Published online 26 May 2021

NAR Genomics and Bioinformatics, 2021, Vol. 3, No. 2 **1**
doi: 10.1093/nargab/lqab047

GeneMark-HM: improving gene prediction in DNA sequences of human microbiome

Alexandre Lomsadze[1], Christophe Bonny[2], Francesco Strozzi[2,†] and Mark Borodovsky [1,3,4,*,†]

[1]Gene Probe, Inc., 1106 Wrights Mill Ct, Atlanta, GA 30324, USA, [2]Enterome, 94/96 avenue Ledru-Rollin, 75011 Paris, France, [3]Wallace H. Coulter Department of Biomedical Engineering, Atlanta, GA 30332, USA and [4]School of Computational Science and Engineering, Georgia Tech, Atlanta, GA 30332, USA

Received November 22, 2020; Revised April 27, 2021; Editorial Decision May 06, 2021; Accepted May 24, 2021

ABSTRACT

Computational reconstruction of nearly complete genomes from metagenomic reads may identify thousands of new uncultured candidate bacterial species. We have shown that reconstructed prokaryotic genomes along with genomes of sequenced microbial isolates can be used to support more accurate gene prediction in novel *metagenomic* sequences. We have proposed an approach that used three types of gene prediction algorithms and found for all contigs in a metagenome nearly optimal models of protein-coding regions either in libraries of precomputed models or constructed de novo. The model selection process and gene annotation were done by the new GeneMark-HM pipeline. We have created a database of the species level pan-genomes for the human microbiome. To create a library of models representing each pan-genome we used a self-training algorithm GeneMarkS-2. Genes initially predicted in each contig served as queries for a fast similarity search through the pan-genome database. The best matches led to selection of the model for gene prediction. Contigs not assigned to pan-genomes were analyzed by crude, but still accurate models designed for sequences with particular GC compositions. Tests of GeneMark-HM on simulated metagenomes demonstrated improvement in gene annotation of human metagenomic sequences in comparison with the current state-of-the-art gene prediction tools.

INTRODUCTION

In recent years a number of sequenced microbial isolates has increased significantly; >310 000 annotated prokaryotic genomes, with the total length ∼1 Tb nt, were listed in GenBank as of February 2021 (www.ncbi.nlm.nih.gov/genome/browse#!/prokaryotes). The number of metagenomes with assembled contigs in the WGS section of GenBank has already reached above 75 000, with the total length of nucleotide sequences exceeding ∼2.1 Tb nt (www.ncbi.nlm.nih.gov/Traces/wgs/?search=metagenome). Recent improvements in sequencing technology and assembly methods have led to increase in the average length of metagenomic contigs (1). Still, a vast majority of them are not long enough to allow for estimation of parameters of the 'contig-specific' high order model with precision typical for the algorithms trained on complete or nearly complete prokaryotic genomes (2–5). Therefore, in gene finders working with short sequences with unknown genomic context special methods for finding effective model parameters have to compensate the lack of information. Several algorithms for gene prediction in metagenomic sequences have been developed: (6–10). Also, recently developed methods of reconstruction of nearly complete genomes from metagenomic datasets added to the picture thousands of genomes of uncultured candidate bacterial species.

The goal of development of the GeneMark-HM pipeline (HM stands for human microbiome) was to improve accuracy of gene prediction in metagenomic contigs, those originated from a shotgun sequencing of human microbiome (11,12), up to a level achievable by gene finding algorithms trained on complete prokaryotic genomes.

Importantly, genomes of sequenced isolates and reconstructed prokaryotic genomes could be used to support more accurate gene prediction in novel metagenomes. The main idea is as follows. Many metagenomic contigs may belong to organisms whose very close relatives have sequenced or nearly sequenced genomes. Still, many other contigs may belong to organisms distant from any known species. We argue that nearly optimal models of protein-coding regions could be determined for annotation of a vast majority of contigs in a metagenome. The solution of the annotation problem is implemented in GeneMark-HM, a pipeline for metagenomic contigs originated from human microbiome.

*To whom correspondence should be addressed. Tel: +1 404 667 8183; Email: borodovsky@gatech.edu
†The authors wish it to be known that, in their opinion, the last two authors should be regarded as Joint Senior Authors.

The pipeline is using three gene prediction algorithms along with two reference databases (Figure 1).

The pipeline components are (i) the newest version of MetaGeneMark, an algorithm using GC content specific models of protein-coding regions (6,5,10). This tool was designed for analysis of anonymous sequence fragments of arbitrary length, even as short as 400nt; (ii) GeneMarkS-2 (5) a self-training algorithm for gene prediction in sufficiently long sequences, up to complete genomes; (iii) GeneMark.hmm-2, an efficient algorithm that requires a model defined externally or by self-training. GeneMark.hmm-2 is an indispensable part of GeneMarkS-2 and the newest version of MetaGeneMark; (iv) a database of pan-genomes of the human microbiome; (v) a database of pan-genome specific models of protein-coding regions; the models are pre-computed by running GeneMarkS-2 on a *representative* genome of each pan-genome. The GeneMark.hmm-2 algorithm was described in the GeneMarkS-2 publication (5).

All contigs in a metagenome are effectively divided into three groups: (i) contigs that can be analyzed by GeneMark-hmm-2 with models from genomes of close relatives, (ii) contigs with no close relatives with sequenced genomes which are sufficiently long to run self-training GeneMarkS-2 and (iii) short anonymous contigs analyzed by MetaGeneMark with pre-computed GC specific models.

The database of the species level pan-genomes was built in the gene space, a non-redundant representation of the entire set of genes of a set of strains of the same species (13). To create a database of *representative* models for the pan-genomes we used the self-training GeneMarkS-2 running on a representative genome of a clade. The genes initially predicted in each metagenomic contig were used for a fast similarity search against the database of genes of pan-genomes. The top hits, if the similarity existed, were used to identify a model for GeneMark.hmm-2 that would make more precise gene predictions. Otherwise, depending on a length of the contig, the GeneMarkS-2 or MetaGeneMark were the right tools to have genes predicted in the anonymous contig (Figure 1).

Testing of GeneMark-HM on simulated metagenomes demonstrated improvement in accuracy of gene annotation in metagenomic sequences from human microbiome in comparison with the current state-of-the-art gene prediction tools.

MATERIALS AND METHODS

Sequence data

Recently as many as 154 723 genomes of prokaryotic species were reconstructed from sequence reads of human metagenomes (14). These genomes, called rGenomes (or MAGs, metagenome assembled genomes), were clustered into 4930 species-level genome bins (SGBs). Almost 2000 SGBs in this collection contained a single rGenome (Supplementary Figure S1 shows a frequency histogram of the number of rGenomes per SGB). Notably, the rGenomes are fairly fragmented, e.g. contigs under 10,000 nt account for 27% of total volume of rGenomes (Supplementary Figure S2). *Representative* rGenomes selected for each SGB in (14), normally the largest in length, varied in total length and GC content (Supplementary Figure S3). For 23% of SGBs, it was observed that a sequenced genome of a prokaryotic species could be found within less than 5% ANI (average nucleotide identity) distance from one of the rGenomes in the SGB. Thus, such SGBs are called 'known SGBs', [kSGBs]. Remaining 77% or 3796 SGBs were deemed to be genome bins of strains of new species and were named 'unknown SGBs' [uSGBs] (14).

Besides the use of SGB rGenomes, we retrieved from the NCBI database genomes of those bacterial and archaeal species that were reported to inhabit human guts (15–20,12,21,22). We grouped genomes of strains having the same Taxa ID into six archaeal and 695 bacterial genome clusters (tClusters). Annotation of the *representative* genomes of these 701 tClusters (Supplementary Figure S4) was available in RefSeq (23). The tClusters and the SGBs could overlap. Merging of the tClusters and the SGBs produced 4930 updated SGBs along with 152 tClusters that did not overlap any of the SGBs (see Supplementary Materials). We argue that after such merging the genomes of strains of a single prokaryotic species would be contained in a single updated SGB or in a single tCluster.

Construction of species level pan-genomes

Each pan-genome was built in the gene space. A pan-genome represented gene sequences from all the genomes from a particular tCluster or SGB. The purpose of making pan-genomes was the reduction of a reference database of gene sequences. The procedure was implemented as follows. In each SGB (tCluster), all the genome sequences were processed by GeneMark.hmm-2 to predict genes. In this analysis one and the same model of protein-coding region, the SGB (tCluster) specific model was used (see below).

The initial set of gene sequences for the pan-genome of an SGB (tCluster) was a whole complement of genes predicted in the SGB (tCluster) *representative genome* (Figure 2). Then the construction process continued by taking one gene set after another in order of diminishing length of the rGenomes (genomes) from a given SGB (tCluster). To decide which genes from a particular gene set should be added to the growing pan-genome, the sequences of genes were used as queries in the BLASTn search against the current set of the pan-genome genes. The search was run in the fast MegaBLAST mode (24). If a given query and its highest scoring target in a pan-genome had an average nucleotide identity (ANI) above 95% (25), along with >90% coverage of the query in the pairwise alignment, then the query gene was deemed to be a close homologue of the target and was not added to the pan-genome. Otherwise, the query was added to the pan-genome as an entry increasing the pan-genome diversity. The process continued until all the genes in all the rGenomes (genomes) were screened. The pan-genome building procedure was repeated for all the SGBs and tClusters. To limit the size of a pan-genome for a given SGB (tCluster) without a noticeable loss of sensitivity in a downstream analysis, only gene sets from the 150 longest rGenomes (genomes) were used in the pan-genome construction. As it was expected, the number of genes in a

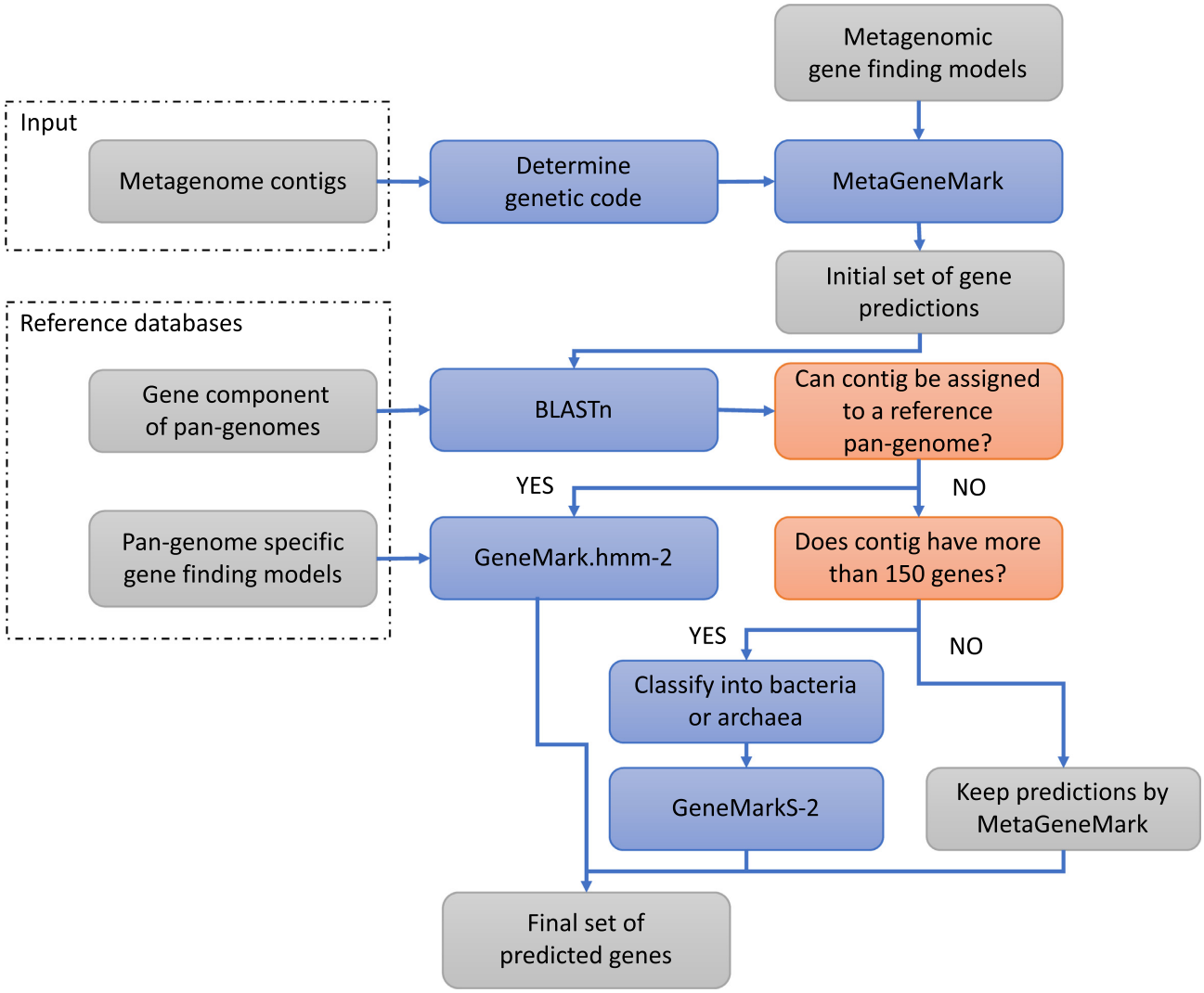

Figure 1. Principal flowchart of the GeneMark-HM pipeline.

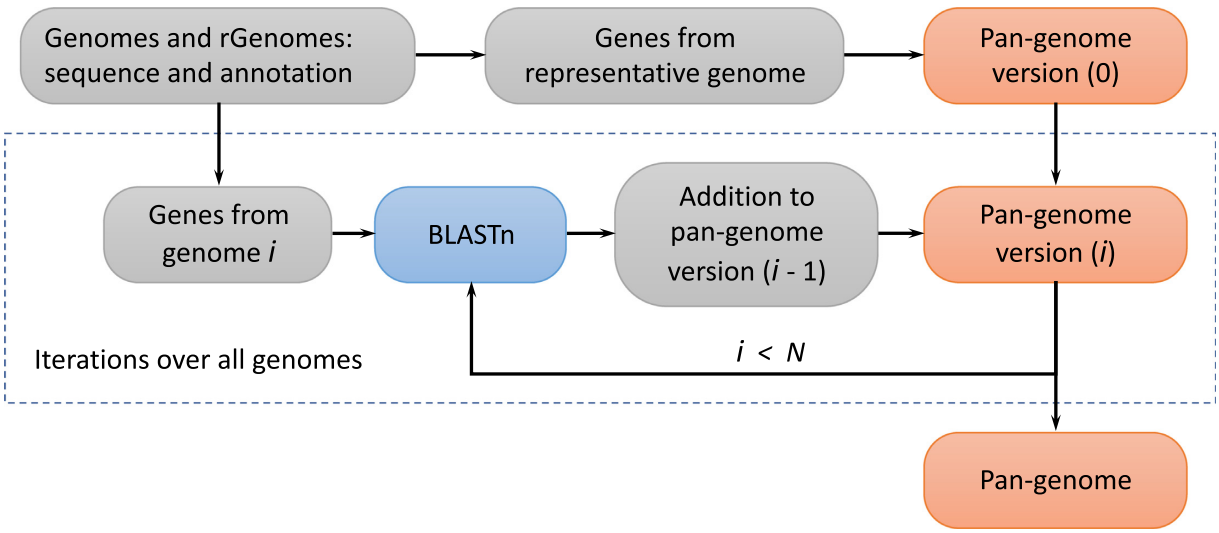

Figure 2. A species level pan-genome construction in gene space.

pan-genome was significantly smaller than the total number of genes predicted in the original rGenomes (genomes) of the SGBs (tClusters).

For gene prediction in human metagenomes, we created a 5082 species level reference pan-genomes.

Models of protein-coding regions and regulatory signals

A *species specific* model of a protein-coding region, formalized as an inhomogeneous Markov chain of an order k (26), is a key component of *ab initio* gene finding algorithms for prokaryotic and eukaryotic genomes (27,5,28). The transition probability values, the model parameters, are derived from the three phase positional frequencies of occurrences of nucleotide (k + 1)-mers in a manually curated, or automatically compiled, set of protein-coding regions. Particularly, these parameters are determined by self-training of GeneMarkS-2 on an anonymous prokaryotic genomic sequence (5). Notably, besides a model of protein-coding region, a full set of parameters of the generalized HMM (hidden Markov model) of a genomic sequence includes a model of non-coding sequence as well as a model of the gene start region with a ribosome binding site (RBS) or a promoter region located close to gene start in case of leaderless transcription (5).

Representative models. The GeneMarkS-2 model trained on a genome *representing* a particular SGB or tCluster we call a *representative* model. Given the evolutionary closeness of the strains comprising an SGB or tCluster we expect that the GeneMark.hmm-2 algorithm using a single *representative* model will be sufficiently accurate in predicting genes in metagenomic contigs of all the strains from this SGB or tCluster. In analysis of metagenomic contigs GeneMark.hmm-2 uses contig-specific *representative* models associated with a contig by a 'matching algorithm'. For contigs 'without match' the algorithm uses either de novo derived GeneMarkS-2 model or, for too short contigs, a pre-computed GC-specific model. We tested the gene prediction accuracy of GeneMark-hmm-2 with the *representative* model in the following way (Figure 3). GeneMarkS-2 was self-trained on each rGenome from an SGB, thus making a set of the rGenome-specific *native* models. Then the sets of gene predictions made for a given rGenome by GeneMark.hmm-2 with the *representative* model and with the *native* model were compared.

The results of the accuracy test were quite satisfactory as it could be seen in the Results section. Therefore, GeneMark.hmm-2 with the SGB (tCluster) specific *representative* model was an accurate tool for gene prediction in metagenomic contigs assigned to the pan-genome corresponding to an SGB or tCluster. The SGB (tCluster) specific gene prediction models were created for each of the corresponding 5,082 pan-genomes.

Notably, in a process of metagenome annotation some contigs may not be assigned to any pan-genome. Then, the gene prediction is done as follows (Figure 1). If the contig is relatively short (see more details below), the genes are predicted by MetaGeneMark with a pre-computed *GC specific models* (10). If the contig is long, the model parameters estimation and the simultaneous gene prediction is done by GeneMarkS-2 (5).

Initial gene prediction in an anonymous contig

For accurate gene prediction, GeneMark-HM had to recognize the type of genetic code of an anonymous contig. Most frequent genetic code was one with the three standard stop codons TAA, TAG, TGA (genetic code 11), while the second in frequency was the one with two stop codons, TAA and TAG, while TGA was reassigned to code for either Tryptophan (genetic code 4) or Glycine (genetic code 25) (www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi). Other types of genetic code, e.g. with TAG or TAA reassignments, have not yet been observed in prokaryotes (29). Parameters of the model of protein-coding region change depending on the role of TGA.

GeneMark-HM determines the role of the TGA triplet by the following approach. The GC-specific models of MetaGeneMark are diversified not only by 1% increment in GC content of contigs but also by the role of TGA (genetic code 11 or 4/25). Therefore, at the first stage of analysis, MetaGeneMark is run twice on each contig, with models corresponding to one or another role of TGA. The log-odds scores of a contig computed in the two runs are used in a classifier trained on 10 000 nt long contigs with known genetic codes (Figure 4). Assessment of the classifier accuracy on sets of 10 000 nt long genomic fragments not overlapping with the training set demonstrated that the role of TGA (a stop codon or a sense codon) was correctly identified in 99.9% cases.

Gene predictions made with a predicted role of TGA are retained as the initial gene predictions in the contig.

Improvement of initial gene predictions

Each initially predicted gene is used as a query in BLASTn working in the MegaBLAST mode (24) for similarity search against the gene database of pan-genomes. Among the hits with scores having E-value better than 10^{-5} we select those in which at least 90 nt of the query sequence and the target in the BLASTn alignment has the ANI value above 95%. Such hits are used to assign the query gene to a pan-genome. If, for a given query, the ANI values of several BLASTn alignments exceed the 95% threshold, the target that has an alignment with the highest BLASTn score is selected. A query gene that has no high enough scores and, thus, is not assigned to a pan-genome is not used in the contig classification. A contig is assigned to a pan-genome to which the majority of the contig's genes have been assigned. Next, the *representative* model of the selected pan-genome is used by GeneMark.hmm-2 to update the initial predictions of protein-coding genes (Figure 1). For contigs not assigned to a pan-genome we proceed as follows. If the contig is relatively short, the number of genes initially predicted by MetaGeneMark is smaller than 150, then the set of genes initially predicted by MetaGeneMark is reported as the final annotation. Otherwise, the self-training GeneMarkS-2 runs on the contig to deliver the final gene predictions.

The above-mentioned restrictions, the alignment length >90 nt as well as the ANI > 95%, were justified as follows. We selected at random four rGenomes from a set of rGenomes not used in a pan-genome construction (any SGB with more than 154 rGenomes would contain such rGenomes). Genes predicted in these rGenomes were used

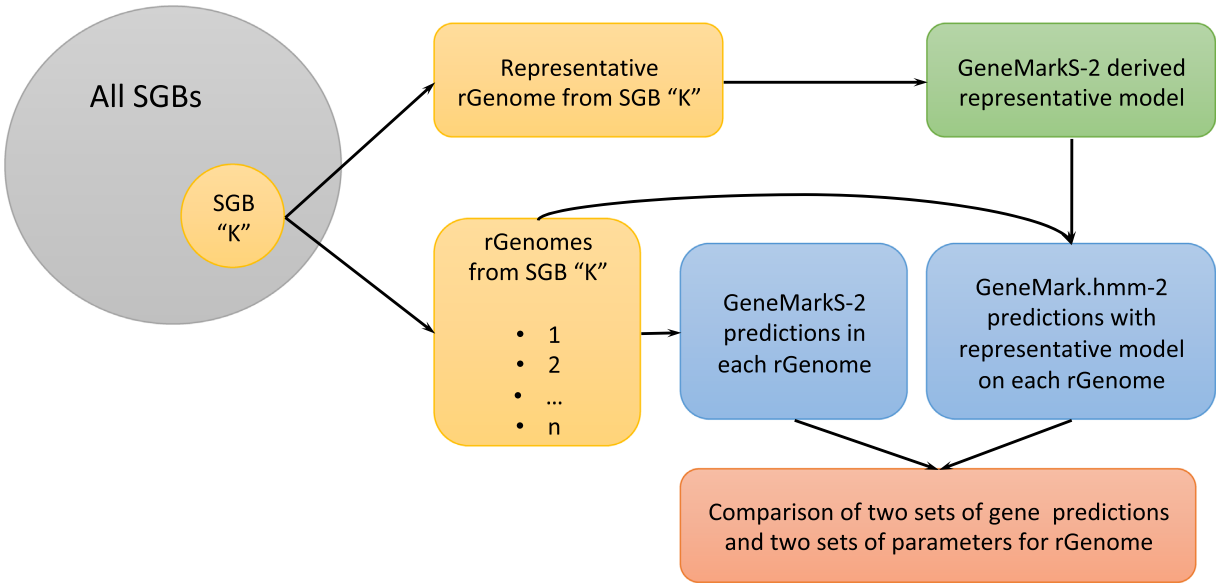

Figure 3. Assessment of accuracy of GeneMark.hmm-2 with the *SGB representative* models on rGenomes from the same SGBs. Gene predictions made by self-trained GeneMarkS-2 in a particular rGenome were considered as 'annotation'. The gene co-ordinates predicted by GeneMark.hmm-2 with the SGB *representative* model were compared with the annotation.

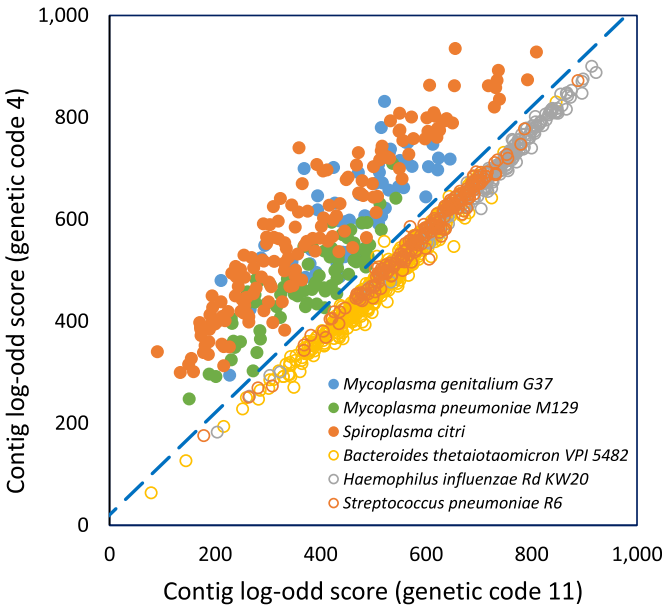

Figure 4. Separation of the log-odds score vectors computed for 10 000 nt contigs.

as queries in BLASTn search against all the pan-genomes. We observed that use of the top BLASTn hits without considering the length of alignment generated many false positives in the contig assignment to the pan-genome. Use of the combined restriction '>90 nt alignment length AND >95% ANI' produced sufficiently high sensitivity with low number of false positives.

Classification of long anonymous contigs as bacterial or archaeal

The GeneMarkS-2 branch of analysis of anonymous long contigs (Figure 1) has an extra step 'Classify into bacteria or archaea', a step made prior to application of GeneMarkS-2. The accuracy of GeneMarkS-2 improves if information about the species domain is provided. To make the domain assignment for a contig, GeneMark-HM uses results of the BLASTn similarity search for genes initially predicted in the contig. If majority of the significant hits (with *E*-value < 10^{-5}), that also have > 60% ANI with the targets and at least 90nt long alignments, have matches to genes of bacterial pan-genomes, the contig is classified as bacterial, otherwise, as archaeal. On contigs with no assignment to a domain, GeneMarkS-2 runs without a domain assignment (5).

The other branch of analysis, for contigs associated with pan-genomes, runs GeneMark.hmm-2 that uses *representative* models for which the type of domain is known from the SGB or tCluster specification. The domain information was used in the derivation of the *representative* model by GeneMarkS-2. In the MetaGeneMark branch analyzing short contigs, the domain of each predicted gene is determined, albeit not precisely, by the model, bacterial or archaeal, delivering higher log-odds score (5).

Tests on simulated metagenomes

Some SGBs were classified as 'known' SGBs (14) if there was a species whose genome in the NCBI database had a high ANI (>95%) with one of the rGenomes in the SGB. We used 625 NCBI genomes associated with 'known' SGBs to generate a set of contigs with lengths distribution like a contig length distribution observed in assembled rGenomes (Supplementary Figure S2). To construct the test set of annotated genes we selected from the 625 genomes only those genes whose protein products had similarity to proteins in

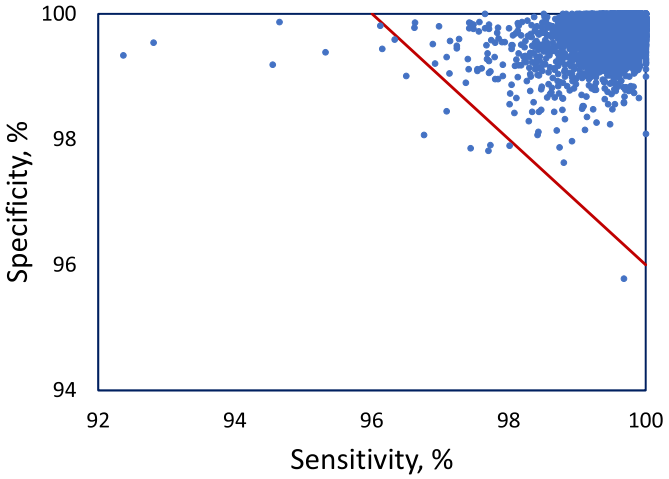

Figure 5. Gene prediction accuracy of GeneMark.hmm-2 with the SGB representative models was assessed on 6599 rGenomes selected from 1849 SGBs with two and more rGenomes. Excluded from the test sets were incomplete genes shorter than 200 nt.

other organisms; such genes were supposed to be more reliably annotated by PGAP (30). The genes co-ordinates determined by PGAP were transferred to the individual contigs. Minimum gene length of both the complete and incomplete genes in the test set was 90 nt. The final test set had 1,640,483 genes, with 83% complete and 17% were incomplete.

RESULTS

Evaluation of accuracy of GeneMark.hmm-2 with the representative models

By the procedure described in Methods, we assessed the accuracy of gene prediction by running GeneMark.hmm-2 with the SGB *representative* models on rGenomes from 1849 SGBs, those having two and more rGenomes. After computing numbers of true positives, false negatives and false positives (Tp, Fn and Fp, respectively) for each rGenome based on comparison with the 'native' model-based annotation we found that most of the time the (Sn + Sp)/2 value (Sn = Tp/(Tp + Fn), Sp = Tp/(Tp + Fp)) was above 98% (Figure 5). The accuracy below 98% (the dots below the red line) was observed in rGenomes whose fragmentation into short contigs was significantly higher than average. Similar results were observed in the same type of experiments with the *representative* models of tClusters.

Among the *representative* models determined for 4930 SGBs we observed the following frequency distribution of types of the regulatory signals, canonical and non-canonical RBS and promoters situated upstream to gene starts (Table 1).

All the three gene finding algorithms employed in GeneMark-HM had to make predictions of the genes with nucleotide composition deviated from a balk of the genes in each genome (atypical genes). This was done with the use of the GC content specific alternative models of protein-coding regions (5). We saw that in the *representative* rGenomes of 4930 SGBs the atypical genes could comprise

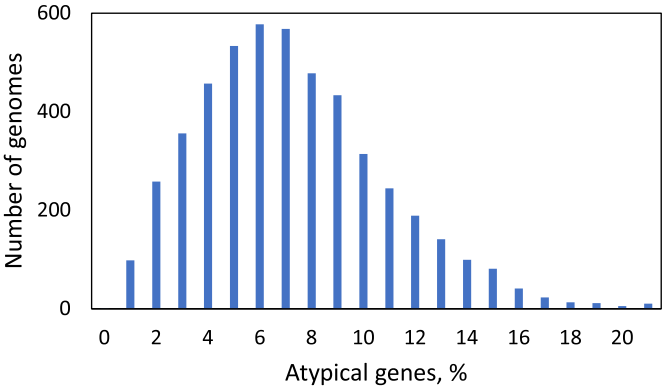

Figure 6. Distribution of the percentage of atypical genes in the *representative* rGenomes of 4930 SGBs.

up to 20% of all the genes in an rGenome (Figure 6); the mode of the distribution was at ∼6%.

Identification of a role of the TGA codon in a metagenomic contig

A role of TGA codon (code 11 or code 4/25) was identified for all the 4930 *representative* rGenomes. Just 18 of them had TGA coding for an amino acid. The predictions were confirmed by the MegaBLAST alignments of the eighteen relatively short (less 1 Mb nt in length) *representative* rGenomes to prokaryotic genomes in GenBank (31) (https://blast.ncbi.nlm.nih.gov/Blast.cgi). All the 18 *representative* rGenomes (with predicted codes 4 or 25) had best hits to genomes known to have genetic codes 4 or 25. Coversely, all the genomes with genetic code 4/25 from GenBank were aligned by MegaBLAST against 4912 rGenomes (4930 – 18). The alignments did not indicate a presence of any additional genomes with genetic code 4/25. Thus, all 4912 rGenomes were identified as ones with TGA coding for the end of translation (code 11).

In GeneMark-HM, a role of TGA in a contig was determined immediately upon the assignment of a contig to a pan-genome. If no such assignment to a pan-genome was made, the role of TGA was defined by the analysis of MetaGeneMark predictions as described in Methods. The accuracy of the TGA role prediction by MetaGeneMark was tested on contigs from the 18 *representative* rGenomes with genetic code 4/25 (a total of 1040 contigs) as well as on the contigs from the *representative* rGenomes with genetic code 11. A prediction was called false positive when a contig with TGA being a stop codon was identified as s contig with TGA coding for an amino acid. The same set of contigs was used to assess the accuracy of identification of the role of TGA by MetaProdigal (7).

If a contig does not carry genes containing TGA (either as a stop or a sense codon) then the result of gene prediction will not change regardless of what role of TGA has a model of a protein-coding region used for gene prediction in the contig Therefore, in addition to counting contigs with correctly and incorrectly predicted genetic code we counted the numbers of genes predicted with errors due to the TGA misinterpretation (Table 2). It was observed that MetaGeneMark made less errors (2170) in contigs with code

Table 1. Types of regulatory signals observed in the representative models for 4930 SGBs. Estimation of the type and parameters of the signal model was done in the GeneMarkS-2 self-training (5)

Description	Number of models	Genome group
Canonical RBS	3481	A
Non-canonical RBS	641	B
Leaderless – Promoters and RBS – bacteria	717	C
Leaderless – Promoters and RBS – archaea	12	D
Unclassified start signal	68	X

11 than MetaProdigal (25 729), (Table 2, top panel) and made more correct predictions than MetaProdigal in contigs with code 4, where TGA codes for Trp (Table 2, bottom panel).

Evaluation of the accuracy of gene prediction (as a whole)

We used 625 complete genomic sequences (see Methods) to create simulated metagenomes. Annotation of 1 640 483 genes made by the NCBI PGAP (30) was transferred to the metagenomics contigs. We emphasize that only a subset of all genes, those annotated with support of protein homology, was used and not the genes *ab initio* predicted by GeneMarkS-2, being a part of PGAP. Along with GeneMark-HM we have tested FragGeneScan, MetaGeneAnnotator, MetaProdigal and MetaGeneMark (Table 3). The results show that GeneMark-HM missed less genes in the test set than other gene finders; GeneMark-HM also made less errors in predictions of gene starts. Since the test set was a subset of the whole complement of genes in any given genome, accurate assessment of false positive rates was a difficult task. However, the results showed that GeneMark-HM predicted fewer overall number of genes than other gene finders, except MetaGeneAnnotator; a tool that had a higher threshold on minimal length of predicted genes. This observation suggests that the lower error rate in gene prediction (higher sensitivity) of GeneMark-HM was not a result of gene over-prediction.

Notably, GeneMark-HM was the only one tool in this comparison that was using a pan-genome database of reference genes. The average accuracy of gene predictions made by GeneMark-HM in each metagenome would, therefore, depend on the proportion of sequences originated from species distant from the ones carried in the reference database. To assess effects of this variation, we used the following approach. While we used one and the same simulated metagenome made from sequences of 625 genomes, we changed the size of the reference database of genes, D, which largest instance was the set of genes in all 5082 pan-genomes (Table 3).

Four instances of database D were as follows: (a) D4—genes of randomly selected 40% of the 625 genomes were excluded from the set of genes of 5082 pan-genomes; (b) D3—genes of randomly selected 30% of the 625 genomes were excluded; (c) D2—genes of randomly selected 20% of the 625 genomes were excluded; (d) D1— all 5082 pan-genomes were used without exclusions. As a result, we saw the percentage of genes predicted via the pan-genome path increasing as the database was getting larger (from a to d); at the same time the percentage of genes predicted by MetaGeneMark was decreasing.

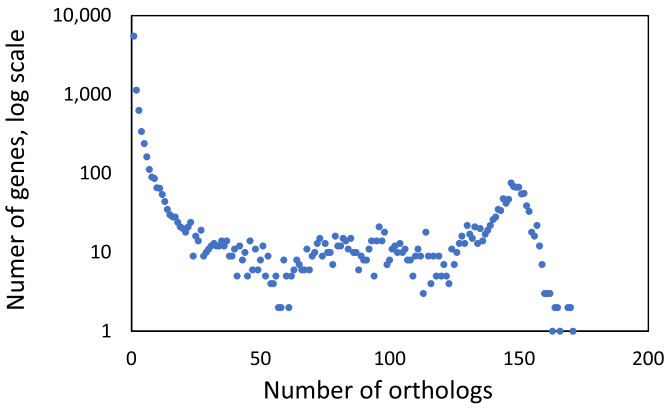

Figure 7. Distribution of frequency of sizes of the orthologous groups in SGB # 1472 (150 rGenomes out of total 200 rGenomes were used in the pan-genome construction).

In the set of 625 complete genomes, 410 were classified by GeneMarkS-2 (5) as genomes with the canonical Shine-Dalgarno RBS, 99 were classified as genomes with non-canonical RBS and 110 were classified as ones with frequent leaderless transcription. GeneMark-HM used the regulatory signal model in the gene start prediction. All over, GeneMark-HM had the lowest number of missed genes and incorrectly predicted gene starts in each class. We show the statistics of the gene start prediction errors for GeneMark-HM and for the runner-up, MetaProdigal (Table 4). The difference in the number of errors in gene start prediction was the smallest for genomes with canonical RBS and the largest for genomes with frequent leaderless translation initiation.

Computational complexity of running GeneMark-HM

Frequency distributions of sizes of the sets of orthologous genes in SGBs and tClusters did show a typical pattern illustrated here for SGB # 1472 (14), a large SGB with 150 rGenomes (Figure 7). A spike at the low numbers of orthologues corresponds to rather unique genes that appear in a single or in a few rGenomes (the highest peak relates to the truly unique genes). A spike at the large number of orthologues corresponds to universal genes present in each of 150 rGenomes. The middle part of the graph relates to disposable genes present in some rGenomes and absent in others.

Most of the time, upon construction of pan-genomes in gene space, the initial gene set defined by a *representative* rGenome was expanded three- or four-fold towards the end of the process (Supplementary Figure S5). However, in some pan-genomes the increase could be more significant (Supplementary Figure S6). To keep the size of a pan-

Table 2. Accuracy of genetic code identification by MetaGeneMark and MetaProdigal

932 969 contigs with TGA being a stop codon (code 11)		
Algorithm	# of contigs predicted to use code 4	# of genes with TGA incorrectly interpreted
MetaGeneMark	**1168**	**2170**
MetaProdigal	26 305	25 729
1040 contigs with TGA coding for Trp (code 4)		
Algorithm	# of contigs predicted to use code 4	# of genes with TGA correctly predicted as Trp
MetaGeneMark	**942**	**11 292**
MetaProdigal	820	9627

Table 3. Assessment of gene finding accuracy of GeneMark-HM on simulated metagenomic sequences. The shortest contig length was 1500 nt. The protein-coding regions reported by MetaGeneAnnotator and FragGeneScan were longer than 120 nt; the three other tools reported predicted coding regions longer than 90 nt

1 640 483 genes in the test set	# of missed genes	% of missed genes	total # of genes predicted	# of wrong starts	% of wrong starts
FragGeneScan	71 242	4.3	1 861 953	291 400	17.8
MetaGeneAnnotator	37 451	2.3	1 794 502	284 555	17.3
MetaProdigal	22 071	1.3	1 824 182	194 188	11.8
MetaGeneMark	21 861	1.3	1 830 949	330 985	20.2
(a) GeneMark-HM with D4 (81% in pan-genome path, 19% in MetaGeneMark path)	15 747	1.0	1 813 984	204 159	12.4
(b) GeneMark-HM with D3 (84% in pan-genome path, 16% in MetaGeneMark path)	15 409	0.9	1 813 830	199 871	12.2
(c) GeneMark-HM with D2 (88% in pan-genome path, 12% in MetaGeneMark path)	15 122	0.9	1 815 335	193 395	11.8
(d) GeneMark-HM with D1 (96% in pan-genome path, 4% in MetaGeneMark path)	**14 162**	**0.9**	1 817 669	**178 190**	**10.9**

Table 4. The gene start prediction errors in genomes of each group: (A) with canonical RBS, (B) with non-canonical RBS and (C) with frequent leaderless transcription

GeneMarkS-2 groups		A	B	C
Number of genes in each class		1 087 637	284 694	244 773
Missed genes	by MetaProdigal	14 332	3795	3561
	by GeneMark-HM	**8797**	**2760**	**2353**
Difference as % of all genes in the test set		0.51	0.36	0.49
Wrong starts	by MetaProdigal	115 201	35 127	40 263
	by GeneMark-HM	**106 993**	**31 370**	**36 374**
Difference as % of all genes in the test set		0.75	1.32	1.59

genome under control we had to restrict addition of the genes with low frequencies of orthologues. An effective way to impose this restriction was to admit at most N rGenomes to a pan-genome construction (here $N = 150$).

Another option to control the size of a pan-genome was to change the length of the shortest gene. However, we observed that the size of a pan-genome practically did not change when the minimum gene length was selected anywhere between 90nt and 200 nt; therefore, the gene length threshold was set to 90 nt.

We also observed that under the restriction of using for the pan-genome construction no more than 150 largest rGenomes per SGB, the whole set of the genes would grow three-fold in comparison with the number of genes in the SGB *representative* rGenome. At the same time, the size of the pan-genome gene sequence database grew >4-fold (due to the asymmetry of the comparison rule between query and target (see Methods) we made more frequent selection of genes longer than queries). Speaking of computational efficiency, a running time of processing a single query increased just 2-fold between using a single *representative* rGenome and a full pan-genome. This effect was partly due to running the fast mode of BLASTn with the *k*-mer size of 28 nt chosen under assumption of high similarity of query and targets (24). This fast mode showed a sub-linear dependence of the run-time on a size of the sequence database. Use of pan-genomes has significantly—by factor 15—reduced the database volume in comparison with a number of all the genes in a non-redundant set of SGBs and tClusters (Table 5).

Construction of the pan-genomes significantly reduced (by factor 15) the memory footprint necessary for the GeneMark-HM operation (Table 5). Although a few erroneously predicted genes could be included in the pan-genome database upon its construction, these 'targets' could rarely appear as hits upon similarity searches. Thus, such errors were not likely to influence accuracy of a contig assignment to a species level clade.

Similarity searches with the query genes against the pan-genome database was the most time-consuming step in the

Table 5. Statistics of the pan-genomes database construction

Source	Number of pan-genomes	Number of genes in pan-genomes	Volume of sequence
tClusters	701	5.7 million	5.3 Gb
SGBs	4930	20.4 million	19.2 Gb
SGBs + non-overlapping tClusters	5082	21.6 million	20.4 Gb

Table 6. Genes predicted in the four human metagenomes with indication of numbers of genes predicted by the three alternative branches of GeneMark-HM (Figure 1). Contigs shorter than 1500 nt were not included in this analysis

Assembly ID	Site	Size, Mb	# of contigs	# of genes predicted	Pan-genome path	MetaGeneMark path	Gene MarkS-2 path	% by pan-genome path
GCA_013297495	skin	44.5	7569	49 285	39 154	10 131	375	79.4
GCA_900415335	gut	49.7	5364	44 476	44 211	265	0	99.4
GCA_003604395	gut	85.9	12 353	88 015	87 305	710	0	99.2
GCA_003640265	oral	296.0	81 048	318 534	283 412	35 122	170	89.0

GeneMark-HM operation. To reduce the run-time, given that we searched for a hit with >95% ANI identity of target to a query, we used a fast MegaBLAST algorithm (24) with long k-mers ($k = 28$). Even in this case, the similarity search step took on average ∼80% of the GeneMark-HM run-time (using a single CPU).

We should mention that, for comparison, we implemented an alternative variant of GeneMark-HM using pan-genomes in protein space and the Diamond algorithm for similarity searches (32). Strikingly, the run-time increased by the two orders of magnitude. A particular reason for the increase was having large number of high scoring alignments to proteins from more remotely related pan-genomes, as the protein translations of genes mask DNA differences in synonymous codon positions.

Gene prediction in real metagenomes

We tested GeneMark-HM on human metagenomes originated from microbiomes sampled from the skin (GCA_013297495), the gut (GCA_900415335 and GCA_003604395) and the oral cavity (GCA_003640265). We observed that most gene predictions in each metagenome, from 79.4% to 99.4%, was made by GeneMark.hmm with the *representative* models (the pan-genome path, Table 6 and Figure 1). A percentage of the predictions made through this path was especially high (>99%) in the human gut metagenomes. A sizable number of relatively short contigs that were not assigned to pan genomes and require use of MetaGeneMark was observed in the oral and in the skin metagenomes, where a few unassigned to a pan-genome contigs were sufficiently long for *de novo* training of GeneMarkS-2 (Table 6). Such long unassigned contigs were absent in the gut metagenomes.

In the tests on real metagenomes with total sequence length of 45 Mb we observed the following run-times (using a single CPU): MetaGeneMark – 10 s, MetaProdigal – 3.5 min, GeneMark-HM – 12 min. The time of GeneMark-HM included the time required for the BLASTn scanning the pan-genome database (∼80% of the time). This time could be significantly improved by parallel processing.

DISCUSSION

In the Results section we demonstrated that the GeneMark-HM pipeline did improve the accuracy of gene prediction in metagenomic sequences. Particularly, we observed improvements in the whole gene prediction sensitivity as well as in accuracy of the gene starts prediction (Table 3).

The observed improvement was mainly due to *selecting an accurate model* for GeneMark.hmm-2 by means of assignment of a metagenomic contig to a pan-genome in the pan-genome database. The optimal model selection made GeneMark.hmm-2 capable to annotate genes in short contigs no less accurately than GeneMarkS-2 annotates complete genomes (Figure 1, Table 3). The pan-genome database for the human microbiome was built from arguably the largest collection of genomes of sequenced isolates and prokaryotic genomes reconstructed *in silico* from metagenomics reads.

Experiments with real metagenomes from human sites demonstrated, that up to 99.4% of all contigs could be assigned to pan-genomes made at a species level. If such an assignment was in place, a *representative* model of the pan-genome could be used in GeneMark.hmm-2 to predict genes in the contig. The highest rate of a contig to a pan-genome assignment, 99.4%, was observed for human gut metagenomes. The lower rate, down to 89.0% and 79.4% was observed for oral and skin metagenomes respectively (Table 6). These results could be expected as the oral and skin microbiomes might be mixed with microbes from environment. The contigs not matched to pan-genomes were processed by either MetaGeneMark, the shorter ones, and by GeneMarkS-2, the longer ones (Figure 1). Interestingly, the MetaGeneMark performance was quite close to GeneMarkS-2 in terms of accurate prediction of the 3′ ends of genes, with 1.3% and 0.9% error, respectively (Table 3).

A significant improvement was observed in gene start prediction (Table 3). Types of sequence signals controlling translation initiation in prokaryotic genomes vary significantly (5). A contig assignment to a pan-genome triggered an automatic selection of a suitable model of the sequences near translation start sites for the whole contig.

For comparison, the gene start models in MetaProdigal are aware of canonical RBS (group A in Table 3) and non-canonical RBS (group B in Table 3), but not of frequently

observed sequence patterns associated with the leaderless translation initiation (group C in Table 3). In this group the decrease in the error rate of GeneMark-HM in comparison with the one of MetaProdigal was the largest (Table 4).

Incorrect identification of a role of the TGA codon would affect the accuracy of gene prediction in an anonymous contig. Determination of this role is, therefore, is made in GeneMark-HM. It was shown that GeneMark-HM identified the role of TGA in anonymous contigs with fewer errors than MetaProdigal (Table 2). Particularly, GeneMark-HM made no errors in contigs with more than 100 genes.

The GeneMark-HM accuracy does depend on the correct assignment of a metagenomic contig to a reference pan-genome. Overall improvement of the gene prediction accuracy confirms that this step was done efficiently. Moreover, the assignment to a pan-genome would also facilitate accurate functional annotation of the predicted genes and proteins by narrowing down the phylogenetic search space.

DATA AVAILABILITY

The GeneMark-HM pipeline software is freely available for academic and non-commercial users: https://github.com/GeneProbe/GeneMark-HM.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

We would like to thank Prof. Nicola Segata for sharing the initial SGB data as well as for useful discussions on metagenomic data analysis.

FUNDING

Enterome, Paris, France; Gene Probe, Inc. Atlanta, USA. Funding for open access charge: Gene Probe, Inc. Atlanta, USA.

Conflict of interest statement. The authors have no conflicts of interest to declare aside from the fact that several of the authors are employees of private companies, Gene Probe, Inc. and Enterome, as stated in the title page of the manuscript.

REFERENCES

1. Vollmers,J., Wiegand,S. and Kaster,A.K. (2017) Comparing and evaluating metagenome assembly tools from a microbiologist's perspective - not only size matters! *PLoS One*, **12**, e0169662.
2. 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.*, **29**, 2607–2618.
3. Delcher,A.L., Bratke,K.A., Powers,E.C. and Salzberg,S.L. (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics*, **23**, 673–679.
4. Hyatt,D., Chen,G.L., Locascio,P.F., Land,M.L., Larimer,F.W. and Hauser,L.J. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, **11**, 119.
5. Lomsadze,A., Gemayel,K., Tang,S. and Borodovsky,M. (2018) Modeling leaderless transcription and atypical genes results in more accurate gene prediction in prokaryotes. *Genome Res.*, **28**, 1079–1089.
6. Besemer,J. and Borodovsky,M. (1999) Heuristic approach to deriving models for gene finding. *Nucleic Acids Res.*, **27**, 3911–3920.
7. Hyatt,D., LoCascio,P.F., Hauser,L.J. and Uberbacher,E.C. (2012) Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics*, **28**, 2223–2230.
8. Noguchi,H., Taniguchi,T. and Itoh,T. (2008) MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. *DNA Res.*, **15**, 387–396.
9. Rho,M., Tang,H. and Ye,Y. (2010) FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res.*, **38**, e191.
10. Zhu,W., Lomsadze,A. and Borodovsky,M. (2010) Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res.*, **38**, e132.
11. Almeida,A., Mitchell,A.L., Boland,M., Forster,S.C., Gloor,G.B., Tarkowska,A., Lawley,T.D. and Finn,R.D. (2019) A new genomic blueprint of the human gut microbiota. *Nature*, **568**, 499–504.
12. Qin,J., Li,R., Raes,J., Arumugam,M., Burgdorf,K.S., Manichanh,C., Nielsen,T., Pons,N., Levenez,F., Yamada,T. *et al.* (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, **464**, 59–65.
13. Vernikos,G., Medini,D., Riley,D.R. and Tettelin,H. (2015) Ten years of pan-genome analyses. *Curr. Opin. Microbiol.*, **23**, 148–154.
14. Pasolli,E., Asnicar,F., Manara,S., Zolfo,M., Karcher,N., Armanini,F., Beghini,F., Manghi,P., Tett,A., Ghensi,P. *et al.* (2019) Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell*, **176**, 649–662.
15. Dubin,K., Callahan,M.K., Ren,B., Khanin,R., Viale,A., Ling,L., No,D., Gobourne,A., Littmann,E., Huttenhower,C. *et al.* (2016) Intestinal microbiome analyses identify melanoma patients at risk for checkpoint-blockade-induced colitis. *Nat. Commun.*, **7**, 10391.
16. Human Microbiome Project Consortium (2012) Structure, function and diversity of the healthy human microbiome. *Nature*, **486**, 207–214.
17. Karlsson,F.H., Tremaroli,V., Nookaew,I., Bergstrom,G., Behre,C.J., Fagerberg,B., Nielsen,J. and Backhed,F. (2013) Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature*, **498**, 99–103.
18. Le Chatelier,E., Nielsen,T., Qin,J., Prifti,E., Hildebrand,F., Falony,G., Almeida,M., Arumugam,M., Batto,J.M., Kennedy,S. *et al.* (2013) Richness of human gut microbiome correlates with metabolic markers. *Nature*, **500**, 541–546.
19. Nielsen,H.B., Almeida,M., Juncker,A.S., Rasmussen,S., Li,J., Sunagawa,S., Plichta,D.R., Gautier,L., Pedersen,A.G., Le Chatelier,E. *et al.* (2014) Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.*, **32**, 822–828.
20. Obregon-Tito,A.J., Tito,R.Y., Metcalf,J., Sankaranarayanan,K., Clemente,J.C., Ursell,L.K., Zech Xu,Z., Van Treuren,W., Knight,R., Gaffney,P.M. *et al.* (2015) Subsistence strategies in traditional societies distinguish gut microbiomes. *Nat. Commun.*, **6**, 6505.
21. Rampelli,S., Schnorr,S.L., Consolandi,C., Turroni,S., Severgnini,M., Peano,C., Brigidi,P., Crittenden,A.N., Henry,A.G. and Candela,M. (2015) Metagenome Sequencing of the Hadza Hunter-Gatherer Gut Microbiota. *Curr. Biol.*, **25**, 1682–1693.
22. Zeller,G., Tap,J., Voigt,A.Y., Sunagawa,S., Kultima,J.R., Costea,P.I., Amiot,A., Bohm,J., Brunetti,F., Habermann,N. *et al.* (2014) Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.*, **10**, 766.
23. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
24. 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.
25. Jain,C., Rodriguez,R.L., Phillippy,A.M., Konstantinidis,K.T. and Aluru,S. (2018) High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.*, **9**, 5114.
26. Borodovsky,M. and McIninch,J. (1993) GENMARK: parallel gene recognition for both DNA strands. *Comput. Chem.*, **17**, 123–133.
27. Burge,C. and Karlin,S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, 78–94.

28. Stanke,M. and Waack,S. (2003) Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*, **19**(Suppl. 2), ii215–ii225.
29. Ivanova,N.N., Schwientek,P., Tripp,H.J., Rinke,C., Pati,A., Huntemann,M., Visel,A., Woyke,T., Kyrpides,N.C. and Rubin,E.M. (2014) Stop codon reassignments in the wild. *Science*, **344**, 909–913.
30. Haft,D.H., DiCuccio,M., Badretdin,A., Brover,V., Chetvernin,V., O'Neill,K., Li,W., Chitsaz,F., Derbyshire,M.K., Gonzales,N.R. *et al.* (2018) RefSeq: an update on prokaryotic genome annotation and curation. *Nucleic Acids Res.*, **46**, D851–D860.
31. Morgulis,A., Coulouris,G., Raytselis,Y., Madden,T.L., Agarwala,R. and Schaffer,A.A. (2008) Database indexing for production MegaBLAST searches. *Bioinformatics*, **24**, 1757–1764.
32. Buchfink,B., Xie,C. and Huson,D.H. (2015) Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*, **12**, 59–60.