Bioinformatics Program
M1 Genomics, Informatics and Mathematics for Health and Environment

Ala Eddine BOUDEMIA

Internship Report

Integrated Analysis of Co-expression Patterns of Histone Variants and Their Chaperones Across Tissues from Publicly Available RNA-Seq Data

Supervisor	Geneviève ALMOUZNI - Tina KARAGYOZOVA
Laboratory	Curie Institute - Nuclear Dynamics Unit UMR3664 – Chromatin Dynamics

2020-2021

Table of Contents

TABLE OF FIGURES III

ABBREVIATIONS: V

1. INTRODUCTION: 2

1.1. HISTONE VARIANTS AND THEIR CHAPERONES: 2

1.1.1. H3-H4 incorporation: 2

1.1.2. H2A-H2B incorporation: 4

1.2. IMPORTANCE: 6

1.3. AIM OF THE PROJECT: 8

2. MATERIALS AND METHODS: 10

2.1. WORKING ENVIRONMENT: 10

2.2. DATASETS: 10

2.2.1. Compile the list of genes 10

2.2.2. Data acquisition: 10

2.2.3. Quality Control: 12

2.3. DIMENSIONALITY REDUCTION: 12

2.3.1. Principal Component Analysis (PCA): 12

2.3.2. Uniform Manifold Approximation and Projection (UMAP): 14

2.4. HIERARCHICAL CLUSTERING AND HEATMAPS: 14

2.5. WEIGHTED GENE CO-EXPRESSION NETWORK ANALYSIS (WGCNA): 14

3. RESULTS: 16

3.1. GTEX DATASET: 16

3.2. TCGA DATASET: 22

4. DISCUSSION: 26

5. CONCLUSION: 30

REFERENCES: 31

Table of Figures

FIGURE 1 MEAN VARIANCE PLOT ON THE GTEX DATASET USING THE LOG2 OF MEAN CPMS FOR EACH GENE ---------- 9

FIGURE 2 PCA ON THE NORMALIZED GTEX DATASET USING ONLY THE LOG2 COUNTS FROM THE HISTONE VARIANTS / CHAPERONES GENES---------- 9

FIGURE 3 UMAP ON THE NORMALIZED GTEX DATASET USING ONLY THE LOG2 COUNTS FROM THE HISTONE VARIANTS/CHAPERONES GENES----------11

FIGURE 4 CLUSTER MAP BASED ON ALL THE SAMPLES IN GTEX BY THE LOG2 COUNTS FROM THE HISTONE VARIANTS / CHAPERONES GENES----------13

FIGURE 5 CLUSTER MAP BASED ON THE PEARSON CORRELATION BETWEEN HISTONE VARIANTS / CHAPERONES GENES FROM THE GTEX DATASET ----------15

FIGURE 6 CO-EXPRESSION NETWORK GENERATED BY THE ITERATIVE WGCNA TOOL USING THE HISTONE VARIANTS / CHAPERONES COUNTS FROM THE GTEX DATASET ----------17

FIGURE 7 MEAN VARIANCE PLOT ON THE TCGA DATASET USING THE LOG2 OF MEAN CPMS FOR EACH GENE ----------17

FIGURE 8 PCA ON THE NORMALIZED TCGA DATASET USING ONLY THE LOG2 COUNTS FROM THE HISTONE VARIANTS / CHAPERONES GENE ----------19

FIGURE 9 UMAP ON THE NORMALIZED TCGA DATASET USING ONLY THE LOG2 COUNTS FROM THE HISTONE VARIANTS/CHAPERONES GENES----------21

FIGURE 10 CLUSTER MAP BASED ON ALL THE SAMPLES IN TCGA BY THE LOG2 COUNTS FROM THE HISTONE VARIANTS / CHAPERONES GENES----------23

FIGURE 11 CLUSTER MAP BASED ON THE PEARSON CORRELATION BETWEEN HISTONE VARIANTS / CHAPERONES GENES FROM THE TCGA DATASET ----------25

FIGURE 12 CO-EXPRESSION NETWORK THAT WAS GENERATED BY THE ITERATIVE WGCNA TOOL USING THE HISTONE VARIANTS / CHAPERONES COUNTS FROM THE TCGA DATASET --------27

Abbreviations:

ANP32E: Acidic Leucine-Rich Nuclear Phosphoprotein 32 family member E
ASF: Anti Slicing Factor
ATRX: α-Thalassemia/Mental Retardation Syndrome X-linked
CABIN: Calcineurin Binding Protein
CAF: Chromatin Assembly Factor
CPM: Counts Per Million
DAXX: Death Domain–Associated Protein 6
DNA: Deoxyribonucleic Acid
DSC: DNA Synthesis-Coupled
DSI: DNA Synthesis-Independent
ERV: Endogenous Retroviruses
FACT: Facilitates Chromatin Transcription
GTEx: Genotype-Tissue Expression
HIRA: Histone Regulator A
HIST: Human Histone Cluster
HJURP: Holiday Junction Recognition Protein
HLB: Histone Locus Body
NASP: Nuclear Autoantigenic Sperm Protein
NCP: Nucleosome Core Particle
PCA: Principal Component Analysis
PCNA: Polymerase Sliding Clamp Proliferating Cell Nuclear Antigen
PTM: Posttranslational Modification
RNA: Ribonucleic Acid
SRCAP: Snf2- Related CREBBP CBP Activator Protein
TCGA: The Cancer Genome Atlas
TSV: Tabula Separated Values
UBN: Ubinuclein
UMAP: Uniform Manifold Approximation and Projection
UTR: Untranslated Regions
WGCNA: Weighted Gene Co-expression Network Analysis

1. Introduction:

The nucleosome is the basic unit of chromatin, a complex of deoxyribonucleic acid (DNA), ribonucleic acid (RNA) and proteins that allows eukaryotic cells to organize their genetic material, and pack it into the nucleus. Each nucleosome comprises a nucleosome core particle (NCP) with 147 bp of DNA sequence wrapped around a core protein particle comprising a $(H3\text{-}H4)_2$ tetramer flanked by 2 x (H2A-H2B) dimers. NCPs are further connected to one another by a linker DNA of variable length (Yadav et al., 2018). Histones exist as distinct variants which display different posttranslational modifications (PTMs) associated with specific chromatin states. Throughout their cellular life, histone variants are escorted by histone chaperones that can be either dedicated or rather permissive to interact with various variants. Histone chaperones control the supply of histones, including their deposition and/or eviction from chromatin (reviewed in Filipescu et al., 2014).

1.1. Histone variants and their chaperones:

Histone variants, except for the H4 family, can be divided into two categories: replicative and replacement. Their gene organization and transcriptional regulation is distinct and they are engaged in distinct deposition pathways. Replicative histones, as indicated their name, show a peak of expression are in S phase and are deposited in a DNA synthesis-coupled (DSC) manner mainly during replication (Tagami et al., 2004). The expression of replicative histones relies on a shared transcriptional regulation mechanism. Encoded by intronless genes with short untranslated regions (UTRs) they harbor a conserved 3' stem-loop structure instead of a polyadenylation signal. Furthermore, their organization in clusters that contain several copies of each family contributes to their co-regulation. In humans, they are distributed in three clusters: human histone cluster 1 (HIST1) is found on chromosome 6 (6p22) and has more than 55 genes equivalent to approximately 55% of the genes, HIST2 is located on chromosome 1 (1q21) and has 10 genes and finally HIST3 which is on the same chromosome (1q42) has 3 genes. One important aspect of these clusters is the fact that they are grouped together in the nuclear space, forming histone locus bodies (HLBs), where the factors required for replicative histone gene transcription and processing are concentrated. These unique features allow the coordinated transcription of replicative histones to provide vast amounts needed during S phase, to ensure the re-assembly in chromatin of the duplicated genome (Mendiratta et al., 2019).

Conversely, the replacement variants are expressed independently of S phase and deposited in a DNA synthesis-independent (DSI) manner at any time during different cell cycle phases (Filipescu et al., 2014). Replacement variants are encoded by individual genes which exhibit a typical gene structure, containing introns and polyadenylation signals. Notably, the expression and deposition of replacement variant can be highly regulated with respect to the cell cycle. For example, the centromeric H3 variant CENP- A is specifically expressed in late G2/M and deposited in early G1 by its chaperone holiday junction recognition protein (HJURP) (Dunleavy et al., 2009, Foltz et al., 2009).

1.1.1. H3-H4 incorporation:

In humans, to date, there are eight known H3 variants. Of which two are replicative variants: H3.1 and H3.2 and differ from the replacement variant H3.3 only by five and four amino acids, respectively. The other replacement variants include the centromeric variant $CenH3^{CENP\text{-}A}$, the testis-specific variants H3.4 and H3.5, and finally the H3.Y.1 and H3.Y.2 variants (reviewed in Filipescu et al., 2014).

The replicative H3 variants as dimers with H4 interact specifically with chromatin assembly factor 1 (CAF-1), a complex made of three subunits: p150, p60 and p48. CAF-1 promotes the deposition of replicative histones in a DSC manner during replication and DNA repair. Phosphorylation events promote CAF-1 recruitment to replication/repair sites via the interaction with the DNA polymerase sliding clamp proliferating cell nuclear antigen (PCNA) (Moggs et al., 2000). CAF-1 is important for the maintenance of cell fate by propagating chromatin states including H3K9me3 pericentromeric chromatin (Ishiuchi et al., 2015, Cheloufi et al., 2015).

The replacement variant, H3.3, is encoded by two paralogous genes (H3F3A and H3F3B), giving an identical protein product. H3.3 can be deposited at different genomic locations where it is associated with distinct chromatin states. Histone regulator A (HIRA), a complex composed of HIRA, calcineurin binding protein 1 (CABIN1), ubinuclein1 (UBN1) or ubinuclein2 (UBN2), is responsible for depositing H3.3 at gene bodies, promoters, and regulatory elements (Goldberg et al., 2010). This includes both de novo deposition, as well as recycling of H3.3, which can contribute to maintaining the chromatin state of actively transcribed regions (Torne et al., 2020). In addition, HIRA has a higher affinity for naked DNA *in vitro* and this has led to propose a 'gap-filling' mechanism for H3.3 deposition, important for maintaining chromatin integrity (Ray-Gallet et al., 2011). H3.3 accumulation at telomeres and pericentric regions is however dependent on the death domain–associated protein 6 (DAXX)/α-thalassemia/mental retardation syndrome X-linked (ATRX) complex (Goldberg et al., 2010, Banaszynski et al., 2013). Finally, H3.3 accumulation at repetitive regions containing endogenous retroviruses (ERV) in mouse embryonic stem cells is key for their silencing which is important during early development (Elsasser et al., 2015).

The centromeric variant, $CenH3^{CENP-A}$, is encoded by a single multi-exon gene (CENPA). $CenH3^{CENP-A}$ is expressed in G2/M phases of the cell cycle but deposited on chromatin in G1 by its chaperone HJURP (Sullivan et al., 1994, Régnier 2003). As the epigenetic mark that defines the centromere, a tight control of $CenH3^{CENP-A}$ localization is crucial for maintaining genome integrity (Lacoste et al., 2014). Indeed, $CenH3^{CENP-A}$ and HJURP are co-regulated, as the exogenous over-expression of either the variant or the chaperone leads to an increase in the levels of the other partner (Heo et al., 2013, Shrestha et al., 2017, in Filipescu et al., 2014)

Apart from the H3 variant-specific chaperones discussed above, general H3-H4 chaperones also exist. Anti-slicing factor 1 (ASF1) a chaperone encoded by two paralog genes (ASF1A and ASF1B) is responsible for handling H3 variants and passing them to their respective variant-specific chaperone. The two paralogs are not functionally equivalent and have different expression patterns. Of note, ASF1a prefers interacting with HIRA and ASF1b selectively interacts with CAF-1 (Abascal et al., 2013). Finally, nuclear autoantigenic sperm protein (NASP) is a general chaperone that has two forms: somatic, which is ubiquitously expressed, and testicular, which is specific to testes and ovaries. NASP fine-tunes the levels of soluble H3-H4 by protecting them from histone-mediated autophagy. This function is critical during replication as it can ensure an appropriate supply of histones is available when there is an acute demand for them (Cook et al., 2011).

1.1.2. H2A-H2B incorporation:

In human, in addition to the replicative H2A, there are four types of replacement H2A variants, which are H2A.X, H2A.Z, macroH2A and testis-specific variants (H2A.B, H2A.L, H2A.P and H2A.Q).

H2A.Z is encoded by two paralogs (H2AFZ & H2AFV) giving rise to two isoforms (H2A.Z.1 and H2A.Z.2) which differ by three amino acids. Furthermore, H2AFV can be alternatively spliced resulting in two isoforms H2A.Z.2.1 and H2A.Z.2.2 (Matsuda et al., 2010). H2A.Z is deposited by two ATP-dependent chromatin remodelers, p400 and Snf2-related CREBBP activator protein (SRCAP). Its eviction however is mediated by the INO80 remodeler and acidic leucine-rich nuclear phosphoprotein 32 family member E (ANP32E) (reviewed in Martire and Banaszynski, 2020). H2A.Z is involved in several biological processes, including the regulation of gene expression. It is enriched at promoters and other regulatory elements such as enhancers, where acetylated H2A.Z correlate with high levels of gene expression and vice versa (Valdés-Mora et al., 2012).

MacroH2A is also encoded by two paralogs (H2AFY & H2AFY2), resulting in two isoforms macroH2A.1 and macroH2A.2. Like to H2AFV, H2AFY can be alternatively spliced giving rise to macroH2A.1.1 and macroH2.1.2. MacroH2A had been proposed to participate in the transcriptional repression of the inactive X chromosome and inactive genes where it accumulates. To date, no chaperone has been identified to be specific for macroH2A (reviewed in Ghiraldini et al., 2021).

Finally, H2A.X is encoded by H2AFX, a gene that combines characteristics of both the replicative and the replacement histone genes. H2AFX is intronless and could be transcribed in two alternative ways; either with a stem-loop or with a poly(A) tail (Mannironi et al., 1989). H2A.X has been associated with the DNA damage response, as its phosphorylation at Ser139 amplifies the DNA damage signal. The phosphorylated form γH2A.X, is thus a marker of damaged chromatin. Like macroH2A, no dedicated chaperone has been found for H2A.X. However, H2A.X can be deposited at repair sites by the general chaperone facilitates chromatin transcription (FACT) (Piquet et al. 2018). FACT is a complex made up of two subunits, SSRP1 and SPT16, which can mediate the deposition of H2A-H2B dimers (Orphanides et al., 1999). However, it remains unclear how this can contribute to obtain the distinct occupancy patterns of H2A variants along the genome, since general chaperones do not discriminate between the different variants. Interestingly, FACT initially identified as a H2A-H2B chaperone, can also serve as a chaperone for H3-H4 dimers (reviewed in Martire and Banaszynski, 2020), thereby broadening its promiscuous function.

1.2. Importance:

Appropriate expression and incorporation of histone variants throughout the cell cycle is critical for normal cell function and organismal development. Indeed, depletion of various histone variants and chaperones leads to embryonic lethality or sterility in mice and other model organisms. For instance, developmental arrest occurs in vertebrate model organisms at the morula stage upon the loss of H3.3 variant. In particular, due to its DSI mode of incorporation, H3.3 plays an important role upon fertilization, as it is the variant used to reconstitute paternal chromatin after protamine removal. Furthermore, its K27 residue is required for the appropriate establishment of paternal pericentric heterochromatin and further developmental progression (reviewed in Filipescu et al., 2014).

Mutations, as well as misexpression of histone variants or their chaperones could lead to changes in genome stability and chromatin organization and has been linked to disease development and progression (reviewed in Ghiraldini et al., 2021, Amatori et al., 2021, Ferrand et al., 2020). For example, CENP-A is overexpressed in several cancer types, and this has been associated with patient prognosis and tumor metastasis (Sun et al., 2016). Moreover, mutation of histone variants

in tumors can alter chromatin state leading to gene dysregulation (Bagert et al., 2021). For example, the famous H3K27M mutation commonly found in pediatric glioblastoma leads to increased levels of H3K27ac and reduced levels H3K27me3 and H3K27me2 leading to DNA hypomethylation and consequently to a loss of gene silencing (reviewed in Amatori et al., 2021).

1.3. Aim of the project:

Biochemical studies have allowed the characterization of several interactions between histone variants and their chaperones. However, their regulation across and within different tissues is still not fully characterized neither in healthy nor in disease context.

The objective of this project was to perform a co-expression analysis using publicly accessible transcriptomics data to investigate how histone chaperones and histone variants are expressed across and within tissues and how these expression patterns would change in the context of distinct cancers.

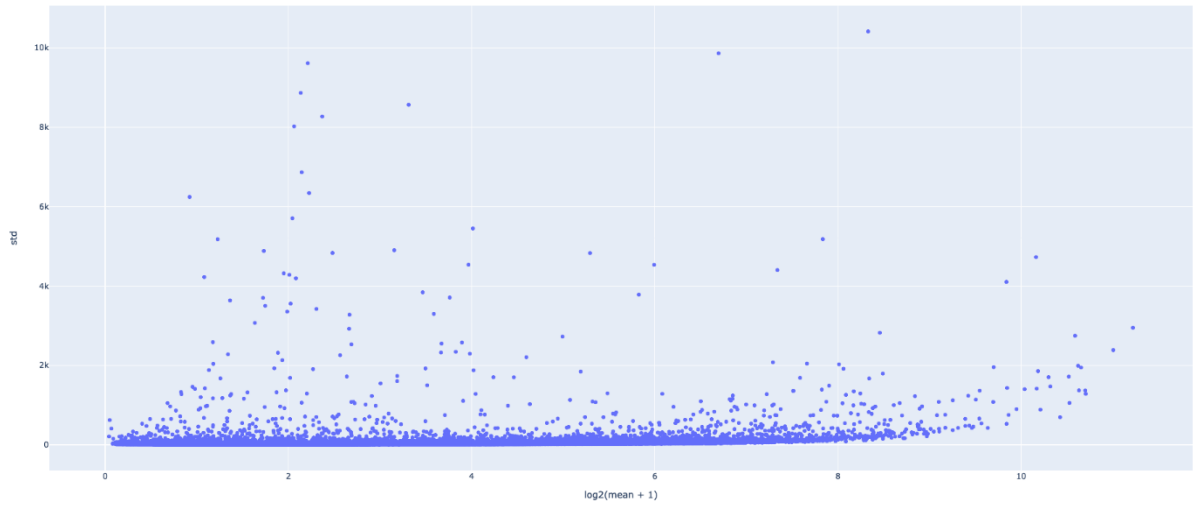


Figure 1 Mean Variance Plot on the GTEx Dataset Using the log2 of Mean CPMs for Each Gene
Full size image available at: https://github.com/Ala-Eddine-BOUDEMIA/Chromatin-Dynamics/blob/main/Images/GTEx/CPM/MV_Plots/Normalized/Full/mv.png
Interactive plot available at: https://github.com/Ala-Eddine-BOUDEMIA/Chromatin-Dynamics/blob/main/Plotly_HTML_Files/GTEx/CPM/MV_Plots/Normalized/Full/mv.html

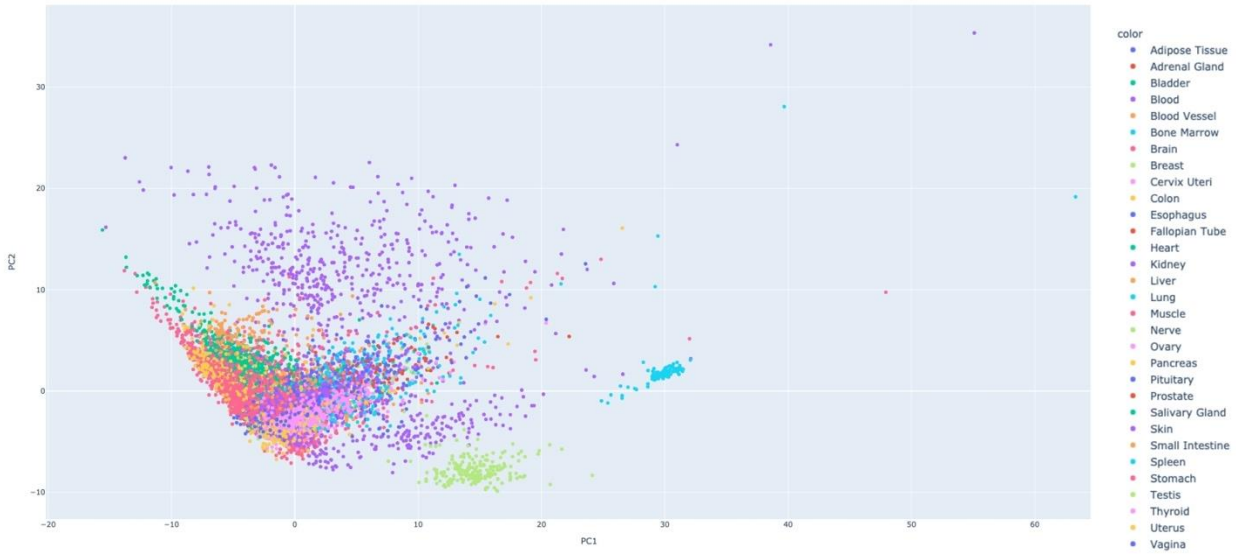


Figure 2 PCA on the Normalized GTEx Dataset Using Only the log2 Counts from the Histone Variants / Chaperones Genes
On the x-axis figures the first principal component and, on the y-axis, figures the second principal component.
Each point represents a sample, and it is colored by the tissue type it corresponds to.
Full size image with detailed sub-tissues available at:
https://github.com/Ala-Eddine-BOUDEMIA/Chromatin-Dynamics/blob/main/Images/GTEx/CPM/PCA/variants_chaperones/Full/pca.png
Interactive plot available at: https://github.com/Ala-Eddine-BOUDEMIA/Chromatin-Dynamics/blob/main/Plotly_HTML_Files/GTEx/CPM/PCA/variants_chaperones/Full/pca.html

2. Materials and Methods:

2.1. Working environment:

This analysis was conducted on an iMAC running on MacOS Catalina version 10.15.7 Equipped with a third-generation core i5 intel processor, with 1 Tb of storage and 8 Gb of RAM extendable virtually using the hard drive.

All the scripts used to perform this analysis are available on the GitHub that I established in the repository named Chromatin-Dynamics (https://github.com/Ala-Eddine-BOUDEMIA/Chromatin-Dynamics). Most of these scripts are written in python3 programming language using both Visual Studio Code (code editor) and Sublime Text (text editor).

The required packages necessary for this analysis are listed in the "requirements.txt" file in the GitHub repository.

To install these packages, type the following command in the folder that contains the "requirements.txt" file: **sudo pip3 install -r requirements.txt**

Few scripts are written in R programming language using R-Studio. The main R library used for this analysis is the recount library (Leonardo Collado-Torres et al., 2017).

2.2. Datasets:

2.2.1. Compile the list of genes

A 'comprehensive' list of genes that code for histone variants, histone chaperones and chromatin remodelers thought to have histone variant-specific activity was compiled from the literature (Filipescu et al., 2014, Martire and Banaszynski, 2020).

2.2.2. Data acquisition:

In order to investigate the co-expression patterns of histone variants and histone chaperones genes across and within healthy tissues, RNA-seq data from the Genotype-Tissue Expression (GTEx) project was used. GTEx v6 dataset comprises 9662 samples distributed across 53 tissues with 58037 genes detected. Sequencing libraries were built using the Illumina TruSeq library construction protocol (non-stranded, polyA+ selection), and have a median coverage of about 82M reads per samples (Lonsdale et al., 2013).

To investigate patterns across and within cancer contexts, data from The Cancer Genome Atlas (TCGA) was used. The TCGA dataset contains 11284 samples from 33 different cancer types. The sequencing libraries were established by either poly(A) selection or ribosomal RNA depletion (National Cancer Institute).

It is important to note that constructing the sequencing libraries using poly(A) selection would lead to a loss of the majority of the replicative variants' transcripts as they do not exhibit a poly(A) tail.

The RNA-seq data described above were obtained from the recount2 project (Leonardo Collado-Torres et al., 2017). Recount2 provides gene expression data in the form of coverage per base counts matrices that have been identically re-processed.

Both GTEx and TCGA RNA-seq datasets were downloaded from the recount2 website as *Rdata* files. Using the recount package, the coverage counts were transformed into read counts using the "read_counts" method and were then saved as *tsv* files. Metadata files for both datasets were acquired as *tsv* files from the recount2 website.

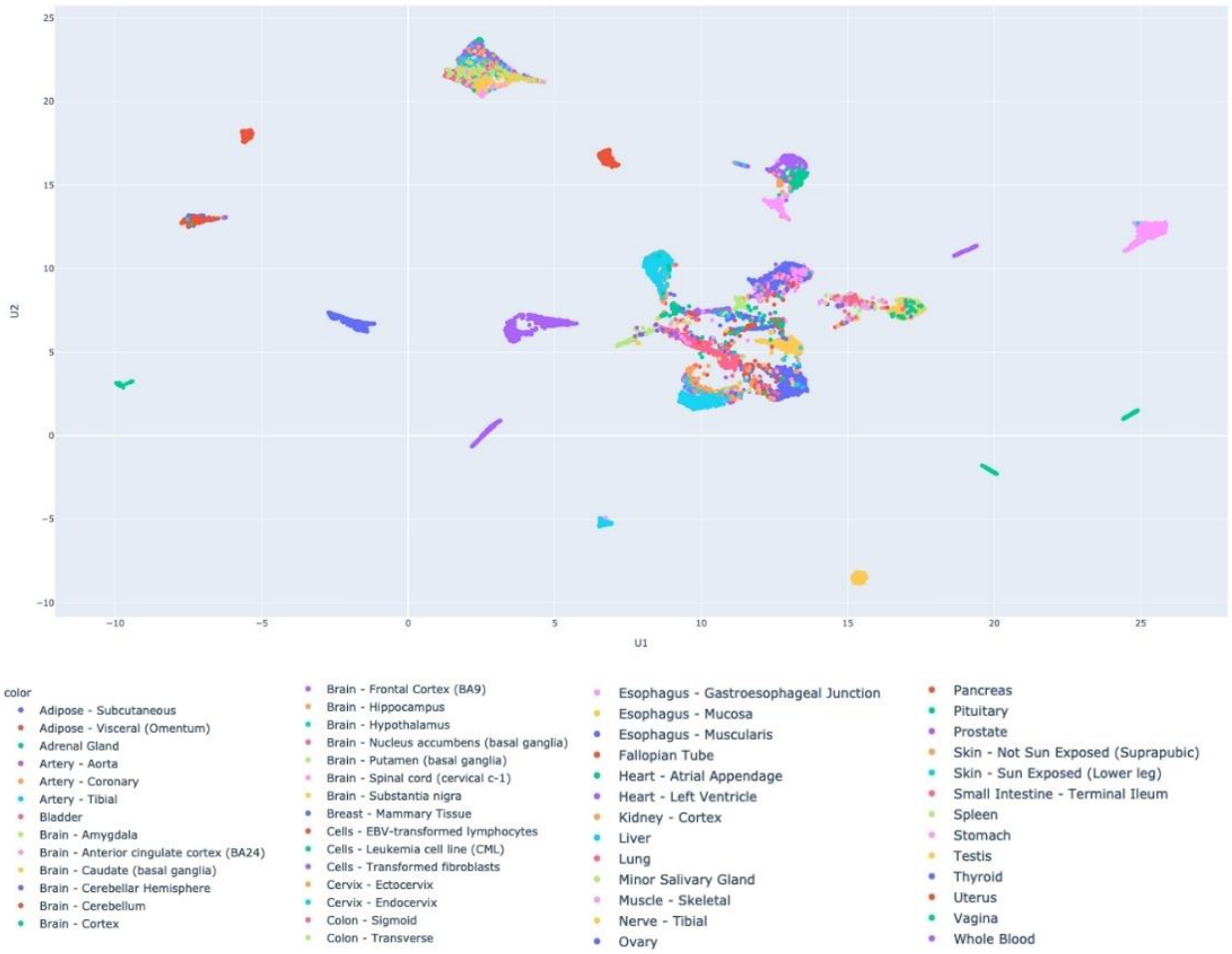


Figure 3 UMAP on the Normalized GTEx Dataset Using Only the log2 Counts from the Histone Variants/Chaperones Genes
On the x-axis figures the first component, and on the y-axis figures the second component.
Each point represents a sample, and it is colored by the sub-tissue it corresponds to.
Full size image available at: https://github.com/Ala-Eddine-BOUDEMIA/Chromatin-Dynamics/blob/main/Images/GTEx/CPM/UMAP/variants_chaperones/Full/umap.png
Interactive plot available at: https://github.com/Ala-Eddine-BOUDEMIA/Chromatin-Dynamics/blob/main/Plotly_HTML_Files/GTEx/CPM/UMAP/variants_chaperones/Full/umap.html

2.2.3. Quality Control:

2.2.3.1. Data Normalization:

To make gene expression levels comparable between samples, raw reads counts were normalized for sequencing depth. Using Pandas package in python3, the count tables previously saved as *.tsv* files were read and scaled by the library size of each sample resulting in counts per million (CPM).

This normalization is applied in two steps:

1. Calculate the scaling factor for each sample, by computing the sum of all the reads in a sample and divide it by 10^6.
2. Divide each count in a sample by the scaling factor corresponding to that sample.

2.2.3.2. Data Filtering:

This step was performed by the same script used for normalizing the data to omit genes with no detectable or very low expression from the analysis. Thus, genes close to the limit of detectability which are highly variable would be excluded from the analysis. After transforming raw counts to CPMs, all genes with an expression level mean below 1 CPM or over 4000 CPM were removed.

TCGA dataset contains different types of samples such as normal tissue, recurrent tumor or metastatic. For the sake of studying the expression patterns of histone variants and their chaperones in cancer context, the samples from the normal tissues were filtered. The other types of samples (recurrent tumour, 50 samples; metastasis, 394 samples, primary blood-derived cancer, 126 samples) were filtered as well to keep this study homogenous as these different types of tumors could exhibit different expression patterns. Therefore, keeping only the primary tumors samples reduced the dataset into 9662 samples.

2.2.3.3. Mean-Variance Trend:

In RNA-seq data, there is typically a positive correlation between the mean and the variance per gene, hence the trends detected in the downstream analysis may be dependent on a small set of highly expressed, highly variable genes. To account for this, the logarithm base 2 of CPMs plus a pseudo-count of 1 (to avoid providing 0 values) was used (Love., 2019).
Using Numpy package, the normalized counts were transformed to log2(CPM + 1).
Using Pandas package, the mean and the standard deviation were computed row wise.
Using Plotly package, an interactive scatter plot of the standard deviation by the mean was generated.

2.3. Dimensionality reduction:

Due to the high dimensionality of RNA-seq data, dimensionality reduction methods are used to help visualize and evaluate the similarity between samples based on their gene expression patterns. This enables important patterns in the data to emerge, whereas noise and / or unnecessary details are often removed. As these methods work best when the variance is independent on the mean, log2(CPM + 1) was used instead of the normalized counts.

2.3.1. Principal Component Analysis (PCA):

PCA is an unsupervised linear dimensionality reduction algorithm that aims to explain the relationship between the majority of the data while preserving its overall structure. It projects the data into smaller subspaces known as principal components (PCs) and looks for the PCs that explain the most variance in the dataset. Each PC must be independent of the others, hence there will be as many PCs as there are samples or genes, whichever is less (Lever et al., 2017).
Only the first two PCs were computed using PCA method in the decomposition module from the Scikit-learn package.

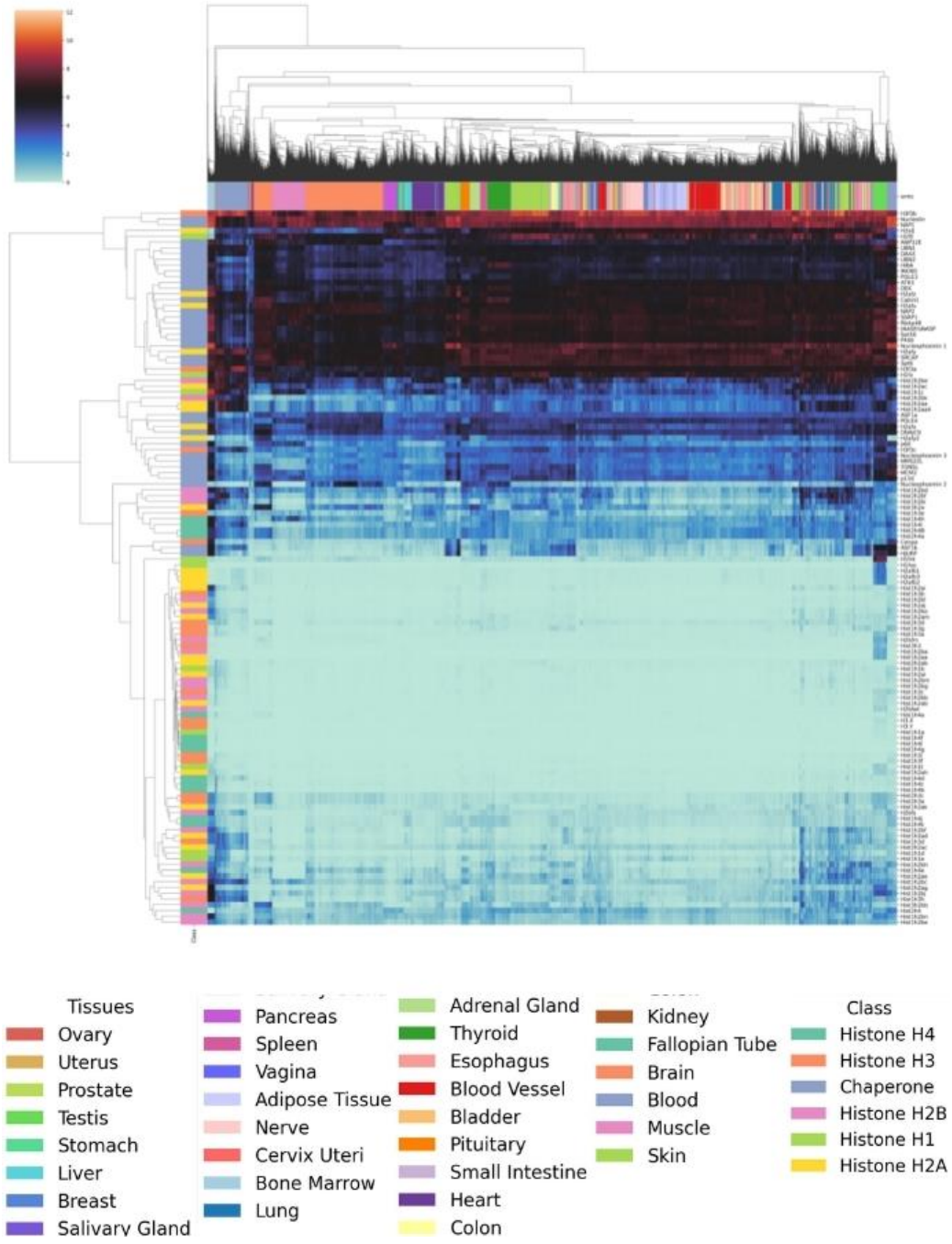


Figure 4 Cluster map based on All the Samples in GTEx by the log2 Counts from the Histone Variants / Chaperones Genes
The horizontal line is represented by all the samples in the GTEx dataset, and each cluster is colored by tissue type.
The vertical line is represented by all the histone variants / chaperones genes, and each cluster is colored by histone variant / chaperones class
Full size image available at: https://github.com/Ala-Eddine-BOUDEMIA/Chromatin-Dynamics/blob/main/Images/GTEx/CPM/Clustermap/Samples_Genes/euclidean/variants_chaperones/Full/clustermap.png

2.3.2. Uniform Manifold Approximation and Projection (UMAP):

UMAP is a nonlinear dimensionality reduction technique that preserves primarily the local structure of the data, but it does a much better job of preserving the global structure compared with other methods such as T-SNE. UMAP is especially useful for locating clusters in the dataset and observing intra-heterogeneity within each cluster. It accomplishes this by approximating the manifold on which the data lies and focuses on similarity rather than variance (McInnes et al., 2018).

UMAP was implemented using the umap package that was developed by the writers of the method themselves. The random seed was also fixed to 0 to ensure reproducibility.

2.4. Hierarchical Clustering and Heatmaps:

To investigate the similarities between gene expression patterns of different samples, hierarchical clustering is commonly used in combination with heatmaps to generate cluster maps. Hierarchical clustering is a machine learning algorithm that groups data points based on their distances, which allows to organize samples by similarity, whereas heatmaps are used to depict gene expression patterns across samples. Thus, expression patterns distinct for particular sample groups become visible.

Cluster maps were generated using the cluster map method from the Seaborn package. The distance metric was set to Euclidean which assumes a normal distribution of the data therefore the log2(CPM + 1) were given to the method. The dendrograms were arranged using the average method which was observed to yield better results after testing the most common methods (complete, weighted, single, ward), in agreement with (Jaskowiak et al., 2018).

2.5. Weighted Gene Co-expression Network Analysis (WGCNA):

To describe the transcriptional relationship between genes and identify gene modules likely to be co-regulated, co-expression networks can be built from RNA-seq counts. A frequently used approach to achieve this is weighted gene co-expression network analysis (WGCNA), which works as follows. First, highly correlated pairs of genes which have a similar expression pattern across samples are identified. The correlation matrix between genes is then used to construct a graph in which each gene is represented by a node and the co-expressed genes are linked by edges. Finally, modules are identified using hierarchical clustering; each module represents a group of genes with similar expression patterns (Van Dam et al., 2018).

The iterative WGCNA tool is a Python extension for the R WGCNA library that improves the robustness of RNA-seq derived networks (Greenfest-Allen et al., 2017). Iterative WGCNA was used to compute the Topology Overlap Matrices (TOMs). The TOMs were later transformed to adjacency matrices using the Igraph library in R. The adjacency matrix was saved as a tsv file, and interactive networks were created using the Networkx package in conjunction with the Pyvis package. The genes were colored differently depending on which module they belong to using the membership list provided by the iterative WGCNA program.

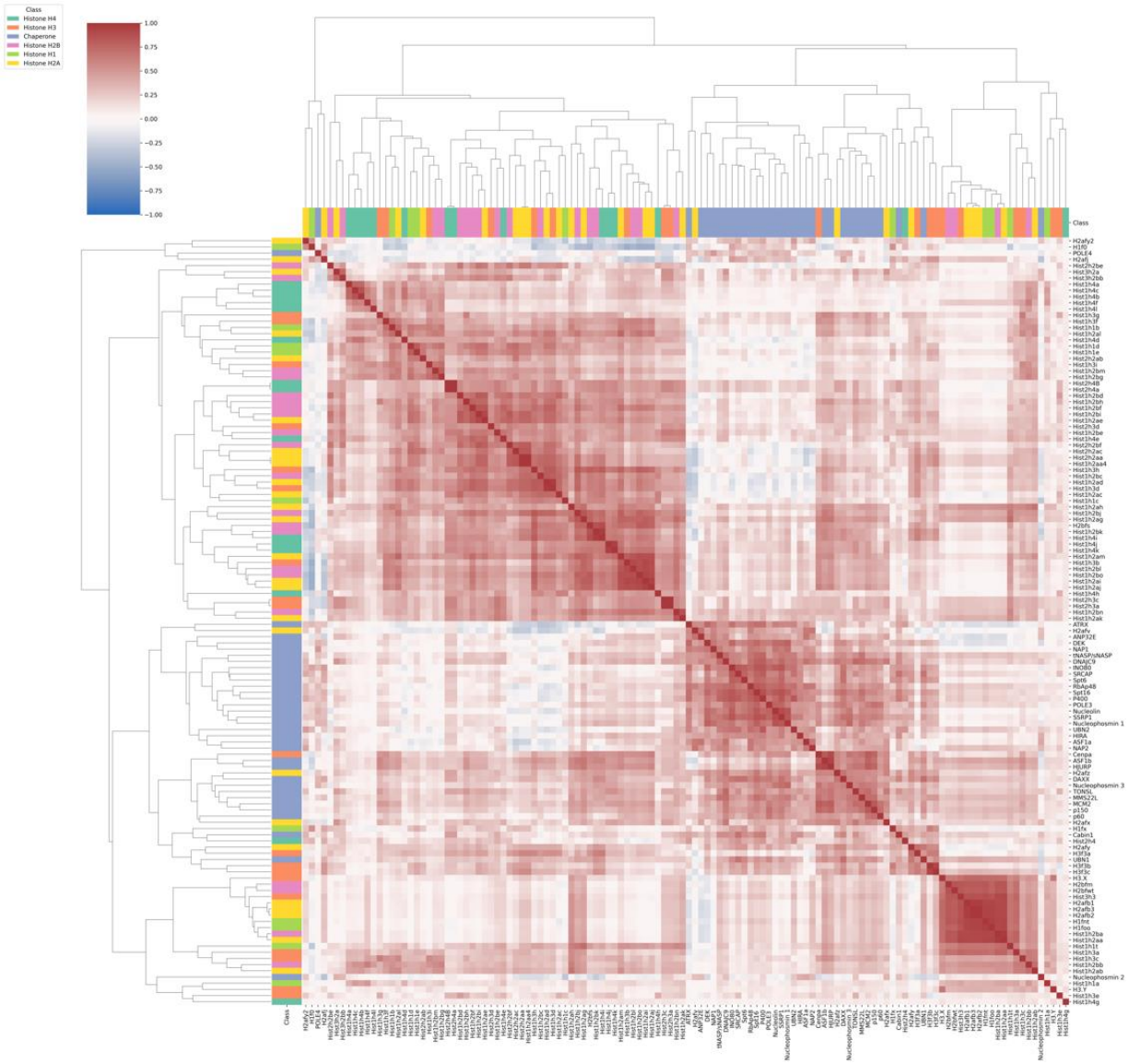


Figure 5 Cluster map Based on the Pearson Correlation Between Histone Variants / Chaperones Genes from the GTEx Dataset

Both the horizontal and the vertical lines are represented by all the histone variants / chaperones genes, and each cluster is colored by histone family or as chaperone

Full size image available at: https://github.com/Ala-Eddine-BOUDEMIA/Chromatin-Dynamics/blob/main/Images/GTEx/CPM/Clustermap/Genes/euclidean/variants_chaperones/Full/clustermap.png

3. Results:

To investigate the co-expression patterns of histone variants and chaperones across healthy tissues and across cancer types, a list of 125 genes of interest was compiled and the expression levels of the respective genes from the GTEx and TCGA were analyzed.

3.1. GTEx dataset:

To ensure the filtering and normalization of the data was effective, the mean-variance trend of the log2(CPM +1) was plotted (Figure 1), which demonstrates that the standard deviation of the CPMs is independent of their mean. It also shows that the data contains some highly variable genes that are mainly mitochondrial.

In order to evaluate if histone gene and chaperone expression can be used to discriminate between tissue types, PCA based on the counts from the histone variants and chaperones gene list was performed (Figure 2). The scatter plot shows that the bone marrow (leukemia cell line), the testis, and the blood can be distinguished based on histone variants / chaperones genes, whereas the rest of the tissues cannot. As a control, PCA based on the top 1000 expressed genes was performed (Supplementary Figure 1). This demonstrated that besides the aforementioned tissues, the pancreas and the liver also separated from the rest, as did the brain and the heart which were overlapped with each other. Overall, PCA does not work well to separate tissue types from the GTEx data. Nevertheless, using only the histone chaperone/variant gene set can separate some tissue types by PCA, although it performs worse than using the top 1000 expressed genes.

The poor separation of tissues by PCA could be due to the fact that it is a linear method, and it is challenging for such method to separate a huge number of points coming from different sources (tissues). Similar issues have been encountered in the analysis of single-cell RNA-seq data, where gene expression for thousands of cells is detected. To address them, non-linear dimensionality methods (e.g., t-SNE, UMAP) have been implemented and shown to perform better.

In the context of this work, individual bulk RNA-seq samples can be thought of as separate cells from a single-cell RNA-seq experiment, so to separate samples by tissue type, UMAP was used. It shows that the histone variants and histone chaperone genes can separate the tissues into different clusters, although some samples do not cluster within their respective tissues (Figure 3). The UMAP also allows to see the samples intra-heterogeneity within each group. For instance, the adipose tissue could be separated into subcutaneous and visceral whereas the brain's sub-tissues cannot be separated, which was confirmed by applying PCA within tissues. As for PCA, a control with the top 1000 expressed genes was performed (Supplementary Figure 2), which yielded similar results to using only histone chaperone/variant genes but with better separation of the clusters. Globally, UMAP separated the samples by tissue type clearly regardless of the gene set used.

This analysis demonstrates that distinct tissue types can be separated based on the expression patterns of the histone variant and chaperone genes. To study how these patterns vary across and within tissues, cluster maps of gene expression by sample were generated for the gene list (Figure 4). To investigate how the genes themselves are co-expressed across tissues, cluster maps based on the Pearson correlation between their expression were also generated (Figure 5).

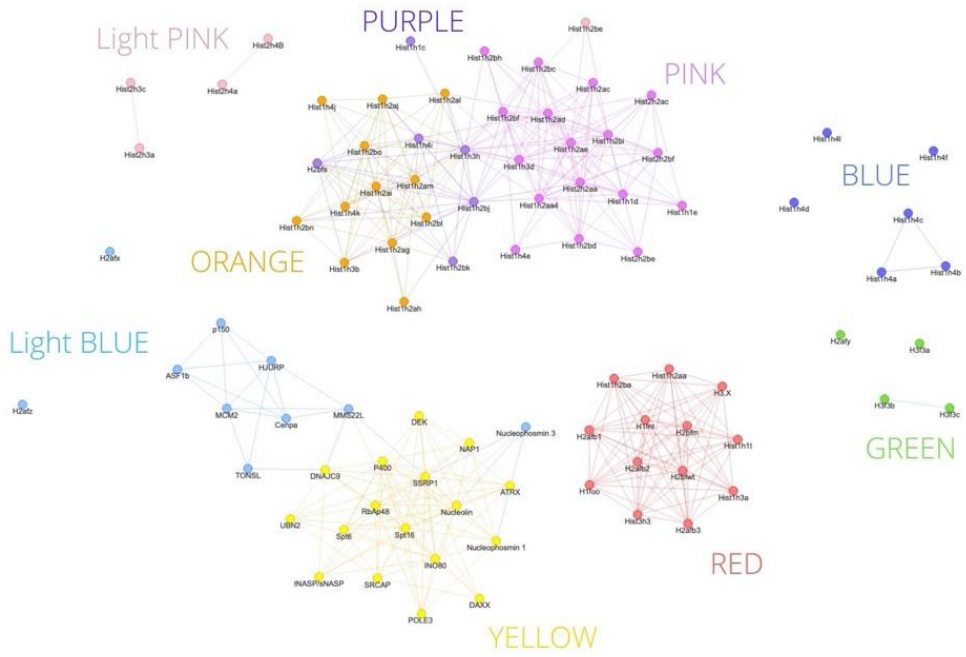


Figure 6 Co-expression Network generated by the iterative WGCNA Tool Using the Histone Variants / Chaperones Counts from the GTEx Dataset
Each gene is represented by a node and each edge links two co-expressed genes.
Each set of genes that constitute a module are colored by the same color.
Download interactive network from: https://github.com/Ala-Eddine-BOUDEMIA/Chromatin-Dynamics/blob/main/WGCNA/GTEx/Networks/graph_gtex.html

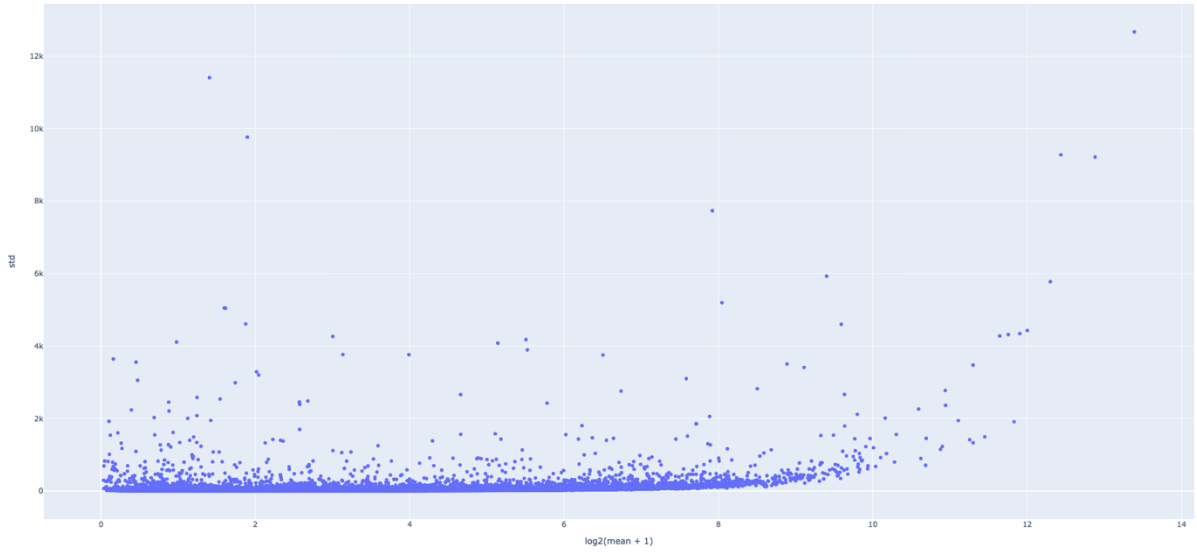


Figure 7 Mean Variance Plot on the TCGA Dataset Using the log2 of Mean CPMs for Each Gene
Full size image available at: https://github.com/Ala-Eddine-BOUDEMIA/Chromatin-Dynamics/blob/main/Images/TCGA_PT/CPM/MV_Plots/Normalized/Full/mv.png
Interactive Plot available at: https://github.com/Ala-Eddine-BOUDEMIA/Chromatin-Dynamics/blob/main/Plotly_HTML_Files/TCGA_PT/CPM/MV_Plots/Normalized/Full/mv.html

In agreement with the UMAP, the cluster map shows that the expression patterns of histone chaperone and variant genes across tissues can distinguish individual tissues as the samples cluster mainly by tissue type (Figure 4). Additionally, the cluster map demonstrates how the genes in this set are co-expressed across tissues.

The majority of the replicative histones are detected as having very low expression and cluster together as part of a larger cluster of histone variant/chaperone genes with low expression. These include testis-specific variants, as well as CENP-A, HJURP and ASF1B, which form their own subgroup among lowly expressed genes. However, several replicative histone genes (Hist2h2be, Hist1h2ac, Hist1h1c, Hist1h2bk, Hist2h2aa, Hist1h2aa4) with higher expression levels are also detected. They cluster separately from the rest of the replicative histones, but together with members of the replisome (MCM2, POLE4) and the replicative H3 chaperone, CAF-1 (p60, p150).

On the other hand, the replacement variants and their chaperones also cluster together. Notably, HIRA's subunits (HIRA, UBN1, UBN2) are forming one cluster with DAXX, but separately from CABIN-1 that shows higher levels of expression. However, H3F3A and H3F3B are not clustering directly together as H3F3B is more highly expressed than H3F3A and forms a subgroup with Nucleolin and NAP1, which appear ubiquitously highly expressed across all tissues.

The results of this heatmap demonstrate that different subunits of the same complex and the respective histone variants they handle can be expressed independently at different levels across tissues. This indicates they are unlikely co-regulated at the transcriptional level, with the exception of replicative histone genes and the CENP-A/HJURP pair.

Notably, the clustering of genes when using their counts as input will be strongly influenced by their baseline expression levels. Hence, to investigate how similar histone variant and chaperone genes behave in terms of expression patterns, clustering of genes based on the Pearson correlation between them was performed (Figure 5). In accordance with the previous results, two main clusters of genes can be identified based on the dendrograms. The majority of the replicative variants (except Hist1h1a, Hist1h2aa and Hist1h2ba), together with H2bfs comprise one highly correlated group, whereas the second major cluster is composed mainly of replacement variants and histone chaperones, whose expression patterns are also highly correlated between each other. Conversely, the correlation between genes from the two clusters seems to vary from positively to negatively moderate.

In the replacement histone group, several sub-clusters can be observed. CENP-A, HJURP and ASF1B are forming one cluster directly related to the cluster containing CAF-1 subunits (p60, p150), with mainly replication-associated histone chaperone genes. Conversely, the CABIN1 and UBN1 members of the HIRA complex also cluster together with the replacement H3 variants. Surprisingly, the HIRA and UBN2 members clusters most closely with CENPA and HJURP. Finally, testis specific variants cluster together forming a highly correlated group.

Overall, this shows that expression patterns of replacement histone H3 variants are similar despite their differences in gene expression levels. It also indicates that some histone variants do not necessarily show a co-transcription with their respective chaperones, which also applies to subunits of the same chaperone complex.

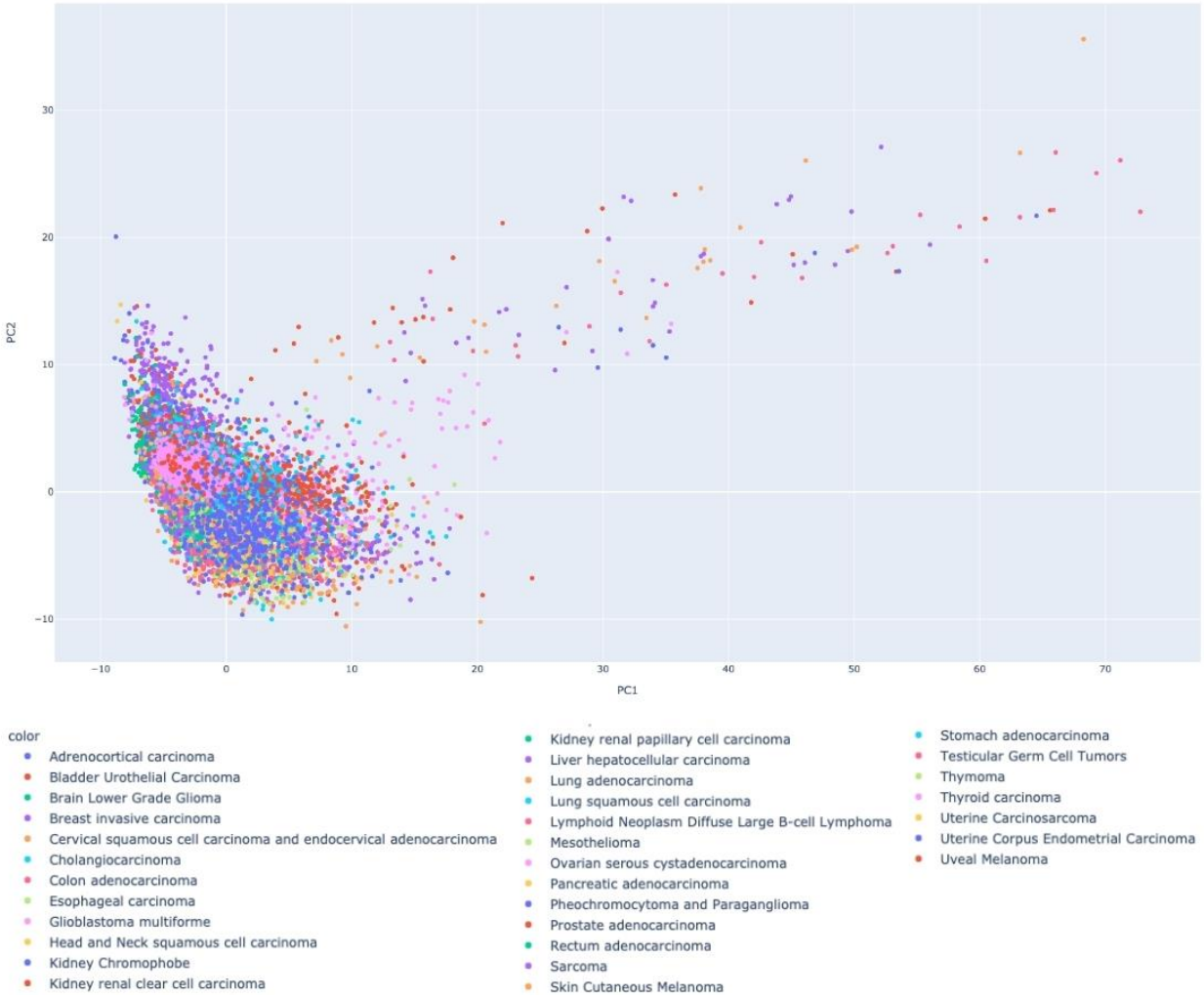


Figure 8 PCA on the Normalized TCGA Dataset Using Only the log2 Counts from the Histone Variants / Chaperones Gene
On the x-axis figures the first principal component and, on the y-axis, figures the second principal component.
Each point represents a sample, and it is colored by the tissue type it corresponds to.
Full size Image available at: https://github.com/Ala-Eddine-BOUDEMIA/Chromatin-Dynamics/blob/main/Images/TCGA_PT/CPM/PCA/variants_chaperones/Full/pca.png
Interactive plot available at: https://github.com/Ala-Eddine-BOUDEMIA/Chromatin-Dynamics/blob/main/Plotly_HTML_Files/TCGA_PT/CPM/PCA/variants_chaperones/Full/pca.html

To visualize the co-transcription relationships between the histone variant and chaperone genes more clearly and investigate if the groups identified from the hierarchical clustering can be presented into co-transcribed modules, co-transcription networks were built using iterative WGCNA.

Figure 06 shows the 91 genes that iterative WGCNA was able to fit into the network, forming 9 modules. These genes include the majority (51 out of 61) of the replicative histone genes and more than half of the replacement and chaperone genes.

Five out of the 9 modules are composed exclusively out of different combinations of replicative histone variants. For two of those (blue and light pink), the genes within them are part of the same histone locus (Hist1 and Hist2, respectively), but comprise different core histone genes, whereas the other 3 (purple, orange and pink) contain mainly H2A and H2B variants from different histone gene clusters. The 6th module (red) contains mostly testis specific variants such as H3.X and H1fnt.

The remaining 3 modules comprise the rest of the replacement histone genes and histone chaperones. The 7th module (green) contains the replacement variants H3F3B and H3F3C (co-expressed), as well as H3f3a and H2AFY. The 8th module (light blue) is composed of 10 genes, of which 5 are replicative histone chaperones, accompanied by the H3 variant CENP-A, its chaperone HJURP and finally, the H2AFX and H2AZ variants which are not connected to the rest. This module is connected to the 9th (yellow) module (composed of highly interconnected 17 chaperone genes) through TONSL/MMS22L-DNAJC9 or through MMS22L-SSRP1. It should be noted that Nucleophosmin3, despite being part of the light blue module it is in fact connected to the genes in the yellow module and only related to its module through its connection with SSRP1.

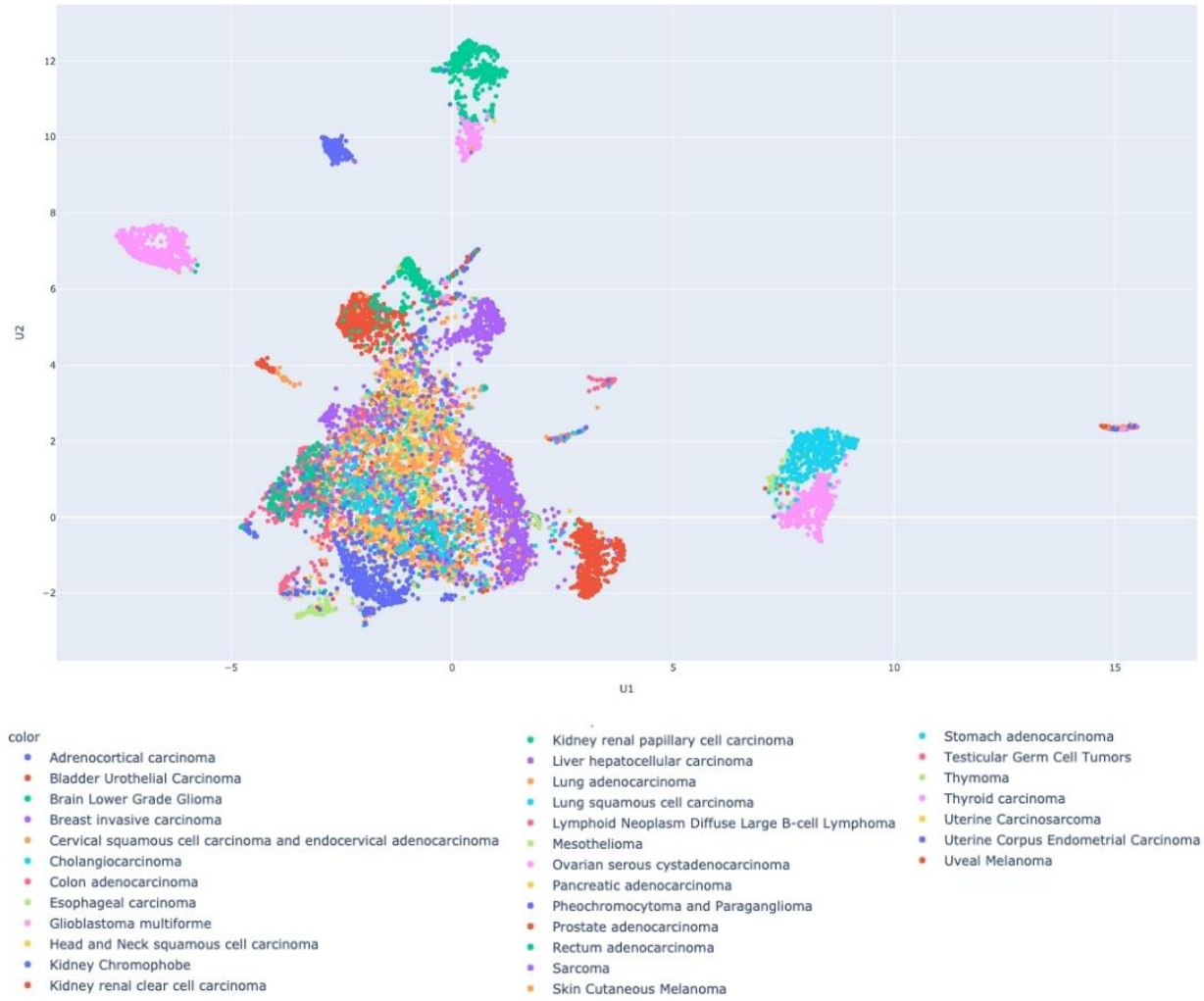


Figure 9 UMAP on the Normalized TCGA Dataset Using Only the log2 Counts from the Histone Variants/Chaperones Genes
On the x-axis figures the first component, and on the y-axis figures the second component.
Each point represents a sample, and it is colored by the sub-tissue it corresponds to.
Full size image available at: https://github.com/Ala-Eddine-BOUDEMIA/Chromatin-Dynamics/blob/main/Images/TCGA_PT/CPM/UMAP/variants_chaperones/Full/umap.png
Interactive plot available at: https://github.com/Ala-Eddine-BOUDEMIA/Chromatin-Dynamics/blob/main/Plotly_HTML_Files/TCGA_PT/CPM/UMAP/variants_chaperones/Full/umap.html

3.2. TCGA dataset:

As for the data from GTEx, plotting the mean-variance trend of the log2(CPM +1) from TCGA (Figure 7) showed that the standard deviation of the CPMs is independent of their mean, and the few highly variable genes that are mainly mitochondrial.

In order to evaluate if histone gene and chaperone expression can be used to discriminate between cancer types, PCA based on the counts from the histone variants and chaperones gene list was performed (Figure 8). The scatter plot shows no discrimination between any particular type as everything is mixed. As a control, PCA based on the top 1000 expressed genes was performed (Supplementary Figure 3), which yielded similar results. Overall, PCA is not well-suited to separate cancer types from the TCGA data, which might be due to the high variability of samples in cancer.
Following the same approach used to analyze GTEx data, and as PCA did not separate cancer types, UMAP was used to separate samples by cancer type.

The UMAP shows that the histone variants and histone chaperones genes can separate some cancer types distinctly, namely: acute myeloid leukemia, esophageal carcinoma, ovarian serous cystadenocarcinoma, pancreatic adenocarcinoma, thymoma, thyroid carcinoma, brain lower grade glioma, glioblastoma multiforme, liver hepatocellular carcinoma, prostate adenocarcinoma, pheochromocytoma and paraganglioma. However, the rest of samples did not cluster well and grouped towards the center of the plot (Figure 9). To compare these results, a control with the top 1000 expressed genes was performed (Supplementary Figure 4). This demonstrated a better separation of the samples as they distinctly grouped by cancer type except for several cancer types of epithelial origin (bladder, cervix, head and neck, lung, uterus), which overlapped with each other. Overall, the top 1000 expressed genes enabled the UMAP to correctly group more samples than the histone variants and histone chaperones genes.

This analysis demonstrates that some cancer types can be separated based on the expression patterns of the histone variant and chaperone genes, although the performance using this gene list compared to the Top 1000 genes is worse when using cancer compared to normal tissue data.

To study how these patterns vary across and within the different cancer types, cluster maps of gene expression by sample were generated for the gene list (Figure 10). To investigate how the genes themselves are co-expressed across cancers, cluster maps based on the Pearson correlation between their expression were also generated (Figure 11).

In agreement with the UMAP, the cluster map shows that the expression patterns of histone chaperone and variant genes across cancer types can distinguish some cancer types as depicted in (Figure 10). Additionally, the cluster map demonstrates how the genes in this set are co-expressed across cancers. Replicative histones form one big cluster that contains two subclusters. The first subcluster contains the majority of replicative histones and some testis-specific replacement histones. The second subcluster contains several replicative histone genes that are more expressed than the rest, along with one replacement H2B (H2bfs) variant. Replacement variants with chaperones form the second big group which is also composed of two subclusters. Of these, one contains mediumly-expressed replicative histones (HIST3h2a, HIST2h4a/b, HIST2h2be, HIST1h4h/i, HIST1h2bd/k, HIST2h2aa, HIST2h2aa4) clustering with replication-associated chaperones (CAF-1 subunits, POLE4, TONSL, DNAJC9, MMS22L), ASF1B, as well as the CENPA/HJURP pair, which is not replication-associated, but still regulated in cell cycle-

dependent manner.

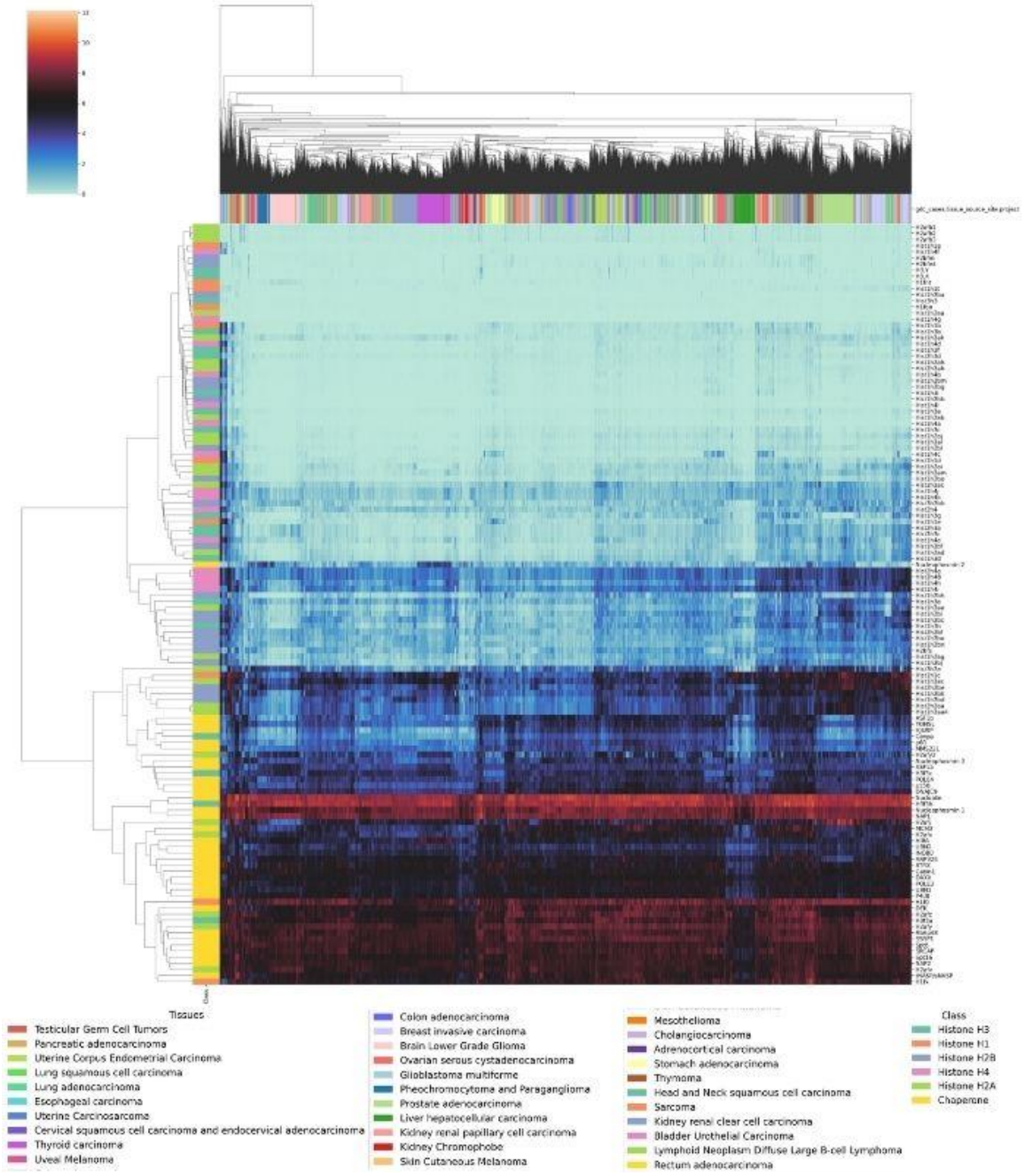


Figure 10 Cluster map based on All the Samples in TCGA by the log2 Counts from the Histone Variants / Chaperones Genes
The horizontal line is represented by all the samples in the TCGA dataset, and each cluster is colored by tissue type.
The vertical line is represented by all the histone variants / chaperones genes, and each cluster is colored by histone family or as chaperone.
Full size image available at: https://github.com/Ala-Eddine-BOUDEMIA/Chromatin-Dynamics/blob/main/Images/TCGA_PT/CPM/Clustermap/Samples_Genes/euclidean/variants_chaperones/Full/clustermap.png

The second subcluster contains highly-expressed replacement variants and their respective chaperones. Some of these are ubiquitously highly expressed here as in GTEx (Nucleolin, NAP1, NPM1 and H3F3b). Conversely to the situation in normal tissues, H3f3a is more highly expressed in cancer and does not cluster with H3f3b and H3F3c, but rather with H2A.Z and DEK alongside SRCAP, Spt6, FACT subunits and p48. Moreover, HIRA subunits are going in the same cluster with MCM2, INO80, P400, ANP32E and H2A.X

The results of this heatmap demonstrate that different subunits of CAF-1 and the respective histone variants it handles can be transcribed at different levels across cancer types, indicating they are not co-regulated on the transcriptional level. However, the different subunits of HIRA seem to be co-regulated in cancers, although not with the histone variants associated to HIRA. Similarly, to the normal tissue, the CENP-A/HJURP pair still demonstrate correlated levels of expression, but they are more highly expressed in cancer.

To investigate how similar histone variant and chaperone genes behave in terms of expression patterns (as opposed to expression levels), clustering of genes based on the Pearson correlation between them was performed (Figure 11). Conversely to the results shown in Figure 10, all the replicative histones cluster together and the only replacement variants that goes along with them is H2bfs. Within this cluster, a set of genes is distinct from the rest as it is highly correlated. This set is composed of 25 genes of which 23 are part of Hist1 locus and contains several copies of each histone family in similar proportions: H1 (4 genes), H2A (5 genes), H2B (4 genes), H3 (6 genes) and H4 (6 genes).

In the cancer context, the rest of the genes form a second cluster which is split into three main groups. The first subcluster contains mainly H3 variants (b, c) and replacement chaperones such as HIRA subunits, ANP32E, INO80, SRCAP, P400, ATRX, NAP1 and NAP2. The second subcluster contains mainly replication associated chaperones such as CAF-1 subunits, ASF1B and replisome-associated genes (TONSL, DNAJC9, MCM2, POLE4). However, this subcluster contains some replacement variants and chaperones as well, this is the case for CENP-A and HJURP, DAXX, H3f3a and H2A.X. Finally, the last subcluster comprises testis-specific variants, although they do not correlate neither between themselves nor with the rest of genes.

Similarly, to GTEx co-expression networks were built using iterative WGCNA to better visualize the co-expression relationships between the histone variant and chaperone genes.

Figure 12 shows the 70 genes that iterative WGCNA was able to fit into the network, forming 5 modules. These genes include the majority (55 out of 61) of the replicative histone genes, 12 chaperone genes and 4 replacement variants.

Four out of the five modules are interconnected and composed exclusively out of different combinations of replicative histone variants. All the genes in the red module are part of the same histone locus (Hist1), whereas the other three modules (yellow, green and pink) contain genes that are mainly part of one histone gene cluster with few of them that are from another cluster.

The only replacement variant that is part of those four modules is H2bfs and it is part of the green module, whereas the other three replacement variants (H2A.X, H2A.Z and CENP-A), are part of the light blue module. The latter is mainly composed of replication-related chaperones such as CAF-1 subunits (p60, p150), MCM2, TONSL, DNAJC9. This suggests that in the cancer context, there is a strong co-regulation of replicative, but not replacement histone variants and chaperones.

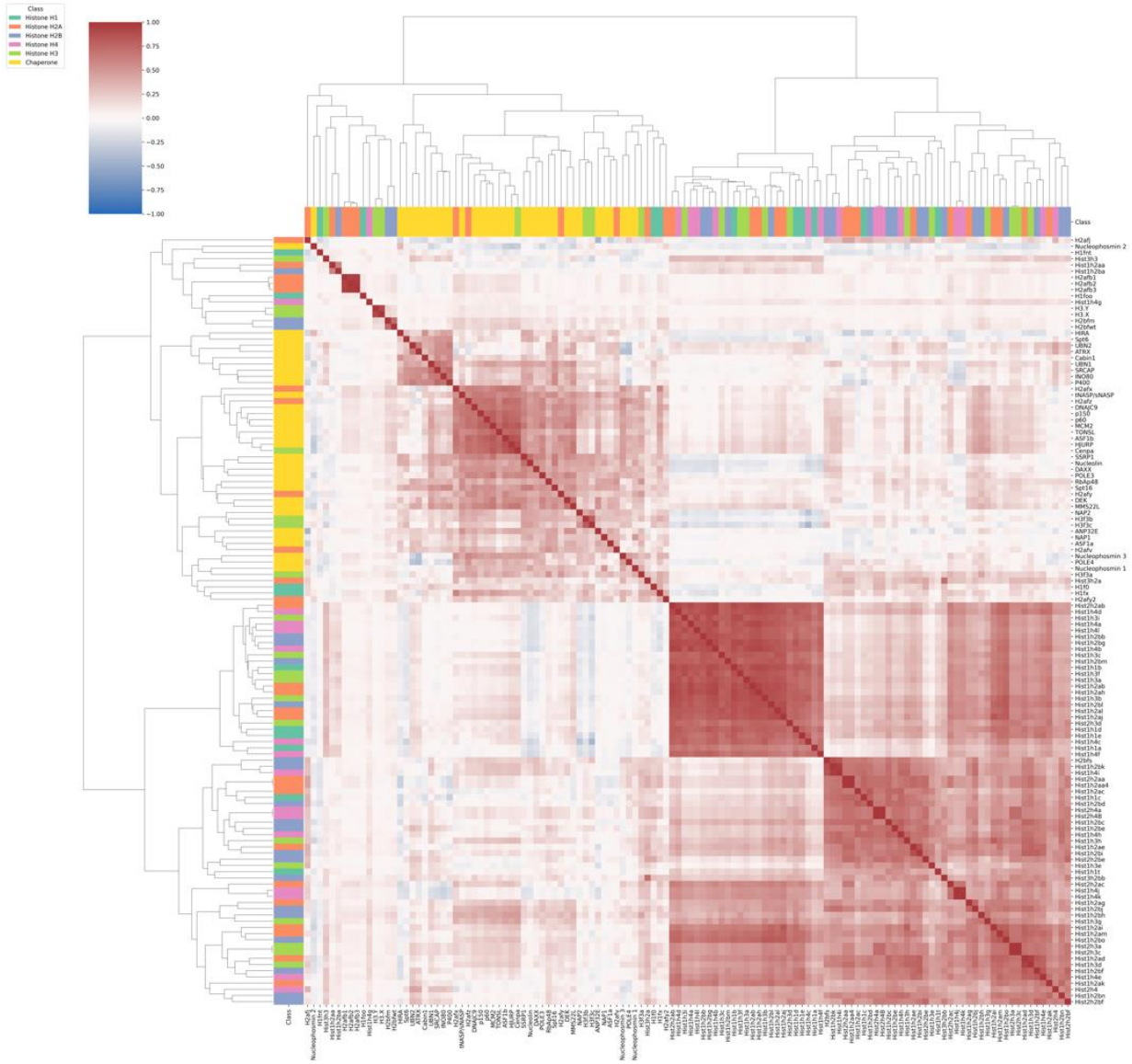


Figure 11 Cluster map Based on the Pearson Correlation Between Histone Variants / Chaperones Genes from the TCGA Dataset
Both the horizontal and the vertical lines are represented by all the histone variants / chaperones genes, and each cluster is colored by histone family or as chaperone
Full size image available at: https://github.com/Ala-Eddine-BOUDEMIA/Chromatin-Dynamics/blob/main/Images/TCGA_PT/CPM/Clustermap/Genes/euclidean/variants_chaperones/Full/clustermap.png

4. Discussion:

The analysis performed in this study shows that histone variant and chaperone genes exhibit distinct expression patterns across tissue types. This is clearly visualized by the fact they can separate samples by tissue of origin by UMAP (Figure 3) and hierarchical clustering (Figure 4), which had a great capacity in grouping similar samples together. More specifically, UMAP outperformed PCA in this particular context where the dataset is composed of a large number of samples. PCA has failed to explain the differences between the samples probably due to the variety of identities that each sample contribute to, as they do not come from the same individual. Conversely, UMAP was able to find the similarities between each sample and grouped them by tissue of origin in a very distinct manner which indicates that the relationships between samples is a non-linear one. However, it is important to note that even though UMAP does a good job at preserving the global structure, it is still not good enough to pursue the analysis based on the UMAP components.

Applying UMAP and hierarchical clustering on the histone variant and chaperone genes provided sample separation that was not perfect and some intermixing between samples of different tissues were observed. However, it is important to note that these tend to belong to tissues of the same system that share similar structure and therefore the same cell types in similar proportions (e.g., esophagus, stomach, small intestine, colon). Furthermore, some of them arise from the same embryonic germ layer suggesting that they shared the same transcriptional profile during some developmental stages.

Conversely, in a cancer context, the distinction of samples based on cancer type was worse as only 12 types out of 33 were separated based on histone variant and chaperone gene expression. This may be due to the fact that cancer cells are more highly proliferative and tend to de-differentiate with disease progression (Hanahan & Weinberg 2011). This may lead to a loss of the distinct histone variant/chaperone expression patterns of the normal tissue of origin. Indeed, while in normal testis, expression of testis-specific variants, such as H1fnt, H1foo and the H2A.B variants (Figure 4) can clearly be detected, this is not anymore, the case for cancers of testicular origin (Figure 10).

When examining the clustering patterns of the histone genes and chaperones across tissues it is critical to keep in mind that the library construction protocol was based on poly(A) selection for all GTEx and most TCGA samples. This is likely to influence the apparent expression levels of the replicative histones, as their transcripts will be lowly detected without a clear distinction if it is due to biological characteristics or to technical effects. Accordingly, replicative histones form a distinct group clearly observed in both normal and cancer context when clustering the genes either by expression (Figure 4, 10) or by correlation between them (Figure 5, 11).

Analyzing the cluster maps revealed that histone variants do not tend to be co-transcribed with their respective chaperones. In addition to the replicative histone gene group, two other major subclusters are present in the data. One is comprised mostly of testis-specific genes, which exhibit high correlation between each other in normal, but not cancer data, corresponding to the loss of tissue-specific expression. The other one contains the rest of the replacement variants and histone chaperones. Within this group, CENP-A, HJURP and ASF1B cluster and correlate with replication associated genes such as CAF-1 subunits (p60, p150), MCM2 and TONSL (Figure 5). This suggests that they are regulated in a cell-cycle dependent manner, especially as they also cluster closely to replicative histones based on their expression levels (Figure 4).

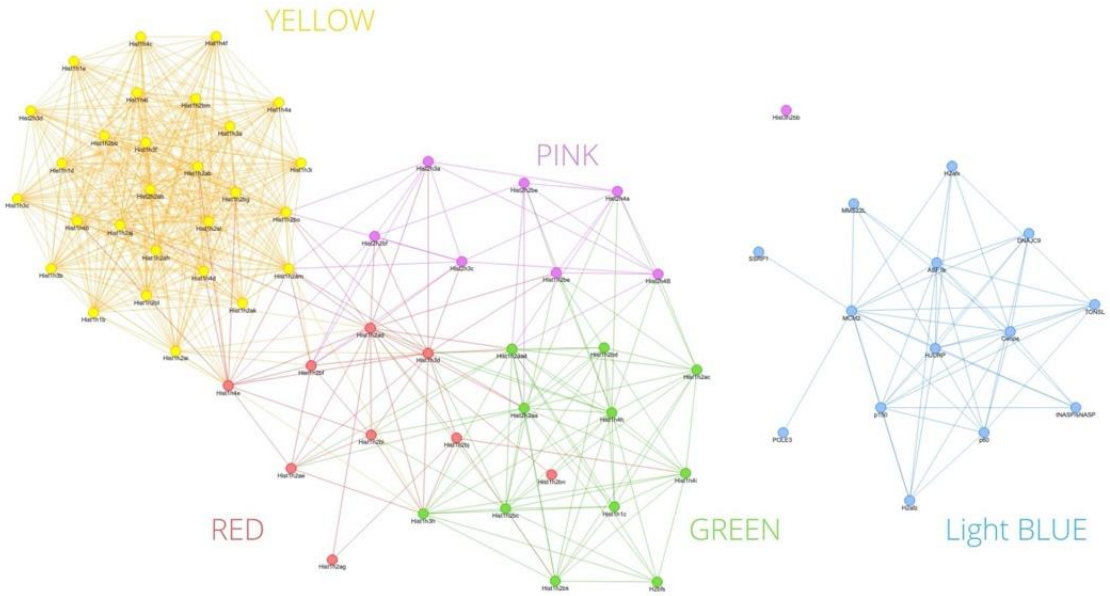


Figure 12 Co-expression Network that Was Generated by the iterative WGCNA Tool Using the Histone Variants / Chaperones Counts from the TCGA Dataset
Download interactive plot from: https://github.com/Ala-Eddine-BOUDEMIA/Chromatin-Dynamics/blob/main/WGCNA/TCGA/Networks/graph_tcga.html

In a cancer context, the cluster of the aforementioned genes also includes the H2A.X variant (Figure 11), which is known to have an important role in DNA damage response and replicative stress (Piquet et al. 2018), although they do not cluster with it by expression levels (Figure 10). Finally, CENP-A and HJURP present an increase of expression in a cancer context and become closer to replacement variants rather than replicative histones (Figure 10). This is in line with what is known about the association of CENP-A and HJURP with replicative status and disease state (Jeffery et al., 2021, Sun et al., 2016, Zasadzińska et al., 2018). However, it would be interesting to see if the observed increased of CENP-A/HJURP pair is more associated with metastatic cancers, using the metastatic samples from TCGA matched to their primary samples.

The replacement H3 variants do not necessarily cluster with their respective chaperones, which is also confirmed by the co-expression networks that group them together but do not connect them to their respective chaperones (Figure 6, 12). In normal tissue, H3F3A and H3F3B do not cluster together based on their expression levels (Figure 4), but they cluster together based on the expression patterns as depicted by the cluster map based on the Pearson correlation between genes which suggest that these two genes may be co-regulated (Figure 5). Notably, H3F3A shows a decreased level of expression in the brain tissue and increased level of expression in cancer context, suggesting that its transcriptional regulation may be related to cell type or to the replicative state of the cells.

When interpreting the results provided by this analysis it is important to keep in mind the limitations related to the datasets that have been used. Firstly, the unbalanced distribution of samples by tissue/cancer type leads to different amount of representation which could bias the hierarchical clustering towards the most represented type. To make sure that this did not affect the analysis it may be useful to filter the tissue types that have few samples and randomly sub-section the rest of the tissues to have the same minimal number of samples per class. However, this approach is not ideal as it would lead to a significant loss of the number of samples, and thus the amount of information, and it would take a significant amount of time to iterate over the different randomly selected sets. Secondly, the GTEx dataset provides metadata that is useful only for quality control, but not for biological interpretations. For instance, it would have been helpful to provide metadata about the proliferative state of the sample and its cellular composition. Conversely, the TCGA dataset provided metadata that is difficult to parse through as it was collected from three different portals making it redundant and contains a lot of missing values. Finally, the library construction protocol makes it difficult to interpret replicative histone genes patterns as these genes' transcripts don't exhibit a polyadenylated tail. This is an important point that should be considered for future large-scale projects as many researchers in the field would be interested in analyzing the expression levels of genes that do not exhibit a poly-adenylation signals.

The results provided by this analysis raise several questions that could be addressed to get a better view concerning histone variants and histone chaperone co-expression patterns and their importance for normal – or cancer cells.

For future directions, it would be interesting to determine the main genes that are driving the differences between tissue types, between healthy and cancer states and between different cancer subtypes arising from the same tissue. This would be helpful to evaluate if particular chaperones and/or histones could represent potential therapeutical targets. It would also be useful to characterize the histone chaperones and variants co-expression across different developmental stages using other publicly available RNA-seq data (ArrayExpress E-MTAB-6814). This would

enable to try and understand when and how the patterns change and what are the potential developmental mechanisms related to these changes. Finally, single cell RNA-seq could be used to characterize expression patterns within tissue types, as bulk RNA-seq assays millions of cells from the same tissue in different proportions that may vary from one sample to another. Hence, scRNA-Seq could be useful to reveal the cell type heterogeneity of normal and tumor tissues and give information on their distinct histone chaperones and variants expression patterns (ENA PRJNA515497, EBI E-GEOD-130148, ArrayExpression E-MTAB-7316, …etc.).

5. Conclusion:

This study provides a first broad approach into characterizing the histone variants and histone chaperones co-regulation across and within tissues in healthy as well as in diseased tissues. This analysis unveils how transcription of these genes vary across and within tissues and how their patterns change across different cancer cases. The results presented here could be used and deepened to build a comprehensive resource to interrogate the transcription of histone variants and their chaperones in healthy tissues and matched primary tumors, identify common regulatory principles. This would enable to evaluate whether general principle can emerge for their genomic organization and regulation as well as the potential relevance of individual variants and chaperones in disease.

References:

Abascal, F., Corpet, A., Gurard-Levin, Z. A., Juan, D., Ochsenbein, F., Rico, D., Valencia, A., & Almouzni, G. (2013). Subfunctionalization via adaptive evolution influenced by genomic context: the case of histone chaperones ASF1a and ASF1b. *Molecular biology and evolution*, *30*(8), 1853–1866. https://doi.org/10.1093/molbev/mst086

Amatori, S., Tavolaro, S., Gambardella, S., & Fanelli, M. (2021). The dark side of histones: genomic organization and role of oncohistones in cancer. Clinical epigenetics, 13(1), 71. https://doi.org/10.1186/s13148-021-01057-x

Bagert, J. D., Mitchener, M. M., Patriotis, A. L., Dul, B. E., Wojcik, F., Nacev, B. A., Feng, L., Allis, C. D., & Muir, T. W. (2021). Oncohistone mutations enhance chromatin remodeling and alter cell fates. *Nature chemical biology*, *17*(4), 403–411. https://doi.org/10.1038/s41589-021-00738-1

Banaszynski, L. A., Wen, D., Dewell, S., Whitcomb, S. J., Lin, M., Diaz, N., Elsässer, S. J., Chapgier, A., Goldberg, A. D., Canaani, E., Rafii, S., Zheng, D., & Allis, C. D. (2013). Hira-dependent histone H3.3 deposition facilitates PRC2 recruitment at developmental loci in ES cells. *Cell*, *155*(1), 107–120. https://doi.org/10.1016/j.cell.2013.08.061

Cheloufi, S., Elling, U., Hopfgartner, B., Jung, Y. L., Murn, J., Ninova, M., Hubmann, M., Badeaux, A. I., Euong Ang, C., Tenen, D., Wesche, D. J., Abazova, N., Hogue, M., Tasdemir, N., Brumbaugh, J., Rathert, P., Jude, J., Ferrari, F., Blanco, A., Fellner, M., … Hochedlinger, K. (2015). The histone chaperone CAF-1 safeguards somatic cell identity. *Nature*, *528*(7581), 218–224. https://doi.org/10.1038/nature15749

Collado-Torres, L., Nellore, A., Kammers, K., Ellis, S. E., Taub, M. A., Hansen, K. D., Jaffe, A. E., Langmead, B., & Leek, J. T. (2017). Reproducible RNA-seq analysis using recount2. *Nature biotechnology*, *35*(4), 319–321. https://doi.org/10.1038/nbt.3838

Cook, A. J., Gurard-Levin, Z. A., Vassias, I., & Almouzni, G. (2011). A specific function for the histone chaperone NASP to fine-tune a reservoir of soluble H3-H4 in the histone supply chain. *Molecular cell*, *44*(6), 918–927. https://doi.org/10.1016/j.molcel.2011.11.021

Dunleavy, E. M., Roche, D., Tagami, H., Lacoste, N., Ray-Gallet, D., Nakamura, Y., Daigo, Y., Nakatani, Y., & Almouzni-Pettinotti, G. (2009). HJURP is a cell-cycle-dependent maintenance and deposition factor of CENP-A at centromeres. *Cell*, *137*(3), 485–497. https://doi.org/10.1016/j.cell.2009.02.040

Elsässer, S. J., Noh, K. M., Diaz, N., Allis, C. D., & Banaszynski, L. A. (2015). Histone H3.3 is required for endogenous retroviral element silencing in embryonic stem cells. *Nature*, *522*(7555), 240–244. https://doi.org/10.1038/nature14345

Zasadzińska, E. et al. (2018). Inheritance of CENP-A nucleosomes during DNA replication requires HJURP. https://doi.org/10.1016/j.devcel.2018.09.003

Ferrand, J., Rondinelli, B., & Polo, S. E. (2020). Histone Variants: Guardians of Genome Integrity. Cells, 9(11), 2424. https://doi.org/10.3390/cells9112424

Filipescu, D., Müller, S., & Almouzni, G. (2014). Histone H3 variants and their chaperones during development and disease: contributing to epigenetic control. Annual review of cell and developmental biology, 30, 615–646. https://doi.org/10.1146/annurev-cellbio-100913-013311

Foltz, D. R., Jansen, L. E., Bailey, A. O., Yates, J. R., 3rd, Bassett, E. A., Wood, S., Black, B. E., & Cleveland, D. W. (2009). Centromere-specific assembly of CENP-a nucleosomes is mediated by HJURP. *Cell*, *137*(3), 472–484. https://doi.org/10.1016/j.cell.2009.02.039

Ghiraldini, F. G., Filipescu, D., & Bernstein, E. (2021). Solid tumours hijack the histone variant network. Nature reviews. Cancer, 21(4), 257–275. https://doi.org/10.1038/s41568-020-00330-0

Giaimo, B. D., Ferrante, F., Herchenröther, A., Hake, S. B., & Borggrefe, T. (2019). The histone variant H2A.Z in gene regulation. *Epigenetics & chromatin*, *12*(1), 37. https://doi.org/10.1186/s13072-019-0274-9

Goldberg, A. D., Banaszynski, L. A., Noh, K. M., Lewis, P. W., Elsaesser, S. J., Stadler, S., Dewell, S., Law, M., Guo, X., Li, X., Wen, D., Chapgier, A., DeKelver, R. C., Miller, J. C., Lee, Y. L., Boydston, E. A., Holmes, M. C., Gregory, P. D., Greally, J. M., Rafii, S., … Allis, C. D. (2010). Distinct factors control histone variant H3.3 localization at specific genomic regions. *Cell*, *140*(5), 678–691. https://doi.org/10.1016/j.cell.2010.01.003

Greenfest-Allen et. al 2017. iterativeWGCNA: iterative refinement to improve module detection from WGCNA co-expression networks. https://doi.org/10.1101/234062

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell*, *144*(5), 646–674. https://doi.org/10.1016/j.cell.2011.02.013

Heo, J. I., Cho, J. H., & Kim, J. R. (2013). HJURP regulates cellular senescence in human fibroblasts and endothelial cells via a p53-dependent pathway. *The journals of gerontology. Series A, Biological sciences and medical sciences*, *68*(8), 914–925. https://doi.org/10.1093/gerona/gls257

Ishiuchi, T., Enriquez-Gasca, R., Mizutani, E., Bošković, A., Ziegler-Birling, C., Rodriguez-Terrones, D., Wakayama, T., Vaquerizas, J. M., & Torres-Padilla, M. E. (2015). Early embryonic-like cells are induced by downregulating replication-dependent chromatin assembly. *Nature structural & molecular biology*, *22*(9), 662–671. https://doi.org/10.1038/nsmb.3066

Jaskowiak, P. A., Costa, I. G., & Campello, R. (2018). Clustering of RNA-Seq samples: Comparison study on cancer data. *Methods (San Diego, Calif.)*, *132*, 42–49. https://doi.org/10.1016/j.ymeth.2017.07.023

Jeffery, D., Gatto, A., Podsypanina, K., Renaud-Pageot, C., Ponce Landete, R., Bonneville, L., Dumont, M., Fachinetti, D., & Almouzni, G. (2021). CENP-A overexpression promotes distinct fates in human cells, depending on p53 status. *Communications biology*, *4*(1), 417. https://doi.org/10.1038/s42003-021-01941-5

Lacoste, N., Woolfe, A., Tachiwana, H., Garea, A. V., Barth, T., Cantaloube, S., Kurumizaka, H., Imhof, A., & Almouzni, G. (2014). Mislocalization of the centromeric histone variant CenH3/CENP-A in human cells depends on the chaperone DAXX. *Molecular cell*, *53*(4), 631–644. https://doi.org/10.1016/j.molcel.2014.01.018

Lonsdale, J., Thomas, J., Salvatore, M. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45,** 580–585 (2013). https://doi.org/10.1038/ng.2653

Lever, J., Krzywinski, M. & Altman, N. Principal component analysis. *Nat Methods* **14,** 641–642 (2017). https://doi.org/10.1038/nmeth.4346

Love, I. M. (2019, October 16). RNA-seq workflow: gene-level exploratory analysis and differential expression. Bioconductor. https://www.bioconductor.org/packages/devel/workflows/vignettes/rnaseqGene/inst/doc/rnaseqGene.html

Mannironi, C., Bonner, W. M., & Hatch, C. L. (1989). H2A.X. a histone isoprotein with a conserved C-terminal sequence, is encoded by a novel mRNA with both DNA replication type and polyA 3' processing signals. *Nucleic acids research*, *17*(22), 9113–9126. https://doi.org/10.1093/nar/17.22.9113

Martire, S., & Banaszynski, L. A. (2020). The roles of histone variants in fine-tuning chromatin organization and function. Nature reviews. Molecular cell biology, 21(9), 522–541. https://doi.org/10.1038/s41580-020-0262-8

Matsuda, R., Hori, T., Kitamura, H., Takeuchi, K., Fukagawa, T., & Harata, M. (2010). Identification and characterization of the two isoforms of the vertebrate H2A.Z histone variant. *Nucleic acids research*, *38*(13), 4263–4273. https://doi.org/10.1093/nar/gkq171

McInnes, L., Healy, J., Saul, N., & Großberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection. Journal of Open-Source Software, 3(29), 861. https://doi.org/10.21105/joss.00861

Mendiratta, S., Gatto, A., & Almouzni, G. (2019). Histone supply: Multitiered regulation ensures chromatin dynamics throughout the cell cycle. The Journal of cell biology, 218(1), 39–54. https://doi.org/10.1083/jcb.201807179

Moggs, J. G., Grandi, P., Quivy, J. P., Jónsson, Z. O., Hübscher, U., Becker, P. B., & Almouzni, G. (2000). A CAF-1-PCNA-mediated chromatin assembly pathway triggered by sensing DNA damage. *Molecular and cellular biology*, *20*(4), 1206–1218. https://doi.org/10.1128/mcb.20.4.1206-1218.2000

National Cancer Institute. (2019, March 6). The Cancer Genome Atlas - Data Types Collected. https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/using-tcga/types

Orphanides, G., Wu, W. H., Lane, W. S., Hampsey, M., & Reinberg, D. (1999). The chromatin-specific transcription elongation factor FACT comprises human SPT16 and SSRP1 proteins. *Nature*, *400*(6741), 284–288. https://doi.org/10.1038/22350

Piquet, S., Le Parc, F., Bai, S. K., Chevallier, O., Adam, S., & Polo, S. E. (2018). The Histone Chaperone FACT Coordinates H2A.X-Dependent Signaling and Repair of DNA Damage. *Molecular cell*, *72*(5), 888–901.e7. https://doi.org/10.1016/j.molcel.2018.09.010

Ray-Gallet, D., Woolfe, A., Vassias, I., Pellentz, C., Lacoste, N., Puri, A., Schultz, D. C., Pchelintsev, N. A., Adams, P. D., Jansen, L. E., & Almouzni, G. (2011). Dynamics of histone H3 deposition in vivo reveal a nucleosome gap-filling mechanism for H3.3 to maintain chromatin integrity. *Molecular cell*, *44*(6), 928–941. https://doi.org/10.1016/j.molcel.2011.12.006

Régnier, V., Novelli, J., Fukagawa, T., Vagnarelli, P., & Brown, W. (2003). Characterization of chicken CENP-A and comparative sequence analysis of vertebrate centromere-specific histone H3-like proteins. *Gene*, *316*, 39–46. https://doi.org/10.1016/s0378-1119(03)00768-6

Shrestha, R. L., Ahn, G. S., Staples, M. I., Sathyan, K. M., Karpova, T. S., Foltz, D. R., & Basrai, M. A. (2017). Mislocalization of centromeric histone H3 variant CENP-A contributes to chromosomal instability (CIN) in human cells. *Oncotarget*, *8*(29), 46781–46800. https://doi.org/10.18632/oncotarget.18108

Sullivan, K. F., Hechenberger, M., & Masri, K. (1994). Human CENP-A contains a histone H3 related histone fold domain that is required for targeting to the centromere. *The Journal of cell biology*, *127*(3), 581–592. https://doi.org/10.1083/jcb.127.3.581

Sun, X., Clermont, P. L., Jiao, W., Helgason, C. D., Gout, P. W., Wang, Y., & Qu, S. (2016). Elevated expression of the centromere protein-A(CENP-A)-encoding gene as a prognostic and predictive biomarker in human cancers. *International journal of cancer*, *139*(4), 899–907. https://doi.org/10.1002/ijc.30133

Tagami, H., Ray-Gallet, D., Almouzni, G., & Nakatani, Y. (2004). Histone H3.1 and H3.3 complexes mediate nucleosome assembly pathways dependent or independent of DNA synthesis. *Cell*, *116*(1), 51–61. https://doi.org/10.1016/s0092-8674(03)01064-x

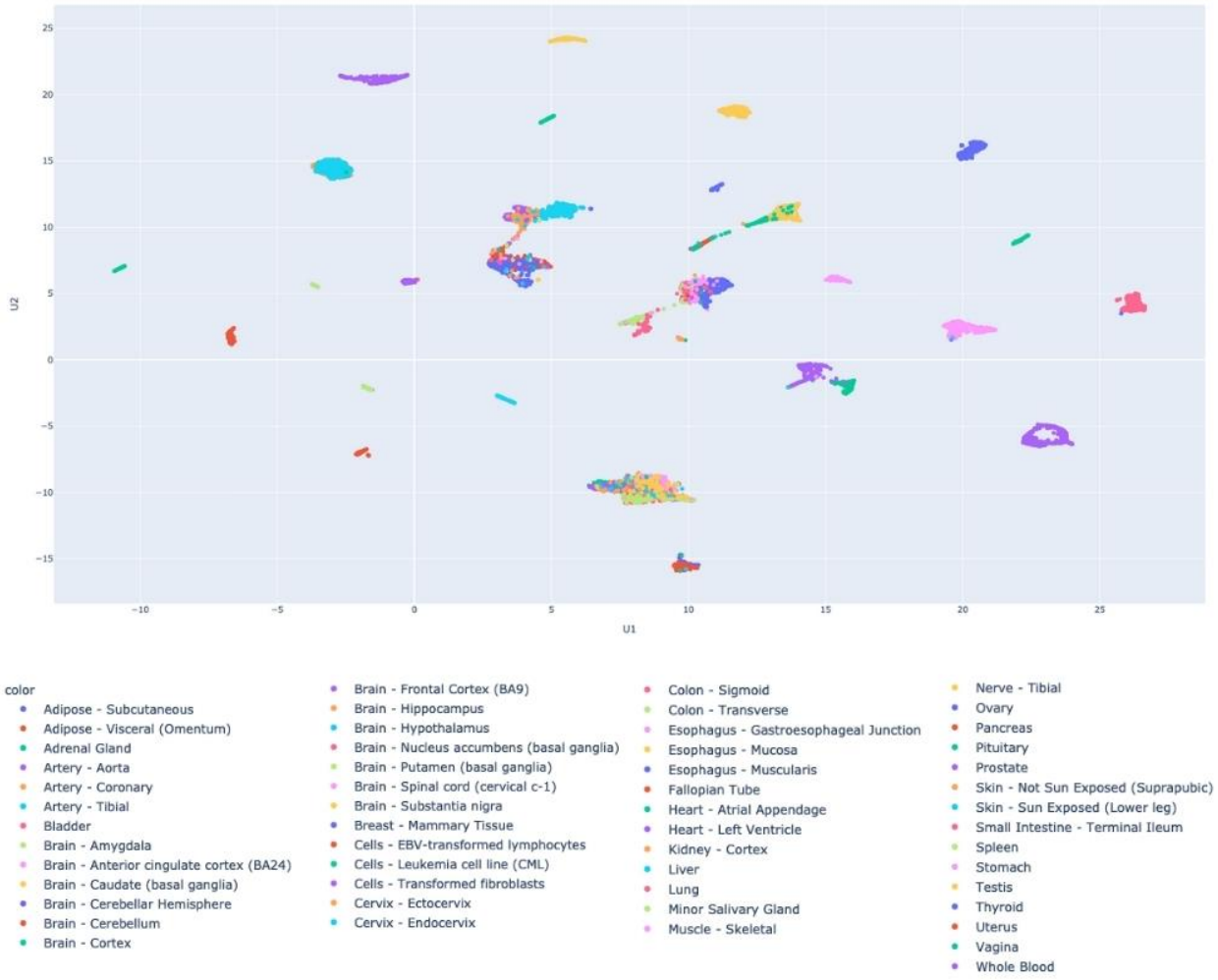
Torné, J., Ray-Gallet, D., Boyarchuk, E., Garnier, M., Le Baccon, P., Coulon, A., Orsi, G. A., & Almouzni, G. (2020). Two HIRA-dependent pathways mediate H3.3 de novo deposition and recycling during transcription. *Nature structural & molecular biology*, *27*(11), 1057–1068. https://doi.org/10.1038/s41594-020-0492-7

Valdés-Mora, F., Song, J. Z., Statham, A. L., Strbenac, D., Robinson, M. D., Nair, S. S., Patterson, K. I., Tremethick, D. J., Stirzaker, C., & Clark, S. J. (2012). Acetylation of H2A.Z is a key epigenetic modification associated with gene deregulation and epigenetic remodeling in cancer. *Genome research*, *22*(2), 307–321. https://doi.org/10.1101/gr.118919.110

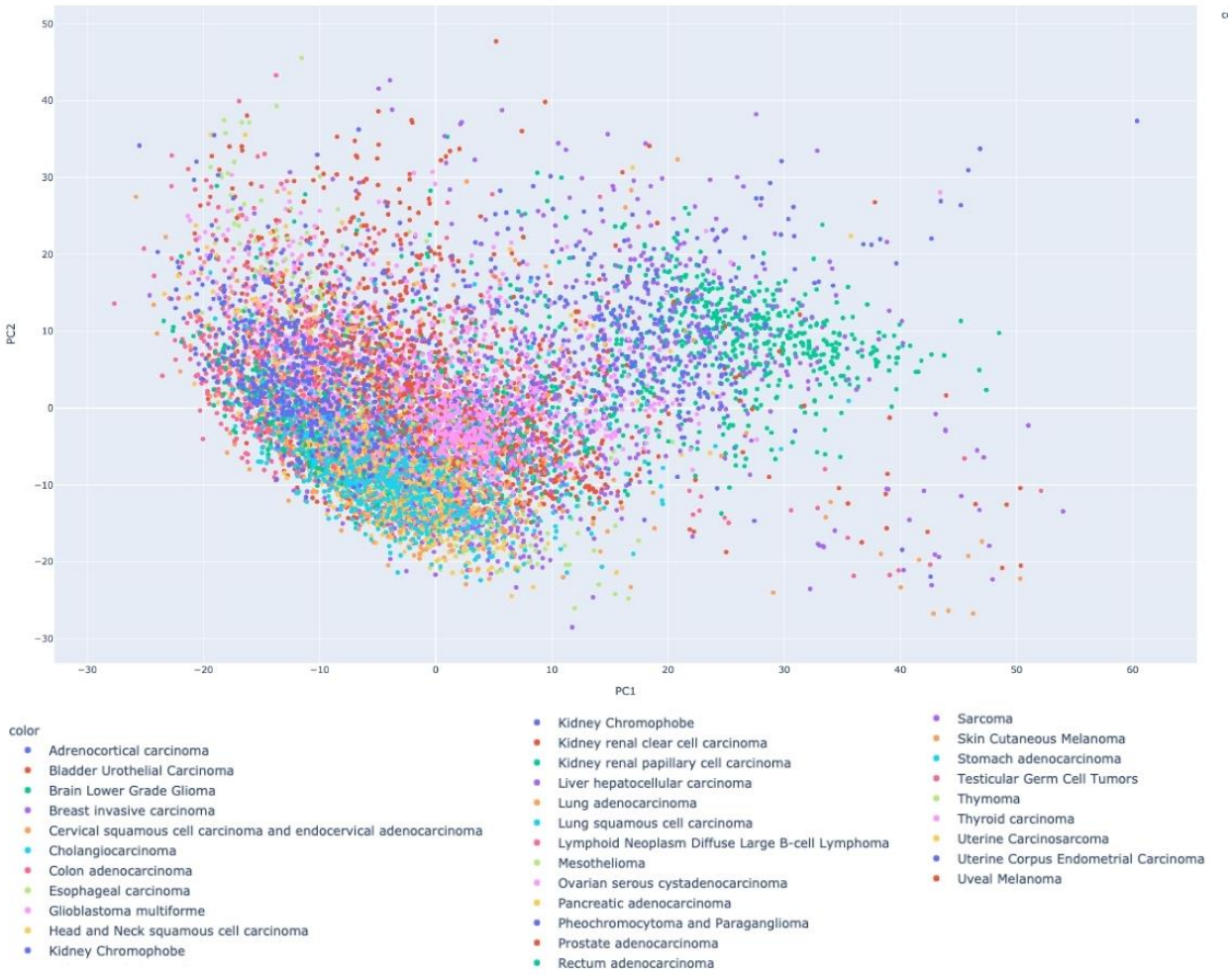
Van Dam, S., Võsa, U., van der Graaf, A., Franke, L., & de Magalhães, J. P. (2018). Gene co-expression analysis for functional classification and gene-disease predictions. *Briefings in bioinformatics*, *19*(4), 575–592. https://doi.org/10.1093/bib/bbw139

Yadav, T., Quivy, J. P., & Almouzni, G. (2018). Chromatin plasticity: A versatile landscape that underlies cell fate and identity. *Science (New York, N.Y.)*, *361*(6409), 1332–1336. https://doi.org/10.1126/science.aat8950

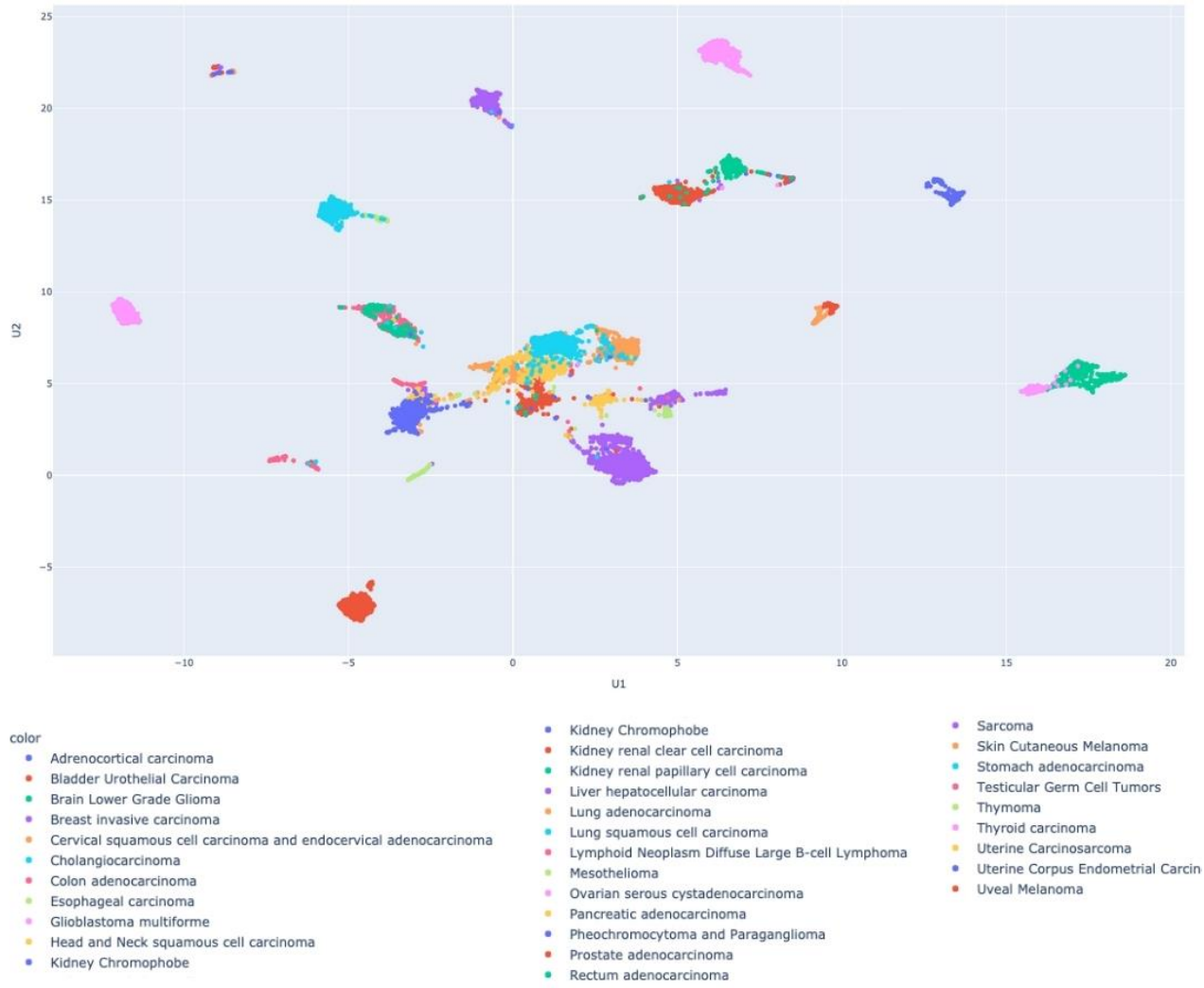
Supplementary Figure 1 PCA on the Normalized GTEx Dataset Using Only the log2 Counts from the Top 1000 Expressed Genes
On the x-axis figures the first principal component and, on the y-axis, figures the second principal component.
Each point represents a sample, and it is colored by the tissue type it corresponds to.
Full size image with detailed sub-tissues available at: https://github.com/Ala-Eddine-BOUDEMIA/Chromatin-Dynamics/blob/main/Images/GTEx/CPM/PCA/Top1000/pca.png
Interactive plot available at: https://github.com/Ala-Eddine-BOUDEMIA/Chromatin-Dynamics/blob/main/Plotly_HTML_Files/GTEx/CPM/PCA/Top1000/pca.html



Supplementary Figure 2 UMAP on the Normalized GTEx Dataset Using Only the log2 Counts from the Top 1000 Expressed Genes
On the x-axis figures the first component, and on the y-axis figures the second component.
Each point represents a sample, and it is colored by the sub-tissue it corresponds to.
Full size image available at: https://github.com/Ala-Eddine-BOUDEMIA/Chromatin-Dynamics/blob/main/Images/GTEx/CPM/UMAP/Top1000/umap.png
Interactive plot available at: https://github.com/Ala-Eddine-BOUDEMIA/Chromatin-Dynamics/blob/main/Plotly_HTML_Files/GTEx/CPM/UMAP/Top1000/umap.html



Supplementary Figure 3 PCA on the Normalized TCGA Dataset Using Only the log2 Counts from the Top 1000 Expressed Genes
Full size image available at: https://github.com/Ala-Eddine-BOUDEMIA/Chromatin-Dynamics/blob/main/Images/TCGA_PT/CPM/PCA/Top1000/pca.png
Interactive plot available at: https://github.com/Ala-Eddine-BOUDEMIA/Chromatin-Dynamics/blob/main/Plotly_HTML_Files/TCGA_PT/CPM/PCA/Top1000/pca.html



Supplementary Figure 4 UMAP on the Normalized TCGA Dataset Using Only the log2 Counts from the Top 1000 Expressed Genes
On the x-axis figures the first component, and on the y-axis figures the second component.
Each point represents a sample, and it is colored by the sub-tissue it corresponds to.
Full size image available at: https://github.com/Ala-Eddine-BOUDEMIA/Chromatin-Dynamics/blob/main/Images/TCGA_PT/CPM/UMAP/Top1000/umap.png
Interactive plot available at: https://github.com/Ala-Eddine-BOUDEMIA/Chromatin-Dynamics/blob/main/Plotly_HTML_Files/TCGA_PT/CPM/UMAP/Top1000/umap.html

Summary

The aim of the project was to characterize in silico the co-expression patterns of histone variants and histone chaperones across tissues and disease states. Histones are the building blocks of chromatin, the complex in which DNA is packaged in the nucleus of eukaryotic cells. Each of the five families of histones (H1, H2A, H2B, H3, H4) comprises different variants that are associated with distinct nuclear processes and contribute to the regulation of genome function (reviewed in Szenker et al., 2014, Martire and Banaszynski, 2020). Thus, expression and incorporation of histone variants into chromatin is regulated to ensure normal cell function throughout the cell cycle and in development (reviewed in Mendiratta et al., 2019). Furthermore, throughout their cellular life, histones are escorted by a network of histone chaperones to ensure their proper usage from synthesis to degradation and coordinate their dynamics in and out of chromatin. Histone chaperones can be dedicated to distinct histone families or histone variants (reviewed in Gurard-Levin et al., 2014, Hammond et al., 2017). This study allowed to characterize the histone variants and histone chaperones transcriptional co-regulation patterns across and within tissues in healthy as well as in diseased tissues. It demonstrated that the histone variant and chaperone genes exhibit distinct patterns across tissues and similar patterns within tissues of the same system or arising from the same embryonic origin. This analysis could be pushed further by investigating the set of genes that are responsible for the differences between tissues and to characterize the patterns across different developmental stages.

Résumé

L'objectif de ce projet était de caractériser in silico les patterns de co-expression des variantes d'histone et des chaperons d'histone à travers les différents tissus et états pathologiques. Les histones sont les éléments constitutifs de la chromatine, cette dernière compacte l'ADN dans le noyau des cellules eucaryotes. Chacune des cinq familles d'histones (H1, H2A, H2B, H3, H4) comprend différentes variantes qui sont associées à des processus nucléaires distincts et contribuent à la régulation de la fonction génomique (Szenker et al., 2014, Martire et Banaszynski, 2020). Ainsi, l'expression et l'incorporation de variantes d'histone dans la chromatine sont régulées pour assurer une fonction cellulaire normale tout au long du cycle cellulaire et durant le développement (Mendiratta et al., 2019). De plus, à travers leur vie cellulaire, les histones sont escortées par un réseau de chaperons d'histones pour assurer leur bon usage. Les chaperons d'histones peuvent être spécifique vers une famille d'histones particulière ou à des variantes d'histones précises comme ils peuvent être général (Gurard-Levin et al., 2014, Hammond et al., 2017). Cette étude a permis de caractériser les patterns de co-régulation transcriptionnelle des variantes d'histones et des chaperons d'histones à travers et au sein des tissus sains et malades. Cette analyse a démontré que les variantes d'histones et les gènes chaperons présentent des patterns distincts à travers les tissus et des patterns similaires entre les tissus du même système ou provenant du même tissue embryonnaire. Finalement, Cette analyse pourrait être poussée plus loin en étudiant l'ensemble des gènes qui sont responsables des différences entre les tissus ou en caractérisant les modèles à travers les différents stades de développement.