Лекция «Исследование функциональных характеристик генома».

Первая часть.

Первая лекция посвящена определению терминов «функциональная геномика» и «транскриптом», выявлению места транскриптома среди других предметов исследования науки о функциональной организации генома, а также обзору современных методов получения качественных и количественных характеристик генома через исследование транскриптома.

Слайд 2. Функции генома

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

Функции генома состоят в том, чтобы содержать в виде кодов или сигналов генетическую информацию:

– для обеспечения сохранения этой информации в процессах жизнедеятельности одной клетки в течение одного клеточного цикла и в процессе пролиферации клеток при передаче через митоз (репликация, репарация, упаковка в хроматин, и т.д.);

– для обеспечения сохранения качества этой информации в ряду поколений особей при мейозе (сегрегация хромосом, рекомбинация);

– для обеспечения развертывания генетических программ развития, т.е. определять условия для формирования транскриптома, предопределенного генетическими программами развития для клеток определеных типов, т.е. обеспечение дифференциальной экспрессии генов в результате процессов транскрипции и сплайсинга.

Слайд 3. Задачи транскриптомики

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. Транскриптом – первый уровень фенотипа, т.е. первый уровень развертывания и реализации генетической информации, заключенной в геноме. Исследование транскриптома – одна из задач функциональной геномики и основная задача транскриптомики.

Задачи транскриптомики состоят в исследовании:

(1) структуры транскриптов и изоформ транскриптов, образованных в процессах альтернативной транскрипции и альтернативного сплайсинга, транс-сплайсинга, РНК-редактирования и т.д.,;

(2) их дифференциального временного и пространственного распределения в клеточных типах или тканях, 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 в транскриптомике позволяет:

- реконструировать структуру генома и коды, заключенные в геноме (взаимодействие с геномикой);
- выявлять информацию в виде сигналов и кодов, необходимых для формирования протеома (взаимодействие с протеомикой).

Слайд 4. Требования к методам

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

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

– возможность измерения относительного и абсолютного содержания транскриптов определенного гена в клетках разных типов, т.е. возможность сравнения результатов разных экспериментов при разных модификациях методов;

- возможность одновременного измерения для как можно большого количества генов соотношения транскриптов
- возможность детекции транскриптов очень слабо или очень специализированно экспрессирующихся генов (чувствительность и достаточно широкий динамический диапазон)
- высокая производительность и эффективность для производства достаточно большого массива данных.

Слайд 5. Характеристика методов

Характеристики методов исследования отдельных элементов транскриптома – транскриптов, а именно качественные и полуколичественные методы исследования структуры транскриптов и их дифференциального временного и пространственного распределения я обобщил для наглядности в виде таблицы. По каждому из требований для метода поставлена условная оценка близости к идеалу. Чем больше плюсов, тем метод лучше с точки зрения возможности производить цифровую информацию, пригодную для обработки методами биоинформатики.

Сначала идут «классические» методы прямой детекции молекул РНК с помощью молекулярной гибридизации ДНК-ДНК или РНК-ДНК. Это методы «нозерн-блот-гибридизация» и «Анализ с помощью защиты от рибонуклеазы» (Ribonuclease protection assay).

Затем представлен более обширный класс методов детекции «слепков» с молекул мРНК – молекул кДНК, полученных с применением обратной транскрипции. 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 – это количественная ОТ-ПЦР (Обратная Транскрипция+Полимеразная Цепная Реакция) (quantitative RT-PCR, qRT-PCR) с детекцией продуктов на элестрофореграмме.

Следующий метод - Дифференциальный дисплей (Differential display) также с детекцией продуктов на элестрофореграмме

Наконец – метод олигонуклеотидных или кДНКовых микробиочипов, в котором сочетаются ОТ-ПЦР для амплификации «слепков» с молекул мРНК и идентификация и соотнесение этих кДНК с генами с помощью молекулярной гибридизации. Этот метод является самым мощным из высокопродуктивныхи высокоинформативныхметодовсследования транскриптома.

Существуют и другие методы массового исследования транскриптома, т.е. высокопродуктивные и высокоинформативные. Эти методы, как правило, основаны на массовом производстве сиквенсов кДНК и их компьютерном анализе, выполняя как бы «in silico гибридизацию» транскриптов с последовательностями генов, представленных в геномных базах данных.

Слайд 6. Высокопродуктивные методы

Итак, в рамках настоящего цикла лекций мы будем рассматривать подробнее только последний класс методов для высокопродуктивных и высокоинформативных исследований структуры транскриптов и их дифференциального временного и пространственного распределения. Это методы:

- Прочитанные фрагменты экспрессированных последовательностей (Expressed Sequence Tags - ESTs);
- Серийный анализ экспрессии генов (Serial analysis of gene expression - SAGE);
- Массовое одновременное секвенирование идентифицирующих фрагментов (Massively Parallel Signature Sequencing - MPSS);
- Экспрессионные ДНК-биочипы (Expression microarrays, biochips).

Слайд 7. EST: история

Рассмотрим первый метод, название которого я перевел как «прочитанные фрагменты экспрессированных последовательностей (Expressed Sequence Tags (ESTs)», далее я буду назвать его «ИэСТи»-метод.

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 8. EST: определение и принципы

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.

Основные принципы метода ESTs:

- анализ совокупности прочитанных EST, ассоциированных с каким-либо геном, позволяет реконструировать структуру транскриптов гена, а после сравнения ее со структурой соответствующего геномного района – структуру самого гена – генную модель;
- для каждой библиотеки кДНК частотное распределение кДНК клонов в целом соответствует исходному распределению транскриптов в популяции мРНК, из которой была приготовлена библиотека;
- чтобы получить достоверную информацию о структуре генов и паттерне его экспрессии EST должно быть получено очень много;
- получение многочисленных EST должно быть относительно дешевым.

Слайд 9. EST: приготовление кДНК

Рассмотрим схему этого метода. Основные этапы – синтез кДНК, однонаправленное клонирование, создание библиотеки кДНК клонов, частичное секвенирование каждого клона с двух сторон – получение EST, кластеризация и выравнивание EST, генерирование консенсусов, реконструирующих структуру транскриптов.

Первый этап - реконструкция структуры транскриптов гена.

А. Экспериментальная фаза: 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 10. EST: кДНК библиотека

На втором этапе идет подготовка полученных кДНК к клонированию, собственно клонирование и трансформация бактериальных клеток.

При подготовке полученных кДНК к клонированию происходит:

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

При клонировании встраивание двухцепочечных молекул кДНК, обработанных рестриктазами, сайты для которых расположены в адаптерах, в специально разработанный вектор, дает 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 NN дадут идентификатор клона NN570306. 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 возможно возникновение 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 11. EST: секвенирование

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 12. EST: dbEST

На этом слайде показано, как успешно развиваются EST-проекты для разных организмов. Видно, что лидирующим видом является человек разумный – уже почти 8 миллионов EST, за один год прибавилось более 800000 новых сиквенсов! 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.* Однако эти объекты постепенно опускаются в списке. Еще четыре года назад *Drosophila melanogaster* была третьей по числу депонированных EST.

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. 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 13. EST: модель гена 1

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 14. EST: модель гена 2

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 15. Альт. сплайсинг

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 16. EST: альт.сплайсинг

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.

Слайд 17. EST: базы данных по AC

Для облегчения работы с выяснением структуры продуктов альтернативных тарнскрипции и сплайсинга, выведенных из анализа структуры EST, созданы специальные базы данных.

Одна из них 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/).

Слайд 18. EST: USCS-визуализация

На этом слайде показано, как с помощью санта-крузовской BLAT-визуализации, о которой я рассказывал в лекции о геномных интернет-ресурсах, можно быстро проанализировать структуру изоформ транскриптов какого-либо гена. Здесь представлены референсные последовательности транскриптов гена, набор полноразмерно «прочитанных» кДНК и набор EST, разнообразно начинающихся, заканчивающихся и составленных из экзонов.

Слайд 19. EST: вычит. гибридизация

Можно сразу заметить, что если стандартно применять EST-технологию, то исследователи очень скоро обнаружат, что основные усилия затрачиваются на выявление одних и тех же (и поэтому неинтересных) транскриптов от генов домашнего хозяйства – актинов, тубулинов, гистонов и т.д. И чтобы выявить действительно интересные различия между разными биологическими объектами, необходимы специальные экспериментальные приемы для подавления генерации данных про эти гены домашнего хозяйства.

Такие экспериментальные приемы заключаются в особенностях приготовления библиотек клонов кДНК, и эти особенности очень важны для применимости этого метода для получения количественной информации о дифференциальной экспрессии генов.

Первый особая процедура приготовления библиотек клонов кДНК – вычитательная (или истощающая) гибридизация (Subtractive Hybridization). На слайде приведена схема. Не буду вдаваться в подробности, подчеркну только главное – удаляя из образца бо̀льшую часть транскриптов от генов домашнего хозяйства, такая процедура позволяет выявлять очень редкие ткане- или стадия-специфичные транскрипты, но изменяет соотношение между транскриптами.

Слайд 20. EST: самоистощение

Другая процедура – нормирование библиотек клонов кДНК. Имеется в виду приведение количеств транскриптов к тому численному соотношению, в каком находятся гены в геноме.

Один из способов такого онрмирования - самоистощение (self-substraction). Т.е. чем больше какой-либо ген экспрессирует транскриптов, тем более вероятно они будут удалены из пула. Рассмотрим схему вариации метода предложенного в 1994 г. Soares M.B. *et al*., (Construction and characterization of a normalized cDNA library. Proc Natl Acad Sci U S A.;91(20):9228-32). Использование специальных адаптеров позволяет превращать в кольцо молекулы кДНК. Затем проводят частичный синтез одноцепочечной цепи ДНК с праймеров, узнающих сайты в векторе. Пул кольцевых молекул кДНК короткими синтезированными фрагментами отделяется от остальных, подвергается денатурации и новому отжигу-гибридизации. Все образовавшиеся частичные дуплексы, продукты перекрестной гибридизации, удаляются, а оставшиеся идут для создания нормированной библиотек клонов кДНК.

Слайд 21. EST: нормир. геномными посл-тями

Второй способ нормирования - истощение геномными последовательностями. Этот способ отличается от вышеприведенного тем, что для истощения используется иммобилизованная фрагментированная и денатурированная геномная ДНК. На графике внизу (где по оси абсцисс отложены разные гены по мере снижения их транскрипционной активности, а по оси ординат число транскриптов для каждого гена) показано, что эта процедура позволяет как бы срезать верхнюю часть распределения транскриптов. Этот прием также позволяет выявлять очень редкие транскрипты, но изменяет т соотношение между обильными и редкими транскриптами.

Слайд 22. EST: DDD

С учетом всех вышеизложенных модификаций в целом EST-метод позволяет с достаточной точностью получать цифровые данные о распределении транскриптов. Основанный на учете EST способ определения профиля экспрессии генов был назван цифровым дифференциальным дисплеем (Digital Differential Display). На слайде показан пример гистограммы распределения транскриптов, полученной одной из французских биоинформатических групп (http://igs-server.cnrs-mrs.fr). Например, они опубликовали статью о генах, проявляющих специфическую экспрессию в сердце (Mégy K, Audic S, Claverie JM. 2002; Heart-specific genes revealed by expressed sequence tag (EST) sampling. Genome Biol. 3(12): research0074.1-research0074.11.). Эти данные послужили для последующего исследования регуляторных районов этих генов-кандидатов и выявления специфических мотивов и их специфической организации.

Слайд 23. EST: NCBI-цифр. нозерн

Существует и другое название для методов компьютерной количественной обработки EST-данных - Электронный или цифровой нозерн. Изображения, которое вы видите на слайде, генерируются сервером UniGene, подразделением в системе GenBank, специализированном на компьютерном анализе EST. Существует много работ, эксплуатирующих EST-метод, и много интернет-ресурсов, посвященным сбору и анализу EST-данных.

Слайд 24. EST: allgenes-цифр. нозерн

На этом слайде показана страница для результатов "цифрового нозерна", предоставляемых сайтом www.allgenes.org для какого-нибудь гена.

Слайд 25. EST: интернет-ресурсы

В заключение я приведу некоторые интернет-ресурсы, посвященные анализу структуры и распределения транскриптов, выведенных из EST-данных.

На этом слайде показана главная страница сайта «The EST Machine» (http://www.tigem.it/ESTmachine.html). Пожалуй, это самая представительная коллекция ссылок по этой теме. Вы видите, сколько есть специализированных сайтов по разным организмам и разным биологическим и медицинским проблемам. Видно также, что все геномные проекты, о которых я рассказывал на лекции о геномных интернет-ресурсах, имеют подразделения, занятые приготовлением и секвенированием EST.

Также приведу список сайтов наиболее важных EST-организаций:

http://www.ncbi.nlm.nih.gov/dbEST
http://www.tigem.it/ESTmachine.html
http://cgap.nci.nih.gov/
http://industry.ebi.ac.uk/~muilu/EST/EST_links.html
http://image.llnl.gov/
http://www.cs.jhu.edu/labs/compbio/morgan.html

Слайд 26. SAGE: история и принципы

Следующий метод «**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).

Основные принципы этого метода:

1. Небольшой фрагмент из определенного места транскрипта достаточно информативен для идентификации этого транскрипта. (Информационная ёмкостьметода: 10 п.о. => 410=1 048 576 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 п.о. и составу – на 70% состоят из адаптеров)

Слайд 27. SAGE: схема

На слайде изображена общая схема метода SAGE. Его основные этапы – выделение тагов, соединение их в протяженные ряды и их массовое секвенирование. Полученные тексты/сигнатуры обрабатывают компьютерными методами и с помощью подсчета одинаковых сигнатур получают в цифровом виде количественные данные о распределении транскриптов многих тысяч генов для разных образцов, например, из нормальной и видоизмененной болезнью тканей.

Слайд 28. SAGE: стадии 1 и 2

Рассмотрим подробнее процедуру получения SAGE-данных. На этом слайде изображены первые стадии 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 29. SAGE: стадия 3

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 30. SAGE: стадия 4

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 31. SAGE: стадия 5

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 32. SAGE: вычисление

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 33. SAGE: достоинства и ограничения

Сформулируем 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
- невыявление сплайсированных и других изоформ
- 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 34. SAGE: пример

В качестве примера представлено исследование, проведенное разработчиком метода SAGE – Виктором Вескулеску (Velculescu V.E. et al., 1997 Analysis of Yeast Transcriptome // Cell 88: 243-251).

В статье описан анализ транскриптов из дрожжевых клеток трех состояний (лог-фаза, задержанные в S- и G2/M- фазах). Всего было проанализировано 60,633 тагов, при этом было выявлено 4,665 генов с уровнями экспрессии от 0.3 до более 200 транскриптов на клетку.

Объем выборки тагов, которые необходимо проанализировать, вычислялся исходя из требований достижения определенного уровня достоверности выявления редких мРНК. Отталкиваясь от полученной ранее оценки, что на клетку приходится 15,000 молекул мРНК, было определено, что секвенирование 20,000 тагов должно дать 72% вероятности обнаружить одну молекулу мРНК на клетку.

Авторы объединили свои данные с данными о расположении генов в хромосомах *Saccharomyces cerevisiae*, что позволило построить хромосомные карты экспрессии с районами транскрипционной активности и выявить гены, существование которых было невозможно предсказать, исходя из знания только о последовательности геномного района.

Слайд 35. SAGE: развитие

В последнее время предпринимаются большие усилия по преодолению основного ограничения метода SAGE – его малой информационной ёмкости. Действительно, для высших эукариот SAGE применен только для сравнения тканевых образцов, т.к. у метода есть серьезное ограничение – количество надежно идентифицирумых транскриптов не настолько велико (1 048 576 транскриптов в идеале), чтобы анализировать неизмеримо более сложный транскриптом высших эукариот.

Однако, в последнее время появились модификации метода, когда используются рестриктазы, делающие разрезы на удалении 21 нуклеотида от сайта узнавания. Это значительно повышает точность идентификации транскриптов в более мощном транскриптоме. Этот метод назван LongSAGE.

Также пытаются преодолеть ограничения из-за привязки SAGE-тагов к 3’ области транскриптов - предложены модификации метода и для тагов в 5’ области. Другое направление для развития – приспособление метода к малым образцам биологического материала для решения более тонких биологических задач.

Слайд 36. SAGE: интернет-ресурсы 1

В заключение части лекции, посвященной методу SAGE, пройдемся по интернет-ресурсам для этого метода. На слайде показана главная страница сайта SageNet. С этой страницы можно перейти на многие другие подразделы сайта.

Слайд 37. SAGE: интернет-ресурсы 2

На сайте NCBI есть также специальные разделы, посвященные данным SAGE-анализа - SAGEmap (http://www.ncbi.nlm.nih.gov/SAGE/).

Слайд 38. MPSS: история и принципы

Теперь рассмотрим третий метод высокопродуктивного и высокоинформативного исследования структуры транскриптов и их дифференциального временного и пространственного распределения – «Массовое одновременное секвенирование идентифицирующих фрагментов» (Massively Parallel Signature Sequencing - MPSS).

Этот метод появился совсем недавно – первая публикация появилась в 2000 году. (Brenner S, Johnson M, Bridgham J, et al., (Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotechnol. 2000; 18(6):630-634.).

Этот метод позволяет идентифицировать почти все ДНК-фрагменты в образце. Основные характеристики этого метода состоят в том, что он:

- Продуцирует количественный профиль экспрессии генов на основе подсчета всех мРНК в образце, и эти цифровые данные удобны для создания реляционных баз данных
- Обладает высокой чувствительностью, способностью детектировать редкие транскрипты. В каждый MPSS-анализ вовлечены более миллиона транскриптов, что обеспечивает исключительный динамический диапазон: от менее чем 10 транскриптов на миллион до 50,000 транскриптов на миллион.
- Не нуждается в предварительных молекулярно-генетических знаниях об организме

Слайд 39. MPSS: клонирование и амплификация

Рассмотрим схему выполнения этой технологии по материалам статьи: Brenner S, Johnson M, Bridgham J, et al., Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotechnol. 2000; 18(6):630-634..

Сначала готовят кДНК из мРНК и клонируют их в вектор, содержащий набор из 1.67 x 107 разных 32-мерных олигонуклеотидных меток. Пул кДНК матриц, представляющих 3-4 x 104 разных транскриптов, после клонирования формирует пул 5-7 x 1011 конъюгатов.

Далее берется образец, включающий только 1% (~1.6 x 105) от общего числа меток. Каждый транскрипт, таким образом, конъюгирован с уникальной меткой. Даже единичный транскрипт представлен с вероятностью более 99%.

Образец из конъюгатов амплифицируется в ПЦР, метки делают одноцепочечными. Конъюгаты гибридизуются с популяцией микрошариков, к которым присоединены все комплементы меток. 1% микрошариков будет связан с конъюгатами. Только эти будут отсортированы с помощью FACS (Fluorescence-Activated Cell Sorter).

Слайд 40. MPSS: лигирование - идентификация

Определение последовательности на основе лигирования и применения эндонуклеазы рестрикции второго типа (BbvI). Смесь адаптеров, включающих все возможные четырехбуквенные выступы, отжигается к целевой последовательности так, чтобы лигировался только тот, который имеет совершенный комплементарный выступ. Каждый из этих 256 адаптеров имеет уникальный мотив, Fn, который может быть обнаружен после лигирования.

Например, на рисунке изображена последовательность выступа матрицы, взаимодействующая с адаптером, маркированным F126, что означает, что последовательность выступа матрицы "TTAC".

Следующий цикл начинается с расщепления BbvI, чтобы высвободить следующие четыре основания матрицы.

В составе кодирующих адаптеров предусмотрены декодирующие мотивы.

Слайд 41. MPSS: схема идентификации

Использование кодирующих адаптеров, чтобы идентифицировать четыре основания в каждом цикле рестрикции-лигирования. После того как микрошарики, загруженные флуоресцентно помеченными (F) кДНК, изолированы с помощью FACS, кДНК разрезаются рестриктазой DpnII, чтобы высвободить четыре основания в виде выступа. Этот выступ достраивается до выступа с тремя основаниями.

Флуоресцентно помеченные (F) инициирующие адаптеры, содержащие BbvI-сайты, лигируются к кДНК в отдельной реакции, после чего микрошарики загружаются в специальные ячейки.

Затем кДНК разрезаются BbvI, кодирующие адаптеры, прогибридизовавшиеся с ними, лигируются. Шестнадцать помеченных фикоэритрином (PE) зонда по отдельности гибридизуются с декодирующими мотивами кодирующих адаптеров. После каждой гибридизации ячейка сканируется, изображение рядов микрошариков анализируется для идентификации оснований выступов.

Кодирующие адаптеры снова обрабатываются BbvI, который расщепляет кДНК и высвобождает четыре новых основания для следующего цикла лигирования и расщепления.

Слайд 42. MPSS: камера с рядами микробусин

На этом слайде показано, как выглядит проточная ячейка с микрошариками и как она используется.

Слева сверху – продольный разрез ячейки. Она сделана так, чтобы удерживать плоские ряды из микрошариков. Загрузка шариков и реагентов идет через входное отверстие. Шарики задерживаются специальным барьером, а растворы проходят насквозь. Небольшие перемещения шариков, зафиксированные сканером, учитываются при анализе изображений.

Слайд 43. MPSS: система детекции

Здесь представлена общая схема устройства для проведения MPSS.

Проточная ячейка монтируется на терморегулируемом блоке Пельтье в конфокальном микроскопе, снабженном ксеноновой лампой, фильтрами и CCD-камерой.

Для сканирования изображения ряды микрошариков разделены на 18 секций с 62,000 шариков в каждой. Специальное программное обеспечение обрабатывает изображения, выделяя центр каждого микрошарика и суммируя флуоресцентный сигнал с определенных пикселов изображения.

Как правило, удается получить надежные данные о последовательности 16-20 нуклеотидов кДНК-матрицы. Далее начинается сказываться «шум» из-за ошибок рестрикции и лигирования.

Полученные 16-20-нуклеотидные мотивы потом преобразуются в сигнатуры с помощью специальных программ.

Слайд 44. MPSS: расшифровка

Так выглядит псевдоцветное компьютерное изображение некоторого количества микрошариков и изображения гистограмм значений флуоресцентного сигнала и расшифрованной последовательности некоего микрошарика.

Слайд 45. MPSS: интернет ресурсы 1

Для MPSS-проектов созданы специализированные интернет-ресурсы, как правило, в рамках геномных проектов. На слайде приведена страница для ссылок на MPSS-проекты для нескольких растений.

Слайд 46. MPSS: интернет ресурсы 2

А на этом слайде показана страница MPSS-проекта для биомедицинских исследований на человеке.

Таким образом, данных о структуре транскриптов и их дифференциальном временном и пространственном распределении, полученных за посление 10 лет и с помощью разных высокопродуктивных и высокоинформативных методов, более чем достаточно. Тем самым созданы очень благоприятные возможности для решения очень сложных проблем функциональной геномики – выявления условий или кодов и сигналов для формирования транскриптома в результате дифференциальной транскрипции генов и сплайсинга транскриптов.