Computational and Structural Biotechnology Journal 18 (2020) 2945–2952

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Prioritizing long range interactions in noncoding regions using GWAS and deletions perturbed TADs

Xuanshi Liu [a,b,c], Wenjian Xu [a,b,c], Fei Leng [a,b,c], Chanjuan Hao [a,b,c], Sree Rohit Raj Kolora [d], Wei Li [a,b,c,*]

[a] Beijing Key Laboratory for Genetics of Birth Defects, Beijing Pediatric Research Institute, Beijing, China
[b] MOE Key Laboratory of Major Diseases in Children, Beijing, China
[c] Genetics and Birth Defects Control Center, Beijing Children's Hospital, Capital Medical University, National Center for Children's Health, Beijing, China
[d] Department of Integrative Biology, University of California Berkeley, Berkeley, CA, USA

ARTICLE INFO

Article history:
Received 28 July 2020
Received in revised form 8 October 2020
Accepted 12 October 2020
Available online 21 October 2020

Keywords:
Noncoding region interpretation
Genome wide association study
Deletion
Topological associated domain
Enhancer

ABSTRACT

Genome-wide association studies (GWAS) have contributed significantly to predisposing the disease etiology by associating single nucleotide polymorphisms (SNPs) with complex diseases. However, most GWAS-SNPs are in the noncoding regions that may affect distal genes via long range enhancer-promoter interactions. Thus, the common practice on GWAS discoveries cannot fully reveal the molecular mechanisms underpinning complex diseases. It is known that perturbations of topological associated domains (TADs) lead to long range interactions which underlie disease etiology. To identify the probable long range interactions in noncoding regions via GWAS and TADs perturbed by deletions, we integrated datasets from GWAS-SNPs, enhancers, TADs, and deletions. After ranking and clustering, we prioritized 201,132 high confident pairs of GWAS-SNPs and target genes. In this study, we performed a systematic inference on noncoding regions via GWAS-SNPs and deletion-perturbed TADs to boost GWAS discovery power. The high confident pairs of GWAS-SNPs and target genes (SE-Gs) provide the promising candidates to understand the molecular mechanisms underlying complex diseases with emphasis on the three-dimensional genome.

1. Introduction

Genome-wide association study (GWAS) is a widely adopted approach to define single nucleotide polymorphisms (SNPs) associated with complex diseases [1,2]. However, GWAS-SNPs predominantly fall into noncoding regions [3]. Despite efforts that have been made [4–6], the challenge of translating noncoding GWAS-SNPs into underlying biological mechanisms remains. Interpretations of GWAS findings are further complicated by noncoding GWAS-SNPs which can affect distal genes through long range enhancer-promoter interactions, e.g. an *FTO* intronic variant embedded in an enhancer regulating *IRX3* in ~ 490 kb away [7], an intergenic schizophrenia-associated SNP regulating *FOXG1* gene ~ 760 kb away [8], an intronic type 2 diabetes associated SNP regulating *ACSL5* gene ~ 624 kb away [9]. Moreover, large deletions (DELs) likely occur around GWAS-SNPs affecting distal target genes [10]. Therefore, the common practice on mapping SNPs to the nearest genes or finding causal variants by linkage disequilibrium (LD) can generate false positive results.

The advanced technologies and growing number of functional genomics data could narrow this gap of knowledge. Hi-C and related technologies have discovered the spatial genome structure, topological associated domain (TAD), which is relatively stable across cell types and species [11,12]. Perturbations of TADs can lead to long range interactions and cause diseases, such as the dysregulation of *IRS4* in sarcoma and squamous cancer is associated with DELs at one specific TAD boundary [13]; a type of limb malformations (brachydactyly) is caused by DELs disrupted TAD borders and produced abnormal gene expressions [14]. Mechanistic studies collectively suggest that probable long range interpretations can be prioritized from GWAS-SNPs that embedded in enhancers and genes within DELs-perturbed TADs. Although emerging

Abbreviations: 3D, three-dimensional; GWAS, genome wide association study; SNP, single nucleotide polymorphism; DEL, deletion; TAD, topological associated domain; GWAS-SNP, SNP significantly associated with diseases or traits in GWAS; DEL-TAD, TAD borders were interrupted by DEL border interrupted by DEL; SE-G, a pair of GWAS-SNP and target gene.

* Corresponding author: Children's Hospital, Capital Medical University, Beijing 100045, China.
E-mail address: liwei@bch.com.cn (W. Li).

https://doi.org/10.1016/j.csbj.2020.10.014

methods or databases have added TADs to gain insights into noncoding regions, in much the same way as 3Disease [15] aims to investigate the chromosome translocations with TADs, GWAS4D integrates Hi-C data and functional annotations on noncoding variants [16]. Thus far, a systematic study on noncoding GWAS-SNPs and genes within DELs-perturbed TADs is still lacking.

Here, we describe a scoring system to decipher GWAS findings at noncoding regions using DELs-perturbed TADs. After integrating massive data, we ranked GWAS-SNPs based on their potential regulatory functions and DELs-perturbed TADs based on their consistencies. Finally, we established the connection between GWAS-SNPs and target genes within DELs-perturbed TADs based on their closest genomic distances. Our work could provide new insights into GWAS discovery by locating functional GWAS-SNPs and linking them to the potential affected genes inferred from three-dimensional genome context.

2. Materials and methods

2.1. Data collections

We collected GWAS-SNPs, enhancers, TADs, and DELs data from 11 different sources listed in Table 1.

The GWAS-SNPs were aggregated from the GWAS Catalog [1] (1) and PhenoScanner V2 [2] (2). We retained SNPs with rs numbers and with p value<$1*10^{-5}$, in order to include SNPs with a potential biological significant as well as to minimize the potential false positive. In total, we got 2,640,328 diseases/traits associated non-redundant SNPs for further analysis (Fig. S1A). For enhancers, we obtained 65,423 enhancers from the Functional ANnoTation Of the Mammalian genome (FANTOM) [17] and 2,255,761 enhancers from Chromatin State Segmentation by HMM (ChromHMM) marked by 4_Strong_Enhancer, 5_Strong_Enhancer, 6_Weak_Enhancer, 7_Weak_Enhancer [18]. Furthermore, we downloaded TAD data generated by Hi-C Seq under 40 kb resolutions from 20 cell lines in Job Dekker's laboratory (https://www.encodeproject.org/data/). Additionally, we downloaded 20,124 protein coding genes from GENCODE (v30lift37) to locate target genes within DELs perturbed TADs. As for DELs (one large type of structural variations), we collected a comprehensive list of structural variations from various sources [21–24] and extracted 818,716 DELs out of all sources.

2.2. Scoring scheme

We hypothesize the presence of long range interactions between enhancers and closest genes through DELs-perturbed TADs. To model it, we designed a metric covering an enhancer confident score and a DEL-TAD score. The complete workflow is represented in Fig. 1.

We first calculated the enhancer confident score by combining the sum of weighted regulatory function scores and numbers of overlapped enhancers. For each GWAS-SNP, we calculated the regulatory function score by summing up available scores generated by eight algorithms if pre-defined thresholds were met (Table S1). The following eight algorithms integrated in SNPnexus tool [25] were used: CADD [26], GWAVA [27], fitCons [28], DeepSEA [29], EIGEN [30], FunSeq2 [31], FATHMM-MKL [32] and ReMM [33]. After annotating, the remaining GWAS-SNPs were 2,639,858. The overlapped enhancers were generated through the following steps: If there was an enhancer found within 10 bp flanking regions of GWAS-SNPs, we recorded it as 1, otherwise as 0. We then marked each GWAS-SNP by the number of overlapped enhancers and used the enhancer confident score to reflect the possible enhancer function. Together, the enhancer confident score is calculated as follows:

$$S = W_{(hits/8)} \times \sum_{i}^{8} S_i + S_{enh}$$

The $W_{(hits/8)}$ stands for the number of algorithms which have scores on GWAS-SNP divided by totally eight algorithms. S_i is the regulatory function score generated by the i^{th} algorithm. S_{enh} refers to the counts of overlapped enhancers on each GWAS-SNP. A cut-off of 0.557 was used since it gives the best performance, and higher than 0.557 meant the GWAS-SNP carried potential regulatory function.

We then defined a DEL-TAD score to measure the genome wide possibility that TAD boundaries affected by DELs. For each TAD, a DEL-TAD score ($S_{DEL\text{-}TAD}$) was defined as the TADs consistency multiplied by the DEL-TAD frequency:

$$S_{DEL-TAD} = 2 \times S_{TAD-freq} \times (S_{DEL-TAD-left} + S_{DEL-TAD-right})$$

The $S_{TAD\text{-}freq}$ refers to the overlapped number of TADs from cell lines. $S_{DEL\text{-}TAD\text{-}left}$ and $S_{DEL\text{-}TAD\text{-}right}$ are the min–max normalized values over the number of overlapping DELs detected at the left or right boundaries of TADs, respectively.

Finally, we combined enhancers and affected genes by connecting GWAS-SNPs to DEL-TADs based on the genomic proximity, i.e. the affected genes within DEL-TADs were assigned to the closest GWAS-SNPs. We kept only pairs of GWAS-SNP and target gene (SE-Gs) on either side of the border where DELs-perturbed TADs.

2.3. External data

To evaluate the enhancer confident scores, we compared the GWAS-SNPs with 1,339 enhancers documented at VISTA [34] by

Table 1
Summary of data sources.[a]

	Database	Total number of inputs	Average length (bp)	Coverage of total genome[b] (%)
GWAS-SNP	GWAS catalog	58,134	1	0.0000188
	PhenoScanner 2.0	2,629,046	1	0.000849
Enhancer	ChromHMM	2,255,761	532	0.388
	FANTOM5	65,423	281	0.00594
TAD	ENCODE	44,177	810,640	11.6
DEL	1000 Genome	42,279	9,444	0.129
	Audano *et al.*	34,211	449	0.00496
	Chaisson *et al.*	37,172	7,343	0.0882
	Ensembl	1,686,961	8,453	4.61
	GnomAD	176,222	7,483	0.426
	GoNL	40,550	1,138	0.0149

[a] Date at data access: GWAS-Catalog (Jan. 2019), PhenoScanner 2.0 (Jul. 2019), ChromHMM (Jan. 2019), FANTOM5 (Jan. 2019), TAD-ENCODE (Jan. 2019), 1000 Genome (Aug. 2019), Audano *et al.* (Aug. 2019). Chaisson *et al.* (Aug. 2019), Ensembl (Jul. 2019), GnomeAD (Aug. 2019), GoNL (Aug. 2019).
[b] Total genome refers to the length of genome from chromosome 1 to chromosome Y.

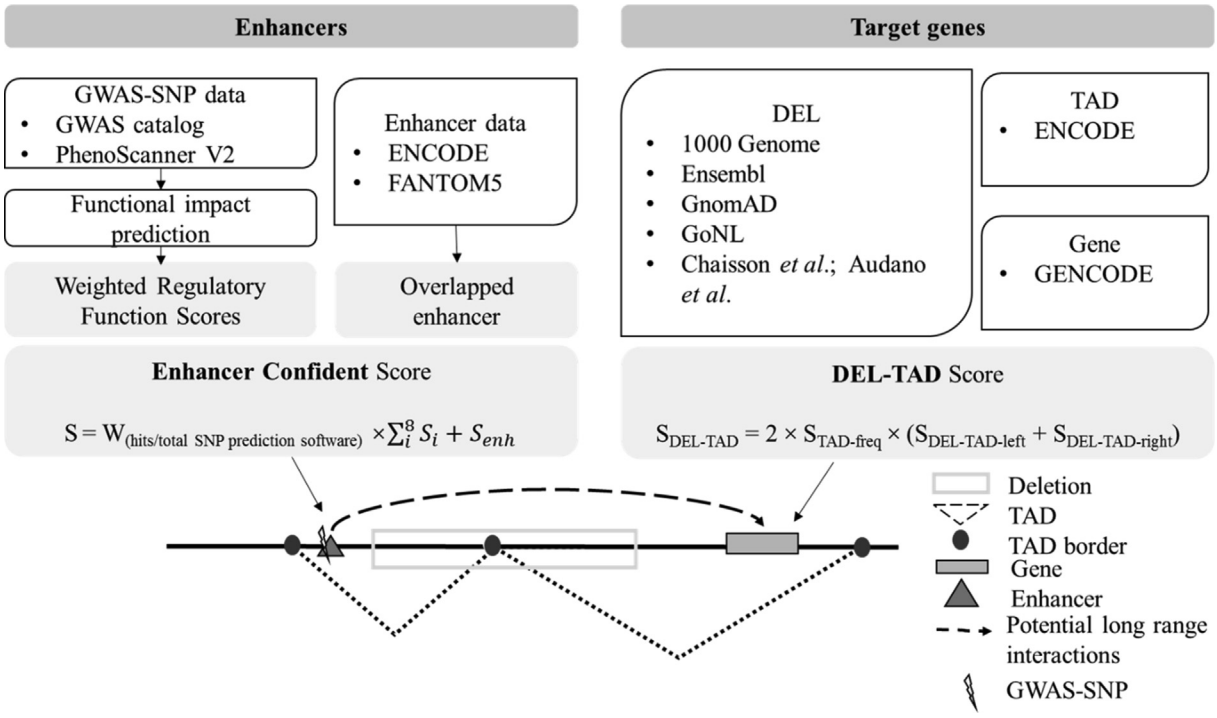

Fig. 1. An overview of analysis pipeline. A relatively comprehensive resource of GWAS-SNPs, enhancers, DELs, TADs and protein coding genes were collected from databases and publications. Pairs of SE-Gs were ranked according to enhancer confident scores and DEL-TAD scores, where an enhancer confident score for each GWAS-SNP was calculated by summing up weighted regulatory function scores and the numbers of overlapped enhancers, and DEL-TAD was based on conservation. GWAS-SNPs and target genes were associated by the closest genomic distances between GWAS-SNPs and DELs perturbed TADs. GWAS: Genome Wide Association Study; SNP: Single Nucleotide Polymorphism; DEL: Deletion; TAD; Topological Associated Domain; SE-Gs, pairs of the GWAS-SNP and the target gene.

calculating the area under the curve (AUC). Since enhancers from VISTA are experimentally validated, GWAS-SNPs with enhancer confident scores and located in enhancer regions were considered true positives (TP). False positives (FP) were defined as those with enhancer confident score, but not in VISTA enhancer regions. True negatives (TN) were those not predicted by enhancer confident score and not found by VISTA enhancers, and false negatives (FN) were those not predicted by enhancer confident score, but overlapped with VISTA enhancers.

To further assess the performance, we calculated the numbers of enhancer-gene pairs between SE-Gs with data from the DiseaseEnhancer database (version 1.0.2) [35] and generated by promoter capture Hi-C (pcHi-C) experiment [36]. We retained 1,122 unique one-to-one enhancer target gene pairs, and 131,843 GWAS-SNPs target genes pairs for validation, respectively.

2.4. Statistical analysis

Statistical analyses and plots were generated by R 3.6.1, notably using the package ggplots and UpSetR. Data integration and mining were done by in house shell scripts, Bedtools (v2.26.0) and Perl v5.16.3. All genomic data were mapped to the hg19 genome assembly. The performance was assessed by:

$$Sensitivity = TP/(TP + FN)$$

$$Specificity = TN/(TN + FP)$$

Enrichment analyses were conducted by an R package ClusterProfiler. *P* values from enrichment analyses were multiple corrected by the Benjamín-Hochberg method to calculate *q* values. For ranked comparisons, we used the Wilcoxon Signed-Rank Test for paired samples. To evaluate the enhancer confident scores and the high confident SE-Gs, we took one sided Pearson's Chi-squared test.

3. Results

3.1. Enhancer confident scores prioritize GWAS-SNPs associated enhancer functions

According to the design, enhancer confident scores consist of weighted regulatory function scores and overlapped enhancers. As for regulatory function predictions, 66.42% of GWAS-SNPs were scored as functionally relevant. We used a combination of eight algorithms because that the computational methods behind differed to a certain extent, and one algorithm alone could not comprise all the possibilities. In our data, we observed that the scored GWAS-SNPs were found at most by three algorithms (Fig. 2).

To dissect enhancers from regulatory elements, we then intersected scored GWAS-SNPs with enhancers documented in ChromHMM and FAMTON in order to calculate the number of overlapped enhancers. We found that 23.07% of GWAS-SNPs at 10 bp flanking regions overlapped with at least one enhancer suggesting that these GWAS-SNPs were probably embedded in the enhancers. Considering the overlapped enhancers in each database, 271,918 GWAS-SNPs (22.85%) overlapped with at least one enhancer in ChromHMM, and 8,530 GWAS-SNPs (0.73%) in FAMTON (Fig. S2). This difference in the number of GWAS-SNPs indicates that enhancers identified through machine learning models with omics data and CAGE experiments have different coverages. Thus, relying on one type of data would result in low sensitivity in enhancer identification.

Finally, we combined weighted regulatory function scores and overlapped enhancers in order to generate the enhancer confident scores. We observed a rather similar distribution between enhancer confident scores and weighted regulatory function score suggesting a combination of these scores could help to prioritize enhancers (Fig. S1B).

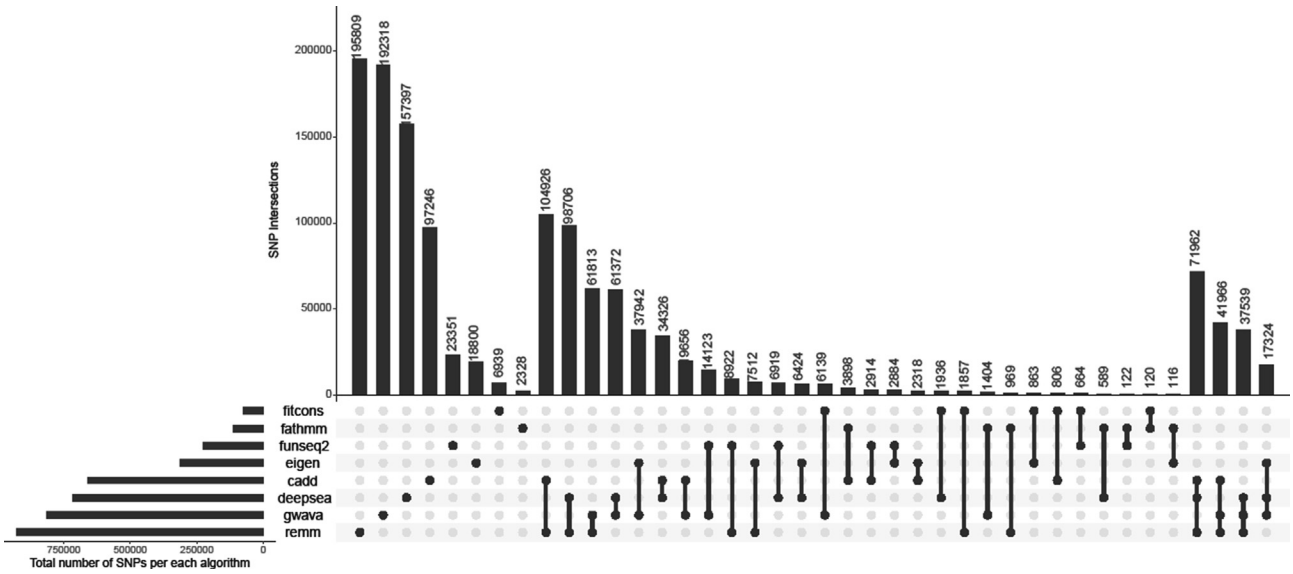

Fig. 2. UpSet plot of interactions among sets of scored GWAS-SNPs from eight algorithms. The bar chart from the left indicates the total number of scored GWAS-SNPs in each algorithm. The upper panel bar chart reflects the intersection size between sets of scored GWAS-SNPs from algorithms. The dark connected dots on the bottom panel show which algorithms are considered for each intersection.

3.2. DEL-TAD scores pinpoint the target genes

To detect DELs interrupted TADs in a consensus way, we first defined the TAD conservation score which is equal to the number of identical TAD boundaries across 20 cell lines. Among 44,177 non-redundant TADs, 168 identical non-redundant TADs were found among 20 cell lines. The median number of identical TADs found from 20 cell lines was 4, which suggests that TADs have certain degrees of conservation. This is in line with previous findings that TADs are preferentially invariant, but it can be varied by tissues and developmental stages [12]. Then we checked the distribution of TADs at chromosome level (Fig. S3A). The TADs span over the entire genome. This proved that our work covered the whole genome level.

Next, we analyzed the breadth and depth of DELs to ensure the detection of overlapped TAD borders in genome wide. The DELs were ranging from 50 bp to 223,214,370 bp and spanning over the genome (Fig. S3B). Then, we performed the analysis on the depth of DELs at base level and observed a mean depth of 18.28 (Table S2). Here, we used the depth of DEL as an analogue of the frequency of a DEL occurring in population because GWAS was built on "common disease common variant" and a rare DEL in the population scale, suggesting that it might have a lower possibility of developing common diseases. Subsequently, we considered DEL-TADs as DELs present within TAD boundaries. In total, 99% of TAD boundaries overlapped with at least one DEL (Fig. S4).

Combining conservation scores and overlapped DEL-TAD, we furthermore generated DEL-TAD scores and we set the cut-off of DEL-TAD score as 2 based on performance. A score greater than 2 may lead to a possible DEL perturbed TAD event.

3.3. Potential associations between GWAS-SNPs and target genes are evaluated by high confident SE-Gs

We associated SE-Gs by the closest genomic distances between GWAS-SNPs and target genes in DEL-TADs. In total, 3,245,076 pairs of SE-Gs were identified and the average distance between SE-Gs was 436,494 bp. Among all pairs, we defined high confident SE-Gs as enhancer confident score greater than 0.557 and DEL-TAD score greater than 2, resulting in 201,132 pairs. These SE-Gs included 162,421 GWAS-SNPs and 2,587 genes with an average distance of 403,329 bp. A complete list of high confident SE-Gs is provided in Table S3.

To decipher noncoding regions, it is obvious to investigate the implications from high confident SE-Gs in GWAS-SNPs and target genes, respectively. We first explored the GWAS-SNPs associated diseases. where we compared associated diseases between original GWAS-SNPs and high confident GWAS-SNPs. In doing so, we noticed a significant difference ($p < 0.22 \times 10^{-15}$, Wilcoxon Signed-Rank Test), indicating that GWAS-SNPs with potential enhancer functions might be enriched in certain diseases. After performing enrichment analyses in disease ontology (DO), Kyoto Encyclopedia of Genes and Genomes (KEGG), Gene Ontology (GO) on target genes, significant values ($q < 0.05$) were generated by GO, and genes were enriched at developmental processes, morphogenesis and leukemia (Fig. 3). We detected *Epha4*, *Pax3*, *Wnt6* from high confident SE-Gs around perturbed TADs. This is in line with the previous study which experimentally validated distal interactions between enhancers and these three genes (*Epha4*, *Pax3*, *Wnt6*) after structural variations, including DELs rewired TAD structures and causing limb malformations [14]. We also identified upstream and downstream enhancer regions of *MYC* via DELs interrupted TADs from high confident SE-Gs, which correlates with the study on T cell acute lymphoblastic leukemia [37].

As we have gathered a relatively comprehensive list of DELs and connected enhancers and affected target genes via DEL-TADs, we further investigate the DELs in noncoding regions. Within all DELs we collected, 41% of DELs did not overlap with any known genes, where direct impacts on these DELs are unknown. Through our high confident SE-Gs, we located 22,576 DELs of this kind. We explored the potential biological implicants on these DELs by studying the target genes where these DELs were found. In the enrichment tests, target genes were also significantly enriched ($q < 0.05$) for developmental processes (Fig. 3). Specifically, we observed that these genes were enriched for several developmental processes on this subset of high confident SE-G pairs. In conclusion, these results all support the role of DELs in developmental processes and embryonic developments.

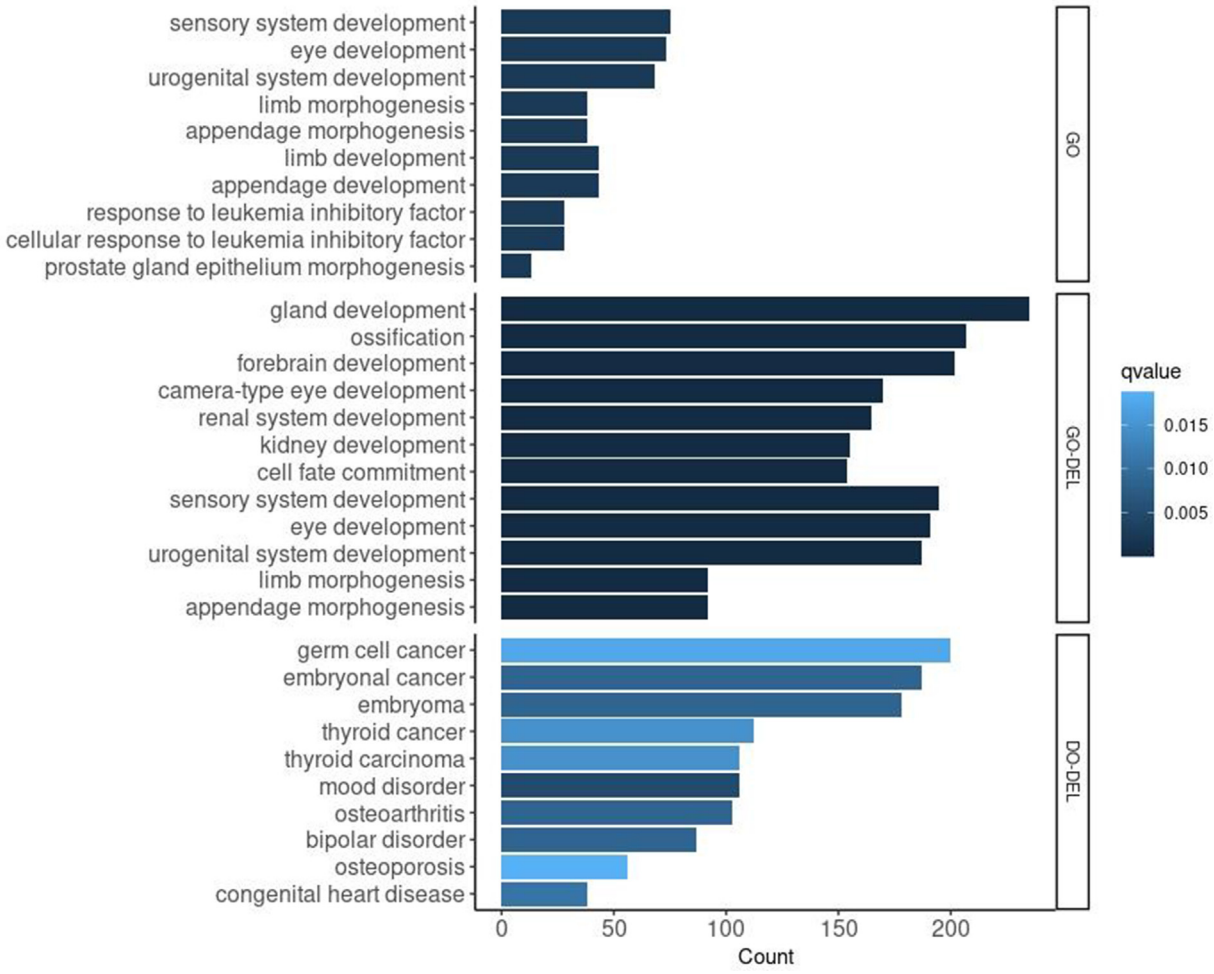

Fig. 3. GO and DO enrichment analyses on different sets of high confident target genes. Upper panel: the enrichment result of gene ontology (GO) in biological process was performed on genes from high confident SE-Gs. Middle and bottom panels: enrichment results of gene ontology (GO-DEL) in biological process and disease ontology (DO-DEL) were carried out by a subset of genes from high confident SE-Gs after considering DELs in noncoding regions. The color represents the FDR value, the y axis shows top 10 significant categories from each ontology, and X-axis represents the number of genes.

3.4. GWAS-SNPs with enhancer confident scores are suggestive of known enhancers

To evaluate the performance of enhancer confident scores, we computed the AUC by comparing GWAS-SNPs with enhancer confident scores and VISTA documented enhancers. By gradually changing the threshold of enhancer confident scores, a series of sensitivity and specificity were computed and these values were used to plot a receiver operating characteristic curve (ROC). The AUC was computed accordingly. Comparison between enhancer confident scores and VISTA gave an AUC of ROC curve of 0.767 (Fig. 4). The result indicated that the enhancer confident score is effective in identifying enhancers. The best performance was reached at the threshold of 0.557, where the specificity was 0.69 and the sensitivity was 0.73.

3.5. Identified SE-Gs are found from the manually curated database and the experimental data

To illustrate whether SE-Gs correlated with previous work, we first compared our results with manually curated data in DiseaseEnhancer database. There were 6,595 out of 2,639,858 GWAS-SNPs covered by 81.47% (598/734) enhancer regions documented at DiseaseEnhancer. We further examined if both GWAS-SNPs and their target genes fell into the enhancers and target gene regions, respectively. In total, 782 SE-Gs were identified, and 33 pairs remained after applying the cut-offs of enhancer confident score and DEL-TAD score to 0.557 and 2, respectively.

To further evaluate the validity of SE-Gs predicted by our method, we took one external omics dataset [36]. Given that our hypothesis is focusing on genes next to TAD borders and version differences in naming SNPs, we cleaned the data from Jung *et al.* by filtering 7,583 genes and 11,268 SNPs. According to our criterion that GWAS-SNPs, DELs and genes must be present on both sides around the TAD border, it occurred that 6,707 pairs from Jung's result were also removed. Finally, we compared the SE-Gs between two datasets using Pearson's Chi-squared test and the high confident SE-Gs were significantly enriched in pcHi-C data ($p = 0.22 \times 10^{-15}$).

4. Discussion

Identification and interpretations of causal variants and affected genes are an enduring challenge in GWAS. Thus, we developed a scoring system focusing on downstream functional dissection of noncoding GWAS-SNPs in three-dimensional context. We compared GWAS-SNPs with enhancer confident scores and SE-Gs to public datasets, which have led to significant results. Moreover,

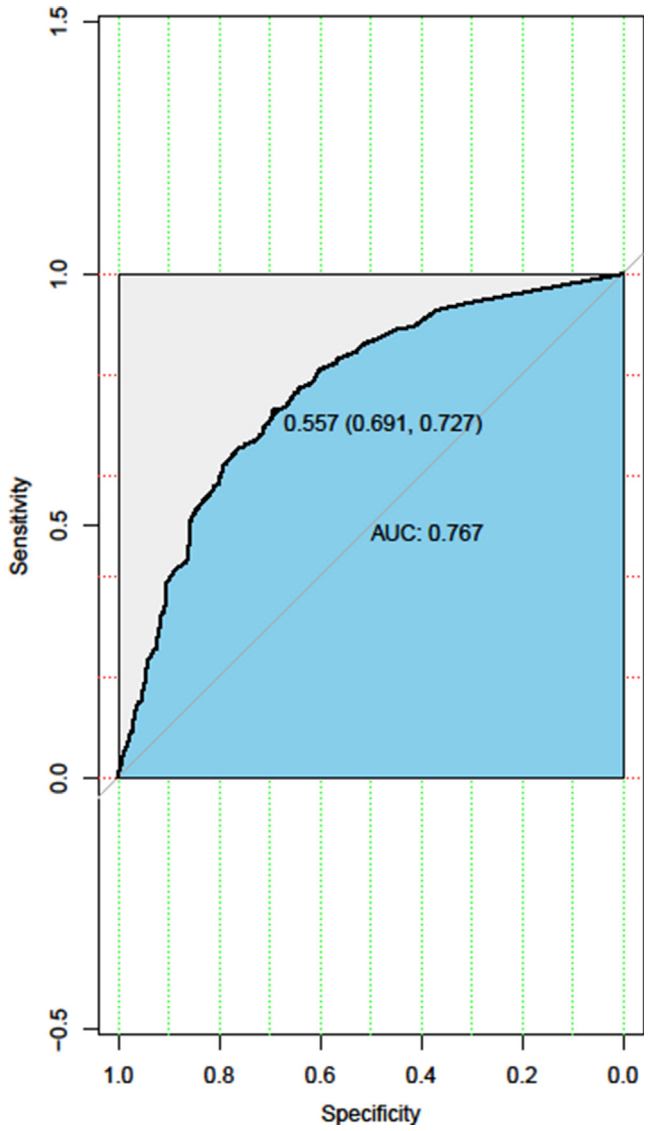

Fig. 4. An AUC of ROC curve between enhancer confident scores and VISTA. The x axis is specificity and y axis represents sensitivity. The AUC is 0.767. At the threshold of 0.557, the best performance is reached where specificity is 0.691 and sensitivity is 0.727.

to our knowledge this is the first attempt in leveraging noncoding GWAS findings with target genes by DELs perturbed TADs.

By integrating DELs, TADs with GWAS-SNPs, we identified 201,132 high confident SE-Gs pairs that play roles in a "long-range" manner. Furthermore, our work on the analysis of high confident SE-Gs uncovered that target genes were enriched in several developmental processes, leukemia and morphogenesis in line with previous studies that explored both structural changes and long range interactions [14,37–40]. Our results could also be extended to explaining DELs that devoid genes, since direct impact on these DELs are difficult to interpret. In total, we detected 22,576 high confident SE-Gs by means of this kind of DELs. "Enhancer hijacking" is a known event in cancer which is sensitive to perturbations. Our study has shown that *MYC* was enriched in several types of cancers where formation of neo-TADs may be involved in *MYC* activation as described by Dixon *et al.* [41].

Although our purpose is to generate consensus results, expanding our analyses to various types of structural variations, cell lines and developmental periods could aid the prioritization of critical regulatory regulatory regions and affected genes. For example, Javierre *et al.* has revealed the cell type specific enhancer-promoter contacts [42]. This is definitely warranted for an important follow-up. Next, we focused on mapping noncoding GWAS-SNPs to target genes in DEL-TADs, however genome-wide studies under this hypothesis are not available to this date, therefore direct assessment on such interactions are challenging. Follow-up experiments, such as reporter assays and chromatin immunoprecipitation sequencing (ChIP-Seq), will be helpful to validate the interactions between enhancers and target genes.

In conclusion, we performed a systematic inference on noncoding regions via GWAS-SNPs and DEL-TADs to boost GWAS discovery power. Our work can be used to locate the functional GWAS-SNPs as well as to uncover affected candidate genes. Moreover, with the rapid development in genome sequencing technologies, our work can also be extended to interpret DELs in noncoding regions. The high confident SE-Gs provide valuable resources to elucidate the biological insights behind complex diseases with emphasis on three-dimensional genome.

Funding

This work was partially supported by grants from the Ministry of Science and Technology of China (2016YFC1000306), the National Natural Science Foundation of China (31830054), the Beijing Municipal Health Commission (JingYiYan 2018-5).

CRediT authorship contribution statement

Xuanshi Liu: Conceptualization, Methodology, Writing - original draft. **Wenjian Xu:** Validation, Writing - review & editing. **Fei Leng:** Writing - review & editing. **Chanjuan Hao:** Supervision, Writing - review & editing. **Sree Rohit Raj Kolora:** Investigation, Writing - review & editing. **Wei Li:** Conceptualization, Supervision, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2020.10.014.

References

[1] Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucl Acids Res 2019;47(D1): D1005-D12.
[2] Kamat MA, Blackshaw JA, Young R, Surendran P, Burgess S, Danesh J, et al. PhenoScanner V2: an expanded tool for searching human genotype-phenotype associations. Bioinformatics 2019;35(22):4851–3.
[3] Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, Shafer A, Neri F, Lee K, Kutyavin T, Stehling-Sun S, Johnson AK, Canfield TK, Giste E, Diegel M, Bates D, Hansen RS, Neph S, Sabo PJ, Heimfeld S, Raubitschek A, Ziegler S, Cotsapas C, Sotoodehnia N, Glass I, Sunyaev SR, Kaul R, Stamatoyannopoulos JA. Systematic localization of common disease-associated variation in regulatory DNA. Science 2012;337 (6099):1190–5. https://doi.org/10.1126/science.1222794.
[4] Wu Y, Zheng Z, Visscher PM, Yang J. Quantifying the mapping precision of genome-wide association studies using whole-genome sequencing data. Genome Biol 2017;18(1). https://doi.org/10.1186/s13059-017-1216-0.
[5] Wu Y, Zeng J, Zhang F, Zhu Z, Qi T, Zheng Z, Lloyd-Jones LR, Marioni RE, Martin NG, Montgomery GW, Deary IJ, Wray NR, Visscher PM, McRae AF, Yang J. Integrative analysis of omics summary data reveals putative mechanisms

underlying complex traits. Nat Commun 2018;9(1). https://doi.org/10.1038/s41467-018-03371-0.
[6] Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang Li, Issner R, Coyne M, Ku M, Durham T, Kellis M, Bernstein BE. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 2011;473(7345):43–9. https://doi.org/10.1038/nature09906.
[7] Smemo S, Tena JJ, Kim K-H, Gamazon ER, Sakabe NJ, Gómez-Marín C, Aneas I, Credidio FL, Sobreira DR, Wasserman NF, Lee JH, Puviindran V, Tam D, Shen M, Son JE, Vakili NA, Sung H-K, Naranjo S, Acemel RD, Manzanares M, Nagy A, Cox NJ, Hui C-C, Gomez-Skarmeta JL, Nóbrega MA. Obesity-associated variants within FTO form long-range functional connections with IRX3. Nature 2014;507(7492):371–5. https://doi.org/10.1038/nature13138.
[8] Won H, de la Torre-Ubieta L, Stein JL, Parikshak NN, Huang J, Opland CK, Gandal MJ, Sutton GJ, Hormozdiari F, Lu D, Lee C, Eskin E, Voineagu I, Ernst J, Geschwind DH. Chromosome conformation elucidates regulatory relationships in developing human brain. Nature 2016;538(7626):523–7. https://doi.org/10.1038/nature19847.
[9] Xia Q, Chesi A, Manduchi E, Johnston BT, Lu S, Leonard ME, Parlin UW, Rappaport EF, Huang P, Wells AD, Blobel GA, Johnson ME, Grant SFA. The type 2 diabetes presumed causal variant within TCF7L2 resides in an element that controls the expression of ACSL5. Diabetologia 2016;59(11):2360–8. https://doi.org/10.1007/s00125-016-4077-2.
[10] Brodie A, Azaria JR, Ofran Y. How far from the SNP may the causative genes be?. Nucl Acids Res 2016;44(13):6046–54. https://doi.org/10.1093/nar/gkw500.
[11] Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 2012;485(7398):376–80. https://doi.org/10.1038/nature11082.
[12] Schmitt A, Hu M, Jung I, Xu Z, Qiu Y, Tan C, Li Y, Lin S, Lin Y, Barr C, Ren B. A compendium of chromatin contact maps reveals spatially active regions in the human genome. Cell Rep 2016;17(8):2042–59. https://doi.org/10.1016/j.celrep.2016.10.061.
[13] Weischenfeldt J, Dubash T, Drainas AP, Mardin BR, Chen Y, Stutz AM, et al. Pan-cancer analysis of somatic copy-number alterations implicates IRS4 and IGF2 in enhancer hijacking. Nat Genet 2017;49(1):65–74.
[14] Lupiáñez Darío G, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, Horn D, Kayserili H, Opitz J, Laxova R, Santos-Simarro F, Gilbert-Dussardier B, Wittler L, Borschiwer M, Haas S, Osterwalder M, Franke M, Timmermann B, Hecht J, Spielmann M, Visel A, Mundlos S. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell 2015;161(5):1012–25. https://doi.org/10.1016/j.cell.2015.04.004.
[15] Li R, Liu Y, Li T, Li C. 3Disease Browser: a Web server for integrating 3D genome and disease-associated chromosome rearrangement data. Sci Rep 2016;6(1). https://doi.org/10.1038/srep34651.
[16] Huang D, Yi X, Zhang S, Zheng Z, Wang P, Xuan C, et al. GWAS4D: multidimensional analysis of context-specific regulatory variant for human complex diseases and traits. Nucl Acids Res 2018;46(W1):W114–W20.
[17] 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, Bagger FO, Jørgensen M, Andersen PR, Bertin N, Rackham O, Burroughs AM, Baillie JK, Ishizu Y, Shimizu Y, Furuhata E, Maeda S, Negishi Y, Mungall CJ, Meehan TF, Lassmann T, Itoh M, Kawaji H, Kondo N, Kawai J, Lennartsson A, Daub CO, Heutink P, Hume DA, Jensen TH, Suzuki H, Hayashizaki Y, Müller F, Forrest ARR, Carninci P, Rehli M, Sandelin A. An atlas of active enhancers across human cell types and tissues. Nature 2014;507(7493):455–61. https://doi.org/10.1038/nature12787.
[18] Ernst J, Kellis M. Chromatin-state discovery and genome annotation with ChromHMM. Nat Protoc 2017;12(12):2478–92. https://doi.org/10.1038/nprot.2017.124.
[19] Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Hsi-Yang Fritz M, Konkel MK, Malhotra A, Stütz AM, Shi X, Paolo Casale F, Chen J, Hormozdiari F, Dayama G, Chen K, Malig M, Chaisson MJP, Walter K, Meiers S, Kashin S, Garrison E, Auton A, Lam HYK, Jasmine Mu X, Alkan C, Antaki D, Bae T, Cerveira E, Chines P, Chong Z, Clarke L, Dal E, Ding Li, Emery S, Fan X, Gujral M, Kahveci F, Kidd JM, Kong Yu, Lameijer E-W, McCarthy S, Flicek P, Gibbs RA, Marth G, Mason CE, Menelaou A, Muzny DM, Nelson BJ, Noor A, Parrish NF, Pendleton M, Quitadamo A, Raeder B, Schadt EE, Romanovitch M, Schlattl A, Sebra R, Shabalin AA, Untergasser A, Walker JA, Wang M, Yu F, Zhang C, Zhang J, Zheng-Bradley X, Zhou W, Zichner T, Sebat J, Batzer MA, McCarroll SA, Mills RE, Gerstein MB, Bashir A, Stegle O, Devine SE, Lee C, Eichler EE, Korbel JO. An integrated map of structural variation in 2,504 human genomes. Nature 2015;526(7571):75–81. https://doi.org/10.1038/nature15394.
[20] Audano PA, Sulovari A, Graves-Lindsay TA, Cantsilieris S, Sorensen M, Welch AE, Dougherty ML, Nelson BJ, Shah A, Dutcher SK, Warren WC, Magrini V, McGrath SD, Li YI, Wilson RK, Eichler EE. Characterizing the major structural variant alleles of the human genome. Cell 2019;176(3):663–675.e19. https://doi.org/10.1016/j.cell.2018.12.019.
[21] Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, Gardner EJ, Rodriguez OL, Guo Li, Collins RL, Fan X, Wen J, Handsaker RE, Fairley S, Kronenberg ZN, Kong X, Hormozdiari F, Lee D, Wenger AM, Hastie AR, Antaki D, Anantharaman T, Audano PA, Brand H, Cantsilieris S, Cao H, Cerveira E, Chen C, Chen X, Chin C-S, Chong Z, Chuang NT, Lambert CC, Church DM, Clarke L, Farrell A, Flores J, Galeev T, Gorkin DU, Gujral M, Guryev V, Heaton WH, Korlach J, Kumar S, Kwon JY, Lam ET, Lee JE, Lee J, Lee W-P, Lee SP, Li S, Marks P, Viaud-Martinez K, Meiers S, Munson KM, Navarro FCP, Nelson BJ, Nodzak C, Noor A, Kyriazopoulou-Panagiotopoulou S, Pang AWC, Qiu Y, Rosanio G, Ryan M, Stütz A, Spierings DCJ, Ward A, Welch AE, Xiao M, Xu W, Zhang C, Zhu Q, Zheng-Bradley X, Lowy E, Yakneen S, McCarroll S, Jun G, Ding Li, Koh CL, Ren B, Flicek P, Chen K, Gerstein MB, Kwok P-Y, Lansdorp PM, Marth GT, Sebat J, Shi X, Bashir A, Ye K, Devine SE, Talkowski ME, Mills RE, Marschall T, Korbel JO, Eichler EE, Lee C. Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat Commun 2019;10(1). https://doi.org/10.1038/s41467-018-08148-z.
[22] Collins RL, Brand H, Karczewski KJ, Zhao X, Alföldi J, Khera AV, et al. An open resource of structural variation for medical and population genetics. bioRxiv 2019;578674.
[23] Boomsma DI, Wijmenga C, Slagboom EP, Swertz MA, Karssen LC, Abdellaoui A, Ye K, Guryev V, Vermaat M, van Dijk F, Francioli LC, Hottenga JJ, Laros JFJ, Li Q, Li Y, Cao H, Chen R, Du Y, Li N, Cao S, van Setten J, Menelaou A, Pulit SL, Hehir-Kwa JY, Beekman M, Elbers CC, Byelas H, de Craen AJM, Deelen P, Dijkstra M, den Dunnen JT, de Knijff P, Houwing-Duistermaat J, Koval V, Estrada K, Hofman A, Kanterakis A, Enckevort DV, Mai H, Kattenberg M, van Leeuwen EM, Neerincx PBT, Oostra B, Rivadeneira F, Suchiman EHD, Uitterlinden AG, Willemsen G, Wolffenbuttel BH, Wang J, de Bakker PIW, van Ommen G-J, van Duijn CM. The genome of the Netherlands: design, and project goals. Eur J Hum Genet 2014;22(2):221–7. https://doi.org/10.1038/ejhg.2013.118.
[24] Hunt SE, McLaren W, Gil L, Thormann A, Schuilenburg H, Sheppard D, et al. Ensembl variation resources. Database 2018;2018.
[25] Dayem Ullah AZ, Oscanoa J, Wang J, Nagano A, Lemoine NR, Chelala C. SNPnexus: assessing the functional relevance of genetic variation to facilitate the promise of precision medicine. Nucl Acids Res 2018;46(W1):W109–W13.
[26] Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 2014;46(3):310–5. https://doi.org/10.1038/ng.2892.
[27] Ritchie GRS, Dunham I, Zeggini E, Flicek P. Functional annotation of noncoding sequence variants. Nat Methods 2014;11(3):294–6. https://doi.org/10.1038/nmeth.2832.
[28] Gulko B, Hubisz MJ, Gronau I, Siepel A. A method for calculating probabilities of fitness consequences for point mutations across the human genome. Nat Genet 2015;47(3):276–83. https://doi.org/10.1038/ng.3196.
[29] Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning–based sequence model. Nat Methods 2015;12(10):931–4. https://doi.org/10.1038/nmeth.3547.
[30] Ionita-Laza I, McCallum K, Xu B, Buxbaum JD. A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nat Genet 2016;48(2):214–20. https://doi.org/10.1038/ng.3477.
[31] Fu Y, Liu Z, Lou S, Bedford J, Mu XJ, Yip KY, Khurana E, Gerstein M. FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. Genome Biol 2014;15(10). https://doi.org/10.1186/s13059-014-0480-5.
[32] Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, Day IN, et al. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics 2015;31(10):1536–43.
[33] Smedley D, Schubach M, Jacobsen JB, Köhler S, Zemojtel T, Spielmann M, Jäger M, Hochheiser H, Washington N, McMurry J, Haendel M, Mungall C, Lewis S, Groza T, Valentini G, Robinson P. A whole-genome analysis framework for effective identification of pathogenic regulatory variants in Mendelian disease. Am J Hum Genet 2016;99(3):595–606. https://doi.org/10.1016/j.ajhg.2016.07.005.
[34] Visel A, Minovitsky S, Dubchak I, Pennacchio LA. VISTA enhancer Browser–a database of tissue-specific human enhancers. Nucl Acids Res 2007;35(Database):D88–92. https://doi.org/10.1093/nar/gkl822.
[35] Zhang G, Shi J, Zhu S, Lan Y, Xu L, Yuan H, et al. DiseaseEnhancer: a resource of human disease-associated enhancer catalog. Nucl Acids Res 2018;46(D1):D78–D84.
[36] Jung I, Schmitt A, Diao Y, Lee AJ, Liu T, Yang D, et al. A compendium of promoter-centered long-range chromatin interactions in the human genome. Nat Genet 2019;51(10):1442–9.
[37] Kloetgen A, Thandapani P, Ntziachristos P, Ghebrechristos Y, Nomikou S, Lazaris C, et al. Three-dimensional chromatin landscapes in T cell acute lymphoblastic leukemia. Nat Genet 2020;52(4):388–400.
[38] Giorgio E, Robyr D, Spielmann M, Ferrero E, Di Gregorio E, Imperiale D, et al. A large genomic deletion leads to enhancer adoption by the lamin B1 gene: a second path to autosomal dominant adult-onset demyelinating leukodystrophy (ADLD). Hum Mol Genet 2015;24(11):3143–54.
[39] Flöttmann R, Wagner J, Kobus K, Curry CJ, Savarirayan R, Nishimura G, Yasui N, Spranger J, Van Esch H, Lyons MJ, DuPont BR, Dwivedi A, Klopocki E, Horn D, Mundlos S, Spielmann M. Microdeletions on 6p22.3 are associated with mesomelic dysplasia Savarirayan type. J Med Genet 2015;52(7):476–83. https://doi.org/10.1136/jmedgenet-2015-103108.
[40] Ibn-Salem J, Köhler S, Love MI, Chung H-R, Huang Ni, Hurles ME, Haendel M, Washington NL, Smedley D, Mungall CJ, Lewis SE, Ott C-E, Bauer S, Schofield PN, Mundlos S, Spielmann M, Robinson PN. Deletions of chromosomal regulatory boundaries are associated with congenital disease. Genome Biol 2014;15(9). https://doi.org/10.1186/s13059-014-0423-1.
[41] Dixon JR, Xu J, Dileep V, Zhan Ye, Song F, Le VT, Yardımcı GG, Chakraborty A, Bann DV, Wang Y, Clark R, Zhang L, Yang H, Liu T, Iyyanki S, An L, Pool C, Sasaki T, Rivera-Mulia JC, Ozadam H, Lajoie BR, Kaul R, Buckley M, Lee K, Diegel M, Pezic D, Ernst C, Hadjur S, Odom DT, Stamatoyannopoulos JA, Broach JR, Hardison RC, Ay F, Noble WS, Dekker J, Gilbert DM, Yue F. Integrative detection

and analysis of structural variation in cancer genomes. Nat Genet 2018;50 (10):1388–98. https://doi.org/10.1038/s41588-018-0195-8.

[42] Javierre BM, Burren OS, Wilder SP, Kreuzhuber R, Hill SM, Sewitz S, Cairns J, Wingett SW, Várnai C, Thiecke MJ, Burden F, Farrow S, Cutler AJ, Rehnström K, Downes K, Grassi L, Kostadima M, Freire-Pritchett P, Wang F, Stunnenberg HG, Todd JA, Zerbino DR, Stegle O, Ouwehand WH, Frontini M, Wallace C, Spivakov M, Fraser P, Martens JH, Kim B, Sharifi N, Janssen-Megens EM, Yaspo M-L, Linser M, Kovacsovics A, Clarke L, Richardson D, Datta A, Flicek P. Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. Cell 2016;167(5):1369–1384.e19. https://doi.org/10.1016/j.cell.2016.09.037.