Final Report

Master 2 GenE2

2022/2023

Rand Fatouh

Shedding light on cancer resistance based on histone variants and chaperones networks.

Supervisor: Geneviève Almouzni - Sébastien Lemaire
Curie Institute - Nuclear Dynamics Unit UMR3664 – Chromatin Dynamics
E-mail : genevieve.almouzni@curie.fr
Tel : 01 56 24 67 01
Address: 26 rue d'Ulm 75248 PARIS 05EME FRANCE

Table of Contents:

Abbreviations:

ATRX: α-thalassemia/mental retardation syndrome protein.

ASF1: anti-silencing factor 1.

CABIN1: calcineurin binding protein 1.

CAF-1: chromatin assembly factor-1.

CENP-A: centromeric protein A.

DAXX: death domain–associated protein.

DNA: deoxyribonucleic acid.

DSB: double strand break.

FACT: facilitated chromatin transcription.

GLM: Generalized Linear Model.

GTEx: Genotype-Tissue Expression.

HCC: hepatocellular carcinoma.

HIRA: histone cell cycle regulator A.

HJURP: Holliday junction recognition protein.

hPTM: histone post-translational modifications

KIRC: Kidney renal clear cell carcinoma.

KIRP: Kidney renal papillary cell carcinoma.

LAML: Acute Myeloid Leukemia.

LGG: Brain Lower Grade Glioma.

LUAD: Lung adenocarcinoma.

LUSC: Lung squamous cell carcinoma.

mRNA: Messenger RNA.

NASP: nuclear autoantigenic sperm protein.

ncRNA: ncRNAnon-coding ribonucleic acid.

NPM1: nucleophosmin 1.

NPM2: nucleophosmin 2.

PCA: Principal component analysis.

PDX: Patient derived xenografts.

PolyA: polyadenylated.

PR: progesterone.

PRAD: Prostate adenocarcinoma.

RNA: ribonucleic acid.

SRCAP: Snf2-related CREBBP activator protein.

TCGA: The Cancer Genome Atlas.

TGCT: Testicular Germ Cell Tumors.

THCA: Thyroid carcinoma.

THYM: Thymoma.

TMM: Trimmed Mean of the M-values.

TNBC: Triple negative breast cancer.

TSS: transcription start site.

UBN1: ubinuclein 1.

UBN2: ubinuclein 2.

UMAP: Uniform Manifold Approximation and Projection.

UVM: Uveal Melanoma.

Introduction

While genetic mutations are a major factor in the onset and progression of diseases, specifically cancer, and they are an important driving force to improve patient health, more recent work has brought forward epigenetic alterations as another important parameter in the disease state. Here, I will refer to epigenetic modifications as inheritable and reversible changes in chromatin states that can affect genome function without altering the DNA sequence *(Yadav et al, 2018).* They include DNA methylation, histone post-translational modifications (hPTM), regulatory non-coding ribonucleic acids (ncRNA), and histone variants regulation *(Ghiraldini et al., 2021)*. A role for epigenetic modifications in the resistance to therapy in cancer has involved for example either downregulating tumor suppressors or upregulating oncogenic factors (Figures 1 & 2). This has led to developing drugs to reverse these alterations as effective in reducing resistance especially when combined with other agents (*Morel et al., 2020)*. While several epigenetic drugs target the regulation of hPTM for instance, histone variants are less investigated although previous studies linked their misregulation with cancer grades and sensitivity to treatment (Figure 1).

Nucleosomes are the building blocks of chromatin. This unit comprises a core particle, namely 147 bp of DNA wrapped around a histone octamer consisting of two copies of the core histones H2A, H2B, H3, and H4. In addition to the linker histone H1 (Martire et al., 2020). Histones exist as distinct variants that I will describe in particular in humans and which are encoded by distinct genes. Histone variants are broadly split in two groups. The first group is the DNA-synthesis coupled histones, which are deposited by designated chaperones into nucleosomes wrapping newly synthesized DNA during the S phase (expression peak) of the cell cycle *(Mendiratta et al., 2019)*. They are encoded by intronless genes and their cognate pre-mRNA not matured by polyadenylation *(Marzluff, 2005; review Marzluff, 2008)*. The other histone variants, whose expression is not coupled with DNA synthesis, differ in their amino acid sequences. They are matured by polyadenylation and their encoding genes have introns, which may result in different isoforms. Most of histone variants are expressed in all cells but only few of them are tissue specific. Additionally, the expression of variants start at different stages of development (Table 1).

Histone group	Histone variant	Tissue Specificity	Replication dependent	Chaperone	Development*
H3	H3.1 and H3.2	Global	Yes	CAF-1, MCM2, ASF1, NASP.	2-cell embryonic
	H3.3	Global	No	HIRA, DAXX-ATRX, ASF1, NASP.	(Zygote)
	CENP-A	Global	No	HJURP, ASF1.	Unknown
	H3.4	Testes	No	NAP2.	Not relevant
	H3.5	Testes	No	Unknown.	Not relevant
H2A	H2A.Z	Global	No	p400–TIP60, SRCAP.	Early blastocyst
	MacroH2A	Global	No	FACT.	8-cell stage
	H2A.B	Testes and brain	No	NAP1.	Not relevant
	H2A.X	Global	No	FACT.	(Zygote)
H4	H4.7	Global	Yes	NPM1.	Unknown.
H1	H1.1-H1.5	Global	Yes	NPM1, NASP, NAP1.	H1.1-H1.5, H1.0, and H1.8 are highly expressed in Zygotes. H1.6 and H1.7 are Testes specific. H1.10, unknown.
	H1.0, H1.10	Global	No		
	H1.6, H1.7	Testes	No		
	H1.8	Oocytes	No		

Table 1 - Histone Variants, their expression pattern across tissues, their chaperones, and their expression in embryonic stages. This includes the variants of H3, H2A, H4, and linker H1. **reviewed in Buschbeck et al., 2017 and Pan et al., 2016, Kawamura et al., 2021*

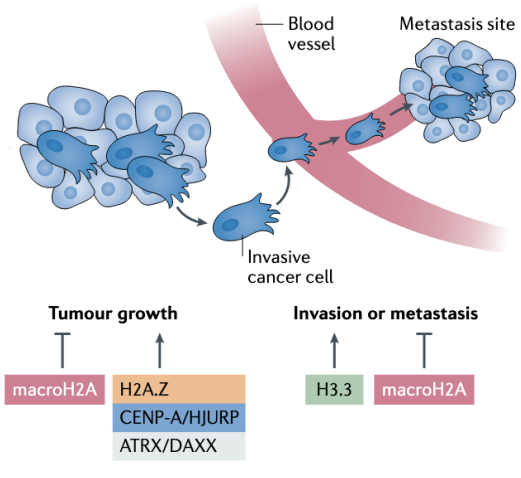


Figure 1 - Contribution of altered histone variants and their chaperones to the stages of tumor development.

Core histones, their variants, and their chaperones:

H3 Variants: H3 histone variant tetramerizes with H4 $(H3\text{-}H4)_2$ to form a nucleosome. It has eight variants: the 2 replicative variants H3.1 and H3.2, and the non-replicative variants H3.3, CENP-A, H3.4, H3.5, H3.Y.1, and H3.Y.2 *(Martire et al., 2020)*. They are incorporated by the specific chromatin assembly factor-1 (CAF-1) complex, composed of three subunits: p150, p60, and p48 *(Volk et al., 2015)*, mediated by the replication-associated helicase Mcm2 *(Richet et al., 2015).* The two paralogous genes H3-3A and H3-3B encode the H3.3 variant *(Martire et al., 2020).* H3.3 is found in different types of locations according to the chaperone which takes it in charge: the histone cell cycle regulator A (HIRA) complex containing in addition to Hira, ubinuclein 1 or ubinuclein 2 (Ubn1 or Ubn2) and the calcineurin binding protein 1 (Cabin1) subunits in the transcribed regions, the death domain–associated protein (Daxx) associated with the chromatin remodeling factor α-thalassemia/mental retardation syndrome protein (Atrx) *(Goldberg et al., 2010)* in pericentric regions and in telomeres. CenpA is a H3 variant restricted to centromeres by its dedicated chaperone Holliday junction recognition protein (Hjurp) (*Lacoste et al., 2014)*. While the previous H3 variants are expressed in every tissue, H3.4 and H3.5 are restricted to testis *(Urahama et al., 2016)*. Also, the anti-silencing factor 1 (ASF1) is a chaperone that carries all the H3 variants from their translation location to the depositing chaperone. The two paralogs ASF1a and ASF1b encode Asf1: which have distinctive gene expression patterns *(Abascal et al., 2013)*. Finally, previous studies associated the nuclear autoantigenic sperm protein (NASP) chaperone with H3.1/H3.2 and H3.3 *(Wang et al., 2008)*.

H2A Variants: Concerning H2A histone, there are four different classes of variants in mammals: H2A.X, H2A.Z, macroH2A and the testis-specific short H2A variants (H2A.B, H2A.L, H2A.P and H2A.Q) *(Martire et al 2020)*. H2A.Z, incorporated by p400–Tip60 and Snf2-related CREBBP activator protein (SRCAP), is usually present at transcription start sites (TSSs) and enhancers regions *(*Ku et al 2012, *Ghiraldini et ., 2021, Ku et al., 2012)*. Two homologous genes encode this variant, H2AZ1 and H2AZ2, yielding two functionally redundant isoforms H2A.Z.1 and H2A.Z.2 despite differing by three amino acids (*Matsuda et al., 2010*). MacroH2A has an additional C-terminal macro domain that resembles the H1 linker in properties and function *(Chakravarthy et al*

2012, Sun et al., 2019). MacroH2A1 and MacroH2A2 paralogous genes encode two isoforms, macroH2A.1 and macroH2A.2. While macroH2A does not have a dedicated chaperone, the genomic regions where it is incorporated are enriched with the facilitated chromatin transcription (FACT) complexes. MacroH2A's function relates to gene repression and compressibility of chromatin. The H2A.B variant is only expressed in testis and brain. It is the smallest variant encoded by three different genes, H2A.B1, H2A.B2 and H2A.B3, and is believed to destabilize nucleosomes, which promotes transcription *(Soboleva et al., 2011)*. H2A.X is encoded by one gene and plays a role in DNA double strand break (DSB) repair upon its phosphorylation, then called γH2A.X. FACT is proposed to deposit H2A.X, a variant that does not have any dedicated chaperone *(Piquet et al., 2018)*.

H2B and H4 Variants: The H2B histone gets incorporated through its dimerization with H2A.Therefore, there are no known chaperones to date directly in contact with H2B. H2B also has variants however their study is still limited *(Martire et al., 2020).* H4 has only one variant in humans, H4.7, in addition to the canonical replicative variants, with 85% of identity. It is also expressed during S phase yet *Long et al., 2019* detected it matured through polyadenylation, and localized in the nucleolus. It interacts with the chaperone NPM1 and the difference in amino acids impairs its ability to form a nucleosome in vitro.

H1 Variants: The linker histone H1 has 11 variants with one encoding gene for each. Some of them are replication-dependent (H1.1-H1.6), two are replication-independent (H1.0 and H1.10), in addition to two testis-specific (H1.6, H1.7) and one oocyte-specific (H1.8) subtypes. They all lack a polyA tail *(Scaffidi, 2016)*. The chaperones known to bind H1 variants are NASP, NAP1, and nucleophosmin (NPM1) *(Wang et al,. 2008, Zhang et al., 2015, and Gadad et al., 2011).*Table 1 provides a summary of histone variants and their chaperones along with their expression profile regarding tissue specificity, replication and development.

Context, aims and objectives of the internship:

Many histone variants and their chaperones are altered in cancerous tissues. They can be modified, mutated or misregulated *(Ghiraldini et ., 2021)*. For example, the downregulation of macroH2A variants in many tumors like breast cancer, melanoma,

and colorectal cancer has been associated with the increased proliferation of cancerous cells *(reviewed in Hsu et ., 2021)*. In contrast, there is an overall decrease in the expression of H1 variants in ovarian cancer with differences on the variant level. H1.1 appears to be the most down-regulated *(Medrzycki et al,. 2012)*. Moreover, the variant/chaperone couple CenpA/Hjurp is overexpressed in many cancers (e.g. liver, pancreatic, breast, …) and linked to tumor progression and low survival rate. Both are currently considered as unfavorable prognostic markers *(reviewed in Mahlke et ., 2020)*. Furthermore, overexpression of CenpA may limit the capacity of its chaperone Hjurp, then, Daxx *(Lacoste et al., 2014)* and Hira *(Nye et al., 2018)* are involved in the ectopic deposition of CenpA on chromosome arms *(Lacoste et al, 2014, Renaud-Pageot et al., 2022).* CenpA expression and distribution in the nucleus has also been linked to therapy outcome and the prediction of tumor sensitivity/ resistance (*Verrelle et al., 2021; Jeffery et al., 2021*). Finally, H3.1 and H3.3 regulation in cancer settings were respectively correlated with the expression of their dedicated chaperones, CAF-1 and Hira. Hira can also compensate for the CAF-1 and DAXX-ATRX deficiencies contributing to Epithelial-mesenchymal transition (EMT) in the first case and to the maintenance of telomeres in the second (Figure 2) *(Ray-Gallet and Almouzni, 2022).* Hence, histone variants and their chaperones form a network which is rewired upon perturbations in cancer (Figure 3) *(Mendiratta et al., 2019).* While previous published studies on cancer showed that the expression of some histone variants and chaperones was abnormal, there is a lack of a systematic review on the topic. Furthermore, how the histone/chaperone network is influenced by these abnormalities has not been investigated yet at a broad scale. So, our question is:

"How do the histone/chaperone networks characterize the healthy or cancerous status of tissues and whether it can predict resistance to therapy?"

We will address the question by breaking it into three steps (Figure 4):

- Characterizing tissue signatures by analyzing the network of histone variants and chaperones in the normal state.
- Examining how the histone variants/chaperones networks are altered in cancer by comparing cancer tissues with their healthy equivalents.
- Looking for histone/chaperones relationships that can be predictive of resistance

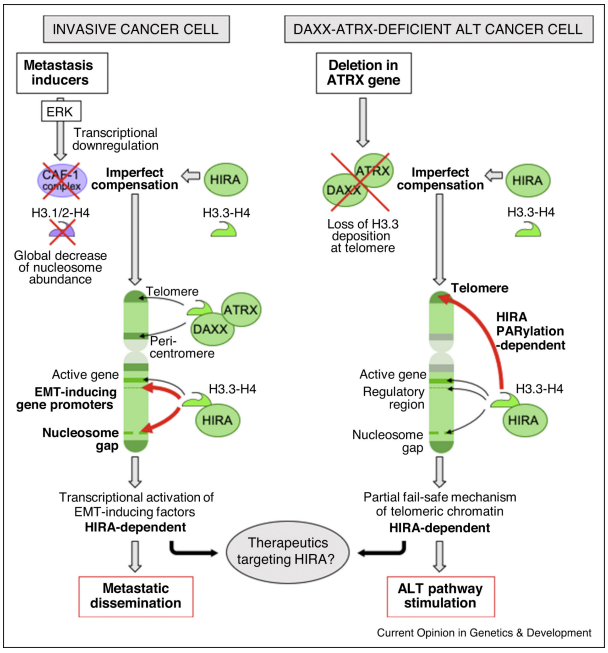


Figure 2 - HIRA chaperone compensates for CAF-1 and DAXX-ARTX during cancer. This leads to metastatic dissemination in the first case (left), and to alternative lengthening of telomere (ALT) pathway stimulation in the second (right). The figure illustrates the mechanism in which that happens (*Ray-Gallet and Almouzni, 2022)*.

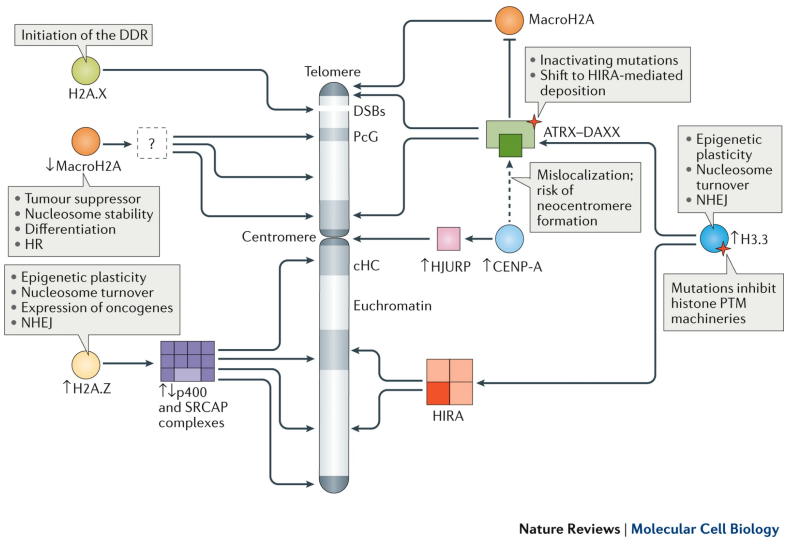


Figure 3 - Histone variants and chaperones networks deregulation in cancer. The network presented involves H2A.X, H2A.Z, MacroH2A, H3.3, and CENP-A histone variants and the following chaperones: HIRA, HJURP, ARTX-DAXX, SRCAP, and p400. The effects of misregulation are also presented *(Buschbeck et al., 2017).*

Materials and Methods:

This projects follows the work of Alla Aladine and Sebastian Lemaire

Environment:

Most of the analysis was done using python version 3.10. Some parts were performed with R version 4.2.2.

Datasets:

To study histone variants and chaperones expression across tissues in the normal state, we exploited RNA seq data from the Genotype-Tissue Expression (GTEx) database, a publicly available dataset containing multi-omics data for around 54 types of healthy organ tissues from 948 consenting donors. The samples were collected postmortem *(GTEx Consortium, 2015)* yet (*Melé et al. 2015*) showed that postmortem ischemia has limited impact. The main advantages of the GTEx database is the large number of samples provided (22951), the collection of multiple types of data from the same tissue allowing integrative analysis, and that it utilizes primary tissues in place of cell lines which multi-omics experiments are usually performed in *(eGTEx Project, 2017)*. The internship project followed what was previously done including filtering the raw count data to discard genes with a too low resolution to be quantitatively evaluated. Since the smallest class in the GTEx dataset contains 80 samples, genes having at least 6 reads in at least 80 samples each were kept, as previously done in *(Yab et al., 2021)*. To examine how the expression of histones variants and chaperones is altered in cancer, and how the network they form is altered, we exploited RNA-seq data from The Cancer Genome Atlas (TCGA) database, a publicly funded database with the initial aim of understanding and categorizing genetic alterations in cancer. Several datasets were made available on the platform for different types of omics data including genomic sequencing, gene expression, DNA methylation, and copy number variation. This data is available for over 30 types of cancer coming from 11000 patients. The experiments were performed on autopsies taken from both the cancers and the surrounding tissues which gives indications about the tumor microenvironment as well *(Wang et al., 2016)*.

In both datasets, Only polyA-matured RNAs have been sequenced.

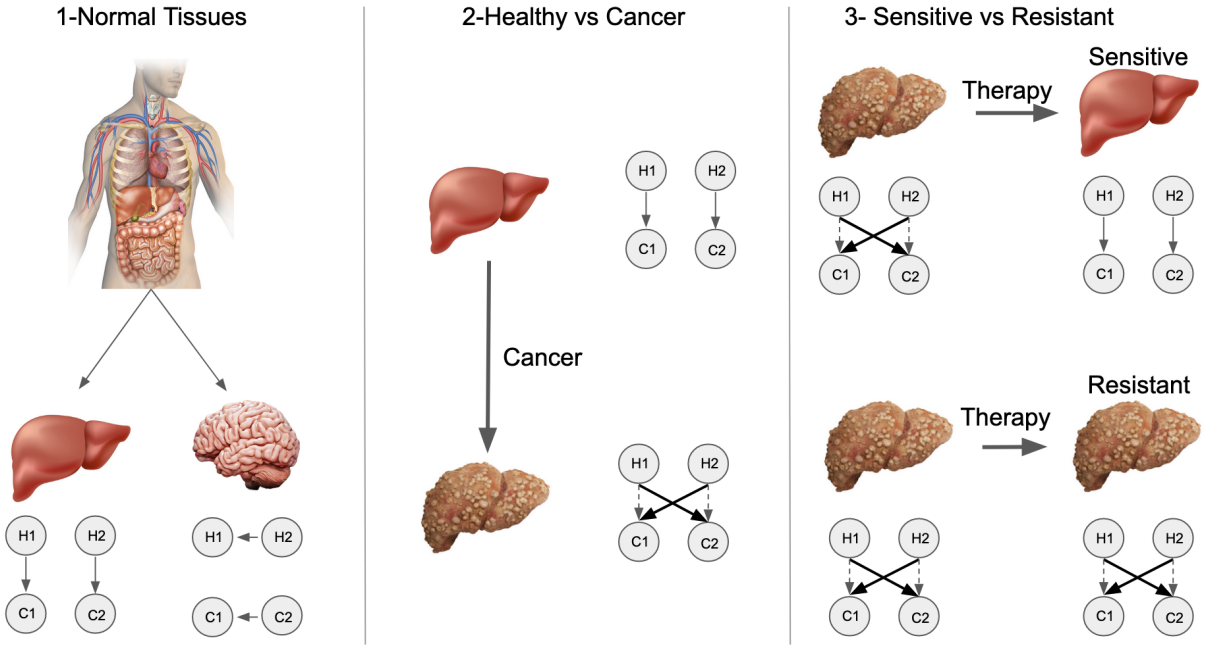


Figure 4 - The different comparisons that we will perform during the analysis. H1, H2: Histone variants 1 and 2. C1, C2: Chaperones 1 and 2. Arrows represent co-expression relationships. Bold and dashed arrows are acquired and lost relationships, respectively. 1) The Histone/Chaperone network will be established for every normal tissue, like liver and brain in the display. 2) Cancer-specific Histone/Chaperone networks will be analyzed in respect of the tissue of origin to identify cancer-associated rewirings in the network. 3) Comparison between Histone/Chaperone networks of resistant and sensitive samples to identify relationships characterizing resistance to treatment.

Data Normalization :

Before studying and analyzing the datasets, we normalized the counts using the Trimmed Mean of the M-values (TMM) method and log transformation, ensuring that the variance in expression is no longer correlated with the average expression of the gene, a necessary assumption for applying most of the statistical tests. Tmm was performed using the conorm package in Python.

Geneset definitions:

We defined highly variable genes (HVGs) by mean expression above 7 and standard deviation above 1.2. The set of histone variants and chaperones is based on our current knowledge as exposed in the introduction. When analyzing GTEx and TCGA datasets, replicative variants were excluded as they are not polyadenylated.

Principal component analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP):

The first step of the analysis was exploring the distribution of the samples in the space of the histone variant and chaperone genes using PCA and UMAP. PCA and UMAP are used because of the huge amount of samples with known classification available for transcriptomic data and to exploit the power of the available variables in correctly classifying the samples.

Characterizing co-expression relationships, tissue signatures and alterations in cancer:

Pearson correlation is the approach commonly used for characterizing co-expression relationships between two genes. We applied it as a reference for comparison on the whole GTEx dataset to compare with our second approach and find modules of co-expressed genes.Since our aim is to look for changes in co-expression relationships between tissues, and between healthy and cancer tissues, we used the generalized linear model (GLM) which explains a variable by the linear combination of one or more variables. It thus allows the recovery of gene-to-gene correlation by assessing how the expression of a gene can explain the expression of another gene, and it can integrate other information as additional terms. Moreover, compared to Pearson correlation, GLM provides statistical tests for how significantly changed are the relationships. First, we built a simple model to compare with Pearson correlation and

check the logic behind this approach. In this model, we explained the expression of one gene by another gene for all possible pairs after scaling and centering the whole dataset. The aim of scaling and centring the data is to keep the gene coefficient between 1 and -1 so it is comparable to pearson correlation.

$$\lambda_y = \mu + \alpha_x + \varepsilon$$

λ_y : expression of y μ : mean expression of y

α_x : Gene coefficient of x ε : Error

Second, to integrate tissue information, we repeated the comparisons including a tissue and an interaction term after centering and scaling the data per tissue. We repeated the procedure with all possible tissues as a referent category. The interaction term is an indication of the change in the relationship between gene x and y in the tissue t compared to the reference tissue.

$$\lambda_y = \cancel{\mu} + \alpha_x + \cancel{\beta_t} + (\alpha\beta)_{x,t} + \varepsilon$$

β_t : tissue term. $(\alpha\beta)_{x,t}$: gene - tissue interaction term.

Scaling and centring per tissue allows to eliminate the basal expression level and the difference between tissues in mean expressions, focusing the model on the change in co-expression relationships. This step results in the canceling of mean expression and tissue term as the centering is done around zero.

Finally, to compare between the healthy and cancerous samples, we repeated the analysis for each tissue by setting the healthy samples as the reference and comparing TCGA samples to it.

$$\lambda_y = \cancel{\mu} + \alpha_x + \cancel{\beta_d} + (\alpha\beta)_{x,d} + \varepsilon$$

β_d : dataset term. $(\alpha\beta)_{x,d}$: gene - dataset interaction term.

To find relationships that are significantly reversed we took into consideration both the gene and interaction coefficients. After removing the coefficients with insignificant p-value for any or both terms, we looked for gene pairs with interaction coefficient away from zero by being above 0.1. Lastly, we considered an absolute difference between gene and interaction coefficients that is above 0.8 to be a significant change in a co-expression relationship.

Results:

Datasets Characterization:

After normalizing the data from healthy tissues (GTEx), I looked at the distribution of the genes based on their mean of expression and standard deviation in healthy tissues (Figure 5). Based on mean expression, the genes follow a bimodal distribution with a distinction between the two modes at 7. Based on standard deviation, the genes follow a unimodal distribution with an extended shoulder for values higher than 1.2.(Figure 5). Therefore, I defined highly variable genes (HVGs) by having a high mean expression and an extended standard deviation. Focusing on histone variants and chaperone genes, we observe that they vary in their mean expression levels but most of them have a low standard deviation. Cancer tissues (TCGA) have similar distribution patterns. However, several histone variants and chaperone genes turned to be more expressed and with higher standard deviation compared to healthy tissues (NPM2, ASF1B, HJURP, MCM2, H2AJ, MACROH2A2, CENPA), indicating some distinctive alterations in cancer (Figure 5).

Histone variants and chaperones power in distinguishing tissues:

I performed PCA on both the highly variable and the histone/chaperone genes. In the first case, starting with 4540 genes, 214 PCs explained 95% of the variance. Based on the 50 histones variants/chaperones genes only, 31 components summarize the same percentage of variance. In both cases, visualizing PCA components revealed the separation of a few tissues based on one or two PCs, for example: testis, whole blood, brain, cerebellum and muscle (Figure 6).

To assess the maximum power of the datasets in segregating the samples by their tissue of origin, I applied UMAP. In the visualization of the The UMAP based on the HVGs-derived genes, all the tissues were completely separated.

On the other hand, many tissues, but not all of them, were segregated in the visualization of the UMAP components based on the histone variants and chaperones gene expression. While histone variants and chaperones are mostly ubiquitous, putting aside brain- and testis-specific genes, we found unexpected tissue signatures based on these genes. The tissues that were not separated are mostly proliferative ones and

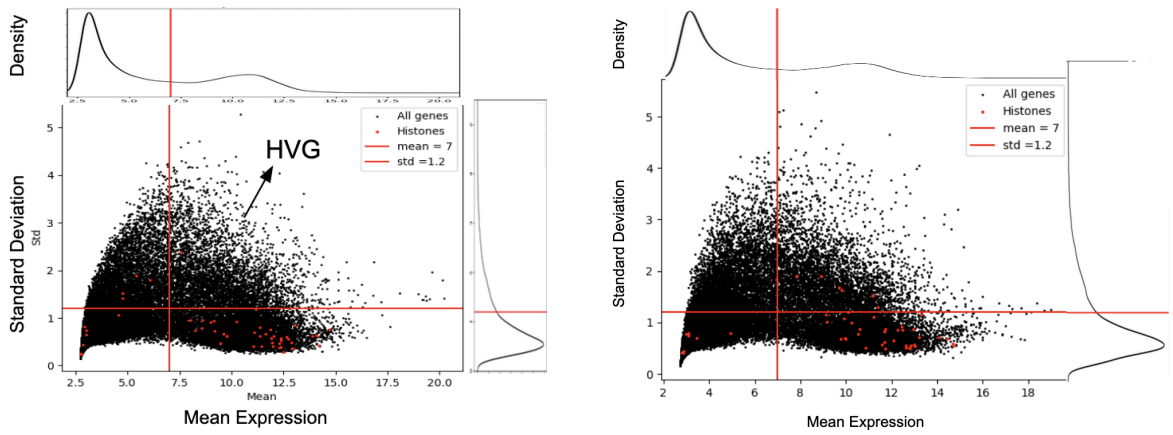


Figure 5 - Mean and standard deviation of expression for genes in GTEx (left) and TCGA (right) datasets.

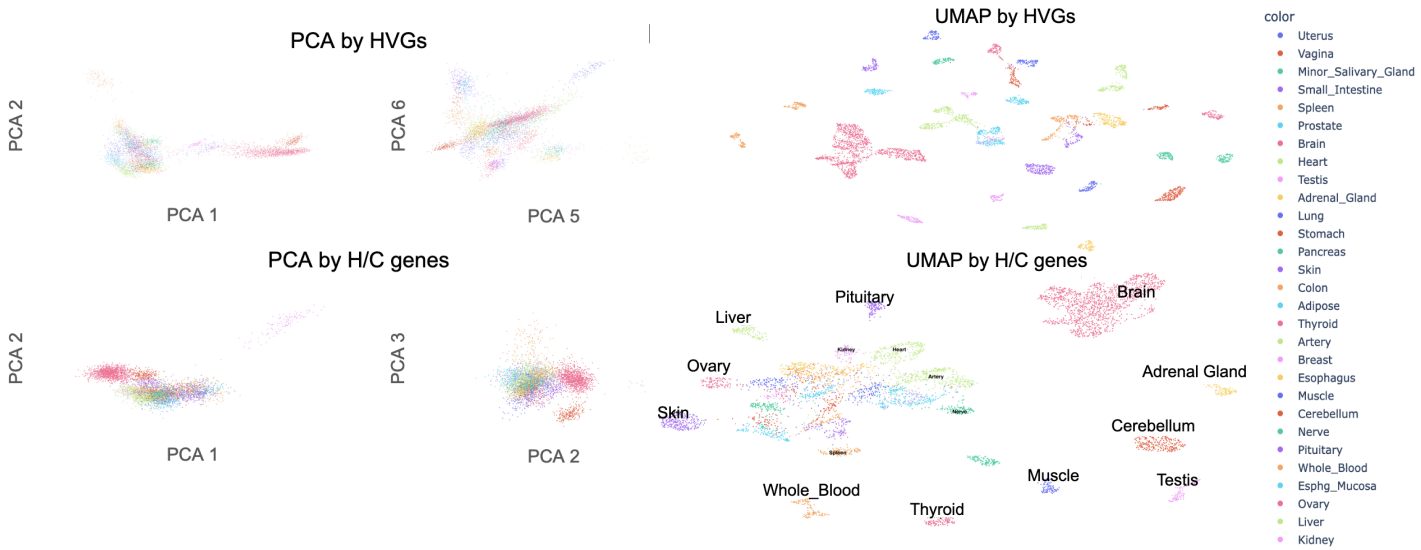


Figure 6 - Left Visualization of some PCA components of GTEx dataset using highly variable genes (top), or histone/chaperone gene (bottom).
Figure 6 - Right Visualization of UMAP components of GTEx dataset using highly variable genes (top), or histone/chaperone gene (bottom).

contain squamous cells like the stomach and esophagus. While the tissues that PCA distinguished, in addition to glands, formed separate clusters (Figure 6).

Histone variants and chaperones power in distinguishing cancer and healthy tissue:

I performed PCA on both the highly variable genes and the histone/chaperone genes. The general trend for component analysis is that the different types of cancer are harder to segregate as they become all highly proliferative. This is reflected in the amount of genes that is needed to explain the variation and in the visualization. In the first case, 820 genes out of 4584 explained 95% of the variance. For histones variants/chaperones genes 36/50 components to explain 95% of the variance in the data. In both cases, visualizing PCA components revealed the separation of a few cancer types based on one or two PCs, but less than those with GTEx (Figure 7).
In the visualization of the UMAP components of the highly variable genes all the cancers were separated as expected. On the other hand, in the visualization of the UMAP components of the histone variants and chaperones genes few tissues were separated compared to GTEx which is in accordance with the proliferative nature of the cancerous tissues, a feature that we observed associated with poor segregation of the normal tissues. Well defined cancer types are LAML (Acute Myeloid Leukemia), LGG (Brain Lower Grade Glioma), THCA (Thyroid carcinoma), PRAD (Prostate adenocarcinoma), UVM (Uveal Melanoma), THYM (Thymoma), and TGCT (Testicular Germ Cell Tumors) (Figure 7).

Histone variants and chaperones expression based tissue signatures:

Moving on with the histone variants and chaperones genes only, I applied Pearson correlation to the GTEX dataset. Looking at the results of pearson correlation in the heatmap (Figure 8) for each pair of genes, we notice variations as some are positively correlated, others are anticorrelated, or not correlated at all.

By applying GLM on the whole dataset to explain the expression of one gene by another for all pairs of genes we got gene coefficients that are identical to the Pearson correlation. The only exception for that is the diagonal as in pearson correlation a gene is completely correlated with itself while in GLM a gene can not be explained by itself (Figure 8). I scaled and centered the dataset to normalize the effect of mean

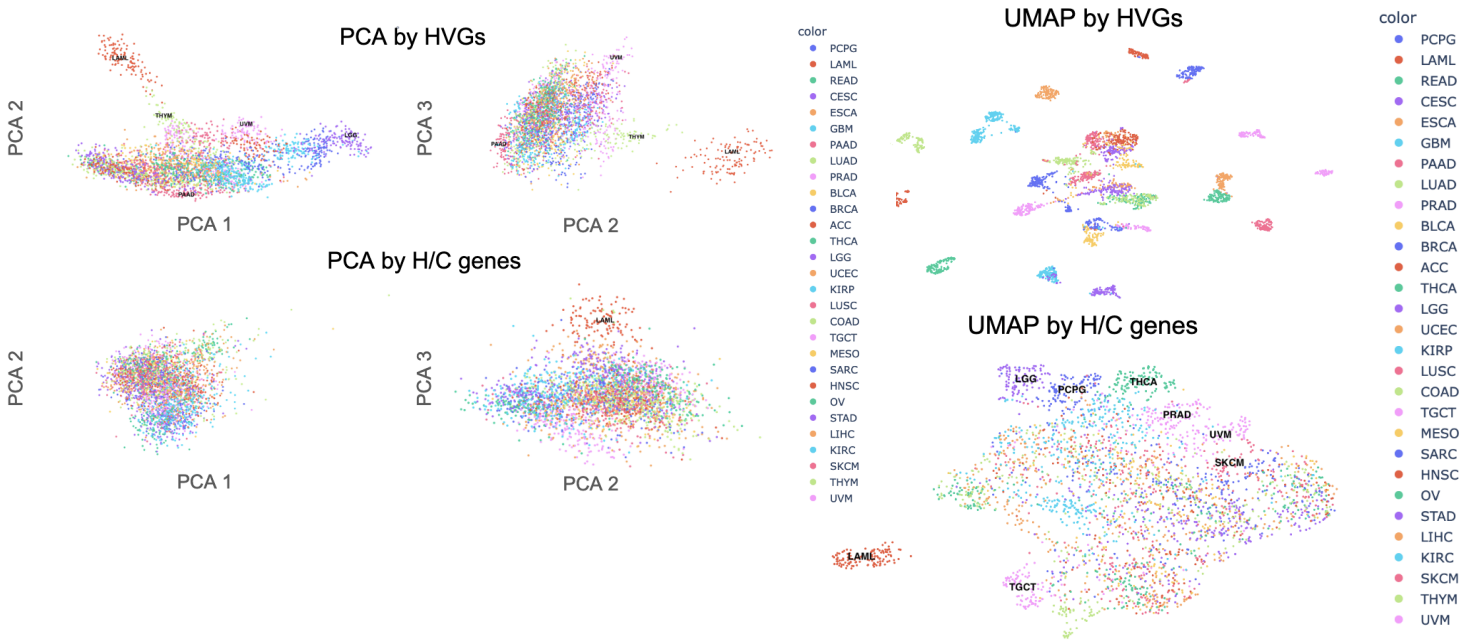


Figure 7 - Left Visualization of some PCA components of TCGA dataset using highly variable genes (top), or histone/chaperone gene (bottom). **Right** Visualization of UMAP components of TCGA dataset using highly variable genes (top), or histone/chaperone gene (bottom).

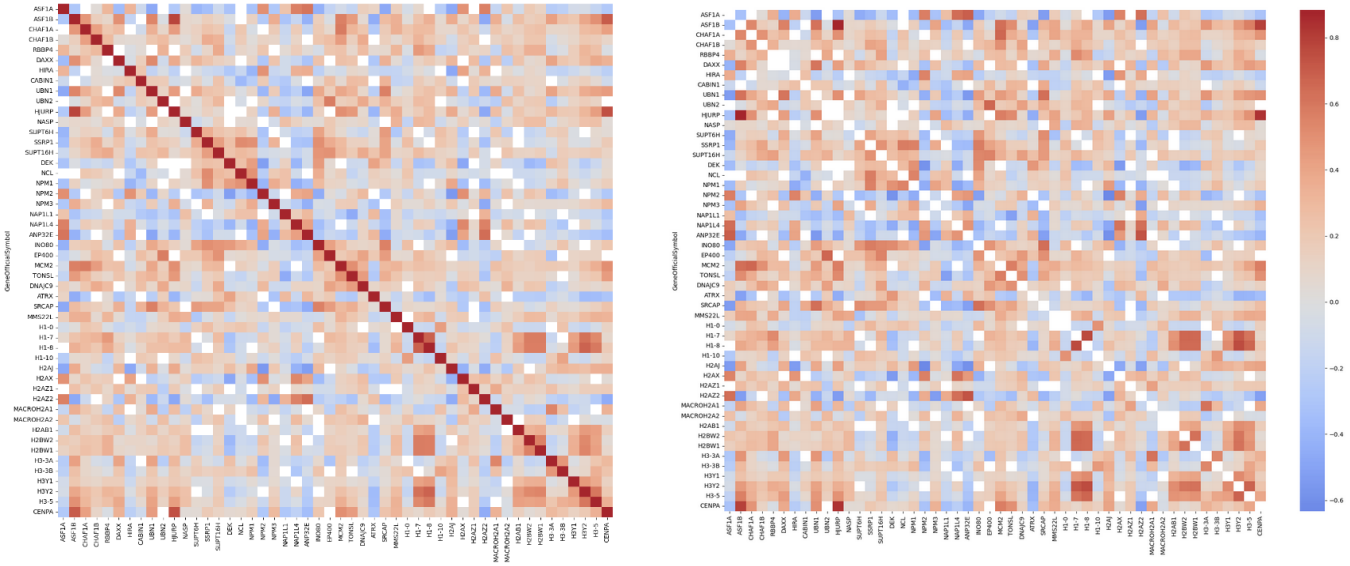


Figure 8 - Heatmap of the Pearson correlation (left) and GLM basic model gene coefficients (right) for all pairs of H/C genes on the whole set of GTEx samples.

expression and limit the variation to the standard deviation. In the clustermap of the results of Pearson correlation and GLM basic model, we can see the formation of biologically logical groups of genes as they exhibit similar expression patterns (Figure 9). For example, the testis specific genes, H3-5, H1-7, H1-8, H2BW1, and H2BW2 were in one cluster.

Co-expression relationships in tissue context:

Some tissues exhibit strong trends when set as a referent category and compared with others, either negatively or positively as seen in the heatmap of the interaction terms in the advanced model. For instance, testis and cerebellum had mostly negative coefficients while whole blood and brain had positive ones. These tissues were the ones separated in PCA and UMAP. On the other hand, some tissues had mixed coefficients, especially the ones that did not separate in component analysis, like Esophagus Mucosa (Figure 10). It is important to keep in mind that the interaction term represents solely the change in relationship between the two genes and is not affected by the level of expression. Besides, we can notice that the the pairs involving the testis specific genes were always insignificant except for when we are comparing other tissues with testis as they are only expressed there. Finally, pairs including NPM2 have the strongest and most variable interaction coefficients as it is the only highly variable gene between the histone variants and chaperones.

Looking at the histogram of the absolute differences between gene and interaction coefficients, we notice that most instances are close to zero and only few of them have large differences (Figure 11). Figure 11 presents two examples of co-expression relationships for a gene pair with a low difference and for another with a high difference alongside the coefficient values.

Histone variants and chaperones expression network alterations in cancer:

Looking at the interaction terms heatmap of the model comparing cancerous and healthy tissues of the same kind, we notice that some tissues exhibit some general strong trends, either negatively or positively. For instance, skin had mostly negative coefficients while colon had positive ones (Figure 12). Also, the pairs containing testis specific genes are mostly insignificant, and the pairs with highly variable genes have strong coefficients.

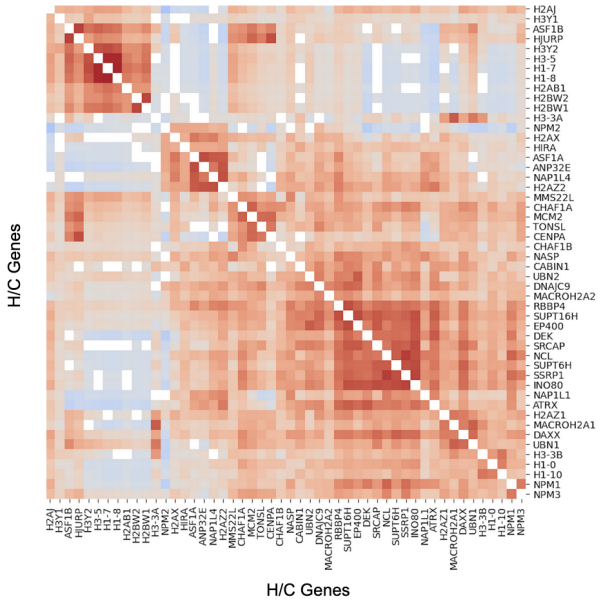


Figure 9 - Clustermap of the genes based on the gene coefficients estimated from the basic model applied on all pairs of H/C genes in GTEx.

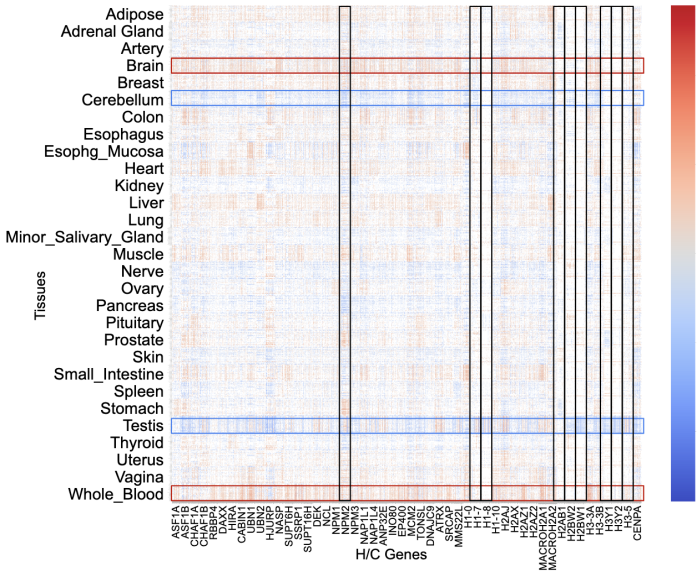
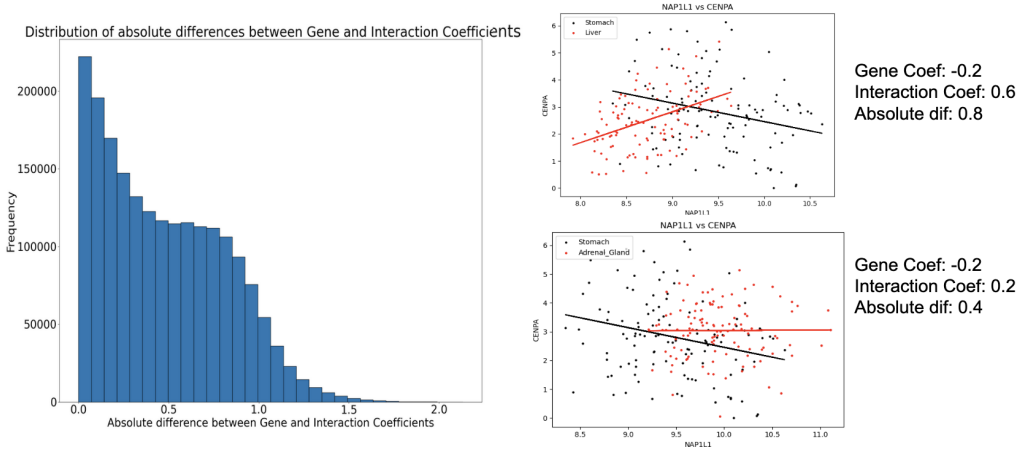


Figure 10 - Heatmap of the interaction coefficients for every pair of H/C genes estimated on healthy tissue samples. Tissues and genes with a strong pattern are highlighted. Cerebellum and testis with negative coefficients, brain and whole blood with positive ones. Testis specific genes (H3-5, H3Y2, H3Y1, H2BW1, H2BW2, H2AB1, H1-7, H1-8) with insignificant coefficients are marked in black. In green, we see highly variable genes.



Gene Coef: -0.2
Interaction Coef: 0.6
Absolute dif: 0.8

Gene Coef: -0.2
Interaction Coef: 0.2
Absolute dif: 0.4

Figure 11 - Histogram of the difference between gene and interaction coefficients of GLM advanced model in GTEX (left) and illustration of two different examples with their values (right).

Looking at the histogram of the absolute differences between gene and interaction coefficients, we notice that most instances are close to zero and only few of them have large differences (Figure 13). Figure 13 presents two examples of relationships for a gene pair with a low difference and for another with a high difference.

The altered relationships among healthy tissues are usually different in one tissue compared to all the others. For example, CENPA and H3-3A are only anti-correlated in whole blood but are positively correlated in all other tissues. The same applies for H2AJ and ATRX in testis.

In cancer, CENPA and H3-3A relationship is only different in whole blood where it becomes positively correlated as in the normal state of all other tissues. For H2AJ and ATRX, the relationship is reversed in cancer for testis and it becomes positively correlated as the normal state for all the other tissues. At the same time, in liver, skin, thyroid, kidney, pancreas, and minor salivary gland, these two genes become anti-correlated in cancer (Figure 14).

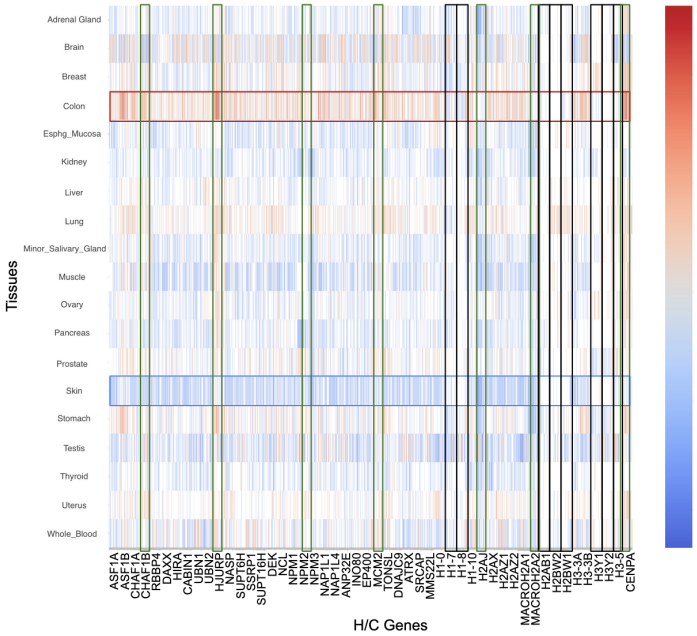


Figure 12 - Heatmap of the interaction coefficients of all pairs of H/C genes in cancer compared to healthy tissues. Tissues and genes with a strong pattern are highlighted. Skin with negative coefficients, colon with positive ones. Testis specific genes (H3-5, H3Y2, H3Y1, H2BW1, H2BW2, H2AB1, H1-7, H1-8) with insignificant coefficients are marked in black. In green, we see highly variable genes.

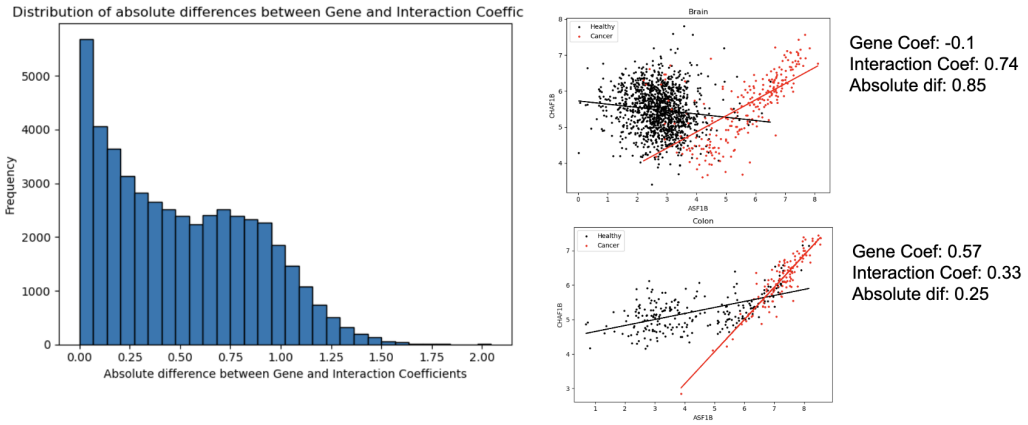


Figure 13 - Histogram of the difference between gene and interaction coefficients of GLM advanced model between GTEX and TCGA (left) and expression relationships of ASF1B and CHAF1B between healthy and cancer samples in brain and colon (right).

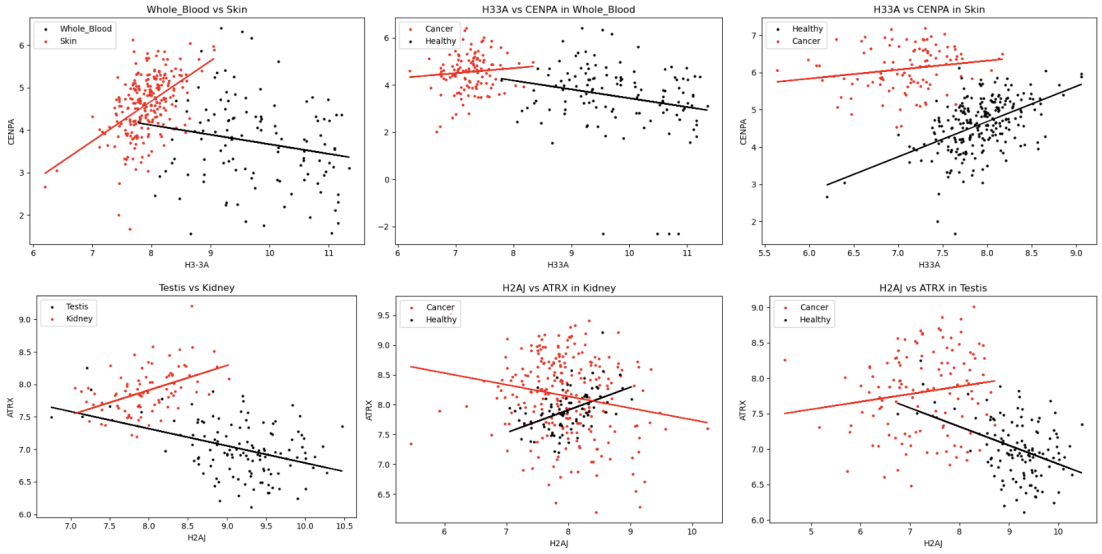


Figure 14 - Co-expression relationships comparison between different healthy tissues and against cancerous equivalents. Top, H33A vs CENPA in whole blood and skin. Bottom, H2AJ and ATRX in testis and kidney.

Discussion:

In the main results, we see from the main and standard deviation distribution that only one gene from the histone variants and chaperones is among the HVGs, nucleophosmin 2 (NPM2). One question that may be asked is whether NPM2, given that it is the only highly variable gene in GTEx, and tissue specific genes are the elements effective in the analysis and are biasing the results. But, looking at the PCA results, we see that 31 PCs were required to explain 95% variation in the data, which is much higher than the number of these genes that sums to 10.

On the other hand, in the TCGA dataset, more genes are among the highly variable and the mean expression is generally higher. The list includes in addition to NPM2, ASF1B, HJURP, MCM2, H2AJ, MACROH2A2, CENPA. Some of these genes had high standard deviation but low mean expression in GTEx (CENPA, HJURP, ASF1B). This overexpression is consistent with what is known in literature. As mentioned in the introduction, CENPA and HJURP overexpression is considered as an unfavorable prognostic marker in cancer (reviewed in Mahlke et ., 2020). Also, a pan-analysis revealed that ASF1B was highly expressed in 22 types of cancers (Hu et al., 2021) and another pan analysis showed that MCM2 is over-expressed in all cancer types in the human cancer atlas (Yuan et al., 2022), MCM2 expression in cancer in addition to its standard deviation. For H2AJ, there was a slight increase in its mean expression and standard deviation since it was so close to being among HVG in healthy state. However, MACROH2A2 is usually underexpressed in cancer (Hsu et al., 2021) contrary to what we found here. Its mean expression is over 7 for all types of cancer in our dataset although it is fluctuating (Table 2). Besides, the standard deviation dramatically increased from 0.9 to 1.6. Thus, the expression of MACROH2A2 in cancer is an interesting project for a future study in the lab.

By comparing the results of component analysis between PCA and UMAP, we notice that UMAP is a much more effective method in spotting the variation. We could see by looking at the principal components the maximum power of the datasets in segregating the samples according to their tissue of origin. The fact that many healthy tissues were separated based on histone and chaperone genes although they are not

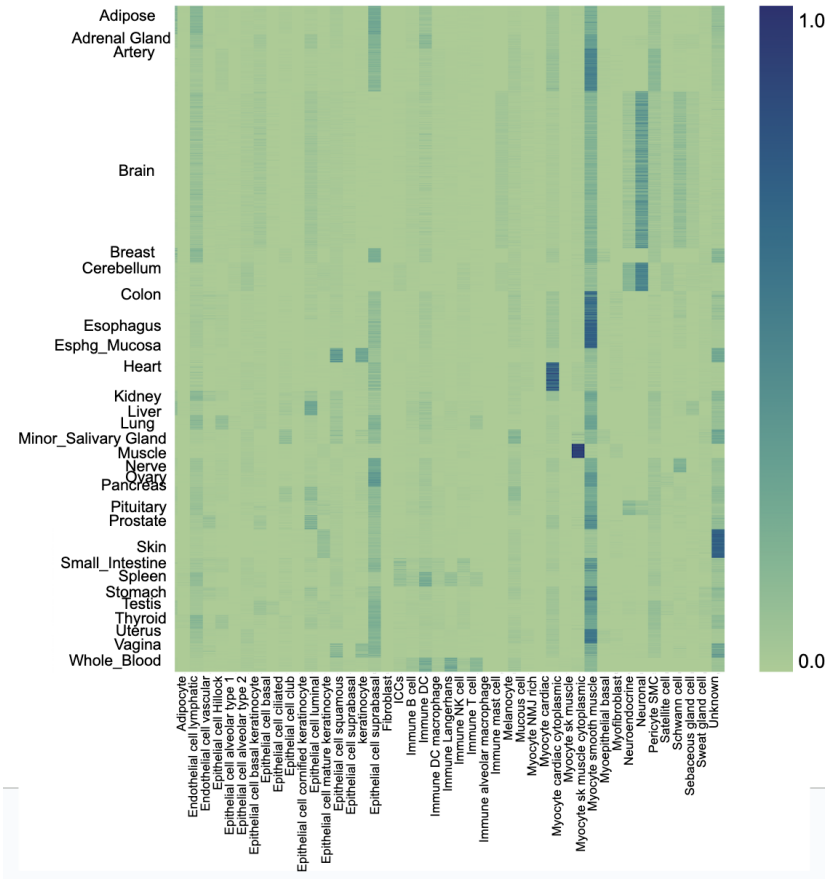


Figure 15 - Cell type composition for each tissue obtained by deconvolution, some tissues have more samples than the others as they have more subtypes. Cell types were based on the PanglaoDB database.

TissueType	Mean Expression	Cancer Type	Corresponding tissue	Mean Expression
Adipose	7.685272	ACC	Adrenal Gland	9.673007
Adrenal_Gland	7.175578	BLCA		10.810447
Artery	8.404717	BRCA	Breast	9.811624
Brain	9.328495	CESC	Minor Salivary Gland	10.260203
Breast	8.585692	COAD	Colon	9.806086
Cerebellum	10.39844	ESCA	Esphg_Mucosa	9.08256
Colon	9.575129	GBM	Brain	8.393037
Esophagus	9.512312	HNSC		10.514447
Esphg_Mucosa	10.343053	KIRC	Kidney	8.86847
Heart	9.641648	KIRP	Kidney	10.194057
Kidney	9.606434	LAML	Whole_Blood	7.351689
Liver	8.462403	LGG	Brain	10.476401
Lung	8.57616	LIHC	Liver	9.079627
Minor_Salivary_Gland	9.082192	LUAD	Lung	8.881784
Muscle	8.58704	LUSC	Lung	10.356909
Nerve	7.743745	MESO		9.747885
Ovary	9.828336	OV	Ovary	10.903965
Pancreas	9.143332	PAAD	Pancreas	10.140421
Pituitary	10.130475	PCPG		10.587973
Prostate	9.892873	PRAD	Prostate	10.606387
Skin	9.377733	READ		9.930543
Small_Intestine	9.289617	SARC	Muscle	9.609765
Spleen	7.591299	SKCM	Skin	7.675843
Stomach	9.262787	STAD	Stomach	8.481409
Testis	10.096377	TGCT	Testis	11.041733
Thyroid	9.175299	THCA	Thyroid	10.440831
Uterus	10.007777	THYM		10.362256
Vagina	9.787382	UCEC	Uterus	11.46563
Whole_Blood	7.207715	UVM		9.176935

Table 2 - Mean expression of MACROH2A2 in different types of tissues and cancers, the expression was TMM normalized then log transformed.

among the highly variable genes and present in every cell supports the idea of a tissue signature based on the histone variants and chaperones network. The tissues that were not separated and are merged into one cloud are mostly proliferative tissues.

As for the cancerous samples, we were not able to reach good results in separating all the cancer types as with the healthy counterparts. This may be due to the fact that cancerous tissues are highly proliferative.

The similarity between the results of Pearson correlation and GLM basic model ensures that the approach used is sound and consistent with widely used statistical tests. The groups of genes based on GLM's gene coefficients (or Pearson correlation) correspond to biological function and can be used as independent sub-networks to be investigated or to simplify the whole network of genes into a network of modules. In addition, the results of the GLM model that integrates tissue information, and the variation in interaction terms further supports the hypothesis that different tissues have different co-expression relationships among histone variants and chaperons. The approach we used to find altered relationships is reusable and can be applied in different cases as seen in the model comparing healthy and cancerous tissues. In future steps, it will be applied to resistant and sensitive samples.

Lastly, by looking at the results of the final model comparing healthy and cancer samples from the same tissue type, we see that many of the altered relationships are shared among all types of tissues while very few of them are specific to one or two origins. Since we see in the UMAP that the tissues are harder to separate in cancer samples, we expect that the samples are moving from more distinct expression patterns to become more similar as they all have some shared properties, mainly that they are highly proliferative and as the tumor progresses, they resemble more embryonic cells.

The reason we choose bulk RNA-seq data instead of single cell RNA (scRNA) is that we are looking for quantitative variations in gene expression in every sample. Single cell RNA is commonly used for studying heterogeneity and profiling cell types. Since the size of the starting material is small, it is less sensitive in detecting lowly or intermediary expressed genes as well as marginal differences in expression (Li et al.,

2021). The large number of samples available in GTEx and TCGA provide enough statistical power for the analysis that is comparable with the number of cells in scRNA-seq experiments. However, the RNA sequencing procedures for TCGA and GTEx were oligo-dT-based so only polyadenylated (poly(A)) RNA transcripts were sequenced and replication histone genes, which are not polyA-matured, could not be included in our study.

As bulk RNA is more suitable to our study, we still need to consider whether the variations we see in the expression can be attributed to cell composition in each tissue. To account for this, we deconvoluted each tissue to its probable cell types and looking at the results in figure 15, we see that cell type composition of each tissue is mostly as expected. Comparing PCA components of samples by H/C genes and by cell composition did not show much correlation between the two. Thus, cell type composition is not enough to explain variation between tissues. The step of deconvolution was done using the Cibersort package in R.

Another issue that may arise in the analysis is the fact that we used two different datasets which can cause biases in the results of the analysis. To avoid this problem, we used Trimmed Mean of the M-values (TMM) normalization.

Conclusions and perspectives:

Conclusions:

Although the expression of histone variants and chaperone genes is lowly variable, and the histones are building blocks of the chromatin in every cell, they are unexpectedly effective in distinguishing samples based on their tissue of origin for many tissues. This indicates that some tissues have distinct expression patterns and signatures based on these genes.

In addition, the GLM is an effective approach in studying gene co-expression relationships. It is comparable to the more widely used Pearson correlation and allows the integration of additional information to the transcriptome analysis like the tissue origin, cancer status, or whether the sensitivity to treatment. Also, it gives insights about the changed relationships and interaction between mentioned factors. By clustering the gene coefficients resulting from the GLM model, we were able to find groups of genes that are correlated and exhibit similar expression patterns. The interaction coefficients of the advanced models provide a systematic way to identify co-expression relationships that are changed between different tissues, and those that are altered in cancer.

Perspectives:

First of all, our results need to be validated biologically with the expertise in the team at the RNA level (RT-qPCR) as well as at the protein level (Western Blot), and by having access to cancer samples being at Curie. Also, to ensure that they are not biased due to different datasets and procedures, we will use other public datasets, Expression atlas for instance, to see whether we will get the same results.

The next steps for the future include, first, integrating resistance information and samples in a new GLM model to study the rewiring of histone variants and chaperone networks in them (Figure 16). We will use patient-derived xenografts (PDXs) for this step. PDXs are models for cancer studies where cancer tissues derived from patients are implanted in immunodeficient mice (Abdolahi et al., 2022). They present a great opportunity because they allow us to perform experiments on a tumor still growing in a mammalian environment and obtain results with the highest reproducibility in humans.

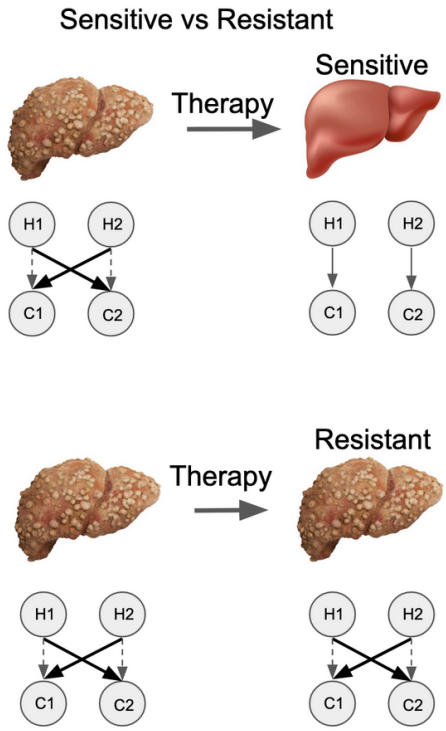


Figure 16 - Rewiring of histone variants and chaperones network in resistant tissues compared to sensitive ones. H1, H2: Histone variants 1 and 2. C1, C2: Chaperones 1 and 2. Arrows represent co-expression relationships. Bold and dashed arrows are acquired and lost relationships, respectively.

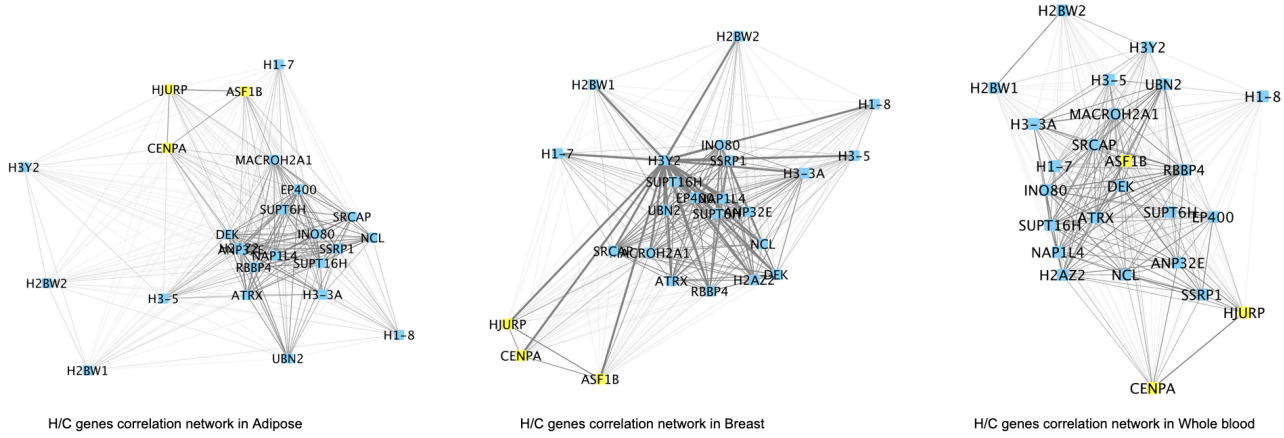


Figure 17 - Pearson correlation based networks in different tissues using Cytoscape. In yellow, the group of the three genes (CENPA, HJURP, ASF1B) whose expression is known to be correlated. In whole blood (Right), ASF1B did not cluster with the other two genes as opposed to Breast (Middle) and Adipose (Left).

Since the sample can be split and be subjected to different treatments, it also provides the means for having a proper control. From patients treated at the Institut Curie, Marangoni lab generated 80 PDXs of breast cancers (Coussy et al., 2019). They sequence total RNA from these PDXs and the same PDXs have been assessed for their sensitivity to three different chemotherapies (capecitabine, cisplatin/carboplatin, adriamycin + cyclophosphamide). Background information about patients' treatment history as well as the outcomes can be integrated. Since the number of samples is much smaller than the datasets used before, we will look at pairwise ratios of expression.

Also, it is important for us to move from gene pairs co-expression relationships to the network concept. One possible approach is to look at the modules of genes based on the gene coefficient and compare them between different situations. Besides, we can look at groups of variants, their dedicated chaperones, as well as chaperone subunits, which are already known as biologically related . Ultimately, we are looking for a way to investigate the whole network of histone variants and chaperones genes. One approach we tested before is using Cytoscape software, which builds co-expression networks based on correlation (Figure 17).

Whether the network of histone variants and chaperones would help to distinguish different types of cancers corresponding to the same healthy tissue can also be investigated. We notice that in the two types of cancer corresponding to kidney, kidney renal clear cell carcinoma (KIRC) and kidney renal papillary cell carcinoma (KIRP), the MACROH2A2 was underexpressed and overexpressed respectively compared to the healthy equivalent (Table 2). The same type of different pattern applies to the lung, as the mean expression in lung adenocarcinoma (LUAD) is close to the healthy lung samples while there is an overexpression in lung squamous cell carcinoma (LUSC). This in turn affects co-expression relationships.We see that if we compare MACROH2A2 with NASP in healthy kidney with all kidney related cancer samples, the correlation is entirely lost and we have a random relationship (Figure 18). If we look at each cancer subtype individually, we notice a slight anti-correlation in KIRC while a positive relationship in KIRP is reserved but still weaker than the healthy case.

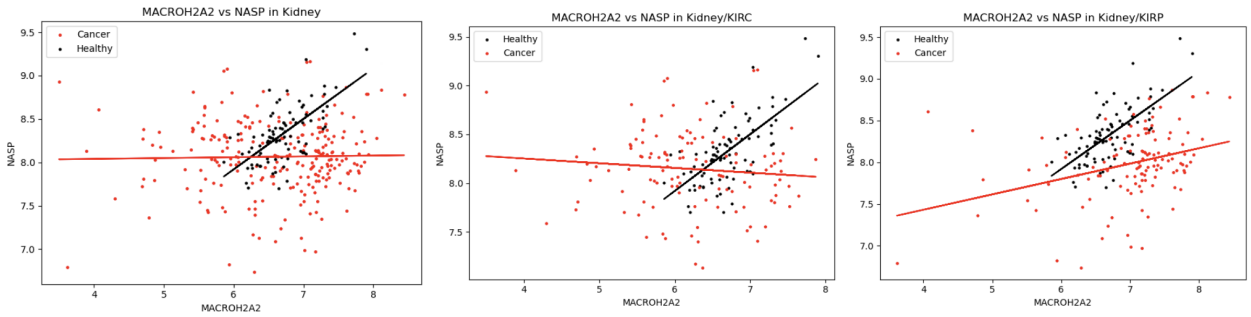


Figure 18 - Co-expression relationship between MACROH2A2 and NASP in healthy kidney samples against: all corresponding cancer samples (left), only KIRC (middle), and KIRP (right).

Bibliography

- Abascal F, Corpet A, Gurard-Levin ZA, Juan D, Ochsenbein F, Rico D, Valencia A, Almouzni G. Subfunctionalization via adaptive evolution influenced by genomic context: the case of histone chaperones ASF1a and ASF1b. Mol Biol Evol. 2013 Aug;30(8):1853-66. doi: 10.1093/molbev/mst086. Epub 2013 May 2. PMID: 23645555.
- Abdolahi S, Ghazvinian Z, Muhammadnejad S, Saleh M, Asadzadeh Aghdaei H, Baghaei K. Patient-derived xenograft (PDX) models, applications and challenges in cancer research. J Transl Med. 2022 May 10;20(1):206. doi: 10.1186/s12967-022-03405-8. PMID: 35538576; PMCID: PMC9088152.
- Bönisch C, Schneider K, Pünzeler S, Wiedemann SM, Bielmeier C, Bocola M, Eberl HC, Kuegel W, Neumann J, Kremmer E, Leonhardt H, Mann M, Michaelis J, Schermelleh L, Hake SB. H2A.Z.2.2 is an alternatively spliced histone H2A.Z variant that causes severe nucleosome destabilization. Nucleic Acids Res. 2012 Jul;40(13):5951-64. doi: 10.1093/nar/gks267. Epub 2012 Mar 29. PMID: 22467210; PMCID: PMC3401452.
- Buschbeck M, Hake SB. Variants of core histones and their roles in cell fate decisions, development and cancer. Nat Rev Mol Cell Biol. 2017 May;18(5):299-314. doi: 10.1038/nrm.2016.166. Epub 2017 Feb 1. PMID: 28144029.
- Chakravarthy S, Patel A, Bowman GD. The basic linker of macroH2A stabilizes DNA at the entry/exit site of the nucleosome. Nucleic Acids Res. 2012 Sep 1;40(17):8285-95. doi: 10.1093/nar/gks645. Epub 2012 Jun 29. PMID: 22753032; PMCID: PMC3458575.
- Chen D, Jin C. Histone variants in environmental-stress-induced DNA damage repair. Mutat Res Rev Mutat Res. 2019 Apr-Jun;780:55-60. doi: 10.1016/j.mrrev.2017.11.002. Epub 2017 Nov 21. PMID: 31395349; PMCID: PMC6690500.
- Coussy F, de Koning L, Lavigne M, Bernard V, Ouine B, Boulai A, El Botty R, Dahmani A, Montaudon E, Assayag F, Morisset L, Huguet L, Sourd L, Painsec P, Callens C, Chateau-Joubert S, Servely JL, Larcher T, Reyes C, Girard E, Pierron G, Laurent C, Vacher S, Baulande S, Melaabi S, Vincent-Salomon A, Gentien D, Dieras V, Bieche I, Marangoni E. A large collection of integrated genomically characterized patient-derived xenografts highlighting the heterogeneity of triple-negative breast cancer. Int J Cancer. 2019 Oct 1;145(7):1902-1912. doi: 10.1002/ijc.32266. Epub 2019 Apr 4. PMID: 30859564.
- Diaz-Papkovich A, Anderson-Trocmé L, Gravel S. A review of UMAP in population genetics. J Hum Genet. 2021 Jan;66(1):85-91. doi: 10.1038/s10038-020-00851-4. Epub 2020 Oct 14. PMID: 33057159; PMCID: PMC7728596.
- Dunleavy EM, Roche D, Tagami H, Lacoste N, Ray-Gallet D, Nakamura Y, Daigo Y, Nakatani Y, Almouzni-Pettinotti G. HJURP is a cell-cycle-dependent maintenance and deposition factor of CENP-A at centromeres. Cell. 2009 May 1;137(3):485-97. doi: 10.1016/j.cell.2009.02.040. PMID: 19410545.
- eGTEx Project. Enhancing GTEx by bridging the gaps between genotype, gene expression, and disease. Nat Genet. 2017 Dec;49(12):1664-1670. doi: 10.1038/ng.3969. Epub 2017 Oct 11. PMID: 29019975; PMCID: PMC6636856.
- Gadad SS, Senapati P, Syed SH, Rajan RE, Shandilya J, Swaminathan V, Chatterjee S, Colombo E, Dimitrov S, Pelicci PG, Ranga U, Kundu TK. The multifunctional protein nucleophosmin (NPM1) is a human linker histone H1 chaperone. Biochemistry. 2011 Apr 12;50(14):2780-9. doi: 10.1021/bi101835j. Epub 2011 Mar 22. PMID: 21425800.
- Ghiraldini FG, Filipescu D, Bernstein E. Solid tumors hijack the histone variant network. Nat Rev Cancer. 2021 Apr;21(4):257-275. doi: 10.1038/s41568-020-00330-0. Epub 2021 Feb 10. PMID: 33568791; PMCID: ,PMC8092022.
- Goldberg AD, Banaszynski LA, Noh KM, Lewis PW, Elsaesser SJ, Stadler S, Dewell S, Law M, Guo X, Li X, Wen D, Chapgier A, DeKelver RC, Miller JC, Lee YL, Boydston EA, Holmes MC, Gregory PD, Greally JM, Rafii S, Yang C, Scambler PJ, Garrick D, Gibbons RJ, Higgs DR, Cristea IM, Urnov FD, Zheng D, Allis CD. Distinct factors control histone variant H3.3 localization at specific genomic regions. Cell. 2010 Mar 5;140(5):678-91. doi: 10.1016/j.cell.2010.01.003. PMID: 20211137; PMCID: PMC2885838.
- Greaves IK, Rangasamy D, Ridgway P, Tremethick DJ. H2A.Z contributes to the unique 3D structure of the centromere. Proc Natl Acad Sci U S A. 2007 Jan 9;104(2):525-30. doi: 10.1073/pnas.0607870104. Epub 2006 Dec 28. PMID: 17194760; PMCID: PMC1766418.
- GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015 May 8;348(6235):648-60. doi: 10.1126/science.1262110. Epub 2015 May 7. PMID: 25954001; PMCID: PMC4547484.
- Guberovic I, Farkas M, Corujo D, Buschbeck M. Evolution, structure and function of divergent macroH2A1 splice isoforms. Semin Cell Dev Biol. 2023 Feb 15;135:43-49. doi: 10.1016/j.semcdb.2022.03.036. Epub 2022 Apr 11. PMID: 35422391.

- Hoang SM, Kaminski N, Bhargava R, Barroso-González J, Lynskey ML, García-Expósito L, Roncaioli JL, Wondisford AR, Wallace CT, Watkins SC, James DI, Waddell ID, Ogilvie D, Smith KM, da Veiga Leprevost F, Mellacharevu D, Nesvizhskii AI, Li J, Ray-Gallet D, Sobol RW, Almouzni G, O'Sullivan RJ. Regulation of ALT-associated homology-directed repair by polyADP-ribosylation. Nat Struct Mol Biol. 2020 Dec;27(12):1152-1164. doi: 10.1038/s41594-020-0512-7. Epub 2020 Oct 12. PMID: 33046907; PMCID: PMC7809635.
- Hsu CJ, Meers O, Buschbeck M, Heidel FH. The Role of MacroH2A Histone Variants in Cancer. Cancers (Basel). 2021 Jun 15;13(12):3003. doi: 10.3390/cancers13123003. PMID: 34203934; PMCID: PMC8232725.
- Huang T, Yuan S, Gao L, Li M, Yu X, Zhan J, Yin Y, Liu C, Zhang C, Lu G, Li W, Liu J, Chen ZJ, Liu H. The histone modification reader ZCWPW1 links histone methylation to PRDM9-induced double-strand break repair. Elife. 2020 May 6;9:e53459. doi: 10.7554/eLife.53459. PMID: 32374261; PMCID: PMC7314539.
- Hu X, Zhu H, Zhang X, He X, Xu X. Comprehensive analysis of pan-cancer reveals potential of ASF1B as a prognostic and immunological biomarker. Cancer Med. 2021 Oct;10(19):6897-6916. doi: 10.1002/cam4.4203. Epub 2021 Sep 2. PMID: 34472711; PMCID: PMC8495294.
- Jeffery D, Gatto A, Podsypanina K, Renaud-Pageot C, Ponce Landete R, Bonneville L, Dumont M, Fachinetti D, Almouzni G. CENP-A overexpression promotes distinct fates in human cells, depending on p53 status. Commun Biol. 2021 Mar 26;4(1):417. doi: 10.1038/s42003-021-01941-5. PMID: 33772115; PMCID: PMC7997993.
- Kanwal R, Gupta K, Gupta S. Cancer epigenetics: an introduction. Methods Mol Biol. 2015;1238:3-25. doi: 10.1007/978-1-4939-1804-1_1. PMID: 25421652.
- Kawamura M, Funaya S, Sugie K, Suzuki MG, Aoki F. Asymmetrical deposition and modification of histone H3 variants are essential for zygote development. Life Sci Alliance. 2021 Jun 24;4(8):e202101102. doi: 10.26508/lsa.202101102. PMID: 34168076; PMCID: PMC8321678.
- Ku M, Jaffe JD, Koche RP, Rheinbay E, Endoh M, Koseki H, Carr SA, Bernstein BE. H2A.Z landscapes and dual modifications in pluripotent and multipotent stem cells underlie complex genome regulatory functions. Genome Biol. 2012 Oct 3;13(10):R85. doi: 10.1186/gb-2012-13-10-r85. PMID: 23034477; PMCID: PMC3491413.
- Lacoste N, Woolfe A, Tachiwana H, Garea AV, Barth T, Cantaloube S, Kurumizaka H, Imhof A, Almouzni G. Mislocalization of the centromeric histone variant CenH3/CENP-A in human cells depends on the chaperone DAXX. Mol Cell. 2014 Feb 20;53(4):631-44. doi: 10.1016/j.molcel.2014.01.018. Epub 2014 Feb 13. PMID: 24530302.
- Latreille D, Bluy L, Benkirane M, Kiernan RE. Identification of histone 3 variant 2 interacting factors. Nucleic Acids Res. 2014 Apr;42(6):3542-50. doi: 10.1093/nar/gkt1355. Epub 2014 Jan 6. PMID: 24393775; PMCID: PMC3973350.
- Li X, Wang CY. From bulk, single-cell to spatial RNA sequencing. Int J Oral Sci. 2021 Nov 15;13(1):36. doi: 10.1038/s41368-021-00146-0. PMID: 34782601; PMCID: PMC8593179.
- Long M, Sun X, Shi W, Yanru A, Leung STC, Ding D, Cheema MS, MacPherson N, Nelson CJ, Ausio J, Yan Y, Ishibashi T. A novel histone H4 variant H4G regulates rDNA transcription in breast cancer. Nucleic Acids Res. 2019 Sep 19;47(16):8399-8409. doi: 10.1093/nar/gkz547. PMID: 31219579; PMCID: PMC6895281.
- Mahlke MA, Nechemia-Arbely Y. Guarding the Genome: CENP-A-Chromatin in Health and Cancer. Genes (Basel). 2020 Jul 16;11(7):810. doi: 10.3390/genes11070810. PMID: 32708729; PMCID: PMC7397030.
- Martire S, Banaszynski LA. The roles of histone variants in fine-tuning chromatin organization and function. Nat Rev Mol Cell Biol. 2020 Sep;21(9):522-541. doi: 10.1038/s41580-020-0262-8. Epub 2020 Jul 14. PMID: 32665685; PMCID: PMC8245300.
- Marzluff WF. Metazoan replication-dependent histone mRNAs: a distinct set of RNA polymerase II transcripts. Curr Opin Cell Biol. 2005 Jun;17(3):274-80. doi: 10.1016/j.ceb.2005.04.010. PMID: 15901497.
- Matsuda R, Hori T, Kitamura H, Takeuchi K, Fukagawa T, Harata M. Identification and characterization of the two isoforms of the vertebrate H2A.Z histone variant. Nucleic Acids Res. 2010 Jul;38(13):4263-73. doi: 10.1093/nar/gkq171. Epub 2010 Mar 18. PMID: 20299344; PMCID: PMC2910051.
- 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
- Melé M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, Young TR, Goldmann JM, Pervouchine DD, Sullivan TJ, Johnson R, Segrè AV, Djebali S, Niarchou A; GTEx Consortium; Wright FA, Lappalainen T, Calvo M, Getz G, Dermitzakis ET, Ardlie KG, Guigó R. Human genomics. The human transcriptome across tissues and individuals. Science. 2015 May 8;348(6235):660-5. doi: 10.1126/science.aaa0355. PMID: 25954002; PMCID: PMC4547472.
- Mendiratta S, Gatto A, Almouzni G. Histone supply: Multitiered regulation ensures chromatin dynamics throughout the cell cycle. J Cell Biol. 2019 Jan 7;218(1):39-54. doi: 10.1083/jcb.201807179. Epub 2018 Sep 26. PMID: 30257851; PMCID: PMC6314538.

- Montalvo-Casimiro M, González-Barrios R, Meraz-Rodriguez MA, Juárez-González VT, Arriaga-Canon C, Herrera LA. Epidrug Repurposing: Discovering New Faces of Old Acquaintances in Cancer Therapy. Front Oncol. 2020 Nov 18;10:605386. doi: 10.3389/fonc.2020.605386. PMID: 33312959; PMCID: PMC7708379.
- Morel D, Jeffery D, Aspeslagh S, Almouzni G, Postel-Vinay S. Combining epigenetic drugs with other therapies for solid tumours - past lessons and future promise. Nat Rev Clin Oncol. 2020 Feb;17(2):91-107. doi: 10.1038/s41571-019-0267-4. Epub 2019 Sep 30. PMID: 31570827.
- Nye J, Sturgill D, Athwal R, Dalal Y. HJURP antagonizes CENP-A mislocalization driven by the H3.3 chaperones HIRA and DAXX. PLoS One. 2018 Oct 26;13(10):e0205948. doi: 10.1371/journal.pone.0205948. Erratum in: PLoS One. 2018 Nov 12;13(11):e0207631. PMID: 30365520; PMCID: PMC6203356.
- Odhiambo Omuya, E., Onyango Okeyo, G. and Waema Kimwele, M. (2021) "Feature selection for classification using Principal Component Analysis and Information Gain," Expert Systems with Applications, 174, p. 114765. Available at: https://doi.org/10.1016/j.eswa.2021.114765.
- Pan C, Fan Y. Role of H1 linker histones in mammalian development and stem cell differentiation. Biochim Biophys Acta. 2016 Mar;1859(3):496-509. doi: 10.1016/j.bbagrm.2015.12.002. Epub 2015 Dec 13. PMID: 26689747; PMCID: PMC4775330.
- Piquet S, Le Parc F, Bai SK, Chevallier O, Adam S, Polo SE. The Histone Chaperone FACT Coordinates H2A.X-Dependent Signaling and Repair of DNA Damage. Mol Cell. 2018 Dec 6;72(5):888-901.e7. doi: 10.1016/j.molcel.2018.09.010. Epub 2018 Oct 18. PMID: 30344095; PMCID: PMC6292839.
- Renaud-Pageot C, Quivy JP, Lochhead M, Almouzni G. CENP-A Regulation and Cancer. Front Cell Dev Biol. 2022 Jun 2;10:907120. doi: 10.3389/fcell.2022.907120. PMID: 35721491; PMCID: PMC9201071.
- Rhyu JJ, Yun JW, Kwon E, Che JH, Kang BC. Dual effects of human adipose tissue-derived mesenchymal stem cells in human lung adenocarcinoma A549 xenografts and colorectal adenocarcinoma HT-29 xenografts in mice. Oncol Rep. 2015 Oct;34(4):1733-44. doi: 10.3892/or.2015.4185. Epub 2015 Aug 7. PMID: 26252638.
- Richet N, Liu D, Legrand P, Velours C, Corpet A, Gaubert A, Bakail M, Moal-Raisin G, Guerois R, Compper C, Besle A, Guichard B, Almouzni G, Ochsenbein F. Structural insight into how the human helicase subunit MCM2 may act as a histone chaperone together with ASF1 at the replication fork. Nucleic Acids Res. 2015 Feb 18;43(3):1905-17. doi: 10.1093/nar/gkv021. Epub 2015 Jan 23. PMID: 25618846; PMCID: PMC4330383.
- Scaffidi P. Histone H1 alterations in cancer. Biochim Biophys Acta. 2016 Mar;1859(3):533-9. doi: 10.1016/j.bbagrm.2015.09.008. Epub 2015 Sep 18. PMID: 26386351.
- Soboleva TA, Nekrasov M, Pahwa A, Williams R, Huttley GA, Tremethick DJ. A unique H2A histone variant occupies the transcriptional start site of active genes. Nat Struct Mol Biol. 2011 Dec 4;19(1):25-30. doi: 10.1038/nsmb.2161. PMID: 22139013.
- Sun Z, Bernstein E. Histone variant macroH2A: from chromatin deposition to molecular function. Essays Biochem. 2019 Apr 23;63(1):59-74. doi: 10.1042/EBC20180062. PMID: 31015383.
- Tagami H, Ray-Gallet D, Almouzni G, Nakatani Y. Histone H3.1 and H3.3 complexes mediate nucleosome assembly pathways dependent or independent of DNA synthesis. Cell. 2004 Jan 9;116(1):51-61. doi: 10.1016/s0092-8674(03)01064-x. PMID: 14718166.
- Urahama T, Harada A, Maehara K, Horikoshi N, Sato K, Sato Y, Shiraishi K, Sugino N, Osakabe A, Tachiwana H, Kagawa W, Kimura H, Ohkawa Y, Kurumizaka H. Histone H3.5 forms an unstable nucleosome and accumulates around transcription start sites in human testis. Epigenetics Chromatin. 2016 Jan 15;9:2. doi: 10.1186/s13072-016-0051-y. PMID: 26779285; PMCID: PMC4714512.
- Verrelle P, Meseure D, Berger F, Forest A, Leclère R, Nicolas A, Fortas E, Sastre-Garau X, Lae M, Boudjemaa S, Mbagui R, Calugaru V, Labiod D, De Koning L, Almouzni G, Quivy JP. CENP-A Subnuclear Localization Pattern as Marker Predicting Curability by Chemoradiation Therapy for Locally Advanced Head and Neck Cancer Patients. Cancers (Basel). 2021 Aug 4;13(16):3928. doi: 10.3390/cancers13163928. PMID: 34439087; PMCID: PMC8391827.
- Volk A, Crispino JD. The role of the chromatin assembly complex (CAF-1) and its p60 subunit (CHAF1b) in homeostasis and disease. Biochim Biophys Acta. 2015 Aug;1849(8):979-86. doi: 10.1016/j.bbagrm.2015.05.009. Epub 2015 Jun 9. PMID: 26066981; PMCID: PMC4515380.
- Wang H, Walsh ST, Parthun MR. Expanded binding specificity of the human histone chaperone NASP. Nucleic Acids Res. 2008 Oct;36(18):5763-72. doi: 10.1093/nar/gkn574. Epub 2008 Sep 9. PMID: 18782834; PMCID: PMC2566879.
- Wang Z, Jensen MA, Zenklusen JC. A Practical Guide to The Cancer Genome Atlas (TCGA). Methods Mol Biol. 2016;1418:111-41. doi: 10.1007/978-1-4939-3578-9_6. PMID: 27008012.
- Yuan J, Lan H, Huang D, Guo X, Liu C, Liu S, Zhang P, Cheng Y, Xiao S. Multi-Omics Analysis of MCM2 as a Promising Biomarker in Pan-Cancer. Front Cell Dev Biol. 2022 May 25;10:852135. doi: 10.3389/fcell.2022.852135. PMID: 35693940; PMCID: PMC9174984.

- Zhang Q, Giebler HA, Isaacson MK, Nyborg JK. Eviction of linker histone H1 by NAP-family histone chaperones enhances activated transcription. Epigenetics Chromatin. 2015 Sep 4;8:30. doi: 10.1186/s13072-015-0022-8. PMID: 26339295; PMCID: PMC4558729.

Abstract:

In 2020, the new cases of cancer reached around 19.3 million and led to 10 million deaths worldwide *(Ferlay et al., 2020)*. Understanding mechanisms underlying cancer progression has helped to develop new treatments and to apply new therapies. However, one major bottleneck is the issue of resistance and relapse *(Vasan et al., 2019)*, a resistance that can either be present in the original tumor or induced by the therapy (*Wang et al., 2020)*. To date, the cause of cancer is considered as a combination of genetic and epigenetic defects. Notably, while genetic defects are fixed, the reversibility of epigenetic features offers means for intervention *(Montalvo-Casimiro et al., 2020)*. For this project, we aim to carry out an in-silico analysis focused on a particular set of epigenetic features involving the regulation of histone variants and their chaperones. We wish to characterize their expression patterns in the normal state across tissues, and their alteration in cancer. Given the known partnerships between variant and chaperone, we wish to explore the co-expression relationships between them, their disruption and changes in cancer cases as potential markers to predict resistance to therapy.

Keywords: Histone variants/chaperones network, Cancer resistance, RNA seq analysis, Generalized Linear model.

Résumé:

En 2020, dans le monde, les nouveaux cas de cancer ont atteint environ 19,3 millions et entraîné 10 millions de décès (Ferlay et al., 2020). La compréhension des mécanismes sous-jacents à la progression du cancer a permis de développer de nouveaux traitements et d'appliquer de nouvelles thérapies. Cependant, une limitation majeure est la question de la résistance et de la résurgence des cancers (Vasan et al., 2019), une résistance qui peut être présente dans la tumeur d'origine ou bien induite par la thérapie (Wang et al., 2020). À ce jour, la cause du cancer est considérée comme une combinaison de défauts génétiques et épigénétiques. Notamment, alors que les défauts génétiques ne peuvent être modifiés, la réversibilité des caractéristiques épigénétiques offre des moyens de restaurer une situation saine (Montalvo-Casimiro et al., 2020). Pour ce projet, nous visons à réaliser une analyse in-silico axée sur un

ensemble particulier de caractéristiques épigénétiques impliquant la régulation des variants d'histone et de leurs chaperons. Nous souhaitons caractériser leurs modes d'expression à l'état normal à travers les tissus, et leur altération dans le cancer. Compte tenu des partenariats connus entre variant et chaperon, nous souhaitons explorer les relations de co-expression entre eux, leur perturbation et les changements dans les cas de cancer en tant que marqueurs potentiels pour prédire la résistance au traitement.

Mots-clés: Réseaux d'histones/chaperons, Résistance au cancer, Analyse des séquences d'ARN, Modèle linéaire généralisé.