Characterizing the cis-regulatory landscape of the mammalian allantois

Julie Chen
Department of Computer Science
Stanford University
jchen80@stanford.edu

Abstract

The umbilical cord develops from the allantois and performs critical functions to support embryo survival in placental mammals. However, to date, relatively little is known about the gene regulation underlying the development of the allantois. This study identifies a candidate set of 977 allantois cis-regulatory elements, drawn from accessible peaks present in mouse allantois cells that are associated with genes relevant for allantois function. Moreover, we highlight candidate CREs near the *Lipt2* and *Supt20* genes as lone regulatory peaks for future validation in experimental deletion studies. Evolutionary conservation of these putative allantois cis-regulatory elements is notably higher up through the clade comprising eutherian (placental) mammals compared to other tissues such as heart and brain. Sequence conservation is relatively lower in clades deeper than non-placental mammals. This finding suggests that the functional role of these candidate allantois cis-regulatory elements may have been under greater selective pressure beginning in eutheria. Enrichment analysis for transcription factor motifs revealed that putative allantois CREs are enriched for GATA, FOX and posterior HOX transcription factors. Candidate heart CREs are similarly enriched for GATA family motifs but additionally for anterior HOX motifs. More detailed analysis of enrichment in HOX family motifs found a clear distinction in anterior vs posterior HOX motif enrichment in the heart and allantois respectively. Although motif enrichment is not direct evidence for active binding events, the differential enrichment for motifs suggests that different HOX factor binding at heart vs allantois candidate CREs may drive gene expression differences in heart vs allantois along the anterior-posterior axis.

1 Introduction

In mammals, the umbilical cord connects the developing embryo to the placenta and is vital for life Inman and Downs (2007). The umbilical cord arises from the allantois and suffuses the embryo with oxygen and nutrients. Evolutionarily speaking, the specialization of the allantois to become umbilical cord is relatively unique to placental mammals. However, the emergence of an allantois-like structure is observed first in amniotes, where it seems to perform similar functions of gas exchange and waste storage for the developing embryo Ferner and Mess (2011). Yet despite its critical importance in supporting embryonic development, relatively little is known about allantois development and evolution compared to that of other organs.

Past research studying how new and diverse tissues evolve has emphasized the role of cis-regulatory elements (CREs) Wray (2007). Briefly, cis-regulatory elements are regions of the genome which unlike genes, do not directly encode proteins, but instead regulate the expression level of nearby genes, i.e. how much mRNA transcript is synthesized Wittkopp and Kalay (2012). Compared to regions of the genome encoding genes, CREs exhibit greater variation between species and increasing evidence seems to support a model whereby changes in CREs are the primary drivers in evolving new tissues Wittkopp and Kalay (2012).

Although the allantois is an amniote innovation, it bears transcriptional and morphological similarities to a subtype of mesoderm known as lateral plate mesoderm conserved throughout vertebrates, which gives rise to heart, blood, blood vessels and limbs Pijuan-Sala et al. (2019). Our unpublished work differentiating human pluripotent stem cells into heart and allantois cell-types *in vitro* has additionally found that both cell fates require BMP signaling to successfully differentiate. Moreover, we found that different sub-types of primitive streak (the stage preceding lateral mesoderm) are competent to give rise to either heart or allantois, but not both suggesting that diversification of the heart and allantois lineages occurs early. However, despite this apparent early lineage divergence, there are important developmental genes shared in common between the heart and allantois such as HAND, ISL, and GATA transcription factors (TFs). How then do "sister" lineages such as the heart and allantois regulate these shared key genes during development?

Thus, the overall aim of this study is to characterize putative CREs important for allantois development. To do so, we first identify CREs relevant for the functional development of the mammalian allantois and select candidate CREs for future deletion studies. Second, we assess the evolutionary sequence conservation of CREs relevant to the allantois vs other tissues. Lastly, we compare the CRE landscape of the developing allantois and heart lineages by performing transcription factor motif enrichment.

2 Related Work

Historically, putative CREs have been identified based on coincidence with certain chromatin modifications, such as H3K27ac and H3K4me1, as well as by binding of certain transcription factors such as p300 Calo and Wysocka (2013). However, recent work has shown that CREs can also be identified by the assay for accessible chromatin (ATAC) sequencing, which identifies regions of the genome that can be tagged by a Tn5 transposase Buenrostro et al. (2015). Although ATAC-seq is less specific in identifying CREs than histone marks or TF binding, existing work in Daugherty et al. (2017) has shown that CREs predicted based on ATAC-seq data of the developing *C. elegans* correspond well with enhancer chromatin predicted from Hidden Markov models trained on histone marks and exhibit some degree of sequence conservation. Moreover, several of these putative CREs were experimentally validated to drive tissue-specific expression. we will apply a similar approach and use ATAC-seq chromatin accessibility data from whole-mouse embryos to identify putative CREs relevant for allantois function.

A major open question in gene regulation is how to test whether candidate CREs exert experimentally measurable effects on a biological phenotype. One method to narrow down on the most promising candidates for validation by experimental genetic deletions raised in Wenger et al. (2013) is to select CREs at vulnerable loci where a gene may have only one or a few regulatory elements identified. As such, we will apply this criterion to select "lone" elements near allantois genes for future experimental validation.

Lastly, sequence conservation of CREs suggests functional significance, although the biological function of these CREs may not be conserved. Past studies have shown that some CREs appear to be relatively well-conserved across diverse species and can be identified using a comparative genomics approach as in Pennacchio et al. (2006). Thus, a relatively higher degree of sequence conservation of putative CREs for the allantois within a given clade might indicate stronger evolutionary pressure and thus greater functional importance.

3 Methods

In this study, we identify a set of putative cis-regulatory elements important for the development of the mouse allantois using single-nuclei ATAC-seq data from whole-mouse embryos. Then, we map accessible peaks to genes and select candidate elements for deletion studies if they fall within the regulatory domain of functional allantois genes. Next, we assess the relative evolutionary sequence conservation of CREs for allantois compared to other tissues such as the heart and brain, and to nonexonic basepairs. Lastly, we compare putative allantois and heart CREs by performing transcription factor motif enrichment of the CREs compared to background genomic sequences.

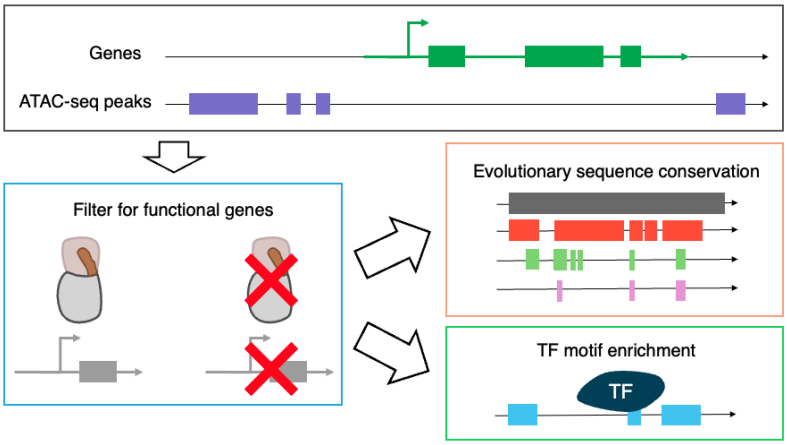


Figure 1: Overview of workflow to select and characterize candidate CREs for a given tissue of interest, e.g. allantois.

3.1 Data

3.1.1 Single-cell -omics datasets

This study will leverage a published multi-omic dataset comprising single-nucleus snRNA and ATAC-seq profiles of whole mouse embryos collected from embryonic day E7.5 to E8.5 Argelaguet et al. (2022). To collect this dataset, whole mouse embryos were staged, dissociated into single cells, and the combined snRNA/ATAC-seq assay made measurements of the RNA transcripts present in the nuclei of these cells (snRNA-seq) and the regions of transposase-accessible chromatin (snATAC-seq) for the same cell. The authors performed peak calling of the snATAC-seq data using ArchR Granja et al. (2021). Subsequently, the authors of this study clustered the transcriptomic profiles of the cells, performed cell-type annotation based on markers found in Pijuan-Sala et al. (2019), and then based on the transcriptomic cell-type assignment, compiled sets of chromatin-accessible regions for each cell-type. we downloaded the peak calls for allantois and heart cell types from the publicly available repository.

3.1.2 Genotype-phenotype annotations

Additionally, this study will use the publicly available Mouse Genome Informatics (MGI) database, which reports phenotypes associated with mouse models of gene deletions and variations Baldarelli et al. (2024). To collect all genotypes associated with functional defects in allantois development, we used the annotation terms "absent allantois" and "abnormal allantois morphology." To gather all genotypes associated with functional defects in heart development, we used the annotation parent term "abnormal heart morphology". Likewise, to collect all genotypes associated with functional defects in brain development, we used the annotation terms "abnormal brain morphology" and "abnormal nervous system development."

3.1.3 Pairwise cross-species alignments

Pairwise net alignments of the mouse mm10 reference genome to the reference genomes of lizard (anoCar2), zebrafish (danRer7), chicken (galGal4), elephant (loxAfr3), Medaka (oryLat2), Tasmanian devil (sarHar1), cow (bosTau6), armadillo (dasNov3), stickleback (gasAcu1), opposum (monDom5), chimpanzee (panTro6), tetraodon (tetNig2), dog (canFam3), frog (xenTro3), human (hg19), platypus (ornAna1), macaque (rheMac2) and fugu (fr3) was downloaded from the UCSC genome browser.

3.2 Experimental details

3.2.1 Identification of candidate CREs relevant for cell-type development

Peak calls from single nuclei ATAC-seq cells for a given cell-type, e.g. allantois, were mapped to the regulatory domain of a gene based on the default association rule used in GREAT McLean et al.

(2010) which consists of a basal domain extending 1kb downstream and 5kb upstream, with up to 1 Mb extension in either direction up to the neighboring genes' basal domains. Then, candidate CREs were selected if they were associated with genes with functional relevance for the cell-type of interest, i.e. knockout of the gene had reported functional defects in the tissue.

3.2.2 Evolutionary sequence conservation

Sequence conservation analysis was performed similarly to as in Wenger et al. (2013). Briefly, conservation maps were generated across the following clades: euarchontoglires (human hg19, chimp panTro6, rhesus rheMac2), eutheria (elephant loxAfr3, cow bosTau6, dog canFam3, armadillo dasNov3), mammalia (platypus ornAna1, opposum monDom5, Tasmanian devil sarHar1); amniota (chicken galGal3, lizard anoCar2), tetrapoda (frog xenTro3); gnathostomata (tetraodon tetNig2, fugu fr3, zebrafish danRer7, stickleback gasAcu1, medaka oryLat2). Conserved regions were selected from levels 1 and 2 of pairwise alignment nets to the mouse mm10 genome after removing genomic duplications, pseudogenes and known exons. A basepair was considered conserved across a clade if it was conserved in the preceding clade and if it overlapped with any of the representatives of the clade, except for gnathostomes where two matches are required.

Putative CREs were considered conserved if there are at least 100bps conserved. For comparison, non-exonic genomic regions were drawn from the mm10 genome by excluding genomic duplications, pseudogenes, known exons and gaps.

3.2.3 TF motif enrichment

841 TF motif matrices from the "vertebrates" taxonomic group were downloaded from the JASPAR 2022 database. Background genomic sequences were generated for each query set of candidate CRE peaks using homer2 to match the relative nucleotide dimer frequencies of the query sequences. Motif matching was performed using motifmatchr for both query and background sequences and a hypergeometric test was performed to test for relative enrichment which is reported as odds ratio. P-values were adjusted for multiple hypothesis testing using the false discovery rate (FDR) method.

4 Results & Discussion

4.1 Identifying a set of candidate CREs important for allantois development

To date, there are 144 reported genes which when mutated result in defects in allantois development. Filtering for peak calls in allantois cells associated with these functional genes produced a set of 977 putative CREs important for allantois development. Although most allantois functional genes appear to be regulated by a few (<5) candidate CREs, the most heavily regulated genes are associated with as many as 40 putative CREs (Figure 2). The genes with the greatest number of associated CREs include *Foxf1*, *Kdr*, *Trp53* and *Lef1* which encode key transcription factors and signaling pathway components (Table 1).

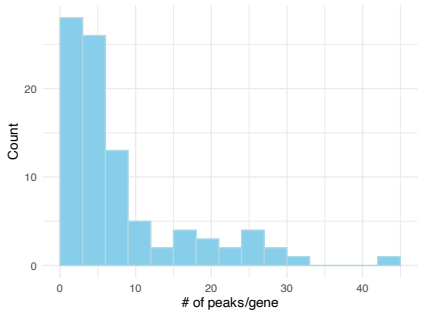


Figure 2: Histogram of peaks per gene for the allantois functional gene subset

Subsequently, we looked for genes where there are relatively few associated CREs, hypothesizing that these may represent vulnerable regulatory points that could be validated experimentally in the future. Two such genes are *Lipt2*, a mitochondrial metabolic enzyme, and *Supt20*, a histone acetyltransferase. Besides the peak near the TSS, likely corresponding to the promoter region, both of these functional

Table 1: Top 8 genes with the highest number of candidate CREs in their regulatory domains

Gene	# of candidate CREs
Foxf1	42
Kdr	38
Trp53	35
Lef1	33
Tead1	31
Plpp3	29
Hand1	27
Snai1	26

allantois genes only have one other candidate CRE in their regulatory domains, making those CREs likely to have an effect on gene expression level if deleted experimentally (Figure 3).

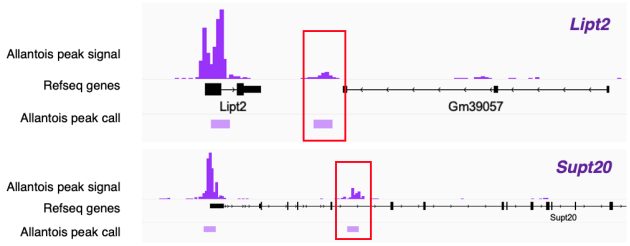


Figure 3: Genome browser tracks for *Lipt2* and *Supt20* genes showing allantois ATAC-seq peak signal, peak calls and gene annotations.

4.2 Evolutionary sequence conservation of allantois cis-regulatory elements

We considered a CRE to be conserved if at least 100bp overlapped with conserved basepairs for a given clade. Applying this criterion to candidate allantois CREs, we found that a large fraction of allantois CREs are conserved across euarchontoglires (77.9%) and eutherian mammals (72.8%), and to a lesser degree across all mammals (35.3%) (Figure 4). This degree of sequence conservation is much higher than that of random non-exonic basepairs reflecting some amount of selective pressure on these candidate allantois CREs, possibly because they perform a biological function in gene regulation.

We then wondered how well-conserved allantois CREs are compared to candidate CREs important for the development of other tissues. We selected the heart because it is a "sister" lineage to the allantois and belongs within the same larger grouping of lateral mesoderm, as well as the brain which is from a completely different germ layer (ectoderm). Candidate CREs for all three tissues are more highly conserved compared to random non-exonic basepairs. As previously observed in Wenger et al. (2013), brain CREs are generally the among the most well-conserved relative to other tissues, and in our study, we found this to also be the case for the mammalia, amniota, tetrapoda and gnasthostomata clades. However, in the shallowest clades, euarchontoglires and eutheria, candidate allantois CREs were the most highly conserved compared to brain and heart CREs. This observation may indicate that the candidate allantois CREs identified in this study were under strongest purifying selection in eutheria, so the associated biological function of these regulatory elements may be most relevant in placental mammals. In other words, although the allantois is present in amniotes, the candidate CREs identified in this study may be most functionally relevant towards the development of the allantois and umbilical cord in placental mammals. Additionally, this result hints at divergence in elements regulating allantois genes before and after the evolution of placental mammals.

4.3 Comparing the cis-regulatory landscape between allantois and heart lineages

Since our unpublished experimental work suggests that the allantois and heart lineages diverge early in development, one possibility is that allantois and heart functional genes are regulated by transcription factors in the preceding stage. To explore this hypothesis, we performed enrichment

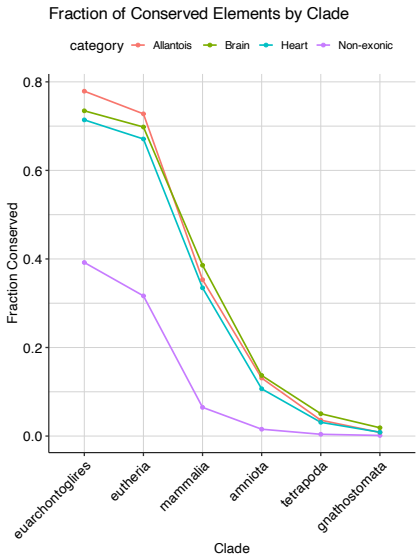
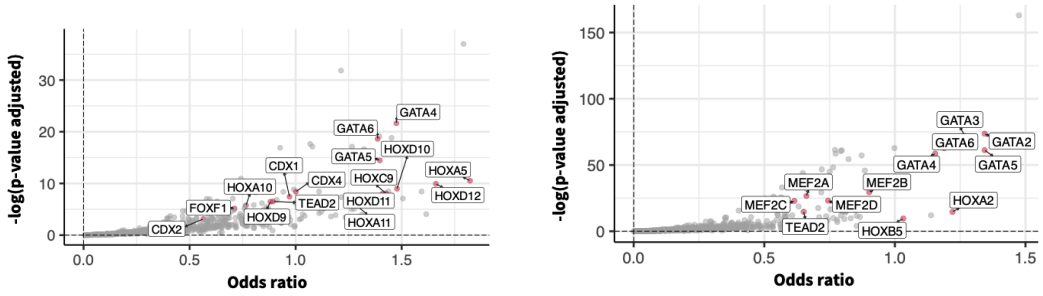


Figure 4: Fraction of conserved CREs by phylogenetic clade for allantois, heart and brain CREs vs genomic non-exonic basepairs.

testing for transcription factor motifs in the candidate allantois and heart CREs. It is worth mentioning that the presence of strong motif matches is not necessarily indicative of a binding event and that TF binding may be developmentally relevant even with sub-optimal motif matches as shown in Farley et al. (2015). However, in the absence of experimental data measuring actual binding events, relative enrichment of a given TF's motif nonetheless provides some indication that binding of that TF in a CRE is more likely.

TF motif enrichment for allantois CREs revealed that the most significantly enriched motifs belonged to the GATA family (GATA4, GATA5, GATA6) as well as posterior HOX genes (HOXD10, HOXA11, HOXD11, HOXD12) and CDX (CDX1, CDX4), which share similar consensus motifs. Candidate heart CREs were similarly enriched for GATA family motifs (GATA2, GATA3, GATA5, GATA6) but by contrast were also enriched for MEF2 family motifs and more anterior HOX genes (HOXA2, HOXB5) (Figure 5).



(a) TF motifs enriched in allantois CREs

(b) TF motifs enriched in heart CREs

Figure 5: Relative enrichment of TF motif sequences in allantois or heart CREs vs matched background genomic sequences, odds ratio calculated with hypergeometric test and p-value significance values adjusted with the FDR method.

HOX genes play a well-documented role in patterning along the anterior-posterior axis, so we delved further into looking at TF motif enrichment specifically for HOX family genes. As shown in Figure 6, candidate allantois CREs are enriched for anterior and posterior HOX genes but heart CREs are only enriched for anterior HOX genes. This result suggests that allantois vs heart CREs are activated by different HOX genes delimiting and concomitant with their position on the anterior-posterior axis.

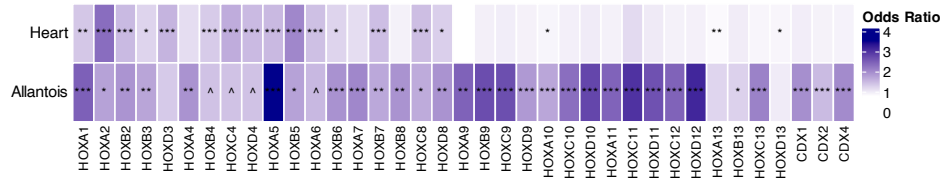


Figure 6: Heatmap displaying odds ratio and significance (FDR adjusted p-values) of enrichment for HOX and CDX transcription factor motifs within allantois and heart candidate CREs.

5 Conclusions

In this study, we identified a set of candidate allantois CREs from mouse embryo single-nuclei ATAC-seq data and reported mouse mutant allantois phenotype data. Using this set of putative allantois CREs, we revealed highly regulated genes for allantois development as assessed by the number of CREs associated with that gene. Additionally, we pinpointed two CREs near the *Lipt2* and *Supt20* gene loci which represent isolated regulatory elements for future deletion studies. Candidate allantois CREs displayed an interesting trend in sequence conservation, whereby they appeared unusually well-conserved up to eutherian mammals, suggesting greater functional importance for this set of CREs specifically within the context of placental mammal allantois development vs for amniotes more broadly. Lastly, TF motif enrichment revealed that allantois vs heart candidate CREs are enriched for some shared motifs including GATA but diverge especially with respect to posterior HOX genes.

Future studies might further explore the role of HOX factors in delineating allantois vs heart fate. First, one should measure experimentally whether candidate allantois vs heart CREs are in fact differentially bound by posterior vs anterior HOX genes. Second, one might conduct perturbative studies assessing whether posterior vs anterior HOX factor binding is necessary or sufficient in driving expression of allantois vs heart genes. Although computational enrichment for TF motifs is a powerful tool for generating hypotheses, these analyses are limited in their ability to provide insight as to whether observed enrichment is functionally significant or relevant.

6 Data and Code Availability

Processed data and code required to reproduce the results of this study are available upon request.

References

Ricard Argelaguet, Tim Lohoff, Jingyu Gavin Li, Asif Nakhuda, Deborah Drage, Felix Krueger, Lars Velten, Stephen J. Clark, and Wolf Reik. 2022. Decoding gene regulation in the mouse embryo using single-cell multi-omics. *bioRxiv*.

Ruth M. Baldarelli, Cynthia L. Smith, Martin Ringwald, Joel E. Richardson, Carol J. Bult, and Mouse Genome Informatics Group. 2024. Mouse genome informatics: an integrated knowledgebase system for the laboratory mouse. *Genetics*, 227(1):iyae031.

Jason D Buenrostro, Beijing Wu, Howard Y Chang, and William J Greenleaf. 2015. ATAC-seq: A method for assaying chromatin accessibility genome-wide. *Curr. Protoc. Mol. Biol.*, 109(1):21.29.1–21.29.9.

Eliezer Calo and Joanna Wysocka. 2013. Modification of enhancer chromatin: what, how, and why? *Mol. Cell*, 49(5):825–837.

Aaron C Daugherty, Robin W Yeo, Jason D Buenrostro, William J Greenleaf, Anshul Kundaje, and Anne Brunet. 2017. Chromatin accessibility dynamics reveal novel functional enhancers in c. elegans. *Genome Res.*, 27(12):2096–2107.

Emma K Farley, Katrina M Olson, Wei Zhang, Alexander J Brandt, Daniel S Rokhsar, and Michael S Levine. 2015. Suboptimization of developmental enhancers. *Science*, 350(6258):325–328.

Kirsten Ferner and Andrea Mess. 2011. Evolution and development of fetal membranes and placentation in amniote vertebrates. *Respiratory Physiology & Neurobiology*, 178(1):39–50. Energetics and Oxygen Transport Mechanisms in Embryos.

Jeffrey M Granja, M Ryan Corces, Sarah E Pierce, S Tansu Bagdatli, Hani Choudhry, Howard Y Chang, and William J Greenleaf. 2021. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat. Genet.*, 53(3):403–411.

Kimberly E Inman and Karen M Downs. 2007. The murine allantois: emerging paradigms in development of the mammalian umbilical cord and its relation to the fetus. *Genesis*, 45(5):237–258.

Cory Y McLean, Dave Bristor, Michael Hiller, Shoa L Clarke, Bruce T Schaar, Craig B Lowe, Aaron M Wenger, and Gill Bejerano. 2010. GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.*, 28(5):495–501.

Len A Pennacchio, Nadav Ahituv, Alan M Moses, Shyam Prabhakar, Marcelo A Nobrega, Malak Shoukry, Simon Minovitsky, Inna Dubchak, Amy Holt, Keith D Lewis, Ingrid Plajzer-Frick, Jennifer Akiyama, Sarah De Val, Veena Afzal, Brian L Black, Olivier Couronne, Michael B Eisen, Axel Visel, and Edward M Rubin. 2006. In vivo enhancer analysis of human conserved non-coding sequences. *Nature*, 444(7118):499–502.

Blanca Pijuan-Sala, Jonathan A Griffiths, Carolina Guibentif, Tom W Hiscock, Wajid Jawaid, Fernando J Calero-Nieto, Carla Mulas, Ximena Ibarra-Soria, Richard C V Tyser, Debbie Lee Lian Ho, Wolf Reik, Shankar Srinivas, Benjamin D Simons, Jennifer Nichols, John C Marioni, and Berthold Göttgens. 2019. A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature*, 566(7745):490–495.

Aaron M Wenger, Shoa L Clarke, James H Notwell, Tisha Chung, Geetu Tuteja, Harendra Guturu, Bruce T Schaar, and Gill Bejerano. 2013. The enhancer landscape during early neocortical development reveals patterns of dense regulation and co-option. *PLoS Genet.*, 9(8):e1003728.

Patricia J. Wittkopp and Gizem Kalay. 2012. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nature Reviews Genetics*, 13(1):59–69.

Gregory A. Wray. 2007. The evolutionary significance of cis-regulatory mutations. *Nature Reviews Genetics*, 8(3):206–216.