Topological Data Analysis of High-dimensional Correlation Structures with Applications in Epigenetics

- SARA PRADA ALONSO

8 November 2021

Antonio Gómez Tato
María de los Ángeles Casares de Cal

Novel analytical tools

Link between mathematical thinking and biological observation

The Forefront of Genomics. Nature, October 2020

CHALLENGES

- How do we model a dataset with more than 400,000 variables?
- How do we calculate the correlation matrix and interpret it?
- How can we create standard genomic analytical tools ready to be used?
- How is the epigenetic network working?

OBJECTIVES

- Develop novel analytical tools to study high-dimensional correlation structures
- Efficient application and diagnostic power
- Understand complex epigenetic mechanisms (3D-correlation)
- Present mathematical strategies valid other research fields

PROPOSAL

- Topological approach on epigenetic data
- Study high-dimensional correlation matrices through related correlation networks
- Model and computational algorithm

Novel analytical tools

Link between mathematical thinking and biological observation

The Forefront of Genomics. Nature, October 2020

CHALLENGES

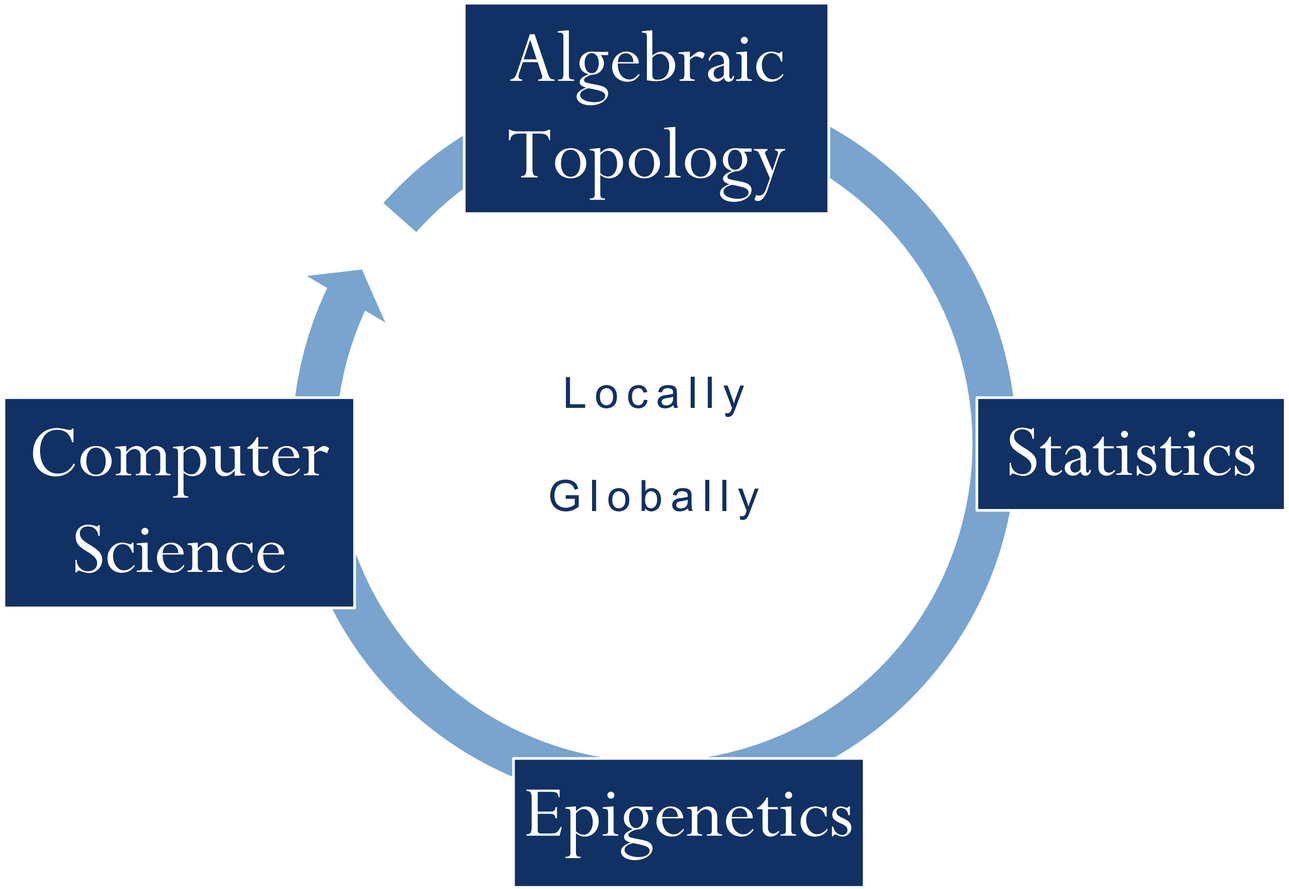
- How do we model a dataset with more than 400,000 variables?
- How do we calculate the correlation matrix and interpret it?
- How can we create standard genomic analytical tools ready to be used?
- How is the epigenetic network working?

OBJECTIVES

- Develop novel analytical tools to study high-dimensional correlation structures
- Efficient application and diagnostic power
- Understand complex epigenetic mechanisms (3D-correlation)
- Present mathematical strategies valid other research fields

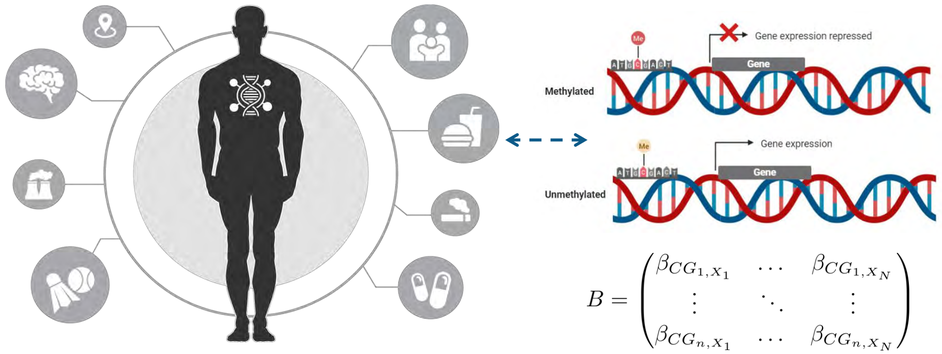
PROPOSAL

- Topological approach on epigenetic data
- Study high-dimensional correlation matrices through related correlation networks
- Model and computational algorithm



Beyond the DNA sequence

EPIGENETICS

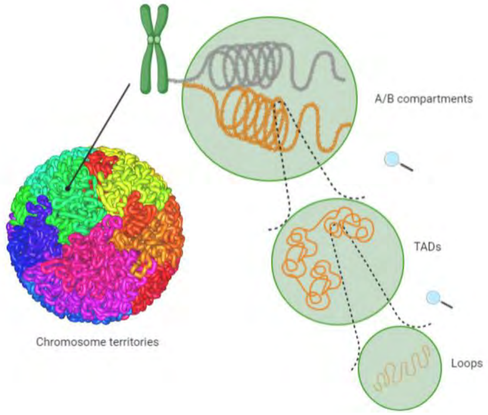


$$B = \begin{pmatrix} \beta_{CG_1,X_1} & \cdots & \beta_{CG_1,X_N} \\ \vdots & \ddots & \vdots \\ \beta_{CG_n,X_1} & \cdots & \beta_{CG_n,X_N} \end{pmatrix}$$

DNA

?

3D GENOME ARCHITECTURE



Molecular biology is also a mathematical challenge

Topological data analysis (TDA)

The translation of data into the language of algebraic topology to study its shape and invariants

PERSISTENT HOMOLOGY

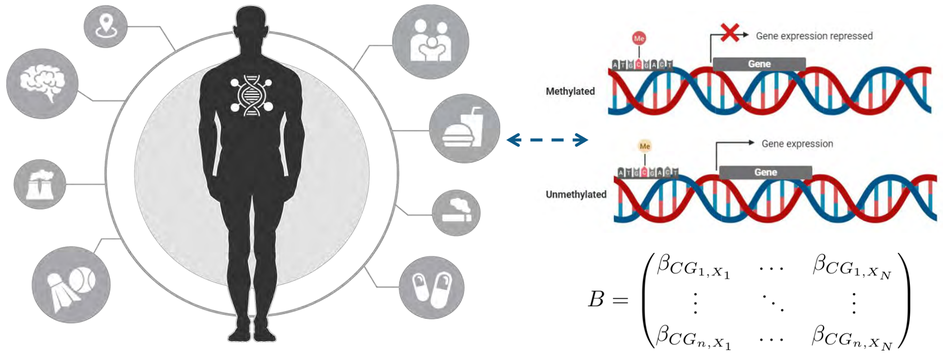
MAPPER

Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition, G. Carlsson et al, 2007.

"To let the data speak"

Beyond the DNA sequence

EPIGENETICS

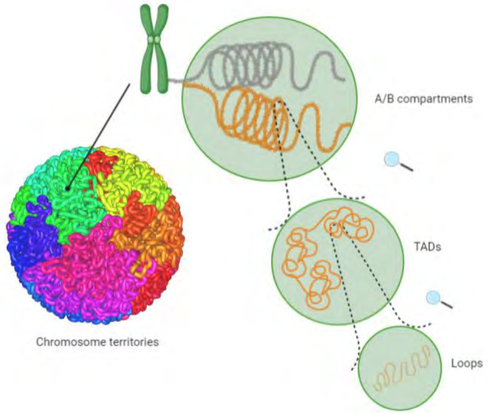


$$B = \begin{pmatrix} \beta_{CG_1,X_1} & \cdots & \beta_{CG_1,X_N} \\ \vdots & \ddots & \vdots \\ \beta_{CG_n,X_1} & \cdots & \beta_{CG_n,X_N} \end{pmatrix}$$

DNA

?

3D GENOME ARCHITECTURE



Molecular biology is also a mathematical challenge

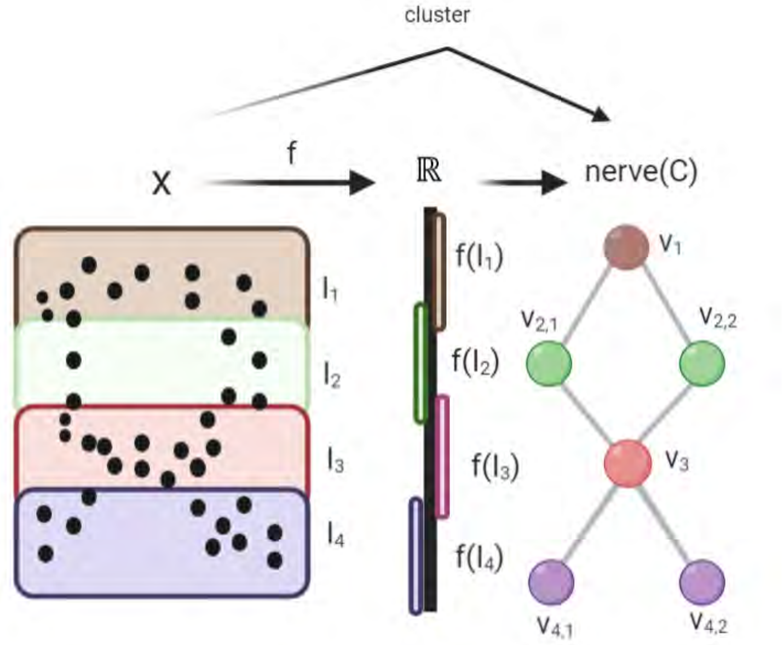
Topological data analysis (TDA)

The translation of data into the language of algebraic topology to study its shape and invariants

PERSISTENT HOMOLOGY

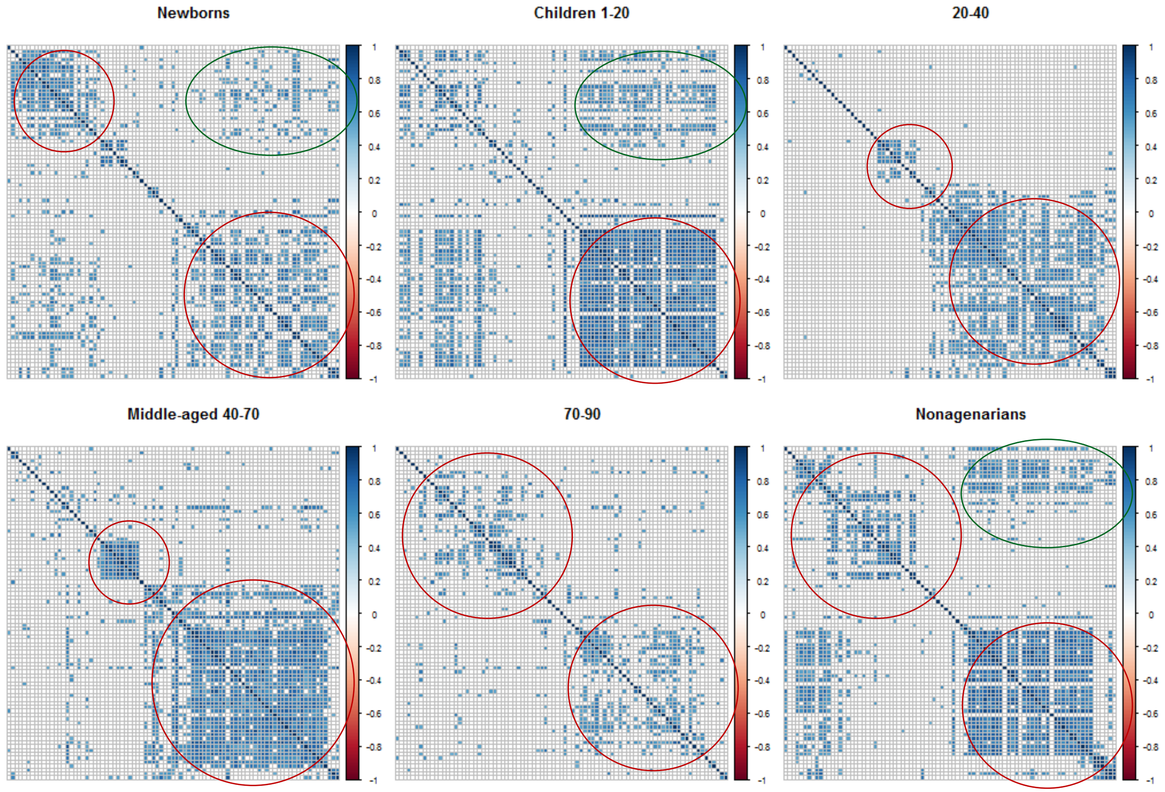
MAPPER

Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition, G. Carlsson et al, 2007.



"To let the data speak"

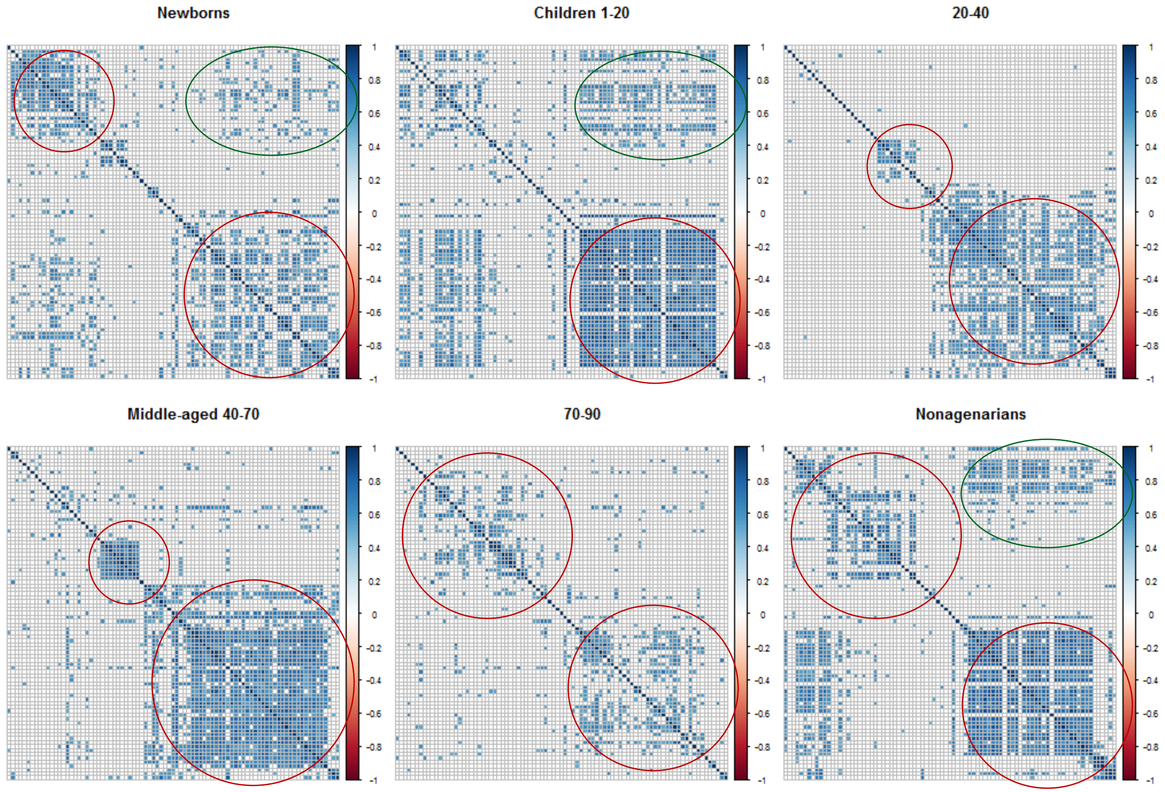
Local Correlation – Aging process

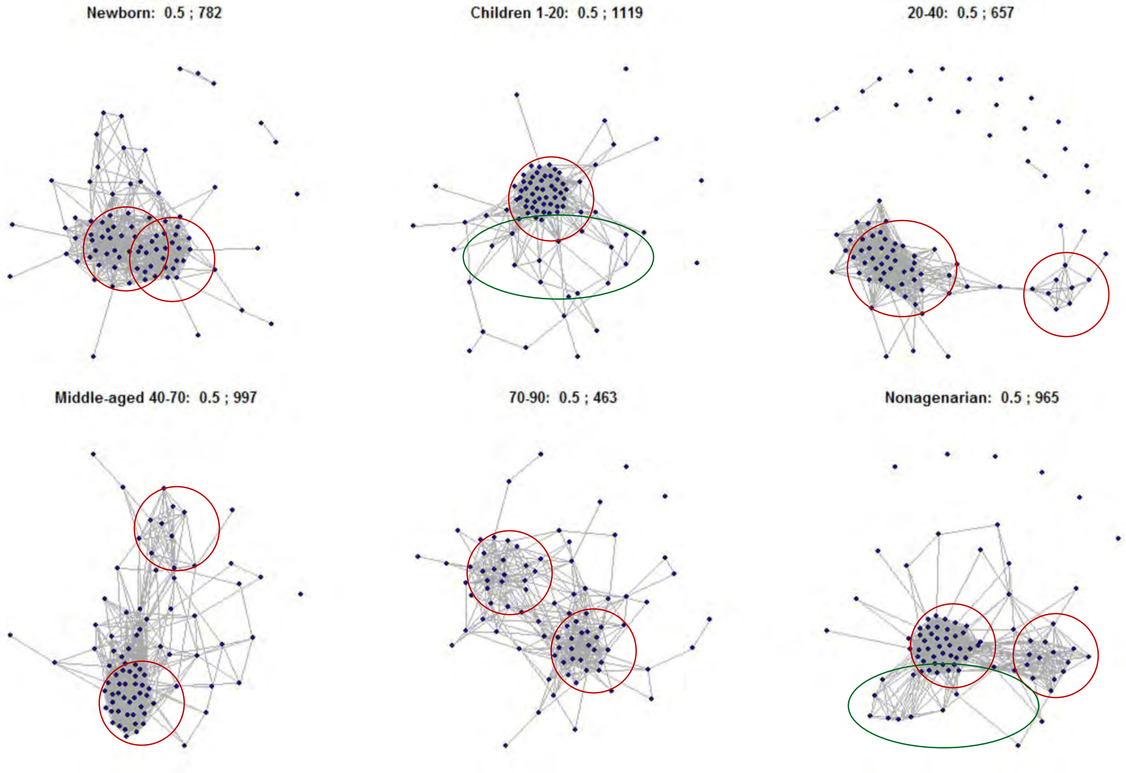


Two network descriptors, the distance and the correlation

The spatial design given by the distance between the nodes affects the topology of the graph

Local Correlation – Aging process





Two network descriptors, the distance and the correlation

The spatial design given by the distance between the nodes affects the topology of the graph

A stochastic **block** model with **distance**

Modeling the community structure

Let $G = (V, E)$ be a graph, where V is the node set of dimension n and E is the set of edges.

Let K be the number of groups of nodes defined taking into account the distance D between the nodes, i.e., nodes are grouped initially based on their genomic distance on a partition of K consecutive intervals.

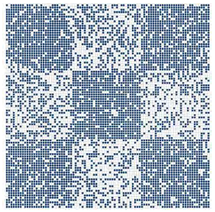
The matrix Z of dimension $n \times K$, is defined such that each row $Z_i = 0$ except exactly once that takes the value $Z_i = 1$ (it represents the group whose the node belongs).

$$Z = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \dots & \dots & \dots & \dots & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 0 & 1 \\ 0 & \dots & 0 & 0 & 1 \end{pmatrix}$$

The matrix C represents the connections among the groups. Each element of the block matrix C_{ij} of dimension $K \times K$ represents the probability of occurrence of an edge between a node in group k_i and a node in group k_j.

$$C = \begin{pmatrix} p_{k_1,k_1} & \dots & p_{k_1,k_K} \\ \vdots & \ddots & \vdots \\ p_{k_K,k_1} & \dots & p_{k_K,k_K} \end{pmatrix}$$

$$C = \begin{pmatrix} 0.8 & 0.2 & 0.8 \\ 0.2 & 0.8 & 0.2 \\ 0.8 & 0.2 & 0.8 \end{pmatrix}$$



Are we observing the secret of longevity?

A stochastic **block** model with **distance**

Modeling the community structure

Let $G = (V, E)$ be a graph, where V is the node set of dimension n and E is the set of edges.

Let K be the number of groups of nodes defined taking into account the distance D between the nodes, i.e., nodes are grouped initially based on their genomic distance on a partition of K consecutive intervals.

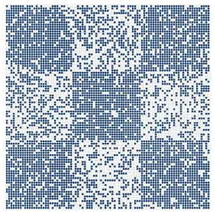
The matrix Z of dimension $n \times K$, is defined such that each row $Z_i = 0$ except exactly once that takes the value $Z_i = 1$ (it represents the group whose the node belongs).

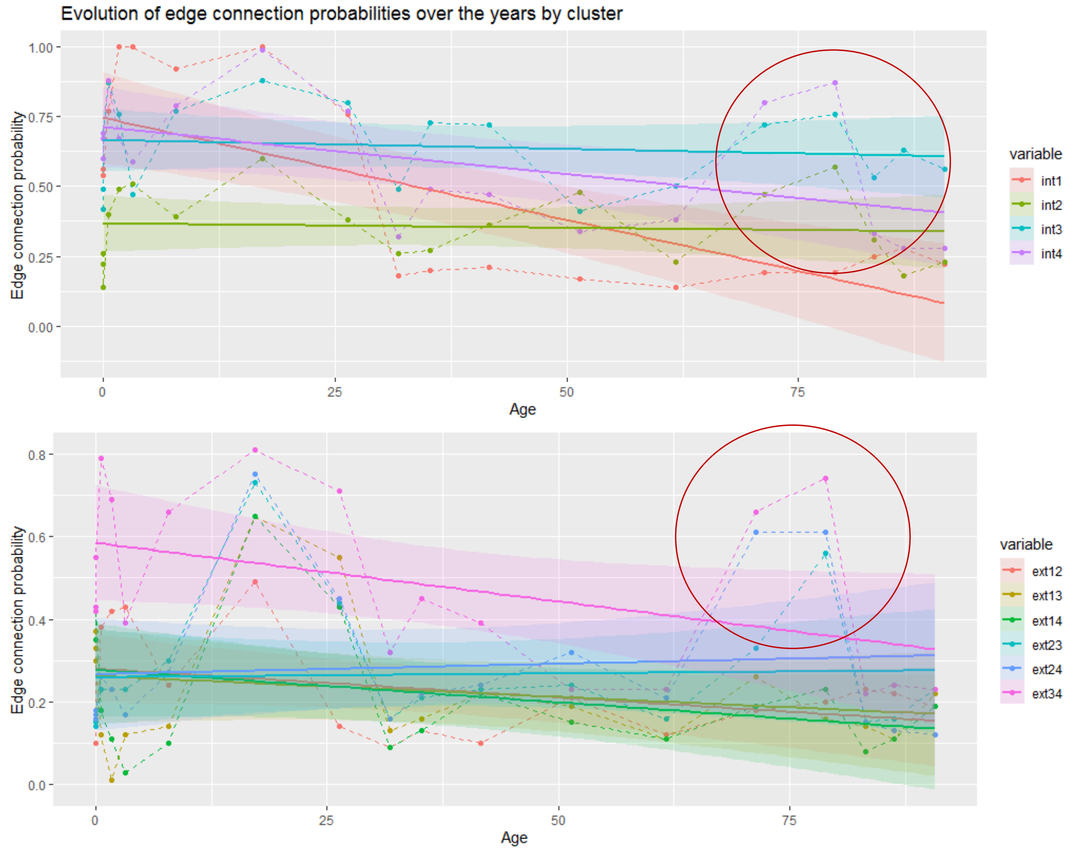
$$Z = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \dots & \dots & \dots & \dots & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 0 & 1 \\ 0 & \dots & 0 & 0 & 1 \end{pmatrix}$$

The matrix C represents the connections among the groups. Each element of the block matrix C_{ij} of dimension $K \times K$ represents the probability of occurrence of an edge between a node in group k_i and a node in group k_j.

$$C = \begin{pmatrix} p_{k_1,k_1} & \dots & p_{k_1,k_K} \\ \vdots & \ddots & \vdots \\ p_{k_K,k_1} & \dots & p_{k_K,k_K} \end{pmatrix}$$

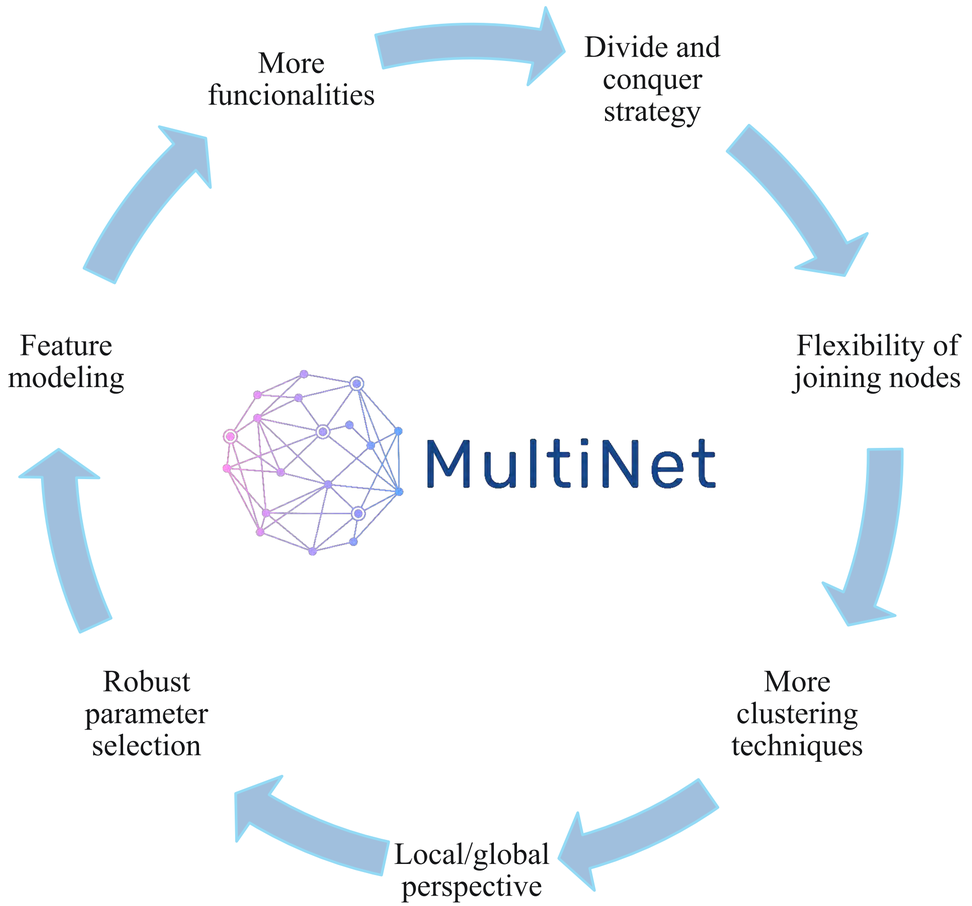
$$C = \begin{pmatrix} 0.8 & 0.2 & 0.8 \\ 0.2 & 0.8 & 0.2 \\ 0.8 & 0.2 & 0.8 \end{pmatrix}$$





Are we observing the secret of longevity?

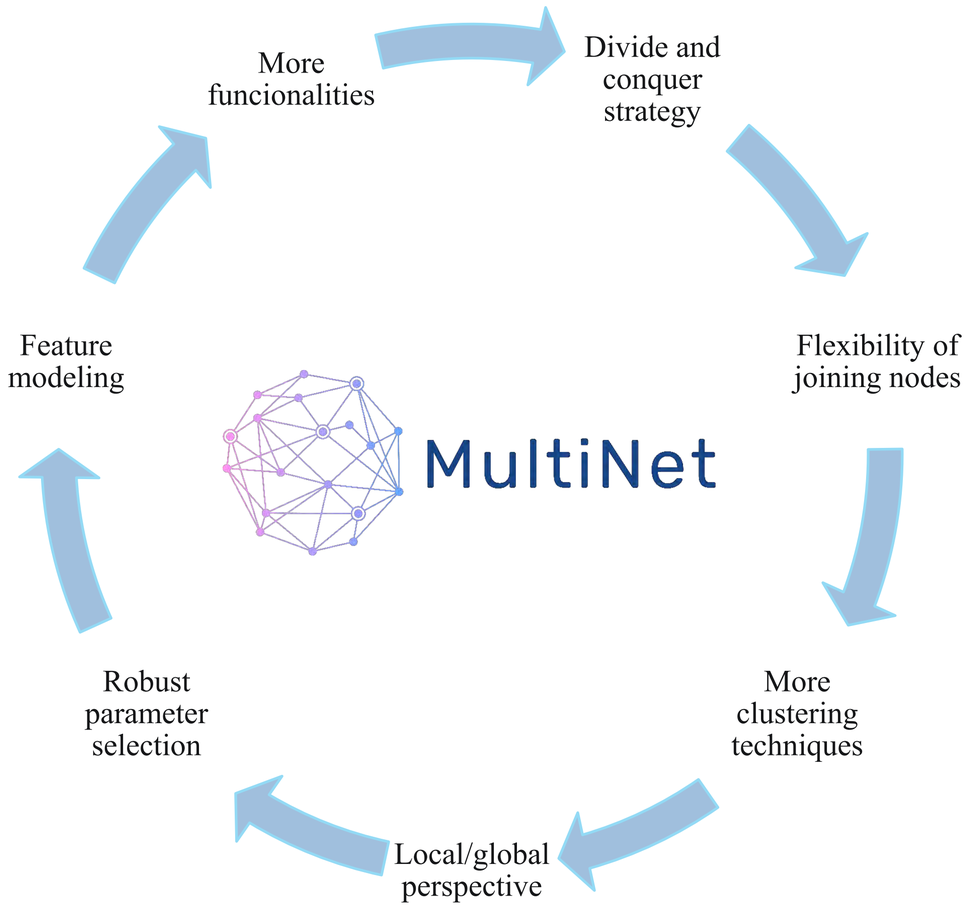
Global Correlation with MultiNet

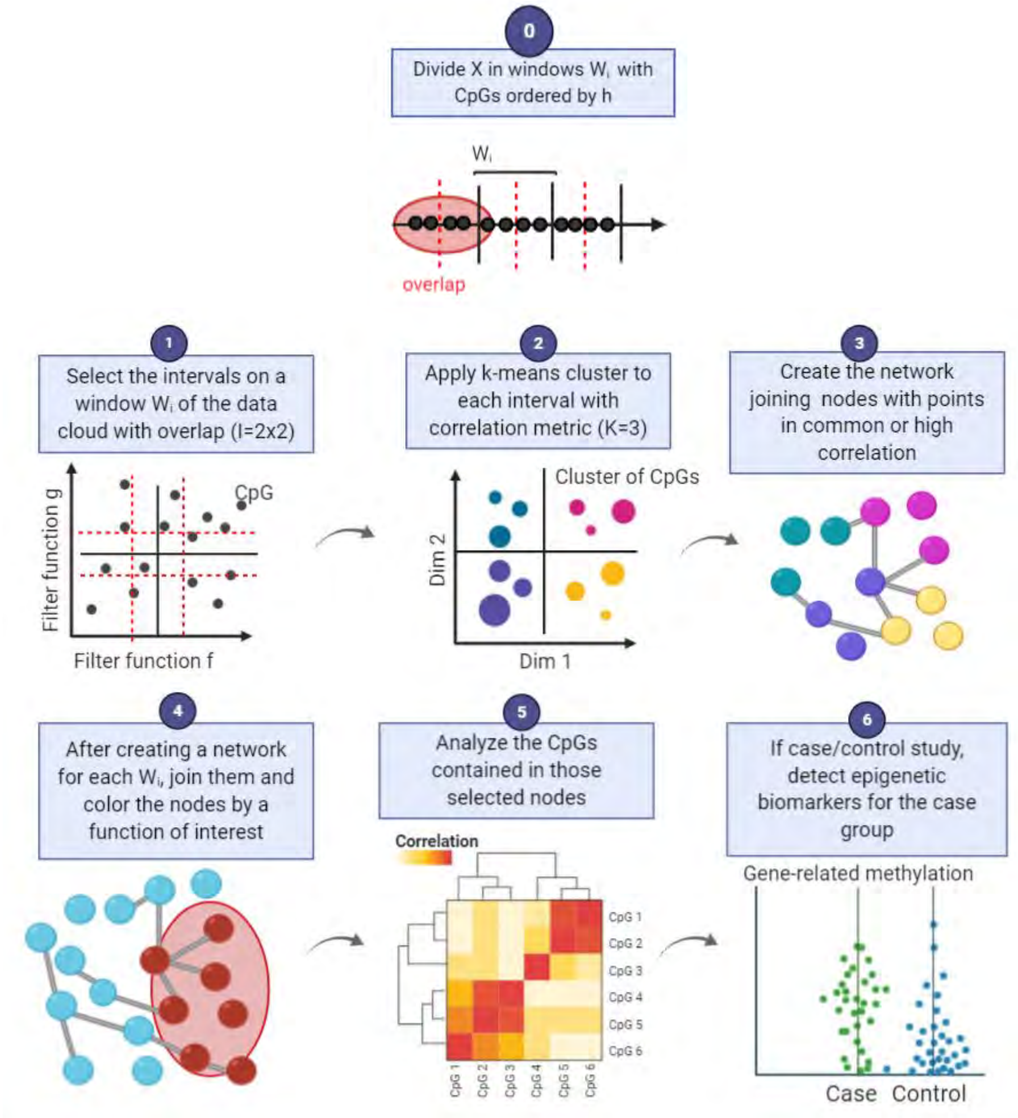


https://github.com/SPRADA1

Process the entire array in less than 50 min!

Global Correlation with MultiNet



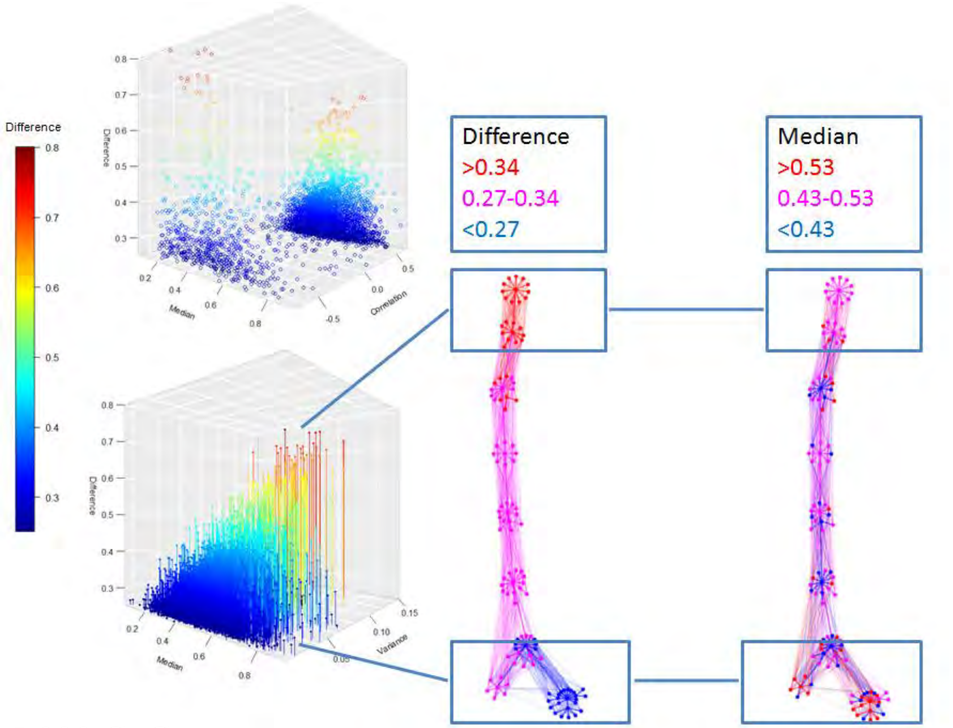


https://github.com/SPRADA1

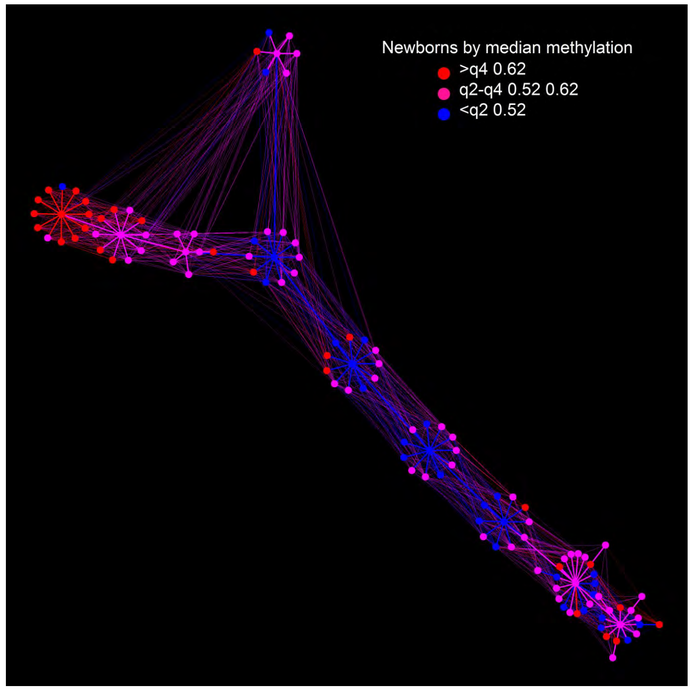
Process the entire array in less than 50 min!

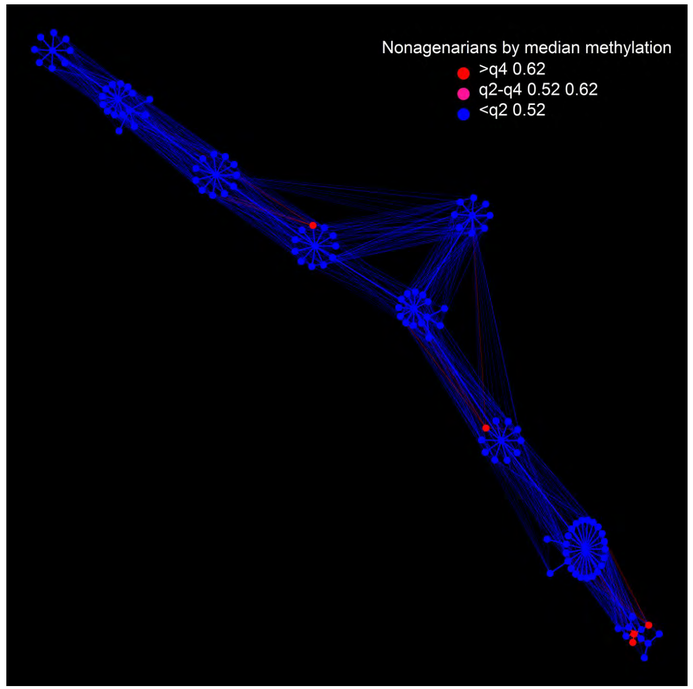
USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Newborns vs. Nonagenarians networks



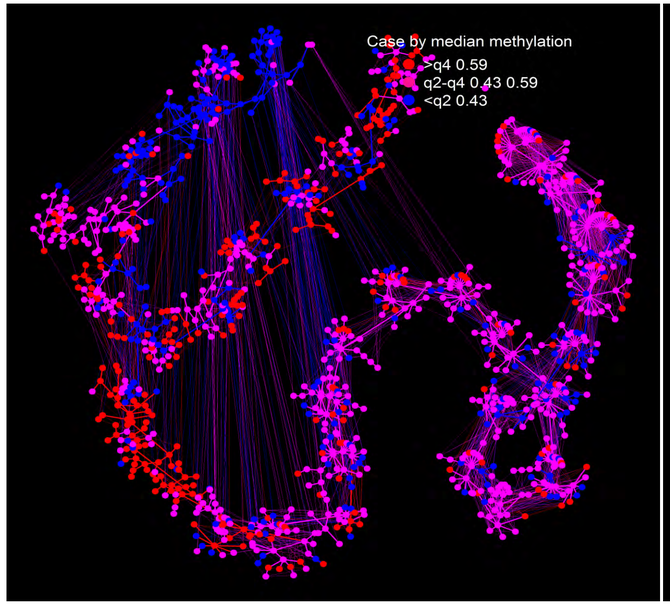
5,000 CpGs

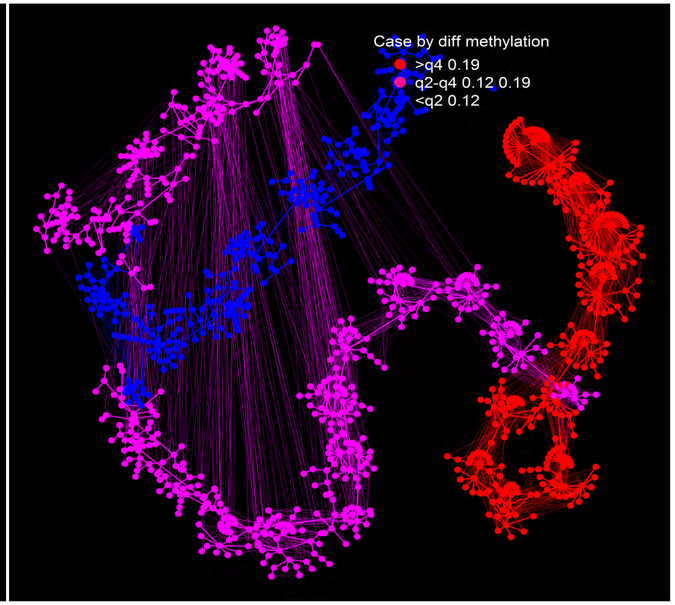


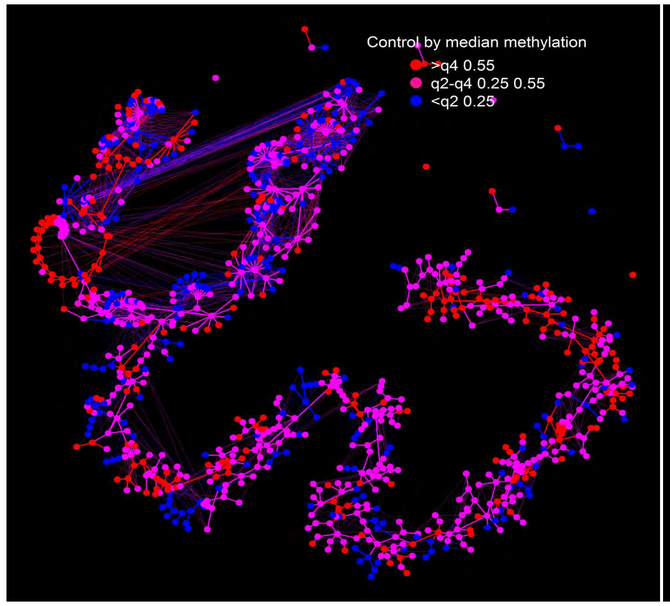


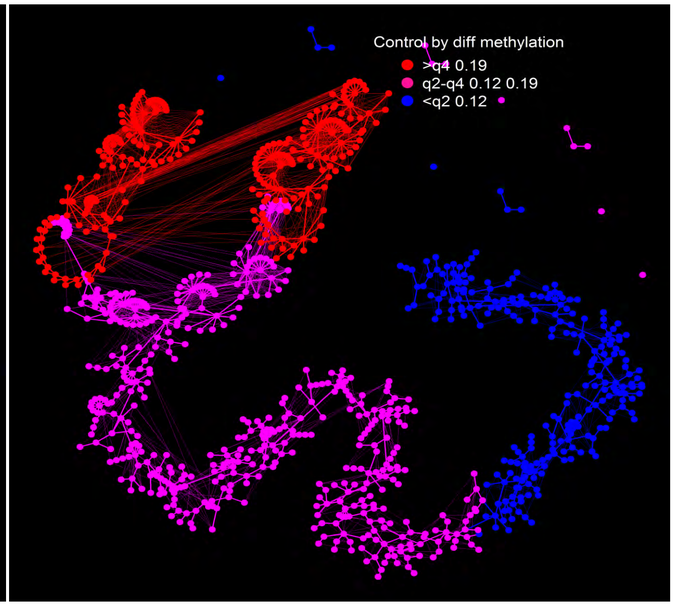
MultiNet detects prostate cancer-status

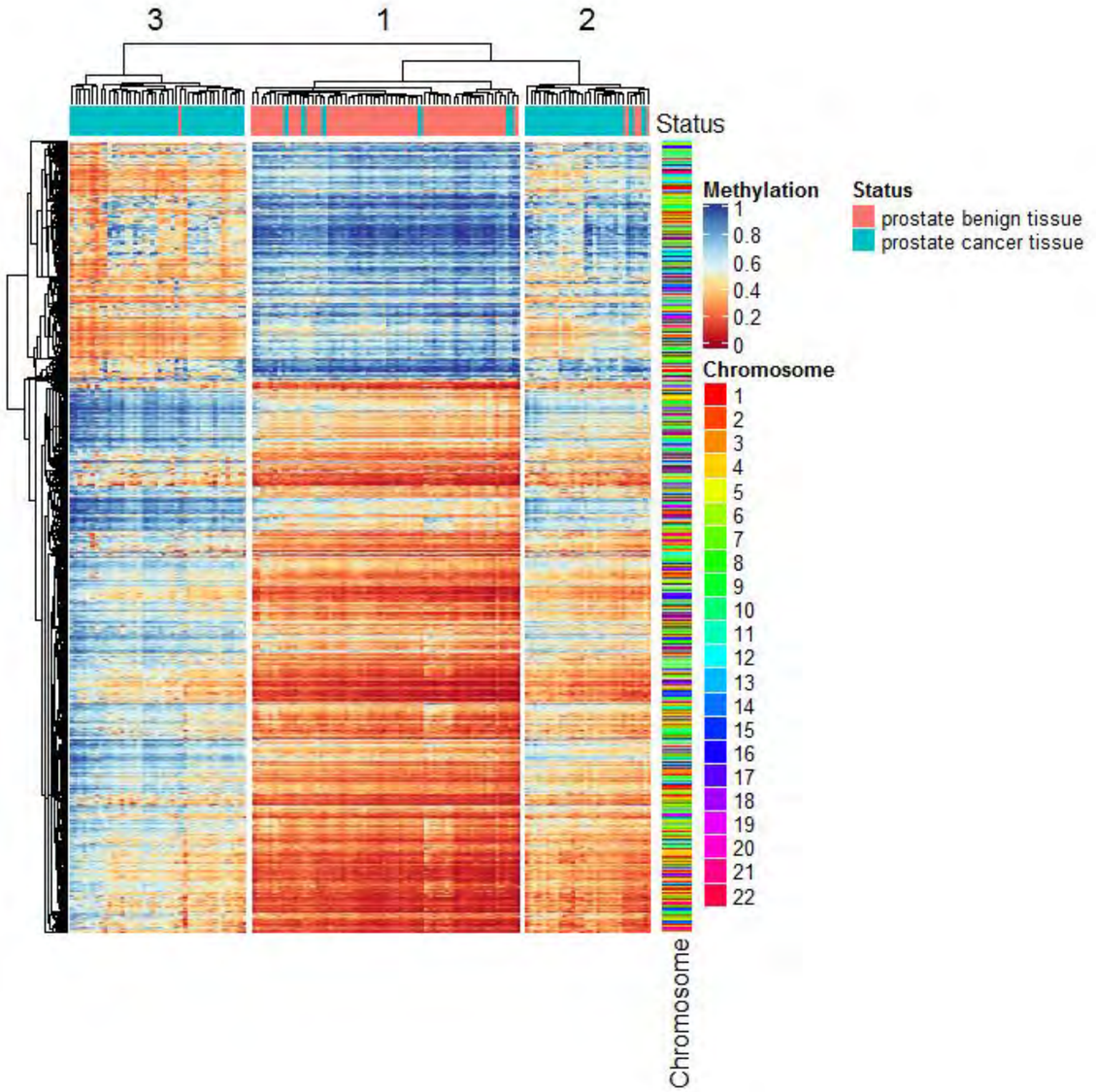
50,000 CpGs



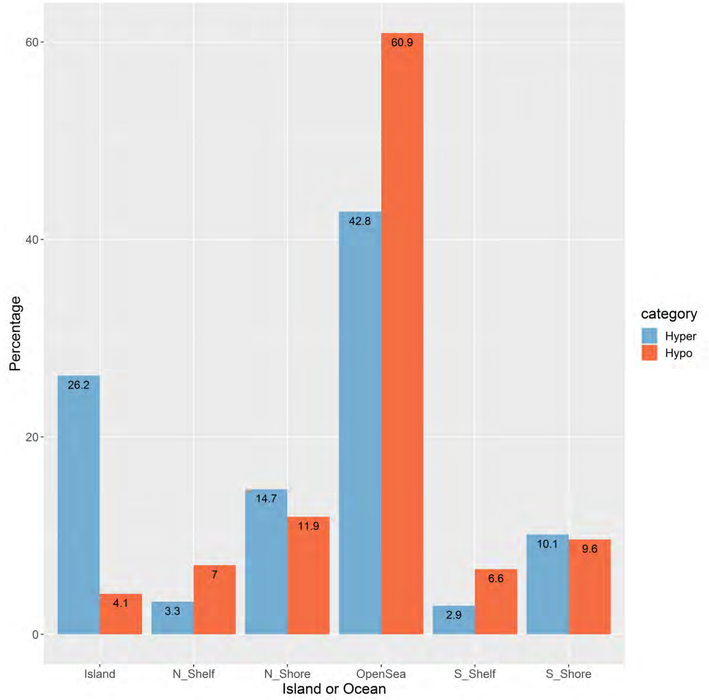


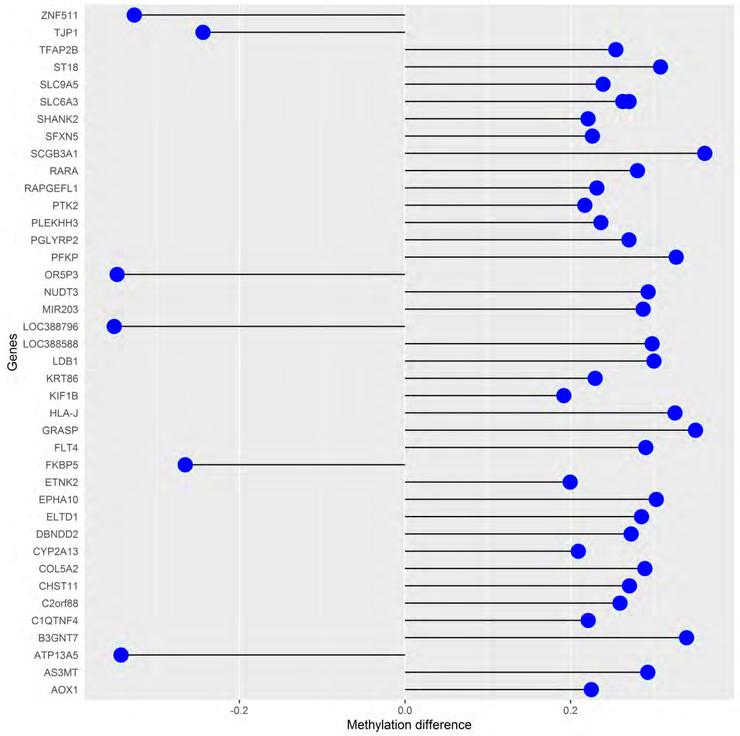






Prostate cancer markers





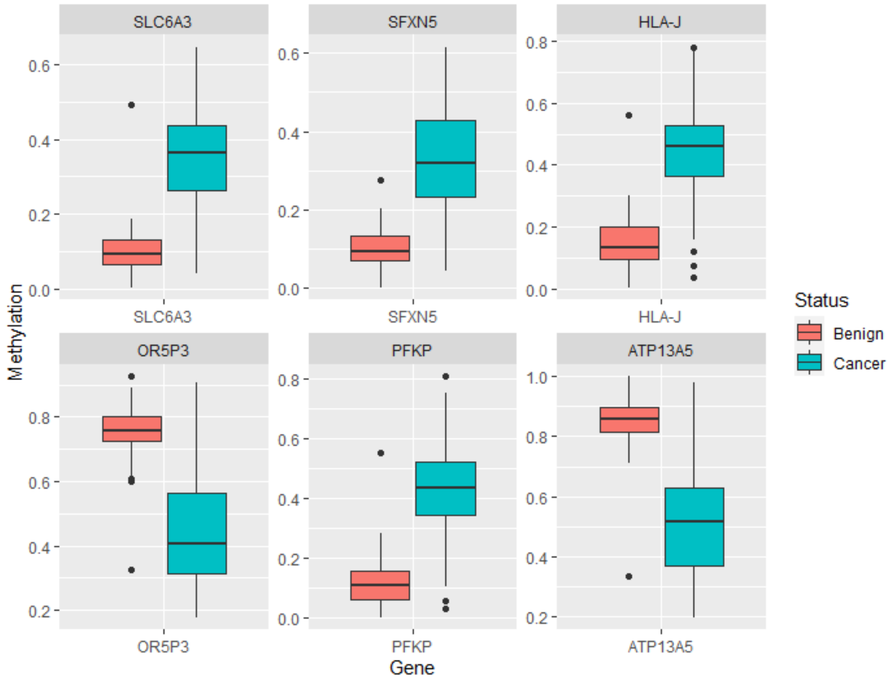


Table 1: 10 of the most relevant CpG sites selected.

CpG ID	Gene; Region	Recent Publication
cg07198194 (hyper)	PFKP; TSS1500	2019
cg21523564 (hyper)	<NA>	<NA>
cg23396786 (hyper)	SFXN5; TSS200	2017
cg16107322 (hyper)	<NA>	<NA>
cg06092265 (hyper)	<NA>	<NA>
cg01748263 (hypo)	ATP13A5; TSS1500	2017
cg15726260 (hyper); cg16794576 (hyper)	HLA-J; Body	2017
cg09729613 (hyper)	AOX1; TSS200	2018
cg04178787 (hyper)	<NA>	<NA>

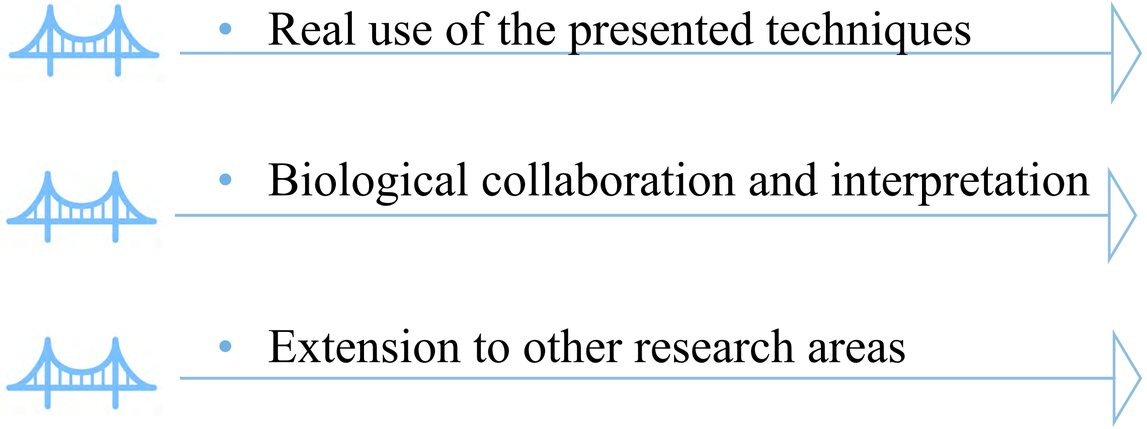
Hypotheses tested with novel techniques

- **New TDA methods**: a mathematical model and a computational algorithm designed to describe and predict the correlation design.
- **Transversality** to succeed: techniques from different mathematical/biological areas were used to understand completely the data structure and model it.
- **Multidisciplinary** design: generating biological hypothesis from observation to be solved analytically with advanced mathematical techniques, spotting a great evolution in both fields which complement each other.

- Real use of the presented techniques
- Biological collaboration and interpretation
- Extension to other research areas

Hypotheses tested with novel techniques

- **New TDA methods**: a mathematical model and a computational algorithm designed to describe and predict the correlation design.
- **Transversality** to succeed: techniques from different mathematical/biological areas were used to understand completely the data structure and model it.
- **Multidisciplinary** design: generating biological hypothesis from observation to be solved analytically with advanced mathematical techniques, spotting a great evolution in both fields which complement each other.



Thank you

SARA PRADA ALONSO
sara.prada.alonso@gmail.com
https://github.com/SPRADA1