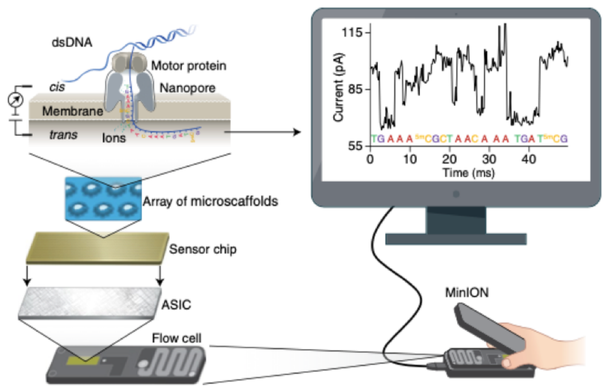
Challenges in the Next Generation (Oxford Nanopore) direct RNA Sequencing (dRNA-seq) Data Processing: Coping With Normalisation of Gene Expression Levels by Principal Component Analysis

A.Kriščiukaitis[1,2], R.Petrolis[1,2], R.Dragunaitė[2], R.Stakaitis[2], D.Skiriutė[2]

[1] Dept. Physics, Mathematics and Biophysics, Lithuanian University of Health Sciences

[2] Neuroscience Institute, Lithuanian University of Health Sciences

Nanopore sequencing technology

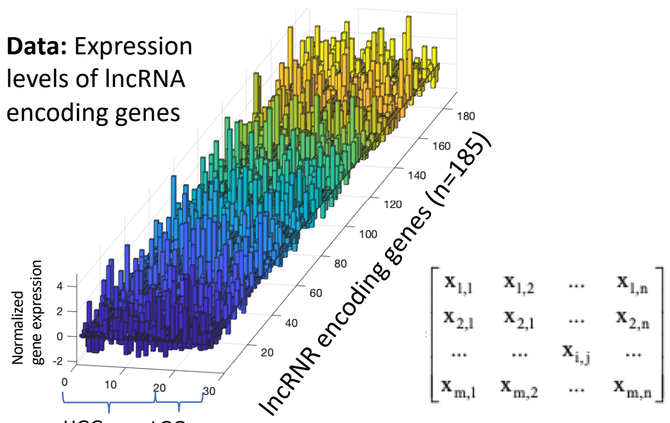


Modified from Wang, Y., Zhao, Y., Bollas, A. et al. Nanopore sequencing technology, bioinformatics and applications. Nat Biotechnol 39, 1348–1365 (2021)

Example of method application – questions to answer:

- Does expression of non-coding-RNA (ncRNA) encoding genes is different in Glioblastoma *vs* Low-grade-glioma (*GBM vs LGG*) cases?
- If yes, of which genes?

Data: Expression levels of lncRNA encoding genes



Covariation matrix of normalized gene expression levels data set: $R_x = E\left[X \cdot X^T\right]$

$$\begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,n} \\ x_{2,1} & x_{2,1} & \dots & x_{2,n} \\ \dots & \dots & x_{i,j} & \dots \\ x_{m,1} & x_{m,2} & \dots & x_{m,n} \end{bmatrix}$$

Introduction: Next generation sequencing (NGS) is getting widely applicable for the detection of molecular markers in modern medical diagnostics. However, processing and evaluation of RNA-seq data is challenging due to some inescapable physical factors causing certain bias in the analyzed data.

Problem: Technical differences ("batch effects") caused by differences in sample processing (up to RNA extraction, RNA-seq library preparation or the number of live pores) may significantly affect the ability to draw generalizable conclusions from such studies.

The Aim: Elaboration of the method for gene expression level analysis avoiding bias caused by inescapable physical factors in RNA-seq data.

The list of HKG: *[Trends in Genetics 29 (2013), 569–574]*

Gene Name	***Gene description***
C1orf43	chromosome 1 open reading frame 43
CHMP2A	charged multivesicular body protein 2A
EMC7	ER membrane protein complex subunit 7
GPI	glucose-6-phosphate isomerase
PSMB2	proteasome subunit, beta type, 2
PSMB4	proteasome subunit, beta type, 4
RAB7A	member RAS oncogene family
REEP5	receptor accessory protein 5
SNRPD3	small nuclear ribonucleoprotein D3
VCP	valosin containing protein
VPS29	vacuolar protein sorting 29 homolog

Solution: The key solution to normalization of gene expression levels estimated across different samples lays in so called ***housekeeping genes (HKG)***, which are required for the maintenance of the basal cell functions. Thus, they are expected to be equally expressed in all cells.

PCA aggregates the information from correlated variables into a smaller set of uncorrelated variables (the principal components). If HKG are expected to be equally expressed in all cells, their expression levels should be aggregated into the principal component, representing technical differences ("batch effects"). On the other hand, we can expect that expression levels of genes, which significantly differ between investigated groups (GBM vs LGG) will be aggregated into some other PC.

Eigenvectors of covariation matrix R:

$$eig(R) \Longrightarrow [\varphi_1, \varphi_2, \dots, \varphi_n]$$

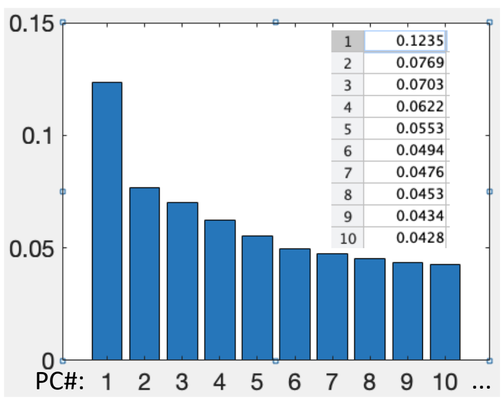
Ordinary eigenvector φ_i (PCi):

$$\varphi_i = [w_{i1}, w_{i2}, \dots\ w_{in}]$$

Expected results: Ideally, the PC representing maximal differences between investigated pathology groups and having lowest contribution of any HKG to it, should be considered for search of genes - candidate biomarkers of investigated pathology.

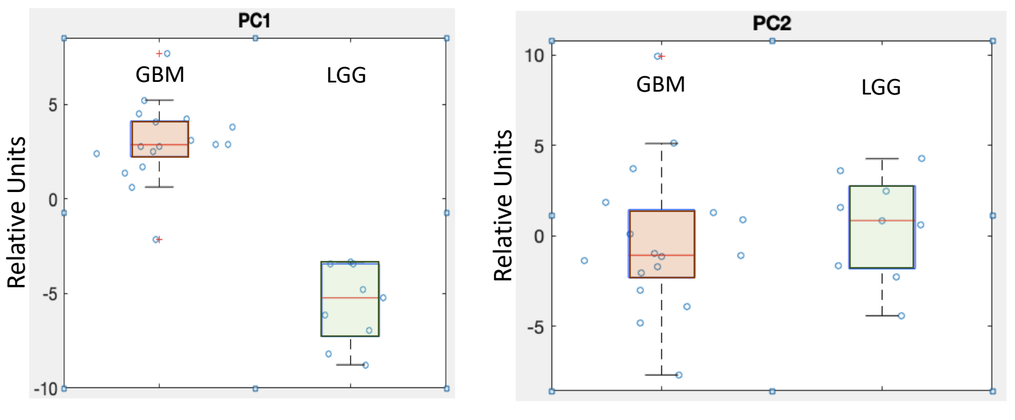
Internal coefficients of each PC represent contribution of each variable to the component. So, the highest coefficients show the most important variables (expression of genes encoding lncRNA) in this component. Therefore, in real situation, we expect that contribution of HKG expression to the sought PC will be substantially lesser than candidate biomarkers of investigated pathology.

Contribution of first eigenvectors of covariation matrix in representation of total variance in analyzed data



Truncated representation of initial data by particular eigenvector:

$PC1\_X_i = w_{11}x_{1i} + w_{12}x_{2i} + \dots + w_{1n}x_{ni}$ $\quad PC2\_X_i = w_{21}x_{1i} + w_{22}x_{2i} + \dots + w_{2n}x_{ni}$

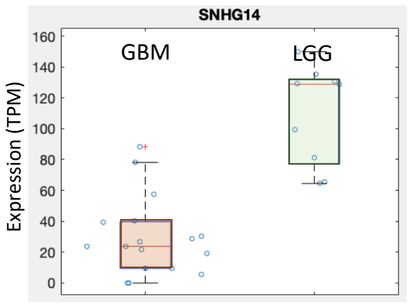


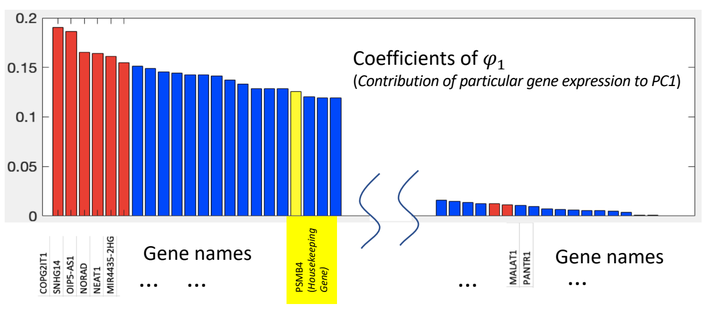
Results: PC1, representing 12.35% of total variance shoved the biggest difference in regard to the pathology type. Contribution of HKG expression in it was negligible. At the same time, PC2, representing 7.69% of total variance, showed non-significant difference between pathology type. The contributions of expression of two HKG were among Top3 in the whole list of this component. So, PC1 was used to search for candidate biomarkers.

Results: ncRNAs encoding genes which expression differs between GBM and LGG cases.

Content of φ_1

Gene Name	PC1 Koef.
COPG2IT1	-0,190234
SNHG14	-0,1861258
OIP5-AS1	-0,1650589
NORAD	-0,1639362
NEAT1	0,1611103
MIR4435-2HG	0,15489642
...	...
MALAT1	0,0124793
PANTR1	-0,011318

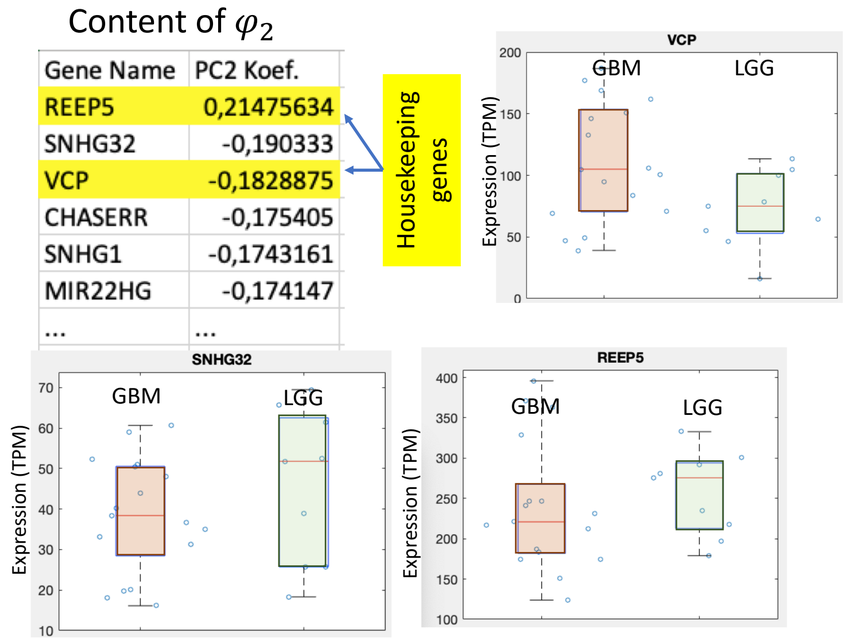




Content of φ_2

Gene Name	PC2 Koef.
REEP5	0,21475634
SNHG32	-0,190333
VCP	-0,1828875
CHASERR	-0,175405
SNHG1	-0,1743161
MIR22HG	-0,174147
...	...

Housekeeping genes



Conclusions: Application of Principal Component Analysis to RNA-seq data revealed particular ncRNA encoding genes most differently expressed in the samples of Glioblastoma and Low-grade-glioma. The *ncRNAs encoded by* COPG2IT1 and SNGH14, together with OIP5-AS1, NORAD, NEAT1, and MIR4435-2HG could be the candidates for biomarkers differentiating Glioblastoma and Low-grade-glioma cases.

Aknowledgements: This study was funded from Research Council of Lithuania (LMT) Grants: S-SEN-20-7 and MIP-20-51.

Research Council of Lithuania