



**HAL**  
open science

# Comprehensive mapping of human genetic variation at Epromoters

Jing Wan

► **To cite this version:**

Jing Wan. Comprehensive mapping of human genetic variation at Epromoters. Life Sciences [q-bio]. Aix Marseille university, 2024. English. NNT : . tel-04646809

**HAL Id: tel-04646809**

**<https://hal.science/tel-04646809v1>**

Submitted on 12 Jul 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

.....

# THÈSE DE DOCTORAT

Soutenue à Aix-Marseille Université  
21/06/2024

## Jing WAN

### Cartographie des variations génétiques humaines dans les Epromoters

**Discipline**

Biologie santé

**Spécialité**

Génomique et Bioinformatique

**École doctorale**

ED 62 SCIENCES DE LA VIE ET DE LA SANTE

**Laboratoire/Partenaires de recherche**

Theories and Approaches of Genomic

Complexity (TAGC)

Inserm 1090



**Composition du jury**

Mikhail SPIVAKOV                      Rapporteur  
Imperial College, London, UK

Anaïs FLORE BARDET                      Rapporteuse  
University of Strasbourg, Illkirch, France

Denis PUTHIER                              Président du jury  
Aix-Marseille University, Marseille, France

Salvatore SPICUGLIA                      Directeur de thèse  
TAGC-Inserm, Marseille, France

.....

## DOCTORAL THESIS

AIX-MARSEILLE UNIVERSITY  
21/06/2024

# Jing WAN

## Comprehensive mapping of human genetic variation at Epromoters

**Discipline**

Biologie santé

**Specialty**

Genomics and Bioinformatics

**Doctoral School**

ED 62 LIFE AND HEALTH SCIENCES

**Laboratory/Research Partners**

Theories and Approaches of Genomic

Complexity (TAGC)

Inserm 1090



**Composition of jury**

Mikhail SPIVAKOV

Reporter

Imperial College, London, UK

Anaïs FLORE BARDET

Reporter

University of Strasbourg, Illkirch, France

Denis PUTHIER

President of jury

Aix-Marseille University, Marseille, France

Salvatore SPICUGLIA

Supervisor

TAGC-Inserm, Marseille, France

# Acknowledgments

I would like to express my deepest gratitude to all those who have supported and guided me throughout my PhD journey.

First and foremost, I am grateful to my supervisor, **Dr. Salvatore Spicuglia**, for his invaluable support, guidance, and suggestions from the beginning of my PhD to the end thesis. His mentorship has been instrumental in shaping both my research and professional development over the past four years.

I would also like to thank my thesis jury members, **Dr. Mikhail Spivakov, Dr. Anaïs Flore Bardet, and Dr. Denis Puthier**, for their constructive comments and suggestions during the thesis review process. Their expertise has greatly enhanced the quality of my work.

I deeply thank my colleagues, **Antoinette van Ouwerkerk** and **Juliette Malfait**, whose collaboration has been essential, particularly in the experimental aspects of my research. Their contributions, assistance, and unwavering support have significantly enriched my study, thesis, and presentations.

I would like to acknowledge **Dr. Benoit Ballester, Jean-Christophe Mouren, and Pierre de Langen** for their provision of the remap analysis and RNAPII resources. I am especially grateful to Benoit for his continuous encouragement and support. I also thank **Carla Heredia and Dr. Jean-Christophe Andrau** for conducting the G4 analysis. I thank **Aurelie Bergon, Nori Sadouni, and Guillaume Charbonnier** for their computational support.

I am appreciative of the guidance and career advice provided by the thesis advisory board members from ENHPATHY, **Robin Andersson, Gioacchino Natoli, and Dariusz Plewczynski**, during network meetings.

I also thank my colleagues at TAGC, including **Francesco Leonetti, José David Abad, Junhua Su, Tannia Uribe, Yannan Fan, Gaele Farah, Iris Manosalva, Yasmina Kermezli, Aitor Gonzalez, Lydie Pradel, Nathalie Arquier, Himanshu Singh, Saadat Hussain, Charbel Souaid, Davide Cavalieriand**, for their support and kindness in the work.

I acknowledge the generous research and financial support from TAGC, Inserm, Aix-Marseille University, the ENHPATHY network, the Marie Curie fellowship, and the MARMARA grant, all of which have been crucial in the completion of my PhD.

I would like to extend my heartfelt thanks to another important group of people—my Chinese friends. Thank you all for the past three years of wonderful companionship and support, which made my life in Marseille truly memorable.

Lastly, I am deeply grateful to my parents and family. I also dedicate this work to the memory of my grandmother. Your unwavering love and support have been invaluable to me. Thank you all from the bottom of my heart.

# Affidavit

I, undersigned, Jing Wan, hereby declare that the work presented in this manuscript is my own work, carried out under the scientific supervision of Salvatore Spicuglia, in accordance with the principles of honesty, integrity and responsibility inherent to the research mission. The research work and the writing of this manuscript have been carried out in compliance with both the French national charter for Research Integrity and the Aix-Marseille University charter on the fight against plagiarism.

This work has not been submitted previously either in this country or in another country in the same or in a similar version to any other examination body.

Place Marseille, date 19 avril 2024

Jing Wan



Cette œuvre est mise à disposition selon les termes de la [Licence Creative Commons Attribution - Pas d'Utilisation Commerciale - Pas de Modification 4.0 International](https://creativecommons.org/licenses/by-nc-nd/4.0/).

# List of publications/patents and conference participations

## Publications

1. Héctor Castillo, Patricia Hanna, Laurent M. Sachs, Nicolas Buisine, Francisco Godoy, Clément Gilbert, Felipe Aguilera, David Muñoz, Catherine Boisvert, Mélanie Debiais-Thibaud, **Jing Wan**, Salvatore Spicuglia, Sylvain Marcellini. "Xenopus tropicalis osteoblast-specific open chromatin regions reveal promoters and enhancers involved in human skeletal phenotypes and shed light on early vertebrate evolution." **Cells & Development** (2024): 203924.
2. Juliette Malfait, **Jing Wan**, and Salvatore Spicuglia. "Epromoters are new players in the regulatory landscape with potential pleiotropic roles." **BioEssays** 45, no. 10 (2023): 2300012.
3. Yongjun Tan, Xiaohao Yan, Jialei Sun, **Jing Wan**, Xinxin Li, Yingzhang Huang, Li Li, Longjian Niu, and Chunhui Hou. "Genome-wide enhancer identification by massively parallel reporter assay in Arabidopsis." **The Plant Journal** 116, no. 1 (2023): 234-250.
4. Longjian Niu, Wei Shen, Zhaoying Shi, Yongjun Tan, Na He, **Jing Wan**, Jialei Sun et al. "Three-dimensional folding dynamics of the Xenopus tropicalis genome." **Nature Genetics** 53, no. 7 (2021): 1075-1087.
5. Longjian Niu, Wei Shen, Yingzhang Huang, Na He, Yuedong Zhang, Jialei Sun, **Jing Wan** et al. "Amplification-free library preparation with SAFE Hi-C uses ligation products for deep sequencing to improve traditional Hi-C analysis." **Communications biology** 2, no. 1 (2019): 267.

## Conference participations

1. MarMaRa ANNUAL SYMPOSIUM 2024, Rare diseases and beyond: navigating the societal and environmental challenges, 18 & 19 APRIL 2024, Marseille, France. (poster presentation)
2. EMBO Workshop, Enhanceropathies: Understanding enhancer function to understand human disease II, 17 – 20 October 2023, Marseille, France. (poster presentation)
3. EMBO Workshop, Enhanceropathies: Understanding enhancer function to understand human disease I, 6–9 October 2021, Santander, Spain. (poster presentation)

# Résumé

La transcription représente le processus biologique le plus complexe du dogme central. La régulation transcriptionnelle contrôle l'expression des gènes, déterminant si les gènes sont transcrits, leur timing et leur abondance. Les principaux acteurs de la régulation transcriptionnelle comprennent l'ARN polymérase, les facteurs de transcription et les éléments régulateurs de l'ADN. Cette thèse se concentre sur les éléments régulateurs, avec un accent particulier sur les activateurs et les promoteurs, qui sont essentiels et largement distribués dans les génomes.

Il existe de plus en plus de preuves qu'un large éventail de maladies humaines et de traits physiologiques sont influencés par la variation génétique des éléments cis-régulateurs. Des études antérieures ont montré qu'un sous-ensemble d'éléments promoteurs, appelés Epromoters, sont capables de réguler les gènes proximaux et distaux en cis. Cela ouvre un paradigme dans l'étude des variantes régulatrices, car les polymorphismes mononucléotidiques (SNP) au sein des Epromoters pourraient influencer l'expression de plusieurs gènes (distaux) en même temps, ce qui pourrait faciliter l'identification des gènes cibles.

L'objectif principal de mon doctorat était de créer une ressource complète des Epromoteurs humains à l'aide de tests rapporteurs à haut débit nouvellement générés et accessibles au public. Nous montrons que les Epromoters présentaient des caractéristiques intrinsèques et épigénétiques qui les distinguent des promoteurs typiques. En intégrant les études d'association pangénomique (GWAS), les loci de traits quantitatifs d'expression (eQTL) et les interactions 3D de la chromatine, nous avons constaté que les variantes régulatrices des Epromoters sont simultanément associées à davantage de maladies et de traits physiologiques, par rapport aux promoteurs typiques. Pour disséquer l'impact réglementaire des variantes d'Epromoter, nous avons évalué leur impact sur l'activité régulatrice en analysant des tests de rapporteurs à haut débit spécifiques alléliques et avons fourni des exemples fonctionnels caractérisés d'Epromoters pléiotropes. En résumé, ce travail a fourni une ressource complète de variantes de régulation soutenant un rôle pléiotrope des Epromoters.

Mots clés: enhancer, promoteur, variante génétique, GWAS, maladie humaine, pléiotropie

# Abstract

Transcription represents the most intricate biological process within the central dogma. Transcriptional regulation controls gene expression, determining whether genes are transcribed, their timing and abundance. Key participants in transcriptional regulation include RNA polymerase, transcription factors, and DNA regulatory elements. This thesis focuses on the regulatory elements, with a particular emphasis on enhancers and promoters, which are essential and widely distributed across genomes.

There is growing evidence that a wide range of human diseases and physiological traits are influenced by genetic variation of cis-regulatory elements. Previous studies have shown that a subset of promoter elements, termed Epromoters, are able to regulate both proximal and distal genes in cis. This opens a paradigm in the study of regulatory variants, as single nucleotide polymorphisms (SNPs) within Epromoters might influence the expression of several (distal) genes at the same time, which could facilitate the identification of target genes.

The main goal of my PhD was to build a comprehensive resource of human Epromoters using newly generated and publicly available high-throughput reporter assays. We show that Epromoters displayed intrinsic and epigenetic features that distinguish them from typical promoters. By integrating Genome-Wide Association Studies (GWAS), expression Quantitative Trait Loci (eQTLs) and 3D chromatin interactions, we found that regulatory variants at Epromoters are concurrently associated with more diseases and physiological traits, compared with typical promoters. To dissect the regulatory impact of Epromoter variants, we evaluated their impact on regulatory activity by analyzing allelic-specific high-throughput reporter assays and provided functional characterized examples of pleiotropic Epromoters. In summary, this work provided a comprehensive resource of regulatory variants supporting a pleiotropic role of Epromoters.

Keywords: enhancer, promoter, genetic variant, GWAS, human disease, pleiotropy



# Contents

<b>Affidavit</b>	<b>3</b>
<b>List of publications/patents and conference participations</b>	<b>5</b>
<b>Résumé</b>	<b>6</b>
<b>Abstract</b>	<b>7</b>
<b>Contents</b>	<b>8</b>
<b>INTRODUCTION</b>	<b>11</b>
<b>Chapter 1. The cis-regulatory elements in the genome</b>	<b>14</b>
1.1 Transcription regulation in eukaryote	14
1.1.1 Participants of transcription	14
1.1.2 Transcription steps	14
1.1.3 Types of transcription regulation	15
1.2 Genome structure	16
1.2.1 Functional structure	16
1.2.2 3D organization structure	17
1.3 Cis-regulatory elements	19
1.3.1 Promoters	20
1.3.2 Enhancers	21
1.3.3 Silencers	23
1.3.4 Insulators	23
1.3.5 Similarities and differences between promoter and enhancer	24
1.3.6 Enhancer-promoter interactions	26
1.3.7 Multiple regulatory roles of DNA sequence	27
1.4 Technologies to identify regulatory elements	27
1.4.1 Chromatin features	28
1.4.2 High-throughput reporter assays	29
1.4.3 Genome editing	32
1.5 Computational strategies to predict regulatory element	34
<b>Chapter 2. Impact of regulatory elements in disease</b>	<b>37</b>

2.1 Genomic variation in the human genome	37
2.1.1 Different types of genomic variation	37
2.1.2 common and rare SNPs	38
2.2 Regulatory element variation in diseases	39
2.3 Link genomic variations with diseases	43
2.3.1 Genome-wide association studies	43
2.3.2 Polygenic risk score	45
2.3.3 Whole-genome sequencing	45
2.4 GWAS SNPs enriched in regulatory elements	46
2.5 Complex effect of regulatory element variation in diseases	47
2.5.1 Redundancy	47
2.5.2 Pleiotropy	49
<b>Chapter 3. Strategies to study the impact of genetic variation in cis-regulatory elements</b>	<b>51</b>
3.1 Evaluation of regulatory sequence variation	51
3.1.1 Genome editing	51
3.1.2 Reporter assay	52
3.1.3 Evaluation of transcription factor binding effect	53
3.2 Target gene identification strategies	55
3.2.1 eQTL	55
3.2.2 High-throughput genetic perturbation	57
3.2.3 Promoter capture Hi-C	59
3.2.4 ABC model	60
3.3 Database and resources to study regulatory element variation	61
<b>Chapter 4. Role of Epromoters in disease</b>	<b>63</b>
4.1 Epromoters identification	63
4.2 General features and potential mechanism(s) of Epromoters	66
4.3 Human diseases associated with Epromoters	67
<b>RESULTS</b>	<b>73</b>
<b>Chapter 5. Background and objectives of the PhD work</b>	<b>74</b>
<b>Chapter 6. Comprehensive mapping of genetic variation at Epromoters reveals pleiotropic association with multiple disease traits</b>	<b>77</b>

6.1 Manuscript	77
Abstract	77
Introduction	78
Results	80
Methods	91
Data availability	101
Code availability	101
Contributions	101
Figures	102
Supplemental Figures	114
Supplemental information 1	118
6.2 Data resource and sharing	125
6.3 Additional work	126
<b>Chapter 7. Epromoters function as a hub to recruit key transcription factors required for the regulation of stress-response clusters</b>	<b>129</b>
<b>DISCUSSION AND PERSPECTIVES</b>	<b>132</b>
<b>Chapter 8. Discussion</b>	<b>133</b>
<b>Chapter 9. Perspectives</b>	<b>139</b>
<b>REFERENCES</b>	<b>141</b>
<b>ANNEXES</b>	<b>163</b>

# **INTRODUCTION**

The central dogma, proposed by Francis Crick in 1958, is still the basic framework of molecular biology today. It described the sequential information transfer between DNA, RNA, and protein (Figure 1). These sequential information transfers include replication, transcription, and translation, which are the most important biological processes in cells. Among the three biological processes, transcription has been considered the most complex and variable process. Accordingly, there are three fundamental questions in transcription: Whether a gene is transcribed? When a gene is transcribed? How much a gene is transcribed?

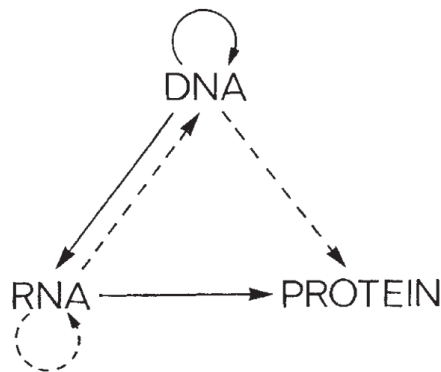


Figure 1 The central dogma proposed by Francis Crick (Crick, 1970).

The answers to the three questions are determined by the working status of the transcription machine, RNA polymerase. RNA polymerase works by binding to DNA, which the binding requires intermediate proteins: transcription factors. The combination of thousands of transcription factors binding with DNA determines whether and when genes are transcribed, that is, which genes are transcribed in which cells.

This also involves the issue of how much transcription is required. Assuming that the working efficiency of RNA polymerase is fixed, the transcription output would be determined by the working time, that is, the total length of time that RNA polymerase binds to DNA in the cell. The total length of RNA polymerase binding time, which can be inferred from the above, is determined by the transcription factors binding affinity and the probability of encountering the transcription factors in nucleus space. The binding affinity of transcription factors is determined by the DNA sequence. The probability of encountering transcription factors in the nucleus, that is, the density of transcription factor aggregation in a location, is determined by the number of DNA binding sites and proteins with high affinity in nearby space.

Considering that chromatin is folding in the nucleus, one mechanism to create more binding opportunities for transcription factors is to pull the DNA locations with the same binding sites to close together. Thus, this can attract more transcription factors to aggregate. One of the mechanisms for this pulling action is the cohesin model, which is a ring-shaped protein complex for chromatin loop forming. These DNA fragments in the distal location that attract transcription factors and thereby enhance transcription are called enhancers.

The advantage of this mechanism is that it can control the amount of gene transcription more precisely and stably. Accordingly, the important questions are: What are the mechanisms that pull DNA together? Whether it's possible to find the enhancers of a promoter by their common transcription factors binding sites? Are there other mechanisms that increase the probability of transcription factor binding? That's opening the door for transcription regulation study.

# Chapter 1. The cis-regulatory elements in the genome

## 1.1 Transcription regulation in eukaryote

### 1.1.1 Participants in transcription

Transcription in eukaryotes happens in the cell nucleus during interphase. The entire process of transcription involves a large number of molecules, mainly including DNA, RNA, RNA polymerase, transcription factors, transcription cofactors, etc. Among them, the RNA polymerase complex is the core machine of transcription and is responsible for the synthesis of RNA. In 1969, Robert G. Roeder discovered the nuclear RNA polymerases I, II, and III in diverse eukaryotes (Roeder & Rutter, 1969). The biochemical identification of three nuclear RNA polymerases fundamentally altered our understanding of gene regulation. Transcription factors are essential for initiating gene transcription. They help RNA polymerase locate the correct location on the DNA and facilitate its binding and initiation of RNA synthesis. There are many types of transcription factors, more than 1,600 in human cells. Transcription cofactors do not directly bind DNA but interact with other transcription factors to enhance their activity, thereby increasing transcription efficiency.

### 1.1.2 Transcription steps

The transcription process can be divided into three main steps: initiation, elongation, and termination (Figure 1.1.2) (Cramer, 2019). During the initiation stage, transcription factors bind to the promoter region and combine with RNA polymerase II to form a pre-initiation complex. This complex recognizes and unwinds the DNA, forming a small open complex that prepares the template strand for RNA synthesis. During the elongation stage, RNA polymerase begins to synthesize RNA along the DNA template strand, adding nucleotides complementary to the DNA template strand one by one from the 5' to 3' direction. As RNA polymerase moves, a transcription bubble forms where the DNA locally unwinds, allowing the newly synthesized RNA strand to detach from the template DNA. During the termination stage, the transcription process nears completion when RNA polymerase encounters a termination sequence on the DNA.

The newly synthesized RNA molecule (pre-mRNA) is released from RNA polymerase and undergoes a series of post-transcriptional modifications, such as capping, splicing, and polyadenylation, ultimately resulting in the formation of mature mRNA.

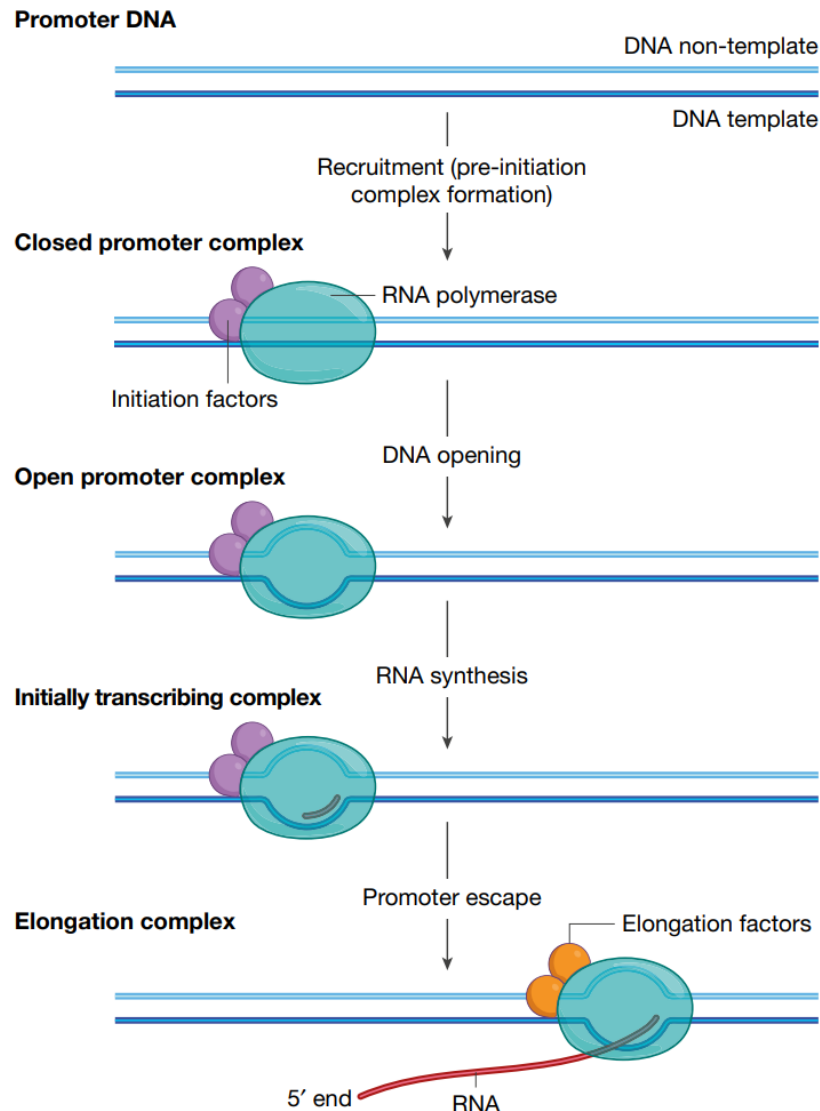


Figure 1.1.2 Key steps of gene transcription (Cramer, 2019).

### 1.1.3 Types of transcription regulation

The regulation of transcription is primarily concentrated in the initiation stage. Firstly, the promoter sequences themselves can influence the efficiency of transcription by affecting the binding of transcription factors. Then, transcription factors can specifically bind to promoters to initiate the assembly of RNA polymerase and promote the initiation of transcription. And transcription factors can act as activators (enhancing gene



expression) or repressors (reducing gene expression). Regulatory elements located distal to the gene can also enhance or inhibit transcription through interactions with the promoter. Additionally, since DNA is wrapped around histones forming chromatin, the structure of chromatin can be modulated through chemical modifications (such as methylation and acetylation), thereby influencing the accessibility of transcription factors and RNA polymerase. Methylation of DNA at specific CpG islands associated with gene silencing can also affect the binding of transcription factors.

## **1.2 Genome structure**

### **1.2.1 Functional structure**

In 2003, the completion of the Human Genome Project (HGP) provided a blueprint for the human genome. Its primary goals were to determine the sequence of the human genome and to identify all human genes. Genes are fundamental to understanding the genome's functional structure. A typical gene structure includes exons, introns, 5' untranslated region (5' UTR), transcription start sites (TSS), 3' untranslated region (3' UTR), and transcription termination sites (TTS) (Figure 1.2.1). Beyond the genes, the intergenic regions were thought as "junk DNA". Building upon the foundational knowledge provided by the HGP, the ENCODE project commenced in 2003 intending to catalog all functional elements in the human genome. While the HGP identified the letters of the genomic code, ENCODE sought to understand the language—how these letters are used to direct cellular function through the regulation of gene expression. ENCODE's findings have dramatically expanded our understanding of the genome: a significant revelation was that approximately 80% of the genome is involved in at least one biochemical activity, suggesting that regions previously dismissed as "junk" DNA have functional importance. These DNA sequences help to control when, where, and how much a gene expressed, which are called regulatory elements. They include promoters, enhancers, silencers, and insulators.

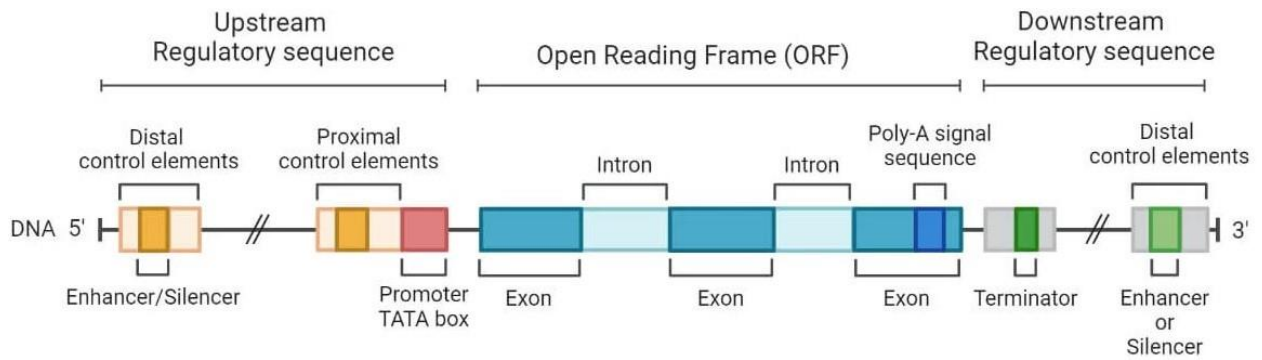


Figure 1.2.1 Eukaryotic Gene Structure. Adapted from Microbe Notes.

(<https://microbenotes.com>)

## 1.2.2 3D organization structure

For a long time, research mainly focused on genome sequences, while ignoring the genome structure in physical space. With the widespread application of Hi-C technology, people have realized the three-dimensional structure of chromatin at the genome-wide scale across species, and these structures play an important role in gene regulation. In 2009, while processing the first Hi-C data in human, Lieberman et al. discovered that the genome is divided into two compartments, representing open and closed chromatin respectively, and are highly correlated with the activity of gene transcription (Lieberman-Aiden et al., 2009). With the improvement of Hi-C data resolution, several studies have found that topologically associating domains (TADs) widely exist in the genome in mammals and *Drosophila*, where genomic interactions are strong within a domain but are sharply depleted on crossing the boundary between two TADs (Dixon et al., 2012) (Nora et al., 2012) (Sexton et al., 2012) (Hou et al., 2012). At the same time, these studies also found enrichment of CTCF at the boundaries of TADs. In 2014, Rao et al. identified genome-wide chromatin loops through higher-resolution human Hi-C data, and these loops often reveal enhancer-promoter interactions (Rao et al., 2014). More precisely, Javierre et al. deployed a promoter capture Hi-C approach to generate high-resolution maps of promoter interactions in 17 primary blood cell types. Therefore, a hierarchical 3D chromatin organization model has been widely mentioned, which consists of compartments at megabase level, TADs at sub-megabase level, and loops between genomic sites (Figure 1.2.2a).

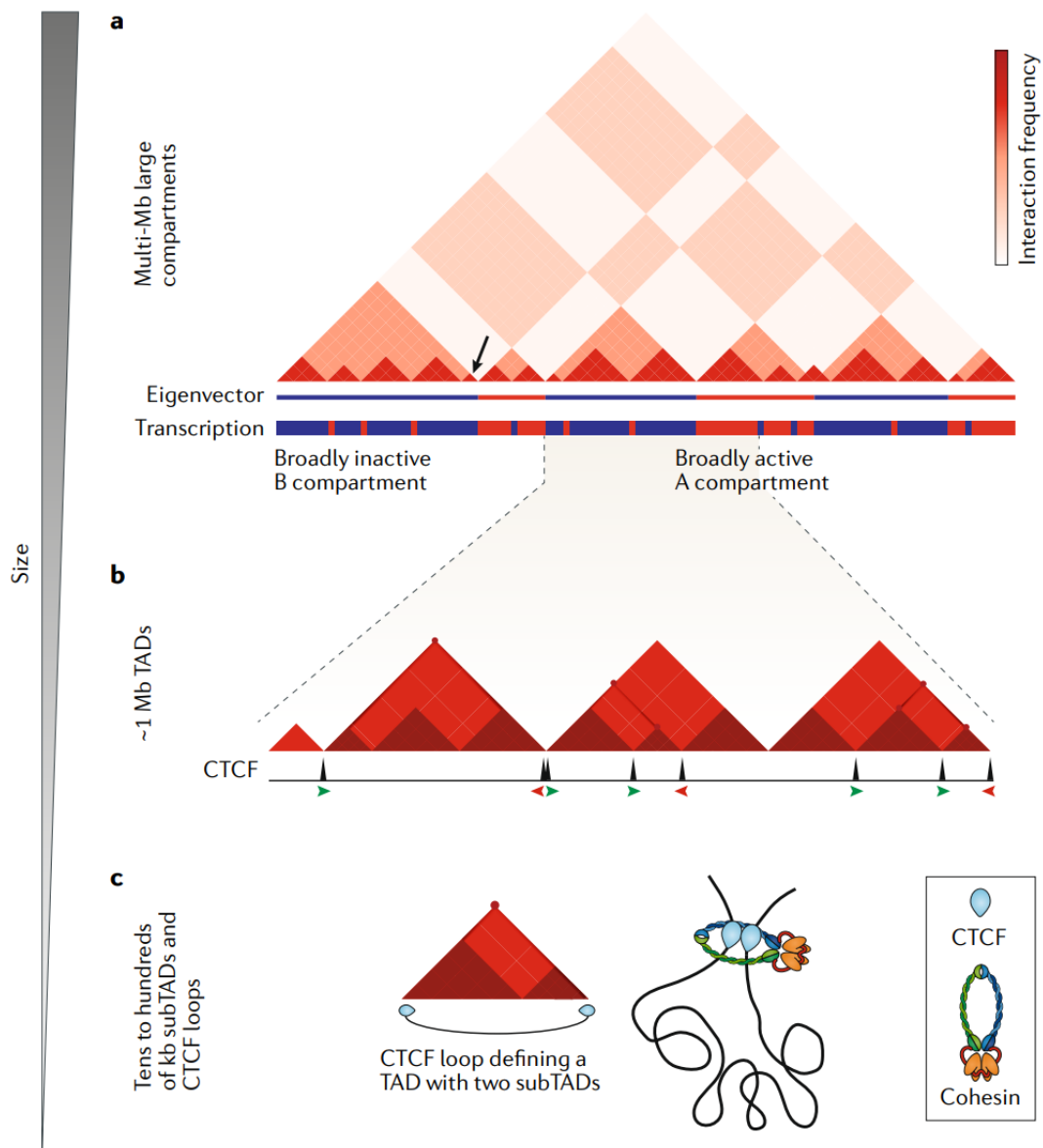


Figure 1.2.2a A hierarchical 3D chromatin organization model has been widely mentioned (Rowley & Corces, 2018).

However, the latest chromatin imaging results provide evidence to support the view that TAD regions do not show clear and stable chromatin structure at the single-cell level (Bhat et al., 2021). TAD may be a statistical characteristic of Hi-C in population cells, rather than a physical cell structure. With studies based on liquid phase separation, another physical model was proposed. The transcription regulation model based on Liquid-Liquid Phase Separation (LLPS) suggests that transcription factors and other regulatory molecules can undergo phase separation, creating concentrated microenvironments within the cell nucleus (Figure 1.2.2b) (Bhat et al., 2021). These microenvironments facilitate enhanced interactions between transcription machinery

and specific genomic regions, thereby regulating gene expression more efficiently. Therefore, nuclear compartmentalization has the potential to serve as a mechanism for quantitative control of gene expression.

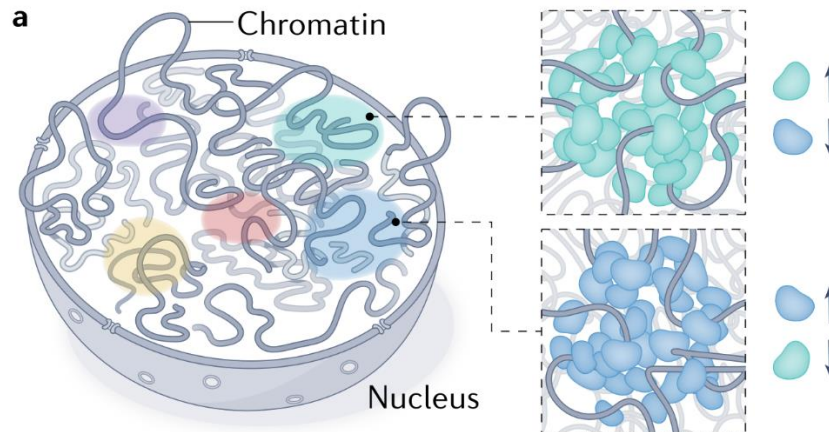


Figure 1.2.2b Nuclear compartments contain high local concentration of specific molecules in 3D space (Bhat et al., 2021). Upward arrows represent higher concentration and downward arrows represent lower concentration.

## 1.3 Cis-regulatory elements

The cis-regulatory elements are DNA sequences that regulate gene transcription. It is generally believed that regulatory elements are located in regions outside the gene, that is intergenic regions. Currently, the main regulatory elements are divided into transcriptional activating elements, such as promoters and enhancers, and transcriptional repressive elements, such as silencers and insulators, according to their functions (Figure 1.3). The results of the second phase of ENCODE show that the vast majority (80.4%) of the human genome participates in at least one biochemical RNA- and/or chromatin-associated event in at least one cell type (Consortium, 2012). This means that the function of non-coding DNA sequences was underestimated. The following introduction will focus on the two most common activating regulatory elements: promoter and enhancer.

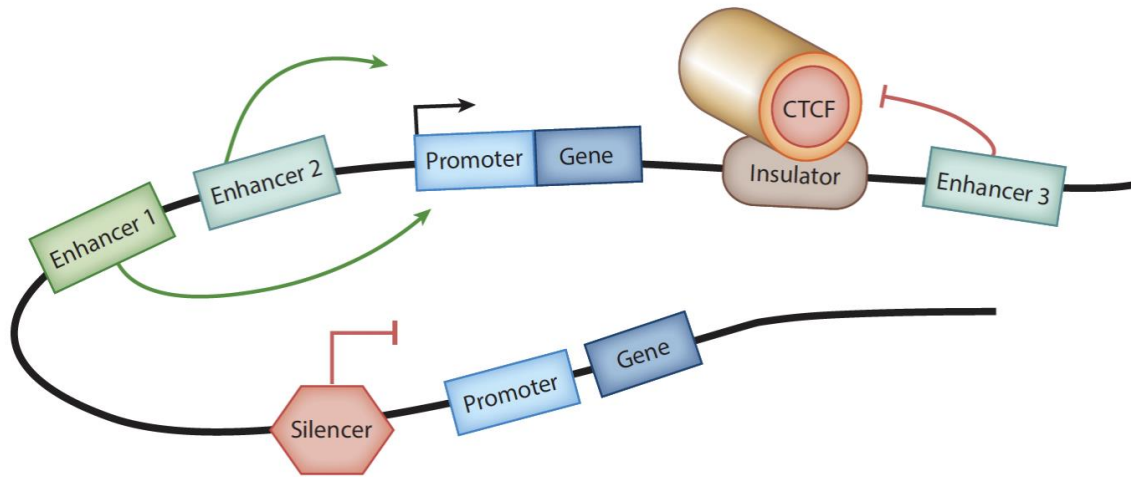


Figure 1.3 Different types of cis-regulatory elements (Chatterjee & Ahituv, 2017).

### 1.3.1 Promoters

Promoters are defined as the DNA region that RNAPII binds for transcription initiation. They are located at the 5' end of the gene, which is the beginning of the gene transcription. More precisely, the core promoter is typically defined as the  $\pm 50$  base pair (bp) region surrounding the transcription start site (TSS). The most well-known core promoter elements are the TATA box and the initiator (INR) element, which are binding sites for the pre-initiation complex. Because some TFs bind proximally to, but not within, the core promoter region, a larger, arbitrarily sized region around the TSS encompassing the core promoter and this 'proximal promoter' region is often referred to as the 'promoter' (Andersson & Sandelin, 2020).

Because TSSs are central to the identification of core promoters, techniques based on the sequencing of RNAs have been highly instrumental in this task, like cap analysis of gene expression (CAGE) and global run-on sequencing (GRO-seq). Based on these RNA detection studies, Andersson et al (Andersson & Sandelin, 2020) made the following summary of promoter features: First, RNAPII initiation is often dispersed over a local area, within the same nucleosome-depleted region (NDR), resulting in multiple close-by TSSs with varying initiation frequencies. Second, most genes have many distinct TSS clusters (termed 'alternative promoters'). The choice of an alternative promoter may alter the final protein product. Third, RNAPII pauses downstream of the TSS before entering a state of active elongation. Fourth, the vast majority of gene TSSs

are accompanied by an additional, proximal, upstream TSS on the opposite strand. Generally, the upstream, divergent TSS produces short (<500 bp), unspliced transcripts, termed 'promoter-upstream transcripts' (PROMPTs) (Preker et al., 2008).

### **1.3.2 Enhancers**

In 1981, enhancer was firstly reported by Banerji et al in a non-coding region of the simian virus 40 (SV40) genome (Banerji et al., 1981), which increased expression at a distance remote from the reporter gene's promoter and independent of the enhancing region's orientation (Gasperini et al., 2020). This became the original definition of an enhancer. Gasperini et al summarized a history of operational definitions of enhancers, which described four stages of enhancer studies according to the detection technologies (Figure 1.3.2). In their opinion, an operational definition is not what an enhancer is, but rather follows from the practical framework that we use to distinguish biological enhancers from other sequences. They proposed the biological definition of an enhancer should meet three criteria: first, deletion from its native genomic context results in altered expression of a potential target gene; second, evidence for a cis-acting mechanism; and last, one line of orthogonal evidence that the underlying sequence is an enhancer.

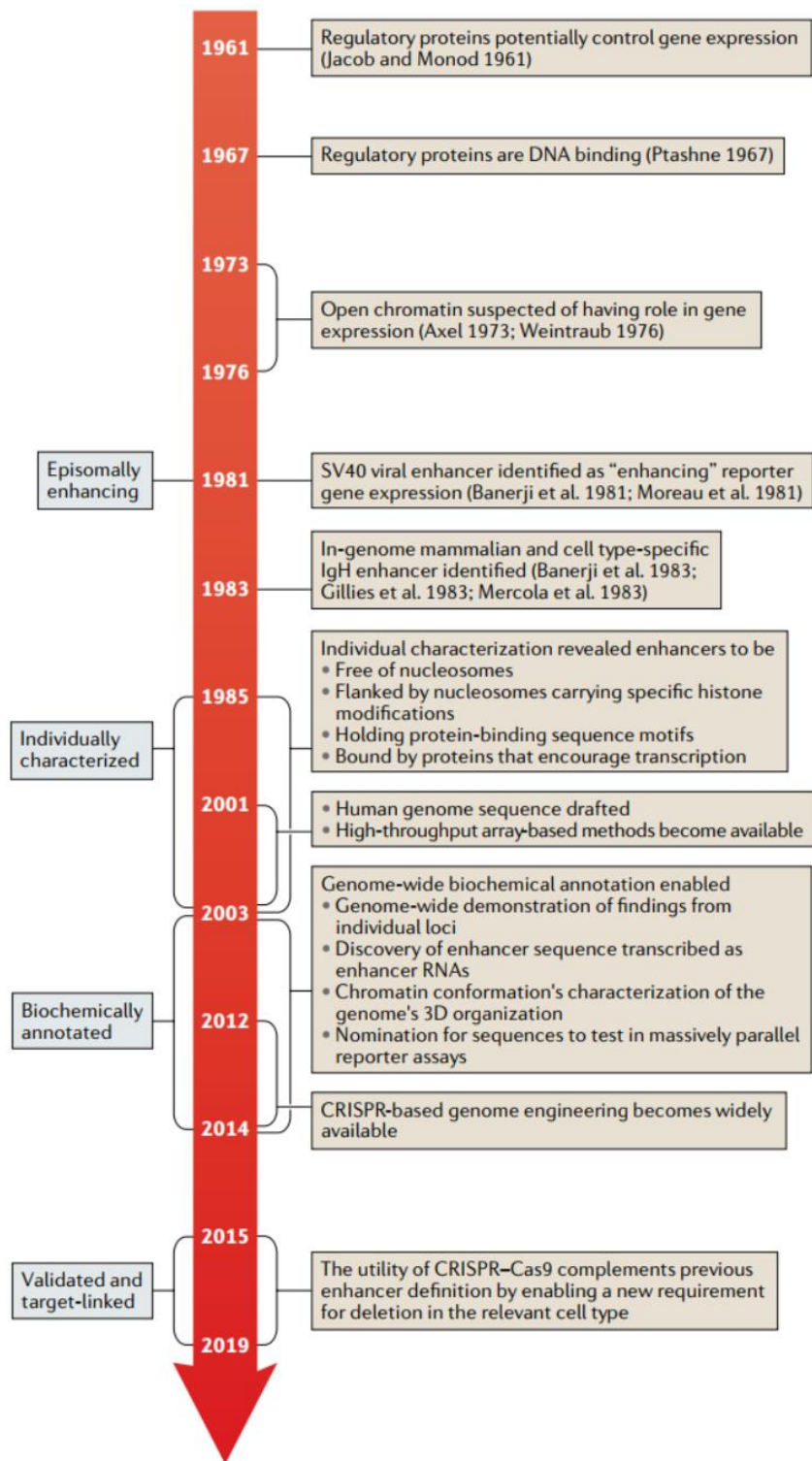


Figure 1.3.2 A history of operational definitions of enhancers (Gasperini et al., 2020).

With the development of high-throughput sequencing technology, the identification of enhancers has also shifted from individual-level to genome-wide. The number of human enhancers identified through characterization and regulatory activity has reached millions (EnhancerAtlas 2.0) (Gao & Qian, 2020). Enhancers are also widely distributed in the genome. In addition to being enriched in intergenic regions and introns, enhancers are also found around TSS and in exons. The sequence features of enhancers are characterized by dense binding of transcription factor binding sites, cell type specificity, and evolutionary conservation. Enhancers are biochemically characterized by enrichment of H3K4me1 and H3K27ac, accompanied by enhancer RNA (eRNA) transcription.

### **1.3.3 Silencers**

Silencers were initially defined as sequence elements that are capable of repressing promoter activity in an orientation and position independent fashion, in the context of a native or a heterologous promoter (Brand et al., 1985). Compared with our knowledge of promoters and enhancers, our knowledge of silencers is largely lacking. A recent review from Pang et al. discussed the biology of silencers, including methods for their discovery, epigenomic and other characteristics, and modes of function of silencers (Pang et al., 2023). Hussain et al. assessed silencer activity from DNase hypersensitive sites by a high-throughput reporter strategy in a mouse T cell line, which provided a general strategy for genome-wide identification and characterization of silencer elements (Hussain et al., 2023).

### **1.3.4 Insulators**

Insulators (also known as boundary elements) function to block genes from being affected by the transcriptional activity of neighboring genes. They thus limit the action of regulatory elements to defined domains and partition the genome into discrete realms of expression. Insulators can block enhancer-promoter communication and prevent the spread of repressive chromatin. Many genome-wide conformation capture studies have helped reveal that insulators are necessary for proper genome-wide organization of topologically associating domains (Chen & Lei, 2019). For example, CTCF-mediated long-range interactions are integral for a multitude of topological features of interphase chromatin, such as the formation of topologically associated

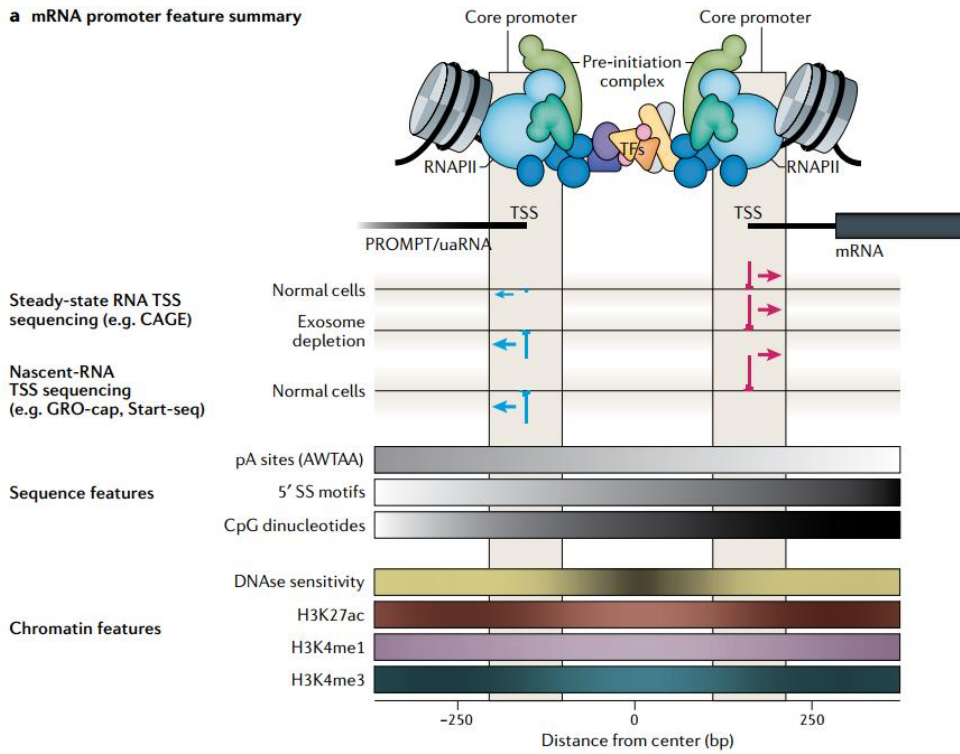


domains, domain insulation, enhancer blocking, and even enhancer function (Ali et al., 2016).

### **1.3.5 Similarities and differences between promoter and enhancer**

Emerging research shows that promoters and enhancers share many similarities. For example: both promoters and enhancers display high chromatin accessibility due to nucleosome degradation; similar histone modifications: H3K4me1, H3K4me3, and H3K27ac; accompanied by the production of short unstable RNAs (Figure 1.3.3). The difference is that promoters show higher H3K4me3 intensity and higher CpG content; enhancers show higher H3K4me1 and H3K27ac intensity, and bidirectional transcription at similar levels (Figure 1.3.3). There have been several studies from different groups showing that some promoters displayed enhancer activity (Engreitz et al., 2016) (Rajagopal et al., 2016) (Diao et al., 2017) (Dao et al., 2017) (See more detail in Chapter 4). This suggests that regulatory elements may have varying degrees of promoter and enhancer activity, essentially affecting proximal or distal transcription initiation.

**a mRNA promoter feature summary**



**b Enhancer feature summary**

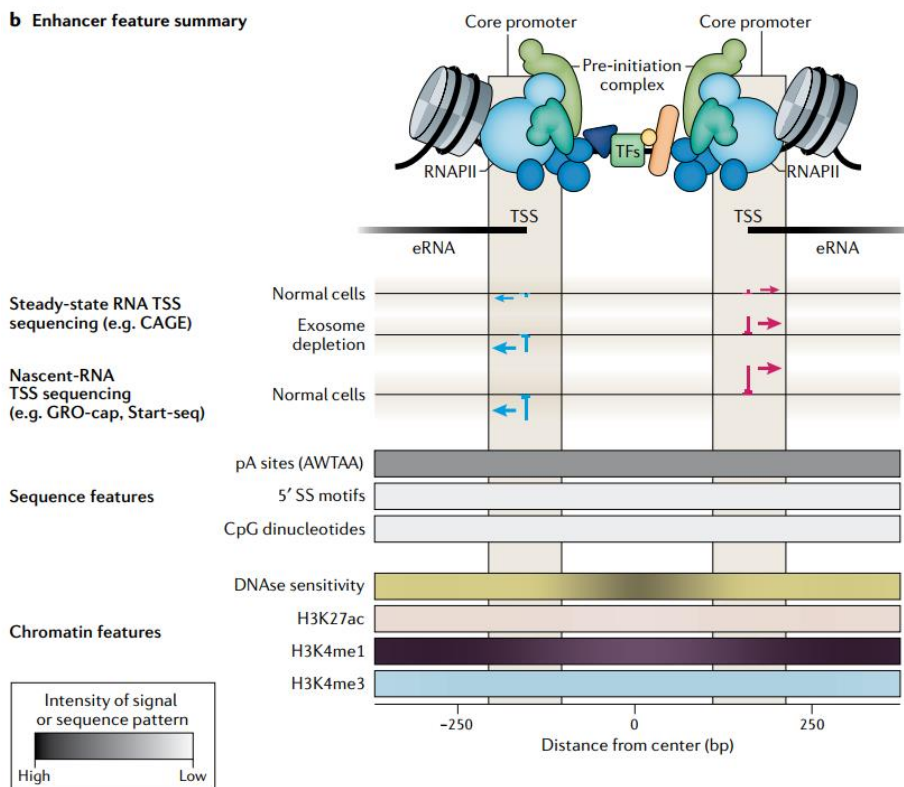


Figure 1.3.3 Features used to distinguish promoters and enhancers (Andersson & Sandelin, 2020)

### 1.3.6 Enhancer-promoter interactions

How do enhancers and promoters interact? A model based on chromatin loop extrusion has been widely proposed (Figure 1.3.4). The model is based on cohesin, a ring-shaped chromatin structural protein complex. Between two CTCF sites in the genome, chromatin forms a loop through cohesin, which provides conditions for the physical proximity of enhancers and genes. Enhancers and promoters form a more stable interaction through the binding of transcription factors, thereby promoting gene transcription. At the same time, such a mechanism also facilitates the interaction between specific enhancers and specific genes. Other cohesin-independent mechanisms to mediate the contacts between promoters and enhancers could exist but not clear yet. For example, Thiecke et al. found that a significant minority of promoter-enhancer contacts are maintained after rapid degradation of cohesin and CTCF (Thiecke et al., 2020).

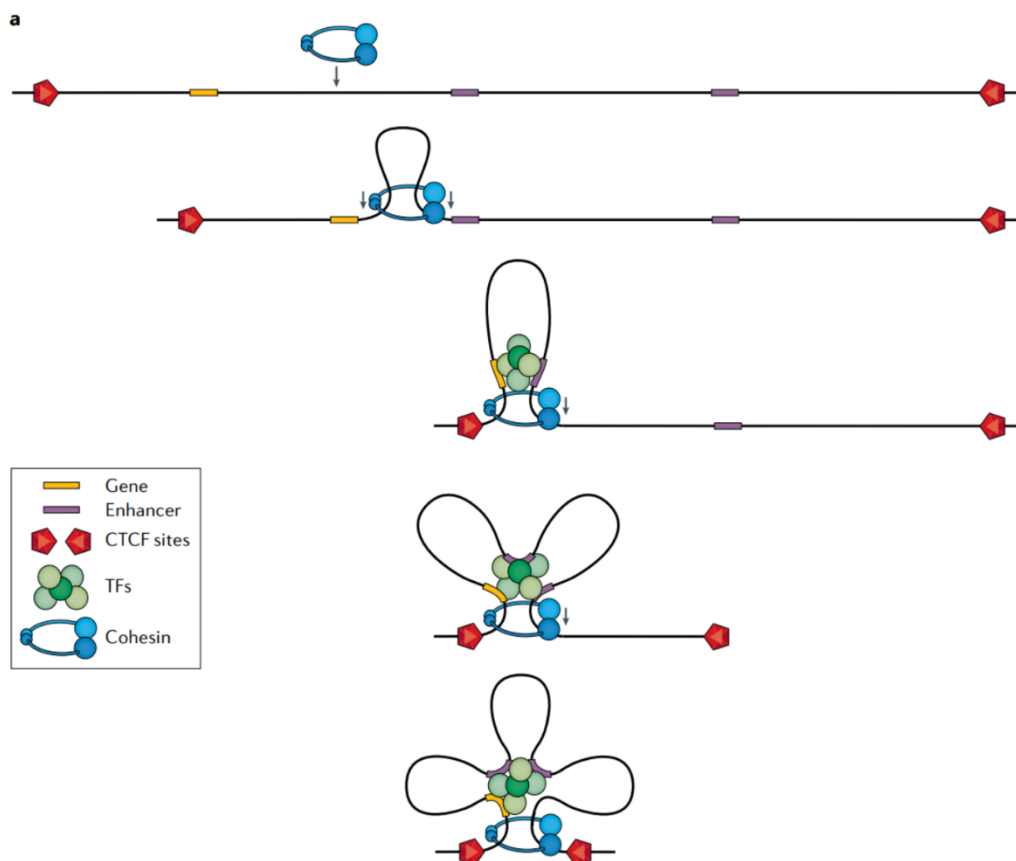


Figure 1.3.4 Formation of enhancer-promoter interactions by loop extrusion and affinity (Oudelaar & Higgs, 2021).

### **1.3.7 Multiple regulatory roles of DNA sequence**

Emerging studies are finding that a single DNA sequence can play multiple regulatory functions. For example, as mentioned before, some promoters displayed enhancer functions. Two recent studies also systematically studied the enhancer function of silencers in *Drosophila* and human genomes respectively (Gisselbrecht et al., 2020) (Huang & Ovcharenko, 2022), and estimated that there are a large number of such dual-functional elements in the genome. In addition, exons that work for protein coding were also found to display enhancer functions (Ahituv, 2016). The multiple roles of regulatory elements have led to the thought that the essence of regulatory elements is the binding affinity between DNA molecules and different proteins. This affinity should be a continuous physical variable, which implies we might consider regulatory elements as a continuous definition instead of artificially dividing them into different types.

## **1.4 Technologies to identify regulatory elements**

Techniques for identifying regulatory elements can be divided into three main categories (Figure 1.4). The first category is to identify regulatory elements through associated chromatin features, including ChIP-seq of histone marks or transcription factors, DNase-seq and ATAC-seq for chromatin accessibility, and nascent RNA detection like GRO-seq and CAGE (Figure 1.4.1). The second category uses reporter assays to detect regulatory activity, such as MPRA, STARR-seq, SURE-seq, etc (Figure 1.4.2). The third category is genome editing, such as CRISPR/Cas9 and CRISPRi.

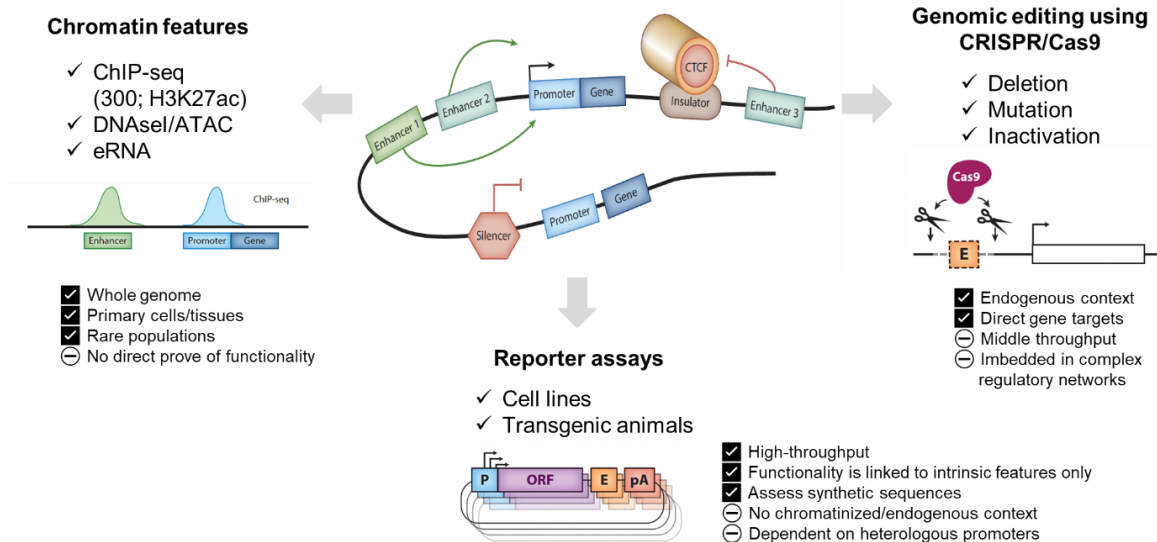


Figure 1.4 Technologies to identify regulatory elements and their advantages and disadvantages.

## 1.4.1 Chromatin features

### Histone marks and TFs identification: ChIP-seq

ChIP-seq, a method used to analyze protein interactions with DNA, combines chromatin immunoprecipitation with massively parallel DNA sequencing to identify the binding sites of DNA-associated proteins. ChIP-seq has been widely used to characterize regulatory elements, including histone modifications and transcription factors. For example, H3K4me1, H3K27ac, and P300 are considered to be associated with enhancers, and H3K4me3 is associated with promoters. A large number of ChIP-seq data sets have been generated. There are over 8,000 ChIP-seq data sets associated with transcription factors in human tissues and cell lines (ReMap2022).

### Chromatin accessibility detection: DNase-seq, ATAC-seq

DNase-I hypersensitive site sequencing (DNase-seq) and Assays for Transposase-Accessible Chromatin sequencing (ATAC-seq) are two widely used protocols for genome-wide identification of open chromatin. DNase-seq and ATAC-seq are based on the use of cleavage enzymes (DNase-I and Tn5, respectively), which recognize and cleave DNA in open chromatin regions. Currently, a large number of DNase-seq and ATAC-seq data sets are also collected in ENCODE.

### Transcription features: CAGE, GRO-seq

The cap analysis of gene expression (CAGE) allows high-throughput identification of sequence tags corresponding to 5' ends of mRNA at the cap sites and the identification of the TSS. CAGE has been widely used in FANTOM5 to detect promoters and enhancers. Global run-on sequencing (GRO-seq) is the most widely used method to measure nascent RNA, which has been used to detect promoter-proximal pausing of RNAP, bidirectional transcription, and enhancer RNA (eRNA).

### Advantages and disadvantages

The advantages of the above technologies for identifying regulatory elements are: genome-wide and high-throughput; library construction technology and bioinformatics analysis are very mature; relatively cheap; easy to apply to a large number of tissues and cell lines of various species; and there are already a large number of available datasets. The disadvantage is that they only detected the chromatin features associated with regulatory elements, not directly measuring the function and activity of regulatory elements. In many cases, these chromatin features do not represent the activity of regulatory elements.

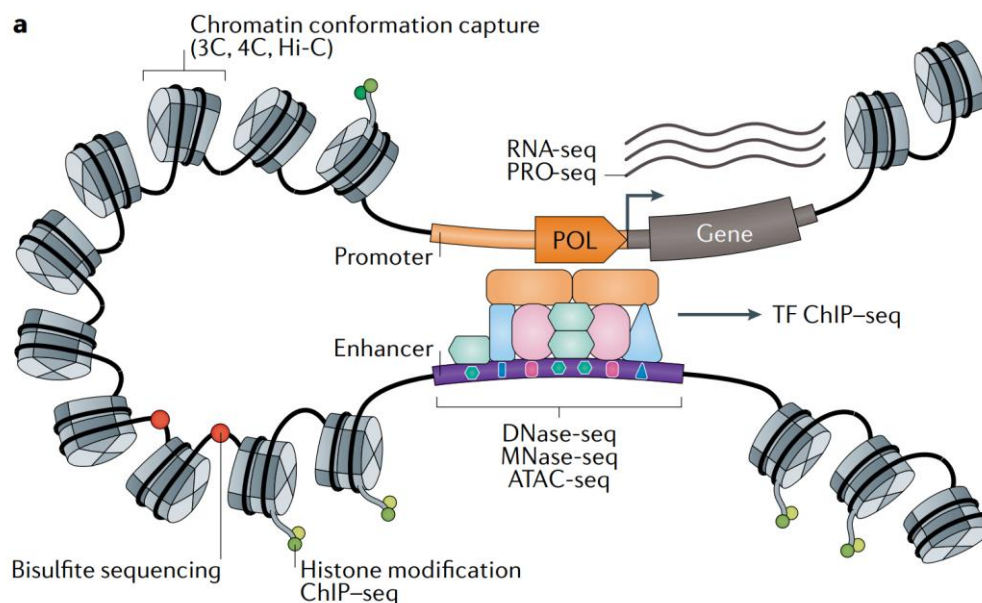


Figure 1.4.1 Technologies to identify regulatory element: based on chromatin features (Gasparini et al., 2020).

## 1.4.2 High-throughput reporter assays

### MPRA

The first massively parallel reporter assay (MPRA) was developed by Patwardhan et al (Patwardhan et al., 2009). They quantitatively assayed the effects of all possible single-nucleotide mutations for three bacteriophage promoters and three mammalian core promoters. The MPRA method consists of the generation of a library of reporter constructs based on microarray synthesis of DNA sequences and unique sequence tags or barcodes (placed in the 3' UTR of the reporter gene). The reporter library is then transfected into cell lines of interest and RNA sequencing of the barcodes is performed, thus providing a quantitative readout of the regulatory activity of the tested regions. Because of the synthetic ability in the library, MPRA was widely used for testing regulatory activity variation of synthetic sequences. That also opened the applications for variants of functional validation in human diseases (See more details in Chapter 3).

### **STARR-seq**

The self-transcribing active regulatory region sequencing (STARR-seq) was developed by Arnold et al. (Arnold et al., 2013), which quantitatively assessed enhancer activity for millions of DNA fragments across the entire *Drosophila* genome (Figure 1.4.2). This method does not require synthesized “barcodes” since the DNA sequences are cloned into the 3' UTR of the reporter gene. The active enhancer will transcribe the reporter gene and themselves, becoming part of the reporter transcript. Thus, the advantage over the classical MPRA is that the tested sequence itself is used as a “barcode”, substantially simplifying the whole procedure to quantify enhancer activity (Figure 1.4.2).

With the complexity and size of mammalian genomes, this technique is not easily implemented, making the formulation of representative libraries a challenge and a very high sequencing depth a necessity. To avoid this issue, a capture-based approach (CapSTARR-seq) to assess a subset of mouse DNase I hypersensitive sites (DHSs) found in developing thymocytes was developed (Vanhille et al., 2015). Here, the regions of interest are captured using custom designed microarrays and cloned into the STARR-seq vector, thus providing cost-effective and accurate quantification of enhancer activity in mammals. Barakat et al. use a combination of chromatin immunoprecipitation and a massively parallel reporter assay (ChIP-STARR-seq) to identify functional enhancers in primed and naive human embryonic stem cells (Barakat et al., 2018). Wang et al. developed HiDRA (High-resolution Dissection of

Regulatory Activity) by coupling accessible chromatin extraction with self-transcribing episomal reporters (ATAC-STARR-seq) (X. Wang et al., 2018). This method allowed them to identify high-resolution driver elements of enhancers in lymphoblastoid cells.

## **SuRE**

The Survey of Regulatory Elements (SuRE) was developed by Arensbergen et al. (van Arensbergen et al., 2017), which tested human promoter activity genome-wide. SuRE can assay more than  $10^8$  DNA fragments with sizes of 0.2–2kb by using a 20bp barcode in the plasmid, which is long enough to include most elements that constitute fully functional promoters. SuRE can work as a high-throughput tool to functionally deconstruct large genomes and systematically identify elements that drive autonomous transcription activity. SuRE operates on a 100- to 1,000-fold larger scale than previous high-throughput promoter assays, sufficient to survey the entire human genome at >50× coverage (van Arensbergen et al., 2017).

## **Advantages and disadvantages**

The advantages of high-throughput reporter assays are: high throughput; functionality linking to intrinsic features only; and the ability to test synthetic sequences outside the genome. Disadvantages are: limited to testing short sequences; most enhancer MPRA studies use the same minimal promoter to test all candidate enhancer sequences, but compatibility between enhancers and core promoters is limited; plasmid-based reporter assays are not testing regulatory elements in endogenous context. However, the last point can be solved in a certain extent through lentivirus-based massively parallel reporter assay (lentiMPRA), which randomly integrates reporter assays into the genome through lentivirus (Inoue et al., 2017).



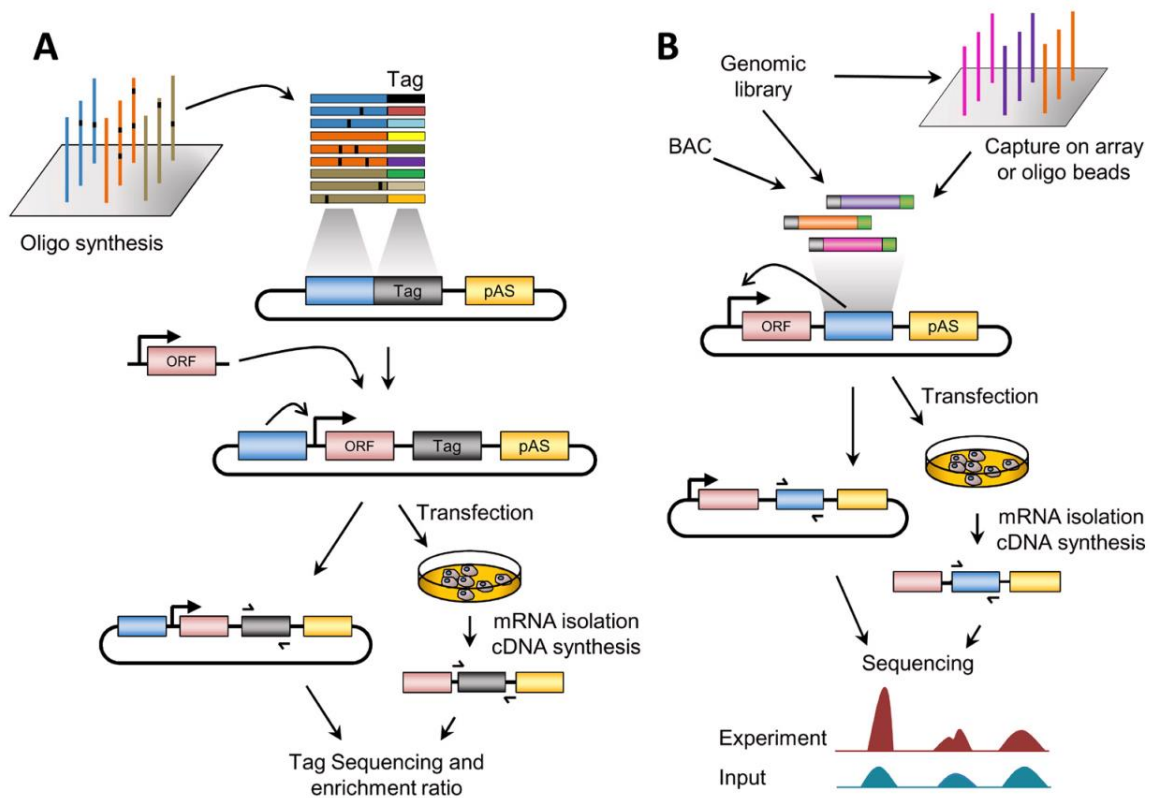


Figure 1.4.2 High-throughput reporter assays to detect regulatory activity (Santiago-Algarra et al., 2017). (A) Overview of massively parallel reporter assay (MPRA). The test sequences (wild-type, variants, etc.) are generally synthesized *in silico* by massive oligonucleotide synthesis with unique barcode tags and cloned into the plasmid backbone. (B) Overview of self-transcribing active regulatory region sequencing (STARR-Seq). A genomic or bacterial artificial chromosome (BAC) library is cloned in the reporter plasmid, downstream of the ORF and upstream of the polyadenylation site (pAS). Alternatively, the regions of interest might be enriched by a capture approach. The reporter library is transfected into cultured cells. Subsequently, mRNA is isolated, and cDNA is synthesized. The cloned regions are sequenced from the plasmid library pool (input) and the cDNA. Differences in the enrichment with respect to the input are proportional to the enhancer activity.

## 1.4.3 Genome editing

### CRISPR/Cas9 based technologies

Genome editing has also been used to validate the function of regulatory elements *in vivo*. The CRISPR system guides the nuclease Cas9 to cut specific DNA through single

guide RNA (sgRNA), thereby enabling precise and accurate genetic perturbation of regulatory elements. Later variants based on the CRISPR system were also developed for functional validation. For example, CRISPR interference (CRISPRi) or CRISPR activation (CRISPRa) was used to inhibit or activate the regulatory elements by the modified Cas9, the link to the gene expression variation (Figure 1.4.3). Recent studies have enabled large-scale functional validation for regulatory elements or variants by combining CRISPRi and single-cell RNA-seq (Gasperini et al., 2019) (Replogle et al., 2022) (Morris et al., 2023).

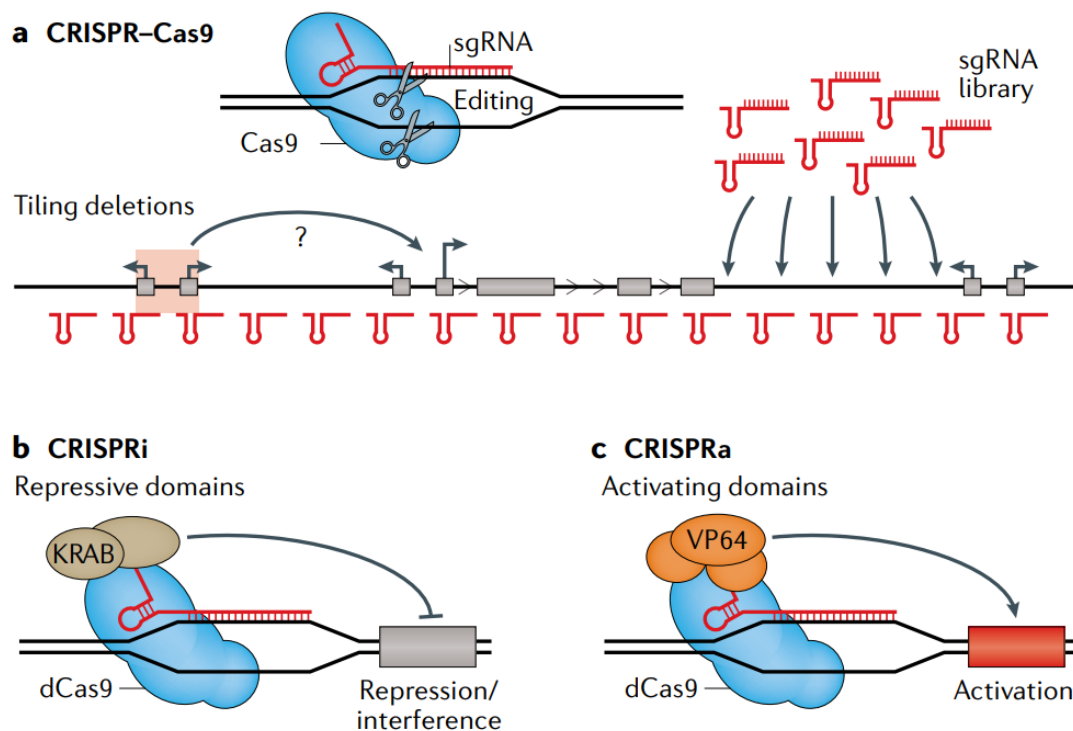


Figure 1.4.3 CRISPR/Cas9 based technologies for perturbing regulatory element (Andersson & Sandelin, 2020)

### Advantages and disadvantages

The advantage of the CRISPR/Cas9 system is that it is based on endogenous context, including local chromatin background and interactions between multiple regulatory elements; and it can directly observe the link between regulatory elements and target genes. The disadvantages are: it is still difficult to conduct high-throughput assessment on a genome-wide scale; the perturbation results in vivo may involve other regulatory factors, such as multiple redundant regulatory elements locally.

# 1.5 Computational strategies to predict regulatory element

From the beginning of regulatory element characterization, many computational strategies have been developed to identify and dissect regulatory elements. Most computational strategies focus on identifying enhancers and promoters. These computational strategies can be classified based on data sources: evolutionary data, histone marks, chromatin accessibility data, transcription factor binding, sequence features, functional screening data, transcription features, etc. The general strategy of traditional methods is: first preprocess these data to extract and select features; then use clustering, classification, regression, and other methods to train the algorithm model; and finally identify and dissect the regulatory features.

In recent years, machine learning has achieved great success in predicting protein structure, but predicting gene regulatory rules is much more difficult (Kim & Wysocka, 2023) (de Boer & Taipale, 2024). Because once the amino acid sequence is determined, the protein structure is almost certain (Figure 1.5a). However, as the basic grammatical unit of the regulatory element sequence, TF binding motif is not 100% fixed. Different TF combinations and distances will affect the activity of the regulatory element. And it is also necessary to consider the relationship between regulatory elements, such as the specificity and compatibility of promoters and enhancers. In addition, gene transcription regulation is quantitative rather than simply on and off. Finally, the context of the regulatory activity must be considered, as the same sequence may have different functions in different cell types. The above factors will all affect the prediction of regulatory elements.

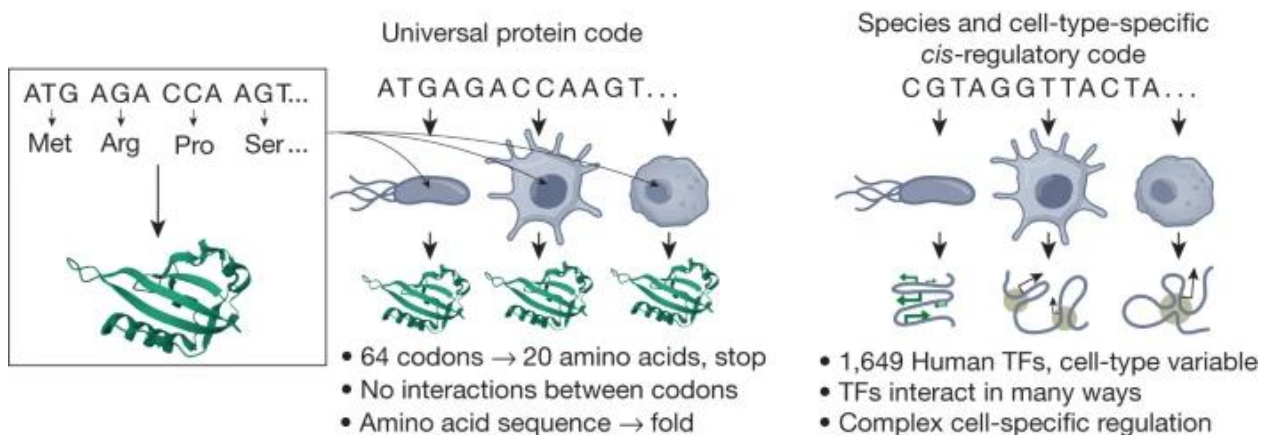


Figure 1.5a Protein versus cis-regulatory codes (de Boer & Taipale, 2024). Although the code that translates codons to folded proteins is nearly universal (left), the cis-regulatory code is both species and cell-type specific (right).

Initial machine learning models thus represented cis-regulatory sequences based on features such as TFBS, k-mers, in vivo protein binding and epigenetic marks to predict enhancer activity or gene expression (Seyres et al., 2016) (Lee et al., 2015). However, recent developments in deep learning, and, in particular, convolutional neural networks, do not require previous biological knowledge and can learn accurate models directly from raw data (i.e. DNA sequences). Most importantly, once trained on raw data, these models allow the extraction and interpretation of the learned rules. These interpretable rules can then be used to decode the regulatory “syntax” or “grammar”, providing detailed information about the arrangement of TFBS, including their number, order, orientation and spacing. For instance, recent studies have used deep learning to predict TF binding in vivo (Avsec, Weilert, et al., 2021), chromatin accessibility (Avsec, Agarwal, et al., 2021), transcriptional reporter activity (Figure 1.5b) (de Almeida et al., 2022) (Sahu et al., 2022) and effect of genetic variants on gene expression (Lu et al., 2022). These studies showcase how large-scale functional measurements can be used in combination with the power of deep learning models to define regulatory syntax.

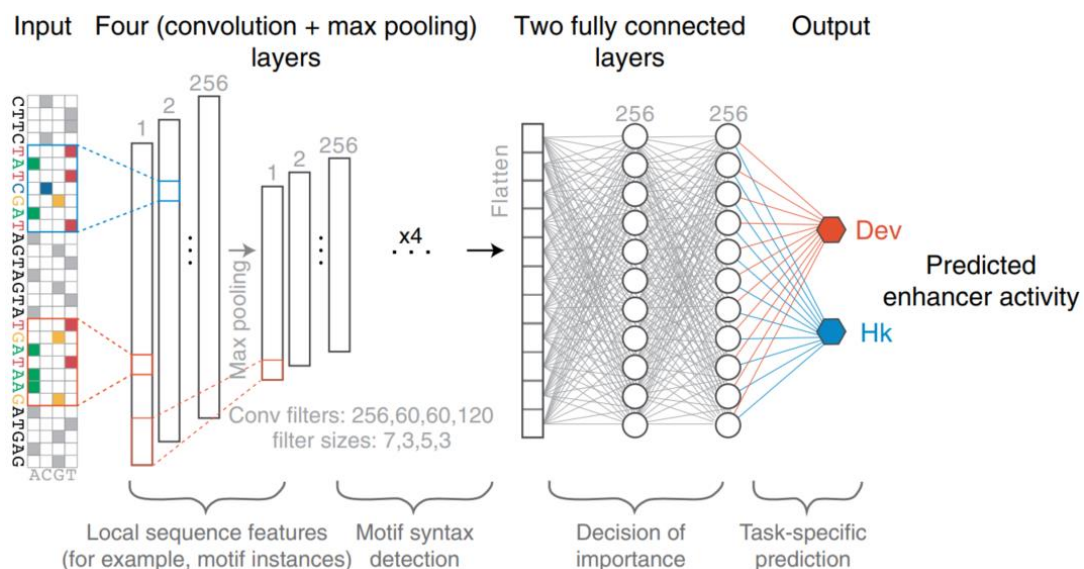


Figure 1.5b An example of deep learning model quantitatively predicts enhancer activity genome wide from DNA sequence (de Almeida et al., 2022). This is the architecture of the multitask convolutional neural network DeepSTARR that was

trained to simultaneously predict quantitative Dev and Hk enhancer activities from 249-bp DNA sequences.

In addition to the algorithm model itself, establishing a gold standard data set and standardized data processing are necessary. Systematic evaluation and testing of model performance can help optimize algorithm strategies. For example, the random promoter DREAM Challenge presented a unique opportunity for participants to propose novel model architectures and training strategies for modeling regulatory sequences (Rafi et al. 2023). It is foreseeable that there will be more and more such algorithm competitions and benchmarks in the coming years.

# Chapter 2. Impact of regulatory elements on disease

ENCODE results have shown that the number and proportion of regulatory elements in the genome far exceed that of gene coding regions. A large number of genome-wide association study (GWAS) results also show that most phenotype-related genetic variations in the population are located in non-coding regions. Therefore, a considerable proportion of research on human diseases has shifted to regulatory elements in non-coding regions, leading to the question: Which diseases are associated with the variations of regulatory elements? What are the roles and mechanisms of regulatory elements in disease? How do identify the causal regulatory variants?

## 2.1 Genomic variation in the human genome

### 2.1.1 Different types of genomic variation

Before discussing human genetic diseases, what types of genetic variation should be known in the human genome? And how many? According to the size and alignment structure of the variants, they are divided into Single Nucleotide Polymorphisms (SNPs), insertions, deletions, and repeats, as well as chromosome structural variations and number variations (Figure 2.1.1). Chromosome structural variation is a change in a large segment of the genome, which can be divided into deletion, duplication, inversion, insertion, and translocation (Figure 2.1.1). For an individual, the genomic variations in the body are divided into germline variants and somatic variants. Germline variants are inherited from parents and are usually heritable. Somatic variants only exist in somatic cells and are not heritable. The heritable variants are often used in the study of genetic diseases. The 1000 Genomes Project sequenced the genomes of 2,504 individuals from 26 populations, including 84.7 million SNPs, 3.6 million short insertion/deletions (indels), and 60,000 structural variants (Genomes Project et al., 2015). With the increase in large-scale population sequencing and the improvement of

genomes (Nurk et al., 2022), the number of variants will keep increasing (Liao et al., 2023).

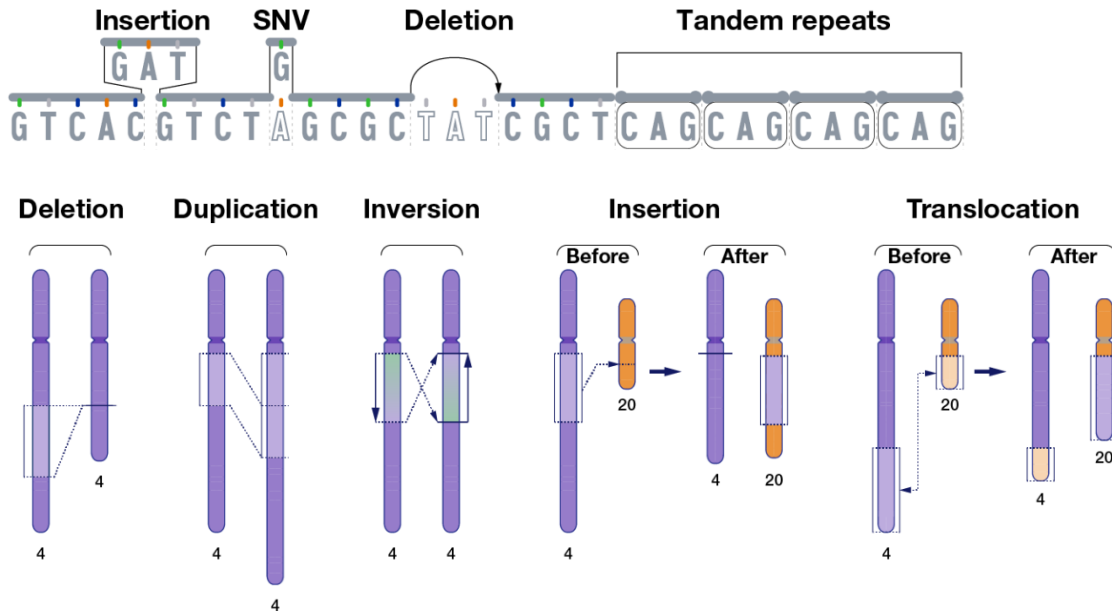


Figure 2.1.1 Different genomic variants and larger chromosome structural variation.

Adopted from NHGRI website: <https://www.genome.gov>

## 2.1.2 common and rare SNPs

In population level, SNPs can be divided into common and rare SNPs according to the minor allele frequency (MAF). Normally the common SNPs were defined as MAF more than 1% and rare SNPs were defined as MAF less than 1%. The variants usually captured in GWAS studies are common SNPs. Due to natural selection, most common SNPs have low genetic effects on diseases, and only a few examples show high effects (Figure 2.1.2). Usually, rare SNPs exhibit higher genetic effects, especially in Mendelian diseases. Therefore, whole genome sequencing of individuals with specific phenotypes and individuals with familial disease will be ideal models for detecting rare variants in rare diseases.

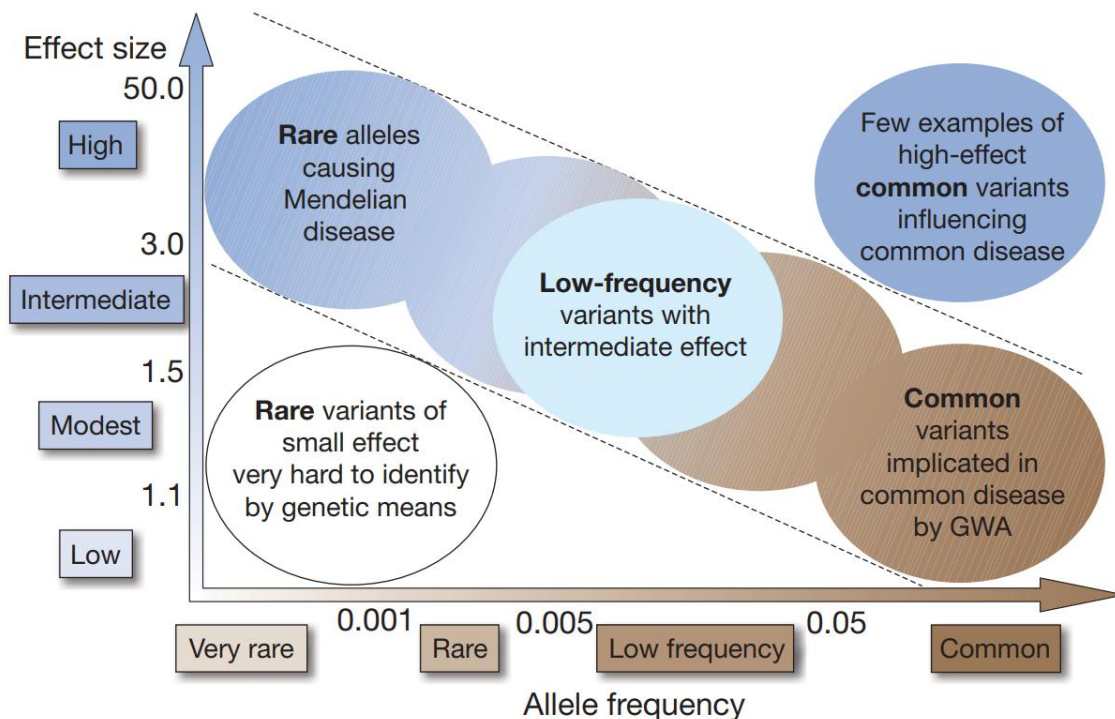


Figure 2.1.2 Feasibility of identifying genetic variants by risk allele frequency and strength of genetic effect (Manolio et al., 2009). Most emphasis and interest lie in identifying associations with characteristics shown within diagonal dotted lines.

## 2.2 Regulatory element variation in diseases

In regulatory elements (enhancers, promoters, insulators, silencers), enhancers have become the main research object due to their large number and wide distribution. There are many examples of diseases caused by changes in regulatory elements (Table 2.2), which can be into two main categories: regulatory function change caused by sequence variation, and regulation target change caused by location variation.

There are three main situations of functional variation of regulatory elements: loss, gain, and switch of regulatory element functions (Figure 2.2a). For example, enhancer deletions in  $\beta$ -globin genes lead to  $\beta$ -thalassemia (Kioussis et al., 1983). In the T-ALL (T-cell acute lymphoblastic leukemia), somatic insertions introduced a MYB binding site and induced the formation of a Neo-enhancer, which activated the oncogene TAL1's expression (Mansour et al., 2014). In obesity, multiple variants on a common



haplotype increase the activity of several enhancers (Smemo et al., 2014). In type 2 diabetes, SNP disrupts the binding of NeuroD1 and decreases enhancer activity to affect ZFAND3 gene expression (Pasquali et al., 2014). Bozhilov et al. showed that a single base change in the human  $\alpha$ -globin cluster creates a new promoter. This promoter acts as an orientation-dependent enhancer blocker, downregulating  $\alpha$ -globin expression, leading to  $\alpha$ -thalassemia (Bozhilov et al., 2021).

#### Enhancer disruption

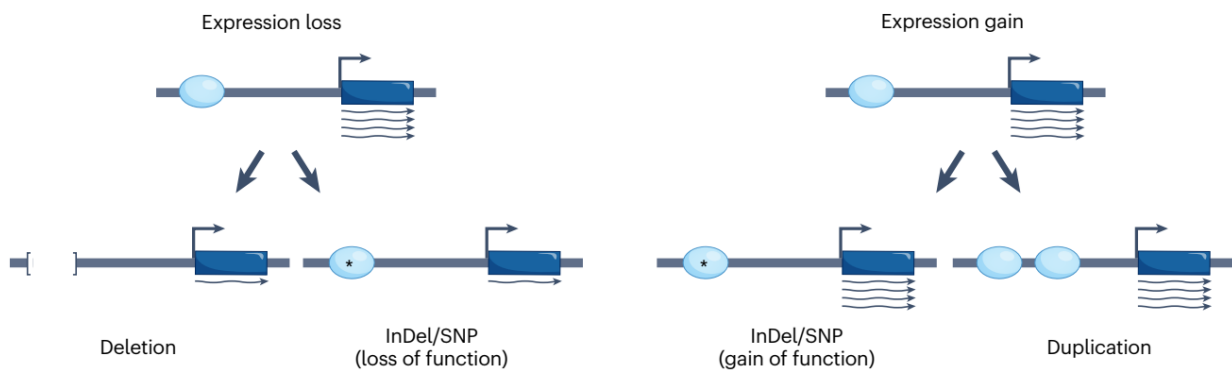


Figure 2.2a Function variation of enhancers (Zaugg et al., 2022). Enhancer deletion caused by deletion or indel may result in loss of distal gene expression. Enhancer gain caused by Indel/SNP or duplication may lead to activation of distal genes.

In addition to variations in the regulatory elements, changes in the connection between the regulatory elements and the target gene are also important. This is mostly caused by changes in chromatin structure, resulting in the loss of connections with original target genes and the establishment of connections with new target genes (Figure 2.2b). For example, Lupiáñez et al. found disruptions of TADs in limb malformations lead to de novo enhancer-promoter interactions and misexpression (Lupianez et al., 2015). Gröschel et al. found in leukemia structural rearrangements involving the chromosomal repositioning of a single enhancer can cause deregulation of two unrelated distal genes (Groschel et al., 2014). Wang et al. identified tens of oncogenes associated with enhancer hijacking in Hi-C data from 50 cancer cell lines and primary tumors (Wang et al., 2021).

Altered enhancer-gene connectivity

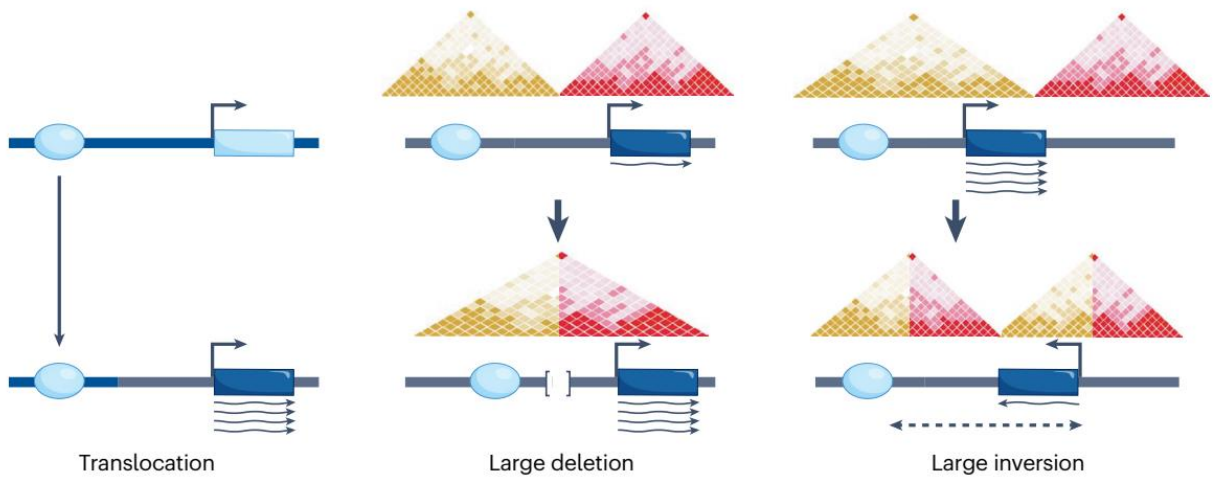


Figure 2.2b Enhancer-gene connectivity altered by chromatin structural variation (Zaugg et al., 2022). Large structural variations can distort or merge topologically associating domains (TADs). As a consequence, enhancer-gene connectivity can be lost or gained, resulting in dysregulated gene expression.

Type of disease	Disease	Affected gene(s)	Enhancer <sup>a</sup>	Type of disruption	Effect on gene expression	Ref.
Monogenic (Mendelian)	β-thalassemia	β-globin genes	LCR	Enhancer deletions	LOE	103,104
	α-thalassemia	α-globin genes	α-globin enhancers	Deletion or insertion of promoters alter enhancer-gene connectivity	LOE and GOE	105,106
	PDD2	SHH	ZRS	Rare variant introducing a TFBS	GOE	107
	HPE	SHH	SBE2	Rare variant disrupting a TFBS	LOE	108
	Limb malformations	PAX3, IHH, WNT6	EPH4 enhancers	Deletions, duplications and inversions disrupt the boundaries of a TAD containing the EPH4 enhancer and rewire the connectivity with different genes	GOE	13
	5q14.3 microdeletion syndrome	MEF2C	MEF2C enhancers	TAD disruption disconnects MEF2C from the associated enhancer	LOE	109
	Pierre Robin syndrome	SOX9	SOX9 enhancer	A point mutation in a conserved enhancer disrupts the binding of MSX1	LOE	74
	Cooks syndrome	SOX9, KCNJ2	SOX9 enhancers	Duplication of a TAD boundary at the SOX9 locus causes neo-TAD formation and KCNJ2 misexpression	GOE	110
	Isolated atrial defect	TBX5	90kb downstream	Rare variant abrogates heart-specific enhancer activity	LOE	111
Isolated pancreatic agenesis	PTF1A	25kb downstream	Rare variants abolish enhancer activity and disrupt the binding of FOXA2 and PDX1	LOE	41	
Common (multifactorial)	Obesity	IRX3, IRX5	FTO intronic	Multiple variants on a common haplotype increase the activity of several enhancers	GOE	97
	Type 2 diabetes	ZFAND3	Upstream	SNP disrupts the binding of NeuroD1 and decreases enhancer activity	LOE	112
	Vascular diseases	EDN1	PHACTR1 intronic	SNP located in a distal region interacting with the EDN1 enhancer	LOE	96
	HBF level	BCL11A	Downstream	SNP disrupts TF binding and diminishes expression in erythroid cells	LOE	113
	Cardiac disorders	SNC5A	SNC10A intronic	SNP in SNC10A modulates SNC5A expression in the heart	LOE	114
	Hirschsprung disease	RET	Several enhancers	Several SNPs located in RET enhancers act synergistically to reduce gene expression	LOE	115
	Parkinson	SNCA	Intronic	SNP alters the binding of EMX2 and NKX6-1	LOE	40
	Cancer	Burkitt lymphoma	MYC	IgH enhancer	Somatic translocation (enhancer hijacking)	GOE
Lung adenocarcinoma		MYC	450kb downstream	Somatic duplication of the enhancer	GOE	118
T-ALL		TAL1	7kb upstream	Somatic insertions introduce a MYB binding site and induce the formation of a Neo-enhancer	GOE	119,120
Ph-like ALL		GATA3	Intronic	A rare variant increases enhancer activity	GOE	121
CLL		AXIN2	Upstream	Common variation in the AXIN2 enhancer modulates CLL susceptibility via differential MEF2 binding	GOE	122
AML		GATA2, EVI1	GATA2 enhancer	Large somatic inversion relocates the GATA2 enhancer in the vicinity of EVI1	LOE and GOE	14
Prostate cancer		PCAT19, CEACAM21	PCAT19 Epromoter	Common variant changes the affinity of TFs and switches promoter and enhancer activities	GOE	123,124

<sup>a</sup>Enhancer location relates to the regulated gene unless otherwise stated. LOE, loss of expression; GOE, gain of expression; LCR, locus control region; PDD2, preaxial polydactyly type II; HPE, holoprosencephaly; HBF, fetal hemoglobin; T-ALL, T cell acute lymphoblastic leukemia; CLL, chronic lymphoblastic leukemia; AML, acute myeloid leukemia; Ph-like ALL, Philadelphia chromosome-like acute lymphoblastic leukemia; TF, transcription factor; TFBS, TF-binding site; IgH, immunoglobulin heavy chain.

Table 2.2 Representative examples of enhancer dysfunction driving disease. (Zaugg et al., 2022)

## **2.3 Link genomic variations with diseases**

### **2.3.1 Genome-wide association studies**

Genome-wide association studies (GWAS) have been widely used to detect disease-associated variants across the whole genome. GWAS are designed to determine genotype-phenotype associations by testing differences in allele frequencies of genetic variants among individuals with different phenotypes. The experimental workflow of a GWAS involves several steps (Figure 2.3.1), including the collection of DNA and phenotypic information from a group of individuals; genotyping of each individual using available GWAS arrays or sequencing strategies; quality control; imputation of untyped variants using haplotype phasing and reference populations; conducting the statistical test for association; conducting a meta-analysis; seeking an independent replication; and interpreting the results by conducting multiple post-GWAS analyses. As the cost of GWAS decreases, more than 5,000 GWAS studies have been applied to more than 3,000 traits (Uffelmann et al., 2021). Many disease studies have made significant progress through GWAS, such as type 2 diabetes, auto-immune diseases, and schizophrenia (Visscher et al., 2017).

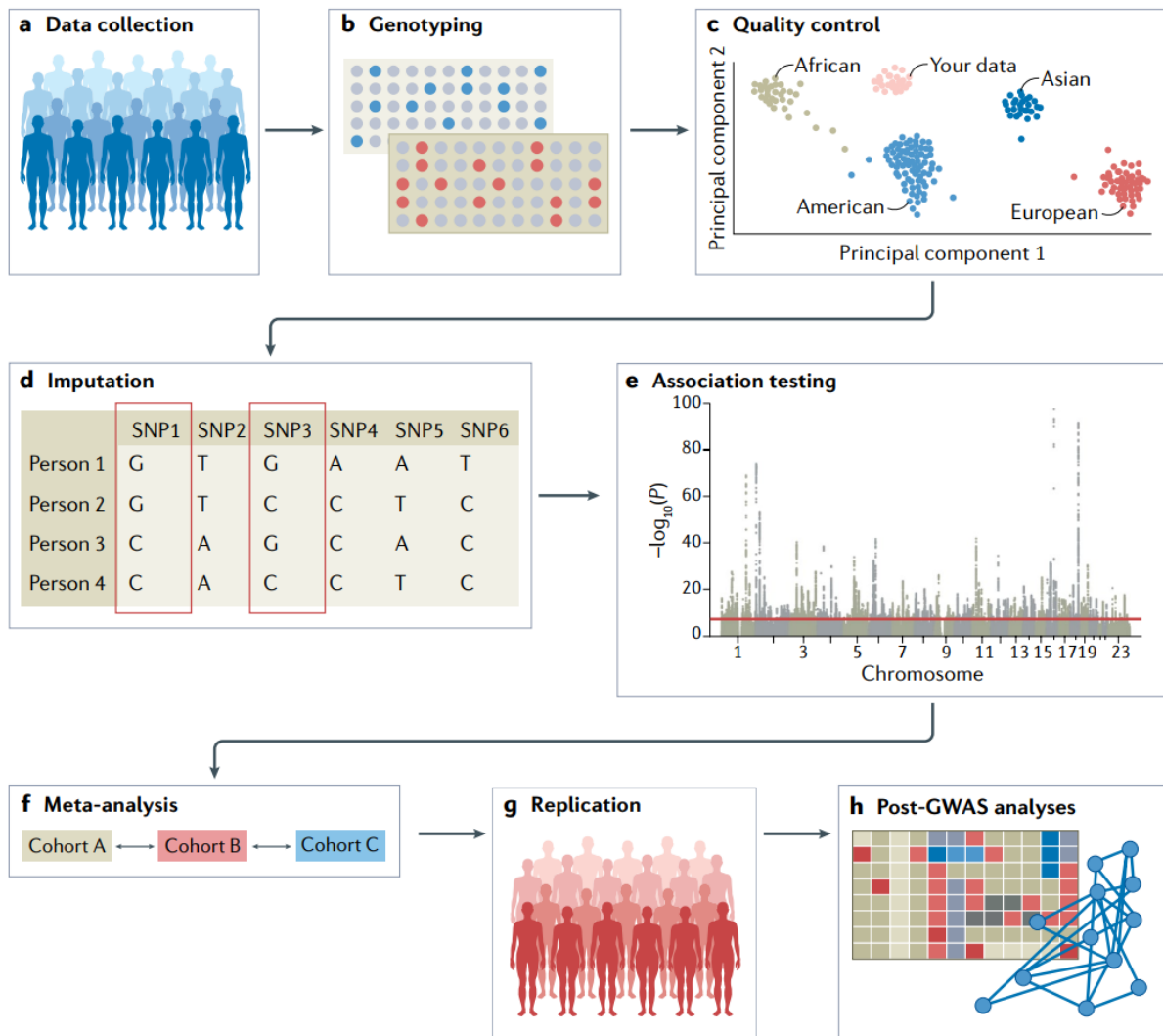


Figure 2.3.1 Overview of steps for conducting GWAS (Uffelmann et al., 2021).

Significant SNPs in GWAS are derived from predetermined tag SNPs in the microarray. However, due to the existence of linkage disequilibrium, the significance of these tag SNPs only represents correlation with phenotype rather than causality. To determine which SNPs are causal SNPs for the phenotype, one method is experimental verification, and the other method is fine-mapping. Fine-mapping is a computational process designed to prioritize variants that are most likely to be causally related to the phenotype of a GWAS based on linkage disequilibrium patterns and association statistics (Schaid et al., 2018).

## 2.3.2 Polygenic risk score

GWAS results indicate that genetic loci affecting complex diseases are often widely distributed in the genome. How to evaluate individual disease risk based on GWAS results? A polygenic risk score (PRS) is an estimate of an individual's genetic liability to a trait or disease. It is calculated by taking the sum of risk alleles for an individual and weighting the risk allele effect size based on the GWAS (Figure 2.3.2). As GWAS sample sizes increase, polygenic scores are likely to play a central role in the future of biomedical research and personalized medicine (S. W. Choi et al., 2020).

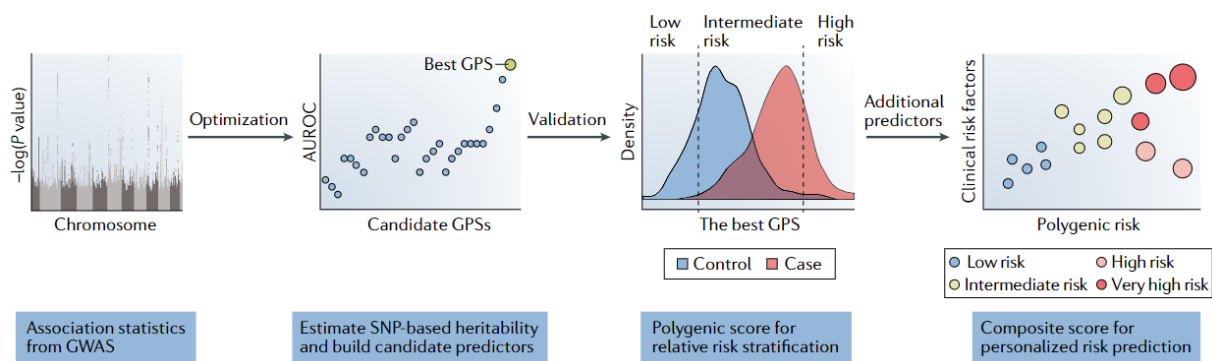


Figure 2.3.2 A genome-wide polygenic risk score (GPS) is based on genome-wide association study (GWAS) summary statistics (Liu & Kiryluk, 2018). The optimization step enables selection of the best method according to the genetic architecture of a disease under study. The validation step requires an external cohort and is critical to obtaining reliable metrics of performance. Clinical predictors of absolute risk will require incorporation of additional demographic, clinical or lifestyle factors into composite risk models. AUROC, area under receiver operating characteristic.

## 2.3.3 Whole-genome sequencing

As the cost of sequencing decreases, exome sequencing and whole-genome sequencing (WGS) are also used to detect disease variants. For example, the 1000 Genomes Project, which has sequenced 2,504 people from different populations around the world, and the UK Biobank (Halldorsson et al., 2022), which sequenced the genomes of 150,119 British people, provided a rich resource of human genetic variation. The whole-genome sequencing can comprehensively detect various types

of genetic variation, including common and rare variants, insertion, deletion, CNVs, and chromatin structural variations.

## 2.4 GWAS SNPs enriched in regulatory elements

GWAS studies have shown that most common genetic variants fall in regulatory regions. For example, Maurano et al. found the majority (~93%) of disease- and trait-associated variants lie within noncoding sequences according to hundreds of GWAS studies (Figure 2.4). Many evidences suggested the involvement of a proportion of such variants in transcriptional regulatory mechanisms, including modulation of promoter and enhancer elements (Figure 2.4). Andersson et al. used the FANTOM5 panel of samples, covering the majority of human tissues and cell types, to produce an atlas of active, in vivo-transcribed enhancers. Their results showed that disease-associated SNPs were over-represented in regulatory regions to a greater extent than in exons.

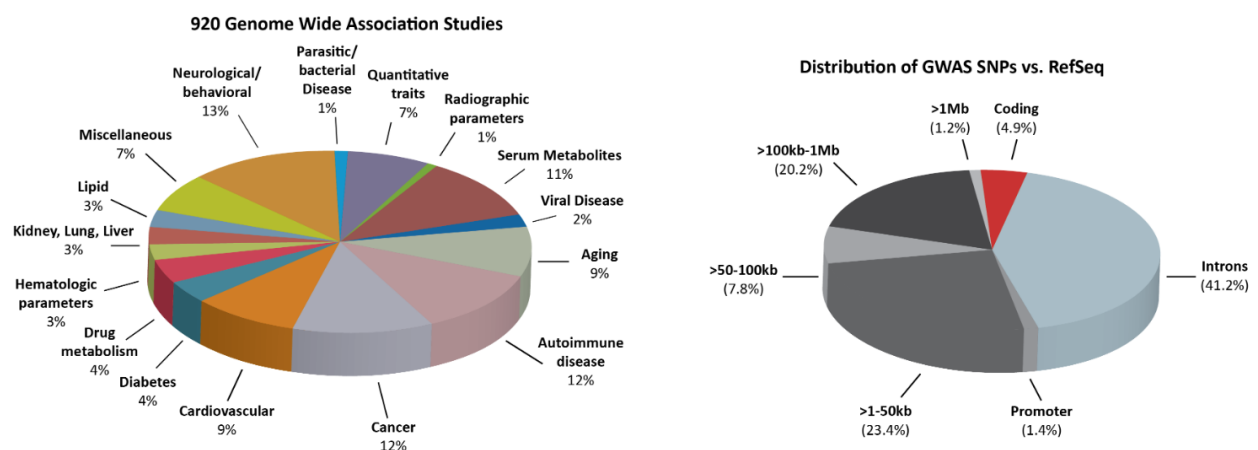


Figure 2.4 Diseases and traits studied by GWAS and distribution of GWAS variants (Maurano et al., 2012). The left chart shows the percentage of GWAS SNPs by disease/trait class. The right chart shows the location of GWAS SNPs relative to genic features.

## 2.5 Complex effect of regulatory element variation in diseases

### 2.5.1 Redundancy

One of the complex effects of regulatory elements is redundancy. For example, multiple enhancers regulate the same gene. This phenomenon was also called shadow enhancers, which was proposed by Mike Levine and colleagues in 2008 (Hong et al., 2008). In this classic study, the shadow enhancers produce gene expression patterns that overlap those produced by the primary enhancers in dorsal-ventral patterning of the *Drosophila* embryo (Hong et al., 2008). Frankel et al. also tested this conception by generating a deficiency that removes two shadow enhancers of *shavenbaby* gene in the *Drosophila* embryo, which lead to extensive loss of trichomes in extreme temperatures (Frankel et al., 2010). In mammals, enhancers are remarkably abundant, which has led to many studies pointing to shadow enhancers. For example, the mouse limb development gene *Gli3* and the eye development gene *Pax6* are regulated by shadow enhancers (Figure 2.5.1) (Kvon et al., 2021). Shadow enhancers might provide an important mechanism for buffering gene expression against mutations in non-coding regulatory regions of genes implicated in human disease.

But how do multiple enhancers work together? Are they additive or synergistic? Thomas et al. found that multiple weak enhancers strongly induce endogenous target gene expression (Thomas et al., 2021). This shows the additivity between multiple enhancers. But there are also studies showing that introducing a weak enhancer between a strong enhancer and the promoter strongly increases reporter gene expression, which displayed synergistic of enhancers (Thomas et al. 2023). We still need more evidence.



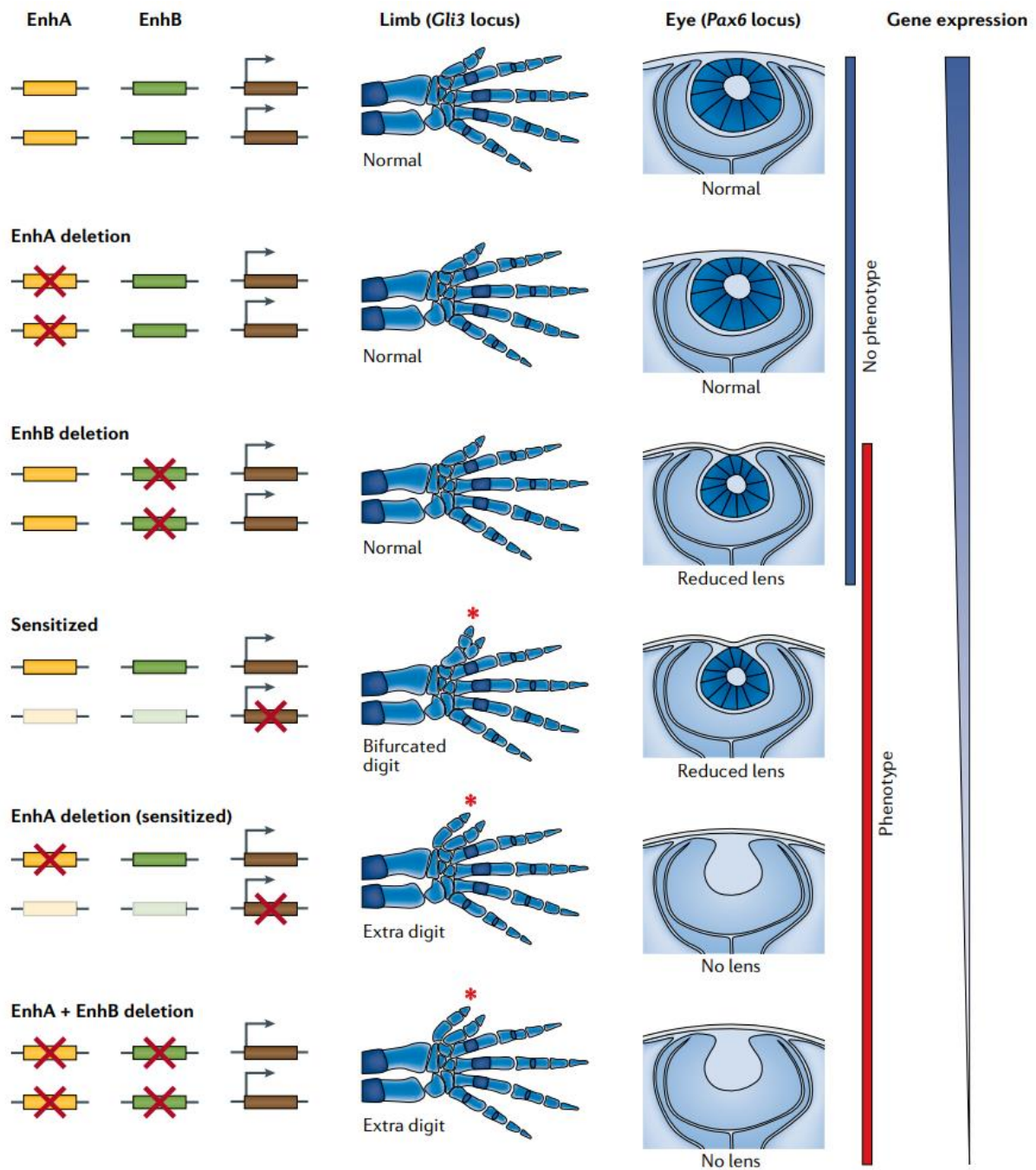


Figure 2.5.1 Shadow enhancers confer phenotypic robustness in mammals (Kvon et al., 2021). The figure shows shadow enhancer deletions in *GLI3* and *Pax6* locus yield no observable phenotypes in mice. *GLI3* is critical for proper limb development, and knockout of the encoding gene causes the formation of extra digits. *Pax6*-deficient mice have arrested eye development and no lens formation.

## 2.5.2 Pleiotropy

Pleiotropy was initially defined as the phenomenon whereby a single gene independently affects two or more phenotypes (Mackay & Anholt, 2024). Mackay et al. proposed different types of pleiotropy (Figure 2.5.2): pleiotropy indicates shared genetic architecture affecting the traits and can occur when a polymorphism independently affects more than one trait ('horizontal' pleiotropy) or when a polymorphism affects one trait which, in turn, affects another ('mediating' pleiotropy).

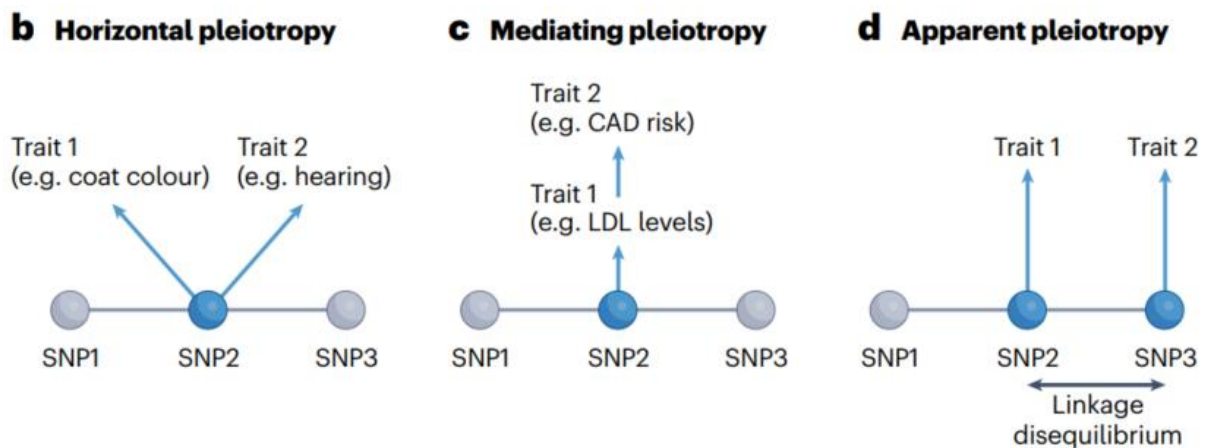


Figure 2.5.2 Different types of pleiotropy (Mackay & Anholt, 2024). Horizontal pleiotropy: the same SNP independently affects two (or more) quantitative traits. Mediating pleiotropy: a SNP affects one trait, which in turn affects a second trait. Apparent pleiotropy: when the traits are not caused by a common SNP but, rather, by two (or more) SNPs that are in linkage disequilibrium and each affects one of the traits.

Much evidence has been supported the pleiotropic effect of regulatory elements. Chromatin interactions and CRISPRi data have shown that a single enhancer can regulate multiple genes (Uyehara & Apostolou, 2023) (Fulco et al., 2016). Laiker et al. analyzed active enhancers across human organs based on the analysis of both eRNA transcription (FANTOM5 consortium data sets) and chromatin architecture (ENCODE consortium data sets) and found that more than 40% of enhancers in the human genome are pleiotropic (active in different organs) (Laiker & Frankel, 2022). Watanabe et al. analyzed 4,155 GWAS and found that 90% of loci were associated with multiple traits (Watanabe et al., 2019). And most of these GWAS-associated variants are located in non-coding regions, which are mostly regulatory elements. Therefore, the

pleiotropy of regulatory elements means the temporal pleiotropy in tissue development and the pleiotropy of multiple regulatory functions, which in turn leads to the pleiotropy of pathological traits. For example, some promoters were found that can function as enhancers to regulate multiple distal genes, which could be associated with multiple phenotypes or diseases (Malfait et al., 2023) (see ANNEX 1).

# Chapter 3. Strategies to study the impact of genetic variation in cis-regulatory elements

Studies examining the impact of genetic variations within regulatory elements are generally categorized into two distinct approaches: one focuses on the effect of sequence changes within the regulatory elements themselves on regulatory activity, and the other investigates the impact of sequence variations in regulatory elements on gene expression.

## 3.1 Evaluation of regulatory sequence variation

The tools for studying the impact of regulatory sequence variations mainly include genome editing-based techniques such as CRISPR/Cas9, base editing, prime editing; reporter gene-based methods like luciferase reporter assay and MPRA; transcription factor binding effect evaluation tools like SNP-SELEX and computational tools.

### 3.1.1 Genome editing

As previously mentioned, CRISPR/Cas9 primarily investigates by observing phenotypic changes or gene expression variations before and after deletions in regulatory regions. Base editing (Figure 3.1.1) is a novel genome editing strategy that precisely alters DNA bases without double-stranded breaks, minimizing off-target effects and enabling targeted gene correction for therapeutic and research applications (Komor et al., 2016). But base editing has a limited range of targetable base pairs and depends on the PAM sequence recognized by the CRISPR system. While reduced compared to CRISPR-Cas9, off-target edits can still occur, necessitating thorough off-target analysis (Rees & Liu, 2018).

Prime editing (Figure 3.1.1) offers unprecedented precision and flexibility for rewriting genetic sequences, enabling the introduction of insertions, deletions, and all types of base-to-base conversions without requiring double-stranded DNA breaks (DSBs) or donor DNA templates (Anzalone et al., 2019). Approximately 90% of human

pathogenic genetic variants are single-base mutations or insertions and deletions of fewer than a dozen base pairs, which are types of DNA change that are well within the capabilities of prime editing systems (Chen & Liu, 2023).

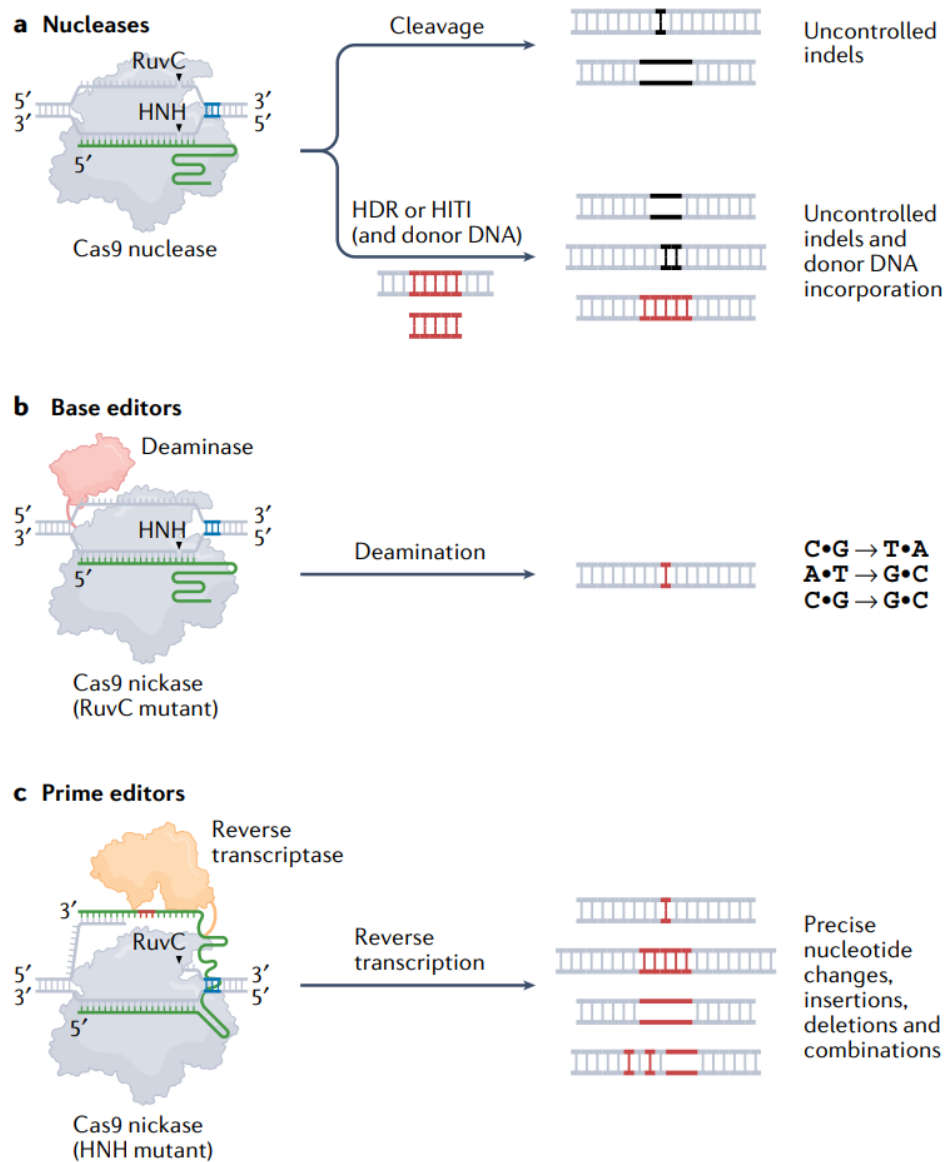


Figure 3.1.1 Three types of genome editing tools (Chen & Liu, 2023).

### 3.1.2 Reporter assay

Luciferase reporter assay is a commonly used reporter assay technology that assesses the function of regulatory elements such as promoters and enhancers by measuring the activity of a reporter gene (usually the luciferase gene) (Naylor, 1999). However, Luciferase reporter assay can usually only test one or a few regulatory

sequences at a time and is not suitable for large-scale screening. When quantitatively measuring reporter gene activity, differences in background signal, cell status, and transfection efficiency may affect the accuracy and reproducibility of the results. Therefore, Luciferase Reporter Assay is suitable as a low-cost method for functional verification of known sites.

MPRA (Massively Parallel Reporter Assay) is a high-throughput technique that can simultaneously test the impact of thousands to millions of DNA sequences on gene expression (as introduced in Chapter 1). Many studies have utilized MPRA to study how specific genetic variations affect the activity of regulatory elements. For example, Tewhey et al. used MPRA to evaluate of 32,373 variants associated with eQTLs in lymphoblastoid cell lines and find 842 variants showing differential gene expression between alleles (Tewhey et al., 2016b). Arensbergen et al. leveraged SuRE to survey the effect of 5.9 million SNPs on enhancer and promoter activity (van Arensbergen et al., 2019). They identified more than 30,000 SNPs that alter the activity of putative regulatory elements, partially in a cell-type-specific manner. Abell et al. applied MPRA to functionally evaluate genetic variants in high, local LD for independent cis-expression quantitative trait loci (eQTL) (Abell et al., 2022). They found that 17.7% of eQTLs exhibit more than one major allelic effect in tight LD.

To better access these MPRA resources, I provide a summary table which include 24 published MPRA studies and 37829 allelic impact SNPs (see supplemental data of results part, Chapter 6). Recently, Zhao et al. developed a database called MPRAbase (<http://www.mprabase.com>), which provides regulatory scores associated with sequences by re-processing all the published MPRA data (10.1101/2023.11.19.567742). This database will be a powerful resource for developing machine learning models to predict regulatory activity.

### **3.1.3 Evaluation of transcription factor binding effect**

Some computational tools have been developed to predict the impact of genomic variation on transcription factor binding. The first strategy utilizes position weight matrices (PWMs) or transcription factor flexible models (TFFMs) of motifs to predict the impact of SNPs on transcription factor binding, for example SNP2TFBS (Kumar et al., 2017), RSAT variation-scan (Santana-Garcia et al., 2019), FABIAN-variant

(Steinhaus et al., 2022). Another strategy evaluates transcription factor binding effect based on ChIP-seq or other epigenomic data, like ANANASTRA (Boytssov et al., 2022) and RegulomeDB (Dong et al., 2023). Since the prediction of transcription factor binding impact is variable between different strategies, tools like MEME Suite (Bailey et al., 2015) and TFmotifView (Leporcq et al., 2020) to view the motifs in some specific regions will be more intuitive. An application study also showed how to predict the functional effects of genetic variants by analyzing allelic variation in TF binding affinity in human lymphoblastoid cell lines (Joanna Mitchelmore et al., 2020).

Many models based on machine learning, especially deep learning, have also been designed for variant effect prediction, such as DeepBind (Alipanahi et al., 2015), DeepSEA (Zhou & Troyanskaya, 2015), BpNet (Avsec, Weilert, et al., 2021), etc. Recently, Han et al. evaluated the performance of 14 computational models that can predict the effects of non-coding variants on TF binding using large-scale in vitro (i.e., SNP-SELEX) and in vivo (i.e., allele-specific binding, ASB) TF binding data (Han et al., 2024). Their evaluation results showed that: for in vitro variant impact prediction, kmer/gkm-based machine learning methods (deltaSVM (Yan et al., 2021), QBiC-Pred (Martin et al., 2019)) trained on in vitro datasets performed the best; for in vivo variant impact prediction, DNN-based multitask models (DeepSEA (Zhou & Troyanskaya, 2015), Sei (Chen et al., 2022), Enformer (Avsec, Agarwal, et al., 2021)) trained on the ChIP-seq datasets exhibited the best performance.

SNP-SELEX is a high-throughput testing system designed to evaluate the impact of genetic variations on transcription factor binding (Figure 3.1.3). It is based on the SELEX technique (Systematic Evolution of Ligands by Exponential enrichment), a method for selecting nucleic acid molecules from a large random sequence library that bind with high affinity and specificity to specific targets such as proteins, small molecules, cell surface markers (Jolma et al., 2013). Through SNP-SELEX, Yan et al. assessed the binding of 270 human transcription factors to 95,886 non-coding variants in the human genome (Yan et al., 2021). This provides a rich experimental resource for evaluating the sequence effects on regulatory elements.

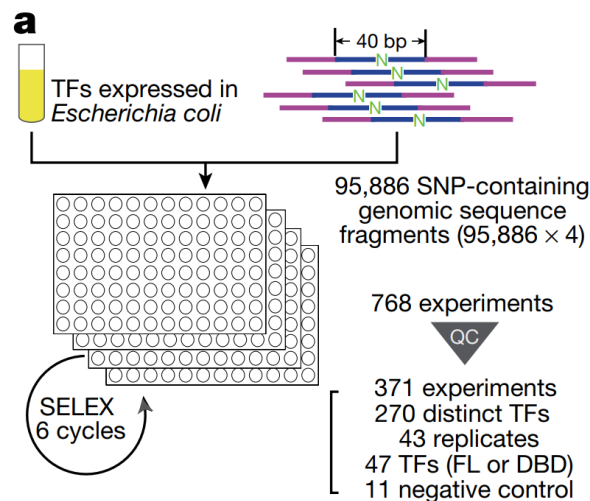


Figure 3.1.3 An overview of the SNP-SELEX experimental procedure (Yan et al., 2021).

## 3.2 Target gene identification strategies

Strategies for identifying the target genes of regulatory elements fall into two main categories: functional-based and structural-based approaches. The functional-based approach examines changes in gene expression triggered by variations in regulatory elements, directly revealing the effects of these elements on target genes. However, this approach faces challenges in distinguishing between the cis and trans effects. The structural-based approach focuses on mapping the physical connections between regulatory elements and distant genes via chromatin interactions, but not all chromatin interactions necessarily reflect biological function. Therefore, these two strategies can be complementary.

### 3.2.1 eQTL

Expression Quantitative Trait Loci (eQTLs) are regions in the genome where genetic variants have a statistically significant association with the levels of gene expression. The basic steps involved in generating eQTL data include (Figure 3.2.1a): Generation of gene expression data: Firstly, tissue samples are collected from various individuals, and the expression levels of thousands of genes within these samples are determined using high-throughput sequencing technologies (such as RNA-Seq) or microarray technologies. Genotype analysis: At the same time, a comprehensive genotyping analysis of these individuals is performed to identify genetic variants. Statistical association analysis: Statistical methods are utilized to analyze the association



between gene expression levels and individual genotypes to identify which genetic variations are related to changes in the expression levels of specific genes. These associated genetic loci are referred to as eQTLs.

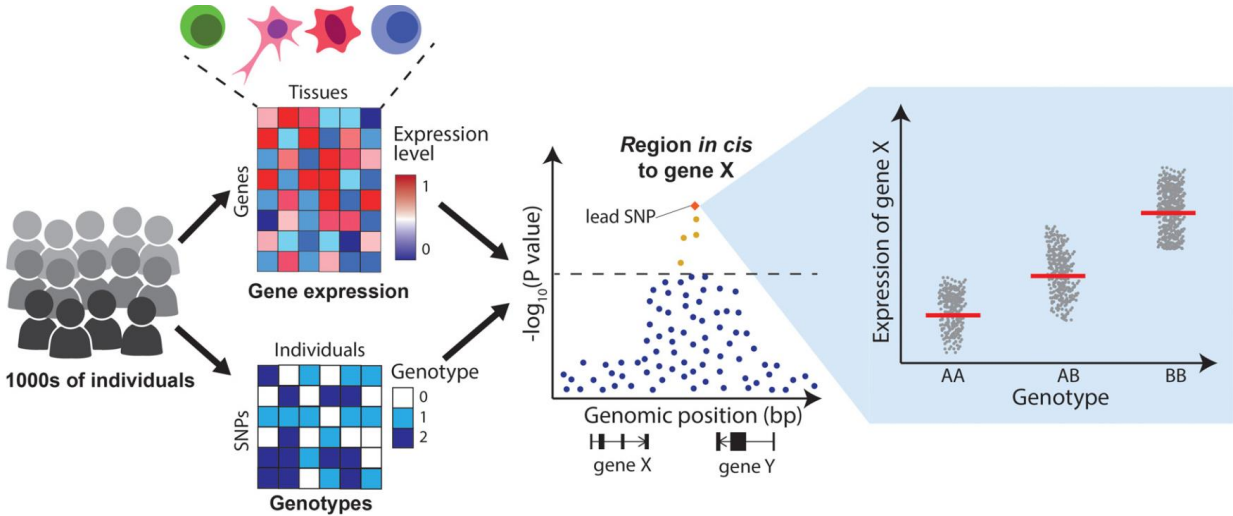


Figure 3.2.1a Overview of eQTL-mapping (Cano-Gamez & Trynka, 2020). In eQTL-mapping gene expression is profiled in thousands of individuals and the expression level of each gene is tested for association with genotypes at SNPs.

Based on the physical location relationship between genetic variations and the gene expression they affect, eQTLs can be divided into cis-eQTLs and trans-eQTLs (Liu et al., 2019) (Vosa et al., 2021) (Westra et al., 2013). Cis-eQTLs are typically located on the same chromosome as the gene expression they affect and are relatively close (usually within 1Mb), while trans-eQTLs may be located on different chromosomes or far away on the same chromosome (Figure 3.2.1b). It's important to distinguish between cis-eQTLs and trans-eQTLs while identifying target genes of regulatory elements using eQTLs.

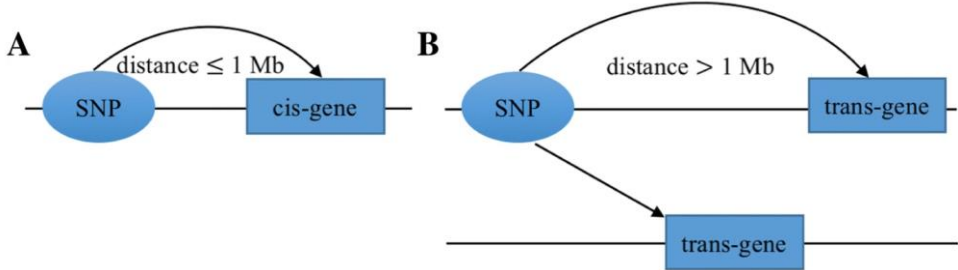


Figure 3.2.1b Scheme of cis-eQTL and trans-eQTL (Shan et al., 2019).

The strengths of eQTL analysis include providing a comprehensive strategy for identifying the target genes of regulatory elements on a large scale and elucidating the connection between an individual's genetic variations and gene expression within the full context of human genetics, offering advantages over genome editing in cell lines or animal models. The limitations are: eQTLs were focused on common genetic variants, most of which have a minor impact on gene expression; it's challenging to distinguish between the effects of different variations within the same LD block using eQTL analysis; eQTL analysis requires extensive genomic and transcriptomic sequencing across a large number of individuals, which can be cost-prohibitive.

### **3.2.2 High-throughput genetic perturbation**

Another functional-based approach to identifying target genes of regulatory elements is detecting gene expression changes by genetic perturbation. For example, Perturb-seq combines CRISPR-mediated gene perturbations with single-cell RNA sequencing (scRNA-seq) (Adamson et al., 2016; Dixit et al., 2016). Through CRISPR and its variant systems, researchers can systematically knock out, knock down, or activate hundreds or thousands of genes in a single experiment, and then use single-cell RNA sequencing to analyze the impact of these perturbations on single-cell gene expression patterns. Combining the CRISPR system with single-cell sequencing provides a powerful strategy for large-scale in vivo validation of regulatory element functions and target gene identification.

For instance, Fulco et al. assessed sequences spanning over 1 megabase around two essential transcription factors, MYC and GATA1, and identified nine distal enhancers controlling gene expression and cell proliferation (Fulco et al., 2016). Gasperini et al. performed CRISPRi perturbations on 5,920 human candidate enhancers, identifying 664 cis human enhancer-gene pairs, providing a large-scale framework for mapping enhancer-gene regulatory interactions (Gasperini et al., 2019). Replogle et al. performed genome-scale Perturb-seq targeting all expressed genes with CRISPR interference (CRISPRi) across millions of human cells (Figure 3.2.2a) (Replogle et al., 2022). This study provides the first genome-wide scale resource for transcriptional effects of genetic perturbations. Morris et al. combined biobank-scale GWASs,

massively parallel CRISPR screens, and single-cell sequencing to discover target genes of noncoding variants for blood trait loci with systematic targeting and inhibition of noncoding GWAS loci with single-cell sequencing (STING-seq) (Morris et al., 2023). This approach can identify target genes in cis and trans, measure dosage effects, and decipher gene-regulatory networks.

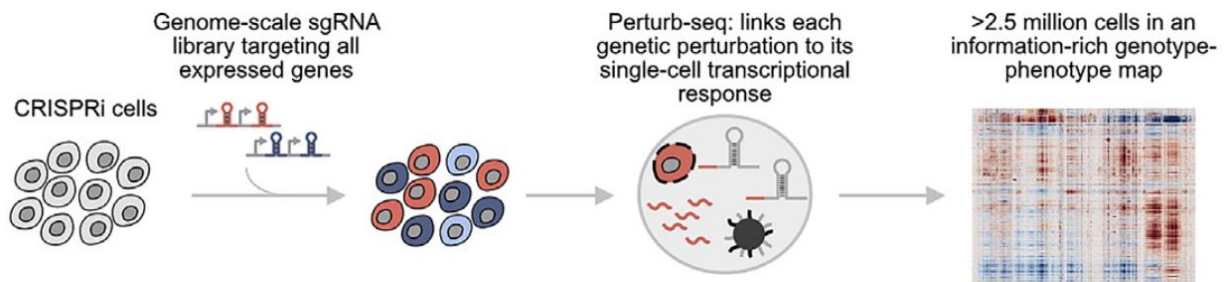


Figure 3.2.2a Genome-scale Perturb-seq constructs a comprehensive genotype-phenotype map (Replogle et al., 2022).

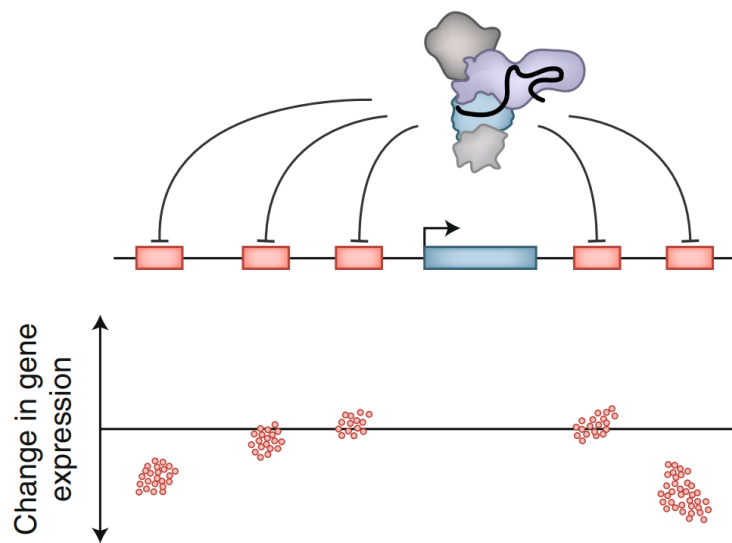


Figure 3.2.2b High-throughput mapping of enhancer-gene interactions with CRISPRi-FlowFISH (Fulco et al., 2019). Candidate regulatory elements are silenced with CRISPR interference, and the outcome is measured by quantifying the expression of an endogenous gene.

There are still some limitations in Perturb-seq: high cost, lack of efficiency for lowly expressed genes and small effects, and complex analysis. Consequently, some relatively low-cost alternatives have been developed. Targeted Perturb-seq (TAP-seq)

is a sensitive, inexpensive, and platform-independent method that focuses on single-cell RNA-seq coverage of genes of interest (Schraivogel et al., 2020). CRISPRi-FlowFISH (Figure 3.2.2b) combines CRISPR interference (CRISPRi) with flow cytometry-based fluorescence in situ hybridization (FlowFISH) technology (Fulco et al., 2019). Compared to Perturb-seq, which provides a comprehensive view of the impact on the entire transcriptome at the single-cell level, CRISPRi-FlowFISH focuses on the expression levels of specific target genes, making it more suitable for detailed investigation of the functions of specific genes or regulatory elements.

### **3.2.3 Promoter capture Hi-C**

Chromatin interaction can provide direct physical contact between regulatory elements and target genes. 3C and its related derivative techniques are widely used to detect chromatin interactions, such as 4C, Hi-C, ChIA-PET, HiChIP, and capture Hi-C. Among these, promoter capture Hi-C is used to capture chromatin interactions between promoters and distant genomic regions, and these contacts are enriched with enhancer-promoter interactions (Figure 3.2.3). For example, Javierre et al. used promoter capture Hi-C to identify interacting regions of 31,253 promoters in 17 human primary hematopoietic cell types (Javierre et al., 2016). They show that promoter interactions are highly cell type specific and enriched for links between active promoters and epigenetically marked enhancers. Jung et al. generated maps of long-range chromatin interactions centered on 18,943 well-annotated promoters for protein-coding genes in 27 human cell/tissue types (Jung et al., 2019). These large-scale promoter capture Hi-C datasets provide comprehensive resources for the connections between regulatory elements and target genes.

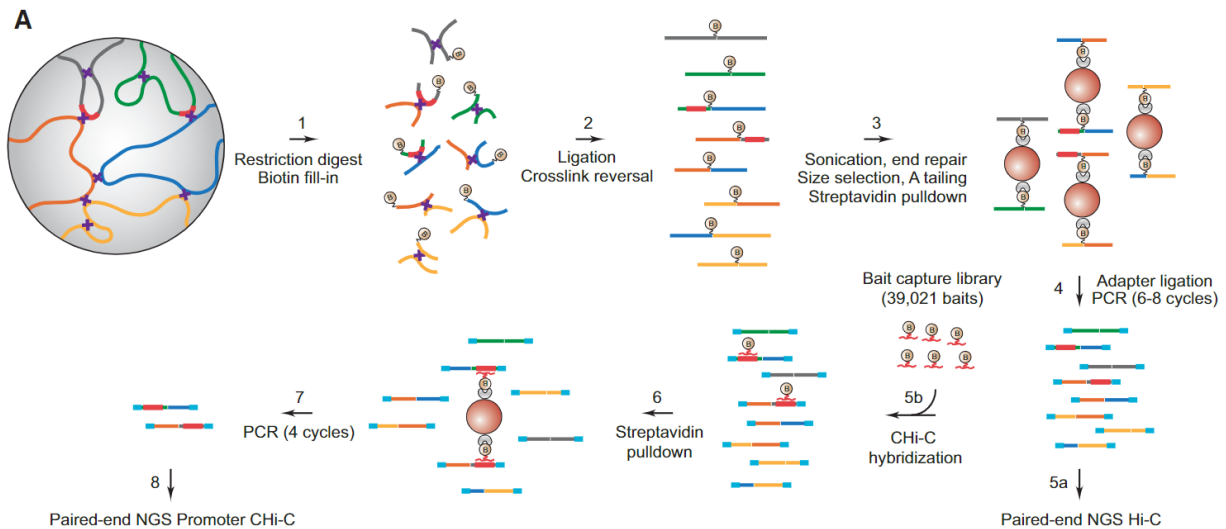


Figure 3.2.3 Promoter capture Hi-C experimental strategy (Schoenfelder et al., 2015).

### 3.2.4 ABC model

Fulco et al. developed the activity-by-contact (ABC) model to predict enhancer–gene connections (Figure 3.2.4) (Fulco et al., 2019). The ABC model is a simple yet powerful computational framework. In predicting distal element–gene connections within CRISPR datasets, the ABC model performs significantly better than other models. This model is based on the simple biochemical notion that an element’s quantitative effect on a gene should depend on its strength as an enhancer (Activity) weighted by how often it comes into 3D contact with the promoter of the gene (Contact). And the contribution of an element to a gene’s expression should depend on that element’s effect divided by the total effect of all elements.

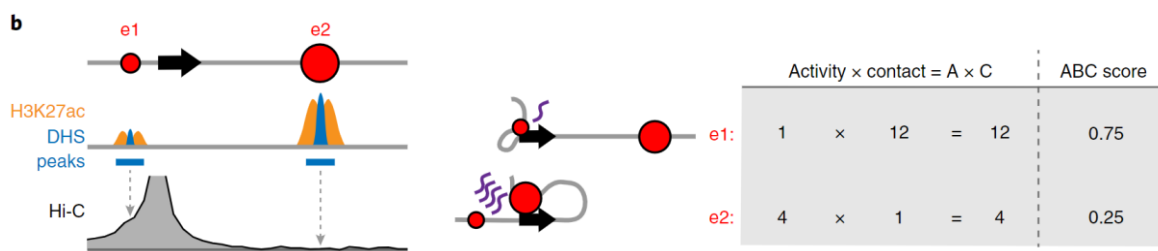


Figure 3.2.4 Calculation of the ABC score (Fulco et al., 2019). e1 and e2 (red circles) represent two arbitrary enhancers for the gene (black arrow).

Because it can make genome-wide predictions in a given cell type that are based on readily obtained epigenomic datasets, the ABC model provides a framework for

mapping enhancer–gene connections across many cell types. Nasser et al. applied the ABC model to create a genome-wide atlas of over six million enhancer-gene connections across 131 human cell types and tissues, using these atlases to interpret the functions of GWAS variants (Nasser et al., 2021). In 72 diseases and complex traits, the ABC model linked 5,036 GWAS signals to 2,249 unique genes.

### 3.3 Database and resources to study regulatory element variation

In the study of regulatory elements, it is important to know which databases and resources are available. Table 3.3 summarizes some commonly used databases and resources for investigating variations in regulatory elements.

Table 3.3 Database and resource to study regulatory element variation

Research purpose	Database/resource	Mainly collected data	links
Regulatory element	ENCODE	Comprehensive genome annotation	<a href="https://www.encodeproject.org">https://www.encodeproject.org</a>
Regulatory element	Roadmap	Histone modifications, DNA methylation	<a href="https://egg2.wustl.edu/roadmap/web_portal/index.html">https://egg2.wustl.edu/roadmap/web_portal/index.html</a>
Regulatory element	FANTOM5	CAGE, RNA-seq	<a href="https://fantom.gsc.riken.jp/5/data/">https://fantom.gsc.riken.jp/5/data/</a>
Regulatory element	EnhancerAtlas	enhancer annotation	<a href="http://www.enhanceratlas.org/index.php">http://www.enhanceratlas.org/index.php</a>
Regulatory element	VISTA Enhancer Browser	experimentally validated enhancers	<a href="https://enhancer.lbl.gov">https://enhancer.lbl.gov</a>
Human genetic variation	dbSNP	SNP data	<a href="https://www.ncbi.nlm.nih.gov/snp">https://www.ncbi.nlm.nih.gov/snp</a>
Human genetic variation	1000 Genomes Project	variant data	<a href="https://www.internationalgenome.org">https://www.internationalgenome.org</a>
Human genetic variation	GWAS Catalog	GWAS associations and summaries	<a href="http://www.ebi.ac.uk/gwas">www.ebi.ac.uk/gwas</a>
Human genetic variation	UK Biobank	Large scale biomedical database	<a href="https://www.ukbiobank.ac.uk">https://www.ukbiobank.ac.uk</a>
Human genetic variation	ClinVar	Variant annotations and clinical significance	<a href="https://www.ncbi.nlm.nih.gov/clinvar/">https://www.ncbi.nlm.nih.gov/clinvar/</a>
eQTL resource	GTEX	Tissue-specific gene expression and eQTLs	<a href="http://www.gtexportal.org">www.gtexportal.org</a>
eQTL resource	eQTL Catalogue	eQTL data from multiple studies	<a href="https://www.ebi.ac.uk/eqtl">https://www.ebi.ac.uk/eqtl</a>

Chromatin interactions	4DN Data Portal	4D nucleome data	<a href="https://data.4dnucleome.org">https://data.4dnucleome.org</a>
Chromatin interactions	WashU Epigenome Browser	Epigenomics and chromatin state data	<a href="https://epigenomegateway.wustl.edu/browser/">https://epigenomegateway.wustl.edu/browser/</a>
Chromatin interactions	3D Genome Browser	Hi-C	<a href="http://3dgenome.org/">http://3dgenome.org/</a>
TF binding	JASPAR	TF binding motif based on sequences	<a href="https://jaspar.elixir.no/">https://jaspar.elixir.no/</a>
TF binding	REMAP	TF binding peaks based on ChIP-seq	<a href="https://remap2022.univ-amu.fr/">https://remap2022.univ-amu.fr/</a>
TF binding	TFmotifView	visualization of transcription factor motifs	<a href="http://bardet.u-strasbg.fr/tfmotifview/">http://bardet.u-strasbg.fr/tfmotifview/</a>
TF binding effect	SNP2TFBS	TF binding effect prediction based on PWM	<a href="https://epd.expasy.org/snp2tfbs/">https://epd.expasy.org/snp2tfbs/</a>
TF binding effect	ANANAstra	TF binding effect prediction based on ChIP-seq	<a href="https://ananastra.autosome.org/">https://ananastra.autosome.org/</a>
TF binding effect	RSAT variation-scan	TF binding effect prediction based on PWM	<a href="http://rsat.sb-roscoff.fr/retrieve-variation-seq_form.cgi">http://rsat.sb-roscoff.fr/retrieve-variation-seq_form.cgi</a>
TF binding effect	FABIAN	TF binding effect prediction based on PWM and TFMM	<a href="https://www.genecascade.org/fabian/">https://www.genecascade.org/fabian/</a>
TF binding effect	RegulomeDB	Annotations of non-coding DNA variants	<a href="https://regulomedb.org/">https://regulomedb.org/</a>
MPRA resource	MPRAbase	Massive parallel reporter assays data	<a href="https://pavlopoulos-lab.org/shinyapps/app/mprabase">https://pavlopoulos-lab.org/shinyapps/app/mprabase</a>

# Chapter 4. Role of Epromoters in disease

## 4.1 Epromoters identification

Although enhancers were initially distinguished from promoters due to their distal regulatory functions, Banerji et al.'s paper in 1981 (Banerji et al., 1981) revealed that a 72bp segment defined as an enhancer was located about 200bp upstream of a gene's transcription start site. By today's definition, this would be considered within the promoter region, meaning the first discovered enhancer is also an Epromoter which we will discuss here. In recent years many studies found similar chromatin structural features (such as H3K4me1) and bidirectional transcription between promoters and enhancers (Andersson, 2015). This has given rise to concerns about the enhancer function of promoters. Nguyen et al. conducted high-throughput comparisons of promoter and enhancer activity in mouse neurons using MPRA, observing a clear positive correlation between the activities of enhancers and promoters (Nguyen et al., 2016). Engreitz et al. knocked out promoters of 12 lncRNAs and 6 protein-coding genes in mouse embryonic stem cells, noting that deletions at 9 sites (50%) significantly affected the expression of nearby genes (Engreitz et al., 2016). Diao et al. performed tiling-deletion across the 2-Mb POU5F1 locus in human embryonic stem cells, identifying 17 enhancer-like promoters that exhibit significant long-range interactions with the POU5F1 promoter (Diao et al., 2017). At the same time, Dao et al. systematically tested enhancer activity of coding gene promoters in HeLa and K562 cell lines by CapStarr-seq (Figure 4.1a). And 632 (3%) and 493 (2.37%) Epromoters among 20,719 promoters were identified in K562 and HeLa cells, respectively (Dao et al., 2017). In this study, Epromoters were defined as a subset of promoters that display enhancer activity to regulate distal gene expression. Dao et al. and Diao et al. demonstrate that promoter regions can enhance the expression of distal genes in vivo, cautioning against the inference of single proximal target genes for these regions (Figure 4.1b) (Catarino et al., 2017). The overlapping between enhancers and promoter regions illustrated that the sequence flexibility of different DNA elements allows different functions to map to the same sequence (Figure 4.1b).



We provided a review to discuss Epromoters which is entitled “Epromoters are new players in the regulatory landscape with potential pleiotropic roles” (see ANNEX 1). In the review, we discuss the different observations pointing to an important role of Epromoters in the regulatory landscape and summarize the evidence supporting a pleiotropic impact of these elements in disease.

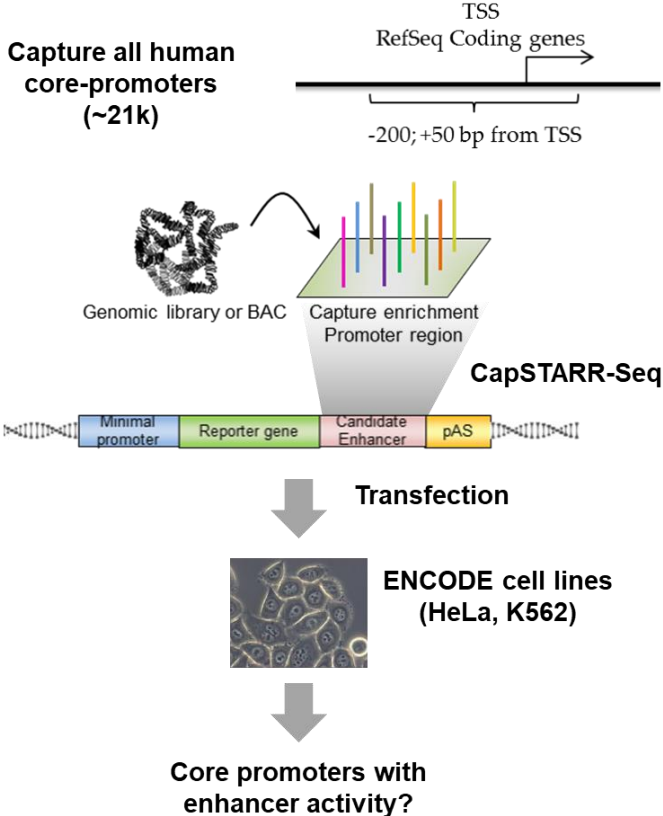


Figure 4.1a Epromoters were systematically identified by CapStarr-seq in K562 and HeLa cells.

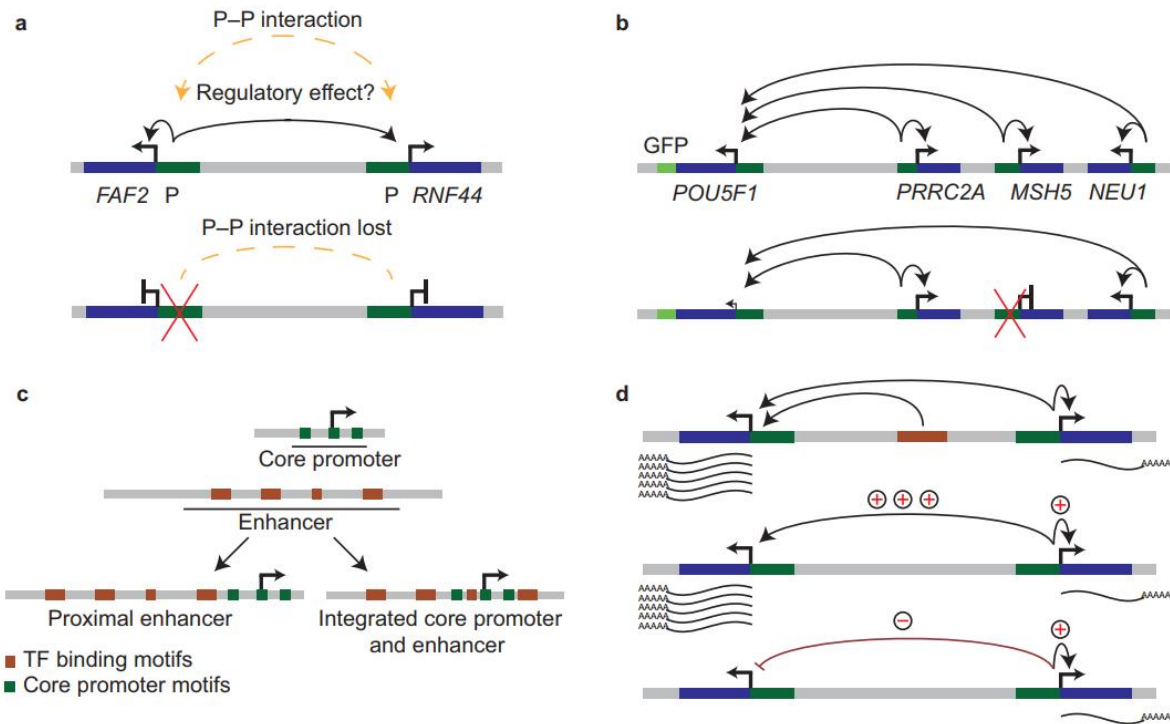


Figure 4.1b Promoters regulate distal gene transcription (Catarino et al., 2017). (a) Deletion of the *FAF2* promoter led to loss of distal *RNF44* expression. P, promoter (Dao et al., 2017). (b) A CRISPR–Cas9-mediated deletion screen using a *POU5F1* allele encoding GFP-tagged protein was used to identify *POU5F1* enhancers, including the core promoters of the *PRRC2A*, *MSH5*, and *NEU1* genes (Diao et al., 2017). (c) The sequence flexibility of different DNA elements allows different functions to map to the same sequence, such that the combination of enhancer and core promoter sequences can lead to proximal enhancers (left) or tightly integrated (or interwoven) elements (right) TF, transcription factor. (d) Enhancer activity in promoters does not necessarily correlate with proximal gene expression (Dao et al., 2017). This might stem from the integration of multiple enhancer inputs at promoters (top), the preferential regulation of distal genes (middle), or elements having distinct activating (+) or repressing (–) effects (bottom).

## 4.2 General features and potential mechanism(s) of Epromoters

Dao et al. have shown the genomic and epigenomic properties of Epromoters: gene expression associated with Epromoters is significantly higher than that associated with non-Epromoters, but the enhancer activity of Epromoters is not strictly correlated with the expression levels of associated genes; Epromoters exhibit a higher ratio of H3K27ac to H3K4me3; Epromoters have a higher density of various bound transcription factors and motifs; there are more frequent promoter-promoter interactions; and they more frequently overlap with eQTLs that affect the expression of genes at distal interactions.

Santiago-Algarra et al. explored the function of Epromoters in response to type I interferon. They find that clusters of IFN $\alpha$ -induced genes are frequently associated with Epromoters and that these regulatory elements preferentially recruit the STAT1/2 and IRF transcription factors and distally regulate the activation of interferon-response genes. A remarkable example is provided by the OAS locus where the OAS3 Epromoter regulates the interferon responses of the OAS1 and OAS2 genes (Figure 4.2b). Furthermore, they identified and validated the involvement of Epromoter-containing clusters in the regulation of LPS-stimulated macrophages. These findings suggest that Epromoters function as a local hub recruiting the key TFs required for coordinated regulation of gene clusters during the inflammatory response.

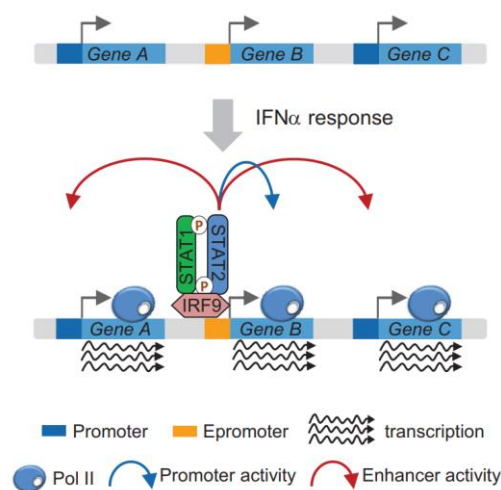


Figure 4.2a Epromoters might function as regulatory hubs for the coordinated induction of gene clusters (Santiago-Algarra et al., 2021).

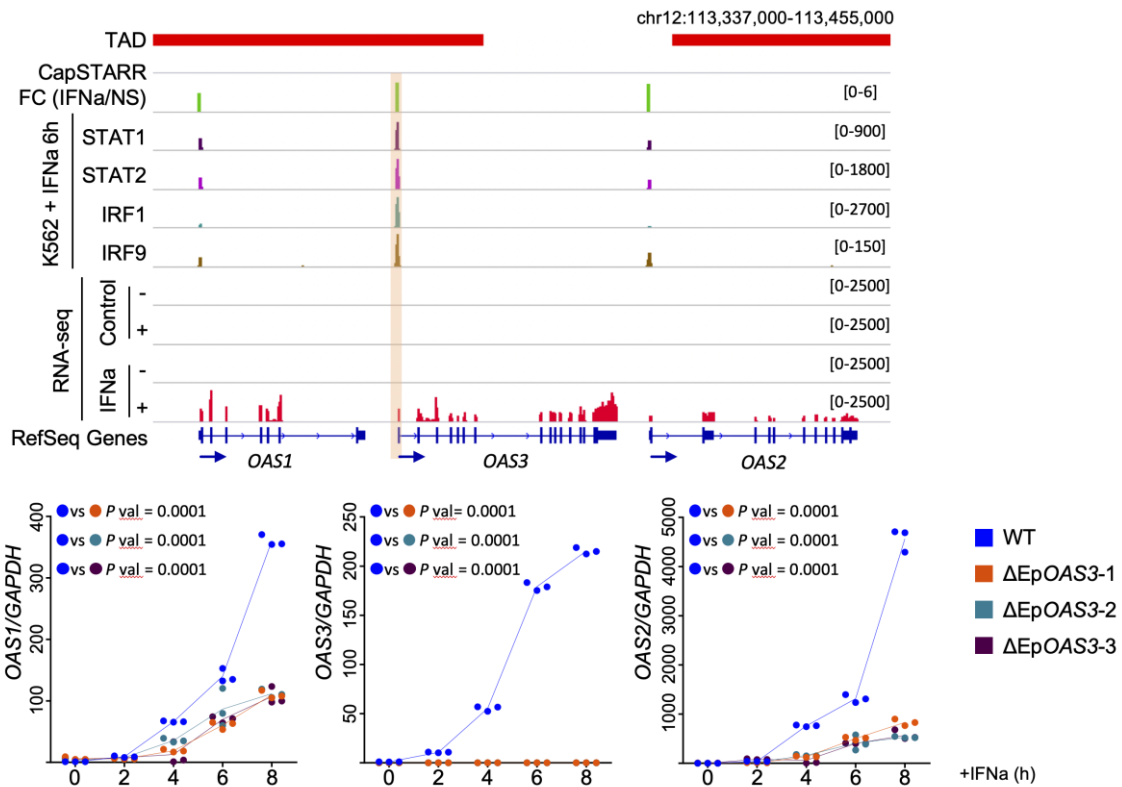


Figure 4.2b The deletion of the OAS3 Epromoter resulted in a dramatic reduction of OAS1 and OAS2 induction after IFNa stimulation (Santiago-Algarra et al., 2021).

## 4.3 Human diseases associated with Epromoters

The discovery of Epromoters opens a new paradigm in the study of regulatory variants as a mutation in a promoter could potentially influence the expression of several genes or change the relative ratio of promoter *versus* enhancer activity. Our previous studies demonstrated that human genetic variation within Epromoters influences distal gene expression (Dao et al., 2017; S. Nisar et al., 2022). Subsequent studies have suggested that SNPs affecting distal gene expression are enriched within Epromoters (Jung et al., 2019; J. Mitchelmore et al., 2020; M. Saint Just Ribeiro et al., 2022; D. Wang et al., 2018), while specific examples highlight the distal impact of disease-associated variants within Epromoters (Chandra et al., 2021; S. Nisar et al., 2022; V. Rusu et al., 2017; I. A. Sergeeva et al., 2016; Yagihara et al., 2016). For example, TF binding variation at the promoter of CLOCK gene does not affect CLOCK expression,

instead associates with the expression of distal gene SRD5A3, whose promoter it contacts in 3D as detected by PCHi-C (Figure 4.3a) (Joanna Mitchelmore et al., 2020). Two studies demonstrated that an alternative variant associated with prostate cancer increases the enhancer activity of the promoter leading to increased expression of two distal transcripts directly involved in cancer progression (Gao et al., 2018; Hua et al., 2018).

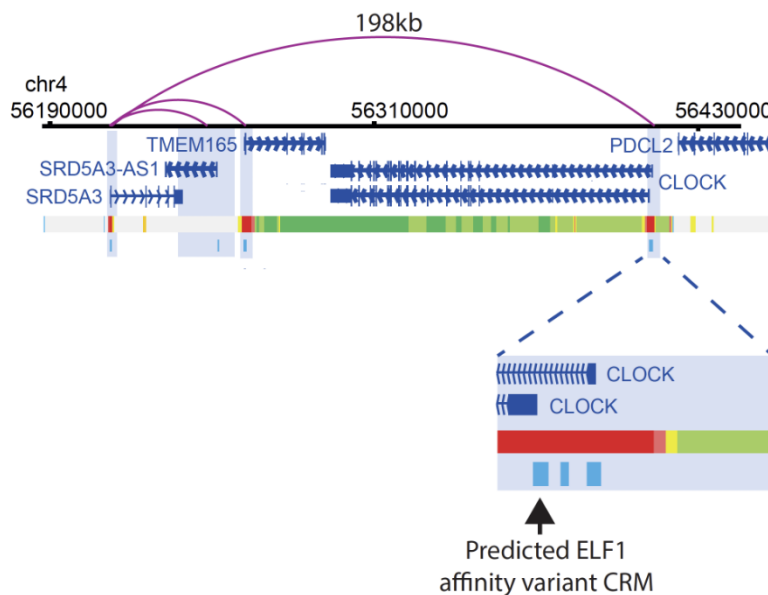


Figure 4.3a The variation in ELF1 binding affinity at promoter of CLOCK gene affects the expression of distal gene SRD5A3 (Joanna Mitchelmore et al., 2020).

Based on the above studies, we proposed the hypothesis that Epromoters might have a pleiotropic effect on the diseases by perturbing the expression of several genes at the same time. It might be envisioned that GWAS variants lying within Epromoters might regulate the expression of distal disease-causing genes. As suggested by the Epromoter’s finding, a mutation in a promoter could potentially influence the expression of several genes or change the relative ratio of promoter *versus* enhancer activity, thus, resulting in a variety of potential changes in the relative expression of neighboring genes (Figure 4.3b). The complex regulation by Epromoters might have two predicted consequences. On the one hand, there might be a general underestimation of the impact of Epromoter variation in disease because the causal gene might not be the closest one and therefore the link between genotype and phenotype might be missed in many case studies. On the other hand, as Epromoters potentially regulate several

genes at the same time and have the ability to efficiently recruit essential TFs, mutations in these regulatory elements are expected to have a stronger (higher penetrance) and/or pleiotropic (be involved in several diseases/traits) impacts on disease. Genetic variants might influence the intrinsic activity of Epromoters to primarily work as a promoter or as an enhancer, which might have an important impact on phenotypic traits and diseases. For instance, a genetic variant lying within an Epromoter might impact the expression of the proximal but also of one or several distal genes. The variant can generally affect the regulatory activity of the Epromoter or differentially impact the promoter or enhancer activity. These complex genetic regulations might result in either synergistic or additive (all affected genes are involved in the same disease/trait) or pleiotropic (each affected gene is involved in a different disease/trait) effects (Figure 4.3b).

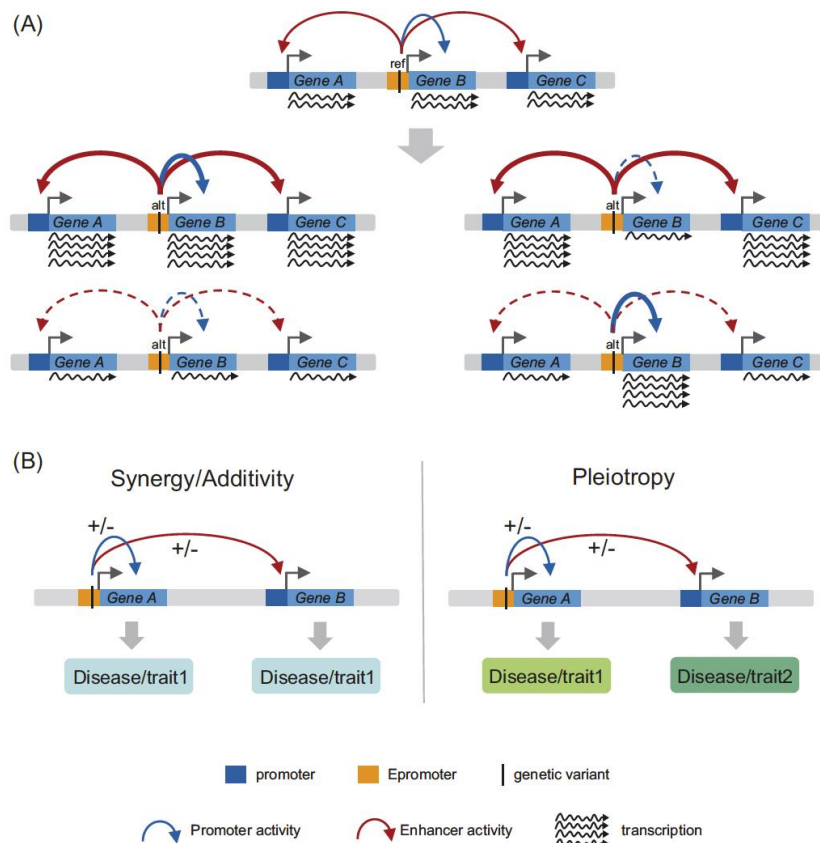


Figure 4.3b Effects of Epromoter variation on gene regulation (Malfait et al., 2023). (A) Different potential impacts of Epromoter genetic variation on proximal and distal gene expression. In the reference (ref) haplotype, proximal and distal genes transcription are regulated by the Epromoter. In the alternative (alt) haplotypes, the promoter (blue arrows) and enhancer (red arrows) activities could increase (thicker arrows) or

decrease (thinner and dashed arrows), resulting in up-or down-regulation of the associated genes. (B) Genetic variants at Epromoters might result in either synergistic/additive (all affected genes are involved in the same disease/trait) or pleiotropic (each affected gene is involved in a different disease/trait) effects.

As discussed in our review (Malfait et al., 2023) (see ANNEX 1), we have found some representative instances that show genetic variants at Epromoters affect distal gene expression to lead to diseases. For example, as shown in Figure 4.3c (A), the variant rs11672691 within the internal PCAT19 promoter is associated with prostate cancer. The alternative variant switches the relative promoter and enhancer activity resulting in up-regulation of the most upstream PCAT19 promoter and the distal gene CEACAM21. Figure 4.3c (B) shows the variant rs1046496 within the BAZ2B promoter is associated with hypothyroidism. The alternative variant decreases the transcription of the MARCHF7 gene. Figure 4.3c (C) shows the variant rs922483 within the BLK promoter is associated with systemic lupus erythematosus. The alternative variant decreases the transcription of the BLK gene while increasing the expression of the FAM167A gene. Figure 4.3c (D) shows a haplotype of five variants containing the lead variant rs10900585 within the internal promoter of ATP2B4 is associated with severe malaria. The alternative variant switches the relative promoter and enhancer activity resulting in up-regulation of the most upstream ATP2B4 promoter. More diseases associated variants in Epromoters have been listed in Table 4.3.

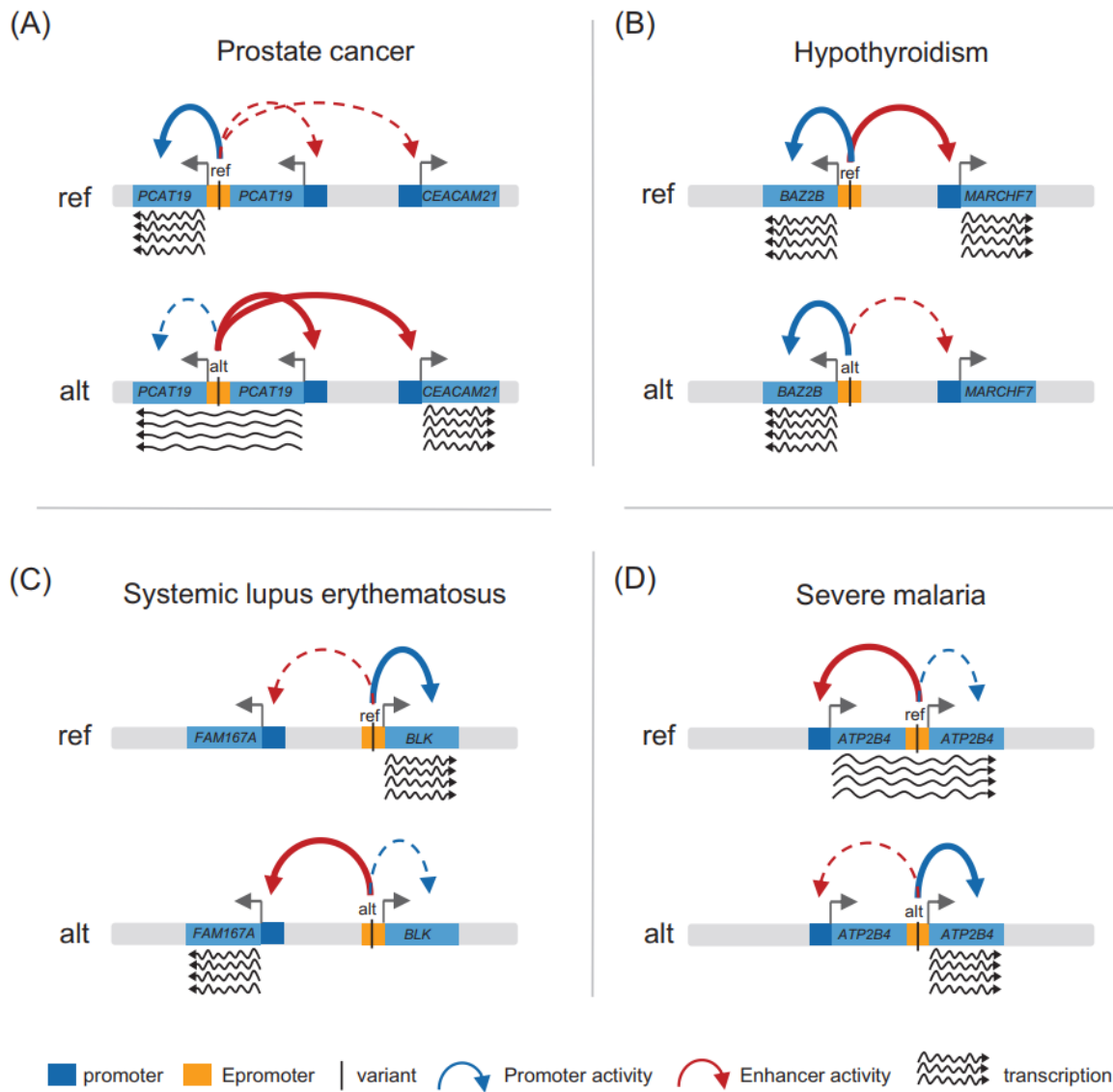


Figure 4.3c Examples of Epromoter variation associated with diseases (Malfait et al., 2023).



Table 4.3 List of diseases associated variants in Epromoters (updated from Malfait et al., 2023)

Disease	Epromoter-containing variant	Affected gene(s)	Evidence	Ref.
Prostate Cancer	<i>PCAT19</i> -short isoform	<i>PCAT19</i> -long & short isoforms, <i>CEACAM21</i>	P-P interaction, Enhancer & promoter reporter assays, CRISPR-Cas9 genome editing, CRISPR interference/activation	(Gao et al., 2018; Hua et al., 2018)
Systemic lupus erythematosus	<i>BLK</i>	<i>BLK</i> , <i>FAM167A</i>	P-P interaction, CRISPR interference,	(Mariana Saint Just Ribeiro et al., 2022)
Severe Malaria	<i>ATP2B4</i> -short isoform	<i>ATP2B4</i> -long & short isoforms	Enhancer & promoter reporter assays, CRISPR/Cas9 genome editing	(Samia Nisar et al., 2022)
Cardiovascular diseases	<i>Nppb</i>	<i>Nppa</i>	P-P interaction, Mouse transgenic models	(Man et al., 2018; Irina A. Sergeeva et al., 2016)
Type 2 diabetes	<i>ARAP1</i>	<i>PDE2A</i>	CapSTARR-seq, eQTL	(Kulzer et al., 2014; Medina-Rivera et al., 2018)
Rheumatoid arthritis	<i>CCR6</i>	<i>RNASET2</i>	P-P interaction, eQTL	(Chandra et al., 2021)
Systemic lupus erythematosus	<i>TREH</i>	<i>CXCR5</i>	P-P interaction, eQTL	(Su et al., 2020)
Crohn disease	<i>SMAD3</i>	<i>SMAD3</i> , <i>AAGAB</i>	P-P interaction, eQTL	(Mumbach et al., 2017; Y. Wang et al., 2018)
Coronary artery disease	<i>CDKN2B</i>	<i>IFNA2</i>	P-P interaction	(Li et al., 2019)
Multiple cancers	<i>TERT</i>	<i>CLPTM1L</i>	Somatic mutations, correlation with gene expression	(Fredriksson et al., 2014)
Type 2 diabetes	<i>INS</i>	<i>SYT8</i>	P-P interaction, siRNA	(Xu et al., 2011)
Schizophrenia	<i>VSP45</i>	<i>AC244033.2</i> , <i>C1orf54</i>	P-P interaction, CRISPR-Cas9 genome editing, CRISPR interference	(Zhang et al., 2023)

# RESULTS

# Chapter 5. Background and objectives of the PhD work

In our team's previous work, we developed CapSTARR-seq for high-throughput detection of enhancer activity in specific genomic regions (Vanhille et al., 2015). The team then used CapSTARR-seq to systematically detect the enhancer activity of the human core promoter in K562 and Hela cell lines (Dao et al., 2017). They demonstrated that a subset of gene-promoters, termed Epromoters, actually works also as bona fide enhancers and regulates distal gene expression. More recent results suggested that Epromoters might play an essential role in the coordination of rapid gene induction during the inflammatory response (Santiago-Algarra et al., 2021). It appears that Epromoters work as a hub for recruiting the essential transcription factors (TFs) required for gene activation in different stress conditions and establishing connections with the other distal response genes.

According to these previous studies, my PhD work was primarily focused on the impact of genetic variants at Epromoters on human diseases. The basic hypothesis of my PhD work is that genetic variants at Epromoters might have a multi-effect on regulatory networks and diseases (Figure 5.1). This means that genetic variants at Epromoters could affect the proximal and/or distal gene expression, which leads to different phenotype variations and diseases. Therefore, the research questions are: How many Epromoters are in the human genome? What are the genomic features that distinguish Epromoters from typical promoters? Which diseases are associated with Epromoters? Does genetic variation at one Epromoter associate with multiple diseases? How do variants at Epromoters affect multiple gene expressions associated with different diseases?

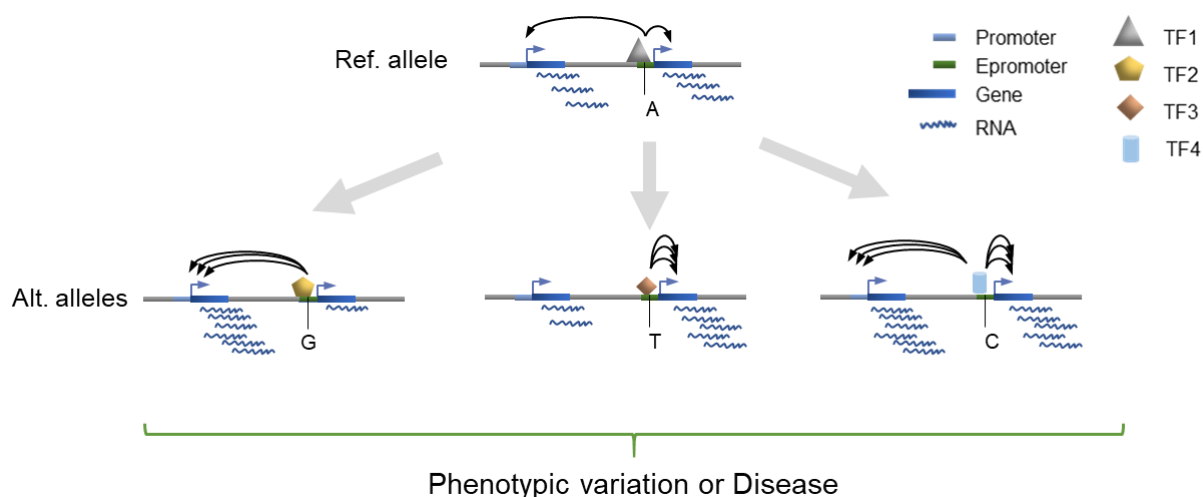


Figure 5.1 Hypothesis: genetic variants at Epromoters might have a multi-effect on regulatory networks and diseases.

Based on the hypothesis and research questions, the goal of my PhD work was: First, to build a comprehensive resource of Human Epromoters; Second, to describe the general features of Epromoters; Third, to evaluate the association between genetic variants and Epromoters in diseases. To accomplish these goals, we established the following general strategy (Figure 5.2): First, we collected STARR-seq datasets from published studies or databases and generated CapSTARR-seq resources in different human cell lines; Second, we identified Epromoters based on the enhancer activity regions from STARR-seq datasets; Third, we analyzed the genomic features of Epromoters as comparing with typical promoters; Then we utilized the GWAS resource to overlap the genetic variants with Epromoters; We used the eQTL datasets to identify the target genes of genetic variants at Epromoters; Last, the variants at Epromoters were validated by an MPRA resource.

The results are described in the draft entitled “Comprehensive mapping of genetic variation at Epromoters reveals pleiotropic associations with multiple disease traits” (Chapter 6). I also participated in a second study which focused on the role of Epromoters in the Stress response (under preparation; Chapter 7). Finally, the genomic resources generated during my PhD are currently used in different ongoing collaborations inside and outside our institute (see Perspective section; Chapter 6, section 6.2), including a study recently published (Castillo et al., 2024) (ANNEX 2).

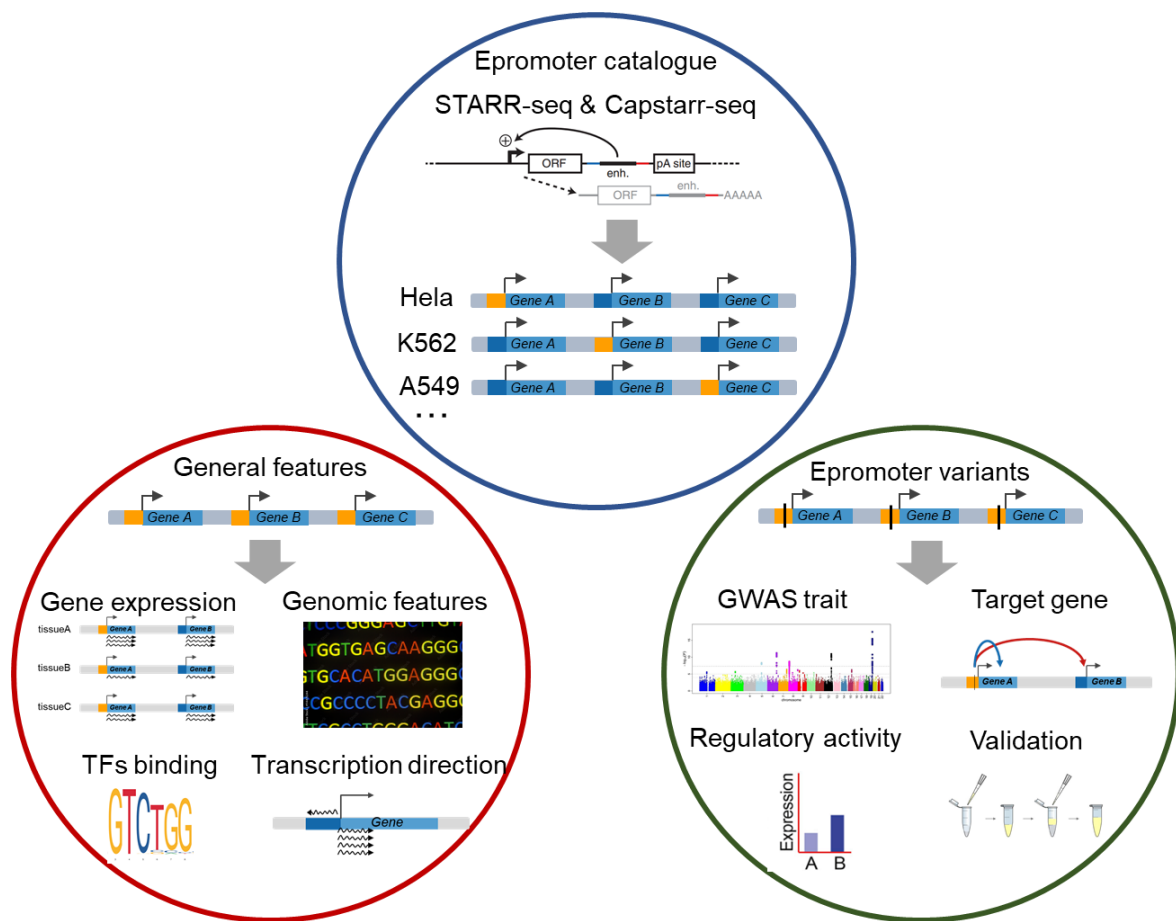


Figure 5.2 The general strategy to study the impact of genetic variants at Epromoters on human diseases.

# Chapter 6. Comprehensive mapping of genetic variation at Epromoters reveals pleiotropic association with multiple disease traits

## 6.1 Manuscript

### Comprehensive mapping of genetic variation at Epromoters reveals pleiotropic associations with multiple disease traits

Jing Wan<sup>1,2,#</sup>, Antoinette van Ouwkerk<sup>1,2,#</sup>, Jean-Christophe Mouren<sup>1</sup>, Carla Heredia<sup>3</sup>, Lydie Pradel<sup>1,2</sup>, Benoit Ballester<sup>1</sup>, Jean-Christophe Andrau<sup>3</sup>, Salvatore Spicuglia<sup>1,2,\*</sup>

<sup>1</sup>Aix-Marseille University, INSERM, TAGC, UMR 1090, Marseille, France.

<sup>2</sup>Equipe Labellisée LIGUE 2023, Marseille, France.

<sup>3</sup>Institut de Génétique Moléculaire de Montpellier, University of Montpellier, CNRS-UMR 5535, Montpellier, France

# These authors contribute equally

\* Corresponding author: Salvatore Spicuglia (salvatore.spicuglia@inserm.fr)

### Abstract

There is growing evidence that a wide range of human diseases and physiological traits are influenced by genetic variation of cis-regulatory elements. We and others have shown that a subset of promoter elements, termed Epromoters, also function as enhancer regulators of distal genes. This opens a paradigm in the study of regulatory variants, as single nucleotide polymorphisms (SNPs) within Epromoters might influence the expression of several (distal) genes at the same time, which could disentangle the identification of disease-associated genes. Here, we built a

comprehensive resource of human Epromoters using newly generated and publicly available high-throughput reporter assays. We showed that Epromoters display intrinsic and epigenetic features that distinguish them from typical promoters. By integrating Genome-Wide Association Studies (GWAS), expression Quantitative Trait Loci (eQTLs), and 3D chromatin interactions, we found that regulatory variants at Epromoters are concurrently associated with more diseases and physiological traits, as compared with typical promoters. To dissect the regulatory impact of Epromoter variants, we evaluated their impact on regulatory activity by analyzing allelic-specific high-throughput reporter assays and provided reliable examples of pleiotropic Epromoters. In summary, our study represents a comprehensive resource of regulatory variants supporting the pleiotropic role of Epromoters.

## Introduction

In higher eukaryotes, gene transcription is regulated through the involvement of regulatory elements that are located near the transcription start site (TSS), called promoters, and those that are located far from TSS, called enhancers. This classical definition implies that enhancers activate gene expression at a distance while promoters induce local gene expression. However, several lines of evidence have now established that some coding-gene promoters, termed Epromoters, also work as *bona fide* enhancers in different cellular contexts from drosophila to humans (Arnold et al., 2013; Corrales et al., 2017; Dao et al., 2017; Diao et al., 2017; Engreitz et al., 2016; Malfait et al., 2023; Nguyen et al., 2016; Rajagopal et al., 2016; Santiago-Algarra et al., 2021; Zabidi et al., 2015). These elements can regulate distal promoters when assessed in episomal reporter systems, as well, and more importantly, in their natural context. Subsequent studies have shown that Epromoters work as a hub for recruiting essential transcription factors (TFs) required for gene activation in different inflammatory and stress conditions and establishing connections with other distal response genes within the same clusters to ensure a rapid coordinated expression response (Dao et al., 2017; Santiago-Algarra et al., 2021). Although typical enhancers and promoters are generally distinguished by their relative location to the TSS of genes that they regulate, their shared architectural properties have suggested a unifying model of gene regulation by *cis*-regulatory elements (Andersson & Sandelin, 2020;

Core et al., 2014; Kim & Shiekhattar, 2015; Malfait et al., 2023; Medina-Rivera et al., 2018; Tippens et al., 2018). Previous studies have suggested that Epromoters share functional and architectural properties with both types of *cis*-regulatory elements (Andersson & Sandelin, 2020; Malfait et al., 2023; Medina-Rivera et al., 2018), but the intrinsic features driving the specific enhancer and promoter function of Epromoters are not yet elucidated.

There is growing evidence that a wide range of human diseases is influenced by the dysfunction of *cis*-regulatory elements caused by genetic, structural, or epigenetic mechanisms (Zaugg et al., 2022). These processes frequently underpin the susceptibility to common diseases but can be also directly involved in cancer or Mendelian diseases. The advent of genome-wide association studies (GWASs) in the past decade has been a great endeavor in genomic research toward identifying genetic variants associated with candidate genes for common diseases. The majority of these genetic variants are found in non-coding regions and, therefore are likely to be involved in regulatory mechanisms controlling gene expression (Deplancke et al., 2016; MacArthur et al., 2017; Maurano et al., 2012). However, a major challenge in interpreting the impact of genetic mutation or variation in disease is to identify the targets that are impacted by the genomic alteration, which might not necessarily be the closest genes and might have confounding features (Xia et al., 2016). Despite this, most studies select the closest gene to the associated GWAS variant to establish possible causal mechanisms, namely when the variant lies in the vicinity of a TSS or within an intronic region. However, GWAS variants might regulate the expression of distal disease-causing genes, in particular when lying within Epromoters.

The discovery of Epromoters thus opens a new paradigm in the study of regulatory variants. A mutation in a promoter could potentially influence the expression of several genes or change the relative ratio of promoter versus enhancer activity. This could result in a variety of potential changes in the relative expression of neighboring genes. In addition, it is plausible that the same *cis*-regulatory element displays preferential promoter activity in some tissues while displaying increased enhancer activity in other tissues, depending on the expressed combination of TFs and the epigenetic context (Chandra et al., 2021; Dao et al., 2017; Leung et al., 2015). Given the potential



regulation of proximal and distal genes by Epromoters, we hypothesized that genetic variation or mutation at Epromoters might therefore impact several physiological and pathological traits simultaneously.

To better assess the functional properties of Epromoters and the impact of genetic variation on physiological traits and diseases, we generated a comprehensive resource of human Epromoters by combining published and newly generated STARR-seq data from different cell lines and conditions. Epromoters displayed intrinsic genomics and epigenomics features that distinguish them from typical promoters. Furthermore, we found that Epromoters have a higher probability of being associated with multiple different GWAS traits, suggesting they are more pleiotropic. Strikingly, Epromoter pleiotropy was found to be associated with distal gene regulation and functional regulatory variants. Our finding supports the hypothesis of an important and pleiotropic role of Epromoter variation on the ontogeny of different diseases and physiological traits.

## Results

### **A comprehensive resource of human Epromoters**

To recover active enhancer regions in different cell types we recovered whole genome STARR-seq, ChIP-STARR-seq and CapSTARR-seq experiments from 28 datasets comprising 11 human cell lines and stimulatory conditions, including interferon-alpha (IFN $\alpha$ ) and multiple drug treatments (Supplemental Table S1). We retrieved a total of 58,388 non-redundant STARR-seq enhancers. We defined Epromoters as genomic regions of 500 bp upstream of the TSS of any coding gene that overlapped an active enhancer as defined by the STARR-seq assays (Figure 1a). The percentage of active enhancers that were defined as Epromoters ranged from 2.3% to 35.0% depending on the STARR-seq dataset (Supplemental Figure S1a). This resulted in a non-redundant set of 5,743 Epromoters, associated with 5,546 genes, and representing 15.4% of total coding-gene promoters (Supplemental Table S2). The percentage of Epromoters in each cell type/condition ranged from 0.3% to 2.9% (Figure 1b; Supplemental Table S3), with, on average, 1.5% of total coding gene promoters per experimental dataset (Figure 1c). For the majority of cell lines, more than 50% of Epromoters were also an

Epromoter in at least one other cell line (Figure 1d). Overall, 36.2% of Epromoters were shared between at least two cell lines (Figure 1e), supporting a physiologically diverse role of Epromoters.

We compared the average expression and tissue-specificity between non-redundant Epromoter-associated genes and the total set of genes using a comprehensive RNA-seq dataset across 30 tissues (Uhlen et al., 2016). We observed that Epromoter-associated genes were significantly more expressed (Figure 1f) and less tissue-specific (Figure 1g) than genes not associated with Epromoters. In order to compare our model Epromoters with a relevant set of typical promoters, we retrieved, for each of the 5,743 Epromoters, a typical promoter associated with a gene with a matching expression pattern to the Epromoter-associated gene across different tissues (hereafter termed “control promoters”,  $n=5,743$ ; Supplemental Figure S1b-c and Methods section). As shown in Figures 1f-g, genes associated with control promoters displayed similar average expression and tissue-specificity as Epromoter-associated genes, justifying the use of this control set as a proxy for typical promoters with similar promoter activity as Epromoters.

We assessed the transcriptional complexity of Epromoter-associated genes (i.e., number of TSS per gene) (Figure 1h). We observed that Epromoters were associated with genes harboring on average more TSS than other promoters (median value for Epromoters= 4). This suggests that in some cases, Epromoters might regulate an alternative promoter of the same gene, as previously suggested (Dao et al., 2017). We then assessed the 3D interactions between Epromoters and other distal promoters. We retrieved promoter-promoter (P-P) interactions based on published promoter-capture HiC (Javierre et al., 2016; Jung et al., 2019) and ABC models (Nasser et al., 2021) across a wide set of tissues. We observed that Epromoters and control promoters displayed a higher number of promoter interactions as compared to typical promoters, and to a lesser extent, to control promoters (Figure 1i), supporting the idea that Epromoters are more likely to be involved in distal gene regulation.

Finally, we predicted that the inactivation of Epromoters should affect the expression of neighboring genes. To assess the impact of Epromoters on distal gene expression, we analyzed a comprehensive Perturb-seq dataset in which all coding-gene promoters

had been repressed by CRISPR-based inactivation (CRISPRi) followed by single-cell RNA-seq analysis (Replogle et al., 2022). We first identified a set of 5,054 promoters that had been efficiently inactivated (i.e. the associated gene is among the top 2 of repressed genes). We then identified the promoters for which CRISPRi resulted in the repression of *cis*-distal genes (< 1 Mb) (Supplemental Table S4). We found that Epromoters significantly overlapped with the set of promoters associated with distal-gene regulation (hypergeometric test,  $P$  value = 0.02), while the control promoters did not. For example, CRISPRi repression of *DNAJC9*- and *ATP5MC1*-associated Epromoters resulted in downregulation of P-P interacting genes *MRPS16* and *UBE2Z*, respectively (Figure 1j). Similar results were found using a CRISPRi screen with a more restricted dataset (Gasperini et al., 2019) ( $P$  value = 0.01 for Epromoters, non-significant for control promoters). These results confirmed the potential regulation of distal genes by the identified Epromoters.

Overall, we have generated a comprehensive resource of human Epromoters based on STARR-seq data and confirmed their functional relevance as distal *cis*-regulatory elements.

### **Epromoters display specific genomic and epigenomic features**

We then asked whether there are specific genomic features that distinguish Epromoters from typical promoters. To make sure that the observed differences are not due to intrinsic promoter activity, but related to the enhancer activity, we further compared the Epromoters to the set of control promoters defined above.

First, we looked at phylogenetic conservation between Epromoters and typical promoters by comparing sequence conservation across placental mammals (Zoonomia, 2020). Most promoters from both subsets are under positive evolutionary constraint (Figure 2a; conservation score > 0), however, Epromoters are significantly more conserved than control promoters, potentially indicating that changes to Epromoters are unfavorable, i.e. they have indispensable function.

CpG islands (CGIs) are an important component of mammalian promoters. We found that 63% of Epromoters overlapped with CGIs as compared with 57% of control promoters (Figure 2b) (Chi-squared test,  $P$  value =  $3.5 \times 10^{-12}$ ), in agreement with the

ubiquitous expression of Epromoter-associated genes. CGIs are naturally enriched for G-quadruplexes (G4), which are secondary DNA structures suggested to play an important role in defining the chromatin structure and regulatory activity of *cis*-regulatory elements (Cyril Esnault et al., 2023; C. Esnault et al., 2023; Matos-Rodrigues et al., 2023). We, therefore, assessed whether G4 predictions were enriched at Epromoters (Figure 2c), using the G4hunter tool (Bedrat et al., 2016). While G4s were not enriched at Epromoters-overlapping CGIs, we found that non-CGI Epromoters harbor significantly more G4 as compared with control non-CGI promoters (Kolmogorov-Smirnov test). Similar results were obtained using different G4 prediction metrics (Supplemental Figure S2a-b). This suggests that beyond the CpG content, the density of G4 might have an important contribution to the Epromoter activity, reminiscent of a potential role of G4 structure at distal enhancers (Lyu et al., 2022).

To assess the complexity of transcription factor binding sites (TFBS) in Epromoters compared to the control promoter set we retrieved the overlap between the family of TFBS (non-redundant) based on the JASPAR database (Castro-Mondragon et al., 2022) and promoter elements. We found that Epromoters displayed higher density (i.e., number of TFBS per promoter; Figure 2d) and diversity (i.e., number of different TFBS families per promoter; Figure 2e) of TFBS as compared to control promoters. We then assessed the number of different TF-binding peaks (non-redundant) *per* promoter, using the ChIP-seq catalog from ReMap (Hammal et al., 2022). We found that Epromoters were bound by a higher number of TFs (Figure 2f), and across a higher number of biotypes (i.e., different cell types; (Figure 2g) (Chi-squared test). These findings align with the understanding that Epromoters are more complex *cis*-regulatory elements and the broader expression of their associated genes. To determine whether TF binding could distinguish between Epromoters and typical promoters, we conducted a nonlinear dimensionality-reduction using uniform manifold approximation and projection (UMAP) analysis using the TF binding information from ReMap (Figure 2h). We found that the primary dimension (UMAP1) was tightly associated with TFBS density (Figure 2h). Strikingly, three promoter groups could be identified based on the UMAP1 dimension that roughly separated Epromoters from control promoters (Figure 2i;  $P$  value =  $1.1 \times 10^{-45}$ ; Chi-Squared test comparing group 1 versus group 3; Supplemental Figure S2c), with the Epromoter-enriched cluster (group 1) displaying higher TF-binding density. We then identified the TFs that were

specifically enriched in Epromoters as compared to control promoters (Figure 2j). Among the top 25 enriched TFs, we found several inducible TFs such as the AP1 family (JUN, JUND and FOS), NfκB (RELA), STAT3 and ATF3. In addition, we also found EP300 (Rada-Iglesias et al., 2011), SMARCA4 (Hodges et al., 2018) and BRD2 (Cheung et al., 2017) in the top 5 enriched TFs, which are general co-factors associated with enhancer function (Supplemental Figure S2d-f).

Finally, we investigated the association with transcription initiation using the CAGE resource from FANTOM5 (Andersson et al., 2014). It has been previously shown that the strength of transcription initiation correlated with enhancer activity at both proximal and distal regulatory elements (Core et al., 2014; Dao et al., 2017; Henriques et al., 2018; Mikhaylichenko et al., 2018; Rennie et al., 2018). We observed that Epromoters were more frequently associated with divergent transcription (Figure 2k; Chi-squared test,  $P$  value =  $4.9 \times 10^{-3}$ ), as well as with more forward and reverse CAGE signals (Figure 2l). This latter observation suggests that Epromoters are associated with increased (bidirectional) transcription initiation, potentially reflecting the enhancer activity of Epromoters. Additionally, we assessed the recruitment of RNA-Polymerase II (RNAPII) to Epromoters and controls using a comprehensive RNAPII binding atlas (de Langen et al., 2023). Consistent with the CAGE results, we observed that Epromoters display more RNAPII binding (Figure 2m). This is reminiscent of a recent study suggesting a prominent role of RNAPII binding on the stabilization of distal interaction between *cis*-regulatory elements (Barshad et al., 2023).

Overall, we found that Epromoters display specific genomics and chromatin features compared to control promoters with similar transcriptional activity.

### **Genetic variation associated with Epromoters.**

First, we overlapped Epromoters and control promoters with both rare and common variants (SNP) from the SNPdb (NCBI) database. Both common and rare variants were significantly enriched at Epromoters ( $P$  values =  $4.7 \times 10^{-15}$  and  $6.2 \times 10^{-71}$ , respectively; Chi-Squared test). We then extracted 186,120 variants associated with 4,138 Genome-Wide Association Studies (GWAS) from the GWAS catalog (Sollis et al., 2023) and retrieved over 2.4 million common SNPs in high linkage disequilibrium (LD;  $r^2 > 0.8$ ; 1000 Genomes project) with GWAS tag SNPs (Figure 3b; hereafter, GWAS-

SNPs). We obtained 4,330 and 4,062 GWAS-SNPs overlapping 2,301 Epromoters and 2,241 control promoters, respectively (Supplemental Table S5). In fact, 40% of Epromoters and 39% of control promoters harbored at least one GWAS-SNP (Figure 3c).

We further investigated the enrichment of GWAS in Epromoters. In total, 1,251 GWAS traits are associated with 4,330 SNPs at Epromoters (Supplemental Table S5). We found that 184 GWAS traits were significantly enriched at Epromoters compared to the genome background ( $P$  value  $< 0.001$ ; hypergeometric test) while 12 GWAS traits were differentially enriched as compared with control promoters ( $P$  value  $< 0.05$ ; Chi-squared test; Supplemental Table S6; Figure 3d). To assess the heritability of Epromoters for GWAS, we calculated the partitioned heritability of 176 GWAS summary statistics using the LD score regression model (Finucane et al., 2015) for Epromoters, control promoters, FANTOM-enhancers (Andersson et al., 2014), and UCSC-defined promoters and coding regions (Supplemental Figure S3). We observed that certain GWAS traits displayed high heritability either in Epromoters, control promoters, or enhancers, but low heritability in total promoters and coding regions. Overall, we found that Epromoters were associated with specific physiological traits or diseases.

Further analysis focused on the association of Epromoters with multiple GWAS traits. Epromoters exhibited more GWAS traits per GWAS-SNP and *per* promoter than control promoters (Figure 3e and 3f, respectively). This observation suggested that Epromoters are associated with a broader range of traits, possibly indicating pleiotropy, referred here as to a single *cis*-regulatory element affecting more than one trait independently (Cano-Gamez & Trynka, 2020). Additionally, we investigated whether pleiotropic GWAS-SNPs were associated with different GWAS categories. Indeed, Epromoters and their associated GWAS-SNPs were found to be more frequently associated with different GWAS categories (Figure 3g and 3h, respectively), supporting the hypothesis that Epromoters play a more pleiotropic role than typical promoters in influencing diverse traits.

**Epromoter's pleiotropy is associated with the regulation of multiple target genes**

To identify potential target genes associated with Epromoter variation, we integrated expression Quantitative Trait Loci (eQTL) datasets obtained from the fine-mapped credible sets within the EBI eQTL Catalog (Kerimov et al., 2021). From 9,137,260 fine-mapped eQTLs, we found 5,843 associated with 2,768 Epromoters and 5,644 associated with 2684 control promoters (Supplemental Table S5). In general, Epromoter and control promoter-associated GWAS-SNPs were found to be enriched in eQTLs (Figure 4a) ( $P$  values =  $6.2 \times 10^{-59}$  and  $2.6 \times 10^{-53}$  for Epromoter and control sets, respectively; Chi-squared test), highlighting the regulatory potential of these variants. Specifically, 48.2% and 47.7% of GWAS-associated Epromoters and control promoters overlapped with at least one eQTL, respectively. Among 2,768 Epromoters and 2,684 control promoters with eQTLs, approximately half exhibited at least two eQTLs.

Furthermore, our analysis delved into the association of eQTLs with proximal, distal, or both proximal and distal target genes (Figure 4b). As expected, Epromoter eQTLs were less associated with proximal genes as compared with control eQTLs (Figure 4b) ( $P$  value = 0.006 for all eQTLs,  $P$  value = 0.05 for GWAS eQTLs, Chi-squared test). Surprisingly, GWAS SNPs were depleted of proximal eQTLs ( $P$  value =  $3.3 \times 10^{-9}$ , Wilcoxon test) and enriched in proximal-distal eQTLs ( $P$  value =  $1.4 \times 10^{-9}$ , Wilcoxon test) for both Epromoter and control promoters sets. However, this category might represent a mixture of *cis* and *trans* effects (Liu et al., 2019; Vosa et al., 2021; Westra et al., 2013), likely combining a *cis* effect on the proximal gene and *trans* effects on distal genes (see below).

Next, we compared the pleiotropic impact on diseases of eQTLs associated with either proximal, distal or proximal and distal genes (Figure 4c). On the one hand, both Epromoters and control eQTLs with both proximal and distal targets were highly pleiotropic, with no significant differences between the two sets (median of GWAS traits = 3; Wilcoxon test). As mentioned above, we believe most of these pleiotropic eQTLs are associated with both *cis* and *trans* effects. On the other hand, Epromoters with proximal- or distal-only eQTLs demonstrated higher pleiotropy than corresponding control eQTLs, suggesting a stronger role of Epromoter variants on *cis*-regulatory functions. To confirm that the pleiotropy associated with distal eQTLs from Epromoters was due to *cis* interaction with distal targets (as opposed to *trans* effects), we analyzed

their consistency with P-P interactions (Figure 4d). Strikingly, Epromoters with consistent distal targets displayed a significant increase in pleiotropy, affirming the link between Epromoter variants and the actual regulation of distal genes. Control promoters did not exhibit the same trend, emphasizing the unique regulatory role of Epromoters in distal gene interactions.

As we found that Epromoters are frequently associated with genes harboring multiple TSSs (Fig. 1h), we wondered whether the higher pleiotropy observed with proximal eQTLs might be linked to distal regulation by alternative promoters as previously suggested (Dao et al., 2017). Indeed, we observed a higher pleiotropy only at Epromoters associated with multiple TSSs (Figure 4e). In fine, the higher pleiotropy observed at Epromoters appears to be linked to the actual regulation of distal targets, including either alternative promoters or distal genes.

### **A pleiotropic Epromoter variant associated with COVID-19 shows enhancer/promoter switch**

Among the promoters that contain disease-associated SNPs, we identified six Epromoters that we previously demonstrated by CRISPR-Cas9 genetic deletion to regulate distal genes, including *OAS3*, *ISG15*, *IFIT3*, *IL15R*, *METTL21* and *BAZ2B* (Dao et al., 2017; Santiago-Algarra et al., 2021) (Supplemental Table S5). This supported a functional link between the genetic variants at these Epromoters and the regulation of distal genes. Among those, the *OAS3* Epromoter provided a remarkable example of a pleiotropic locus. The *OAS3* gene is embedded in a cluster that also includes *OAS1* and *OAS2* (Figure 5a), which all encode for the oligoadenylate synthetase (OAS) family of proteins and play an important role in antiviral immunity (Hornung et al., 2014). The *OAS1/2/3* locus is a highly pleiotropic locus associated with several diseases, including asthma, blood protein measurement, chronic leukemia, systemic lupus erythematosus, and severe COVID-19 (Supplementary Table 5). Furthermore, the minor allele haplotype of the *OAS1/2/3* locus is a Neanderthal haplotype, first introduced into the modern human population by interbreeding with Neanderthals around 50,000 years ago (Zeberg & Pääbo, 2021). This haplotype spans a 75 kb region, and variants of this haplotype have been associated with protection



against West Nile Virus (Lim et al., 2009), increased resistance to hepatitis C infection (El Awady et al., 2014), and protection against SARS-CoV (He et al. 2006), and most recently with reduced risk of becoming severely ill upon SARS-CoV-2 infection (Initiative, 2021; Pairo-Castineira et al., 2021). The *OAS3* promoter showed IFN $\alpha$ -dependent enhancer activity in HeLa, K562 and CCRF-CEM cell lines (Supplementary Table 3). Strikingly, we previously showed that deletion of the *OAS3* Epromoter resulted in impaired induction of the entire *OAS* cluster after IFN $\alpha$  stimulation (Santiago-Algarra et al., 2021), suggesting this element is a master regulator of the interferon response of the locus. Since there is no indication of other regulatory regions within the *OAS1/2/3* locus, except the promoters of the three genes (Santiago-Algarra et al., 2021), we assumed that *cis*-regulatory variants mainly reside in the *OAS3* Epromoter.

We initially identified rs1156361 (located 352 bp upstream of the *OAS3* TSS) as a GWAS-SNP within the *OAS3* Epromoter (Supplementary Table 5). eQTL data of the GTEx database indicates that the minor allele of rs1156361 is associated with lower expression of all three *OAS* genes in multiple tissues (Figure 5b), consistent with the role of this Epromoter as a master regulator of the *OAS* locus. We realized that the promoter library used for the CapSTARR-seq experiments contains both alleles of the rs1156361 SNP. We therefore assessed the allele-specific activity of this SNP in the K562 and CCRF-CEM cell lines with or without IFN $\alpha$  stimulation (Figure 5c). We observed that the *OAS3* Epromoter harboring the major allele (C) displayed a significantly higher enhancer activity upon stimulation with IFN $\alpha$ . Upon closer inspection of the *OAS3* promoter [-500bp; 250bp], we found 4 SNPs in high LD ( $r^2 > 0.97$  in the European population) with rs1156361: rs3815178 and rs1859331 (5' UTR variants), rs1859330 (missense variant) and rs1859329 (synonymous variant). These additional 4 SNPs are also in eQTLs with *OAS1/2/3* with the same directionality as rs1156361 (Supplemental Figure S4). To assess the contribution of the two haplotypes on the relative promoter and enhancer activity of the *OAS3* Epromoter, we performed luciferase reporter assays in K562 cells using a 726 bp genomic region containing the 5 SNPs. We observed that the major haplotype confers both a stronger promoter and enhancer activity in K562 after IFN $\alpha$  stimulation (Figure 5d). We also performed the luciferase reporter assays in A549 cells, a lung epithelial cell line commonly used as a model for COVID-19 (Plaze et al., 2021; Pyrc et al., 2021). In this cell line, the major

haplotype similarly conferred stronger promoter activity, but the minor haplotype displayed stronger enhancer activity (Figure 5e). Interestingly, there was a higher absolute promoter activity in IFN $\alpha$ -treated K562 cells compared to the enhancer activity, while the opposite was observed in A549 cells. Overall, these results suggest that the pleiotropic association of the *OAS1/2/3* locus with multiple diseases, including severe COVID-19, might be explained, at least partially, by transcriptional deregulation of all three *OAS* genes by regulatory variants lying within the *OAS3* Epromoter. Our results also highlight the differential impact of genetic variants on enhancer versus promoter activity of Epromoters.

### **Functional assessment of pleiotropic Epromoter's variants**

To globally assess the functional impact of Epromoter's variants, we compiled the results from 24 published Massive Paralleled Reporter Assays (MPRA) experiments (Supplemental Table S7), which have assessed the regulatory impact of genetic variants. From 37,829 SNPs with significant allelic impact on regulatory activity (allelic-skewed SNPs), 292 and 209 overlapped with GWAS-SNPs from Epromoter and control promoters, respectively (Figure 6a). Strikingly, Epromoter GWAS-SNPs with MPRA-validated allelic impact displayed significantly higher pleiotropy (Figure 6b), while control GWAS-SNPs did not. To further explore the functional relevance of Epromoter GWAS-SNPs, we assessed the impact on TF binding by interrogating the SNP-SELEX dataset (Yan et al., 2021), which systematically assessed the binding of 270 human transcription factors to 95,886 noncoding variants in the human genome using an ultra-high-throughput multiplex protein-DNA binding assay (Figure 6c). We found that Epromoter GWAS-SNPs that impact TF binding (skewed TF binding) displayed higher pleiotropy than the remaining Epromoter GWAS-SNPs, while there were no significant differences in the case of control promoters. Similar results were observed when analyzing allelic-specific TF binding *in vivo* using the ANANASTRA resource (Boytssov et al., 2022) (Figure 6d). Altogether, these results suggest that the observed pleiotropic effects are due to the functional impact of Epromoter's variants in terms of skewed *cis*-regulatory activity and TF binding.

We next integrated the different levels of validation resources to retrieve a list of 156 Epromoter overlapping GWAS-SNPs with consistent distal eQTLs, P-P interactions, allelic-skewed MPRA activity and TF binding (Figure 6e). From this list, 123 (79%)

SNPs were associated with more than one GWAS trait, thus representing a resource of bona fide pleiotropic Epromoters (Figure 6f; Supplemental Table S5). Figure 7 provides four examples of pleiotropic Epromoters (*SETD1A*, *COASY*, *ORMDL3* and *PPIL3*) with consistent 3D interaction and eQTL target genes (Figure 7a-b), significant differences on allelic regulatory activity (Figure 7c) and predicted perturbation of TF binding (Figure 7c; Supplemental Table S5). Careful examination of proximal and distal target genes suggested that the association with multiple GWAS traits might be explained by the combination of the individual gene functions (see Supplemental Information 1 for a detailed description of each locus). For example, The *SETD1A* Epromoter is a highly pleiotropic locus involved in over 30 diverse GWAS traits, including immune-associated diseases (Graves, psoriasis, Crohn's, eosinophil count), neurological diseases (Parkinson's, epilepsy, anxiety) and heart disease risk factors (BMI, blood and pulse pressure, triglycerides and LDL cholesterol measurements). The associated rs4889599 SNP is predicted to affect the binding of the HTATIP2 (Figure 7d; Supplemental Table S5) and displayed allelic-skewed MPRA activity (Figure 7c, top panel). Interestingly, *SETD1A* and the *STX1B* distal target are both associated with neurological disorders, while the *HSD3DB7* and *STX4* distal targets are associated with immune-related and cardiometabolic diseases, respectively (Supplemental Information 1). Overall, we concluded that the pleiotropic association of Epromoters with multiple diseases and traits is linked to the *cis*-regulatory impact of the genetic variants and the combination of the physiological functions of proximal and distal target genes.

## Methods

### Cell culture

K562, CCRF-CEM and RPMI cells were maintained in RPMI 1640 medium GlutaMAX (Gibco, 61870010) supplemented with 10% FBS (Gibco, Fetal Bovine Serum A5256701) (inactivated at 55°C for 1 hour) between  $0.3 \times 10^6$  and  $1 \times 10^6$  cells per mL, incubated at 37°C with 5% CO<sub>2</sub>. GM12878 cells were maintained in the same conditions but with 15% instead of 10% FBS. Cells were tested for mycoplasma infection once a month and tested negative. A549 cells were maintained in DMEM/F12 GlutaMAX (Gibco, 10565018) supplemented with 10% deactivated FBS, incubated at 37°C with 5% CO<sub>2</sub>. When 90% confluent, the medium was aspirated and cells rinsed with PBS, followed by trypsinization (Trypsin-EDTA (0.05%), phenol red, Gibco, 25300-062) at 37°C for 5 minutes. 5x the volume of medium is added to detach the cells from the dish, the cells are centrifuged, resuspended in the medium and split at the appropriate density into a new dish. Cells were tested for mycoplasma infection once a month and tested negative.

### CapSTARR-seq

The human promoter CapSTARR-seq library used in this study has been generated previously (Dao et al., 2017; Santiago-Algarra et al., 2021). The STARR-seq protocol was performed in CCRF-CEM (without stimulation and with IFN $\alpha$  stimulation), RPMI and GM12878 cell lines. 100 million cells were transfected with 1.25 mg of CapSTARR-seq promoter library using the Neon transfection system (Thermo Fisher Scientific) using the following settings: voltage V 1300, pulse width 20 and pulse number 3. After 24 hours of incubation, either the STARR-seq protocol was performed as published before (Dao et al., 2017), or (for CCRF-CEM cells) interferon alpha (IFN $\alpha$ ) was used to induce interferon response (100 ng/mL, Sigma Aldrich, SRP4594) for 6 hours followed by the STARR-seq protocol (Santiago-Algarra et al., 2021). cDNA and input libraries were sequenced on an Illumina NextSeq500, and mapping and analysis were performed as published (Santiago-Algarra et al., 2021).

### STARR-seq and CapStarr-seq data processing

Human enhancers were retrieved from 19 whole-genome STARR-seq, 2 ChIP-STARR-seq and 7 CapStarr-seq datasets (Supplemental Table S1). Seventeen whole

genome STARR-seq datasets (A549, MCF-7, HCT116, SH-SY5Y, HepG2 and K562 with different stimulation) were obtained from ENCODE and were already processed by STARRPeaker (Lee et al., 2020) or MACS2 (Zhang et al., 2008) for peak calling defining the active enhancer regions. The peak files in bed format were directly downloaded from ENCODE (ENCODE accessions in Supplemental Table S1). To recover high-quality peaks, we took common peaks from different replicates for each dataset and averaged the enhancer activity values. Common peaks were ranked by the average values and peaks with the values higher than the inflection point (inflection R package) were taken as enhancers in this study. Two whole genome STARR-seq datasets in HeLa were collected from supplementary data (GSE100432) of Muerdter et al. (Muerdter et al., 2018). Two ChIP-STARR-seq datasets in hESC were collected from supplementary data of Barakat et al. (Barakat et al., 2018) (GSE99631). The two hESC datasets were filtered by at least one of the active regions of NANOG, OCT4, H3K27ac, and H3K4me1, with the enhancer activity score RPP (reads per plasmid) over 256 according to the original analysis described in (Barakat et al., 2018). Three Capstarr-seq datasets in HeLa and K562 were collected as Epromoters from supplementary data of Dao et al. (Dao et al., 2017) and Santiago et al. (Santiago-Algarra et al., 2021). Four Capstarr-seq datasets in GM12878, CCRF-CEM (with and without IFN $\alpha$  stimulation) and RPMI were generated in this study (GEO accession numbers are provided in Supplemental Table S1) and processed as previously described (Dao et al., 2017; Santiago-Algarra et al., 2021). Briefly, fastq files were trimmed using sickle with -q 20 option and mapped to the hg19 reference genome using Bowtie2 with default parameters. Sam files were converted using SamTools and bed files were generated with bedtools "BamToBed" command. Fragment reads were extended to 314 nt, corresponding to the average size of the captured fragments. Coverage of captured regions was computed using bedtools "coverage" command for both transfected and non-transfected libraries. The coverage was normalized by Fragments per kilobase per million reads mapped (FPKM). Promoter regions with an FPKM < 1 in the input library were removed. The ratio of the Capstarr-seq coverage over the input (fold-change) was computed for each sample. Promoter regions with enhancer activity were defined using the inflection point of the ranked fold-change as a threshold. Finally, all the enhancer regions from the 28 datasets were converted to hg38 coordinates and merged into a single non-redundant list in bed format. This

resulted in 58,388 non-redundant enhancers in 11 cell lines (Supplemental Tables S2 and S3).

### **Epromoter identification**

To identify Epromoters in the human genome, first, we defined the promoter region according to the hg38 genome annotation file from Ensembl (release-103, [http://ftp.ensembl.org/pub/release-103/gtf/homo\\_sapiens/](http://ftp.ensembl.org/pub/release-103/gtf/homo_sapiens/)). The promoters were defined as 500bp region upstream of the TSS (transcription start site) of each protein-coding transcript. The promoter regions were overlapped with no-redundant enhancers by bedtools intersect (v2.28.0) (Quinlan & Hall, 2010), with at least 50% overlap (bedtools intersect -wa -wb -f 0.5 -F 0.5 -e). The enhancer-overlapping promoters were defined as Epromoters. The Epromoter regions were merged if they overlapped by at least 1 nt. Finally, 5743 non-redundant Epromoters were defined (Supplemental Tables 2 and 3).

### **Gene expression and tissue specificity calculation**

Gene expression data was downloaded from the supplementary data of Uhlén et al. (Uhlen et al., 2016) (Table EV1). The study provided a gene expression matrix of 18684 genes across 30 human tissues from GTEx. The tissue specificity was calculated according to Yanai et al. (Yanai et al., 2005), using the following formula:

$$Tissue\ specificity\ index = \frac{\sum_{i=1}^N (1 - x_i)}{N - 1},$$

where N is the number of tissues and  $x_i$  is the expression profile component normalized by the maximal component value. The tissue specificity index varies from 0 to 1, where 0 means broad expression and 1 means high specificity.

### **Control promoter set**

We generated a control promoter set associated with genes displaying the most similar expression patterns as the Epromoter-associated genes. First, all coding genes were clustered according to the gene expression across 30 tissues using the expression matrix from (Uhlen et al., 2016) (Hierarchy cluster was performed with the “Euclidean” method in R4.3.2). For each Epromoter-associated gene, the gene that is nearest to the Epromoter gene in the cluster results was assigned as a control gene. The control promoter regions were defined as described for Epromoters.

### **Promoter-promoter interactions analysis**

The promoter-promoter interaction data were collected from 2 promoter capture-Hi-C studies (Javierre et al., 2016; Jung et al., 2019) and the ABC model predictions (Nasser et al., 2021). We downloaded the processed high-confidence interactions (CHICAGO score $\geq$ 5) from Supplemental Data S1 of Javierre et al., which was generated by promoter capture-Hi-C from 17 blood cell types. The data from Jung et al. was downloaded from their Supplementary Table 4, which includes processed significant promoter-promoter promoter capture-Hi-C interactions from 26 human tissues. Nasser et al. provided a comprehensive element-gene connections resource across 131 human cell types and tissues by the ABC model, which is a high-performance prediction model based on measurements of chromatin accessibility, H3K27ac, and Hi-C data (Nasser et al., 2021). The ABC predictions in 131 cell types and tissues were downloaded from (<https://www.engreitzlab.org/resources>). After converting to hg38, all interactions from the three datasets were overlapped with total promoters (5' upstream 500bp of TSS, Ensembl) in both anchors as the total promoter-promoter interactions. The target genes of Epromoters and control promoters were identified by overlapping their coordinates with the total promoter-promoter interactions, which include the target genes associated with each promoter. The circular visualization and promoter-promoter interactions in the Epromoter instances (Figure 7) was performed by R package circlize (Gu et al., 2014).

### **CRISPRi screen analysis**

CRISPRi screen data was collected from Replogle et al. (Replogle et al., 2022) and Gasperini et al. (Gasperini et al., 2019). Replogle et al. generated genome-scale CRISPRi screen data in K562 by Perturb-seq. We downloaded the processed Perturb-seq file of K562 genome-scale sample in h5ad format (<https://gwps.wi.mit.edu/>, gemgroup Z-normalized pseudo-bulk expression data). The processed Perturb-seq file was processed by Seurat (V5) (Hao et al., 2024) into a normalized expression matrix of all gRNAs and effect genes. We first identified a set of 5,054 promoters that had been efficiently inactivated (i.e. the associated gene is among the top 2 of repressed genes). The top 30 repressed genes were taken as regulated genes of these promoters. We then identified the promoters for which CRISPRi resulted in the repression of *cis*-distal genes (< 1 Mb). This CRISPRi result was intersected with

Epromoters and control promoters to identify their *cis*-regulated genes. Gasperini et al. performed CRISPRi perturbations in K562, which include target sites on TSS as positive controls. We downloaded the CRISPRi screen results from the pilot and scale experiments (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE120861>). The results include the target sites' position and expression-affected genes. The target sites of positive controls were extracted to overlap with Epromoters. The expression-affected genes that were not on the target sites were taken as distal effect genes of Epromoters. The same analysis was performed for control promoters.

### **Sequence conservation analysis**

The sequence conservation data were downloaded from the Zoonomia Placental Mammals track (including 241 vertebrate species) (Zoonomia, 2020) in the UCSC genome browser (<https://hgdownload.soe.ucsc.edu/goldenPath/hg38/cactus241way/cactus241way.phyloP.bw>). The conservation scores included in the bigwig file were computed by phyloP (Pollard et al., 2010) from the PHAST package (Hubisz et al., 2011) at each single nucleotide level. In this conservation score, each base with positive scores was predicted as conserved, and negative scores were predicted as fast-evolving. The bigwig file of conservation scores was converted into wig format by “bigWigToWig” and then into bed file by “wig2bed”. The conservation score of each Epromoter was calculated by the sum of all the bases.

### **CpG island and G4 analysis**

The CpG islands (CGIs) annotations have been recovered from UCSC ([https://genome.ucsc.edu/cgi-bin/hgTables?hgta\\_doMainPage=1&hgta\\_group=regulation&hgta\\_track=cpgIslandExt&hgta\\_table=cpgIslandExt&hgsid=1956573466\\_K6emxI9N7oynnWsuT8Zjnyk9XX6n](https://genome.ucsc.edu/cgi-bin/hgTables?hgta_doMainPage=1&hgta_group=regulation&hgta_track=cpgIslandExt&hgta_table=cpgIslandExt&hgsid=1956573466_K6emxI9N7oynnWsuT8Zjnyk9XX6n)) in hg38 genome version. This dataset contains CGIs “masked” that do not contain repetitive elements. CGIs in Epromoter or control promoters were identified by using the Bedtools (2.31.0). The coverage of G-quadruplexes (G4) in Epromoters or control promoters was calculated as the percent of base pairs covered by predicted G4 annotations. These annotations and G4 Hunter scores are obtained from the G4Hunter algorithm described in Bedrat et al. (Bedrat et al., 2016) by using the threshold score 1. The statistical significance was calculated by R with the Kolmogorov-Smirnov test.



## **TF binding analysis**

Transcription factor (TF) binding sites data were collected from the JASPAR (2022) database (Castro-Mondragon et al., 2022), which was downloaded in bigbed format (<http://hgdownload.soe.ucsc.edu/gbdb/hg38/jaspar/JASPAR2022.bb>) from the UCSC genome track with a score of P-value for each binding site. All the TF binding sites were filtered by a score higher than 400 ( $P\text{-value} \leq 10e-4$ ). The filtered TF binding sites were overlapped with Epromoters by bedtools intersect. Each TF binding site was associated with a corresponding TF family according to the supplemental data from Castro-Mondragon et al. (Castro-Mondragon et al., 2022). The TF binding site family density was calculated by the binding sites of TF families at each Epromoter. The TF binding site family diversity was calculated by the number of TF families at each Epromoter. The same analysis was performed for control promoters. The TF binding data was collected from ReMap (2022) (Hammal et al., 2022). We used the ReMap datasets which include 68.2 million non-redundant ChIP-seq peaks from 1210 TFs in humans (<https://zenodo.org/records/10527088>). The non-redundant ChIP-seq peaks were overlapped with Epromoters to quantify the number of peaks per Epromoter. 737 cell lines and tissues associated with the ChIP-seq peaks were classified into 18 biotypes to describe TF diversity (Hammal et al., 2022). The same analysis was also performed for control promoters. The odds ratio and P-value were calculated for each TF between Epromoters and control promoters by the number of ChIP-seq peaks, as the description of TFs binding enrichment at Epromoters. The UMAP analysis was performed by the R package umap, which is based on a matrix of each TF binding state (ReMap) at each Epromoter or control promoter (Value 1 is defined as binding, and value 0 is defined as no-binding). Then each Epromoter or control promoter was quantified by the TF binding peak density.

## **CAGE data analysis**

The CAGE data were collected from FANTOM5 (Consortium et al., 2014; Kanamori-Katayama et al., 2011; Lizio et al., 2015). The CAGE peaks were downloaded in bed format with hg38 ([https://fantom.gsc.riken.jp/5/datafiles/reprocessed/hg38\\_v7/extra/CAGE\\_peaks/](https://fantom.gsc.riken.jp/5/datafiles/reprocessed/hg38_v7/extra/CAGE_peaks/)), which was identified by DPI (decomposition-based peak identification, Forrest et al 2014) across all the tissues in FANTOM5. The CAGE signal data were downloaded

from the UCSC track in bigwig format (<https://hgdownload.soe.ucsc.edu/gbdb/hg38/fantom5/ctssTotalCounts.fwd.bw>, <https://hgdownload.soe.ucsc.edu/gbdb/hg38/fantom5/ctssTotalCounts.rev.bw>), which include the total reads count by strand across all tissues from FANTOM5. In the CAGE signal analysis, we defined forward signal as direction (strand) consistent between the CAGE signal and genes and reverse signal as inconsistent. Epromoters were extended to 500bp upstream and downstream of TSS to cover the forward and reverse signals around TSS. The stranded sense and antisense CAGE peaks were overlapped with the extended regions of Epromoters to address the directionality. The CAGE signal in bigwig was overlapped with the extended regions of Epromoters by strand separately to quantify the transcription initiation. And the same analysis was performed for control promoters.

### **RNAPII data analysis**

RNAPII data was collected from de Langen et al. (de Langen et al., 2023) (<https://zenodo.org/records/8091826>), which include RNAPII consensus peaks identified from 900 RNAPII ChIP-seq experiments in normal tissues and cancer samples. The RNAPII consensus peaks were overlapped with Epromoters to quantify the RNAPII enrichment from different tissues and samples. The same analysis was performed for control promoters.

### **Common SNPs and rare SNPs collection**

The total SNPs (660,146,174 SNPs) were downloaded from SNPdb in VCF format in hg38 ([https://ftp.ncbi.nlm.nih.gov/snp/organisms/human\\_9606/VCF/00-All.vcf.gz](https://ftp.ncbi.nlm.nih.gov/snp/organisms/human_9606/VCF/00-All.vcf.gz)). The common (37,302,978) and rare (45,894,070) SNPs were filtered by minor allele frequency (MAF) of more or less than 1% according to 1000 genomes allele frequency, respectively. The common and rare SNPs were overlapped with Epromoters and control promoters by bedtools intersect.

### **GWAS analysis**

186120 GWAS variants associated with 4138 GWAS traits were collected from the NHGRI-EBI GWAS Catalog (v1.0.2) (<https://www.ebi.ac.uk/gwas/api/search/downloads/alternative>) (Sollis et al., 2023). SNPs without rsID and genomic coordinates were removed. The human common

SNPs were downloaded from 1000 Genomes Project (v5a) in vcf format ([http://ftp.ensembl.org/pub/data\\_files/homo\\_sapiens/GRCh38/variation\\_genotype/](http://ftp.ensembl.org/pub/data_files/homo_sapiens/GRCh38/variation_genotype/)) (Genomes Project et al., 2015), which were filtered by Plink (v1.9) (Chang et al., 2015) from 5 super populations (European, African, American, East Asian, South Asian) using the following parameters --geno 0.05 --maf 0.01 --hwe 1e-6. The lead SNPs from GWAS Catalog were linked with common SNPs from 1000 Genomes Project by Plink with parameters of --ld-window-kb 1000 --ld-window-r2 0.8, allowing to retrieve SNPs within 1 Mb in high linkage disequilibrium ( $r^2 > 0.8$ ) of each lead SNP. Then these linkage disequilibrium SNPs associating with GWAS (GWAS-SNPs) were overlapped with Epromoters and control promoters. Each GWAS study is assigned a unique GWAS trait with an Experimental Factor Ontology (EFO) ID and a corresponding GWAS category from the EFO database (<https://www.ebi.ac.uk/ols4/ontologies/efo>) (Malone et al., 2010). Each GWAS trait was mapped into a GWAS category (parent trait), here including 17 categories according to the EFO database. The number of GWAS traits associated with each promoter was counted by the total non-redundant GWAS traits of different SNPs at the same promoter. The GWAS trait enrichment was calculated by the ratio of SNPs associating each GWAS trait between Epromoters or control promoters versus whole genome (hypergeometric test). The GWAS trait enrichment was also compared between Epromoters and control promoters (Chi-Squared test). Partitioned heritability was calculated by LD score regression (LDSC) (Finucane et al., 2015). We calculated the partitioned heritability of 176 GWAS summary statistics (<https://console.cloud.google.com/storage/browser/broad-alkesgroup-public-requester-pays/LDSCORE?pageState>) in Epromoters and control promoters, also including the partitioned regions Enhancer\_Andersson, Promoter\_UCSC, Coding\_UCSC annotated by the baseline model of LDSC. For each GWAS study, the partitioned heritability described how much genetic contribution by different partitioned regions.

### **eQTL data analysis**

The eQTL data was downloaded from the fine mapped credible sets in eQTL Catalogue (Kerimov et al., 2021) ([https://www.ebi.ac.uk/eqtl/Data\\_access/](https://www.ebi.ac.uk/eqtl/Data_access/)), which used the fine mapping model SuSiE (Wallace, 2021). The eQTL data include 9137260 eQTLs identified from 96 tissues or cell types. These eQTLs overlapped with Epromoters and control promoters. The eQTLs associated with different target genes

from different tissues were merged into a non-redundant eQTL list. Then, the eQTLs were associated with the GWAS traits by the coordinates overlapping between eQTLs and GWAS-SNPs. We classified the merged eQTL list into 3 categories by the distance between eQTLs and the TSS of target genes, including proximal eQTLs, distal eQTLs, and proximal and distal eQTLs. The proximal eQTLs were defined as located less than 2 kb from the TSS of all target genes. The distal eQTLs were defined as located more than 2 kb from the TSS of all target genes. The proximal and distal eQTLs were defined as including both proximal and distal target genes. The eQTL heatmap in the Epromoter instances was performed according to the z-score of effect genes associating with each eQTL in different tissues from eQTL Catalogue.

### **MPRA resource collection**

Massively parallel reporter assays (MPRA) data were collected from 17 published studies (Abell et al., 2022; Bourges et al., 2020; J. Choi et al., 2020; Cooper et al., 2022; Hansen et al., 2023; Kalita et al., 2018; Khetan et al., 2021; Liu et al., 2017; Long et al., 2022; Lu et al., 2021; Mattioli et al., 2019; Mouri et al., 2022; Myint et al., 2020; Tewhey et al., 2016a; Ulirsch et al., 2016; van Arensbergen et al., 2019; Zhang et al., 2018), including 24 MPRA datasets from 14 human cell lines (Supplemental Table S7). We collected the SNPs tested in MPRA from the supplemental data of each study. The assessed SNPs were filtered by the allelic impact thresholds described in the original studies. This resulted in 37829 SNPs with significant allelic impact overlapped.

### **SNP-SELEX collection**

The SNP-SELEX data was collected from (Yan et al., 2021), which systematically assessed the binding of 270 human transcription factors to 95,886 noncoding variants in the human genome using an ultra-high-throughput multiplex protein–DNA binding assay. In the original results, 11,079 SNPs exhibited significantly differential binding to at least one transcription factor. We collected these SNPs with transcription factor binding effect to overlap with Epromoters and control promoters.

### **TF binding effect analysis**

The TF binding effect analysis of SNPs was analyzed by ANANASTRA (Boytssov et al., 2022) (<https://ananastra.autosome.org/>) which is based on allele-specific binding data from ChIP-Seq. The SNPs at Epromoters and control promoters were loaded into

ANANASTRA for analysis by rsID. The parameter of ANANASTRA was the default on the website. Additionally, we used SNP2TFBS (Kumar et al., 2017) (<https://epd.expasy.org/snp2tfbs/>) and FABIAN-variant (Steinhaus et al., 2022) (<https://www.genecascade.org/fabian/>) which based on position weight matrix (PWM) to predict the TF binding effect. The parameters of SNP2TFBS were used default in the websites. The results of FABIAN-variant were filtered by the absolute value of the prediction score over 0.5 for each motif.

### **Luciferase reporter assays**

Luciferase vectors were generated by GeneCust, inserting the 726 bp OAS3 promoter region (hg38 chr12:112,938,128-112,938,853) with the 5 minor alleles or the 5 major alleles into pGL4.12 luc2cp using KpnI-XhoI sites to assess promoter activity, and into pGL4 sv40 luc2cp (Santiago-Algarra et al., 2021) using BamHI-Sall sites to assess enhancer activity. Sequences of the plasmids are available in Supplemental Table S8. For K562,  $3 \times 10^6$  cells were spun down per plasmid transfection (3 replicates), and cells were washed with PBS and resuspended in 30  $\mu$ l Buffer R of the Neon transfection kit (Thermo Fisher Scientific). 1  $\mu$ g of the plasmid to be tested, and 200 ng of Renilla was transfected per  $1 \times 10^6$  cells in triplicate with the 10  $\mu$ l NEON tip using the following settings; Voltage: 1450, ms: 10, pulses: 3.  $1 \times 10^6$  transfected cells were transferred to 2 mL prewarmed medium in a 12-well plate. After 18 hours, 1 mL of each transfection was transferred to a new 12-well plate, allowing 1 mL of cells as non-stimulated control, and 1 mL to be treated with human recombinant IFN $\alpha$  protein (100 ng/mL) (Abcam ab9642) for 6 hours. For A549,  $0.25 \times 10^6$  of cells were seeded in a 12-well plate 24 hours before transfection. At 90% confluence, the following day, 1  $\mu$ g of each of the 4 plasmids (promoter and enhancer tests of major and minor OAS3 haplotype) and 200 ng Renilla were transfected in 6 wells using the Lipofectamine 3000 (Thermo Fisher Scientific, L3000008) protocol. 24 hours after transfection, 3 wells were treated with IFN $\alpha$  (100 ng/mL) for 6 hours, leaving 3 wells per plasmid untreated as non-stimulated controls. After 6 hours of IFN $\alpha$  stimulation, the cells were washed with 1X PBS, and resuspended in 350  $\mu$ l lysis buffer of the Dual-Glo Luciferase Assay kit (Promega E2920). After 15 minutes of incubation in lysis buffer, cells were spun down and 20  $\mu$ l supernatant was transferred to the luminescence plate reader. Luciferase signal was measured by the addition of 100  $\mu$ l Luciferin, followed by Renilla signal measurement by the addition of 100  $\mu$ l STOP&GLO (Promega E2920). The transfection of 3

replicates was repeated once in a separate experiment (to give a total of 6 samples per construct). For data analysis of the luciferase assays, luciferase values were normalized to the Renilla luciferase activity to control for between-well transfection efficiency. For each construct, readings from different days were merged by normalizing the activity of reporters to the minor allele only vector (reference allele).

### **Allelic-specific CapSTARR-seq analysis**

Using the BAM files of the CapSTARR-seq data from K562 and CCRF-CEM cell lines with and without IFN $\alpha$  stimulation, the number of reads containing the minor (T) or major (C) allele of the rs1156361 SNP was quantified using the IGV web tool (Thorvaldsdottir et al., 2013). Average read numbers from two replicates were calculated, and the reads were normalized to the no-stimulation condition for each allele.

### **Data availability**

The raw sequencing data and processed files of CapSTARR-seq generated in this study have been submitted to the Gene Expression Omnibus (GEO) under the accession GSE268615. All generated and publicly available datasets are listed in the Supplemental Table S1. The supplemental tables and descriptions can be accessed by the cloud link: <https://amubox.univ-amu.fr/s/FcapcWqFM8gED3E>.

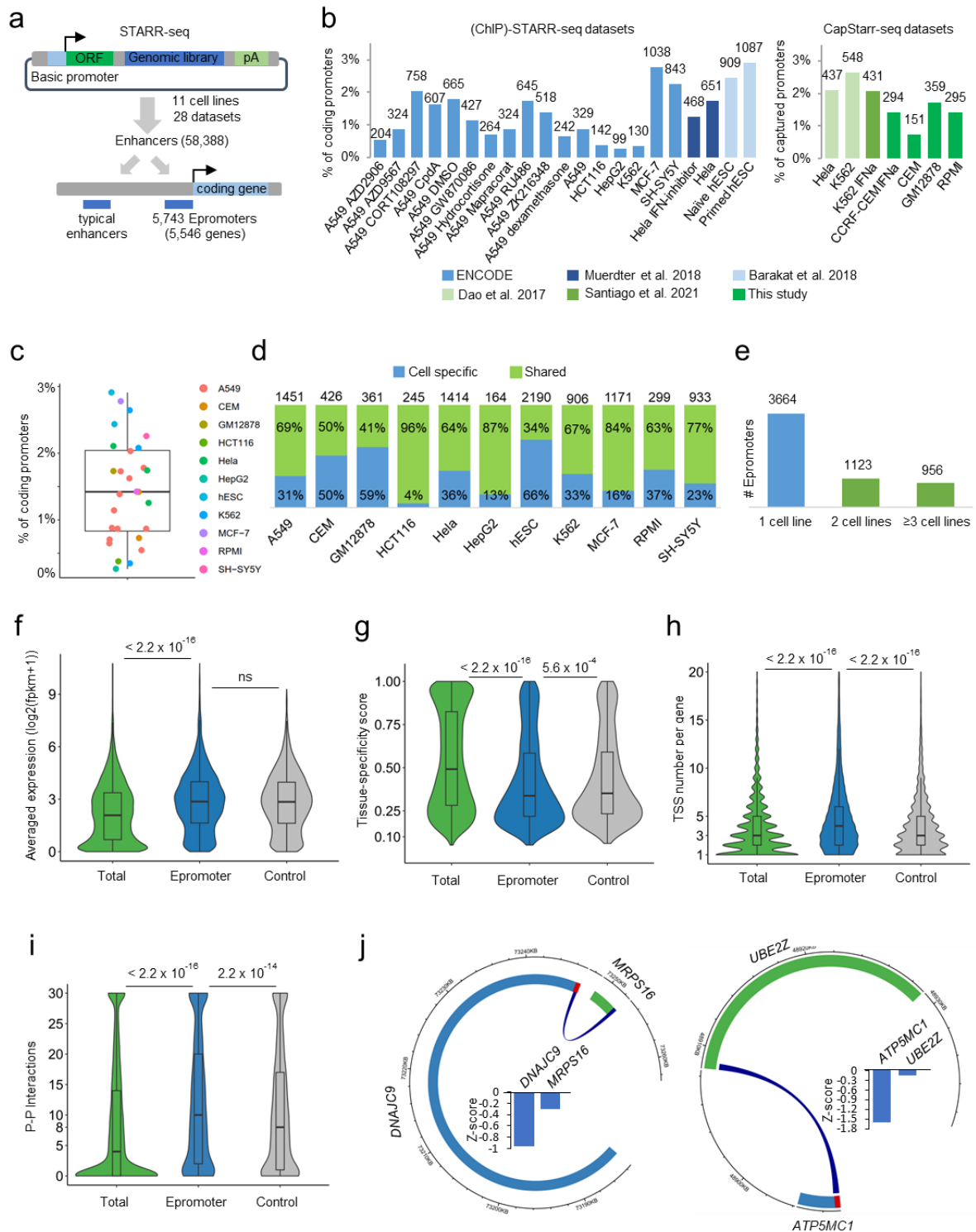
### **Code availability**

All custom scripts used in this study are available at GitHub ([https://github.com/jingwan/Epromoter\\_GWAS](https://github.com/jingwan/Epromoter_GWAS)).

### **Contributions**

SS, JW and AvO designed the study. JW performed all bioinformatics work. AvO performed all experimental work. JCM and BB provided ReMap resources. CH and JCA provided genomic G4 annotations. LP performed capSTARR-seq in CCRF-CEM. SS, JW and AvO analyzed the results and wrote the manuscript.

# Figures



**Figure 1. A comprehensive dataset of human Epromoters.**

(a) A schematic diagram illustrating the strategy to identify Epromoters from the (Cap)STARR-seq data.

(b) The percentage and number of promoters identified as Epromoters identified in each (Cap)STARR-seq dataset are indicated. The legend at the bottom describes the source of datasets in corresponding colors.

(c) The boxplot shows the percent distribution of promoters identified as Epromoters in each dataset. Each dot means one dataset. The colors represent the cell lines of each dataset.

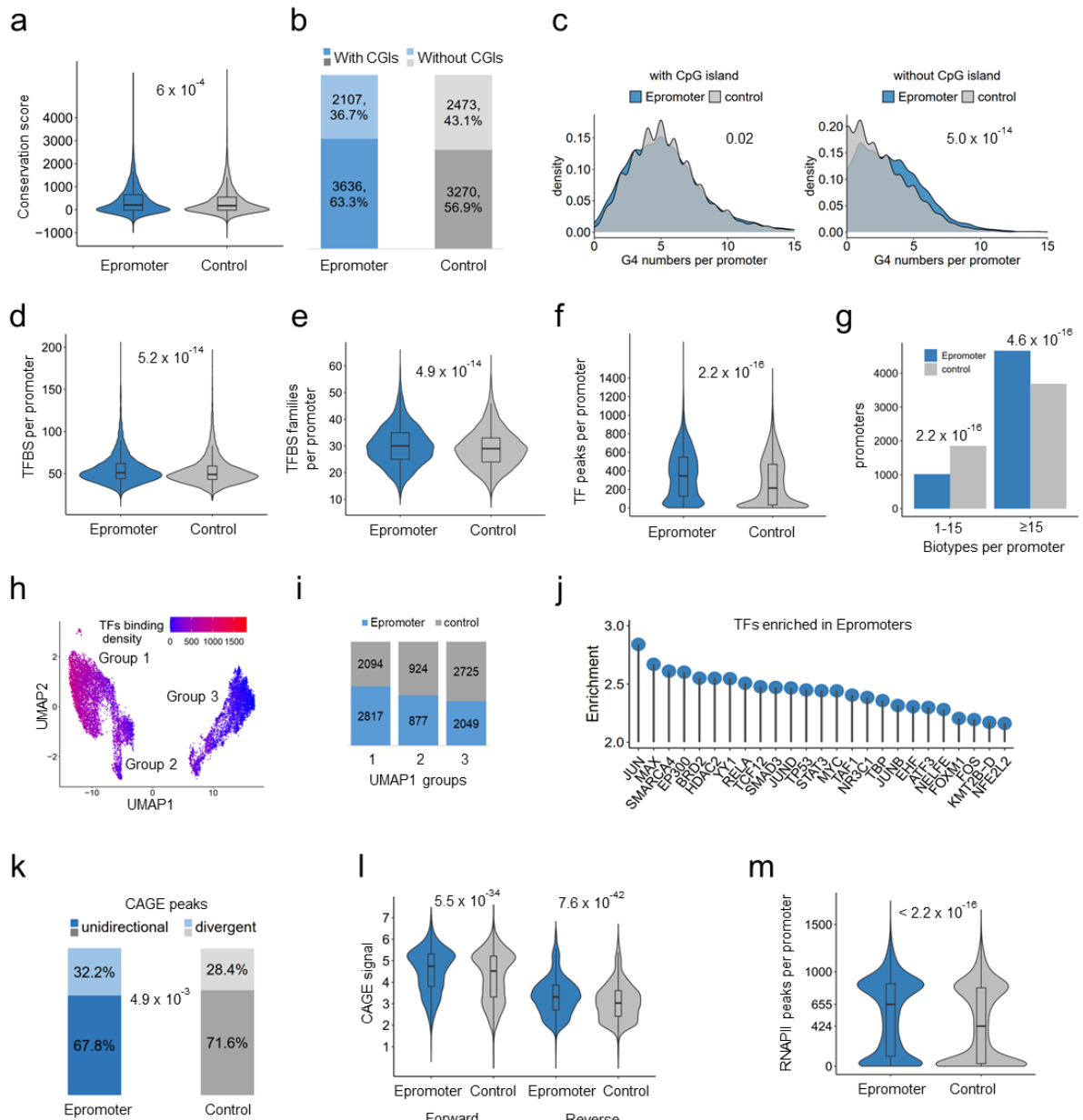
(d) The bar plots show the percentage of Epromoters found in only one cell line (blue) or shared between two or more cell lines (green). The number of Epromoters in each cell line is shown at the top of each bar.

(e) The bars show the number of Epromoters found in the indicated number of cell lines.

(f-i) Violin plots displaying the average gene expression level (f), tissue specificity score (g), the number of TSS per gene (h) and the promoter-promoter interactions (i) of all protein-coding (Total), Epromoter-associated, and control genes. The expression for each gene in (f) was calculated by the average level across 30 human tissues from GTEX. *P*-values were calculated by a Wilcoxon test (ns: not significant).

(j) Two examples of consistent P-P interactions and CRISPRi-mediated regulation of distal genes by Epromoters. The plots show the circular visualization of Epromoters and interacting genes based on their genomic locations. The blue bar is the Epromoter-associated gene. The other genes are shown in green. The short red bar represents the Epromoter. Genes in the outer circle are in the positive strand. Genes in the inner circle are in the negative strand. The blue curves are P-P interactions. The inset plots display the Z score values of the Perturb-seq experiments (Replogle et al., 2022).





**Figure 2. Epromoters display specific genomic/epigenomic features.**

(a) Conservation score of Epromoters and control promoters, which were retrieved from 241 Zoonomia Placental Mammals. The conservation score of each promoter was calculated as the sum of all the bases in the region. The positive scores were predicted as conserved. The negative scores were predicted as fast-evolving. Statistical significance was assessed by a Wilcoxon test.

(b) CpG islands (CGIs) enriched in Epromoters. The bar plots show the percentage and number of Epromoters and control promoters with CpG islands (CGIs) and without CGIs. Statistical significance was assessed by a Chi-squared test.

(c) G4 numbers per promoter of Epromoters and control promoters with or without CGI. The density means the distribution of Epromoters or control promoters, which display

the enrichment of Epromoters or control promoters. Statistical significance was assessed by a Kolmogorov-Smirnov test.

(d-e) Violin plots displaying the number of TF binding site (TFBS) families per promoter (d; i.e., density) and the number of different TFBS families per promoter (e; i.e. diversity) using the JASPAR database. Statistical significance was assessed by a Wilcoxon test.

(f) Violin plots displaying the number of TF binding peaks per promoter identified by ChIP-seq using the ReMap resource. Statistical significance was assessed by a Wilcoxon test.

(g) Number of different tissues (Biotypes) of ChIP-seq peaks associated with Epromoters and control Epromoters as classified by ReMap. *P-values* were calculated by a Chi-squared test.

(h) Dimension reduction by Uniform Manifold Approximation and Projection (UMAP) based on TF binding (ReMap) at each promoter. The color is displaying the TF binding density at each promoter. Three groups were manually separated based on UMAP1 dimension.

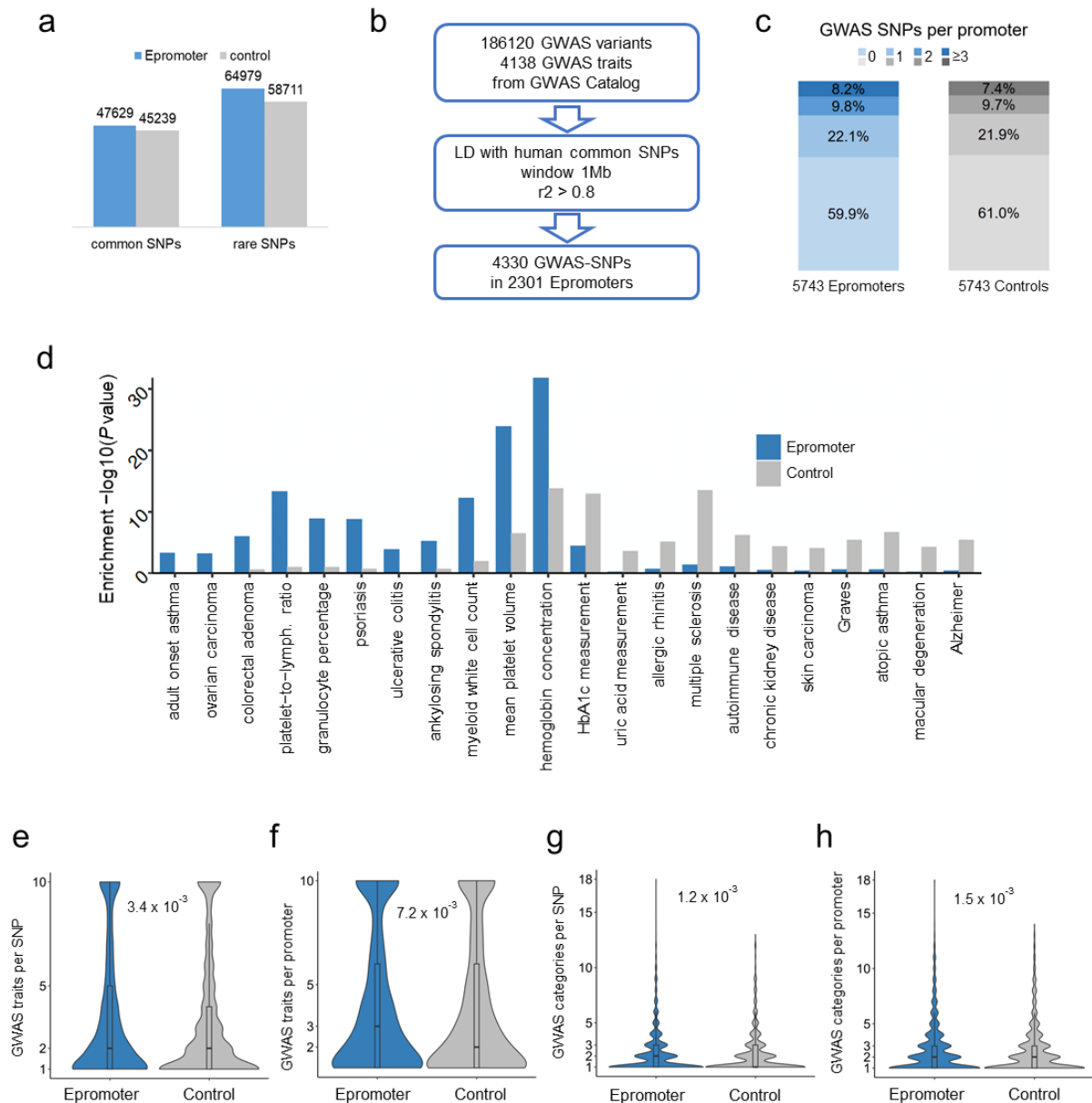
(i) Number of Epromoters and control promoters in each UMAP group defined in Figure 2h.

(j) Top 25 TFs enriched at Epromoters, compared with control promoters. The height of the lollipop represents the odds ratio of TFs binding frequency between Epromoters and control promoters.

(k) The percentage of unidirectional and divergent promoters as assessed by CAGE peaks. *P-values* were calculated by Chi-squared test.

(l) Violin plots displaying the forward and reverse CAGE signal in function of the genomic orientation of the promoters. Statistical significance was assessed by a Wilcoxon test.

(m) Violin plots displaying the number of RNAPII ChIP-seq peaks overlapping Epromoters and control promoters. Statistical significance was assessed by a Wilcoxon test.



**Figure 3. Genetic variation associated with Epromoters.**

(a) Number of common and rare SNPs overlapped with Epromoters and control promoters.

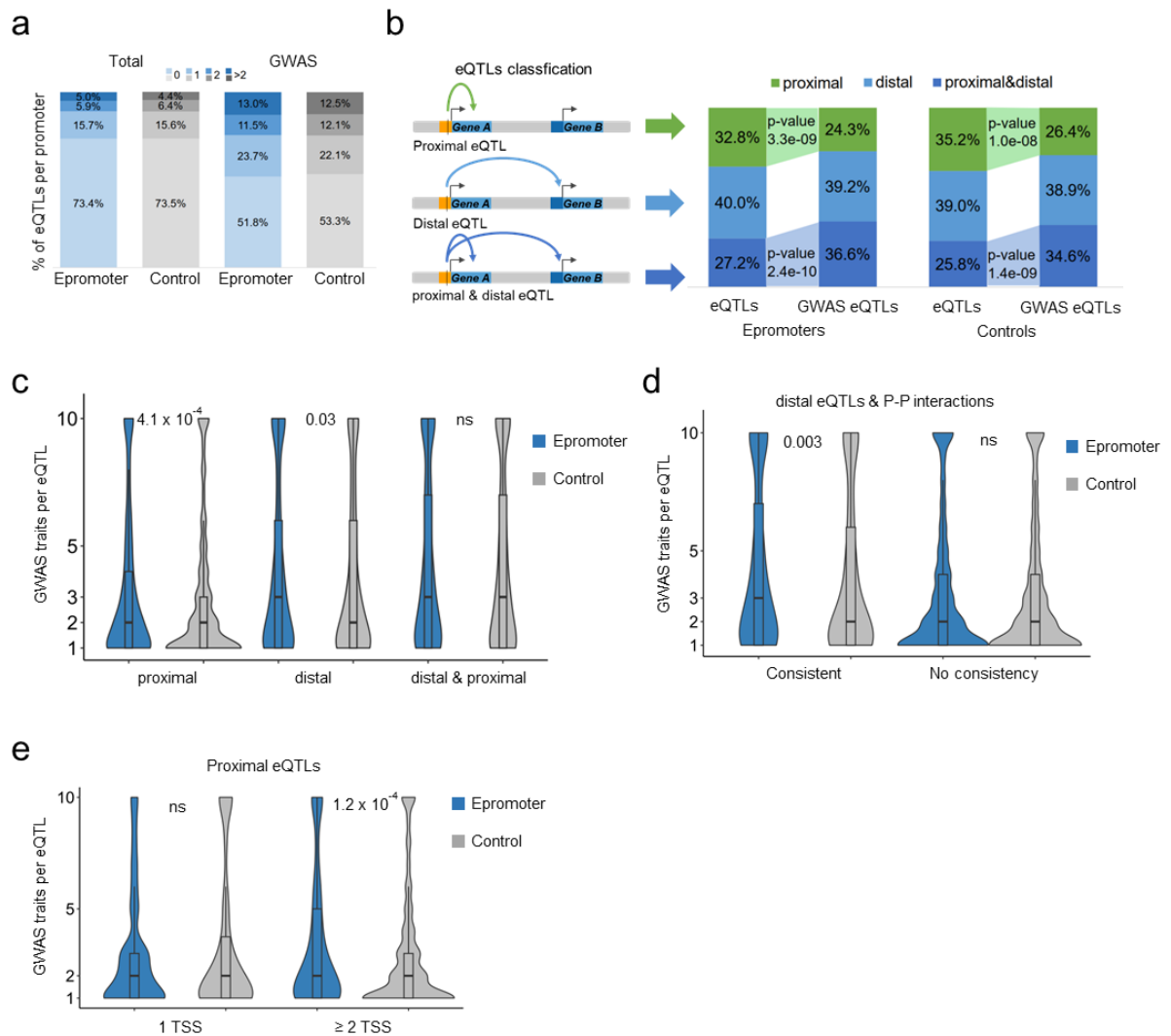
(b) Scheme to identify GWAS-SNPs in Epromoters. First, 186120 tag SNPs associated with 4138 GWAS traits were collected from the GWAS Catalog. The SNPs from the GWAS Catalog were linked with common SNPs by a stringent linkage disequilibrium (LD) threshold ( $r^2 > 0.8$ ) within 1 Mb. Then the LD SNPs associated with GWAS (GWAS-SNPs) were overlapped with Epromoters. Finally, 4330 GWAS-SNPs were found in 2301 Epromoters.

(c) Distribution of GWAS-SNPs per Epromoters or control promoters.

(d) GWAS traits differentially enriched in Epromoters. The GWAS trait enrichment was calculated by the ratio of SNPs associating each GWAS trait between Epromoters or control promoters *versus* the whole genome. The  $P$  values for enrichment were calculated by the hypergeometric test. Only differentially enriched GWAS traits between Epromoters and control promoters and associated with a known GWAS category are shown in the plot. Statistical significance for the difference was assessed by Chi-squared test.

(e-f) Violin plots displaying the number of GWAS traits per SNP (e) and per promoter (f). Statistical significance was assessed by a Wilcoxon test.

(g-h) Violin plots displaying the number of GWAS categories per SNP (g) and per promoter (h). Statistical significance was assessed by a Wilcoxon test.

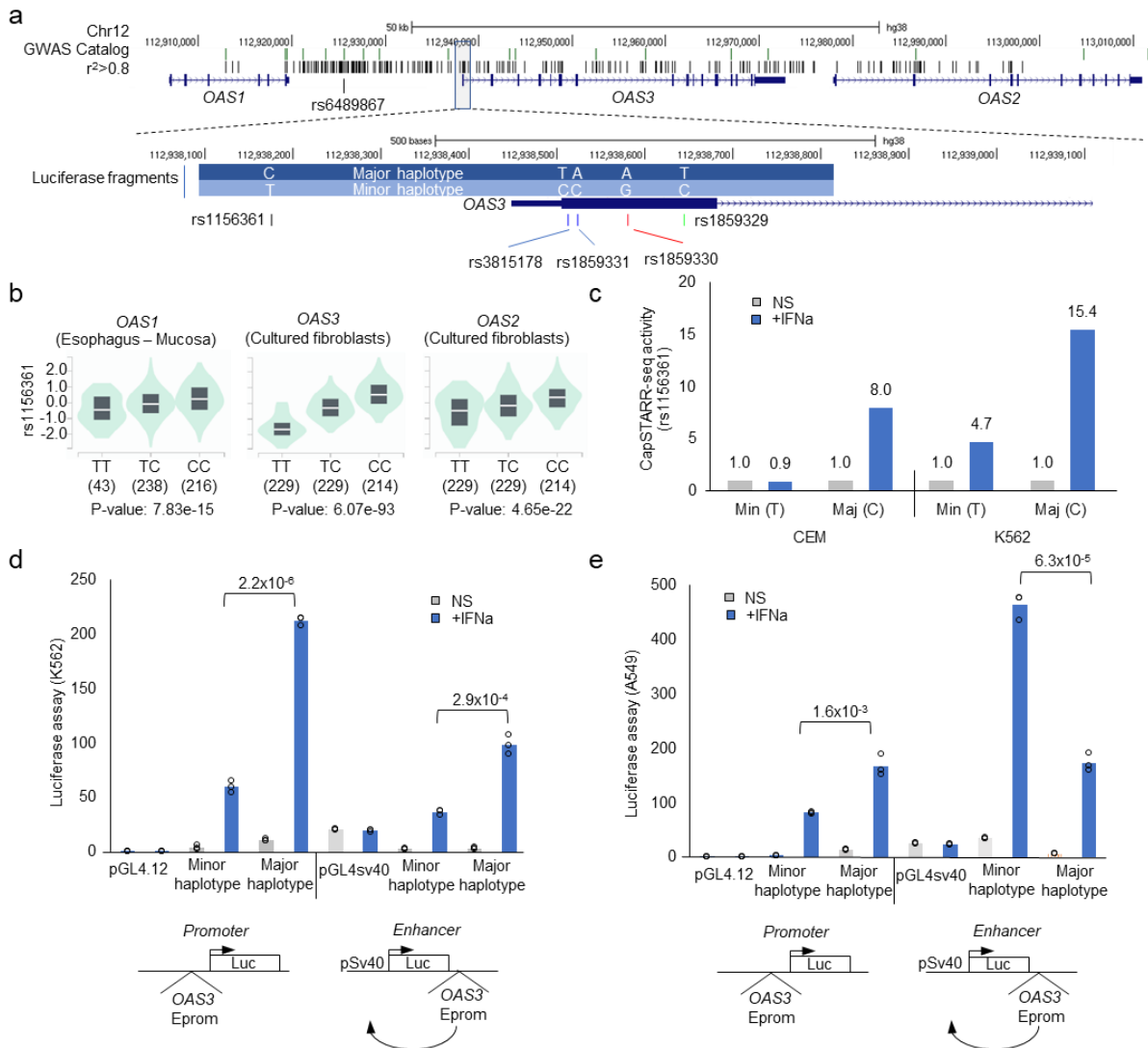


**Figure 4. Link between pleiotropy and target genes.**

(a) Percentages of Epromoters and control promoters according to the number of eQTLs per promoter and considering either all SNPs or only the GWAS-associated SNPs.

(b) The eQTLs were classified into proximal (green), distal (light blue), or proximal & distal eQTLs (dark blue) as indicated in the left panel. The right panels indicate the percentages of promoters associated with the different types of eQTLs. P-values were calculated by a Chi-Squared test.

(c-e) Violin plots displaying the number of GWAS traits per eQTL in the function of the eQTL type (c), eQTLs with distal targets consistent or inconsistent with P-P interactions (d), and the number of TSS per gene associated with proximal eQTLs (e). Statistical significance was assessed by a Wilcoxon test.



**Figure 5. A pleiotropic Epromoter variant associated with COVID-19 shows enhancer/promoter switch**

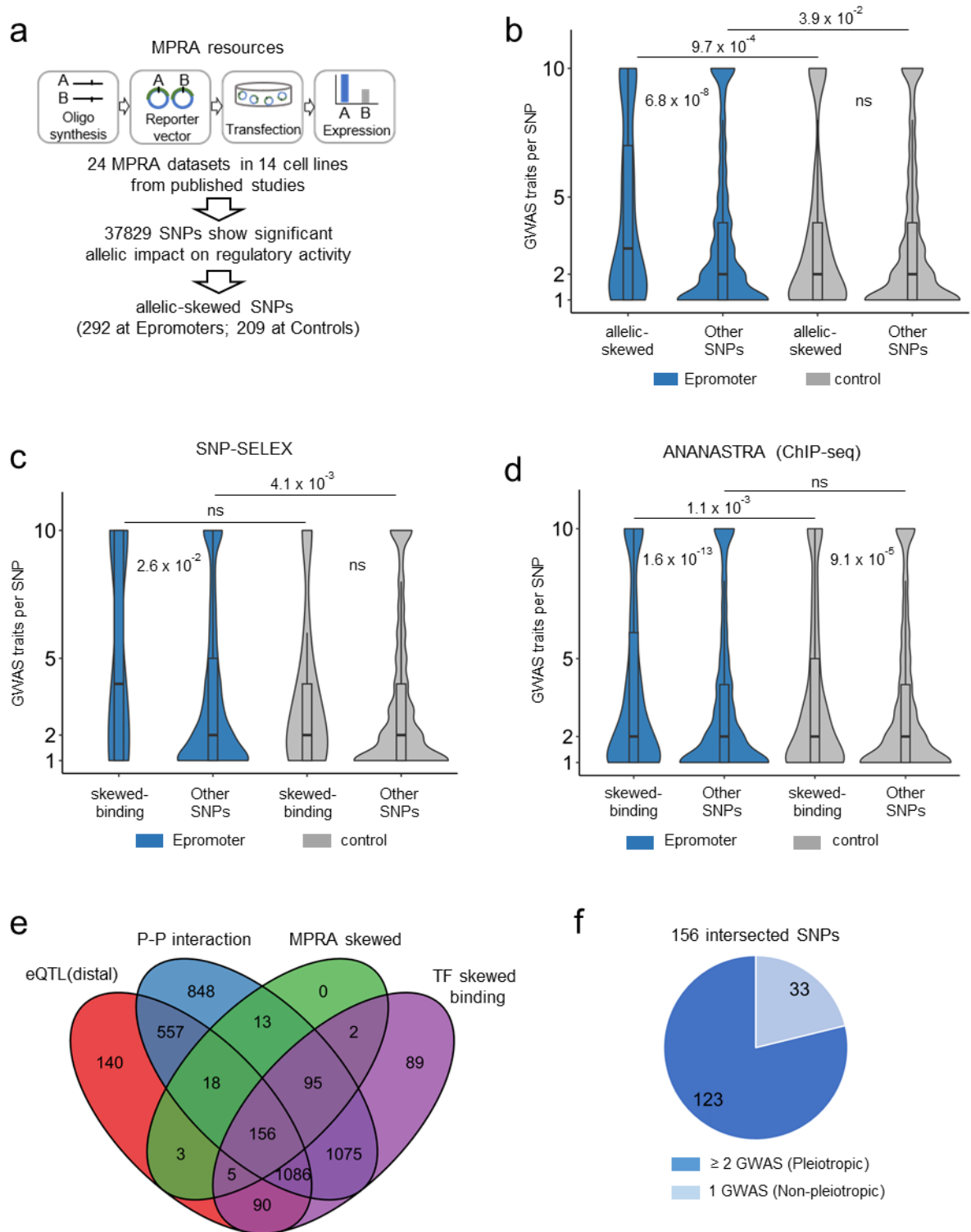
(a) UCSC browser view of OAS1/2/3 locus, with lead COVID-19 SNP rs6489867 and SNPs in LD ( $r^2 > 0.8$ ), as well as the location of the 726 bp region containing 5 SNPs in the OAS3 Epromoter analyzed in Figure 5d/e.

(b) eQTL (GTEx) of rs1156361, showing decreased expression of the OAS1/2/3 of the minor allele.

(c) CapSTARR-seq activity of the OAS3 Epromoter containing the rs1156361 minor (Min T) or major (Maj C) alleles in the CCRF-CEM and K562 cell lines with no stimulation (NS) and with 6 hours of IFNa stimulation showing increased regulatory activity of the major allele upon IFNa stimulation as compared to the minor allele in both cell lines.

(d)-(e) Luciferase reporter assays assessing the promoter (left panel) or enhancer (right panel) activity of the OAS3 Epromoter harboring the minor or major haplotypes

before and after IFNa stimulation for 6h in the K562 (d) and A549 (e) cell lines. Experiments were performed in triplicate and statistical significance was assessed by Students' t-test.



## **Figure 6. Functional validation of pleiotropic Epromoter SNPs.**

(a) Schematic strategy to identify GWAS-SNPs with allelic-skewed regulatory activity. First, 24 MPRA datasets in 14 cell lines were collected from published studies. 37831 SNPs in total show a significant allelic impact on regulatory activity. Finally, 292 allelic-skewed SNPs were overlapped within Epromoters.

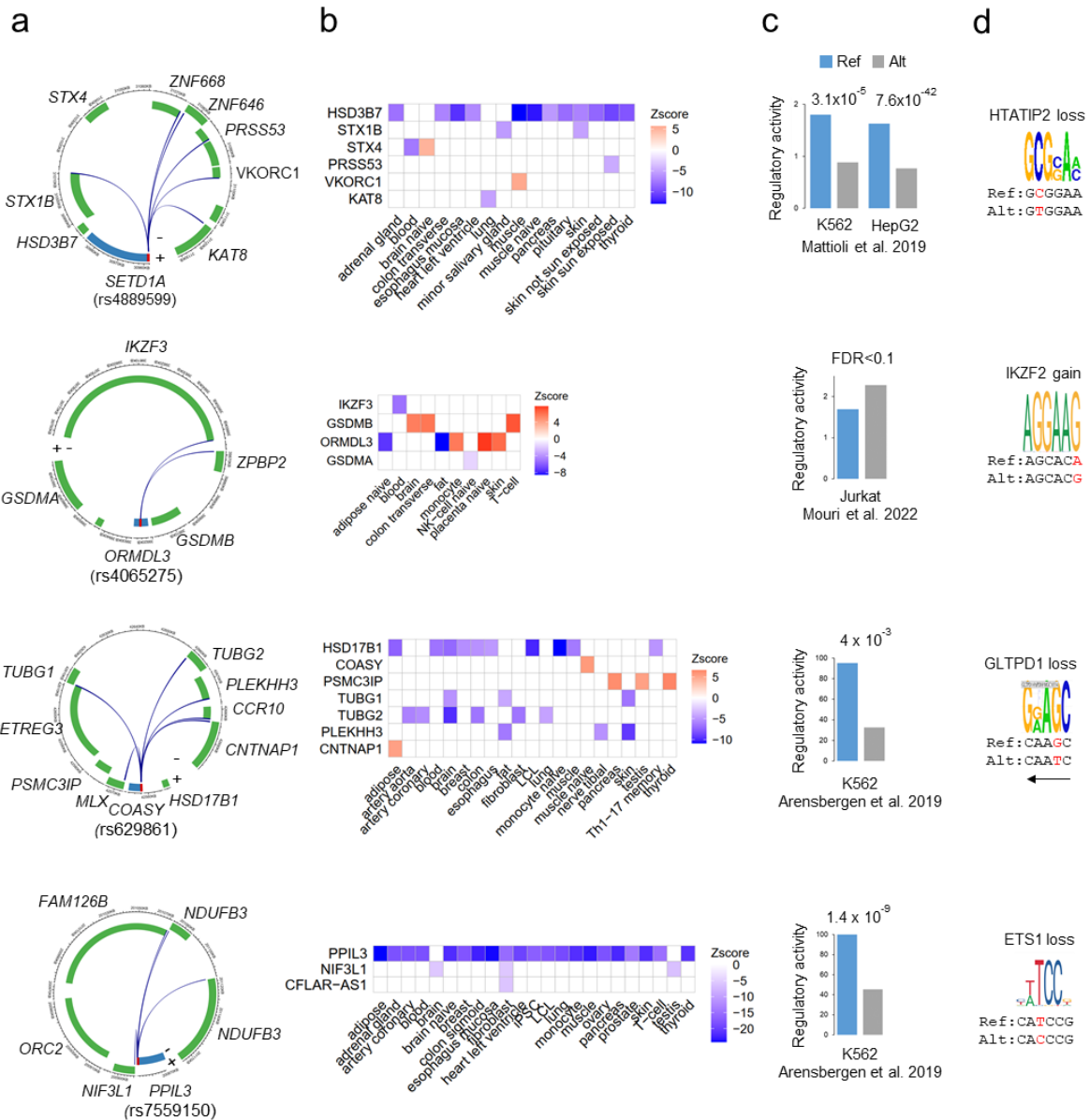
(b) Violin plots displaying the number of GWAS traits per SNP in the function of whether the SNP had an allelic-skewed regulatory activity or not (other SNPs) based on MPRA experiments. Statistical significance was assessed by a Wilcoxon test.

(c-d) Violin plots displaying the number of GWAS traits per SNP in the function of whether the SNP had a skewed TF binding based on SNP-SELEX assays (c) and ANANASTRA (d). Statistical significance was assessed by a Wilcoxon test.

(e) The Venn diagram illustrates the intersections of SNPs located at Epromoters among four categories: eQTLs with distal effects, promoter-promoter interactions, allelic-skewed SNPs identified by MPRA, and SNPs exhibiting skewed transcription factor binding.

(f) The pie chart shows the number of non-pleiotropic (1 GWAS trait) and pleiotropic ( $\geq 2$  GWAS traits) SNPs from the 156 intersected SNPs.





**Figure 7. Examples of pleiotropic Epromoters**

(a) The circular visualization of Epromoters and interacting genes based on their genomic locations. The blue bar is the Epromoter-associated gene. The other genes were shown in green. The short red bar represents the Epromoter. Genes in the outer circle are in the positive strand. Genes in the inner circle are in the negative strand. The selected SNP is indicated under the Epromoter-gene name. The blue curves are promoter-promoter interactions.

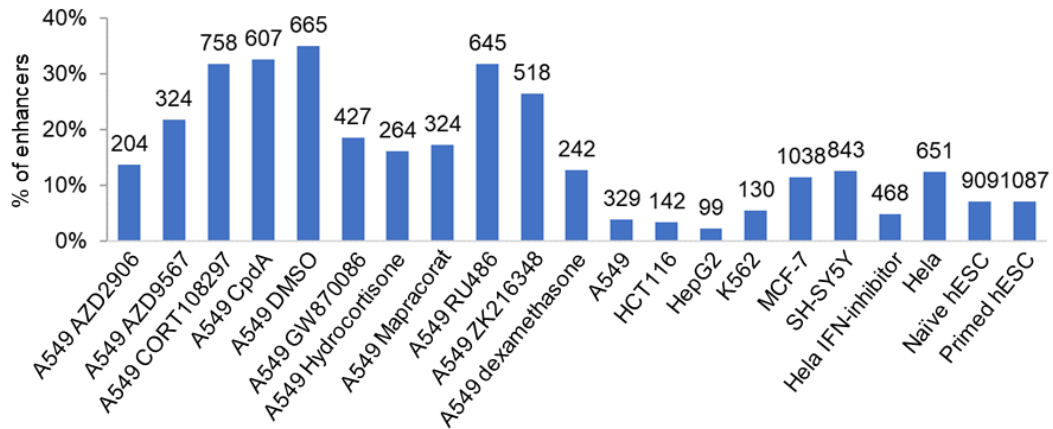
(b) The heatmaps show the eQTLs effect of the selected SNPs on target genes in different tissues from the eQTL Catalogue. Each row in the heatmap represents the gene associated with the eQTL. Each column represents the tissue of eQTL. The color bar shows the z-score of the eQTL effect on target genes.

(c) The bar plots show the allelic-skewed regulatory activity of the selected SNPs validated by MPRA. The blue bar shows the regulatory activity of the Epromoter with reference allele. The gray bar shows the regulatory activity of the Epromoter with the alternative allele. *P* values or FDR according to original studies are shown on the top.

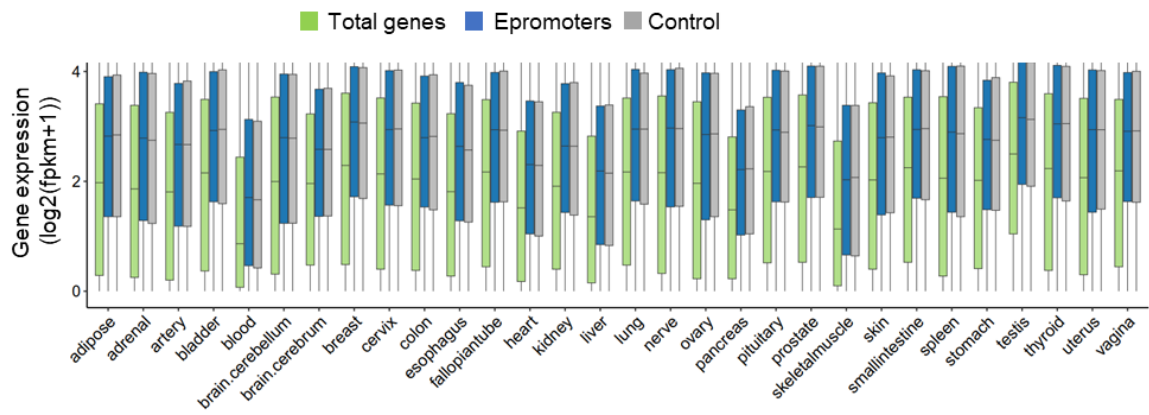
(d) Representative TF binding sites affected by the selected SNPs. The predicted consequences of the SNPs (from reference to alternative alleles) are shown at the top. The sequences of reference and alternative alleles are shown at the bottom. The SNP is shown in red. The arrow indicated the sequence is in the reverse complement.

# Supplemental Figures

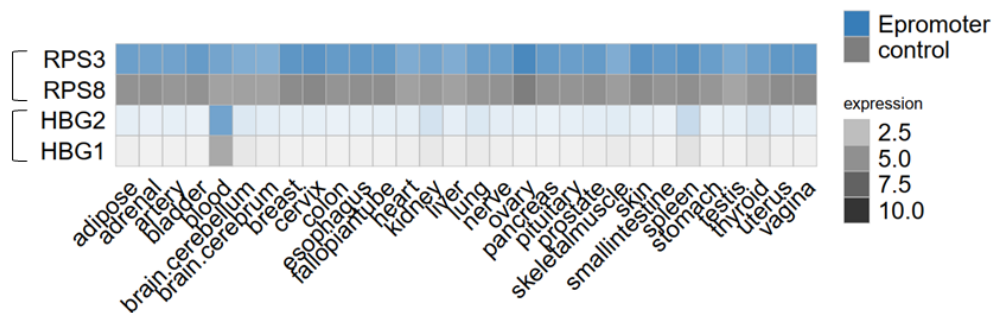
a



b



c

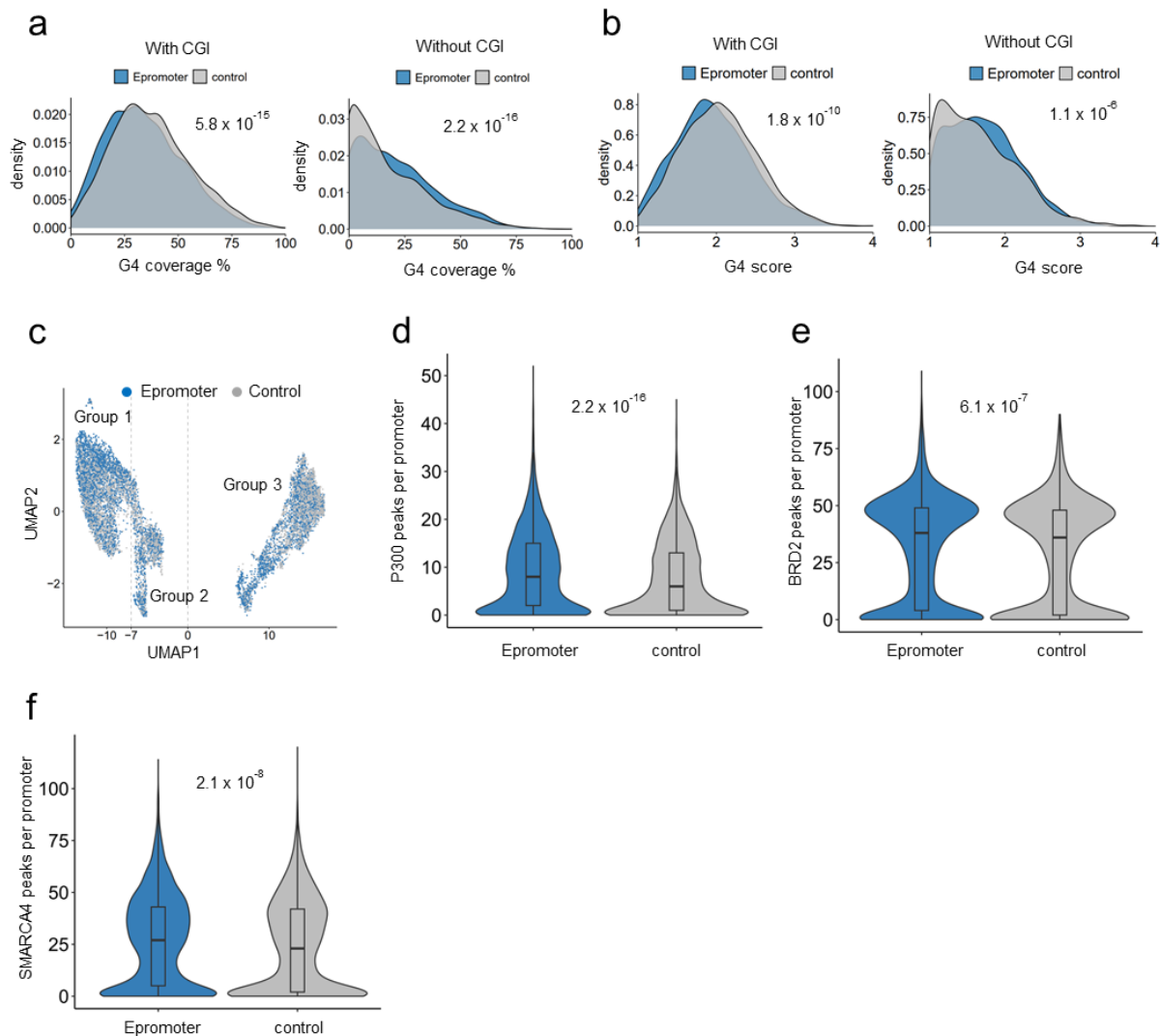


## Supplemental Figure S1

(a) Epromoters identified in 21 STARR-seq datasets. The numbers on the bars are showing the number of Epromoters in each dataset. The height of bar represents that the percentage of enhancers were identified as Epromoters.

(b) Gene expression of total genes, Epromoter genes and control genes across 30 human tissues. All tissue-specific differences in expression between the groups total genes and Epromoters are significant, and none of the expression differences between the groups Epromoters and controls are significant as tested by Wilcoxon test.

(c) Two examples show that each control gene displays the same expression profile across 30 tissues as its Epromoter counterpart.

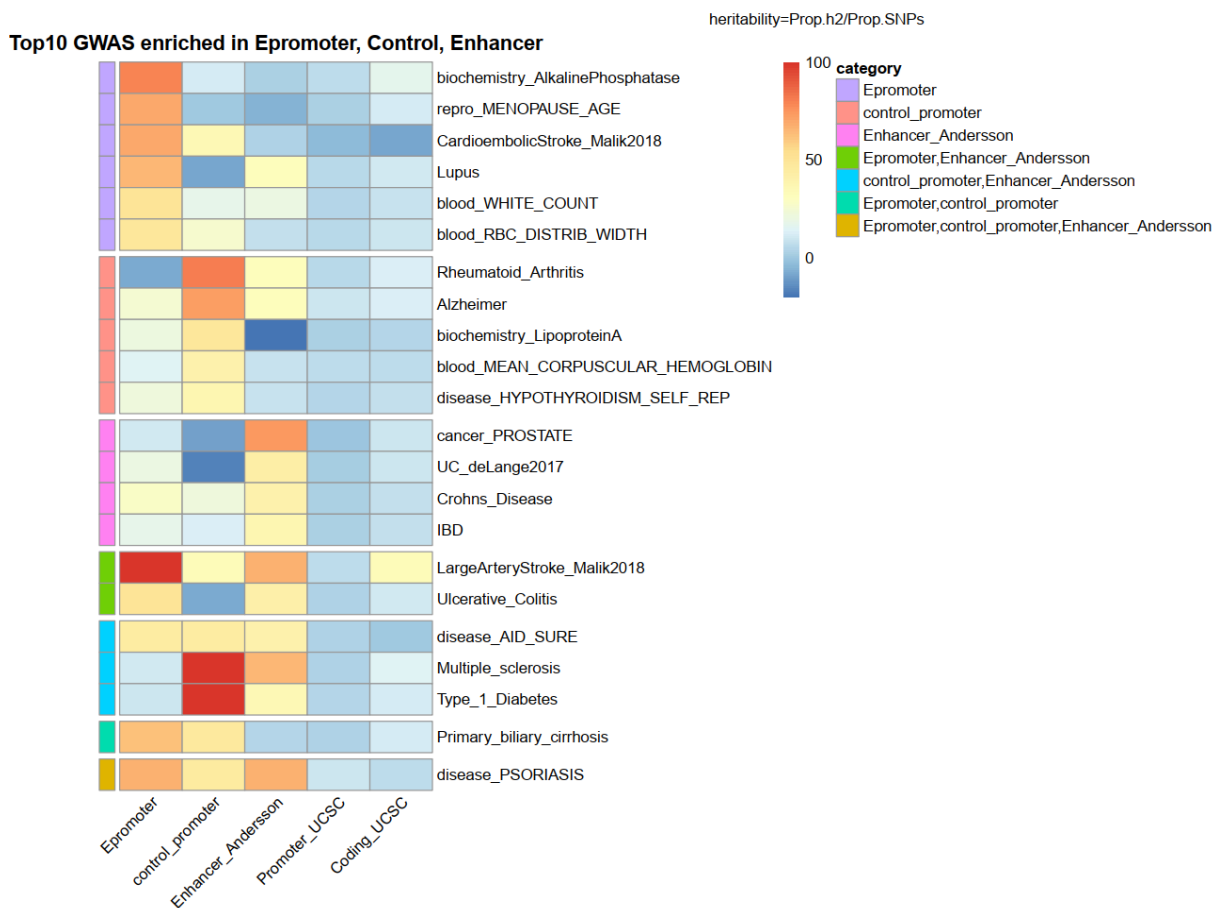


### Supplemental Figure S2

(a-b) The distribution of G4 coverage and G4 hunting scores at Epromoters and control promoters with or without CpG islands (CGI). The G4 coverage means the percentage of the promoter region covered by G4 structures. The G4 score was assessed by the G4Hunter tool which provide the likeness of DNA sequences to form a G4 structure. In panel b, only promoters with predicted G4s (G4 score  $\geq 1$ ) are shown. Statistical significance was assessed by a Kolmogorov-Smirnov test.

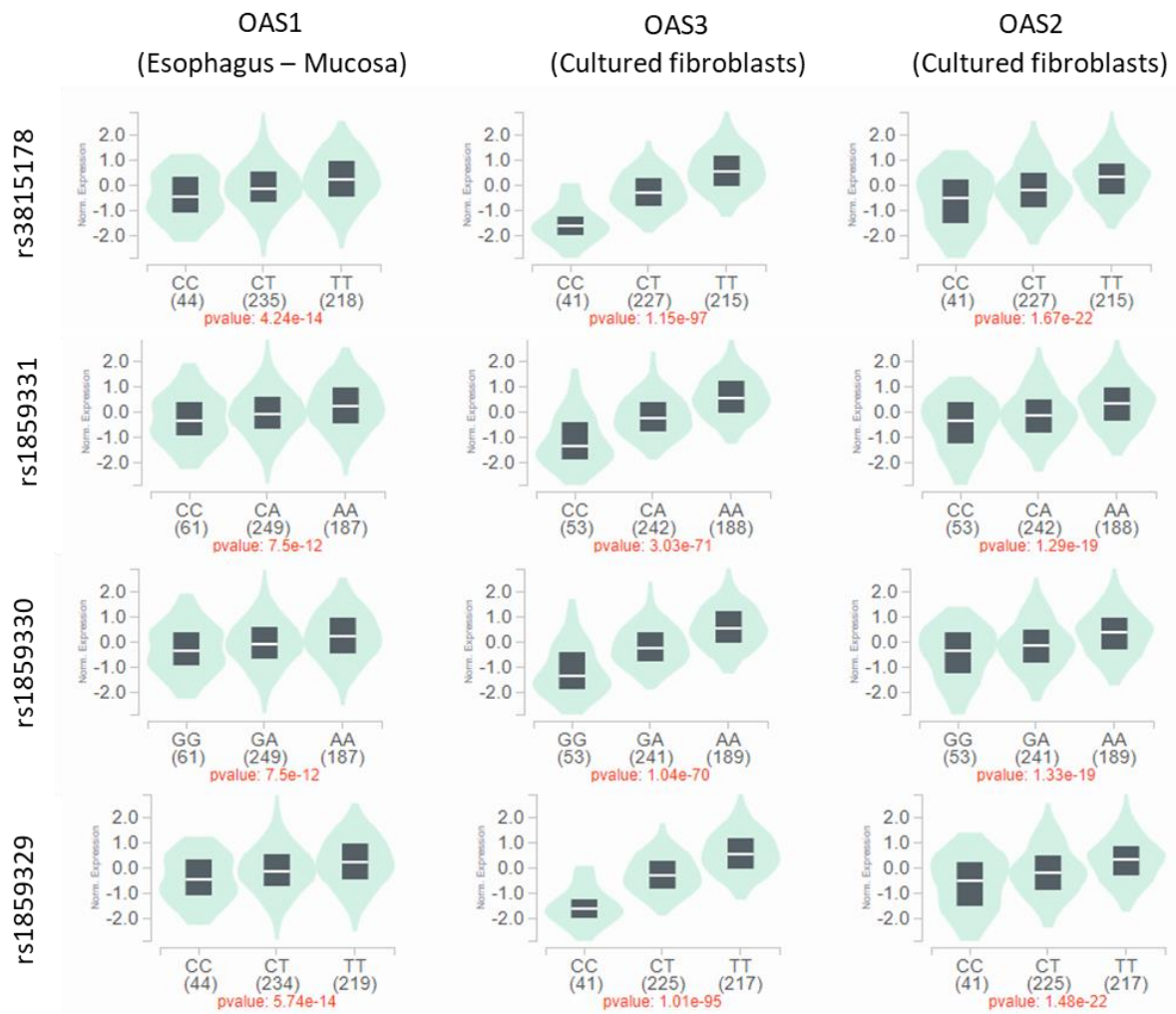
(c) Dimension reduction by Uniform Manifold Approximation and Projection (UMAP) of Epromoters (blue dots) and control promoters (grey dots) based on TF binding (ReMap) at each promoter. Three groups were manually separated based on UMAP1 dimension.

(d-f) The binding density of P300/BRD2/SMARCA4 at Epromoters and control promoters. The peaks of P300/BRD2/SMARCA4 were collected from ReMap, including 43 P300 ChIP-seq datasets, 57 BRD2 ChIP-seq datasets, and 84 SMARCA4 ChIP-seq datasets in different human cell lines or tissues. Statistical significance was assessed by a Wilcoxon test.



### Supplemental Figure S3

The partitioned heritability of GWAS enriched in Epromoters, control promoters or enhancers. Each line of heatmap is a GWAS summary statistics from LDSC. The squares of each line show the heritability of Epromoters, control promoters, enhancers from Andersson et al, promoters from UCSC, coding regions from UCSC, which calculated by LDSC baseline model. Each category shows the top 10 enriched GWAS in one or different regions.



### Supplemental Figure S4

eQTL for the 4 SNPs in high LD with rs1156361 (see also Figure 5a) in the OAS3 Epromoter.

# Supplemental information 1

## **SETD1A locus**

The *SETD1A* Epromoter is active in A549 upon 4 different conditions (DMSO, CORT108297, CpdA and RU486), as well as in naïve hESCs. Rs4889599 lies in this Epromoter, and is in LD with SNPs that are associated with 30 different GWAS (Graves' disease, Juvenile myoclonic epilepsy, Parkinson's disease, ankylosing spondylitis, psoriasis, ulcerative colitis, Crohn's disease, sclerosing cholangitis, aspartate aminotransferase measurement, serum alanine aminotransferase measurement, low density lipoprotein triglyceride measurement, body fat percentage, high density lipoprotein cholesterol measurement, sex hormone-binding globulin measurement, body fat distribution, body fat percentage, body height, body mass index, diastolic blood pressure, eosinophil count, erythrocyte count, gait measurement, heel bone mineral density, hematocrit, hip circumference, irritability measurement, lifestyle measurement, mean corpuscular hemoglobin, mean corpuscular volume, mean reticulocyte volume, multiple sclerosis, neuroticism measurement, psoriasis, psoriatic arthritis, pulse pressure measurement, response to anticoagulant, sex hormone-binding globulin measurement, total blood protein measurement, triglyceride measurement, visceral adipose tissue measurement, vitamin D measurement, waist circumference, tonsillectomy risk measurement), including immune-associated diseases (Graves, psoriasis, Crohn's, eosinophil count), neurologic diseases (Parkinson's, epilepsy, anxiety) and heart disease risk factors (BMI, blood and pulse pressure, triglycerides and LDL cholesterol measurements).

There are 30 P-P-interactions in different tissues (*AC135050.5*, *ZNF668*, *ARMC5*, *COX6A2*, *FUS*, *RP11-388M20.6*, *ITGAD*, *KAT8*, *RP11-196G11.4*, *PRSS36*, *PRSS53*, *RP11-196G11.1*, *VKORC1*, *RP11-170L3.2*, *ZNF267*, *ZNF646*, *ZNF668*, *ARMC5*, *COX6A2*, *FUS*, *ITGAD*, *ITGAX*, *PRSS36*, *RNF40*, *SLC5A2*, *STX1B*, *TGFB111*, *ZNF267*, *ZNF629*, *ZNF646*, *ZNF668*, *ZNF771*, *ZNF843*, *Orai3*, *SETD1A*), and 6 genes in eQTL with the SNP (*HSD3B7*, *STX1B*, *STX4*, *PRSS53*, *VKORC1* and *KAT8*) (Figure 7a-b). The eQTLs show that the alternative allele decreases expression of all of the eQTLs except *VKORC1*, which shows increased expression. However, *SETD1A* is not an eQTL. Moreover, MPRA from the Mattioli et col. study (Mattioli et al., 2019) found consistent decreased activity of the alternative allele of rs4889599 in K562 and

HepG2 cell lines (Figure 7c). Interestingly, the variant changes TFBS for several TFs, including HTATIP2 (Figure 7d), which is a positive regulator of transcription by RNA polymerase II, and EGR1 and EGR2 (Supplemental Table 5), transcriptional regulators that are involved in ischemia response and repression of inflammatory enhancers (Trizzino et al., 2021), which could contribute to the immune-related disease associations.

The *SETD1A* gene encodes a chromatin modifier protein involved in synaptic function and development of neurons, and mutations in this gene are associated with early-onset epilepsy (Yu et al., 2019), which makes this gene also a candidate for contribution to other neurodevelopmental disorder associations like Parkinson's and anxiety.

HSD3B7 is an enzyme involved in the synthesis of bile acids from cholesterol, but also plays a role in lymphoid cell movement by regulating a chemotactic receptor (Yi et al., 2012). As such this eQTL target could be involved in any of the immune related associations.

STX1B Syntaxin 1B plays a role in exocytosis and synaptic vesicles, and has been associated with a spectrum of epilepsy syndromes (Schubert et al., 2014), as well as Parkinson's disease (Nalls et al., 2014).

*STX4* Syntaxin 4 is a paralog of *STX1B*, is also involved in synaptic vesicle biology, and is associated with a range of cardiometabolic pathologies, including triglyceride level systolic blood pressure and body fat percentage (Martin et al., 2021; Richardson et al., 2020; Sakaue et al., 2021), which could also be the link with the cardiometabolic GWAS linked to rs4889599 and Epromoter *SETD1A*. Additionally, *STX4* is expressed in pancreatic B cells, promotes islet function (Oh et al., 2018) and could thus play a role in the diabetes association. Moreover, *STX4* is involved in cytotoxic T-lymphocyte immunological synapse formation (Spessott et al., 2017), which could mediate the auto-immune associations (e.g. Graves', Crohn's, psoriasis, but also neurodegenerative diseases like Parkinson's).

PRSS53 is predicted to have endopeptidase activity and was identified to play a role in maintaining the health of pancreatic islet b cells (Mizusawa et al., 2022). This gene is associated with psoriasis, and it is the most highly over-expressed gene in psoriatic skin (Stuart et al., 2010), potentially mediated through the allele-specific *SETD1A* Epromoter.



Overexpression of *VKORC1* was identified to lead to increased VKOR activity, which is the target of anticoagulants, thus leading to its association “response to anticoagulant drugs” that could be mediated by the SNP in Epromoter *SETD1A* (Rost et al., 2004).

### ***ORMDL3* locus**

The *ORMDL3* Epromoter is active in naïve hESCs, and SNPs rs4065275 and rs8076131 in this region are in LD with SNPs associated with over 30 GWAS hits (Crohn's disease, Eczema, allergic rhinitis, Glucocorticoid use measurement, Inhalant adrenergic use measurement, Oral ulcer, acute lymphoblastic leukemia, age at onset of asthma, allergic rhinitis, allergy, allergy age at onset, asthma, ankylosing spondylitis, psoriasis, ulcerative colitis, sclerosing cholangitis, asthma, asthma exacerbation measurement, atopic asthma, atrial fibrillation, autoimmune disease, autoimmune thyroid disease, type I diabetes mellitus, Common variable immunodeficiency, ankylosing spondylitis, psoriasis, celiac disease, ulcerative colitis, juvenile idiopathic arthritis, systemic lupus erythematosus, biliary liver cirrhosis, bipolar I disorder, blood protein measurement, cervical carcinoma, childhood onset asthma, atopic eczema, atopic march, dermatomyositis, juvenile dermatomyositis, eosinophil count, eosinophil percentage of leukocytes, inflammatory bowel disease, leukocyte count, lymphocyte count, mathematical ability, monocyte percentage of leukocytes, multiple sclerosis, nitric oxide exhalation measurement, platelet-to-lymphocyte ratio, primary biliary cirrhosis, respiratory system disease, rheumatoid arthritis, selective IgA deficiency disease, self-reported educational attainment, serum IgM measurement, serum gamma-glutamyl transferase measurement, serum non-albumin protein measurement, systemic scleroderma, ulcerative colitis), the majority of which are a wide range of autoimmune related diseases (e.g. asthma, eczema, allergy, Crohn's disease, type I diabetes, SLE, RA).

eQTLs of rs4065275 are *AC090844.2*, *GSDMA*, *GSDMB*, *IKZF3*, *ORMDL3*. eQTLs of rs8076131 are *PGAP3*, *IKZF3*, *GSDMB*, *GSDMA* and *ORDML3*. P-P interactions are found with 8 genes (*AC087491.2*, *PPP1R1B*, *ERBB2*, *PGAP3*, *GRB7*, *IKZF3*, *ZPBP2*, *MIEN1*), which include common targets with eQTL *IKZF3* and *PGAP3* (Figure 7a-b). CRISPRi data has shown that *KRT10* and its antisense-RNA (TMEM99) are regulatory targets of the *ORMDL3* promoter. *KRT10* (keratin) is a component of the cytoskeleton

of skin epithelial cells, and it plays a role in microbial infection in the nose and lung (Shivshankar et al., 2011). The alternative allele of rs4065275 was found to increase regulatory activity in an MPRA in Jurkat cells (Figure 7c) (Mouri et al., 2022), while the alternative allele of rs8076131 showed a decrease in regulatory activity in an MPRA in HEK293T cells (Liu et al., 2017). A link with the GWAS could be the fact that the alternative variant of rs4065275 changes TFBS for IKZF2, a member of the IKAROS transcription factor family which is involved in the regulation of lymphocyte development and controls T cell apoptosis in an IL2-dependent manner (Figure 7d) (Heizmann et al., 2018; Morgan et al., 1997).

*ORDML3* plays a role in innate immunity, explaining its involvement in most of the auto-immune related disease-associations. Additionally, increased expression of *ORMDL3* has been associated with asthma (Nowakowska et al., 2023). The alternative allele of the variants is associated with increased expression of *ORMDL3* as well as *GSDMA* and *GSDMB* in several tissues (eQTL), but interestingly with decreased expression of *IKZF3*, another member of the IKAROS transcription factor family. Indeed, altered expression of this transcription factor due to the variants in the *ORDML3* Epromoter could explain the wide range of immune-related diseases associated with the variants. *PGAP3* encodes a glycosylphosphatidylinositol-specific phospholipase. Mutations in this gene cause neurologic hyperphosphatasia with cognitive disability (Abdel-Hamid et al., 2018), and might be linked with cognitive-associated GWAS like bipolar I disorder, mathematical ability, and educational attainment via a similar biological mechanism.

*GSDMA* (Gasdermin A) and *GSDMB* (Gasdermin B) are involved in inflammatory cell death (pyroptosis), necessary for the recruitment of immune cells to infected sites in the skin and intestine (Deng et al., 2022; Zhou et al., 2020). Altered expression of these genes as indicated by eQTL for both variants could be linked to eczema and intestine-related autoimmune disease association.

### **COASY locus**

The *COASY* promoter shows enhancer activity in 5 different cell lines (A549, CCRF CEM with IFN $\alpha$  stimulation, HCT116, MCF-7, and SH-SY5Y). SNP rs629861 in the *COASY* promoter is associated with 17 different GWAS traits (Drugs used in diabetes use measurement, Eczema, Parkinson's disease-age at diagnosis, type II diabetes

mellitus, atopic asthma, body mass index, body weight, colorectal cancer, endometrial neoplasm, cortical surface area measurement, high density lipoprotein cholesterol measurement, lymphocyte count, self-reported educational attainment, serum gamma-glutamyl transferase measurement, tea consumption measurement, vitamin D measurement). Disease ontology includes Parkinson's disease, asthma, type II diabetes and body mass index.

SNP rs629861 has 7 eQTLs within 100kb (*CNTNAP1*, *COASY*, *HSD17B1*, *PLEKHH3*, *PSMC3IP*, *TUBG1*, *TUBG2*), and the *COASY* promoter has 22 P-P interactions in over 30 tissues (*AC003104.1*, *ATP6V0A1*, *CCR10*, *CNTNAP1*, *CTD-3193K9.4*, *PLEKHH3*, *CNP*, *CNTD1*, *COA3*, *CTD-2132N18.3*, *RAB5C*, *DNAJC7*, *NKIRAS2*, *FAM134C*, *TUBG1*, *HMG2P15*, *PTRF*, *TUBG2*, *WNK4*, *COASY*, *LOC108783654*, *MLX*). As many as 5 gene targets overlap between the eQTL and P-P interactions (*CNTNAP1*, *COASY*, *PLEKHH3*, *TUBG1*, *TUBG2*) (Figure 7a-b), increasing the likelihood that these genes are potential targets of the *COASY* Epromoter. The SNP shows allelic-skewed activity in K562 cells as assessed by MPRA (van Arensbergen et al., 2019) (Figure 7c). The SNP alters the binding sites of several TFBS, including loss of the GLTPD1 binding site (Figure 7d), which is involved in negative regulation of NLRP3 inflammasome complex assembly and interleukin-1 beta production. Moreover, the SNP also alters the binding of HCFC1 (Supplemental Table 5), which is required for certain types of insulin secretion (Iwata et al., 2013), as well as for the recruitment of epigenetic activators to promoters of lipogenic genes to promote *de novo* lipogenesis (Lane et al., 2019). These functions of HCFC1 could link the SNP to the metabolic-related GWAS type II diabetes, body weight and body mass index.

*COASY* encodes protein coenzyme A synthase, which plays an important role in synthetic and degradative metabolic pathways, in particular of vitamin B5 (Daugherty et al., 2002). Mutations in this gene are associated with neurodegeneration with brain iron accumulation (Dusi et al., 2014; Rosati et al., 2023; van Dijk et al., 2018), which could imply similar mechanisms are at play in explaining the association of rs629861 with Parkinson's disease. Furthermore, contactin-associated protein (*CNTNAP1*) is also associated with hypomyelination and nervous system development (Laquérière et al., 2014), and could therefore as well be a distal target of the *COASY* Epromoter explaining the association with Parkinson's disease. *HSD17B1* and *PSMC3IP*, both

eQTLs for rs629861, are involved in estrogen metabolism and activation (Puranen et al., 1997; Zangen et al., 2011). Because estrogen is a known player in insulin sensitivity and gluconeogenesis (Yan et al., 2019), this gene could be a distal target of the COASY Epromoter involved in the metabolic- and sex-hormone related traits (endometrial neoplasm, body weight, lipoprotein cholesterol, and type II diabetes).

### ***NIF3L1/PPIL3* locus**

The shared *NIF3L1/PPIL3* promoter shows enhancer activity in 2 different cell types (SH-SY5Y\_normal and primed hESC). A variant in this Epromoter is rs7559150, which is associated with 3 different GWAS (outer ear morphology trait, parathyroid hormone measurement, response to triptolide). Triptolide is a compound from the bark of a plant root that has anti-inflammatory properties and it has been used in the treatment of autoimmune diseases, fibrosis and neurodegeneration.

There are 3 genes in eQTL with rs7559150 (*NIF3L1*, *PPIL3* and *CFLAR*). Additionally, 21 genes show P-P interaction with the Epromoter in different tissues (*ALS2*, *ALS2CR12*, *CASP10*, *CASP8*, *CFLAR*, *CFLAR-AS1*, *FAM126B*, *NDUFB3*, *RNU6-1206P*, *KCTD18*, *SGOL2*, *NOP58*, *SNORD70*, *STRADB*, *TRAK2*, *BZW1*, *CLK1*, *FAM126B*, *LOC101927795*, *NIF3L1*, *PPIL3*) (Figure 7a-b). Transcription of two genes is impacted by CRISPRi on the Epromoter (*PPIL3* and *CFLAR*), giving additional evidence for Epromoter activity and target genes. The minor allele of the SNP is associated with a significant decrease in the expression of *PPIL3*, *NIF3L1* and *CFLAR*, as well as several other genes in the vicinity. *CFLAR* is an eQTL of the SNP with an increase in expression in the esophagus, but a decrease in expression in e.g. thyroid. Similarly, the minor allele of rs7559150 showed decreased regulatory activity in a SuRE assay in K562 cells (Figure 7c)(van Arensbergen et al., 2019). Several TFBS are disrupted by rs7559150, including several members of the ETS family of TFs (Figure 7d; Supplemental Table 5), which are involved in a wide range of functions including inflammation and apoptosis, which could be at the basis of the association with triptolide cytotoxicity.

The rs7559150 SNP was identified to reduce *NIF3L1* expression in memory T cells after 16h of stimulation (Soskic et al., 2022). Moreover, the gene *NIF3L1* itself is associated with Williams-Beuren syndrome, in which patients often have elevated

blood calcium levels (hypercalcemia) (Merla et al., 2004). This is in agreement with the fact that rs7559150 is associated with parathyroid hormone measurement. The rs7559150 minor allele could lead to decreased *NIF3L1* expression which in turn leads to decreased negative regulation of transcription of the target genes, resulting in increased parathyroid activity and increased blood calcium levels. Additionally, *NIF3L1* lies in the ALS2-critical region, a neurodegenerative disease with loss of motor neurons. Interestingly, one of the P-P interactions of this Epromoter is with the *ALS2* gene, mutations in which are also associated with ALS. Potentially the *NIF3L1/PPIL3* Epromoter could play a role in the regulation of *ALS2* (over 600kb away), adding to the ALS-association of both genes.

*CFLAR* is a gene regulator of apoptosis and inflammation (Xiao et al., 2012; Xiaohong et al., 2019). Moreover, high expression of *CFLAR* was found to positively regulate immune response to soft tissue sarcoma in the tumor microenvironment (Liu et al., 2024). Rs7559150 in Epromoter *NIF3L1/PPIL3* shows P-P interaction with *CFLAR*, and could thus potentially regulate and alter expression of *CFLAR*, resulting in an altered systemic apoptosis and inflammation regulation, which could culminate in altered cytotoxicity and its association with triptolide response.

The PPIL3 protein is a member of the cyclophilin family, and was shown to be a negative feedback regulator of NF- $\kappa$ B signaling pathway in homeostasis of innate immunity (Sheng et al., 2018), which could similarly suggest involvement in the association to triptolide cytotoxicity.

## 6.2 Data resource and sharing

In this work, we generated a comprehensive human enhancer and Epromoter resource and an Epromoter variants resource which characterized with GWAS, eQTLs, promoter-promoter interactions, affected TF binding, and MPRA. Additionally, we provide CRISPRi perturbed promoters with cis-regulated genes collected from published studies. We also provide an MPRA resource with allelic impact SNPs collected from published studies. These data resources will be useful not only for Epromoter studies but also for other regulatory element studies. Currently, these data resources can be accessed by the cloud link: <https://amubox.univ-amu.fr/s/FcapcWqFM8qED3E>.

For example, the data resources were utilized in collaborative works from several teams. The enhancers and Epromoters resource were also shared with an exonic-enhancers project (collaboration with Benoit Ballester's team from TAGC, Marseille, France), a G-quadruplex project (collaboration with Jean-Christophe Andrau's team from IGMM, Montpellier, France), short tandem repeats project (collaboration with Charles Lecellier's team from IGMM, Montpellier, France). The GWAS trait resource and analysis in the PhD work was also utilized by a collaborative work with Sylvain Marcelline's team from the University of Concepción in Chili (See Annex 2; <https://doi.org/10.1016/j.cdev.2024.203924>).

## 6.3 Additional work

### Assessment of allelic impact on promoter versus enhancer activity of Epromoters

One way to assess the allelic impact of genetic variants on cis-regulatory activity is to perform allelic-specific MPRA where reference and alternative alleles are tested in parallel. Comprehensive assessments of allelic-skewed variants from published MPRA allowed us to demonstrate that functional cis-regulatory variants associated with Epromoters are significantly more pleiotropic (Figure 6 of part 6.1; Results section). However, an underlying question is whether allelic variants differentially impact on enhancer or promoter activity of Epromoters. To systematically address this question, we designed an MPRA strategy to assess the impact of allelic variants on promoter and enhancer activity in parallel (Figure 6.3.1).

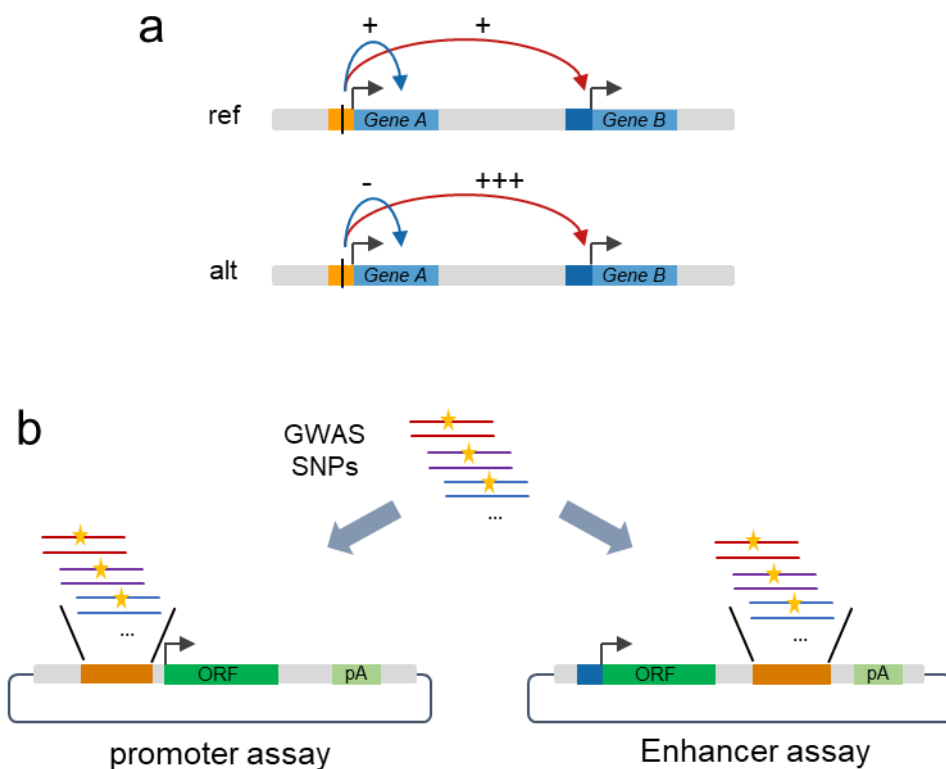


Figure 6.3.1 The MPRA strategy to assess the impact of allelic variants on promoter and enhancer activity of Epromoters. Figure 6.3.1a shows the impact of allelic variants on promoter and enhancer activity of Epromoters. Figure 6.3.1b shows the impact of allelic variants on promoter and enhancer activity of Epromoters can be assessed by promoter assay and enhancer assay.

Together with Antoinette van Ouwkerk (a post-doc in Spicuglia's lab), we built an MPRA library to assess all identified GWAS-SNPs associated with Epromoters in our study (Wan et al. draft; Results section). More specifically, we took the 4330 GWAS-SNPs that lie in 2301 Epromoters, resulting in 8660 variant sequences with the two alleles, for which we extracted a 220bp genetic sequence with the variant at the 110th position. Additionally, 561 variants that lie within 500bp of the TSSs of two divergent Epromoters, were added in both orientations corresponding to the direction of the TSS, resulting in a total of 9782 variant sequences. In addition, we selected in total 686 positive control sequences, of which 104 sequences of Epromoters are active in >5 cell lines, and 582 sequences of Epromoters active in K562 or GM12878 cell lines. Additionally, we selected 800 positive control SNPs, based on the top 100 SNPs ranked by allelic effect p-value in two published studies ((Abell et al., 2022) performed in GM12878 and (van Arensbergen et al., 2019) performed in K562). Furthermore, we selected 698 negative control sequences, divided into 298 sequences with low regulatory activity (fold change between 0.9 and 1.1) in K562 and GM12878 CapSTARR-seq and 200 randomly shuffled sequences from positive control tested SNPs, in forward and reverse orientation (resulting in 400 sequences). In total, we generated a library of 11966 sequences that were cloned in enhancer and promoter MPRA vectors in parallel (Figure 6.3.2a).

Both enhancer and promoter MPRA experiments have been performed in the K562 cell line with or without IFN $\alpha$  stimulation, by Antoinette van Ouwkerk. As for now, we have performed preliminary analyses of the enhancer MPRA. We performed DE-seq analyses to assess the differential allelic activity of the Epromoter's variants (Figure 6.3.2b-c). We observed 47 and 39 significant allelic-skewed SNPs in non-stimulated and IFN $\alpha$ -stimulated K562 cells, respectively (adjusted  $P$  value < 0.05). Further analysis will be explored based on this dataset. For example, whether these functional allelic SNPs display pleiotropic effect? whether is a correlation between allelic impact and pleiotropic effect? And whether the same SNPs have a similar or opposite impact on promoter and enhancer activities.



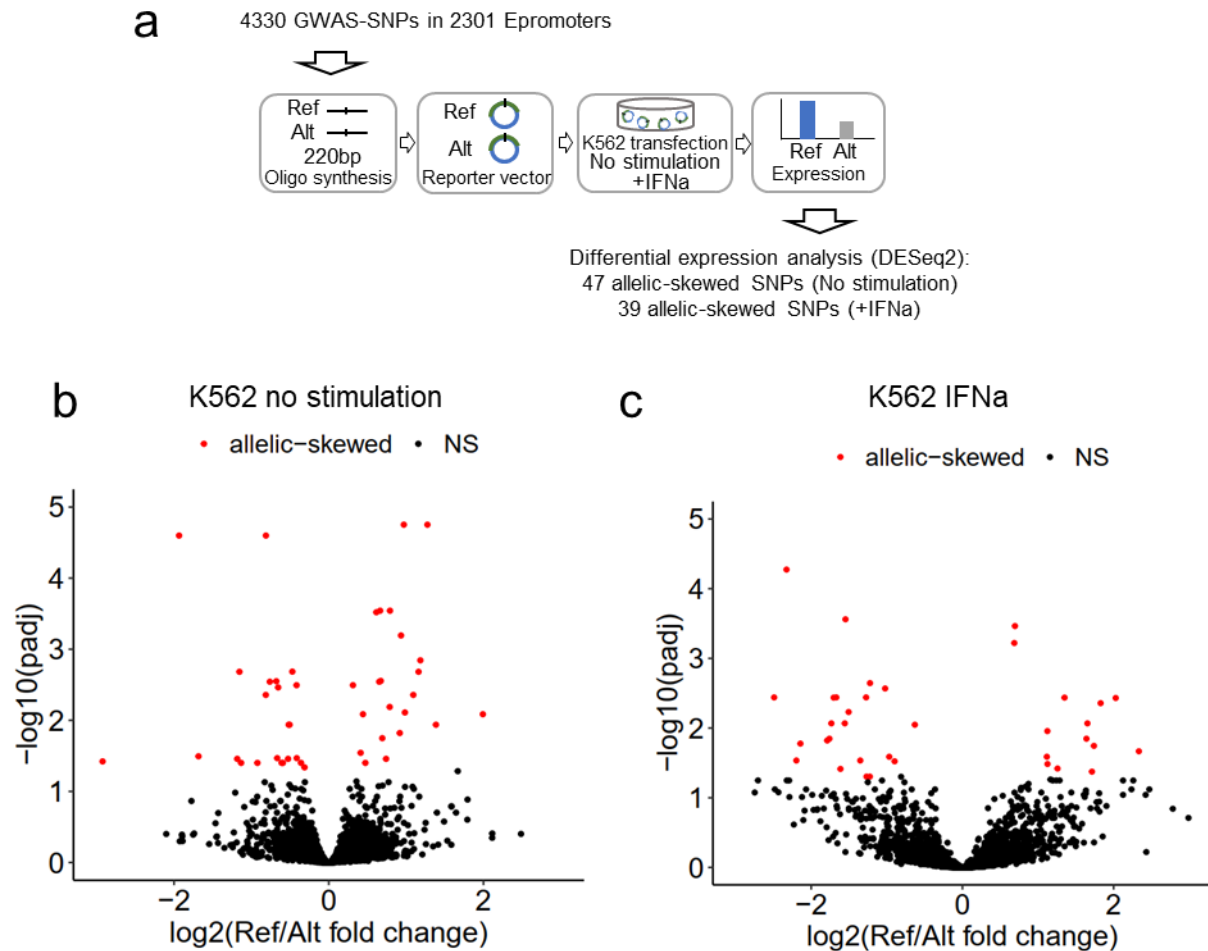


Figure 6.3.2 Assess allelic impact on enhancer activity of Epromoters. Figure 6.3.2a. The overall design of the enhancer MPRA. The oligo sequences were designed as 220bp with reference or alternative alleles of SNPs in the center. Then the oligo sequences were integrated into the reporter vectors. The reporter vectors were transfected into K562 in two conditions (No stimulation and IFNa stimulation). The quantification and differential expression analysis were initially performed in two conditions of MPRA library. Figure 6.3.2b-c. The volcano plots show the allelic impact results in two MPRA experiments (K562 without stimulation and IFNa stimulation). Each dot represents a SNP. The red dots are allelic-skewed SNPs (adjusted  $P$  value  $< 0.05$ ). The black dots are not significant.

# Chapter 7. Epromoters function as a hub to recruit key transcription factors required for the regulation of stress-response clusters

This is a collaborative work led by Juliette Malfait (PhD student) in Salvatore Spicuglia's team which the manuscript is in preparation. Currently, my contribution to this work was the bioinformatic analysis in Figure 7, and results cleaning from a pipeline of Epromoter cluster prediction.

Previous results have shown that Epromoters function as a local hub recruiting the key TFs required for coordinated regulation of gene clusters during the inflammatory response (Santiago-Algarra et al., 2021). Following this study, the research question is: whether Epromoters coordinate regulation of gene clusters in other stress response processes? Whether Epromoters are function as hub in more general cellular response to intra- and extra-cellular signals?

The collaborative work is investigating whether stress response genes are regulated as clusters. The work is based on the comprehensive stress response datasets collected by Juliette and colleagues in different stimulatory or stress conditions (heat shock, serum response, DNA damage, TNF stimulations) in human and mouse from published studies.

The strategy is investigating the genomic distances between stress response genes and comparing with developmental genes and randomly selected genes (Figure 7A). We calculated the gene distance distribution between induced genes (including stress response datasets and differentiation datasets) and random genes. As the three example datasets shown in Figure 7B-C, the gene distance distribution is mostly enriched within 100kb in stress response datasets (eg. Heat-Shock) by comparing with random genes but not in differentiation datasets be it in vitro or embryonic (eg. Fibroblast growth factor (FGF), Mesoderm CD56). The deviation scores were calculated to quantify the distance distribution differences (Figure 7C). In general, the deviation scores are higher in stress response datasets than differentiation datasets (Figure 7D). According to gene distance distribution and deviation scores, the stress

response datasets generally display higher deviation scores associated with smaller distances, which are separated from the differentiation datasets (Figure 7E).

The distance distribution results indicate that induced genes are more clustered in stress response datasets but not in developmental datasets. Based on these findings, a bioinformatic pipeline was developed to predict the Epromoters at play in different stimulatory or stress conditions. The systematic validation and analysis (performed by Juliette Malfait) suggested that Epromoters function as a local hub to recruit key transcription factors required for the regulation of nearby co-induced genes in response to different inflammatory and stress conditions.

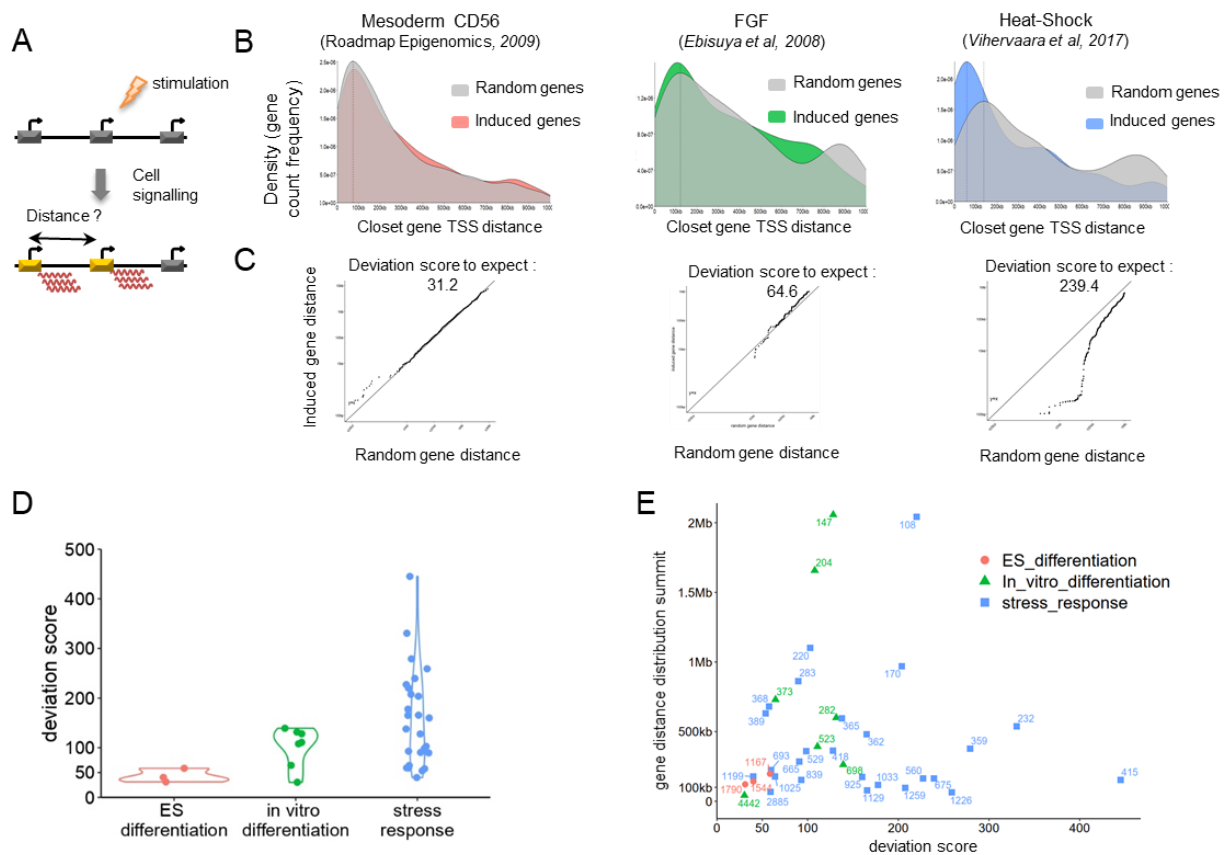


Figure 7. Stress response genes are distributed as clusters in mammalian genomes.

A. The scheme to investigate the genomic distances between induced genes after stress stimulation.

B. Gene distances distribution of induced genes and random genes in three examples' datasets. The summit of distribution is marked by a dashed line.

C. The QQ-plot illustrated the difference between the distance distribution of induced and random genes in three examples' datasets. The  $y=x$  line means that the expectation between induced and random genes is no difference. The deviation scores are quantified as the difference between the QQ-plot and expectation line.

D. The summary of deviation score in all the differentiation and stress response datasets. Each point represents one dataset.

E. The correlation between gene distance distribution and deviation score. Each point represents one dataset.

# **DISCUSSION AND PERSPECTIVES**

# Chapter 8. Discussion

Genome-wide studies have become pivotal in unraveling the genetic basis of complex traits through the identification of single nucleotide polymorphisms (SNPs) associated with specific phenotypes. In this work, we employed a comprehensive approach to investigate the genetic landscape of Epromoters, an unconventional type of cis-regulatory element harboring both enhancer and promoter functions. We examined their associations with genetic variants, particularly focusing on SNPs identified in GWAS. Our comprehensive analysis provides novel insights into the genetic variation within Epromoters and control promoters, highlighting their potential roles in complex trait regulation. The enrichment of specific GWAS traits and the increased pleiotropy observed in Epromoters, as compared with typical promoters, suggest their importance in the genetic architecture of complex traits and diseases.

Our findings underscore the intricate relationship between Epromoter-associated genetic variation, eQTLs, and pleiotropy, unraveling the potential regulatory impact on both proximal and distal target genes. The identified link between Epromoters and distal gene regulation provides valuable insights into the functional genomics of complex traits. It paves the way for a deeper understanding of the molecular mechanisms underlying pleiotropy.

## **Epromoters' intrinsic features are distinguished from typical promoters**

A major paradigm in the field of gene regulation is to understand what are the molecular bases of proximal (promoter) versus distal (enhancer) functions (Andersson & Sandelin, 2020). Although a unified model of cis-regulatory functions has been proposed (Core et al., 2014), several studies, including ours, have suggested that intrinsic (binding sites, nucleotide composition, etc) and extrinsic (transcription factors (TFs), genomic context, etc) features that drive enhancer and promoter activities are not the same (Core et al., 2014) (Henriques et al., 2018) (Rennie et al., 2018) (Mikhaylichenko et al., 2018) (Nguyen et al., 2016) (Santiago-Algarra et al., 2021) (Dao et al., 2017) (Malfait et al., 2023). Previous studies have shown that the type of TF that binds a cis-regulatory element might influence the relative enhancer or promoter activity (Andersson & Sandelin, 2020; Nguyen et al., 2016). Similarly, we showed that interferon-response Epromoters have a higher density and better quality of Interferon-

Stimulated Response Elements (ISRE), as compared with typically induced promoters, which, in turn, results in the Epromoter-specific recruitment of STAT1/2 and IRF TFs and activation of neighbor genes (Santiago-Algarra et al., 2021).

Here, we took advantage of the comprehensive Epromoter resource we have built to perform a thorough comparison of Epromoters with typical promoters displaying similar promoter activity. Our results revealed several intrinsic differences between Epromoters and typical promoters (Figure 8.1). First, Epromoters are associated with genes that are less tissue-specific and harbor multiple alternative promoters. Second, they are involved in a higher number of interactions with other promoters. Third, their sequences are more conserved and display a higher number of G4 elements. Fourth, Epromoters have a higher density and complexity of TF binding sites, which is reflected by a high density of TF binding. Finally, Epromoters display a higher level of sense and antisense transcription initiation which is reflected by a higher overlap with RNAPII binding.

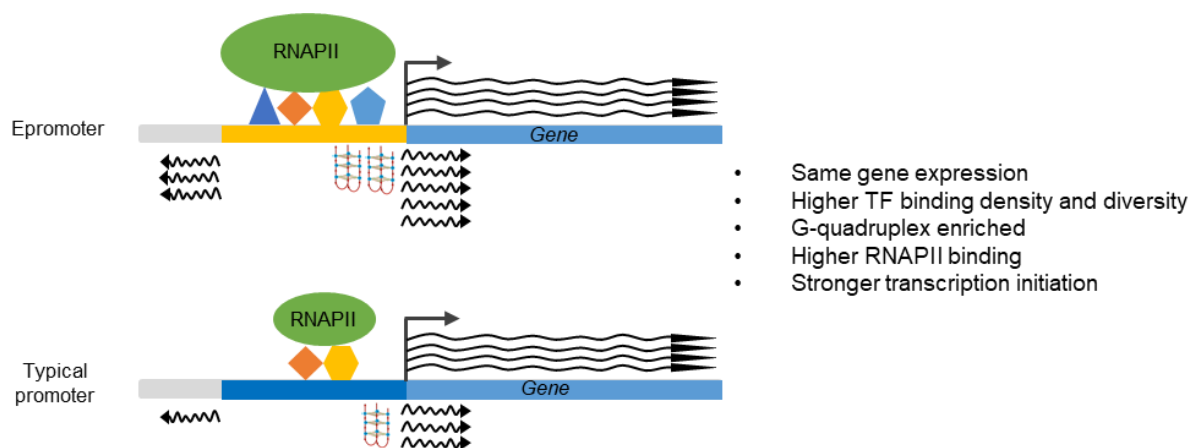


Figure 8.1. The summary of Epromoters' intrinsic features which are distinguished from typical promoters.

Based on these findings, we speculate that Epromoters represent a combination of the two types of cis-regulatory elements, thus combining features associated with enhancer and promoter activities within an enhancer-promoter continuum of cis-regulatory elements. This intermediated position implies that Epromoters might display a higher density and complexity of TFBS because it has to accommodate the binding of TFs for both enhancer and promoter functions. In this scenario, typical promoters are enriched in binding sites for TFs conferring promoter activity and enhancers enriched in binding sites for TFs conferring enhancer activity, while Epromoters will be

enriched for both types of binding sites leading to a higher density of TFBS. Future works should systematically assess the contribution of TFBS and associated TFs to the enhancer and promoter activity to better understand the molecular features that determine the intrinsic promoter and enhancer potentials of cis-regulatory elements, and in particular of Epromoters. This, in turn, might help to better predict the impact of mutations or natural variants of Epromoters that might affect either proximal or distal gene regulation.

### **Pleiotropic impact of Epromoter's variants**

Several studies, including ours, have demonstrated that human genetic variation within Epromoters influences distal gene expression (Dao et al., 2017) (X. Wang et al., 2018) (Joanna Mitchelmore et al., 2020) (Jung et al., 2019) (Mariana Saint Just Ribeiro et al., 2022). Moreover, specific examples highlight the distal impact of disease-associated variants within Epromoters (Chandra et al., 2021) (Victor Rusu et al., 2017) (Irina A. Sergeeva et al., 2016) (Yagihara et al., 2016) (Samia Nisar et al., 2022) (Hua et al., 2018) (Gao et al., 2018) (Malfait et al., 2023). The complex regulation by Epromoters might therefore have two predicted consequences. On the one hand, there might be a general underestimation of the impact of Epromoter variation in disease because the causal gene might not be the closest one and therefore the link between genotype and phenotype might be missed in many case studies. On the other hand, as Epromoters potentially control several genes at the same time and efficiently recruit key TFs, mutations in these regulatory elements are expected to have a stronger pathological impact, as compared to typical promoters. This might result from the regulation of multiple genes either involved in the same (additive or synergistic effects) or different (pleiotropy) pathways (Figure 8.2). Indeed, our present work reveals that genetic variants within Epromoters linked to GWAS are significantly associated with multiple diseases as compared with typical promoters, supporting the hypothesis whereby Epromoters might have a pleiotropic effect in disease by perturbing the expression of several genes at the same time.



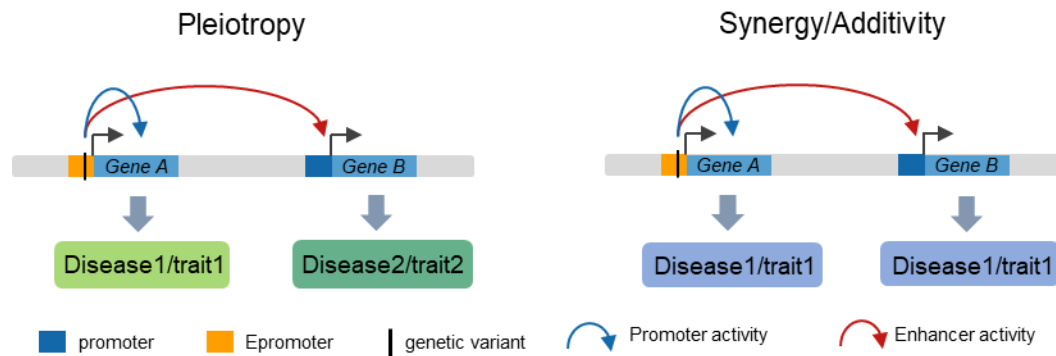


Figure 8.2. The complex effect of genetic variation at Epromoters.

Pleiotropy, referred as to a single cis-regulatory element affecting more than one trait independently (Cano-Gamez & Trynka, 2020), could be due to the perturbation of a single gene playing multiple functions in different tissues (Gupta et al., 2017) (Sinnott-Armstrong et al., 2021) or the regulation of multiple genes in the same or different tissues (Sobreira & Nóbrega, 2021) (Joslin et al., 2021). Our results rather point to the latter possibility. On the one hand, we observed that pleiotropy is associated with an increased number of target genes, as assessed by consistent eQTL and promoter-promoter interactions. While it is difficult to ensure that all Epromoter variants are bona fide distal regulators, we noticed that taking into consideration functional assessment of allelic-specific activity by MPRA allows for significant enrichment of pleiotropic Epromoters. On the other hand, a careful examination of several pleiotropic Epromoters, reveals that the different target genes play a role in different physiological functions that might explain the association with the different diseases. For instance, a SNP in the Epromoter of SETD1A gene affected STX1B, PRSS53, VKORC1 and KAT8 genes downstream, which are associated with brain diseases and warfarin dose effect in heart diseases (Figure 8.3; Figure 7 in section 6.1). Another example in our study shows a SNP at Epromoter of OAS3 associated with COVID-19 severity by affecting three OAS family genes expression in antiviral function (Figure 8.3; Figure 5 in results section). In line with our finding, a schizophrenia-risk SNP within the promoter of the VSP45 gene was shown to cis-regulate three genes via allele-specific chromatin looping (Figure 8.3). These genes act in a non-additive synergistic fashion to enhance dendritic complexity and neuronal activity (Zhang et al., 2023).

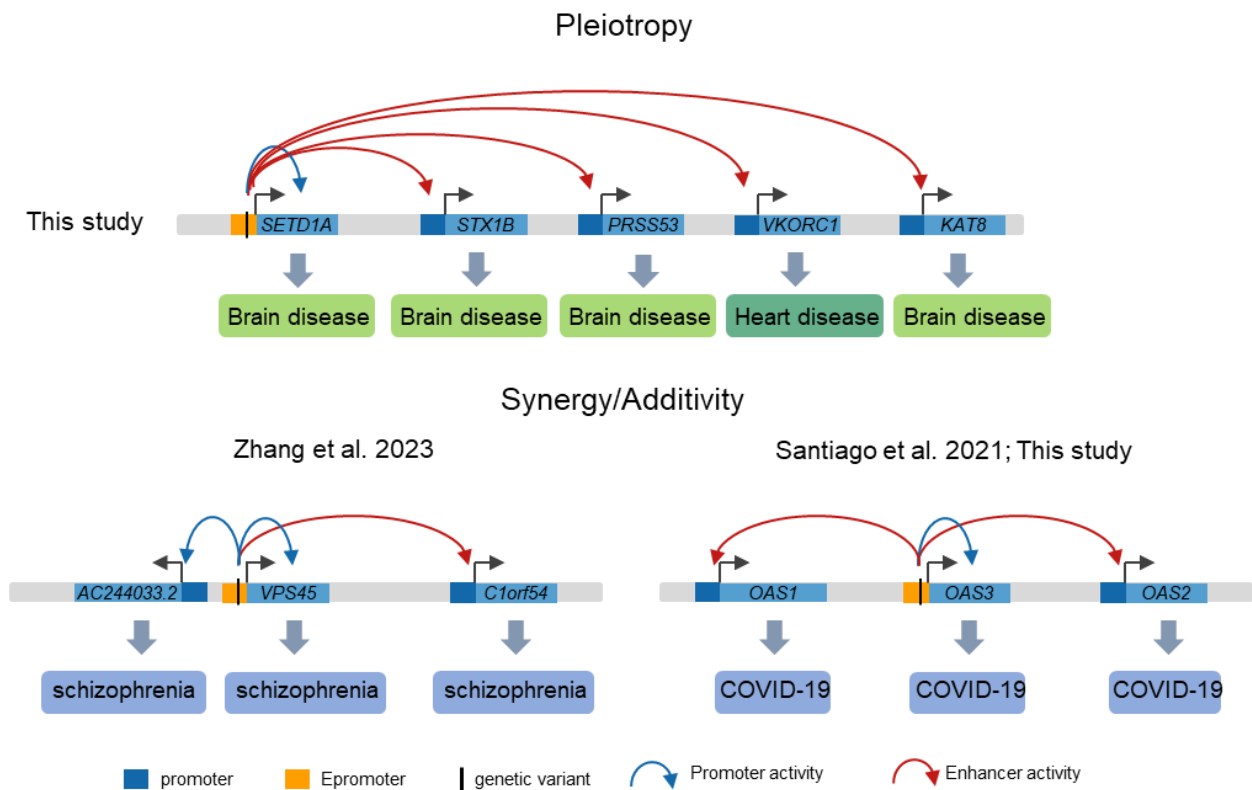


Figure 8.3. The instances of pleiotropic and synergistic/additive effect of Epromoters variation.

Genetic variation might impact the expression of neighboring genes in the same (e.g., enhancer and promoter activity are equally affected) or opposite (e.g., the genetic variant induces an enhancer/promoter switch) directions. For instance, two studies demonstrated that an alternative variant associated with prostate cancer increases the enhancer activity of the promoter leading to decreased expression of the proximal transcript but increased expression of two distal transcripts directly involved in cancer progression (Hua et al., 2018) (Gao et al., 2018). Moreover, the genetic variants might differently impact enhancer and/or promoter activity in different tissues. For instance, Leung et al. found frequent examples of dynamic epigenetic switches where active promoters in one tissue displayed a histone modification signature of enhancers in other tissues/cell types (Leung et al., 2015). Similarly, Chandra et al. found a substantial number of promoter-promoter interactions involving transcriptionally inactive genes, suggesting that non-transcribing promoters may function as active enhancers for distal genes (Chandra et al., 2021). An enlightening example is provided by the OAS1/2/3 locus, where genetic variation at the OAS3 Epromoter affects Interferon-dependent enhancer and promoter activity in both K562 and A549 cell lines.

However, the relative impact on enhancer and promoter activities switches between the two cell lines.

Overall, by leveraging extensive genomic and functional datasets, our study explores the intricate relationship between Epromoter variation, pleiotropy, and target gene regulation, shedding light on the complex regulatory mechanisms underlying the genetic architecture of complex traits.

### **Limitations of this study**

There are some limitations in this PhD work. First, the STARR-seq datasets collected in this study are up to 2021. Some of the latest STARR-seq data may be released or published. For the sake of consistency of the results of previous and subsequent analyses, these data are not collected in this study for the time being. Second, compared with other epigenomic data, genome-wide STARR-seq applied to human cell lines is still scarce. Therefore, the number of all Epromoters on a genome-wide scale may still be underestimated. We look forward to the availability of STARR-seq datasets in more human cell lines. Third, due to the differences in STARR-seq and CapSTARR-seq libraries and data processing, we cannot provide a universal standard enhancer activity value for all Epromoters. Therefore, in the supplementary datasets, we provide the genomic coordinates of the Epromoter, corresponding genes, cell lines, and conditions. But for the CapSTARR-seq generated in this study, we provided the enhancer activity for each dataset. Fourth, in integrating Epromoter, eQTL, promoter-promoter interactions, and MPRA data, we did not analyze the consistency of cell lines between different datasets. The reason is that there are not many available datasets from each data type in the same cell line. And the data from different cell types still could be provided as potential evidence. Fifth, in the prediction of transcription factor binding effects, we found that the prediction results of different tools didn't display a high consistency. Therefore, we provide the prediction results of different tools to keep more potential evidence.

# Chapter 9. Perspectives

## **Assess the impact of allelic variants on Epromoters by MPRA**

We have systematically characterized the GWAS SNPs in Epromoter and collected previous MPRA datasets to functionally validate a subset of these SNPs. However, a more comprehensive evaluation of all GWAS SNPs within Epromoters remains necessary. MPRA is an ideal technology for large-scale evaluation of the effects of SNPs on regulatory activity. As detailed in Chapter 6 (Section 6.3), our MPRA will be divided into two strategies to evaluate promoter activity and enhancer activity respectively. By leveraging this data, we can systematically assess the effects of SNPs on promoter and enhancer activity of Epromoters. This comprehensive analysis will provide a deeper understanding of the distinct roles and interconnected relationships between enhancers and promoters.

## **Apply AI-based strategy to dissect grammar rules of Epromoter**

Machine learning, particularly deep learning strategies, has demonstrated significant power in the study of enhancers and promoters. Another PhD student in the host lab has set out to use deep learning to explore the grammar of Epromoters. The STARR-seq and Epromoter datasets generated in this study will serve as training data for deep learning models. Preliminary results indicate that larger training datasets and quantitative data substantially enhance the predictive performance of deep learning models. Therefore, the MPRA resources collected in this study, along with the MPRA data to be generated in the future, can also serve as a foundation for training these models. As the quality and scale of the data improve, it will become possible to predict the regulatory activity of Epromoters and to decipher their sequence syntax rules. Ultimately, this will enable the de novo design of regulatory elements.

## **Contribution of Epromoter's target genes on disease**

The results obtained during the thesis suggested that Epromoters might have an important contribution to disease because genetic variants can have an impact on the expression of multiple neighbor genes. However, to validate this hypothesis, functional experiments should be performed to assess the contribution of individual target genes on the disease(s) associated with the Epromoter's variants. A recent study that focused

on GWAS regulatory variants linked to schizophrenia provided a clear example of how to systematically validate the complex genetic effect of a single Epromoter variant (Zhang et al., 2023). By combining analyses of allelic chromatin accessibility, CRISPRi screening, precise SNP editing, chromatin interaction, and cellular phenotypes, they show that multiple genes in a single GWAS risk locus act in a non-additive synergistic fashion (Figure 9.1).

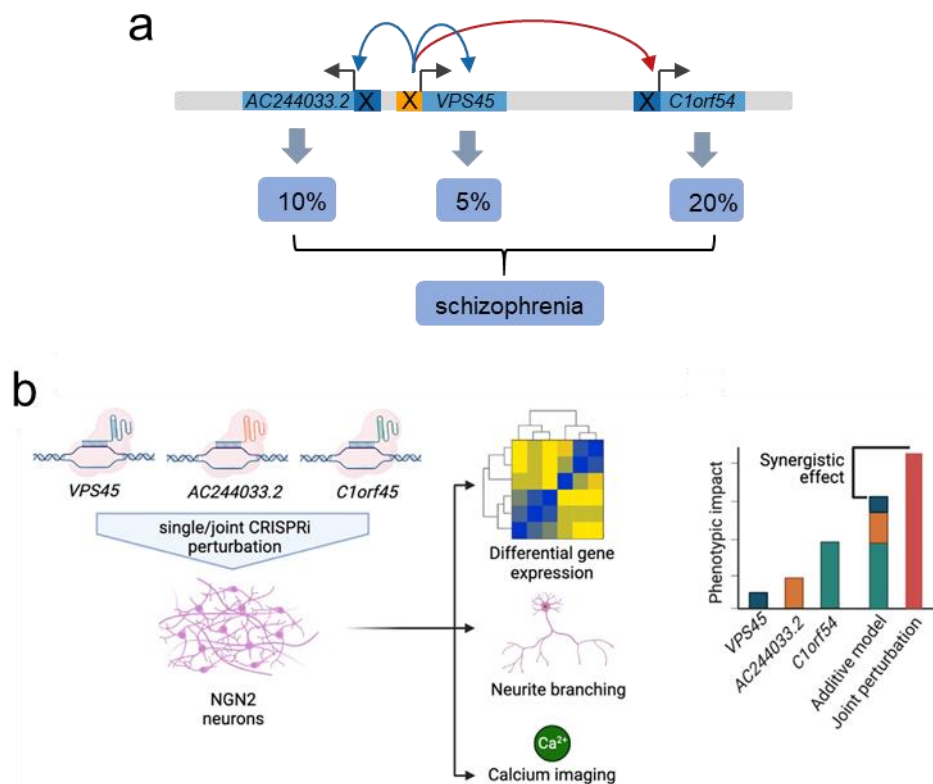


Figure 9.1. An example shows how to systematically validate the complex genetic effect (Zhang et al., 2023). Figure 9.1a shows the proximal gene and distal genes contributed differently to schizophrenia. Figure 9.1b (adopted from (Deans & Brennand, 2023)) shows individual and joint knockdown of VPS45, AC244033.2 and C1orf54 expression in NGN2-neurons results in altered gene expression, neurite branching, and neuronal activity. Joint perturbations can result in more or fewer synergistic (non-additive) effects than predicted by the additive model.

# REFERENCES

- Abdel-Hamid, M. S., Issa, M. Y., Otaify, G. A., Abdel-Ghafar, S. F., Elbendary, H. M., & Zaki, M. S. (2018). PGAP3-related hyperphosphatasia with mental retardation syndrome: Report of 10 new patients and a homozygous founder mutation. *Clin Genet*, *93*(1), 84-91. <https://doi.org/10.1111/cge.13033>
- Abell, N. S., DeGorter, M. K., Gloudemans, M. J., Greenwald, E., Smith, K. S., He, Z., & Montgomery, S. B. (2022). Multiple causal variants underlie genetic associations in humans. *Science*, *375*(6586), 1247-1254. <https://doi.org/10.1126/science.abj5117>
- Adamson, B., Norman, T. M., Jost, M., Cho, M. Y., Nunez, J. K., Chen, Y., Villalta, J. E., Gilbert, L. A., Horlbeck, M. A., Hein, M. Y., Pak, R. A., Gray, A. N., Gross, C. A., Dixit, A., Parnas, O., Regev, A., & Weissman, J. S. (2016). A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell*, *167*(7), 1867-1882 e1821. <https://doi.org/10.1016/j.cell.2016.11.048>
- Ahituv, N. (2016). Exonic enhancers: proceed with caution in exome and genome sequencing studies. *Genome Med*, *8*(1), 14. <https://doi.org/10.1186/s13073-016-0277-0>
- Ali, T., Renkawitz, R., & Bartkuhn, M. (2016). Insulators and domains of gene expression. *Curr Opin Genet Dev*, *37*, 17-26. <https://doi.org/10.1016/j.gde.2015.11.009>
- Alipanahi, B., Delong, A., Weirauch, M. T., & Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol*, *33*(8), 831-838. <https://doi.org/10.1038/nbt.3300>
- Andersson, R. (2015). Promoter or enhancer, what's the difference? Deconstruction of established distinctions and presentation of a unifying model: Prospects & Overviews. *Bioessays*, *37*(3), 314-323. <https://doi.org/10.1002/bies.201400162>
- Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., Ntini, E., Arner, E., Valen, E., Li, K., Schwarzfischer, L., Glatz, D., Raithel, J., Lilje, B., Rapin, N., . . . Sandelin, A. (2014). An atlas of active enhancers across human cell types and tissues [Research Support, Non-U.S. Gov't]. *Nature*, *507*(7493), 455-461. <https://doi.org/10.1038/nature12787>
- Andersson, R., & Sandelin, A. (2020). Determinants of enhancer and promoter activities of regulatory elements. *Nat Rev Genet*, *21*(2), 71-87. <https://doi.org/10.1038/s41576-019-0173-8>
- Anzalone, A. V., Randolph, P. B., Davis, J. R., Sousa, A. A., Koblan, L. W., Levy, J. M., Chen, P. J., Wilson, C., Newby, G. A., Raguram, A., & Liu, D. R. (2019). Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature*, *576*(7785), 149-157. <https://doi.org/10.1038/s41586-019-1711-4>
- Arnold, C. D., Gerlach, D., Stelzer, C., Boryn, L. M., Rath, M., & Stark, A. (2013). Genome-wide quantitative enhancer activity maps identified by STARR-seq [Research Support, Non-U.S. Gov't]. *Science*, *339*(6123), 1074-1077. <https://doi.org/10.1126/science.1232542>
- Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J. R., Grabska-Barwinska, A., Taylor, K. R., Assael, Y., Jumper, J., Kohli, P., & Kelley, D. R. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. *Nat Methods*, *18*(10), 1196-1203. <https://doi.org/10.1038/s41592-021-01252-x>
- Avsec, Ž., Weilert, M., Shrikumar, A., Krueger, S., Alexandari, A., Dalal, K., Fropf, R., McAnany, C., Gagneur, J., Kundaje, A., & Zeitlinger, J. (2021). Base-resolution

- models of transcription-factor binding reveal soft motif syntax. *Nat Genet*, 53(3), 354-366. <https://doi.org/10.1038/s41588-021-00782-6>
- Bailey, T. L., Johnson, J., Grant, C. E., & Noble, W. S. (2015). The MEME Suite. *Nucleic Acids Res*, 43(W1), W39-49. <https://doi.org/10.1093/nar/gkv416>
- Banerji, J., Rusconi, S., & Schaffner, W. (1981). Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences [Research Support, Non-U.S. Gov't]. *Cell*, 27(2 Pt 1), 299-308. <http://www.ncbi.nlm.nih.gov/pubmed/6277502>
- Barakat, T. S., Halbritter, F., Zhang, M., Rendeiro, A. F., Perenthaler, E., Bock, C., & Chambers, I. (2018). Functional Dissection of the Enhancer Repertoire in Human Embryonic Stem Cells. *Cell Stem Cell*, 23(2), 276-288 e278. <https://doi.org/10.1016/j.stem.2018.06.014>
- Barshad, G., Lewis, J. J., Chivu, A. G., Abuhashem, A., Krietenstein, N., Rice, E. J., Ma, Y., Wang, Z., Rando, O. J., Hadjantonakis, A. K., & Danko, C. G. (2023). RNA polymerase II dynamics shape enhancer-promoter interactions. *Nat Genet*, 55(8), 1370-1380. <https://doi.org/10.1038/s41588-023-01442-7>
- Bedrat, A., Lacroix, L., & Mergny, J. L. (2016). Re-evaluation of G-quadruplex propensity with G4Hunter. *Nucleic Acids Res*, 44(4), 1746-1759. <https://doi.org/10.1093/nar/gkw006>
- Bhat, P., Honson, D., & Guttman, M. (2021). Nuclear compartmentalization as a mechanism of quantitative control of gene expression. *Nat Rev Mol Cell Biol*, 22(10), 653-670. <https://doi.org/10.1038/s41580-021-00387-1>
- Bourges, C., Groff, A. F., Burren, O. S., Gerhardinger, C., Mattioli, K., Hutchinson, A., Hu, T., Anand, T., Epping, M. W., Wallace, C., Smith, K. G., Rinn, J. L., & Lee, J. C. (2020). Resolving mechanisms of immune-mediated disease in primary CD4 T cells. *EMBO Mol Med*, 12(5), e12112. <https://doi.org/10.15252/emmm.202012112>
- Boytsov, A., Abramov, S., Aiusheeva, A. Z., Kasianova, A. M., Baulin, E., Kuznetsov, I. A., Aulchenko, Y. S., Kolmykov, S., Yevshin, I., Kolpakov, F., Vorontsov, I. E., Makeev, V. J., & Kulakovskiy, I. V. (2022). ANANASTRA: annotation and enrichment analysis of allele-specific transcription factor binding at SNPs. *Nucleic Acids Res*, 50(W1), W51-W56. <https://doi.org/10.1093/nar/gkac262>
- Bozhilov, Y. K., Downes, D. J., Telenius, J., Marieke Oudelaar, A., Olivier, E. N., Mountford, J. C., Hughes, J. R., Gibbons, R. J., & Higgs, D. R. (2021). A gain-of-function single nucleotide variant creates a new promoter which acts as an orientation-dependent enhancer-blocker. *Nat Commun*, 12(1), 3806. <https://doi.org/10.1038/s41467-021-23980-6>
- Brand, A. H., Breeden, L., Abraham, J., Sternglanz, R., & Nasmyth, K. (1985). Characterization of a "silencer" in yeast: a DNA sequence with properties opposite to those of a transcriptional enhancer [Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, P.H.S.]. *Cell*, 41(1), 41-48. <http://www.ncbi.nlm.nih.gov/pubmed/3888409>
- Cano-Gamez, E., & Trynka, G. (2020). From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Frontiers in Genetics*, 11, 424. <https://doi.org/10.3389/fgene.2020.00424>
- Castillo, H., Hanna, P., Sachs, L. M., Buisine, N., Godoy, F., Gilbert, C., Aguilera, F., Munoz, D., Boisvert, C., Debais-Thibaud, M., Wan, J., Spicuglia, S., & Marcellini, S. (2024). *Xenopus tropicalis* osteoblast-specific open chromatin regions reveal promoters and enhancers involved in human skeletal phenotypes and shed light on early vertebrate evolution. *Cells Dev*, 203924. <https://doi.org/10.1016/j.cdev.2024.203924>

- Castro-Mondragon, J. A., Riudavets-Puig, R., Rauluseviciute, I., Lemma, R. B., Turchi, L., Blanc-Mathieu, R., Lucas, J., Boddie, P., Khan, A., Manosalva Perez, N., Fornes, O., Leung, T. Y., Aguirre, A., Hammal, F., Schmelter, D., Baranasic, D., Ballester, B., Sandelin, A., Lenhard, B., . . . Mathelier, A. (2022). JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res*, *50*(D1), D165-D173. <https://doi.org/10.1093/nar/gkab1113>
- Catarino, R. R., Neumayr, C., & Stark, A. (2017). Promoting transcription over long distances. *Nat Genet*, *49*(7), 972-973. <https://doi.org/10.1038/ng.3904>
- Chandra, V., Bhattacharyya, S., Schmiedel, B. J., Madrigal, A., Gonzalez-Colin, C., Fotsing, S., Crinklaw, A., Seumois, G., Mohammadi, P., Kronenberg, M., Peters, B., Ay, F., & Vijayanand, P. (2021). Promoter-interacting expression quantitative trait loci are enriched for functional genetic variants. *Nat Genet*, *53*(1), 110-119. <https://doi.org/10.1038/s41588-020-00745-3>
- Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*, *4*, 7. <https://doi.org/10.1186/s13742-015-0047-8>
- Chatterjee, S., & Ahituv, N. (2017). Gene Regulatory Elements, Major Drivers of Human Disease. *Annu Rev Genomics Hum Genet*. <https://doi.org/10.1146/annurev-genom-091416-035537>
- Chen, D., & Lei, E. P. (2019). Function and regulation of chromatin insulators in dynamic genome organization. *Curr Opin Cell Biol*, *58*, 61-68. <https://doi.org/10.1016/j.ceb.2019.02.001>
- Chen, K. M., Wong, A. K., Troyanskaya, O. G., & Zhou, J. (2022). A sequence-based global map of regulatory activity for deciphering human genetics. *Nat Genet*, *54*(7), 940-949. <https://doi.org/10.1038/s41588-022-01102-2>
- Chen, P. J., & Liu, D. R. (2023). Prime editing for precise and highly versatile genome manipulation. *Nat Rev Genet*, *24*(3), 161-177. <https://doi.org/10.1038/s41576-022-00541-1>
- Cheung, K. L., Zhang, F., Jaganathan, A., Sharma, R., Zhang, Q., Konuma, T., Shen, T., Lee, J. Y., Ren, C., Chen, C. H., Lu, G., Olson, M. R., Zhang, W., Kaplan, M. H., Littman, D. R., Walsh, M. J., Xiong, H., Zeng, L., & Zhou, M. M. (2017). Distinct Roles of Brd2 and Brd4 in Potentiating the Transcriptional Program for Th17 Cell Differentiation. *Mol Cell*, *65*(6), 1068-1080 e1065. <https://doi.org/10.1016/j.molcel.2016.12.022>
- Choi, J., Zhang, T., Vu, A., Ablain, J., Makowski, M. M., Colli, L. M., Xu, M., Hennessey, R. C., Yin, J., Rothschild, H., Grawe, C., Kovacs, M. A., Funderburk, K. M., Brossard, M., Taylor, J., Pasaniuc, B., Chari, R., Chanock, S. J., Hoggart, C. J., . . . Brown, K. M. (2020). Massively parallel reporter assays of melanoma risk variants identify MX2 as a gene promoting melanoma. *Nat Commun*, *11*(1), 2718. <https://doi.org/10.1038/s41467-020-16590-1>
- Choi, S. W., Mak, T. S., & O'Reilly, P. F. (2020). Tutorial: a guide to performing polygenic risk score analyses. *Nat Protoc*, *15*(9), 2759-2772. <https://doi.org/10.1038/s41596-020-0353-1>
- Consortium, F., the, R. P., Clst, Forrest, A. R., Kawaji, H., Rehli, M., Baillie, J. K., de Hoon, M. J., Haberle, V., Lassmann, T., Kulakovskiy, I. V., Lizio, M., Itoh, M., Andersson, R., Mungall, C. J., Meehan, T. F., Schmeier, S., Bertin, N., Jorgensen, M., . . . Hayashizaki, Y. (2014). A promoter-level mammalian expression atlas. *Nature*, *507*(7493), 462-470. <https://doi.org/10.1038/nature13182>
- Consortium, T. E. P. (2012). An integrated encyclopedia of DNA elements in the human genome [Research Support, American Recovery and Reinvestment Act



- Research Support, N.I.H., Extramural  
 Research Support, N.I.H., Intramural  
 Research Support, U.S. Gov't, Non-P.H.S.]. *Nature*, 489(7414), 57-74.  
<https://doi.org/10.1038/nature11247>
- Cooper, Y. A., Teyssier, N., Drager, N. M., Guo, Q., Davis, J. E., Sattler, S. M., Yang, Z., Patel, A., Wu, S., Kosuri, S., Coppola, G., Kampmann, M., & Geschwind, D. H. (2022). Functional regulatory variants implicate distinct transcriptional networks in dementia. *Science*, 377(6608), eabi8654.  
<https://doi.org/10.1126/science.abi8654>
- Core, L. J., Martins, A. L., Danko, C. G., Waters, C. T., Siepel, A., & Lis, J. T. (2014). Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers [Research Support, N.I.H., Extramural]. *Nat Genet*, 46(12), 1311-1320. <https://doi.org/10.1038/ng.3142>
- Corrales, M., Rosado, A., Cortini, R., van Arensbergen, J., van Steensel, B., & Filion, G. J. (2017). Clustering of Drosophila housekeeping promoters facilitates their expression. *Genome Res*, 27(7), 1153-1161.  
<https://doi.org/10.1101/gr.211433.116>
- Cramer, P. (2019). Organization and regulation of gene transcription. *Nature*, 573(7772), 45-54. <https://doi.org/10.1038/s41586-019-1517-4>
- Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227(5258), 561-563.  
<https://doi.org/10.1038/227561a0>
- Dao, L. T. M., Galindo-Albarran, A. O., Castro-Mondragon, J. A., Andrieu-Soler, C., Medina-Rivera, A., Souaid, C., Charbonnier, G., Griffon, A., Vanhille, L., Stephen, T., Alomairi, J., Martin, D., Torres, M., Fernandez, N., Soler, E., van Helden, J., Puthier, D., & Spicuglia, S. (2017). Genome-wide characterization of mammalian promoters with distal enhancer functions. *Nat Genet*, 49(7), 1073-1081.  
<https://doi.org/10.1038/ng.3884>
- Daugherty, M., Polanuyer, B., Farrell, M., Scholle, M., Lykidis, A., de Crécy-Lagard, V., & Osterman, A. (2002). Complete reconstitution of the human coenzyme A biosynthetic pathway via comparative genomics. *J Biol Chem*, 277(24), 21431-21439. <https://doi.org/10.1074/jbc.M201708200>
- de Almeida, B. P., Reiter, F., Pagani, M., & Stark, A. (2022). DeepSTARR predicts enhancer activity from DNA sequence and enables the de novo design of synthetic enhancers. *Nat Genet*, 54(5), 613-624. <https://doi.org/10.1038/s41588-022-01048-5>
- de Boer, C. G., & Taipale, J. (2024). Hold out the genome: a roadmap to solving the cis-regulatory code. *Nature*, 625(7993), 41-50. <https://doi.org/10.1038/s41586-023-06661-w>
- de Langen, P., Hammal, F., Guéret, E., Mouren, J. C., Spinelli, L., & Ballester, B. (2023). Characterizing intergenic transcription at RNA polymerase II binding sites in normal and cancer tissues. *Cell Genom*, 3(10), 100411.  
<https://doi.org/10.1016/j.xgen.2023.100411>
- Deans, P. J. M., & Brennand, K. J. (2023). Better together: Non-additive interactions between schizophrenia risk genes. *Cell Genom*, 3(9), 100403.  
<https://doi.org/10.1016/j.xgen.2023.100403>
- Deng, W., Bai, Y., Deng, F., Pan, Y., Mei, S., Zheng, Z., Min, R., Wu, Z., Li, W., Miao, R., Zhang, Z., Kupper, T. S., Lieberman, J., & Liu, X. (2022). Streptococcal pyrogenic exotoxin B cleaves GSDMA and triggers pyroptosis. *Nature*, 602(7897), 496-502.  
<https://doi.org/10.1038/s41586-021-04384-4>

- Deplancke, B., Alpern, D., & Gardeux, V. (2016). The Genetics of Transcription Factor DNA Binding Variation [Research Support, Non-U.S. Gov't Review]. *Cell*, 166(3), 538-554. <https://doi.org/10.1016/j.cell.2016.07.012>
- Diao, Y., Fang, R., Li, B., Meng, Z., Yu, J., Qiu, Y., Lin, K. C., Huang, H., Liu, T., Marina, R. J., Jung, I., Shen, Y., Guan, K. L., & Ren, B. (2017). A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. *Nat Methods*, 14(6), 629-635. <https://doi.org/10.1038/nmeth.4264>
- Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C. P., Jerby-Arnon, L., Marjanovic, N. D., Dionne, D., Burks, T., Raychowdhury, R., Adamson, B., Norman, T. M., Lander, E. S., Weissman, J. S., Friedman, N., & Regev, A. (2016). Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell*, 167(7), 1853-1866 e1817. <https://doi.org/10.1016/j.cell.2016.11.038>
- Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., & Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398), 376-380. <https://doi.org/10.1038/nature11082>
- Dong, S., Zhao, N., Spragins, E., Kagda, M. S., Li, M., Assis, P., Jolanki, O., Luo, Y., Cherry, J. M., Boyle, A. P., & Hitz, B. C. (2023). Annotating and prioritizing human non-coding variants with RegulomeDB v.2. *Nat Genet*, 55(5), 724-726. <https://doi.org/10.1038/s41588-023-01365-3>
- Dusi, S., Valletta, L., Haack, T. B., Tsuchiya, Y., Venco, P., Pasqualato, S., Goffrini, P., Tigano, M., Demchenko, N., Wieland, T., Schwarzmayr, T., Strom, T. M., Invernizzi, F., Garavaglia, B., Gregory, A., Sanford, L., Hamada, J., Bettencourt, C., Houlden, H., . . . Tiranti, V. (2014). Exome sequence reveals mutations in CoA synthase as a cause of neurodegeneration with brain iron accumulation. *Am J Hum Genet*, 94(1), 11-22. <https://doi.org/10.1016/j.ajhg.2013.11.008>
- El Awady, M. K., Bader El Din, N. G., Abdel Aziz Riad, M., Omran, M. H., Abdelhafez, T. H., Elbaz, T. M., Hunter, S. S., Dawood, R. M., & Abdel Aziz, A. O. (2014). Predictors of disease recurrence post living donor liver transplantation in end stage chronic HCV patients. *Dis Markers*, 2014, 202548. <https://doi.org/10.1155/2014/202548>
- Engreitz, J. M., Haines, J. E., Perez, E. M., Munson, G., Chen, J., Kane, M., McDonel, P. E., Guttman, M., & Lander, E. S. (2016). Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature*, 539(7629), 452-455. <https://doi.org/10.1038/nature20149>
- Esnault, C., Garcia-Oliver, E., El Aabidine, A. Z., Robert, M.-C., Magat, T., Gawron, K., Basyuk, E., Karpinska, M., Pigeot, A., Cucchiarini, A., Luo, Y., Verga, D., Mourad, R., Radulescu, O., Mergny, J.-L., Bertrand, E., & Andrau, J.-C. (2023). G-quadruplexes are promoter elements controlling nucleosome exclusion and RNA polymerase II pausing. *bioRxiv*, 2023.2002.2024.529838. <https://doi.org/10.1101/2023.02.24.529838>
- Esnault, C., Magat, T., Zine El Aabidine, A., Garcia-Oliver, E., Cucchiarini, A., Bouchouika, S., Lleres, D., Goerke, L., Luo, Y., Verga, D., Lacroix, L., Feil, R., Spicuglia, S., Mergny, J. L., & Andrau, J. C. (2023). G4access identifies G-quadruplexes and their associations with open chromatin and imprinting control regions. *Nat Genet*, 55(8), 1359-1369. <https://doi.org/10.1038/s41588-023-01437-4>
- Finucane, H. K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P. R., Anttila, V., Xu, H., Zang, C., Farh, K., Ripke, S., Day, F. R., ReproGen, C., Schizophrenia Working Group of the Psychiatric Genomics, C., Consortium, R., Purcell, S., Stahl, E., Lindstrom, S., Perry, J. R., . . . Price, A. L. (2015). Partitioning heritability by

- functional annotation using genome-wide association summary statistics. *Nat Genet*, 47(11), 1228-1235. <https://doi.org/10.1038/ng.3404>
- Frankel, N., Davis, G. K., Vargas, D., Wang, S., Payre, F., & Stern, D. L. (2010). Phenotypic robustness conferred by apparently redundant transcriptional enhancers. *Nature*, 466(7305), 490-493. <https://doi.org/10.1038/nature09158>
- Fredriksson, N. J., Ny, L., Nilsson, J. A., & Larsson, E. (2014). Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nature Genetics*, 46(12), 1258-1263. <https://doi.org/10.1038/ng.3141>
- Fulco, C. P., Munschauer, M., Anyoha, R., Munson, G., Grossman, S. R., Perez, E. M., Kane, M., Cleary, B., Lander, E. S., & Engreitz, J. M. (2016). Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science*, 354(6313), 769-773. <https://doi.org/10.1126/science.aag2445>
- Fulco, C. P., Nasser, J., Jones, T. R., Munson, G., Bergman, D. T., Subramanian, V., Grossman, S. R., Anyoha, R., Doughty, B. R., Patwardhan, T. A., Nguyen, T. H., Kane, M., Perez, E. M., Durand, N. C., Lareau, C. A., Stamenova, E. K., Aiden, E. L., Lander, E. S., & Engreitz, J. M. (2019). Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nature Genetics*, 51(12), 1664-1669. <https://doi.org/10.1038/s41588-019-0538-0>
- Gao, P., Xia, J. H., Sipeky, C., Dong, X. M., Zhang, Q., Yang, Y., Zhang, P., Cruz, S. P., Zhang, K., Zhu, J., Lee, H. M., Suleman, S., Giannareas, N., Liu, S., Consortium, P., Tammela, T. L. J., Auvinen, A., Wang, X., Huang, Q., . . . Wei, G. H. (2018). Biology and Clinical Implications of the 19q13 Aggressive Prostate Cancer Susceptibility Locus. *Cell*, 174(3), 576-589 e518. <https://doi.org/10.1016/j.cell.2018.06.003>
- Gao, T., & Qian, J. (2020). EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. *Nucleic Acids Res*, 48(D1), D58-D64. <https://doi.org/10.1093/nar/gkz980>
- Gasperini, M., Hill, A. J., McFaline-Figueroa, J. L., Martin, B., Kim, S., Zhang, M. D., Jackson, D., Leith, A., Schreiber, J., Noble, W. S., Trapnell, C., Ahituv, N., & Shendure, J. (2019). A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. *Cell*, 176(1-2), 377-390 e319. <https://doi.org/10.1016/j.cell.2018.11.029>
- Gasperini, M., Tome, J. M., & Shendure, J. (2020). Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nat Rev Genet*, 21(5), 292-310. <https://doi.org/10.1038/s41576-019-0209-0>
- Genomes Project, C., Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., & Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, 526(7571), 68-74. <https://doi.org/10.1038/nature15393>
- Gisselbrecht, S. S., Palagi, A., Kurland, J. V., Rogers, J. M., Ozadam, H., Zhan, Y., Dekker, J., & Bulyk, M. L. (2020). Transcriptional Silencers in *Drosophila* Serve a Dual Role as Transcriptional Enhancers in Alternate Cellular Contexts. *Mol Cell*, 77(2), 324-337 e328. <https://doi.org/10.1016/j.molcel.2019.10.004>
- Groschel, S., Sanders, M. A., Hoogenboezem, R., de Wit, E., Bouwman, B. A. M., Erpelinck, C., van der Velden, V. H. J., Havermans, M., Avellino, R., van Lom, K., Rombouts, E. J., van Duin, M., Dohner, K., Beverloo, H. B., Bradner, J. E., Dohner, H., Lowenberg, B., Valk, P. J. M., Bindels, E. M. J., . . . Delwel, R. (2014). A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia [Research Support, Non-U.S. Gov't]. *Cell*, 157(2), 369-381. <https://doi.org/10.1016/j.cell.2014.02.019>

- Gu, Z., Gu, L., Eils, R., Schlesner, M., & Brors, B. (2014). circlize Implements and enhances circular visualization in R. *Bioinformatics*, *30*(19), 2811-2812. <https://doi.org/10.1093/bioinformatics/btu393>
- Gupta, R. M., Hadaya, J., Trehan, A., Zekavat, S. M., Roselli, C., Klarin, D., Emdin, C. A., Hilvering, C. R. E., Bianchi, V., Mueller, C., Khera, A. V., Ryan, R. J. H., Engreitz, J. M., Issner, R., Shores, N., Epstein, C. B., de Laat, W., Brown, J. D., Schnabel, R. B., . . . Kathiresan, S. (2017). A Genetic Variant Associated with Five Vascular Diseases Is a Distal Regulator of Endothelin-1 Gene Expression. *Cell*, *170*(3), 522-533 e515. <https://doi.org/10.1016/j.cell.2017.06.049>
- Hammal, F., de Langen, P., Bergon, A., Lopez, F., & Ballester, B. (2022). ReMap 2022: a database of Human, Mouse, Drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments. *Nucleic Acids Res*, *50*(D1), D316-D325. <https://doi.org/10.1093/nar/gkab996>
- Han, D., Li, Y., Wang, L., Liang, X., Miao, Y., Li, W., Wang, S., & Wang, Z. (2024). Comparative analysis of models in predicting the effects of SNPs on TF-DNA binding using large-scale in vitro and in vivo data. *Brief Bioinform*, *25*(2). <https://doi.org/10.1093/bib/bbae110>
- Hansen, T., Fong, S., Capra, J. A., & Hodges, E. (2023). Human gene regulatory evolution is driven by the divergence of regulatory element function in both cis and trans. *bioRxiv*. <https://doi.org/10.1101/2023.02.14.528376>
- Hao, Y., Stuart, T., Kowalski, M. H., Choudhary, S., Hoffman, P., Hartman, A., Srivastava, A., Molla, G., Madad, S., Fernandez-Granda, C., & Satija, R. (2024). Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat Biotechnol*, *42*(2), 293-304. <https://doi.org/10.1038/s41587-023-01767-y>
- Heizmann, B., Kastner, P., & Chan, S. (2018). The Ikaros family in lymphocyte development. *Curr Opin Immunol*, *51*, 14-23. <https://doi.org/10.1016/j.coi.2017.11.005>
- Henriques, T., Scruggs, B. S., Inouye, M. O., Muse, G. W., Williams, L. H., Burkholder, A. B., Lavender, C. A., Fargo, D. C., & Adelman, K. (2018). Widespread transcriptional pausing and elongation control at enhancers. *Genes Dev*, *32*(1), 26-41. <https://doi.org/10.1101/gad.309351.117>
- Hodges, H. C., Stanton, B. Z., Cermakova, K., Chang, C. Y., Miller, E. L., Kirkland, J. G., Ku, W. L., Veverka, V., Zhao, K., & Crabtree, G. R. (2018). Dominant-negative SMARCA4 mutants alter the accessibility landscape of tissue-unrestricted enhancers. *Nat Struct Mol Biol*, *25*(1), 61-72. <https://doi.org/10.1038/s41594-017-0007-3>
- Hong, J. W., Hendrix, D. A., & Levine, M. S. (2008). Shadow enhancers as a source of evolutionary novelty. *Science*, *321*(5894), 1314. <https://doi.org/10.1126/science.1160631>
- Hornung, V., Hartmann, R., Ablasser, A., & Hopfner, K. P. (2014). OAS proteins and cGAS: unifying concepts in sensing and responding to cytosolic nucleic acids. *Nat Rev Immunol*, *14*(8), 521-528. <https://doi.org/10.1038/nri3719>
- Hou, C., Li, L., Qin, Z. S., & Corces, V. G. (2012). Gene density, transcription, and insulators contribute to the partition of the Drosophila genome into physical domains. *Mol Cell*, *48*(3), 471-484. <https://doi.org/10.1016/j.molcel.2012.08.031>
- Hua, J. T., Ahmed, M., Guo, H., Zhang, Y., Chen, S., Soares, F., Lu, J., Zhou, S., Wang, M., Li, H., Larson, N. B., McDonnell, S. K., Patel, P. S., Liang, Y., Yao, C. Q., van der Kwast, T., Lupien, M., Feng, F. Y., Zoubeidi, A., . . . He, H. H. (2018). Risk SNP-Mediated Promoter-Enhancer Switching Drives Prostate Cancer through lncRNA PCAT19. *Cell*, *174*(3), 564-575 e518. <https://doi.org/10.1016/j.cell.2018.06.014>

- Huang, D., & Ovcharenko, I. (2022). Enhancer-silencer transitions in the human genome. *Genome Res*, 32(3), 437-448. <https://doi.org/10.1101/gr.275992.121>
- Hubisz, M. J., Pollard, K. S., & Siepel, A. (2011). PHAST and RPHAST: phylogenetic analysis with space/time models. *Brief Bioinform*, 12(1), 41-51. <https://doi.org/10.1093/bib/bbq072>
- Hussain, S., Sadouni, N., van Essen, D., Dao, L. T. M., Ferré, Q., Charbonnier, G., Torres, M., Gallardo, F., Lecellier, C. H., Sexton, T., Sacconi, S., & Spicuglia, S. (2023). Short tandem repeats are important contributors to silencer elements in T cells. *Nucleic Acids Res*. <https://doi.org/10.1093/nar/gkad187>
- Initiative, C.-H. G. (2021). Mapping the human genetic architecture of COVID-19. *Nature*, 600(7889), 472-477. <https://doi.org/10.1038/s41586-021-03767-x>
- Inoue, F., Kircher, M., Martin, B., Cooper, G. M., Witten, D. M., McManus, M. T., Ahituv, N., & Shendure, J. (2017). A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res*, 27(1), 38-52. <https://doi.org/10.1101/gr.212092.116>
- Iwata, T. N., Cowley, T. J., Sloma, M., Ji, Y., Kim, H., Qi, L., & Lee, S. S. (2013). The transcriptional co-regulator HCF-1 is required for INS-1 beta-cell glucose-stimulated insulin secretion. *PLoS ONE*, 8(11), e78841. <https://doi.org/10.1371/journal.pone.0078841>
- Javierre, B. M., Burren, O. S., Wilder, S. P., Kreuzhuber, R., Hill, S. M., Sewitz, S., Cairns, J., Wingett, S. W., Varnai, C., Thiecke, M. J., Burden, F., Farrow, S., Cutler, A. J., Rehnstrom, K., Downes, K., Grassi, L., Kostadima, M., Freire-Pritchett, P., Wang, F., . . . Fraser, P. (2016). Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell*, 167(5), 1369-1384 e1319. <https://doi.org/10.1016/j.cell.2016.09.037>
- Jolma, A., Yan, J., Whittington, T., Toivonen, J., Nitta, K. R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., Palin, K., Vaquerizas, J. M., Vincentelli, R., Luscombe, N. M., Hughes, T. R., Lemaire, P., Ukkonen, E., Kivioja, T., & Taipale, J. (2013). DNA-binding specificities of human transcription factors. *Cell*, 152(1-2), 327-339. <https://doi.org/10.1016/j.cell.2012.12.009>
- Joslin, A. C., Sobreira, D. R., Hansen, G. T., Sakabe, N. J., Aneas, I., Montefiori, L. E., Farris, K. M., Gu, J., Lehman, D. M., Ober, C., He, X., & Nóbrega, M. A. (2021). A functional genomics pipeline identifies pleiotropy and cross-tissue effects within obesity-associated GWAS loci. *Nature Communications*, 12(1), 5253. <https://doi.org/10.1038/s41467-021-25614-3>
- Jung, I., Schmitt, A., Diao, Y., Lee, A. J., Liu, T., Yang, D., Tan, C., Eom, J., Chan, M., Chee, S., Chiang, Z., Kim, C., Masliyah, E., Barr, C. L., Li, B., Kuan, S., Kim, D., & Ren, B. (2019). A compendium of promoter-centered long-range chromatin interactions in the human genome. *Nat Genet*, 51(10), 1442-1449. <https://doi.org/10.1038/s41588-019-0494-8>
- Kalita, C. A., Brown, C. D., Freiman, A., Isherwood, J., Wen, X., Pique-Regi, R., & Luca, F. (2018). High-throughput characterization of genetic effects on DNA-protein binding and gene transcription. *Genome Res*, 28(11), 1701-1708. <https://doi.org/10.1101/gr.237354.118>
- Kanamori-Katayama, M., Itoh, M., Kawaji, H., Lassmann, T., Katayama, S., Kojima, M., Bertin, N., Kaiho, A., Ninomiya, N., Daub, C. O., Carninci, P., Forrest, A. R., & Hayashizaki, Y. (2011). Unamplified cap analysis of gene expression on a single-molecule sequencer. *Genome Res*, 21(7), 1150-1159. <https://doi.org/10.1101/gr.115469.110>

- Kerimov, N., Hayhurst, J. D., Peikova, K., Manning, J. R., Walter, P., Kolberg, L., Samoviča, M., Sakthivel, M. P., Kuzmin, I., Trevanion, S. J., Burdett, T., Jupp, S., Parkinson, H., Papatheodorou, I., Yates, A. D., Zerbino, D. R., & Alasoo, K. (2021). A compendium of uniformly processed human gene expression and splicing quantitative trait loci. *Nat Genet*, 53(9), 1290-1299. <https://doi.org/10.1038/s41588-021-00924-w>
- Khetan, S., Kales, S., Kursawe, R., Jillette, A., Ulirsch, J. C., Reilly, S. K., Ucar, D., Tewhey, R., & Stitzel, M. L. (2021). Functional characterization of T2D-associated SNP effects on baseline and ER stress-responsive beta cell transcriptional activation. *Nat Commun*, 12(1), 5242. <https://doi.org/10.1038/s41467-021-25514-6>
- Kim, S., & Wysocka, J. (2023). Deciphering the multi-scale, quantitative cis-regulatory code. *Mol Cell*, 83(3), 373-392. <https://doi.org/10.1016/j.molcel.2022.12.032>
- Kim, T. K., & Shiekhattar, R. (2015). Architectural and Functional Commonalities between Enhancers and Promoters [Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Review]. *Cell*, 162(5), 948-959. <https://doi.org/10.1016/j.cell.2015.08.008>
- Kioussis, D., Vanin, E., deLange, T., Flavell, R. A., & Grosveld, F. G. (1983). Beta-globin gene inactivation by DNA translocation in gamma beta-thalassaemia. *Nature*, 306(5944), 662-666. <https://doi.org/10.1038/306662a0>
- Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A., & Liu, D. R. (2016). Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature*, 533(7603), 420-424. <https://doi.org/10.1038/nature17946>
- Kulzer, J. R., Stitzel, M. L., Morken, M. A., Huyghe, J. R., Fuchsberger, C., Kuusisto, J., Laakso, M., Boehnke, M., Collins, F. S., & Mohlke, K. L. (2014). A common functional regulatory variant at a type 2 diabetes locus upregulates ARAP1 expression in the pancreatic beta cell. *Am J Hum Genet*, 94(2), 186-197. <https://doi.org/10.1016/j.ajhg.2013.12.011>
- Kumar, S., Ambrosini, G., & Bucher, P. (2017). SNP2TFBS - a database of regulatory SNPs affecting predicted transcription factor binding site affinity. *Nucleic Acids Res*, 45(D1), D139-D144. <https://doi.org/10.1093/nar/gkw1064>
- Kvon, E. Z., Waymack, R., Gad, M., & Wunderlich, Z. (2021). Enhancer redundancy in development and disease. *Nat Rev Genet*, 22(5), 324-336. <https://doi.org/10.1038/s41576-020-00311-x>
- Laiker, I., & Frankel, N. (2022). Pleiotropic Enhancers are Ubiquitous Regulatory Elements in the Human Genome. *Genome Biol Evol*, 14(6). <https://doi.org/10.1093/gbe/evac071>
- Lane, E. A., Choi, D. W., Garcia-Haro, L., Levine, Z. G., Tedoldi, M., Walker, S., & Danial, N. N. (2019). HCF-1 Regulates De Novo Lipogenesis through a Nutrient-Sensitive Complex with ChREBP. *Mol Cell*, 75(2), 357-371 e357. <https://doi.org/10.1016/j.molcel.2019.05.019>
- Laquérière, A., Maluenda, J., Camus, A., Fontenas, L., Dieterich, K., Nolent, F., Zhou, J., Monnier, N., Latour, P., Gentil, D., Héron, D., Desguettes, I., Landrieu, P., Beneteau, C., Delaporte, B., Bellesme, C., Baumann, C., Capri, Y., Goldenberg, A., . . . Melki, J. (2014). Mutations in CNTNAP1 and ADCY6 are responsible for severe arthrogyrosis multiplex congenita with axogial defects. *Hum Mol Genet*, 23(9), 2279-2289. <https://doi.org/10.1093/hmg/ddt618>
- Lee, D., Gorkin, D. U., Baker, M., Strober, B. J., Asoni, A. L., McCallion, A. S., & Beer, M. A. (2015). A method to predict the impact of regulatory variants from DNA sequence. *Nat Genet*, 47(8), 955-961. <https://doi.org/10.1038/ng.3331>

- Lee, D., Shi, M., Moran, J., Wall, M., Zhang, J., Liu, J., Fitzgerald, D., Kyono, Y., Ma, L., White, K. P., & Gerstein, M. (2020). STARRPeaker: uniform processing and accurate identification of STARR-seq active regions. *Genome Biol*, 21(1), 298. <https://doi.org/10.1186/s13059-020-02194-x>
- Leporcq, C., Spill, Y., Balaramane, D., Toussaint, C., Weber, M., & Bardet, A. F. (2020). TFmotifView: a webserver for the visualization of transcription factor motifs in genomic regions. *Nucleic Acids Res*, 48(W1), W208-W217. <https://doi.org/10.1093/nar/gkaa252>
- Leung, D., Jung, I., Rajagopal, N., Schmitt, A., Selvaraj, S., Lee, A. Y., Yen, C. A., Lin, S., Lin, Y., Qiu, Y., Xie, W., Yue, F., Hariharan, M., Ray, P., Kuan, S., Edsall, L., Yang, H., Chi, N. C., Zhang, M. Q., . . . Ren, B. (2015). Integrative analysis of haplotype-resolved epigenomes across human tissues [Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't]. *Nature*, 518(7539), 350-354. <https://doi.org/10.1038/nature14217>
- Li, W., Wong, W. H., & Jiang, R. (2019). DeepTACT: predicting 3D chromatin contacts via bootstrapping deep learning. *Nucleic Acids Res*, 47(10), e60. <https://doi.org/10.1093/nar/gkz167>
- Liao, W. W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., Lu, S., Lucas, J. K., Monlong, J., Abel, H. J., Buonaiuto, S., Chang, X. H., Cheng, H., Chu, J., Colonna, V., Eizenga, J. M., Feng, X., Fischer, C., Fulton, R. S., . . . Paten, B. (2023). A draft human pangenome reference. *Nature*, 617(7960), 312-324. <https://doi.org/10.1038/s41586-023-05896-x>
- Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S., & Dekker, J. (2009). Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science*, 326(5950), 289-293. <https://doi.org/10.1126/science.1181369>
- Lim, J. K., Lisco, A., McDermott, D. H., Huynh, L., Ward, J. M., Johnson, B., Johnson, H., Pape, J., Foster, G. A., Krysztof, D., Follmann, D., Stramer, S. L., Margolis, L. B., & Murphy, P. M. (2009). Genetic variation in OAS1 is a risk factor for initial infection with West Nile virus in man. *PLoS Pathog*, 5(2), e1000321. <https://doi.org/10.1371/journal.ppat.1000321>
- Liu, L., & Kiryluk, K. (2018). Genome-wide polygenic risk predictors for kidney disease. *Nat Rev Nephrol*, 14(12), 723-724. <https://doi.org/10.1038/s41581-018-0067-6>
- Liu, S., Liu, Y., Zhang, Q., Wu, J., Liang, J., Yu, S., Wei, G. H., White, K. P., & Wang, X. (2017). Systematic identification of regulatory variants associated with cancer risk. *Genome Biol*, 18(1), 194. <https://doi.org/10.1186/s13059-017-1322-z>
- Liu, X., Li, X., & Yu, S. (2024). CFLAR: A novel diagnostic and prognostic biomarker in soft tissue sarcoma, which positively modulates the immune response in the tumor microenvironment. *Oncol Lett*, 27(4), 151. <https://doi.org/10.3892/ol.2024.14284>
- Liu, X., Li, Y. I., & Pritchard, J. K. (2019). Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. *Cell*, 177(4), 1022-1034 e1026. <https://doi.org/10.1016/j.cell.2019.04.014>
- Lizio, M., Harshbarger, J., Shimoji, H., Severin, J., Kasukawa, T., Sahin, S., Abugessaisa, I., Fukuda, S., Hori, F., Ishikawa-Kato, S., Mungall, C. J., Arner, E., Baillie, J. K., Bertin, N., Bono, H., de Hoon, M., Diehl, A. D., Dimont, E., Freeman, T. C., . . . consortium, F.

- (2015). Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol*, 16(1), 22. <https://doi.org/10.1186/s13059-014-0560-6>
- Long, E., Yin, J., Funderburk, K. M., Xu, M., Feng, J., Kane, A., Zhang, T., Myers, T., Golden, A., Thakur, R., Kong, H., Jessop, L., Kim, E. Y., Jones, K., Chari, R., Machiela, M. J., Yu, K., Melanoma Meta-Analysis, C., Iles, M. M., . . . Choi, J. (2022). Massively parallel reporter assays and variant scoring identified functional variants and target genes for melanoma loci and highlighted cell-type specificity. *Am J Hum Genet*, 109(12), 2210-2229. <https://doi.org/10.1016/j.ajhg.2022.11.006>
- Lu, F., Sossin, A., Abell, N., Montgomery, S. B., & He, Z. (2022). Deep learning-assisted genome-wide characterization of massively parallel reporter assays. *Nucleic Acids Res*, 50(20), 11442-11454. <https://doi.org/10.1093/nar/gkac990>
- Lu, X., Chen, X., Forney, C., Donmez, O., Miller, D., Parameswaran, S., Hong, T., Huang, Y., Pujato, M., Cazares, T., Miraldi, E. R., Ray, J. P., de Boer, C. G., Harley, J. B., Weirauch, M. T., & Kottyan, L. C. (2021). Global discovery of lupus genetic risk variant allelic enhancer activity. *Nat Commun*, 12(1), 1611. <https://doi.org/10.1038/s41467-021-21854-5>
- Lupianez, D. G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J. M., Laxova, R., Santos-Simarro, F., Gilbert-Dussardier, B., Wittler, L., Borschiwer, M., Haas, S. A., Osterwalder, M., Franke, M., Timmermann, B., Hecht, J., . . . Mundlos, S. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, 161(5), 1012-1025. <https://doi.org/10.1016/j.cell.2015.04.004>
- Lyu, J., Shao, R., Kwong Yung, P. Y., & Elsasser, S. J. (2022). Genome-wide mapping of G-quadruplex structures with CUT&Tag. *Nucleic Acids Res*, 50(3), e13. <https://doi.org/10.1093/nar/gkab1073>
- MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., Pendlington, Z. M., Welter, D., Burdett, T., Hindorff, L., Flicek, P., Cunningham, F., & Parkinson, H. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res*, 45(D1), D896-D901. <https://doi.org/10.1093/nar/gkw1133>
- Mackay, T. F. C., & Anholt, R. R. H. (2024). Pleiotropy, epistasis and the genetic architecture of quantitative traits. *Nat Rev Genet*. <https://doi.org/10.1038/s41576-024-00711-3>
- Malfait, J., Wan, J., & Spicuglia, S. (2023). Epromoters are new players in the regulatory landscape with potential pleiotropic roles. *Bioessays*, 45(10), e2300012. <https://doi.org/10.1002/bies.202300012>
- Malone, J., Holloway, E., Adamusiak, T., Kapushesky, M., Zheng, J., Kolesnikov, N., Zhukova, A., Brazma, A., & Parkinson, H. (2010). Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics*, 26(8), 1112-1118. <https://doi.org/10.1093/bioinformatics/btq099>
- Man, J., Barnett, P., & Christoffels, V. M. (2018). Structure and function of the Nppa–Nppb cluster locus during heart development and disease. *Cellular and Molecular Life Sciences*, 75(8), 1435-1444. <https://doi.org/10.1007/s00018-017-2737-0>
- Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., . . . Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265), 747-753. <https://doi.org/10.1038/nature08494>



- Mansour, M. R., Abraham, B. J., Anders, L., Berezovskaya, A., Gutierrez, A., Durbin, A. D., Etchin, J., Lawton, L., Sallan, S. E., Silverman, L. B., Loh, M. L., Hunger, S. P., Sanda, T., Young, R. A., & Look, A. T. (2014). Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element [Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S.]. *Science*, 346(6215), 1373-1377. <https://doi.org/10.1126/science.1259037>
- Martin, S., Cule, M., Basty, N., Tyrrell, J., Beaumont, R. N., Wood, A. R., Frayling, T. M., Sorokin, E., Whitcher, B., Liu, Y., Bell, J. D., Thomas, E. L., & Yaghootkar, H. (2021). Genetic Evidence for Different Adiposity Phenotypes and Their Opposing Influences on Ectopic Fat and Risk of Cardiometabolic Disease. *Diabetes*, 70(8), 1843-1856. <https://doi.org/10.2337/db21-0129>
- Martin, V., Zhao, J., Afek, A., Mielko, Z., & Gordan, R. (2019). QBiC-Pred: quantitative predictions of transcription factor binding changes due to sequence variants. *Nucleic Acids Res*, 47(W1), W127-W135. <https://doi.org/10.1093/nar/gkz363>
- Matos-Rodrigues, G., Hisey, J. A., Nussenzweig, A., & Mirkin, S. M. (2023). Detection of alternative DNA structures and its implications for human disease. *Mol Cell*, 83(20), 3622-3641. <https://doi.org/10.1016/j.molcel.2023.08.018>
- Mattioli, K., Volders, P. J., Gerhardinger, C., Lee, J. C., Maass, P. G., Mele, M., & Rinn, J. L. (2019). High-throughput functional analysis of lncRNA core promoters elucidates rules governing tissue specificity. *Genome Res*, 29(3), 344-355. <https://doi.org/10.1101/gr.242222.118>
- Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J., Shafer, A., Neri, F., Lee, K., Kutuyavin, T., Stehling-Sun, S., Johnson, A. K., Canfield, T. K., Giste, E., Diegel, M., . . . Stamatoyannopoulos, J. A. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science*, 337(6099), 1190-1195. <https://doi.org/10.1126/science.1222794>
- Medina-Rivera, A., Santiago-Algarra, D., Puthier, D., & Spicuglia, S. (2018). Widespread Enhancer Activity from Core Promoters. *Trends in Biochemical Sciences*, 43(6), 452-468. <https://doi.org/10.1016/j.tibs.2018.03.004>
- Merla, G., Howald, C., Antonarakis, S. E., & Reymond, A. (2004). The subcellular localization of the ChoRE-binding protein, encoded by the Williams-Beuren syndrome critical region gene 14, is regulated by 14-3-3. *Hum Mol Genet*, 13(14), 1505-1514. <https://doi.org/10.1093/hmg/ddh163>
- Mikhaylichenko, O., Bondarenko, V., Harnett, D., Schor, I. E., Males, M., Viales, R. R., & Furlong, E. E. M. (2018). The degree of enhancer or promoter activity is reflected by the levels and directionality of eRNA transcription. *Genes & Development*, 32(1), 42-57. <https://doi.org/10.1101/gad.308619.117>
- Mitchelmore, J., Grinberg, N. F., Wallace, C., & Spivakov, M. (2020). Functional effects of variation in transcription factor binding highlight long-range gene regulation by epromoters. *Nucleic Acids Research*, 48(6), 2866-2879. <https://doi.org/10.1093/nar/gkaa123>
- Mitchelmore, J., Grinberg, N. F., Wallace, C., & Spivakov, M. (2020). Functional effects of variation in transcription factor binding highlight long-range gene regulation by epromoters. *Nucleic Acids Res*, 48(6), 2866-2879. <https://doi.org/10.1093/nar/gkaa123>

- Mizusawa, N., Harada, N., Iwata, T., Ohigashi, I., Itakura, M., & Yoshimoto, K. (2022). Identification of protease serine S1 family member 53 as a mitochondrial protein in murine islet beta cells. *Islets*, *14*(1), 1-13. <https://doi.org/10.1080/19382014.2021.1982325>
- Morgan, B., Sun, L., Avitahl, N., Andrikopoulos, K., Ikeda, T., Gonzales, E., Wu, P., Neben, S., & Georgopoulos, K. (1997). Aiolos, a lymphoid restricted transcription factor that interacts with Ikaros to regulate lymphocyte differentiation. *EMBO J*, *16*, 2004-2013.
- Morris, J. A., Caragine, C., Daniloski, Z., Domingo, J., Barry, T., Lu, L., Davis, K., Ziosi, M., Glinos, D. A., Hao, S., Mimitou, E. P., Smibert, P., Roeder, K., Katsevich, E., Lappalainen, T., & Sanjana, N. E. (2023). Discovery of target genes and pathways at GWAS loci by pooled single-cell CRISPR screens. *Science*, *380*(6646), eadh7699. <https://doi.org/10.1126/science.adh7699>
- Mouri, K., Guo, M. H., de Boer, C. G., Lissner, M. M., Harten, I. A., Newby, G. A., DeBerg, H. A., Platt, W. F., Gentili, M., Liu, D. R., Campbell, D. J., Hacohen, N., Tewhey, R., & Ray, J. P. (2022). Prioritization of autoimmune disease-associated genetic variants that perturb regulatory element activity in T cells. *Nat Genet*, *54*(5), 603-612. <https://doi.org/10.1038/s41588-022-01056-5>
- Muerdter, F., Boryń, Ł. M., Woodfin, A. R., Neumayr, C., Rath, M., Zabidi, M. A., Pagani, M., Haberle, V., Kazmar, T., Catarino, R. R., Schernhuber, K., Arnold, C. D., & Stark, A. (2018). Resolving systematic errors in widely used enhancer activity assays in human cells. *Nature Methods*, *15*(2), 141-149. <https://doi.org/10.1038/nmeth.4534>
- Mumbach, M. R., Satpathy, A. T., Boyle, E. A., Dai, C., Gowen, B. G., Cho, S. W., Nguyen, M. L., Rubin, A. J., Granja, J. M., Kazane, K. R., Wei, Y., Nguyen, T., Greenside, P. G., Corces, M. R., Tycko, J., Simeonov, D. R., Suliman, N., Li, R., Xu, J., . . . Chang, H. Y. (2017). Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. *Nature Genetics*, *49*(11), 1602-1612. <https://doi.org/10.1038/ng.3963>
- Myint, L., Wang, R., Boukas, L., Hansen, K. D., Goff, L. A., & Avramopoulos, D. (2020). A screen of 1,049 schizophrenia and 30 Alzheimer's-associated variants for regulatory potential. *Am J Med Genet B Neuropsychiatr Genet*, *183*(1), 61-73. <https://doi.org/10.1002/ajmg.b.32761>
- Nalls, M. A., Pankratz, N., Lill, C. M., Do, C. B., Hernandez, D. G., Saad, M., DeStefano, A. L., Kara, E., Bras, J., Sharma, M., Schulte, C., Keller, M. F., Arepalli, S., Letson, C., Edsall, C., Stefansson, H., Liu, X., Pliner, H., Lee, J. H., . . . Group, A. G. A. (2014). Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nat Genet*, *46*(9), 989-993. <https://doi.org/10.1038/ng.3043>
- Nasser, J., Bergman, D. T., Fulco, C. P., Guckelberger, P., Doughty, B. R., Patwardhan, T. A., Jones, T. R., Nguyen, T. H., Ulirsch, J. C., Lekschas, F., Mualim, K., Natri, H. M., Weeks, E. M., Munson, G., Kane, M., Kang, H. Y., Cui, A., Ray, J. P., Eisenhaure, T. M., . . . Engreitz, J. M. (2021). Genome-wide enhancer maps link risk variants to disease genes. *Nature*, *593*(7858), 238-243. <https://doi.org/10.1038/s41586-021-03446-x>
- Naylor, L. H. (1999). Reporter gene technology: the future looks bright. *Biochem Pharmacol*, *58*(5), 749-757. [https://doi.org/10.1016/s0006-2952\(99\)00096-9](https://doi.org/10.1016/s0006-2952(99)00096-9)
- Nguyen, T. A., Jones, R. D., Snavely, A. R., Pfenning, A. R., Kirchner, R., Hemberg, M., & Gray, J. M. (2016). High-throughput functional comparison of promoter and enhancer

- activities. *Genome Research*, 26(8), 1023-1033. <https://doi.org/10.1101/gr.204834.116>
- Nisar, S., Torres, M., Thiam, A., Pouvelle, B., Rosier, F., Gallardo, F., Ka, O., Mbengue, B., Diallo, R. N., Brosseau, L., Spicuglia, S., Dieye, A., Marquet, S., & Rihet, P. (2022). Identification of ATP2B4 Regulatory Element Containing Functional Genetic Variants Associated with Severe Malaria. *International Journal of Molecular Sciences*, 23(9), 4849. <https://doi.org/10.3390/ijms23094849>
- Nisar, S., Torres, M., Thiam, A., Pouvelle, B., Rosier, F., Gallardo, F., Ka, O., Mbengue, B., Diallo, R. N., Brosseau, L., Spicuglia, S., Dieye, A., Marquet, S., & Rihet, P. (2022). Identification of Identification of ATP2B4 Regulatory Element Containing Functional Genetic Variants Associated with Severe Malaria. *Int J Mol Sci*, 23(9). <https://doi.org/10.3390/ijms23094849>
- Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N. L., Meisig, J., Sedat, J., Gribnau, J., Barillot, E., Blüthgen, N., Dekker, J., & Heard, E. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, 485(7398), 381-385. <https://doi.org/10.1038/nature11049>
- Nowakowska, J., Olechnowicz, A., Langwiński, W., Koteluk, O., Lemańska, Ż., Józwiak, K., Kamiński, K., Łosiewski, W., Stegmayr, J., Wagner, D., Alsafadi, H. N., Lindstedt, S., Dziuba, M., Bielicka, A., Graczyk, Z., & Szczepankiewicz, A. (2023). Increased expression of ORMDL3 in allergic asthma: a case control and in vitro study. *J Asthma*, 60(3), 458-467. <https://doi.org/10.1080/02770903.2022.2056896>
- Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bizkadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., . . . Phillippy, A. M. (2022). The complete sequence of a human genome. *Science*, 376(6588), 44-53. <https://doi.org/10.1126/science.abj6987>
- Oh, E., Ahn, M., Afelik, S., Becker, T. C., Roep, B. O., & Thurmond, D. C. (2018). Syntaxin 4 Expression in Pancreatic  $\beta$ -Cells Promotes Islet Function and Protects Functional  $\beta$ -Cell Mass. *Diabetes*, 67(12), 2626-2639. <https://doi.org/10.2337/db18-0259>
- Oudelaar, A. M., & Higgs, D. R. (2021). The relationship between genome structure and function. *Nat Rev Genet*, 22(3), 154-168. <https://doi.org/10.1038/s41576-020-00303-x>
- Pairo-Castineira, E., Clohisey, S., Klaric, L., Bretherick, A. D., Rawlik, K., Pasko, D., Walker, S., Parkinson, N., Fourman, M. H., Russell, C. D., Furniss, J., Richmond, A., Gountouna, E., Wrobel, N., Harrison, D., Wang, B., Wu, Y., Meynert, A., Griffiths, F., . . . Baillie, J. K. (2021). Genetic mechanisms of critical illness in COVID-19. *Nature*, 591(7848), 92-98. <https://doi.org/10.1038/s41586-020-03065-y>
- Pang, B., van Weerd, J. H., Hamoen, F. L., & Snyder, M. P. (2023). Identification of non-coding silencer elements and their regulation of gene expression. *Nat Rev Mol Cell Biol*, 24(6), 383-395. <https://doi.org/10.1038/s41580-022-00549-9>
- Pasquali, L., Gaulton, K. J., Rodriguez-Segui, S. A., Mularoni, L., Miguel-Escalada, I., Akerman, I., Tena, J. J., Moran, I., Gomez-Marin, C., van de Bunt, M., Ponsa-Cobas, J., Castro, N., Nammo, T., Cebola, I., Garcia-Hurtado, J., Maestro, M. A., Pattou, F., Piemonti, L., Berney, T., . . . Ferrer, J. (2014). Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants [Research Support, Non-U.S. Gov't]. *Nat Genet*, 46(2), 136-143. <https://doi.org/10.1038/ng.2870>
- Patwardhan, R. P., Lee, C., Litvin, O., Young, D. L., Pe'er, D., & Shendure, J. (2009). High-resolution analysis of DNA regulatory elements by synthetic saturation

- mutagenesis. *Nature Biotechnology*, 27(12), 1173-1175. <https://doi.org/10.1038/nbt.1589>
- Plaze, M., Attali, D., Prot, M., Petit, A. C., Blatzer, M., Vinckier, F., Levillayer, L., Chiaravalli, J., Perin-Dureau, F., Cachia, A., Friedlander, G., Chrétien, F., Simon-Loriere, E., & Gaillard, R. (2021). Inhibition of the replication of SARS-CoV-2 in human cells by the FDA-approved drug chlorpromazine. *Int J Antimicrob Agents*, 57(3), 106274. <https://doi.org/10.1016/j.ijantimicag.2020.106274>
- Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., & Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res*, 20(1), 110-121. <https://doi.org/10.1101/gr.097857.109>
- Preker, P., Nielsen, J., Kammler, S., Lykke-Andersen, S., Christensen, M. S., Mapendano, C. K., Schierup, M. H., & Jensen, T. H. (2008). RNA exosome depletion reveals transcription upstream of active human promoters. *Science*, 322(5909), 1851-1854. <https://doi.org/10.1126/science.1164096> [pii]
- Puranen, T., Poutanen, M., Ghosh, D., Vihko, P., & Vihko, R. (1997). Characterization of structural and functional properties of human 17 beta-hydroxysteroid dehydrogenase type 1 using recombinant enzymes and site-directed mutagenesis. *Mol Endocrinol*, 11(1), 77-86. <https://doi.org/10.1210/mend.11.1.9872>
- Pyrć, K., Milewska, A., Duran, E. B., Botwina, P., Dabrowska, A., Jedrysik, M., Benedyk, M., Lopes, R., Arenas-Pinto, A., Badr, M., Mellor, R., Kalber, T. L., Fernandez-Reyes, D., Schätzlein, A. G., & Uchegbu, I. F. (2021). SARS-CoV-2 inhibition using a mucoadhesive, amphiphilic chitosan that may serve as an anti-viral nasal spray. *Sci Rep*, 11(1), 20012. <https://doi.org/10.1038/s41598-021-99404-8>
- Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6), 841-842. <https://doi.org/10.1093/bioinformatics/btq033>
- Rada-Iglesias, A., Bajpai, R., Swigut, T., Bruggmann, S. A., Flynn, R. A., & Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. *Nature*, 470(7333), 279-283. <https://doi.org/10.1038/nature09692>
- Rajagopal, N., Srinivasan, S., Kooshesh, K., Guo, Y., Edwards, M. D., Banerjee, B., Syed, T., Emons, B. J. M., Gifford, D. K., & Sherwood, R. I. (2016). High-throughput mapping of regulatory DNA. *Nature Biotechnology*, 34(2), 167-174. <https://doi.org/10.1038/nbt.3468>
- Rao, S. S., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S., & Aiden, E. L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7), 1665-1680. <https://doi.org/10.1016/j.cell.2014.11.021>
- Rees, H. A., & Liu, D. R. (2018). Base editing: precision chemistry on the genome and transcriptome of living cells. *Nat Rev Genet*, 19(12), 770-788. <https://doi.org/10.1038/s41576-018-0059-1>
- Rennie, S., Dalby, M., van Duin, L., & Andersson, R. (2018). Transcriptional decomposition reveals active chromatin architectures and cell specific regulatory interactions. *Nat Commun*, 9(1), 487. <https://doi.org/10.1038/s41467-017-02798-1>
- Replogle, J. M., Saunders, R. A., Pogson, A. N., Hussmann, J. A., Lenail, A., Guna, A., Mascibroda, L., Wagner, E. J., Adelman, K., Lithwick-Yanai, G., Iremadze, N., Oberstrass, F., Lipson, D., Bonnar, J. L., Jost, M., Norman, T. M., & Weissman, J. S. (2022). Mapping information-rich genotype-phenotype landscapes with genome-

- scale Perturb-seq. *Cell*, 185(14), 2559-2575 e2528.  
<https://doi.org/10.1016/j.cell.2022.05.013>
- Richardson, T. G., Sanderson, E., Palmer, T. M., Ala-Korpela, M., Ference, B. A., Davey Smith, G., & Holmes, M. V. (2020). Evaluating the relationship between circulating lipoprotein lipids and apolipoproteins with risk of coronary heart disease: A multivariable Mendelian randomisation analysis. *PLoS Med*, 17(3), e1003062.  
<https://doi.org/10.1371/journal.pmed.1003062>
- Roeder, R. G., & Rutter, W. J. (1969). Multiple forms of DNA-dependent RNA polymerase in eukaryotic organisms. *Nature*, 224(5216), 234-237.  
<https://doi.org/10.1038/224234a0>
- Rosati, J., Johnson, J., Stander, Z., White, A., Tortorelli, S., Bailey, D., Fong, C. T., & Lee, B. H. (2023). Progressive brain atrophy and severe neurodevelopmental phenotype in siblings with biallelic COASY variants. *Am J Med Genet A*, 191(3), 842-845.  
<https://doi.org/10.1002/ajmg.a.63076>
- Rost, S., Fregin, A., Ivaskevicius, V., Conzelmann, E., Hörtnagel, K., Pelz, H. J., Lappégard, K., Seifried, E., Scharrer, I., Tuddenham, E. G., Müller, C. R., Strom, T. M., & Oldenburg, J. (2004). Mutations in VKORC1 cause warfarin resistance and multiple coagulation factor deficiency type 2. *Nature*, 427(6974), 537-541.  
<https://doi.org/10.1038/nature02214>
- Rowley, M. J., & Corces, V. G. (2018). Organizational principles of 3D genome architecture. *Nat Rev Genet*, 19(12), 789-800. <https://doi.org/10.1038/s41576-018-0060-8>
- Rusu, V., Hoch, E., Mercader, J. M., Tenen, D. E., Gymrek, M., Hartigan, C. R., DeRan, M., von Grotthuss, M., Fontanillas, P., Spooner, A., Guzman, G., Deik, A. A., Pierce, K. A., Dennis, C., Clish, C. B., Carr, S. A., Wagner, B. K., Schenone, M., Ng, M. C. Y., . . . Consortium, S. T. D. (2017). Type 2 Diabetes Variants Disrupt Function of SLC16A11 through Two Distinct Mechanisms. *Cell*, 170(1), 199-212.e120.  
<https://doi.org/10.1016/j.cell.2017.06.011>
- Rusu, V., Hoch, E., Mercader, J. M., Tenen, D. E., Gymrek, M., Hartigan, C. R., DeRan, M., von Grotthuss, M., Fontanillas, P., Spooner, A., Guzman, G., Deik, A. A., Pierce, K. A., Dennis, C., Clish, C. B., Carr, S. A., Wagner, B. K., Schenone, M., Ng, M. C. Y., . . . Cortes, M. L. (2017). Type 2 Diabetes Variants Disrupt Function of SLC16A11 through Two Distinct Mechanisms. *Cell*, 170(1), 199-212.e120.  
<https://doi.org/10.1016/j.cell.2017.06.011>
- Sahu, B., Hartonen, T., Pihlajamaa, P., Wei, B., Dave, K., Zhu, F., Kaasinen, E., Lidschreiber, K., Lidschreiber, M., Daub, C. O., Cramer, P., Kivioja, T., & Taipale, J. (2022). Sequence determinants of human gene regulatory elements. *Nature Genetics*, 54(3), 283-294. <https://doi.org/10.1038/s41588-021-01009-4>
- Saint Just Ribeiro, M., Tripathi, P., Namjou, B., Harley, J. B., & Chepelev, I. (2022). Haplotype-specific chromatin looping reveals genetic interactions of regulatory regions modulating gene expression in 8p23.1. *Front Genet*, 13, 1008582.  
<https://doi.org/10.3389/fgene.2022.1008582>
- Saint Just Ribeiro, M., Tripathi, P., Namjou, B., Harley, J. B., & Chepelev, I. (2022). Haplotype-specific chromatin looping reveals genetic interactions of regulatory regions modulating gene expression in 8p23.1. *Frontiers in Genetics*, 13, 1008582.  
<https://doi.org/10.3389/fgene.2022.1008582>
- Sakaue, S., Kanai, M., Tanigawa, Y., Karjalainen, J., Kurki, M., Koshihara, S., Narita, A., Konuma, T., Yamamoto, K., Akiyama, M., Ishigaki, K., Suzuki, A., Suzuki, K., Obara, W., Yamaji, K., Takahashi, K., Asai, S., Takahashi, Y., Suzuki, T., . . . FinnGen. (2021). A cross-

- population atlas of genetic associations for 220 human phenotypes. *Nat Genet*, 53(10), 1415-1424. <https://doi.org/10.1038/s41588-021-00931-x>
- Santana-Garcia, W., Rocha-Acevedo, M., Ramirez-Navarro, L., Mbouamboua, Y., Thieffry, D., Thomas-Chollier, M., Contreras-Moreira, B., van Helden, J., & Medina-Rivera, A. (2019). RSAT variation-tools: An accessible and flexible framework to predict the impact of regulatory variants on transcription factor binding. *Comput Struct Biotechnol J*, 17, 1415-1428. <https://doi.org/10.1016/j.csbj.2019.09.009>
- Santiago-Algarra, D., Dao, L. T. M., Pradel, L., España, A., & Spicuglia, S. (2017). Recent advances in high-throughput approaches to dissect enhancer function. *F1000Research*, 6, 939. <https://doi.org/10.12688/f1000research.11581.1>
- Santiago-Algarra, D., Souaid, C., Singh, H., Dao, L. T. M., Hussain, S., Medina-Rivera, A., Ramirez-Navarro, L., Castro-Mondragon, J. A., Sadouni, N., Charbonnier, G., & Spicuglia, S. (2021). Epromoters function as a hub to recruit key transcription factors required for the inflammatory response. *Nature Communications*, 12(1), 6660. <https://doi.org/10.1038/s41467-021-26861-0>
- Schaid, D. J., Chen, W., & Larson, N. B. (2018). From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat Rev Genet*, 19(8), 491-504. <https://doi.org/10.1038/s41576-018-0016-z>
- Schoenfelder, S., Furlan-Magaril, M., Mifsud, B., Tavares-Cadete, F., Sugar, R., Javierre, B.-M., Nagano, T., Katsman, Y., Sakthidevi, M., Wingett, S. W., Dimitrova, E., Dimond, A., Edelman, L. B., Elderkin, S., Tabbada, K., Darbo, E., Andrews, S., Herman, B., Higgs, A., . . . Fraser, P. (2015). The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Research*, 25(4), 582-597. <https://doi.org/10.1101/gr.185272.114>
- Schraivogel, D., Gschwind, A. R., Milbank, J. H., Leonce, D. R., Jakob, P., Mathur, L., Korbel, J. O., Merten, C. A., Velten, L., & Steinmetz, L. M. (2020). Targeted Perturb-seq enables genome-scale genetic screens in single cells. *Nat Methods*, 17(6), 629-635. <https://doi.org/10.1038/s41592-020-0837-5>
- Schubert, J., Siekierska, A., Langlois, M., May, P., Huneau, C., Becker, F., Muhle, H., Suls, A., Lemke, J. R., de Kovel, C. G., Thiele, H., Konrad, K., Kawalia, A., Toliat, M. R., Sander, T., Rüschemdorf, F., Caliebe, A., Nagel, I., Kohl, B., . . . Consortium, E. R. (2014). Mutations in STX1B, encoding a presynaptic protein, cause fever-associated epilepsy syndromes. *Nat Genet*, 46(12), 1327-1332. <https://doi.org/10.1038/ng.3130>
- Sergeeva, I. A., Hooijkaas, I. B., Ruijter, J. M., van der Made, I., de Groot, N. E., van de Werken, H. J., Creemers, E. E., & Christoffels, V. M. (2016). Identification of a regulatory domain controlling the Nppa-Nppb gene cluster during heart development and stress [Research Support, Non-U.S. Gov't]. *Development*, 143(12), 2135-2146. <https://doi.org/10.1242/dev.132019>
- Sergeeva, I. A., Hooijkaas, I. B., Ruijter, J. M., van der Made, I., de Groot, N. E., van de Werken, H. J. G., Creemers, E. E., & Christoffels, V. M. (2016). Identification of a regulatory domain controlling the *Nppa-Nppb* gene cluster during heart development and stress. *Development*, dev.132019. <https://doi.org/10.1242/dev.132019>
- Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., & Cavalli, G. (2012). Three-dimensional folding and functional organization principles of the Drosophila genome [Research Support, Non-U.S. Gov't]. *Cell*, 148(3), 458-472. <https://doi.org/10.1016/j.cell.2012.01.010>
- Seyres, D., Darbo, E., Perrin, L., Herrmann, C., & González, A. (2016). LedPred: an R/bioconductor package to predict regulatory sequences using support vector

- machines. *Bioinformatics*, 32(7), 1091-1093. <https://doi.org/10.1093/bioinformatics/btv705>
- Shan, N., Wang, Z., & Hou, L. (2019). Identification of trans-eQTLs using mediation analysis with multiple mediators. *BMC Bioinformatics*, 20(Suppl 3), 126. <https://doi.org/10.1186/s12859-019-2651-6>
- Sheng, C., Yao, C., Wang, Z., Chen, H., Zhao, Y., Xu, D., Huang, H., Huang, W., & Chen, S. (2018). Cyclophilin J limits inflammation through the blockage of ubiquitin chain sensing. *Nat Commun*, 9(1), 4381. <https://doi.org/10.1038/s41467-018-06756-3>
- Shivshankar, P., Boyd, A. R., Le Saux, C. J., Yeh, I. T., & Orihuela, C. J. (2011). Cellular senescence increases expression of bacterial ligands in the lungs and is positively correlated with increased susceptibility to pneumococcal pneumonia. *Aging Cell*, 10(5), 798-806. <https://doi.org/10.1111/j.1474-9726.2011.00720.x>
- Sinnott-Armstrong, N., Sousa, I. S., Laber, S., Rendina-Ruedy, E., Nitter Dankel, S. E., Ferreira, T., Mellgren, G., Karasik, D., Rivas, M., Pritchard, J., Guntur, A. R., Cox, R. D., Lindgren, C. M., Hauner, H., Sallari, R., Rosen, C. J., Hsu, Y.-H., Lander, E. S., Kiel, D. P., & Claussnitzer, M. (2021). A regulatory variant at 3q21.1 confers an increased pleiotropic risk for hyperglycemia and altered bone mineral density. *Cell Metabolism*, 33(3), 615-628.e613. <https://doi.org/10.1016/j.cmet.2021.01.001>
- Smemo, S., Tena, J. J., Kim, K. H., Gamazon, E. R., Sakabe, N. J., Gómez-Marín, C., Aneas, I., Credidio, F. L., Sobreira, D. R., Wasserman, N. F., Lee, J. H., Puvindran, V., Tam, D., Shen, M., Son, J. E., Vakili, N. A., Sung, H. K., Naranjo, S., Acemel, R. D., . . . Nóbrega, M. A. (2014). Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature*, 507(7492), 371-375. <https://doi.org/10.1038/nature13138>
- Sobreira, D. R., & Nóbrega, M. A. (2021). Regulatory Landscapes of *Nppa* and *Nppb*. *Circulation Research*, 128(1), 130-132. <https://doi.org/10.1161/CIRCRESAHA.120.318495>
- Sollis, E., Mosaku, A., Abid, A., Buniello, A., Cerezo, M., Gil, L., Groza, T., Güneş, O., Hall, P., Hayhurst, J., Ibrahim, A., Ji, Y., John, S., Lewis, E., MacArthur, Jacqueline A. L., McMahon, A., Osumi-Sutherland, D., Panoutsopoulou, K., Pendlington, Z., . . . Harris, Laura W. (2023). The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Research*, 51(D1), D977-D985. <https://doi.org/10.1093/nar/gkac1010>
- Soskic, B., Cano-Gamez, E., Smyth, D. J., Ambridge, K., Ke, Z., Matte, J. C., Bossini-Castillo, L., Kaplanis, J., Ramirez-Navarro, L., Lorenc, A., Nakic, N., Esparza-Gordillo, J., Rowan, W., Wille, D., Tough, D. F., Bronson, P. G., & Trynka, G. (2022). Immune disease risk variants regulate gene expression dynamics during CD4. *Nat Genet*, 54(6), 817-826. <https://doi.org/10.1038/s41588-022-01066-3>
- Spessott, W. A., Sanmillan, M. L., Kulkarni, V. V., McCormick, M. E., & Giraudo, C. G. (2017). Syntaxin 4 mediates endosome recycling for lytic granule exocytosis in cytotoxic T-lymphocytes. *Traffic*, 18(7), 442-452. <https://doi.org/10.1111/tra.12490>
- Steinhaus, R., Robinson, P. N., & Seelow, D. (2022). FABIAN-variant: predicting the effects of DNA variants on transcription factor binding. *Nucleic Acids Res*, 50(W1), W322-W329. <https://doi.org/10.1093/nar/gkac393>
- Stuart, P. E., Nair, R. P., Ellinghaus, E., Ding, J., Tejasvi, T., Gudjonsson, J. E., Li, Y., Weidinger, S., Eberlein, B., Gieger, C., Wichmann, H. E., Kunz, M., Ike, R., Krueger, G. G., Bowcock, A. M., Mrowietz, U., Lim, H. W., Voorhees, J. J., Abecasis, G. R., . . . Elder, J. T. (2010). Genome-wide association analysis identifies three psoriasis susceptibility loci. *Nat Genet*, 42(11), 1000-1004. <https://doi.org/10.1038/ng.693>

- Su, C., Johnson, M. E., Torres, A., Thomas, R. M., Manduchi, E., Sharma, P., Mehra, P., Le Coz, C., Leonard, M. E., Lu, S., Hodge, K. M., Chesi, A., Pippin, J., Romberg, N., Grant, S. F. A., & Wells, A. D. (2020). Mapping effector genes at lupus GWAS loci using promoter Capture-C in follicular helper T cells. *Nature Communications*, *11*(1), 3294. <https://doi.org/10.1038/s41467-020-17089-5>
- Tewhey, R., Kotliar, D., Park, D. S., Liu, B., Winnicki, S., Reilly, S. K., Andersen, K. G., Mikkelsen, T. S., Lander, E. S., Schaffner, S. F., & Sabeti, P. C. (2016a). Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. *Cell*, *165*(6), 1519-1529. <https://doi.org/10.1016/j.cell.2016.04.027>
- Tewhey, R., Kotliar, D., Park, D. S., Liu, B., Winnicki, S., Reilly, S. K., Andersen, K. G., Mikkelsen, T. S., Lander, E. S., Schaffner, S. F., & Sabeti, P. C. (2016b). Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. *Cell*, *165*(6), 1519-1529. <https://doi.org/10.1016/j.cell.2016.04.027>
- Thiecke, M. J., Wutz, G., Muhar, M., Tang, W., Bevan, S., Malysheva, V., Stocsits, R., Neumann, T., Zuber, J., Fraser, P., Schoenfelder, S., Peters, J. M., & Spivakov, M. (2020). Cohesin-Dependent and -Independent Mechanisms Mediate Chromosomal Contacts between Promoters and Enhancers. *Cell Rep*, *32*(3), 107929. <https://doi.org/10.1016/j.celrep.2020.107929>
- Thomas, H. F., Kotova, E., Jayaram, S., Pilz, A., Romeike, M., Lackner, A., Penz, T., Bock, C., Leeb, M., Halbritter, F., Wysocka, J., & Buecker, C. (2021). Temporal dissection of an enhancer cluster reveals distinct temporal and functional contributions of individual elements. *Mol Cell*, *81*(5), 969-982.e913. <https://doi.org/10.1016/j.molcel.2020.12.047>
- Thorvaldsdottir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration [Research Support, N.I.H., Extramural Research Support, Non-U.S. Gov't]. *Brief Bioinform*, *14*(2), 178-192. <https://doi.org/10.1093/bib/bbs017>
- Tippens, N. D., Vihervaara, A., & Lis, J. T. (2018). Enhancer transcription: what, where, when, and why? [Review]. *Genes Dev*, *32*(1), 1-3. <https://doi.org/10.1101/gad.311605.118>
- Trizzino, M., Zucco, A., Deliard, S., Wang, F., Barbieri, E., Veglia, F., Gabrilovich, D., & Gardini, A. (2021). EGR1 is a gatekeeper of inflammatory enhancers in human macrophages. *Sci Adv*, *7*(3). <https://doi.org/10.1126/sciadv.aaz8836>
- Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., & Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, *1*(1). <https://doi.org/10.1038/s43586-021-00056-9>
- Uhlen, M., Hallstrom, B. M., Lindskog, C., Mardinoglu, A., Ponten, F., & Nielsen, J. (2016). Transcriptomics resources of human tissues and organs. *Mol Syst Biol*, *12*(4), 862. <https://doi.org/10.15252/msb.20155865>
- Ulirsch, J. C., Nandakumar, S. K., Wang, L., Giani, F. C., Zhang, X., Rogov, P., Melnikov, A., McDonel, P., Do, R., Mikkelsen, T. S., & Sankaran, V. G. (2016). Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits. *Cell*, *165*(6), 1530-1545. <https://doi.org/10.1016/j.cell.2016.04.048>
- Uyehara, C. M., & Apostolou, E. (2023). 3D enhancer-promoter interactions and multi-connected hubs: Organizational principles and functional roles. *Cell Reports*, 112068. <https://doi.org/10.1016/j.celrep.2023.112068>



- van Arensbergen, J., FitzPatrick, V. D., de Haas, M., Pagie, L., Sluimer, J., Bussemaker, H. J., & van Steensel, B. (2017). Genome-wide mapping of autonomous promoter activity in human cells. *Nature Biotechnology*, 35(2), 145-153. <https://doi.org/10.1038/nbt.3754>
- van Arensbergen, J., Pagie, L., FitzPatrick, V. D., de Haas, M., Baltissen, M. P., Comoglio, F., van der Weide, R. H., Teunissen, H., Vosa, U., Franke, L., de Wit, E., Vermeulen, M., Bussemaker, H. J., & van Steensel, B. (2019). High-throughput identification of human SNPs affecting regulatory element activity. *Nat Genet*, 51(7), 1160-1169.
- van Dijk, T., Ferdinandusse, S., Ruiters, J. P. N., Alders, M., Mathijssen, I. B., Parboosingh, J. S., Innes, A. M., Meijers-Heijboer, H., Poll-The, B. T., Bernier, F. P., Wanders, R. J. A., Lamont, R. E., & Baas, F. (2018). Biallelic loss of function variants in COASY cause prenatal onset pontocerebellar hypoplasia, microcephaly, and arthrogryposis. *Eur J Hum Genet*, 26(12), 1752-1758. <https://doi.org/10.1038/s41431-018-0233-0>
- Vanhille, L., Griffon, A., Maqbool, M. A., Zacarias-Cabeza, J., Dao, L. T. M., Fernandez, N., Ballester, B., Andrau, J. C., & Spicuglia, S. (2015). High-throughput and quantitative assessment of enhancer activity in mammals by CapStarr-seq. *Nature Communications*, 6(1), 6905. <https://doi.org/10.1038/ncomms7905>
- Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet*, 101(1), 5-22. <https://doi.org/10.1016/j.ajhg.2017.06.005>
- Vosa, U., Claringbould, A., Westra, H. J., Bonder, M. J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Yazar, S., Brugge, H., Oelen, R., de Vries, D. H., van der Wijst, M. G. P., Kasela, S., Pervjakova, N., Alves, I., Fave, M. J., Agbessi, M., . . . Franke, L. (2021). Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat Genet*, 53(9), 1300-1310. <https://doi.org/10.1038/s41588-021-00913-z>
- Wallace, C. (2021). A more accurate method for colocalisation analysis allowing for multiple causal variants. *PLoS Genet*, 17(9), e1009440. <https://doi.org/10.1371/journal.pgen.1009440>
- Wang, D., Liu, S., Warrell, J., Won, H., Shi, X., Navarro, F. C. P., Clarke, D., Gu, M., Emani, P., Yang, Y. T., Xu, M., Gandal, M. J., Lou, S., Zhang, J., Park, J. J., Yan, C., Rhie, S. K., Manakongtreecheep, K., Zhou, H., . . . Gerstein, M. B. (2018). Comprehensive functional genomic resource and integrative model for the human brain. *Science*, 362(6420). <https://doi.org/10.1126/science.aat8464>
- Wang, X., He, L., Goggin, S. M., Saadat, A., Wang, L., Sinnott-Armstrong, N., Claussnitzer, M., & Kellis, M. (2018). High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human. *Nature Communications*, 9(1), 5380. <https://doi.org/10.1038/s41467-018-07746-1>
- Wang, X., Xu, J., Zhang, B., Hou, Y., Song, F., Lyu, H., & Yue, F. (2021). Genome-wide detection of enhancer-hijacking events from chromatin interaction data in rearranged genomes. *Nat Methods*, 18(6), 661-668. <https://doi.org/10.1038/s41592-021-01164-w>
- Wang, Y., He, H., Liyanarachchi, S., Genutis, L. K., Li, W., Yu, L., Phay, J. E., Shen, R., Brock, P., & de la Chapelle, A. (2018). The role of SMAD3 in the genetic predisposition to papillary thyroid carcinoma. *Genetics in Medicine*, 20(9), 927-935. <https://doi.org/10.1038/gim.2017.224>
- Watanabe, K., Stringer, S., Frei, O., Umicevic Mirkov, M., de Leeuw, C., Polderman, T. J. C., van der Sluis, S., Andreassen, O. A., Neale, B. M., & Posthuma, D. (2019). A global

- overview of pleiotropy and genetic architecture in complex traits. *Nat Genet*, 51(9), 1339-1348. <https://doi.org/10.1038/s41588-019-0481-0>
- Westra, H. J., Peters, M. J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., Christiansen, M. W., Fairfax, B. P., Schramm, K., Powell, J. E., Zernakova, A., Zernakova, D. V., Veldink, J. H., Van den Berg, L. H., Karjalainen, J., Withoff, S., Uitterlinden, A. G., Hofman, A., Rivadeneira, F., . . . Franke, L. (2013). Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat Genet*, 45(10), 1238-1243. <https://doi.org/10.1038/ng.2756>
- Xia, Q., Chesi, A., Manduchi, E., Johnston, B. T., Lu, S., Leonard, M. E., Parlin, U. W., Rappaport, E. F., Huang, P., Wells, A. D., Blobel, G. A., Johnson, M. E., & Grant, S. F. A. (2016). The type 2 diabetes presumed causal variant within TCF7L2 resides in an element that controls the expression of ACSL5. *Diabetologia*, 59(11), 2360-2368. <https://doi.org/10.1007/s00125-016-4077-2>
- Xiao, J., Moon, M., Yan, L., Nian, M., Zhang, Y., Liu, C., Lu, J., Guan, H., Chen, M., Jiang, D., Jiang, H., Liu, P. P., & Li, H. (2012). Cellular FLICE-inhibitory protein protects against cardiac remodelling after myocardial infarction. *Basic Res Cardiol*, 107(1), 239. <https://doi.org/10.1007/s00395-011-0239-z>
- Xiaohong, W., Jun, Z., Hongmei, G., & Fan, Q. (2019). CFLAR is a critical regulator of cerebral ischaemia-reperfusion injury through regulating inflammation and endoplasmic reticulum (ER) stress. *Biomed Pharmacother*, 117, 109155. <https://doi.org/10.1016/j.biopha.2019.109155>
- Xu, Z., Wei, G., Chepelev, I., Zhao, K., & Felsenfeld, G. (2011). Mapping of INS promoter interactions reveals its role in long-range regulation of SYT8 transcription. *Nature Structural & Molecular Biology*, 18(3), 372-378. <https://doi.org/10.1038/nsmb.1993>
- Yagihara, N., Watanabe, H., Barnett, P., Duboscq-Bidot, L., Thomas, A. C., Yang, P., Ohno, S., Hasegawa, K., Kuwano, R., Chatel, S., Redon, R., Schott, J. J., Probst, V., Koopmann, T. T., Bezzina, C. R., Wilde, A. A., Nakano, Y., Aiba, T., Miyamoto, Y., . . . Makita, N. (2016). Variants in the SCN5A Promoter Associated With Various Arrhythmia Phenotypes. *J Am Heart Assoc*, 5(9). <https://doi.org/10.1161/JAHA.116.003644>
- Yan, H., Yang, W., Zhou, F., Li, X., Pan, Q., Shen, Z., Han, G., Newell-Fugate, A., Tian, Y., Majeti, R., Liu, W., Xu, Y., Wu, C., Allred, K., Allred, C., Sun, Y., & Guo, S. (2019). Estrogen Improves Insulin Sensitivity and Suppresses Gluconeogenesis via the Transcription Factor Foxo1. *Diabetes*, 68(2), 291-304. <https://doi.org/10.2337/db18-0638>
- Yan, J., Qiu, Y., Ribeiro Dos Santos, A. M., Yin, Y., Li, Y. E., Vinckier, N., Nariai, N., Benaglio, P., Raman, A., Li, X., Fan, S., Chiou, J., Chen, F., Frazer, K. A., Gaulton, K. J., Sander, M., Taipale, J., & Ren, B. (2021). Systematic analysis of binding of transcription factors to noncoding variants. *Nature*, 591(7848), 147-151. <https://doi.org/10.1038/s41586-021-03211-0>
- Yanai, I., Benjamin, H., Shmoish, M., Chalifa-Caspi, V., Shklar, M., Ophir, R., Bar-Even, A., Horn-Saban, S., Safran, M., Domany, E., Lancet, D., & Shmueli, O. (2005). Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics*, 21(5), 650-659. <https://doi.org/10.1093/bioinformatics/bti042>
- Yi, T., Wang, X., Kelly, L. M., An, J., Xu, Y., Sailer, A. W., Gustafsson, J. A., Russell, D. W., & Cyster, J. G. (2012). Oxysterol gradient generation by lymphoid stromal cells guides activated B cell movement during humoral responses. *Immunity*, 37(3), 535-548. <https://doi.org/10.1016/j.immuni.2012.06.015>

- Yu, X., Yang, L., Li, J., Li, W., Li, D., Wang, R., Wu, K., Chen, W., Zhang, Y., Qiu, Z., & Zhou, W. (2019). De Novo and Inherited SETD1A Variants in Early-onset Epilepsy. *Neurosci Bull*, 35(6), 1045-1057. <https://doi.org/10.1007/s12264-019-00400-w>
- Zabidi, M. A., Arnold, C. D., Schernhuber, K., Pagani, M., Rath, M., Frank, O., & Stark, A. (2015). Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature*, 518(7540), 556-559. <https://doi.org/10.1038/nature13994>
- Zangen, D., Kaufman, Y., Zeligson, S., Perlberg, S., Fridman, H., Kanaan, M., Abdulhadi-Atwan, M., Abu Libdeh, A., Gussow, A., Kisslov, I., Carmel, L., Renbaum, P., & Levy-Lahad, E. (2011). XX ovarian dysgenesis is caused by a PSMC3IP/HOP2 mutation that abolishes coactivation of estrogen-driven transcription. *Am J Hum Genet*, 89(4), 572-579. <https://doi.org/10.1016/j.ajhg.2011.09.006>
- Zaugg, J. B., Sahlén, P., Andersson, R., Alberich-Jorda, M., de Laat, W., Deplancke, B., Ferrer, J., Mandrup, S., Natoli, G., Plewczynski, D., Rada-Iglesias, A., & Spicuglia, S. (2022). Current challenges in understanding the role of enhancers in disease. *Nature Structural & Molecular Biology*. <https://doi.org/10.1038/s41594-022-00896-3>
- Zeberg, H., & Pääbo, S. (2021). A genomic region associated with protection against severe COVID-19 is inherited from Neandertals. *Proc Natl Acad Sci U S A*, 118(9). <https://doi.org/10.1073/pnas.2026309118>
- Zhang, P., Xia, J. H., Zhu, J., Gao, P., Tian, Y. J., Du, M., Guo, Y. C., Suleman, S., Zhang, Q., Kohli, M., Tillmans, L. S., Thibodeau, S. N., French, A. J., Cerhan, J. R., Wang, L. D., Wei, G. H., & Wang, L. (2018). High-throughput screening of prostate cancer risk loci by single nucleotide polymorphisms sequencing. *Nat Commun*, 9(1), 2022. <https://doi.org/10.1038/s41467-018-04451-x>
- Zhang, S., Zhang, H., Forrest, M. P., Zhou, Y., Sun, X., Bagchi, V. A., Kozlova, A., Santos, M. D., Piguel, N. H., Dionisio, L. E., Sanders, A. R., Pang, Z. P., He, X., Penzes, P., & Duan, J. (2023). Multiple genes in a single GWAS risk locus synergistically mediate aberrant synaptic development and function in human neurons. *Cell Genom*, 3(9), 100399. <https://doi.org/10.1016/j.xgen.2023.100399>
- Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., & Liu, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol*, 9(9), R137. <https://doi.org/10.1186/gb-2008-9-9-r137>
- Zhou, J., & Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods*, 12(10), 931-934. <https://doi.org/10.1038/nmeth.3547>
- Zhou, Z., He, H., Wang, K., Shi, X., Wang, Y., Su, Y., Li, D., Liu, W., Zhang, Y., Shen, L., Han, W., Ding, J., & Shao, F. (2020). Granzyme A from cytotoxic lymphocytes cleaves GSDMB to trigger pyroptosis in target cells. *Science*, 368(6494). <https://doi.org/10.1126/science.aaz7548>
- Zoonomia, C. (2020). A comparative genomics multitool for scientific discovery and conservation. *Nature*, 587(7833), 240-245. <https://doi.org/10.1038/s41586-020-2876-6>

# **ANNEXES**

## **Annexes description**

The Annexes include 5 publications which I was a co-author. Annex 1 is a review about Epromoters, for which I made all the figures and wrote the manuscript together with other authors. Annex 2 is an article for which collaborated with Sylvain Marcellini's group from University of Concepcion in Chile. My contribution in this article was to provide the GWAS resource and perform the GWAS enrichment analysis. Annex 3 is an article related to my master project and thesis but published during my PhD. My contribution to this article included manuscript writing and figures preparation, STARR-seq data analysis, enhancer identification, epigenomics data analysis, TF data collection, and methods comparison. Annex 4 is an article about Hi-C analysis in *Xenopus* embryos, which I contributed to Hi-C data analysis and TAD identification during my master. Annex 5 is an article about an improved Hi-C strategy, for which I contributed to Hi-C data analysis and TAD border calculation during my master.

## HYPOTHESES

## Insights &amp; Perspectives

# Epromoters are new players in the regulatory landscape with potential pleiotropic roles

Juliette Malfait<sup>1,2</sup> | Jing Wan<sup>1,2</sup> | Salvatore Spicuglia<sup>1,2</sup> 

<sup>1</sup>Aix-Marseille University, Inserm, TAGC, UMR1090, Marseille, France

<sup>2</sup>Equipe Labélisée Ligue Contre le Cancer, LIGUE, Marseille, France

**Correspondence**

Salvatore Spicuglia, Aix-Marseille University, Inserm, TAGC, Marseille, UMR1090, France.  
Email: salvatore.spicuglia@inserm.fr

**Funding information**

Institut National de la Santé et de la Recherche Médicale (INSERM); Ligue contre le Cancer (Equipe Labélisée Ligue 2023); ANR, Grant/Award Numbers: ANR-18-CE12-0019, ANR-17-CE12-0035; Bettencourt Schueller Foundation (Prix coup d'élan pour la recherche française); French Ministry of Education and Aix-Marseille University; Horizon 2020 research and innovation program, Grant/Award Number: 860002

**Abstract**

Precise spatiotemporal control of gene expression during normal development and cell differentiation is achieved by the combined action of proximal (promoters) and distal (enhancers) *cis*-regulatory elements. Recent studies have reported that a subset of promoters, termed Epromoters, works also as enhancers to regulate distal genes. This new paradigm opened novel questions regarding the complexity of our genome and raises the possibility that genetic variation within Epromoters has pleiotropic effects on various physiological and pathological traits by differentially impacting multiple proximal and distal genes. Here, we discuss the different observations pointing to an important role of Epromoters in the regulatory landscape and summarize the evidence supporting a pleiotropic impact of these elements in disease. We further hypothesize that Epromoter might represent a major contributor to phenotypic variation and disease.

**KEYWORDS**

diseases, enhancer, epromoter, gene regulation, pleiotropy, promoter, variants

**INTRODUCTION**

In higher eukaryotes, gene transcription is regulated through the involvement of regulatory elements that are located near the transcription start site (TSS), called promoters, and those that are located far from TSS, called enhancers. This classical definition implies that enhancers activate gene expression at a distance while promoters induce local gene expression. Although these two elements are distinguishable by their genomics and epigenomics characteristics, a strict dichotomy between *cis*-regulatory elements is being challenged by the broad mechanistic similarities between promoters and enhancers.<sup>[1,2]</sup> On one hand, active enhancers are able to initiate transcription and recruit general transcription factors as promoters do. On the other hand, several lines of evidence have shown that a subset of coding-gene promoters is able to function as bona fide enhancers (hereafter named Epromoters), as detailed below.

Given the potential regulation of proximal and distal genes by Epromoters, we hypothesized that genetic variation or mutation at Epromoters might play an important role in physiological and patho-

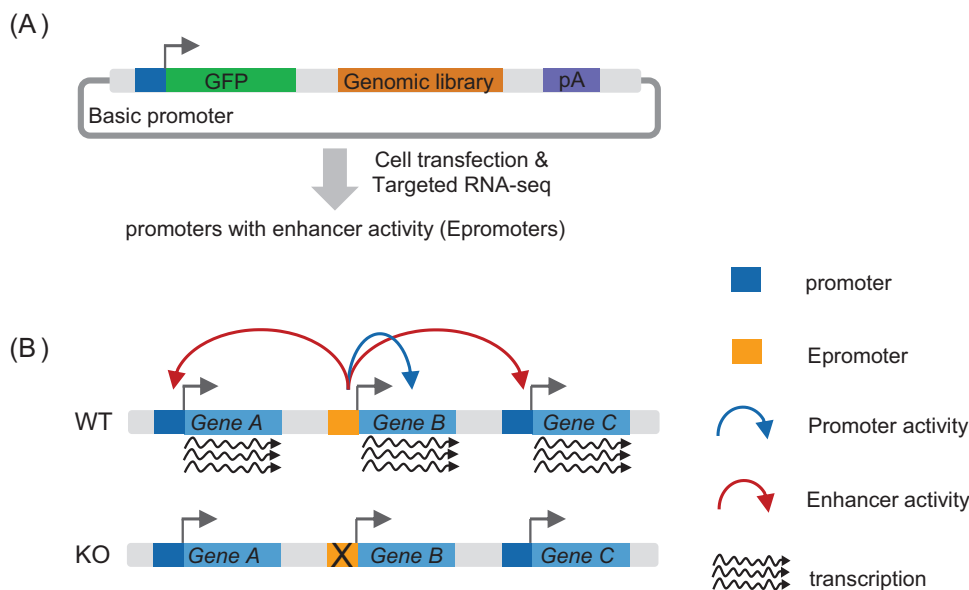
logical traits. In this review, we first describe the different studies supporting the physiological relevance of genomic regulation by Epromoters. We then discuss the intrinsic features that might drive the enhancer and promoter activity of Epromoters and whether these are shared or specific properties as compared with typical enhancers and promoters. Finally, we provide current observations supporting the hypothesis of an important and pleiotropic role of Epromoter variation on the ontogeny of different diseases.

**SHORT AND LONG-RANGE GENE REGULATION BY EPROMOTERS**

Dissection of *cis*-regulatory elements is classically based on gene reporter assays where the tested DNA regions are either placed immediately before the reporter gene (to assess for promoter activity) or placed upstream or downstream of a basic promoter (to assess for enhancer activity). Surprisingly, many of the early characterized enhancers in the 80s and 90s overlapped the promoter of inducible

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2023 The Authors. *BioEssays* published by Wiley Periodicals LLC.



**FIGURE 1** Proximal and distal gene regulation by Epromoters. (A) Example of high-throughput reporter assays to assess enhancer activity of genomic regions. A genomic library is cloned downstream of a basic promoter and a GFP reporter gene. With this construct, enhancers will activate their own transcription. The transcripts are then quantified to assess the enhancer activity. In particular, this type of strategy allows identifying of promoter elements with intrinsic enhancer activity (i.e., Epromoters). pA: Poly-adenylation. (B) In wild-type (WT) cells, the Epromoter (yellow) regulates a proximal gene (promoter activity; blue arrow) and distal genes (enhancer activity; red arrows). In Epromoter-deleted cells (KO: knock out), both the proximal and distal gene expressions are impaired.

genes,<sup>[2,3]</sup> including the first identified enhancer which corresponded to the promoter of the simian virus 40 (SV40) early gene.<sup>[4]</sup> Currently, powerful techniques incorporating high-throughput sequencing into reporter assays enable systematic and straightforward quantification of enhancer activity of *cis*-regulatory elements. Two similar high-throughput reporter assays have been widely used in recent years: Massively Parallel Reporter Assay (MPRA) and Self-Transcribing Active Regulatory Region sequencing (STARR-seq).<sup>[5]</sup> One striking observation of these episomal reporter assays, when assessing large genomic regions from *Drosophila* to different mammal cell types, is that many promoters display enhancer activity<sup>[6–15]</sup> (Figure 1A). In particular, by assessing all human core promoters of coding genes by STARR-seq, we found that ~3% of promoters exhibited enhancer activity in a given cell line.<sup>[8]</sup>

The presence of enhancer activity in some promoters, when tested in episomal reporter assays, does not necessarily mean that they could regulate other promoters in their endogenous context. Thus, whether gene promoters may function as *bona fide* enhancers by controlling distal gene expression is a critical issue. Several independent studies using mouse transgenics or CRISPR/Cas9 genome editing have demonstrated that the deletion or mutation of some promoters reduces the expression of a distally located gene,<sup>[8,16–19]</sup> implying they function as enhancers in their natural context (Figure 1B). So far, around 20 Epromoters have been validated experimentally at their endogenous loci by the different genome editing approaches in mouse or human cell lines (detailed in ref.[2]). High-throughput mapping of regulatory DNA by CRISPR-based screens also found evidence of distal gene regulation by gene promoters.<sup>[20–23]</sup> In particular, the repression of 30 (~8%) promoters out of 359 tested by CRISPR inactivation screen, resulted

in reduced expression of a distal gene,<sup>[22]</sup> supporting the idea that a substantial amount of promoters display enhancer function at their endogenous loci. Besides coding-gene promoters, promoters of long non-coding RNAs (lncRNAs) have also been shown to display enhancer activity independently of the transcript itself.<sup>[16,24,25]</sup> Nevertheless, in such cases, it is challenging to determine whether the tested regulatory element is a distal enhancer associated with a long transcript or rather the promoter of a “functional” lncRNA, which indirectly controls the expression of the neighbor gene.

Further evidence of Epromoter function comes from the analysis of 3D and genetic interactions. In addition to the enhancer-promoter interactions, the analyses of capture Hi-C and similarly derived 3C-based methods have shown that promoter-promoter (P-P) interactions are highly frequent.<sup>[26–32]</sup> Similarly, studies of expression quantitative trait loci (eQTLs) have revealed enrichment for genetic variants laying within gene promoters and associated with the regulation of distal genes.<sup>[8,17,29,33,34]</sup> Noticeably, some promoters interacting with other promoters, either at the 3D or genetic levels, indeed displayed enhancer activity.<sup>[8,17,23,27,33,35]</sup> An important outcome of these observations is that Epromoters often regulate several distal genes (in addition to the proximal one), including clusters of inducible genes,<sup>[18,32,36]</sup> suggesting that Epromoters might function as regulatory hubs for the coordinated regulation of gene clusters.

The impact on distal gene expression caused by the deletion or mutation of a promoter might, in principle, be caused by different mechanisms that are independent of a direct enhancer-like function. For instance, a given promoter might regulate distal genes by trans-effects, involving either the transcript itself or the protein-coding gene. However, re-expression of the *FAF2*<sup>[8]</sup> or *OAS3*<sup>[18]</sup> genes

associated with Epromoters did not rescue the expression of distal genes perturbed by the deletion of these Epromoters. Similarly, genetic dissection combining promoter deletion and the introduction of polyadenylation signals also provided evidence of direct enhancer-like functions.<sup>[16]</sup> Another possibility is related to the notion of transcription factories, whereby transcriptional hubs are known to contain many genes and their promoters, and are thought to share limited resources for their expression.<sup>[36–38]</sup> Here, the deletion of a promoter within the hub could affect the expression of other distal genes within that hub, without necessarily acting as an enhancer. Although this type of mechanism might be at play in some instances, we have demonstrated that in the case of the aforementioned *FAF2* and *OAS3* Epromoters, the deletion of the promoters associated with the distally regulated genes did not impact the expression of the Epromoter-associated genes, supporting the directionality of the regulation by these Epromoters.<sup>[8,18]</sup> Alternatively, the predicted Epromoter might be involved in a complex 3D organization, for example, involving the sequestering of another distal *cis*-regulatory element and, when deleted, it could indirectly affect the expression of a distal gene without acting as an enhancer for that gene. To date, few examples have unambiguously demonstrated the direct enhancer function of Epromoters, and whether the majority of Epromoters have a direct impact on distal gene expression or work in combination with the aforementioned mechanisms, will need further investigation.

All in all, the discovery that a subset of promoters also functions as bona fide enhancers, has important implications for our understanding of complex gene regulation in normal development and offers a rationale for the frequent repurposing of promoters and enhancers during mammalian evolution.<sup>[1,39,40]</sup> Additionally, they raise the intriguing possibility that sequence variation found within Epromoters may have an impact on diseases or physiological traits by directly impacting distal gene expression or changing their relative promoter and enhancer activities. A concomitant question is to understand which intrinsic features drive the specific enhancer and promoter function of Epromoters.

## EPROMOTERS SHARE ENHANCER AND PROMOTER PROPERTIES

Although typical enhancers and promoters are logically distinguished by their relative location with respect to the TSS of genes that they regulate, their shared architectural properties have suggested a unifying model of gene regulation by *cis*-regulatory elements.<sup>[1,2,41–43]</sup> Alike promoters, enhancers recruit RNA-Polymerase II (Pol II) and General Transcription Factors (GTF), and transcribe non-coding RNAs (eRNAs).<sup>[41,44–48]</sup> Promoters and enhancers are both demarcated by divergent transcription initiation, surrounded by a well-positioned array of nucleosomes, and enriched in core promoter elements.<sup>[41,45,48–51]</sup> However, transcripts generated by enhancers are generally bidirectional and less stable. Being generally depleted in CpG islands, enhancers recruit master regulators like CpG-poor promoters,<sup>[48]</sup> while some developmental enhancers require a prox-

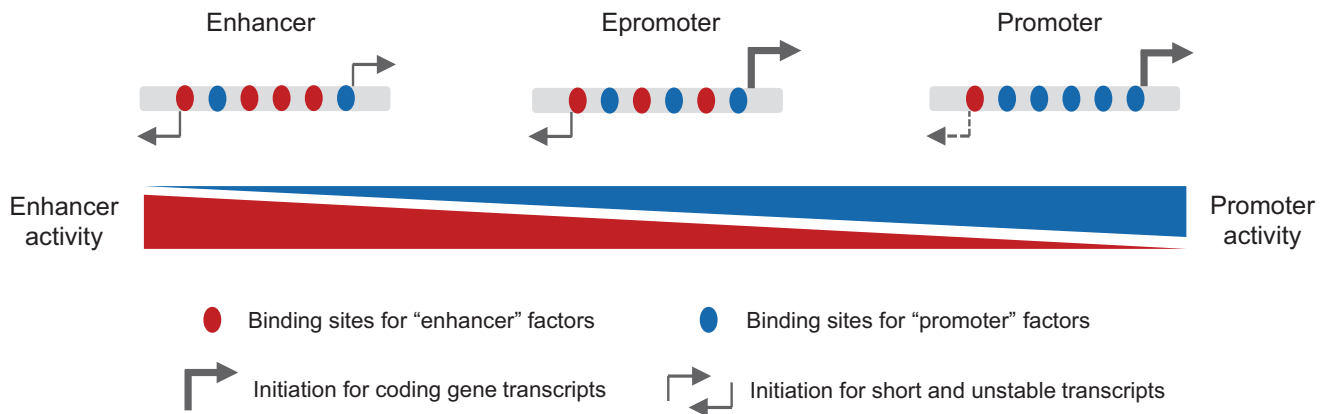
imal CpG island to function.<sup>[52]</sup> Histone post-translational modifications have been used to discriminate between enhancers and promoters.<sup>[53–55]</sup> For instance, gene promoters typically exhibit trimethylation of the histone H3 Lys4 (H3K4me3), while enhancers were found to be enriched in monomethylated H4K4 (H3K4me1) and acetylated H4K27 (H3K27ac). As a consequence, the presence of H3K27ac along with high levels of H3K4me1 and low H3K4me3 is commonly used as a proxy for active enhancers.<sup>[54]</sup> However, the level of H3K4me3 is positively correlated with the enhancer strength and eRNA level,<sup>[9,41,45,47,48,56,57]</sup> and, therefore, the presence of H3K4me3 is fully compatible with the enhancer activity. Thus, the relative enrichment in epigenetic modifications might simply indicate differences in transcriptional levels between the two types of elements, rather than reflecting mutually exclusive functions.

A main intrinsic difference between enhancers and promoters relates to the composition of transcription factor binding sites (TFBS). High promoter activity is associated with a high density of overlapping binding sites for different TFs, in particular, ubiquitously expressed ones, while enhancers are less constrained.<sup>[1]</sup> In addition to binding site complexity, the type of TF that binds a *cis*-regulatory element might influence the relative enhancer or promoter activity.<sup>[58]</sup> For instance, the binding of AP1 and NFY is associated with enhancer activity, whereas the binding of CREB, ETS and SP1 is preferentially associated with promoter activity.<sup>[13,59–62]</sup> This suggests that the nature of bound TFs might directly contribute to the enhancer and promoter properties of *cis*-regulatory elements.

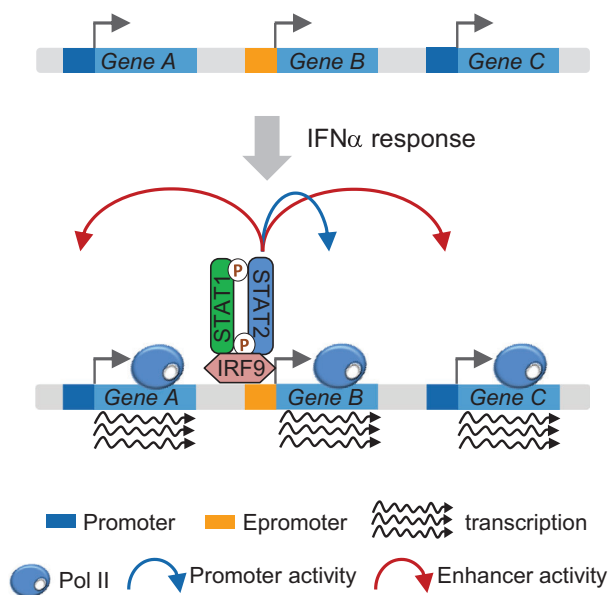
Given the aforementioned duality of *cis*-regulatory elements, current models do not propose that promoter and enhancer activities are mutually exclusive, rather different regulatory elements might accommodate different proportions of these activities<sup>[1]</sup> (**Figure 2**). In such scenarios, Epromoters represent remarkable examples of *cis*-regulatory elements that share functional and architectural properties with both types of *cis*-regulatory elements.<sup>[1,2]</sup>

High-throughput reporter assays have allowed a systematic comparison of *cis*-regulatory elements independently of their relative proximity with TSS.<sup>[5]</sup> These studies have revealed specific properties of Epromoters that distinguish them from typical enhancers and promoters. When compared to distal enhancers, Epromoters mainly differ by the type of associated TFBS. While distal enhancers are preferentially associated with TFBS for developmental and tissue-specific transcription factors (TFs), Epromoters appear to be associated with TFBS for ubiquitous or inducible TFs.<sup>[7,10,12,15,18]</sup> When compared to typical promoters, Epromoters differ by the higher level of unstable bidirectional transcripts,<sup>[8,45,50,51,63]</sup> association with co-activators, such as p300<sup>[8]</sup> and more frequent P-P interactions.<sup>[8,29]</sup> Moreover, Epromoters appear to be associated with a higher density of TF binding and a higher quality of binding sites.<sup>[8,18]</sup> The above results are consistent with the fact that Epromoters are preferentially associated with housekeeping and stress response genes,<sup>[7,8,10,12,15]</sup> including interferon-response genes.<sup>[8,12,18,64]</sup> For instance, we found that Epromoters are significantly associated with the induction of gene clusters during the inflammatory response and that induced Epromoters are characterized by a higher density and quality of Interferon-Stimulated





**FIGURE 2** A general model defining cis-regulatory function. Regulatory elements are composed of different ranges of binding sites for transcription factors associated with either enhancer (red) or promoter (blue) properties. Enhancers are mostly composed of binding sites associated with enhancer activity and initiate bidirectional transcription of short and unstable transcripts. Promoters are principally composed of binding sites associated with promoter activity and initiate strong unidirectional transcription towards the coding gene. Epromoters share structural features of both promoters and enhancers.



**FIGURE 3** Epromoters might function as regulatory hubs for the coordinated induction of gene clusters. During the interferon response, clusters of induced genes are frequently associated with Epromoters. In this context, the Epromoter recruits the key interferon-response TFs STAT1, STAT2 and IRF9, and simultaneously regulates co-induced genes within the same cluster. Legend is as in Figure 1.

Response Elements (ISRE), as compared with typically induced promoters, which, in turn, results in the Epromoter-specific recruitment of STAT1/2 and IRF TFs<sup>[18]</sup> (Figure 3). Moreover, inhibition of interferon signaling in HeLa cells drastically reduced the number of active Epromoters without affecting distal enhancers.<sup>[12,18,64]</sup> These results suggest that at least a subset of Epromoters plays an essential role in the coordination of rapid gene induction upon cellular response to

intra- and extra-cellular signals which might require a high efficiency of TF recruitment.

We speculate that Epromoters represent a combination of the two types of cis-regulatory elements, thus combining features that are associated with enhancer and promoter activities within an enhancer-promoter continuum of cis-regulatory elements (Figure 2). This intermediated position implies that Epromoters might display a higher density and complexity of TFBS because it has to accommodate the binding of TFs for both enhancer and promoter functions. In this scenario, typical promoters are enriched in binding sites for TF conferring promoter activity and enhancers enriched in binding sites for TF conferring enhancer activity, while Epromoters will be enriched for both types of binding sites leading to a higher density of TFBS. An alternative model will imply that Epromoters are associated with a unique combination of TFBS providing specific enhancer-promoter features. Future works should apply systematic assess the contribution of TFBS and associated transcription factors to enhancer and promoter activity in combination with machine learning models to understand the molecular features that determine the intrinsic promoter and enhancer potentials of cis-regulatory elements, and in particular of Epromoters.<sup>[1,10,65]</sup> This, in turn, might help to better predict the impact of mutations or natural variants of Epromoters that might affect either proximal or distal gene regulation.

## GENETIC VARIATION AT EPROMOTERS MIGHT HAVE PLEIOTROPIC ROLES

There is growing evidence that a wide range of human diseases is influenced by dysfunctions of cis-regulatory elements caused by genetic, structural, or epigenetic mechanisms.<sup>[66]</sup> These processes frequently underpin the susceptibility to common diseases but can be also directly involved in cancer or Mendelian diseases. The advent of genome-wide association studies (GWASs) in the past decade has been one of

the great endeavors in genomic research toward identifying genetic variants associated with candidate genes for common diseases. The majority of these genetic variants are found in non-coding regions, therefore are likely to be involved in regulatory mechanisms controlling gene expression.<sup>[67–69]</sup> However, a major challenge in interpreting the impact of genetic mutation or variation in disease is to identify the targets that are impacted by the genomic alteration, which are not necessarily the closest genes and might have confounding features.<sup>[70]</sup> Despite this, most studies select the closest gene to the associated GWAS variant to establish possible causal mechanisms, namely when the variant lies in the vicinity of a TSS or within an intronic region. However, this assumption has been shown to be biased in examples like the *FTO*<sup>[71]</sup> and *TCF7L2*<sup>[72]</sup> loci, in which obesity and type 2 diabetes GWAS data, respectively, were interpreted to implicate the nearest genes, while 3D epigenomics and functional follow-up showed that the disease variants reside in elements that regulate distant genes. In a similar way, it might be envisioned that GWAS variants lying within Epromoters might regulate the expression of distal disease-causing genes.

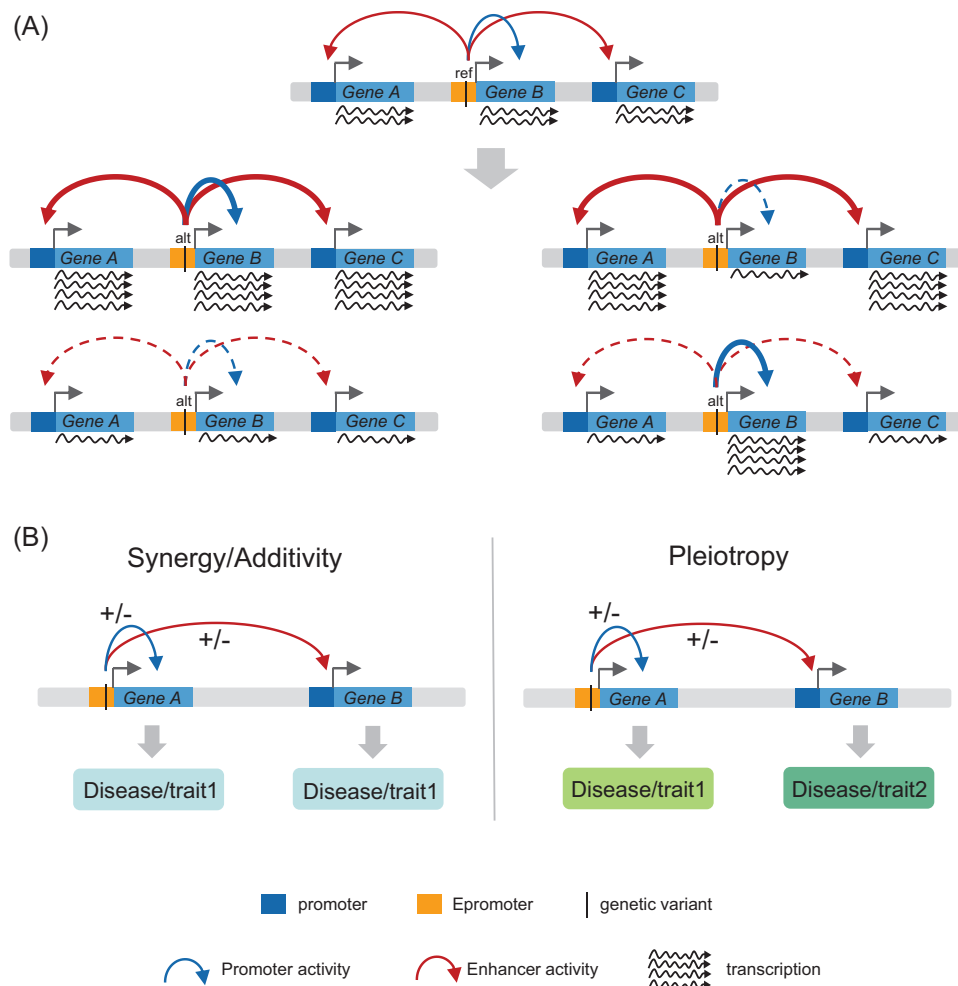
The discovery of Epromoters thus opens a new paradigm in the study of regulatory variants as a mutation in a promoter could potentially influence the expression of several genes or change the relative ratio of promoter versus enhancer activity. Thus, resulting in a variety of potential changes in the relative expression of neighboring genes (**Figure 4A**). In addition, it is plausible that the same *cis*-regulatory element displays preferential promoter activity in some tissues while displaying increased enhancer activity in other tissues, depending on the expressed combination of TFs and the epigenetic context.<sup>[8,33,73]</sup> For instance, Leung et al. found frequent examples of dynamic epigenetic switches where active promoters in one tissue displayed a histone modification signature of enhancers in other tissues/cell types.<sup>[73]</sup> Similarly, Chandra et al. found that a substantial number of promoter-promoter interactions involved transcriptionally inactive genes, suggesting that non-transcribing promoters may function as active enhancers for distal genes.<sup>[33]</sup>

The complex regulation by Epromoters might therefore have two predicted consequences. On one hand, there might be a general underestimation of the impact of Epromoter variation in disease because the causal gene might not be the closest one and therefore the link between genotype and phenotype might be missed in many case studies. On the other hand, as Epromoters potentially control several genes at the same time and efficiently recruit key TFs, mutations in these regulatory elements are expected to have a stronger pathological impact, as compared to typical promoters. This might result from the regulation of multiple genes either involved in the same (additive or synergistic effects) or different (pleiotropy) pathways (**Figure 4B**). Here, pleiotropy refers to a single *cis*-regulatory element affecting more than one trait independently.<sup>[74]</sup> Pleiotropy could be due to the perturbation of a single gene playing multiple functions in different tissues<sup>[75,76]</sup> or the regulation of multiple genes in the same or different tissues.<sup>[77,78]</sup> Several genomic features, such as a higher number of regulated genes and more abundance and diversity of encoded TFBS, are indicative of increasing variant pleiotropy.<sup>[78–80]</sup> Notably, these two features are

readily associated with both Epromoters and typical enhancers. Thus, it is fair to hypothesize that genetic alterations affecting Epromoters are likely to be involved in disease, as previously suggested,<sup>[2,18,27]</sup> and that Epromoters might have a stronger impact on the regulation of disease-associated genes, as compared with typical promoters. Although the hypothesis is not fully validated yet, several lines of observations support this assumption, as detailed below.

There are several pieces of genetic evidence indicating that promoter variants affecting distal genes are physiologically relevant. One way to connect GWAS-reported genetic variants with effects on gene function is to associate the genetic polymorphisms with eQTLs.<sup>[81]</sup> eQTLs with a higher probability to directly impact gene expression variation tend to be found in open chromatin regions, such as promoters and enhancers,<sup>[82]</sup> supporting the hypothesis of a possible effect through changes on regulatory mechanisms. Studies of natural genetic variation through eQTLs thus provide important insights into the mechanisms of specific diseases and gene control, and can point to the possible gene regulatory function of specific sequences based on their allelic associations with gene expression.<sup>[66]</sup>

Several studies have observed significant enrichment for eQTLs located within gene promoters which are associated with the regulation of distal genes,<sup>[8,17,29,33,83]</sup> pointing out to Epromoter-like regulation. A study, integrating predicted allelic variation in TF binding affinity in human lymphoblastic cell lines with their putative target genes inferred from Promoter Capture Hi-C, observed that a large proportion of regulatory variants associated with distal gene expression localized to the promoter regions of other genes, supporting the notion of Epromoters.<sup>[29]</sup> Interestingly, some of these variants were co-associated with the expression of both proximal and distal genes, while others were uniquely associated with distal genes. Using a set of Epromoters identified by STARR-seq, we observed that eQTLs lying within Epromoters are more likely to be associated with the expression of a distal gene as compared to other promoters and tend to have stronger effects on distal gene expression.<sup>[8]</sup> By analyzing promoter-centered long-range chromatin interactions in the human genome, Jung et al.<sup>[17]</sup> and Chandra et al.<sup>[33]</sup> found that P-P interactions were significantly enriched in eQTLs where a genetic variant in one of the interacting promoters was associated with the expression of the other interacting promoter. In these three studies, CRISPR-mediated gene editing recapitulated the predicted function of the promoter-associated eQTL variants in the regulation of distal gene expression.<sup>[8,17,33]</sup> A similar study found that the majority of genetic variation affecting TF binding at Epromoters in Lymphoblastoid cell lines was associated with the expression of distal genes alone, independently of whether the proximal gene was transcriptionally active.<sup>[29]</sup> While it is difficult to ensure that all promoter-distal eQTLs are bona fide distal regulators, we noticed that taking into consideration functional assays for enhancer activity (STARR-seq) allows to significantly enrich for promoter-eQTLs regulating distal genes with a higher probability of perturbing TF binding affinity.<sup>[8]</sup> Taken together, these findings support the functional significance of long-range transcriptional regulation by Epromoters and imply that regulatory variants within these elements may have both independent



**FIGURE 4** Effects of Epromoter variation on gene regulation. (A) Different potential impacts of Epromoter genetic variation on proximal and distal gene expression. In the reference (ref) haplotype, proximal and distal genes transcription are regulated by the Epromoter. In the alternative (alt) haplotypes, the promoter (blue arrows) and enhancer (red arrows) activities could increase (thicker arrows) or decrease (thinner and dashed arrows), resulting in up- or down-regulation of the associated genes. (B) Genetic variants at Epromoters might result in either synergistic/additive (all affected genes are involved in the same disease/trait) or pleiotropic (each affected gene is involved in a different disease/trait) effects.

and shared effects on the expression of their proximal and distal target genes.

Besides the global genomic evidence, several specific examples are pointing toward the relationship between disrupted Epromoters and variants associated with a variety of diseases, including autoimmunity (Crohn's disease, Lupus),<sup>[30,33,34,83,84]</sup> cardiovascular diseases,<sup>[19,85,86]</sup> diabetes,<sup>[2,35,87]</sup> infection diseases,<sup>[88,89]</sup> and cancer<sup>[90-93]</sup> (Table 1). However, to our best knowledge only in four cases, have the link between the disease-associated variant and the Epromoter function been experimentally validated<sup>[84,88,91,92]</sup> (Figure 5). In the first case, the alternative variant of a promoter-overlapping SNP associated with prostate cancer changes the relative affinity for two transcription factors resulting in promoter-enhancer switching and the corresponding increase of the expression of two distal transcripts directly involved in cancer progression (Figure 5A).<sup>[91,92]</sup> In the second case, the promoter of the *BAZ2B* gene was identified as Epromoter based on STARR-seq, while CRISPR-mediated deletion resulted in decreased expression of the *MARCHF7* gene located 95 kb

away.<sup>[8]</sup> The *BAZ2B* promoter overlaps with an SNP in eQTL with *MARCHF7*<sup>[8]</sup> and is associated with hypothyroidism (Ref.[94]; unpublished observation). Haplotype replacement by CRISPR-mediated homology recombination resulted in reduced expression of *MARCHF7*, but not *BAZ2B*,<sup>[8]</sup> suggesting it is the distal gene regulation of the identified Epromoter's SNP that is involved in the disease (Figure 5B). In the third case, haplotype-specific chromatin looping implicating genetic variants associated with Systemic Lupus Erythematosus (SLE) revealed that the alternative haplotype laying within the promoter of *BLK* gene decrease the promoter activity while increasing the long-range interaction with the *FAM167A* promoter resulting in the up-regulation of *FAM167A* expression (Figure 5C).<sup>[84]</sup> In the last case, a haplotype of five genetic variants associated with severe Malaria and laying within the internal promoter of the *ATP2B4* gene was found to switch the relative promoter and enhancer activity by luciferase reporter assays resulting in the increased expression of the long *ATP2B4* isoform initiated from the upstream promoter (Figure 5D).<sup>[88]</sup>

**TABLE 1** List of diseases associated variants in Epromoters.

Disease	Epromoter-containing variant <sup>a</sup>	Affected gene(s)	Evidence <sup>b</sup>	Ref.
Prostate cancer	PCAT19-short isoform	PCAT19-long & short isoforms, CEACAM21	P-P interaction, enhancer & promoter reporter assays, CRISPR-Cas9 genome editing, CRISPR interference/activation	[91, 92]
Hypothyroidism	BAZ2B	MARCHF7	P-P interaction, CRISPR-Cas9 genome editing, eQTL	[8]
Systemic lupus erythematosus	BLK	BLK, FAM167A	P-P interaction, CRISPR interference,	[84]
Severe malaria	ATP2B4-short isoform	ATP2B4-long & short isoforms	Enhancer & promoter reporter assays, CRISPR/Cas9 genome editing	[88]
Cardiovascular diseases	Nppb	Nppa	P-P interaction, Mouse transgenic models	[19, 86]
Type 2 diabetes	ARAP1	PDE2A	CapSTARR-seq, promoter reporter assays, eQTL	[2, 87]
Rheumatoid arthritis	CCR6	RNASET2	P-P interaction, eQTL	[33]
Systemic lupus erythematosus	TREH	CXCR5	P-P interaction, eQTL	[34]
Crohn disease	SMAD3	SMAD3, AAGAB	P-P interaction, eQTL	[30, 99]
Coronary artery disease	CDKN2B	IFNA2	P-P interaction	[85]
Multiple cancers	TERT	CLPTM1L	Somatic mutations, correlation with gene expression	[90]
Type 2 diabetes	INS	SYT8	P-P interaction, siRNA	[35]

<sup>a</sup>Name of the proximal gene associated with the Epromoter.

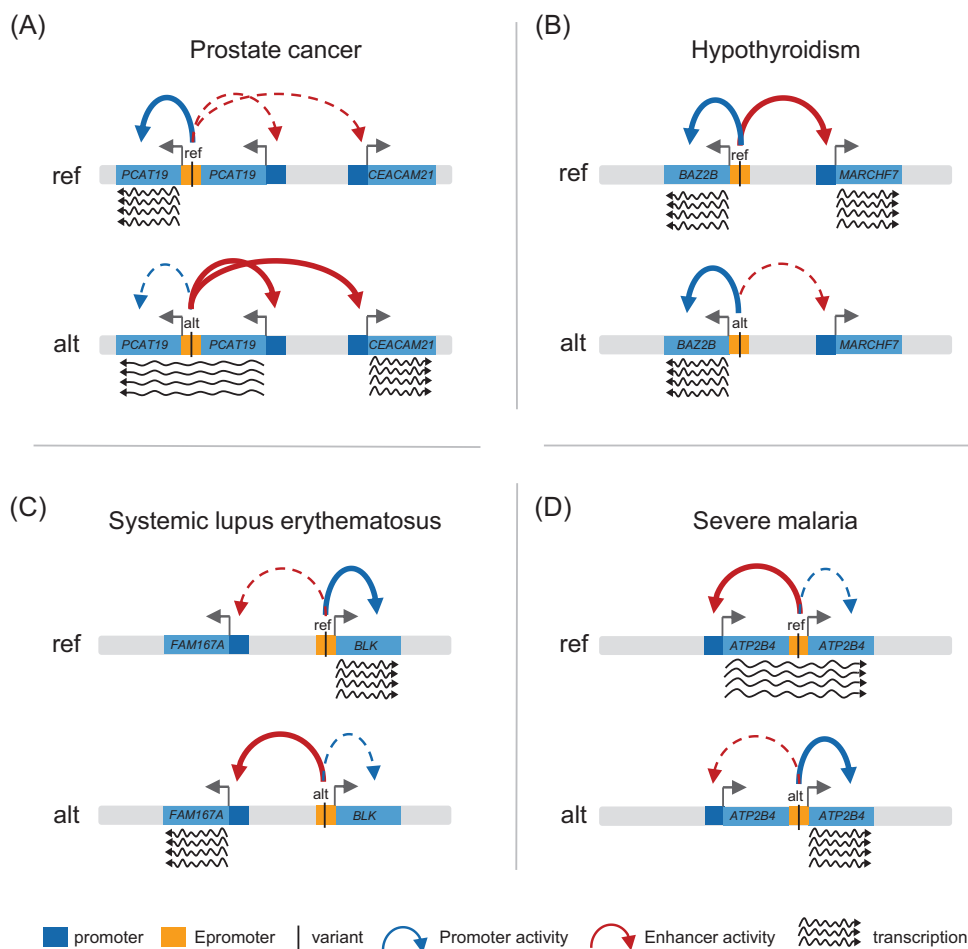
<sup>b</sup>Nature of the evidence suggesting an Epromoter-type of regulation.

In addition to small genetic variants, other types of genomic alterations involving enhancer repositioning (a phenomenon also named “enhancer hijacking”) by chromosomal translocations, genomic rearrangements, or insulator disruption, have been described as common molecular mechanisms resulting in disease-related gene deregulation, including overexpression of oncogenes.<sup>[66]</sup> Epromoter hijacking could likely impact diseases through related mechanisms. For instance, in T-acute lymphoblastic leukemia, a large intergenic deletion replaces the *TAL1* locus into the vicinity of the *CMPK1* promoter, which displays features of an active enhancer, resulting in *TAL1* oncogenic expression.<sup>[93]</sup> More generally, a study, using randomly integrated reporter constructs, found that chromosomal contacts with endogenous promoters

of housekeeping genes is required for the expression of the reporter gene,<sup>[95]</sup> supporting the idea whereby genomic repositioning due to structural variations might result in gene expression deregulation by Epromoter hijacking.

## CONCLUDING REMARKS AND FUTURE DIRECTIONS

Current results point to an important role of Epromoters in the regulatory landscape. These findings also open up the possibility that disease-associated variants or developmental traits lying within a



**FIGURE 5** Examples of Epromoter variation associated with diseases. (A) The variant rs11672691 is associated with prostate cancer and locate within the internal *PCAT19* promoter. The alternative variant switches the relative promoter and enhancer activity resulting in up-regulation of the most upstream *PCAT19* promoter and the distal gene *CEACAM21*. (B) The variant rs1046496 is associated with hypothyroidism and locate within the *BAZ2B* promoter. The alternative variant decreases the transcription of the *MARCHF7* gene. (C) The variant rs922483 is associated with systemic lupus erythematosus and locate within the *BLK* promoter. The alternative variant decreases the transcription of the *BLK* gene while increasing the expression of the *FAM167A* gene. (D) A haplotype of five variants containing the lead variant rs10900585 is associated with severe malaria and is located within the internal promoter of *ATP2B4*. The alternative variant switches the relative promoter and enhancer activity resulting in up-regulation of the most upstream *ATP2B4* promoter.

subset of promoters directly impact distal gene expression. Indeed, the recent observations support the hypothesis whereby Epromoters have a pleiotropic effect on diseases by perturbing the expression of several genes at the same time. Future works, including additional experimental settings where the Epromoter function is assessed in their endogenous loci, should tell us how commonly promoters are used as distal enhancers and better describe the physiological contexts where they are at play. Equally important will be to assess the physiological relevance of genes regulated by Epromoters. Are both proximal and distal gene deregulation directly involved in diseases? In particular, given the preponderant role of Epromoters in the regulation of interferon-response genes, it is expected that they might have an important impact on the etiology of inflammatory diseases.

Although several works have provided clear examples of natural genetic variation within promoters affecting the expression of a distal gene or an isoform regulated by a distal promoter, more

systematic studies are required to ascertain whether Epromoters indeed play a pleiotropic role in disease. While sequence-based models perform well in predicting how the impact of genetic variants in promoters affects local gene expression, they still perform very low to predict distal gene effects.<sup>[96]</sup> Therefore, we suggest that promoter-associated variants should systematically be tested for their proximal and distal effects. One possibility might be to simultaneously test the same DNA fragment in high-throughput reporter assays designed to assess the enhancer and the promoter activity in parallel.<sup>[13,50]</sup> This type of approach will help to elucidate whether enhancer and promoter activity of Epromoters are generally correlated amongst different cell types/tissues or whether they exhibit tissue-specific context. Similarly, it will help to assess whether genetic variants generally influence the global regulatory activity of Epromoters or rather affect their relative function to primarily work as a promoter or as an enhancer.

A central goal of biology is to decipher the cis-regulatory code that governs when and how much each gene is transcribed in a given genome and cellular state.<sup>[97]</sup> Two major advancements provide a paradigm shift in our capacity to integrate mechanistically informed, quantitative models of transcriptional regulation toward cracking the cis-regulatory code. On one side, the development of high-throughput reporter assays allows systematic and quantitative measurements of cis-regulatory activity. On the other side, the ability of recent deep learning models to learn the most relevant features from genomic data and to interpret and extract the features (e.g., DNA sequence) that underlie the predictions. These interpretable rules can then be used to decode the regulatory “syntax” or “grammar,” providing detailed information about the arrangement of TFBS, including their number, order, orientation and spacing. Several examples, disentangling cis-regulatory functions by combining high-throughput reporter assays with deep learning methods, have provided remarkable results.<sup>[62,65,98]</sup> Similar approaches aiming to dissect the intrinsic DNA features required for promoter and enhancer activities should be applied to the study of Epromoters. As Epromoters share features of both enhancer and promoter functions, a clear disentangling of both activities should lead to a better definition of the molecular bases governing gene regulation. In turn, understanding the genetic features driving proximal and distal activities will allow a better prediction of the impact of genetic variants with a pleiotropic effect on disease.

#### ACKNOWLEDGMENTS

Work in the laboratory of Salvatore Spicuglia was supported by recurrent funding from Institut National de la Santé et de la Recherche Médicale (INSERM) and Aix-Marseille University, and by specific grants from Ligue contre le Cancer (Equipe Labellisée Ligue 2023), ANR (ANR-18-CE12-0019 and ANR-17-CE12-0035) and Bettencourt Schueller Foundation (Prix coup d'élan pour la recherche française). Juliette Malfait is supported by a fellowship from the French Ministry of Education and Aix-Marseille University. Jing Wan is a fellow of the EU-funded Innovative Training Network “Molecular Basis of Human Enhanceropathies” and received funding from the Horizon 2020 research and innovation program under Marie Skłodowska-Curie grant agreement no. 860002.

#### CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

#### DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

#### ORCID

Salvatore Spicuglia  <https://orcid.org/0000-0002-8101-7108>

#### REFERENCES

- Andersson, R., & Sandelin, A. (2020). Determinants of enhancer and promoter activities of regulatory elements. *Nature Reviews Genetics*, 21, 71–87.
- Medina-Rivera, A., Santiago-Algarra, D., Puthier, D., & Spicuglia, S. (2018). Widespread enhancer activity from core promoters. *Trends in Biochemical Sciences*, 43, 452–468.
- Schaffner, W. (2015). Enhancers, enhancers – from their discovery to today's universe of transcription enhancers. *Biological Chemistry*, 396, 311–327.
- Banerji, J., Rusconi, S., & Schaffner, W. (1981). Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell*, 27, 299–308.
- Gasparini, M., Tome, J. M., & Shendure, J. (2020). Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nature Reviews Genetics*, 21, 292–310.
- Arnold, C. D., Gerlach, D., Stelzer, C., Boryń, Ł. M., Rath, M., & Stark, A. (2013). Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science*, 339, 1074–1077.
- Barakat, T. S., Halbritter, F., Zhang, M., Rendeiro, A. F., Perenthaler, E., Bock, C., & Chambers, I. (2018). Functional dissection of the enhancer repertoire in human embryonic stem cells. *Cell Stem Cell*, 23, 276–288.e8.
- Dao, L. T. M., Galindo-Albarrán, A. O., Castro-Mondragon, J. A., Andrieu-Soler, C., Medina-Rivera, A., Souaid, C., Charbonnier, G., Griffon, A., Vanhille, L., Stephen, T., Alomairi, J., Martin, D., Torres, M., Fernandez, N., Soler, E., Van Helden, J., Puthier, D., & Spicuglia, S. (2017). Genome-wide characterization of mammalian promoters with distal enhancer functions. *Nature Genetics*, 49, 1073–1081.
- Ernst, J., Melnikov, A., Zhang, X., Wang, L., Rogov, P., Mikkelsen, T. S., & Kellis, M. (2016). Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. *Nature Biotechnology*, 34, 1180–1190.
- Glaser, L. V., Steiger, M., Fuchs, A., Chung, H-R., Vingron, M., & Meijnsing, S. H. (2021). Assessing genome-wide dynamic changes in enhancer activity during early mESC differentiation by FAIRE-STARR-seq. *Nucleic Acids Research*, 49(21), 12178–12195.
- Liu, Y., Yu, S., Dhiman, V. K., Brunetti, T., Eckart, H., & White, K. P. (2017). Functional assessment of human enhancer activities using whole-genome STARR-sequencing. *Genome Biology*, 18, 219.
- Muerdter, F., Boryń, Ł. M., Woodfin, A. R., Neumayr, C., Rath, M., Zabidi, M. A., Pagani, M., Haberle, V., Kazmar, T., Catarino, R. R., Schernhuber, K., Arnold, C. D., & Stark, A. (2018). Resolving systematic errors in widely used enhancer activity assays in human cells. *Nature Methods*, 15, 141–149.
- Nguyen, T. A., Jones, R. D., Snavely, A. R., Pfenning, A. R., Kirchner, R., Hemberg, M., & Gray, J. M. (2016). High-throughput functional comparison of promoter and enhancer activities. *Genome Research*, 26, 1023–1033.
- Wang, X., He, L., Goggin, S. M., Saadat, A., Wang, L., Sinnott-Armstrong, N., Clausnitzer, M., & Kellis, M. (2018). High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human. *Nature Communications*, 9, 5380.
- Zabidi, M. A., Arnold, C. D., Schernhuber, K., Pagani, M., Rath, M., Frank, O., & Stark, A. (2015). Enhancer–core-promoter specificity separates developmental and housekeeping gene regulation. *Nature*, 518, 556–559.
- Engreitt, J. M., Haines, J. E., Perez, E. M., Munson, G., Chen, J., Kane, M., McDonel, P. E., Guttman, M., & Lander, E. S. (2016). Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature*, 539, 452–455.
- Jung, I., Schmitt, A., Diao, Y., Lee, A. J., Liu, T., Yang, D., Tan, C., Eom, J., Chan, M., Chee, S., Chiang, Z., Kim, C., Maslah, E., Barr, C. L., Li, B., Kuan, S., Kim, D., & Ren, B. (2019). A compendium of promoter-centered long-range chromatin interactions in the human genome. *Nature Genetics*, 51, 1442–1449.
- Santiago-Algarra, D., Souaid, C., Singh, H., Dao, L. T. M., Hussain, S., Medina-Rivera, A., Ramirez-Navarro, L., Castro-Mondragon, J. A., Sadouni, N., Charbonnier, G., & Spicuglia, S. (2021). Epromoters

- function as a hub to recruit key transcription factors required for the inflammatory response. *Nature Communications*, *12*, 6660.
19. Sergeeva, I. A., Hooijkaas, I. B., Ruijter, J. M., Van Der Made, I., De Groot, N. E., Van De Werken, H. J. G., Creemers, E. E., & Christoffels, V. M. (2016). Identification of a regulatory domain controlling the *Nppa-Nppb* gene cluster during heart development and stress. *Development*, *143*(12), 2135–2146.
  20. Fulco, C. P., Nasser, J., Jones, T. R., Munson, G., Bergman, D. T., Subramanian, V., Grossman, S. R., Anyoha, R., Doughty, B. R., Patwardhan, T. A., Nguyen, T. H., Kane, M., Perez, E. M., Durand, N. C., Lareau, C. A., Stamenova, E. K., Aiden, E. L., Lander, E. S., & Engreitz, J. M. (2019). Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nature Genetics*, *51*, 1664–1669.
  21. Rajagopal, N., Srinivasan, S., Kooshesh, K., Guo, Y., Edwards, M. D., Banerjee, B., Syed, T., Emons, B. J. M., Gifford, D. K., & Sherwood, R. I. (2016). High-throughput mapping of regulatory DNA. *Nature Biotechnology*, *34*, 167–174.
  22. Gasperini, M., Hill, A. J., Mcfaline-Figueroa, J. L., Martin, B., Kim, S., Zhang, M. D., Jackson, D., Leith, A., Schreiber, J., Noble, W. S., Trapnell, C., Ahituv, N., & Shendure, J. (2019). A genome-wide framework for mapping gene regulation via cellular genetic screens. *Cell*, *176*, 377–390.e19.e19.
  23. Diao, Y., Fang, R., Li, B., Meng, Z., Yu, J., Qiu, Y., Lin, K. C., Huang, H., Liu, T., Marina, R. J., Jung, I., Shen, Y., Guan, K. L., & Ren, B. (2017). A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. *Nature Methods*, *14*, 629–635.
  24. Kowalczyk, M. S., Hughes, J. R., Garrick, D., Lynch, M. D., Sharpe, J. A., Sloane-Stanley, J. A., MCGowan, S. J., De Gobbi, M., Hosseini, M., Vernimmen, D., Brown, J. M., Gray, N. E., Collavin, L., Gibbons, R. J., Flint, J., Taylor, S., Buckle, V. J., Milne, T. A., Wood, W. G., & Higgs, D. R. (2012). Intragenic enhancers act as alternative promoters. *Molecular Cell*, *45*, 447–458.
  25. Paralkar, V. R., Taborda, C. C., Huang, P., Yao, Y. u., Kossenkov, A. V., Prasad, R., Luan, J., Davies, J. O. J., Hughes, J. R., Hardison, R. C., Blobel, G. A., & Weiss, M. J. (2016). Unlinking an lncRNA from its associated cis element. *Molecular Cell*, *62*, 104–110.
  26. Javierre, B. M., Burren, O. S., Wilder, S. P., Kreuzhuber, R., Hill, S. M., Sewitz, S., Cairns, J., Wingett, S. W., Várnai, C., Thiecke, M. J., Burden, F., Farrow, S., Cutler, A. J., Rehnström, K., Downes, K., Grassi, L., Kostadima, M., Freire-Pritchett, P., Wang, F., & Flicek, P. (2016). Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. *Cell*, *167*, 1369–1384.e19.e19.
  27. Li, G., Ruan, X., Auerbach, R. K., Sandhu, K. S., Zheng, M., Wang, P., Poh, H. M., Goh, Y., Lim, J., Zhang, J., Sim, H. S., Peh, S. Q., Mulawadi, F. H., Ong, C. T., Orlov, Y. L., Hong, S., Zhang, Z., Landt, S., Raha, D., & Ruan, Y. (2012). Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, *148*, 84–98.
  28. Mifsud, B., Tavares-Cadete, F., Young, A. N., Sugar, R., Schoenfelder, S., Ferreira, L., Wingett, S. W., Andrews, S., Grey, W., Ewels, P. A., Herman, B., Happe, S., Higgs, A., Leproust, E., Follows, G. A., Fraser, P., Luscombe, N. M., & Osborne, C. S. (2015). Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nature Genetics*, *47*, 598–606.
  29. Mitchelmore, J., Grinberg, N. F., Wallace, C., & Spivakov, M. (2020). Functional effects of variation in transcription factor binding highlight long-range gene regulation by e-promoters. *Nucleic Acids Research*, *48*, 2866–2879.
  30. Mumbach, M. R., Satpathy, A. T., Boyle, E. A., Dai, C., Gowen, B. G., Cho, S. W., Nguyen, M. L., Rubin, A. J., Granja, J. M., Kazane, K. R., Wei, Y., Nguyen, T., Greenside, P. G., Corces, M. R., Tycko, J., Simeonov, D. R., Suliman, N., Li, R., Xu, J., & Chang, H. Y. (2017). Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. *Nature Genetics*, *49*, 1602–1612.
  31. Schoenfelder, S., Furlan-Magaril, M., Mifsud, B., Tavares-Cadete, F., Sugar, R., Javierre, B.-M., Nagano, T., Katsman, Y., Sakthidevi, M., Wingett, S. W., Dimitrova, E., Dimond, A., Edelman, L. B., Elderkin, S., Tabbada, K., Darbo, E., Andrews, S., Herman, B., Higgs, A., & Fraser, P. (2015). The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Research*, *25*, 582–597.
  32. Wen, J., Lagler, T. M., Sun, Q., Yang, Y., Chen, J., Harigaya, Y., Sankaran, V. G., Hu, M., Reiner, A. P., Raffield, L. M., & Li, Y. (2022). Super interactive promoters provide insight into cell type-specific regulatory networks in blood lineage cell types. *Plos Genetics*, *18*, e1009984.
  33. Chandra, V., Bhattacharyya, S., Schmiel, B. J., Madrigal, A., Gonzalez-Colin, C., Fotsing, S., Crinklaw, A., Seumo, G., Mohammadi, P., Kronenberg, M., Peters, B., Ay, F., & Vijayanand, P. (2021). Promoter-interacting expression quantitative trait loci are enriched for functional genetic variants. *Nature Genetics*, *53*, 110–119.
  34. Su, C., Johnson, M. E., Torres, A., Thomas, R. M., Manduchi, E., Sharma, P., Mehra, P., Le Coz, C., Leonard, M. E., Lu, S., Hodge, K. M., Chesi, A., Pippin, J., Romberg, N., Grant, S. F. A., & Wells, A. D. (2020). Mapping effector genes at lupus GWAS loci using promoter Capture-C in follicular helper T cells. *Nature Communications*, *11*, 3294.
  35. Xu, Z., Wei, G., Chepelev, I., Zhao, K., & Felsenfeld, G. (2011). Mapping of INS promoter interactions reveals its role in long-range regulation of SYT8 transcription. *Nature Structural & Molecular Biology*, *18*, 372–378.
  36. Zhu, I., Song, W., Ovcharenko, I., & Landsman, D. (2021). A model of active transcription hubs that unifies the roles of active promoters and enhancers. *Nucleic Acids Research*, *49*, 4493–4505.
  37. Feuerborn, A., & Cook, P. R. (2015). Why the activity of a gene depends on its neighbors. *Trends in Genetics*, *31*, 483–490.
  38. Ueyehara, C. M., & Apostolou, E. (2023). 3D enhancer-promoter interactions and multi-connected hubs: Organizational principles and functional roles. *Cell Reports*, *42*(4), 112068.
  39. Carelli, F. N., Liechti, A., Halbert, J., Warnefors, M., & Kaessmann, H. (2018). Repurposing of promoters and enhancers during mammalian evolution. *Nature Communications*, *9*, 4066.
  40. Wu, X., & Sharp, P. A. (2013). Divergent transcription: A driving force for new gene origination? *Cell*, *155*, 990–996.
  41. Core, L. J., Martins, A. L., Danko, C. G., Waters, C. T., Siepel, A., & Lis, J. T. (2014). Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nature Genetics*, *46*, 1311–1320.
  42. Kim, T. K., & Shiekhattar, R. (2015). Architectural and functional commonalities between enhancers and promoters. *Cell*, *162*, 948–959.
  43. Tippens, N. D., Vihervaara, A., & Lis, J. T. (2018). Enhancer transcription: What, where, when, and why? *Genes & Development*, *32*, 1–3.
  44. De Santa, F., Barozzi, I., Mietton, F., Ghisletti, S., Polletti, S., Tusi, B. K., Muller, H., Ragoussis, J., Wei, C. L., & Natoli, G. (2010). A large fraction of extragenic RNA Pol II transcription sites overlap enhancers. *Plos Biology*, *8*, e1000384.
  45. Henriques, T., Scruggs, B. S., Inouye, M. O., Muse, G. W., Williams, L. H., Burkholder, A. B., Lavender, C. A., Fargo, D. C., & Adelman, K. (2018). Widespread transcriptional pausing and elongation control at enhancers. *Genes & Development*, *32*, 26–41.
  46. Kim, T. K., Hemberg, M., Gray, J. M., Costa, A. M., Bear, D. M., Wu, J., Harmin, D. A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., Markenscoff-Papadimitriou, E., Kuhl, D., Bito, H., Worley, P. F., Kreiman, G., & Greenberg, M. E. (2010). Widespread transcription at neuronal activity-regulated enhancers. *Nature*, *465*, 182–187.
  47. Koch, F., Fenouil, R., Gut, M., Cauchy, P., Albert, T. K., Zacarias-Cabeza, J., Spicuglia, S., De La Chapelle, A. L., Heidemann, M., Hintermair, C., Eick, D., Gut, I., Ferrier, P., & Andrau, J.-C. (2011). Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters. *Nature Structural & Molecular Biology*, *18*, 956–963.
  48. Consortium, T. F., Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., & Sandelin, A. (2014). An atlas of active enhancers across human cell types and tissues. *Nature*, *507*, 455–461.

49. He, H. H., Meyer, C. A., Shin, H., Bailey, S. T., Wei, G., Wang, Q., Zhang, Y., Xu, K., Ni, M., Lupien, M., Mieczkowski, P., Lieb, J. D., Zhao, K., Brown, M., & Liu, X. S. (2010). Nucleosome dynamics define transcriptional enhancers. *Nature Genetics*, *42*, 343–347.
50. Mikhaylichenko, O., Bondarenko, V., Harnett, D., Schor, I. E., Males, M., Viales, R. R., & Furlong, E. E. M. (2018). The degree of enhancer or promoter activity is reflected by the levels and directionality of eRNA transcription. *Genes & Development*, *32*, 42–57.
51. Rennie, S., Dalby, M., Lloret-Llinares, M., Bakoulis, S., Vaagensø, C. D., Jensen, T. H., & Andersson, R. (2018). Transcription start site analysis reveals widespread divergent transcription in *D. melanogaster* and core promoter-encoded enhancer activities. *Nucleic Acids Research*, *46*(11), 5455–5469.
52. Pachano, T., Sánchez-Gaya, V., Ealo, T., Mariner-Faullí, M., Bleckwehl, T., Asenjo, H. G., Respuela, P., Cruz-Molina, S., Muñoz-San Martín, M., Haro, E., Van Ijcken, W. F. J., Landeira, D., & Rada-Iglesias, A. (2021). Orphan CpG islands amplify poised enhancer regulatory activity and determine target gene responsiveness. *Nature Genetics*, *53*, 1036–1049.
53. Creighton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., Hanna, J., Lodato, M. A., Frampton, G. M., Sharp, P. A., Boyer, L. A., Young, R. A., & Jaenisch, R. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences of the United States of America*, *107*, 21931–21936.
54. Heintzman, N. D., & Ren, B. (2009). Finding distal regulatory elements in the human genome. *Current Opinion in Genetics & Development*, *19*, 541–549.
55. Rada-Iglesias, A., Bajpai, R., Swigut, T., Bruggmann, S. A., Flynn, R. A., & Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. *Nature*, *470*, 279–283.
56. Pekowska, A., Benoukraf, T., Zacarias-Cabeza, J., Belhocine, M., Koch, F., Holota, H., Imbert, J., Andrau, J. C., Ferrier, P., & Spicuglia, S. (2011). H3K4 tri-methylation provides an epigenetic signature of active enhancers: Epigenetic signature of active enhancers. *The EMBO Journal*, *30*, 4198–4210.
57. Vanhille, L., Griffon, A., Maqbool, M. A., Zacarias-Cabeza, J., Dao, L. T. M., Fernandez, N., Ballester, B., Andrau, J. C., & Spicuglia, S. (2015). High-throughput and quantitative assessment of enhancer activity in mammals by CapStarr-seq. *Nature Communications*, *6*, 69 05.
58. Seipel, K., Georgiev, O., & Schaffner, W. (1992). Different activation domains stimulate transcription from remote ('enhancer') and proximal ('promoter') positions. *Embo Journal*, *11*, 4961–4968.
59. Colbran, L. L., Chen, L., & Capra, J. A. (2019). Sequence characteristics distinguish transcribed enhancers from promoters and predict their breadth of activity. *Genetics*, *211*, 1205–1217.
60. Kheradpour, P., Ernst, J., Melnikov, A., Rogov, P., Wang, L., Zhang, X., Alston, J., Mikkelsen, T. S., & Kellis, M. (2013). Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Research*, *23*, 800–811.
61. Kwasniewski, J. C., Fiore, C., Chaudhari, H. G., & Cohen, B. A. (2014). High-throughput functional testing of ENCODE segmentation predictions. *Genome Research*, *24*, 1595–1602.
62. Sahu, B., Hartonen, T., Pihlajamaa, P., Wei, B., Dave, K., Zhu, F., Kaasinen, E., Lidschreiber, K., Lidschreiber, M., Daub, C. O., Cramer, P., Kivioja, T., & Taipale, J. (2022). Sequence determinants of human gene regulatory elements. *Nature Genetics*, *54*, 283–294.
63. Scruggs, B. S., Gilchrist, D. A., Nechaev, S., Muse, G. W., Burkholder, A., Fargo, D. C., & Adelman, K. (2015). Bidirectional transcription arises from two distinct hubs of transcription factor binding and active chromatin. *Molecular Cell*, *58*, 1101–1112.
64. Dao, L. T. M., & Spicuglia, S. (2018). Transcriptional regulation by promoters with enhancer function. *Transcription*, *9*, 307–314.
65. De Almeida, B. P., Reiter, F., Pagani, M., & Stark, A. (2022). DeepSTARR predicts enhancer activity from DNA sequence and enables the de novo design of synthetic enhancers. *Nature Genetics*, *54*, 613–624.
66. Zaugg, J. B., Sahlén, P., Andersson, R., Alberich-Jorda, M., De Laat, W., Deplancke, B., Ferrer, J., Mandrup, S., Natoli, G., Plewczynski, D., Rada-Iglesias, A., & Spicuglia, S. (2022). Current challenges in understanding the role of enhancers in disease. *Nature Structural & Molecular Biology*, *29*(12), 1148–1158.
67. Deplancke, B., Alpern, D., & Gardeux, V. (2016). The genetics of transcription factor DNA binding variation. *Cell*, *166*, 538–554.
68. MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., Pendlington, Z. M., Welter, D., Burdett, T., Hindorf, L., Flicek, P., Cunningham, F., & Parkinson, H. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS catalog). *Nucleic Acids Research*, *45*, D896–D901.
69. Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J., Shafer, A., Neri, F., Lee, K., Kutayavin, T., Stehling-Sun, S., Johnson, A. K., Canfield, T. K., Giste, E., Diegel, M., ... Stamatoyannopoulos, J. A. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science*, *337*, 1190–1195.
70. Brandes, N., Weissbrod, O., & Linnal, M. (2022). Open problems in human trait genetics. *Genome Biology*, *23*, 131.
71. Smemo, S., Tena, J. J., Kim, K.-H., Gamazon, E. R., Sakabe, N. J., Gómez-Marín, C., Aneas, I., Credidio, F. L., Sobreira, D. R., Wasserman, N. F., Lee, J. H., Puvion-Dran, V., Tam, D., Shen, M., Son, J. E., Vakili, N. A., Sung, H.-K., Naranjo, S., Acemel, R. D., ... Nóbrega, M. A. (2014). Obesity-associated variants within FTO form long-range functional connections with IIRX3. *Nature*, *507*, 371–375.
72. Xia, Q., Chesi, A., Manduchi, E., Johnston, B. T., Lu, S., Leonard, M. E., Parlin, U. W., Rappaport, E. F., Huang, P., Wells, A. D., Blobel, G. A., Johnson, M. E., & Grant, S. F. A. (2016). The type 2 diabetes presumed causal variant within TCF7L2 resides in an element that controls the expression of ACSL5. *Diabetologia*, *59*, 2360–2368.
73. Leung, D., Jung, I., Rajagopal, N., Schmitt, A., Selvaraj, S., Lee, A. Y., Yen, C.-A., Lin, S., Lin, Y., Qiu, Y., Xie, W., Yue, F., Hariharan, M., Ray, P., Kuan, S., Edsall, L., Yang, H., Chi, N. C., Zhang, M. Q., ... Ren, B. (2015). Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature*, *518*, 350–354.
74. Cano-Gamez, E., & Trynka, G. (2020). From GWAS to function: Using functional genomics to identify the mechanisms underlying complex diseases. *Frontiers in Genetics*, *11*, 424.
75. Gupta, R. M., Hadaya, J., Trehan, A., Zekavat, S. M., Roselli, C., Klarin, D., Emdin, C. A., Hilvering, C. R. E., Bianchi, V., Mueller, C., Khera, A. V., Ryan, R. J. H., Engreitz, J. M., Issner, R., Shores, N., Epstein, C. B., De Laat, W., Brown, J. D., Schnabel, R. B., ... Kathiresan, S. (2017). A genetic variant associated with five vascular diseases is a distal regulator of endothelin-1 gene expression. *Cell*, *170*, 522–533.e15.e15.
76. Sinnott-Armstrong, N., Sousa, I. S., Laber, S., Rendina-Ruedy, E., Nitter Dankel, S. E., Ferreira, T., Mellgren, G., Karasik, D., Rivas, M., Pritchard, J., Guntur, A. R., Cox, R. D., Lindgren, C. M., Hauner, H., Sallari, R., Rosen, C. J., Hsu, Y.-H., Lander, E. S., Kiel, D. P., & Clausnitzer, M. (2021). A regulatory variant at 3q21.1 confers an increased pleiotropic risk for hyperglycemia and altered bone mineral density. *Cell Metabolism*, *33*, 615–628.e13.e13.
77. Sobreira, D. R., & Nóbrega, M. A. (2021). Regulatory landscapes of *Nppa* and *Nppb*. *Circulation Research*, *128*, 130–132.
78. Joslin, A. C., Sobreira, D. R., Hansen, G. T., Sakabe, N. J., Aneas, I., Montefiori, L. E., Farris, K. M., Gu, J., Lehman, D. M., Ober, C., He, X., & Nóbrega, M. A. (2021). A functional genomics pipeline identifies pleiotropy and cross-tissue effects within obesity-associated GWAS loci. *Nature Communications*, *12*, 5253.
79. Ribeiro, D. M., Rubinacci, S., Ramisch, A., Hofmeister, R. J., Dermitzakis, E. T., & Delaneau, O. (2021). The molecular basis, genetic control and



- pleiotropic effects of local gene co-expression. *Nature Communications*, 12, 4842.
80. Singh, D., & Yi, S. V. (2021). Enhancer pleiotropy, gene expression, and the architecture of human enhancer–gene interactions. *Molecular Biology and Evolution*, 38, 3898–3909.
  81. Flynn, E. D., & Lappalainen, T. (2022). Functional characterization of genetic variant effects on expression. *Annual Review of Biomedical Data Science*, 5, 119–139.
  82. GTEx Consortium. (2017). Genetic effects on gene expression across human tissues. *Nature*, 550, 204–213.
  83. Wang, D., Liu, S., Warrell, J., Won, H., Shi, X., Navarro, F. C. P., Clarke, D., Gu, M., Emani, P., Yang, Y. T., Xu, M., Gandal, M. J., Lou, S., Zhang, J., Park, J. J., Yan, C., Rhie, S. K., Manakongtreecheep, K., Zhou, H., & PsychENCODE Consortium. (2018). Comprehensive functional genomic resource and integrative model for the human brain. *Science*, 362, eaat8464.
  84. Saint Just Ribeiro, M., Tripathi, P., Namjou, B., Harley, J. B., & Chepelev, I. (2022). Haplotype-specific chromatin looping reveals genetic interactions of regulatory regions modulating gene expression in 8p23.1. *Frontiers in Genetics*, 13, 1008582.
  85. Li, W., Wong, W. H., & Jiang, R. (2019). DeepTACT: Predicting 3D chromatin contacts via bootstrapping deep learning. *Nucleic Acids Research*, 47, e60–e60.
  86. Man, J., Barnett, P., & Christoffels, V. M. (2018). Structure and function of the Nppa–Nppb cluster locus during heart development and disease. *Cellular and Molecular Life Sciences*, 75, 1435–1444.
  87. Kulzer, J. R., Stitzel, M. L., Morken, M. A., Huyghe, J. R., Fuchsberger, C., Kuusisto, J., Laakso, M., Boehnke, M., Collins, F. S., & Mohlke, K. L. (2014). A common functional regulatory variant at a type 2 diabetes locus upregulates ARAP1 expression in the pancreatic beta cell. *The American Journal of Human Genetics*, 94, 186–197.
  88. Nisar, S., Torres, M., Thiam, A., Pouvelle, B., Rosier, F., Gallardo, F., Ka, O., Mbengue, B., Diallo, R. N., Brosseau, L., Spicuglia, S., Dieye, A., Marquet, S., & Rihet, P. (2022). Identification of ATP2B4 regulatory element containing functional genetic variants associated with severe malaria. *International Journal of Molecular Sciences*, 23, 4849.
  89. Stikker, B. S., Stik, G., Van Ouwerkerk, A. F., Trap, L., Spicuglia, S., Hendriks, R. W., & Stadhouders, R. (2022). Severe COVID-19-associated variants linked to chemokine receptor gene control in monocytes and macrophages. *Genome Biology*, 23, 96.
  90. Fredriksson, N. J., Ny, L., Nilsson, J. A., & Larsson, E. (2014). Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nature Genetics*, 46, 1258–1263.
  91. Gao, P., Xia, J.-H., Sipeky, C., Dong, X.-M., Zhang, Q., Yang, Y., Zhang, P., Cruz, S. P., Zhang, K., Zhu, J., Lee, H.-M., Suleman, S., Giannareas, N., Liu, S., Tammela, T. L. J., Auvinen, A., Wang, X., Huang, Q., Wang, L., ... Wei, G.-H. (2018). Biology and clinical implications of the 19q13 aggressive prostate cancer susceptibility locus. *Cell*, 174, 576–589.e18.e18.
  92. Hua, J. T., Ahmed, M., Guo, H., Zhang, Y., Chen, S., Soares, F., Lu, J., Zhou, S., Wang, M., Li, H., Larson, N. B., McDonnell, S. K., Patel, P. S., Liang, Y., Yao, C. Q., Van Der Kwast, T., Lupien, M., Feng, F. Y., Zoubeidi, A., ... He, H. H. (2018). Risk SNP-mediated promoter-enhancer switching drives prostate cancer through lncRNA PCAT19. *Cell*, 174, 564–575.e18.e18.
  93. Hnisz, D., Weintraub, A. S., Day, D. S., Valton, A.-L., Bak, R. O., Li, C. H., Goldmann, J., Lajoie, B. R., Fan, Z. P., Sigova, A. A., Reddy, J., Borges-Rivera, D., Lee, T. I., Jaenisch, R., Porteus, M. H., Dekker, J., & Young, R. A. (2016). Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science*, 351, 1454–1458.
  94. Sollis, E., Mosaku, A., Abid, A., Buniello, A., Cerezo, M., Gil, L., Groza, T., Güneş, O., Hall, P., Hayhurst, J., Ibrahim, A., Ji, Y., John, S., Lewis, E., Macarthur, J. A. L., McMahon, A., Osumi-Sutherland, D., Panoutsopoulou, K., Pendlington, Z., ... Harris, L. W. (2023). The NHGRI-EBI GWAS Catalog: Knowledgebase and deposition resource. *Nucleic Acids Research*, 51, D977–D985.
  95. Corrales, M., Rosado, A., Cortini, R., Van Arensbergen, J., Van Steensel, B., & Filion, G. J. (2017). Clustering of *Drosophila* housekeeping promoters facilitates their expression. *Genome Research*, 27, 1153–1161.
  96. Karollus, A., Mauermeier, T., & Gagneur, J. (2023). Current sequence-based models capture gene expression determinants in promoters but mostly ignore distal enhancers. *Genome Biology*, 24, 56.
  97. Kim, S., & Wysocka, J. (2023). Deciphering the multi-scale, quantitative cis-regulatory code. *Molecular Cell*, 83, 373–392.
  98. Movva, R., Greenside, P., Marinov, G. K., Nair, S., Shrikumar, A., & Kundaje, A. (2019). Deciphering regulatory DNA sequences and non-coding genetic variants using neural network models of massively parallel reporter assays. *PLoS ONE*, 14, e0218073.
  99. Wang, Y., He, H., Liyanarachchi, S., Genutis, L. K., Li, W., Yu, L., Phay, J. E., Shen, R., Brock, P., & De La Chapelle, A. (2018). The role of SMAD3 in the genetic predisposition to papillary thyroid carcinoma. *Genetics in Medicine*, 20, 927–935.

**How to cite this article:** Malfait, J., Wan, J., & Spicuglia, S. (2023). Epromoters are new players in the regulatory landscape with potential pleiotropic roles. *BioEssays*, 45, e2300012. <https://doi.org/10.1002/bies.202300012>



Imprint logo

Contents lists available at ScienceDirect

Cells &amp; Development

journal homepage: [www.journals.elsevier.com/cells-and-development](http://www.journals.elsevier.com/cells-and-development)

C&amp;D

Full Length Article

## *Xenopus tropicalis* osteoblast-specific open chromatin regions reveal promoters and enhancers involved in human skeletal phenotypes and shed light on early vertebrate evolution

Héctor Castillo<sup>a,\*</sup>, Patricia Hanna<sup>a,1</sup>, Laurent M. Sachs<sup>b</sup>, Nicolas Buisine<sup>b</sup>, Francisco Godoy<sup>a</sup>, Clément Gilbert<sup>c</sup>, Felipe Aguilera<sup>a</sup>, David Muñoz<sup>a</sup>, Catherine Boisvert<sup>d</sup>, Mélanie Debiais-Thibaud<sup>e</sup>, Jing Wan<sup>f,g</sup>, Salvatore Spicuglia<sup>f,g</sup>, Sylvain Marcellini<sup>a,\*</sup>

<sup>a</sup> Group for the Study of Developmental Processes (GDeP), School of Biological Sciences, University of Concepción, Chile

<sup>b</sup> UMR7221, Physiologie Moléculaire et Adaptation, CNRS, MNHN, Paris Cedex 05, France

<sup>c</sup> Université Paris-Saclay, CNRS, IRD, UMR Évolution, Génomes, Comportement et Écologie, 12 route 128, 91190 Gif-sur-Yvette, France

<sup>d</sup> School of Molecular and Life Sciences, Curtin University, Perth, WA, Australia

<sup>e</sup> Institut des Sciences de l'Évolution de Montpellier, ISEM, Univ Montpellier, CNRS, IRD, Montpellier, France

<sup>f</sup> Aix-Marseille University, INSERM, TAGC, UMR 1090, Marseille, France

<sup>g</sup> Equipe Labelisée LIGUE contre le Cancer, Marseille, France

## ARTICLE INFO


## Keywords:

*Xenopus tropicalis*  
Osteoblasts  
ATAC-Seq  
Regulatory regions  
Bone diseases  
Skeletal evolution

## ABSTRACT

While understanding the genetic underpinnings of osteogenesis has far-reaching implications for skeletal diseases and evolution, a comprehensive characterization of the osteoblastic regulatory landscape in non-mammalian vertebrates is still lacking. Here, we compared the ATAC-Seq profile of *Xenopus tropicalis* (*Xt*) osteoblasts to a variety of non mineralizing control tissues, and identified osteoblast-specific nucleosome free regions (NFRs) at 527 promoters and 6747 distal regions. Sequence analyses, Gene Ontology, RNA-Seq and ChIP-Seq against four key histone marks confirmed that the distal regions correspond to *bona fide* osteogenic transcriptional enhancers exhibiting a shared regulatory logic with mammals. We report 425 regulatory regions conserved with human and globally associated to skeletal genes. Of these, 35 regions have been shown to impact human skeletal phenotypes by GWAS, including one *trps1* enhancer and the *runx2* promoter, two genes which are respectively involved in trichorhinophalangeal syndrome type I and cleidocranial dysplasia. Intriguingly, 60 osteoblastic NFRs also align to the genome of the elephant shark, a species lacking osteoblasts and bone tissue. To tackle this paradox, we chose to focus on *dlx5* because its conserved promoter, known to integrate regulatory inputs during mammalian osteogenesis, harbours an osteoblast-specific NFR in both frog and human. Hence, we show that *dlx5* is expressed in *Xt* and elephant shark odontoblasts, supporting a common cellular and genetic origin of bone and dentine. Taken together, our work (i) unravels the *Xt* osteogenic regulatory landscape, (ii) illustrates how cross-species comparisons harvest data relevant to human biology and (iii) reveals that a set of genes including *bnc2*, *dlx5*, *ebf3*, *mir199a*, *nfia*, *runx2* and *zfhx4* drove the development of a primitive form of mineralized skeletal tissue deep in the vertebrate lineage.

## Genome-wide enhancer identification by massively parallel reporter assay in *Arabidopsis*

Yongjun Tan<sup>1,2,†</sup>, Xiaohao Yan<sup>2,†</sup>, Jialei Sun<sup>2,†</sup>, Jing Wan<sup>3</sup>, Xinxin Li<sup>4,5</sup>, Yingzhang Huang<sup>2</sup>, Li Li<sup>3</sup>, Longjian Niu<sup>4,5,\*</sup> and Chunhui Hou<sup>1,\*</sup> 

<sup>1</sup>State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650201, China,

<sup>2</sup>Department of Biology, Southern University of Science and Technology, Shenzhen 518055, China,

<sup>3</sup>Department of Bioinformatics, Huazhong Agricultural University, Wuhan 430070, China,

<sup>4</sup>School of Public Health and Emergency Management, Southern University of Science and Technology, Shenzhen 518055, China, and

<sup>5</sup>Shenzhen Key Laboratory of Cardiovascular Health and Precision Medicine, Southern University of Science and Technology, Shenzhen 518055, China

Received 29 November 2022; revised 29 May 2023; accepted 27 June 2023.

\*For correspondence (e-mail [niulongjian@126.com](mailto:niulongjian@126.com); [houchunhui@mail.kiz.ac.cn](mailto:houchunhui@mail.kiz.ac.cn)).

<sup>†</sup>These authors contributed equally to this work.

### SUMMARY

Enhancers are critical cis-regulatory elements controlling gene expression during cell development and differentiation. However, genome-wide enhancer characterization has been challenging due to the lack of a well-defined relationship between enhancers and genes. Function-based methods are the gold standard for determining the biological function of cis-regulatory elements; however, these methods have not been widely applied to plants. Here, we applied a massively parallel reporter assay on *Arabidopsis* to measure enhancer activities across the genome. We identified 4327 enhancers with various combinations of epigenetic modifications distinctively different from animal enhancers. Furthermore, we showed that enhancers differ from promoters in their preference for transcription factors. Although some enhancers are not conserved and overlap with transposable elements forming clusters, enhancers are generally conserved across thousand *Arabidopsis* accessions, suggesting they are selected under evolution pressure and could play critical roles in the regulation of important genes. Moreover, comparison analysis reveals that enhancers identified by different strategies do not overlap, suggesting these methods are complementary in nature. In sum, we systematically investigated the features of enhancers identified by functional assay in *A. thaliana*, which lays the foundation for further investigation into enhancers' functional mechanisms in plants.

**Keywords:** plant enhancer, STARR-seq, epigenetic modification, transcription factor binding sites, evolutionary conservation, transposable elements.



OPEN

# Three-dimensional folding dynamics of the *Xenopus tropicalis* genome

Longjian Niu<sup>1,6</sup>, Wei Shen<sup>2,3,6</sup>, Zhaoying Shi<sup>1,6</sup>, Yongjun Tan<sup>1</sup>, Na He<sup>1</sup>, Jing Wan<sup>2,3</sup>, Jialei Sun<sup>1</sup>, Yuedong Zhang<sup>1</sup>, Yingzhang Huang<sup>1</sup>, Wenjing Wang<sup>1</sup>, Chao Fang<sup>1,4</sup>, Jiashuo Li<sup>1</sup>, Piaopiao Zheng<sup>1</sup>, Edwin Cheung<sup>5,6</sup>, Yonglong Chen<sup>1</sup>, Li Li<sup>2,3</sup> and Chunhui Hou<sup>1</sup>

**Animal interphase chromosomes are organized into topologically associating domains (TADs). How TADs are formed is not fully understood. Here, we combined high-throughput chromosome conformation capture and gene silencing to obtain insights into TAD dynamics in *Xenopus tropicalis* embryos. First, TAD establishment in *X. tropicalis* is similar to that in mice and flies and does not depend on zygotic genome transcriptional activation. This process is followed by further refinements in active and repressive chromatin compartments and the appearance of loops and stripes. Second, within TADs, higher self-interaction frequencies at one end of the boundary are associated with higher DNA occupancy of the architectural proteins CTCF and Rad21. Third, the chromatin remodeling factor ISWI is required for de novo TAD formation. Finally, TAD structures are variable in different tissues. Our work shows that *X. tropicalis* is a powerful model for chromosome architecture analysis and suggests that chromatin remodeling plays an essential role in de novo TAD establishment.**

Interphase chromosomes are partitioned into TADs<sup>1–4</sup>, segregating into the compartments of active or repressive chromatin<sup>5–7</sup>. The structure of TADs is relatively stable and resilient to environmental perturbations<sup>8,9</sup> and their architecture is evolutionarily conserved in eukaryotes<sup>4,10,11</sup>. Disruption of TAD borders can lead to developmental disorders and even tumorigenesis; this underlines the importance of three-dimensional (3D) genome organization in gene regulation<sup>12–15</sup>.

The establishment of chromatin architecture during embryogenesis provides an initial spatial frame that may guide proper genome organization, chromatin interaction and gene regulation<sup>16</sup>. In fruit flies, mice and humans, TADs form at the zygotic genome activation (ZGA) stage and continually consolidate through early embryo development<sup>16–20</sup>. However, in zebrafish, TADs are already reformed before ZGA, subsequently lost and then reestablished in later developmental stages<sup>21</sup>. The difference in TAD formation between species thus raises the question of whether this process is evolutionarily conserved.

DNA loop extrusion mediated by the cohesin complex was recently reported in several in vitro studies<sup>22,23</sup> and proposed as a functional mechanism underlying TAD establishment<sup>24–26</sup>. In cultured cells, deletion of the cohesin complex component double-strand-break repair protein rad21 homolog (Rad21) alone was enough to abolish the establishment of TADs<sup>27</sup>. Other proteins, including CCCTC-binding factor (CTCF), the cohesin antagonist Wings apart-like protein homolog (WAPL) and its partner PDS5, also participate in TAD regulation and loop structure formation<sup>28,29</sup>. CTCF loss disrupts TAD insulation but not higher-order genomic compartmentalization<sup>30</sup>. Likewise, TAD formation during mouse<sup>31</sup> and human embryogenesis<sup>32</sup> requires Rad21 and CTCF, respectively. These findings suggest that TAD formation in cultured

and embryonic nuclei is conserved and may require both factors through cohesin-mediated extrusion that stops at convergent CTCF binding sites<sup>11,33,34</sup>.

Interestingly, transcription appears to be dispensable for TAD formation at ZGA in fruit flies and mice<sup>16–20</sup> but not in humans<sup>32</sup>. Heinz et al.<sup>35</sup> showed that transcription disrupts TAD borders by displacing cohesin and CTCF during influenza A virus infection, while others found that transcription drives the formation of domain borders in *Caulobacter* cells<sup>36</sup>. These opposing findings suggest that the role of transcription in TAD formation is likely context-dependent or regulated by undefined factors.

How TADs are formed during embryogenesis is still not fully clear. During *X. tropicalis* embryogenesis, major ZGA occurs after 12 synchronous cell cycles<sup>37</sup> at the mid-blastula transition (MBT) (stage 8+) stage when S and gap phases appear and interphase lengthens<sup>38,39</sup>. More than 1,000 genes are activated before MBT<sup>40,41</sup>, while most of the zygotic genome is transcriptionally silent. To examine and assess the role of specific factors in the de novo establishment of chromatin architecture in the *Xenopus* zygote, morpholinos can be used to block the new translation of target proteins. In this study, we examined chromosome conformation change across multiple developmental stages in wild-type (WT) *X. tropicalis* embryos and embryos where RNA polymerase II (Pol II), CTCF, Rad21 or the chromatin remodeling factor ISWI translation was inhibited. Our work revealed that in *Xenopus*, TADs appear at ZGA and are followed by the sequential establishment of loop and stripe structures in later developmental stages. We found that TAD formation requires CTCF and Rad21. We also demonstrated that ISWI is required for both the establishment of TADs and embryo development. Interestingly, we showed that chromatin interaction directionality is almost always stronger on one side of the TAD






<sup>1</sup>Department of Biology, School of Life Sciences, Southern University of Science and Technology, Shenzhen, China. <sup>2</sup>Department of Bioinformatics, Huazhong Agricultural University, Wuhan, China. <sup>3</sup>Hubei Key Laboratory of Agricultural Bioinformatics, Huazhong Agricultural University, Wuhan, China. <sup>4</sup>Cancer Centre, Faculty of Health Sciences, University of Macau, Taipa, China. <sup>5</sup>Centre of Precision Medicine Research and Training, Faculty of Health Sciences, University of Macau, Taipa, China. <sup>6</sup>These authors contributed equally: Longjian Niu, Wei Shen, Zhaoying Shi. ✉e-mail: [echung@um.edu.mo](mailto:echung@um.edu.mo); [chenyl@sustech.edu.cn](mailto:chenyl@sustech.edu.cn); [li.li@hzau.edu.cn](mailto:li.li@hzau.edu.cn); [houch@sustech.edu.cn](mailto:houch@sustech.edu.cn)

## ARTICLE

<https://doi.org/10.1038/s42003-019-0519-y>

OPEN

## Amplification-free library preparation with SAFE Hi-C uses ligation products for deep sequencing to improve traditional Hi-C analysis

Longjian Niu <sup>1,2,5</sup>, Wei Shen <sup>3,4,5</sup>, Yingzhang Huang<sup>1,5</sup>, Na He<sup>1</sup>, Yuedong Zhang<sup>1</sup>, Jialei Sun<sup>1</sup>, Jing Wan <sup>3,4</sup>, Daxin Jiang <sup>1</sup>, Manyun Yang<sup>1</sup>, Yu Chung Tse <sup>1</sup>, Li Li<sup>3,4</sup> & Chunhui Hou<sup>1</sup>

PCR amplification of Hi-C libraries introduces unusable duplicates and results in a biased representation of chromatin interactions. We present a simplified, fast, and economically efficient Hi-C library preparation procedure, SAFE Hi-C, which generates sufficient non-amplified ligation products for deep sequencing from 30 million *Drosophila* cells. Comprehensive analysis of the resulting data shows that amplification-free Hi-C preserves higher complexity of chromatin interaction and lowers sequencing depth for the same number of unique paired reads. For human cells which have a large genome, SAFE Hi-C recovers enough ligated fragments for direct high-throughput sequencing without amplification from as few as 250,000 cells. Comparison with published in situ Hi-C data from millions of human cells demonstrates that amplification introduces distance-dependent amplification bias, which results in an increased background noise level against genomic distance. With amplification bias avoided, SAFE Hi-C may produce a chromatin interaction network more faithfully reflecting the real three-dimensional genomic architecture.

<sup>1</sup>Department of Biology, Southern University of Science and Technology, 518055 Shenzhen, China. <sup>2</sup>Department of Biology, Nankai University, 300071 Tianjin, China. <sup>3</sup>Department of Bioinformatics, Huazhong Agricultural University, 430070 Wuhan, China. <sup>4</sup>Hubei Key Laboratory of Agricultural Bioinformatics, Huazhong Agricultural University, 430070 Wuhan, China. <sup>5</sup>These authors contributed equally: Longjian Niu, Wei Shen, Yingzhang Huang. Correspondence and requests for materials should be addressed to L.L. (email: [li.li@mail.hzau.edu.cn](mailto:li.li@mail.hzau.edu.cn)) or to C.H. (email: [houch@sustech.edu.cn](mailto:houch@sustech.edu.cn))