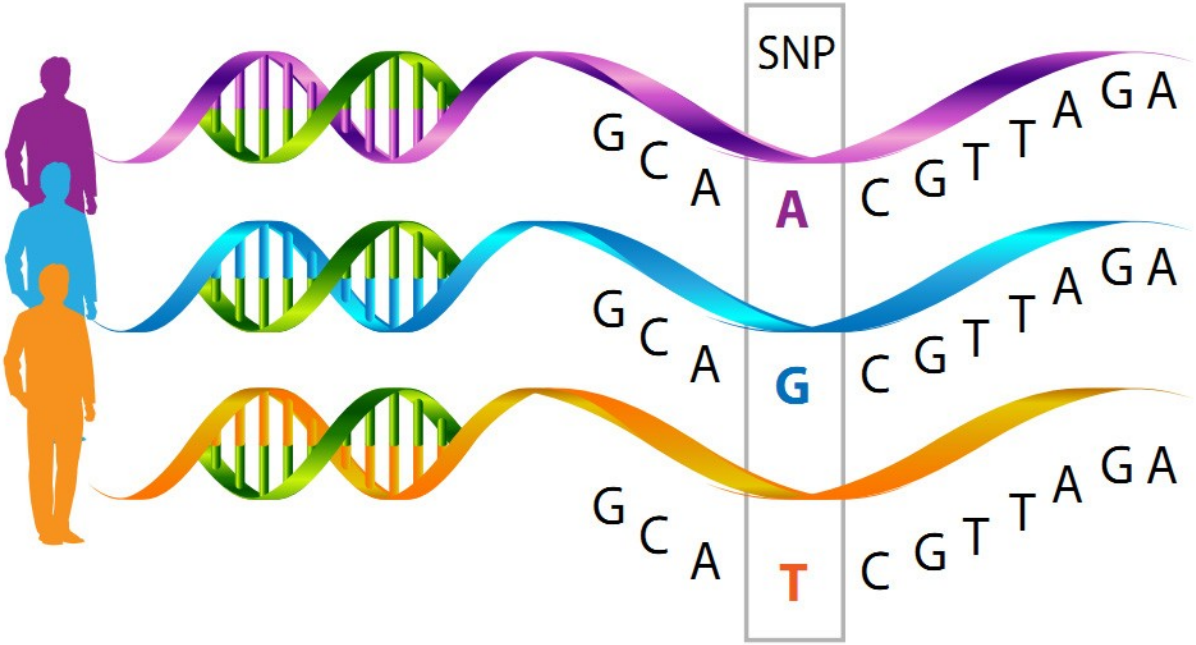
Фонд интеллект

Машинное обучение в иммунологии и медицине

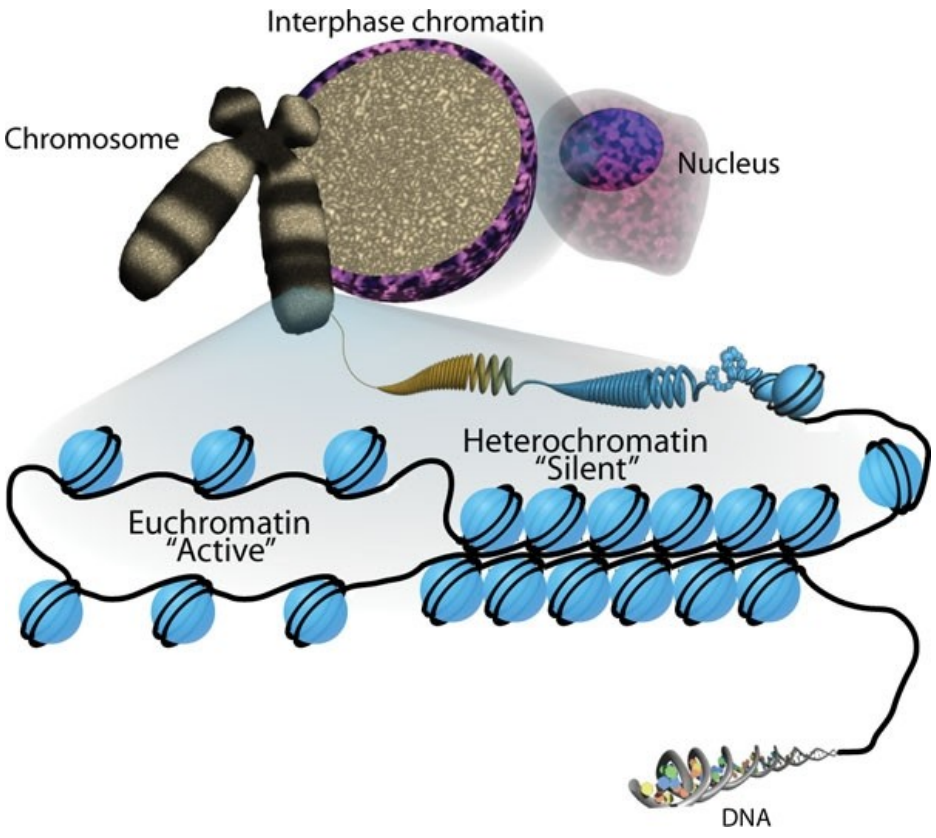
Получение данных в молекулярной биологии. Система iScan для поиска SNP. Чипы для измерения уровней экспрессии генов. Хаускипинг гены. Нормировка.

SNPs - Single nucleotide polymorphisms



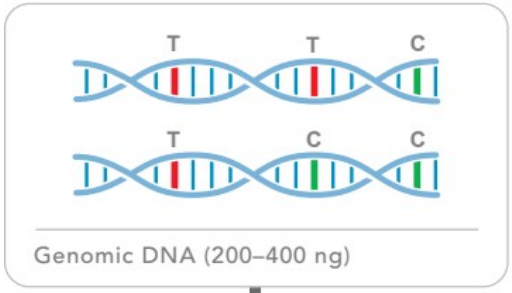
SNPs (single nucleotide polymorphism) - однонуклеотидные отличия последовательности ДНК в геноме представителей одного между гомологичными участками гомологичных хромосом.

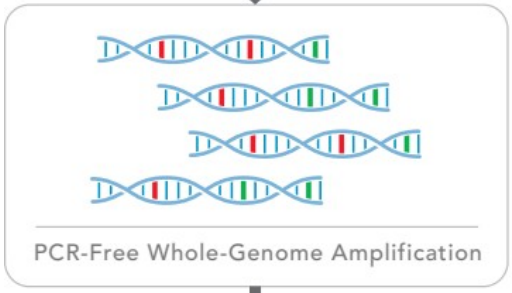
Уровни экспрессии генов

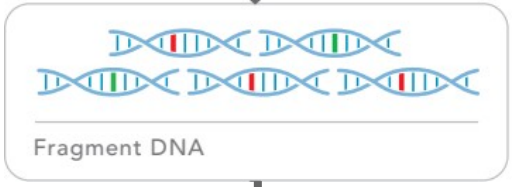


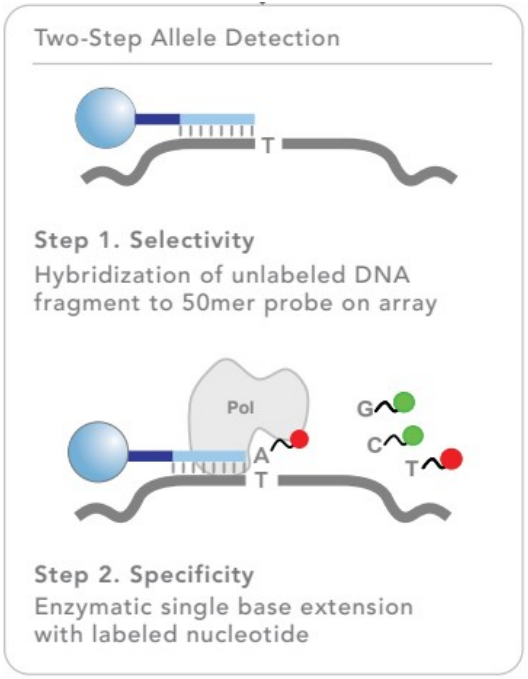
Степень активности гена (количество производимого белка) регулируется преимущественно на уровне транскрипции. Задача определения активности гена сводится к подсчету количества мРНК, соответствующей исследуемому гену.

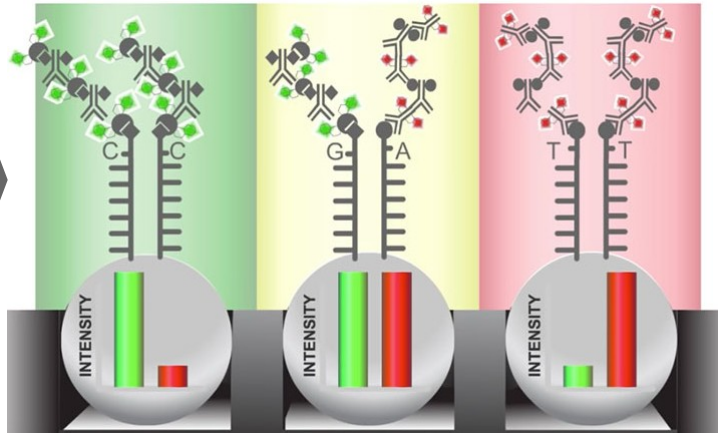
Определение SNP iScan Illumina

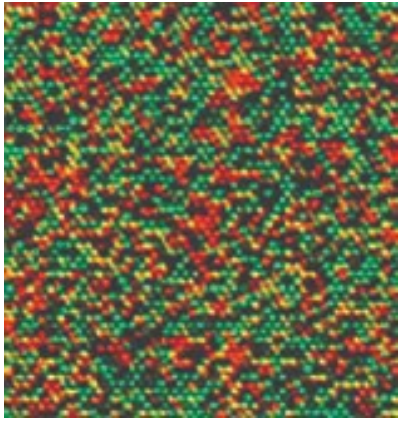












Определение SNP iScan Illumina

Определение SNP iScan Illumina

Feature	Description	
Species	Human	
Total Number of Markers	642,824	
Capacity for Custom Bead Types	50,000	
Number of Samples per BeadChip	24	
DNA Input Requirement	200 ng	
Assay Chemistry	Infinium HTS	
Instrument Support	iScan or HiScan® System	
Sample Throughput[a]	~ 2304 samples/week	
Scan Time per Sample	iScan System 2.5 min	HiScan System 2.0 min

a. Estimate assumes 1 iScan System, 1 AutoLoader, 2 Tecan robots, and a 5-day work week.

Marker Categories			No. of Markers
Exonic Markers[a]			66,199
Intronic Markers[a]			256,673
Nonsense Markers[b]			3232
Missense Markers[b]			43,342
Synonymous Markers[b]			5109
Mitochondrial Markers[c]			137
Indels[c]			3836
Sex Chromosomes[c]	X 16,927	Y 1456	PAR/Homologous 576

Content	No. of Markers	Research Application/Note
ADME Core and Extended Genes	5816	Drug metabolism and excretion
ADME Core and Extended Genes +/- 10 kb	7246	Drug metabolism and excretion (plus regulatory regions)
APOE	17	Cardiovascular disease, Alzheimer's disease, immunoregulation, and cognition
Blood Phenotype Genes	1984	Blood phenotypes
COSMIC Genes	276,149	Somatic mutations in cancer
GO CVS Genes	82,984	Cardiovascular conditions
Database of Genomic Variants	494,268	Genomic structural variation
eQTLs	2680	Genomic loci regulating mRNA expression levels
Fingerprint SNPs	385	Human identification
HLA Genes	439	Disease defense, transplant rejection, and autoimmune disorders
Extended MHC[a]	8608	Disease defense, transplant rejection, and autoimmune disorders
KIR Genes	27	Autoimmune disorders and disease defense
Neanderthal SNPs	765	Neanderthal ancestry and human population migration
NHGRI GWAS Catalog	6988	Markers from published genome-wide association studies
RefSeq 3' UTRs	10,808	3' untranslated regions of known genes
RefSeq 5' UTRs	5268	5' untranslated regions of known genes
RefSeq All UTRs	15,614	All untranslated regions of known genes
RefSeq	310,926	All known genes
RefSeq +/- 10 kb	367,210	All known genes plus regulatory regions
RefSeq Promoters	13,567	2 kb upstream of all known genes to include promoter regions
RefSeq Splice Regions	1714	Variants at splice sites in all known genes

Анализ данных SNP чипов

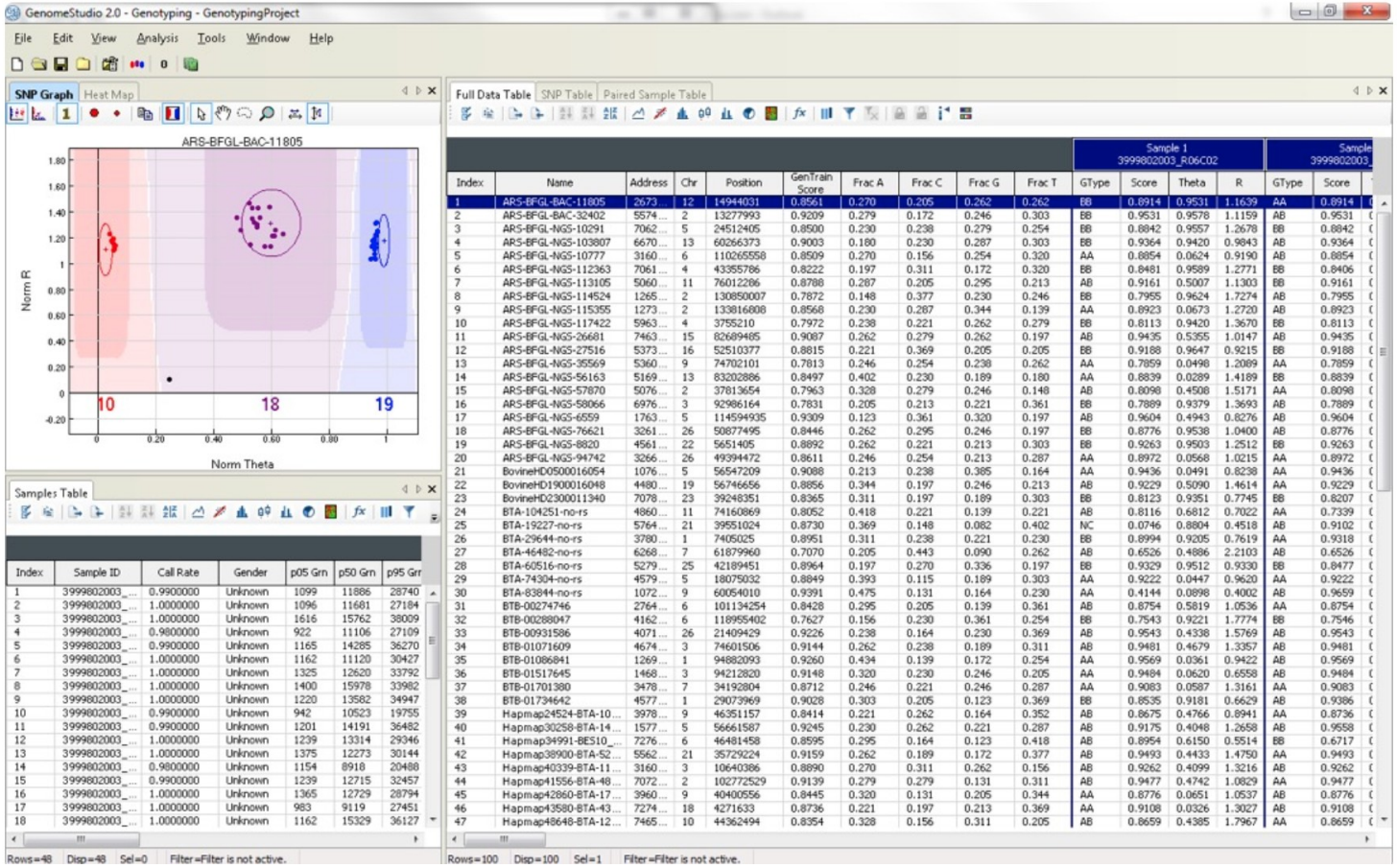
iScan

↓

IDAT file

↓

GenomeStudio



Задачи, решаемые с помощью SNP чипов

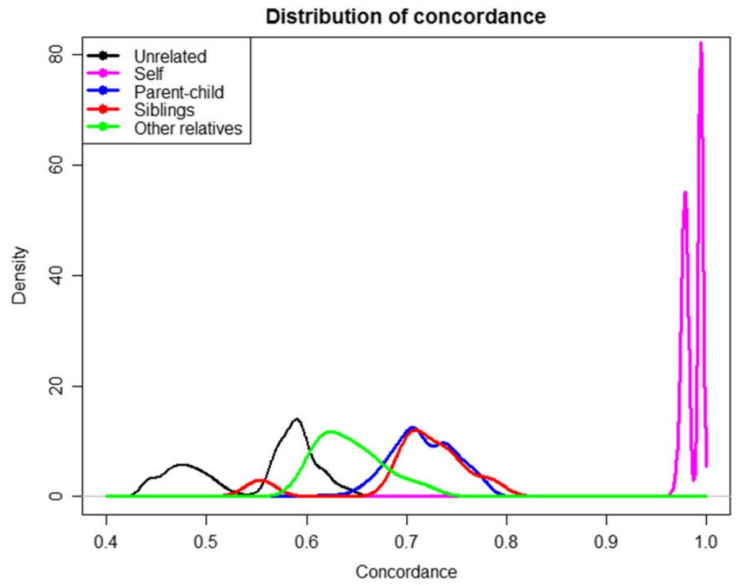


Fig. 1 Concordance for same samples between Infinium QC and 1000 Genomes data (*purple*), and for different unrelated samples (*black*), between parent and child, siblings, and other relatives

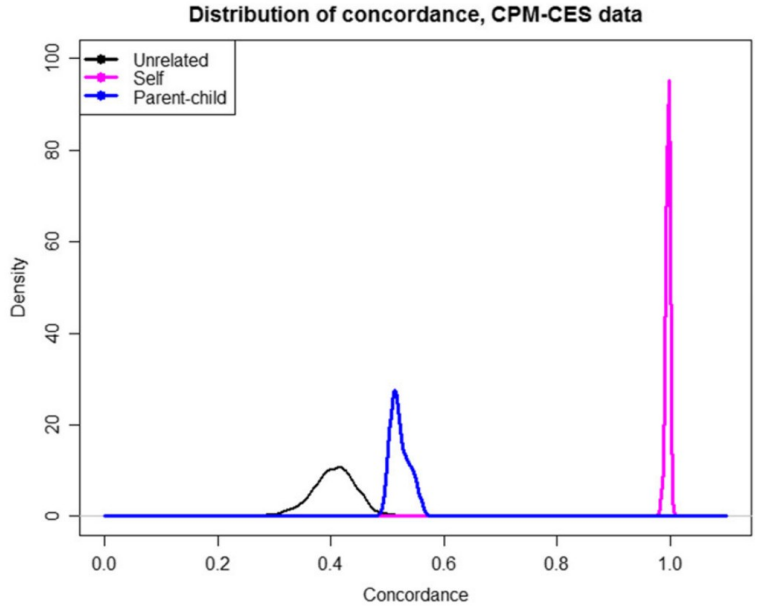


Fig. 2 Concordance histogram for all possible pairs of samples from iScan and CES experiments

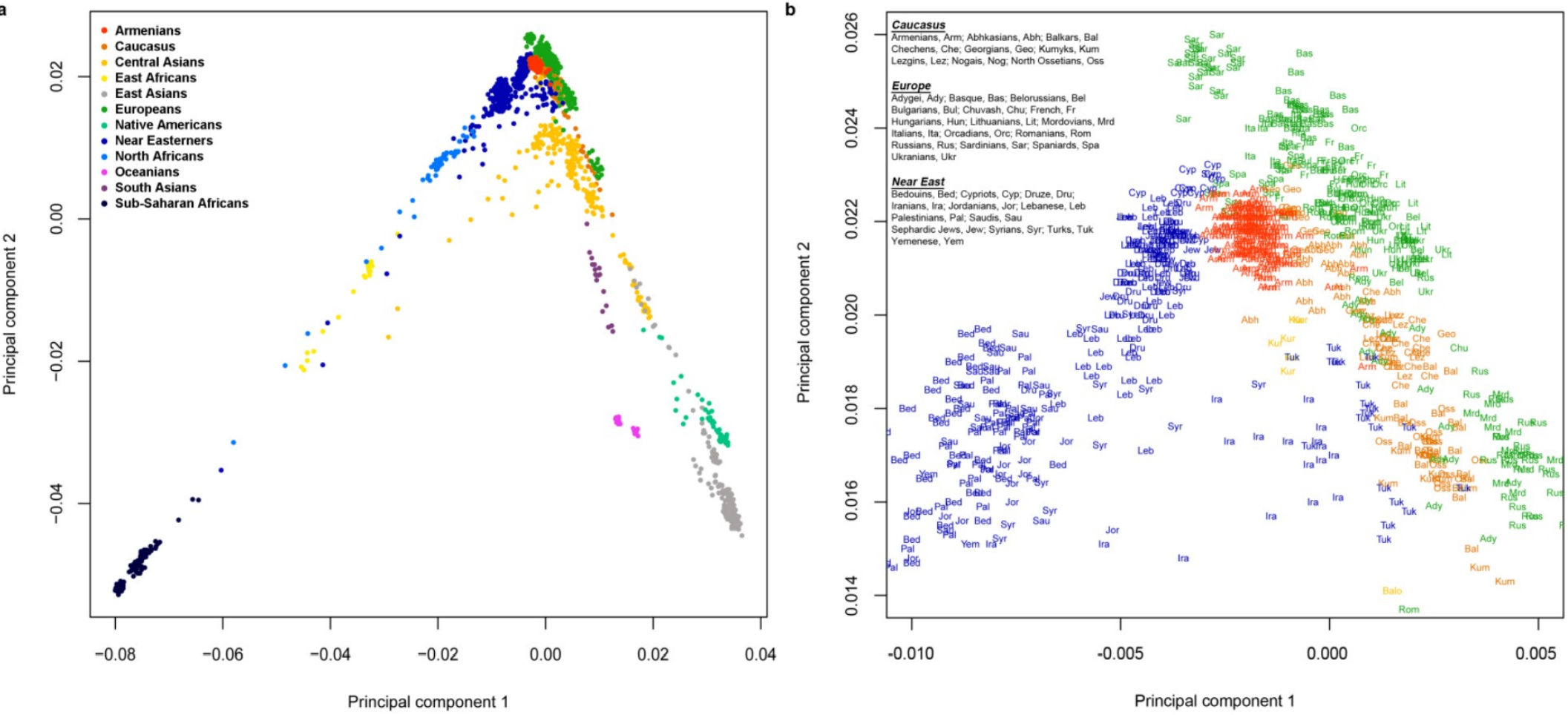
Источник: Clinical utility of the low-density Infinium QC genotyping Array in a genomics-based diagnostics laboratory; October 2017BMC Medical Genomics 10(1); DOI:10.1186/s12920-017-0297-7

Задачи, решаемые с помощью SNP чипов

- Выясните ваши риски заболеваний
- Проверьте персональную непереносимость продуктов
- Узнайте о своем происхождении
- NEW Найдите родственников
- Оцените вероятность спортивных травм
- Изучите особенности организма
- Бесплатно проконсультируйтесь с генетиком

Источник: https://atlas.ru/dna

Этнические различия SNP



Источник: Genetic evidence for an origin of the Armenians from Bronze Age mixing of multiple populations; February 2015 DOI:10.1101/015396

Уровень экспрессии генов

THE HUMAN PROTEIN ATLAS

MENU HELP NEWS

ACE2 Search Fields »

Search result (18 genes): **ACE2** | BACE2 | RCE1 ...

TISSUE

PRIMARY DATA

TISSUES

GENE/PROTEIN

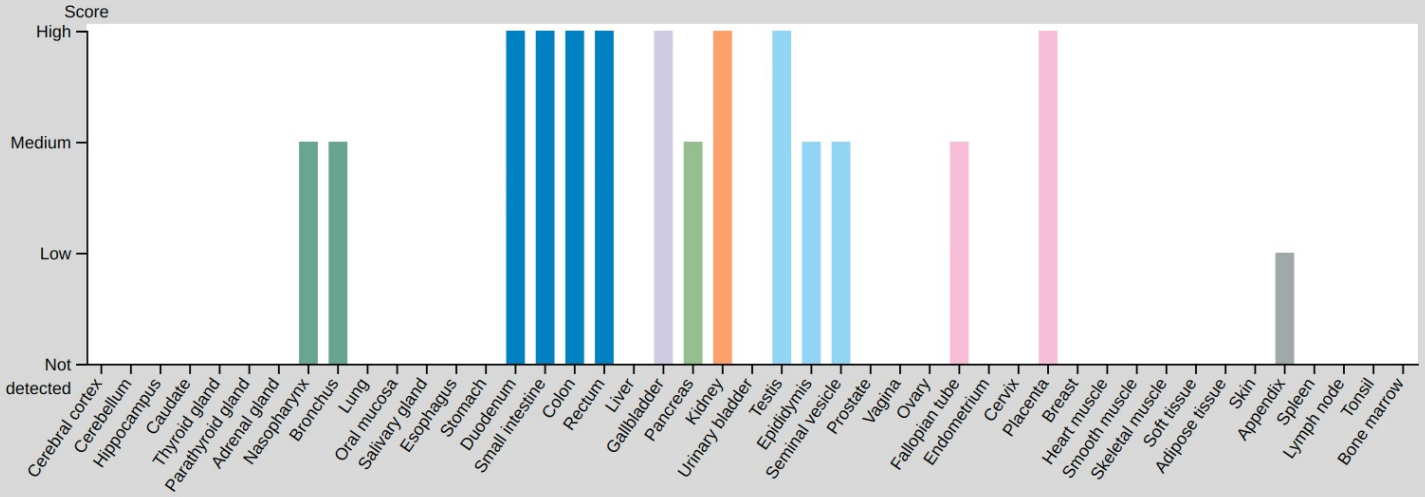
ANTIBODIES AND VALIDATION

Dictionary

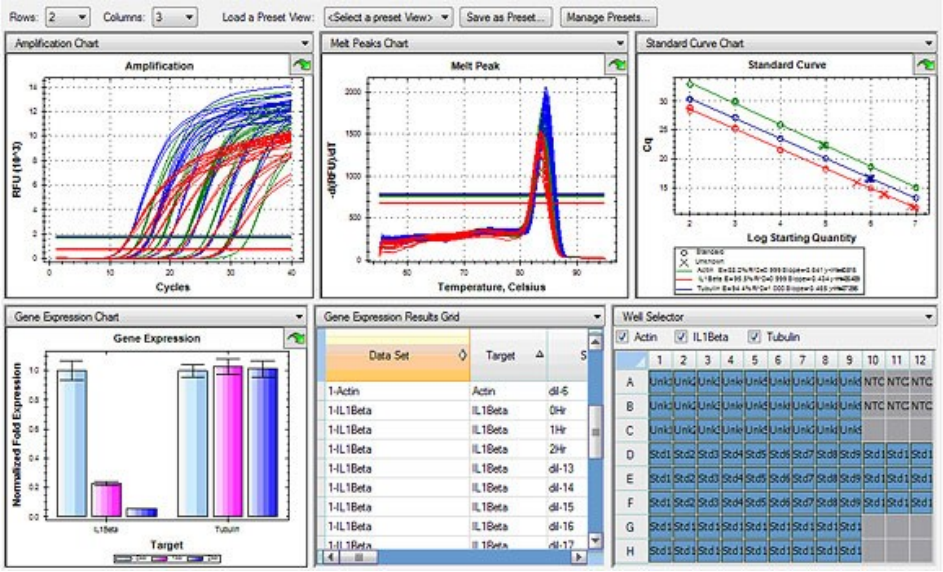
Tissue proteome

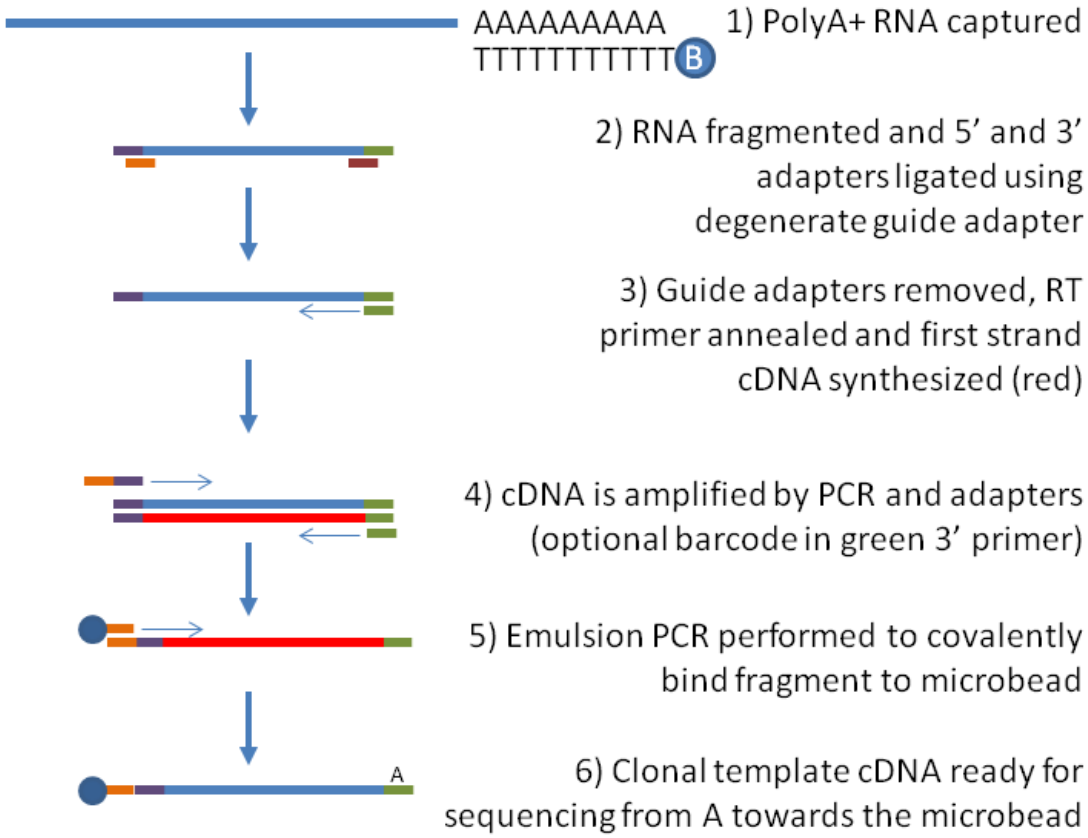
PROTEIN EXPRESSION OVERVIEW

Organ | Expression | Alphabetical

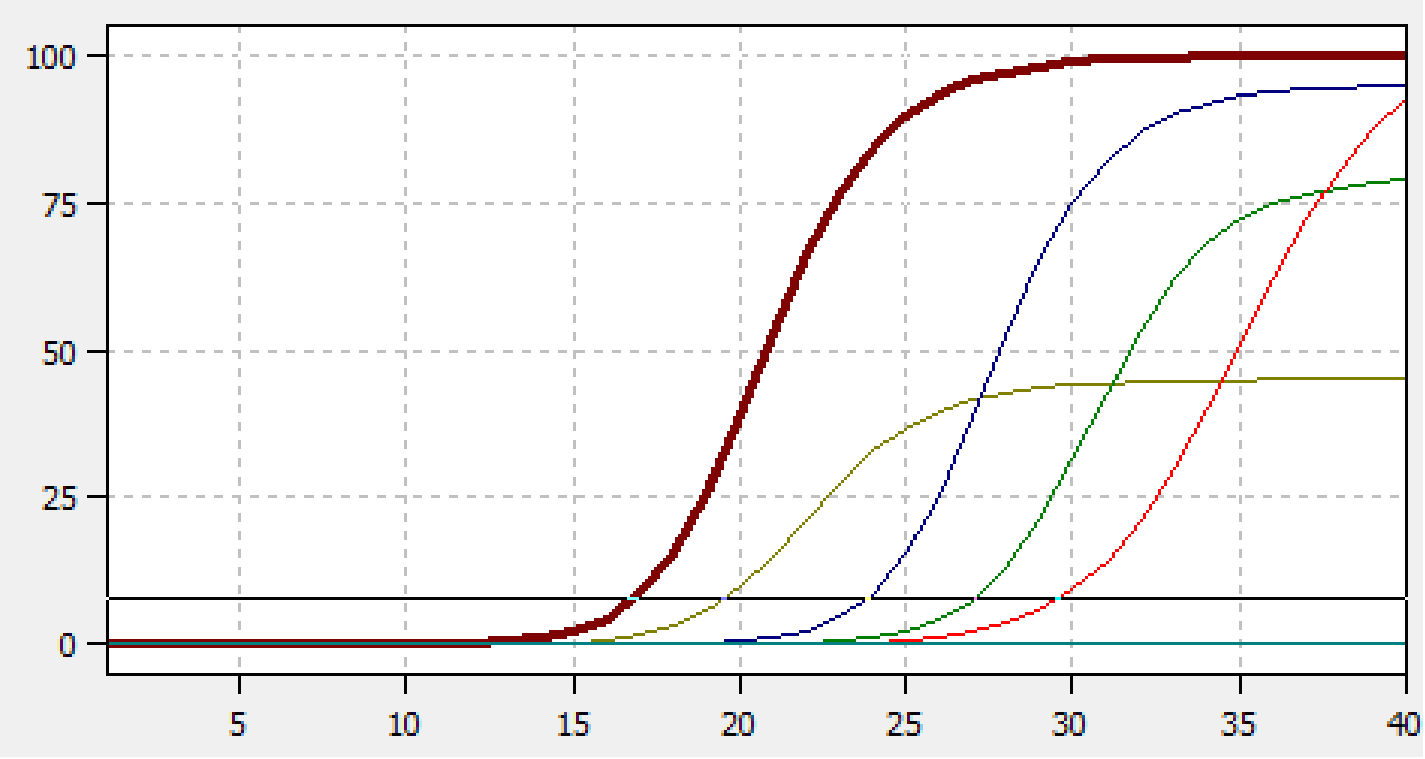


Уровень экспрессии генов

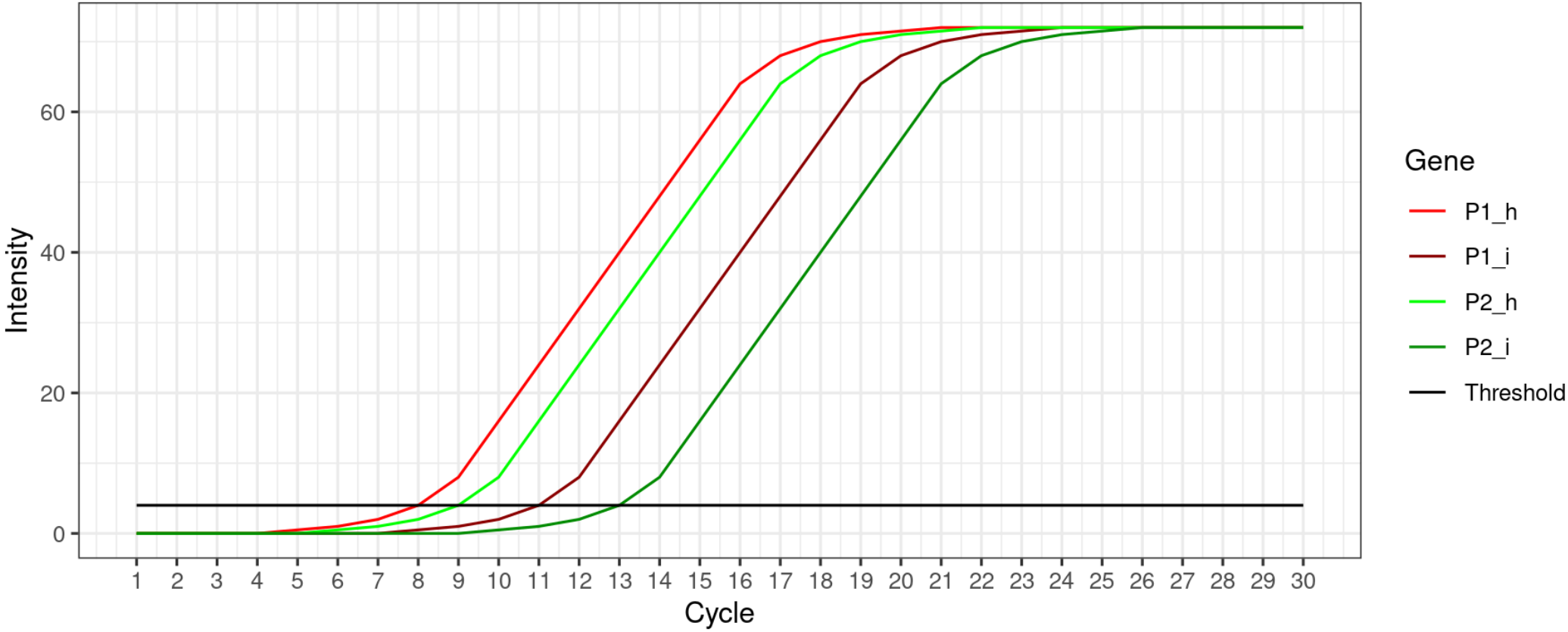




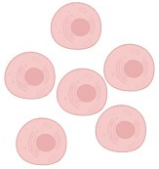
Уровень экспрессии генов



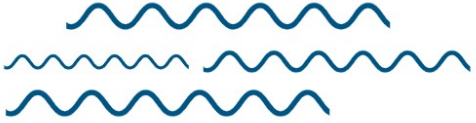
Задачка


Intensity
0
20
40
60
Cycle
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
Gene
P1_h
P1_i
P2_h
P2_i
Threshold

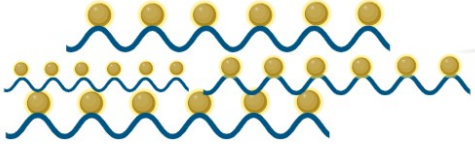
Принцип работы чипов для измерения уровня экспрессии



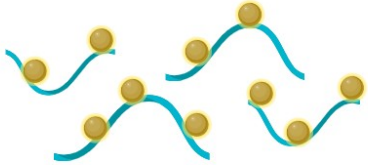
Выделение мРНК,
приготовление кДНК



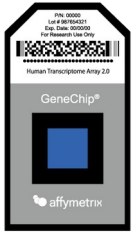
Биотинилирование кДНК

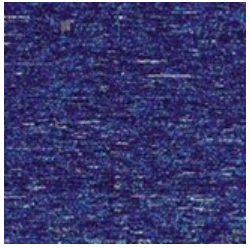


фрагментация



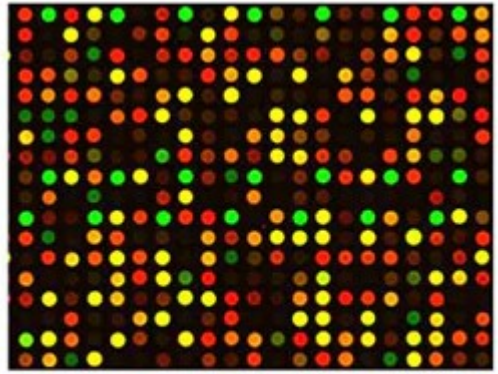
Нанесение на чип,
гибридизация, окрашивание
стрептавидин-фикоэритрин





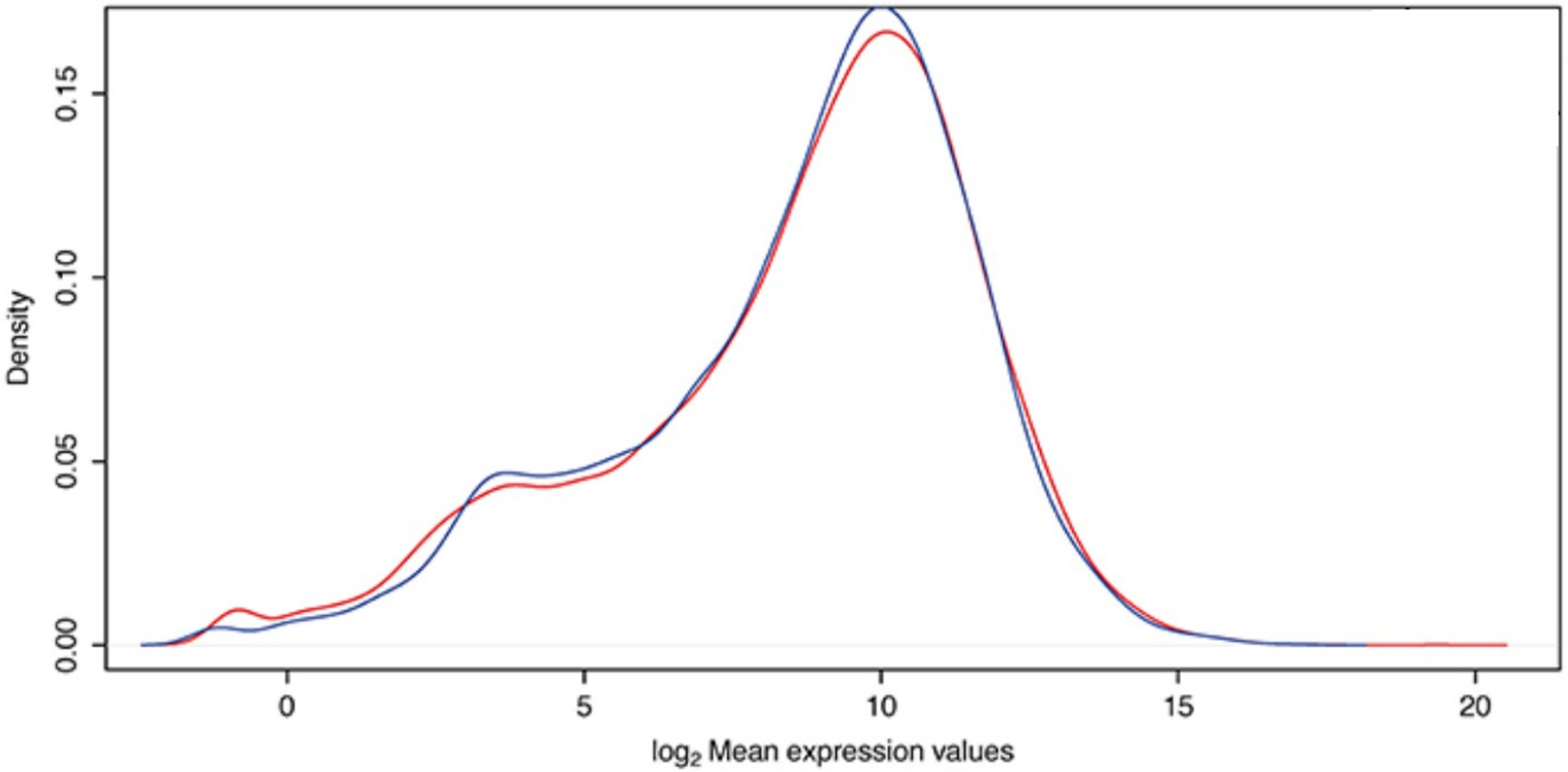
Чипы для измерения уровня экспрессии генов

Agilent Scanner



DNA Microarray

Нормализация сигнала



Анализ данных чипов

Фильтрация ячеек
с низким качеством сигнала

↓

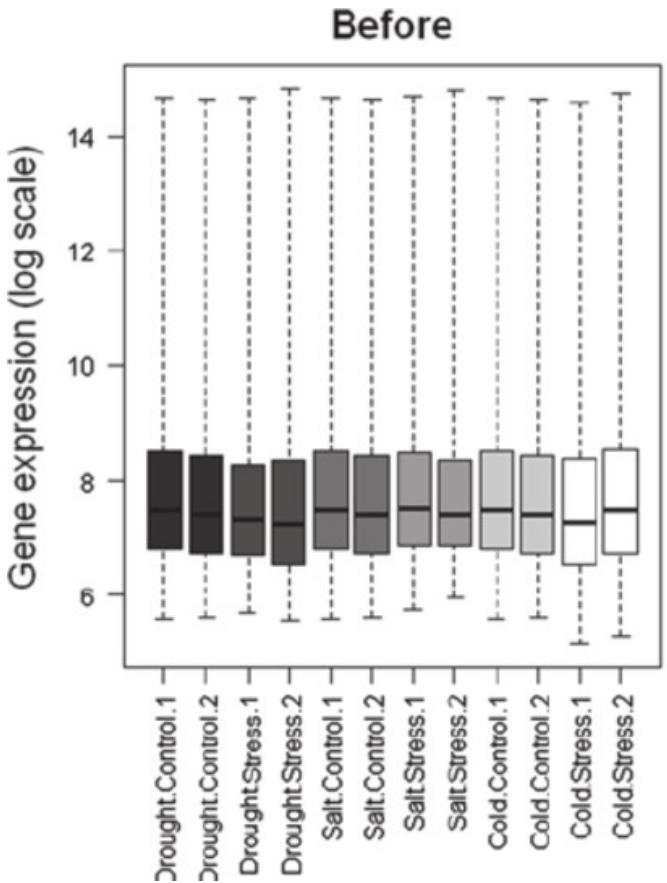
Нормализация образцов

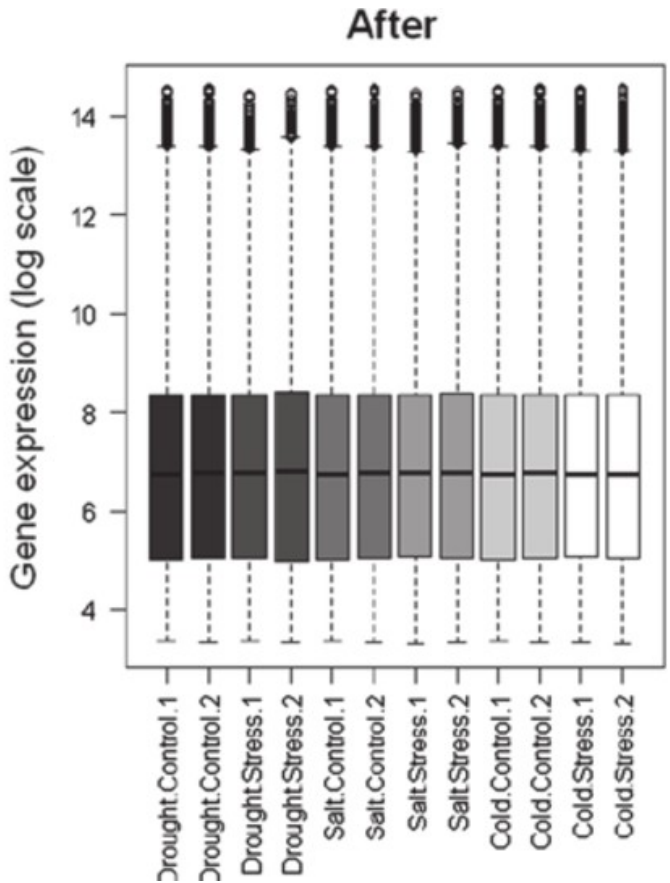
↓

Выделение статистически
значимых генов

↓

Кластерный анализ





Entry type
DataSets (53)
Series (587)
Samples (14,318)
Platforms (26)

Organism
Customize ...

Study type
Expression profiling by array
Methylation profiling by array
Customize ...

Author
Customize ...

Attribute name
tissue (6,369)
strain (410)
Customize ...

Publication

Summary 20 per page Sort by Default order Send to:

Search results

Items: 1 to 20 of 14984

<< First < Prev Page 1 of 750 Next > Last >>

1. MiR-155 deficiency alters heterogeneity of lung immune cells in a mouse model of severe **asthma**

(Submitter supplied) The lung is a complex tissue comprised of more than 40 cell populations. The interplay among the heterogeneous pulmonary cells is required for proper function and long-term maintenance of lung homeostasis. Because each of the cell populations has a distinct transcriptome, bulk RNA-seq could mask cell-cell crosstalk, cell-type-specific contributions, and relevant networking changes in the lung. To overcome the limitations of bulk tissue analysis, we performed single cell RNA sequencing (scRNA-seq) to investigate the transcriptional profiles of lungs affected by miR-155. more...

Organism: Mus
Type: Expression profiling by high throughput sequencing
Platform: GPL31136 6 Samples
Download data: MTX, TSV
Series Accession: GSE203656 ID: 200203656

Filters: Manage Filters

Top Organisms [Tree]
Homo sapiens *(14308)*
Mus musculus *(639)*
Macaca mulatta *(21)*
viruses *(21)*
Rattus norvegicus *(20)*
More...

Find related data
Database: Select
Find items

Search details

```
"asthma"[MeSH Terms] OR
asthma[All Fields]
```

NCBI GEO database

NCBI

Gene Expression Omnibus

GEO Publications | FAQ | MIAME | Email GEO

NCBI » GEO » GEO2R » GSE195996

Login

Use GEO2R to compare two or more groups of Samples in order to identify genes that are differentially expressed across experimental conditions. Results are presented as a table of genes ordered by significance. Full instructions

GEO accession GSE195996 Set Allergic sensitization and exposure to ambient air pollution beginning early in life leads to a COPD-like phenotype in young adult mice

Samples Define groups Selected **9** out of **12** samples

Enter a group name: List

Cancel selection

control (3 samples)

ova (3 samples)

PM (3 samples)

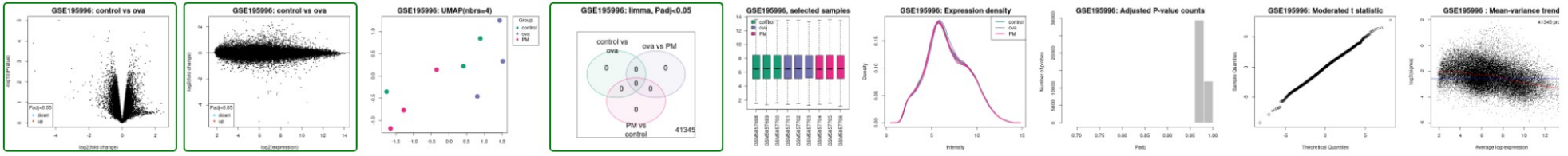
Columns Set

Group	Accession	Title	Source name	Treatment	Exposure
control	GSM5857698	Lung, Control1	Lung, Control	No ovalbumin	No PM2.5
control	GSM5857699	Lung, Control2	Lung, Control	No ovalbumin	No PM2.5
control	GSM5857700	Lung, Control3	Lung, Control	No ovalbumin	No PM2.5
ova	GSM5857701	Lung, OVA1	Lung, OVA	Ovalbumin	No PM2.5
ova	GSM5857702	Lung, OVA2	Lung, OVA	Ovalbumin	No PM2.5
ova	GSM5857703	Lung, OVA3	Lung, OVA	Ovalbumin	No PM2.5
PM	GSM5857704	Lung, PM1	Lung, PM	No ovalbumin	PM2.5
PM	GSM5857705	Lung, PM2	Lung, PM	No ovalbumin	PM2.5
PM	GSM5857706	Lung, PM3	Lung, PM	No ovalbumin	PM2.5
-	GSM5857707	Lung, OVA+PM1	Lung, OVA+PM	Ovalbumin	PM2.5
-	GSM5857708	Lung, OVA+PM2	Lung, OVA+PM	Ovalbumin	PM2.5
-	GSM5857709	Lung, OVA+PM3	Lung, OVA+PM	Ovalbumin	PM2.5

GEO2R | Options | Profile graph | R script

Reanalyze if you changed any options.

Visualization

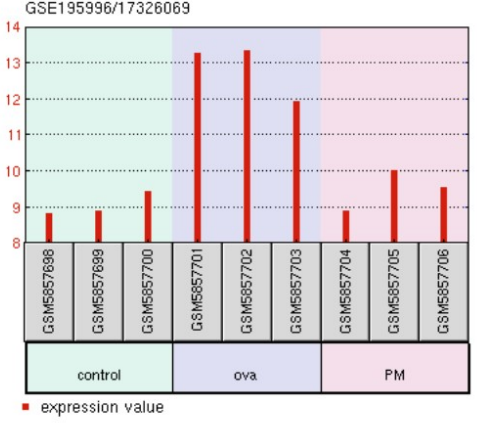


Top differentially expressed genes

Download full table Select columns

ID	adj.P.Val	P.Value	F	SPOT_ID
17326069	0.689	0.0000173	50.58	chr16(+):48842552-48844461
17257045	0.689	0.0000381	41.38	--not in current csv annotation file
17205805	0.689	0.0000695	35.45	--normgene->intron
17337583	0.689	0.00009	33.13	chr17(+):37672025-37672990

Sample values



GEO2R | Options | Profile graph | R script

```
# Version info: R 3.2.3, Biobase 2.30.0, GEOquery 2.40.0, limma 3.26.8
################################################################
#   Differential expression analysis with limma
library(GEOquery)
library(limma)
library(umap)

# load series and platform data from GEO

gset <- getGEO("GSE195996", GSEMatrix =TRUE, AnnotGPL=FALSE)
if (length(gset) > 1) idx <- grep("GPL16570", attr(gset, "names")) else idx <- 1
gset <- gset[[idx]]

# make proper column names to match toptable
fvarLabels(gset) <- make.names(fvarLabels(gset))

# group membership for all samples
gsms <- "000111222XXX"
sml <- strsplit(gsms, split="")[[1]]

# filter out excluded samples (marked as "X")
sel <- which(sml != "X")
sml <- sml[sel]
gset <- gset[ ,sel]

# log2 transformation
ex <- exprs(gset)
qx <- as.numeric(quantile(ex, c(0., 0.25, 0.5, 0.75, 0.99, 1.0), na.rm=T))
LogC <- (qx[5] > 100) ||
          (qx[6]-qx[1] > 50 && qx[2] > 0)
if (LogC) { ex[which(ex <= 0)] <- NaN
  exprs(gset) <- log2(ex) }

# assign samples to groups and set up design matrix
gs <- factor(sml)
groups <- make.names(c("control","ova","PM"))
levels(gs) <- groups
gset$group <- gs
design <- model.matrix(~group + 0, gset)
colnames(design) <- levels(gs)

fit <- lmFit(gset, design)  # fit linear model

# set up contrasts of interest and recalculate model coefficients
cts <- paste(groups, c(tail(groups, -1), head(groups, 1)), sep="-")
cont.matrix <- makeContrasts(contrasts=cts, levels=design)
fit2 <- contrasts.fit(fit, cont.matrix)

# compute statistics and table of top significant genes
fit2 <- eBayes(fit2, 0.01)
tT <- topTable(fit2, adjust="fdr", sort.by="B", number=250)
```