2126–2137 *Nucleic Acids Research, 2014, Vol. 42, No. 4* *Published online 23 November 2013*
doi:10.1093/nar/gkt1186

chroGPS, a global chromatin positioning system for the functional analysis and visualization of the epigenome

Joan Font-Burgada[1,2,*], Oscar Reina[2], David Rossell[2,3,*] and Fernando Azorín[1,2,*]

[1]Institute of Molecular Biology of Barcelona, CSIC, Baldiri Rexac, 10, 08028 Barcelona, Spain, [2]Institute for Research in Biomedicine, IRB Barcelona, Baldiri Reixac, 10, 08028 Barcelona, Spain and [3]Department of Statistics, University of Warwick, Coventry CV4 7AL, UK

Received July 26, 2013; Revised October 31, 2013; Accepted November 2, 2013

ABSTRACT

Development of tools to jointly visualize the genome and the epigenome remains a challenge. chroGPS is a computational approach that addresses this question. chroGPS uses multidimensional scaling techniques to represent similarity between epigenetic factors, or between genetic elements on the basis of their epigenetic state, in 2D/3D reference maps. We emphasize biological interpretability, statistical robustness, integration of genetic and epigenetic data from heterogeneous sources, and computational feasibility. Although chroGPS is a general methodology to create reference maps and study the epigenetic state of any class of genetic element or genomic region, we focus on two specific kinds of maps: chroGPS^{factors}, which visualizes functional similarities between epigenetic factors, and chroGPS^{genes}, which describes the epigenetic state of genes and integrates gene expression and other functional data. We use data from the modENCODE project on the genomic distribution of a large collection of epigenetic factors in *Drosophila,* a model system extensively used to study genome organization and function. Our results show that the maps allow straightforward visualization of relationships between factors and elements, capturing relevant information about their functional properties that helps to interpret epigenetic information in a functional context and derive testable hypotheses.

INTRODUCTION

Understanding how genomic information is translated into cellular functions constitutes a main challenge in Biology. The eukaryotic genome exists as chromatin, a nucleoprotein complex composed by DNA, regulatory RNAs and a variety of histone and non-histone proteins that are often modified and regulate expression of the genetic information contained in DNA (1–3). Chromatin contains both genetic information encoded in the DNA sequence and epigenetic instructions that, residing in DNA-associated factors and modifications, regulate its expression. Full understanding of the functional content of the genome requires description of the epigenetic information contained in chromatin or, in other words, the epigenome.

In recent years, after sequencing the genomes of several model organisms, large amounts of data have been gathered regarding different aspects of genome functioning, from gene expression and non-coding RNAs to the genomic distribution of epigenetic factors, namely DNA methylation, histone modifications and chromatin associated proteins. There are also numerous databases describing gene functions and interactions. Tools to analyze, visualize and integrate genomic data at a functional level are available. However, integrating experimental results and databases on epigenetic factors and genetic elements in a user-friendly manner, amenable to the non-specialist, remains a challenge [reviewed in (4)]. In this context we developed chroGPS, a global chromatin positioning system to integrate and visualize the associations between epigenetic factors and their relation to functional genetic elements in low-dimensional maps.

chroGPS belongs to the family of dimensionality reduction techniques that have proven successful in analyzing

*To whom correspondence should be addressed. Tel: +858 534 0848; Fax: +858 534 8158; Email: jfontburgada@ucsd.edu
Correspondence may also be addressed to David Rossell. Tel: +44 2476523062; Fax: +44 2476524532; Email: D.Rossell@warwick.ac.uk
Correspondence may also be addressed to Fernando Azorín. Tel: +34 934034958; Fax: +34 9334934979; Email: fambmc@ibmb.csic.es
Present address:
Joan Font-Burgada, Laboratory of Gene Regulation and Signal Transduction, Departments of Pharmacology and Pathology, School of Medicine, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA.

The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

genomic data (5–9). The basic rationale is to measure similarity between epigenetic factors or between genetic elements on the basis of their epigenetic state and using multidimensional scaling (MDS) represent the similarities in 2D/3D reference maps. Emphasis is placed on interpretability, computational feasibility and statistical considerations to guarantee reliable representations and integration of data from multiple sources (studies, technologies, genetic backgrounds, etc.). A key feature of the approach lies in its generality: rather than producing a map in a specific condition, we provide a map-generating tool applicable to a wide range of situations. We illustrate the potential with two specific types of maps: chroGPS$^{\text{factors}}$, which describes similarities between epigenetic factors based on their genomic distribution profiles and informs about their functional association, and chroGPS$^{\text{genes}}$, which integrates epigenetic marks at the gene level and describes the epigenetic context of gene expression and function. As a proof of principle, we generated chroGPS maps using data from the modENCODE project in *Drosophila* (10), which constitutes the most comprehensive dataset on epigenetic factors available to date.

MATERIALS AND METHODS

Data access

ChIP-chip data from the modEncode project are freely available at www.modencode.org. Supplementary Tables S1 and S2 provide the sample identifiers. ChIP-seq data were obtained from the NCBI GEO repository at http://www.ncbi.nlm.nih.gov/geo/ (GSE19325, GSE24115 and GSE27078). See Supplementary Section S1 for details on data acquisition and formatting.

Generation, integration and annotation of chroGPS maps

chroGPS is based on two steps. First, numeric distances between objects are measured with a user-specified metric. Second, MDS is applied to generate a low-dimensional map where Euclidean distances between objects approximate the calculated distances. Therefore, the main challenges are defining an appropriate distance metric and generating a high-quality map in a reasonable computational time. Below we address these issues separately for chroGPS$^{\text{factors}}$ and chroGPS$^{\text{genes}}$. Another important challenge is integrating data from different sources, as strong source-specific biases that hamper analysis are usually present. We propose methods to adjust these biases. Finally, it is important to annotate the maps to maximize their usefulness. We discuss multivariate statistical techniques such as projections, clustering or non-parametric density estimation that help interpreting the maps. The practical use of these methodologies is described in the ‘Results’ section, pointing to detailed descriptions in the Supplementary Materials whenever appropriate.

We provide the open-source Bioconductor package chroGPS (11) (http://www.bioconductor.org). The manuals illustrate the functionality, data import and export for integration with Cytoscape (12), and the code required for our analyses.

RESULTS

Generating chroGPS maps for epigenetic factors

The inputs for chroGPS$^{\text{factors}}$ are lists of genomic intervals with predicted binding sites for each epigenetic factor, e.g. inferred from ChIP-chip or ChIP-seq experiments. We also allow using continuous scores (e.g. probe intensities, sequencing coverage) as input data. By default we use genomic intervals, as these are comparable across technologies and usually provide similar maps to those from continuous scores. Furthermore, genome annotations (genes, promoters, transposons, etc.) are also expressed as genomic intervals, facilitating data integration.

The first step in chroGPS$^{\text{factors}}$ is defining a similarity measure between two factors based on their genomic profile overlap. Although in principle MDS can be used with any similarity measure, it is important to assess its performance and effects on the final results. To address this question, we considered three metrics: interval Tanimoto (iTanimoto), average interval overlap (iOverlap) and chi-square (Supplementary Section S3). When the input are continuous scores we defined distances as $d = (1-r)/2$, where r is the Pearson correlation between two genome-wide profiles. Continuous scores were also used to perform Principal Component Analysis (PCA).

The second step is generating 2D/3D maps where distances between elements are as similar as possible to the calculated distances. Obviously, the larger the number of elements the harder it is to represent all distances accurately. We focused on two popular MDS techniques: classical metric MDS (13) and isoMDS (14). The former produces a map where pairwise Euclidean distances between elements approximate the original distances in squared norm. isoMDS is a non-metric MDS that considers monotone transformations of the input distances. We measure the map accuracy with the squared Pearson correlation (R^2) between original and approximated distances and the classical stress-1 function (s) (Supplementary Section S2).

Based on simulation studies and observations in experimental data, by default we recommend iOverlap combined with isoMDS. This choice exhibited a robust behavior when the number of binding sites was unbalanced across factors, and gave good representation accuracy and biological meaningful maps. See Supplementary Section S3.2 and Supplementary Figures S1–S3.

chroGPS$^{\text{factors}}$ describes main functional chromatin states

Figure 1A shows iOverlap distances between factors in *Drosophila* S2 cells (rows and columns sorted by hierarchical clustering). While the plot suggests several clusters, interpretation is not straightforward and the relative similarity between clusters is unclear. Instead, representing the distance matrix in 2D/3D reference maps (Figure 1B and C, Supplementary Figure S14 and Video S1) provides a directly interpretable representation. isoMDS maps show higher R^2 than their classicMDS counterparts, and for

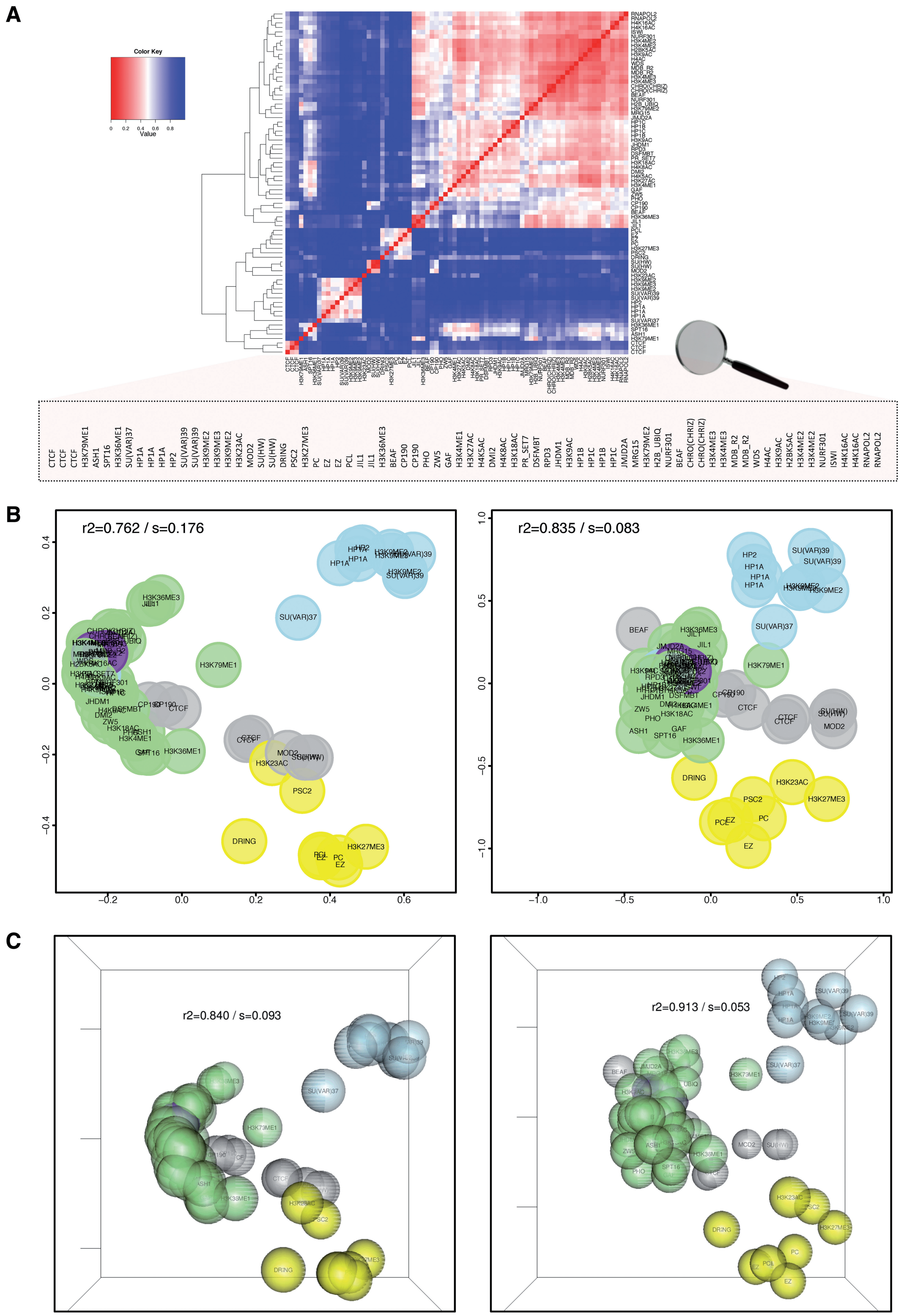

Figure 1. *chroGPS*^{factors} visualizes functional associations between epigenetic factors. Similarity between 76 individual epigenetic factors (Supplementary Table S1), as determined from their genomic profiles with iOverlap, is represented in a heatmap with hierarchical clustering dendrogram (**A**) or 2D (**B**) and 3D (**C**) reference maps using classical MDS (left) or isoMDS (right). R^2 and stress (s) are indicated. Factors are colored according to their biological activity: RNApol II (purple), regulation of transcription (green), boundary/insulator function (gray), HP1-(blue) and Polycomb (PC)-dependent silencing (yellow). See also Supplementary Video S1 for visualization of the 3D-map in motion (Supplementary Section S10.1).

both metrics 3D maps provide a non-negligible accuracy improvement. Therefore, although for convenience we use 2D representations throughout the manuscript, we generally recommend 3D maps for better visualization.

The map describes main functional chromatin states. To facilitate interpretation, we colored factors according to their biological activity: RNApol II (purple), regulation of transcription (green), boundary/insulator function (gray), HP1-dependent (blue) and Polycomb (PC)-dependent silencing (yellow). As shown in Figure 1B and C, the map has a characteristic funnel configuration. Factors involved in transcription regulation seat at the apex and define the 'active chromatin' domain, while boundary/insulator, HP1- and PC-chromatin domains localize in the wider zone. In the core of the 'active chromatin' domain resides RNApol II (purple), and several branches emerge from there (Supplementary Section S8).

chroGPS^{factors} integrates data from different sources

The practical usefulness of chroGPS strongly depends on its ability to integrate data from different sources; an extreme case being datasets obtained using completely different methodologies. For instance, at present genomic profiling of epigenetic factors is largely determined through ChIP-seq experiments, which identify binding sites at higher accuracy than ChIP-chip. Due to intrinsic technical differences, ChIP-seq and ChIP-chip data tend to appear on separate layers in the map (Supplementary Figure S5).

We propose two bias-adjustment methods: Procrustes and peak width adjustment (PWA). Procrustes superimposes two sets of points by altering their center, scale and orientation, while preserving relative distances within each set. It is a general adjustment that accounts for fairly general biases. A limitation is requiring the datasets to share common points (in practice, ≥ 3), which may not be available. As an important difference across methodologies lies in peak-calling resolution, we propose PWA as an alternative. PWA selects the source with widest peaks and increases the peak width in the remaining sources until the mean and standard deviation of the peak width distribution are equal. See Supplementary Section S5.1 for further details and results.

We illustrate the power of chroGPS^{factors} to integrate data from different sources using four factor profiles obtained via ChIP-seq. For the three factors, ChIP-chip data are also available (H3K4me3, H3K27me3 and H3K36me3), the remaining one being the active RNApol II form Pol IIo^{ser5} (15) (see 'Data Access' section). A joint ChIP-chip/ChIP-seq map including these data (Figure 2A) locates the three common ChIP-seq elements as an external layer to their ChIP-chip counterparts. As shown in Figure 2B, applying Procrustes to Figure 2A using the three common elements effectively matches ChIP-seq and ChIP-chip locations. The adjustment also brings the Pol IIo^{ser5} ChIP-seq position in close proximity to whole RNApol II and other transcription activation factors. Applying PWA to Figure 2A provides similar results (Figure 2C).

chroGPS^{factors} assesses conservation of functional co-operation and genomic location of epigenetic factors

Functional co-operation between epigenetic factors is largely conserved, implying that chroGPS^{factors} maps obtained separately in different cellular backgrounds should present similar configurations. However, the actual genomic locations of the factors under consideration may well differ across backgrounds (e.g. binding different genes). These observations prompt a direct use of chroGPS maps to assess conservation of both genomic locations and factor interactions, e.g. identify cell-type specific interactions or disease-related alterations. Whenever the genomic locations of a given factor are conserved across backgrounds, directly merging all data into a joint map reveals a similar location for that factor across all backgrounds. If only interactions between factors are conserved (but not factor-binding sites), backgrounds appear separated in the map but the relative factor positions within each background are conserved. See Supplementary Section S5.2 for further details.

We integrated ChIP-chip data from *Drosophila* BG3 cell line (Supplementary Table S2) with the S2 data shown above. BG3 is a third instar larval stage CNS-derived cell line while S2 is a late embryonic cell line. As anticipated, a separate BG3 map (Figure 3A) retains the general features of the S2 map, which indicates that functional associations between factors are largely conserved. Figure 3B merges the two datasets, calculating a joint S2/BG3 distance matrix and representing it via isoMDS. Remarkably, the joint map is biologically meaningful and shows a similar configuration to the individual maps. This finding indicates that genomic locations of most factors are highly conserved, i.e. the S2/BG3 distance for each factor is relatively small. Indeed, S2/BG3 distances are of the same magnitude as distances between functionally related factors (intra-domain) and much smaller than those between unrelated factors (inter-domain) (Figure 3C). Furthermore, in the joint S2/BG3 map intra- and inter-domain distances are similar to those in the S2 map (Figure 3C), i.e. the joint map describes functional domains as accurately as the separate maps.

These results indicate that chroGPS maps identify cases where genome-wide factor location is conserved. We now investigate their ability to detect situations where functional interplays are conserved but genomic locations differ across backgrounds. We performed a simulation study by artificially increasing/decreasing the similarities between S2 and BG3, while not altering similarities within each cell line (Supplementary Section S5.2). That is, we emulated a situation where interplays within cell lines are conserved but genomic locations are not conserved across cell lines. The map preserved its general configuration whenever S2/BG3 distances remained smaller than inter-domain distances (Supplementary Figure S8). As desired, when S2/BG3 distances grew further the map split S2 from BG3 factors into two subsets, each with similar configuration. These results show that chroGPS^{factors} maps assess conservation of functional interactions and genomic locations in a straightforward manner.

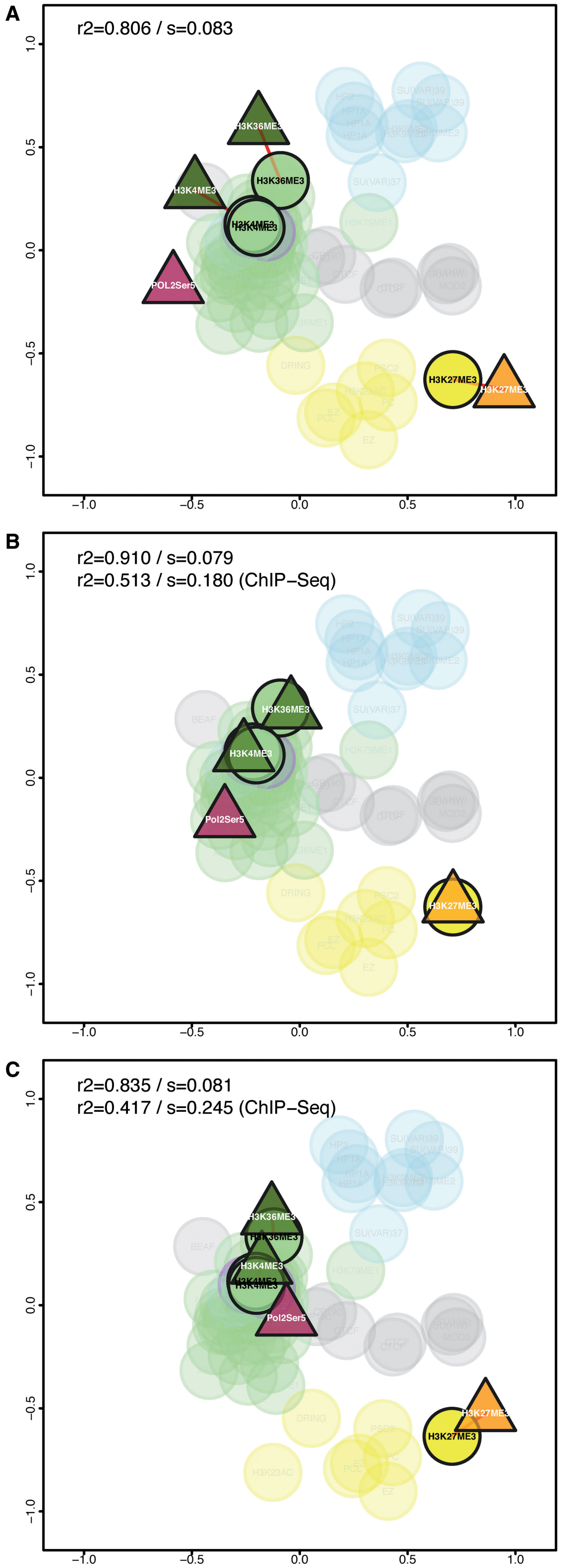

Figure 2. Integration of ChIP-chip and ChIP-seq data in chroGPS^{factors} maps. ChIP-seq data (triangles) for H3K4me3, H3K27me3 H3K36me3 and Pol IIo^{ser5} is integrated into the chroGPS^{factors} map generated from modENCODE ChIP-chip data (circles). The joint ChIP-chip/ChIP-seq 2D-map is presented without any adjustment (**A**) and after Procrustes (**B**) or PWA (**C**) adjustment. ChIP-Seq elements and the corresponding ChIP-chip counterparts are highlighted. Common factors in both technical backgrounds (replicates) are joined by a red line. Maps were generated from iOverlap distances and represented using isoMDS. R^2 and stress (s) are indicated for both the joint map and the ChIP-Seq map before (A) and after integration (B and C).

chroGPS^{factors} functionally catalogs novel epigenetic factors

Here we use chroGPS^{factors} to interrogate about the functions of novel factors in epigenetic terms. This operation is straightforward when the novel factor is studied in an experimental system with abundant experimental data from other factors. It simply requires including the new data and generating a joint map. As an example, Figure 4A highlights the position of JMJD2A, a histone demethylase of unknown function that is capable of demethylating both H3K9me3 and H3K36me3 (16,17). JMJD2A localizes in the active chromatin domain close to H3K36me3, suggesting that *in vivo* it demethylates H3K36me3 and contributes to transcription regulation.

Learning the function of a novel factor in a system with limited experimental data poses a bigger challenge. To illustrate the potential of chroGPS maps in this situation, we used data generated in the *Drosophila* wing imaginal disc that only contain seven factors, four common to the S2 map (H3K4me3, H3K27me3, H3K36me3 and Pol IIo^{ser5}) and three unique (Pol IIo^{ser2}, ASH2 and dKDM5/LID) (18,19) (see 'Data Access' section). These data were generated by ChIP-seq in a larval structure formed by a heterogeneous population of cells and, thus, differs strongly from embryonic S2 cells ChIP-chip data. A raw map generated with these seven elements contains very low structural information (Figure 4B). In Figure 4C we merged these data with the S2 factors into a joint distance matrix and used Procrustes to match the positions of the four common factors. The unique WING-factors localize to the active chromatin domain, close to other functionally related factors. For instance, elongating Pol IIo^{ser2} lies close to H3K36me3, a modification that is enriched at the coding region of elongating genes. Similarly, dKDM5/LID, which is involved in the regulation of Pol IIo^{ser5} at promoters (18), is close to Pol IIo^{ser5}. On the contrary, ASH2, which has also been shown to regulate Pol IIo^{ser5} (19), maps at a more external position, suggesting it plays additional roles. That is, integrating wing imaginal disc and S2 factors into a joint map provided a richer environment to study functions of the former.

Generating chroGPS maps for genetic elements

chroGPS^{genes} maps display the epigenetic state of genes, as well as their expression and function. The maps are based on measuring similarity between genes according to their shared epigenetic factors. The input is a binary matrix with genes in rows and factors in columns, where 1 indicates that the factor binds that gene and 0 otherwise (Supplementary Section S1). We considered three similarity metrics: chi-square, Tanimoto and average overlap.

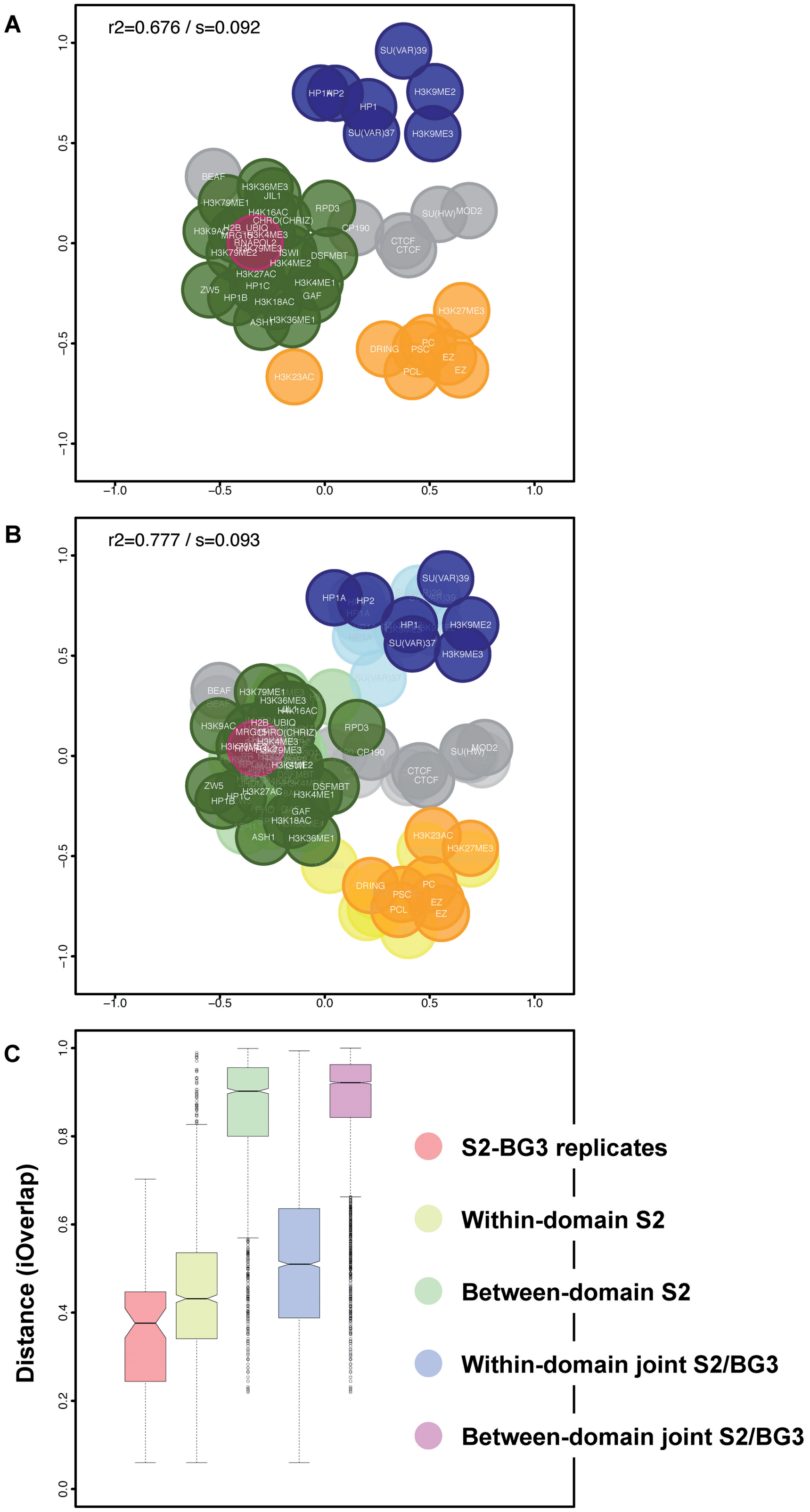

Figure 3. Integration of data from different cellular backgrounds in chroGPS^{factors} maps. (**A**) Map generated from modENCODE ChIP-chip data obtained in BG3 cells for 46 different epigenetic factors (Supplementary Table S2). (**B**) Joint S2/BG3 map obtained by merging S2 and BG3 datasets. Color code is as in Figure 1 and BG3 factors are highlighted. Maps were generated from iOverlap distances and represented using isoMDS. R^2 and stress (s) are indicated. (**C**) iOverlap distances between S2-BG3 replicates are compared to those between functionally related (within-domain) and unrelated (between-domain) factors determined in both the S2 and the joint S2/BG3 chroGPS^{factors} maps.

The former is the basis of correspondence analysis, a standard dimensionality reduction technique for binary matrices. The latter two measure overlap in an intuitive manner and weight all epigenetic factors equally. As some users might consider that sharing a scarce factor can be more biologically meaningful than sharing a frequent one, we defined a weighted Tanimoto distance as a fourth metric. This metric weights co-occurrences inversely to the number of genes with that factor (Supplementary Section S4).

Representing tens of thousands of points via MDS is a high-dimensional problem posing two important challenges. First, classical MDS may fail to numerically minimize the stress function and result in poor R^2 coefficients. Second, the required computation time for alternative solutions can be substantial. To overcome these limitations we propose BoostMDS, a novel two-step procedure (Supplementary Section S6). Briefly, BoostMDS finds an initial solution by splitting the distance matrix into smaller overlapping sub-matrices. The sub-problems are computationally tractable, can be run in parallel and the solutions are stitched into an overall map using Procrustes. This initial solution is then refined by formally maximizing the R^2 coefficient via a gradient search algorithm with automatic step size selection. By default we recommend applying BoostMDS to Tanimoto distances, as this option achieves high representation accuracy at a substantially reduced computation time (Supplementary Figure S4).

Annotating chroGPS^{genes} maps

chroGPS^{genes} maps contain thousands of points, which hampers interpretation. Fortunately, they can integrate genetic data to visualize the relationship between epigenetic states and gene functions. For instance, gene expression or other continuous measurements can be shown by coloring the map. Another basic operation is to highlight genes bound by a given epigenetic factor, related to some genetic pathway or function. We use non-parametric contours to indicate high-density regions for any given gene set. Contours indicate the overlap between gene sets in a manner analogous to Venn diagrams, with the advantage of providing a functional context. See Supplementary Section S7 for details.

Beyond representing individual gene sets, we provide tools to display the results of clustering analyses to help interpret each area in the map. Clusters not showing good separation in the map are merged using posterior expected correct classification (CCR) and probabilistic overlap (PO) criteria. We provide an automatic procedure to stop merging clusters based on change-point analysis.

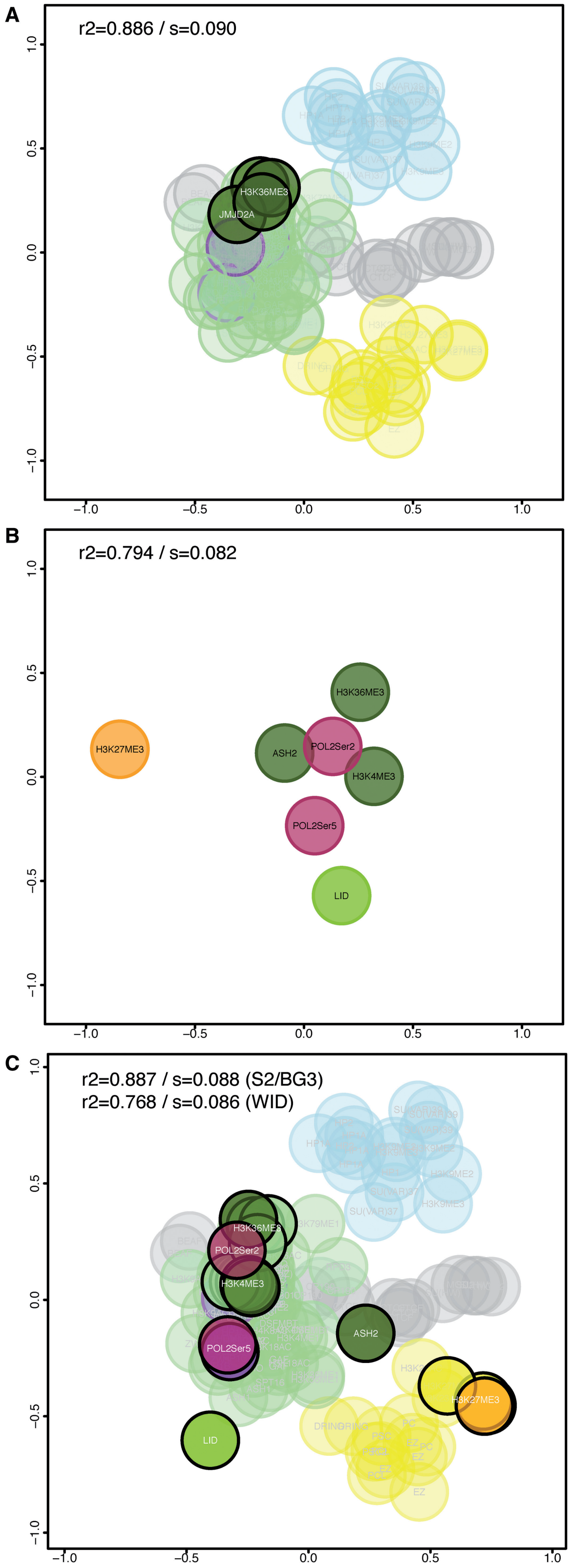

Figure 4. Using chroGPS^{factors} maps to analyze novel epigenetic factors of unknown function. (**A**) The position corresponding to JMJD2A is highlighted in the map integrating S2 ChIP-chip and ChIP-seq data, and BG3 ChIP-chip data. The positions of three different datasets for H3K36me3 are also indicated. (**B**) chroGPS^{factors} map generated from ChIP-seq data obtained in the wing imaginal disc (WID) for the seven indicated factors. The map was generated from

Figure 4. Continued
iOverlap distances and represented using isoMDS. R^2 and stress (s) are indicated. (**C**) Joint S2/WID chroGPS^{factors} map integrating WID ChIP-seq data, and S2 ChIP-chip and ChIP-seq data. The map was generated from iOverlap distances, represented using isoMDS and adjusted using Procrustes. Pearson correlation (R^2) and MDS stress function (s) values are indicated for both the joint map and the WID map after integration. Color code is as in Figure 1.

Finally, the robustness of the map interpretation afforded by the clusters is assessed via bootstrap. See Supplementary Section S7 for details.

chroGPS^{genes} describes the epigenetic context of gene function

Figure 5A (left) shows the chroGPS^{genes} map of *Drosophila* S2 cells based on BoostMDS and Tanimoto distances (for a 3D representation see Supplementary Video S2). Contours in Figure 5A indicate clusters of genes with similar epigenetic profiles (average between-cluster distance threshold = 50%). These clusters overlap strongly and do not provide a clear description of the map. Accordingly, the CCR is low (Supplementary Figure S12). Complexity of the map is highly reduced after merging clusters until their PO drops swiftly (Figure 5A, middle and Supplementary Figure S11). The 37 clusters merged to 12 that are specifically enriched/depleted in particular epigenetic factors (Figure 5B) and describe distinct epigenetic states. This 12-clusters configuration was found to be reproducible in a bootstrap analysis (Supplementary Figure S9 and Supplementary Section S7). Clusters 1–2 correspond to the 'active chromatin region', clusters 3 and 4 to the 'HP1-chromatin region', clusters 5, 6, 7, 8 and 10 to the 'PC-chromatin region', and clusters 9 and 11 show a peculiar combination of 'silencing' and 'active' factors, and cluster 12 has both PC and HP1 silencing marks. We decreased the between-cluster distance threshold <50% and obtained smaller clusters (Supplementary Figure S10). After cluster merging the CCR was high (>75%) and largely independent of the distance threshold (Supplementary Figure S13).

The chroGPS^{genes} map can be used as the canvas to paint other 'omics' information, delivering a global picture of the biological scenarios. As an illustration, Figure 5A (right) shows gene expression and Supplementary Figure S15 the distribution of certain epigenetic factors. Highly expressed genes concentrate in clusters 1–2, lowly expressed in clusters 3–9, while clusters 10–12 show a wider expression range. In good agreement, genes bound by RNApol II and marked with 'active' histone modifications (H3K4me3 and H3K36me3) concentrate in clusters 1 and 2, while silenced clusters 3–9 are generally enriched in 'repressive' factors [HP1a/Su(var)3–9 and PC/EZ] and clusters 10, 11 and 12 appears to correspond to intermediate states, as they contain both 'active' and 'repressive' marks (see Supplementary Section S9 and Supplementary Figure S15 for details).

chroGPS^{genes} complements the information provided by chroGPS^{factors} on the function of novel epigenetic factors, facilitating interpretation of the biological context of their action. Figure 6A shows contours for genes bound by JMJD2A and JHDM1 (16,17,20). These two histone

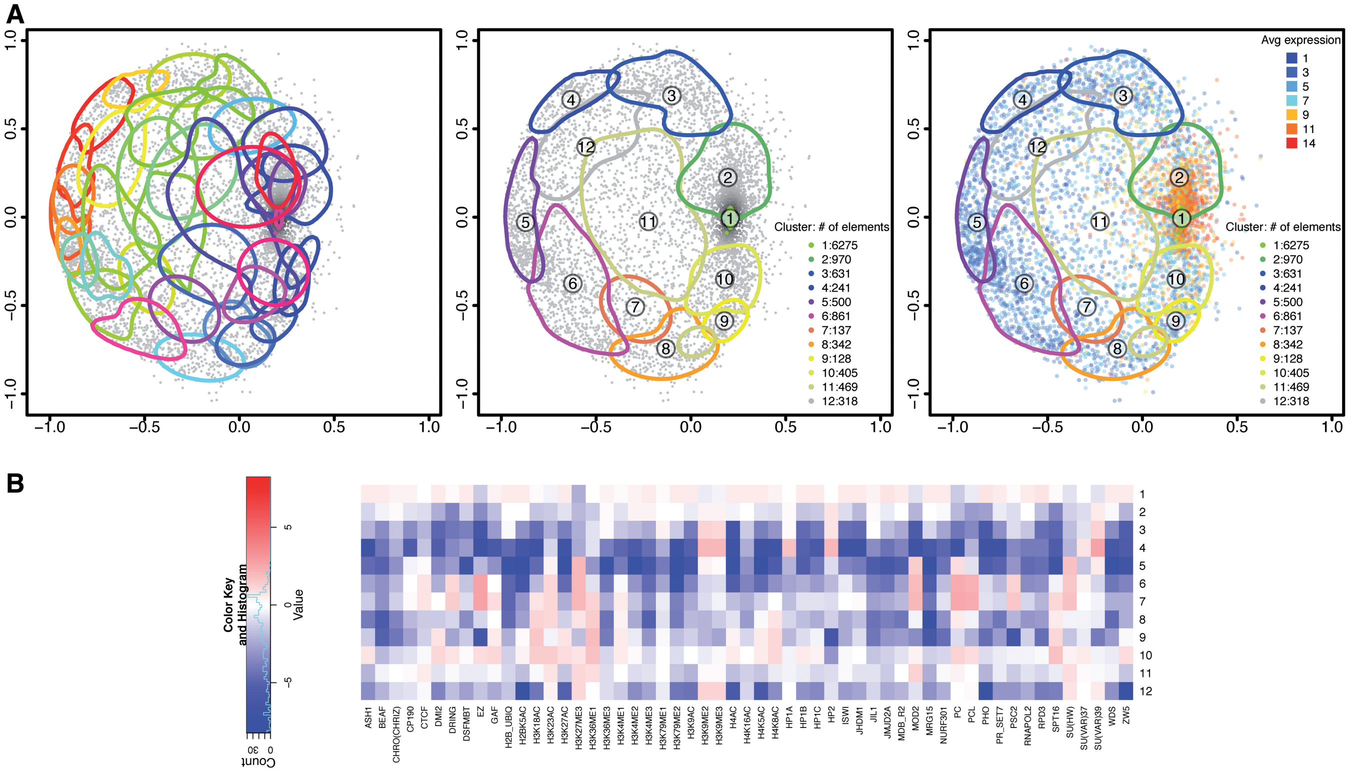

Figure 5. chroGPS^{genes} map of S2 cells. The map was generated using Tanimoto distances and BoostMDS for representation. (**A**) Analysis of the map based on hierarchical clustering with average linkage. Clusters corresponding to 50% between-cluster distance before (left) and after unsupervised merging (center). On the right, the average log2 gene expression level is indicated. For each cluster, the 75% density contour is shown. The number of elements in each cluster is indicated. (**B**) The epigenetic profiles of the 12 merged clusters in Figure 5A (center) are presented. For each cluster, the $\log_2$ enrichment/depletion ratio of each factor with respect to its average distribution in the whole map are indicated. See also Supplementary Video S2 for visualization of the 3D-map in motion (Supplementary Section S10.1).

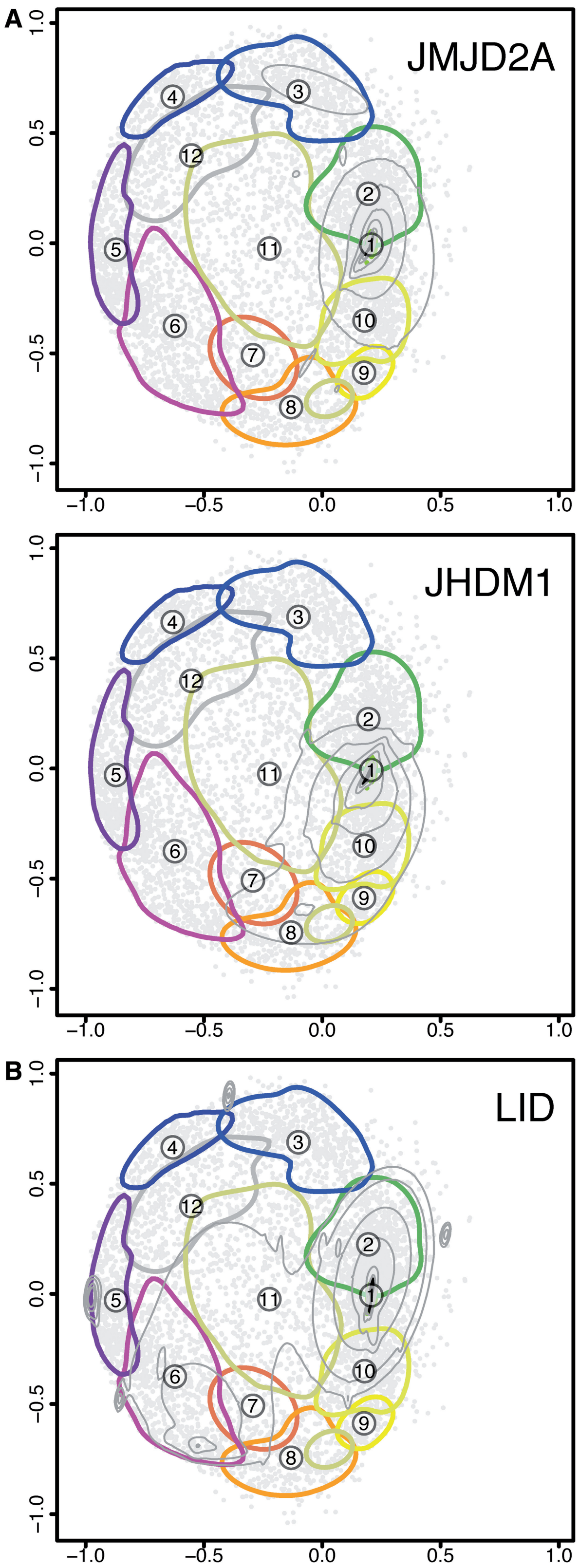

Figure 6. Using chroGPS^{genes} maps to analyze novel epigenetic factors of unknown function. (**A**) The distribution of JMJD2A (top) and JHDM1 (bottom) are shown on the 12-cluster configuration of the S2 chroGPS^{genes} map. (**B**) The distribution of genes bound by dKDM5/LID in the wing imaginal disc is shown on the 12-cluster configuration of the S2 chroGPS^{genes} map. For each factor, concentric density contours (from 10% to 95%) are presented. For each cluster, the 75% density contour is shown.

demethylases target H3K36me3 (an 'active' mark) and were anticipated to act as repressive factors. Opposite to this hypothesis, they map to cluster 1, suggesting that they regulate H3K36me3 in active genes, likely during transcription elongation.

Although chroGPS^{genes} maps are cell-type specific, they can also provide useful information for data generated in a different cell-type. For instance, genes bound by dKDM5/LID in the wing imaginal discs are active (18) and they mostly locate in the active cluster 1 in the S2 map (Figure 6B), indicating that they maintain the active state in S2. However, some of them are also distributed across other clusters, which likely represent genes that are bound by dKDM5/LID and active in the wing disc but are repressed, or less active, in S2 cells. Interestingly, this set of genes mostly concentrates in the PC-clusters 5 and 6, suggesting that dKDM5/LID-target genes tend to be repressed by PC.

chroGPS^{genes} allows epigenetic analysis of complex biological networks

Locating genes involved in *Toll* signaling in the map provides an integrated view of the network's epigenetic state. We consider 15 genes involved in the extracellular *Toll* cascade, whose expression would eventually change depending of the cell type and developmental stage. Figure 7B (top) evidences their complex and diverse epigenetic regulation, since seven lie in the active chromatin clusters 1–2, three locate at PC-chromatin cluster 5, three at the HP1-chromatin clusters 3–4 and two in the interphase between clusters 11, 8 and 10. Importantly, this analysis provides additional information beyond the experimental data on the factors binding each gene. For instance, only a few epigenetic factors have been experimentally found to bind the three genes lying at cluster 5 (*necrotic*, *spheroide* and *sphinx2*). For *necrotic* only H3K27me3 has been detected, for *sphinx2* Su(Hw) and for *spheroide* both H3K27me3 and Su(Hw). The algorithm maps these three genes in cluster 5 which is not only enriched in H3K27me3 and Su(Hw), but also in PC and MOD2. These findings are interesting in that, despite the little factor binding information for the three genes, they suggest testable hypotheses.

Genes of the intracellular signaling cascade, known to be ubiquitously expressed, concentrate at the active chromatin cluster 1 characterized by high expression and multiple epigenetic factors (Figure 7B, center). Finally, Toll immunity pathway target genes show a complex pattern (Figure 7B, bottom). None map to the active chromatin cluster 1, in agreement with the expected weak engagement of Toll pathway in ideal cell-culture conditions. Interestingly, several target genes lie in the interphase between clusters 9 and 10, which corresponds to an intermediate epigenetic state sharing marks of both active and repressed chromatin.

DISCUSSION

Large efforts are currently devoted to describing the epigenome (21–30). The development of tools to integrate,

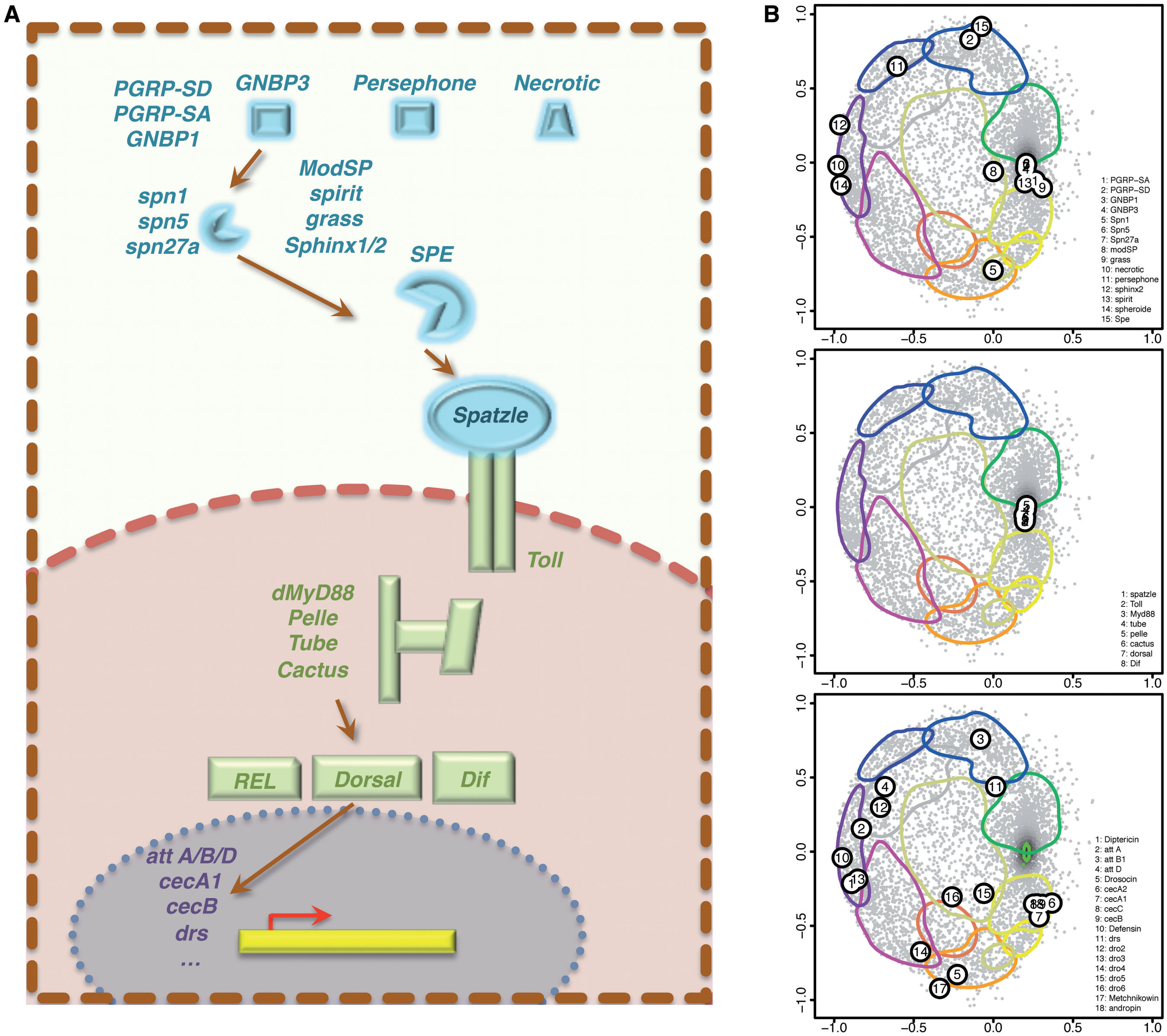

Figure 7. Using chroGPS^{genes} maps for the epigenetic analysis of the Toll innate immune pathway. (**A**) Schematic overview of the pathway. (**B**) Extracellular (top), intracellular (center) and nuclear-target (bottom) genes of the pathway are highlighted on the 12-cluster configuration of the S2 chroGPS^{genes} map.

analyze and visualize large amounts of epigenetic and genetic data is a priority in the field. In particular, hidden Markov models have been used to describe and characterize epigenetic states in chromatin (24,31). While these approaches are useful, they only focus on combinations of epigenetic factors and setups where all data are of the same kind and no systematic biases are present. An approach based on self-organizing maps, again aimed at portraying associations between epigenetic factors based on single-type data, has also been proposed (8). Instead, chroGPS is the first dimensionality reduction technique specifically designed to explore combinations of multiple data types, accounting for systematic biases and can focus both on genetic elements and epigenetic factors. Our main contribution is enabling the integration and interpretation of massive heterogeneous epigenetic data in a visually appealing and context-rich manner. We assessed the adequacy of multiple distance metrics, provided algorithms to represent a large number of objects at high resolution and a computational effort manageable by a desktop computer, and strategies to annotate the maps in order to enhance their interpretability.

chroGPS maps proved useful in a variety of situations, such as understanding functional interplays between epigenetic factors in *Drosophila*, assess conservation across S2 and BG3 cells, deriving testable hypotheses for novel factors, studying chromatin states at genes and the epigenetic regulation of complex pathways. While these examples illustrate the broad potential, we envision further possible uses. For instance, chroGPS maps that consider only overlaps at specific locations (e.g. promoters, exons, origins of replication, transposons, selected gene-sets, etc.) would inform about epigenetic states and functional interactions occurring at the investigated elements. In addition, integrating Hi-C data would provide direct information about the relative position of the analyzed elements. Another interesting venue is to generate maps for a particular developmental process or disease, as these would portray genetic/epigenetic changes in the analyzed conditions. For instance, comparing chroGPS^{genes} maps from normal and affected cells in a given disease condition (i.e. neurons in Rett syndrome) could identify which genes are changing epigenetic status and in what direction(s). Similarly, a joint chroGPS^{factors} plot containing data from normal and disease status could identify altered interactions between epigenetic factors (i.e. MeCP2 in Rett syndrome), at the whole-genome level or in specific genomic regions.

Overall, chroGPS maps should prove a valuable approach to explore these complex questions by combining large amounts of data, serving as a hypothesis generating tool and the starting point for further in-depth analyses.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors are thankful to Dr J. Graffelman for useful comments.

FUNDING

The Spanish 'Ministerio de Economia y Competitividad' [CSD2006-49, BFU2009-07111 and BFU2012-30724] and the 'Generalitat de Catalunya' [SGR2009-1023]; This work was carried out within the framework of the 'Centre de Referència en Biotecnologia' of the 'Generalitat de Catalunya'. Funding for open access charge: Spanish 'Ministerio de Economia y Competitividad' [BFU2012-30724].

Conflict of interest statement. None declared.

REFERENCES

1. Lange,M., Demajo,S., Jain,P. and Di Croce,L. (2011) Combinatorial assembly and function of chromatin regulatory complexes. *Epigenomics*, **3**, 567–580.
2. Lee,J.T. (2012) Epigenetic regulation by long noncoding RNAs. *Science*, **338**, 1435–1439.
3. Musselman,C.A., Lalonde,M.E., Côté,J. and Kutateladze,T.G. (2012) Perceiving the epigenetic landscape through histone readers. *Nat. Struct. Mol. Biol.*, **19**, 1218–1227.
4. Karnik,R. and Meissner,A. (2013) Browsing (Epi)genomes: a guide to data resources and epigenome browsers for stem cell researchers. *Cell Stem Cell*, **13**, 14–21.
5. Boulesteix,A.-L. and Strimmer,K. (2007) Partial least squares: a versatile tool for the analysis of high-dimensional data. *Brief. Bioinform.*, **8**, 32–44.
6. Miclaus,K., Wolfinger,R. and Czika,W. (2009) SNP selection and multidimensional scaling to quantify population structure. *Genet. Epidemiol.*, **33**, 488–496.
7. Motsinger,A.A. and Ritchie,M.D. (2006) Multifactor dimensionality reduction: an analysis strategy for modeling and detecting gene-gene interaction in human genetics and pharmacogenomics studies. *Hum. Genom.*, **2**, 318–328.
8. Steiner,L., Hopp,L., Wirth,H., Galle,J., Binder,H., Prohaska,S.J. and Rohlf,T. (2012) A global genome segmentation method for exploration of epigenetic patterns. *PLoS One*, **7**, e46811.
9. van de Wiel,M.A. and van Wieringen,W.N. (2007) CGHregions: dimension reduction for array CGH data with minimal information loss. *Cancer Inform.*, **3**, 55–63.
10. Celniker,S.E., Dillon,L.A., Gerstein,M.B., Gunsalus,K.C., Henikoff,S., Karpen,G.H., Kellis,M., Lai,E.C., Lieb,J.D., MacAlpine,D.M. *et al.* (2009) Unlocking the secrets of the genome. *Nature*, **459**, 927–930.
11. Gentleman,R.C., Carey,V.J., Bates,D.M., Bolstad,B., Dettling,M., Dudoit,S., Ellis,B., Gautier,L., Ge,Y., Gentry,J. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.
12. Smoot,M.E., Ono,K., Ruscheinski,J., Wang,P.L. and Ideker,T. (2011) Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, **27**, 431–432.
13. Torgerson,W.S.P. (1952) Multi-dimensional scaling: I, theory and method. *Psychometrika*, **17**, 401–419.
14. Venables, W.N. and Ripley, B.D. (2002) Modern applied statistics with S, 4th edn. Springer, New York, ISBN 0387954570.
15. Gan,Q., Schones,D.E., Ho Eun,S., Wei,G., Cui,K., Zhao,K. and Chen,X. (2010) Monovalent and unpoised status of most genes in undifferentiated cell-enriched Drosophila testis. *Genome Biol.*, **11**, R42.
16. Lin,C.H., Li,B., Swanson,S., Zhang,Y., Florens,L., Washburn,M.P., Abmayr,S.M. and Workman,J.L. (2008) Heterochromatin protein 1a stimulates histone H3 lysine 36 demethylation by the Drosophila KDM4A demethylase. *Mol. Cell*, **32**, 696–706.
17. Lloret-Llinares,M., Carré,C., Vaquero,A., de Olano,N. and Azorín,F. (2008) Characterisation of Drosophila melanogaster JmjC + N histone demethylases. *Nucleic Acids Res.*, **36**, 2852–2863.
18. Lloret-Llinares,M., Pérez-Lluch,S., Rossell,D., Morán,T., Ponsa-Cobas,J., Auer,H., Corominas,M. and Azorín,F. (2012) dKDM5/LID regulates H3K4me3 dynamics at the transcription-start site (TSS) of actively transcribed developmental genes. *Nucleic Acids Res.*, **40**, 9493–9505.
19. Pérez-Lluch,S., Blanco,E., Carbonell,A., Raha,D., Snyder,M., Serras,F. and Corominas,M. (2011) Genome-wide chromatin occupancy analysis reveals a role for ASH2 in transcriptional pausing. *Nucleic Acids Res.*, **39**, 4628–4639.
20. Pedersen,M.T. and Helin,K. (2010) Histone demethylases in development and disease. *Trends Cell Biol.*, **20**, 662–671.
21. Dunham,I., Kundaje,A., Aldred,S.F., Collins,P.J., Davis,C.A., Doyle,F., Epstein,C.B., Frietze,S., Harrow,J., Kaul,R. *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
22. Ernst,J. and Kellis,M. (2010) Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.*, **28**, 817–825.
23. Ernst,J., Kheradpour,P., Mikkelsen,T.S., Shoresh,N., Ward,L.D., Epstein,C.B., Zhang,X., Wang,L., Issner,R., Coyne,M. *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43–49.
24. Filion,G.J., van Bemmel,J.G., Braunschweig,U., Talhout,W., Kind,J., Ward,L.D., Brugman,W., de Castro,I.J., Kerkhoven,R.M., Bussemaker,H.J. *et al.* (2010) Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. *Cell*, **143**, 212–214.
25. Gerstein,M.B., Lu,Z.J., Van Nostrand,E.L., Cheng,C., Arshinoff,B.I., Liu,T., Yip,K.Y., Robilotto,R., Rechtsteiner,A., Ikegami,K. *et al.* (2011) Integrative analysis of the Caenorhabditis

elegans genome by the modENCODE project. *Science*, **330**, 1775–1787.

26. Kharchenko,P.V., Alekseyenko,A.A., Schwartz,Y.B., Minoda,A., Riddle,N.C., Ernst,J., Sabo,P.J., Larschan,E., Gorchakov,A.A., Gu,T. *et al.* (2011) Comprehensive analysis of the chromatin landscape in *Drosophila melanogaster*. *Nature*, **471**, 480–485.
27. Liu,T., Rechtsteiner,A., Egelhofer,T.A., Vielle,A., Latorre,I., Cheung,M.S., Ercan,S., Ikegami,K., Jensen,M., Kolasinska-Zwierz,P. *et al.* (2011) Broad chromosomal domains of histone modification patterns in *C. elegans*. *Genome Res.*, **21**, 227–236.
28. Roudier,F., Ahmed,I., Bérard,C., Sarazin,A., Mary-Huard,T., Cortijo,S., Bouyer,D., Caillieux,E., Duvernois-Berthet,E., Al-Shikhley,L. *et al.* (2011) Integrative epigenomic mapping defines four main chromatin states in *Arabidopsis*. *EMBO J.*, **30**, 1928–1938.
29. Roy,S., Ernst,J., Kharchenko,P.V., Kheradpour,P., Negre,N., Eaton,M.L., Landolin,J.M., Bristow,C.A., Ma,L., Lin,M.F. *et al.* (2010) Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science*, **330**, 1787–1797.
30. van Bemmel,J., Filion,G., Rosado,A., Talhout,W., de Haas,M., van Welsem,T., van Leeuwen,F. and van Steensel,B. (2013) A network model of the molecular organization of chromatin in *Drosophila*. *Mol. Cell*, **49**, 759–771.
31. Ernst,J. and Kellis,M. (2012) ChromHMM: automatic chromatin-state discovery and characterization. *Nat. Methods*, **9**, 215–216.