Лекция «Исследование качественных и количественных характеристик транскриптома».
Первая часть.

Первая лекция посвящена определению термина «транскриптом», выявлению места транскриптома среди других предметов исследования науки о функциональной организации генома, а также обзору современных методов получения качественных и количественных данных о нем.

Слайд 2.

Транскриптом – Особенность транскриптома в том, что, поскольку он является первым уровнем фенотипа, т.е. первым уровнем развертывания и реализации генетической информации, заключенной в геноме.

Его структура сложно организована, поскольку отражает зависимость от стадий клеточного цикла, от типа клеток и тканей, от стадий развития организма, от наличия внешних сигналов как самой транскрипции генов, так и различных пост-транскрипционных процессов – сплайсинга, редактирования, взаимодействия с микроРНК и короткими интерферирующими РНК

Иными словами, транскриптому изначально присуща пространственная и временна́я дифференциальность в распределении транскриптов разных генов и изоформ транскриптов отдельного гена.

Исследование транскриптома – одна из задач функциональной геномики. Напомню эти задачи, которых несколько в соответствии с тем набором функций, который выполняет геном.

Функции генома:

– содержать/кодировать генетическую информацию и генетические программы развития,
– сохранять их в процессах жизнедеятельности одной особи - (1) в течение одного клеточного цикла - репликация, репарация, генная конверсия, борьба с транспозабельными элементами и т.д., (2) в процессе пролиферации клеток - передача через ряд митотических делений – компактизация и декомпактизация хромосом, митоз и т.д.,
– передавать их в ряду поколений – процессы мейоза, сегрегации хромосом, рекомбинации/кроссинговера, поддержания целостности теломер и т.д.,
– создавать/поддерживать условия для формирования транскриптома, т.е. набора транскриптов, характерного для клеток разных типов и предопределенного генетическими программами развития - процессы (1) транскрипции (сайленсеры, энхансеры, граничные элементы/инсуляторы, организация промоторов для конститутивной или индуцибельной экспрессии, сигналы завершения синтеза транскрипта РНК-полимеразой и т.д.) и (2) сплайсинга (сигналы сплайсинга, регуляция альтернативного сплайсинга).

Таким образом, мы видим, что исследование транскриптома является одной из задач функциональной геномики в той мере, в какой информация, необходимая для формирования его структуры, закодирована в геноме и может быть выявлена при его изучении.

Слайд 3.

Еще раз коротко определим наш объект: Транскриптом – совокупность всех транскриптов всех генов, экспрессирующихся в какой-либо клетке, или группе однотипных клеток – т.е. ткани, или во всех клетках организма в определенные моменты функционирования и/или развития.

В соответствии с функциями самого транскриптома есть несколько задач транскриптомики - науки, изучающей транскриптом:

– исследование структуры транскриптов и изоформ транскриптов, образованных в процессах альтернативной транскрипции и альтернативного сплайсинга, транс-сплайсинга, РНК-редактирования и т.д.,
– исследование дифференциального временно́го и пространственного распределения транскриптов в клетках и организмах, сформированного в результате процессов их транскрипции, их транспорта из ядра, их транспорта и запасания в цитоплазме, их miRNA-опосредованной деградации и деградации, связанной с их трансляцией.

Применение методов биоинформатики на уровне транскриптома для исследования структуры транскриптов и их дифференциального временно́го и пространственного распределения в клетках и организмах позволяет:

– реконструировать коды, заключенные в геноме (кооперация с геномикой)

– выявлять информацию в виде сигналов и кодов, необходимую для формирования протеома (кооперация с протеомикой)

Слайд 4.

Применение методов биоинформатики к транскриптомным экспериментальным данным выдвигает ряд требований к этим данным и методам получения их.

Требования к методам исследования структуры транскриптов и их дифференциального временно́го и пространственного распределения в клетках и организмах:

- возможность измерения относительного и абсолютного содержания транскриптов определенного гена в клетках разных типов;
- возможность одновременного измерения соотношения транскриптов как можно большого количества генов;
- возможность детекции транскриптов очень слабо или очень специализированно экспрессирующихся генов (достаточно широкий динамический диапазон);
- высокая производительность и эффективность для производства достаточно большого массива данных.

Слайд 5.

Рассмотрим сначала методы исследования отдельных элементов транскриптома – транскриптов, а именно качественные и полуколичественные методы исследования структуры транскриптов и их дифференциального временного и пространственного распределения

Обзор методов исследования структуры и распределения транскриптов начнем с методов **прямой детекция молекул РНК**.

Метод Нозерн-блот-гибридизация является одним из самых первых, разработанных для этих целей, но он до сих пор не потерял своей актуальности и очень часто применяется. Название его не имеет никакого отношения к «северу», оно было придумано в шутку по аналогии с названием метода «Саузерн-блот-гибридизация», предложенного исследователем Саузерном, и затем прижилось.

На схеме изображен принцип метода:

- сначала образцы РНК (суммарной или очищенной поли(А)$^+$РНК) фракционируются в агарозном геле под действием электрического поля;
- затем фракции РНК переносятся (блоттинг) в ортогональном направлении (по сравнению с направлением фракционирования) на какую-либо мембрану (нитроцеллюлозную или нейлоновую) и иммобилизируются на ней.
- Наконец, проводится гибридизация со специфическими мечеными ДНК- или РНК-пробами (зондами). Именно процесс гибридизации между одноцепочечными полинуклеотидными молекулами позволяет в определенных условиях удержать в месте локализации на мембране определенных фракций РНК комплементарные меченые молекулы и тем самым выявить первые. Метка может быть радиоактивная или нерадиоактивная, это определяет методы выявления меченых молекул.

В результате получаются двумерные изображения в виде полос на дорожках, соответствующих длине транскриптов какого-либо гена. Используя различные специфические пробы можно выявить несколько наборов полос или паттернов для нескольких генов.

Слайд 6.

Нозерн-блот-гибридизация является стандартным методом для детекции отдельных мРНК и количественной оценки их относительного содержания в 5-20 образцах.

Достоинства этого метода: (1) это единственный метод, позволяющий определять размер транскриптов и выявлять изоформы, образованные в результате альтернативных процессов транскрипции и сплайсинга; (2) отражает реальное соотношение количеств мРНК и дает точные данные об относительном содержании разновидности РНК в образце.

Ограничения: (1) метод практически не дает возможности получить данные об абсолютном содержании разновидности РНК в образце; (2) малочувствительный, т.е. не позволяет выявлять транскрипты очень слабо или очень специализированно экспрессирующихся генов; (3) очень трудоемкий, долгий, и малопроизводительный.

Слайд 7.

Метод «**Анализ с помощью защиты от рибонуклеазы**» (Ribonuclease protection assay).

На схеме изображен принцип метода:

- сначала с образцами РНК проводят гибридизацию в растворе со специфической антисмысловой меченой РНК-пробой (необходимо отметить, что гибридизация нуклеиновых кислот в растворе на два порядка эффективнее, чем иммобилизованных);
- затем продукты гибридизации обрабатываются РНКазой, которая деградирует все одноцепочечные молекулы РНК, но оставляет в растворе РНК-дуплексы;
- оставшиеся РНК-дуплексы анализируются с помощью гель-электрофореза и гибридизации с меченными зондами

В результате получаются двумерные изображения в виде полос на дорожках, соответствующих длине дуплексов в соответствии с длиной использованной РНК-пробы. Используя различные специфические пробы можно выявлять продукты гибридизации одновременно для нескольких генов. А используя знание о точной концентрации пробы и титрование ее, можно получить данные об абсолютном содержании определенной разновидности транскрипта в образце.

Слайд 8.

Метод «Анализ с помощью защиты от рибонуклеазы» (Ribonuclease protection assay) пригоден для одновременной детекции 10-15-ти мРНК и количественной оценки их содержания в 5-20-ти образцах.

Достоинства: отражает реальное соотношение количеств мРНК, высокочувствительный; позволяет точно картировать 5’- и 3’- окончания транскриптов и экзон-интронные стыки.

Ограничения: метод еще более трудоемкий, долгий и малопроизводительный.

Слайд 9.

Теперь перейдем к более обширному классу методов детекция молекул РНК с применением **обратной транскрипции и ПЦР.**

Открытие процесса обратной транскрипции, в результате которого на РНК-матрице синтезируется ДНК-цепь, существенно изменило стратегию исследования транскриптома, т.к. позволило переносить информацию от нестабильных молекул РНК к более стабильным молекулам ДНК (с учетом последующего достраивания второй цепи), способным, во-первых, долговременно храниться и воспроизводиться с помощью клонирования в бактериальных клетках и, во-вторых, амплифицироваться до необходимых для детекции количеств (с помощью клонирования в бактериальных клетках или полимеразной цепной реакции – ПЦР).

Первый метод, который мы рассмотрим, - количественная ОТ-ПЦР (Обратная Транскрипция+Полимеразная Цепная Реакция) (quantitative RT-PCR, qRT-PCR).

На схеме изображен принцип этого метода:

- сначала с образцами РНК проводят реакцию обратной транскрипции, запускаемой с полиА-тракта с помощью олигоТ-праймеров (часто снабженными с 5’-конца специфическими короткими последовательностями, необходимыми для последующих клонирования или

амплификации), в результате чего получают первую цепь комплементарной ДНК (кДНК, cDNA);
- затем, используя первую цепь как матрицу и пару специфических праймеров, сайты для которых расположены где-либо в представляющем интерес транскрипте, проводят ПЦР, в результате чего амплифицируется фрагмент транскрипта до количеств, позволяющих детектировать его простым гель-электрофорезом;
- количественный характер данных достигается с помощью параллельного проведения всей процедуры с тестируемым транскриптом и неким маркерным/стандартным транскриптом, точная концентрация которого была определена ранее.

На электрофореграмме, приведенной внизу и справа схемы, видна кинетика накопления продуктов ПЦР некоего фрагмента транскрипта.

Существует много разновидностей этого метода, в основном направленных на достижение условий надежной сравнимости кинетики ПЦР для тестируемого и маркерного транскриптов и надежного преобразования получаемых экспериментальных данных в числовые для последующей математической обработки.

Слайд 10.

Метод количественной ОТ-ПЦР является самым чувствительным для детекции мРНК и количественной оценки их содержания в образцах.

Достоинства: высокая чувствительность; может выявлять несколько разных транскриптов одновременно; высокая производительность;

Ограничения метода полностью проистекают от особенностей ПЦР: (1) большой риск неспецифической амплификации, (2) риск амплификации оставшихся в образцах фрагментов геномной ДНК, (3) при неравномерности кинетик амплификации тестируемого и маркерного транскриптов или при неудачном выборе диапазона сравнения в нелинейной области оценки соотношений между разными продуктами реакции будут искажены, (4) количество одновременно выявляемых транскриптов ограничено.

Слайд 11.

Следующий метод - **Дифференциальный дисплей** (Differential display).

В ранее описанных методах, как можно было заметить, для выявления транскриптов и подсчета разницы в экспрессии каких-либо генов требовалось предварительное знание о структуре исследуемых транскриптов. Однако такое знание добывается в результате очень трудоемких и долгих экспериментальных процедур, что сильно сдерживало исследование дифференциальной экспрессии генов. Чтобы преодолеть это ограничение, был предложен метод дифференциального дисплея, позволяющего выявлять разницу в экспрессии неизвестных генов.

Суть этого метода изображена на схеме:
- сначала с образцами РНК проводят реакцию обратной транскрипции как и в других методах, в результате чего получают первую цепь кДНК;
- затем, используя первую цепь как матрицу и особые праймеры, один из которых является олигоТ плюс еще один нуклеотид A,C или G на 3’-конце (чтобы «заякорить» праймер именно на границу транскрипта и полиА-тракта), а другой – олигонуклеотидом случайного состава размером 10-12 н.о., проводят ПЦР в присутствии меченных нуклеотидтрифосфатов, в результате чего амплифицируются меченые небольшие фрагменты транскриптов многих генов, имеющие в 3’-области сайт для второго праймера;
- меченые продукты амплификации для сравниваемых образцов одновременно анализируются полиакриламидным денатурирующим гель-электрофорезом, что позволяет сразу обнаруживать разницу по присутствию/отсутствию какой-либо фракции или по ее интенсивностям между образцами;
- использование трех вариаций олигоТ-праймера с «якорями» и большого набора случайных олигонуклеотидов позволяет вовлечь в сравнение по представительству транскриптов довольно большое число генов;

- извлечение из геля отдельных заинтересовавших фракций позволяет клонировать и идентифицировать ген.

Слайд 12.

Метод дифференциального дисплея с 1992 стал самым распространенным для быстрого, точного и чувствительного выявления различий в относительном количестве множества видов мРНК во многих образцах.

Достоинства: высокая чувствительность; высокая производительность.

Ограничения Ограничения: анонимность выявляемых полос и необходимость дальнейшей трудоемкой идентификации полос; много фальшивых позитивных полос; можно одновременно анализировать не очень большое количество образцов.

Слайд 13.

Следующий метод - **вычитательная (или истощающая) гибридизация** (Subtractive Hybridization).

Этот метод скорее вспомогательный и применяется в сочетании с другими для сокращения затрат труда и материальных ресурсов в исследованиях транскриптома. Этот метод позволяет с помощью предварительной обработки образцов РНК удалять транскрипты генов «домашнего хозяйства», т.е. заведомо не дифференциально экспрессированные, но составляющие до 95% от всей массы мРНК.

На схеме изображен принцип этого метода:

- сначала из исследуемой ткани (целевой, target) выделяют РНК проводят реакцию обратной транскрипции как и в других методах, в результате чего получают первую цепь кДНК, затем вторую цепь, полученные дуплексы клонируют, т.е. встраивают в бактериальные векторы, совокупность полученных клонов образует целевую библиотеку ДНК;
- используя расположенные в векторе сайты, узнаваемые РНК-полимеразой Т3 фага, получают со всего пула клонированных кДНК целевой пул одноцепочечных антисмысловых фрагментов ДНК;
- из ткани маркерной (направляющей, driver), относительно которой будет проведено вычитание, также аналогично готовится направляющая библиотека ДНК;
- аналогично получают направляющий пул смысловых фрагментов ДНК, которые метятся с помощью внедрения биотинилированных нуклеотидов;
- затем оба пула объединяют и проводят гибридизацию в растворе;
- образовавшиеся дуплексы, соответствующие общим для двух тканей транскриптам (как правило, это транскрипты генов актинов, тубулинов, гистонов и т.д.) удаляют, используя свойство высокоаффинного связывания биотина со стрептавидином;
- оставшиеся фрагменты целевого пула используют или для создания истощенной библиотеки кДНК или для приготовления истощенной пробы для других экспериментов.

Слайд 14.

Теперь переходим к методам массового исследования транскриптов, т.е. высокопродуктивным и высокоинформативным методам исследования структуры транскриптов и их дифференциального временнòго и пространственного распределения.

Начнем с метода «**Серийный анализ экспрессии генов**» - Serial analysis of gene expression (SAGE). Этот мощный метод для глобального компьютерного анализа уровней экспрессии генов был предложен Victor E. Velculescu в 1995 (Velculescu et al. Science 270 (5235) : 484-487).

Слайд 15.

На этом слайде изображен принцип этого метода:

1. Синтез биотинилированной двухцепочечной кДНК. Из образца ткани выделяется РНК и готовится пул биотинилированных двухцепочечных кДНК, за счет использования меченного биотином олигоТ-праймера;

2. Рестрикция по *Nla*III сайтам и отделение самых 3'-крайних фрагментов. Сайт узнавания рестриктазы *Nla*III представляет собой тетрамер CATG, поэтому довольно часто встречается в транскриптах, в частности, в 3'-области транскриптов. С помощью магнитных частиц, покрытых стрептавидином, взаимодействующим с биотиновой меткой, продукты рестрикции отделяются от других молекул.

Слайд 16.

3. Присоединение специфичных адаптеров и вырезание меток, тагов (tag), длиной 10 п.о. с помощью рестриктазы IIs типа. Пул продуктов рестрикции делится на две равные части, к молекулам одного субпула к липким концам *Nla*III сайта присоединяется один адаптер, а к молекулам второго – другой адаптер. Затем оба субпула обрабатываются рестриктазой IIs типа, т.е. способными делать двухцепочечный разрез на некотором удалении от сайта узнавания, в данной случае – *Bsm*FI, делающий разрез на удалении 10 н.о.. Участок между сайтом узнавания *Bsm*FI и точкой разреза называется «меткой», «тагом» (tag). Остатки 3'-концов кДНК удаляются за биотиновую метку.

Слайд 17.

4. Образование дитагов, их амплификация и удаление адаптеров. Два субпула объединяются и обрабатываются лигазой, которая соединяет молекулы таги+адаптеры так, что таги примыкают друг к другу. Образуются молекулы дитаги+адаптеры, которые амплифицируются при использовании праймеров, специфичных к участкам адаптеров. Образовавшийся пул молекул снова обрабатывается рестриктазой *Nla*III, после чего продукты рестрикции разделяются и адаптеры удаляются.

Слайд 18.

5. Формирование конкатемеров и их секвенирование. Дитаги лигируются и образуют длинные ряды, которые затем клонируются, и индивидуальные клоны, наконец, тотально секвенируются. В полученных текстах таги выявляются по наличию сайтов *Nla*III, разнесенных на строго определенное расстояние.

Слайд 19.

6. Вычисление уровня экспрессии транскрипта исходя из числа встречаемости определенного тага. После секвенирования всех клонов проводят идентификацию тагов, соотнесение к транскриптам генов и их подсчет для каждого идентифицированного гена. Таким образом создается профиль экспрессии генов в тестируемом образце ткани.

Слайд 20.

Суммируем основные принципы метода SAGE:

1. Небольшой фрагмент из определенного места транскрипта достаточно информативен для идентификации этого транскрипта. (9 п.о. => 4^9=262 144 транскрипта);
2. Конкатенация и параллельный анализ повышают продуктивность (секвенирование 96 колоний за один прогон => 2400 прочитанных тагов);
3. Минимизация связанных с амплификацией искажений в соотношении тагов (ампликоны почти равны по размеру ~ 100 н.о. и составу, т.к. на 70% состоят из адаптеров)

Слайд 21.

Сформулируем, как мы условились, характеристики метода с точки зрения выдвинутых требований.

Достоинства:

- высокая продуктивность;
- возможность определить соотношение транскриптов тысяч генов;
- почти отсутствует наведенные амплификацией количественные искажения;
- возможность определять абсолютные значения уровней экспрессии генов;
- возможность сравнивать результаты разных экспериментов.
- возможность выявлять слабо экспрессирующиеся гены;
- возможность выявлять новые экспрессирующиеся гены;

Ограничения:

- зависимость от предварительного знания последовательности значительной части генов организма
- не выявление сплайсированных форм
- зависимость от качества синтеза первой цепи кДНК и от расстояния между сайтом рестрикции и полиА-трактом
- применимость только полиаденилированным транскриптам эукариот
- требует большое количество РНК и поэтому не позволяет профилировать экспрессию генов с высоким разрешением.

Слайд 22.

Теперь рассмотрим другой высокопродуктивный и высокоинформативный метод исследования структуры транскриптов и их дифференциального временно̀го и пространственного распределения.
Это метод, который я бы перевел как «прочитанные фрагменты экспрессированных последовательностей (Expressed Sequence Tags (ESTs)», а далее буду назвать его «ИСТ»-метод.

Слайд 23.

Это очень мощный метод для глобального компьютерного анализа структуры транскриптов, реконструкции структуры генов и измерения уровней экспрессии генов.

Принцип был предложен в статье Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, et al. 1991 Complementary DNA sequencing: expressed sequence tags and human genome project. Science, 252:1651-1656.

Количественное приложение разработано в статье Okubo K, Hori N, Matoba R, Niiyama T, Fukushima A, Kojima Y, Matsubara K. Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. Nature Genetics 1992, 2:173-179.

Слайд 24.

Определим, что такое, собственно, Expressed Sequence Tags (ESTs). Это короткие (обычно около 300-500 bp), прочитанные за один раз (single-pass reads) фрагменты кДНК.
Они представляют собой «слепок», «отпечаток» (snapshot) с продуктов гена, проэкспрессированных в определенной ткани или на определенной стадии развития. Они являются метками (tags) экспрессии гена для определенной библиотеки кДНК.

Принципы метода ESTs:

- анализ совокупности прочитанных фрагментов, ассоциированных с каким-либо геном, позволяет реконструировать структуру транскриптов гена, а после сравнения ее со структурой геномного района – структуру самого гена;
- в пределах определенным образом приготовленной библиотеки кДНК частотное распределение кДНК клонов в целом соответствует исходному распределению транскриптов в популяции мРНК, из которой приготовлена библиотека;
- многочисленность прочитанных фрагментов;
- относительная дешевизна прочтения фрагментов.

Слайд 25.

Рассмотрим схему этого метода.

РЕКОНСТРУКЦИЯ СТРУКТУРЫ ТРАНСКРИПТОВ ГЕНА.

А. Экспериментальная фаза А1. Приготовление кДНК библиотеки

Рассмотрим некий ген, с которого транскрибируется пре-мРНК. Направление транскрибирования от промотора до сигнала полиаденилирования задает стандартное положение 5' и 3'–концов. Пре-мРНК после сплайсирования интронов превращается в мРНК. По стандартной процедуре с использованием олигоТ-затравки, несущей адаптер Ad1 для клонирования, проводится приготовление первой цепи кДНК.

На схеме видно, что образовавшиеся в процессе обратной транскрипции молекулы отличаются от изначальных транскриптов. Источник отличий от исходных природных молекул:

- деградация мРНК с 5'-конца в процессе выделения РНК;
- случайный обрыв синтеза кДНК;
- внутреннее праймирование с любого полиА-богатого участка транскрипта.

Кроме этого в саму последовательность вносятся отличия – замены нуклеотидов в результате ошибок обратной транскриптазы.

Слайд 26.

Далее идет подготовка полученных кДНК к клонированию:

- присоединение второго адаптера Ad2;
- замещение РНК второй цепью кДНК;

Встраивая обработанные рестриктазами, сайты для которых расположены в адаптерах, двухцепочечные молекулы кДНК в специально разработанный вектор, получают кольцевые молекулы, способные бесконечно воспроизводиться в бактериальных клетках, – клоны кДНК. С помощью трансформации бактериальных клеток этими кольцевыми молекулами и рассева колоний, содержащих плазмиду, удается индивидуализировать клоны посредством их перемещения в матрично-организованные плашки. Собственно, это и есть библиотека клонов кДНК, как набор индивидуальных клонов. Получаются целые стопки таких плашек, содержащих каждая по 96 ячеек, а есть сейчас и 384-ячейный формат. Положение колонии, в плашке дает как бы координаты – ряд 3 и колонка 6 в 57 плашке библиотеки XX дадут идентификатор клона XX570306. Все операции с библиотеками осуществляются без участия человека с помощью специальных роботов. Это необходимо, во-первых, для поддержания стерильности, и, во-вторых, для минимизации ошибок при отслеживании происхождения образцов в последующих процедурах. Прямо специальным манипулятором с 96-тью стержнями (при 96-ячейном формате) проводится пересев клонов на свежую среду и отбор материала для выделения ДНК параллельно сразу из всего набора клонов.

На этих этапах также возможны внесения отличий от исходных природных молекул мРНК:

- образование артефактных химерных молекул - при недостаточной рестрикции адаптеров и их частичной деградации происходит лигирование по тупым концам молекул, происходящих от совершенно разных генов;
- некоторые ошибки в последовательность вносятся Taq полимеразой.

Слайд 27.

С помощью специфических праймеров, комплементарных к сайтам в векторе, окаймляющем встройку кДНК, проводят секвенирование ближайших к векторной оправе частей встройки. При секвенировании с 5'-фланка вектора получают 5'-EST, и, соответственно, с 3'-фланка вектора получают 3'-EST. Необходимо напомнить, что в методе не предусматривается секвенирование всей встройки и при этом качественно, т.к., как правило, получают многие и многие тысячи клонированных встроек. Напротив, принцип EST-метода – одноразовый сиквенс двух небольших участков встройки – в среднем 300-400, не больше 700 н.о., но зато массовый сиквенс!

И на этом этапе вносятся отличия от исходных природных молекул мРНК в виде ошибок Taq полимеразы или секвеназы. Кроме этого возможны еще ошибки информатического свойства – уже как ошибки оперирования данными. Это ошибки в обозначении идентификатора клона (id_clone

identification error), ошибки в обозначении направления секвенирования клона (5'/3' identification error), смещение дорожек при параллельном электрофорезе продуктов сиквенсной реакции (lane slipping), приводящее к возникновению «виртуальных» химер (сейчас использование капиллярных автоматических секвенаторов позволяет избежать этого).

Полученные короткие сиквенсы подаются в специальный подраздел GenBank – dbEST (или в EMBL bank) с сохранением информации об идентификаторе библиотеки, способе ее приготовления, идентификаторе клона, направлении секвенирования.

Слайд 28.

На этом слайде показано, как успешно развиваются EST-проекты для разных организмов. Видно, что лидирующим видом является человек разумный – более 7 миллионов EST! За ним следуют виды с большим биомедицинским значением, затем идут многие сельскохозяйственно-значимые объекты. В списке есть и объекты фундаментально-биологической значимости - *Danio rerio, Xenopus laevis, Drosophila melanogaster, Caenorhabditis elegans, Arabidopsis thaliana.* Как правило, EST-проекты являются составной частью геномных проектов, о которых я рассказывал на лекции, посвященной информационным ресурсам по анализу структуры и функции геномов эукариот. Однако эти объекты постепенно опускаются в списке. Еще три года назад *Drosophila melanogaster* была третьей по числу депонированных EST. Процесс перераспределения видов в списке отражает характер финансирования подобных проектов, все-таки очень дорогостоящих. Основными спонсорами EST-проектов являются мощные интернациональные фармацевтические и агротехнические фирмы. Между ними была достигнута договоренность помещать в GeneBank для публичного доступа все EST, полученные для своих исследуемых объектов, не позже чем через год, чтобы академическое сообщество также могло их использовать в своих исследованиях и предлагать свои биологически значимые интерпретации.

Слайд 29.

После получения EST для клонов библиотеки начинается применение методов собственно информатики, это компьютерная фаза EST-технологии. Сначала проводится кластеризация и выравнивание EST.

На схеме показана схема этих процедур. Для наглядности снова повторены структуры исходного абстрактного гена, идеальная схема его мРНК, набор реально выделенных транскриптов и синтезированных с них кДНК. А ниже показаны фрагменты, соответствующие просеквенированным частям клонированных кДНК. В процедуре кластеризации на основании результатов БЛАСТ-анализа, выявляющего все EST (как 3’-, так и 5’-), имеющие на достаточно длинном протяжении достаточно высокое сходство между собой и с мРНК какого-либо гена, формируется группа EST, для которой при процедуре выравнивания возможно построить общий консенсус. Теперь мы видим, что те недостатки, о которых страдали многие ранее описанные методы - деградация мРНК с 5’-конца в процессе выделения РНК, случайный обрыв синтеза кДНК, внутреннее праймирование с любого полиА-богатого участка транскрипта – в случае обилия EST превращаются в достоинства, поскольку позволяют «внедряться» вовнутрь мРНК, т.е. расшифровывать ее первичную последовательность без применения очень трудоемких и дорогих методов пошагового секвенирования со специфических праймеров или субклонирования. Напротив, применяются простые стандартные вектора, праймеры и требования к процедуре секвенирования. При это видно, как эффективно удаляются все внесенные на предыдущих этапах ошибки за счет взаимной компенсации. Остаются только ошибки, расположенные в местах, слабо «подкрепленных» разными EST. В целом получается, что частота ошибок, достигающая для EST 5%, снижается в несколько раз при генерировании консенсусов.

Слайд 30.

На предыдущем слайде был показан хороший случай, когда EST для какого-либо гена так много, что они счастливым образом распределились по всей длине мРНК этого гена. В случае неполного «перекрытия» всей длины некоей мРНК с помощью EST возникает брешь, особенно часто это происходит, если ген продуцирует очень длинные транскрипты, а экспрессия этого гена или низка или очень специфична. Если эта мРНК ранее уже была известна, то это не приводит к серьезным затруднениям в реконструкции структуры мРНК и гена. А если нет, то возникают проблемы – один

это ген или два? В большинстве случаев именно информация об идентификаторах клонов помогает разрешать этот вопрос. На схеме видно, что оба консенсуса как бы «связаны» наличием в своем составе EST с одинаковыми идентификаторами клонов. Такие консенсусы становятся LinkTS – связанными транскрибируемыми последовательностями. На основании информации о такой связанности формируется кластер консенсусов для какого-либо гена. В нем одному консенсусу, как правило, самому длинному или сформированному из самой многочисленной популяции EST, придается статус Reference TS, т.е. транскрибируемой последовательностью, с которой соотносятся все остальные.

Слайд 31.

Рассмотрим, как с помощью EST реконструируется структура изоформ транскриптов генов. Эти изоформы образуются в основном за счет двух процессов - альтернативной транскрипции и альтернативного сплайсинга.

Альтернативная транскрипция, как правило, затрагивает крайние экзоны и заключается в (1) использовании альтернативных стартов транскрипции, расположенных или внутри первого экзона или в начале нескольких первых экзонов, или (2) использовании альтернативных сигналов полиаденилирования, которые также могут быть или внутри последнего экзона или в конце нескольких последних экзонов.

Альтернативный сплайсинг затрагивает внутренние экзоны. Различают несколько типов:

- пропуск экзона (exon skipping);
- удлинение экзона в 3'-область (exon 3'-extention);
- удлинение экзона в 5'-область (exon 5'-extention);
- удержание интрона (intron retention) или несплайсированная незрелая форма (premature mRNA);
- альтернативное включение экзонов (alternative exon usage, casette exons)

Слайд 32.

Рассмотрим случай альтернативных стартов транскрипции в альтернативных 5’-экзонах. Мы видим, что в процессе кластеризации выравнивания генерируется два консенсуса TS1 и TS2. Причем в третьем экзоне они имеют один общий участок сходства. На этом основании они образуют кластер консенсусов.

С помощью различных программ визуализации результатов поиска сходства – например, NCBI "2 Seq BLAST" сервера - мы можем видеть, в каких местах в одной и в другой TS расположен участок сходства, какова доля этого участка сходства в той и другой TS, наконец, каково его расположение относительно несхожих участков. Анализ этой картины позволяет выдвинуть предположение, что эти две TS представляют собой изоформы транскриптов, образованных с двух стартов транскрипции в альтернативных 5’-экзонах.

Слайд 33.

В заключение я коротко покажу некоторые интернет-ресурсы, посвященные анализу структур транскриптов, выведенных из EST.

Это германская база данных “Database of alternative splice forms” по альтернативно сплайсированным изоформам транскриптов для генов девяти модельных организмов - *Arabidopsis thaliana, Bos taurus, Caenorhabditis elegans, Danio rerio, Drosophila melanogaster, Mus musculus, Rattus norvegicus, Xenopus laevi*s и *Homo sapiens* (http://medseq.bioinf.mdc-berlin.de/imap/splicelib/).

Ниже показана страница американской базы данных по альтернативно сплайсированным изоформам транскриптов для генов нескольких организмов – “Alternative splicing DB” (http://devnull.lbl.gov:8888/alt/).