BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 8 2012, pages 1062–1069
doi:10.1093/bioinformatics/bts085

Genome analysis

Advance Access publication February 17, 2012

Signal analysis for genome-wide maps of histone modifications measured by ChIP-seq

Dominik Beck[1,2,3,*], Miriam B. Brandl[3,4], Lies Boelen[1,2], Ashwin Unnikrishnan[1,2], John E. Pimanda[1,2,*] and Jason W. H. Wong[1,2,*]

[1]Lowy Cancer Research Centre, [2]Prince of Wales Clinical School, University of New South Wales, Sydney, NSW 2052, [3]School of Engineering and Information Technology, University of New South Wales, Canberra, ACT, 2600 and [4]Children's Cancer Institute Australia, Lowy Cancer Research Centre, University of New South Wales, Sydney, NSW 2052, Australia

Associate Editor: Alex Bateman

ABSTRACT

Motivation: Chromatin structure, including post-translational modifications of histones, regulates gene expression, alternative splicing and cell identity. ChIP-seq is an increasingly used assay to study chromatin function. However, tools for downstream bioinformatics analysis are limited and are only based on the evaluation of signal intensities. We reasoned that new methods taking into account other signal characteristics such as peak shape, location and frequencies might reveal new insights into chromatin function, particularly in situation where differences in read intensities are subtle.

Results: We introduced an analysis pipeline, based on linear predictive coding (LPC), which allows the capture and comparison of ChIP-seq histone profiles. First, we show that the modeled signal profiles distinguish differentially expressed genes with comparable accuracy to signal intensities. The method was robust against parameter variations and performed well up to a signal-to-noise ratio of 0.55. Additionally, we show that LPC profiles of activating and repressive histone marks cluster into distinct groups and can be used to predict their function.

Availability and implementation: http://www.cancerresearch.unsw.edu.au/crcweb.nsf/page/LPCHP A Matlab implementation along with usage instructions and an example input file are available from: http://www.cancerresearch.unsw.edu.au/crcweb.nsf/page/LPCHP

Contact: d.beck@student.unsw.edu.au; jpimanda@unsw.edu.au; jason.wong@unsw.edu.au

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

Received on October 14, 2011; revised on January 20, 2011; accepted on February 13, 2012

1 INTRODUCTION

Gene expression is controlled at multiple levels, including factors regulating DNA accessibility (Hobert, 2008). A combinatorial code of post-translational modifications of the nucleosome proteins H2A, H2B, H3 and H4, can render their associated DNA, accessible or inaccessible for transcription and splicing (Watson, 2003). Acetylation is generally associated with activation, while the affect of histone methylations is condition and position dependent (Wang *et al*., 2008). Chromatin immunoprecipitation (ChIP), originally developed to enrich DNA fragments bound by specific proteins, is now frequently used in studies of histone modifications (HMs) (Barski *et al*., 2007). In brief, a target protein, such as a transcription factor (TF) or histone with a particular modification is first cross-linked and immunoprecipitated using an antibody. The associated DNA is then extracted and detected using methods such as a hybridization array (ChIP-chip) or deep sequencing (ChIP-seq). The later approach shows advantages in resolution and following bioinformatics processing provides genome-wide maps of TF binding and HMs.

A large number of algorithms for downstream processing of ChIP-seq datasets have been developed (Pepke *et al*., 2009). These tools generally evaluate signal intensities, with the aim of detecting sparse, highly localized and enriched peaks. However, while this is a common characteristic for TF binding, the signals from HM ChIP-seq are more variable. For example, H3K4me3 and H2A.Z, localize at the transcription start site (TSS) whereas others, such as H3K36me3 and H3K20me1, spread over many base pairs covering the promoter and full gene body (Barski *et al*., 2007; Wang *et al*., 2008). In addition, these signals often differ in peak shape and location, as well as frequency. By focusing the analysis on signal intensity alone, existing methods are potentially limiting and constrain our ability to identify common motifs or patterns.

A number of genome-wide studies have recently been published and show correlations between HMs and gene expression. While each study varied slightly in processing of the ChIP-seq data, in all cases, the read intensities that mapped to an arbitrary region around the TSS of a gene were extracted and normalized. A vector containing quantification values from each dataset was constructed for each gene, and these vectors were subsequently correlated with gene expression using different multivariate statistical approaches. The earliest work analyzed ChIP-seq data for 20 histone methylations (Yu *et al*., 2008) and identified two major gene groups by hierarchical cluster analysis. The first contained active marks and genes expressed above average while the second cluster contained repressive marks and genes expressed below average. A later study used the same methylation data but also included measures of 19 histone acetylation profiles (Karlic *et al*., 2010). Here, a linear regression model was used and again revealed that the HM ChIP-seq data was predictive for gene expression.

*To whom correspondence should be addressed.

A smaller study of the two methylations, H3K4me3 and H3K27me3, applied a mixture of linear regression models and found that HMs were more predictive for gene expression compared to TF binding (Costa *et al.*, 2011).

Interestingly, these articles reported good predictability of gene expression by three different computational models, all using transformations of read counts as input. However, purely based on intensity information, these approaches neglect other key signal characteristics such as peak shape and location, as well as signal frequencies. Incorporating peak shape information will be essential to identify HMs present on adjacent nucleosomes. The specific locations of the histone modification, e.g. in the promoter or within the transcribed region, have been correlated with different functions, including transcription initiation, promoter clearance or transcriptional elongation (Karlic *et al.*, 2010; Wang *et al.*, 2008). In addition, recent data from ChIP-seq of H3K27me3 identified three different regions relative to the TSS of gene, which independently correlate with different gene expression levels (Young *et al.*, 2011). Motivated by the biology and the shortcoming in available algorithms, we propose a new strategy that quantifies the ChIP-seq profile, making use of the pattern and location of the signal.

After data pre-processing, the linear predictive coding (LPC) model, a method widely used in speech recognition, was applied to optimally parameterize the signal. The derived coefficients were used as quantitative features replacing signal intensities. With focus on the correlation between histone methylation and gene expression, we first validated our approach and show that it is robust, tolerates noise and performs with comparable accuracy to read intensities in a general and large-scale cross-validation. We then show that our method can be used to successfully predict the function of HMs.

2 METHODS

2.1 Datasets and gene selection

In this study, gene expression and HM data of resting CD4$^+$ T cells was analyzed. The gene expression data, measured on an Affymetrics whole-genome HG-U133A expression array, was obtained from (Su *et al.*, 2004). Genome-wide ChIP-seq data were acquired on a Solexa Genome 1G platform and obtained from Barski *et al.* (2007). These included, 20 histone modifications, the histone variant H2Z.A, as well as PolII and CTCF binding. The recently established consensus coding sequence (CCDS) database (Pruitt *et al.*, 2009) was used to map between these datasets.

2.2 Pre-processing of histone modification data

The datasets from 21chromatin modification, PolII and CTCF were referred to as $C=\{c_1,c_2,c_3,...,c_M\}$ and $M=23$. On an average, each ChIP-seq experiment contained 8.19×10^6 sequence reads with a length of 20 bp and the total read counts that uniquely mapped to the genome were denoted as r_c.

The CCDS database was used to identify human protein coding genes g that have high-quality annotations for their genome location. A genomic window w_g was then defined for every gene, stretching equal distances from its TSS and having a total length $l(w_g)$. The positions $w_g(i)$, $i=1,2,3,\ldots,l(w_g)$ represent the base pairs within the genomic loci.

Next, signal profiles at w_g were extracted from all ChIP-seq experiments. Therefore, the forward (+) and reverse (−) strands were first considered independently and aligned sequence tags at each genome coordinate i were summed up into the read profiles r_{cg}^+ and r_{cg}^- respectively (Fig. 1, Step 1). These profiles were then joined to avoid strand-specific bias (Valouev *et al.*, 2008):

$$r_{cg}=r_{cg}^+(i-\lambda)+r_{cg}^-(i+\lambda) \tag{1}$$

where λ is the peak shift parameter, which was calculated for individual regions of 300 bp that exceed a threshold of 600 aligned sequence reads. In each region, we first indentified local maxima that correspond between the forward and reversed strands, λ was then calculated as the average distance between these peaks (Fig. 1, Step 2).

The resulting signal profiles were transformed, using kernel density estimators, into density profiles further smoothening the signal and increasing resolution, which both benefits the LPC estimation procedure (Pepke *et al.*, 2009; Silverman, 1998). In addition, this step removed the overall intensity information. The density profile z_{cg} was then calculated as follows:

$$z_{cg}(i)=\frac{1}{nh}\sum_{j=1}^{n}K(u_{cg}) \tag{2}$$

$$u_{cg}=\frac{r_{cg}(i)-r_{cg}(j)}{h} \tag{3}$$

where n is the number of reads, h is a smoothing parameter and K the kernel function. Here, we use the Gaussian kernel which is given by $K(u)=\frac{1}{\sqrt{2\pi}}e^{-0.5u^2}$ and the bisquare kernel given by $K(u)=\frac{15}{16}\left(1-u^2\right)^2$ if $1\leq u\leq 1$ and 0 otherwise (Fig. 1, Step 3).

2.3 Computation of LPC features from ChIP-seq

LPC is a spectral analysis method that optimally characterizes a given wave signal by a set of LPC parameters (Rabiner and Juang, 1993). It has a number of desirable characteristics including a mathematically precise solution, straightforward and simple implementation as well as low computational costs. In addition, the LPC model is widely applied in speech recognition systems, where it performs well in variety of applications. Interestingly, the waveforms and large variance that are characteristic for human speech are similarly found in ChIP-seq signals of different HM [compare figure 1 in Pinkowski (1993)].

Therefore, we have investigated the application of LPC to parameterize ChIP-seq data into spectral vectors, and use these features for downstream bioinformatics analysis (Fig. 1, Steps 4 and 5).The basic idea behind the application of LPC to ChIP-seq data is that the kernel density estimate $z_{cg}(i)$ of the signal c at gene g and genome coordinate i can be approximated from a linear combination of the previous signals $i-1,i-2,i-3,...,i-p$ following (Rabiner and Juang, 1993):

$$\hat{z}_{cg}(i)=\sum_{k=1}^{p}a_{cg}(k)z_{cg}(i-k) \tag{4}$$

where $a_{cg}(k)$ with $k=1,2,3,...,p$ are the prediction coefficients to be determined, and p also referred to as poles denotes the number of earlier signals used. The approximation error $e_{cg}(i)$ is given by: $e_{cg}(i)=z_{cg}(i)-\hat{z}_{cg}(i)$ and thus the mean squared error E can be expressed as:

$$E_{cg}=\sum_{i=1}^{l(w_g)}\left[z_{cg}(i)-\sum_{k=1}^{p}a_{cg}(k)z_{cg}(i-k)\right]^2 \tag{5}$$

Having obtained an expression for the mean square error, we can derive an optimal solution for each $a_{cg}(k)$ that minimizes E:

$$\frac{\partial E_{cg}}{\partial a_{cg}}=0,\quad k=1,2,3,...,p \tag{6}$$

which can be solved by the autocorrelation method:

$$a_{cg}=R_{cg}^{-1}r_{cg} \tag{7}$$

where a_{cg} is a $p\times 1$ vector that contains the LPC coefficients, R_{cg} is a $p\times p$ Toepliz matrix and r_{cg} is a $p\times 1$ vector both containing the autocorrelation values.

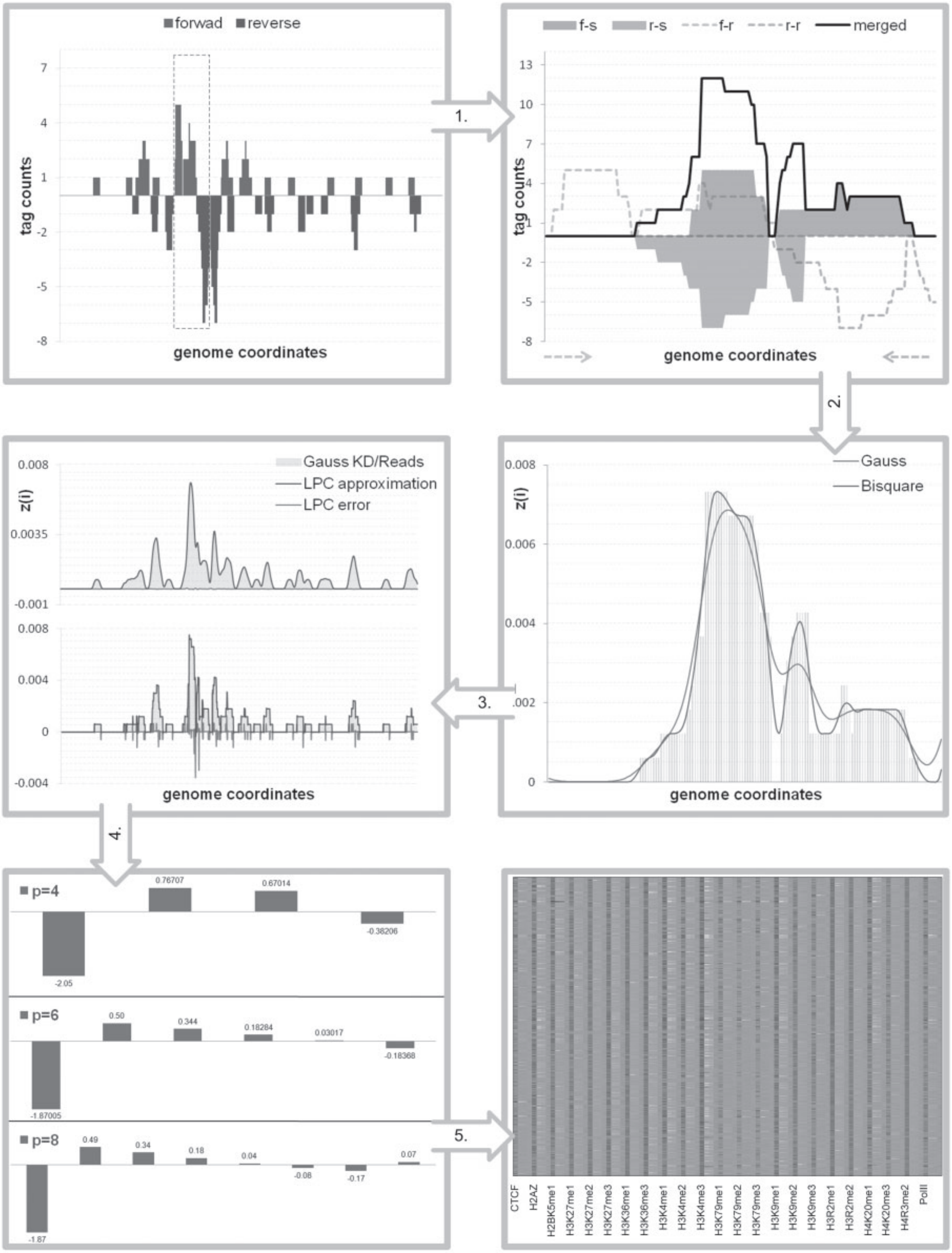

Fig. 1. Overview of the data processing steps from raw reads to the final LPC parameterization for an arbitrary example of a histone mark 1.5 kb around a TSS. First, raw read alignments for the forward and reverse strand are shown. After applying the peak shifting strategy the signals from both strands were merged (Step 1). The merged signal was smoothened using kernel density estimates, and we show the fit of two kernel functions here (Step 2). The LPC approximations (Step 3, black) and their estimation errors (read) are shown for a smooth Gaussian density estimate (top) and a density estimate based on raw reads (bottom). LPC parameterizations of the smooth Gaussian density signal are shown for three different parameters (Step 4). A heatmap of the LPC parameterization ($p=4$) of the bisquare kernel functions ($h=10$) derived from 23 ChIP-Seq datasets of 200 highly expressed genes (Step 5).

2.4 Histone profiles derived from LPC and read intensities

The LPC features a_{cg}, obtained from each of the analyzed experiments of gene g, were summarized to the feature vector $f_g = [a_1, a_2, a_3, ..., a_M]$. We denote f_g the LPC histone profile (LPCHP) of the gene g as it describes its overall histone pattern, including PolII and CTCF, measured by 23 ChIP-seq experiments as described earlier.

In addition, histone profiles from read intensities (RIHP) were calculated. Therefore, the signal profiles r_{cg} for a given gene were summed as following:

$$I_{cg} = \log_2 \left(\frac{1}{r_c} \sum_{i=1}^{l(w_g)} r_{cg}(i) \right). \quad (8)$$

The final values from all datasets M were summarized to the feature vector $f'_g = [I_1, I_2, I_3, ..., I_M]$. We denote f'_g the read intensity profile of the gene g.

2.5 Cross-validation accuracies for validation and parameter analysis of LPCHP

In order to identify reasonable model parameters and to validate the proposed approach, we performed cross-validation analysis on several pre-labeled gene groups. Therefore, the LPCHP f_g, and for comparison reason the RIHP f'_g, were derived for all pre-labeled genes and a supervised classification (two and three groups) was performed using the support vector machine approach (Chang and Lin, 2001; Hastie *et al*., 2009).

In brief, for each cross-validation the full dataset is first randomly split into k equally sized parts, each part is used once as a validation set while the other $k-1$ parts are used as a training set (Hastie *et al*., 2009). In each experiment, k supervised classifications tasks are executed and the mean classification accuracy over all runs is calculated.

In our analysis the classifications were run 100 times to allow for statistical assessment of the results.

2.6 Cluster congruence for parameter comparison of LPCHP

Cluster analysis using the average linkage method and Euclidean distance metric was performed on the LPCHP with different parameters. In order to quantitatively evaluate the congruence of these clusterings, we used a popular measure called the Adjusted Rand Index (ARI) (see Supplementary Material 1 of detailed description) (Hubert and Arabie, 1985), which has previously been used in bioinformatics applications (Thalamuthu *et al*., 2006). It measures the agreement between two clusterings U and V, and is bound to $ARI \in [0,1]$ with $ARI(U,V) = 1$ in case of maximal agreement (e.g. identical clusterings) and $ARI(U,V) = 0$ when the clusters are independent (e.g. no agreement between the clusterings).

2.7 Histone function prediction through similarity analysis of LPCHP

We investigated if functionally similar methylations also show similarities in their ChIP-seq structure. Therefore, the full dataset C was split into two non-overlapping sets, K containing well characterized transcriptional activators or repressors, and the set X with histone modifications that are less studied in CD4$^+$ cells. In each set the histone marks were represented by the LPC parameterization of some pre-selected genes, covering high and baseline expression levels.

In order to identify structural similarities in K, we used a standard hierarchical clustering algorithm based on the average linkage method and Euclidean distance metric.

For functional prediction of histone methylations, we split K into a two class training sets containing the transcriptional activators and repressors A and R, respectively, and defined as follows (compare also Fig. 3):

$$\forall A_i \{A_{i,H}, A_{i,L}\}, i = 1,2,3,...,\alpha$$
$$\forall R_i \{R_{j,H}, R_{j,L}\}, j = 1,2,3,...,\beta \quad (9)$$

where the indices H and L were used to represent the genes in G_H and G_L. In order to predict the function of each methylation in $x \in X$ we used a simple decision rule, defined by the mean distance of x and the training sets following:

$$D_A = \frac{1}{2\alpha} \sum_{i=1}^{\alpha} \left(\sum_{\substack{x \in X_H \\ a \in A_{i,H}}} \frac{d(x,a)}{|x_H| \, |A_{i,H}|} + \sum_{\substack{x \in X_L \\ a \in A_{i,L}}} \frac{d(x,a)}{|x_L| \, |A_{i,L}|} \right)$$

$$D_R = \frac{1}{2\beta} \sum_{j=1}^{\beta} \left(\sum_{\substack{x \in X_H \\ r \in R_{j,H}}} \frac{d(x,r)}{|x_H| \, |R_{j,H}|} + \sum_{\substack{x \in X_L \\ r \in R_{j,L}}} \frac{d(x,r)}{|x_L| \, |R_{j,L}|} \right) \quad (10)$$

where d is the Euclidean distance. The histone methylation x is assigned to the activation or repressing group following:

$$x \rightarrow \begin{cases} A \ \ if \ \ D_A < D_R \\ R \ \ if \ \ D_R < D_A \ . \end{cases} \quad (11)$$

Note that in case $D_A = D_R$ no prediction can be made for the specific methylation. The greater the difference between D_A and D_R the better the prediction.

2.8 Implementation

The algorithms were implemented in Matlab (version 7.11.0.584). All datasets including the ChIP-seq and expression array, as well as the annotation files from the CCDS and affymetrix were imported to and mapped between using MySql (version 5.5.14). The LibSVM implementation of the support vector machine was used with default parameters (Chang and Lin, 2001). A Matlab script implementing the procedure outlined in Figure 1, along with example input files and usage instructions are available from: http://www.cancerresearch.unsw.edu.au/crcweb.nsf/page/LPCHP

3 RESULTS

3.1 Evaluation of LPCHP and comparison with RIHP

In order to validate the proposed methodology, we set up several cross-validation experiments based on the support vector machine approach with standard parameters (Hastie *et al*., 2009). Following the idea of the histone code theory, HM profiles are expected to be similar within and different between gene groups expressed at different levels. Hence, a logical way to evaluate our methodology is to determine how well the LPCHP distinguishes different gene groups and how this accuracy compares to a more traditional analysis based on RIHP.

In an initial exploratory study, we cross-validated a large number of all model parameters on a small gene set. This revealed that the bisquare kernel with parameter $h = 10$ performed best across most combinations. Therefore, this parameter was fixed for the remainder of this article. In contrast, the comparison for the parameter p in the LPC model identified a number of parameters with good performance, indicating the robustness of this model. Reasonable values were found between 8 and 16, which is the suggested range for speech recognition application, as well as smaller values including 4 and 6, and larger values like 22 and 24 (Supplementary Material 2).

For a detailed evaluation of our approach on a larger dataset we first defined three gene groups for high (G_H), medium (G_M) and

Table 1. Cross-validation comparison

Gene set	k	RIHP	LPCHP $p=4$	LPCHP CUSTOM $p=25$	LPCHP CUSTOM $p=28$
(G_H,G_L)	2	(0.89, 0.87)	(0.87, 0.85)	(0.86, 0.85)	(0.84, 0.82)
	5	(0.89, 0.88)	(0.87, 0.86)	(0.86, 0.85)	(0.84, 0.83)
	10	(0.89, 0.89)	(0.87, 0.86)	(0.86, 0.85)	(0.84, 0.83)
	L1O	0.89	0.87	0.85	0.85
(G_H,G_M)	2	(0.72, 0.69)	(0.69, 0.66)	(0.67, 0.65)	(0.69, 0.66)
	5	(0.73, 0 .71)	(0.69, 0.67)	(0.67, 0.67)	(0.69, 0.67)
	10	(0.73, 0.71)	(0.69, 0.67)	(0.67, 0.66)	(0.69, 0.67)
	L1O	0.71	0.67	0.65	0.65
(G_M,G_L)	2	(0.88, 0.87)	(0.86, 0.85)	(0.86, 0.85)	(0.84, 0.82)
	5	(0.88, 0.871)	(0.86, 0.85)	(0.86, 0.85)	(0.84, 0.83)
	10	(0.88, 0.87)	(0.85, 0.85)	(0.85, 0.85)	(0.84, 0.83)
	L1O	0.87	0.85	0.85	0.85
(G_H,G_M, G_L)	2	(0.72, 0.70)	(0.69, 0.67)	(0.67, 0.66)	(0.66, 0.64)
	5	(0.72, 0.71)	(0.69, 0.68)	(0.68, 0.67)	(0.67, 0.66)
	10	(0.72, 0.71)	(0.69, 0.68)	(0.68, 0.67)	(0.67, 0.66)
	L1O	0.72	0.68	0.66	0.66

This table shows the cross-validation results in two different experiments. In the first case the parameter p was kept constant for all datasets included in the LPCHP and we report the best performing parameter $p=4$. In the second experiment, the parameter p was adapted for each dataset in the LPCHP using the variogram approach. In this case we show the results for two different parameters for H3K4me3 $p=25$ and $p=28$. In each case the k-fold cross-validation was run 100 times and the best (first number) and mean (second number) classification rates are shown.

low (G_L) expressed genes. For each set, 500 genes were identified based on expression measurements from the HG-U133A expression array (Su *et al.*, 2004). The mean expression levels were $G_H=$ 3613.6, $G_M=209.66$ and $G_L=2.39$. The LPCHP and RIHP were calculated and we visualized the first three principle components in Supplementary Figure S1. The RIHP clearly separated the three gene groups, as expected from the literature. However, strikingly the LPCHP also showed a significant grouping, similar to the RIHP.

In order to quantify this trend we performed a k-fold cross-validation with the objective of classifying the whole dataset (e.g. 1500 genes) into the three expression groups. The results for $k=10,5$ and 2 as well as using the leave-one-out strategy are shown in Table 1, with more details shown in Supplementary Table S1. As expected from the principle component analysis the RIHP and the LPCHP achieved similar classification rates. For example, the overall differences in mean classification from RIHP to LPCHP were ~2.5% for (G_H,G_L), ~4.5% for (G_M,G_L), ~1.8% for (G_H,G_M) and ~3.3% for (G_H,G_M,G_L).

In addition, the LPCHP approach performed well and achieved mean classification accuracy >75% over all four experiments. The classification accuracies between (G_H,G_L) were on average ~81%, between (G_M,G_L) were 85% and (G_H,G_M) achieved only 65%. However, this trend was expected as genes in G_H and G_M are expressed with relatively high copy numbers and should be actively marked on the histone level (e.g. $G_H=3613.6$ and $G_M=209.66$).

Therefore, we further examined the predictive power of any single HM to distinguish the gene groups G_H and G_M. We notice that in this comparison the LPCHP outperformed the classification rates of RIHP for H3K79me1, H3K79me2, H3K79me3, H3K9me1, H3K9me2, H3K9me3 and H4K20me (Supplementary Fig. S3).

Noise tolerance analysis further suggested that the approach remains accurate up to a signal-to-noise ratio of 0.55 and 0.3 for Gaussian and uniform noises, respectively (Supplementary Fig. S4).

3.2 Evaluation and robustness of LPCHP under different parameters

3.2.1 Evaluation of LPCHP with parameter settings customized to each dataset In the last section, the same p was used for all datasets included in the LPCHP. The cross-validation analysis resulted in consistent classification accuracies for a number of different p. Here, we further investigated whether a customized parameter p should be selected for each dataset. Unfortunately, due to the large number of potential combinations, it is not computationally feasible to cross validate different p for each datasets included into the LPCHP. As an example, to test only five different values, for each of the 23 ChIP-seq experiments, a total of 5^{23} cross-validation experiments are required.

Therefore, we utilized the variogram approach (see Supplementary Material 3 for detailed description) to estimate parameters for a number of selected genes from each ChIP-seq experiment (Supplementary Table S2 and Supplementary Material 3 and 4). We found that the estimates were stable for the genes within each dataset. An exception was H3K4me3, where the variogram suggested two different parameters for high and medium marked genes. In addition, for 19 out of 23 (~83%) experiments the parameter was either 25 or 26. This stability is in line with the cross-validation presented above. It further provides evidence that p can be selected constant for the analysis of different ChIP-seq datasets, hence making its application more straightforward in future applications.

However, the values derived from the variogram analysis differed from those determined in the cross-validation of the previous section. Hence, we revisited the cross-validation results for the LPC parameters 25 and 26 (Supplementary Table S1). Further, we investigated the cross-validation performance of LPCHP, when a customized p (Table 1) was set for each dataset included in the analysis. When compared to the previous LPCHP with $p=4$, we did not find significant improvements on the classification accuracy.

The result indicates that LPCHP is robust against changes in p, including cases where it was customized to each dataset. In addition, the variogram analysis estimated constant parameters p from all ChIP-seq datasets. Together this suggests that it is appropriate to use the same parameter for the datasets analyzed here.

3.2.2 Analysis of robustness of the LPCHP parameters In the analysis of the past two sections, our approach performed well in cross-validation for multiple parameters. However, it is unclear if different parameter sets extract the same structural information from the underlying ChIP-seq datasets. Therefore, we further evaluated the LPCHP using the recurring task of de novo gene groupings. We reasoned that the proposed approach is robust in its parameters if for different parameterizations p the clustering results are identical or similar. This contrasts with clustering results that are highly dissimilar or close to random.

The ARI was used to compare the clusterings for the genesets G_H,G_L,G_M when different parameterizations p and k-means clustering ($k=3$) are used (Fig. 2). In general, the ARI values were high, indicating excellent congruence between parameterizations.

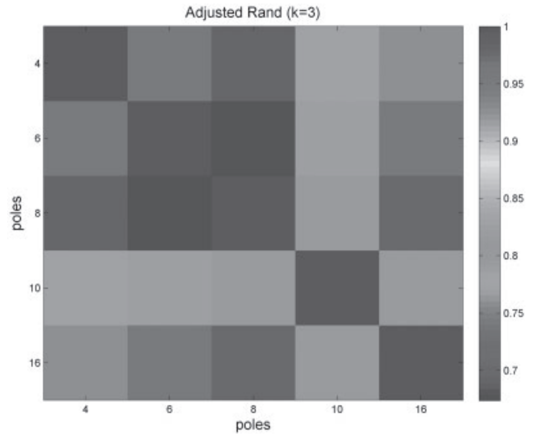

Fig. 2. Heatmap showing the congruence of clusters generated by different LPC parameters, p, using the ARI. The histone profiles (f_g) of 1500 genes (high, low and medium expressed) obtained under different p were clustered into three groups ($k=3$) using the k-means algorithm and compared using the ARI.

The greatest similarity was found between three of the five tested values $p=4,6$ and 16 with ARI of 0.95 ($p=4,6$), 0.93 ($p=4,16$) and 0.95 ($p=6,16$), respectively. This suggests that further downstream analysis would be marginally influenced if p was to be selected among these values. Similarly, the smallest ARI0.67 found for $p=6$ and $p=8$, was still relatively high.

The high congruence suggests that the proposed methodology is robust for tested paramters p. It further argues that the different parameterizations tested recover similar signal characteristics, and that the choice for p is less critical. In addition to the previous section, this suggests that the parameter $p=4$ is ideal for the analyzed datasets. Therefore, we fixed this parameter, and applied our method to the prediction of histone function in the next section.

3.3 Histone function prediction using LPCHP

In the second application, we investigated if structural features obtained through LPCHP can identify commonalities between different histone modifications. In particular, we hypothesized that similarity information would be most apparent in the two sets (G_H, G_L). Therefore, we performed hierarchical cluster analysis of the five well characterized histone modifications associated with transcriptional activation: H3K4me1, H3K4me2, H3K4me3, H3K36me3 and PolII as well as three repressive marks: H3K27me3, H3K9me2 and H3K9me3. This analysis (Fig. 3A) indicated a clear clustering between the activating and repressing histone marks.

Motivated by this finding we hypothesized that the LPCHP could be utilized to characterize or predict the function of less well-defined HMs. The LPCHP profiles calculated from the two gene groups (G_H, G_L) were used as features for the HM (Fig. 3B). A training set was built using the eight well-characterized active and repressive chromatin marks, as described above. Note that methylations that did not clearly correlate with gene expression in human CD4$^+$ T-cells, e.g. different methylations of H3K79, H4K20me3 or the variant H2A.Z that was previously associated with activation (TSS) and repression (gene body) depending on the genome region, were not used in the training set.

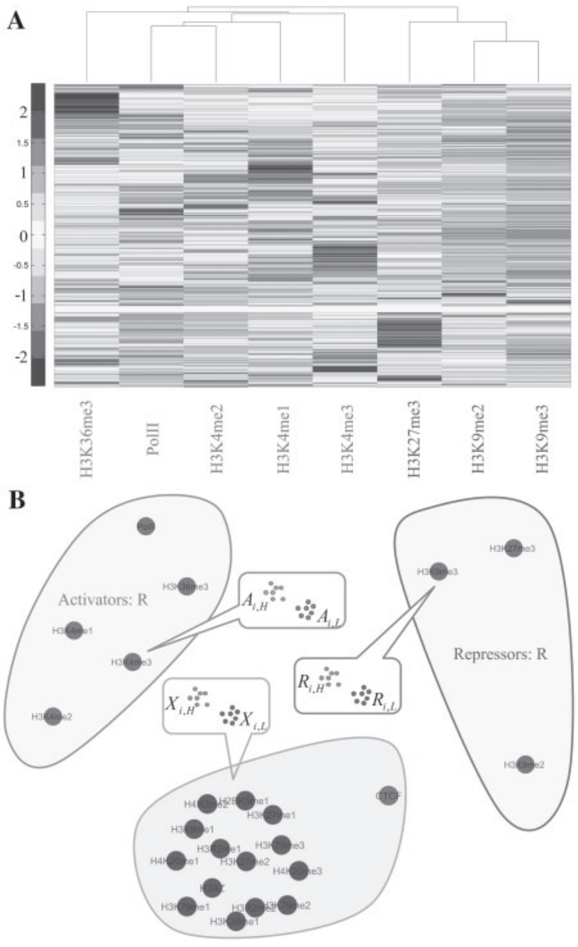

Fig. 3. In (**A**) The LPCHP for a set of seven well characterized histone marks and PolII were analyzed by hierarchical clustering. We show the dendogram for the 4th component of the LPCHP ($p=4$), while all other dendograms are given in Supplementary Figure S5. In each plot a clear clustering between the activating (green) and repressive (red) marks can be seen. In (**B**) we illustrate the histone function prediction as detailed in the text. Note, that in each group of histone modifications, e.g. A, R and X, the distances between nodes were calculated using the force-weighted layout within Cytoscape. Nodes were colored according to the predicted function of the HM they present. We used red for repressive and green for active marks.

The training data were then used to predict the potential function of the other 14 histone ChIP-seq profiles in Barski *et al.* (2007) and Wang *et al.* (2008). Within this set, seven marks were consistent with active chromatin and the other seven marks were associated with repressed chromatin around the TSS. Two of the seven inactive marks, (H3K27me2, H2AZ) were also assigned as such by Barski *et al.*, (2007) and Yu *et al.* (2008) while five (H3K79me2, H3K36me1, H3R2me2, H4K20me3, H4R3me2) were in agreement with their known repressive role in gene expression (Xu *et al.*, 2010; Yu *et al.*, 2008). Within the chromatin marks predicted as active by LPCHP, four (H3K9me1, H2BK5me1, H4K20me1, H3K27me1) were in agreement with Barski *et al.* (2007) and Yu *et al.* (2008) while two (H3K79me1 and H3K79me3) were consistent with the predictions made by Xu *et al.* (2010). The LPCHP approach also predicted an active role for H3R2me1. While this mark was assigned as a modification of repressed genes by Xu *et al*. (2010) and Yu *et al*. (2008), it is known to accumulate across the coding regions of active genes in yeast, where it correlates with active transcription (Kirmizis *et al*., 2009).

Taken together, LPCHP performs robustly in predicting chromatin function when analyzing enrichment profiles at TSS.

4 DISCUSSION

The analysis of enrichment profiles is currently based on read intensities and do not take into account signal structure. We show that modeled signal structures (LPCHP) from a set of 23 ChIP-seq experiments correlate well with gene expression. We also show that LPCHP can be used to gain insights into chromatin function. It is important to note that the binding profiles obtained from HM ChIP-seq are a superposition of signals from the histone modification and underlying location of the nucleosome. Therefore, the reported correlations likely reflect the influence of both factors, and normalization in the presence of data on total H2, H3 and H4 levels is necessary to avoid this bias.

LPCHP is better suited for comparative studies of chromatin function, particular in situations where the signal structure is expected to change, while the signal intensity remains constant. Importantly, nucleosome free regions (NFR) mark general sites of transcription initiation for coding and non-coding RNAs, and are influenced by a combination of DNA sequence and transcription factors (Radman-Livaja and Rando, 2010; Schones *et al.*, 2008). If nucleosome repositioning is assumed in ChIP-seq libraries, our approach is expected to detect the associated peak shift, while no differences would be detected from the signal intensities.

The LPC approach can also be extended for the analysis of other sequencing protocols such as the distinction of nucleosome positioning between methylated and non-methylated CpG islands (Choi, 2010). Incorporating peak shape information will further facilitate the identification of specific TF binding to one of multiple, closely situated binding sites, as typically seen in promoters of critical developmentally regulated genes.

Furthermore, our analysis pipeline automatically reduces noise and handles data scaling and normalization, while obtaining spectral features, which can be further analyzed using a variety of previously developed spectral pattern comparison techniques (Rabiner and Juang, 1993).

5 CONCLUSION

In conclusion, both applications show that the LPCHP and the signal structural that they represent, provide a useful feature that is currently neglected in the literature. Since we show that the LPCHP can be used as an alternative to read intensities, its utility may extend beyond ChIP-seq to other next-generation sequencing applications. It will be particularly useful in situations where read intensity information is either insufficient or non-informative. In addition, the transformation of the original signal into LPC feature vectors is a useful intermediate allowing for the application of various machine learning algorithms. As recently noted, a particular important task in bioinformatics is the identification of all possible chromatin states (Baker, 2011), which could be obtained from a simple vector quantization analysis of the LPCHP. Other potential applications include the identification of enhancer or regulatory regions in the genome. In addition, further methodological development is needed to extend the application of LPCHP to the analysis of more complex genomic features. This includes, for example, the comparison of HM binding structures in gene body regions that are variable in length, and the number and length of exons, introns and enhancers.

ACKNOWLEDGEMENTS

We thank Xiaobo Zhou and Tuan Pham for initial discussions and supervision of D.B.

Funding: Postgraduate research scholarships from the University of New South Wales (to D.B. and M.B.B.); Cancer Institute NSW Early Career and UNSW Vice-Chancellor's Post-Doctoral Fellowships (to J.W.H.W); National Health and Medical Research Council of Australia NHMRC-APP1024364/APP1008515, Australian Research Council ARC- DP0984701 and the Leukaemia Foundation 20080917CEG (for the work in J.E.P.'s laboratory).

Conflict of Interest: none declared.

REFERENCES

Baker,M. (2011) Making sense of chromatin states.*Nat. Methods*, **8**, 717–722.

Barski,A. *et al.* (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.

Chang,C. and Lin,C. (2001) LIBSVM: a library for support vector machines. Available at http://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf (last accessed date May 2011).

Choi,J.K. (2010) Contrasting chromatin organization of CpG islands and exons in the human genome. *Genome Biol.*, **11**, R70.

Costa,I.G. *et al.* (2011) Predicting gene expression in T-cell differentiation from histone modifications and transcription factor binding affinities by linear mixture models. *BMC Bioinformatics*, **12** (Suppl. 1), S29.

Hastie,T. *et al*. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics, Springer, San Francisco, California, USA.

Hobert,O. (2008) Gene regulation by transcription factors and microRNAs. *Science*, **319**, 1785–1786.

Hubert,L. and Arabie,P. (1985) Comparing partitions. *J Classif.*, **2**, 193–218.

Karlic,R. *et al.* (2010) Histone modification levels are predictive for gene expression. *Proc. Natl Acad. Sci. USA*, **107**, 2926–2931.

Kirmizis,A. *et al.* (2009) Distinct transcriptional outputs associated with mono- and dimethylated histone H3 arginine 2. *Nat. Struct. Mol. Biol.*, **16**, 449–451.

Pepke,S. *et al*. (2009) Computation for ChIP-seq and RNA-seq studies. *Nat. Methods*, **6**, S22–S32.

Pinkowski,B. (1993) LPC spectral moments for clustering acoustic transients. *IEEE T. Speech Audi. P.*, **1**, 362–368.

Pruitt,K.D. *et al.* (2009) The consensus coding sequence (CCDS) project: identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res.*, **19**, 1316–1323.

Rabiner,L. and Juang,B.-H. (1993) *Fundamentals of Speech Recognition*. Prentice Hall, Upper Saddle River, NJ, USA.

Radman-Livaja,M. and Rando,O.J. (2010) Nucleosome positioning: how is it established, and why does it matter? *Dev. Biol.*, **339**, 258–266.

Schones,D.E. *et al.* (2008) Dynamic regulation of nucleosome positioning in the human genome. *Cell*, **132**, 887–898.

Silverman,B.W. (1998) *Density Estimation for Statistics and Data Analysis*. Monographs on statistics and applied probability 26. Chapman & Hall/CRC, Boca Raton.

Su,A.I. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA*, **101**, 6062–6067.

Thalamuthu,A. *et al.* (2006) Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics*, **22**, 2405–2412.

Valouev,A.. *et al.* (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat. Meth.*, **5**, 829–834.

Wang,Z. *et al.* (2008) Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat. Genet.*, **40**, 897–903.

Watson,J.D. (2003) *Molecular Biology of the Gene*. Cold Spring Harbor Laboratory Press, New York, USA.

Xu,X. *et al.* (2010) Application of machine learning methods to histone methylation ChIP-Seq data reveals H4R3me2 globally represses gene expression. *BMC Bioinformatics*, **11**, 396.

Young,M.D. *et al.* (2011) ChIP-seq analysis reveals distinct H3K27me3 profiles that correlate with transcriptional activity. *Nucleic Acids Res.*, **39**, 7415–7427.

Yu,H. *et al.* (2008) Inferring causal relationships among different histone modifications and gene expression. *Genome Res.*, **18**, 1314–1324.