Московский физико-технический институт (государственный университет)
Факультет биологической и медицинской физики
Кафедра кафедра молекулярной и трансляционной медицины

Диссертация допущена к защите
зав. кафедрой

_____ Лазарев В.Н.

«____» _____ 2017 г.

Выпускная квалификационная работа на соискание степени МАГИСТРА

Тема: **Протеогеномный анализ штамма *Mycobacterium tuberculosis W-148***

Направление: 010900 – Прикладные математика и физика
Магистерская программа: 010982 – Физико-химическая биология и биотехнологи

Выполнил студент гр. 1114 _____ Смоляков А.В.

Научный руководитель,
к. б. н. _____ Шитиков Е.А.

Работа выполнена в ФГБУ ФНКЦ ФХМ ФМБА России
Москва – 2017

Оглавление

1. Список сокращений

GSSP — Genome Search Specific Peptides

PSM — Peptide Spectrum Match

2. Введение

Mycobacterium tuberculosis является возбудителем тяжелой болезни - туберкулеза, наносящей большой вред здоровью. Особенно остро проблема стоит в развивающихся странах. Появление резистентных и более вирулентных штаммов усугубило ситуацию. Было сделано множество работ по исследованию протеома этого патогена [1, 2]. Есть несколько работ, посвященных аннотации генома *M.tubeculosis*. Полногеномное секвинирование штамма *Mycobacterium tuberculosis H37Rv* было проведено в 1998 году, затем был отсеквенирован штамм *CDC1551* и несколько других [3, 4]. Точная аннотация белок-кодирующих генов является постоянно меняющимся и модернизирующимся техническим процессом. Это особено заметно в случае *M.tubeculosis*. В работе Коула и его колег, было показано наличие 3924 ORF в геноме *H37Rv* [3]. При повторной аннотации, проведенной теми же авторами, число генов стало 3995 [5]. В марте 2011 база TubercuList содеражала 4012 аннотированых белок-кодирующих генов. де Соуза с коллегами провели сравнение двух различных аннотаций для *H37Rv*, сделанных группами Sanger, TIGR, и обнаружили, что 50% генов имеют различные стартовые сайты [6]. В этой же работе, используя протеомные данные 449 культур, авторы смогли исправить аннотацию 24 генов. Так же, возможность существования CDSs, которые ещё не были аннотированы в *H37Rv*, были предложены в работе Лью и коллег [7].

3. Обзор литературы

3.1. Mycobacterium tuberculosis

Mycobacterium tuberculosis – возбудитель туберкулеза (ТБ), заболевания известного еще с древности и являющегося одной из главных причин смертности во всем мире. По данным Всемирной организации здравоохранения за 2014 год выявлено 9.6 миллионов новых и 1.5 миллиона смертельных случаев ТБ (WHO, 2015). При этом Россия является 1 из 22 стран с наивысшем бременем туберкулеза, на долю которых приходиться 80% всех случаев ТБ.

Mycobacterium tuberculosis являются грамположительными палочками, длиной 1-10 мкм и диаметром 0.2-0.6 мкм. Морфологически выделяют, как прямые, так и слегка изогнутые формы.

Особенности генома

Последовательность генома M. tuberculosis штамма H37Rv была полностью расшифрована в Сенгеровском институте в 1998 году (Cole et al., 1998) (Рисунок 1). Это был третий опубликованный бактериальный геном после Haemophilus influenzae (Fleischmann et al., 1995) и Mycoplasma genitalium (Fraser et al., 1995).

Следует отметить, что ключевые особенности организации генома одинаковы для всех штаммов патогена. Геномы представлены кольцевой молекулой ДНК протяженностью около 4400 тысяч пар оснований (т.п.о.) и характеризуются высоким содержанием GC пар ($\tilde{6}5.5\%$). При этом существует несколько регионов, отличающихся по GC составу.

Углубленный анализ генома *M. tuberculosis* штамма *H37Rv* выявил около 3906 генов (аннотация RefSeq, NCBI Reference Sequence: NC_000962.3),

кодирующих белки. При этом следует отметить, что альтернативный старт трансляции GTG встретился в 35% случаев, что существенно чаще, чем 14% и 9% в геномах *Bacillus subtilis* и *Escherichia coli*, соответственно. Был найден один набор рибосомных генов и 45 транспортных РНК.

Типирование на основе однонуклеотидных полиморфизмов в последние годы все чаще используется для генотипирования микобактерий туберкулезного комплекса. Данный маркер является практически идеальным для классификации штаммов и отнесения их к тем или иным филогенетическим линиям. При этом полиморфизмы могут быть идентифицированы как путем сравнения in silico опубликованных полных геномов штаммов M. tuberculosis (Fleischmann et al., 2002; Gutacker et al., 2002; Alland et al., 2003; Garnier et al., 2003; Baker et al., 2004), так и de novo анализом (Dos Vultos et al., 2008; Hershberg et al., 2008), когда первоначально выбираются гены, а потом в них проводится поиск SNP. Первый вариант является менее предпочтительным, так как использование полиморфизмов, подобранных на основе сравнения полных геномов только нескольких штаммов, причем, в основном, одной линии, содержит риск получения непредставительной коллекции маркеров, и как результат коллапсирование филогенетических ветвей (Alland et al., 2003; Pearson et al., 2004; Achtman, 2008; Smith et al., 2009). Использование SNP маркеров, идентифицированных de novo, представляется наиболее удачным вариантом для определения филогенетических взаимоотношений между штаммами, а в некоторых случаях и для молекулярной эпидемиологии. Здесь следует отметить, что в последние годы для построения углубленных филогений все чаще применяется полногеномное секвенирование с последующим анализом однонуклеотидных полиморфизмов. Публикации последних лет также подчеркивают, что секвенирование геномов становится своего рода новым «золотым стандартом» для исследо-

вания молекулярной эпидемиологии. Так, например, показано, что оно обладает значительно большей дискриминирующей способностью, чем любые из стандартных методов генотипирования, и что образцы, имеющие одинаковые генетические профили, могут иметь существенные отличия на геномном уровне, что в дальнейшем оказывается важным для интерпретации паттернов трансмиссии (Niemann et al., 2009; Gardy et al., 2011; Casali et al., 2012) или для изучения смешанных инфекций (Comas et al., 2011; Saunders et al., 2011). В случае филогенетического анализа сравнительная геномика позволяет установить степень неоднородности популяции и определить генетические дистанции. Основываясь на недавно опубликованных исследованиях геномных последовательностей было выделено шесть основных филогенетических линий для микобактерий туберкулезного комплекса, ассоциированных с туберкулезом человека (Comas et al., 2010; Bentley et al., 2012) (Рисунок 4). Четыре линии относятся к M. tuberculosis, а две к M. africanum. Как упоминалось ранее, анализ геномных данных позволяет вычислить эволюционные дистанции. Было установлено, что на уровне геномов отличия между линиями составляют в среднем 2000 SNPs, что эквивалентно, к примеру, эволюционной дистанции между M. tuberculosis и M. bovis (Garnier et al., 2003), а разнообразие мировой популяции МБТ в целом выше, чем разнообразие всех остальных представителей комплекса при сравнении друг с другом (без учета M. canettii и других «гладких» микобактерий).

Генетическое семейство Beijing

Впервые представители генотипа Beijing были обнаружены в 90х годах XX века, в двух независимых исследованиях, проведенных группами исследователей из Голландии и Америки (van Soolingen et al., 1995; Bifani

et al., 1996). В ходе IS6110 RFLP анализа и сполиготипирования коллекции изолятов M. tuberculosis, полученных от больных ТБ в 1992-1994 годах в Китайской Народной Республике и Монголии, van Soolingen с соавт. выявили доминирующий генотип. При этом наиболее часто представители генотипа встречались в окрестностях Пекина (англ. Beijing), отчего и получили свое название (van Soolingen et al., 1995). Параллельно этому исследованию, Bifani с соавт. из Научно-исследовательского института общественного здравоохранения США (Public Health Research Institute, NY, США), в 1996 году методами молекулярно-генетической эпидемиологии описали вспышку лекарственно-устойчивого туберкулеза, произошедшую в Нью-Йорке в начале 1990х. Выявленные штаммы характеризовались крайне схожими паттернами IS6110 профилей и были названы «W» (Bifani et al., 1996). В дальнейшем эти названия были объединены в W/Beijing или просто Beijing (Kurepina et al., 1998; Van Soolingen, 2001). При этом название как нельзя лучше отражает реальное место зарождения генотипа. Его представители наиболее часто встречаются в Восточной Азии и, по мнению Мокроусова с соавт., генотип Beijing возник в Северном Китае более 1,000 лет назад. Дальнейшее его распространение было связано с миграционными потоками: со средневековых времен в Россию, совсем недавно в ЮАР (с XVII века) и в Австралию (в XIX веке) (Mokrousov et al., 2008). В свою очередь, согласно Merker с соавт., генотип в целом возник более 6,000 лет назад в географической зоне, включающей в себя Северо-Восток Китая, Корею и Японию (Merker et al., 2015).

Согласно международной базе данных сполиготипирования SpolDB4, штаммы Beijing присутствуют в наибольшем количестве стран на глобальном уровне (13% от мирового количества изолятов), являясь по этому показателю уникальным генотипом (Brudey et al., 2006). Здесь также

следует отметить ассоциацию представителей генотипа с многочисленными вспышками заболеваний во всем мире, многие из которых были лекарственно устойчивые (Frieden et al., 1996; Agerton et al., 1999; Caminero et al., 2001; Affolabi et al., 2009). В структуре популяции возбудителя туберкулеза в России доля штаммов Beijing составляет от 50% до 80% (Mokrousov et al., 2003), причем, крайне выражена ассоциация штаммов с лекарственной устойчивостью (Casali et al., 2014). Исходя из сказанного выше, предполагается, что у штаммов данной эволюционной линии, возможно, развились уникальные свойства, которые позволили им распространиться по всему миру (клональная экспансия). По мнению многих авторов, этими свойствами являются: 1) способность «ускользать» от БЦЖ-вакцинирования 2) способность штаммов относительно быстро приобретать устойчивость к противотуберкулезным препаратам.

Характеристика кластера B0/W148

При описании генетического семейства Beijing в главе выше довольно часто речь заходила об «успешных» представителях генотипа. Одним из таких «успешных» кластеров является Beijing B0/W148. Впервые штаммы кластера были выявлены на рубеже XX и XXI веков. В независимых исследованиях Нарвской, Курепиной и Portaels с использованием IS6110 RFLP типирования обнаружили группы кластеризующихся образцов, названные B0 (отечественная систематика) и W148 (иностранная систематика). Отличительной особенностью этих образцов было наличие двойной полосы (7.1 и 9.2 Kb) в верхней части профиля. В 2008 году штаммы Beijing B0/W148 с sensu stricto профилем и характерной двойной полосой были отнесены к Beijing B0/W148 (Mokrousov et al., 2008), а метод IS6110-RFLP типирования был признан «золотым стандартом»

для выявления изолятов данной клональной группы. Дополнительными названиями кластера могут считаться CladeB (Casali et al., 2014) и ECDC0002 (de Beer et al., 2014).

На сегодняшний день опубликовано достаточно много данных об ассоциации штаммов кластера с лекарственной устойчивостью, что является достаточным, для оценки степени опасности, исходящей от циркуляции изолятов B0/W148. Первая статья, описывающая российские штаммы M. tuberculosis, выделенные в середине 1990-х, уже показала широкую распространенность МЛУ среди изолятов данного кластера (Marttila et al., 1998). Недавно изоляты B0/W148 с множественной лекарственной устойчивостью были выявлены в эпидемиологически значимой выборке пациентов с впервые выявленным ТБ в Ленинградской (Narvskaya O. 2003. Genome polymorphism of Mycobacterium tuberculosis and its role in epidemic process. D.Sc. dissertation. Institute of Experimental Medicine, St. Petersburg, Russia. (In Russian.)), Тульской (Dubiley et al., 2010), Самарской и других областях. Следует отметить, что в исследовании более 1,000 штаммов из Самары 119 штаммов относилось к кластеру B0/W148 и все они были лекарственно устойчивы. В Абхазии, 22 из 23 изолятов B0/W148 были МЛУ, в то время как другие включенные в исследование изоляты генотипа Beijing были чувствительны к противотуберкулезным препаратам (10 из 55) (Pardini et al., 2009). В Эстонии изоляты кластера B0/W148 составили 37.2% от всех лекарственно устойчивых штаммов генотипа Beijing, в то время как ни одного чувствительного штамма B0/W148 в данном исследовании выявлено не было, включая штаммы, выделенные в 1994 году (Kruuner et al., 2001). Также следует отметить, что в исследовании 2,092 образцов из 24 стран Европы методом VNTR кластер B0/W148 был выявлен в 470 случаях (17 стран, преимущественно Восточная Европа). Согласно исследованию

этот кластер называется ECDC0002 и в крайней степени ассоциирован с лекарственной устойчивостью. Довольно интересной является гипотеза о происхождении штаммов и первичном их распространении. По мнению Мокроусова штаммы кластера зародились в Сибири до 1960х годов, что в целом согласуется с исследованием Merker с соавт. (касательно даты возникновения). В дальнейшем, в ходе программы по освоению целины (1955-1960 годы), миграционные потоки были направлены в Казахстан и в Сибирь. Следует отметить, что в Казахстане представленность кластера B0/W148 крайне мала и составляет около 4%. Это подтверждает гипотезу автора о том, что в европейской части России штаммов кластера в те годы еще не было. В свою очередь вторая волна миграции, 1960-1980 годы, напротив, из Сибири по всей стране могла повлечь массовое распространение представителей кластера. Согласно Мокроусову, триггером распространения именно устойчивого клона могло послужить повсеместное использование открытого в 1963 году рифампицина.

Анализ эпидемиологических и литературных данных показал, что в настоящее время туберкулез остается одним из социально значимых заболеваний в мире. Разработка лекарственных препаратов и вакцины позволили лишь на время снизить угрозу распространения туберкулеза по миру и количество смертельных случаев. Сегодня же, наблюдается некоторое ухудшение эпидемиологической обстановки в связи с активным распространением лекарственноустойчивых форм M. tuberculosis, в том числе устойчивых ко всем известным лекарственным препаратам. Таким образом, исследования возбудителя туберкулеза активно проводятся учеными во всем мире.

3.2. Применение масс-спектрометрии в протеомике

Протеомика исследует всю совокупность белков, синтезируемых организмом/клеткой в данной среде и на конкретном этапе клеточного цикла. Она описывает их качественный состав, относительную представленность, взаимодействие с другими макромолекулами, а так же посттрансляционные модификации (ПТМ) (Hakkinen et al., 2000; Molloy and Witzmann, 2002; Monteoliva and Albar, 2004). Белки играют важную роль почти во всех биологических процессах, соответственно в клетках существуют тысячи белков, каждый из которых подвергается взаимодействию, как с другими белками, так и с целыми клеточными компартментами.

DODELAT

3.3. Проведение масс-спектрометрического эксперимента

Стандартный эксперимент по исследованию белков с использованием масс-спектрометра состоит из пяти основных этапов. На первом шаге белки из клеточного лизата или ткани извлекаются и очищаются. Как правило, за этим следует разделение полученной смеси гель-электрофорезом. Следующим этапом является фрагментирование белка на пептиды. Обычно это происходит при помощи фермента трипсина. На третьем этапе пептиды, предварительно разделенные жидкостной хроматографией на одну или несколько фракций, подвергаются ионизации и поступают в масс анализатор. После MS-анализа пептиды могут подвергнуться повторной фрагментации. Новые ионы так же подвергаются анализу. Это пятый этап (MS/MS или тандемная масс-спектрометрия). МС-основанная протеомика зарекомендовала себя как незаменимая технология для определения закодированной в геноме информации. На сегодняш-

ний момент белковый анализ (первичная структура, посттранскрипционные модификации или белок-белковые взаимодействия) при помощи МС является наиболее успешным при работе с небольшим (по сравнению с другими методами белкового анализа) количеством белка, выделенным из различных образцов. Системный анализ большого количества генов, экспрессированных в клетки, является основной целью протеомики; эта область сейчас быстро развивается, в основном, благодаря разработке новых экспериментальных подходов.

3.4. Методы протеогеномики

Для получения протеомных данных обычно используют метод «shotgun proteomics» - комбинация жидкостной хроматографии и тандемной масс-спектрометрии [8]. Одним из ключевых шагов в протеомике является идентификация пептидов на основе полученных MS/MS спектров. В отличие от геномных технологий, вроде ДНК или РНК секвенирования, где происходит непосредственное секвенирование исходной последовательности, в протеомике, как правило, пептиды идентифицируются за счет сопоставления экспериментальных MS/MS спектров и теоретических спектров всех пептидов, представленных в базе, против которой осуществляется поиск [9]. При таком поиске используются следующие исходные предположения: 1. все белок-кодирующие последовательности генома точно известны и аннотированы 2. все эти последовательности включены в базу, против которой осуществляется поиск . Весь последующий анализ, включая идентификацию, количественный анализ и прочий статистический анализ, основывается на этих предположениях [10].

Проблема такого подхода заключается в том, что не все пептиды представлены в текущей поисковой базе или какой-либо другой. Пепти-

ды могут содержать мутации, находиться в новых генах, перед неверно аннотированным стартом или в альтернативных сплайсформах. Один из способов идентификации пептидов с мутациями заключается в масс-тег подходе. При этом подходе происходит идентификация коротких участков пептида, после чего осуществляется поиск в более широком диапазоне масс прекурсеров [11].

Более общим подходом является протеогеномика. Термин впервые был использован в 2004 и изначально использовался в исследовании, где протеомные данные использовались для улучшения качества аннотации [12]. С тех пор этот термин используется в более общем смысле. В протеогеномном подходе, пептиды идентифицируются за счет идентификации MS/MS спектров против специальной базы, включающей в себя последовательности новых, предсказанных белков и различные варианты последовательности белка. Такие базы получаются за счет использования геномной и транскриптомной информации. Таким образом, протеогеномика позволяет не только подтвердить текущую аннотацию, но так же уточнить её [13].

Типы пептидов, идентифицируемых при протеогеномном исследовании

Пептиды, идентифицируемые при протеогеномном поиске, соответствуют различным участкам генома. Такие пептиды можно разделить на межгенные (находятся между аннотированными генами) и внутрегенные (находятся полностью или частично в областях, где содержится аннотированный ген). Внутрегенные можно разделить на 1. находящиеся в белок-кодирующих генах 2. находящиеся в длинных некодирующих РНК 3. находящиеся в псевдогенах [14]. Большинство пептидов, иденти-

фицируемых в протеогеномике, уже известны и относятся к аннотированным генам. В эукариотах (в которых присутствует интроно-экзоная структура) большинство пептидов относятся к экзомам, и, как правило, меньше 20% относятся к экзон-экзон участкам. Новые пептиды, не идентифицируемые против какой-либо базы, могут находится в неаанотированных участках генома, быть результатами одно-аминокислотной замены (SAP), находится в нетранслируемых регионах (3' или 5' UTR) или интронах, является результатам альтернативного сплайсинга [13].

Подходы к созданию поисковых баз

Идентификация пептидов против кастомных баз является ключевым шагом в протеогеномике. Обычно база состоит из известных аннотированных последовательностей и предсказанных последовательностей. При протеогеномном поиске следует внимательно относиться к размеру базы: увеличение размера влечет за собой увлечение времени поиска и FDR. Оптимальный выбор зависит от того, что требуется в эксперименте: точность или чувствительность [13].

Транслирование в шести рамках генома - такая база может получена в результате транслирования в шести рамках генома [15]. Недостатком такого подхода является гигантский размер итоговой базы (в основном состоящей из последовательностей несуществующих белков) и невозможности поиска экзом-экзом пептидов, в случае эукариот. Например транслированный таким образом геном человека приводит к базе в 3.2 гигабазы белковых последовательностей, что в 70 раз больше, чем референс в 45 мегабаз [16]. Для уменьшения размера базы могут применены различные вычислительные методы: выбор последовательностей, имеющих гомологии с уже известными белками; использование методов

предсказания кодирующего потенциала; исключение слишком коротких последовательностей (например, меньших, чем 30 аминокислот) [17].

Ab initio предсказание генов

Expressed sequence tag (EST) data

Аннотированные РНК-транскрипты Белковые последовательности могут быть получены в результате шестирамочного транслирования аннотированных РНК-транскриптов, например Enselbl или RefSeq. Это позволяет идентифицировать альтернативные сайты инициации трансляции. База GENCODE содержит 84408 мРНК аннотированных белков. В результате транслирования такой базы получается белковая база в 200 мегабазы, что всего в 4.5 раза больше референса [16]. Так же такие базы могут содержать последовательность, аннотированные как псевдогены или длинные некодирующие РНК [18].

RNA-seq данные

Различные вариации последовательностей Белковые последовательности в рефернсной базе могут быть дополнены последовательностями, являющими вариациями референсных последовательностей (как правило, это одноаминокислотные полиморфизмы, делеции и инсерции). Для каждой вариации, берется большая область вокруг вариации и добавляется в базу, как независимая последовательность. Одно аминокислотные замены можно скачать из различных баз данных: NCBI dbSNP, Online Mendelian Inheritance, Protein Mutant Database [19].

Прочие специализированные базы

Влияние размера базы

Возможность идентифицировать спектры пептидов, полученные в результате MS/MS экспериментов, используя поисковые базы данных,

зависит от многих факторов. Во-первых, пептид должен присутствовать в поисковой базе. Однако, чем больше сравнивается теоретических спектров с экспериментальным, тем больше вероятность того, что лучший результат будет у неверного теоретического спектра, и тем труднее различать верные и неверные идентификации [9]. В результате, поиск против большой базы может дать несколько новых идентификаций белков или пептидов, но при этом общее количество идентификаций будет меньше, в сравнении с поиском против референсного сиквенса [17, 20]. Так же увеличение базы приводит к увеличению машинного времени, необходимого для этого поиска. Таким образом, одним из ключевых моментов в протеогеномике является поиск баланса между размером базы и её содержимым.

Способы увелечение чувствительности идентификации пептидов

Подходы, используемые в протеомике для увелечения числа идентификаций, включают в себя идентификацию с помощью нескольких поисковых машин одного и того же набора данных [21]. После поиска происходит изменение скорингов идентификации пептидов за счет комбинации нескольких источников информации, используя подходы машинного обучения [9]. Одной из дополнительных стратегий для сокращения пространства поиска является фракционирование смеси пептидов, проводимое до LS-MS/MS анализа. Фракционирование может проходить за счет определенных физико-химических свойств пептида или свойств последовательности пептида. Примером фракционирования может служить изоэлектрическая фокусировка. Спектры, полученные от фракции с определенной изоэлектрической точкой, можно искать против пептидов с при-

мерно такой же теоретически предсказанной изоэлектрической точкой [22].

Чувствительность идентификации так же можно повысить за счет многоступенчатого анализа данных. На первом этапе можно проводить поиск против "эталонной" базы, наиболее точно описывающей исследуемый организм и позволяющей идентифицировать большинство спектров. На втором этапе происходит поиск против расширенной базы для идентификации дополнительных спектров [23, 24]. При таком подходе результаты первоначального поиска используются для уточнения второй базы, используемой при дальнейшем анализе.

Точность идентификации

В протеогеномике, как и в протеомике, для предотвращения накопления ошибок при переходе с PSM на уровень уникальных пептидов, избыточные PSM должны быть свернуты в один PSM с наивысшим скорингом [9]. Пептиды, идентифицированные в различных состояниях (например, двух- или трех-зарядные ионы; модифицированные и не модифицированные формы) так же должны быть объединены или обрабатываться вероятностно, с учетом удельных весов [25]. При использовании многоэтапного поиска, когда результаты первой идентификации используются при подготовки базы для дальнейшей, более специфической идентификации, необходимо на каждом шагу добавлять соответствующие количество decoy-последовательностей в базу [9]. Кроме контроля глобального FDR, необходимо следить за достоверностью каждой отдельной идентификации (например, при глобальном пептидном FDR в 5% ошибка идентификации отдельных пептидов может превышать это значение). Вероятность идентификации белка или события (новый ген,

альтернативный старт, сплайс форма, в случае протеогеномики, может быть рассчитана на основании вероятностей соотвествующих уникальных пептидов [25, 26].

Класс-специфический анализ идентификаций и FDR

Ложная неслучайная идентификация пептидов

Ошибочная идентификация пептида происходит в одном из двух случаев: либо случайное совпадение с высоким скором MS/MS-спектра и несвязанного с ним пептидом из базы, либо в результате гомологичности последовательности пептида из базы и истинной последовательности пептида. Вне зависимости от типа используемой при поиске decoy-базы (например реверсивное или случайное прочтение исходных последовательностей), ложные идентификации второго типа, скорей всего, будут недооценены [9]. Часто ложная идентификация происходит в результате химической модификации высоко представленного пептида, если в результате сдвига масс из-за модификации, масса пептида становится эквивалента массе некоторого другого пептида из базы [27, 28]. Для исключения таких идентификаций можно, например с использованием BLAST, проверить схожесть последовательности каждого нового идентифицированного пептида, с последовательностями всех пептидов, представленных в референсной базе, и исключить (или отдельо контролировать) все высоко гомологичные. Если нужно сохранить для дальнейшего анализа такие пептиды (например при поиске одно аминокислотных замен), нужно проверить, что наблюдаемая разница масс между новым и референсынм пептидом не совпадает с массой какой-либо распространенной химической или пост-трансляционной модификацией [19]. Список наиболее частных модификаций, специфичных для исследуемого биологичесо-

го объекта, можно получить, используя 'blind' поисковый алгоритм [29]. Кроме того, замена лейцина на изолейцин и обратная не может быть идентифицирована с помощью масс-спектрометрии. Пептиды, содержащие такие замены, должны быть исключены из дальнейшего анализа.

Интерпретация данных и новых событий

В протеомике результатом идентификации является список идентифицированных белков или генов, а так же список идентифицированных пептидов, с определенным уровнем FDR. В протеогеномике к таким результатам, так же добавляются списки новых событий, вроде "новый ген "новый кодирующий регион "альтернативный старт"и так далее, с соответствующими пептидами, подтверждающие эти события. Пептиды с одинаковыми последовательностями, могут придти из различных областей генома (например, от паралогов, идентичных сайтов разных белков или из псевдогенов). Такие пептиды не могут быть доказательством экспресии с какого-то определенного участка генома [10]. Кроме того, новые пептиды могут быть интерпретированы разными способами, например как новый транскрипт этого гена или как пептид из интрона или нетранслируемого региона того же гена.

Идентификация новых пептидов

Одним из результатов протеогеномного анализа является список новых пептидов. Этот список зависит от выбора референсной базы и её версии. Как обсуждалось ранее, для многих организмов существуют несколько версий белковых баз, и эти базы отличаются объемом и качественном аннотации. Более того, эти базы регулярно обновляются, в результате в них добавляются и удаляются последовательности. Таким

образом, в протеогеномным исследованиях, пептиды идентифицируемые против специфичных баз, должны быть проверены на вхождение в основанные базы, существующие для данного организма.

3.5. Применение протеогеномики

Поиск новых белок-кодирующих регионов

Возможность применение масс-спектрометрических протеомных данных для поиска новых белок-кодирующих регионов и подтверждения границ уже аннотированных, обсуждалась начиная с первых дней существования протеомики, как науки в её современном виде [30, 31]. Такие результаты чаще всего достигаются за счет поиска против специальных баз данных, полученных в результате шести фреймового транслированя генома, трех фреймового транслирования предсказанных различными методами белок-кодирующих участков генома или шести (трех, в случае стрнед-специфичного секвенирования) фреймового транслирования данных РНК-секвенирования. Как правило, новые белок-кодирующие регионы находят у новых, немодельных организмов [32–34]. Даже для хорошо исследованых эукариот есть работы, в которых находят новые гены. Например, исследование белкового профиля человека и мыши с использованием шестирамочного транслирования, позволило идентифицировать 98 и 52, соотвественно, ранее не аннотированных белок-кодирующих областей [22].

3.6. Идентификация коротких рамок считывания и сайтов инициации трансляции

3.7. Протеогеномика *M.tuberculosis H37Rv*

Для проверки и корректорвки аннотации Келрат с коллегами провели протегеномное исследование штамма *H37Rv* с использованием масс-спектромертрических данных [35]. В своей работе, в качестве геномной базы, они использовали транслированный в шести фреймах геном. В качестве стартовых кодонов использовались GTG и TTG, которые транслируются как метеонин, а не валин и лейцин соотвественно, в случае, если они являются стартовыми сайтами [3].

GSSP пептиды были разделены на 3 группы: 1. относящиеся к межгенным областям 2. частично пересекающиеся с аннотрованным геномом 3. полностью относящиеся к аннотированым генам .

Они обнаружили 41 новый ген, и кореектировку рамки для 79 генов. Из 79 генов с измененной рамкой: для 78 было исправлено положение стартового кодона, и два гены были объеденины в один. Для подтверждения этих результатов, использовались альтернативные программы для аннотации генома (FgeneSB, GeneMark 2.5), а так же поиск гомологий среди известных генов, используя алгоритм blast.

Авторы работы были удивлены, что они обнаружили новые гены с учетом того, что геном был отсеквенированн более 10 лет назад, и аннотирован множеством независимых групп.

4. Материалы и методы

Общая схема эксперимента приведелена на рисунке 1.

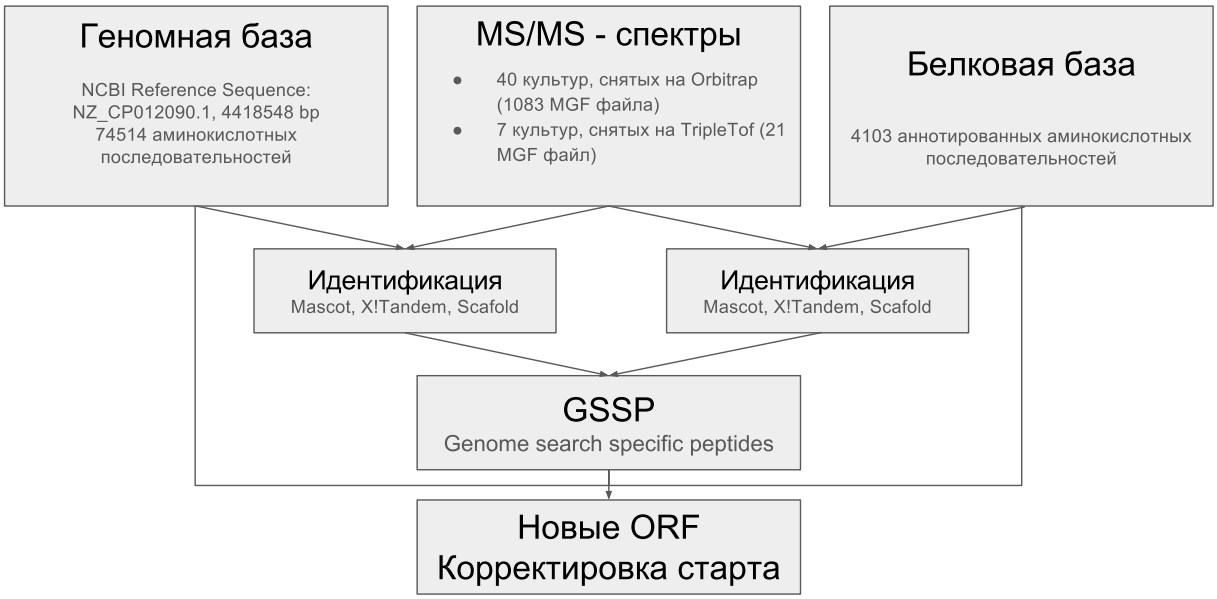

Рис. 1. Схема анализа масс-спектрометрических данных. Можно выделить следующие ключевые шаги: *a)* идентификация против белковой базы *b)* идентификация против геномной базы *c)* нахождение GSSP *d)* интерпретация GSSP.

4.1. Получение бактерий

Бактерии были выделены из мокроты, больных туберкулёзом пациентов.

4.2. Проведение масс-спектрометрического эксперимента

Для протеогеномного анализа использовались масс-спектры белковых клеточных лизатов для *M.tuberculosis*, полученных с прибора AB SCIEX TripleTOF 5600 в лаборатории протеомного анализа НИИ ФХМ и ThermoFisher Q Exactive Hybrid Quadrupole-Orbitrap в ИБМХ.

4.3. Контроль качества

Для всех масс-спектров был проведен контроль качества масс-спектров с использованием программного решения реализованного в лаборатории биоинформатики НИИ ФХМ. В ходе контроля качества были проверены следующие факторы: качество трипсинолиза, распределение зарядов родительских ионов, ошибка измерения m/z для родительских и дочерних ионов, распределение идентифицированных пептидов по времени удерживания пептидов в хроматографической колонке.

4.4. Создание поисковых баз

В работе использовалось 2 типа баз: белковая и геномная. Белковая база - аннотированные последовательности, для данного штамма. Геноманя - база, полученная в результате транслирования генома в шести рамках. Белковые базы для *M.tuberculosis W-148* и *M.tuberculosis H37Rv* были составлены из аннотированных белков штаммов (*W-148*: NCBI Reference Sequence: NZ_CP012090.1, версия от 11 марта 2017 года, 4103 аминокислотных последовательности, 137 псевдогена; *H37Rv*: NCBI Reference Sequence: NC_000962.3, версия от 2 августа 2016 года, 3932 аминокислотных последовательности). Геномные базы были получены в результате 6 рамочного транслирования от стоп- до стоп-каднона геномов штаммов *M.tuberculosis W-148* и *M.tuberculosis H37Rv*, используя программу Artemis версия 16.0.0 [36]. При транслировании использовалась 11 трансляционная таблица NCBI. Минимальная длинна рамки была установлена в 100 нуклеиновых кислот. Пример транслированного в шести рамках учатска генома представлен на рисунке 2. К каждой базе были добавлены последовательности 26 контаминантных белков (кератины, альбумины, трипин) и decoy-последовательности, полученные в

результате прочтения аминокислотных последовательностей с конца, за исключением стартового метеонина.

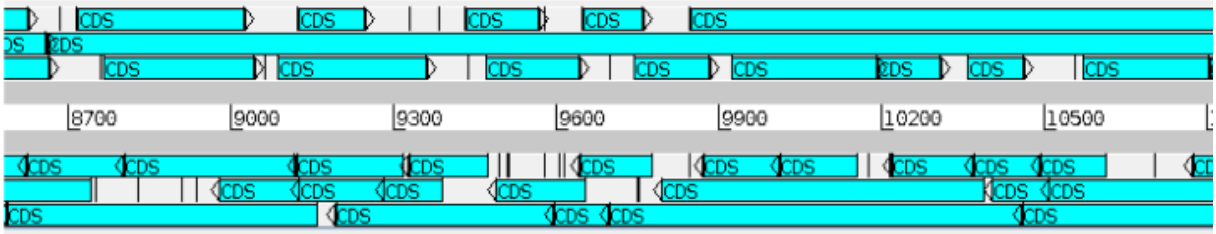

Рис. 2. Транслированный в шести рамках геном. По центру обозначены геномные координаты. Сверху - 1-3 фрейм, снизу 4-6 фреймы. Вертикальными линиями обозначены стоп-кодоны. Голубым - ORF, длинной более 100 пар оснований.

4.5. Идентификация пептидов и белков

Данные полученные в результате LC-MS/MS эксперимента (Raw формат) были сконвертированы в пик-лист (MGF формат), используя ProteoWizard msconvert [37]. Идентификация проходила против двух белковых и двух геномных баз с использованием Mascot Search Engine version 2.5.1 [38] и X!Tandem version 3.4.3 [39]. Результаты идентификации двух программ были объеденины в Scaffold version 4.2.1.

Параметры поиска Mascot были следующими: триптические пептиды, не более одного пропущенного сайта трипсинолиза, ошибка массы прекурсера 20 ppm, ошибка массы фрагментов 0.04 Да, заряды прекурсера 2+, 3+, 4+. Oxidation(M) была устанолвена как возможнная модификация пептидов, Carbamidomethylation(C) как фиксированная.

Параметры X!Tandem были следующими: триптические пептиды, не более одного пропущенного сайта трипсинолиза, ошибка массы прекурсера 20 ppm, ошибка массы фрагментов 50 ppm, проверка не моноизотопных масс, Carbamidomethylation(C) - фиксированная модификация, Oxidation(M) возможная модификация.

Результаты работы поисковых машин были объединины в Scaffold с параметрами: 1:1 forward/decoy ratio, LFDR scoring, стандартные белковые группы, не проводить GO-аннотацию. Белковый и пептидный FDR был установлен на уровне 1%, 1 и более пептидов на белок. Результаты были экспортированны в виде листа идентифицированных пептидов.

4.6. Протеогеномика *W-148*

Координаты аннотированных генов были пересечены с учетом стренда и фрейма с координатами ORF, полученными в результате шестирамочного транслирования. Для поиска GSSP из результатов поиска против геномной базы *W-148* были исключены пептиды, идентифицированные против белковой базы *W-148*. Так же были исключены пептиды идентифицируемые против геномной базы и представленные в аннотации, и пептиды, идентифицируемые только в одной культуре. Для дальнейшего анализа были выбраны ORF, в которых произошло одно из следующих событий: 1. идентифицировано два и более уникальных GSSP 2. идентифицирован GSSP и присутствует аннотированный ген 3. идентифицирован GSSP и есть пересечение по координатам с псевдогеном в пределах стренда .

Валидация результатов идентификации

Для каждого GSSP была проведена проверка времени выхода и ошибки масс при идентификации; потенциальной остаточной контаминации на приборе; ошибки интерпретации модифицированной акминокислоты, как другой немодифицированной.

Проверка ошибки идентификации масс проходила на уровне PSM. В каждом ране, для каждого PSM, соответсвующему GSSP, находилась

среднее значение и стандартное отклонение ошибки идентификации всех PSM в диапазоне +/- 5 минут от времени выхода данного PSM. Из дальнейшенго анализа исключались все PSM, ошибка идентификации масс которых отличалась более чем на 3 стандартных отклоения от средей ошибки в установленном временом интервале. В каждом ране были отфильтрованны 5% самых ранних и поздних по време выхода PSM.

Для проверки времени выхода результаты идентификации были разделены на две груупы: PSM, относящиеся к идентифицированным против белковой базы, и PSM относящиеся к GSSP. Для каждого пептида была вычислена его гидрофобность, используя библиотеку qqman языка R. Пр PSM, относящиеся к белковой идентификации была линейная регрессия, где в качестве зависимой переменной использовалась время выхода, а независимой - гидрофобность. Полученная модель была применена к PSM, относящимся к GSSP. 10% пептидов, показавших наибольшее отклонение были исключены из дальнейшего анализа.

Был проведен поиск точного и с учетом одной возможной замены вхождения GSSP в другие белки, представленные в базе NCBInr.

Идентификация новых белков

Рассматривались ORF, в которых было идентифицировано два и более уникальных GSSP-пептида, прошедших все этапы валидации, и в которых не содержится аннотированный ген. Для проверки потенциала кодирующей способности рамки, был проведён blastp против базы nr.

Уточнение N-концов

Рассматривались ORF, в которых было идентифицировано два и более уникальных GSSP-пептида и в которых содержится аннотиванный

ген. Новые рамки сравнивались с аналогичными генами в *H37Rv*.

4.7. Визуалиция данных

Для визуализации данных использовался Gbrowse. Были выделены следующие глифы: 1. аннотированные гены 2. идентифицированные пептиды 3. псевдогены 4. ORF с новыми генами 5. ORF с пептидами, идентифицируемые перед аннотированным стартом 6. GSSP-пептиды. Результаты идентификации были обработаны и экспортированы в gff3 формате.

5. Результаты и обсуждение

5.1. Подготовка и проведение масс-спектрометрических измерений

Тут что-нибудь про бактерий напиать. Были отсняты масс-спектры 30 культур. Из них 6 были не фракционированы, остальные были разделены на 6 фракций. Разделение на фракции позволило идентифицировать пептиды, которые не были бы идентифицированы при обычном поиске, из-за того, что на измерения MS/MS спектра были бы отобраны более высоко представленные пептиды. Количество отснятых спектров для каждой культуры представлены в таблице.

5.2. Подготовка баз

Скаченная с NCBI аннотация содержала 4103 аннотированных белок-кодирующих последовательностей. После добавления контаминант и decoy-последовательностей получилась белковая база, состоящая из 8258 последовательностей. После транслирования генома в 6 рамках и исключения коротких последовательностей получилось 74488 белок-кодирующих последовательностей. В результате добавления контаминант и decoy-последовательностей, получилась геномная база, состоящая из 149028 последовательностей. Исключение коротких последовательностей из геномной базы позволило ускорить поиск; снизить пороговые значения идентификации, тем самым повысив чувствительность подхода. Минимальная длинна аннотированного белка *W-148* составляет 84 аминокислоты, таким образом, исключение рамок длинной меньше чем 33 аминокислоты не должно привести к потерям при идентификации.

5.3. Идентификация белков и пептидов

Поиск проходил при помощи поисковых маших Mascot и X!Tandem. Объединение результатов поиска и перерасчет FDR был произведен в Scaffold. Против белковой базы было идентифицировано 32054 пептида (1041059 psm), против геномной базы 36502 уникальных пептида (1131085 psm). Пересечение идентифицированных пептидов представлено на рисунке 3. Часть пептидов идентифицирована против белковой базы и не идентифицирована против более полной полной геномной базы. Это связано с различными пороговыми скорингами при поиске против баз разных размеров. После вычитания результатов поиска против белковой базы из результатов поиска против геномной базы получилось 6015 GSSP (Genome search specific peptides). После исключения пептидов представленных в аннотированном геноме и идентифицированных только против геномной базы, осталось 1397 GSSP. Наличие таких пептидов, идентифицируемых только против геномной базы и представленных в аннотации, связано с пересчетом FDR: при отдельном рассмотрении результатов идентификации каждой поисковой машины таких эффектов не возникает. После исключения GSSP, представленных только в одной культуре, осталось 425 пептидов. Результаты интерпретации GSSP при таких фильтрах: 16 новый ORF, и 304 гена с корректированным положением старта. Пептидов, интерпретируемых как пептиды перед аннотированным стартом примерно в 8 раз больше, чем пептидов, относящихся к новым генам. Такой разброс может быть связан с тем, что пептидам относящимся к корректировке рамки проще пройти порог FDR. В самом деле, при пептидном и белковом FDR в 1% и, примерно, 1000000 psm при размере базы в 4000 аминокислотных последовательностей, 10000 пептидов будут ложно-положительно идентифицированы. Если брать критерий 2 и

более пептидов для идентификации белка, то такого количества пептидов будет достаточно для идентификации 5000 белков, если не учитывать белковой FDR. С учетом белкового FDR количество ложно-положительно идентифицированных белков должно быть не боле 40. Для этого достаточно 80 пептидов. Таким образом, в экспериментах с большим количеством исходных данных, белковый FDR становится более жестким критерием, чем пептидный. Соответственно, GSSP относящимся к корректировке рамки проще пройти белковый FDR, так как в этой рамке так же присутствуют пептиды из аннотированной части последовательности. В случае нового гена в "прохождение" белкового FDR участвуют только GSSP пептиды.

Для подтверждения результатов были применены дополнительные критерии. После удаления PSM, ошибка идентификации которых составляет более трех стандартных отклонений в предположении о нормальном распределении ошибки идентификации всех PSM в интервале +/- 5 минут, остался 331 уникальный GSSP. Следует отметить, что на этом шаге отсекались пептиды, ошибка идентификации которых меньше, чем реальная точность приборов. Поэтому данный критерий отнесен к дополнительным. Затем была проведена фильтрация по времени выхода. После исключения 10% наиболее отклоняющихся пептидов, остался 147 уникальный GSSP.

Все GSSP были проверены на точное вхождение в базу NCBInr. Среди белков, в которых нашлись GSSP, не было найдено таких, которые бы относились к организмам, которые ранее снимались на используемом масс-спектрометре. Таким образом можно исключить остаточную контаминацию на приборе и пробоподготовке. Так же были исключены 8 пептидов, которые присутствуют в аннотированных последовательность *W-148* с учетом одной замены.

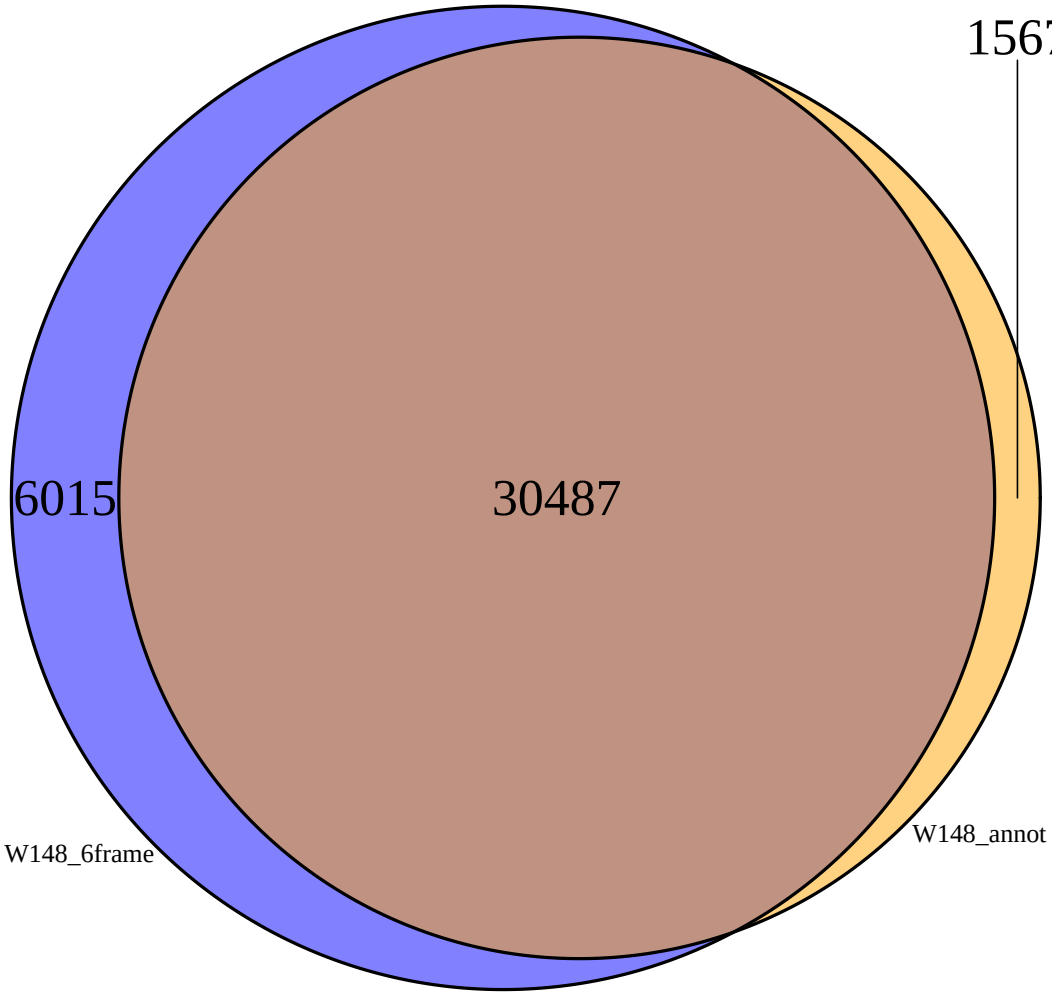

Рис. 3. Сравнение идентификаций против различных баз: *a)* annot - количество уникальных пептидов идетифицированных против белковой базы *b)* 6frame - количество уникальных пептидов идетифицированных против геномной базы.

5.4. Интерпретация новых событий

Новые ORF

Было идентифицировано 16 новых ORF, в которых присутствует 2 и более уникальных GSSP. Из шестнадцати ORF пять пересекаются с аннотированными псевдогенами, одиннадцать лежат на комплиментарной цепи участков с аннотированными генами. У пяти из шестнадцати есть гомолог в *H37Rv*. Гены всех *M.tuberculosis* плотно расположены, и межгенные области либо отсутствуют, либо их длинна намного меньше длины гена, либо, если межгенник большой, в нем находится псевдолен. Поэтому не найдено новых генов, которые не относились бы к псевдогенам и лежали в межгенном пространстве.

Следует отметить, что причины по которым ген становится "псевдоегном" с точки зрения биологии и системы аннотации NCBI различны. Так, наиболее частыми причинами из-за которых участок генома аннотируется как псевдоген являются: фреймшифт (потеря или вставка не кратного трем числа нуклеотидов, в результате чего нарушается белковая последовательность), неполный ген (присутвствует только часть гена, в сравнении с гамологами), стоп-кодон по середине последовательности, низкое качество сборки (например, если ген находится на стыке контигов). Для всех идентифицированных псевдогенов была найдена "техническая" и "биологическая" причина, из-за которой они получили статус псевдогена. Пример идентифицированно псевдогена представлен на рисунке 4.

В результате поиска против базы NCBInr при помощи алгоритма blust для десяти из одинналцати ORF, лежащих на комплиментарной цепи (пример такого ORF предствлен на рисунке 5) к аннотированным генам, были найдены гомологи. Эти гомологи были аннотированы как

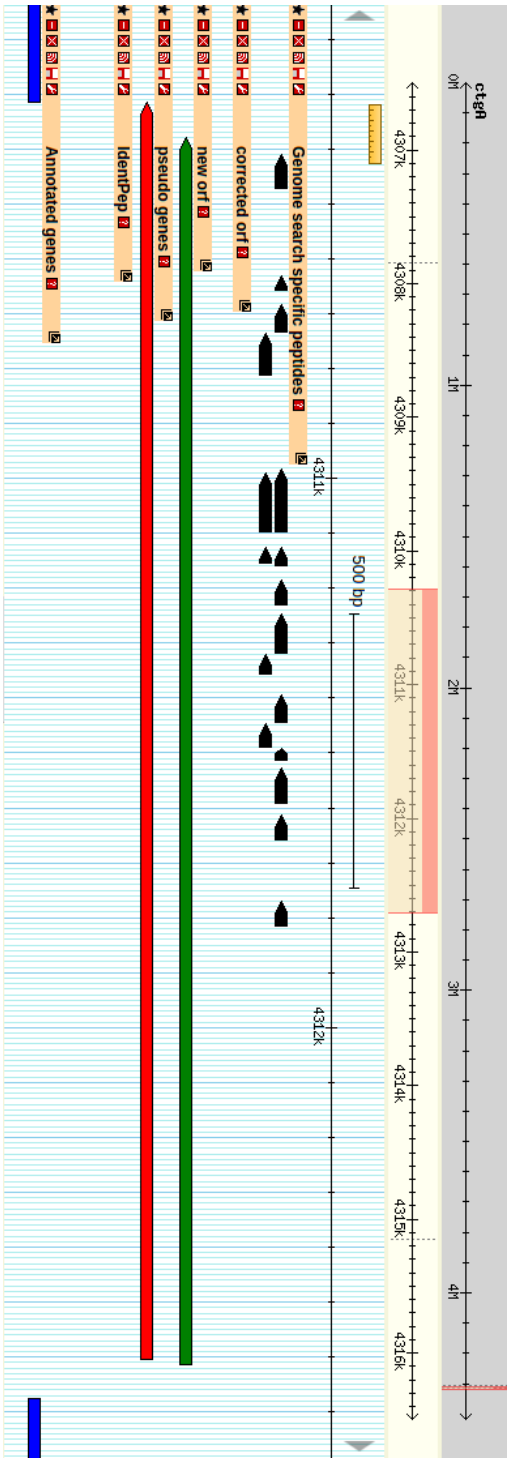

Рис. 4. Идентифицированный псевдоген. Сними обозначены аннотированые гены, красным - псевдоген, зеленым - открытая рамка считывания, Genome search specific peptides - пептиды, идентифицируемы только про поиски против геномной базы

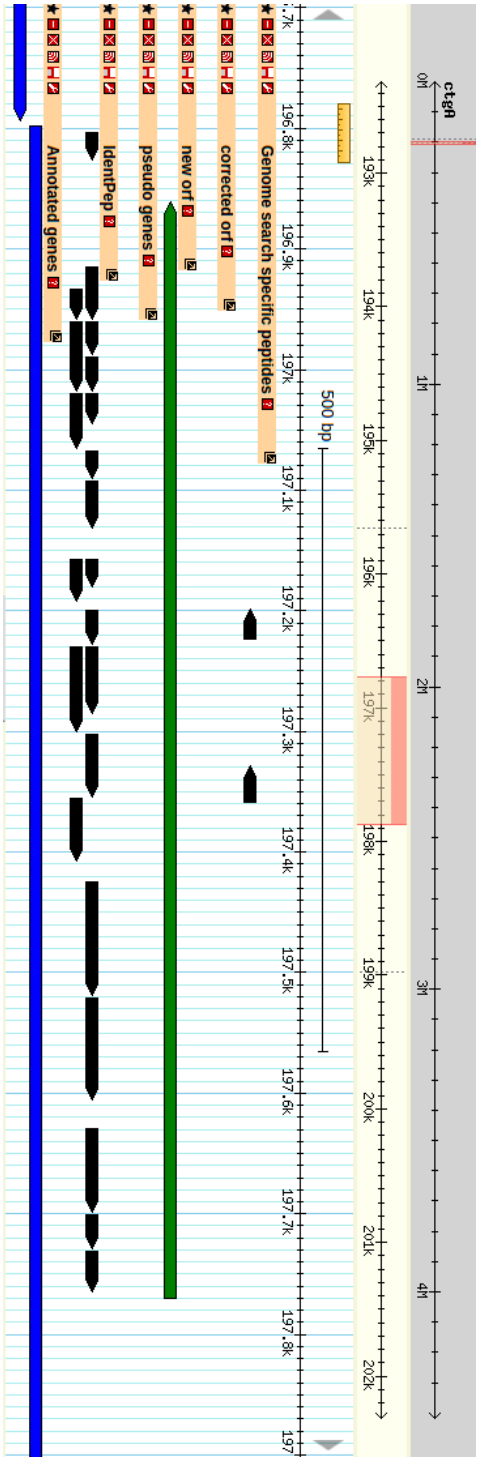

Рис. 5. ORF, лежащий на комплиментраной цепи к аннотированому и экспрессирующемуся гену. Сними обозначены аннотированые гены, IdentPep - идентифицрованные при поиске против белковой базы пептиды, красным - псевдоген, зеленым - открытая рамка считывания, Genome search specific peptides - пептиды, идентифицируемы только про поиски против геномной базы

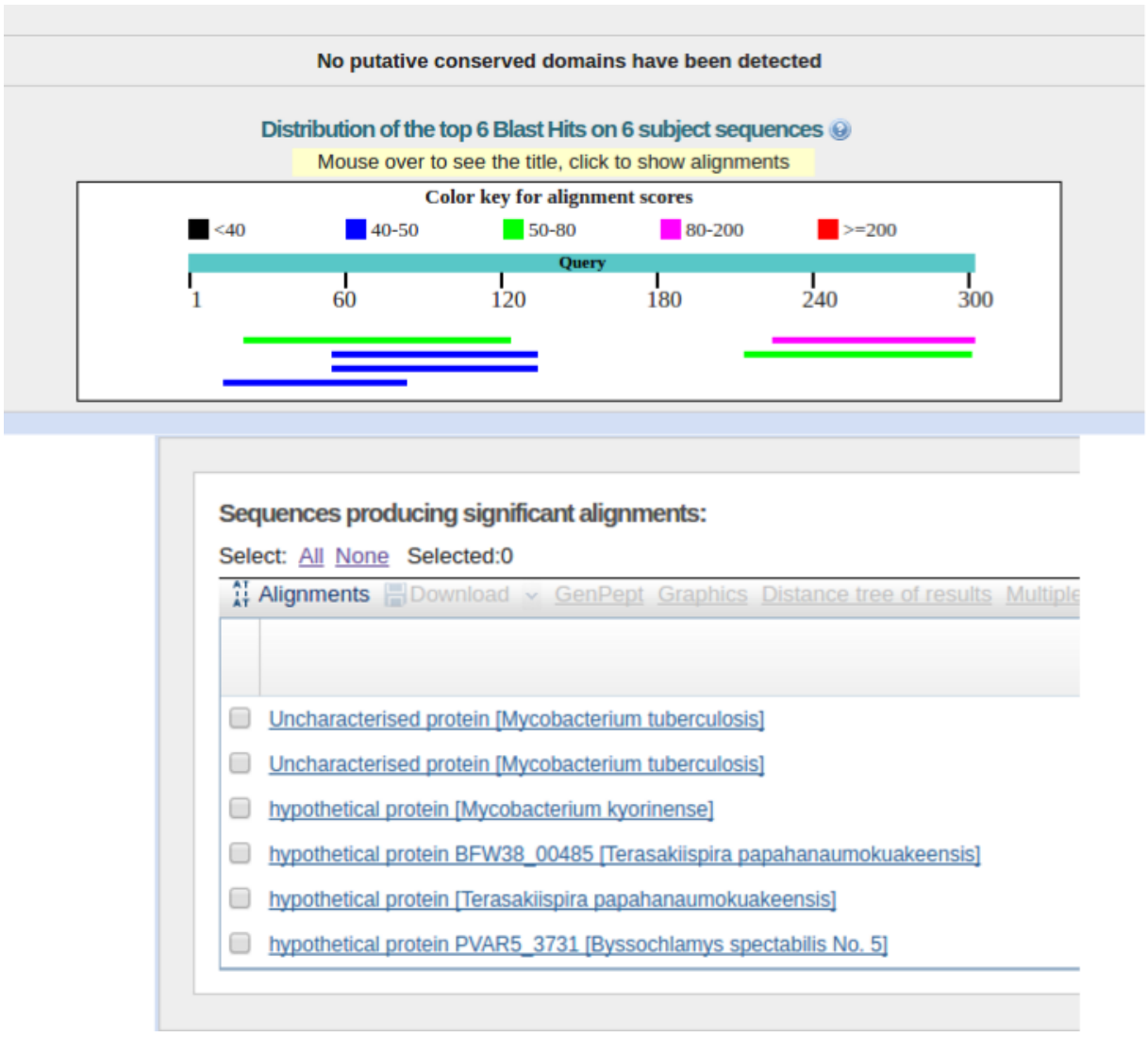

Рис. 6. Результат поиска гомолога ORF, лежащей на комплиментарной цепи и представленной на рисунке 5) против базы NCBInr с ипользованием алгоритма blast.

“гипотетический/предсказанный/непроверенный белок” другого штамма *M.tuberculosis*. Косвенно результаты идентификации подтверждает тот факт, что все новые ORF лежат на комплиментарной цепи, а не в другом фрейме аннотированной. В самом деле, с пространственной точки зрения предположение, что транскрипт снимается с комплиментарной цепи выглядит более вероятным, чем предположение, что с одного транскрипта идет трансляция в двух фреймах двух разных аминокислотных последовательностей.

После применения дополнительных критериев фильтрации GSSP осталось шесть ORF с псевдогенами и один ORF на комплиментарной цепи.

Корректировка положения аннотированного старта

Всего 304 рамки содержат аннотированный ген и GSSP пептид. Из 304 308 содержат два и более GSSP. Эти рамки сравниили с гомологичными генами в *H37Rv*. Из 38 36 совпадают с точностью до SAP, 2 рамки длинней у *W-148*, чем у *H37Rv*. Для этих 38 рамок были проверены пептиды, пересекающиеся аннотированный старт. Для 17 из 38 были найдены пептиды, содержащие в себе аннотированный старт (overlap-пептиды), для 7 были найдены стартовые пептиды из аннотации, для 3 были найдены как стартовые, так и аннотированные пептиды.

Пептиды, содержащие аннотированный старт

Для 17 из 38 рамок, содержащих аннотированный ген и два более GSSP, были найдены пептиды, содержащие в себе стартовую аминокислоту для аннотированного гена. Такой пептид представлен на рисунке 7. Для 7 из 38 были найдены пептиды, являющиеся стартовыми для анноти-

рованного гена. У 3 из 38 найдены как стартовые, так и overlap-пептиды, причем в этом месте не было сайта трипсинолиза.

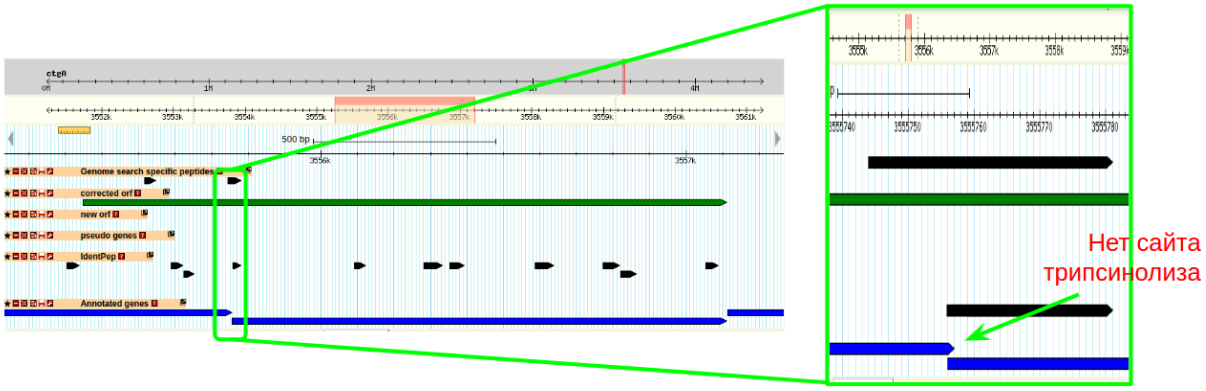

Рис. 7. Ген с измененой рамкой считывания. Черным обозначены идентифицированные пептиды. Верхний глиф - GSSP, нижний - пептиды, идентифицированные при поиске против белковой базы. Идентифицирован как стартовый пептиы, так и overlap-пептид. Сайта трипсинолиза в этом месте нет.

6. Выводы

В результате протеогеномного анализа штамма *Mycobacterium tuberculosis W-148* были идентифицированы 16 новых генов и у 249 (24) открытых рамок считывания были скорректированы старты трансляции

Проведенная реаннотация генома штамма Mycobacterium tuberculosis W-148 была подтверждена несколькими биоинформатическими подходами

Список литературы

1. Jungblut P., Schaible U., Mollenkopf H.-J. et al. Comparative proteome analysis of Mycobacterium tuberculosis and Mycobacterium bovis BCG strains: towards functional genomics of microbial pathogens // Molecular microbiology. 1999. Vol. 33, no. 6. P. 1103–1117.
2. Mattow J., Siejak F., Hagens K. et al. Proteins unique to intraphagosomally grown Mycobacterium tuberculosis // Proteomics. 2006. Vol. 6, no. 8. P. 2485–2494.
3. Cole S., Brosch R., Parkhill J. et al. Erratum: Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence // Nature. 1998. Vol. 396, no. 6707. P. 190.
4. Fleischmann R., Alland D., Eisen J. A. et al. Whole-genome comparison of Mycobacterium tuberculosis clinical and laboratory strains // Journal of bacteriology. 2002. Vol. 184, no. 19. P. 5479–5490.
5. Camus J.-C., Pryor M. J., Médigue C., Cole S. T. Re-annotation of the genome sequence of Mycobacterium tuberculosis H37Rv // Microbiology. 2002. Vol. 148, no. 10. P. 2967–2973.
6. De Souza G. A., Målen H., Søfteland T. et al. High accuracy mass spectrometry analysis as a tool to verify and improve gene annotation using Mycobacterium tuberculosis as an example // Bmc Genomics. 2008. Vol. 9, no. 1. P. 316.
7. Lew J. M., Kapopoulou A., Jones L. M., Cole S. T. TubercuList–10 years after // Tuberculosis. 2011. Vol. 91, no. 1. P. 1–7.
8. Bantscheff M., Lemeer S., Savitski M. M., Kuster B. Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present // Analytical and bioanalytical chemistry. 2012. Vol. 404, no. 4. P. 939–965.

9. Nesvizhskii A. I. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics // Journal of proteomics. 2010. Vol. 73, no. 11. P. 2092–2123.
10. Nesvizhskii A. I., Aebersold R. Interpretation of shotgun proteomic data the protein inference problem // Molecular & Cellular Proteomics. 2005. Vol. 4, no. 10. P. 1419–1440.
11. Dasari S., Chambers M. C., Slebos R. J. et al. TagRecon: high-throughput mutation identification through sequence tagging // Journal of proteome research. 2010. Vol. 9, no. 4. P. 1716.
12. Jaffe J. D., Berg H. C., Church G. M. Proteogenomic mapping as a complementary method to perform genome annotation // Proteomics. 2004. Vol. 4, no. 1. P. 59–77.
13. Nesvizhskii A. I. Proteogenomics: concepts, applications and computational strategies // Nature methods. 2014. Vol. 11, no. 11. P. 1114–1125.
14. Harrow J., Frankish A., Gonzalez J. M. et al. GENCODE: the reference human genome annotation for The ENCODE Project // Genome research. 2012. Vol. 22, no. 9. P. 1760–1774.
15. Baerenfaller K., Grossmann J., Grobei M. A. et al. Genome-scale proteomics reveals Arabidopsis thaliana gene models and proteome dynamics // Science. 2008. Vol. 320, no. 5878. P. 938–941.
16. Khatun J., Yu Y., Wrobel J. A. et al. Whole human genome proteogenomic mapping for ENCODE cell line data: identifying protein-coding regions // BMC genomics. 2013. Vol. 14, no. 1. P. 141.
17. Blakeley P., Overton I. M., Hubbard S. J. Addressing statistical biases in nucleotide-derived protein databases for proteogenomic search strategies // Journal of proteome research. 2012. Vol. 11, no. 11. P. 5221–5234.
18. Derrien T., Johnson R., Bussotti G. et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution,

and expression // Genome research. 2012. Vol. 22, no. 9. P. 1775–1789.

19. Li J., Su Z., Ma Z.-Q. et al. A bioinformatics workflow for variant peptide detection in shotgun proteomics // Molecular & Cellular Proteomics. 2011. Vol. 10, no. 5. P. M110–006536.
20. Krug K., Carpy A., Behrends G. et al. Deep coverage of the Escherichia coli proteome enables the assessment of false discovery rates in simple proteogenomic experiments // Molecular & Cellular Proteomics. 2013. Vol. 12, no. 11. P. 3420–3430.
21. Shteynberg D., Nesvizhskii A. I., Moritz R. L., Deutsch E. W. Combining results of multiple search engines in proteomics // Molecular & Cellular Proteomics. 2013. Vol. 12, no. 9. P. 2383–2393.
22. Branca R. M., Orre L. M., Johansson H. J. et al. HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics // Nature methods. 2014. Vol. 11, no. 1. P. 59–62.
23. Ning K., Fermin D., Nesvizhskii A. I. Computational analysis of unassigned high-quality MS/MS spectra in proteomic data sets // Proteomics. 2010. Vol. 10, no. 14. P. 2712–2718.
24. Helmy M., Sugiyama N., Tomita M., Ishihama Y. Mass spectrum sequential subtraction speeds up searching large peptide MS/MS spectra datasets against large nucleotide databases for proteogenomics // Genes to Cells. 2012. Vol. 17, no. 8. P. 633–644.
25. Shteynberg D., Deutsch E. W., Lam H. et al. iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates // Molecular & Cellular Proteomics. 2011. Vol. 10, no. 12. P. M111–007690.
26. Castellana N., Bafna V. Proteogenomics to discover the full coding content of genomes: a computational perspective // Journal of proteomics. 2010. Vol. 73, no. 11. P. 2124–2135.

27. Nesvizhskii A. I., Roos F. F., Grossmann J. et al. Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides // Molecular & Cellular Proteomics. 2006. Vol. 5, no. 4. P. 652–670.
28. Abraham P., Adams R. M., Tuskan G. A., Hettich R. L. Moving away from the reference genome: Evaluating a peptide sequencing tagging approach for single amino acid polymorphism identifications in the genus Populus // Journal of proteome research. 2013. Vol. 12, no. 8. P. 3642–3651.
29. Tsur D., Tanner S., Zandi E. et al. Identification of post-translational modifications via blind search of mass-spectra // Computational Systems Bioinformatics Conference, 2005. Proceedings. 2005 IEEE / IEEE. 2005. P. 157–166.
30. Choudhary J. S., Blackstock W. P., Creasy D. M., Cottrell J. S. Interrogating the human genome using uninterpreted mass spectrometry data // Proteomics. 2001. Vol. 1, no. 5. P. 651–667.
31. Andersen J. S., Mann M. Mass spectrometry allows direct identification of proteins in large genomes // Proteomics. 2001. Vol. 1, no. 5. P. 641g650.
32. Castellana N. E., Shen Z., He Y. et al. An automated proteogenomic method uses mass spectrometry to reveal novel genes in Zea mays // Molecular & Cellular Proteomics. 2014. Vol. 13, no. 1. P. 157–167.
33. Castellana N. E., Payne S. H., Shen Z. et al. Discovery and revision of Arabidopsis genes by proteogenomics // Proceedings of the national academy of sciences. 2008. Vol. 105, no. 52. P. 21034–21038.
34. Yang X., Tschaplinski T. J., Hurst G. B. et al. Discovery and annotation of small proteins using genomics, proteomics, and computational

approaches // Genome research. 2011. Vol. 21, no. 4. P. 634–641.

35. Kelkar D. S., Kumar D., Kumar P. et al. Proteogenomic analysis of Mycobacterium tuberculosis by high resolution mass spectrometry // Molecular & Cellular Proteomics. 2011. Vol. 10, no. 12. P. M111–011627.
36. Rutherford K., Parkhill J., Crook J. et al. Artemis: sequence visualization and annotation // Bioinformatics. 2000. Vol. 16, no. 10. P. 944–945.
37. Chambers M. C., Maclean B., Burke R. et al. A cross-platform toolkit for mass spectrometry and proteomics // Nature biotechnology. 2012. Vol. 30, no. 10. P. 918–920.
38. Cottrell J. S., London U. Probability-based protein identification by searching sequence databases using mass spectrometry data // electrophoresis. 1999. Vol. 20, no. 18. P. 3551–3567.
39. Fenyö D., Beavis R. C. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes // Analytical chemistry. 2003. Vol. 75, no. 4. P. 768–774.