole-genome alignments and creating tight (clonal) genome groups.

2.1. NCBI hardware and software

The hardware available at NCBI includes a Univa Grid Engine (UGE) Grid-Engine-based computer farm and PanFS scalable storage system connected through a powerful router. The most recent version UGE 8.2.1 includes support for Linux GROUPs, support for Window server execution nodes, and a beta version of DRMMA2 (Distributed Resource Management and Application API 2). A large weakly coupled distributed computer system like this requires coarse-grained parallelization approaches with minimal communication between the processes. Many processing steps, such as computing BLAST hits, are naturally parallel.

2.2. Data snapshot

A given data snapshot represents a collection of genome (and protein) sequences and metadata available at the time. Navigating through millions of nucleotide sequences in public archives to find a set that comprises a whole-genome collection can be sometimes challenging. GenBank release 207 contains 182,188,746 sequences, and 189,739,230,107 nucleotides. The traditional NCBI sequence repository was designed for GenBank records in the early 1990s. It is organized as a collection of single-nucleotide sequence records with annotated sequences stored as nucleotide–protein sets. By GenBank requirements, each sequence record should be associated

with the organisms registered in the NCBI Taxonomy Database. For the first 10 years of microbial genome sequencing, each species has a unique genome representation in public sequence archives. When sequencing costs decreased, researchers began to explore microbial population structure and the intraspecies differences. NCBI Taxonomy group began assigning Taxonomy ID for strain level nodes as proxies of unique genome identifiers. More recently, next-generation sequencing and rapid pathogen detection approaches have shifted the paradigm from a single isolate representing an organism to multiisolate projects often representing almost identical isolates from the outbreak analysis. These closely related genomes differ by metadata only: patient information, date, and place of sample collection. NCBI has created new resources that capture the sequence data and metadata information: BioProject, BioSample, and Assembly [16]. A triplet of these identifiers uniquely defines a genome with the metadata that can be used for further comparative analysis.

NCBI internal database UniCol is used to store collections of the nucleotide and protein sequence data associated with every BioProject, BioSample, Assembly triplet. The database provides a tracking history for a given snapshot with the sequence assembly and metadata available at the time.

Clade_id	**Name**	**Genomes**	**Clonal groups**	**Taxonomy**
19988	*Staphylococcus aureus*	4182	118	species
19668	*Escherichia, Shigella*	2479	986	multiple
20829	*Mycobacterium tuberculosis*	1844	11	species
19669	*Salmonella*	971	139	genus
19507	*Acinetobacter*	846	306	genus
19252	*Helicobacter pylori*	432	258	species
20104	*Streptococcus*	394	154	genus
19672	*Enterobacter, Klebsiella*	384	149	multiple
20137	*Enterococcus*	354	161	genus
19921	*Brucella*	335	9	genus

Table 1. Calculated clades may include a single species, a single genus, or multiple genera for closely related species.

2.3. Genome quality assessment

There are several criteria that are used to evaluate the quality of genome assembly.

The N50/L50 metrics are automatically calculated for each genome. Acceptable values are dependent on genome size, and genomes which do not meet the criteria are not processed for Refseq. For known clades, the genome size is expected to fall within 2 standard deviations from the mean for clades, which have at least 10 members. This standard allows for the identification of partial genomes and unusually large genomes, which may indicate a bad assembly or contamination.

http://dx.doi.org/10.5772/62125

Some genomes submitted to GenBank represent an assembly from a mixed culture (accession # AKNF01000000 is a mixed culture of *Shigella flexneri* 1235-66 and an unknown *Shigella* species) or a hybrid of different species or a chimera genome (accession # AP012495 chimera genome constructed by cloning the whole genome of *Synechocystis* strain PCC6803 into the *Bacillus subtilis* 168 genome). Partial and "anomalous" assemblies are clearly flagged in NCBI assembly database and not included in clade analysis.

2.4. Marker to genome alignment

Genome distance is defined as an average of pairwise protein distances of universally conserved single-copy proteins as defined in [8] (Table 2).

Genomic markers (E.coli K12 accessions)	Genomic markers
NP_417801	ribosomal protein S12
NP_417800	ribosomal protein S7
NP_414564	ribosomal protein S2
NP_418410	ribosomal protein L11
NP_418411	ribosomal protein L1
NP_417779	ribosomal protein L3
NP_417774	ribosomal protein L22
NP_417773	ribosomal protein S3
NP_417769	ribosomal protein L14
NP_417767	ribosomal protein L5
NP_417765	ribosomal protein S8
NP_417100	ribosomal protein l6p/L9E
NP_417762	ribosomal protein S5
NP_417757	ribosomal protein S13
NP_417756	ribosomal protein S11
NP_417698	ribosomal protein L13
NP_417697	ribosomal protein S9
NP_417634	ribosomal protein S15P/S13E
NP_417770	ribosomal protein S17
NP_417772	ribosomal protein L16/L10E
NP_417760	ribosomal protein L15
NP_417763	ribosomal protein L18
NP_417755	ribosomal protein S4

Table 2. List of genomic markers used in genomic analysis. Escherichia coli K-12 accessions are given as an example. Each marker has a corresponding protein cluster which is used in the analysis.

2.5. Genome distance

Protein marker distances and genomic distance are designed to be robust while remaining appropriately sensitive. Protein distance measuring dissimilarity between markers of the same

type is designed to ignore differences in protein lengths and tuned to measure dissimilarity in internal parts of the sequences. The subsequent genomic distance averages over the majority of marker-distances, ignoring the outliers.

2.5.1. Protein distances

Consider proteins i and j, having the best aggregated BLAST alignment of length L_{ij} with aggregated score S_{ij}. Assume that the proteins have lengths L_i and L_j and self-scores S_{ii} and S_{jj}. Define normalized scores: $s_{ij}=S_{ij}/L_{ij}$, $s_{ii}=S_{ii}/L_i$, $s_{jj}=S_{jj}/L_j$.

Then define protein distances:

$$d_{ij}=1-\min\left(1,\frac{s_{ij}}{\min\left(s_{ii},\ s_{jj}\right)}\right) \tag{1}$$

Distance (1) is an identity-like characteristic calculated from the aggregated BLAST [17] scores (using positives based on BLOSUM62 matrix [22]). For full-length alignment, it can be reduced to $1-\frac{S_{ij}}{\min(S_{ii},\ S_{jj})}$. However, when lengths are different; distance (1) avoids penalizing nonaligned ends of the proteins, taking into account only mutation events.

2.5.2. Genomic distances

Suppose that genomes *i* and *j* have N_{ij}^{a} types of markers found in both genomes, with N_{ij}^{h} of them having acceptable BLAST hits.

Define the offset $\Delta_{ij}=\max\left(3,\ \frac{N_{ij}^{h}}{4},\ 1+N_{ij}^{a}-N_{ij}^{h}\right)$. Order marker distances in the ascending order: $d_{ij}^{(0)}\le d_{ij}^{(1)}\le\ldots\le d_{ij}^{\left(N_{ij}^{h}-1\right)}$. Then robust genomic distance is defined by the formula:

$$D_{ij}=\frac{\sum_{p=\Delta_{ij}}^{p=N_{ij}^{h}-\Delta_{ij}-1}d_{ij}^{(p)}l_{ij}^{(p)}}{\sum_{p=\Delta_{ij}}^{p=N_{ij}^{h}-\Delta_{ij}-1}l_{ij}^{(p)}}, \tag{2}$$

where $l_{ij}^{(p)}$ are corresponding alignment length. The marker-protein distances are weighted by alignment lengths $l_{ij}^{(p)}$ in order to provide where possible results similar to the original method in [8] based on concatenation of proteins. However, the use of offset Δ_{ij} allows filtering out outliers since the averaging in (2) is performed over $N_{ij}^{h}-2\Delta_{ij}$ distances in the middle. For each phylum level group, an agglomerative hierarchical clustering tree is built using the complete linkage clustering algorithm [19, 20].

http://dx.doi.org/10.5772/62125

2.6. Genome clustering pipeline

The pipeline for calculating genome clades consists of three major components (see Figure 1). The first is the collection of the input data from NCBI main sequence repositories. The genomic data are dynamic: hundreds of new genomes and assembly updates are submitted to NCBI each day. We create a snapshot of all live genome assemblies and their nucleotide sequence components (chromosomes, scaffolds, and contigs) and store them in an internal relational database: UniCol. The genome data set is organized into large groups (phyla and superphyla defined by NCBI Taxonomy). The assemblies are then filtered by quality and passed to the processing script. Ribosomal protein markers are predicted in every genome to overcome problems with the genome annotations (missing and/or incorrect annotations) and to normalize markers' data set. Marker predictions are performed by aligning reference protein markers against full genome assemblies. Assemblies with at least 17 markers are passed to the next step. Genome distance is calculated as an average of pairwise protein distances of markers shared in a pair of genomes. Finally, agglomerative hierarchical clustering trees are built within phylum-level groups. Clades at the species level are calculated using species-aware algorithm. Superclade trees are constructed by sectioning the trees at the distance of 0.25.

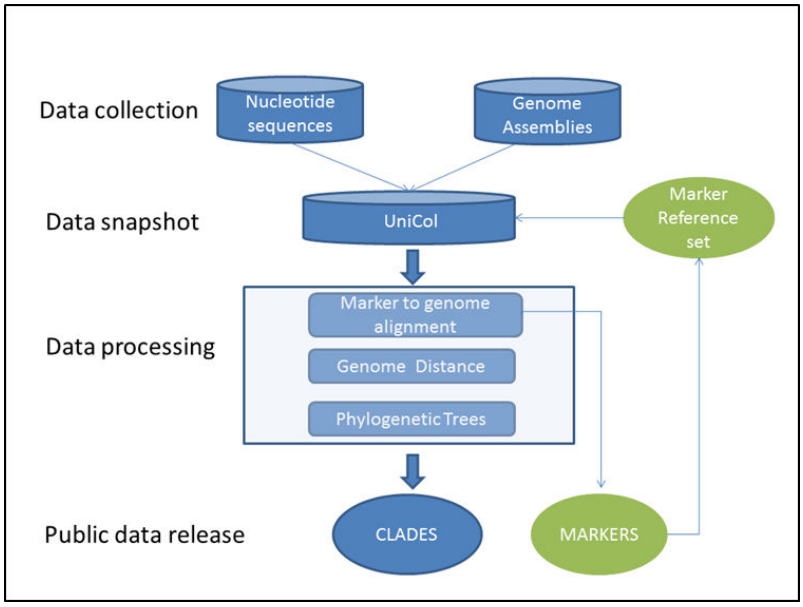

Figure 1. Dataflow of ribosomal-marker-based clade (genome group) processing. Ribosomal markers (in green) are maintained outside of the main pipeline (in blue). Clades and markers are available on NCBI FTP site: ftp://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/CLADES/ ftp://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/MARKERS/

2.7. Clades and superclades

Due to biological, historical, and sampling reasons, microbial organisms have very different levels of strain variation within species. Using the genome data available in public archives we have calculated the diameter of the species defined by NCBI Taxonomy (see Figure 2).

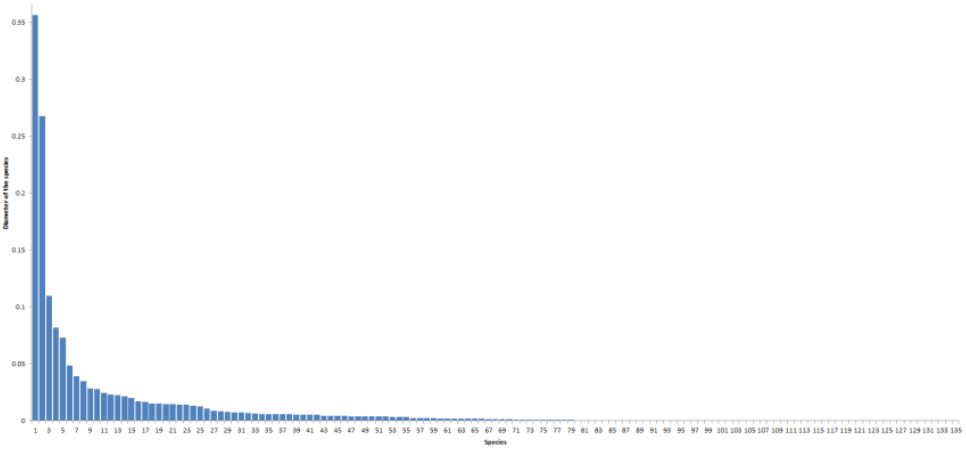

Figure 2. Distribution of Taxonomy-defined species diameter. Y axes – diameter of species, X axes – species numbered in the descending diameter order.

Instead of using one fixed threshold, we utilize a taxonomy-aware algorithm that allows increasing the size of a genomic group in certain circumstances. Two distance threshold, the lower threshold d_lower and the upper threshold d_upper, are established (currently, we use values d_lower = 0.015 and d_upper = 0.025). Genomes with the lowest common ancestor with height d_lower or below are always in the same group, while genomes with the lowest common ancestor with height above d_upper are never placed together. In between d_lower and d_upper, taxonomic information is used: two subgroups are merged in a larger group if any pair of species in a group is already together in one of two subgroups (i.e., there are no new merges of species). Species are defined according to the NCBI taxonomic records [16].

Phylum-level trees are not practical for presentation and evaluation of closely related genomes. However, it is important to see the relationships (distance) between close clades (see Figure 3).

2.8. Genome groups

Species-level clades are further refined by whole-genome alignments using megablast with default parameters [18]. The genome groups are defined by clustering the genomes at 95% identity and 90% coverage. An example of genome groups for *Klebsiella pneumonia* clade is shown in Figure 4. For each group a representative genome with the highest level of assembly and annotation quality is selected.

http://dx.doi.org/10.5772/62125

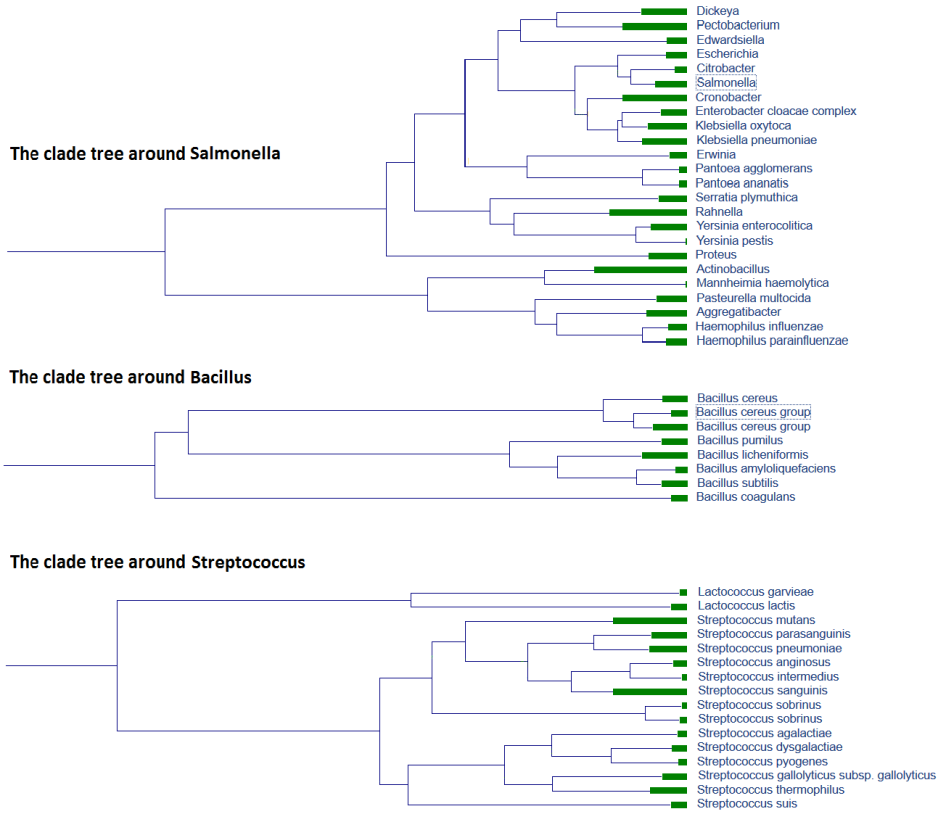

Figure 3. Superclade tree for three abundant groups: A – Salmonella, B – Bacillus, C – Streptococcus. Green boxes represent clades; box size is proportional to the number of genome in a clade.

3. Results and discussion

Large clades obtain additional members in each subsequent snapshot (see Figure 5). The process assigns related genomes to the same clade consistently. There is also a large growth in singleton clades, reflecting an increasing interest in sequencing taxonomically distinct organisms.

We have developed an infrastructure for grouping all whole-genome sequence assemblies at various proximity levels. By using universally conserved ribosomal genes we define the species-level groups. We propose a set of 23 single-copy marker gene families that have consistent evolutionary histories. The proposed ribosomal protein-marker distance and genomic distance are tailored to achieve robustness, while remaining appropriately sensitive.

The major objective of our approach is to generate and actively maintain the target sets for pan-genome analysis. These ribosomal-marker-based groups (clades) roughly correspond to

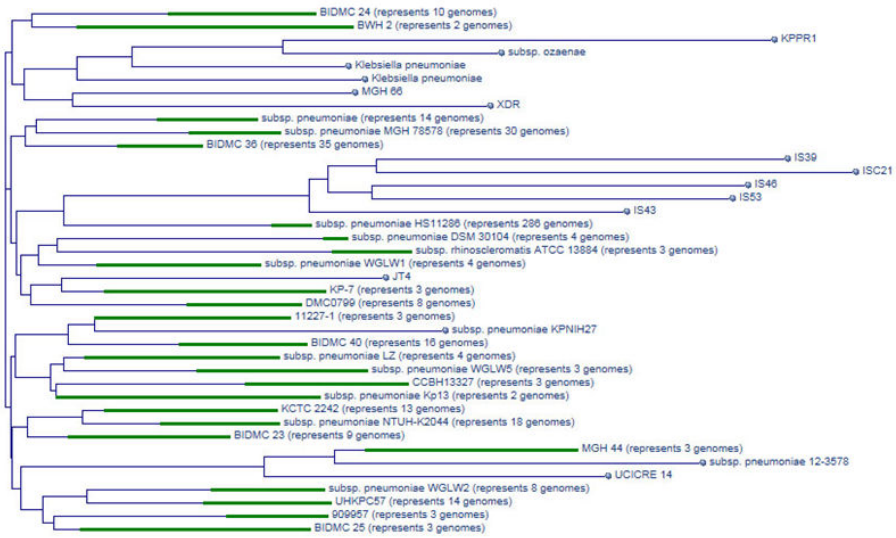

Figure 4. *Klebsiella pneumonia* clade contains 534 full genome assemblies organized in 25 closely related genomic groups. Blue circles at the end of the branch represent a single genome; green boxes represent a group of genomes with the box size proportional to the number of genomes.

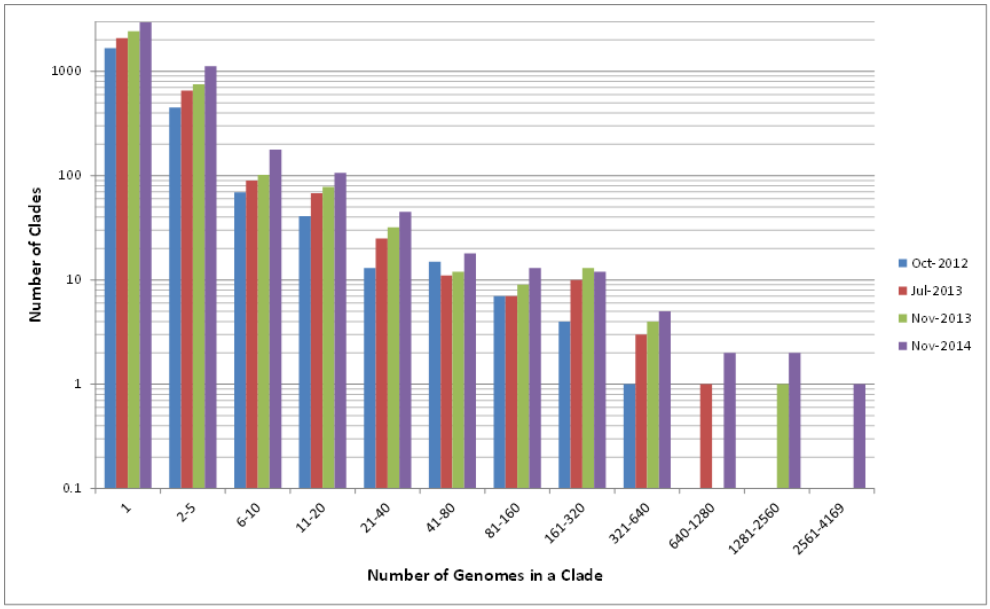

Figure 5. Clade growth in four sequential snapshots.

http://dx.doi.org/10.5772/62125

the species level as defined by NCBI Taxonomy. The subclades are calculated to show the closeness of the groups at the higher level. The relationship within the species-level group is further refined with whole pairwise genome alignment performed by megablast [18]. Tight genomic groups are defined at the level of 95% identity over the 95% genome coverage. By using the representative genomes from the tight groups, we can reduce the redundancy in comparative genomic studies. Other targets can be used for more refined population variation studies within species or SNP analysis for pathogen outbreak detection. These target sets require more accurate distance measure such as whole genomic alignments, K-mer distance [21].

3.1. Clades and species

Using a taxonomy-aware clustering algorithm does not completely solve the discrepancies between the species-level clades and traditional species. Genome sequences provide great opportunity to refine the classical taxonomic description of prokaryotes [23]. All cases of discrepancy were manually evaluated; most of them have been resolved by literature support. Some examples are described below.

3.1.1. Different species merged into a single clade

Escherichia coli and some *Shigella* species are combined in a single clade by ribosomal marker distance. *Shigella*, which is recognized as a genus with four species in most situations, taxonomically belongs to the diverse *E. coli* group, but the genus-level distinction has been retained due to historical recognition of its medical significance. *Shigella* has adapted to higher primates as the only natural hosts.

The genus *Brucella* consists of 10 classically recognized species [http://icsp.org/subcommittee/brucella/] based on antigenic/biochemical characteristics and primary host species*: Brucella abortus*(cattle); *Brucella canis* (dogs); *Brucella ceti* (marine mammals); *Brucella inopinata*; *Brucella melitensis* (sheep and goats); *Brucella microti*; *Brucella neotomae* (rodents); *Brucella ovis* (sheep); *Brucella pinnipedialis* (marine mammals); *Brucella suis* (swine, cattle, rodents, wild ungulates), and recently described in [24] *Brucella papionis* isolated from baboons (*Pappio* spp.). The wave of Next-Generation Sequencing brought in almost a hundred new isolates from a population of *Brucella*, which are clearly distinct from currently recognized species that are tentatively designated at the species level. These unnamed isolates have not yet been characterized using traditional methods, or the species name has not yet been validly published. *Brucella* genus–level clade is shown in Figure 6.

Single species represented by multiple clades

Prochlorococcus and marine *Synechococcus* organisms are small marine cyanobacteria, their genomes are characterized by small size and an evolutionary trend toward low GC content [25]. Whereas many shared derived characters define *Prochlorococcus* as a clade, many genome-based analyses recover them as paraphyletic. The single species, *Prochlorococcus marinus*, comprises six named ecotypes. Our ribosomal marker analysis and whole-genome alignment (described above in section on Methods) analysis suggests that this species should be repre-

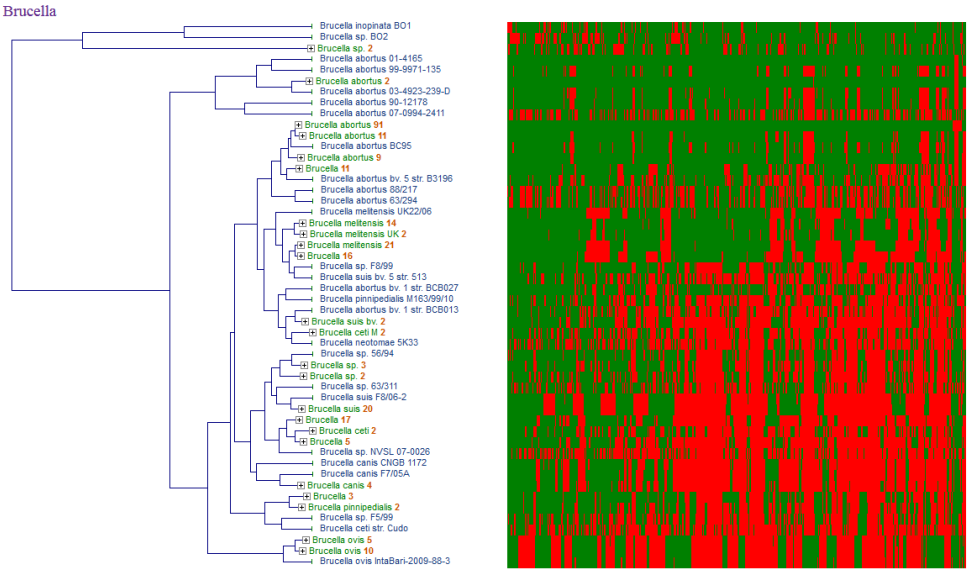

Figure 6. Ribosomal-marker-based clade comprises various species of Brucella. The pairwise genome distance is defined by the number of shared proteins in the core set of Brucella pan-genome. Green dots – proteins present in CORE set; red dots – proteins absent in CORE set.

Figure 7. *Prochlorococcus marinus* interspecies diversity. The dendrogram is calculated using blast genome alignment score (%identity). The leaf nodes displayed as circles represent genomes of individual isolates/strains.

sented by 11 different clades (see Figure 7.) These results are supported by recent genomic analysis of the genus of *Prochlorococcus* [26].

Novel species from noncultured not-isolated single cell and metagenome assemblies and new unclassified isolates (<genus> sp.) from clinical and epidemiological studies can be organized in hierarchical groups by genome sequence comparison methods. These groups can be used for downstream analysis: 1) pan-genome by clades not species; 2) groups of closely related genomes below species that can be calculated by nucleotide whole-genome comparison like K-mer or BLAST; 3) classification validation; 4) visualization of large data sets by selecting the

http://dx.doi.org/10.5772/62125

genome representatives. Some of the applications marker-based clades and tight genome groups have been previously briefly described in [27,28].

4. Conclusions

No matter how impressive the numbers of genome sequencing projects are, they represent a miniscule fraction of the total number of bacterial species. The future genomic analysis tools will have to take into consideration the uncertain origin of the DNA sequences during analysis. Making sense of genomic data is one of the goals that are aided by the genome clustering procedure. The hierarchical infrastructure provides the foundation for further development of genome analysis and visualization tools.

Acknowledgements

This research was supported by the Intramural Research Program of the NIH, National Library of Medicine. The authors are thankful to David J. Lipman, James Ostell, Scott Federhen, Michael Galperin, Eugene Koonin, Yury I. Wolf for productive discussions and to Vyacheslav Brover for the database support.

Author details

Leonid Zaslavsky, Stacy Ciufo, Boris Fedorov, Boris Kiryutin, Igor Tolstoy and Tatiana Tatusova*

*Address all correspondence to: tatiana@ncbi.nlm.nih.gov

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

References

[1] Yarza P, Yilmaz P, Pruesse E, Glöckner FO, Ludwig W, Schleifer KH, Whitman WB, Euzéby J, Amann R, Rosselló-Móra R. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat Rev Microbiol*. 2014;12(9):635-45.

[2] Rinke C, Schwientek P, Sczyrba A, Ivanova NN, et al. Insights into the phylogeny and coding potential of microbial dark matter. *Nature*. 2013;499(7459):431-7.

[3] Audic S, Robert C, Campagna B, Parinello H, Claverie JM, Raoult D, Drancourt M. Genome analysis of Minibacterium massiliensis highlights the convergent evolution of water-living bacteria. *PLoS Genet*. 2007;3(8):138.

[4] MacDonald SJ, Thomas GH, Douglas AE. Genetic and metabolic determinants of nutritional phenotype in an insect-bacterial symbiosis. *Mol Ecology*. 2011;20(10):2073-84.

[5] Manzano-Marín A, Latorre A. Settling down: the genome of Serratia symbiotica from the aphid Cinara tujafilina zooms in on the process of accommodation to a cooperative intracellular life. *Genome Biol Evol*. 2014;6(7):1683-98.

[6] Pamp SJ, Harrington ED, Quake SR, Relman DA, Blainey PC.. Single-cell sequencing provides clues about the host interactions of segmented filamentous bacteria (SFB). *Genome Res*. 2012;22(6):1107-19.

[7] Rosselló-Móra R, Amann R. Past and future species definitions for Bacteria and Archaea. *Syst Appl Microbiol*. 2015;Feb 20:22-3.

[8] Ciccarelli FD, Doerks T, von Mering C, et al. Toward automatic reconstruction of a highly resolved tree of life. *Science*. 2006;311(5765):1283-7.

[9] Mende DR, Sunagawa S, Zeller G, Bork P. Accurate and universal delineation of prokaryotic species. *Nat Methods*. 2013;10(9):881-4.

[10] Lang JM, Darling AE, Eisen JA. Phylogeny of bacterial and archaeal genomes using conserved genes: supertrees and supermatrices. *PLoS One*. 2013;8(4):{PT PageRange}e62510{/PageRange PT}.

[11] Wu D, Jospin G, Eisen JA. Systematic identification of gene families for use as "markers" for phylogenetic and phylogeny-driven ecological studies of bacteria and archaea and their major subgroups. *PLoS One*. 2013;8(10):{PT PageRange}e77033{/PageRange PT}.

[12] Darling AE, Jospin G, Lowe E, Matsen FA IV, Bik HM, Eisen JA. PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ*. 2014;2:e243.

[13] Tanabe AS, Toju H. *PLoS One*. 2013;8:e76910.

[14] Larsen MV, Cosentino S, Lukjancenko O, Saputra D, Rasmussen S, Hasman H, Sicheritz-Pontén T, Aarestrup FM, Ussery DW, Lund O. Benchmarking of methods for genomic taxonomy. *J Clin Microbiol*. 2014;52(5):1529-39.

[15] Vernikos G, Medini D, Riley DR, Tettelin H. Ten years of pan-genome analyses. *Curr Opin Microbiol*. 2015;Feb(23):148-54.

[16] NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. 2014;42(Database issue):D7-17.

[17] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215:403-10.

http://dx.doi.org/10.5772/62125

[18] McGinnis S, Madden TL. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res*. 2004;32(Web Server Issue):W20-5.

[19] Everitt BS, Landau S, Leese M, Stahl D. Cluster Analysis. 5th ed. Wiley; 2011.

[20] Felsenstein J. *Inferring Phylogenies*. 2nd ed. Sinauer Associates; 2004.

[21] Compeau P, Pevzner P, Teslar G. How to apply de Bruijn graphs to genome assembly. *Nature Biotechnol*. 2011;29(11):987-91.

[22] Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA*. 1992;89(22):10915-19.

[23] Whitman WB. Genome sequences as the type material for taxonomic descriptions of prokaryotes. *Syst Appl Microbiol*. 2015;Feb 20:24-7.

[24] Whatmore AM, Davison N, Cloeckaert A, Al Dahouk S., Zygmunt MS, Brew S D, Perrett LL, Koylass MS, Vergnaud G, Quance C, Scholz HC, Dick, EJ, Hubbard G, Schlabritz-Loutsevitch NE. Brucella papionis sp. nov., isolated from baboons (Papio spp.). *Int J Sys Evolution Microbiol*. 2014;64:4120-28.

[25] Olga Zhaxybayeva, W. Ford Doolittle, R. Thane Papke†, J. Peter Gogarten. Intertwined evolutionary histories of marine Synechococcus and Prochlorococcus marinus. *Genome Biol Evol*. 2009;1:325-339.

[26] Thompson CC, Silva GG, Vieira NM, Edwards R, Vicente AC, Thompson FL. Genomic taxonomy of the genus prochlorococcus. *Microb Evol*. 2013;66(4):752-62.

[27] Zaslavsky L, Tatusova TA. Clustering analysis of proteins from microbial genomes at multiple levels of resolution. *Lect Notes Comp Sci*. 2015;9096:438-9.

[28] Tatusova T, Ciufo S, Federhen S, Fedorov B, McVeigh R, O'Neill K, Tolstoy I, Zaslavsky L. Update on RefSeq microbial genomes resources. *Nucleic Acids Res*. 2015;43(Database Issue):D599-605.