Evidence for compensatory evolution within pleiotropic regulatory elements

Zane Kliesmete, Peter Orchard, Victor Yan Kin Lee, et al.

Genome Res. published online September 10, 2024
Access the most recent version at doi:10.1101/gr.279001.124

Evidence for compensatory evolution within pleiotropic regulatory elements

Zane Kliesmete[1], Peter Orchard[1,2], Victor Yan Kin Lee[1,3], Johanna Geuder[1], Simon M. Krauß[1,4], Mari Ohnuki[1,5], Jessica Jocher[1], Beate Vieth[1], Wolfgang Enard[1], Ines Hellmann[1,*]

[1] Anthropology and Human Genomics, Faculty of Biology, Ludwig-Maximilians Universit¨at Mu¨nchen, Munich, Germany

[2] Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA

[3] Section for Molecular Ecology and Evolution, Globe Institute, University of Copenhagen, Copenhagen, Denmark

[4] Department of Hematology, Cell Therapy, Hemostaseology and Infectious Diseases, University Leipzig Medical Center, Leipzig, Germany

[5] Faculty of Medicine Bldg.B, Institute for the Advanced Study of Human Biology (ASHBi), Kyoto University, Kyoto, Japan

[*] correspondence:

Dr. Ines Hellmann

Telefon +49 (0)89 2180-74336

Telefax +49 (0)89 2180-74331

hellmann@bio.lmu.de, www.anthropologie.bio.lmu.de

Keywords

Regulatory element evolution, pleiotropy, compensation, primates, cross-species, accessibility, expression, ATAC-seq, RNA-seq

Abstract

Pleiotropy, measured as expression breadth across tissues, is one of the best predictors for protein sequence and expression conservation. In this study, we investigated its effect on the evolution of *cis*-regulatory elements (CREs). To this end, we carefully reanalyzed the Epigenomics Roadmap data for nine fetal tissues, assigning a measure of pleiotropic degree to nearly half a million CREs. To assess the functional conservation of CREs, we generated ATAC-seq and RNA-seq data from humans and macaques. We found that more pleiotropic CREs exhibit greater conservation in accessibility, and the mRNA expression levels of the associated genes are more conserved. This trend of higher conservation for higher degrees of pleiotropy persists when analyzing the transcription factor binding repertoire. In contrast, simple DNA sequence conservation of orthologous sites between species tends to be even lower for pleiotropic CREs than for species-specific CREs. Combining various lines of evidence, we propose that the lack of sequence conservation in functionally conserved pleiotropic CREs is due to within-element compensatory evolution. In summary, our findings suggest that pleiotropy is also a good predictor for the functional conservation of CREs, even though this is not reflected in the sequence conservation of pleiotropic CREs.

Introduction

One of the initial perplexing revelations of the human genome project was the seemingly limited number of genes, which did not align with the increase in complexity compared to organisms such as yeast, worms, and flies. It became evident that this complexity must stem from gene regulation, with the probability that most genes play roles in multiple contexts throughout development and in various tissues. Considering the varying contexts of utilization in terms of location as well as

timing, it follows that mutations within the same gene can exert influence on multiple traits. This phenomenon is widely recognized as pleiotropy. In a molecular context, pleiotropy is frequently measured as the number of tissues in which a gene is expressed, a metric called expression breadth (Hastings, 1996; Duret and Mouchiroud, 2000).

The advent of microarrays and subsequent RNA-seq technology allowed for an impartial, genome-wide evaluation of expression breadth. As data accumulated, it became evident that expression breadth is in fact a very good predictor of the conservation of protein sequences. In particular, the ratio of the non-synonymous over synonymous substitution rate shows that pleiotropic genes tend to be more conserved than tissue-specific genes (Hastings, 1996; Duret and Mouchiroud, 2000; Zhang and WH Li, 2004). Moreover, the amount of sequence constraint associated with expression in a given tissue can vary considerably: Genes expressed in the brain tend to be more conserved than genes specific to other tissues, such as the liver (Kuma et al., 1995; HY Wang et al., 2007; Khaitovich et al., 2005). A similar pattern emerges in terms of expression level conservation: Brain-expressed as well as pleiotropic genes tend to have more similar expression levels across species than other genes (Khaitovich et al., 2005; Brawand et al., 2011; ZY Wang et al., 2020).

Naively, one would expect that a higher level of expression conservation would be achieved via a higher level of activity and sequence conservation of the associated cis-regulatory elements (CREs). However, most enhancers are tissue-specific (Gasperini et al., 2020) and show low activity conservation across species, although the target gene expression appears conserved (Paris et al., 2013; Villar et al., 2014; Berthelot et al., 2018). In contrast to the majority of tissue-specific CREs, the activity of more pleiotropic CREs indeed appears to be more conserved across species (Roller et al., 2021).

In line with this pattern, previously identified highly active pleiotropic enhancers in humans were found to have higher sequence conservation than tissue-specific enhancers across a large phylogeny (Andersson, Gebhard, et al., 2014; Singh and Yi, 2021). Of note, this analysis included only a couple of hundred enhancers, using a rather stringent definition of pleiotropy. Similar trends were

found analyzing human population data, i.e. a much shorter evolutionary time scale (Huang et al., 2017).

Promoters are much more likely to be functionally conserved than enhancers (Berthelot et al., 2018). In addition, promoters are more pleiotropic than enhancers, which is probably due to the fact that core promoters are more restricted in their spatial genomic location than enhancers which can be located megabases away from the targeted transcription start sites (TSS) (Villar et al. 2015; Andersson and Sandelin, 2020). Promoters are further distinguished by their shape: Broad promoters are large, thought to harbor multiple TSS and tend to be more pleiotropic. In contrast, narrow promoters are small, probably have only one TSS and are more likely to be tissue-specific (Andersson and Sandelin, 2020). Furthermore, evidence suggests that expression from broad promoters is less noisy and more robust towards mutations (Carninci et al., 2006; Schor et al., 2017; Sigalova et al., 2020; Floc'hlay et al., 2020) and in humans these broad promoters also show strong enrichment for CpG islands (Morgan and Marioni, 2018). At least in flies, this results in the counter-intuitive observation that although broad promoters are more robust and thus also more likely to be functionally conserved across species, overall they exhibit lower sequence conservation between species than narrow promoters (Schor et al., 2017). In summary, the relationship between pleiotropy and sequence conservation for CREs appears to be much more complicated than that between pleiotropy and protein-coding sequence conservation.

Here, we investigate the impact of pleiotropy on sequence and functional conservation in primates. To gauge pleiotropy, we thoroughly re-analyzed DNase hypersensitivity data from 9 primary fetal tissues (Bernstein et al., 2010), integrating across seven or more biological replicates per tissue to also identify tissue-specific CREs robustly. To assess functional conservation of the identified CREs, we obtained RNA-seq and ATAC-seq data from two human and two cynomolgus macaque neural progenitor cell lines. Furthermore, we obtained four different measures of sequence conservation: 1) a population genomic measure, 2) a conservation measure for the human lineage since the most recent common ancestor of humans and chimpanzees (Gronau et al., 2013), 3) a conservation score calculated for the primate phylogeny (Pollard et al., 2010) and 4) a scaled

measure of transcription factor binding site (TFBS) conservation.

Results

To investigate different aspects associated with varying degrees of regulatory pleiotropy, we identified putative CREs as DNase hypersensitive sites (DHS) in the Roadmap Epigenomics data, which provides comparable experiments for a wide selection of tissues (Bernstein et al., 2010). To ensure reproducibility, we included only tissues for which at least seven biological replicates of DNase-seq data were available, resulting in a set of nine tissues: adrenal gland, brain, heart, kidney, large intestine, lung, muscle, stomach and thymus (Fig. 1A,B). Hence, in this study, we focus on the regulatory evolution across somatic tissues. We called DHS for each tissue separately using a peak caller that utilizes replicate information to gauge certainty (Ibrahim et al., 2015), resulting in >1.1 million DHS ranging from ~80,000 sites detected in the large intestine to ~175,000 sites detected in the stomach (Fig. 1C). Of note, the number of detected DHS does not correlate strongly with the number of replicates across tissues (Pearson's $\rho = 0.23$, p-value = 0.55).

In analogy to how expression breadth has been used as a proxy for pleiotropy of genes, we merge overlapping DHS from different tissues and define the Pleiotropic Degree (PD) as the number of tissues in which we found a DHS if this DHS was significantly differentially accessible (DA) compared to tissues without a DHS at this location, resulting in ~443,000 union CREs stratified by PD (Supplemental Methods). We distinguish promoters and enhancers based on genomic distance, where we designate CREs within 2 kb of an active, annotated TSS (GENCODE v.32) as promoters and all other CREs within 1 Mb as enhancers (Fishilevich et al., 2017; McLean et al., 2010). Consistent with expectations, the majority of enhancers are tissue-specific (PD1) (Gasperini et al., 2020), while promoters are more likely to be pleiotropic (PD9) and CREs with an intermediate PD (1<PD<9) are generally rare (Fig. 1D). With a median size of 1.2 kb, PD9 promoters are the largest CREs (Fig. 1E, Supplemental Fig. S1A). Also the overlap among DHS is highest in PD9 promoters (Supplemental Fig. S1B), suggesting that their larger size is due to a higher information content

rather than being an artifact of concatenation. A large proportion of the pleiotropic promoters are CpG islands (76.7%) and the proportion of CpG island promoters increases with increasing PD (Fig. 1F). The same is true for enhancers, although enhancers are only very rarely CpG islands (3.2%).

Next, we wanted to investigate whether the PD of CREs has an impact on the expression of the associated genes. To this end, we integrated DHS with gene expression estimates from matching samples from Epigenomics Roadmap Project (Fig. 1G, Supplemental Fig. S2). As expected, we find a good correspondence between promoter PD and gene expression PD (χ^2-test p-value$<2.2^{-16}$). Specifically, PD9 promoters are associated with genes that are expressed in all 9 tissues (Pearson residual = 34.1, rank 1) and we also find tissue-specific promoters to be associated with tissue-specific genes (Pearson residual = 14.7, rank 2) (Fig. 1G, Supplemental Fig. S1C). We further modeled the associated gene expression using the presence of different CpG and non-CpG island CREs of varying PDs as predictors (Fig. 1H; Supplemental Fig. S1D). The total amount of variation in expression levels that can be explained by the model is 24% (CI: 23.8-24.3%), while the same model with a shuffled PD label can only explain 19% (CI: 18.8-19.8%) (see Methods). Inspecting the scaled coefficients of the mixed effects model suggests that PD9 promoters have the largest activating effect on expression, followed by PD9 and PD1 enhancers.

Characterization of transcription factor binding site repertoire across pleiotropic degrees

Under the premise that CREs regulate gene expression by binding transcription factors, we continued by characterizing TFBS associated with CREs of varying pleiotropic degrees. To this end, we collected non-redundant position weight matrices (PWMs) of 643 binding motifs (Fornes et al., 2020) belonging to 561 TFs that we found to be expressed in at least one of the investigated tissues (Fig. 2A). Almost half of all expressed TFs (237 out of 561, 42%) were present in all tissues, i.e. pleiotropic, while 94 (17%) showed tissue-specific expression. In agreement with others (Vaquerizas et al. 2009), we found that the brain has the highest proportion of tissue-specific TFs

(8.5%).

Next, we evaluated the overall binding potential of a TF to a CRE using Cluster-Buster (Frith et al., 2003) (see Methods). We found that TFBS diversity increases with pleiotropy for both enhancers and promoters (Fig. 2B). When normalizing TFBS diversity by the CRE width yielding a measure for TFBS density, we observe decreasing density with increasing PD (Supplemental Fig. S6A, see also Supplemental Fig. S6B for GC content normalization). This suggests that the complex binding repertoire in PD9 CREs is mainly achieved through their increasing width. Next, we investigated whether tissue-specific and pleiotropic CREs have binding sites for the same TFs or whether preferences exist. For most TFs (62%) we do not find an enrichment of predicted TFBS in any PD category (Fig. 2C). Of the remaining motifs, 159 (24.9%) are over-represented in CREs that are specific for one of the tissues and 84 (13.1%) motifs are enriched in the PD9 CREs. Gene-set enrichment analysis shows that over-represented TF motifs in brain-specific CREs are associated with neural differentiation (Fig. 2E). Most prominently, this is driven by OLIG2 that is essential for oligodendrocyte development (Zhou and Anderson, 2002; Jakovcevski et al., 2009; Yu et al., 2013), as well as by NEUROD1, NEUROD2 and NEUROG1 that are important for neuron development (Olson et al., 2001; Sun et al., 2001; Messmer et al., 2012; Pataskar et al., 2016). Other tissues also show an enrichment for TFBS of tissue-specific TFs: For example, TFBS that are over-represented in heart-specific CREs include motifs of MEF2C, TBX20 and NKX2-5 (Fig. 2F), which are essential for cardiac muscle development (He et al., 2011; Schlesinger et al., 2011; Grunert et al., 2016).

In contrast, PD9 CREs contain more predicted binding sites for TFs associated with basic cellular processes, such as transcription regulation in connection with cell cycle and stress response (Fig. 2D). These motifs are more GC-rich and tend to have a higher information content than PD1-enriched motifs or motifs without any preference (Supplemental Fig. S6C,D). In addition, these elements are enriched for the so-called ‘Stripe’ TFs (odds ratio = 10.53, Fisher’s exact p-value = 3^{-12}) (Zhao et al., 2022), that include SP, KLF and ZBTB family members.

The impact of pleiotropy on the evolutionary conservation of regulatory activity

To investigate the relationship between CRE pleiotropic degree and their evolutionary conservation, we generated RNA-seq and ATAC-seq data from iPSC-derived neural progenitor cell lines (NPCs) from humans and cynomolgus macaques (Supplemental Fig. S3A,B, Supplemental Methods). We intersected the detected genes and accessible peaks with the processed Epigenomics Roadmap data to assign a pleiotropic degree to genes and peaks (Fig. 3A). Of ~66,000 and ~69,000 called peaks in human and macaque NPCs, 76.8% and 64.3% show an overlap with our tissue-based CRE annotations, respectively (Supplemental Fig. S3C) (see Methods for the cross-species liftOver strategy). In line with the expectation that pleiotropic CREs are more conserved, the amount of CRE overlap with NPC ATAC-seq peaks increases with increasing PD and is generally higher for promoters than for enhancers (Fig. 3B). Moreover, the activity of PD9 CREs is also more conserved between humans and macaques: Of all overlapping PD9 CREs, 88% were detected to be active in NPCs from both species, while this was the case for 15% of the PD1 CREs. Of note, the observed association between PD and activity conservation in human and macaque NPCs is not only due to the increased regulatory activity that might generate a higher probability for PD9 elements being detected as peaks (Fig. 3B). Instead, even without stratifying by whether a peak was called, we observe a decrease in differential activity with increasing PD, measured by absolute log_2-fold changes using DESeq2 (Love et al., 2014) (Fig. 3E). A very similar conservation pattern emerges, when we leverage the data from Roller et al. (2021) that encompasses CRE activity measures for 10 mammalian species in 4 tissues. To begin with, the assigned PD in our study that was solely based on human tissues is similar to the CRE PD in other primates, rhesus macaque and marmoset, and is also consistent with the average level of pleiotropy across all 10 mammalian species (Supplemental Fig. S4A-D). Moreover, also the CRE activity conservation across the 10 species confirms that a higher PD is associated with evolutionarily conserved activity (Fig. 3C,D).

To investigate the effect of differentially active CREs on the expression of associated genes, we test whether genes with a DA CRE (BH-adjusted p-value<0.1) are also more likely to be DE (BH-adjusted p-value<0.1; $|log_2$-fold change|>1) in our NPCs. When comparing the effect of CREs

from different PD categories, PD9 promoters have a larger impact on the DE-status of the associated gene than PD1 promoters, while the difference between enhancers from different PD categories is non-significant (Fig. 3F). In summary, in agreement with Roller et al., 2021, we also find that the activity of pleiotropic CREs is evolutionarily more conserved between species than the activity of tissue-specific CREs. Moreover, if the activity of a pleiotropic promoter changes, such changes are more likely to have downstream effects, i.e. to impact the expression of associated genes.

Cross-species sequence conservation is lowest in pleiotropic CREs

Up to this point, pleiotropy shows the expected effect on gene regulation and CRE activity conservation. Here, we investigate how this functional conservation is reflected in the underlying DNA sequence. We focus on three measures of sequence conservation: 1) recent selection quantified as the number of weakly deleterious sites in humans E.W. (Gronau et al., 2013) (Fig. 4A,B), 2) selection on the lineage since the most recent common ancestor of humans and chimpanzees, quantified as the fraction of sites under (strong) negative selection ρ (Gronau et al., 2013) (Fig. 4C,D) and 3) the average phyloP and phastCons scores across a primate phylogeny (Pollard et al., 2010) (Supplemental Fig. S5A,B, Supplemental Methods). The main difference among the three measures is the evolutionary time across which sequence conservation is averaged. Since PD was assessed in human samples, the E.W. measure provides the closest match to our measure of pleiotropy. For ρ, phyloP and phastCons we average the strength of selection over longer evolutionary times. It should be noted that variants emerging within a population may undergo recombination, whereas mutations occurring after speciation remain on separate haplotypes. In line with our expectations, we indeed find that the number of weakly deleterious sites (E.W.) increases with the pleiotropic degree for both promoters and enhancers (Fig. 4A). This observation aligns with the conservation of CRE accessibility: Across all PD categories, we observe a higher prevalence of weakly deleterious sites in CREs that are open in both species (Fig. 4B). In contrast, when using ρ as a measure of conservation, we only find a higher sequence conservation

for tissue-specific CREs (PD1-3) with conserved accessibility, while it appears that accessibility conservation is not reflected in the sequence conservation of pleiotropic CREs (Fig. 4D). Overall ρ suggests that PD9 CREs have the lowest fraction of negatively selected sites compared to other PD categories (Fig. 4C). These patterns persist also when inferring E.W. and ρ for the CGI and non-CGI fractions of the PD categories separately (Supplemental Fig. S5C,D). This result also remains when we use the average phyloP or average phastCons score across a 10-species primate phylogeny as a measure of conservation, which confirms PD9 CREs as the PD category with the lowest cross-species sequence conservation (Supplemental Fig. S5A,B). In summary, even though the number of weakly deleterious sites within a CRE increases with pleiotropy in humans, this is not reflected in sequence conservation across species.

Tissue-specific effects on CRE sequence conservation

So far, we have not considered the possibility that different tissues might impose different amounts of constraint to the DNA sequence of the CRE. By separating CREs by the tissues in which they are utilized, we find that brain CREs are clearly under more constraint than those of other tissues. Nevertheless, also for the brain the number of weakly deleterious sites increases with PD, showing that, although to smaller amounts, activity in other tissues still adds to the overall constraint (Fig. 5A). Again, this is not true when considering substitutions on the human lineage as used in the measure ρ (Fig. 5B). Here, brain-specific CREs show the most constraint, much more than pleiotropic PD9 CREs, which by definition are utilized also in the brain. To exclude the possibility that the brain effect on the PD9 elements is diluted by the merging of DHS across tissues, we contrast the ρ of the brain peak sequence with adjacent sequences that are part of the same merged CRE but are open only in other tissues (Fig. 5C). For PD9 CREs, brain peak sequences show lower sequence conservation on the human lineage than the adjacent sequence utilized only by other tissues, while for less pleiotropic CREs the part that is used in the brain is under much more constraint (Fig. 5D). In summary, even though we find tissue-specific effects, this cannot explain the overall pattern of the relatively low sequence conservation of pleiotropic CREs. It remains that

for pleiotropic CREs there is no simple relationship between sequence and functional conservation between species.

Pleiotropic CRE TF repertoire is conserved, not the binding sites

To explain the apparent mismatch between functional and sequence conservation in PD9 CREs across primates, we continued to analyze levels that are intermediate between sequence conservation (farther from function) and accessibility conservation (closer to function), which are CpG content, TFBS repertoire and position conservation between human CREs and their orthologous sequences in cynomolgus macaques. To begin with, we find that the conservation of CpG and GC content increases with PD and is highest for pleiotropic promoters (Supplemental Fig. S5E,F). This coincides with the increase of the proportion of CpG island CREs (Fig. 1F) and suggests that the CpG island property at high PD CREs is conserved across species, thereby showing patterns closer to the functional side. Next, for each CRE we calculated the binding potential for all expressed TF motifs and approximated TFBS repertoire conservation using the average pairwise Canberra distance between species ($1 - \overline{d}_{C_{MH}}$). When we contrast CREs with conserved and non-conserved openness between humans and macaques, we find that functionally conserved CREs also show higher repertoire conservation across all PD categories (Fig. 6B).

TFBS repertoire conservation generally increases with pleiotropy in all tissues (Fig. 6C). There is a simple relationship for promoters for which repertoire conservation is highest for PD9 and lowest for PD1 CREs (Fig. 6A). Also for enhancers, albeit less pronounced, the average repertoire conservation in PD9 CREs (0.66, SEM 0.166) is considerably higher than in PD1 CREs (0.62, SEM 0.241), which is in contrast to what we observed for sequence conservation, again showing higher similarity to the patterns of functional conservation (Fig. 3E). To validate our findings about TFBS repertoire conservation, we integrated our PD scores with ChIP-seq data on 4 liver TFs in five mammalian species (Ballester et al., 2014) (Supplemental Fig. S7A,B, Supplemental Methods). Consistent with our estimates, experimentally inferred TF binding is also more

conserved if it resides within a pleiotropic as compared to binding in liver-specific CREs (Supplemental Fig. S7C,D).

To understand how the high repertoire conservation is achieved for PD9 CREs in spite of having the lowest sequence conservation, we analyzed the positional conservation of TFBS as a third intermediate metric using Jaccard similarity index (IoU_{MH}) (Fig. 6D). We find that the average repertoire conservation appears to be unrelated to the positional conservation in high PD categories (Fig. 6E). The pattern of positional conservation across PD categories resembles the pattern that we observe for sequence conservation (Fig. 6F). In summary, while CRE sequence and TFBS positions are least conserved in PD9 elements, CpG content and TFBS repertoire are in agreement with the more functional metrics - accessibility and expression conservation - in that they show the highest conservation in PD9 elements. These patterns are consistent with a mechanism of compensatory evolution (Fig. 6G).

The PD9 promoter of *ATXN-3* gene as an example

To illustrate the within-CRE compensatory evolution of TFBS within a PD9 promoter, we took a closer look at the promoter of the ubiquitously-expressed protein-coding gene *ATXN3*. It is an important factor for the regulation of the degradation of damaged proteins (Schmitt et al., 2007; Gao et al., 2015; Feng et al., 2018). This gene plays a role for the brain, as its malfunction can lead to neurodegenerative diseases such as spinocerebellar ataxia (Evers et al., 2014). The *ATXN3* promoter has relatively low sequence conservation (34th percentile) and low TFBS binding site conservation (49th percentile), but high TFBS repertoire (77th percentile), accessibility and expression conservation (Fig. 7A-E).

To investigate a few likely relevant TFs closer, we overlapped our TFBS data with published ChIP-seq data from human neural cells available in the GTRD database (Yevshin et al., 2018) and visualized the binding sites of the 2 TFs (MYCN, POU3F2) annotated to be involved in neurogenesis (Gene Ontology Biological Process term GO:0022008) (Fig. 7H). Both of their motifs are moderately complex as shown by their information content (MYCN: IC = 11.8, POU3F2: IC =

13.7) (Fig. 7I,J). Both promoter orthologues show high binding potential for both TFs. Humans have 6 and macaques 5 MYCN binding sites and both have one POU3F2 binding site, which is also reflected in similar ATAC-seq peak-shapes (Fig. 7F,G). However, only 3 of the 10 binding sites are positionally conserved between the species. This serves as an example of how the large disagreement between sequence, TF binding site conservation and TFBS repertoire might co-occur.

Discussion

Pleiotropy has been shown to be the best predictor of both protein coding sequence conservation (Hastings, 1996; Duret and Mouchiroud, 2000; Zhang and WH Li, 2004) and gene expression levels (Khaitovich et al., 2005; Brawand et al., 2011; ZY Wang et al., 2020). Here, we investigate the effect of pleiotropy on the evolution of *cis*-regulatory elements (CREs). In agreement with the findings by Roller et al. (2021), we find that measures close to CRE function, such as accessibility and TFBS repertoire conservation, indeed show the expected higher conservation for more pleiotropic CREs. Similarly, a measure of conservation based on human diversity data also shows a trend for higher conservation in more pleiotropic CREs. However, we found that this is not reflected in the sequence and TFBS positional conservation between species (Fig. 6H). These observations imply that a simple model of purifying selection alone is insufficient to explain the effect of pleiotropy on CRE evolution and suggest a role for compensatory evolution.

Zooming into tissue effects, in line with previous investigations on brain evolution (Kuma et al., 1995; HY Wang et al., 2007; Brawand et al., 2011; Roller et al., 2021), we find that brain-specific CREs show by far the highest sequence conservation irrespective of the measure. Therefore, we would expect that pleiotropic CREs, which are by definition also open in the brain, show the highest sequence conservation. However, looking at between-species sequence conservation, the sub-sequences of PD9 CREs that are open in the brain are even less conserved than the adjacent sequences (Fig. 5D). This confirms the notion that the structure and evolution of PD9 CREs is inherently different, in that it allows for functional conservation without much sequence conservation.

Indeed, several basic structural properties of PD9 elements distinguish them from less pleiotropic CREs. They tend to be larger, have more CpGs and a higher GC content. Moreover, PD9 elements show an over-representation of GC-rich motifs that are associated with TFs that tend to be involved in more basic cellular processes. Among those, we also find enrichment for binding sites of a recently described group of highly cooperative TFs that prolong CRE openness (Zhao et al., 2022). It should also be noted that the majority of PD9 CREs are promoters and PD9 promoters share many properties with broad promoters that were defined via the size of CAGE peaks (Andersson, Gebhard, et al., 2014). Even though this classification is based on a completely different concept, broad promoters were also shown to be more pleiotropic, active and CpG-rich. Indeed, as observed for PD9 CREs, broad promoters also have an increased substitution rate. Moreover, broad promoters have been shown to be more robust than narrow promoters, in that they show less expression noise across haplotypes in *Drosophila* (Floc'hlay et al., 2020; Schor et al., 2017). Similarly, CpG island promoters have been found to induce more stable expression in humans (Morgan and Marioni, 2018). Mechanistically, this picture fits with the notion that GC-rich regions facilitate combinatorial binding (Zhao et al. 2022), which has been shown to lead to evolutionarily more stable TF binding across closely as well as distantly related mammalian species (Stefflova et al., 2013; Ballester et al., 2014). In the same vein, Hagai et al. (2018) found that the regulatory response of genes associated with CpG islands to an immune stimulus is more conserved than that of genes associated with a TATA-box. In summary, there is ample evidence that large CpG island promoters are functionally robust while having high substitution rates.

We also find high substitution rates in PD9 enhancers, which share many features with PD9 promoters. Moreover, promoter and enhancer functionality frequently switches over evolutionary time (Roller et al., 2021). Hence, we suggest that the main differences in the evolutionary patterns observed for promoters and enhancers are closely linked to their degree of pleiotropy. Most enhancers show strong tissue preferences, placing them in our PD1 category. Consistent with multiple other studies (Danko et al., 2018; Berthelot et al., 2018; Roller et al., 2021; Schmidt et al., 2010), we find that only a relatively small fraction of enhancers is conserved between species in

terms of accessibility and that this fraction is strongly enriched for pleiotropic CREs irrespective of their classification as enhancers or promoters. In fact, CRE conservation across the genome is so low (Doniger and Fay, 2007; Horton et al., 2023), that a simple evolutionary model cannot explain it. Instead, the presence of proto-enhancers is required (Tuğrul et al., 2015; Emera et al., 2016).

In addition, the observed high CRE turnover rates appear to be inconsistent with the relatively low rates of change in gene expression levels. This discrepancy has prompted the proposal of compensatory evolution as a prevalent mechanism for CREs (Schmidt et al., 2010; Berthelot et al., 2018). The phenomenon of CREs at non-orthologous genomic positions in different species exhibiting the same function and being able to compensate for one another has been documented in detail for several cases (Ludwig et al., 2000; Arnold et al., 2014; Domené et al., 2013; Emera et al., 2016). It is related to the observation that a lot of function is encoded redundantly within a gene's regulatory landscape by so-called shadow enhancers (Hong et al., 2008; Osterwalder et al., 2018; Wunderlich et al., 2016). Osterwalder et al. (2018) showed that the deletion of one strong enhancer did not have an effect on the phenotype as long as the shadow enhancer was still active. This clearly demonstrates the presence of epistasis. If multiple similarly fit haplotypes co-exist and different ones can get fixed in different species, this would facilitate compensatory evolution across CREs.

Other properties of CREs suggest that there is also a lot of epistasis within the same CRE. The billboard model (Kulkarni and Arnosti, 2003; Arnosti and Kulkarni, 2005) and the TF-collective model (Junion et al., 2012) of enhancer activity suggest that two CRE haplotypes with shifted but similar TFBSs should be functionally equivalent. It follows that the mutations that create these two haplotypes will also have non-additive effects on fitness. Moreover, some studies showed that binding to a high-affinity site is facilitated by many neighboring low-affinity binding sites (Crocker et al. 2016), thus providing the raw material for high TFBS turnover rates (Tuğrul et al., 2015).

Combining all the evidence, we suggest that within-element compensation of TFBS is a common mode of evolution for pleiotropic CREs. This mode of evolution would explain the apparent disparity between the cross-species and within-species sequence conservation (Fig. 4). Moreover, it would also explain the disparity between the low sequence and the high functional conservation

between species as observed in our ATAC-seq and RNA-seq data: If different, functionally equivalent haplotypes got fixed in different species, this should lead to high sequence divergence while the open chromatin state and downstream gene expression remain conserved (Fig. 6H). Furthermore, we show that this is achieved through binding repertoire, not binding site conservation. In summary, we think that compensatory evolution is a prevalent mode for evolution of regulatory elements. While for tissue-specific CREs compensation occurs between CREs over larger genomic distances, for pleiotropic CREs compensation mainly occurs within the same element leading to high sequence divergence.

Methods

CRE effect on gene expression across tissues

GENCODE v.32 (Harrow et al., 2012) was used to identify TSS (5' of transcripts) of the expressed genes (RNA-seq RPKM >1 in 50% of the samples) in each tissue. CREs within 2 kb of a TSS are designated promoters and associated with all TSSs within that distance. CREs within 1 Mb of a TSS are deemed enhancers and associated with the 2 closest TSSs in each direction, unless the distance to one TSS is >10x smaller than to the other TSS - in that case only the closest TSS is assigned. In total, 397,228 CREs (89.6%) were assigned.

Log mean expression is estimated using a linear mixed effects model with tissues as a random effect and the distance to TSS weighted (*d*) numbers of CpG Island and non-CpG Island promoters and enhancers as fixed effects:

$$log_2(\bar{e}) \sim \sum_{i \in PD} \sum_{P/E} \sum_{CGI} \beta_i \sum_{CREs_g} \frac{1}{log_2(d+2)} + Zb_{tissue}$$

To assess the effect of PD, we shuffled the PD labels across all CREs.

Cross-species gene expression and accessibility analysis

Previously generated iPSCs from human (*Homo sapiens*) and cynomolgus macaque (*Macaca fascicularis*) (Geuder et al., 2021) were differentiated to neural progenitor cells via dual-SMAD

inhibition (Chambers et al., 2009; Ohnuki et al. 2014).

To generate RNA-seq data, we used 3 clones of 3 human individuals and 4 clones of 2 cynomolgus macaque individuals. After 5 days of differentiation the cells showed characteristics of neural progenitors and were harvested at days 5, 7 and 9. cDNA libraries were generated using prime-seq (Janjic et al., 2022) and processed with zUMIs (Parekh et al., 2018). Human samples were mapped to GRCh38, GENCODE v.32. Cynomolgus samples were mapped to macFas6 (Jayakumar et al., 2021) and for gene annotation, we transferred human GENCODE v.32 to macFas6 using Liftoff (Shumate and Salzberg, 2021). Only genes from GRCh38 detected also in macFas6 were further included. Genes with UMI counts in at least 28.57% (6/21) samples were kept, resulting in a 14,608 gene set.

To generate ATAC-seq data, iPSCs of 2 clones from 2 human individuals and 2 clones from 2 cynomolgus macaque individuals were differentiated as described above. Libraries were generated using the Omni-ATAC protocol (Corces et al., 2017) with minor modifications. Reads were mapped to GRCh38 and macFas6 using BWA-MEM2 (Vasimuddin et al., 2019). Peak calling was done using Genrich (https://github.com/jsh58/Genrich). Reciprocal liftOver (RLO) was used to convert CRE coordinates from GRCh37 to GRCh38 and from GRCh38 to macFas6. For example, coordinates from GRCh38 were converted to macFas6 coordinates and defragmented by merging all hits within 40 bp. The resulting macFas6 region was then reciprocally converted to GRCh38 coordinates and we kept only CREs for which those coordinates overlapped with the original ones. We identified RLO matches for 99.7% CREs in GRCh38 and 88% in macFas6 and removed CREs with RLO match width beyond [1.2 × GRCh37; 0.8 × GRCh37] and 31 CREs that contained Ns. This resulted in an orthologous coordinate set containing 385,111 CREs.

Macaque ATAC-seq reads in macFas6 and human reads in GRCh38 were counted within the RLO PD-CRE coordinates. Only CREs that overlapped with an ATAC-seq peak by 10% relative to the width of both the PD-CRE and the ATAC-seq peak in at least one species were kept for differential accessibility (DA) analysis (n=65,753). Differential gene expression (DGE) and DA analyses were performed separately using DESeq2 (Love et al., 2014), using species as the predictor (BH-adjusted

p-value < 0.1). For DGE we also require $|log_2\text{-FC}| > 1$.

To analyze the relationship between CRE DA and gene DE, for each gene's landscape we quantified the number of DA and non-DA CREs in each of the 6 categories: assignment (promoter / enhancer) × PD1 / PD2-8 / PD9. To investigate what effect DA of a certain CRE group has on the odds of being associated with a DE gene, we chose matching landscapes that differed only by the DA status of 1 CRE in the respective category and used a BH-corrected Fisher's exact test to detect association with DE genes. This resulted in comparing 6,882 gene landscapes.

For analysis where we used the state of the ATAC-seq peak (open / closed) as an indicator for peak conservation, we required that conserved peaks overlap by 10% of their width between human LO peaks and macaque peaks in macFas6, keeping only 1-to-1, 0-to-1, 1-to-0 and 0-to-0 overlaps.

Estimation of CRE phylogenetic conservation and PD across mammals

We used cross-species histone modification ChIP-seq data from Roller et al. (2021). It contains samples from 4 tissues of 10 mammalian species. Since it does not contain human data, we used the macaque genome space as an anchor. Coordinates of our human PD annotations were RLO to rheMac10, yielding 426,102 RLO CREs. All macaque CRE coordinates (174,860) were collected across the 9 pairwise species files. Overlapping macaque CRE coordinates carrying different functional annotations in different tissues were merged, resulting in 141,649 merged CREs. Orthologous human and macaque CREs from the two datasets were identified by requiring that the overlapping part is >10% relative to the width of both species CREs. This resulted in 47,377 (66%) 1-to-1 orthologues, 580 (1%) 1-human-to-many-macaque and 24,160 (33.4%) many-human-to-1-macaque CREs. PD of the orthologous CRE in the other species was inferred using macaque orthologue annotation tables from Roller et al. The phylogenetic tree from Bininda-Emonds et al. (2007) was trimmed to the 10 species. Activity conservation for each CRE was calculated using relative branch length of the phylogenetic tree for that CRE versus the full 10-species tree.

Quantification of transcription factor binding

Two sets of TF Position Weight Matrices (PWMs) of the expressed TFs in 1) 9 human tissues (643 motifs from 561 TFs) and 2) our human and cynomolgus macaque NPCs (521 motifs from 446 TFs) were extracted from the JASPAR core vertebrate set (Fornes et al., 2020). These PWMs were provided to Cluster-Buster (Frith et al., 2003). In each species for each TFBS cluster of a CRE, we ranked TF motifs based on their strongest binding site. For all subsequent analyses, for each CRE we only considered TF binding motifs that were among the 10% strongest in at least 1 cluster in at least 1 species.

Calculation of TFBS diversity

For each CRE in each species, we measured TFBS diversity by Shannon entropy (*H*) (Shannon, 1948) where we considered a CRE as a collection of $i = 1, 2, .., n$ motifs of varying frequency (p):

$$H = -\sum_{i}^{n} p_i \ln ln\, p_i$$

where $p_i = \frac{S_i}{\sum_i^n S_i}$ and S_i is the cumulative motif score per motif i per CRE. H is further converted to the true diversity e^H (Hill, 1973; Jost, 2006).

CRE PD ranking per motif to detect over-represented motifs

Per tissue. We considered only expressed TF binding to CREs that are open in that tissue. For each PD category and motif, the relative binding frequency was obtained as the fraction of CREs that have binding sites for that motif:

$$f_{PD,i} = \frac{C_{PD,i}}{C_{PD}}$$

where PD indicates a PD category, i indicates a motif, $C_{PD,i}$ is the count of CREs with motif i binding site(s) present, C_{PD} is the total CRE count. Having obtained relative frequencies per PD, we ranked PD categories for each motif. Fold changes of the binding fraction of rank-1 PD relative to the average fraction for each motif i:

$$FC_{PD_{(1)},i} = \frac{f_{PD_{(1)},i}}{\frac{1}{9}\sum_{rank=1}^{9} f_{PD_{(rank)},i}}$$

Across tissues. We focused on motifs that had the highest binding fractions (rank-1) to either PD9 or PD1. To obtain the PD9-enriched motifs, we identified TF motifs for which PD9 CREs had rank-1 in all tissues. As the PD1-tissue-specific motifs we considered the ones that have PD1 with rank-1 only in that particular tissue, but not in the others.

We selected universal 'Human Stripe Factors' from Zhao et al. (2022), using a detection rate cutoff of 0.9 across samples.

Calculation of TFBS repertoire conservation

The average Canberra distance for each CRE across the i = 1, 2, .., n motif cumulative scores (S) was calculated as follows:

$$\overline{d}_{C_{MH}} = \frac{1}{n}\sum_{i}^{n} \frac{|S_{M,i} - S_{H,i}|}{(S_{M,i} + S_{H,i})}$$

where M indicates the orthologous CRE in macaque and H in human. As a control, we shuffled CRE identifiers of the macaque profiles within the respective PD class and calculated the average random TFBS profile similarity between species (Supplemental Fig. S6E,F).

TFBS position overlap between human and macaque orthologous CREs

Orthologous human and macaque CRE sequences were aligned with MAFFT (Katoh and Standley, 2013). Using the alignment of a CRE, the positions of TFBS with motif binding score of >=3 in either species were projected onto the common alignment space. Binding site agreement per motif i was calculated as the intersection of binding positions in bp between species over the union (Jaccard index) and summarized by taking the mean across all i = 1, 2, .., n motifs that bind to the CRE:

$$\overline{IoU}_{MH} = \frac{1}{n}\sum_{i}^{n} \frac{B_{M,i} \cap B_{H,i}}{B_{M,i} \cap B_{H,i}}$$

where B is a set of positions in the alignment that overlap with a binding site of motif i in the respective species macaque M or human H.

Quantification and Statistical Analyses

Data visualizations and statistical analysis was performed using R (version 4.2.3) (R Core Team, 2023), session info is on GitHub. Details of the statistical tests performed in this study can be found in the main text and Supplemental Materials. Schematics were made using bioRender.com

Data Access

RNA-seq and ATAC-seq data generated in this study have been submitted to ArrayExpress (https://www.ebi.ac.uk/biostudies/arrayexpress) under accession numbers E-MTAB-13494 and E-MTAB-13373.

A compendium containing processing scripts, tables and detailed instructions to reproduce the analysis is available as Supplemental Code and on GitHub: https://github.com/Hellmann-Lab/Evidence-for-compensatory-evolution-within-pleiotropic-regulatory-elements

Data files and tables are available as Supplemental Material and on zenodo: https://doi.org/10.5281/zenodo.11244831

Competing Interest

The authors declare no competing interests.

Acknowledgements

We thank Lucas Esteban Wange, Johannes Bagnoli, Aleks Janjic for helping with bulk RNA-seq preparation. We also thank Paulina Spurk for contributing to the initial ATAC-seq analysis, Diego Emiliano Ruiz Navarro for contributing to the initial RNA-seq analysis, Swati Parekh for assistance with Perl scripts and Eva Briem for helping with a schematic. We also thank Boyan Bonev for generating the liftOver files from GRCh38 to macFas6 and Andrea Betancourt for helpful discussions. We furthermore thank the Reviewers whose insightful comments helped to improve the

manuscript. This work was supported by DFG project HE 7669/2-1 (project number 458247426) and DFG project HE 7669/1-2 (project number 407541155).

Author Contributions

I.H. proposed the project and conceived the approaches of this study. W.E. provided the resources for data generation and helpful discussions. P.O. processed the human tissue accessibility data. B.V. provided expertise during initial steps. V.Y.K.L. contributed to TFBS evolutionary analyses. J.G. and M.H generated the primate cell lines and the expression data. S.K. and J.G. generated the primate accessibility data. I.H. supervised the work and provided guidance in data analysis. Z.K. collected, integrated and analyzed all data. Z.K. and I.H. wrote the manuscript. All authors read, corrected and approved the final manuscript.

References

Alexa, A and J Rahnenfuhrer (n.d.). *Gene set enrichment analysis with topGO*. https://bioconductor.statistik.tu-dortmund.de/packages/3.3/bioc/vignettes/topGO/inst/doc/topGO.pdf. Accessed: 2023-9-19.

Amemiya, HM, A Kundaje, and AP Boyle (June 2019). "The ENCODE Blacklist: Identification of Problematic Regions of the Genome". en. In: *Sci. Rep.* 9.1, p. 9354.

Andersson, R, C Gebhard, et al. (Mar. 2014). "An atlas of active enhancers across human cell types and tissues". en. In: *Nature* 507.7493, pp. 455–461.

Andersson, R and A Sandelin (Feb. 2020). "Determinants of enhancer and promoter activities of regulatory elements". en. In: *Nat. Rev. Genet.* 21.2, pp. 71–87.

Arnold, CD, D Gerlach, D Spies, JA Matts, YA Sytnikova, M Pagani, NC Lau, and A Stark (July 2014). "Quantitative genome-wide enhancer activity maps for five *Drosophila* species show functional enhancer conservation and turnover during *cis*-regulatory evolution". en. In: *Nat.*

Genet. 46.7, pp. 685–692.

Arnosti, DN and MM Kulkarni (Apr. 2005). "Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards?" en. In: *J. Cell. Biochem.* 94.5, pp. 890–898.

Ballester, B et al. (Oct. 2014). "Multi-species, multi-transcription factor binding highlights conserved control of tissue-specific biological pathways". en. In: *Elife* 3, e02626.

Bernstein, BE, JA Stamatoyannopoulos, JF Costello, B Ren, A Milosavljevic, A Meissner, M Kellis, MA Marra, AL Beaudet, JR Ecker, et al. (2010). "The NIH roadmap epigenomics mapping consortium". In: *Nat. Biotechnol.* 28.10, pp. 1045–1048.

Berthelot, C, D Villar, JE Horvath, DT Odom, and P Flicek (Jan. 2018). "Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression". en. In: *Nat Ecol Evol* 2.1, pp. 152–163.

Bininda-Emonds, ORP, M Cardillo, KE Jones, RDE MacPhee, RMD Beck, R Grenyer, SA Price, RA Vos, JL Gittleman, and A Purvis (Mar. 2007). "The delayed rise of present-day mammals". en. In: *Nature* 446.7135, pp. 507–512.

Bradley, RK, XY Li, C Trapnell, S Davidson, L Pachter, HC Chu, LA Tonkin, MD Biggin, and MB Eisen (Mar. 2010). "Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related *Drosophila* species". en. In: *PLoS Biol.* 8.3, e1000343.

Brawand, D et al. (Oct. 2011). "The evolution of gene expression levels in mammalian organs". en. In: *Nature* 478.7369, pp. 343–348.

Buenrostro, JD, PG Giresi, LC Zaba, HY Chang, and WJ Greenleaf (Dec. 2013). "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position". en. In: *Nat. Methods* 10.12, pp. 1213–1218.

Bushnell, B (Mar. 2014). *BBMap: A fast, accurate, splice-aware aligner*. en. Tech. rep. LBNL-7065E. Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States).

Carninci, P et al. (June 2006). "Genome-wide analysis of mammalian promoter architecture and

evolution". en. In: *Nat. Genet.* 38.6, pp. 626–635.

Chambers, SM, CA Fasano, EP Papapetrou, M Tomishima, M Sadelain, and L Studer (2009). "Highly efficient neural conversion of human ES and iPS cells by dual inhibition of SMAD signaling". In: *Nature Biotechnology 2009 27:3* 27, pp. 275–280.

Corces, MR et al. (Oct. 2017). "An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues". en. In: *Nat. Methods* 14.10, pp. 959–962.

Crocker, J, EPB Noon, and DL Stern (Jan. 2016). "The Soft Touch: Low-Affinity Transcription Factor Binding Sites in Development and Evolution". en. In: *Curr. Top. Dev. Biol.* 117, pp. 455–469.

Danko, CG et al. (Mar. 2018). "Dynamic evolution of regulatory element ensembles in primate CD4+ T cells". en. In: *Nat Ecol Evol* 2.3, pp. 537–548.

Domené, S, VF Bumaschny, FSJ de Souza, LF Franchini, S Nasif, MJ Low, and M Rubinstein (Dec. 2013). "Enhancer turnover and conserved regulatory function in vertebrate evolution". en. In: *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 368.1632, p. 20130027.

Doniger, SW and JC Fay (May 2007). "Frequent gain and loss of functional transcription factor binding sites". en. In: *PLoS Comput. Biol.* 3.5, e99.

Duret, L and D Mouchiroud (Jan. 2000). "Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate". en. In: *Mol. Biol. Evol.* 17.1, pp. 68–74.

Emera, D, J Yin, SK Reilly, J Gockley, and JP Noonan (May 2016). "Origin and evolution of developmental enhancers in the mammalian neocortex". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 113.19, E2617–26.

Evers, MM, LJA Toonen, and WMC van Roon-Mom (2014). "Ataxin-3 protein and RNA toxicity in spinocerebellar ataxia type 3: current insights and emerging therapeutic strategies". In: *Mol. Neurobiol.*

Feng, Q et al. (July 2018). "ATXN3 Positively Regulates Type I IFN Antiviral Response by Deubiquitinating and Stabilizing HDAC3". en. In: *J. Immunol.* 201.2, pp. 675–687.

Fishilevich, S et al. (Jan. 2017). "GeneHancer: genome-wide integration of enhancers and target genes in GeneCards". en. In: *Database* 2017.

Floc'hlay, S, E Wong, B Zhao, RR Viales, M Thomas-Chollier, D Thieffry, DA Garfield, and EEM Furlong (Dec. 2020). "Cis-acting variation is common across regulatory layers but is often buffered during embryonic development". en. In: *Genome Res.* 31.2, pp. 211–224.

Fornes, O et al. (Jan. 2020). "JASPAR 2020: update of the open-access database of transcription factor binding profiles". en. In: *Nucleic Acids Res.* 48.D1, pp. D87–D92.

Frith, MC, MC Li, and Z Weng (2003). "Cluster-Buster: Finding dense clusters of motifs in DNA sequences". In: *Nucleic Acids Res.* 31.13, pp. 3666–3668.

Gao, R et al. (Jan. 2015). "Inactivation of PNKP by mutant ATXN3 triggers apoptosis by activating the DNA damage-response pathway in SCA3". en. In: *PLoS Genet.* 11.1, e1004834.

Gasperini, M, JM Tome, and J Shendure (May 2020). "Towards a comprehensive catalogue of validated and target-linked human enhancers". en. In: *Nat. Rev. Genet.* 21.5, pp. 292–310.

Geuder, J, LE Wange, A Janjic, J Radmer, P Janssen, JW Bagnoli, S Müller, A Kaul, M Ohnuki, and W Enard (2021). "A non-invasive method to generate induced pluripotent stem cells from primate urine". In: *Scientific Reports 2021 11:1* 11, pp. 1–13.

Gronau, I, L Arbiza, J Mohammed, and A Siepel (May 2013). "Inference of natural selection from interspersed genomic elements based on polymorphism and divergence". In: *Mol. Biol. Evol.* 30.5, pp. 1159–1171.

Grunert, M, C Dorn, and S Rickert-Sperling (2016). "Cardiac Transcription Factors and Regulatory Networks". In: *Congenital Heart Diseases: The Broken Heart: Clinical Features, Human Genetics and Molecular Pathways*. Ed. by S Rickert-Sperling, RG Kelly, and DJ Driscoll. Vienna: Springer Vienna, pp. 139–152.

Hagai, T et al. (Nov. 2018). "Gene expression variability across cells and species shapes innate

immunity". en. In: *Nature* 563.7730, pp. 197–202.

Harrow, J et al. (Sept. 2012). "GENCODE: the reference human genome annotation for The ENCODE Project". en. In: *Genome Res.* 22.9, pp. 1760–1774.

Hastings, KE (June 1996). "Strong evolutionary conservation of broadly expressed protein isoforms in the troponin I gene family and other vertebrate gene families". en. In: *J. Mol. Evol.* 42.6, pp. 631–640.

He, A, SW Kong, Q Ma, and WT Pu (Apr. 2011). "Co-occupancy by multiple cardiac transcription factors identifies transcriptional enhancers active in heart". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 108.14, pp. 5632–5637.

Hill, MO (Mar. 1973). "Diversity and Evenness: A Unifying Notation and Its Consequences". In: *Ecology* 54.2, pp. 427–432.

Hong, JW, DA Hendrix, and MS Levine (Sept. 2008). "Shadow enhancers as a source of evolutionary novelty". en. In: *Science* 321.5894, p. 1314.

Horton, CA et al. (Sept. 2023). "Short tandem repeats bind transcription factors to tune eukaryotic gene expression". en. In: *Science*, p. 2022.05.24.493321.

Huang, YF, B Gulko, and A Siepel (Apr. 2017). "Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data". en. In: *Nat. Genet.* 49.4, pp. 618–624.

Ibrahim, MM, SA Lacadie, and U Ohler (2015). "JAMM: a peak finder for joint analysis of NGS replicates". In: *Bioinformatics* 31.1, pp. 48–55.

Jakovcevski, I, R Filipovic, Z Mo, S Rakic, and N Zecevic (June 2009). "Oligodendrocyte development and the onset of myelination in the human fetal brain". en. In: *Front. Neuroanat.* 3, p. 5.

Janjic, A et al. (Mar. 2022). "Prime-seq, efficient and powerful bulk RNA sequencing". en. In: *Genome Biol.* 23.1, p. 88.

Jayakumar, V et al. (June 2021). "Chromosomal-scale de novo genome assemblies of

Cynomolgus Macaque and Common Marmoset". en. In: *Sci Data* 8.1, p. 159.

Jost, L (May 2006). *Entropy and diversity*.

Junion, G, M Spivakov, C Girardot, M Braun, EH Gustafson, E Birney, and EEM Furlong (Feb. 2012). "A transcription factor collective defines cardiac cell fate and reflects lineage history". en. In: *Cell* 148.3, pp. 473–486.

Katoh, K and DM Standley (Apr. 2013). "MAFFT multiple sequence alignment software version 7: improvements in performance and usability". en. In: *Mol. Biol. Evol.* 30.4, pp. 772–780.

Kent, WJ, R Baertsch, A Hinrichs, W Miller, and D Haussler (Sept. 2003). "Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 100.20, pp. 11484–11489.

Khaitovich, P, I Hellmann, W Enard, K Nowick, M Leinweber, H Franz, G Weiss, M Lachmann, and S Pääbo (Sept. 2005). "Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees". en. In: *Science* 309.5742, pp. 1850–1854.

Kulkarni, MM and DN Arnosti (Dec. 2003). "Information display by transcriptional enhancers". en. In: *Development* 130.26, pp. 6569–6575.

Kuma, K, N Iwabe, and T Miyata (Jan. 1995). "Functional constraints against variations on molecules from the tissue level: slowly evolving brain-specific genes demonstrated by protein kinase and immunoglobulin supergene families". en. In: *Mol. Biol. Evol.* 12.1, pp. 123–130.

Li, H and R Durbin (May 2009). "Fast and accurate short read alignment with Burrows–Wheeler transform". en. In: *Bioinformatics* 25.14, pp. 1754–1760.

Love, MI, W Huber, and S Anders (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". en. In: *Genome Biol.* 15.12, p. 550.

Ludwig, MZ, C Bergman, NH Patel, and M Kreitman (Feb. 2000). "Evidence for stabilizing selection in a eukaryotic enhancer element". en. In: *Nature* 403.6769, pp. 564–567.

McLean, CY, D Bristor, M Hiller, SL Clarke, BT Schaar, CB Lowe, AM Wenger, and G Bejerano (May 2010). "GREAT improves functional interpretation of *cis*-regulatory regions". en. In: *Nat. Biotechnol.* 28.5, pp. 495–501.

Messmer, K, WB Shen, M Remington, and PS Fishman (Apr. 2012). "Induction of neural differentiation by the transcription factor neuroD2". en. In: *Int. J. Dev. Neurosci.* 30.2, pp. 105–112.

Morgan, MD and JC Marioni (June 2018). "CpG island composition differences are a source of gene expression noise indicative of promoter responsiveness". en. In: *Genome Biol.* 19.1, p. 81.

Nakagawa, S, PCD Johnson, and H Schielzeth (Sept. 2017). "The coefficient of determination R 2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded". en. In: *J. R. Soc. Interface* 14.134, p. 20170213.

Ohnuki, M et al. (Aug. 2014). "Dynamic regulation of human endogenous retroviruses mediates factor-induced reprogramming and differentiation potential". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 111.34, pp. 12426–12431.

Olson, JM, A Asakura, L Snider, R Hawkes, A Strand, J Stoeck, A Hallahan, J Pritchard, and SJ Tapscott (June 2001). "NeuroD2 is necessary for development and survival of central nervous system neurons". en. In: *Dev. Biol.* 234.1, pp. 174–187.

Osterwalder, M et al. (Feb. 2018). "Enhancer redundancy provides phenotypic robustness in mammalian development". en. In: *Nature* 554.7691, pp. 239–243.

Parekh, S, C Ziegenhain, B Vieth, W Enard, and I Hellmann (2018). "zUMIs - A fast and flexible pipeline to process RNA sequencing data with UMIs". In: *Gigascience* 7.

Paris, M, T Kaplan, XY Li, JE Villalta, SE Lott, and MB Eisen (Sept. 2013). "Extensive divergence of transcription factor binding in *Drosophila* embryos with highly conserved gene expression". en. In: *PLoS Genet.* 9.9, e1003748.

Pataskar, A, J Jung, P Smialowski, F Noack, F Calegari, T Straub, and VK Tiwari (Jan. 2016).

"NeuroD1 reprograms chromatin and transcription factor landscapes to induce the neuronal program". en. In: *EMBO J.* 35.1, pp. 24–45.

Pollard, KS, MJ Hubisz, KR Rosenbloom, and A Siepel (Jan. 2010). "Detection of nonneutral substitution rates on mammalian phylogenies". en. In: *Genome Res.* 20.1, pp. 110–121.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. Vienna, Austria.

Roller, M, E Stamper, D Villar, O Izuogu, F Martin, AM Redmond, R Ramachanderan, L Harewood, DT Odom, and P Flicek (Feb. 2021). "LINE retrotransposons characterize mammalian tissue-specific and evolutionarily dynamic regulatory regions". en. In: *Genome Biol.* 22.1, p. 62.

Schlesinger, J, M Schueler, M Grunert, JJ Fischer, Q Zhang, T Krueger, M Lange, M Tönjes, I Dunkel, and SR Sperling (Feb. 2011). "The cardiac transcription network modulated by Gata4, Mef2a, Nkx2.5, Srf, histone modifications, and microRNAs". en. In: *PLoS Genet.* 7.2, e1001313.

Schmidt, D et al. (May 2010). "Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding". en. In: *Science* 328.5981, pp. 1036–1040.

Schmitt, I, M Linden, H Khazneh, BO Evert, P Breuer, T Klockgether, and U Wuellner (Oct. 2007). "Inactivation of the mouse Atxn3 (ataxin-3) gene increases protein ubiquitination". en. In: *Biochem. Biophys. Res. Commun.* 362.3, pp. 734–739.

Schor, IE et al. (Apr. 2017). "Promoter shape varies across populations and affects promoter evolution and expression noise". en. In: *Nat. Genet.* 49.4, pp. 550–558.

Schwartz, S, WJ Kent, A Smit, Z Zhang, R Baertsch, RC Hardison, D Haussler, and W Miller (Jan. 2003). "Human-mouse alignments with BLASTZ". en. In: *Genome Res.* 13.1, pp. 103–107.

Sedlazeck, FJ, P Rescheneder, and A von Haeseler (Nov. 2013). "NextGenMap: fast and accurate read mapping in highly polymorphic genomes". en. In: *Bioinformatics* 29.21, pp. 2790–2791.

Shannon, CE (July 1948). “A Mathematical Theory of Communication”. In: *Bell System Technical Journal* 27.3, pp. 379–423.

Shumate, A and SL Salzberg (July 2021). “Liftoff: accurate mapping of gene annotations”. en. In: *Bioinformatics* 37.12, pp. 1639–1643.

Sigalova, OM, A Shaeiri, M Forneris, EE Furlong, and JB Zaugg (Aug. 2020). “Predictive features of gene expression variation reveal mechanistic link with differential expression”. en. In: *Mol. Syst. Biol.* 16.8, e9539.

Singh, D and SV Yi (Mar. 2021). “Enhancer Pleiotropy, Gene Expression, and the Architecture of Human Enhancer–Gene Interactions”. en. In: *Mol. Biol. Evol.* 38.9, pp. 3898–3909.

Stefflova, K et al. (Aug. 2013). “Cooperativity and rapid evolution of cobound transcription factors in closely related mammals”. en. In: *Cell* 154.3, pp. 530–540.

Sun, Y, M Nadal-Vicens, S Misono, MZ Lin, A Zubiaga, X Hua, G Fan, and ME Greenberg (Feb. 2001). “Neurogenin promotes neurogenesis and inhibits glial differentiation by independent mechanisms”. en. In: *Cell* 104.3, pp. 365–376.

Tan, G and B Lenhard (May 2016). “TFBSTools: an R/bioconductor package for transcription factor binding site analysis”. en. In: *Bioinformatics* 32.10, pp. 1555–1556.

Tuğrul, M, T Paixão, NH Barton, and G Tkačik (Nov. 2015). “Dynamics of Transcription Factor Binding Site Evolution”. en. In: *PLoS Genet.* 11.11, e1005639.

Vasimuddin, M, S Misra, H Li, and S Aluru (May 2019). “Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems”. In: *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, pp. 314–324.

Villar, D, P Flicek, and DT Odom (Apr. 2014). “Evolution of transcription factor binding in metazoans — mechanisms and functional implications”. en. In: *Nat. Rev. Genet.* 15.4, pp. 221–233.

Wang, HY, HC Chien, N Osada, K Hashimoto, S Sugano, T Gojobori, CK Chou, SF Tsai, CI Wu, and CKJ Shen (Feb. 2007). “Rate of evolution in brain-expressed genes in humans and other

primates". en. In: *PLoS Biol.* 5.2, e13.

Wang, ZY et al. (Dec. 2020). "Transcriptome and translatome co-evolution in mammals". en. In: *Nature* 588.7839, pp. 642–647.

Wunderlich, Z, MDJ Bragdon, BJ Vincent, JA White, J Estrada, and AH DePace (Mar. 2016). "Krüppel Expression Levels Are Maintained through Compensatory Evolution of Shadow Enhancers". en. In: *Cell Rep.* 14.12, p. 3030.

Yevshin, I, R Sharipov, S Kolmykov, Y Kondrakhin, and F Kolpakov (Nov. 2018). "GTRD: a database on gene transcription regulation—2019 update". en. In: *Nucleic Acids Res.* 47.D1, pp. D100–D105.

Yu, Y et al. (Jan. 2013). "Olig2 targets chromatin remodelers to enhancers to initiate oligodendrocyte differentiation". en. In: *Cell* 152.1-2, pp. 248–261.

Zhang, L and WH Li (Feb. 2004). "Mammalian housekeeping genes evolve more slowly than tissue-specific genes". en. In: *Mol. Biol. Evol.* 21.2, pp. 236–239.

Zhao, Y et al. (Sept. 2022). ""Stripe" transcription factors provide accessibility to co-binding partners in mammalian genomes". en. In: *Mol. Cell* 82.18, 3398–3411.e11.

Zhou, Q and DJ Anderson (Apr. 2002). "The bHLH transcription factors OLIG2 and OLIG1 couple neuronal and glial subtype specification". en. In: *Cell* 109.1, pp. 61–73.

Figure 1. Study overview. (*A*) Open chromatin and expression data from the Roadmap Epigenomics Project (Bernstein et al., 2010) were used to infer the effect of pleiotropy on sequence and TFBS evolution, and associated gene expression in primates. Overlapping DHS peaks between tissues were merged to determine the degree of tissue-specificity per CRE. (*B*) DHS-data from 9 human fetal tissues. The number of biological replicates per tissue varies between 7 and 34. (*C*) The number of CREs per tissue varies 2.3-fold. There is no association between the number of replicates and the number of accessible regions per tissue. (*D*) Most enhancers (dotted line) are tissue-specific, while promoters (solid line) are mostly pleiotropic. (*E*) CRE length increases with the number of tissues, particularly at the promoters. This increase was also observed at the peak level prior to merging (Supplemental Fig. S1A). (*D, E*) The colors represent the tissues as introduced in (*A, B*). (*F*) The majority of PD9 CREs are CpG island promoters (solid blue), while tissue-specific elements are rarely CpG islands and mainly enhancers (transparent green). (*G*) Pleiotropic promoters are more commonly associated with pleiotropic gene expression patterns. The promoter PD indicates the highest PD of the associated promoters per gene. The y-axis shows the proportions of those x-categories (promoter PD) with associated gene expression pleiotropy ranging from 1 to 9. EPD = expression pleiotropic degree. (*H*) Scaled coefficients of a linear mixed model to predict gene expression levels using distance scaled CRE counts of different types.

Figure 2. TFBS repertoire diversity and enrichment across tissue-specific and pleiotropic CREs. (*A*) An overview of the pleiotropic degree of TF expression (EPD) across tissues. (*B*) TFBS repertoire diversity increases with PD, particularly across promoters. Depicted are mean values +/- SEM. (*C*) Overview of the over-represented motifs in PD9 and PD1 CRE sequences. Depicted are binding site preferences for 643 motifs, the respective TFs of which are detected in the expression data. If a motif showed the highest binding proportion to PD9 CREs consistently within each tissue-CREs, it was assigned as over-represented in PD9. If a motif showed the highest binding proportion to a PD1 of a particular tissue, but not that of others, it was assigned as PD1 enriched for that tissue. (*D*) Top 5 categories of gene set enrichment analysis of PD9-enriched motifs using all

motifs as background (Gene ontology, Biological Process, Fisher's exact p-value<0.05). (*E, F*) Top 4 categories of gene set enrichment analysis of tissue-specific PD1 enriched motifs using all motifs as background (Gene ontology, Biological Process, Fisher's exact p-value<0.05). (*E*) Brain-specific PD1 over-represented motifs. (*F*) Heart-specific PD1 over-represented motifs. (*E, F*) Odds ratio is based on the proportion of tissue-specific PD1 CREs with the motif over the global average proportion for that motif.

Figure 3. Pleiotropic degree and evolutionary conservation of expression and regulatory activity. (*A, B*) The fraction of enhancers and promoters of different pleiotropic degrees (PD) as defined using data from 9 tissues from the Epigenomics Roadmap project, which overlapped with ATAC-seq peaks called in neural progenitor cell lines (NPCs) from cynomolgus macaques and humans. The colors indicate whether a human DHS-derived CRE overlapped with an NPC ATAC-seq peak from humans, cynomolgus macaques, both or none. (*C*) The phylogeny from Bininda-Emonds et al. (2007) of species for which Roller et al. (2021) provide activity estimates for 4 tissues based on histone marks. Scale: Million years ago. (*D*) Activity conservation for orthologous CREs based on Roller et al. (2021) (n=47,377). We calculate the ratio subtree length λ_S for species in which the CRE is active over the total tree length λ_T . (*E*) Mean absolute log_2-fold changes of gene expression and activities between humans and cynomolgus macaques. Error bars: 95% bootstrap CIs. PD9 genes (CREs) show more conserved expression (activity) than more tissue-specific genes (CREs). (*F*) Enrichment (odds ratio>1) of differentially accessible CREs with differentially expressed genes between humans and cynomolgus macaques using matched regulatory landscapes (n=6,882). Error bars: 95% CIs of the odds ratio. Asterisks: BH-adjusted Fisher's test p-value ($^{*} < 0.05$, $^{**} < 0.01$, $^{***} < 0.001$).

Figure 4. CRE sequence conservation patterns across varying degrees of pleiotropy. (*A, B*) Number of weakly deleterious sites in humans (E.W.) inferred from human polymorphisms. It increases with increasing PD. (*A*) Separated by enhancers / promoters. (*B*) Separated by human-macaque

accessibility conservation in NPCs. (*C, D*) Fraction of sites under (strong) negative selection on the lineage since the MRCA of humans and chimpanzees is the highest at the intermediately-specific CREs and lowest in the pleiotropic CRE sequences. (*C*) Separated by enhancers / promoters. (*D*) Separated by human-macaque accessibility conservation in NPCs. (*A, B, C, D*) Depicted are mean values +/- SEM.

Figure 5. CRE sequence conservation patterns per tissue. (*A*) Weak negative selection inferred from human polymorphisms, separated by the tissue that utilizes the CREs. (*B*) Fraction of sites under (strong) negative selection, separated by the tissue that utilizes the CREs. (*C*) Brain CRE sequences were separated into peak and adjacent sequences. (*D*) The part of the sequence that is used by the brain shows a considerably higher fraction of sites under negative selection than the respective adjacent sequences. (*A, B, D*) Depicted are mean values +/- SEM.

Figure 6. TFBS repertoire and position conservation between orthologous human and macaque CREs. (*A, B, C*) TFBS repertoire conservation across PDs. Depicted are mean values +/- SEM. (*A*) TFBS repertoire conservation increases with higher PD among promoters and decreases slightly at high PD enhancers. (*B*) CREs that overlap NPC peaks with conserved openness show higher TFBS repertoire conservation than species-specific NPC peaks. (*C*) TFBS repertoire conservation differs across tissues, where the brain shows the highest conservation at lower PDs. (*D*) Simplified schematic of the scenarios of repertoire and position conservation. (*E*) TFBS position conservation versus repertoire conservation across PD categories. Depicted are mean values +/- SEM. (*F*) Standardized scores (z-scores) of sequence (primate phyloP), TFBS repertoire and binding site conservation between human and cynomolgus macaque. (*G*) A schematic depicting how lower sequence conservation might lead to higher TFBS repertoire conservation through compensatory mechanisms. (*H*) A summary of the scaled average conservation metric scores across PDs. Sequence: primate phyloP scores, TFBS position: $\overline{IoU}_{MH}$, TFBS repertoire: $1 - \bar{d}_{C_{MH}}$, CpG

observed/expected: 1-$|CpG_{\frac{o}{e},H} - CpG_{\frac{o}{e},M}|$, accessibility: 1-|LFC| of NPC-DA, downstream expression: 1-|LFC| of NPC-DE.

Figure 7. Ranks of *ATXN3* PD9 promoter compared to other CREs in terms of (*A*) sequence conservation (mean phastCons), (*B*) TFBS binding site conservation, (*C*) TFBS repertoire conservation, (*D*) CRE openness conservation between human and cynomolgus macaque in NPCs and (*E*) *ATXN3* gene expression conservation between human and cynomolgus macaque in NPCs. (*F*) *ATXN3* PD9 promoter ATAC-seq read coverage in human NPCs. (*G*) *ATXN3* PD9 promoter ATAC-seq read coverage in macaque NPCs. (*F, G*) *ATXN3* PD9 promoter is accessible in NPCs from both species. Gray background area indicates the positions of the called ATAC-seq peak in the respective species. (*H*) *ATXN3* promoter shows diverged TFBS positions between species among two TFs (MYCN, POU3F2) with validated binding based on ChIP-seq neural cell data that are involved in neurogenesis. Motif binding sites and scores are depicted in the aligned sequence space, human on the top, cynomolgus macaque on the bottom. (*I, J*) PWM logos from JASPAR2020 vertebrate core set of the investigated TF motifs with ChIP-seq data available: MYCN (*I*), POU3F2 (*J*).