Bioinformatics, 36, 2020, i455–i463
doi: 10.1093/bioinformatics/btaa445
ISMB 2020

Chromatin network markers of leukemia

N. Malod-Dognin[1,2], V. Pancaldi[1,3,4], A. Valencia[1,5,6] and N. Pržulj[1,2,5,*]

[1]Department of Life Sciences, Barcelona Supercomputing Center (BSC), Barcelona 08034, Spain, [2]Department of Computer Science, University College London, London WC1E 6BT, UK, [3]Centre de Recherches en Cancérologie de Toulouse (CRCT), Toulouse 31037, France, [4]University Paul Sabatier III, Toulouse 31330, France, [5]ICREA, Pg. Lluís Companys 23, Barcelona 08010, Spain and [6]Coordination Node, Spanish National Bioinformatics Institute, ELIXIR-Spain (INB, ELIXIR-ES), Madrid 28029, Spain

*To whom correspondence should be addressed.

Abstract

Motivation: The structure of chromatin impacts gene expression. Its alteration has been shown to coincide with the occurrence of cancer. A key challenge is in understanding the role of chromatin structure (CS) in cellular processes and its implications in diseases.

Results: We propose a comparative pipeline to analyze CSs and apply it to study chronic lymphocytic leukemia (CLL). We model the chromatin of the affected and control cells as networks and analyze the network topology by state-of-the-art methods. Our results show that CSs are a rich source of new biological and functional information about DNA elements and cells that can complement protein–protein and co-expression data. Importantly, we show the existence of structural markers of cancer-related DNA elements in the chromatin. Surprisingly, CLL driver genes are characterized by specific local wiring patterns not only in the CS network of CLL cells, but also of healthy cells. This allows us to successfully predict new CLL-related DNA elements. Importantly, this shows that we can identify cancer-related DNA elements in other cancer types by investigating the CS network of the healthy cell of origin, a key new insight paving the road to new therapeutic strategies. This gives us an opportunity to exploit chromosome conformation data in healthy cells to predict new drivers.

Availability and implementation: Our predicted CLL genes and RNAs are provided as a free resource to the community at https://life.bsc.es/iconbi/chromatin/index.html.

Contact: natasha@bsc.es

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

1.1 Chronic lymphocytic leukemia

Chronic lymphocytic leukemia (CLL) is the most common leukemia in adults (National Cancer Institute, 2019). The bone marrow produces blood stem cells (immature cells) that mature over time. To become white blood cells, blood stem cells first become lymphoid stem cells, which in turn become either B lymphocytes (antibodies that fight infections), T lymphocytes (that help B lymphocytes to fight infections) or natural killer cells (that attack cancer cells and viruses) (National Cancer Institute, 2019). However, in CLL, abnormal lymphocytes that are called leukemia cells build-up in the bone marrow, lymph nodes and blood, and crowd out healthy blood cells (Kipps *et al.*, 2017; National Cancer Institute, 2019). These changes often occur during or after middle age and gradually get worse. When leukemia cells take over the healthy blood cells, the patients become subjects to infection, anemia and easy bleeding (National Cancer Institute, 2019).

Whole exome sequencing has revealed a high degree of genetic variability in somatically muted genes across CLL patients (Fabbri *et al.*, 2011; Puente *et al.*, 2011; Wang *et al.*, 2011). On the other hand, about 82% of patients with CLL carry at least one of five common chromosomal alterations (Döhner *et al.*, 2000): a deletion in 13q (55%), a deletion in 11q (18%), a trisomy in 12q (16%), a deletion in 17p (7%) and a deletion in 6q (6%). Furthermore, epigenetic changes have also been reported in CLL. Notably, methylation signatures can classify distinct clinical CLL subgroups (Bhoi *et al.*, 2016; Kulis *et al.*, 2012), and there is an association between methylation evolution and adverse clinical outcomes (Landau *et al.*, 2014; Oakes *et al.*, 2014).

1.2 Chromatin structure

In the nucleus of an eukaryotic cell, DNA molecules are packed with other molecules to form the chromatin. The development and application of high throughput chromosome conformation capture technologies, such as Hi-C (Lieberman-Aiden *et al.*, 2009) and Capture Hi-C (Mifsud *et al.*, 2015), allowed for comprehensive captures of 3D structures of chromatin (also called 3D genomes), giving new insight into the chromatin folding principles and its organization (Bonev and Cavalli, 2016). The organization of the chromatin structure (CS) has been shown to be hierarchical (Bonev and Cavalli, 2016). First, DNA is packed around histones to form nucleosomes (Bonev and Cavalli, 2016). Then, regions that are far away in the DNA can come into contact in 3D space to form chromatin loops (Bonev and Cavalli, 2016). Several loops can assemble to form Topologically Associating Domains (TADs) (Bonev and Cavalli, 2016). These TADs are further organized into promoting and

repressing regions (A/B compartments) and finally into chromosome territories (Bonev and Cavalli, 2016). These different levels of CS are functionally important. The CS is known to be involved in the regulation of gene expression, despite controversies about the underlying mechanisms. For instance, chromatin loops bring into close spatial proximity an enhancer with its target promoter [e.g. the locus control region of the β-globin cluster that interacts with its target genes via long range chromatin contacts (Palstra *et al.*, 2003)] and TADs are thought to be the regulatory units of the genome (Dixon *et al.*, 2016). Furthermore, the CS has a role in diseases. Alterations in the CS have been shown to have pathogenic effects, e.g. altering limb development (Lupiáñez *et al.*, 2015), and coinciding with the occurrence of cancer (Ferraro, 2016). Thus, uncovering the relationships between the local structure of the chromatin around genes and their biological functions is a key to understanding fundamental regulatory processes and the role of chromatin in diseases.

CSs have been modeled as networks in which nodes represent DNA fragments and in which two nodes are connected by an edge if the corresponding DNA fragments are in contact in the CS. The wiring patterns (topology) of these networks have been studied with various topological descriptors including graphlets (Pržulj *et al.*, 2004)(detailed below), which revealed connectivity differences between regulatory regions (Thibodeau *et al.*, 2017).

1.3 Graphlet-based analysis of network data

System-level molecular datasets are frequently modeled and analyzed as networks. For example the protein–protein interactions (PPIs) of a cell are modeled as the PPI network in which nodes represent proteins and two nodes are connected by an edge if the corresponding proteins can physically bind. It has long been known that the exact comparison between large networks is computationally intractable (Cook, 1971). Thus, the topological analyses of networks use approximate (heuristic) comparisons. These heuristics include network properties, such as node degrees and betweenness centralities, to approximately say whether the structures of networks are similar (Newman, 2010). Among the most sensitive network properties are graphlets, small, connected, induced subgraphs of large networks (Pržulj *et al.*, 2004).

Graphlets have been used to characterize and compare the local wiring patterns around nodes in a PPI network (Milenković and Pržulj, 2008), that revealed that molecules involved in similar functions tend to be similarly wired (Davis *et al.*, 2015). These topological similarities between nodes have also been used to define the graphlet correlation distance (GCD), which is the most sensitive measure of topological similarity between networks (Yaveroğlu *et al.*, 2014, 2015a). Graphlets have also been used to compare protein 3D structures represented by networks (Faisal *et al.*, 2017; Malod-Dognin and Pržulj, 2014), which allowed for transferring annotations. Finally, graphlets have been successfully used to guide the node mapping process of network alignment methods (Pržulj, 2019), which allowed for transferring annotations from well-studied proteins to less-studied ones.

1.4 Contributions

In this study, we propose a framework to perform topological, graphlet-based comparative analysis of tissue-specific CSs modeled as networks. We use this framework to study the CS network of 17 healthy blood cell types and of one CLL cell type, and we show that the global wiring patterns of CS networks capture the modular organization of the chromatin and relate to the functioning of the corresponding cells. We show that the local wiring patterns around the DNA elements in CS networks relate to their biological functions. The large functional enrichments that we obtain confirm that the topology of CS networks, captured by graphlets, is a new source of biological information that complement the information that can be obtained from other types of molecular data, such as PPI and gene co-expression networks. The comparison between the CS networks of control naive B (nB) cells and CLL cells suggests that leukemia cells have large structural changes in the chromatin network, reduced modularity and reduced functional coherence. Importantly, we uncover that CLL driver genes are characterized by specific local wiring patterns in the CS networks of not only CLL cells, but also of healthy control cells, a key insight with profound implication, as it enables finding new cancer-related genes from the CS data of healthy cells. We use these specific local wiring patterns to successfully predict new leukemia-related genes and non-coding RNAs.

2 Materials and methods

2.1 Datasets

Data networks. We collected the CSs of 17 different human blood cell types from Javierre *et al.* (2016), which include control nB cells, and the CS of CLL cells from R. Beekman (personal communication). All chromatin contacts were captured using Capture HiC and processed in the same laboratory and using the same experimental protocol (Beekman *et al.*, 2018; Javierre *et al.*, 2016). We model CS dataset of each blood cell type as a CS network, in which nodes represent DNA elements (genes and non-coding RNAs) and in which edges connect nodes whose DNA elements are in contact in the CS. Although inter-individual variability and potential batch effects may be present in general CS data, the data used in this study are the best CS datasets on blood cell types to date. Also, we confirmed that the CS data of the CLL cells used in this study contain no genetic rearrangements (Beekman, 2020). From IID (v.04-2018) (Kotlyar *et al.*, 2016), we collected the experimentally validated PPIs of human (generic, not tissue-specific). We model them as a PPI network in which nodes represent genes and in which edges connect genes whose proteins can bind. From COXPRESdb (v.6.0) (Okamura *et al.*, 2015), we collected the co-expressions of the human genes (generic, not tissue-specific). We model them as a COEX network in which nodes represent genes and in which edges connect genes that are co-expressed. To only consider the most co-expressed genes, we select the top 1% mutual rank correlations and keep them as edges in the COEX network. The statistics of the networks are presented in Supplementary Table S1.

Differential gene expressions. From Feirreira *et al.* (2014), we collected the gene expression profiles of 122 CLL samples and of 20 controls samples of healthy B cells. The gene expression measurements were obtained using Affymetrix Human Genome U219 Array Plates, and raw CEL files were preprocessed and normalized using Robust Multi-array Average. We computed the corresponding differential expressions using Limma (Ritchie *et al.*, 2015). A gene is significantly differentially expressed if the corresponding differential expression *P*-value after Benjamini–Hochberg correction for multiple hypothesis testing is lower than or equal to 0.05.

Biological annotations of genes. From the Reactome database (Fabregat *et al.*, 2018), we collected the reactome reaction (RR) and reactome pathway (RP) annotations of the human genes. We also collected from Gene Ontology (Ashburner *et al.*, 2000) the Biological Process (GO-BP) and Molecular Function (GO-MF) annotations of the genes. All annotations were collected in June 2018.

Lists of CLL driver genes and of mutated genes. From intOgene (v.2014.12) (Gonzalez-Perez *et al.*, 2013), we collected the list of 38 driver genes for CLL, out of which 31 are in the CS network of nB cells, 32 are in the CS network of CLL cells and 30 are found in both networks. Also, we collected from intOgen the list of all genes that are known to be mutated in CLL.

Model networks. To investigate the organizational principles of CS networks, we compare their wiring patterns to the ones of synthetic networks generated according to nine commonly used random graph models of the size and edge density of the CS networks: the Erdös–Rènyi random graph model (ER) (Erdös and Rényi, 1959), the Generalized random graph model (ER-DD, also called the configuration model) (Newman, 2010), the Geometric random graph model (GEO) (Penrose, 2003), the Geometric random graph with gene duplication model (Pržulj *et al.*, 2010), the Barabàsi–Albert Scale-free model (SF, also called the preferential attachment model) (Barabási and Albert, 1999), the Scale-free with gene duplication and divergence model (Vázquez *et al.*, 2003), the Watts–Strogatz

small-world model (Watts and Strogatz, 1998), the Stickiness-index-based model (STICKY) (Pržulj and Higham, 2006) and the non-uniform Popularity Similarity Optimization model (nPSO) (Muscoloni and Cannistraci, 2018). For each data network and for each random model, we generated 20 random networks having the same characteristics as the input data network (the same number of nodes and edges, the same degree distribution in the case of ER-DD and the same number of communities in the case of nPSO).

2.2 Methods

Capturing the wiring patterns of biological networks. Because graphlets are the most sensitive measure of network topology to date (Yaveroğlu *et al.*, 2015a,b), we use them to capture the local wiring patterns around nodes in networks. *Graphlets* are small, connected, non-isomorphic, induced subgraphs of a large network that appear at any frequency (Pržulj *et al.*, 2004). Within a graphlet, symmetry groups of nodes, called automorphism *orbits*, are used to characterize different topological positions of a node. These orbits are used to generalize the notion of a node degree: the *graphlet degrees* of a node are the numbers of times a node touches each graphlet orbit in the network (Pržulj, 2007). Following the methodology of Yaveroğlu *et al.* (2015a), we use the 11 non-redundant orbits of 2- to 4-node graphlets, which have been shown to perform better than higher order graphlets. Thus, each node in a network is characterized by an 11-dimensional vector, called the *Graphlet Degree Vector* (*GDV*), or *GDV signature*, which captures the 11 non-redundant 2- to 4-node graphlet degrees of the node.

Within a network, we quantify the similarity between the wiring patterns of two nodes by using the *Graphlet Degree Vector Distance* (*GDVD*) (Milenković and Pržulj, 2008) between their GDVs, which we compute as follows. Given two GDV vectors, u and v, the distance between their ith coordinates is defined as: $D_i(u,v) = w_i \times \frac{|\log(u_i+1)-\log(v_i+1)|}{\log(\max u_i,v_i+2)}$, where w_i is the weight of orbit i that accounts for dependencies between orbits (Milenković and Pržulj, 2008). Then, GDVD is defined as: $\mathrm{GDVD}(u,v) = \frac{\sum_{i=1}^{11} D_i(u,v)}{\sum_{i=1}^{11} w_i}$. GDVD is a distance in [0,1), such that a distance equal to 0 means that the two GDVs are identical.

We characterize the global wiring patterns of a network with its *Graphlet Correlation Matrix* (*GCM*) (Yaveroğlu *et al.*, 2015a), which is an 11×11 symmetric matrix encoding the Spearman's correlations between non-redundant orbit counts over all nodes of the network Yaveroğlu *et al.* (2015a) We measure the distance between two networks with their *Graphlet Correlation Distance* (*GCD-11*), which is the Euclidean distance of the upper triangle values of their GCMs (Yaveroğlu *et al.*, 2015a).

Additionally, we also used the following topological descriptors (detailed in Supplementary Materials): node degree, clustering coefficient, shortest path length, betweenness centrality, eccentricity, diameter, radius, pagerank and modularity.

Investigating the organizational principles of CS networks. To measure the similarity between a real network (e.g. the CS network of naive B cells) and a given random model (e.g. ER model), we generate for each real network 20 random networks from the given model that have sizes and edge densities equal to the real network. We assess the quality of the fit between the data and the network model by the overlap between two distributions: the distribution of GCD-11s between the data and the model networks and the distribution of GCD-11s between the model networks. A data network is not fitted by a network model if the P-value of the Wilcoxon–Mann–Whitney U-test (MWU) between the two distributions of distances (real-to-model and model-to-model) is lower than or equal to 5% (threshold for which the two distributions are statistically significantly different).

Also, we use GCD-11 to assess if topological similarities between CS networks relate to the phenotypes of the corresponding cells (see Section 3.1).

Enrichment analysis. In Section 3.2, we assess if chromatin contacts tend to connect DNA elements whose proteins interact (PPI), that are co-expressed (COEX), or that share at least one common biological annotation term (RR, RP, GO-BP or GO-MF, defined below) using the following enrichment analysis. For a given CS network and a given annotation dataset (e.g. pairs of DNA elements sharing at least one GO-BP annotation), our analysis covers the set of n DNA elements that are both in the CS network and in the annotation dataset (i.e. we exclude the elements that have no chromatin contacts and the elements that do not share biological annotation with any other elements). These DNA elements define the background of $M = n(n-1)/2$ pairs of DNA elements, out of which K are interacting in the annotation dataset. We focus on N pairs of DNA elements that are interacting in the CS network, out of which X are also interacting in the annotation dataset. The fold enrichment of chromatin contacts in terms of annotation data is: $fold = \frac{X/N}{K/M}$, with a fold enrichment greater than one indicating that the chromatin contacts are enriched in annotation data. The probability of observing a fold enrichment greater than or equal to *fold* by chance, using a permutation test, is: $p = \frac{r+1}{n+1}$, where r is the number of permutations that have a fold enrichment greater than or equal to *fold*, and $n = 100$ is the number of permutations that we used. We consider a fold enrichment to be statistically significant if the corresponding P-value is lower than or equal to 5%.

Note that in this specific enrichment analysis, we globally assess if the genes that are in contact in the chromatin tend to share biological annotations (e.g. GO-BP). Instead of considering each annotation term separately, we consider that a pair of genes is similarly annotated if the two genes share some annotation terms. Then, we compute if the pairs of genes that are in contact in the chromatin are enriched in pairs of genes that are similarly annotated. Since this statistical test by its nature can only be applied once, there is no need to correct for multiple hypothesis testing.

Clustering and enrichment analysis. In Section 3.2, we assess if the wirings of DNA elements in CS networks relate to their biological functions using the following clustering and enrichment analysis. We cluster DNA elements in a CS network using two different k-means clusterings. We apply k-means on the k first eigenvectors of the Laplacian matrix of the CS network (so-called, spectral clustering) to cluster DNA elements that are densely connected to each other in the CS network. Also, we apply k-means directly on the GDVs of nodes in the CS network to cluster together DNA elements that have similar wiring patterns in the CS network, captured by graphlets. In both cases, the number of clusters, k, is chosen according to the rule of thumb (Kodinariya and Makwana, 2013): $k = \sqrt{\binom{n}{2}}$, where n is the number of nodes in the network. To account for the randomness of k-means, each clustering is repeated 10 times. Then, we measure the percentages of the produced clusters that are enriched in GO-BP, GO-MF, RR or RP annotations. The probability that an annotation is enriched is computed using sampling without replacement test (also called the hyper-geometric test): $p = 1 - \sum_{i=0}^{X-1} \binom{K}{i}\binom{M-K}{N-i} / \binom{M}{N}$, where N is the size of the cluster (only annotated DNA elements from the cluster are taken into account), X is the number of DNA elements in the cluster that are annotated with the annotation in question, M is the number of annotated DNA elements in the network and K is the number of DNA elements in the network that are annotated with the annotation in question. A cluster is significantly enriched if the enrichment P-value, after Benjamini–Hochberg correction for multiple hypothesis testing, is lower than or equal to 5%.

3 Results and discussion

3.1 CS network organization relates to the functioning of the corresponding cells

Because modularity is an important feature of CSs (Bonev and Cavalli, 2016), first we illustrate the CS networks of naive B (nB) cells and of CLL cells in two-dimensional space using spring embedding (Kamada and Kawai, 1989), which is known to reveal the modular organization of a network by grouping together nodes that are densely connected. We present a quantitative verification of this in the next paragraph. As illustrated in Figure 1A, the CS network

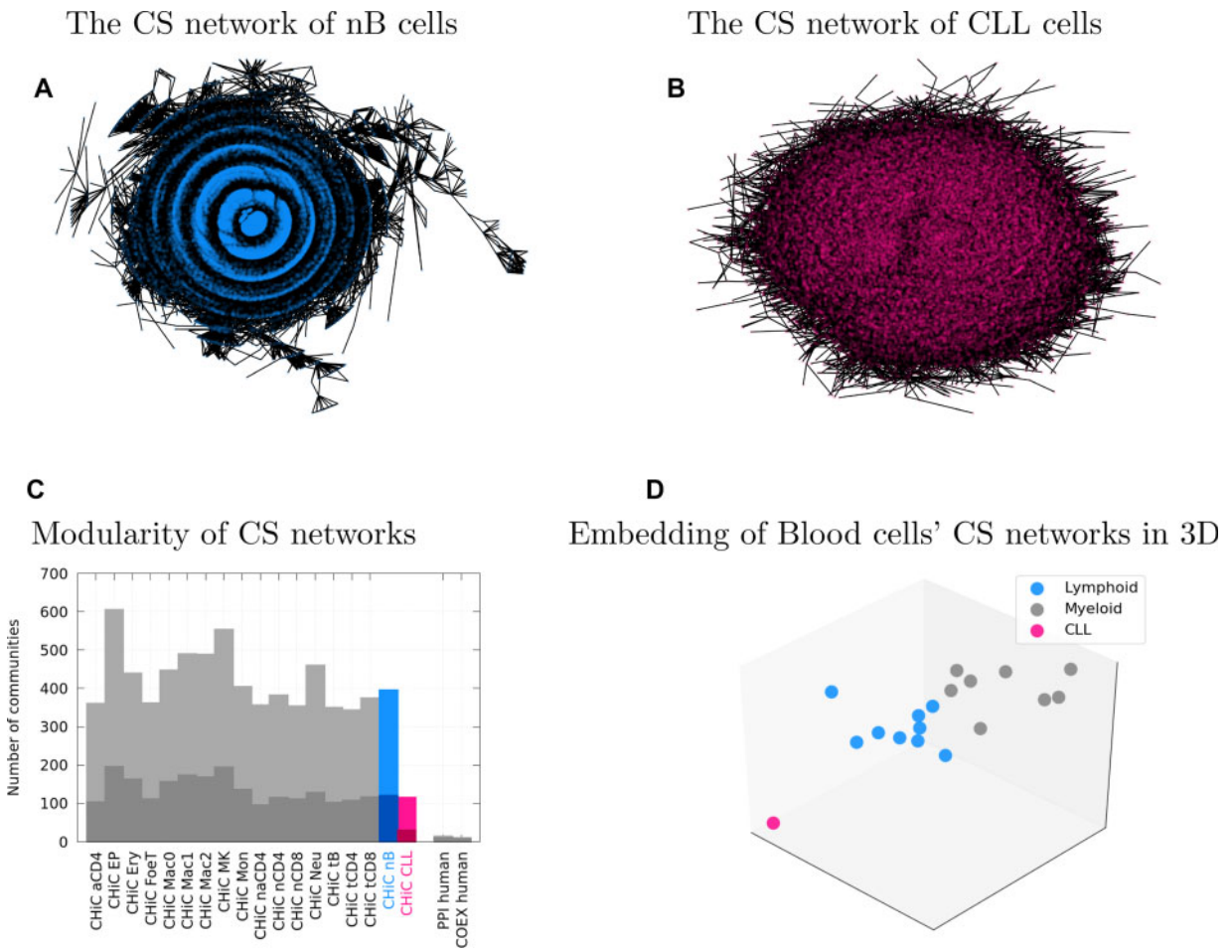

Fig. 1. Topological features of CS networks. (**A**) The two-dimensional spring embedding of the largest connected component of the CS network of nB cells, illustrating its organization. (**B**) The two-dimensional spring embedding of the largest connected component of the CS network of CLL cells. (**C**) The numbers of communities that are found in the networks. The darker bars present the number of communities that are found in the largest connected components of the networks. (**D**) The CS networks of blood cells (points) are embedded into 3D space according to their pairwise GCDs using MDS. The CS networks are colored as follows: blue for lymphoid cells (among which are nB cells), grey for myeloid cells and pink for the CLL cells

of nB cells has a modular organization, having well-defined clusters of nodes (illustrated by blue contiguous sets of points in Fig. 1A). On the other hand, as illustrated in Figure 1B, the CS network of CLL cells does not have a modular organization, since the red nodes are dispersed throughout the network and do not form any clusters.

We formerly assess this difference in modularity by computing the number of communities in the two networks [using the Louvain algorithm (Blondel *et al.*, 2008), which is a state of the art community detection algorithm]. As presented in Figure 1C, the CS networks of healthy blood cells have about 422 communities on average (or 136 when focusing on their largest connected components). This number is much smaller than the expected number of TADs in human cells (≈1000), which suggests that Capture HiC technology captures higher-order organization of the CS that may correspond to meta-TADs (grouping of TADs with coherent expression changes) (Fraser *et al.*, 2015), or compartments. In contrast, the CS network has reduced modularity in CLL, going from 395 communities in nB cells to only 115 communities in CLL cells (or from 121 to 30 when considering only the largest connected components of the CS networks). Furthermore, the distributions of sizes of the communities are also significantly different between the two networks, with Mann–Whitney *U*-test *P*-value $= 3.49 \times 10^{-4}$ (see Supplementary Fig. S14). The modularity in a molecular network is thought to be related with adaptability of the cells (Csermely *et al.*, 2015), the underlying idea being that stronger modularity allows for more specialized functioning and optimization of multiple objective functions. Thus, we hypothesize that this structural change of the chromatin in CLL, in which communities that are well separated in CS networks of healthy blood cells are merged together in CLL cells, allows the CLL cells to become less specialized, acquiring larger adaptability through plasticity and possibly heterogeneity, becoming closer to stem cells.

To further investigate the topological differences between the CS network of CLL compared to the CS network of nB cells, we perform a comparative analysis of the CS networks using standard network statistics (Supplementary Fig. S13). It reveals that the CS network of CLL cells has statistically significantly smaller mean node degrees, significantly smaller mean clustering coefficients, significantly larger mean betweenness centralities of its nodes, and significantly smaller mean Pageranks of its nodes than the CS network of nB cells (Supplementary Fig. S13 panels A, B, C and D, respectively). Then, we then focus on the differences between driver genes and background elements. Our results suggest that in CLL cells, driver genes become chromatin hubs by connecting chromatin modules that were poorly interacting in the control cells (measured by increases in the betweenness centrality of driver genes and decreases in their clustering coefficients in CLL CS network compared to the nB CS network, see Supplementary Fig. S10).

We test if the organization of CS networks relate to the phenotypes of the corresponding cells. To this aim, we use GCD-11 (Yaveroğlu *et al.*, 2014), one of the most sensitive measures of topological similarity between networks, to measure the similarity between the CS networks of control and CLL cell types. We illustrate the clustering it produces by embedding the CS networks as points in

three dimensions using multi-dimensional scaling (MDS), so that the pairwise distances between the points best approximate the GCD-11 distances between the networks. In the hematopoietic tree, our 17 blood cell types are classified into two main classes: the lymphoid cells (aCD4, FoeT, naCD4, nB, nCD4, nCD8, tB, tCD4, tCD8) and the myeloid cells (EP, Ery, Mac0, Mac1, Mac2, MK, Mon, Neu). As presented in Figure 1D, GCD-11 groups CS networks according to the hematopoietic origin of the corresponding cells (lymphoid or myeloid), while the CS network of CLL cells appears as an outlier. This demonstrates that the topological organization of the CS networks captured by graphlets relates to the cell's phenotype and ontology, consistent with findings in Javierre *et al.* (2016).

Finally, we ask whether the structure of our CS networks are close to the structure of random networks, by assessing how well nine random graph models that are commonly used to model biological networks recapitulate the features of our CS networks (see Supplementary Material) (Barabási and Albert, 1999; Erdös and Rényi, 1959; Muscoloni and Cannistraci, 2018; Newman, 2010; Penrose, 2003; Pržulj and Higham, 2006; Pržulj *et al.*, 2010; Vázquez *et al.*, 2003; Watts and Strogatz, 1998). We show that CS networks are not random, as they are not fitted well by Erdös–Rényi random graphs. Interestingly, the CS network of CLL cells is not more random than the one of control nB cells. To the opposite, the distances between the CS network of CLL cells and the ER networks are larger than those between the CS network of nB cells and the ER networks, as presented in Supplementary Figure S1A. That is CLL chromatin seems to be getting less modular, but more structured. The best fitting model to each CS data network is the non-uniform popularity similarity optimization model (Muscoloni and Cannistraci, 2018), which models geometrically and modularly organized networks (Supplementary Fig. S3).

3.2 CS network as a new source of functional information

In this section, we quantify the relationships between the wiring patterns (topology) of the DNA elements in the CS networks and their biological functions. First, we test if DNA elements that are in contact in the CS (i.e. that are connected by an edge in a CS network) tend to share biological functions. To do that, we assess if they tend to form PPIs, to be co-expressed or to share biological annotations (see Section 2: 'Chromatin interaction enrichment analysis'). We find that the DNA elements that are in contact in the CS are most likely to participate in common biochemical reactions (share at least one common RR annotation), to be co-expressed and to participate in common biological pathways (share at least one RP annotation) (see Fig. 2A) compared to random pairs of DNA elements, with *P*-values $\leq$ 0.01. To a lesser extent, DNA elements that are in contact in the CS are also likely to have protein products forming PPIs, to have similar molecular functions (to share at least one common GO-MF annotation) and to participate in common biological processes (to share at least one GO-BP annotation) (see Fig. 2B) compared to random pairs of DNA elements, with *P*-values $\leq$0.01. Interestingly, the DNA elements that are in contact in the CS of CLL cells are less functionally enriched in any of the tested interaction types and annotations. These results not only corroborate the current knowledge

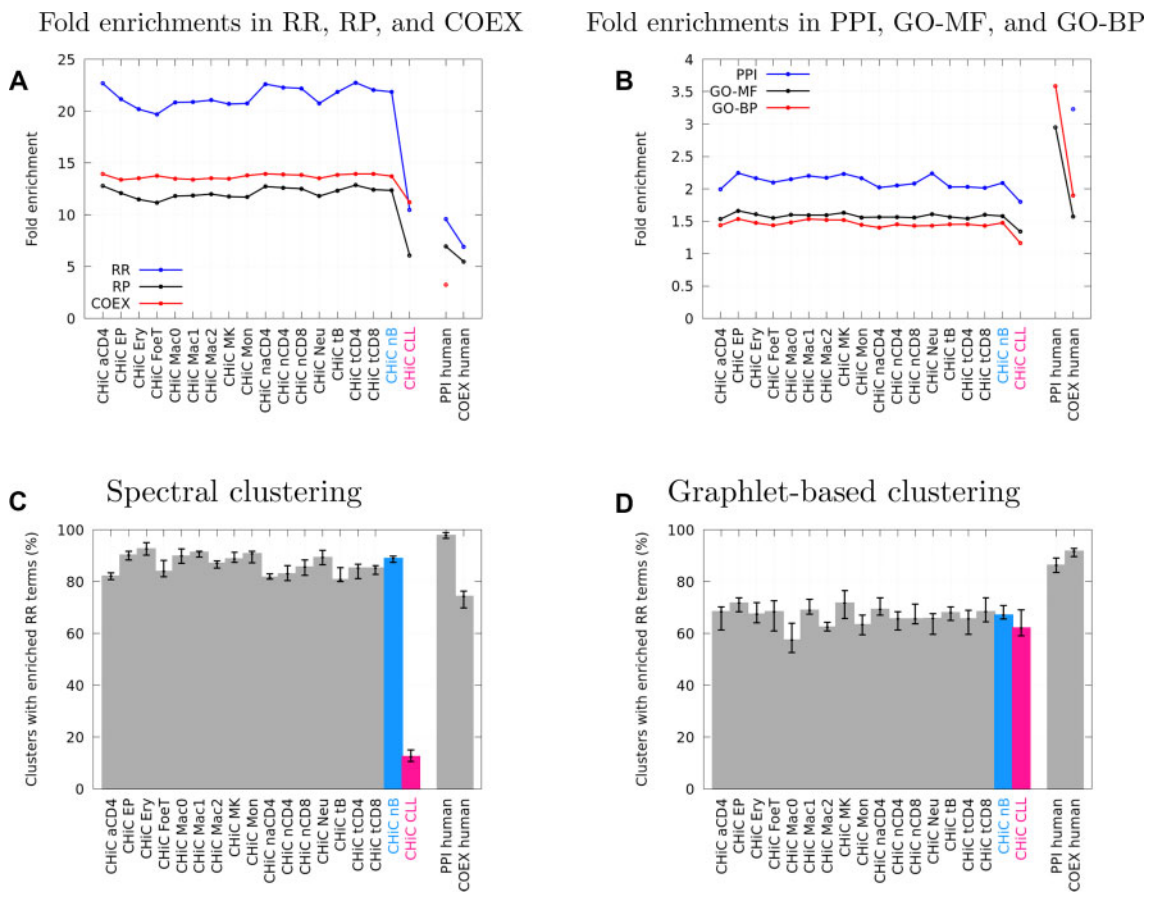

Fig. 2. Functional analysis of CS networks. (**A**) For each network (*x*-axis), we consider the genes connected by edges and report how many times they are more likely (fold enrichment) to share a RR (in blue) annotation, to share a RP annotation (in black), or to be co-expressed (COEX, in red) than expected by random. (**B**) The same as panel A, but for the fold enrichment in genes that form a PPI (in blue), that share a GO-MF annotation (in black), or that share a GO-BP annotation (in red). All the fold enrichments in panels A and B are statistically significant, with enrichment *P*-values $\leq$ 5%. (**C**) For each network (*x*-axis), we apply spectral clustering to group together genes that are densely connected to each other. For each clustering, the bar charts report the median percentage of the clusters that are statistically significantly enriched in at least one RR annotation term (median values over 10 runs of spectral clustering), and the error-bars present the corresponding 15.9th and 84.1th percentiles (*y*-axis). (**D**) The same as panel C, but when applying graphlet-based clustering to group together genes that have similar wiring patterns

on the role of the chromatin proximity in the regulation of expression (Dixon *et al.* 2016), confirming that chromatin spatially groups DNA elements forming biological functional units so that they could be co-expressed, but also suggests that the CS of CLL cells is less functionally coherent.

In addition, we investigate if the wirings of DNA elements in CS networks relate to their biological functions using the following clustering and enrichment analysis (detailed in Section 2: 'Clustering and enrichment analysis'). We cluster DNA elements in a network using two different k-means clustering methods. To cluster DNA elements that are densely connected to each other in a network, we apply k-means on the k first Eigen-vectors from the network's Laplacian matrix (i.e. we perform the spectral clustering). To cluster DNA elements that have similar wiring patterns in the network, independent of them being in the same network region, we apply k-means directly on the GDVs of nodes (DNA elements) in the network (GDVs are feature vectors that describe the local wiring patterns around nodes according to their participation in graphlets, detailed in Section 2: 'Capturing the wiring patterns of biological networks'). For each clustering and each annotation type, we report the number of clusters that are significantly enriched in one or more annotations, with enrichment P-values after Benjamini–Hochberg correction for multiple hypothesis testing ≤ 0.05.

As presented in Figure 2C and D, and in Supplementary Figures S4 and S5, we find that the resulting clusters in CS networks are mostly enriched in RR and RP annotations, which is consistent with the above observation that these two types of biological annotations are the most enriched on chromatin contacts. Also, the clusters obtained on the CS network of CLL cells are less functionally enriched than those obtained from the CS network of control nB cells, suggesting again that the CS network of CLL cells is less functionally coherent. This reduced functional coherence is more visible when using spectral clustering, which is more sensitive to network rewiring because it groups together nodes that are densely connected as its sole criterion. On the other hand, graphlet-based approaches are known to be more robust to network rewiring (Yaveroğlu *et al.*, 2015a), which may explain the smaller differences in functional enrichments between the CS networks of CLL and control nB cells obtained using graphlets as opposed to spectral clustering.

Furthermore, we observe that the highest enrichments are generally obtained when using spectral clustering, i.e. in the clusters based on densely connected neighborhoods, rather than when using graphlet-based clustering (Fig. 2C and D), which corroborates the above observation that chromatin spatially groups DNA elements that form functional units. The same is observed when computing the enrichments of the modules found by the Louvain community detection algorithm (see Supplementary Fig. S6). However, because enrichment analysis tends to favor smaller clusters (communities), the comparison of community enrichments may be biased toward the network having larger number of smaller communities (i.e. the CS network of nB cells). Finally, the clusters based on wiring pattern similarities, independent on the nodes being in the same network location, are also found to be significantly enriched (Supplementary Fig. S5). This highlights the importance of the local wiring patterns in CS networks, i.e. of the local three-dimensional shapes of the chromatin around the DNA elements, as a new source of biological information, detailed in the next section.

The CS networks are produced by a different omics biotechnology and hence capture different functional information than PPI and COEX networks. We find support for this also by observing that the functional enrichments in CS networks differ from those in PPI and COEX networks: this is evidenced by small overlaps of the biological annotations that are found to be enriched in the CS network of nB cells, in the PPI network and in the COEX network (Supplementary Fig. S7). Hence, and as expected from very different omics biotechnology producing systems-level molecular data of different types, CS networks are a new source of biological information that complements the information contained in PPI and COEX networks.

In this section, we quantified the relationships between the wiring patterns of nodes in the networks and their biological functions. In the next section, we investigate if there is anything specific to the GVDs of cancer driver genes as opposed to other DNA elements in the CS networks and if that could be exploited to predict new cancer-related genes. Further exploring the specifics of GVDs of all DNA elements in CS networks in relation to their biological annotations is a topic of future research.

3.3 Structural markers of leukemia in CS networks

We investigate if CLL driver genes are characterized by specific local wiring patterns in the CS networks of leukemia and control cells, and test whether these wiring patterns could be used to uncover new leukemia-related genes. We use the more descriptive GDVs to characterize the local wiring patterns around nodes in a CS network and the GDV distance to measure the differences between the local wiring patterns of two nodes (see Section 2). As presented in Supplementary Figure S11A, the average GDV distances between the CLL driver genes in the CS network of CLL cells are significantly smaller than the average GDV distances between the background elements (with P-value $\leq 5\%$). This suggests that CLL driver genes have similar wiring patterns in the CS network of CLL cells, i.e. that in the CS of CLL cells there exists specific local patterns characteristic to CLL-related genes. Surprisingly, such cancer 'hot-spots' can also be identified in the chromatin network of healthy cells. Indeed, as presented in Supplementary Figure S11B, the average GDV distances between the CLL driver genes in the CS network of nB cells are significantly smaller than the average GDV distances between the background elements as is the case in the cancer cells. The corresponding GDV signatures are presented in Figure 3A and B.

Since GDVs of CLL driver genes are similar, for each DNA element that is currently not known to be a driver, we compute its average GDV distance from the known driver genes using the GDVs from the CS networks of CLL cells and also of nB cells. First, we present the results when using the network of CLL cells. We use the obtained average distance to prioritize DNA elements with wirings the most similar to the wirings of the known driver genes. To this aim, GDV distances are computed by considering only the graphlet orbits that have significantly different values between background and CLL driver genes (highlighted by black circles in Fig. 3A). We globally assess if the prioritized DNA elements, which we assume could be new CLL-related genes and non-coding RNAs, tend to be associated to cancer by measuring the percentage of our prioritized DNA elements that are dysregulated in CLL patients (see Section 2), or that are known to be mutated in CLL (mutation data from intOgene). As presented in Figure 3C, our top 5000 prioritized DNA elements in the CS network of CLL cells (having GDVs the most similar to those of CLL drivers) are indeed more frequently dysregulated and mutated. We formally measure this ability of prioritizing dysregulated and mutated genes using Receiver Operating Characteristic (ROC) curve analysis. We find that DNA elements having small GDV distances to CLL driver genes in the CS network of CLL cells are statistically significantly more frequently dysregulated (with Area Under the ROC Curve, AUC $= 0.61$ and P-value $= 1.4 \times 10^{-186}$) and mutated (with AUC $= 0.54$ and P-value $= 5.5 \times 10^{-25}$) than randomly chosen DNA elements (Supplementary Fig. S15A).

To further validate our predictions, we investigate the top 50 prioritized genes in the CS network of CLL cells and find literature evidence that at least 62% of them have a role in cancer (third column of Supplementary Table S3). Furthermore, 70% of them are significantly dysregulated in patients with CLL. Because of a lack of samples, we could not assess if the prioritized genes are prognostic markers of patient survival in blood cancer, so we used other available cancers in the Human Protein Atlas (Uhlén *et al.*, 2015) and found that 74% of these prioritized genes are statistically significant prognostic markers of patient survival in various cancers. We also investigate non-coding RNAs. However, due to the scarcity of the available knowledge on the roles of RNAs in cancer, the relevance of the prioritized RNAs cannot be assessed directly. Thus, we validate them indirectly by checking if their closest genes along the DNA are known to be involved in cancer in the literature. In this way, we find that 70% of our prioritized non-coding RNAs in CLL cells may be relevant to leukemia (Supplementary Table S4).

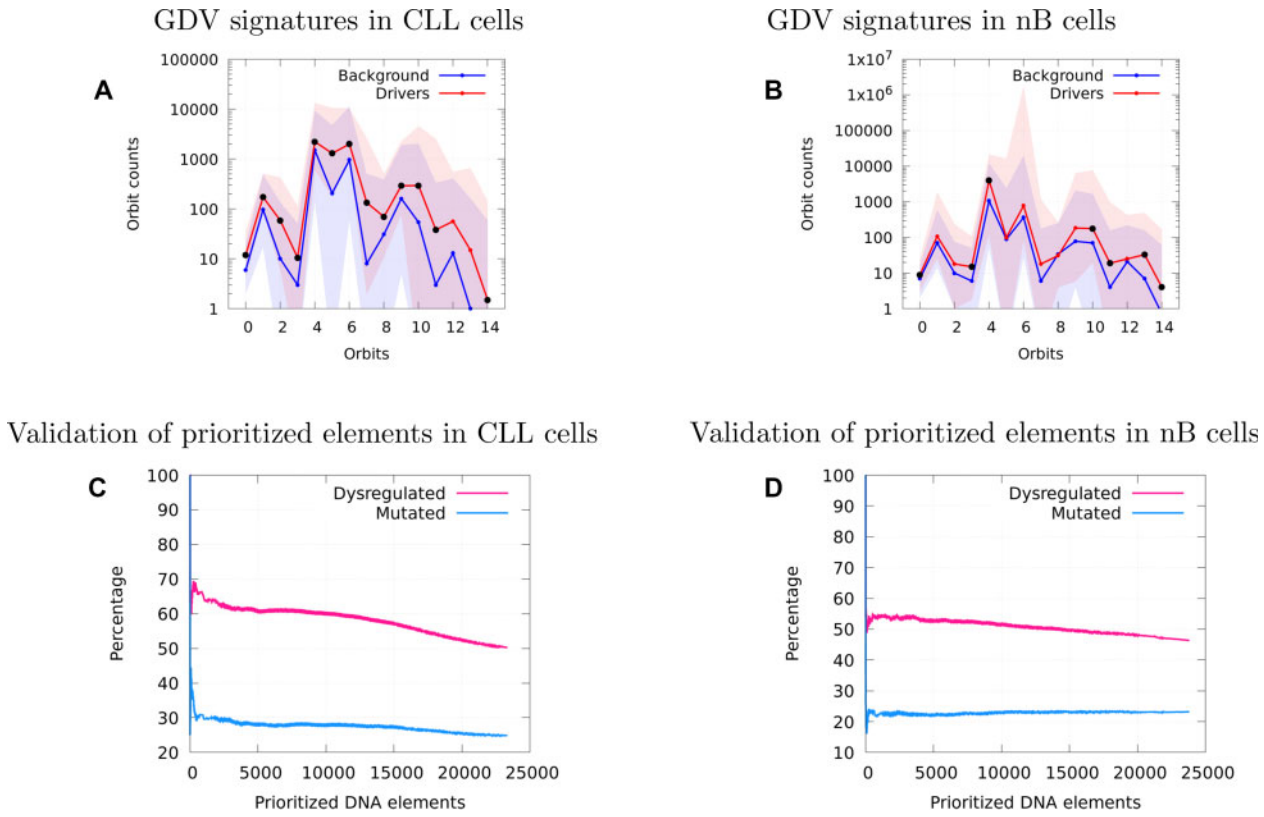

Fig. 3. Uncovering new CLL-related elements in CS networks. (**A**) The median GDV signatures of CLL cancer genes (in pink) and of background genes (in blue) in the CS network of CLL cells. Areas around the curves indicate the corresponding 15.9th and 84.1th percentiles. Graphlet orbits circled in black have statistically significantly different values for CLL genes than for background genes (with MWU-test *P*-values $\leq$ 5%). (**B**) shows the same, but in the CS network of nB cells. (**C**) For the top scoring prioritized DNA elements according to their GDV similarities to the driver genes in the CS network of CLL cells, we report the percentage of them that are dysregulated in CLL cancer (pink line) or that are mutated in CLL (blue line). (**D**) Shows the same, but for the DNA elements that are prioritized according to their GDV similarities to the driver genes in the CS network of nB cells

Importantly, we obtain similarly large validation rates when applying the same methodology on the CS network of nB cells. As presented in Figure 3D and Supplementary Figure S15B, DNA elements having small GDV distances to CLL driver genes in the CS network of nB cells are statistically significantly more frequently dysregulated (with AUC = 0.56 and *P*-value = 1.5×10^{-58}), but not mutated (with AUC = 0.49 and *P*-value = 0.08). When focusing on the top 50 prioritized genes, we find that 70% of the prioritized genes have literature support of their role in cancer, 72% of them are dysregulated in patients with CLL and 74% of them are prognostic markers of cancers according to the Human Protein Atlas (Uhlén *et al.*, 2015) (Supplementary Table S5). Furthermore, when focusing on the top 20 prioritized non-coding RNAs, we find that 65% may be relevant to leukemia (Supplementary Table S6).

The large validation rates that we obtain for predicting new CLL-related genes and non-coding RNAs from both CLL and nB cells suggest that the other top scoring elements may also be involved in cancer. For instance, the top scoring prioritized gene in CLL cells that has no direct support of its role in cancer in the literature is SLC35A5. Our results show that it is significantly differentially expressed in CLL patients (*P*-value $\approx 8.75 \times 10^{-23}$) and that its expression level is a prognostic biomarker in renal cancer (Supplementary Table S3). SLC35A5 encodes a nucleotide sugar transporter protein from the solute carrier family SLC35, which is known to be involved in many diseases including cancer (Ishida and Kawakita, 2004). While little is known about the two top scoring prioritized genes in nB cells that have no direct support of their role in cancer in the literature (C2ORF74 and C2ORF162), the third one, ZFYVE27, is likely to be related to cancer. Our results show that that it is significantly differentially expressed in CLL patients (*P*-value $\approx 3.78 \times 10^{-7}$) and that its expression level is a prognostic biomarker in colorectal, stomach and urothelial cancers (Supplementary Table S5). ZFYVE27 possess a Rab-11 binding domain that is known to determine the surface expression of adhesion proteins, which are critical parameters for adhesion, migration and invasion processes (Welz *et al.*, 2014). In particular, RAB25 (another Rab-11 protein) is a known oncogene in ovarian and breast cancers (Welz *et al.*, 2014).

Overall, our results confirm the existence of specific structural markers of CLL in the chromatin. That is, in the CS, there exist specific local patterns by which cancer related genes are wired. Such 'hot-spots' exist in the CSs of both cancer (CLL) and healthy (nB) cells, so both can be mined and used for finding new leukemia-related DNA elements.

To visually inspect these leukemia hot spots in the CS, we embed the largest connected component of the CS networks of CLL and nB cells in two-dimensional space and highlight the driver and prioritized genes. As presented in Supplementary Figure 9A and B, driver genes tend to be in the 'core' of the networks (this is particularly visible in the CS network of nB cells). We formally measure if drivers and prioritized DNA elements are more central in the networks using eccentricity (see Section 2). As presented in Supplementary Figure 9C and D, driver genes and prioritized elements are closer to the center of the network than to the periphery. However, driver genes, as well as our newly prioritized genes, are not forming a single 'core' in the CS, as evidenced by large average shortest path lengths between them (Supplementary Fig. S12). That is despite being central in the CS network, the cancer hot-spots are not in the same network region. As they are characterized by similar wiring and are central, we hypothesize that they are on local and central structural motifs that are repeated along the chromatin.

4 Concluding remarks

We present a comparative topological analysis of CS networks of CLL and healthy blood cells. We observe that the topologies of CS networks capture the modular organization of the chromatin that is associated with the phenotypes of the cells. By analyzing the wiring of DNA elements in CS networks, we find that in the nucleus, functional units are spatially organized so that they can be co-expressed, which is consistent with the current knowledge on the role of chromatin in expression regulation. We also find that the local wiring patterns around DNA elements in CS networks, captured by graphlets (independent of their location in the networks), also relate to their function and uncover new biological function that is different from the function found in PPI and co-expression data. These results confirm that CS networks are a new source of biological information that complements the information contained in other molecular networks.

While the 17 CS networks of healthy blood cell types have similar topologies (as measured with simple network statistics and with graphlets; see Fig. 1 and Supplementary Fig. S13) and functional coherences (as reported in our clustering and enrichment analyses; see Section 3.2, Fig. 2 and Supplementary Figs S4–S6), the comparison between the CS network of CLL cells and healthy control cells demonstrates that CLL induces a large rewiring in CS networks, making the CS network less modular and less functionally coherent. Our study also finds the role of CLL driver genes in these topological changes: they become hubs that connect modules that are disconnected in normal cells, they become more central and make the CS network less modular. As we already mentioned, potential rewiring biases coming from inter-individual variability and batch effects may exist in the data, so further work will be needed once larger datasets become available, ideally on control and CLL CSs coming from the same patients. However, as graphlet-based used here are very robust to noise, we do not expect the results to qualitatively change.

Finally and importantly, we show the existence of structural markers of cancer-related DNA elements in the chromatin of both control and cancer cells. CLL driver genes are characterized by specific local wiring patterns in the CS network of both healthy and CLL cells and these specific local wiring patterns allowed us to successfully predict new CLL-related genes and non-coding RNAs. Surprisingly, this indicates that we can identify cancer-related DNA elements in other cancer types by investigating the CS networks of their healthy cells of origin. Since it can be hard to establish the accuracy of chromatin contact maps in cells whose genomes are highly rearranged, as is the case for most cancer cells, this gives us an opportunity to exploit chromatin conformation data in healthy cells to predict new cancer-related DNA elements, which is our key insight, observed for the first time in this study, which may lead to profound new insights into disease and function. Our results are consistent with the observations that the chromatin organization contributes to regional variations in germline and somatic mutation rates (Makova and Hardison, 2015). A future research is to identify the mechanisms that connect these specific local wiring patterns in CS networks and the mutation, dysregulation and dysfunction of the corresponding genomic elements to ultimately pave the road to new therapeutic strategies.

Funding

This work was supported by the European Research Council (ERC) Consolidator Grant 770827, the Serbian Ministry of Education and Science Project III44006, the Slovenian Research Agency project J1-8155, The Prostate Project, and the Foundation Toulouse Cancer Santé and Pierre Fabre Research Institute as part of the Chair of Bio-Informatics in Oncology of the CRCT.

Conflict of Interest: none declared.

References

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Barabási,A.-L. and Albert,R. (1999) Emergence of scaling in random networks. *Science*, **286**, 509–512.

Beekman,R. *et al.* (2018) The reference epigenome and regulatory chromatin landscape of chronic lymphocytic leukemia. *Nat. Med.*, **24**, 868–880.

Bhoi,S. *et al.* (2016) Prognostic impact of epigenetic classification in chronic lymphocytic leukemia: the case of subset# 2. *Epigenetics*, **11**, 449–455.

Blondel,V.D. *et al.* (2008) Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.*, **2008**, P10008.

Bonev,B. and Cavalli,G. (2016) Organization and function of the 3D genome. *Nat. Rev. Genet.*, **17**, 661–678.

Cook,S.A. (1971) The complexity of theorem-proving procedures. In: *Proceedings of the Third Annual ACM Symposium on Theory of Computing*, May 3-5, pp. 151–158. ACM, OH, USA.

Csermely,P. *et al.* (2015) Cancer stem cells display extremely large evolvability: alternating plastic and rigid networks as a potential mechanism: network models, novel therapeutic target strategies, and the contributions of hypoxia, inflammation and cellular senescence. *Semin. Cancer Biol.*, 30, 42-51.

Davis,D. *et al.* (2015) Topology-function conservation in protein–protein interaction networks. *Bioinformatics*, **31**, 1632–1639.

Dixon,J.R. *et al.* (2016) Chromatin domains: the unit of chromosome organization. *Mol. Cell*, **62**, 668–680.

Döhner,H. *et al.* (2000) Genomic aberrations and survival in chronic lymphocytic leukemia. *N. Engl. J. Med.*, **343**, 1910–1916.

Erdös,P. and Rényi,A. (1959) On random graph. *Publ. Math.*, **6**, 290–297.

Fabbri,G. *et al.* (2011) Analysis of the chronic lymphocytic leukemia coding genome: role of notch1 mutational activation. *J. Exp. Med.*, **208**, 1389–1401.

Fabregat,A. *et al.* (2018) The reactome pathway knowledgebase. *Nucleic Acids Res.*, **46**, D649–D655.

Faisal,F.E. *et al.* (2017) Grafene: graphlet-based alignment-free network approach integrates 3d structural and sequence (residue order) data to improve protein structural comparison. *Sci. Rep.*, **7**, 14890.

Ferraro,A. (2016) Altered primary chromatin structures and their implications in cancer development. *Cell. Oncol.*, **39**, 195–210.

Ferreira,P.G. *et al.* (2014) Transcriptome characterization by RNA sequencing identifies a major molecular and clinical subdivision in chronic lymphocytic leukemia. *Genome Res.*, **24**, 212–226.

Fraser,J. *et al.* (2015) Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. *Mol. Syst. Biol.*, **11**, 852.

Gonzalez-Perez,A. *et al.* (2013) IntOGen-mutations identifies cancer drivers across tumor types. *Nat. Methods*, **10**, 1081–1082.

Ishida,N. and Kawakita,M. (2004) Molecular physiology and pathology of the nucleotide sugar transporter family (SLC35). *Pflügers Archiv.*, **447**, 768–775.

Javierre,B.M. *et al.* (2016) Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. *Cell*, **167**, 1369–1384.

Kamada,T. and Kawai,S. (1989) An algorithm for drawing general undirected graphs. *Inf. Process. Lett.*, **31**, 7–15.

Kipps,T.J. *et al.* (2017) Chronic lymphocytic leukaemia. *Nat. Rev. Dis. Primers*, **3**, 16096.

Kodinariya,T.M. and Makwana,P.R. (2013) Review on determining number of cluster in k-means clustering. *Int. J. Adv. Res. Comput. Sci. Management Stud.*, **1**, 90–95.

Kotlyar,M. *et al.* (2016) Integrated interactions database: tissue-specific view of the human and model organism interactomes. *Nucleic Acids Res.*, **44**, D536–D541.

Kulis,M. *et al.* (2012) Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia. *Nat. Genet.*, **44**, 1236–1242.

Landau,D.A. *et al.* (2014) Locally disordered methylation forms the basis of intratumor methylome variation in chronic lymphocytic leukemia. *Cancer Cell*, **26**, 813–825.

Lieberman-Aiden,E. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.

Lupiáñez,D.G. *et al.* (2015) Disruptions of topological chromatin domains cause pathogenic rewiring of gene–enhancer interactions. *Cell*, **161**, 1012–1025.

Makova,K.D. and Hardison,R.C. (2015) The effects of chromatin organization on variation in mutation rates in the genome. *Nat. Rev. Genet.*, **16**, 213–223.

Malod-Dognin,N. and Pržulj,N. (2014) GR-align: fast and flexible alignment of protein 3D structures using graphlet degree similarity. *Bioinformatics*, **30**, 1259–1265.

Mifsud,B. *et al.* (2015) Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.*, **47**, 598–606.

Milenković,T. and Pržulj,N. (2008) Uncovering biological network function via graphlet degree signatures. *Cancer Inf.*, **6**, CIN.S680.

Muscoloni,A. and Cannistraci,C.V. (2018) A nonuniform popularity-similarity optimization (nPSO) model to efficiently generate realistic complex networks with communities. *N. J. Phys.*, **20**, 052002.

National Cancer Institute. (2019) *Chronic Lymphocytic Leukemia Treatment*. https://www.cancer.gov/types/leukemia/patient/cll-treatment-pdq (03 March 2019, date last accessed).

Newman,M. (2010) *Networks: An Introduction*. Oxford University Press, Oxford.

Oakes,C.C. *et al.* (2014) Evolution of DNA methylation is linked to genetic aberrations in chronic lymphocytic leukemia. *Cancer Discov.*, **4**, 348–361.

Okamura,Y. *et al.* (2015) COXPRESdb in 2015: coexpression database for animal species by DNA-microarray and RNAseq-based expression data with multiple quality assessment systems. *Nucleic Acids Res.*, **43**, D82–D86.

Palstra,R.-J. *et al.* (2003) The β-globin nuclear compartment in development and erythroid differentiation. *Nat. Genet.*, **35**, 190–194.

Penrose,M. (2003) *Random Geometric Graphs. Number 5*. Oxford University Press, Oxford.

Pržulj,N. (2007) Biological network comparison using graphlet degree distribution. *Bioinformatics*, **23**, e177–e183.

Pržulj,N. and Higham,D.J. (2006) Modelling protein–protein interaction networks via a stickiness index. *J. R. Soc. Interface*, **3**, 711–716.

Pržulj,N. *et al.* (2004) Modeling interactome: scale-free or geometric? *Bioinformatics*, **20**, 3508–3515.

Pržulj,N. *et al.* (2010) Geometric evolutionary dynamics of protein interaction networks. In *Proceedings of the Pacific Symposium on Biocomputing*, January 4-8, pp. 178–189. World Scientific, Hawaii, USA.

Pržulj,N. (2019) *Analyzing Network Data in Biology and Medicine: An Interdisciplinary Textbook for Biological, Medical and Computational Scientists*. Cambridge University Press, Cambridge.

Puente,X.S. *et al.* (2011) Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature*, **475**, 101–105.

Ritchie,M.E. *et al.* (2015) Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**, e47.

Thibodeau,A. *et al.* (2017) Chromatin interaction networks revealed unique connectivity patterns of broad H3K4me3 domains and super enhancers in 3D chromatin. *Sci. Rep.*, **7**, 14466.

Uhlén,M. *et al.* (2015) Tissue-based map of the human proteome. *Science*, **347**, 1260419–1260419.

Vázquez,A. *et al.* (2003) Modeling of protein interaction networks. *Complexus*, **1**, 38–44.

Wang,L. *et al.* (2011) SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. *N. Engl. J. Med.*, **365**, 2497–2506.

Watts,D.J. and Strogatz,S.H. (1998) Collective dynamics of 'small-world' networks. *Nature*, **393**, 440–442.

Welz,T. *et al.* (2014) Orchestration of cell surface proteins by rab11. *Trends Cell Biol.*, **24**, 407–415.

Yaveroğlu,Ö.N. *et al.* (2015a) Revealing the hidden language of complex networks. *Sci. Rep.*, **4**, 4547.

Yaveroğlu,Ö.N. *et al.* (2015b) Proper evaluation of alignment-free network comparison methods. *Bioinformatics*, **31**, 2697–2704.