Задача оценивания экспрессии генов по данным микрочипового анализа

Рябенко Е. А., Когадеева М. С., ВМК МГУ

ММРО-15, Петрозаводск

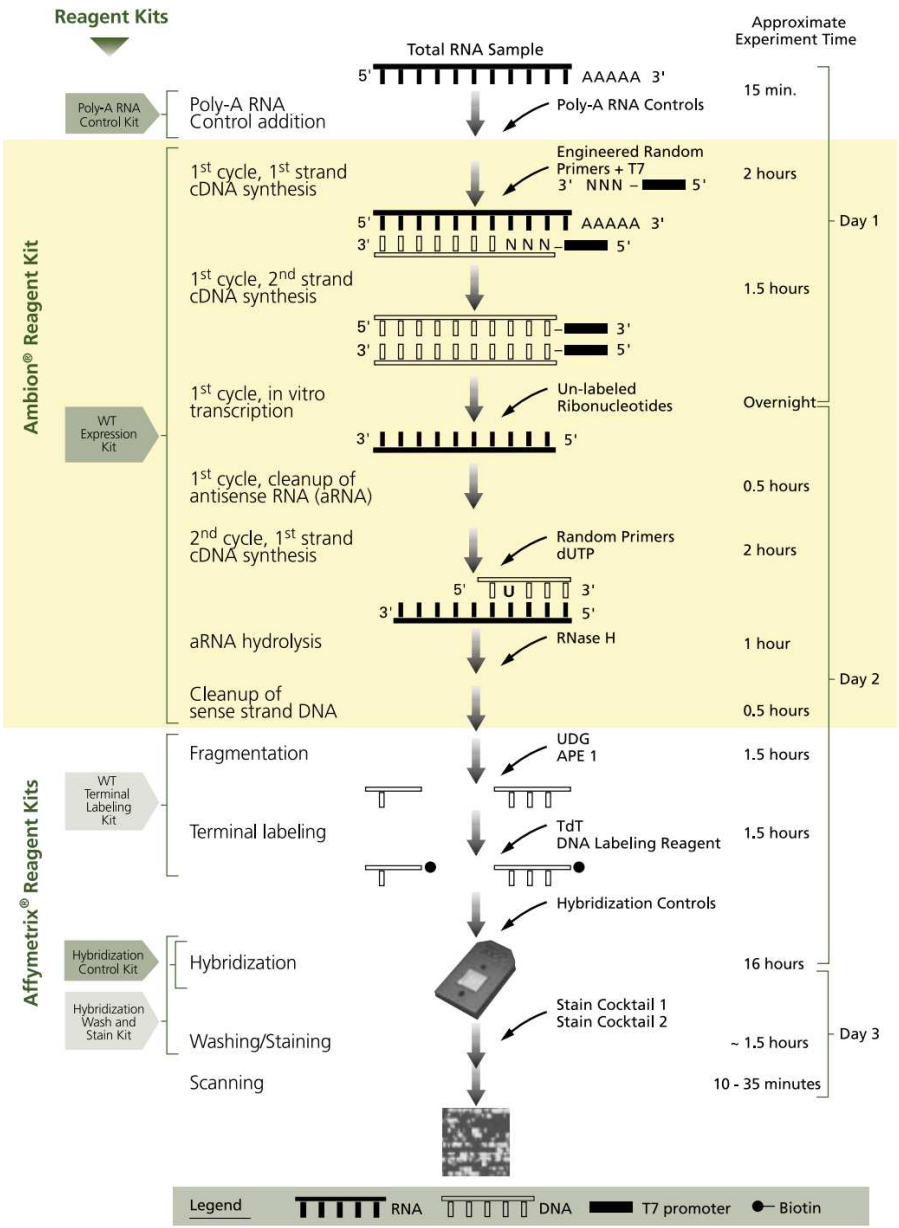
16 сентября 2011 г.

Генетика за 30 секунд

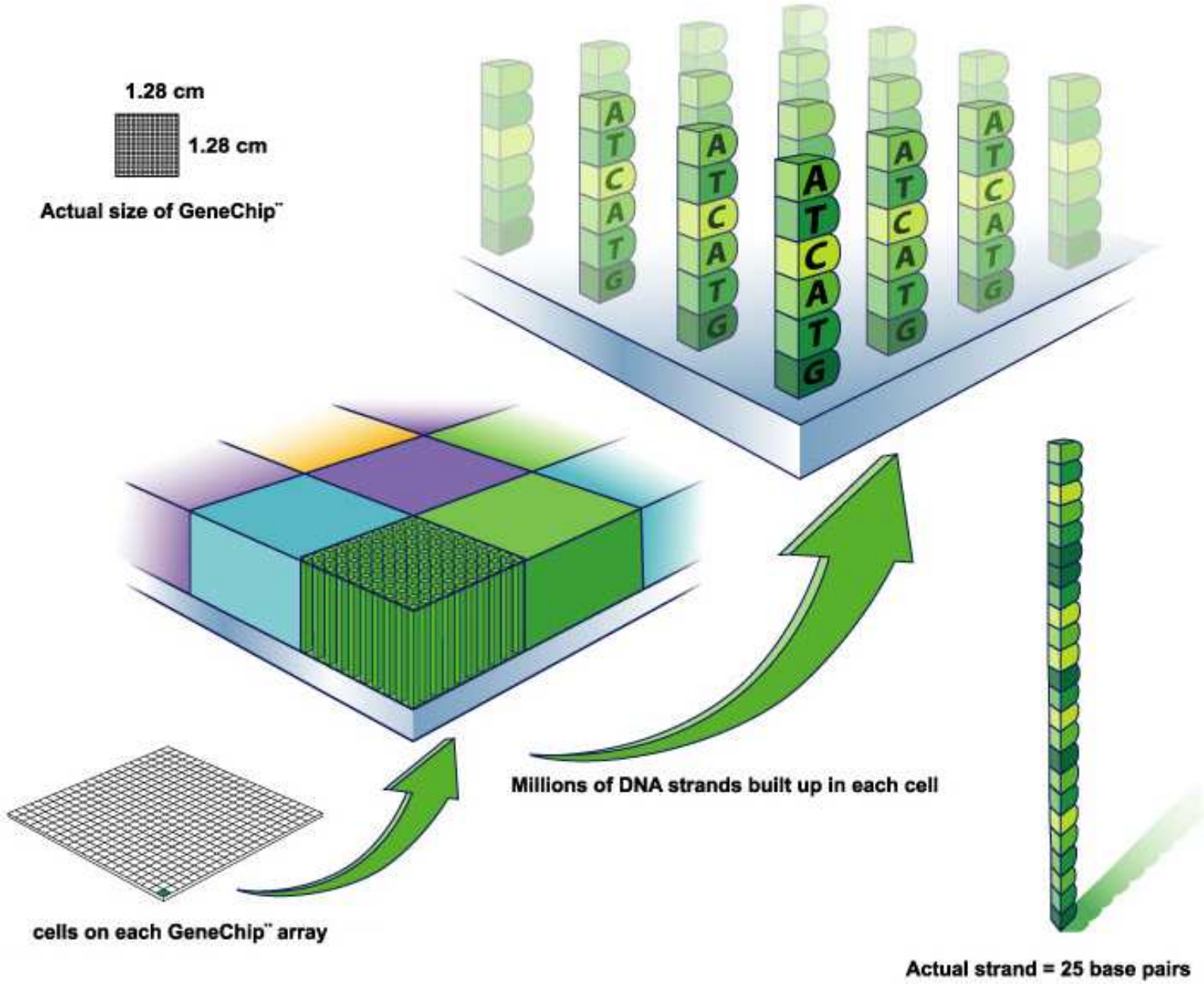
ДНК — молекула, содержащая всю информацию, необходимую для функционирования клетки.
Ген — участок ДНК, несущий какую-либо целостную функциональную информацию.
Экспрессия гена — процесс преобразования информации, содержащейся в гене, в функциональный продукт.
РНК — молекула-посредник, передающий информацию о гене структурам клетки, отвечающим за синтез белка; однозначно соответствует гену.

Количество молекул РНК в клетке служит мерой активности гена (оценкой экспрессии).

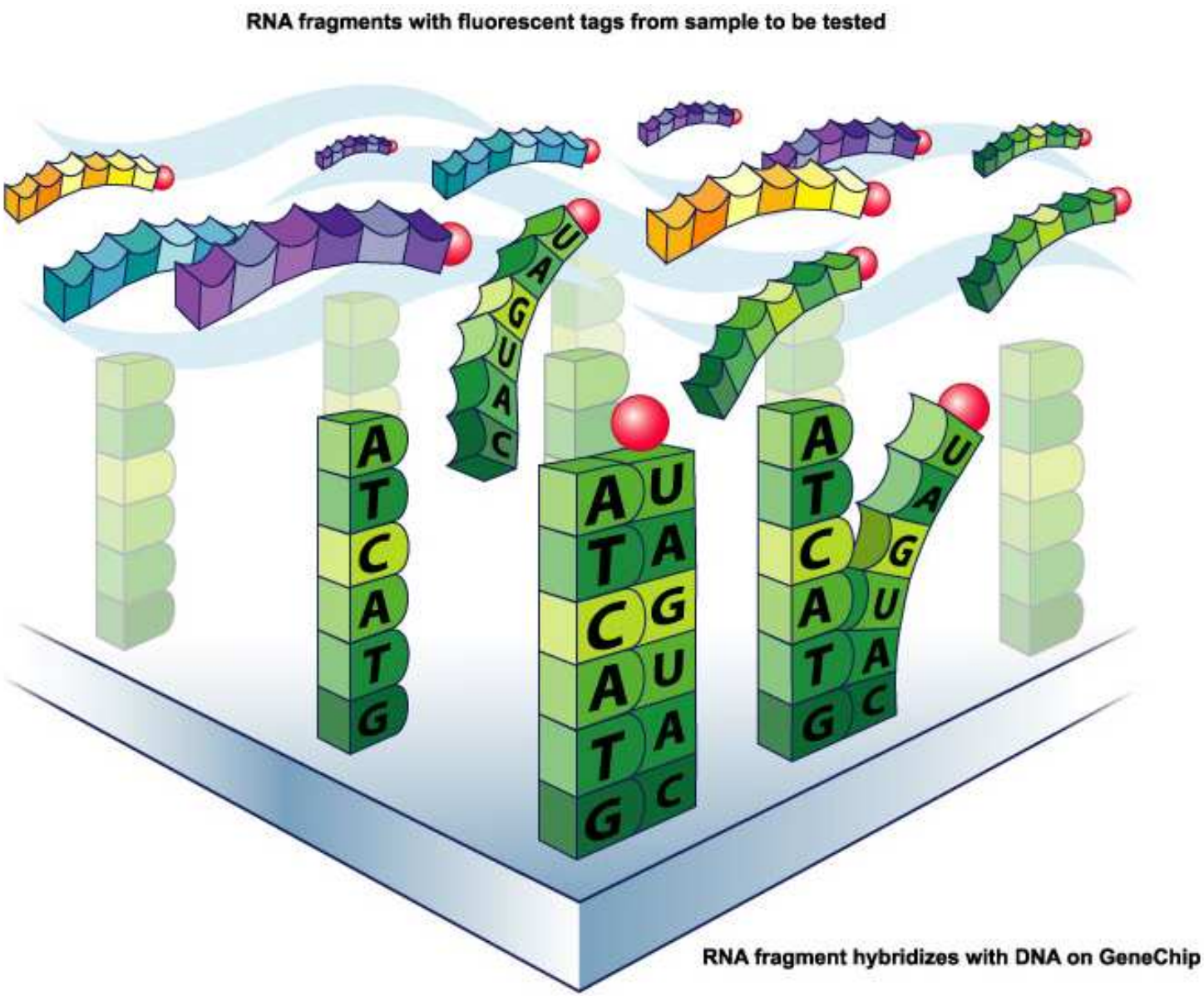
Ход эксперимента



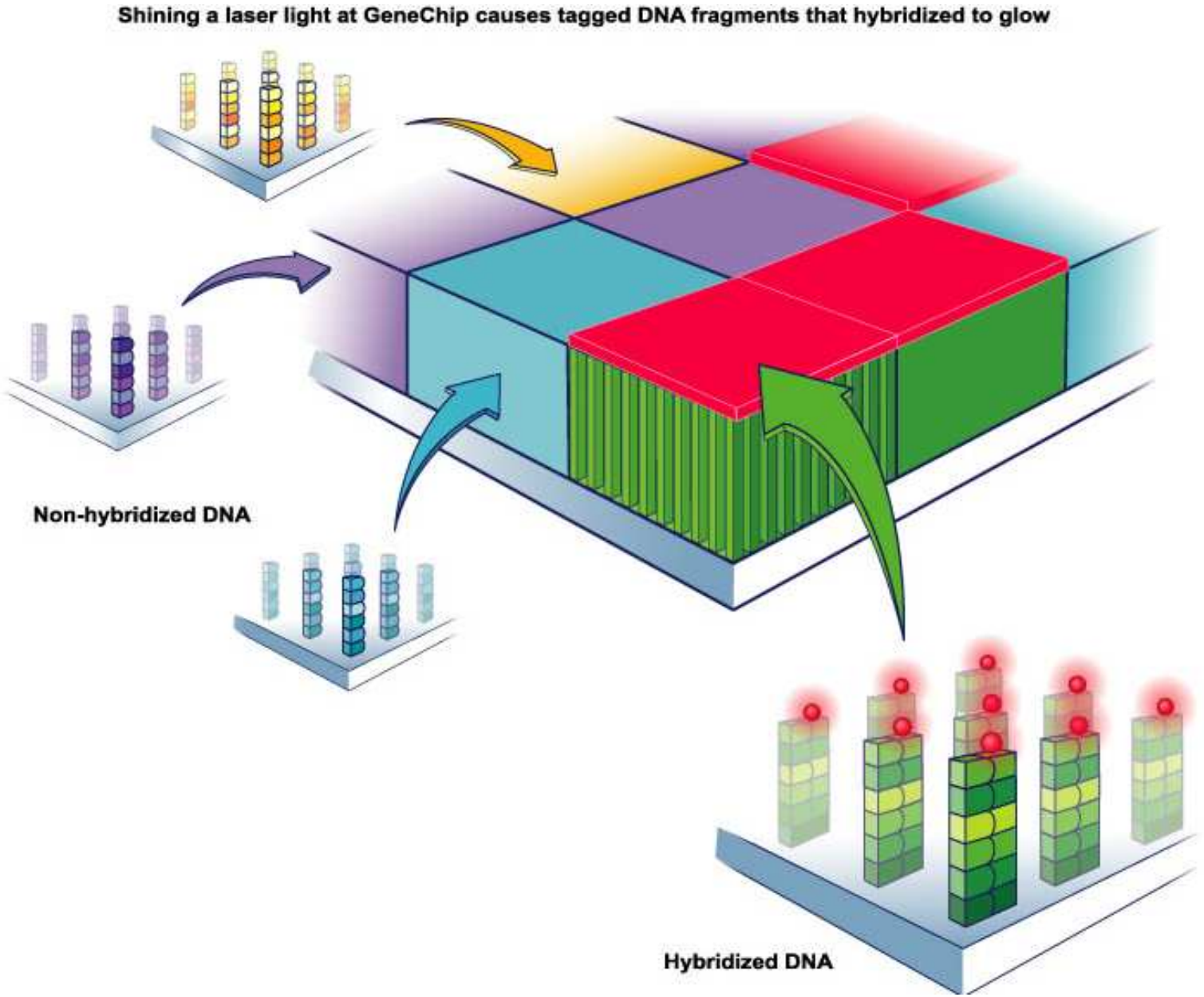
Вид микрочипа



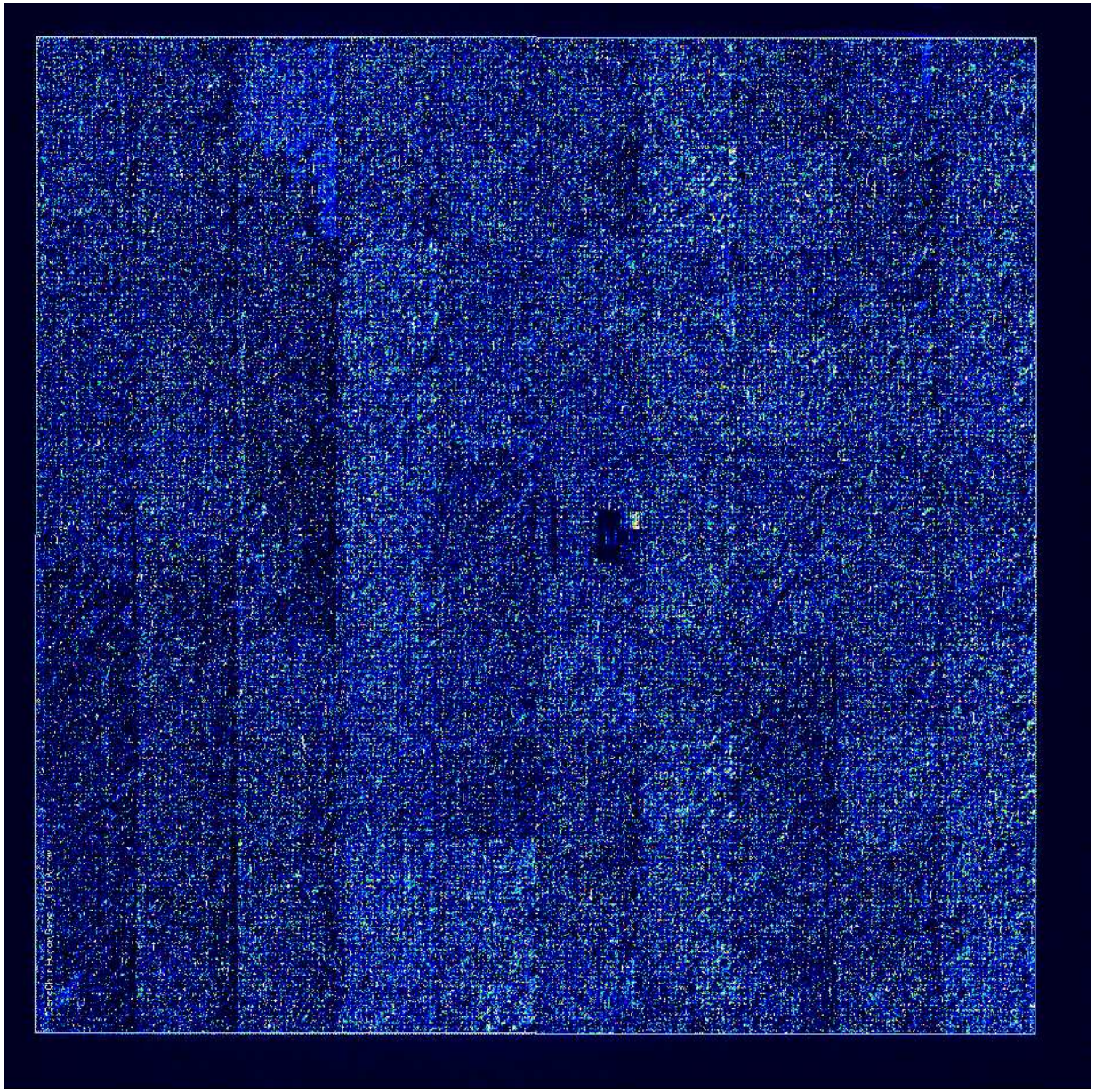
Гибридизация



Сканирование

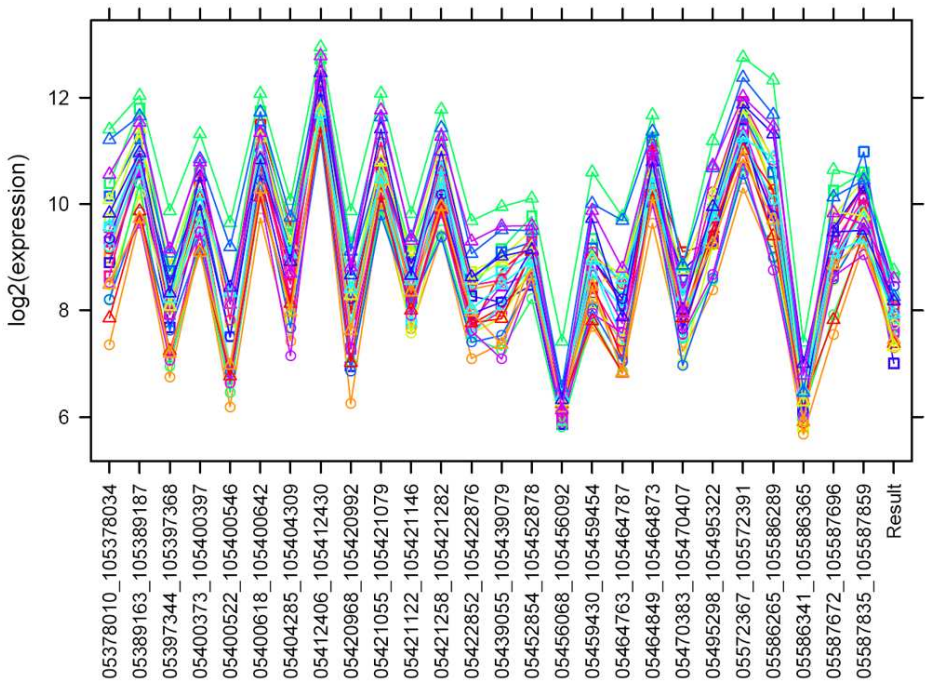


Результат сканирования



Получение оценок экспрессии

1. Изображение со сканера оцифровывается, получаем вектор значений интенсивности флуоресценции проб.
2. Проводится предобработка интенсивностей.
3. Значения предобработанных интенсивностей всех проб каждого гена **усредняются** (median polish), давая оценку экспрессии.



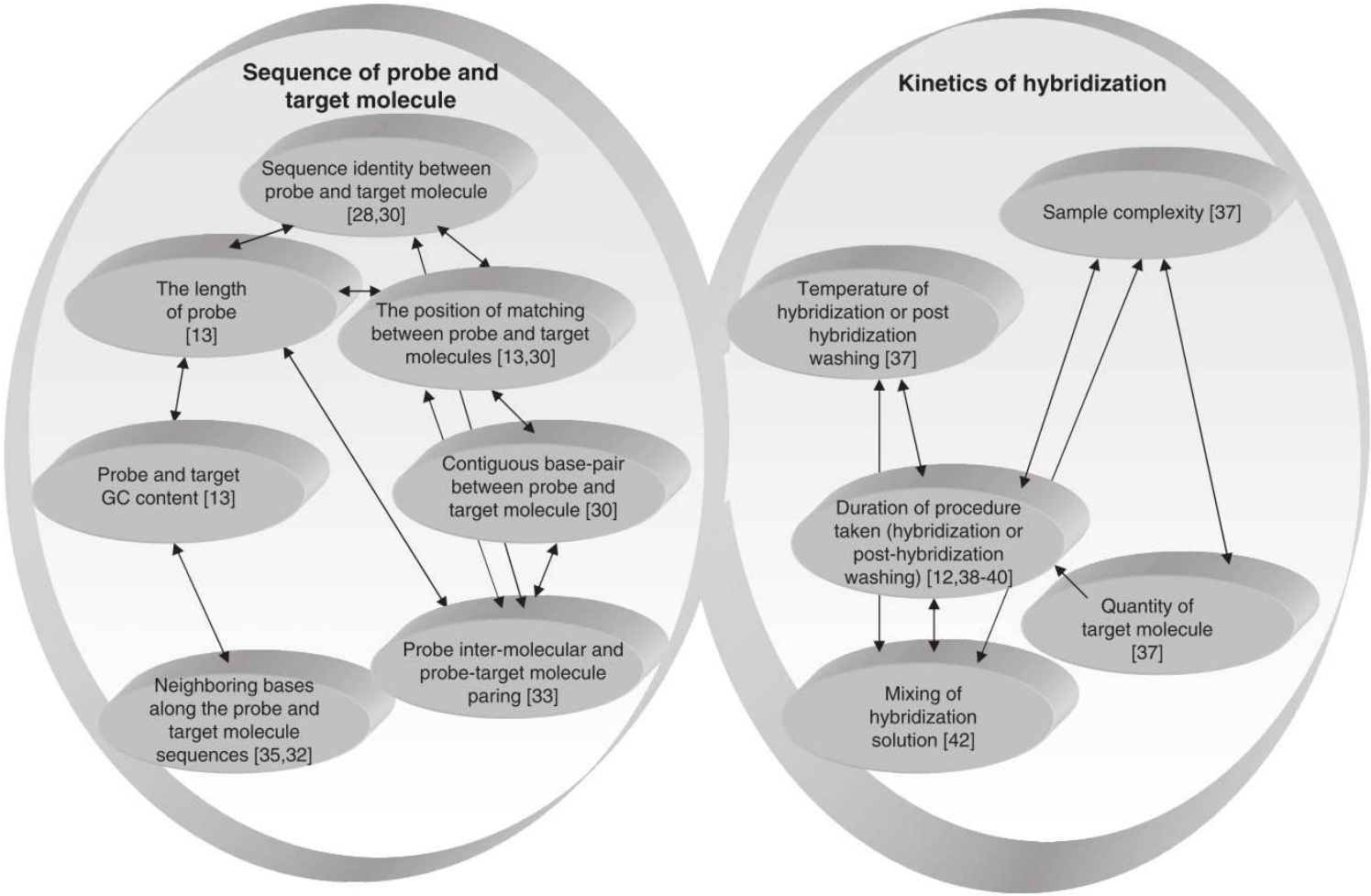
Мешающие факторы

На интенсивность флуоресценции каждой пробы дополнительно влияют: её расположение на гене, нуклеотидный состав, наличие несовпадений с последовательностью гена, их положение, а также наличие совпадений с последовательностями других генов.

Кроме того, на высоких концентрациях генов проявляется эффект насыщения — их взаимосвязь с интенсивностью становится нелинейной.

Влияние этих факторов меняется в зависимости от конкретного вида микрочипов (длина пробы, плотность проб, структура подложки, используемые реагенты и т. д.

Взаимосвязь факторов



Koltai, 2008

Попытки построения моделей гибридизации

Работы, в которых предлагались физические модели гибридизации:

Bloomfield, 2000;
Dai, 2002;
Yuen, 2002;
Hekstra, 2003;
Held, 2003;
Mei, 2003;
Nielsen, 2003;
Rouillard, 2003;
Zhang, 2003;
Zucker, 2003;
Binder 2004;
Rouillard, 2004;
SantaLucia, 2004;
Shen, 2004;
Vallone, 2004;
Wu, 2004;
Binder, 2005;
Levicky, 2005;
Schaupp, 2005;
Wu, 2005;
Sorokin, 2007;
Suzuki, 2007;
Wernersson, 2007;
Kroll, 2008;
Naiser, 2008;
Ono, 2008;
Pozhitkov, 2008;
Wei, 2008;
Ferrantini, 2009;
Hooyberghs, 2009;
Kroll, 2009;
Mulders, 2009;
Hooyberghs, 2010;
Mueckstein, 2010;
Xia, 2010;
Ambia-Garrido, 2011

Модель

Будем учитывать взаимодействие каждой пробы с каждым геном:

$$I_i^k = \frac{a_1 \sum_l C_l}{1 + a_2 \sum_l C_l} + \sum_j A_{ij} C_j^k$$

I_i — интенсивность флуоресценции i-й пробы, $i = \overline{1, N}$,
C_l, C_j — концентрации РНК генов, $l = \overline{1, M_i}$, $j = \overline{1, M}$,
a_1, a_2 определяют характер взаимодействия проб с генами, содержащими полностью комплементарные участки,
$A_{ij} \geq 0$ характеризует силу неспецифического взаимодействия между пробой i и геном j,
k — индекс номера микрочипа в выборке, $k = \overline{1, K}$.

Требуется восстановить элементы матрицы A_{ij}, характерной для данного вида микрочипа.

Характерные размерности

Микрочипы Affymetrix Human Gene 1.0 содержат $N = 861493$ проб к $M = 28869$ генам.

То есть, выписанная модель содержит от 28 миллиардов неизвестных параметров.

На NK известных величин приходится $M\,(N + K + 1)$ параметров, т. е. неизвестных становится меньше при размере выборки примерно равном среднему числу параметров для одной пробы.

Дополнительная информация

К счастью, для каждой пробы взаимодействия с большинством генов невозможны: если последовательности пробы и гена имеют слишком мало общего, гибридизация гарантированно не произойдёт.

Имея информацию о последовательностях проб и генов, можно заранее обнулить большую часть коэффициентов A_{ij}.

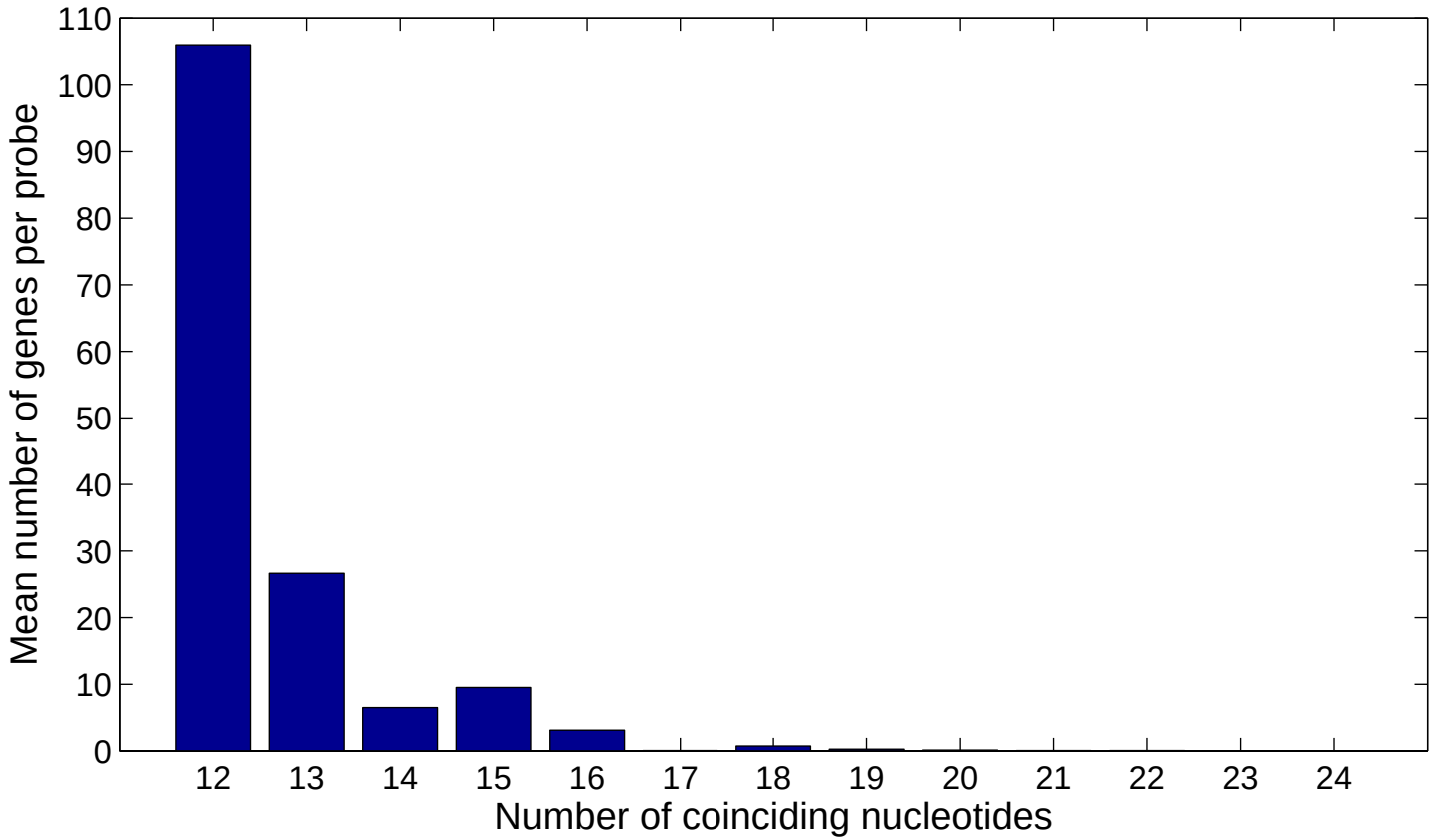
Используем алгоритм BLASTN для поиска максимального совпадения между последовательностями каждой пары проба–ген.

Минимальная значимая длина совпадения

Длина проб Affymetrix Human Gene 1.0 — 25 нуклеотидов.
Вполне корректно предположить, что гибридизация не будет происходить при максимальном совпадении длиной меньше 12, и положить $A_{ij} = 0$ при $BLASTN\,(Gene_j, Probe_i) < 12$.

При таком ограничении каждой пробе соответствует в среднем 150 ненулевых коэффициентов.
Тогда число известных величин превышает число неизвестных начиная с $K = 150$.

Распределение среднего числа комплементарных генов

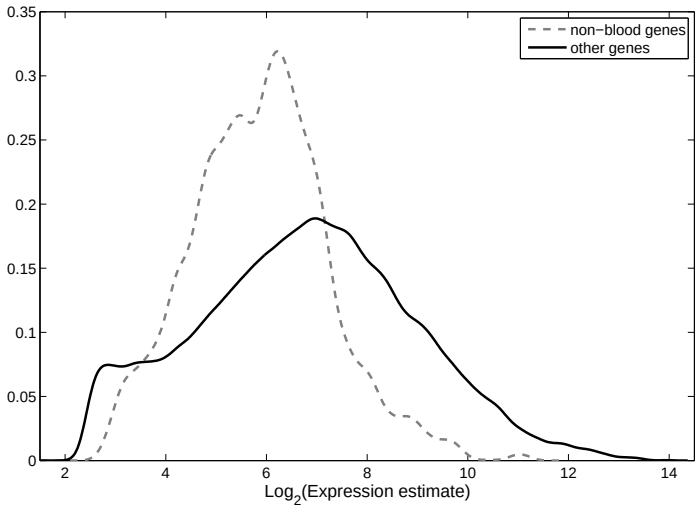


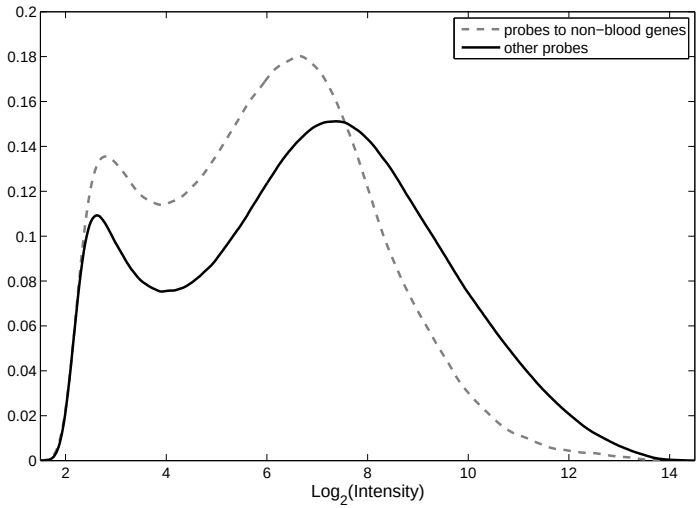
Можно ли уточнить порог включения коэффициента в модель?

Тканеспецифичность

Используем информацию о тканеспецифичности генов.

Имеется 70 микрочипов Affymetrix Human Gene 1.0, источник РНК — клетки крови.
При помощи баз данных Gene Expression Barcode и TiGER были выделены списки генов, экспрессия которых нехарактерна для клеток крови.
Пересечение этих списков содержит 787 генов, к которым на чипах имеется 24835 проб.
Для этих проб уравнения содержат только слагаемое, соответствующее кросс-гибридизации.





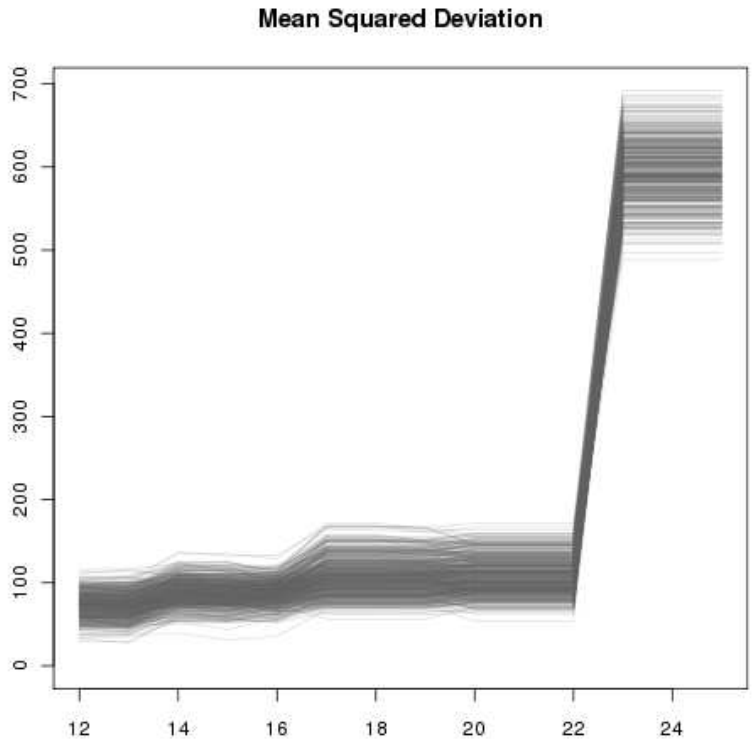
Задача

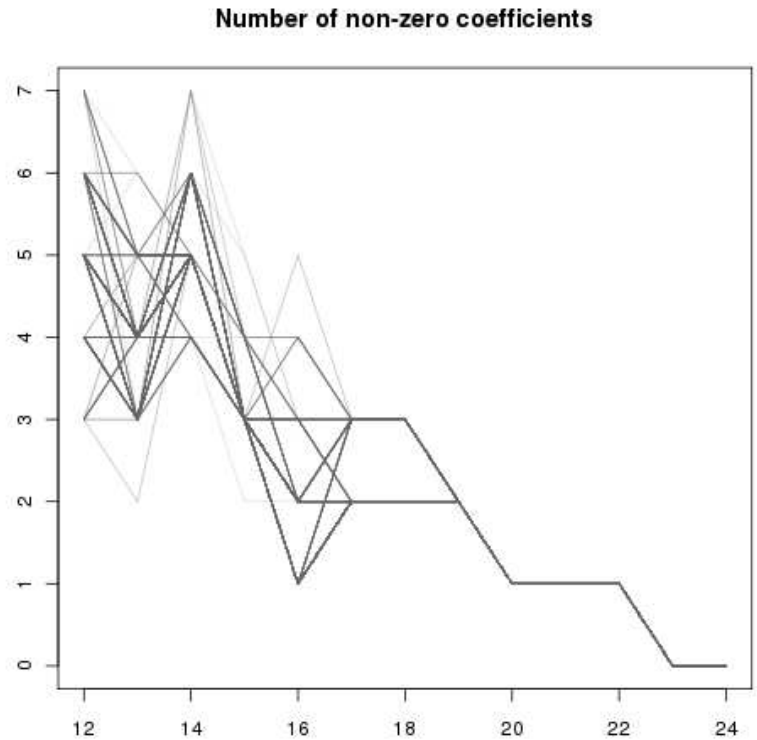
Для каждой пробы будем решать задачу

$$\begin{cases} \left(I_i^k - \sum_j A_{ij} C_j^k\right)^2 \to \min, \\ A_{ij} \geq 0, A_{ij} = 0 \text{ при } BLASTN\left(Gene_j, Probe_i\right) < t. \end{cases}$$

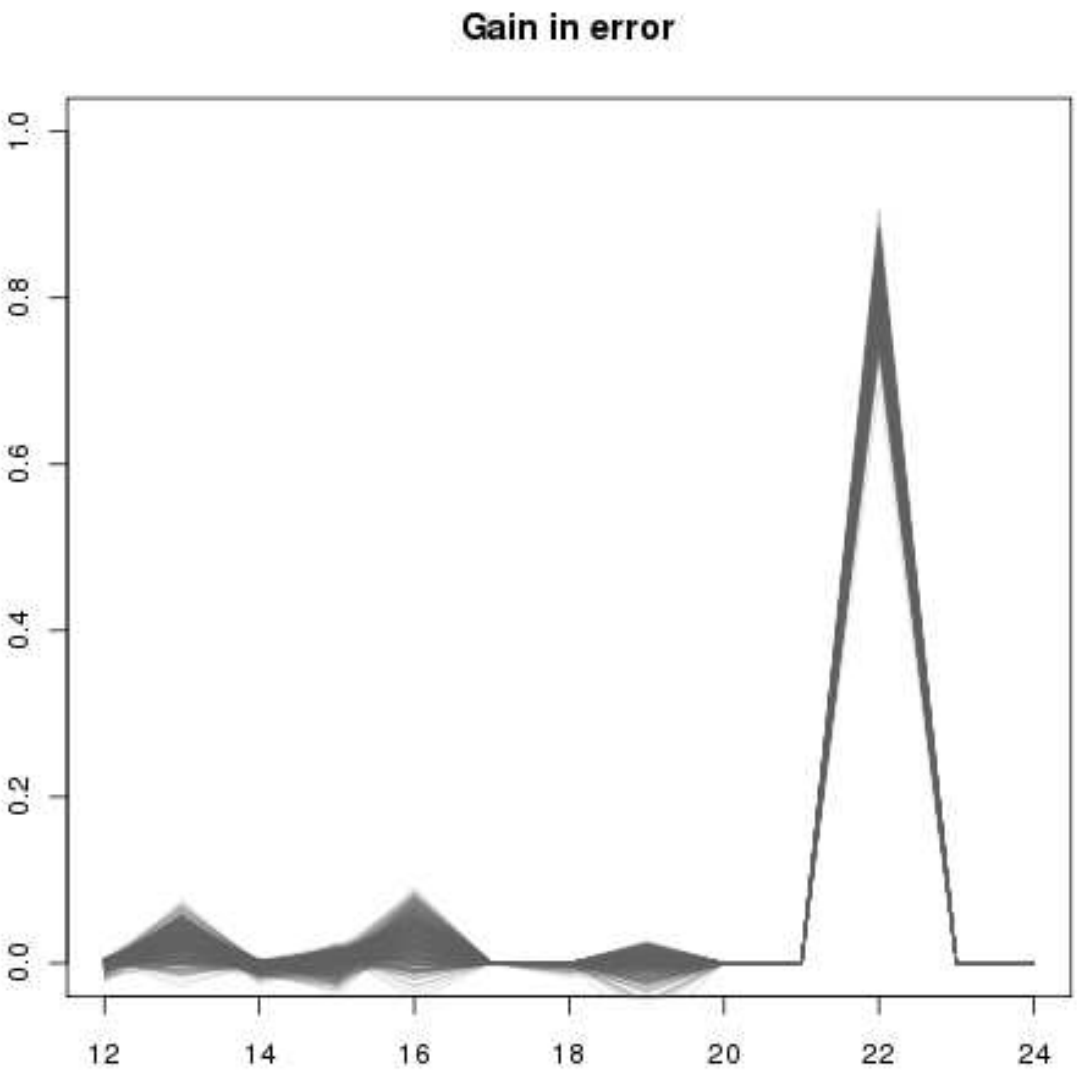
для всех $t = 12, \ldots, 24$ и различных подвыборок затем сравним качество моделей на контроле.

Пример

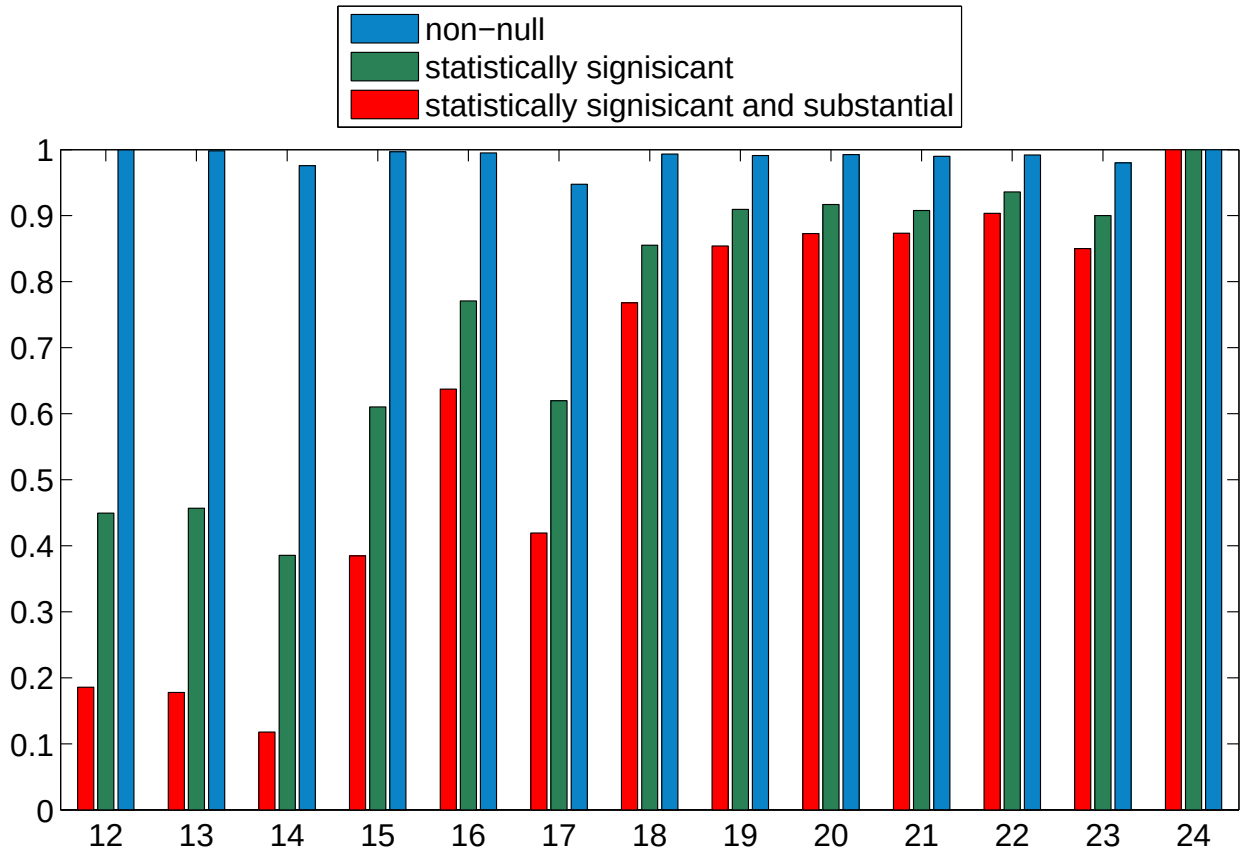

Mean Squared Deviation
0
100
200
300
400
500
600
700
12
14
16
18
20
22
24


Number of non-zero coefficients
0
1
2
3
4
5
6
7
12
14
16
18
20
22
24

Пример



Результат


non−null
statistically signisicant
statistically signisicant and substantial
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
12
13
14
15
16
17
18
19
20
21
22
23
24

Другие виды дополнительной информации

1. Тканеспецифичность для выбора отсутствующих генов.
2. Гены, проб к которым нет на данной модели чипов.
3. Данные spike-in экспериментов, в которых известна концентрация РНК некоторых генов.
4. Технические репликанты — микрочипы, сделанные из одних и тех же образцов.
5. Tissue mixtures — микрочипы, на которые нанесены смеси нескольких образцов в известных пропорциях.
6. * Микрочипы, сделанные из образцов, исследованных одновременно при помощи RNA-Seq.

Итого

- Предложена многопараметрическая модель данных микрочиповых экспериментов и способ сокращения числа параметров, при использовании которого задача их определения становится адекватной.
- Приведён пример использования дополнительной информации о тканеспецифичности генов для исследования свойств данных.