*6614–6624* *Nucleic Acids Research, 2016, Vol. 44, No. 14* *Published online 24 June 2016*
*doi: 10.1093/nar/gkw569*

# NCBI prokaryotic genome annotation pipeline

**Tatiana Tatusova[1,†], Michael DiCuccio[1,†], Azat Badretdin[1], Vyacheslav Chetvernin[1], Eric P. Nawrocki[1], Leonid Zaslavsky[1], Alexandre Lomsadze[2], Kim D. Pruitt[1], Mark Borodovsky[2,3,*,‡] and James Ostell[1,‡]**

[1]National Center for Biotechnology Information, U.S. National Library of Medicine, Bethesda, MD 20894, USA, [2]Wallace H. Coulter Department of Biomedical Engineering, Georgia Tech, Atlanta, GA 30332, USA and [3]School of Computational Science and Engineering, Georgia Tech, Atlanta, GA 30332, USA

Received March 18, 2016; Revised June 08, 2016; Accepted June 13, 2016

## ABSTRACT

**Recent technological advances have opened unprecedented opportunities for large-scale sequencing and analysis of populations of pathogenic species in disease outbreaks, as well as for large-scale diversity studies aimed at expanding our knowledge across the whole domain of prokaryotes. To meet the challenge of timely interpretation of structure, function and meaning of this vast genetic information, a comprehensive approach to automatic genome annotation is critically needed. In collaboration with Georgia Tech, NCBI has developed a new approach to genome annotation that combines alignment based methods with methods of predicting protein-coding and RNA genes and other functional elements directly from sequence. A new gene finding tool, GeneMarkS+, uses the combined evidence of protein and RNA placement by homology as an initial map of annotation to generate and modify *ab initio* gene predictions across the whole genome. Thus, the new NCBI's Prokaryotic Genome Annotation Pipeline (PGAP) relies more on sequence similarity when confident comparative data are available, while it relies more on statistical predictions in the absence of external evidence. The pipeline provides a framework for generation and analysis of annotation on the full breadth of prokaryotic taxonomy. For additional information on PGAP see https://www.ncbi.nlm.nih.gov/genome/annotation_prok/ and the NCBI Handbook, https://www.ncbi.nlm.nih.gov/books/NBK174280/.**

## INTRODUCTION

The first version of the NCBI Prokaryotic Genome Automatic Annotation Pipeline (PGAAP) combining HMM-based gene prediction algorithms with protein sequence similarity search methods was developed in 2001-2002. The initial pipeline used a combination of automatic protein-coding gene model prediction via two prediction methods, GeneMarkS (1) and Glimmer (2). These predictions were then augmented using information on evolutionarily conserved proteins from Clusters of Orthologous Groups or COGs (3) and NCBI Prokaryotic Clusters (4). Proteins from these clusters were mapped to the genome in order to search for genes missed by the *ab initio* predictors. Genes of ribosomal RNAs were predicted either by the BLASTn sequence similarity search using the entries from the RNA sequence database as queries or by running specialized tools, such as Infernal (5,6) and Rfam (7). Genes of transfer RNAs were predicted using tRNAscan-SE (8). The Standard Operating Procedure (SOP) for the first version of the NCBI genome annotation pipeline was published in 2008 (9). Here, we describe a new design of the NCBI Prokaryotic Annotation Pipeline (PGAP).

With the new version of PGAP, we introduced several new features. First and most importantly, the pipeline now uses a pan-genome approach to protein annotation with pan-genome proteins defined for a specific clade (see below). We have built a collection of clusters of homologous proteins; the proteins are used to generate a map of protein footprints in the genomic sequence submitted for annotation. We assume that most of the proteins conserved in a given clade—the core proteins—should be encoded in a genome of a new species in the clade. We have observed that in highly populated clades the core genes comprise up to 75% of the total number of annotated genes in a single genome (Table 1). Second, the pipeline incorporates additional specialized search tools to identify novel non-protein-coding functional elements, including CRISPR regions. Third, to identify protein-coding genes the pipeline

*To whom correspondence should be addressed. Tel: +1 404 894 8432; Fax: +1 404 894 4243; Email: borodovsky@gatech.edu
†These authors contributed equally to the work as the first authors.
‡Senior authors.

uses a two-pass approach designed to detect frameshifted genes and pseudogenes. Fourth, we have replaced the original *ab initio* predictors with a new software tool, GeneMarkS+, that integrates extrinsic information (alignment based protein predictions, predicted RNA genes, etc.) with intrinsic information on genome-specific sequence patterns of protein-coding regions. Finally, for better process control, PGAP has migrated into a new specialized application framework now widely used inside NCBI that provides enhanced computational performance, reliability and a comprehensive tracking system.

The major difference with other published pipelines (10–13), including the previous version of the NCBI pipeline (9), is that we calculate a set of alignment-based hints for protein-coding and non-protein-coding regions prior to executing *ab initio* prediction. The hints are incorporated by GeneMarkS+ into gene prediction (as described below). Other pipelines (10–13) attempt to run *ab initio* algorithms first to reduce computational load on alignment-based searching. As a consequence, such pipelines face an issue of reconciling the *ab initio* and alignment based predictions. Because GeneMarkS+ can reconcile these hints with *ab initio* predictions internally, PGAP can make more effective use of its computational resources to provide high-quality protein alignments first and guarantee that these alignments will not be compromised by conflicting *ab initio* predictions in subsequent processing.

## MATERIALS AND METHODS

### PGAP general structure and input

The PGAP pipeline is designed to annotate both complete genomes and draft genomes comprising multiple contigs. PGAP is deeply integrated into NCBI infrastructure and processes, and uses a modular software framework, GPipe, developed at NCBI for execution of all annotation tasks, from fetching of raw and curated data from public repositories (the Sequence and Assembly databases) through sequence alignment and model-based gene prediction, to submission of annotated genomic data to public NCBI databases.

On input, PGAP accepts an assembly (either draft or complete) with a predefined NCBI Taxonomy ID that defines the genetic code of the organism. PGAP also accepts a predetermined clade identifier, matching the genome in question to a species-specific clade. Clade IDs are computed using a series of 23 universal ribosomal protein markers and are independent of taxonomy. In the absence of a clade ID, we can infer the ID from taxonomy in the majority of cases. The clade ID determines the realm of core proteins used as the target protein set. PGAP annotation of a new genomic sequence can be requested at the time of submission to GenBank. Taxonomic and clade identifiers are determined outside of the annotation pipeline, and are influenced by GenBank curatorial decisions. The clade-dependent sets of protein clusters as well as sets of curated structural ribosomal RNAs (5S, 16S and 23S) are generated and maintained outside of PGAP. More details on the PGAP workflow are provided below.

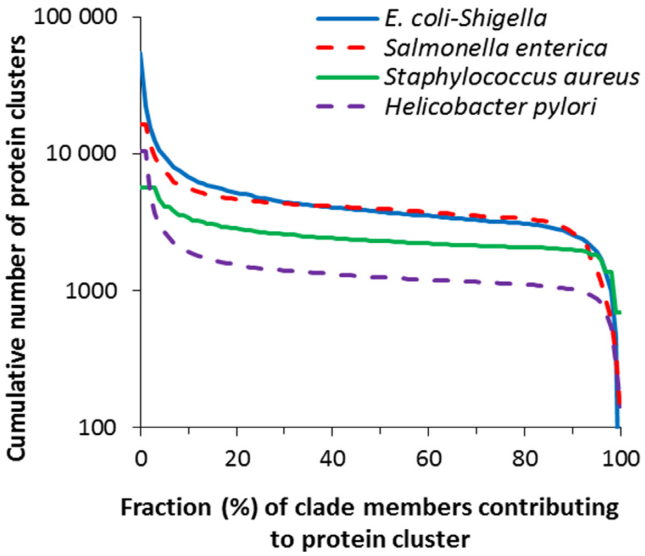


**Figure 1.** Cumulative number of protein clusters (Y) is defined for a given X (%) as the number of clusters containing proteins from fraction x ≥ X of all members of the clade. Data are presented for the four well studied clades.

### Pan-genome approach

The genome sequencing revolution has radically altered the field of microbiology. Whole-genome sequencing for prokaryotes became a standard method of study ever since the first complete genome of free-living organism, *Haemophilus influenza*, was sequenced in 1995 (14). Due to the widespread use of the next generation sequencing (NGS) techniques, thousands of genomes of prokaryotic species are now available, including genomes of multiple isolates of the same species, typically human pathogens. Thus, the mere density of comparative genomic information for high interest organisms provides an opportunity to introduce a pan-genome based approach to prediction of the protein complement of a species.

The collection of prokaryotic genomes available at NCBI is growing exponentially and shows no signs of abating: as of January 2016 NCBI's assembly resource contains 57 890 genome assemblies representing 8047 species (see genome browser https://www.ncbi.nlm.nih.gov/genome/browse/, for the up-to-date information). Notably, genomes of different strains of the same species can vary considerably in size, gene content and nucleotide composition. In 2005, Tettelin *et al*. (15) introduced the concept of *pan-genome*, aiming to provide a compact description of the full complement of genes of all the strains of a species. Genes common to all pan-genome members (or to the vast majority of them) are called *core genes*; those present in just a few clade members are termed *accessory* or *dispensable genes*; genes specific to a particular genome (strain) are termed *unique genes* (16).

In PGAP we define the pan-genome of a clade at a species or higher level (17). To be included as a *core gene* for a species-level pan-genome, we require the gene to be present in the vast majority—at least 80%—of all genomes in the clade. A set of *core genes* gives rise to a set of *core proteins*. We show in Figure 1 how the number of protein clusters,

**Table 1.** Statistics of genomes, genes and core protein clusters in the 10 largest clades

| Clade name | # Genomes | # CDS Total | Median #CDS/Genome | # Core protein clusters |
|---|---|---|---|---|
| *Escherichia - Shigella* | 1502 | 7 594 943 | 4990 | 3220 |
| *Salmonella* | 527 | 2 334 839 | 4511 | 3393 |
| *Staphylococcus aureus* | 445 | 1 195 744 | 2672 | 2066 |
| *Streptococcus* | 334 | 714 947 | 2150 | 1223 |
| *Brucella* | 283 | 886 682 | 3120 | 1704 |
| *Helicobacter pylori* | 268 | 433 955 | 1631 | 1200 |
| *Streptococcus agalactiae* | 254 | 523 389 | 2038 | 1595 |
| *Acinetobacter* | 212 | 796 523 | 3785 | 2755 |
| *Neisseria* | 194 | 402 822 | 1997 | 1540 |
| *Leptospira interrogans* | 186 | 778 660 | 4062 | 3024 |

for each of four well studied large clades, depends on the fraction of the clade members that contribute proteins to the cluster. There are three critical regions in this analysis: (i) unique genes, present in less than 1% of all clade members; (ii) dispensable genes, present in 1–20% of genomes; and (iii) core genes, found in at least 80% of the represented genomes. Based on our analysis, there are very few clusters appearing in at least 20% of the members of a clade but no more than 80% of the members. The use of a cutoff of 80% was chosen to capture a wide set of genes conserved within the whole clade while eliminating genes having less abundant representation. We further subject the *core proteins* to clustering using USearch to reduce the total number of proteins required to represent the full protein complement of the pan-genome (18). We use the representative *core proteins* to infer genes for homologous core proteins in a newly sequenced genome (19).

The notion of the *pan-genome* can be generalized beyond a species level and applies, in fact, to any taxonomy level (from genus to phylum to kingdom). Notably, in the pan-genomes of Archaea and Bacteria, the universally conserved ribosomal genes make a group of core genes. The main practical value of the pan-genome approach is in formulating an efficient framework for comparative analysis of large groups of closely related organisms separated by small evolutionary distances as defined by ribosomal protein markers (20,21).

### Prediction of RNA genes (structural RNA, tRNA, small ncRNA)

Structural rRNAs (5S, 16S and 23S) are highly conserved in closely related prokaryotic species. The NCBI RefSeq Targeted Loci collection (22) contains curated sets of the three types of rRNA gene sequences, which serve as reference sets for PGAP (https://ncbi.nlm.nih.gov/RefSeq/targetedloci/). To identify genes for 16S and 23S rRNAs PGAP uses members of the reference sets as queries in BLASTn (23). Hits that correspond to partial alignments are dropped if they fall below a certain coverage and identity thresholds with respect to the average length of the corresponding rRNA (50% coverage and 70% identity for 16S rRNA; 50% coverage and 60% identity for 23S rRNA). Borders of predicted rRNA genes are defined by a voting mechanism similar to the one mentioned below for identifying gene starts among several alternative start codons.

For prediction of 5S rRNAs and small ncRNAs, PGAP uses *cmsearch* (ver. 1.1.1) along with covariance models, score thresholds and recommended command line options from the Rfam database (release 12.0 (7)). Current execution of this *cmsearch* version has been optimized to permit direct use of the tool without a preliminary BLASTn search (5–7).

For prediction of tRNA sequences, PGAP relies on tRNAscan-SE. The input genomic sequence is split into overlapping fragments long enough to cover a tRNA gene with possible introns. These fragments are used as inputs to tRNAscan-SE (8), currently one of the most widely used tRNA gene identification tools. Domain specific parameters of tRNAscan-SE are selected automatically for each genome (8).

All predicted RNA genes from the above steps are collected and presented to GeneMarkS+ as a set of RNA gene 'footprints' (Figure 2). GeneMarkS+ has several labels ('M', 'N' and 'R') for RNA gene footprints; the labels specify different types of possible overlaps between protein-coding genes and RNA genes.

### Repetitive regions (CRISPRs) and mobile genetic elements (prophages)

Clustered regularly interspaced short palindromic repeats (CRISPRs), along with associated proteins, comprise a prokaryotic defense system. Interest in CRISPRs has recently blossomed due to the opportunity to use the system as programmable restriction enzymes. CRISPRs are commonly found in almost all archaeal genomes while they are less frequent in bacterial genomes (24,25). For CRISPRs prediction and annotation PGAP uses a combination of the CRISPR recognition tool (CRT) (26) and PILER-CR (27).

Phage and plasmid genes are frequent subjects of horizontal transfer and may be difficult to predict by gene finding tools that are tuned to identify *native* genes or *foreign* genes adapted to the genomic context in evolution to (1). Some phage and plasmid genes can be identified by alignment based methods. In PGAP, we utilize a curated set of phage and plasmid proteins and map these proteins to a genome in question using tBLASTn and ProSplign as described in the section below. The reference set of phage proteins was created earlier in a project on inference and curation of phage protein clusters (4). High-scoring phage and plasmid protein alignments form another set of footprints in the input to GeneMarkS+.

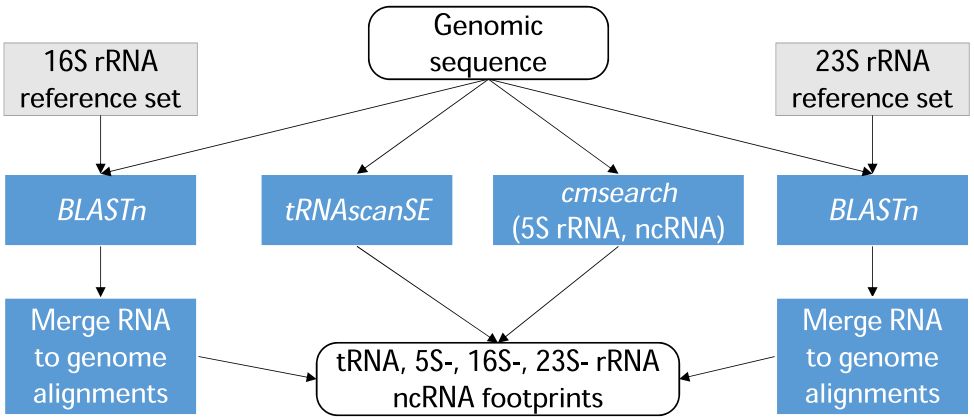


**Figure 2.** A fragment of the PGAP execution graph: prediction of structural RNA genes (ncRNA, tRNA, 5S-, 16S-, 23S- rRNA).

**Protein alignments and integration of multiple evidence types into genome annotation with GeneMarkS+**

GeneMarkS+, a self-training gene finder, was designed to detect protein-coding patterns in genomic sequence and to infer gene locations compatible with externally supplied evidence, the *hints* or *footprints*, indicating presence of various functional elements in specific sequence intervals. Examples of *hints* include protein-coding intervals suggested by protein-to-genome alignments; intervals covering RNA genes identified by specialized tools; or intervals recognized as repetitive sequences, e.g. elements of CRISPR. The core of the integration algorithm is the *ab initio* gene finding tool GeneMarkS (1), which implements the Viterbi algorithm for a hidden semi-Markov model (HSMM) of prokaryotic genomic sequence (28). External hints transformed into sequence labels restrict possible parses of a genome in question into protein-coding and non-protein-coding regions. Of note, these constraints reduce the set of possible parses and, thus, help accelerate execution of the Viterbi algorithm. Species specific parameters of GeneMarkS are determined by iterative unsupervised training (1) on the whole genomic sequence submitted for annotation (see Figure 3); thus the training occurs without any hints defined restrictions.

In the first round (or pass) of annotation we align representative proteins from the clade specific core clusters using tBLASTn (22) to the genome. High scoring protein alignments are further refined by ProSplign, a frameshift-aware protein to genome aligner. An added benefit of ProSplign is its ability to identify cases in which newly defined genes for core proteins are frameshifted (https://ncbi.nlm.nih.gov/sutils/static/prosplign/prosplign.html) . The core protein alignments are then transformed into footprints (interval hints) for consumption by GeneMarkS+. In the presence of a frameshift, the pipeline generates a 'corrected' sequence of the genome, adding or subtracting bases to repair the translation frame. This repaired sequence is consumed by GeneMarkS+. During this first pass (top half of Figure 3) the core protein footprints are used as interval hints (with the label 'C') that carry information on the frame of either an intact mapped gene, a repaired pseudogene or a repaired gene with sequencing error. Notably, the frameshifted regions are 'repaired' not only in the first pass but in the second pass as well, when the focus moves to identification of genes of non-core proteins. The first pass of the pipeline concludes with execution of the first run of GeneMarkS+ making the initial prediction of the genome-wide complement of genes and proteins.

In the second pass (bottom half of Figure 3), proteins predicted by GeneMarkS+ in locations that were not covered by the footprints of core proteins are considered as the 'seeds' or 'prototypes' of the non-core genes. These seeds are searched via BLASTp against a broader database of RefSeq proteins to identify a subset of seeds supported by evolutionarily conserved non-core proteins. Each identified seed protein is aligned back to the genome using ProSplign. The ProSplign non-core protein alignments are combined with the core protein alignments from the first pass and the predictions of all protein-coding regions are transformed into footprints (intervals with 'C' label) defined on the repaired genomic sequence.

Importantly, the core genes predicted in the first round as well as the first round *ab initio* predicted genes are evaluated in the second round with BLASTp searching against a broader set of representative proteins from all prokaryotic protein clusters. High-scoring BLASTp hits are then mapped back to the genome, and proteins that map incompletely or with disruptions are realigned to the genome using tBLASTn and ProSplign. The rationale of repeating this analysis is that the larger set of all protein clusters provides not only new but also more extensive evidence for predicted gene models thus improving the accuracy. To give an example (Figure 4) in place of overlapping CDS features predicted by *ab initio* in the first round (panel A), frameshifts can be identified when, in the second round, proteins homologous to the predicted CDS sequences are aligned by ProSplign to genomic sequence (panel B). In another example (Figure 5) one of the alternative start codons (panel A) may get more supporting evidence than the other in the second round of alignments (panel B). Genomic intervals predicted to be protein-coding by the two rounds of alignments make the set of 'protein footprints' used as input in the second run of GeneMarkS+.

These final protein-coding footprints are combined with information on other types of hints, including RNA genes

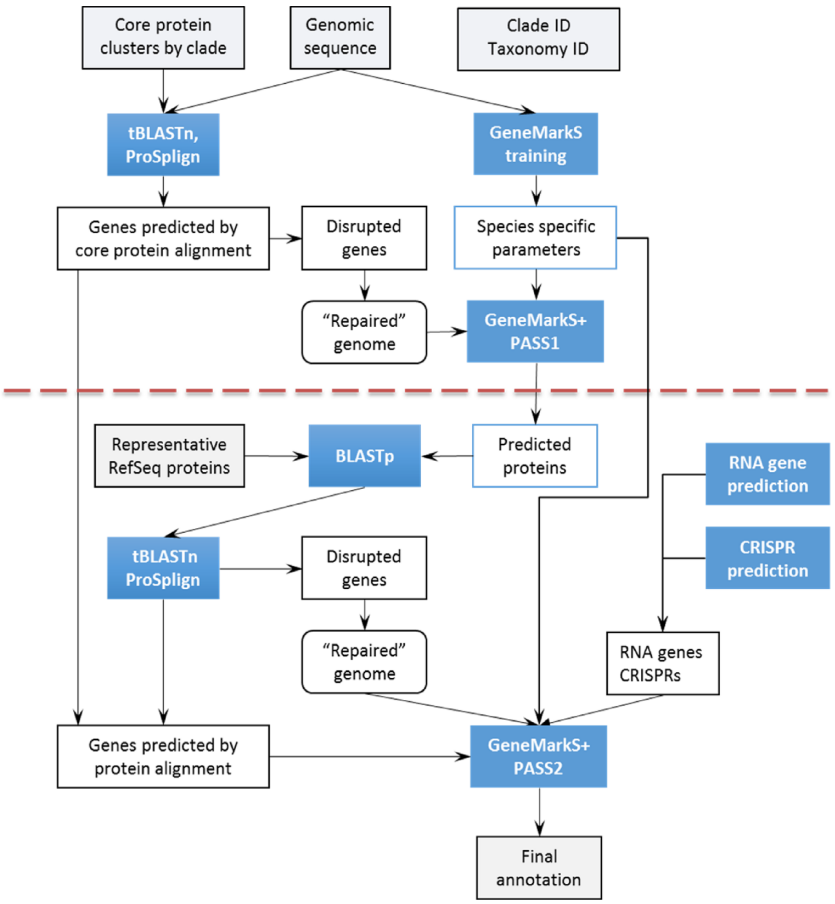


**Figure 3.** Flowchart of PGAP. The red dotted line indicates separation between pass one and pass two (see text for details).

and CRISPR sequences, prior to the second pass of GeneMarkS+. The presence of RNA genes (rRNA, tRNA, small ncRNA) prohibits prediction of protein-coding genes in the same location, less small overlaps. This restriction is particularly important for low GC content genomes in which sequences of RNA genes frequently trigger prediction of false protein-coding genes due to the relatively high GC content of RNA genes.

We also considered other types of hints that occur rarely and may affect one or two genes in a whole genome. Examples include non-canonical start sites or sites of non-conventional amino acid translation, such as selenocysteine. While these hints do not have matching labels, GeneMarkS+ is able to integrate such hints outside the labeling routine by assigning a non-canonical start codon or by making a stop codon at specified location a part of the protein-coding region (stop codon read-through).

In PGAP, the second run of GeneMarkS+ produces the final structural gene annotation in which all identified protein-coding and non-coding regions are compatible with the whole set of externally defined hints (labels). One critical aspect of the algorithm revolves around protein-coding region start site selection. Within PGAP, we determine start sites using a combination of *ab initio* and alignment evidence. In the presence of even a single aligned protein with sufficiently strong score we give more weight to the start site determined by the alignment; in the absence of protein alignment information, GeneMarkS+ prediction determines the start site. There are two specific cases that we must then consider within pipeline execution. First, the protein evidence we use, as described above, is generated by a sample of representatives of the clustered pan-genome proteins. As a result, each such protein carries a weight proportional to the number of proteins it represents. Therefore, when alignments suggest alternative start sites we use a simple voting algorithm to favor the start site with the largest weight. Second, we must filter out cases when the protein homology in some regions is too weak to reliably support protein-coding hints. These weak alignments are filtered out to allow GeneMarkS+ to operate in such regions in *ab initio* mode to identify gene models in all possible frames.

In practice, for genomes from well-studied species such as *E. coli* and *S. typhimurium* that have large sets of core genes, evidence from core proteins is abundant and the role of *ab initio* prediction is limited. Conversely, in less well studied species where the full complement of core proteins less well known or reduced to clusters of ribosomal proteins, the protein-coding regions are predicted mostly by GeneMarkS+ running in the GeneMarkS mode (*ab initio*). Thus, the two-pass approach outlined here is robust in the face of a wide variety of taxa and genome features.


**Figure 4.** A region in the *Deinococcus radiodurans* R1 genome assembly (GCA_000008565.1) contains three overlapping ORFs predicted ab initio as CDSs in the first pass of PGAP. Automatic evaluation of the cross-species protein evidence through the second pass of PGAP reveals proteins bearing homology to all three fragments. Alignment of the proteins to the genome reveals otherwise unpredicted frameshifts. Green bars represent genes, red bars – coding regions; grey bars – alignments with red vertical bars indicating mismatches. (**A**) A region of Chromosome 1 of *D. radiodurans* (AE000513.1) containing the three CDS features is displayed alongside the six-frame translation. (**B**) The same region, updated to include final annotation markup with a frameshifted CDS as well as supporting proteins that demonstrate a consistent pattern and location of two frameshifts (marked by arrows at positions 100 733 and 100 959).

As new organisms are sequenced and annotated, PGAP updates clusters of core proteins. In turn, the use of newly updated clusters allows for iterative improvement of annotation of all prokaryotic genomes in the database.

## Protein naming

Protein naming is a critical part of the PGAP pipeline, since the protein name has to reflect the protein function. We use a BLASTp search for all newly identified protein products against a specialized database comprised of representatives of all automatically derived prokaryotic protein clusters (4), reviewed proteins from the UniProt-SwissProt Protein Knowledgebase (29) and all curated bacteriophage proteins from the RefSeq collection. The resulting BLASTp alignments are filtered to remove those that fall below thresholds of identity (25%) and coverage (the alignment region should comprise at least 70% of both query and subject (target) length). Each BLASTp alignment is assigned a distance value, varying between 0 (identical) and 1 (entirely dissimilar) based on the ratio of the BLASTp score to the expected BLASTp score for a perfect match alignment. A query protein is assigned to the *closest* identification cluster situated at least at 0.5 distance from the query protein. However, no such identification is made if within a distance of 0.1 from the query protein there is another 'competing' identification cluster. Thus, we favor unique and unambiguous assignments: if a query protein matches best to one and only one identification cluster, we accept the match; if there is an ambiguous placement (i.e. a close unrelated identification cluster), we do not accept the match. The NCBI protein naming conventions use similar rules that have been adopted by the UniProt-SwissProt Protein Knowledgebase (29,30) (http://www.uniprot.org/docs/proknameprot).

Assignment of protein names, however, important this part of the pipeline is, is not in the focus of this paper. The full set of rules of functional annotation along with any additional related information will be described in a forthcoming publication.

## The PGAP execution system

The new PGAP uses a robust, high-performance execution framework (GPipe) developed for in-house use at NCBI. This system provides distributed parallel computing, robust tracking of all execution tasks and optimization of compute-intensive steps. Tracking of execution and decision-making is a critical feature that permits easy retrieval of the evidence behind key algorithmic decisions. The modular structure of PGAP allows for easy incorporation of new algorithms.

At its heart, GPipe is a workflow management tool that describes collections of tasks connected by dataflow between programs. In the GPipe model, execution consists of generating a *build* (statement of intent to complete a workflow). Each build owns one or more *build-runs* (an actual attempt to complete the workflow). A build-run contains an execution graph, which connects execution of specific tasks to other tasks using strongly-type data connections as edges. Each task in turn contains potentially multiple executions of a series of programs. All parameters and dataflow connections for all executions are tracked in a relational database and can be queried to identify historical usage patterns and deviations from expected executions. Workflows

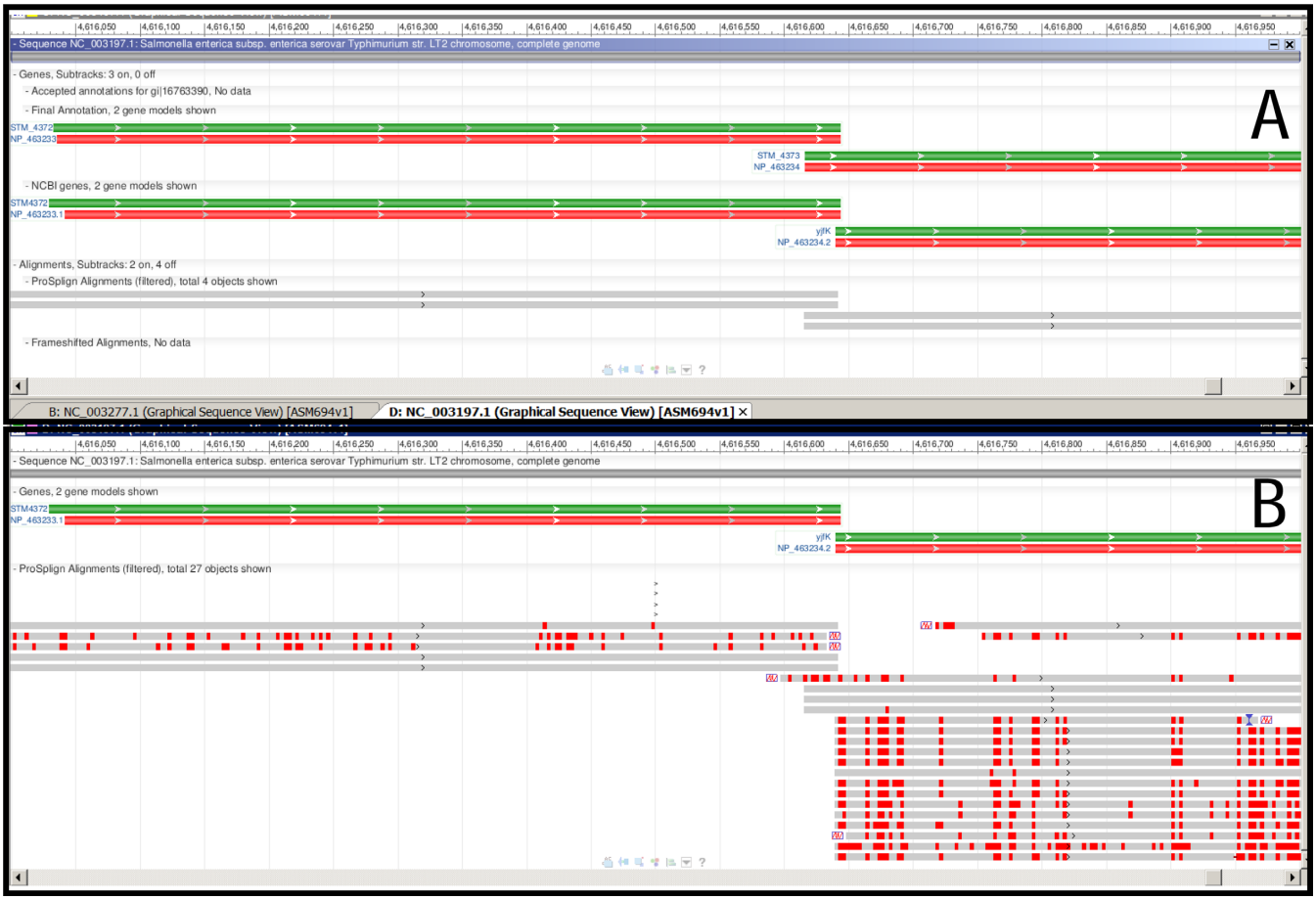


**Figure 5.** Annotation of genome of *Salmonella enterica* subsp. *enterica* serovar *Typhimurium* str. LT2 (NC_003197). Protein alignment provides support for gene start selection. See legend to Figure 4 for description of the meaning of green, red and gray bars. (**A**) the first round of alignments of protein representatives from the 'core' protein clusters doesn't give enough evidence for gene start selection. (**B**) the second round of alignments clearly supports a shorter gene model which does not overlap with the upstream gene.

can be subdivided into specialized subgraphs that execute repetitive tasks; these subgraphs can be created many times in each build-run to handle varying types of evidence using identical sets of actions (e.g. Figure 2).

The pipeline execution environment consists of four major components: (i) a database in which tasks are organized as builds and build-runs; (ii) a series of graphs and graph templates used to organize execution tasks; (iii) C++ object code implementing the execution tasks; and (iv) an execution application that reads build definitions from the database and executes the appropriate tasks.

## PGAP output: validation and final annotation

In addition to identifying structural components and establishing functional identification, PGAP produces reports in a wide variety of formats used by curators, submitters, as well as internal submission tools. Included in the reports produced are: (i) the primary annotated genome objects, represented in NCBI's ASN.1 data model and suitable for direct submission into GenBank; (ii) a report on annotation markup discrepancies requiring submitter or curator attention; (iii) genome annotation in GenBank flat file format ready for manual review and public display; and (iv) statistics from the annotation process along with citation of supporting evidence for each gene model. Concomitant with the ASN.1 markup produced in PGAP, we record the evidence selected for each model to support the selection of the specific start/stop site as well as evidence used to support the functional identification of each inferred object (e.g. Figure 6).

In recent years NCBI, in conjunction with INSDC, has developed validation procedures designed to ensure that the new records represent biologically valid and consistently formatted data. These quality control (QC) procedures have been incorporated into the PGAP pipeline. The quality of annotation of a prokaryotic genome can be assessed by several metrics. In an effort to develop standards for prokaryotic genome annotation, NCBI has established a collaboration with other major archive databases and major sequencing centers. This collaboration resulted in a set of annotation standards approved and accepted by all major annotation pipelines (20). Genome annotation should follow INSDC submission guidelines (GenBank/ENA/DDBJ), https://www.ncbi.nlm.nih.gov/genbank/genomesubmit/.

In addition to the standards for genomes mentioned above, PGAP implements rules defining complete genome

```
##Genome-Annotation-Data-START##
Annotation Provider               :: NCBI
Annotation Date                   :: 02/05/2016 09:49:57
Annotation Pipeline               :: NCBI Prokaryotic Genome
                                     Annotation Pipeline
Annotation Method                 :: Best-placed reference protein
                                     set; GeneMarkS+
Annotation Software revision      :: 3.1
Features Annotated                :: Gene; CDS; rRNA; tRNA; ncRNA;
                                     repeat_region
Genes (total)                     :: 3,034
CDS (total)                       :: 2,945
Genes (coding)                    :: 2,921
CDS (coding)                      :: 2,921
Genes (RNA)                       :: 89
rRNAs                             :: 6, 6, 6 (5S, 16S, 23S)
complete rRNAs                    :: 6, 6, 6 (5S, 16S, 23S)
tRNAs                             :: 67
ncRNAs                            :: 4
Pseudo Genes (total)              :: 24
Pseudo Genes (ambiguous residues) :: 0 of 24
Pseudo Genes (frameshifted)       :: 9 of 24
Pseudo Genes (incomplete)         :: 6 of 24
Pseudo Genes (internal stop)      :: 9 of 24
##Genome-Annotation-Data-END##
```

**Figure 6.** A summary of PGAP genome annotation process is provided in the COMMENT section of GenBank and RefSeq records. The example is given for *Listeria monocytogenes* strain CFSAN010068, complete genome NZ_CP014250.1.

annotation. Minimally complete annotation must contain the following genes with locus-tags: (i) at least one copy of rRNAs genes (5S, 16S, 23S); (ii) at least one copy of tRNAs for each amino acid; and (iii) for all protein coding genes with a corresponding CDS, protein naming should follow UniProt guidelines available at http://www.uniprot.org/docs/proknameprot.

The annotation process in PGAP incorporates multiple QC checks within the annotation pipeline itself. Results of these QC metrics are made available to curators and are used in automated decisions to accept or reject annotation based on quality standards. For GenBank submission, quality requirements are permissive to allow the majority of genome annotations to pass; only annotations with serious quality problems are filtered out. For RefSeq submission, rules are stricter, permitting RefSeq to represent a more trusted set of annotation of higher quality genomes.

## RESULTS

### Case study: comparison of PGAP generated annotations with GenBank genome annotations

We have executed PGAP on the GenBank versions of genome assemblies of the following eight species: *Bacillus subtilis subsp. subtilis str. 168* (GCA_000009045.1); *Chlamydia trachomatis D/UW-3/CX* (GCA_000008725.1); *E. coli str. K 12 substr. MG1655* (GCA_000005845.2); *Francisella tularensis* (GCA_000008985.1); *Mycobacterium leprae* (GCA_000195855.1); *Mycobacterium tuberculosis H37Rv* (GCA_000195955.2); *Neisseria meningitidis MC58* (GCA_000008805.1); *Pseudomonas aeruginosa PAO1* (GCA_000006765.1); *Salmonella enterica subsp. enterica serovar Typhimurium str. LT2* (GCA_000006945.1); and *Yersinia pestis CO92* (GCA_000009065.1). This selection of genomes represents a cross-section of organism types in high quality finished assemblies. While the annotation on these genomes is not perfect, these 10 examples are among the highest quality available annotation sets curated to community-accepted standards.

Comparison of structural annotations generated through PGAP pipeline with the GenBank annotations (Table 2) indicates that PGAP can automatically match 3′ ends of genes in more than 98% of cases on average (98.2%). Sources of differences could be related to errors either in automatic or in GenBank annotation. For instance, some gene fusion and fission events may lead to predicting some genes as pseudogenes, which could potentially decrease the number of genes matched at the 3′ end. Arguably, a difference of more than 2% may indicate some issues with the GenBank record (such as absence of continuous curation; e.g. the last updates of the GenBank annotation records of *N. meningitidis MC58* and *B. subtilis* were made in 2005 and 2009, respectively). Based on curator review at NCBI, some genes present in the GenBank annotations of the 10 genomes were removed in the RefSeq annotation versions.

Comparing full annotated CDS matches at both 5′ and 3′ ends, PGAP matches GenBank annotation in 89.9% of cases. Of note, gene start positions (5′ ends) are considered less accurately annotated than 3′ ends. Therefore, some fraction of mismatch in start positions can also be attributed to less than perfect GenBank annotation.

### GenBank submission service

By design, the PGAP pipeline aims to provide accurate, robust automated annotation. To further aims of generating consistent annotation, NCBI offers execution of the PGAP pipeline for all prokaryotic GenBank submissions, and it

**Table 2.** Comparison of the genome annotations generated by PGAP-3.1 with the GenBank annotations of the same genomes (snapshot from February 2016)

| Data point | # of genes in GenBank annotation | # of predictions matching annotation in 3′ end | % of annotated genes missed in predictions | # of predictions matching annotation in 5′ and 3′ ends | % of predictions with mismatch in 5′ end | # of hypothetical genes in GenBank annotation |
|---|---|---|---|---|---|---|
| *Bacillus subtilis* | 4185 | 4044 | 3.4 | 3768 | 6.8 | 232 |
| *Chlamydia trachomatis* | 892 | 886 | 0.7 | 822 | 7.2 | 285 |
| *E. coli* | 4140 | 4093 | 1.1 | 3915 | 4.3 | 0 |
| *Franciscella tularensis* | 1602 | 1589 | 0.8 | 1330 | 16.3 | 202 |
| *Mycobacterium leprae* | 1605 | 1599 | 0.4 | 1391 | 13.0 | 14 |
| *Mycobacterium tuberculosis* | 4018 | 3954 | 1.6 | 3342 | 15.5 | 508 |
| *Neisseria meningitidis* | 2063 | 1958 | 5.1 | 1705 | 12.9 | 529 |
| *Pseudomonas aeruginosa* | 5571 | 5531 | 0.7 | 5051 | 8.7 | 1693 |
| *Salmonella enterica* | 4554 | 4485 | 1.5 | 4031 | 10.1 | 10 |
| *Yersinia pestis* | 4083 | 4031 | 1.3 | 3429 | 14.9 | 332 |

is integrated into the submission system. GenBank submission standards require genomic sequences to meet specific quality levels. Submitted sequences should pass contamination screening to eliminate known foreign sequences; the sequences require proper formatting and attribution and should contain information necessary for annotation, including organism/taxonomic information, genetic code and unique locus-tag prefix. More details on GenBank submission standards are provided at https://www.ncbi.nlm.nih.gov/genome/annotation_prok/.

### RefSeq

Historically, RefSeq annotation of prokaryotic genomes relied on submitter-supplied annotation available in public archives. In recent years, several papers and working groups have identified cases of inconsistent structural and functional annotation (31,32). It was demonstrated that annotation of closely related genomes may vary in number of coding genes, positions of gene starts and assignment of protein function. Inconsistent annotation was observed also beyond protein coding genes; annotation of RNA genes as well as pseudogenes could be inconsistent or entirely absent. RefSeq quality control rules include more stringent criteria than the rules of GenBank submission. Genomes that do not pass quality control are not accepted into RefSeq. Of note, RefSeq rules include comparative analysis of all genomes in a clade (33). To improve consistency of annotation all prokaryotic RefSeq genomes were re-annotated by PGAP in December 2014 through March 2015 (for details see https://ncbi.nlm.nih.gov/refseq/about/prokaryotes/reannotation/). Subsequent to this, RefSeq quality criteria were strengthened resulting in removal of some genomes, and all RefSeq prokaryotic genomes were re-annotated again in August 2015 using a single software version, PGAP-3.0 (see release notes https://ncbi.nlm.nih.gov/genome/annotation_prok/release_notes/) to further improve consistency.

### Community curated genomes and genes

Some bacterial species with significant community interest have been manually curated by expert biologists. Annotation of these genomes is evaluated by the RefSeq curators and is updated as new information becomes available from community experts. Additional efforts are planned to strengthen collaboration with the research community in order to provide accurate and up-to-date annotation for high interest species and metabolic pathways. Active community curated species include *E. coli K-12 MG165* (34), *Mycobacterium tuberculosis* strains (35) and *Pseudomonas aeruginosa* strains (36).

### New data model for proteins

As a result of deep sequencing of multiple strains of pathogens, several prokaryotic clades grew significantly and created a flood of identical proteins. To decrease redundancy in annotated proteins, particularly bacterial proteins, the RefSeq collection introduced a new protein data type signified by a 'WP' accession prefix. The new protein record type represents a group of identical protein sequences annotated in genomes of various isolates, strains or species. This data type is managed independently of the genome sequence record. More details are available at https://ncbi.nlm.nih.gov/refseq/about/nonredundantproteins/.

## CONCLUSIONS

The PGAP pipeline, in its new implementation, uses multiple approaches to annotation to strike a balance in performance whether dealing with well-known bacterial clades or with less-well-described taxonomic lineages. Combining the best features of the pan-genome approach in highly abundant clades with well-described and well-tested *ab initio* methods, PGAP now presents a flexible and extensible framework for prokaryotic annotation needs. As shown in Figure 7, the majority of predicted prokaryotic protein-coding genes are supported by homology to known proteins. Two factors explain this observation. First, in large clades, the corpus of 'core' genes can support 70–80% of the average number of genes in each clade member (Table 1; e.g. *S. aureus, S. enterica, L. monocytogenes, V. cholera* in Figure 7). Second, in small clades, while the number of isolates may not be sufficient to calculate pan-genome 'core' genes and proteins, the number of conserved gene and protein families defined within 'expanded' clades that amount to higher level taxonomic units can still be high (e.g. *S. globosa* in Figure

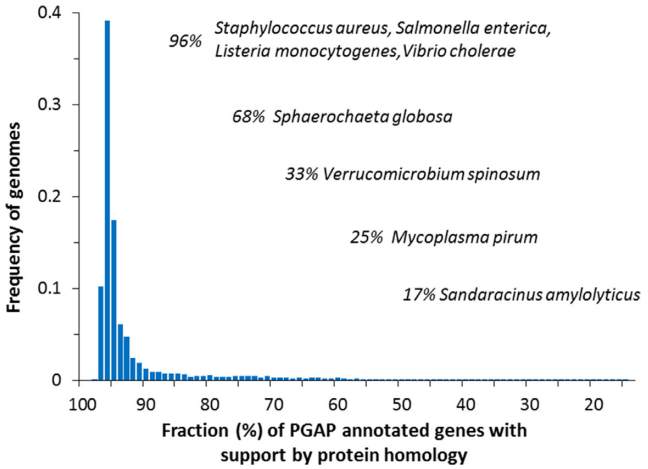


**Figure 7.** Frequency histogram of genomes with respect to the fraction of the whole complement of genes supported by similarity to proteins in RefSeq. In about 50% of the total set of genomes in consideration, mostly from highly populated clades, more than 95% of protein-coding genes are supported by protein sequence similarity.

7). Of note, *ab initio* gene prediction methods can contribute most in annotation of genomes from novel taxonomic lineages (e.g. *V. sominosum*, *M. pirum*, *S. amilolyticus* in Figure 7). While prediction of protein coding genes is of high priority, identification of non-protein-coding elements cannot be overlooked. PGAP incorporates robust tools developed by the community for prediction of such elements, and accurately combines this information with protein coding elements.

The new PGAP version employs a robust automatic system optimized for high throughput. PGAP is able to annotate more than 1200 genomes a day, an increase of two orders of magnitude over its semi-automatic predecessor. As the tide of prokaryotic submissions is ever-increasing, NCBI will continue to focus on throughput and accuracy to meet the needs of the community. More than 8000 new GenBank submissions were annotated by PGAP by the end of February 2015; also, more than 30 000 RefSeq genomes were re-annotated (RefSeq release 70). The latest RefSeq release 74 (January 2016) includes 54 242 annotated prokaryotic genomes.

NCBI's PGAP is an evolving system. This manuscript describes major features implemented in PGAP version 2.0–3.1. In July 2015, additional features were implemented and released in version 3.0, and version 3.1 of PGAP went into production in January 2016 (see for details: https://www.ncbi.nlm.nih.gov/genome/annotation_prok/release_notes/). The pipeline will continue to be extended with new features and novel algorithms as research into automated annotation continues.

## ACKNOWLEDGEMENTS

The authors would like to thank Dr David Lipman for many fruitful discussions about prokaryotic biology, insightful suggestions on improving the annotation results and his continuous support for developing the NCBI prokaryotic annotation pipeline.

## FUNDING

Intramural Research Program of the NIH National Library of Medicine (in part); the work of M.B. and A.L. was supported in part by NIH grant HG000783 to M.B. Funding for open access charge: Intramural Research Program of the NIH National Library of Medicine.
*Conflict of interest statement.* None declared.

## REFERENCES

1. Besemer,J., Lomsadze,A. and Borodovsky,M. (2001) GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res.*, **26**, 1107–1115.
2. Delcher,A.L., Harmon,D., Kasif,S., White,O. and Salzberg,S.L. (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.*, **23**, 4636–4641.
3. Tatusov,R.L., Natale,D.A., Garkavtsev,I.V., Tatusova,T.A., Shankavaram,U.T., Rao,B.S., Kiryutin,B., Galperin,M.Y., Fedorova,N.D. and Koonin,E.V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.*, **29**, 22–28.
4. Klimke,W., Agarwala,R., Badretdin,A., Chetvernin,S., Ciufo,S., Fedorov,B., Kiryutin,B., O'Neill,K., Resch,W., Resenchuk,S. *et al.* (2009) The National Center for Biotechnology Information's Protein Clusters Database. *Nucleic Acids Res.*, **37**, D216–D223.
5. Nawrocki,E.P. and Eddy,S.R. (2013) Infernal 1.1: 100-fold faster RNA Homology searches. *Bioinformatics*, **29**, 2933–2935.
6. Nawrocki,E.P. (2014) Annotating functional RNAs in genomes using Infernal. *Methods Mol. Biol.*, **1097**, 163–197.
7. Nawrocki,E.P., Burge,S.W., Bateman,A., Daub,J., Eberhardt,R.Y., Eddy,S.R., Floden,E.W., Gardner,P.P., Jones,T.A., Tate,J. *et al.* (2015) Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.*, **43**, D130–D138.
8. Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.
9. Angiuoli,S., Gussman,A., Klimke,W., Cochrane,G., Field,D., Garrity,G., Kodira,C.D., Kyrpides,N., Madupu,R., Markowitz,V. *et al.* (2008) Toward an online repository of Standard Operating Procedures (SOPs) for (meta) genomic annotation. *OMICS*, **12**, 137–141.
10. Overbeek,R., Olson,R., Pusch,G.D., Olsen,G.J., Davis,J.J., Disz,T., Edwards,R.A., Gerdes,S., Parrello,B., Shukla,M. *et al.* (2014) The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res.*, **42**, D206–D214.
11. Seemann,T. (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, **30**, 2068–2069.
12. Pati,A., Ivanova,N.N., Mikhailova,N., Ovchinnikova,G., Hooper,S.D., Lykidis,A. and Kyrpides,N.C. (2010) GenePRIMP: a gene prediction improvement pipeline for prokaryotic genomes. *Nat. Methods*, **7**, 455–457.
13. Pareja-Tobes,P., Manrique,M., Pareja-Tobes,E., Pareja,E. and Tobes,R. (2012) BG7: a new approach for bacterial genome annotation designed for next generation sequencing data. *PLoS One*, **7**, e49239.
14. Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlavage,A.R., Bult,C.J., Tomb,J.F., Dougherty,B.A., Merrick,J.M. *et al.* (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, **269**, 496–512.
15. Tettelin,H., Masignani,V., Cieslewicz,M.J., Donati,C., Medini,D., Ward,N.L., Angiuoli,S.V., J.,Crabtree., Jones,A.L., Durkin,A.S. *et al.* (2005) Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial 'pan-genome'. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 13950–13955.
16. Mira,A., Martín-Cuadrado,A.B., D'Auria,G. and Rodríguez-Valera,F. (2010) The bacterial pan-genome: a new paradigm in microbiology. *Int. Microbiol.*, **13**, 45–57.
17. Zaslavsky,L., Ciufo,S., Fedorov,B., Kiryutin,B., Tolstoy,I. and Tatusova,T. (2016) Dealing with the data deluge: new strategies in prokaryotic genome analysis. In: Kulski,JK (ed.) *Next Generation Sequencing - Advances, Applications and Challenges*. InTech, Croatia.

18. Edgar,R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.
19. Zaslavsky,L. and Tatusova,T. (2015) Clustering analysis of proteins from microbial genomes at multiple levels of resolution. In: *Bioinformatics Research and Applications: 11th International Symposium, ISBRA 2015, June 7-10, Proceedings*. Norfolk, pp. 438–440.
20. Ciccarelli,F.D., Doerks,T., von Mering,C., Creevey,C.J., Snel,B. and Bork,P. (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science*, **311**, 1283–1287.
21. Mende,D.R., Sunagawa,S., Zeller,G. and Bork,P. (2013) Accurate and universal delineation of prokaryotic species. *Nat. Methods*, **10**, 881–884.
22. Tatusova,T., Ciufo,S., Federhen,S., Fedorov,B., McVeigh,R., O'Neill,K., Tolstoy,I. and Zaslavsky,L. (2015) Update on RefSeq microbial genome resources. *Nucleic Acids Res.*, **43**, D599–D605.
23. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
24. Grissa,I., Vergnaud,G. and Pourcel,C. (2007) The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics*, **8**, 172.
25. Makarova,K.S., Haft,D.H., Barrangou,R., Brouns,S.J., Charpentier,E., Horvath,P., Moineau,S., Mojica,F.J., Wolf,Y.I., Yakunin,A.F. *et al.* (2011) Evolution and classification of the CRISPR-Cas systems. *Nat. Rev. Microbiol.*, **9**, 467–467.
26. Bland,C., Ramsey,T.L., Sabree,F., Lowe,M., Brown,K., Kyrpides,N.C. and Hugenholtz,P. (2007) CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics*, **8**, 209.
27. Biswas,A., Gagnon,J.N., Brouns,S.J., Fineran,P.C. and Brown,C.M. (2013) CRISPRTarget: Bioinformatic prediction and analysis of crRNA targets. *RNA Biol.*, **10**, 817–827.
28. Lukashin,A.V. and Borodovsky,M. (1998) GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.*, **26**, 1107–1115.
29. UniProt Consortium (2015) UniProt: a hub for protein information. *Nucleic Acids Res.*, **43**, D204–D212.
30. Klimke,W., O'Donovan,C., White,O., Brister,J.R., Clark,K., Fedorov,B., Mizrachi,I., Pruitt,K.D. and Tatusova,T. (2011) Solving the problem: Genome annotation standards before the data deluge. *Stand. Genomic Sci.*, **5**, 168–193.
31. Poptsova,M.S. and Gogarten,J.P. (2010) Using comparative genome analysis to identify problems in annotated microbial genomes. *Microbiology*, **156**, 1909–1917.
32. Radivojac,P., Clark,W.T., Oron,T.R., Schnoes,A.M., Wittkop,T., Sokolov,A., Graim,K., Funk,C., Verspoor,K., Ben-Hur,A. *et al.* (2013) A large-scale evaluation of computational protein function prediction. *Nat. Methods*, **10**, 221–227.
33. Tatusova,T., Ciufo,S., Fedorov,B., O'Neill,K. and Tolstoy,I. (2014) RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic Acids Res.*, **42**, D553–D559.
34. Zhou,J., Richardson,A.J. and Rudd,K.E. (2013) EcoGene-RefSeq: EcoGene tools applied to the RefSeq prokaryotic genomes. *Bioinformatics*, **29**, 1917–1918.
35. Lew,J.M., Mao,C., Shukla,M., Warren,A., Will,R., Kuznetsov,D., Xenarios,I., Robertson,B.D., Gordon,S.V., Schnappinger,D. *et al.* (2013) Database resources for the tuberculosis community. *Tuberculosis (Edinb)*, **93**, 12–17.
36. Winsor,G.L., Lam,D.K., Fleming,L., Lo,R., Whiteside,M.D., Yu,N.Y., Hancock,R.E. and Brinkman,F.S. (2011) Pseudomonas Genome Database: improved comparative analysis and population genomics capability for Pseudomonas genomes. *Nucleic Acids Res.*, **39**, D596–D600.