Canonical Correlation- based bioinformatic analysis for effective melanoma biomarker discovery

Sonia Wróbel [1], Ewa Stępień [1], Cezary Turek [2], Monika Piwowar [2]

1 Department of Medical Physics, Jagiellonian University, Marian Smoluchowski Institute of Physics, Kraków, Poland, 2 Department of Bioinformatics and Telemedicine, Jagiellonian University–Medical College, Krakow, Poland

ABSTRACT

- Here we introduce a new method based on canonical correlation analysis (CCA) that uses real-life dataset to meet the challenge of **melanoma biomarker discovery** [1-2]. The bioinformatics pipeline was successfully applied to human skin **melanoma multi-OMICS datasets** containing: **(1)** microvesicle micro-RNA transcriptomics, **(2)** microvesicle proteomics, **(3)** cell-total-RNA transcriptomics.
- The method applies a **sparse CCA (sCCA)** to three matrices, starting from features correlation across integrated experimental data [3].
- Validation using clinical data as well as supporting meta-data from extracellular vesicle dedicated databases allows the identification of evidence-based candidates for highly **significant molecular signatures** like **melanoma-associated microRNAs and oncoproteins**.

CHALLENGE

- Next Generation Sequencing (NGS) and other advanced large-scale experimental methods provide enormous amounts of **multi-dimensional biological data**. Understanding the interactions between transcriptomics, proteomics and other types of data generated using different platforms is fundamental. In such analyzes, the **integration of multiple OMICS** datasets together and selection of variables is a key to obtain **interpretable results**.
- **Canonical Coronation Analysis (CCA) is one of the most powerful method** for this bioinformatic challenge. Over the last years, a number of promising results for implementing CCA in the integration of OMICS data have been proposed [4-5].

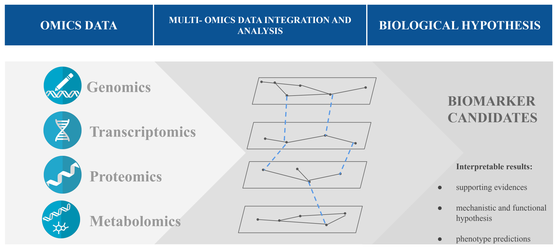


Fig. 1 Multi-omics data integration and analyses as effective method for identification of the biomarker candidates using information of biological interrelationships, bioactive molecules and their functions.

MELANOMA MODEL

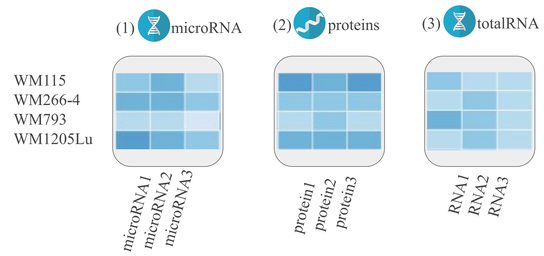
- We used two melanoma cell line models:
  - WM115: a primary vertical growth phase cell line and WM266-4: a lymph node metastasis vertical growth phase cell line. Bothe established from the same patient.
  - WM793: a primary vertical growth phase cell line and WM1205Lu: a metastatic vertical growth phase cell line. First established from patient and second from nude mice lung metastases.

Fig. 2 Melanoma cell lines: WM115, WM266-4, WM793, WM1205Lu originated from the European Searchable Tumour Cell Line and Data Bank (ESTDAB)- A Collection of Immunologically Characterised Melanoma Cell Lines and Databank (Tübingen, Germany).

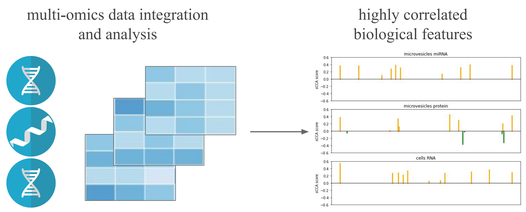
METHOD OVERVIEW

- As an input data we used proprietary microvesicle micro-RNA transcriptome and open source datasets for microvesicle proteome and cells total-RNA transcriptome [6-7]. Each data type was derived for standardized cell lines: WM115, WM266-4, WM793 and WM1205-Lu.
- Data analysis and interpretation was done using method based on sparse canonical correlation bioinformatics method developed in our research group (Fig. 2).
- To conduct sparse CCA we use matrices which represent different sets of features (1) microvesicle micro-RNA transcripts, (2) microvesicle proteins and (3) cell-total-RNA transcripts. on the same set of melanoma cell lines samples. Multi-OMICS dataset has samples in rows and the features on columns. Prepared matrices always had the same number of rows, but had different numbers of columns.
- In next step there was the visualization of highest correlated features and a list of this features with respective ranks.
- Last step provided pathways analysis and annotations supporting each functional insight from extracellular dedicated databases.

a. Input data:



b. Sparse Canonical Correlation Analysis (sCCA)



c. Results: biomarker candidates with supporting biological findings

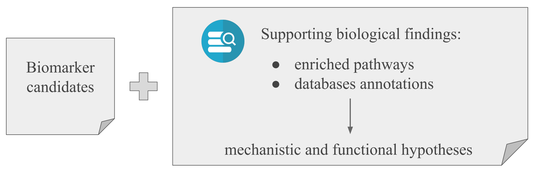


Fig. 3 Method overview. a) Method requires three input matrices for different genomics features for the same set of samples. In this study we used (1) microvesicle-micro-RNA transcripts, (2) microvesicle proteins and (3) cell-total-RNA transcripts for four melanoma cell lines models: WM115, WM266-4, WM793 and WM1207Lu. **b)** Method provides visualization of highest correlated features and a list of this features with ranks. **c)** Last step provides pathways analysis and annotations supporting each functional insight from extracellular dedicated databases for example: ExoCarta (www.exocarta.org), Vesiclepedia (www.microvesicles.org) [8].

RESULTS

- We identified highly correlated microRNA, proteins and totalRNA (Fig. 4 and Table 1). The top 30 highest ranked by the algorithm were selected for further analysis steps (five each with the highest negative and positive correlation from each of the data types).

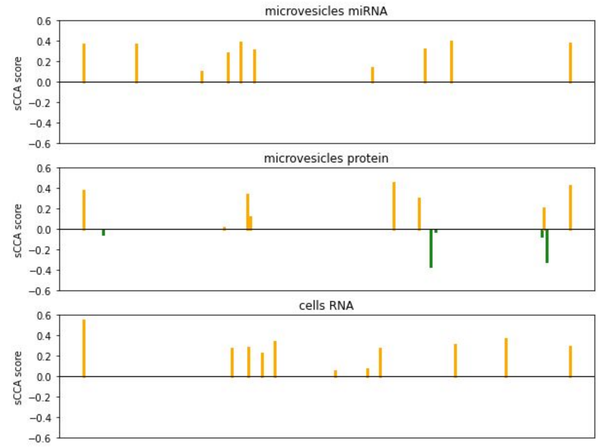


Fig. 4 Visualization of sCCA results for melanoma: microvesicles miRNA, microvesicles proteins and cell totalRNA. The x-axis shows features, while the y-axis shows the cCCA score. Presented bioinformatic method allows to adjust the number of displayed features, starting with the most important ones.

Table 1. Results for 30 top scored sCCA melanoma 1) microvesicles miRNA, 2) microvesicles proteins and 3) cell totalRNA with sCCA scores.

miRNA ID	sCCA score	protein ID	sCCA score	RNA (Gene) ID	sCCA score
MIMAT0002866	3,95E-01	Q15029	4,45E-01	AMIGO2	5,45E-01
MIMAT0002837	3,86E-01	Q14103	4,17E-01	SVEP1	3,60E-01
MIMAT0004687	3,73E-01	P25788	3,73E-01	IL31RA	3,38E-01
MIMAT0000724	3,67E-01	P27695	3,30E-01	RPS14P8	3,07E-01
MIMAT0000281	3,58E-01	Q6DD88	2,98E-01	ZNF812P	2,88E-01
MIMAT0002859_1	3,15E-01	O95232	1,99E-01	HEATR4	2,81E-01
MIMAT0002838	3,01E-01	P11717	1,15E-01	GFRA1	2,72E-01
MIMAT0002835	2,76E-01	Q9Y6E0	6,95E-02	NRP1	2,67E-01
MIMAT0002855	1,31E-01	P07195	3,17E-01	HRH1	2,24E-01
MIMAT0002833	9,78E-02	Q16186	3,61E-01	NCLP1	6,58E-02

- Selected top 30 highest ranked biological features were used for functional analysis starting with finding the most important interactions. We combine RNA interactome: http://www.rna-society.org/rnainter/ with protein interactome: https://string-db.org/ . We use only strongest experimental evidences with highest confidence score (>0.9).
- The three most important connection clusters were selected (Fig. 6). The clusters were supplemented with information from databases dedicated to extracellular microbes. Based on these data, two very significant protein with strong evidence for melanoma were found: IGF2R (protein ID: P11717, ExoCarta ID: ExoCarta_3482) and EFTUD2 (protein ID: Q15029, ExoCarta ID: ExoCarta_9343).
- The interactome study based on top 30 features also showed functional molecular enrichments like telomeric and damaged DNA binding or protein tyrosine kinase related pathways.

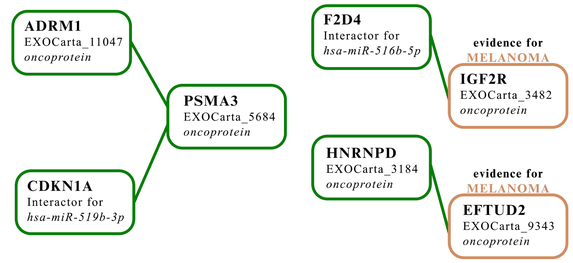


Fig. 5 Interactome analysis. We identify two oncoproteins with strong evidence for extracellular vesicles derived melanoma processes: IGF2R (protein ID: P1171) and EFTUD2 (protein ID: Q15029).

Table 2. Functional enrichments in study network.

Molecular Function (Gene Ontology)	
GO term	description
GO:0042162	telomeric DNA binding
GO:0004714	transmembrane receptor protein tyrosine kinase activity
GO: 0003684	damaged DNA binding
GO:0019955	cytokine binding
GO:0004713	protein tyrosine kinase activity

DISCUSSION

- Proposed method detected important signatures in multi-omics datasets and identified biomarkers candidates like circulating cancer-associated microRNAs and oncoproteins.
- Pipeline ranked significant biological features using sCCA score.
- Method allowed to examine the biological processes related with melanoma progression by selecting molecular signatures that have supporting evidence in databases.
- Method is dedicated to extracellular melanoma biomarker identification but it is elastic and can be adapted to research on other data and cancer types.

REFERENCES

1. Ryan Van Laar, Mitchel Lincoln, and Barton Van Laar. Development and validation of a plasma-based melanoma biomarker suitable for clinical use. British Journal of Cancer, 118(6):857–866, January 2018.
2. Su Yin Lim, Jenny H. Lee, Russell J. Diefenbach, Richard F. Kefford, and Helen Rizos. Liquid biomarkers in melanoma: detection and discovery. Molecular Cancer, 17(1), January 2018.
3. Daniela Witten, Robert Tibshirani, and Trevor Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis, Biostatistics, 10(3):515-534, July 2009.
4. Theodoulos Rodosthenous, Vahid Shahrezaei, and Marina Evangelou. Integrating multi-OMICS data through sparse canonical correlation analysis for the prediction of complex traits: a comparison study. Bioinformatics, 36(17) May 2020.
5. Helian Feng, Nicholas Mancuso, Alexander Gusev, Arunabha Majumdar, Megan Major, Bogdan Pasaniuc and Peter Kraft. Leveraging expression from multiple tissues using sparse canonical correlation analysis and aggregate tests improve the power of transcriptome-wide association studies. July 2020.
6. Magdalena Surman , Sylwia Kędracka-Krok, Dorota Hoja-Łukowicz, Urszula Jankowska, Anna Drożdż, Ewa Ł. Stępień and Małgorzata Przybyło. Mass Spectrometry-Based Proteomic Characterization of Cutaneous Melanoma Ectosomes Reveals the Presence of Cancer-Related Molecules. International Journal of Molecular Sciences , 21(8), 2934, March 2020
7. Dieudonne van der Meer, Syd Barthorpe, Wanjuan Yang, Howard Lightfoot, Caitlin Hall, James Gilbert, Hayley E Francies and Mathew J Garnett. Cell Model Passports—a hub for clinical, genetic and functional datasets of preclinical cancer models. Nucleic Acids Research, 47(D1):D923–D929, January 2019.
8. Website: http://www.microvesicles.org/ and http://www.exocarta.org/ . Date of access: 11.11.2020

REST and ZBTB33 in glioma REST ChIP-seq peaks - partners or competitors?

Marta Jardanowska [1,2 *], Bartosz Wojtas[3], Małgorzata Perycz[3], Bożena Kaminska[3], Michal J. Dabrowski [1]

1 Institute of Computer Science, Computational Biology Lab, Polish Academy of Sciences, Warsaw, Poland
2 Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Warsaw, Poland
3 Nencki Institute of Experimental Biology, Warsaw, Poland
* To whom the correspondence should be addressed: marta.jardanowska@ipipan.waw.pl

Introduction

Gliomas are one of the most common and deadly cancers and because of that are intensively studied. At the same time, one of the most promising and still unfathomable issue is the role of the REST transcription factor in brain carcinogenesis processes. On the other hand, the canonical role of REST is regulation of neurogenesis and glial cells development and participation in the neurosecretion process. REST is the main repressor of transcription in neurodegenerative diseases and is associated with the regulation of ion channels and cytoskeletal proteins, but also other transcription factors (TFs). Therefore REST is described as both, activator and repressor of transcription depending on physiological or pathophysiological context. The purpose of this study was to check whether any TF motifs overlap or are in close proximity to REST Transcription Factor Binding Sites (TFBS).

Materials and methods

For REST ChIP-seq peaks from U87 cell line we assigned their summits within the 200bp sequence around the summit (+/- 100bp), using open source bioinformatic tools. For that purpose we used Position Weight Matrices (PWMs) of TF motifs from HOCOMOCO[1] database and 14 additional REST PWMs, mainly from ENCODE[2].
The search of TF motifs was performed using PWMEnrich[3] Bioconductor R package. To identify specific transcription factor binding sites with the corresponding q-values, we used online FIMO[4] tool from MEME Suite 5.0.5. Additionally, peaks were assigned to gene promoters and based on TCGA glioma RNA-seq and in-house REST ChIP-seq data it was specified whether REST represses or activates the expression of the particular genes based on the correlation results, negative or positive, respectively.

Results

Rank	Target	PWM	P-value
1.5	KAISO_HUMAN.H11MO.0.A		0
1.5	KAISO_HUMAN.H11MO.1.A		0
3	KAISO_HUMAN.H11MO.2.A		1.39e-139
4	E2F4_HUMAN.H11MO.1.A		3.8e-82
5	REST_HUMAN.H11MO.0.A		4.85e-78
6	REST_m14_known_matrix		1.67e-77
7	FEV_HUMAN.H11MO.0.B		3.52e-73
8	ZBED1_HUMAN.H11MO.0.D		2.47e-72
9	E2F5_HUMAN.H11MO.0.B		1.44e-68
10	ZBT14_HUMAN.H11MO.0.C		4.7e-67
11	E2F2_HUMAN.H11MO.0.B		1e-65
12	E2F1_HUMAN.H11MO.0.A		3.24e-64
13	E2F4_HUMAN.H11MO.0.A		6.39e-64
14	SP1_HUMAN.H11MO.1.A		7.2e-64
15	REST_m4_GM12878_encode		2.36e-62

Fig. 1 Ranking of TOP15 motifs for REST activated genes.

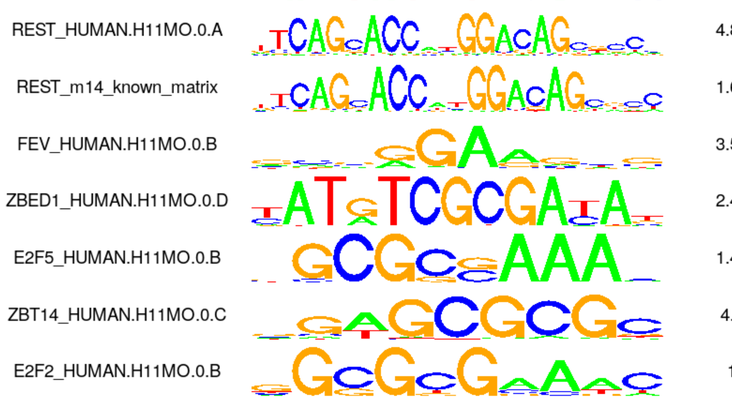


Fig. 2 Clustering of TF motifs characteristic for REST activated genes, REST repressed genes and common motifs based on DNA sequences.

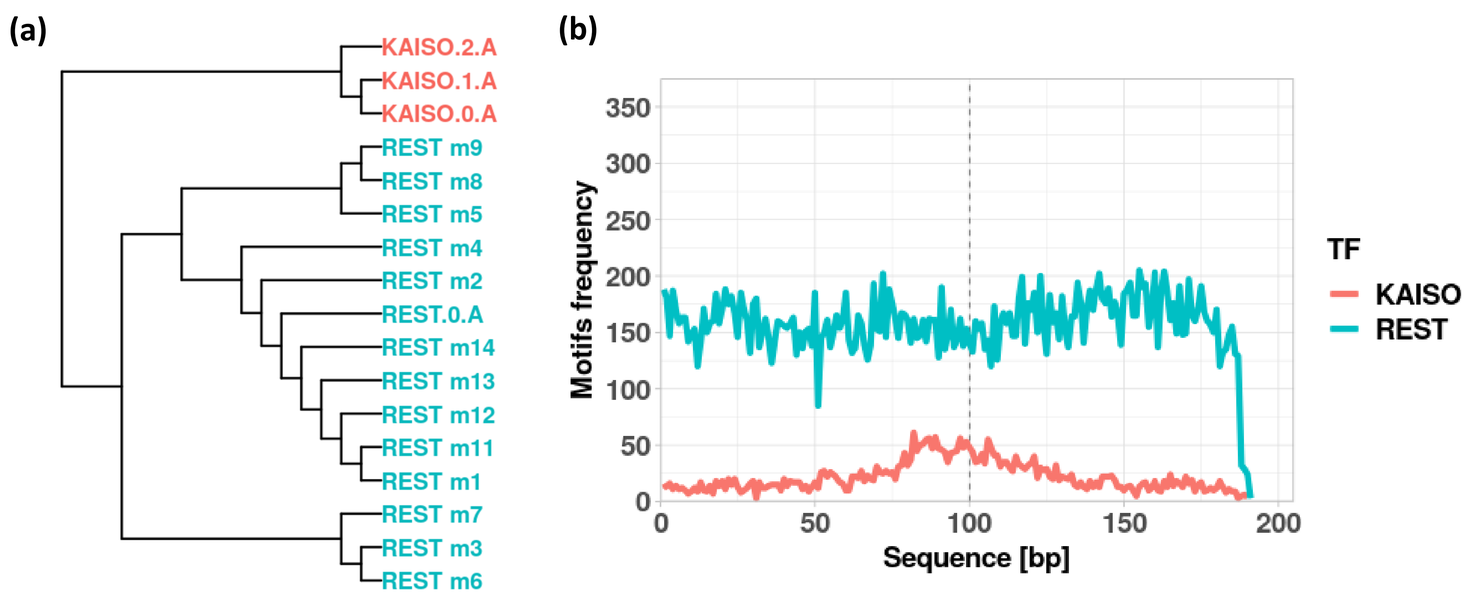


Fig. 3 REST and KAISO motifs (a) clustering based on DNA sequences (b) occurrence dependent from the localization in the activated genes sequences.

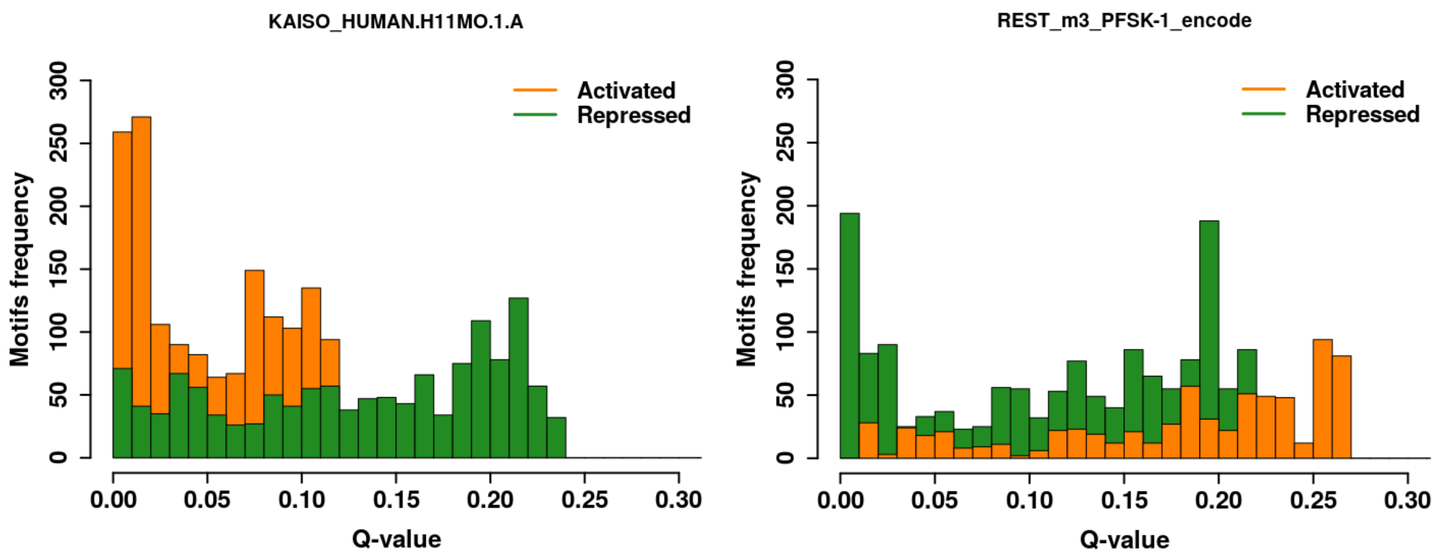


Fig. 4 Q-value and frequency relation for selected KAISO and REST motif for REST ChIP-seq peaks for activated and repressed genes.

Conclusions

- We identified 202 TF motifs (12 REST motifs) in the 200bp sequences surrounding REST ChIP-seq peaks for the activated genes sequences and 237 TF (14 REST motifs) motifs for the repressed genes sequences. Top places in the motifs ranking for the REST activated genes were occupied by the KAISO motifs, characteristic for the ZBTB33 transcription factor. (Fig. 1)
- Motifs characteristic for activated (n = 21) and repressed (n = 56) genes clustered separately. (Fig. 2)
- Analysis of the nucleotide sequences of the identified motifs showed that they significantly differed between REST and ZBTB33, meaning that the co-occurrence of these TF motifs within the examined sequences was not due to sequence similarity. (Fig.3a)
- We observed that in the REST activated genes, KAISO motifs were significantly more frequent in the proximity to the peak summits than in the rest of the examined 200bp sequence. (Fig. 3b)
- ZBTB33 motifs occurred with higher frequency and lower q-value in the REST activated genes, while the majority of REST motifs were within the repressed genes. (Fig. 4)
- These results may suggest that while the main REST role may be repressive, its role within the activated genes promoters can be at least co-dependent on ZBTB33.

Supported by National Science Centre, Poland grant number DEC- 2015/16/W/NZ2/00314.

References

[1] I. V Kulakovskiy, I. E Vorontsov, I. S Yevshin, R. N Sharipov, A. D Fedorova, E. I Rumynskiy, Y. A Medvedeva, A. Magana-Mora, V. B Bajic, D. A Papatsenko, F. A Kolpakov, V. J Makeev, *HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis,* Nucleic Acids Research, Volume 46, Issue D1, 4 January 2018, Pages D252–D259 **[2]** https://www.encodeproject.org **[3]** R. Stojnic, D . Diez, *PWMEnrich: PWM enrichment analysis. R package version 4.18.0,* 2018. **[4]** C. E. Grant, T. L. Bailey and W. Stafford Noble, *FIMO: Scanning for occurrences of a given motif*, Bioinformatics 27(7):1017–1018, 2011.

DNA methylation patterns of active enhancers specific for *pilocytic astrocytoma* and Higher Grade Glioma samples

Agata Dziedzic[1], Marta Jardanowska[1], Marcin Grynberg[2] and Michał J Dąbrowski[1]

[1] Institute of Computer Science, Polish Academy of Sciences, Poland
[2] Institute of Biochemistry and Biophysics, Polish Academy of Sciences

Aim

- To study molecular differences in enhancers of different glioma grades: *pilocytic astrocytoma* and Higher Grade Glioma.
- To detect specific methylation sites in Transcription Factor motifs responsible for changes of its transcription factors binding affinity and as a result - changes of target gene expression.

Fig.1. Schematic representation of target gene expression regulation via enhancer.

Materials & Methods

Experiment	Type of data	Analysis maperformed on data
Chip-seq for H3K27ac	Genome coordinates of active enhancers	Motif search
Bisulphite seq	Methylation level per single cytosine (~3.5 mln sites per sample)	**DM** cytosines calling
RNA-seq	Read counts per gene	**DE** genes calling

Tab.1. Analysis performed on three layers of biological information for the set of 7 PA and 10 HGG samples.

Results

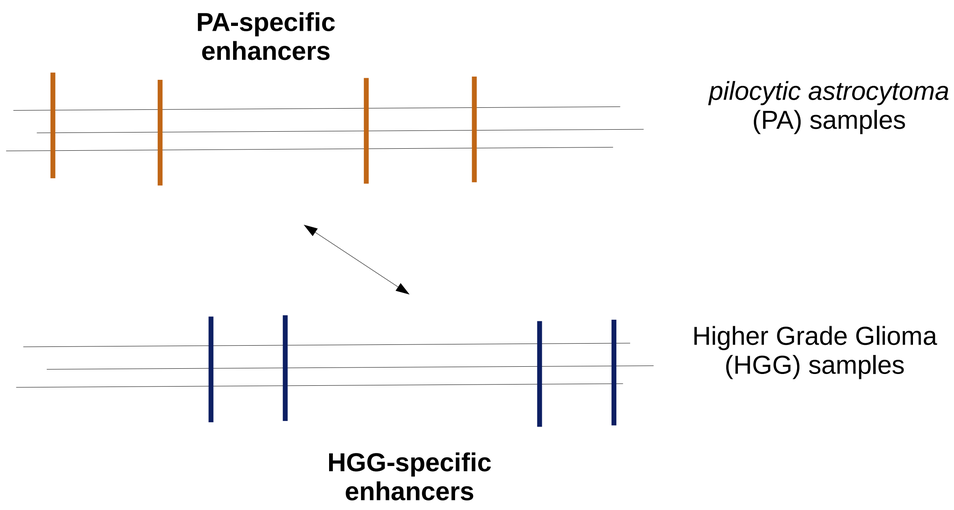


Fig.2. Schematic representation of enhancers methylation levels comparisones.

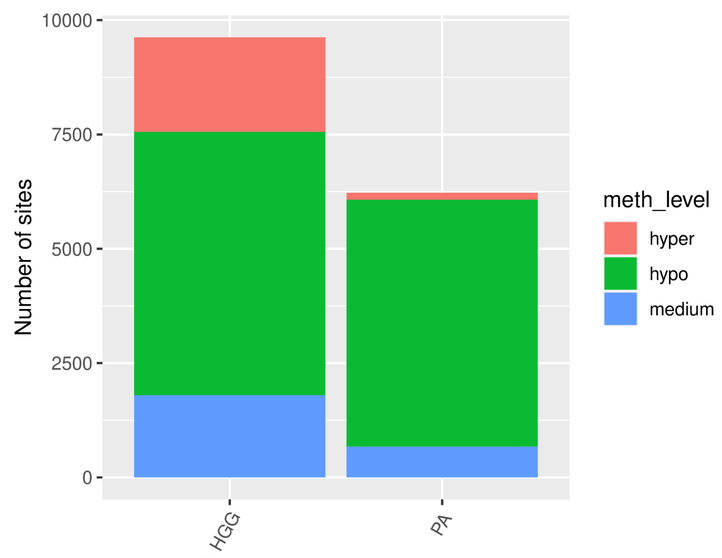


Fig.3. Number of CpG sites devided into three ranges of methylation level.There are more hypermethylated sites in HGG-spec. Enhancers comparing to PA-spec. enhancers (X-squared = 1309.9, df = 1, p-value < 2.2e-16).

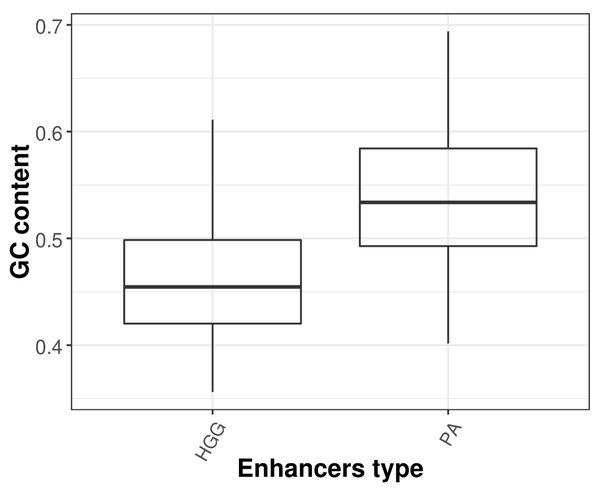


Fig.4. Mean GC content was 46 % for HGG and 54 % for PA – difference was statiscally important (HGG n = 124, PA n = 114, Mann-Whitney U test: p-value = 4.292992e-17, W = 11528)

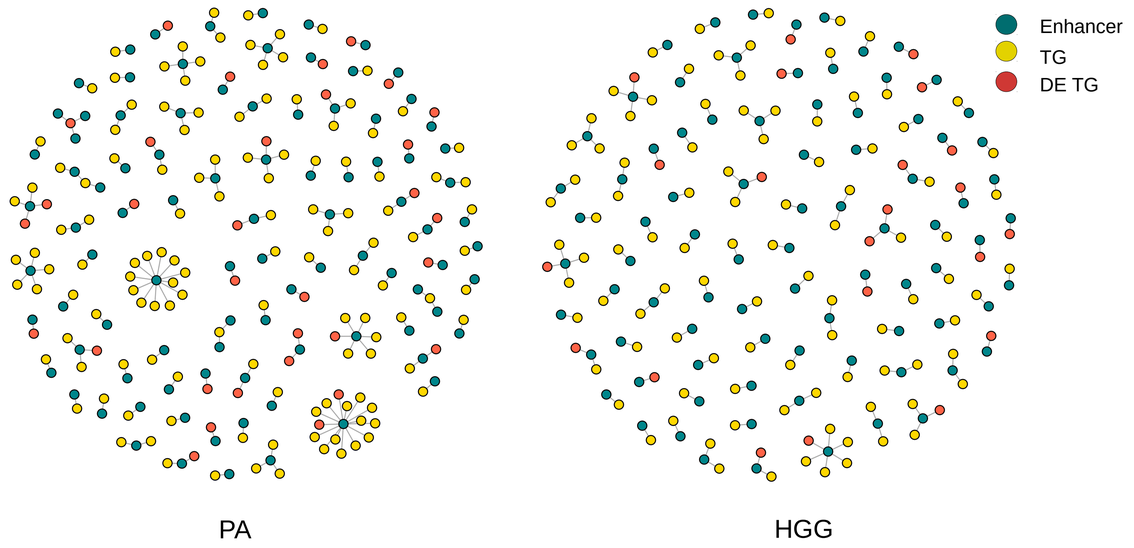


Fig.5. PA: 92 enhancers targeting 161 TG (32 DE). HGG: 84 enhancers targeting 120 TG (22 DE).

Conclusions

- HGG-specific enhancers had **lower frequency of guanine and cytosine nucleotides** then PA-specific enhancers and higher global DNA methylation level.
- Methylation pattern of **14 TF motifs was confirmed to be consequently hypermethylated in HGG** compared to PA samples and all of this motifs were found in at least one ehnancer with differentially expressed target gene.
- These results indicate specific TF motifs whose **methylation may have an influence on regulation of TG expression** and therefore contribute to gliomagenesis.

a)

b)

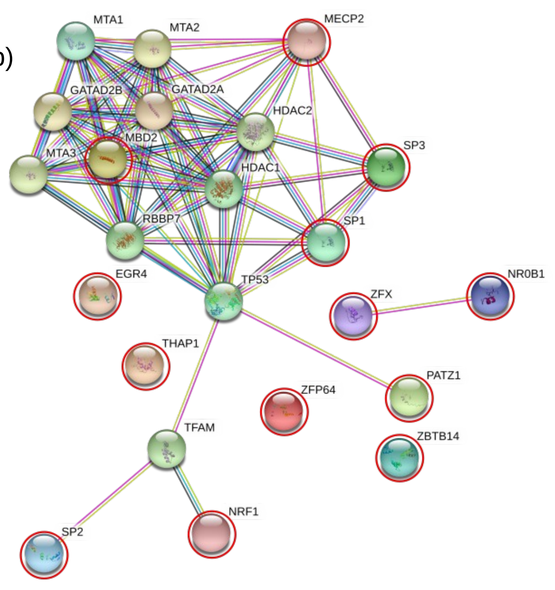


Fig.6. a) Selected 14 TF motifs & their nucleotide sequence; b) Graph of 13 TFs together with additional proteins they interact with.

Funding:
National Science Center, SYMFONIA3: DEC 2015/16/W/NZ2/00314

Novel alternative splicing events detection in human genome with Spladder

Agata Muszyńska[1,2],Paweł Łabaj[1,3]

[1] Małopolska Centre of Biotechnology UJ, Kraków, Poland
[2] Institute of Automatic Control, Electonics and Computer Science, Silesian University of Technology, Gliwice, Poland
[3]Chair of Bioifnormatics RG, Boku University, Vienna, Austria

Abstract

Although human genome is widely studied since many years its complexity remains not fully understood. One of the mechanisms that stands for that is alternative splicing, which is a process of joining exons in multiple ways, so that novel mRNA and, in fact, novel proteins are produced. Currently we are not fully aware of all of the splicing events that might be present in a given genome. One of the tools that provides the possibility to investigate that is Spladder. It builds an augmented splicing graph, based on current annotation and than expands it with novel events. Currently Spladder supports detecting six different types of such events. We used Spladder software on data from SEQC consortium project [1][2].

We investigated 3 samples (A- mixture of 10 different cancer cell lines, B- healthy individual and C- A and B samples mixed in 1:1 ratio) run on different RNA targeting panels, as well as on whole transcriptome sequencing data obtained with two protocols- ribo-depletion and polyA selection. Preliminary results show that there is a fraction of genes containing novel events, which seems to be cancer or sample specific, but majority is the same irrespective of sample. It seems that the current gene model can be extended by this data. Spladder also revealed that the fraction of intron retention events is higher for whole transcriptome sequencing data than for targeted approach and is higher for ribo-depletion protocol than for polyA selection, what is expected after comparing sample processing and library preparation for these approaches.

These results show that there is still a lot of work ahead of us to fully describe our genome but at the same time that Spladder might be a good tool, not only for that challenge, but also for others like detecting cancer specific events.

Study design

4 x A, B, C → Panel1, Panel2, Panel3, WTS PolyA, WTS riboZero

Samples

A: mixture od 10 different cancer cel lines
B: healthy individual
C: A and B mixed in 1:1 ratio

Panels

Panel1: around 1000 genes
Panel2: around 2000 genes
Panel3: around 2000 genes

Spladder pipeline

Annotattion + Alignment Data → Augumented Splicing Graph → Detected Splice Events → Quantified Splice Events → Differentail Analysis/ sQTL Tests

Events detected

- Alternative 5 prime end
- Alternative 3 prime end
- Exon skip
- Intron retention
- Mutually exclusive exons
- Multiple exon skip

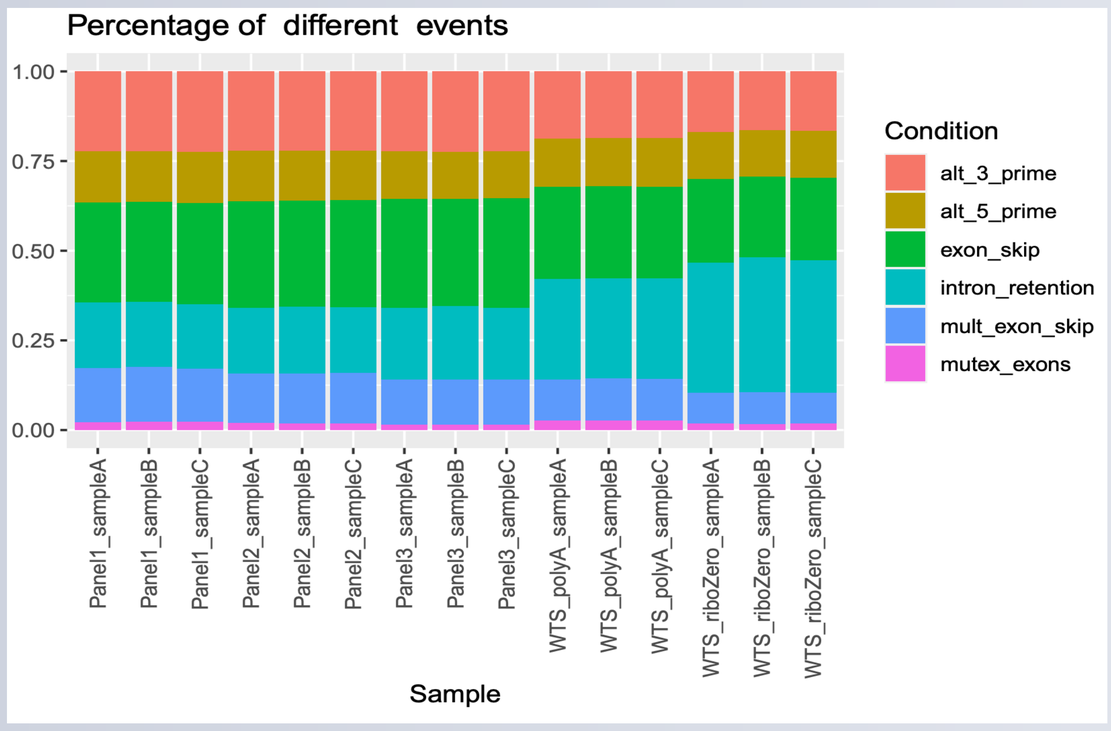
PSI value

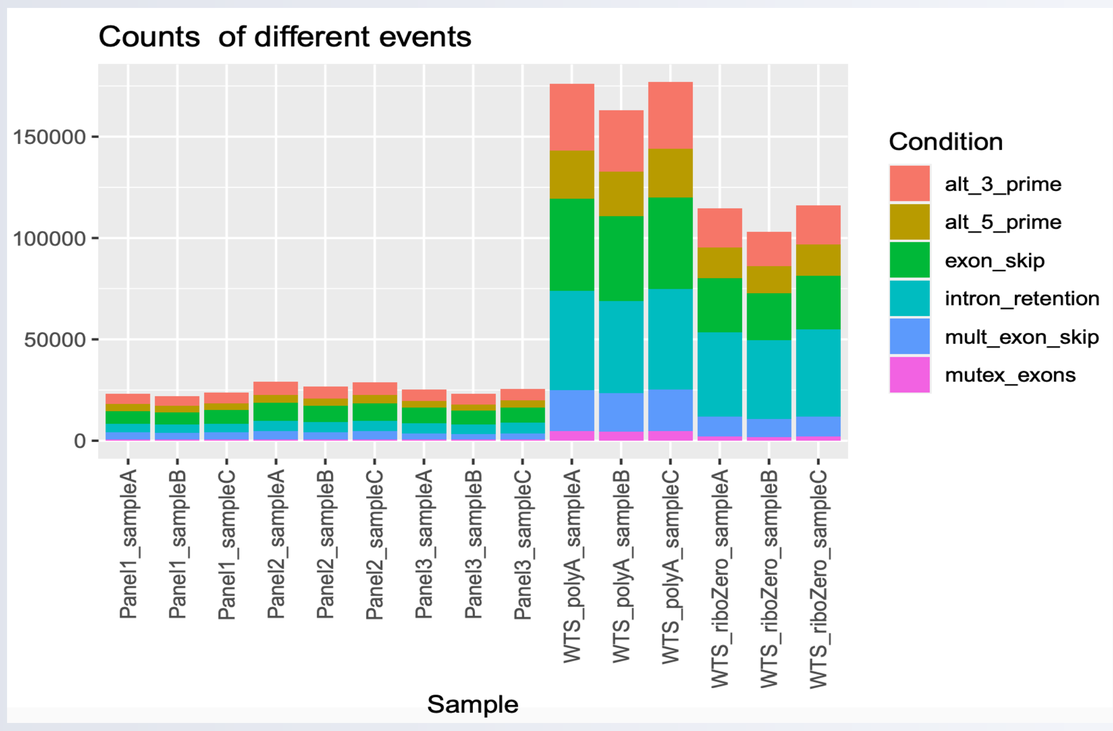
$$PSI = \frac{IR}{IR+ER}$$

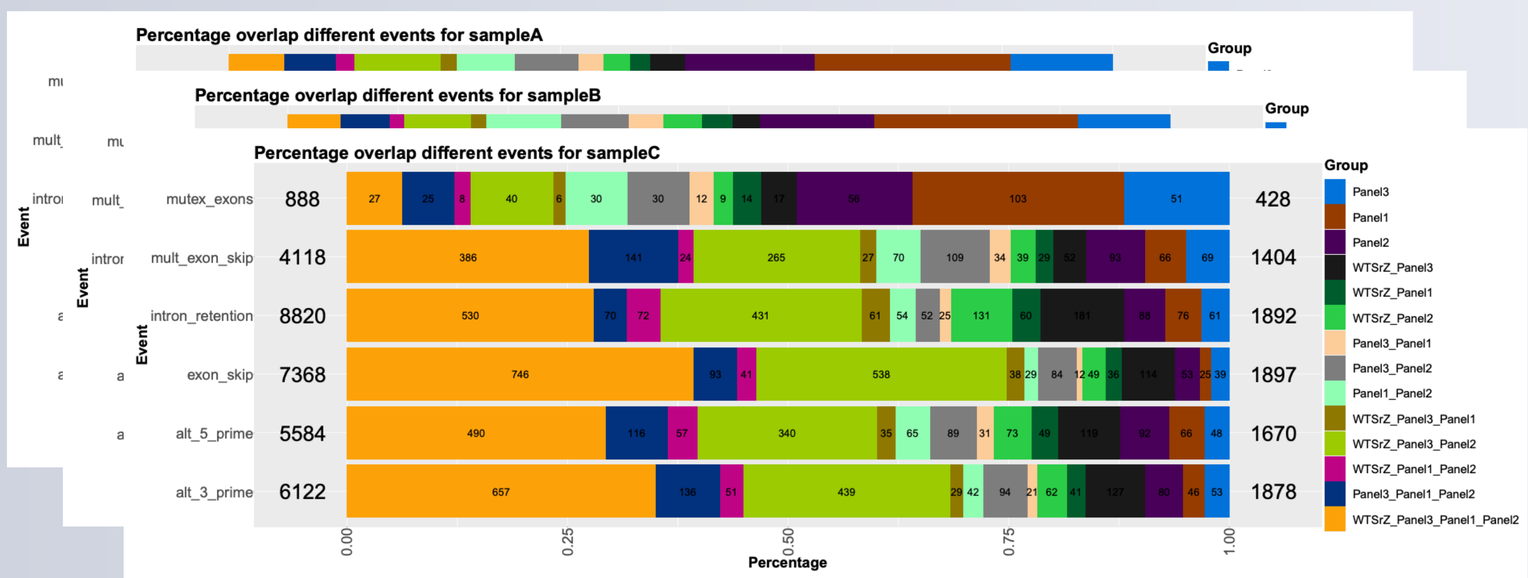
Valid events

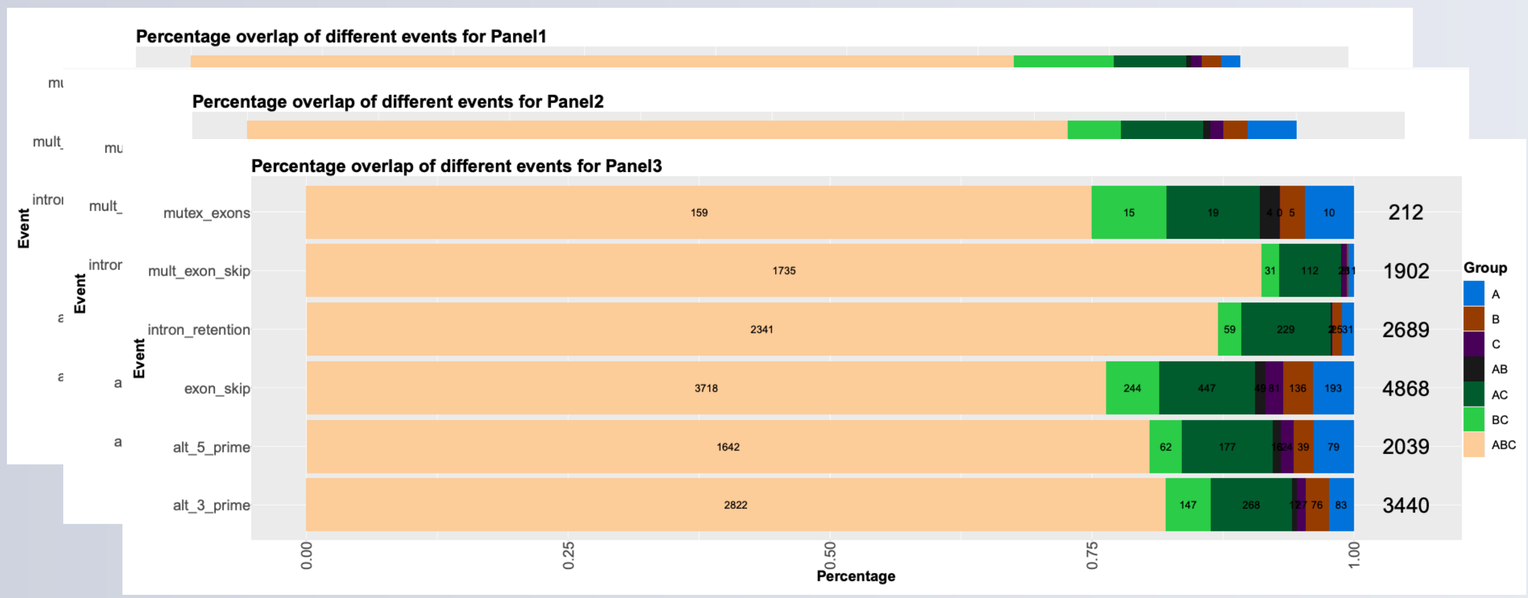
Events with PSI value above 0.1 for all 4 replicates

Results









Conclusions

- We were able to detect all splicing events in our data, among which the most prevalent were exon skip and intron retention, whereas the least- mutually exclusive exons.
- Although there were some events, which seems to be cancer or sample specific, majority is common- this suggest that current gene model might be expanded.
- Intron retention events occur more often in whole transcriptome sequencing data, than in any of the panels and also often in ribo-depletion than in polyA. This reflects differences in library preparation for these approaches.
- WTS with polyA protocol detects more events than riboZero.

References

[1] Su, Zu et al. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Consortium. Nature Biotechnology 32, 903-914(2014)
[2] Kahles A, Ong CS, Zhong Y, Rätsch G. SplAdder: identification, quantification and testing of alternative splicing events from RNA-Seq data. Bioinformatics. 2016 Jun 15;32(12):1840-7.