chroGPS2: differential analysis of epigenome maps in R.

Oscar Reina[1], Fernando Azorin[1,2] and Camille Stephan-Otto Attolini[1].

[1]Institute for Research in Biomedicine, IRB Barcelona, Baldiri Reixac, 10, 08028 Barcelona, Spain.
[2]Institute of Molecular Biology of Barcelona, CSIC, Baldiri Reixac, 10, 08028 Barcelona, Spain.

December 21, 2018

1 Abstract

In the last years, after systematic mapping of epigenomics data from multiple organisms, tissues and cell lines, the ability to efficiently integrate, visualize and compare such information remains a challenge. Here we present chroGPS version 2, a major update of our previously developed software chroGPS, for visualization and differential analysis of epigenomes. Methods are provided for efficient integration and comparison of data from different conditions or biological backgrounds, accounting and adjusting for systematic biases in order to provide an efficient and statistically robust base for differential analysis. We also include functionalities for general data assessment and quality control prior to comparing maps, such as functions to study chromatin domain conservation between epigenomic backgrounds, to detect gross technical outliers and also to help in the selection of candidate marks for de-novo epigenome mapping.

- Availability: `https://www.bioconductor.org/packages/release/bioc/html/chroGPS.html` - Contact: oscar.reina@irbbarcelona.org

2 Introduction

Recently, we have assisted to an unprecedented increase in availability of epigenomics data related to whole-genome distribution of transcription factors, histone modifications and other DNA binding proteins [Celniker et al., 2009; Dunham et al., 2012; Bernstein et al., 2010], which helped to establish a deeper knowledge about chromatin states and topologically associated domains [Filion et al., 2010; Serra et al., 2016], and to offer a broad scope of genomic regulation from both a functional and structural point of view. However, there remains a need for efficient tools for visualization, functional analysis and comparison of epigenomics data [Marx, 2015]. Previously, we developed chroGPS [Font-Burgada et al., 2014], an R package based on dimensionality reduction techniques (namely Multidimensional Scaling, MDS) to measure and visualize genome-wide associations between epigenetic factors and their relationship to functional genetic elements in low dimensional maps. Now we extend this software

and introduce novel features to perform differential analysis of epigenome maps using Procrustes [Lisboa et al., 2014], hierarchical clustering and Bayesian density estimation methods [Jara et al., 2011]. Additionally, we provide functions for general data assessment, quality control and to help in the selection of epigenetic factors for de-novo epigenome mapping. ChroGPS2 is integrated in Bioconductor [Huber et al., 2015], an open-source collection of R packages for computational analysis of omics data. We illustrate our approach using publicly available datasets containing extensively mapped fruit fly epigenomes. Supplementary material includes detailed information, workflows and additional examples in several biological scenarios.

3 Data assessment and Quality Control

We offer functions to perform a first general assessment of functional relationships between epigenetic factors and conformation of chromatin domains across datasets, and to highlight major differences between them. After measuring pairwise similarities between epigenetic factors based on genome-wide overlap of their binding profiles with the *distGPS* function, the *domainDist* one evaluates and compares these results at factor and domain level, pointing towards major changes that may suggest strong biological effects but also gross technical problems or annotation errors (i.e. sample mislabeling, truncated or empty data, inefficient immunoprecipitation, see Supplementary section 2.1).

4 Select candidates for de-novo epigenome experiments

Selection of candidate factors when designing high-throughput ChIP-Seq, Dam-ID or Hi-C experiments in order to compare a known scenario with an novel one poses important experimental and economical challenges. The *rankFactors* function effectively ranks a set of factors based on how much they contribute towards conservation of their respective chromatin domain identity. Additionally, if domain identity is unknown or incomplete we rely on relative conservation of functional relationships between individual epigenomic marks and genetic elements (i.e. genes, promoters, etc) and how information from experimentally mapped factors can be used to successfully impute presence/absence of others Ernst and Kellis [2015]. In detail, we offer methods based on linear and logistic regression to rank marks based on their ability to predict others, helping the researcher to refine selection of potential candidates for experimental mapping (See Supplementary section 2.2).

5 Comparing epigenomic factor maps.

chroGPS factor maps use MDS to represent genome-wide epigenetic factor co-localization based on similarities between their binding profile overlaps according to a metric of choice. This core methodology is also used as a starting point to identify differences between epigenomic factor associations at different conditions, time points, cell lines, or between different species.

Comparing two epigenomic factor maps from the same species with a rich number of common factors is relatively straightforward, as they share a genomic background over which pairwise similarities *s(i,j)* between their genome-wide binding profile overlaps can be computed using the *distGPS* function, and represented together in a low dimensional space using

MDS. This already produces a joint map which accurately represents putative functional relationships between all elements from both datasets Font-Burgada et al. [2014]. Further adjustment and differential analysis is performed with the function *diffFactors*, which uses Procrustes to identify potential biological differences between both backgrounds after adjusting for possible technical and biological biases. Procrustes matches two sets of points represented in a low-dimensional space by using the information from common elements (landmarks) to compute a transformation involving scaling, translation and rotation to minimize euclidean distances between landmark points while preserving relative spatial configuration within each set. General goodness of fit and magnitude of the observed changes between common factors is measured via Procrustes sum of squares, and statistical assessment of the changes in epigenomic overlap can be performed via permutation tests [Gel et al., 2016]. This approach has been successfully applied to integrate and compare the epigenomic landscape obtained at different time points during genome replication in *Drosophila* S2 cells (See Figure 1 and supplementary section 3.3).

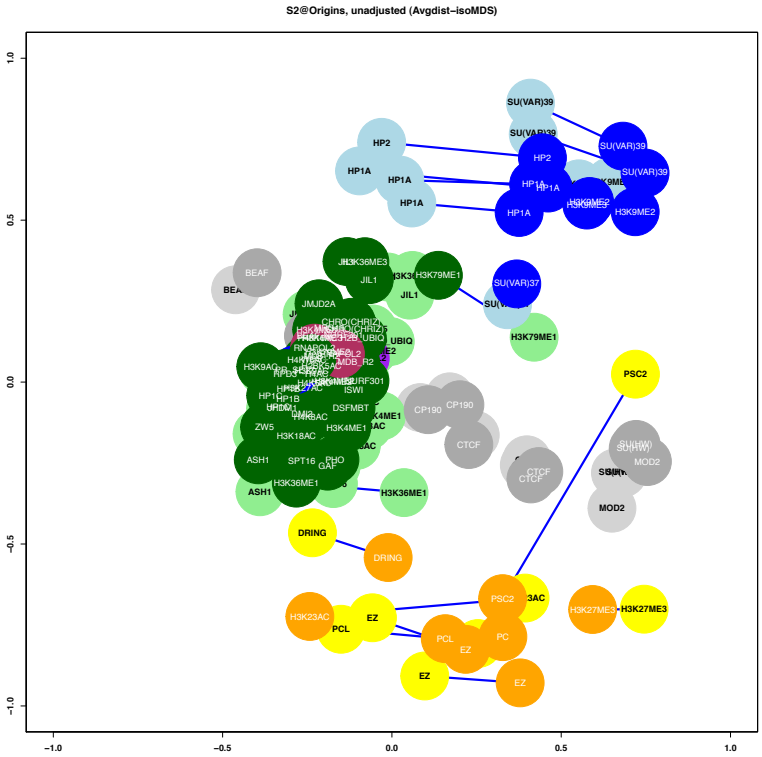

Figure 1: Differential chroGPS-factors map showing the epigenomic landscape differences at *Drosophila melanogaster* Early (light) and Late (dark) time points during genome replication.

6 Comparing epigenomic gene maps

ChroGPS-genes maps represent similarities between genomic features by computing pairwise distances between the binary vectors defined by their epigenetic footprint, that is, their col-

lection of nearby epigenetic marks, and projecting them in a low dimensional space via MDS Font-Burgada et al. [2014]. Downstream functional analysis can be performed by using hierarchical clustering with the *clusGPS* function or using the original and approximated distances with other machine learning or class discovery methods of choice.

- **Detecting epigenomic footprint differences:** The *diffGenes* function uses essentially this strategy to compare two sets of mapped factors from the same or different species using epigenomic footprint information of both datasets simultaneously. First, common factors between both sets are selected. Then, pairwise similarities between their whole-genome epigenetic footprints are computed to generate a map comprising all common genes from both datasets, relying on inter-species gene homology information if necessary. Identification of genes presenting changes in their epigenomic footprints is then straightforward, as they will be indicated by those genes being represented by different points for each set.

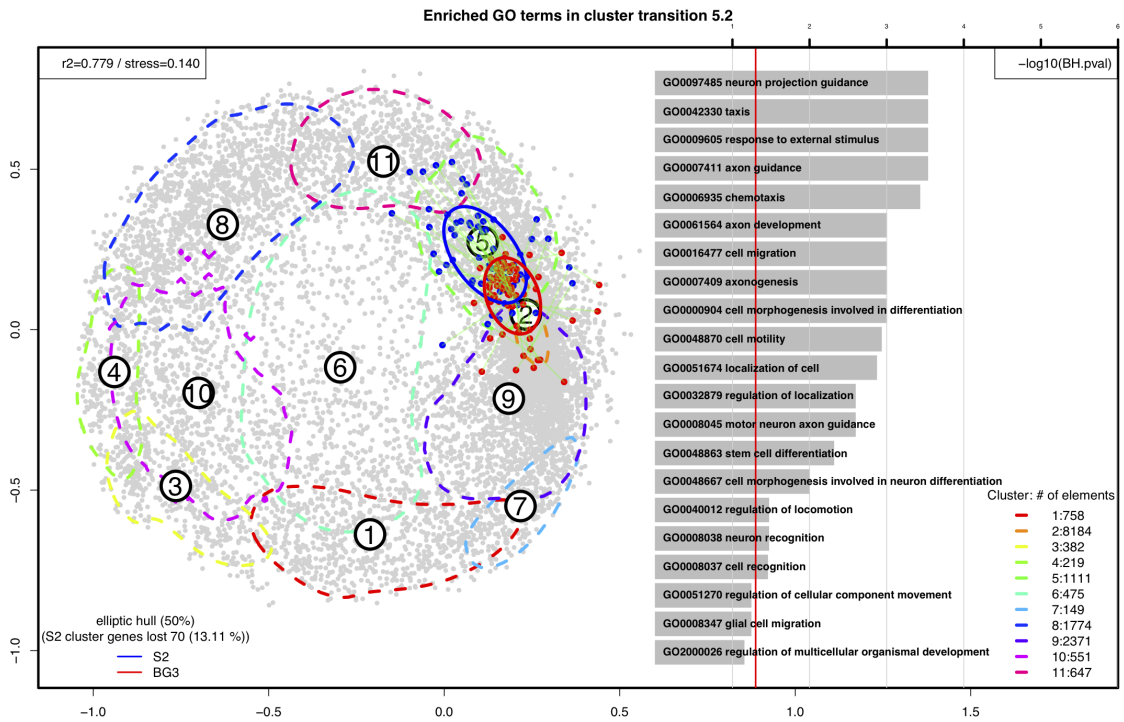

Figure 2: Differential chroGPS-genes map of *Drosophila melanogaster* S2 and BG3 cell lines. Focus is put on genes changing from cluster 5 in S2 (blue, moderate HP1 repression) to cluster 2 in BG3 (red, transcriptionally active). Dashed lines indicate probability contours containing 50 percent of genes for each cluster. On the right, the top 20 enriched Gene Ontology terms for genes involved in the analyzed cluster transition are shown. Selected genes are specially enriched in nervous system and cell differentiation categories.

- **Interpreting loss of epigenomic footprint identity:** A first approach to study those differences is to rank them using Procrustes to compute the residual sum of

squares between points representing differentially located genes, but this offers little insight on the observed changes from a functional perspective. We use unsupervised hierarchical clustering followed by computation of Bayesian posterior probability of cluster classification using a Dirichlet Process mixture of normals Jara et al. [2011] to identify and characterize clusters of genes with similar epigenetic patterns, and to report genes found to be strongly classified under two distinct clusters between both conditions. Akin with results obtained from other high-throughput technologies, candidate gene lists can be used for further downstream functional analysis, such as Gene Ontology enrichment or Gene Set Enrichment Analysis, as well as compared with other genomics data such as changes in gene expression or regulation to further explore interesting hypothesis. We illustrate the use of this strategy to identify and characterize genes potentially involved in epigenomic changes between third instar larvae *Drosophila* S2 cells and BG3 neuronal tissue. (See Figure 2 and Supplementary section 4.3).

References

Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., et al. (2010). The NIH Roadmap Epigenomics Mapping Consortium. Nat. Biotechnol., 28(10):1045–1048.

Celniker, S. E., Dillon, L. A., Gerstein, M. B., Gunsalus, K. C., et al. (2009). Unlocking the secrets of the genome. Nature, 459(7249):927–930.

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature, 489(7414):57–74.

Ernst, J. and Kellis, M. (2015). Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. Nat. Biotechnol., 33(4):364–376.

Filion, G. J., van Bemmel, J. G., Braunschweig, U., Talhout, W., et al. (2010). Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. Cell, 143(2):212–224.

Font-Burgada, J., Reina, O., Rossell, D., and Azorin, F. (2014). chroGPS, a global chromatin positioning system for the functional analysis and visualization of the epigenome. Nucleic Acids Res., 42(4):2126–2137.

Gel, B., Diez-Villanueva, A., Serra, E., Buschbeck, M., Peinado, M. A., and Malinverni, R. (2016). regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. Bioinformatics, 32(2):289–291.

Huber, W., Carey, V. J., Gentleman, R., Anders, S., et al. (2015). Orchestrating high-throughput genomic analysis with Bioconductor. Nat. Methods, 12(2):115–121.

Jara, A., Hanson, T. E., Quintana, F. A., Muller, P., and Rosner, G. L. (2011). DPpackage: Bayesian Non- and Semi-parametric Modelling in R. J Stat Softw, 40(5):1–30.

Lisboa, F. J. G., Peres-Neto, P. R., Chaer, G. M., Jesus, E. d. C., et al. (2014). Much beyond mantel: Bringing procrustes association metric to the plant and soil ecologist's toolbox. PLOS ONE, 9(6):1–9.

Marx, V. (2015). Visualizing epigenomic data. Nat. Methods, 12(6):499–502.

Serra, F., Baù, D., Filion, G., and Marti-Renom, M. A. (2016). Structural features of the fly chromatin colors revealed by automatic three-dimensional modeling. bioRxiv.