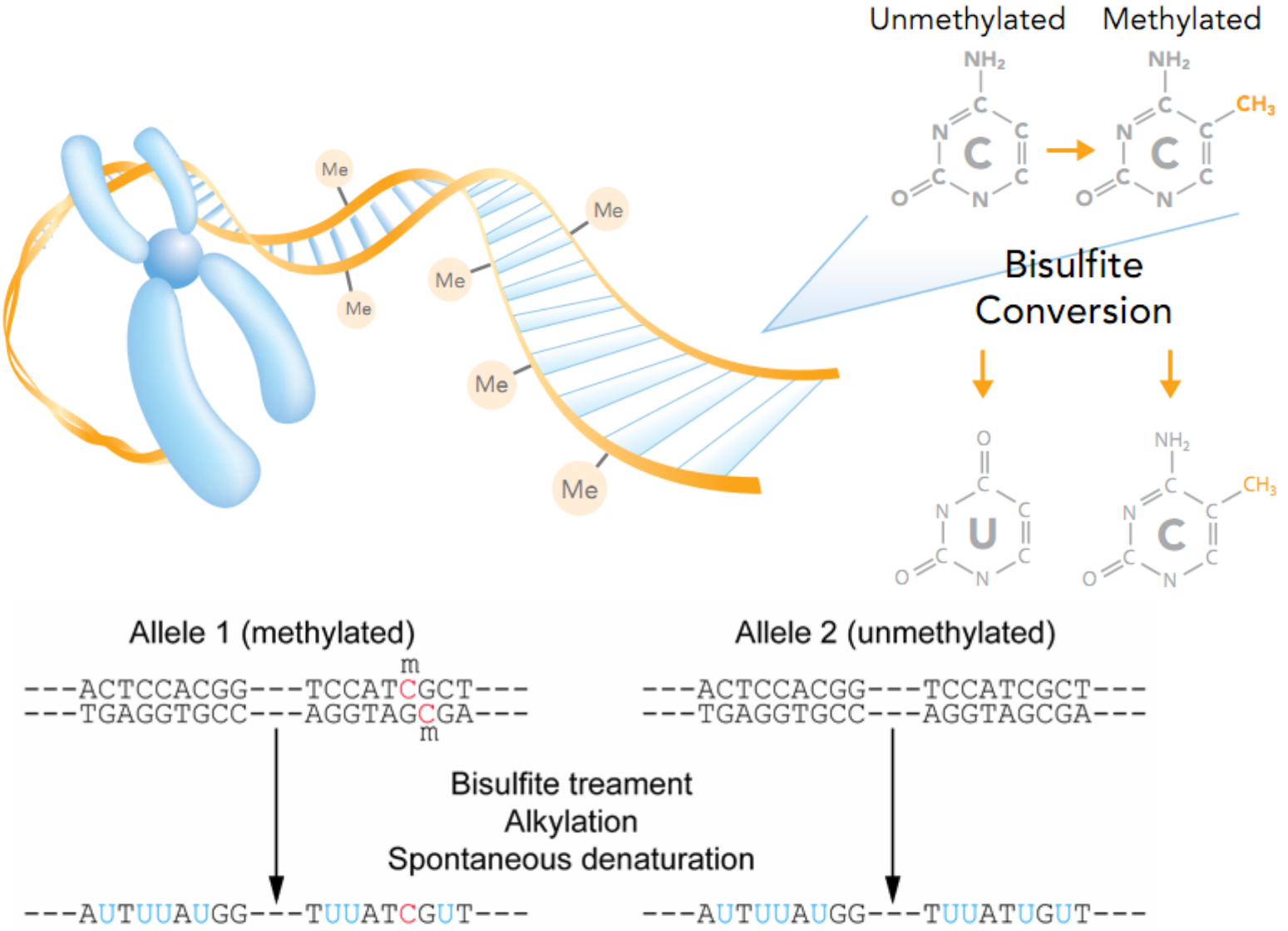
DNA co-methylation analysis of breast cancer samples

Presented by Ruslan Akulenko

Outline

- Introduction:
  - DNA methylation;
  - Co-methylation example;
- DNA methylation data processing:
  - Tumor data;
  - Data processing;
  - Comparison against the randomized data;
- Results:
  - Functional similarity of co-methylated genes (CMG);
  - Disease candidate mapping;
  - Distance vs methylation;
  - Pathway enrichment;
- Summary
- Assignment 5 - introduction

Recall – DNA Methylation



Recall – Cancer & DNA methylation

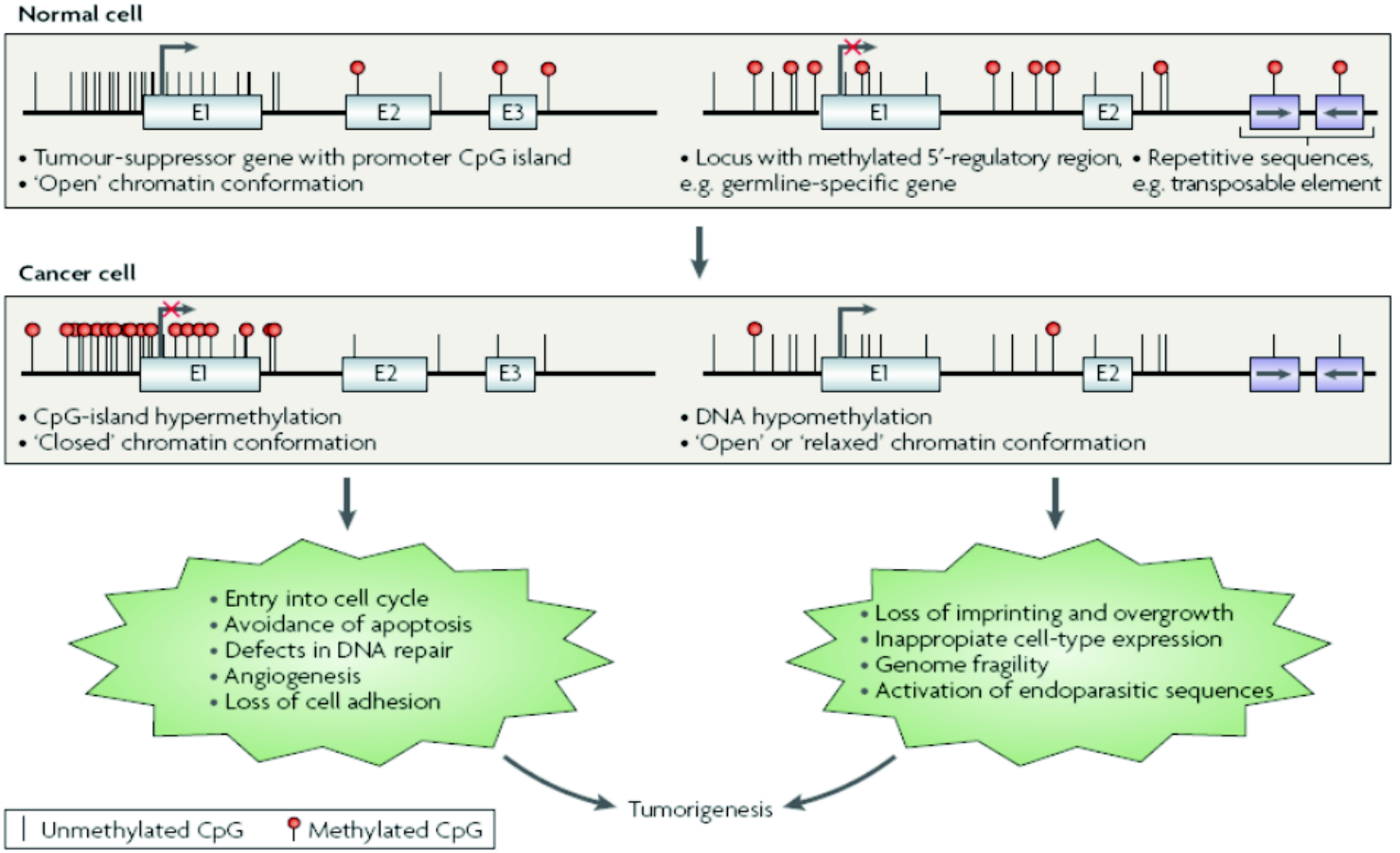


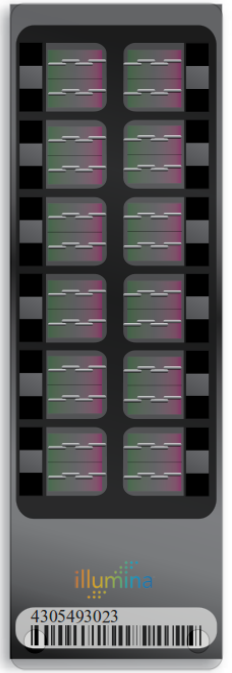
Figure 1 | **Altered DNA-methylation patterns in tumorigenesis.** The hypermethylation of CpG islands of tumour-suppressor genes is a common alteration in cancer cells, and leads to the transcriptional inactivation of these genes and the loss of their normal cellular functions. This contributes to many of the hallmarks of cancer cells. At the same time, the genome of the cancer cell undergoes global hypomethylation at repetitive sequences, and tissue-specific and imprinted genes can also show loss of DNA methylation. In some cases, this hypomethylation is known to contribute to cancer cell phenotypes, causing changes such as loss of imprinting, and might also contribute to the genomic instability that characterizes tumours. E, exon.

Esteller, Nat. Rev. Gen. 8, 286 (2007)

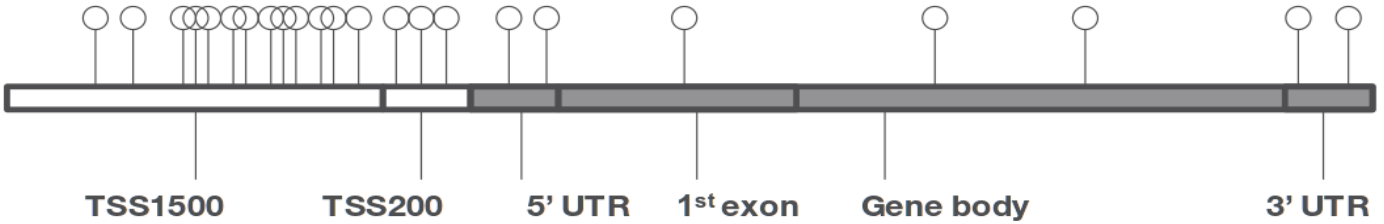
How to measure DNA methylation?

Infinium HumanMethylation27

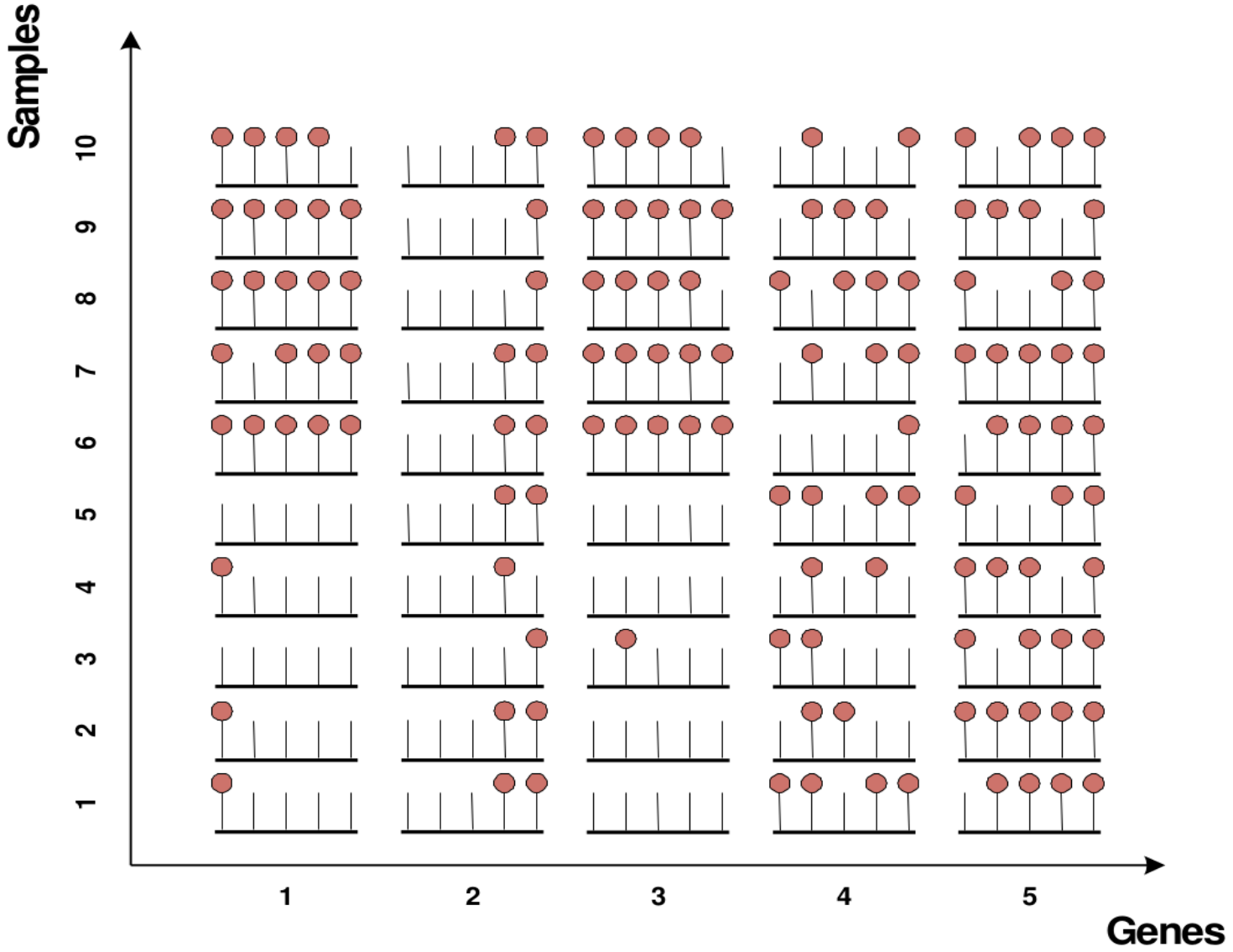
Infinium HumanMethylation450



Cluster CG#	CHR #	Coordinate	Genome Build	Sequence
cg00009407	14	88,360,674	36	...GGCG[CG]CTGC...
cg00003994	7	15,692,387	36	...TCTT[CG]TTGG...
cg00000292	16	28,797,601	36	...AATA[CG]GCCT...
cg00002426	3	57,718,583	36	...ACCA[CG]CTCT...
cg00005847	2	176,737,319	36	...ATGG[CG]CTTT...
cg00006414	7	148,453,770	36	...GGCG[CG]ATCC...



The co-methylation example



Outline

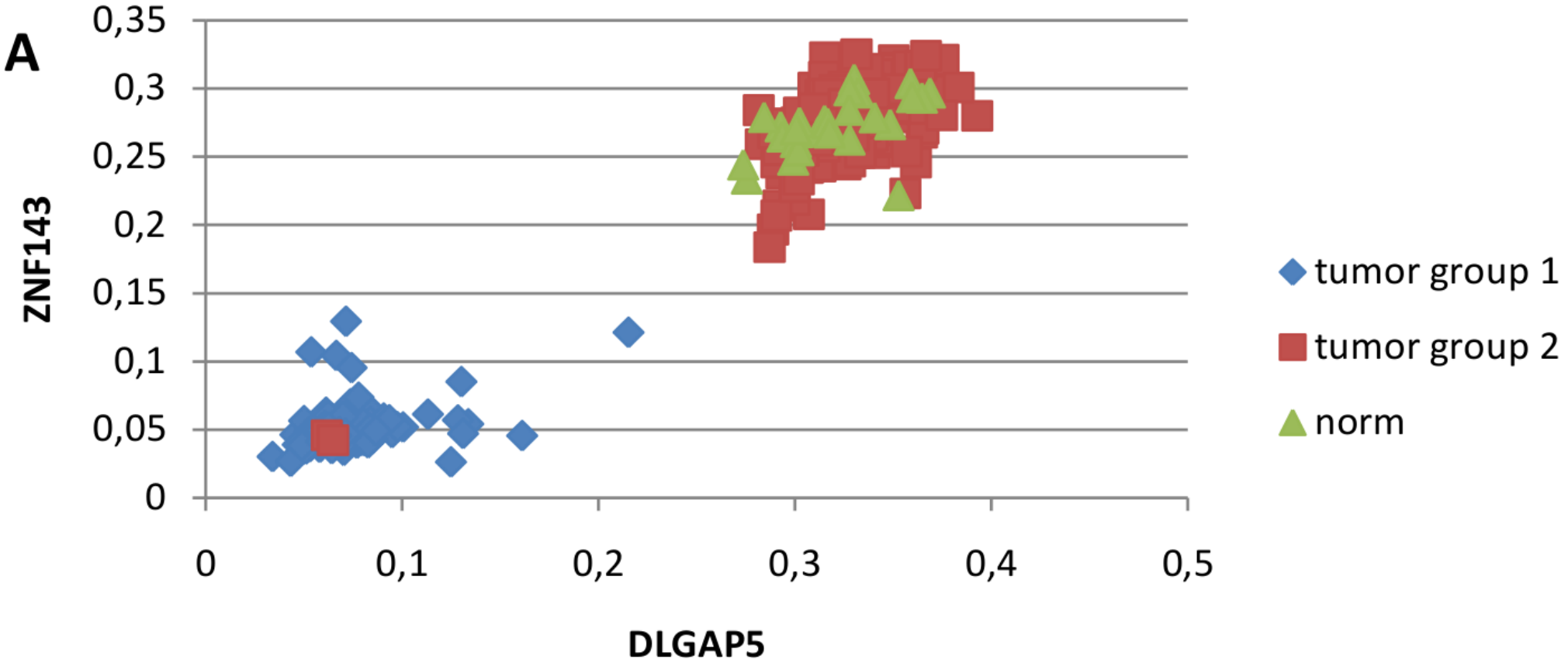
- Introduction:
  - DNA methylation;
  - Co-methylation example;
- DNA methylation data processing:
  - Tumor data;
  - Data processing;
  - Comparison against the randomized data;
- Results:
  - Functional similarity of co-methylated genes (CMG);
  - Disease candidate mapping;
  - Distance vs methylation;
  - Pathway enrichment;
- Summary
- Assignment 5 - introduction

Tumor data

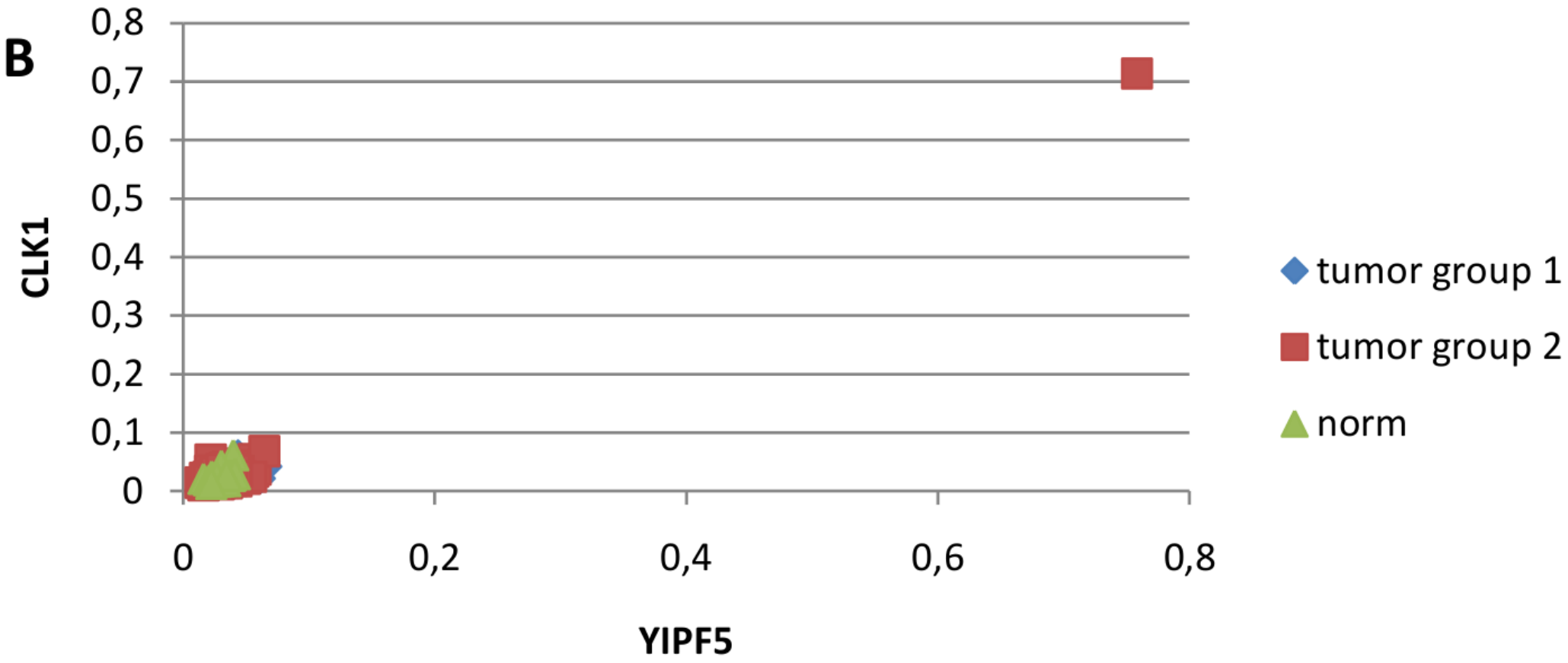
Data Type (Base-Specific)	Level 1 (Raw Data)	Level 2 (Normalized/ Processed)	Level 3 (Segmented/ Interpreted)	Level 4 (Summary Finding/ROI)
DNA Methylation	Raw signals per probe	Normalized signals per probe or probe set and allele calls	Methylated sites/genes per sample	Statistically significant methylated sites/genes across samples

- 183 available tumor samples deposited in September 2011 (tumor group 1);
- 134 tumor samples (tumor group 2) and
- 27 matched normal that were both deposited in October 2011.

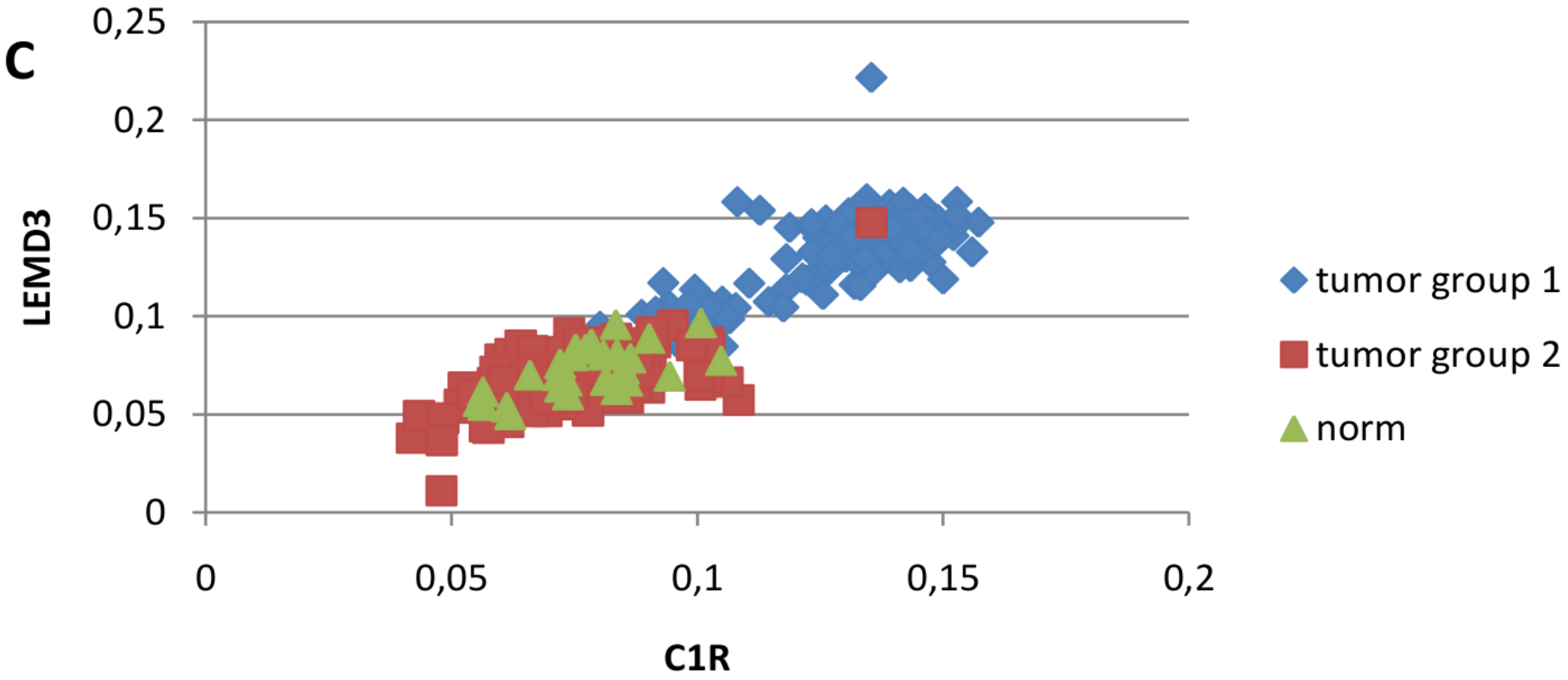
Difficulties: batch effect



Difficulties: outliers



Difficulties: low variance



Solution

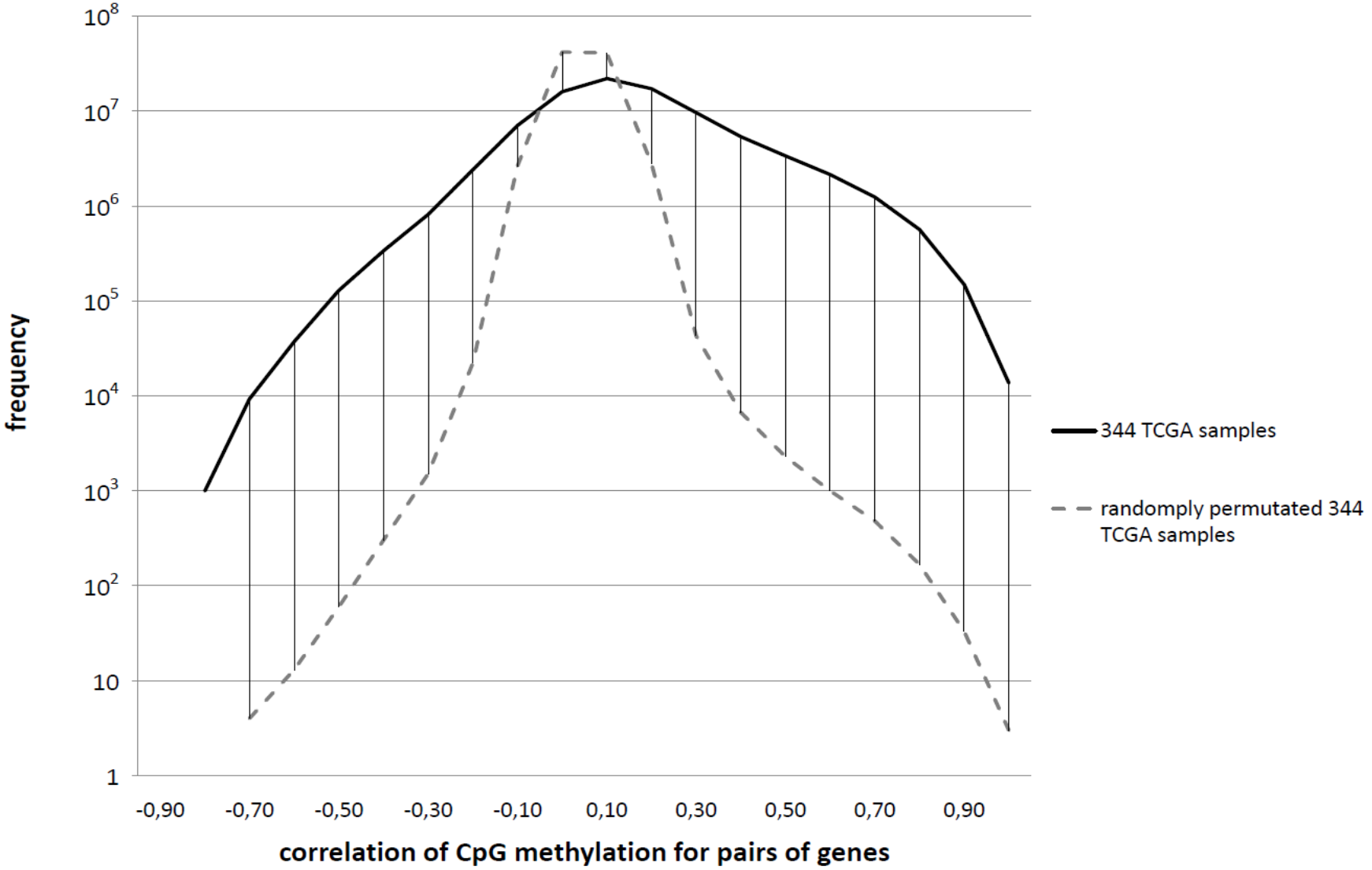
- Preprocessing:
  - Importing files to the SQL database;
  - Cleaning data;
- Filtering/removing:
  - Batch effect;
  - Genes with at least one outlier according to boxplot.stats;

$$\underset{i \in T}{\text{outlier}}(beta_i) = 0$$

  - Large interquartile range;

$$\underset{i \in T}{quartile3}(beta_i) - \underset{i \in T}{quartile1}(beta_i) \geq 0{,}1$$

Comparison against the randomized data



Outline

- Introduction:
  - DNA methylation;
  - Co-methylation example;
- DNA methylation data processing:
  - Tumor data;
  - Data processing;
  - Comparison against the randomized data;
- Results:
  - Functional similarity of co-methylated genes (CMG);
  - Disease candidate mapping;
  - Distance vs methylation;
  - Pathway enrichment;
- Summary
- Assignment 5 - introduction

The Gene Ontology

GO provides an ontology of defined terms representing gene product properties.

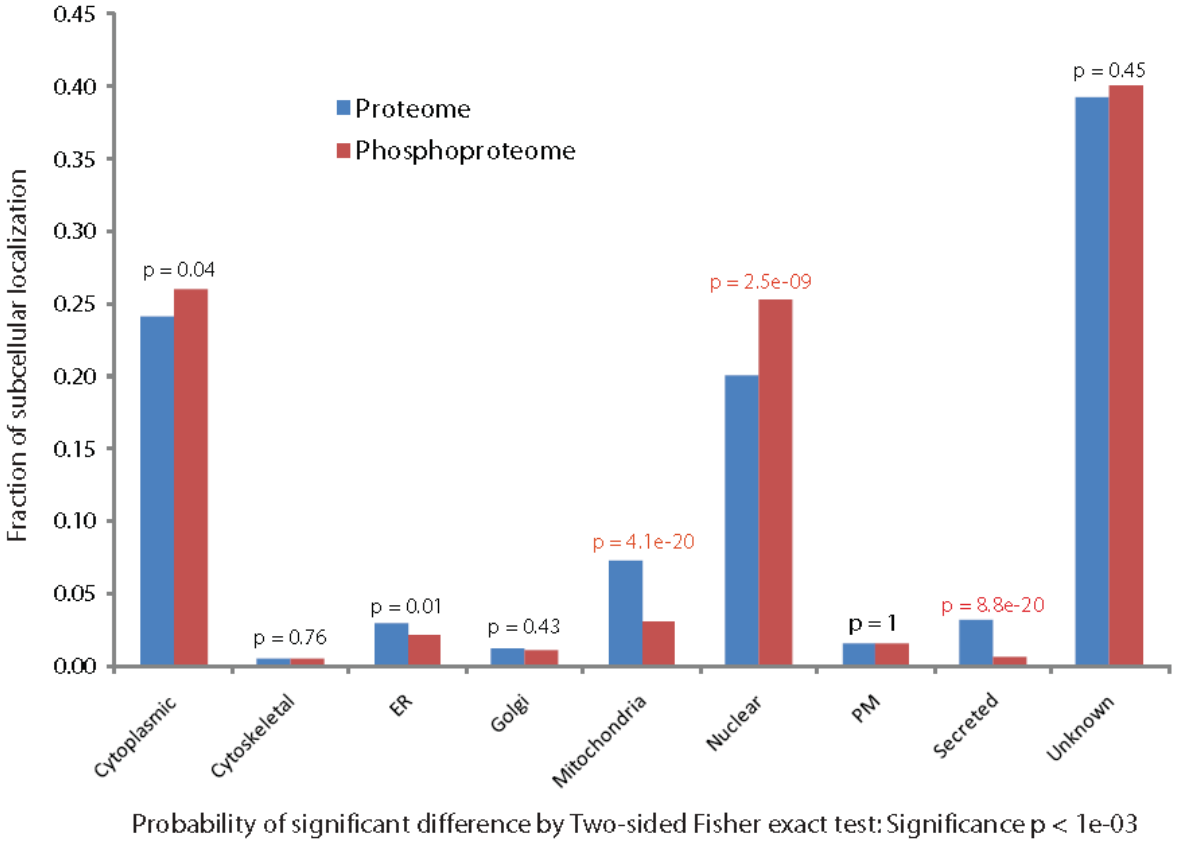
The ontology covers three domains:

- cellular component, the parts of a cell or its extracellular environment;
- molecular function, the elemental activities of a gene product at the molecular level, such as binding or catalysis;
- biological process, operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units: cells, tissues, organs, and organisms.

  The GO ontology is structured as a directed acyclic graph, and each term has defined relationships to one or more other terms in the same domain, and sometimes to other domains.

Recall - Differential phosphorylation

Gene ontology (GO) analysis of protein and phosphoproteins subcellular localization. All proteins identified by MS were clustered according to their GO annotation for sub-cellular localization (Blue bars). The same clustering was done for all phosphoproteins (Red bars).

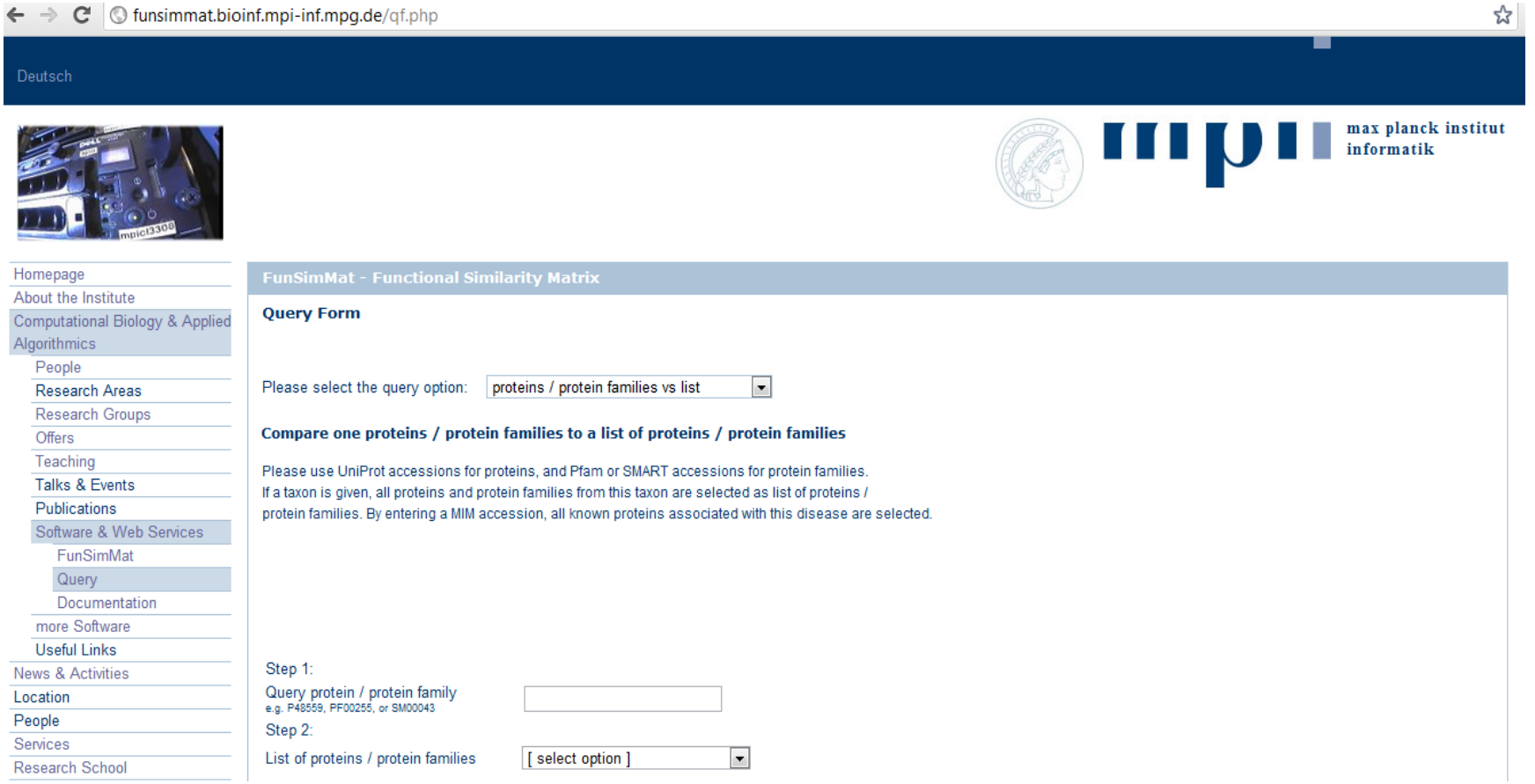


y-axis : percentage of the indicated sub-cellular fractions from the total.

Compared to the proteome distribution, phosphorylated proteins are over-represented in the nucleus and under-represented amongst mitochondrial and secreted proteins.

Olsen Science Signaling 3 (2010)

Functional similarity

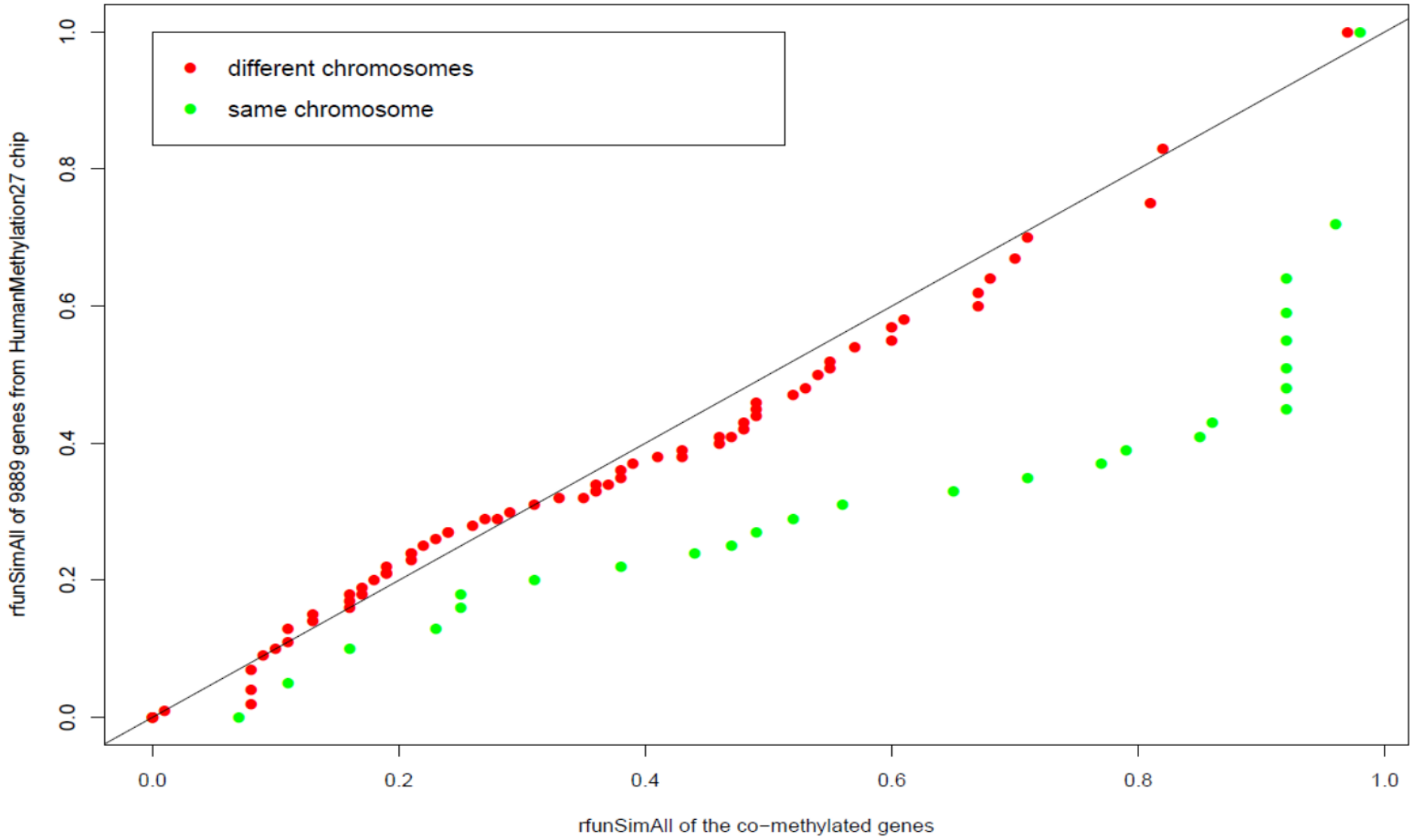


$$funSimAll(p,q) = \frac{1}{3} \cdot \left[\left(\frac{BPscore(p,q)}{max_{BPscore}} \right)^2 + \left(\frac{MFscore(p,q)}{max_{MFscore}} \right)^2 + \left(\frac{CCscore(p,q)}{max_{CCscore}} \right)^2 \right]$$

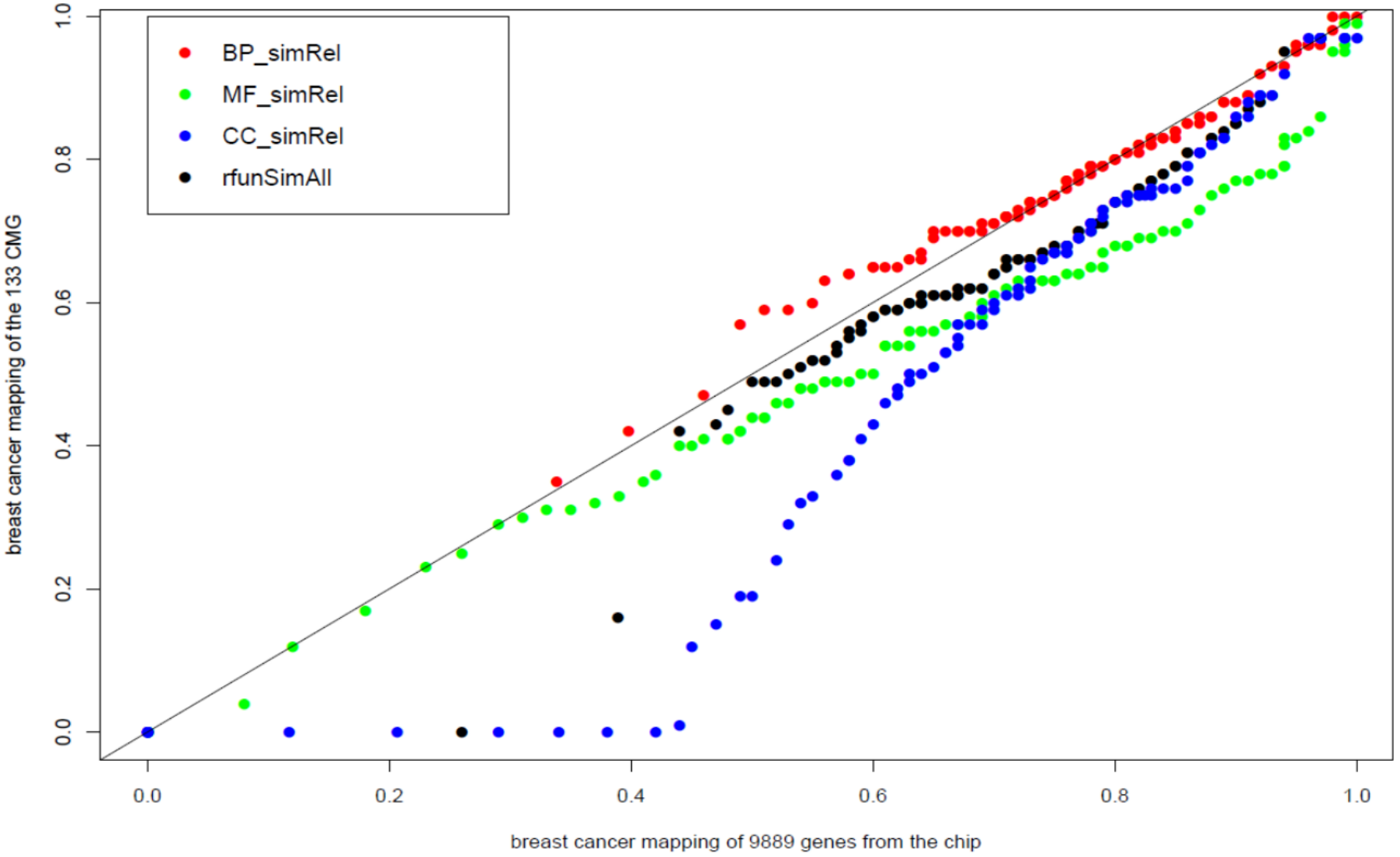
Interesting observation – top 10 gene pairs

First gene	Second gene	Pearson correlation
SPRR1B	SPRR1A	0,872
FCN2	FCN1	0,870
CD244	CD48	0,866
SPRR1B	SPRR4	0,862
TAS2R13	PRB4	0,859
F7	TFF1	0,856
SH3TC2	SPARCL1	0,853
ABCE1	SC4MOL	0,849
REG1B	REG1P	0,846
SPRR3	SPRR4	0,843

rfunSimAll of CMG vs all genes



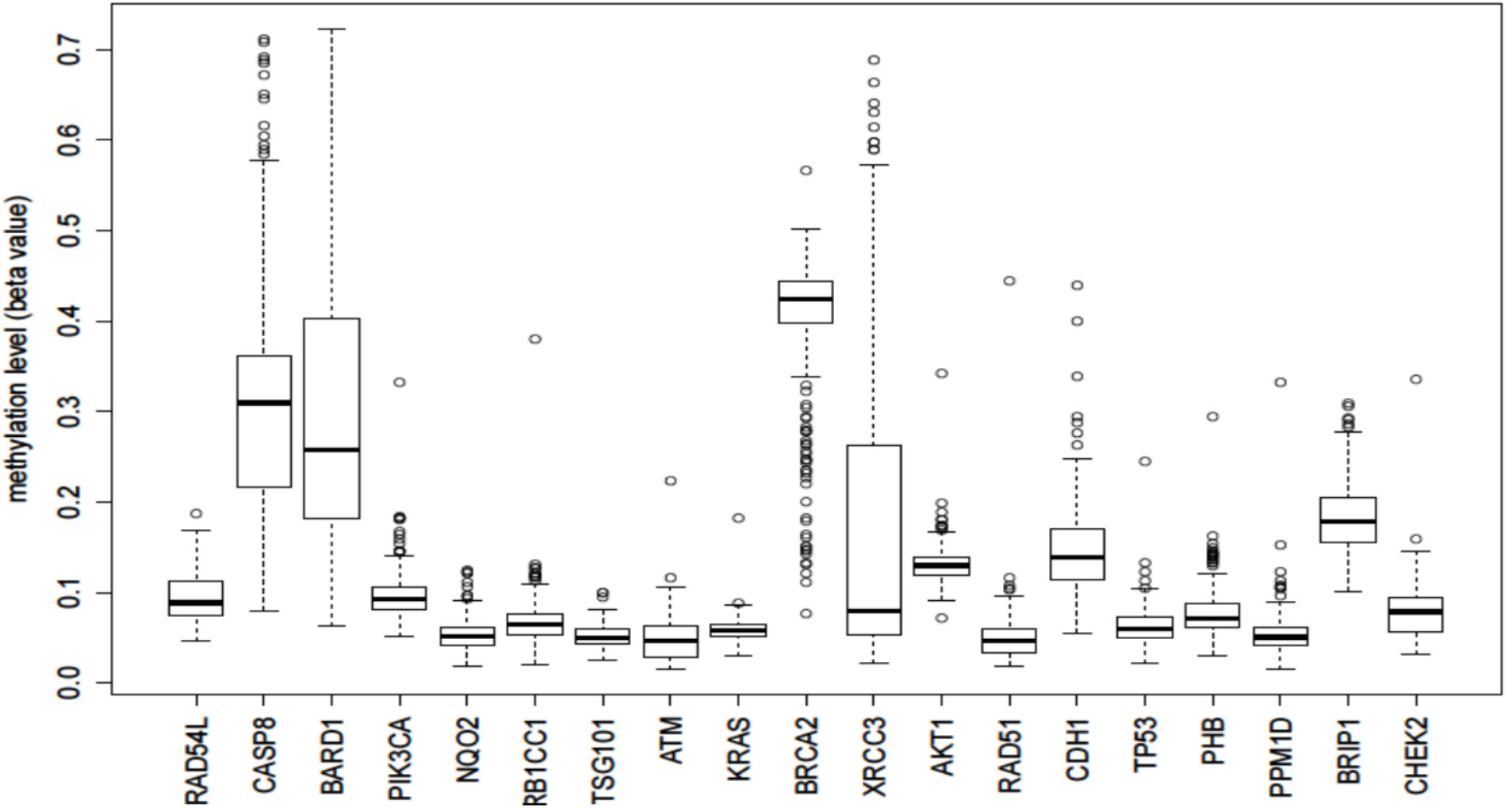
Disease mapping



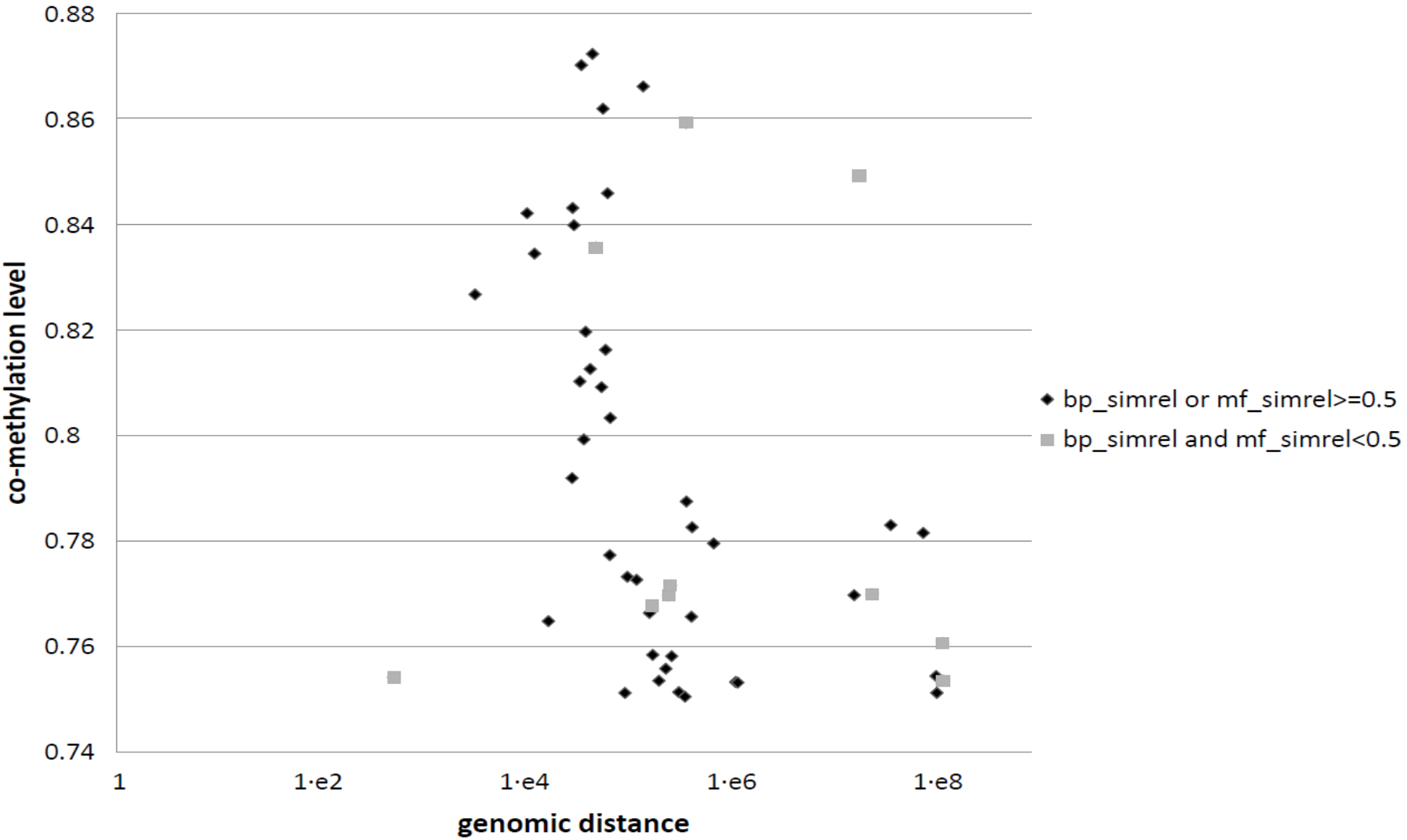
OMIM breast cancer genes, 114480

Location	Phenotype	Phenotype MIM number	Gene/Locus	Gene/Locus MIM number
1p34.1	{Breast cancer, invasive ductal}	114480	RAD54L	603615
2q33.1	{Breast cancer, protection against}	114480	CASP8	601763
2q35	{Breast cancer, susceptibility to}	114480	BARD1	601593
3q26.32	Breast cancer, somatic	114480	PIK3CA	171834
5q34	{Breast cancer, susceptibility to}	114480	HMMR	600936
6p25.2	{?Breast cancer susceptibility}	114480	NQO2	160998
8q11.23	Breast cancer, somatic	114480	RB1CC1	606837
11p15.4	Breast cancer, somatic	114480	SLC22A1L	602631
11p15.1	Breast cancer, somatic	114480	TSG101	601387
11q22.3	{Breast cancer, susceptibility to}	114480	ATM	607585
12p12.1	Breast cancer, somatic	114480	KRAS	190070
13q13.1	{Breast cancer, male, susceptibility to}	114480	BRCA2	600185
14q32.33	{Breast cancer, susceptibility to}	114480	XRCC3	600675
14q32.33	Breast cancer, somatic	114480	AKT1	164730
15q15.1	{Breast cancer, susceptibility to}	114480	RAD51A	179617
16p12.2	{Breast cancer, susceptibility to}	114480	PALB2	610355
16q22.1	{Breast cancer, lobular}	114480	CDH1	192090
17p13.1	Breast cancer	114480	TP53	191170
17q21.33	{Breast cancer, susceptibility to}	114480	PHB	176705
17q23.2	Breast cancer	114480	PPM1D	605100
17q23.2	Breast cancer, early-onset	114480	BRIP1	605882
22q12.1	{Breast cancer, susceptibility to}	114480	CHEK2	604373

DNA methylation of “OMIM genes”



Co-methylation and genomic distance



Pathway enrichment

KEGG pathways	p-value	Genes involved in pathways	FDR
hsa04950:Maturity onset diabetes of the young	0.003	HNF1B, FOXA2, NEUROD1	2.622
hsa04640:Hematopoietic cell lineage	0.009	CD1A, CD1E, CD1D	6.229
hsa04730:Long-term depression	0.004	GRM5, C7ORF16, PRKG2	2.952
hsa04512:ECM-receptor interaction	0.005	COL5A2, COL11A1, SPP1	3.500

Summary

- New three step filtering of DNA methylation data;
- Unexpectedly highly correlated DNA methylation levels are found in gene pairs from breast cancer samples.
- Correlated gene pairs show strong combined functional similarity.
- DNA methylation corelates with some pathways;

Literature: Akulenko R, Helms V. DNA co-methylation analysis suggests novel functional associations between gene pairs in breast cancer samples. Hum Mol Genet. 2013 Apr 16.

http://hmg.oxfordjournals.org/content/early/2013/04/16/hmg.ddt158.full.pdf+html

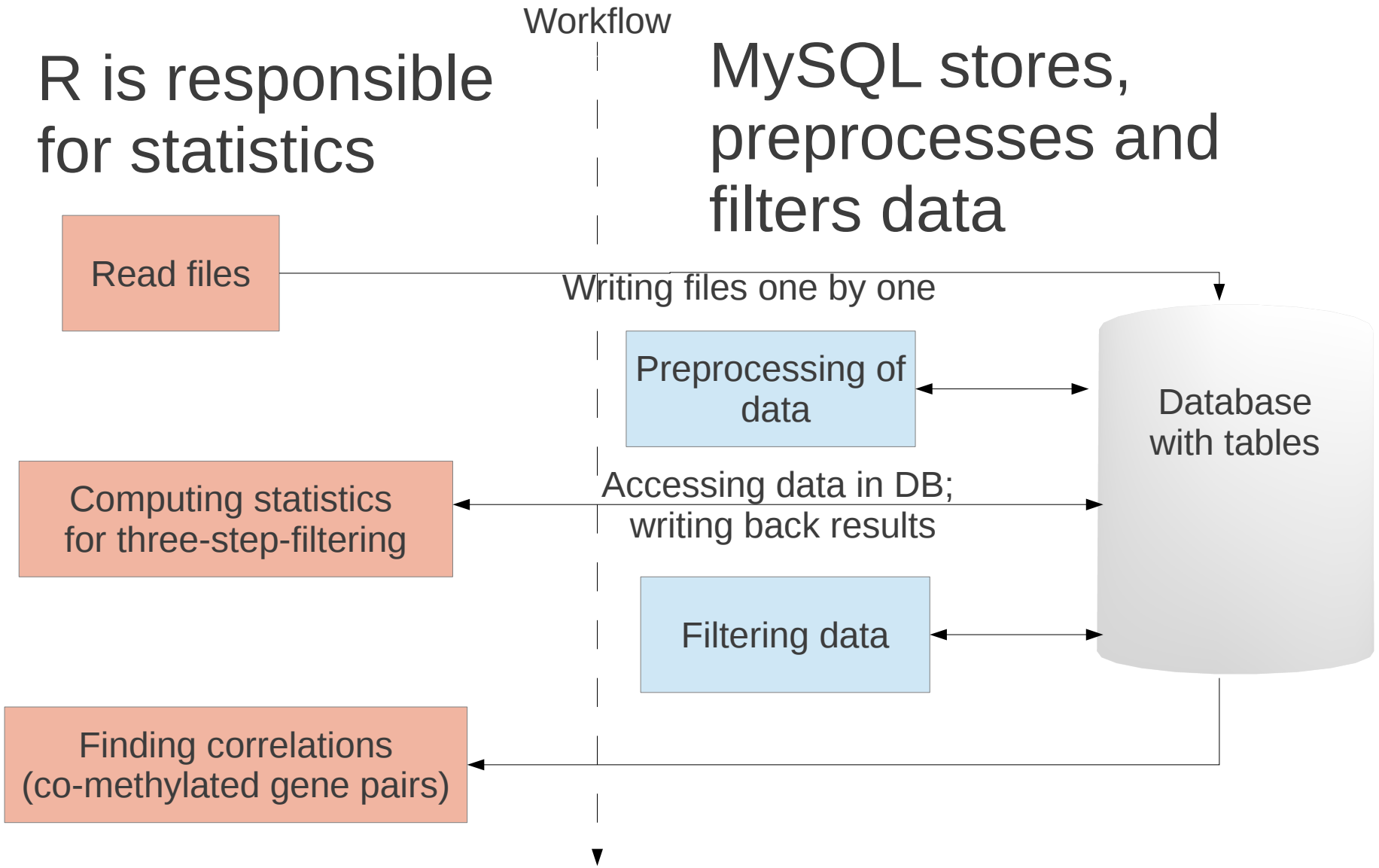
Assignment 5 - introduction

- You need R together with RMySQL package and MySQL under Ubuntu (preferred).
- Alternative solution is R + MySQL + any of the following packages: DBI, RODBC.
- Installing MySQL in Ubuntu:

  sudo apt-get install mysql-server mysql-client libmysqlclient15-dev
- Recommended IDE for MySQL is MySQL Workbench

  http://www.mysql.de/products/workbench/

Working with MySQL from R



Useful functions

- RMySQL package in R:
  - con <- **dbConnect**(dbDriver("MySQL"), dbname = "my_database", username = "my_username", password = "my_password", client.flag = CLIENT_MULTI_STATEMENTS)
  - query <- **sprintf**("SELECT * FROM my_table WHERE gene = '%s'", gene_name)
  - **dbSendQuery**(con, query)
  - **dbWriteTable**(con, 'my_table', my_data_frame, append = TRUE, row.names = FALSE)
  - **dbDisconnect**(con)
- MySQL: **INSERT, SELECT, DELETE, UPDATE, CREATE TABLE,** etc.