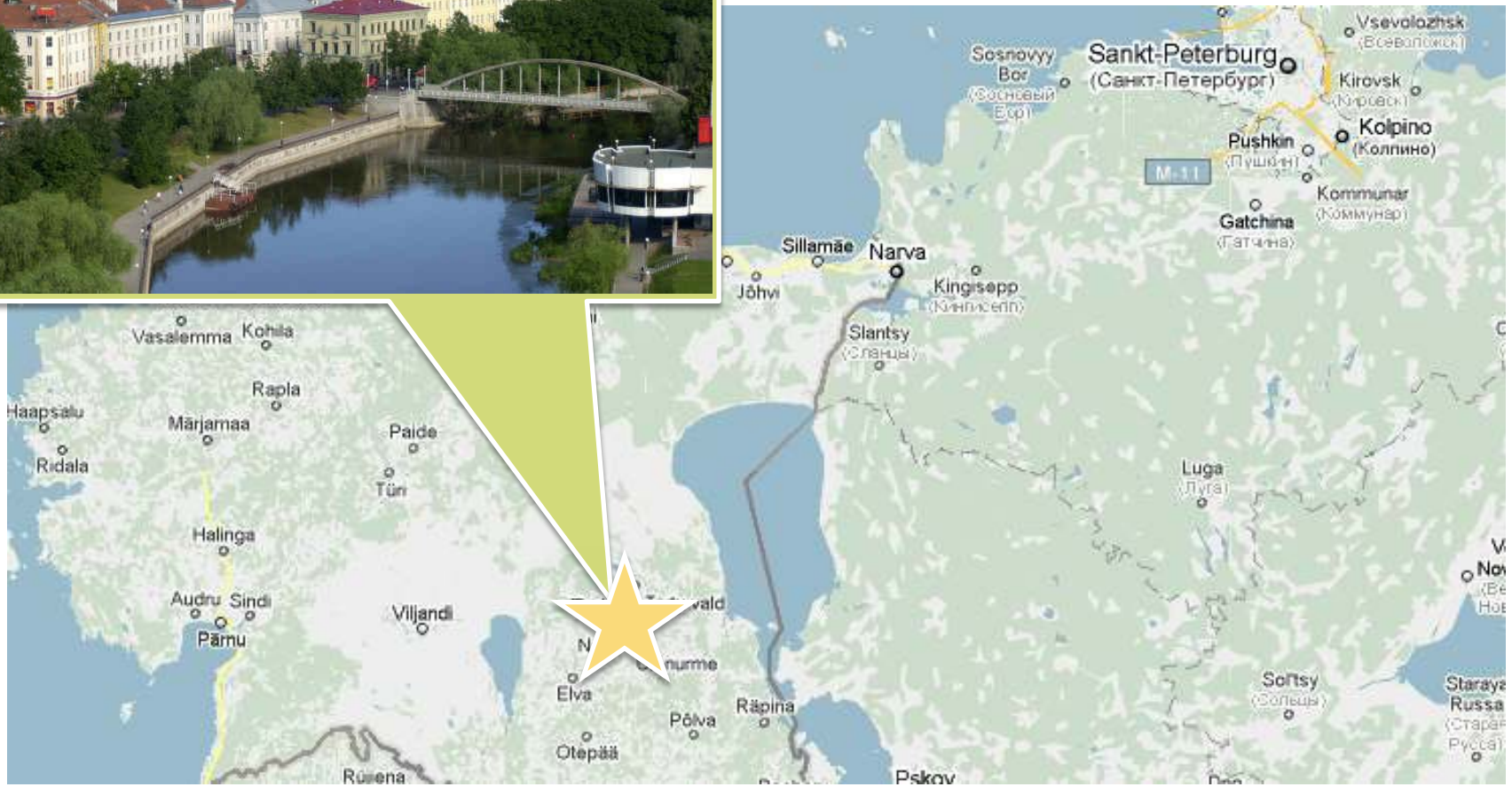
# Анализ экспрессии генов

Константин Третьяков (kt@ut.ee)

**Тартуский Университет**
BIIT Research group (http://biit.cs.ut.ee)

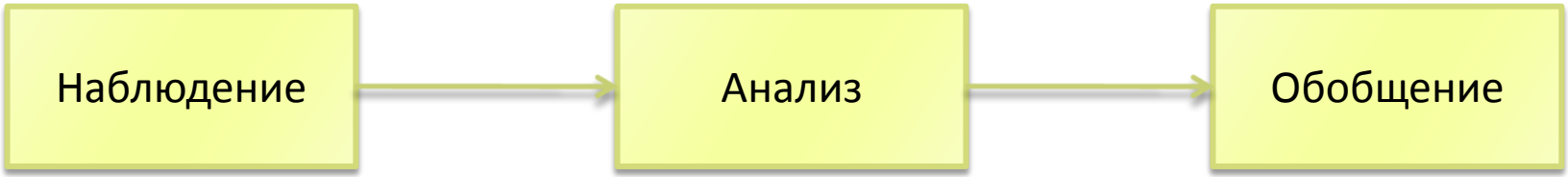

Sankt-Peterburg
(Санкт-Петербург)
Sosnovyy Bor
(Сосновый Бор)
Vsevolozhsk
(Всеволожск)
Kirovsk
(Кировск)
Kolpino
(Колпино)
Pushkin
(Пушкин)
M-11
Kommunar
(Коммунар)
Gatchina
(Гатчина)
Sillamäe
Narva
Jõhvi
Kingisepp
(Кингисепп)
Slantsy
(Сланцы)
Luga
(Луга)
Vasalemma
Kohila
Rapla
Haapsalu
Märjamaa
Paide
Rõdala
Türi
Halinga
Audru
Sindi
Pärnu
Viljandi
Elva
Põlva
Räpina
Otepää
Rõuena
Pskov
Soltsy
(Сольцы)

# http://biit.cs.ut.ee

**Bioinformatics, Algorithmics and Data Mining Group**

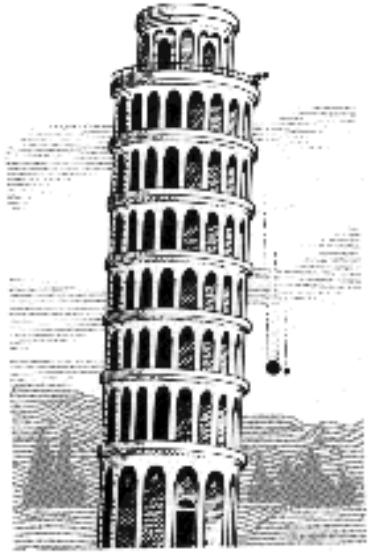
# http://stacc.ee

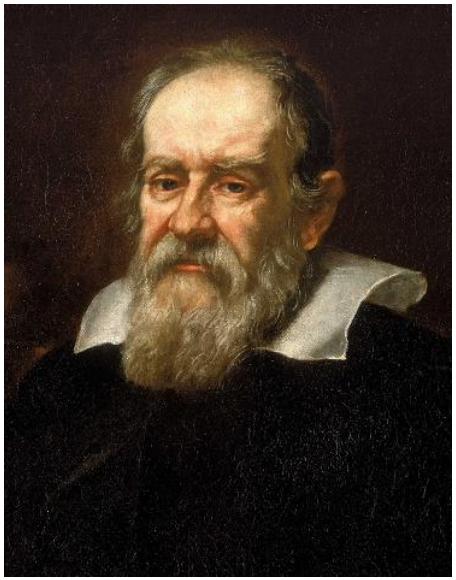
# Наука



# Наука

Наблюдение → Анализ → Обобщение

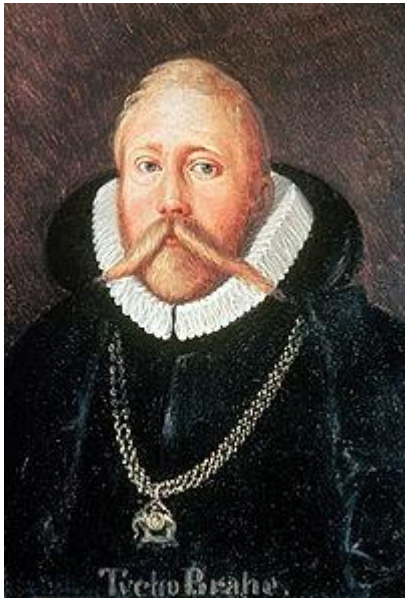


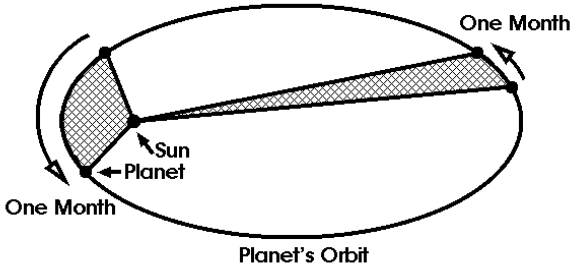


$$s = \frac{a}{2}t^2$$

# Наука

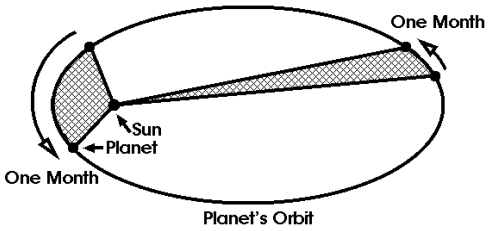
Наблюдение → Анализ → Обобщение

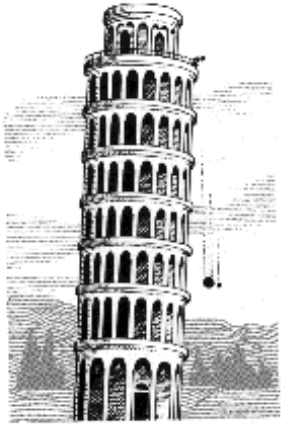


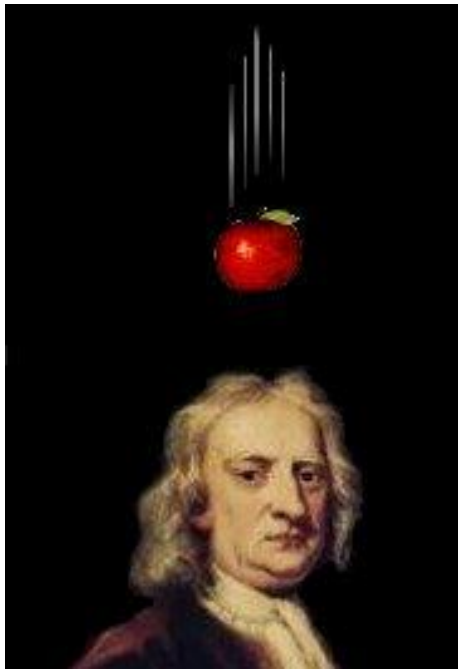


# Наука

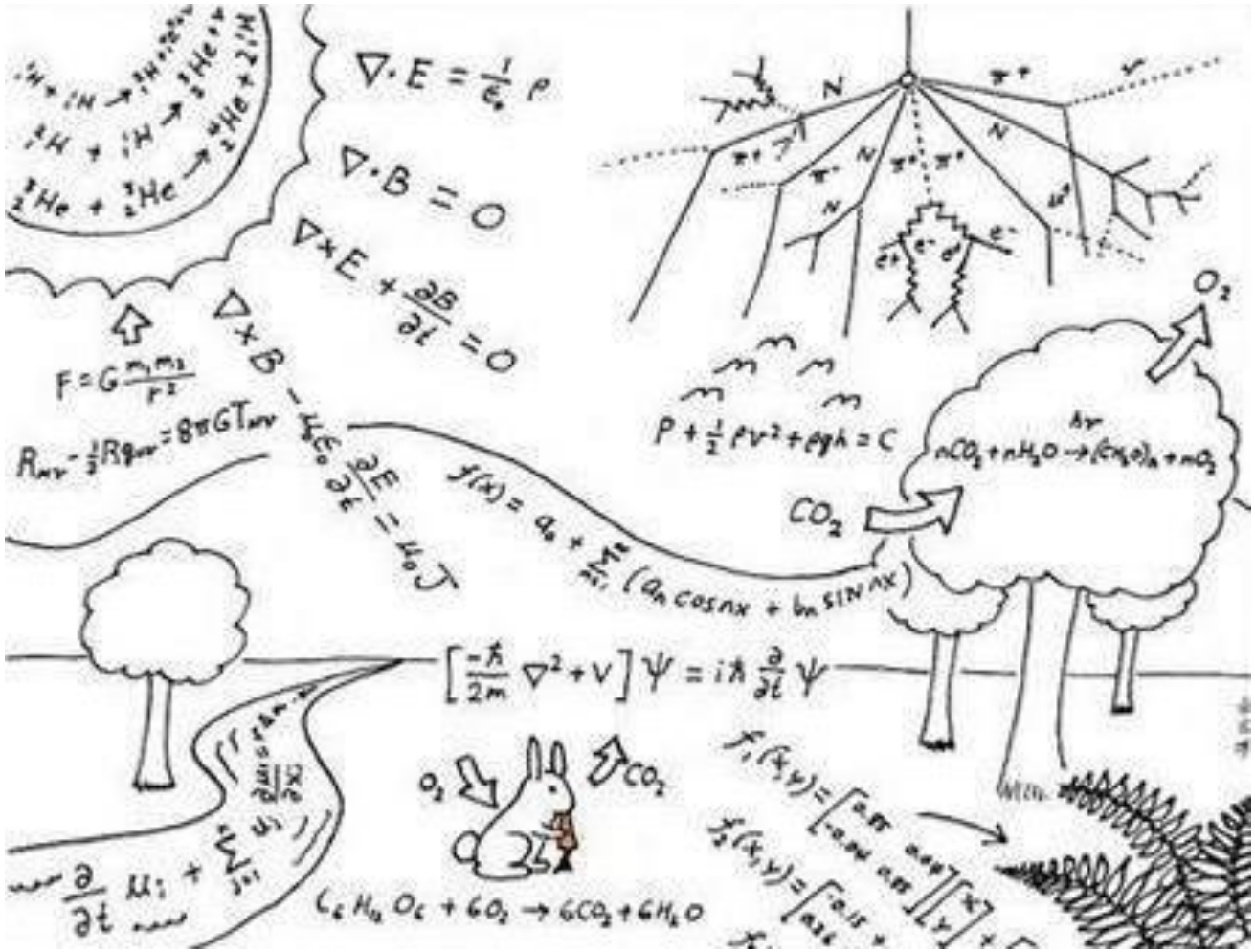
Наблюдение → Анализ → Обобщение

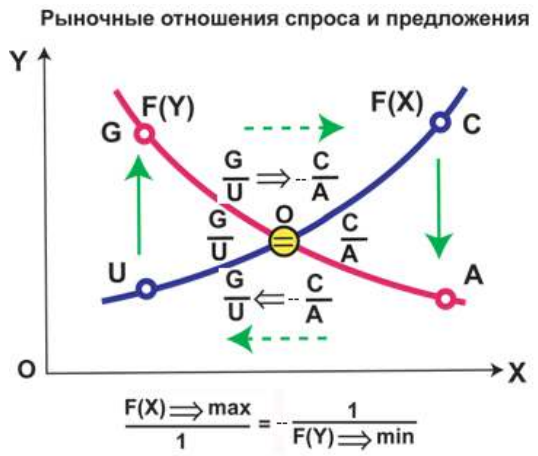


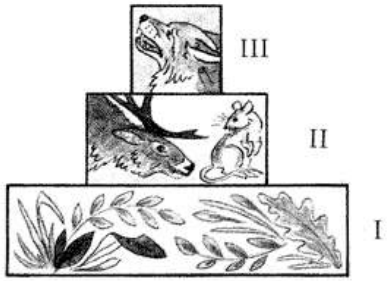



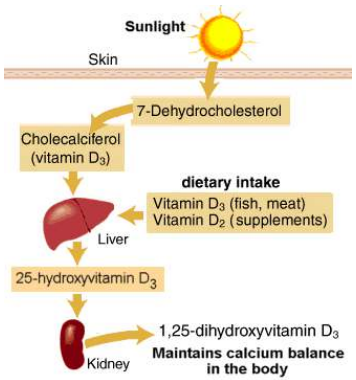


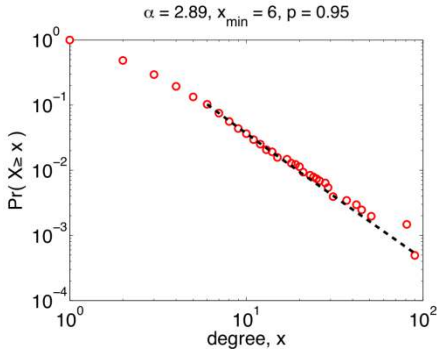
$$\vec{F} = m\vec{g}$$

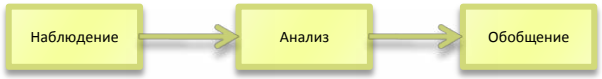

$\nabla \cdot E = \frac{1}{\epsilon_0}\rho$
$\nabla \cdot B = 0$
$\nabla \times E + \frac{\partial B}{\partial t} = 0$
$\nabla \times B - \mu_0\epsilon_0\frac{\partial E}{\partial t} = \mu_0 J$
$F = G\frac{m_1 m_2}{r^2}$
$R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu} = 8\pi G T_{\mu\nu}$
$P + \frac{1}{2}\rho v^2 + \rho g h = C$
$f(x) = a_0 + \sum_{n=1}^{\infty}(a_n \cos nx + b_n \sin nx)$
$\left[\frac{-\hbar}{2m}\nabla^2 + V\right]\Psi = i\hbar\frac{\partial}{\partial t}\Psi$
$C_6H_{12}O_6 + 6O_2 \rightarrow 6CO_2 + 6H_2O$
$CO_2$
$O_2$


Рыночные отношения спроса и предложения
Y
F(Y)
F(X)
G
C
U
A
O
X
G/U ⇒ -- C/A
G/U
O
C/A
G/U ⇐ -- C/A
F(X) ⇒ max / 1 = -- 1 / F(Y) ⇒ min

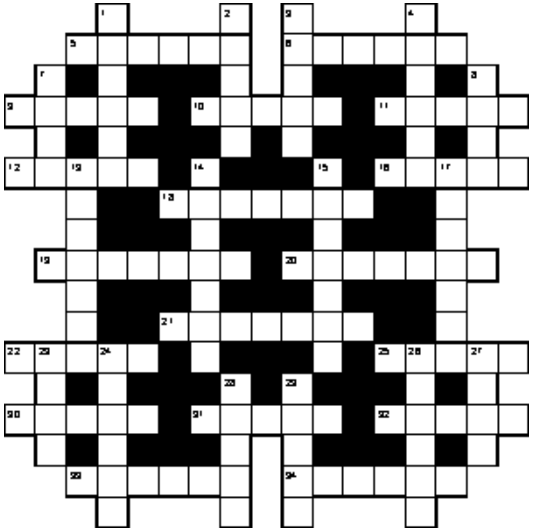

III
II
I


Sunlight
Skin
7-Dehydrocholesterol
Cholecalciferol (vitamin D3)
dietary intake
Vitamin D3 (fish, meat)
Vitamin D2 (supplements)
Liver
25-hydroxyvitamin D3
1,25-dihydroxyvitamin D3
Kidney
Maintains calcium balance in the body


α = 2.89, x_min = 6, p = 0.95
Pr( X≥ x )
degree, x

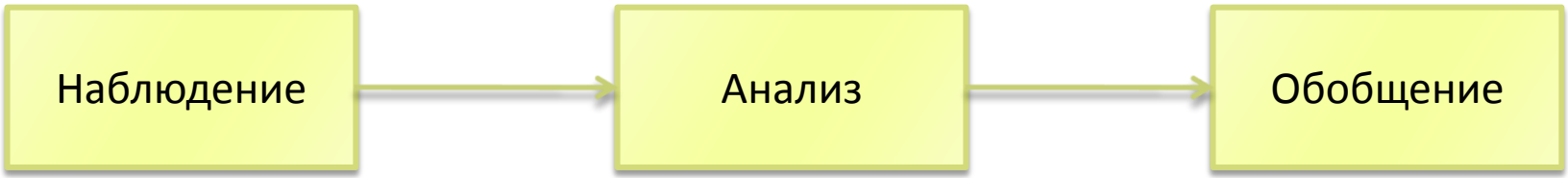

Наблюдение
Анализ
Обобщение

# Зачем?



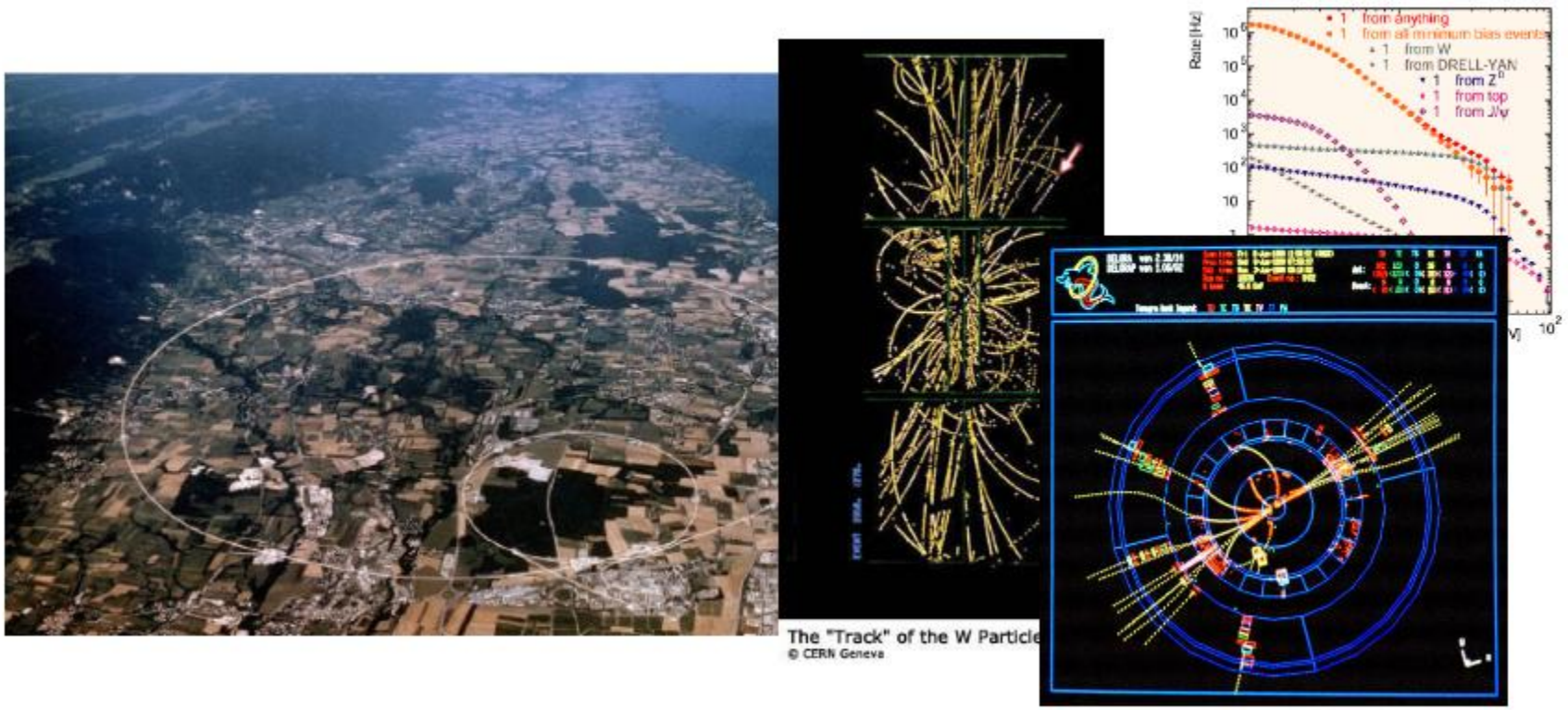
# Зачем?

# Современная наука

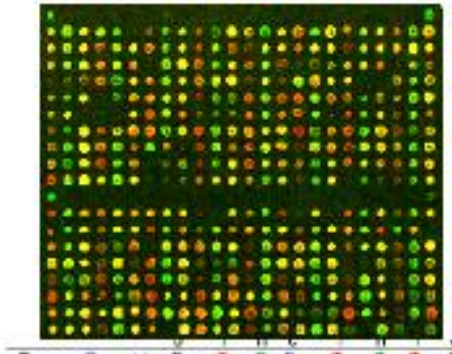


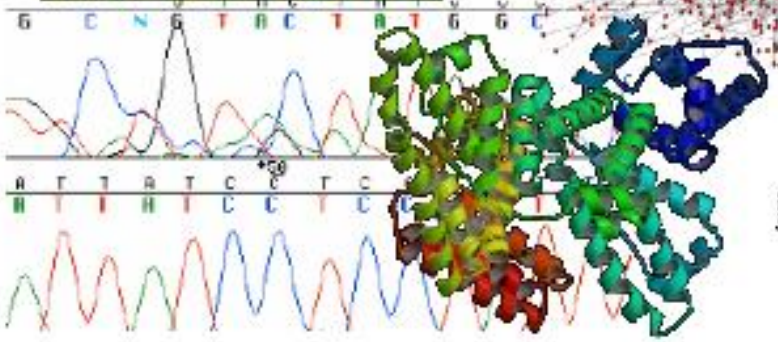
# Современная наука

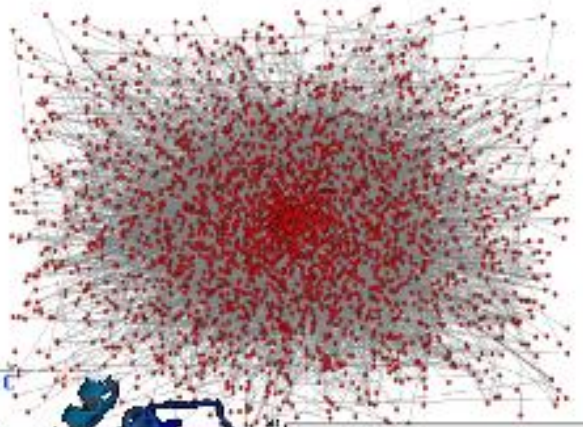
# Современная наука

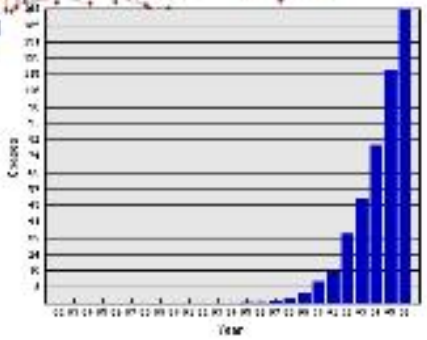


# Современная наука



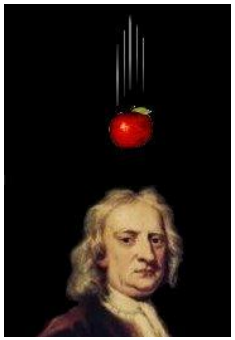




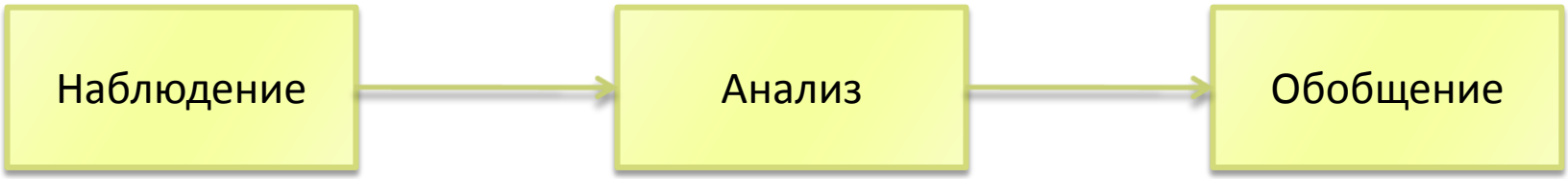


# Современная наука

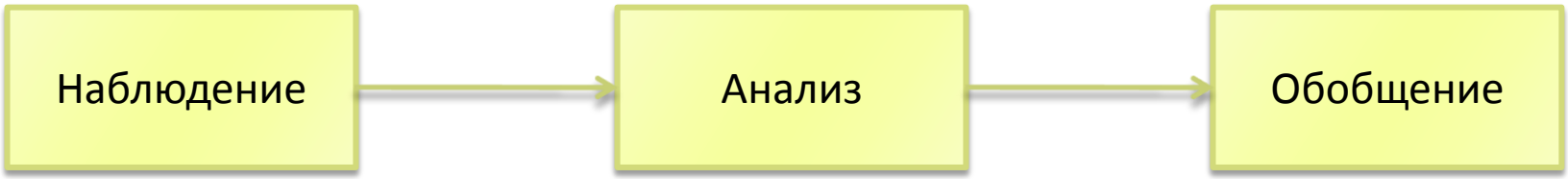
Наблюдение → Анализ → Обобщение



# Современная наука



# Современная наука



**Data mining**,

Data analysis, Statistical analysis,

Pattern discovery, Statistical learning,

**Machine learning**, Predictive analytics,

Business intelligence, Data-driven statistics

Inductive reasoning, **Pattern analysis**,

Knowledge discovery from databases,

Analytical processing,

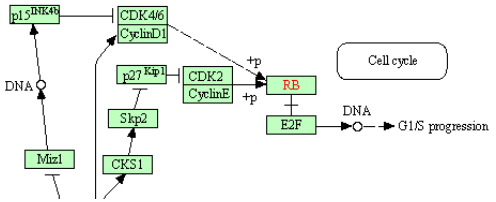
…

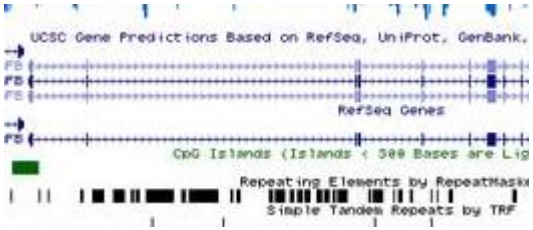
# Современная наука

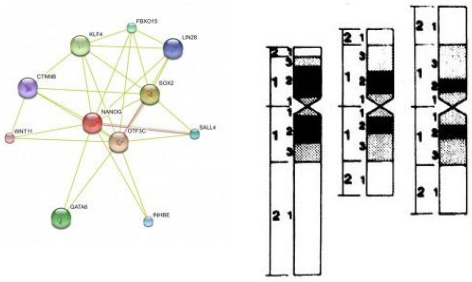
Наблюдение → Анализ → Обобщение

- **Нуклеотидные последовательности**
- **Молекулярные взаимодействия**
- **Экспрессия генов**
- **Наследственность**
- **Эффекты медикаментов**
- **…**

**Биоинформатика**



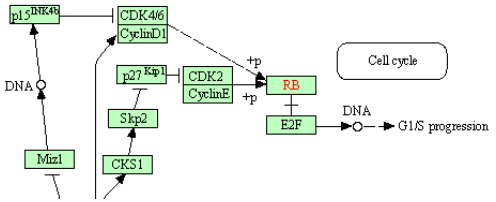



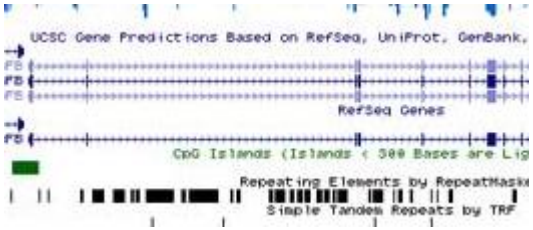
# Современная наука

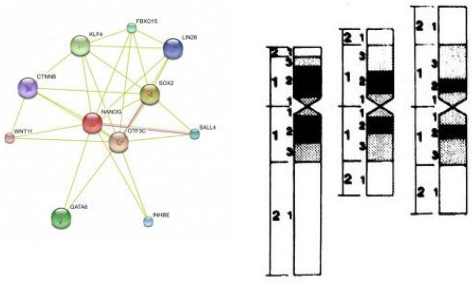
Наблюдение → Анализ → Обобщение

- **Нуклеотидные последовательности**
- **Молекулярные взаимодействия**
- **Экспрессия генов**
- **Наследственность**
- **Эффекты медикаментов**
- **…**

**Биоинформатика**



о-в. Приветствия
море Демагогии
м.Технологий
респ. Нуклеотидов
земля Смысла
б-о Проблем
а.о. Многомерного анализа
страна Экспрессии

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

# Bioconductor

**Bioconductor version 2.7 (Release)**

- AnnotationData (510)
- ExperimentData (68)
- Software (408)
  - Annotation (55)
    - GO (11)
    - Pathways (17)
    - ProprietaryPlatforms (1)
    - ReportWriting (15)
  - AssayDomains (161)
  - AssayTechnologies (254)
  - Bioinformatics (240)
    - Classification (27)
    - Clustering (32)
    - MultipleComparisons (29)
    - Preprocessing (78)
    - QualityControl (36)
    - SequenceMatching (6)
    - TimeCourse (10)
  - BiologicalDomains (39)
  - Infrastructure (176)

**Packages**

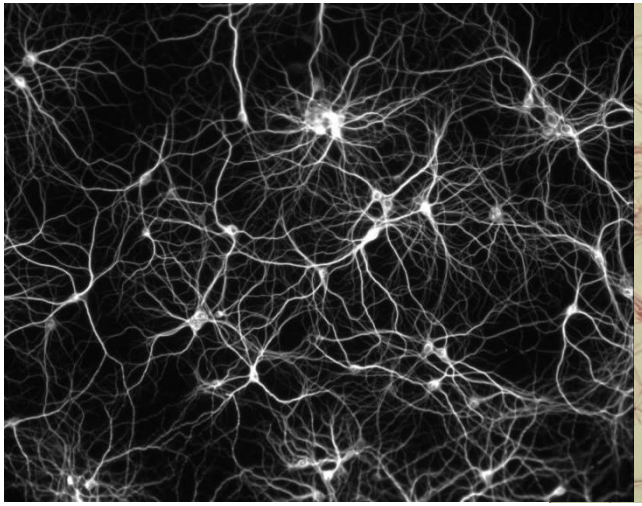
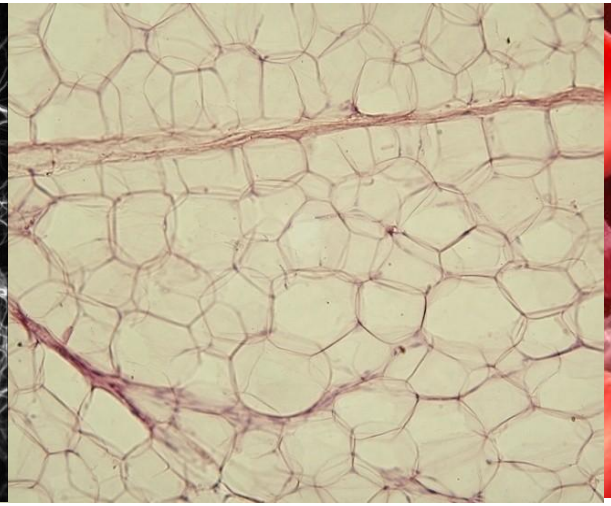
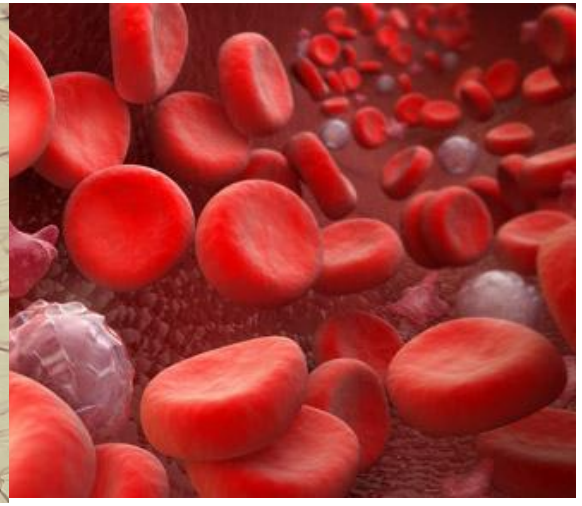
Software

- ABarray • aCGH • ACME • ADaCGH2 • adSplit • affxparser • affy • affycomp • Affy
- affyContam • affycoretools • AffyExpress • affyILM • affyio • affylmGUI • affyPara •
- affyPLM • affyQCReport • AffyTiling • Agi4x44PreProcess • AgiMicroRna • altcdfenvs •
- annotate • AnnotationDbi • annotationTools • apComplex • aroma.light • ArrayExpres
- arrayQuality • arrayQualityMetrics • ArrayTools • attract • BAC • BayesPeak • baySe
- beadarray • beadarraySNP • BeadDataPackR • betr • bgafun • BGmix • bgx • BHC •
- BiocCaseStudies • biocDatasets • biocGraph • biocViews • bioDist • biomaRt • BioMV
- BioSeqClass • Biostrings • bridge • BSgenome • BufferedMatrix • BufferedMatrixMeth
- CALIB • CAMERA • Category • cellHTS • cellHTS2 • CGEN • CGHbase • CGHcall • cg
- CGHnormaliter • CGHregions • charm • ChemmineR • ChIPpeakAnno • chipseq • Ch
- ChromHeatMap • clippda • clusterStab • CMA • CNTools • CNVtools • CoCiteStats •
- CoGAPS • ConsensusClusterPlus • convert • copa • coRNAi • CORREP • cosmo • cos
- CRImage • crlmm • CSAR • ctc • cycle • daMA • ddCt • DEDS • DEGraph • DEGsec
- diffGeneAnalysis • DNAcopy • domainsignatures • dualKS • dyebias • DynDoc • EBIr
- edd • edgeR • eisa • exonmap • explorase • ExpressionView • externalVector • fabi
- farms • fdrame • flagme • flowClust • flowCore • flowFlowJo • flowFP • flowMeans •
- flowStats • flowTrans • flowUtils • flowViz • frma • frmaTools • gaga • gage • gaggle
- genArise • gene2pathway • GeneAnswers • genefilter • GeneGA • GeneMeta • genep
- geneRecommender • GeneRegionScan • GeneRfold • GeneSelectMMD • GeneSelecto
- GeneticsDesign • GeneticsPed • GeneTraffic • genoCN • GenomeGraphs • genomeIn
- GenomicFeatures • GenomicRanges • GEOmetadb • GEOquery • GEOsubmission • G
- girafe • GLAD • GlobalAncova • globaltest • goProfiles • GOSemSim • goseq • GOst
- gpls • graph • GraphAlignment • GraphAT • GSEABase • GSEAlm • GSRI • Harshligh
- HELP • HEM • HilbertVis • HilbertVisGUI • hopach • HTqPCR • HTSanalyzeR • hyperd
- Icens • iChip • idiogram • imageHTS • impute • IRanges • iSeq • IsoGeneGUI • ITA
- iterativeBMA • iterativeBMAsurv • KCsmart • KEGGgraph • keggorthology • KEGGSO
- les • limma • limmaGUI • LiquidAssociation • LMGene • logicFS • logitT • LPE • LPEa
- LVSmiRNA • maanova • macat • maCorrPlot • maDB • made4 • maigesPack • make
- makePlatformDesign • MANOR • MantelCorr • marray • maSigPro • MassArray • Mas

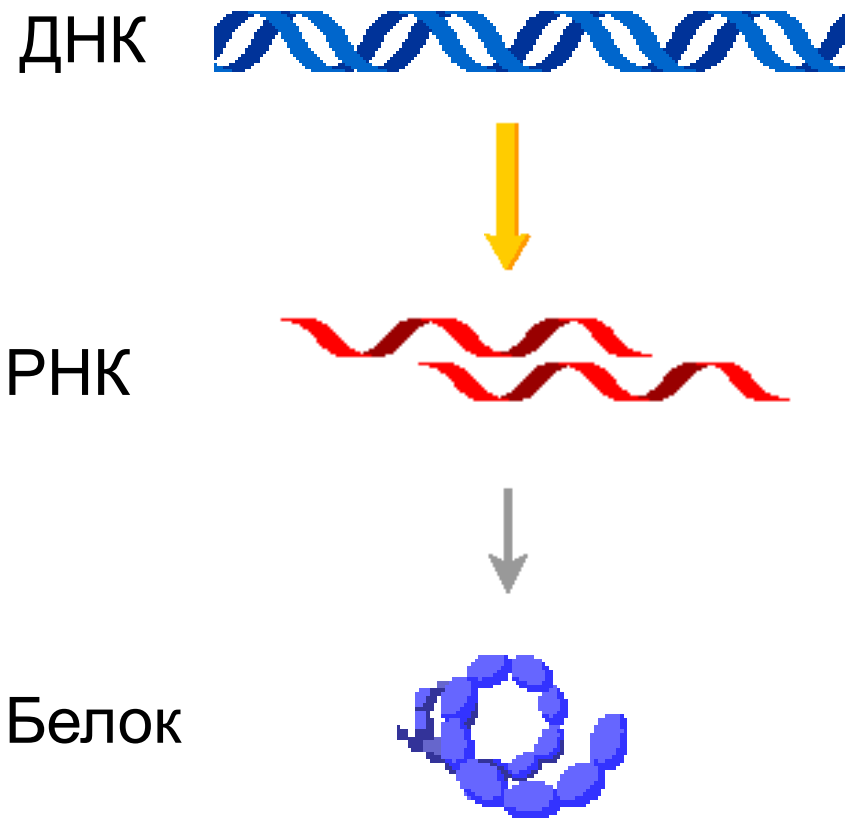
**http://biit.cs.ut.ee/~kt/spb**

о-в. Приветствия
море Демагогии
м.Технологий
респ. Нуклеотидов
земля Смысла
б-о Проблем
а.о. Многомерного анализа
страна Экспрессии

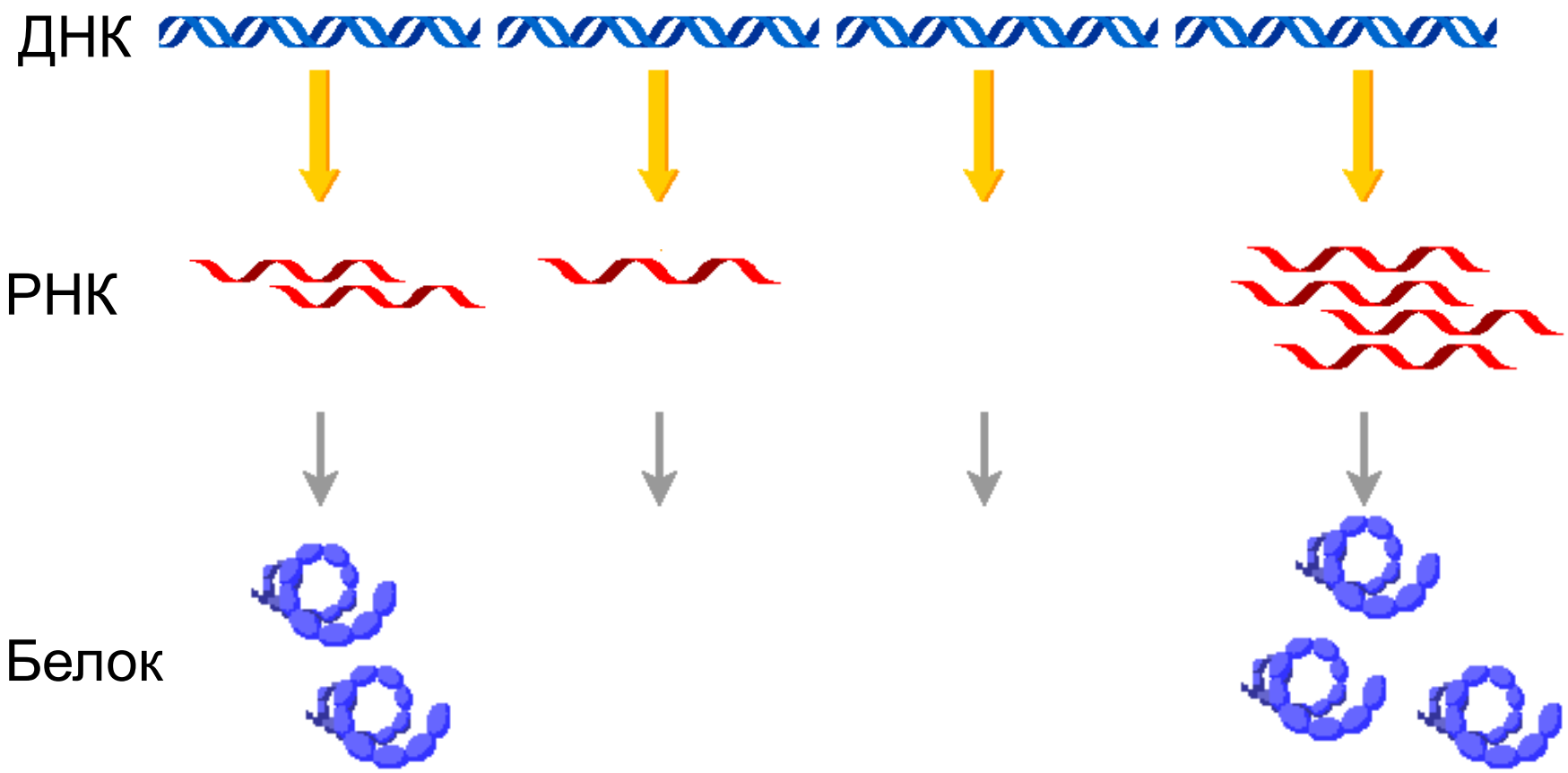
# Для чего измерять экспрессию?

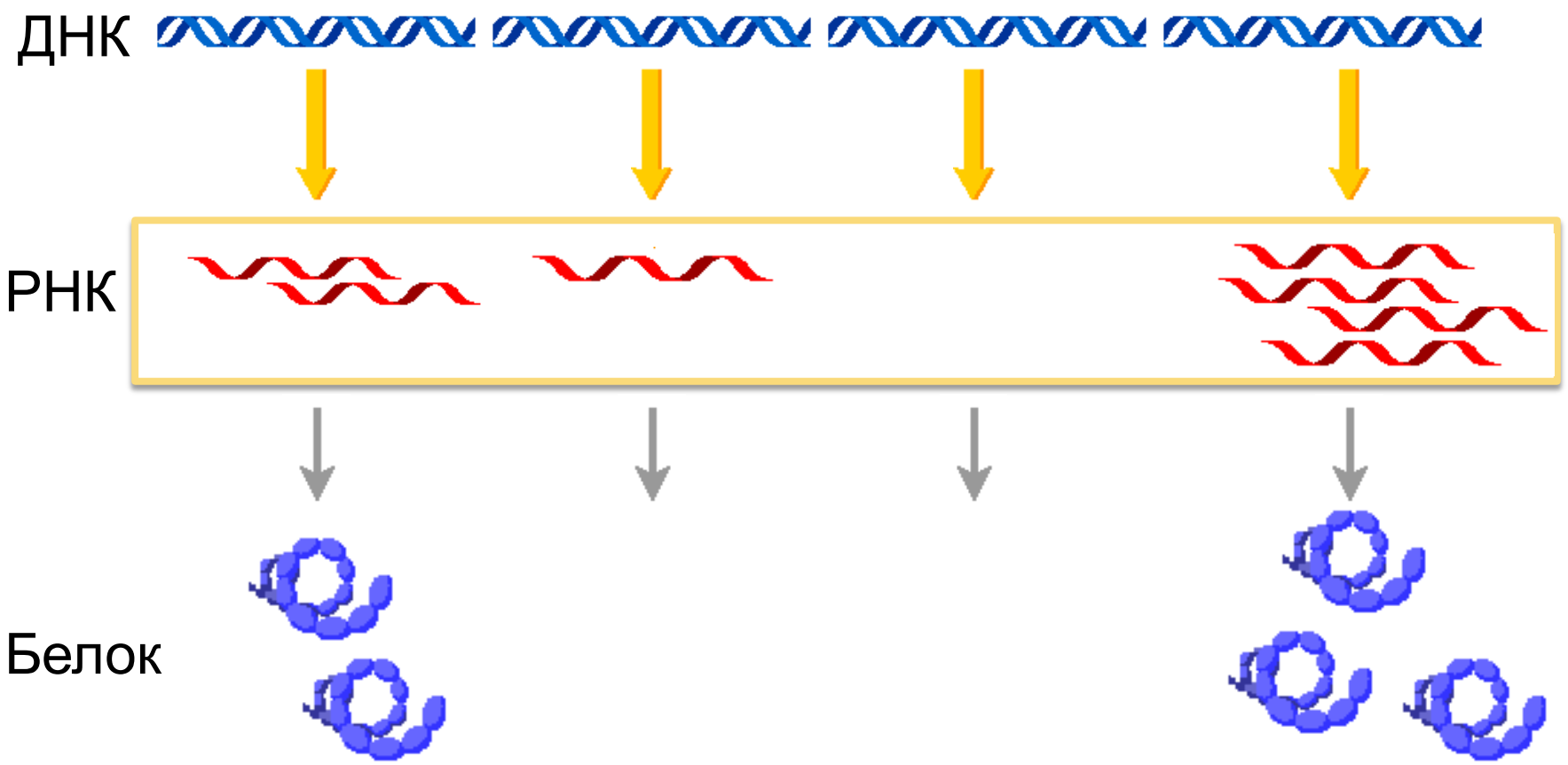
# Как измерять экспрессию?



# Как измерять экспрессию?



# Как измерять экспрессию?

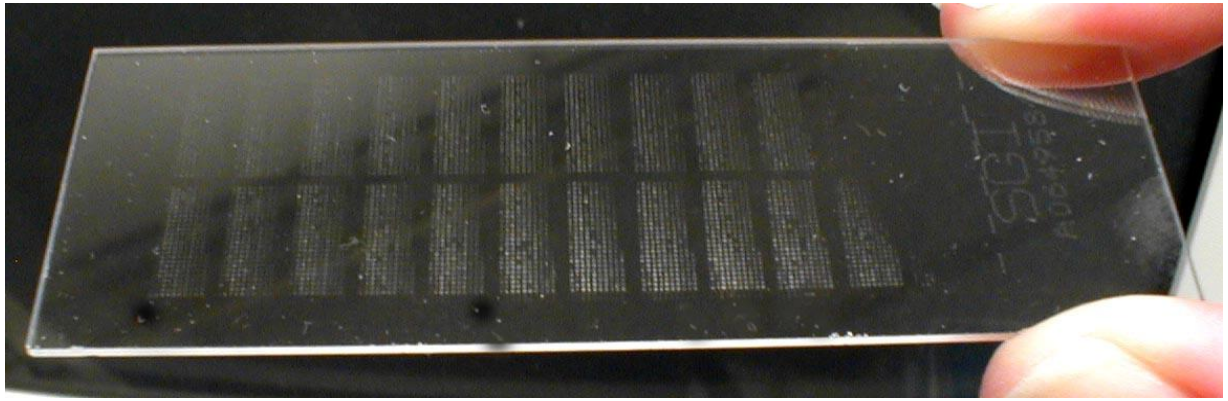


# Как измерять экспрессию?

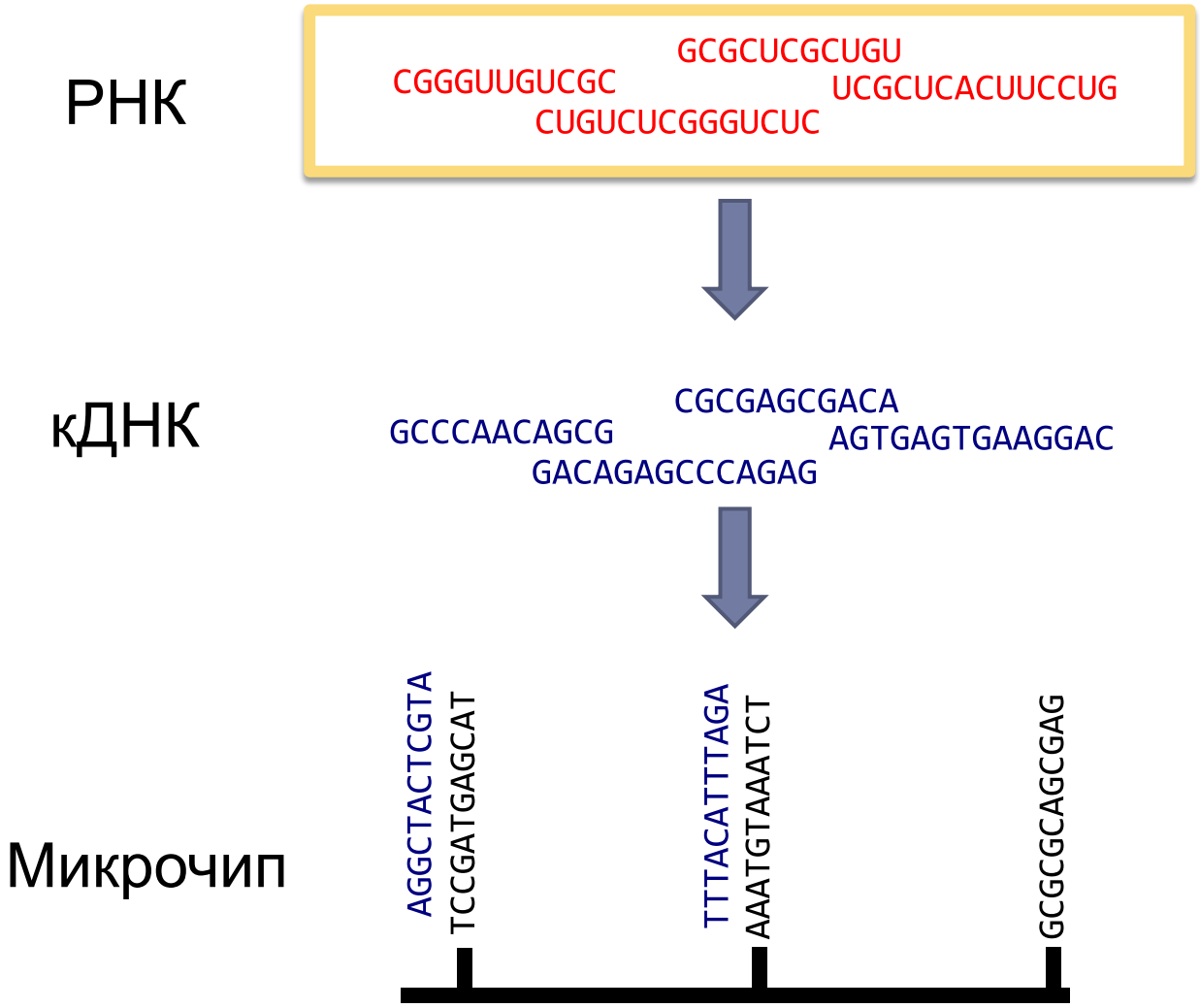
РНК

# Как измерять экспрессию?

РНК

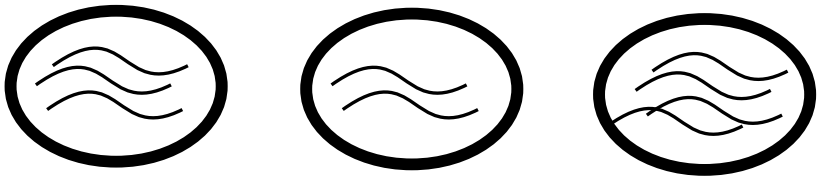


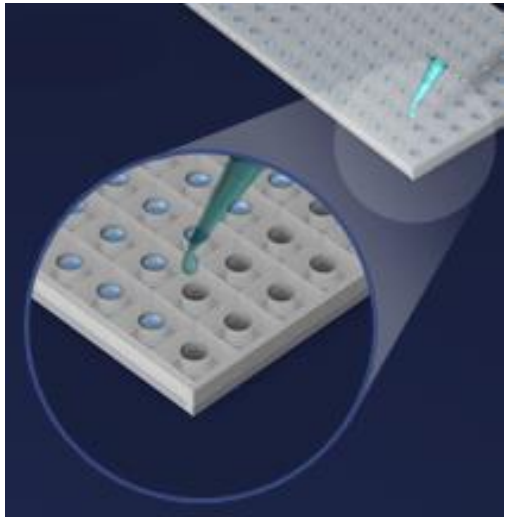
# Микрочип (Microarray)



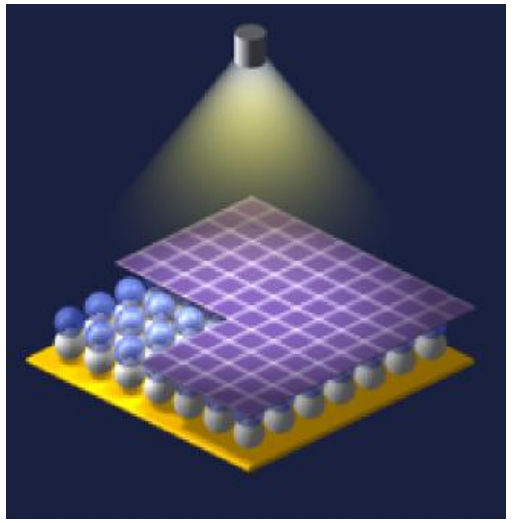
# Технологии микрочипов

- кДНК

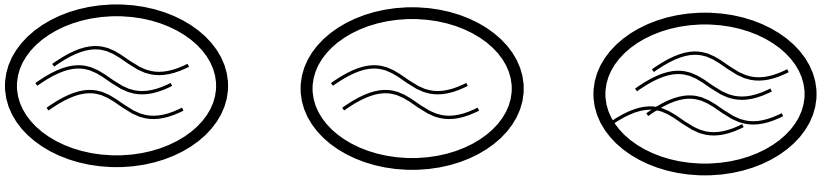


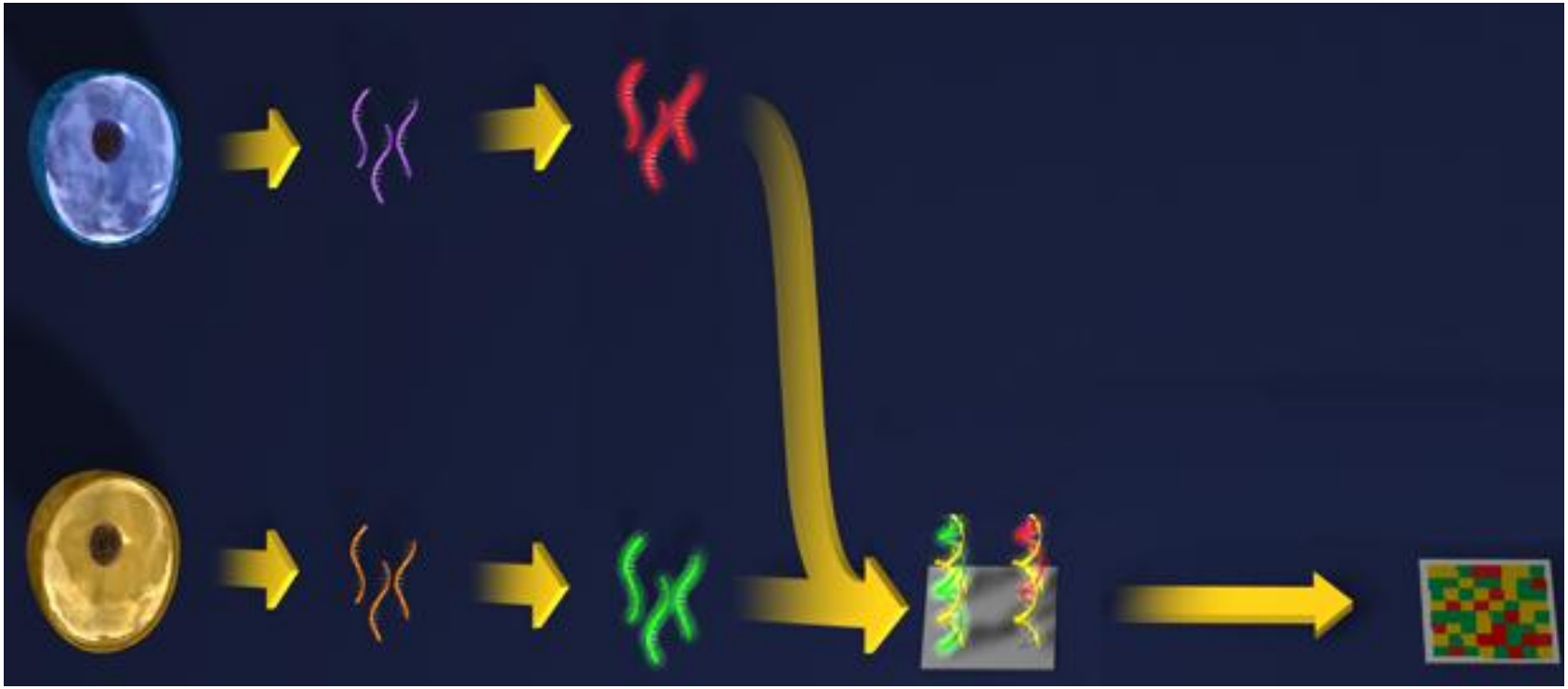


- Олигонуклеотидный

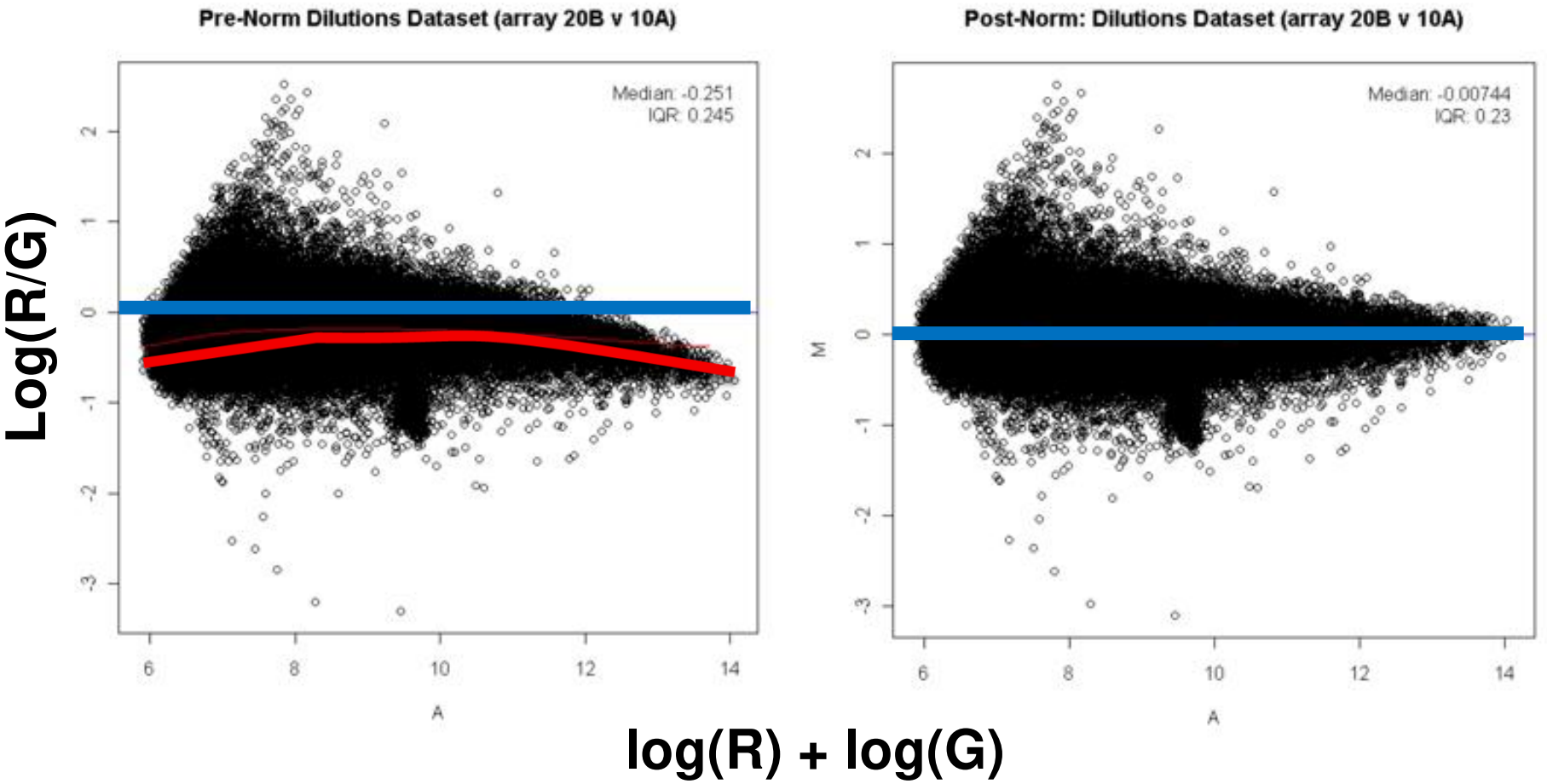


# кДНК: Двукрасочный чип (2-dye)



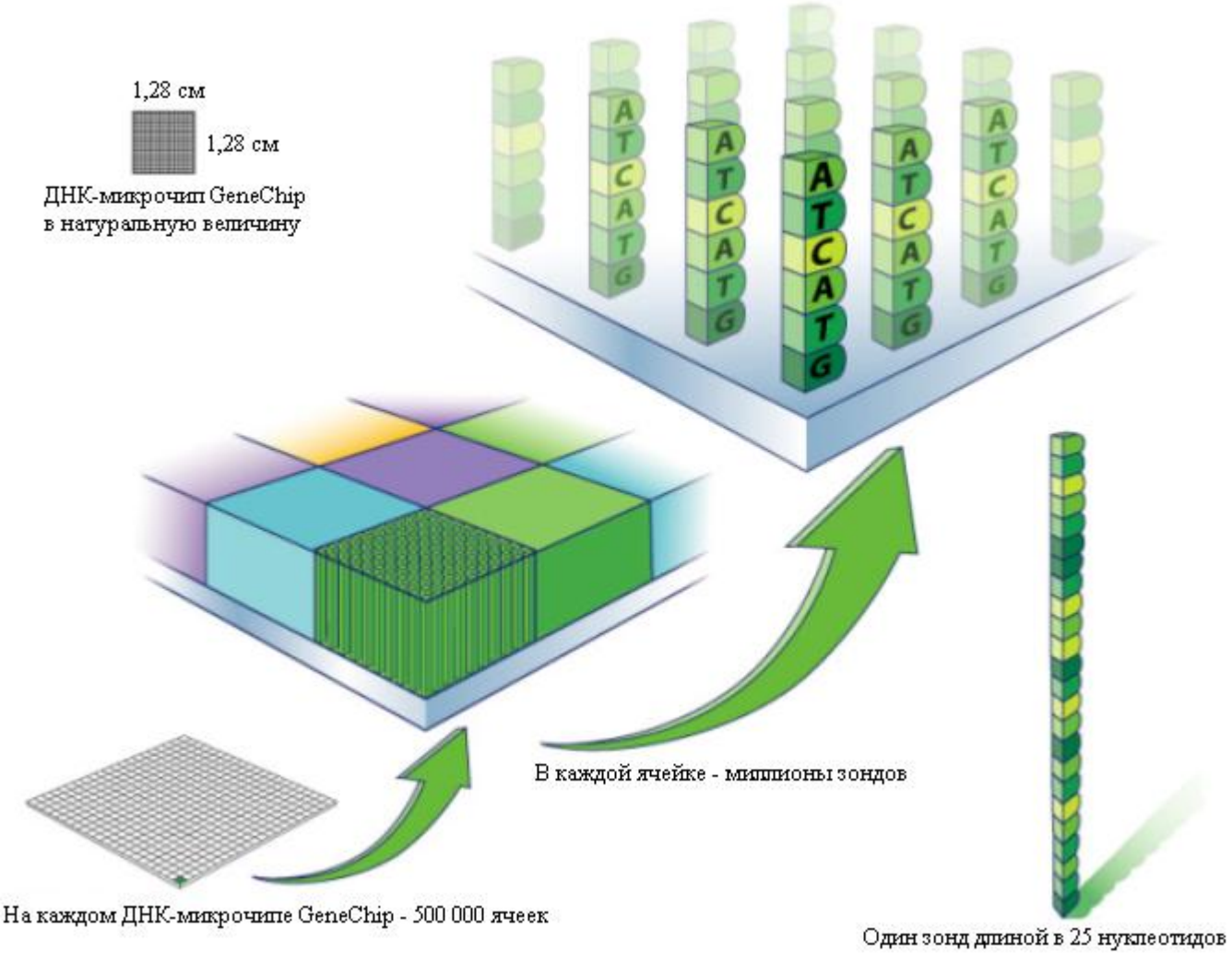


# Нормализация двукрасочных чипов



# Affymetrix GeneChip

**$100-850**



# Affymetrix GeneChip

**Ген**

**10-20 зондов, каждый длиной 25 нуклеотидов**

**+ столько же “mismatch” зондов**

# Affymetrix GeneChip

```
> source("http://www.bioconductor.org/biocLite.R")
> biocLite("ygs98probe")
> library(ygs98probe)
> as.data.frame(ygs98probe[1:10,1:4])
                      sequence  x   y Probe.Set.Name
1  CGATATGGAGGCAATGCTGGTAGAC 90 121        10000_at
2  GGAGGCAATGCTGGTAGACGAACTC 91 121        10000_at
3  AATGCTGGTAGACGAACTCGTATGT 92 121        10000_at
4  GGTAGACGAACTCGTATGTGACACG 93 121        10000_at
5  CGAACTCGTATGTGACACGAGGGAC 94 121        10000_at
6  CGTATGTGACACGAGGGACCTACTT 95 121        10000_at
7  GAGGGACCTACTTGATGTCGATGAA 96 121        10000_at
8  TGAGTCAGCACTTGAGGAAGAAACC 97 121        10000_at
9  TGAAGAACTAACTCTACTTACGGAA 98 121        10000_at
```

# Affymetrix GeneChip

```
> biocLite("ygs98.db")
> library(ygs98.db)

> ygs98ORF[["10000_at"]]
[1] "YLR331C"
> ygs98CHRLOC[["10000_at"]]
    12
-790669
> ygs98CHRLOCEND[["10000_at"]]
    12
-791046
```

# Affymetrix GeneChip

```
> biocLite("BSgenome.Scerevisiae.UCSC.sacCer1")
> library(BSgenome.Scerevisiae.UCSC.sacCer1)

> Scerevisiae$chr12[790669:791046]
  378-letter "DNAString" instance
seq: TCATAGTATGTTGTCTTTCACAACCAAGAATAGTT...CTTCAGATTCTTCA

> seq = Scerevisiae$chr12[790669:791046]
> probe = ygs98probe[1,1]

> library(Biostrings)
> matchPattern(probe, seq)
  Views on a 378-letter DNAString subject
subject: TCATAGTATGTTGTCTTTCACAACCAAGAATAG...T
views: NONE
```

# Affymetrix GeneChip

```
> matchPattern(probe, reverseComplement(seq))
  Views on a 378-letter DNAString subject
subject: ATGGAAAAGGATGAGGAGGATGAAGAATCTGAA...TATTCTTGGTTG
views:
    start end width
[1]    90 114    25 [CGATATGGAGGCAATGCTGGTAGAC]
```

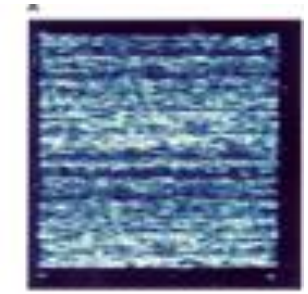
```
> as.data.frame(ygs98probe[1,4:5])
  Probe.Set.Name Probe.Interrogation.Position
1       10000_at                          102
```

# Affymetrix GeneChip

**Ген**

**10-20 зондов, каждый длиной 25 нуклеотидов**

**+ столько же “mismatch” зондов**



**.DAT**

**Зонд1 = 2.0**
**Зонд2 = 1.0**
**Зонд3 = 1.5**
**…**

**.CEL**

**Ген1 = -1**
**Ген2 = 0.2**
**Ген3 = 1.2**
**…**

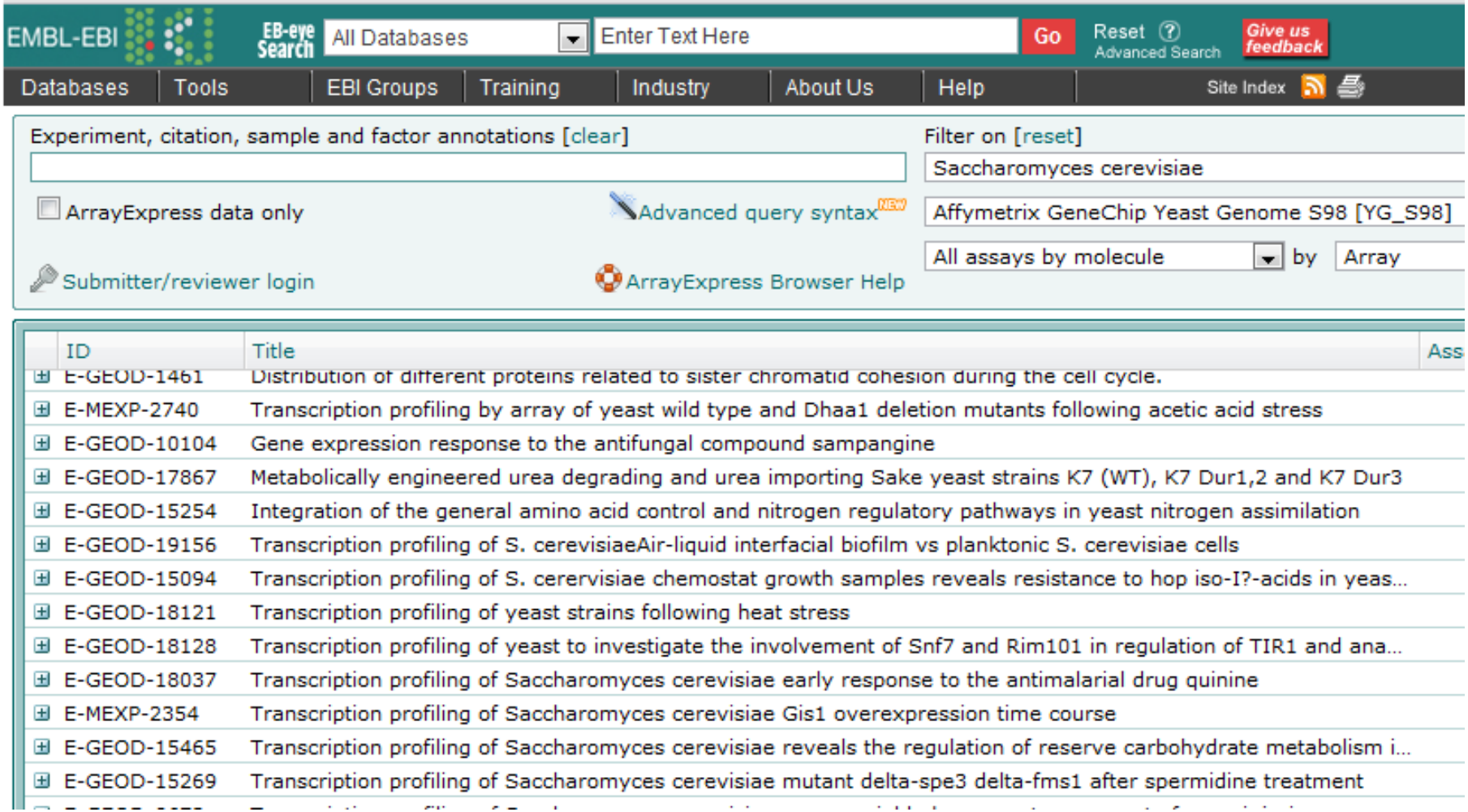
**.CSV**

# Базы данных экспрессии

**Gene Expression Omnibus**: http://www.ncbi.nlm.nih.gov/geo/

**ArrayExpress**: http://www.ebi.ac.uk/arrayexpress/

# ArrayExpress



# ArrayExpress

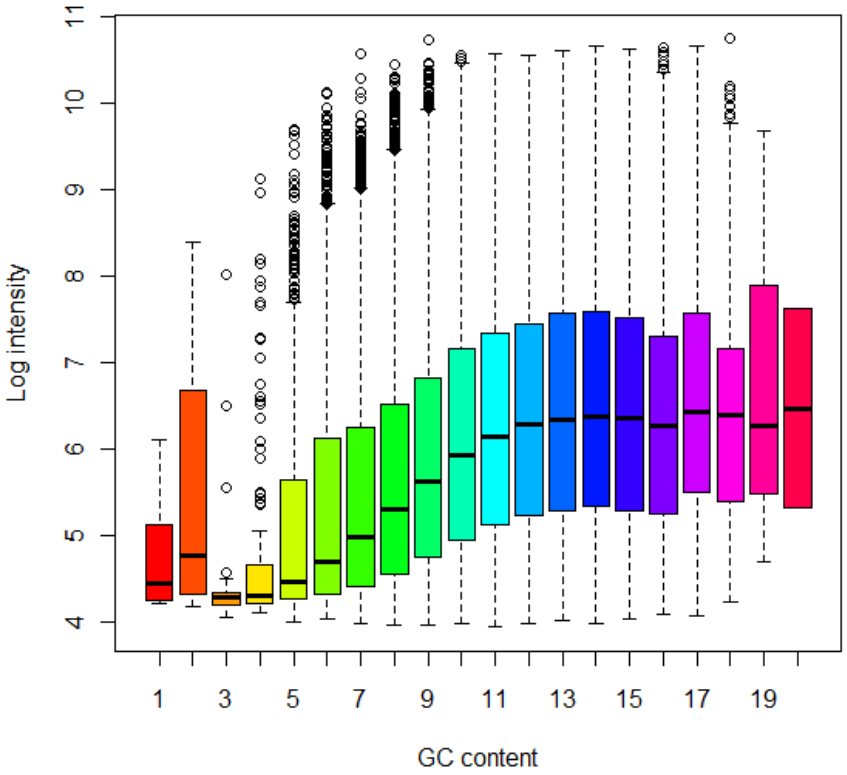
```
> library(ArrayExpress)
> library(affy)
> affydata = ArrayExpress("E-GEOD-18037")

> affydata
AffyBatch object
size of arrays=534x534 features (18 kb)
cdf=YG_S98 (9335 affyids)
number of samples=4
number of genes=9335
annotation=ygs98
notes=E-GEOD-18037
        E-GEOD-18037
        NA
        c("unknown_experiment_design_type"
```

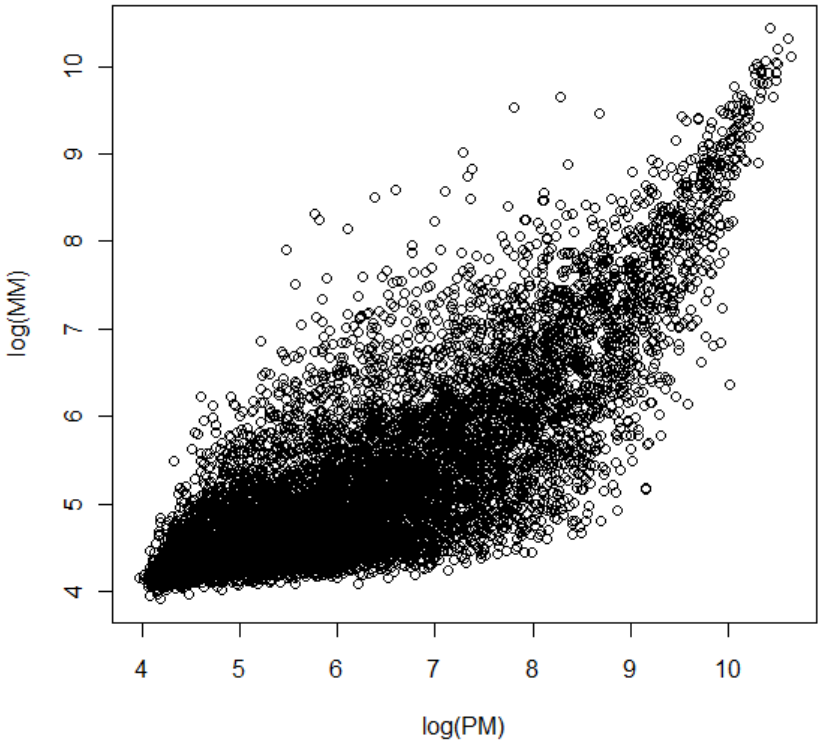
# AffyBatch

```
> pm(affydata)[1:5,]
      GSM451051.CEL GSM451049.CEL GSM451050.CEL GSM451048.CEL
64704       10066.3       10329.3        4719.0        6084.0
64705         112.3         105.8          81.3          96.0
64706         112.8         112.3         106.3         105.5
64707          85.8         112.5          70.0          90.5
64708          69.5          80.3          58.0          76.0
> mm(affydata)[1:5,]
      GSM451051.CEL GSM451049.CEL GSM451050.CEL GSM451048.CEL
65238        7597.0        8203.3        2949.0        4438.3
65239         113.3         144.5         100.0         131.0
65240          84.0          80.8          60.0          76.0
65241          73.3          91.5          55.0          80.5
65242          84.3          92.5          61.3          72.0

> dim(pm(affydata))
[1] 138412      4
> dim(ygs98probe)
[1] 138412      6
```

о-в. Приветствия
море Демагогии
м.Технологий
респ. Нуклеотидов
земля Смысла
б-о Проблем
а.о. Многомерного анализа
страна Экспрессии

# Нормализация

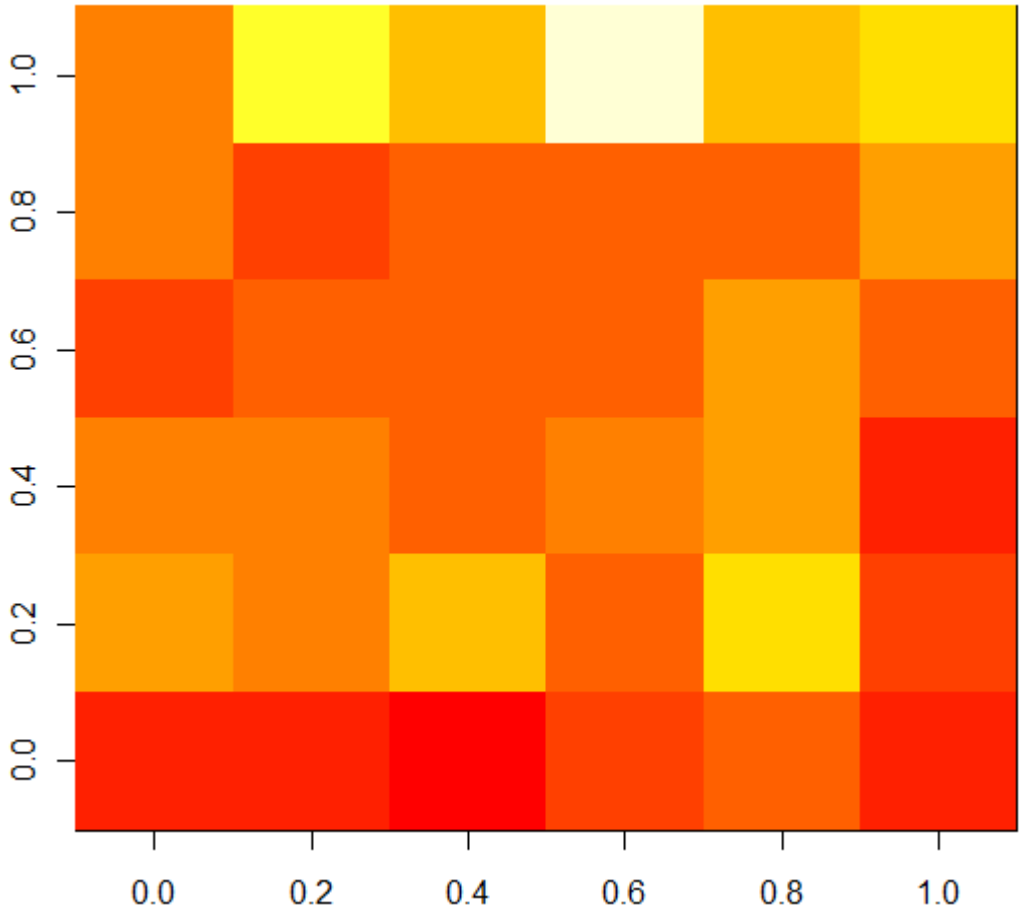



```
> probes = DNAStringSet(ygs98probe[,1])
> actg = alphabetFrequency(probes)
> x = actg[,"C"] + actg[,"G"]
> y = log(pm(affydata)[,1])
> boxplot(y ~ x)
```

```
> x = log(pm(affydata))
> y = log(mm(affydata))
> plot(x,y)
```

# Нормализация



# Нормализация

- Доверяем ли мы замеренным значениям?
- Что делать с PM/MM?
- Как сопоставить Probeset и Ген?
- Как нормализовать сразу несколько чипов?

# Нормализация

- Доверяем ли мы замеренным значениям?
- Что делать с PM/MM?
- Как сопоставить Probeset и Ген?
- Как нормализовать сразу несколько чипов?

```
> library(affy)
```

```
> expdata = mas5(affydata)
```

```
> expdata = rma(affydata)
```

```
> exptable = get("exprs", assayData(expdata))
```

# Gene expression data

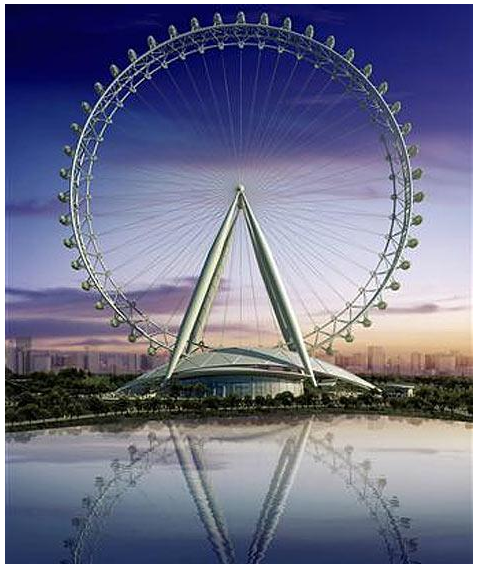
```
> expdata = read.table("http://genome-www.stanford.edu/ce

> expdata = expdata[,26:49]
> names(expdata)
 [1] "cdc15_10"  "cdc15_30"  "cdc15_50"  "cdc15_70"
 [5] "cdc15_80"  "cdc15_90"  "cdc15_100" "cdc15_110"
 [9] "cdc15_120" "cdc15_130" "cdc15_140" "cdc15_150"
[13] "cdc15_160" "cdc15_170" "cdc15_180" "cdc15_190"
[17] "cdc15_200" "cdc15_210" "cdc15_220" "cdc15_230"
[21] "cdc15_240" "cdc15_250" "cdc15_270" "cdc15_290"

> dim(expdata)
[1] 6178   24
```

```
> expdata[1:5,1:5]
        cdc15_10 cdc15_30 cdc15_50 cdc1
YAL001C    -0.16     0.09    -0.23
YAL002W     0.09       NA       NA    -
YAL003W    -0.37    -0.22    -0.16
```

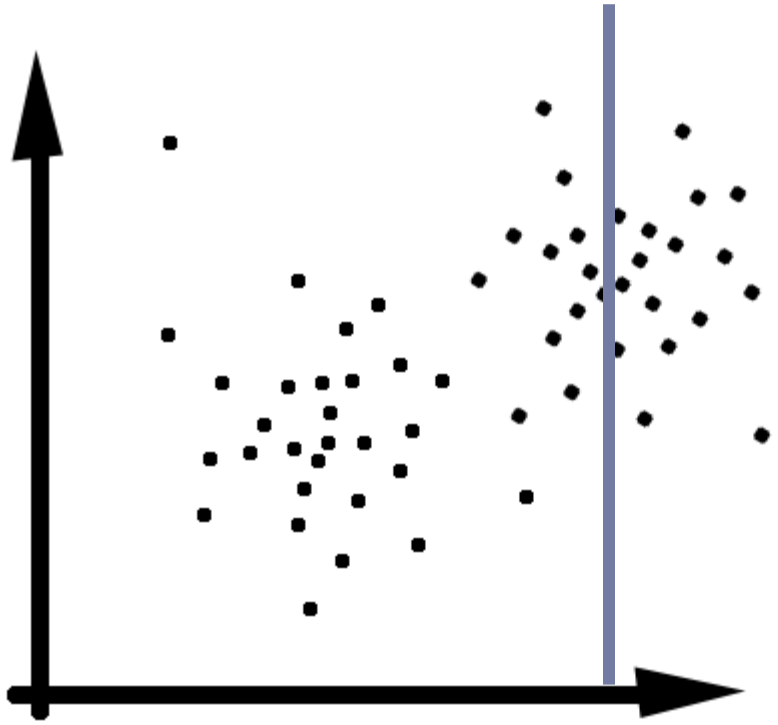
Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridization.
**Paul Spellman, et al. (1998)**

о-в. Приветствия
море Демагогии
м.Технологий
респ. Нуклеотидов
земля Смысла
б-о Проблем
а.о. Многомерного анализа
страна Экспрессии

# Welcome!

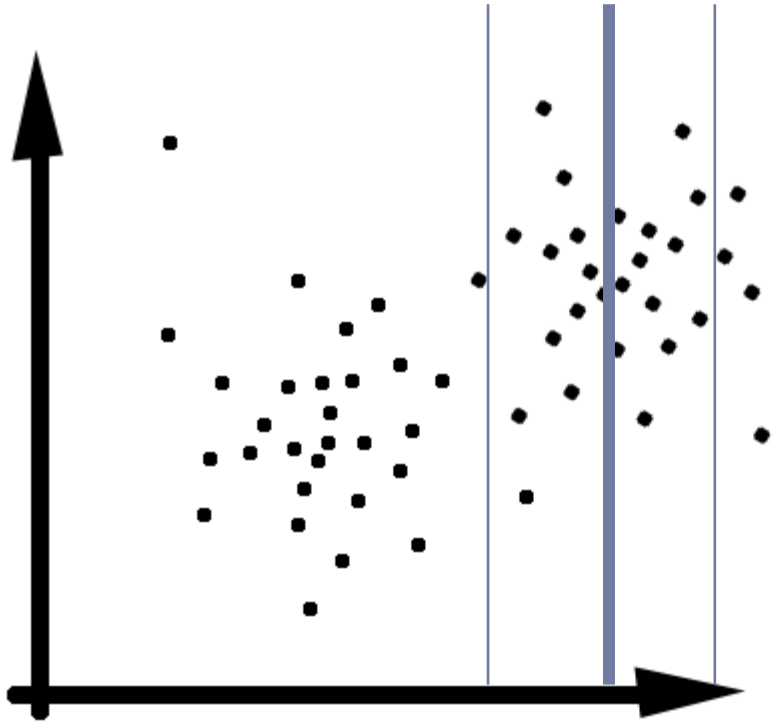


```
> expdata[1:5,1:5]
        cdc15_10 cdc15_30 cdc15_50 cdc15_70 cdc15_80
YAL001C    -0.16     0.09    -0.23     0.03    -0.04
YAL002W       NA       NA       NA    -0.58     0.23
```



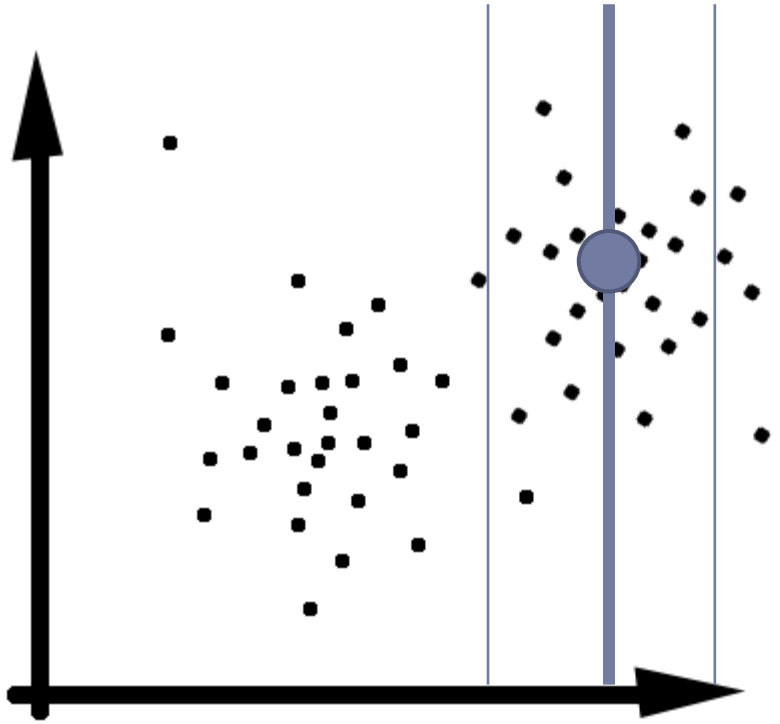
| x | y |
|---|---|
| 0.50 | 0.21 |
| 0.49 | 0.42 |
| 0.67 | 0.34 |
| 1.02 | 0.55 |
| 1.53 | 0.71 |
| 0.31 | 0.33 |
| 0.35 | 0.70 |
| 0.21 | 1.20 |
| ... | ... |
| **2.00** | **NA** |

# KNN-impute



| x | y |
|---|---|
| 0.50 | 0.21 |
| 0.49 | 0.42 |
| 0.67 | 0.34 |
| 1.02 | 0.55 |
| 1.53 | 0.71 |
| 0.31 | 0.33 |
| 0.35 | 0.70 |
| 0.21 | 1.20 |
| ... | ... |
| **2.00** | **NA** |

# KNN-impute



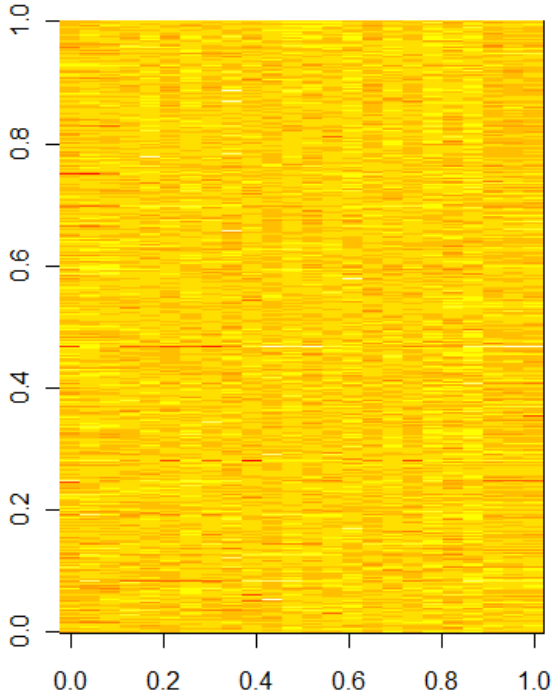
| x | y |
|---|---|
| 0.50 | 0.21 |
| 0.49 | 0.42 |
| 0.67 | 0.34 |
| 1.02 | 0.55 |
| 1.53 | 0.71 |
| 0.31 | 0.33 |
| 0.35 | 0.70 |
| 0.21 | 1.20 |
| ... | ... |
| **2.00** | **1.90** |

Missing value estimation methods for DNA microarrays
**Olga Troyanskaya, et al. (2000)**

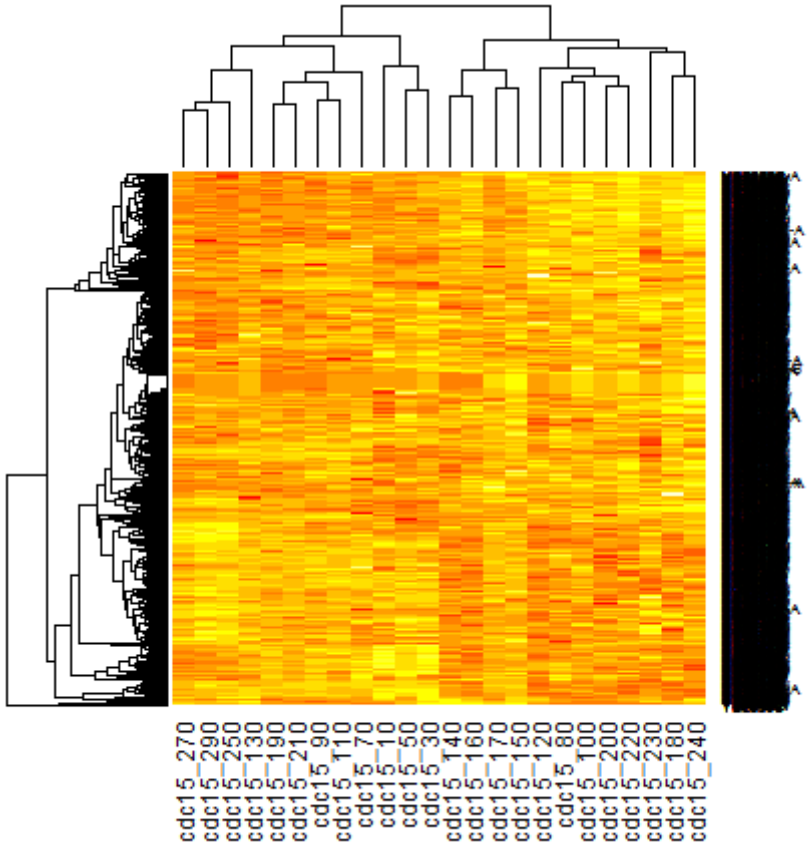
# KNN-impute

```
> library(impute)
> expdata = impute.knn(as.matrix(expdata))$data
```

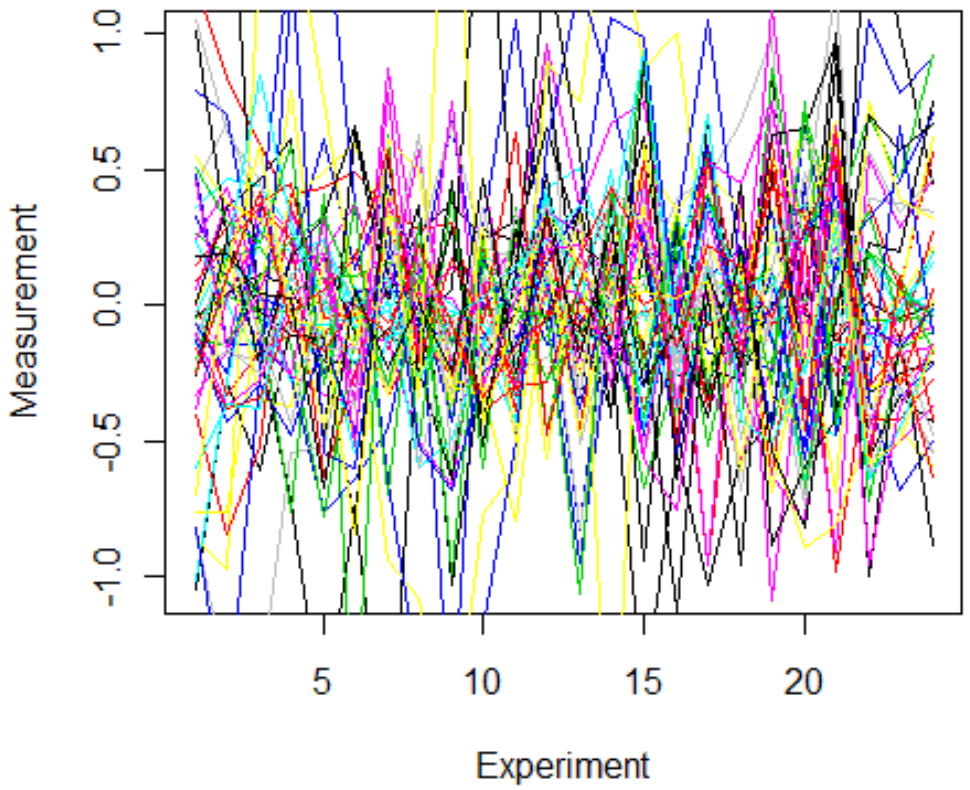
# Как увидеть 24 измерения?




```
> rsamp = sample(1:nrow(expdata), 1000)
> expdata.sample = expdata[rsamp,]

> image(t(expdata.sample))
```

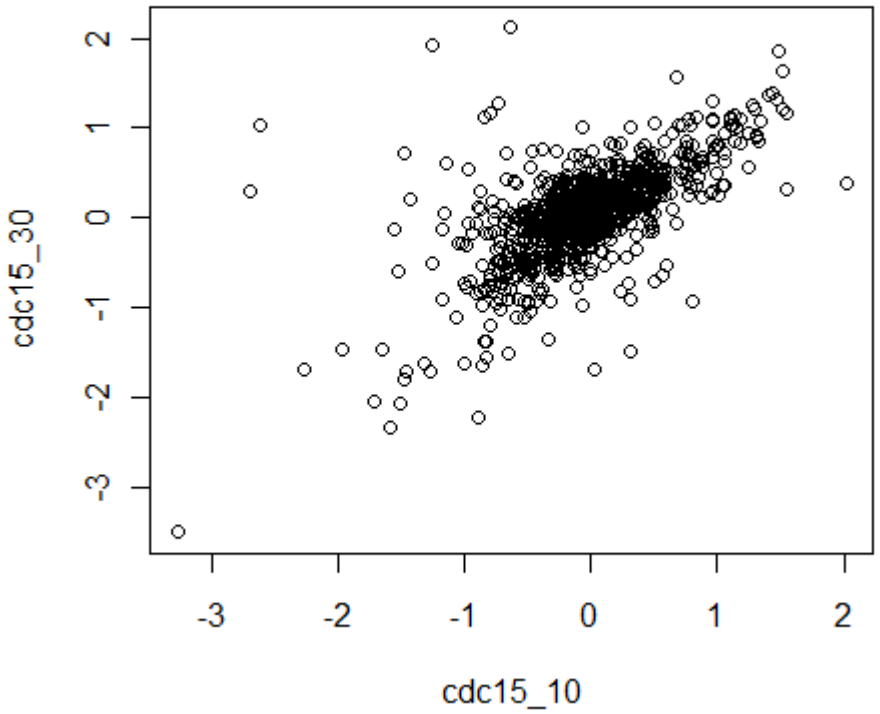



```
> heatmap(expdata.sample)
```

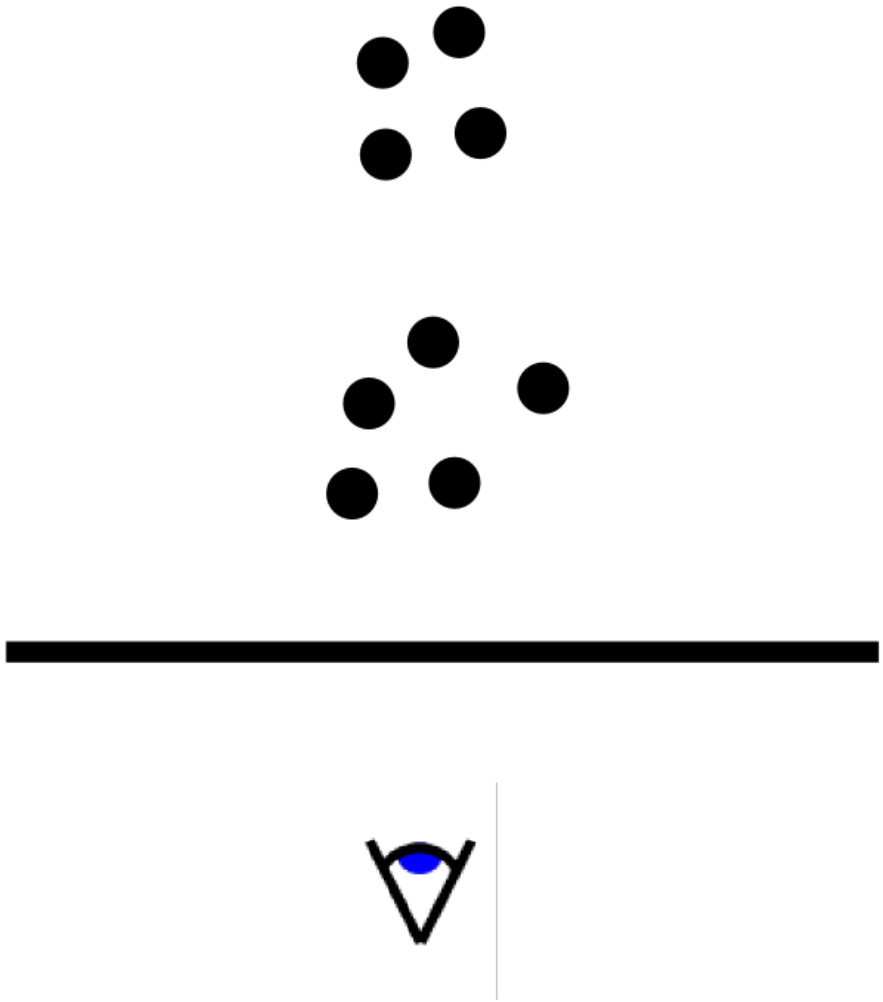


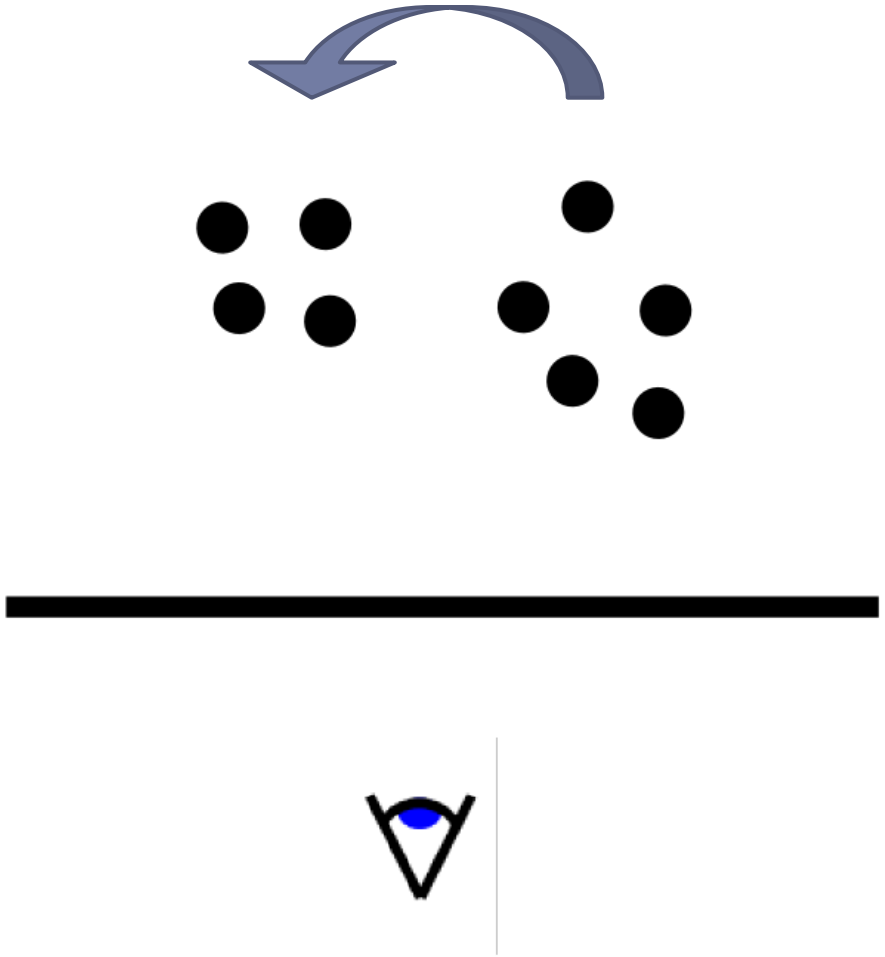

```
> plot(expdata.sample[1,],type='l')
> for (i in 2:50) lines(expdata.sample[i,],col=i)
```

# Спроецируем на 2 измерения

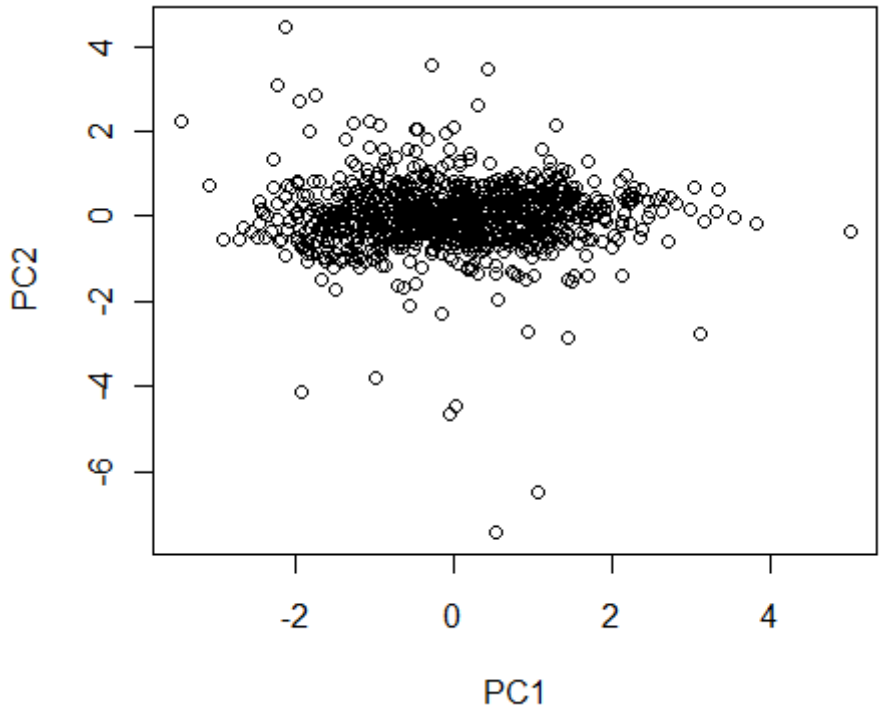



```
> plot(expdata.sample[,1], expdata.sample[,2])
```

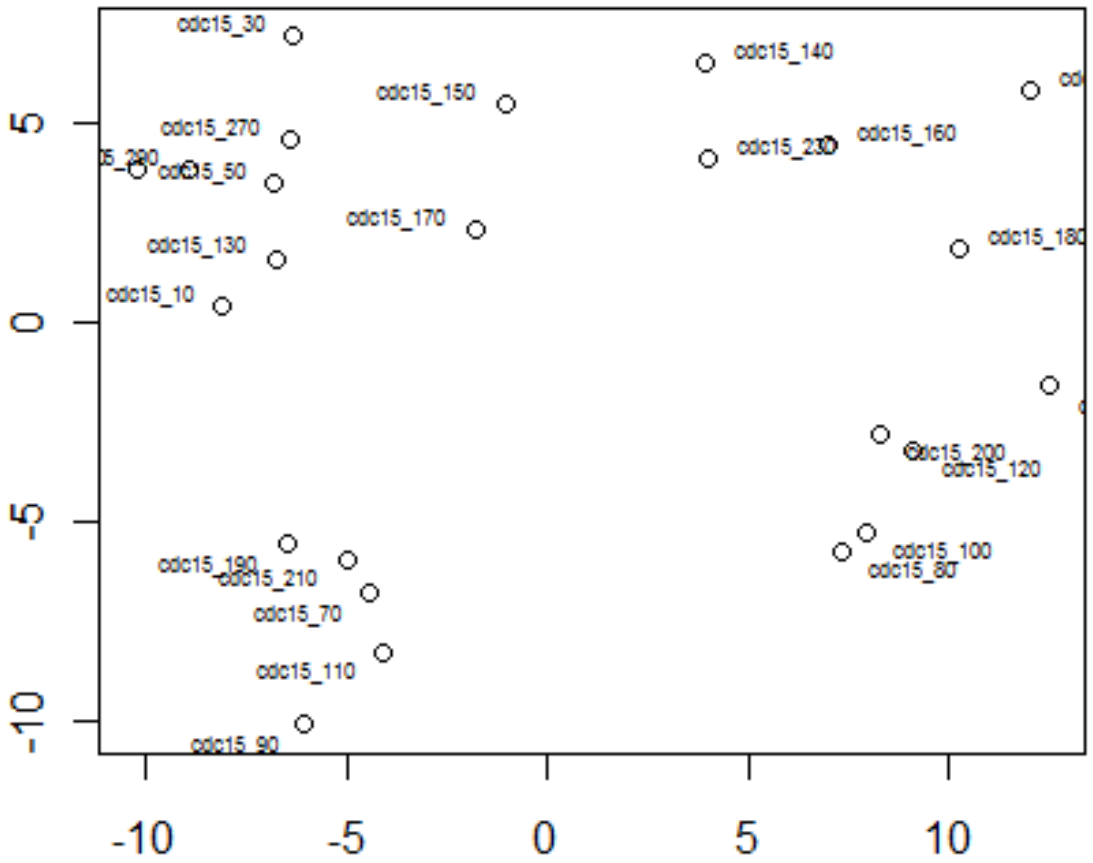




# Метод главных компонент (PCA)

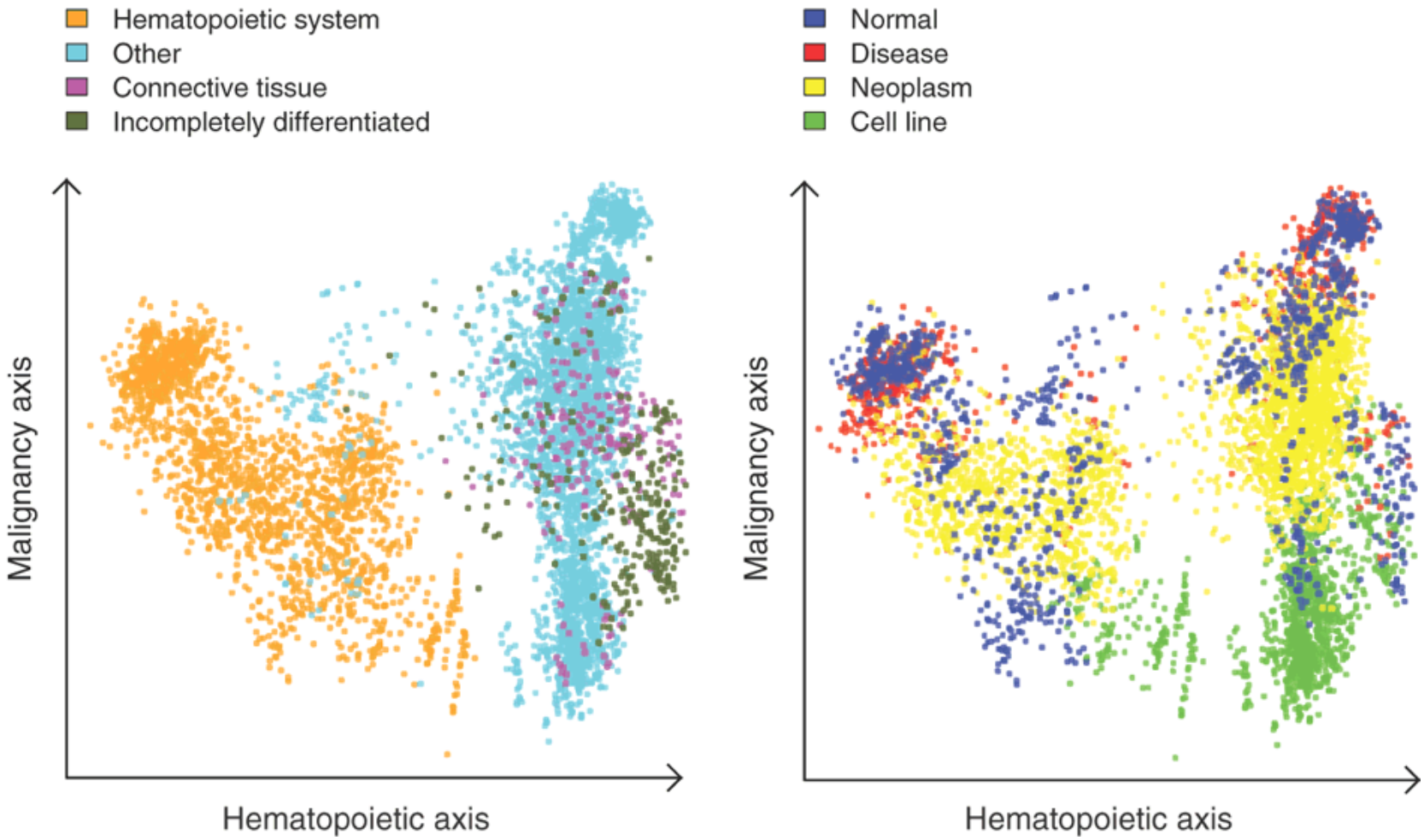



```
> pc = prcomp(expdata.sample, retx=TRUE)
> plot(pc)
> plot(pc$x[,1], pc$x[,2])
```




```
> pc = prcomp(t(expdata.sample), retx=TRUE)
```

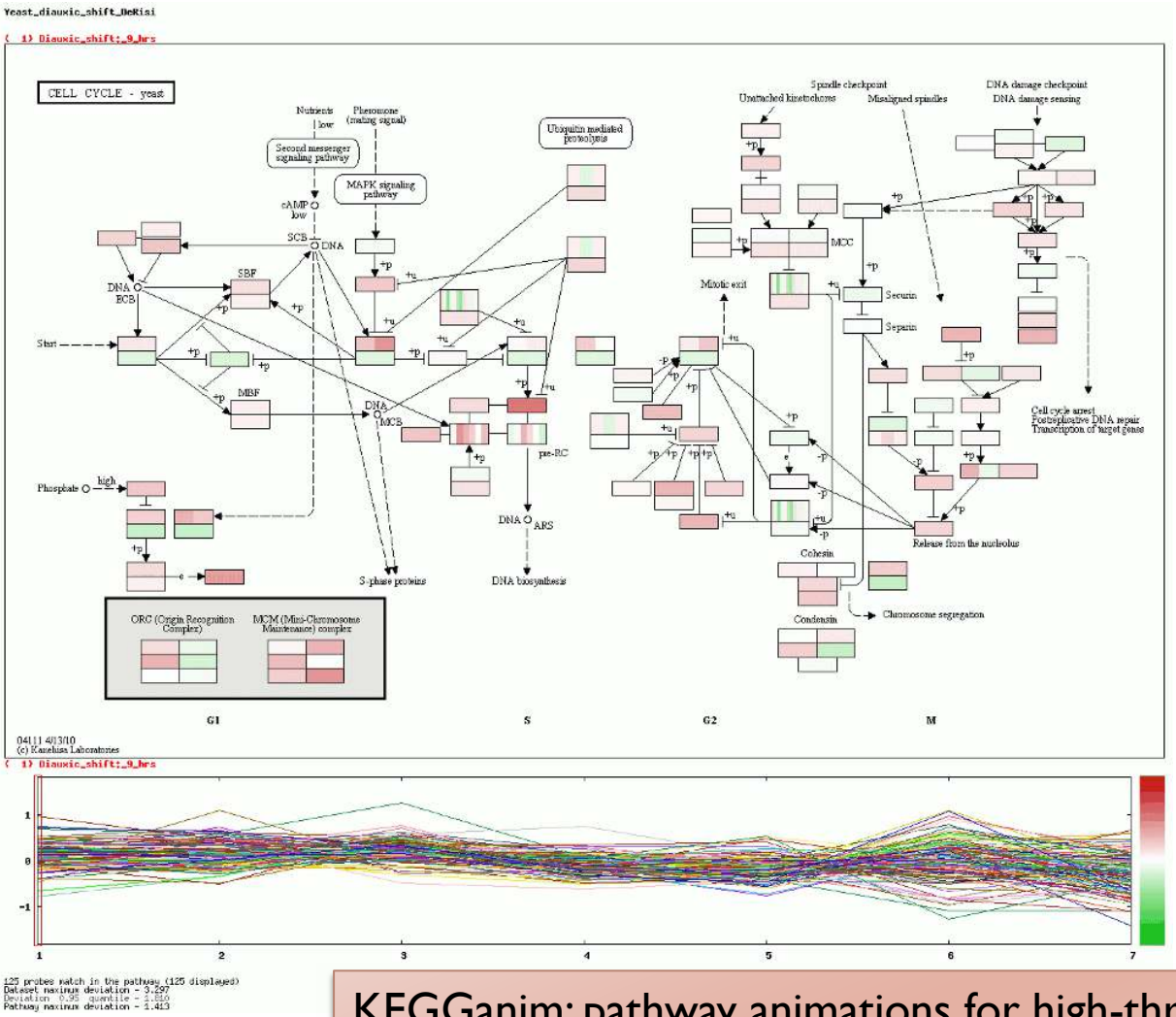
# PCA (5372 микрочипа)




A global map of human gene expression
**Margus Lukk, Misha Kapushesky, et al. (2010)**

# KeggAnim

**http://biit.cs.ut.ee/kegganim**




KEGGanim: pathway animations for high-throughput data. **Priit Adler, et al. (2008)**

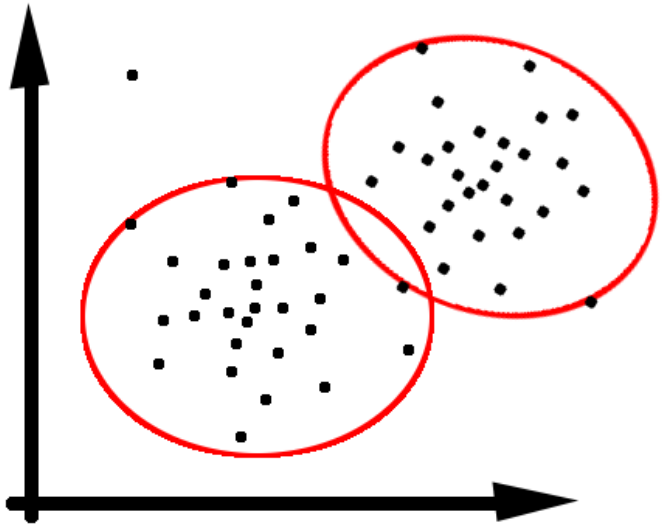
# Мораль

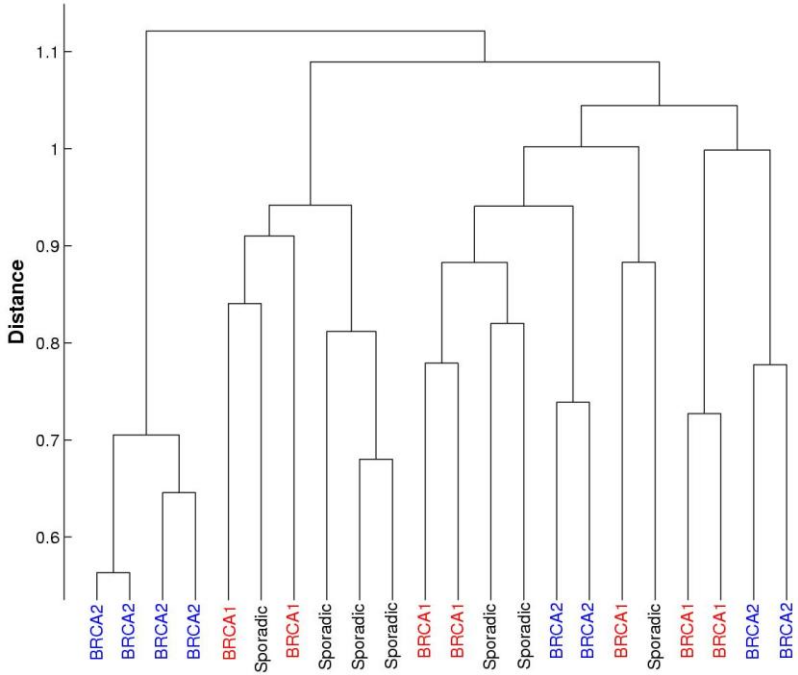
- Есть множество способов визуализации

- Идеального нет. Придумывайте еще!

о-в. Приветствия
море Демагогии
м.Технологий
респ. Нуклеотидов
земля Смысла
б-о Проблем
а.о. Многомерного анализа
страна Экспрессии

# Кластеризация





# Расстояние между генами

- Евклидово

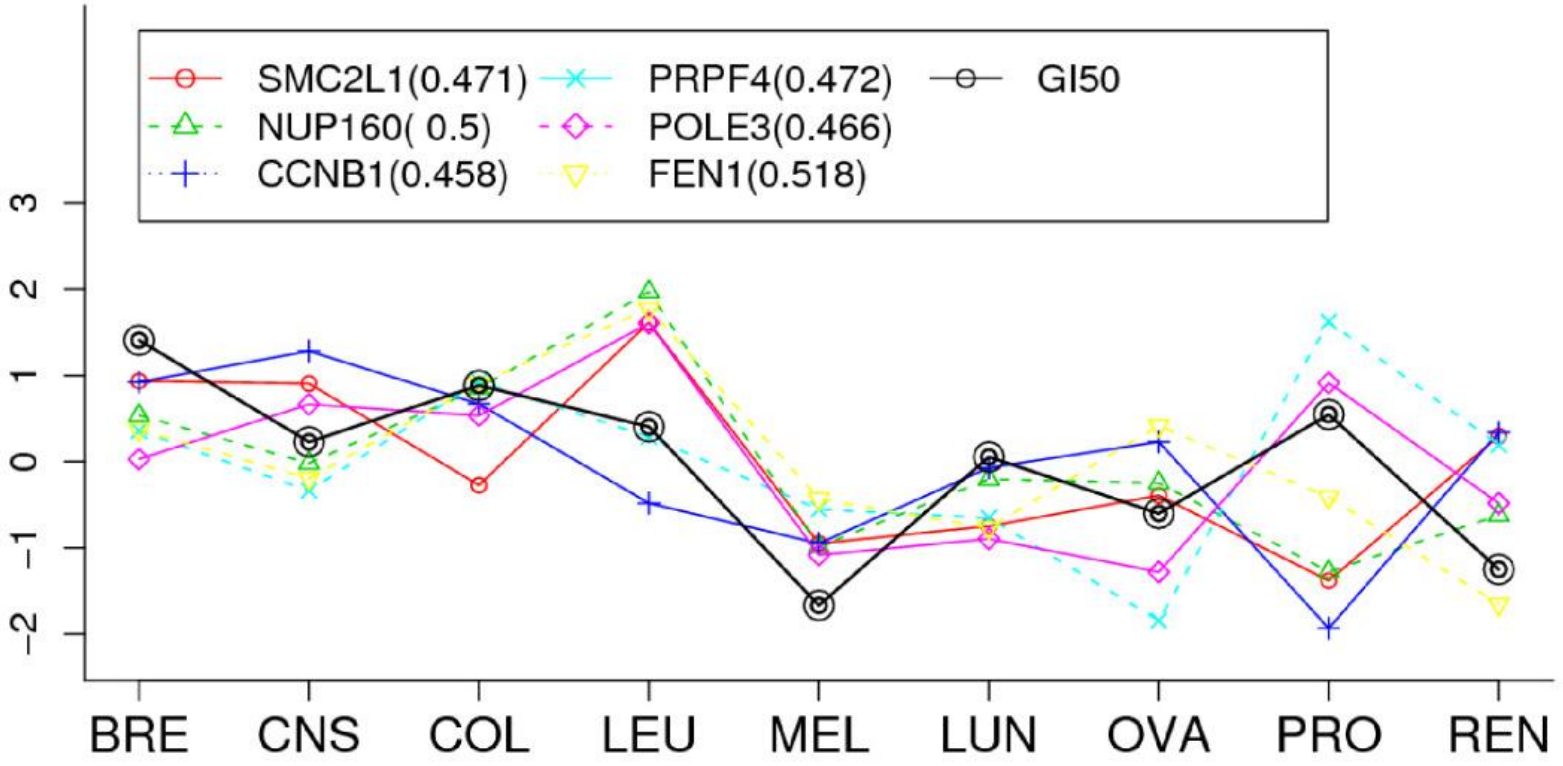
$$D(x,y) = (x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_m - y_m)^2$$

- Корреляция

$$\tilde{x}_i = \frac{x_i - \mathrm{mean}(x)}{\mathrm{stdev}(x)}$$

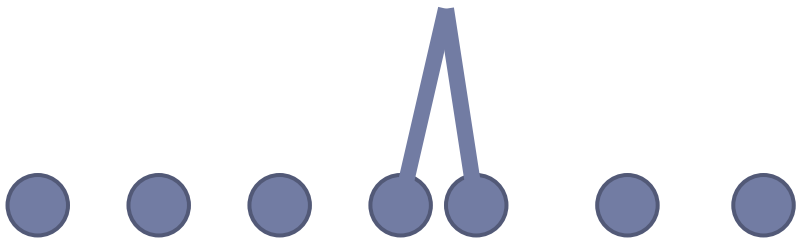
$$1 - D(x,y) = \tilde{x}_1\tilde{y}_1 + \tilde{x}_2\tilde{y}_2 + \ldots + \tilde{x}_m\tilde{y}_m$$

# Корреляция

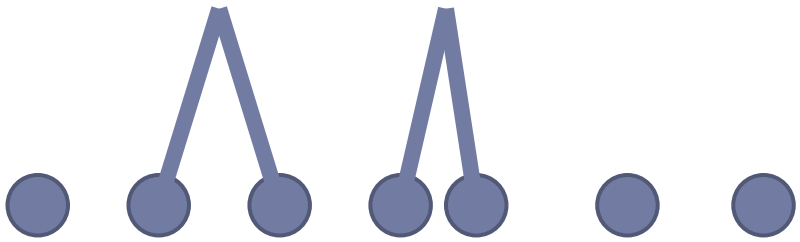

SMC2L1(0.471)
PRPF4(0.472)
GI50
NUP160( 0.5)
POLE3(0.466)
CCNB1(0.458)
FEN1(0.518)
3
2
1
0
−1
−2
BRE
CNS
COL
LEU
MEL
LUN
OVA
PRO
REN

# Иерархическая кластеризация

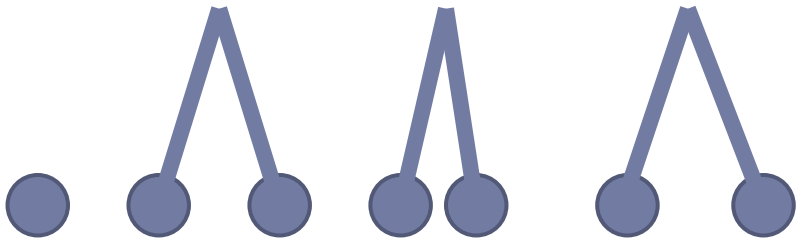
# Иерархическая кластеризация



# Иерархическая кластеризация

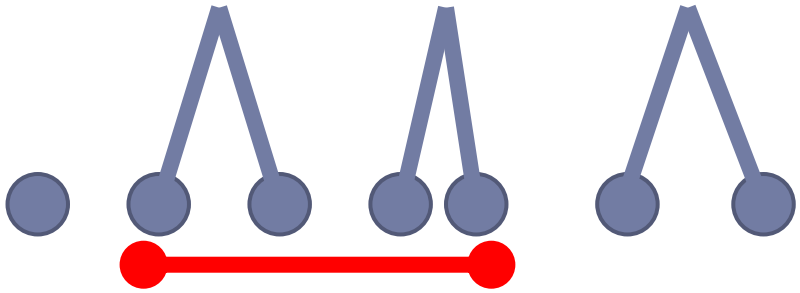


# Иерархическая кластеризация

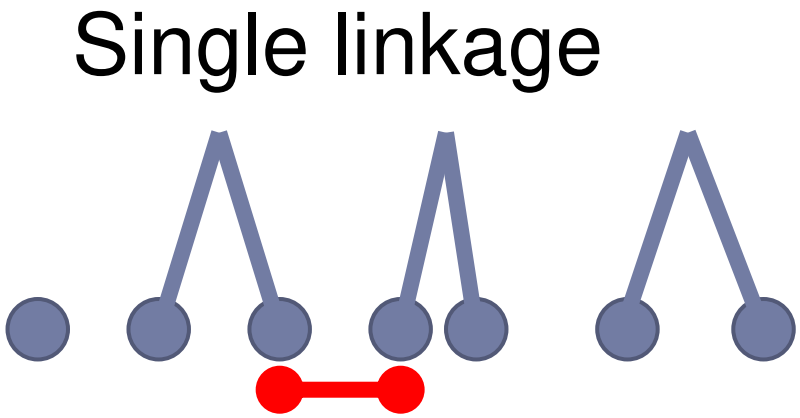


# Иерархическая кластеризация

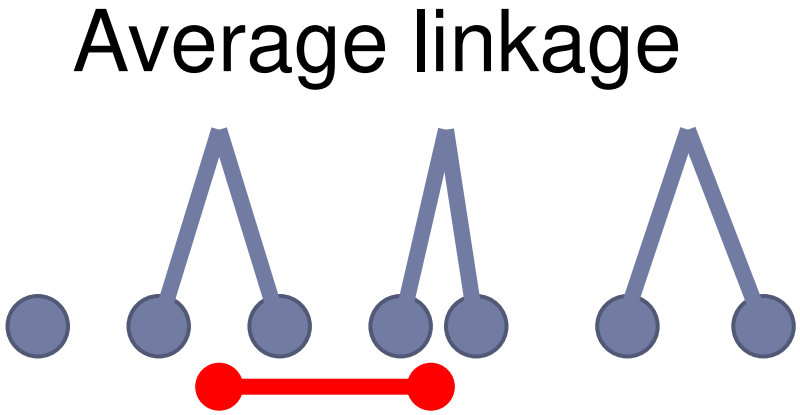
Complete linkage



# Иерархическая кластеризация

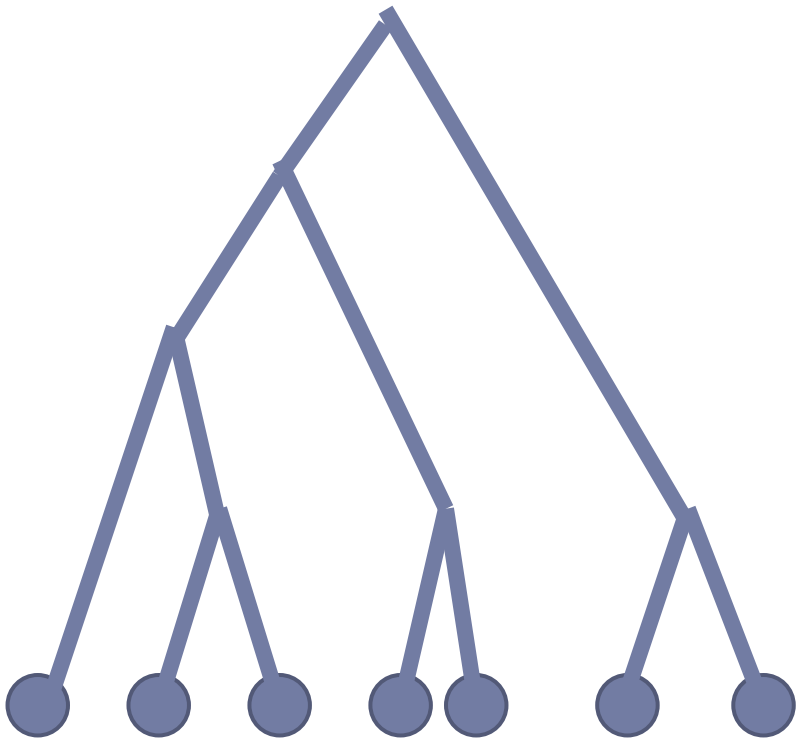


# Иерархическая кластеризация

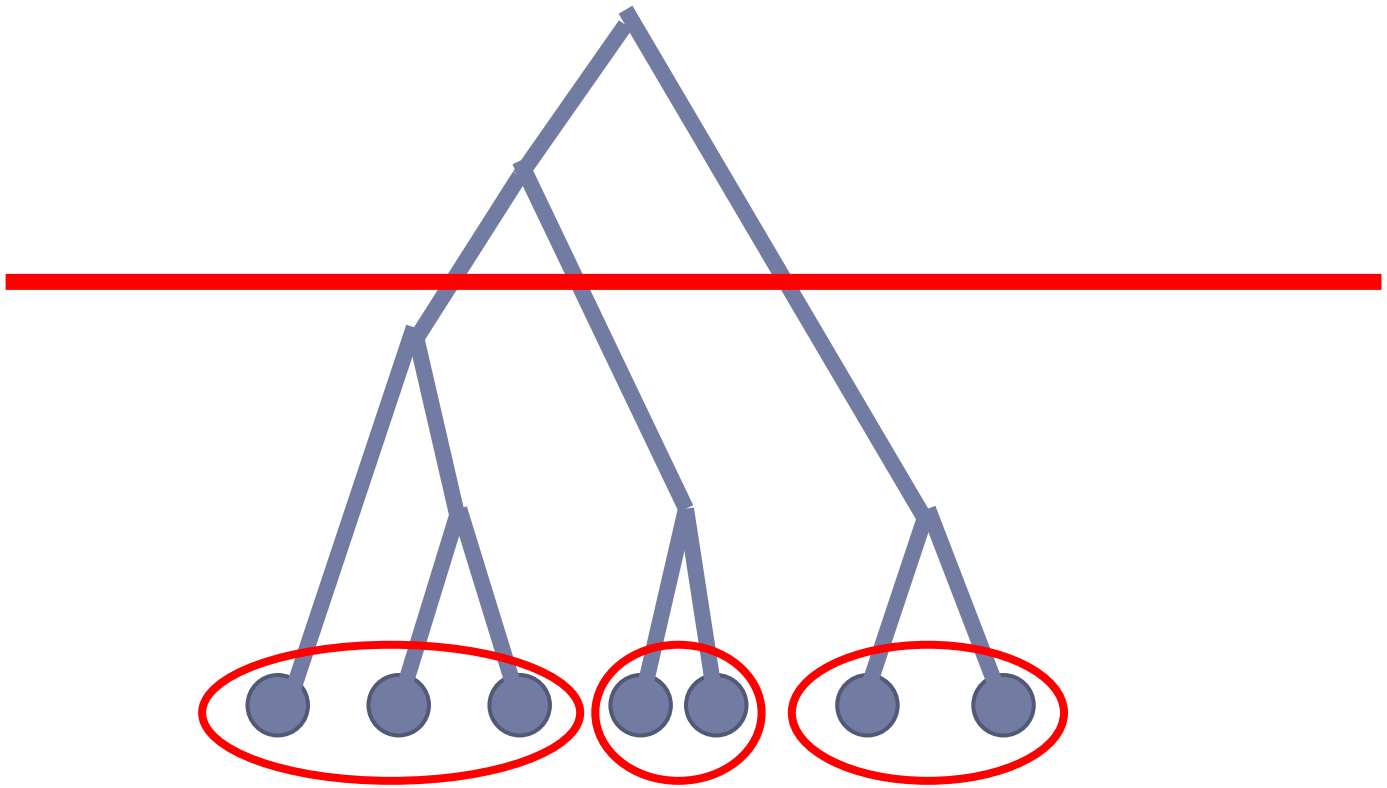


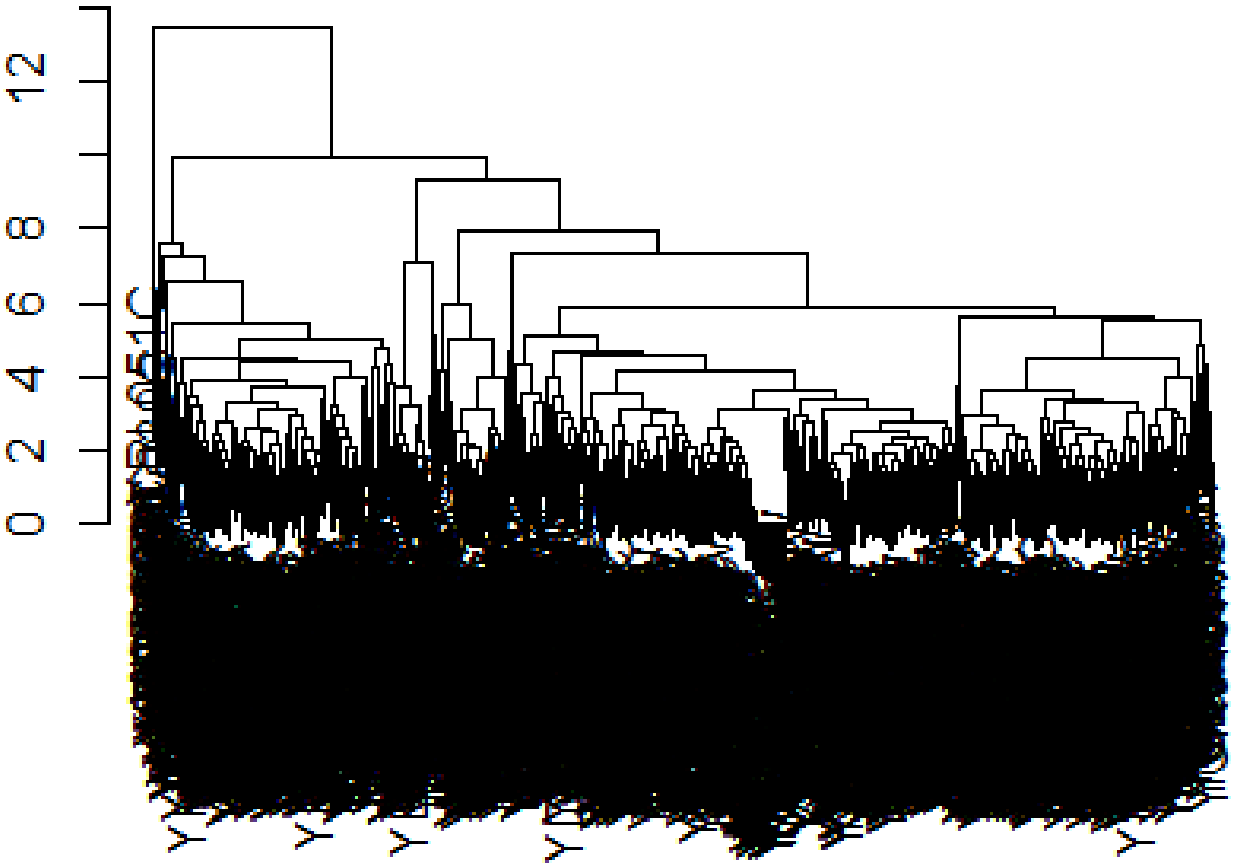
# Иерархическая кластеризация

# Иерархическая кластеризация



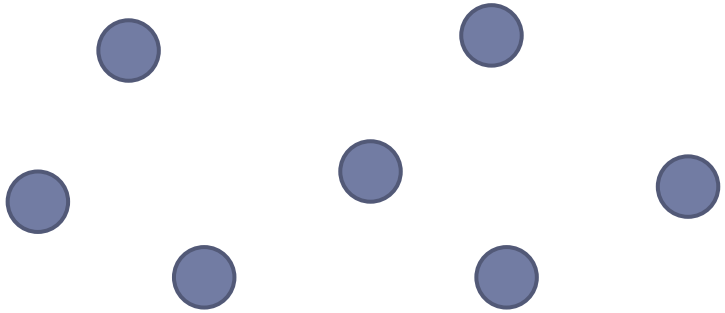
# Иерархическая кластеризация



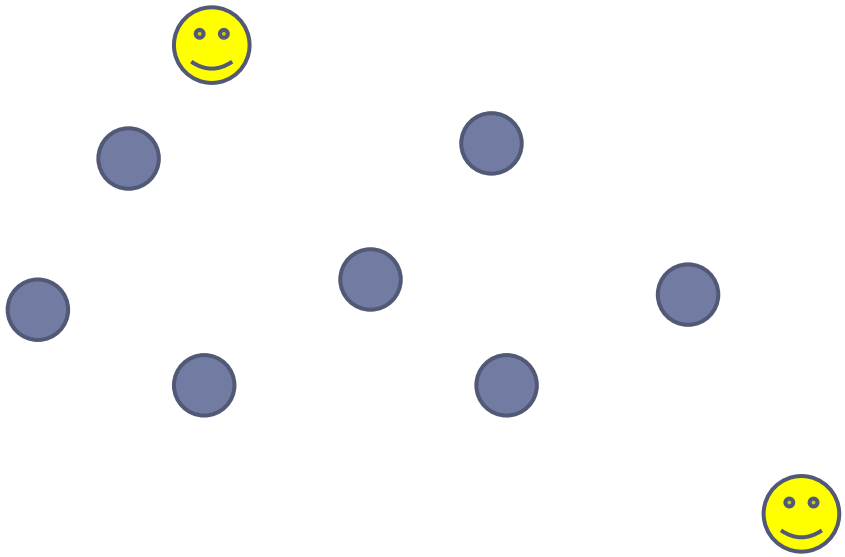



```
> c = hclust(dist(expdata.sample, "euclidean"), "complete")
> plot(c)
```

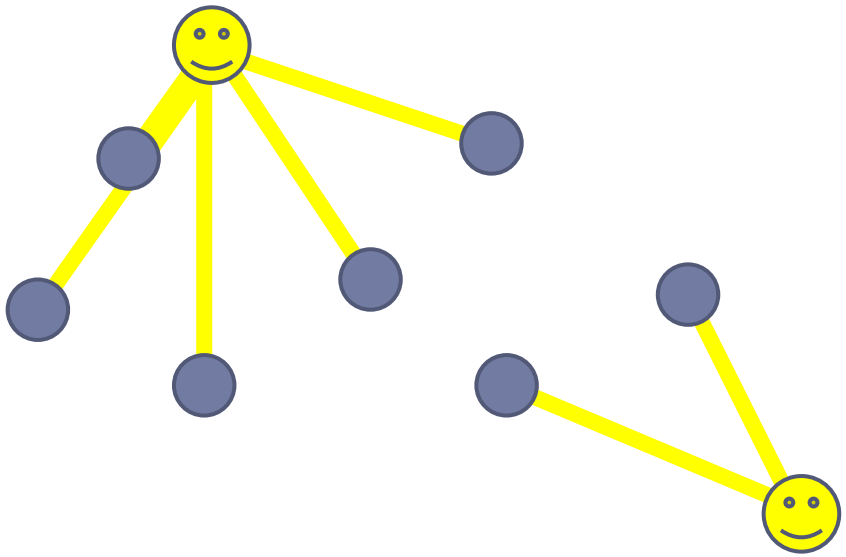
# Метод К-средних



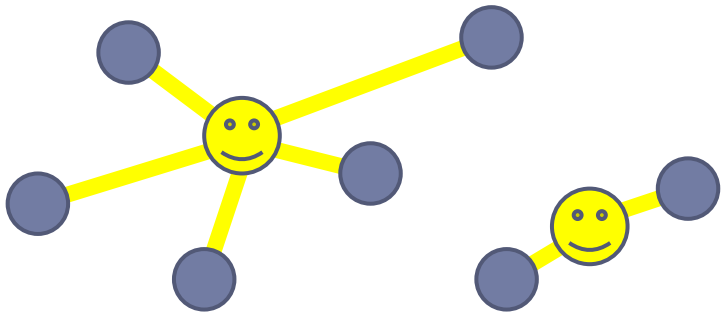
# Метод К-средних



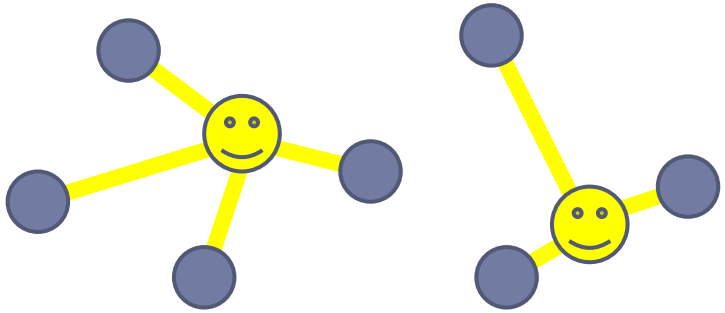
# Метод К-средних



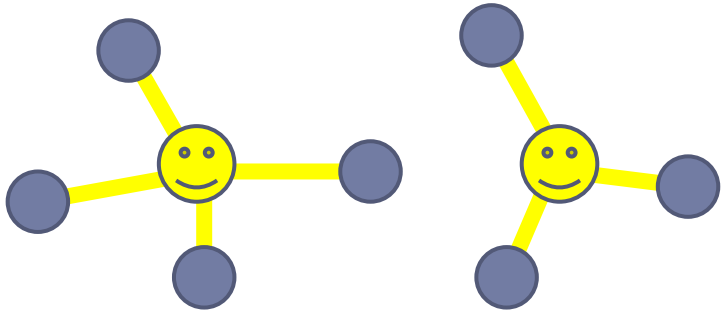
# Метод К-средних



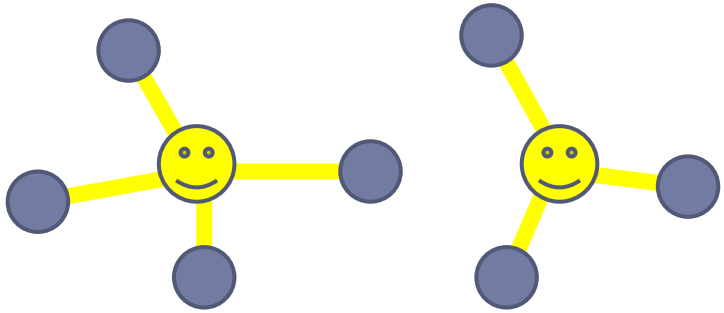
# Метод К-средних



# Метод К-средних

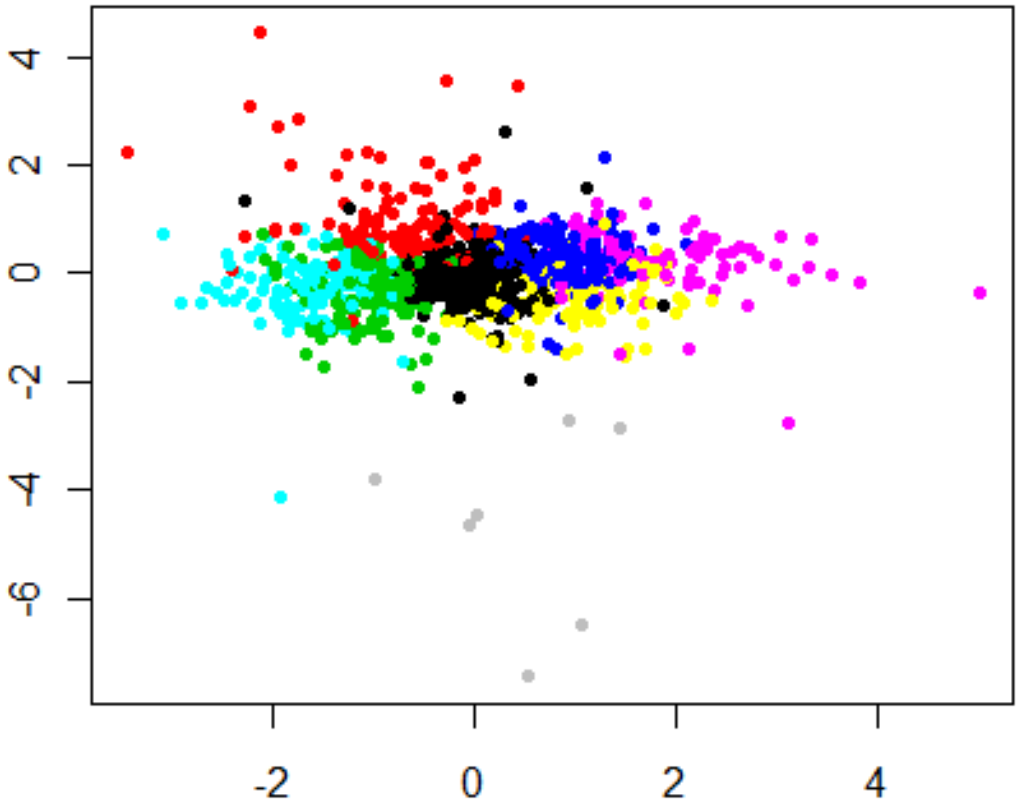


# Метод К-средних



**Метод локально минимизирует сумму квадратов (евклидовых) расстояний до центров кластеров**

# Метод К-средних




```
> c = kmeans(expdata.sample, 10)
> plot(pc$x[,1], pc$x[,2], col=c$cluster)
```

# Есть еще куча способов

**Bioconductor version 2.7 (Release)**

- AnnotationData (510)
- ExperimentData (68)
- Software (408)
  - Annotation (55)
  - AssayDomains (161)
  - AssayTechnologies (254)
  - Bioinformatics (240)
    - Classification (27)
    - Clustering (32)
    - MultipleComparisons (29)
    - Preprocessing (78)
    - QualityControl (36)
    - SequenceMatching (6)

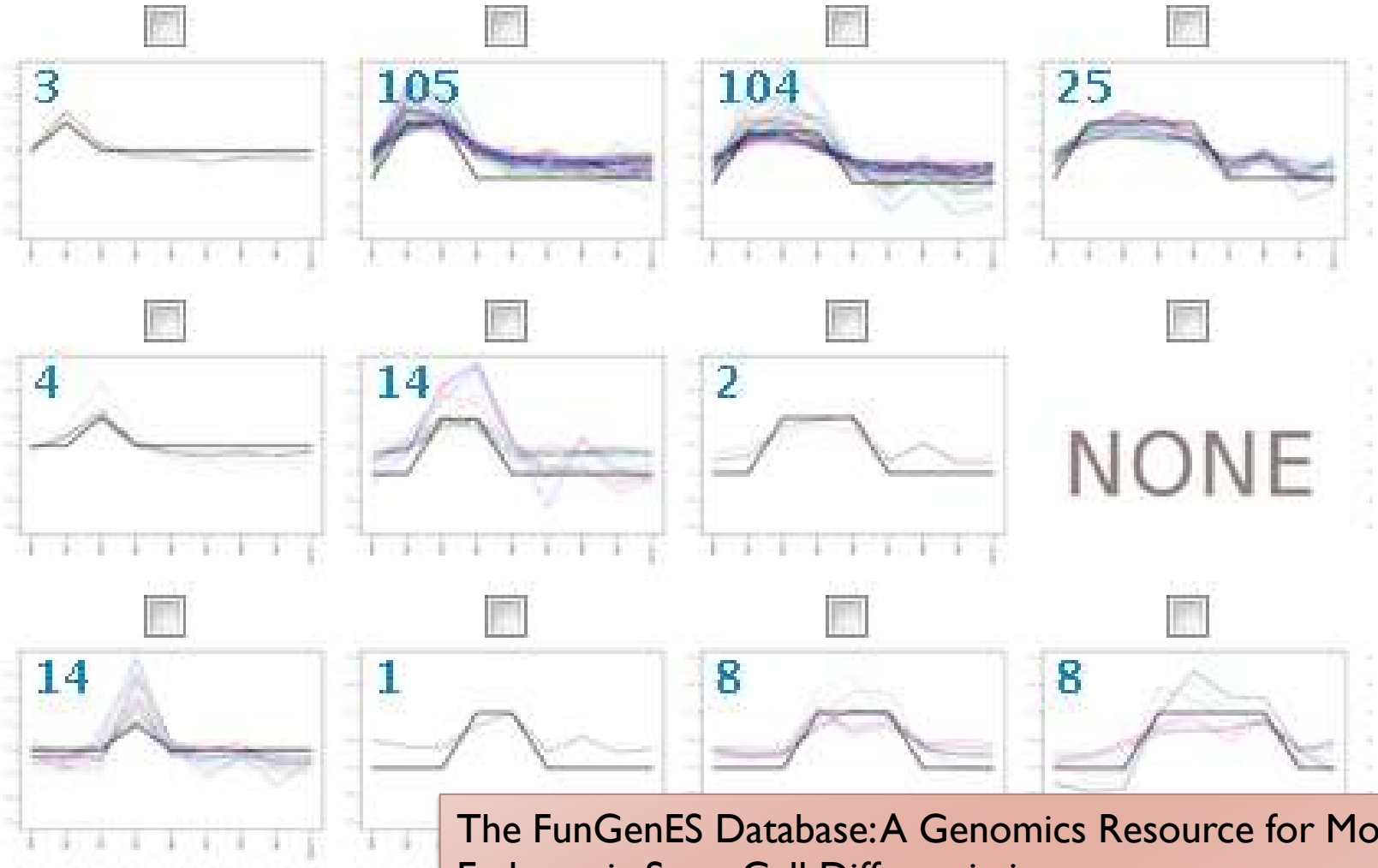
**Packages**

Software > Bioinformatics > Clustering

adSplit • BHC • BicARE • CGEN • ChemmineR • clusterStab • ConsensusClusterPlus • CORREP • ctc • fabia • flowClust • flowFP • flowMeans • flowMerge • geneRecommender • GOSemSim • hopach • maanova • made4 • maigesPack • MantelCorr • methVisual • Mfuzz • MLInterfaces • netresponse • PICS • puma • RBioinf • RTools4TB • SAGx • SamSPECTRAL • SpeCond

http://biit.cs.ut.ee/fungenes

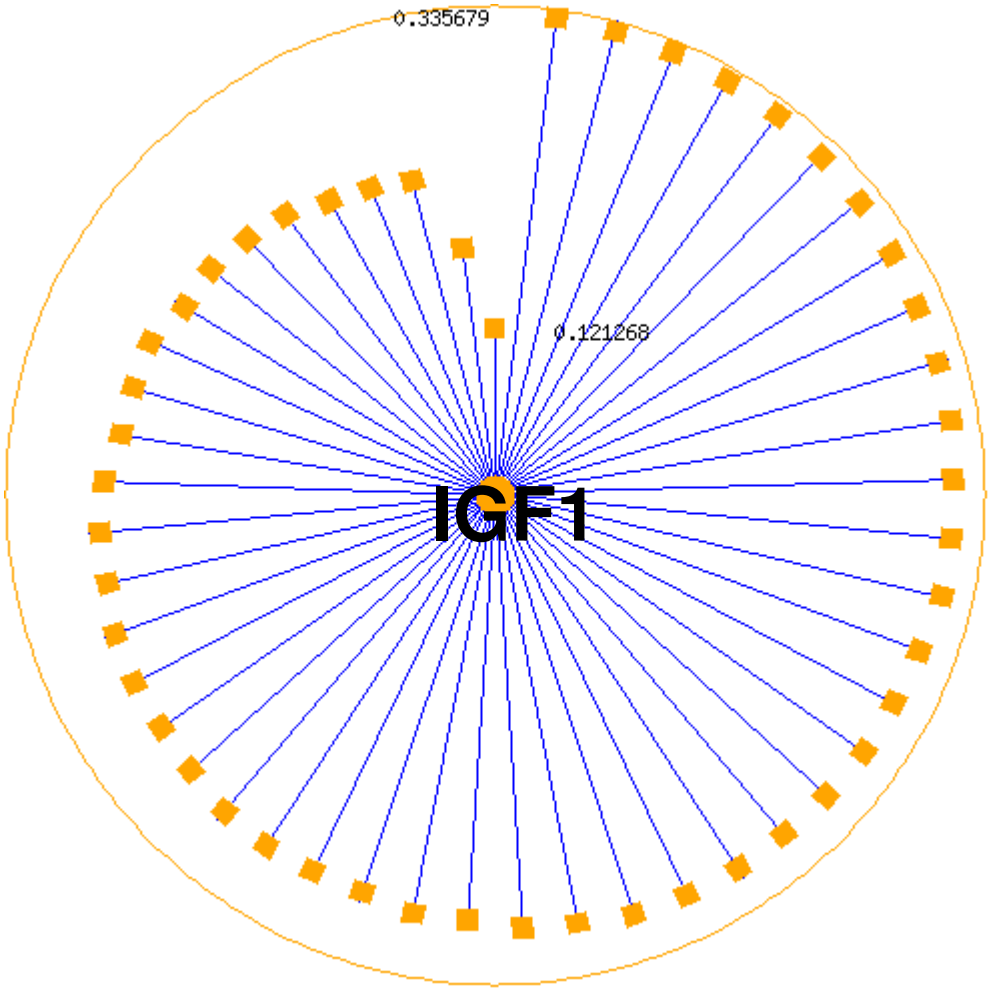
# Иногда можно группировать проще




The FunGenES Database: A Genomics Resource for Mouse Embryonic Stem Cell Differentiation.
**Herbert Shulz, Raivo Kolde, Priit Adler, et al. (2009)**

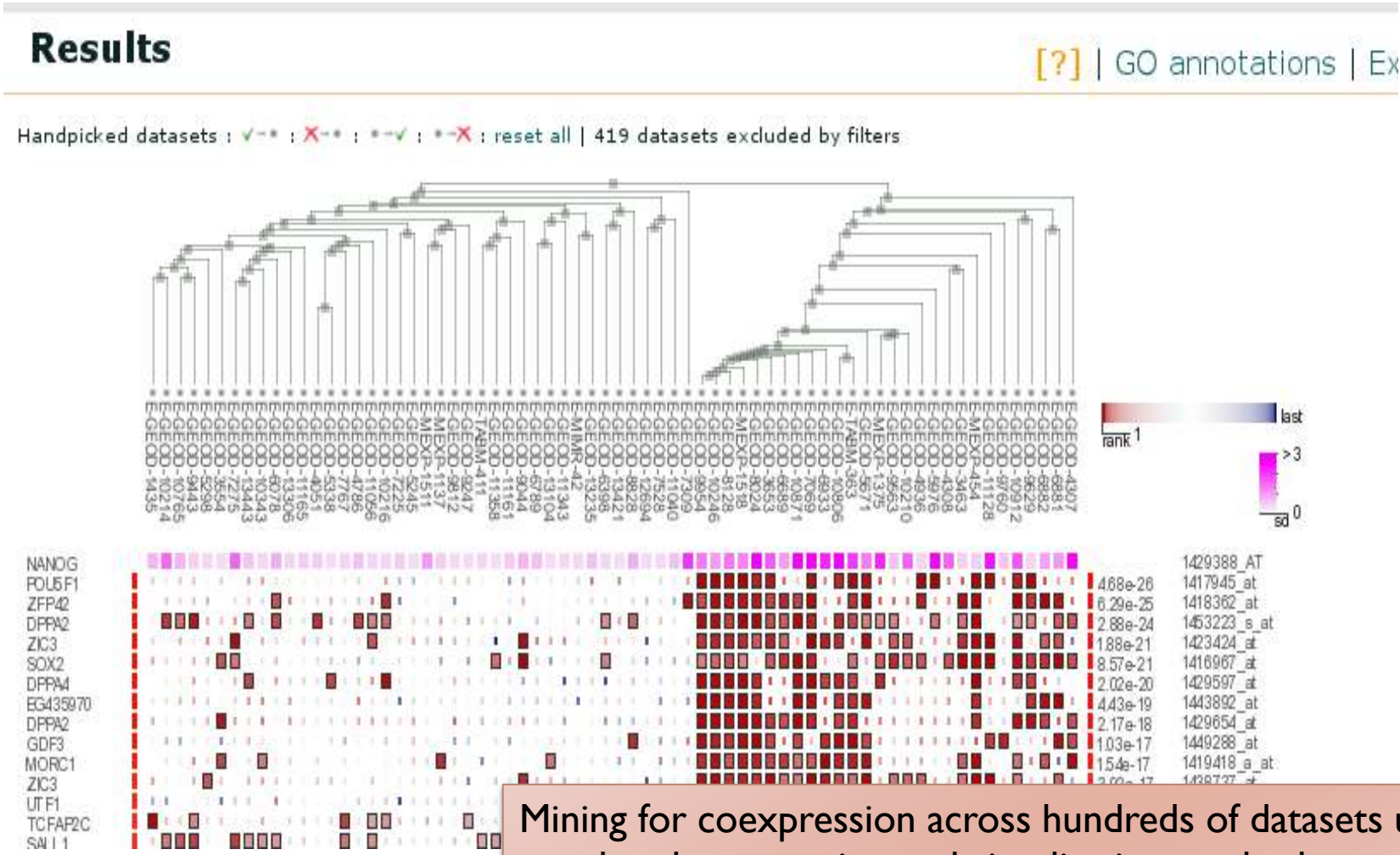
http://biit.cs.ut.ee/gprofiler/gsorter.cgi

# Или еще проще



http://biit.cs.ut.ee/mem

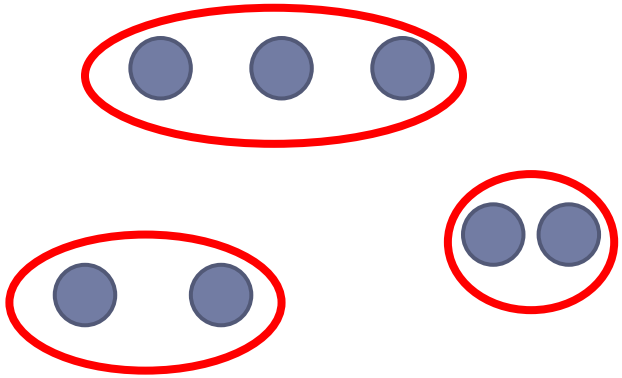
# ..или с *большим* размахом




Mining for coexpression across hundreds of datasets using novel rank aggregation and visualisation methods
**Priit Adler, et al. (2009)**

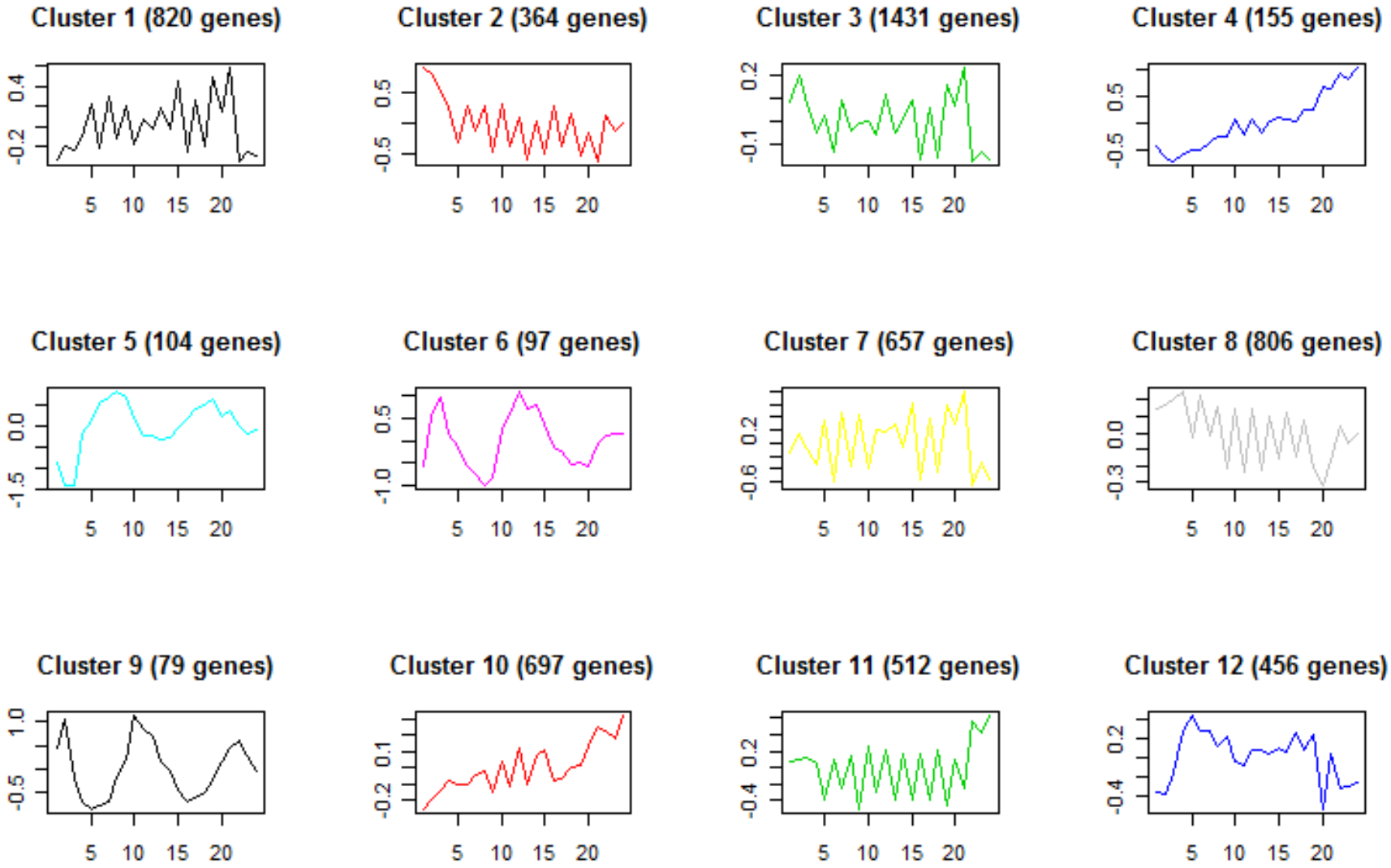
# Но в конце концов…

Имеем набор групп

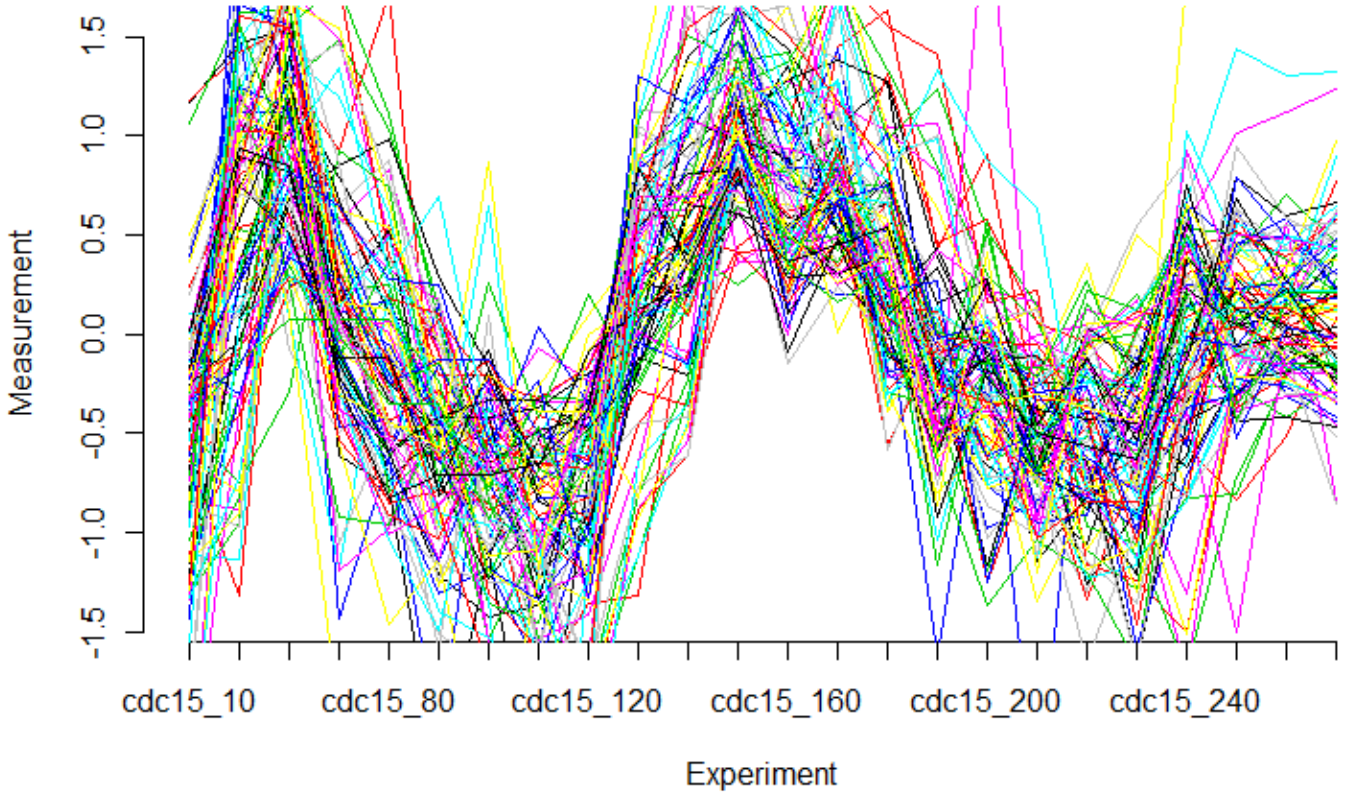


о-в. Приветствия
море Демагогии
м.Технологий
респ. Нуклеотидов
земля Смысла
б-о Проблем
а.о. Многомерного анализа
страна Экспрессии

```
> c = kmeans(expdata, 12)
> csize = table(c$cluster)
   1    2    3    4    5    6    7    8    9   10   11   12
 820  364 1431  155  104   97  657  806   79  697  512  456

> par(mfrow=c(3,4))
> for (i in 1:12) plot(c$centers[i,], type='l')
```




```
> my = which(c$cluster == 6)
> expdata.my = expdata[my,]
```


Measurement
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
cdc15_10
cdc15_80
cdc15_120
cdc15_160
cdc15_200
cdc15_240
Experiment

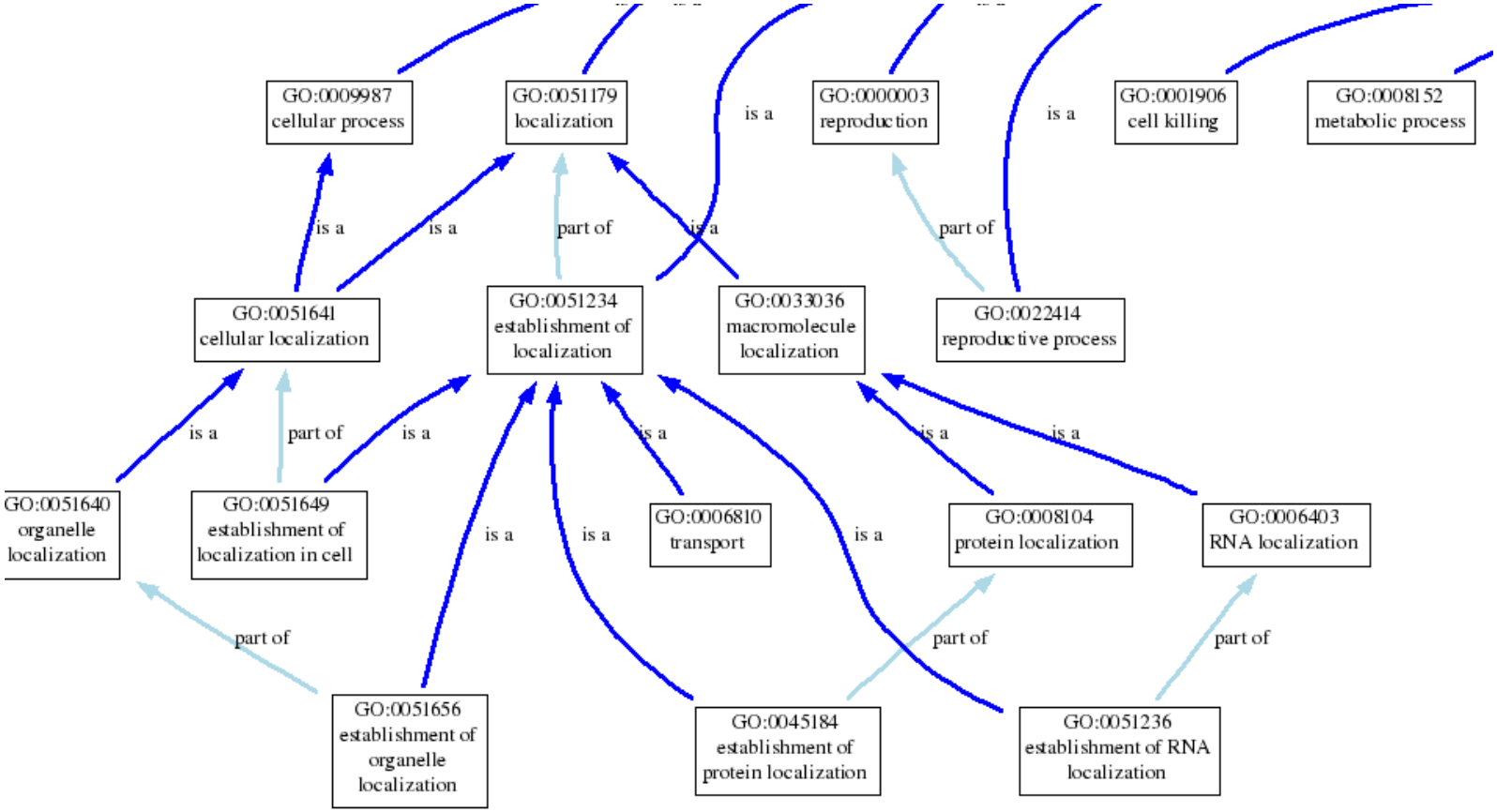
о-в. Приветствия
море Демагогии
м.Технологий
респ. Нуклеотидов
земля Смысла
б-о Проблем
а.о. Многомерного анализа
страна Экспрессии

```
> gene_names = rownames(expdata.my)
  [1] "YAR007C" "YBL002W" "YBL003C" "YBL035C" "YBR009C"
  [6] "YBR010W" "YBR073W" "YBR088C" "YBR089W" "YCL060C"
 [11] "YCR065W" "YDL003W" "YDL018C" "YDL055C" "YDL101C"
 [16] "YDR097C" "YDR113C" "YDR224C" "YDR225W" "YDR309C"
 [21] "YDR353W" "YDR356W" "YDR503C" "YDR528W" "YER001W"
 [26] "YER003C" "YER091C" "YER095W" "YER124C" "YGL163C"
 [31] "YGR055W" "YGR109C" "YGR151C" "YGR152C" "YGR189C"
 [36] "YGR221C" "YHR110W" "YHR143W" "YIL066C" "YIL140W"
 [41] "YIL141W" "YJL074C" "YJL078C" "YJL115W" "YJL173C"
 [46] "YJL187C" "YJR148W" "YKL001C" "YKL008C" "YKL045W"
 [51] "YKL066W" "YKL067W" "YKL101W" "YKL127W" "YKR012C"
 [56] "YKR013W" "YLL002W" "YLL022C" "YLR103C" "YLR121C"
 [61] "YLR183C" "YLR212C" "YLR286C" "YLR300W" "YLR313C"
 [66] "YLR326W" "YLR383W" "YML027W" "YMR179W" "YMR199W"
 [71] "YNL030W" "YNL031C" "YNL262W" "YNL263C" "YNL300W"
 [76] "YNR066C" "YOL007C" "YOL017W" "YOL019W" "YOL090W"
```

# Cell division genes

```
> gene_names = rownames(expdata.my)
 [1] "YAR007C" "YBL002W" "YBL003C" "YBL035C" "YBR009C"
 [6] "YBR010W" "YBR073W" "YBR088C" "YBR089W" "YCL060C"
[11] "YCR065W" "YDL003W" "YDL018C" "YDL055C" "YDL101C"
[16] "YDR097C" "YDR113C" "YDR224C" "YDR225W" "YDR309C"
[21] "YDR353W" "YDR356W" "YDR503C" "YDR528W" "YER001W"
[26] "YER003C" "YER091C" "YER095W" "YER124C" "YGL163C"
[31] "YGR055W" "YGR109C" "YGR151C" "YGR152C" "YGR189C"
[36] "YGR221C" "YHR110W" "YHR143W" "YIL066C" "YIL140W"
[41] "YIL141W" "YJL074C" "YJL078C" "YJL115W" "YJL173C"
[46] "YJL187C" "YJR148W" "YKL001C" "YKL008C" "YKL045W"
[51] "YKL066W" "YKL067W" "YKL101W" "YKL127W" "YKR012C"
[56] "YKR013W" "YLL002W" "YLL022C" "YLR103C" "YLR121C"
[61] "YLR183C" "YLR212C" "YLR286C" "YLR300W" "YLR313C"
[66] "YLR326W" "YLR383W" "YML027W" "YMR179W" "YMR199W"
[71] "YNL030W" "YNL031C" "YNL262W" "YNL263C" "YNL300W"
[76] "YNR066C" "YOL007C" "YOL017W" "YOL019W" "YOL090W"
```

# Генетическая Онтология




**http://www.geneontology.org**

# Генетическая Онтология

```
> library(GO.db)
> GOTERM[["GO:0007117"]]
GOID: GO:0007117
Term: budding cell bud growth
Ontology: BP
Definition: The process by which the bud portion of a cell
    reproduces by budding irreversibly increases in size o
    by accretion and biosynthetic production of matter sim
    that already present.
Synonym: bud growth
```

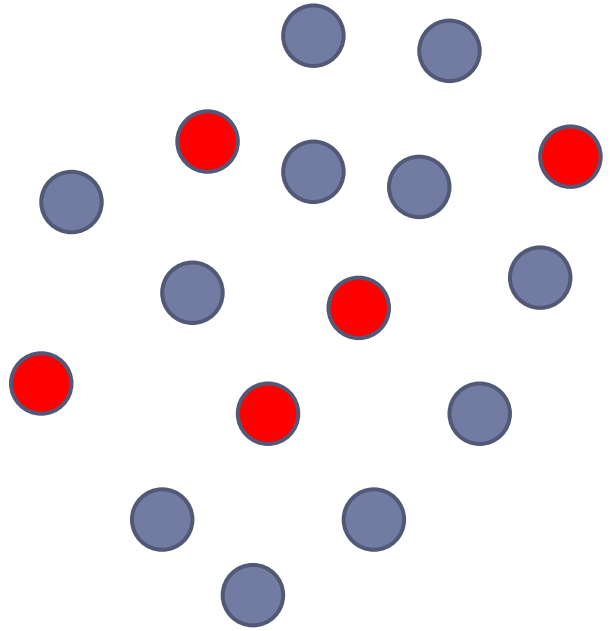
# Аннотации

```
> library(org.Sc.sgd.db)
> names(org.Sc.sgdGO[["YGR221C"]])
[1] "GO:0007117" "GO:0007117" "GO:0005886" "GO:0005934" "GO
[6] "GO:0016020" "GO:0016021" "GO:0033101" "GO:0003674"
```

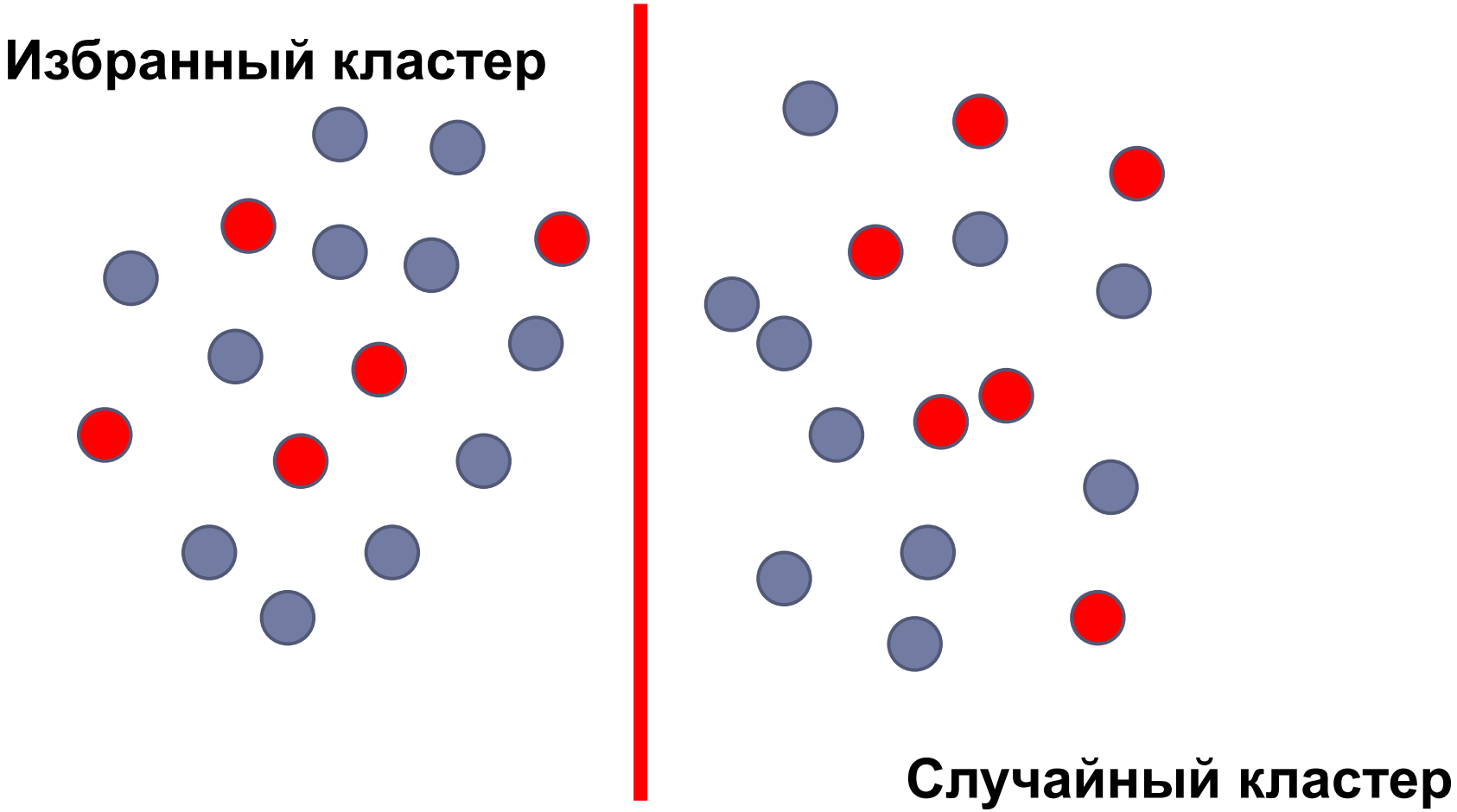
# Аннотации

```
> library(org.Sc.sgd.db)
> names(org.Sc.sgdGO[["YGR221C"]])
[1] "GO:0007117" "GO:0007117" "GO:0005886" "GO:0005934" "GO
[6] "GO:0016020" "GO:0016021" "GO:0033101" "GO:0003674"


> org.Sc.sgdGO[["YGR221C"]][[1]]
$GOID
[1] "GO:0007117"

$Evidence
[1] "IMP"

$Ontology
[1] "BP"
```

# Статистическая значимость?



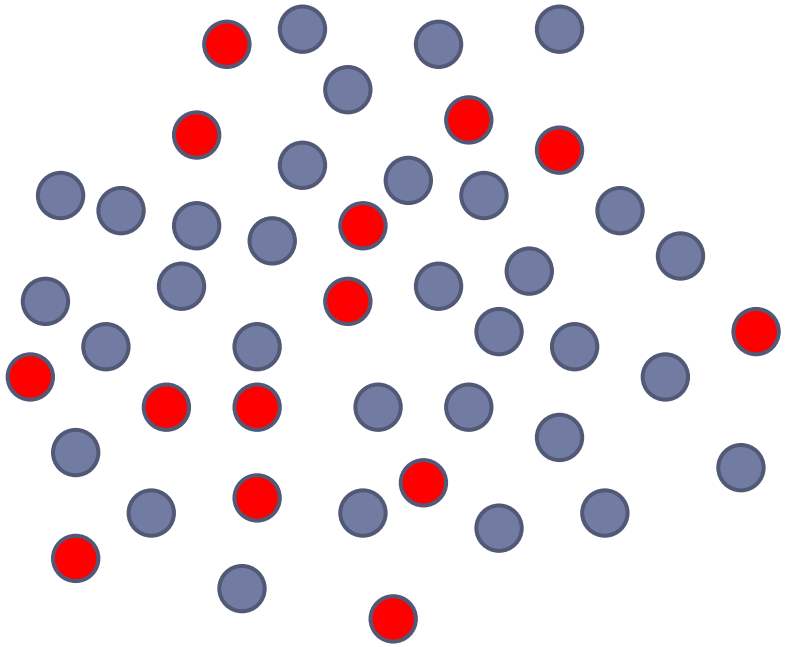
# Статистическая значимость?



# Мера интересности

**P**[в случайно взятом кластере будет как минимум столько же представителей интересующей нас функции]

# Теория вероятностей

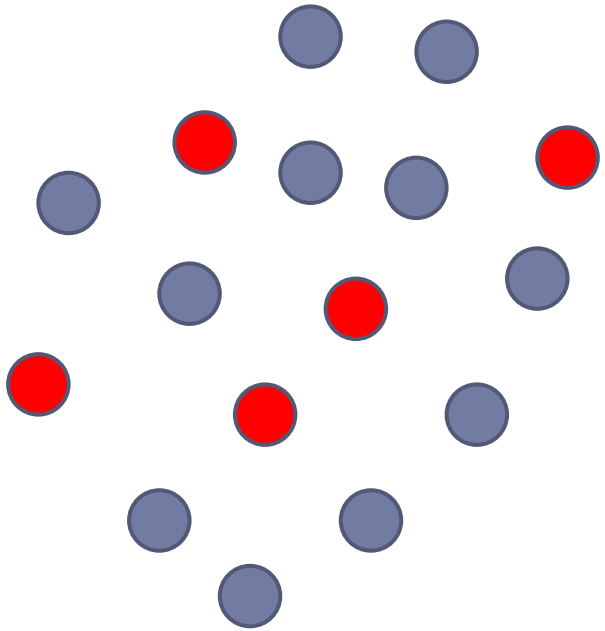


**Популяция**

**6178 генов,**
**из которых 40 относятся**
**к данной функции**

```
> nrow(expdata)
[1] 6178

> go7117 = unique(org.Sc.sgdGO2ALLORFS[["GO:0007117"]])
> length(go7117)
[1] 40
```

# Теория вероятностей



**Кластер**

**97 генов,**
**из которых 1 относится**
**к данной функции**

```
> length(gene_names)
[1] 97

> length(intersect(go7117, gene_names))
[1] 1
```

# Гипергеометрическое распределение

**Популяция**

**6178 генов,**
**из которых 40 относятся**
**к данной функции**

**Кластер**

**97 генов,**
**из которых 1 относится**
**к данной функции**

$$P(X = k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}$$

40 6178-40
1 96
6178
97

# Гипергеометрическое распределение

**Популяция**

**6178 генов,**
**из которых 40 относятся**
**к данной функции**

**Кластер**

**97 генов,**
**из которых 1 относится**
**к данной функции**

```
> dhyper(1, 40, 6178-40, 97)
[1] 0.3402958
```

```
> 1-phyper(0, 40, 6178-40, 97)
[1] 0.4700858
```

# Аннотация кластера

- Для каждой функциональной категории
  - Считаем сколько раз она присутствует **всего**
  - Сколько раз она присутствует **в кластере**
  - Оцениваем вероятность получить **такой же или лучший** результат **случайно** (P-value)
  - Оставляем функциональные категории, у которых
    - P-value < 0.01

# Множественное тестирование

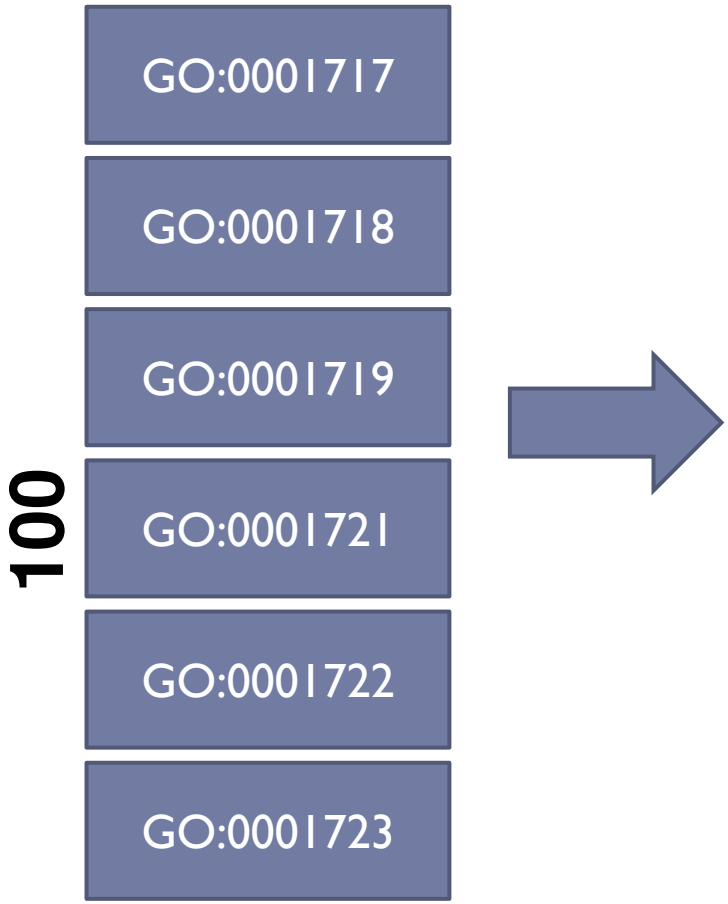
GO:0001717

**С вероятностью 0.01 мы присвоим эту категорию случайному кластеру.**

# Множественное тестирование

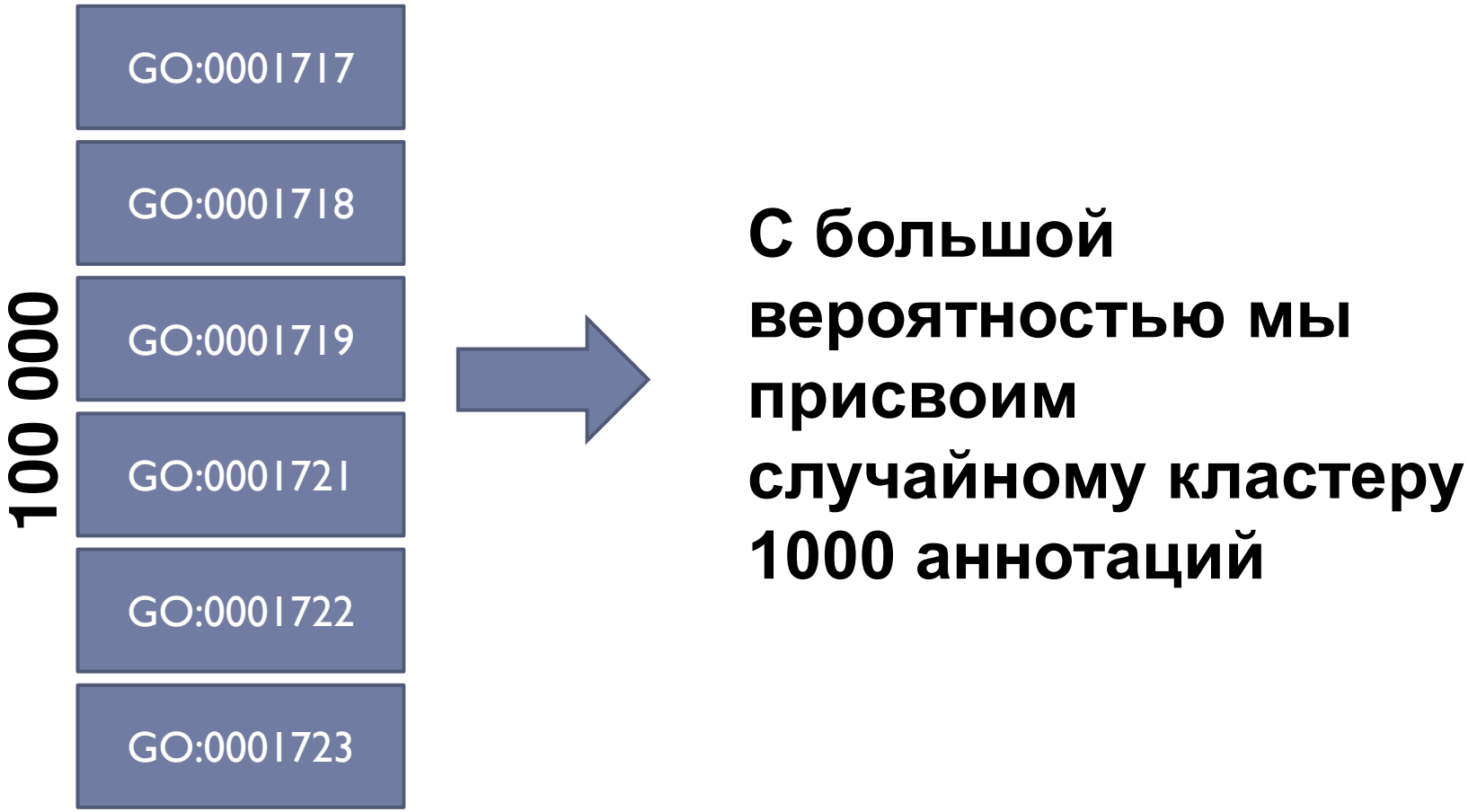
| | |
|---|---|
| GO:0001717 | **С вероятностью 0.01 мы присвоим эту категорию случайному кластеру.** |
| GO:0001718 | **С вероятностью 0.01 мы присвоим эту категорию случайному кластеру.** |
| GO:0001719 | **С вероятностью 0.01 мы присвоим эту категорию случайному кластеру.** |
| GO:0001721 | **С вероятностью 0.01 мы присвоим эту категорию случайному кластеру.** |
| GO:0001722 | **С вероятностью 0.01 мы присвоим эту категорию случайному кластеру.** |
| GO:0001723 | **С вероятностью 0.01 мы присвоим эту категорию случайному кластеру.** |

# Множественное тестирование




**С большой вероятностью мы присвоим случайному кластеру хотя бы одну аннотацию**

# Множественное тестирование




**С большой вероятностью мы присвоим случайному кластеру 1000 аннотаций**

# Коррекция Бонферрони

**Засчитываем все категории с P-value < 0.01 / 100 000**

- Для 100 000 терминов GO
  - Аннотация значительна если

**P-value < $10^{-7}$**

# False Discovery Rate

- Предположим мы делаем 100 000 тестов, и **все они** имеют

  **P-value = 0.01.**

# False Discovery Rate

- Предположим мы делаем 100 000 тестов, и **все они** имеют

  **P-value = 0.01.**

- Но для случайного кластера мы ожидаем только **1000** тестов с таким значением.

- **Если мы решим принять все эти тесты, мы рискуем что лишь 0.01 из них неверны.**

# False Discovery Rate

- Находим $P_{fdr}$ для которого

$$\frac{100\,000\ P_{fdr}}{\{\text{количество тестов, у которых P-value} < P_{fdr}\}} < 0.05$$

И засчитываем все категории с P-value < $P_{fdr}$

# Аннотация кластера

- Для каждой функциональной категории
  - Считаем сколько раз она присутствует **всего**
  - Сколько раз она присутствует **в кластере**
  - Оцениваем вероятность получить **такой же или лучший** результат **случайно** (P-value)
  - Оставляем функциональные категории, у которых
    - P-value < ПОПРАВКА(0.01)

```
> library(GOstats)
> params = new("GOHyperGParams",
+     geneIds=gene_names,
+     universeGeneIds=rownames(expdata),
+     annotation="org.Sc.sgd.db",
+     ontology="BP",
+     pvalueCutoff=0.01/1000,
+     testDirection="over",
+     conditional=TRUE)
> hgTest = hyperGTest(params)
```

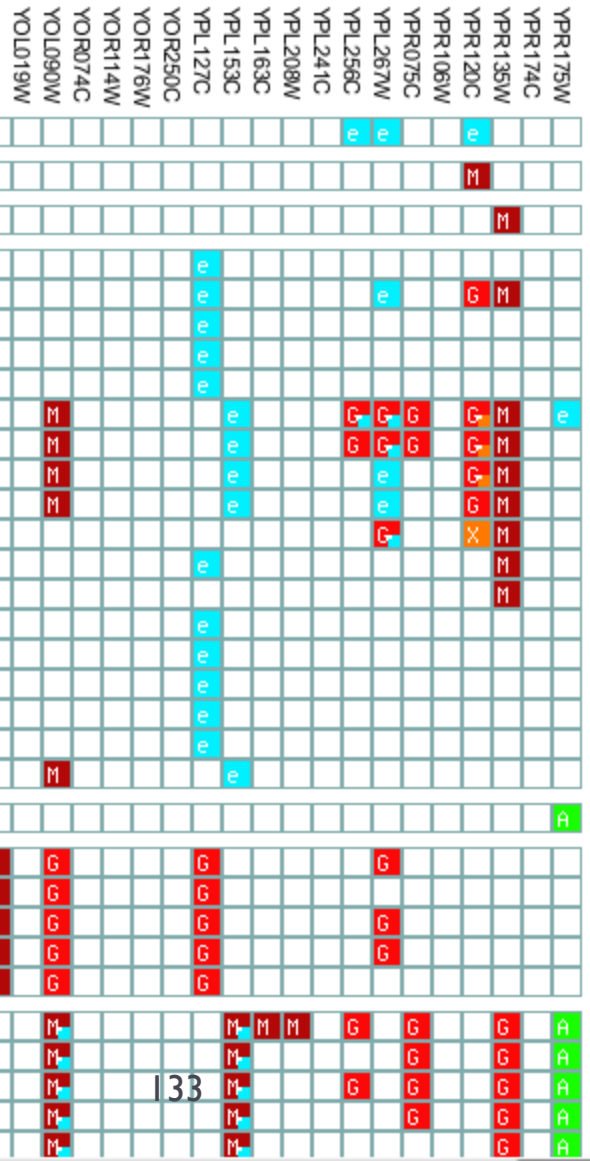
```
Gene to GO BP Conditional test for over-representation
733 GO BP ids tested (22 have p < 1e-05)
```

```
> paste(gene_names, collapse=" ")
[1] "YAR007C YBL002W YBL003C YBL035C YBR009C
```

# http://biit.cs.ut.ee/gprofiler

g:Profiler -- a web-based toolset for functional profiling of gene lists from large-scale experiments
**Jyri Reimand, et al. (2007)**

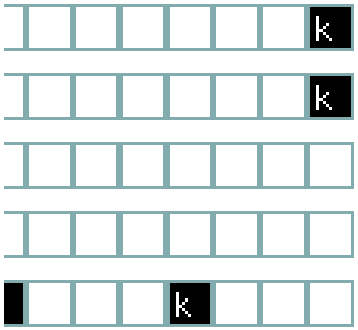
# http://biit.cs.ut.ee/gprofiler




| P-value | T | Q | Q&T | Q&T/Q | Q&T/T | term ID | term domain and name |
|---|---|---|---|---|---|---|---|
| 1.96e-06 | 293 | 97 | 17 | 0.175 | 0.058 | GO:0051301 | BP cell division (1) |
| 1.99e-05 | 76 | 97 | 8 | 0.082 | 0.105 | GO:0007346 | BP regulation of mitotic |
| 1.74e-05 | 22 | 97 | 5 | 0.052 | 0.227 | GO:0007064 | BP mitotic sister chromat |
| 6.62e-12 | 77 | 97 | 14 | 0.144 | 0.182 | GO:0071103 | BP DNA conformation change |
| 1.39e-05 | 1654 | 97 | 45 | 0.464 | 0.027 | GO:0016043 | BP cellular component orga |
| 2.41e-12 | 46 | 97 | 12 | 0.124 | 0.261 | GO:0006323 | BP DNA packaging (2) |
| 1.92e-05 | 211 | 97 | 13 | 0.134 | 0.062 | GO:0034622 | BP cellular macromolecular |
| 2.43e-11 | 55 | 97 | 12 | 0.124 | 0.218 | GO:0065004 | BP protein-DNA complex |
| 3.29e-11 | 609 | 97 | 33 | 0.340 | 0.054 | GO:0007049 | BP cell cycle (1) |
| 3.04e-10 | 514 | 97 | 29 | 0.299 | 0.056 | GO:0022402 | BP cell cycle process |
| 3.31e-08 | 407 | 97 | 23 | 0.237 | 0.057 | GO:0022403 | BP cell cycle phase |
| 5.74e-06 | 317 | 97 | 17 | 0.175 | 0.054 | GO:0000279 | BP M phase (4) |
| 6.87e-06 | 287 | 97 | 16 | 0.165 | 0.056 | GO:0000278 | BP mitotic cell cycle |
| 3.40e-08 | 552 | 97 | 27 | 0.278 | 0.049 | GO:0051276 | BP chromosome organizatio |
| 4.28e-05 | 62 | 97 | 7 | 0.072 | 0.113 | GO:0000819 | BP sister chromatid seg |
| 3.15e-05 | 221 | 97 | 13 | 0.134 | 0.059 | GO:0006325 | BP chromatin organizati |
| 2.66e-11 | 43 | 97 | 11 | 0.113 | 0.256 | GO:0034728 | BP nucleosome organi |
| 1.99e-10 | 51 | 97 | 11 | 0.113 | 0.216 | GO:0006333 | BP chromatin assembly |
| 2.14e-12 | 35 | 97 | 11 | 0.113 | 0.314 | GO:0031497 | BP chromatin assem |
| 6.87e-13 | 32 | 97 | 11 | 0.113 | 0.344 | GO:0006334 | BP nucleosome a |
| 1.33e-05 | 72 | 97 | 8 | 0.082 | 0.111 | GO:0007127 | BP meiosis I (1) |
| 5.94e-06 | 18 | 97 | 5 | 0.052 | 0.278 | GO:0006273 | BP lagging strand elongat |
| 5.38e-05 | 200 | 97 | 12 | 0.124 | 0.060 | GO:0009892 | BP negative regulation of |
| 7.40e-05 | 175 | 97 | 11 | 0.113 | 0.063 | GO:0051172 | BP negative regulation |
| 5.12e-05 | 199 | 97 | 12 | 0.124 | 0.060 | GO:0031324 | BP negative regulation |
| 2.35 | | | | | | | of |
| 7.40 | | | | | | | of |
| 3.14 | | | | | | | s ( |
| 1.28 | | | | | | | ss |
| 2.08 | | | | | | | o st |
| 1.16 | | | | | | | e to |
| 7.63e-11 | 271 | 97 | 22 | 0.227 | 0.081 | GO:0006974 | BP response to DNA da |

g:Profiler -- a web-based toolset for functional profiling of gene lists from large-scale experiments
**Jyri Reimand, et al. (2007)**

# http://biit.cs.ut.ee/gprofiler

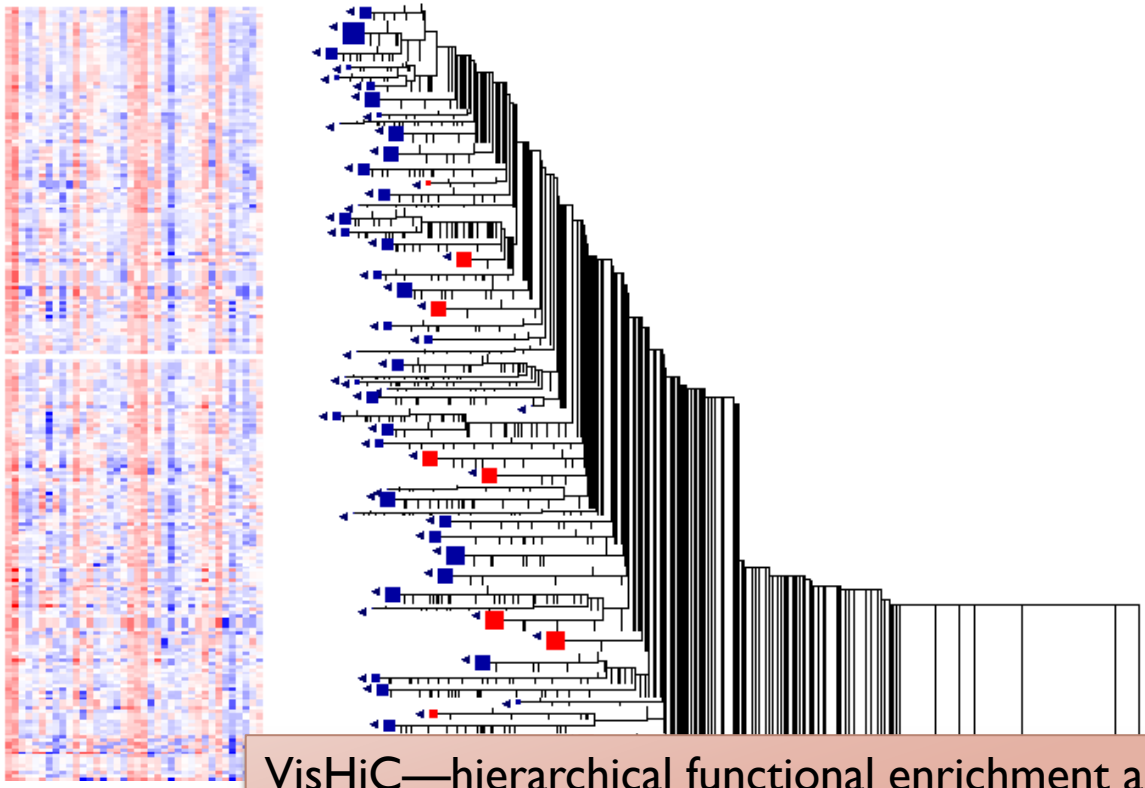


| | | |
|---|---|---|
| KEGG:00240 | ke | Pyrimidine metabolism (1) |
| KEGG:03030 | ke | DNA replication (1) |
| KEGG:03430 | ke | Mismatch repair (1) |
| KEGG:03440 | ke | Homologous recombination (1) |
| KEGG:04111 | ke | Cell cycle - yeast (1) |

g:Profiler -- a web-based toolset for functional profiling of gene lists from large-scale experiments
**Jyri Reimand, et al. (2007)**

http://biit.cs.ut.ee/vishic

- Аннотация помогает отличить «осмысленные» группы от «неинтересных»



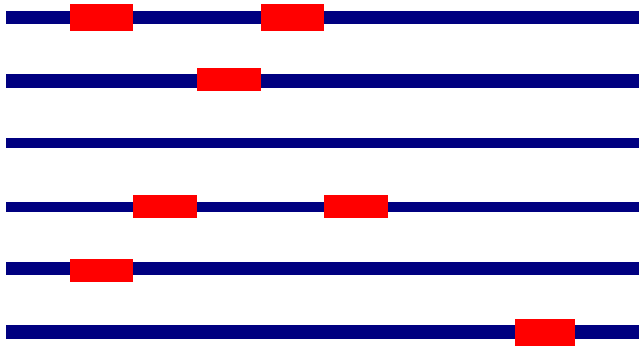
VisHiC—hierarchical functional enrichment analysis of microarray data
**Darya Krushevskaya, et al. (2009)**

# Не только гены и онтологии

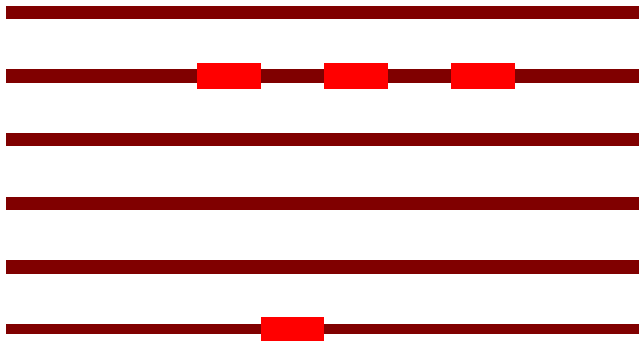
Text mining for automatic annotation of microarray experiment clusters.
**Aleksandr Tkachenko, et al. (2008)**

# Мотивы

**Кластер генов**

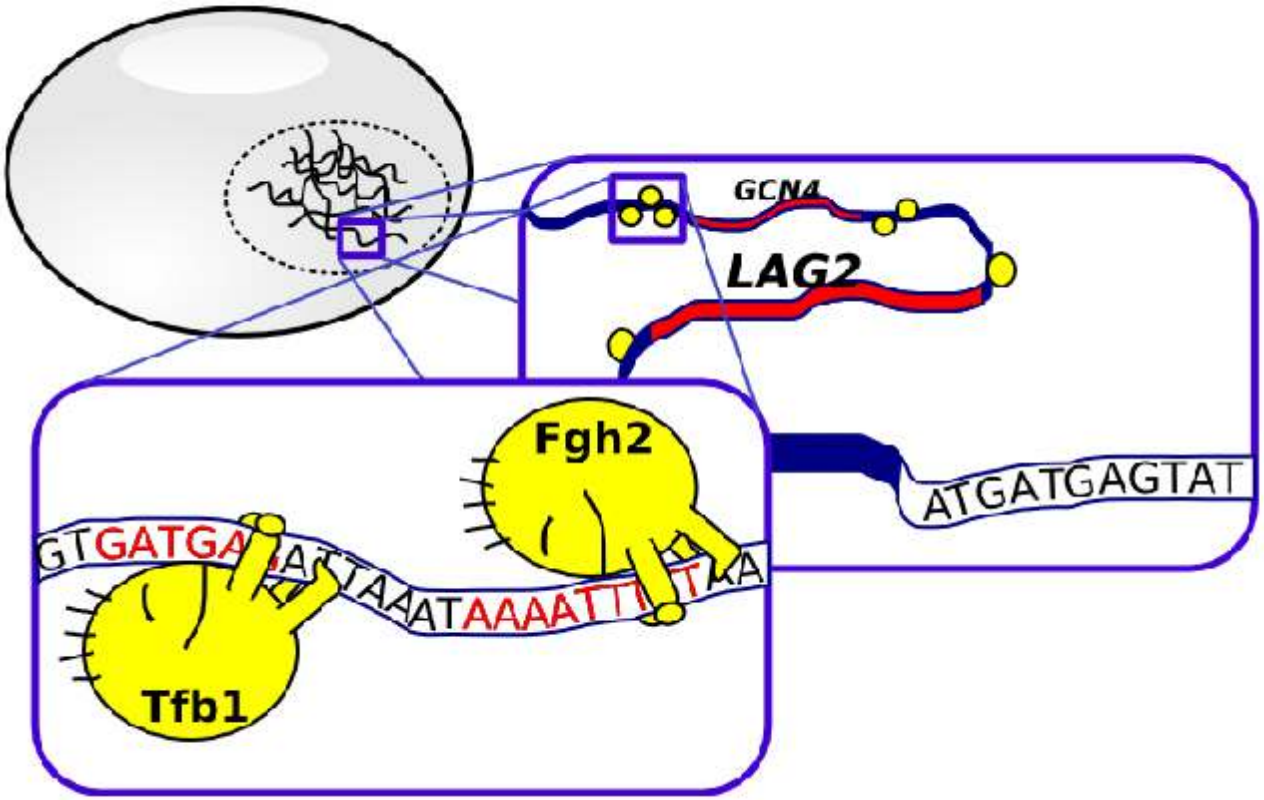


**Другие гены**



о-в. Приветствия
море Демагогии
м.Технологий
респ. Нуклеотидов
земля Смысла
б-о Проблем
а.о. Многомерного анализа
страна Экспрессии

# Мотивы



# Поиск мотивов

- Для каждого возможного мотива
  - Считаем сколько промотеров **всего** содержат его
  - Сколько промотеров **в кластере** его содержат
  - Оцениваем вероятность получить **такой же или лучший** результат **случайно** (P-value)
  - Оставляем мотивы, у которых
    - P-value < ПОПРАВКА(0.01)

# Пример

```
> library(BSgenome.Scerevisiae.UCSC.sacCer1)
```

```
> getPromoter = function(g) {
+     chr = org.Sc.sgdCHR[[g]]
+     if (is.null(chr)) return(NULL)
+     g_start = org.Sc.sgdCHRLOC[[g]]
+     chrdata = Scerevisiae[[sprintf("chr%s", chr)]]
+     p_start = g_start - 499
+     if (g_start < 0)
+         return(complement(chrdata[abs(p_start):abs(g_start
+     else
+         return(chrdata[p_start:g_start])
+ }
```

# Пример

```
> ps = apply(as.matrix(gene_names),1,getPromoter)


> random_genes = sample(rownames(expdata), 500)
> ns = apply(as.matrix(random genes),1,getPromoter)



> for (i in 1:length(ps))
+     for (j in (1:(length(ps[[i]])-7)))
+         strs = c(strs, toString(ps[[i]][j:(j+7)]))

> strs = unique(strs)
> length(strs)
[1] 28843

> pd = PDict(strs, tb.start=3, tb.width=3)
```

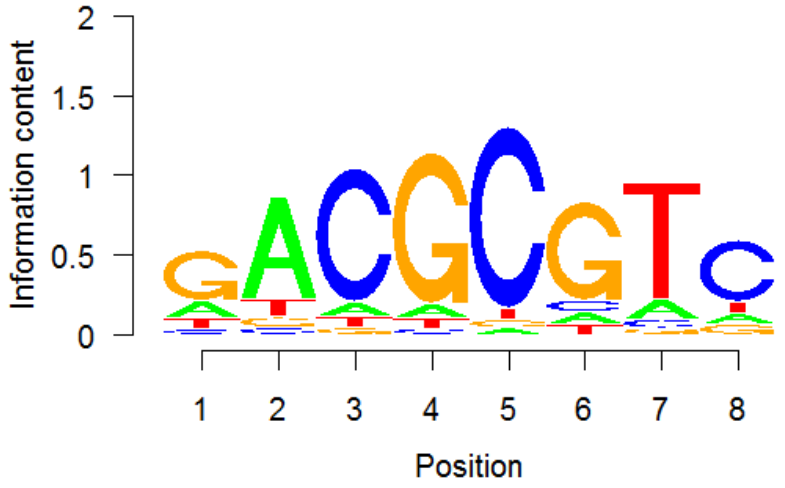
# Пример

```
> countPos = rep(0, length(strs)) # Нули
> for (i in 1:length(ps))
+     countPos = countPos +
+         (countPDict(pd, ps[[i]], max.mismatch=2) > 0)

> countNeg = rep(0, length(strs))
> for (i in 1:length(ns))
+     countNeg = countNeg +
+         (countPDict(pd, ns[[i]], max.mismatch=2) > 0)

> pvalues = c()
> for (i in 1:length(strs))
+     pvalues[i] = 1-phyper(countPos[i]-1, length(ps), length(ns

> pv = p.adjust(pvalues, "fdr")
> motifs = which(pv < 0.001)
> strs[motifs]
[1] "GACGCGTC" "TGCGCGTT" "GACGCGTT"
```

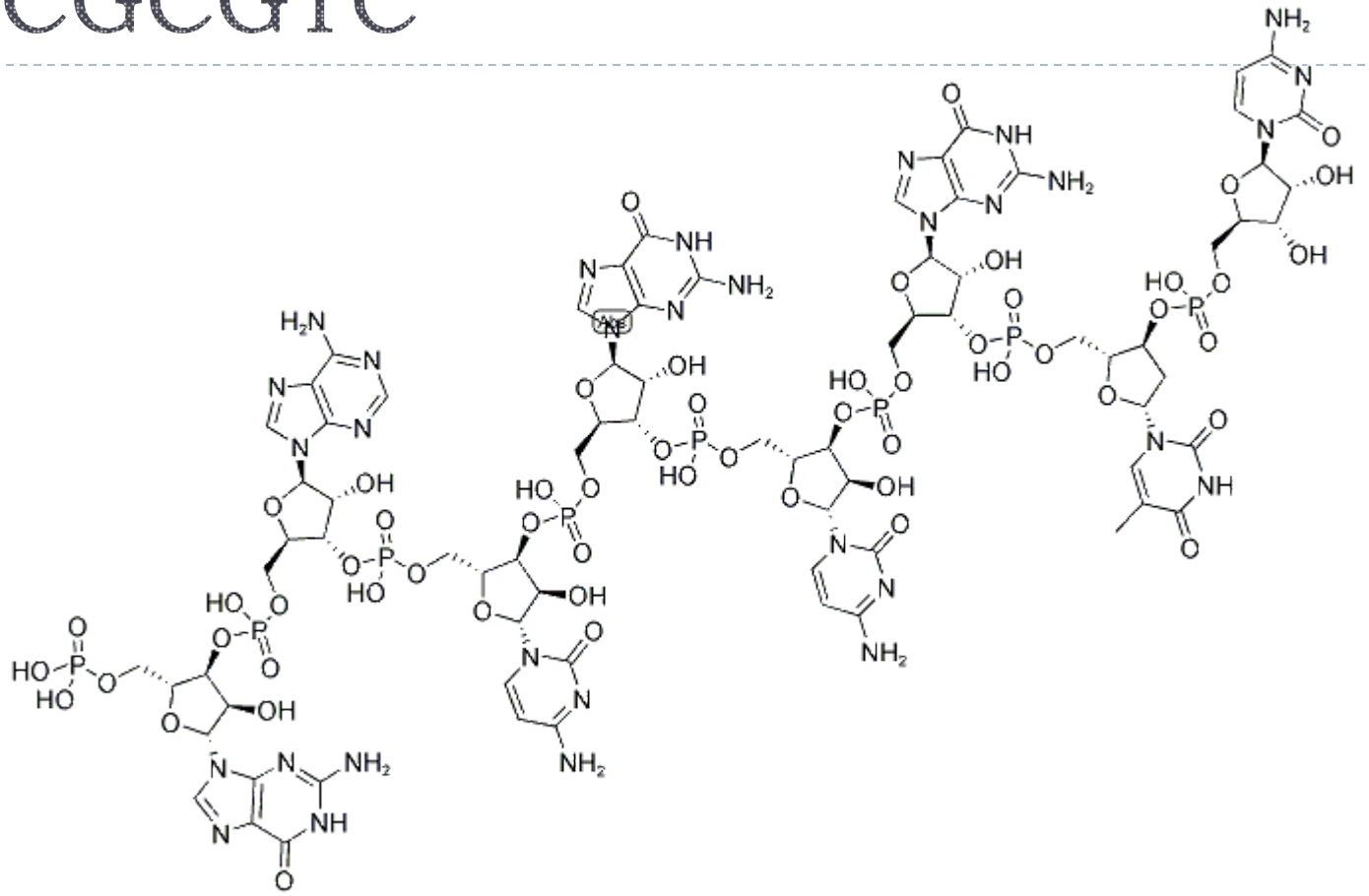
# Пример

```
> m = strs[motifs[1]]


> all_matches = as.character(
+     matchPattern(m, ps[[1]], max.mismatch=2))
> for (i in 2:length(ps))
+     all_matches = c(all_matches,
+         as.character(matchPattern(m, ps[[i]], max.mismat
> cm = consensusMatrix(all_matches, as.prob=TRUE)[1:4,]
```

```
> library(seqLogo)
> seqLogo(cm)
```



# GACGCGTC

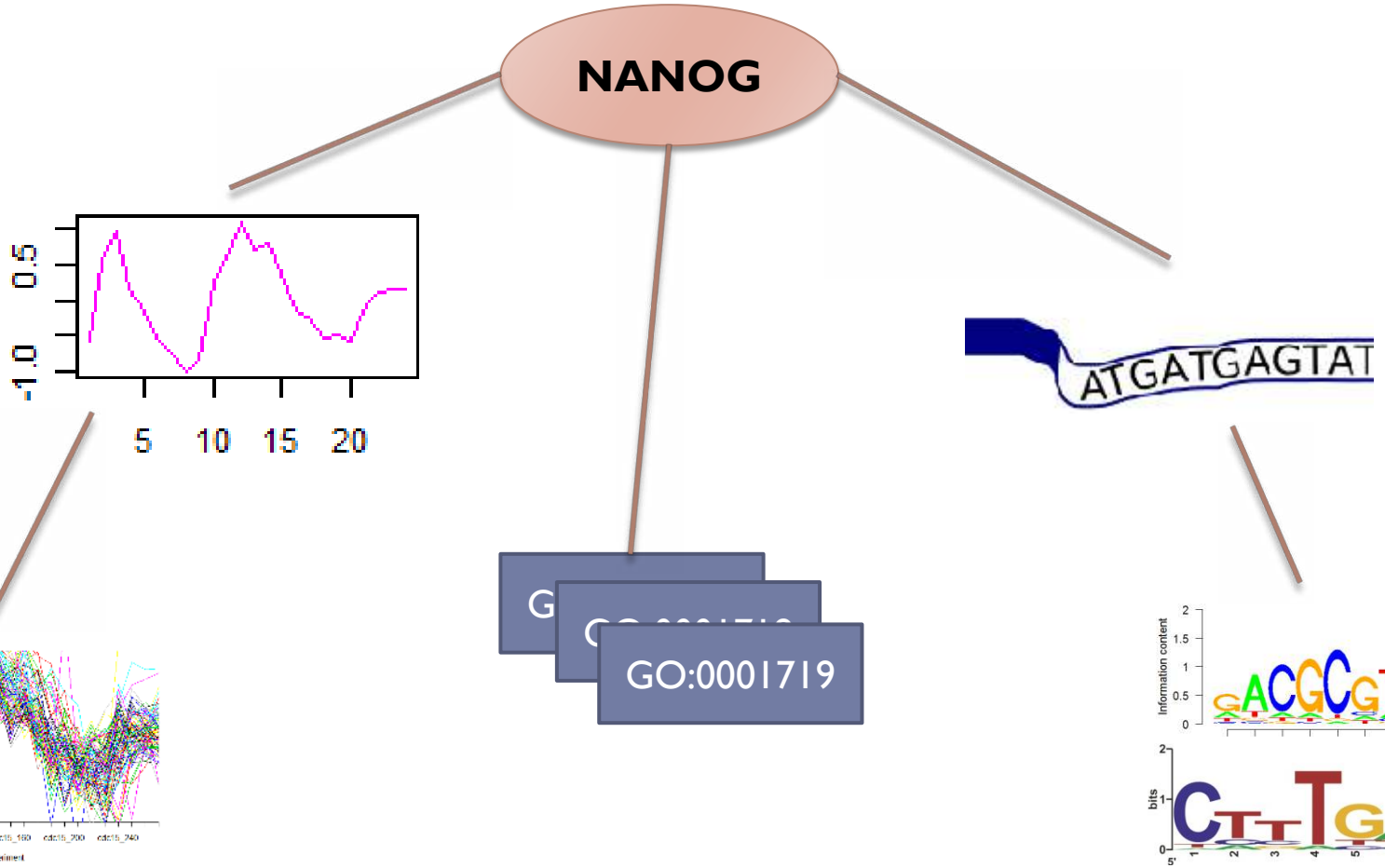


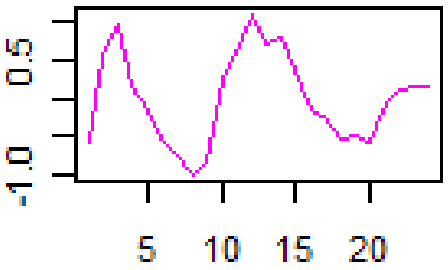
GACGCGTC, 5'-PHOSPHORYLATED Suppliers  Global( 1)Suppliers  USA 1

| Supplier | Tel | Fax | Email |
|---|---|---|---|
| The Midland Certified Reagent Company, Inc. | 800 247 8766 | 432 694 2387 | orders@oligos.com |

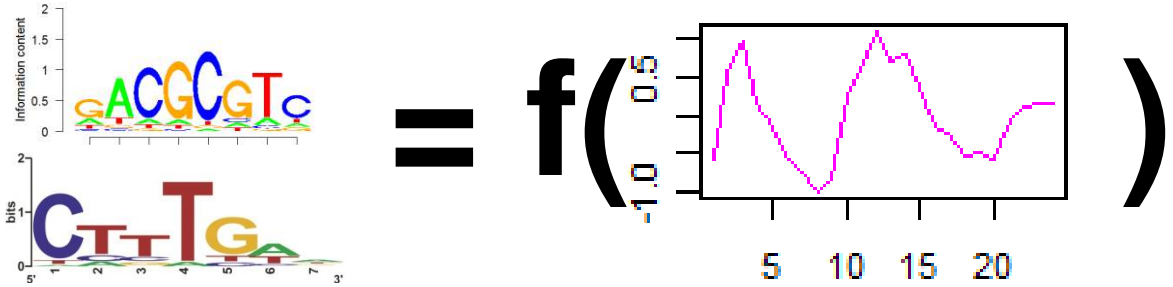
о-в. Приветствия
море Демагогии
м.Технологий
респ. Нуклеотидов
земля Смысла
б-о Проблем
а.о. Многомерного анализа
страна Экспрессии

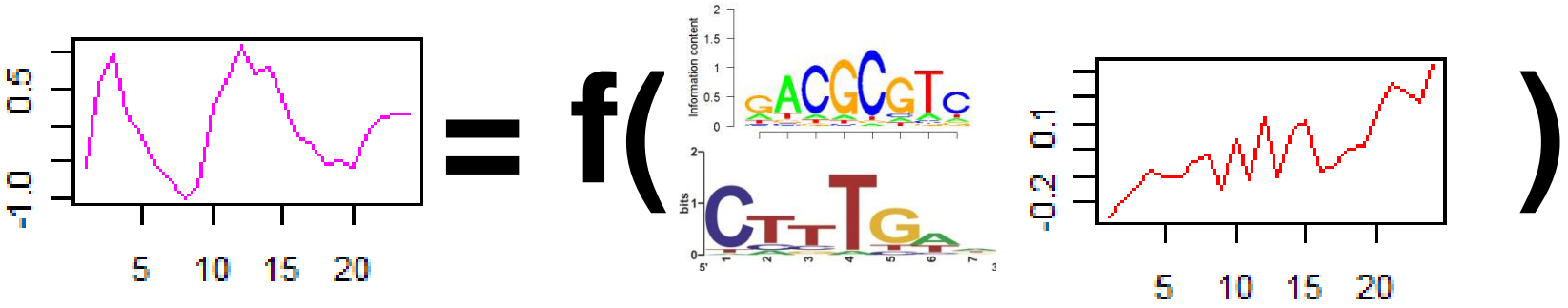
# Посмотрим под другим углом






= f( ATGATGAGTAT )

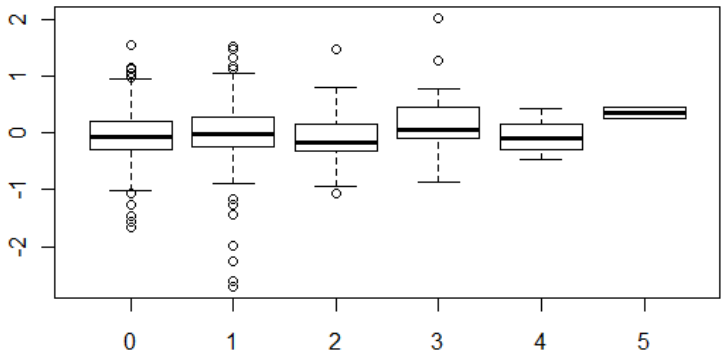

Information content
2
1.5
1
0.5
0
bits
2
1
0
5'
3'
=
f(
0.5
-1.0
5
10
15
20
)


-1.0 0.5
5 10 15 20
=
f(
Information content
bits
-0.2 0.1
5 10 15 20
)

# Пример

```
> y = expdata[names(ns),1]
> x1 = sapply(ns, function(s) { countPattern(m, s, max.mis
> x2 = sapply(ns, function(s) { countPattern("GATTACAG", s
```

```
> boxplot(y~x1)
```




```
> lm(y ~ x1 + x2)

(Intercept)           x1           x2
   -0.02466      0.03667     -0.01961
```

# Пример

- Предположим что экспрессия гена g, Expr[g], в выбранном нами эксперименте выражается как линейная комбинация присутствующих в промотере гена мотивов:

$$\mathrm{Expr[g]} = \alpha_1 m_{g1} + \alpha_2 m_{g2} + \ldots + \alpha_k m_{gk}$$

$$m_{gi} = \mathrm{HasMatch}(\mathrm{Motif}_i, \mathrm{Promoter}[g])$$

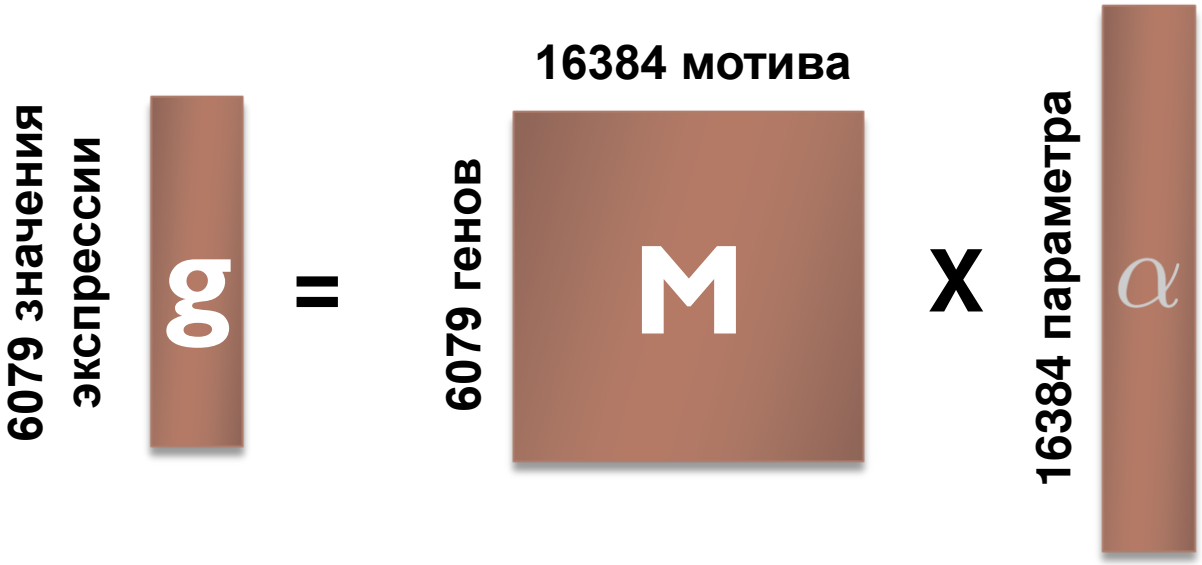
- Т.к. мы не знаем какие мотивы на самом деле важны, возьмем *все*, и по подобранным коэффициентам $\alpha_i$ найдем значимые.

# Пример

- Возьмем здесь для простоты в качестве мотивов-кандидатов все последовательности длиной 7. Для каждой такой последовательности проверим ее наличие в промотере: получим матрицу M из нулей и единиц:

$$\mathrm{Expr}[\mathrm{g}] = \alpha_1 m_{g1} + \alpha_2 m_{g2} + \ldots + \alpha_k m_{gk}$$

**В матричной форме (для всех генов сразу) модель записывается как**



# Пример (весь код см. на сайте)

```
> M[1:5,1:5]
        AAAAAAA CAAAAAA TAAAAAA GAAAAAA ACAAAAA
YAL001C       0       0       0       0       1
YAL002W       1       1       1       1       1
YAL003W       1       1       1       1       1
YAL004W       0       1       0       0       1
YAL005C       0       1       0       1       0
```

```
> as.matrix(g[1:5], ncol=1)
                [,1]
YAL001C -0.1600000
YAL002W -0.2150000
YAL003W -0.3700000
YAL004W -0.4433333
YAL005C -0.4300000
```

- Задача сводится к оценке значений 16384 параметров линейной модели.
- Т.к. генов всего 6078, задача плохо определена и обычная линейная регрессия не поможет.
- Однако существуют методы подбора параметров (attribute selection) и регуляризации, которые справятся.
  - LARS, Ridge Regression, Boosting, Greedy search, etc…

- Самое простое – найти корреляцию каждого мотива с экспрессией. Можно предположить, что если наличие мотива коррелирует со значением экспрессии сильнее других – данный мотив важен.

```
> motifCorrs = rep(0,ncol(M))
> for (i in 1:ncol(M)) motifCorrs[i] = abs(cor(M[,i], g))

> sort(motifCorrs, decreasing=TRUE)[1:5]
   AGATGAG    GATGAGG    TGATGAG    GGAGATG    GCCAATC
0.06426586 0.06394384 0.06220971 0.06072638 0.06054394
```

Regulatory element detection using correlation with expression.
**Harmen Bussemaker, et al. (2001)**

# High-resolution DNA binding specificity analysis of yeast transcription factors

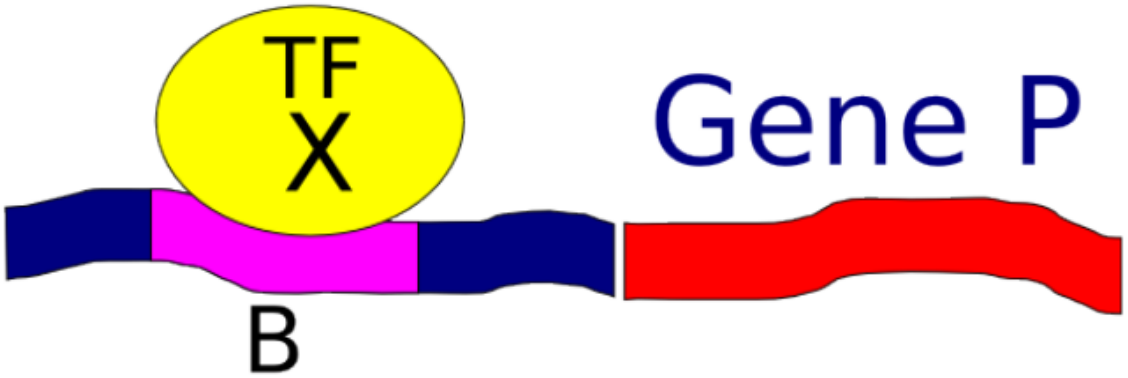
Cong Zhu[1], Kelsey Byers[1], Rachel McCord[2], Zhenwei Shi[3],

...

experimentally determined DNA binding specificities. Among other novel regulators, we discovered proteins that bind the PAC (Polymerase A and C) motif (GATGAG) and regulate rRNA transcription and processing, core cellular

- Обратите внимание, что сейчас мы рассматривали лишь корреляцию с экспрессией в первом микрочипе.
- Что можно сделать, имея данные нескольких микрочипов?

# Существует много вариаций этой идеи

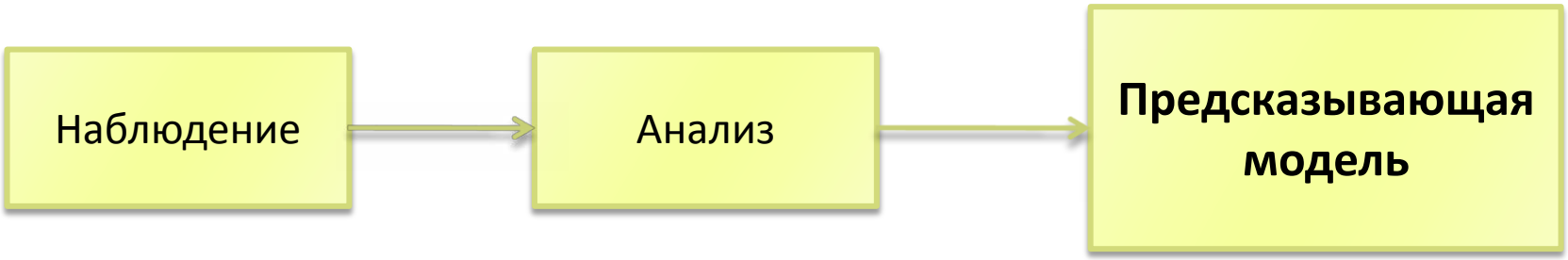



$$g_P \approx \alpha_P m_B t_X$$

G=MAT: Linking Transcription Factor Expression with DNA Binding
**Konstantin Tretyakov, et al. (2008)**

# Заключение

- Наука (в идеальном мире)



о-в. Приветствия
море Демагогии
м.Технологий
респ. Нуклеотидов
земля Смысла
б-о Проблем
а.о. Многомерного анализа
страна Экспрессии

# Спасибо за внимание!

**Bioinformatics, Algorithmics and Data Mining Group**

# Полезное чтение

- Биоинформатика
  - Введение: **Introduction to Computational Genomics: A Case Studies Approach.** Nello Cristianini and Matthew W. Hahn.
  - Продолжение: много хороших книг, см например http://www.bioinformatics.org/wiki/Recommended_books#Computational.2FMathematical_aspects
- Data Mining
  - Введение в : **Data Mining: Practical Machine Learning Tools and Techniques.** Ian H. Witten, Eibe Frank
  - Продолжение: сложно посоветовать, вот очень правильный список: http://www.thearling.com/books.htm#mining
  - См также: **Kernel Methods for Pattern Analysis**. John Shawe-Taylor, Nello Cristianini.
- Классическая статистика
  - Введение: **Introductory Statistics with R.** Peter Dalgaard
  - Продолжение: **Statistical Inference.** George Casella, Roger L. Berger.
  - Линейные модели: **An Introduction to Generalized Linear Models.** Anette J. Dobson
  - Множественное тестирование:
    **Multiple Testing Procedures with Applications to Genomics**. Sandrine Duduit, Mark van der Laan.
  - Обратите внимание, что «гипергеометрический» тест в статистике называется «тест Фишера» (Fisher' Exact Test)

# Полезное чтение

- Обработка и анализ данных экспрессии
  - **Microarray Data Analysis: Methods and Applications.** Michael J. Korenberg (editor)
- Поиск мотивов
  - **Motif Discovery on Promotor Sequences.** Maximilian Haeussler. ftp://ftp.inria.fr/INRIA/publication/publi-pdf/RR/RR-5714.pdf
  - Pattern Discovery from Biosequences. Jaak Vilo http://ethesis.helsinki.fi/julkaisut/mat/tieto/vk/vilo/
- R & Bioconductor
  - **Bioconductor Case Studies.** F. Hahne, W. Huber.
  - Вообще про R много книг: http://www.amazon.com/s/ref=nb_sb_noss?url=search-alias%3Dstripbooks&field-keywords=statistics+R

# Image references

The following images by various authors are used in this presentation:

- Slide 1: http://en.wikipedia.org/wiki/File:The_Scream.jpg
- Slide 2:
  - http://maps.google.com/
  - http://www.smarttravel.ee/images/tartu.jpg
- Slide 3:
  - http://www.histrodamus.ee/upload/files/0.jpg
  - http://www.bbsed2007.ttu.ee/public/UT.jpg
  - http://www.flickr.com/photos/40397630@N00/558145359/ © Triin Noorkõiv
- Slide 4: Photo © Jaak Vilo
- Slide 5: http://www.stacc.ee/ © STACC
- Slide 7:
  - http://brendanbody.blogspot.com/2010/04/weight-problem.html
  - http://en.wikipedia.org/wiki/File:Justus_Sustermans_-_Portrait_of_Galileo_Galilei,_1636.jpg
- Slide 8:
  - http://en.wikipedia.org/wiki/File:Tycho_Brahe.JPG
  - http://en.wikipedia.org/wiki/File:Johannes_Kepler_1610.jpg
  - http://www.astro.psu.edu/users/stark/outreach/Kepler/
- Slide 9: http://www.foundalis.com/phi/WhyTimeFlows.htm
- Slide 10: Derived from http://demotivators.ru/posters/72755/smotri-na-mir-prosche.htm

# Image references

- Slide 11:
  - http://milogiya.narod.ru/mir-antimir.htm
  - http://ekosait.21429s01.edusite.ru/p29aa1.html
  - http://www.scientificpsychic.com/health/vitamins.html
- Slide 12: http://www.ubcbotanicalgarden.org/potd/2006/02/brassica_oleracea_botrytis_group_romanesco.php © Kimberly T
- Slide 13: http://www.moonbattery.com/archives/2009/12/go-crystal-ball.html
- Slide 14: http://www.computertim.net/HappyComputer.gif
- Slide 15:
  - http://vintagedancers.org/newport/n_07rg_bro.html
  - http://www.pulsar.com/
- Slide 16: Photo © D. Bertola, CERN
- Slide 17: http://www.ars.usda.gov/is/graphics/photos/dec97/k7807-1.htm © Ken Hackman
- Slide 23: Derived from
  - http://www.biotaq.com/Affymetrix/index.htm
  - http://www.iconarchive.com/show/forum-faces-icons-by-afterglow/Statistician-icon.html © Cian Walsh
  - http://www.fotosearch.com/UNC212/u10606813/ © FotoSearch.com
  - http://www.iconarchive.com/show/nature-icons-by-fasticon/Mountain-icon.html © FastIcon.com
  - http://www.iconarchive.com/show/transport-icons-by-aha-soft/sailing-ship-icon.html © Aha-Soft

# Image references

- Slide 28:
  - http://www.greenspine.ca/Images © Paul De Konick
  - http://en.wikipedia.org/wiki/File:Yellow_adipose_tissue_in_paraffin_section_-_lipids_washed_out.jpg © Department of Histology, Jagiellonian University Medical College
  - http://www.getfreeimage.com/image/81/blood-cells-in-vein-erythrocyte-thrombocyte-leukocyte © GetFreeImage.com
- Slide 29: Derived from http://www.phschool.com/science/biology_place/biocoach/transcription/overview.html
- Slide 33: Source unknown
- Slide 35: Derived from http://www.scq.ubc.ca/spot-your-genes-an-overview-of-the-microarray/ © Jiang Long
- Slide 38: http://www.machinelearning.ru/wiki/index.php?title=%D0%98%D0%B7%D0%BE%D0%B1%D1%80%D0%B0%D0%B6%D0%B5%D0%BD%D0%B8%D0%B5:Chip.PNG
- Slide 45: http://www.dnavision.com/resequencing-with-affymetrix.php
- Slide 57:
  - http://www.ve-group.ru/products40_102.html
  - http://www.globalconstructionwatch.com/?p=158
  - http://familion.ru/zoos/126-zoopark-skazka.html
  - http://www.ay-tour.ru/moscow-region/Foresta-Festival-Park
- Slide 58: Photo © Steve Cole

All other illustrations are © Konstantin Tretyakov. Free reuse allowed under CC/Attribution+ShareAlike.