Differential principal component analysis of ChIP-seq

Hongkai Ji[a,1], Xia Li[b,c], Qian-fei Wang[b], and Yang Ning[a]

[a]Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD 21205; [b]CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100029, People's Republic of China; and [c]University of Chinese Academy of Sciences, Beijing 100049, People's Republic of China

Edited by Hongyu Zhao, Yale University, New Haven, CT, and accepted by the Editorial Board February 25, 2013 (received for review March 18, 2012)

We propose differential principal component analysis (dPCA) for analyzing multiple ChIP- sequencing datasets to identify differential protein–DNA interactions between two biological conditions. dPCA integrates unsupervised pattern discovery, dimension reduction, and statistical inference into a single framework. It uses a small number of principal components to summarize concisely the major multiprotein synergistic differential patterns between the two conditions. For each pattern, it detects and prioritizes differential genomic loci by comparing the between-condition differences with the within-condition variation among replicate samples. dPCA provides a unique tool for efficiently analyzing large amounts of ChIP-sequencing data to study dynamic changes of gene regulation across different biological conditions. We demonstrate this approach through analyses of differential chromatin patterns at transcription factor binding sites and promoters as well as allele-specific protein–DNA interactions.

allele-specific binding | differential binding | histone modification | next-generation sequencing | RNA-seq

ChIP coupled with high-throughput sequencing (seq) is increasingly used for studying transcription factor (TF) binding sites and histone modifications (HMs) (1–3). A fundamental but unsolved problem in ChIP-seq data analysis is to compare quantitative binding signals between two biological conditions with respect to multiple proteins. This problem is often raised when researchers study changes of gene regulatory programs between different conditions (e.g., normal and cancer cells, different developmental time points). Fig. 1 shows a representative data structure. For each of the two conditions, data for multiple TFs and/or HMs are available. Each protein may have multiple replicate samples carrying information about the biological or technical variation. Given these data, one often asks three questions: (*i*) What are the major patterns of protein–DNA interaction (PDI) differences between the two conditions? (*ii*) How do I detect and prioritize differential genomic loci for follow-up studies? (*iii*) How do I evaluate statistical significance of the observed differences, given the background variation among replicate samples?

These fundamental questions cannot be answered by existing ChIP-seq data analysis tools. Most peak-calling algorithms are developed for finding PDI locations in one cell type (4, 5) (see also *SI Appendix*, *Text S1*). They do not characterize quantitative differences between two cell types. Although one can compare two cell types based on the binary peak calls, this comparison is qualitative and cannot replace a quantitative comparison of the continuous binding signals. For instance, a peak called present in both conditions may have dramatically different binding intensities (6). Analyzing ChIP-seq quantitatively allows one to study differences between conditions better, prioritize genomic loci for follow-up experiments, and predict other genomic signals better (*SI Appendix*, *Text S1* and Fig. S1 *A* and *B*). A few methods can compare binding signals from two conditions (6–9), but they only analyze one protein at a time and do not consider replicate samples. Recently, several methods for analyzing multiple ChIP datasets have been developed for improving peak calling (10–12) and identifying combinatorial binding patterns of multiple proteins within a cell type (13–16). However, none of these methods have been developed for comparing quantitative binding signals of multiple proteins between two biological conditions while also considering the background variation among replicate samples.

We propose to solve this problem by developing differential principal component analysis (dPCA). We consider a scenario in which a list of candidate genomic loci (e.g., DNA motif sites, binding regions obtained from ChIP-seq peak-calling algorithms) is given for analyzing differences (*SI Appendix*, Fig. S1*C*).

Define a dataset to be a collection of replicate samples generated by one laboratory for one particular protein in both conditions (Fig. 1). One simple approach to characterize differences between the two conditions is to analyze each dataset separately to find differential loci, similar to identifying differentially expressed genes from microarray or RNA-seq data (17, 18). Unfortunately, this approach has two major drawbacks. First, analyzing each dataset separately ignores the correlation among proteins that may provide insight on multiprotein synergy. Second, if there are M datasets, this approach will produce M differential loci lists and 3^M combinatorial patterns (because each locus has three possible states in each dataset: up, down, and no change). As M grows, the results will become difficult to report, interpret, and use. For example, if one wants to choose some differential loci to follow up experimentally, which of the M lists or 3^M patterns should be followed up first, given the finite resource?

To address these issues, dPCA integrates unsupervised pattern discovery, dimension reduction, and statistical tests into a single framework. It first summarizes main patterns of differences between the two biological conditions using a small number of differential principal components (dPCs). Each dPC represents a covariation pattern of quantitative signals among multiple proteins. dPCs can simplify description of the data. The analysis then identifies differential genomic loci for each major dPC and prioritizes these loci based on their magnitude of differences. For each locus, statistical significance is evaluated by comparing the between-condition differences with the background variation among the replicate samples. We will demonstrate dPCA using both simulations and real data. dPCA is implemented in ANSI C and is freely available at www.biostat.jhsph.edu/dpca.

Results

dPCA. Consider two biological conditions ($i = 1, 2$), each with M datasets. In condition i, dataset m has K_{im} replicates. There are G genomic loci. Typically $G \gg M$ (Fig. 1). dPCA takes coordinates of these loci and aligned ChIP-seq reads as input. After preprocessing, normalization, and $\log_2$ transform (*SI Appendix*, *Text S1*), PDI intensity for locus g, condition i, dataset m, and replicate k will be summarized into one value: x_{gimk}. We assume that x_{gimk}s are generated by adding independent Gaussian noises ε_{gimk} to true binding levels μ_{gim}. The variance of ε_{gimk}, σ^2, characterizes the variability among replicates and is unknown. Taking an average over replicates gives $\bar{x}_{gim} = \sum_k x_{gimk}/K_{im}$ and

Author contributions: H.J. designed research; H.J. performed research; H.J. and Y.N. contributed new reagents/analytic tools; H.J., X.L., and Q.-f.W. analyzed data; and H.J. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. H.Z. is a guest editor invited by the Editorial Board.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. E-mail: hji@jhsph.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1204398110/-/DCSupplemental.

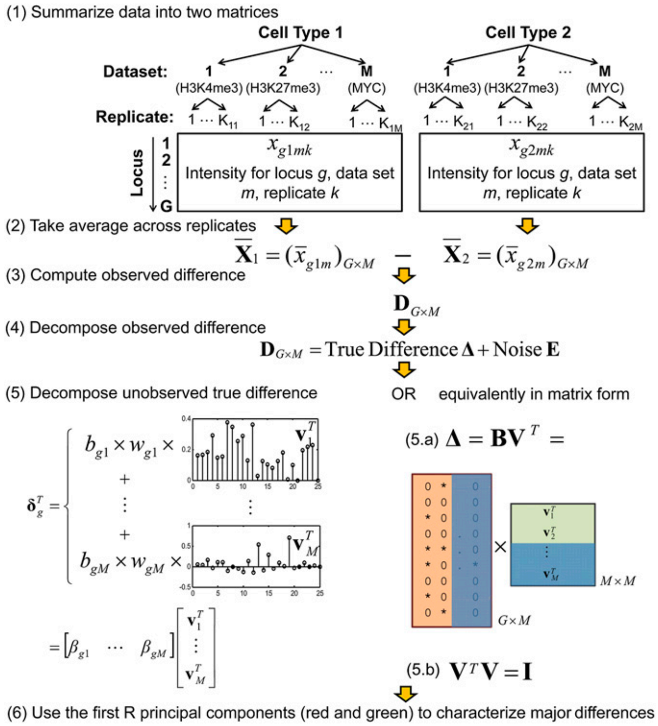
(1) Summarize data into two matrices
Cell Type 1
Cell Type 2
Dataset: 1 2 … M
(H3K4me3) (H3K27me3) (MYC)
Replicate: $1 \cdots K_{11}$ $1 \cdots K_{12}$ $1 \cdots K_{1M}$
$1 \cdots K_{21}$ $1 \cdots K_{22}$ $1 \cdots K_{2M}$
Locus 1 2 ⋮ G
x_{g1mk}
Intensity for locus g, data set m, replicate k
x_{g2mk}
Intensity for locus g, data set m, replicate k
(2) Take average across replicates
$\overline{\mathbf{X}}_1 = (\overline{x}_{g1m})_{G\times M}$ − $\overline{\mathbf{X}}_2 = (\overline{x}_{g2m})_{G\times M}$
(3) Compute observed difference
$\mathbf{D}_{G\times M}$
(4) Decompose observed difference
$\mathbf{D}_{G\times M}$ = True Difference $\boldsymbol{\Delta}$ + Noise $\mathbf{E}$
(5) Decompose unobserved true difference
OR equivalently in matrix form
$\boldsymbol{\delta}_g^T =$
$b_{g1} \times w_{g1} \times$ $\mathbf{v}_1^T$
+
⋮
+
$b_{gM} \times w_{gM} \times$ $\mathbf{v}_M^T$
$= [\beta_{g1} \cdots \beta_{gM}] \begin{bmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_M^T \end{bmatrix}$
(5.a) $\boldsymbol{\Delta} = \mathbf{B}\mathbf{V}^T =$
$G \times M$
$M \times M$
(5.b) $\mathbf{V}^T\mathbf{V} = \mathbf{I}$
(6) Use the first R principal components (red and green) to characterize major differences

Fig. 1. dPCA. The objective of dPCA is to compare two conditions. Each condition has multiple TF or HM ChIP-seq datasets. Each dataset has several replicates. dPCA attempts to characterize differences at a list of user-specified genomic loci. The plot shows the major steps of dPCA.

$d_{gm} = \overline{x}_{g1m} - \overline{x}_{g2m}$. The observed difference between the two conditions for locus g and dataset m is d_{gm}.

Organize d_{gm} into a matrix $\mathbf{D} = (d_{gm})_{G\times M}$. Each row of $\mathbf{D}$ corresponds to a locus, and each column corresponds to a dataset. We consider applications in which, within each column, d_{gm}s can be both positive and negative and fluctuate around zero (*SI Appendix, Text S1* and Fig. S1 *D–G*). We decompose $\mathbf{D}$ into two matrices: $\mathbf{D} = \boldsymbol{\Delta} + \mathbf{E}$, where $\boldsymbol{\Delta} = (\delta_{gm})_{G\times M}$ represents the unobserved true differences between the two conditions and $\mathbf{E} = (e_{gm})_{G\times M}$ corresponds to random sampling noise. Here, $\delta_{gm} = \mu_{g1m} - \mu_{g2m}$, and $e_{gm} \sim N(0, \sigma^2(1/K_{1m} + 1/K_{2m}))$. The e_{gm}s are independent and reflect replicate variation. The gth row of matrices $\mathbf{D}$, $\boldsymbol{\Delta}$, and $\mathbf{E}$, denoted by $\mathbf{d}_g^T$, $\boldsymbol{\delta}_g^T$, and $\mathbf{e}_g^T$, comprises the observed difference, unobserved true difference, and sampling noise at locus g, respectively. Here, T means vector or matrix transpose.

Our primary interest is the unobserved truth $\boldsymbol{\Delta}$. Intuitively, dPCA attempts to characterize $\boldsymbol{\Delta}$ by a few principal components (PCs). This is different from the conventional principal component analysis (PCA) that studies PCs of the observed data matrix $\mathbf{D}$. Distinguishing $\boldsymbol{\Delta}$ from $\mathbf{D}$ is important because it allows one to assess how much variation in the observed $\mathbf{D}$ is due to the underlying truth ($\boldsymbol{\Delta}$) as opposed to the variation among replicate samples ($\mathbf{E}$). Subsequently, one will be able to infer whether the estimated PCs can accurately match the truth, assess the statistical significance of the observed differences, and more efficiently reduce the data dimension. These are functions not provided by PCA.

dPCA is based on assuming that there exist M orthogonal differential patterns $\mathbf{v}_1, \ldots, \mathbf{v}_M$, such that the true difference $\boldsymbol{\delta}_g$ at each locus can be represented by a linear combination:

$$\boldsymbol{\delta}_g = \sum_{j=1}^{M} b_{gj} \times w_{gj} \times \mathbf{v}_j = \sum_{j=1}^{M} \beta_{gj}\mathbf{v}_j = \mathbf{V}\boldsymbol{\beta}_g. \qquad [1]$$

Each $\mathbf{v}_j$ is a $M \times 1$ vector with unitary length, representing a covariation pattern of binding intensities among multiple ChIP-seq datasets (Fig. 1). $\mathbf{V}_{M\times M} = (\mathbf{v}_1, \ldots, \mathbf{v}_M)$ is an orthogonal matrix (i.e., $\mathbf{V}^T\mathbf{V} = \mathbf{I}$). We treat $\mathbf{V}$ as fixed but unknown parameters. Given $\mathbf{V}$, $\boldsymbol{\delta}_g$s are assumed to be randomly and independently generated as follows. First, 0/1 valued binary indicators b_{gj} are independently drawn from Bernoulli distributions with success probability $Pr(b_{gj} = 1) = \pi_j$. Second, real-valued coefficients w_{gj} are drawn independently from some unknown distributions $H_j(w; 0, \tau_j^2)$ with zero mean and unknown variance τ_j^2. The products $\beta_{gj} \equiv b_{gj} \times w_{gj}$ are then used as the coefficients to combine $\mathbf{v}_j$s to generate $\boldsymbol{\delta}_g$. In Eq. **1**, $\mathbf{V}$ is common to all loci, but β_{gj}s are locus-specific. This model implies that each locus can be differential with respect to some patterns (when $\beta_{gj} \neq 0$) but nondifferential for the others ($\beta_{gj} = 0$). For each pattern $\mathbf{v}_j$, the data consist of a mixture of differential and nondifferential loci, and π_j is the prior probability for a locus to be differential.

Group coefficients β_{gj} into a matrix $\mathbf{B} = (\beta_{gj})_{G\times M}$. $\boldsymbol{\beta}_g^T$ is the gth row of the matrix. It contains all coefficients β_{gj}s for locus g. The jth column of $\mathbf{B}$ contains all coefficients for pattern $\mathbf{v}_j$. In matrix form, Eq. **1** is equivalent to $\boldsymbol{\Delta} = \mathbf{B}\mathbf{V}^T$ (Fig. 1). Based on our model, β_{gj}s in column j of $\mathbf{B}$ have a mean $E(\beta_{gj}) = 0$ and variance $Var(\beta_{gj}) = \pi_j\tau_j^2 \equiv \lambda_j$; hence, λ_j characterizes the variation in $\boldsymbol{\Delta}$ contributed by pattern $\mathbf{v}_j$. We assume that λ_js are unequal and, without loss of generality, arranged in descending order $\lambda_1 > \ldots > \lambda_M \geq 0$. Under these assumptions, $Var(\boldsymbol{\delta}_g) = E(\boldsymbol{\Delta}^T\boldsymbol{\Delta}/G) = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T$, where $\boldsymbol{\Lambda} = diag(\lambda_1, \ldots, \lambda_M)$ is a diagonal matrix and λ_js are its diagonal elements. Thus, λ_js are the unique eigenvalues of $Var(\boldsymbol{\delta}_g)$ due to the uniqueness of eigendecomposition, and $\mathbf{v}_j$s are the unique eigenvectors up to a multiplier of ±1, which will not affect the two-sided hypothesis tests below. Each $\mathbf{v}_j$ essentially corresponds to a PC of $Var(\boldsymbol{\delta}_g)$ and is called a dPC. In real data, the first few dPCs often explain the main variation in $\boldsymbol{\Delta}$ (i.e., one can find an integer $R \ll M$ such that $\sum_{j=1}^{R}\lambda_j / \sum_{j=1}^{M}\lambda_j$ is big). Consequently, one can use the first R dPCs instead of all M patterns to summarize the major changes between conditions (i.e., $\boldsymbol{\Delta} \approx \tilde{\mathbf{B}}_{G\times R}\tilde{\mathbf{V}}_{R\times M}^T$, where $\tilde{\mathbf{B}}$ and $\tilde{\mathbf{V}}$ contain the first R columns of $\mathbf{B}$ and $\mathbf{V}$, respectively. Reducing the dimension from M to R can greatly reduce the complexity of data interpretation and make follow-up studies more manageable.

In the model above, $\mathbf{V}$, $\mathbf{B}$, σ^2, π_j, and $H_j(.; 0, \tau_j^2)$ are all unknown. Our primary interest is $\mathbf{V}$ and $\mathbf{B}$. dPCA has three goals: (*i*) find the major differential patterns $\tilde{\mathbf{V}}$; (*ii*) for each locus g and pattern $\mathbf{v}_j$, estimate β_{gj} by projecting data to the estimated $\hat{\mathbf{v}}_j$ (because $\mathbf{v}_j^T\mathbf{d}_g = \mathbf{v}_j^T\boldsymbol{\delta}_g + \mathbf{v}_j^T\mathbf{e}_g = \beta_{gj} + \epsilon_{gj}$) and infer whether the locus is differential or not (i.e., test H_0: $\beta_{gj} = 0$ vs. H_1: $\beta_{gj} \neq 0$); and (*iii*) for each pattern $\mathbf{v}_j$, rank genomic loci based on the magnitude of difference $|\beta_{gj}|$ for follow-up studies. We developed a computationally efficient algorithm to achieve these goals (*Methods*). For the examples below, the algorithm only takes 1–2 min on a laptop computer with a 2.2 GHz central processing unit (CPU) and 4 GB of random access memory after data preprocessing, which takes a much longer time.

To determine which dPCs to report, we project data to each dPC and define a signal-to-noise ratio (SNR) measure $SNR_j = Var(\mathbf{v}_j^T\mathbf{d}_g)/Var(\mathbf{v}_j^T\mathbf{e}_g)$. We estimate SNR_j and report leading dPCs for which $\widehat{SNR}_j > 5$. This is based on observing that the dPC estimates and statistical inference on β_{gj}s are not reliable when the SNR is small (examples I–III).

The basic model above analyzes differences without considering the total amount of absolute binding at each locus. Users often want to analyze differences more specifically at locations where there are significant binding activities that might be easier to interpret or to study experimentally. We provide multiple options to do so. For instance, one can filter out loci not bound in any dataset before dPCA. This and other more sophisticated options are discussed in detail in *SI Appendix, Text S1*.

dPCA is different from factor analysis (*SI Appendix, Text S1*). Unlike methods requiring random initialization or ad hoc choice of parameters (e.g., k-means clustering), dPCA is a deterministic algorithm and patterns discovered by dPCA are reproducible from one investigator to another.

Example I: Analysis of Differential Chromatin Patterns at TF Binding Sites. We first demonstrate dPCA using 18 ENCODE (3) chromatin datasets consisting of 70 ChIP-seq, DNase-seq, and Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE)-seq samples from two cell lines: K562 and human umbilical vein endothelial cell (Huvec) (*SI Appendix*, Table S1). Each dataset had one to three replicates in each cell line. We mapped the MYC (E-box) motif to the human genome and obtained 138,325 MYC motif sites. After excluding sites not associated with significant chromatin signal(s) in any dataset, dPCA was applied to explore differential chromatin patterns at the remaining 58,997 motif sites (Fig. 2 and *SI Appendix*, *Text S1* and Fig. S2).

We begin with asking whether the differential patterns discovered by dPCA are biologically meaningful. The top two dPCs passed the cutoff of SNR > 5. They explained 58.7% and 17.2% of the variance in $\mathbf{\Delta}$, respectively (Fig. 2 *A* and *B*). In dPC1, differences between the two cell types are mainly driven by H3K4me1, H3K4me2, H3K4me3, H3K9ac, and H3K27ac. These HMs are known marks for active transcription or enhancer activities. dPC2 mainly captures the difference in H3K27me3, which is a mark for gene repression. Thus, without using prior knowledge, these two dPCs automatically summarized 18 datasets into two biologically meaningful modules corresponding to gene activation and repression, respectively.

We then asked whether dPCA provides a meaningful way to rank differential loci for follow-up studies. Using independent ENCODE MYC ChIP-seq data, we computed $\log_2$ fold changes ($\log_2 FC$) of MYC ChIP-seq signals between the two cell lines (*SI Appendix*, *Text S1*). Interestingly, although our dPCA analysis did not involve any MYC ChIP-seq data, the coefficients $\hat{\beta}_{g1}$ for dPC1 strongly correlated with the differential MYC ChIP-seq signal (Fig. 2*D*; Pearson's correlation, $\rho = 0.65$). We further defined 6,433 motif sites bound by MYC in at least one cell type and with MYC ChIP-seq $|\log_2 FC| > 1.5$ as true differential MYC binding sites (*SI Appendix*, *Text S1*). Fig. 2*E* shows that a significant fraction of the top motif sites ranked by dPC1 (e.g., 2,895 of 5,000 top sites) were indeed differentially bound by MYC. Importantly, compared with motif site rankings based on each individual dataset (using $|d_{gm}|$ as the ranking criterion), the dPC1 ranking predicted differential MYC binding better (Fig. 2*E* and *SI Appendix*, *Text S1* and Fig. S2*D*). Moreover, if one were to use the best single dataset-based ranking to choose differential loci for follow-up study, one would have to determine which of the 18 datasets is the best. If there were no prior knowledge or independent benchmark data, such as MYC ChIP-seq, this would be difficult. The unsupervised dPCA produced the best ranking without using any prior knowledge. It is able to integrate information automatically from multiple datasets and to prevent one from being overwhelmed by having too many datasets. Unlike dPC1, dPC2 only had a weak negative correlation with differential MYC binding ($\rho = -0.06$; Fig. 2 *D* and *E*). This weak correlation mainly reflects the nature of the H3K27me3 data and could have many possible explanations (*SI Appendix*, *Text S1*). Jointly, dPC1 and dPC2 were able to explain the differential MYC binding slightly better (*SI Appendix*, *Text S1* and Fig. S2 *B* and *E*).

At the 5% false discovery rate (FDR) level, dPCA reported 34,034 (57.7% of 58,997 and 24.6% of 138,325) and 28,379 (48.1% of 58,997 and 20.5% of 138,325) differential motif sites for dPC1 and dPC2, respectively, with 16,906 common sites (Fig. 2*C*). This amounts to a total of 45,507 (77.1% of 58,997 and 32.9% of 138,325) differential sites. Without knowing the truth, it is difficult to evaluate how accurate the dPC and FDR estimates are. To shed light on the performance of these estimates, we performed simulations by retaining the main characteristics of real data, which may deviate from the assumptions made by dPCA (e.g., normality, common σ^2 for all loci). Simulations were performed in different global SNR settings (Fig. 3 and *SI Appendix*, Fig. S3), with details described in *SI Appendix*, *Text S1*. Fig. 3 provides a representative example to illustrate the results. Fig. 3*A* shows the estimated SNR for each dPC. Fig. 3*B* shows the accuracy of the $\mathbf{v}_j$ estimates. The accuracy was measured by the cosine distance $d(\mathbf{v}_j, \hat{\mathbf{v}}_j) = 1 - |\langle \mathbf{v}_j, \hat{\mathbf{v}}_j \rangle|$. A small d means accurate. The vertical bars show the variability of d, measured by its SD across 10 independent simulations. Fig. 3*C* shows the error of λ_j estimates (i.e., $\hat{\lambda}_j - \lambda_j$). Fig. 3*D* shows the percentage of variance explained by the top dPCs. Fig. 3 *E*–*G* compares the true FDR with the estimated FDR for the first three dPCs, respectively. These results show that the accuracy of dPC estimates decreases with decreasing $\widehat{SNR}_j$ (Fig. 3 *A* and *B*). For dPCs with $\widehat{SNR}_j > 10$, the estimated $\hat{\mathbf{v}}_j$ matched the true $\mathbf{v}_j$ well and the claimed FDR provided reasonable estimates for the true FDR for testing $\beta_{gj} = 0$ vs. $\beta_{gj} \neq 0$ even if the data were projected to $\hat{\mathbf{v}}_j$ instead of $\mathbf{v}_j$ (Fig. 3 *B* and *E*). The performance deteriorated as $\widehat{SNR}_j$ decreased (Fig. 3 *B* and *F*). For dPCs with $\widehat{SNR}_j < 5$, the estimates were off the mark (Fig. 3 *B* and *G*). For dPCs with a small SNR, $\hat{\mathbf{v}}_j$ also had high variability, as evidenced by the wide error bars in Fig. 3*B* and additional simulations in *SI Appendix*, *Text S1* and Fig. S4 *A*–*C*, which show that if two laboratories independently generate similar data and run dPCA, they may not discover the same patterns, causing a reproducibility issue. Intuitively, when SNR_j is small, the geometric direction represented by $\hat{\mathbf{v}}_j$ in the $\mathbf{R}^M$ space can be easily rotated by noise. When $\hat{\mathbf{v}}_j$ is biased, it is not reliable to draw conclusions about $\beta_{gj} = \mathbf{v}_j^T \boldsymbol{\delta}_g$ by projecting data to $\hat{\mathbf{v}}_j$, because $\mathbf{v}_j$ and $\hat{\mathbf{v}}_j$ represent different geometric directions. We observed similar phenomena in all simulations (*SI Appendix*, Fig. S3). Therefore, although all nonzero elements in $\mathbf{B}$ are assumed to be true differences not explained by cross-sample variation, we only report dPCs with $\widehat{SNR}_j > 5$ (dPC1 and dPC2 in this example), because the differential patterns with a smaller SNR cannot be accurately and

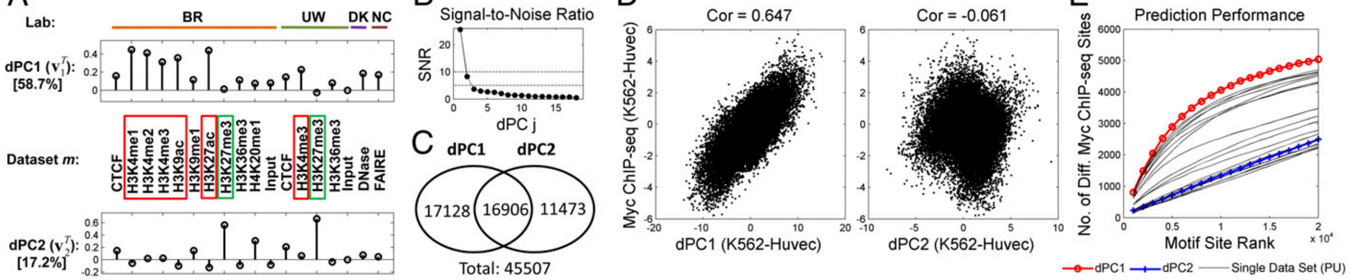

Fig. 2. dPCA analysis of MYC motif sites. (*A*) First two dPCs. Vertical bars show values of $\hat{v}_{jm}$ in $\hat{\mathbf{v}}_j$. The data are from four laboratories (BR, Broad Institute; UW, University of Washington; DK, Duke University; and NC, University of North Carolina). The percentages of variance explained are shown in square brackets. (*B*) Estimated SNR_j for each dPC. (*C*) Numbers of differential loci. (*D*) MYC ChIP-seq $\log_2$ fold changes at the 58,997 analyzed motif sites are plotted against dPC1 ($\hat{\beta}_{g1}$) and dPC2 ($\hat{\beta}_{g2}$), respectively. Cor, Pearson correlation coefficient. (*E*) Numbers (No.) of top-ranked motif sites that are truly differentially (Diff.) bound by MYC are shown for different ranking methods (*SI Appendix*, *Text S1*). PU, the Peak Union method (*SI Appendix*, Text S1).

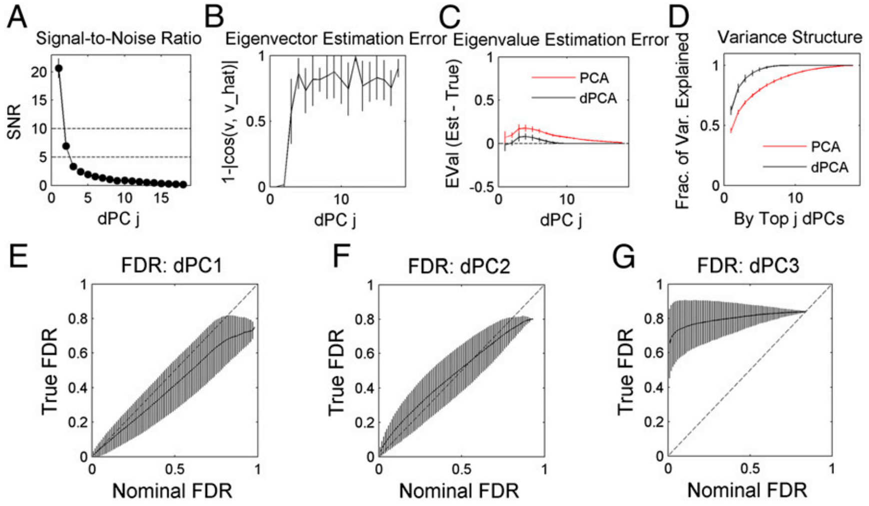

Fig. 3. Simulation results (π_0 = 0.25; *SI Appendix, Text S1*). (*A*) Estimated SNR for each dPC. (*B*) Accuracy of $\mathbf{v}_j$ estimates, measured by the cosine (cos) distance. (*C*) Error of eigenvalue (EVal) estimates (Est) is $\hat{\lambda}_j - \lambda_j$. (*D*) Percentage of variance (Var.) explained by top dPCs (dPCA) or PCs (PCA). Frac., Fraction. (*E*–*G*) True FDR at different levels of the estimated FDR for the first three dPCs. All plots show the average performance of 10 simulations. Vertical bars indicate ±1 SD of the 10 simulations.

reproducibly discovered. In the real data, dPC2 has SNR > 5. Based on the simulations, it is very likely to be a true differential pattern despite its weak correlation with differential MYC binding.

In our data, the top two dPCs had patterns similar to the top two PCs in PCA (Fig. 2*A* and *SI Appendix*, Fig. S2*G*). However, the top two PCs in PCA only explained 57% of the variance in **D**, whereas the top two dPCs explained 76% of the variance in **Δ**. To explain the ≥76% of the variance in **D**, PCA needs six PCs. To reduce dimension, the conventional PCA often chooses the number of PCs based on the percentage of variance explained. Using this criterion, dPCA is more efficient for dimension reduction. This is confirmed by simulations showing that the eigenvalues in PCA tend to be bigger than those in dPCA (Fig. 3*C*), which results in a smaller percentage of variance explained by the top PCs (Fig. 3*D*; an intuitive explanation is provided in *SI Appendix, Text S1*). Unlike PCA, dPCA also provides SNR_j to help one choose which dPCs to report based on judging whether the dPC and FDR estimates are close to the truth and whether dPCs are reproducible in future studies.

Our analysis suggests that one can combine motif analysis with surrogate experiments, such as HM ChIP-seq to infer dynamic changes of TF binding. Analyses of several other TFs, cell types, and data combinations confirmed this observation (*SI Appendix*, Table S1 and Fig. S2 *L*–*N*). Performing ChIP-seq experiments for all TFs is currently not feasible due to a lack of antibodies and the high cost. However, good antibodies for many HMs are available, and among the 1,400+ human TFs, ∼500 have known DNA binding motifs. Therefore, dPCA analysis of multiple surrogate datasets (e.g., HM ChIP-seq) provides a solution to unsupervised characterization of gene regulation dynamics, and it allows one to infer differential binding of many TFs simultaneously using the same set of experiments. Unlike several recent studies that use surrogates to predict TF binding in one condition (19, 20), dPCA allows one to predict dynamic changes of TF binding across conditions.

Example II: Analysis of Differential Promoters. We also analyzed 24,376 human promoters using the same 18 datasets in K562 and Huvec lines (Fig. 4 and *SI Appendix, Text S1*, Table S1, and Fig. S5). Applying dPCA to the 22,368 promoters bound in at least one dataset, two dPCs passed the cutoff of $\widehat{SNR}_j > 5$. They were similar to the ones found in the MYC analysis, except that H3K4me1 played a weaker role in dPC1 in the promoter analysis (Figs. 2*A* and 4*A*). This is consistent with the knowledge that H3K4me1 preferentially marks enhancers rather than promoters (21). At the 5% FDR level, 16,990 (76.0% of 22,368 and 69.7% of 24,376) and 13,735 (61.4% of 22,368 and 56.4% of 24,376) differential promoters (common = 10,818, total = 19,907) were found for dPC1 and dPC2, respectively, reflecting a global change of chromatin landscape between the two cell lines (Fig. 4*B*). Simulations again show that the dPC and FDR estimates were reasonable when $\widehat{SNR}_j > 10$ and clearly biased when $\widehat{SNR}_j < 5$ (*SI Appendix*, Fig. S3).

The dPC1 coefficients $\hat{\beta}_{g1}$ strongly correlated with differential gene expression (DE) determined by RNA-seq (Fig. 4*C*; ρ = 0.67), which is an independent technology. Promoter ranking based on dPC1 predicted DE better than or as good as rankings based on each individual dataset (Fig. 4*D* and *SI Appendix*, Fig. S5*D*). Again, even though some datasets individually performed comparably to dPC1, in a hypothetical future application, where no prior knowledge or benchmark data are available, determining which individual dataset can provide the best ranking, and hence should be used to choose differential loci for follow-up studies, remains difficult. In that scenario, dPCA will provide a solution to integrating information automatically from multiple datasets to produce optimal or near-optimal ranking. dPC2 had a weak negative correlation with DE (*SI Appendix*, Fig. S5*C*). However, dPC1 and dPC2 jointly explained more DE than each dPC alone (*SI Appendix*, Fig. S5*E*). When promoters were grouped into nine classes based on their dPC1 and dPC2 differential states, the classes in which $\hat{\beta}_{g1}$ (i.e., dPC1) and $\hat{\beta}_{g2}$ (i.e., dPC2) had opposite signs had both the largest magnitude of DE (*SI Appendix*, Fig. S5*G*) and the strongest correlation between $\hat{\beta}_{g1}$ and DE (Fig. 4*E*). This is consistent with the activation and repression nature of dPC1 and dPC2.

Example III: Analysis of Allele-Specific Events. ChIP-seq provides new opportunities to study allele-specific binding (ASB) and HM (22–24). ASB detection often suffers from low statistical power because only reads mapped to heterozygote SNPs contain allelic information. Also, whether or how ASB of different proteins is correlated is often unknown. One can treat the two alleles, the allele consistent with the reference genome and the nonreference allele, as paired samples from two biological conditions. dPCA can be modified to handle the paired sample data (*SI Appendix, Text S1*). Using the modified dPCA, we analyzed ASB in 20 ChIP-seq datasets (44 samples) from the ENCODE GM12878 cells (*SI Appendix*, Table S1). Genotypes for a collection of 5,504 heterozygote SNPs were obtained from a study by Rozowsky et al. (23). After removing various read mapping biases (22, 24) and applying dPCA to 2,584 bound SNPs (*SI Appendix, Text S1* and Fig. S6 *A*–*E*), one dPC passed the cutoff of $\widehat{SNR}_j > 5$ (Fig. 5 *A* and *B*). This dPC is mainly driven by correlated ASB of H3K27ac, H3K4me2, H3K4me3, H3K9ac, Pol2, and c-Myc (Fig. 5*A*), and it positively correlates with allele-specific expression (ASE) (*SI Appendix, Text S1* and Fig. S6*E*). At the 5% FDR level, 725 (28.1% of 2,584 and 13.2% of 5,504) SNPs were differential for dPC1. Simulations confirmed that the

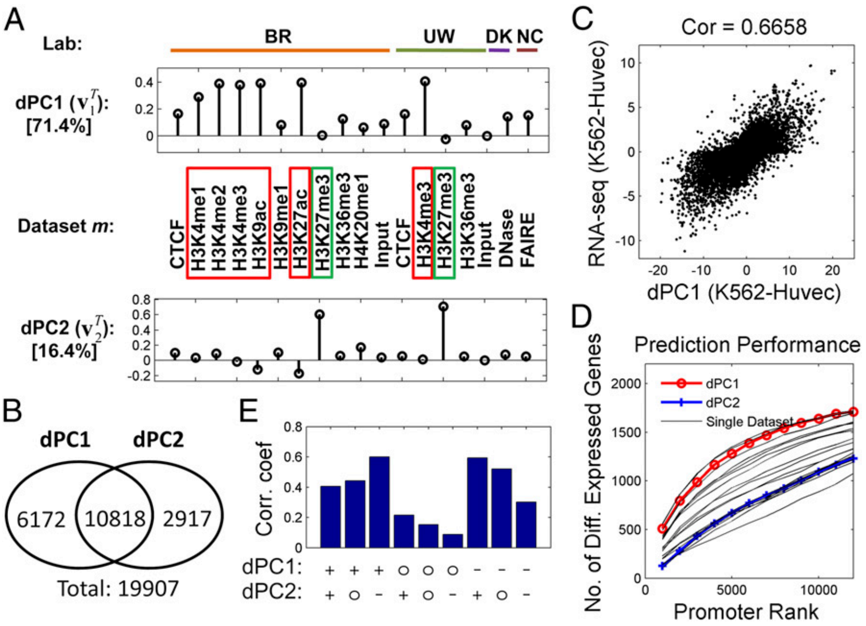

Fig. 4. dPCA analysis of human promoters. (*A*) First two dPCs, explaining 71.4% and 16.4% of variance, respectively. The data are from four laboratories (BR, UW, DK, and NC). (*B*) Numbers of differential loci. (*C*) DE measured by log$_2$ RNA-seq fold change is plotted against dPC1 ($\hat{\beta}_{g1}$). (*D*) Numbers of top-ranked promoters that are differentially expressed. (*E*) Promoters are grouped into nine classes based on dPC1 and dPC2 differential status (+, up in K562; –, down in K562; ○, nondifferential). For each class, the correlation coefficient (Corr. coef) between dPC1 and DE is shown.

dPC and FDR estimates were reasonably accurate if $\widehat{SNR}_j > 10$ and biased when $\widehat{SNR}_j < 5$ (*SI Appendix*, Fig. S3). In real data, $\widehat{SNR}_j$ was between 5 and 10. The $\mathbf{v}_1$ estimate is expected to be slightly biased. This will not affect its usefulness for ranking SNPs, but the FDR estimates may be inaccurate.

We benchmarked SNP ranking in two ways. First, GM12878 is a female. Due to X-inactivation, only one allele of chromosome X (chrX) is expected to be active. Here, chrX refers to nonpseudoautosomal regions of the X chromosome. We therefore compared different ranking methods based on counting how many top-ranked SNPs were in chrX (Fig. 5*C* and *SI Appendix*, Fig. S6*F*). Second, using independent RNA-seq data, we obtained exonic SNPs with ASE (*SI Appendix*, *Text S1*). We compared different methods by counting how many top-ranked SNPs were in the neighborhood of exonic ASE SNPs (Fig. 5*D* and *SI Appendix*, Fig. S6*G*). We also did the same analysis after excluding all SNPs in chrX (*SI Appendix*, Fig. S6*H*). In all analyses, dPC1 predicted ASB better than the rankings based on individual datasets. Thus, dPCA not only allows one to explore the unknown correlation patterns of ASB across multiple proteins but improves ASB detection by using this correlation to integrate information from multiple datasets. Sometimes, the improvement can be significant. For instance, suppose one only has the nine datasets from the Broad Institute; then, the best ranking based on individual datasets shown in Fig. 5*E* and *SI Appendix*, Fig. S5*I* only detected 54 chrX SNPs among the top 500 SNPs, whereas dPCA on these nine datasets detected 69 chrX SNPs (28% improvement).

Functional Interpretation and Absolute Binding. After dPCA, one may use other existing "omics" data to help with interpreting dPCs if their biological meanings are not immediately clear by looking at the $\hat{\mathbf{v}}_j$ patterns (*SI Appendix*, *Text S1*). For instance, in both the MYC and promoter examples, analyses of enriched gene sets were able to connect dPC1 and dPC2 to gene activation

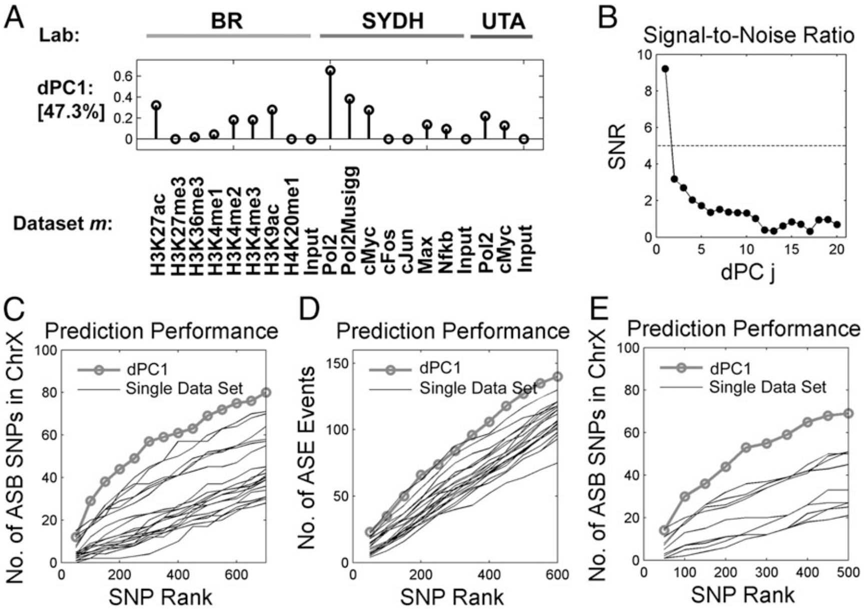

Fig. 5. dPCA analysis of ASB. (*A*) First dPC, explaining 47.3% of the total variance. SYDH, Stanford/Yale/Davis/Harvard; UTA, University of Texas, Austin. (*B*) Estimated SNR. (*C*) Number of top-ranked SNPs that are chrX SNPs. (*D*) Number of top-ranked SNPs that are in the neighborhood of exonic ASE SNPs. (*E*) Analysis in *C* was repeated by using only the nine datasets from the BR.

and repression, respectively (*SI Appendix*, Text S1 and Fig. S2 *H* and *K* and Fig. S5 *J* and *M*).

In all examples, we also repeated the dPCA analyses by incorporating the absolute binding information in different ways to identify differences that co-occur with significant binding activities. These analyses produced similar dPC patterns and largely similar biological findings (*SI Appendix*, Text S1, Fig. S2 *C*, *F*, and *H–K*, Fig. S4 *J–L*, S5 *F*, *H*, and *J–M*, Fig. S6 *L–N*, and Fig. S7).

Discussion

dPCA provides a unique tool that integrates unsupervised pattern discovery, dimension reduction, and statistical tests to explore and summarize concisely the major quantitative differences between two conditions in multiple datasets. The computational efficiency and broad applicability make it very suitable for exploratory analysis of large ChIP-seq data. In principle, one may also use it to analyze other data types, such as RNA-seq. dPCA rankings of differential loci can guide design of follow-up experiments. Patterns discovered by dPCA may inform directions for improving analytical tools in various applications. For example, the correlation patterns found in the ASB analysis may provide a basis for developing new specialized tools to optimize the ASB detection power. The statistical tests in the current dPCA are based on model assumptions, such as normality and equal variance (i.e., common σ^2 for all loci and datasets), which only provide a first-order approximation to the real data. Therefore, instead of providing rigorous FDR control, the tests in dPCA often are "approximate" in nature. Empirically, this approximation worked well in our test data (*SI Appendix*, *Text S1* and Figs. S3 and S4 *E–I*). In the future, the statistical tests may be improved by incorporating better data distribution assumptions tailored to specific applications.

dPCA attempts to find major patterns of differences in the data and the associated loci. Patterns with a small SNR cannot be reliably discovered, and therefore are not reported (*SI Appendix*, *Text S1*). Thus, dPCA reports main differences rather than all differences. We implicitly assume that there are some common patterns shared by many loci. If no such pattern exists, or if one wants to study loci with unique patterns, dPCA may not directly help. For detecting all loci and loci with unique patterns, a simple approach is to detect differential loci in each dataset (e.g., by *t* test), take their union, and find those not reported by dPCA (*SI Appendix*, *Text S1*). dPCA uses replicate variability to help with statistical inference. When there is no replicate, a variant of dPCA may be used by introducing additional assumptions (*SI Appendix*, *Text S1* and Fig. S8). In practice, dPCs can be used or interpreted either separately or jointly depending on the available resources, and one may use other types of omics data (e.g., gene sets) to help with interpreting dPCs (*SI Appendix*, *Text S1*). Currently, absolute binding is handled by dPCA through pre- and postprocessing. How to integrate the absolute binding optimally into the model to improve the analysis of differences is still an open problem worth further investigation (*SI Appendix*, *Text S1*). *SI Appendix*, *Text S1* also includes discussions about the zero mean [i.e., $E(\boldsymbol{\delta}_g) = 0$] and equal variance (i.e., common σ^2) assumptions in dPCA. We show that these assumptions, although not perfect, are reasonable and can produce useful results.

Our data show that it is feasible to infer differential TF binding without ChIP-seq data for the TF of interest and to improve ASB analysis by exploiting correlation among multiple datasets. These examples not only demonstrate the value of dPCA but highlight the importance of developing new tools for integrative analysis of ChIP-seq data.

Methods

Data processing and analysis details for the three examples are provided in *SI Appendix*, *Text S1*. Below, we outline the dPCA algorithm and leave the mathematical details in *SI Appendix*, *Text S1*.

i) Estimate **V**. We first estimate σ^2 using replicate information. Then, $E(\mathbf{\Delta}^T\mathbf{\Delta}/G) = E(\mathbf{D}^T\mathbf{D}/G) - E(\mathbf{E}^T\mathbf{E}/G) \approx \mathbf{D}^T\mathbf{D}/G - \hat{\sigma}^2\mathbf{\Omega}$, where $\mathbf{\Omega}$ = *diag*((1/K_{11} + 1/K_{21}), …, (1/K_{1M} + 1/K_{2M})) is a diagonal matrix. We use the eigenvalues $\hat{\lambda}_j$ and eigenvectors $\hat{\mathbf{v}}_j$ of the estimated $E(\mathbf{\Delta}^T\mathbf{\Delta}/G)$ to estimate λ_j and $\mathbf{v}_j$. The proportion of variance explained by the *j*th dPC is computed as $\hat{\lambda}_j/\sum_{j'}\hat{\lambda}_{j'}$.

ii) Infer β_{gj}. We have $\mathbf{v}_j^T\mathbf{d}_g = \mathbf{v}_j^T\boldsymbol{\delta}_g + \mathbf{v}_j^T\mathbf{e}_g = \beta_{gj} + \epsilon_{gj}$, where $\epsilon_{gj} \sim N(0, \sigma^2\mathbf{v}_j^T\mathbf{\Omega}\mathbf{v}_j)$. If $\mathbf{v}_j$ is known, one could estimate β_{gj} by $\mathbf{v}_j^T\mathbf{d}_g$ and test whether β_{gj} is zero by comparing the *t*-statistic $T_{gj} = \mathbf{v}_j^T\mathbf{d}_g/\sqrt{\hat{\sigma}^2\mathbf{v}_j^T\mathbf{\Omega}\mathbf{v}_j}$ with a *t*-distribution. This yields a two-sided *P* value p_{gj}. In reality, $\mathbf{v}_j$ is unknown. Thus, we project $\mathbf{d}_g$ to the estimated $\mathbf{v}_j$ to obtain $\hat{\beta}_{gj} = \hat{\mathbf{v}}_j^T\mathbf{d}_g$ and $\hat{T}_{gj} = \hat{\mathbf{v}}_j^T\mathbf{d}_g/\sqrt{\hat{\sigma}^2\hat{\mathbf{v}}_j^T\mathbf{\Omega}\hat{\mathbf{v}}_j}$. We obtain the estimated *P* value $\hat{p}_{gj}$ by comparing $\hat{T}_{gj}$ with a *t*-distribution. Subsequently, for each dPC, $\hat{p}_{gj}$s are converted to FDRs using the method of Storey and Tibshirani (25). Our simulations show that if the SNR for the *j*th dPC is big enough, the FDR computed using $\hat{p}_{gj}$ can estimate the true FDR for testing $\beta_{gj} = 0$ reasonably well even if the data are projected to $\hat{\mathbf{v}}_j$ instead of $\mathbf{v}_j$.

iii) For each pattern $\mathbf{v}_j$, rank genomic loci based on $|\hat{T}_{gj}|$.

iv) Determine which dPCs to report. We report dPCs for which $\widehat{SNR}_j = \hat{\mathbf{v}}_j^T(\mathbf{D}^T\mathbf{D}/G)\hat{\mathbf{v}}_j/(\hat{\sigma}^2\hat{\mathbf{v}}_j^T\mathbf{\Omega}\hat{\mathbf{v}}_j) > 5$.

ACKNOWLEDGMENTS. This research is supported by National Institutes of Health Grant R01HG006841.

1. Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science* 316(5830):1497–1502.
2. Barski A, et al. (2007) High-resolution profiling of histone methylations in the human genome. *Cell* 129(4):823–837.
3. Dunham I, et al.; ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489(7414):57–74.
4. Laajala TD, et al. (2009) A practical comparison of methods for detecting transcription factor binding sites in ChIP-seq experiments. *BMC Genomics* 10:618.
5. Wilbanks EG, Facciotti MT (2010) Evaluation of algorithm performance in ChIP-seq peak detection. *PLoS ONE* 5(7):e11471.
6. Xu H, Wei CL, Lin F, Sung WK (2008) An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data. *Bioinformatics* 24(20):2344–2349.
7. Taslim C, et al. (2009) Comparative study on ChIP-seq data: Normalization and binding pattern characterization. *Bioinformatics* 25(18):2334–2340.
8. Johannes F, et al. (2010) Comparing genome-wide chromatin profiles using ChIP-chip or ChIP-seq. *Bioinformatics* 26(8):1000–1006.
9. Shao Z, Zhang Y, Yuan GC, Orkin SH, Waxman DJ (2012) MAnorm: A robust model for quantitative comparison of ChIP-Seq data sets. *Genome Biol* 13(3):R16.
10. Choi H, Nesvizhskii AI, Ghosh D, Qin ZS (2009) Hierarchical hidden Markov model with application to joint analysis of ChIP-chip and ChIP-seq data. *Bioinformatics* 25(14):1715–1721.
11. Wu H, Ji H (2010) JAMIE: Joint analysis of multiple ChIP-chip experiments. *Bioinformatics* 26(15):1864–1870.
12. Chen Y, et al. (2011) MM-ChIP enables integrative analysis of cross-platform and between-laboratory ChIP-chip or ChIP-seq data. *Genome Biol* 12(2):R11.
13. Hon G, Ren B, Wang W (2008) ChromaSig: A probabilistic approach to finding common chromatin signatures in the human genome. *PLOS Comput Biol* 4(10):e1000201.
14. Jaschek R, Tanay A (2009) Spatial clustering of multivariate genomic and epigenomic information. *Proceedings of the 13th Annual International Conference on Research in Computational Molecular Biology*, ed Batzoglou S (Springer-Verlag, Berlin, Heidelberg), pp 170–183.
15. Ernst J, Kellis M (2010) Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* 28(8):817–825.
16. Kharchenko PV, et al. (2011) Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. *Nature* 471(7339):480–485.
17. Smyth GK (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 3: Article3.
18. Anders S, Huber W (2010) Differential expression analysis for sequence count data. *Genome Biol* 11(10):R106.
19. Pique-Regi R, et al. (2011) Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res* 21(3):447–455.
20. Boyle AP, et al. (2011) High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Res* 21(3):456–464.
21. Heintzman ND, et al. (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* 39(3):311–318.
22. Degner JF, et al. (2009) Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics* 25(24):3207–3212.
23. Rozowsky J, et al. (2011) AlleleSeq: Analysis of allele-specific expression and binding in a network framework. *Mol Syst Biol* 7:522.
24. Skelly DA, Johansson M, Madeoy J, Wakefield J, Akey JM (2011) A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. *Genome Res* 21(10):1728–1737.
25. Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. *Proc Natl Acad Sci USA* 100(16):9440–9445.