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

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

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- 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 attempt to formalize a hypothesis according to which the topological properties of the chromatin network can impact phenotypes, such as plasticity and variability. Our observations lead us to reconsider the theory according to which cancer cells re-acquire pluripotency through a reactivation of early evolutionary programs characteristic of species lacking complex body plans or even multicellularity.
Recent work investigating gene variability across tissues and individuals in multiple species has shown that promoter sequence characteristics can be predictive of gene expression variability. 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 then investigate whether 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[1], 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[2,3], 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 lifes were sequenced and revealed that molecular pathways are remarkably well conserved[4]. 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[18].

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 [34]. Animals require a greater division of labor, represented by their greater number of cell types compared to plants, fungi, and algae[35]. 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[34]. If several species share this cancer phenotype, it becomes possible that some mechanisms are in common. As written above, 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[36]. 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[37]. Three chromatin-marks (DNA methylation, H3K36me3, and H3K27ac) in mammal ESCs are independent of DNA sequence evolution[37].

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[4,5], which enabled complex body plans. At the same time, species won the ability to regulate their cell cycles, modulate their growth and form complex networks of communication[6]. 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[6]. 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[7]. 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[6,8]. More interesting, cancer cells show expression of genes that are specific to embryogenesis and are not expressed in healthy adult tissues[9]. Indeed, it seems that epigenetic alterations occur very early in oncogenesis[10]. 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[11]. 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[12].

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[13]. Studies show that the variability in genome across various cell types and species is responsible for shaping innate immunity[14]. Moreover the variability of 3D architecture can be directly linked to hematological malignancies due to dysregulation of the lymphopoietic transcriptional program[15].
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[16].

Results and discussion

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[17] 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[18]. 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)(**Fig. 1A**)[7]. Throughout the manuscript we will consider either the 16 specific ages 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 (**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. 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.

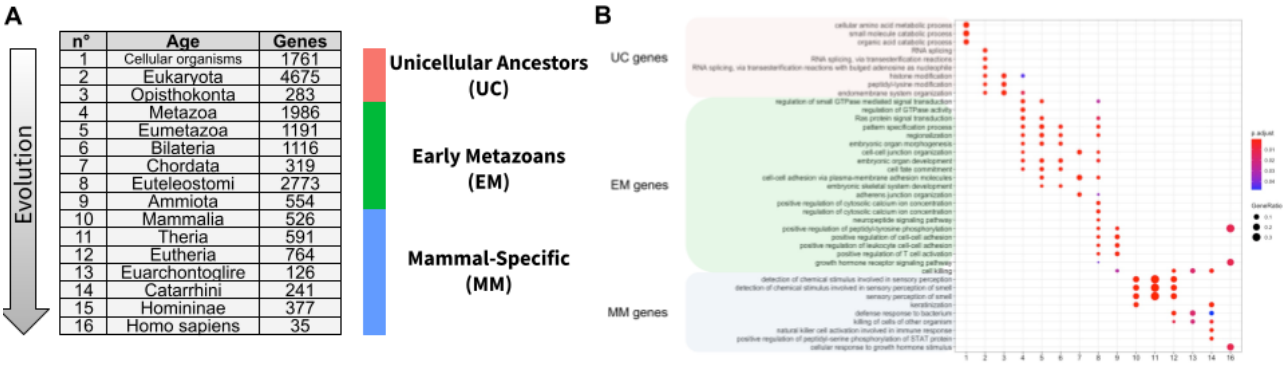

Figure 1: Overview of genes according to their ages and definition of main age categories. A) Number of genes in each age class. **B)** Functional enrichment of genes in different age classes.

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[19] and successively used by Ecker *et al*. to estimate the variability of gene expression across

individuals[20], 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[21]. 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) (**Fig. 2A**). 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[22], we next looked at the relation between expression variability and age (**Fig. 2B**). We found that gene expression variability is also dependent on gene ages, showing a lower variability for unicellular genes compared to newer genes.

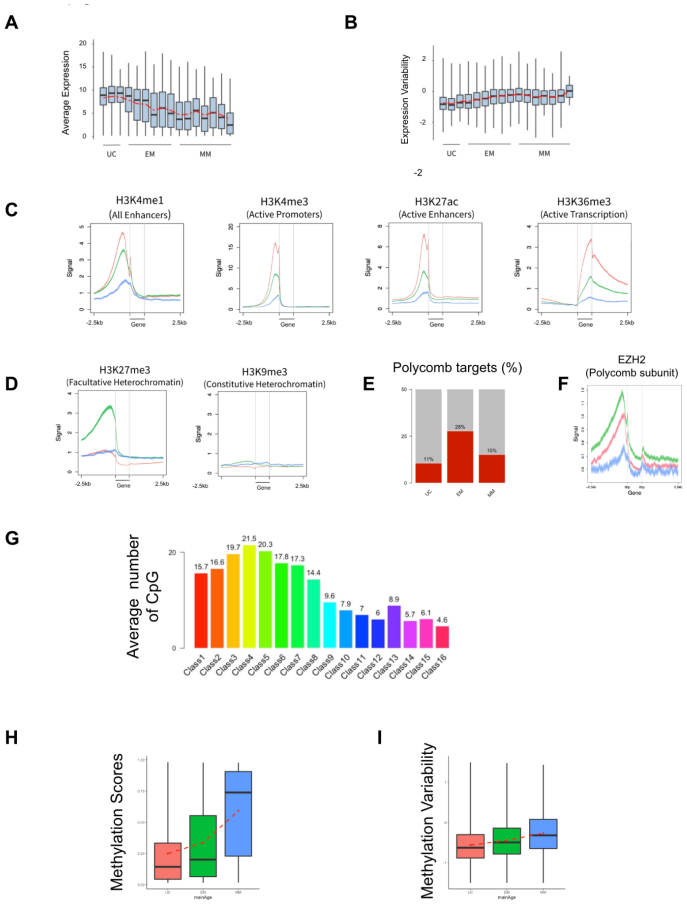

Figure 2: Genes of different evolutionary age classes display specific genetic and epigenomic features A) Boxplots of average inter-individual gene expression levels across 16 gene age classes. **B**) Boxplots of inter-individual expression variability across 16 gene age classes.

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 and H3K9me3 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 and **F**) Average EZH2 protein profile around the 3 main gene age classes (Red Unicellular genes, Green Early metazoan genes, Blue Mammal Specific genes) of several Transcription Factor categories (Architectural, Pluripotency, Pioneer, Polycomb). **G**) Average number of potential CpG islands by gene across the age categories. **H**) Methylation scores in monocytes across the three main

age classes. **I**) DNA methylation variability across the three main age classes.

To further test our findings on a separate dataset, we considered the features from Sigalova *et al*[23] 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 several active histone modifications in Monocyte. We analyze average signal profiles around genes from the 3 main gene age classes (**Fig. 2C**). We 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.

We also analyzed H3K27me3 and H3K9me3 histone modifications that are known to be repressive (**Fig. 2D**) with H3K27me3 being the mark associated with polycomb protein complexes. 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. This would suggest 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 compute the proportion of genes from each main age class that are enriched with a H3K27me3 peak on the promoter in Monocyte (**Fig. 2E**). 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). In order to confirm this specific behavior of polycomb proteins toward metazoan genes, we analyzed signal enrichment of the EZH2 binding profiles (obtained by ChIP-seq) around the 3 main gene classes in Monocyte (**Fig. 2F**). This polycomb subunit is a methyltransferase enzyme that participates in histone methylation and transcriptional repression[24]. Results show again a higher enrichment around genes from this specific age class.
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.

We finally decided to focus on DNA methylation as another important epigenetic feature. Chen *et al*'s dataset contains a methylation score for each potential CpG across the entire human genome in Monocyte across 125 healthy individuals[21]. Results show that half of all 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 . CpG islands are known to mostly repress gene expression when they are located in promoters. We find that the average number of potential CpG islands increases from gene age class 1 to 4 and then progressively decreases until class 16 (**Fig. 2G**).
We then analyze methylation scores computed by Chen *et al*. on Monocyte gene promoters depending on their respective ages (**Fig. 2H**). Results show a progressively increasing methylation through the 3 main evolutionary ages. This suggests a higher regulation by DNA methylation for the MM genes. We finally computed the inter-individual DNA methylation variability (**Fig. 2I**). This variability roughly linearly increases across all gene age classes, suggesting a more stable methylation profile for old genes, and a more variable methylation profile around the newest genes.
These findings suggest a progressive higher DNA methylation variability related to the new genes.

These first results showed that evolutionary gene ages are related to many 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 active histone modification. 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.

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.

Local gene expression regulatory processes in relation to evolutionary ages

We first asked what evolutionary ages are the most impacted by the change of gene expression profiles during cell differentiation and cancer. In order to study gene expression deregulation through cell differentiation, we analyzed a dataset of gene expression from hESC undifferentiated cells to differentiated cardiomyocytes[25]. We applied a differential gene expression analysis and computed proportions of UP and DOWN deregulated genes in each evolutionary age class (**Fig. 3A**). 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 with 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[7]. 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.

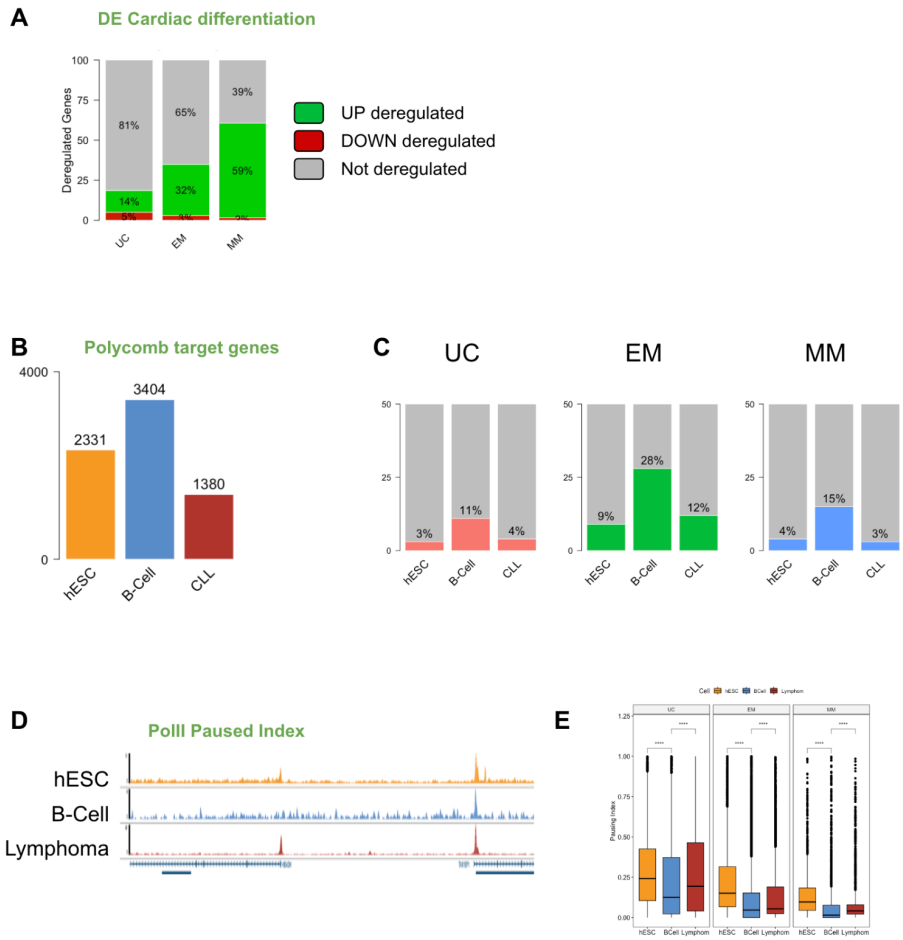

Figure 3. Local gene expression regulatory processes in relation to evolutionary ages
A) Proportion of Down (red) and Up (green) and no (grey) deregulated genes in Cardiomyocyte compared to hESC for the 3 main gene ages. **B**) Number of Polycomb target genes in hESC (yellow), B-Cell (blue) and CLL (red). **C**) Percentage of Polycomb target genes from the 3 cells in UC (left), EM (middle) and MM (right) gene age categories. **D**) Browser snapshots of PolII tracks in hESC (yellow), B-Cell (blue) and CLL (red). **E**) Paused PolII Index distributions from the 3 cells for UC (left), EM (middle) and MM (right) genes. Statistical comparisons between groups were done by the Wilcoxon-Mann-Whitney test. n.s = not significant, **** = p<0.0001.

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 stochasticity[26].

We computed the total number of polycomb targeted genes across the genome in each cell state (**Fig 3B**), showing an increasing number of targeted genes through cell differentiation and a strong decrease through oncogenesis. We then computed in which age categories we could find the polycomb targeted genes (**Fig. 3C**). Our previous results showed metazoan genes in the EM class were the most regulated by polycomb in Monocyte (c.f. **Fig. 2D, Fig. 2E, Fig. 2F**), and we found that this is also the case in stem cells, B-Cells and 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[27]. 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. 3D**). 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.

In this second part, we focused on gene expression and gene expression regulation related to evolutionary gene ages and across cell states. 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 compared to the stem state during differentiation and then decrease in cancer state. 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.

Taken together, our findings reinforce 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. Interestingly, 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[28]. Chromatin organization is known to be dramatically re-organised both through differentiation[29] and lineage definition, and in cancer, leading us to investigate whether gene age and expression variability might also be associated to localisation of genes in the 3D genome.

Evolutionary gene age and expression variability are strongly associated withgenome 3D structure

In previous research we and others have shown that many different epigenomic features such as many transcription factor binding peaks[30], gene expression[30], DNA methylation, histone modifications[31] and even replication timing and origins of replication[32] 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[30] more than would be expected based on genomic distances separating the contacting fragments.

Having observed a strong relationship between gene expression regulation and 3D chromatin structure led us to study the gene age classes and EV levels in a 3D context.

We constructed promoter-centered chromatin networks starting from PCHi-C datasets from Javierre *et al*.[33]. In these networks, each node is a DNA fragment and an edge represents a significant physical contact between two DNA fragments detected by the conformation capture protocol. We can either consider only contacts involving promoters (PP) or also contacts joining a promoter with a regulatory region, or other end (PO)[31].

We started by mapping gene ages on promoters on the monocyte PP subnetwork. As can be clearly seen in **Fig. 4A**, genes of specific gene age classes tend to form clusters, implying a significant assortativity of the gene age feature, which was confirmed by the ChAs z-scores for each age category being significantly above 2 (equivalent to p=0.05) (**Fig. 4B**) and highest for UC (ChAs Z-score =11.2).
We then calculated assortativity of the 3 broad gene age classes interpreted as categorical values and observed UC genes to be the most inter-connected genes (z-score=10.6) in monocytes, followed by EM (z-score=7.9) and MM (z-score=1.9). These findings suggest that the oldest unicellular genes are more highly connected and form clusters in the chromatin network, closely followed by EM genes who also form preferential clusters. New genes do form small clusters but their preferential interactions seem less important. These findings are consistent with the identification of polycomb group proteins as very highly assortative in mESCs (Pancaldi et al. 2016) and in monocytes, neutrophils and T cells (Madrid-Mencia et al. 2020).

Interestingly, we notice an increase in the distance between promoters and their regulatory regions from older to younger genes (**Fig. 4C**) and an increase of overlap between younger genes and Laminar Associated Domains (LADs) as annotated by Kind *et al*[34] (**Fig. 4D**). We further investigated whether the assortativity of gene age classes would change across differentiation and oncogenesis. As shown by **Fig. 4E**, in hESC cells UC genes are the most assortative with limited ChAs for EM and MM genes, in B cells EM genes greatly increase their ChAs reaching values similar to UC genes, while in cancer (CLL cells) the UC genes return to be the most assortative, with EM genes reducing their ChAs value.
As expected due to the relation between gene ages and EV, we also find that genes with similar EV values tend to form preferential contacts with each other (see method for details). As seen in **Fig. 4F,** genes in each of the EV categories are significantly assortative in monocytes, neutrophils and T cells, especially in the low and high expression range, suggesting that genes with the same expression variability level interact more together than with other genes.

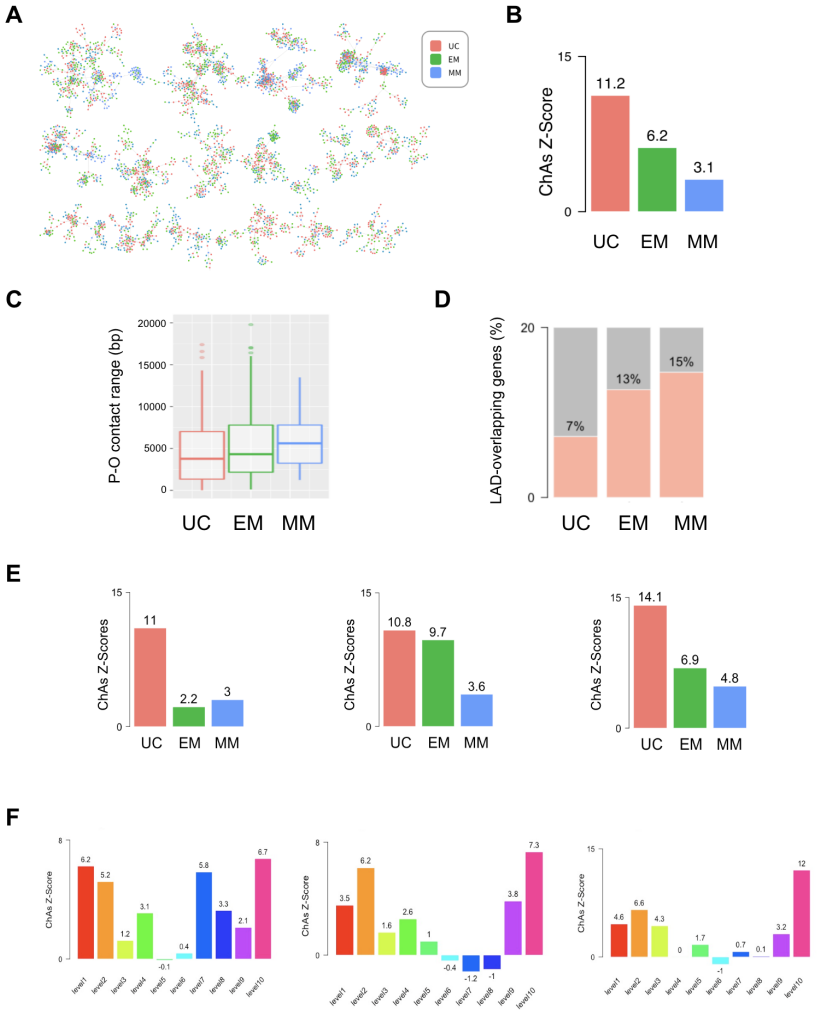

Figure 4: Relation between gene age, gene expression variability and 3D structure.

A) 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).
B) Assortativity z-scores compared amongst the 3 age classes in monocyte promoter-promoter interaction network.
C) Genomic distance spanned by Promoter-Other end (PO) contacts of genes in the different age classes, from hESC PCHi-C data.
D) Overlap with Lamina Associated Domains (LAD) annotated in multiple cell lines[34] for genes belonging to the different age classes.
E) Comparison of ChAs z-score of genes from different gene classes between hESCs, B cells and CLL B cells.
F) Comparison of assortativity z-scores of genes in each EV level across monocytes, neutrophils and T cells (from left to right).

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[31], we showed a relationship between gene age and chromatin structure, which goes beyond existing observations relating gene ages to the TAD structure[39,40]. 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[30]. 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, suggesting that 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 variabilty, gene regulation characteristics and age, with variability decreasing with gene age, we surmise that the reshaping of the organisation 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.

Material and Methods

Gene Ages

We used Gene ages dataset from Trigos public published work which has estimated gene ages by phylostratigraphy[7]. 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 published work[20]. Variability has been computed from 125 healthy individuals with method accounted for confounding effects due to the correlation between mean and variability measurements. Fisher exact tests have been computed by using geneOverlap (v1.32.0) R package[41].

ChIP-Seq Processing

Histone modifications and proteins ChIP-Seq datasets in Monocyte have been downloaded from the ENCODE portal (https://www.encodeproject.org/)[42] in bigwig format with identifiers summarized in the following table:

Target	Cell	Identifier
H3K4me1	Monocyte	ENCFF929UOK (FoldChange over control)
H3K4me3	Monocyte	ENCFF231OTU
H3K27ac	Monocyte	ENCFF437JSB
H3K36me3	Monocyte	ENCFF064OOI
H3K27me3	Monocyte	ENCFF557YRZ
H3K9me3	Monocyte	ENCFF256KXQ
EZH2	B-Cell (GM12878)	ENCFF410SZR
RNA-Seq	hESC	GSE69618
RNA-Seq	CM	GSE69618
H3K27me3	hESC (h9 /WA09)	SRR11185316
Input	hESC (h9 /WA09)	SRR11185368
H3K27me3	B-Cell	ENCFF864CRK
Input	B-Cell	ENCFF000AMM
H3K27me3	CLL B-Cell	Blueprint
Input	CLL B-Cell	Blueprint
PolII	hESC (h9 /WA09)	SRR7741345
PolII	B-Cell (GM12878)	ENCFF000OBB
PolII	Lymphoma B-Cell (LY1)	GSM1133652

EZH2 polycomb subunit ChIP-Seq data on GM12878 B-Cell cell line have been downloaded from the ENCODE portal (https://www.encodeproject.org/) in bigwig format with ENCFF410SZR identifier. Sequenced reads were mapped to the human reference genome (built hg19/GRCh37) by using Burrows-Wheeler Aligner bwa (v0.7.17-r1188)[43]. Bam files were obtained, sorted and indexed by using samtools utilities (v1.15.1)[44]. ChIP-Seq peak calling was done using MACS2 (v.2.2.7.1) with respective inputs[45]. H3K27me3 peak calling was done with –broad parameter on.

ChIP-Seq Analysis

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. Super-Enhancer associated genes were considered as genes interacting with a SE overlapping fragment according to PCHi-C chromatin networks. ChIP-Seq profiles were computed with the seqplots (v1.23.3) R package[46]. CpG localizations were studied using the ChIPSeeker (v1.32.1) R package[47].

Differential gene expression 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[48] (v3.52.1) and clusterProfiler[49] (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 indeces across cells, we normalized each distribution such that the normalized data falls between 0 and 1.

PCHi-C chromatin networks

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

Cell	Publication
hESC	Freire-Pritchett *et al*[29]
B-Cell	Javierre *et al*[33]
CLL	Beekman et al, 2018[50]

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[51]. We computed Chromatin Assortativity by using the ChAseR (v0.0.0.9) R package available on bitbucket[31]. Gene ages and EV level 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

1. Hanahan, D. Hallmarks of Cancer: New Dimensions. *Cancer Discov.* **12**, 31–46 (2022).
2. Boveri, T. Concerning the Origin of Malignant Tumours by Theodor Boveri. Translated and annotated by Henry Harris. *J. Cell Sci.* **121**, 1–84 (2008).
3. Busund, L.-T. & Rasmussen. 10. Hypotheses of Carcinogenesis—The AtavisticTheory. in *Advancing Systems Epidemiology in Cancer* (Universitetsforlaget, 2020). doi:10.18261/9788215041193-2020-10.
4. Domazet-Lošo, T. & Tautz, D. Phylostratigraphic tracking of cancer genes suggests a link to the emergence of multicellularity in metazoa. *BMC Biol.* **8**, 66 (2010).
5. Zhang, L., Tan, Y., Fan, S., Zhang, X. & Zhang, Z. Phylostratigraphic analysis of gene co-expression network reveals the evolution of functional modules for ovarian cancer. *Sci. Rep.* **9**, 2623 (2019).
6. Benton, M. L. *et al.* The influence of evolutionary history on human health and disease. *Nat. Rev. Genet.* **22**, 269–283 (2021).
7. Trigos, A. S., Pearson, R. B., Papenfuss, A. T. & Goode, D. L. 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 (2017).
8. Bussey, K. J., Cisneros, L. H., Lineweaver, C. H. & Davies, P. C. W. Ancestral gene

regulatory networks drive cancer. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 6160–6162 (2017).

9. Schoenhals, M. *et al.* Embryonic stem cell markers expression in cancers. *Biochem. Biophys. Res. Commun.* **383**, 157–162 (2009).
10. Patani, H. *et al.* Transition to naïve human pluripotency mirrors pan-cancer DNA hypermethylation. *Nat. Commun.* **11**, 3671 (2020).
11. Bonev, B. & Cavalli, G. Organization and function of the 3D genome. *Nat. Rev. Genet.* **17**, 772–772 (2016).
12. Pancaldi, V. Network models of chromatin structure. *Curr. Opin. Genet. Dev.* **80**, 102051 (2023).
13. Chaicharoenaudomrung, N., Kunhorm, P. & Noisa, P. Three-dimensional cell culture systems as an *in vitro* platform for cancer and stem cell modeling. *World J. Stem Cells* **11**, 1065–1083 (2019).
14. Hagai, T. *et al.* Gene expression variability across cells and species shapes innate immunity. *Nature* **563**, 197–202 (2018).
15. Scourzic, L., Salataj, E. & Apostolou, E. Deciphering the Complexity of 3D Chromatin Organization Driving Lymphopoiesis and Lymphoid Malignancies. *Front. Immunol.* **12**, 669881 (2021).
16. Pancaldi, V. Chromatin Network Analyses: Towards Structure-Function Relationships in Epigenomics. *Front. Bioinforma.* **1**, 742216 (2021).
17. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
18. Domazet-Lošo, T., Brajković, J. & Tautz, D. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet.* **23**, 533–539 (2007).
19. Alemu, E. Y., Carl, J. W., Corrada Bravo, H. & Hannenhalli, S. Determinants of expression variability. *Nucleic Acids Res.* **42**, 3503–3514 (2014).
20. BLUEPRINT Consortium *et al.* Genome-wide analysis of differential transcriptional and epigenetic variability across human immune cell types. *Genome Biol.* **18**, 18 (2017).

21. Chen, L. *et al.* Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell* **167**, 1398-1414.e24 (2016).

22. Ecker, S., Pancaldi, V., Valencia, A., Beck, S. & Paul, D. S. Epigenetic and Transcriptional Variability Shape Phenotypic Plasticity. *BioEssays* **40**, 1700148 (2018).

23. Sigalova, O. M., Shaeiri, A., Forneris, M., Furlong, E. E. & Zaugg, J. B. Predictive features of gene expression variation reveal mechanistic link with differential expression. *Mol. Syst. Biol.* **16**, e9539 (2020).

24. Piunti, A. & Shilatifard, A. The roles of Polycomb repressive complexes in mammalian development and cancer. *Nat. Rev. Mol. Cell Biol.* **22**, 326–345 (2021).

25. Choy, M.-K. *et al.* Promoter interactome of human embryonic stem cell-derived cardiomyocytes connects GWAS regions to cardiac gene networks. *Nat. Commun.* **9**, 2526 (2018).

26. Kar, G. *et al.* Flipping between Polycomb repressed and active transcriptional states introduces noise in gene expression. *Nat. Commun.* **8**, 36 (2017).

27. Adelman, K. & Lis, J. T. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nat. Rev. Genet.* **13**, 720–731 (2012).

28. Schoenfelder, S. *et al.* Polycomb repressive complex PRC1 spatially constrains the mouse embryonic stem cell genome. *Nat. Genet.* **47**, 1179–1186 (2015).

29. Freire-Pritchett, P. *et al.* Global reorganisation of cis-regulatory units upon lineage commitment of human embryonic stem cells. *eLife* **6**, e21926 (2017).

30. Pancaldi, V. *et al.* Integrating epigenomic data and 3D genomic structure with a new measure of chromatin assortativity. *Genome Biol.* **17**, 152 (2016).

31. Madrid-Mencía, M., Raineri, E., Cao, T. B. N. & Pancaldi, V. Using GARDEN-NET and ChAseR to explore human haematopoietic 3D chromatin interaction networks. *Nucleic Acids Res.* **48**, 4066–4080 (2020).

32. Jodkowska, K. *et al.* 3D chromatin connectivity underlies replication origin efficiency in mouse embryonic stem cells. *Nucleic Acids Res.* **50**, 12149–12165 (2022).

33. Javierre, B. M. *et al.* Lineage-Specific Genome Architecture Links Enhancers and

Non-coding Disease Variants to Target Gene Promoters. *Cell* **167**, 1369-1384.e19 (2016).

34. Kind, J. *et al.* Genome-wide Maps of Nuclear Lamina Interactions in Single Human Cells. *Cell* **163**, 134–147 (2015).

35. Athena Aktipis, C. *et al.* Cancer across the tree of life: Cooperation and cheating in multicellularity. *Philos. Trans. R. Soc. B Biol. Sci.* **370**, (2015).

36. Bell, G. & Mooers, A. O. Size and complexity among multicellular organisms. *Biol. J. Linn. Soc.* **60**, 345–363 (1997).

37. Endo, Y., Kamei, K. & Inoue-Murayama, M. Genetic Signatures of Evolution of the Pluripotency Gene Regulating Network across Mammals. *Genome Biol. Evol.* **12**, 1806–1818 (2020).

38. Xiao, S. *et al.* Comparative Epigenomic Annotation of Regulatory DNA. *Cell* **149**, 1381–1392 (2012).

39. Acemel, R. D. & Lupiáñez, D. G. Evolution of 3D chromatin organization at different scales. *Curr. Opin. Genet. Dev.* **78**, 102019 (2023).

40. James, C., Trevisan-Herraz, M., Juan, D. & Rico, D. *Vertebrate whole genome duplications shaped the current 3D genome architecture*. http://biorxiv.org/lookup/doi/10.1101/2021.06.11.448047 (2021) doi:10.1101/2021.06.11.448047.

41. Li Shen, M. S. <Shenli S. C. GeneOverlap. (2017) doi:10.18129/B9.BIOC.GENEOVERLAP.

42. Sloan, C. A. *et al.* ENCODE data at the ENCODE portal. *Nucleic Acids Res.* **44**, D726–D732 (2016).

43. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).

44. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

45. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).

46. Stempor, P. & Ahringer, J. SeqPlots - Interactive software for exploratory data analyses, pattern discovery and visualization in genomics. *Wellcome Open Res.* **1**, 14 (2016).

47. Yu, G., Wang, L.-G. & He, Q.-Y. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* **31**, 2382–2383 (2015).

48. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47–e47 (2015).

49. Wu, T. *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation* **2**, 100141 (2021).

50. Beekman, R. *et al.* The reference epigenome and regulatory chromatin landscape of chronic lymphocytic leukemia. *Nat. Med.* **24**, 868–880 (2018).

51. Csárdi, G. *et al.* igraph for R: R interface of the igraph library for graph theory and network analysis. (2023) doi:10.5281/ZENODO.7682609.