Long-range promoter-enhancer contacts are conserved during evolution and contribute to gene expression robustness

Alexandre Laverré[1*], Eric Tannier[1,2], Anamaria Necsulea[1,*]

[1] Univ Lyon, Université Claude Bernard Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive, F-69100, Villeurbanne, France.

[2] INRIA Grenoble Rhône-Alpes, Montbonnot, France

[*] Corresponding authors:
alexandre.laverre@univ-lyon1.fr
anamaria.necsulea@univ-lyon1.fr

Abstract

Gene expression is regulated through complex molecular interactions, involving *cis*-acting elements that can be situated far away from their target genes. Data on long-range contacts between promoters and regulatory elements is rapidly accumulating. However, it remains unclear how these regulatory relationships evolve and how they contribute to the establishment of robust gene expression profiles. Here, we address these questions by comparing genome-wide maps of promoter-centered chromatin contacts in mouse and human. We show that there is significant evolutionary conservation of *cis*-regulatory landscapes, indicating that selective pressures act to preserve regulatory element sequences and their interactions with target genes. The extent of evolutionary conservation is remarkable for long-range promoter-enhancer contacts, illustrating how the structure of regulatory interactions constrains large-scale genome evolution. Notably, we show that the evolution of *cis*-regulatory landscapes, measured in terms of distal element sequences, synteny or contacts with target genes, is tightly linked to gene expression evolution.

Keywords : *cis*-regulatory landscapes, gene expression evolution, large-scale genome evolution, Promoter Capture Hi-C

Introduction

The evolution of gene expression and the evolution of expression regulatory mechanisms have attracted considerable attention ever since the proposal that phenotypic differences between species may be driven by changes in gene activity rather than by changes in gene

products (King and Wilson, 1975). In the past decade, these two topics have been extensively scrutinized through comparative “omics” studies, facilitated by high-throughput sequencing technologies (Brawand et al., 2011; Cardoso-Moreira et al., 2019; Gilad et al., 2006; Khaitovich, 2005). These studies showed that gene expression patterns are generally well conserved across species, while the sequences and the activities of regulatory elements (in particular those of expression enhancers) evolve rapidly. These paradoxical observations warrant further exploration, to better understand the determinants of the gene expression robustness in the presence of rapidly evolving regulatory landscapes. However, so far only few attempts were made to directly connect the evolution of gene expression to the evolution of the underlying regulatory mechanisms, at a genome-wide scale (Berthelot et al., 2018; Wong et al., 2017). These studies proposed that the presence of complex regulatory landscapes, involving numerous expression enhancers with potentially redundant roles, is the key driver of the robustness of gene expression levels (Berthelot et al. 2018).

To date, identifying the regulatory elements that control each gene remains a challenging task. Transcriptional activity is regulated through complex interactions between *trans*-acting factors and *cis*-acting regulatory elements. Genes are typically associated with multiple *cis*-acting elements, which can refine their expression level, control different expression domains, or confer expression robustness through partial redundancy (Kvon et al., 2021). Conversely, each regulatory element can influence multiple genes, either concomitantly or in a context-specific manner (Schoenfelder and Fraser, 2019). The prevalence of long-distance associations between genes and regulatory elements further complicates the picture. There is now extensive evidence that *cis*-acting regulatory elements can be located millions of base pairs away from their target genes (Montavon et al., 2011; Smemo et al., 2014). The existence of such long-range interactions challenges an assumption commonly made in the literature, namely that the targets of *cis*-regulatory elements are the neighboring genes in the genome, within a certain genomic distance (Berthelot et al., 2018; McLean et al., 2010; Villar et al., 2015; Wong et al., 2017).

A new dimension of the complexity of *cis*-acting regulatory landscapes can now be perceived thanks to the development of chromatin conformation capture techniques, which identify pairs of genomic segments found in physical proximity in the nucleus (Dekker et al., 2002; Schoenfelder et al., 2015; Zhao et al., 2006). These techniques can reveal long-range chromatin contacts between promoters and distal regulatory elements, which are needed to modulate gene expression (de Laat and Duboule, 2013). The accumulation of chromatin contact datasets available for multiple species offers the opportunity to study the evolution of *cis*-regulatory interactions, at a genome-wide scale, based on experimental data. Here, we

perform a comparative analysis of *cis*-regulatory landscapes, based on high-resolution promoter-centered chromatin interaction maps for human and mouse (Choy et al., 2018; Comoglio et al., 2018; Freire-Pritchett et al., 2017; Javierre et al., 2016; Koohy et al., 2018; Mifsud et al., 2015; Novo et al., 2018; Pan et al., 2018; Rubin et al., 2017; Schoenfelder et al., 2015, 2018; Siersbæk et al., 2017). By contrasting this data with simulated regulatory landscapes, we show that there is significant evolutionary conservation of promoter-enhancer chromatin contact maps, even at large genomic distances, suggesting that genomic rearrangements that perturb regulatory interactions are counter-selected. Using transcriptome sequencing data from multiple organs and developmental stages for human and mouse, we show that the rate of regulatory landscape evolution is significantly and positively correlated with the rate of gene expression evolution. We therefore provide decisive elements towards understanding the constraints imposed by the three-dimensional structure of *cis*-regulatory landscapes on genome evolution, and into the consequences of regulatory landscape changes on gene expression evolution.

Results

Promoter Capture Hi-C data includes pre-formed chromatin contacts

To examine the evolution of *cis*-regulatory landscapes, we processed and analyzed promoter capture Hi-C (PCHi-C) data derived from 16 human cell types and 7 mouse cell types (Methods, Supplementary Table 1). To evaluate the biological significance of the observations obtained with this dataset, we simulated interactions that mimic the numbers of contacts and the distribution of distances between contacting regions (Methods). We first verified to what extent chromatin interactions are shared among cell types. Overall, 56% of all interactions are detected in at least 2 cell types in human, compared to only 27% in the corresponding simulated data (Figure 1a, chi-squared test p-value < 1e-10). The enrichment compared to simulated data is even stronger for interactions that occur at large genomic distances (>=500kb), which are shared among cell types in 48% of the cases, compared to 10% in simulated data (Figure 1b, chi-squared test p-value < 1e-10). We obtained similar results for mouse (Supplementary Material). Within each species, chromatin contact landscapes cluster by cell type (Figure 1c, Supplementary Figure 1). Moreover, we confirm the presence of pre-formed chromatin interactions, that is, interactions that are detected in most PCHi-C samples even in the absence of the gene expression. This is illustrated by the interaction landscape of the developmental gene *SHH,* which contacts enhancers situated in the introns of the neighboring *LMBR1* gene (Lettice et al. 2003) (Figure 1c). Overall, genes that are involved in chromatin interactions shared across cell types are not necessarily ubiquitously expressed (Supplementary Material), confirming that contacts can be detected

even in the absence of transcription (Schoenfelder et al., 2015).

Thus, these PCHi-C datasets provide a broad overview of promoter-centered chromatin interactions mainly shared across cell types, including pre-formed contacts and thus extending beyond regulatory interactions that function exclusively in the sampled cell types. Although similar cell types were not always available for human and mouse, we are therefore confident that this dataset is well-suited for between-species comparisons. Furthermore, we could analyze three cell types (embryonic stem cells, B lymphocytes and pre-adipocytes) that were sampled for both species, thus further ensuring data comparability.

PCHi-C chromatin contacts describe regulatory landscapes

The promoter-centered chromatin contacts defined with PCHi-C data are known to be enriched in regulatory interactions (Schoenfelder et al., 2015). We validated this observation by jointly analyzing PCHi-C contact maps and genome-wide enhancer predictions (Methods). For all enhancer datasets, the average restriction fragment length that is covered by enhancers is significantly higher in the PCHi-C data than in the simulated data (Figure 2a, Wilcoxon rank sum test, p-value < 1e-10). For example, for the ENCODE enhancer dataset, the average fraction of the restriction fragment length that is covered by enhancers is 3.5% in the PCHi-C data, compared to 2.3% in the simulated data. Moreover, the proportion of restriction fragments that overlap with at least one enhancer is also significantly higher in the PCHi-C data than in simulations. For the ENCODE dataset, these proportions are 35% and 26% in the PCHi-C data and in the simulations, respectively (chi-squared test, p-value < 1e-10). The overlap with enhancers decreases when the distance between contacting regions increases, for both real and simulated PCHi-C data (Figure 2b). The number of cell types in which an interaction is observed is also positively correlated with the presence of enhancers in the restriction fragments contacted by promoters (Figure 2c). Similar results were obtained for other predicted enhancer datasets, for both species (Supplementary Material).

To further test the presence of regulatory interactions, we evaluated the correlations between gene expression and enhancer activity across samples (Methods). Due to the complexity of gene regulatory mechanisms, as well as due to inherent noise in activity measurements, these correlations are expected to be weak, even for genuine regulatory interactions (Hait et al., 2018). Nevertheless, we found that the activity levels of genes and ENCODE-predicted enhancers connected in PCHi-C data are significantly more correlated than expected by chance, as estimated with our simulated data (mean correlation coefficient 0.08 in the PCHi-C dataset and 0.04 in the simulations, Wilcoxon rank sum test p-value <

1e-10). We note that for both datasets the correlations decrease with the distance between promoters and enhancers, which could suggest better associations between genes and enhancers found in close vicinity in the gene (Figure 2d). However, the difference between PCHi-C data and simulations increases with the distance (Figure 2d). Similar results were obtained using other enhancer datasets for both human and mouse (Supplementary Material).

Taken together, these results confirm that the gene-enhancer pairs connected by PCHi-C data are enriched in genuine regulatory interactions. These chromatin contact maps thus provide a good basis to study *cis*-regulatory landscapes, at a genome-wide scale.

Genomic sequences contacted by promoters are conserved during evolution

We analysed the sequence conservation level of restriction fragments and enhancers contacted by promoters by calculating the length of the aligned sequence in nine vertebrate genomes (Methods, Figure 3a). We masked exonic regions for this analysis (Methods). The median fraction of contacted fragments aligned between human and mouse is 27% in PCHi-C data, which is significantly higher than the 23% observed in the simulated dataset (Wilcoxon rank sum test, p-value < 1e-10, Figure 3b). This observation holds for all pairwise comparisons between vertebrates, although conservation scores are expectedly weak for comparisons between divergent species. However, the enhancer elements that are contacted by gene promoters in our dataset are not significantly more conserved than other predicted enhancers (e.g., median aligned fraction 64.9% for PCHi-C data, 64.2% for simulated data, for human and mouse, Wilcoxon rank sum test p-value 0.75, Figure 3c), which may be missing from our interaction dataset because they regulate other genes or function in other cell types. Thus, the higher evolutionary conservation of PCHi-C contacted fragments compared to simulated data is not explained by overlap with better conserved enhancers.

Interestingly, the extent of sequence conservation increases with the distance from gene promoters, for contacted restriction fragments as well as for enhancers (Figure 3d,e, Supplementary Material). This observation could be in part explained by a higher rate of spurious contacts among regions separated by small genomic distances, despite statistical corrections implemented to account for this effect (Cairns et al., 2016). However, the distance between promoters and contacted regions or enhancers also co-varies with other relevant factors, such as the overlap with repetitive sequences and with the gene density, which are themselves correlated (Supplementary Figures 2,3). The overlap with repetitive

elements explains most of the difference in sequence conservation levels between PCHi-C and simulated data (Figure 3f,g). Remarkably, the sequence conservation levels of restriction fragments and enhancers are negatively correlated with the number of genes found within a 500 kb window (Figure 3f,g). In gene-dense regions, restriction fragments and enhancers are significantly more conserved for PCHi-C data than for simulated data, irrespective of their overlap with repeats (Figure 3f,g). However, this is not true for gene-poor regions (Figure 3f,g).

These results illustrate the complexity of the factors affecting sequence conservation levels for *cis*-regulatory elements. The frequency of repetitive elements seems to explain the difference in sequence conservation levels between observed and simulated PCHi-C data. We note that this may in itself be an indication of purifying selective pressures acting on genuine regulatory elements. Indeed, the most repeat-poor regions in the human and mouse genomes are the HOX gene clusters, which are crucial for embryonic development (International Human Genome Sequencing Consortium, 2001). Likewise, the negative association between sequence conservation levels and gene density may be explained by the presence of functionally constrained “gene deserts”, rich in regulatory elements (Ovcharenko et al., 2005). Consistent with these observations, genes that are in contact with highly conserved enhancers are enriched in functional categories related to developmental processes (Supplementary Table 2).

Pairs of promoters and enhancers involved in chromatin contacts are maintained in synteny

Functional interactions between gene promoters and regulatory elements are thought to occur predominantly in *cis*, at linear genomic distances within the Mb-scale (de Laat and Duboule, 2013). Thus, genomic rearrangements that separate promoter-enhancer pairs to different chromosomes or to contact-prohibiting distances are expected to be counter-selected. To test this hypothesis, we assessed the proportion of promoter-enhancer pairs that are maintained in synteny (on the same chromosome and within a maximum distance of 2Mb) through pairwise comparisons between human/mouse and other vertebrate species. We focused on genes that are 1-to-1 orthologous between species and on enhancers for which putative homologues could be predicted in the target species based on sequence conservation (Methods). We further restricted this analysis to promoters-enhancers pairs that are separated by distances between 100 kb and 1.5 Mb in the reference species (Methods). With this convention, synteny "breaks" should be interpreted as evolutionary events that add at least 500 kb to the distance between promoters and enhancers. We found

that the pairs of promoters-enhancers that are in contact in our PCHi-C dataset are maintained in synteny significantly more than expected based on the simulated dataset (Figure 4a, Supplementary Material). For example, 96% of pairs between promoters and ENCODE enhancers are maintained in synteny between human and mouse, compared to 94.8% in the simulated dataset (Figure 4a, chi-squared test p-value < 1e-10). At larger evolutionary distances, there is no significant difference between the PCHi-C data and the simulated interactions (chi-squared test p-value = 0.1 for the comparison between human and chicken, Figure 4a). The excess of synteny conservation compared to simulated data is mainly visible at large genomic distances, which are likely to accumulate genomic rearrangements with time if these are not counter-selected (Figure 4b,c, Supplementary Material).

These observations are consistent with the presence of a selective pressure against major rearrangements that would perturb interactions between genes and *cis*-regulatory elements, as previously hypothesised (Mongin et al. 2009; Lemaitre et al. 2009).

Promoter-enhancer contact maps are conserved during evolution

While breaks of synteny are likely to preclude the formation of chromatin contacts, conserved synteny does not guarantee that promoter-enhancer interactions are functionally conserved. To further test the presence of selective pressures to maintain chromatin contacts between promoters and enhancers, we directly compared PCHi-C interaction landscapes between human and mouse. We tested for contact conservation between pairs of cell types in human and mouse (Methods). For all cell types pairs, the frequency of conserved contacts is higher in the PCHi-C data than in the simulated data and particularly high for more comparable cell types (embryonic stem cells, pre-adipocytes and B-lymphocytes) (Figure 5a, Supplementary Material). The extent of promoter-enhancer contact conservation is positively correlated with the number of cell types in which the contact is detected (Figure 5b). This result is expected, given that the PCHi-C datasets do not represent the same cell types in human and mouse, and thus cell type-specific regulatory interactions may appear as species-specific.

For both PCHi-C and simulated data, the proportion of conserved contacts between promoters and enhancers decreases as the genomic distance between the two increases (Figure 5c, Supplementary Material), as expected given the over-representations of contacts at relatively short distances in both species (Supplementary Material). However, here again, the excess of contact conservation compared to simulated data is stronger at large genomic distances: around 1 Mb, the proportion of conserved contacts is less than 0.5% for simulated

7

data, while for PCHi-C data the value is 14% (Figure 5c).

We note that there is a strong enrichment for functional categories related to developmental processes among genes with highly conserved promoter-enhancer contact maps (Supplementary Table 3). This observation is consistent with the presence of strong functional constraints on the *cis*-regulatory landscapes of developmental genes.

The complexity of *cis*-acting regulatory landscapes is associated with gene expression characteristics and with the rates of gene expression evolution

To better understand the functional relevance and the phenotypic implications of the promoter-enhancer interactions predicted with PCHi-C data, we examined their relationship with the patterns of gene expression evolution. We evaluated gene expression patterns using a recently published comparative transcriptome collection spanning five major organs and numerous developmental stages (Cardoso-Moreira et al., 2019). This dataset allowed us to measure the breadth of gene expression and to identify changes in expression profiles, such as changes in organ or developmental stage “preference”, the gain or loss of an expression domain, etc. We measured the extent of expression conservation through the Spearman correlation coefficient between relative expression values for pairs of orthologous genes (Methods, Supplementary Material). Similar results were obtained when contrasting orthologous gene expression profiles with the Euclidean distance (Methods, Supplementary Figure 4, Supplementary Material). Interestingly, genes with the highest expression profile conservation levels are enriched in processes related to expression regulation (Supplementary Table 4).

We observe that genes that are in contact with a large number of predicted enhancers in PCHi-C data exhibit higher average expression levels (Figure 6a, Supplementary Material). This confirms previous observations showing that the density of enhancers in the gene vicinity is positively correlated with the expression levels (Berthelot et al., 2018). We also show that these genes generally have broad expression patterns, that is, they are expressed in large numbers of organs and developmental stages (Figure 6b, Supplementary Material). This observation suggests that the contact with a large number of enhancers may enable gene activation in a wide variety of spatio-temporal contexts. Interestingly, we observe different patterns for contacts at moderate (less than 500 kb) or at large (at least 500 kb) genomic distances (Figure 6a,b). Gene expression levels are significantly correlated with the numbers of contacts occurring at less than 500 kb distances (Kruskal-Wallis test, p-value < 1e-10), but less so with long-range contacts (Figure 6a, p-value = 0.038). In contrast,

8

expression specificity is well correlated with the number of contacts, irrespective of the distances at which they take place (Kruskal-Wallis test p-value < 1e-10 for the three distance classes). Consistent with the functional enrichment observed for genes with conserved expression profiles, we note that gene ontology categories associated with regulation of gene expression are over-represented among genes with a large number of contacted enhancers (Supplementary Table 5).

Our results thus confirm that the complexity of the regulatory landscape is linked to the pattern of gene expression (Berthelot et al., 2018). Given that gene expression levels and gene expression breadth are correlated with our estimates of the rate of expression profile evolution (Supplementary Figure 4), we corrected for the effect of these two factors with a multiple regression model (Methods). We found that this corrected expression conservation measure is weakly but significantly correlated with the number of contacted enhancers (Kruskal-Wallis test for all contacts, p-value = 0.0013, Figure 6c). Significant associations between expression conservation levels and the number of contacts are present for both moderate-range (Kruskal-Wallist test, p-value 0.0031) and long-range contacts (Kruskal-Wallis test, p-value 0.0029). These observations confirm that the complexity of *cis*-acting regulatory landscapes contributes to the robustness of gene expression during evolutionary time (Berthelot et al., 2018), and that this is not simply explained by the association between the number of regulatory elements, expression level and expression breadth.

We also evaluated quantitative differences in gene expression levels using published transcriptome sequencing data for the three cell types sampled in both human and mouse PCHi-C datasets (Methods). We confirm that there is a strong positive correlation between the number of contacted enhancers and gene expression levels in each cell type (Supplementary Figure 5a-b, Kruskal-Wallis test, p-value < 1e-10). However, we do not find a significant association between the number of contacted enhancers and quantitative gene expression differences, after correcting for the effect of the gene expression level (Kruskal-Wallis test, p-value > 0.05 for the three cell types, Supplementary Figure 5c).

The evolution of *cis*-regulatory landscapes is correlated with the evolution of gene expression profiles

We next investigated the relationship between the evolution of *cis*-regulatory landscapes and the evolution of gene expression patterns. We evaluated the evolutionary conservation of PCHi-C-predicted regulatory landscapes at three different levels: the extent of sequence conservation of the contacted enhancers, the conservation of the synteny between

9

promoters and enhancers, and the conservation of their chromatin contacts. We correlated these measures of regulatory landscape evolution with the evolution of gene expression patterns, corrected for the effect of gene expression level and expression breadth as described above (Methods).

We found significant positive associations between the rate of regulatory landscape evolution and the rate of gene expression profile evolution (Figure 6d-f, Supplementary Material). Specifically, genes that contact enhancers with highly conserved sequences tend to have well-conserved expression profiles (Kruskal-Wallis test for all contacts, p-value 8.52e-6). This association is strong for enhancers contacted at moderate distances (p-value 1.6e-4), but not significant for long-range contacts (p-value 0.80, Figure 6d). Moreover, genes that underwent synteny breaks in their regulatory landscapes tend to have less conserved expression profiles than genes with regulatory landscapes in conserved synteny (Kruskal-Wallis test for all contacts, p-value 1.83e-3). The effect of synteny breaks is significant for contacts that occur at moderate distances (p-value 0.002), but not at large distances (p-value 0.85, Figure 6e). The latter could be explained by the fact that synteny breaks that perturb medium-range interactions are likely disruptive to the entire regulatory landscape of the genes, while rearrangements that affect long-range interactions do not necessarily affect those regulatory contacts that occur in the immediate vicinity of the promoter. Finally, we tested the correlation between the proportion of conserved contacts and the rate of expression profile conservation. Although we observe a positive tendency, we could not conclude on the presence of a significant signal (Kruskal-Wallis test, p-value 0.045 for the pairwise comparison between genes with fewer than 10% conserved contacts and genes with more than 40% conserved contacts). Similar results are found when we contrast gene expression profiles between species with the Euclidean distance (Supplementary Material).

We note that we did not observe significant correlations between quantitative expression level differences and *cis*-regulatory landscape evolution, in any of the three cell types that were sampled for both mouse and human (Supplementary Figure 5d-f).

Taken together, these results illustrate the complex dependency between *cis*-regulatory landscape evolution and gene expression evolution. Although we could not find any correlation between regulatory changes and quantitative expression differences for individual cell types, we were able to uncover significant associations between the rate of *cis*-regulatory evolution and the rate of expression evolution, measured as the divergence of gene expression profiles among comparable organs and developmental stages (Cardoso-

Moreira et al., 2019).

Discussion

In this study, we investigated the evolution of gene expression *cis*-regulatory landscapes at a genome-wide scale, using experimentally determined promoter-enhancer contacts. The use of chromatin contact data enabled us to overcome the simplification made by most previous studies, which inferred target genes based on genomic proximity (Berthelot et al., 2018; Wong et al., 2017). Importantly, this data allowed us to evaluate long-range promoter-enhancer interactions, which are thought to be critical part of *cis*-regulatory networks (Montavon and Duboule 2012).

We were thus able to evaluate the evolution of *cis*-regulatory landscapes at multiple levels. Starting with primary sequence analyses, we showed that distant enhancers have higher levels of sequence conservation than those found in the immediate vicinity of their putative target genes. Enforcing the idea that long-range contacts between promoters and enhancers are an important mode of gene expression regulation, we show that synteny breaks that would prohibit these interactions are under-represented, and thus potentially counter-selected. Moreover, we observed substantial contact conservation for long-range promoter-enhancer pairs, though very little is expected by chance. These results confirm previous hypotheses stating that the presence of long-range regulatory interactions constrains the large-scale evolution of vertebrate genomes (Mongin et al. 2009; Lemaitre et al. 2009). We thus validate previous computational studies, which predicted long-range regulatory interactions based on the long-term conservation of synteny (Mongin et al., 2009; Clément et al., 2020).

We re-evaluated the relationship between gene expression evolution and regulatory evolution. Although it seems intuitive that evolutionary changes in *cis*-regulatory landscapes should affect gene expression, it is well established that, overall, protein-coding gene expression patterns evolve slowly (Necsulea and Kaessmann, 2014), while *cis*-regulatory elements such as enhancers evolve rapidly (Cheng et al., 2014; Villar et al., 2015). Moreover, previous attempts to connect the rate of enhancer evolution with the rate of expression evolution of putative target genes revealed only mild associations between the two (Berthelot et al., 2018). The robustness of gene expression regulatory networks, achieved through the presence of enhancers with redundant functions, is a plausible explanation for this observation (Cannavò et al., 2016; Frankel et al., 2010; Osterwalder et

al., 2018). Nevertheless, in our study we uncover a significant positive association between the rate of regulatory landscape evolution and the pattern of gene expression evolution. In contrast with previous studies, our work uses experimental predictions of the regulatory interactions between genes and enhancers, which were traditionally inferred based on genomic proximity alone. The more precise inference of target genes could explain our ability to better link gene expression evolution with regulatory evolution. Furthermore, here we analyze relative expression profiles across comparable organs and developmental stages, rather than expression levels in a single tissue (Berthelot et al., 2018). This approach has the advantage of reducing technical biases linked to gene expression comparisons between distant species, as well as of providing a broader overview of the pattern of expression evolution. Despite the functional redundancy of *cis*-regulatory networks, by jointly analyzing numerous biological conditions we increase the likelihood of observing the molecular consequences of enhancer evolution. Indeed, our analysis of quantitative expression differences between species, in individual cell types, did not reveal any significant association with rate of *cis*-regulatory landscape evolution.

Importantly, we show that genomic rearrangements that affect *cis*-regulatory landscapes are associated with increased divergence of expression profiles. By partially restructuring *cis*-regulatory landscapes, genomic rearrangements likely contribute to gene expression evolution, not just by disrupting existing regulatory relationships, but also by redistributing regulatory elements and thus allowing their adoption by other genes (Lettice et al., 2011). These effects on gene expression explain why such rearrangements are generally counter-selected, as indicated by our synteny conservation analyses. Our findings offer an intermediate point of view between reports that large-scale rearrangements that perturb regulatory landscapes can have strong phenotypic consequences, in mouse models of human diseases (Lupiáñez et al., 2015), and reports that multiple chromosomal rearrangements in *Drosophila* laboratory strains have little to no effects on gene expression (Ghavi-Helm et al., 2019). We note that our work, like previous studies, does not provide a complete overview of the phenotypic consequences of regulatory landscape rearrangements. On one hand, studies that were motivated by the need to understand the genomic underpinnings of human diseases are necessarily biased towards events with strongly deleterious phenotypic consequences (Lupiáñez et al., 2015). On the other hand, studies of *Drosophila* strains are likely biased towards genomic alterations with little impact on organism fitness (Ghavi-Helm et al., 2019). Here, we can only observe those genomic rearrangements that were maintained during evolution, and thus also exclude events with highly deleterious effects. This inherent limitation may explain why we do not observe stronger correlations between gene expression evolution and regulatory evolution.

12

Importantly, here we only analyse genes that are kept as orthologs between human and mouse. We speculate that cases where promoter-enhancer interactions are affected by major evolutionary events, such as large-scale genome rearrangements, could often lead to loss of function or pseudogenization, rather than gene expression profile changes.

In this manuscript, we focused on *cis*-interactions between promoters and enhancers, for which regulatory principles are well described. We thus only capture a fraction of the complexity of vertebrate regulatory landscapes. Together with new genome-wide data on other types of regulatory elements, such as expression silencers (Pang and Snyder, 2020), promoter-centered chromatin contact maps will provide a valuable starting point to investigate the roles of diverse regulatory components on gene expression evolution.

Methods

Promoter Capture HiC data processing

To obtain a global overview of the chromatin interactions that take place between gene promoters and other genomic regions, across cell types and species, we collected and processed publicly available Promoter Capture Hi-C (PCHi-C) data for human (Choy et al., 2018; Freire-Pritchett et al., 2017; Javierre et al., 2016; Mifsud et al., 2015; Pan et al., 2018; Rubin et al., 2017) and mouse samples (Comoglio et al., 2018; Koohy et al., 2018; Novo et al., 2018; Schoenfelder et al., 2015, 2018; Siersbæk et al., 2017). To ensure data comparability, we selected PCHi-C datasets that were generated with experimental procedures similar to those described by Schoenfelder and co-authors (Schoenfelder et al., 2015). We processed PCHi-C data for 26 samples and 16 cell types for human and 14 samples and 8 cell types for mouse (Supplementary Table 1). The data include three cell types (embryonic stem cells, adipocytes and B cells) that are in principle comparable across species, although cell culture procedures and differentiation stages may differ (Supplementary Table 1).

To homogenize computational analysis procedures across datasets and species, we re-processed all PCHi-C raw data with the same computational approaches. We first used the HiCup pipeline (Wingett et al., 2015), which aligns reads, filters artifactual fragments (such as circularized reads and re-ligations), and removes duplicates. We mapped PCHi-C reads to the human hg38 (GRCh38.p12) and mouse mm10 (GRCm38.p6) genome assemblies, downloaded from Ensembl release 94 (Cunningham et al., 2019), using Bowtie2 version 2.3.4.3 (Langmead and Salzberg, 2012). We then called interactions with the CHiCAGO pipeline (Cairns et al., 2016). We selected chromatin contacts with a CHiCAGO score

13

greater or equal to 5 in at least one sample (Supplementary Table 1). We obtained 910,180 unique chromatin interactions between 19,389 baited restriction fragments and 308,359 other fragments, for human; and 824,406 unique chromatin interactions between 21,858 baited fragments and 247,668 other fragments, for mouse. These data are provided in Supplementary Dataset 1. We further focused on intra-chromosomal contacts *(cis*-interactions*)* separated by a genomic distance of 25 kb to 2 Mb between the midpoint of baited and contacted regions (representing 83% and 84% of total interactions in human and mouse datasets respectively). With this filter, we exclude short-range interactions that have high levels of background noise (Cairns et al., 2016), as well as interactions that are beyond the typical size observed for topologically-associating domains (Dixon et al., 2012). We discarded interactions that involved restriction fragments that were smaller than 150 bp or larger than 50 kb. We also discarded interactions involving restriction fragments that showed aberrant interaction patterns probably due to structural variation (*i.e.*, in contact with more than 15 baited fragments, with a mean CHiCAGO score over 10 and at a median distance of more than 1Mb). The full list of excluded restriction fragments is provided in Supplementary Dataset 1. Finally, we restricted our analyses to contacts between baited and non-baited restriction fragments, which are more likely to include “classical” regulatory interactions, such as those between promoters and enhancer elements. All these filters resulted in 662,562 human and 564,678 mouse unique chromatin interactions.

Simulated interactions

As a control for our evolutionary analyses, we generated simulated interaction landscapes that reproduce the observed distribution of distances between baited restriction fragments and non-baited contacted fragments, as well as the number of contacted fragments *per* bait, for each sample (Supplementary Material). To do this, we computed the absolute linear genomic distance between the center position of each baited fragment and each interacting fragment. We then divided the observed interactions into 5 kb distance classes, from 25 kb to 2 Mb upstream and downstream of the baited region. There were 395 classes in total. We computed the fraction of contacts that are found in each distance class, across all baited fragments, for each sample. This distance distribution was used to simulate contacts, as follows: for each baited fragment, we computed the contact probability for all fragments found on the same chromosome, within the 25 kb - 2 Mb distance window, as the average probability of the overlapping distance classes (a single fragment can overlap with multiple distance classes). We then randomly drew contacts based on this empirical probability distribution. We respected the number of interactions observed in the real PCHiC interactions, for each bait. The simulated landscape thus mimics the average distribution of distances between baits and non-baited contacted fragments, as well as the number of

14

contacts *per* bait (though not *per* contacted fragment), for each sample. The simulated interaction data are available in Supplementary Dataset 2.

Baited region annotation

We re-annotated PCHiC baited restriction fragments with respect to transcription start sites, using annotations from the Ensembl database, release 94. We downloaded transcript coordinates from Ensembl using the BioMart interface (Kinsella et al., 2011) (Supplementary Dataset 3). For each baited fragment, we extracted all transcription start sites of protein coding genes that were found within at most 1 kb of the fragment. Each baited restriction fragment could thus be associated with zero, one or several protein coding genes. Conversely, each gene could be represented by zero, one or several baited fragments. The baited fragment annotation is available in Supplementary Dataset 1.

Genomic characteristics of the PC-HiC contacted sequences

We evaluated the characteristics of the genomic sequences contacted by baited promoters. To do this, we extracted the repeat-masked DNA sequence of all baits and contacted regions, and used BLAT (Kent et al., 2012) to search for sequence similarity in the same genome. For each sequence, we counted the number of BLAT hits corresponding to at least 80% of their repeat-masked length with at least 95% sequence identity, elsewhere in the genome. We considered that sequences with more than 1 BLAT hit were potentially duplicated in the reference genome and we discarded them from further analyses. We evaluated the GC content and the repetitive sequence content of all contacted regions, using RepeatMasker annotation tables provided by the UCSC Genome Browser (Hinrichs, 2006). We also evaluated the protein coding gene density within a 500 kb maximum distance, upstream and downstream of the contacted sequences. These genomic characteristics are available in Supplementary Dataset 5.

Predicted enhancer elements

We evaluated the presence of predicted enhancer elements in the PCHiC contacted regions, using different sources and types of data. First, we extracted enhancers predicted with the Cap Analysis of Gene Expression method by the FANTOM5 consortium (Andersson et al., 2014). This dataset consisted of 63,285 human enhancers and 49,797 mouse enhancers. Enhancer coordinates on hg38 (Dalby et al. 2017) and mm10 (Dalby et al. 2018) genome assemblies were downloaded from FANTOM5 consortium: https://zenodo.org/record/556775#.X3Gvf5rgprl and https://zenodo.org/record/1411211#.X3Gvq5rgprm .

15

Second, we used data from a recent publication, which inferred enhancer-promoter links based on their correlated activity across numerous samples (Hait et al., 2018). From this resource, we downloaded enhancer positions predicted with DNase I hypersensitivity (DHS) assays by the ENCODE consortium (Thurman et al., 2012) or by the Roadmap Epigenomics consortium (Roadmap Epigenomics Consortium et al., 2015), for the human genome: http://acgt.cs.tau.ac.il/focs/download.html. There were 408,802 putative enhancers in the ENCODE dataset and 470,546 putative enhancers in the Roadmap Epigenomics dataset. We also downloaded 255,925 putative enhancer positions predicted with global run-on sequencing (GRO-seq) for the human genome, from the same source (Hait et al., 2018). For all three datasets, we converted enhancer coordinates from the hg19 to the hg38 genome assembly using liftOver and associated genome alignment files downloaded from the UCSC Genome Browser (Kent et al., 2012). After genome coordinate conversion, we had 408,737 enhancers in the ENCODE DHS dataset, 470,479 in the RoadmapEpigenomics DHS dataset and 255,812 in the GRO-seq dataset.

Third, we extracted enhancer positions predicted based on the presence of H3K4me1, H3K4me3 and H3K27ac histone modifications in the mouse genome, from the ENCODE consortium (The Mouse ENCODE Consortium et al., 2014). We downloaded the midpoint coordinates of predicted enhancer elements for 23 tissues or cell lines from http://promoter.bx.psu.edu/ENCODE/download.html and we extended them by 75 bp on each side. We then combined the enhancer coordinates across all samples, obtaining 740,058 enhancer regions. We converted enhancer coordinates from the mm9 to the mm10 mouse genome assembly using liftOver and associated genome alignment files from the UCSC Genome Browser (Kent et al., 2012). After conversion, there were 739,996 enhancer regions in the mouse mm10 assembly.

Finally, we applied the same procedure described above for restriction fragments to discard duplicated enhancers and to evaluate their genomic characteristics (GC content, repetitive sequences, proximal gene density). The predicted enhancers data and their genomic characteristics are available in Supplementary Dataset 4.

Prediction of contacts between gene and enhancers

For most evolutionary analyses, we analyzed pairs of gene promoters and enhancers predicted to be in contact by PCHi-C data. We constructed gene-enhancer pairs by associating to each “baited” protein-coding gene all the predicted enhancers contained in the restriction fragments contacted by its baits. We applied this procedure to each predicted

16

enhancer dataset described above, obtaining four different contact datasets for human and two datasets for mouse (Supplementary Dataset 4).

Correlated activity of gene-enhancer pairs

To verify whether the promoter-enhancer interactions predicted with PCHi-C data are enriched in genuine regulatory relationships, we tested whether their activity was significantly correlated across samples. To do this, we used estimates of promoter and enhancer activity based on DHS, CAGE and GRO-seq assays, provided by the ENCODE, FANTOM5, RoadmapEpigenomics consortia and other publications and processed by Hait and co-authors in the framework of the FOCS method (Hait et al., 2018). We downloaded normalized activity profiles for promoters (or genes) and enhancers from http://acgt.cs.tau.ac.il/focs/download.html. We computed pairwise Spearman correlations, based on normalized activity profiles across samples, for all pairs of promoters - enhancers predicted to be in contact in the real PCHi-C data or in the simulated data (Supplementary Dataset 8). We then tested whether the correlation coefficient distributions were significantly different, using the Wilcoxon rank-sum test for median comparisons and the Kolmogorov-Smirnov test for global distribution comparisons, as implemented in R (R Core Team, 2014).

Sequence conservation

To evaluate the evolutionary conservation of sequences that are found in contact with gene promoters in the PCHiC data, we first identified putative homologous regions using liftOver on pairwise whole-genome alignments between a reference species (human or mouse) and a target species (human, macaque, mouse, rat, rabbit, cow, elephant, dog, opossum or chicken), downloaded from the UCSC Genome Browser (Hinrichs, 2006). We set a low threshold (10%) for the minimum ratio of bases that must remap in the liftOver genomic coordinates conversion. We discarded regions that were duplicated or split in the target genome. The regions who did not remap in the target genome were considered as non-conserved and were given a conservation score of 0. The predicted homologous regions were then aligned with Pecan (Paten et al., 2008). We computed the sequence conservation score as the ratio between the total number of aligned (without gaps) base pairs and the total number of bases in the alignment (union between reference and target species). To avoid the confounding factor of the presence of gene regions, we computed the sequence conservation score on non-exonic regions alone. To do this, we extracted exonic coordinates from the Ensembl database. Both coding and non-coding exons were masked prior to evaluating sequence conservation. To evaluate the sequence conservation of potential regulatory elements, we applied the same alignment procedure for predicted enhancers

17

described above. The alignements statistics are available in Supplementary Dataset 7.

Synteny conservation

We defined synteny conservation between a reference and a target species as the presence of both genes and predicted enhancer on the same chromosome and at a distance of less than 2 Mb in both species. As previously, the estimated size of interaction domains is rarely above 2Mb, it is estimated that beyond this distance functional interactions are difficult. We restricted this analysis to protein-coding genes that have a 1-to-1 orthologue gene in the target species, as predicted in the Ensembl database (release 94 for all but one species, release 99 for macaque). We also selected only those interacting enhancers that had predicted homologous regions in the target genome, as defined above. For each pair of species, we discarded the 10% least conserved enhancers based on the non-exonic alignment score. We further restricted this analysis to promoters-enhancers pairs that are separated by distances between 100 kb and 1.5 Mb in the reference species. With this convention, synteny "breaks" should be interpreted as evolutionary events that add at least 500 kb to the distance between promoters and enhancers. We then asked whether the predicted homologue of the contacted enhancer was on the same chromosome and within 2 Mb of the transcription start site of the predicted 1-to-1 orthologue of the gene. If the 1-to-1 orthologous gene in the target species had more than one transcription start site, we considered the minimum distance between these and the contacted region homologue. The resulting homologous gene-enhancer pairs are available in Supplementary Dataset 7.

Chromatin contact conservation

Similar to the synteny conservation analysis, we restricted this analysis to pairs of gene-predicted enhancer for which the protein coding gene was associated to a 1-to-1 orthologue gene in the target species, and for which the contacted enhancer had a predicted homologue in the target species, defined using liftOver and Pecan as described above. We further restricted the analysis to cases where the 1-to-1 orthologous gene was also baited in the PCHiC data of the target species (human or mouse). We then asked whether the predicted homologue of the contacted enhancer overlapped any of the regions found in contact with the bait(s) associated with the 1-to-1 orthologous gene, in the PCHiC data of the target species. The proportion of conserved contacts was computed with respect to the number of pairs satisfying all previous criteria (*i.e.,* homology prediction for the baited gene and the contacted enhancer, presence of baits for the orthologous gene in the target species). We evaluated the chromatin contact conservation between each pair of PCHiC or simulated human and mouse samples. The resulting analyses are available in

18

Supplementary Dataset 7.

Gene expression data

To evaluate the relationship between gene expression patterns and regulatory landscape characteristics, we analyzed gene expression data for human and mouse. First, we used expression data across multiple organs and developmental stages, from a recent publication (Cardoso-Moreira et al., 2019). We downloaded gene-level RPKM (reads *per* kilobase of exon *per* million mapped reads) values from this publication. For evolutionary analyses, we restricted the expression table to protein coding genes predicted as 1-to-1 orthologs for human and mouse in the Ensembl database, release 94, and to organ/developmental stage combinations that were deemed to be directly comparable between the two species (Cardoso-Moreira et al., 2019). We also restricted our analysis to somatic organs, excluding the fast-evolving testes samples (Brawand et al., 2011, Cardoso-Moreira et al., 2019) and the ovary, for which fewer samples were available. We re-normalized the data across samples and species using a previously published median-scaling procedure, based on the genes that vary the least in terms of ranks across samples (Brawand et al., 2011). We required genes to be expressed above a minimum threshold (RPKM=2) in at least three samples. The resulting expression data and sample details are provided in Supplementary Dataset 6.

Second, we aimed to compare gene expression levels between human and mouse, in cell types that are comparable between the two species and for which PCHiC data is also available. We thus focused on three cell types: embryonic stem cells (ESC), adipocytes and B cells. We note that the corresponding samples in our PCHi-C collection are not expected to be perfectly comparable between the two species, as they may still differ in terms of differentiation stages, cell culture protocols, *etc.* (Supplementary Table 1). We searched the Sequence Read Archive (SRA) database for RNA-seq samples derived from these cell types, obtaining between 4 and 20 samples for each species and cell type. We computed gene expression levels using Kallisto (Bray et al., 2016), on cDNA sequences downloaded from Ensembl 94. For each gene, we computed mean and median TPM (transcript *per* million) expression levels, across all replicates for each cell type. These values were used to compute cell type-specific gene expression divergence values between human and mouse. The resulting expression data and sample details are provided in Supplementary Dataset 6.

Gene expression characteristics

We used transcriptome data from multiple organs and developmental stages (Cardoso-

Moreira et al., 2019) to broadly characterize gene expression patterns. We computed a tissue/developmental stage specificity index with the formula tau = sum $(1 - r_i)/(n-1)$, where r_i represents the ratio between the expression level in sample i and the maximum expression level across all samples, and n represents the total number of samples (Liao et al., 2006). We computed this index on RPKM values, averaged across all replicates for a given species / organ / developmental stage combination.

Evolutionary divergence of gene expression profiles

To evaluate the divergence of gene expression patterns between two species, we first computed relative expression profiles, by dividing the RPKM values (averaged across biological replicates) by the maximum value observed among samples (organ/developmental stage combinations), for each gene. To do this, we used transcriptome data from multiple organs and developmental stages (Cardoso-Moreira et al., 2019). We then defined a measure of expression divergence by calculating the Spearman correlation coefficient and the Euclidean distance between the relative expression profiles of orthologous genes for a pair of species. As these measures of expression divergence are significantly correlated with the average gene expression level and the expression specificity index of reference species, we built linear regressions that model the relationship between expression divergence, expression specificity and expression levels (averaged across all samples and species, for each gene), and extracted the residual values. We referred to these values as “corrected expression divergence” in the supplementary figures. The gene expression divergence data is provided in Supplementary Dataset 6.

Quantitative differences in gene expression levels between species

We measured quantitative gene expression differences between species for the three cell types that were present in the PCHi-C datasets of both human and mouse: B lymphocytes, pre-adipocytes and embryonic stem cells. To do this, we computed average TPM levels across all replicates available for a given cell type, for each species. We then estimated expression divergence as the absolute value of the difference between human and mouse, divided by the maximum of the two values, for each gene. We note that this expression divergence measure is strongly correlated with gene expression levels. To correct for this effect, we constructed a linear regression that models the relationship between the expression divergence and the average expression levels, across both species, and extracted the residual values. We referred to these values as “corrected expression divergence”. Gene expression conservation measures are simply defined as 1-expression divergence values. The gene expression divergence data is provided in Supplementary

20

Dataset 6.

Evolutionary divergence of regulatory landscapes

We measured different levels of evolutionary divergence of gene regulatory landscapes defined by PCHi-C interactions between promoters and predicted enhancers. We performed analysis on 1-to-1 orthologous protein-coding genes between human and mouse. For each reference gene, we evaluated the following aspects: the total number of enhancers found in contact with the reference gene promoter(s); the non-exonic sequence alignment score of contacted enhancers; the number of enhancers with conserved sequences in the target species (*i.e.*, successfully projected with liftOver in the target genome and with a non-exonic sequence alignment score greater or equal to 40% corresponding to the average alignment between human and mouse genomes) and found in conserved synteny with the orthologous gene in the target genome (*i.e.*, on the same chromosome and within a distance of 2 Mb); the number of enhancers with conserved sequences, whose predicted homologue was also in PCHiC-supported contact with the orthologous gene in the target genome. We could thus evaluate for each pair of orthologous genes between the two species, the median alignment score of contacted enhancers, the percentage of conserved enhancers maintained in synteny and the percentage of conserved contacts. These two last measures were calculated only for genes who have at least 5 and at most 100 conserved enhancers or enhancers maintained in synteny. We then correlated these measures to the divergence of gene expression calculated above.

Gene ontology enrichment

We performed gene ontology enrichment analyses on single-ranked lists of protein coding genes present in human PCHi-C data and with an 1-to-1 orthologous gene in mouse genome with GOrilla (Eden et al., 2009). Gene lists were ranked by the following factors: the number of contacted ENCODE enhancers, the median distance between these enhancers and the gene TSS, the mean alignment score of contacted enhancers, the proportion of these enhancers that are conserved in synteny and in contact, their expression profile conservation. The ontology enrichment results are provided in Supplementary Dataset 9 and Supplementary Tables 2-5.

Statistics and graphical representations

We used R 3.5.2 for all statistical analyses and graphical representations (R Core Team, 2014). Given that the variables we analyze are often not normally-distributed, we used non-parametric statistical tests as a general rule. We computed 95% confidence intervals for

21

mean values using the BCa non-parametric bias-corrected and accelerated bootstrap method (DiCiccio and Efron, 1996), implemented in the bootBCa 1.0/r2 package in R (Tibshirani et al., 2014). We performed 100 bootstrap replicates. We computed 95% confidence intervals for median values using the formula +/- 1.58 IQR/sqrt(n), where IQR is the inter-quartile range of the distribution and n the total number of values. This formula is implemented in the boxplot.stats function in the grDevices package in R (Chambers et al., 1983). We performed pairwise comparisons of distributions with the Wilcoxon rank sum test and multiple comparisons with the Kruskal-Wallis test, both implemented in the stats package in R. We compared proportions with the chi-squared test and computed two-sided 95% confidence intervals with the prop.test function in R.

Code availability

All scripts are available at : https://github.com/AlexandreLaverre/Regulatory_Landscape

Acknowledgements

We thank C. Berthelot, Y. Ghavi-Helm, F. Picard and D. Mouchiroud for discussions and advice on the project. We also thank T. Latrille and T. Tricou for their valuable advice and for assistance with bioinformatics and P. Veber for advice on statistics. Computational analyses were performed using the computing facilities of the CC LBBE/PRABI and the Core Cluster of the Institut Français de Bioinformatique (IFB) (ANR-11-INBS-0013). This work was funded by the French National Research Agency (ANR-17-CE12-0019-01 «LncEvoSys»).

Author contributions

The project was originally designed by A.L., E.T and A.N.; A.L. conducted data analyses with help from A.N; A.L., E.T. and A.N. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Corresponding authors

Alexandre Laverré: alexandre.laverre@univ-lyon1.fr

Anamaria Necsulea: anamaria.necsulea@univ-lyon1.fr

Additional information

Processed data are available at

22

http://pbil.univ-lyon1.fr/members/necsulea/RegulatoryLandscapes

References

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., et al. (2014). “An atlas of active enhancers across human cell types and tissues”. Nature *507*, 455–461.

Berthelot, C., Villar, D., Horvath, J.E., Odom, D.T., and Flicek, P. (2018). “Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression”. Nature Ecology & Evolution *2*, 152–163.

Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csárdi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M., et al. (2011). “The evolution of gene expression levels in mammalian organs”. Nature *478*, 343–348.

Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nature Biotechnology *34*, 525–527.

Cairns, J., Freire-Pritchett, P., Wingett, S.W., Várnai, C., Dimond, A., Plagnol, V., Zerbino, D., Schoenfelder, S., Javierre, B.-M., Osborne, C., et al. (2016). “CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data”. Genome Biology *17*.

Cannavò, E., Khoueiry, P., Garfield, D.A., Geeleher, P., Zichner, T., Gustafson, E.H., Ciglar, L., Korbel, J.O., and Furlong, E.E.M. (2016). “Shadow Enhancers Are Pervasive Features of Developmental Regulatory Networks”. Current Biology *26*, 38–51.

Cardoso-Moreira, M., Halbert, J., Valloton, D., Velten, B., Chen, C., Shao, Y., Liechti, A., Ascenção, K., Rummel, C., Ovchinnikova, S., et al. (2019). “Gene expression across mammalian organ development”. Nature *571*, 505–509.

Chambers, J., M., Cleveland, W.S., Kleiner, B., and Tukey, P.A. (1983). “Graphical Methods for Data Analysis” (Wadsworth & Brooks/Cole).

Cheng, Y., Ma, Z., Kim, B.-H., Wu, W., Cayting, P., Boyle, A.P., Sundaram, V., Xing, X., Dogan, N., Li, J., et al. (2014). “Principles of regulatory information conservation between mouse and human”. Nature *515*, 371–375.

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., et al. (2018). “Promoter interactome of

23

human embryonic stem cell-derived cardiomyocytes connects GWAS regions to cardiac gene networks". Nature Communications *9*.

Clément, Y., Torbey, P., Gilardi-Hebenstreit, P., and Crollius, H.R. (2020). "Enhancer–gene maps in the human and zebrafish genomes using evolutionary linkage conservation". Nucleic Acids Research *48*, 2357–2371.

Comoglio, F., Park, H.J., Schoenfelder, S., Barozzi, I., Bode, D., Fraser, P., and Green, A.R. (2018). "Thrombopoietin signaling to chromatin elicits rapid and pervasive epigenome remodeling within poised chromatin architectures". Genome Research *28*, 295–309.

Cunningham, F., Achuthan, P., Akanni, W., Allen, J., Amode, M.R., Armean, I.M., Bennett, R., Bhai, J., Billis, K., Boddu, S., et al. (2019). "Ensembl 2019". Nucleic Acids Research *47*, D745–D751.

Dalby, M., Rennie, S., and Andersson, R. (2017). "FANTOM5 transcribed enhancers in hg38" (Zenodo).

Dalby, M., Rennie, S., and Andersson, R. (2018). "FANTOM5 transcribed enhancers in mm10" (Zenodo).

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). "Capturing Chromosome Conformation". Science *295*, 1306–1311.

DiCiccio, T.J., and Efron, B. (1996). "Bootstrap Confidence Intervals". Statistical Science *11*, 189–228.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). "Topological domains in mammalian genomes identified by analysis of chromatin interactions". Nature *485*, 376–380.

Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009). "GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists". BMC Bioinformatics *10*.

Frankel, N., Davis, G.K., Vargas, D., Wang, S., Payre, F., and Stern, D.L. (2010). "Phenotypic robustness conferred by apparently redundant transcriptional enhancers". Nature *466*, 490–493.

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., et al. (2017). "Global reorganisation of cis-regulatory units upon lineage commitment of human embryonic stem

24

cells". ELife *6*.

Ghavi-Helm, Y., Jankowski, A., Meiers, S., Viales, R.R., Korbel, J.O., and Furlong, E.E.M. (2019). "Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression". Nat Genet *51*, 1272–1282.

Gilad, Y., Oshlack, A., Smyth, G.K., Speed, T.P., and White, K.P. (2006). "Expression profiling in primates reveals a rapid evolution of human transcription factors". Nature *440*, 242–245.

Hait, T.A., Amar, D., Shamir, R., and Elkon, R. (2018). "FOCS: a novel method for analyzing enhancer and gene activity patterns infers an extensive enhancer–promoter map". Genome Biology *19*.

Hinrichs, A.S. (2006). "The UCSC Genome Browser Database: update 2006". Nucleic Acids Research *34*, D590–D598.

International Human Genome Sequencing Consortium (2001). "Initial sequencing and analysis of the human genome". Nature *409*, 860–921.

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., et al. (2016). "Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters". Cell *167*, 1369-1384.e19.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2012). "The Human Genome Browser at UCSC". Genome Research *12*, 996–1006.

Khaitovich, P. (2005). "Parallel Patterns of Evolution in the Genomes and Transcriptomes of Humans and Chimpanzees". Science *309*, 1850–1854.

King, M.-C., and Wilson, A.C. (1975). "Evolution at two levels in humans and chimpanzees". Science *188*, 107–116.

Kinsella, R.J., Kahari, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., Almeida-King, J., Staines, D., Derwent, P., Kerhornou, A., et al. (2011). "Ensembl BioMarts: a hub for data retrieval across taxonomic space". Database *2011*, bar030–bar030.

Koohy, H., Bolland, D.J., Matheson, L.S., Schoenfelder, S., Stellato, C., Dimond, A., Várnai, C., Chovanec, P., Chessa, T., Denizot, J., et al. (2018). "Genome organization and chromatin analysis identify transcriptional downregulation of insulin-like growth factor

signaling as a hallmark of aging in developing B cells". Genome Biology *19*.

Kvon, E.Z., Waymack, R., Elabd, M.G., and Wunderlich, Z. (2021). "Enhancer redundancy in development and disease". Nature Reviews Genetics.

de Laat, W., and Duboule, D. (2013). "Topology of mammalian developmental enhancers and their regulatory landscapes". Nature *502*, 499–506.

Langmead, B., and Salzberg, S.L. (2012). "Fast gapped-read alignment with Bowtie 2". Nature Methods *9*, 357–359.

Lemaitre, C., Zaghloul, L., Sagot, M.-F., Gautier, C., Arneodo, A., Tannier, E., and Audit, B. (2009). "Analysis of fine-scale mammalian evolutionary breakpoints provides new insight into their relation to genome organisation". BMC Genomics *10*, 335.

Lettice, L.A., Heaney, S.J.H., Purdie, L.A., Li, L., de Beer, P., Oostra, B.A., Goode, D., Elgar, G., Hill, R.E., and de Graaff, E. (2003). "A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly". Human Molecular Genetics *12*, 1725–1735.

Lettice, L.A., Daniels, S., Sweeney, E., Venkataraman, S., Devenney, P.S., Gautier, P., Morrison, H., Fantes, J., Hill, R.E., and FitzPatrick, D.R. (2011). "Enhancer-adoption as a mechanism of human developmental disease". Hum. Mutat. *32*, 1492–1499.

Liao, B.-Y., Scott, N.M., and Zhang, J. (2006). "Impacts of Gene Essentiality, Expression Pattern, and Gene Compactness on the Evolutionary Rate of Mammalian Proteins". Molecular Biology and Evolution *23*, 2072–2080.

Lupiáñez, D.G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J.M., Laxova, R., et al. (2015). "Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions". Cell *161*, 1012–1025.

McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010). "GREAT improves functional interpretation of cis-regulatory regions". Nature Biotechnology *28*, 495–501.

Mifsud, B., Tavares-Cadete, F., Young, A.N., Sugar, R., Schoenfelder, S., Ferreira, L., Wingett, S.W., Andrews, S., Grey, W., Ewels, P.A., et al. (2015). "Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C". Nature Genetics *47*, 598–606.

Mongin, E., Dewar, K., and Blanchette, M. (2009). "Long-range regulation is a major driving

26

force in maintaining genome integrity". BMC Evolutionary Biology *9*, 203.

Montavon, T., and Duboule, D. (2012). "Landscapes and archipelagos: spatial organization of gene regulation in vertebrates". Trends in Cell Biology *22*, 347–354.

Montavon, T., Soshnikova, N., Mascrez, B., Joye, E., Thevenet, L., Splinter, E., de Laat, W., Spitz, F., and Duboule, D. (2011). "A Regulatory Archipelago Controls Hox Genes Transcription in Digits". Cell *147*, 1132–1145.

Necsulea, A., and Kaessmann, H. (2014). "Evolutionary dynamics of coding and non-coding transcriptomes". Nature Reviews Genetics *15*, 734–748.

Novo, C.L., Javierre, B.-M., Cairns, J., Segonds-Pichon, A., Wingett, S.W., Freire-Pritchett, P., Furlan-Magaril, M., Schoenfelder, S., Fraser, P., and Rugg-Gunn, P.J. (2018). "Long-Range Enhancer Interactions Are Prevalent in Mouse Embryonic Stem Cells and Are Reorganized upon Pluripotent State Transition". Cell Reports *22*, 2615–2627.

Osterwalder, M., Barozzi, I., Tissières, V., Fukuda-Yuzawa, Y., Mannion, B.J., Afzal, S.Y., Lee, E.A., Zhu, Y., Plajzer-Frick, I., Pickle, C.S., et al. (2018). "Enhancer redundancy provides phenotypic robustness in mammalian development". Nature.

Ovcharenko, I., Loots, G.G., Nobrega, M.A., Hardison, R.C., Miller, W., and Stubbs, L. (2005). "Evolution and functional classification of vertebrate gene deserts". Genome Research *15*, 137–145.

Pan, D.Z., Garske, K.M., Alvarez, M., Bhagat, Y.V., Boocock, J., Nikkola, E., Miao, Z., Raulerson, C.K., Cantor, R.M., Civelek, M., et al. (2018). "Integration of human adipocyte chromosomal interactions with adipose gene expression prioritizes obesity-related genes from GWAS". Nature Communications *9*.

Pang, B., and Snyder, M.P. (2020). "Systematic identification of silencers in human cells". Nat Genet *52*, 254–263.

Paten, B., Herrero, J., Beal, K., Fitzgerald, S., and Birney, E. (2008). "Enredo and Pecan: Genome-wide mammalian consistency-based multiple alignment with paralogs". Genome Research *18*, 1814–1828.

R Core Team (2014). "R: A Language and Environment for Statistical Computing" (Vienna, Austria: R Foundation for Statistical Computing).

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al. (2015). "Integrative

analysis of 111 reference human epigenomes". Nature *518*, 317–330.

Rubin, A.J., Barajas, B.C., Furlan-Magaril, M., Lopez-Pajares, V., Mumbach, M.R., Howard, I., Kim, D.S., Boxer, L.D., Cairns, J., Spivakov, M., et al. (2017). "Lineage-specific dynamic and pre-established enhancer–promoter contacts cooperate in terminal differentiation". Nature Genetics *49*, 1522–1528.

Schoenfelder, S., and Fraser, P. (2019). "Long-range enhancer–promoter contacts in gene expression control". Nature Reviews Genetics.

Schoenfelder, S., Furlan-Magaril, M., Mifsud, B., Tavares-Cadete, F., Sugar, R., Javierre, B.-M., Nagano, T., Katsman, Y., Sakthidevi, M., Wingett, S.W., et al. (2015). "The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements". Genome Research *25*, 582–597.

Schoenfelder, S., Mifsud, B., Senner, C.E., Todd, C.D., Chrysanthou, S., Darbo, E., Hemberger, M., and Branco, M.R. (2018). "Divergent wiring of repressive and active chromatin interactions between mouse embryonic and trophoblast lineages". Nature Communications *9*.

Siersbæk, R., Madsen, J.G.S., Javierre, B.M., Nielsen, R., Bagge, E.K., Cairns, J., Wingett, S.W., Traynor, S., Spivakov, M., Fraser, P., et al. (2017). "Dynamic Rewiring of Promoter-Anchored Chromatin Loops during Adipocyte Differentiation". Molecular Cell *66*, 420-435.e5.

Smemo, S., Tena, J.J., Kim, K.-H., Gamazon, E.R., Sakabe, N.J., Gómez-Marín, C., Aneas, I., Credidio, F.L., Sobreira, D.R., Wasserman, N.F., et al. (2014). "Obesity-associated variants within FTO form long-range functional connections with IRX3". Nature *507*, 371–375.

The Mouse ENCODE Consortium, Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., Ryba, T., Sandstrom, R., Ma, Z., Davis, C., et al. (2014). "A comparative encyclopedia of DNA elements in the mouse genome". Nature *515*, 355–364.

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., et al. (2012). "The accessible chromatin landscape of the human genome". Nature *489*, 75–82.

Tibshirani, R., Leisch, F., and Flater, D. (2014). "bootBCa: Function to find nonparametric BCa intervals". R package version 1.0/r2.

Villar, D., Berthelot, C., Aldridge, S., Rayner, T.F., Lukk, M., Pignatelli, M., Park, T.J.,

Deaville, R., Erichsen, J.T., Jasinska, A.J., et al. (2015). “Enhancer Evolution across 20 Mammalian Species”. Cell *160*, 554–566.

Wingett, S., Ewels, P., Furlan-Magaril, M., Nagano, T., Schoenfelder, S., Fraser, P., and Andrews, S. (2015). “HiCUP: pipeline for mapping and processing Hi-C data”. F1000Research *4*, 1310.

Wong, E.S., Schmitt, B.M., Kazachenka, A., Thybert, D., Redmond, A., Connor, F., Rayner, T.F., Feig, C., Ferguson-Smith, A.C., Marioni, J.C., et al. (2017). “Interplay of cis and trans mechanisms driving transcription factor binding and gene expression evolution”. Nature Communications *8*.

Zhao, Z., Tavoosidana, G., Sjölinder, M., Göndör, A., Mariano, P., Wang, S., Kanduri, C., Lezcano, M., Singh Sandhu, K., Singh, U., et al. (2006). “Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions”. Nature Genetics *38*, 1341–1347.

29

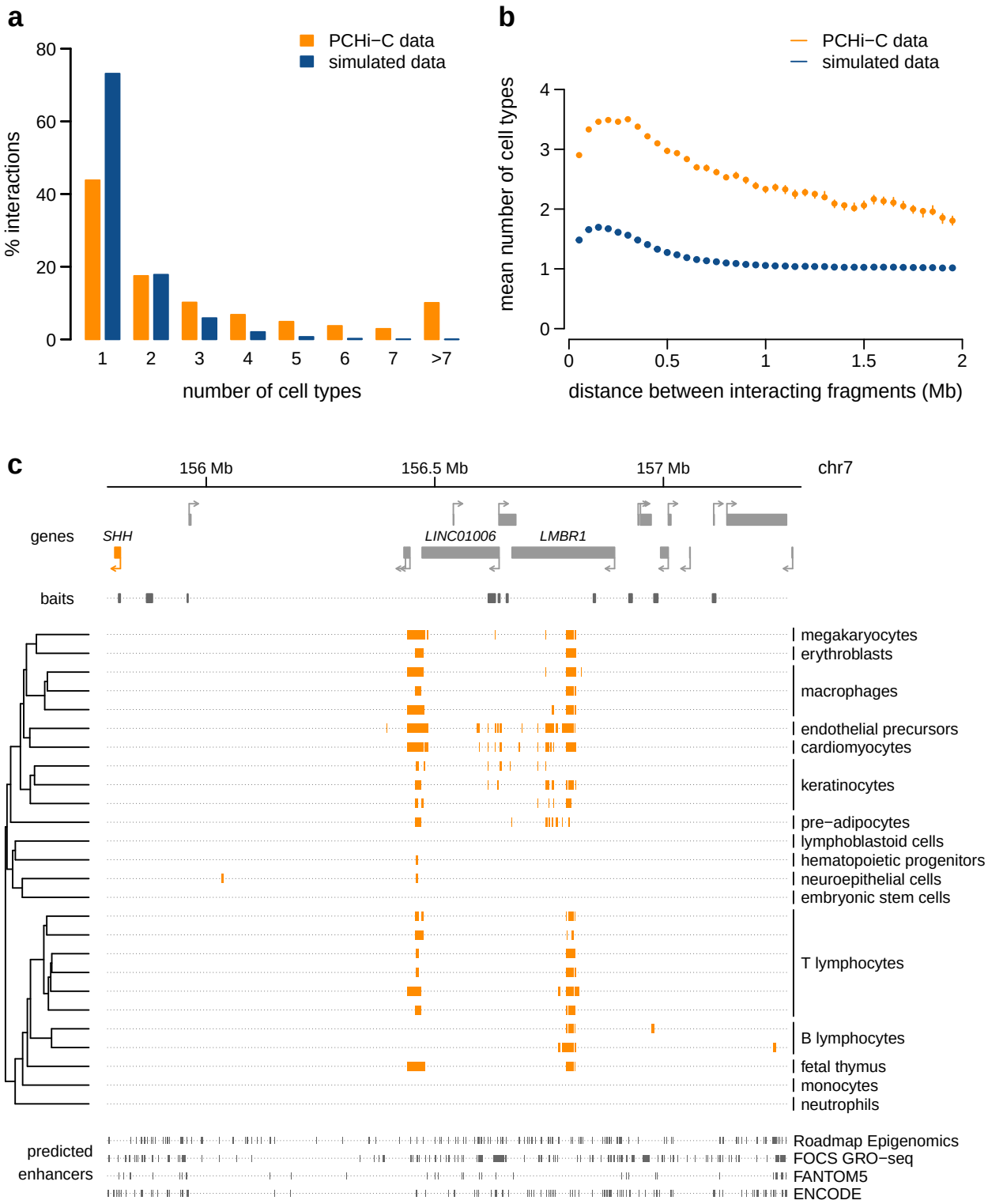

Figure 1: PCHi-C contacts are shared among cell types and include pre-formed chromatin contacts. **a)** Histogram of the number of cell types in which interactions are observed, for human PCHi-C data (orange) and simulated data (blue). **b)** Average number of cell types in which interactions are observed, as a function of the distance between baited and contacted fragments. Dots represent mean values, vertical segments represent 95% confidence intervals of the mean, obtained with a non-parametric bootstrap approach (Methods). **c)** Chromatin contacts between the *SHH* gene promoter and other genomic regions. From top to bottom: gene coordinates; localization of PCHi-C baits; localization of restriction fragments contacted by the *SHH* bait in different samples; localization of predicted enhancers. Left: hierarchical clustering of samples, based on the percentage of shared interactions between pairs of samples, in PCHi-C data *vs.* simulated data (Methods).

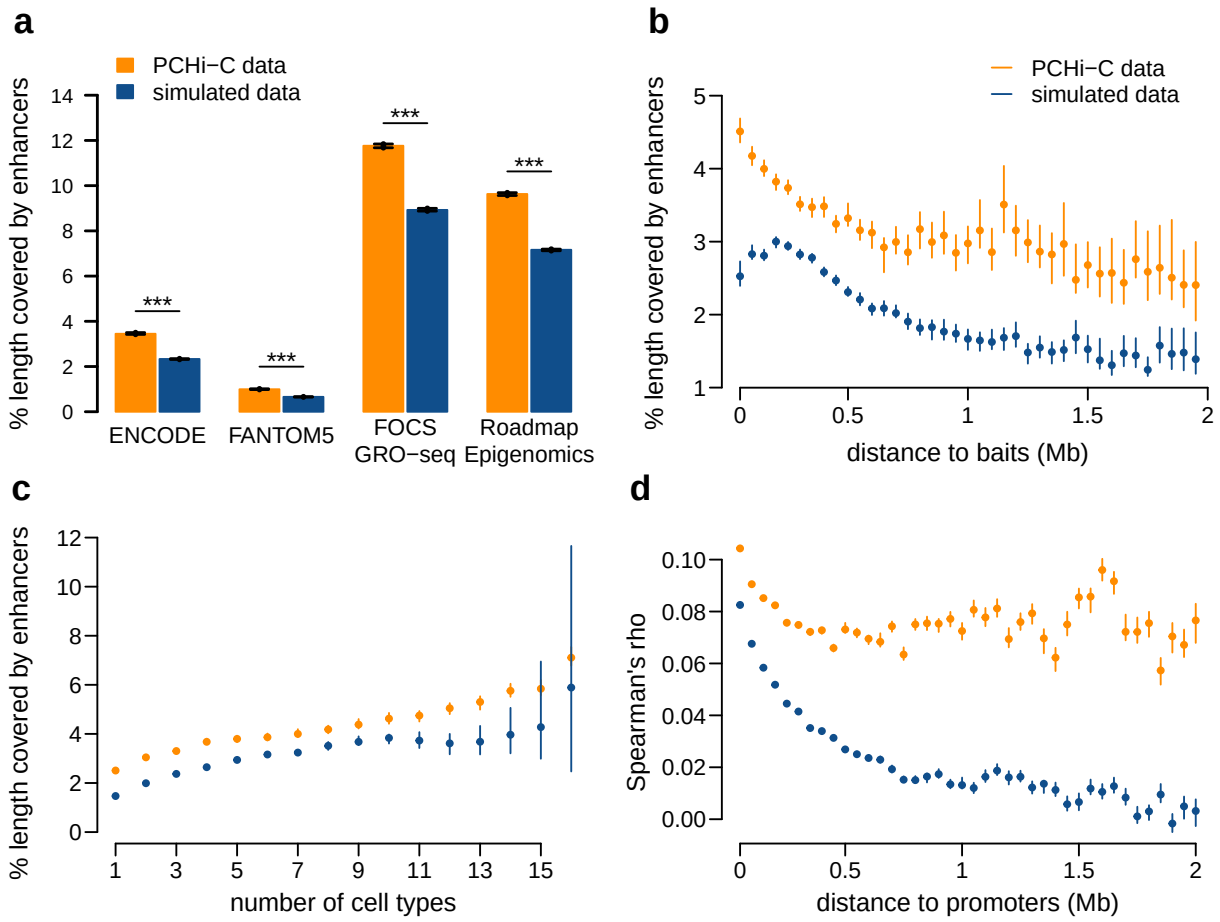

Figure 2: PCHi-C chromatin contacts describe regulatory landscapes. **a)** Average length fraction covered by predicted enhancers, for restriction fragments contacted in human PCHi-C data (orange) and simulated data (blue). **b)** Average length fraction covered by ENCODE enhancers, as a function of the distance between baits and contacted restriction fragments. **c)** Average length fraction covered by ENCODE enhancers, as a function of the number of cell types in which interactions are observed, for human restriction fragments. **d)** Distribution of Spearman's correlation coefficient between promoter and enhancer activity levels, for promoter-enhancer pairs in contact in PCHi-C data or in simulated data, according to the distance between them (Methods). **a-d)** Bars and dots represent mean values, vertical segments represent 95% confidence intervals of the mean, obtained with a non-parametric bootstrap approach (Methods).

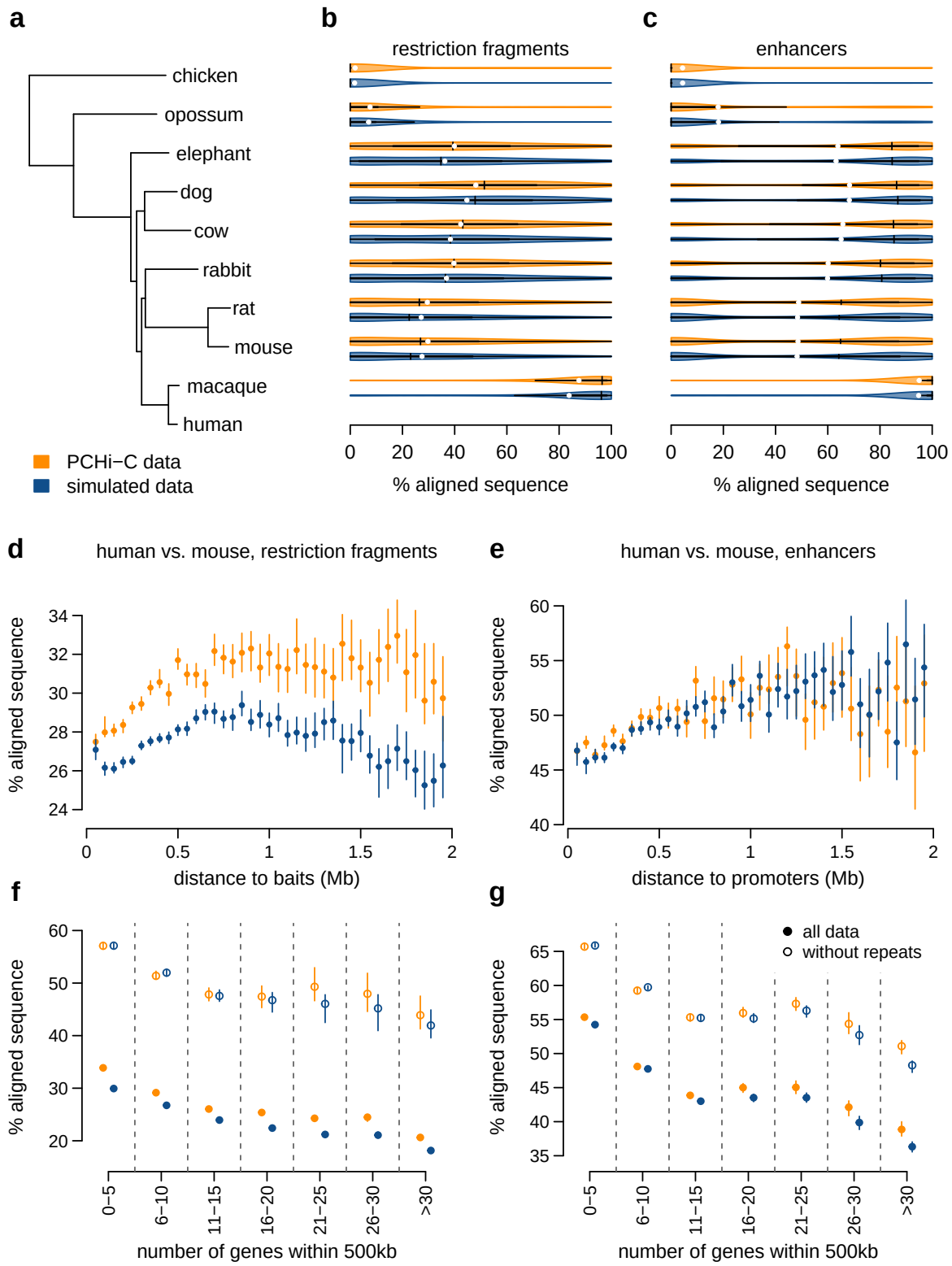

Figure 3: Genomic sequences contacted by promoters are conserved during evolution. **a)** Phylogenetic tree for the analyzed species. **b)** Sequence conservation levels for contacted restriction fragments. Violin plots represent the distribution of the percentage of aligned nucleotides, for pairwise comparisons between human and other species. Vertical segments represent median values; white dots represent mean values. **c)** Sequence conservation levels for contacted ENCODE enhancers. **d)** Conservation of contacted restriction fragments between human and mouse, as a function of the median genomic distance between restriction fragments and contacting baits. **e)** Same as **d**, for ENCODE enhancers. **f)** Conservation of contacted restriction fragments between human and mouse, as a function of the number of genes found within a maximum distance of 500kb from the restriction fragment. Filled dots represent all data; empty dots represent fragments that do not overlap with repetitive sequences. **g)** Same as **f**, for ENCODE enhancers. **d-g)** Dots represent mean values, vertical segments represent 95% confidence intervals of the mean, obtained with a non-parametric bootstrap approach (Methods).

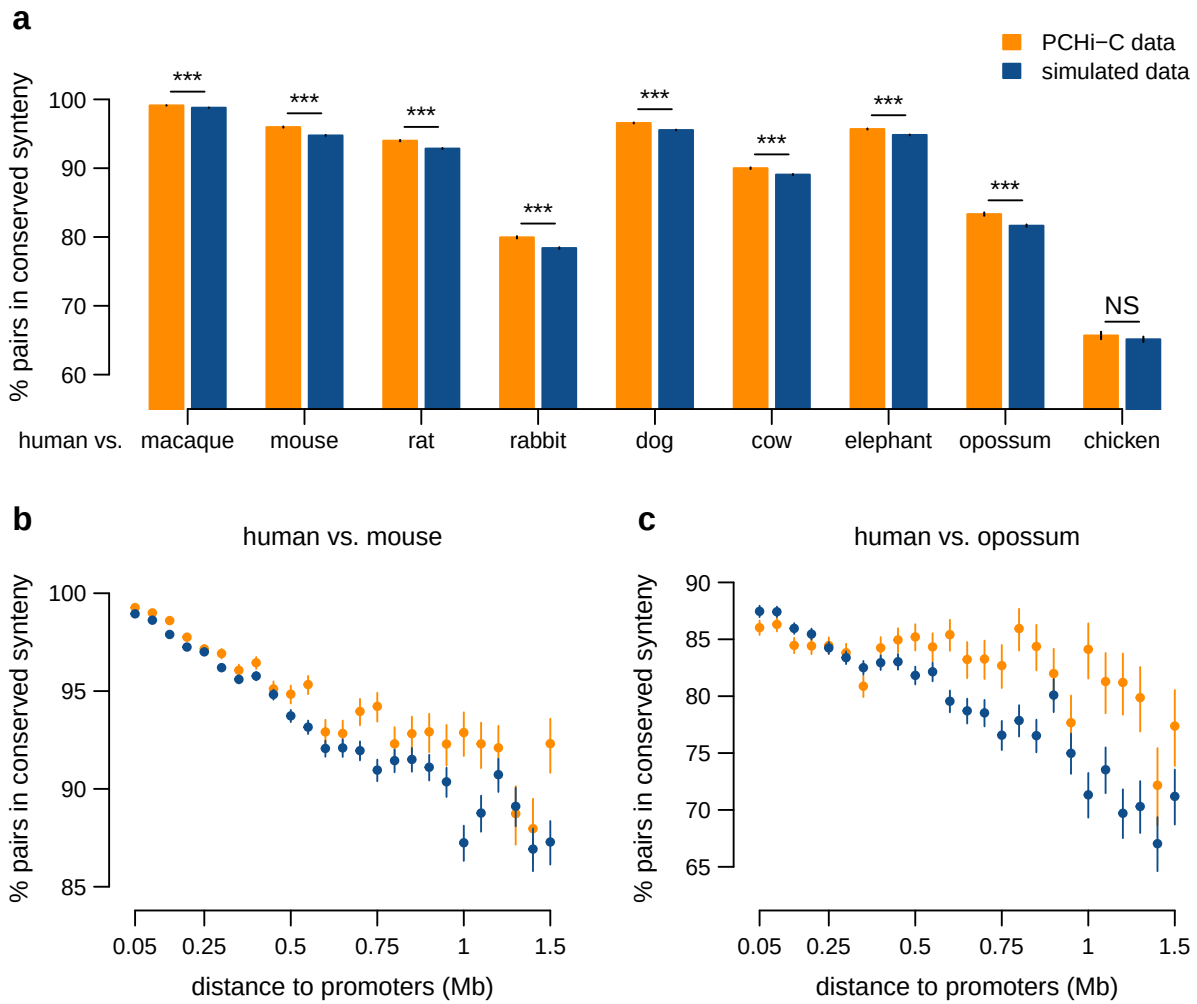

Figure 4: Pairs of promoters and enhancers involved in chromatin contacts are maintained in synteny. **a)** Proportion of human promoter-enhancer (ENCODE) pairs maintained in synteny (*i.e.*, found on the same chromosome, within a maximum distance of 2 Mb) in other vertebrate genomes, for PCHi-C data (orange) and simulated data (blue). **b)** Proportion of human promoter-enhancer pairs maintained in synteny in the mouse genome, as a function of the distance between them in human genome. **c)** Same as **d)**, for the comparison between human and opossum. **b,c)** Bars represent 95% two-sided confidence intervals for the proportions (Methods).

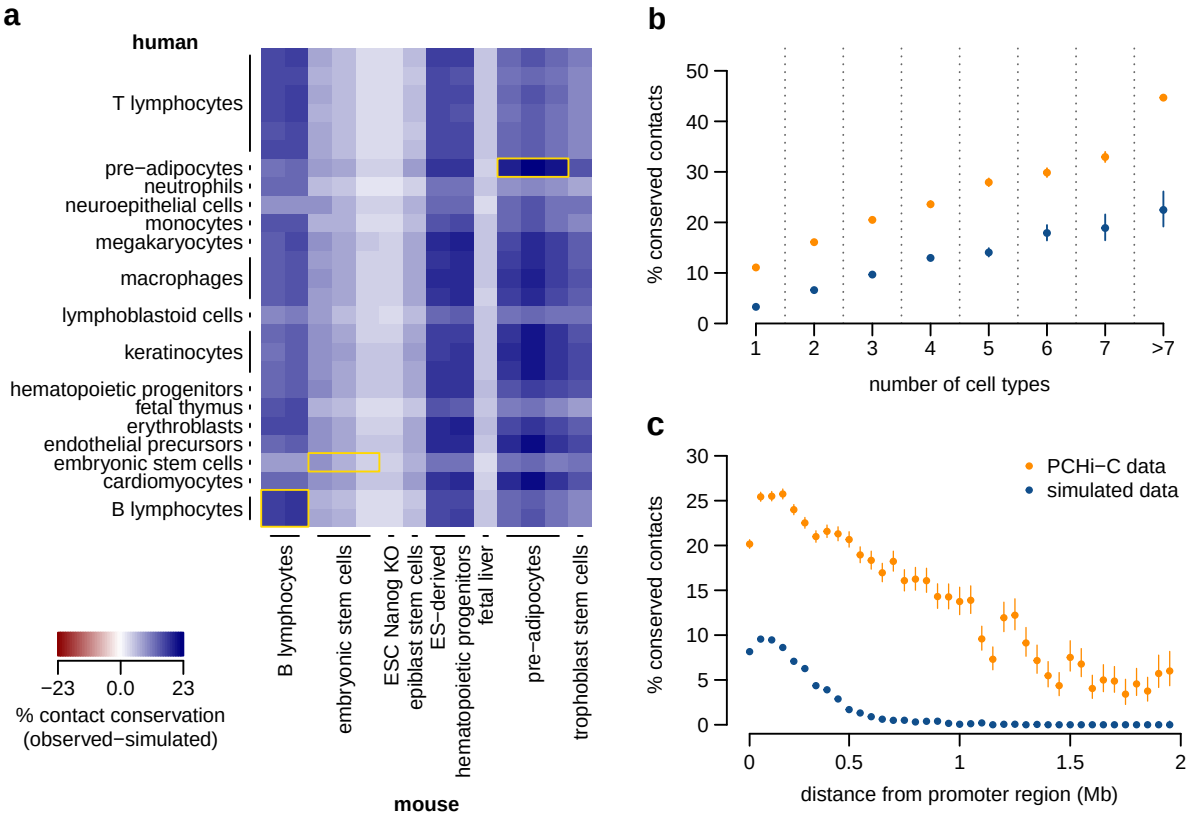

Figure 5: Promoter-enhancer contact maps are conserved during evolution. **a)** The heatmap displays comparisons between pairs of samples (one human sample and one mouse sample). For each pair of samples, we compute the difference between the percentage of conserved contacts in PCHi-C data and in simulated data. Yellow squares highlight comparisons between cell types sampled for both species (embryonic stem cells, B lymphocytes and pre-adipocytes). **b)** Proportion of human promoter-enhancer contacts conserved in mouse for PCHi-C data (orange) and simulated data (blue), as a function of the number of human cell types in which interactions are observed. **c)** Proportion of human promoter-enhancers contacts conserved in mouse, as a function of the distance between the two elements in the human genome. **b,c)** Bars represent 95% two-sided confidence intervals for the proportions (Methods).

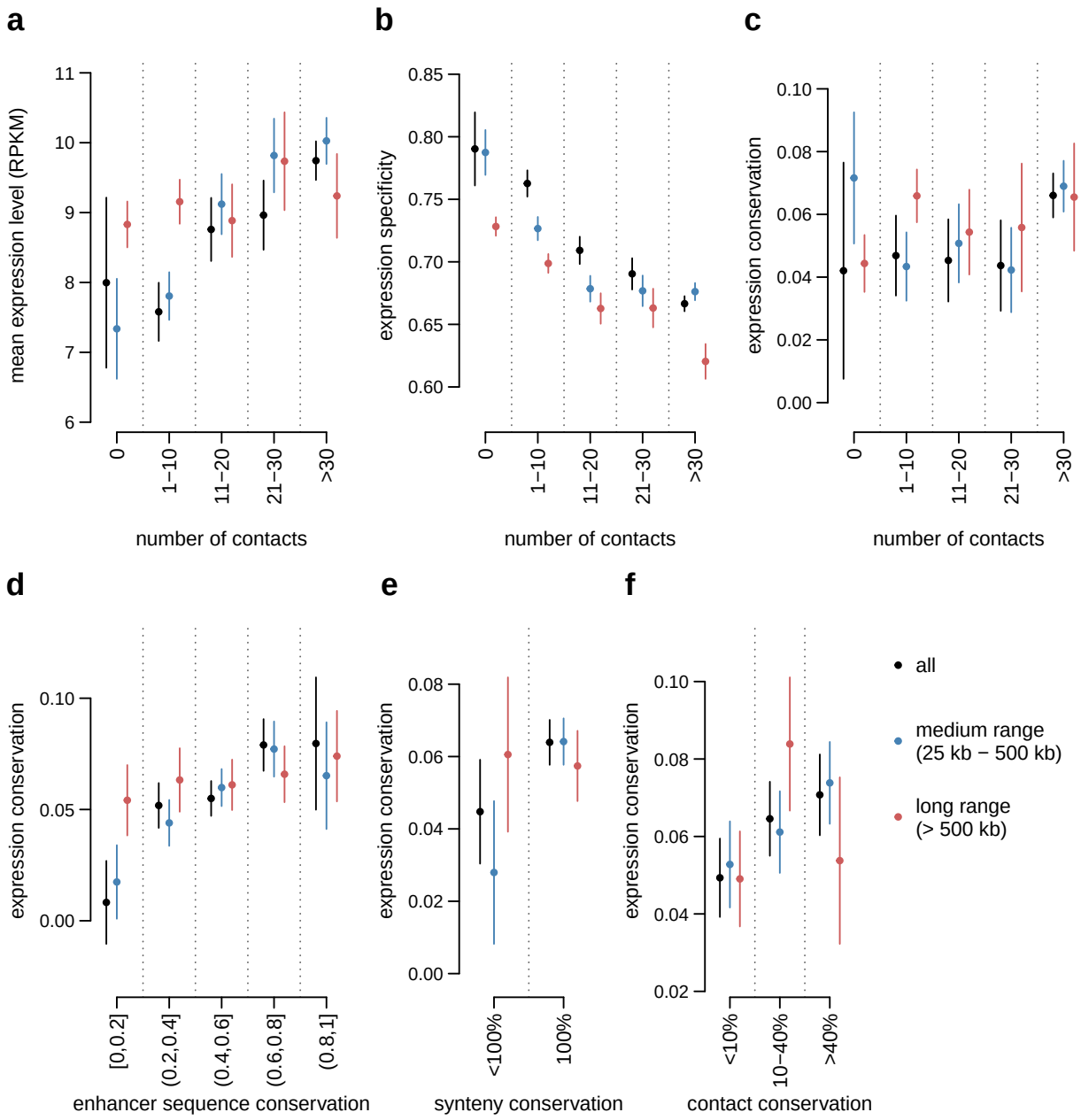

Figure 6: *Cis*-regulatory landscapes co-evolve with gene expression profiles. Gene expression conservation is measured with Spearman's correlation coefficient between human and mouse relative expression profiles, for pairs of 1-to-1 orthologous genes, across organs and developmental stages (expression data from Cardoso-Moreira et al., 2019). Expression conservation is further corrected to account for the effect of expression levels and of expression specificity with a multiple linear regression model (Methods). Enhancer predictions are taken from ENCODE. **a)** Average expression levels as a function of the number of contacted enhancers in PCHi-C data. Promoter-enhancer contacts are divided into two classes according to the distance between them: medium range (25kb-500kb) in blue, long range (above 500Kb) in red. The black dots represent the full set of promoter-enhancer contacts. **b)** Human gene expression specificity as a function of the number of contacted enhancers. **c)** Expression conservation as a function of the number of contacted enhancers in human PCHi-C data. **d)** Expression conservation as a function of the average sequence conservation of human contacted ENCODE enhancers. **e)** Expression conservation, depending on whether or not genes underwent least one break of synteny with the contacted enhancers between human and mouse genomes. **f)** Expression conservation as a function of the proportion of human promoter-enhancers contacts conserved in mouse PCHi-C data. **a-f)** Dots represent median values across all genes in a class, vertical segments represents 95% confidence intervals for the median.