BMC Genomics

PROCEEDINGS Open Access

# A nucleosomal approach to inferring causal relationships of histone modifications

Ngoc Tu Le[1*], Tu Bao Ho[1], Bich Hai Ho[2], Dang Hung Tran[3]

*From* The Twelfth Asia Pacific Bioinformatics Conference (APBC 2014)
Shanghai, China. 17-19 January 2014

## Abstract

**Motivation:** Histone proteins are subject to various posttranslational modifications (PTMs). Elucidating their functional relationships is crucial toward understanding many biological processes. Bayesian network (BN)-based approaches have shown the advantage of revealing causal relationships, rather than simple cooccurrences, of PTMs. Previous works employing BNs to infer causal relationships of PTMs require that all confounders should be included. This assumption, however, is unavoidably violated given the fact that several modifications are often regulated by a common but unobserved factor. An existing non-parametric method can be applied to tackle the problem but the complexity and inflexibility make it impractical.

**Results:** We propose a novel BN-based method to infer causal relationships of histone modifications. First, from the evidence that nucleosome organization *in vivo* significantly affects the activities of PTM regulators working on chromatin substrate, hidden confounders of PTMs are selectively introduced by an information-theoretic criterion. Causal relationships are then inferred from a network model of both PTMs and the derived confounders. Application on human epigenomic data shows the advantage of the proposed method, in terms of computational performance and support from literature. Requiring less strict data assumptions also makes it more practical. Interestingly, analysis of the most significant relationships suggests that the proposed method can recover biologically relevant causal effects between histone modifications, which should be important for future investigation of histone crosstalk.

## Background

Genomes of higher organisms are organized into chromatin, a condensed structure of nucleosome units. Each of these units comprises a short piece of DNA wrapping around an octamer histone, containing two proteins of each type: H2A, H2B, H3, and H4 [1]. The histone protein is subject to various biochemical modifications, a.k.a. post-translational modifications (PTMs), which have been shown to play crucial roles in many cellular processes, such as transcription and replication [2]. Defects of PTMs have also been implicated in determining cell fate and oncogenesis [3,4]. The facts that PTMs may cause combinatorial effects on downstream events, and, by forming stable chromatin domains, properly pass modified states to the next generation [5,6] suggest the existence of "histone codes" [7]. Therefore, revealing genome-wide PTM patterns and related functional implications would help increase our understanding of different DNA-mediated processes. For example, [8] discovered a common modification module of 17 modifications in human, suggesting their critical roles in gene regulation.

Advances in profiling techniques, such as ChIP-Chip and ChIP-Seq, have enabled the availability of genome-scale PTM data [8,9], thus providing an unprecedented opportunity to decipher histone codes and their associated *cis*-regulatory elements. However, it also poses a great requirement for methods to understand such data. Many methods, ranging from clustering- to Hidden Markov Model (HMM)- to Bayesian network (BN)-based, have been developed to identify histone modifications patterns from ChIP-Chip and ChIP-Seq data

[1] Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan
Full list of author information is available at the end of the article

[10-14]. Among them, BN-based approaches may help discover not only the cooccurrence but also the causal relationships of histone modifications [15]. This is especially important to understand histone crosstalk, a phenomenon that often occurs among different PTM events [16-18].

Bayesian network is a family of graphical models representing conditional independence of multiple variables [19]. First introduced to model gene regulatory networks (GRNs) from expression data [20], it has been widely used in reconstructing various biological networks, such as protein-protein interactions, protein signaling networks [21-23]. Likewise, there have been attempts to employ BNs to analyze histone modification data, in which compelled edges of the resulting models were considered causal relationships between PTMs [14,24,25]. Though useful, these works have a significant drawback: they require *causal sufficiency* assumption, i.e., all confounders of PTMs should be observed [26,27]. This assumption, however, is unavoidably violated given the fact that some modifications can be regulated by enzymatic activity of a common but unobserved modifier [2].

Therefore, in order to reveal causal relationships of PTMs the existence of hidden confounders should be taken into account. Basically, there are two choices for network topology containing hidden confounders: overlapping and hierarchical [28]. In the overlapping (Figure 1a), each hidden variable is a parent of several observed variables, and several hidden variables can share a common observed variable as their child. In the hierarchical (Figure 1b), hidden variables form a tree structure, in which each of them is a parent of several other variables (either observed or hidden) and serves to capture the dependencies among its children and between its children and other nodes in the network. Biological evidences have showed that some modifications can be regulated by a common regulator and vice versa [2,29]. Overlapping topology, therefore, is more suitable to describe the relationships between PTMs and their hidden regulators. Thus, the problem of learning network models representing causal relationships of PTMs can be formulated as learning two adjacency matrices, one representing the relationships among observed variables (PTMs), denoted as $X$, and the other representing the relationships between PTMs and their hidden causes, denoted as $Z$, as proposed by [30]. However, their non-parametric approach to learning the models requires strict data assumptions and employs a time-consuming procedure to infer $Z$. These drawbacks make it inflexible and inefficient in practice.

In this work, we propose a novel BN-based method to infer causal relationships of PTMs that accounts for the existence of hidden confounders. First, an information-theoretic criterion is proposed to selectively introduce a *pairwise hidden confounder* (PHC) for each pair of PTMs. *General hidden confounders* (GHCs) are then derived from PHCs. The idea of deriving GHCs from PHCs has been presented in [31] to learn two-layer BNs with hidden variables. Differently, we based our approach on the evidence that chromatin *in vivo* imposes regulatory effects on the activities of PTM regulators. Thus, the criterion is proposed exploiting information about chromatin structure, i.e., nucleosome positioning. Matrix $X$ is separately learned by a BN structure learning method. Compelled edges, i.e., causal relationships, are then derived from a network model of both PTMs and GHCs. Application on human epigenomic data of 38 histone modifications and histone variant H2A.Z, shows that the proposed method outperformed the non-parametric (*Np*) and the traditional one, which does not account for hidden confounders (*noHidden*), in terms of computational performance and literature support. Moreover, analysis of the most significant relationships shows that the proposed method can recover biologically relevant causal effects between histone modifications, such as $H3K27Me3 \rightarrow H3K9Me3$, $H3K4Me3 \rightarrow H2AK5Ac$, $H4K8Ac \rightarrow H2AZ$. This is important for future investigation of histone crosstalk.

## Methods

### Information theory

Mutual information (MI) has been increasingly used in reverse engineering, especially to reconstruct GRNs [32-35]. It is a more general measure compared to correlation in estimating the dependency between two

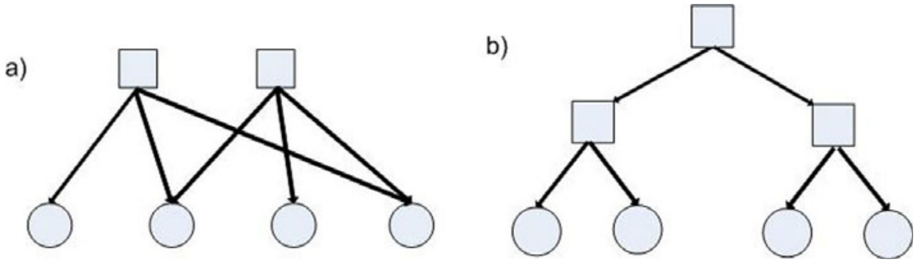

**Figure 1 Overlapping (a) and hierarchical (b) topologies**. The circles represent observed variables, the squares represent hidden ones.

variables. Given two random variables, $x$ and $y$, MI is computed by:

$$MI(x,y) = \int\int p(x,y)\log\frac{p(x,y)}{p(x)p(y)}dxdy \qquad (1)$$

where $p(x, y)$, $p(x)$, and $p(y)$ are joint density function and marginal density functions of $x$ and $y$, respectively.

Likewise, conditional mutual information (CMI) is introduced to measure conditional dependency between two variables given the other(s). CMI of $x$ and $y$ given $\mathbf{z}$ (uni- or multi-variate) is computed by:

$$CMI(x,y|\mathbf{z}) = \int\int\int p(x,y,\mathbf{z})\log\frac{p(x,y|\mathbf{z})}{p(x|\mathbf{z})p(y|\mathbf{z})}dxdyd\mathbf{z} \qquad (2)$$

If $x$, $y$, $\mathbf{z}$ are discrete variables, the integrals are replaced by the sum over all of their values. It is, however, difficult to compute the integrals given the limited number of samples in general cases. Thus, in practice, probability density functions are often approximated by density estimation methods. Given N samples of a variable $\mathbf{x}$, density function can be approximated by:

$$\hat{p}(\mathrm{x}) = \frac{1}{N}\prod_{i=1}^{N}\delta(\mathrm{x}-\mathrm{x}_i, h) \qquad (3)$$

where $\delta(.)$ is the Parzen window function, $\mathbf{x}_i$ is the $i$th sample, and $h$ is the window width. In our work, $\delta(.)$ was chosen as Gaussian function:

$$\delta(\mathbf{z},h) = \exp(-\frac{\mathbf{z}^T\Sigma^{-1}\mathbf{z}}{2h^2})/\left\{(2\pi)^{d/2}h^d|\Sigma|^{1/2}\right\} \qquad (4)$$

where $\mathbf{z} = \mathbf{x} - \mathbf{x}_i$, $d$ is the dimension of $\mathbf{x}$, and $\Sigma$ is the covariance matrix of $\mathbf{z}$. When $d = 1$, equation (3) returns the estimated marginal density. When $d = 2$, it can be used to estimate the joint density function of bivariate variable $(x, y)$. In our work, MI and CMI values were computed using a software package provided by [36].

## Bayesian networks

### *Definition*

A Bayesian network is a directed graph representing conditional independence of multiple variables by a set of conditional probability distributions [19,37]. Joint probability distribution of a variable set $\mathbf{x}$ encoded by the model can be factorized as:

$$p(\mathrm{x}) = \prod_{i=1}^{n}p(x_i|\mathbf{Pa}_i) \qquad (5)$$

in which $p\ (x_i|\mathbf{Pa}_i)$ corresponds to the local probability distribution of variable $x_i$, and $\mathbf{Pa}_i$ are $x_i$'s parents.

### *D-separation property*

In a BN, there are three fundamental local structures, namely *serial*, *diverging*, and *converging* connections (Figure 2). These structures are associated with a set of rules, which is independent of any particular calculus for certainty, to assess how a change of certainty in one variable may change the certainty for other variables in the networks. These rules form d-separation property of a BN. If two variables are d-separated, change in the certainty of one variable has no impact on the other. Two variables are called d-connected if they are not d-separated [19]. Thus, d-separation property can be used as a general assessment of the dependencies among nodes of a BN.

### *BN structure learning*

BN structure can be learned by score-based methods, aiming to identify the structure(s) that "best" describe the data. In this work, BDe score [37,38] with uniform prior was used to measure the fitness of a candidate network. Because it is infeasible to search though all possible structures [39], greedy hill-climbing search combined with simulated annealing algorithm to avoid local maxima was employed.

## Criterion for introducing PHCs

It has been widely shown that the binding of chromatin modifiers, and the large multiprotein complexes in which they reside, to chromatin is greatly affected by chromatin structure, i.e., nucleosome organization [7,40-44]. From this observation, the relationships among two PTMs, their hidden regulator(s), and NucPos can be described by two local causal structures, illustrated in Figure 3. The following results can be easily proved based on d-separation properties:

**Proposition 1** *Consider two PTMs, if each has its own (hidden) regulator, they will be d-separated given evidence on nucleosome positioning.*

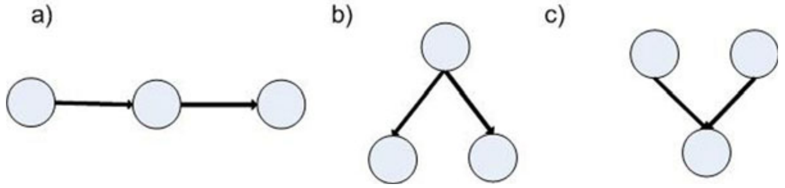

**Figure 2 Fundamental causal structures in BN models: serial (a), diverging (b), and converging (c).**

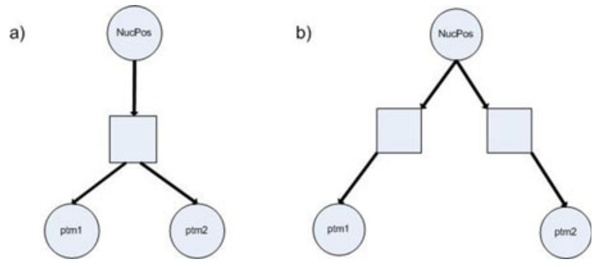

**Figure 3 Causal structures when two PTMs share a hidden confounder (regulator) (a) or not (b)**.

**Proposition 2** *Consider two PTMs, if they share a hidden regulator (in other words, a confounder), their d-separation property does not change upon the availability of nucleosome positioning evidence.*

The results suggest that, given evidence on NucPos, the dependency level between two PTMs would not change if they share a hidden confounder, and would change (becoming “less” dependent) if each has its own (hidden) regulator. Using MI and CMI as the measures of dependency levels between two PTMs, we derive the following criterion for introducing a PHC for a pair of modifications, *ptm*1 and *ptm*2:

Define *Mutual Information Gain (MIG)* of two PTMs as:

$$MIG(ptm1, ptm2) = |MI(ptm1, ptm2) - CMI(ptm1, ptm2|NucPos)| \qquad (6)$$

Then, a PHC is introduced if the following conditions are satisfied:

$$\begin{cases} MIG(ptm1, ptm2) \leq \alpha \\ MI(ptm1, ptm2) \leq \beta \end{cases} \qquad (7)$$

where $\alpha$, $\beta$ >0 are significant thresholds. These criteria will be used to derive PHCs for all pairs of PTMs.

### Derivation of GHCs

From a set of PHCs derived in previous step, we define *hidden confounder graph*, an undirected graph whose nodes correspond to PTMs and edges to PHCs, implying that two nodes are connected if they share a PHC. Maximal clique algorithm is then applied on this graph, resulting in a set of maximal cliques, each corresponding to a GHC.

### Causal relationship inference

To derive causal relationships of PTMs, we first combine BN received from structure learning step with GHCs, forming a network of PTMs and their hidden confounders. The edges among PTMs that share a GHC are then removed. Finally, the algorithm for finding compelled edges [26] is applied to the resulting structure, producing a set of compelled edges representing causal relationships of PTMs.

### Data

**Chromatin modification**. CD4+ T cell data containing 20 methylations, 18 acetylations, and histone variant H2A.Z were retrieved from [9] and [8].

**Nucleosome positioning** data of resting CD4+ T cell was obtained from [45].

**Gene set**. UCSC Known Genes were retrieved from UCSC Genome Browser [46]. After removing genes with duplicated or without U133P2 probe IDs, 12456 genes were kept for analysis.

## Results

### Derivation of hidden confounders

Tag count profiles of 38 PTMs and histone variant H2A.Z, taken at the promoters (*TSS* ± 1*kb*) of 12456 selected genes, were first discretized into 3-category values. Tag count profiles of NucPos were transformed into logarithm scale. Then, all were used to compute *MI* and *MIG* values for all pairs of modifications. In Figure 4, the distributions of these values are illustrated in red. Permutation method [47] was employed to evaluate the significance of these distributions. By which, PTM profiles were permuted 1000 times and the distributions of the new *MI* and *MIG* values for all pair of PTMs were computed for each permutation. The averages of 1000 permuted *MI* and *MIG* distributions are illustrated in blue (Figure 4). The result showed that when *MIG* ≤ 0.0007 and *MI* ≥ 0.002, permutation was unable to create any association with the original *MIG* and *MI* distributions. The significant thresholds $\alpha$ and $\beta$ were thus assigned to 0.0007 and 0.002, respectively. This resulted in a hidden confounder graph of 39 nodes and 63 edges. 50 maximal cliques were derived from this graph, corresponding to the same number of GHCs. The list of GHCs and their belonging modifications is given in supplementary information (http://www.jaist.ac.jp/~s1060011/SI.zip). Although it is hard to show that all GHCs are biologically relevant, we did find supporting evidences for some, whose child nodes are well-characterized modifications. For example, CBP is known to have enzymatic activity on both lysines 14 and 27 of histone H3 [2,48], thus may play the role of confounder for H3K14Ac and H3K27Ac. The same observations were also reported for histone acetyltransferase GCN5, which may be the confounder of H3K14Ac and H3K36Ac [2,49], or of H3K4Ac and H3K14Ac [2,50]. Also, JMJD2C/GASC1 or JMJD2A/JHDM3A may be confounder of H3K9 and H3K36 methylation, though histone methyltransferases often target to specific residues [2].

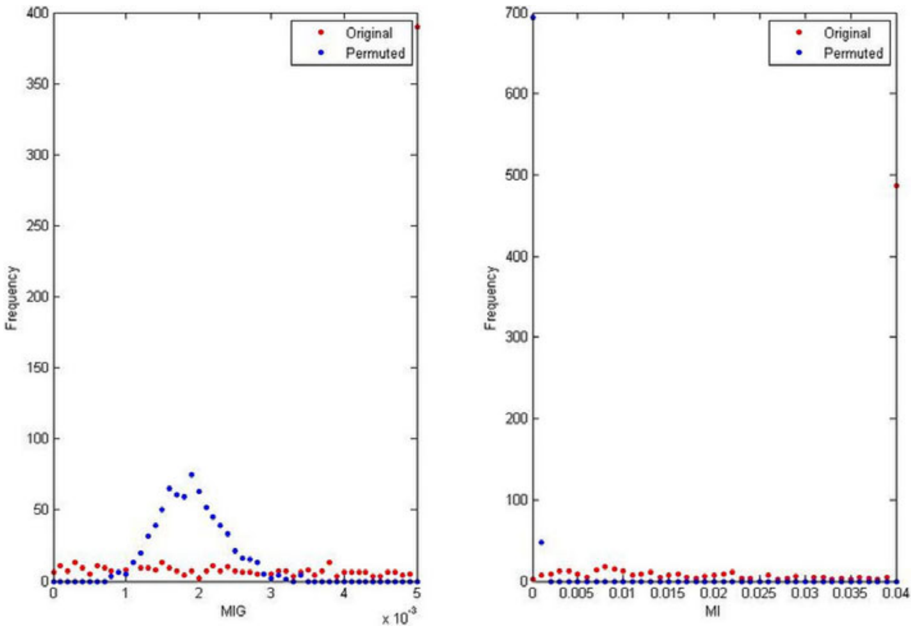

**Figure 4 Distributions of original and average permuted MIG values.**

## Inference of PTM causal relationships

### *General scheme*

BN structures were learned by Banjo (http://www.cs.duke.edu/~amink/software/banjo/), limited to 1, 300, 000 iterations because no significant improvement was achieved in further iteration (data not shown). The resulted structures were combined with 50 GHCs derived in previous step to produce a set of causal relationships. Significance scores were evaluated by bootstrapping method [20]. By which, original data was randomly bootstrapped $N$ times, generating $N$ bootstrapped datasets, and a set of causal relationships was derived for each. Significance score of each relationship was defined as the frequency of its appearance in $N$ bootstrapped sets. In our experiment, $N$ was set to 100.

For comparison, the implementation of *Np* by [30] was run on the same data. Because it only works with binary variables, the data were discretized into binary values by three schemes, using 70 (Scheme 1), 80 (Scheme 2), and 90 (Scheme 3) percentiles as thresholds. After receiving hidden confounders, the above procedure was employed to generate three sets of causal relationships, corresponding to each scheme.

## Comparison

### *Performance*

Table 1 presents the running time and number of hidden confounders derived by the two methods when running on a server machine (Intel Xeon X5570 2.93GHz (4 CPUs), 6GB RAM, Windows Server 2008 OS). It shows that, our method (denoted as *hidden*) worked faster than *Np* (converged after ≈ 200 iterations, data not shown) no matter what discretization scheme was employed. Moreover, to compute MIs and MIGs, it does not require any additional assumption on input data, thus more flexible and practical.

### *Literature-based comparison*

Because it does not exist a list of confirmed causal relationships that could be used as a "gold standard", we resorted to literature to compare the results given by different methods. Biomedical literature represents almost all of our existing knowledge about biological entities and their relationships. For the analysis presented here, we employed a simple but effective way to derive potential associations between PTMs from literature, the cooccurrence approach, which was previously applied for GRN reconstruction [51-53]. Simply, if two PTMs appear in an article abstract indexed in PubMed, we assume an association between them. However, in addition to the associations extracted based on direct cooccurrence, we also assume an association between two PTMs if they share some directly associated biomedical concepts. This assumption is based on the fact that PTMs often functionally interact with each other

**Table 1 Performances of Np and our method.**

| | *Np* (200 iterations) | | | *hidden* |
|---|---|---|---|---|
| | **Scheme1** | **Scheme2** | **Scheme3** | |
| Running time (sec.) | 5.0e+02 | 3.8e+02 | 2.8e+02 | 8.41 |
| #Confounders | 22 | 30 | 17 | 50 |

#Confounders is the number of hidden confounders.

through intermediary proteins [2,54]. To extract these indirect "associations" we employed FACTA+ [55], a state-of-the-art biomedical text mining system which supports both directly and indirectly related (pivot and target, respectively, so called in FACTA+) biomedical term search. Thus, two kinds of literature-based PTM associations were derived with the association weight defined as following. Regarding cooccurrence-based association, we took the weight definition from [52]:

$$w_{C_o}(ptm_1, ptm_2) = \frac{freq(ptm_1, ptm_2)}{max\{freq(ptm_1), freq(ptm_2)\}} \quad (8)$$

in which $freq(ptm1, ptm2)$ is the frequency that both PTM terms appear together in PubMed abstracts, and $freq(ptm_i)$ is the frequency of each individually.

Regarding indirect association based on shared pivot concepts, i.e., proteins/genes in this case:

$$w_{I_n}(ptm_1, ptm_2) = N \times \frac{1}{\sqrt{\sum_{i=1}^{N}(sig_{1_i} - sig_{2_i})^2}} \quad (9)$$

in which $N$ is the number of the most significant shared concepts between two PTMs, $sig_{1i}$ and $sig_{2i}$ are the significant levels, assigned as point-wise mutual information values, of the associations between the *ith* shared concept and the two PTMs. All of these were retrieved through FACTA+ search with the list of the search terms given in supplementary information.

We define a measure, named *literature support*, for comparison purpose. It is the sum of literature-derived weights of $N$ most significant associations (edges) of a resulting model $M$:

$$LS_M(N) = \sum_{i=1}^{N} w(e_i) \quad (10)$$

where $w(e_i)$ is the literature-derived weight of the edge $e_i$ ($i = 1 \cdots N$). Figure 5 illustrates literature supports for the top 50 significant relationships given by three methods. It shows that, in case of both direct (left figure) and indirect (right figure) associations, the most significant relationships given by our method have comparable literature support to the ones given by *noHidden*, and both are better than the result given by *Np*.

An alternative way for comparison is to assess the significance scores of PTM pairs previously reported as highly correlated [52]. [11] developed a biclustering method to search for combinatorial patterns of PTMs on the same data. From the resulting bilusters, they found three most frequently cooccurred PTM pairs: *(H3K27Ac, H3K4Me3)*, *(H2AZ, H2BK120Ac)*, and *(H3K9Ac, H3K36Ac)*. Also, we selected 10 most correlated PTM pairs ($r \geq 0.7$) reported by [8] in their pairwise correlation analysis on the data. Comparison on these two sets of highly correlated PTM pairs shows that the confidence scores assigned by our method are significantly higher than or at least equal to the ones assigned by the other two methods (Tables 2, 3, and

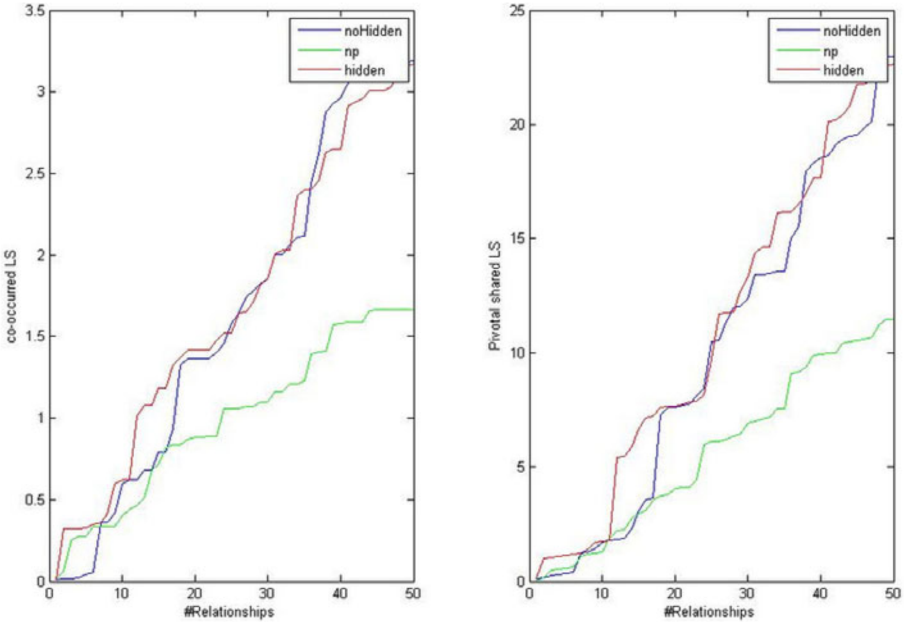

**Figure 5 Literature supports for the top 50 significant relationships given by our method (red), *Np* (green) (scheme 3), and *noHidden* (blue)**.

**Table 2 Comparison on the significance scores of three highly correlated PTM pairs reported in [11].**

| PTM pairs | *hidden* | *noHidden* | *p − value* |
|---|---|---|---|
| H3K27Ac-H3K4Me3 | 0.866 | 0.724 | 2.1e-10 |
| H2AZ-H2BK120Ac | 0.002 | 0.002 | *Nd* |
| H3K9Ac-H3K36Ac | 0.195 | 0.155 | *Nd* |

nd means no difference.

**Table 3 Comparison on the significance scores of 10 most correlated PTM pairs reported in [8].**

| PTM pairs | *hidden* | *noHidden* | *p − value* |
|---|---|---|---|
| H2BK5ac-H3K27ac | 0.677 | 0.481 | 6.26e-10 |
| H2BK120ac-H2BK5ac | 0.594 | 0.301 | 6.11e-10 |
| H2BK120ac-H4K91ac | 0.843 | 0.336 | 1.81e-15 |
| H2BK5ac-H3K9ac | 0.524 | 0.416 | 2.36e-08 |
| H3K79me2-H3K79me3 | 0.794 | 0.793 | *Nd* |
| H2BK120ac-H3K27ac | 0.623 | 0.207 | 3.24e-13 |
| H2BK120ac-H3K18ac | 0.61 | 0.196 | 1.55e-14 |
| H3K18ac-H3K27ac | 0.453 | 0.19 | 1.28e-09 |
| H2BK5ac-H3K18ac | 0.047 | 0.004 | 4.31e-08 |
| H2BK5ac-H4K91ac | 0.294 | 0.283 | *Nd* |

nd means no difference.

supplementary information). This means, taking into account the existence of hidden confounders significantly increases our ability to recover highly correlated pairs of histone modifications.

## Analysis and discussions

Finally, we assessed whether the proposed method can produce biologically meaningful causal relationships by deriving a network model consisted of the most confident relationships (significance score ≥ 0.7). At this threshold, a network of 49 relationships was created (Figure 6).

We investigated biological characteristics of the resulting network by assessing its *dominant modifications* and the most significant *Markov relations* employing the method described in [20]. By which, dominance score of each modification $X$ is calculated by $dScore(X) = \Sigma C_0(X, Y)^k$, where $C_0(X, Y)$ denotes the significance score of $X$ being an ancestor of $Y$, $k$ is the constant to reward highly significant features. Table 4 shows 10 most dominant modifications ($k = 2$, for other values of $k$ only the orders were changed) and significant Markov relations, with the corresponding scores given by our method.

Analyzing the top dominant modifications, we found that 8 out of 10 PTMs, $\{H3K4Me3, H3K27Ac, H2BK120Ac, H4K8Ac, H4K5Ac, H4K91Ac, H3K4Me1, H3K9Ac\}$, have been reported in the original research as important marks that appeared in the modification back-bone at promoters [8]. For the other two, $H3K27Me3$ is known as an important repressive mark, and $H3K27Me1$ as an active mark at promoters [9]. Interestingly, the result suggested the significant role of $H2BK120Ac$ and its regulatory effect on $H3K4Me3$, an important modification mark of active promoters, through the chain $H2BK120Ac \rightarrow H3K18Ac \rightarrow H3K4Me3$. For a long time, the functions of $H2B$ modifications, particularly $H2BK120Ac$, have remained obscure compared to other modifications [56]. Just recently there has been an indication that $H2BK120Ac$ appears as an early modification mark in TSS regions and affects $H2BK120Ub$ [57], a modification that regulates $H3K4Me3$ [58,59], providing support for our finding. Investigation of the most significant Markov relations revealed that well-characterized modifications are mostly functionally related. For example, the

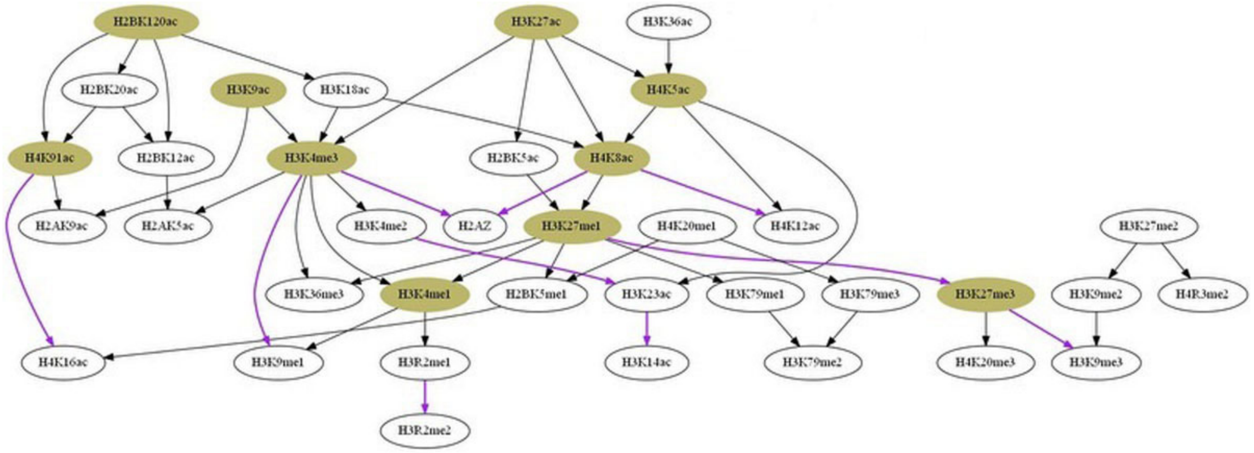

**Figure 6 A network model of highly significant causal relationships given by our method**. 10 most dominant modifications and highest confidence Markov relations are illustrated by filled nodes and purple edges, respectively.

**Table 4 The top dominant histone modifications and significant Markov relations with corresponding dominance and significance scores (*dScore* and $C_0$, respectively) given by our method.**

| Modifications | *dScore* | Markov relations | $C_0$ |
|---|---|---|---|
| *H3K4Me3* | 4.473 | $H3K23Ac \rightarrow H3K14Ac$ | 1 |
| *H4K8Ac* | 2.8836 | $H4K8Ac \rightarrow H2AZ$ | 1 |
| *H3K27Me1* | 2.7993 | $H4K8Ac \rightarrow H4K12Ac$ | 1 |
| *H3K27Ac* | 2.6593 | $H4K91Ac \rightarrow H4K16Ac$ | 1 |
| *H4K5Ac* | 2.3603 | $H3K4Me2 \rightarrow H3K23Ac$ | 1 |
| *H2BK120Ac* | 2.2473 | $H3K4Me3 \rightarrow H2AZ$ | 1 |
| *H4K91Ac* | 1.7744 | $H3K4Me3 \rightarrow H3K9Me1$ | 1 |
| *H3K4Me1* | 1.6105 | $H3R2Me1 \rightarrow H3R2Me2$ | 0.99 |
| *H3K27Me3* | 1.5533 | $H3K27Me3 \rightarrow H3K9Me3$ | 0.98 |
| *H3K9Ac* | 1.3325 | $H3K27Me1 \rightarrow H3K27Me3$ | 0.96 |

N-terminal tail of histone H4 has four acetylated lysines: K5, K8, K12, K16, of which H4 K5Ac/K8Ac/K12Ac play a non-specific, cumulative regulatory role different from that of H4K16Ac [60]. In consistence with this observation, these modifications were predicted to be closely linked and separated from $H4K16Ac$ in the resulting model: $H4K5Ac \rightarrow H4K8Ac$, $H4K5Ac \rightarrow H4K12Ac$, and $H4K8Ac \rightarrow H4K12Ac$ (one of the top 10 Markov relations). For other less well-known modifications, such as $H3R2$ methylations or H3K27 mono-methylation, the links might suggest novel biological understanding. While the relationship between $H3R2Me1 \rightarrow H3R2Me2$ might reflect a directional equilibrium between mono- and di-methyl $H3R2$, the one between $H3K27Me1 \rightarrow H3K27Me3$ might reflect their functional association through G9a methyltransferase, as recently reported by [61]. More interestingly, 4 out of 10 most significant Markov relations have already been reported to be causal in literature. [14] have shown evidences for causal relationships of $H3K27Me3 \rightarrow H3K9Me3$ and $H3K4Me3 \rightarrow H2AZ$. In [62], $H3K9Me1/2$ was shown to be demethylated by $P\ HD$ finger protein 8 ($PHF8$), whose catalytic activity is in turn stimulated by $H3K4Me3$, suggesting the causal effect of $H3K4Me3$ on $H3K9Me1$, represented by the link $H3K4Me3 \rightarrow H3K9Me1$. Also, the deposition of histone variant $H2A.Z$ by $SWR1$ complex is known to be triggered by $NuA4$-mediated acetylation of histone $H4$ [63,64]. Our model supported this observation with the relationship $H4K8Ac \rightarrow H2AZ$. Additionally, causal effects have also been observed to support other relationships of the resulting model. For example, [14] have given evidence for the relationship $H3K4Me3 \rightarrow H3K36Me3$. [65] have reported that the recruitment of $MLL1$, a histone methyltransferase responsible for $H3K4$ methylation, is required for the binding of $TIP60$ histone acetyltransferase, which catalytically acetylates $H2AK5$. In agreement, our model predicted the relationship $H3K4Me3 \rightarrow H2AK5Ac$, suggesting causal effect of $H3K4Me3$ on $H2AK5Ac$.

## Conclusion

Elucidation of functional relationships among histone modifications is crucial to understanding important chromatin-mediated processes. Previous BN-based approaches, however, have not taken into account the existence of hidden regulators when inferring causal relationships of PTMs. We tackled the problem by proposing a novel approach that exploits chromatin organizational information to capture the effect of PTM hidden regulators. Application on human epigenomic data showed the advantage of the proposed method over the previous ones. Moreover, it could recover biologically relevant causal relationships between histone modifications, which may be useful for future investigation of histone crosstalk.

### Competing interests

The authors declare that they have no competing interests.

### Authors' contributions

NTL and TBH defined the research problem and proposed the solution. NTL, BHH, and DHT designed and implemented the experiment. All authors contributed to and approved the final version of the manuscript.

### Acknowledgements

This work was supported by Vietnam's National Foundation for Science and Technology Development (NAFOSTED Project No.102.01−2011.05).

### Declarations

The publication costs for this article were funded by Vietnam's National Foundation for Science and Technology Development (NAFOSTED Project No.102.01−2011.05).

This article has been published as part of *BMC Genomics* Volume 15 Supplement 1, 2014: Selected articles from the Twelfth Asia Pacific Bioinformatics Conference (APBC 2014): Genomics. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcgenomics/supplements/15/S1.

### Authors' details

[1]Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan. [2]Vietnam Academy of Science and Technology, 18 Hoang Quoc Viet, Hanoi, Vietnam. [3]Hanoi National University of Education, 36 Xuan Thuy, Cau Giay, Hanoi, Vietnam.

Published: 24 January 2014

### References

1. Luger K, Ku M, Jaffe DB, Issac B: **Crystal structure of the nucleosome core particle at 2.8 A resolution.** *Nature* 1997, **389**:251-60.
2. Kouzarides T: **Chromatin modifications and their function.** *Cell* 2007, **128(4)**:693-705.
3. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie A, Meissner A, Wernig M, Jaenisch R, Nusbaum C, Lander ES, Bernstein BE: **Genome-wide maps of chromatin state in pluripotent and lineage-committed cells.** *Nature* 2007, **448(7151)**:553-60.
4. Fraga MF, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, P : **Loss of acetylation at Lys16 and trimethylation at Lys20 of histone H4 is a common hallmark of human cancer.** *Nature Genet* 2005, **37**:391-400.
5. Bernstein BE, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, P : **The mammalian epigenome.** *Cell* 2007, **128(4)**:669-81.

6. Hawkins RD, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, P : **Distinct epigenomic landscapes of pluripotent and lineage-committed human cells.** *Cell* 2010, **6(5)**:479-91.
7. Strahl BD, Allis CD: **The language of covalent histone modifications.** *Nature* 2000, **403(6765)**:41-5.
8. Wang Z, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, P : **Combinatorial patterns of histone acetylations and methylations in the human genome.** *Nature Genet* 2008, **40(7)**:897-903.
9. Barski A, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, P : **High-resolution profiling of histone methylations in the human genome.** *Cell* 2007, **129(4)**:823-37.
10. Hon G, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, P : **ChromaSig: a probabilistic approach to finding common chromatin signatures in the human genome.** *PLoS Comput Biol* 2008, **4(10)**:e1000201.
11. Ucar D, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, P : **Combinatorial chromatin modification patterns in the human genome revealed by subspace clustering.** *Nucleic Acids Res* 2011, **39(10)**:4063-75.
12. Jaskchek R, Tanay A: **Spatial clustering of multivariate genomic and epigenomic information.** *Proceedings of the 13th Annual International Conference on Research in Computational Molecular Biology, Springer-Verlag* 2009, 170-183.
13. Ernst J, Kellis M: **Discovery and characterization of chromatin states for systematic annotation of the human genome.** *Nat Biotechnol* 2010, **28(8)**:817-25.
14. Yu H, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, P : **Inferring causal relationships among different histone modifications and gene expression.** *Genome Res* 2008, **18(8)**:1314-1324.
15. Hawkins RD, Hon GC, Ren B: **Next-generation genomics: an integrative approach.** *Nat Rev Genet* 2010, **11(7)**:476-86.
16. Berger SL: **The complex language of chromatin regulation during transcription.** *Nature* 2007, **447**:407-12.
17. Latham JA, Dent SYR: **Cross-regulation of histone modifications.** *Nat Struct Mol Biol* 2007, **14(11)**:1017-24.
18. Suganuma T, Workman JL: **Crosstalk among Histone Modifications.** *Cell* 2008, **135(4)**:604-7.
19. Jensen FV, Nielsen TD: *Bayesian Networks and Decision Graphs (second edition)* New York: Springer-Verlag; 2001.
20. Friedman N, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, P : **Using Bayesian networks to analyze expression data.** *J Comput Biol* 2000, **7(3-4)**:601-20.
21. Jansen R, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, P : **A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data.** *Science* 2003, **302(5644)**:449-53.
22. Steensel BV, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, P : **Bayesian network analysis of targeting interactions in chromatin.** *Genome Res* 2009, **20(2)**:190-200.
23. Sachs K, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, P : **Causal protein-signaling networks derived from multiparameter single-cell data.** *Science* 2005, **308(5127)**:523-9.
24. Lv J, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, P : **Discovering Cooperative Relationships of Chromatin Modifications in Human T Cells Based on a Proposed Closeness Measure.** *PLoS One* 2010, **5(12)**:e14219.
25. Le NT, Ho TB: **Reconstruction of histone modification network from next-generation sequencing data.** *IEEE International Conference on Bioinformatics and Bioengineering* 2011.
26. Chickering DM: **A transformational characterization of equivalent Bayesian network structures.** *Proceedings of 11th Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers* 1995, 87-89.
27. Zhang J, Spirtes P: **A Transformational Characterization of Markov Equivalence between DAGs with Latent Variables.** *Proceedings of 21th Conference on Uncertainty in Artificial Intelligence, UAUI Press* 2005, 667-674.
28. Koller D, Friedman N: **Probabilistic Graphical Models: Principles and Techniques.** *MIT Press* 2009.
29. Bannister AJ, Kouzarides T: **Regulation of chromatin by histone modifications.** *Cell Res* 2011, **21(3)**:381-95.
30. Wood F, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, P : **A Non-Parametric Bayesian Method for Inferring Hidden Causes.** *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence* 2006, 536-543.
31. Martin JD, VanLehn K: **Discrete factor analysis: Learning hidden variables in Bayesian networks.** *Technical report, Department of Computer Science, University of Pittsburgh* 1995.
32. Basso K, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, P : **Reverse engineering of regulatory networks in human B cells.** *Nat Genet* 2005, **37(4)**:382-90.
33. Margolin AA, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, P : **ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context.** *BMC Bioinformatics* 2006, **7(S7)**.
34. Meyer PE, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, P : **minet: A R/Bioconductor package for inferring large transcriptional networks using mutual information.** *BMC Bioinformatics* 2008, **9(461)**.
35. Zhang X, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, P : **Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information.** *Bioinformatics* 2012, **28**:98-104.
36. Peng H, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, P : **Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy.** *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2005, **27(8)**:1226-38.
37. Heckerman D, Geiger D, Chickering DM: **Learning Bayesian networks: The combination of knowledge and statistical data.** *Machine Learning* 1995, **20**:197-243.
38. Cooper G, Herskovits E: **A Bayesian method for the induction of probabilistic networks from data.** *Machine Learning* 1992, **9**:309-47.
39. Chickering DM: **Learning Bayesian Networks is NP hard.** *Technical report, Redmond, WA: Microsoft Research* 1994.
40. Brown CE, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, P : **The many HATs of transcription coactivators.** *Trends Biochem Sci* 2000, **25**:15-9.
41. Hager GL, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, P : **Influence of chromatin structure on the binding of transcription factors to DNA.** *Cold Spring Harb Symp Quant Biol* 1993, **58**:63-71.
42. Kwon H, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, P : **Nucleosome disruption and enhancement of activator binding by a human SW1/SNF complex.** *Nature* 1994, **370(6489)**:477-81.
43. Floer M, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, P : **A RSC/Nucleosome Complex Determines Chromatin Architecture and Facilitates Activator Binding.** *Cell* 2010, **141(3)**:407-18.
44. Wang X, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, P : **Nucleosomes and the accessibility problem.** *Trends Genet* 2011, **27(12)**:487-92.
45. Schones DE, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, P : **Dynamic regulation of nucleosome positioning in the human genome.** *Cell* 2008, **132(5)**:887-98.
46. Karolchik D, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, P : **The UCSC Table Browser data retrieval tool.** *Nucleic Acids Res* 2004, **132**: D493-6.
47. Butte AJ, Kohane IS: **Mutual Information Relevance Networks: Functional Genomic Clustering Using Pairwise Entropy Measurements.** *Pacific Symposium on Biocomputing* 2000, **5**:415-26.
48. Tie F, A , B : **CBP-mediated acetylation of histone H3 lysine 27 antagonizes Drosophila Polycomb silencing.** *Development* 2009, **136(18)**:3131-41.
49. Morris SA, Banerjee R, Stratton CA: **Identification of histone H3 lysine 36 acetylation as a highly conserved histone modification.** *J Biol Chem* 2007, **282(10)**:7632-40.
50. Guillemette B, Banerjee R, Stratton CA: **H3 lysine 4 is acetylated at active gene promoters and is regulated by H3 lysine 4 methylation.** *PLoS Genet* 2011, **7(3)**:e1001354.
51. Jenssen TK, Higuchi T, Goto T, Jaakkola TS: **A literature network of human genes for high-throughput analysis of gene expression.** *Nat Genet* 2001, **28**:21-28.
52. Djebbari A, Quackenbush J: **Seeded Bayesian Networks: Constructing genetic networks from microarray data.** *BMC Systems Biology* 2008, **2(57)**.
53. Steele E, Tucker A, Schuemie PHM: **Literature-based priors for gene regulatory networks.** *Bioinformatics* 2009, **25(14)**:1768-17748.
54. Izzo A, Schneider R: **Chatting histone modifications in mammals.** *Brief Funct Genomics* 2010, **9(5-6)**:429-43.
55. Tsuruoka Y, Higuchi T, Goto T, Jaakkola TS: **Discovering and visualizing indirect associations between biomedical concepts.** *Bioinformatics* 2011, **27(13)**:i111-i119.
56. Wyrick JJ, Parra MA: **The role of histone H2A and H2B post-translational modifications in transcription: a genomic perspective.** *Biochim Biophys Acta* 2009, **1789**:37-44.

57. Gatta R, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, P : **An acetylation-mono-ubiquitination switch on lysine 120 of H2B.** *Epigenetics* 2011, **6(5)**:630-7.
58. Lee JS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, P : **Histone crosstalk between H2B monoubiquitination and H3 methylation mediated by COMPASS.** *Cell* 2007, **131(6)**:1084-96.
59. Shukla A, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, P : **Histone methylation and ubiquitination with their cross-talk and roles in gene expression and stability.** *Cell Mol Life Sci* 2009, **66(8)**:1419-33.
60. Dion MF, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, P : **Genomic characterization reveals a simple histone h4 acetylation code.** *Proc Natl Acad Sci USA* 2005, **102(15)**:5501-6.
61. Yoo K, Hennighausen L: **EZH2 methyltransferase and H3K27 methylation in breast cancer.** *Int J Biol Sci* 2012, **8**:59-65.
62. Feng W, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, P : **PHF8 activates transcription of rRNA genes through H3K4me3 binding and H3K9me1/2 demethylation.** *Nat Struct Mol Biol* 2010, **17(4)**:445-50.
63. Altaf M, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, P : **NuA4-dependent Acetylation of Nucleosomal Histones H4 and H2A Directly Stimulates Incorporation of H2A.Z by the SWR1 Complex.** *J Biol Chem* 2010, **285(21)**:15966-77.
64. Zhou BO, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, P : **SWR1 complex poises heterochromatin boundaries for antisilencing activity propagation.** *Mol Cell Biol* 2010, **30(10)**:2391-400.
65. Jeong KW, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, P : **Recognition of enhancer element specific histone methylation by TIP60 in transcriptional activation.** *Nat Struct Mol Biol* 2011, **18(12)**:1358-65.

doi:10.1186/1471-2164-15-S1-S7
**Cite this article as:** Le *et al.*: A nucleosomal approach to inferring causal relationships of histone modifications. *BMC Genomics* 2014 **15**(Suppl 1): S7.