

# Machine Learning Genetic Signatures and Cancer Data

- „small data” modelling in oncology,
- „big data” modelling in oncology.

**Przemysław Biecek**  
**University of Warsaw**  
**Warsaw University of Technology**

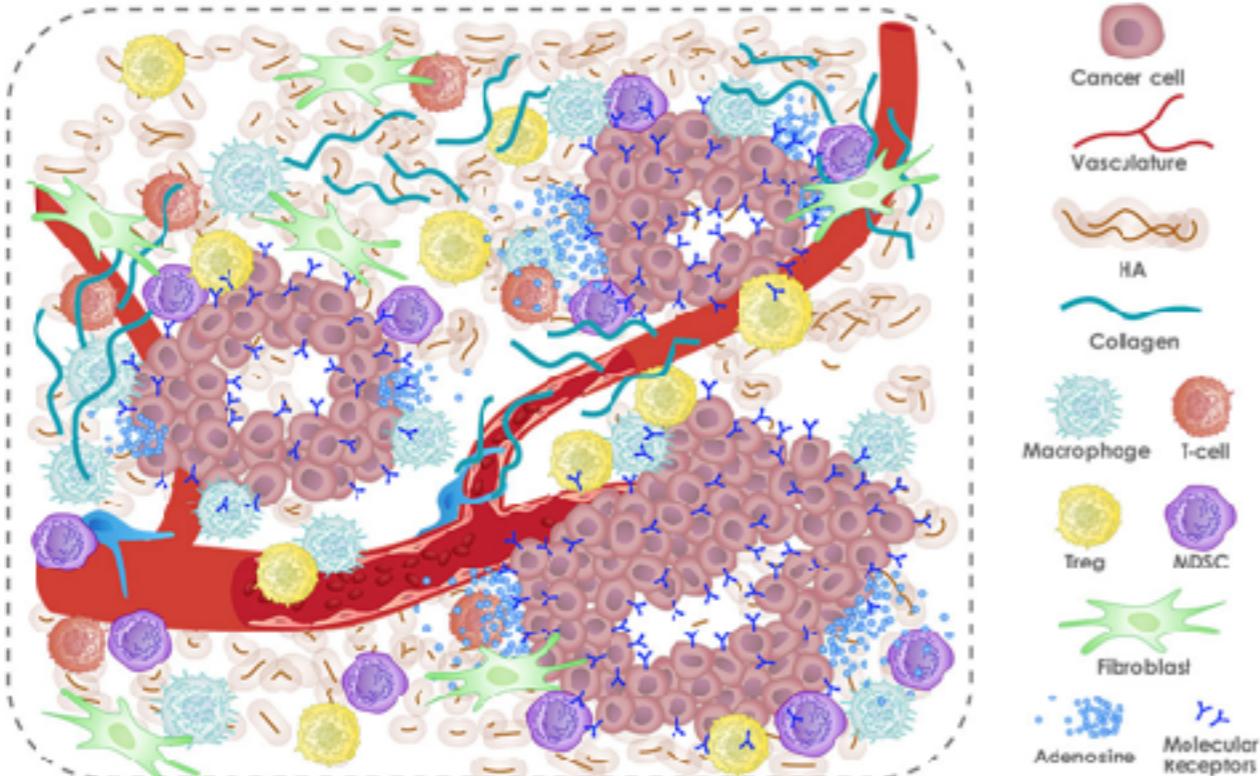


How to build a „small” predictive model?

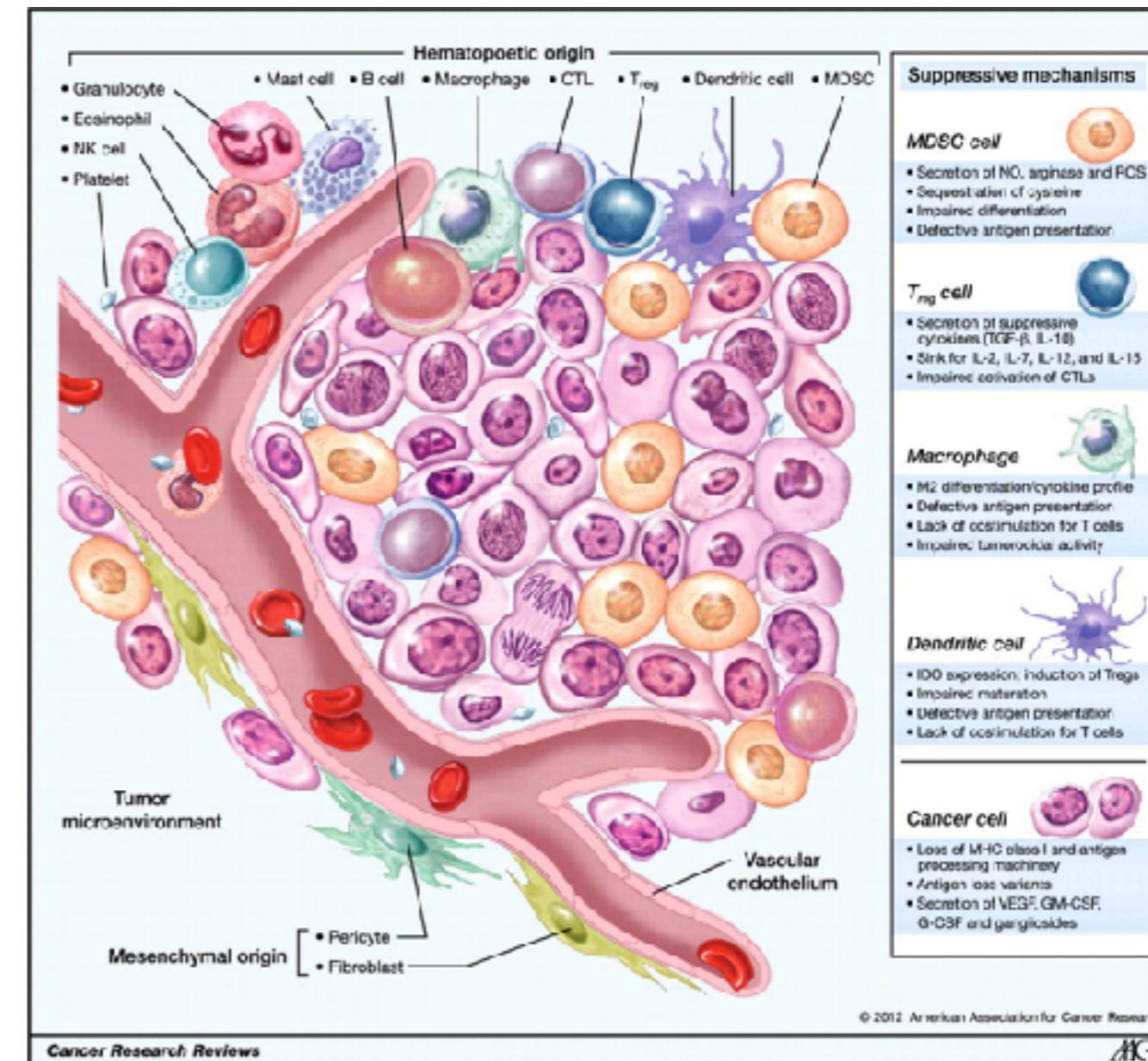
UseCase  
Collaboration with  
Lower-Silesian Oncology Centre

Cancer is not homogenous,  
microenvironment of heterogeneous cells with  
super fast evolution - makes treatment difficult

### The Tumor Microenvironment



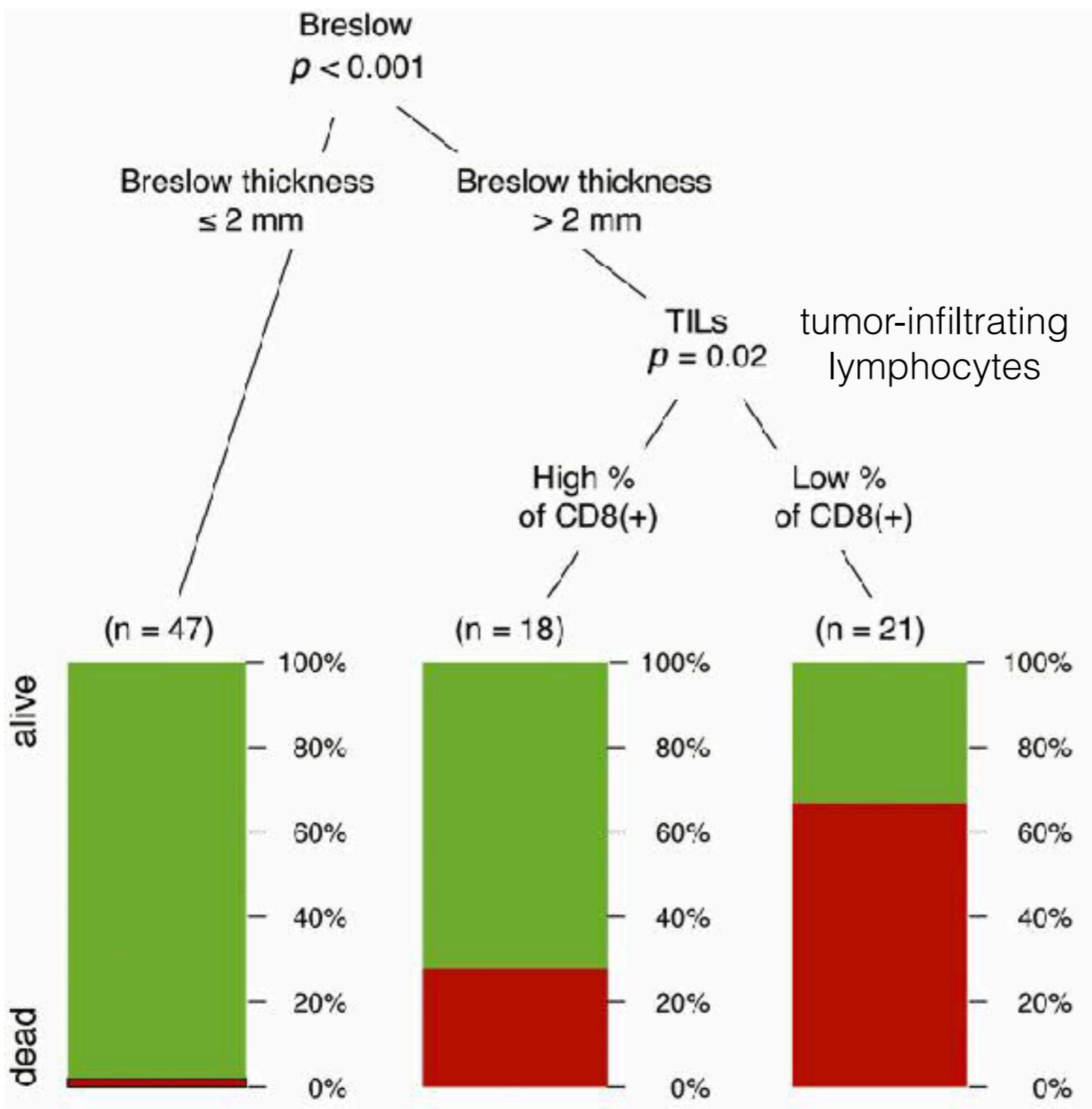
Whatcott et al.  
Clin Cancer Res. 21:15 (2015)



Cellular Constituents of Immune Escape  
within the Tumor Microenvironment,  
Cancer Research

# ,,small data” modelling - data fits into a single Excel file

	CSOS tygodnie 5 years	DFS tygodnie last_rec 5y (bez nawrctu)	nawrót lub czerniakowy zgon	nawrot	PM (obecność przeszutów odległych)	pN (przerzuty w węzłach chłonnych)	AJCC staging	SNLB_plus (z przerzutami)	regression	owrzodzenie (ULCER)	wiek w momencie diagnozy	plec	breslow w mm	Breslow (pogrupowany w 4 grupy)	clark	hist_type	mitotic rate
224	224	0	0	0	0	0	1	1	1	0	58	0	1.1	1	4	1	0
48	48	0	0	0	0	0	1	1	1	0	79	1	0.5	1	4	2	1
186	186	0	0	0	0	0	1	1	1	0	70	1	0.5	1	4	3	1
158	155	0	0	0	0	0	1	1	1	0	55	1	0.4	1	3	3	1
173	89	0	0	0	0	0	1	1	1	0	26	0	0.75	1	3	2	1
226	226	0	0	0	0	0	1	1	1	0	61	1	0.4	1	4	4	1
260	260	0	0	0	0	0	1	1	1	0	30	0	0.9	1	4	3	1
260	260	0	0	0	0	0	1	1	1	0	49	0	0.4	1	4	3	1
258	258	0	0	0	0	0	1	1	1	0	45	1	0.4	1	4	3	1
166	166	0	0	0	0	0	1	1	1	0	61	0	0.9	1	4	3	1
260	260	0	0	0	0	0	1	1	1	0	39	0	0.5	1	3	2	1
204	204	0	0	0	0	0	1	1	1	0	71	1	0.4	1	3	2	1
247	247	0	0	0	0	0	1	1	1	0	77	1	0.4	1	3	2	1
138	138	0	0	0	0	0	1	1	1	0	50	1	0.2	1	3	3	1
260	260	0	0	0	0	0	1	1	1	0	57	0	0.6	1	3	3	0
170	170	0	0	0	0	0	1	1	1	0	69	1	0.2	1	3	3	0
257	257	0	0	0	0	0	1	1	1	0	40	1	0.2	1	3	3	1
231	231	0	0	0	0	0	1	1	1	0	57	1	0.3	1	3	3	1
118	118	0	0	0	0	0	1	1	1	0	75	1	0.3	1	3	3	1
99	99	0	0	0	0	0	1	1	1	0	73	0	1.8	1	3	3	1
202	97	0	1	1	0	0	1	1	1	0	76	1	0.8	1	4	5	0
260	260	0	0	0	0	0	1	1	1	0	72	1	0.4	1	3	4	1
200	180	0	1	1	0	0	1	1	1	0	71	1	0.2	1	3	3	0
231	231	0	0	0	0	0	1	1	1	0	73	1	1.9	1	3	3	1
190	190	0	0	0	0	0	1	1	1	0	53	1	1.2	1	3	3	1
35	35	1	0	0	0	0	1	1	1	0	76	1	2.5	1	3	4	2
186	186	0	0	0	0	0	1	1	1	0	71	1	0.5	1	2	2	0
251	251	0	0	0	0	0	1	1	1	0	33	1	1.2	1	2	2	0
243	243	0	0	0	0	0	1	1	1	0	33	1	1.1	1	3	3	0
5	5	1	0	0	0	0	1	1	1	0	37	0	6	1	4	4	2
167	167	0	0	0	0	0	1	1	1	0	55	1	0.6	1	3	3	1
252	252	0	0	0	0	0	1	1	1	0	40	1	1.5	1	3	3	1

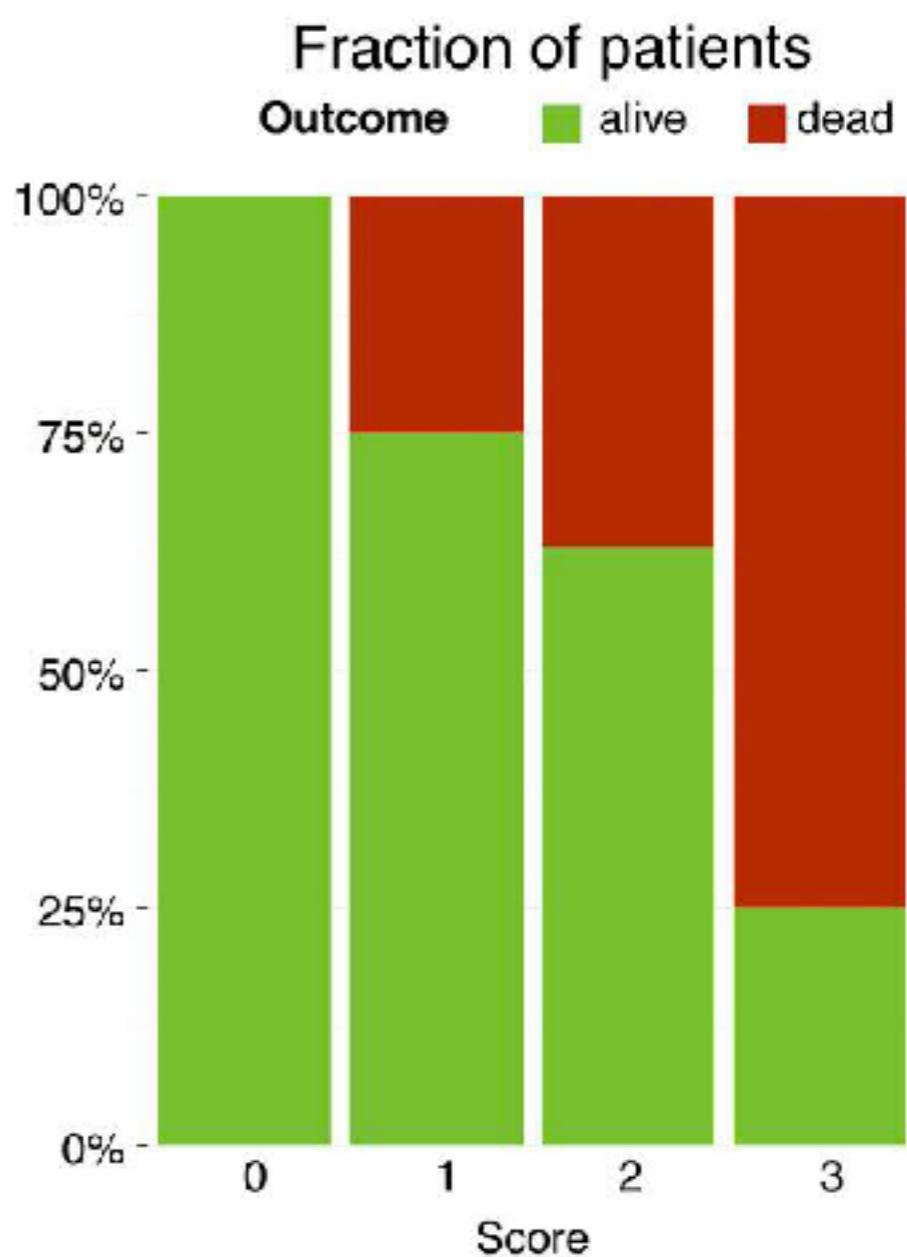


*BILLCD8 - A Multivariable Survival Model as a Simple and Clinically Useful Prognostic Tool to Identify High-risk Cutaneous Melanoma Patients.* Donizy P, Biecek P, Halon A, Matkowski R. **Anticancer Res.** 2016 Sep;36(9):4739-47.

# Regression model (Cox or logistic)

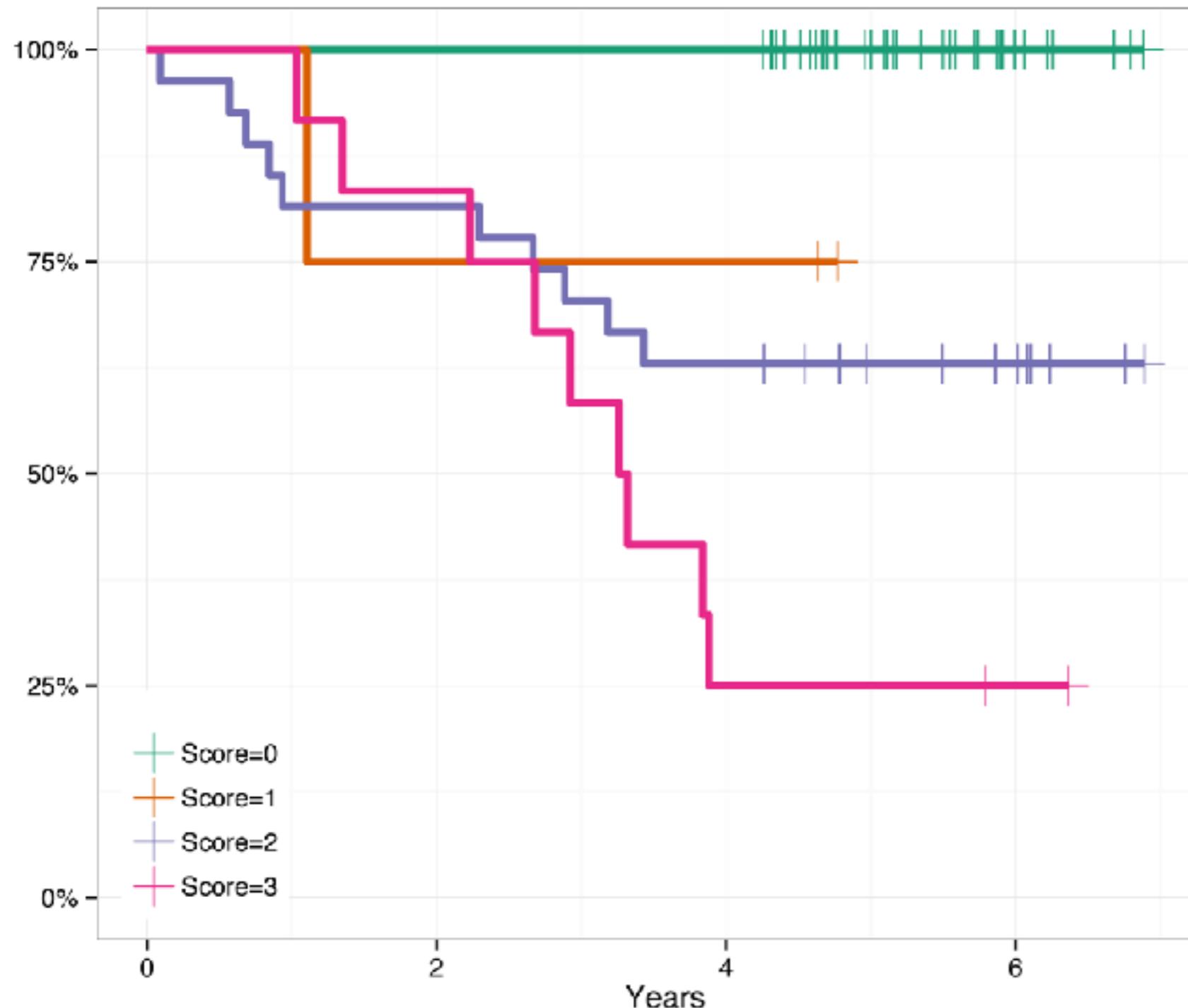
Table II. Full and refined survival model.

Pathological parameters	Estimate effect	p-Value
(Intercept) – full model	-2.27	0.12
Breslow: $\geq 2$ mm	2.7	0.043*
TILs: non-brisk, brisk	-3.07	0.0041**
Mitotic rate: $\geq 1/\text{mm}^2$	-1.72	0.24
LYI: present	1.38	0.074
Ulceration: present	1.85	0.11
Clark: IV, V	1.45	0.11
(Intercept) – refined model	-2.22	0.05*
Breslow: $\geq 2$ mm	3.61	0.00089***
TILs: non-brisk, brisk	-2.02	0.0088**



BILLCD8 - A Multivariable Survival Model as a Simple and Clinically Useful Prognostic Tool to Identify High-risk Cutaneous Melanoma Patients. Donizy P, Biecek P, Halon A, Matkowski R. **Anticancer Res.** 2016 Sep;36(9):4739-47.

# Validation with a new data set



*BILLCD8 - A Multivariable Survival Model as a Simple and Clinically Useful Prognostic Tool to Identify High-risk Cutaneous Melanoma Patients.* Donizy P, Biecek P, Halon A, Matkowski R. **Anticancer Res.** 2016 Sep;36(9):4739-47.

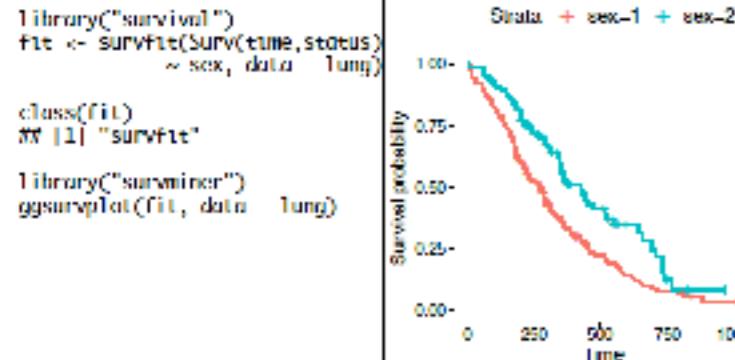
Tools that facilitate the understanding of a model

# Creating Survival Plots

## Informative and Elegant with survminer

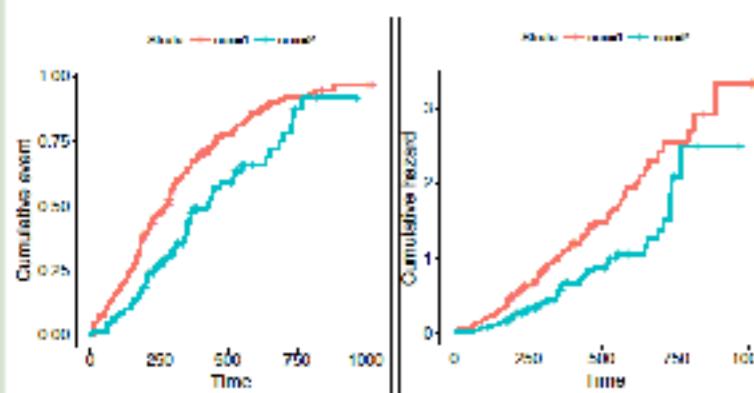
### Survival Curves

The `ggsurvplot()` function creates ggplot2 plots from `survfit` objects.



Use the `fun` argument to set the transformation of the survival curve. E.g. "event" for cumulative events, "cumhaz" for the cumulative hazard function or "pct" for survival probability in percentage.

```
ggsurvplot(fit, data = Lung, fun = "event")
ggsurvplot(fit, data = Lung, fun = "cumhaz")
```



With lots of graphical parameters you have full control over look and feel of the survival plots; position and content of the legend; additional annotations like p-value, title, subtitle.

```
ggsurvplot(fit, data = Lung,
conf.int = TRUE,
pval = TRUE,
fun = "pct",
risk.table = TRUE,
size = 1,
linetype = "strata",
palette = c("#F781BF",
            "#2E9FDF"),
legend = "bottom",
legend.title = "Sex",
legend.labs = c("Male",
              "Female"))
```

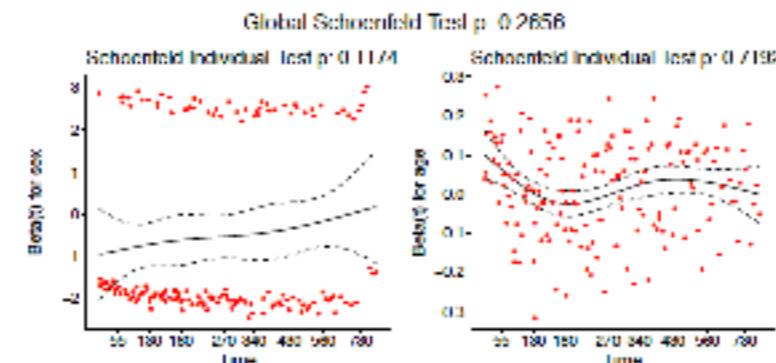
### Diagnostics of Cox Model

The function `cox.zph()` from `survival` package may be used to test the proportional hazards assumption for a Cox regression model fit. The graphical verification of this assumption may be performed with the function `ggcoxzph()` from the `survminer` package. For each covariate it produces plots with scaled Schoenfeld residuals against the time.

```
library("survival")
fit <- coxph(Surv(time, status) ~ sex + age, data = Lung)
ftest <- cox.zph(fit)
ftest
```

	rho	chisq	p
sex	0.1236	7.457	0.117
age	0.0725	0.129	0.719
GLOBAL	NA	2.651	0.266

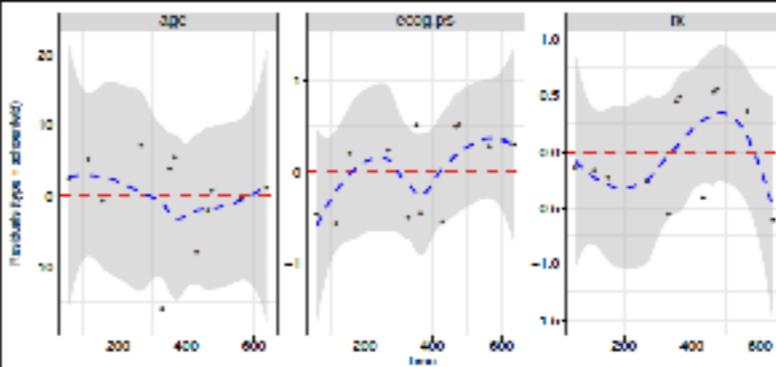
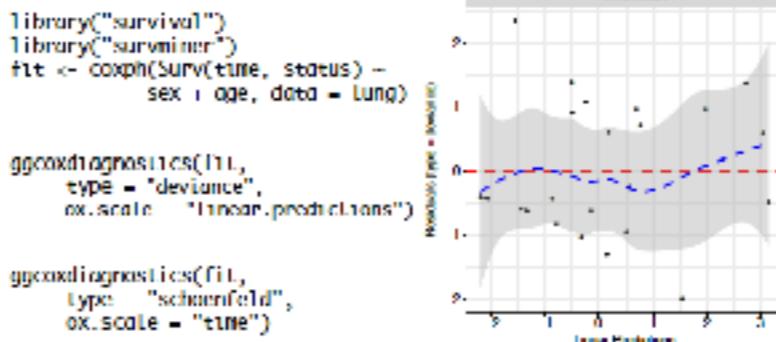
```
library("survminer")
ggcoxzph(ftest)
```



The function `ggcoxdiagnostics()` plots different types of residuals as a function of time, linear predictor or observation id. The type of residual is selected with `type` argument. Possible values are "martingale", "deviance", "score", "schoenfeld", "dlbeta", "dlnbeta", and "scaledsch".

The `ox.scale` argument defines what shall be plotted on the OX axis. Possible values are "linear.predictions", "observation.id", "time".

Logical arguments `hline` and `sline` may be used to add horizontal line or smooth line to the plot.

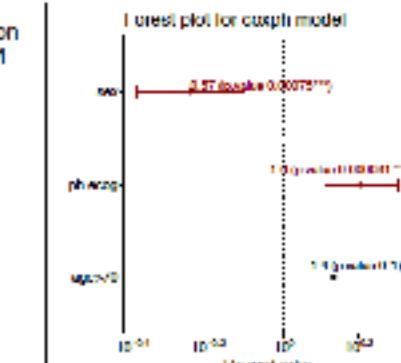


### Summary of Cox Model

The function `ggforest()` from the `survminer` package creates a forest plot for a Cox regression model fit. Hazard ratio estimates along with confidence intervals and p values are plotted for each variable.

```
library("survival")
library("survminer")
Lung$age <- ifelse(Lung$age > 70, ">70", "<=70")
fit <- coxph(Surv(time, status) ~ sex + ph.ecog + age, data = Lung)
fit
```

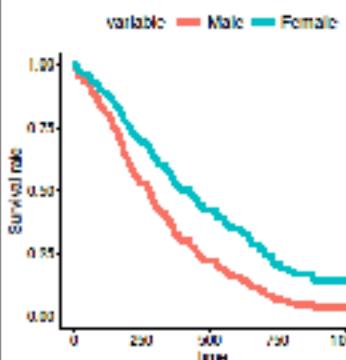
```
## Call:
## coxph(formula = Surv(time, status) ~ sex + ph.ecog + age, data=Lung)
##
##          coef exp(coef) se(coef)    z      p
## sex     -0.567    0.567   0.168 -3.37 0.00075
## ph.ecog  0.470    1.600   0.113  4.16 3.1e-05
## age>70  0.307    1.359   0.187  1.64 0.10175
##
## Likelihood ratio test=31.6 on
## n = 227, number of events= 164
ggforest(fit)
```



The function `ggcoxadjustedcurves()` from the `survminer` package plots Adjusted Survival Curves for Cox Proportional Hazards Model. Adjusted Survival Curves show how a selected factor influences survival estimated from a Cox model.

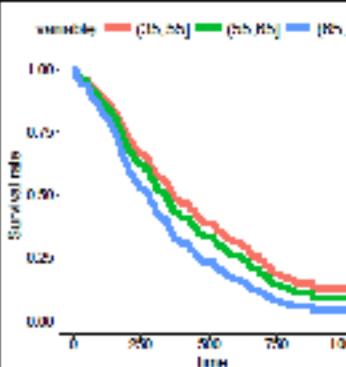
Note that these curves differ from Kaplan Meier estimates since they present expected survival based on given Cox model.

```
library("survival")
library("survminer")
Lung$sex <- ifelse(Lung$sex == 1,
                    "Male", "Female")
fit <- coxph(Surv(time, status) ~ sex + ph.ecog + age,
             data = Lung)
ggcoxadjustedcurves(fit, data = Lung,
                     variable = Lung$sex)
```



Note that it is not necessary to include the grouping factor in the Cox model. Survival curves are estimated from Cox model for each group defined by the factor independently.

```
Lung$age3 <- cut(Lung$age,
                   c(35, 55, 65, 85))
ggcoxadjustedcurves(fit, data = Lung,
                     variable = Lung$age3)
```



# factorMerger: an extended hierarchical post-hoc testing

Project Lead:  
Agnieszka Sitko

**FactorMerger - set of tools to support results from post hoc testing**

### Level fusing plot

The top-left plot shows level fusing paths (merging paths). With arguments **family=**, **show=**, **fuse=**, **spacing=**, one can select how to merge factors and what shall be presented on OX/OY axes.

Argument	Summary
panel = "all"	All panels
panel = "left"	Only left two panels
panel = "top"	Only top two panels
panel = "merging"	Merging path plot

### Group summaries

The top-right panel shows group characteristics. Use the parameter **summary=** to select the most suitable presentation.

Argument	Summary
summary = "heatmap"	For m gaussian
summary = "profile"	For m gaussian
summary = "boxplot"	For gaussian
summary = "means"	For gaussian
summary = "survival"	For Cox regression
summary = "proportions"	For binomial regression

**Merging path plot**

Optimal G/C partition: (I:(H)(E)(G)(D)(C):(B)(A)(F))(J)

**Model statistics**

Group success proportion  
Success: 1 (green), failure: 0 (violet)

**heat map**

Heatmap  
Group intensity variables

**profile**

Profile plot  
Variable means ratio

**box plot**

Boxplot  
Summary statistic: mean

**means**

Summary statistics: mean  
Means and standard deviations of coefficient estimates

**survival**

Survival plot  
Adjusted survival curves for single model

**proportions**

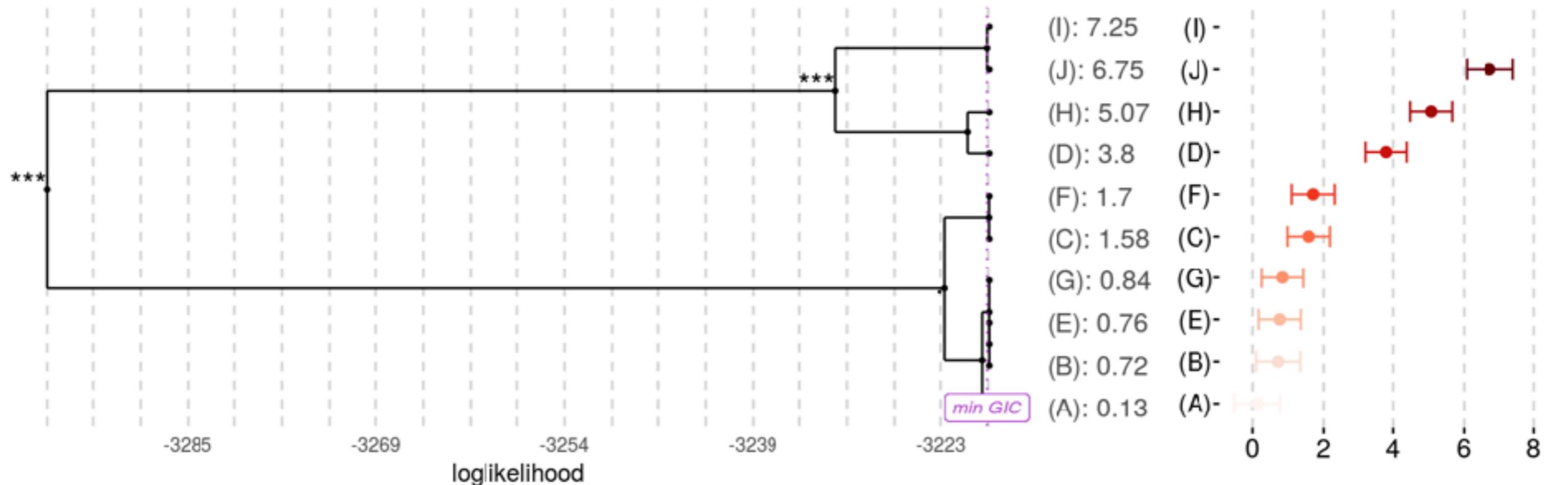
Group success proportion  
Success: 1 (green), failure: 0 (violet)

# FactorMerger: an extended hierarchical post-hoc testing

Project Lead:  
Agnieszka Sitko

## Merging path plot

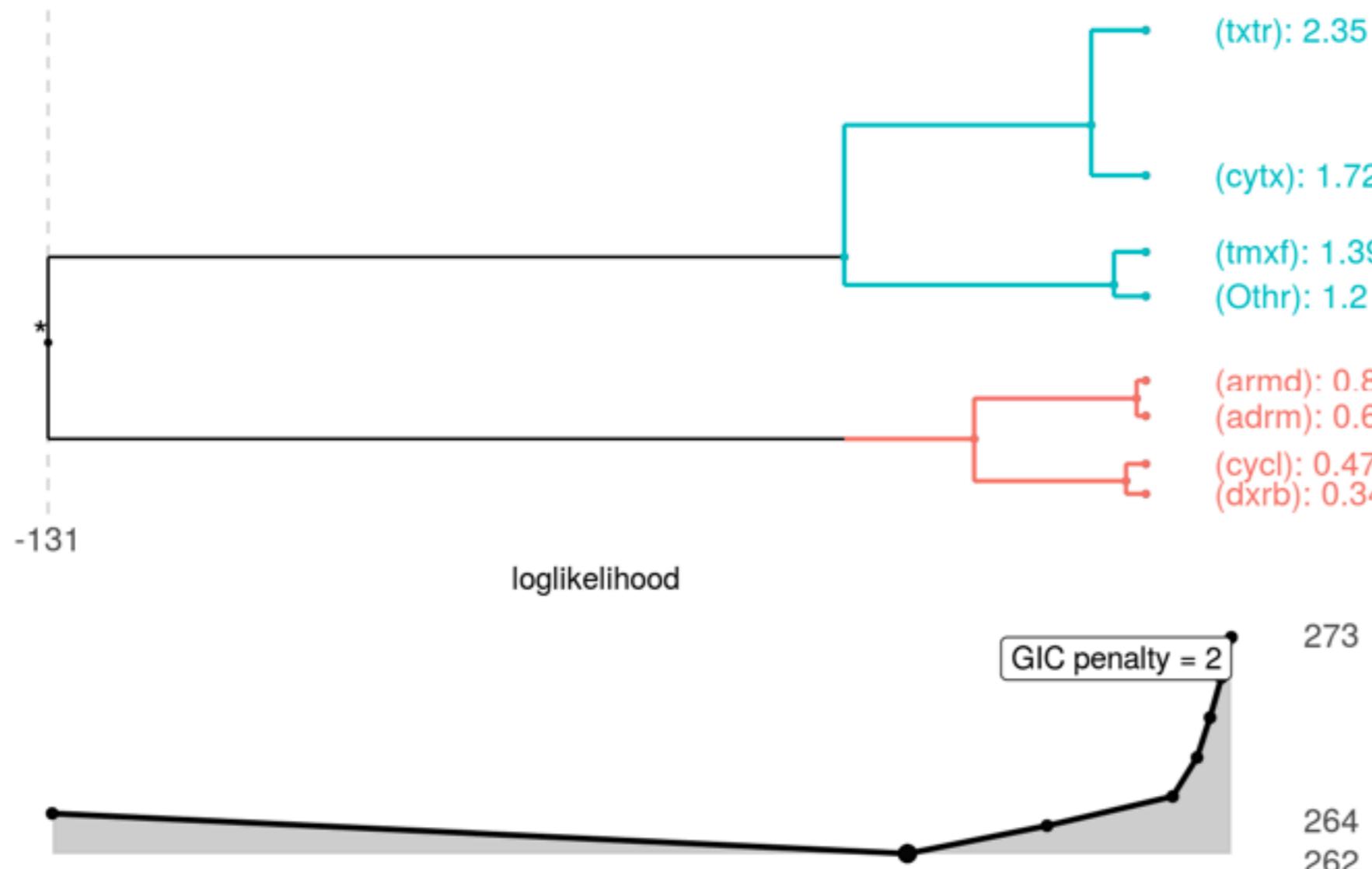
Optimal GIC partition: (A)(B)(E)(G):(C)(F):(D):(H):(J):(I)



# FactorMerger: an extended hierarchical post-hoc testing

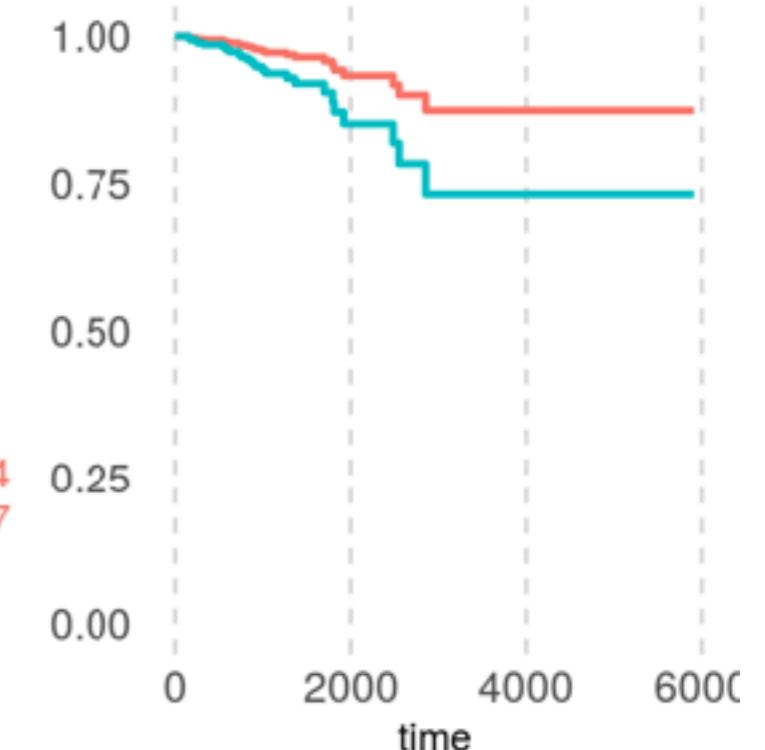
Project Lead:  
Agnieszka Sitko

Factor Merger Tree



Survival plot

Adjusted survival curves for co:



Tools that facilitate the validation of a model

BILLCD8 - A Multivariable Survival Model as a Simple and Clinically Useful Prognostic Tool to Identify High-risk Cutaneous Melanoma Patients. Donizy P, Biecek P, Halon A, Matkowski R. **Anticancer Res.** 2016 Sep; 36(9):4739-47.

Ductal carcinoma *in situ* on stereotactic biopsy of suspicious breast microcalcifications: Expression of SPARC (Secreted Protein, Acidic and Rich in Cysteine) can predict postoperative invasion. Szynglarewicz B, Kasprzak P, Donizy P, Biecek P, Halon A, Matkowski R. **J Surg Oncol.** 2016 Oct;114(5):548-556.

Golgi-Related Proteins GOLPH2 (GP73/GOLM1) and GOLPH3 (GOPP1/MIDAS) in Cutaneous Melanoma: Patterns of Expression and Prognostic Significance. Donizy P, Kaczorowski M, Biecek P, Halon A, Szkudlarek T, Matkowski R. **Int J Mol Sci.** 2016 Oct 1;17(10).

Screen-detected ductal carcinoma *in situ* found on stereotactic vacuum-assisted biopsy of suspicious microcalcifications without mass: radiological-histological correlation. Szynglarewicz B, Kasprzak P, Biecek P, Halon A, Matkowski R. **Radiol Oncol.** 2016 Apr 23;50(2):145-52.

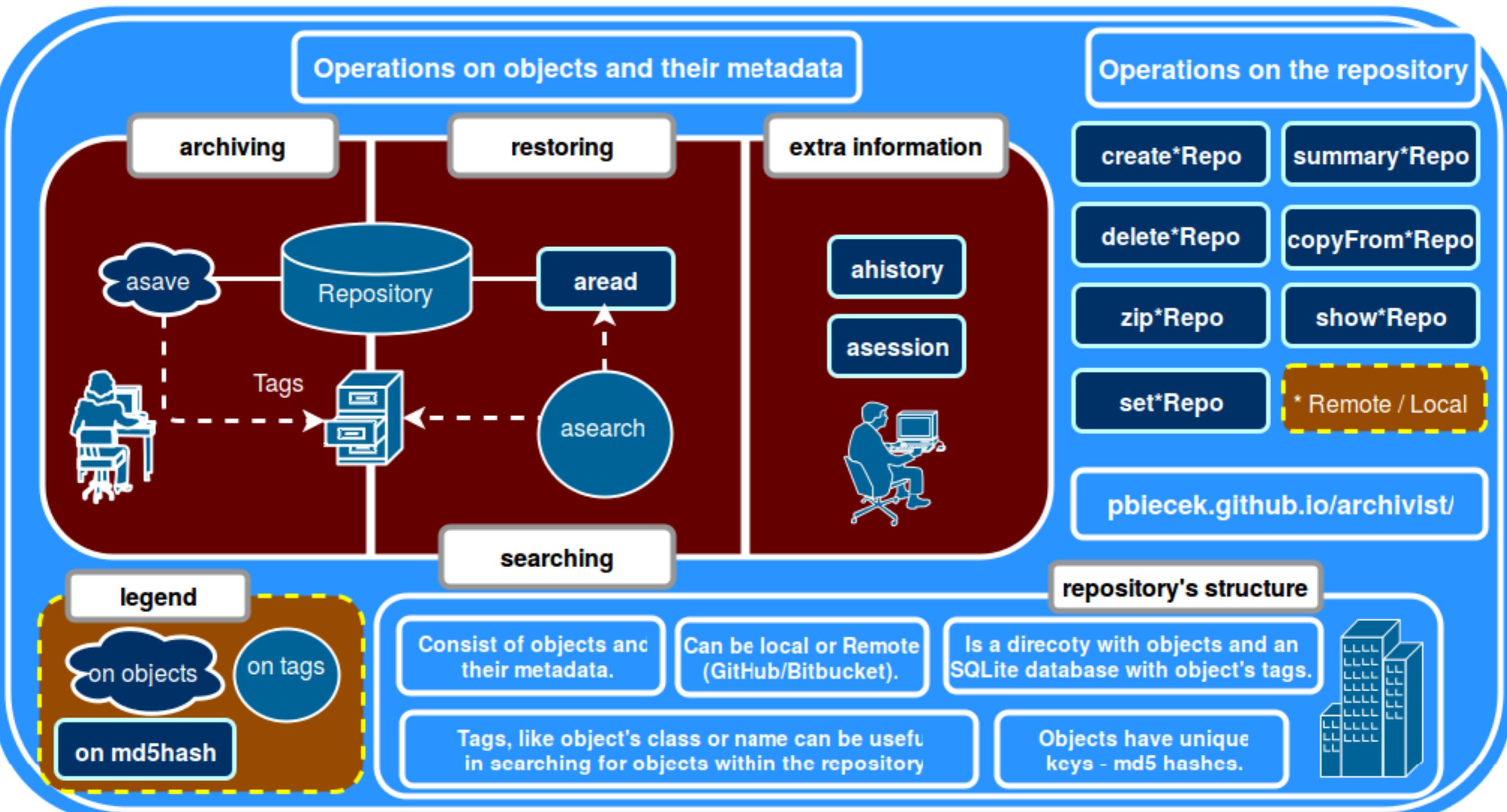
HER-2 expression in immunohistochemistry has no prognostic significance in gastric cancer patients. Halon A, Donizy P, Biecek P, Rudno-Rudzinska J, Kielan W, Matkowski R. **ScientificWorldJournal.** 2012;2012:941259.

Prognostic role of c-met expression in breast cancer patients. Gisterek I, Lata E, Halon A, Matkowski R, Szelachowska J, Biecek P, Kornafel J. **Rep Pract Oncol Radiother.** 2011 Jun 8;16(5):173-7.

Correlation between hepatocyte growth factor receptor and vascular endothelial growth factor-A in breast carcinoma. Gisterek I, Matkowski R, Suder E, Hałoń A, Szelachowska J, Łata E, Łacko A, Biecek P, Kornafel J. **Folia Histochem Cytobiol.** 2010 Jan 1;48(1):78-83.

Serum vascular endothelial growth factors a, C and d in human breast tumors. Gisterek I, Matkowski R, Lacko A, Sedlaczek P, Szewczyk K, Biecek P, Halon A, Staszek U, Szelachowska J, Pudelko M, Bebenek M, Harlozinska-Szmyrka A, Kornafel J. **Pathol Oncol Res.** 2010 Sep;16(3):337-44.

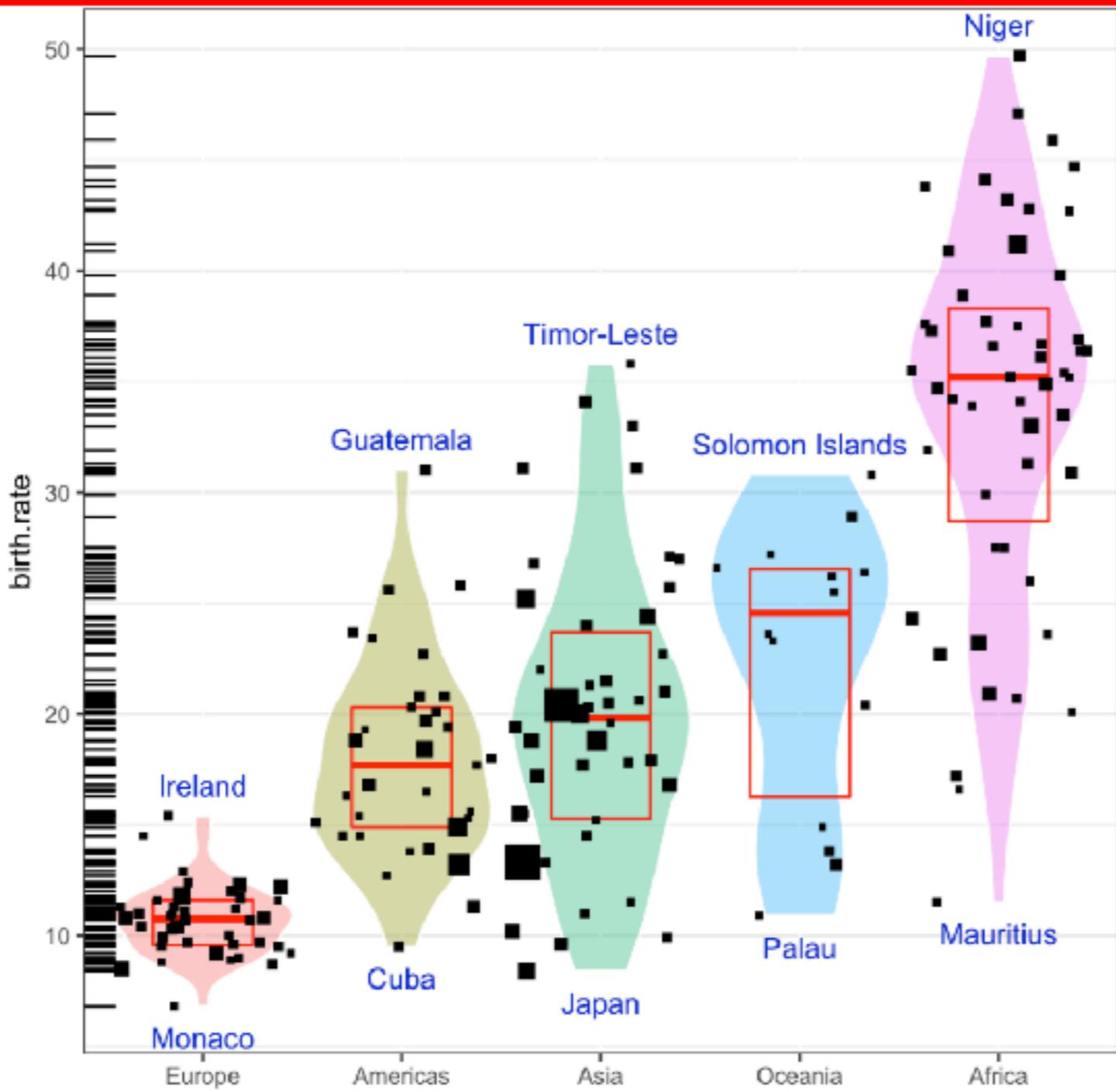
# archivist - reproducible and recordable research



<https://github.com/pbiecek/archivist>

```
ggplot(countries, aes(x=continent, y=birth.rate, label=country)) +  
  geom_violin(scale="width", aes(fill=continent), color="white", alpha=0.4) +  
  stat_summary(fun.data = "q3", geom = "crossbar",  
    colour = "red", width = 0.5) +  
  geom_jitter(aes(size=(population)^0.9), position=position_jitter(width = .45, height = 0),  
    shape=15) +  
  geom_rug(sides = "1") +  
  geom_text(data=countriesMin, vjust=2, color="blue3") +  
  geom_text(data=countriesMax, vjust=-1, color="blue3") +  
  theme_bw() + xlab("") + theme(legend.position="none", panel.grid.major.x = element_line(color="white"))
```

Load: archivist::aread('pbiecek/Eseje/arepo/ba7f58faf7373420e3ddce039558140')



# How to build a „large” predictive model?

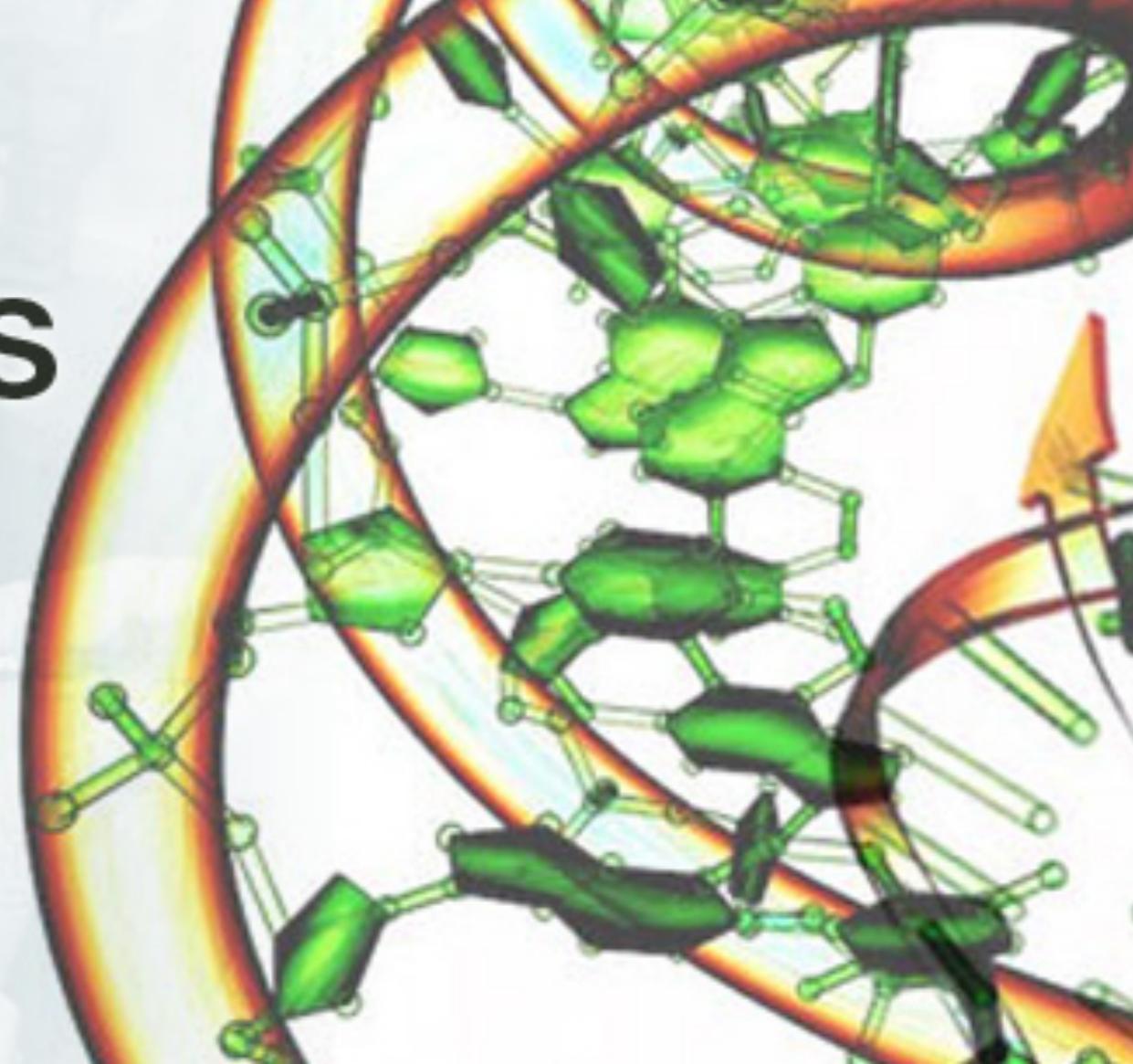
UseCase  
Collaboration with  
Greater Poland Cancer Centre

The Cancer Genome Atlas

# The Cancer Genome Atlas



*Understanding  
genomics  
to improve  
cancer care*

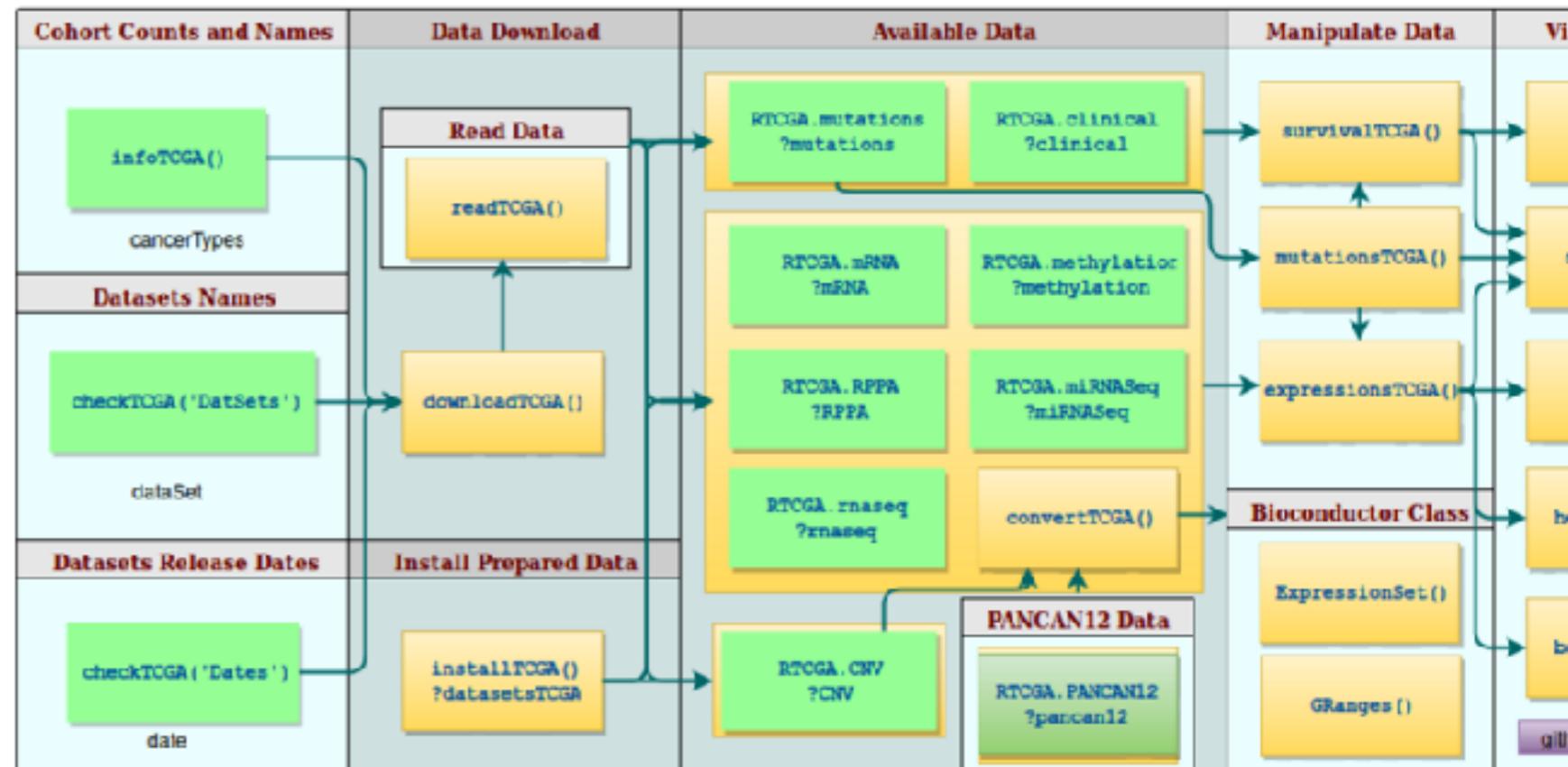


## The Cancer Genome Atlas (TCGA)

Multi-dimensional maps of the key genomic changes in **33 types of cancer**. **2.5 petabytes of data** describing tumour tissue and matched normal tissues from more than **11,000 patients** is **publicly available**. The data have contributed to more than a thousand studies of cancer by independent researchers.

Source: <http://cancergenome.nih.gov/abouttcga/overview>

# Workflow of RTCGA package



# RTCGA

Factory of R packages  
with TCGA data

Home

Web Application

Cox Survival A

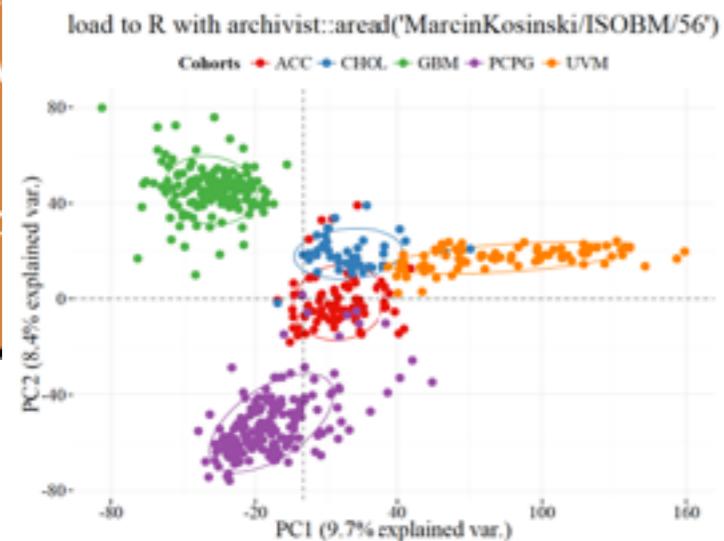
TCGA Data Dow

Visualizations

GitHub project

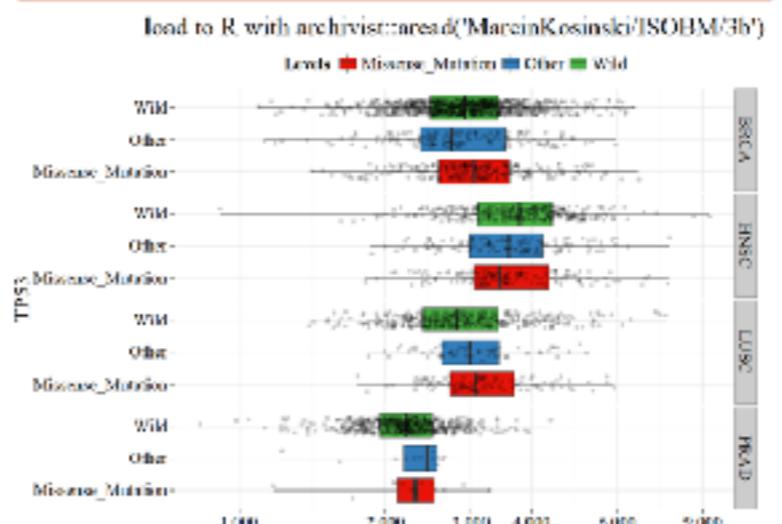
This can be reproduced in R with

```
library(RTCGA.rnaseq)
rnaseqBiplot(c("ACC", "CHOL", "GBM", "PCPG", "UVM"))
```



This can be reproduced in R with

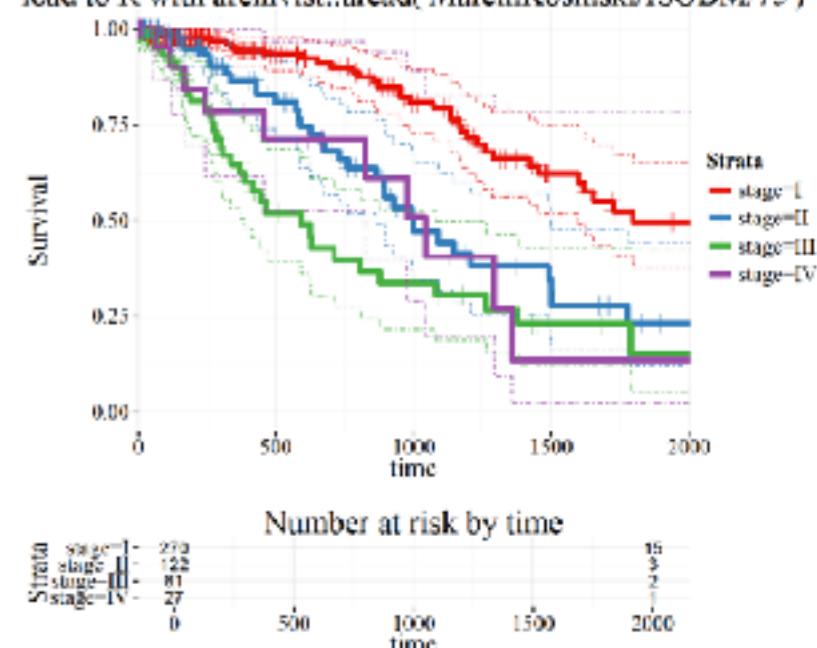
```
library(RTCGA.mutations);library(RTCGA.rnaseq)
mutationsBox(c('BRCA', 'HNSC', 'LUSC', 'PRAD'), 'TP53', 'ETFE1')
```



This can be reproduced in R with

```
library(RTCGA.clinical)
clinicalStageSurvival(LUAD.clinical, xlims = c(0,2000))
```

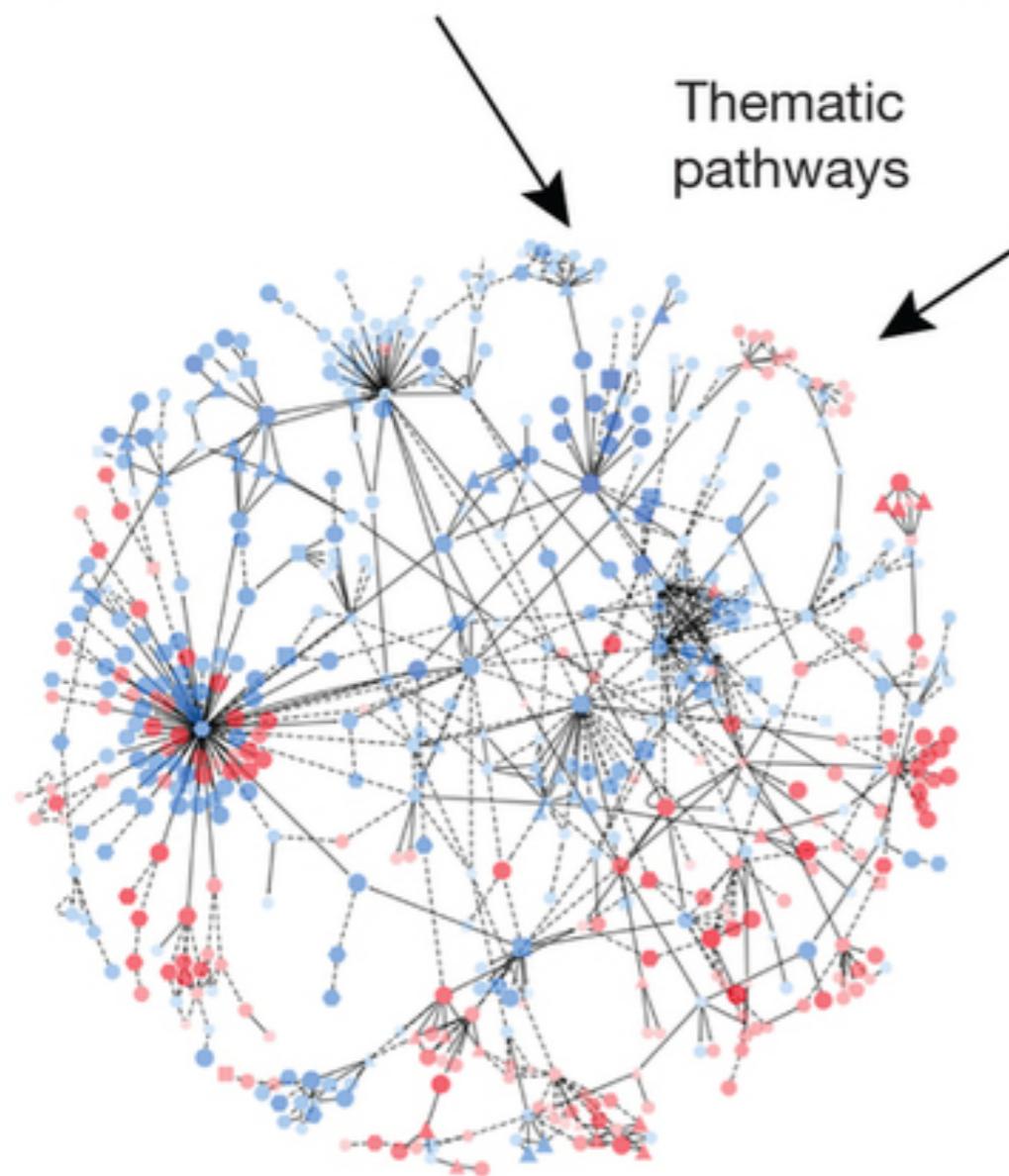
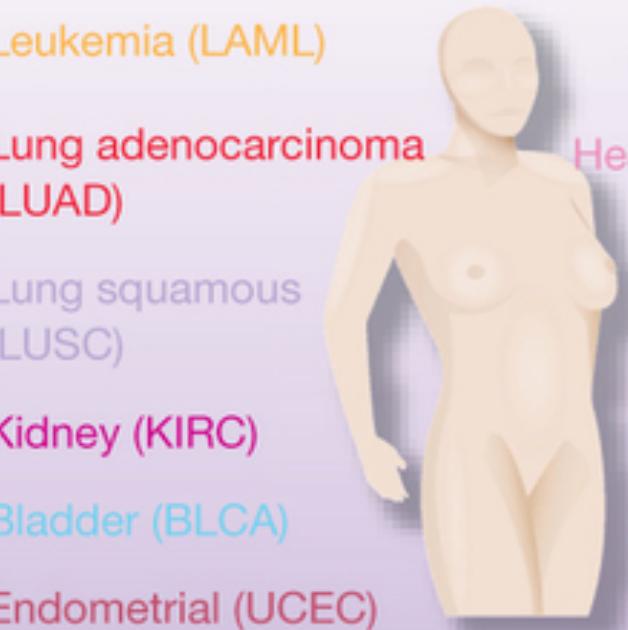
load to R with archivist::aread('MarcinKosiński/ISOBM/75')



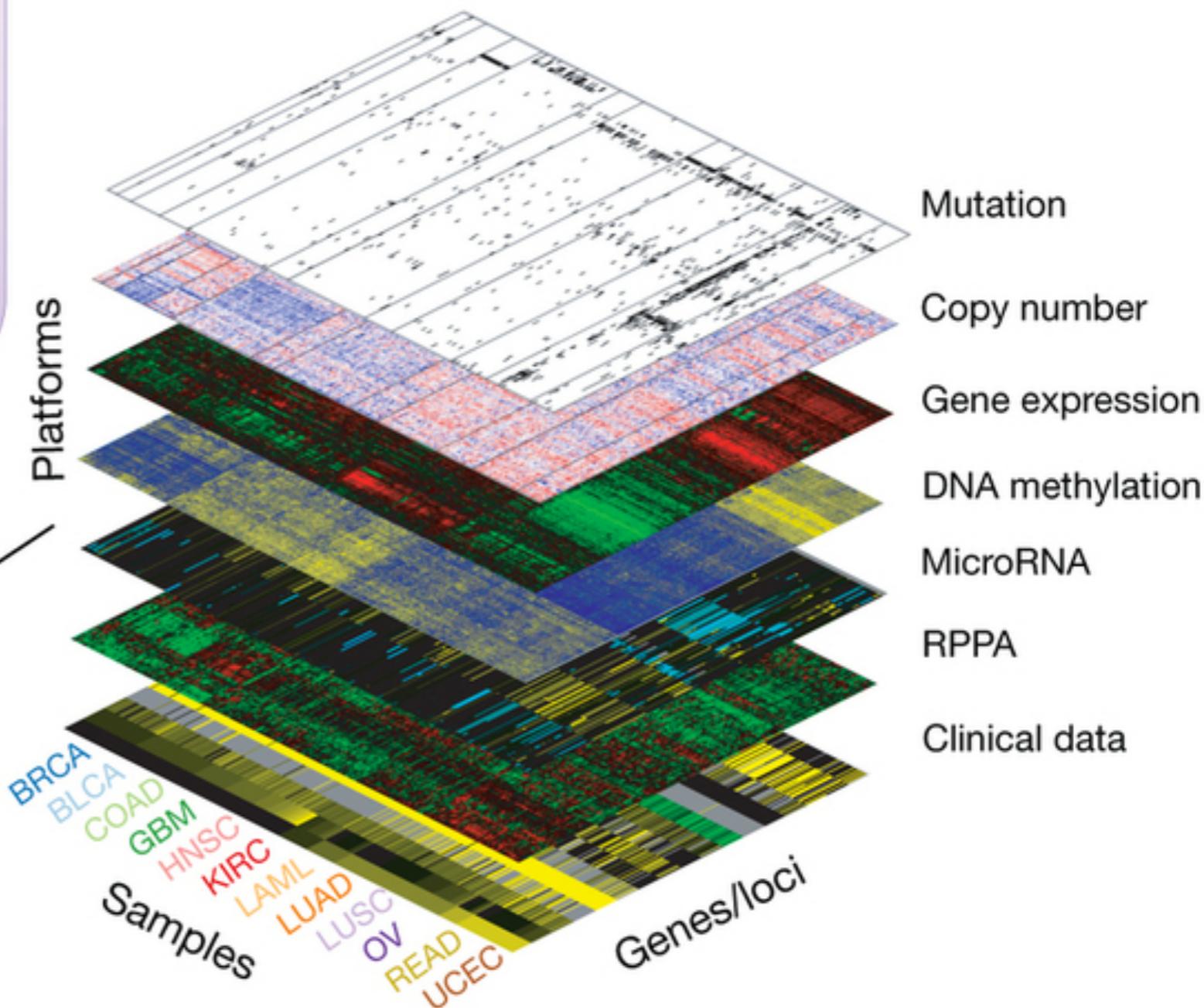
Maintainer: Marcin Kosiński

Leukemia (LAML)  
Lung adenocarcinoma (LUAD)  
Lung squamous (LUSC)  
Kidney (KIRC)  
Bladder (BLCA)  
Endometrial (UCEC)

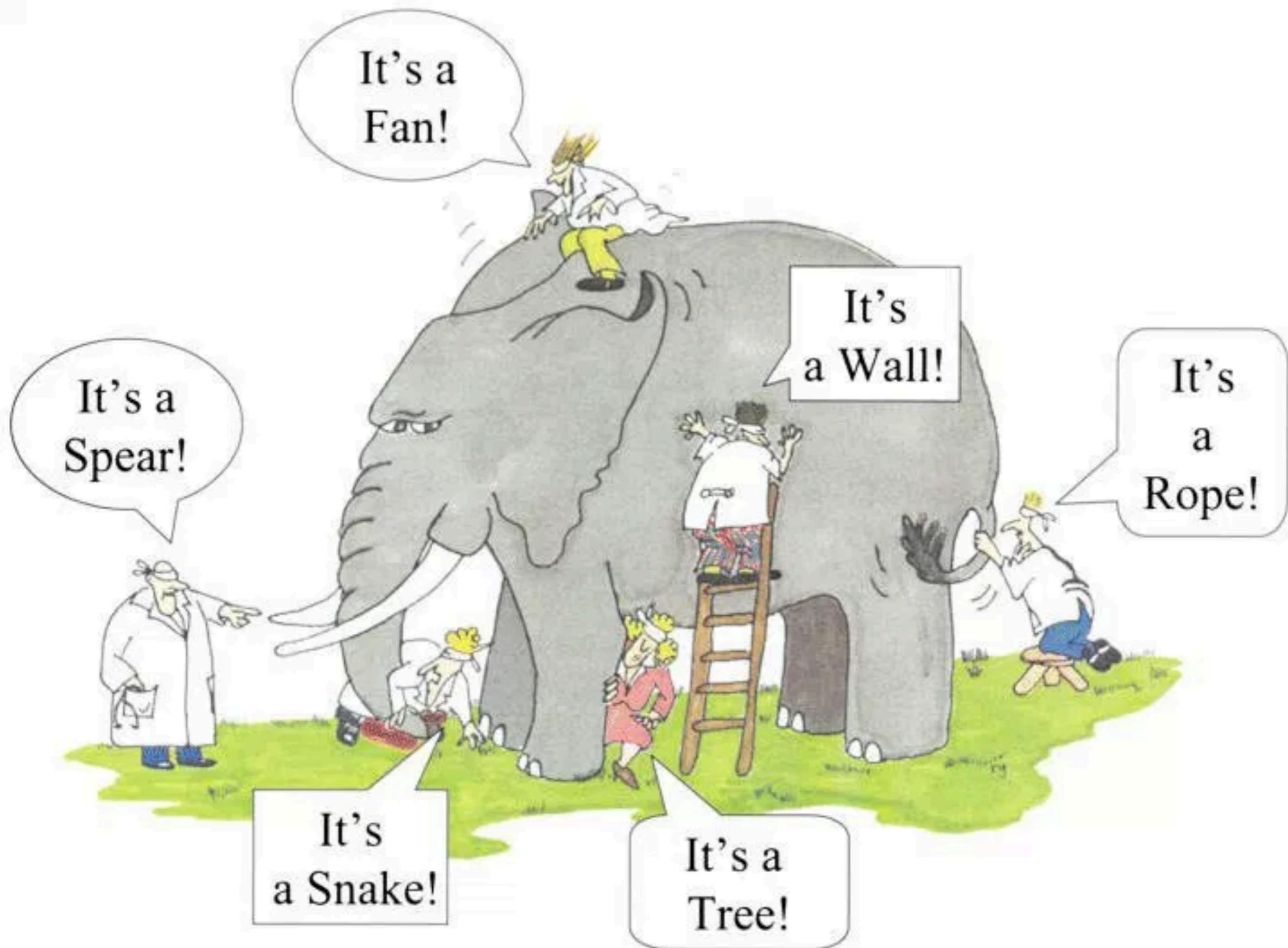
Glioblastoma (GBM)  
Head and neck (HNSC)  
Breast (BRCA)  
Ovarian (OV)  
Colon (COAD)  
Rectum (READ)



## Omics characterizations

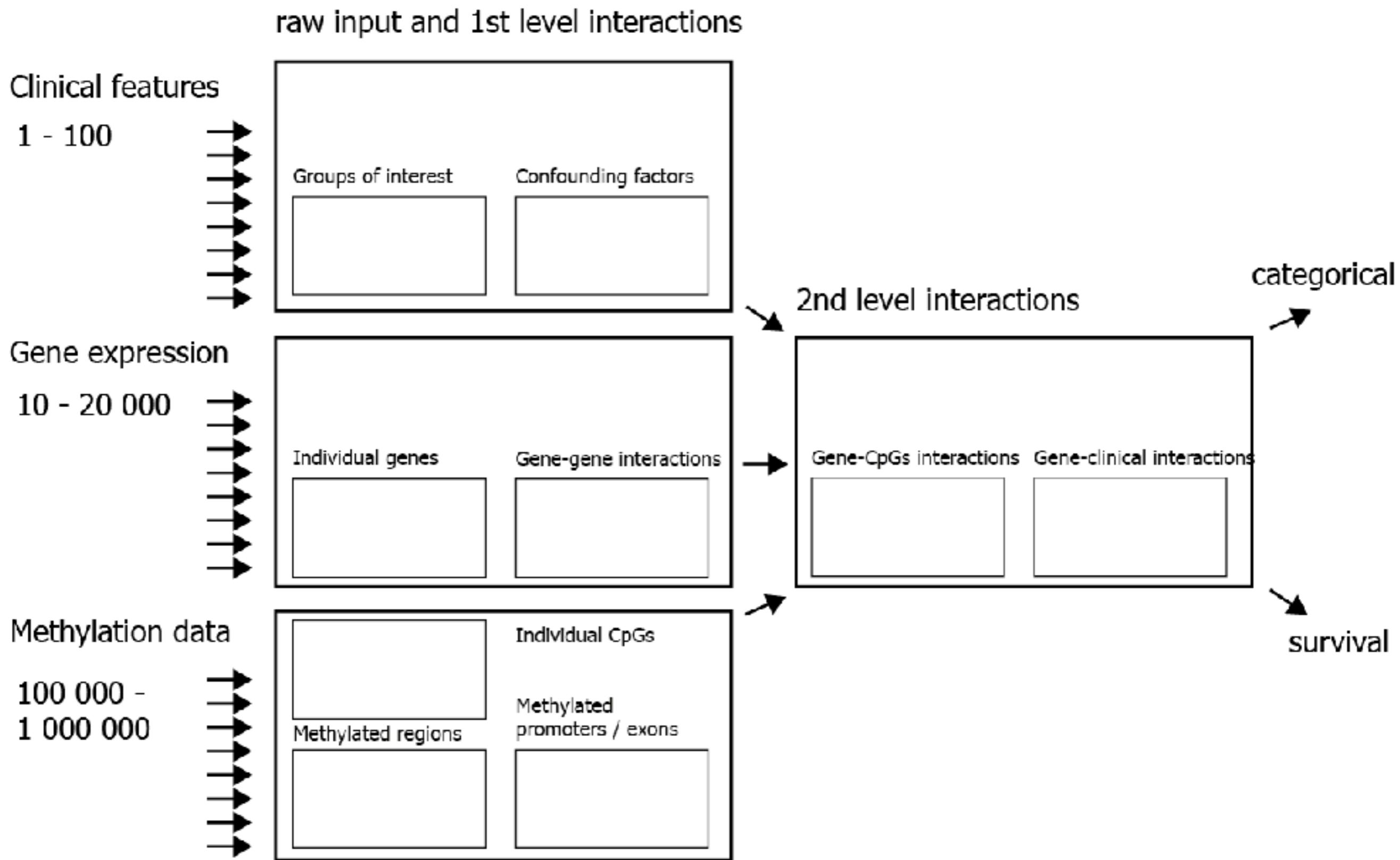


Kyle Chang et al, The Cancer Genome Atlas Pan-Cancer analysis project, Article in Nature Genetics  
45(10):1113-20

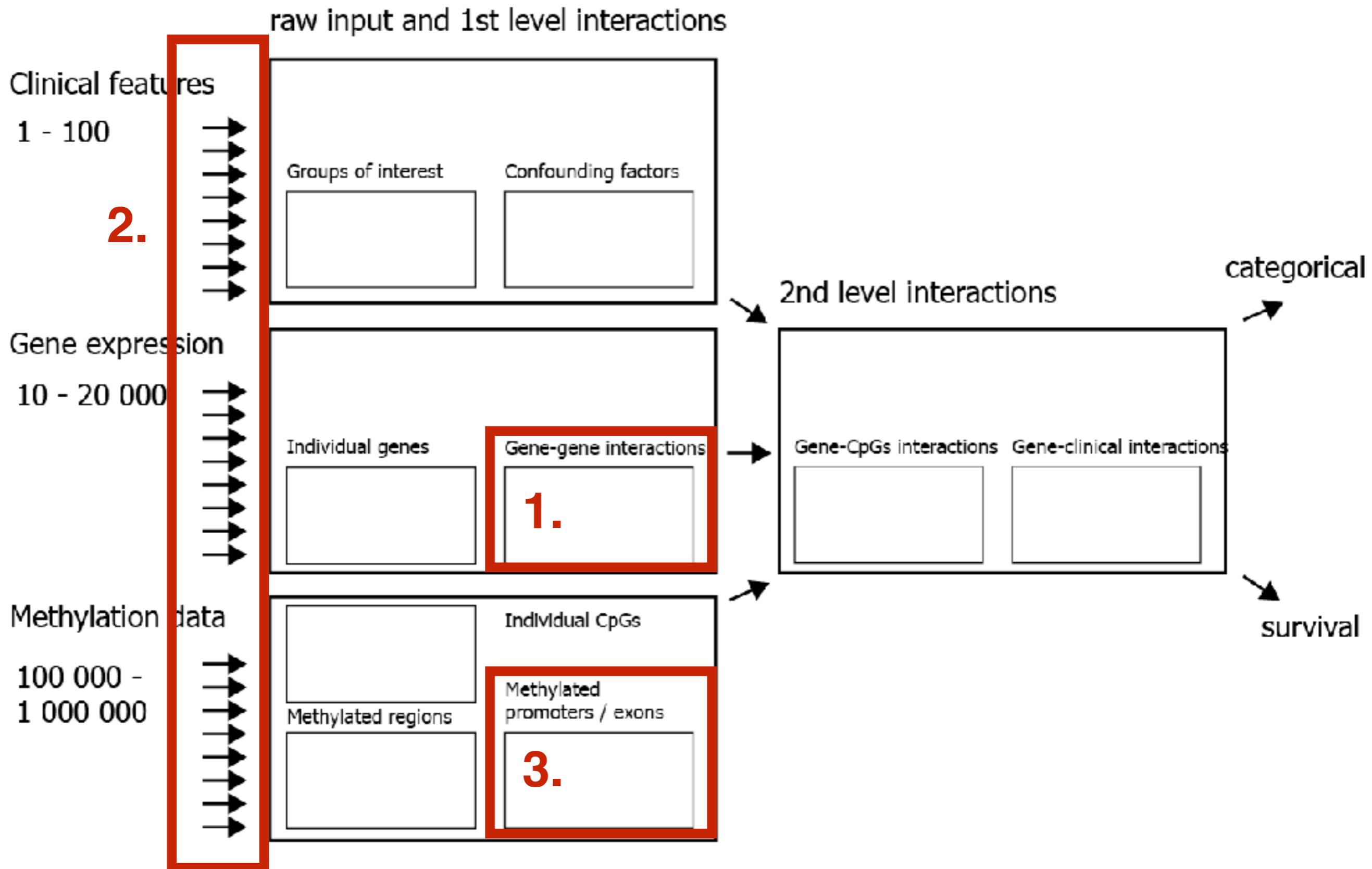


<http://www.philipchircop.com/post/25783275888/seeing-the-full-elephant-its-a-tree-its-a>

# Learning with Structure: MLGenSig (Machine Learning Genetic Signatures)



# Learning with Structure: MLGenSig (Machine Learning Genetic Signatures)



## Problem 1

Gene2Vec = word2vec for genes

- 20 000 genes -> 200 000 000 pairwise interactions
- We cannot test 200 millions of hypotheses !!!
- How to choose candidates for interacting genes?

## Problem 1

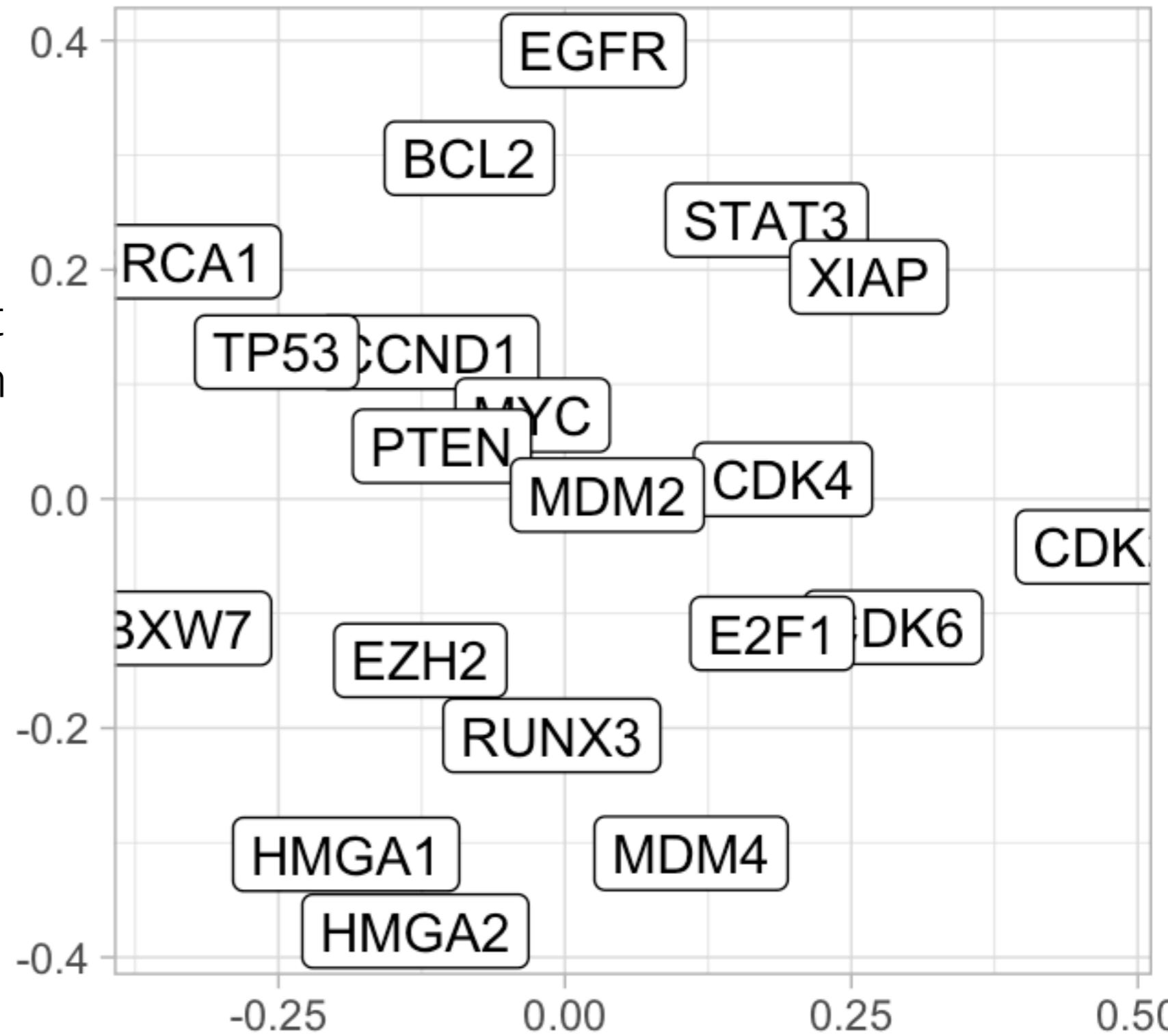
Gene2Vec = word2vec for genes

- 20 000 genes -> 200 000 000 pairwise interactions
- We cannot test 200 millions of hypotheses !!!
- How to choose candidates for interacting genes?
- Word2vec produces word embeddings: d-dimensional representation of words (where  $d \sim 100$ ).
- *Words that occur in similar contexts tend to have similar meanings (Harris, 1954; Firth, 1957; Deerwester et al., 1990). – If words have similar row vectors in a word–context matrix, then they tend to have similar meanings.*

*You shall know a word by the company it keeps*

# Problem 1

1. Download abstracts from PUBMED [~25GB compressed data]
2. Calculate the word context matrix - which word occurs in what context
3. Use GloVe or other word2vec algorithm to find 50-, 200, 600-dimensional representation of words
4. Find genes close to the gene of interest



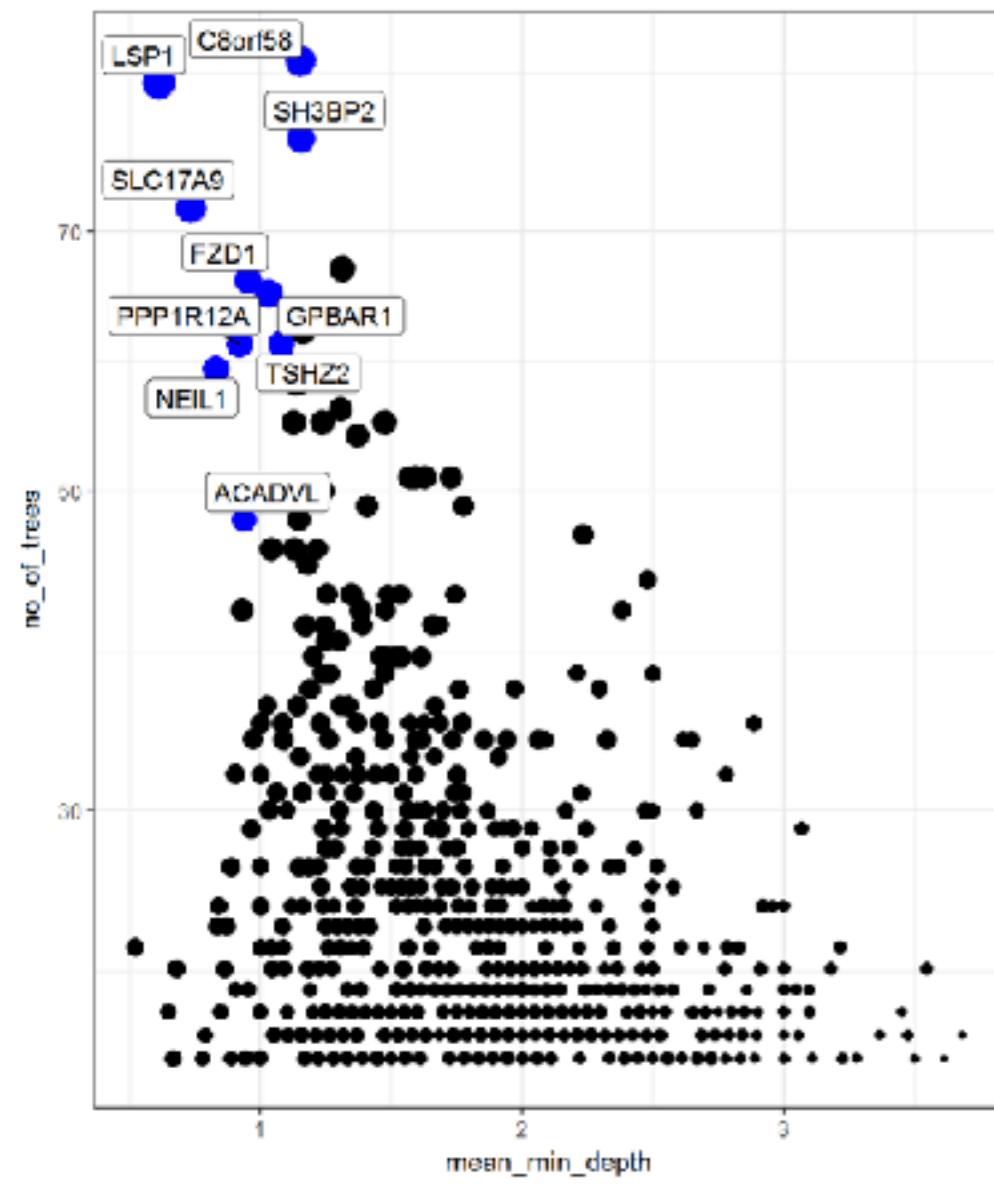
Maintainer: Kornel Kiełczewski

# Problem 2

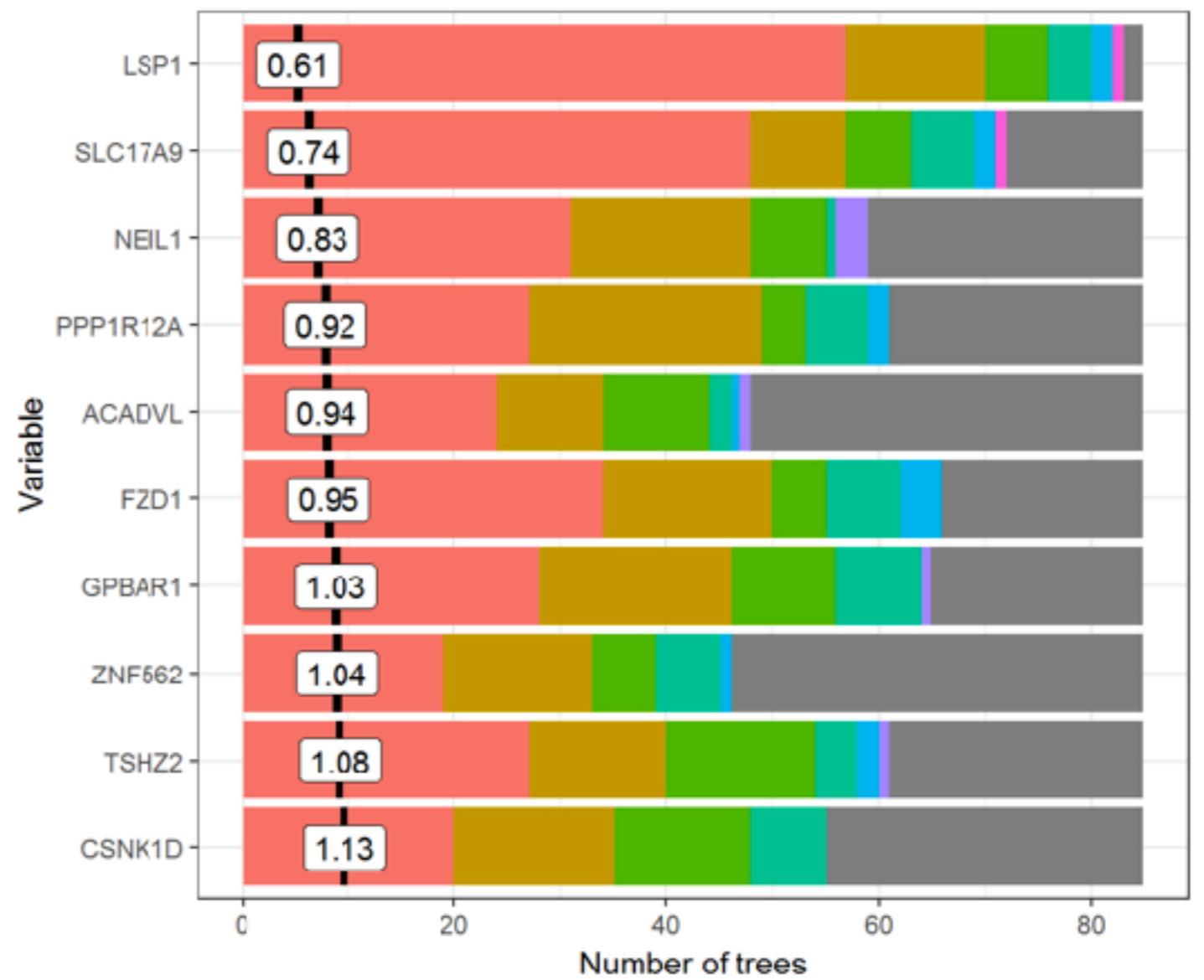
randomForestExplainer:  
What is in the (random) forest?

Project Lead:  
Ola Paluszynska

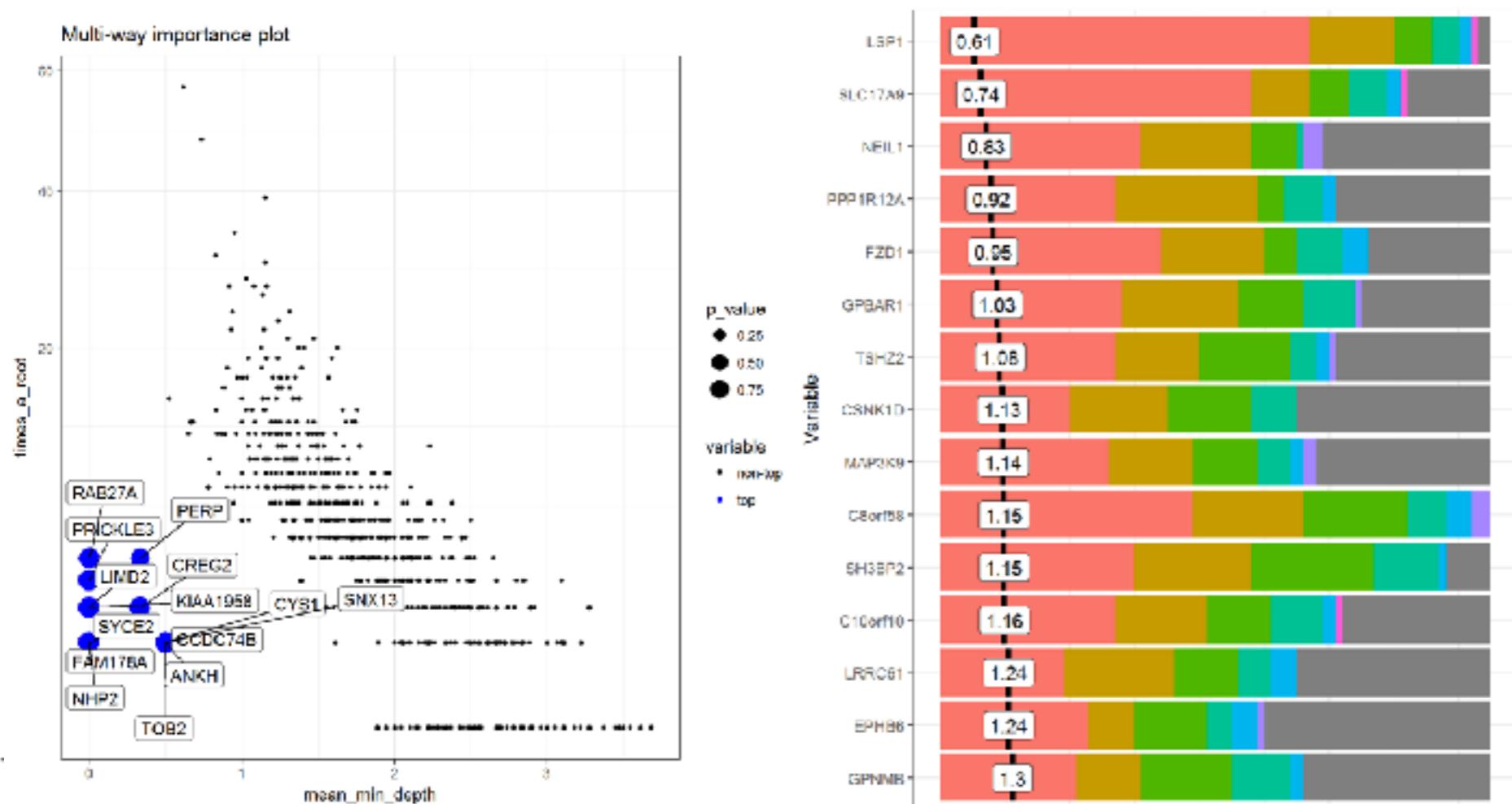
Multi-way importance plot



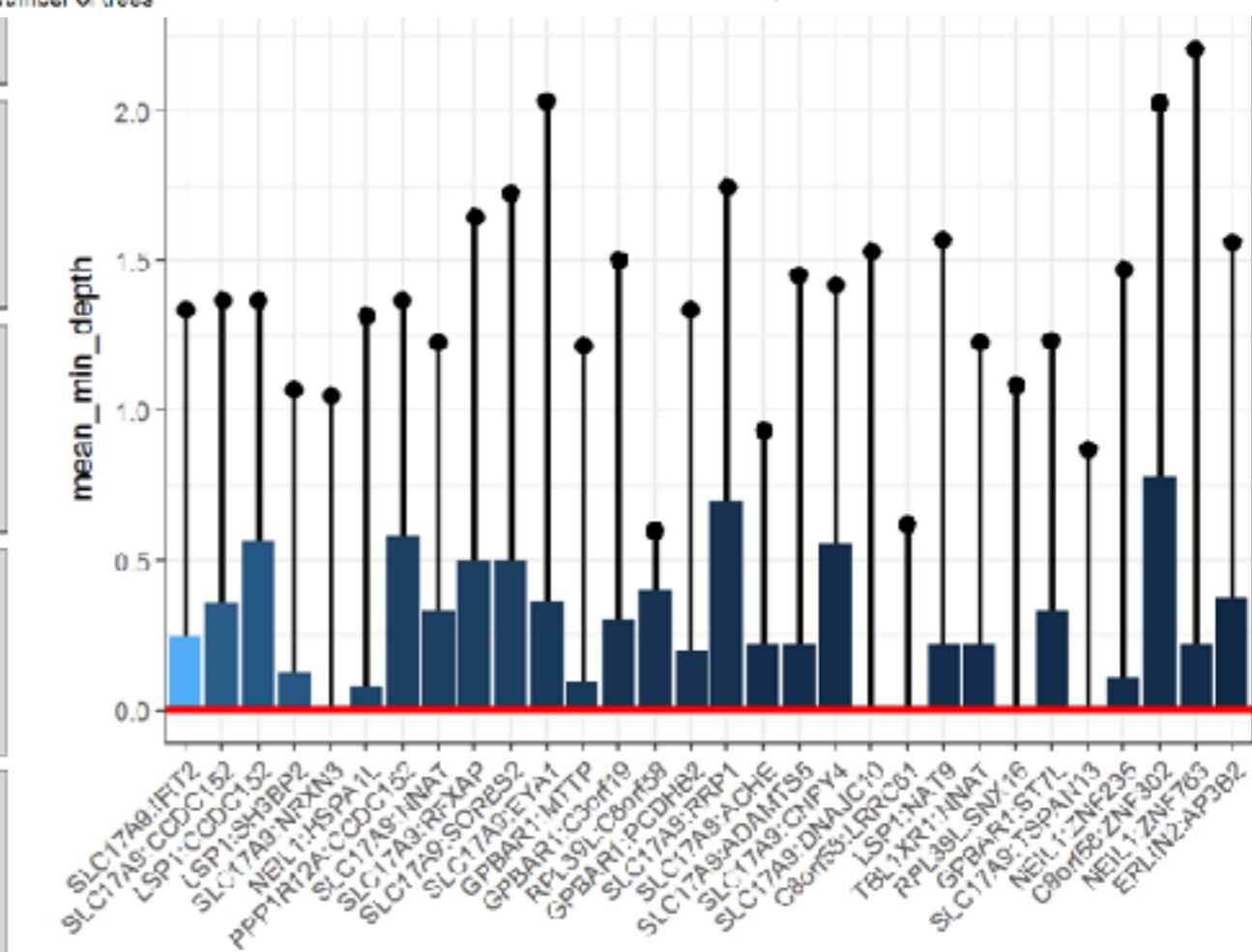
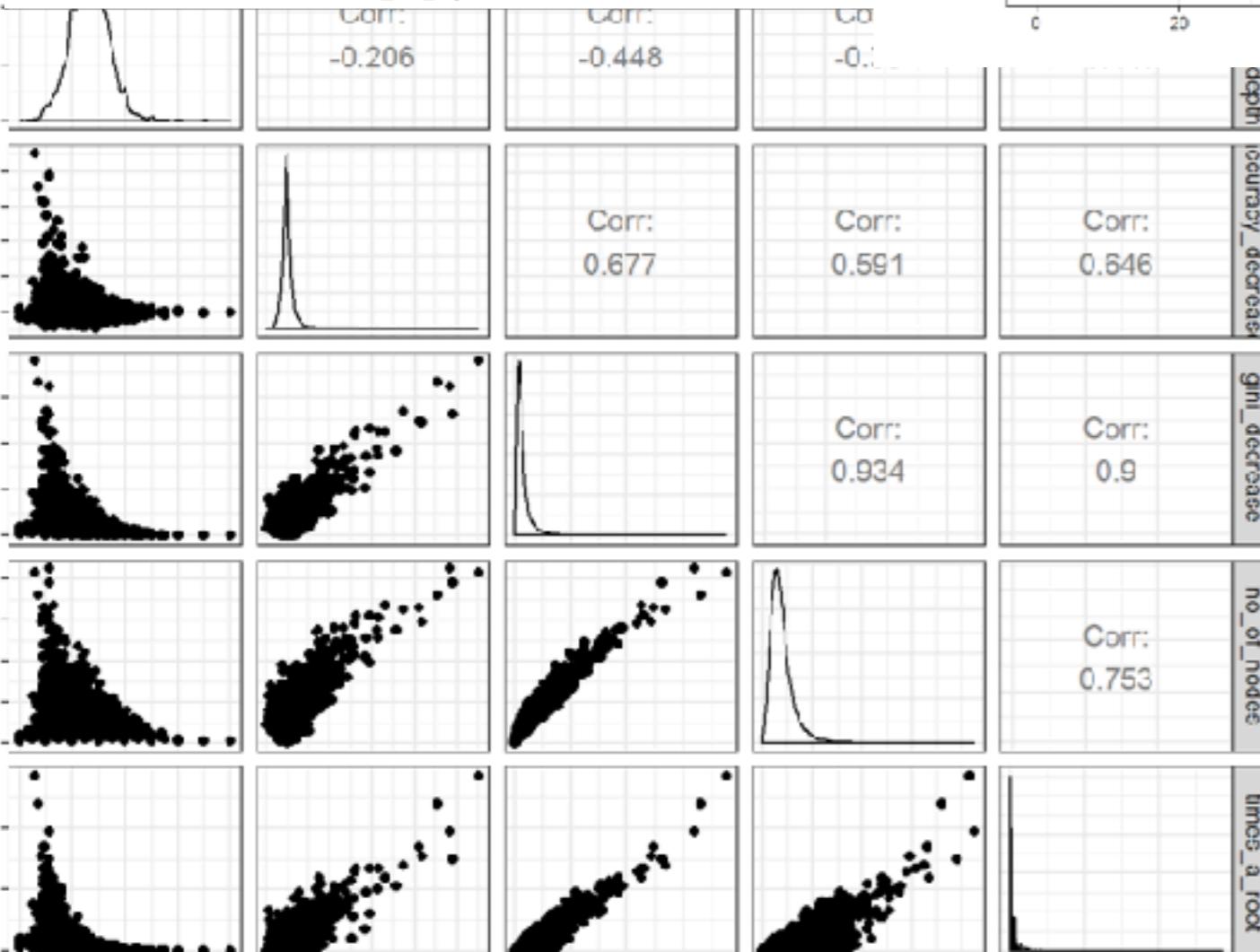
The distribution of minimal depth



Project Lead:  
Ola Paluszynska

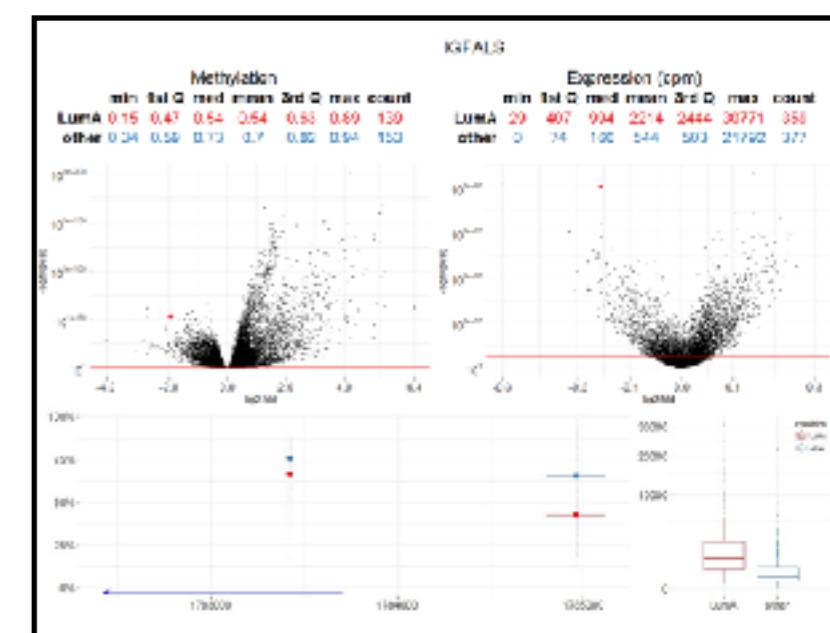
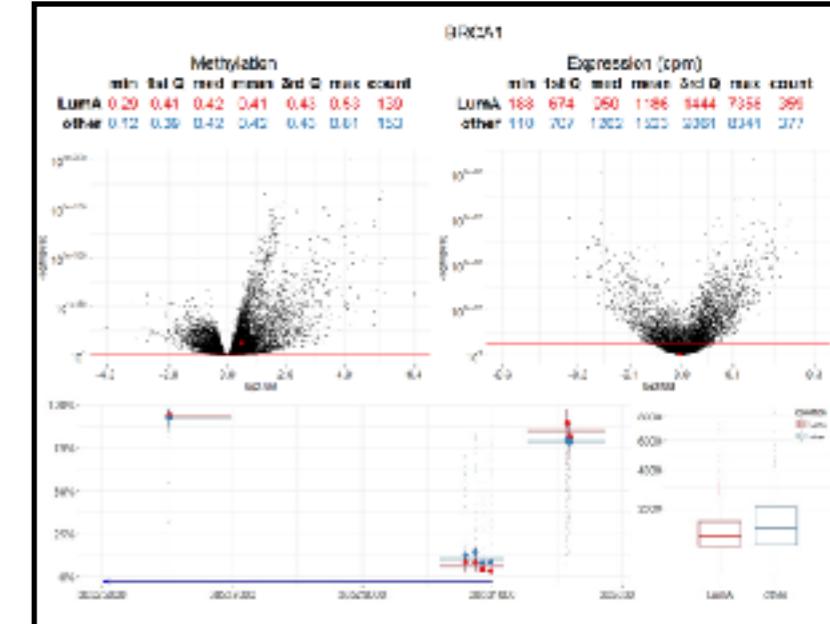
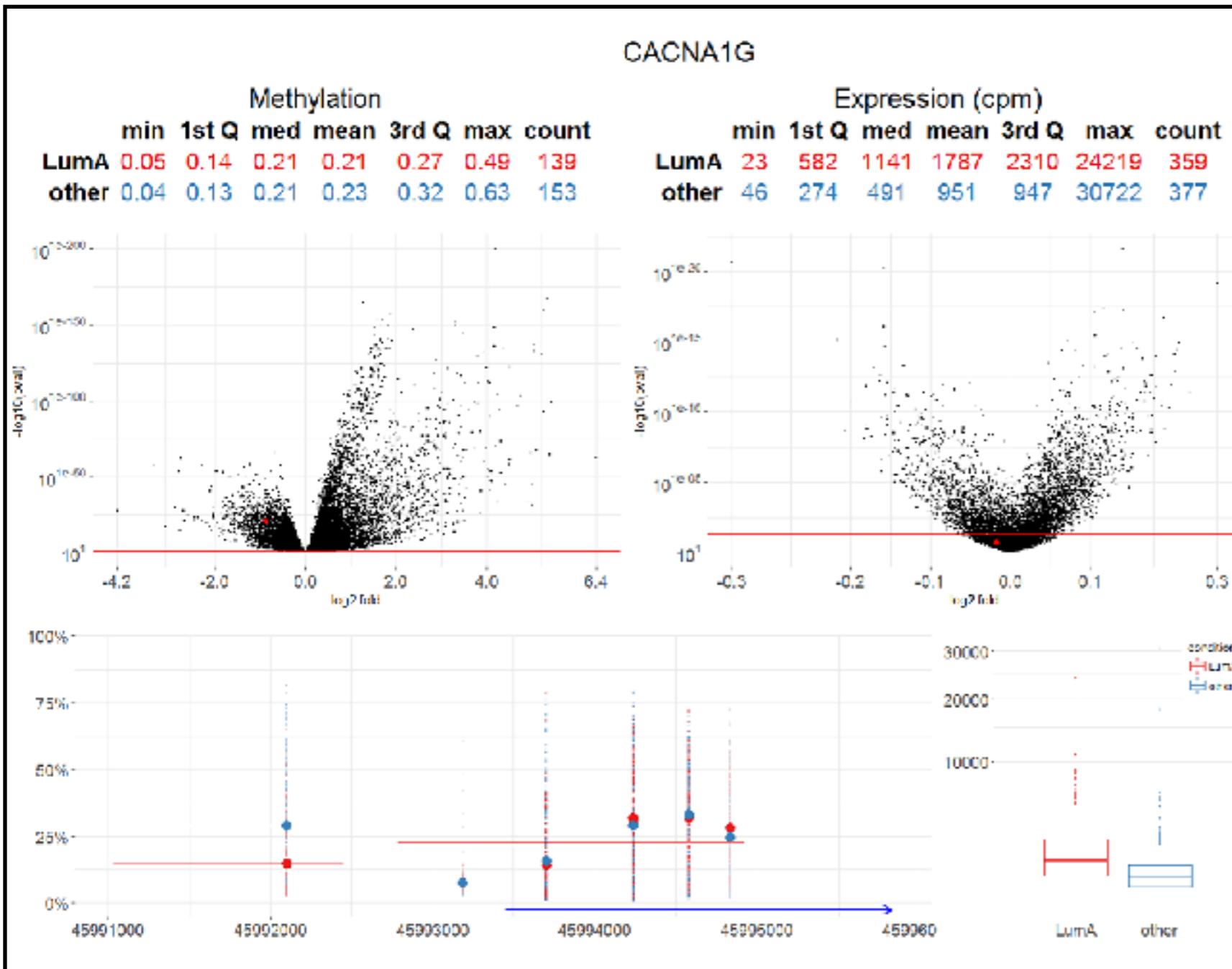


### most frequent interactions



# MLGenSig

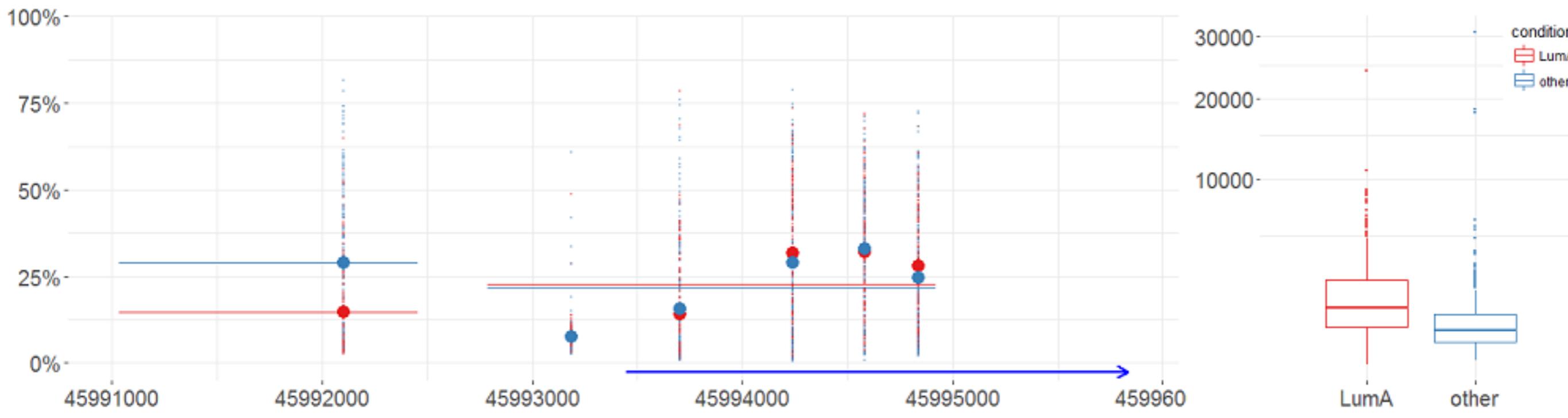
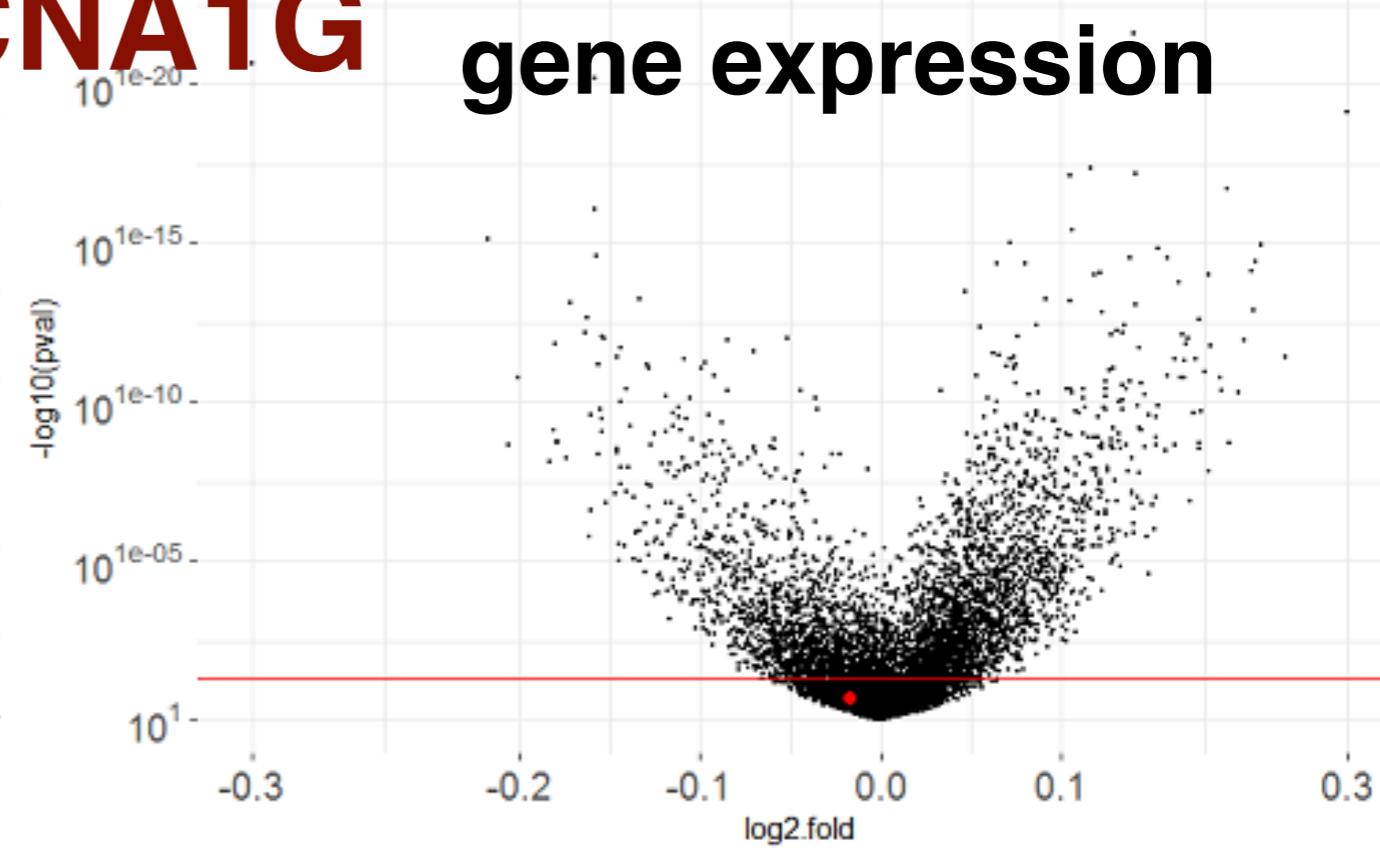
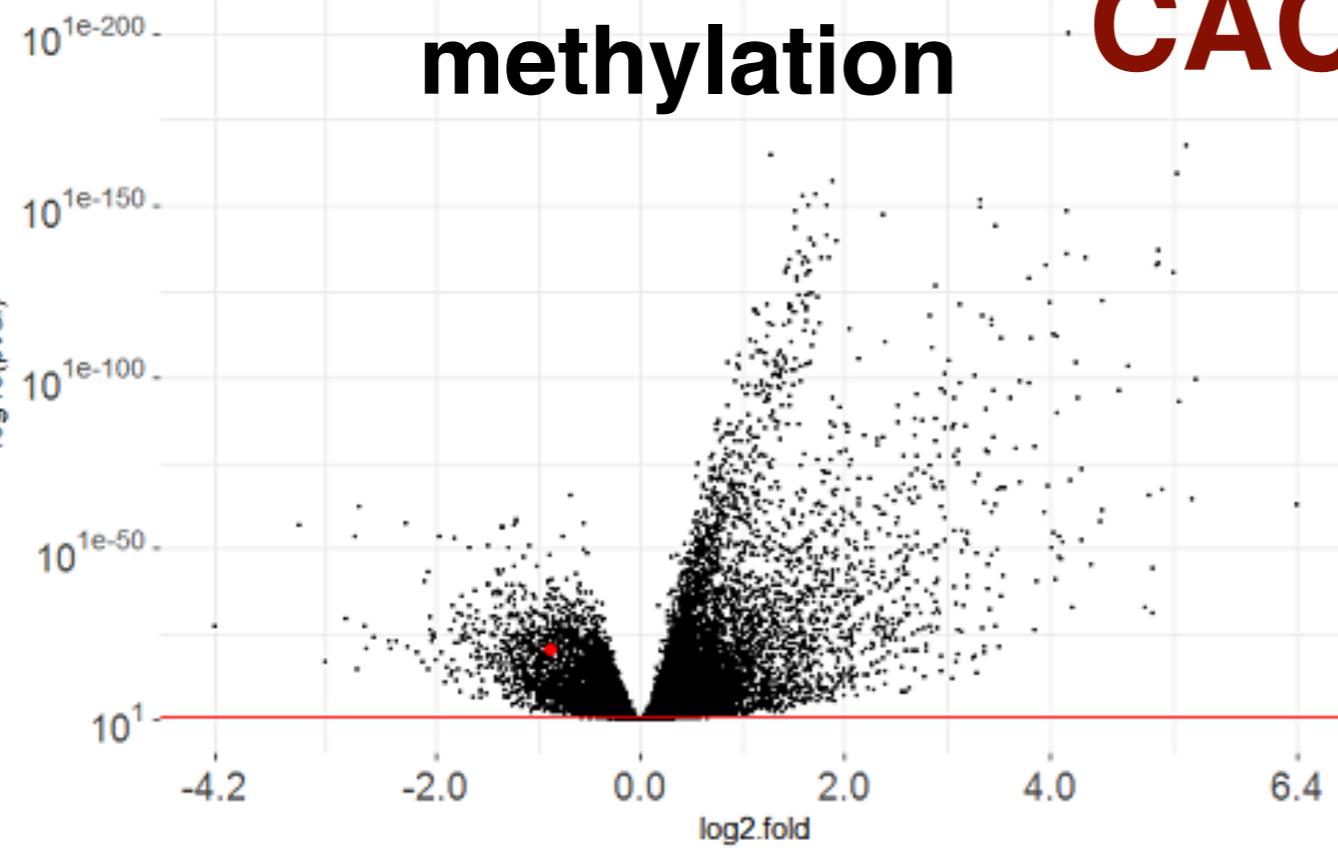
Project Lead:  
Aleksandra Dąbrowska  
Alicja Gosiewska



# methylation

# CACNA1G

# gene expression



# Problem 4

Feature Importance

Random Forest

Clinical features

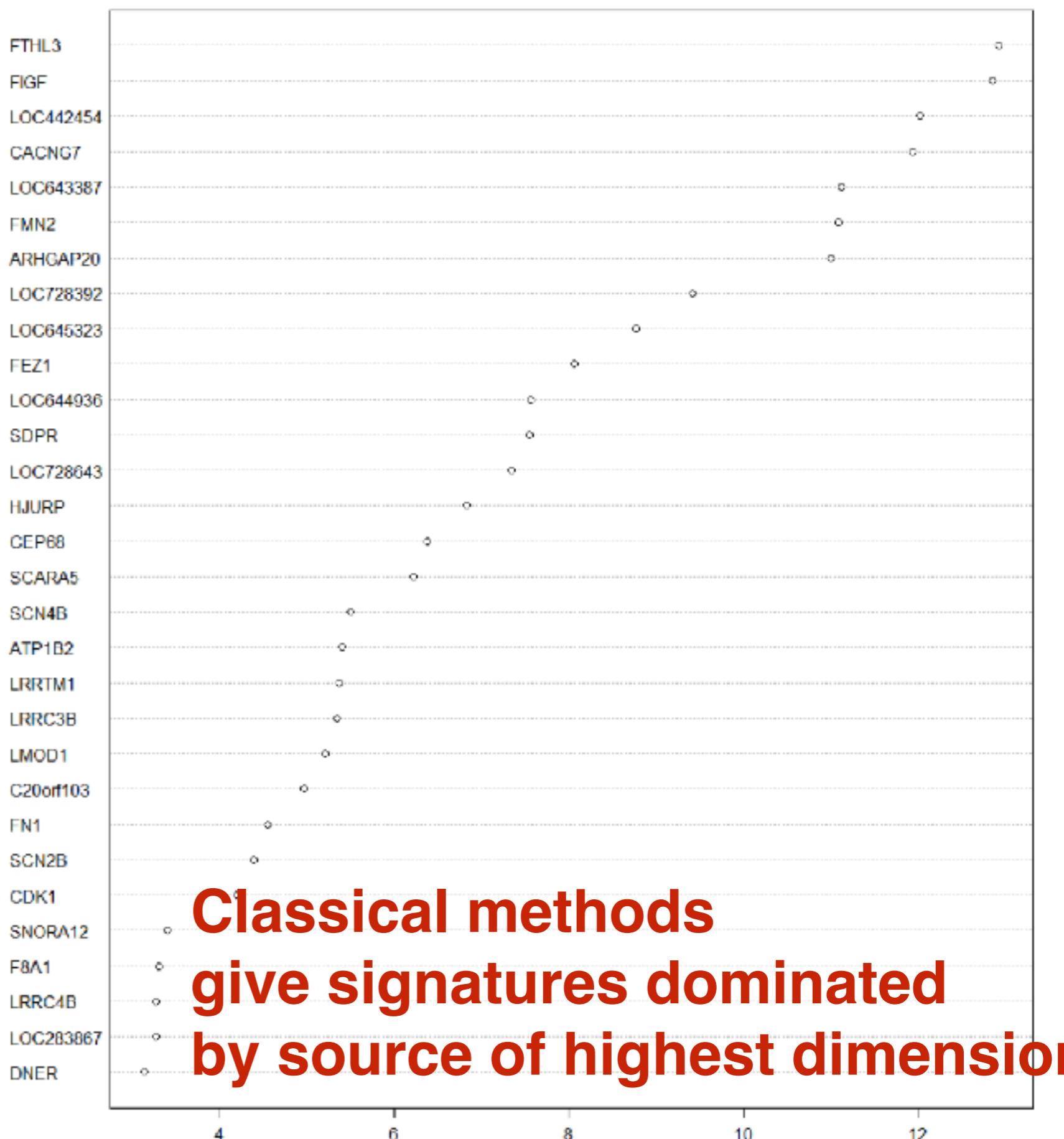
size: 1-100

mRNA seq

size: 100 - 20 000

Methylation

size: 10 000 - 1 000 000



## Problem 4

## Modified Bayesian Information Criteria

$$BIC = -2 \log L(M_i, \hat{\theta}_i) + k_i \log(n)$$

$$mBIC = -2 \log L(M_i, \hat{\theta}_i) + k_i^{cli} p_{cli} + k_i^{gen} p_{gen} + k_i^{met} p_{met}$$

Model fitness

penalty for clinical features

penalty for mRNA seq

penalty for methylation

Other penalties for other data sources?

## Problem 4

## Modified Bayesian Information Criteria

$$mBIC = -2 \log L(M_i, \hat{\theta}_i) + k_i^{cli} p_{cli} + k_i^{gen} p_{gen} + k_i^{met} p_{met}$$

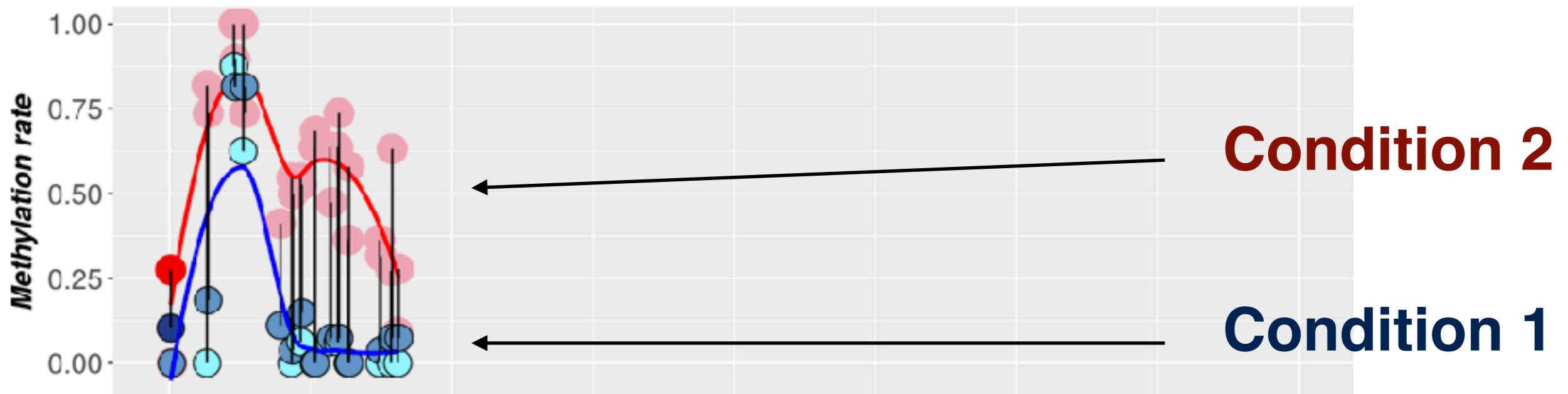
*Locating multiple interacting quantitative trait loci using robust model selection.* Andreas Baierl, Andreas Futschik, Małgorzata Bogdan, Przemysław Biecek. **Computational Statistics & Data Analysis, 51 (12) p. 6423–6434**

*Extending the Modified Bayesian Information Criterion (mBIC) to Dense Markers and Multiple Interval Mapping.* Małgorzata Bogdan, Florian Frommlet, Przemysław Biecek, Riyan Cheng, Jayanta K. Ghosh, R.W. Doerge.  
**Biometrics 64 (4) p. 1162–1169 (2008)**

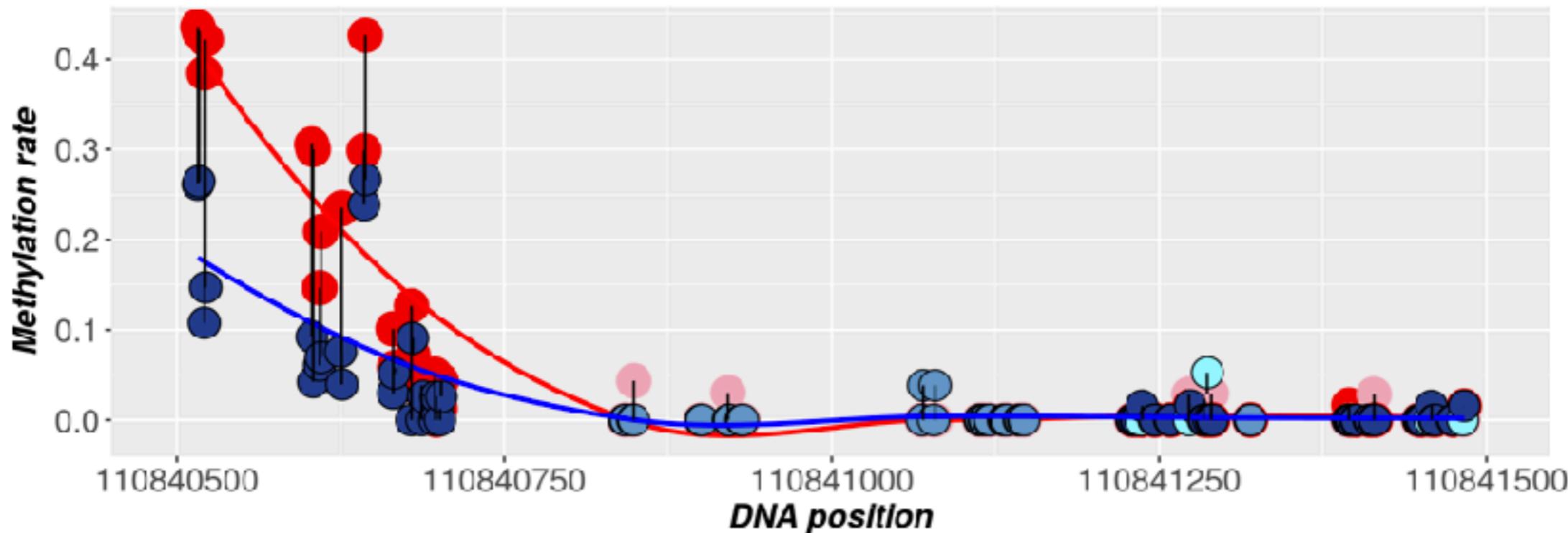
*The R Package bgmm: Mixture Modeling with Uncertain Knowledge.*  
Przemysław Biecek, Ewa Szczurek, Martin Vingron and Jerzy Tiuryn **Journal of Statistical Software, 47(3), pp. 1-31 (2012)**

# Problem 5

**Methylation rate within two samples in chr7**

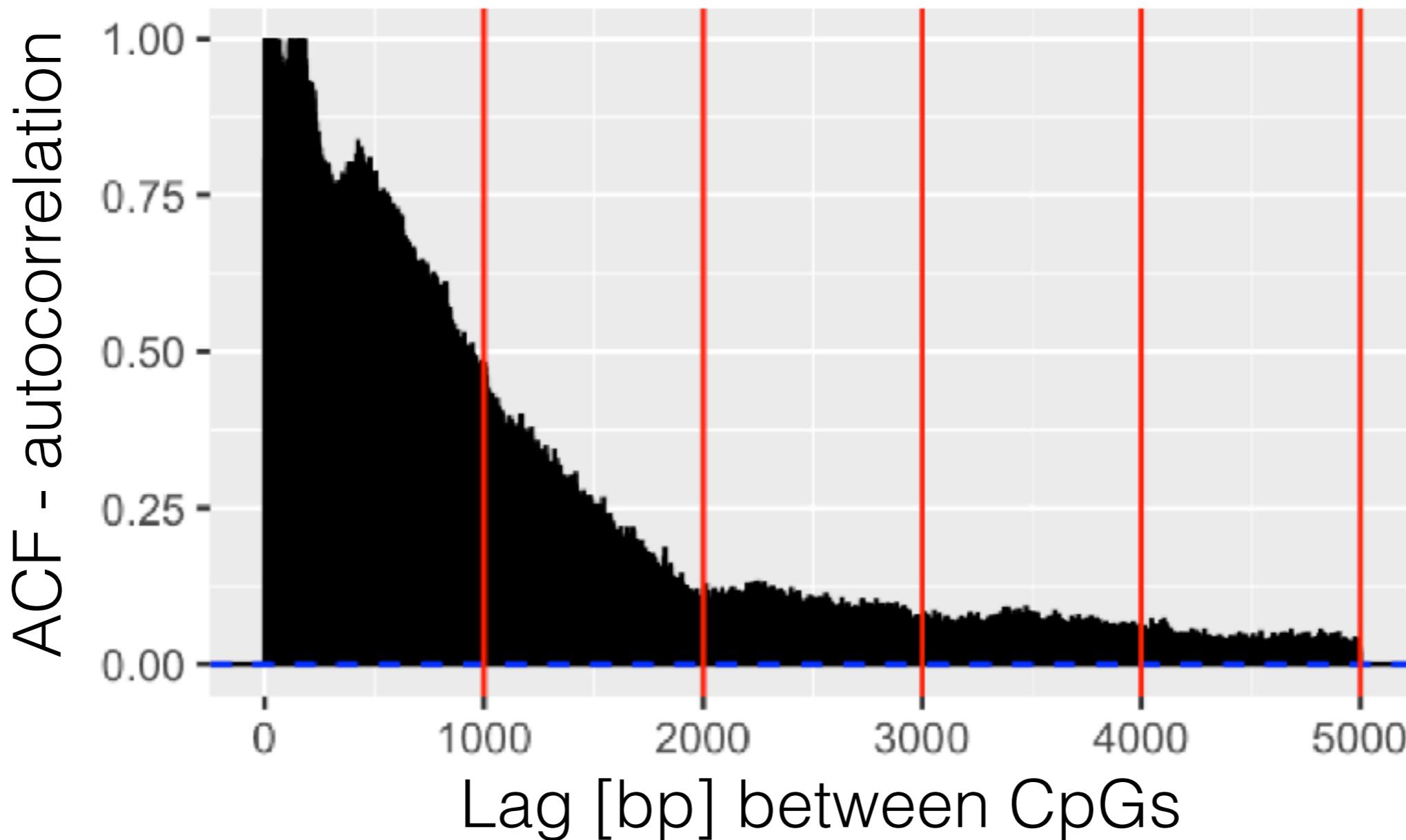


**Methylation rate within two samples in chr12**



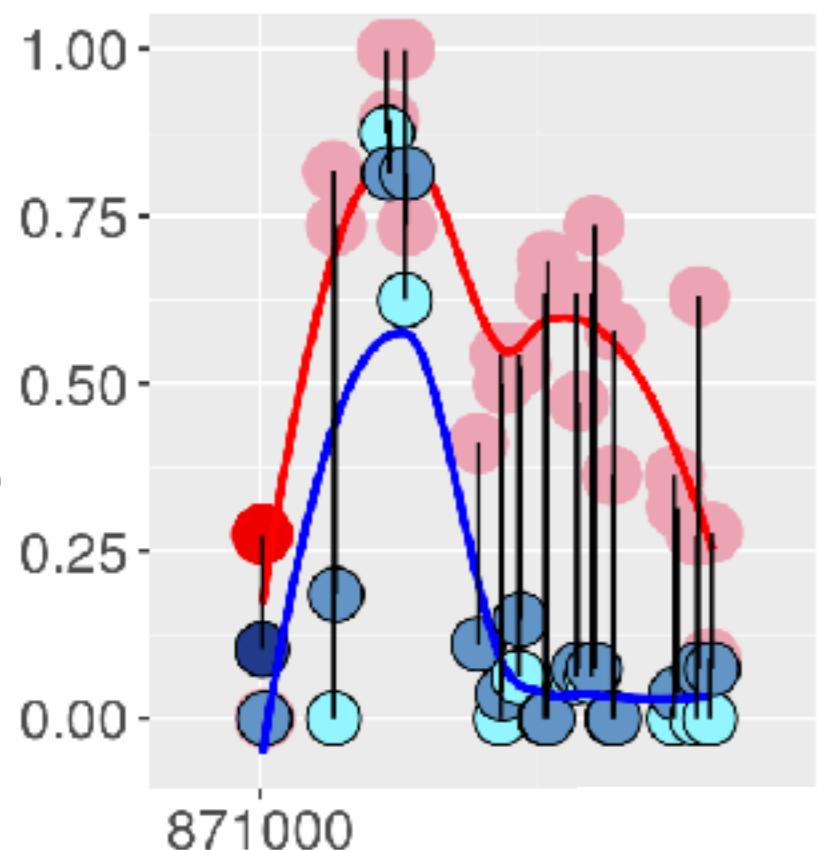
## Problem 5

Methylation is strongly autocorrelated.  
The autocorrelation shall be taken into account



## Problem 5

**Methylation rate**



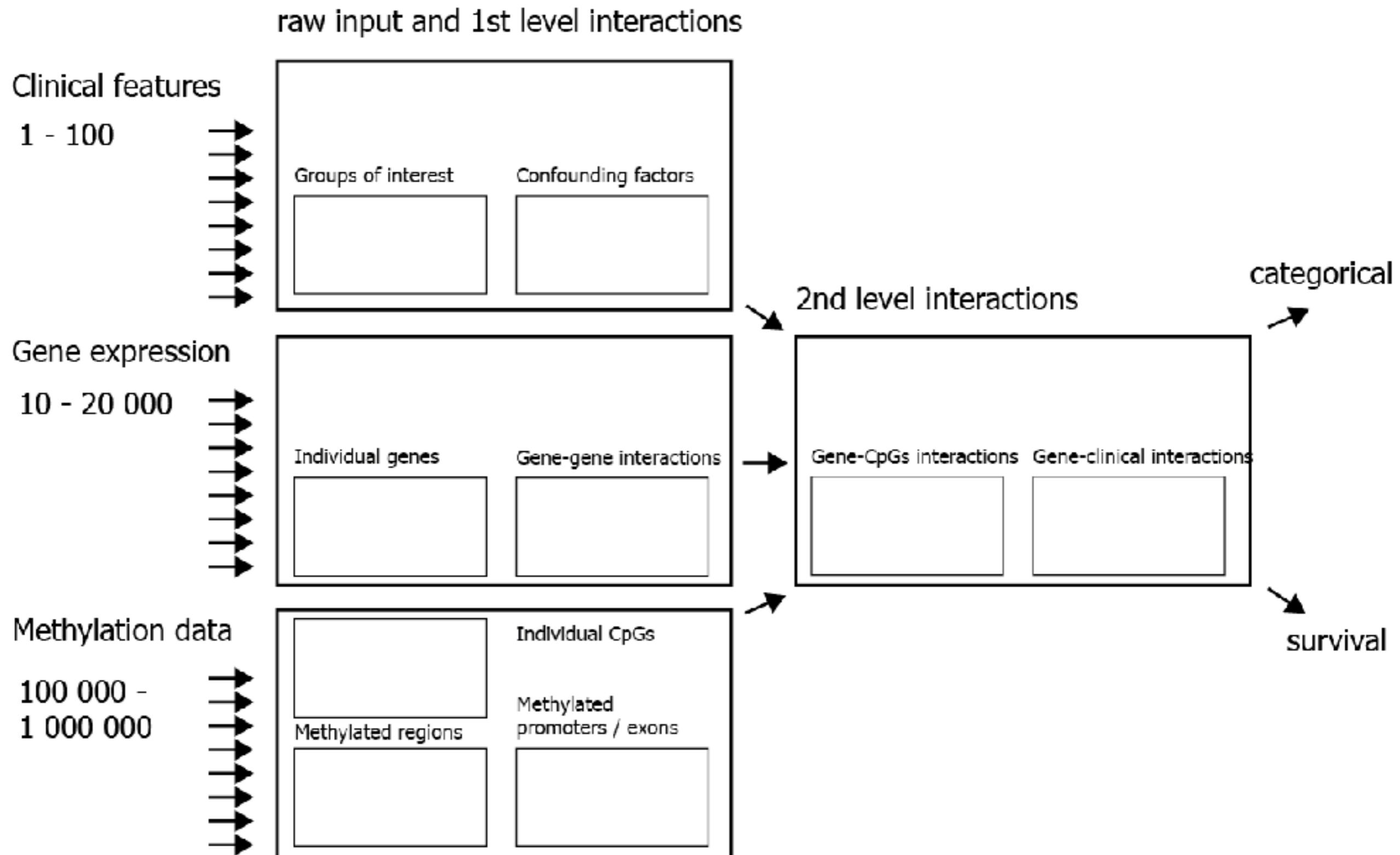
Again, generalised mixed models

$$y = X\beta + Zu + \varepsilon$$

$$\varepsilon \sim \mathcal{N}(0, I\sigma^2)$$

$$u \sim \mathcal{N}\left(0, \begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_3 & \dots & \dots \\ \rho_1 & 1 & \rho_1 & \rho_2 & \rho_3 & \dots \\ \rho_2 & \rho_1 & 1 & \rho_1 & \rho_2 & \rho_3 \\ \rho_3 & \rho_2 & \rho_1 & 1 & \rho_1 & \rho_2 \\ \dots & \rho_3 & \rho_2 & \rho_1 & 1 & \rho_1 \\ \dots & \dots & \rho_3 & \rho_2 & \rho_1 & 1 \end{bmatrix}\right)$$

# Learning with Structure: MLGenSig (Machine Learning Genetic Signatures)



Thank you for your attention