BCVapp: analyzing population sequencing chromatograms for pathogen identification

Central Research Institute of Epidemiology, Moscow, Russia
2014

BCV – Base Caller with Vocabulary

Fantin YS et al. 2013, PLoS ONE 8(1): e54835

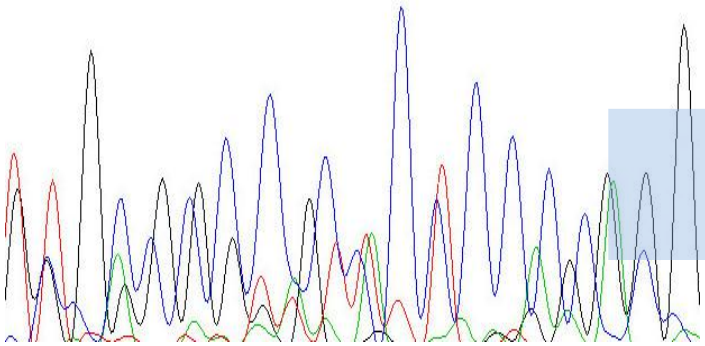
BCV is intended for analysis of direct (population) sequencing chromatograms using a vocabulary of sequences similar to the target DNA.

BCVapp is a web application, which uses a collection of 16S rRNA sequences from human microbiome as a vocabulary and is intended for processing .ab1 chromatograms obtained from clinical samples by direct sequencing of 16S rRNA gene. It also assigns definitive taxonomic classification to the predicted sequences with the help of STAP (An Automated Phylogenetic Tree-Based Small Subunit rRNA Taxonomy and Alignment Pipeline)

BCV-pipeline: BCV

BCV performs mixture deconvolution, generating a set of sequences that could presumably produce such a chromatogram.

1. **Input:** Sanger chromatogram in . ab1 format, obtained from mixed DNA sample



2. **Output:** .fasta file with probable components of the mixture

```
>cluster_1 0.176055
CTTTCTGGTAGTTACCGTCCTGTGTGAACAATCACTCTCACACACGTTCTTC

>cluster_2 0.364073
CTTTCTGGTAGTTACCGTCCT-GTGTGAACACTCACTCTCACACACGTTCTT

>cluster_3 0.432982
CTTTCTGGTAGTTACCGTCATGTGTGAACAATCCCTCTCACACACGTTCTTC

>cluster_4 0.0268912
CTTTCTGGTAGTTACCGTCCTGTGTGAACACTACCTCTCACACACGTTCTTC
```

0.0268912 - expectation to be a component of the mixture. Sequences with expectation > 0.05 are chosen for further analysis.

BCV-pipeline: taxonomic identification is generally performed by STAP

3. Taxonomy is assigned to each sequence with expectation > 0.05 (detailed output).

Corynebacterium unclassified

Staphylococcus haemolyticus

Staphylococcus haemolyticus

Staphylococcus

4. Simple output: a list of taxons present in the sample.

Corynebacterium unclassified

Staphylococcus haemolyticus

Taxonomy: GreenGenes & STAP

GreenGenes database offers annotated, chimera-checked, full-length 16S rRNA gene sequences and provides highly reliable taxonomy for them.

DeSantis, T. Z., P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and G. L. Andersen. 2006. Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. Appl Environ Microbiol 72:5069-72.

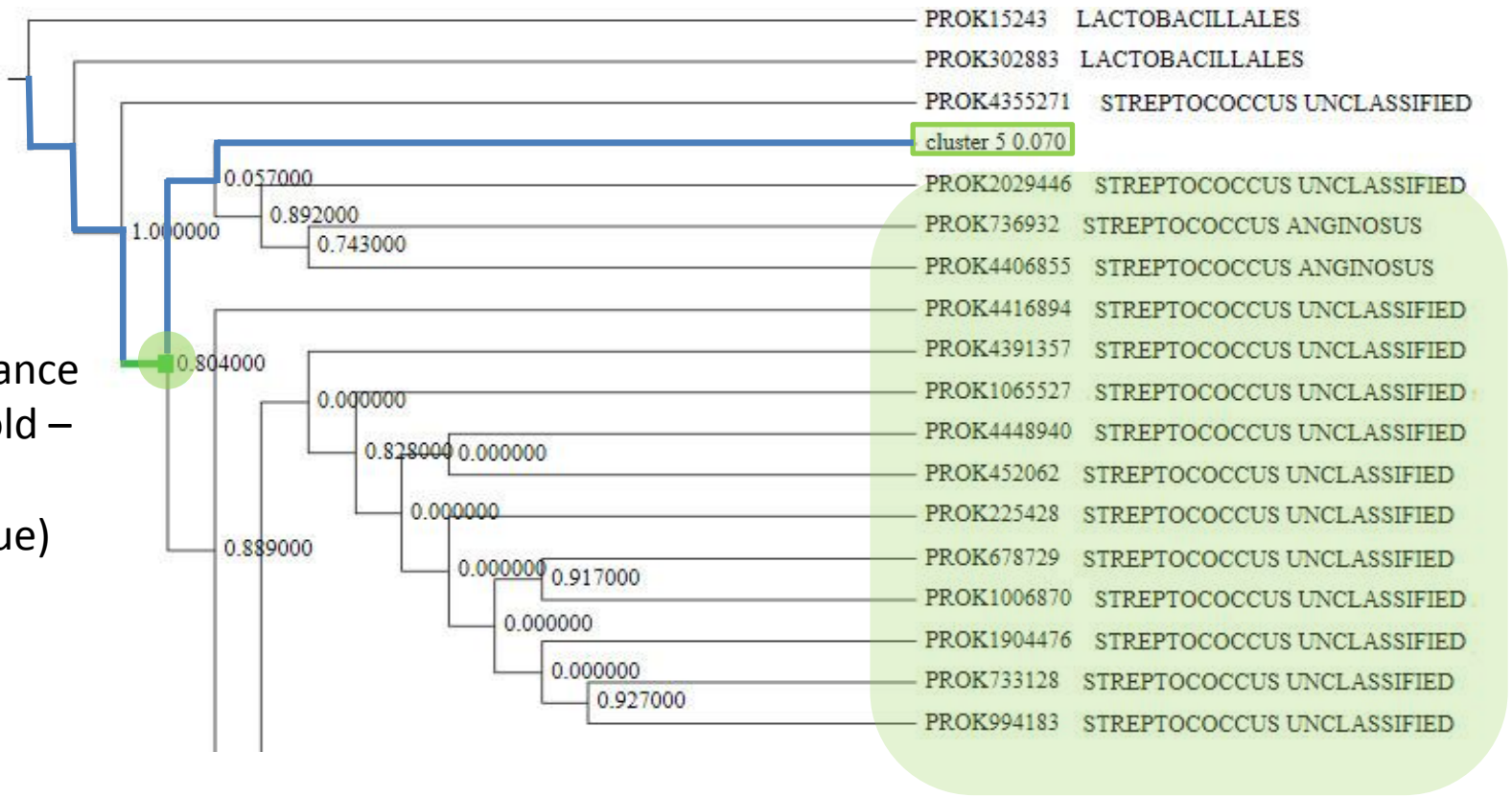
http://greengenes.lbl.gov/cgi-bin/nph-index.cgi

STAP assigns taxonomy to BCV-generated sequences by building up ML phylogenetic trees from the most similar sequences found in GreenGenes. Original version of STAP considers only one nearest neighbour in the tree to assign taxonomy; to make the assignment more credible we use a slightly modified algorithm.

Wu D, Hartman A, Ward N, Eisen JA (2008) An Automated Phylogenetic Tree-Based Small Subunit rRNA Taxonomy and Alignment Pipeline (STAP). PLoS ONE 3(7): e2566. doi:10.1371/journal.pone.0002566

Taxonomy: modified STAP

We look for the smallest sub-tree containing the sequence of interest (in the green rectangle) that is separated by a statistically significant branch. Taxonomy of the sequence is defined by the taxonomic affiliations of all the sequences in that sub-tree.



Significance threshold – 80% (SH-value)

Maximum distance between taxons < 3% => *S. anginosus* or *S.* unclassified
Max dist between taxons ≥ 3% => *Streptococcus* (taxonomy rank is lifted).
Max dist could be specified in the corresponding field.

Simple output

For each chromatogram only the most deeply classified sequences from each taxon are shown in the simple output.

Streptococcus unclassified
Streptococcus infantis
Streptococcus infantis
Streptococcus

→ *Streptococcus infantis*
*S.*unclassified

Streptococcus infantis or *S.* unclassified
Streptococcus infantis

→ *Streptococcus infantis*

File name	BCV expect	STAP classification	BLAST best hit	BLAST %id	BLAST %cov	PROKMSA names for relatives
A_43_3_534R	0.92	*Streptococcus* unclassified	*S.* unclassified	99,35	90,41	*Streptococcus intermedius, Streptococcus constellatus*
A53-1_534R	0.34	*S.* unclassified	*S.* unclassified	98,66	89,72	*Streptococcus sp.*
	0.07	*S. infantis*	*S.* unclassified	98,3	84,57	*Streptococcus sp.*
A53-2_534R	0.70	*S. infantis* or *S.* unclassified	*S.* unclassified	97,77	89,07	*Streptococcus sp., Streptococcus mitis*
A53-3_534R	0.27	*S. infantis* or *S.* unclassified	*S. infantis*	98,67	90,73	*Streptococcus sp., Streptococcus mitis*
	0.08	*Bacillaceae*	*Bacillus* unclassified	95,79	76,54	*Oceanobacillus sojae, Oceanobacillus picturae, Bacillus sp., Virgibacillus sp.*

Bacillaceae, expectation – 0.08. Probable contamination?

Detailed output

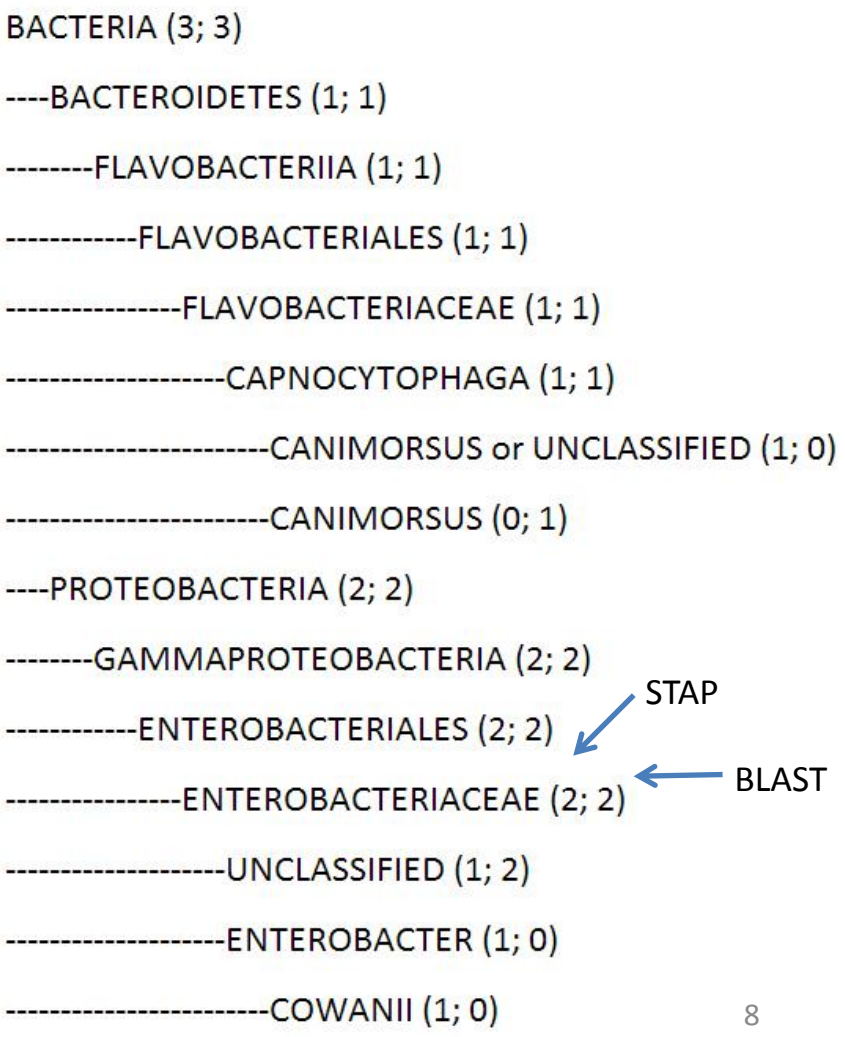
Detailed output for the chromatogram expands on click on the corresponding line in the table.

Chromatogram 2_Un162

Length: 920 nt

Total expectation (of all the sequences with expectation > 0.05): 0.532

Taxonomic tree showing all the taxons recovered from this chromatogram. The number of sequences assigned to this taxon (by STAP and by best BLASTN hit) is specified in brackets.



Detailed output

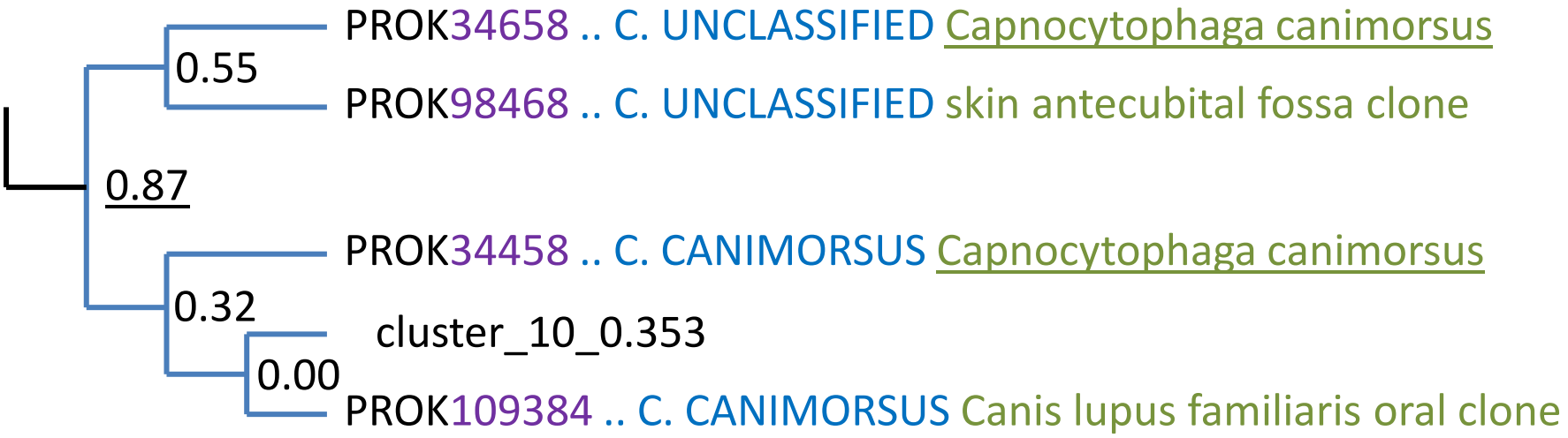
Sequence 1 (cluster_20_0.178)

Expectation	0.178
Length	776 nt
Blast best hit taxonomy	BACTERIA|PROTEOBACTERIA|GAMMAPROTEOBACTERIA|ENTEROBACTERIALES|ENTEROBACTERIACEAE|UNCLASSIFIED
Blast best hit information	Identity = 94.65 Coverage = 94.85 E-value = 0.0
Blast best hit PROKid	300595
Blast best hit PROKname	African elephant feces clone AFEL_aai30b09
STAP taxonomy	**BACTERIA|PROTEOBACTERIA|GAMMAPROTEOBACTERIA|ENTEROBACTERIALES|ENTEROBACTERIACEAE|UNCLASSIFIED**
PROKnames for relatives	Escherichia coli (3)
STAP confidence	0.996000

PROKid – GreenGenes id of the most similar sequence found by BLASTN.
PROKname – description provided by the authors who uploaded the sequence into NCBI
PROKnames for relatives – species present in the sub-tree (with the number of sequences belonging to these species). These names are cut from PROKnames.
STAP confidence – significance of the branch separating the sub-tree.

Detailed output

A fragment of the phylogenetic tree constructed by STAP (xxx.mltree2.svg)



BLAST best hit taxonomy and STAP taxonomy are based on GreenGenes taxonomy.

PROKid – GreenGenes sequence id.

PROKnames for relatives – species present in the sub-tree (if you choose full taxonomic database, there is a chance that there will be no meaningful PROKnames in the subtree; in this case the corresponding field will be empty).

STAP confidence – branch significance.

Interpretation guidelines

- The most reliable prediction is the list of taxons in **STAP taxonomy** field**.**

- **Expectation** value roughly estimates the proportion of different microorganisms in the sample.

- **PROKMSA names** can enhance taxonomic resolution within the confines of STAP prediction. However, you should keep in mind that PROKnames are error prone.

- If STAP taxonomy of some sequence ends with *unclassified,* it means that GreenGenes couldn't explicitly associate its close relatives with any taxon of lower rank. In this case **PROKMSA names** can be especially useful (e.g., *Enterobacteriaceae* unclassified may turn out to be *Escherichia/Shigella,*)f you look at this field).

- You can use best BLASTN hit to check the reliability of STAP prediction: if there is a contradiction between these two predictions, the sequence in question should be omitted.