Сравнение эффективности биоинформационных алгоритмов в анализе геномов штаммов возбудителя чумы

Носов Н.Ю., Оглодин Е.Г., Куклева Л.М., Матвеева Ж.В.,
Ерошенко Г.А., Краснов Я.М., Кутырев В.В.

1. В международных генетических базах данных представлено более 300 полногеномных последовательностей штаммов возбудителя чумы различного происхождения, депонированных в том числе специалистами РосНИПЧИ «Микроб».

2. Для анализа этих последовательностей используются различные биоинформационные программы, что приводит к различиям в интерпретации данных полногеномного секвенирования.

3. Стоит задача разработки оптимального биоинформационного алгоритма применения коммерческих программ и программ свободного доступа, который обеспечивает проведение молекулярного типирования, дифференциации и филогенетического анализа штаммов *Y.pestis*.

Для анализа эффективности различных биоинформационных алгоритмов нами использованы полногеномные последовательности 70 штаммов *Yersinia pestis* из природных очагов чумы России и сопредельных государств, а также 170 штаммов из баз данных NCBI GenBank, PATRIC и EMBL и штамма IP32953 *Yersinia pseudotuberculosis.*

При сравнении полногеномнных последовательностей штаммов микроорганизмов используется метод анализа единичных нуклеотидных замен в коровом геноме. Для поиска единичных нуклеотидных замен используются такие программы как Mummer; Kodon; Wombac и т.д. Из этих программ для проведения полногеномного SNP анализа нами была выбрана программа Wombac 2.0, т.к. она сочетает простоту использования с высокой эффективностью. При помощи этой программы на платформе BioLinux нами был проведен поиск единичных нуклеотидных замен (SNP) корового генома

использованных штаммов. В результате был получен файл, содержащий SNP профили каждого включенного в анализ штамма.

По ранее использованной нами в ряде публикаций схеме мы загружали полученный файл в программу Bionumerics 7.5, где проводили филогенетический анализ, используя метод наибольшей экономии (Maximum Parsimony). Данный подход обладает рядом неоспоримых преимуществ. Все манипуляции, начиная с создания базы данных SNP профилей штаммов, заканчивая филогенетическим анализом, проводятся в одной программе, с весьма удобным графическим интерфейсом. А сам метод построения не чувствителен к аппаратной составляющей. В его основе лежит принцип, что наиболее вероятный эволюционный сценарий содержит наименьшее количество событий (в нашем случае нуклеотидных замен).

Построенная таким образом дендрограмма довольно четко отображает все основные филогенетические линии возбудителя чумы в порядке их отхождения от общего ствола эволюции. Но в тоже время данный подход не способен четко дифференцировать популяционную структуру внутри крупных ветвей, например, такой как ветвь средневекового биовара 2.MED.

Штаммы этой ветви являются эволюционно молодыми и генетически мономорфными, обладая небольшим количество различий в коровых единичных нуклеотидных заменах, и метод наибольшей экономии ввиду простоты своей математической модели не способен качественно их кластеризовать, в ряде случаев выделяя штамм в отдельную субветвь, а так же не позволяя кластеризовать штаммы, выделенные из одного природного очага.

Ввиду актуальности исследования штаммов средневекового биовара *Y.pestis*, как наиболее распространенных на территории природных очагов Российской Федерации и сопредельных государств, нами было проведено существенное усовершенствование действующего алгоритма, которое позволило бы проводить точный филогенетический анализ как крупных внутривидовых категорий (таких как подвиды и биовары), так и более тонкий

анализ, касающийся популяционной структуры отдельно взятых филогенетических линий, какой является 2MED1, к которой относятся все штаммы средневекового биовара выделенные на территории России и сопредельных государств.

На основе полученных нами ранее знаний о филогенетическом разнообразии возбудителя чумы и мирового опыта биоинформационных исследований нами были внесены следующие изменения в действующий биоинформационный алгоритм.

Из полученных SNP профилей штаммов были удалены единичные нуклеотидные замены в количестве 28, находящиеся в областях гомоплазии генома возбудителя чумы, т.к они не несут никакого эволюционного смысла. Далее редактированный файл из формата FASTA путем использования программ MEGA 6.0 и PAUP 4.0 конвертировали в формат PHYLIP.

Для построения филогенетического дерева использовали программу PhyML (Институт молекулярной генетики Монтпелье, Франция), основывающуюся на принципе максимального правдоподобия (Maximum Likelihood). Нами была выбрана 5 параметрическая модель замен HKY 85 (Hasegawa; Япония) и бустрэп подкрепление в размере 300 повторностей. Программа преобразовывает нуклеотидные последовательности в цифровую матрицу подобий и поэтому работает значительно быстрее, чем другие программы с аналогичным методом построения (в том чилсе Biomunerics).

На последнем этапе проводили визуализацию филогенетического дерева в программе Dendroscope.

В результате нами впервые в России построено наиболее полное и точное филогенетическое дерево *Y.pestis*, отражающее филогенетические связи между всеми входящими в него внутривидовыми группами (Подвиды, биовары, филогенетические линии).

На полученном филогенетическом дереве в формате радиальной кладограммы четко кластеризуются штаммы возбудителя чумы в порядке их отхождения от штамма *Y.pseudotuberculosis*. Наиболее древними являются

представители линии O.PE7 включающей в себя китайские штаммы, выделенные на территории Тибета и штаммы кавказского подвида линии O.PE2. Далее отходят штаммы гиссарского, алтайского подвида, штаммы microtus и таласской группы объединяющиеся в центрально-азиатский подвид и ветвь O.PE5 улегейского подвида.

Наибольший интерес в новом филогенетическом дереве представляет филогенетическая линия 2MED, состоящая из штаммов *Y.pestis* средневекового биовара, т.к штаммы именно этого биовара циркулируют в большей части очагов России и сопредельных государств.

Ветвь 2.MED имеет сложную популяционную структуру. В основании ветви лежит штамм из центрально кавказского высокогорного очага C-627 относящийся к линии 2.MED0. Далее последовательно отходят кластеры линий 2.MED2 и 2.MED3, циркулирующие исключительно в очагах Китая.
Все штаммы средневекового биовара из природных очагов России относятся к линии 2MED`1.

Штаммы линии 2.MED1 делятся на 2 основные ветви — кавказско-каспийскую и среднеазиатско-китайскую, названные в соответствии с территориями очагов на которых входящие в них штаммы были выделены. В свою очередь эти ветви внутри также делятся на ряд следующих крупных кластеров.

Среднеазиатская-Китайская ветвь делится на 2 кластера: кластер штаммов выделенных в провинции Синьцзян (Китай) и кластер штаммов, выделенных на территории Средней Азии (Казахстан, Узбекистан, Киргизия).

Кавказско-каспийская ветвь включает в себя 4 крупных кластера.

Наиболее крупные кластер образован штаммами, выделенными на территории очагов зоны Прикаспия. Точность разработанного алгоритма подтверждает факт объединения штаммов из Прикаспийского песчаного очага в единую группу. Прежде они лежали разрозненно. Штаммы из очагов Кавказа и Ирана также образовали отдельную ветвь.

Отдельный интерес представляют штаммы, выделенных из Волго-Уральского степного и песчаного очагов (относящиеся к трансграничным) от больных людей в начале 20ого века. Они сформировали два отдельных кластера, не кластеризуясь со штаммами, выделенных от природных носителей возбудителя чумы.

Таким образом, разработанный нами биоинформационный алгоритм анализа полногеномных последовательностей штаммов возбудителя чумы позволяет проводить эффективный филогенетический анализ, четко дифференцируя при этом крупные внутривидовые категории, и позволяя проводить анализ популяционной структуры отдельных ветвей.

На данном слайде мы провели сравнение основных этапов нашего алгоритма с подходом китайских исследователей (Cui Yu; Yu C), так же активно занимающихся данным направлением. Основные отличия заключаются в методе поиска коровых единичных нуклеотидных замен и методе проведения анализа.

Наши результаты совпадают с таковыми у китайских коллег. Но в силу отсутствия у них штаммов ряда филогенетических групп, а также разнообразия по линии 2.MED1, наше итоговое филогенетическое древо является более полным, превосходя выборку китайских исследователей практически в два раза.

Выводы

1. Разработан простой и высокоэффективный биоинформационный алгоритм анализа полногеномных последовательностей штаммов возбудителя чумы доступный для использования в учреждениях Роспотребнадзора

2. С его помощью проведен филогенетический анализ всех представленных в международных базах данных геномов штаммв Y.pestis, а также геномов штаммов секвенированы в РосНИПЧИ Микроб

3. Определена современная популяционная структура вида Y.pestis

4. Впервые выполнен развернутый филогенетический анализ штаммов средневекового биовара из природных очагов России и сопредельных стран.