Chromatin network rearrangements in differentiation and cancer reveal connections between evolutionary processes and gene regulation

Benoît Aliaga[1*], Flavien Raynal[1*], Kaustav Sengupta[2,3,4], Dariusz Plewczynski[2,3] and Vera Pancaldi[1,5]

1- Centre de Recherches en Cancérologie de Toulouse, INSERM, CNRS, Université Toulouse III-Paul Sabatier, Toulouse, France;
2- Laboratory of Functional and Structural Genomics, Center of New Technologies (CeNT), University of Warsaw, Mazowieckie, Poland
3- Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland
4- The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA
5- Barcelona Supercomputing Center, Barcelona, Spain
* Co-first authors

Abstract

The availability of unprecedented datasets detailing the organization of chromatin across species, developmental stages, cell types and diseases invites an integrated analysis of genome organization in three dimensions and its biological significance.
In this study we hypothesize that topological properties of the chromatin network can impact plasticity and variability of phenotypes. Our observations lead us to reconsider the theory according to which cancer cells re-acquire pluripotency through reactivation of early evolutionary programs characteristic of species lacking complex body plans or even multicellularity.
Recent work has shown that promoter sequence characteristics, including general GC content, TATA-box motifs and number of TF binding motifs, can be predictive of gene expression variability across tissues and individuals in multiple species. Exploiting this dataset and others, we show that genes with specific promoter features and associated similar values of expression variability tend to have specific epigenomic marks and form preferential interactions in 3D, as quantified by chromatin assortativity. We further demonstrate that rearrangements of the 3D topology of the chromatin network in differentiation and oncogenesis can change these associations between gene regulation and evolutionary age. We observed that evolutionary older genes, also broadly corresponding to the least variable genes, are strongly clustered in the chromatin network.
We characterize the topology of promoter-centered chromatin networks as well as assortativities of gene ages and expression variability in multiple datasets capturing topological changes in differentiation (embryonic stem cells, B-cells) and find some of these changes to be reversed during oncogenesis.

Introduction

Researchers have identified several hallmarks of cancer describing characteristic cancer cell phenotypes (Hanahan, 2022), but the oncogenesis process remains partially understood. In the last decades, the atavistic theory, which suggests that cancer cells are normal cells that have lost their function by reverting to ancestral and unicellular phenotypes (Boveri, 2008; Busund and Rasmussen, 2020), has brought some cancer researchers to look at cancer from an evolutionary perspective.

In recent years, some researchers have become interested in the atavistic theory proposed in 1914 by Theodor Boveri, who saw the malignant tumor cell as a previously normal and “altruistic” tissue cell that changed into an “egotistical” mode with a loss of its specific functions. Over the years, this theory was forgotten, but in 2017, Trigos et al showed that the gene expression of evolutionary old genes increased in cancer, raising a new interest in this theory, but not explaining the underlying mechanism.

The human genome is the result of billions of years of evolution, and thanks to the high-throughput genomic technologies, the genome from several species across the tree of life were sequenced and revealed that molecular pathways are remarkably well conserved (Meadows and Lindblad-Toh, 2017). The approximately 20,000 coding human genes appeared at different evolutionary ages and evolved together. These genes have progressively been integrated across time, adding new functions. With the genomic phylostratigraphy method, it becomes possible to date the genes by looking at their homologs across species (Domazet-Lošo et al., 2007).

Recently, several studies explored cancer mechanisms across the tree of life. Cancers and cancer-like phenomena are present in the majority of the branches of the Tree of Life (Athena Aktipis et al., 2015). Animals require a greater division of labor, represented by their greater number of cell types compared to plants, fungi, and algae (Bell and Mooers, 1997). Evolutionary biologists hypothesize that cancer is the result of a breakdown of multicellular cooperation, and five aspects of cooperation are necessary to maintain a multicellular body: (i) proliferation inhibition, (ii) controlled cell death, (iii) division of labor, (iv) resource allocation, and (v) extracellular environment maintenance (Athena Aktipis et al., 2015). If several species share this cancer phenotype, it becomes possible that some mechanisms are in common. Multicellularity is characterized by cell cooperation for the development, maintenance, and reproduction of the multicellular organism. The differentiation process that occurs during development is shared with multicellular species. For example, some papers showed a conservation of fundamental biological processes of mammalian pluripotent stem cells across species (Endo et al., 2020). According to Waddington, this differentiation is explained by epigenetics. A subset of chromatin marks, like Polycomb deposited H3K27me3 and promoter-associated H3K4me3, were conserved more often than expected at orthologous regions with low substitution rates. Other chromatin marks can be conserved without any underlying sequence conservation (Xiao et al., 2012). Three chromatin-marks (DNA methylation, H3K36me3, and H3K27ac) in mammal ESCs are independent of DNA sequence evolution (Xiao et al., 2012).

Previous works used phylostratigraphy to study the origin of cancer in humans. The first study revealed that the appearance of most cancer genes dates back to the time of the

origin of unicellular organisms and of the transition to multicellularity (Domazet-Lošo and Tautz, 2010; Zhang et al., 2019), which enabled complex body plans. At the same time, species acquired the ability to regulate their cell cycles, modulate their growth and form complex networks of communication (Benton et al., 2021). Caretakers and gatekeepers are the two groups of genes that regulate cell cycle control. The caretaker genes are implied in basic control of the cell cycle and DNA repair. If they are mutated, the mutation rates and/or the genomic instability will increase the cancer risk. The origin of these genes are from the first evolutionary cells. Unlike the caretaker genes, the gatekeepers genes appeared later in evolutionary time (metazoan multicellularity). They play a role in regulating cell growth, death, and communication (Benton et al., 2021). Transcriptomics analyses from seven solid tumors showed that genes which originated from unicellular organisms were strongly up-regulated, and genes from metazoan origin were inactivated in cancer (Trigos et al., 2017). The increased unicellular gene expression seems to be a general phenomenon, while specific up-regulation depends on the function of the associated processes, suggesting that cancer atavism is not a chaotic or random process but an ancient choreography of correlated gene expression (Benton et al., 2021; Trigos et al., 2017). More interesting, cancer cells show expression of genes that are specific to embryogenesis and are not expressed in healthy adult tissues (Schoenhals et al., 2009). Indeed, it seems that epigenetic alterations occur very early in oncogenesis (Patani et al., 2020). In this paper we investigate whether this ancient choreography of correlated gene expression, which leads to expression of embryogenesis genes, may be reflected in 3D chromatin structure.

The 2-meter long chain of DNA folds into 3D space and coils inside the nucleus of each cell. This folding gives rise to various hierarchical structures of the genome. These structures are different in each cell but their overall characteristics can be specific to each cell type and also reflect cell state. Despite our incomplete knowledge about the mechanisms behind these processes, chromatin contacts can be associated with the cell's phenotype.

DNA loops are formed when two distant parts of the chromatin fiber are brought together in 3D space by various transcription factors (like CTCF, cohesin, RNA-POL II, etc.). A set of loops or a region which has a higher concentration of such loops within it than the number of loops crossing two regions is called a Topologically associated Domain (TAD) or Chromatin Contact Domain (CCD). A number of reviews describe our current understanding of genome organization in the nucleus (Bonev and Cavalli, 2016). Despite consistent progress in our understanding of chromatin organization since the advent of HiC related techniques over 10 years ago and non-sequencing based alternative experimental protocols, there are still open questions about how the characteristics of the 3D epigenome can be related to cellular processes and phenotypes.

A useful framework to represent the organization of chromatin inside the nucleus is that of network theory. Chromatin can be represented as a graph in which nodes are chromatin fragments and edges are experimentally detected contacts between fragments (through either sequencing based chromosome capture or any other 3D epigenome reconstruction technique). In this framework, TADs appear as modules in the network and chromatin interactions can be studied with network theory tools, to extract their statistical properties (Pancaldi, 2023).

To understand the functionality of the cell we need to decipher variability of 3D structure in the cell nucleus for various cell types. Studying the changes in 3D structure and variability

can shed light into the differentiation mechanism of cells (Chaicharoenaudomrung et al., 2019). Studies show that the variability in genome across various cell types and species is responsible for shaping innate immunity (Hagai et al., 2018). Moreover the variability of 3D architecture can be directly linked to hematological malignancies due to dysregulation of the lymphopoietic transcriptional program (Scourzic et al., 2021).

Whether and how the regulation of phenotypes in differentiation and cancer is driven or accompanied by rearrangements in chromatin structure is still not fully elucidated, but it is possible to speculate that the structure-function relationship connecting chromatin organization to phenotype could be key to understand differentiation and oncogenesis (Pancaldi, 2021).

Results

Gene expression and expression variability are strongly associated to evolutionary ages

To start exploring genes' evolutionary history in relation to their nuclear context, we downloaded a gene age estimation dataset, in which gene ages were assigned based on phylostratigraphy. Briefly, this method uses the BLAST sequence alignment algorithm (Altschul et al., 1990) for each gene in a reference species to infer the presence or the absence of ancestral homologs. With a predetermined phylogeny, an age will be assigned to each gene in the reference species when a homolog is detected and corresponds to the most distantly related species (Domazet-Lošo et al., 2007). A novel gene is detected when the algorithm is not able to detect homolog outside of its own genome, or close relatives. This dataset contains 19.404 genes that are categorized into 16 gene evolutionary ages, which can be grouped into 3 main age classes, unicellular (UC, 7.397 genes), Early metazoan (EM, 8.682 genes) and Mammal specific (MM, 3.325 genes) (**Supplementary Fig. 1A**) (Trigos et al., 2017). Throughout the manuscript, we will consider either the 16 specific age classes or the more global 3 groups, depending on the specific analysis.
We started by performing a functional enrichment analysis of the genes contained in each category (**Supplementary Fig. 1B**) and confirmed that genes in the UC class are enriched in basal cellular processes (catabolism, RNA splicing, histone modification, …) as opposed to genes assigned to later phylostrata involved in more complex cellular functions (sensory perception, keratinization, …).
We also tested enrichment of housekeeping genes with all the gene ages classes (**Supplementary Fig. 1C**). As expected, most housekeeping genes belong to UC genes class. Among all, 18% of UC genes are found to be housekeeping, against 6% of EM genes and only 1% of MM genes.

While the transformation of a healthy cell into a cancer one produces a total overhaul of a cell's transcriptome, partially due to the appearance of genetic alterations, we know that even healthy cells display a large amount of variability and plasticity in their phenotype based on changes of their epigenome. We therefore proceed to investigate the relationship of gene age with mean level of expression as well as with expression variability. In the following we will define variability as variation of values across instances (for example across individuals or across single cells) and plasticity as regulation of expression levels in response to external conditions changing in time.
More specifically, we employed a measure of expression variability (EV) defined by Alemu (Alemu et al., 2014) and successively used by Ecker *et al*. to estimate the variability of gene expression across individuals (Ecker et al., 2017), taking into account the dependence of this quantity on the mean of expression. Thus the EV value estimates the excess or reduced variability compared to what would be expected for the corresponding mean value of the feature.
We first investigated the link between gene ages and gene expression in 3 immune cell types (monocytes, neutrophils and T-cells) across 125 healthy individuals (Chen et al., 2016) (**Fig. 1A, Supplementary Fig. 1D**). By dividing genes into the 16 evolutionary age classes, we saw that gene age is inversely correlated with average gene expression (Spearman test,

Rho = -0.95, *p*-value < 2.2e-16 in Monocyte). Old genes have a higher average expression level than newer ones, which is consistent due to the basal cellular processes they are involved in.

Since expression variability is an important phenotype to understand the processes underlying gene regulation (Ecker et al., 2018), we next looked at the relation between expression variability and age in those 3 immune cell types (**Fig. 1B, Supplementary Fig. 1E**). We found that gene expression variability is also associated with gene ages, showing a lower variability for unicellular genes compared to newer genes (Spearman test, Rho = 0.77, *p*-value = 0.0007 in Monocyte).

We next confirmed statistically the link between gene expression variability and gene evolutionary ages. To do so, we divided genes into 10 equally sized groups according to their expression variability across individuals and tested enrichment of each main gene age category in these classes using Fisher's Exact Test (**Supplementary Fig. 1F**). We find that for each of the three immune cell types considered, UC genes are significantly enriched in the low variability groups (groups 1 to 6) while EM genes are significantly enriched in high variability groups (groups 8 to 10). These results confirm that old genes tend to have lower variability compared to newer genes.

In order to prove that this trend in variability is not just reflecting a dependence on the mean, we defined 10 equally sized gene groups based on expression distribution deciles in Monocyte. For each of the 10 expression level groups, we computed Spearman correlation across the 16 evolutionary ages (**Supplementary Fig. 1G**). We found that for the most variable gene groups (levels 7 to 10), the correlation of expression variability with gene age is progressively increasing across the 16 phylostrata with significant Spearman correlations. This shows that within a group composed of genes with the same variability level, lower variable genes correspond to older genes, while higher expression variability is related to evolutionary newer genes. This proves that the global variability distribution is not dependent on gene expression averages.

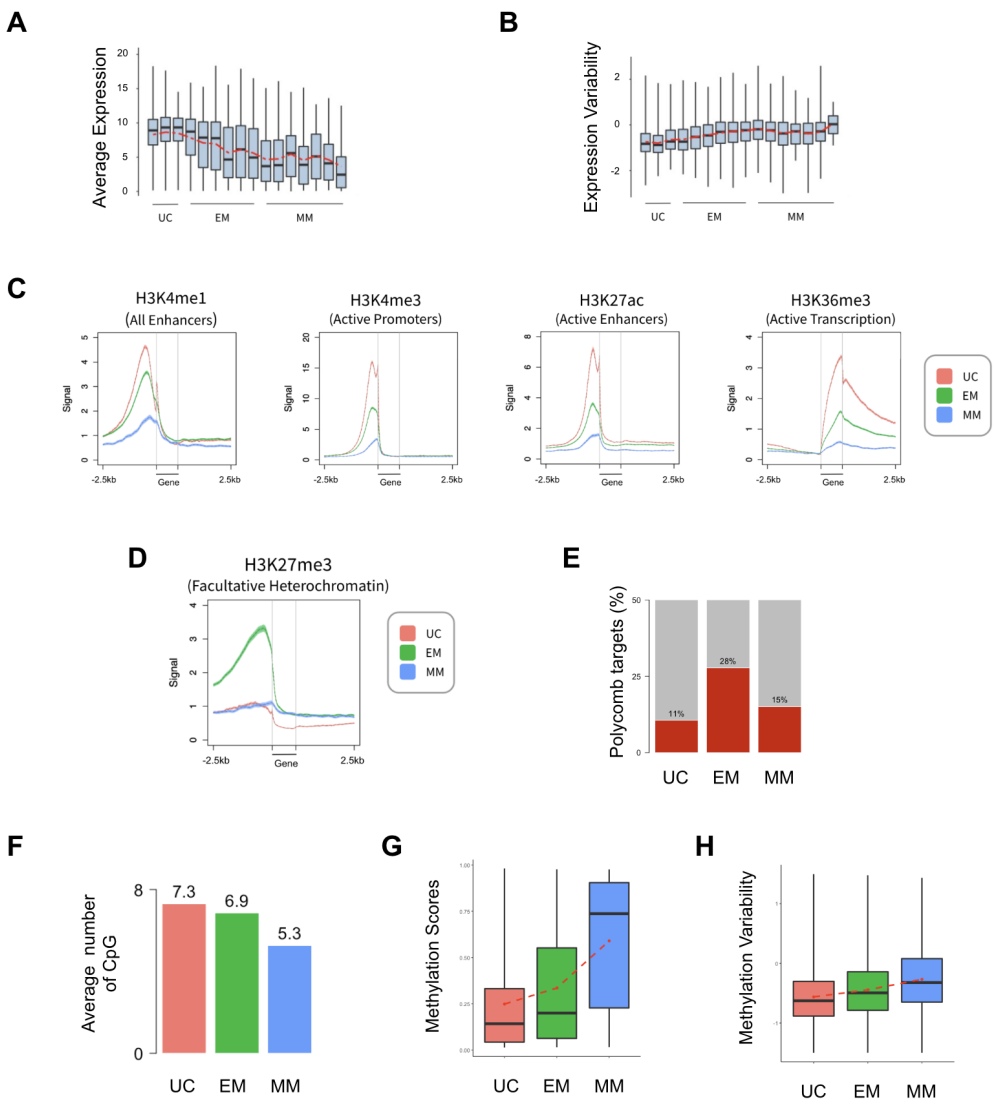


Figure 1: Genes of different evolutionary age classes display specific genetic and epigenomic features. A) Boxplot of average inter-individual gene expression levels in Monocyte across 16 gene age classes (Spearman Rho = -0.95, *p*-value < 2.2e-16). **B**) Boxplot of inter-individual expression variability in Monocyte across 16 gene age classes (Spearman Rho = 0.77, *p*-value = 0.00074). **C**) Average epigenetic profiles around the 3 main gene age classes (Red Unicellular genes, Green Early metazoan genes, Blue Mammal Specific genes). **D**) Average H3K27me3 profile around the 3 main gene age classes (Red Unicellular genes, Green Early metazoan genes, Blue Mammal Specific genes). **E**) Percentages of H3K27me3 repressed gene promoters in UC, EM and MM ages. **F**) Average number of potential CpG islands by gene across the 3 main age categories. **G**) Average Methylation scores in monocytes across the 3 main age categories. **H**) Average DNA methylation variability across the 3 main age categories.

To further test our findings on a separate dataset, we considered the features from Sigalova *et al* (Sigalova et al., 2020) (**Supplementary Fig. 1H**) which include estimates of variability across isogenic Drosophila lines as well as across human tissues. Interestingly, Sigalova *et al*. identified an impact of gene promoter features (sequence-based and epigenomic) in defining the level of gene expression variation. Analyzing these features related to variability in combination with the previously introduced age classes, we found correlations between

gene age and active TSS (Spearman Rho = -0.37, *p*-value < 2.2e-16), the number of TF binding motifs in their promoters (Spearman Rho = -0.34, *p*-value < 2.2e-16), the number of gene exons (Spearman Rho = -0.27, *p*-value < 2.2e-16), promoter size (Spearman Rho = -0.13, *p*-value < 2.2e-16) and GC content (Spearman Rho = -0.12, *p*-value < 2.2e-16). Thus old genes fall into the category that Sigalova *et al*. define as having low-variability thanks to broad promoters and multiple regulatory motifs.

Having observed the high correlation between evolutionary ages, gene expression levels and expression variability, we then decided to ask whether genes of specific age classes would display specific epigenomic patterns at the local gene scale. To address this question, we first downloaded ChIP-Seq datasets for the main histone modifications associated to gene activation in Monocyte. We analyzed average signal profiles around genes from the 3 main gene age classes (**Fig. 1C**) and found that the average signal across those epigenetic marks is always higher around UC genes, and decreases around metazoan EM genes. Signals show a lower enrichment across MM genes landscapes. High signal of active histone modification enrichments is consistent with gene expression levels of those genes age classes previously shown. These results suggest that genes' epigenetic environment is also related to their respective evolutionary ages.

Polycomb proteins target mainly metazoan genes

We then analyzed histone modifications associated to transcriptional repression. The results for the H3K27me3 mark associated with polycomb protein complexes are shown in (**Fig. 1D**). Interestingly, in Monocyte this mark shows enrichment around metazoan EM genes while both UC and MM genes show no enrichment. On the other hand, the H3K9me3 repressive mark shows no enrichment for any gene age class (**Supplementary Fig. 1I**). This suggests a regulation by polycomb complexes for this specific gene age class.

This observation led us to focus on genes targeted by polycomb proteins. To this end, we computed the proportion of genes from each main age class that are enriched with a H3K27me3 peak on the promoter in Monocyte (**Fig. 1E**). Results show a higher proportion of EM genes (28%) targeted by polycomb complexes while UC and MM both show a lower number of targeted genes (11% and 15%, respectively).

Taken together, these findings suggest that polycomb protein complexes regulate mainly genes that appeared at the metazoans period, which is in line with these genes being involved in the establishment of body plans and also requiring suppression in the early phases of development when the body plan is not final.

CpG islands and DNA methylation variability are associated to evolutionary ages

We finally decided to focus on DNA methylation as one of the main epigenomic features. Chen *et al*'s dataset contains a methylation score for each of the 440,905 detected methylated CpG across the entire human genome in Monocyte across 200 healthy individuals estimated by Whole Genome Bisulfite Sequencing (Chen et al., 2016). We first investigated the genome regions that are related to those methylated CpGs

(**Supplementary Fig. 1J**). Results show that half of all studied CpG islands occur around gene promoters (52.52% closer than 3kb to the TSS, 42% are around less than 1kb).

We therefore explored the relationship between gene ages and DNA methylation (**Fig. 1F**). Methylated CpG islands are known to mostly repress gene expression when they are located in promoters (Bird, 2002). We find that the average number of potential methylated CpG islands on gene promoter regions tend to progressively decrease through evolutionary gene age classes.

We then analyzed methylation scores computed by Chen *et al*. on Monocyte gene promoters classified according to their respective ages (**Fig. 1G**). Briefly, methylation scores were assigned to the 440,905 detected methylated CpG islands and calculated as the ratio between the intensity of the methylated probe and the global intensity within each individual. Results show progressively increasing methylation through the 3 main evolutionary ages. This suggests a preferential regulation by DNA methylation for the MM genes.

We finally computed the inter-individual DNA methylation variability (**Fig. 1H**). This variability increases roughly linearly across all gene age classes. This suggests a more stable DNA methylation profile for old genes, and a more variable DNA methylation profile around the newest genes.

These first results show that evolutionary gene ages correlate with different expression and epigenomic features. More precisely, each specific gene age class taken independently shows specific properties. The oldest genes, in class UC, have a higher expression level, have the lowest inter-individual expression variability, and are surrounded by the highest levels of histone modifications reminiscent of active genes. Metazoan genes, in class EM, are shown as the most regulated and repressed genes by polycomb protein complexes. Finally, the newest genes, in class MM, have the highest methylation scores on CpGs located on promoters and also have the most variable CpG methylation regulation. Genes' epigenomic environment is therefore linked to evolutionary ages and seems to be well established throughout the entire genome.

Most genes deregulated through cell differentiation appeared recently in evolution

These results led us to focus on several gene expression regulation processes that could potentially explain the organization and maintenance of age-related gene expression programs. We therefore decided to further explore which regulatory processes could be responsible for this difference in expression and variability between genes of different ages. Our observations about global expression, variability, and epigenomic profiles might suggest that this evolutionary-related organization could be required for the cell basal functioning. In parallel, we asked whether these programs are stable across global phenotypic cell changes, like cell differentiation and oncogenesis, to find out whether this organization is pre-programmed in stem cells, appears with differentiation, and whether it is altered during oncogenesis.

To do so, we decided to focus our analysis on 3 well defined and studied cell states for comparison: the human embryonic stem cell (hESC), as undifferentiated cell state, CD19+ primary B-cells, as a normal differentiated state, and B-cells from Chronic Lymphocytic Leukemia patients (B-CLL), as an example of a cancerous B-cell state.

We first asked whether genes with specific evolutionary ages are the most regulated during cell differentiation and oncogenesis. In order to study gene expression deregulation through cell differentiation, we analyzed a gene expression dataset from hESC undifferentiated cells to differentiated cardiomyocytes (Choy et al., 2018) (**Supplementary Fig. 2A**). We applied a differential gene expression analysis and computed proportions of up- and down-regulated genes in each evolutionary age class (**Fig. 2A, Supplementary Fig. 2B**). Results show that most deregulated genes are up-regulated through cell differentiation (from 5.382 differentially expressed genes, 89% are found up-regulated) which is consistent with acquisition of a cell-type specific gene expression profile. Interestingly, the proportion of up-regulated genes increases across evolutionary age classes (14% of UC genes, 32% of EM genes and 59% of MM genes). Few genes are found down-regulated in cardiac differentiation. This proportion tends to even decrease across gene age classes (5% of UC genes, 3% of EM genes and 2% of MM genes).

These findings could also be related to the fact that pluripotent cells have higher variability of gene expression across cells, related to the poised state of many genes (indicated mostly by the presence of Polycomb marks), leading to low-level expression of several genes resulting from different cells expressing different genes, especially in the old gene class implicated in basal cellular processes. Differentiation would elicit the up-regulation of specific genes in all or most of the cells, leading to strong up-regulation at the population level, especially involving genes that were less expressed because related to specific functions (in the EM and MM class).

In human cancer cells, gene expression profiles are known to be reshaped, including through deregulation of cancer-specific genes that are commonly involved in crucial basal cell functions. In 2017, Trigos *et al* showed that the oldest genes, in the UC age class, tend to have a higher gene expression level across 7 different tumor samples (Trigos et al., 2017). These findings from Trigos suggest that oncogenesis might re-establish a balance with higher expression of old genes more typical of undifferentiated states, potentially accompanying the regain of pluripotency which is typically observed in cancer cells.

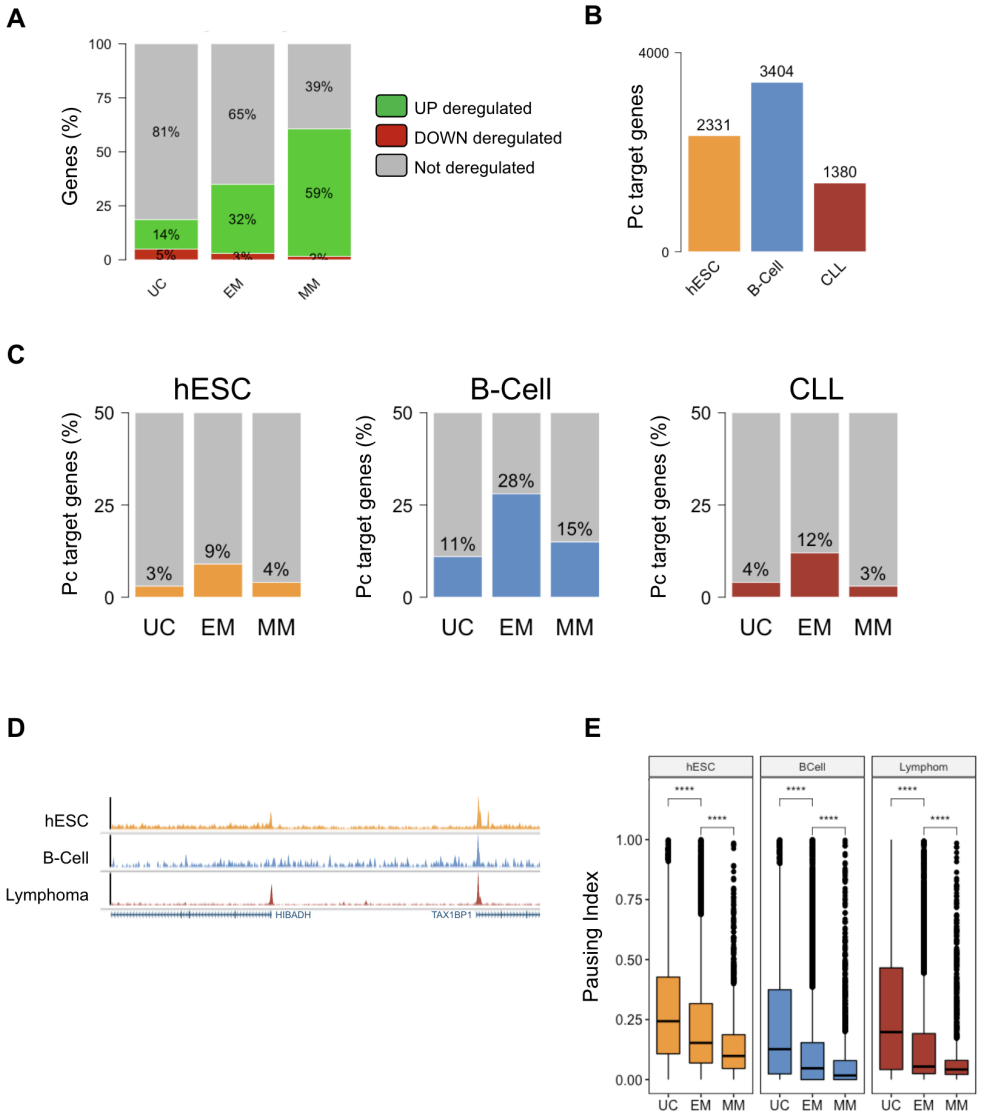


Figure 2. Gene expression regulation of different evolutionary ages is altered in differentiation and cancer.
A) Proportion of down (red), up (green) and no (grey) deregulated genes in Cardiomyocyte compared to hESC for the 3 main gene ages. **B**) Number of Polycomb target genes in hESC (yellow), B-cell (blue) and CLL (red). **C**) Percentage of Polycomb target genes for the 3 main gene age categories in hESC (left), B-cell (middle) and CLL (right) cell states. **D**) Browser snapshots of PolII tracks in hESC (yellow), B-cell (blue) and CLL (red). **E**) PolII Pausing Index distributions in hESC (yellow), B-cell (blue) and CLL (red) for each of the 3 main age groups. Statistical comparisons between groups were done by the Wilcoxon-Mann-Whitney test. n.s = not significant, **** = p<0.0001.

Gene expression regulation of genes from different evolutionary ages is altered in differentiation and cancer

In order to study how gene expression regulation is affected in cell differentiation and oncogenesis, we first focused on polycomb protein complexes, analyzing the polycomb-associated H3K27me3 mark in hESC, B-cell and CLL. It is well known that polycomb marks can impact the levels of expression noise (Kar et al., 2017).

We computed the total number of polycomb targeted genes across the genome in each cell state (**Fig. 2B**), showing an increasing number of targeted genes through cell differentiation and a strong decrease through oncogenesis. We then computed proportions of polycomb target genes in each main gene age category and cell stage (**Fig. 2C, Supplementary Fig. 2C**). Our previous results showed metazoan genes in the EM class were the most regulated by polycomb in Monocyte (c.f. **Fig. 1D, Fig. 1E, Fig. 1F**), and we found that this is also the case in stem cells, B-cells and CLL cancer cells. However, the effect of gene age class is strongest in B-cells and seems less pronounced in stem or cancer cells.

We then decided to study gene expression regulation by RNA polymerase II (PolII) binding, which could also impact the levels and variability of transcription. To do so, we analyzed Pausing RNA Polymerase II, which is a mechanism by which cells can control and synchronize transcription. The pausing PolII status can be quantified with the Pausing Index (PI), originally defined by Adelman & Lis in 2012 (Adelman and Lis, 2012). Briefly, the pausing index is calculated as the ratio of PolII signal density around gene TSS over the PolII signal density along the gene body.

In order to study the Pausing PolII Index, we downloaded PolII ChIP-Seq datasets in hESC, B-cell and B-CLL (**Fig. 2D**). We then computed PolII Pausing Indices for each gene and plotted distributions of this index in genes grouped by age class and in each cell state (**Fig. 2E**). We show a decreasing global distribution of Pausing Index across evolutionary classes, suggesting that oldest genes are more tightly regulated by the PolI pausing mechanism. In terms of cell states, global index distributions are shown to significantly decrease through cell differentiation and significantly increase through oncogenesis for every gene age class.

To summarise, we found that the newest genes are the most expressed across cell differentiation, while oldest genes are the most expressed across oncogenesis. Metazoan genes are the most targeted by polycomb complexes in any of our cell states. Polycomb activity tends to increase during stem cell differentiation and then decrease in cancer. We also showed that PolII Pausing is more prominent in older genes, potentially impacting their expression regulation. Opposite to polycomb activity, this mechanism tends to decrease in cell differentiation and increase in cancerogenesis.

The nuclear periphery is mostly harbouring newer genes

Gene regulation, especially through polycomb binding, has been strongly associated with the 3D chromatin structure inside the nucleus, which was shown to be deeply compromised in polycomb units mutants (Schoenfelder et al., 2015). Chromatin organization is known to be dramatically re-organised both in cell differentiation through lineage definition (Freire-Pritchett et al., 2017), and in cancer (Deng et al., 2022; Wang et al., 2022). This led us to investigate whether gene evolutionary ages might also be associated with gene localisations across the 3D genome conformation.

In order to start exploring the link between gene evolutionary ages and genome 3D structure, we asked whether genes from a specific gene age category would be more located at the nucleus periphery. To answer this question, we downloaded the dataset of computed Lamina Associated Domains (LADs) as annotated by Kind *et al*. (Kind et al.,

2015). This dataset was obtained from nuclear lamina interacting domains collected across nine different cell lines from multiple human tissues. It describes lamina interacting domains that are either classified as “facultative” cell type-specific or “constitutive” cell type-invariant considering the set of studied cell lines.
We considered genes contained in facultative and constitutive LADs either together or separately and examined their age (**Fig. 3A, Supplementary Fig. 3A**). Interestingly, we found a higher proportion of younger genes associated with LAD domains (15% and 13% of MM and EM age categories, respectively) compared to older genes (7% of UC genes). Results were similar considering only constitutive LADs. These results might suggest that most of the genes found at the nucleus periphery are newer in origin.
Considering that genes interacting with nuclear lamina are known to have a repressed expression (Reddy et al., 2008; Van Steensel and Belmont, 2017), this finding is consistent with the low expression level of the youngest genes (c.f. **Fig. 1A**).

We next downloaded Promoter Capture Hi-C (PCHi-C) datasets from Javierre *et al*. (Javierre et al., 2016) in order to study gene promoter interacting regions across the entire genome. We first considered separately contacts involving two gene promoters (PP) and contacts joining a promoter with a regulatory region, called other-end (PO). We computed contact ranges of PO interactions considering evolutionary ages assigned to promoters (**Fig. 3B**). In hESC, we noticed contact range tends to globally increase slightly through evolutionary ages, suggesting youngest gene promoters are located farther from their distal regulatory regions compared to oldest genes.

Genes of specific gene ages establish preferential contacts in the 3D chromatin network

In previous research, we and others have shown that many different epigenomic features such as many transcription factor binding peaks (Pancaldi et al., 2016), gene expression, DNA methylation, histone modifications (Madrid-Mencía et al., 2020) and even replication timing and origins of replication (Jodkowska et al., 2022) display strong patterns of clustering in 3D. We have used a network framework to represent chromatin interactions inside the nucleus and the concept of chromatin assortativity (ChAs) to identify which features cluster on the chromatin network (Pancaldi et al., 2016) more than would be expected based on genomic distances separating the contacting fragments.

We exploited promoter-centered chromatin networks derived from PCHi-C datasets in which each node is a DNA fragment and an edge represents a significant physical contact (or proximity) between two DNA fragments detected by the conformation capture protocol. We can either consider subnetworks composed of interacting promoters only (PP subnetworks) or composed of Promoter-OtherEnd contacts only (PO subnetworks) (Pancaldi et al. 2016, Madrid-Mencía et al., 2020).

We started by mapping gene ages on promoters on monocyte, neutrophil and T-cell PP subnetworks (**Fig. 3C, Supplementary Fig. 3B**). As can be seen on the genome-wide interaction promoters network, genes of specific gene age classes tend to group together, forming age-specific clusters. We next confirmed this tendency using ChAs computation with the 3 broad gene age classes interpreted as categorical values (**Fig. 3D, Supplementary**

Fig. 3C) (Madrid-Mencía et al., 2020). We observed ChAs z-scores for some age category being significantly above 2 (equivalent to *p*-value = 0.05). Interestingly, UC genes appear to be the most assortative gene class in the three cell types, always followed by EM. The most recent gene age category shows lower assortativity in the three immune cell types.

These findings suggest that the oldest unicellular genes are more highly connected amongst them and form clusters in the chromatin network, closely followed by EM genes who also form preferential interactions. Newest genes do form small clusters, but their preferential interactions seem less important, potentially because these small clusters are not connected amongst each other, as shown by their lower ChAs z-score.

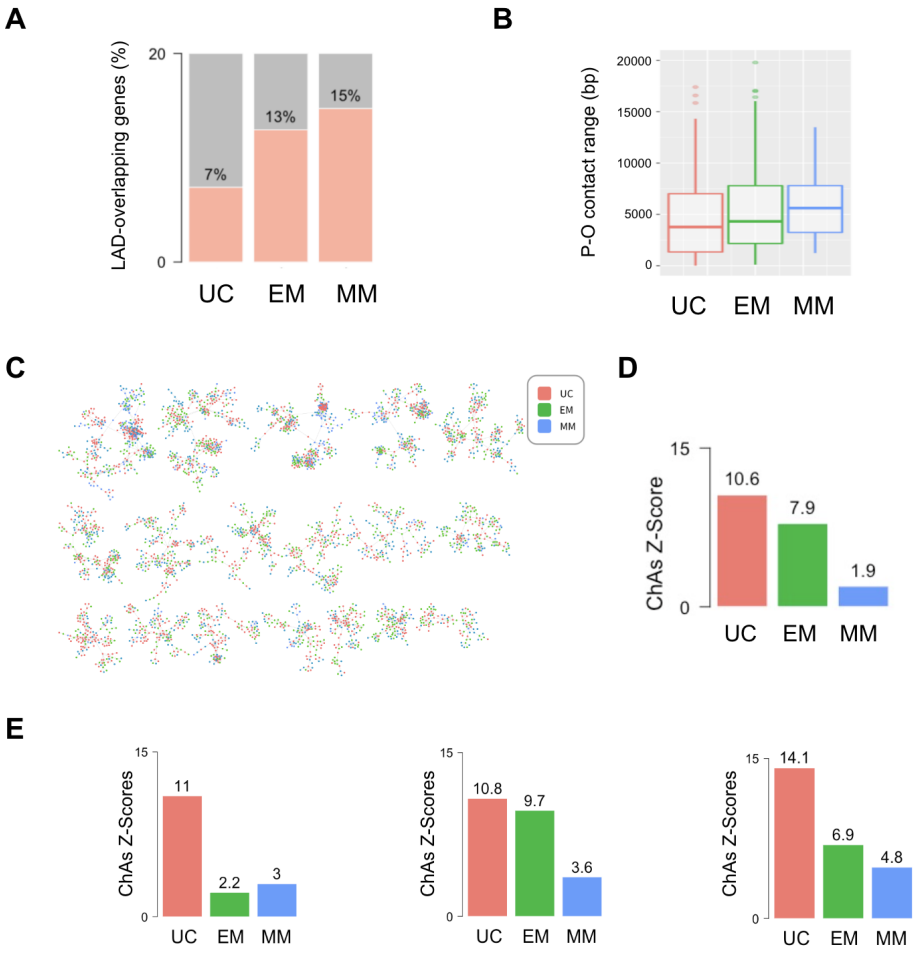


Figure 3: Genes of specific gene ages establish preferential contacts in the 3D chromatin network. A) Overlap with Lamina Associated Domains (LADs) annotated in multiple cell lines (Kind et al., 2015) for genes belonging to the different age classes. **B**) Genomic distance spanned by Promoter-Other end (PO) contacts of genes in the different age classes, from hESC PCHi-C data. **C**) Visualization of the 3D promoter-promoter interaction network in monocyte (Javierre *et al*. 2016) with node color representing gene age classes (red = UC, green = EM, blue = MM). **D**) Assortativity z-scores compared amongst the 3 age classes in monocyte promoter-promoter interaction network. **E**) Comparison of ChAs z-score of genes from different gene classes between hESCs, B-cells and CLL B-cells.

Observing those specific interactions between genes belonging to the same evolutionary age in Monocyte, we next asked to what extent this age-related 3D conformation might be impacted through cell differentiation and oncogenesis.

To answer this question, we downloaded PCHi-C datasets in hESC, normal B-cells and cancerous Chronic Lymphocytic Leukemia cells, that is transformed B-cells. As previously, we turned each set of interactions into a cell specific promoter-centered contact network. As was done on monocyte, we computed ChAs z-scores in each of our cell-specific PP subnetworks using the 3 main age classes as categorical features (**Fig. 3E**). In all cell states, we found the oldest UC gene category shows the highest assortativity values compared to other gene age classes, suggesting those genes keep strong clustering formation through cell differentiation and cancer. Interestingly, while MM genes always show low assortativities, we found EM genes having the most variable ChAs z-scores across cell states, with a clear increase in differentiated normal B-cells. This finding suggests a high reorganization of early metazoan genes contacts between themself across differentiation and oncogenesis. Those genes might contact together through cell differentiation, forming clusters at differentiated cell state, which might be lost through the oncogenesis process.

In the 3D context, our results suggest a specific genome conformation related to gene evolutionary ages. First, our lamina interacting domains analysis revealed that nucleus periphery is mainly enriched in recent genes from EM and MM classes. Our findings also prove the existence of preferential gene promoter interactions between genes belonging to the same evolutionary class on a genome wide scale. Among them, the oldest genes are shown to form the most preferential interactions, and this is observed across cell types, with no major differences observed in stem or cancer cells. Early metazoan genes also form preferential interactions but this is particularly prominent in differentiated cells, where they reach assortativity levels comparable to the UC genes, suggesting that an important reorganization of metazoan gene interactions accompanies cell differentiation and is lost in cancer.
Interestingly those EM genes are also the most targeted by polycomb proteins (as shown in **Fig. 2C**). In previous works, Polycomb features have been already shown to be highly assortative in promoter-centered chromatin networks (Pancaldi et al., 2016). Therefore, one can think that EM ChAs is closely linked to polycombs targets. Metazoan gene assortativity changes through cell differentiation and oncogenesis processes could be also related to changes in terms of polycomb activity.

Conclusion

Through integrated analysis of epigenomic marks, chromatin structure and gene expression, our results show that genes' epigenomic environment is linked to evolutionary ages. Interestingly, we found UC, EM and MM genes to have progressively more overlap with Lamina Associated Domains, which are typically localized on the nuclear membrane, suggesting that UC genes might occupy a more central position, far from the repressive chromatin marks present in LADs.

Using a chromatin network approach (Madrid-Mencía et al., 2020), we showed a relationship between gene age and chromatin structure, which goes beyond existing observations relating gene ages to the TAD structure (Acemel and Lupiáñez, 2023; James et al., 2021). In hESC, UC genes are more assortative, meaning that they are more preferentially interconnected, than EM and MM genes. The low value of ChAs z-score in EM was expected, since at this stage the cell is undifferentiated and the EM genes are not targeted by PcG, a complex which is found to be highly assortative, especially in stem cells (Pancaldi et al., 2016). In differentiated B-cells, the ChAs z-score of the EM genes is comparable to that of the UC genes, consistent with the PcG based repression of the cell-type specific EM genes. In CLL cancer cells, the ChAs z-score of the EM genes decrease compared to differentiated cells (c.f. **Fig. 3E**), while UC genes maintain their high assortativity. These results suggest that the UC genes are central, both geographically and topologically, in hESC and cancer, forming clusters in the chromatin network. In differentiated cells, tissue specific functions are probably related to the high assortativity of EM genes, while, during oncogenesis dedifferentiation leads towards an embryonic stem cell-like phenotype, as cells lose their positional information.

Since we found a strong relationship between gene expression variability, gene regulation characteristics and age, with variability decreasing with gene age, we surmise that the reshaping of the organization of genes in the nucleus during differentiation and oncogenesis might be related to the variability and plasticity of the phenotypes, which is reduced in fully differentiated cells and remains high in stem and cancer cells.

Despite the potential interest of the hypotheses raised in this work, we must mention several limitations of this study. First and foremost, we have investigated published datasets integrating chromatin contact maps with epigenomic marks from different publications, potentially referring to cells in slightly different conditions and undoubtedly being affected by the biases specific to each type of experimental assay. Secondly, the chromatin assortativity measures depend, to a certain extent, on the size and characteristics of the contact networks compared. We therefore tried to limit ourselves to comparing ChAs z-score values of gene age categories within a single chromatin network (for example, hESC, B-cell or CLL cells) rather than comparing across the different networks.

Further work will have to be devoted to understand whether the regulation of specific genes or a global rearrangement in the way that broad gene categories are distributed in 3D can be the cause of the modification in variability (across cells) and plasticity (across time) alteration during differentiation and oncogenesis. Taken together, our results suggest that the increased centrality of older genes in cancer observed by Trigos et al. in protein-protein interaction networks could have a counterpart in chromatin 3D structure networks.

Taken together, our findings support the atavistic theory hypothesis, according to which older genes reacquire an important role in cancer, eliciting programs that are reminiscent of early stages of development and of unicellular phenotypes.

Material and Methods

Gene Ages

We used Gene ages dataset from Trigos *et al* public published work which has estimated gene ages by phylostratigraphy (Trigos et al., 2017). 19.404 genes are classified into 16 age classes from unicellular genes to Homo sapiens. Those classes can be redefined into 3 main gene age categories : Unicellular Ancestor genes (UC) with 7.397 genes, Early Metazoan genes (EM) with 8.682 genes and Mammal-Specific genes (MM) with 3.325 genes.

Expression and DNA methylation Variability

Expression and DNA methylation Variability data in immune cells comes from previous Simone Ecker *et al* published work (Ecker et al., 2017). Variability has been computed from 125 healthy individuals with method accounted for confounding effects due to the correlation between mean and variability measurements. Fisher's exact tests have been computed by using geneOverlap (v1.32.0) R package (Li Shen, 2017).

ChIP-Seq & RNA-Seq dataset accession

Datasets used in this study with corresponding accession numbers are summarized in the following table:

Target	Cell	Database	Identifier
H3K4me1	Monocyte	ENCODE	ENCFF929UOK
H3K4me3	Monocyte	ENCODE	ENCFF231OTU
H3K27ac	Monocyte	ENCODE	ENCFF437JSB
H3K36me3	Monocyte	ENCODE	ENCFF064OOI
H3K27me3	Monocyte	ENCODE	ENCFF557YRZ
H3K9me3	Monocyte	ENCODE	ENCFF256KXQ
EZH2	B-cell (GM12878)	ENCODE	ENCFF410SZR
RNA-Seq	hESC	GEO	GSE69618
RNA-Seq	CM	GEO	GSE69618
H3K27me3	hESC (h9 /WA09)	GEO	SRR11185316
Input	hESC (h9 /WA09)	GEO	SRR11185368
H3K27me3	B-cell	BLUEPRINT	EGAX00001215956

Input	B-cell	BLUEPRINT	EGAX00001215601
H3K27me3	CLL B-cell	BLUEPRINT	EGAX00001231899
Input	CLL B-cell	BLUEPRINT	EGAX00001223301
PolII	hESC (h9 /WA09)	GEO	SRR7741345
PolII	B-cell (GM12878)	ENCODE	ENCFF000OBB
PolII	Lymphoma B-cell (LY1)	GEO	GSM1133652

ChIP-Seq Processing

Histone modifications ChIP-Seq datasets in Monocyte have been downloaded from the ENCODE portal (https://www.encodeproject.org/) (Sloan et al., 2016) in bigwig format with ENCODE identifiers shown in the table above.
For other ChIP-Seq datasets, sequenced reads were mapped to the human reference genome (built hg19/GRCh37) by using Burrows-Wheeler Aligner bwa (v0.7.17-r1188) (Li and Durbin, 2010). Bam files were obtained, sorted and indexed by using samtools utilities (v1.15.1) (Li et al., 2009). ChIP-Seq peak calling was done using MACS2 (v.2.2.7.1) with respective inputs (Zhang et al., 2008). H3K27me3 peak calling was done with --broad parameter on.

ChIP-Seq Analysis

ChIP-Seq profiles were computed with the seqplots (v1.23.3) R package (Stempor and Ahringer, 2016). Polycomb target genes were defined as genes with strict TSS localisation (according to EnsDb.Hsapiens.v75 Ensembl hg19 gene annotations) that overlaps with a H3K27me3 peak. CpG localizations were studied using the ChIPSeeker (v1.32.1) R package (Yu et al., 2015).

RNA-Seq Analysis

Raw reads count matrix of hESC and hESC-derived Cardiomyocyte was downloaded from GEO processed file under accession number GSE69618. Differential gene expression analysis was processed using the DESeq2 (v1.36.0) R package. Differential gene expression with log2FC > 1.2 and *p*-value < 0.05 were considered as deregulated.
Gene Ontology enrichment analysis was done using limma (Ritchie et al., 2015) (v3.52.1) and clusterProfiler (Wu et al., 2021) (v4.4.3) R packages.

PolII Pausing Index

We downloaded RNA Polymerase II ChIP-Seq raw Fastq files for hESC (H9/WA09 cell line), B-cell (GM12878 cell line) and Lymphoma B-cell (LY1 cell line) and processed as previously

described. PolII Bigwig density files were obtained using the bamCoverage tool from deeptools (v3.0.2) utilities and normalized using the flag –normalizeUsing RPKM on mode.
In order to compute PolII Pausing Indexes, we defined for each gene the promoter region from TSS-30bp to TSS+300bp and the gene body region from TSS+300bp to TES according to EnsDb.Hsapiens.v75 gene annotation. PolII scores of every overlapping bins to both regions were averaged separately. The Pausing Index was defined as the ratio of PolII density score on the promoter region over the density on the gene body region.
For each cell index distribution, we detected and removed outliers using the InterQuartile Range (IQR) method. We removed genes such that the sum of PolII pausing index in the 3 cells is lower than 0.2, defined as unexpressed. Genes with length lower than 1kb according to EnsDb.Hsapiens.v75 gene annotation were also discarded. In order to compare indices across cells, we normalized each distribution such that the normalized data falls between 0 and 1.

LAD-associated genes

We used the LADs dataset Kind *et al*, 2015 considering 'constitutive' and 'facultative' LADs either together or separately (Kind et al., 2015). Nuclear lamina interacting genes were considered as genes for which at least a part of the gene body (according to EnsDb.Hsapiens.v75 Ensembl hg19 gene annotations) overlapped with LADs.

PCHi-C chromatin networks ChAs

We used multiple PChi-C datasets for comparisons, with GEO number accession summarized in the following table:

Cell	Publication
hESC	Freire-Pritchett *et al* (Freire-Pritchett et al., 2017)
B-cell	Javierre *et al* (Javierre et al., 2016)
CLL	Beekman *et al* (Beekman et al., 2018)

From the significant called chromatin interactions, we first turn such data into networks by considering DNA restriction fragments as nodes and significant physical interaction between fragments as edges. We then loaded those networks on R and computed topological measurements with the igraph (v1.3.2) R package (Csárdi et al., 2023). We computed Chromatin Assortativity by using the ChAseR (v0.0.0.9) R package available on bitbucket (Madrid-Mencía et al., 2020). Gene ages ChAs was computed as classes with the categorical ChAs mode. We used 100 randomizations preserving genomic distances (dist.match = T) for z-scores calculations. We used Cytoscape (v3.8.0) and Gephi (v0.10.1) for chromatin network visualization.

References

Acemel, R.D., Lupiáñez, D.G., 2023. Evolution of 3D chromatin organization at different scales. Curr. Opin. Genet. Dev. 78, 102019. https://doi.org/10.1016/j.gde.2022.102019

Adelman, K., Lis, J.T., 2012. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. Nat. Rev. Genet. 13, 720–731. https://doi.org/10.1038/nrg3293

Alemu, E.Y., Carl, J.W., Corrada Bravo, H., Hannenhalli, S., 2014. Determinants of expression variability. Nucleic Acids Res. 42, 3503–3514. https://doi.org/10.1093/nar/gkt1364

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Athena Aktipis, C., Boddy, A.M., Jansen, G., Hibner, U., Hochberg, M.E., Maley, C.C., Wilkinson, G.S., 2015. Cancer across the tree of life: Cooperation and cheating in multicellularity. Philos. Trans. R. Soc. B Biol. Sci. 370. https://doi.org/10.1098/rstb.2014.0219

Beekman, R., Chapaprieta, V., Russiñol, N., Vilarrasa-Blasi, R., Verdaguer-Dot, N., Martens, J.H.A., Duran-Ferrer, M., Kulis, M., Serra, F., Javierre, B.M., Wingett, S.W., Clot, G., Queirós, A.C., Castellano, G., Blanc, J., Gut, M., Merkel, A., Heath, S., Vlasova, A., Ullrich, S., Palumbo, E., Enjuanes, A., Martín-García, D., Beà, S., Pinyol, M., Aymerich, M., Royo, R., Puiggros, M., Torrents, D., Datta, A., Lowy, E., Kostadima, M., Roller, M., Clarke, L., Flicek, P., Agirre, X., Prosper, F., Baumann, T., Delgado, J., López-Guillermo, A., Fraser, P., Yaspo, M.-L., Guigó, R., Siebert, R., Martí-Renom, M.A., Puente, X.S., López-Otín, C., Gut, I., Stunnenberg, H.G., Campo, E., Martin-Subero, J.I., 2018. The reference epigenome and regulatory chromatin landscape of chronic lymphocytic leukemia. Nat. Med. 24, 868–880. https://doi.org/10.1038/s41591-018-0028-4

Bell, G., Mooers, A.O., 1997. Size and complexity among multicellular organisms. Biol. J. Linn. Soc. 60, 345–363. https://doi.org/10.1111/j.1095-8312.1997.tb01500.x

Benton, M.L., Abraham, A., LaBella, A.L., Abbot, P., Rokas, A., Capra, J.A., 2021. The influence of evolutionary history on human health and disease. Nat. Rev. Genet. 22, 269–283. https://doi.org/10.1038/s41576-020-00305-9

Bird, A., 2002. DNA methylation patterns and epigenetic memory. Genes Dev 16. https://doi.org/10.1101/gad.947102

Bonev, B., Cavalli, G., 2016. Organization and function of the 3D genome. Nat. Rev. Genet. 17, 772–772. https://doi.org/10.1038/nrg.2016.147

Boveri, T., 2008. Concerning the Origin of Malignant Tumours by Theodor Boveri. Translated and annotated by Henry Harris. J. Cell Sci. 121, 1–84. https://doi.org/10.1242/jcs.025742

Busund, L.-T., Rasmussen, 2020. 10. Hypotheses of Carcinogenesis—The AtavisticTheory, in: Advancing Systems Epidemiology in Cancer. Universitetsforlaget. https://doi.org/10.18261/9788215041193-2020-10

Chaicharoenaudomrung, N., Kunhorm, P., Noisa, P., 2019. Three-dimensional cell culture systems as an *in vitro* platform for cancer and stem cell modeling. World J. Stem Cells 11, 1065–1083. https://doi.org/10.4252/wjsc.v11.i12.1065

Chen, L., Ge, B., Casale, F.P., Vasquez, L., Kwan, T., Garrido-Martín, D., Watt, S., Yan, Y., Kundu, K., Ecker, S., Datta, A., Richardson, D., Burden, F., Mead, D., Mann, A.L., Fernandez, J.M., Rowlston, S., Wilder, S.P., Farrow, S., Shao, X., Lambourne, J.J., Redensek, A., Albers, C.A., Amstislavskiy, V., Ashford, S., Berentsen, K., Bomba, L., Bourque, G., Bujold, D., Busche, S., Caron, M., Chen, S.-H., Cheung, W., Delaneau, O., Dermitzakis, E.T., Elding, H., Colgiu, I., Bagger, F.O., Flicek, P., Habibi, E., Iotchkova, V., Janssen-Megens, E., Kim, B., Lehrach, H., Lowy, E., Mandoli, A., Matarese, F., Maurano, M.T., Morris, J.A., Pancaldi, V., Pourfarzad, F., Rehnstrom,

K., Rendon, A., Risch, T., Sharifi, N., Simon, M.-M., Sultan, M., Valencia, A., Walter, K., Wang, S.-Y., Frontini, M., Antonarakis, S.E., Clarke, L., Yaspo, M.-L., Beck, S., Guigo, R., Rico, D., Martens, J.H.A., Ouwehand, W.H., Kuijpers, T.W., Paul, D.S., Stunnenberg, H.G., Stegle, O., Downes, K., Pastinen, T., Soranzo, N., 2016. Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. Cell 167, 1398-1414.e24. https://doi.org/10.1016/j.cell.2016.10.026

Choy, M.-K., Javierre, B.M., Williams, S.G., Baross, S.L., Liu, Y., Wingett, S.W., Akbarov, A., Wallace, C., Freire-Pritchett, P., Rugg-Gunn, P.J., Spivakov, M., Fraser, P., Keavney, B.D., 2018. Promoter interactome of human embryonic stem cell-derived cardiomyocytes connects GWAS regions to cardiac gene networks. Nat. Commun. 9, 2526. https://doi.org/10.1038/s41467-018-04931-0

Csárdi, G., Nepusz, T., Müller, K., Horvát, S., Traag, V., Zanini, F., Noom, D., 2023. igraph for R: R interface of the igraph library for graph theory and network analysis. https://doi.org/10.5281/ZENODO.7682609

Deng, S., Feng, Y., Pauklin, S., 2022. 3D chromatin architecture and transcription regulation in cancer. J. Hematol. Oncol.J Hematol Oncol 15, 49. https://doi.org/10.1186/s13045-022-01271-x

Domazet-Lošo, T., Brajković, J., Tautz, D., 2007. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. Trends Genet. 23, 533–539. https://doi.org/10.1016/j.tig.2007.08.014

Domazet-Lošo, T., Tautz, D., 2010. Phylostratigraphic tracking of cancer genes suggests a link to the emergence of multicellularity in metazoa. BMC Biol. 8, 66. https://doi.org/10.1186/1741-7007-8-66

Ecker, S., Chen, L., Pancaldi, V., Bagger, F.O., Fernández, J.M., Carrillo De Santa Pau, E., Juan, D., Mann, A.L., Watt, S., Casale, F.P., Sidiropoulos, N., Rapin, N., Merkel, A., BLUEPRINT Consortium, Stunnenberg, H.G., Stegle, O., Frontini, M., Downes, K., Pastinen, T., Kuijpers, T.W., Rico, D., Valencia, A., Beck, S., Soranzo, N., Paul, D.S., 2017. Genome-wide analysis of differential transcriptional and epigenetic variability across human immune cell types. Genome Biol. 18, 18. https://doi.org/10.1186/s13059-017-1156-8

Ecker, S., Pancaldi, V., Valencia, A., Beck, S., Paul, D.S., 2018. Epigenetic and Transcriptional Variability Shape Phenotypic Plasticity. BioEssays 40, 1700148. https://doi.org/10.1002/bies.201700148

Endo, Y., Kamei, K., Inoue-Murayama, M., 2020. Genetic Signatures of Evolution of the Pluripotency Gene Regulating Network across Mammals. Genome Biol. Evol. 12, 1806–1818. https://doi.org/10.1093/gbe/evaa169

Freire-Pritchett, P., Schoenfelder, S., Várnai, C., Wingett, S.W., Cairns, J., Collier, A.J., García-Vílchez, R., Furlan-Magaril, M., Osborne, C.S., Fraser, P., Rugg-Gunn, P.J., Spivakov, M., 2017. Global reorganisation of cis-regulatory units upon lineage commitment of human embryonic stem cells. eLife 6, e21926. https://doi.org/10.7554/eLife.21926

Hagai, T., Chen, X., Miragaia, R.J., Rostom, R., Gomes, T., Kunowska, N., Henriksson, J., Park, J.E., Proserpio, V., Donati, G., Bossini-Castillo, L., Vieira Braga, F.A., Naamati, G., Fletcher, J., Stephenson, E., Vegh, P., Trynka, G., Kondova, I., Dennis, M., Haniffa, M., Nourmohammad, A., Lässig, M., Teichmann, S.A., 2018. Gene expression variability across cells and species shapes innate immunity. Nature 563, 197–202. https://doi.org/10.1038/s41586-018-0657-2

Hanahan, D., 2022. Hallmarks of Cancer: New Dimensions. Cancer Discov. 12, 31–46. https://doi.org/10.1158/2159-8290.CD-21-1059

James, C., Trevisan-Herraz, M., Juan, D., Rico, D., 2021. Vertebrate whole genome duplications shaped the current 3D genome architecture (preprint). Genomics. https://doi.org/10.1101/2021.06.11.448047

Javierre, B.M., Burren, O.S., Wilder, S.P., Kreuzhuber, R., Hill, S.M., Sewitz, S., Cairns, J., Wingett, S.W., Várnai, C., Thiecke, M.J., Burden, F., Farrow, S., Cutler, A.J., Rehnström, K., Downes, K., Grassi, L., Kostadima, M., Freire-Pritchett, P., Wang, F.,

Martens, J.H., Kim, B., Sharifi, N., Janssen-Megens, E.M., Yaspo, M.-L., Linser, M., Kovacsovics, A., Clarke, L., Richardson, D., Datta, A., Flicek, P., Stunnenberg, H.G., Todd, J.A., Zerbino, D.R., Stegle, O., Ouwehand, W.H., Frontini, M., Wallace, C., Spivakov, M., Fraser, P., 2016. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. Cell 167, 1369-1384.e19. https://doi.org/10.1016/j.cell.2016.09.037

Jodkowska, K., Pancaldi, V., Rigau, M., Almeida, R., Fernández-Justel, J.M., Graña-Castro, O., Rodríguez-Acebes, S., Rubio-Camarillo, M., Carrillo-de Santa Pau, E., Pisano, D., Al-Shahrour, F., Valencia, A., Gómez, M., Méndez, J., 2022. 3D chromatin connectivity underlies replication origin efficiency in mouse embryonic stem cells. Nucleic Acids Res. 50, 12149–12165. https://doi.org/10.1093/nar/gkac1111

Kar, G., Kim, J.K., Kolodziejczyk, A.A., Natarajan, K.N., Torlai Triglia, E., Mifsud, B., Elderkin, S., Marioni, J.C., Pombo, A., Teichmann, S.A., 2017. Flipping between Polycomb repressed and active transcriptional states introduces noise in gene expression. Nat. Commun. 8, 36. https://doi.org/10.1038/s41467-017-00052-2

Kind, J., Pagie, L., de Vries, S.S., Nahidiazar, L., Dey, S.S., Bienko, M., Zhan, Y., Lajoie, B., de Graaf, C.A., Amendola, M., Fudenberg, G., Imakaev, M., Mirny, L.A., Jalink, K., Dekker, J., van Oudenaarden, A., van Steensel, B., 2015. Genome-wide Maps of Nuclear Lamina Interactions in Single Human Cells. Cell 163, 134–147. https://doi.org/10.1016/j.cell.2015.08.040

Li, H., Durbin, R., 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26, 589–595. https://doi.org/10.1093/bioinformatics/btp698

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li Shen, M.S. <Shenli S.C., 2017. GeneOverlap. https://doi.org/10.18129/B9.BIOC.GENEOVERLAP

Madrid-Mencía, M., Raineri, E., Cao, T.B.N., Pancaldi, V., 2020. Using GARDEN-NET and ChAseR to explore human haematopoietic 3D chromatin interaction networks. Nucleic Acids Res. 48, 4066–4080. https://doi.org/10.1093/nar/gkaa159

Meadows, J.R.S., Lindblad-Toh, K., 2017. Dissecting evolution and disease using comparative vertebrate genomics. Nat. Rev. Genet. 18, 624–636. https://doi.org/10.1038/nrg.2017.51

Pancaldi, V., 2023. Network models of chromatin structure. Curr. Opin. Genet. Dev. 80, 102051. https://doi.org/10.1016/j.gde.2023.102051

Pancaldi, V., 2021. Chromatin Network Analyses: Towards Structure-Function Relationships in Epigenomics. Front. Bioinforma. 1, 742216. https://doi.org/10.3389/fbinf.2021.742216

Pancaldi, V., Carrillo-de-Santa-Pau, E., Javierre, B.M., Juan, D., Fraser, P., Spivakov, M., Valencia, A., Rico, D., 2016. Integrating epigenomic data and 3D genomic structure with a new measure of chromatin assortativity. Genome Biol. 17, 152. https://doi.org/10.1186/s13059-016-1003-3

Patani, H., Rushton, M.D., Higham, J., Teijeiro, S.A., Oxley, D., Cutillas, P., Sproul, D., Ficz, G., 2020. Transition to naïve human pluripotency mirrors pan-cancer DNA hypermethylation. Nat. Commun. 11, 3671. https://doi.org/10.1038/s41467-020-17269-3

Reddy, K.L., Zullo, J.M., Bertolino, E., Singh, H., 2008. Transcriptional repression mediated by repositioning of genes to the nuclear lamina. Nature 452, 243–247. https://doi.org/10.1038/nature06727

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., Smyth, G.K., 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 43, e47–e47. https://doi.org/10.1093/nar/gkv007

Schoenfelder, S., Sugar, R., Dimond, A., Javierre, B.-M., Armstrong, H., Mifsud, B., Dimitrova, E., Matheson, L., Tavares-Cadete, F., Furlan-Magaril, M., Segonds-Pichon, A., Jurkowski, W., Wingett, S.W., Tabbada, K., Andrews, S.,

Herman, B., LeProust, E., Osborne, C.S., Koseki, H., Fraser, P., Luscombe, N.M., Elderkin, S., 2015. Polycomb repressive complex PRC1 spatially constrains the mouse embryonic stem cell genome. Nat. Genet. 47, 1179–1186. https://doi.org/10.1038/ng.3393

Schoenhals, M., Kassambara, A., Vos, J.D., Hose, D., Moreaux, J., Klein, B., 2009. Embryonic stem cell markers expression in cancers. Biochem. Biophys. Res. Commun. 383, 157–162. https://doi.org/10.1016/j.bbrc.2009.02.156

Scourzic, L., Salataj, E., Apostolou, E., 2021. Deciphering the Complexity of 3D Chromatin Organization Driving Lymphopoiesis and Lymphoid Malignancies. Front. Immunol. 12, 669881. https://doi.org/10.3389/fimmu.2021.669881

Sigalova, O.M., Shaeiri, A., Forneris, M., Furlong, E.E., Zaugg, J.B., 2020. Predictive features of gene expression variation reveal mechanistic link with differential expression. Mol. Syst. Biol. 16, e9539. https://doi.org/10.15252/msb.20209539

Sloan, C.A., Chan, E.T., Davidson, J.M., Malladi, V.S., Strattan, J.S., Hitz, B.C., Gabdank, I., Narayanan, A.K., Ho, M., Lee, B.T., Rowe, L.D., Dreszer, T.R., Roe, G., Podduturi, N.R., Tanaka, F., Hong, E.L., Cherry, J.M., 2016. ENCODE data at the ENCODE portal. Nucleic Acids Res. 44, D726–D732. https://doi.org/10.1093/nar/gkv1160

Stempor, P., Ahringer, J., 2016. SeqPlots - Interactive software for exploratory data analyses, pattern discovery and visualization in genomics. Wellcome Open Res. 1, 14. https://doi.org/10.12688/wellcomeopenres.10004.1

Trigos, A.S., Pearson, R.B., Papenfuss, A.T., Goode, D.L., 2017. Altered interactions between unicellular and multicellular genes drive hallmarks of transformation in a diverse range of solid tumors. Proc. Natl. Acad. Sci. U. S. A. 114, 6406–6411. https://doi.org/10.1073/pnas.1617743114

Van Steensel, B., Belmont, A.S., 2017. Lamina-Associated Domains: Links with Chromosome Architecture, Heterochromatin, and Gene Repression. Cell 169, 780–791. https://doi.org/10.1016/j.cell.2017.04.022

Wang, M., Sunkel, B.D., Ray, W.C., Stanton, B.Z., 2022. Chromatin structure in cancer. BMC Mol. Cell Biol. 23, 35. https://doi.org/10.1186/s12860-022-00433-6

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., Fu, X., Liu, S., Bo, X., Yu, G., 2021. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. The Innovation 2, 100141. https://doi.org/10.1016/j.xinn.2021.100141

Xiao, S., Xie, D., Cao, X., Yu, P., Xing, X., Chen, C.-C., Musselman, M., Xie, M., West, F.D., Lewin, H.A., Wang, T., Zhong, S., 2012. Comparative Epigenomic Annotation of Regulatory DNA. Cell 149, 1381–1392. https://doi.org/10.1016/j.cell.2012.04.029

Yu, G., Wang, L.-G., He, Q.-Y., 2015. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. Bioinformatics 31, 2382–2383. https://doi.org/10.1093/bioinformatics/btv145

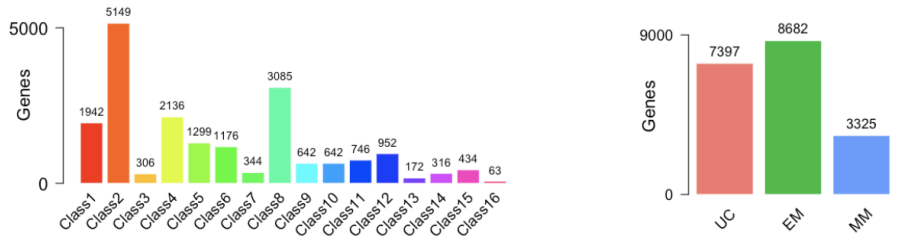
Zhang, L., Tan, Y., Fan, S., Zhang, X., Zhang, Z., 2019. Phylostratigraphic analysis of gene co-expression network reveals the evolution of functional modules for ovarian cancer. Sci. Rep. 9, 2623. https://doi.org/10.1038/s41598-019-40023-9

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., Liu, X.S., 2008. Model-based Analysis of ChIP-Seq (MACS). Genome Biol. 9, R137. https://doi.org/10.1186/gb-2008-9-9-r137

Supplementary Figures

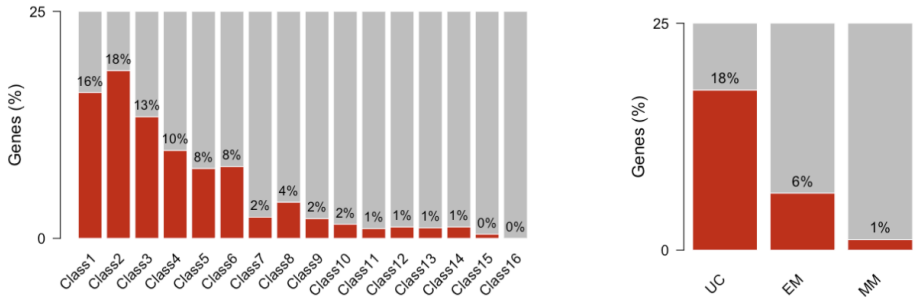
Supplementary Figure 1

A

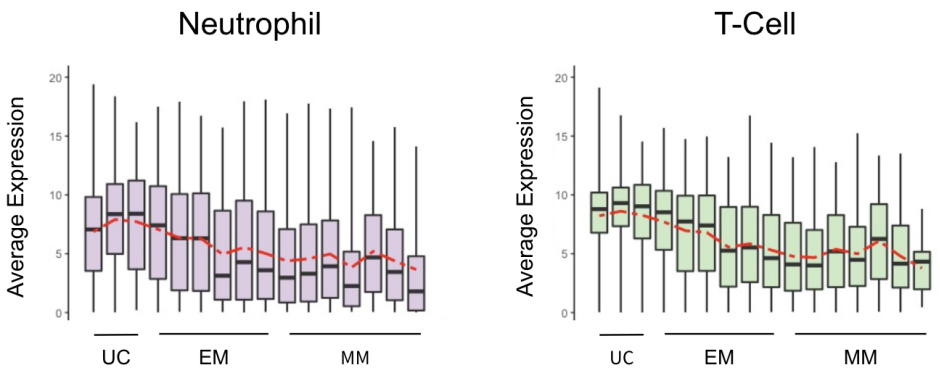


B

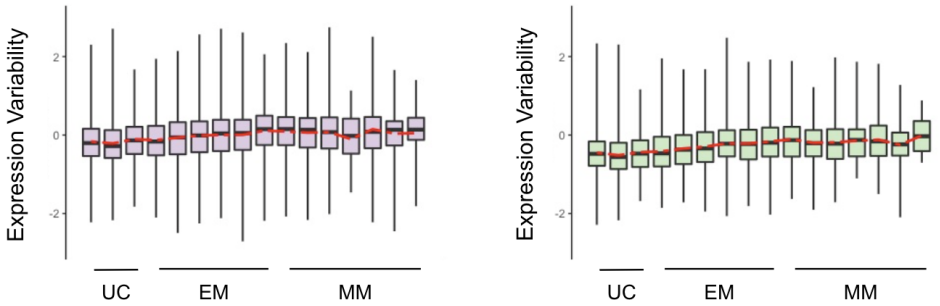
C



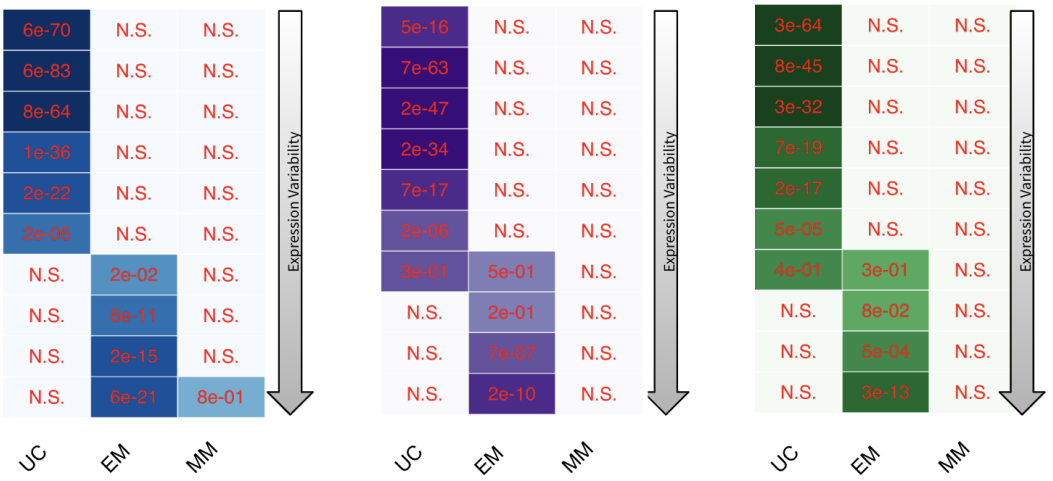
D

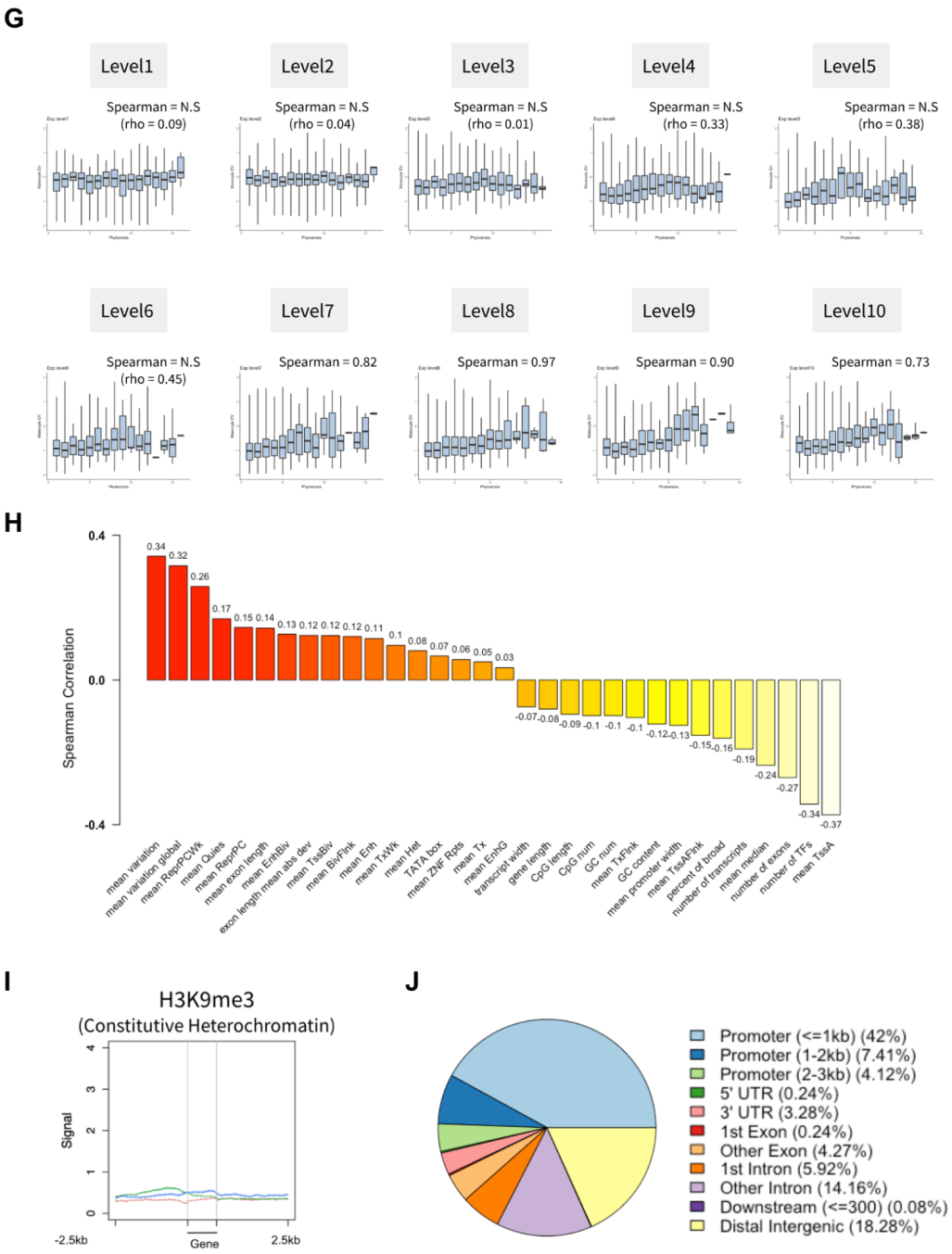

Neutrophil
T-Cell
Average Expression
20
15
10
5
0
UC
EM
MM

E


Expression Variability
2
0
-2
UC
EM
MM

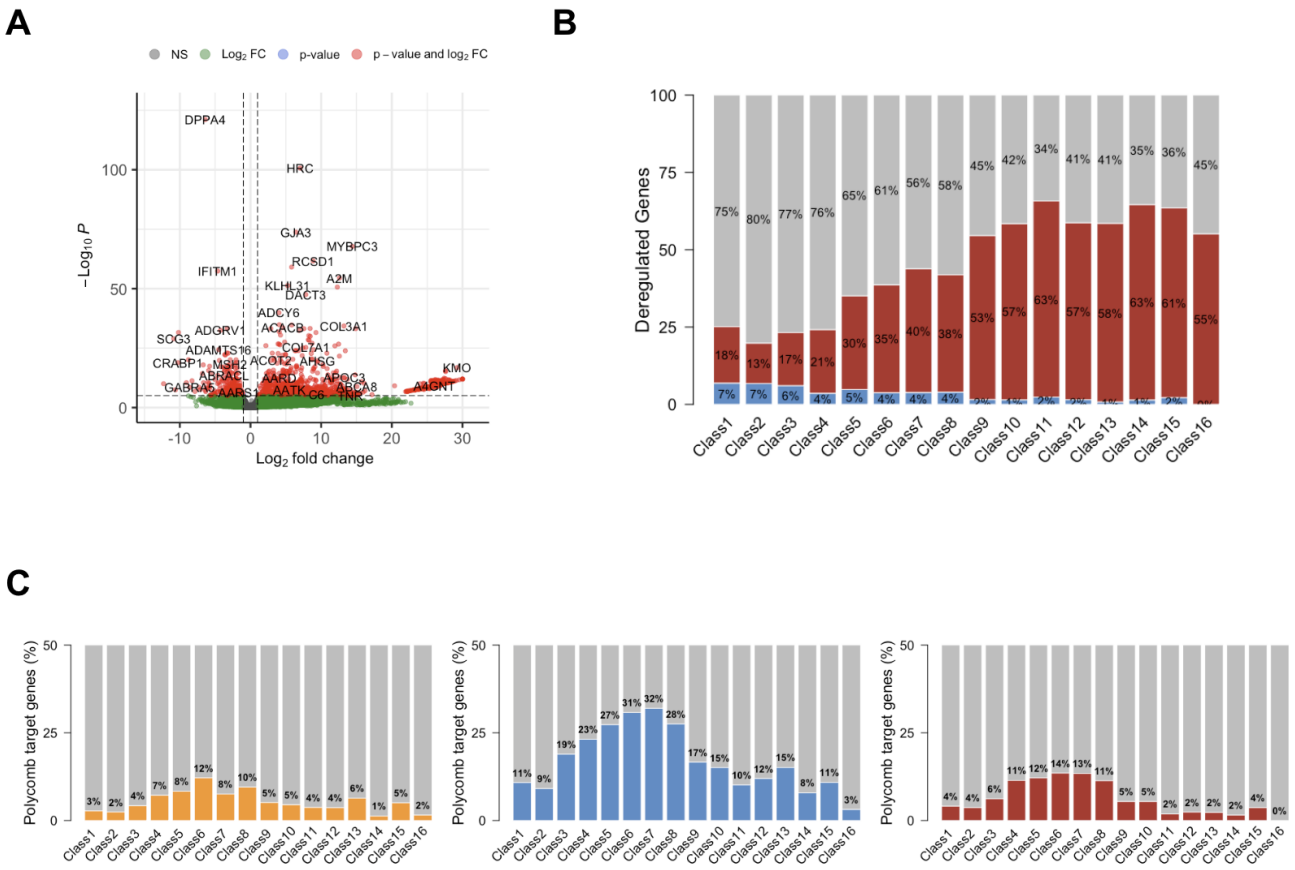
F


6e-70 N.S. N.S.
6e-83 N.S. N.S.
8e-64 N.S. N.S.
1e-36 N.S. N.S.
2e-22 N.S. N.S.
2e-05 N.S. N.S.
N.S. 2e-02 N.S.
N.S. 6e-11 N.S.
N.S. 2e-15 N.S.
N.S. 6e-21 8e-01
5e-16 N.S. N.S.
7e-63 N.S. N.S.
2e-47 N.S. N.S.
2e-34 N.S. N.S.
7e-17 N.S. N.S.
2e-06 N.S. N.S.
3e-01 5e-01 N.S.
N.S. 2e-01 N.S.
N.S. 7e-07 N.S.
N.S. 2e-10 N.S.
3e-64 N.S. N.S.
8e-45 N.S. N.S.
3e-32 N.S. N.S.
7e-19 N.S. N.S.
2e-17 N.S. N.S.
5e-05 N.S. N.S.
4e-01 3e-01 N.S.
N.S. 8e-02 N.S.
N.S. 5e-04 N.S.
N.S. 3e-13 N.S.
Expression Variability
UC
EM
MM



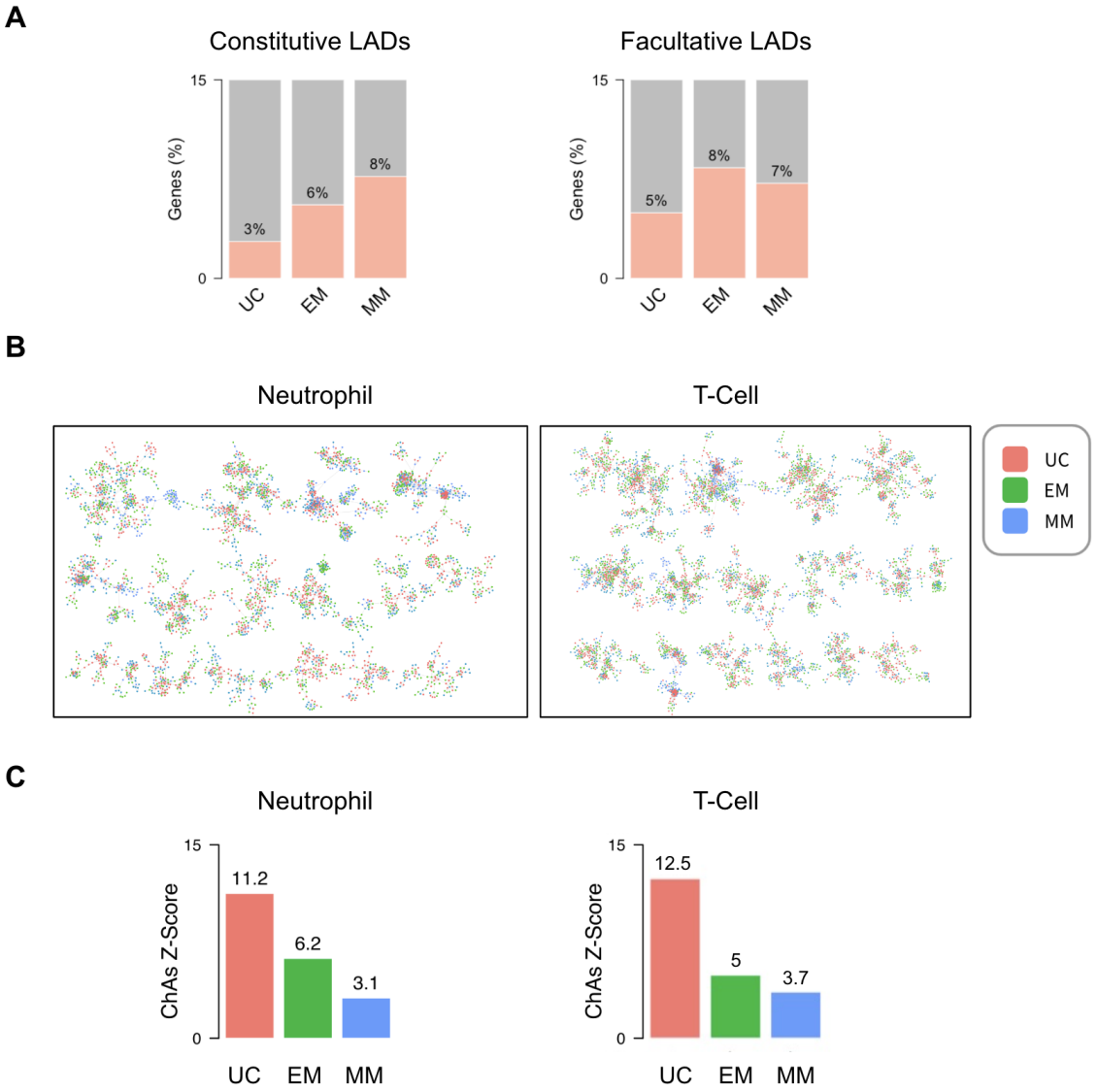
Supplementary Fig1. (related to Fig.1)
A) Number of genes in each phylostratum (left) and in each main age class (right). **B**) Gene Ontology enrichment of genes from each phylostratum. **C**) Proportion of housekeeping genes in each phylostratum (left) and in each main age class (right). **D**) Boxplot of average inter-individual gene expression levels in Neutrophil (left) and T-cell (right) across 16 gene age classes (Spearman test in neutrophil, Rho = -0.88, *p*-value < 2.2e-16. Spearman test in T-cell, Rho -0.85, *p*-value = 1.023e-05). **E**) Boxplots of inter-individual expression variability in Neutrophil (left) and T-cell (right) across 16 gene age classes (Spearman test in neutrophil, Rho = 0.74, *p*-value = 0.001703. Spearman test in T-cell, Rho = 0.84, *p*-value = 1.932e-05). **F**) Fisher's Exact Tests of 10 EV decile levels and 3 main gene ages in Monocyte, Neutrophil and T-cell, from left to right. Color shades show Jaccard index. **G**) Boxplots of inter-individual expression variability in Monocyte across the 16 phylostrata for 10 gene expression level classes. **H**) Spearman correlations of various DNA variability properties from Sigalova *et al*. with 16 gene age classes. **I**) Average H3K9me3 profile in Monocyte around the 3 main gene age classes (Red Unicellular genes, Green Early metazoan genes, Blue Mammal Specific genes). **J**) Genomic localisations of all detected CpG islands in monocyte.

Supplementary Figure 2



Supplementary Fig2. (related to Fig.2)
A) Volcano plot of differentially expressed genes in Cardiomyocyte compared to hESC. **B**) Proportion of Down (blue), Up (red) and no (grey) deregulated genes in Cardiomyocyte compared to hESC across the 16 gene age categories. **C**) Proportion of Polycomb target genes across the 16 gene age categories in hESC (left), B-cell (middle) and CLL (right).

Supplementary Figure 3



Supplementary Fig3. (related to Fig.3)
A) Overlaps with Lamina Associated Domains (LADs) annotated in multiple cell lines (Kind et al., 2015) for genes belonging to the different age classes. **B**) Visualizations of the 3D promoter-promoter interaction networks in neutrophil (left) and T-cell (right) with node color representing gene age classes (red = UC, green = EM, blue = MM). **C**) ChAs z-scores compared amongst the 3 age classes in neutrophil (left) and T-cell (right) promoter-promoter interaction networks.