Sauta et al. BMC Bioinformatics (2020) 21:219
https://doi.org/10.1186/s12859-020-3510-1

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

A Bayesian data fusion based approach for learning genome-wide transcriptional regulatory networks

Elisabetta Sauta[1*], Andrea Demartini[1], Francesca Vitali[2], Alberto Riva[3] and Riccardo Bellazzi[1]

* Correspondence: elisabetta.sauta01@universitadipavia.it
[1]Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Via Ferrata 5, 27100 Pavia, Italy
Full list of author information is available at the end of the article

Abstract

Background: Reverse engineering of transcriptional regulatory networks (TRN) from genomics data has always represented a computational challenge in System Biology. The major issue is modeling the complex crosstalk among transcription factors (TFs) and their target genes, with a method able to handle both the high number of interacting variables and the noise in the available heterogeneous experimental sources of information.

Results: In this work, we propose a data fusion approach that exploits the integration of complementary *omics*-data as prior knowledge within a Bayesian framework, in order to learn and model large-scale transcriptional networks. We develop a hybrid structure-learning algorithm able to jointly combine TFs ChIP-Sequencing data and gene expression compendia to reconstruct TRNs in a genome-wide perspective. Applying our method to high-throughput data, we verified its ability to deal with the complexity of a genomic TRN, providing a snapshot of the synergistic TFs regulatory activity.
Given the noisy nature of data-driven prior knowledge, which potentially contains incorrect information, we also tested the method's robustness to false priors on a benchmark dataset, comparing the proposed approach to other regulatory network reconstruction algorithms. We demonstrated the effectiveness of our framework by evaluating structural commonalities of our learned genomic network with other existing networks inferred by different DNA binding information-based methods.

Conclusions: This Bayesian omics-data fusion based methodology allows to gain a genome-wide picture of the transcriptional interplay, helping to unravel key hierarchical transcriptional interactions, which could be subsequently investigated, and it represents a promising learning approach suitable for multi-layered genomic data integration, given its robustness to noisy sources and its tailored framework for handling high dimensional data.

Keywords: Genomic transcriptional networks, *omics*-data fusion, Bayesian networks, Hybrid structure learning algorithm

Background

The transcriptional regulatory machinery consists of cooperative interactions among transcription factors (TFs) responsible for regulating the spatial and temporal expression of genes in response to different cellular stimuli. Dysregulation of such transcriptional programs is one of the key hallmarks of cancer, affecting the clinical progression and the therapeutic responsiveness of the disease phenotype [1]. A promising approach for investigating the altered transcriptional response underlying cancer is to reconstruct the transcriptional dependencies among TFs and their target genes as a network, exploiting the genome-wide scale and the complementary data types offered by high-throughput technologies, to mine the resulting regulatory structure and extract interactions pattern from the genomic transcriptional hierarchy of the considered phenotype [2, 3].

Modelling such complex transcriptional regulatory networks (TRNs) represents one of the most challenging task in Computational Biology, given the high dimensionality of involved interactors and that their molecular dynamics are not fully understood [4]. For this reason, computational modeling is an essential component in reverse engineering of transcriptional networks. As He and Tan pointed out in their recent review [5], among current computational approaches for constructing TRNs, there is a lack of integrative genome-wide methods which combine *omics*-data sources to strengthen the accuracy of the obtained models and to provide novel insights from the inferred network structure. A particularly important issue is to find a method able to deal with the biological complexity of these systems, and that is sufficiently robust to scale their genomic dimension allowing multiple data integration.

During the last years, this aspect has been increasingly emphasized along with the rapid growth of high-throughput genomic data types. A variety of approaches have been exploited to predict the interaction of regulatory elements [6–8], including models focused on reconstructing physical locations of transcription factors through analysis of DNA sequence information, either using TFs binding site motifs, or chromatin accessibility data, as measured by DNase I hypersensitivity sites sequencing (DNase-Seq) or by transposase-accessible chromatin sequencing (ATAC-Seq) [7, 9–11]. Nevertheless, the regulatory activity that can be predicted from these binding affinities is limited to the set of TFs whose specific molecular sequences have been characterized, without taking into account the capability of certain TFs to recognize multiple motifs and the interaction with other cofactors, losing a proportion of potential transcriptional dependencies, that may help to depict a more comprehensive regulatory schema [12, 13].

Other computational strategies rely on learning the wiring transcriptional architecture, which orchestrates cellular gene expression, from transcriptomic data (in particular obtained by microarray experiments, widely available in different conditions and contexts) as a sole or primary source [14], using mathematical approaches including Boolean networks, information theoretic or correlation-based methods, differential equations systems, Bayesian and Neural networks [15–17]. Among these, Bayesian Networks (BNs) have become the prominent technique to model TRNs for their probabilistic formalism that can reflect the stochastic and combinatorial nature of gene regulation and for their ability to handle incomplete noisy data [18–20]. In this way, the network structure, constituted of causal and non-causal regulatory relationships among biological factors, is learned from genomic expression profiles, within a static or a dynamic schema. Friedman et al. [21] and Murphy and Mian [22] were among the

first to apply a Bayesian structure learning strategy on time-series data, trying to capture transcriptional dynamics in the temporal domain. The limited number of monitored time points is nevertheless statistically insufficient for reconstructing even a moderately-sized network, making this approach not suitable for human genome-wide transcriptional networks. Other methods [20, 23, 24] have focused their learning procedure only on static gene expression profiles that could produce unreliable biological transcriptional regulations, due to the noisy nature of this experimental source. Moreover, learning networks from a single data type gives a partial picture of the regulatory mechanisms, affecting the truthfulness of inferred results [25]. Data integration can overcome these limitations, allowing to build more accurate models, that are less prone to overfitting and more robust to noise and parameters perturbation [26].

To this aim, BNs provide an ideal probabilistic framework to handle heterogeneous data integration, and to incorporate biological functional information into the model as *prior knowledge*. Several structure learning methods have been tested to include prior knowledge in their search process, since the reconstruction of regulatory networks is computationally expensive [27, 28]. For instance, Imoto et al. [29] and Werhli et al. [30] represented biological priors in terms of energy function to evaluate the fitness of each learned network to the prior structure. Hartemink et al. [31], instead, included genomic location data as a model prior, forcing the search procedure to add arcs in a specific position, and discarding all graphs lacking these recommended edges. However, application of these algorithms is limited to small networks due to their complexity and high computational cost [32].

In this work, we present a data fusion approach for learning transcriptional Bayesian Networks in a high-dimensional space, exploiting heterogeneous *omics*-data integration, to determine the transcriptional architecture on a genome-wide scale. Our method implements a hybrid structure learning algorithm able to draw structural priors from Chromatin ImmunoPrecipitation followed by deep sequencing (ChIP-Seq) data. This type of epigenomic data produces a binding profile for each considered TF, consisting of all the target genes for which the TF is the transcriptional regulator. The integration of multiple genome-wide TF binding profiles allows reconstructing the circuitry of a regulatory network which captures the natural directionality of transcriptional flow.

Moreover, the algorithm exploits integrated gene expression data as evidence for both assigning prior probabilities to each individual transcriptional relation, and for learning the model parameters during its search process. This multi-layered -omics data integration can reveal topological hierarchies as a reflection of the transcriptional impact on gene regulation, which, to our knowledge, have not been investigated with a Bayesian learning strategy on a genomic scale.

We apply our novel framework to a chronic myeloid leukemia (CML) ChIP-Seq dataset, for gathering a data-driven prior knowledge to model the underlying transcriptional genomic interplay, whose overall structural consistency was further assessed through existing networks within the hematopoietic context. The performance of the proposed approach is then evaluated with other inference methods using as a benchmark a literature-derived transcriptional network of the yeast *Saccharomyces cerevisiae* that, for our purpose, is the only eukaryotic TRN available as gold standard with transcriptional cooperation level sufficiently complex if compared with ours.

Results

We applied our data fusion approach to a Chronic Myeloid Leukemia (CML) dataset, using data-driven prior knowledge gathered from the integration of TFs ChIP-Seq binding profiles, in order to prove its ability to handle a real genome-wide transcriptional network. Given the noise linked to this experimental data source, we then tested the robustness of our hybrid learning algorithm to incorrect prior information, evaluating it on a gold standard regulatory network, from yeast *Saccharomyces cerevisiae,* and comparing its learning performance to other inference strategies, as described in the Methods section. Moreover, to further assess the structural consistency of the CML transcriptional model, we examined its regulatory patterns comparing them to other hematopoietic networks derived by another class of inference methods, which use DNA sequence information to predict TF regulations [6, 9].

Performance assessment on high-throughput data

CML dataset

A collection of 65 TFs ChIP-Seq alignment data was retrieved from the Encyclopedia of DNA Elements (ENCODE) database [33] for the CML reference cell line K562. The binding profile of each TF was obtained through a bioinformatics pipeline, which included MACS2 peak calling [34], replicates consistency evaluation and peak-to-target assignment, in order to evaluate the statistical significance of the detected binding signals along the genome and to identify target genes. To further assess the consistency of the TF-gene interactions, each regulatory relationship was quantitatively weighted through a score-based method, which reflects the confidence of the considered binding event [35], discarding spurious interactions. The computational integration of all of the obtained genome-wide TFs profiles generated a genomic TRN composed of 20,876 nodes (65 TFs and 20,811 target genes), and 478,558 directed edges. Each edge was also weighted, using the score previously mentioned as a measure of the binding strength.

As first step, we dissected this genomic network into a TF-TF Component, characterized by 1827 edges between the 65 TFs, and a TF-Genes Component, which included the remaining network edges. Applying the BN design process (see Methods Section - Bayesian model definition) to the TF-TF Component, all weights (i.e. binding scores) associated with the arcs were sorted in decreasing order and evaluated within an iterative process aimed at finding a minimal connected directed acyclic graph (DAG). We obtained a whitelist of 1763 transcriptional relations and a DAG defined by 65 nodes and 64 interactions. This DAG was then combined with the TF-Genes Component to obtain a genomic transcriptional BN (TBN).

As a second omics data source, a compendium of microarray data from 122 CML patients was generated through the integration of five GE datasets, retrieved from GEO and ArrayExpress databases (GEO accessions GSE13159 [36], GSE47927 [37], GSE24739 [38]) (ArrayExpress accessions E-MTAB-2581 [39], E-MEXP-480 [40]). These data were normalized with the Robust Multi-array Average (RMA) technique [41], retaining the expression of those genes expressed in all the considered experiments. The TBN and the related whitelist were then integrated with this genomic expression panel and only relations among nodes for which the expression information was available were retrieved. The purpose was achieving a fully observable network, whose underlying distribution will be

modeled as a joint multivariate Gaussian, where the conditional density of each variable given its regulators can be represented as a linear Gaussian model (see Methods Section).

The resulting BN consisted of 11,986 nodes (of which 60 TFs) and 282,533 edges represented the initial structural input of our hybrid learning algorithm, together with the TF-TF arcs whitelist (1587 edges) for which the Pearson correlation between each TFs pair was estimated. After 100 runs of the algorithm, we collected 100 transcriptional BN models; the computational time required for learning all the obtained genomic networks on a single multi-core machine is shown in Fig. 1.

In order to obtain a transcriptional consensus BN and find consistencies across the learned structures, we estimated the robustness of each TF-TF edge from all the learned networks as a weight using Eq. (8), in order to rank these transcriptional relations, following the approach described in Methods Section - Consensus Transcriptional BN definition. We chose as a strict confidence threshold the weight value corresponding to the $5th$ percentile of the arcs weights distribution, to avoid the inclusion of edges with low confidence. The resulting consensus network was defined by 70 TF-TF edges; of these, 6 were present in the initial DAG but their directionality in the final TBN was reversed by the algorithm, as an effect of TRN regulatory loops.

The connectivity of each consensus node was then analyzed computing topological statistics as out-degree and in-degree that evaluate the number of incoming and outcoming edges for a node, respectively. In particular, these measures were used to calculate for each TF the hierarchy height metric [6], to topologically mine the chain of command underlying the transcriptional flow of the network. We identified a three-layered hierarchy, as illustrated in Fig. 2 (II), representing the regulator activity of different TF classes, composed of 20 master regulator TFs, at the top, 16 brokers or middle managers, and the remaining 24 workhorses TFs, at the bottom.

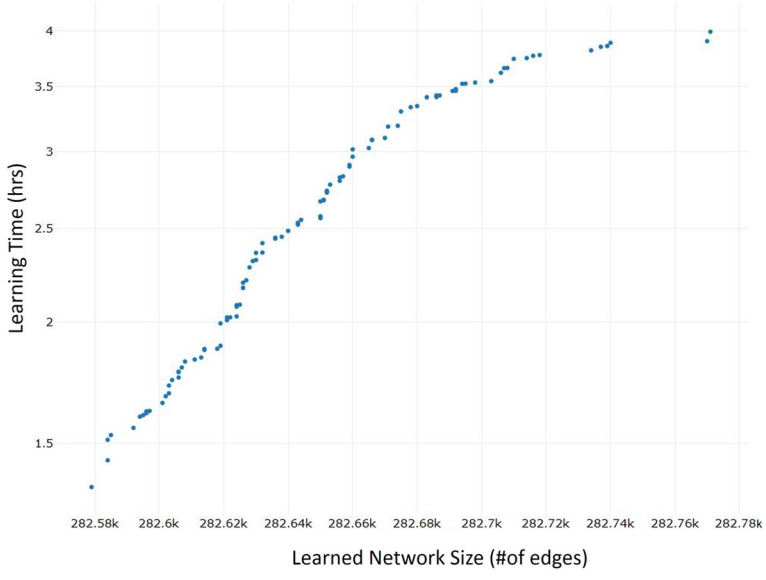

Fig. 1 Performance evaluation of the hybrid structure learning algorithm on CML network. The execution time of the search strategy to learn the BN model structures is plotted against the number of edges of each learned genomic network

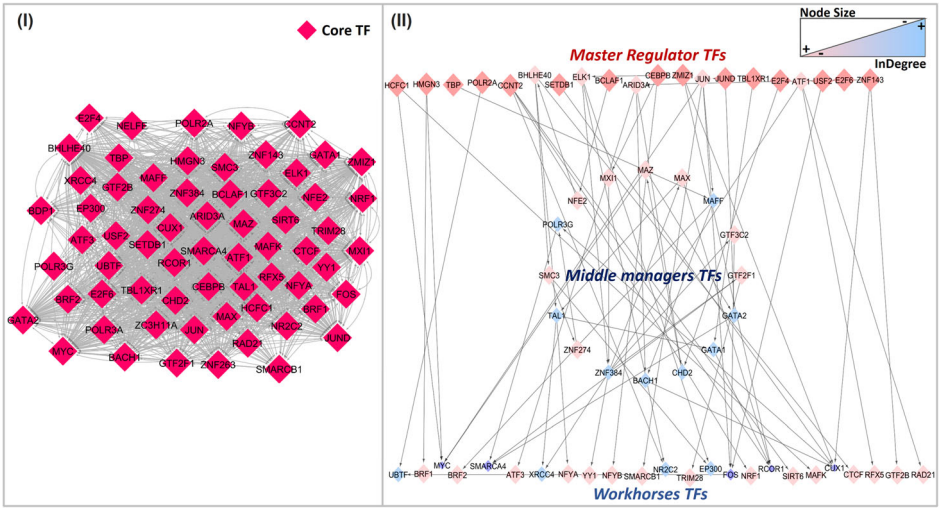

Fig. 2 Comparison of transcriptional interactions models. Panel (I). Transcriptional regulations characterizing the TF-TF Component derived from ChIP-Seq data integration. Panel (II). Transcriptional hierarchy underlying the Consensus Bayesian Network. The color intensity and the size of TFs nodes are proportional to the incoming connectivity (e.g. small size combined with a darker color for a high number of incoming edges)

This hierarchical organization was not detectable in the initial TF-TF component, derived from the sole integration of ChIP-Seq binding profiles, and due to the intrinsic complexity and compactness, the network has a typical "hairball" representation, as shown in Fig. 2 (I). As demonstrated by its high average node connectivity [42] ($\overline{k}(TF\_TF_{Component})$ =20), all nodes are consistently interconnected to each other without a specific topological order. Clearly, the resulting regulatory schema emerged thanks to the transcriptomics data integration within the structure learning framework, where gene expression guides the learning phase in two steps of the algorithm search process. First, the sampling process is led by the arc extraction probability equivalent to the Pearson correlation calculated between each TFs pair, ensuring that edges tied to high correlated TFs have a greater probability to be included in the structural model, as reported in Supplementary Figure 2 in Additional File 1. Second, during the learning of the model structure, gene expression values were used for the estimation of parameters which define the probability distribution of each node given its parents regulators, as explained in Methods Section.

Robustness evaluation to false prior information

Benchmark datasets

We retrieved all available transcriptional regulations in yeast among known TFs and target genes, which map to verified Open Reading Frames (ORFs), from the YEASTRACT [43] and Saccharomyces Genome Database (SGD) [44] repositories.

We used the normalized data from Spellman et al [45] as gene expression (GE) information, considering only those genes identified by the authors as cell-cycle regulated, and with a missing values rate less than 10%. We performed a *k*-nearest-neighbor imputation, obtaining a final complete dataset of 473 cell-cycle related genes expressed in 77 samples.

Combing the validated transcriptional binding information with GE data, we defined as ground truth a yeast regulatory network (yTRN) composed of 33 TFs and 437 target genes, and 3299 transcriptional regulations, 249 of which were TF-TF interactions.

To test the robustness of our method to incorrect prior information, we randomly added an increasing number of false edges to the yTRN, from 10 to 60% of the total number of TF-TF regulations. We considered each known interaction as true positive (TP), and every additional incorrect arc as false positive (FP). The performance of our method compared to the one of BANJO (Bayesian network interference with Java objects) [46] and ARACNe-AP [47] was evaluated for each FPs percentage, considering both the number of false edges included in the final model, and the fraction of true interactions among all inferred ones (precision). BANJO is a structure learning algorithm which, combining simulating annealing and a greedy search, finds and scores candidate networks inferring them from discretized expression data. ARACNe-AP uses instead the mutual information metric estimated from gene expression data input and data processing inequality to infer relations from a predefined list of TFs to their targets (see Competing Methods).

The yTRN underwent the BN definition procedure and was decomposed into a TF-TF Component of 33 TFs and 249 interactions, and a TF-Genes Component with 470 nodes and 3050 interactions. Using the option for unweighted transcriptional data, the TF-TF component was submitted to the iterative process, and a DAG with 33 nodes and 32 TF-TF interactions was then obtained. Combining it with the other Component, we defined the structure of the initialized model, consisting of 470 nodes and 3082 edges. This starting TRN and the arcs whitelist, whose dimension varied according to the considered FP rate, were used to test the proposed approach in all of the six incorrect prior conditions. We collected 100 learned transcriptional BNs for each tested FP percentage, and we evaluated the computational performance of our learning method on them, considering the time used by our algorithm to learn all the obtained networks, as illustrated in Fig. 3. The average computational time estimated on the total number of transcriptional BN models for all FP levels varied from 1.61 min to 2.00 min.

We then applied the “consensus” approach, described in Methods Section - Consensus Transcriptional BN definition, on each set of learned networks. For all the analyzed FP percentages, we selected the $25th$ percentile of the arcs weights distribution to find the confidence threshold for including only high confidence arcs in the related yeast Consensus BNs. The performance of our algorithm throughout all tested FP levels is reported in Table 1A and in Supplementary Figure 1.

BANJO was evaluated in each incorrect prior scenario taking as input data the discretized GE yeast dataset, the same initial DAG structures exploited by our approach on the yeast dataset, and a blacklist, to avoid gene-gene interactions and unrealistic regulations from genes to TFs. We ran BANJO using default parameters, and a fixed search time (5 h) as a stop criterion. All results are summarized in Table 1B.

ARACNe-AP cannot be evaluated under these incorrect prior conditions since it infers the network structure using GE data and a list of regulators (the considered 33 yeast TFs). Its Consensus network was obtained after 100 bootstraps from gene expression samples, using a MI threshold of 0.2989 estimated on the provided GE data. The number of FP edges calculated on the total consensus arcs is shown in Table 1C.

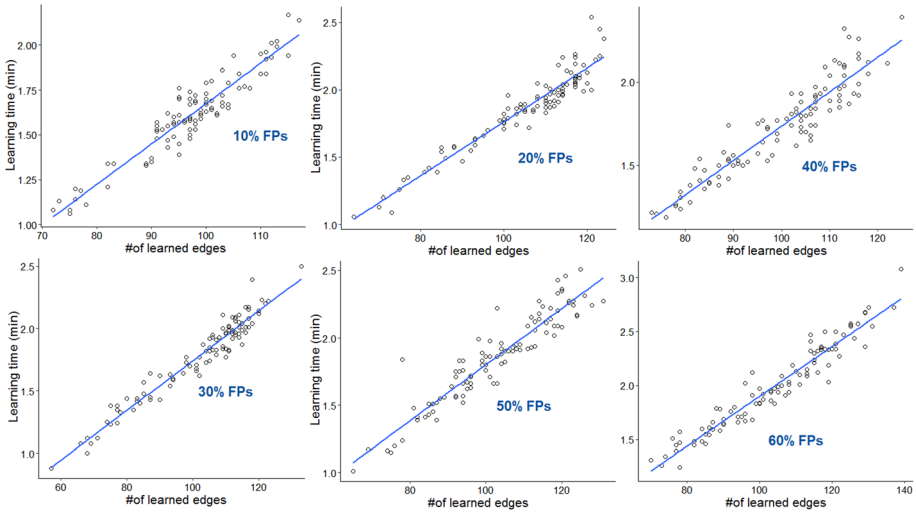

Fig. 3 Performance evaluation of the hybrid structure learning algorithm on yeast network. For all the considered FP levels, we analyzed the execution time of the search strategy to learn the BN model structures, comparing it to the number of edges of each learned network

Table 1 Summary of all results obtained from the comparison of described methods: (A) Data Fusion approach, (B) BANJO and (C) ARACNe-AP

Tested FP rate	Consensus BN size (#of edges)	% FPs added	Precision
A. Data Fusion Performance Results			
10%	50	8%	0.92
20%	58	12%	0.88
30%	56	10%	0.88
40%	60	12%	0.88
50%	69	11%	0.88
60%	76	12%	0.88
B. BANJO Performance Results			
10%	69	60%	0.41
20%	69	60%	0.41
30%	69	60%	0.41
40%	69	40%	0.59
50%	69	40%	0.59
60%	69	60%	0.41
C. ARACNe-AP Performance Results			
	1003	70%	0.3

Comparing the Data Fusion (DF) approach with BANJO, which, as our method, exploits prior knowledge for guiding the learning phase, DF showed a higher precision and robustness despite the progressively higher FP rate included in each prior. Moreover, it does not require a blacklist, as instead for BANJO, to avoid the inclusion of unreliable regulations (i.e. relations from gene to TF), that, for this benchmark interactome of moderated size is composed of 205,391 edges. Clearly, for a genomic network with thousands of nodes, the blacklist definition composed of all relations from genes to TFs and from gene to gene would become more computationally onerous. DF outperformed also ARACNe-AP, which reached the lowest precision, highlighting how the expression alone makes difficult to discriminate between a highly correlated regulator-target gene pair and a true causal relationship. This learning method indeed relies on a single data source, and the only "prior" allowed is the list of regulators among which DPI procedure is applied to infer dependencies.

Comparison with other hematopoietic transcriptional networks

To further assess the effectiveness of our framework and the overall quality of our transcriptional data-driven prior used for determining the final TBN structure, we next examined the extent to which ChIP-derived transcriptional patterns agreed on existing networks obtained with a different inference methodology class. Considering the hematopoietic context, we compared our learned CML interactome with other networks derived from K562 cell line data through inference methods that exploit genomic locations information to derive the network backbone [6, 9] but without a structure learning schema. In this analysis, structural comparisons were performed evaluating all the transcriptional relationships shared by common regulators among the considered networks.

DNaseI-footprinted hematopoietic transcriptional networks

We exploited the transcriptional network obtained from DNaseI footprints of K562 data integrated with a predictive motif-based search of known TFs binding sites [9]. We extracted from this interactome all relations driven by the regulators in common with our TFs set, constituting a subnetwork of 38 TFs nodes and 165 edges, as depicted in Fig. 4, panel I. Considering all transcriptional dependencies shared by our ChIP-Seq derived regulations and the resulting DNaseI subnetwork, we reached a structural regulatory homology of about 60%, represented with blue edges in Fig. 4 (I), despite the difference of the applied inference techniques. In this set of common arcs, transcriptional relations of the final TBN are also included (red colored edges in Fig. 4 (I)). They represent high confidence arcs as a result of the selection process to which they underwent within our framework. The choice of transcriptional relationships is indeed determined by a trade-off between edge prior probabilities and the inherent ability to explain expression of a target gene in the learning phase of the hybrid algorithm and within the procedure for the consensus model outlining, by a quantitative score as a strength measure of dependencies across all learned models.

Moreover, to further demonstrate the consistency of the transcriptional information enclosed in our prior, the comparative analysis was extended considering cross-regulatory interactions among the major transcriptional factors TAL1/SCL, SP1/PU.1, ELF1, HES1, MYB, GATA1 and GATA2, which have been extensively characterized for their lineage commitment role on hematopoietic cells [48, 49]. Examining the related networks

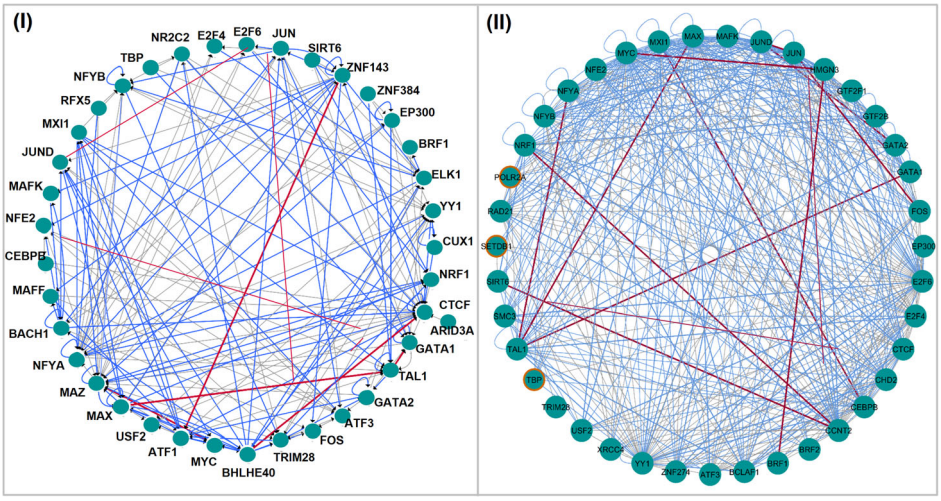

Fig. 4 Structural comparison of K562 data derived transcriptional networks. (I). DNaseI-footprinted subnetwork. (II). TFs co association-based subnetwork. Common transcriptional dependencies are highlighted with blue edges and with red edges if the relationship is also included in the final TBN model. Orange bordered nodes are TFs considered only as targets in the TFs co-association network and not as source regulators like in the TBN

obtained from the DNaseI footprinted interactome and our data, we obtained a structure similarity of 59%, as shown in Fig. 5. For SP1/PU.1, ELF1, HES1 and MYB, we cannot infer regulations (dashed grey arcs in Fig. 5), since we do not have the related ChIP-Seq binding profiles. On the other hand, some transcriptional relationships reconstructed only from our data (solid grey arcs in Fig. 5) are verified in literature within the hematopoietic context. For example, GATA2 play an essential role for the maintenance and proliferation of hematopoietic progenitor cells through tightly regulated interactions with other hematopoietic-associated TFs, including HES1 [50, 51], TAL1/SCL and MYB [52–54], patterns that, if altered, commonly lead to leukemogenesis. Another key factor is TAL1/SCL whose regulatory circuit, in which ELF1 and MYB take part, directs the expression of genes involved in the differentiation of blood cells types [48, 55–57]. These transcriptional dependencies are instead absent in the DNaseI derived network, highlighting the limitation of a knowledge-based transcriptional network reconstruction that, in this case, is constrained by the availability of known TF recognition sequences.

Given these considerations, we have also evaluated the subnetwork composed of all transcriptional interactions involving TAL1/SCL, GATA1 and GATA2, for which the binding profiles can be mined from both our data and the DNaseI derived network. Examining common regulations from shared TFs, we gained a 63% of structural concordance as represented in Supplementary Figure 3 in Additional File 1, and a 75% of edges as further source of transcriptional information included in our prior but missing in the DNaseI inferred regulations. Details about this analysis are reported in Additional file 1 (Major hematopoietic regulators subnetwork comparison).

TF co-association based hematopoietic transcriptional network

In order to perform this comparison, we considered the transcriptional network reconstructed from a TF co-association model learned from K562 ChIP-Seq co-binding

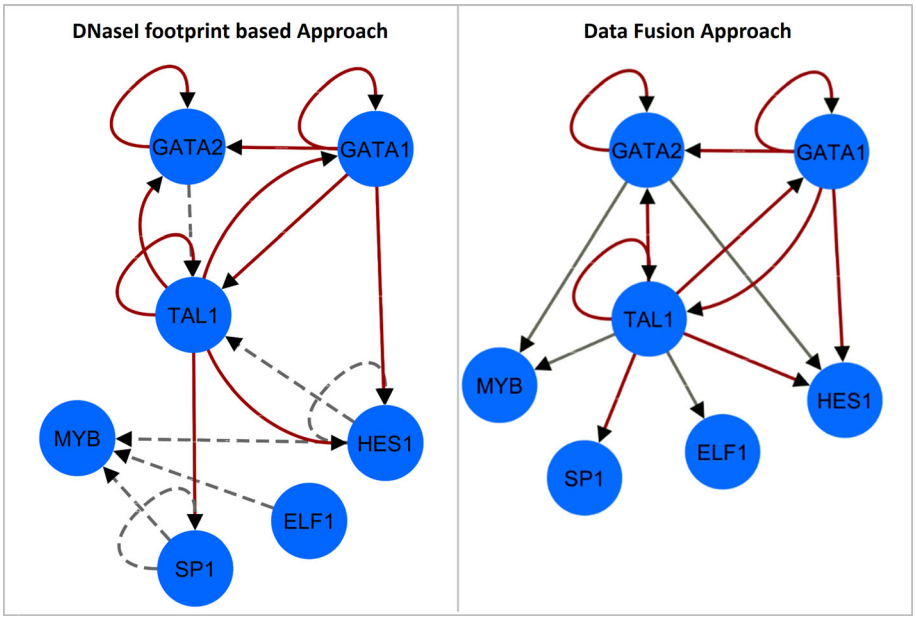

Fig. 5 Major hematopoietic TFs network comparison. Shown are regulatory networks comprising edges among seven hematopoietic regulators derived by DNaseI footprints (left panel) and our proposed integrative framework (right panel). Transcriptional relationships are red colored to indicate common regulations or grey colored if they are specific for the considered network. Dashed grey arcs represent dependencies for which binding data are not available

matrix with a discriminative machine learning approach [6]. Evaluating common regulators between this network and our TFs group, we defined a core of 39 TFs and 464 edges, as depicted in Fig. 4, panel II, in which almost the entire interactions (~ 93%) are shared between the considered transcriptional models. To further highlight this structural overlap, we have also investigated the underlying regulatory hierarchy using both the hierarchical score, applied on the TFs co-association derived network, and the hierarchy metric, used on our model. We obtained again a high degree of similarity in terms of level assignment, confirming our structural regulatory diagram. Results of this analysis are described in Additional file 1 (TF co-association based hematopoietic transcriptional network: Hierarchical comparison).

Among common interactions, consensus relationships verified in the previous analyses were also supported by this comparison. TAL1-CEBPB-GATA2 are primary-interacting partners of GATA1 which guide lineage-specific differentiation of hematopoietic cells. GATA1 has been shown to recruit TAL1 at several erythroid enhancers [58], regulating gene expression after being directed to a distinct subset of genomic binding sites in multi-lineage cells via its association with different complexes containing master regulators such as GATA2, CEBPB and RUNX1 [59, 60]. Moreover, GATA1 has a competitive behavior for the binding in regulatory complexes with GATA2, with which it is frequently co-associated [61]. Together with MAX, these factors constituted a regulatory circuit involved in erythrocyte and myeloid differentiation [62]. In addition to TAL1-CEBPB, another novel pairing CCNT2-HMGN3, identified and validated in the Gerstein et al. study [6], is also present in our final model. The Activator protein-1 (AP-1) complex consists of JUND-JUN-FOS factors, all of which are known to form heterodimeric protein aggregates with each other, which promote

myeloid differentiation, and genetic lesions affecting their expression have been associated to the leukemogenesis process [63, 64].

Furthermore, in the TF co association-based network, three of the 39 core TFs were considered as targets (POLR2A, TBP and SETDB1, orange bordered nodes in Fig. 4, II) and not as source TFs which can regulate other nodes like in our network, missing a part of regulatory information. POLR2A and TBP are key component of the core transcriptional machinery whose interaction with other hematopoietic co-regulators such as CEBPB, SP1, RUNX-related factors can modulate gene expression programs during myelopoiesis [65]. SETDB1 is instead a gatekeeper of tumor survival whose chromatin remodeler role is recently emerged as a potential therapeutic target for immunotherapy to avoid leukemic cells evasion from immune system [66, 67].

Discussion

In the era of 'Omics', data integration represents a challenging tool to deliver more comprehensive insights into the biological system under study, helping to translate novel molecular knowledge into improved disease understanding. In particular, going deeper into cancer deregulated gene expression programs, investigating their first level of regulation, where the transcriptional determinants act on a genome-wide scale, may help to define the molecular signatures driving the patient's phenotype. To this aim, the development of a robust computational approach able to deal with omics data heterogeneity and with this biological complexity is mandatory.

In this work, we proposed a data fusion approach which exploits multi-layered genomic data integration, allowing to model large-scale transcriptional networks within a Bayesian formalism. Our hybrid structure learning strategy allows to use ChIP-Seq transcriptional binding profiles as prior information, to both initialize the model structure and to constrain the search space. In particular, it models the natural directionality of the transcriptional flow, also evaluating edges which belong to feedback regulatory loops, and whose direction may be reversed by the algorithm. The learning procedure also exploits gene expression (GE) data integration, which acts on the initial search phase, (*i*) with the correlation, as a sampling probability tied to the arcs whitelist, allowing that a higher correlation will be translated into a greater prior probability for a transcriptional dependency to be included in the final model; and then (*ii*) on the estimation of the model parameters, specifying how combination of TFs functionally regulates the expression of their targets. We proposed a prior-based approach that it first reconstructs the regulatory skeleton and then refines the network structure using condition-specific expression data, prioritizing the underlying regulators. Using this joint learning schema, we obtain increased accuracy of the reconstructed transcriptional networks compared to those approaches which rely only on a single data source (i.e. GE data), such as ARACNe-AP. Despite these inference methods are widely used, as already pointed out [68], the expression alone makes it difficult to gain mechanistic insight between a highly correlated regulator-gene pair and a true causal relationship between them. In this framework, learnt models that are generated from each learning run of the algorithm underwent to a "Consensus" definition procedure, which ensures that only consistent dependencies appear in the final transcriptional model, reducing the occurrence of weak relationships. In addition, the estimated correlation is converted

in probability linked to each edge, as a further measure of 'robustness' of the considered binding event.

We applied our data fusion method to a CML -omics dataset, to test its computational ability to learn a genomic transcriptional network and model its complex transcriptional interplay. Although the mutational causative event of the considered disease is known to be the BCR-ABL1 gene fusion, the underlying transcriptional architecture has not been deeply investigated yet. Therefore, we wanted to mine the molecular mechanisms linked to the considered disease maintaining a genome-wide overview. To this end, we used integrated ChIP-Seq binding profiles from K562 cell line that is a representative in-vitro model of the CML, combining the resulting functional readouts with gene expression data from untreated CML patients, allowing to emerge only disease linked processes. The proposed learning strategy enabled us to identify a stratified hierarchy in the final consensus transcriptional Bayesian Network, representing the overall system-level regulatory wiring, which was undetectable in the initial CML transcriptional network. Indeed, the starting TRN, obtained from the integration of a single data type, the ChIP-Seq binding profiles, showed a high compactness and a complex connectivity, emerging from TF-TF interactions, due to the cooperative behavior of TFs, difficult to translate into a meaningful biological inference. The three-layered hierarchy instead can be interpreted as the effect of regulator impact of different TF classes (master regulators, middle managers and workhorses) on gene expression cellular programs, since the learning phase of the hybrid algorithm is driven by the transcriptome expression. These specified TF levels collectively control the non-regulator gene targets, lying in a lowest fourth layer that, due to its large size, cannot be graphically showed.

The correlation between the topological and functional aspects of TF, established within this hierarchy in a genome-wide perspective, represents an interesting novel result for the considered disease that could be further experimentally investigated. A pivotal role of the epigenetic regulation is also emerged from these transcriptional interactions, whose importance and implications for leukemia have been recently emphasized [69]. Moreover, most of these transcriptional dependencies were confirmed in other hematopoietic transcriptional networks differently inferred from K562 data sources. For example, the SETDB1 TF that in this context has been topologically classified as a master regulator (MR), is characterize also by an epigenetic activity, regulating gene expression via chromatin remodeling. Aberrant SETDB1 functionality and the related altered epigenetic changes have been shown to promote silencing of tumor suppressor genes, and thus contributes to enhance tumor growth and metastasis [70]. Moreover, SETDB1 maintains hematopoietic stem cells, restricting the activation of non-hematopoietic genes in normal conditions [71] while, when deregulated, it enables leukemia cells to evade innate immune controls allowing them to expand [66]. Our result highlights the importance of this TF, confirmed by its emerging role as a promising therapeutic target for several types of cancer [72, 73], including other forms of acute and chronic leukemia.

Another MR, the CEBPB TF, within the hematopoietic system is effectively indicated in the literature with as playing the role of MR, expressed at high levels to regulate genes involved in immune and inflammatory responses. Under stress conditions, such as cancer microenvironments, CEBPB is involved in BCR–ABL1 mediated myeloid expansion and leukemic stem cell exhaustion in the CML chronic-phase [74]. The MR

ZNF143 was observed to bind CEBPB and other C/EBP factors, whose interactions are required for a balanced expression in myeloid cells and for granulocytic differentiation of myeloid progenitors [75]. Members of the Jun family (JUN and JUND), that are key subunits of the transcription factor AP-1, are designated as MRs in healthy and cancer cells [76], given their crucial role in cell cycle progression, differentiation and programmed cell death. Not surprisingly, they are frequently overexpressed in leukemia, and their leukemogenesis actions are BCR-ABL1-induced [77]. Despite RAD21 and SMC3 TFs belonging to the same cohesin complex involved in DNA damage repair and whose composing genes are frequently mutated in myeloid neoplasms [78], these regulators are located at different network layers as a result of their different effects on their regulating modules. The same observation can be drawn for the heterodimeric complex composed of MAX and MYC genes, situated in the central and lower part of the hierarchy, representing their sequential recruitment necessary for regulating hematopoietic homeostasis [79]. Furthermore, MYC maps in the same layer of YY1, another known cooperating partner, whose expression alteration impacts on the MYC oncogene function. CTCF also co-localizes with RAD21 and together with SMC3 are commonly associated with insulator elements to mediate long-range interactions affecting the higher-order chromatin structure [80]. SMARCB1 and SMARCA4 interacting TFs lie in the same hierarchical level; both belong to the SWI/SNF complex as chromatin remodelers, playing an important function in pluripotency and cellular reprogramming. Recently, their involvement in maintaining oncogenic gene expression program in myeloid leukemia, in particular for the tumor suppressor SMARCB1, have been demonstrated [81]. GATA1 and GATA2 are two fundamental TFs which play a crucial role in gene regulation during development and differentiation of hematopoietic cells. They belong to the same hierarchical layer, reflecting their sequential molecular recruitment; it is indeed know that GATA2 binds the promoter region of GATA1 whose expression can be repressed in the hematopoietic stem and progenitor cells [53]. ATF1 is a master regulator, as detected in our model, often traslocated or overexpressed in blood malignancies, promoting leukemic cells expansion and resistance [82]. TBL1XR1 is a factor required for the activation of multiple intracellular signaling pathways important for hematopoietic cells fate, not surprisingly identified in this context as master regulator. A variety of genomic alterations was identified on this gene in several forms of leukemia, and its loss observed in recent studies has been proposed as a potential therapeutic target [83]. E2F4 and E2F6 play an essential function in specifying lymphoid subtype, orchestrating the activation of essential cell cycle progression genes and other key TFs, such as EBF1, required for normal and malignant B-lymphocyte development [84, 85]. Moreover, these E2F proteins have been found to co-associate with HCFC1/HCF-1, another TF that, in our hierarchy, was classified as master regulator, inducing histone methylation and transcriptional activation and contributing to leukemogenesis [86]. Recently, it has been shown that expression of the BCLAF1 MR is increased in leukemia blasts relative to normal precursor populations and suppression of this TF highlighted its potential in neoplastic self-renewal program, causing reduced proliferation and leading to induction of differentiation to a dendritic cell fate [87, 88]. HMGN3 belongs to a family of chromatin remodeling proteins that are enriched in aggressive cancers and stem cells, due to their role in maintaining nuclear organization critical for stem cell properties, both during development and oncogenesis. These factors are

frequently overexpressed in leukemia, enhancing aberrant gene transcription [89, 90]. Underlying this proposed regulatory schema clearly emerged a key role of the epigenetic regulation, whose involvement in leukemia-related processes has become of clinical relevance in the last years.

We then benchmarked our procedure against the yeast transcriptional network, demonstrating the robustness of the method to an increasing amount of false positive prior information that can also interpreted as a noisy source, intrinsic characteristic of experimental data. The validation of our procedure was performed only on the yeast dataset, because only a few experimentally verified eukaryotic transcriptional networks are commonly available as gold standards, like yeast *S. cerevisiae* and *E. coli*. This last one has a transcriptional network not sufficiently large and complex to apply our hybrid learning strategy, since the TF-TF counterpart with 214 TF-TF interactions annotated with strong evidence in RegulonDB database [91] is poorly enriched of TFs coregulations. The proposed data fusion technique is tailored for investigating complex transcriptional networks enriched of many coregulatory interactions, as human transcriptional networks, since these co-binding events are initially exploited by the algorithm to define its structural priors and the Bayesian model, and during the learning phase to add or reverse an existent arc. This implicates that the final probabilistic model will include only high confidence and consistent relations across all learned structures, as a selection of all initial coregulations.

Given the peculiarity of the proposed structure learning strategy, it is difficult to find similar published approaches for the validation process. Despite some common features, we excluded a recent Bayesian structure learning tool, bnlearn, since it forces arcs designed as structural prior (specified through a *whitelist*) to be included in the final model, preventing the addition of any other extra transcriptional regulation. For these reasons, we performed a comparison with the hybrid search algorithm implemented in BANJO, and with ARACNEe-AP, an ARACNe based approach for reconstructing transcriptional regulatory network. Our algorithm outperforms both methods, producing a significant improvement in structural accuracy, even with a progressively higher FP rate.

ARACNe-AP bases its structural reconstruction only on a single source of data (GE data), and this penalizes the correctness of the inferred transcriptional relations, 70% of which are FP predicted interactions. On the other hand, BANJO allows specifying a structural prior, but its implemented constraints and parameters, whose setting is not trivial (i.e. initialTemperature, coolingFactor, reannealingTemperature, etc.), does not enable to perform an accurate learning.

From a computational perspective, our approach is fast and scales well, thanks to its search method, particularly appropriate for parallel computation, and for the learning phase, based on local learning, while most Bayesian reconstruction methods, which use prior knowledge, are not practical for large networks. Our algorithm does not constrain the number of interacting variables or the maximum number of parents for each variable, as done in other methods [31, 92] and in BANJO, for which is advised to set this threshold less than seven, due to memory requirements needed for the learning. Moreover, it does not require a list of forbidden arcs (*blacklist*), like BANJO or bnlearn, whose definition for large-scale transcriptional networks is equivalent to 2^n (where n is the number of genes) interactions to exclude.

This data fusion method is designed to exploit data integration at different levels analyzing the resulting combined information in a unified framework, to gain insights into molecular signatures potentially driving disease phenotypes. In this study, we used high-throughput genomic datasets as the case of greatest complexity to present the ability of the approach in handling genome-wide transcriptional networks, whose modeling represents a novelty in the field of Bayesian structure learning algorithms.

The proposed methodology was conceived to use a genomic transcriptional interactome, classically enriched of co-regulations among TFs, as a primary integrative source for Bayesian model initialization and structural priors definition. Interactions among TFs, the main property of transcriptional networks, is the only feature required for the input transcriptomics data, making the learning algorithm highly adaptable to different transcriptional sources that infer TF binding events either with a direct approach (e.g. ChIP-Seq) or with an indirect one. Indeed, given the experimental heterogeneity and availability of published transcriptional data, the starting transcriptional network can be built in different ways, not only pooling together several ChIP-binding profiles, as was done in this study, but for example integrating ChIP data with TF-TF regulations derived from biological interactions databases such as STRING [93] or BIOGRID [94], or with motif-based search to find core regulators of the network, also combined with accessible chromatin profiles obtained from ATAC-Seq and DNase-Seq assays. The adaptability of our prior-based approach allows to integrate different –omics types of regulatory evidences to infer a genome-wide transcriptional network, whose structure, defined by TF dependencies, is then refined with condition-specific expression data, performing a comprehensive characterization of the potential factors driving disease transcriptional signatures.

Conclusions

In this work, we proposed a data fusion-based approach that, exploiting -omics data integration, is able to reconstruct genome-wide transcriptional networks and, using Bayesian modeling, enables a probabilistic assessment of the underlying structure within the hybrid learning algorithm. An innovative aspect of our method is that the structural properties of the initial reconstructed network are defined from ChIP-Seq data and are used as prior knowledge. Combining such informative priors with a search and score schema both at the local and global levels of the structure model, the algorithm efficiently handles genome-scale networks. Moreover, the “consensus” approach allows including only high confidence learned interactions in the final transcriptional BN.

We studied the transcriptional landscape of chronic myeloid leukemia, which, to our knowledge, has not been investigated at a genome-wide level with a multi-layered Bayesian framework. The obtained findings demonstrate that our method uncovers interesting transcriptional interactions, relating the effective regulatory impact of TFs on gene expression with topological network properties.

The validation results showed the robustness of the proposed approach to noisy data, such as omics sources, and to prior knowledge with limited reliability, here given as increasing fractions of false transcriptional interactions in the priors.

The developed method is divided into well-defined steps in order to be applicable to other case studies, e.g. adapting the iterative procedure for finding the initial DAG to

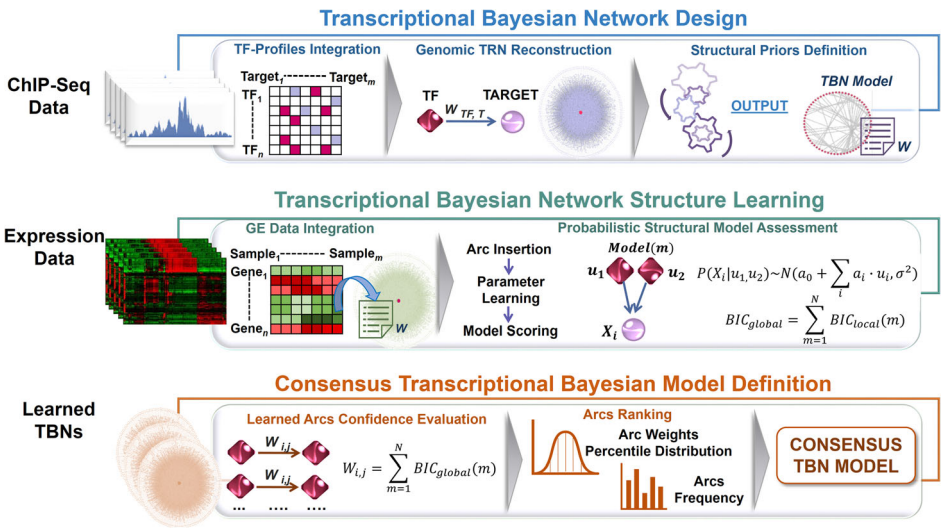

Fig. 6 Overview of the Data Fusion approach. The procedure is structured on three main parts defined as (i) the reconstruction of the transcriptional network and its Bayesian modelling (upper panel), (ii) the model integration with transcriptomics data for its probabilistic assessment through the hybrid structure learning algorithm (central panel), and (iii) the consensus definition process aimed at identifying high confidence interactions among all learned transcriptional dependencies

weighted or unweighted arcs while the TRN reconstruction phase, or the score threshold as ending criterion of the learning algorithm, varying the sampling edges set size, or the confidence threshold for the "consensus" approach.

This makes the developed data fusion approach an ideal framework for integrating potentially noisy complementary data, and a data-driven platform for transcriptional regulatory network inference.

Methods

In this section, we present our data fusion approach, whose main steps are depicted in Fig. 6. Its multi-step procedures rely on –omics data integration, exploiting data complementary to jointly investigate transcriptional regulations at genomic level under a unified Bayesian framework. The proposed method firstly uses ChIP-Seq data to reconstruct a genomic regulatory skeleton, from which the implemented hybrid algorithm draws structural priors and evidence from integrated expression data to probabilistically assess the transcriptional network structure. The final evaluation on transcriptional relations obtained from all the learned models ensures that only consistent dependencies will characterize the final transcriptional Bayesian model.

The following sections provide a description of the aforementioned Bayesian framework in which our developed hybrid algorithm lies, of its structural constraints definition and of the learning strategy.

Bayesian modeling framework

A Bayesian network model is a graphical representation of the joint probability distribution of a set of random variables $X = \{X_1, \ldots, X_n\}$. The encoding of this probability distribution is defined by a network structure S and a set of model parameters Θ, which

describe the probability distribution of model's variables [95]. Model structure S is represented as a directed acyclic graph (DAG), whose vertices (or nodes) are the random variables, and whose conditional dependencies are described by directed edges. In particular, each variable is assumed to be independent of its non-descendants given its set of parents, denoted as $\boldsymbol{pa}(X_n)$. Under this Markov assumption, the joint probability distribution of all nodes of the model is given as

$$P(X) = \prod_{i=1}^{n} P(X_i|\boldsymbol{pa}(X_i) = \prod_{i=1}^{n} \theta_{X_{i|pa(X_i)}} \quad (1)$$

where each variable X_i is described by a set of parameters (θ_i) which defines the variable distribution conditional on its parents.

Within our transcriptional network context, a BN model represents the regulatory relationships among transcription factors (TFs) and from TFs to genes. An edge denotes an observed transcriptional regulation relationship between the considered nodes. All the variables of the model are real valued, and the joint distribution is assumed to be a multivariate Gaussian [96]. The conditional density of each variable X_i given its parents $\boldsymbol{pa}(X_i) = \{U_1, \ldots, U_k\}$, can be represented as a linear Gaussian model Eq. (2).

$$P(X_i|u_1, \ldots, u_k) \sim N\left(a_0 + \sum\nolimits_i a_i \cdot u_i, \sigma^2\right) \quad (2)$$

That is, X_i is normally distributed around a mean that depends linearly on the values of $\boldsymbol{pa}(X_i)$; the variance of this Normal distribution is independent of the parents' values. In this representation $\theta_{X_i|\{u_1,\ldots,u_k\}} = \langle, a_0, \ldots, a_k, \sigma\rangle$.

Given a dataset $D = \{D_1, \ldots, D_n\}$ where D is an instantiation of all the variables in X, learning BN structure from D corresponds to finding a model structure that best fits the observed data.

Finding the optimal BN represents an NP-hard (nondeterministic polynomial-time) problem that has been approached with constraints-based and score-based structure learning methods [97]. The former strategy exploits conditional independence tests to construct a partially oriented graph, retaining or rejecting candidate edges; the latter uses a scoring function to assign a network score reflecting its goodness of fit, which the algorithm then attempts to maximize. Both strategies scale to large networks poorly, because the number of possible graph structures or tests rises exponentially as the size of the network increases.

Hybrid algorithms, which are another class of structure learning methods, combine the characteristics of both aforementioned approaches to maximize their advantages. Typically, they start with a constraint-based search to find the skeleton of the network and then employ a score-based scheme to identify a high-scoring network structure. Our algorithm follows this search and score paradigm, optimizing the learning in order to manage genome-scale networks.

The hybrid Bayesian network structure learning algorithm

The learning procedure is preceded by two fundamental phases of *omics*-data fusion in order to gather informative priors from genome-wide binding data and to define the parameters space from integrated gene expression profiles.

The starting point of this framework is a Transcriptional Regulatory Network (TRN) obtained from the integration of TF binding data. It is defined as a directed graph $TRN = \langle V, E \rangle$, where V is the set of TFs and genes vertices, and E is a set of ordered pairs of edges composed in turn by two subsets, describing the regulatory interactions between TFs (E_1) and from TFs to genes (E_2).

$$TRN = \langle V, E \rangle where \quad E = \begin{Bmatrix} V = \{TF_1, ..., TF_i, G_1, ..., G_k\} \\ E_1 = \{(TF_1, TF_2), ..., (TF_i, TF_j)\} \forall_i \forall_j, i \neq j \\ E_2 = \{(TF_1, G_1), ..., (TF_i, G_k)\} \end{Bmatrix}$$

First, the TRN is converted into a Bayesian model, defining its structural constraints. The obtained transcriptional BN is then integrated with an expression data compendium, to achieve a fully observable network whose structure and parameters will be learned by our algorithm.

Bayesian model definition

The Bayesian learning procedure starts from an initial directed acyclic graph (DAG), which does not allow loops. Transcriptional networks are characterized by many loops of regulation, a typical characteristic of the dynamic crosstalk among TFs, through which they modulate both the expression and the activity of other TFs [98]. Since the TRN is enriched in these regulatory patterns, to match the acyclicity Bayesian constraint, the proposed approach exploits the property of TRNs, whose regulations can be divided in turn in two subsets, defining the interactions between TFs and between TFs and genes, respectively.

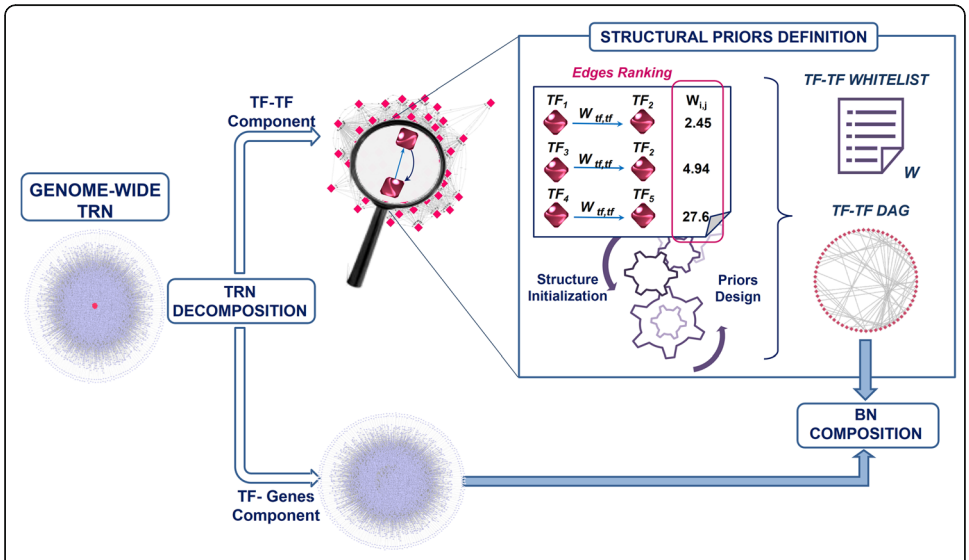

Fig. 7 Transcriptional BN definition. Decomposition of a genome-wide Transcriptional Regulatory Network (TRN) allows to operate on the TF-TF Component, characterized by many regulatory loops, e.g. feedback loop (as shown in the magnifying glass) to initialize the BN structure model and its structural constraints. The obtained DAG is then combined with the TF-Genes Component to define a genome-scale transcriptional BN

$$TF_C = \langle V_{TF_C}, E_{TF_C} \rangle \quad where \quad \left\{ E_{TF_C} = \begin{matrix} TF\text{–}TF_{Component} \\ V_{TF_C} = \{TF_1, ..., TF_i\} \\ \{(TF_1, TF_2), ..., (TF_i, TF_j)\} \quad \forall_i \forall_j, i \neq j \end{matrix} \right\}$$

$$TF_{G_C} = \langle V_{TF_{G_C}}, E_{TF_{G_C}} \rangle where \quad \left\{ \begin{matrix} TF\text{–}Genes_{Component} \\ V_{TF_{G_C}} = \{TF_1, ..., TF_i; G_1, ..., G_k\} \\ E_{TF_{G_C}} = \{(TF_1, G_1), ..., (TF_i, G_k)\} \end{matrix} \right\}$$

As shown in Fig. 7, the TRN is decomposed into its fundamental parts, in order to be transformed into a BN: a TF-TF Component, consisting in TF-TF edges (which may contain regulatory loops), and a TF-Genes Component, consisting of edges from TFs to genes. The former then undergoes an iterative process aimed at initializing the model structure and defining the priors of the algorithm, while removing loops. Within this scheme, the procedure first evaluates the edges between TFs, ranking and sorting them in decreasing order if they are weighted, otherwise it shuffles all the arcs and assigns an equal weight to them. The process then removes one arc at a time, starting from edges with lower weight (if arcs are weighted), otherwise randomly extracting an arc, and checking, at every iteration, if the TF-TF Component is still full connected. The procedure ends when at least a minimal connected DAG is found. All the TF-TF edges excluded from this structure initialization constituted an *arcs whitelist* (W), which will represent the search space of the algorithm.

The resulting DAG is joined with the TF-Genes Component, to obtain again a genomic transcriptional network, but designed as a Bayesian model (TBN). As a second step, the TBN is then integrated with gene expression (GE) data, as shown in Fig. 8 below, in order to obtain a fully observable BN. This transcriptomic data source is also used to calculate the correlation between TFs included in the whitelist, since this measure will be exploited as a sampling probability for each whitelisted arc extracted by the hybrid algorithm and evaluated in the TBN.

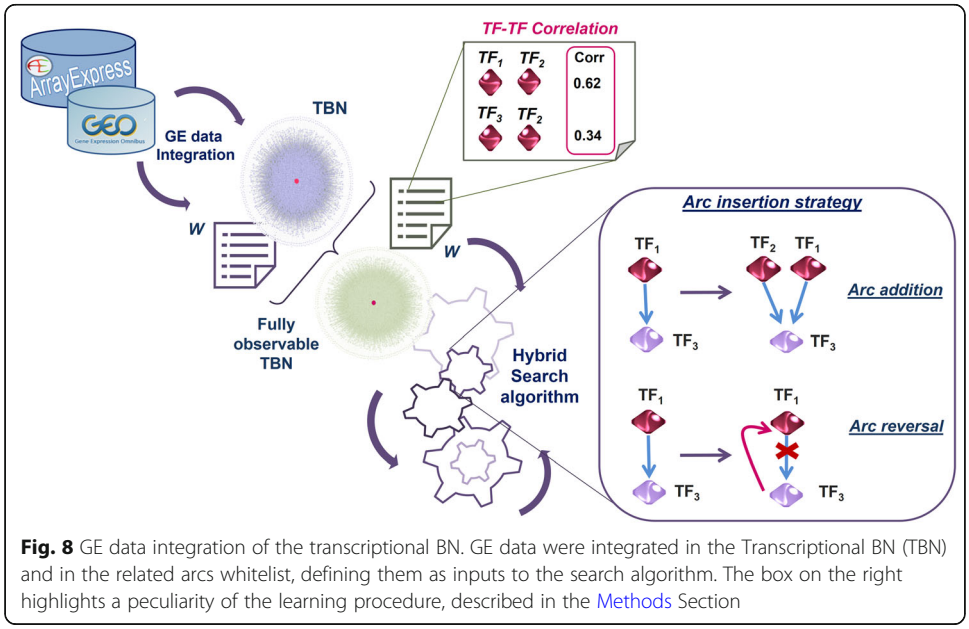

Fig. 8 GE data integration of the transcriptional BN. GE data were integrated in the Transcriptional BN (TBN) and in the related arcs whitelist, defining them as inputs to the search algorithm. The box on the right highlights a peculiarity of the learning procedure, described in the Methods Section

The structure learning strategy

The hybrid algorithm developed in the current study proposes a heuristic search over the space of all possible structures derived from the whitelist, which encloses the informative structure priors concerning the TF-TF relations. All steps of the learning process are referenced below and are presented in Fig. 9, in which the pseudo code of the algorithm, implemented in Matlab language, is reported.

The learning procedure is designed for parallel computing and it has been tested both on a single multi-core machine (P7 CPU 4.0 GHz, 32 GB RAM) and on a high-performance computing environment, the University of Florida HiPerGator 2.0 cluster (52,000 cores, with 4GB RAM per core).

Hybrid Structure Learning Algorithm

1: **Procedure** Hybrid Struct. Learning (*TBN*, *D*, *W*)
Inputs: *TBN*, transcriptional BN model containing the genomic variables X_i, *i=1, ..., n*
D, Gene Expression Dataset representing the evidence for all X_i
W, arcs whitelist
Output: TBN*, learned TBN

2: $BICglobal_{old} \leftarrow BIC_{global}$ *estimation on TBN*
3: cnt=0
4: **while** $(BICglobal_{new} - BICglobal_{old})/BICglobal_{new} >$ threshold
%Phase I: Whitelisted arcs evaluation
5: extract $\underline{w} \subset W$, *where* $(TF_i, TF_j) \in W$
6: $B = \emptyset$
7: **for all** arc ∈ $\underline{w}$ **do**
7.1**: insert the arc in the *TBN*
7.2: learn model parameters from *D*
7.3: B(*arc*) ← BIC_{local} *estimation*
7.4: **end for**
%Phase II: Update the *TBN* model
8: $(TF_i, TF_j)_{best} = \max(B)$
9: $tBN^* = tBN \cup (TF_i, TF_j)_{best}$
10: delete $(TF_i, TF_j)_{best}$ from *W*
11: $BICglobal_{new} \leftarrow BIC_{global}$ *update on TBN**
%Phase III: Escape from Local Minimum
13: **if** $(BICglobal_{new} > BICglobal_{old})$ **then**
13.1: **if** cnt < 10 **then**
13.2: extract $\underline{w_i} \subset W$ *where size*($\underline{w_i}$)=2**size*($\underline{w}$)
13.3: cnt= cnt+1
13.4: $\underline{w_i} = \underline{w}$
13.5: **continue** %go to step 6
13.6: **elseif** cnt == 10
13.7: **end procedure**
13.8: **end if**
13.9: **else**
14: cnt =0
14.1: $BICglobal_{old} = BICglobal_{new}$
14.2: **continue** %go to step 5
14.3: **end if**
15: **end while**

**Possible arc operations $(\forall_i \forall_j, i \neq j)$:
1. *arc addition:* $TBN' \leftarrow TBN \cup (TF_i, TF_j)$
2. *arc reversal:* $TBN' \leftarrow TBN \setminus (TF_i, TF_j) \cup (TF_j, TF_i)$

Fig. 9 Pseudo code of the hybrid algorithm for learning a transcriptional BN structure

At each iteration, the algorithm randomly draws a set of arcs ($\underline{w}$) (i.e. one hundred) from the whitelist to test them in the transcriptional BN (step 5). This sampling process is guided by correlation, which is exploited as an extraction probability associated to each whitelisted edge. The algorithm adds every sampled arc, one by one, to the BN model, learns the model parameters from gene expression (GE) data, and evaluates the newly obtained BN using the *Bayesian Information Criterion* (BIC) scoring metric (steps 6–7). Exploiting the decomposability property of this scoring function [99], the score of a network (G) given the data (D) can be written as the sum of scores of individual variables, where the score of each variable is calculated considering only the variable and its parents, as reported in Eq. (3).

$$Score(G \mid D) = \sum_{i=1}^{n} FamScore\left(X_i \middle| pa^G(X_i) \middle| D\right) \tag{3}$$

In particular, since the TBN distribution is assumed to be jointly multivariate Gaussian, the BIC score can be expressed in terms of the residual sum of squares (RSS)

$$BIC = nlog(RSS/n) + klog(n) \tag{4}$$

where n is the number of observations (the GE dataset size), and k is the number of parameters in the model.

Determining the optimal structure G* from a finite set of model structures requires selecting a model that maximizes Eq. (3), as

$$G^* = \underset{G}{argmax}\, Score(G \mid D) \tag{5}$$

Assuming that each variable of the BN model is linearly dependent upon its continuous parents, we consider the BN as the sum of all local models.

Thus, we modeled two BIC scores, a *local* one that is used to assess the local improvement in the network before and after a whitelisted arc addition, and a *global* one which represents the BN score computed as the sum of all BIC scores from local models, as shown in Eq. (6) and Eq. (7), respectively.

$$BIC_{local} = \Delta BIC = BIC_{old} - BIC_{new} = nlog\left(\frac{RSS_{old}}{RSS_{new}}\right) - \Delta k * \log(n) \tag{6}$$

$$BIC_{global} = \sum_{i=1}^{m} BIC \tag{7}$$

where m denotes the number of local models composing the transcriptional BN.

The second term in Eq. (6) is a penalty term that takes into account the edge changes; since many of the whitelisted arcs come from TRN regulatory loops, the algorithm can add a new arc between two nodes ($\Delta k = 1$) or reverse the directionality of an existing BN arc ($\Delta k = 0$), as illustrated in the box of Fig. 8.

Our learning scheme is designed for parallel computing, allowing to test all the arcs extracted from the whitelist simultaneously. The algorithm evaluates all the computed BIC_{local}, selects as best model the solution that maximizes Eq. (6) (step 8), and then includes the corresponding arc into the model (step 9). The BN structure and its new score ($BICglobal_{new}$) are updated, and the process moves forward (steps 9–11) until the

stop criterion (defined at the step 4) is met. The algorithm ends its iterations when the new model score does not improve more than a fixed threshold compared to the score of the previous network ($BICglobal_{old}$). This threshold is estimated as $10 * (BIC_{global})^{-1}$ calculated on the initial TRN, given as input of the hybrid algorithm.

The learning procedure also provides a strategy to prevent the search phase from getting trapped in a local optimal network (steps 13–14). When the stop condition is verified, the algorithm tries to move out of this potential local minimum for 10 consecutive times, combining an increased arc sampling size ($\underline{w_i}$) (for instance, if dim($\underline{w}$) = 100, the dimension of $\underline{w_i}$ is doubled as dim (w_i) = 200) with a correspondingly augmented proportion of arcs to test. We save the BIC_{global} computed on the model before starting this procedure as $BICglobal_{old}$; if in any of these steps the BIC_{global} for the new solution ($BICglobal_{new}$) is not better than the $BICglobal_{old}$, at the last iteration, the algorithm stops, otherwise it accepts the new model structure and continues the search process. At the end of each algorithm run, the heuristic procedure returns as output a learned transcriptional BN.

Consensus transcriptional BN definition

In order to obtain a final robust transcriptional model including the regulatory dependencies consistently found across all the learned TBNs, as pointed out in [100], we defined a "consensus approach" to identify structural consistencies among all the models gained from several runs of the learning algorithm. We determined a confidence threshold, as the minimum degree of confidence for an edge to be significantly accepted in a final Consensus Bayesian Network. For each learned TF-TF edge (e_{ij}), we compute its strength (w_{ij}) considering the BN models (m), in which this transcriptional relationship appeared, and their related scores (BIC_{global}).

$$w_{ij} = \sum_{m=1}^{n} \left(BIC_{global}(m) \right) \tag{8}$$

Edges with high confidence (significant edges present in more than half of the learned network structures, and in the best scenario, present in all the network structures) are strongly weighted and more likely to be included in the final consensus model.

The percentile distribution of the edge weights combined with the edge frequencies were used to rank all the considered arcs and to assess a confidence threshold, ensuring that the obtained transcriptional consensus BN is acyclic and fully connected.

Competing methods

The peculiarities of our novel approach optimized for learning large-scale transcriptional BNs make finding other similar methods difficult, especially in the class of hybrid BN learning algorithms, which exploit prior knowledge, directed regulations, transcriptomics and epigenomics data. To evaluate the performance of our method, we selected the SAGA (Simulated Annealing with a Greedy Algorithm) algorithm [46], the only approach with some common grounds with our strategy, and ARACNe (Algorithm for the Reconstruction of Accurate Cellular Networks), which is the most widely used technique for regulatory network reconstruction from gene expression data [101]. Another tool for learning BN structures and estimating their parameters is the R

package bnlearn [102], which however cannot be used for our purposes. Bnlearn implements the Max-Min Hill-Climbing as hybrid algorithm, but to reconstruct the network from GE data, it forces all the arcs of the structural prior, specified as a DAG within a *whitelist*, to be included in the final network, preventing the addition of any other extra transcriptional relationship. This constraint makes this approach not appropriate to handle transcriptional networks, and in particular our type of whitelist given the presence of regulatory loops.

SAGA is a hybrid Bayesian learning algorithm, implemented in the BANJO (Bayesian Network Inference with Java Objects) software [103], which combines Simulated Annealing with a greedy search, using Bayesian Dirichlet equivalence as a scoring metric to evaluate the generated network. It allows arc addition and reversal, and the possibility to specify a structural prior as well as a list of forbidden arcs that must not be added (blacklist) to the model. This method does not exploit an arcs whitelist strategy, but it infers the network structure from discretized gene expression data. BANJO ends its search when one of the termination criteria are met (i.e. fixed number of explored networks, search threshold time, maximum number of restarts reached), and returns as output the learned network with the best score.

ARACNe is an information-theoretic based approach that implements Data Processing Inequality on each connected gene triplet from the GE dataset, to remove the least significant edge in mutual-information (MI) relevant networks. For our test, we used the last version of this algorithm, ARACNe-AP [47], that works on reconstructing transcriptional networks taking as inputs a GE dataset and a predefined list of regulators (TFs). Its strategy consists of computing MI only for every TF/target pair, and reconstructing MI networks from bootstrapped GE samples. A consensus network is then generated from the significant edges detected across all bootstrap runs.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12859-020-3510-1.

Additional file 1.

Abbreviations

ATAC-Seq: Assay for Transposase-accessible Chromatin Sequencing; BIC: Bayesian Information Criterion; BN: Bayesian Network; ChIP-Seq: Chromatin ImmunoPrecipitation followed by Sequencing; DNase-Seq: DNase I hypersensitivity sites Sequencing; CML: Chronic Myeloid Leukemia; DAG: Directed Acyclic Graph; GE: Gene Expression; HM: Hierarchy metric; HSA: Hierchical Simulating Annealing algorithm; MR: Master Regulator; TBN: Transcriptional Bayesian Network; TF: Transcription Factor; TRN: Transcriptional Regulatory Network

Acknowledgements

We would like to acknowledge the University of Florida Research Computing (URL: http://researchcomputing.ufl.edu) for providing computational resources and support that have contributed to the research results reported in this publication.

Authors' contributions

ES and RB conceived the study and designed the method. ES implemented it and performed data analysis. AD and AR helped with the algorithm implementation. FV supported the initial network modeling phase. ES wrote the manuscript and RB reviewed it. All the authors read the manuscript and approved the final version.

Funding

The work is part of the project "Rete Ematologica Lombarda (REL) biotechnology cluster for the implementation of genomic analysis and the development of innovative treatments in hematological malignancies", jointly funded by the Fondazione Cariplo and the Regione Lombardia (project ID 2013–0387). The funding bodies did not play any roles in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

All the datasets used and analyzed during the current study are publicy available in the following repositories.

• Chronic Myeloid leukemia datasets:
ENCODE ChIP-Seq K562 raw data https://www.encodeproject.org/search/?type=Experiment&status=released&biosample_ontology.term_name=K562&assay_title=ChIP-seq&biosample_ontology.classification=cell+line&files.file_type=bam&assay_title=TF+ChIP-seq
Chronic myeloid leukemia Gene Expression data:
GEO accessions: GSE1315, GSE47927, and GSE24739; ArrayExpress accessions: E-MTAB-2581, E-MEXP-480.
• Yeast transcriptional datasets:
SGD Project transcriptional interactions data https://yeastgenome.org/
YEASTRACT transcriptional interactions data http://yeastract.com/download.php
• Hematopoietic transcriptional networks:
The K562 DNaseI derived network is available in the supplementary data of Neph et al study [9].
The K562 TFs co-association network is available at http://encodenets.gersteinlab.org/
A Matlab implementation of our algorithm is available at https://github.com/esauta/TBN_learning

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Via Ferrata 5, 27100 Pavia, Italy. [2]Center for Biomedical Informatics and Biostatistics, Dept. of Medicine, The University of Arizona Health Sciences, 1230 Cherry Ave, Tucson, AZ 85719, USA. [3]Bioinformatics Core, Interdisciplinary Center for Biotechnology Research, University of Florida, Gainesville, FL 32610, USA.

Received: 9 January 2020 Accepted: 22 April 2020
Published online: 29 May 2020

References
1. Bradner JE, Hnisz D, Young RA. Transcriptional addiction in Cancer. Cell. 2017;168:629–43. https://doi.org/10.1016/j.cell.2016.12.013.
2. Huang S, Chaudhary K, Garmire LX. More is better: recent Progress in multi-Omics data integration methods. Front Genet. 2017;8. https://doi.org/10.3389/fgene.2017.00084.
3. Gallo Cantafio ME, Grillone K, Caracciolo D, Scionti F, Arbitrio M, Barbieri V, et al. From single level analysis to multi-Omics integrative approaches: A powerful strategy towards the precision oncology. High-Throughput. 2018;7. https://doi.org/10.3390/ht7040033.
4. Lefebvre C, Rieckhof G, Califano A. Reverse-engineering human regulatory networks. Wiley Interdiscip Rev Syst Biol Med. 2012;4:311–25. https://doi.org/10.1002/wsbm.1159.
5. He B, Tan K. Understanding transcriptional regulatory networks using computational models. Curr Opin Genet Dev. 2016;37:101–8. https://doi.org/10.1016/j.gde.2016.02.002.
6. Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan K-K, Cheng C, et al. Architecture of the human regulatory network derived from ENCODE data. Nature. 2012;489:91–100. https://doi.org/10.1038/nature11245.
7. Doane AS, Elemento O. Regulatory elements in molecular networks. Wiley Interdiscip Rev Syst Biol Med. 2017;9. https://doi.org/10.1002/wsbm.1374.
8. Boyle AP, Song L, Lee B-K, London D, Keefe D, Birney E, et al. High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. Genome Res. 2011;21:456–64. https://doi.org/10.1101/gr.112656.110.
9. Neph S, Stergachis AB, Reynolds A, Sandstrom R, Borenstein E, Stamatoyannopoulos JA. Circuitry and dynamics of human transcription factor regulatory networks. Cell. 2012;150:1274–86. https://doi.org/10.1016/j.cell.2012.04.040.
10. Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, Pritchard JK. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome Res. 2011;21:447–55. https://doi.org/10.1101/gr.112623.110.
11. Gusmao EG, Allhoff M, Zenke M, Costa IG. Analysis of computational footprinting methods for DNase sequencing experiments. Nat Methods. 2016;13:303–9. https://doi.org/10.1038/nmeth.3772.
12. Inukai S, Kock KH, Bulyk ML. Transcription factor–DNA binding: beyond binding site motifs. Curr Opin Genet Dev. 2017;43:110–9. https://doi.org/10.1016/j.gde.2017.02.007.
13. Siggers T, Gordân R. Protein–DNA binding: complexities and multi-protein codes. Nucleic Acids Res. 2014;42:2099–111. https://doi.org/10.1093/nar/gkt1112.
14. Liu Z-P. Reverse engineering of genome-wide gene regulatory networks from gene expression data. Curr Genomics. 2015;16:3–22. https://doi.org/10.2174/1389202915666141110210634.
15. Bansal M, Belcastro V, Ambesi-Impiombato A, di Bernardo D. How to infer gene networks from expression profiles. Mol Syst Biol. 2007;3:78. https://doi.org/10.1038/msb4100120.
16. Chai LE, Loh SK, Low ST, Mohamad MS, Deris S, Zakaria Z. A review on the computational approaches for gene regulatory network construction. Comput Biol Med. 2014;48:55–65. https://doi.org/10.1016/j.compbiomed.2014.02.011.
17. Yaghoobi H, Haghipour S, Hamzeiy H, Asadi-Khiavi M. A review of modeling techniques for genetic regulatory networks. J Med Signals Sens. 2012;2:61–70.

18. Murrugarra D, Aguilar B. Chapter 5 - Modeling the Stochastic Nature of Gene Regulation With Boolean Networks. In: Robeva R, Macauley M, editors. Algebr. Comb. Comput. Biol., Academic Press; 2019. p. 147–73. https://doi.org/10.1016/B978-0-12-814066-6.00005-2.
19. Hecker M, Lambeck S, Toepfer S, van Someren E, Guthke R. Gene regulatory network inference: data integration in dynamic models-a review. Biosystems. 2009;96:86–103. https://doi.org/10.1016/j.biosystems.2008.12.004.
20. Hartemink AJ, Gifford DK, Jaakkola TS, Young RA. Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. Pac Symp Biocomput Pac Symp Biocomput. 2001:422–33.
21. Friedman N. Inferring cellular networks using probabilistic graphical models. Science. 2004;303:799–805. https://doi.org/10.1126/science.1094068.
22. Murphy K, Mian S, others. Modelling gene expression data using dynamic Bayesian networks. Berkeley: Technical report, Computer Science Division, University of California; 1999.
23. Penfold CA, Wild DL. How to infer gene networks from expression profiles, revisited. Interface Focus. 2011;1:857–70. https://doi.org/10.1098/rsfs.2011.0053.
24. Emmert-Streib F, Glazko GV, Altay G, de Matos Simoes R. Statistical inference and reverse engineering of gene regulatory networks from observational expression data. Front Genet. 2012;3. https://doi.org/10.3389/fgene.2012.00008.
25. Ghanbari M, Lasserre J, Vingron M. Reconstruction of gene networks using prior knowledge. BMC Syst Biol. 2015;9. https://doi.org/10.1186/s12918-015-0233-4.
26. Thomas SA, Jin Y. Reconstructing biological gene regulatory networks: where optimization meets big data. Evol Intell. 2014;7:29–47. https://doi.org/10.1007/s12065-013-0098-7.
27. Chickering DM. Learning Bayesian Networks is NP-Complete. Learn. New York: Data, Springer; 1996. p. 121–30. https://doi.org/10.1007/978-1-4612-2404-4_12.
28. Chickering DM, Heckerman D, Meek C. Large-sample learning of Bayesian networks is NP-hard. J Mach Learn Res. 2004;5:1287–330.
29. Imoto S, Higuchi T, Goto T, Tashiro K, Kuhara S, Miyano S. Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks. J Bioinforma Comput Biol. 2004;2:77–98.
30. Werhli AV, Husmeier D. Reconstructing gene regulatory networks with Bayesian networks by combining expression data with multiple sources of prior knowledge. Stat Appl Genet Mol Biol. 2007;6. https://doi.org/10.2202/1544-6115.1282.
31. Hartemink AJ, Gifford DK, Jaakkola TS, Young RA. Combining location and expression data for principled discovery of genetic regulatory network models. Pac Symp Biocomput. 2002;7:437–49.
32. Greenfield A, Hafemeister C, Bonneau R. Robust data-driven incorporation of prior knowledge into the inference of dynamic regulatory networks. Bioinforma Oxf Engl. 2013;29:1060–7. https://doi.org/10.1093/bioinformatics/btt099.
33. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57–74. https://doi.org/10.1038/nature11247.
34. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9:R137. https://doi.org/10.1186/gb-2008-9-9-r137.
35. Sikora-Wohlfeld W, Ackermann M, Christodoulou EG, Singaravelu K, Beyer A. Assessing computational methods for transcription factor target gene identification based on ChIP-seq data. PLoS Comput Biol. 2013;9:e1003342. https://doi.org/10.1371/journal.pcbi.1003342.
36. Kohlmann A, Kipps TJ, Rassenti LZ, Downing JR, Shurtleff SA, Mills KI, et al. An international standardization programme towards the application of gene expression profiling in routine leukaemia diagnostics: the microarray innovations in LEukemia study prephase. Br J Haematol. 2008;142:802–7. https://doi.org/10.1111/j.1365-2141.2008.07261.x.
37. Personalized synthetic lethality induced by targeting RAD52 in leukemias identified by gene mutation and expression profile | Blood Journal n.d. http://www.bloodjournal.org/content/122/7/1293?sso-checked=true (accessed 14 Oct 2017).
38. Affer M, Dao S, Liu C, Olshen AB, Mo Q, Viale A, et al. Gene expression differences between enriched Normal and chronic Myelogenous leukemia quiescent stem/progenitor cells and correlations with biological abnormalities. J Oncol. 2011. https://doi.org/10.1155/2011/798592.
39. Abraham SA, Hopcroft LE, Carrick E, Drotar ME, Dunn K, Williamson AJ, et al. Dual targeting of p53 and c-Myc selectively eliminates leukaemic stem cells. Nature. 2016;534:341–6. https://doi.org/10.1038/nature18288.
40. Zheng C, Li L, Haak M, Brors B, Frank O, Giehl M, et al. Gene expression profiling of CD34+ cells identifies a molecular signature of chronic myeloid leukemia blast crisis. Leukemia. 2006;20:1028–34. https://doi.org/10.1038/sj.leu.2404227.
41. Exploration, normalization, and summaries of high density oligonucleotide array probe level data | Biostatistics | Oxford Academic n.d. doi:https://doi.org/10.1093/biostatistics/4.2.249 (accessed 9 Oct 2017).
42. Beineke LW, Oellermann OR, Pippert RE. The average connectivity of a graph. Discret Math. 2002;252:31–45. https://doi.org/10.1016/S0012-365X(01)00180-7.
43. Teixeira MC, Monteiro PT, Guerreiro JF, Gonçalves JP, Mira NP, Santos D, et al. The YEASTRACT database: an upgraded information system for the analysis of gene and genomic transcription regulation in Saccharomyces cerevisiae. Nucleic Acids Res. 2014;42:D161–6. https://doi.org/10.1093/nar/gkt1015.
44. Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, et al. Saccharomyces genome database: the genomics resource of budding yeast. Nucleic Acids Res. 2012;40:D700–5. https://doi.org/10.1093/nar/gkr1029.
45. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, et al. Comprehensive identification of cell cycle–regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell. 1998;9:3273–97.
46. Adabor ES, Acquaah-Mensah GK, Oduro FT. SAGA: A hybrid search algorithm for Bayesian network structure learning of transcriptional regulatory networks. J Biomed Inform. 2015;53:27–35. https://doi.org/10.1016/j.jbi.2014.08.010.
47. Lachmann A, Giorgi FM, Lopez G, Califano A. ARACNe-AP: gene network reverse engineering through adaptive partitioning inference of mutual information. Bioinformatics. 2016;32:2233–5. https://doi.org/10.1093/bioinformatics/btw216.
48. Wilson NK, Calero-Nieto FJ, Ferreira R, Göttgens B. Transcriptional regulation of haematopoietic transcription factors. Stem Cell Res Ther. 2011;2:6. https://doi.org/10.1186/scrt47.
49. Göttgens B, Nastos A, Kinston S, Piltz S, Delabesse ECM, Stanley M, et al. Establishing the transcriptional programme for blood: the SCL stem cell enhancer is regulated by a multiprotein complex containing Ets and GATA factors. EMBO J. 2002;21:3039–50. https://doi.org/10.1093/emboj/cdf286.

50. Rodrigues NP, Boyd AS, Fugazza C, May GE, Guo Y, Tipping AJ, et al. GATA-2 regulates granulocyte-macrophage progenitor cell function. Blood. 2008;112:4862–73. https://doi.org/10.1182/blood-2008-01-136564.
51. Guiu J, Shimizu R, D'Altri T, Fraser ST, Hatakeyama J, Bresnick EH, et al. Hes repressors are essential regulators of hematopoietic stem cell development downstream of notch signaling. J Exp Med. 2013;210:71–84. https://doi.org/10.1084/jem.20120993.
52. Wlodarski MW, Collin M, Horwitz MS. GATA2 deficiency and related myeloid neoplasms. Semin Hematol. 2017;54:81–6. https://doi.org/10.1053/j.seminhematol.2017.05.002.
53. Guo Y, Fu X, Huo B, Wang Y, Sun J, Meng L, et al. GATA2 regulates GATA1 expression through LSD1-mediated histone modification. Am J Transl Res. 2016;8:2265–74.
54. Tijssen MR, Cvejic A, Joshi A, Hannah RL, Ferreira R, Forrai A, et al. Genome-wide analysis of simultaneous GATA1/2, RUNX1, FLI1, and SCL binding in megakaryocytes identifies hematopoietic regulators. Dev Cell. 2011;20:597–609. https://doi.org/10.1016/j.devcel.2011.04.008.
55. Sanda T, Lawton LN, Barrasa MI, Fan ZP, Kohlhammer H, Gutierrez A, et al. Core transcriptional regulatory circuit controlled by the TAL1 complex in human T cell acute lymphoblastic leukemia. Cancer Cell. 2012;22:209–21. https://doi.org/10.1016/j.ccr.2012.06.007.
56. Göttgens B, Broccardo C, Sanchez M-J, Deveaux S, Murphy G, Göthert JR, et al. The scl +18/19 stem cell enhancer is not required for hematopoiesis: identification of a 5′ Bifunctional hematopoietic-endothelial enhancer bound by Fli-1 and Elf-1. Mol Cell Biol. 2004;24:1870–83. https://doi.org/10.1128/MCB.24.5.1870-1883.2004.
57. Swiers G, Patient R, Loose M. Genetic regulatory networks programming hematopoietic stem cells and erythroid lineage specification. Dev Biol. 2006;294:525–40. https://doi.org/10.1016/j.ydbio.2006.02.051.
58. Wu W, Cheng Y, Keller CA, Ernst J, Kumar SA, Mishra T, et al. Dynamics of the epigenetic landscape during erythroid differentiation after GATA1 restoration. Genome Res. 2011;21:1659–71. https://doi.org/10.1101/gr.125088.111.
59. Wu W, Morrissey CS, Keller CA, Mishra T, Pimkin M, Blobel GA, et al. Dynamic shifts in occupancy by TAL1 are guided by GATA factors and drive large-scale reprogramming of gene expression during hematopoiesis. Genome Res. 2014;24:1945–62. https://doi.org/10.1101/gr.164830.113.
60. Goode DK, Obier N, Vijayabaskar MS, Lie-A-Ling M, Lilly AJ, Hannah R, et al. Dynamic gene regulatory networks drive hematopoietic specification and differentiation. Dev Cell. 2016;36:572–87. https://doi.org/10.1016/j.devcel.2016.01.024.
61. Martowicz ML, Grass JA, Bresnick EH. GATA-1-mediated transcriptional repression yields persistent transcription factor IIB-chromatin complexes. J Biol Chem 2006;281:37345–37352. doi: https://doi.org/10.1074/jbc. M605774200.
62. Vagapova ER, Spirin PV, Lebedev TD, Prassolov VS. The role of TAL1 in hematopoiesis and Leukemogenesis. Acta Nat. 2018;10:15–23.
63. Lord KA, Abdollahi A, Hoffman-Liebermann B, Liebermann DA. Proto-oncogenes of the fos/Jun family of transcription factors are positive regulators of myeloid differentiation. Mol Cell Biol. 1993;13:841–51.
64. Lee S-Y, Yoon J, Lee M-H, Jung SK, Kim DJ, Bode AM, et al. The role of Heterodimeric AP-1 protein comprised of JunD and c-Fos proteins in hematopoiesis. J Biol Chem. 2012;287:31342–8. https://doi.org/10.1074/jbc.M112.387266.
65. Burda P, Laslo P, Stopka T. The role of PU.1 and GATA-1 transcription factors during normal and leukemogenic hematopoiesis. Leukemia. 2010;24:1249–57. https://doi.org/10.1038/leu.2010.104.
66. Cuellar TL, Herzner A-M, Zhang X, Goyal Y, Watanabe C, Friedman BA, et al. silencing of retrotransposons by SETDB1 inhibits the interferon response in acute myeloid leukemia. J Cell Biol. 2017;216:3535–49. https://doi.org/10.1083/jcb.201612160.
67. Robbez-Masson L, Tie CHC, Rowe HM. Cancer cells, on your histone marks, get SETDB1, silence retrotransposons, and go! SETDB1 suppresses antitumor immunity. J Cell Biol. 2017;216:3429–31. https://doi.org/10.1083/jcb.201710068.
68. Marbach D, Costello JC, Küffner R, Vega N, Prill RJ, Camacho DM, et al. Wisdom of crowds for robust gene network inference. Nat Methods. 2012;9:796–804. https://doi.org/10.1038/nmeth.2016.
69. Ntziachristos P, Mullenders J, Trimarchi T, Aifantis I. Mechanisms of epigenetic regulation of leukemia onset and progression. Adv Immunol. 2013;117. https://doi.org/10.1016/B978-0-12-410524-9.00001-3.
70. Chen K, Zhang F, Ding J, Liang Y, Zhan Z, Zhan Y, et al. Histone methyltransferase SETDB1 promotes the progression of colorectal Cancer by inhibiting the expression of TP53. J Cancer. 2017;8:3318–30. https://doi.org/10.7150/jca.20482.
71. Setdb1 maintains hematopoietic stem and progenitor cells by restricting the ectopic activation of nonhematopoietic genes | Blood Journal 2019. http://www.bloodjournal.org/content/128/5/638.long?sso-checked=true (accessed 28 Apr 2019).
72. Karanth AV, Maniswami RR, Prashanth S, Govindaraj H, Padmavathy R, Jegatheesan SK, et al. Emerging role of SETDB1 as a therapeutic target. Expert Opin Ther Targets. 2017;21:319–31. https://doi.org/10.1080/14728222.2017.1279604.
73. Silencing of retrotransposons by SETDB1 inhibits the interferon response in acute myeloid leukemia n.d. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5674883/ (accessed 28 Apr 2019).
74. Hirai H, Yokota A, Tamura A, Sato A, Maekawa T. Non-steady-state hematopoiesis regulated by the C/EBPβ transcription factor. Cancer Sci. 2015;106:797–802. https://doi.org/10.1111/cas.12690.
75. Gonzalez D, Luyten A, Bartholdy B, Zhou Q, Kardosova M, Ebralidze A, et al. ZNF143 protein is an important regulator of the myeloid transcription factor C/EBPα. J Biol Chem. 2017;292:18924–36. https://doi.org/10.1074/jbc.M117.811109.
76. Vaňhara P, Šmarda J. Jun: the master regulator in healthy and cancer cells. J Appl Biomed. 2006;4:163–70.
77. JUN is a key transcriptional regulator of the unfolded protein response in acute myeloid leukemia | Leukemia n.d. https://www.nature.com/articles/leu2016329 (accessed 28 Apr 2019).
78. Leeke B, Marsman J, O'Sullivan JM, Horsfield JA. Cohesin mutations in myeloid malignancies: underlying mechanisms. Exp Hematol Oncol. 2014;3:13. https://doi.org/10.1186/2162-3619-3-13.
79. Hoffman B, Amanullah A, Shafarenko M, Liebermann DA. The proto-oncogene c- myc in hematopoietic development and leukemogenesis. Oncogene. 2002;21:3414–21. https://doi.org/10.1038/sj.onc.1205400.
80. Zuin J, Dixon JR, van der Reijden MIJA, Ye Z, Kolovos P, Brouwer RWW, et al. Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. Proc Natl Acad Sci U S A. 2014;111:996–1001. https://doi.org/10.1073/pnas.1317788111.
81. Chatterjee SS, Biswas M, Boila LD, Banerjee D, Sengupta A. SMARCB1 deficiency integrates epigenetic signals to oncogenic gene expression program maintenance in human acute myeloid leukemia. Mol Cancer Res. 2018;16:791–804. https://doi.org/10.1158/1541-7786.MCR-17-0493.

82. D'Auria F, Pietro RD. Role of CREB protein family members in human Haematological malignancies. Cancer Treat - Conv Innov Approaches. 2013. https://doi.org/10.5772/55368.
83. Li JY, Daniels G, Wang J, Zhang X. TBL1XR1 in physiological and pathological states. Am J Clin Exp Urol. 2015;3:13–23.
84. Yu J, Guo X-L, Bai Y-Y, Yang J-J, Zheng X-Q, Ruan J-C, et al. Genome-wide profiling of lncRNA expression patterns in patients with acute promyelocytic leukemia with differentiation therapy. Oncol Rep. 2018;40:1601–13. https://doi.org/10.3892/or.2018.6521.
85. Takeda S, Chen DY, Westergard TD, Fisher JK, Rubens JA, Sasagawa S, et al. Proteolysis of MLL family proteins is essential for Taspase1-orchestrated cell cycle progression. Genes Dev. 2006;20:2397–409. https://doi.org/10.1101/gad.1449406.
86. Tyagi S, Chabes AL, Wysocka J, Herr W. E2F activation of S phase promoters via association with HCF-1 and the MLL family of histone H3K4 Methyltransferases. Mol Cell. 2007;27:107–19. https://doi.org/10.1016/j.molcel.2007.05.030.
87. Dell'Aversana C, Giorgio C, D'Amato L, Lania G, Matarese F, Saeed S, et al. miR-194-5p/ BCLAF1 deregulation in AML tumorigenesis. Leukemia. 2017;31:2315–25. https://doi.org/10.1038/leu.2017.64.
88. White LS, Soodgupta D, Johnston RL, Magee JA, Bednarski JJ. Bclaf1 promotes maintenance and self-renewal of fetal hematopoietic stem cells. Blood 2018;132:1269ssss. doi: https://doi.org/10.1182/blood-2018-99-114144.
89. Resar L, Chia L, Xian L. Lessons from the crypt: HMGA1—Amping up Wnt for stem cells and tumor progression. Cancer Res. 2018;78:1890. https://doi.org/10.1158/0008-5472.CAN-17-3045.
90. Kang R, Chen R, Zhang Q, Hou W, Wu S, Cao L, et al. HMGB1 in health and disease. Mol Asp Med 2014;0:1–116. doi: https://doi.org/10.1016/j.mam.2014.05.001.
91. Gama-Castro S, Salgado H, Santos-Zavaleta A, Ledezma-Tejeida D, Muñiz-Rascado L, García-Sotelo JS, et al. RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond. Nucleic Acids Res. 2016;44:D133–43. https://doi.org/10.1093/nar/gkv1156.
92. Bernard A, Hartemink AJ. Informative structure priors: joint learning of dynamic regulatory networks from multiple types of data. Pac Symp Biocomput Pac Symp Biocomput. 2005:459–70.
93. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019;47:D607–13. https://doi.org/10.1093/nar/gky1131.
94. Oughtred R, Stark C, Breitkreutz B-J, Rust J, Boucher L, Chang C, et al. The BioGRID interaction database: 2019 update. Nucleic Acids Res. 2019;47:D529–41. https://doi.org/10.1093/nar/gky1079.
95. Jensen FV. Introduction to Bayesian networks. 1st ed. Secaucus: Springer-Verlag New York, Inc.; 1996.
96. Lauritzen SL, Wermuth N. Graphical models for associations between variables, some of which are qualitative and some quantitative. Ann Stat. 1989;17:31–57.
97. Neapolitan RE. Learning Bayesian networks. Pearson Prentice Hall; 2004.
98. Qi Y, Ge H. Modularity and dynamics of cellular networks. PLoS Comput Biol. 2006;2. https://doi.org/10.1371/journal.pcbi.0020174.
99. Heckerman D, Geiger D, Chickering DM. Learning Bayesian networks: the combination of knowledge and statistical data. Mach Learn. 1995;20:197–243.
100. Steele E, Tucker A. Consensus and meta-analysis regulatory networks for combining multiple microarray gene expression datasets. J Biomed Inform. 2008;41:914–26. https://doi.org/10.1016/j.jbi.2008.01.011.
101. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera RD, et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics. 2006;7:S7. https://doi.org/10.1186/1471-2105-7-S1-S7.
102. Learning Bayesian Networks with the bnlearn R Package | Scutari | Journal of Statistical Software 2017. doi: https://doi.org/10.18637/jss.v035.i03.
103. Banjo: Bayesian Network Inference with Java Objects n.d. https://users.cs.duke.edu/~amink/software/banjo/ (accessed 16 Oct 2017).

Publisher's Note