Задача от к.б.н. Н.Ю. Опариной (Стокгольм, Каролинский институт)

Долгое время то, как экспрессируются гены человека в различных органах и тканях, изучали с помощью микропанелей с заранее подготовленными и нанесенными зондами, специфичными для отдельных участков мРНК. Развитие методов массового секвенирования нового поколения позволило получать информацию о полном наборе экспрессирующихся РНК в конкретном препарате, в том числе и для ранее неизвестных генов или альтернативных изоформ. В результате альтернативного сплайсинга один ген может кодировать несколько различных по структуре и функции белков. Но в получаемых данных очень много шума, значительная часть описанных вариаций альтернативного сплайсинга не консервативна при сравнении даже эволюционно близких видов, а предсказанные на основе анализа транскриптомов изменения чаще всего не удается подтвердить, анализируя протеом. Ну а если мы можем анализировать транскрипты данного гена не только в одном образце, а в нескольких? Десятки или сотни образцов одной и той же ткани или клеточной линии, полученных из разных лабораторий и просеквенированных в разных центрах? И сравнить их со столь же разнообразными данными для другой ткани? Или этой же клеточной линии, обработанной низкомолекулярными соединениями, например, лекарственными препаратами?

База данных Intropolis предоставляет самые масштабные на сегодняшний день данные анализа более 20 тысяч транскриптомов человека с помощью программы Rail-RNA. В результате к геномным координатам для каждого исследуемого транскриптома "привязаны" экзон-экзонные границы, подтвержденные по крайней мере одним "ридом".

Данные представлены в базе https://github.com/nellore/int и содержат следующую информацию:

chromosome
intron start position (1-based; inclusive)
intron end position (1-based; inclusive)
strand (+ or -)
donor dinucleotide (e.g., GT)
acceptor dinucleotide (e.g., AG)
comma-separated list of indexes of samples in which junction was found
comma-separated list of corresponding numbers of reads mapping across junction in samples from field 7

В других файлах базы есть информация о соответствии внутреннего идентификатора образца (Intropolis) и его идентификатора во внешнем источнике (SRA, идентификатор "чтения" конкретного транскриптома; идентификатор "образца"). Несколько образцов могут входить в одно и то же "исследование", например, здоровая и опухолевая ткань. В отдельном файле присутствуют метаданные: текстовые описания, названия тканей или клеток, тип обработки, если она проводилась, комментарии и т.д.

Нетрудно заметить, что в таком виде ресурс неудобен и не позволяет с ним работать биологам-экспериментаторам. Нами подобран список интересных генов, например, ассоциированных с аутоммунными заболеваниями и их геномных координат. Мы хотим узнать: есть ли среди них те, для которых характерен тканеспецифичный или регулируемый альтернативный сплайсинг? Можем ли мы использовать эти данные для суммарной оценки уровня экспрессии гена или его известного "канонического" транскрипта и оценки дифференциальной экспрессии генов в наборах образцов?

(Для каждого SRA-"чтения" в метаданных присутствует общее число "ридов", для каждого гена известны геномные координаты). Можем ли мы оценить уровень шума и потенциальных ошибок сплайсинга по набору "нестабильных", не подвержденных в большом наборе образцов экзон-экзонных границ? Это может говорить о том, например, что наблюдаемые изменения уровня мРНК не соответствуют изменениям уровня белка.

Приветствуются способы фильтрации шума, визуализации данных, интерактивные подходы, позволяющие, например, группировать транскриптомные "чтения" по метаданным и анализировать стабильные и альтернативные экзон-экзонные границы для выбранных генов (заданы геномными координатами).

Дополнительные методы и метрики оценки можно использовать на основе Portcullis (http://portcullis.readthedocs).

Результаты выполнения проекта могут стать как частью исследования генетических ассоциаций с аутоиммунными заболеваниями, для ряда которых показано влияние на альтернативный сплайсинг регулируемых генов, так и самостоятельным исследованием или методом, пригодным для широкого практического использования в работе биологов-экспериментаторов и биоинформатиков.