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: **Количественный протеогеномный анализ туберкулеза и ещё чего-нибудь**

Направление: 010900 – Прикладные математика и физика

Магистерская программа: 010982 – Физико-химическая биология и биотехнология

Выполнил студент гр. 1114 _____ Смоляков А.В.

Научный руководитель,

к. б. н. _____ Лазарев В.Н.

Работа выполнена в ФГБУ ФНКЦ ФХМ ФМБА России

Москва – 2017

Оглавление

1. Список сокращений 4
2. Введение 5
3. Обзор литературы 6
   - 3.1. Mycobacterium tuberculosis 6
   - 3.2. Применение масс-спекртрометрии в протеомике 6
   - 3.3. Orbitrap 6
   - 3.4. Протеогеномика 6
     - Подходы к созданию баз 6
     - Поиск новых генов и корректировка рамок 6
4. Материалы и методы 7
   - 4.1. Получение бактерий 7
   - 4.2. Проведение масс-спектрометрического эксперимента 7
   - 4.3. Контроль качества 7
   - 4.4. Создание поисковых баз 7
   - 4.5. Идентификация пептидов и белков 7
   - 4.6. Протеогеномика *W-148* 8
     - Идентификация новых белков 8
     - Уточнение N-концов 8
   - 4.7. Сравнение идентификаций против *W-148* и *H37Rv* 8
     - Поиск новый генов 8
     - Уточнение N-концов 8
     - Анализ SAP 8
5. Результаты и обсуждение 9
   - 5.1. Протеогеномика *W-148* 9
     - Идентификация 9
     - Новые гены и их валидация 9
     - Уточнение N-концов 9
   - 5.2. Сравнение идентификаций против *W-148* и *H37Rv* 9
     - Новые гены и их валидация 9
     - Уточнение N-концов 9

Валидация SAP 9
6. Выводы 10

Список литературы 11

1. Список сокращений

GSSP - Genome Search Specific Peptides. Это пептиды, индетифициуремые при поиске против геномной базы, и не идентифицируемые при происке против протеомной.

2. Введение

3. Обзор литературы

3.1. Mycobacterium tuberculosis

3.2. Применение масс-спекртрометрии в протеомике

3.3. Orbitrap

3.4. Протеогеномика

Подходы к созданию баз

Поиск новых генов и корректировка рамок

4. Материалы и методы

4.1. Получение бактерий

4.2. Проведение масс-спектрометрического эксперимента

4.3. Контроль качества

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 (NCBI Reference Sequence: NZ_CP012090.1, 4244 аминокислотных последовательностей для *W-148* и). 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 [1]. При транслировании использовалась 11 трансляционная таблица NCBI. Минимальная длинна рамки была установлена в 100 нуклеиновых кислот. К каждой базе были добавлены последовательности 26 контаминантных белков (кератины, альбумины, трипин).

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 [2]. Идентификация проходила против двух белковых и двух геномных баз с использованием Mascot Search Engine version 2.5.1 [3]. Параметры поиска были следующими: триптические пептиды, не более двух пропущенных сайтов трипсинолиза, ошибка массы прекурсера 20 ppm, ошибка массы фрагментов 0.5 Да, заряды прекурсера 2+, 3+, 4+. Oxidation(M), Carbamidomethylation(C) and Deamidated(NQ) были установлены как возможнные модификации пептидов. Для подсчета FDR и порогового скоринга использовался поиск против decoy-базы, полученной в результате реверса исходной базы. FDR был выбран на уровне 5%. Пептид считался идентифицированным, если его скор выше порогового скоринга и ранг равен еденице. Белок считался

идентифицированным, если для него нашлось два и более уникальых 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*.

Идентификация новых белков

Рассматривались ORF, в которых было идентифицированно два и более уникальных пептидов.

Уточнение N-концов

4.7. Сравнение идентификаций против *W-148* и *H37Rv*

Поиск новый генов

Уточнение N-концов

Анализ SAP

5. Результаты и обсуждение

5.1. Протеогеномика *W-148*

Идентификация

Новые гены и их валидация

Уточнение N-концов

5.2. Сравнение идентификаций против *W-148* и *H37Rv*

Новые гены и их валидация

Уточнение N-концов

Валидация SAP

6. Выводы

Список литературы

1. Rutherford K., Parkhill J., Crook J. et al. Artemis: sequence visualization and annotation // Bioinformatics. 2000. Vol. 16, no. 10. P. 944–945.
2. 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.
3. 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.