

FUTURE LEADER REVIEW

The importance of considering regulatory domains in genome-wide analyses – the nearest gene is often wrong!

Ellora Hui Zhen Chua¹, Samen Yasar¹ and Nathan Harmston^{1,2,*}

ABSTRACT

The expression of a large number of genes is regulated by regulatory elements that are located far away from their promoters. Identifying which gene is the target of a specific regulatory element or is affected by a non-coding mutation is often accomplished by assigning these regions to the nearest gene in the genome. However, this heuristic ignores key features of genome organisation and gene regulation; in that the genome is partitioned into regulatory domains, which at some loci directly coincide with the span of topologically associated domains (TADs), and that genes are regulated by enhancers located throughout these regions, even across intervening genes. In this review, we examine the results from genome-wide studies using chromosome conformation capture technologies and from those dissecting individual gene regulatory domains, to highlight that the phenomenon of enhancer skipping is pervasive and affects multiple types of genes. We discuss how simply assigning a genomic region of interest to its nearest gene is problematic and often leads to incorrect predictions and highlight that where possible information on both the conservation and topological organisation of the genome should be used to generate better hypotheses.

The article has an associated Future Leader to Watch interview.

KEY WORDS: Chromatin structure, Enhancers, Gene regulation

Introduction

The precise regulation of gene expression is necessary for both development and homeostasis, with dysregulation leading to developmental disorders and disease (Lee and Young, 2013). In multicellular organisms, a number of genes are not primarily regulated by their core and proximal promoter, but are under the control of regulatory elements located both proximally and distally, often referred to as enhancers (Miguel-Escalada et al., 2015). Chromatin looping between enhancers and promoters places these elements into close physical proximity with their cognate target genes, with this spatial colocalisation being necessary for their role in regulating gene expression (Tolhuis et al., 2002). While genes and their regulatory elements are organised on a linear chromosome, within the nucleus they are part of a complex, hierarchical and non-random three-dimensional structure (Lieberman-Aiden et al., 2009). The development of experimental techniques to investigate this topological organisation has provided insights into how the spatial

conformation of chromatin across multiple levels directly affects the regulation of gene expression.

Only ~2% of the human genome is involved in coding for proteins, while at least 8.2% is under some level of selective pressure (Rands et al., 2014), indicating the importance of non-coding regions. Depending on the transcription factor (TF) or histone modification investigated, a large proportion of ChIP-seq peaks are located outside of coding regions (i.e. are intergenic or intronic), and 93% of single nucleotide polymorphisms (SNPs) identified by genome-wide association studies (GWAS) have been found to be located in non-coding regions (Maurano et al., 2012). Maurano et al. reported that 76.6% of GWAS SNPs lie within a regulatory region defined by DNase I hypersensitivity or were in linkage disequilibrium (LD) with SNPs overlapping a DNase hypersensitive site (DHS) (Maurano et al., 2012).

Enhancers drive expression in a cell-type specific and spatiotemporal manner, potentially regulating only one specific gene in one specific context (Long et al., 2016). These elements are bound by TFs which provide regulatory logic (Jindal and Farley, 2021), and are associated with distinct patterns of histone modifications (Creyghton et al., 2010). Unlike SNPs in protein coding genes, which may affect splicing or protein structure/function, SNPs in regulatory elements can result in changes in expression of the gene they are regulating. SNPs frequently alter allelic chromatin state and disrupt TF binding sites (TFBSs), directly implicating non-coding variation in regulatory elements as drivers of both phenotypic diversity and disease (Kasowski et al., 2010; Spielmann and Mundlos, 2016). The regulatory domain, or regulatory landscape (Acemel et al., 2017), of a gene is understood to be the region of the genome which contains all of the regulatory information and elements which allows the gene to be expressed correctly (Bolt and Duboule, 2020). With enhancers responsible for regulating the expression of one or several genes, in different contexts, being located throughout this region.

One commonly used method for assigning a non-coding locus of interest (i.e. enhancer or SNP) to its target/relevant gene is to assign it to the nearest gene on the physical chromosome. Here, we discuss the existing experimental evidence, at individual loci and genome-wide, and using insights from comparative genomics, that nearest gene assignment is often misleading, and highlight the importance of considering the topological architecture of the genome when annotating and interpreting the results from genome-wide experiments.

The topological organisation of the genome determines the range of regulatory interactions

Within the nucleus, chromosomes are located in spatially distinct chromosome territories (Fig. 1A) (Bolzer et al., 2005; Cremer and Cremer, 2010). Chromatin is organised into active (A) and inactive (B) compartments (Fig. 1B), with the genome further partitioned into a set of preferentially interacting regions, known as

¹Science Division, Yale-NUS College, Singapore 138527, Singapore. ²Program in Cancer and Stem Cell Biology, Duke-NUS Medical School, Singapore 169857, Singapore.

*Author for correspondence (nathan.harmston@yale-nus.edu.sg)

 E.H.Z.C., 0000-0002-5520-2254; N.H., 0000-0002-8589-2938

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution and reproduction in any medium provided that the original work is properly attributed.

Table 1. (Vastly incomplete) summary of studies that have identified different SNP-target gene relationships from those proposed by nearest gene assignment.

SNPs	Nearest gene	Target gene	Context	References
rs1421085	<i>FTO</i>	<i>IRX3/IRX5</i>	Pancreas, Brain, Adipocytes	(Ragvin et al., 2010; Smemo et al., 2014; Claussnitzer et al., 2015)
rs1191551	<i>PRKD1</i>	<i>FOXG1</i>	Cortex	(Won et al., 2016)
rs7754840	<i>CDKAL1</i>	<i>SOX4</i>	Pancreas	(Ragvin et al., 2010; Miguel-Escalada et al., 2019)
rs9349379	<i>PHACTR1</i>	<i>EDN1</i>	Aorta	(Gupta et al., 2017)
rs1736020, rs1297265, rs2823286	<i>USP25</i>	<i>NRIP1</i>	CD34+	(Mifsud et al., 2015)
rs6983267	<i>POU5F1B</i>	<i>MYC, CCAT1</i>	S174T, LoVo, Colo205	(Jäger et al., 2015)
rs3806624	<i>EOMES</i>	<i>AZI2</i>	GM12878, Jurkat	(Martin et al., 2015)
rs9603616, rs7993214	<i>COG6</i>	<i>FOXO1</i>	GM12878, Jurkat	(Martin et al., 2015)
rs11203032	<i>CH25H, LIPA</i>	<i>ACTA2</i>	Cardiomyocytes	(Montefiori et al., 2018)
rs12740374	<i>CELSR2, PSRC1</i>	<i>SORT1</i>	Cardiomyocytes	(Montefiori et al., 2018)

topologically associating domains (TADs) (Shopland et al., 2006; Lieberman-Aiden et al., 2009; Dixon et al., 2012, 2015) (Fig. 1C). The partitioning of the genome into compartments occurs because of preferential interactions between regions of chromatin which have a similar state, with compartment B being associated with inactive, gene-poor regions of the genome, which are preferentially found in close proximity to the nuclear lamina, whereas compartment A is associated with active, gene-rich regions of the genome that are often located towards the centre of the nucleus and are often associated with transcription factories. During differentiation, chromatin can switch compartment type, from A

to B or vice versa, and this is associated with changes in expression of the genes in this region (Dixon et al., 2015). Genes and regulatory elements located within a TAD preferentially interact with each other but show a depletion for interactions with chromatin located in adjacent TADs, suggesting that TAD boundary regions act as insulators and that they function to constrain the range and activity of regulatory elements (Dixon et al., 2012).

TADs have been found to represent coherent functional blocks in terms of their association with replication domains, patterns of CTCF conservation and lamina associated domains (Paulsen et al., 2017; Pope et al., 2014; Vietri Rudan et al., 2015), and are largely

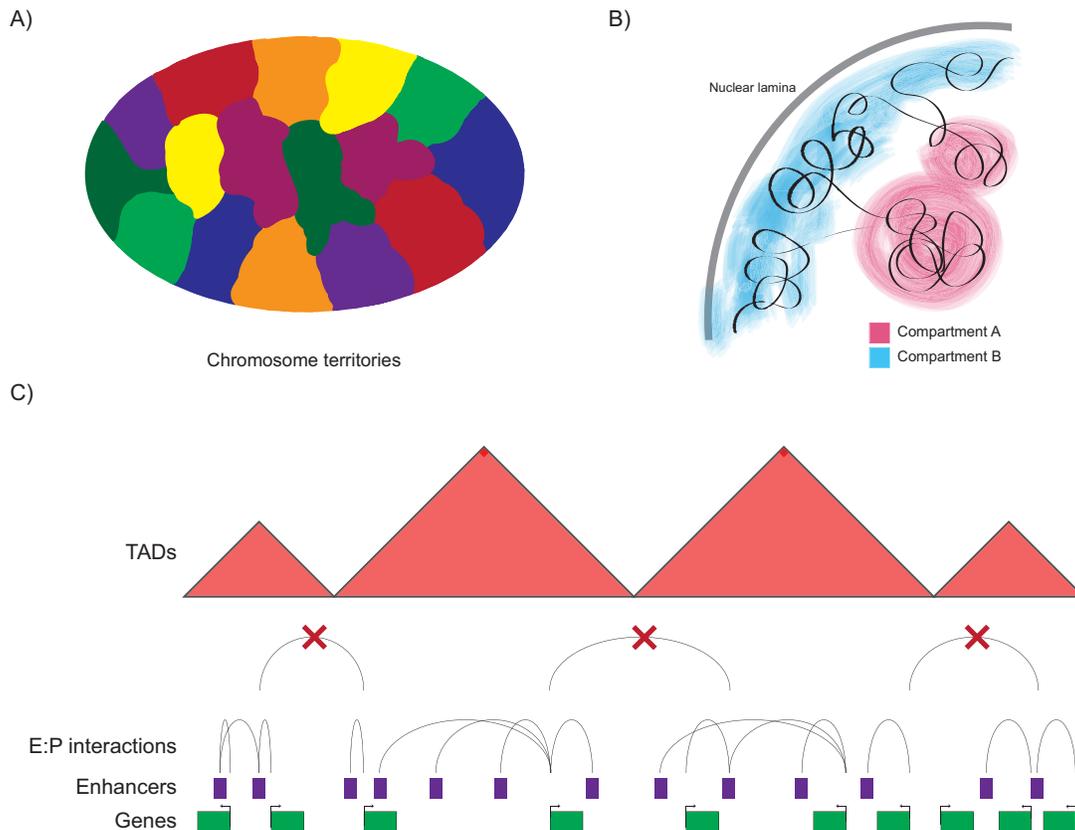


Fig. 1. Chromatin within a nucleus is hierarchically organised. (A) Chromosomes occupy distinct regions or territories within the nucleus known as chromosome territories. (B) Chromosomes are partitioned into compartments, with compartment A being associated with the interior of the nucleus and active chromatin, compartment B being associated with inactive chromatin located near the nuclear lamina. (C) The genome is further partitioned into TADs, with interactions between enhancers and their target genes preferentially occurring within TADs. At several TADs, enhancers regulating the expression of a gene can be located throughout its TAD – indicating that TADs can directly correspond to gene regulatory domains.

invariant across cell types (Siersbæk et al., 2017). Several studies support the direct concordance between TADs and gene regulatory domains. Genome-wide studies of enhancer–promoter interactions have found that the clear majority of interactions occur within a TAD (Montefiori et al., 2018; Jung et al., 2019). Random insertion of enhancer reporter constructs into the mouse genome found that the patterns of reporter activation during development reflected the organisation of a subset of TADs and the expression of their constituent genes (Symmons et al., 2014). At several loci, perturbation of TAD boundaries, via deletion of CTCF boundary sites and other techniques, has been demonstrated to lead to ectopic enhancer–promoter interactions and dysregulation of gene expression (Lupiáñez et al., 2015; Franke et al., 2016; Narendra et al., 2016; de Bruijn et al., 2020). However, other studies have found that perturbing TAD structure and boundaries does not have a large effect on gene expression (Despang et al., 2019; Ghavi-Helm et al., 2019). Mutations affecting TAD boundaries have been associated with ectopic enhancer–promoter interactions and altered gene expression in a number of neurological and developmental disorders (Lupiáñez et al., 2015; Franke et al., 2016). In tumors, structural variation affecting TAD organisation has been identified in multiple types of cancer (Taberlay et al., 2016; Akdemir et al., 2020), with experimental evidence that the deletion or rearrangement of boundary regions can directly lead to the dysregulation of important oncogenes by allowing ectopic enhancer–promoter interactions (Vicente-García et al., 2017; Weischenfeldt et al., 2017).

Therefore, at multiple loci, TADs appear to function as structural units of the genome whose purpose is to increase the probability that regulatory elements meet their target promoters within a specific domain, whilst decreasing the probability of interacting with elements and genes outside of the domain (Flyamer et al., 2017), helping to ensure that genes are turned on and off by the correct enhancers. However, it is likely that TADs may have multiple other functions (Austena et al., 2015; Hu et al., 2015), and that the correspondence between TADs and regulatory domains is likely only seen at a distinct subset of loci.

Patterns of microsynteny in metazoan genomes as a consequence of gene-regulatory constraints

During evolution, genomes can be rearranged leading to differences in the order and location of genes between species. However, the observed patterns of gene order (microsynteny) are not random (Oliver and Misteli, 2005), with hundreds of genes being found to be physically linked together over large evolutionary distances (Irimia et al., 2012). This has been proposed to be the result of either the co-expression/co-regulation of these genes, or that one member of the pair (the *bystander gene*) contains regulatory elements within its introns or exons that are necessary for the proper regulation of the other (the *target gene*).

This pattern of microsyntenic conservation, reflecting the need to keep regulatory elements in *cis* with their target genes was found to overlap with loci that have a high density of conserved non-coding elements (CNEs) (Bejerano, 2004). CNEs have a high percentage identity over a large number of base pairs between evolutionarily distinct groups of species, and often display regulatory activity in reporter assays (Harmston et al., 2013). Several of these CNEs were found to be located within the introns of housekeeping genes but were involved in the regulation of a different gene(s). The combination of these features led to the proposal of the genomic regulatory block (GRB) model (Becker and Lenhard, 2007; Engström et al., 2007; Kikuta et al., 2007b), where genes are physically linked together because of the need to ensure proper

developmental regulation of a specific target gene(s). This has led to the linkage of developmental transcription factors with housekeeping genes over large evolutionary distances, with these linkages spanning large genomic distances. Investigation of these regions using retroviral screens identified that insertions of reporter genes around important transcription factors resulted in the same expression pattern regardless of their position, highlighting that these regions correspond to the regulatory domains of these important transcriptional regulators (Kikuta et al., 2007a; Navratilova et al., 2009).

Studies have found that TADs are syntenic between humans and mice (Dixon et al., 2012) and show conservation in macaques and dogs (Vietri Rudan et al., 2015), with chromosomal rearrangements preferentially occurring at the boundaries of TADs (Berthelot et al., 2015; Liao et al., 2021). A subset of TADs directly corresponds to the location of GRBs, as inferred from the distribution of CNEs, across multiple species (Harmston et al., 2017). The correspondence between GRBs and TADs suggests that this subset of TADs primarily corresponds to the regulatory domain of a developmental TF under long-range regulation and not all of its constituent genes. Recent studies directly comparing TAD organisation in *Drosophila melanogaster* and *Drosophila triauraria* identified conservation of a subset of TADs with distinct features, further supporting this observation (Torosin et al., 2020). Conservation of TADs and maintenance of microsynteny both reflect selective pressure on genome organisation due to gene regulatory constraints, further supporting the notion that TADs reflect ‘regulatory units’ of the genome (Dixon et al., 2016) and gene-regulatory domains.

The identification that elements responsible for regulating the expression of one gene can be located within other genes directly highlights a problem with nearest gene assignment. These enhancers would be annotated as regulating their overlapping gene, which at many loci throughout the genome would be incorrect.

Studies of individual loci illuminate the fallacy of nearest gene assignment

Sonic Hedgehog (*SHH*) is a key developmental transcription factor with important roles in development of several tissues (Briscoe and Théron, 2013). In mammals, *SHH* lies within a large TAD spanning approximately 920 kb (Fig. 2) with a high density of CNEs. *SHH* is under complex enhancer-driven regulation by multiple elements located both proximally and distally (Anderson et al., 2014). Several enhancers have been identified within this TAD that drive *SHH* expression in a variety of tissues, including the brain, laryngotracheal tube, gut and limb bud (Lettice, 2003; Jeong et al., 2008; Sagai et al., 2009; Tsukiji et al., 2014) (Fig. 2). Several mutations within the regulatory domain surrounding *SHH* have been found to cause congenital abnormalities (Hill and Lettice, 2013). Genetic mapping of an interval associated with preaxial polydactyly (Hing et al., 1995) implicated *LMBR1* as a putative regulator of limb development (Clark et al., 2000). However, further studies found that proper expression of *SHH* in the developing limb bud depends on an enhancer (known as ZRS) located within the fifth intron of *LMBR1*, located 850 kb distally from *SHH* (Lettice et al., 2002; Lettice, 2003). This intronic enhancer corresponds to a CNE which has identifiable sequence conservation back to shark (Dahn et al., 2007); with this region of the genome being classified as a GRB. Polymorphisms within this element result in limb defects, including preaxial polydactyly and syndactyly in human (Lettice et al., 2008), with deletion of this element leading to limb truncation in the mouse (Sagai et al., 2005). A number of these mutations are

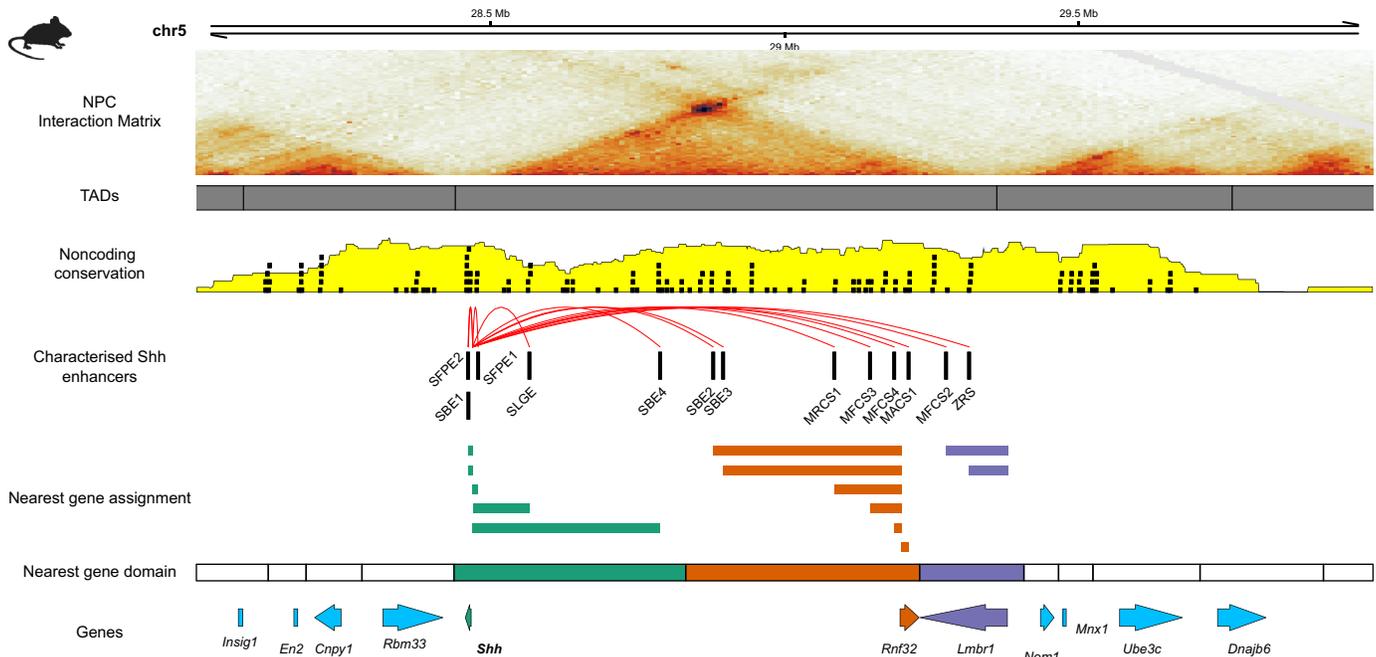


FIGURE 2

Fig. 2. The regulatory domain of the key developmental transcription factor *Shh*. Visualisation of the region chr5:28Mb-30Mb in mouse (mm10) which contains multiple genes; displaying the Hi-C interaction matrix in neural progenitor cells (NPC), TADs and locations of validated *Shh* enhancers. *Shh* is under the control of multiple enhancers located throughout its TAD, some of which are located closer to other genes than *Shh*, which would be erroneously assigned by using nearest gene assignment.

responsible for altering the activity and specificity of TFBSs located within ZRS, perturbing both the level and extent of *SHH* expression in the developing limb bud (Zhao et al., 2009; Lettice et al., 2012). Insufficient expression of *SHH* during brain development results in holoprosencephaly, which can be caused by mutations either affecting the coding region of *SHH* (Roessler et al., 1996) or its regulatory landscape (Belloni et al., 1996). Translocation events in the vicinity of *SHH* have been found to displace enhancers away from the *SHH* promoter (Jeong et al., 2006), and mutations within SBE2 have been shown to directly affect the expression of *SHH* by altering *Six3* binding at this enhancer (Jeong et al., 2008). Investigation of the *Shh* regulatory domain by transposon insertion mapping revealed that the range of action for *Shh* enhancers is coherent with the span of the regulatory domain predicted using Hi-C, and that the neighbours of *Shh* did not respond to long-range regulation (i.e. *Rnf32*) (Anderson et al., 2014). Considering the genomic locations of functionally characterised enhancers of *Shh* ($N=13$, Fig. 2), eight (61%) would be associated with the wrong gene by nearest gene assignment, including ZRS and SBE2.

Variants located within *FTO* (Gerken et al., 2007), fat-mass and obesity-associated gene, have been associated with several obesity-related phenotypes using GWAS (Dina et al., 2007; Frayling et al., 2007). The set of variants identified at this locus have been found to be highly replicable, have a high population frequency and show a strong effect size (Scuteri et al., 2007). Although these SNPs are located within introns 1 and 2 of *FTO*, eQTL studies found no evidence of a link between them and differences in the expression and splicing of *FTO* (Grunnet et al., 2009; Klötting et al., 2008). *FTO* is located within a region enriched for extreme non-coding conservation identified as a GRB (de la Calle-Mustienes et al., 2005), which accurately predicts the boundaries of the topological domain at this locus (Harmston et al., 2017; Hunt et al., 2015). The

predictions of the GRB model indicate that the majority of regulatory elements within this region are directly involved in the regulation of *IRX3* and *IRX5* (the target genes of this GRB), and that *FTO* and *RPGRIP1L* are simply bystander genes that are only regulated by proximal regulatory elements, if at all. This prediction has been confirmed by results from a number of studies. Enhancer screens have demonstrated that the CNE containing *rs1421085*, a SNP associated with Type 2 diabetes (T2D), acts as an enhancer driving reporter expression in the pancreatic area at 48 hpf in zebrafish (Ragvin et al., 2010). This suggested that the GWAS signal was not reflecting a variant associated with regulation of the constitutively expressed *FTO*, but an enhancer variant affecting the expression of *IRX3* and *IRX5*. Smemo et al. demonstrated that *FTO* is not under long range regulation in the mouse brain, but that *IRX3* interacts with intronic elements located within *FTO* (Smemo et al., 2014) (Fig. 3). In addition, studies have found that the *FTO* promoter does not respond to long-range regulation during zebrafish development (Rinkwitz et al., 2015). Recently, it has been shown that *rs1421085* is located within an *ARID5B* binding site, leading to impaired *ARID5B*-mediated repression of *IRX3* and *IRX5* during early adipocyte differentiation (Claussnitzer et al., 2015). This loss of repression leads to a loss of mitochondrial thermogenesis and a shift from fat browning to whitening programs. Therefore, although these variants appeared to implicate *FTO* as the causative gene based on nearest gene assignment, multiple experiments across several species and tissues have provided extensive evidence, and importantly mechanistic explanations, for these variants being involved in affecting the enhancer-driven regulation of *IRX3* and *IRX5*.

Additional studies have also shown the pervasiveness of genes involved in interactions with elements at ranges beyond the nearest gene at a multitude of other loci. *MEIS1* is located within a region containing a large number of CNEs. Several of these CNEs were

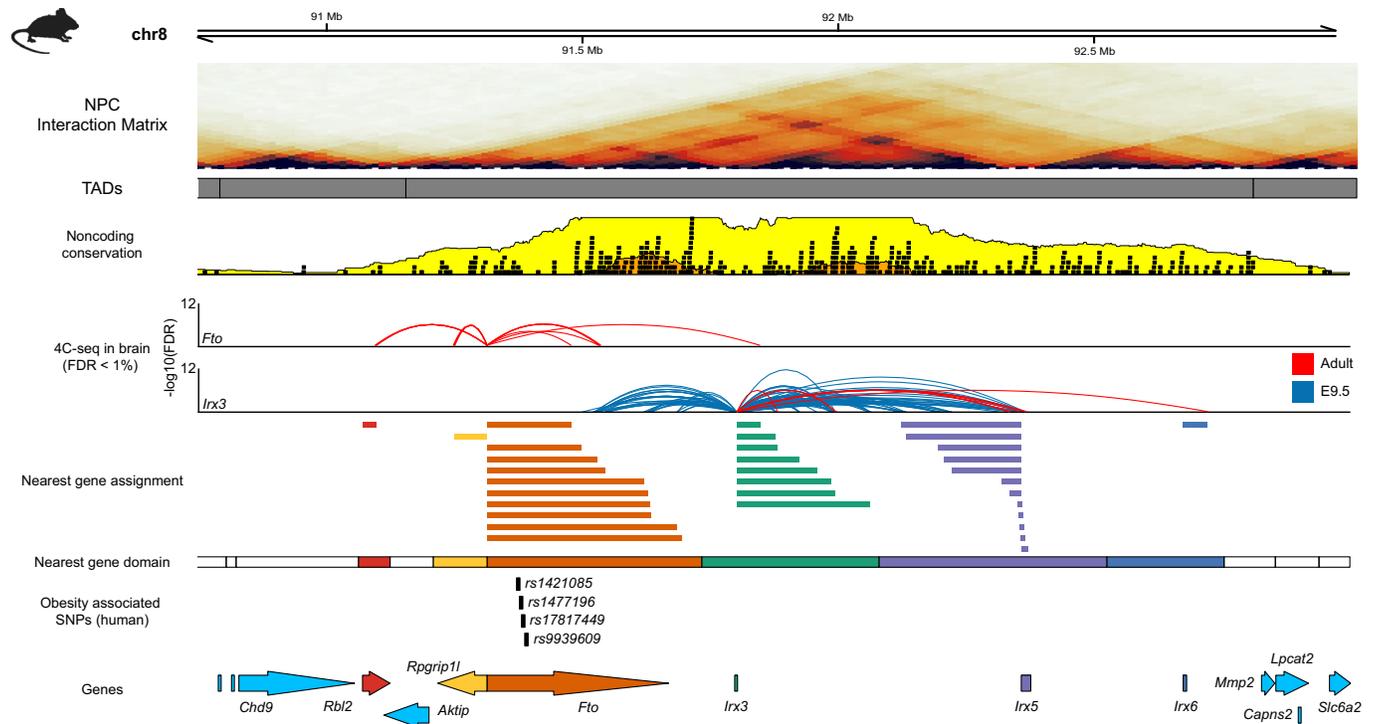


Fig. 3. The regulatory domain of the key developmental transcription factors *Irx3/5* and *6*. Visualisation of the region chr8:90.5Mb-93Mb in mouse (mm10); displaying Hi-C interaction matrix from in neural progenitor cells (NPC), location of interactions as identified 4C involving the promoters of *Fto* and *Irx3* (Smemo et al., 2014). *Irx3* is regulated by multiple regulatory elements located throughout its TAD, including elements located within the introns of *Fto*.

tested in enhancer assays, with 65% (22/34) testing positive and recapitulating the expression patterns of *MEIS1* (Royo et al., 2012). Using the heuristic of nearest gene assignment, eight (36%) of these elements would be assigned to different protein-coding genes. In zebrafish, regulatory elements located within the introns of the skin-specific *slc2a15a* and ubiquitously expressed *fbxw4* are involved in the regulation of *Fgf8a* (Komisarczuk et al., 2009). This study found that the majority of regulatory elements within this region drove expression of *Fgf8a*, with *slc2a15a* and *fbxw4* appearing to be non-responsive to long-range regulation. The ability of elements located within *fbw4* to drive *Fgf8* expression has been confirmed in mice (Marinić et al., 2013). In addition, the loss of exonic enhancers located within *DYNCL11* in humans has been reported to lead to split hand/foot malformations by affecting the expression of *DLX5/6*, a gene located 1Mb from these enhancers (Lango Allen et al., 2014). Therefore, multiple studies investigating regulatory elements at single loci have found that initially annotating the nearest gene as the target of an enhancer and/or genetic variant is often incorrect.

Insights from integrative genome-wide analyses of topological organisation

Although techniques including 3C and 4C permit the assessment of interactions between pre-determined viewpoints and one or multiple genomic regions, a number of techniques have enabled the identification of chromatin interactions genome-wide (McCord et al., 2020). Analyses based on Hi-C data have helped to putatively define the *cis*-regulatory domains and have identified mechanisms and factors involved in regulating chromatin looping (Phillips-Cremins et al., 2013; Van Bortle et al., 2014). However, significantly higher resolution data is required to precisely identify interactions between regulatory elements and promoters, which can be achieved by including an enrichment step (Fullwood et al., 2009; Schoenfelder et al., 2015; Mumbach et al., 2016), using a different

restriction enzyme (i.e. that cuts DNA more frequently) and/or by sequencing to a higher depth (Rao et al., 2014; Hua et al., 2021). Studies of chromatin interactions using these techniques, and integration of these maps with other types of data, have identified a number of important features relevant to understanding nuclear organisation and gene regulation, and have proposed new genes as being involved in disease processes.

Insertional mutagenesis screens use the patterns of recurrent retrovirus insertions, known as common insertion sites, to identify genes potentially involved in tumourigenesis (Uren et al., 2008). These insertions can lead to changes in gene expression with subsequent effects on tumour growth. However, insertions may be located anywhere within the regulatory domain of the gene whose expression it disrupts. By considering retroviral insertions in the context of spatial organisation, as defined using Hi-C, Babaei et al. were able to identify novel target genes that were not originally proposed using nearest gene assignment and re-assigned some insertions as putatively regulating different genes (Babaei et al., 2015). Several of these genes (i.e. *BRCA2*, *FANCS*, *APC*, *JAK1*, *NOTCH1*) were proposed to be more probable targets than the originally reported genes that were located near to insertions, with insertions involved in interactions being found to be more likely to deregulate genes involved in tumourigenesis. Therefore, integrating information on topological organisation with patterns of retroviral insertions led to improvements in sensitivity and specificity, and subsequently to new biological hypotheses.

Integrating Hi-C maps of the developing brain with the results from schizophrenia GWAS (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014) identified 402 genes that were involved in interactions with a region containing a significantly associated SNP, but which were neither adjacent to, nor in LD with this SNP (Won et al., 2016). This set of genes was enriched for processes including neuronal differentiation, and

significantly overlapped with a set of genes known to be downregulated in the prefrontal cortex of schizophrenia patients. *rs1191551*, an SNP associated with schizophrenia, was identified as being located within an enhancer active in the developing cortex, which was interacting with *FOXG1* (located 760 kb away) instead of the nearby *PRKDI* (located 45 kb away). Deletion of this region led to a decrease in expression of *FOXG1*, but had no effect on *PRKDI*, confirming that this region is involved in regulating *FOXG1*.

Capture Hi-C (cHi-C) combines Hi-C with target sequence enrichment (i.e. using baits targeting specific restriction fragments of interest) to improve resolution without requiring the sequencing of exponentially more reads (Schoenfelder et al., 2015). This technique allows the investigation of the interactions that a set of pre-defined genomic features (i.e. promoters or GWAS SNPs) are involved in at lower cost than standard Hi-C, essentially trading library complexity for statistical power. Mifsud et al. performed cHi-C in human blood cell lines and calculated that two thirds of promoter-centred interactions appeared to interact with the nearest gene, with the remainder interacting over longer distances, often across intervening genes (Mifsud et al., 2015). *USP25* is the closest gene (~280 kb away) to three SNPs implicated in inflammatory bowel disease, but all of these SNPs were observed to interact with *NR1P1*, a gene located ~380 kb away.

Targeting loci associated with susceptibility to autoimmune disease in B- and T-cells identified cell-type specific interactions between elements located near disease-associated SNPs and genes which were not adjacent in the genome (Martin et al., 2015). The 3' intronic region of *COG6* contains a number of SNPs associated with rheumatoid arthritis (RA) and juvenile idiopathic arthritis (JIA) which exhibited robust interactions with *FOXO1*, located over 1 Mb away. RA-associated variants located close to *EOMES* were identified as being involved in long-range interactions, spanning approximately 640 kb, with the promoter of *AZI2*. *DEXI* interacted with multiple loci associated with susceptibility to different autoimmune diseases over long distances; interacting with a region associated with type 1 diabetes and JIA located adjacent to *RMI2* (~530 kb away) and an RA-associated locus proximal to *ZC3H7A* (~1.2 Mb away). Three distinct SNPs located at 6q23 have been independently associated with various autoimmune diseases. A study investigating this locus using cHi-C revealed that these SNPs were involved in interactions with multiple genes over a wide range of distances and intervening genes (Martin et al., 2016). Although extensive work has suggested an important role for *TNFAIP3* in the modulation of autoimmune disease, the use of conformation capture techniques implicated additional putative genes and suggested that the primary causal gene may be *IL20RA* (Martin et al., 2016). Both of these studies demonstrate that non-coding variants associated with similar diseases may be involved in regulating the same gene, or genes in the same pathway, despite them being located distally from each other and proximal to other genes.

Promoter capture Hi-C (pcHi-C) maps of iPSCs and iPSC-derived cardiomyocytes found that 90% of interactions involving SNPs associated with cardiovascular disease and their target genes did not involve the nearest gene, with the majority (89%) of these interactions being between genes and regulatory elements located within the same TAD (Montefiori et al., 2018). Several SNPs associated with cardiovascular disease are proximal to both *CELSR2* and *PSRC1*, however the region containing these SNPs was found to interact with *SORT1* located 120 kb away, supporting a previous study that *SORT1* is the target gene of this region

(Musunuru et al., 2010). While *rs11203032* is proximal (<10 kb) to both *CH25H* and *LIPA*, this region was found to be interacting with *ACTA2*, located 220 kb away, indicating that it is actually the correct target gene.

By performing pcHi-C in human pancreatic islets, Escalada et al. were able to propose target genes for regulatory elements overlapping with type 2 diabetes (T2D) and fasting glycemia SNPs (Miguel-Escalada et al., 2019). This identified potential target genes for 53 regions, however only 24 of these regions (45%) interacted with the gene annotated in the original GWAS. 87% of these regions were found to interact with at least one gene not previously annotated by GWAS, including the regulation of *SOX4* by intronic *CDKAL1* enhancers, as previously proposed (Ragvin et al., 2010). Several of these putative SNP:gene interactions were validated experimentally using CRISPR. *rs11257655* was proximal to *CDC123* (~15 kb), however this region was found to be interacting with more distally located genes, *OPTN* and *CAMK1D* (located 834 kb and 84 kb away respectively). CRISPR deletion of the enhancer overlapping *rs11257655* led to downregulation of both *CAMK1D* and *OPTN*, with no effect on *CDC123*. In all of the loci tested, deletion of the regulatory element affected the expression of at least one of the genes that it was interacting with, with four regions affecting the expression of more than one gene (target gene multiplicity). This validation confirmed the functional importance of the relationships proposed by their pcHi-C maps, but also highlights the importance of using orthogonal methods to demonstrate that the spatial co-localisation of specific elements and genes does lead to downstream consequences on gene expression.

Whalen et al. compared maps of linkage disequilibrium (LD) with maps of chromatin architecture, derived from high-resolution Hi-C, in human and observed that they were largely uncorrelated and reflect the result of two distinct processes (Whalen and Pollard, 2019), with LD resulting from recombination events driven by PRDM9 and chromatin structure resulting from transcription, loop extrusion and other processes (Fudenberg et al., 2016; Rowley et al., 2017; Nuebler et al., 2018). LD blocks were found to be significantly smaller than TADs (median size 13 kb versus 840 kb). Therefore regulatory variants are typically located in LD blocks which do not overlap the gene(s) that they are involved in regulating, with only 2-7% of interactions between genes and noncoding elements being located within the same LD block, highlighting that the range of chromatin interactions is often much larger than the span of LD blocks.

Techniques such as ChIA-PET and HiChIP (Fullwood et al., 2009; Mumbach et al., 2016) combine proximity ligation with an antibody enrichment step, which allows the identification of chromatin interactions associated with a specific transcription factor or histone modification. A study investigating how chromatin interactions differ during the differentiation of embryonic stem cells (ESCs) into neural stem cells (NSCs) and neural progenitors (NPCs) found that the majority of putative enhancers (76%, 77%, 54% for ESC, NSCs and NPCs respectively) did not interact with their nearest gene in any of the cell-types investigated (Zhang et al., 2013). Maps of H3K27ac-associated chromatin interactions generated using HiChIP in distinct cell types from the T-cell lineage found that SNPs associated with autoimmune diseases were enriched in loop anchors. Only 14% of these autoimmune disease-associated SNPs interacted with their nearest gene, with the remainder skipping at least one gene to interact with a more distal gene, a feature the authors termed enhancer skipping (Mumbach et al., 2017).

Large-scale studies of the chromatin interactome have found that regulatory elements often interact with genes that are not linearly proximal, and that these interactions are highly cell-type specific and display enrichment for relevant TFBSs and processes (Beagan and Phillips-Cremins, 2020). These studies help to confirm that disease-associated variants do not need to be in LD to have an effect on the same gene, and that at numerous loci, the range of regulatory influence of an enhancer or SNP is primarily determined by topological organisation, and not by genomic distance. The chromatin interaction landscape is highly complex and cannot be easily predicted from the simple linear genome, due to pervasive features such as enhancer skipping and target gene multiplicity.

Just how problematic is nearest gene assignment?

Given a non-coding locus of interest there are several potential heuristics for assigning it to its putative target gene(s) (Fig. 4A). A region can simply be assigned the nearest gene in terms of genomic distance (N.) or to the nearest expressed gene (N.E.). N.E. assumes that observing a distal region to be interacting with a gene leads to the gene being expressed. However, given that TADs serve to demarcate the span of interactions between regulatory elements and their targets, incorporating information from Hi-C can be used to reduce the search space and instead ask what the nearest gene within the same TAD (N. in TAD) and nearest expressed gene within the same TAD (N. E. in TAD) are, as the region of interest (Fig. 4A).

By examining genome-wide maps of promoter centered interactions generated using pcHi-C in cardiac muscle (CM) and iPSCs (Montefiori et al., 2018) it is possible to assess the positive predictive value (PPV) for each of these heuristics. PPV (also known as precision) can be interpreted as the probability that an interaction predicted using one of these heuristics is true. For each distal anchor:promoter pair identified using pcHi-C we assessed whether this distal anchor would be assigned to the same gene using each of the four heuristics described above (Fig. 4B).

A PPV of 15.18% and 12.05%, for cardiac muscle (CM) and iPSCs respectively, was observed if only the nearest gene was considered as the target of an anchor. A minor increase in PPV was observed when considering only those genes which are expressed in the relevant cell type. Restricting the search space to only consider genes located within the same TAD as the distal anchor resulted in an increase in PPV to 16.40% and 13.00% for CMs and iPSCs respectively. This highlights that incorporating information on chromatin structure can improve performance in predicting the target of a non-coding region. An increase in PPV was observed only in CM when incorporating information on gene expression. Poor performance on this task was apparent for all of the heuristics investigated. This analysis indicates that studies that use nearest gene assignment (and other related heuristics) lack predictive power for a large number of genes in the genome.

Various algorithms for predicting enhancer–promoter interactions have been developed; for a comprehensive review see (Hariprakash and Ferrari, 2019). These techniques attempt to accomplish this task by using a combination of features including genomic distance, synteny, and one-dimensional (1D) local chromatin states such as transcription factor (TF) binding, histone modifications, and chromatin accessibility signatures.

IM-PET uses a set of four features: correlation between enhancer and target promoter activity profile, transcription factor and target promoter correlation, coevolution of enhancer and target promoter, and a distance constraint between enhancer and target promoter (He et al., 2014). He et al. showed that by incorporating multiple

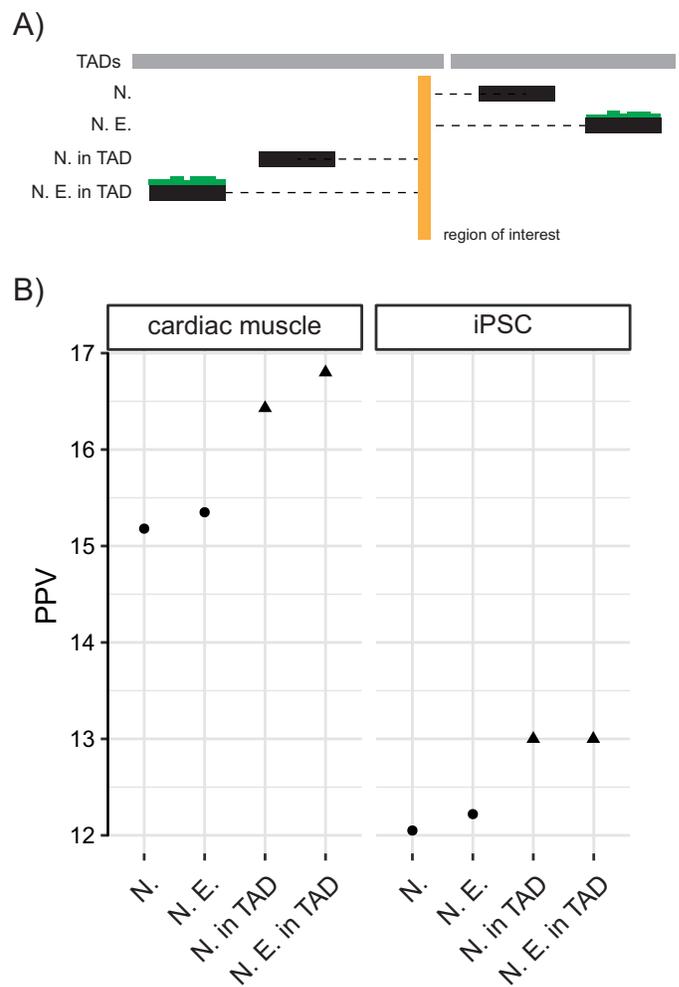


Fig. 4. Performance of nearest gene assignment and associated heuristics in cardiomyocytes. (A) Schematic of heuristics for assigning a region of interest to its potential target gene - nearest gene (N.), nearest expressed gene (N. E.), nearest gene within the same TAD (N. in TAD) and nearest expressed gene within the same TAD (N. E. in TAD). (B) Positive predictive value (PPV) for different heuristics to assign a non-coding region of interest to target genes in cardiac muscle and iPSCs.

features, IM-PET yielded an area under the curve (AUC) ROC of 94%, which was a significant increase in performance from using only nearest promoter as a predictor. GeneHancer uses combination of genomic distance, eQTL, capture Hi-C, eRNA co-expression, and TF co-expression (Fishilevich et al., 2017). While adding a distance as a feature led to the proposal of ~500,000 new gene–enhancer connections, the authors noted that none of the ~40,000 gene–enhancer connections obtained from the most stringent threshold were predicted using distance alone. The activity-by-contact (ABC) model minimally requires a measure of chromatin accessibility in the form of DNase-seq or ATAC-seq data and a measure of enhancer activity, usually H3K27ac (Fulco et al., 2019) and does not consider genomic distance in its predictions, apart from limiting to search space for putative enhancers around potential target genes. Fulco et al. found that predictions based solely on genomic distance achieved an area under the precision–recall curve (AUCPRC) of 0.39, whereas the same metric for the ABC model was 0.65.

In multiple studies attempting to predict enhancer–promoter interactions, genomic distance was found to be a useful feature when

considered in combination with other features, but on its own lacks predictive power and has been shown to be a poor predictor of enhancer–promoter interactions.

Conclusions

The phenomena of enhancers regulating genes other than the one they overlap with or are nearest to is extremely common genome-wide. This has been demonstrated by functional genomics studies dissecting individual gene regulatory domains, as well as from genome-wide studies of enhancer–promoter interactions using chromosome conformation capture techniques. This pattern of long-range regulation is reflected in the conservation of synteny, as observed in comparative genomics studies. All of these studies highlight that for a large number of loci in the genome, nearest gene assignment is wrong. Whilst studies have tried to understand the rules which determine the specificity of enhancer–promoter interactions, we still are lacking a systematic understanding of the features involved (Zabidi et al., 2015; Arnold et al., 2017; Zhou et al., 2020).

When performing genome-wide analysis, it is necessary to consider the impact of the topological organisation of the genome on the robustness of the results and annotation that is being proposed. In attempting to predict the relevant gene given a non-coding locus of interest, not considering information on topological structure and its conservation across both cell types and species and simply assigning a non-coding regulatory element or SNP to the nearest gene will often give misleading results, particularly in the case where the real gene of interest is under long-range regulation. Failure to adequately consider this could lead to incorrect hypotheses about putative causal genes, which would be both time-consuming and expensive. In addition, studies using and developing methods for predicting enhancer–promoter interactions should be first evaluated against nearest gene assignment and other related heuristics to identify whether these machine-learning techniques can first outperform these methods. In addition, it should be remembered that while the prevailing model of enhancer driven regulation is via direct physical interactions, it has been found that some enhancers do not need to be in close physical proximity to regulate gene expression (Benabdallah et al., 2019; Karr et al., 2022).

The development of experimental techniques to assay the regulatory landscape, the development of robust analysis pipelines, and the public availability of high-quality chromosome conformation data will enable researchers to drastically reduce the search space of potential target genes, which when followed by further computational analysis and experimental validation will help improve our mechanistic understanding of gene regulation and how its dysregulation impacts disease.

MATERIALS AND METHODS

Validated *Shh* enhancers were obtained from Jeong et al., Sagai et al. and Tsukiji et al. and lifted over to mm10 using rtracklayer (Jeong et al., 2006; Sagai et al., 2009; Tsukiji et al., 2014). Sets of conserved non-coding elements (CNEs) were obtained from ANCOR and smoothed using a sliding window approach to generate density tracks (Engström et al., 2008).

4C-seq from the brain of embryonic (E14.5) and adult mice was obtained from Smemo et al. and lifted over from mm9 to mm10 using rtracklayer (Smemo et al., 2014).

RNA-seq data for iPSC and cardiac muscle was obtained from E-MTAB-6013 (Montefiori et al., 2018) and aligned against the human genome (hg19 Ensembl 87) using STAR and quantified using RSEM (Dobin et al., 2013; Li and Dewey, 2011).

mESC Hi-C data was obtained from GEO:GSE96107 (Bonev et al., 2017) and aligned using BWA against mm10 and processed using FAN-C (Kruse

et al., 2020). Aligned data was filtered and binned into 40 kb bins and KR-normalised. TADs were identified using TopDom (Shin et al., 2016). Hi-C data for iPSC and cardiac muscle was downloaded from E-MTAB-6014 (Montefiori et al., 2018), aligned using BWA and processed using FAN-C. Aligned data was filtered and binned into 40 kb bins and KR-normalised. TADs were identified using TopDom. Promoter capture Hi-C data for iPSC and cardiac muscle was obtained from Montefiori et al. (2018), and was analysed using GenomicInteractions (Harmston et al., 2015). Visualisations of genomic data were generated using a combination of GViz (Hahne and Ivanek, 2016) and GenomicInteractions (Harmston et al., 2015).

For calculating the positive predictive value of various heuristics, interactions spanning longer than 2Mb were removed from the pHi-C datasets. Genes which were expressed at more than 1 TPM in at least one replicate of cardiac muscle cells or iPSCs were defined to be expressed in that cell type. A true positive was defined if the gene predicted by a heuristic matched one of the genes that the anchor/bait region was found to be interacting with (as in some cases a bait region can overlap the promoters of different genes). For nearest gene (N.) we identified which was the closest gene in terms of genomic distance to each of the anchors of interest, whereas for nearest expressed gene (N.E.), we first filtered out all genes that were not expressed at more than 1 TPM in at least one replicate in the corresponding cell line, we next identified which of these genes were closest to each of the anchors of interest. For assessing the performance of using the nearest gene in the same TAD (N. in TAD), we used TADs identified using TopDom to restrict the search space for potential genes and only calculated distances between genes and anchors of interest for genes which were present in the same TAD. For nearest expressed gene within the same TAD (N.E. in TAD), only those genes with a TAD that were expressed at more than one TPM in at least one replicate were considered. Results obtained from including Hi-C data (TADs) were robust to the choice of the parameter w .

All code necessary to recreate figures and analyses from this manuscript are available from: <https://github.com/harmstonlab/NearestGene>

Acknowledgements

The authors thank Elizabeth Ing-Simmons and Sara Haghani for their critical reading of the manuscript.

Competing interests

The authors declare no competing or financial interests.

Funding

This work was supported by Ministry of Education, National University of Singapore and Yale-NUS College (through Reimagine Research Grant IG20-RRSG-001).

References

- Acemel, R. D., Maeso, I. and Gómez-Skarmeta, J. L. (2017). Topologically associated domains: a successful scaffold for the evolution of gene regulation in animals: topologically associated domains. *WIREs Dev. Biol.* **6**, e265. doi:10.1002/wdev.265
- Akdemir, K. C., Le, V. T., Chandran, S., Li, Y., Verhaak, R. G., Beroukhim, R., Campbell, P. J., Chin, L., Dixon, J. R. and Futreal, P. A. (2020). Disruption of chromatin folding domains by somatic genomic rearrangements in human cancer. *Nat. Genet.* **52**, 294–305. doi:10.1038/s41588-019-0564-y
- Anderson, E., Devenney, P. S., Hill, R. E. and Lettice, L. A. (2014). Mapping the *Shh* long-range regulatory domain. *Development* **141**, 3934–3943. doi:10.1242/dev.108480
- Arnold, C. D., Zabidi, M. A., Pagani, M., Rath, M., Schernhuber, K., Kazmar, T. and Stark, A. (2017). Genome-wide assessment of sequence-intrinsic enhancer responsiveness at single-base-pair resolution. *Nat. Biotechnol.* **35**, 136–144. doi:10.1038/nbt.3739
- Austena, L. M. I., Barozzi, I., Simonatto, M., Masella, S., Della Chiara, G., Ghisletti, S., Curina, A., de Wit, E., Bouwman, B. A. M., de Pretis, S. et al. (2015). Transcription of mammalian cis-regulatory elements is restrained by actively enforced early termination. *Mol. Cell* **60**, 460–474. doi:10.1016/j.molcel.2015.09.018
- Babaei, S., Akhtar, W., de Jong, J., Reinders, M. and de Ridder, J. (2015). 3D hotspots of recurrent retroviral insertions reveal long-range interactions with cancer genes. *Nat. Commun.* **6**, 6381. doi:10.1038/ncomms7381
- Beagan, J. A. and Phillips-Cremins, J. E. (2020). On the existence and functionality of topologically associating domains. *Nat. Genet.* **52**, 8–16. doi:10.1038/s41588-019-0561-1

- Becker, T. S. and Lenhard, B. (2007). The random versus fragile breakage models of chromosome evolution: a matter of resolution. *Mol. Genet. Genomics* **278**, 487–491. doi:10.1007/s00438-007-0287-0
- Bejerano, G. (2004). Ultraconserved elements in the human genome. *Science* **304**, 1321–1325. doi:10.1126/science.1098119
- Belloni, E., Muenke, M., Roessler, E., Traverse, G., Siegel-Bartelt, J., Frumkin, A., Mitchell, H. F., Donis-Keller, H., Helms, C., Hing, A. V. et al. (1996). Identification of Sonic hedgehog as a candidate gene responsible for holoprosencephaly. *Nat. Genet.* **14**, 353–356. doi:10.1038/ng1196-353
- Benabdallah, N. S., Williamson, I., Illingworth, R. S., Kane, L., Boyle, S., Sengupta, D., Grimes, G. R., Therizols, P. and Bickmore, W. A. (2019). Decreased enhancer-promoter proximity accompanying enhancer activation. *Mol. Cell* **76**, 473–484.e7. doi:10.1016/j.molcel.2019.07.038
- Berthelot, C., Muffato, M., Abecassis, J. and Roest Crollius, H. (2015). The 3D organization of chromatin explains evolutionary fragile genomic regions. *Cell Rep.* **10**, 1913–1924. doi:10.1016/j.celrep.2015.02.046
- Bolt, C. C. and Duboule, D. (2020). The regulatory landscapes of developmental genes. *Development* **147**, dev171736. doi:10.1242/dev.171736
- Bolzer, A., Kreth, G., Solovei, I., Koehler, D., Saracoglu, K., Fauth, C., Müller, S., Eils, R., Cremer, C., Speicher, M. R. et al. (2005). Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS Biol.* **3**, e157. doi:10.1371/journal.pbio.0030157
- Bonev, B., Mendelson Cohen, N., Szabo, Q., Fritsch, L., Papadopoulos, G. L., Lubling, Y., Xu, X., Lv, X., Hugnot, J.-P., Tanay, A. et al. (2017). Multiscale 3D genome rewiring during mouse neural development. *Cell* **171**, 557–572.e24. doi:10.1016/j.cell.2017.09.043
- Briscoe, J. and Théron, P. P. (2013). The mechanisms of Hedgehog signalling and its roles in development and disease. *Nat. Rev. Mol. Cell Biol.* **14**, 416–429. doi:10.1038/nrm3598
- Clark, R. M., Marker, P. C. and Kingsley, D. M. (2000). A novel candidate gene for mouse and human preaxial polydactyly with altered expression in limbs of hemimelic extra-toes mutant mice. *Genomics* **67**, 19–27. doi:10.1006/geno.2000.6225
- Claussnitzer, M., Dankel, S. N., Kim, K.-H., Quon, G., Meuleman, W., Haugen, C., Glunk, V., Sousa, I. S., Beaudry, J. L., Puvion-Randall, V. et al. (2015). FTO obesity variant circuitry and adipocyte browning in humans. *N. Engl. J. Med.* **373**, 895–907. doi:10.1056/NEJMoa1502214
- Cremer, T. and Cremer, M. (2010). Chromosome territories. *Cold Spring Harbor Perspect. Biol.* **2**, a003889–a003889. doi:10.1101/cshperspect.a003889
- Creyghton, 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. et al. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl Acad. Sci. USA* **107**, 21931–21936. doi:10.1073/pnas.1016071107
- Dahn, R. D., Davis, M. C., Pappano, W. N. and Shubin, N. H. (2007). Sonic hedgehog function in chondrichthyan fins and the evolution of appendage patterning. *Nature* **445**, 311–314. doi:10.1038/nature05436
- de Bruijn, S. E., Fiorentino, A., Ottaviani, D., Fanucchi, S., Melo, U. S., Corral-Serrano, J. C., Mulders, T., Georgiou, M., Rivolta, C., Pontikos, N. et al. (2020). Structural variants create new topological-associated domains and ectopic retinal enhancer-gene contact in dominant retinitis pigmentosa. *Am. J. Hum. Genet.* **107**, 802–814. doi:10.1016/j.ajhg.2020.09.002
- de la Calle-Mustienes, E., Feijóo, C. G., Manzanares, M., Tena, J. J., Rodríguez-Seguel, E., Letizia, A., Allende, M. L. and Gómez-Skarmeta, J. L. (2005). A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate Iroquois cluster gene deserts. *Genome Res.* **15**, 1061–1072. doi:10.1101/gr.4004805
- Despang, A., Schöpflin, R., Franke, M., Ali, S., Jerković, I., Paliou, C., Chan, W.-L., Timmermann, B., Wittler, L., Vingron, M. et al. (2019). Functional dissection of the Sox9–Kcnj2 locus identifies nonessential and instructive roles of TAD architecture. *Nat. Genet.* **51**, 1263–1271. doi:10.1038/s41588-019-0466-z
- Dina, C., Meyre, D., Gallina, S., Durand, E., Körner, A., Jacobson, P., Carlsson, L. M. S., Kiess, W., Vatin, V., Lecoœur, C. et al. (2007). Variation in FTO contributes to childhood obesity and severe adult obesity. *Nat. Genet.* **39**, 724–726. doi:10.1038/ng2048
- Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S. and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380. doi:10.1038/nature11082
- Dixon, J. R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J. E., Lee, A. Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W. et al. (2015). Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331–336. doi:10.1038/nature14222
- Dixon, J. R., Gorkin, D. U. and Ren, B. (2016). Chromatin domains: the unit of chromosome organization. *Mol. Cell* **62**, 668–680. doi:10.1016/j.molcel.2016.05.018
- Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21. doi:10.1093/bioinformatics/bts635
- Engström, P. G., Ho Sui, S. J., Drivenes, O., Becker, T. S. and Lenhard, B. (2007). Genomic regulatory blocks underlie extensive microsynteny conservation in insects. *Genome Res.* **17**, 1898–1908. doi:10.1101/gr.6669607
- Engström, P. G., Fredman, D. and Lenhard, B. (2008). Ancora: a web resource for exploring highly conserved noncoding elements and their association with developmental regulatory genes. *Genome Biol.* **9**, R34. doi:10.1186/gb-2008-9-2-r34
- Fishilevich, S., Nudel, R., Rappaport, N., Hadar, R., Plaschkes, I., Iny Stein, T., Rosen, N., Kohn, A., Twik, M., Safran, M. et al. (2017). GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database* **2017**, bax028. doi:10.1093/database/bax028
- Flyamer, I. M., Gassler, J., Imakaev, M., Brandão, H. B., Ulianov, S. V., Abdennur, N., Razin, S. V., Mirny, L. A. and Tachibana-Konwalski, K. (2017). Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. *Nature* **544**, 110–114. doi:10.1038/nature21711
- Franke, M., Ibrahim, D. M., Andrey, G., Schwarzer, W., Heinrich, V., Schöpflin, R., Kraft, K., Kempfer, R., Jerković, I., Chan, W.-L. et al. (2016). Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* **538**, 265–269. doi:10.1038/nature19800
- Frayling, T. M., Timpson, N. J., Weedon, M. N., Zeggini, E., Freathy, R. M., Lindgren, C. M., Perry, J. R. B., Elliott, K. S., Lango, H., Rayner, N. W. et al. (2007). A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* **316**, 889–894. doi:10.1126/science.1141634
- Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N. and Mirny, L. A. (2016). Formation of chromosomal domains by loop extrusion. *Cell Rep.* **15**, 2038–2049. doi:10.1016/j.celrep.2016.04.085
- 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. et al. (2019). Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nat. Genet.* **51**, 1664–1669. doi:10.1038/s41588-019-0538-0
- Fullwood, M. J., Liu, M. H., Pan, Y. F., Liu, J., Xu, H., Mohamed, Y. B., Orlov, Y. L., Velkov, S., Ho, A., Mei, P. H. et al. (2009). An oestrogen-receptor- α -bound human chromatin interactome. *Nature* **462**, 58–64. doi:10.1038/nature08497
- Gerken, T., Girard, C. A., Tung, Y.-C. L., Webby, C. J., Saudek, V., Hewitt, K. S., Yeo, G. S. H., McDonough, M. A., Cunliffe, S., McNeill, L. A. et al. (2007). The obesity-associated FTO gene encodes a 2-oxoglutarate-dependent nucleic acid demethylase. *Science* **318**, 1469–1472. doi:10.1126/science.1151710
- Ghavi-Helm, Y., Jankowski, A., Meiers, S., Viales, R. R., Korbel, J. O. and Furlong, E. E. M. (2019). Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression. *Nat. Genet.* **51**, 1272–1282. doi:10.1038/s41588-019-0462-3
- Grunnet, L. G., Nilsson, E., Ling, C., Hansen, T., Pedersen, O., Groop, L., Vaag, A. and Poulsen, P. (2009). Regulation and function of FTO mRNA expression in human skeletal muscle and subcutaneous adipose tissue. *Diabetes* **58**, 2402–2408. doi:10.2337/db09-0205
- 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. et al. (2017). A genetic variant associated with five vascular diseases is a distal regulator of endothelin-1 gene expression. *Cell* **170**, 522–533.e15. doi:10.1016/j.cell.2017.06.049
- Hahne, F. and Ivanek, R. (2016). Visualizing genomic data using gviz and bioconductor, in: statistical genomics. *Methods Mol. Biol.* **1418**, 335–351. doi:10.1007/978-1-4939-3578-9_16
- Hariprakash, J. M. and Ferrari, F. (2019). Computational biology solutions to identify enhancers-target gene pairs. *Comput. Struct. Biotechnol. J.* **17**, 821–831. doi:10.1016/j.csbj.2019.06.012
- Harmston, N., Barešić, A. and Lenhard, B. (2013). The mystery of extreme non-coding conservation. *Phil. Trans. R. Soc. B* **368**, 20130021. doi:10.1098/rstb.2013.0021
- Harmston, N., Ing-Simmons, E., Perry, M., Barešić, A. and Lenhard, B. (2015). GenomicInteractions: an R/Bioconductor package for manipulating and investigating chromatin interaction data. *BMC Genomics* **16**, 963. doi:10.1186/s12864-015-2140-x
- Harmston, N., Ing-Simmons, E., Tan, G., Perry, M., Merkschlager, M. and Lenhard, B. (2017). Topologically associating domains are ancient features that coincide with Metazoan clusters of extreme noncoding conservation. *Nat. Commun.* **8**, 441. doi:10.1038/s41467-017-00524-5
- He, B., Chen, C., Teng, L. and Tan, K. (2014). Global view of enhancer-promoter interactions in human cells. *Proc. Natl. Acad. Sci. USA* **111**, E2191–E2199. doi:10.1073/pnas.1320308111
- Hill, R. E. and Lettice, L. A. (2013). Alterations to the remote control of Shh gene expression cause congenital abnormalities. *Phil. Trans. R. Soc. B* **368**, 20120357. doi:10.1098/rstb.2012.0357
- Hing, A. V., Helms, C., Slauch, R., Burgess, A., Wang, J. C., Herman, T., Downton, S. B. and Donis-Keller, H. (1995). Linkage of preaxial polydactyly type 2 to 7q36. *Am. J. Med. Genet* **58**, 128–135. doi:10.1002/ajmg.1320580208
- Hu, J., Zhang, Y., Zhao, L., Frock, R.-L., Du, Z., Meyers, R. M., Meng, F., Schatz, D. G. and Alt, F. W. (2015). Chromosomal loop domains direct the

- recombination of antigen receptor genes. *Cell* **163**, 947-959. doi:10.1016/j.cell.2015.10.016
- Hua, P., Badat, M., Hanssen, L. L. P., Hentges, L. D., Crump, N., Downes, D. J., Jeziorska, D. M., Oudelaar, A. M., Schwessinger, R., Taylor, S. et al. (2021). Defining genome architecture at base-pair resolution. *Nature* **595**, 125-129. doi:10.1038/s41586-021-03639-4
- Hunt, L. E., Noyvert, B., Bhaw-Rosun, L., Sesay, A. K., Paternoster, L., Nohr, E. A., Davey Smith, G., Tommerup, N., Sørensen, T. I. A. and Elgar, G. (2015). Complete re-sequencing of a 2Mb topological domain encompassing the FTO/IRX5 genes identifies a novel obesity-associated region upstream of IRX5. *Genome Med.* **7**, 126. doi:10.1186/s13073-015-0250-3
- Irimia, M., Tena, J. J., Alexis, M. S., Fernandez-Miñan, A., Maeso, I., Bogdanović, O., de la Calle-Mustienes, E., Roy, S. W., Gómez-Skarmeta, J. L. and Fraser, H. B. (2012). Extensive conservation of ancient microsynteny across metazoans due to cis-regulatory constraints. *Genome Res.* **22**, 2356-2367. doi:10.1101/gr.139725.112
- Jäger, R., Migliorini, G., Henrion, M., Kandaswamy, R., Speedy, H. E., Heindl, A., Whiffin, N., Carnicer, M. J., Broome, L., Dryden, N. et al. (2015). Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. *Nat. Commun.* **6**, 6178. doi:10.1038/ncomms7178
- Jeong, Y., El-Jaick, K., Roessler, E., Muenke, M. and Epstein, D. J. (2006). A functional screen for sonic hedgehog regulatory elements across a 1 Mb interval identifies long-range ventral forebrain enhancers. *Development* **133**, 761-772. doi:10.1242/dev.02239
- Jeong, Y., Leskow, F. C., El-Jaick, K., Roessler, E., Muenke, M., Yocum, A., Dubourg, C., Li, X., Geng, X., Oliver, G. et al. (2008). Regulation of a remote Shh forebrain enhancer by the Six3 homeoprotein. *Nat. Genet.* **40**, 1348-1353. doi:10.1038/ng.230
- Jindal, G. A. and Farley, E. K. (2021). Enhancer grammar in development, evolution, and disease: dependencies and interplay. *Dev. Cell* **56**, 575-587. doi:10.1016/j.devcel.2021.02.016
- Jung, I., Schmitt, A., Diao, Y., Lee, A. J., Liu, T., Yang, D., Tan, C., Eom, J., Chan, M., Chee, S. et al. (2019). A compendium of promoter-centered long-range chromatin interactions in the human genome. *Nat. Genet.* **51**, 1442-1449. doi:10.1038/s41588-019-0494-8
- Karr, J. P., Ferrie, J. J., Tjian, R. and Darzacq, X. (2022). The transcription factor activity gradient (TAG) model: contemplating a contact-independent mechanism for enhancer-promoter communication. *Genes Dev.* **36**, 7-16. doi:10.1101/gad.349160.121
- Kasowski, M., Grubert, F., Heffelfinger, C., Hariharan, M., Asabere, A., Waszak, S. M., Habegger, L., Rozowsky, J., Shi, M., Urban, A. E. et al. (2010). Variation in transcription factor binding among humans. *Science* **328**, 5. doi:10.1126/science.1183621
- Kikuta, H., Fredman, D., Rinkwitz, S., Lenhard, B. and Becker, T. S. (2007a). Retroviral enhancer detection insertions in zebrafish combined with comparative genomics reveal genomic regulatory blocks - a fundamental feature of vertebrate genomes. *Genome Biol.* **8** Suppl. 1, S4. doi:10.1186/gb-2007-8-s1-s4
- Kikuta, H., Laplante, M., Navratilova, P., Komisarczuk, A. Z., Engström, P. G., Fredman, D., Akalin, A., Caccamo, M., Sealy, I., Howe, K. et al. (2007b). Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome Res.* **17**, 545-555. doi:10.1101/gr.6086307
- Klötting, N., Schleinitz, D., Ruschke, K., Berndt, J., Fasshauer, M., Tönjes, A., Schön, M. R., Kovacs, P., Stumvoll, M. and Blüher, M. (2008). Inverse relationship between obesity and FTO gene expression in visceral adipose tissue in humans. *Diabetologia* **51**, 641-647. doi:10.1007/s00125-008-0928-9
- Komisarczuk, A. Z., Kawakami, K. and Becker, T. S. (2009). Cis-regulation and chromosomal rearrangement of the *fgf8* locus after the teleost/tetrapod split. *Dev. Biol.* **336**, 301-312. doi:10.1016/j.ydbio.2009.09.029
- Kruse, K., Hug, C. B. and Vaquerizas, J. M. (2020). FAN-C: a feature-rich framework for the analysis and visualisation of chromosome conformation capture data. *Genome Biol.* **21**, 303. doi:10.1186/s13059-020-02215-9
- Lango Allen, H., Caswell, R., Xie, W., Xu, X., Wragg, C., Turmpenny, P. D., Turner, C. L. S., Weedon, M. N. and Ellard, S. (2014). Next generation sequencing of chromosomal rearrangements in patients with split-hand/split-foot malformation provides evidence for *DVNC11* exonic enhancers of *DLX5/6* expression in humans. *J. Med. Genet.* **51**, 264-267. doi:10.1136/jmedgenet-2013-102142
- Lee, T. I. and Young, R. A. (2013). Transcriptional regulation and its misregulation in disease. *Cell* **152**, 1237-1251. doi:10.1016/j.cell.2013.02.014
- Lettice, L. A. (2003). A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum. Mol. Genet.* **12**, 1725-1735. doi:10.1093/hmg/ddg180
- Lettice, L. A., Horikoshi, T., Heaney, S. J. H., van Baren, M. J., van der Linde, H. C., Breedveld, G. J., Joosse, M., Akarsu, N., Oostra, B. A., Endo, N. et al. (2002). Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. *Proc. Natl Acad. Sci. USA* **99**, 7548-7553. doi:10.1073/pnas.112212199
- Lettice, L. A., Hill, A. E., Devenney, P. S. and Hill, R. E. (2008). Point mutations in a distant sonic hedgehog cis-regulator generate a variable regulatory output responsible for preaxial polydactyly. *Hum. Mol. Genet.* **17**, 978-985. doi:10.1093/hmg/ddm370
- Lettice, L. A., Williamson, I., Wiltshire, J. H., Peluso, S., Devenney, P. S., Hill, A. E., Essafi, A., Hagman, J., Mort, R., Grimes, G. et al. (2012). Opposing functions of the ETS factor family define Shh spatial expression in limb buds and underlie polydactyly. *Dev. Cell* **22**, 459-467. doi:10.1016/j.devcel.2011.12.010
- Li, B. and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323. doi:10.1186/1471-2105-12-323
- Liao, Y., Zhang, X., Chakraborty, M. and Emerson, J. J. (2021). Topologically associating domains and their role in the evolution of genome structure and function in *Drosophila*. *Genome Res.* **31**, 397-410. doi:10.1101/gr.266130.120
- 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. et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-293. doi:10.1126/science.1181369
- Long, H. K., Prescott, S. L. and Wysocka, J. (2016). Ever-changing landscapes: transcriptional enhancers in development and evolution. *Cell* **167**, 1170-1187. doi:10.1016/j.cell.2016.09.018
- Lupiáñez, D. G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J. M., Laxova, R. et al. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**, 1012-1025. doi:10.1016/j.cell.2015.04.004
- Marinić, M., Aktas, T., Ruf, S. and Spitz, F. (2013). An integrated holo-enhancer unit defines tissue and gene specificity of the Fgf8 regulatory landscape. *Dev. Cell* **24**, 530-542. doi:10.1016/j.devcel.2013.01.025
- Martin, P., McGovern, A., Orozco, G., Duffus, K., Yarwood, A., Schoenfelder, S., Cooper, N. J., Barton, A., Wallace, C., Fraser, P. et al. (2015). Capture Hi-C reveals novel candidate genes and complex long-range interactions with related autoimmune risk loci. *Nat. Commun.* **6**, 10069. doi:10.1038/ncomms10069
- Martin, P., McGovern, A., Massey, J., Schoenfelder, S., Duffus, K., Yarwood, A., Barton, A., Worthington, J., Fraser, P., Eyre, S. et al. (2016). Identifying causal genes at the multiple sclerosis associated region 6q23 using capture Hi-C. *PLoS One* **11**, e0166923. doi:10.1371/journal.pone.0166923
- Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J. et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190-1195. doi:10.1126/science.1222794
- McCord, R. P., Kaplan, N. and Giorgetti, L. (2020). Chromosome conformation capture and beyond: toward an integrative view of chromosome structure and function. *Mol. Cell* **77**, 688-708. doi:10.1016/j.molcel.2019.12.021
- Mifsud, B., Tavares-Cadete, F., Young, A. N., Sugar, R., Schoenfelder, S., Ferreira, L., Wingett, S. W., Andrews, S., Grey, W., Ewels, P. A. et al. (2015). Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.* **47**, 598-606. doi:10.1038/ng.3286
- Miguel-Escalada, I., Pasquali, L. and Ferrer, J. (2015). Transcriptional enhancers: functional insights and role in human disease. *Curr. Opin. Genet. Dev.* **33**, 71-76. doi:10.1016/j.gde.2015.08.009
- Miguel-Escalada, I., Bonàs-Guarch, S., Cebola, I., Ponsa-Cobas, J., Mendieta-Esteban, J., Atla, G., Javierre, B. M., Rolando, D. M. Y., Farabella, I., Morgan, C. C. et al. (2019). Human pancreatic islet three-dimensional chromatin architecture provides insights into the genetics of type 2 diabetes. *Nat. Genet.* **51**, 1137-1148. doi:10.1038/s41588-019-0457-0
- Montefiori, L. E., Sobreira, D. R., Sakabe, N. J., Aneas, I., Joslin, A. C., Hansen, G. T., Bozek, G., Moskowitz, I. P., McNally, E. M. and Nóbrega, M. A. (2018). A promoter interaction map for cardiovascular disease genetics. *eLife* **7**, e35788. doi:10.7554/eLife.35788.103
- Mumbach, M. R., Rubin, A. J., Flynn, R. A., Dai, C., Khavari, P. A., Greenleaf, W. J. and Chang, H. Y. (2016). HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat. Methods* **13**, 919-922. doi:10.1038/nmeth.3999
- 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. et al. (2017). Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. *Nat. Genet.* **49**, 1602-1612. doi:10.1038/ng.3963
- Musunuru, K., Strong, A., Frank-Kamenetsky, M., Lee, N. E., Ahfeldt, T., Sachs, K. V., Li, X., Li, H., Kuperwasser, N., Ruda, V. M. et al. (2010). From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* **466**, 714-719. doi:10.1038/nature09266
- Narendra, V., Bulajić, M., Dekker, J., Mazzoni, E. O. and Reinberg, D. (2016). CTCF-mediated topological boundaries during development foster appropriate gene regulation. *Genes Dev.* **30**, 2657-2662. doi:10.1101/gad.288324.116
- Navratilova, P., Fredman, D., Hawkins, T. A., Turner, C., Lenhard, B. and Becker, T. S. (2009). Systematic human/zebrafish comparative identification of cis-regulatory activity around vertebrate developmental transcription factor genes. *Dev. Biol.* **327**, 526-540. doi:10.1016/j.ydbio.2008.10.044
- Nuebler, J., Fudenberg, G., Imakaev, M., Abdennur, N. and Mirny, L. A. (2018). Chromatin organization by an interplay of loop extrusion and compartmental segregation. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E6697-E6706. doi:10.1073/pnas.1717730115

- Oliver, B. and Misteli, T. (2005). A non-random walk through the genome. *Genome Biol.* **6**, 214. doi:10.1186/gb-2005-6-4-214
- Paulsen, J., Sekelja, M., Oldenburg, A. R., Barateau, A., Briand, N., Delbarre, E., Shah, A., Sørensen, A. L., Vigouroux, C., Buendia, B. et al. (2017). Chrom3D: three-dimensional genome modeling from Hi-C and nuclear lamin-genome contacts. *Genome Biol.* **18**, 21. doi:10.1186/s13059-016-1146-2
- Phillips-Cremins, J. E., Sauria, M. E. G., Sanyal, A., Gerasimova, T. I., Lajoie, B. R., Bell, J. S. K., Ong, C.-T., Hookway, T. A., Guo, C., Sun, Y. et al. (2014). Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell* **153**, 1281-1295. doi:10.1016/j.cell.2013.04.053
- Pope, B. D., Ryba, T., Dileep, V., Yue, F., Wu, W., Denas, O., Vera, D. L., Wang, Y., Hansen, R. S., Canfield, T. K. et al. (2014). Topologically associating domains are stable units of replication-timing regulation. *Nature* **515**, 402-405. doi:10.1038/nature13986
- Ragvin, A., Moro, E., Fredman, D., Navratilova, P., Drivenes, Ø., Engstrom, P. G., Alonso, M. E., Mustienes, E. d. I. C., Skarmeta, J. L. G., Tavares, M. J. et al. (2010). Long-range gene regulation links genomic type 2 diabetes and obesity risk regions to HHEX, SOX4, and IRX3. *Proc. Natl. Acad. Sci. USA* **107**, 775-780. doi:10.1073/pnas.0911591107
- Rands, C. M., Meader, S., Ponting, C. P. and Lunter, G. (2014). 8.2% of the human genome is constrained: variation in rates of turnover across functional element classes in the human lineage. *PLoS Genet.* **10**, e1004525. doi:10.1371/journal.pgen.1004525
- Rao, S. S. P., 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. et al. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-1680. doi:10.1016/j.cell.2014.11.021
- Rinkwitz, S., Geng, F.-S., Manning, E., Suster, M., Kawakami, K. and Becker, T. S. (2015). BAC transgenic zebrafish reveal hypothalamic enhancer activity around obesity associated SNP rs9939609 within the human FTO gene. *Genesis* **53**, 640-651. doi:10.1002/dvg.22884
- Roessler, E., Belloni, E., Gaudenz, K., Jay, P., Berta, P., Scherer, S. W., Tsui, L. C. and Muenke, M. (1996). Mutations in the human Sonic Hedgehog gene cause holoprosencephaly. *Nat. Genet.* **14**, 357-360. doi:10.1038/ng1196-357
- Rowley, M. J., Nichols, M. H., Lyu, X., Ando-Kuri, M., Rivera, I. S. M., Hermetz, K., Wang, P., Ruan, Y. and Corces, V. G. (2017). Evolutionarily conserved principles predict 3D chromatin organization. *Mol. Cell* **67**, 837-852.e7. doi:10.1016/j.molcel.2017.07.022
- Royo, J. L., Bessa, J., Hidalgo, C., Fernández-Miñán, A., Tena, J. J., Roncero, Y., Gómez-Skarmeta, J. L. and Casares, F. (2012). Identification and analysis of conserved cis-regulatory regions of the MEIS1 gene. *PLoS One* **7**, e33617. doi:10.1371/journal.pone.0033617
- Sagai, T., Hosoya, M., Mizushina, Y., Tamura, M. and Shiroishi, T. (2005). Elimination of a long-range cis-regulatory module causes complete loss of limb-specific Shh expression and truncation of the mouse limb. *Development* **132**, 797-803. doi:10.1242/dev.01613
- Sagai, T., Amano, T., Tamura, M., Mizushina, Y., Sumiyama, K. and Shiroishi, T. (2009). A cluster of three long-range enhancers directs regional Shh expression in the epithelial linings. *Development* **136**, 1665-1674. doi:10.1242/dev.032714
- Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014). Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421-427. doi:10.1038/nature13595
- Schoenfelder, S., Furlan-Magaril, M., Mifsud, B., Tavares-Cadete, F., Sugar, R., Javierre, B.-M., Nagano, T., Katsman, Y., Sakthidevi, M., Wingett, S. W. et al. (2015). The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Res.* **25**, 582-597. doi:10.1101/gr.185272.114
- Scuteri, A., Sanna, S., Chen, W.-M., Uda, M., Albai, G., Strait, J., Najjar, S., Nagaraja, R., Orrú, M., Usala, G. et al. (2007). Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. *PLoS Genet.* **3**, 11. doi:10.1371/journal.pgen.0030115
- Shin, H., Shi, Y., Dai, C., Tjong, H., Gong, K., Alber, F. and Zhou, X. J. (2016). TopDom: an efficient and deterministic method for identifying topological domains in genomes. *Nucleic Acids Res.* **44**, e70. doi:10.1093/nar/gkv1505
- Shopland, L. S., Lynch, C. R., Peterson, K. A., Thornton, K., Kepper, N., von Hase, J., Stein, S., Vincent, S., Molloy, K. R., Kreth, G. et al. (2006). Folding and organization of a contiguous chromosome region according to the gene distribution pattern in primary genomic sequence. *J. Cell Biol.* **174**, 27-38. doi:10.1083/jcb.200603083
- Siersbæk, R., Madsen, J. G. S., Javierre, B. M., Nielsen, R., Bagge, E. K., Cairns, J., Wingett, S. W., Traynor, S., Spivakov, M., Fraser, P. et al. (2017). Dynamic rewiring of promoter-anchored chromatin loops during adipocyte differentiation. *Mol. Cell* **66**, 420-435.e5. doi:10.1016/j.molcel.2017.04.010
- Smemo, S., Tena, J. J., Kim, K.-H., Gamazon, E. R., Sakabe, N. J., Gómez-Marín, C., Aneas, I., Credidio, F. L., Sobreira, D. R., Wasserman, N. F. et al. (2014). Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature* **507**, 371-375. doi:10.1038/nature13138
- Spielmann, M. and Mundlos, S. (2016). Looking beyond the genes: the role of non-coding variants in human disease. *Hum. Mol. Genet.* **25**, R157-R165. doi:10.1093/hmg/ddw205
- Symmons, O., Uslu, V. V., Tsujimura, T., Ruf, S., Nassari, S., Schwarzer, W., Ettwiller, L. and Spitz, F. (2014). Functional and topological characteristics of mammalian regulatory domains. *Genome Res.* **24**, 390-400. doi:10.1101/gr.163519.113
- Taberlay, P. C., Achinger-Kawecka, J., Lun, A. T. L., Buske, F. A., Sabir, K., Gould, C. M., Zotenko, E., Bert, S. A., Giles, K. A., Bauer, D. C. et al. (2016). Three-dimensional disorganization of the cancer genome occurs coincident with long-range genetic and epigenetic alterations. *Genome Res.* **26**, 719-731. doi:10.1101/gr.201517.115
- Tolhuis, B., Palstra, R.-J., Splinter, E., Grosveld, F. and de Laat, W. (2002). Looping and interaction between hypersensitive sites in the active β -globin locus. *Mol. Cell* **10**, 1453-1465. doi:10.1016/S1097-2765(02)00781-5
- Torosin, N. S., Anand, A., Golla, T. R., Cao, W. and Ellison, C. E. (2020). 3D genome evolution and reorganization in the Drosophila melanogaster species group. *PLoS Genet.* **16**, e1009229. doi:10.1371/journal.pgen.1009229
- Tsukiji, N., Amano, T. and Shiroishi, T. (2014). A novel regulatory element for Shh expression in the lung and gut of mouse embryos. *Mech. Dev.* **131**, 127-136. doi:10.1016/j.mod.2013.09.003
- Uren, A. G., Kool, J., Matentzoglou, K., de Ridder, J., Mattison, J., van Uitert, M., Lagcher, W., Sie, D., Tanger, E., Cox, T. et al. (2008). Large-scale mutagenesis in p19ARF- and p53-deficient mice identifies cancer genes and their collaborative networks. *Cell* **133**, 727-741. doi:10.1016/j.cell.2008.03.021
- Van Bortle, K., Nichols, M. H., Li, L., Ong, C.-T., Takenaka, N., Qin, Z. S. and Corces, V. G. (2014). Insulator function and topological domain border strength scale with architectural protein occupancy. *Genome Biol.* **15**, R82. doi:10.1186/gb-2014-15-5-r82
- Vicente-García, C., Villarejo-Balcells, B., Irastorza-Azcárate, I., Naranjo, S., Acemel, R. D., Tena, J. J., Rigby, P. W. J., Devos, D. P., Gómez-Skarmeta, J. L. and Carvajal, J. J. (2017). Regulatory landscape fusion in rhabdomyosarcoma through interactions between the PAX3 promoter and FOXO1 regulatory elements. *Genome Biol.* **18**, 106. doi:10.1186/s13059-017-1225-z
- Vietri Rudan, M., Barrington, C., Henderson, S., Ernst, C., Odom, D. T., Tanay, A. and Hadjur, S. (2015). Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell Rep.* **10**, 1297-1309. doi:10.1016/j.celrep.2015.02.004
- Weischenfeldt, J., Dubash, T., Drinain, A. P., Mardin, B. R., Chen, Y., Stütz, A. M., Waszak, S. M., Bosco, G., Halvorsen, A. R., Raeder, B. et al. (2017). Pan-cancer analysis of somatic copy-number alterations implicates IRS4 and IGF2 in enhancer hijacking. *Nat. Genet.* **49**, 65-74. doi:10.1038/ng.3722
- Whalen, S. and Pollard, K. S. (2019). Most chromatin interactions are not in linkage disequilibrium. *Genome Res.* **29**, 334-343. doi:10.1101/gr.238022.118
- Won, H., de la Torre-Ubieta, L., Stein, J. L., Parikhshak, N. N., Huang, J., Opland, C. K., Gandai, M. J., Sutton, G. J., Hormozdiari, F., Lu, D. et al. (2016). Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature* **538**, 523-527. doi:10.1038/nature19847
- Zabidi, M. A., Arnold, C. D., Schernhuber, K., Pagani, M., Rath, M., Frank, O. and Stark, A. (2015). Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature* **518**, 556-559. doi:10.1038/nature13994
- Zhang, Y., Wong, C.-H., Birnbaum, R. Y., Li, G., Favaro, R., Ngan, C. Y., Lim, J., Tai, E., Poh, H. M., Wong, E. et al. (2013). Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. *Nature* **504**, 306-310. doi:10.1038/nature12716
- Zhao, J., Ding, J., Li, Y., Ren, K., Sha, J., Zhu, M. and Gao, X. (2009). HnRNP U mediates the long-range regulation of Shh expression during limb development. *Hum. Mol. Genet.* **18**, 3090-3097. doi:10.1093/hmg/ddp250
- Zhou, J., Liu, R., Wu, Z., Zhang, J. and Liu, J. (2020). Exploiting epigenomic and sequence-based features for predicting enhancer-promoter interactions. *Bioinformatics* **33**, i252-i260. doi:10.1051/e3sconf/202021803046