Bioinformatics, 36(5), 2020, 1649–1651
doi: 10.1093/bioinformatics/btz770
Advance Access Publication Date: 11 October 2019
Applications Note

Databases and ontologies

dbInDel: a database of enhancer-associated insertion and deletion variants by analysis of H3K27ac ChIP-Seq

Moli Huang [1,2], Yunpeng Wang[1], Manqiu Yang[1], Jun Yan[3], Henry Yang[4], Wenzhuo Zhuang[1], Ying Xu[2], H. Phillip Koeffler[4,5], De-Chen Lin[5,*] and Xi Chen[6,*]

[1]Department of Bioinformatics, School of Biology and Basic Medical Sciences and [2]Cambridge-Suda Genomic Research Center, Soochow University, Suzhou 215123, China, [3]MOE Key Laboratory of Model Animal for Disease Study, Nanjing University, Nanjing 210061, China, [4]Cancer Science Institute of Singapore, National University of Singapore, Singapore 119074, Singapore, [5]Department of Medicine, Cedars-Sinai Medical Center, Los Angeles, CA 90048, USA and [6]Department of Biology, Southern University of Science and Technology, Shenzhen 518055, China

*To whom correspondence should be addressed.

Associate Editor: Janet Kelso

Received on April 17, 2019; revised on August 24, 2019; editorial decision on October 4, 2019; accepted on October 8, 2019

Abstract

Summary: Cancer hallmarks rely on its specific transcriptional programs, which are dysregulated by multiple mechanisms, including genomic aberrations in the DNA regulatory regions. Genome-wide association studies have shown many variants are found within putative enhancer elements. To provide insights into the regulatory role of enhancer-associated non-coding variants in cancer epigenome, and to facilitate the identification of functional non-coding mutations, we present dbInDel, a database where we have comprehensively analyzed enhancer-associated insertion and deletion variants for both human and murine samples using ChIP-Seq data. Moreover, we provide the identification and visualization of upstream TF binding motifs in InDel-containing enhancers. Downstream target genes are also predicted and analyzed in the context of cancer biology. The dbInDel database promotes the investigation of functional contributions of non-coding variants in cancer epigenome.

Availability and implementation: The database, dbInDel, can be accessed from http://enhancer-indel.cam-su.org/.

Contact: dchlin11@gmail.com or chenx9@sustech.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Human cancer genome contains millions of non-coding mutations of either germline or somatic origins, including single-nucleotide substitutions, small and large insertions and deletions (Groschel *et al.*, 2014). Since estimation of region-specific background mutational rates in tumor samples can be particularly challenging, it is extremely difficult to systematically and unbiasedly assess the selection pressure of non-coding mutations in a genome-wide manner (Polak *et al.*, 2014; Schuster-Bockler and Lehner, 2012). As a result, to date, only a few non-coding variants have been characterized and the functional contribution of non-coding mutations is underappreciated (Garraway and Lander, 2013; Zhang *et al.*, 2018). Recent genome-wide association studies (GWAS) have shown that many cancer-related variants tend to appear in putative enhancer elements (Sur and Taipale, 2016), which can be systematically investigated via techniques such as ChIP-Seq of histone marks or ATAC-Seq. By focusing on mutations in putative enhancers, one can profoundly reduce surveying space and increase statistical power. In this way, we not only capture bona fide InDels, but also detect low frequency variants in the regulatory regions (Abraham *et al.*, 2017). In addition, we could efficiently link candidate variants to target transcripts. More importantly, we have further developed this method to incorporate transcription factor binding motif analysis, which considerably facilitates the identification of potential upstream transcriptional regulators involved.

Here, we present dbInDel, the first comprehensive and interactive database cataloging enhancer-associated insertion and deletion variants for both human and murine samples. Unlike the COSMIC (Tate *et al.*, 2019), an established database centered on the annotation of existing somatic mutations, our database focuses on the construction of bona fide enhancer-associated InDels captured from H3K27ac ChIP-Seq experiments, leveraging recent advances in alignment algorithms (Supplementary Fig. S1A). To facilitate the analysis of target transcripts of InDel-containing enhancers, we integrate mRNA expression profiles and survival analysis in both tumor and normal samples across all human cancer types (Supplementary Fig. S1B). Moreover, our method identifies the potential recruitment of transcription factors as a result of enhancer-associated InDels, promoting the investigation of functional contributions of this important class of non-coding variants (Supplementary Fig. S1C).

2 Materials and methods and results

To identify small InDels (defined as <50 bp) in putative enhancer elements across a broad spectrum of cancers, we computationally reconstructed the published method described in Abraham *et al.*, 2017 and analyzed H3K27ac ChIP-Seq datasets from over 250 samples representing 26 cancer types. The detailed workflow is summarized in Supplementary Figure S2. Briefly, we first mapped ChIP-Seq reads to the reference genome to establish putative enhancer landscapes. Then, from the unmapped reads of ChIP-Seq, we identified and verified enhancer-associated InDels with multiple alignment procedures. Currently, our database contains 640 432 insertions and 157 554 deletions in 593 655 putative enhancers detected in 275 samples (Supplementary Table S1). Within all the InDels, there are 274 995 unique insertions and 71 603 unique deletions (Supplementary Fig. S3). Notably, our results showed that many well-established cancer type-specific drivers harbored unique enhancer-associated InDels in the respective cancer types. For example, AR, a unique prostate cancer oncogene, was detected to have enhancer-associated InDels exclusively in prostate cancer samples. Similarly, TAL1 had enhancer-associated InDels only in leukemia and lymphoma samples (Supplementary Fig. S4). These data strongly suggest that some of the enhancer-associated InDels detected by our pipeline are under selection pressure, and they confer growth/survival advantages to specific types of cells because of the unique functions of the target genes that they regulate.

We next performed integrative and comprehensive analysis focusing on the predicted downstream genes (Supplementary Fig. S1B). Specifically, in the context of cancer biology, we performed survival analysis, tumor/normal differential expression using TCGA/GTEx data and mRNA expression of target genes in cancer cell lines from CCLE, considering this information is particularly relevant for understanding tumor cell biology. Moreover, putative cancer drivers were predicted using Cancer Gene Consensus of Sanger. Notably, 26 out of 100 genes with the highest numbers of enhancer-associated InDels are categorized as cancer drivers; in stark contrast, none of the genes with the lowest numbers of enhancer-associated InDels are designated as drivers (Supplementary Fig. S5). This finding may partially indicate that a faction of predicted InDels were under positive selection pressure during tumor development. In addition, we provided their potential roles in cancers (Bailey *et al.*, 2018), drug-gene interaction (Cotto *et al.*, 2018) and GWAS publications, further facilitating the analysis of putative cancer genes for our users.

To facilitate exploration of the potential mechanism of InDels on regulating the activity of their hosting enhancers, we incorporated motif analysis to predict the TF binding motifs which are generated by the detected InDels. Briefly, using FIMO we scanned canonic TF binding motifs in each InDel-containing sequence, and compared that with results obtained from the reference genome. Motifs and associated TFs that were specific to InDel-containing sequences were compiled and visualized in our database (Supplementary Fig. S1C). Using this method, we successfully recovered the MYB motif, which is generated by the insertions in the super-enhancer region of TAL1 gene in Jurkat cells, as shown previously (Mansour *et al.*, 2014). More importantly, numerous additional recruitments of potentially important TFs were predicted by our method. A particularly notable case was found in the super-enhancer region downstream of FOXA1 in a prostate cancer cell line 22Rv1 (Shukla *et al.*, 2017). We predicted that a CTT deletion generated a canonical HNF4G motif (Supplementary Fig. S6A, P 5.67e-05; Supplementary Tables S2 and S3). Importantly, in 22Rv1 cells, ChIP-Seq data revealed prominent HNF4G binding precisely in our predicted InDel locus within this FOXA1 super-enhancer. In contrast, in the wild-type cell line LnCaP, no such HNF4G binding was observed and FOXA1 had a much weaker H3K27ac signal (Supplementary Fig. S6A). Even forced overexpression of HNF4G in LnCaP cells did not lead to its occupancy to this FOXA1 enhancer region. Interestingly, the FOXA1 expressions are comparable in this two cell lines (Supplementary Fig. S6B). Supportively, FOXA1 was expressed uniquely high in prostate cancer in a pan-cancer analysis (Supplementary Fig. S6C). High FOXA1 level was also associated with poor prognosis in this cancer, albeit without reaching statistical significance (Supplementary Fig. S6D). Our data thus implies that this CTT deletion introduces a canonical HNF4G motif, which may recruit HNF4G binding to the downstream of FOXA1 locus in 22Rv1 cells (Supplementary Fig. S6E). This indicates the mechanisms causing the high expression of FOXA1 may vary in different samples.

3 Discussion

With the field of epigenome advancing rapidly, non-coding mutations identified directly from epigenomic feature is one of the most investigated area. There exists a great need among cancer researchers to understand the functional significance of these variants. To the best of our knowledge, dbInDel is the first database that allows users to perform interactive analysis of non-coding InDels, upstream TF prediction and enhancer features. In addition, we incorporated many cancer-centric analysis of candidate InDel-associated genes, including differential expression between matched normal and tumor samples, survival analysis, function predictions, drug-gene interactions and potential upstream TF binding analysis. Those features greatly facilitate the investigations by cancer researchers. This would result in the identification of important mutations and fast forward novel therapeutics to target the non-coding genome.

A few databases exist curating non-coding somatic mutations (e.g. COSMIC, ICGC, DoCM, IntOGen) or germline variants (ClinVar) in human cancers. However, none of these resources associates non-coding mutations with distal cis-regulatory elements (such as enhancers or other TF binding sites). Therefore, it is extremely difficult to prioritize the functional potential of the large number of non-coding variants in the cancer genome. Our database directly tackles this problem. In addition, with the ever-increasing number of cancer samples profiled by epigenome sequencing (such as ChIP-Seq and ATAC-Seq), the catalog of non-coding InDel variants associated with functional epigenomic domains will be expanding rapidly and we will continue adding the newly identified variants in our database.

Acknowledgements

The authors would like to thank Richard A. Young for sharing calling InDel variants pipeline to this work. They also thank Kar-tong Tan and HPK lab members for the kind help.

Funding

The work was supported by National Key R&D Program of China (2018YFA0801100), National Natural Science Foundation of China (31971117, 81672873) and the Priority Academic Program Development of Jiangsu Higher Education Institutions.

Conflict of Interest: none declared.

References

Abraham,B.J. *et al.* (2017) Small genomic insertions form enhancers that misregulate oncogenes. *Nat. Commun.*, **8**, 14385.

Bailey,M.H. *et al.* (2018) Comprehensive characterization of cancer driver genes and mutations. *Cell*, **174**, 1034–1035.

Cotto,K.C. *et al.* (2018) DGIdb 3.0: a redesign and expansion of the drug-gene interaction database. *Nucleic Acids Res.*, **46**, D1068–D1073.

Garraway,L.A. and Lander,E.S. (2013) Lessons from the cancer genome. *Cell* **153**, 17–37.

Groschel,S. *et al.* (2014) A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia. *Cell* **157**, 369–381.

Mansour,M.R. *et al.* (2014) Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science* **346**, 1373–1377.

Polak,P. *et al*. (2014) Reduced local mutation density in regulatory DNA of cancer genomes is linked to DNA repair. *Nat. Biotechnol*., **32**, 71–75.

Schuster-Bockler,B. and Lehner,B. (2012) Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* **488**, 504–507.

Shukla,S. *et al*. (2017) Aberrant activation of a gastrointestinal transcriptional circuit in prostate cancer mediates castration resistance. *Cancer Cell*, **32**, 792–806.e7.

Sur,I. and Taipale,J. (2016) The role of enhancers in cancer. *Nat. Rev. Cancer*, **16**, 483–493.

Tate,J.G. *et al*. (2019) COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res*., **47**, D941–D947.

Zhang,W. *et al*. (2018) A global transcriptional network connecting noncoding mutations to changes in tumor gene expression. *Nat. Genet*., **50**, 613–620.