Gene Prediction Team 3: Background and Strategy

Pallavi Misra
Sonali Gupta
Ahish Melkote Sujay
Shen-Yi Cheng
Jie Zhou

February 13, 2020

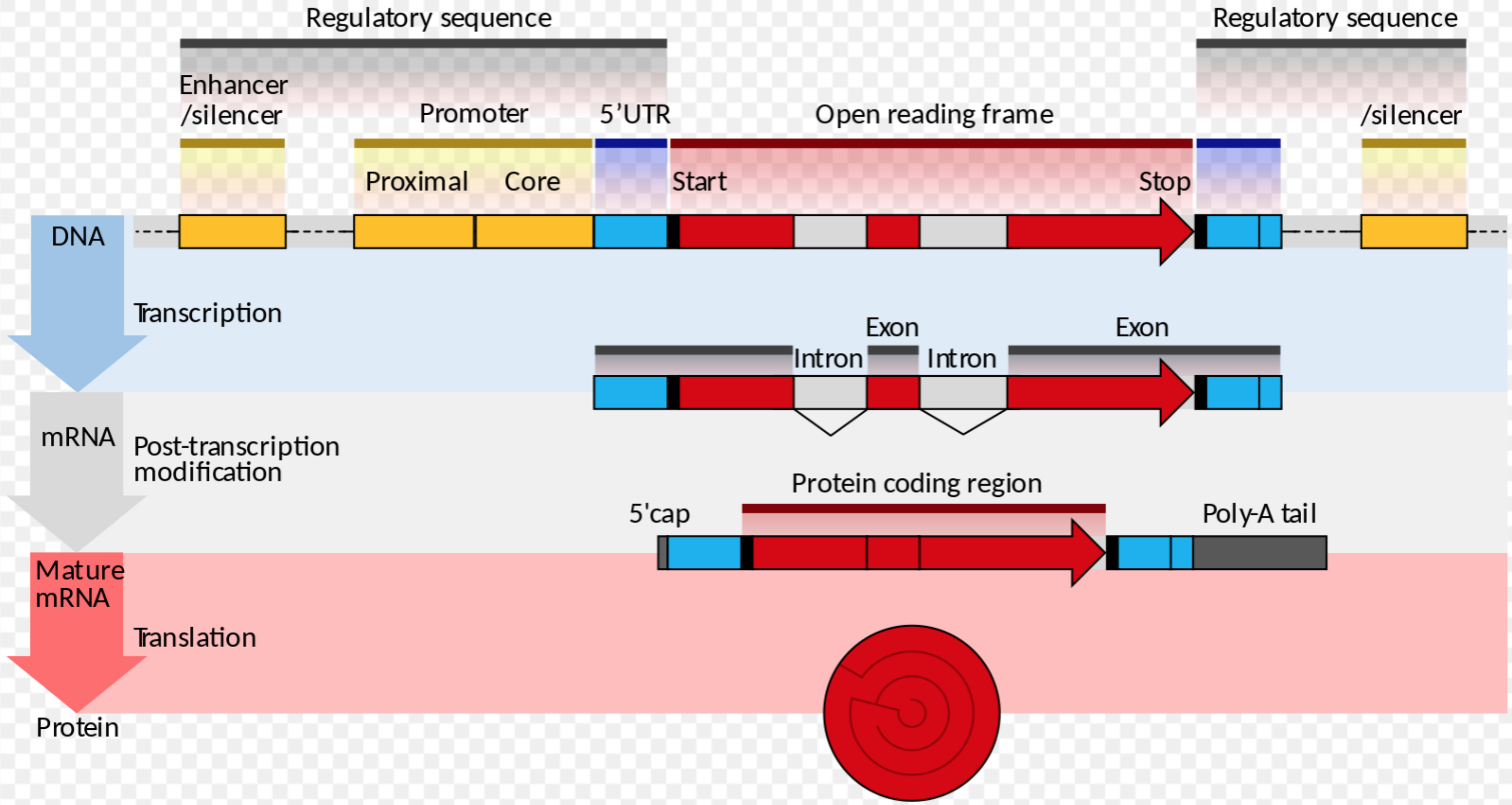
GENE PREDICTION

The process of identifying the regions of genomic DNA that encode genes:

1. Protein-coding genes
    - Ahish Sujay
    - Pallavi Mishra
    - Sonali Gupta

2. Non-coding RNA genes, other regulatory regions
    - Shen-Yi Cheng
    - Jie Zhou

- Challenges:
    - Sequencing errors
    - Quality of assembly
    - Frameshift mutations, overlapping genes



Prokaryotic gene structure

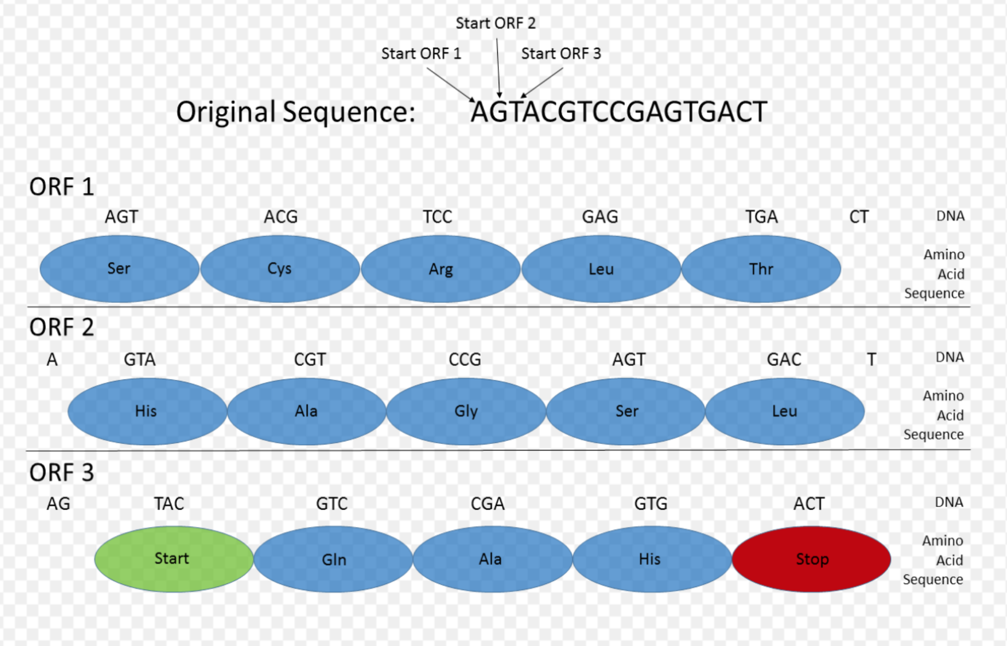
METHODS

1. Ab initio methods

Genomic DNA sequence is systematically searched for # of protein-coding genes

Prokaryotic genes have:

- Transcription binding sites
- Promoter sequences
- Contiguous ORFs
- Compositional domain GC composition : Isochores



Start ORF 2
Start ORF 1 Start ORF 3
Original Sequence: AGTACGTCCGAGTGACT

ORF 1
AGT ACG TCC GAG TGA CT DNA
Ser Cys Arg Leu Thr Amino Acid Sequence

ORF 2
A GTA CGT CCG AGT GAC T DNA
His Ala Gly Ser Leu Amino Acid Sequence

ORF 3
AG TAC GTC CGA GTG ACT DNA
Start Gln Ala His Stop Amino Acid Sequence

Limitations:

- Have to rely on extrinsic evidence to determine if a gene is functional

METHODS

2. Homology based methods

Target genome is searched for sequences that are similar to extrinsic evidence in the form of the known

- Expressed sequence tags
- Messenger RNA (mRNA)
- Protein products

Limitations:

Computationally expensive in complex organisms

Not all genes are expressed at a time; requires an extensive database

Cannot predict Horizontally transferred genes

HOMOLOGY BASED GENE PREDICTION

1. Based on sequence similarity of query sequence with annotated genes present in database

2. Given a database of sequences of the organism, search for a query sequence in the database

3. If the identified sequences are genes, the query sequence is a gene

TOOLS FOR HOMOLOGY BASED GENE PREDICTION

TOOL	YEAR OF PUBLICATION	CITATIONS
BLAST	1990	82,373+
HMMER	2011	1,672
PROCRUSTES	1996	381
DIAMOND	2015	1,308
GENEWISE	2004	1,490

BLAST

1. Before BLAST, alignment programs used dynamic programming algorithms, such as the Needleman-Wunsch and Smith-Waterman algorithms, required long processing times

2. instead of comparing every residue against each other, BLAST uses short "word" (w) segments to create alignment "seeds." : this reduces the search space

3.BLAST extends the alignment in both directions according to a threshold (T) that is set by the user

MAX HSPS & MAX_TARGET_SEQ

- max_hsps = Maximum number of HSPs (alignments) to keep for any single query-subject pair. If this option is not set, BLAST shows all HSPs meeting the e value criteria.

- max_target_seq = Number of aligned sequences per query to keep

- E-value = number of expected hits of similar quality (score) that could be found just by chance

HMMER

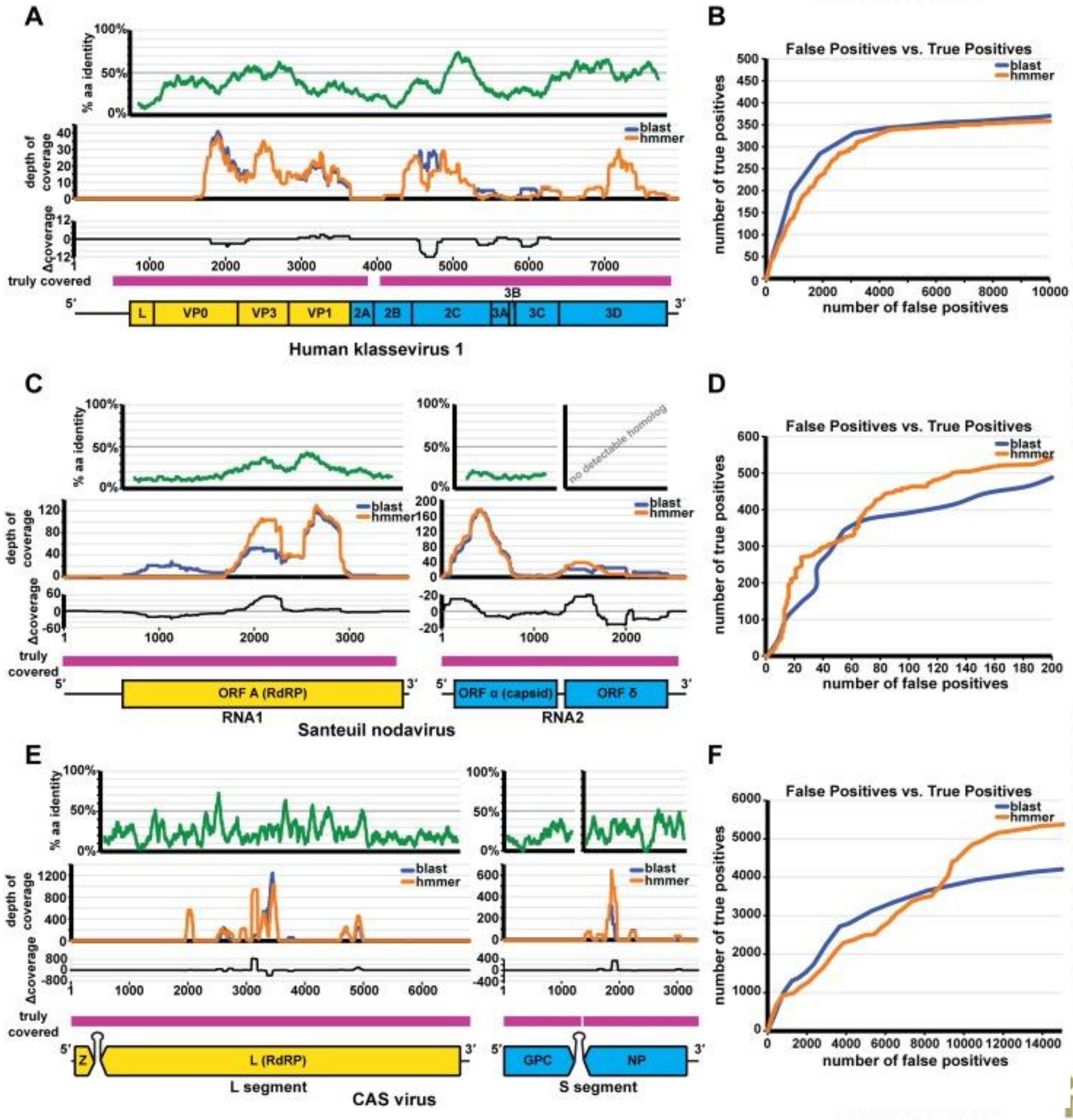
1. It detects homology by comparing a profile-HMM to either a single sequence or a database of sequences

2. Profile HMMs:

- multiple sequence alignment into a position-specific scoring system
- certain positions in a sequence alignment tend to have biases
- one state in HMM corresponds to each consensus column in a sequence alignment
- probability of emitting a particular residue is determined by the frequency at which that residue has been observed in that column of the alignment

3. Sequences that score significantly better to the profile-HMM considered to be homologous to the sequences

A comparison of BLAST vs. HMMER for the detection of Human klassevirus 1, Santeuil nodavirus, and CAS virus.



DIAMOND

- The program is based on the traditional seed-and-extend paradigm for sequence comparison,
- Spaced seeds. A second improvement of the seed step is to use spaced seeds—that is, longer seeds in which only a subset of positions are used
- Double index: DIAMOND uses a double-indexing approach in which both the queries and the references are indexed

DATABASES

RefSeq

Title:RefSeq Genome Database

Description:This database contains NCBI Refseq genomes across all taxonomy groups. It contains only the longest sequences representing any given part of the genomes; contigs are not included

Molecule Type:Genomic

Update date:2016/12/14

Number of sequences:33120025

GenBank

The GenBank archival sequence database includes publicly available DNA sequences submitted from individual laboratories and large-scale sequencing projects. GenBank sequence records are owned by the original submitter and cannot be altered by a third party.

As an archival database, GenBank can be very redundant for some loci.

Nr-nt

Title:Nucleotide collection (nt)

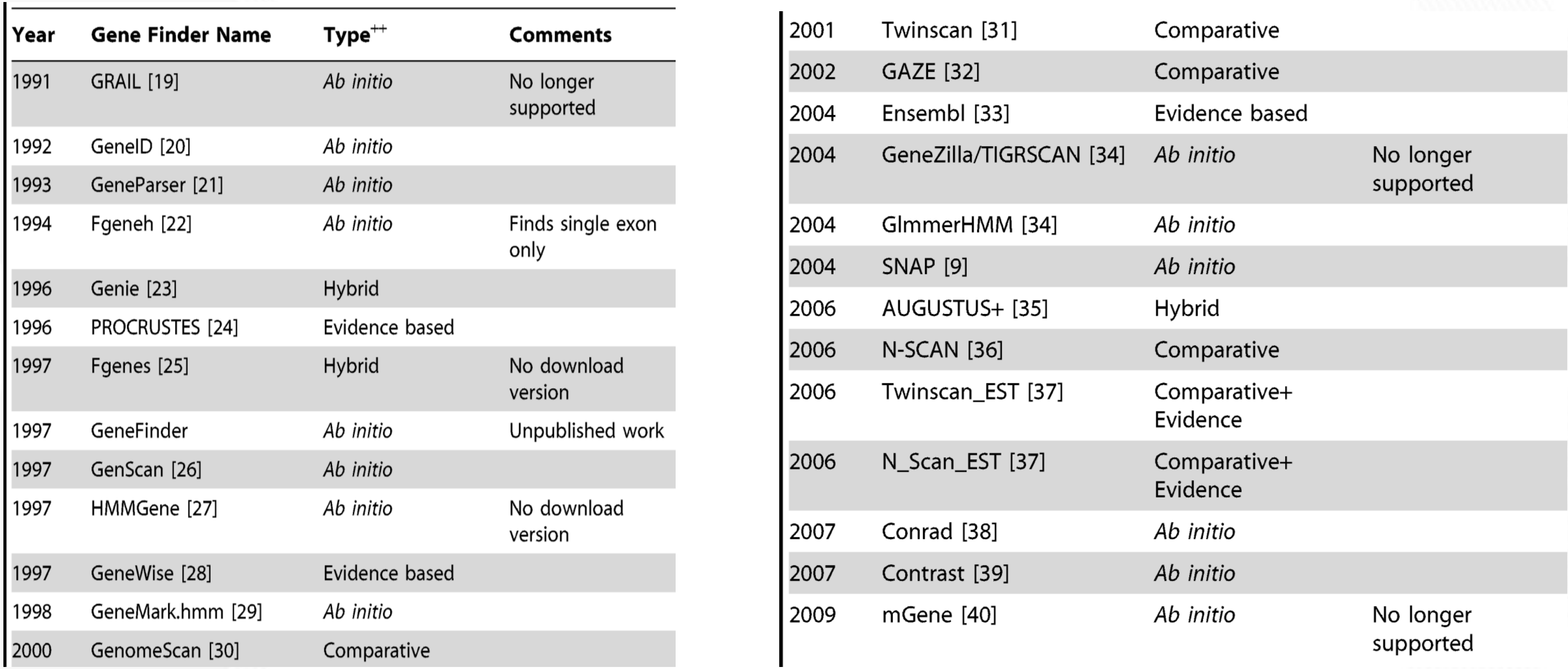
Description:The nucleotide collection consists of GenBank+EMBL+DDBJ+PDB+RefSeq Sequences. The database is non-redundant, annotated and curated. Identical sequences have been merged into one entry, while preserving the accession, GI, title and taxonomy information for each entry.

Molecule Type:mixed DNA

Update date:2019/10/03

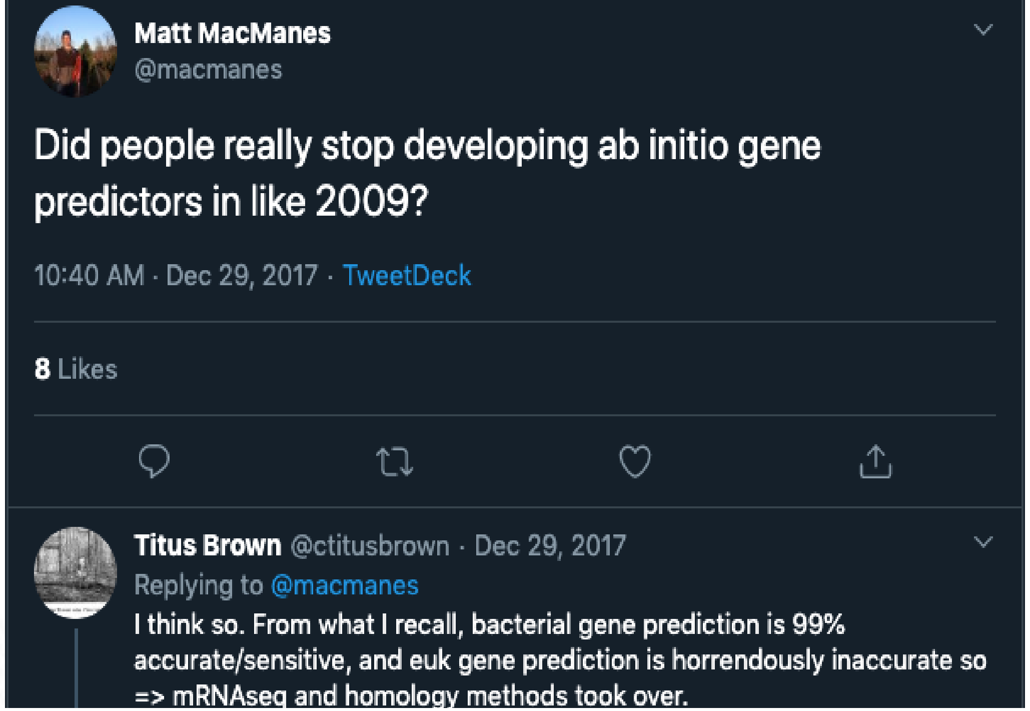
Number of sequences:55908648

Ab Initio Methods



Year	Gene Finder Name	Type^{++}	Comments
1991	GRAIL [19]	*Ab initio*	No longer supported
1992	GeneID [20]	*Ab initio*	
1993	GeneParser [21]	*Ab initio*	
1994	Fgeneh [22]	*Ab initio*	Finds single exon only
1996	Genie [23]	Hybrid	
1996	PROCRUSTES [24]	Evidence based	
1997	Fgenes [25]	Hybrid	No download version
1997	GeneFinder	*Ab initio*	Unpublished work
1997	GenScan [26]	*Ab initio*	
1997	HMMGene [27]	*Ab initio*	No download version
1997	GeneWise [28]	Evidence based	
1998	GeneMark.hmm [29]	*Ab initio*	
2000	GenomeScan [30]	Comparative	
2001	Twinscan [31]	Comparative	
2002	GAZE [32]	Comparative	
2004	Ensembl [33]	Evidence based	
2004	GeneZilla/TIGRSCAN [34]	*Ab initio*	No longer supported
2004	GlmmerHMM [34]	*Ab initio*	
2004	SNAP [9]	*Ab initio*	
2006	AUGUSTUS+ [35]	Hybrid	
2006	N-SCAN [36]	Comparative	
2006	Twinscan_EST [37]	Comparative+ Evidence	
2006	N_Scan_EST [37]	Comparative+ Evidence	
2007	Conrad [38]	*Ab initio*	
2007	Contrast [39]	*Ab initio*	
2009	mGene [40]	*Ab initio*	No longer supported

Goodswen SJ, Kennedy PJ, Ellis JT. Evaluating high-throughput ab initio gene finders to discover proteins encoded in eukaryotic pathogen genomes missed by laboratory techniques. PLoS One. 2012;7(11):e50609. doi: 10.1371/journal.pone.0050609. Epub 2012 Nov 30. PubMed PMID: 23226328; PubMed Central PMCID: PMC3511556.



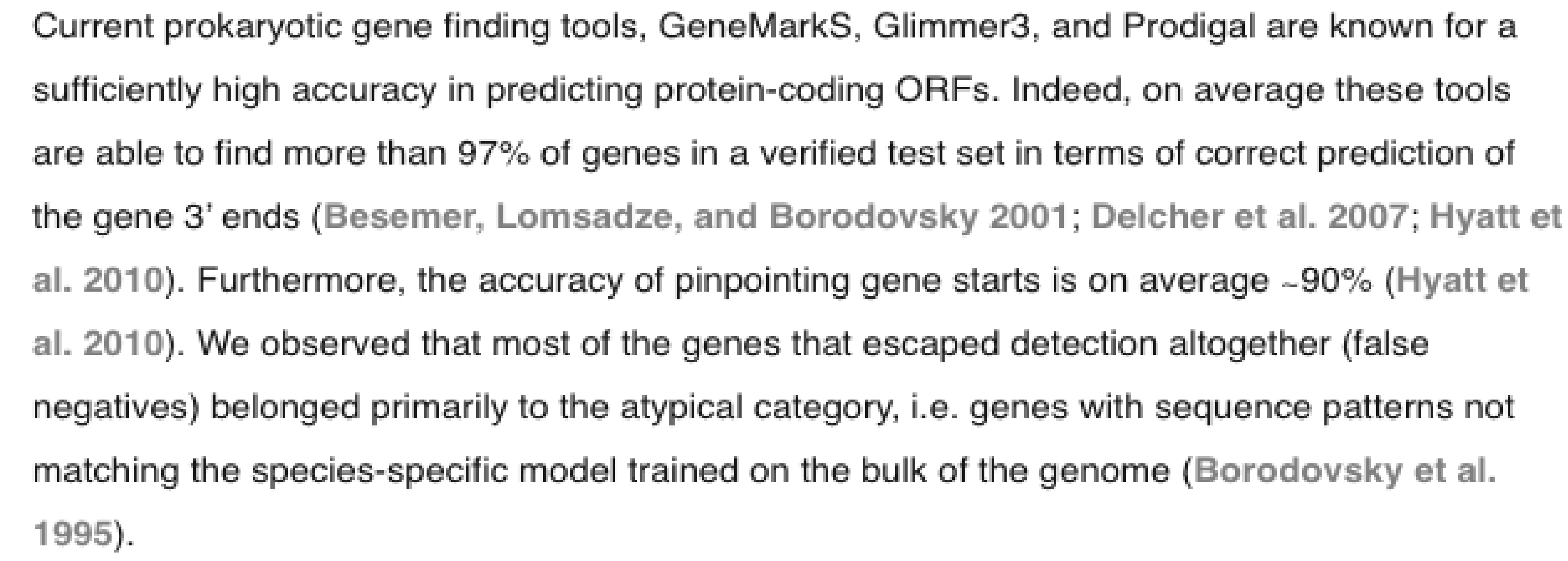
Matt MacManes
@macmanes

Did people really stop developing ab initio gene predictors in like 2009?

10:40 AM · Dec 29, 2017 · TweetDeck

8 Likes

Titus Brown @ctitusbrown · Dec 29, 2017
Replying to @macmanes
I think so. From what I recall, bacterial gene prediction is 99% accurate/sensitive, and euk gene prediction is horrendously inaccurate so => mRNAseq and homology methods took over.



Current prokaryotic gene finding tools, GeneMarkS, Glimmer3, and Prodigal are known for a sufficiently high accuracy in predicting protein-coding ORFs. Indeed, on average these tools are able to find more than 97% of genes in a verified test set in terms of correct prediction of the gene 3' ends (Besemer, Lomsadze, and Borodovsky 2001; Delcher et al. 2007; Hyatt et al. 2010). Furthermore, the accuracy of pinpointing gene starts is on average ~90% (Hyatt et al. 2010). We observed that most of the genes that escaped detection altogether (false negatives) belonged primarily to the atypical category, i.e. genes with sequence patterns not matching the species-specific model trained on the bulk of the genome (Borodovsky et al. 1995).

Improved Prokaryotic Gene Prediction Yields Insights into Transcription and Translation Mechanisms on Whole Genome Scale- (Alexandre Lomsadze, Karl Gemayel, Shiyuyun Tang, Mark Borodovsky)

Comparison of Ab Initio tools

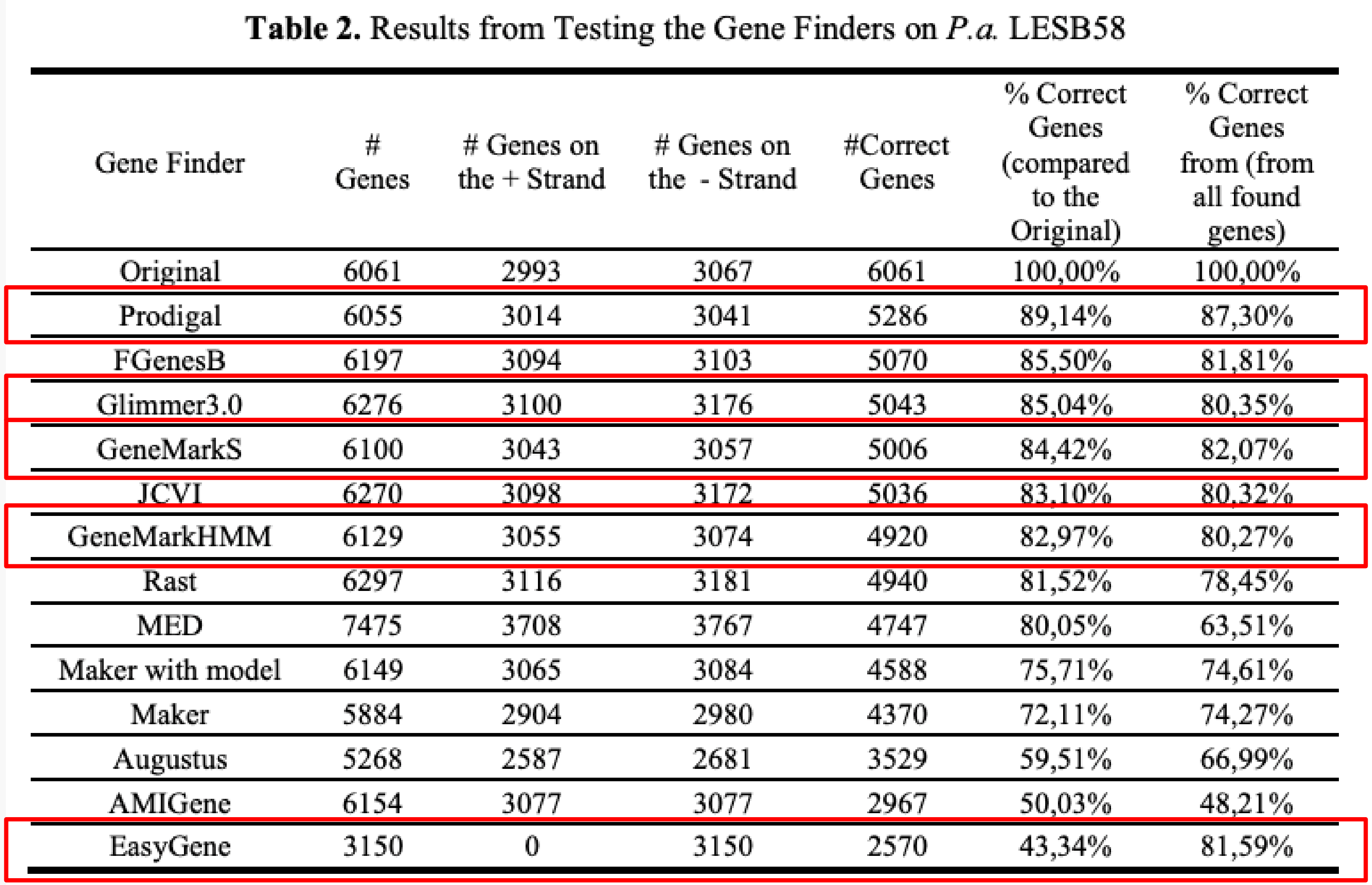


Table 2. Results from Testing the Gene Finders on *P.a.* LESB58

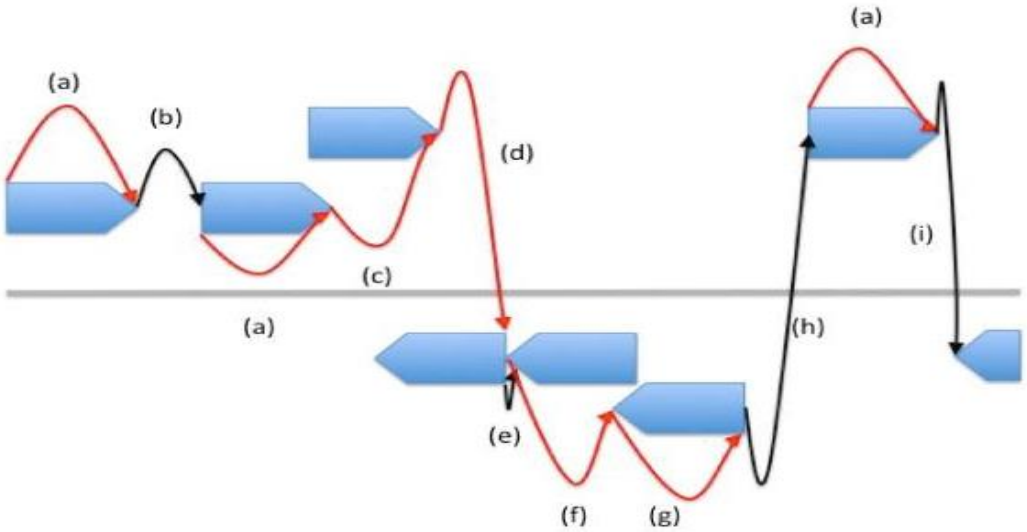
Gene Finder	# Genes	# Genes on the + Strand	# Genes on the - Strand	#Correct Genes	% Correct Genes (compared to the Original)	% Correct Genes from (from all found genes)
Original	6061	2993	3067	6061	100,00%	100,00%
Prodigal	6055	3014	3041	5286	89,14%	87,30%
FGenesB	6197	3094	3103	5070	85,50%	81,81%
Glimmer3.0	6276	3100	3176	5043	85,04%	80,35%
GeneMarkS	6100	3043	3057	5006	84,42%	82,07%
JCVI	6270	3098	3172	5036	83,10%	80,32%
GeneMarkHMM	6129	3055	3074	4920	82,97%	80,27%
Rast	6297	3116	3181	4940	81,52%	78,45%
MED	7475	3708	3767	4747	80,05%	63,51%
Maker with model	6149	3065	3084	4588	75,71%	74,61%
Maker	5884	2904	2980	4370	72,11%	74,27%
Augustus	5268	2587	2681	3529	59,51%	66,99%
AMIGene	6154	3077	3077	2967	50,03%	48,21%
EasyGene	3150	0	3150	2570	43,34%	81,59%

Angelova, Mihaela & Kalajdziski, Slobodan & Kocarev, Ljupco. (2010). Computational Methods for Gene Finding in Prokaryotes. ICT Innovations. 1. 1857-7288

Ab Initio Tools	Algorithm	Citations	Basis
GeneMark.hmm	HMM	1681	Excellent documentation, most widely used and high number of citations
GeneMarkS	HMM	1379	Self training, excellent documentation, most widely used and high number of citations
GeneMarkS2	HMM	20	Self training, excellent documentation, most widely used, superior than S2 (stated by their paper)
Prodigal	DP + Markov Model	3440	Self training, excellent documentation, most widely used and high number of citations
Glimmer	IMM	1212	Self training, excellent documentation, most widely used and high number of citations
SNAP	Semi-HMM	1251	Algorithm needs to be trained on dataset, ZFF format needed (Nobody except the develop uses this format)
AUGUSTUS	HMM	952	Algorithm needs to be trained on dataset, need to upload whole genome data, has been trained on only 3 species of Bacteria
EasyGene	HMM + BLAST	187	Number of citations are low
ChemGenome	Physicochemical characteristics and MD simulation	32	Number of citations are extremely low
MED 2.0	MED Algorithm (Non-supervised)	37	Not maintained anymore

PRODIGAL (PROkaryotic DYnamic programming Gene-finding ALgorithm)

- PRODIGAL scores individual ORFs using various features and scoring rules and then performs dynamic programming on all pairs of start-and-stop triplets to find the maximum scoring path.
- The adopted features Prodigal includes are GC bias in first, second, and third positions of each codon, frequency of hexamers, ORF length, upstream sequence resembling ORF, etc.
- The connection of a start node to its corresponding stop node represents a gene, whereas the connection of a 3' end to a new 5' end represents intergenic space.



The red arrows represent gene connections, and the black arrows represent intergenic connections.
(a) 5' forward to 3' forward: Gene on the forward strand.
(b) 3' forward to 5' forward: Intergenic space between two forward strand genes.
(c) 3' forward to 3' forward: Overlapping genes on the forward strand.
(d) 3' forward to 5' reverse: Forward and reverse strand genes whose 3' ends overlap.
(e) 5' reverse to 3' reverse: Intergenic space between two reverse strand genes.
(f) 3' reverse to 5' reverse: Gene on the reverse strand.
(g) 3' reverse to 3' reverse: Overlapping genes on the reverse strand.
(h) 5' reverse to 5' forward: Intergenic space between two opposite strand genes.
(i) 3' forward to 3' reverse: Intergenic space between two opposite strand genes.

GLIMMER (Gene Locator and Interpolated Markov ModelER)

- GLIMMER searches for long-ORFs and generates a training data set to which it trains all six Markov models of coding and non-coding DNA from zero to eight order.
- After calculating the probabilities from the above data, GLIMMER decides to either use fixed order Markov model or interpolated Markov model. Performed by program “build-imm”.

a. If the no. observation > 400 = Fixed order Markov model

b. If the no. observation < 400 = Interpolated Markov model

- Obtains score for every long-ORF generated and if score if greater than a certain threshold, GLIMMER predicts it as a gene. Performed by program “glimmer”.

GeneMarkS-2

- It uses a model derived by self-training for finding species-specific (native) genes
- Horizontal Gene Transfer detection: It uses precomputed heuristic models designed to identify harder-to-detect genes

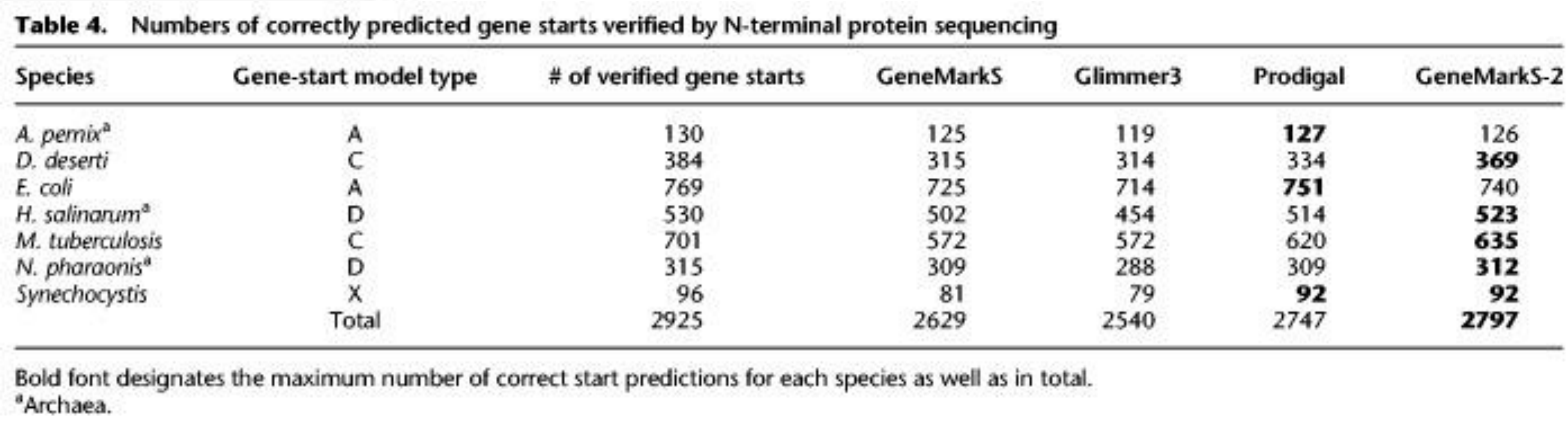


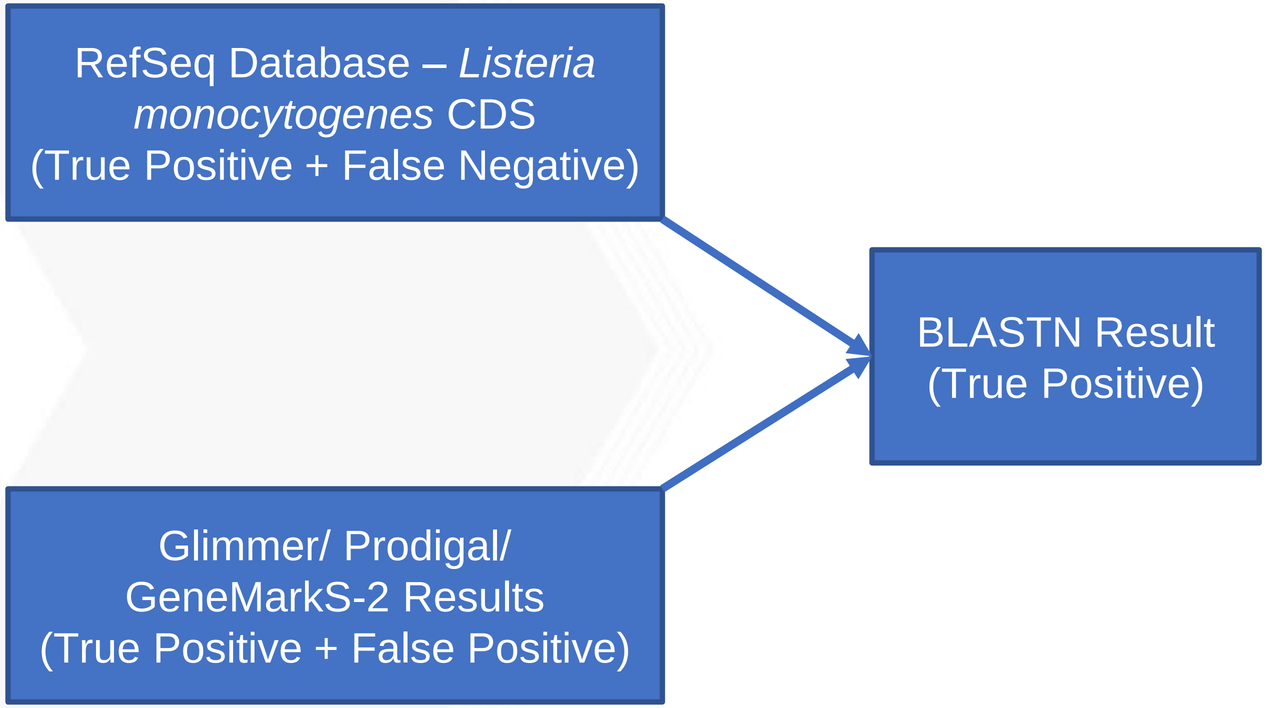
Table 4. Numbers of correctly predicted gene starts verified by N-terminal protein sequencing

Species	Gene-start model type	# of verified gene starts	GeneMarkS	Glimmer3	Prodigal	GeneMarkS-2
A. pernix[a]	A	130	125	119	**127**	126
D. deserti	C	384	315	314	334	**369**
E. coli	A	769	725	714	**751**	740
H. salinarum[a]	D	530	502	454	514	**523**
M. tuberculosis	C	701	572	572	620	**635**
N. pharaonis[a]	D	315	309	288	309	**312**
Synechocystis	X	96	81	79	**92**	**92**
	Total	2925	2629	2540	2747	**2797**

Bold font designates the maximum number of correct start predictions for each species as well as in total.
[a]Archaea.

Lomsadze, Alexandre, et al. "Modeling leaderless transcription and atypical genes results in more accurate gene prediction in prokaryotes." *Genome research* 28.7 (2018): 1079-1089.

Workflow for selection of Gene Prediction tools


RefSeq Database – *Listeria monocytogenes* CDS
(True Positive + False Negative)
Glimmer/ Prodigal/ GeneMarkS-2 Results
(True Positive + False Positive)
BLASTN Result
(True Positive)

Test dataset:

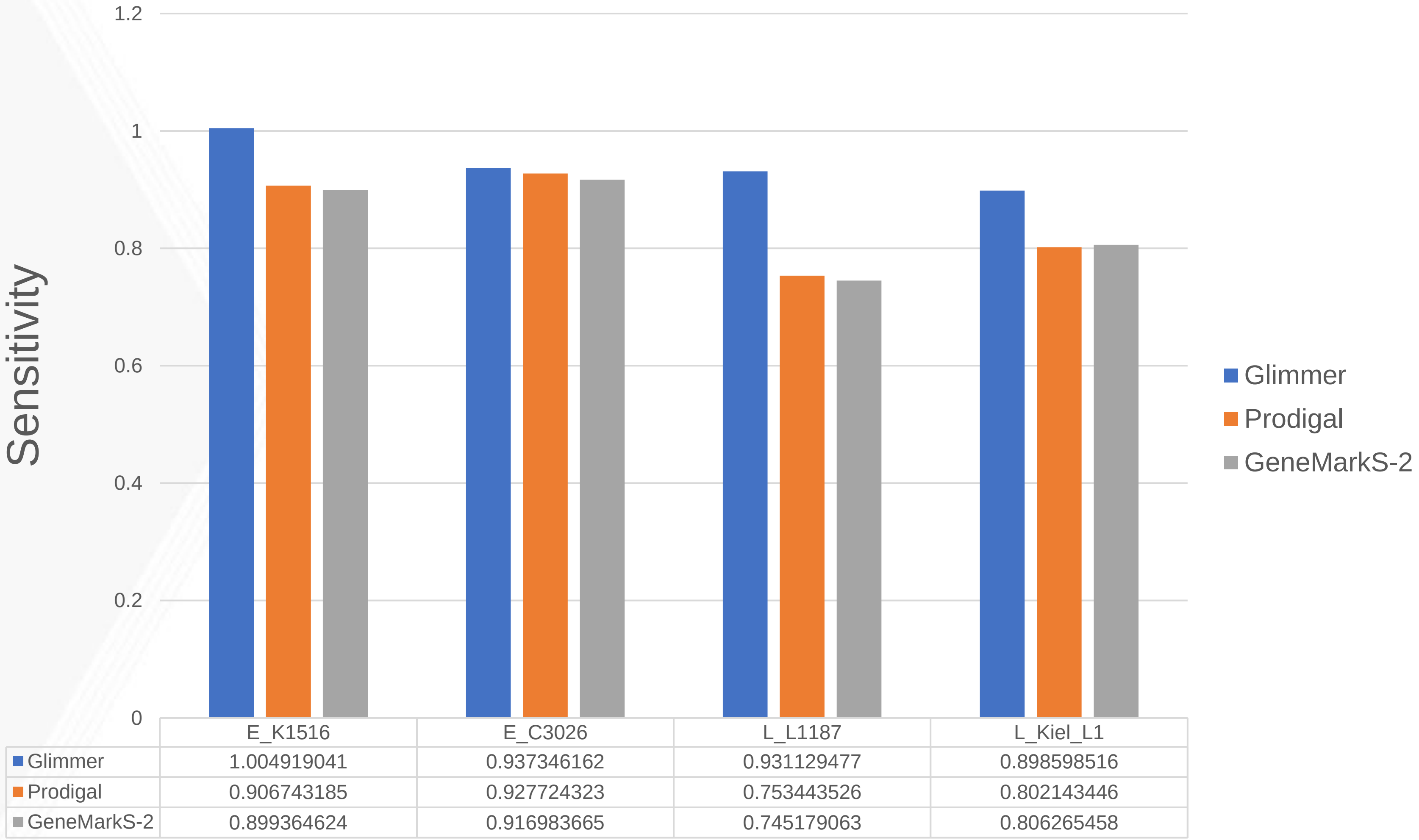
Escherichia coli O15:H18 str. K1516 (E. coli)
Escherichia coli K-12 (E. coli) Strain: *C3026*
Listeria floridensis FSL S10-1187 (firmicutes)
Listeria kieliensis (firmicutes) Strain: *Kiel-L1*

Evaluation metric used:

- Sensitivity: $\frac{True\ Positive}{True\ Positive + False\ Negative}$
- False Discovery Rate: $\frac{False\ Positive}{True\ Positive + False\ Positive}$
- Precision: $\frac{True\ Positive}{True\ Positive + False\ Positive}$

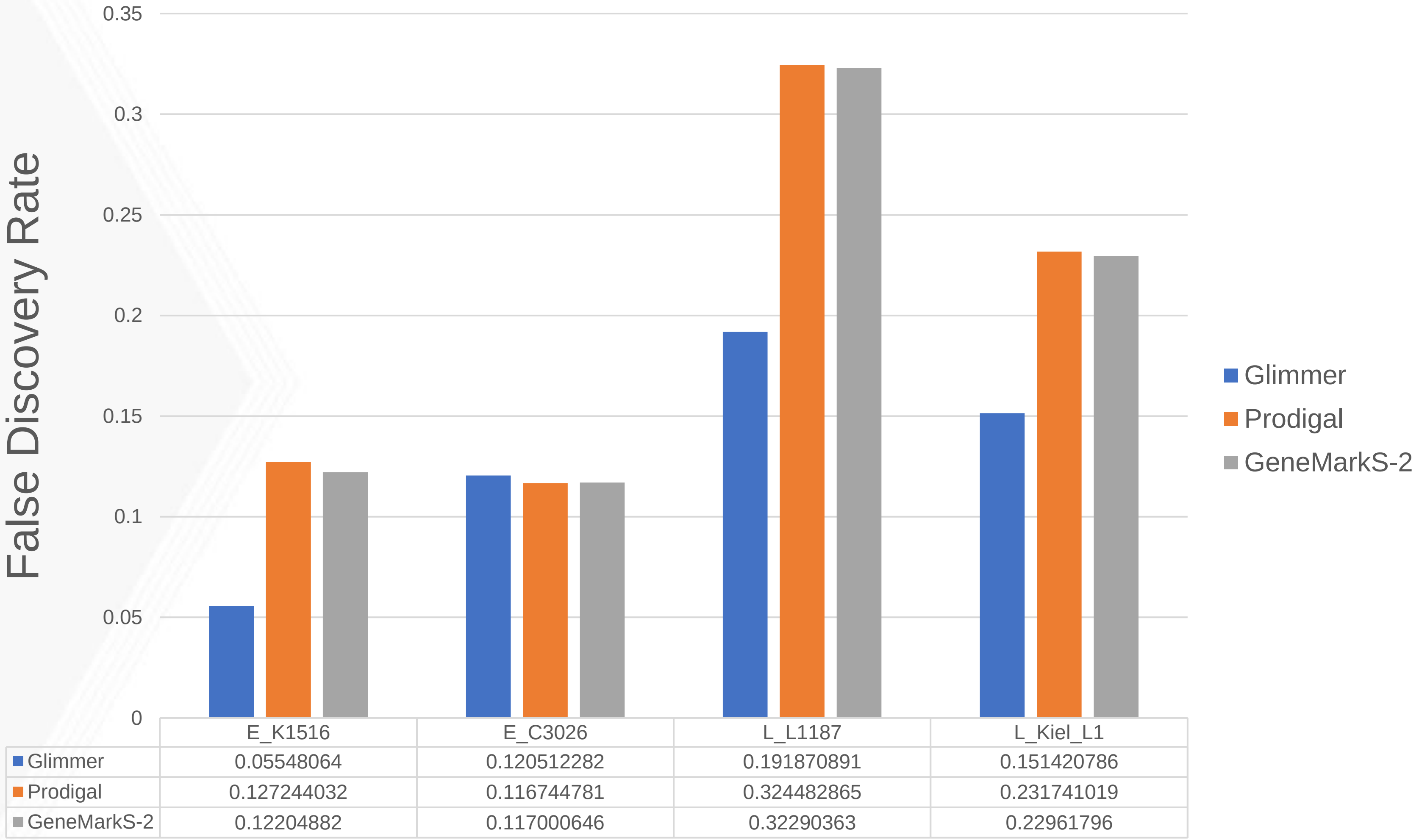
- True positive- predicted genes which matched with protein database
- False positive- predicted genes which did not match with protein database
- False negative- missing protein coding genes from the predicted genes
- True negative- non-protein coding genes

Sensitivity



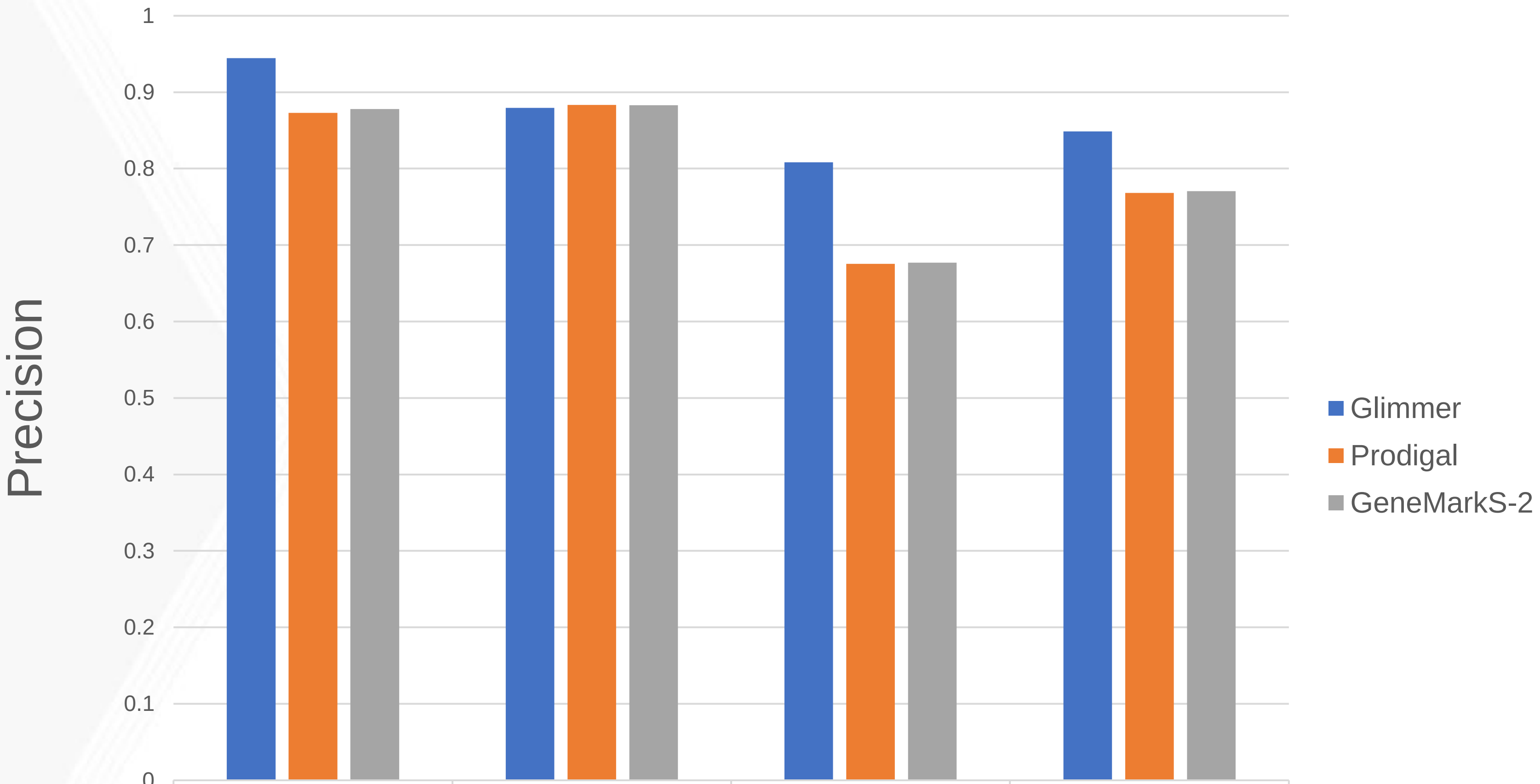
	E_K1516	E_C3026	L_L1187	L_Kiel_L1
🟦 Glimmer	1.004919041	0.937346162	0.931129477	0.898598516
🟧 Prodigal	0.906743185	0.927724323	0.753443526	0.802143446
⬜ GeneMarkS-2	0.899364624	0.916983665	0.745179063	0.806265458

False Discovery Rate



	E_K1516	E_C3026	L_L1187	L_Kiel_L1
■ Glimmer	0.05548064	0.120512282	0.191870891	0.151420786
■ Prodigal	0.127244032	0.116744781	0.324482865	0.231741019
■ GeneMarkS-2	0.12204882	0.117000646	0.32290363	0.22961796

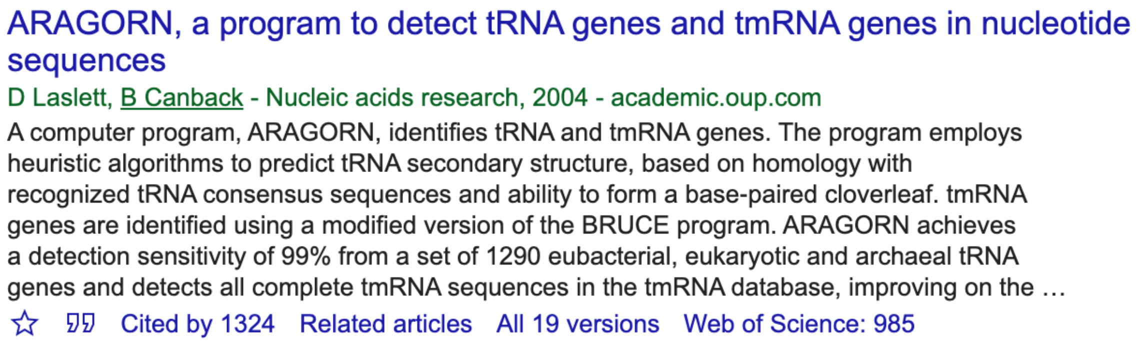
Precision



	E_K1516	E_C3026	L_L1187	L_Kiel_L1
🟦 Glimmer	0.94451936	0.879487718	0.808129109	0.848579214
🟧 Prodigal	0.872755968	0.883255219	0.675517135	0.768258981
◼ GeneMarkS-2	0.87795118	0.882999354	0.67709637	0.77038204

Non-Coding Gene Prediction

ARAGORN



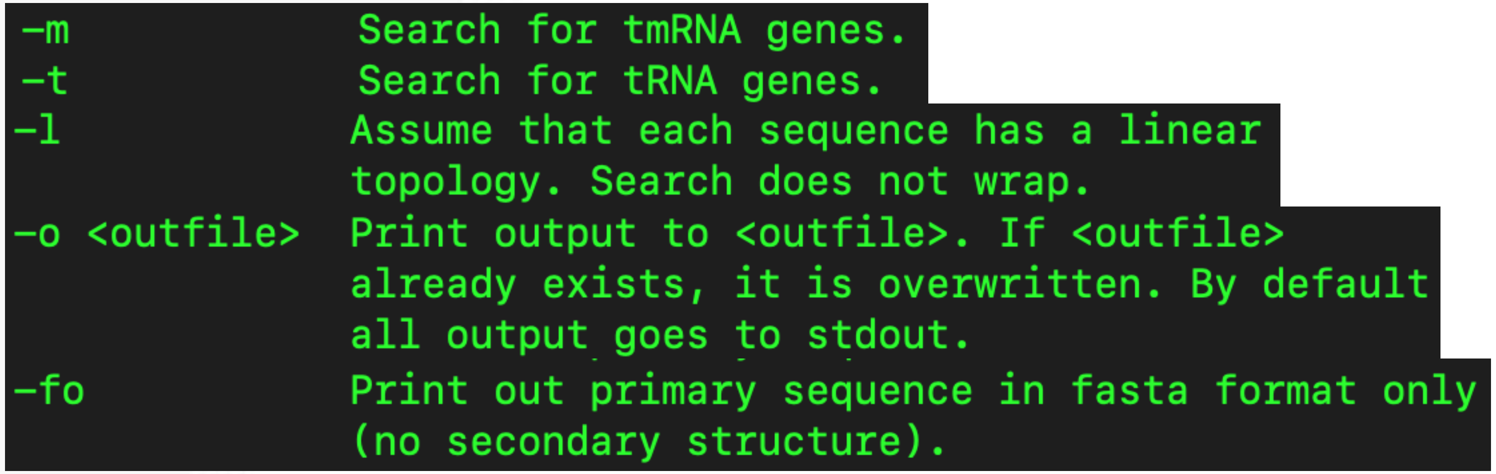
ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences

D Laslett, B Canback - Nucleic acids research, 2004 - academic.oup.com

A computer program, ARAGORN, identifies tRNA and tmRNA genes. The program employs heuristic algorithms to predict tRNA secondary structure, based on homology with recognized tRNA consensus sequences and ability to form a base-paired cloverleaf. tmRNA genes are identified using a modified version of the BRUCE program. ARAGORN achieves a detection sensitivity of 99% from a set of 1290 eubacterial, eukaryotic and archaeal tRNA genes and detects all complete tmRNA sequences in the tmRNA database, improving on the …

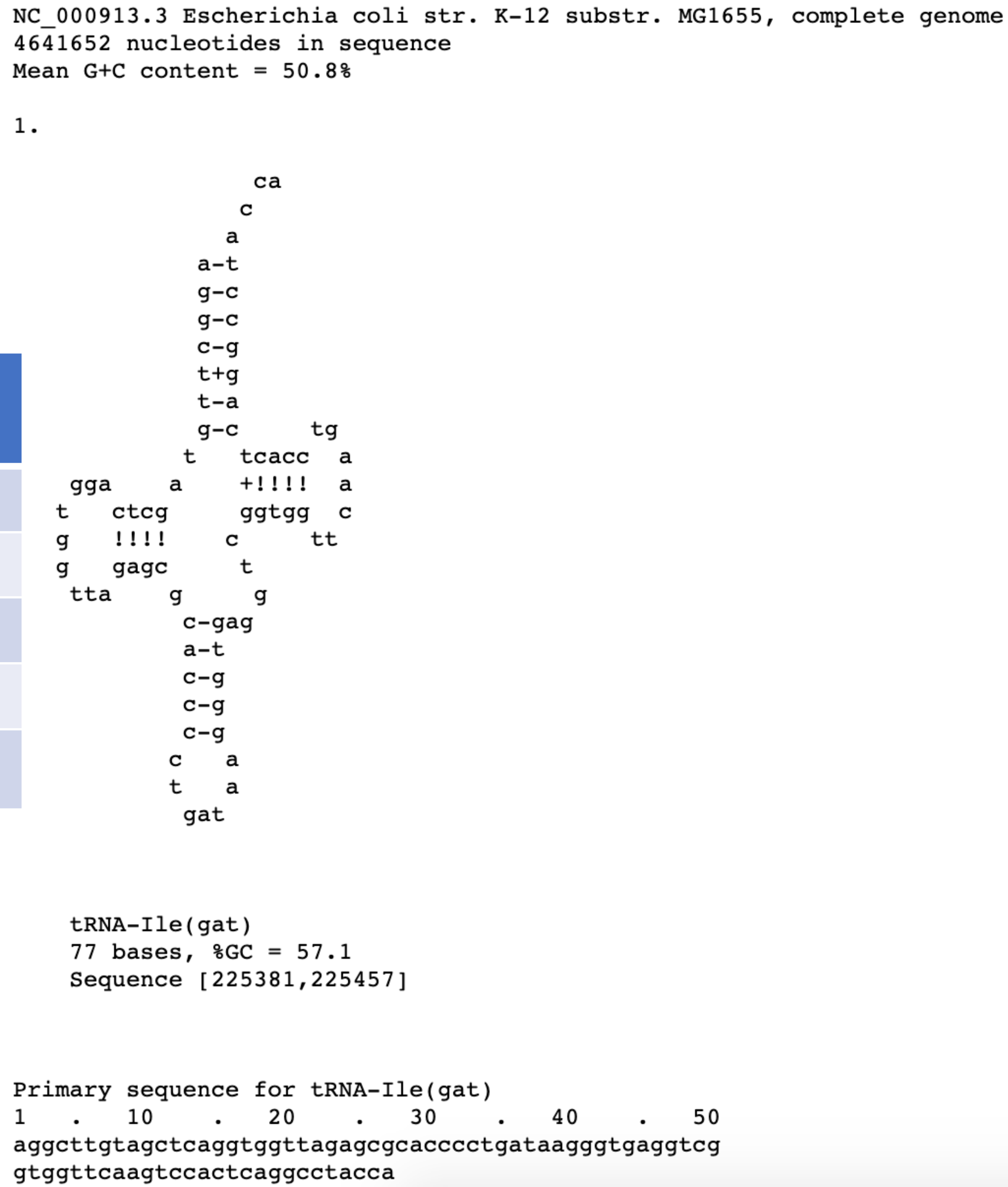
☆ 99 Cited by 1324 Related articles All 19 versions Web of Science: 985

- Identify tRNA and tmRNA genes. (Compare to tRNAscan-SE only identify tRNA)
- The program employs heuristic algorithms to predict tRNA secondary structure
- The output of the program reports the proposed tRNA secondary structure



```
-m              Search for tmRNA genes.
-t              Search for tRNA genes.
-l              Assume that each sequence has a linear
                topology. Search does not wrap.
-o <outfile>    Print output to <outfile>. If <outfile>
                already exists, it is overwritten. By default
                all output goes to stdout.
-fo             Print out primary sequence in fasta format only
                (no secondary structure).
```

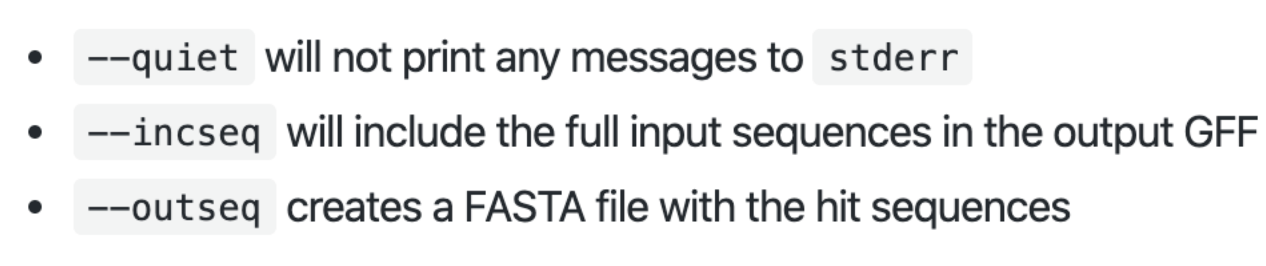
	ARAGORN	tRNAscan	RefSeq
K-12 MG1655	88	87	89
O157:H7 Sakai	105	104	105
IAI39	88	87	88
O83:H1 NRG 857C	84	83	84
O104:H4 2011C-3493	94	93	94



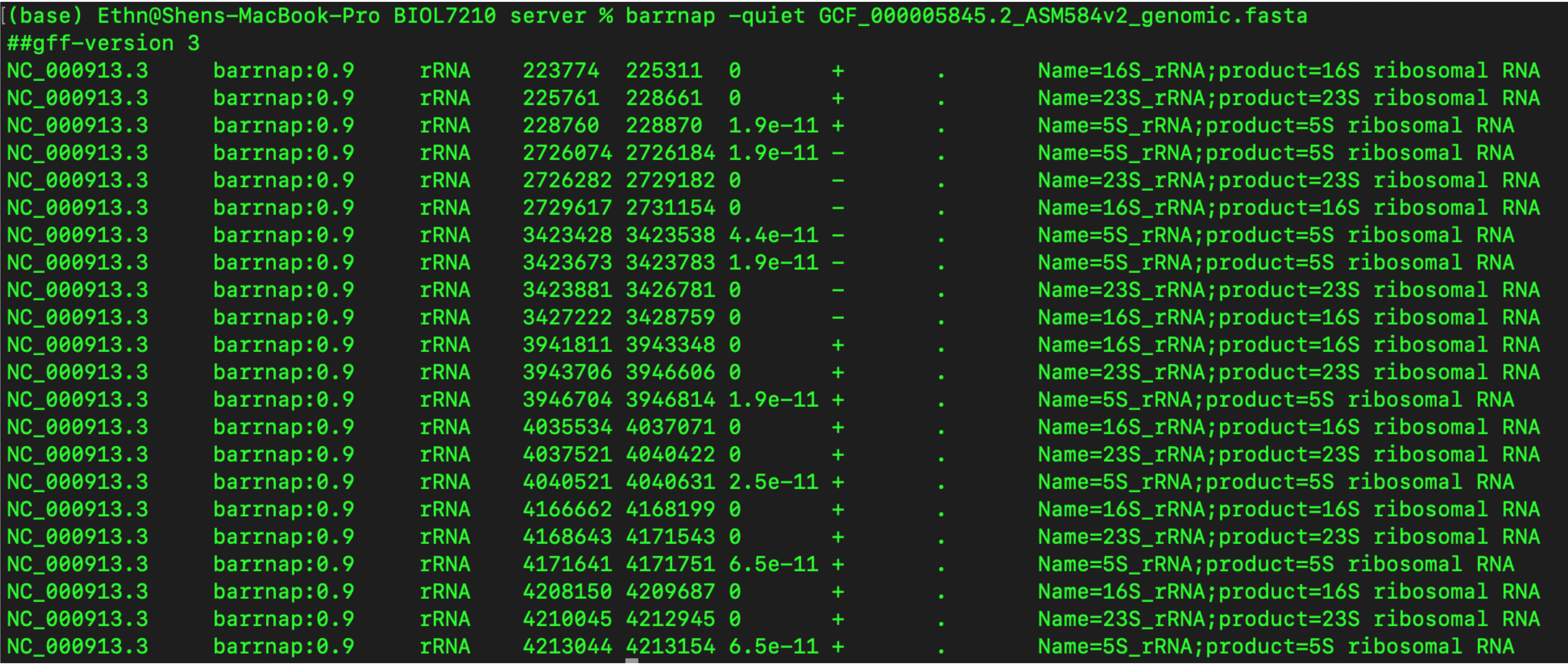
```
NC_000913.3 Escherichia coli str. K-12 substr. MG1655, complete genome
4641652 nucleotides in sequence
Mean G+C content = 50.8%

1.

                 ca
                c
               a
             a-t
             g-c
             g-c
             c-g
             t+g
             t-a
             g-c     tg
            t   tcacc  a
    gga    a    +!!!!  a
   t   ctcg     ggtgg  c
   g   !!!!    c     tt
   g   gagc     t
    tta    g     g
            c-gag
            a-t
            c-g
            c-g
            c-g
           c   a
           t   a
            gat


    tRNA-Ile(gat)
    77 bases, %GC = 57.1
    Sequence [225381,225457]


Primary sequence for tRNA-Ile(gat)
1   .   10    .   20    .   30    .   40    .   50
aggcttgtagctcaggtggttagagcgcacccctgataagggtgaggtcg
gtggttcaagtccactcaggcctacca
```

BARRNAP

- Barrnap predicts the location of ribosomal RNA genes in genomes.
- It takes FASTA DNA sequence as input and write GFF3 as output.
- It uses the new NHMMER tool that comes with HMMER 3.1 for HMM searching in RNA:DNA style.



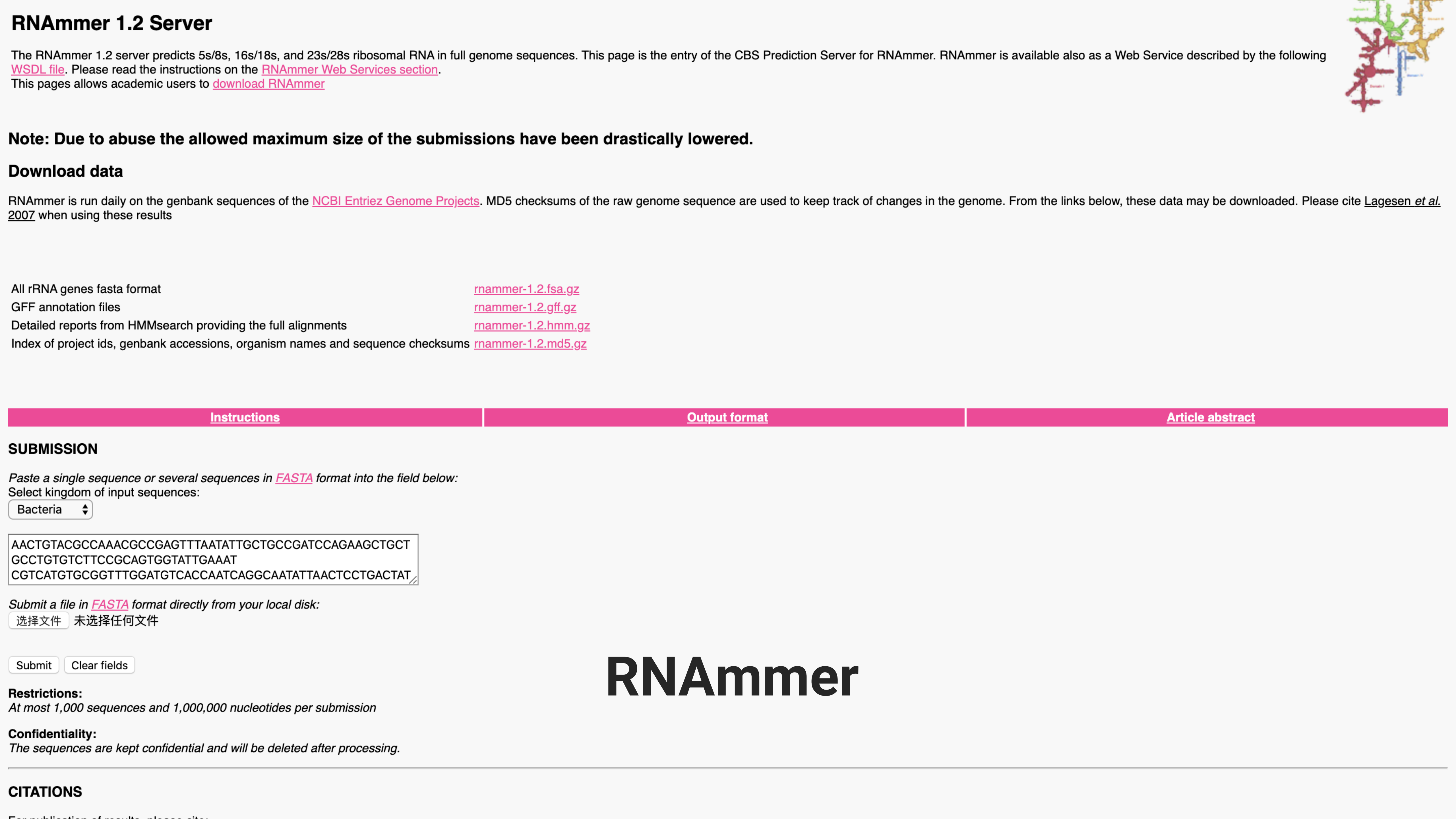
- `--quiet` will not print any messages to `stderr`
- `--incseq` will include the full input sequences in the output GFF
- `--outseq` creates a FASTA file with the hit sequences



```
[(base) Ethn@Shens-MacBook-Pro BIOL7210 server % barrnap -quiet GCF_000005845.2_ASM584v2_genomic.fasta
##gff-version 3
NC_000913.3	barrnap:0.9	rRNA	223774	225311	0	+	.	Name=16S_rRNA;product=16S ribosomal RNA
NC_000913.3	barrnap:0.9	rRNA	225761	228661	0	+	.	Name=23S_rRNA;product=23S ribosomal RNA
NC_000913.3	barrnap:0.9	rRNA	228760	228870	1.9e-11	+	.	Name=5S_rRNA;product=5S ribosomal RNA
NC_000913.3	barrnap:0.9	rRNA	2726074	2726184	1.9e-11	-	.	Name=5S_rRNA;product=5S ribosomal RNA
NC_000913.3	barrnap:0.9	rRNA	2726282	2729182	0	-	.	Name=23S_rRNA;product=23S ribosomal RNA
NC_000913.3	barrnap:0.9	rRNA	2729617	2731154	0	-	.	Name=16S_rRNA;product=16S ribosomal RNA
NC_000913.3	barrnap:0.9	rRNA	3423428	3423538	4.4e-11	-	.	Name=5S_rRNA;product=5S ribosomal RNA
NC_000913.3	barrnap:0.9	rRNA	3423673	3423783	1.9e-11	-	.	Name=5S_rRNA;product=5S ribosomal RNA
NC_000913.3	barrnap:0.9	rRNA	3423881	3426781	0	-	.	Name=23S_rRNA;product=23S ribosomal RNA
NC_000913.3	barrnap:0.9	rRNA	3427222	3428759	0	-	.	Name=16S_rRNA;product=16S ribosomal RNA
NC_000913.3	barrnap:0.9	rRNA	3941811	3943348	0	+	.	Name=16S_rRNA;product=16S ribosomal RNA
NC_000913.3	barrnap:0.9	rRNA	3943706	3946606	0	+	.	Name=23S_rRNA;product=23S ribosomal RNA
NC_000913.3	barrnap:0.9	rRNA	3946704	3946814	1.9e-11	+	.	Name=5S_rRNA;product=5S ribosomal RNA
NC_000913.3	barrnap:0.9	rRNA	4035534	4037071	0	+	.	Name=16S_rRNA;product=16S ribosomal RNA
NC_000913.3	barrnap:0.9	rRNA	4037521	4040422	0	+	.	Name=23S_rRNA;product=23S ribosomal RNA
NC_000913.3	barrnap:0.9	rRNA	4040521	4040631	2.5e-11	+	.	Name=5S_rRNA;product=5S ribosomal RNA
NC_000913.3	barrnap:0.9	rRNA	4166662	4168199	0	+	.	Name=16S_rRNA;product=16S ribosomal RNA
NC_000913.3	barrnap:0.9	rRNA	4168643	4171543	0	+	.	Name=23S_rRNA;product=23S ribosomal RNA
NC_000913.3	barrnap:0.9	rRNA	4171641	4171751	6.5e-11	+	.	Name=5S_rRNA;product=5S ribosomal RNA
NC_000913.3	barrnap:0.9	rRNA	4208150	4209687	0	+	.	Name=16S_rRNA;product=16S ribosomal RNA
NC_000913.3	barrnap:0.9	rRNA	4210045	4212945	0	+	.	Name=23S_rRNA;product=23S ribosomal RNA
NC_000913.3	barrnap:0.9	rRNA	4213044	4213154	6.5e-11	+	.	Name=5S_rRNA;product=5S ribosomal RNA
```

RNAmmer

- An Ab Initio based tool
- Locate rRNA using HMM
- Accepts both prokaryotic and eukaryotic input
- Drawbacks: can not predict lnc-RNA



RNAmmer 1.2 Server

The RNAmmer 1.2 server predicts 5s/8s, 16s/18s, and 23s/28s ribosomal RNA in full genome sequences. This page is the entry of the CBS Prediction Server for RNAmmer. RNAmmer is available also as a Web Service described by the following WSDL file. Please read the instructions on the RNAmmer Web Services section.
This pages allows academic users to download RNAmmer

Note: Due to abuse the allowed maximum size of the submissions have been drastically lowered.

Download data

RNAmmer is run daily on the genbank sequences of the NCBI Entriez Genome Projects. MD5 checksums of the raw genome sequence are used to keep track of changes in the genome. From the links below, these data may be downloaded. Please cite Lagesen *et al.* 2007 when using these results

All rRNA genes fasta format	rnammer-1.2.fsa.gz
GFF annotation files	rnammer-1.2.gff.gz
Detailed reports from HMMsearch providing the full alignments	rnammer-1.2.hmm.gz
Index of project ids, genbank accessions, organism names and sequence checksums	rnammer-1.2.md5.gz

SUBMISSION

Paste a single sequence or several sequences in FASTA format into the field below:
Select kingdom of input sequences:

Bacteria

```
AACTGTACGCCAAACGCCGAGTTTAATATTGCTGCCGATCCAGAAGCTGCT
GCCTGTGTCTTCCGCAGTGGTATTGAAAT
CGTCATGTGCGGTTTGGATGTCACCAATCAGGCAATATTAACTCCTGACTAT
```

Submit a file in FASTA format directly from your local disk:

RNAmmer

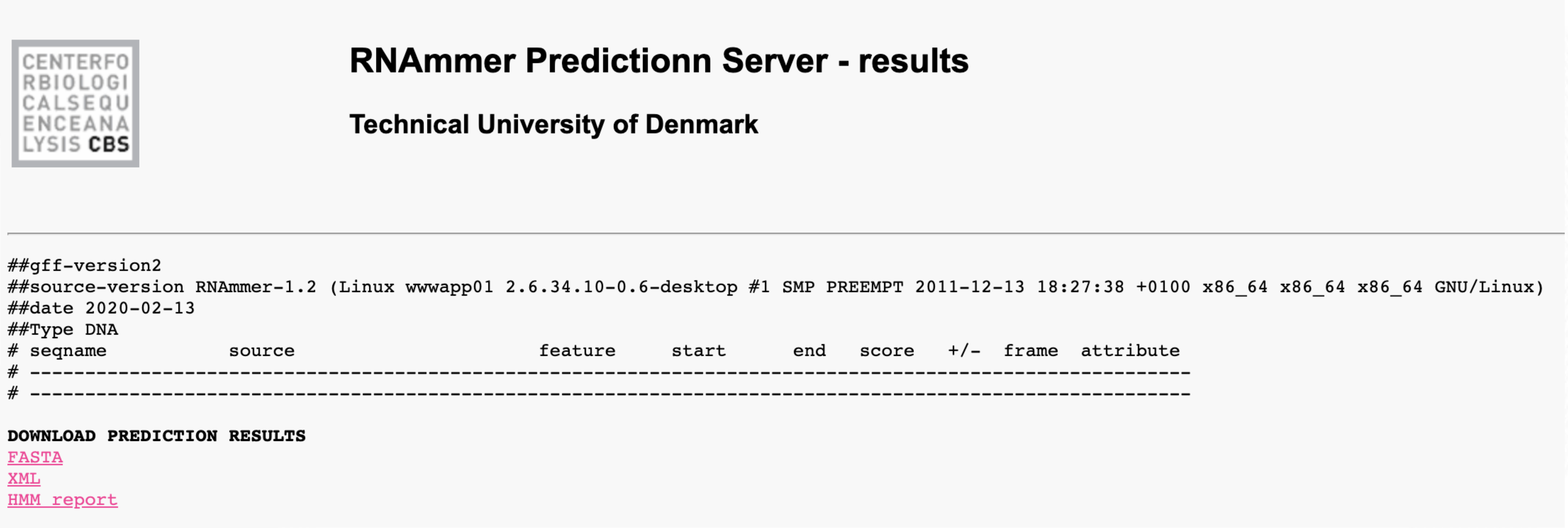
Restrictions:
At most 1,000 sequences and 1,000,000 nucleotides per submission

Confidentiality:
The sequences are kept confidential and will be deleted after processing.

CITATIONS

For publication of results, please cite:

Output

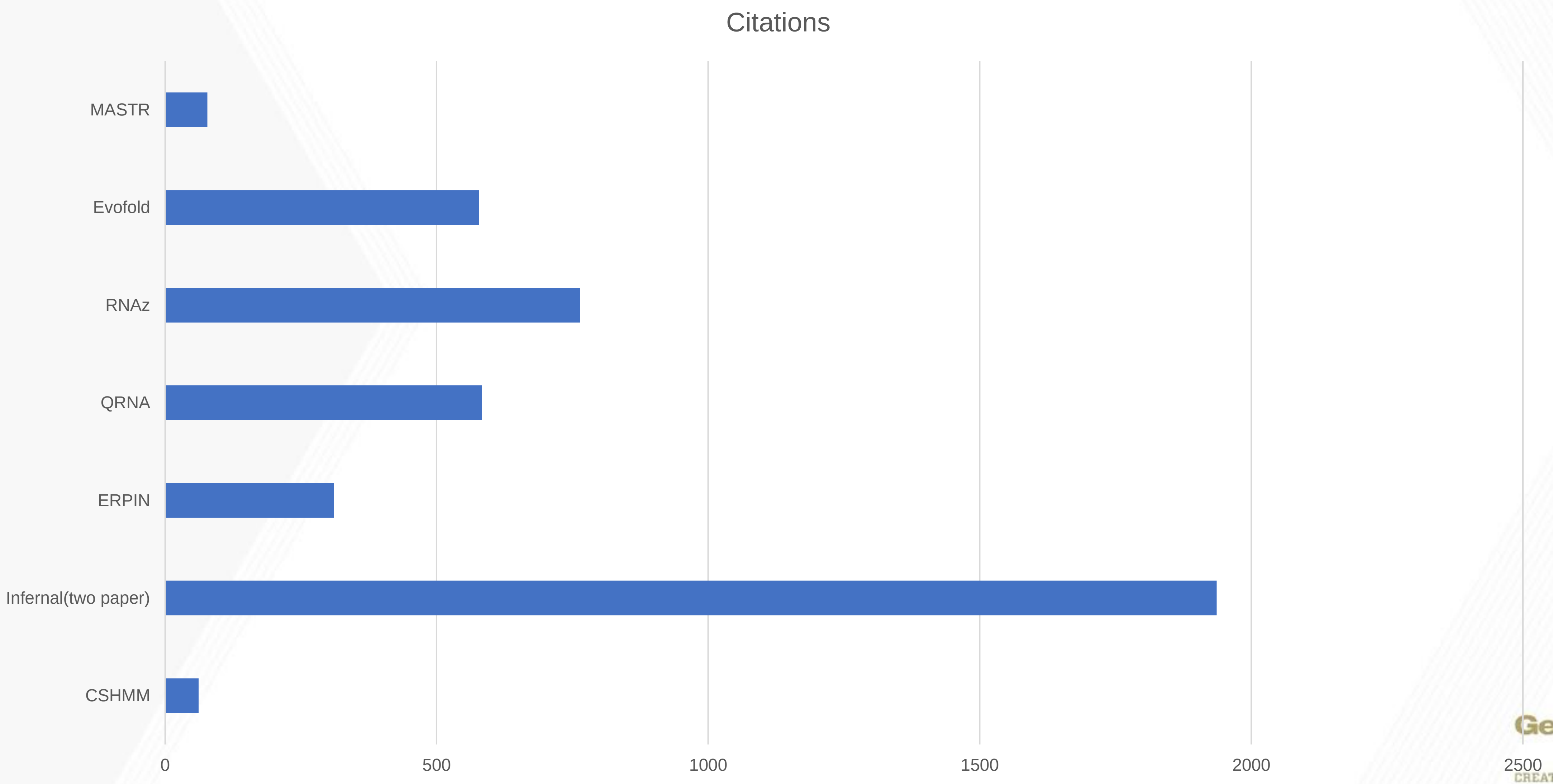


RNAmmer Predictionn Server - results

Technical University of Denmark

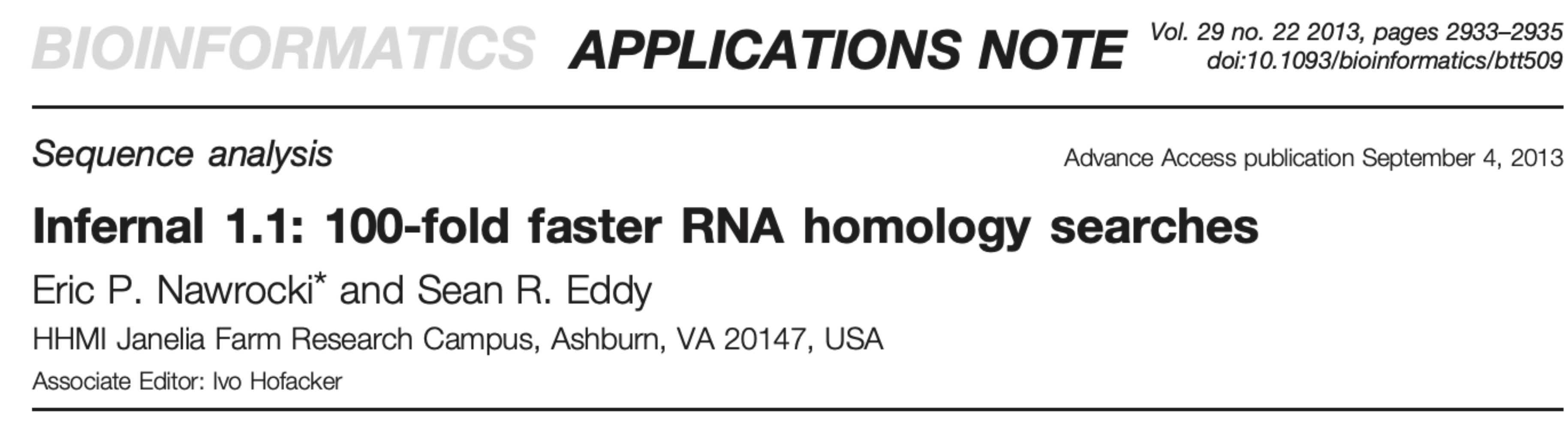
```
##gff-version2
##source-version RNAmmer-1.2 (Linux wwwapp01 2.6.34.10-0.6-desktop #1 SMP PREEMPT 2011-12-13 18:27:38 +0100 x86_64 x86_64 x86_64 GNU/Linux)
##date 2020-02-13
##Type DNA
# seqname           source                      feature     start      end   score   +/-  frame  attribute
# ---------------------------------------------------------------------------------------------------------
# ---------------------------------------------------------------------------------------------------------

DOWNLOAD PREDICTION RESULTS
FASTA
XML
HMM report
```

Homology based non-coding prediction



Infernal

- A homology tools
- Can predict many families of non-coding RNA
- Based a database call Rfam
- Use primary and secondary structure to predict



BIOINFORMATICS **APPLICATIONS NOTE** *Vol. 29 no. 22 2013, pages 2933–2935*
doi:10.1093/bioinformatics/btt509

Sequence analysis Advance Access publication September 4, 2013

Infernal 1.1: 100-fold faster RNA homology searches

Eric P. Nawrocki* and Sean R. Eddy

HHMI Janelia Farm Research Campus, Ashburn, VA 20147, USA

Associate Editor: Ivo Hofacker

Infernal install

-
  ```
  git clone https://github.com/EddyRivasLab/infernal.git infernal
  cd infernal
  git clone https://github.com/EddyRivasLab/easel.git
  git clone https://github.com/EddyRivasLab/hmmer.git
  ```

-
  ```
  ln -s `pwd`/easel/aclocal.m4 hmmer
  ```

-
  ```
  ./configure  --prefix=`pwd`/../infernal_bin
  make
  make install
  cd easel; make install
  ```

Infernal database build

- wget ftp://ftp.ebi.ac.uk/pub/databases/Rfam/12.2/Rfam.cm.gz
  gunzip Rfam.cm.gz
  wget
  ftp://ftp.ebi.ac.uk/pub/databases/Rfam/12.2/Rfam12.2.claninfo
- cmporess Rfam.cm
- We can build a local database for infernal to align

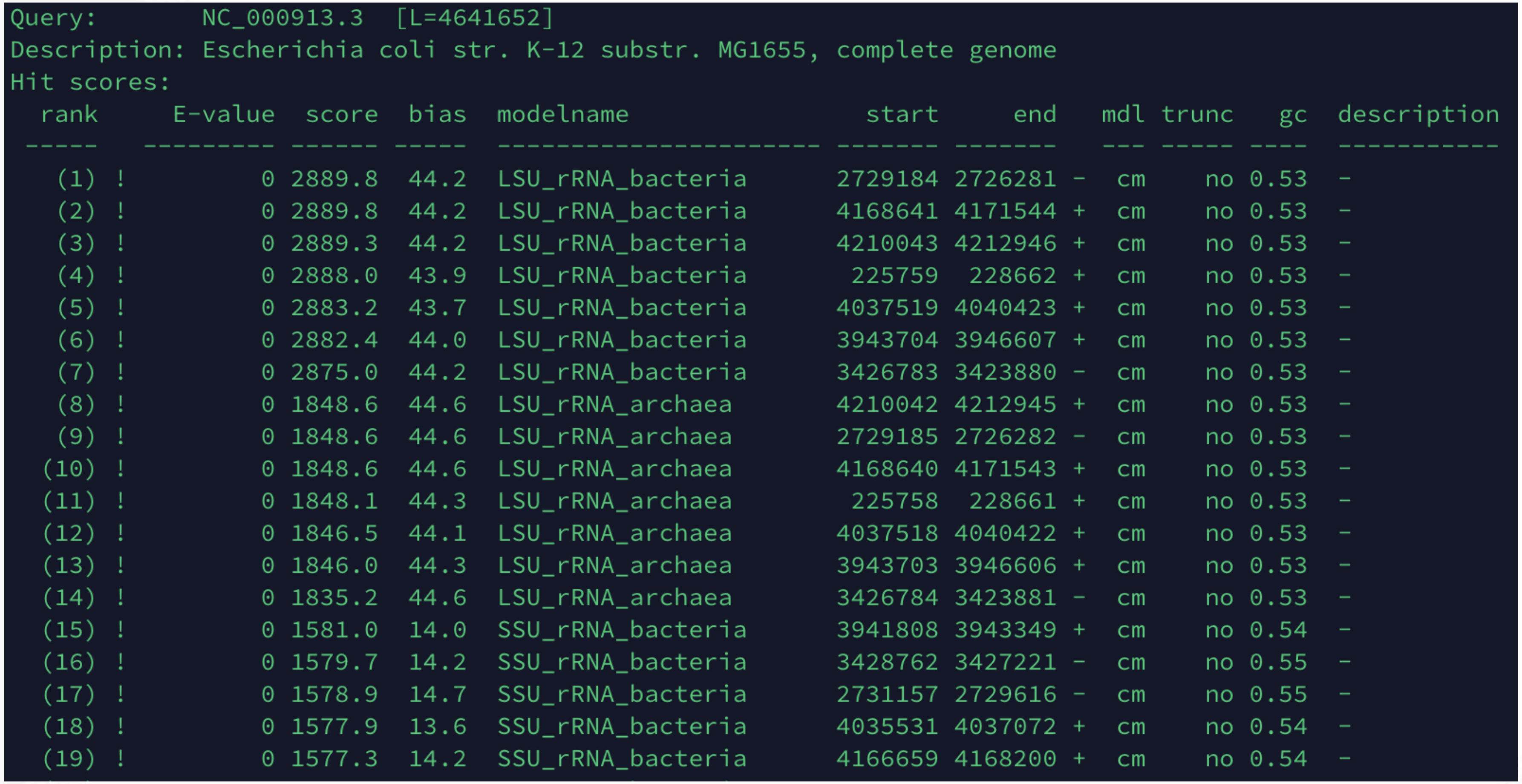
When finished building database

-
  ```
  Working...    done.
  Pressed and indexed 2588 CMs and p7 HMM filters (2588 names
  and 2588 accessions).
  Covariance models and p7 filters pressed into binary
  file:  Rfam.cm.i1m
  SSI index for binary covariance model file:              Rfam.cm.i1i
  Optimized p7 filter profiles (MSV part)  pressed into:     Rfam.cm.i1f
  Optimized p7 filter profiles (remainder) pressed
  into:      Rfam.cm.i1p
  ```

Infernal gene prediction

-
  ```
  cmscan -Z 6 --cut_ga --rfam --nohmmonly --tblout my-genome.tblout --fmt 2 --clanin Rfam12.2.claninfo Rfam.cm my-genome.fa > my-genome.cmscan
  ```

Output



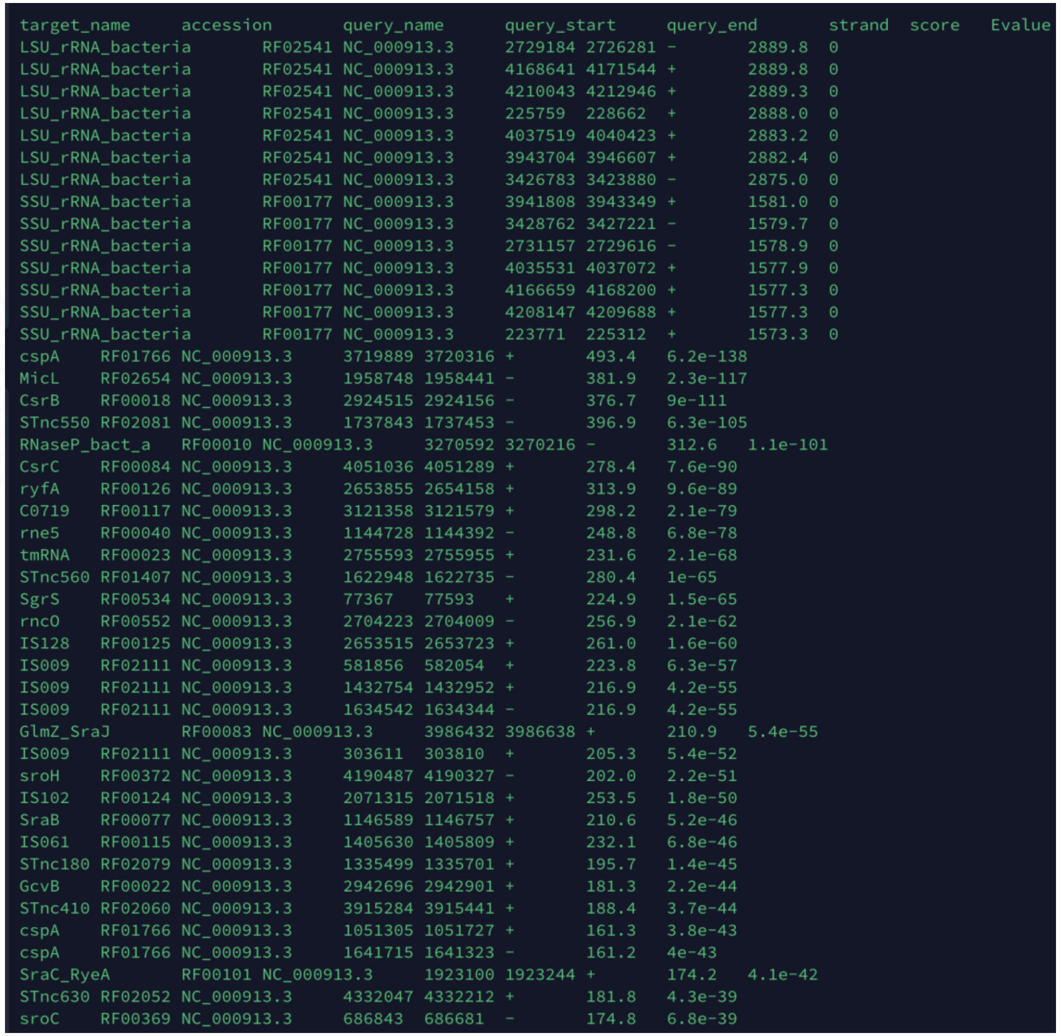
```
Query:       NC_000913.3  [L=4641652]
Description: Escherichia coli str. K-12 substr. MG1655, complete genome
Hit scores:
  rank     E-value  score  bias  modelname                 start     end   mdl trunc   gc  description
 -----   --------- ------ -----  ---------------------- ------- -------   --- ----- ----  -----------
   (1) !         0 2889.8  44.2  LSU_rRNA_bacteria      2729184 2726281 -  cm    no 0.53  -
   (2) !         0 2889.8  44.2  LSU_rRNA_bacteria      4168641 4171544 +  cm    no 0.53  -
   (3) !         0 2889.3  44.2  LSU_rRNA_bacteria      4210043 4212946 +  cm    no 0.53  -
   (4) !         0 2888.0  43.9  LSU_rRNA_bacteria       225759  228662 +  cm    no 0.53  -
   (5) !         0 2883.2  43.7  LSU_rRNA_bacteria      4037519 4040423 +  cm    no 0.53  -
   (6) !         0 2882.4  44.0  LSU_rRNA_bacteria      3943704 3946607 +  cm    no 0.53  -
   (7) !         0 2875.0  44.2  LSU_rRNA_bacteria      3426783 3423880 -  cm    no 0.53  -
   (8) !         0 1848.6  44.6  LSU_rRNA_archaea       4210042 4212945 +  cm    no 0.53  -
   (9) !         0 1848.6  44.6  LSU_rRNA_archaea       2729185 2726282 -  cm    no 0.53  -
  (10) !         0 1848.6  44.6  LSU_rRNA_archaea       4168640 4171543 +  cm    no 0.53  -
  (11) !         0 1848.1  44.3  LSU_rRNA_archaea        225758  228661 +  cm    no 0.53  -
  (12) !         0 1846.5  44.1  LSU_rRNA_archaea       4037518 4040422 +  cm    no 0.53  -
  (13) !         0 1846.0  44.3  LSU_rRNA_archaea       3943703 3946606 +  cm    no 0.53  -
  (14) !         0 1835.2  44.6  LSU_rRNA_archaea       3426784 3423881 -  cm    no 0.53  -
  (15) !         0 1581.0  14.0  SSU_rRNA_bacteria      3941808 3943349 +  cm    no 0.54  -
  (16) !         0 1579.7  14.2  SSU_rRNA_bacteria      3428762 3427221 -  cm    no 0.55  -
  (17) !         0 1578.9  14.7  SSU_rRNA_bacteria      2731157 2729616 -  cm    no 0.55  -
  (18) !         0 1577.9  13.6  SSU_rRNA_bacteria      4035531 4037072 +  cm    no 0.54  -
  (19) !         0 1577.3  14.2  SSU_rRNA_bacteria      4166659 4168200 +  cm    no 0.54  -
```

Georgia Tech
CREATING THE NEXT

Format Transform

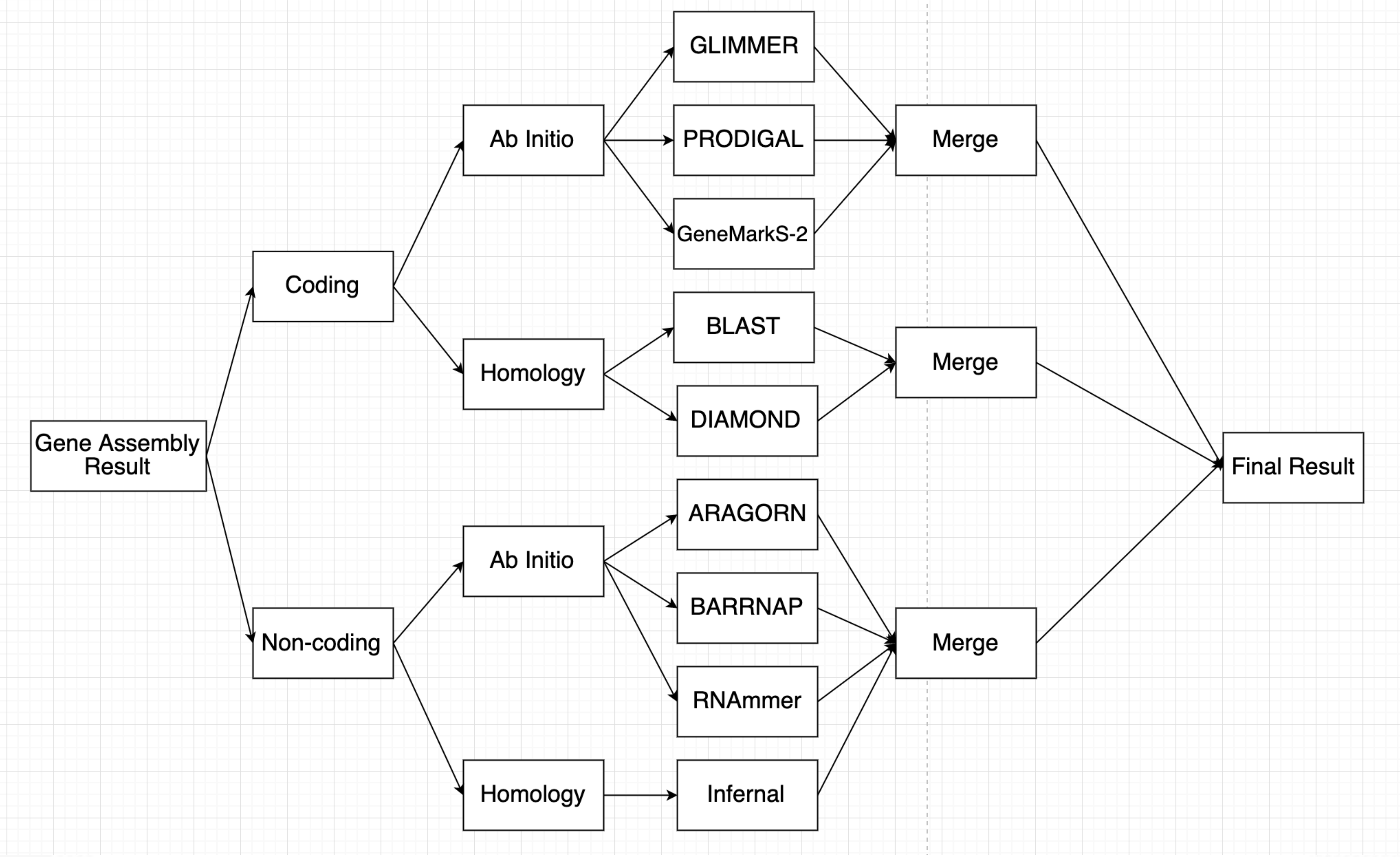
-
  ```
  awk 'BEGIN{OFS="\t";}{if(FNR==1) print "target_name\taccession\tquery_name\tquery_start\tquery_end\tstrand\tscore\tEvalue"; if(FNR>2 && $20!="=" && $0!~/^#/) print $2,$3,$4,$10,$11,$12,$17,$18; }' my-genome.tblout >my-genome.tblout.final.xls
  ```

Final Output



```
target_name     accession       query_name      query_start     query_end       strand  score   Evalue
LSU_rRNA_bacteria       RF02541 NC_000913.3     2729184 2726281 -       2889.8  0
LSU_rRNA_bacteria       RF02541 NC_000913.3     4168641 4171544 +       2889.8  0
LSU_rRNA_bacteria       RF02541 NC_000913.3     4210043 4212946 +       2889.3  0
LSU_rRNA_bacteria       RF02541 NC_000913.3     225759  228662  +       2888.0  0
LSU_rRNA_bacteria       RF02541 NC_000913.3     4037519 4040423 +       2883.2  0
LSU_rRNA_bacteria       RF02541 NC_000913.3     3943704 3946607 +       2882.4  0
LSU_rRNA_bacteria       RF02541 NC_000913.3     3426783 3423880 -       2875.0  0
SSU_rRNA_bacteria       RF00177 NC_000913.3     3941808 3943349 +       1581.0  0
SSU_rRNA_bacteria       RF00177 NC_000913.3     3428762 3427221 -       1579.7  0
SSU_rRNA_bacteria       RF00177 NC_000913.3     2731157 2729616 -       1578.9  0
SSU_rRNA_bacteria       RF00177 NC_000913.3     4035531 4037072 +       1577.9  0
SSU_rRNA_bacteria       RF00177 NC_000913.3     4166659 4168200 +       1577.3  0
SSU_rRNA_bacteria       RF00177 NC_000913.3     4208147 4209688 +       1577.3  0
SSU_rRNA_bacteria       RF00177 NC_000913.3     223771  225312  +       1573.3  0
cspA    RF01766 NC_000913.3     3719889 3720316 +       493.4   6.2e-138
MicL    RF02654 NC_000913.3     1958748 1958441 -       381.9   2.3e-117
CsrB    RF00018 NC_000913.3     2924515 2924156 -       376.7   9e-111
STnc550 RF02081 NC_000913.3     1737843 1737453 -       396.9   6.3e-105
RNaseP_bact_a   RF00010 NC_000913.3     3270592 3270216 -       312.6   1.1e-101
CsrC    RF00084 NC_000913.3     4051036 4051289 +       278.4   7.6e-90
ryfA    RF00126 NC_000913.3     2653855 2654158 +       313.9   9.6e-89
C0719   RF00117 NC_000913.3     3121358 3121579 +       298.2   2.1e-79
rne5    RF00040 NC_000913.3     1144728 1144392 -       248.8   6.8e-78
tmRNA   RF00023 NC_000913.3     2755593 2755955 +       231.6   2.1e-68
STnc560 RF01407 NC_000913.3     1622948 1622735 -       280.4   1e-65
SgrS    RF00534 NC_000913.3     77367   77593   +       224.9   1.5e-65
rncO    RF00552 NC_000913.3     2704223 2704009 -       256.9   2.1e-62
IS128   RF00125 NC_000913.3     2653515 2653723 +       261.0   1.6e-60
IS009   RF02111 NC_000913.3     581856  582054  +       223.8   6.3e-57
IS009   RF02111 NC_000913.3     1432754 1432952 +       216.9   4.2e-55
IS009   RF02111 NC_000913.3     1634542 1634344 -       216.9   4.2e-55
GlmZ_SraJ       RF00083 NC_000913.3     3986432 3986638 +       210.9   5.4e-55
IS009   RF02111 NC_000913.3     303611  303810  +       205.3   5.4e-52
sroH    RF00372 NC_000913.3     4190487 4190327 -       202.0   2.2e-51
IS102   RF00124 NC_000913.3     2071315 2071518 +       253.5   1.8e-50
SraB    RF00077 NC_000913.3     1146589 1146757 +       210.6   5.2e-46
IS061   RF00115 NC_000913.3     1405630 1405809 +       232.1   6.8e-46
STnc180 RF02079 NC_000913.3     1335499 1335701 +       195.7   1.4e-45
GcvB    RF00022 NC_000913.3     2942696 2942901 +       181.3   2.2e-44
STnc410 RF02060 NC_000913.3     3915284 3915441 +       188.4   3.7e-44
cspA    RF01766 NC_000913.3     1051305 1051727 +       161.3   3.8e-43
cspA    RF01766 NC_000913.3     1641715 1641323 -       161.2   4e-43
SraC_RyeA       RF00101 NC_000913.3     1923100 1923244 +       174.2   4.1e-42
STnc630 RF02052 NC_000913.3     4332047 4332212 +       181.8   4.3e-39
sroC    RF00369 NC_000913.3     686843  686681  -       174.8   6.8e-39
```

Proposed workflow



References

1. Goodswen SJ, Kennedy PJ, Ellis JT. Evaluating high-throughput ab initio gene finders to discover proteins encoded in eukaryotic pathogen genomes missed by laboratory techniques. PLoS One. 2012;7(11):e50609. doi: 10.1371/journal.pone.0050609. Epub 2012 Nov 30. PubMed PMID: 23226328; PubMed Central PMCID: PMC3511556.

2. Angelova, Mihaela & Kalajdziski, Slobodan & Kocarev, Ljupco. (2010). Computational Methods for Gene Finding in Prokaryotes. ICT Innovations. 1. 1857-7288

3. Birney E, Durbin R. Using GeneWise in the Drosophila annotation experiment. Genome Res. 2000;10(4):547–548. doi:10.1101/gr.10.4.547

4. Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, David J. Lipman, Basic local alignment search tool, Journal of Molecular Biology,Volume 215, Issue 3

5. Skewes-Cox P, Sharpton TJ, Pollard KS, DeRisi JL (2014) Profile Hidden Markov Models for the Detection of Viruses within Metagenomic Sequence Data. PLoS ONE 9(8): e105067. https://doi.org/10.1371/journal.pone.0105067

6. Lomsadze, Alexandre, et al. "Modeling leaderless transcription and atypical genes results in more accurate gene prediction in prokaryotes." *Genome research* 28.7 (2018): 1079-1089.