Chromatin meta-profiling of healthy and cancer cells using publicly available datasets

Meije MATHÉ[1,2], Victoire BAILLET[1], Claudia CHICA[1]

[1] Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, F-75015 Paris, France
[2] Université Rouen Normandie, UFR Sciences et Techniques, Département de Biologie, Place Émile Blondel, 76821, Mont Saint Aignan

Abstract

The main goal of our study is to find a method to quantify the variability of the epigenomic information and use it to dissect the epigenome's role in cellular differentiation, or the transition from a healthy to a cancer state.

For that purpose, we retrieved from the EpiMap project dataset a selection of epigenome profiles corresponding to distinct cellular states (eg. cancer vs healthy tissue) and including five histone marks associated with activation and repression.

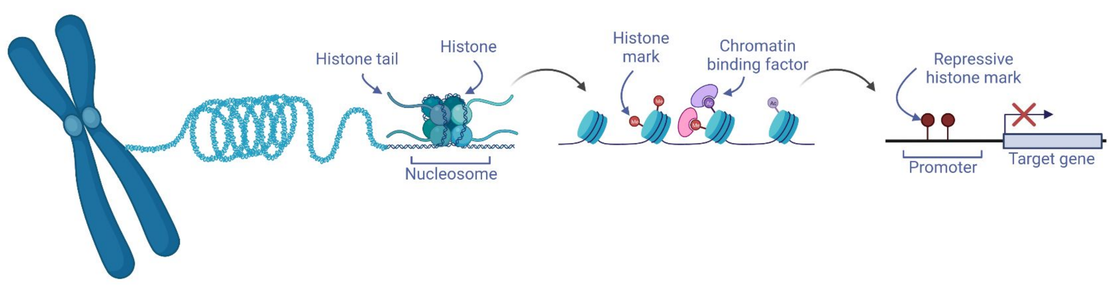
We used Independent Component Analysis to find relations between genomic regions, the epigenomic information they carry, and the biological context. After confirming the validity of our approach, we could show that the distribution of specific marks on active enhancers is diagnostic of the cancer vs. healthy state, and that both the identity and the mark enrichment of these regions differ between the two types of samples. Future work will include the functional analysis of the genomic context for a selection of these regions.

The epigenome, structure and function

The basic unit of chromatin is the nucleosome, a histone octamer around which DNA is wrapped. The N-term (or tail) regions of the histones protrude out of the nucleosome and can undergo chemical modifications, such as the addition of one or several methyl or acetyl groups on some residues.

These modifications, known as histone marks, are involved in the regulation of gene expression by locally modulating the compaction level of DNA, as well as the interaction with chromatin binding factors (transcription factors or remodeling complexes).

Out of the >100 histone marks identified to date, the combination of some of them at particular annotations are diagnostic of specific states of gene activity.

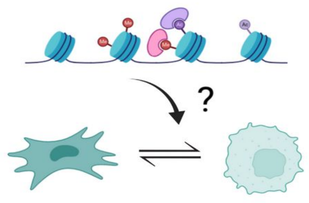


Marks associated with repression	Marks associated with activation
H3K27me3 : - Polycomb-mediated repression (transient repression) **H3K9me3**: - Heterochromatin (long-term repression)	**H3K4me3 (+ H3K27ac)**: - (active) promoters **H3K4me1 (+ H3K27ac)**: - (active) enhancers

Data and approach

Question(s)

What is the epigenome's role (histone marks) on the cellular state variation: from a healthy to a cancer state or from an undifferentiated state to a differentiated one?



How to quantify the variability of the epigenomic information?

- Explore matrix factorisation approaches to summarize (quantify and integrate) the epigenomic information.
- Find relations between genomic regions, the epigenomic information they carry and the cellular state.

Data retrieval

EpiMap project

Cancer samples (lymphoma, leukemia)
Healthy closest matching tissue (B cell, T cell)
5 histone marks

bigWig

ChromHMM [2]

bed — Regions of interest — *promoters, enhancers, polycomb domains*

bed — Histone marks coverage — *200 bp bins, genome-wide*

Data pre-processing

BASH

Remove biased regions
Get common regions
Size threshold

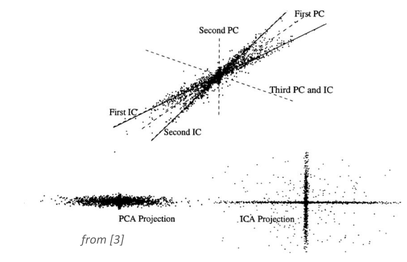
Treat NA values
Concatenate adjacent bins
Log+1 transformation

Batch correction (age + sex)

matrix

ICA input matrix

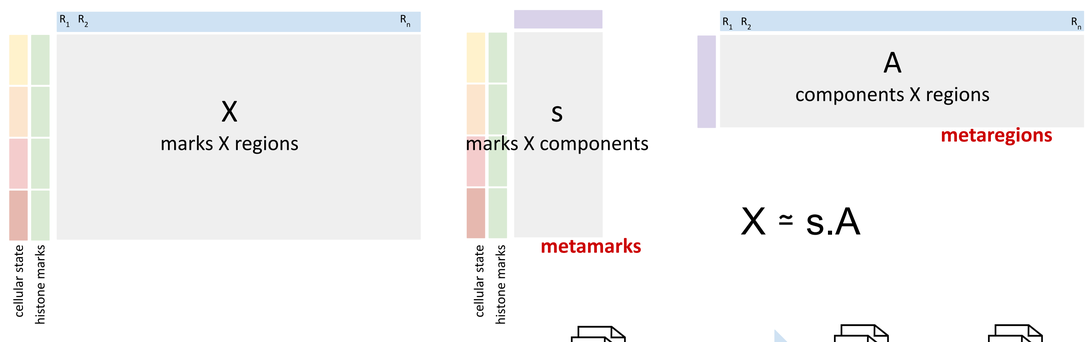
Matrix factorization approaches



from [3]

Complementary approaches to explain data variability:

- Principal Components Analysis (PCA): maximization of the variance, identification of orthogonal components.
- Independent Components Analysis (ICA): maximization of the non-Gaussianity (hence independence between the components), identification of independent components.



Critical ICA Settings:

- Reproducibility of components' identification
- Determination of appropriate number of components

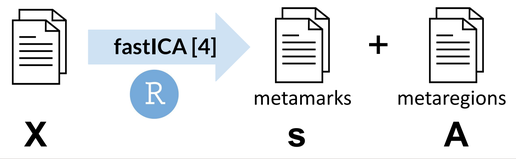
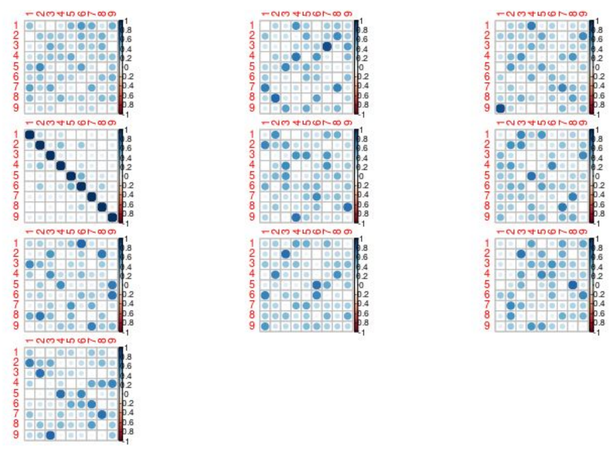


Illustration of ICA approach on lymphoma vs. B cell enhancers

Optimization of the ICA approach for epigenomic data

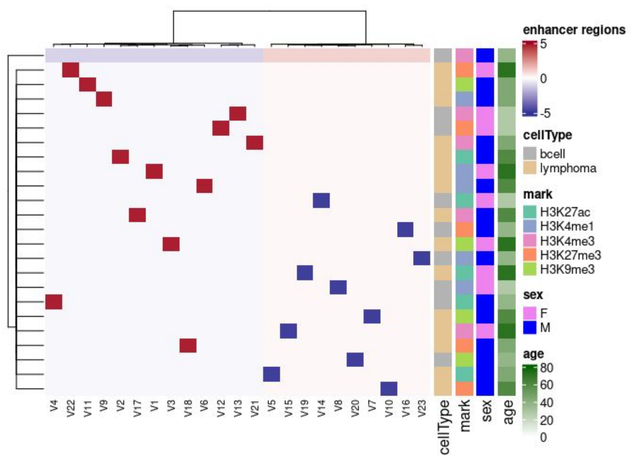
Independent components (ICs) are reproducible among multiple runs with identical parameters on the same dataset.

Correlations across all possible pairwise comparisons between ICs for one run.



For the epigenomic datasets under scrutiny, the maximum number of ICs does not produce relevant results.

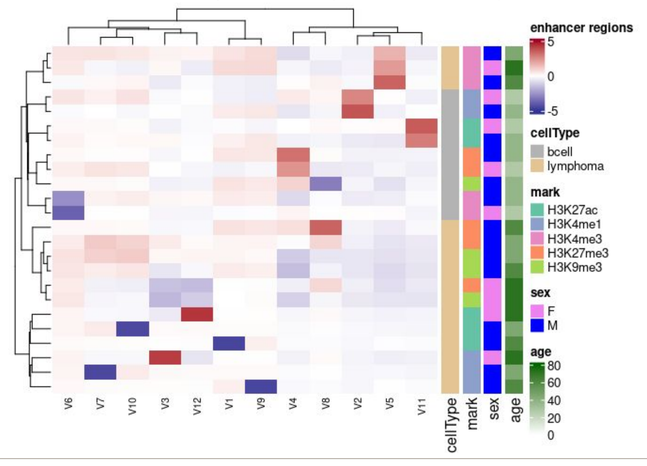
Heatmap of **s** matrix (metamarks, 24 ICs)



Characterization of enhancers associated to different cellular states using ICs

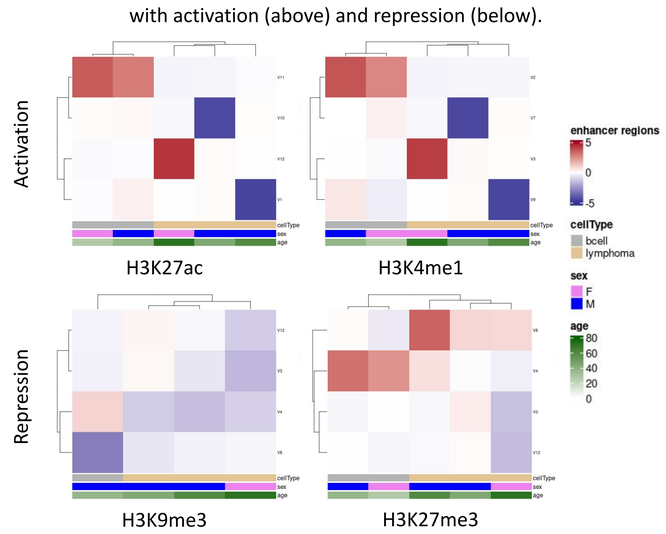
ICs stratify enhancers in cancer vs. healthy samples.

Heatmap of **s** matrix (metamarks, 12 ICs)

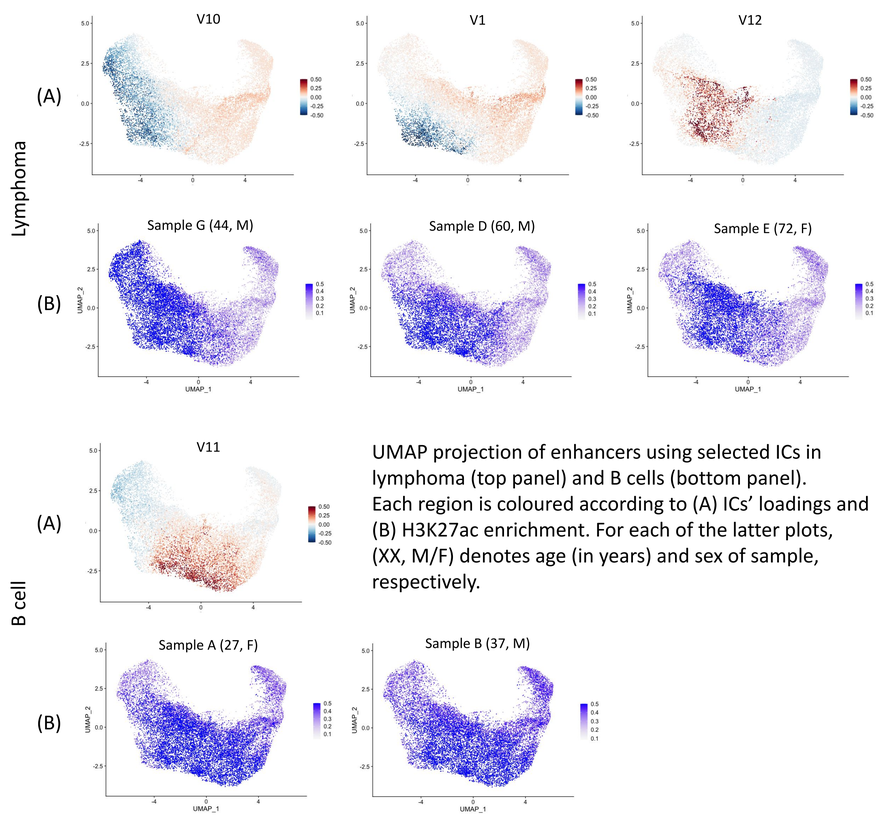


ICs associated to specific histone marks are diagnostic of the cellular state (cancer vs. healthy) **and** the different stages of aging in cancer cells (from young to old). Conversely, ICs describing the marks associated with repression differentiate **only** the cellular states of the samples.

Heatmap of **s** matrix (metamarks) restricted to the samples of marks associated with activation (above) and repression (below).



Active enhancers differ between cancer and healthy samples in both their identity and enrichment. Age adds epigenomic variation only for cancer samples.



UMAP projection of enhancers using selected ICs in lymphoma (top panel) and B cells (bottom panel). Each region is coloured according to (A) ICs' loadings and (B) H3K27ac enrichment. For each of the latter plots, (XX, M/F) denotes age (in years) and sex of sample, respectively.

References

[1] C. A. Boix, B. T. James, Y. P. Park, W. Meuleman, and M. Kellis. (2021) Regulatory genomic circuitry of human disease loci by integrative epigenomics. Nature 590(7845):300–307.
[2] J. Ernst, M. Kellis. (2012) ChromHMM: automating chromatin-state discovery and characterization. Nature Methods, (9):215-216.
[3] M. S. Bartlett, J. R. Movellan and T. J. Sejnowski. (2002) Face recognition by independent component analysis. IEEE Transactions on Neural Networks 13(6):1450-1464.
[4] A. Hyvarinen and E. Oja. (2000) Independent Component Analysis : Algorithms and Applications. Neural Networks 13(4-5):411-430.

UNIVERSITÉ DE ROUEN NORMANDIE — UFR Sciences et Techniques — Master BioInformatique — Institut Pasteur