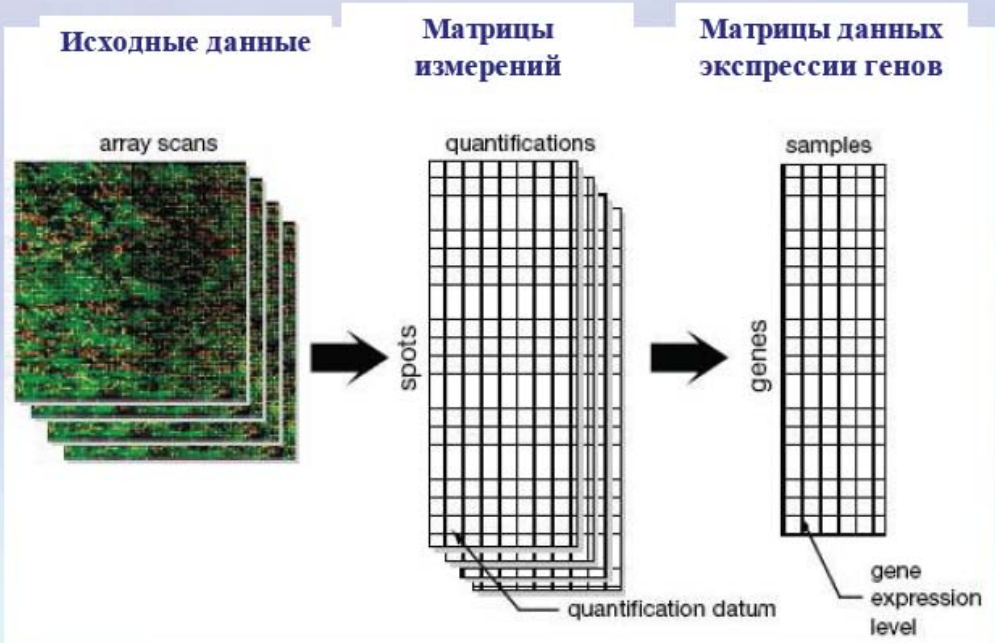
Статистический анализ микрочиповых данных

Ефимов Вадим Михайлович, д.б.н.
Катохин Алексей Вадимович, к.б.н.

Кафедра информационной биологии ФЕН НГУ

Статистический анализ микрочиповых данных

ОБРАБОТКА ДАННЫХ, ПОЛУЧЕННЫХ МЕТОДОМ ДНК-БИОЧИПОВ



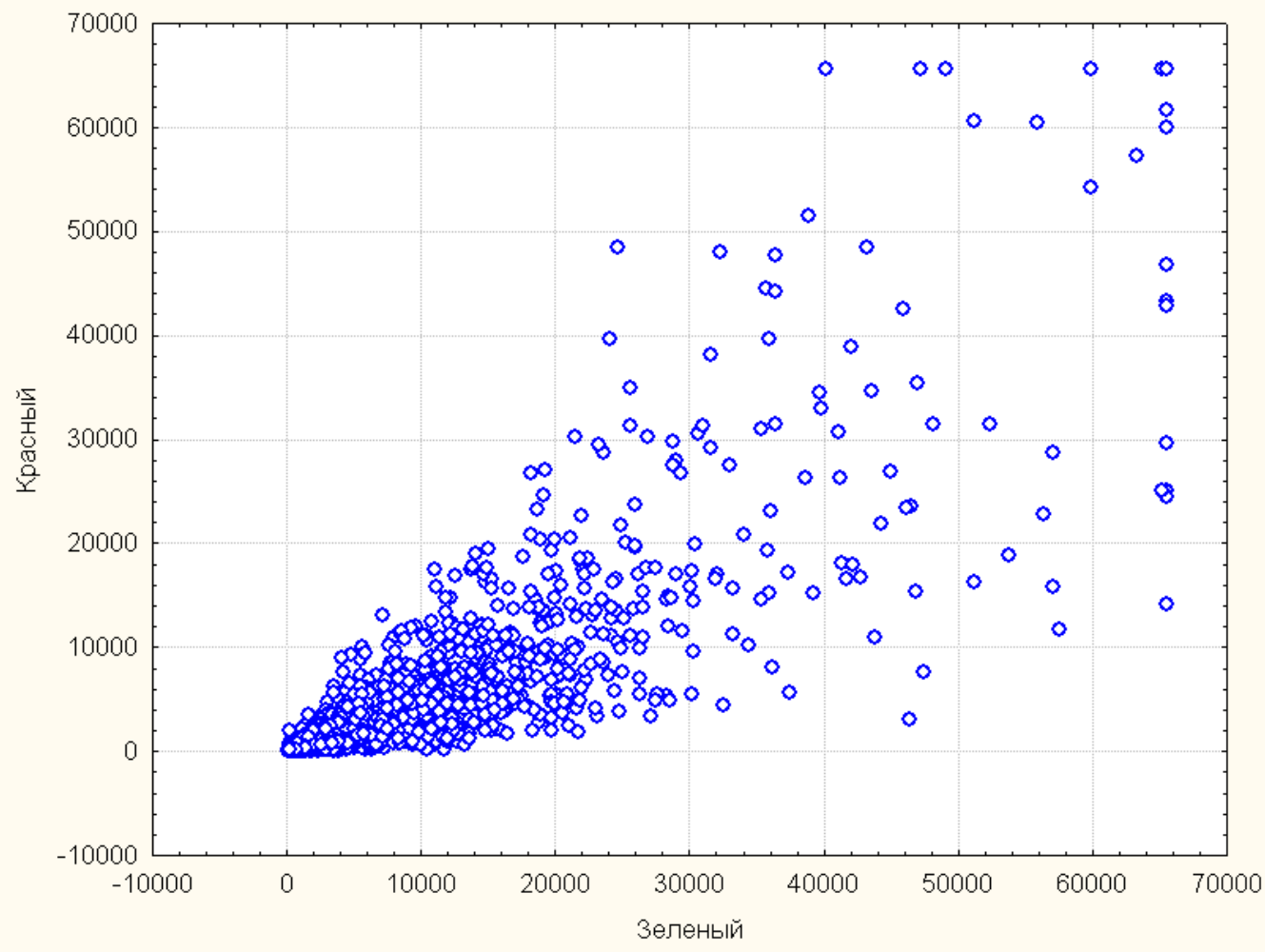
Статистический анализ микрочиповых данных

Фрагмент матрицы данных

Block	Column	Row	Name	ID	OL5636 Cy5	OL5650 Cy5	OL5647 Cy3	OL5238 Cy3
1	1	1	CG_Mm_3000232_1	NM_008080	234	430	760	490
1	2	1	CG_Mm_3000235_1	NM_019735	221	970	1460	524
1	3	1	CG_Mm_3000238_1	NM_008387	288	624	898	506
1	4	1	CG_Mm_3000253_1	J03880	4009	9151	8537	8959
1	5	1	CG_Mm_3000256_1	NM_016893	356	1085	1259	956
1	6	1	CG_Mm_3000259_1	U89924	109	114	287	230
1	7	1	CG_Mm_3000262_1	AF229644	162	369	636	408
1	8	1	CG_Mm_3000277_1	NM_009178	253	663	1290	543
1	9	1	CG_Mm_3000280_1	NM_015828	197	348	564	387
1	10	1	CG_Mm_3000283_1	NM_010293	167	256	437	360
1	11	1	CG_Mm_3000286_1	NM_008194	125	259	472	286
1	12	1	CG_Mm_3000589_1	NM_011561	5049	6659	10409	10063
1	13	1	CG_Mm_3000592_1	NM_016925	495	976	1777	765
1	14	1	CG_Mm_3000595_1	AB041576	2309	3271	4836	3126
1	15	1	CG_Mm_3000598_1	AF176524	178	654	889	314

Статистический анализ микрочиповых данных

Визуализация исходны данных



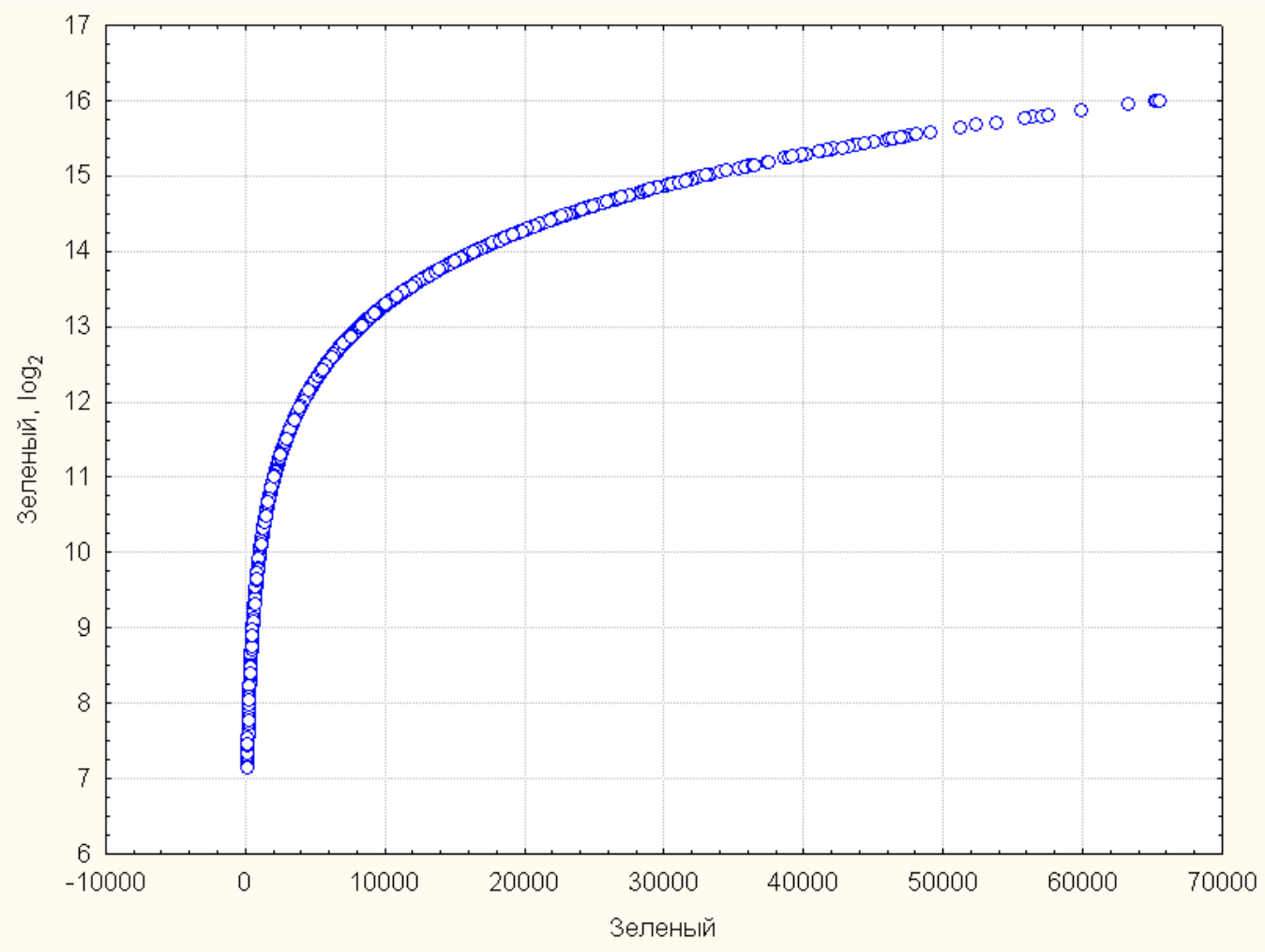
Статистический анализ микрочиповых данных

Этапы анализа микрочиповых данных

1. Логарифмирование
2. Нормализация
3. Нахождение дифференциально экспрессирующихся генов
4. Многомерный анализ

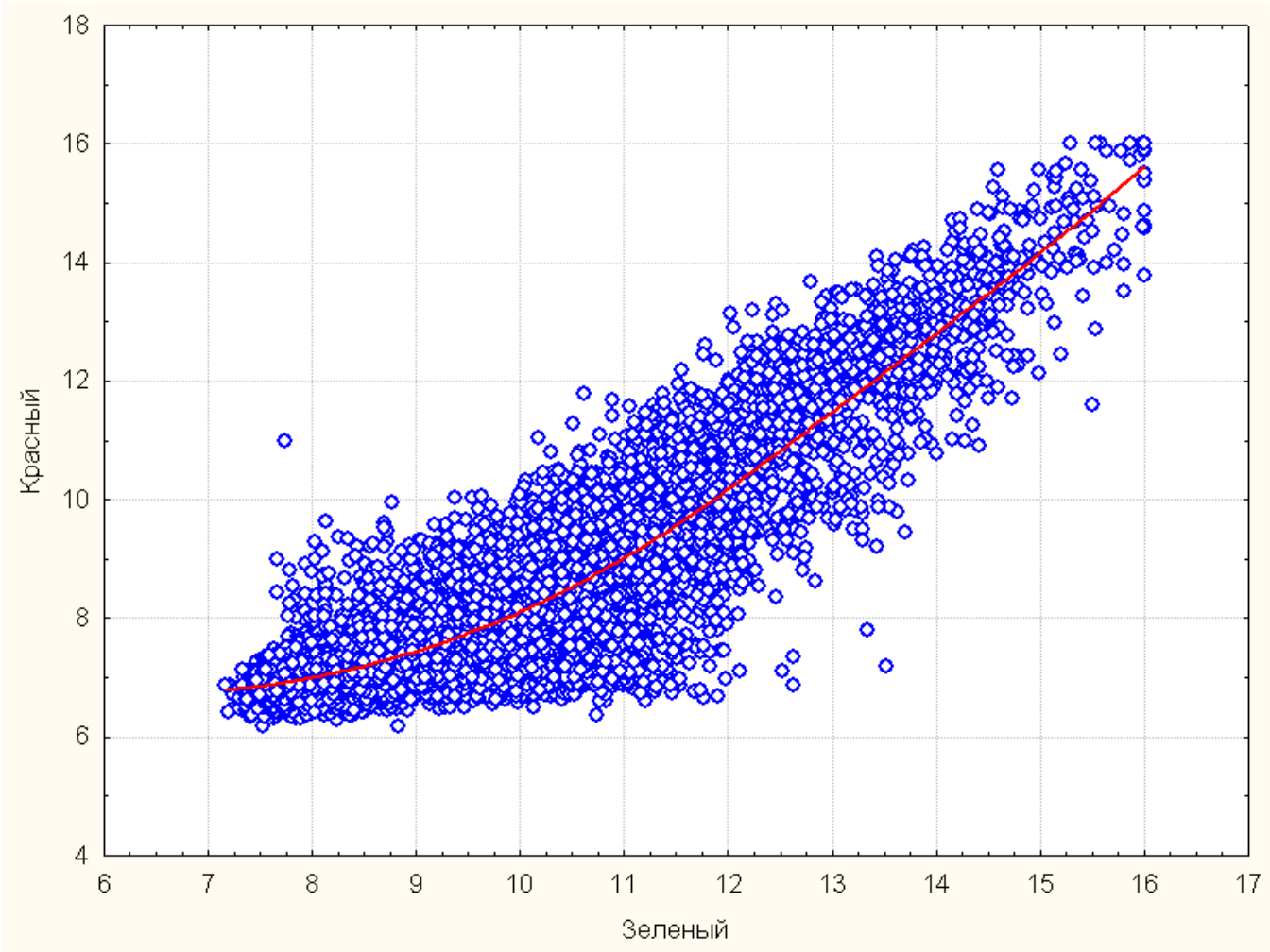
Статистический анализ микрочиповых данных

Предобработка: Логарифмирование



Статистический анализ микрочиповых данных

Визуализация логарифмированных данных



Поворот

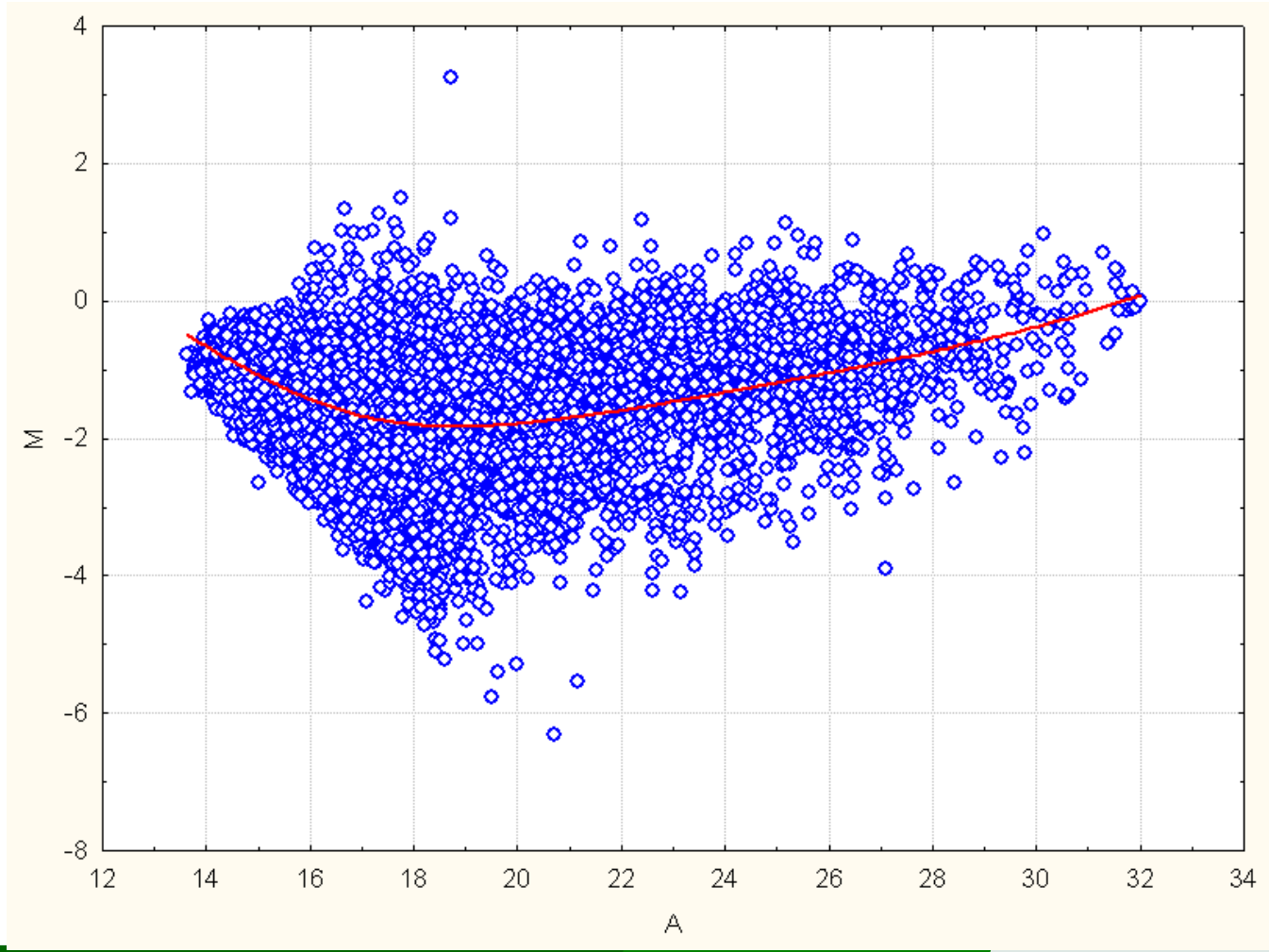
$$M = \log_2 R - \log_2 G$$

$$A = \log_2 R + \log_2 G$$

R – red, красный
G – green, зеленый

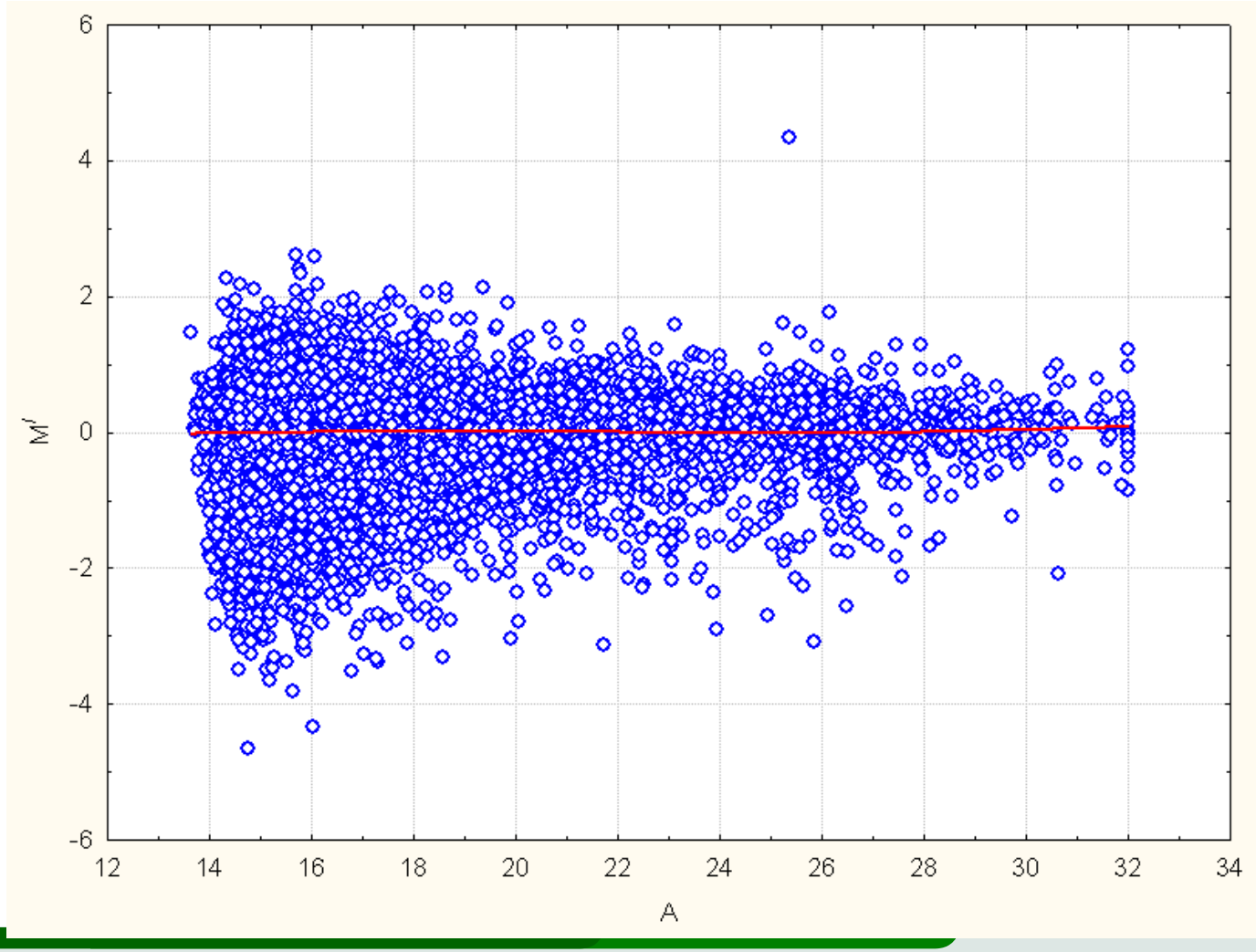
Статистический анализ микрочиповых данных

Локально взвешенная (lowess) регрессия



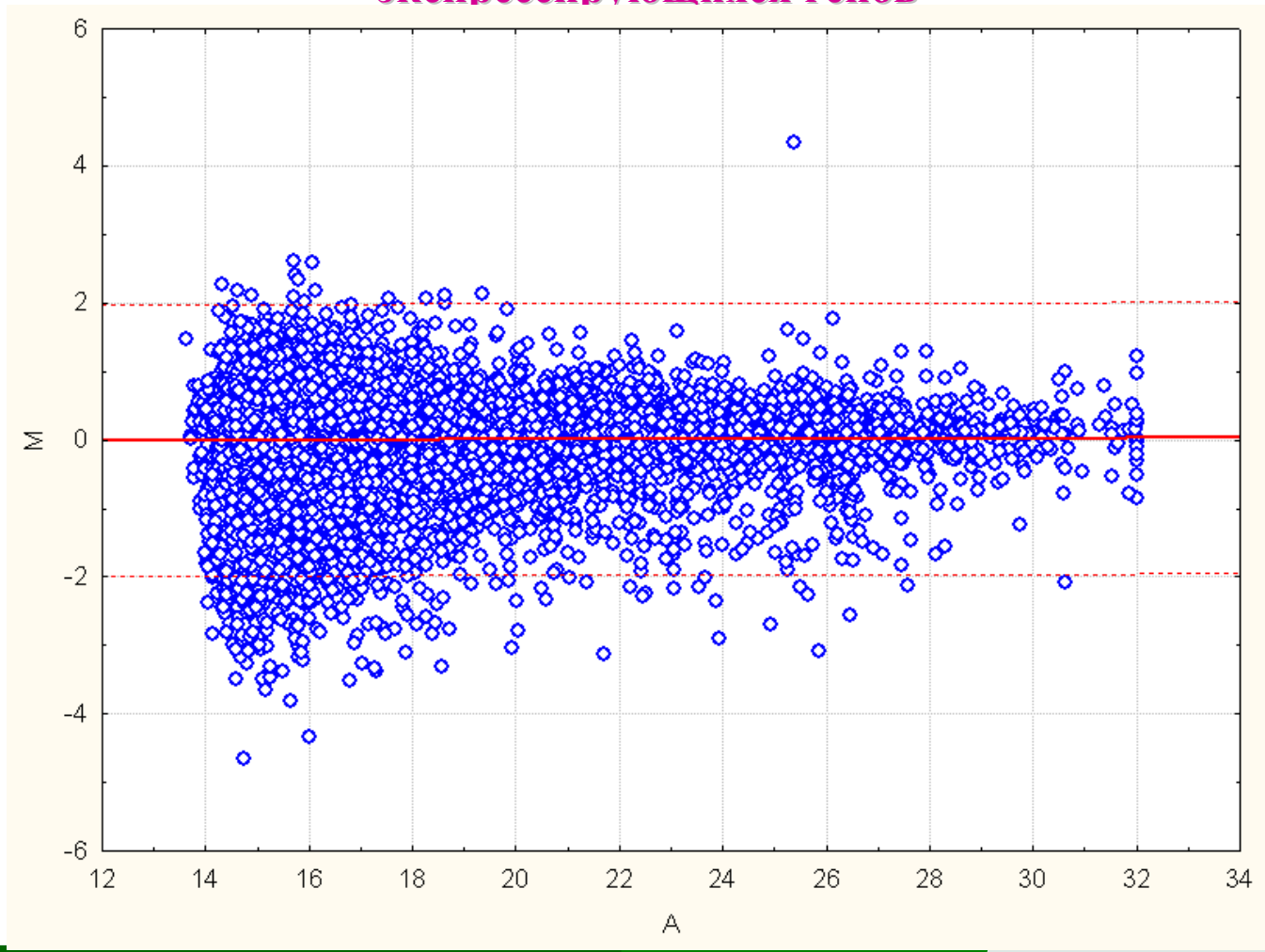
Статистический анализ микрочиповых данных

Локально взвешенная (lowess) регрессия - результат



Статистический анализ микрочиповых данных

Нахождение дифференциально экспрессирующихся генов



Обратный поворот

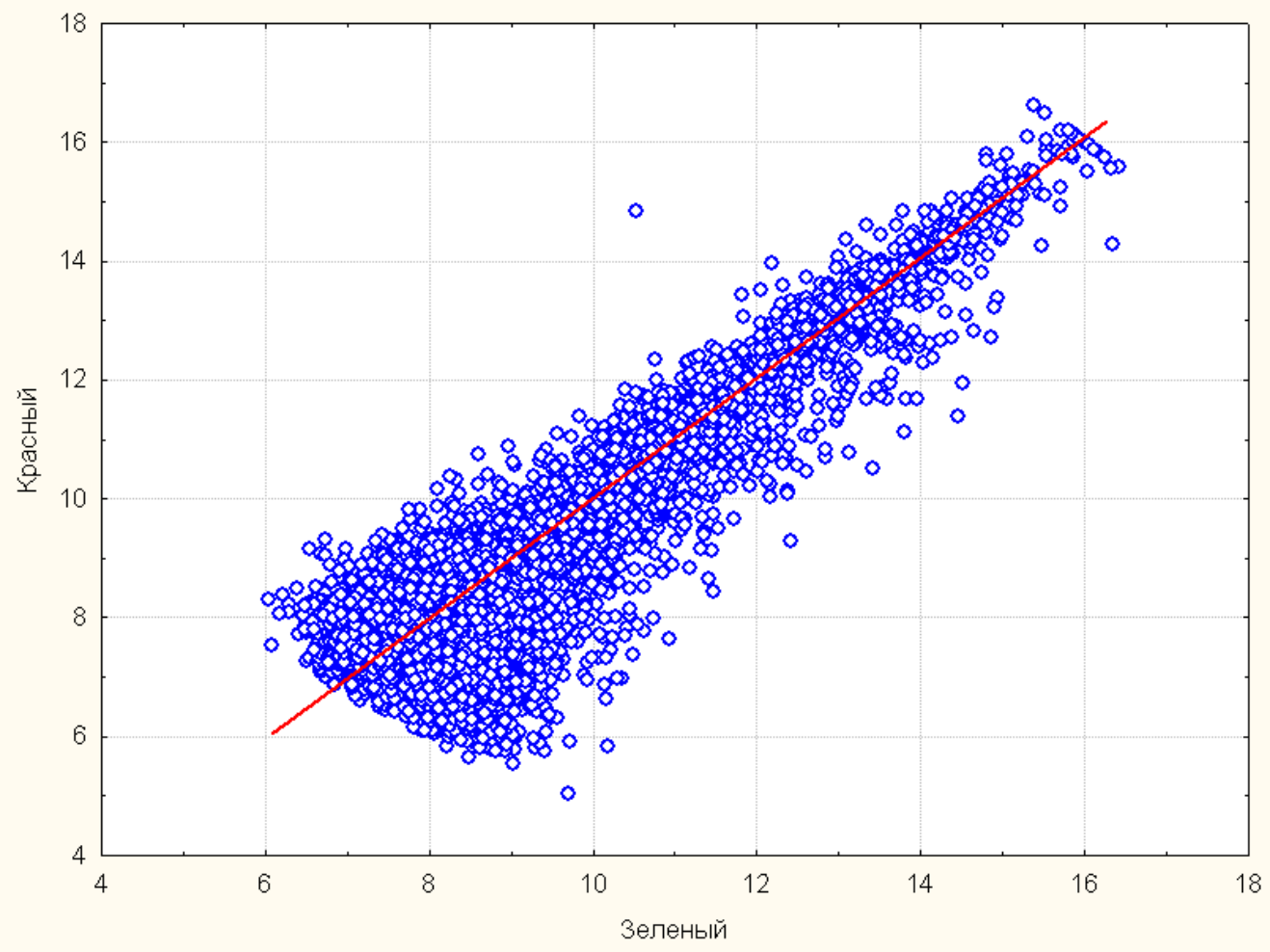
$$log_2R = (A+M)/2$$

$$log_2G = (A-M)/2$$

R – red, красный
G – green, зеленый

Статистический анализ микрочиповых данных

Нормализация (логарифмический масштаб)



Статистический анализ микрочиповых данных

Многомерный анализ: фрагмент матрицы данных

		normal	normal	normal	normal	HD	HD	HD	HD
ID_REF	IDENTIFIER	GSM30580	GSM30581	GSM30582	GSM30583	GSM30538	GSM30539	GSM30540	GSM30541
201012_at	NM_000700	267.2	420.8	277.7	316.7	801.2	628.9	615.1	627.8
200989_at	NM_001530	192.5	205.4	470.8	159.9	881.6	930.6	708.1	572.9
214578_s_at	NM_005406	115.8	76.4	61.7	180.5	167	132.2	213.3	138.3
201023_at	NM_005642	50.1	92.6	129.8	71.1	205.8	285.3	224.2	171.9
218589_at	NM_005767	113.7	62.1	82.6	83.5	219.4	277.8	142	271.9
208374_s_at	NM_006135	161.1	235.9	427	152.3	979.1	1130.9	1091	917.9
217392_at	NM_006135	3.8	6.8	1.9	10.2	5.7	3	3.5	2.9
222300_at	NM_006135	9.8	33.4	36.4	125.3	59.3	42.8	51.5	47.1
201070_x_at	NM_012433	50	53.4	121.1	106.7	121.9	125.7	140.1	147.6
201071_x_at	NM_012433	282.1	357.6	414.2	257.9	1047.9	657.2	975	1025
214305_s_at	NM_012433	61.9	42.4	52.8	87.5	116.2	70.4	147.9	83.6
212287_at	NM_015355	74.2	95.7	80	23.2	252.3	202.3	291.2	254.1
213971_s_at	NM_015355	3.9	11.9	8.2	26	19	9	5.8	23
217783_s_at	NM_016061	260	224.6	255.5	146.6	511.4	422	502.1	345
217816_s_at	NM_020357	27.8	34.3	70.8	21.7	252.6	268.4	251.3	312.1
202653_s_at	NM_022826	48.1	76.3	121.8	59.7	431	416.2	426.4	275.2

Стандартизация: центрирование и нормирование

Если мы умножим значения любого количественного признака на любую ненулевую константу и прибавим к ним любую константу, то это никак не изменит относительных расстояний между объектами по этому признаку. Поэтому мы можем использовать преобразования сдвига и масштаба для приведения разных признаков в соответствие друг с другом. Преобразование:

$$x_i' = (x_i - x^*)$$

где $x^* = \Sigma x_i'/N$ – среднее значение, называется *центрированием*. После центрирования новое среднее признака равно 0 ($x'^* = \Sigma x_i'/N = 0$). Преобразование:

$$x_i' = x_i/s$$

где $s^2 = \Sigma(x_i - x^*)^2/N$ – дисперсия признака (вместо N часто применяется $N-1$), называется *нормированием*. После такого преобразования все признаки становятся безразмерными, а новая дисперсия равна 1 ($s'^2 = \Sigma(x_i' - x'^*)^2/N = 1$).

Мера различий – евклидово расстояние

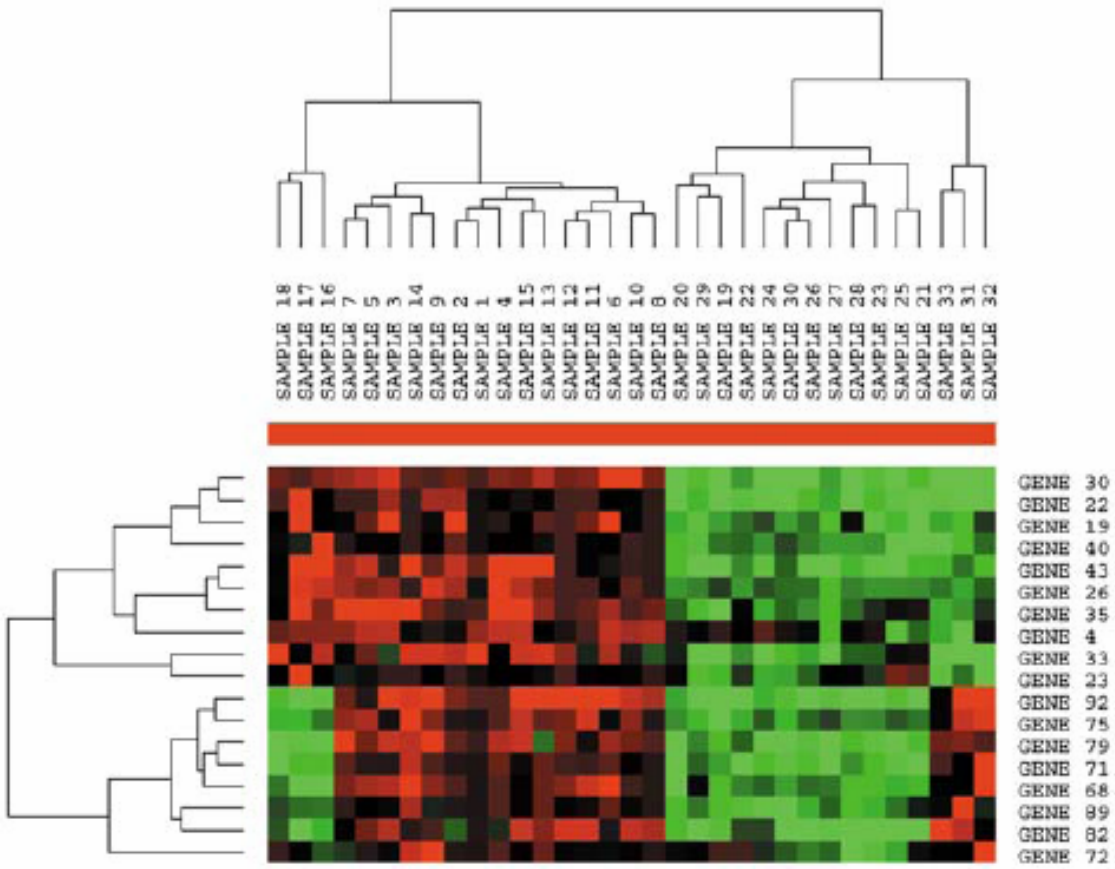
Если мы хотим работать с дистанциями, то за меру различия между профилями можно взять евклидово расстояние между строками, т.е. корень из суммы квадратов покоординатных разностей. После несложных преобразований получим

$$D=\sqrt{\sum(x_i-y_i)^2}=\sqrt{\sum x_i^2+\sum y_i^2-2\sum x_i y_i}=\sqrt{2-2r}$$

т.е., евклидово расстояние между строками является в этом случае монотонной функцией коэффициента корреляции. Чем выше коэффициент корреляции, тем меньше расстояние. Если $r=1$, то $D=0$, как и следовало ожидать. Максимальное расстояние $D=2$ достигается при $r=-1$. Следовательно, результаты кластерного анализа для методов типа ближайшего или дальнего соседа, где используются только сравнения меры сходства-различия, и при выборе коэффициента корреляции и при выборе евклидова расстояния будут совпадать. Для методов, где вычисляется средняя мера сходства-различия, результаты могут несколько различаться. Однако, если в качестве меры различия выбрать квадрат евклидова расстояния $D^2=2(1-r)$ (но заметим, что он уже не является евклидовым расстоянием), то результаты кластеризации будут совпадать и для усредняющих меру сходства-различия методов

Статистический анализ микрочиповых данных

Термокарты (heatmaps)
и иерархическая кластеризация



Статистический анализ микрочиповых данных

Термокарты
и иерархическая классификация (2)

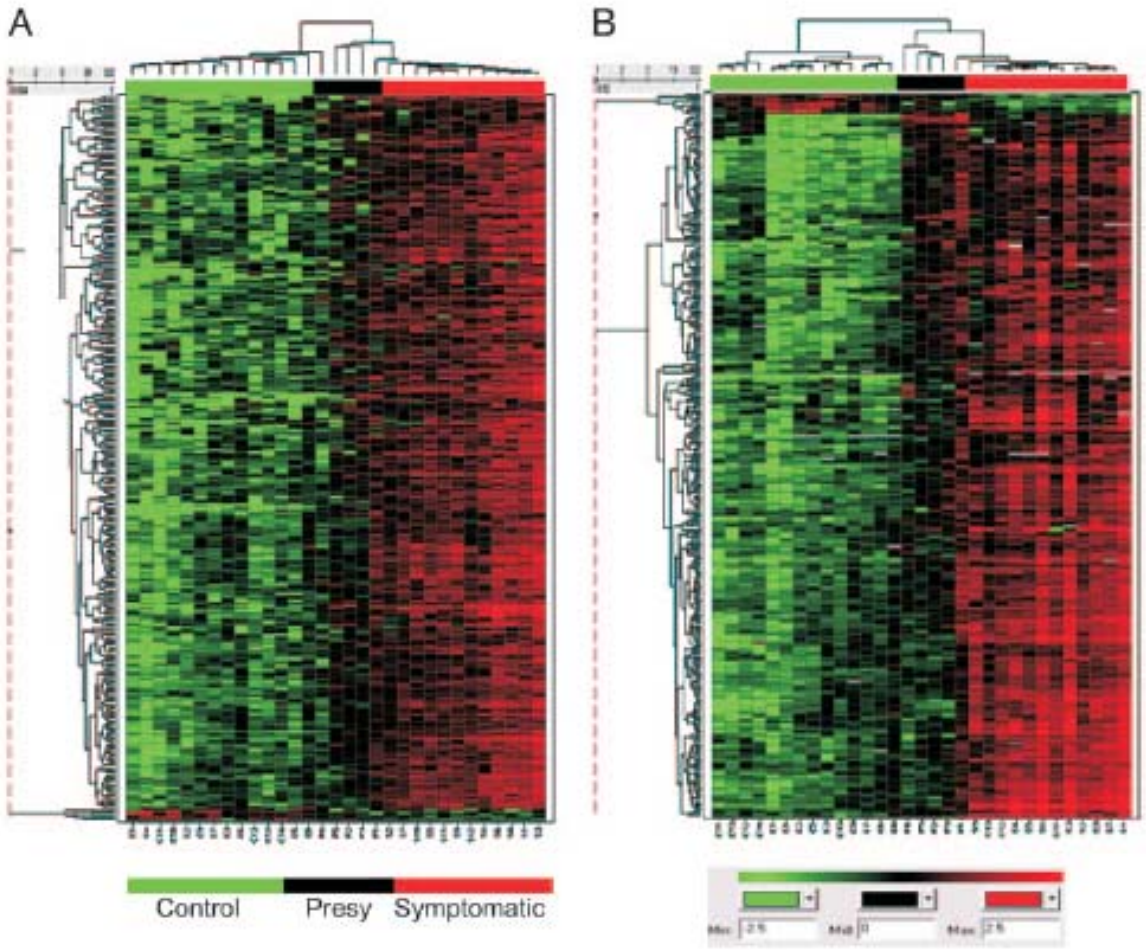
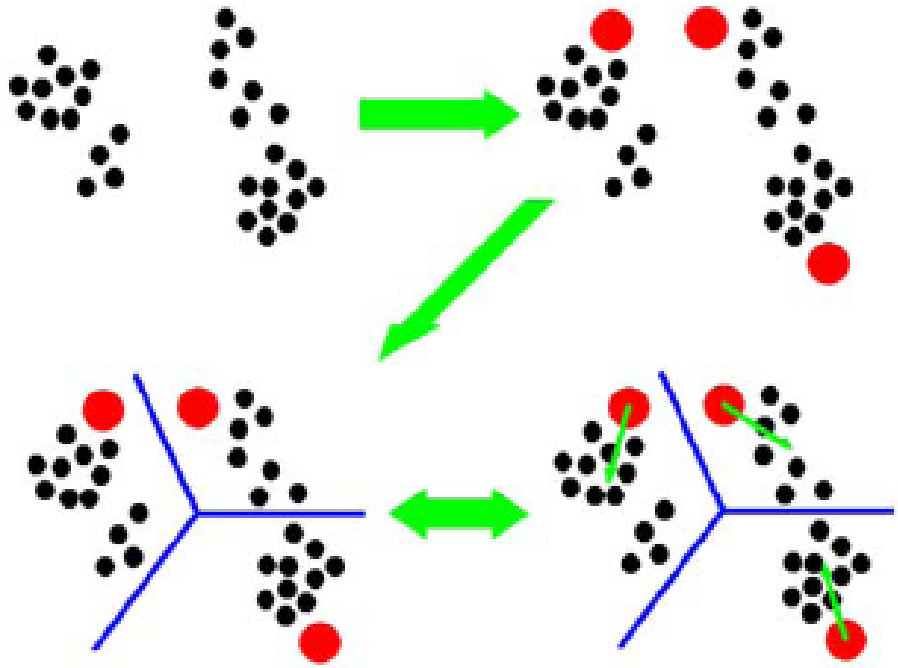


Fig. 1. Altered gene expression in the blood of HD patients. Cluster analysis of the 322 most differentially expressed genes on Affymetrix (*A*) and Amersham Biosciences (*B*) microarrays is shown. The genes were selected from 17 HD-affected subjects and 14 healthy control subjects according to *P* value ($P < 0.0005$), fold change (>1.8 or <0.6), and expression maximum >100 (Affymetrix) or >1 (Amersham Biosciences). Each column represents a sample and each row a gene. The colorgram depicts high (red) and low (green) relative levels of gene expression. The samples were normalized (median-polished) for each platform. Hierarchical clustering using cosine correlation with complete linkage was performed on the pool of all samples from both platforms to determine sample and gene clustering. The two groups were then separated in the display to compare the gene profiles between the two platforms (healthy control subjects C1–C14, late presymptomatic carriers of the HD mutation P1–P5, and symptomatic HD patients S1–S12).

Алгоритм К-средних



Статистический анализ микрочиповых данных

Изображение одного кластера

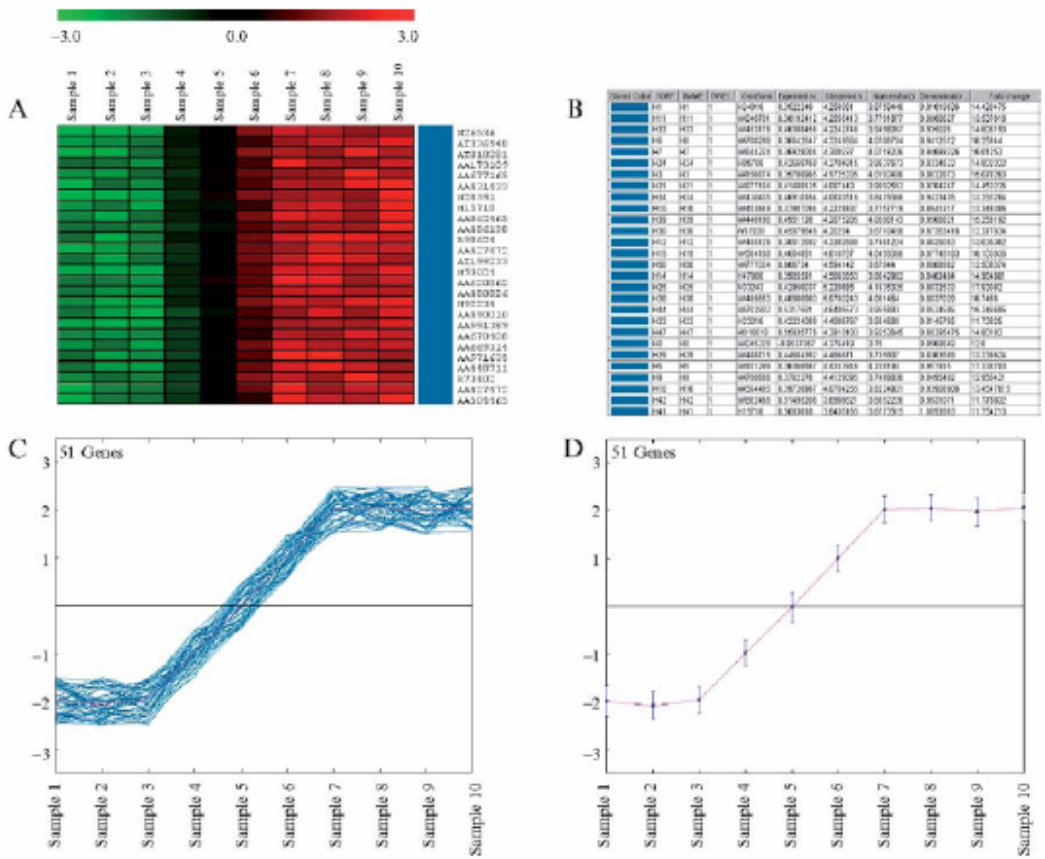


FIG. 13. Cluster viewer examples, (A) Expression image, (B) cluster table viewer, (C) expression graph, and (D) centroid graph.

Статистический анализ микрочиповых данных

Изображение нескольких кластеров

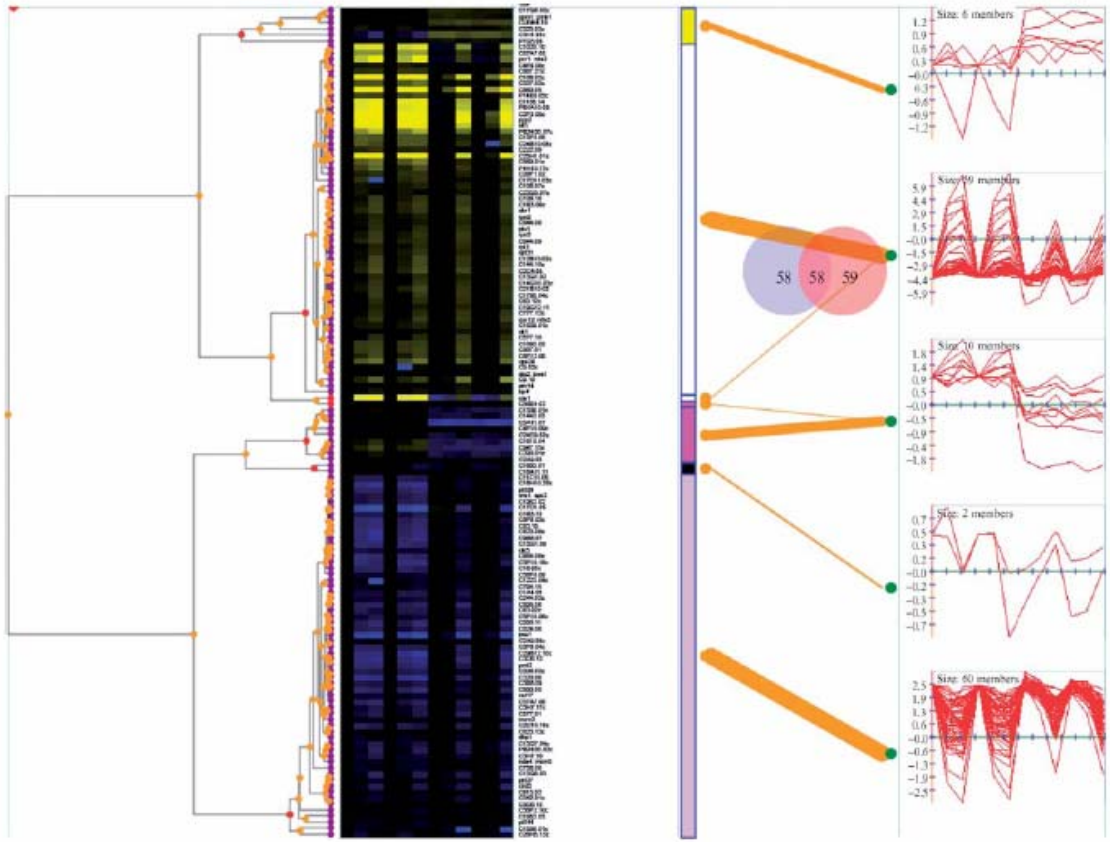
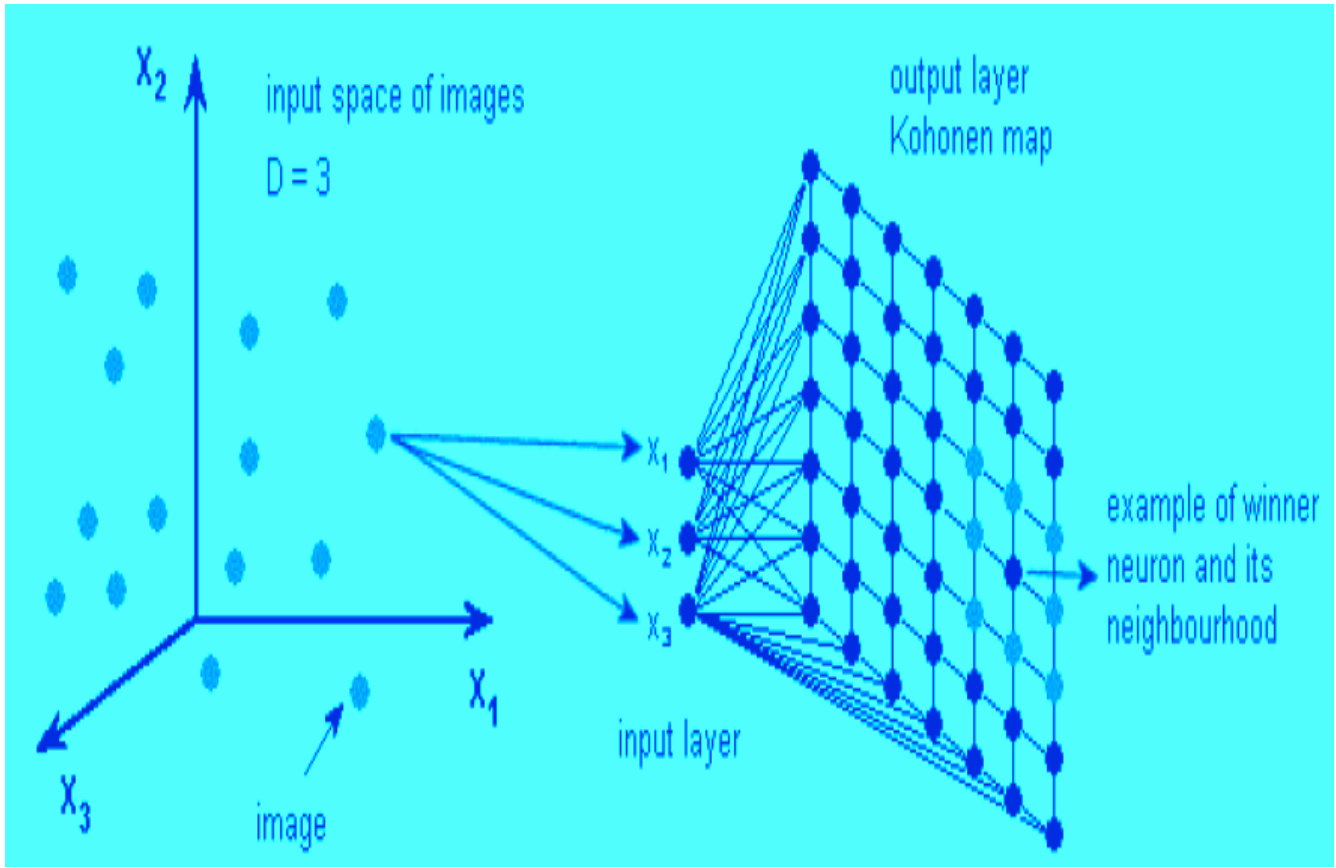


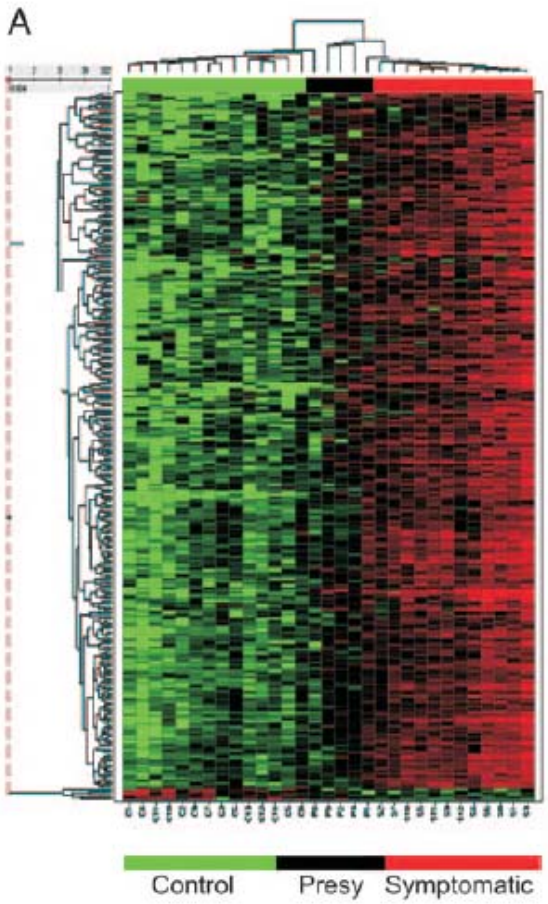
FIG. 8. Expression Profiler clustering comparison visualization. A hierarchical flat comparison, matching a K-means clustering ($K = 5$) to a dendrogram. Overlaps between matching clusters are shown with Venn diagrams.

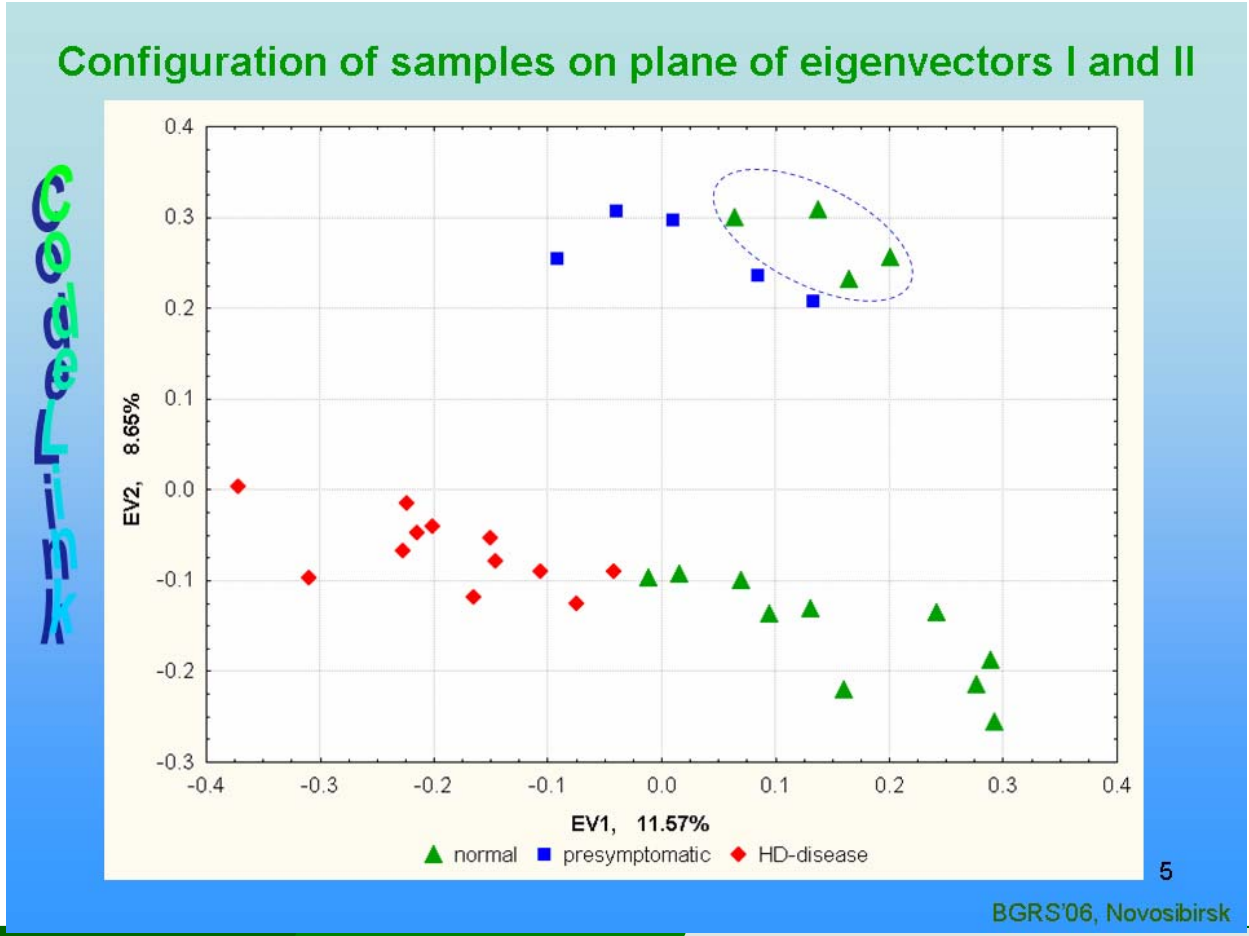
Самоорганизующиеся карты Кохонена (Self-Organizing Maps – SOM)



Статистический анализ микрочиповых данных

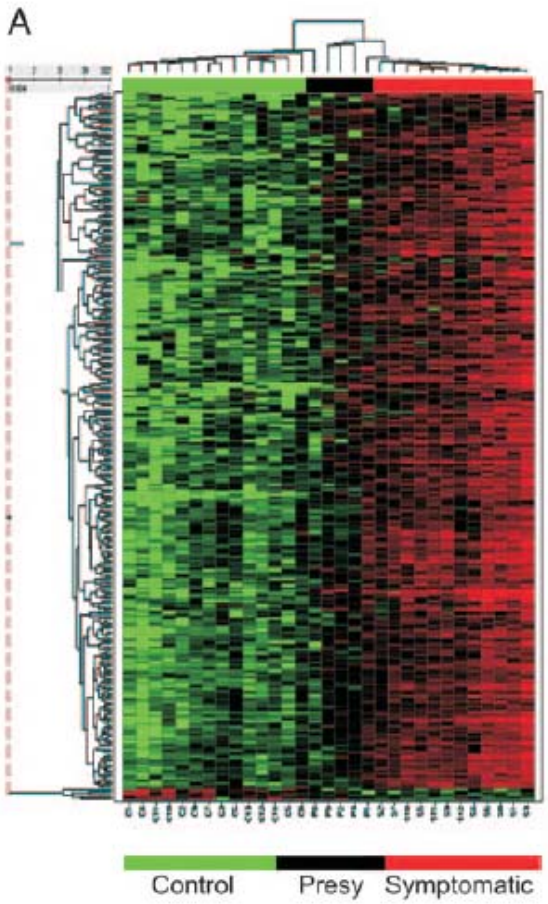
Метод главных компонент 1
(Principal component analysis – PCA)

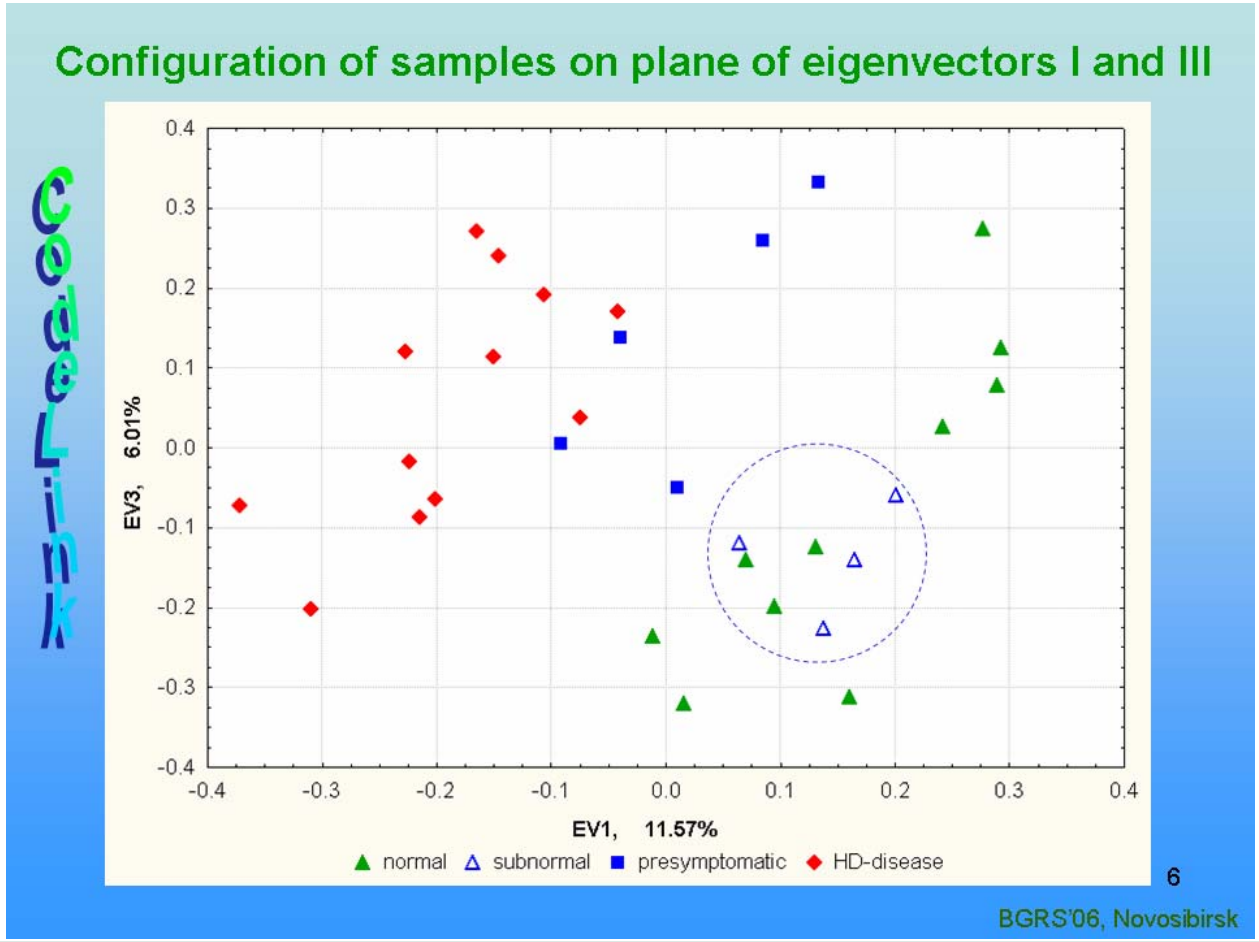




Статистический анализ микрочиповых данных

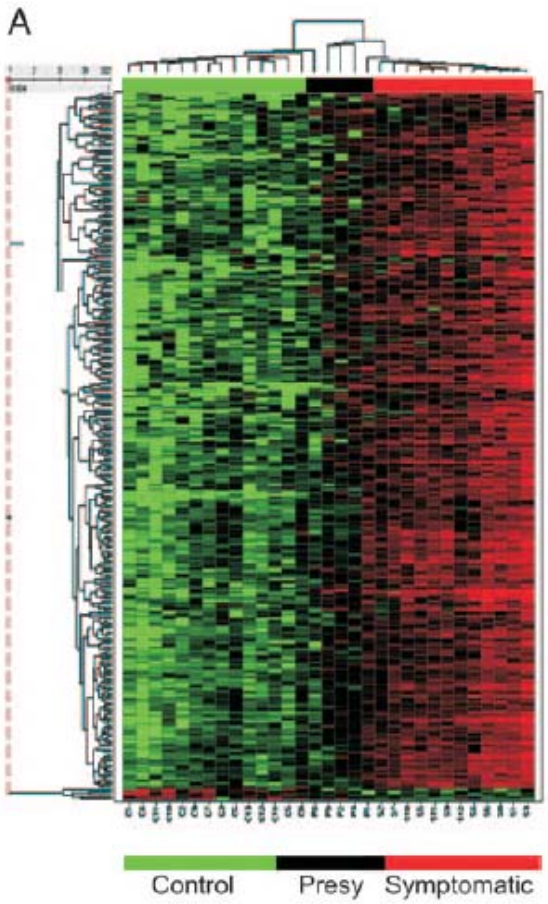
Метод главных компонент 2

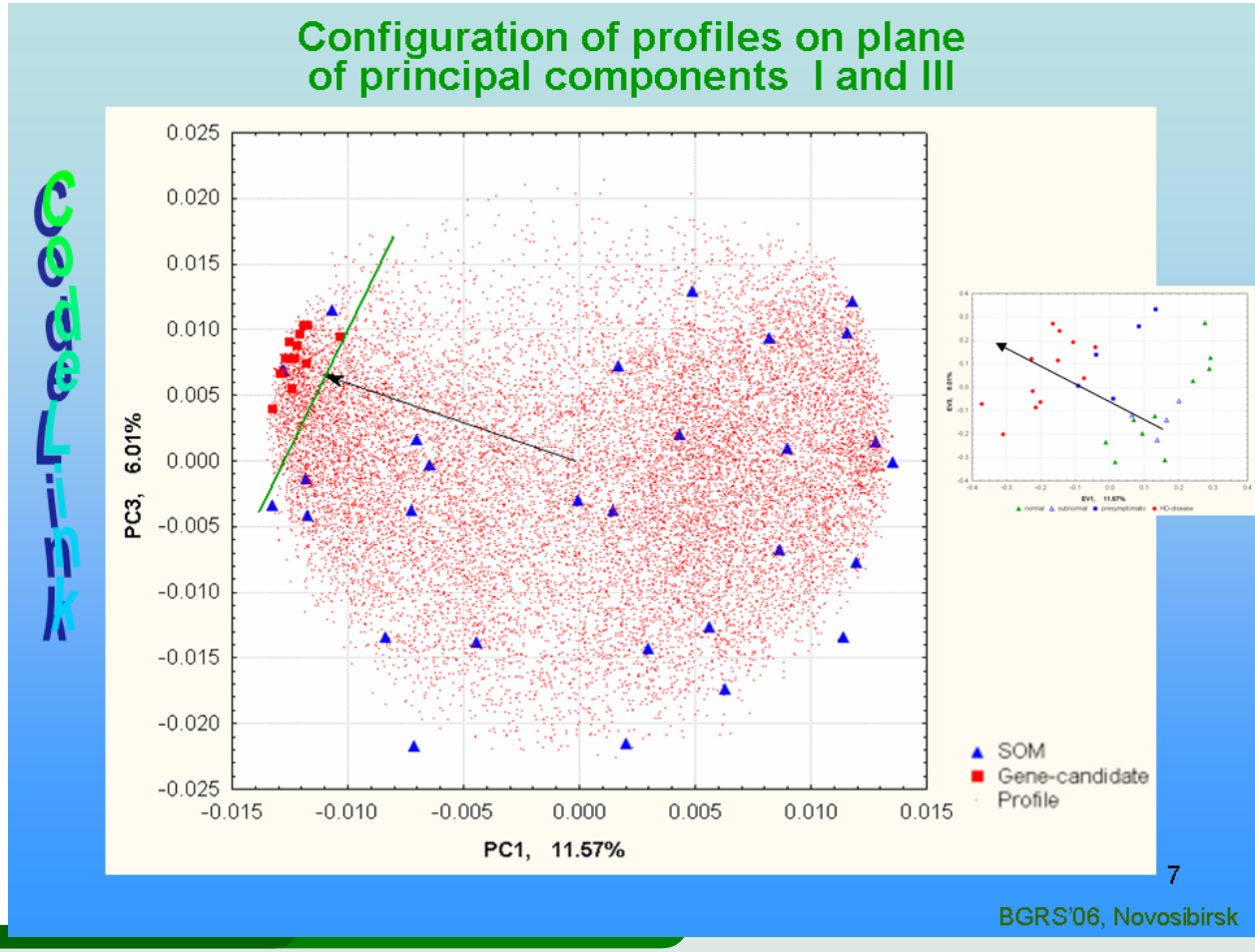




Статистический анализ микрочиповых данных

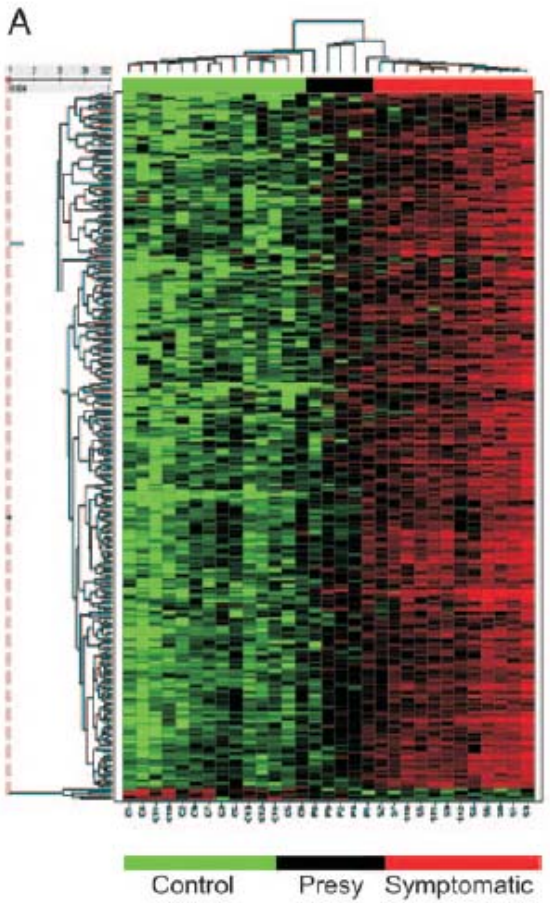
Метод главных компонент 3

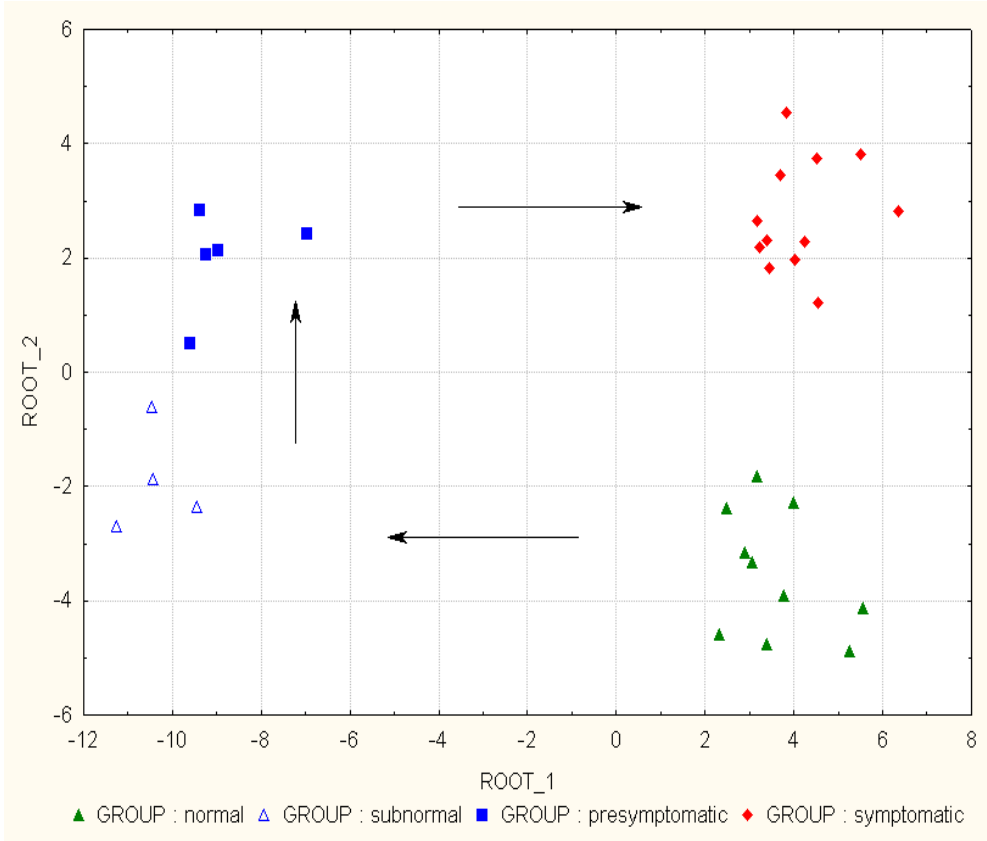




Статистический анализ микрочиповых данных

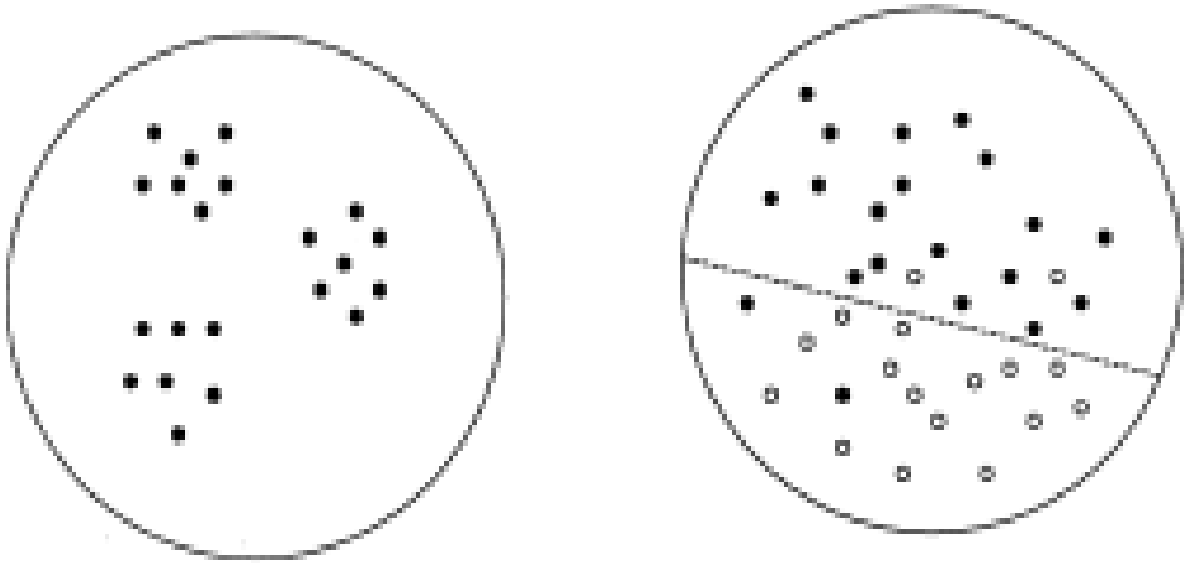
Многомерное шкалирование (Multidimensional scaling - MDS)





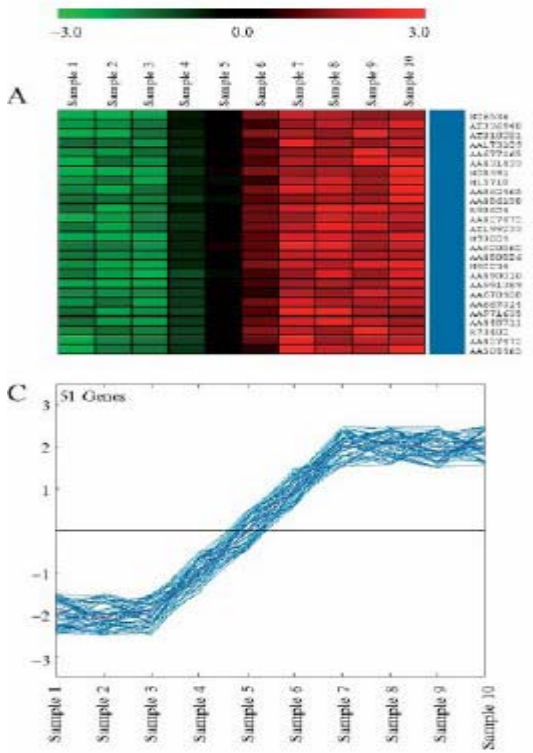
Статистический анализ микрочиповых данных

«Обучение с учителем» (supervised data analysis)



Статистический анализ микрочиповых данных

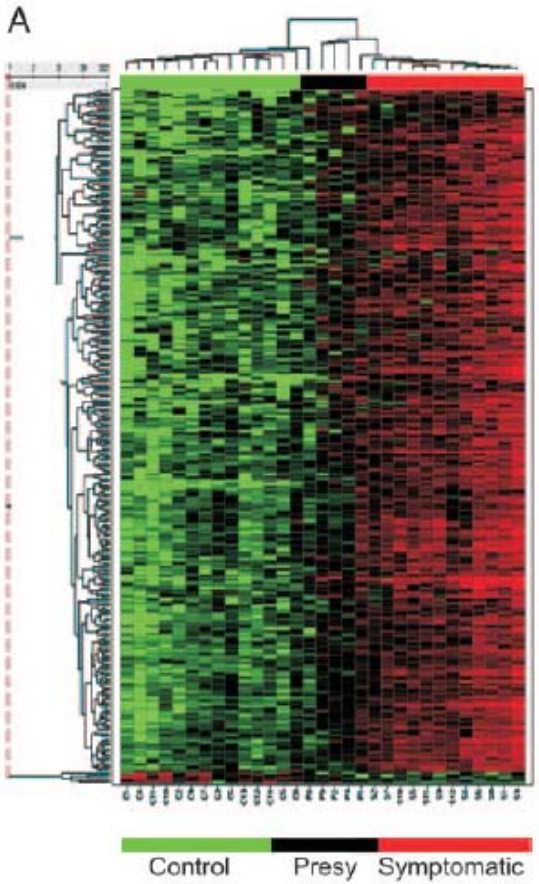
t-критерийУэлша

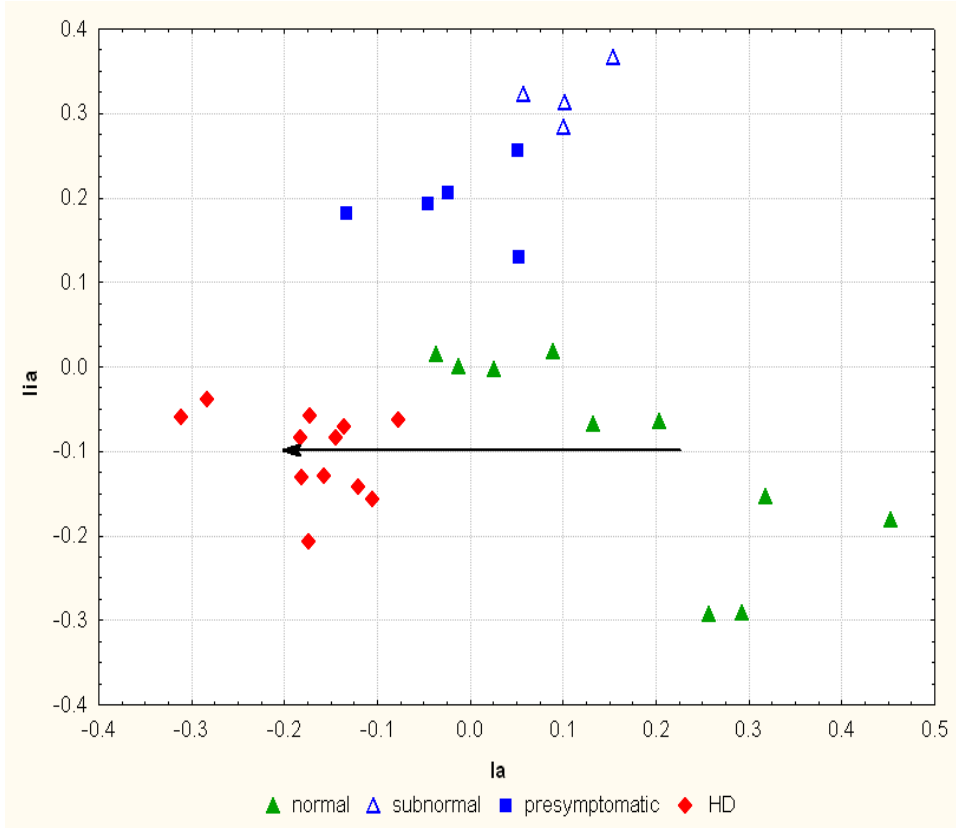


$$t = \frac{|X_i - X_j|}{\sqrt{\frac{s_i^2}{n_i} + \frac{s_j^2}{n_j}}}$$

Статистический анализ микрочиповых данных

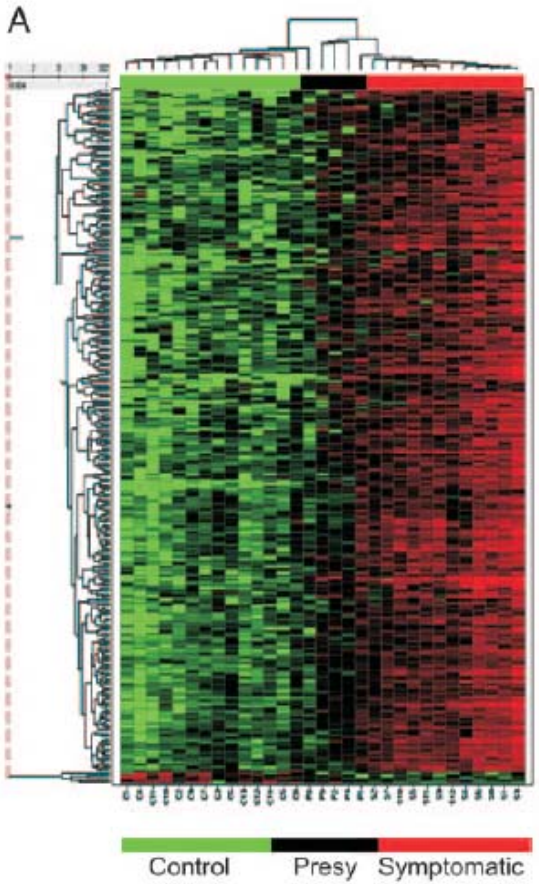
Дискриминантный анализ 1 (Discriminant analysis - DA)

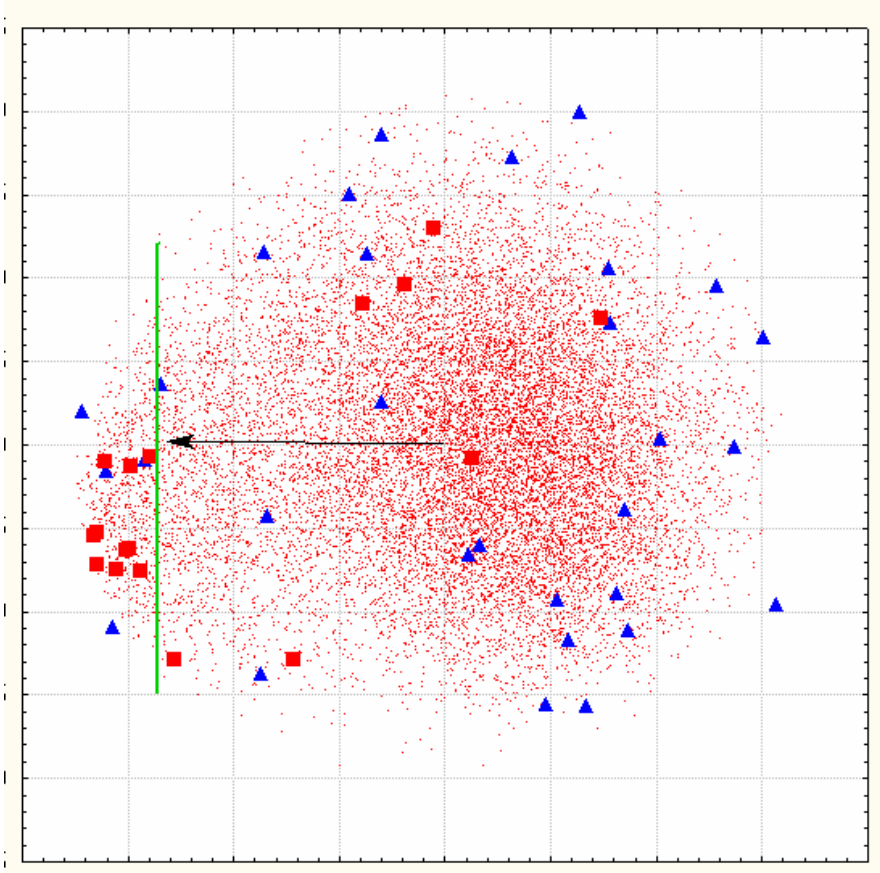


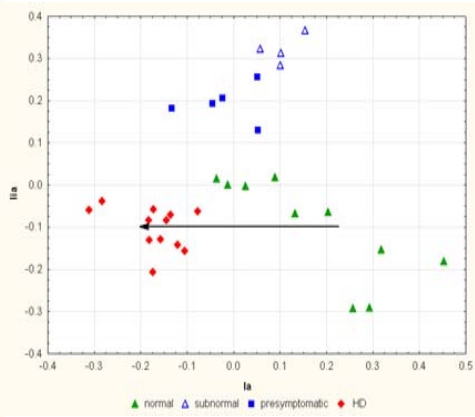


Статистический анализ микрочиповых данных

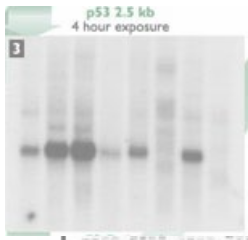
Дискриминантный анализ 2

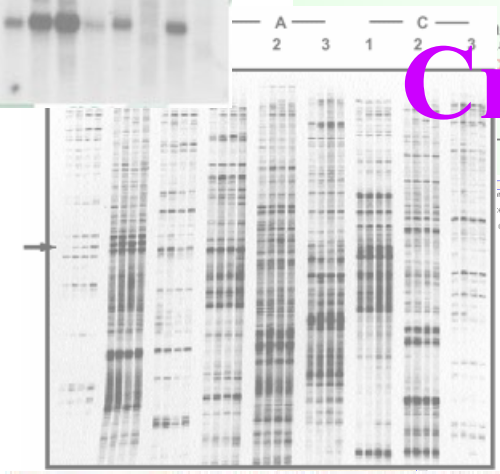


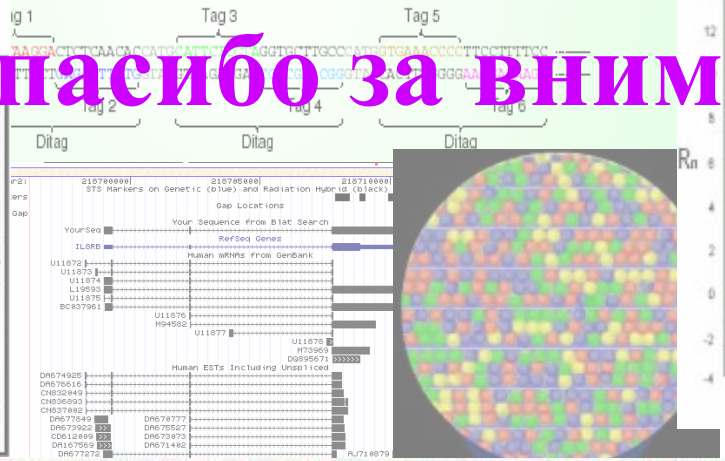




Статистический анализ микрочиповых данных







Спасибо за внимание