Лекция «Применение ДНК-биочипов (микроматриц) для исследования экспрессии генов».
Часть 1. **Компьютерная транскриптомика**

Слайд 3-10. Часть материала, прочитанная проф. Н.А. Колчановым.

Слайд 11.

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 и прикреплены 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.

Развитие 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.

Слайд 12.

Рассмотрим, какие бывают типы ДНК-биочипов. ДНК-биочипы различают в зависимости от материала подложки и метода иммобилизации фрагментов ДНК. Самым первым появился метод иммобилизации фрагментов ДНК в объеме небольших капель или блоков геля – этот метод был разработан в ИМБ РАН российскими учеными во главе с академиком А.Д. Мирзабековым, первые публикации об этом методе - это Lysov et al., 1988 и Khrapko et al., 1989.

Затем был предложен метод иммобилизации фрагментов ДНК на поверхности стекла или реже – полимера. Обычно это микроскопное стекло размером 25 мм х 76 мм.

Различают два основных типы таких ДНК-биочипов в зависимости от природы используемых фрагментов ДНК: для производства олигонуклеотидных биочипов используют химически синтезированные одноцепочечные олигонуклеотиды длиной 20-75 н.о.; а для кДНК-биочипов – двухцепочечные фрагменты ДНК из библиотек кДНК длиной 100-2500 н.о., размноженные в бактериальных клетках или с помощью ПЦР.

Эксперименты с олигонуклеотидными ДНК-биочипами были впервые представлены в публикациях Fodor et al., 1993, Lipshutz et al., 1995 и Lockhart et al., 1996. А работы с кДНК-биочипами – в публикации DeRisi et al., 1996.

ДНК-биочипы с иммобилизацией фрагментов ДНК на твердой поверхности получили наибольшее применение в биологии и медицине, поэтому я в основном буду рассказывать об экспериментах, основанных на их использовании.

Слайд 13.

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 приготовленные таким образом пробы или мишени (targets) распределяются на поверхности стекла специальными аппаратами, называемыми arrayer или spotter.

Второй этап - приготовление образца (sample) для гибридизации. Из двух источников клеток – тестовых (или испытываемых, анализируемых) 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-блот-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 излучают свет 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 псевдоаналоговое 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 разницу в содержании тех или иных молекул в тестовом или референсном пулах кДНК, выделенных из образцов клеток.

Слайд 14.

Теперь рассмотрим типизацию ДНК-биочипов по способу их приготовления. Олигонуклеотидные 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 контактной или бесконтактной, струйная.

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 15.

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 16.

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

Слайд 17.

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 известных 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 в 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..
- олигонуклеотидные биочипы высокой плотности (high-density oligonucleotide arrays) могут содержать до 500,000 пар проб на одном стекле, причем одному гену соответствуют 11-20 проб.

Слайд 18.

Рассмотрим подробнее стадию 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 красителя цианин3 (Cyanine3 (Cy3)) и красного красителя цианин5 (Cyanine5 (Cy5).

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

Слайд 19.

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 отдельных процессов на всех предшествующих этапах, приводящее к значительному статистическому разбросу данных при проведении ДНК-биочиповых экспериментов.

Слайд 20.

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 21.

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 22.

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

Поэтому 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 23.

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 24.

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 25.

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 26.

Таким образом, применение статистических методов обработки сравнительных данных по оцифровке изображений результатов ДНК-биочип-гибридизации позволяет выявлять дифференциально экспрессирующиеся гены в тех или иных экспериментальных условиях. Например, на этом слайде цветом выделены области в распределении данных с разными значениями Z-score, т.е. с разным уровнем статистической значимости наблюдаемых различий.

Слайд 27.

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 28.

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 29.

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 30.

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.

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 31.

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.

Слайд 32.

После такого введения в биоинформатику ДНК-биочип-экспериментов рассмотрим на примерах применение ДНК-биочипов в биологии. Самые распространенные области их применения – это анализ полиморфизма, анализ экспрессии генов и сравнительный анализ геномов.

Слайд 33.

Методы анализа с помощью ДНК-биочипов полиморфизма в геномной ДНК генов, особенно их регуляторных областей, позволяют не напрямую, но опосредованно исследовать изменения в транскриптоме, чему посвящена наша лекция, поэтому я вкратце опишу их.

Миллионы нуклеотидных позиций, вариабельных у разных особей (single nucleotide polymorphism – SNP), могут быть скринированы с помощью специально разработанных ДНК-биочипов. SNP-биочипы используются для исследования (1) сцепления между маркерами, (2) неравновесия по сцеплению (linkage disequilibria), (3) потери гетерозиготности (loss of heterozigosity).

На схеме представлены три экспериментальные стратегии для анализа последовательностей ДНК с помощью биочипов. Одноцветный анализ “gain-of-signal” позволяет после гибридизации с чипом, в котором для каждого SNP предусмотрены четыре различающихся по центральной вариабельной позиции аллель-специфичных олигонуклеотида, обнаруживать присутствие нового аллеля благодаря

появлению нового сигнала. Двухцветный анализ “loss-of-signal” позволяет после конкурентной совместной гибридизации испытуемого и референсного образцов с SNP-биочипом с большей специфичностью обнаруживать присутствие нового аллеля. Наконец, метод минисеквенирования состоит в проведении реакции аллель-специфичного удлинения олигонуклеотида на один из четырех меченых дидезоксинуклеотидов прямо на чипе.

Слайд 34.

Рассмотрим теперь примеры анализа экспрессии генов с помощью ДНК-биочипов. Сформировалось целое направление – создание «молекулярных паспортов» каких-либо клеток или тканей, например, стволовых клеток. Пример взят из статьи: Tanaka T.S., *et al.*, 2002. Gene expression profiling of embryo-derived stem cells reveals candidate genes associated with pluripotency and lineage specificity. Genome Res. 12(12):1921-1928 . Были взяты образцы четырёх клеточных линий: ES (эмбрионально-стволовые), TS3,.5 и TS6,.5 (трофобласт-стволовые), MEF (мышиные эмбриональные фибробласты), и мышиный кДНК-биочип NIA 15K. На графике гены, экспрессия которых достоверно ($P < 0.05$) превышала фоновое значение, выделены цветом. Кластер-анализ методом «k-средних» дифференциально экспрессирующихся генов выявил 15 кластеров, объединяющих гены со сходным профилем экспрессии.

Слайд 35.

На этом слайде показаны результаты иерархической кластеризации 346 генов, специфично экспрессирующихся в исследованных образцах. Видно, что кластеры A и E объединяют профили экспрессии генов, характерных для TS линии; кластер B – для MEF; кластер C – соответствует генам со слабой экспрессией; кластер D – для ES линии.

Каждому гену приписан номер принадлежности к какому-либо кластеру. Кластер 4 соответствует ES специфичным генам, 7 - TS специфичным, 14 - MEF- специфичным, 12 – общим для ES и TS клеток.

Слайд 36.

Кроме непосредственного измерения содержания транскриптов множества генов ДНК-биочип-технология позволяет исследовать зависимость транскрипции от таких фундаментальных процессов, как состояние хроматина множества генов, временем репликации этих генов, уровнем трансляции их транскриптов – и все это в масштабах всего генома.

Рассмотрим, как можно с помощью ДНК-биочипов исследовать связь между экспрессией генов и эпигенетическим состоянием геномных районов вокруг них, например, исследовать степень метилированности ДНК человека и млекопитающих.

Метилирование ДНК – один из эпигенетических механизмов. Основной мишенью для метилирования в геноме человека и млекопитающих является цитозин. Чаще всего метилирование происходит в контексте динуклеотидов CpG, хотя CpNG, CC(a/t)GG, CpA и CpT также могут быть метилированы. Так называемые «CpG-островки», как правило, охватывают промоторы и первые экзоны генов. Метилированое состояние «CpG-островков» часто ведет к подавлению экспрессии генов. Показано во многих работах, что распределение метилированных сайтов ДНК в нормальных и трансформированных клетках значительно различается – в опухолевых клетках наблюдается гипометилирование одних сайтов ДНК и гиперметилирование других.

Слайд 37.

На этом слайде показан принцип исследования связи между экспрессией генов и степенью метилированности их ДНК по материалам статьи Novik K.L., et al., 2002, Epigenomics: genome-wide study of methylation phenomena (Curr Issues Mol Biol. 4(4):111-128). Образцы из фрагментов геномной ДНК обрабатываются бисульфитом натрия, превращающего неметилированные цитозины в урацилы, затем эти образцы метятся в процессе амплификации и гибридизуются с олигонуклеотидными биочипами. Для каждого потенциального сайта метилирования разрабатывается пара олигонуклеотидов. В случае если сайт метилирован, то в анализируемой позиции образца останется

цитозин, и сигнал после гибридизации будет от олигонуклеотида, содержащего в соответствующей позиции гуанозин (правая часть левого рисунка). А если сайт не метилирован, то сигнал будет от олигонуклеотида, содержащего в соответствующей позиции аденозин (левая часть левого рисунка). Если использовать двухцветную совместную гибридизацию испытываемого и референсного образцов, то анализ расположения зеленых, красных или желтых (т.е. образец представляет собой или гетерозиготу, или гетерогенную смесь клеток) сигналов на ДНК-биочипе позволит определить состояние метилированности каждого потенциального сайта метилирования ДНК каждого исследуемого гена. А сопоставление этих данных с результатами параллельного биочип-эксперимента по определению количества мРНК этих исследуемых генов в этих же образцах позволило выявить искомые корреляции между процессами метилирования ДНК генов и их уровнем транскрипции.

Слайд 38.

Рассмотрим, как можно исследовать связь между экспрессией генов и временем репликации соответствующих районов генома на примере статьи Schübeler D., et al., 2002, Genome-wide DNA replication profile for Drosophila melanogaster: a link between transcription and replication timing. Nature Genet. 32:438-442.

На рисунке показан профиль клеточного цикла, выявляемый после импульсного введения в клетки бром-дезокси-уридина (BrdU) и окрашивания клеток ДНК-специфическим красителем (пропидиум иодидом). Используя метод FACS (fluorescence-activated cell sorting), клетки затем сортируются по содержанию ДНК в зависимости от флуоресценции пропидиум иодида. Затем проводят иммунопреципитацию новосинтезированной ДНК антителами против BrdU, амплификацию осажденных фрагментов ДНК и введение в ампликоны флюорометки, в данном случае образцы из клеток, находящихся на стадии ранней репликации, помечены зеленым Cy3, а образцы из клеток, находящихся на стадии поздней репликации, помечены красным Cy5.

Совместная двухцветная гибридизация меченых образцов с кДНК-биочипом, содержащим 6500 генов дает возможность выявить момент репликации каждого конкретного гена. В этом эксперименте были использованы три биологические реплики, т.е. для приготовления образцов все процедуры и манипуляции с клетками и ДНК проводились независимо три раза.

Слайд 39.

После обработки данных гибридизации меченых образцов с кДНК-биочипом был получен репликационный профиль для 6500 генов из секвенированной части генома *D. melanogaster*. А сопоставление этих данных с результатами параллельного биочип-эксперимента по определению количества мРНК этих генов в этих же образцах позволило выявить корреляции между транскрипционной активностью гена и временем его репликации: чем выше уровень транскрипции гена, тем раньше он реплицируется, и наоборот.

Слайд 40.

Рассмотрим, как можно исследовать связь между транскрипционной и трансляционной активностями генов на материале статьи Arava Y., et al., 2003, Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae (Proc Natl Acad Sc USA. 100(7):3898-3894).

Эксперимент базируется на предположении, что чем больше уровень трансляции мРНК, тем больше на ней полисом. Поэтому сначала были экспериментально получены полисомные профили с помощью седиментации мРНК в сахарозном градиенте, отбора образцов из разных фракций, экстракция и обратная транскрипция образцов мРНК, мечение образцов кДНК. Затем проводилась гибридизация образцов кДНК с ДНК-биочипами, содержащими все открытые рамки считывания *Saccharomyces cerevisiae*

Слайд 41.

Таким образом получались количественные данные о содержании отдельных разновидностей мРНК в полисомальных фракциях. Экспрессия некоторых генов была проверена на полуколичественном уровне методом нозерн-блот-гибридизации.

Слайд 42.

Следует учитывать при таком подходе, что количество рибосом на мРНК зависит от длины транскрипта, поэтому необходимо исследовать не только число, но и плотность рибосом. На рисунках показаны: гистограмма А., на которой гены сгруппированы соответственно длине их ОРС, гистограмма В. Число генов как функция плотности рибосом и график С. Плотность рибосом как функция от длины ОРС.

Слайд 43.

В последнее время к таким компонентам транскриптома как мРНК, рибосомальным, транспортным, малым ядерным и т.д. РНК, был добавлен еще один – малые интерферирующие РНК, способные подавлять экспрессии генов через взаимодействие с их мРНК по механизму или расщепления транскриптов или подавлению трансляции с них. Среди этих новых представителей транскриптома особенно интересны микроРНК (или сокращенно – миРНК), поскольку они являются эндогенными продуктами генома, т.е. генерируются посредством транскрипции, и в этом отношении миРНК-гены сходны с белок-кодирующими генами тем, что обладают дифференциальной экспрессией. Поэтому получение данных о содержании конкретных миРНК на определенной стадии или в определенной ткани необходимо для полноценного и точного знания о механизмах, определяющих количество того или иного транскрипта, являющегося мишенью для какой-либо миРНК.

С этой целью разработаны методы профилирования экспрессии миРНК-генов с помощью специализированных ДНК-биочипов. На слайде приведена в качестве примера работа Barad et al., MicroRNA expression detected by oligonucleotide microarrays: system establishment and expression profiling in human tissues. (Genome Res. 2004 Dec;14(12):2486-94), в которой описан профиль экспрессии 150 миРНК человека в пяти органах и клеточной культуре. (Шкала интенсивности представляет log2 значений интенсивности сигнала после вычитания фона и нормализации).

Слайд 44.

На следующем слайде показаны пример работы по экспериментальной аннотация человеческого генома (Shoemaker DD, et al., Experimental annotation of the human genome using microarray technology. Nature. 2001 409:922-927).

В этой работе была предпринята попытка оценки компьютерных предсказаний генов и определение полноразмерных транскриптов с помощью выявления одновременной экспрессии их экзонов. Таким способом были проанализирована экспрессия генов из 22 хромосомы (8,183 экзонов) в 69 парах экспериментальных условий. Было выявлено 572 группы корегулируемых экзонов или 572 проверенных по экспрессии гена (expression-verified genes - EVGs). Это составило 210 (85%) из 247 известных генов 22 хромосомы и 185 (57%) из 325 предсказанных генов.

Слайд 45.

На этом слайде показаны псевдоцветное изображения взвешенного по ошибкам log10-отношения (красный/зеленый) экспрессии каждого из ~8000 экзонов. Можно видеть, как работает алгоритм выявления коэкспрессии экзонов и «сборка» их в ген.

Слайд 46.

Затем в этой же работе была проведена проверка полученных результатов с помощью ДНК-биочипов, содержащих 60-мерные фрагменты, перекрывающие геномные последовательности генов. С помощью этих перекрывающихся рядов проб, соответствующих обеим цепям отдельных геномных районов хромосомы 22, были уточнены границы экзонов и самих транскриптов.

Слайд 47.

В конце концов, в этой работе было проведено сканирование экзонов в масштабах всего генома: было взято 1 090 408 проб, 110 000 обратно-комплементарных проб, 50 биочипов, два экспериментальных условия (РНК из двух клеточных линий). На гистограммах показана в виде красных полос доля экспериментально (для этих условий) проверенных экзонов.

Слайд 48.

Теперь вкраце обрисуем круг области применения ДНК-биочиповых технологий исследования дифференциальной экспрессии генов в медицине.

При исследовании нормальных тканей можно составлять прогнозы, т.е. оценивать предрасположенности к тем или иным заболеваниям.

При исследовании пораженных болезнью тканей можно составлять/уточнять диагнозы, изучать течение и особенности патологии, выявлять мишени для лекарственных средств.

При исследовании тканей, подвергнутых воздействию химических соединений/лекарственных средств, получают информацию об эффективности лекарств, или о токсичности этих соединений.

Слайд 49.

Рассмотрим примеры работ по второму направлению, как наиболее представленному в литературе. Один из аспектов этого направления - молекулярная классификация злокачественных опухолей. Например, в статье Bittner M. et al., Molecular classification of cutaneous malignant melanoma by gene expression profiling (Nature. 2000, 406(6795):536-540) проведена селекция кластеров меланом с помощью двумерного кластерного анализа образцов опухолей и генов. В целях диагностики были выявлены четыре самых контрастных по экспрессии генов в кластерах и характеристические гены в них.

Слайд 50.

В приведенной работе проведена таксономия меланом на основе данных профилирования экспрессии генов с помощью кДНК-биочипа (6,971 генов). Показаны: a. Дендрограмма иерархической кластеризации 19 меланом по данным экспрессии генов; b. Трехмерный график значений MDS для 31 меланомы. Показан кластер из 19 образцов.

Слайд 51.

На этом слайде показано выявление генов, экспрессия которых позволяет различать кластеры меланом. Видны результаты MDS-анализа распределения генов по их вкладу в минимизацию объема кластера и максимизацию межкластерной дистанции; а также выделена экспрессия 22 самых заметных генов в кластерах.

Слайд 52.

В следующей статье, посвященной молекулярной классификации злокачественных опухолей, Khan J. et al., Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks (Nat Med. 2001 7(6):673-679), описана таксономия нейробластом (NB), рабдомиосарком (RMS), лимфом Ходжкина (NHL) и опухолей семейства Эуинга (EWS) на основе данных профилирования экспрессии генов с помощью кДНК-биочипа, включающего последовательности 3789 кДНК и 2778 EST.

Слайд 53. **Некоторые аспекты статистического анализа данных ДНК-чипов**

Слайд 54. **Обработка изображений**

Принято выражать экспрессию гена как отношение между интенсивностью сигнала от исследуемого гена, выявленного, например, Cy5-меченным образцом (red), и интенсивностью сигнала от контрольного гена, выявленного, соответственно, Cy3-меченным образцом (green): Ti=Ri/Gi.
Гены с повышенной в два раза экспрессией будут иметь отношение 2, а с пониженной в два раза – 0.5. Т.е различие будет несимметрично.
Поэтому применяют преобразование отношения в виде логарифма по основанию 2 : M=log2(Ri/Gi) .
При этом понижение интенсивности в 2 раза и повышение в 2 раза имеют симметричны и равны -1 и +1 соответственно.

Слайд 55. **Графическое представление сигналов.**
Существует два наиболее распространенных способа графического представления сигналов. Первый использует соотношение логарифмов яркости зеленого и красного сигналов, которые откладываются на оси абсцисс и ординат, соответственно. Во втором откладываются преобразованные значения M=log(R)-log(G) и A= (log(R)+log(G))/2, что позволяет представлять отклонения сигналов от среднего более наглядно.

Слайд 56. **Систематическое отклонение**
Средняя интенсивность сигнала (красная линия) отклоняется от нулевого значения. Это означает присутствие систематического отклонения значений сигналов. Систематической отклонение вызвано рядом технических причин при проведении анализов и обработке сигналов, и может быть статистически выявлено и скорректировано.

Слайд 57. **Вид распределения после нормализации.**
Красная линия, обозначающая среднее значение, близка к нулю. Точки, лежащие выше нулевой линии, соответствуют генам у которых экспрессия повышена, лежащие ниже – пониженная экспрессия, лежащие вблизи линии – не изменили существенно уровень экспрессии.

Слайд 58. **Обработка данных, полученных методом ДНК-биочипов**
Обработка включает в себя получение исходных данных – сканированные изображения чипов после гибридизации, матрицы измерений – оцифрованные по специальным алгоритмам распознавания значения сигналов, и матрицы данных экспрессии генов, по которым производится дальнейший анализ.

Слайд 59. **Выделение генов с различающимся уровнем экспрессии**
Матрица экспрессии представляет собой записанные в форме таблицы значения экспрессии генов (относительной или абсолютной). Столбцам соответствуют анализы, строкам – отдельные гены.

Слайд 60. **Выделение выборок.**
При сравнении экспрессии гена в двух группах анализов из матрицы выделяется два соответствующих вектора.

Слайд 61. **Сравнение выборок.**
Затем их средние и дисперсии сравниваются с помощью статистических тестов, например критерия Стьюдента.

Слайд 62. **Анализ корреляций профилей экспрессии генов**
Основные меры сходства профилей экспрессии двух генов:

- корреляция Пирсона
- Евклидово расстояние

В зависимости от выбора меры результаты анализа могут сильно различаться, что иногда делает задачу анализа отдельной математической проблемой.

Слайд 63. **Методы интерпретации результатов**
Алгоритмы кластеризации (классификации) данных по экспрессии генов делятся на:

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

Слайд 64. **Пример кластеризации**
Иерархический алгоритм UPGMA (Unweighted Pair Group Method with Arithmetic mean).
Кластеризация генов по экспрессии в течение трех стадий клеточного цикла по результатам измерений по 60 разным временным точкам.

Слайд 65. **Двоякая интерпретация матрицы экспрессии.**
Каждый ген представлен вектором – строкой из матрицы экспрессии.
Аналогично можно построить дерево из векторов-столбцов, и тогда получится разбиение на классы не генов, а образцов.

Слайд 66. **Визуальный анализ**
Позволяет в наглядной форме представить результаты кластеризации, выделить группы генов и образцов со сходной экспрессией.

Слайд 67. **Полезные WWW-ресурсы**
www.r-project.org Свободно распространяемая программная среда для анализа данных. Большие возможности для статистической обработки и графического представления данных, в т.ч. Microarray Analysis.
www.bioconductor.org Проект по созданию библиотек для R для анализа Microarray данных.
http://rana.lbl.gov Страничка лаборатории Майкла Айзена, одного из корифеев в области анализа Microarray данных (университет Беркли). Популярные программы, статьи, данные экспериментов.
http://genome-www5.stanford.edu/ Stanford Microarray Database. В открытом доступе более 200 статей и исходные данные более чем 7700 Microarray экспериментов.
http://www.ncbi.nlm.nih.gov/geo/ NCBI Gene Expression Omnibus, большая база данных по экспрессии генов, открытый доступ.
www.microarrays.org Протоколы проведения экспериментов, программы для обработки данных.
http://www.ebi.ac.uk/arrayexpress/ база данных ArrayExpress EBI, открытый доступ.