Published online 26 October 2016 *Nucleic Acids Research, 2017, Vol. 45, Database issue* ***D643–D649***
doi: 10.1093/nar/gkw1022

3DSNP: a database for linking human noncoding SNPs to their three-dimensional interacting genes

Yiming Lu[1,†], Cheng Quan[1,†], Hebing Chen[1,*], Xiaochen Bo[1,*] and Chenggang Zhang[1,*]

[1]Beijing Institute of Radiation Medicine, State Key Laboratory of Proteomics, Beijing 100850, China

Received August 11, 2016; Revised October 16, 2016; Editorial Decision October 17, 2016; Accepted October 18, 2016

ABSTRACT

The vast noncoding portion of the human genome harbors a rich array of functional elements and disease-causing regulatory variants. Recent high-throughput chromosome conformation capture studies have outlined the principles of these elements interacting and regulating the expression of distal target genes through three-dimensional (3D) chromatin looping. Here we present 3DSNP, an integrated database for annotating human noncoding variants by exploring their roles in the distal interactions between genes and regulatory elements. 3DSNP integrates 3D chromatin interactions, local chromatin signatures in different cell types and linkage disequilibrium (LD) information from the 1000 Genomes Project. 3DSNP provides informative visualization tools to display the integrated local and 3D chromatin signatures and the genetic associations among variants. Data from different functional categories are integrated in a scoring system that quantitatively measures the functionality of SNPs to help select important variants from a large pool. 3DSNP is a valuable resource for the annotation of human noncoding genome sequence and investigating the impact of noncoding variants on clinical phenotypes. The 3DSNP database is available at http://biotech.bmi.ac.cn/3dsnp/.

INTRODUCTION

The vast majority of sequence variants in the genome occur outside of coding regions (1,2). Mutations in coding regions are annotated as different types based on the conserved sequence of protein-coding genes and amino acid changes, however, the relationships between noncoding variants and genes are not straightforward. Efforts by the ENCODE (3) and Roadmap Epigenomics (4) projects as well as individual research groups (5–8) have revealed the landscape of regulatory elements across the human genome. Mapping variants to the whole genome showed that disease-associated single nucleotide polymorphisms (SNPs) are strongly enriched in regulatory elements, especially those activated in relevant cell types (9). However, because most regulatory elements are widely dispersed across the genome, interpreting the effects of noncoding variants at on the regulation process of target genes is a great challenge.

The rapid advances of chromosome conformation capture (3C)-based technologies such as 5C (10), Hi-C (11–13) and ChIA-PET (14,15) are providing increasing data on the 3D architecture of the genomes. 5C and Hi-C identify protein-independent chromatin looping and measure 3D genome organization, while ChIA-PET identifies protein-mediated looping and gives information about the role of proteins in structuring 3D organization. Recent studies based on these technologies have revealed the models of DNA elements regulating the expression of distal target genes through 3D chromatin interactions. Scattered elements such as enhancers, insulators and protein-binding sites are tethered to the promoter regions of genes through chromatin looping to facilitate gene transcription. Rao *et al.* found that chromatin loops identified by Hi-C frequently link promoters and enhancers, correlate with gene expression, and are conserved across cell types (12). Another study showed that transcriptionally active genes frequently contacted enhancer-like elements, whereas transcriptionally inactive genes interacted with potential long-range silencers marked by repressive features. They also showed that the interacting loci were enriched for disease-associated SNPs, suggesting distal mutations may disrupt the regulation of relevant genes (16). In addition, principles of 3D chromatin structuring have also been revealed based on the patterns of chromatin modifications and the distance between the pair of loop anchors (12,17–19). Handoko *et al.* found that chromatin domains demarcated by chromatin loops could be clustered into five types with distinct histone signatures and observed a clear transition of the histone patterns that differentiate active signals from inactive signals at ∼200 kb (17). A transition of gene expression profile was also ob-

*To whom correspondence should be addressed. Tel: +86 10 66931590; Fax: +86 10 68169574; Email: zhangcglab@gmail.com
Correspondence may also be addressed to Hebing Chen. Tel: +86 10 66930242; Email: chb-1012@163.com
Correspondence may also be addressed to Xiaochen Bo. Tel: +86 10 66930242; Email: boxiaoc@163.com
†These authors contributed equally to the paper as first authors.

served at similar distance of promoter-enhancer interactions in our recent study on the multivalent roles of CTCF in 3D chromatin looping (19).

A number of databases and computational tools have been developed for annotating noncoding SNPs based on their local genomic contexts. RegulomeDB (20) database focuses on interpreting the one-dimensional (1D) chromatin signatures, including predicted chromatin states (21), histone modifications, DNase I hypersensitive sites (DHSs) and transcription factor binding sites (TFBSs). HaploReg (22,23) integrates the local chromatin signatures with LD information to facilitate the annotation of noncoding variants. A computational tool GWAS3D (24) includes chromosome interactions as an important local signature together with other local signatures and LD data to reprioritize genetic variants, however, the linkages between noncoding variants and genes haven't been established. FunSeq2 (25) attempts to link regulatory variants with potential target genes by correlating epigenetic modifications with gene expression levels, however, one major limitation of this method is that the target gene prediction can only be made within 1Mb from the regulatory elements so that some distal interactions may be missed. CCSI (26) database collected a large set of 3D chromatin interactions in human, mouse and yeast, and can provide a list of 3D interacting genomic regions, the embedded enhancers and SNPs when given a gene name or a genomic region. However, missing a series of important genomic features including chromatin states, histone modifications, TFBS, TF motif, eQTL and conservation greatly limits the application of this database. Therefore, integrating 3D chromatin interactions and 1D chromatin signatures is essential for linking noncoding regulatory SNPs to target genes.

Here, we provide a database, 3DSNP, which comprehensively annotates the regulatory function of human noncoding SNPs by exploring their 3D interactions with genes mediated by chromatin loops. 3DSNP integrates sequential and genotyping data, 3D chromatin interactions, phenotyping data and a variety of chromatin signatures across a broad range of cell types. 3DSNP provides a series of informative tables, publish-ready figures and a comprehensive scoring system to help researchers discover the regulatory roles of noncoding variants upon their 1D and 3D genomic features.

DATA COLLECTION AND PROCESSING

Sequential and genotyping data

Sequential and genotyping information, including chromosome position, Reference/Alternative alleles and minor allele frequency (MAF), of 149 254 102 SNPs and small INDELs were obtained from NCBI dbSNP build 146. Pairwise LD was calculated for each pair of SNPs within a 200 kb window from the 1000 Genomes Project phase 3 data using VCFTools (27) and PLINK (28), and associated SNPs were determined using a LD threshold of $r^2 = 0.8$ in each of the five super populations (African: AFR, Ad Mixed American: AMR, East Asian: ASN, European: EUR and South Asian: SAS). Gene annotations were obtained from GRCh37/hg19 version of RefSeq genes from the UCSC Genome Browser (29), and the relative position of each variant to the closest annotated gene was determined by overlapping the variant with the 2 kb upstream to 2 kb downstream of the gene using BEDTools (30). The data sources of 3DSNP are shown in Figure 1.

Three-dimensional chromatin interactions

Hi-C technology was chosen as the main source for inferring 3D chromatin interactions in 3DSNP, because it detects 3D genome organization at a truly genome-wide scale and the detection is independent of any specific protein factor. We collected and curated Hi-C datasets from several genome-wide chromosome conformation studies (12,13,31,32). Chromatin loops already available in these studies (12,13,32) were directly used with the finest resolution. For H1 and IMR90 datasets, in which chromatin loops were not available, we used the HOMER pipeline (33) to call chromatin loops with a resolution of 10 kb and False Discovery Rate (FDR) < 0.01. In total, we collected 75,362 high-confidence chromatin loops in 12 human cell types. According to the recently developed models of 3D chromatin structuring, we divided chromatin loops into two different types. For a chromatin loop spans shorter than 200 kb, the corresponding interaction type is 'Within loop', where regulatory variants and genes whose transcription start sites (TSSs) located within can interact with each other. For a chromatin loop longer than 200 kb, the type is 'Anchor-to-anchor', where only regulatory variants and genes whose TSS located at the two anchors are supposed to interact with each other.

One-dimensional chromatin signatures

A variety of local chromatin signatures were used to annotate the regulatory functions of SNPs, including predicted chromatin states, histone modifications, DNase I hypersensitivity sites (DHSs) and TFBSs. Chromatin states were predicted in 127 Roadmap cell lines using the core 15-state ChromHMM model trained on a core set of five histone marks (H3K4me3, H3K4me1, H3K36me3, H3K27me3, H3K9me3) in 60 epigenomes. The 15-state model was used because it can capture all the key chromatin states with sufficiently distinct chromatin marks. States corresponding to enhancers or promoters were assigned to SNPs by intersecting intervals by coordinate. The binding sites of 153 transcription factors (TFs) in 91 human cell lines and DHSs in 125 cell lines were obtained from ENCODE and intersected with SNPs for annotation.

Genome sequence signatures

Two types of sequence signatures were used to annotate regulatory SNPs: TF binding motifs altered by SNPs and sequence conservation. To annotate SNPs by their effects on TF binding motifs, we collected a total of 1207 position weight matrices (PWMs) of TFs from the TRANSFAC (34) and JASPAR (35) databases and used the TFM-Scan software (36) to locate the putative binding motifs of TFs by scanning the two strands of the genomic DNA with these PWMs. A stringent threshold of P-value < e^{-12} (6.1E−06)

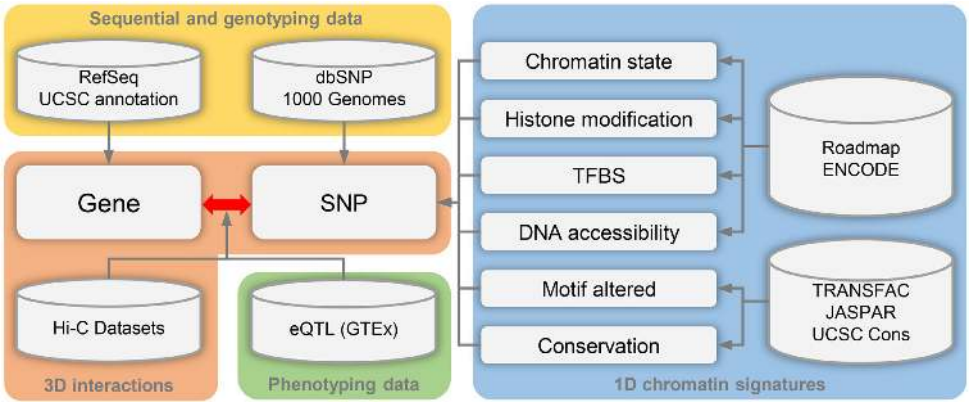

Figure 1. The data sources of 3DSNP (see the description of data collection and processing for details).

was used to determine significantly matched motifs, which were then overlapped with SNPs using BEDTools. The conservation of SNPs was measured by two PhyloP scores (37) obtained from the UCSC Genome Browser. The two PhyloP scores were calculated from multiple alignments of 46 vertebrate genomes and 33 mammal genomes respectively. The absolute values of the PhyloP scores represent −log(*P*-values) under a null hypothesis of neutral evolution, and sites predicted to be conserved are assigned positive scores, while sites predicted to be fast-evolving are assigned negative scores.

Expression quantitative trait loci (eQTLs)

Correlations between genotype and tissue-specific gene expression levels will help annotate the effects of genetic variants on gene regulation. We collected a total of 19 582 729 significant SNP-gene pairs (FDR ≤ 0.05) in 44 human tissues from the GTEx project version 6 (38). To measure the significance of the eQTLs, nominal eQTL p-values and the effect size were obtained for each SNP-gene pair. Nominal eQTL *P*-values were generated using a two-tailed t test, testing the alternative hypothesis that the beta deviates from the null hypothesis of $\beta = 0$. The effect size of the eQTLs is defined as the slope ('β') of the linear regression, and is computed as the effect of the alternative allele (ALT) relative to the reference allele (REF) in the human genome.

DATABASE FEATURES

Scoring system

In 3DSNP, each SNP is scored based on its annotated records on six functional categories: 3D interacting genes, enhancer state, promoter state, transcription factor binding sites, sequence motifs altered and conservation score. Different from the scoring scheme of RegulomeBD, which classifies SNPs into classes based on the combinatorial presence/absence status of functional categories, 3DSNP adopts a quantitative scoring system to evaluate the functional significance of a SNP in different categories. For the first five categories, we used the number of annotated records (hits) to assign score to a SNP in the corresponding category. Specifically, we fitted the numbers of hits of all SNPs in each chromosome to a Poisson distribution model. Considering a SNP has k hits in one functional category F, λ is the fitted parameter of the corresponding Poisson model, then the score of the SNP in this category is defined as follows:

$$Score_F = -\log_{10}\left(\int_k^{+\infty} \frac{\lambda^k e^{-\lambda}}{k!}\right)$$

For the conservation category (F'), we used the PhyloP scores in 33 mammal genomes to assign the conservation scores to SNPs. We found the PhyloP scores of all SNPs in a chromosome follow a Gaussian distribution. Considering a SNP has a conservation score of c, μ and σ are the fitted parameters of the corresponding Gaussian model, then the score of the SNP in the conservation category is defined as follows:

$$Score_{F'} = -\log_{10}\left(\int_c^{+\infty} \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\right)$$

The total score of a SNP is the sum of scores of the six functional categories.

Visualization

We used a regional LD plot to visualize the associated SNPs of a specific SNP in LD. In this plot, the x-axis shows chromosome coordinates, y-axis shows values for r^2, the size of the node represents its total score, and associated SNPs in five populations are shown in different colors, as shown in Figure 2. Associated SNPs in each of the five populations can be removed from or added to the plot by clicking the corresponding circle in the legend. Users can also restrict the range of total score for displaying by adjusting the upper and lower bound of size bar at the right side of the plot. A detailed page will be opened by clicking the node of the corresponding SNP. The regional LD plots can be displayed in the browser and can also be downloaded as high quality, publication-ready PNG files.

To visualize chromosome interactions among noncoding variants, distal regulatory elements and gene promoters, 3DSNP provides both circular and linear plots of chromosome interactions and epigenetic signatures. We first used the Circos (39) software to dynamically generate cir-

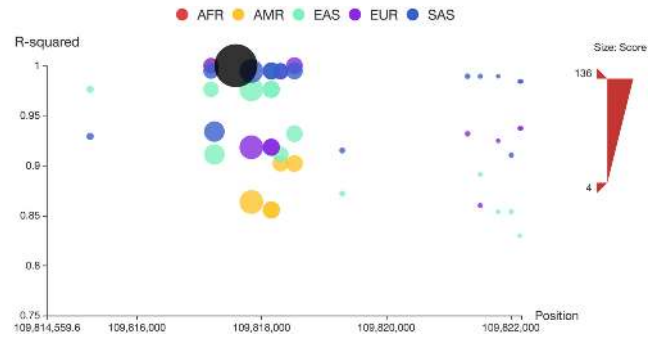

Figure 2. The regional LD plot of associated SNPs of rs12740374. In the plot, x-axis shows chromosome coordinates, y-axis shows values for r^2, the size of the node represents its total score, and associated SNPs in five populations (AFR: African, AMR: Ad Mixed American, ASN: East Asian, EUR: European and SAS: South Asian) are displayed in different colors, and rs12740374 is displayed in black.

cular plots according to the selected cell type and chromatin features by users. As shown in Figure 3A, from outer to inner, the circle represents chromatin states, annotated genes, histone modification set (red), transcription factor set (blue), current SNP and associated SNPs, and 3D chromatin interactions, respectively. The mapping data of chromatin states, DHS and six important histone modifications (H3K4me1, H3K4me3, H3K9me3, H3K27ac, H3K27me3 and H3K36me3) in 127 cell types obtained from the Roadmap Epigenomics project and the binding sites of 153 TFs in 91 cell lines obtained from the ENCODE project are available for visualizing by Circos in 3DSNP. Current SNP (labeled by red) and its top 20 significantly associated SNPs ($r^2 > 0.8$) in LD in all five populations are displayed in the SNP circle. Chromatin interactions in 12 cell types are displayed using colored ribbon links in the central region of the plot. For long-range interactions spanning >1 Mb, the chromosome is cropped by removing irrelevant genomic region. In addition to Circos, 3DSNP also provides a linear plot of chromatin interactions and signatures using UCSC genome browser. As shown in Figure 3B, from top to bottom, the tracks are: genomic coordinates, chromatin interactions, current SNP, UCSC genes, RefSeq genes, histone modifications, CTCF binding sites, DNase Clusters and mammal conservation. Chromatin interactions and SNP tracks are added to the plot using custom track of the genomic coordinates in BED format. All these plots are displayed in the browser and can be downloaded as high quality, publication-ready PNG files.

Web interface

3DSNP provides a series of user-friendly interfaces for the users to search, browse, visualize and export the results. A search box is available on the home page, which can be used for fast searching and browsing. The system accepts three types of input: a set of dbSNP IDs, a genomic region written as 'chrN:start-end' or an annotated gene written as 'gene:SYMBOL'. Users can also upload a text file containing a set of SNPs or genomic regions by clicking the upload icon at the right side of the search box. A dynamic table below the search box displays the summarized information about the important features of SNPs, including total score, 3D interacting genes, linear closest genes, Enhancer/Promoter states, TFBS and Motifs. For the last four features, the numbers of records are displayed for brevity. The summary table can be copied to clipboard or be exported to files in 'Excel' or 'PDF' formats. By clicking the '+' sign at the beginning of each row, users can get a set of SNPs associated to the corresponding SNP in LD in five populations. A table containing the total score, r^2 and D' information and a regional LD plot of SNPs are displayed in the expanded area. By clicking on the SNP ID in these tables, a new page containing all detailed information on the corresponding SNP will be opened. The detailed page is divided into different sections, and a navigation bar on the right side of the web interface can be used to quickly navigate to any required section. These sections are 'Overview', 'Sequential information', 'Visualization', 'Linear closest gene', '3D interacting gene', 'eQTL', '3D interacting SNP', 'Chromatin state', 'TFBS', 'Sequence Motif' and 'Conservation'. Most of the sections contain informative tables and figures displaying the details of the SNP in the corresponding functional category. For the conciseness of the details page, a functional category will only be shown if the SNP has one or more records in the category. The 'Overview' section contains a text describes the scores of six functional categories and the total score of the SNP. A radar chart at the right side of the text can intuitively display the functional tendency of a SNP among six functional categories.

Application program interface (API)

3DSNP provides a more powerful way for users to access the data through the use of API. 3DSNP API enables different applications to access a set of functional categories of request SNPs with programmable interfaces. The API is defined as a HTTP URL request message (http://biotech.bmi.ac.cn/3dsnp/api.do), along with a set of parameters, including: request data type, response message format and request functional category. Two types of request data are supported: SNP IDs and a genomic region. Response messages can be returned in two formats on request: XML or JSON. Example API requests could be defined as: http://biotech.bmi.ac.cn/3dsnp/api.do?id=rs1000&format=json&type=3dgene or http://biotech.bmi.ac.cn/3dsnp/api.do?position=1000000-1000100&chrom=chr1&format=xml&type=3dgene,eqtl.

Usage

To illustrate the usage of 3DSNP, we search the database with a well-studied noncoding SNP at the chromosome 1p13 locus, rs12740374, associated with the risk of myocardial infarction (40). From the summary table below the search box, we can see that rs12740374 is assigned with a high total score of 134.74, interacts with *PSRC1* and other seven genes in 3D, locates in enhancer state in 53 cell lines and promoter state in 20 cell lines, and overlaps with 69 TFBSs. The associated SNPs of rs12740374 in LD can be seen by clicking the '+' sign at the beginning of the corresponding row. The total score, pairwise r^2 and D' of associated SNPs in five populations are listed in a table, and

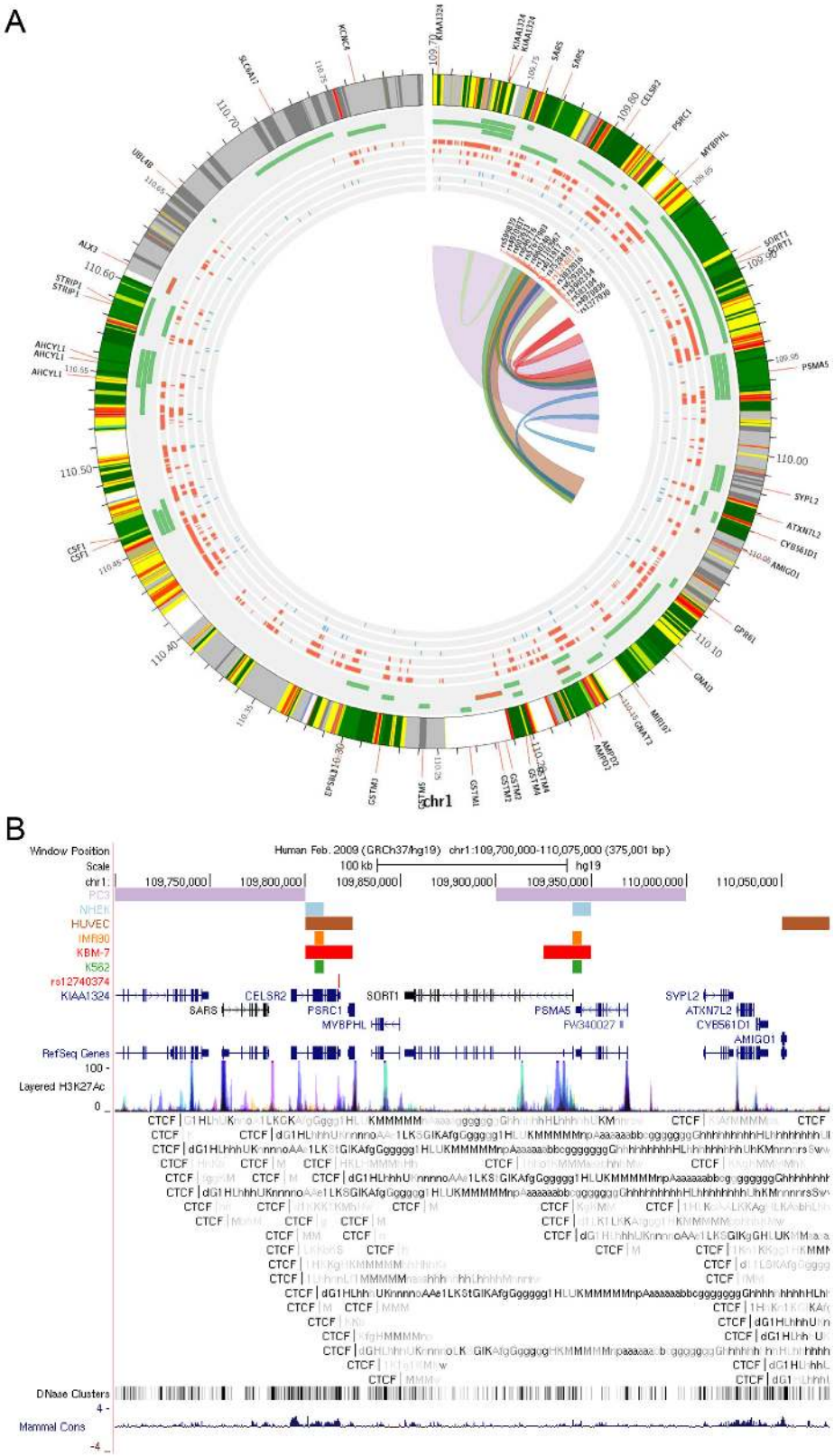

Figure 3. The circular plot (**A**) and linear plot (**B**) of chromosome interactions and epigenetic signatures in HepG2 cells related to rs12740374. In the circular plot, from outer to inner, the circle represents chromatin states, annotated genes, histone modification set (red), transcription factor set (blue), current SNP and associated SNPs, and 3D chromatin interactions, respectively. The three circles in the histone modification set are H3K4me1, H3K4me3, H3K27ac, and the three circles in the TFs set are CTCF, CEBPB and CEBPD. In the linear plot, from top to bottom, the tracks are: genomic coordinates, chromatin interactions, current SNP, UCSC genes, RefSeq genes, histone modifications, CTCF binding sites, DNase Clusters and mammal conservation.

an integrated regional LD plot is also displayed, as shown in Figure 2. In the detailed page of rs12740374, we can see the total score of this variant is mainly attributed to three functional categories: TFBS (86.23), Enhancer (32.13) and Promoter (12.64) in the 'Overview' section. In the 'Visualization' section, the local chromatin signatures and 3D chromatin interactions are displayed in circular and linear plots, and the cell type of the circular plot can be set to HepG2, as shown in Figure 3A,B. In the 'eQTL' section, we can see rs12740374 is significantly correlated with the expression levels of *SORT1* in liver. In the 'TFBS' section, we can see rs12740374 locates at the binding sites of CEBPB and CEBPD in HepG2, IMR90 and HeLa-S3 cells with high DNA accessibilities (1000/1000). These results are highly consistent with a previous study on this noncoding locus (40) reporting that rs12740374 creates a C/EBP (CCAAT/enhancer binding protein) TFBS and alters the hepatic expression of the *SORT1* gene. More importantly, we can see from both the circular and linear plots and the '3D interacting SNP' section, that rs12740374 interacts with the *SORT1* gene mediated by chromatin loops in five different cell types: KBM-7, NHEK, IMR90, K562 and PC3, strongly suggesting that the relationship between rs12740374 and *SORT1* is mediated by 3D chromatin looping.

DISCUSSION

3D chromatin interactions are crucial for decoding the roles of DNA regulatory elements and the embedded SNPs. 3DSNP takes advantage of the rapid development of the Hi-C technology to annotate noncoding variants. Despite that a number of Hi-C studies have been carried out and some important principles on 3D genome have been uncovered, the 3D chromatin architectures in most human cell lines are still unclear. 3DSNP contains all currently available Hi-C datasets, and we will keep updating the database with new Hi-C datasets. In addition, we notice that computational methods for inferring 3D chromatin interactions from 1D epigenetic data have been recently developed (41–44). These algorithms may provide a new data source for 3D chromatin structures, since the 1D epigenomes are available in a wide range of cell lines.

FUNDING

National Basic Research Project (973 program) [2012CB518200] (http://www.973.gov.cn/); General Program [31401141, 81573251] and Major Research plan [U1435222] of the Natural Science Foundation of China (www.nsfc.gov.cn); State Key Laboratory of Proteomics of China [SKLP-Y201303, SKLP-O201104, SKLP- K201004] (www.bprc.ac.cn); Special Key Programs for Science and Technology of China [2012ZX09102301-016]; Program of International S&T Cooperation (2014DFB30020); National High Technology Research and Development Program of China (2015AA020108). Funding for open access charge: General Program [31401141, 81573251] of the Natural Science Foundation of China.
Conflict of interest statement. None declared.

REFERENCES

1. Hindorff,L.A., Sethupathy,P., Junkins,H.A., Ramos,E.M., Mehta,J.P., Collins,F.S. and Manolio,T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 9362–9367.
2. 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.
3. Consortium,E.P. (2004) The ENCODE (ENCyclopedia Of DNA Elements) project. *Science*, **306**, 636–640.
4. Bernstein,B.E., Stamatoyannopoulos,J.A., Costello,J.F., Ren,B., Milosavljevic,A., Meissner,A., Kellis,M., Marra,M.A., Beaudet,A.L., Ecker,J.R. *et al.* (2010) The NIH roadmap epigenomics mapping consortium. *Nat. Biotechnol.*, **28**, 1045–1048.
5. Heintzman,N.D., Stuart,R.K., Hon,G., Fu,Y., Ching,C.W., Hawkins,R.D., Barrera,L.O., Van Calcar,S., Qu,C., Ching,K.A. *et al.* (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.*, **39**, 311–318.
6. Visel,A., Blow,M.J., Li,Z., Zhang,T., Akiyama,J.A., Holt,A., Plajzer-Frick,I., Shoukry,M., Wright,C., Chen,F. *et al.* (2009) ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, **457**, 854–858.
7. Heintzman,N.D., Hon,G.C., Hawkins,R.D., Kheradpour,P., Stark,A., Harp,L.F., Ye,Z., Lee,L.K., Stuart,R.K., Ching,C.W. *et al.* (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, **459**, 108–112.
8. Lu,Y., Qu,W., Shan,G. and Zhang,C. (2015) DELTA: a distal enhancer locating tool based on AdaBoost algorithm and shape features of chromatin modifications. *PLoS One*, **10**, e0130622.
9. Ernst,J., Kheradpour,P., Mikkelsen,T.S., Shoresh,N., Ward,L.D., Epstein,C.B., Zhang,X., Wang,L., Issner,R., Coyne,M. *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43–49.
10. Sanyal,A., Lajoie,B.R., Jain,G. and Dekker,J. (2012) The long-range interaction landscape of gene promoters. *Nature*, **489**, 109–113.
11. 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.
12. Rao,S.S., Huntley,M.H., Durand,N.C., Stamenova,E.K., Bochkov,I.D., Robinson,J.T., Sanborn,A.L., Machol,I., Omer,A.D., Lander,E.S. *et al.* (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**, 1665–1680.
13. 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.
14. Fullwood,M.J. and Ruan,Y. (2009) ChIP-based methods for the identification of long-range chromatin interactions. *J. Cell. Biochem.*, **107**, 30–39.
15. Fullwood,M.J., Wei,C.L., Liu,E.T. and Ruan,Y. (2009) Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome Res.*, **19**, 521–532.
16. 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.
17. Handoko,L., Xu,H., Li,G., Ngan,C.Y., Chew,E., Schnapp,M., Lee,C.W., Ye,C., Ping,J.L., Mulawadi,F. *et al.* (2011) CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat. Genet.*, **43**, 630–638.
18. Guo,Y., Xu,Q., Canzio,D., Shou,J., Li,J., Gorkin,D.U., Jung,I., Wu,H., Zhai,Y., Tang,Y. *et al.* (2015) CRISPR inversion of CTCF sites alters genome topology and enhancer/promoter function. *Cell*, **162**, 900–910.
19. Lu,Y., Shan,G., Xue,J., Chen,C. and Zhang,C. (2016) Defining the multivalent functions of CTCF from chromatin state and

three-dimensional chromatin interactions. *Nucleic Acids Res.*, **44**, 6200–6212.
20. Boyle,A.P., Hong,E.L., Hariharan,M., Cheng,Y., Schaub,M.A., Kasowski,M., Karczewski,K.J., Park,J., Hitz,B.C., Weng,S. *et al.* (2012) Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.*, **22**, 1790–1797.
21. Ernst,J. and Kellis,M. (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods*, **9**, 215–216.
22. Ward,L.D. and Kellis,M. (2016) HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic Acids Res.*, **44**, D877–D881.
23. Ward,L.D. and Kellis,M. (2012) HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.*, **40**, D930–D934.
24. Li,M.J., Wang,L.Y., Xia,Z., Sham,P.C. and Wang,J. (2013) GWAS3D: Detecting human regulatory variants by integrative analysis of genome-wide associations, chromosome interactions and histone modifications. *Nucleic Acids Res.*, **41**, W150–W158.
25. Fu,Y., Liu,Z., Lou,S., Bedford,J., Mu,X.J., Yip,K.Y., Khurana,E. and Gerstein,M. (2014) FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol.*, **15**, 480.
26. Xie,X., Ma,W., Songyang,Z., Luo,Z., Huang,J., Dai,Z. and Xiong,Y. (2016) CCSI: a database providing chromatin-chromatin spatial interaction information. *Database*, **2016**, bav124.
27. Danecek,P., Auton,A., Abecasis,G., Albers,C.A., Banks,E., DePristo,M.A., Handsaker,R.E., Lunter,G., Marth,G.T., Sherry,S.T. *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.
28. Purcell,S., Neale,B., Todd-Brown,K., Thomas,L., Ferreira,M.A., Bender,D., Maller,J., Sklar,P., de Bakker,P.I., Daly,M.J. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.
29. Speir,M.L., Zweig,A.S., Rosenbloom,K.R., Raney,B.J., Paten,B., Nejad,P., Lee,B.T., Learned,K., Karolchik,D., Hinrichs,A.S. *et al.* (2016) The UCSC Genome Browser database: 2016 update. *Nucleic Acids Res.*, **44**, D717–D725.
30. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
31. Sanborn,A.L., Rao,S.S., Huang,S.C., Durand,N.C., Huntley,M.H., Jewett,A.I., Bochkov,I.D., Chinnappan,D., Cutkosky,A., Li,J. *et al.* (2015) Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, E6456–E6465.
32. Taberlay,P.C., Achinger-Kawecka,J., Lun,A.T., 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.
33. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.
34. Matys,V., Fricke,E., Geffers,R., Gossling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E., Kel-Margoulis,O.V. *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
35. Mathelier,A., Fornes,O., Arenillas,D.J., Chen,C.Y., Denay,G., Lee,J., Shi,W., Shyr,C., Tan,G., Worsley-Hunt,R. *et al.* (2016) JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **44**, D110–D115.
36. Liefooghe,A., Touzet,H. and Varre,J.S. (2006) Large scale matching for position weight matrices. *Lect. Notes Comput. Sci.*, **4009**, 401–412.
37. Siepel,A., Bejerano,G., Pedersen,J.S., Hinrichs,A.S., Hou,M., Rosenbloom,K., Clawson,H., Spieth,J., Hillier,L.W., Richards,S. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.
38. Consortium,G.T. (2013) The genotype-tissue expression (GTEx) project. *Nat. Genet.*, **45**, 580–585.
39. Krzywinski,M., Schein,J., Birol,I., Connors,J., Gascoyne,R., Horsman,D., Jones,S.J. and Marra,M.A. (2009) Circos: an information aesthetic for comparative genomics. *Genome Res.*, **19**, 1639–1645.
40. 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.
41. Mourad,R. and Cuvier,O. (2015) Predicting the spatial organization of chromosomes using epigenetic data. *Genome Biol.*, **16**, 182.
42. Zhu,Y., Chen,Z., Zhang,K., Wang,M., Medovoy,D., Whitaker,J.W., Ding,B., Li,N., Zheng,L. and Wang,W. (2016) Constructing 3D interaction maps from 1D epigenomes. *Nat. Commun.*, **7**, 10812.
43. Chen,Y., Wang,Y., Xuan,Z., Chen,M. and Zhang,M.Q. (2016) De novo deciphering three-dimensional chromatin interaction and topological domains by wavelet transformation of epigenetic profiles. *Nucleic Acids Res.*, **44**, e106.
44. Huang,J., Marco,E., Pinello,L. and Yuan,G.C. (2015) Predicting chromatin organization using histone marks. *Genome Biol.*, **16**, 162.