eCAMBer: efficient support for large-scale comparative analysis of multiple bacterial strains

Michal Wozniak[1,2], Limsoon Wong[2] and Jerzy Tiuryn[1]

[1]University of Warsaw
[2]National University of Singapore

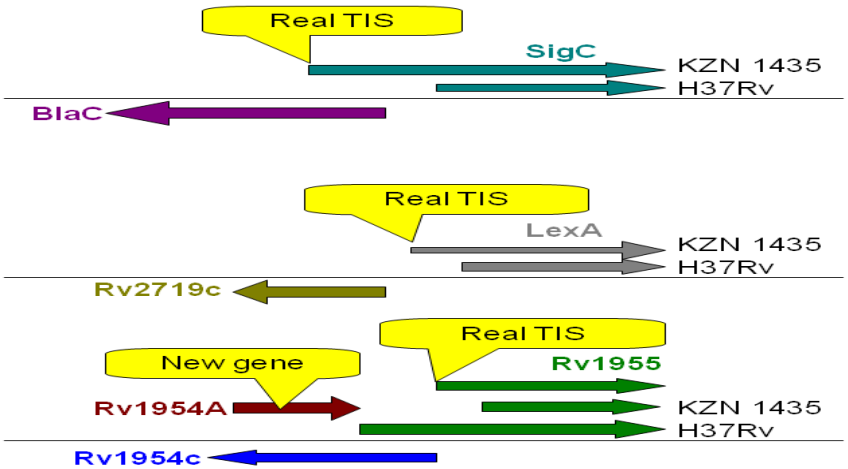
9 October, 2013

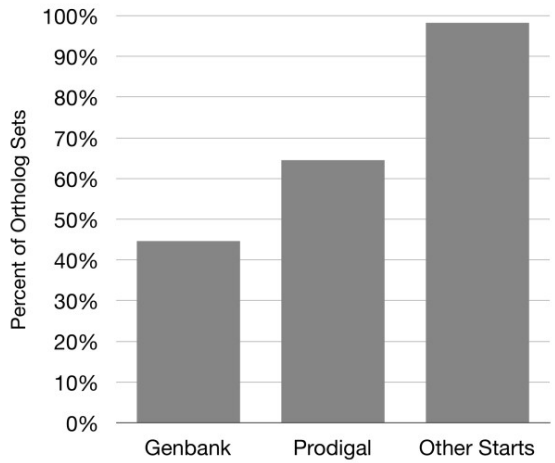
1. Introduction
   - Motivation and goals
2. Methodology
   - General schema of eCAMBer
   - Phase 1 in eCAMBer
   - Phase 2 in eCAMBer
   - Time complexity
3. Results
   - Running times
   - Evalution on the set of 20 E.coli strains
   - Annotation consistency
   - Annotation accuracy
4. Summary
   - Limitations of eCAMBer
   - Summary and conclusions

Annotation inconsistencies

There is a large number of observed inconsistencies are in the genome annotations of bacterial strains. Moreover, it has been shows, that these inconsistencies are often not reflected by sequence discrepancies, but are caused by **wrongly annotated gene starts** as well as **mis-identified gene presence**:

- *Consistency of gene starts among Burkholderia genomes*, BMC Genomics 2011
- *Using comparative genome analysis to identify problems in annotated microbial genomes*, Microbiology 2010





Example of annotation inconsistencies

There are 67 strains of M. tuberculosis in the PATRIC database

- 67 with PATRIC annotations
- 46 with RefSeq annotations

Annotations of the key drug resistance genes:

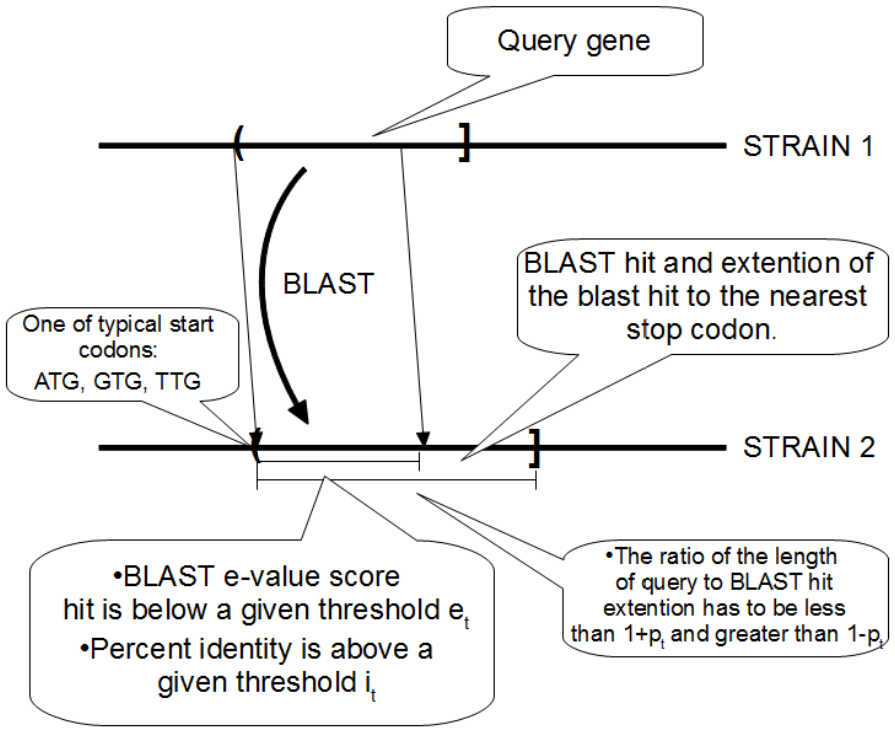
- rpoB: 3 strains with missing annotations in RefSeq
- katG: 5 strains with missing annotations in RefSeq (1 in PATRIC)
- inhA: no strains with missing annotations in RefSeq
- gyrA: no strains with missing annotations in RefSeq
- rpsL: no strains with missing annotations in RefSeq (1 in PATRIC)
- pncA: no strains with missing annotations in RefSeq (1 in PATRIC)

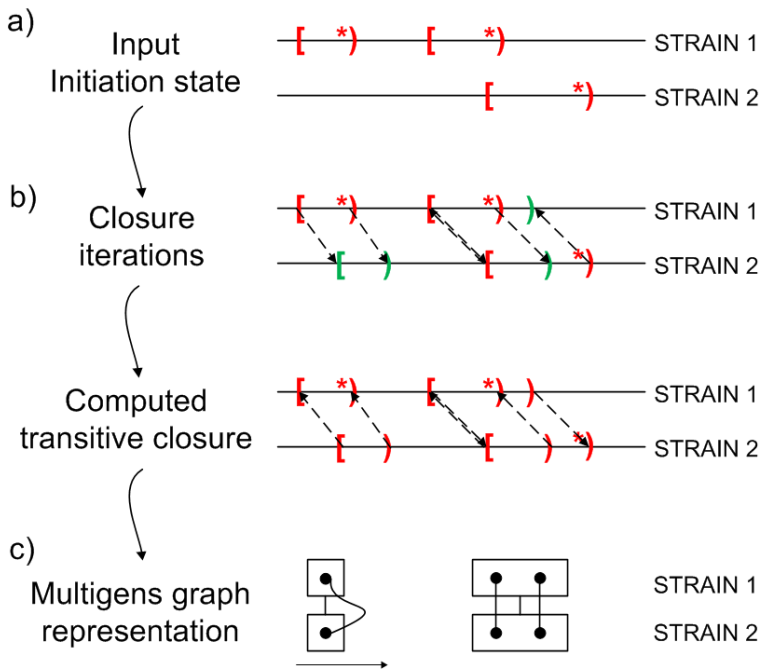
Comparative analysis approaches

It has also been argued, that the consistency and accuracy of annotations may be improved by comparative analysis of these annotations among bacterial strains:

- *Genome majority vote improves gene predictions*,
  PLoS Computational Biology 2011
- *Improving pan-genome annotation using whole genome multiple alignment*,
  BMC Bioinformatics 2011
- *ORFcor: identifying and accommodating ORF prediction inconsistencies for phylogenetic analysis*,
  PLoS ONE 2013
- *CAMBer: an approach to support comparative analysis of multiple bacterial strains*,
  BMC Genomics 2011

Overview of CAMBer





A BLAST hit is acceptable if (default parameters):

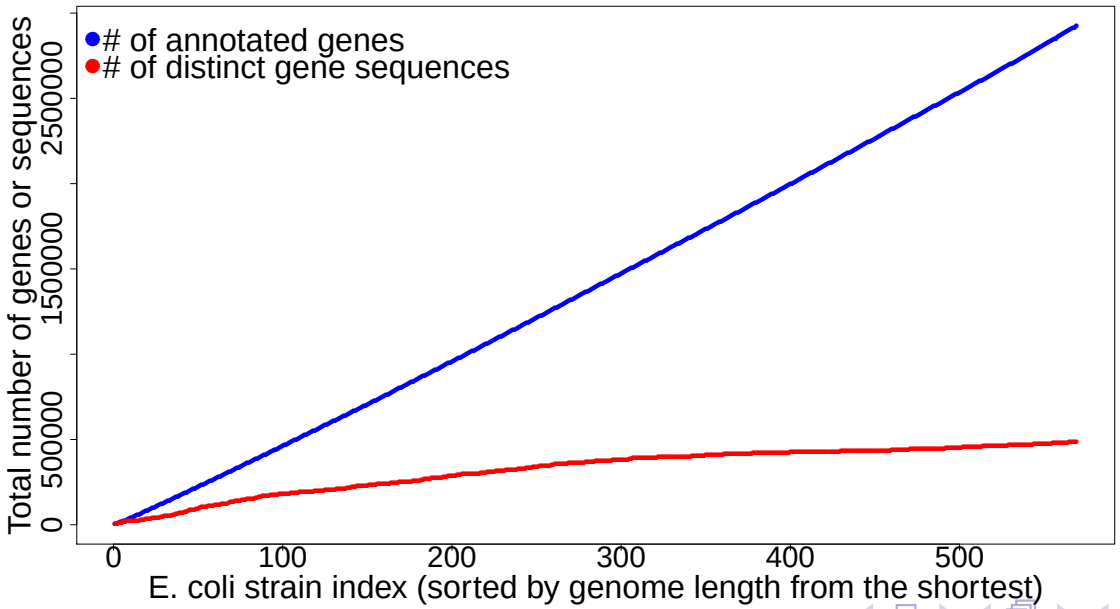
- the hit has one of the appropriate start codons: ATG, GTG, TTG, or the same start codon as in the query sequence,
- BLAST e-value is smaller than 10^{-10},
- the length change is smaller than 0.2,
- the threshold for the percentage of identity is 80% for long sequences and is adjusted for shorter sequences by the HSSP curve.

Major issues with CAMBer

Major issues with CAMBer:

- It propagates annotation errors
- It uses each gene sequence (annotated or predicted) as a BLAST query

The number of gene sequences is much higher than the number of distinct gene sequences!



Goals

Major goals for CAMBer and eCAMBer:

- Goal 1: unification of annotations among bacterial strains,
- Goal 2: identification of annotation inconsistencies.

Major goals for eCAMBer:

- Goal 3: speeding up the closure procedure by avoiding repetitions of sequences used as BLAST queries,
- Goal 4: cleaning up of propagated annotations errors.

General schema of eCAMBer

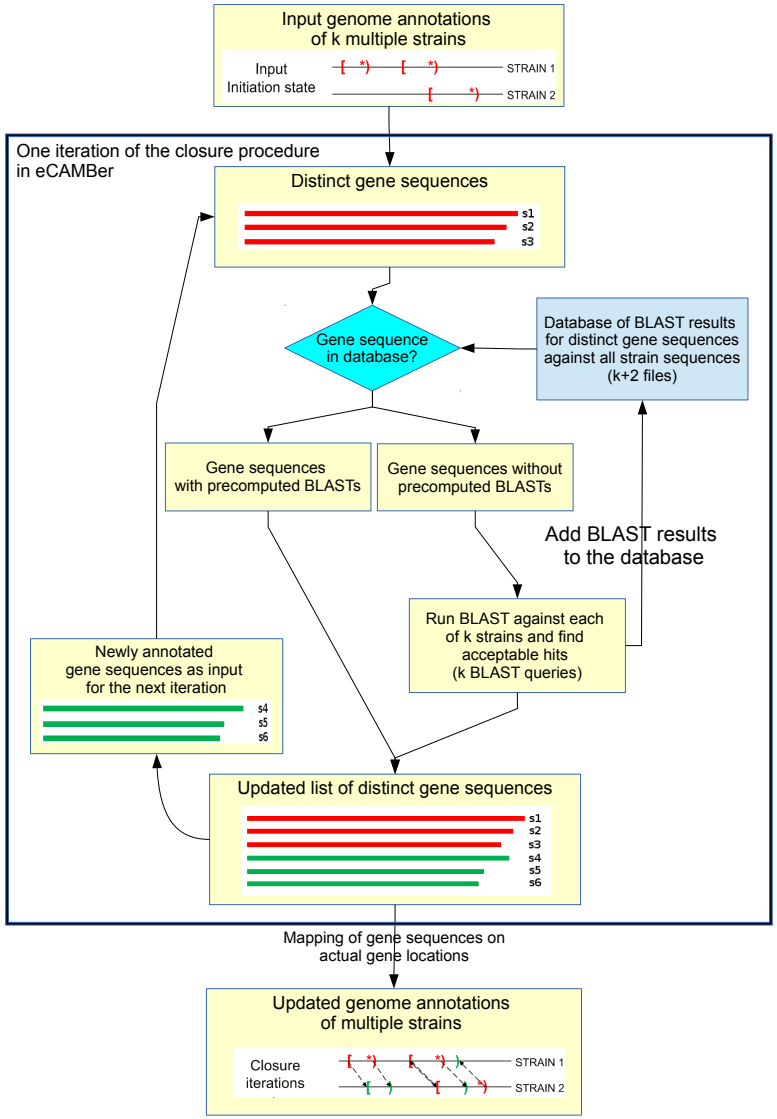
Phase 1:

- modified closure procedure

Phase 2:

- modified *refinement* procedure for splitting homologous gene families into orthologous gene clusters,
- the *TIS voting* procedure for selecting the most reliable TIS,
- the *clean up* procedure procedure for removal of multigene clusters that are likely to be annotation errors propagated during the closure procedure.

Schema of the closure procedure in eCAMBer

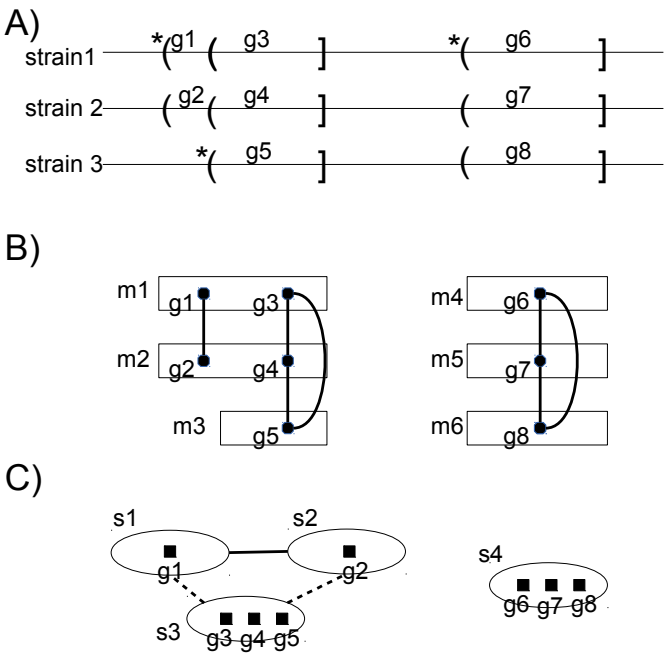


Algorithm 1 The closure procedure (pseudocode)

Require: A set S of bacterial strains; and for each $s \in S$, a set A_s^0 of annotations, a set G_s of sequences constituting the genome of s, and a mapping function $sequences_s(A)$ which returns the set of sequences in the genome G_s corresponding to the set of annotations A.

1: $Q^0 \leftarrow D^0 \leftarrow \bigcup_{s \in S} sequences_s(A_s^0)$
2: $i \leftarrow 0$
3: **while** $Q^i \neq \emptyset$ **do**
4: **for all** $s \in S$ **do**
5: $H_s^i \leftarrow$ acceptable BLAST hit extensions from Q^i on genome G_s
6: $A_s^{i+1} \leftarrow A_s^i \cup H_s^i$
7: **end for**{The above operations are done in parallel for each $s \in S$. Also, for a query sequence $q \in Q^i$, if its BLAST hits are available in a database of precomputed BLAST results, eCAMBer takes results from the database instead.}
8: $H^i \leftarrow \bigcup_{s \in S} sequences_s(H_s^i)$
9: $D^{i+1} \leftarrow D^i \cup H^i$
10: $Q^{i+1} \leftarrow H^i \setminus D^i$
11: $i \leftarrow i + 1$
12: **end while**
13: **return** annotations A_s^i, for all $s \in S$

Sequence consolidation graphs

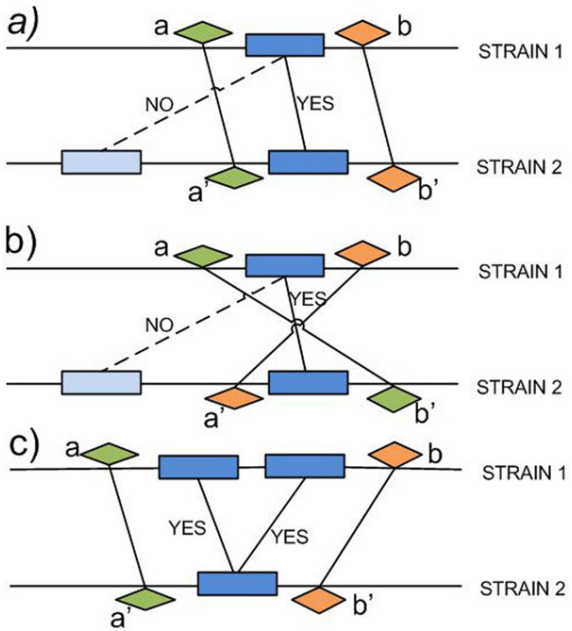


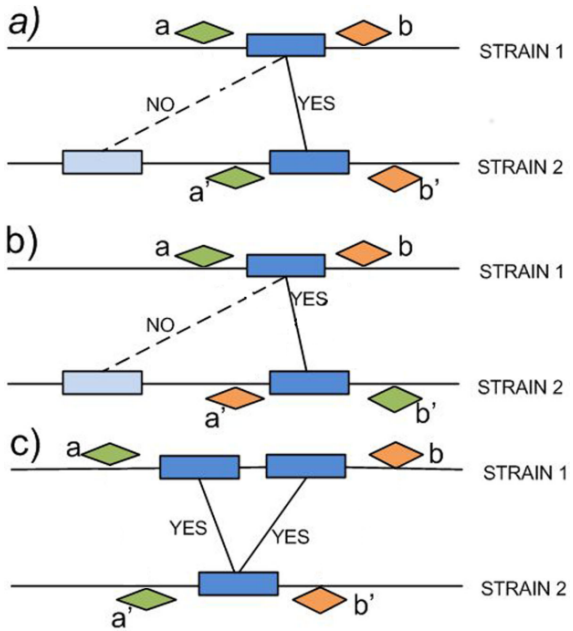
- (A) ORF consolidation graph (V_O, E_O)- nodes represent annotated or predicted ORFs, there is an edge $\{x, y\} \in E_O$ if there was an acceptable BLAST hit between the pair of ORFs,
- (B) multigene consolidation graph (V_M, E_M) - nodes represent multigenes, there is an edge $\{x, y\} \in E_M$ if there was an acceptable BLAST hit between any elements of the pair of multigenes,
- (C) sequence consolidation graph (V_S, E_S, E_B) - nodes represent distinct gene sequences; there is a *shared-end* edge $\{x, y\} \in E_S$ between a pair of sequence nodes if there is a multigene having two elements with these sequences; there is a BLAST-hit edge $\{x, y\} \in E_B$ between a pair of sequence nodes if there is an acceptable BLAST between ORFs x and y.

Refinement procedure

Subsequent steps of the procedure:

- for each strain sort multigenes by positions of stop codons,
- for every pair of strains (s_1, s_2): {in parallel}
  - reconstruct the subgraph of the multigene consolidation graph for non-anchors,
  - for each multigene m on s_1 determine its neighbours, that belong to a multigene cluster with an element on s_2,
  - for each non-anchor edge between a pair of multigenes on s_1 and s_2 check if it is supported (remove if not supported).





TIS voting procedure

For each multigene m in each multigene cluster c, we try to find a TIS (originally annotated or transferred) that belongs to a connected component of the ORF consolidation graph, where the connected component satisfies the following two conditions:

- (i) it has TISs (originally annotated or transferred) present in at least 80% of the multigenes in c; and
- (ii) it has TISs originally annotated in at least 50% of the multigenes in c, or it has TISs originally annotated in at least twice the number of multigenes in c than all other connected components in c.

If such a TIS is found, it is selected as the TIS for m. If such a TIS is not found, but m has an originally annotated TIS, then the originally annotated TIS is selected as the TIS for m. Otherwise, the longest ORF in the multigene m is selected. After the TIS voting procedure, every multigene has exactly one TIS selected.

Clean up procedure

The input for this procedure consists of the set of multigene clusters C^* and multigene annotations M_s, for each strain $s \in S$. For each multigene cluster $c \in C^*$ we compute the following features:

- (i) l, the median multigene length in c,
- (ii) p, the ratio of the number of strains with at least one element from c to the total number of strains;
- (iii) r, the ratio of the number of originally annotated multigenes to the total number of multigenes in c;
- (iv) v, the ratio of the number of multigenes in the cluster that are overlapped by a longer multigene to the total number of multigenes in the cluster.

Then, we update the set of multigene clusters C^*, by removing of multigene clusters for which: $(p < \frac{1}{3} \text{ or } r < \frac{1}{3}) \text{ and } (l < 150 \text{ or } v > 0.5)$.

Closure procedure in CAMBer vs. eCAMBer

- k - number of strains
- n - number of ORF sequences
- d - number of distinct ORF sequences
- $O(kn \cdot k)$ of BLAST computations for CAMBer
- $O(d \cdot k)$ of BLAST computations for eCAMBer

Case study of 64 M. tuberculosis strains

- number of ORFs: 669620
- number of multigenes: 350774
- number of distinct ORF sequences: 60854
- number of edges in ORF consolidation graph: 23177547
- number of edges in multigene consolidation graph: 10875300
- number of edges in sequence consolidation graph: 139885

Comparison of running times

CAMBer vs. eCAMBer on four datasets from the CAMBer paper (using 4 processors):

	CAMBer		eCAMBer	
Dataset	BLASTs	closure	BLASTs	closure
2 str. of S. aureus	1min 47s	2min 5s	8s	18s
9 str. of M. tuberculosis	1h 22min	1h 27min	27s	41s
22 str. of S. aureus	6h	6.5h	3min 15s	4min
41 str. of E. coli	42h	48.5h	22min	25min

CAMBer vs. eCAMBer vs. Mugsy-Annotator on a single processor:

	CAMBer		eCAMBer		
Dataset	BLASTs	closure	BLASTs	closure	Mugsy-Annotator
2 str. of S. aureus	7min 9s	7min 31s	9s	20s	8s
9 str. of M. tuberculosis	4h 10min	4h 12min	1min	2min	7min 42s
22 str. of S. aureus	36h 54min	37h 5min	8min 30s	14min 57s	3h 23min
41 str. of E. coli	272h 30min	273h 22min	1h 1min	1h 40min	8h 56min

Running times on large datasets

Running times of eCAMBer on large datasets (using 20 processors):

Detaset description			Running times					
Dataset desc.	# of genes	# of distinct seq.	BLASTs	closure	graph	refinement	TIS voting	clean-up
E. coli (569)	2923165	487141 (0.17)	7h 46min	12h	59min	2h 51min	14min	10min
S. enterica (293)	1366439	244450 (0.18)	3h 39min	3h 56min	18min	36min	4min	4min
S. agalactiae (250)	517648	56215 (0.11)	5min	29min	2min	5min	37s	53s
S. pneumoniae (238)	529076	99578 (0.19)	2h 16min	2h 29min	5min	9min	1min 30s	1min 10s
S. aureus (195)	523557	98562 (0.19)	59min	1h 7min	3min	4min	1min 50s	1min
H. pylori (163)	267302	208790 (0.78)	1h 36min	1h 42min	12min	5min	5min 10s	2min 10s
L. interrogans (139)	649916	175899 (0.27)	1h 22min	1h 30min	4min	7min	1min 30s	1min 50s
V. cholerae (130)	467413	97258 (0.21)	22min	24min	2min	2min 20s	35s	51s
A. baumannii (131)	487775	129089 (0.27)	31min	34min	3min	2min 30s	52s	58s
B. cereus (104)	602986	395477 (0.66)	58min	1h 13min	6min	3min 50s	2min 57s	1min 52s

Comparison of graph sizes

Selected statistics for the largest dataset of 569 strains of *E. coli*:

- 12.4mln nodes in the ORF consolidation graph (ORF annotations),
- 1.6mln nodes in the sequence consolidation graph (unique ORF sequences),
- 2.8bln edges in the ORF consolidation graph
- 1.3mln shared-end edges in the sequence consolidation graph
- 55.9mln BLAST-hit edges in the sequence consolidation graph

Input dataset

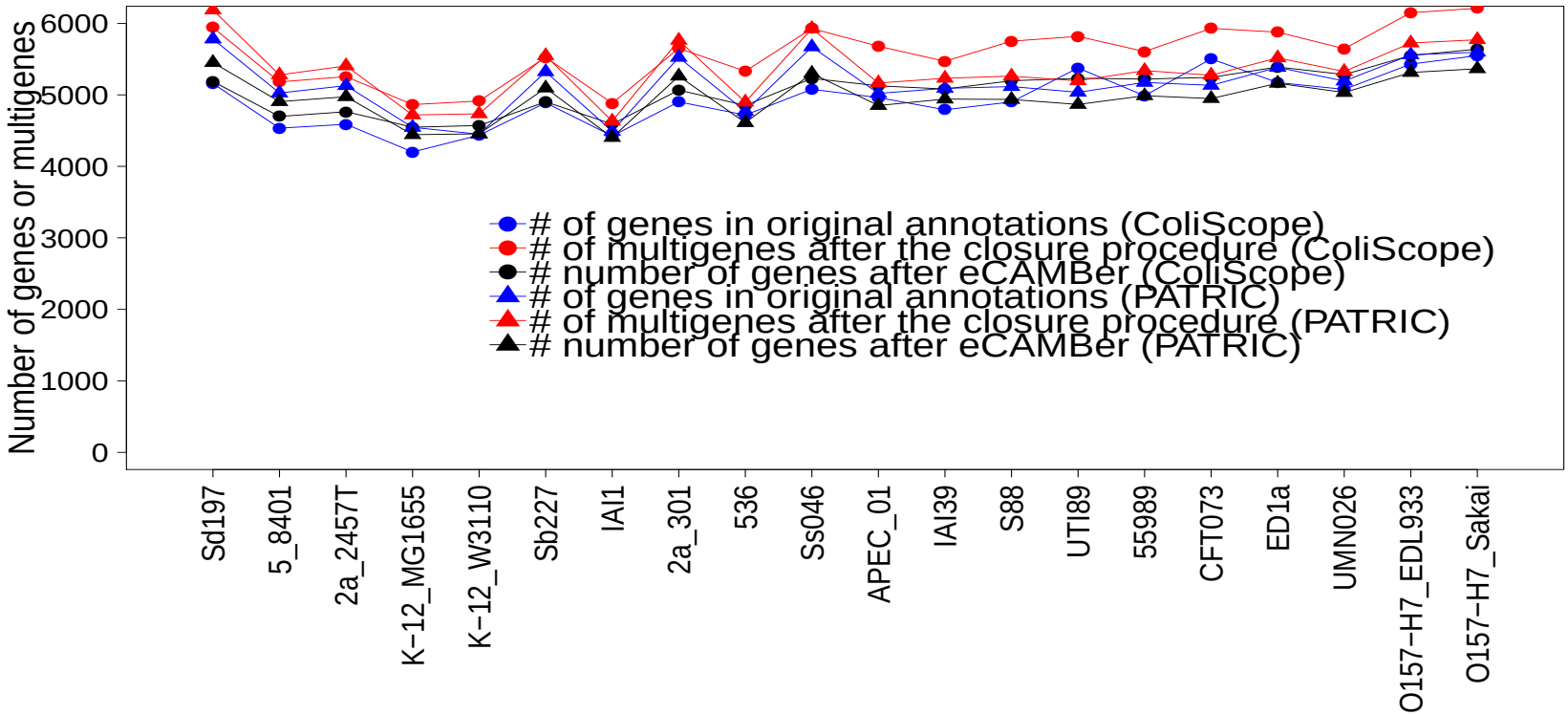
We use a dataset of 20 *E. coli* strains with manually curated annotations, deposited in the ColiScope database. The annotations were published with the work:
Organized genome dynamics in the Escherichia coli species results in highly diverse adaptive paths (PLoS Genet. 2009).

Experimental support

There are 923 genes with experimental support (EcoGene 3 database) for strain K-12 1655, out of which:

- 903 are present in the ColiScope annotations;
- 833 are present in the PATRIC annotations.

Number of genes before and after



The mean absolute difference in the number of annotated multigenes between two neighbour strains (sorted in the order of increasing genome sizes):

- 311 for the ColiScope annotations from ColiScope vs. 181 after eCAMBer
- 409 for the PATRIC annotations and 323 after applying eCAMBer.

Inconsistencies in the ColiScope dataset

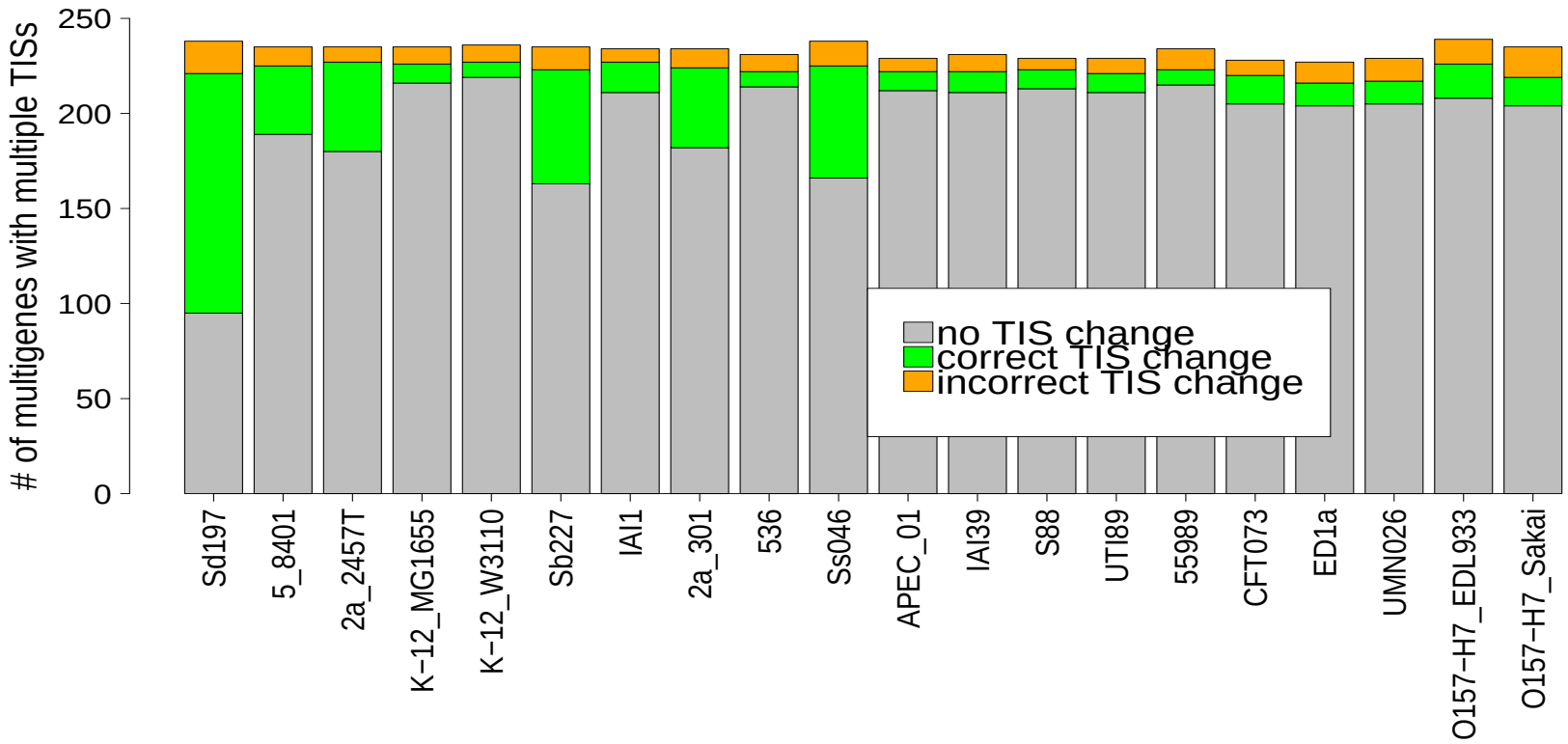
Putative missing gene annotations

There are 73 gene families which have a multigene in every strain, and exactly one missing original annotation. The top four strains with the highest number of missing gene annotations of that type are: Sd197 (12), 2a 2457T (8), 536 (7) and Sb227 (7). The most well-studied strain K-12 MG1655 has four missing annotations of the above type.

Inconsistent TIS annotations

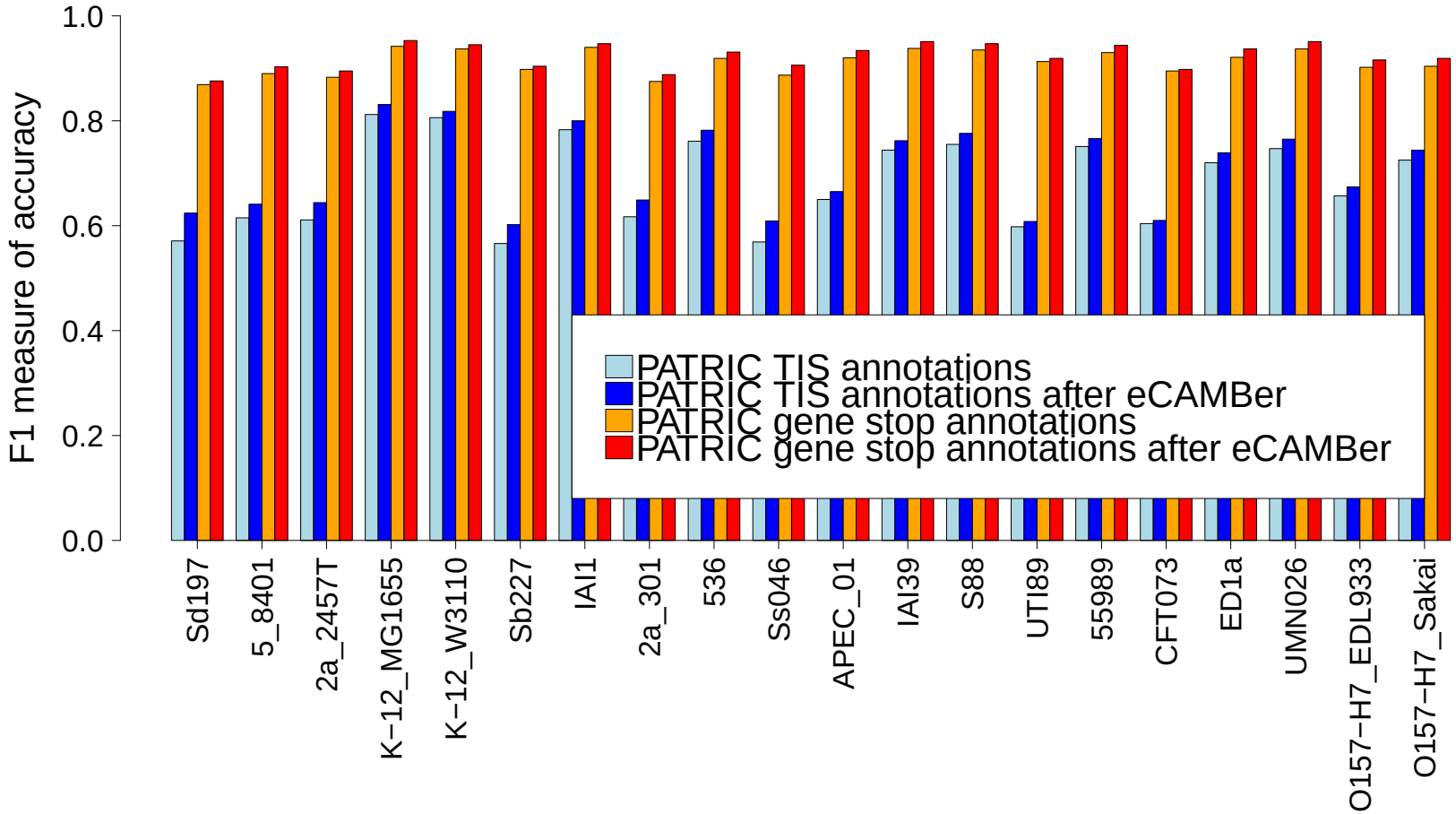
There are 3923 pairs of annotated genes with different TISs, but with identical sequence (including 100bp. upstream region from the TIS of the longer annotation). This number was reduced to 482 after applying the TIS majority voting procedure and the clean up procedure.

Statistics for the TIS voting procedure



There are 1134 gene families in the ColiScope dataset with consistent TIS annotations (gold standard). For about 240 (depending on strain) of them annotations of TIS in the PATRIC database were unambiguous. In total 534 (74% out of 739) changes were correct.

Overall accuracy (f1 measure)



Limitations of eCAMBer

- eCAMBer only purely on the quality of original annotations. Thus, for example, eCAMBer cannot identify genes, whose annotations were missing for all strains;
- eCAMBer excludes pseudogenes and non-protein coding genes from the analysis. This follows from the assumption that eCAMBer considers only genes that start with start codon, end with stop codon, and have length divisible by 3.

Summary and conclusions

- eCAMBer is a tool to unify annotations among bacterial strains within the same species,
- eCAMBer is more efficient than CAMBer and scales up to datasets comprising hundrends of bacterial strains,
- eCAMBer improves overall annotation consistency and accuracy,
- it supports downloading genome sequences and genome annotations from the PATRIC database, for the set of selected strains within a species,
- eCAMBer generates output compatible with CAMberVis, a tool for simultaneous visualization of multiple genome annotations of bacterial strains,
- the project webpage: *http://bioputer.mimuw.edu.pl/ecamber*.

Thank you

Thank you!

You are welcome to give comments or ask questions.