DNA methylation impact on the glioma patients overall survival.

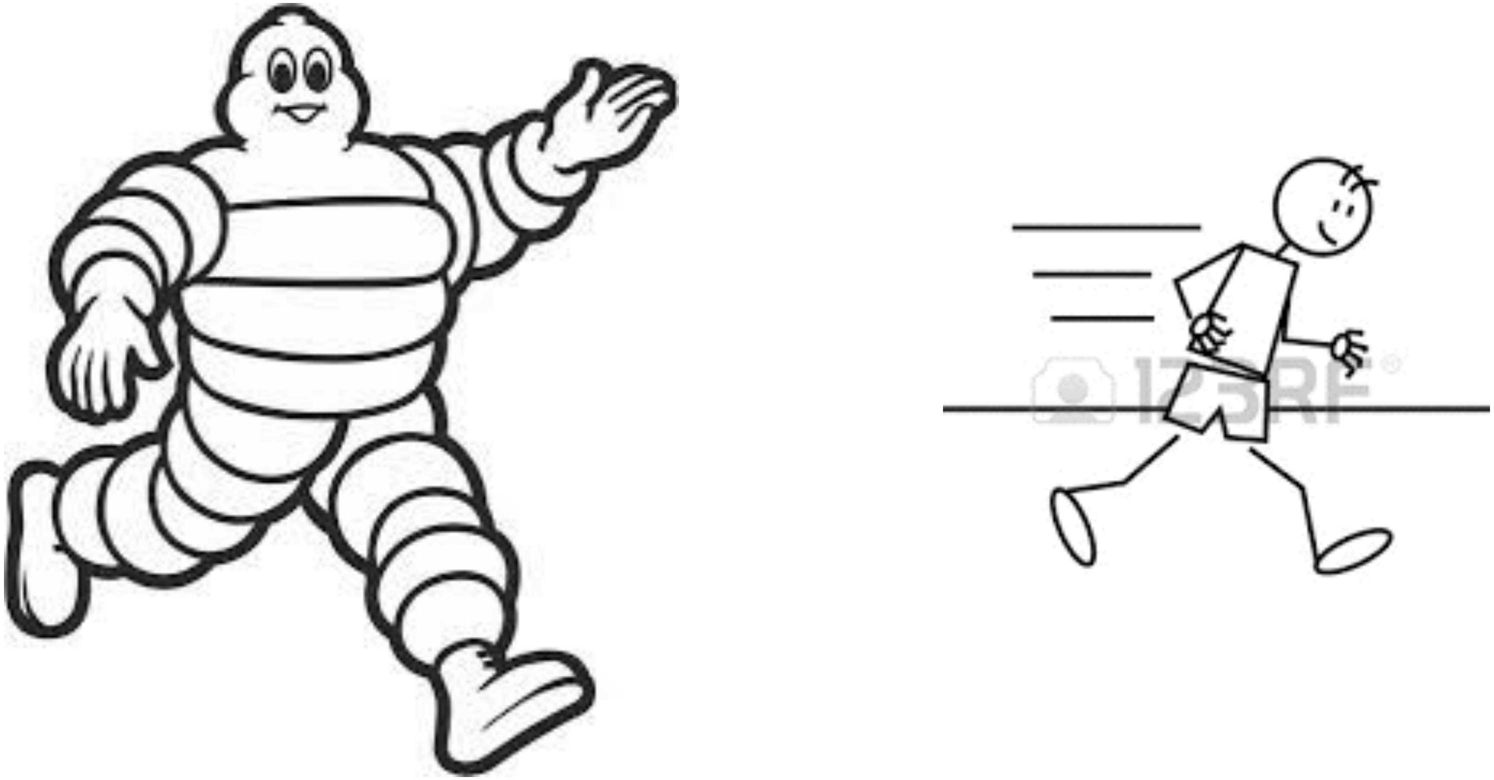
Michał J. Dąbrowski

IPI PAN seminar 2017-06-05

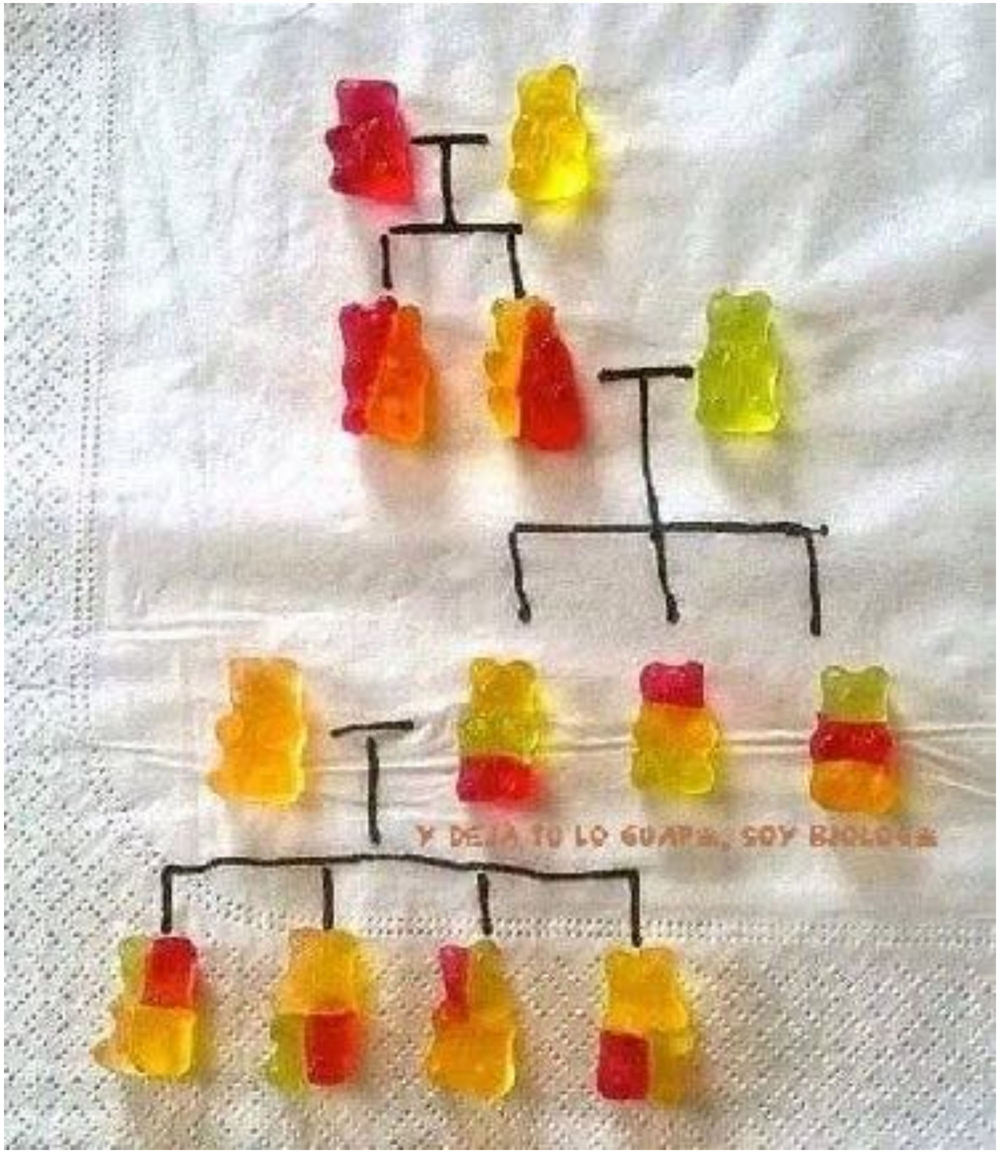
Overview

- General aspects of genetics
- The Cancer Genome Atlas data
- Feature selection
- Interdependencies Discovery
- Significant features
- Putative functions
- Laboratory validation of the results
- Next steps

Genotype, phenotype



Monogenic characteristics



Corn some day ago and today

Czasy Majów/Azteków/Inków
Kukurydza: kłos 5-7cm i 20-30 zia

X
K

Genome size

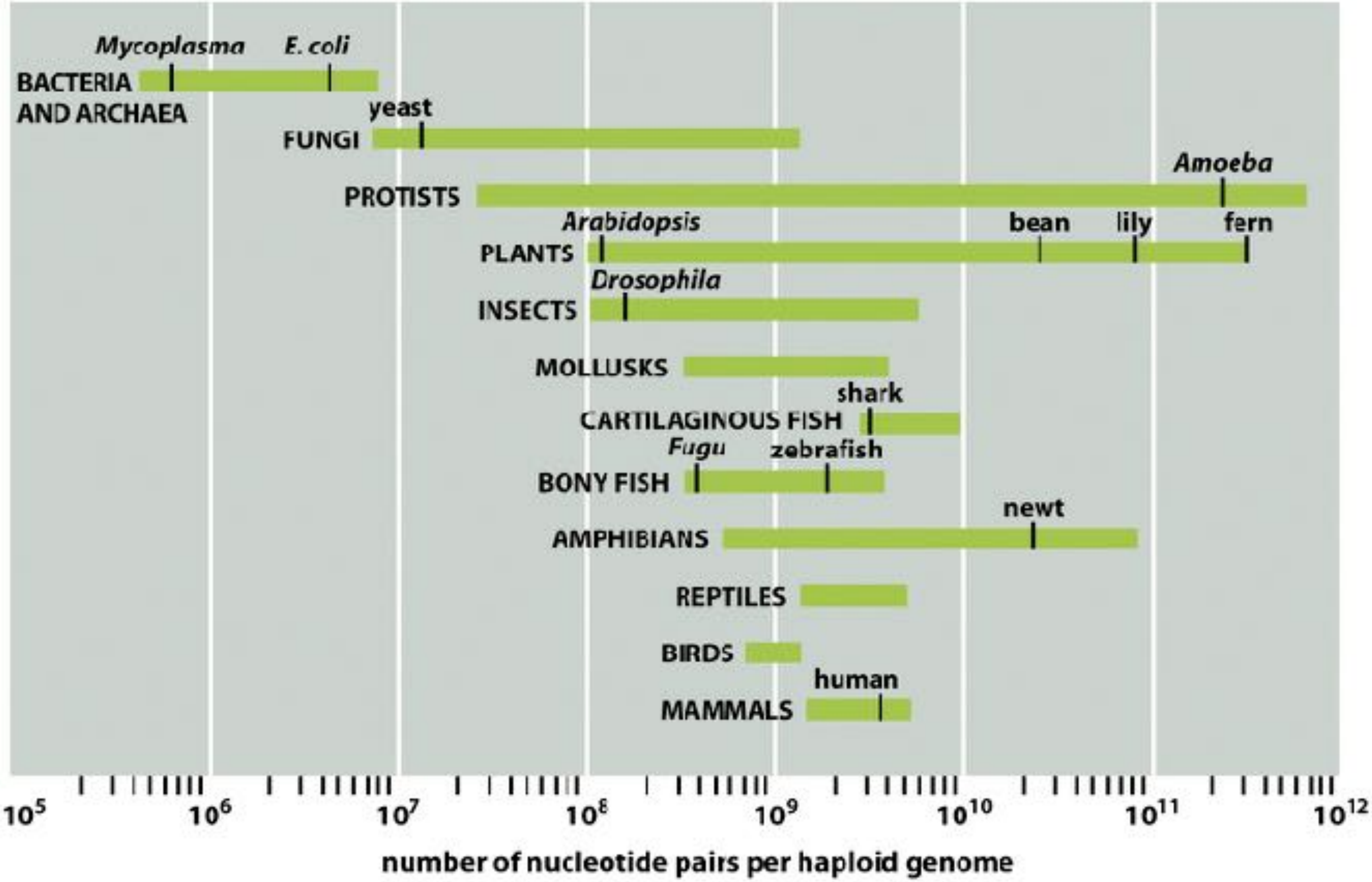


Figure 1-37 Molecular Biology of the Cell 5/e (© Garland Science 2008)

Variability

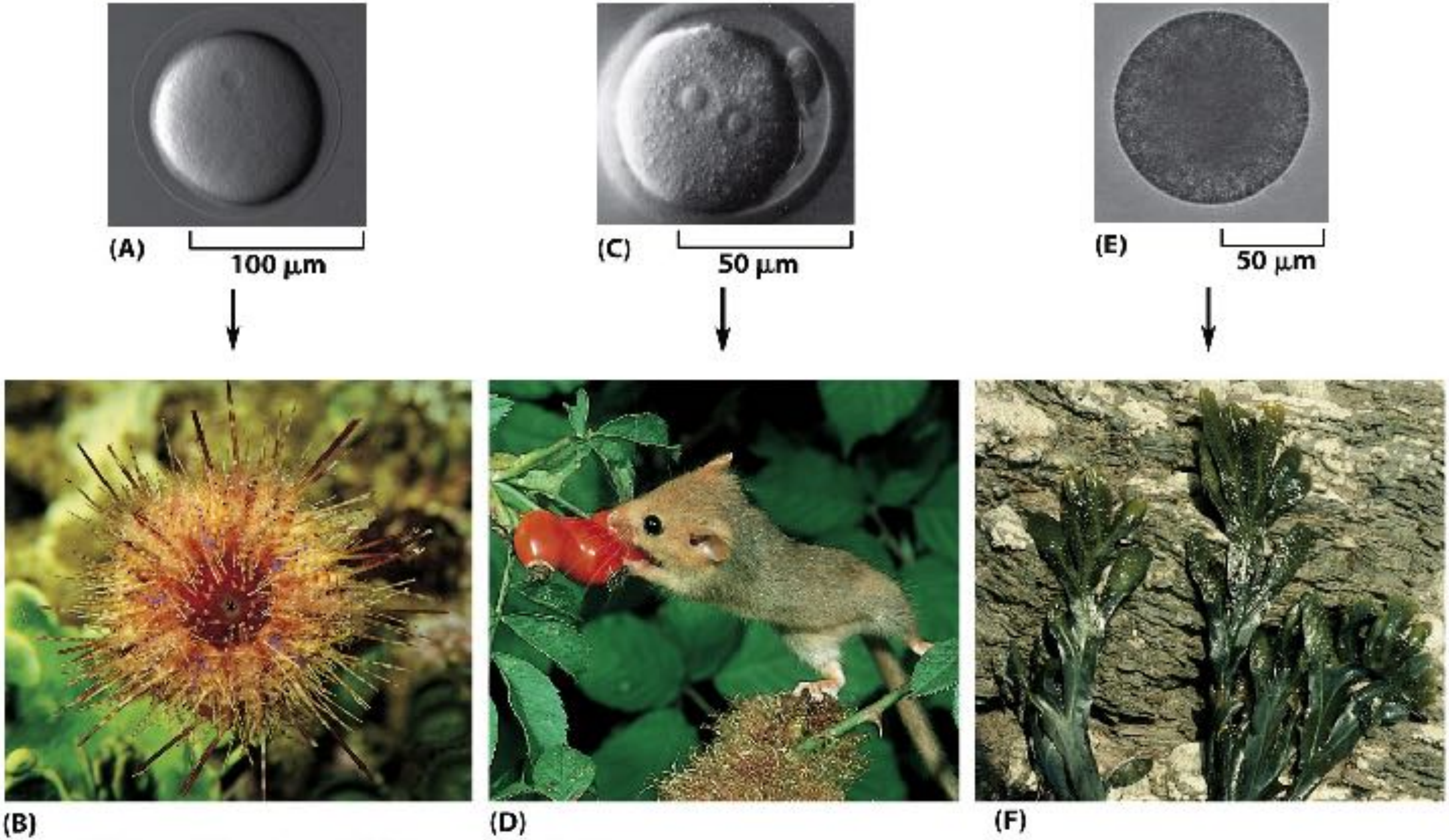
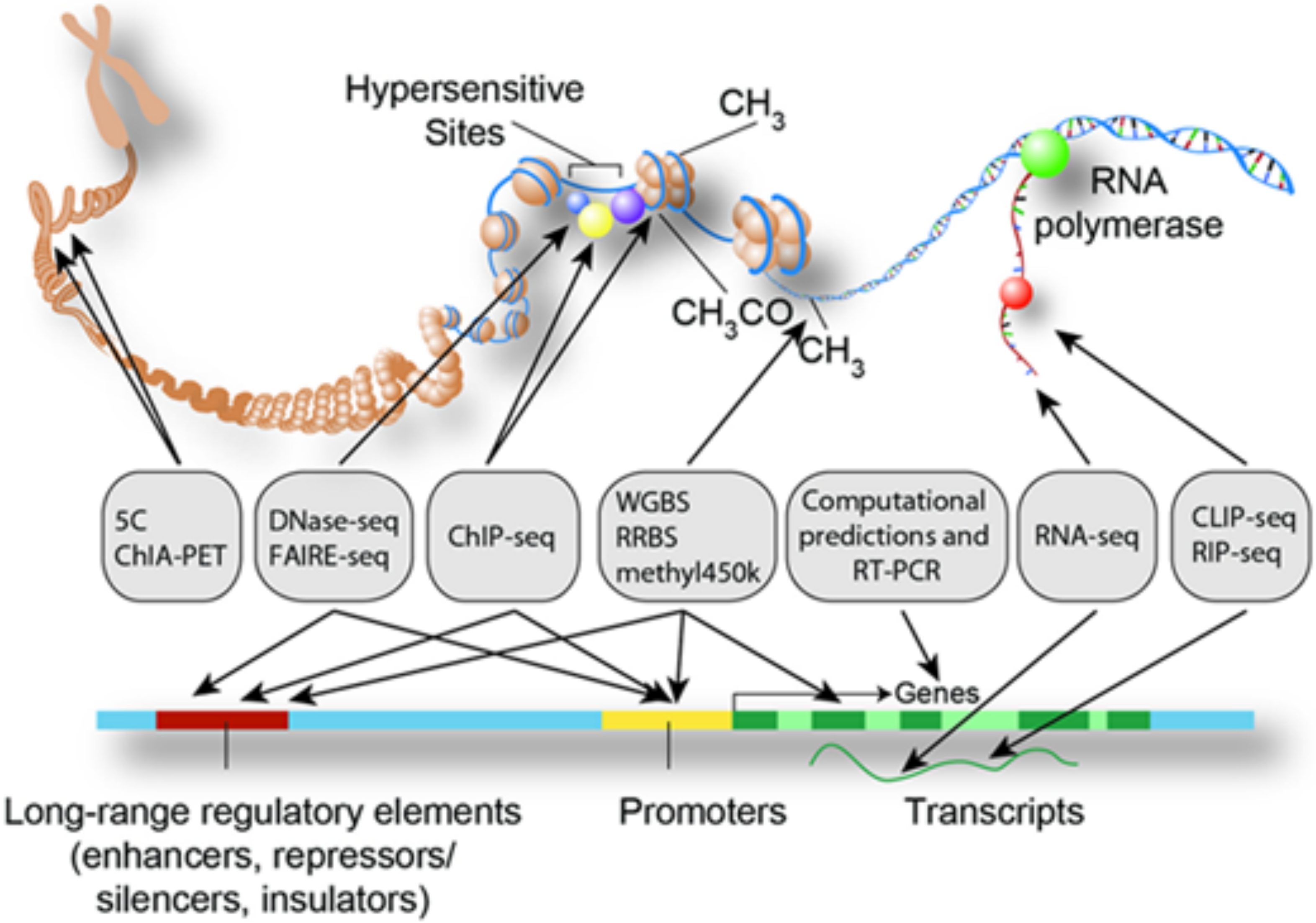
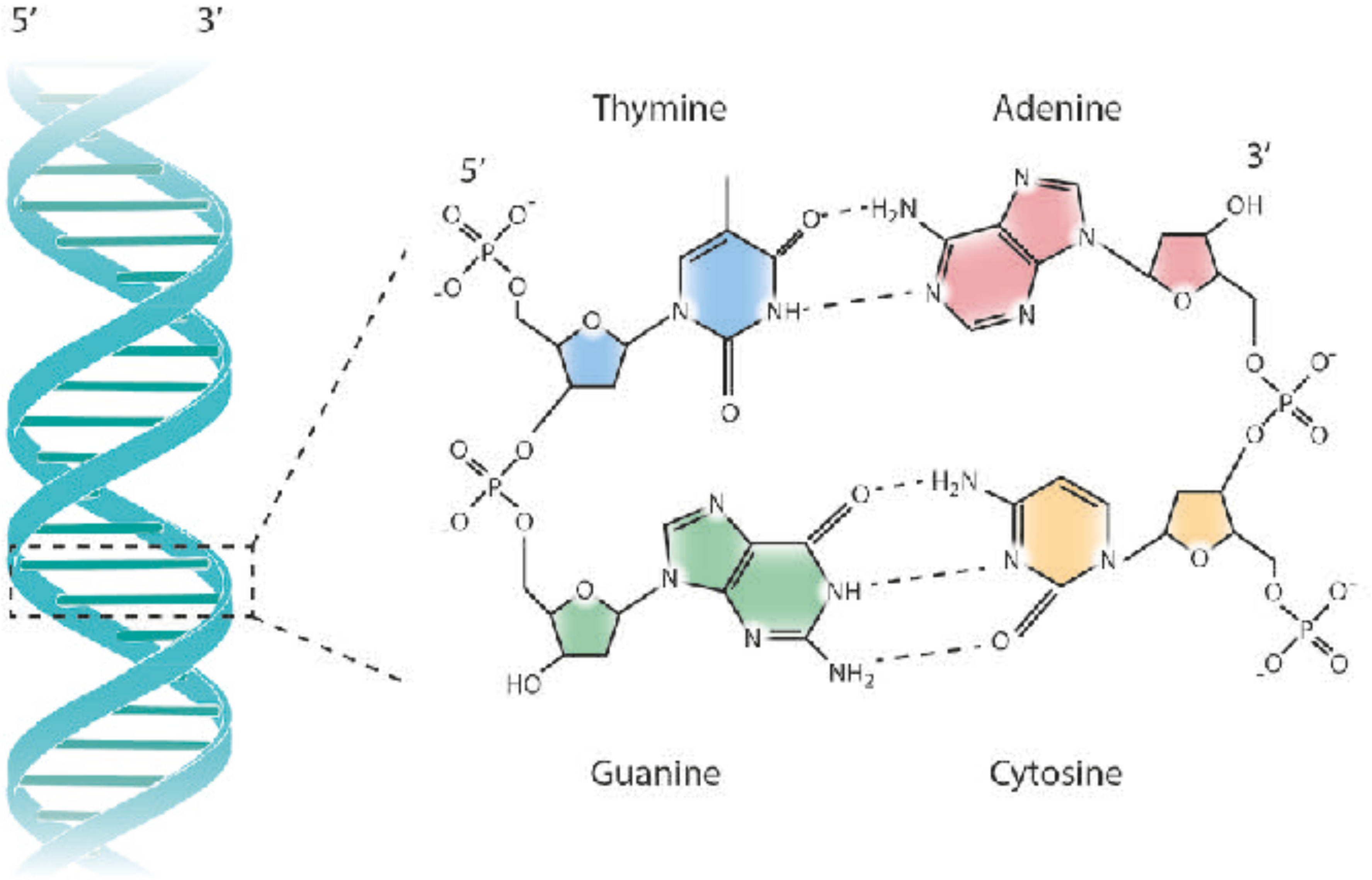


Figure 1-1 Molecular Biology of the Cell 5/e (© Garland Science 2008)

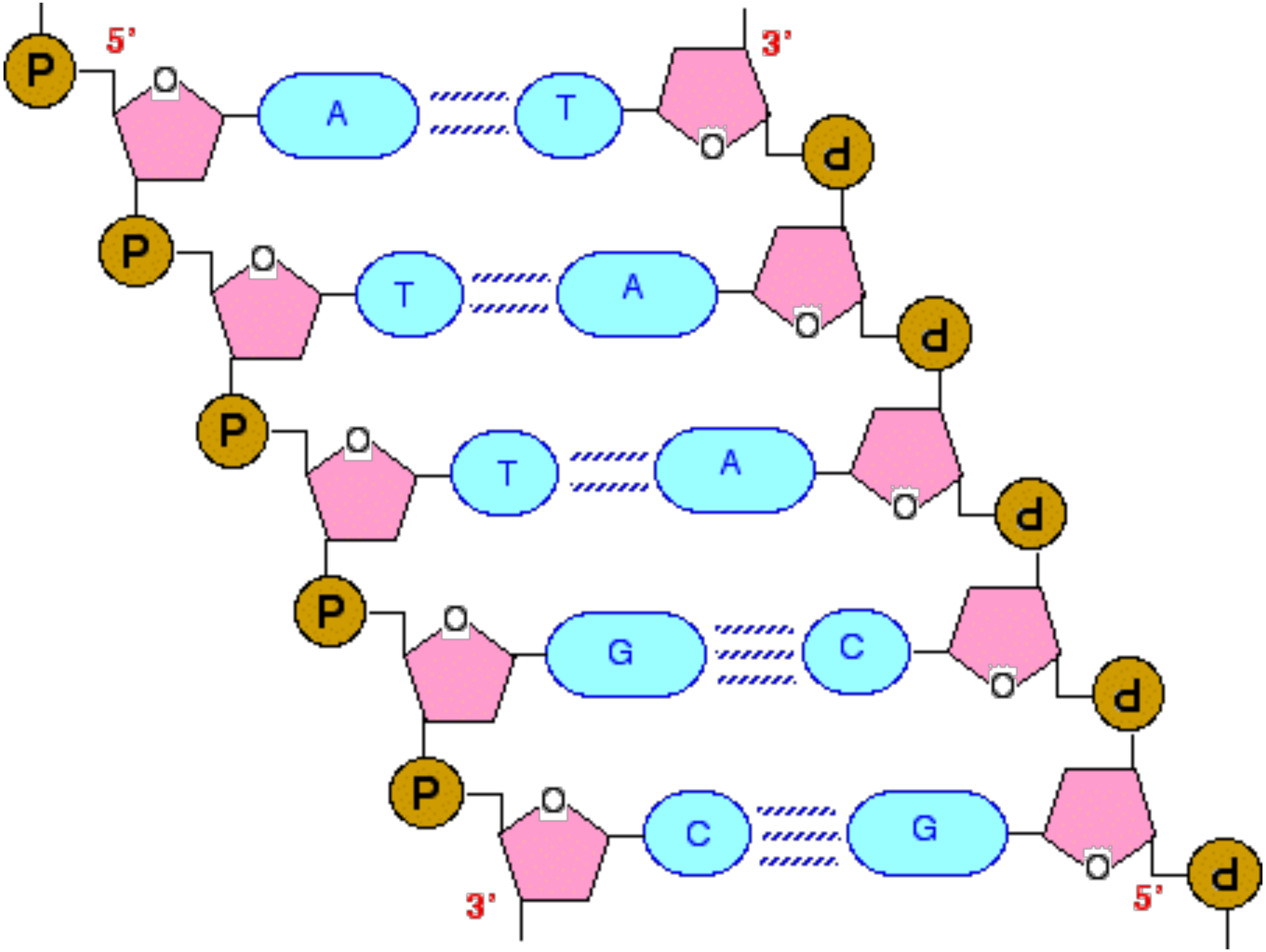
Molecular genetics



Molecular genetics



Polymer



Methylation

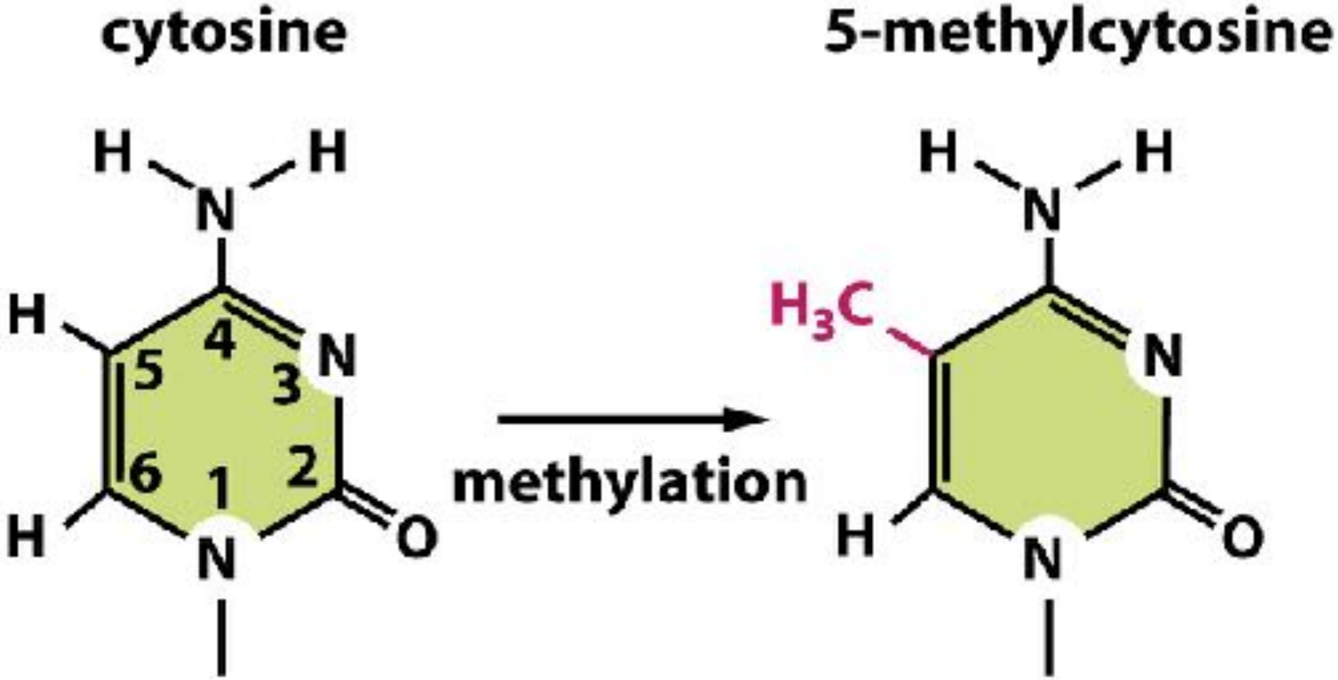
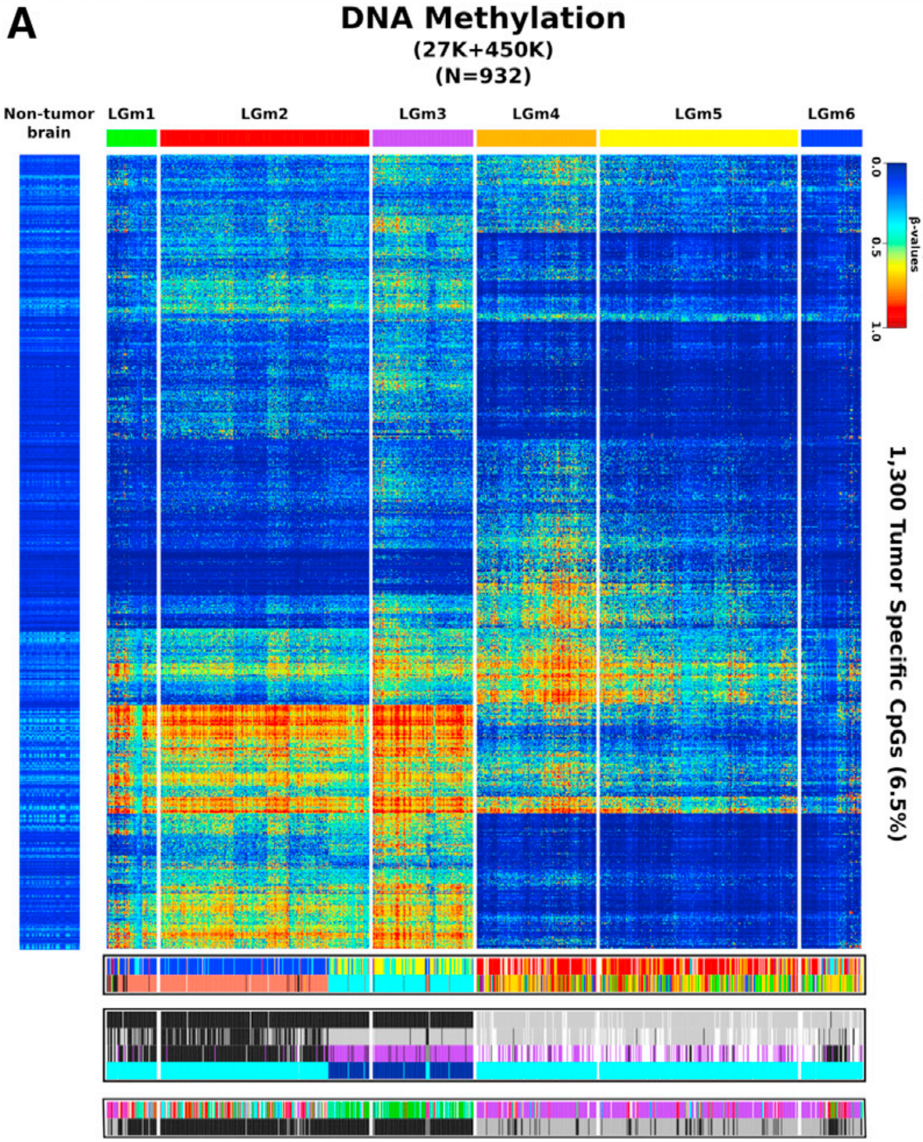


Figure 7-79 Molecular Biology of the Cell 5/e (© Garland Science 2008)

TCGA - Ceccarelli et al. 2016



- Columns represent 932 glioma patients (Human Methylation450/27)
- Grouping according to unsupervised cluster analysis
- Rows represent methylations sorted by hierarchical clustering

Legend for the Heatmaps

DNA Methylation Cluster
🟩 LGm1
🟥 LGm2
🟪 LGm3
🟧 LGm4
🟨 LGm5
🟦 LGm6

***IDH/ATRX* status**
⬜ WT
⬛ Mutant

***TERT* Expression Status**
⬛ Not Expressed
🟪 Expressed

Expression Cluster
🟨 LGr1
🟦 LGr2
🟦 LGr3
🟥 LGr4

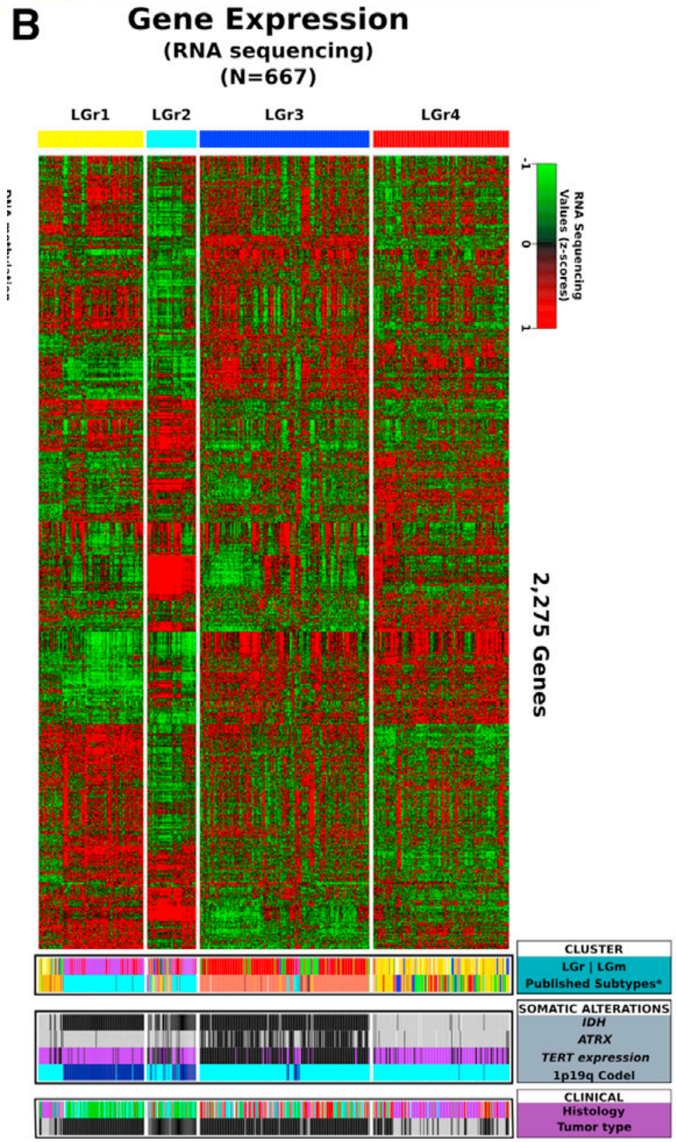
Codel 1p19q
🟦 non-codel
🟦 codel

Histology
🟥 Astrocytoma
🟪 Glioblastoma
🟦 Oligoastrocytoma
🟩 Oligodendroglioma

Published Subtypes
🟥 Classical
⬛ G-CIMP
🟦 *IDH*mut-codel
🟧 *IDH*mut-non-codel
🟨 *IDH*wt
🟩 Mesenchymal
🟦 Neural
🟧 Proneural

Tumor type
⬛ LGG (Grade II/III)
⬜ GBM (Grade IV)

TCGA - Ceccarelli et al. 2016



- Unsupervised clustering analysis for 667glioma samples
- 2,275 most variant genes

Legend for the Heatmaps

DNA Methylation Cluster
- 🟩 LGm1
- 🟥 LGm2
- 🟪 LGm3
- 🟧 LGm4
- 🟨 LGm5
- 🟦 LGm6

***IDH/ATRX* status**
- ⬜ WT
- ⬛ Mutant

***TERT* Expression Status**
- ⬛ Not Expressed
- 🟪 Expressed

Expression Cluster
- 🟨 LGr1
- 🟦 LGr2
- 🟦 LGr3
- 🟥 LGr4

Codel 1p19q
- 🟦 non−codel
- 🟦 codel

Histology
- 🟥 Astrocytoma
- 🟪 Glioblastoma
- 🟦 Oligoastrocytoma
- 🟩 Oligodendroglioma

Published Subtypes
- 🟥 Classical
- ⬛ G−CIMP
- 🟦 *IDH*mut−codel
- 🟧 *IDH*mut−non−codel
- 🟨 *IDH*wt
- 🟩 Mesenchymal
- 🟦 Neural
- 🟧 Proneural

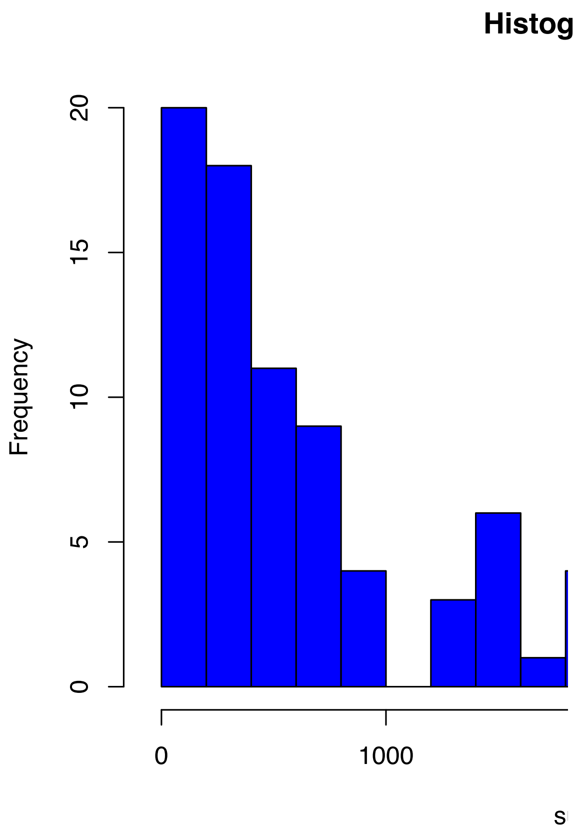
Tumor type
- ⬛ LGG (Grade II/III)
- ⬜ GBM (Grade IV)

The above results confirm IDH status as the major determinant of the molecular footprints of diffuse glioma.

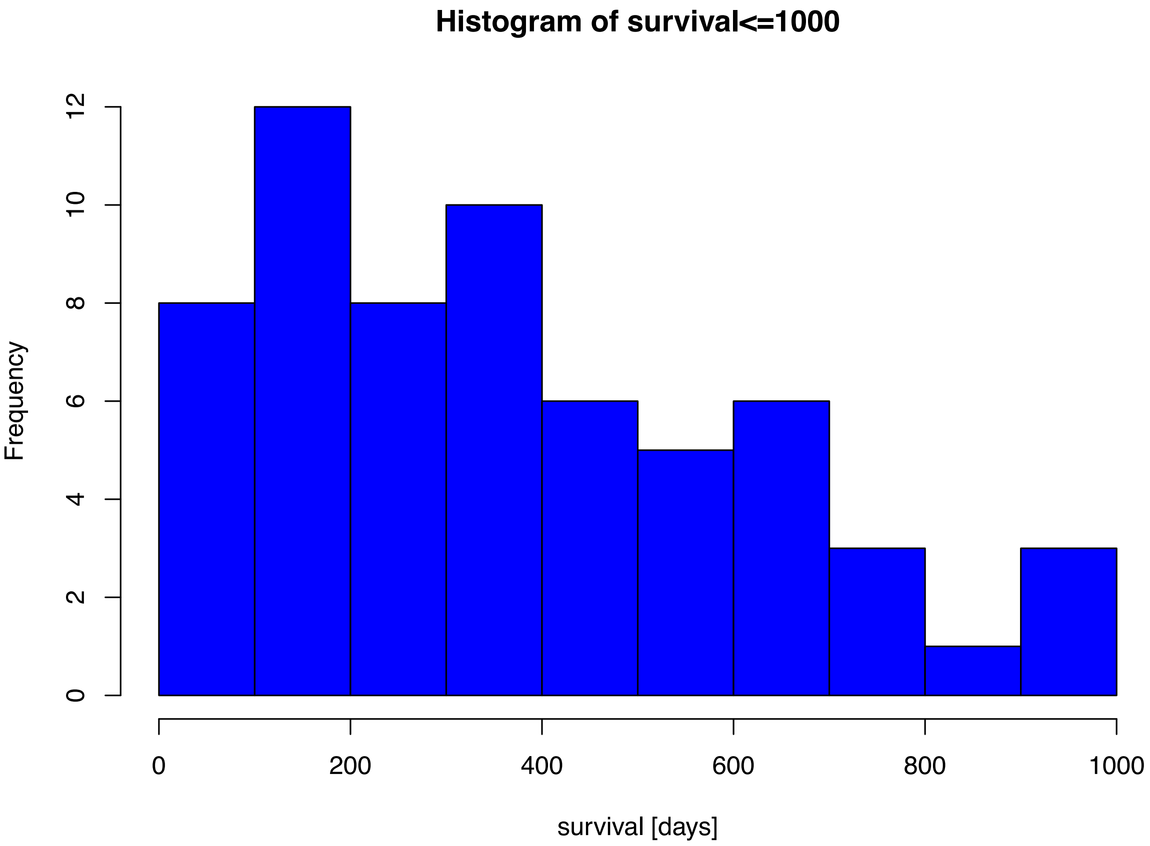
Aim: Detect methylation sites interacting with gene expression levels

- Data: from TCGA 88 glioma patients
    - gene expression,
    - methylation beta values
- Scientific question: Can we model the patients survival?
    - Binary decision “up to 400” and “over 400”

Patients survival



Histogram of survival

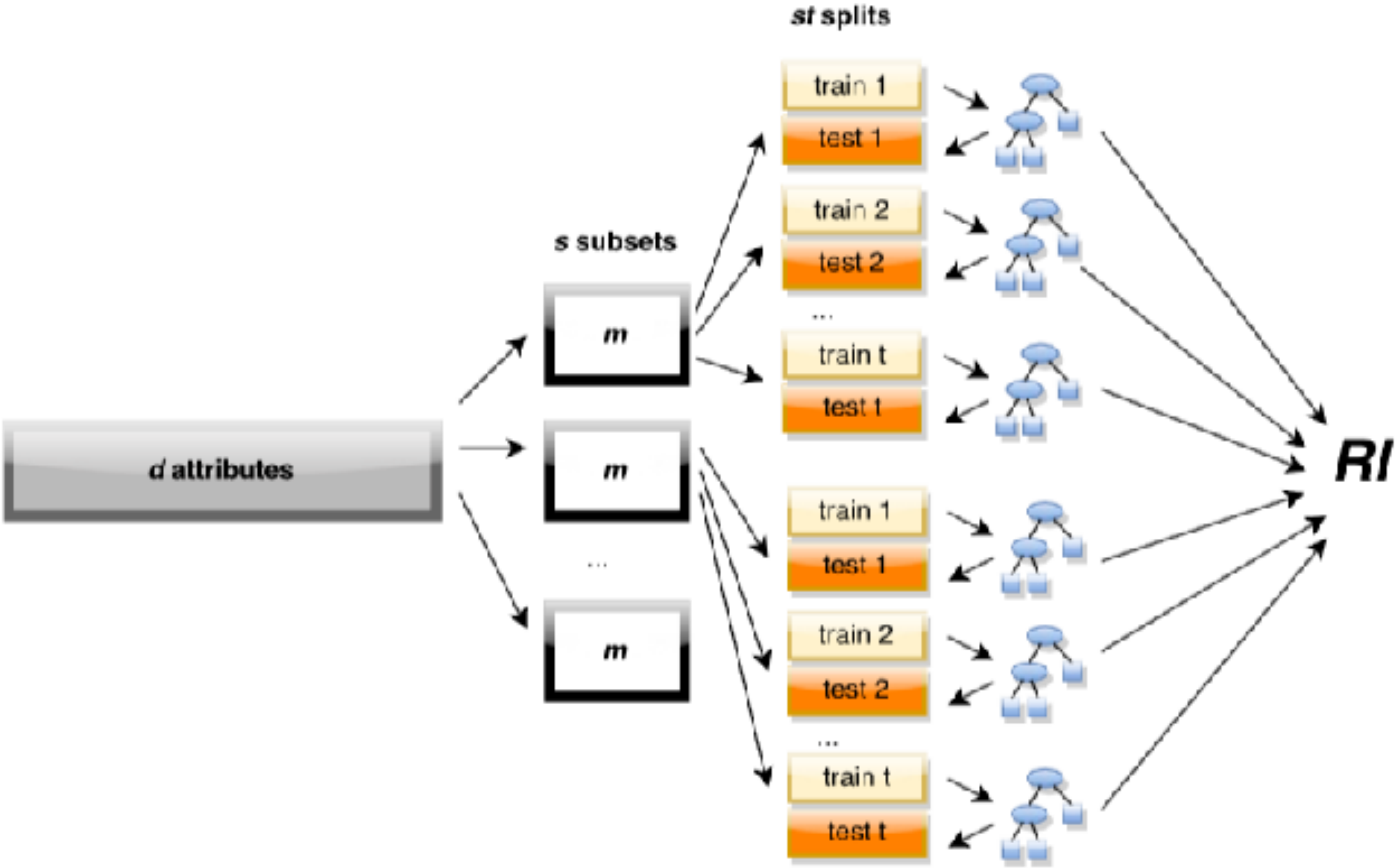


Histogram of survival<=1000

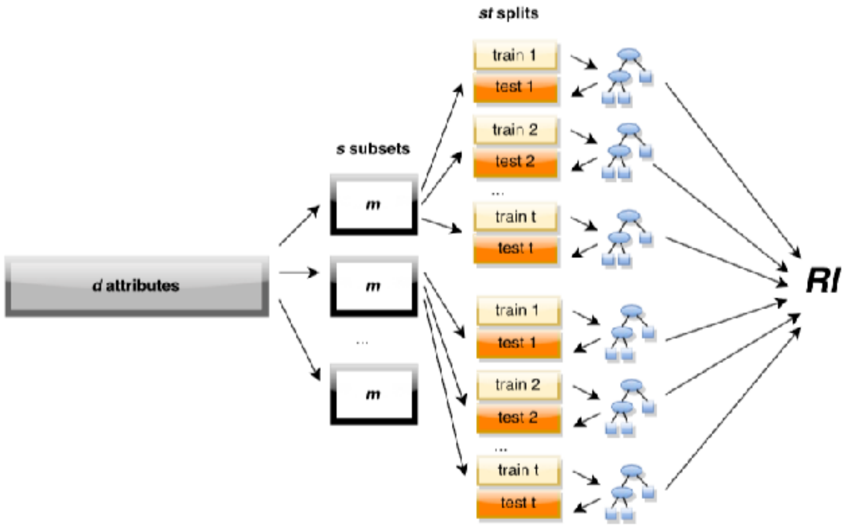
Data analysis scheme

Patient	Cpg1	Cpg2	…	Gene1	Gene2	…	Survival	Decision
TCGA1	0.1	0.5	…	6.9	12	…	560	‘400+’
TCGA2	0.6	0.9	…	5.2	13.6	…	180	‘≤400’
…								
TCGA88	0.34	0.4	…	234	34	…	860	‘400+’

Methods: MCFS-ID



Methods: MCFS-ID

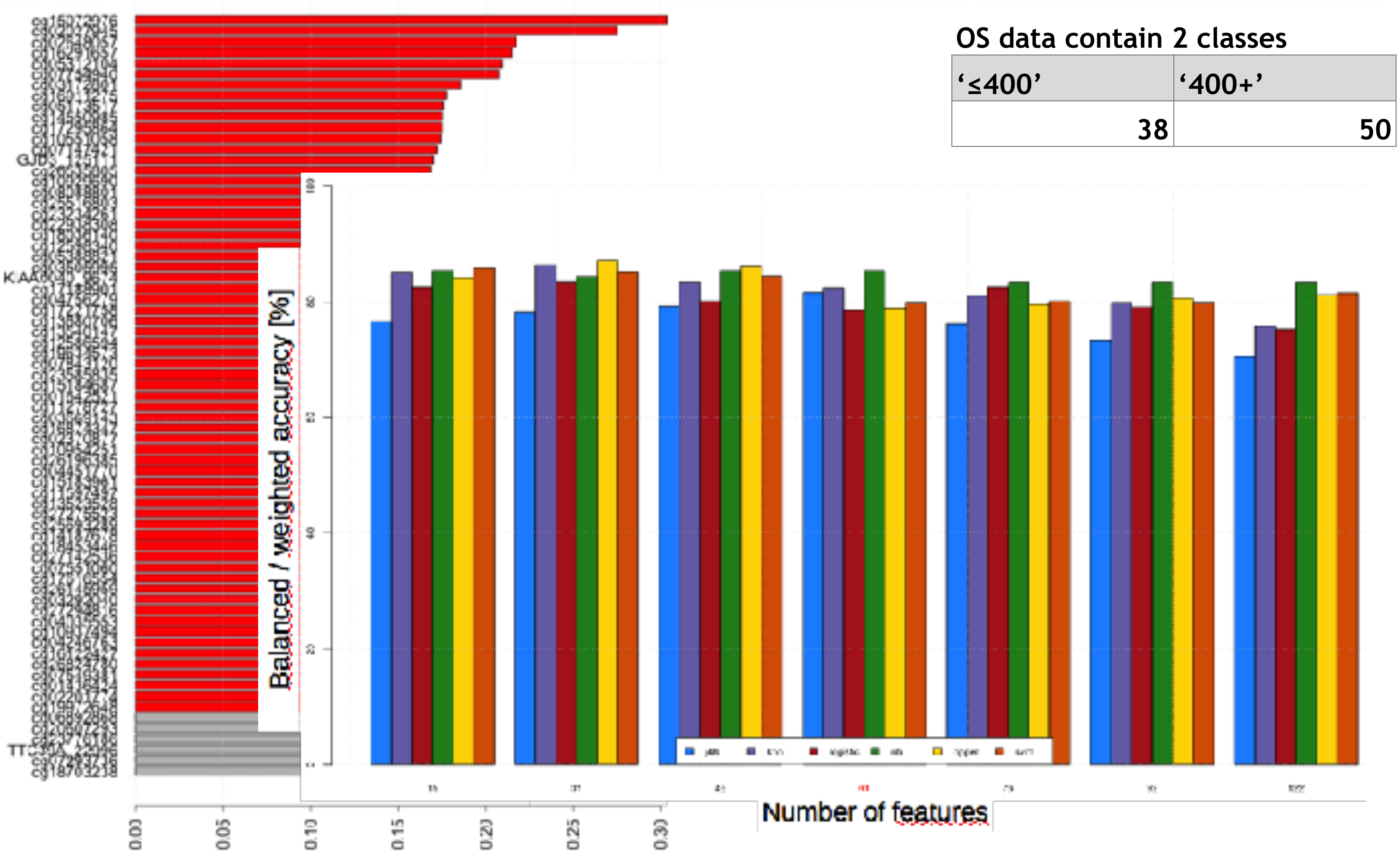


Important attribute:
- occurs in many decision trees (DT),
- is located nearby the root (separates many objects),
- separates classes within the node with high quality,
- DTs based on it perform well on unseen data.

The relative importance of feature g_k, RI_{g_k}, is defined as

$$\mathrm{RI}_{g_k} = \sum_{\tau=1}^{s\cdot t} \mathrm{wAcc}_{\tau}^{u} \sum_{n_{g_k}(\tau)} \mathrm{IG}(n_{g_k}(\tau)) \left(\frac{\text{no. in } n_{g_k}(\tau)}{\text{no. in } \tau}\right)^{v},$$

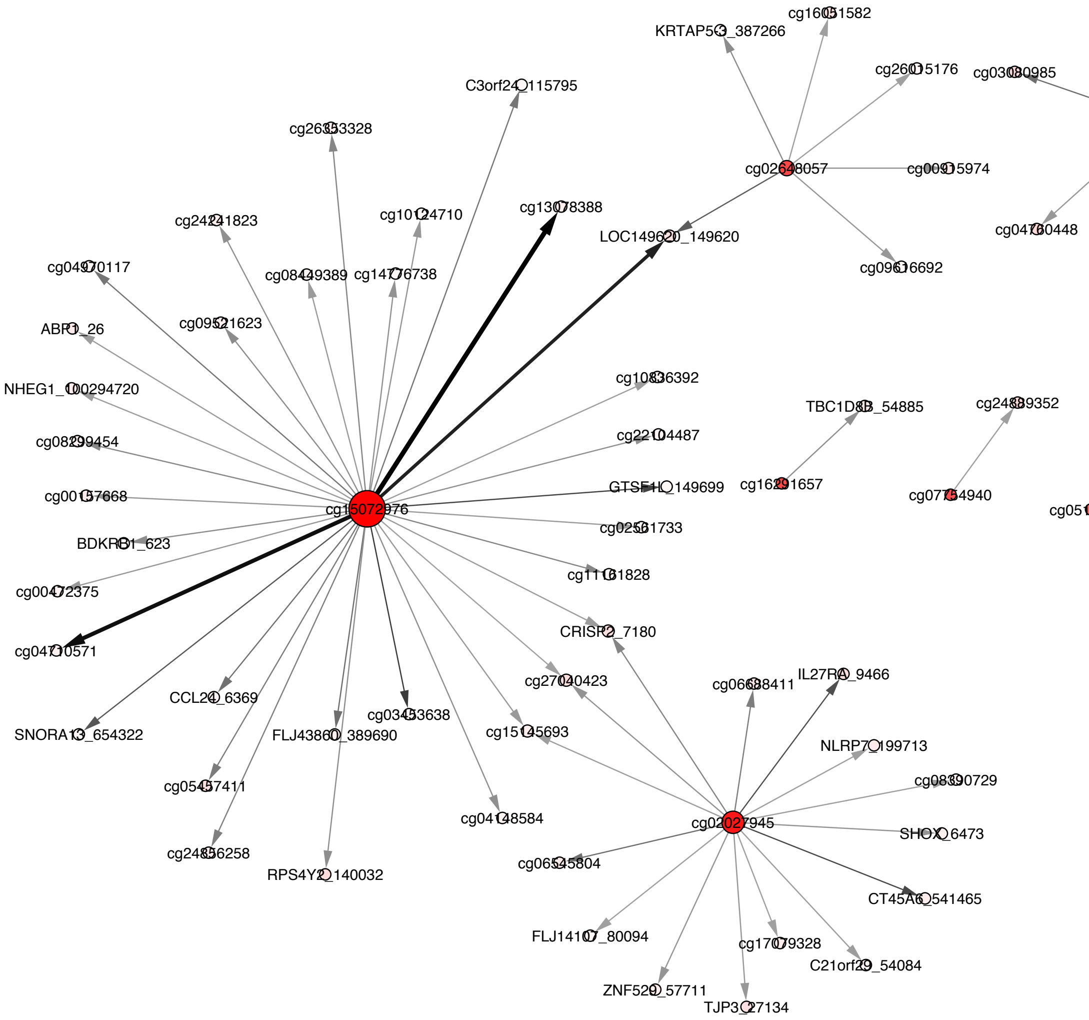
OS - Top Features

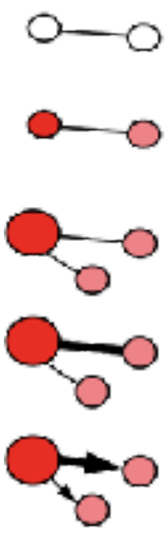


OS data contain 2 classes

'≤400'	'400+'
38	50

Interdependencies Discovery





Two features co-occur along a path in a decision tree, they are considered as interdependent

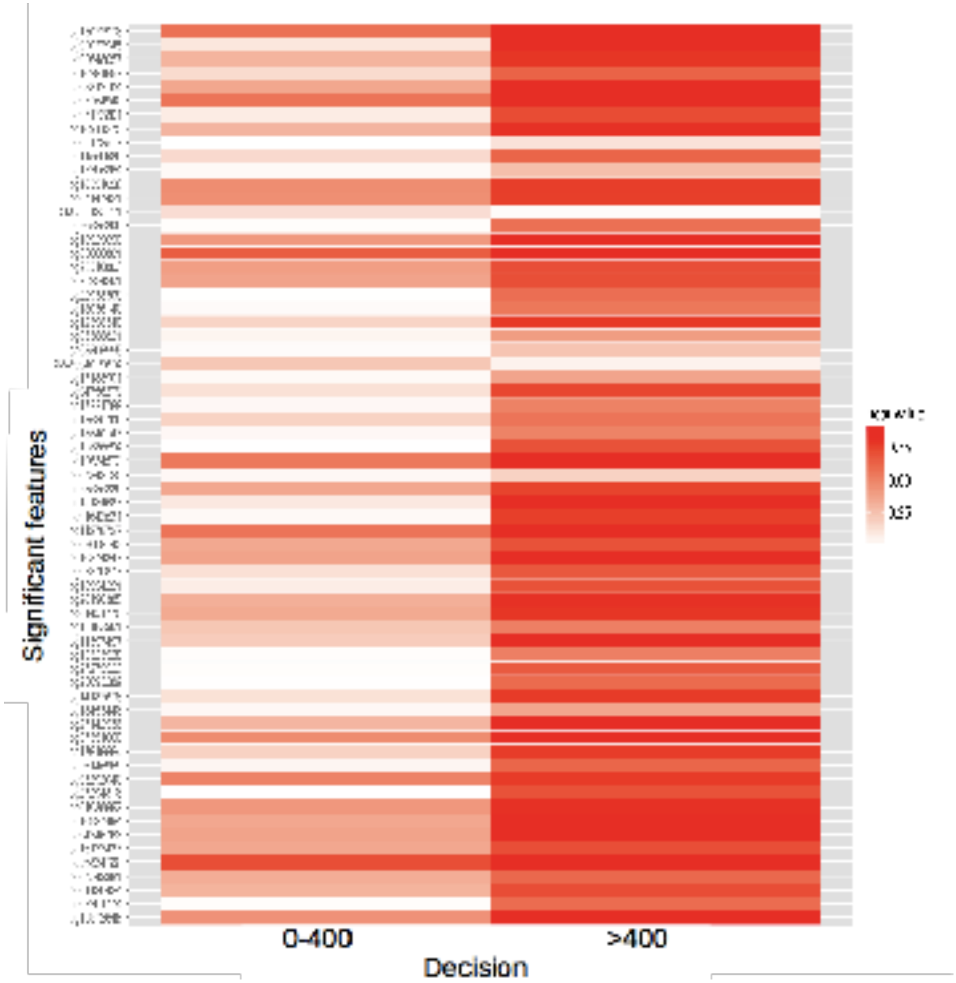
The higher the feature in a decision tree, the stronger it contributes (dark red)

The larger the size of the node the more interdependencies it has

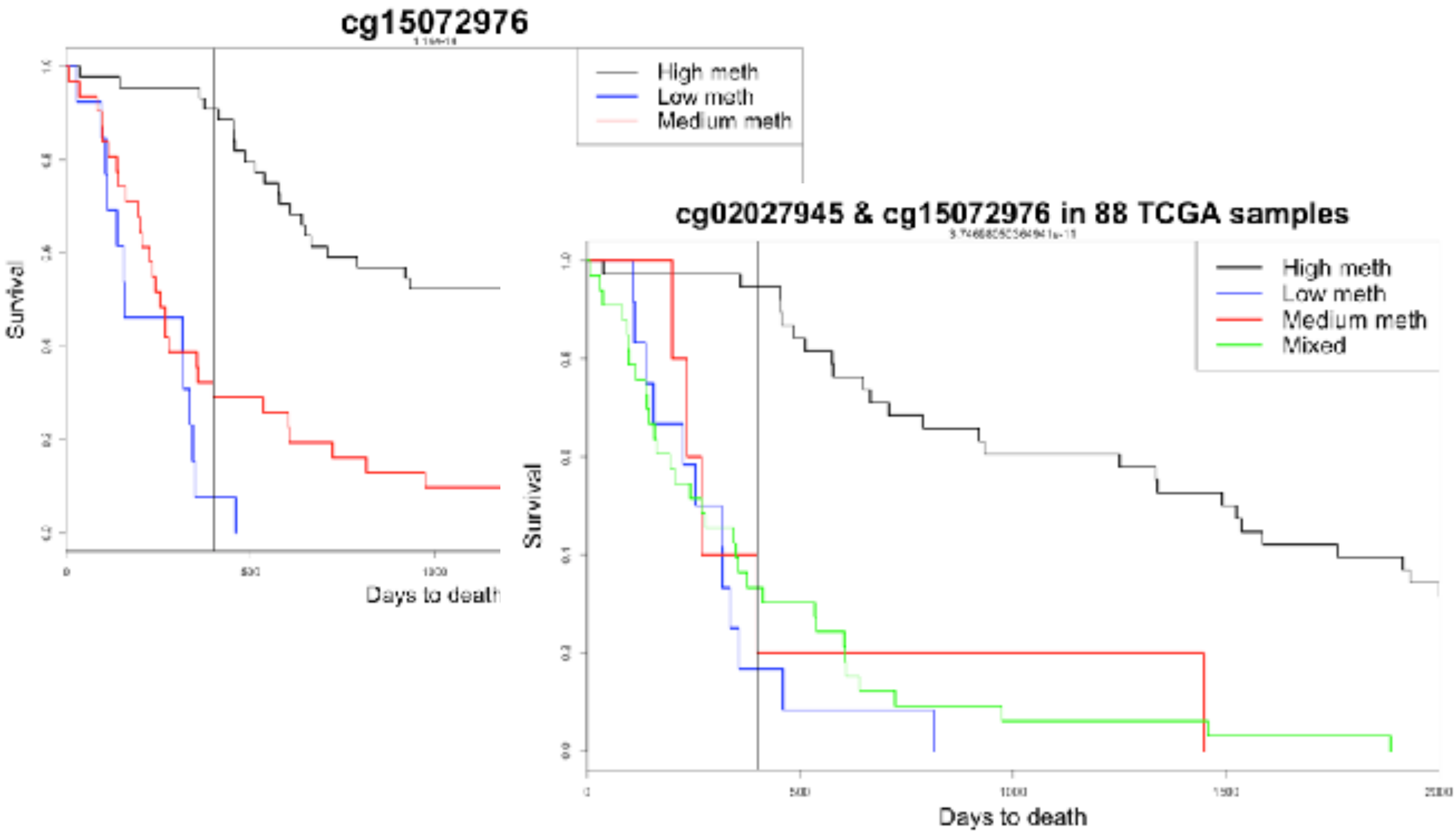
The connection is thicker when a pair occurs more often in the trees

The direction of the connections folows the depth in the three, from higher to lower

Feature values

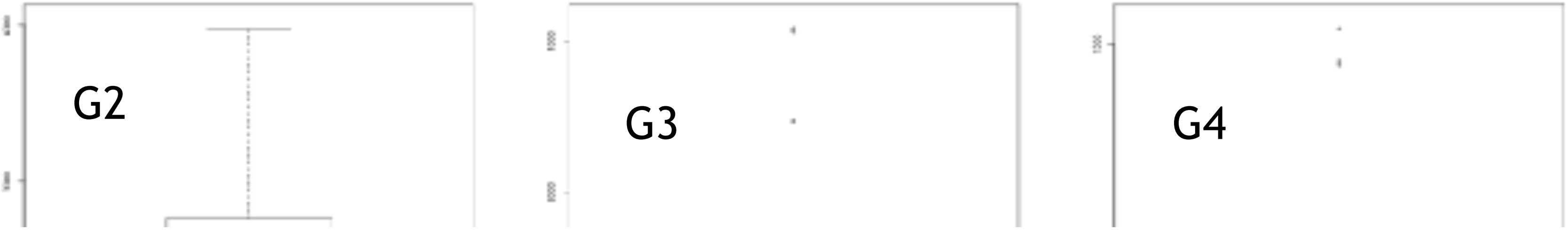


Patients survival


cg15072976

cg02027945 & cg15072976 in 88 TCGA samples

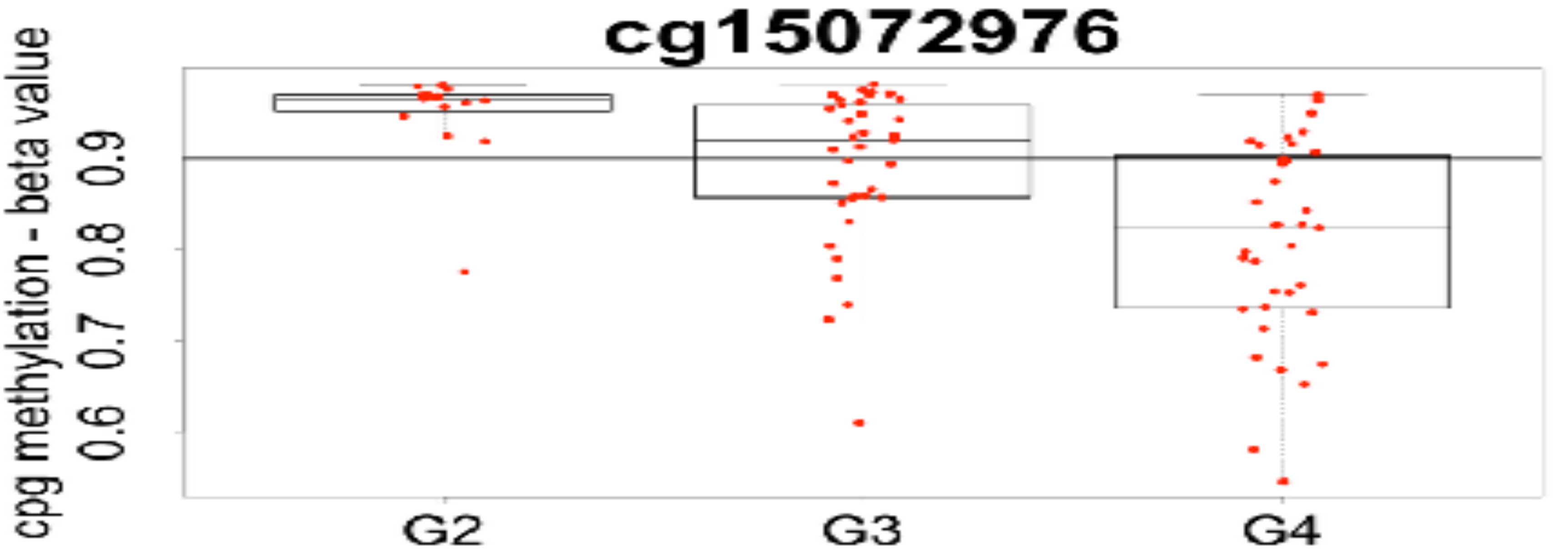
Grade vs top cg



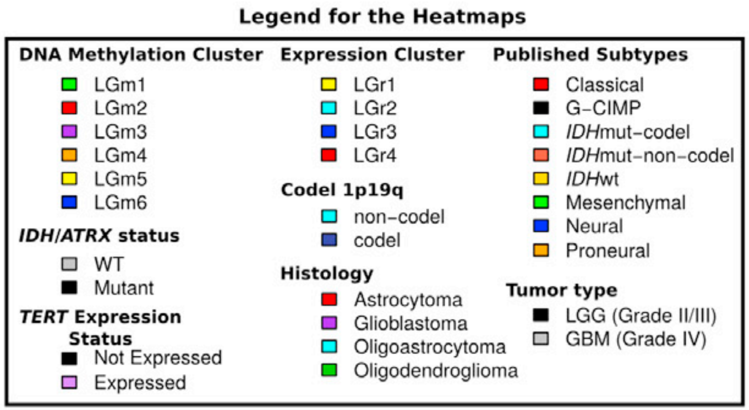
G2

G3

G4



5k top features + markers



Legend for the Heatmaps

DNA Methylation Cluster
- 🟩 LGm1
- 🟥 LGm2
- 🟪 LGm3
- 🟧 LGm4
- 🟨 LGm5
- 🟦 LGm6

***IDH/ATRX* status**
- ⬜ WT
- ⬛ Mutant

***TERT* Expression Status**
- ⬛ Not Expressed
- 🟪 Expressed

Expression Cluster
- 🟨 LGr1
- 🟦 LGr2
- 🟦 LGr3
- 🟥 LGr4

Codel 1p19q
- 🟦 non-codel
- 🟦 codel

Histology
- 🟥 Astrocytoma
- 🟪 Glioblastoma
- 🟦 Oligoastrocytoma
- 🟩 Oligodendroglioma

Published Subtypes
- 🟥 Classical
- ⬛ G-CIMP
- 🟦 *IDH*mut-codel
- 🟥 *IDH*mut-non-codel
- 🟨 *IDH*wt
- 🟩 Mesenchymal
- 🟦 Neural
- 🟧 Proneural

Tumor type
- ⬛ LGG (Grade II/III)
- ⬜ GBM (Grade IV)

OS data contain 2 classes

≤400’	400+’
38	50

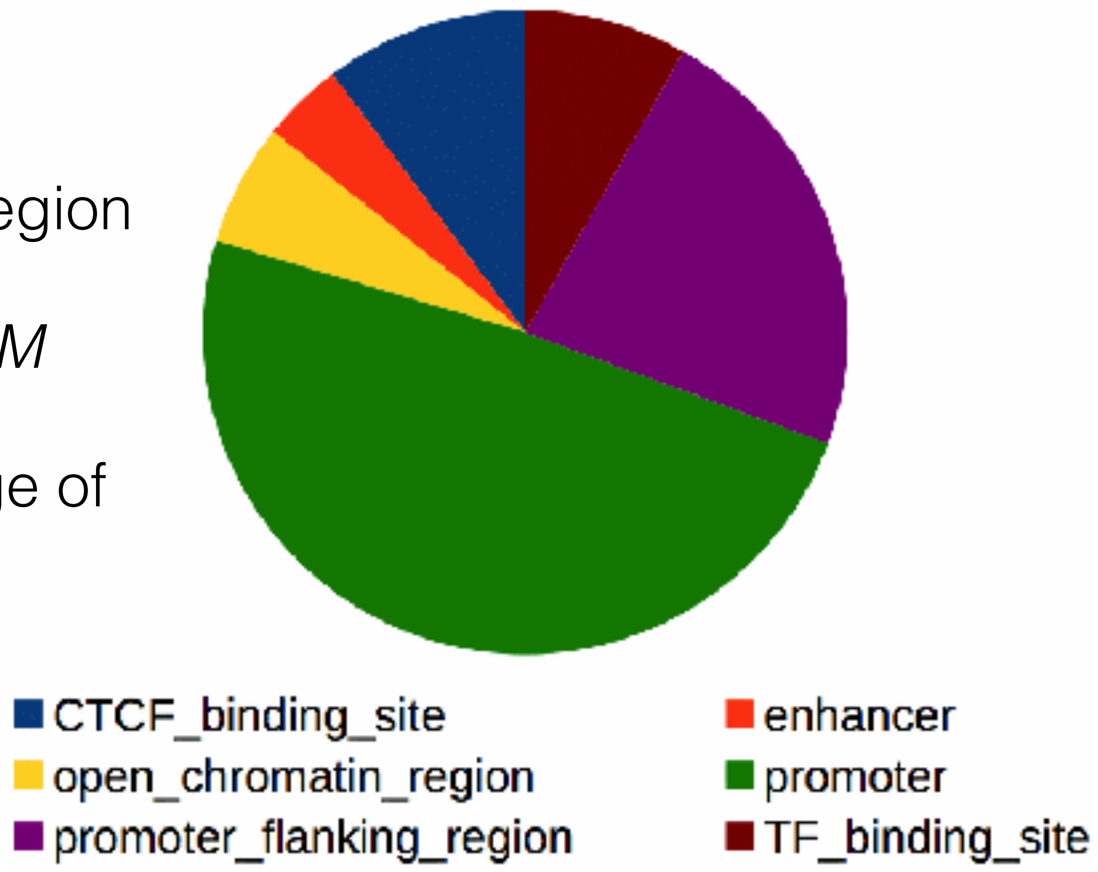
```
      position           attribute projections classifiers nodes      RI_norm
137        137    IDH.codel.subtype          198          239   239 0.035692070
605        605                  Age          203          168   173 0.018395388
835        835          DAXX.status          199          102   102 0.015770035
1143      1143          ATRX.status          193           94    94 0.013475393
1789      1789   X1p.19q.codeletion          202          103   103 0.009829582
1895      1895    BRAF.V600E.status          209           71    71 0.009314941
2030      2030    Chr.19.20.co.gain          224           84    84 0.008713459
2640      2640           IDH.status          180           42    42 0.006419770
3846      3846 BRAF.KIAA1549.fusion          211           28    28 0.002766280
```

Summary

- Methylations dominated the ranking
- Rare cg-gene interactions detected
- Survival was much more strongly associated with methylome patterns than grade
- One dead end reached ~ methylations studied through the function of related genes

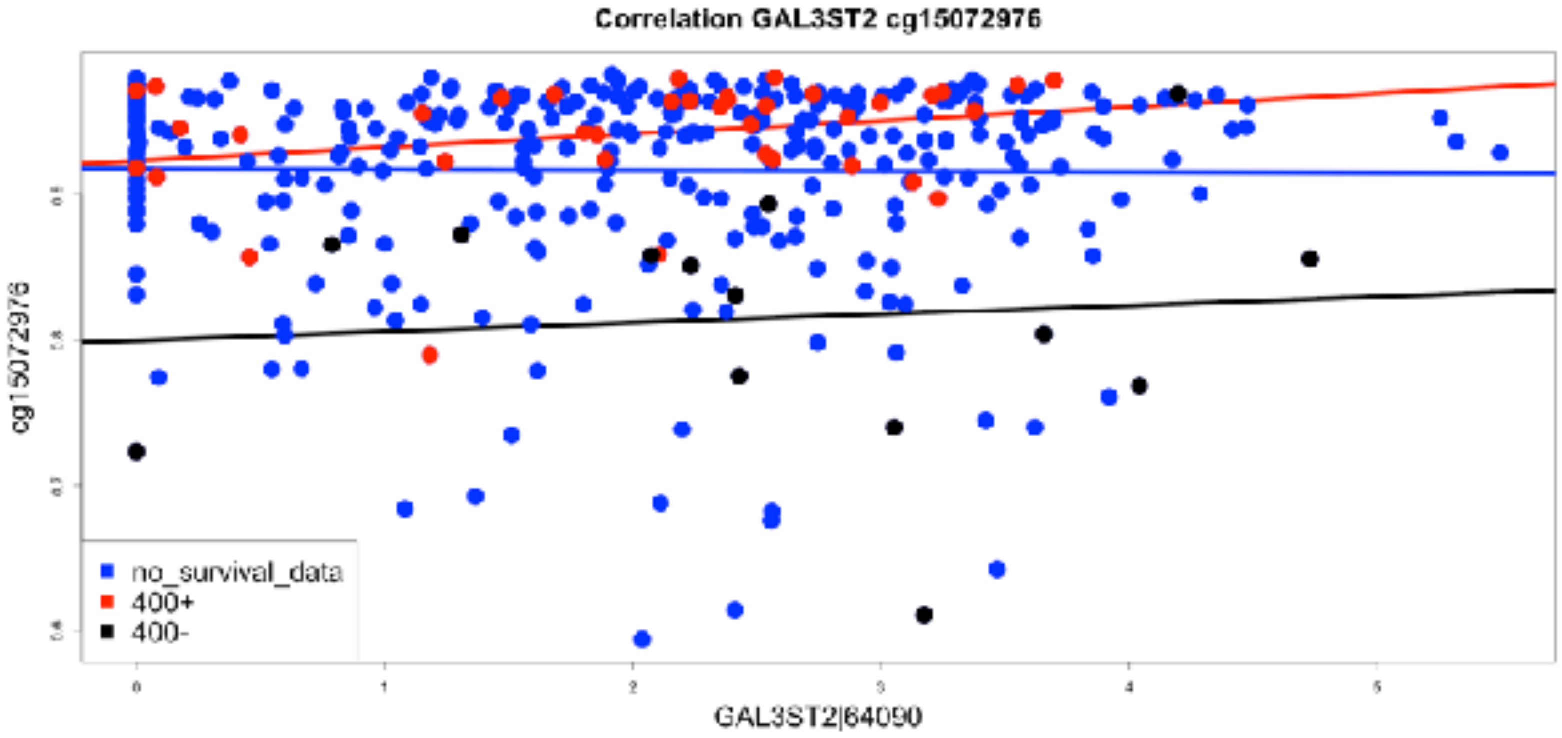
Mapping of 63 methylations

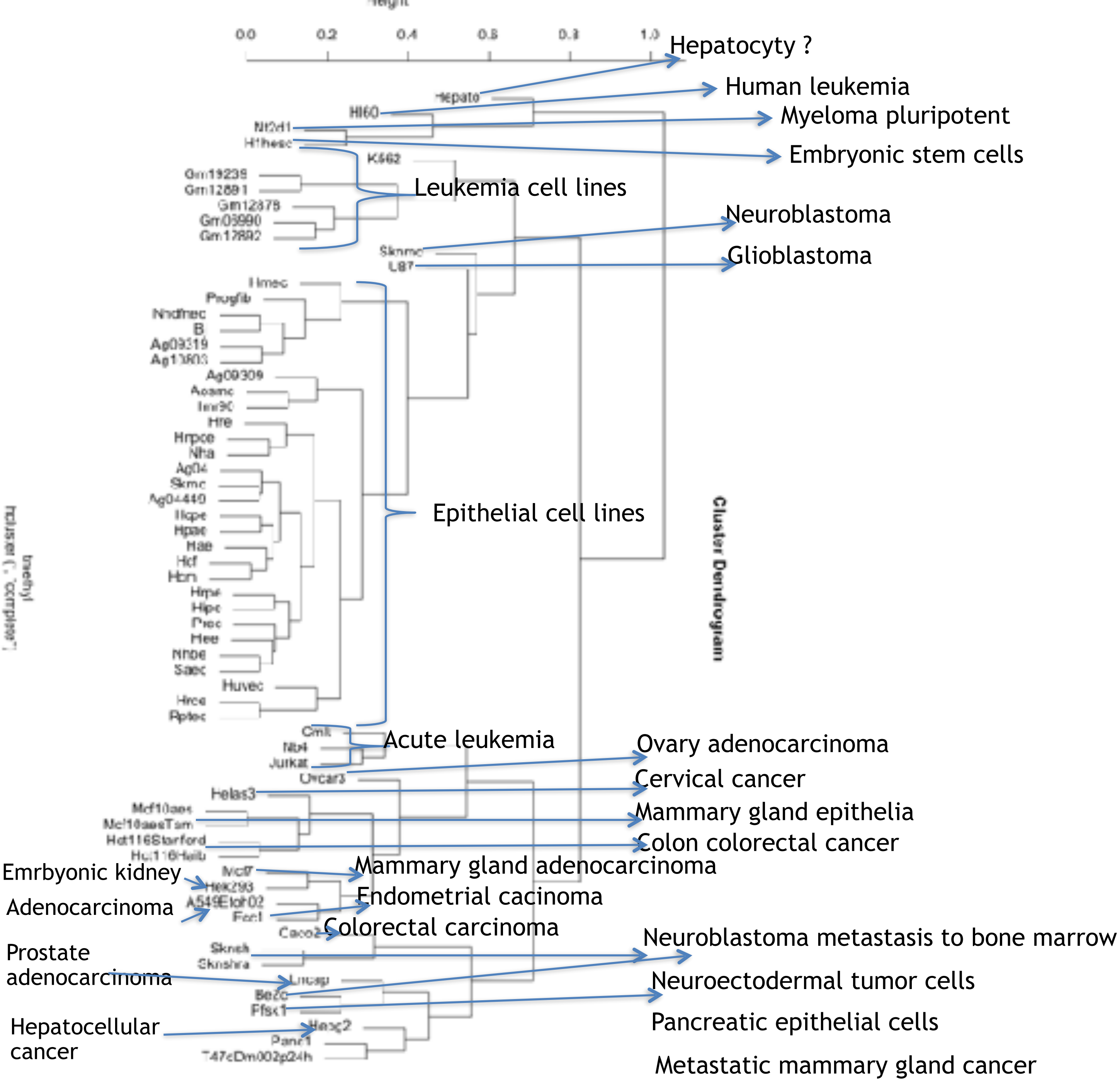
- 44 out of 63 were mapped to a gene
- 7 methylations within a region of 865 bp were mapped to *MYADM*
- 2 methylations in a range of 287 bp were mapped to *TBR1*



- Map cg to active enhancers/promoters

Testing associations





Hepatocyty ?

Human leukemia

Myeloma pluripotent

Embryonic stem cells

Leukemia cell lines

Neuroblastoma

Glioblastoma

Epithelial cell lines

Acute leukemia

Ovary adenocarcinoma

Cervical cancer

Mammary gland epithelia

Colon colorectal cancer

Emrbyonic kidney

Mammary gland adenocarcinoma

Adenocarcinoma

Endometrial cacinoma

Colorectal carcinoma

Prostate adenocarcinoma

Neuroblastoma metastasis to bone marrow

Neuroectodermal tumor cells

Hepatocellular cancer

Pancreatic epithelial cells

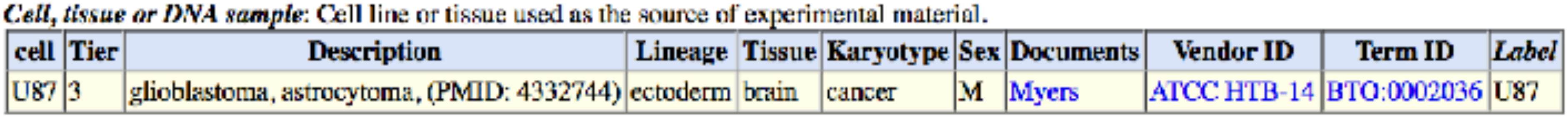
Metastatic mammary gland cancer

TFBS

Assays for REST in Cluster

#	signal	abr	cellType	factor	antibody	treatment	lab	more info
1	172.00	u	U87	REST	NRSF	None	HudsonAlpha	metadata ▾

Neuron-restrictive silencer transcription factor, represses neuronal genes in non-neuronal tissues, member of the Kruppel-type zinc finger family



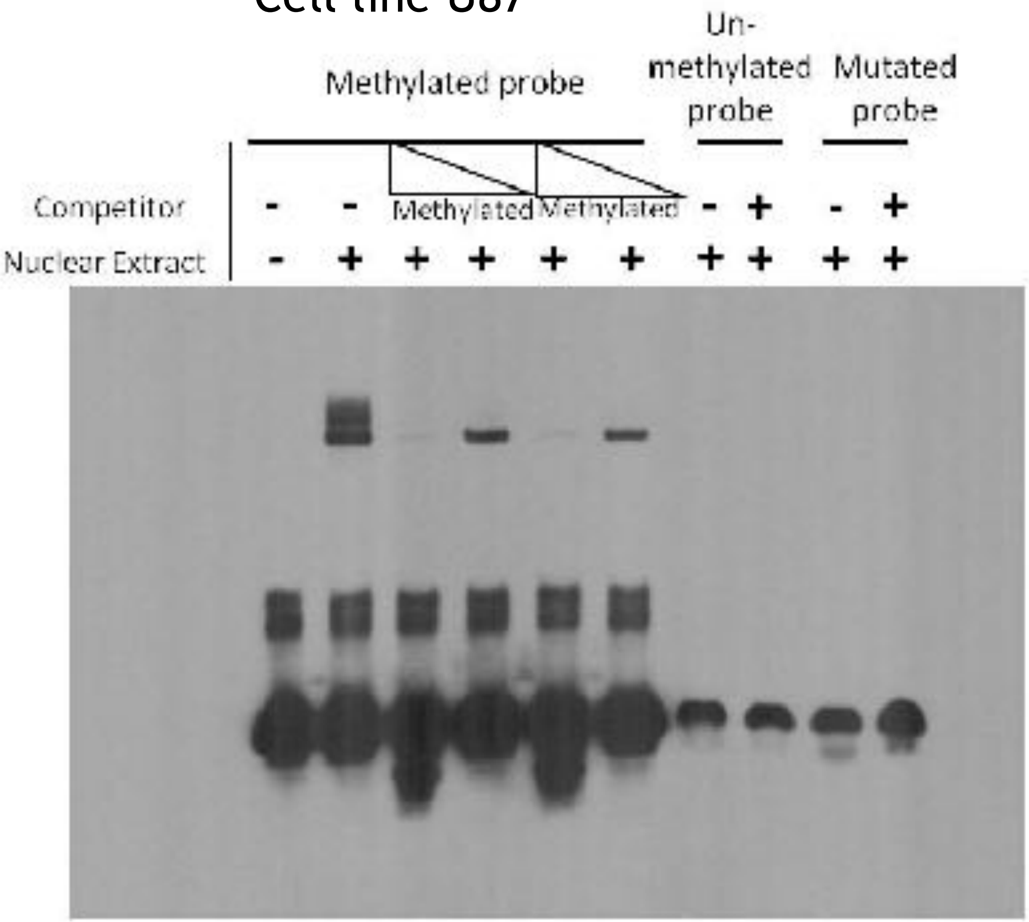
***Cell, tissue or DNA sample*:** Cell line or tissue used as the source of experimental material.

cell	Tier	Description	Lineage	Tissue	Karyotype	Sex	Documents	Vendor ID	Term ID	*Label*
U87	3	glioblastoma, astrocytoma, (PMID: 4332744)	ectoderm	brain	cancer	M	Myers	ATCC HTB-14	BTO:0002036	U87

MOTIF	WIDTH	BEST POSSIBLE MATCH
REST_f1	22	GGGCGCTGTCCATGGTGCTGAA

Electrophoretic mobility shift assay

Cell line U87

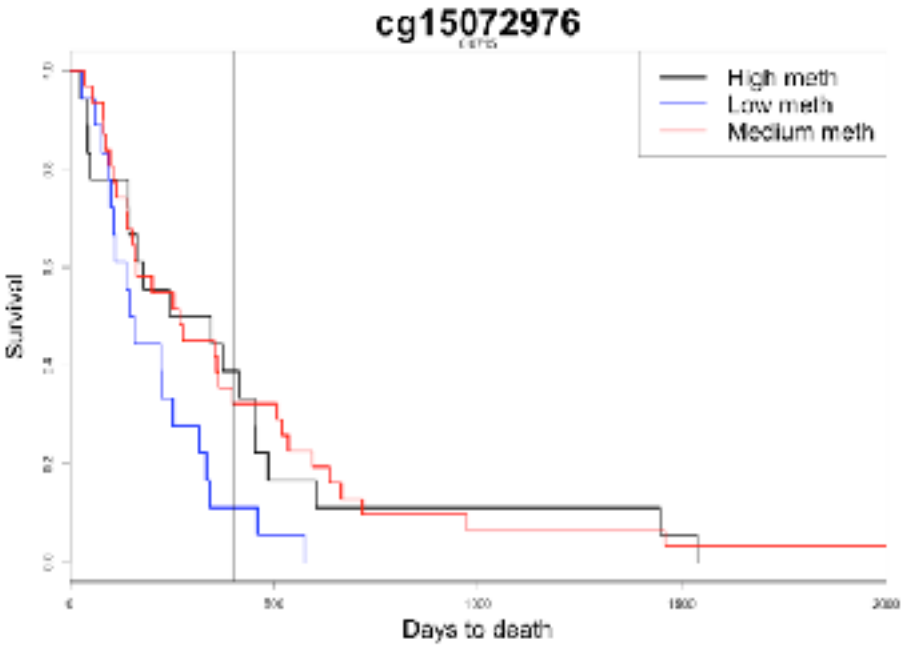


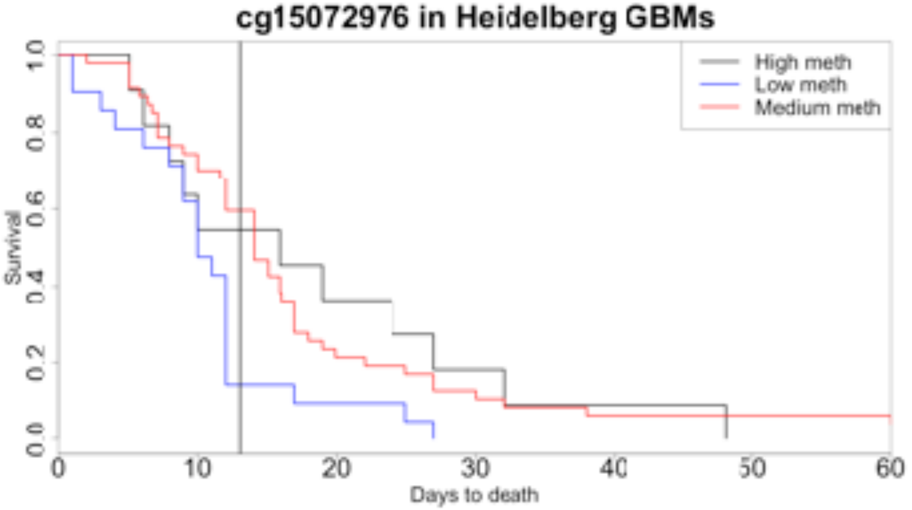
There are additional 9 REST binding sites

Back to TCGA

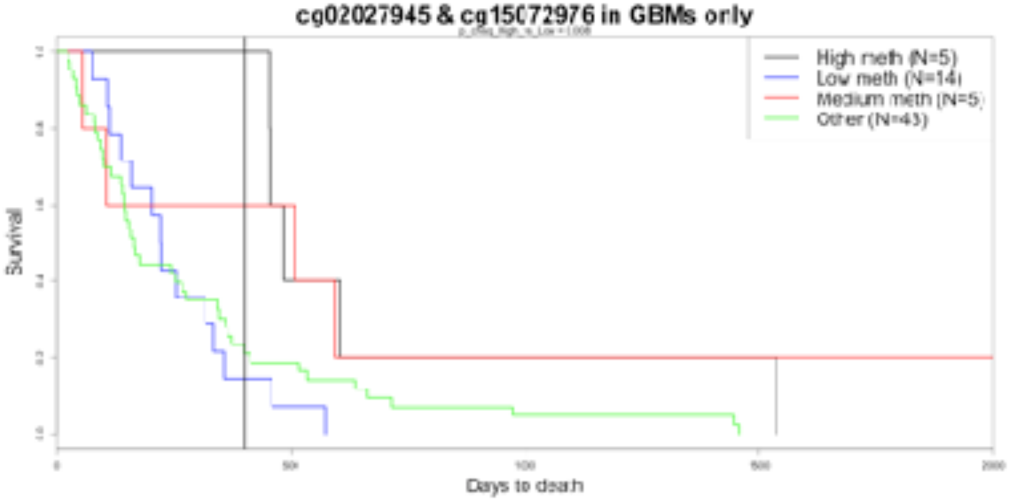
- New data 167 patients with known OS

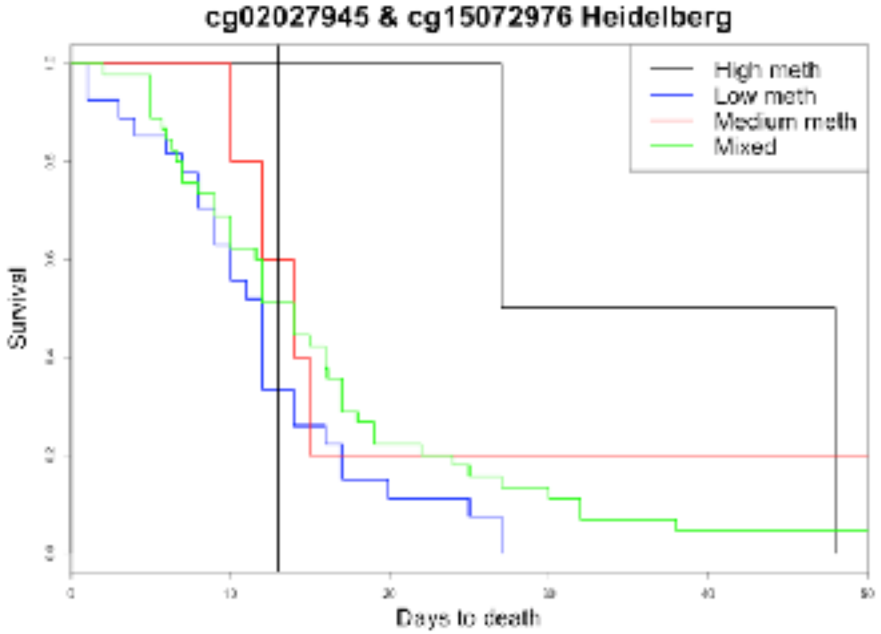
Heildenberg





Heildenberg





The Data

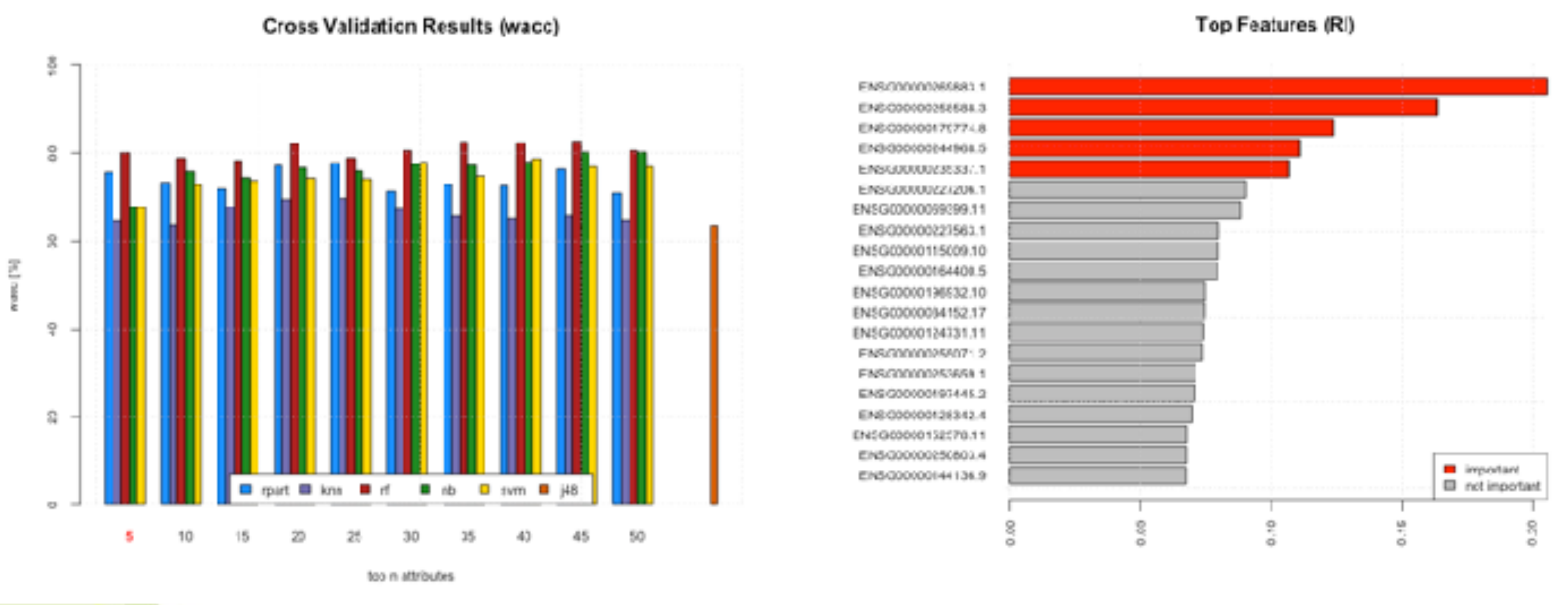
TCGA Data

- Patients: **167**
- Attributes: **58191**
    - Fenotype/genotype features **41**
- Decision
    - **Survival** (0_400, 400+)
    - **IDH.status** (wt, mutant)

Fenotype/genotype features

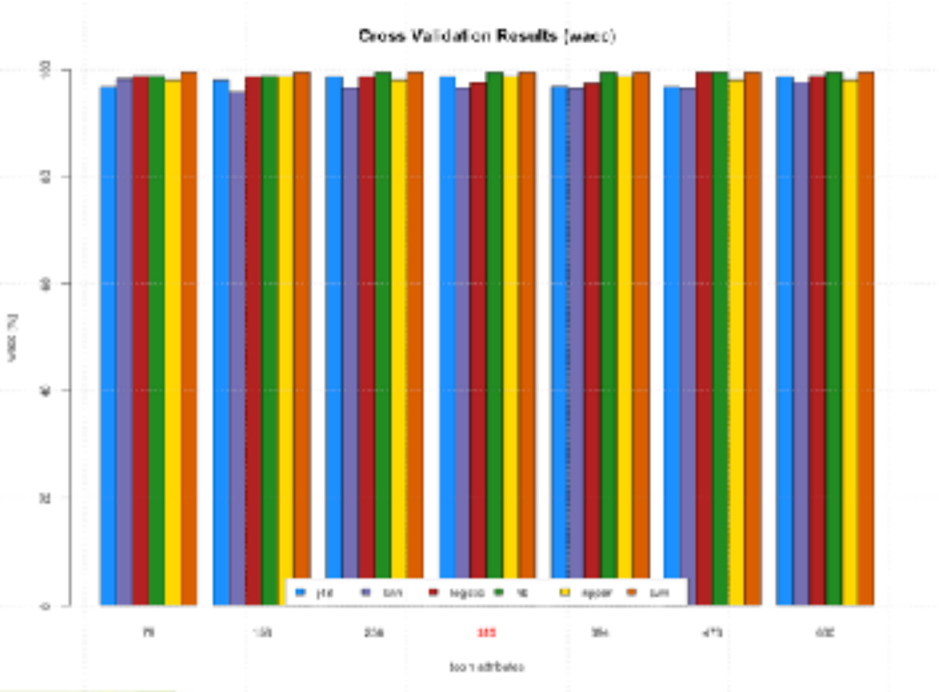
```
 [1] "BCR"                                     "Histology"
 [3] "Grade"                                   "Age"
 [5] "Gender"                                  "Karnofsky.Performance.Score"
 [7] "Mutation.Count"                          "Percent.aneuploidy"
 [9] "IDH.status"                              "X1p.19q.codeletion"
[11] "IDH.codel.subtype"                       "MGMT.promoter.status"
[13] "Chr.7.gain.Chr.10.loss"                  "Chr.19.20.co.gain"
[15] "TERT.promoter.status"                    "TERT.expression.log2"
[17] "TERT.expression.status"                  "ATRX.status"
[19] "DAXX.status"                             "Telomere.Maintenance"
[21] "BRAF.V600E.status"                       "BRAF.KIAA1549.fusion"
[23] "ABSOLUTE.purity"                         "ABSOLUTE.ploidy"
[25] "ESTIMATE.stromal.score"                  "ESTIMATE.immune.score"
[27] "ESTIMATE.combined.score"                 "Original.Subtype"
[29] "Transcriptome.Subtype"                   "Pan.Glioma.RNA.Expression.Cluster"
[31] "IDH.specific.RNA.Expression.Cluster"     "Pan.Glioma.DNA.Methylation.Cluster"
[33] "IDH.specific.DNA.Methylation.Cluster"    "Supervised.DNA.Methylation.Cluster"
[35] "Random.Forest.Sturm.Cluster"             "RPPA.cluster"
[37] "Telomere.length.estimate.in.blood.normal.Kb" "Telomere.length.estimate.in.tumor.Kb"
[39] "diagnosis"                               "race"
[41] "ethinicity”
```

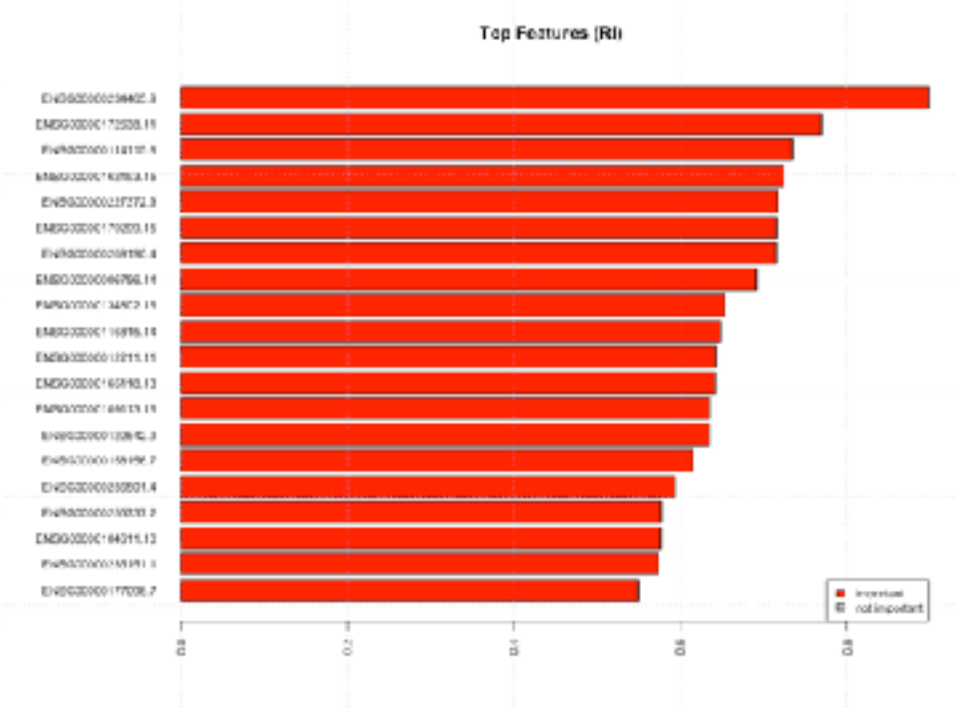
Survival MCFS results



position	attribute	RI
56	**IDH.status**	0.0434345804
557	**IDH.codel.subtype**	0.0151150247
4020	Chr.19.20.co.gain	0.0048719168
4181	TERT.expression.status	0.00477300010
4617	DAXX.status	0.00450833711
5012	BRAF.V600E.status	0.0043099263
5498	X1p.19q.codeletion	0.004068978

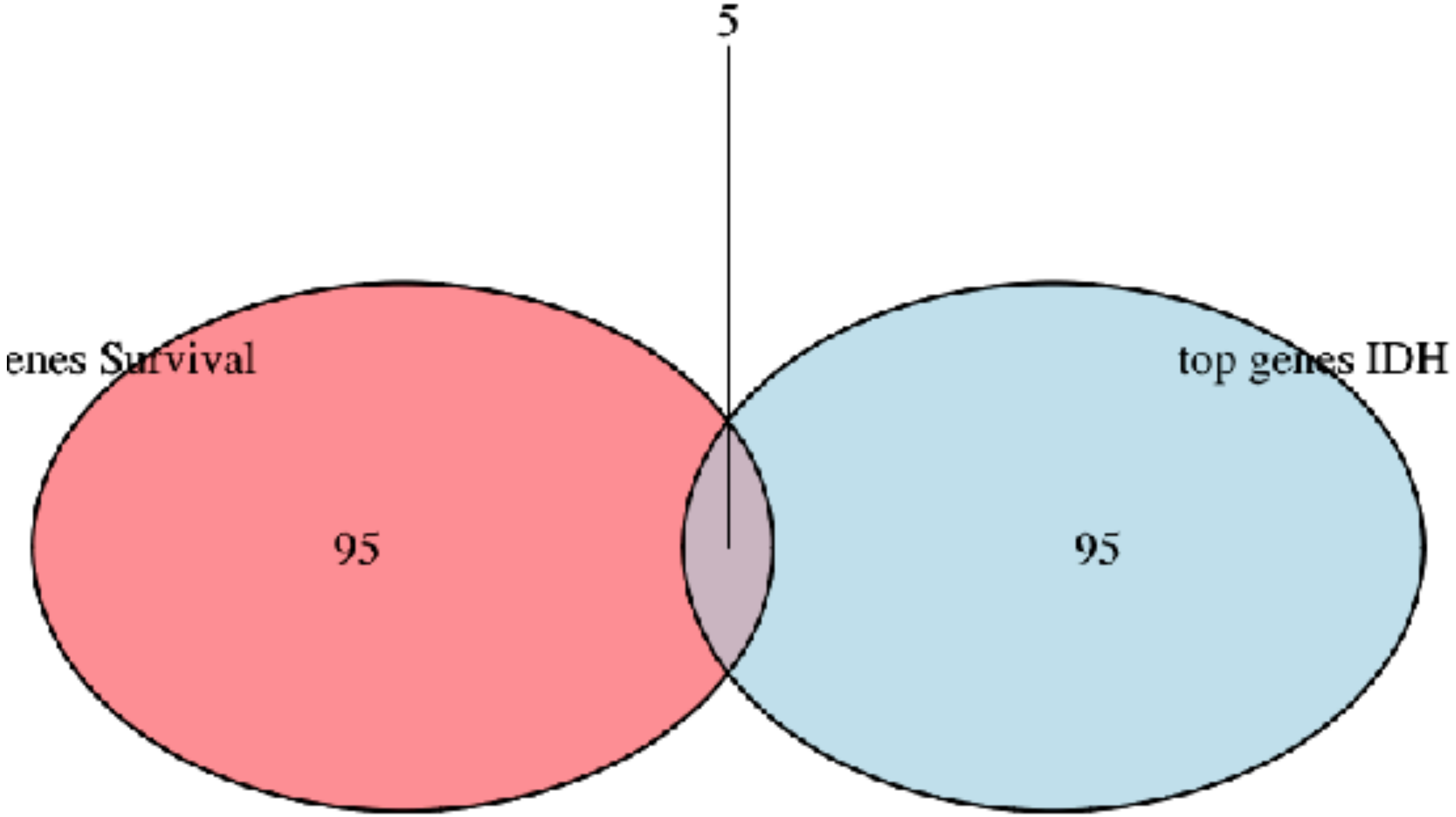
IDH.Status MCFS results



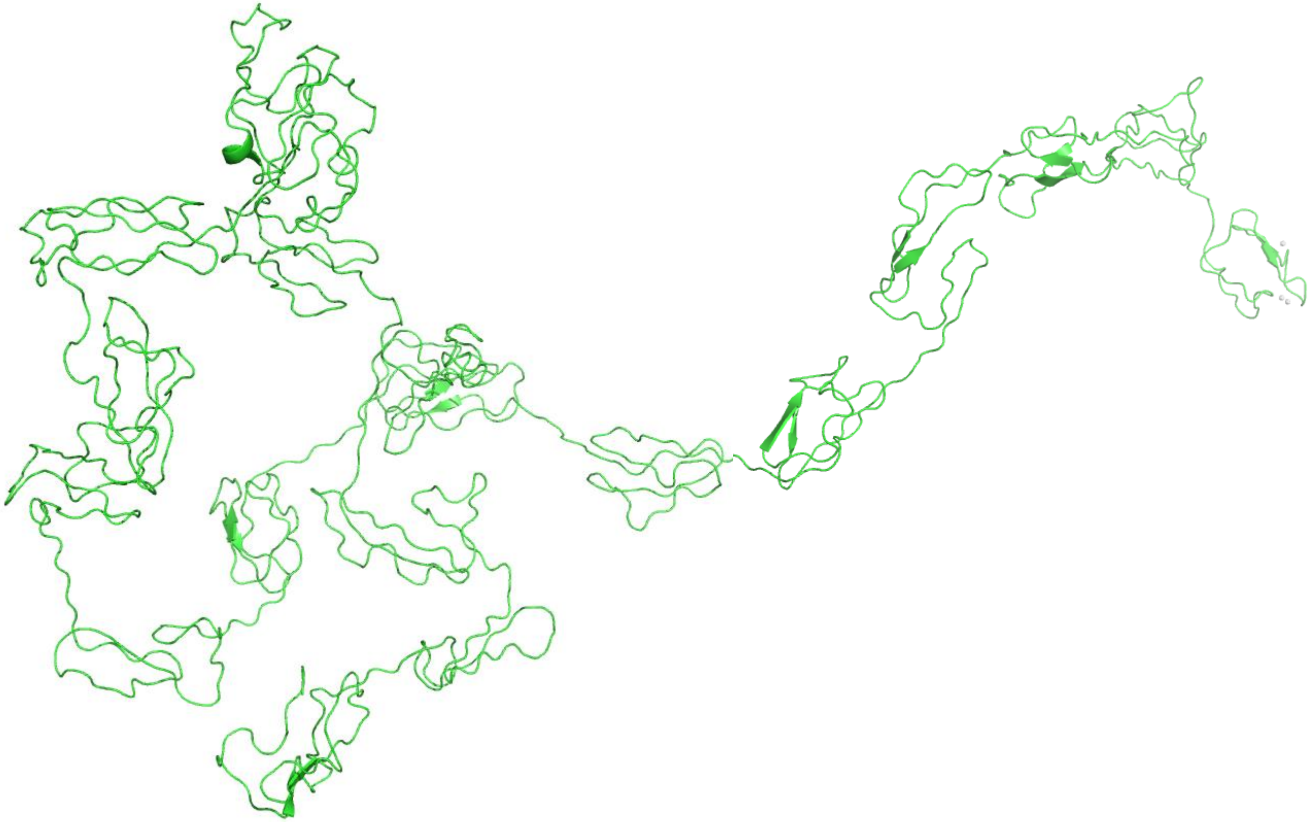


position	attribute	RI
103	Chr.7.gain.Chr.10.loss	0.24287560013
217.	Random.Forest.Sturm.Cluster	0.1204359708
674	Pan.Glioma.RNA.Expression.Cluster	0.0284643536
744	ATRX.status	0.02515611812
812	Supervised.DNA.Methylation.Cluster	0.0220223947
860	Original.Subtype	0.020578600

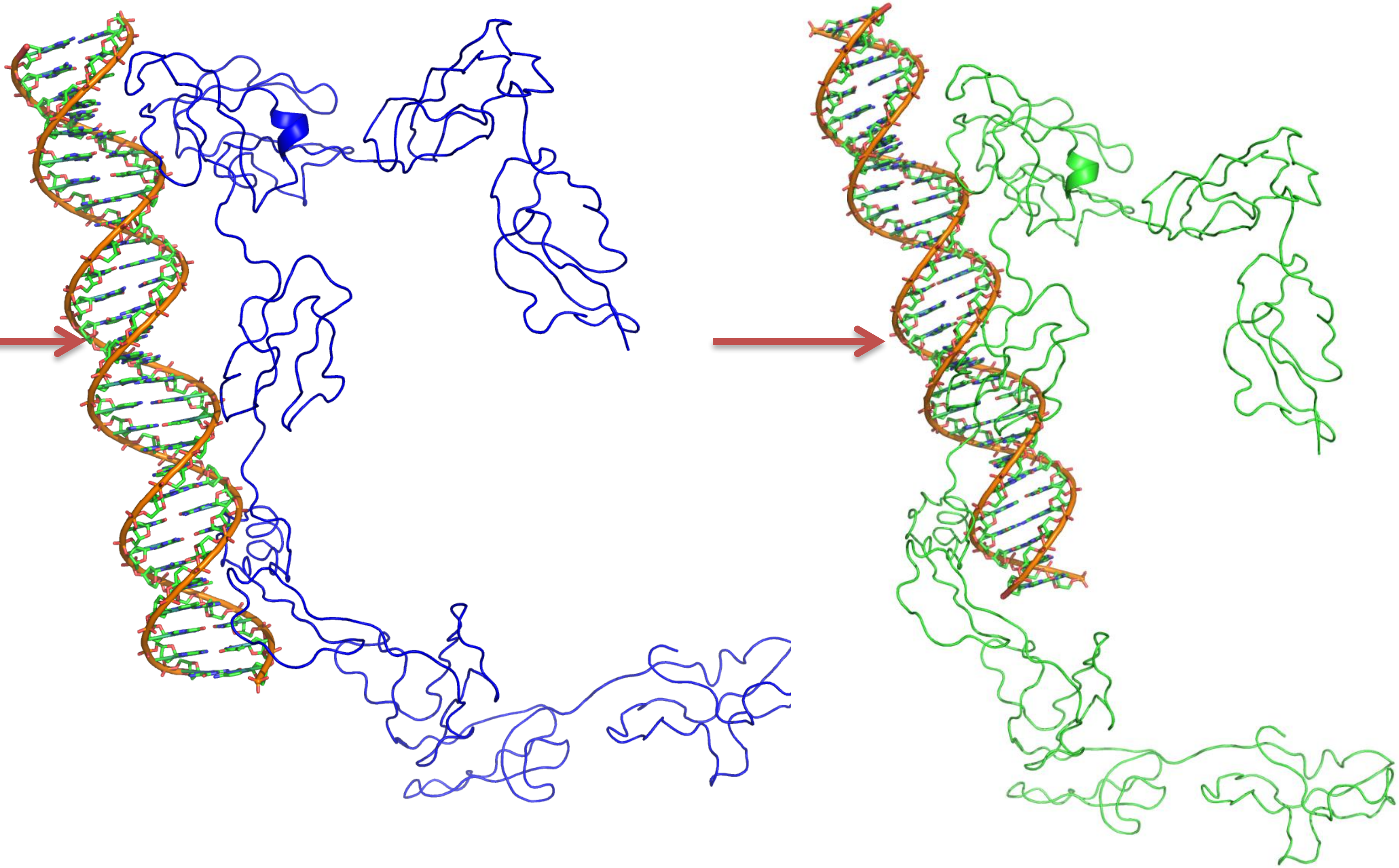
Survival vs IDH.Status



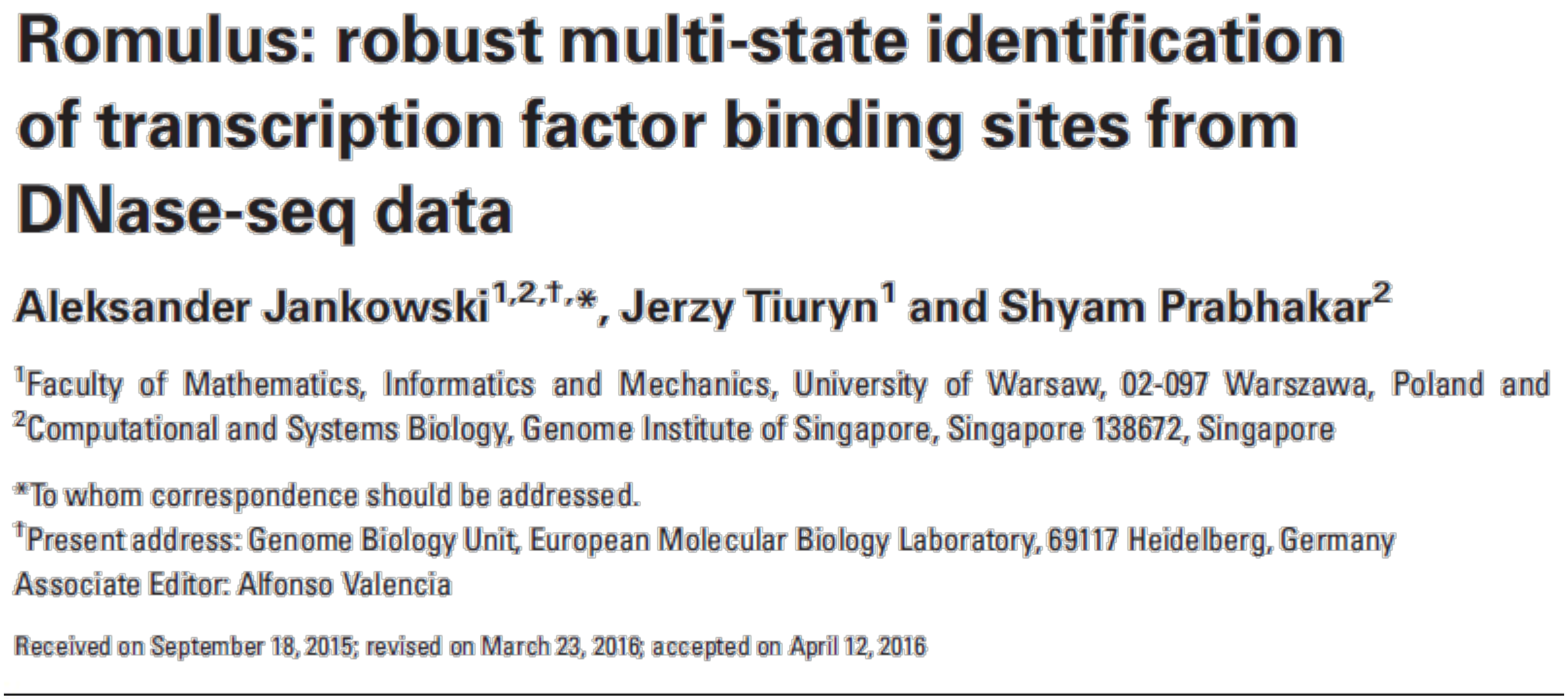
Modeling REST structure



DNA + REST



DNasal and TF binding site



Romulus: robust multi-state identification of transcription factor binding sites from DNase-seq data

Aleksander Jankowski[1,2,†,*], Jerzy Tiuryn[1] and Shyam Prabhakar[2]

[1]Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, 02-097 Warszawa, Poland and [2]Computational and Systems Biology, Genome Institute of Singapore, Singapore 138672, Singapore

*To whom correspondence should be addressed.

†Present address: Genome Biology Unit, European Molecular Biology Laboratory, 69117 Heidelberg, Germany

Associate Editor: Alfonso Valencia

Received on September 18, 2015; revised on March 23, 2016; accepted on April 12, 2016

Acknowledgments

Institute of Computer Science
Michał Dramiński
Magdalena Mozolewska
Jacek Koronacki
Paweł Teisseyre
Jan Komorowski

Uppsala University
Klev Diamanti

Nencki Institute of Experimental Biology
Bartosz Wojtas
Karolina Stępniak
Jakub Mieczkowski
Bożena Kamińska

University of Warsaw, Faculty of Mathematics, Informatics, and Mechanics
Ania Macioszek
Bartosz Wilczyński