ITMO UNIVERSITY

CT Lab
ITMO UNIVERSITY

Microarrays

and introduction to gene expression studies

Kontantin Zaitsev

March 17th, 2020

Microarrays

Materials are at

/mnt/data/microarray

Installing libraries for today

```
if (!requireNamespace("BiocManager", quietly = TRUE)) install.packages("BiocManager")
if (!requireNamespace("affxparser", quietly = TRUE)) BiocManager::install("affxparser")
if (!requireNamespace("affy", quietly = TRUE)) BiocManager::install("affy")
if (!requireNamespace("GEOquery", quietly = TRUE)) BiocManager::install("GEOquery")
if (!requireNamespace("mouse4302.db", quietly = TRUE)) BiocManager::install("mouse4302.db
```

DNA Microarray

- **DNA Microarray** is a collection of microscopic **DNA spots** attached to a solid suface
- **DNA spot** contains copies of specific DNA sequence called **probes** (or oligoes)
- DNA probe is usually specific for a certain DNA region of certain mRNA
- DNA probe hybridizes with complement fluorescent labeled DNA (cDNA) molecule
- After that we can detect fluorescence

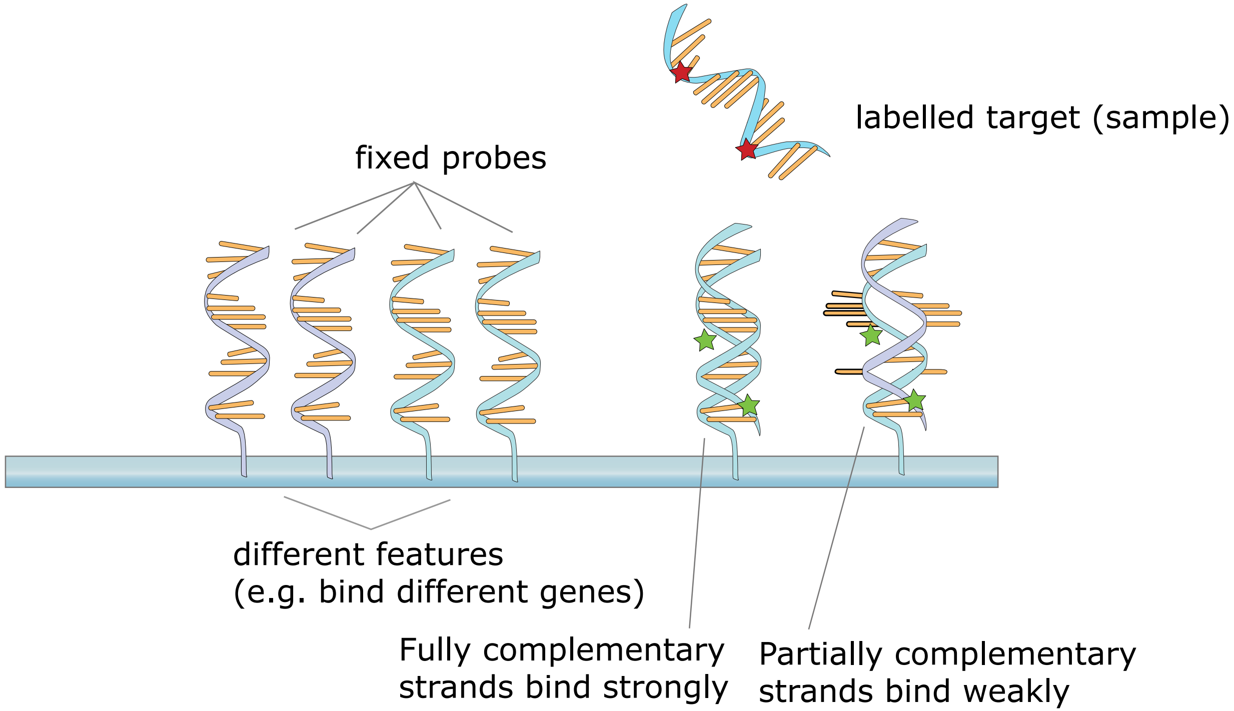
DNA Microarray: usages

- Genotyping: allele specific probes
- Tiling array: you can cover (like a whole chromosome) with overlapping probes to detect expression levels and coverage
- **Gene expression**: if probes are specific for different gene regions -- we can measure relative abundance of RNA within the sample
- Many other

Microarray: gene expression

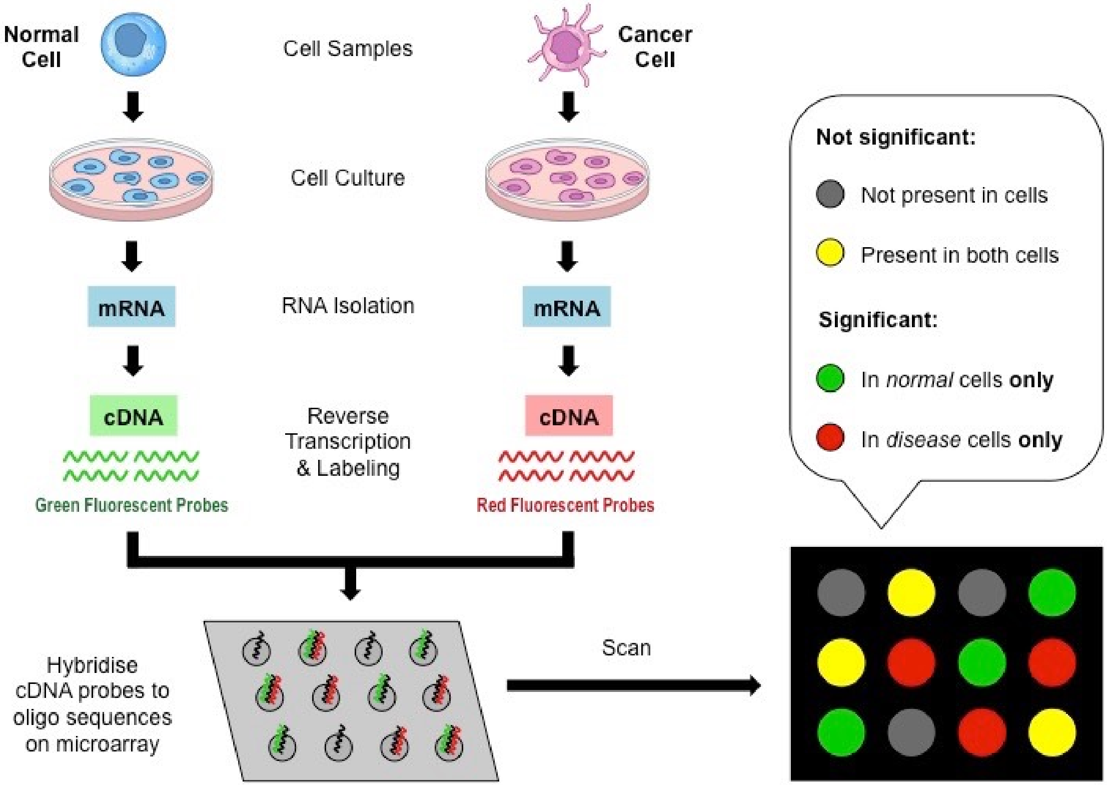
- Microarray can be considered as RNA capture technique
- Microarray consists of thousands of probes
- Probes consist of many oligonucleotides (all of which are the same within the probe)
- When cDNA hybridizes with complementary oligonucleotides, we detect fluoresence

Microarray



Taken from https://en.wikipedia.org/wiki/DNA_microarray

Microarray



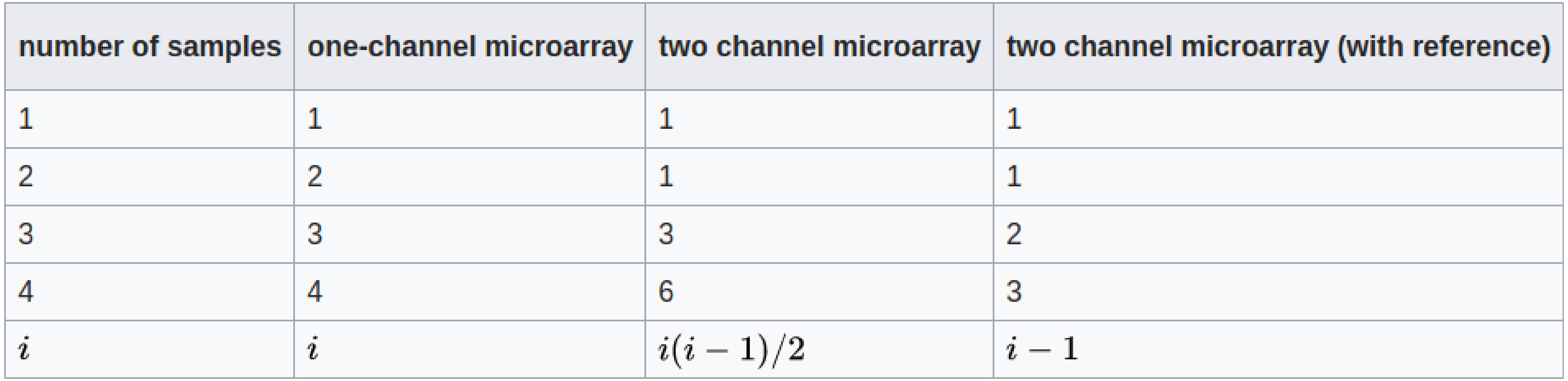
Taken from https://ib.bioninja.com.au/options/untitled/b4-medicine/dna-microarrays.html

Historical remark

- Researchers **used to do two-color microarray**: two samples could be processed with the same DNA chip
- Now most of the array are done in single-color: chips are relatively cheap
- But the legacy is huge:
  - people still do red-green heatmaps
  - all the schematics for microarray will be in red-green colors
  - GEO datasets still use green-magenta color-scheme

Historical remark

- You don't benefit much from two-channel microarray when working with larger number of samples



number of samples	one-channel microarray	two channel microarray	two channel microarray (with reference)
1	1	1	1
2	2	1	1
3	3	3	2
4	4	6	3
i	i	$i(i-1)/2$	$i-1$

Taken from https://en.wikipedia.org/wiki/DNA_microarray

Dataset for today: GSE129260

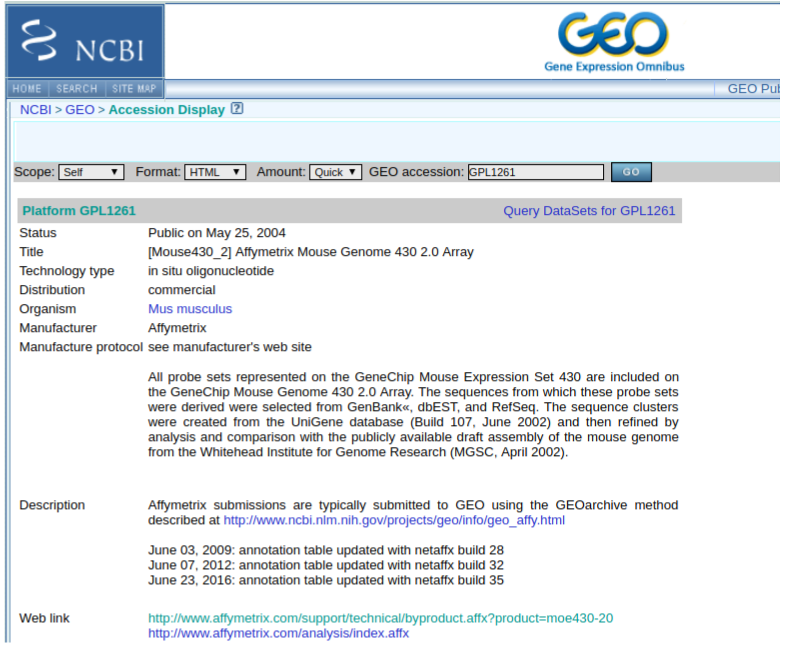
https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE129260

In short:

- B cells
- IL10 positive and negative
- Treated with LPS and anti-CD40
- In total four groups, two replicates in each group

Affymetrix microarrays

Most likely, microarray gene expression data will come from affymetrix microarray.



NCBI > GEO > Accession Display

Scope: Self ▼ Format: HTML ▼ Amount: Quick ▼ GEO accession: GPL1261 GO

Platform GPL1261	Query DataSets for GPL1261
Status	Public on May 25, 2004
Title	[Mouse430_2] Affymetrix Mouse Genome 430 2.0 Array
Technology type	in situ oligonucleotide
Distribution	commercial
Organism	Mus musculus
Manufacturer	Affymetrix
Manufacture protocol	see manufacturer's web site All probe sets represented on the GeneChip Mouse Expression Set 430 are included on the GeneChip Mouse Genome 430 2.0 Array. The sequences from which these probe sets were derived were selected from GenBank«, dbEST, and RefSeq. The sequence clusters were created from the UniGene database (Build 107, June 2002) and then refined by analysis and comparison with the publicly available draft assembly of the mouse genome from the Whitehead Institute for Genome Research (MGSC, April 2002).
Description	Affymetrix submissions are typically submitted to GEO using the GEOarchive method described at http://www.ncbi.nlm.nih.gov/projects/geo/info/geo_affy.html June 03, 2009: annotation table updated with netaffx build 28 June 07, 2012: annotation table updated with netaffx build 32 June 23, 2016: annotation table updated with netaffx build 35
Web link	http://www.affymetrix.com/support/technical/byproduct.affx?product=moe430-20 http://www.affymetrix.com/analysis/index.affx

Taken from ttps://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL1261

Raw microarray files

Different microarray platforms have different specifications about:

- What are the probes
- Where probes are physically located on the chip
- This is usually desrcribed in CDF file (Chip Description File)

Raw microarray files

Raw files for microarray are just fluorescence itensities for a chip (CEL files).

So if you are running microarray from scratch you will have:

- CEL file for each sample
- CDF file for your microarray platform

Raw microarray files: time for some code

```r
library(affxparser)
library(affy)

CELfile <- readCel("GSE129260_RAW/GSM3703675_IL-10_posi_anti-CD40-1.CEL")
```

Raw microarray files: time for some code

```
head(CELfile$header)
```

```
## $filename
## [1] "GSE129260_RAW/GSM3703675_IL-10_posi_anti-CD40-1.CEL"
##
## $version
## [1] 1
##
## $cols
## [1] 1002
##
## $rows
## [1] 1002
##
## $total
## [1] 1004004
##
## $algorithm
## [1] "Feature Extraction Cell Generation"
```

Raw microarray files: time for some code

```
head(CELfile$intensities)
```

```
## [1]   78 4808  106 5064  147   74
```

Converting CEL to features

In most cases we don't need to do that ourselves.

IN MOST CASES YOU REALLY DON'T WANT TO DO THAT

Realistically

Affymetrix arrays come with tools to:

- Get the feature expression values
- Normalize expression levels
- These tools are standartized and available in Bioconductor

Public data

```
files <- list.files("GSE129260_RAW/", full.names = T)
microarrayData <- justRMA(filenames = files)
```

```
## Warning: replacing previous import 'AnnotationDbi::tail' by 'utils::tail'
## when loading 'mouse4302cdf'
```

```
## Warning: replacing previous import 'AnnotationDbi::head' by 'utils::head'
## when loading 'mouse4302cdf'
```

```
##
```

Public data

```r
exprs(microarrayData)[1:5, 1:2]
```

```
##              GSM3703675_IL-10_posi_anti-CD40-1.CEL
## 1415670_at                                8.779487
## 1415671_at                                8.920662
## 1415672_at                                8.976458
## 1415673_at                                6.666717
## 1415674_a_at                              9.006864
##              GSM3703676_IL-10_nega_anti-CD40-1.CEL
## 1415670_at                                8.561089
## 1415671_at                                8.895862
## 1415672_at                                8.956885
## 1415673_at                                6.467003
## 1415674_a_at                              9.099215
```

Normalization

- Raw Affy data contains about twenty probes for the same RNA target
- Half of these are "mismatch spots", which do not precisely match the target sequence
- These can theoretically measure the amount of nonspecific binding for a given target

Normalization

- Robust Multi-array Average (**RMA**) is a normalization approach that does not take advantage of these mismatch spots, but still must summarize the perfect matches through median polish
- The current Affymetrix **MAS5** algorithm, which uses both perfect match and mismatch probes, continues to enjoy popularity and do well in head to head tests

How to get symbols ?

```r
library(mouse4302.db)

symbolAnnotation <- as.list(mouse4302SYMBOL)
head(symbolAnnotation, 3)
```

```
## $`1415670_at`
## [1] "Copg1"
##
## $`1415671_at`
## [1] "Atp6v0d1"
##
## $`1415672_at`
## [1] "Golga7"
```

Public data: GEOquery

We have much easier ways to get annotation for samples/probes with GEOquery

```r
library(GEOquery)
GSE129260 <- getGEO("GSE129260", AnnotGPL = TRUE)[[1]]
```

```
## Warning: 64 parsing failures.
##   row             col           expected    actual         file
## 45038 Platform_SPOTID 1/0/T/F/TRUE/FALSE --Control literal data
## 45039 Platform_SPOTID 1/0/T/F/TRUE/FALSE --Control literal data
## 45040 Platform_SPOTID 1/0/T/F/TRUE/FALSE --Control literal data
## 45041 Platform_SPOTID 1/0/T/F/TRUE/FALSE --Control literal data
## 45042 Platform_SPOTID 1/0/T/F/TRUE/FALSE --Control literal data
## ..... ............... .................. ......... ............
## See problems(...) for more details.
```

Public data: GEOquery

```
dim(exprs(GSE129260))
```

```
## [1] 45101     8
```

Public data: GEOquery

```
head(exprs(GSE129260))
```

```
##              GSM3703675 GSM3703676 GSM3703677 GSM3703678 GSM3703679
## 1415670_at     439.3887  377.51083   597.2262   493.4291   397.1739
## 1415671_at     484.9137  476.33698   674.3707   600.3790   581.9670
## 1415672_at     503.6775  496.95230   501.6765   595.6385   750.4399
## 1415673_at     101.6343   88.44778   644.5211   442.4400   262.5089
## 1415674_a_at   514.6692  548.55630   418.8315   527.6122   549.1731
## 1415675_at     343.5385  373.09020   347.4206   441.6546   380.8810
##              GSM3703680 GSM3703681 GSM3703682
## 1415670_at     382.7747   674.1451   504.8682
## 1415671_at     645.0598   752.7134   816.8667
## 1415672_at     784.1332   840.3690   827.8496
## 1415673_at     298.3548   942.3843   593.4224
## 1415674_a_at   548.4058   516.0414   526.1480
## 1415675_at     374.6690   316.8613   359.3207
```

Public data: GEOquery

```
head(pData(GSE129260)[, 1:2])
```

```
##                                                                   title
## GSM3703675 IL-10 positive B cells, anti-CD40 for 48 h, biological rep1
## GSM3703676 IL-10 negative B cells, anti-CD40 for 48 h, biological rep1
## GSM3703677       IL-10 positive B cells, LPS for 48 h, biological rep1
## GSM3703678       IL-10 negative B cells, LPS for 48 h, biological rep1
## GSM3703679 IL-10 positive B cells, anti-CD40 for 48 h, biological rep2
## GSM3703680 IL-10 negative B cells, anti-CD40 for 48 h, biological rep2
##            geo_accession
## GSM3703675    GSM3703675
## GSM3703676    GSM3703676
## GSM3703677    GSM3703677
## GSM3703678    GSM3703678
## GSM3703679    GSM3703679
## GSM3703680    GSM3703680
```

Public data: GEOquery

```
head(fData(GSE129260)[, 1:2])
```

```
##                         ID
## 1415670_at     1415670_at
## 1415671_at     1415671_at
## 1415672_at     1415672_at
## 1415673_at     1415673_at
## 1415674_a_at 1415674_a_at
## 1415675_at     1415675_at
##                                                      Gene title
## 1415670_at           coatomer protein complex, subunit gamma 1
## 1415671_at    ATPase, H+ transporting, lysosomal V0 subunit D1
## 1415672_at            golgi autoantigen, golgin subfamily a, 7
## 1415673_at                           phosphoserine phosphatase
## 1415674_a_at            trafficking protein particle complex 4
## 1415675_at   dolichol-phosphate (beta-D) mannosyltransferase 2
```