Corrections

NEUROBIOLOGY

The authors note that Figure 3 appeared incorrectly. The corrected figure and its legend appear below. This error does not affect the conclusions of the article.

Fig. 3. Upper and lower attachments of the tip link. (A and B) Freeze-etch images of tip-link upper insertions in guinea pig cochlea (A) and (left to right) two from guinea pig cochlea, two from bullfrog sacculus, and two from guinea pig utriculus (B). Each example shows pronounced branching. (C and D) Freeze-etch images of the tip-link lower insertion in stereocilia from bullfrog sacculus (C) and guinea pig utriculus (D); multiple strands (arrows) arise from the stereociliary tip. (E) Freeze-fracture image of stereociliary tips from bullfrog sacculus; indentations at tips are indicated by arrows. (Scale bars: A = 100 nm, B = 25 nm; C–E = 100 nm.)
**BIOCHEMISTRY**


 The authors note that, due to a printer’s error, the author name Clemens Vonrehin should instead appear as Clemens Vonrhein. The corrected author line appears below. The online version has been corrected.

 Sébastien Igonet, Marie-Christine Vaney, Clemens Vonrhein, Gérard Bricogne, Enrico A. Stura, Hans Hengartner, Bruno Eschli, and Félix A. Rey

**BIOPHYSICS AND COMPUTATIONAL BIOLOGY, STATISTICS**

 Correction for “Differential principal component analysis of ChIP-seq,” by Hongkai Ji, Xia Li, Qian-fei Wang, and Yang Ning, which appeared in issue 17, April 23, 2013, of *Proc Natl Acad Sci USA* (110:6789–6794; first published April 8, 2013; 10.1073/pnas.1204398110).

 The authors note the following statement should be added to the Acknowledgments: “Q.-F.W. is supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDA01010305).”
We propose differential principal component analysis (dPCA) for analyzing multiple ChIP-seq 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-seq 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.

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 loci 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. SI 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. S1C).

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 SM 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 SM 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 Nm replicates. There are G genomic loci. Typically G ≫ M (Fig. 1). dPCA takes coordinates of these loci and aligned ChIP-seq reads as input. After preprocessing, normalization, and log2 transform (SI Appendix, Text S1), PDI intensity for locus g, condition i, dataset m, and replicate k will be summarized into one value: xigmk. We assume that xigmks are generated by adding independent Gaussian noises σgkmk to true binding levels μgm. The variance of xigmks, σ, characterizes the variability among replicates and is unknown. Taking an average over replicates gives xigm = ∑k xigmk/Nmk and p-value of difference between conditions is given by tigm = xigm − xigm/σgkmk

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.

1To 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.

www.pnas.org/cgi/doi/10.1073/pnas.1204398110

PNAS | April 23, 2013 | vol. 110 | no. 17 | 6789–6794
(1) Summarize data into two matrices

<table>
<thead>
<tr>
<th>Dataset:</th>
<th>Cell Type 1</th>
<th>Cell Type 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Replicate:</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(H(klik) [H(klik)])</td>
<td>(M(klik))</td>
</tr>
<tr>
<td></td>
<td>(M(klik)) [H(klik)]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(H(klik)) [M(klik)]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(M(klik))</td>
<td>(H(klik))</td>
</tr>
</tbody>
</table>

(2) Take average across replicates

\[ \bar{X}_1 = (\bar{x}_{1m})_{G \times M} \]

(3) Compute observed difference

\[ \bar{X}_1 = (\bar{x}_{12m})_{G \times M} \]

(4) Decompose observed difference

\[ D_{G \times M} = \text{True Difference} \lambda \times \text{Noise} \ E \]

(5) Decompose unobserved true difference

\[ \delta_i = b_i \times w_i \times \delta_i \]

(6) Use the first R principal components (red and green) to characterize major differences

\[ VV = I \]

**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} = \bar{x}_{1m} - \bar{x}_{2m}. The observed difference between the two conditions for locus g and dataset m is d_{gm}.

Organize d_{gm} into a matrix D = (d_{gm})_{G \times M}. Each row of D corresponds to a locus, and each column corresponds to a dataset. We consider applications in which, within each column, d_{gm} is both positive and negative and fluctuates around zero (SI Appendix, Text S1 D-G). We decompose D into two matrices: D = \Delta + E, where \Delta = (\delta_{gm})_{G \times M} represents the unobserved true differences between the two conditions and E = (\epsilon_{gm})_{G \times M} corresponds to random sampling noise. Here, \delta_{gm} = \sum b_i w_i \delta_i and \epsilon_{gm} \sim N(0, \sigma^2(1/K_{1m} + 1/K_{2m})). The \epsilon_{gm} are independent and reflect replicate variation. The gth row of matrices \Delta, M, and E, denoted by \delta_g, \delta_g^2, and \epsilon_g^2, 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 \Delta. Intuitively, dPCA attempts to characterize \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 D. Distinguishing \Delta from D is important because it allows one to assess how much variation in the observed D is due to the underlying truth (\Delta) as opposed to the variation among replicate samples (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 v_1, \ldots, v_M such that the true difference \delta at each locus can be represented by a linear combination:

\[ \delta_g = \sum_{j=1}^{M} b_j w_j \times v_j = \sum_{j=1}^{M} \beta_j v_j = \mathbf{V} \beta_g \]

Each 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}_{GM} = (v_1, \ldots, v_M) is an orthogonal matrix (i.e., \mathbf{V}^T \mathbf{V} = I). We treat \mathbf{V} as fixed but unknown parameters. Given \mathbf{V}, \delta_g are assumed to be random and independently generated as follows. First, 0/1 valued binary indicators b_j are independently drawn from Bernoulli distributions with success probability P(\beta_j = 1) = \pi_j. Second, real-valued coefficients w_j are drawn independently from some unknown distributions H(w_j | \tau_j^2) with zero mean and unknown variance \tau_j^2. The products \beta_j \equiv b_j \times w_j are then used as the coefficients to combine vs to generate \delta_g. In Eq. 1, \mathbf{V} is common to all loci, but \beta_g are locus-specific. This model implies that each locus can be differential with respect to some patterns (when \beta_j \neq 0) but non-differential for the others (\beta_j = 0). For each pattern v_j, the data consist of a mixture of differential and nondifferential loci, and \pi_j is the prior probability for a locus to be differential.

Group coefficients \beta_j in a matrix B = (\beta_{gj})_{G \times M}. \beta_j is the gth row of the matrix. It contains all coefficients \beta_g for locus g. The jth column of B contains all coefficients for pattern v_j. In matrix form, Eq. 1 is equivalent to \Delta = \mathbf{BV}^T (Fig. 1). Based on our model, \beta_{gj} in column j of B have a mean E(\beta_{gj}) = 0 and variance Var(\beta_{gj}) = \pi_j \tau_j^2 \equiv \lambda_j; hence, \lambda_j characterizes the variation in the data contributed by pattern v_j. We assume that \lambda_j are unequal and, without loss of generality, arranged in descending order \lambda_1 > \lambda_2 \geq \cdots \lambda_M > 0. Under this assumption, \lambda_j is interpreted as the unique eigenvalue of \mathbf{V}^T \mathbf{V} due to the uniqueness of eigendecomposition, and v_j are the unique eigenvectors up to a multiplier of \pm 1, which will not affect the two-sided hypothesis tests below. Each v_j essentially corresponds to a PC of \mathbf{V} \mathbf{V}^T and is called a dPCA. In real data, the first few dPCs often explain the main variation in D (i.e., one can find an integer R \leq M such that \sum_{j=R+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., \Delta \approx B_{1R} \mathbf{V}_{1R}^T; where B and \mathbf{V} contain the first R columns of 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, V, B, \sigma^2, \tau_j, and H(\cdot;\tau_j) are all unknown. Our primary interest is V and B. dPCA has three goals: (i) find the major differential patterns V; (ii) for each locus g and pattern v_j, estimate \beta_g by projecting data to the estimated v_j (here, \mathbf{d}_g \equiv \mathbf{v}_j^T \mathbf{d} = \mathbf{v}_j^T \mathbf{X}_g + \mathbf{v}_j^T \mathbf{e}_g); and infer whether the locus is differential or not (i.e., test H_{0g}: \beta_g = 0 vs. H_{1g}: \beta_g \neq 0); and (iii) for each pattern v_j, rank genomic loci based on the magnitude of difference \left| \beta_g \right| 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 dPCA and define a signal-to-noise ratio (SNR) measure SNR_j = Var(\mathbf{v}_j^T \mathbf{d}_j)/Var(\mathbf{v}_j^T \mathbf{e}_j). We estimate SNR_j and report leading dPCs for which SNR_j > 5. This is based on observing that the dPC estimates and statistical inference on \beta_g 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 D-G.

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.
We then asked whether dPCA provides a meaningful way to rank differential chromatin patterns at TF binding sites. Intuitively, when SNR is small, dPCA becomes a useful tool for differential TF binding sites. 

Using independent ENCODE MYC ChIP-seq data, we computed log₂ fold changes in 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 $\beta_j$ for dPC1 strongly correlated with the differential MYC ChIP-seq signal (Fig. 2D; Pearson’s correlation, $p = 0.65$). We further defined 6,433 motif sites bound by MYC in at least one cell type and with MYC ChIP-seq log₂ $FC > 1.5$ as true differential MYC binding sites (SI Appendix, Text S1).

Fig. 2E 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 $\sigma_{ij}$ as the ranking criterion), the dPC1 ranking predicted differential MYC binding better (Fig. 2E and SI Appendix, Text S1 and Fig. S2D). 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).

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. 2C). 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 $\sigma^2$ 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. 3A shows the estimated FDR for each dPC. Fig. 3B shows the accuracy of the $\rho_v$ estimates. The accuracy was measured by the cosine distance $d(\rho_v, \rho_v) = 1 - \langle \rho_v, \rho_v \rangle$. A small $d$ means accurate. The vertical bars show the variability of $d$, measured by its SD across 10 independent simulations. Fig. 3C shows the error of $\rho_v$ estimates (i.e., $d(\rho_v, \rho_v)$). Fig. 3D shows the percentage of variance explained by the top dPCs. Fig. 3E–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 SNR (Fig. 3 A and B). For dPCs with SNR > 10, the estimated $\rho_v$ matched the true $\rho_v$ well and the claimed FDR provided reasonable estimates for the true FDR for testing $\beta_0 = 0$ vs. $\beta_0 \neq 0$ even if the data were projected to $\rho_v$ instead of $\rho_v$ (Fig. 3 B and E). The performance deteriorated as SNR decreased (Fig. 3 B and F). For dPCs with SNR < 5, the estimates were off the mark (Fig. 3 B and G). For dPCs with a small SNR, $\psi$ also had high variability, as evidenced by the wide error bars in Fig. 3B 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 is small, the geometric direction represented by $\psi$ in the R$^2$ space can be easily rotated by noise. When $\psi$ is biased, it is not reliable to draw conclusions about $\beta_0 = \psi \beta_v$ by projecting data to $\psi$, because $\psi$ and $\psi$ represent different geometric directions. We observed similar phenomena in all simulations (SI Appendix, Fig. S3). Therefore, although all nonzero elements in $\beta$ are assumed to be true differences not explained by cross-sample variation, we only report dPCs with SNR > 5 (dPC1 and dPC2 in this example), because the differential patterns with a smaller SNR cannot be accurately and
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. 24 and SI Appendix, Fig. S2G). 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 A. 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. 3C), which results in a smaller percentage of variance explained by the top PCs (Fig. 3D); an intuitive explanation is provided in SI Appendix, Text S1). Unlike PCA, dPCA also provides SNR 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 SNR > 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. 2A and 4I). 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. 4B). Simulations again show that the dPC and FDR estimates were reasonable when SNR > 10 and clearly biased when SNR < 5 (SI Appendix, Fig. S3).

The dPC1 coefficients β1 strongly correlated with differential gene expression (DE) determined by RNA-seq (Fig. 4C; 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. 4D and SI Appendix, Fig. S5D). 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. S5C). However, dPC1 and dPC2 jointly explained more DE than each dPC alone (SI Appendix, Fig. S5E). When promoters were grouped into nine classes based on their dPC1 and dPC2 differential states, the classes in which β1 (i.e., dPC1) and β2 (i.e., dPC2) had opposite signs had both the largest magnitude of DE (SI Appendix, Fig. S5G) and the strongest correlation between β1 and DE (Fig. 4E). 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 non-reference 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 SNR > 5 (Fig. 5A and B). This dPC is mainly driven by correlated ASB of H3K27ae, H3K4me2, H3K4me3, H3K9ac, Pol2, and c-Myc (Fig. 5A), and it positively correlates with allele-specific expression (ASE) (SI Appendix, Text S1 and Fig. S6E). 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
dPC and FDR estimates were reasonably accurate if \( \text{SNR}_j > 10 \) and biased when \( \text{SNR}_j < 5 \) (SI Appendix, Fig. S3). In real data, \( \text{SNR}_j \) was between 5 and 10. The \( v_1 \) 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 non-pseudautosomal regions of the X chromosome. We therefore compared different ranking methods based on counting how many top-ranked SNPs were in chrX (Fig. 5C and SI Appendix, Fig. S6F). 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. 5D and SI Appendix, Fig. S6G). We also did the same analysis after excluding all SNPs in chrX (SI Appendix, Fig. S6F). 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. 5E and SI Appendix, Fig. S5I 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 \( v_j \) patterns (SI Appendix, Text SI). For instance, in both the MYC and promoter examples, analyses of enriched gene sets were able to connect dPC1 and dPC2 to gene activation...
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 $\sigma^2$ 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 then find those not reported by dPCA (SI Appendix, Text S1). dPCA uses replicative 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., $E(\delta_j) = 0$) and equal variance (i.e., common $\sigma^2$) 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 $\sigma^2$ using replicate information. Then, $E(\Delta^2 G_j) = E(D^T D_j) - E(\Delta^2 G_j) = D^T D_j - \sigma^2 \Omega_j$, where $\Omega_j = diag(\frac{1}{K_{j1}} + \frac{1}{K_{j2}} \ldots, \frac{1}{K_{jN}} + \frac{1}{K_{jM}})$ is a diagonal matrix. We use the eigenvalue $\lambda_j$ and eigenvectors $v_j$ of the estimated $\Delta^2 G_j$ to estimate $\lambda_j$ and $v_j$. The proportion of variance explained by the $j$th dPC is computed as $\lambda_j / \sum_{j} \lambda_j$.

ii) Infer $\rho_j$. We have $v^T \delta_j - v^T \delta_j$ and $v^T \epsilon_j - v^T \epsilon_j$, where $\epsilon_j \sim N(0, \sigma^2 v^T \Omega_j v)$. If $v_j$ is known, one could estimate $\rho_j$ by $v^T \delta_j$ and test whether $\rho_j$ is zero by comparing the t-statistic $T_j = \frac{v^T \delta_j}{\sqrt{\frac{\sigma^2}{v^T \Omega_j v}}}$ with a t-distribution. This yields a two-sided $p$ value $p_{\text{asm}}$. In reality, $v_j$ is unknown. Thus, we project $\delta_j$ to the estimated $v_j$ to obtain $\hat{\beta}_j = v_j^T \delta_j$ and $T_j = \frac{v_j^T \delta_j}{\sqrt{\frac{\sigma^2}{v^T \Omega_j v(j)}}}$. We obtain the estimated $p$ value $p_{\text{dPCA}}$ by comparing $T_j$ with a t-distribution. Subsequently, for each dPC $\rho_j$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 $p_{\text{dPCA}}$ can estimate the true FDR for testing $p_j = 0$ reasonably well even if the data are projected to $v_j$ instead of $v_j$.

iii) For each pattern $v_j$, rank genomic loci based on $|T_j|$.

iv) Determine which dPCs to report. We report dPCs for which $SNR = \sqrt{\frac{\sigma^2}{v^T \Omega_j v}} > 5$.

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