BCVapp: расшифровка смешанных хроматограмм для идентификации патогенных микроорганизмов

Попова А. В.
НГ Биоинформатики
ФБУН ЦНИИ Эпидемиологии Роспотребнадзора
2014 г

BCV – BaseCaller with Vocabulary

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

Имея данные о примерных возможных последовательностях ДНК (словарь), из хроматограммы прямого секвенирования можно извлечь информацию о реальном составе популяции гомологичных молекул ДНК.

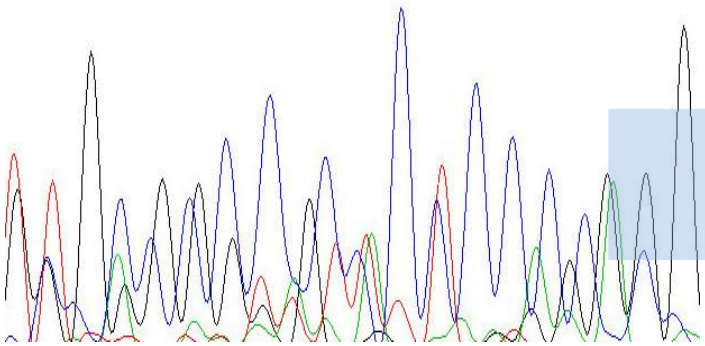
Основное практическое применение BCV –

определение патогенов в стерильных в норме клинических образцах методом прямого секвенирования 16S/18S генов.

BCV-pipeline: BCV

BCV генерирует последовательности, соответствующие смешанной хроматограмме.

1. На вход принимается **смешанная хроматограмма** в формате **ab1**, полученная при секвенировании по Сэнгеру сложного образца.



2. BCV формирует **список последовательностей** в формате **fasta**, которые могли бы породить такую хроматограмму.

```
>cluster_1 0.176055
CTTTCTGGTAGTTACCGTCCTGTGTGAACAATCACTCTCACACACGTTCTTC

>cluster_2 0.364073
CTTTCTGGTAGTTACCGTCCT-GTGTGAACACTCACTCTCACACACGTTCTT

>cluster_3 0.432982
CTTTCTGGTAGTTACCGTCATGTGTGAACAATCCCTCTCACACACGTTCTTC

>cluster_4 0.0268912
CTTTCTGGTAGTTACCGTCCTGTGTGAACACTACCTCTCACACACGTTCTTC
```

0.0268912 - ожидаемая доля в смеси. Значимыми считаются последовательности, доля которых больше 0.05.

BCV-pipeline: таксономическая идентификация

Таксономическая принадлежность последовательностей определяется при помощи программы STAP.

3. Для каждой последовательности, прошедшей порог по ожидаемой доле, определяется **таксономическая принадлежность** (подробная выдача).

Corynebacterium
unclassified

Staphylococcus haemolyticus

Staphylococcus haemolyticus

4. На выходе для хроматограммы формируется **список** породивших ее **таксонов** (компактная выдача).

Corynebacterium
unclassified

Staphylococcus haemolyticus

Таксономия: GreenGenes & STAP

БД **GreenGenes** - специализированная база полных последовательностей генов 16S рРНК, очищенная от химер. GreenGenes предоставляет курируемую таксономию, основанную на филогенетическом анализе de novo.

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.

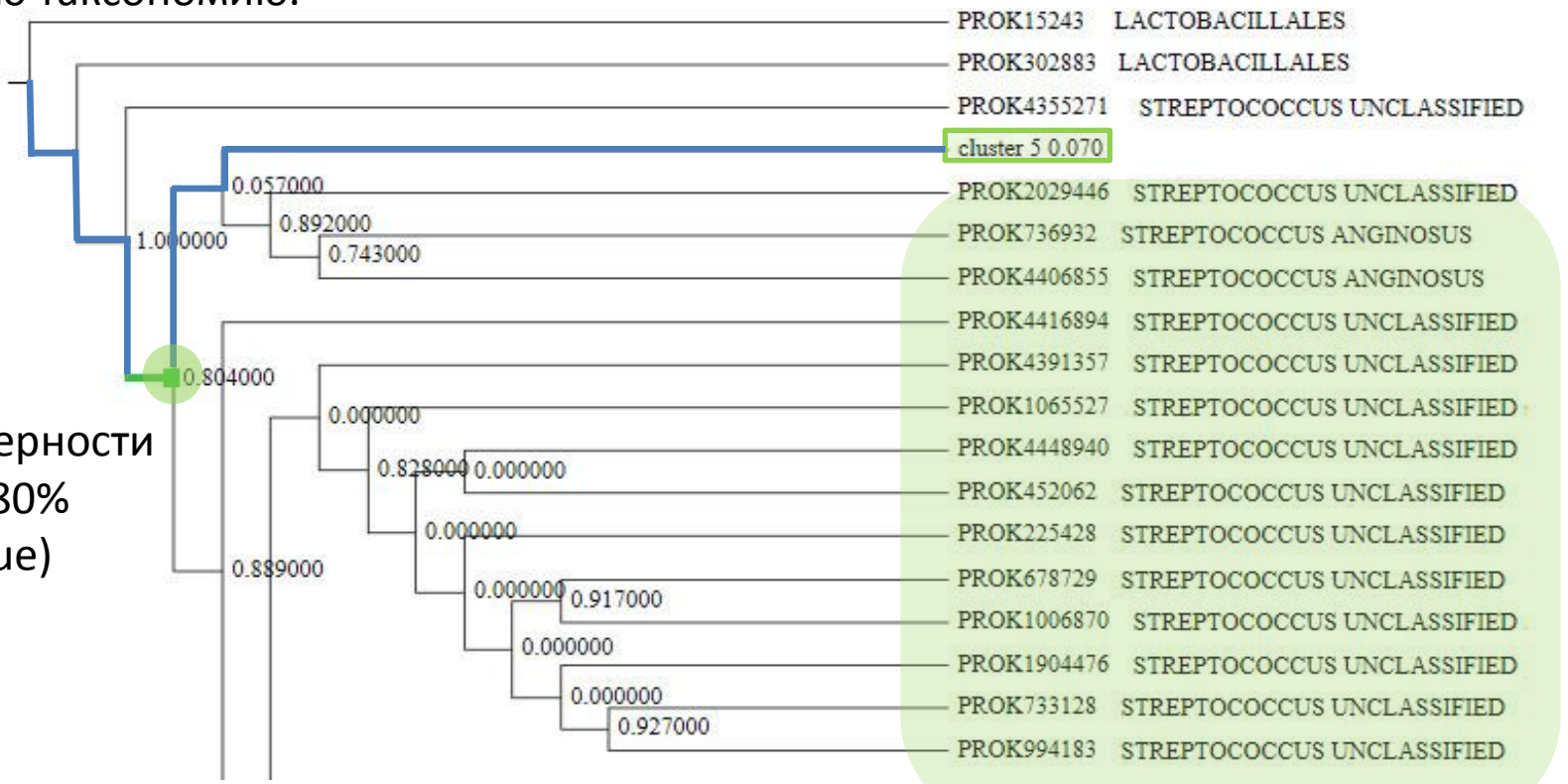
http://greengenes.lbl.gov/cgi-bin/nph-index.cgi

STAP классифицирует полученные BCV последовательности, создавая филогенетическое дерево из наиболее близких последовательностей, присутствующих в GreenGenes (2 раунда отбора последовательностей и построения дерева).

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

Таксономическая идентификация на основе STAP

Содержание наименьшего достоверного поддерева (выделено зеленым), в которое входит интересующая нас последовательность (зеленый прямоугольник), определяет ее итоговую таксономию.



Порог достоверности узла – 80% (SH-value)

Наибольшее расстояние между таксонами < 3% => *S. anginosus* or *S.* unclassified

Наибольшее расстояние между таксонами ≥ 3% => *Streptococcus* (поднимаем ранг до рода). Можно задать другое пороговое расстояние в поле «Maximum distance between close taxons»

Выдача BCV-pipeline (компактная)

Последовательности, относящиеся к одной хроматограмме, группируются по таксонам, и в компактной выдаче каждая группа представлена наиболее глубоко классифицированными последовательностями.

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, ожидаемая доля в смеси – 0.08. Вероятная контаминация?

Выдача BCV-pipeline (развернутая)

Для просмотра полной выдачи - кликнуть на строке нужной хроматограммы. Для каждой из хроматограмм представлена следующая информация:

Chromatogram 2_Un162	Название хроматограммы
Length: 920 nt	Длина хроматограммы
Total expectation: 0.532	Суммарная ожидаемая доля всех значимых последовательностей (т.е. с ожидаемой долей не меньше 5%)

Выдача BCV-pipeline (развернутая)

BACTERIA (3; 3)

----BACTERIOIDETES (1; 1)

--------FLAVOBACTERIIA (1; 1)

----------FLAVOBACTERIALES (1; 1)

----------FLAVOBACTERIACEAE (1; 1)

----------CAPNOCYTOPHAGA (1; 1)

----------CANIMORSUS or UNCLASSIFIED (1; 0)

----------CANIMORSUS (0; 1)

----PROTEOBACTERIA (2; 2)

--------GAMMAPROTEOBACTERIA (2; 2)

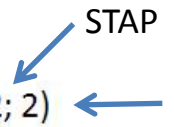
----------ENTEROBACTERIALES (2; 2)

----------ENTEROBACTERIACEAE (2; 2)

----------UNCLASSIFIED (1; 2)

----------ENTEROBACTER (1; 0)

----------COWANII (1; 0)



Таксономическое дерево с указанием числа последовательностей, отнесенных к данному таксону

(согласно классификации STAP и по наилучшему хиту, найденному BLAST)

Выдача BCV-pipeline (развернутая)

Данные для каждой из последовательностей, прошедшей порог

Expectation	0.178
Length	776 nt
Blast best hit taxonomy	BACTERIA|PROT
Blast best hit information	Identity = 94.65
Blast best hit PROKid	300595
Blast best hit PROKname	African elephant
STAP taxonomy	BACTERIA|PROT
PROKnames for relatives	Escherichia coli
STAP confidence	0.996000

Ожидаемая доля в смеси

Длина последовательности

Таксономия лучшего хита BLASTN

Сходство, покрытие, E-value

Идентификатор хита в БД GreenGenes

Описание организма, добавленное авторами, загрузившими последовательность хита в NCBI

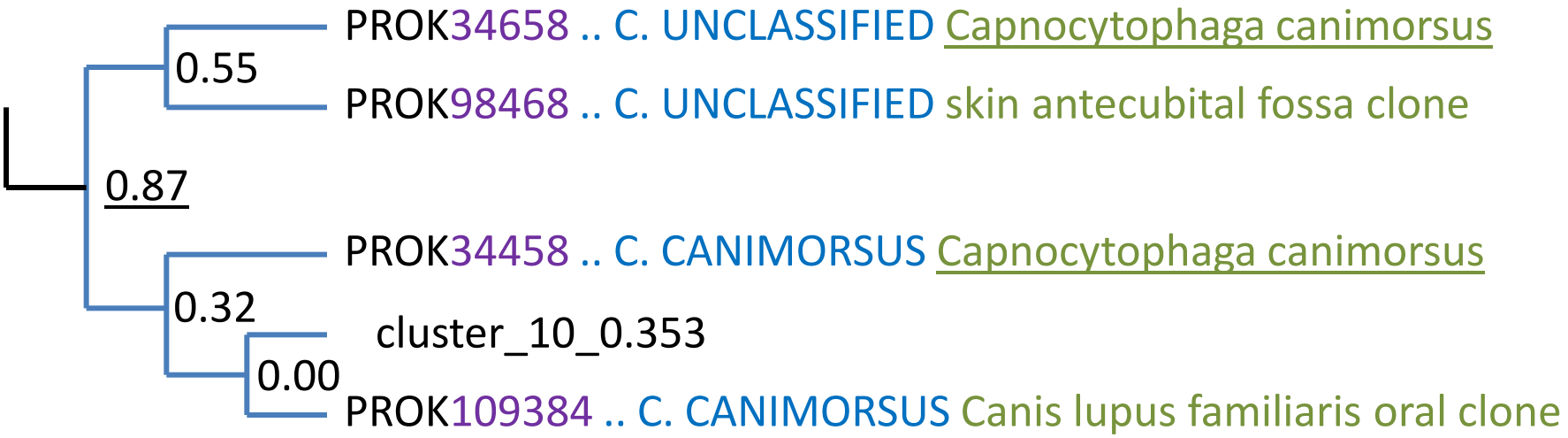
Таксономия, определенная STAP

Названия всех организмов из поддерева, выбранного на этапе классификации (с указанием числа каждого из видов)

Достоверность ветки, отделяющей это поддерево.

Выдача BCV-pipeline (развернутая)

Фрагмент филогенетического дерева, построенного STAP (файл xxx.mltree2.svg)



PROKid – идентификатор последовательности в БД GreenGenes.
PROKnames for relatives – названия всех организмов из поддерева, выбранного на этапе классификации(если выбрана полная таксономическая база (“Taxonomic database” – “full”), в этом поле может не содержаться дополнительной информации о видовой принадлежности).
STAP confidence – достоверность ветки, отделяющей это поддерево.

Интерпретация результатов

- Основной результат - список таксонов в поле **STAP taxonomy.**

- Приблизительно оценить относительное содержание микроорганизмов в образце можно по ожидаемой доле последовательности в смеси (поле **Expectation**). Пример см. на слайде 13, проба №142: ожидаемая доля предсказанного RipSeq *Streptococcus* в три раза меньше, чем доля *Finegoldia magna;* можно предположить, что мажорный микроорганизм и является основным возбудителем заболевания.

- **PROKMSA names** указываются для конкретизации классификации в рамках предсказания STAP. Однако стоит помнить, что достоверность указанных в PROKMSA names данных заранее неизвестна. Имеет смысл обращать внимание на количество представителей того или иного таксона, присутствующих в поддереве (числа в скобках): минорный таксон может оказаться ошибкой аннотации.

- Таксономия, определяемая STAP, может завершаться категорией unclassified. В эту категорию включены те последовательности, которые алгоритмы GreenGenes не смогли однозначно отнести к таксону более низкого ранга. Для уточнения классификации таких последовательностей необходимо посмотреть на PROKMSA names (например, *Enterobacteriaceae* unclassified может определяться по PROKMSA names как *Eshcerichia/Shigella*).

- Лучший хит BLAST предназначен для проверки корректности классификации. Если две классификации существенно противоречат друг другу, то это хороший повод исключить последовательность из рассмотрения. Если показатели identity и coverage у лучшего хита достаточно хороши (обычно в качестве порогового значения для идентификации на уровне вида принимается identity = 97%, coverage = 95%), можно использовать его классификацию для уточнения. Однако стоит принять во внимание, что лучший хит не всегда является ближайшим родственником; прояснить ситуацию можно, посмотрев на ближайшего родственника последовательности в филогенетическом дереве, построенном STAP.