Age-dependent DNA methylation Parenclitic Networks in family-based cohort patients with Down Syndrome

M. Krivonosov[1,*], T. Nazarenko[2], M.G. Bacalini[3], C. Franceschi[1,3], A. Zaikin[1,2,4], and M. Ivanchenko[1]

[1]Department of Applied Mathematics and Laboratory of Systems Biology of Ageing, Lobachevsky University, Nizhny Novgorod, 603950, Russia
[2]Department of Mathematics and Institute for Women's Health, University College London, London, WC1H 0AY, UK
[3]IRCCS Istituto delle Scienze Neurologiche di Bologna (ISNB), Italy
[4]Department of Paediatrics and Paediatric Infectious Diseases, Institute of Child Health, Sechenov First Moscow State Medical University, 119146 Moscow, Russia
[*]krivonosov@itmm.unn.ru

ABSTRACT

Network models are a powerful tool to represent, analyze and unfold the complexity of a large-dimensional data system at the fundamental level. The main advantage of network analysis is the opportunity to identify network disease signatures which we use in this paper for patients with Down Syndrome. One of the new methods based on the reconstruction of relations between system features is a Parenclitic Networks approach enabling setting links even for functionally unlinked features. In our work, we develop and generalize the Parenclitic Networks approach using Down Syndrome as a case study. We present our open-source implementation to make the method more accessible to all researchers. The software includes a complete workflow to construct Parenclitic Networks and demonstrate as a generalization that any machine learning algorithm can be chosen as a kernel to build edges in the network using geometric (SVM) and probabilistic (PDF) approaches as examples. We also present a new approach (PDF-adaptive) that allows to automatically to solve one of the main problems of building a network - choosing a "cut-off threshold". We apply our implementation to the problem of detecting the network signature of Down Syndrome in DNA methylation data from the family-based cohort. We demonstrate the first insights into how Parenclitic Networks can be used not only as constructs to reduce data dimension and solve the classification problem using network characteristics, but also as a network signature of transitional changes.

Introduction

In the era of the accumulation of biological Big Data, the most important tasks of machine learning are not only obtaining prognostic models (for example, the case-control classification for disease studies based on genomic, proteomic, epigenetic, and other multi-omic data), but also obtaining information about the relationships within the data, enhancing the understanding of the system as a whole and the processes occurring in it. One of the most effective and powerful tool for solving such problems is network analysis rised as a transdisciplinary effort to characterized network structure and function from the physics of complex systems[1]. It studies the global topology and structural patterns of interactions among the constituents of complex systems, infrastructures, brain and biological networks[2–5]. One difficulty in interpreting Big Data for defining clinically useful information is that multiple different changes may be responsible for the onset of a disease, as exemplified by the efforts of The Cancer Genome Atlas project[6]. On the other hand, very often a system is characterized by a huge set of parameters with unknown interdependencies. To address this challenge Zanin et al.[7] proposed an algorithm to build parenclitic networks, able to establish links between parameters/nodes without any a priori knowledge of their interactions. This approach is based on consideration of pairwise interactions of system features and determination of deviations in each such pair of the case group from the control group. As a measure of deviation Zanin proposed using the length of the perpendicular from each case point to a linear regression that was build on a control group. Parenclitic networks have been successfully applied to problems of the detection of key genes and metabolites in different diseases[8–11] see[12] for a review and[13] for a discussion of applications for brain research. In[14] we have applied this methodology to implement a machine learning classification of human DNA methylation data carrying signatures of cancer development. Later[15], based on the understanding that the interactions of two features (at least in biological systems of biomarkers) often cannot be described by linear model, it was proposed to use 2-dimensional kernel density estimation (2DKDE) to model the control distribution. The deviation is then defined as a

$1-p$-value for a point lying inside the 2DKDE-grid, where p is the total probability of all points with larger probability as the point under consideration (thus the smaller the $1-p$, the farther the point is from the center of distribution and the greater the distance for it) and as normalized distance to the grid for the points outside the 2DKDE-grid.

The main difficulty in using such approaches is the laboriousness of producing a complete analysis package from the beginning to the end (consisting of many technically-complex nested steps) and presence of difficult-to-solve issues (like making "cut-off thresholds" decisions for network edges design). In our work, we generalize parenclitic approach, explore its capabilities and make it more accessible to all researchers. As a case study we use application of parenclitic network analysis to the patients with Down Syndrome from family-based cohort.

First, we present the implementation of a complete workflow to construct Parenclitic Networks; demonstrated that any machine learning algorithm can be chosen as a kernel to build edges in the network using geometric (SVM) and probabilistic (based on the probability density function - PDF) approaches as examples. We also present a new approach (PDF-adaptive) that allows to automatically to solve one of the main problems of building a network - choosing a “cut-off threshold”. We give all the necessary computational estimates for the kernel and indicate their features in connection with their work with specific data. Secondly, we apply our implementation to the problem of detecting the age dependent network signature of Down Syndrome on methylation data. In this study we demonstrate the first insights into how Parenclitic networks can be used not only as constructs to reduce data dimension and solve the classification problem using network characteristics (like "case-control" that was considered in previous works), but also as a network signature of transitional changes.

It is known that impaired development of patients with DS is often expressed in two different directions: 1) underdevelopment of functions: processes that are inherent in the development of a healthy individual, but do not develop in patients with DS with age (for example, a difference in the development of the nervous system); 2) accelerated functions: processes that occur in healthy individuals more often in old age, but occur in patients with DS earlier in terms of an age-time scale (for example, Alzheimer’s disease).

In our work, with the help of different configurations of Parenclitic Networks, we single out separate classes of DNA methylation signatures that can be interpreted from the point of view of the above features. We show how the constructed network models, on the one hand, characterize the boundary states for the case-control system, but on the other hand demonstrate a continuous transition from one state to another one on the age scale, depending on how the class of control was defined in the system configuration (old or young). We assume that this analysis opens up new possibilities for the study of Down Syndrome disease, as well as new opportunities for using Parenclitic Networks in the problems of continuous processes. So we believe that the approaches demonstrated in this article can be used for a more detailed study of diseases (for example, such as cancer), from the point of view of the transition process in time between two case-control boundary conditions. The results of such an analysis based on Parenclitic Networks will not only help in the early diagnosis of the disease (by identifying critical transition marks) and risk assessment, but also shed light on the process itself, through objects involved in it, from the point of view of increasing interaction between them in time.

As an application and demonstration of our implementation, we use DNA methylation data, since on the one hand, such data is presented in huge volume and is interesting from the point of view of our approaches, and on the other hand it is one of very promising and important topics of biology and medicine research because the molecular basis of Down syndrome remains unknown and and identification of a novel network methylation signature may give insight about previously unknown functional dependencies and possible targets for the medical treatment and control. DNA methylation is a chemical modification of the DNA molecule, by addition of a methyl group to the cytosine (C) base. DNA methylation usually occurs at CpG dinucliotides in somatic cells of adult organisms, but in embryonic stem cells it sometimes also occurs outside the context of CpG dinucleotides. Unmethylated CpG dinucleotides tend to be grouped in so-called "CpG islands" that are present in the 5’-regulatory regions of many genes. Hypermethylation of CpG islands in the promoter regions of genes leads to a stable repression of transcription[16].

In recent decades, researchers have discovered that methylation is an important component in numerous cellular processes, including embryonic development, genomic imprinting, X-chromosome inactivation, and preservation of chromosome stability, see[17] and refs therein. Given the many processes in which methylation plays a part, we can also link deviations in methylation to a variety of severe human diseases, especially to cancer, see for the review[18,19]. A majority of diseases cannot be attributed to a methylation error in one site only since changes usually occur in bulk. The increase in the number of such publications and the growing involvement of research groups in the study of methylation issues is due to the development of Illumina Infinium BeadChips technologies, which offered a more affordable, convenient and inexpensive alternative to obtaining methylation data relative to Whole-Genome Bisulphite Sequencing (WGBS). Continuous improvement of the technology made it possible to increase the number of measured CpG sites from 25 000 (HumanMethylation27K BeadChip (HM27), 2008), first to 485 000 (HumanMethylation450K BeadChip (HM450), 2011), and by the latest version, to over 850 000 CpG sites (MethylationEPIC BeadChip (EPIC), 2016). Single site DNA methylation analysis has demonstrated statistically significant epigenetic changes caused by the syndrome, and identified epigenetic age acceleration[20] but the whole picture of network alterations remains hidden. It is especially unclear how links in the network signature alter with the age of a patient. In this work, we present the

first insights into the age-dependent epigenetic markers network through the studying of the characteristics of Down Syndrom disease.

Results

Implementation of workflow to construct Parenclitic Network

The implementation of a complete workflow to construct Parenclitic Networks is available at https://github.com/mike-live/parenclitic. Parenclitic is an open source python library. It can be distributed through PyPI repository. Current version 0.1.6 provides functionality described in this paper.

Our package provides 3 main features:

- Build, save and load parenclitic network.
- Choose or create kernel to identify edges.
- Compute network metrics based on python-igraph package.

Performance tests are done on machine 256Gb RAM, x2 CPU Intel Xeon Gold 6238 CPU @ 2.10GHz, 22 cores per socket, 2 threads per core, 88 threads in sum. Computation time of CpG dataset with 87 samples and 114674 features is 147562 seconds $\sim$ 41 hours.

Kernels of Parenclitic Network

Our implementation demonstrates that any machine learning classification algorithm can be chosen as a kernel to build edges in the network using both geometric ideas (for example, SVM - see details in Methods, Parenclitic algorithm, Kernel: SVM) and probabilistic ones (for example, PDF - see details in Methods , Parenclitic algorithm, Kernels: PDF reconstruction). We also present a new PDF-adaptive approach (see details in Methods, Parenclitic algorithm, Kernel: PDF-adaptive) that allows to automatically to solve one of the main problems of building a network - choosing a "cut-off threshold".

By default, the PDF-adaptive core is proposed (for reasons of high adaptability to data and a reasonably high calculation speed). However, the kernel can be selected based on the following considerations:

- **Asymptotics of method**: it affects the computational speed, generally it is proportional to $n^2 \cdot \text{Kernel}$, where Kernel can be: SVM $= m^2 ... m^3$; PDF $= m \cdot L + m^2, (L \sim 10^5)$; PDF-adaptive $= m^2$ (m is a number of samples in dataset, n is a number of features);
- **Features of the data:** SVM is applicable only for pairs separated by a hyperplane (possibly with bends); PDF is applicable for complex spatial configurations of classes with the ability to control a common threshold in probability for all pairs of parameters; PDF-adaptive is applicable for complex spatial configurations of classes with the ability to automatically select for each pair of parameters of its adaptive threshold;
- **Features of the Kernels:** SVM is more robust and cannot find multiple areas of the same class; PDF simple takes into account the division of the plane into several groups belonging to the same class and able to find extra groups if there are strange outliers; PDF-adaptive is similar to the previous one, but loses a feature of the probabilistic probability cutoff approach and becomes less robust than the previous one.

Applying to detection of signature of Down Syndrome on methylation data

Empirical data

To understanding the method capabilities the analysis has been performed on DNA methylation data profiles of blood cells in families with Down Syndrome child. Normalized data are publicly available in the repository NCBI Gene Expression Omnibus under accession number GSE52588.[20] We used dataset of 29 families: each one consist of a mother and two children - one of them is healthy and other one has a Down Syndrome disease. We distinguished a separate group of mothers (M), healthy siblings (S) and children with Down Syndrome disease (DS).

The HM450 beadchips collect methylation information using two different mechanisms (two different types of probes). All received methylation signals (from a separate site for one sample) are accumulated in the so-called GREEN channel, and all non-methylation signals - in the RED channel. After that, the β-value is calculated:

$$\beta = \frac{\text{GREEN}}{\text{RED} + \text{GREEN} + \alpha}, \tag{1}$$

where α is a small positive constant, to avoid dividing by zero then GREEN and RED both are equal 0 (usually $\alpha = 100$). If $\beta = 0$, then CpG is unmethylated (there is practically no green signal), if $\beta = 1$, then CpG is methylated (no red signal). Thus, one subject patient (one sample) is characterized by 450k scalar numbers - β-values of methylation level for considered sites.

We performed the preprocessing, which includes 3 steps. First, it was a raw data extraction using Minfi[21] Bioconductor package and normalization using the preprocessFunnorm function implemented in the same package to exclude batch effects in obtaining methylation levels. Second, we performed the removal of cross-reactive probes and the probes whose DNA targets contain SNPs[22] (60467 probes); third was the removal of sex chromosomes X and Y (9808 probes). Finally, 412993 probes from the original array were used. Due to the large number of features it is necessary to reduce dimensionality. One way to reduce the dimensionality of data is to use an average methylation level calculated for a certain area of DNA or considerations of only certain area of DNA. In our study, we are concentrating only on Islands and Shores regions (for the preservation of the biological meaning) and using two types of data:

- **CpGs set** - a subset HM450 probes from Islands and Shores regions (114674 CpGs);
- **Genes set** - a data set in which each gene is represented by averaging the β-values of those probes from the **CpGs set** that lie within the given gene (14756 genes).

Building different types of parenclitic Genes- and CpGs - networks

We analyze separately **Genes set** and **CpGs set** and form the results for them in the *Genes-Networks* and *CpGs-Networks* respectively, applying PDF-adaptive Kernel.

We concentrated on exploring of different age effects. Based on the fact that the population of healthy samples is well divided into groups of age-related categories (see the distribution of ages in Fig.1) through the labels of family members (Mothers and Siblings), we chose three methods for constructing Parenclitic Networks.

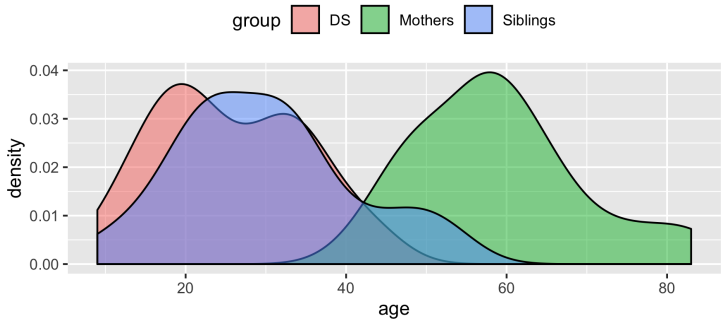

Figure 1. Age-distribution for M/S/DS groups

In one of them (we will later define it as the M-control (Genes/CpGs)-Network) we concentrated on selection of attributes that distinguish DS patients from a healthy older population (Mothers) and then, on the selected attributes, we also considered the behavior of their peers (Siblings). In another construction (we will later define it as the S-control (Genes/CpGs)-Network) we concentrated on selection of distinguishing attributes of DS patients from a healthy population of the same age category (Siblings) and then, on the selected attributes, we also considered the behavior of older people (Mothers). In the third construction (we will later define it as the DS-control (Genes/CpGs)-Network), we aim to select the attributes that distinguish the DS patients from the entire healthy population (without reference to age). Our task is to study the attributes in these three groups (to identify common and unique parts in them), to establish their relationship with age (where possible, that is, in M-control (Genes/CpGs)-Network and S-control (Genes/CpGs)-Network, on the basis of a samples not involving in the attributes selection) and to study them in terms of biological functions and epigenetic features.

We built Parenclitic Networks using three different configurations, described in labeling-terms of the Input Data of our algorithm in accordance with Table.1.

	Y for M group	Y for S group	Y for DS group
M-control (Genes/CpGs)-Network	-1	-2	$+1$
S-control (Genes/CpGs)-Network	-2	-1	$+1$
DS-control (Genes/CpGs)-Network	$+1$	$+1$	-1

Table 1. Y-labeling of M,S and DS groups for building different types of Parenclitic (Genes/CpGs)-Networks

The labeling is structured as the following:

1. **M-control (Genes/CpGs)-Network** — M group (label is $Y = -1$, i.e. *control* group in terms of original Parenclitic Network) is selected as a group for the construction of PDF. Using the solution in the kernel of the PDF-adaptive, the best density threshold is selected that separates the group M from the DS group (label is $Y = +1$, i.e. *case* group in terms of original Parenclitic Network). After choosing this value, any point of the initial set (S, M, and DS together) receives a value of -1 if it is inside the PDF (taking into account the selected threshold) and a value of $+1$ if it is outside it. Further, for M and DS, the metric of Accuracy is considered (for the original Y and new labels assigned by the algorithm). If Accuracy> 0.9, then the corresponding pair of signs is recognized as "significant" and taken into account when constructing the network: the edge will be built for all new labels $+1$ and will be skipped for all new labels -1. Note that in this case, the set S (label is $Y = -2$) is not involved in the construction of the graph and the decision to select of "significant" pairs. Therefore, this set will be *test* in this case, and with it we will be able to interpret the features obtained in this case of Networks.

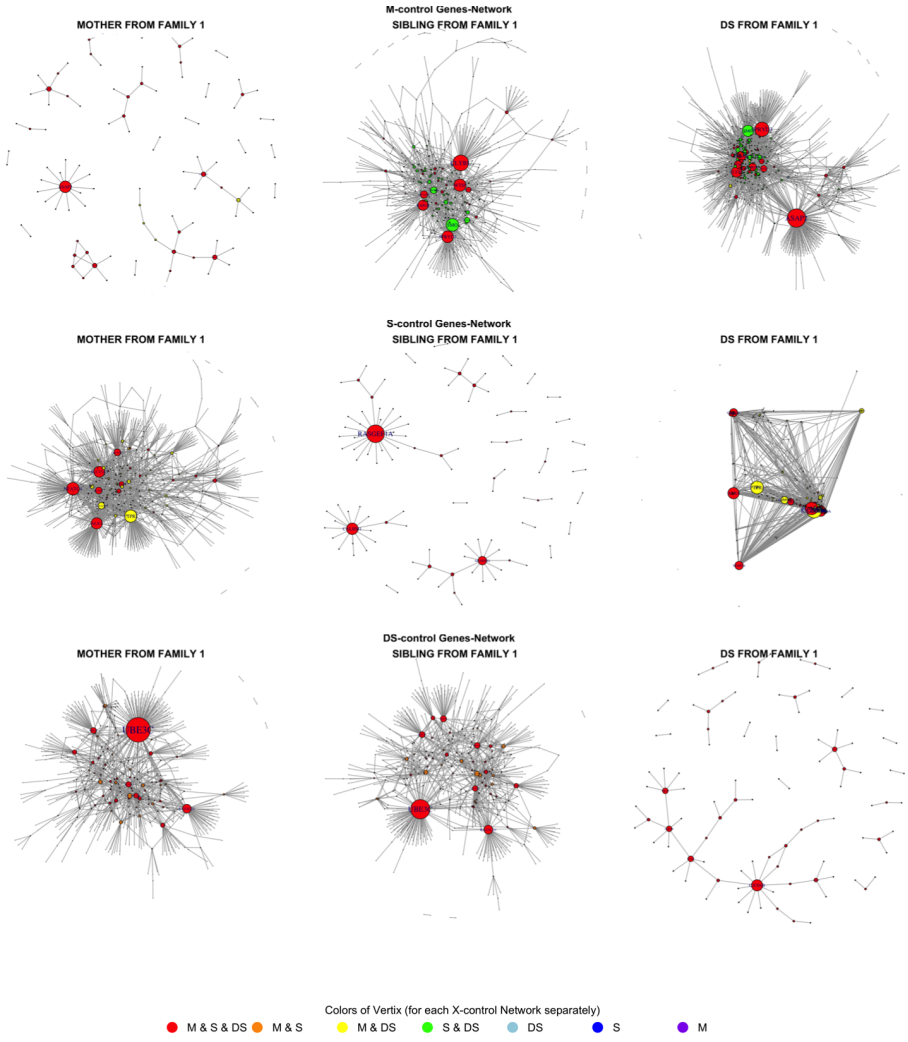

Figure 2. Illustration networks for members of one family in different constructions of Genes-Networks. Rows are the different constructions of Genes-Networks: M-control Genes-Network, S-control Genes-Network, DS-control-Genes Network respectively, columns are the family members: Mother, healthy Sibling, DS. Each case the participant of the control group is characterized by a discharged network structure; the participant of the case group is characterized by a dense and connected network structure, and the participant of the test group (for the first two networks) is characterized by an intermediate state (between discharged and density), and this state varies with age.

2. **S-control (Genes/CpGs)-Network** — S group (label is $Y = -1$, i.e. *control*), DS - deviated group ($+1$, *case*), M is a (-2, *test*). All ideas of construction are similar to those described above for M-control (Genes/CpGs)-Network) (taking into account the changing roles of groups S and M).

3. **DS-control (Genes/CpGs)-Network** — DS group (label is $Y = -1$, *control*), union S and M groups are deviated group (label is $+1$, *test*). In this case, we do not have a *test* set, but we assume that the set of pairs of characters selected in this case will characterize the strict difference of the set of DS from healthy samples.

For the illustration the Genes-Networks examples for the same samples (representatives of the same family) in structures M-control, S-control and DS-control are shown on Fig. 2. We can say that in each construction the Parenclitic Network Algorithm esteblished a multidimensional boundary in the selected set (own for each construction) of attribute pairs through which, the set of the control group is very clearly separated from the case group. Samples of the control group are located on one side of such a multidimensional border (that is, they have almost no edges in their networks), and the group of cases turned out to be on the other side of this multidimensional border, which is indexed by an edge in each pair of selected signs, as an indication that they lie on the other side of control group. Thus, the illustrations of the networks for the participants in the control (see examples for Mother network in 1 row, Fig. 2, Sibling network in 2 row, Fig. 2, DS network in 3 row, Fig. 2) and case groups (see examples for DS network in 1 row, Fig. 2, DS network in 2 row, Fig. 2, Mother and Sibling networks in 3 row,Fig. 2)) clearly demonstrate the global topological differences on the selected attribute systems. It is interesting to consider the participants in the test group in the first two constructions (see examples for Sibling network in 1 row, Fig. 2 and Mother network in 2 row, Fig. 2). This networks are characterized by an intermediate state (between control and case networks). In order to investigate such intermediate states and establish their relationship with age, we turn to the analysis of the topological characteristics of the received networks for each participant in the dataset in each network construction.

In the entire description that follows, under the *X*-control (Genes/CpGs)-Network, we mean the **union** of all edges constructed for all samples in the *case* group (taking into account that the construction method already includes the selection of highly separating features and, as will be shown below (Fig.3-B), each edge of the constructed network belongs to more than 70% of *case* samples). The obtained *X*-control (Genes/CpGs)-Networks are available for further analysis and are contained in the form of lists of edges, indicating for each edge the proportion of samples in each group that have it (Supplementary Results, *Network* folder).

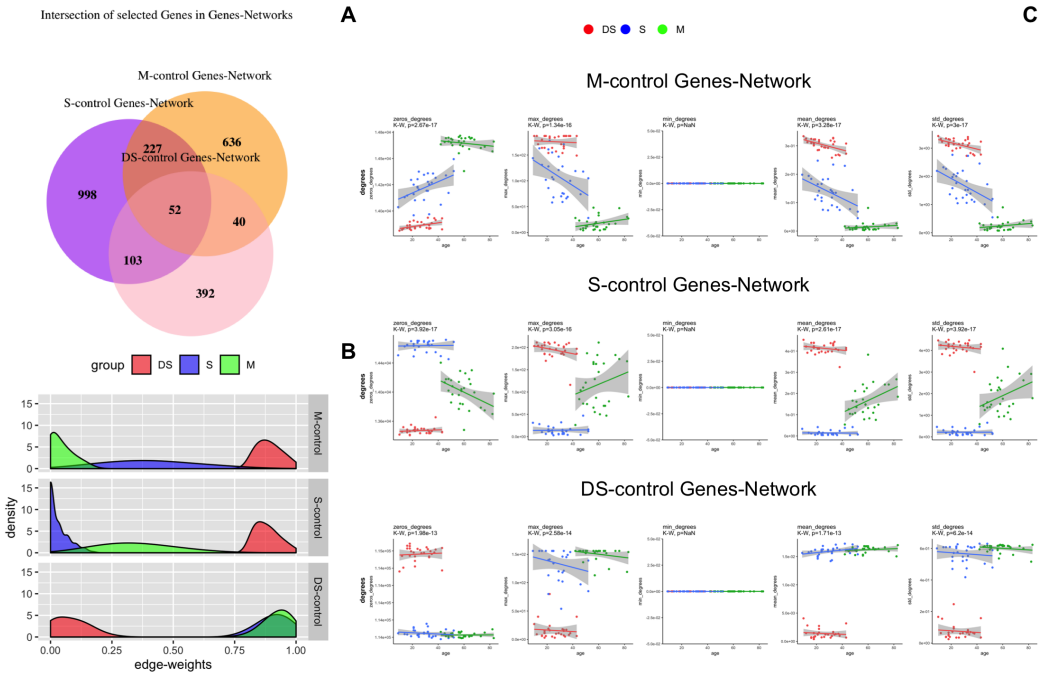

Figure 3. Characteristics of Genes-Networks. **A**: Intersection of genes (Networks vertices) of M/S/Ds-control Genes-Network, demonstrating that most of the attributes selected by different constructions are unique to them; **B**: distributions of edge-weights (proportions of samples in each M/S/DS group which have edge) in each construction M/S/Ds-control Genes-Network; **C**: Vertices degree characteristics for each construction M/S/Ds-control Genes-Network, demonstrating that test set (in the first two constructions) indicates the direction of the age-dependence between the two critical control-case states

Networks analysis and identification of network-age-dependent effects

The approach described above allowed the selection of set of significant pairs of Genes and set of significant pairs of CpGs in each configuration. It is interesting that the selected sets of objects practically do not intersect for the selected configurations (see Fig.3-A for intersection of configurations in Gene-Networks and CpGs-Networks).

It is important to note that the selection of more attributes into group S-control *X*-Network means that group S, has more compactness by a pair of attributes (i.e., less variability within itself), which allows it to better separate from the DS group (then S selected as *control*). On Fig.4, we give an example of a pair of KAZALD1 and PLTP genes (Fig.4-A) and show how PDF is constructed if the *control* group selects as DS (Fig.4-B) and as S (Fig.4-C). It can be seen that the KDE capture region on (Fig.4-C) is tighter then on (Fig.4-B) and allows to precisely separate S group from DS group, but on (Fig.4-B), the capture region becomes wider and includes two points of the opposite group.

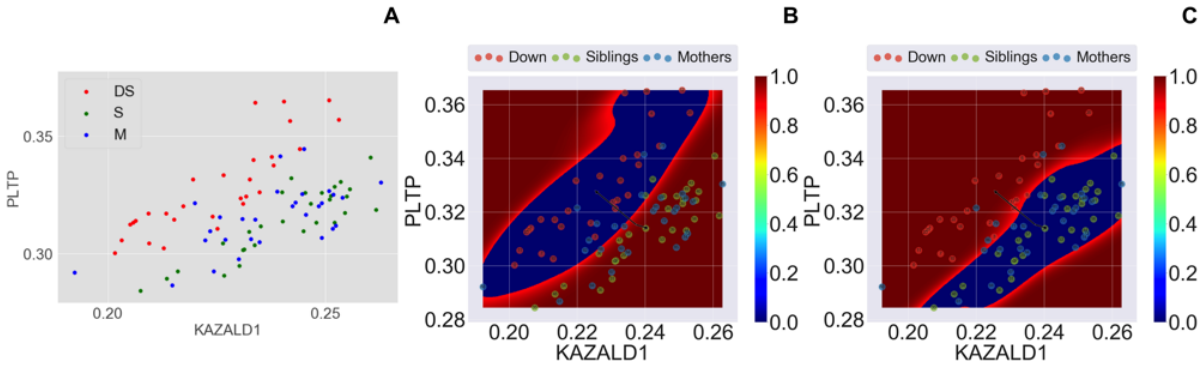

Figure 4. Example of the construction of KDE for pairs of (PLTP and KAZALD1) genes. **A**: Original plot of values, in which group green points (S) is visually separated from the group of red dots (DS). Moreover, several blue dots (M) pass from the S region to the DS region, as our analysis showed, this transition is associated with a higher age for such subjects. **B**: example construction of KDE then DS group is a *control* group, demonstrating that due to the high variability within the control group (DS), the KDE region has a wider structure and captures two points of the opposite class (S) **C**: example construction of KDE then S group is a *control* group, demonstrating that due to the stronger compactness of the group of green points (S), the KDE area is built very accurately, without capturing the points of the opposite case class (DS)

For each configuration and each edge in it, we calculated *edge-weight* - fraction of samples in each group M, S, DS it was built and examined the distribution of these characteristics. We show that the distribution of edge-weights in the *test* groups (M for the S-control and S for the M-control) formed between edge-weights distributions of the *case* and *control* groups on which the network was built (Fig.3-B).

We analyzed all the characteristics of the obtained networks and found that the characteristics for the *test* groups (in S-control and M-control configurations) not only lie between the characteristics of the other two groups, but also demonstrate the transition from one group to other one with age (a full description for Genes-Networks characteristics is given in the Supplementary Materials,Fig.1-3 and for CpGs-Networks characteristics is given in the Supplementary Materials,Fig.4-6), as, for example, vertices degree characteristics in each Genes-configuration (Fig.3-C).

Considering the characteristics of networks relative to age allows us to make the following interpretations for each network configuration:

- **M-control (Genes/CpGs)-Network** - this is a state of the feature system, which is characterized by close relationships for the DS group and the discharged structure for the M group, regardless of the age of the individuals in the groups. That is, in this feature system, the groups DS and M are clearly distinguishable sets. The behavior of the characteristics of *test* group S (healthy Siblings) on the age scale (see, for example, vertices degrees characteristics on Fig. 3-C, 1 row, and full sets of characteristics in Supplementary Materials: Fig.1 for Genes-Network and Fig.4 for CpGs-Network) shows a transition from the values of characteristics close to DS group to values close to the values of characteristics to M group. Apparently, the attributes selected in this system characterize a system of properties that changes more slowly (or does not change at all) with age in the DS population than in a healthy siblings population. Therefore, processes united by the attributes of M-control Network can be called *"Decelerated Age Processes" (DAP)* for DS population. We note that we use the term *Decelerated* only to reflect the fact that in the attributes of this network, DS patients are in a state, which is close to the state characteristic of a very young healthy population, which is rather associated with the term *underdevelopment* for DS pacients. We do not indicate that such processes in DS themselves are slowed down in time

(this issue requires a separate study), but only that the performance of this system does not correspond to the physical age of patients with DS and mostly correspond to younger healthy population.

- **S-control (Genes/CpGs)-Network** - this is a state of the feature system, which is characterized by close relationships for the DS group and the discharged structure for the S group, regardless of the age of the individuals in the groups. That is, in this feature system, the groups DS and S are clearly distinguishable sets. The behavior of the characteristics of *test* group M (Mothers) on the age scale (see, for example, vertices degrees characteristics on Fig. 3-C, 2 row, and full sets of characteristics in Supplementary Materials: Fig.2 for Genes-Network and Fig.5 for CpGs-Network) shows a transition from the values of characteristics close to S group to values close to the value of characteristics to DS group. Apparently, the attributes selected in this system characterize a system of properties that has changed in the DS population in a way that could potentially change in a healthy population in old age. Therefore, the processes combined by the attributes of S-control Network can be called *"Accelerated Age Processes" (AAP)* for DS population. We note that we use the term *Accelerated* only to reflect the fact that in the attributes of this network, DS patients are in a state, which is close to the state characteristic of a very old healthy population. We do not indicate that such processes in DS themselves are accelerated in time (this issue requires a separate study), but only that the performance of this system does not correspond to the physical age of patients with DS and mostly correspond to older healthy population.

- **DS-control (Genes/CpGs)-Network** - this is a state of the feature system, which is characterized by close relationships for the S and M groups and the discharged structure for the DS group, regardless of the age of the individuals in the groups. That is, in this feature system, the unhealthy group (DS) and healthy group (M and S) are clearly distinguishable sets. In this case, we don't have a *test* set, but a high degree of separation of characteristics indicates that the process system selected by this network configuration characterizes a stable system that distinguishes the DS group from the group of healthy people regardless of age. Therefore, the processes combined by the attributes of DS-control Network can be called *"Age-Independent Processes" (AIP)* which go in DS and healthy population groups in different ways and are not associated with age. We note that we use the term textit Age-independent here only to reflect the fact that in the attributes of this network, patients with DS are in a state that distinguishes them from any healthy patients, regardless of age. This does not mean that the processes included in this network cannot be extreme cases for the two processes described above (this issue requires further study).

A visual representation of these interpretations can be shown by Fig.4-A. It can be seen that the green and red clouds (S and DS) are visually distinguishable; the blue cloud (M) is closer to the green, but some blue points are shifted in this pair of signs to the red cloud. That is, the function of a pair of these genes for M shifts with age to that functional region that is stably characteristic of DS at any age. This suggests that the DS group itself in this place has an accelerated age process (not typical of their age in a healthy population). We also note that a pair S-control these signs was recognized as significant and selected by an edge only in configuration S-control Genes-Network (the percentage of groups containing this edge: DS group: 1.0, S group: 0.0, M group: 0.2413793), and was not selected in (M or DS)-control Genes-Network. It confirms our interpretation about association between the age-related transition and states for different types of network configurations.

Analysis of the functional features of networks (histone marks enrichment analysys for CpGs-Networks and gene onthology enrichment analysys for Genes-Networks)

- **Histone marks enrichment analysys:** In each configuration of *X*-control CpGs-Networks, we distinguished two subgroups: *X*-control (hyper) - those CpGs sites in which the average β-values of DS group is greater than the average β-values in M and S groups at the same time) and *X*-control (hypo) - those CpGs sites in which (the average β-values of DS group is less than the average β-values in M and S groups at the same time) and analyzed all selected sets using eFORGE v2.0[23]. The results are shown in Fig.5. It can be seen that all three of our *Process* groups differ in the combinations of enrichment of histone modifications: for the (hyper) M-control group is characteristic the significance of enrichment with modifications HX3K4me1, H3K27me3 and H3K4me3 is high; for the group (hyper) C-control group - the high significance of enrichment with modifications H3K4me1 and H3K27me3; and all three (hypo) groups are characterized by a high significance of enrichment with H3K36me2 modifications. We do not go into a detailed analysis of these results, but simply emphasize that the 3 different groups of processes selected by our study are indeed characterized by different epigenetic patterns. Full eForge Analysis results are given as separate files in Supplementary Results (folder *eForge*).

- **Gene Ontology Analysis:** For each *X*-control Genes-Networks, we performed gene ontology analysis of vertices (genes), using using an online Gene Ontology resource (GO[24]; http://geneontology.org). We selected data from GO biological processes complete (GOBP), GO molecular functions complete (GOMF), GO cellular component complete (GOC), and PANTHER protein class complete (PPC). The general statistics of the identified significant processes (with

	GOBP	GOMF	GOC	PPC
M-control (Genes)-Network	108	11	12	6
S-control (Genes)-Network	70	8	16	4
DS-control (Genes)-Network	–	1	2	-

Table 2. Number of GO identified significant processes (FDR < 0.05) for M/S/DS-control Genes-Networks

FDR-adjusment < 0.05 and indicated as Overrepresented) is as follows: Illustration of the mutual intersections processes of M/S/DS-control Genes-Networks for GOBP, GOC and PPC are given in Supplementary Materials, in the Fig. 7-A,B,C respectively (and full GO Analysis results are given as separate json files in Supplementary Results, folder *GO*). Here we present only an illustration for GO molecular functions complete (Fig. 6).

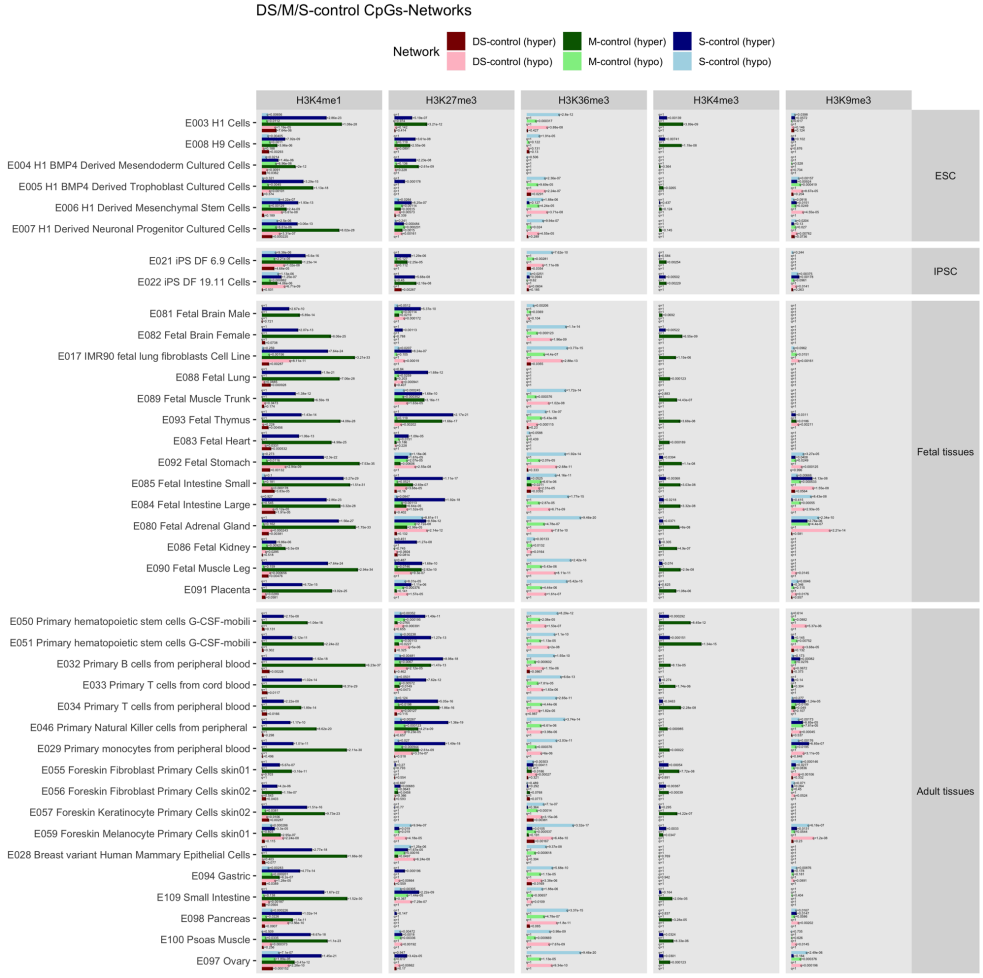

Figure 5. Results of histone marks enrichment analysis. Demonstration of different combination significant enrichments for different constructions of CpGs-Networks (with their division into hyper and hypo subgroups). Results were obtained by eFORGE v2.0 Analysis: The Benjamini–Yekutieli (BY) multiple-testing corrected q-value is evaluated to mark enrichments as significant at q< 0.05

Suggestions for continuing the study of identified processes on maps of built networks using Wnt signaling pathway is an example.

Here we show that the traditional analysis of gene ontology performed can be continued as a part of the study of constructed physical networks of interactions. For the sake of a demonstration, we chose one of the special processes found in the configuration of M-control Genes-Network - Wnt-activated receptor activity process (see orange cloud Fig. 6). In our interpretation, this network was called as "Decelerated Age Processes" (that is, it contains processes that are different on set DS

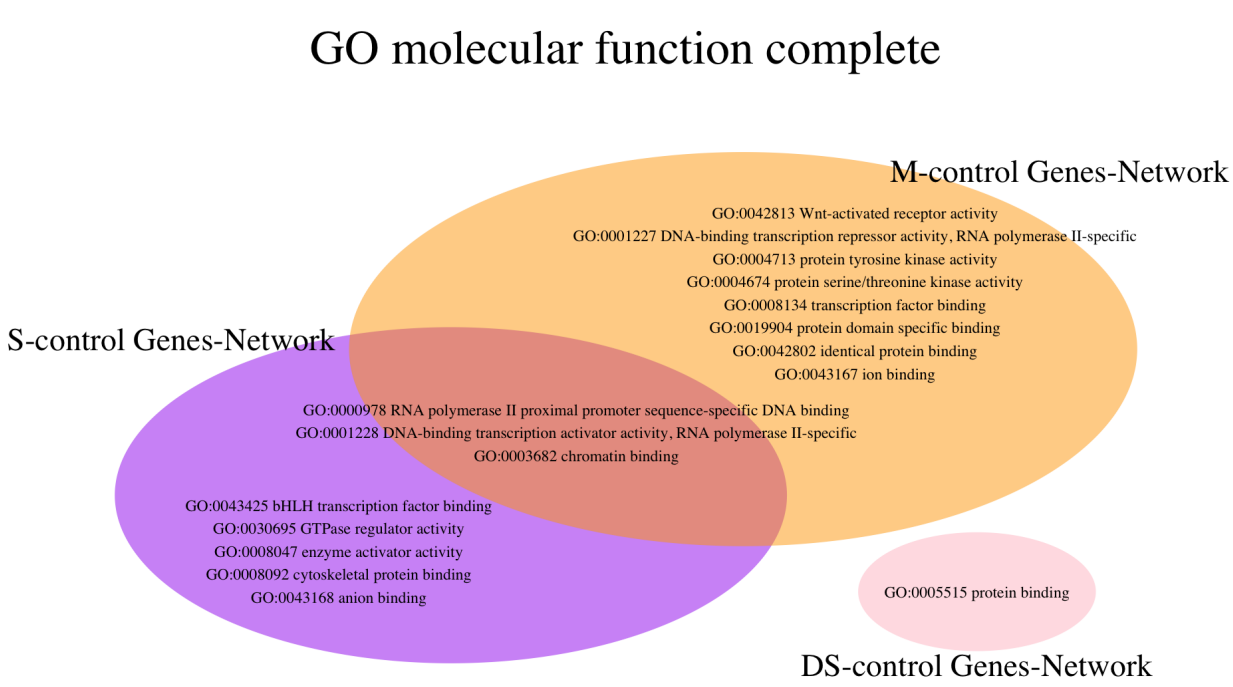

Figure 6. Part of gene ontology analysis - GO molecular functions complete for vertices of each construction M/S/DS-control Genes-Networks, demonstrating that attributes of each network mainly involved in different functional process

and M, but S set (healthy children) with age are transmitting from a state close to DS to a state close to Mothers in this system). The Wnt signalling pathways are very demonstrative in our study. On one hand, the scientific community associates it with oncogenesis[25,26], on the other hand, this signaling pathway plays an important role in the development of the nervous system in adults[27], and is highly associated with disorders of this development, for example, Alzheimer's disease[28].

It is noteworthy that DS patients are a group of patients that specifically unite both of these areas: on the one hand, DS patients are less likely to have a cancer[29,30], and on the other hand they are more likely to develop Alzheimer's disease[31] (and this happens for them at an earlier age than in a healthy population). We suggest that identifying this pathway in our age-specific system can help to reveal not only its role for the DS population, but also, on the whole, to more accurately understand its own functions and features, through further study of the relationships of genes associated with the identified pathway, through connections with other genes in our system.

We examined the initial localization of Wnt-associated genes *(FZD3, LRP6, PKD1, LRP5, FZD6)* in the whole M-control Genes-Network (Fig.7,A-I) and within the subgraph defined by only these Wnt genes vertices (Fig.7,A-II); we considered similar states when connecting the 1st (Fig.7,B-I,II) and 2nd (Fig.7,C-I,II) neighborhood levels. As shown, this is enough for seizing Wnt genes into one connected subgraph.

We checked the genes of the 1st and 2nd levels of neighbourhood in Gene Ontology and found, that the 1st neighborhood level did not produce significant results in any analysis, but the 2nd neighborhood level gave significant results of enrichment in each analysis:

- GO biological process complete: negative regulation of cellular process;
- GO molecular function complete: RNA polymerase II proximal promoter sequence-specific DNA binding, protein kinase activity, protein binding;
- GO cellular component complete: TOR complex, cytoplasm, intracellular membrane-bounded organelle.

Figures 7,C-I,II also show that the 2nd neighborhood level is the heart of the central cluster and, possibly, glues all the signs into one system. In this regard, we note that the 2nd neighborhood level was connected in terms of molecular function called "protein binding" with a DS-control Genes-Network, and in terms of molecular function "RNA polymerase II proximal promoter sequence-specific DNA binding" with a S-control Genes-Network (see Fig.6), what is possibly the main function entailing main violations (age-related or not) for the DS population.

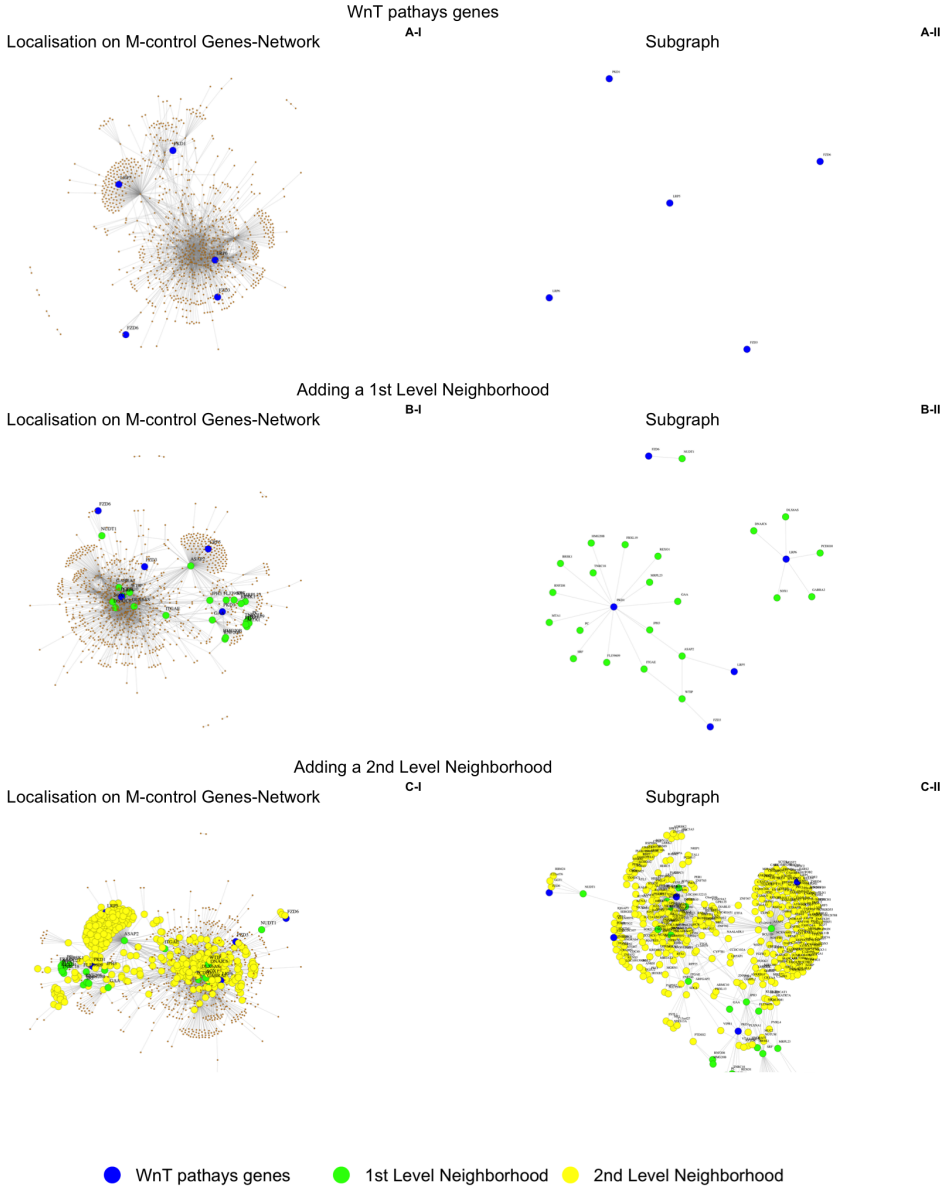

Figure 7. Wnt-associated genes on M-control Genes-Network. **A** Wnt-associated genes on whole network (I), and on within own subgraph (II); **B** Wnt-associated genes with 1st neighborhood level on whole network (I), and on within own subgraph (III); **C** Wnt-associated genes with 1st and 2nd neighborhood levels on whole network (I), and on within own subgraph (II)

Discussion

In this study, we presented our open-source implementation of Generalized Parencltic Network analysis to discuss how we have generalized it and to make it more accessible to all researchers. The main methodological advance is an introducing new kernels and discussing the possibility of their selection based on the goals of the problem. We believe that the simple integration of new kernels into the overall implementation will allow researchers to use not only the proposed methods, but also connect and test their own ideas.

As a case study we applied our implementation to the construction of DNA methylation Parencltic Networks for a family-based cohort patients with Down Syndrome and revealed a network-age-dependence effects. We gave interpretations of the constructed networks as a group of functions *decelerated* and *accelerated* in the age scale for DS patients. Our study allowed for the first time to decompose the epigenetic signature accompanying DS patients into such categories based on the hidden

links between the covariates.

We believe that the networks built in this work are not the end of the study, but only the beginning, because new information obtained in the course of network analysis can be now widely analysed by biologist and clinicians in order to identify molecular mechanisms resulting in and accompanying Down Syndrome as well as identify new molecular targets for treatment of patients with Down syndrome to prevent their accelerated ageing. The networks built in this study require more detailed analysis and can help researchers involved in Down Syndrome research to discover new interpretations based on the interactions detected.

In this work, we especially single out the Wnt signaling pathway process that was logically integrated into this study by the sum of the properties known about it and the features of the patients considered in this work. We believe that it would be interesting to investigate it (as well as other highlighted processes) on a physical network map and examine all associated edges with it. We also believe that the property of the networks themselves in terms of *decelerated* and *accelerated* processes of patients with DS can also help formulate hypotheses more powerfully.

We particularly highlight here that the main idea of Parenclitic Network approach, consisting in the separation states of group case-control, can be also apply for analyzing the transition in time (by age or other continuous scale) from one such state to another. The *test* set (which is located between two critical states within the framework of the process under study) can become an additional improvement in the study of complex systems. In our work, we used different constructions of Parenclitic Networks by determining critical states (control and case) in different ways. For example, in the S-control Network, where peers of DS patients were selected as the control group, and DS patients as the case group. Firstly, we identified attributes in which these two groups were well separated. Then, all the selected attributes of the system were understood as a multidimensional region (consisting of selected pairs), in which a multidimensional boundary separating the case and control groups (obtained by PDF-adaptive algorithm) was fixed (as the union of all boundaries in attribute pairs). Mothers set (as *test* set) was passed through such a fixed construction and it was shown what their place is between the two boundary states. Since it was important in our task to associate the differences between control-case groups with age-related changes, we have investigated the connection between *test*-networks (namely, their topological characteristics) with age and showed that the transition between two critical conditions is highly associated with age. Namely, that transitions across Network borders are associated with a change in age in a healthy population and remain almost unchanged in patients with DS.

This design can be applied to a sample of patients of any disease (for example, cancer), in which, in addition to critical states (healthy - diagnosed patient), there are data on intermediate conditions (for example, analyzes of diagnosed patients at earlier time, when they could be considered as healthy people). We believe that the results of such an analysis based on Generalized Parenclitic Networks will not only help in the early diagnosis of the disease (by identifying critical transition marks) and risk assessment, but also shed light on the process itself, through attributes involved in it.

Methods

Parenclitic algorithm

In this section we describe a framework of the generalized parenclitic algorithm. The main goal of the algorithm is to identify deviations of a sample by pair of features. Final representation of those relations is a network where features corresponds to nodes and feature connections are links between nodes.

Input data

A structure of input data for Network construction is as follows: (X_i, Y_i), where X_i - feature vector of i-th subject, $Y_i \in \pm 1, \pm 2$ - label of i-th subject: number (1 or 2) defines set of subjects involved in construction a network (1 - involved, 2 - not involved) and sign defines set of controls and cases ("$-$": control group for construction KDE area, "$+$": cases - group for study of deviation from control).

Network construction

Network construction consists of 3 steps:

1. First of all, we consider subjects on the two-dimensional plane of couple of features (i, j). There is generally a control, case and testing subject groups, or only a control and case group. For certainty, consider an example where mothers is a control group and others is a testing group (Fig. 8, A).

2. At this step, a special classification method is used to determine the distance from subject to control group. This is kernel method which determines subject deviations from a normal state. Based on subject features it should predict if subject deviated from control group or not and how much it deviates (Fig. 8, B).

3. Finally, we reconstruct individual network for subject such that nodes are a features and links between 2 nodes designate a subject deviation by corresponding features. More precisely there is link if subject deviate from control group. (Fig. 8, C).

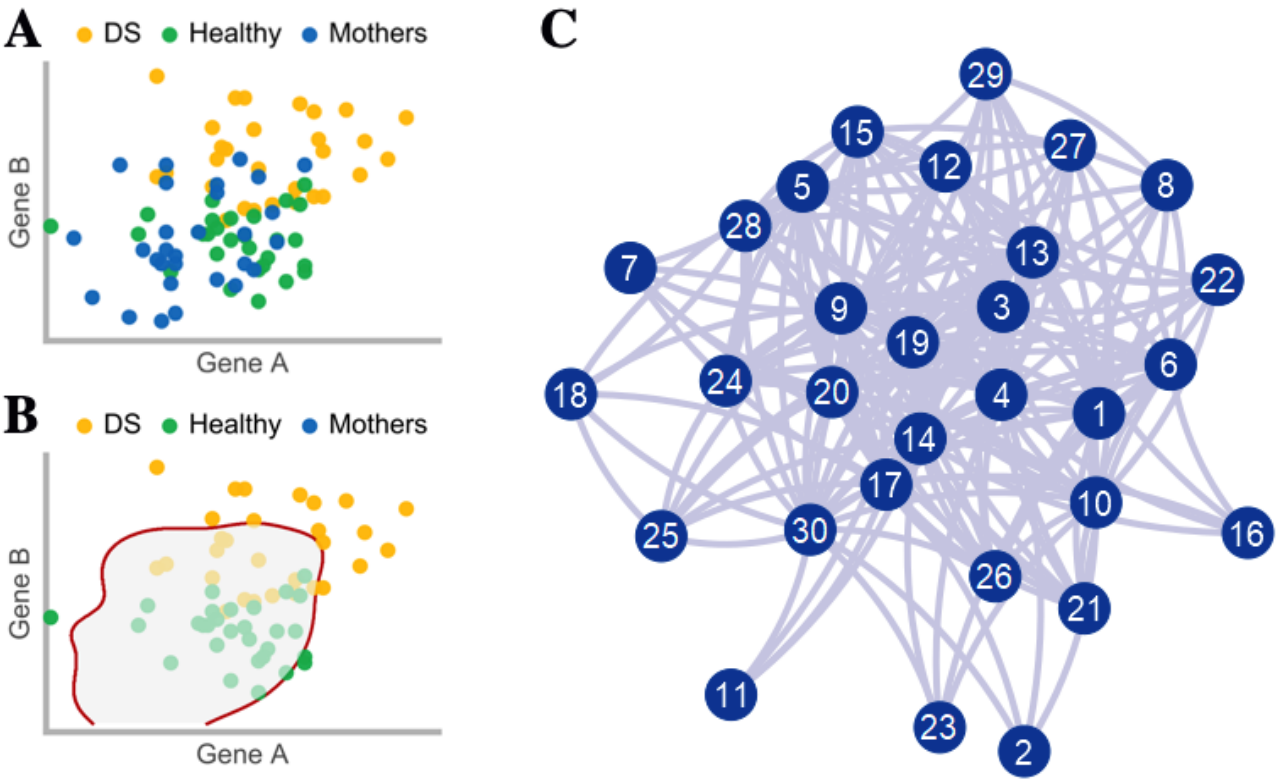

Figure 8. Illustration of Parenclitic Network construction steps. **A**: Example of two-dimensional plane of couple of features $(GeneA, GeneB)$; **B**: Selection of control group (as Mothers for example) and construction KDE; **C**: Reconstruction individual network for subject such that nodes are a features and links between 2 nodes designate a subject deviation by corresponding features

A classification method on the second step is a kernel of parenclitic algorithm. As such a method, we suggest 3 different approaches considered in the next 3 subsections. Samples with 2 selected features and labels represents an input for those algorithms. Algorithms perform classification subjects on 2 classes: deviated from control group and not deviated.

Kernels

- **Kernel: PDF reconstruction.** As one of these classifiers, the distance threshold obtained from a PDF estimation can be used. The idea is to reconstruct a probability density function (PDF) of control group subjects on two dimensional plane of two selected features. Let it be, for example, the KDE (kernel density estimation) method with Gaussian kernel. There are 2 groups: control group and test group. So, algorithm can be described by the following sequence of steps:

  1. Compute parameters of KDE for control group samples.
  2. Calculate distance from subject to PDF. Let us define such distance on the plane of two features:

$$d(x,y) = \int\limits_{\mathrm{PDF}(\eta,\nu)>\mathrm{PDF}(x,y)} \mathrm{PDF}(\eta,\nu)d\eta d\nu, \tag{2}$$

     where (x,y) - values of two selected features for subject. The meaning of this distance formula is a probability for subject to be more probable to control group PDF. So, the closer subject in value of $\mathrm{PDF}(x,y)$ to the maximum value $\mathrm{PDF}(x^*,y^*)$, the smaller the distance $d(x,y)$.
  3. Choose threshold of significance level. As $d(x,y)$ is a probability, it has value in range $[0,1]$. Therefore, one can choose a constant threshold value for the distance. Initially, there is no rule for threshold defining, so

it can be varied in the allowable range with some step. For certainty, let us define set of thresholds $\mathbf{thr} \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$.

4. Identify subject deviation from a control group by a comparison between the distance for this subject and the threshold value. If the subject is far from control group ($d(x,y) > \mathbf{thr}$), this deviation is significant.

- **Kernel: SVM classification.** Due to the long time of computation for the PDF kernel approach, we suggested faster distance classifier method. The methylation data analysis of Down Syndrome blood shows that almost all interesting feature pairs represent simple point cloud configurations. The Support Vector Machine classifier is well applicable to such data. In the general case, one can use different kernels inside the SVM and different parameters for the best results. To finalise all steps for the identification of deviations from the control group one should:

  1. Construct separating hyperplane by SVM algorithm on the samples with labels C_i from both groups: control and test.
  2. Check if SVM provides us with a good separation or not. Division on 2 sets can be assessed by setting a score, and if the score is too small no need to consider those couple of features. So, if $score < \mathbf{thr_s}$ then there is no way to determine correct deviations. Also this score cut off help to reduce number of links.
  3. Calculate signed distance from sample to hyperplane $d(x,y)$.
  4. Choose threshold value for distance (by default is 0). Here one can apply the same approach as for choosing a threshold in the PDF approach.
  5. Identify subject deviation from the control group by a comparing a distance for this subject and the threshold value. If the subject is far from control group ($d(x,y) > \mathbf{thr}$) deviation is significant (the same rule as in PDF approach)

- **Kernel: PDF-adaptive (the best threshold method).** The calculation of a distance in previous approaches helped us to reduce a two-dimensional problem to the one-dimensional one. Based on those distance function a decision rule is defined. The only essential complication in previous kernel algorithms is a choice of the threshold value. To overcome it we propose the following: let's choose a threshold value in an adaptive manner. Choosing a threshold defines the 2D plane division into two complex sets of points: deviated or not. Subjects can be in both sets, but it can be in expected set or not. But on the one-dimensional distances it just a two ranges: $[0; \mathbf{thr}]$ and $(\mathbf{thr}; 1]$. Here, however, the problem of the best set splits arises. To the best of our knowledge it has already been solved in decision trees algorithm, using a measure of an information gain for all possible splits and choosing the one that maximizes it. So, only the third step in the PDF reconstruction method is practically changed.

  There are finite number of threshold separators where subsets of samples are changed. For all of them, one can calculate information gain metric and take the threshold giving the maximum value.

Network analysis

Revealing internals of investigated complex system can be done by applying network analysis. Insight to the structure of networks give us a possibility to measure integral system state observed for different subjects. Also, centrality measures locates the most important or interesting nodes (features for initial system) in network.

In the previous section, the algorithms produced network for each subject, where nodes represent features, and links — deviation by couple of features from it's expected state. In this study the following metrics is considered:

- Metrics that associate a number with each node (feature):Betweenness, Pagerank, Closeness, Eigenvector centrality, Degrees;
- Metrics that associate a number with each links: Distance;
- Metrics that characterize graph as whole: Efficiency, Robustness, Max component size, Number of nodes, Number of links;
- Other metrics: Component sizes;
- Statistics for vector metrics: Minimal value, Maximal value, Mean, Standard deviation, Number of zeros;

Full metric description includes:

- Degree centrality is defined as the number of links incident upon a node. The degree can be interpreted in terms of the immediate risk of a node for catching whatever is flowing through the network (such as a virus, or some information)

- Betweenness centrality is a measure of centrality in a graph based on shortest paths. For every pair of vertices in a connected graph, there exists at least one shortest path between the vertices such that either the number of edges that the path passes through (for unweighted graphs) or the sum of the weights of the edges (for weighted graphs) is minimized. The betweenness centrality for each vertex is the number of these shortest paths that pass through the vertex. *The betweenness centrality type measures how often each graph node appears on a shortest path between two nodes in the graph. Since there can be several shortest paths between two graph nodes s and t, the centrality of node u is:

$$c_u = \sum_{s,t \neq u} \frac{n_{st}(u)}{N_{st}}$$

  $n_{st}(u)$ is the number of shortest paths from s to t that pass through node u, and N_{st} is the total number of shortest paths from s to t. If the graph is undirected, then the paths from s to t and from t to s count only as one path (divide the formula by two).

- PageRank algorithm outputs a probability distribution used to represent the likelihood that a person randomly clicking on links will arrive at any particular page. PageRank of undirected graph statistically close to its degree distribution. *The pagerank centrality type results from a random walk of the network. At each node in the graph, the next node is chosen uniformly from the set of successors of the current node (neighbors for the undirected case). The centrality score is the average time spent at each node during the random walk.

- Closeness centrality of a node is a measure of centrality in a network, calculated as the reciprocal of the sum of the length of the shortest paths between the node and all other nodes in the graph. Thus, the more central a node is, the closer it is to all other nodes. The closeness centrality types use the inverse sum of the distance from a node to all other nodes in the graph. If not all nodes are reachable, then the centrality of node i is:

$$c_i = \left(\frac{A_i}{n-1}\right)^2 \frac{1}{C_i},$$

  A_i is the number of reachable nodes from node i (not counting i), n is the number of nodes in G, and C_i is the sum of distances from node i to all reachable nodes. If no nodes are reachable from node i, then c_i is zero.

- Eigenvector centrality is a measure of the influence of a node in a network. Relative scores are assigned to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes. A high eigenvector score means that a node is connected to many nodes who themselves have high scores. The eigenvector centrality type uses the eigenvector corresponding to the largest eigenvalue of the graph adjacency matrix. The scores are normalized such that the sum of all centrality scores is 1.

- Distance to dividing line from SVM; or a PDF distance defined in PDF kernel description.

- Component is a subgraph in which any two vertices are connected to each other by paths, and which is connected to no additional vertices in the initial graph. There are several components. Size of component is a number of vertices in this subgraph.

- Efficiency of a network is a measure of how effective is the information exchange in the system. Average efficiency is defined as:

$$\frac{1}{n(n-1)} \sum_{s \neq t} \frac{1}{d_{st}}$$

  where n denotes number of nodes and d_{st} denotes the length of the shortest path between a node s and another node t.

- Robustness of the ability to withstand failures and perturbations, is a critical attribute of many complex systems including complex networks. Robustness calculated as number of steps in process of removing nodes with high degrees until links presented in graph.

Complexity analysis

Evaluation of network analysis complexity affecting the computation time provided in this paper can be done using the algorithmic complexity. The main pipeline includes analysis of pairs of features and performing some special kernel estimation for them. So, complexity of a framework algorithm is $O(n^2 \cdot T_{kernel}(m))$, where m denote number of samples and n — number of features. A computational time of the kernel depends on implementation details. Therefore, the following is a detailed description of PDF kernel calculation.

PDF kernel detailed implementation

Algorithm of this kernel supposes that input data consists of $d = 2$ features and m samples.

- First step is a calculation of two-dimensional KDE function. It is represented by a sum of Gaussian functions and can be evaluated as $O(d^3 + m \cdot d^2)$. Hence, in the case $d = 2$, complexity is $O(m)$.
- Next, distance from subjects to distribution is calculated. There are several techniques to compute an integral. Here we suggest Monte Carlo integration method. It does not have the highest accuracy, but it can help calculate all of the integrals at once in the fastest way.

  Let us consider Monte Carlo method to integrate PDF over whole plane. To start we should do sampling from reconstructed distribution and sum up them. Let us define random variables $(X_j, Y_j), j = \overline{1, L}$ equally distributed same as the reconstructed two dimensional PDF. Next, we generate pair of numbers (x_j, y_j) from those distributions.

$$\frac{1}{L}\sum_{j=1}^{L} \mathrm{PDF}(x_j, y_j) \xrightarrow[L\to\infty]{} \int\limits_{\mathbb{R}^2} \mathrm{PDF}(\eta, \nu) d\eta d\nu = 1 \tag{3}$$

  This phase can be done in $O(L \cdot d^2 \cdot m)$ time. But we are interested in computing of integral that select only those PDF values that are greater than some constant value $PDF(x, y)$. In this way we obtain:

$$\frac{1}{L}\sum_{j=1}^{L} \mathrm{PDF}(x_j, y_j) \cdot [\mathrm{PDF}(x_j, y_j) > \mathrm{PDF}(x, y)] \xrightarrow[L\to\infty]{} \int\limits_{\mathrm{PDF}(\eta,\nu) > \mathrm{PDF}(x,y)} \mathrm{PDF}(\eta, \nu) d\eta d\nu \tag{4}$$

  It costs $O(L \cdot d^2 \cdot m^2)$ for computing m distances. First optimization of this formula is a calculating PDF values only once. To achieve this we order (x_j, y_j) by value of $p_j = PDF(x_j, y_j)$ in descending order and compute cumulative sum $s_j = \sum_{k=1}^{j} p_j$. This takes $O(L + L \cdot \log L + L \cdot d^2 \cdot m)$ time. Values of m distances for subjects can be obtained by binary search over p_j from s_j. Overall this step takes $O(L + L \cdot \log L + L \cdot d^2 \cdot m + m \cdot \log m)$. By experiments we suggest $L = 10^4$ or $L = 10^5$ to achieve a balance between computational time and accuracy.
- After the distance values are known, a construction of links using threshold takes linear time in m.

It takes a lot of time to compute distance values. But if only the deviation presence matters it can be performed without heavy integral computations. To achieve this, notice that distance monotonically decreased with PDF value increased and for one constant $p = PDF(x, y)$ value there is only one distance value $d = d(x, y)$. This means that we can solve problem of threshold defining using d or p values in a manner. In the last section is defined adaptive threshold algorithm which gives the same result on both p and d sequences. So, for adaptive threshold technique: computation of p costs $O(m^2)$ and best split identifying based on information gain costs $O(m \cdot \log m + m)$. Full complexity estimation is $O(n^2 \cdot m^2)$.

The last kernel based on SVM method doesn't use any additional computations, consequently, a complexity of the kernel is the same as SVM on 2 features with m samples (also it depends on internal parameters and implementation). Full complexity estimation is $O(n^2 \cdot m^2)$.

References

1. Bullmore, E. & Sporns, O. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat. reviews neuroscience* **10**, 186–198 (2009).
2. Castellani, G. C. *et al.* Systems medicine of inflammaging. *Briefings bioinformatics* **17**, 527–540 (2016).

3. Albert, R. & Barabási, A.-L. Statistical mechanics of complex networks. *Rev. modern physics* **74**, 47 (2002).
4. Newman, M. E. Mixing patterns in networks. *Phys. Rev. E* **67**, 026126 (2003).
5. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M. & Hwang, D.-U. Complex networks: Structure and dynamics. *Phys. reports* **424**, 175–308 (2006).
6. Collins, F. S. & Barker, A. D. Mapping the cancer genome. *Sci. Am.* **296**, 50–57 (2007).
7. Zanin, M. *et al.* Parenclitic networks: uncovering new functions in biological data. *Sci. Reports* **4**, 5112, DOI: 10.1038/srep05112 (2014).
8. Zanin, M. & Boccaletti, S. Complex networks analysis of obstructive nephropathy data. *Chaos: An Interdiscip. J. Nonlinear Sci.* **21**, 033103 (2011).
9. Zanin, M., Menasalvas, E., Sousa, P. A. & Boccaletti, S. Preprocessing and analyzing genetic data with complex networks: An application to obstructive nephropathy. *Networks & Heterog. Media* **7**, 473 (2012).
10. Zanin, M., Menasalvas, E., Boccaletti, S. & Sousa, P. Feature selection in the reconstruction of complex network representations of spectral data. *PloS one* **8** (2013).
11. Zanin, M. *et al.* Knowledge discovery in spectral data by means of complex networks. *Metabolites* **3**, 155–167 (2013).
12. Zanin, M. *et al.* Combining complex networks and data mining: why and how. *Phys. Reports* **635**, 1–44 (2016).
13. Papo, D., Buldú, J. M., Boccaletti, S. & Bullmore, E. T. Introduction: Complex network theory and the brain. *Philos. Transactions: Biol. Sci.* **369**, 1–7 (2014).
14. Karsakov, A. *et al.* Parenclitic network analysis of methylation data for cancer identification. *PloS one* **12**, e0169661 (2017).
15. Whitwell, H. J., Blyuss, O., Menon, U., Timms, J. F. & Zaikin, A. Parenclitic networks for predicting ovarian cancer. *Oncotarget* **9**, 22717–22726, DOI: https://doi.org/10.18632/oncotarget.25216 (2018).
16. Jones, P. A. & Baylin, S. B. The epigenomics of cancer. *Cell* **128**, 683–692 (2007).
17. Riddihough, G. & Zahn, L. M. What is epigenetics? *Science* **330**, 611–611 (2010).
18. Widschwendter, M. *et al.* Epigenome-based cancer risk prediction: rationale, opportunities and challenges. *Nat. reviews Clin. oncology* **15**, 292 (2018).
19. Beck, S. & Olek, A. *The epigenome: Molecular hide and seek* (John Wiley & Sons, 2006).
20. Bacalini, M. G. *et al.* Identification of a dna methylation signature in blood cells from persons with down syndrome. *Aging* **7**, 82–96, DOI: 10.18632/aging.100715 (2015).
21. Aryee, M. J. *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363–1369, DOI: 10.1093/bioinformatics/btu049 (2014). https://academic.oup.com/bioinformatics/article-pdf/30/10/1363/17344721/btu049.pdf.
22. Zhou, W., Laird, P. W. & Shen, H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic Acids Res.* **45**, e22–e22, DOI: 10.1093/nar/gkw967 (2016). https://academic.oup.com/nar/article-pdf/45/4/e22/25364844/gkw967.pdf.
23. Breeze, C. E. *et al.* eFORGE v2.0: updated analysis of cell type-specific signal in epigenomic data. *Bioinformatics* **35**, 4767–4769, DOI: 10.1093/bioinformatics/btz456 (2019). https://academic.oup.com/bioinformatics/article-pdf/35/22/4767/30706685/btz456.pdf.
24. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47**, D330–D338, DOI: 10.1093/nar/gky1055 (2018). https://academic.oup.com/nar/article-pdf/47/D1/D330/27437640/gky1055.pdf.
25. Anastas, J. N. & Moon, R. T. Wnt signalling pathways as therapeutic targets in cancer. *Nat. Rev. Cancer* **13**, 11–26, DOI: 10.1038/nrc3419 (2013).
26. Zhan, T., Rindtorff, N. & Boutros, M. Wnt signaling in cancer. *Oncogene* **36**, 1461–1473, DOI: 10.1038/onc.2016.304 (2017).
27. Inestrosa, N. C. & Arenas, E. Emerging roles of wnts in the adult nervous system. *Nat. Rev. Neurosci.* **11**, 77–86, DOI: 10.1038/nrn2755 (2010).
28. Jia, L., Piña-Crespo, J. & Li, Y. Restoring wnt/b-catenin signaling is a promising therapeutic strategy for alzheimer's disease. *Mol. Brain* **12**, 104, DOI: 10.1186/s13041-019-0525-5 (2019).

29. Sussan, T. E., Yang, A., Li, F., Ostrowski, M. C. & Reeves, R. H. Trisomy represses apcmin -mediated tumours in mouse models of down's syndrome. *Nature* **451**, 73–75, DOI: 10.1038/nature06446 (2008).

30. Nayar, A. Why people with down's syndrome get fewer cancers. *Nature* DOI: 10.1038/news.2009.493 (2009).

31. Granno, S. *et al.* Downregulated wnt/b-catenin signalling in the down syndrome hippocampus. *Sci. Reports* **9**, 7322, DOI: 10.1038/s41598-019-43820-4 (2019).

Acknowledgements (not compulsory)

Authors acknowledge support by the grant of the Ministry of Education and Science of the Russian Federation Agreement No. 074-02- 2018-330. AZ, TN acknowledge support by the MRC grant MR/R02524X/1.

Author contributions statement

AZ, CF, MI designed the study, MK developed software, TN developed methodology of age-related analysis, MK, TN, MGB conducted the graph and data analysis. All authors have written and reviewed the manuscript.

Additional information

Authors declare no competing interests.