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

Вторая лекция посвящена краткому обобщению выводов по сопоставлению характеристик методов, изложенных в прошлой лекции и обзору примеров приложения высокопродуктивных и компьютероемких методов получения качественных и количественных данных о транскриптоме – SAGE и EST.

Слайд 2.

Еще раз повторим те требования, которые мы как биоинформатики выдвигаем к методам исследования транскриптома, а именно дифференциального временного и пространственного распределения транскриптов 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 измерения соотношения транскриптов как можно большого количества генов;
- 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 3.

Характеристики методов, которые мы рассмотрели в прошлой лекции, я обобщил для наглядности в виде таблицы. По каждому из требований методу поставлена условная оценка близости к идеалу. Чем больше плюсов, тем метод лучше с точки зрения возможности производить цифровую информацию, пригодную для обработки методами биоинформатики.

Восклицательные знаки в строках для количественной ОТ-ПЦР и дифференциального дисплея означает предостережение по поводу нелинейных искажений, которые ПЦР-основанные методы могут вносить в исчисление соотношений между транскриптами.

Восклицательный знак в строке для EST-метода означает предостережение по поводу искажений в исчислении соотношений между транскриптами, вносимых модификациями приготовления библиотеки клонов кДНК. Об этом я расскажу позже.

Поскольку мы видим наибольшее число плюсов в строках для методов SAGE и EST, мы еще раз рассмотрим их основные характеристики.

Слайд 4.

На слайде изображен принцип метода “Серийный анализ экспрессии генов” (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 5.

Рассмотрим примеры исследований, выполненных этим методом. В статье Buckhaults P. *et al*., (2001) (Secreted and Cell Surface Genes Expressed in Benign and Malignant Colorectal Tumors // Cancer Res; 61(19):6996-7001) описано применение SAGE-анализа для выявления транскриптов, кодирующих секретируемые или клеточно-поверхностные белки в добро- и злокачественных опухолях колоректума. Всего было проанализировано 290,394 тагов из разных образцов и выявлено 957 транскриптов, дифференциально экспрессируемых в норме и аденомах или в норме и раке.

Свои результаты подсчета тагов авторы оформили в виде графика. На панели А показано распределение значений кратности изменений для достоверно ($P < 0.05$) дифференциально экспрессируемых транскриптов. Отношение выражено в логарифмической шкале. Панель В - увеличенный фрагмент графика А. Поле графика делится на четыре части. Выше диагонали – достоверные значения, полученные при сравнении злокачественных опухолей и нормальных образцов, ниже диагонали – достоверные значения, полученные при сравнении доброкачественных опухолей и нормальных образцов. Видны области рассогласованных значений, а также области согласованной между добро- и злокачественными опухолями повышенной и пониженной экспрессии генов. На основании этой информации выделены гены кандидаты, проявляющие дифференциальную экспрессию в процессе перерождения ткани. Экспрессия транскриптов выявленных генов-кандидатов проверена количественной ОТ-ПЦР. Это важный момент, т.к. считается, что высокопродуктивные массовые методы производят необходимые свидетельства, но не достаточные. Условие достижения необходимых и достаточных доказательств для выводов выполняется, только если проводится исследование экспрессии генов—кандидатов с помощью более надежных и точных методов (или более привычных как заслуживающие доверия для основной массы исследователей) - ОТ-ПЦР, Нозерн-блот-гибридизация и т.д.. Собственно именно поэтому я в прошлой лекции уделил им так много внимания, что они до сих пор считаются неотемлемой частью стратегии исследования дифференциальной экспрессии генов.

Слайд 6.

Второй пример – 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 7.

Теперь немного об интернет-ресурсах, созданных для исследователей, интересующихся методом SAGE и результатами его применения. На слайде показана главная страница сайта http://www.sagenet.org/.

С этой страницы можно перейти на многие другие подразделы сайта.

Слайд 8.

Например, мы можем увидеть, сколько тагов исследовано для каких организмов или образцов из разных тканей или стадий развития. Мы видим, что для высших эукариот SAGE применен только для сравнения тканевых образцов, т.к. у метода есть серьезное ограничение – количество надежно идентифицирумых транскриптов не настолько велико (4^9=262 144 транскрипта 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.

Слайд 9.

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

Есть еще веб-сайт, рекламирующий адаптацию метода для анализа малых образцов. Дело в том, что метод SAGE вообще очень “прожорлив” по отношению к РНК. Это также сдерживает его применение для тонкого сравнительного анализа образцов. Так вот, это обстоятельство побуждает исследователей модифицировать метод SAGE, например, предлагать модификацию SADE или Serial microanalysis (http://www-dsv.cea.fr/thema/get/sade.html)

Слайд 10.

Теперь снова обратимся к методу анализа прочитанных фрагментов экспрессированных последовательностей, или EST-методу. Рассмотрим принципиальную схему. 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.

Слайд 11.

Теперь рассмотрим особенности приготовления библиотек клонов кДНК, поскольку, как я отмечал в таблице восклицательным знаком, эти особенности очень важны для применения этого метода для получения количественной информации о дифференциальной экспрессии генов.

Первая особенность приготовления библиотек клонов кДНК – уже знакомая нам 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 12.

Следующая особенность – нормирование библиотек клонов кДНК. Имеется в виду приведение количеств транскриптов к тому соотношению, в каком находятся количества генов в геноме.

Один из способов - 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 13.

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 14.

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 15.

Существует и другое название для методов компьютерной 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 16.

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

Слайд 17.

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.

Слайд 18.

Как же происходит компьютерная обработка информации, заключенной в EST? Рассмотрим схему. Мы видим, что от экспериментального модуля поступают первичные, грубые данные (raw data) в компьютерный модуль. Здесь происходит фильтрация и очистка материала от остатков последовательностей вектора, с помощью программ RepeatMasker - от последовательностей транспозабельных элементов или диспергированных повторов. Затем уже проводится БЛАСТ-анализ, по результатам которого формируются кластеры EST, которые соотносятся с генами (аннотация кластеров). К структурной аннотация затем добавляется функциональная аннотация, выявляются несоответствия, которые или подтверждаются как новая информация или служат основанием для переделки кластеров и усовершенствования алгоритмов.

Слайд 19.

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

http://www.tigem.it/ESTmachine.html
http://cgap.nci.nih.gov/
http://industry.ebi.ac.uk/~muilu/EST/EST_links.html
http://www.ncbi.nlm.nih.gov/dbEST/how_to_submit.html
http://image.llnl.gov/

http://www.cs.jhu.edu/labs/compbio/morgan.html

Слайд 20.

Теперь рассмотрим относительно новый, но очень перспективный метод количественного исследования транскриптома – метод ДНК-биочипов или ДНК микроматриц.

Слайд 21.

В этих лекциях мы рассматриваем только ДНК-биочипы, но существуют похожие по принципу белковые биочипы (иммобилизированные микроматрицы белков, антител, белковых антигенов) для исследований в области протеомики и интерактомики; клеточные биочипы (иммобилизированные микроматрицы клеток разных типов или штаммов микроорганизмов) и т.д..

Итак, ДНК-биочипы – это миниатюризированные матрицы или подложки, на которых в определенном порядке распределены фрагменты ДНК, соответствующие отдельным генам или их частям. Такие организованные микроматрицы позволяют проводить эксперименты по одновременному анализу структуры и экспрессии тысяч генов с помощью параллельной гибридизации.

Высокий уровень методов преобразования результатов этих экспериментов в цифровые данные и методов компьютерной обработки последних обеспечивает возможность анализировать и сопоставлять экспрессию таких массивов генов во множестве экспериментальных условий.

Таким образом получается статическая информация об экспрессии генов (в какой ткани или типе клеток, на какой стадии, при каком воздействии и т.д.) и динамическая информация об экспрессии генов при сопоставлении данных отдельных экспериментов.

Слайд 22.

Рассмотрим, какие бывают типы ДНК-биочипов. Во-первых, различают ДНК-биочипы в зависимости от структуры подложки: наиболее растпространены ДНК-биочипы на поверхности стекла или реже – полимера, существует также метод иммобилизации фрагментов ДНК в объеме небольших каплях или блоков геля – этот метод был разработан российскими учеными во главе с академиком А.Д. Мирзабековым, но из-за ограниченности результатов, полученных этим методом, я только упомяну о нем, но подробно останавливаться не буду.

Во-вторых, различают типы ДНК-биочипов в зависимости от природы используемых фрагментов ДНК: для производства олигонуклетидных биочипов используют химически синтезированные одноцепочечные олигонуклетиды длиной 20-75 н.о.; а для кДНК-биочипов – размноженные или в бактериальных клетках или с помощью ПЦР двухцепочечные клоны из библиотек кДНК длиной 100-2500 н.о..

Слайд 22.

Прежде чем приступить к подробному описанию отдельных вариаций метода ДНК-биочипы, рассмотрим общую схему, как этот метод, независимо от этих вариаций, применяется для исследования транскриптома на примере кДНК-биочипов.

Первый этап - приготовление биочипов. В автоматическом роботизированном режиме клоны их библиотеки кДНК нарабатываются до необходимых количеств и распределяются на поверхности стекла специальными аппаратами, называемыми arrayer или spotter.

Второй этап - приготовление образца (sample) для гибридизации. Из двух источников клеток – тестовых (или испытываемых) и референсных (или эталонов для сравнения), это могут быть клетки двух разных тканей, или клетки одной ткани в двух разных состояниях и т.д., выделяются пулы мРНК, которые обратно транскрибируются в пулы кДНК. В каждый пул кДНК вводится своя флуоресцентная метка, например, в тестовый пул красного цвета, а в референсный – зеленого. Затем оба пула объединяют и получают образец для гибридизации.

Условия гибридизации почти не отличаются от тех, что используются в методах Саузерн- или нозерн- гибридизации, только проводят ее в очень малых объемах – не больше полумиллилитра – и

поэтому в специальных камерах или устройствах. Меченные молекулы образца гибридизуются с пробами или мишенями (targets), распределенными на биочипе. Затем проводят отмывку неспецифически гибридизовавшихся молекул.

Затем проводят регистрацию сигналов гибридизации с помощью сканирования. В специальном устройстве (scanner или reader) лазеры генерируют излучение света определенной длины волны, в ответ на которое меченные молекулы образца, удержанные молекулами проб, испускают флуоресцентное свечение также определенной длины волны, которое регистрируется в цифровом виде. В результате получают изображение, снимок поверхности стекла. Необходимо подчеркнуть, что регистрация проводится отдельно по каждой длине волны, т.е. каналу, поэтому изображение получается в черно-белой шкале, отражающей интенсивность флуоресцентного ответа. Затем происходит обработка изображений с помощью компьютера, например, псевдораскрашивание изображений в красный и зеленый цвет, суперпозиция изображений, позволяющая наглядно проявить дифференциальное содержание тех или иных молекул в тестовом или референсном пулах кДНК, выделенных из образцов клеток.

Слайд 24.

Продолжим теперь классификацию разновидностей ДНК-биочипов. Типизацию по содержанию мы уже рассмотрели. Теперь рассмотрим типизацию по способу их приготовления. Олигонуклетидные биочипы могут быть приготовлены или способом синтеза олигонуклеотидов in situ, или распечатаны способом, применяемым в принтерах, обычно струйных. кДНК-биочипы, как правило, изготавливаются методом печати и обычно контактной, хотя встречается и струйная, бесконтактная.

Рассмотрим на схемах способы приготовления ДНК-биочипов.

Первый метод – фотолитография, или фотоиндуцируемый синтеза олигонуклеотидов in situ.

Второй - механическое раскапывание или контактная, матричная печать.

Третий – бесконтактная печать по методу ink jet принтеров.

Видно, что в первом случае пробы распределены в виде квадратных пятен, а в двух последних – в виде круглых.

Слайд 25.

Так выглядит стеклянный ДНК-биочип, в данном случае, судя по круглым пятнам, - приготовленный методом печати. Видно, какую маленькую область на предметном стекле занимают ряды проб. Видно также, что морфология пятен проб слегка различается – это один из источников статистического разброса в результатах. А после гибридизации сканированное изображение показывает ряды из реплицированных проб, т.к. таким образом этот разброс в какой-то мере преодолевается.

Слайд 26.

На этом слайде проиллюстрированы различные устройства для приготовления ДНК-биочипов – эррейеры.

Слайд 27.

По поводу терминов, используемых в области ДНК-биочиповых экспериментов для обозначения того, что с чем гибридизуется, до сих пор есть разнобой.

Некоторые авторы заявляют, что они склонны придерживаться принципов, установившихся для классических методов Саузерн- и нозерн-блот-гибридизаций: иммобилизованный партнер гибридизации – это мишень или образец, а растворенный партнер гибридизации – это проба или зонд, набор конкретных молекул, с помощью которых характеризуют мишень или образец. Однако, поскольку в методе ДНК-биочипов ситуация прямо противоположная классической: ведь иммобилизованными становятся именно наборы конкретных молекул, то большинство авторов считают, что термины проба или зонд должны относится к этим молекулам, т.е. факт их предварительной охарактеризованности важнее, чем характер их физического состояния с

эксперименте. Соответственно, термины мишень или образец продолжают относится к совокупности неидентифицированных молекул, которые надо охарактеризовать, хотя они и в виде раствора.

Таким образом, при чтении статей с применением метода ДНК-биочипов, следует сразу разобраться с тем, как авторы используют эти термины.

Теперь продолжим характеристику типов ДНК-биочипов с использованием введенных терминов:

- кДНК-биочипы, как правило, состоят из 40000 кДНК проб длиной 600-2400 н.п..
- Олигонуклетидные биочипы высокой плотности (high-density oligonucleotide arrays) могут содержать до 500,000 пар проб на одном стекле, причем одному гену соответствуют 11-20 проб. Для этого типа ДНК-биочипов есть еще специализированные термины: каждая пара проб состоит из «пробы совершенного сходства» (perfect match (PM) probe) и «пробы нарушенного сходства» (mismatch (MM) probe), в которой посередине введена однонуклеотидная замена (A на G или C на T).

Слайд 28.

Теперь перейдем к стадии приготовления меченного образца или мишени.

Каждый пул кДНК, приготовленный из двух разных образцов, делится на две равные части молекул, затем в молекулы вводятся меченые флуоресцентными группировками нуклеотиды. Это прямое включение. Есть методы и непрямого включения, когда вводятся нуклеотиды, модифицированные активными группировками, к которым затем химически присоединяют флуорохромы. Так поступают для того, чтобы избежать ингибирующего влияния на реакцию полимеризации слишком громоздких флуорохромных группировок или дифференциального влияния разных флуорохромных группировок.

Наиболее распространено использование зеленого цианина3 (Cyanine3 (Cy3)) и красного цианина5 (Cyanine5 (Cy5).

Как правило, получения данных о дифференциальной экспресии генов используют двухцветную совместную конкурентную гибридизацию, когда оба меченых образца объединяются.

Слайд 29.

Теперь перейдем к стадии сканирования результатов гибридизации и получение изображений.

На слайде показаны результаты сканирования нескольких ДНК-биочипов с одинаковым содержанием, но изготовленных по разным технологиям и на разных аппаратах. Видно, насколько разными получаются изображения в результате сканирования результатов гибридизации: на одних картинках большой фон, на других - неоднородный фон, на третьих – неровные пятна и т.д.. Собственно на этом этапе видно суммирование малозаметных случайных технических неравномерностей процессов на всех предшествующих этапах, приводящее к значительному статистическому разбросу данных при проведении ДНК-биочиповых экспериментов.

Слайд 30.

Рассмотрим в общем виде источники разброса характеристик сканированных изображений.

Систематические ошибки происходят от различий в:

- количествах образцов
- эффективности выделения РНК
- эффективности обратной транскрипции
- введения метки
- эффективности детекции сигнала.

Т.е. эти факторы имеют сходный эффект на многие измерения и поддаются коррекции на основе анализа данных для калибровки отдельных этапов технологического процесса ДНК-биочипового эксперимента.

Стохастические ошибки связаны с различиями в:

- успешности ПЦР и качестве ДНК в пробах
- эффективности раскапывания/печати, отражающейся на размере пятен и их морфологии,

а также в присутствии в какой-то мере кросс-гибридизации и неспецифическая гибридизации.

Эти факторы случайны и не учитываемы, представляют собой естественный «шум» и фон, и поддаются коррекции с помощью моделирования ошибки.

Слайд 31.

Рассмотрим теперь в общем виде, как же получаются данные биочип-экспериментов. После сканирования изображений результатов гибридизаций получают сначала матрицы измерений интенсивностей сигналов гибридизаций для набора экспериментов. Это первичные данные, генерируемые сканерами. Затем после процедур математической и статистической обработок получают матрицы данных по экспрессии генов, в которой сведены данные какого-либо набора экспериментов.

Слайд 32.

Итак, рассмотрим этап обработки изображений, с которого начинается анализ цифровой информации.

Исходные данные, получаемые после сканирования по каждому каналу, - это, как вы помните, псевдоаналоговое изображения определенного района поверхности стекла в черно-белой шкале. Обычно это 16-битные TIFF (Tagged Information File Format) изображения.

Они преобразовываются в цифровые данные об интенсивности сигнала гибридизации после:

- определения центра пробы (регистрации)
- выделение пикселей картины, относящихся к пробе и не-пробе (сегментация)
- определение значений интенсивности сигнала от пробы (как суммированной величины значений для пикселей каждого сегмента пробы) и определение значений фона (как суммированной величины значений для пикселей каждого сегмента не-пробы) (квантификация).

Слайд 33.

Необходимо еще раз подчекнуть, что идейной основой исследований транскриптома методом ДНК-биочипов является предположение, что измеренные в биочипе интенсивности для каждого гена отражают их относительный уровень экспрессии.

Поэтому снова отметим, что после сопоставления данных об интенсивностях сигналов между пробами на одном биочипе получается статическая информация о дифференциальной экспрессии генов (в какой ткани или типе клеток, на какой стадии, при каком воздействии и т.д.).

А после сопоставления данных об интенсивностях сигналов между теми же пробами, полученными в результате отдельных гибридизационных экспериментов, получается динамическая информация об экспрессии генов.

Слайд 34.

Экспрессию гена принято выражать как отношение между интенсивностью сигнала от исследуемого гена, выявленного, например, Cy5-меченным образцом, и интенсивностью сигнала от контрольного гена, выявленного, соответственно, Cy3-меченным образцом в условиях двухцветной совместной конкурентной гибридизации.

$T_i=R_i/G_i$.

Тогда гены с повышенной в два раза экспрессией будут иметь отношение 2, а с пониженной в два раза – 0.5. Такая форма учета дифференциальной экспрессии генов неудачна из-за несимметричности результатов.

Поэтому применяют преобразование вышеприведенного отношения в виде логарифма по основанию 2 :

$T_i= \log_2(R_i/G_i)$.

Таким образом дифференциальной экспрессии генов учитывается как «число двукратных различий» между экспрессиями какого-либо гена и контрольного.

Слайд 35.

Далее следует процедура нормализации данных, как стало принято называть процедуры стандартизации и централизации данных биочиповых экспериментов.

Итак, нормализации данных – это процесс выявления и исключения систематических ошибок, не связанных с дифференциальной экспрессией генов.

Обычно строят график ‘R-I’ (ratio/intensity), с помощью которого можно выявить артефакты в измерениях log2(ratio). Особенно такие искажения наблюдаются в области низких значений интенсивности сигналов гибридизаций и высоких.

Слайд 36.

Одна из разновидностей нормализации данных – это нормализация по тотальной интенсивности.

Если считать, что примерно одинаковое количество меченых молекул из образцов гибридизуются в пробами в биочипе, то общая интенсивность гибридизации должна быть одинаковой для окаждого меченного образца. Нормировочный фактор исчисляется как отношение суммированных интенсивностей по обоим каналам детекции.

Существуют методы глобальной или локальной нормализации, параметрической и непараметрической нормализации, и т.д.

Слайд 37.

Для снижения влияния стохастических ошибок и для повышения надежности цифровых данных, получаемых при обработке изображений, применяют распространенный способ – используют повторы экспериментов, образцов и т.д., т.е. разного рода реплики.

Различают биологические реплики, т.е. использование независимо приготовленных меченных образцов. Они дают информацию о естественной изменчивости в изучаемой биологической системе, а также случайные различия в процессе приготовления образцов.

Также есть технические реплики. Это повторы пробы, стёкол, гибридизаций и т.д. Они позволяют при одном и том же образце получить информацию о естественных и систематических ошибках методики.

Распространенный в ДНК-биочиповых экспериментах тип технических реплик – повторная гибридизация с теми же образцами, меченными наоборот (dye-reversal or flip-dye analysis), когда сначала и референсный и анализируемый образцы делятся на две порции. Одна порция какого-либо образца метится Cy3, а другая - Cy5. Потом проводят две гибридизации, в одной участвуют Cy3-референсный и Cy5-анализируемый образцы, а в другой – наоборот. Отцифрованные данные обеих гибридизаций усредняются для каждой пробы.

Слайд 38.

На этом слайде показана естественная вариабельность биологических образцов, в данном случае - образцов РНК от четырех особей, выявленная гибридизацией с ДНК-биочипами одной серии. Понятно, что применение биологических реплик позволяет отделить информацию, связанную с условиями эксперимента, от естественного шума.

Слайд 39.

Непременным этапом в процессе обработка изображений является оценка статистической значимости выявленных различий в интенсивности сигналов. Например, используют фильтр по значению дисперсии.

На графике показана форма распределения данных, если по оси абсцисс отложены суммарные интенсивности по красному Cy5 (R) и зеленому Cy3 (G) каналам, а по оси ординат – разность между ними. Видно, что в при разных значениях суммарной интенсивности наблюдается отклонение формы

«облака значений» от идеальной, когда «облако» распределено равномерно вдоль линии со значением разницы 0. Видно, что в области низких значений суммарной интенсивности происходит сильный разброс значений с преобладанием зеленых сигналов. В области больших значений суммарной интенсивности также наблюдается разброс. На графике линиями показаны области с различными уровнями достоверности различий.

Использование фильтра по значению дисперсии позволяет получить надежные статистически значимые данные по дифференциальной экспрессии генов. Однако всегда существует опасность потерять биологически значимые различия между генами в области малых интенсивностей, а также в области больших интенсивностей из-за насыщения сигнала (обычно для 16-битного сканнера предел измерения - 216–1=65,535 на пиксель).

Слайд 40.

В результате процедур, описанных выше, получают матрицы данных по экспресси генов. Такие матрицы фактически являются профилями экспрессии генов, а процесс получения этих матриц называется профилированием экспрессии генов GEP (Gene Expression Profiling).

Теперь рассмотрим, какие операции производят с этими матрицами в процессе дальнейшего тонкого анализа и биологической интерпретации данных.

Слайд 41.

Самый распространенный и очевидный подход при биологической интерпретации данных – это выявить дифференциально экспрессирующиеся гены с оценкой статистической значимости этой разницы в экспресиях. Для этого вычисляют Z-скор как меру статистической значимости. На графике показано цветовой кодировкой, что зеленые точки относятся к генам с достоверным отклонением вэкспрессии при сравнении двух образцов при двухцветной совместной гибридизации.

Слайд 42.

Более сложные методы анализа результатов серии биочиповых экспериментов заключаются в выделении групп или классов генов (или условий), проявляющих определенное сходство по набору качеств. Иными словами – это математические процедуры кластеризации или классификации данных.

Алгоритмы кластеризации (классификации) данных по экспрессии генов делятся на:

- методы безусловной или неконтролируемой классификации (unsupervised)
- методы контролируемой классификации (supervised)

Слайд 43.

К методам неконтролируемой классификации относятся иерархические алгоритмы кластеризации, которые бывают дивизимные и агломеративные. Первые начинают процесс выделения кластеров с разделения всего набора объектов на две группы, которые, в свою очередь, подразделяются на следующие два кластера. И так далее до исчерпания массива данных. Вторые, наоборот, стартуют с объединения наиболее похожих объектов. Потом к усредненному полученному объекту подбирается следующий наиболее похожий. И так далее.

К методам контролируемой классификации относятся неиерархические алгоритмы кластеризации. Наиболее часто применяют алгоритмы «K-средних» и «Самоорганизующиеся карты (SOM)».

Слайд 44.

На этом слайде показан пример профилирования экспрессии генов и кластеризация профилей с помощью иерархического алгоритма UPGMA (Unweighted Pair Group Method with Arithmetic mean). Показана кластеризация генов по экспрессии в течение трех стадий клеточного цикла по результатам измерений по 60 разным временным точкам.

Слайд 45.

С помощью планирования биочипа и всего эксперимента с помощью технических реплик можно заранее найти способ снижения влияния этих ошибок.

Необходимо учитывать два аспекта планирования:

(i) Определить какие пробы должны быть, должны ли быть реплики, есть ли возможность для множественных реплик, какие контроли и т.д.

(ii) Определить расположение проб – дизайн биочипа.

Важность предварительного планирования дорогостоящих экспериментов с использованием ДНК-биочипов объясняется тем, что значимая биологическая информация в этих экспериментах получается при сравнении результатов или совместной гибридизации, или нескольких последовательных гибридизации, и от того, что с чем будет сравниваться, сильно зависит точность и надежность выводов. Это проиллюстрировано на схеме. При прямом сравнении данных об испытуемом образце с данными контрольного образца дисперсия составляет $\sigma^2/2$. Однако, если нам необходимо испытать несколько образцов, сравнить их между собой, и при этом иметь возможность сравнить все данные с таковыми других экспериментов в других лабораториях, то мы вынуждены использовать референсные данные и работать с дисперсией $2\sigma^2$, т.е. в четыре раза больше.

Слайд 46.

Рассмотрим однофакторный эксперимент с тремя испытуемыми образцами A, B и C. Если мы используем референсный образец при непрямом сравнении, то для достижения приемлемого разброса данных нам необходимо потратить в два раза больше стекол, т.е. биочипов, и, соответственно, потратить в два раза больше материала в виде меченого образца. Поэтому дизайн биочиповых экспериментов определяется такими физическими ограничениями, как число стекол в наличии, или количество исходной мРНК для приготовления меченого образца.

Конечно, можно было бы сравнить напрямую с расходом двойной порции материала, но зато с еще меньшей дисперсией, но тогда мы бы не могли корректно сравнивать свои данные с данными из внешних источников.

Слайд 47.

Таким же образом можно проанализировать «цену» различных дизайнов биочип-экспериментов с четырьмя временными точками, когда нужно сравнить данные о последовательных стадиях какого-либо процесса. Тут главное – определить главную цель: выявить различия между всеми состояниями относительно первого или сравнить развитие экспрессии на всех стадиях.

Слайд 48.

Учитывая большую сложность, масштабность и зависимость от статистических обработок ДНК-биочиповых данных, можно понять, почему ведущие специалисты в этой области озаботились выработкой стандартов для этих данных. Затратив большие усилия на выработку общих согласованных решений, группа ученых организовала специальное Общество «Microarray Gene Expression Data» (MGED) (http://www.mged.org) для установления общих стандартов описания данных по биочип-экспериментам, систем обработки, передачи и хранения данных в общедоступных базах данных.

Слайд 49.

Среди прочих средств унификации и стандартизации был выработан язык «MicroArray Gene Expression Markup Language» (MAGE-ML) для создания общего формата и достижения сравнимости результатов; протокол «Minimum Information About a Microarray Experiment» (MIAME) для определения типа информации и степени подробности, с которой исследователь обязан ее представить; рабочую группу MGED «Society Ontology Working Group» (http://www.mged.org/ontology) для формирования набора контролируемых словарей и онтологий,

необходимых для описания биологических образцов и манипуляций с ними, т.е. экспериментальных процедур.

Слайд 50.

Протокол «Minimum Information About a Microarray Experiment» (MIAME) содержит требования к минимальной информации об опубликованном эксперименте, основанном на ДНК-биочип-методе, и включает шесть типов описаний:

1. План эксперимента - набор отдельных гибридизационных экспериментов
2. План биочипа – содержание пятен/ячеек, компоновка по рядам и т.д.
3. Образцы – источник, приготовление экстрактов, способ мечения
4. Гибридизация – процедура и параметры
5. Измерение – характеристики изображений и сканнеров
6. Нормировка – способ, коэффициенты и т.д.