Bachelor's Degree in Bioinformatics (UPF-UPC-UB)

Final Grade Project

A comprehensive map of predicted enhancers on a large panel of human primary cells, cell lines and tissues

Sara López Ruiz de Vargas[1]

Scientific directors: Trisevgeni Rapakoulia PhD[2] , Professor Dr. Martin Vingron[2].

[1] Bachelor's Degree in Bioinformatics ESCI-UPF, Passeig Pujades 1, 08003 Barcelona, Spain.

[2]Max Planck Institute for Molecular Genetics, Department of Computational Molecular Biology, Transcriptional Regulation Group, Address Ihnestraße 63, 14195 Berlin, Germany.

Abstract

Motivation: Enhancers are genomic regulatory elements that activate or increase the transcription of a target gene. The location of active enhancers differs among cell types and the exact mechanism that they regulate gene expression is not well known. Enhancers are extremely cell type specific and have been found to have critical roles in cell differentiation, morphogenesis and development. Many computational methods predict condition specific enhancers based on the chromatin states but a comprehensive application of these methods in several cell types and conditions is still limited. Mapping enhancer elements in different human cell types can lead us to a better understanding of gene regulation and phenotypic diversity among different cells.

Results: Here we applied an enhancer prediction method to an extensive set of epigenetic profiles from the ENCODE project. We identified enhancer elements in 104 human primary cells, cell lines, and tissues. This extensive generation of enhancer profiles allowed us to identify groups of cell types that are pretty similar in terms of enhancer usage and cell types that differ significantly. A case study on adult brain and embryo placenta tissues with quite different sets of active enhancers revealed cell-type specific enhancers that regulate the expression of genes linked to the tissue phenotype. Further exploitation of additional cell groups can reveal the lineage-specific enhancers. We provide an extensive resource of enhancer elements in different cell types that may support and complement future gene regulation analysis studies.

Supplementary information: Supplementary data and scripts are available at the GitHub link **https://github.com/slrvv/Scripts** , Supplementary figures are given as a separate file.

1 Introduction

Enhancers are short DNA fragments bound by proteins and transcription factors and activate or increase the transcription of a target gene [1]. They are vital elements in gene regulation [2], controlling the cell structure and function, and are the basis for cell differentiation and morphogenesis. Enhancers can regulate the expression of genes independent of orientation and over large genomic distances. However, recently introduced conformation capture methods such as Hi-C revealed [3] the three-dimensional structure of the genome, which brings enhancers into close proximity with target genes. Enhancer-gene interactions are mainly restricted into subchromosomal structures called topological associated domains (TADs)[4] ,

The activity of enhancers is extremely cell type, time-point and condition-specific. This dynamic nature of enhancer activity precisely determines where, when under what conditions and which genes are transcribed [2]. While there are hundreds of thousands of enhancer regions for a particular tissue only a specific number of enhancers are brought into proximity with the promoters that they regulate [2] adding a complexity level to their localization in different cells and conditions.

Thousands of enhancers have been experimentally identified using reporter assays [5]. However, the experimental identification is a cost- and time-consuming process if we consider the different organisms and cell types. Given that, the field of enhancer prediction computational methods for enhancer prediction has become the norm [6].

Enhancer DNA regions have particular properties that can be determined using high-throughput sequencing of histone modifications [7] and other epigenomic marks. Active enhancers act as binding regions for various transcription factors [2], thus are located in accessible chromatin indicated by the enrichment of certain histone marks such as H3K4me1[8] or H3K27ac[9]. They usually have low methylation levels [10] and frequently produce short and unstable RNA fragments (eRNAs)[11].

Several computational methods have been proposed to predict active enhancers from the combinations of the properties [12] mentioned above. They generally use machine-learning algorithms to learn the epigenetic profiles of enhancers active in a given cell type and then predict active enhancers in additional cell types. A prominent example of such methods is ChromHMM [13], which uses a multivariate Hidden Markov Model (HMM). ChromHMM can be classified as an unsupervised learning method. On the other hand, supervised methods rely on a gold-set of known and validated enhancer and non-enhancer regions from which features that distinguish active enhancers can be learned [6]. Examples of these methods are REPTILE [14] and CRUP [15], both of which use random forest classifiers. CRUP, which stands for Condition-specific Regulatory Units Prediction, is an algorithm developed by Vingron's lab in the Max Planck Institute for Molecular Genetics. CRUP can be applied in different cell types and species without retraining. It combines the prediction of active enhancers (CRUP-EP), with condition-specific enhancer dynamics (CRUP-ED) and enhancer-target identification (CRUP-ET), in a three-step pipeline. CRUP-EP was found to have higher and more stable performance in enhancer prediction than ChromHMM and REPTILE across different cell types and species [15].

Although these computational approaches predict condition specific enhancers based on chromatin states, a comprehensive application of these methods in several cell types and tissues is still limited. Mapping enhancer elements in different human cell types can lead us to a better understanding of gene regulation and phenotypic diversity among different cells. The ENCODE consortium [16] recently released the Registry of candidate cis-Regulatory Elements (cCREs), which annotates cell type specific regulatory elements based on DNase-seq, H3K4me3, H3K27ac, and CTCF ChIP-seq data [17]. While all these data may not be available for many cell types, CRUP-EP requires only three histone marks measured by ChIP-seq as input, namely H3K4me1, H3K4me3, and H3K27ac, making it broadly applicable in many available epigenetic datasets and non model organisms.

Objectives

In this project, we are going to use the CRUP-EP algorithm to generate enhancer predictions in all available ENCODE human epigenetic datasets. Our aim is to generate a high quality comprehensive map of cell-type-specific enhancers in various human cells and tissues. The research objectives of the current thesis can be summarized to the:

1)Download the three histone mark Chip-seq profiles for all primary cells, cell lines and tissues provided by the ENCODE consortium [16] .

2)Applying the CRUP-EP module to predict the cell-type-specific enhancers.

3)Analysis of enhancer predictions to identify groups of cell types that are pretty similar in terms of enhancer usage and cell types that differ significantly.

2 Methods

2.1 Building the comprehensive map of enhancer and promoter predictions

2.1.1 Downloading the ENCODE data

The data used in this analysis were downloaded from the ENCODE consortium [16]. We selected samples where the following histone ChIP-seq [7] profiles were available: H3K27ac, H3K4me1 and H3K4me3, as well as the Control ChIP-seq data for the specific cell type. All of the data we downloaded are Homo Sapiens GRCh38 genome assembly released status and not perturbed.

We developed our own script for downloading the data because if we had used the ENCODE tools, we would have to manually check that each HM ChIP-seq came from the same sample.
The available ENCODE data were filtered using PostgreSQL [18] and downloaded using the

ENCODE downloader scripts, developed by the Kundaje lab [19].

In cases where the data could not be downloaded using the ENCODE downloader scripts [19], we utilized the batch downloading tools from the ENCODE, searching directly for the data filtered in the PostgreSQL step.

In case that we had biological or technical replicates for the same sample, we used SAMTOOLS [20] to merge them. All the scripts used to prepare the data and run CRUP over the samples are provided in the supplementary materials [21]. CRUP was run on human genome assembly GRCh38 with five cores.

2.1.2 Running CRUP

CRUP-EP [15] script was used to generate the enhancer prediction scores for every cell type [22]. CRUP, before applying CRUP-EP , does quantile normalization. CRUP-EP uses a combination of two binary random forest classifiers, both consisting of 100 decision trees. The first classifier learns to distinguish between active genomic regions (promoters and enhancers) and inactive genomic regions. The second classifier distinguishes between active promoters and active enhancers. In the end, we have the probability of a region being an enhancer given that the region is active. This probability is assigned to each 100 bp bin in the fragmented genome. The probability is computed as follows :

$$P(bin_x = active\ enhancer) =$$

$$P\ (bin_x = active)P\ (bin_x = active\ enhancer|bin_x = active \quad (1)$$

We made some minor modifications to the original script of CRUP-EP to also generate promoter predictions across the genome, and we called this new script CRUP-PP. CRUP-PP uses the same random forest classifiers but it computes the probability of having an active promoter. Therefore the only change is the second probability that is then assigned to the 100 bp bins. In this case, it is computed as follows :

$$P(bin_x = active\ promoter) =$$

$$P\ (bin_x = active)(1 - (P\ (bin_x = active\ enhancer|bin_x = \quad (2)$$

2.1.3 Grouping overlapping enhancers

We applied a clustering method to group together the overlapping nearby enhancer in different cell types. This method was developed by Emel Comak, a PhD student at the Vingron's Lab. First, the method generates clusters of nearby enhancers using bedtools[23]. After merging, many enhancers are very large and thus the method parses the large enhancer clusters into subclusters. Enhancers are grouped into these subclusters according to the mean and standard deviation of the cluster's enhancer sizes.

2.2 Analysis of predicted enhancers

2.2.1 Principal Components Analysis

Principal Components Analysis (PCA)[24][25] was applied using the prcomp function [26] from the stats [27] R package, as a preliminary analysis to check if the data coming from different labs present bias. The PCA was applied to a contingency table. The columns of this table are the different samples and the rows are the predicted enhancers. Each cell has either a one or a zero, one if the enhancer was predicted in this sample zero otherwise.

2.2.2 Correspondence Analysis and Association Plots

Correspondence Analysis [28] was performed using APL shiny app [29]. The analysis was applied to the same contingency table as the PCA. We also used APL to make the Association Plots [30] and get the enhancers highly associated with clusters of interest. To select the optimal number of dimensions for the Association Plots, we used APL's Elbow Rule option.

Association Plots are two-dimensional plots that depict variables that make every cluster distinctive. The association plot is derived from the Correspondence Analysis and helps visualize information that can become invisible in a biplot, especially if we are working with very high-dimensional data [30].

2.2.3 GO Enrichment analysis

Gene Ontology [31] enrichment analysis [32] was applied on the two most dissimilar clusters, according to the Correspondence analysis and PCA. More precisely, we first mapped the cluster specific enhancers we had from the APL analysis to the closest gene using the GENECODE (gencode.v38.annotation.gtf) annotation files and the GenomicRanges R package [33]. We used clusterProfiler [34] to compute the GO enrichment analysis using the Biological Process Ontology.
The enriched Biological Processes were further clustered into groups of closely related terms, using the REVIGO software [35](allowed similarity = 0.7 and semantic similarity measure Resnik).

3 Results and Discussion

3.1 Generation of enhancer profiles

We downloaded the H3K27ac, H3K4me1 and H3K4me3 profiles for 104 different human primary cells, cell lines and tissues from the ENCODE [16] project (Table 1). Replicates from the same cell type were merged in a single profile for every cell type (Methods). We then applied the CRUP-EP module using as input the ChIP-Seq for every cell type. We obtained enhancer predictions for all the 104 distinct cell types (Fig.1).
We can notice that the highest number of enhancers are predicted in tissues, which have a more complex composition compared to primary cells and cell lines. There is an extremely low number of enhancers predicted in the cell line PC-3, indicating a deficiency in the available epigenomic data for this cell line. The size of CPUP enhancers is 1100 bp and the average number of predicted enhancers per cell type is 62,337±13,806.

Table 1. Number of samples in the repository of predictions

Type of sample	Single-ended	Paired-ended
Tissues	59	0
Cell lines	28	9
Primary cells	5	3

3.2 Clustering nearby enhancers

Once we obtained the whole collection of enhancer profiles, we used them to identify similarities and differences between the cell types in their active enhancers. However, the dimensionality of our data was excessively high, with 104 samples and more than 3 million predicted enhancers in total.

Many of the predicted enhancers were located in very similar genomic positions across the cells. To reduce the number of predicted enhancers per cell type, we used a clustering method, developed by Emel Comak (section 2.1.3), that groups together overlapping enhancer regions in different cells and categorizes them into exclusive enhancer regions. After grouping, we obtained 634,982 enhancers, with a mean enhancer length of 1823 bps ±1175 bps.

3.3 Identifying dissimilar cell types

We applied Principal Component Analysis (as described in section 2.2.1) in our enhancer profiles to observe distinct groups of cell types and any potential batch effects. The PCA biplot (Fig. 2) shows an easily distinguishable cluster of samples consisting of angular gyrus, caudate nucleus, cingulate gyrus, middle frontal area, hippocampus layer, and temporal lobe adult tissues. All these samples belong to brain-related tissues. We can observe another distinct cluster of samples forming with the chorionic villus, chorion, placenta, and placental basal plate embryo tissues on the other side of the PCA plot. We do not observe any apparent bias effect due to the fact that the ChIP Seq experiments provided by ENCODE were generated in different labs. To obtain higher resolution, we also performed PCA analysis per chromosome which is provided in the Supplementary Material.

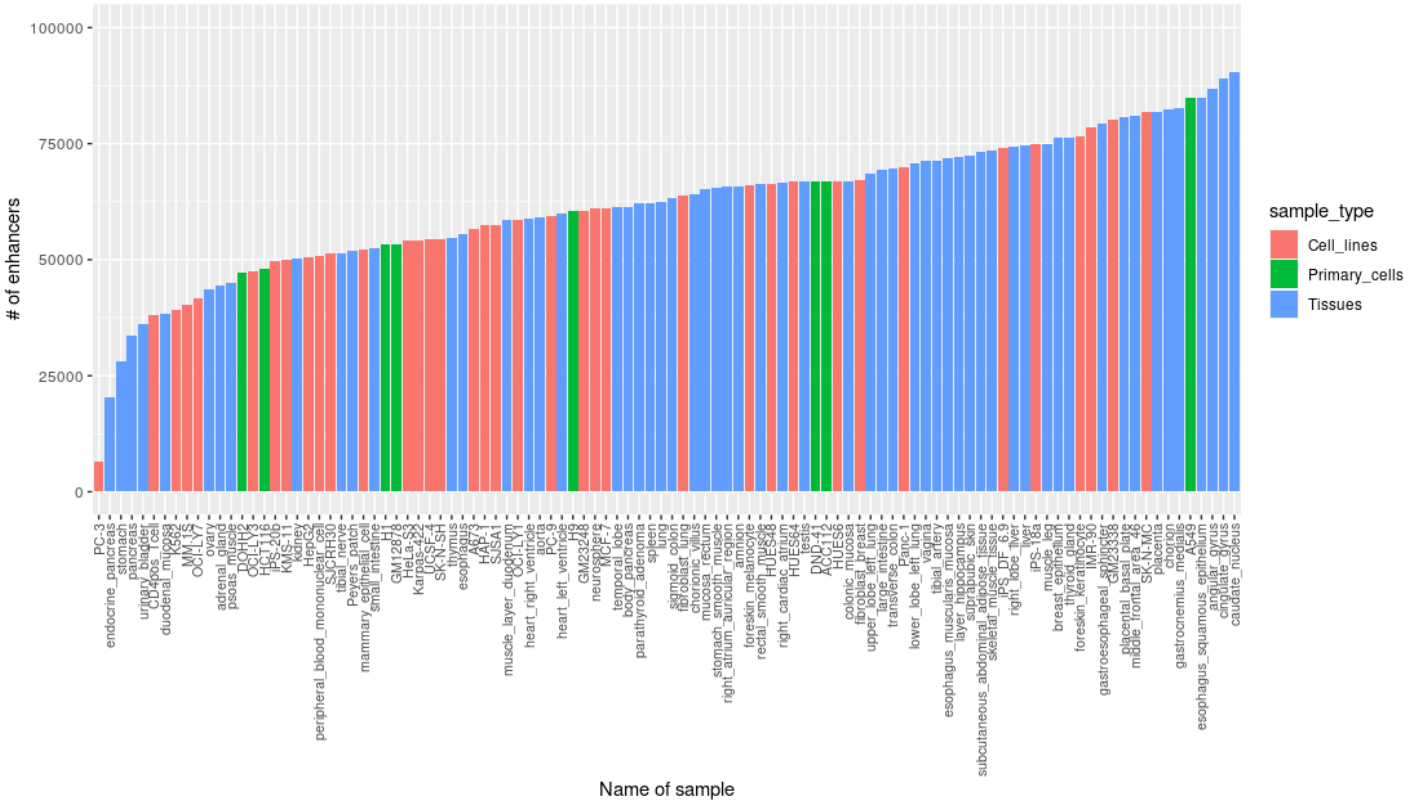

Fig 1. The number of CRUP enhancers predicted for each sample. The type of sample is shown in different colours.

Fig 2. PCA plot of the predicted enhancer profiles. The labs that contributed the data to the ENCODE are shown in different shapes and the type of sample in different colours.

3.4 Identifying cell type specific enhancers

Next, we applied Correspondence Analysis to the predicted enhancer profiles. Unlike Principal Component Analysis, Correspondence Analysis (CA) allows us to detect groups of enhancers specific to a cluster of samples. CA uses chi-square distances instead of Euclidean distances and thus reflects relationships between conditions and observations simultaneously [28]. Moreover, this method also allows us to visualize conditions (cell types) and observations (enhancers) in the same space.

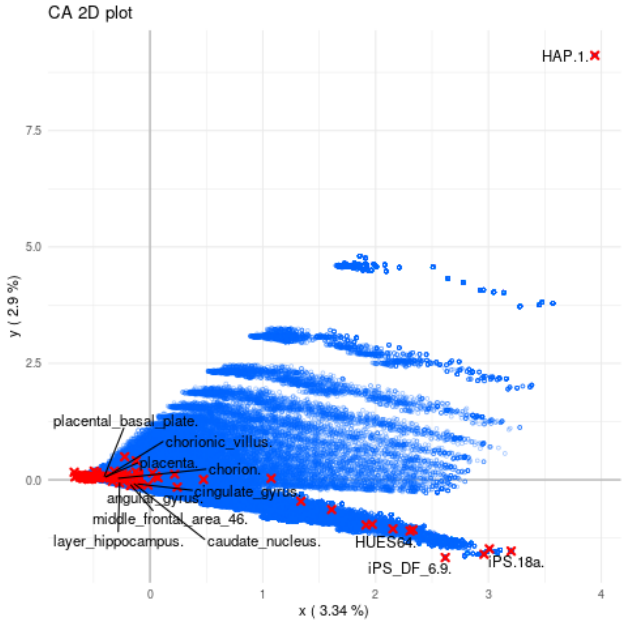

Fig 3. CA biplot on the enhancers. The enhancers are shown as blue points and the samples as red crosses.

In the Correspondence Analysis biplot (Fig. 3) we can see that the samples that differentiate the most from the whole are the cell lines, such as iPS-18a, HUES64 and iPS DF 6.9. It seems that cell lines have more homogeneous enhancer profiles than tissues which is expected since tissues consist of many different cell types. All the tissues seem to cluster together without any observable difference between them. However a more detailed analysis in chromosomes 13 and 14 revealed again the dissimilarity between brain and the placenta/chorion groups, in agreement with the PCA plot of Fig.2. The PCA plot for all chromosomes and CA plots for chromosomes 13 and 21 are provided in the Supplementary Material.

To recover information that may be lost in the above biplots, we used the Association Plots (APL) method. An APL corresponds to a specific cell type/cluster; it is derived from the Correspondence Analysis and depicts variables, in our case enhancers, that make every cell type/cluster distinctive.

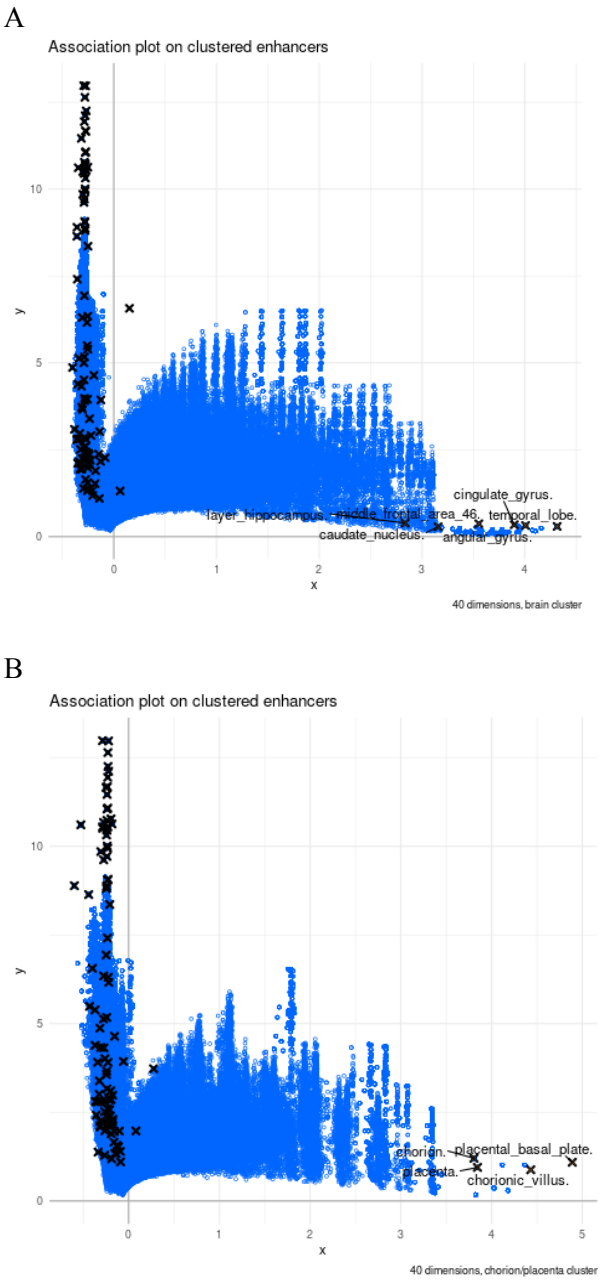

Fig 4. Association plots. A. Association plot for the brain cluster. B. Association plot for the placenta/chorion cluster.Enhancers are shown as blue points and samples as black crosses.

When a cluster of cells is well defined, all ubiquitous enhancers are located on the y-axis while the cluster-specific enhancers are placed along the x-axis. The farther from the x origin are the enhancers; the more specific they are for the cluster [30]. We applied the association plot analysis in the adult brain and embryo placenta tissue clusters (Fig 4A, 4B) and we managed to find specific enhancers for both clusters (enhancers found at the right end of the x-axis).

3.5 Case study : Brain and placenta clusters

As a proof of concept, we focused our further analysis on the brain and placenta tissue clusters, which are the most dissimilar groups based on the PCA and CA analysis. We extracted the list of enhancers that are associated with these two groups from the APL software. APL ranks the enhancers according to a specificity score. The higher the score, the stronger the enhancer is associated with the cluster [30]. Based on the distribution of specificity scores for both clusters (Fig 5A, B), we selected enhancers with a score higher than three as the more specific to the cluster.

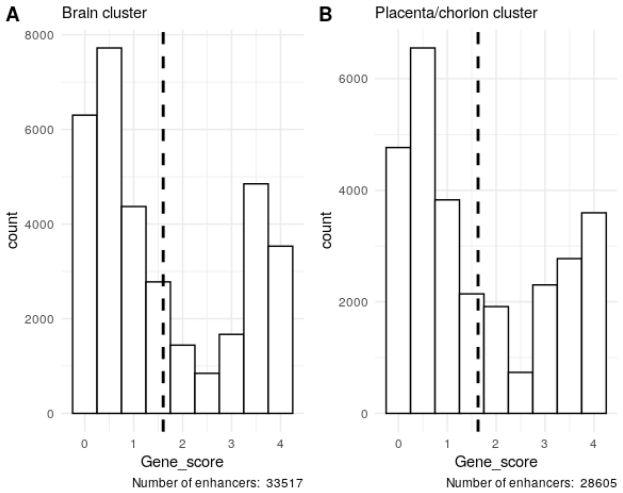

Fig 5. Histogram of APL specificity scores A) for the adult brain-specific-enhancers and, B) embryo placenta-specific-enhancers. The mean of both distributions is shown with a horizontal dashed line.

Next we annotated the specific enhancers for each cluster to the closest gene based on the genomic linear distance. Although enhancers do not always regulate the expression of their nearest gene, this simplification works quite well in the absence of Hi-C conformation data. We performed Gene

Ontology (GO) enrichment analysis of the target genes and we visualized the most enriched biological processes in fig 6A and B for the both clusters.
The enrichment analysis for the adult brain cluster (Fig 6A) shows that most enriched biological processes, such as neuron projection, trans-synaptic signaling, cognition, forebrain development, and learning, are particular to the brain and present very low p-values.
The enrichment analysis for the embryo placenta and chorion cluster, on the other hand, revealed diverse enriched biological processes. However, it seems to be a clear trend towards development, morphogenesis, and differentiation processes, which are expected to be enriched in embryo tissues.
Further clustering of the enriched GO terms using semantic similarity [35] (Supplementary figures) confirmed the domination of brain and development related processes in the enrichment analysis of two clusters.

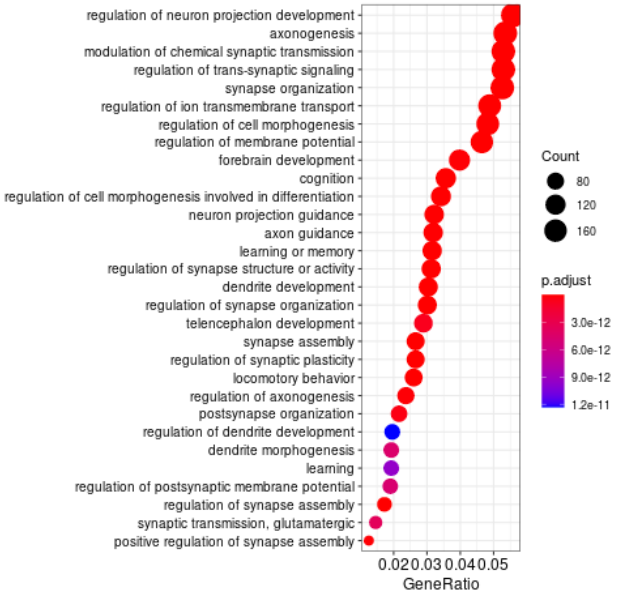

Fig 6 A. Dotplot for the GO enrichment analysis of the brain cluster. On the x-axis we have the Gene ratio (percentage of the total genes enriched in the GO term). On the y-axis we show the enriched terms. Gene count (number of enriched genes in a GO term) is shown with the different sizes of the dots. The dots are colored according to the adjusted p-values.

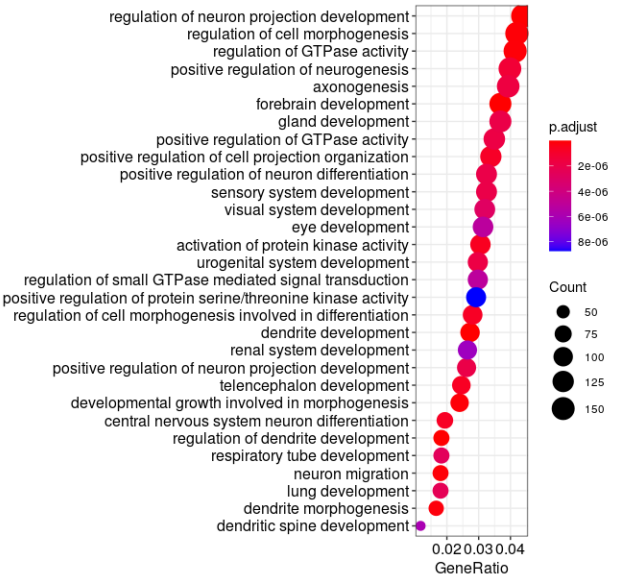

Fig 6 B. Dotplot for the GO enrichment analysis of the placenta/chorion cluster . On the x-axis we have the Gene ratio (percentage of the total genes enriched in the GO term). On the y-axis we show the enriched terms. Gene count (number of enriched genes in a GO term) is shown with the different sizes of the dots. The dots are colored according to the adjusted p-values.

We further noticed that the most enriched biological process for both the adult brain and embryo placenta cluster is the regulation of neural projection development. Since the two clusters do not share any cluster specific enhancer region derived from the association plots, we look into the overlap of the closest genes of the enhancers between these two clusters (Fig 7). The Venn diagram reveals that the two clusters share quite a high number of genes, regulated by different sets of enhancers. This finding is quite interesting, and may confirm the hypothesis that a gene can be regulated by different enhancers in different cells and conditions [2]. However we have to also take into account that the closest gene simplification may mislead in some false enhancer gene associations.

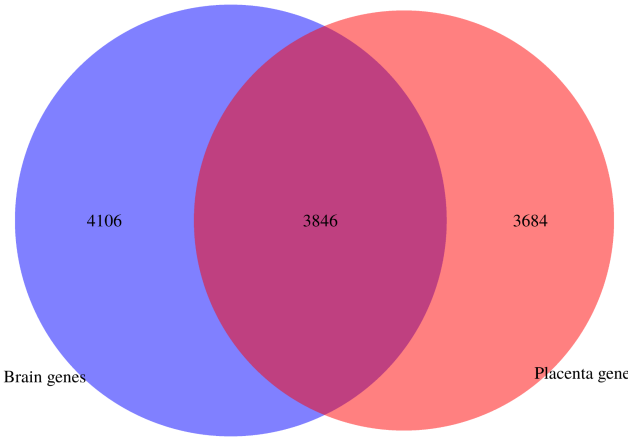

Fig 7. Venn diagram of the genes regulated by the brain enhancer set and the placenta enhancer set. Placenta genes are shown in green and brain genes are shown in orange.

3.7 Discussion

In this thesis, we generated a comprehensive map of predicted enhancers in 104 human primary cells, cell lines, and tissues. This large collection of enhancer profiles allowed us to identify similar cell types in terms of their enhancer usage and cell types that are quite different. We discovered distinct clusters of cell types and enhancers specific to these clusters by performing an extensive analysis using Principal Component Analysis, Correspondence Analysis, and Association Plots. We focused our analysis on two sets of tissues that differed substantially, consisting of adult brain and embryonic placenta samples, respectively. Further analysis of the cluster-specific enhancers revealed that the possible regulated genes are linked to the tissue phenotypes.

Mapping enhancer elements in different human cell types can help us to better understand gene regulation and phenotypic diversity among different cells. Although there are already annotated datasets of cis Regulatory Elements[36] in various human cell types, the lack of available epigenetic and open chromatin experiments makes the annotation incomplete. On the other hand, the CRUP algorithm requires only three histone marks and is more applicable in many cell types and organisms. The same time CPUP outperforms well established enhancer prediction methods[2].

We do not observe any apparent batch effect because the ChIP Seq experiments provided by ENCODE were generated in different labs. CRUP algorithm normilizes the data before the prediction and alleviates any bias coming from different experiments.

We additionally perform a multistep analysis of the predicted enhancers to identify distinct sets of cells that share the same enhancers. We focus mainly on two groups of cells, consisting of adult brain and embryo placenta related tissues.

In the association plot for the embryo placenta and chorion cluster (Fig 5), although we identified enhancers related to this cluster, they are fewer and less strongly related to the cluster compared to the brain cluster association plot. This observation suggests that maybe the embryo placenta cluster is less homogeneous than the brain cluster.

We annotated the cluster specific enhancers with their closest gene according to the linear distance. We understand that this simplification may produce false positive associations, but the identification of gene targets of enhancers is an open research topic [37], out of the scope of this thesis. Future work that applies a more sophisticated mapping method may result in more precise findings.
However, the GO enrichment analysis of the closest genes of the cluster-specific enhancers confirmed the validity of our analysis. The enriched processes in the brain cluster are highly related to the brain " biological program", giving support to our finding that these enhancers are very specific to brain tissue. Developmental processes were found enriched in the embryo placenta tissue, although many other diverse biological processes came up from the analysis. Maybe it can be explained from the heterogeneous composition of the specific clustered, already noticed in the association plots.

There is an interesting observation that the two distinct clusters, although they do not share specific enhancers, seem to regulate a common set of genes. This finding needs further exploration to examine if it confirms the hypothesis that different enhancers regulate the same genes in different tissues, or if it

comes from the approximation of the closest gene annotation.

The current analysis is restricted to the time limitations of the bachelor's degree thesis. Nevertheless, we generated a valuable source of predicted enhancers in a large panel of human cells. Although we focused only on the two most dissimilar cell clusters, based on the association plots, further analysis in the rest of the cell types can reveal more distinct groups of cells in terms of enhancer activity and lineage-specific enhancers. Thus, we provide a comprehensive map of predicted enhancer elements in different cell types that may be analyzed in several ways either by us or by future gene regulation analysis studies

4 Conclusions

This study provides a comprehensive map of predicted enhancers across a large panel of different cell types. Although there have been many computational methods developed in the field of enhancer prediction, a comprehensive application of them on a large panel of cell types is still limited.

We have developed a multi-step analysis to identify distinct sets of cells that share the same enhancers. We also developed a proof of concept that validates our analysis because it links the enhancers specific to the analyzed clusters to possible genes related to the phenotype of the cell types that make up these clusters.

We noticed that although the brain and placenta/chorion tissues do not share common enhancers, they seem to regulate a common set of genes. We understand that, since we mapped the enhancers to the closest gene this simplification may be leading us to false enhancer-target-associations. But, it could also confirm the hypothesis that different enhancers regulate the same genes in different tissues and it needs further examination.

Our contribution can be summarized as follows:

1) Generation of a comprehensive map of predicted enhancers in 104 cell lines, primary cells and tissues.
2) Identification of distinct clusters of cells in terms of their active enhancers.
3) Detection of cluster-specific enhancers.
4) Case study on two clusters that differed substantially (embryo placenta and adult brain cluster) and verification that the possible regulated genes of these enhancers link back to the phenotype of the tissues.

Our analysis provides a valuable resource to understand how enhancers work in different cell types that may complement future gene regulation studies.

Acknowledgements

I would like to express my deepest gratitude to my project tutor, Trisevgeni Rapakoulia, for her guidance, patience, and the amount of time and work that she has put into this project. You made my stay in Berlin in the middle of a global pandemic that much more enjoyable and for that I am very grateful.

I would also like to give my sincere thanks to Martin Vingron for letting me perform my project in his lab, for the hospitality I have been given there and for wanting to continue with this collaboration.

I want to thank Emel Comak and Stefan Haas for their contribution to the project, and being there to discuss and give us ideas. Thanks to Elzbieta Gralinska for letting us use her APL app and always being there to help, answer any questions and discuss results. I also extend my gratitude to the entire team of the Lab for their hospitality and being so welcoming.

Finally, thanks to those that have accompanied me in these four years of Bachelor studies : To my family, my friends, my boyfriend and my two lovely cats.

References

1. Banerji, J., Olson, L. & Schaffner, W. A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. *Cell* **33,** 729–740 (1983).

2. Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A. & Bejerano, G. Enhancers: Five essential questions. *Nature Reviews Genetics* **14,** 288–295 (2013).

3. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159,** 1665–1680 (2014).

4. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485,** 376–380 (2012)

5. Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser - A database of tissue-specific human enhancers. *Nucleic Acids Res.* **35,** (2007).

6. Hariprakash, J. M. & Ferrari, F. Computational Biology Solutions to Identify Enhancers-target Gene Pairs. *Computational and Structural Biotechnology Journal* **17,** 821–831 (2019).

7. Barski, A. *et al.* High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell* **129,** 823–837 (2007).

8. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* **39,** 311–318 (2007).

9. Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 21931–21936 (2010).

10. Schmidl, C. *et al.* Lineage-specific DNA methylation in T cells correlates with histone methylation and enhancer activity. *Genome Res.* **19,** 1165–1174 (2009).

11. Melgar, M. F., Collins, F. S. & Sethupathy, P. Discovery of active enhancers through bidirectional expression of short transcripts. *Genome Biol.* **12,** 1–11 (2011).

12. Kleftogiannis, D., Kalnis, P. & Bajic, V. B. Progress and challenges in bioinformatics approaches for enhancer identification. *Brief. Bioinform.* **17,** 967–979 (2016).

13. Ernst, J. & Kellis, M. Chromatin-state discovery and genome annotation with ChromHMM. *Nat. Protoc.* **12,** 2478–2492 (2017).

14. He, Y. *et al.* Improved regulatory element prediction based on tissue-specific local epigenomic signatures. *Proc. Natl. Acad. Sci. U. S. A.* **114,** E1633–E1640 (2017).

15. Ramisch, A. *et al.* CRUP: a comprehensive framework to predict condition-specific regulatory units. *Genome Biol.* **20,** 227 (2019).

16. Feingold, E. A. *et al.* The ENCODE (ENCyclopedia of DNA Elements) Project. *Science* **306,** 636–640 (2004).

17. Abascal, F. *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583,** 699–710 (2020).

18. PostgreSQL: Documentation: 13: PostgreSQL 13.3 Documentation. Available at: https://www.postgresql.org/docs/13/index.html. (Accessed: 9th June 2021)

19. Kundaje Lab. GitHub - kundajelab/ENCODE_downloader: Downloader for ENCODE. Available at: https://github.com/kundajelab/ENCODE_downloader. (Accessed: 9th June 2021)

20. Samtools. Available at: http://www.htslib.org/. (Accessed: 9th June 2021)

21. López Ruiz de Vargas, S. GitHub - slrvv/Scripts: Supplementary material. Available at: https://github.com/slrvv/Scripts. (Accessed: 9th June 2021)

22. Heinrich, V. GitHub - VerenaHeinrich/CRUP: CRUP collapses different layers of epigenetic information into a single list of regulatory units consisting of dynamically changing enhancers and target genes. Available at: https://github.com/VerenaHeinrich/CRUP. (Accessed: 19th June 2021)

23. bedtools: a powerful toolset for genome arithmetic — bedtools 2.30.0 documentation. Available at: https://bedtools.readthedocs.io/en/latest/. (Accessed: 10th June 2021)

24. Pearson, K. LIII. On lines and planes of closest fit to systems of points in space . *London, Edinburgh, Dublin Philos. Mag. J. Sci.* **2,** 559–572 (1901).

25. Hotelling, H. Relations Between Two Sets of Variates. *Biometrika* **28,** 321 (1936).

26. prcomp function - RDocumentation. Available at: https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/prcomp. (Accessed: 10th June 2021)

27. stats package - RDocumentation. Available at: https://www.rdocumentation.org/packages/stats/versions/3.6.2. (Accessed: 18th June 2021)

28. Hirschfeld, H. O. A Connection between Correlation and Contingency. *Math. Proc. Cambridge Philos. Soc.* **31,** 520–524 (1935).

29. Gralinska, E. GitHub - elagralinska/APL: Explore your data and find cluster-specific genes using Association Plots. Available at: https://github.com/elagralinska/APL. (Accessed: 9th June 2021)

30. Gralinska, E. & Vingron, M. Association Plots: Visualizing associations in high-dimensional correspondence analysis biplots. *bioRxiv* 2020.10.23.352096 (2020). doi:10.1101/2020.10.23.35209

31. Harris, M. A. *et al.* The Gene Ontology project in 2008. *Nucleic Acids Res.* **36,** D440 (2008).

32. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 15545–15550 (2005).

33. GenomicRanges package - RDocumentation. Available at: https://www.rdocumentation.org/packages/GenomicRanges/versions/1.24.1. (Accessed: 19th June 2021)

34. Yu, G., Wang, L. G., Han, Y. & He, Q. Y. ClusterProfiler: An R package for comparing biological themes among gene clusters. *Omi. A J. Integr. Biol.* **16,** 284–287 (2012).

35. Supek, F., Bošnjak, M., Škunca, N. & Šmuc, T. Revigo summarizes and visualizes long lists of gene ontology terms. *PLoS One* **6,** 21800 (2011).

36. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507,** 455–461 (2014).

37. Gasperini, M., Tome, J. M. & Shendure, J. Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nature Reviews Genetics* **21,** 292–310 (2020).