RESEARCH ARTICLE

*e*PIANNO: ePIgenomics ANNOtation tool

Chia-Hsin Liu[1,2,3☯], Bing-Ching Ho[4,5☯], Chun-Ling Chen[1☯], Ya-Hsuan Chang[1], Yi-Chiung Hsu[1], Yu-Cheng Li[1], Shin-Sheng Yuan[1], Yi-Huan Huang[1], Chi-Sheng Chang[5], Ker-Chau Li[1], Hsuan-Yu Chen[1]*

1 Institute of Statistical Science, Academia Sinica, Nangang, Taipei, Taiwan, **2** Bioinformatics Program, Taiwan International Graduate Program, Academia Sinica, Nangang, Taipei, Taiwan, **3** Institute of Biomedical Informatics, National Yang-Ming University, Taipei, Taiwan, **4** Department of Clinical Laboratory Sciences and Medical Biotechnology, College of Medicine, National Taiwan University, Taipei, Taiwan, **5** NTU Center for Genomic Medicine, National Taiwan University College of Medicine, Taipei, Taiwan

☯ These authors contributed equally to this work.
* hychen@stat.sinica.edu.tw

Citation: Liu C-H, Ho B-C, Chen C-L, Chang Y-H, Hsu Y-C, Li Y-C, et al. (2016) *e*PIANNO: ePIgenomics ANNOtation tool. PLoS ONE 11(2): e0148321. doi:10.1371/journal.pone.0148321

Editor: Hodaka Fujii, Osaka University, JAPAN

Received: July 13, 2015

Accepted: January 15, 2016

Published: February 9, 2016

Data Availability Statement: ChIP-seq results were obtained from three sources on the public domain. 1. hmChIP database containing 392 ChIP-seq experiments of human including 129 chromatin modification (CM) and 263 transcription factor binding site (TFBS) experiments (http://jilab.biostat.jhsph.edu/database/cgi-bin/download.pl). 2. ENCODE database containing 1654 ChIP-seq experiments of human including 489 CM and 1165 TFBS experiments (http://genome.ucsc.edu/ENCODE/downloads.html). 3. Roadmap Epigenomics Project containing 1,069 CM experiments (http://genboree.org/EdaccData/Release-9/). Disease-associated SNPs are available from GWAS Catalog of NHGRI (https://www.genome.

Abstract

Recently, with the development of next generation sequencing (NGS), the combination of chromatin immunoprecipitation (ChIP) and NGS, namely ChIP-seq, has become a powerful technique to capture potential genomic binding sites of regulatory factors, histone modifications and chromatin accessible regions. For most researchers, additional information including genomic variations on the TF binding site, allele frequency of variation between different populations, variation associated disease, and other neighbour TF binding sites are essential to generate a proper hypothesis or a meaningful conclusion. Many ChIP-seq datasets had been deposited on the public domain to help researchers make new discoveries. However, researches are often intimidated by the complexity of data structure and largeness of data volume. Such information would be more useful if they could be combined or downloaded with ChIP-seq data. To meet such demands, we built a webtool: ePIgenomic ANNOtation tool (*e*PIANNO, http://epianno.stat.sinica.edu.tw/index.html). *e*PIANNO is a web server that combines SNP information of populations (1000 Genomes Project) and gene-disease association information of GWAS (NHGRI) with ChIP-seq (hmChIP, ENCODE, and ROADMAP epigenomics) data. *e*PIANNO has a user-friendly website interface allowing researchers to explore, navigate, and extract data quickly. We use two examples to demonstrate how users could use functions of *e*PIANNO webserver to explore useful information about TF related genomic variants. Users could use our query functions to search target regions, transcription factors, or annotations. *e*PIANNO may help users to generate hypothesis or explore potential biological functions for their studies.

Introduction

Transcription factors (TFs) and chromatin regulators (CRs) play key roles in eukaryotic gene transcriptional processes[1, 2]. Mapping the genomic locations of these factors is crucial to understand the mechanisms of transcriptional and epigenetic regulation. Recently, with the

gov/26525384). Gene annotation are available from ENCODE (https://www.encodeproject.org/).

Funding: Supported by grants from Academia Sinica, Institute of Statistical Science AS, AS-100-TP-AB2, NSC 98-3112-B-001-034, NSC 99-2314-B-001-003-MY3, NSC 100-2325-B-001-027, NSC 101-2325-B-002-071, NSC 102-2325-B-002-078, NSC 101-2319-B-002-002, NSC 102-2319-B-002 -002, NSC 102-2911-I-002-303, NSC 101-2911-I-002-303, NSC 102-2911-I-002-303, DOH101-TD-B-111-001, 102R7557, NSC 102-2923-B-002-004, NSC 103-2923-B-002-003, MOST 104-2923-B-002-003, and Taiwan Biosignature Project of Lung Cancer. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abbreviations: NGS, next generation sequencing (NGS); ChIP, the combination of chromatin immunoprecipitation; GWASs, genome-wide association studies; TFs, transcription factors; CRs, chromatin regulators; SNPs, single-nucleotide polymorphisms; IBD, inflammatory bowel disease.

development of next generation sequencing (NGS), the combination of chromatin immunoprecipitation (ChIP) and NGS, namely ChIP-seq, has become a powerful technique to capture potential genomic binding sites of regulatory factors, histone modifications and chromatin accessible regions[3, 4]. Many ChIP-seq datasets had been deposited on the public domain to help researchers make new discoveries. However, researches are often intimidated by the complexity of data structure and largeness of data volume.

Currently, many ChIP-seq web servers and databases have been developed to provide solutions for diverse biological issues. For examples, CR Cistrome is a ChIP-seq database for chromatin regulators and histone modification linkages in human and mouse [5] and ChIPBase is a database for decoding the transcriptional regulation of long non-coding RNA and microRNA genes from ChIP-Seq data [6]. NCBI Gene Expression Omnibus (GEO) [7] and Sequence Read Archive (SRA) [8] are the most popular depositories of public domain but they do not provide function for interactively exploring ChIP-seq data. ChIP-X database [9] and Factorbook.org [10] collect targeted genes of TFs from published ChIP-seq data but tools for retrieving data across samples are not supported currently. Several resources provide annotated metadata for ChIP-Seq or DNase-Seq studies such as nuclear receptor CistromeFinder, Cistrome, NCBI Epigenomics, hmChIP, SwissRegulon, and CistromeMap [11–17]. Most of them only support the gene symbol querying method and only reveal background information of samples or studies. Although UCSC browser supported multiple types of query and results are shown in an interactive graph, it currently does not provide any batch query function. In addition, previous studies did provide valuable results and information of ChIP-seq data for researchers [18–20]. However, researchers without computational abilities cannot explore associations between their interesting TFs, binding sites, or genomic variations by themselves.

In this work, we considered two important databases that have been frequently used in the genetic research. The first one was the 1000 Genomes Project which provided an overview of human genomic variation, mainly single-nucleotide polymorphisms (SNPs) and small InDels [21, 22]. A total of 1,092 individuals from 14 populations in four continents of the world were enrolled to this project and phase 1 data was released in 2012. The 1000 Genomes Project reveals the alternative allele frequency (AF) distribution of population. A genomic variant with high AF in human general population is considered as a common variant while one with skewed distribution cross different populations would be inferred as a population specific variant. The second database we considered was Genome-Wide Association Studies (GWAS) Catalog of National Human Genome Research Institute (NHGRI) [23] which mainly provided the SNP-trait association data. A total of 11,912 SNPs and 1,060 diseases from 1,751 published GWAS were enrolled.

For most researchers, additional information including genomic variations on the TF binding site, allele frequency of variation between different populations, variation associated disease, and other neighbour TF binding sites are essential to generate a proper hypothesis or a meaningful conclusion. Such information would be more useful if they could be combined or downloaded with ChIP-seq data. To meet such demands, we build an integrative ChIP-seq webserver: *e*PIANNO (http://epianno.stat.sinica.edu.tw/index.html). *e*PIANNO contained over 1000 samples from over 3000 ChIP-seq experiments of human. *e*PIANNO a had a user-friendly website interface allowing researchers to explore, navigate, and extract data quickly. It provided two query functions including official gene symbol and local region, respectively. User could query and extract peak data across samples and experiments in small regions they interest. Results of each query were annotated by 1000 Genomes Project phase 1 release, GWAS Catalog of NHGRI[23], and annotation of ENCODE[24]. *e*PIANNO may help researches to explore potential biological functions of the TFs in which they were interested, thereby helping them generate hypothesis for future studies.

Materials and Methods

*e*PIANNO was constructed by combining the ChIP-seq data part and the annotation part. In the ChIP-seq data part, data came from three sources. The first one was hmChIP database containing 392 ChIP-seq experiments of human including 129 chromatin modification (CM) and 263 transcription factor binding site (TFBS) experiments (http://jilab.biostat.jhsph.edu/database/cgi-bin/download.pl) [13]. The second was ENCODE database containing 1654 ChIP-seq experiments of human including 489 CM and 1165 TFBS experiments (http://genome.ucsc.edu/ENCODE/downloads.html). The third one was Roadmap Epigenomics Project containing 1,069 CM experiments (http://genboree.org/EdaccData/Release-9/). The peak calling step had been conducted by hmChIP, ENCODE, and ROADMAP Epigenomics, respectively. We only curate the peak files (chromosome, start, and end positions) of ChIP-seq datasets from three projects and provide original sources. In hmChip, datasets were collected from GEO database in BED file format contained chromosome, start, and end positions. However, they stop updating in 2011 and the genome mapping version was hg18. As a consequence, we remapped the BED file provided from hmChip to hg19. In ENCODE, four peak callers, SPP, GEM, PeakSeq, and MACS, were integrated in their uniform peak calling pipeline to generate final peak file. In ROADMAP epigenomics, Pash 3.0 read mapper was used to make reads mapping first. For the histone ChIP-seq data, the MACSv2.0.10 peak caller was used to identify peak region (https://github.com/taoliu/MACS/). For DNase-seq data, the Hotspot algorithm was used to identify fixed-size (150bp) DNase hypersensitive sites and MACSv2.0.10 was used to call peaks using the same settings as the histone mark. Fragment lengths for each dataset were pre-estimated using strand cross-correlation analysis and the SPP peak caller package (https://code.google.com/p/phantompeakqualtools/) and these fragment length estimates were explicitly used as parameters in the MACS2 program (—shift-size = fragment_length/2). Totally, ePIANNO contained potential 253,000,853 peaks in various cell types or experiments. Genomic variant data from 1000 Genomes Project [21], GWAS Catalog of NHGRI[23], and gene annotation of ENCODE[24] were included in the annotation part (Table 1).

Table 1. Databases and data size in the system.

Database Name	Feature	Experiments	Data size
hmChIP	CM[ǂ]	129	3,494,199 peaks
	TF[ǂ]	263	6,235,702 peaks
ENCODE	CM[ǂ]	489	33,690,989 peaks
	TF[ǂ]	1165	33,809,568 peaks
Roadmap Epigenomics	CM[ǂ]	1069	175,770,395 peaks
NHGRI	Number of disease		663
	Number of unique SNP		7,883
1000 Genomes Project	Number of individual		1,092
	Number of SNP		36,820,992
ENCODE annotation	Number of gene		40,686

Databases and data size of the *e*PIANNO. For ChIP-seq data part, peak data from hmChIP, ENCODE, and Roadmap Epigenomics databases were recruited in the system. Totally, the database contained over 253 million peak data in various cell types or experiments. For annotation part, there were 36,820,992 genomic variant data from 1000 Genomes Project, 7,883 unique SNPs and 663 diseases from GWAS Catalog of NHGRI, 40,868 annotations of gene from ENCODE annotation, respectively.

[ǂ] CM: chromatin modification; TF: transcription factor

doi:10.1371/journal.pone.0148321.t001

Results

Most disease-associated SNPs locate in the protein-binding region

Among all potential DNA binding sites in the *e*PIANNO, 2,824,294 (1.11%) potential DNA binding sites had disease-associated SNPs annotated by the GWAS Catalog of NHGRI. Among all disease-associated SNPs annotated by the GWAS Catalog of NHGRI, 7,797 (98.91%) SNPs were found in the potential protein-DNA binding regions (Fig 1A). This high percentage of disease-associated SNPs may imply that the protein-DNA binding event is crucial to reveal the associations between genomic variants and diseases.

We calculate total sites and disease-associated sites covered by all ChIP-seq experiments collected in ePIANNO. Totally 2,668,537,612 bps (86.65%) were covered by Chip-seq experiments based on hg19 genome size (3,079,843,747 bps). This result showed that although the overall lengths of all peaks in ePIANNO was 206 billion bps, most of them (98%) were overlapped with each other. A hypergeometric test showed that disease-associated SNPs were enriched in the ChIP-seq covering regions ($P < 2\text{x}10^{-16}$, Fig 1B). This result is smiliar with a previous analysis study done by Bryzgalov et al. They describe an approach aggregating the whole set of ENCODE ChIP-seq data in order to search for regulatory SNPs and concluded that SNPs located in the protein-DNA region are more likely to influence transcription regulation[19].

Utility and web interface: query and retrieval

Currently, *e*PIANNO provided four functions: searches of transcription factor, target gene, correspondent binding region, and gene annotations. User could search the database by official gene symbol (ex. *EGFR*) or by coordinate (ex. chr7:10000~10000000). When users search specific peaks or genes, both forward and reverse streams of genomic regions will be reported and results will be shown on the browser. *e*PIANNO also provided one-click download function for user to retrieve results in EXCEL CSV format.

Search transcription factors

The first function of *e*PIANNO is the search of TFs. If user wants to know potential TFs that regulate a gene of interest, *e*PIANNO will search TFs whose DNA-binding locations were around the gene of interest based on peak information of ChIP-seq experiments. The first step is to input the specific gene in the query box (Fig 2A). User could choose specific source of experiments (Fig 2B) or a ChIP type (Fig 2C). Basic results including transcription factor/chromatin modification names, location, distance to gene, data source, and sample name will be shown on the same web page (Fig 2D). By clicking the correspondent buttons, a user could expand the detail of results including SNP information in population (Fig 2E), disease related variants (Fig 2F), and peak data of the ChIP-seq experiment (Fig 2G) in peak region. "SNP information in populations" contains information of SNP coordinate, allele call, and minor allele frequency in populations of the four continents. "SNP information in NHGRI" contains SNP-trait association data with correspondent PubMed link from GWAS Catalog of NHGRI. Finally, peak summary shows data of ChIP-seq experiment including peak length, start and end site of peak, and other information in original data source. For example, if a user wants to explore what TFs may regulate *HLA-DQB1*, click "Search Transcription Factors" button at the top-left of the web page and put "*HLA-DQB1*" in the query box. *e*PIANNO will show all TFs binding locations in up or downstream of *HLA-DQB1*. Advance results such as peak data summary and variations could be expanded by clicking the button. In addition, if a user only focuses on specific genomic region instead of gene, *e*PIANNO also shows binding locations and correspondent TFs in the given genomic region. For example, a user can specify genomic

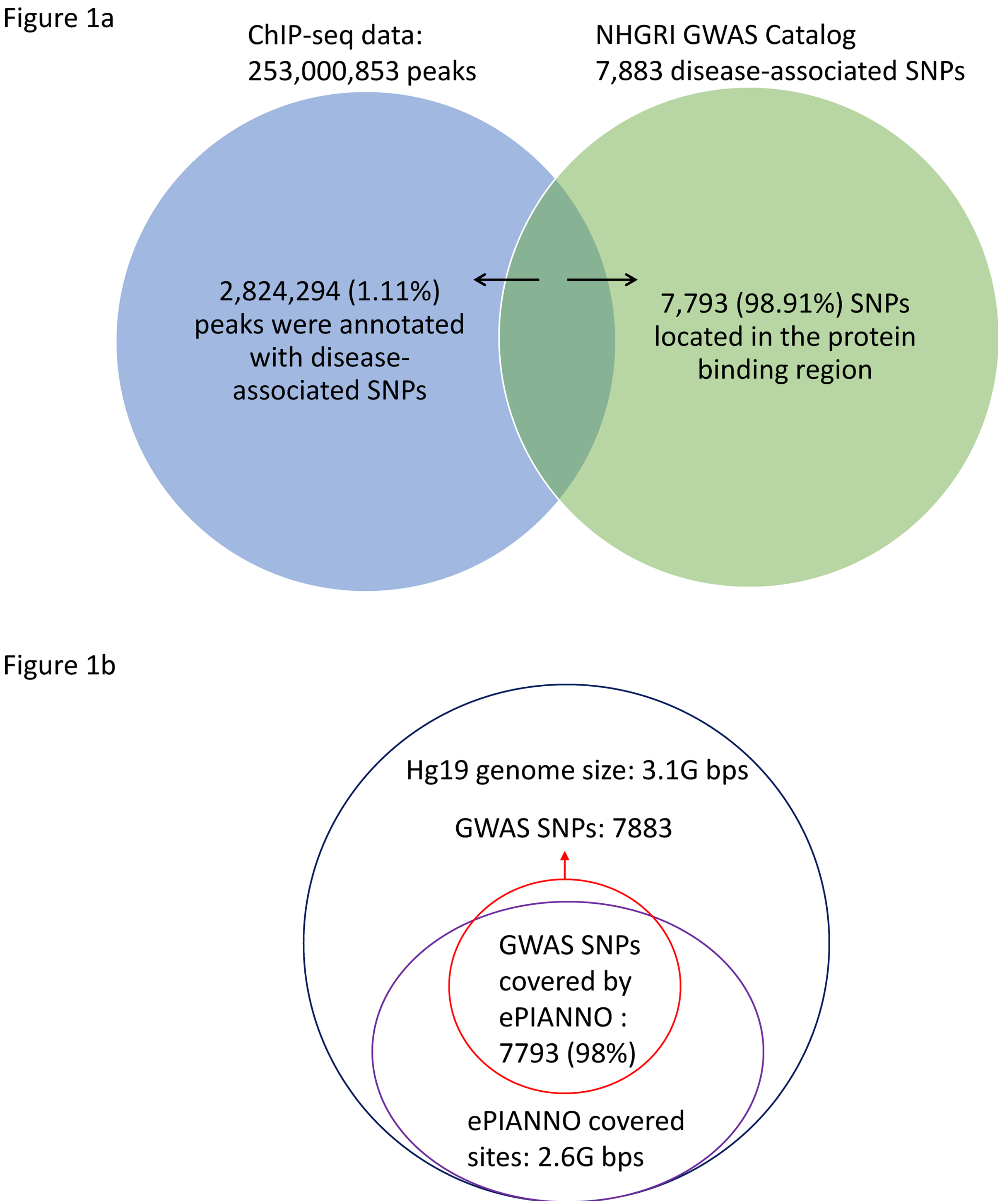
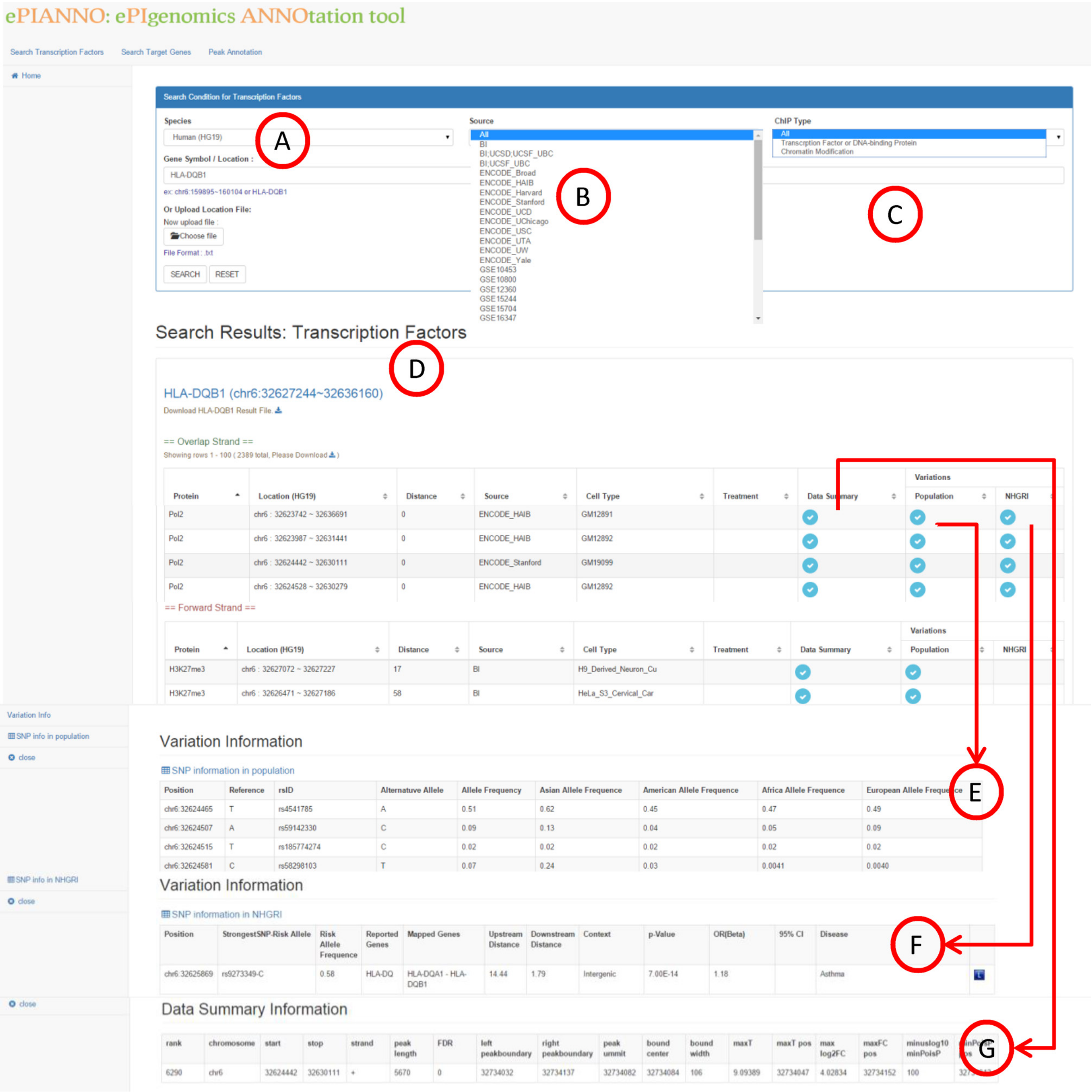


Figure 1a
ChIP-seq data:
253,000,853 peaks
NHGRI GWAS Catalog
7,883 disease-associated SNPs
2,824,294 (1.11%) peaks were annotated with disease-associated SNPs
7,793 (98.91%) SNPs located in the protein binding region
Figure 1b
Hg19 genome size: 3.1G bps
GWAS SNPs: 7883
GWAS SNPs covered by ePIANNO : 7793 (98%)
ePIANNO covered sites: 2.6G bps

Fig 1. Disease-associated SNPs in *e*PIANNO. (a) A total of 7,793 (98.91%) out of 7,883 SNPs were covered by peaks recruited in *e*PIANNO. (b) Totally 2,668,537,612 bps (86.65%) were covered by Chip-seq experiments based on hg19 genome size (3,079,843,747 bps). A hypergeometric test showed that disease-associated SNPs were enriched in the ChIP-seq covering regions (*P* < 2E-16).

doi:10.1371/journal.pone.0148321.g001



Search Condition for Transcription Factors

Species: Human (HG19)

Gene Symbol / Location : HLA-DQB1

ex: chr6:159895~160104 or HLA-DQB1

Or Upload Location File:

Now upload file : Choose file

File Format : .txt

SEARCH RESET

Source: All
BI
BI;UCSD;UCSF_UBC
BI;UCSF_UBC
ENCODE_Broad
ENCODE_HAIB
ENCODE_Harvard
ENCODE_Stanford
ENCODE_UCD
ENCODE_UChicago
ENCODE_USC
ENCODE_UTA
ENCODE_UW
ENCODE_Yale
GSE10453
GSE10800
GSE12360
GSE15244
GSE15704
GSE16347

ChIP Type: All
Transcrption Factor or DNA-binding Protein
Chromatin Modification

Search Results: Transcription Factors

HLA-DQB1 (chr6:32627244~32636160)

Download HLA-DQB1 Result File.

== Overlap Strand ==

Showing rows 1 - 100 (2389 total, Please Download)

							Variations	
Protein	Location (HG19)	Distance	Source	Cell Type	Treatment	Data Summary	Population	NHGRI
Pol2	chr6 : 32623742 ~ 32636691	0	ENCODE_HAIB	GM12891		✓	✓	✓
Pol2	chr6 : 32623987 ~ 32631441	0	ENCODE_HAIB	GM12892		✓	✓	✓
Pol2	chr6 : 32624442 ~ 32630111	0	ENCODE_Stanford	GM19099		✓	✓	✓
Pol2	chr6 : 32624528 ~ 32630279	0	ENCODE_HAIB	GM12892		✓	✓	✓

== Forward Strand ==

							Variations	
Protein	Location (HG19)	Distance	Source	Cell Type	Treatment	Data Summary	Population	NHGRI
H3K27me3	chr6 : 32627072 ~ 32627227	17	BI	H9_Derived_Neuron_Cu		✓	✓	
H3K27me3	chr6 : 32626471 ~ 32627186	58	BI	HeLa_S3_Cervical_Car		✓	✓	

Variation Information

SNP information in population

Position	Reference	rsID	Alternatuve Allele	Allele Frequency	Asian Allele Frequence	American Allele Frequence	Africa Allele Frequence	European Allele Freque ce
chr6:32624465	T	rs4541785	A	0.51	0.62	0.45	0.47	0.49
chr6:32624507	A	rs59142330	C	0.09	0.13	0.04	0.05	0.09
chr6:32624515	T	rs185774274	C	0.02	0.02	0.02	0.02	0.02
chr6:32624581	C	rs58298103	T	0.07	0.24	0.03	0.0041	0.0040

Variation Information

SNP information in NHGRI

Position	StrongestSNP-Risk Allele	Risk Allele Frequence	Reported Genes	Mapped Genes	Upstream Distance	Downstream Distance	Context	p-Value	OR(Beta)	95% CI	Disease	
chr6:32625869	rs9273349-C	0.58	HLA-DQ	HLA-DQA1 - HLA-DQB1	14.44	1.79	Intergenic	7.00E-14	1.18		Asthma	

Data Summary Information

rank	chromosome	start	stop	strand	peak length	FDR	left peakboundary	right peakboundary	peak ummit	bound center	bound width	maxT	maxT pos	max log2FC	maxFC pos	minuslog10 minPoisP	inPoisP os
6290	chr6	32624442	32630111	+	5670	0	32734032	32734137	32734082	32734084	106	9.09389	32734047	4.02834	32734152	100	327[illegible]

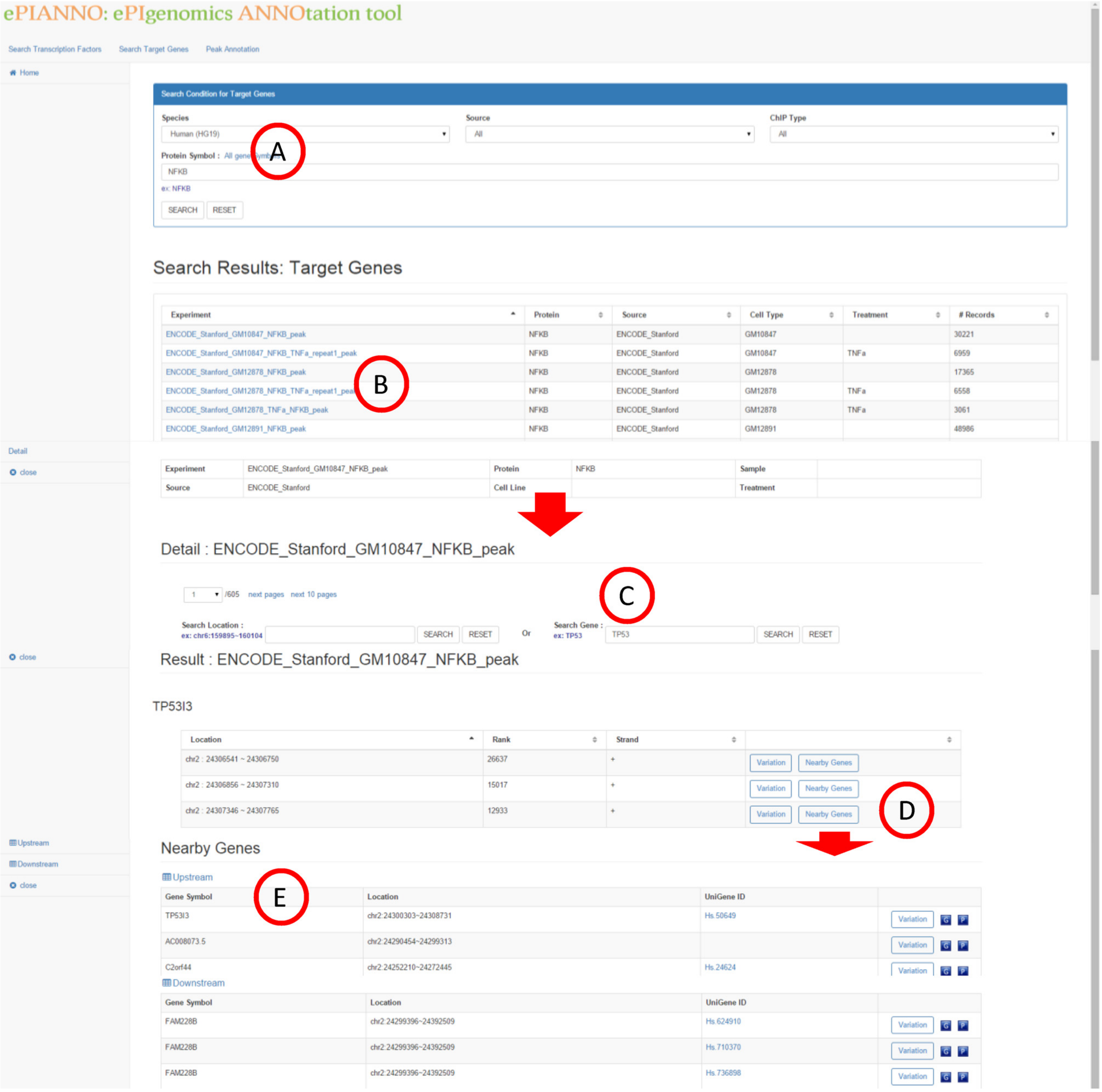
Fig 2. Search transcription factors. If user wants to know potential TFs that regulate gene or genomic region of interest, they could specify gene name, ex., or genomic region, ex. "chr6:159895~160104" or "HLA-DQB1" in the query box (A). User could choose specific source of experiments (B) or ChIP type (C). Results including transcription factor/chromatin modification names, location, distance to gene, data source, and sample name will be shown on the same web page (D). User could expand advanced results by clicking buttons "Variation" for SNP information in populations (E), disease related variants(F), and "Data Summary" for peak data of the ChIP-seq experiment (G).

doi:10.1371/journal.pone.0148321.g002

region, such as "chr6:159895~160104", or multiple regions separated by semi-colon or another line, such as "chr6:159895~160104; chr6:182232~210148", in the query box. *e*PIANNO will show information of TFs, variations, and peak data similar with results mentioned above.

Search target genes

The second function of *e*PIANNO is to search target genes of a TF. Given a TF name of interest (Fig 3A), *e*PIANNO shows DNA-binding locations of this TF based on peak data from ChIP-seq experiments (Fig 3B). A user could find additional information of DNA-binding locations (Fig 3C) through clickable links and the detail information of the location (Fig 3D) will be shown in a new page. However, the candidate binding regions may be over thousands and



Search Condition for Target Genes

Species: Human (HG19)
Source: All
ChIP Type: All
Protein Symbol : All gene symbol
NFKB
ex: NFKB

Search Results: Target Genes

Experiment	Protein	Source	Cell Type	Treatment	# Records
ENCODE_Stanford_GM10847_NFKB_peak	NFKB	ENCODE_Stanford	GM10847		30221
ENCODE_Stanford_GM10847_NFKB_TNFa_repeat1_peak	NFKB	ENCODE_Stanford	GM10847	TNFa	6959
ENCODE_Stanford_GM12878_NFKB_peak	NFKB	ENCODE_Stanford	GM12878		17365
ENCODE_Stanford_GM12878_NFKB_TNFa_repeat1_peak	NFKB	ENCODE_Stanford	GM12878	TNFa	6558
ENCODE_Stanford_GM12878_TNFa_NFKB_peak	NFKB	ENCODE_Stanford	GM12878	TNFa	3061
ENCODE_Stanford_GM12891_NFKB_peak	NFKB	ENCODE_Stanford	GM12891		48986

Experiment	ENCODE_Stanford_GM10847_NFKB_peak	Protein	NFKB	Sample	
Source	ENCODE_Stanford	Cell Line		Treatment	

Detail : ENCODE_Stanford_GM10847_NFKB_peak

1 /605 next pages next 10 pages

Search Location : ex: chr6:159895~160104 Or Search Gene : ex: TP53 TP53

Result : ENCODE_Stanford_GM10847_NFKB_peak

TP53I3

Location	Rank	Strand	
chr2 : 24306541 ~ 24306750	26637	+	Variation Nearby Genes
chr2 : 24306856 ~ 24307310	15017	+	Variation Nearby Genes
chr2 : 24307346 ~ 24307765	12933	+	Variation Nearby Genes

Nearby Genes

Upstream

Gene Symbol	Location	UniGene ID	
TP53I3	chr2:24300303~24308731	Hs.50649	Variation G P
AC008073.5	chr2:24290454~24299313		Variation G P
C2orf44	chr2:24252210~24272445	Hs.24624	Variation G P

Downstream

Gene Symbol	Location	UniGene ID	
FAM228B	chr2:24299396~24392509	Hs.624910	Variation G P
FAM228B	chr2:24299396~24392509	Hs.710370	Variation G P
FAM228B	chr2:24299396~24392509	Hs.736898	Variation G P

Fig 3. Search target genes. Given a TF name of interest in the query box (A), *e*PIANNO shows DNA-binding locations of this TF based on peak data from ChIP-seq experiments (B). Additional information of DNA-binding locations (C) and detail information of the location (D) will be shown in a new window. Two buttons could be further clicked, "Variation" and "Nearby Genes". "Variation" shows all SNPs in the binding location and their association with diseases. "Nearby Genes" shows genes located upstream or downstream of the binding location (E).

doi:10.1371/journal.pone.0148321.g003

further filtering is needed. In this new page, user could further narrow down the candidate list by specifying the target region or gene. Two buttons could be further clicked, "Variation" and "Nearby Genes". "Variation" shows information of all SNPs in the binding location including their association with diseases. "Nearby Genes" shows genes located nearby upstream or downstream of the binding location (Fig 3E). These genes may be potentially regulated by the queried TF (Fig 3A).

For example, inputting the TF name "*NFKB*" in the query box, *e*PIANNO will show all ChIP-seq experiments of *NFKB*. After clicking one of experiment link, candidate binding regions of *NFKB* will be shown in a new page. In this new page, user could further specify a gene or a region of interest to explore whether it was bind by *NFKB* or not. Advanced results of the DNA-binding location showed SNP-disease association information, SNP information of populations, and genes located nearby upstream or downstream of the DNA-binding location. These genes nearby the DNA-binding location of *NFKB* may be potentially regulated by *NFKB*. Furthermore, SNP-disease association information and SNP information of populations of these genes are also shown.

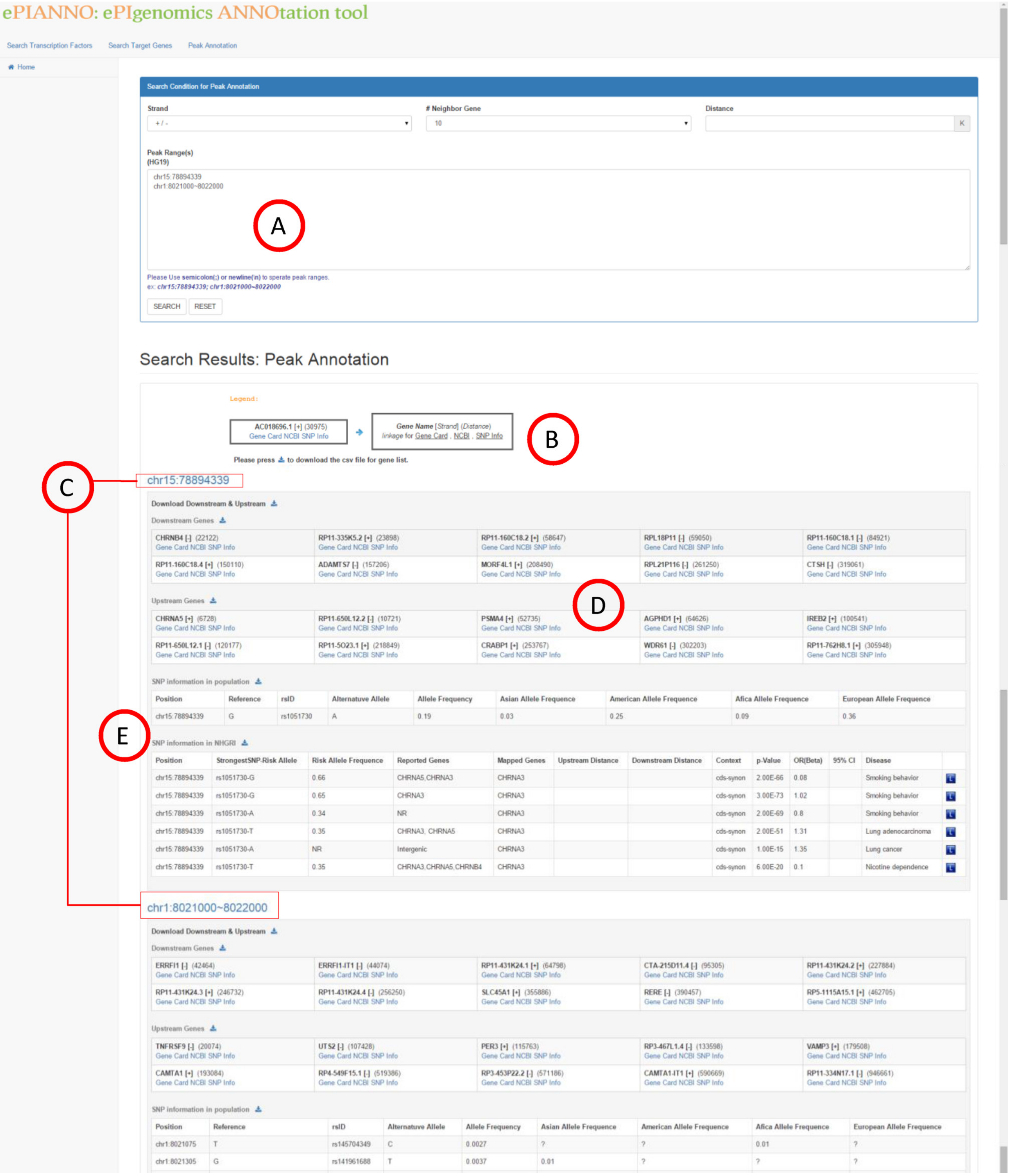
Peak annotations

Users may have their own DNA-binding location (peak data from user's ChIP-seq experiments) in mind and want to know the annotations of neighbour regions of the peak. To meet this need, the third function of *e*PIANNO is to search for neighbour genes and the annotation of specific genomic regions. This function helps a user to explore genes located near to given genomic coordinates of interest. If users have aimed on some specific regions, then this function can help them to explore local information in these regions. User could define region(s) of interest (Fig 4A), the number of neighbour genes, or distance to the given region in the query page. *e*PIANNO shows the genes and SNPs located around the region in query, both upstream and downstream. Gene information including gene name, +/- strand, distance to queried genomic region, and links of annotations will be shown on the browser (Fig 4B). Users could expand links (Fig 4C) and the detail gene information will be shown on the webpage or download them in the EXCEL CSV format (Fig 4D). Population SNP information and associations with disease are also provided (Fig 4E).

Example1: rs1051730(chr15:78894339;C>T) and SMARCA4

In a protein-DNA binding event, both protein and DNA were involved. Alternations of protein and DNA may both affect the binding event. As a consequence, if a genomic variant was found to be associated with some diseases, alternations of the correspondent DNA-binding protein may also be linked to similar diseases. In the following case, we demonstrate that a DNA-binding protein which targeted a disease-associated genomic variant was also a potential disease-associated protein.

In 2008 and 2009, a genomic variant which caused synonymous substitution in *CHRNA3*, rs1051730, was reported to be associated with lung cancer and lung adenocarcinoma in two independent GWASs, respectively [25, 26]. We use "Search transcription factors" function of *e*PIANNO to explore any potential transcription factor may bind on it. First, we just put "chr15:78894339" in the query box and select "Transcription factor or DNA-binding protein" as the ChIP type to conduct the search (Fig 2A). We found the binding regions of three subunits of RNA polymerase complex and two transcription factors, SMARCA4 and YY1, exactly flank this position (Fig 2D). This may imply that these DNA-binding proteins were also potentially associated with lung cancer. *YY1* had been reported as a known lung cancer associated gene [27–30]. In 2014, *SMARCA4* was identified to be one of tumour suppressor gene in the



ePIANNO: ePIgenomics ANNOtation tool
Search Transcription Factors
Search Target Genes
Peak Annotation
Home
SEARCH
RESET
A
B
C
D
E

Search Condition for Peak Annotation

Strand: + / -

Neighbor Gene: 10

Distance: K

Peak Range(s) (HG19)

chr15:78894339
chr1:8021000~8022000

Please Use semicolon(;) or newline(\n) to sperate peak ranges.
ex: *chr15:78894339; chr1:8021000~8022000*

Search Results: Peak Annotation

Legend:

AC018696.1 [+] (30975)
Gene Card NCBI SNP Info
→
Gene Name [*Strand*] (*Distance*)
linkage for Gene Card , NCBI , SNP Info

Please press ⤓ to download the csv file for gene list.

chr15:78894339

Download Downstream & Upstream ⤓

Downstream Genes ⤓

CHRNB4 [-] (22122) Gene Card NCBI SNP Info	RP11-335K5.2 [+] (23898) Gene Card NCBI SNP Info	RP11-160C18.2 [+] (58647) Gene Card NCBI SNP Info	RPL18P11 [-] (59050) Gene Card NCBI SNP Info	RP11-160C18.1 [-] (84921) Gene Card NCBI SNP Info
RP11-160C18.4 [+] (150110) Gene Card NCBI SNP Info	ADAMTS7 [-] (157206) Gene Card NCBI SNP Info	MORF4L1 [+] (208490) Gene Card NCBI SNP Info	RPL21P116 [-] (261250) Gene Card NCBI SNP Info	CTSH [-] (319061) Gene Card NCBI SNP Info

Upstream Genes ⤓

CHRNA5 [+] (6728) Gene Card NCBI SNP Info	RP11-650L12.2 [-] (10721) Gene Card NCBI SNP Info	PSMA4 [+] (52735) Gene Card NCBI SNP Info	AGPHD1 [+] (64626) Gene Card NCBI SNP Info	IREB2 [+] (100541) Gene Card NCBI SNP Info
RP11-650L12.1 [-] (120177) Gene Card NCBI SNP Info	RP11-5O23.1 [+] (218849) Gene Card NCBI SNP Info	CRABP1 [+] (253767) Gene Card NCBI SNP Info	WDR61 [-] (302203) Gene Card NCBI SNP Info	RP11-762H8.1 [+] (305948) Gene Card NCBI SNP Info

SNP information in population ⤓

Position	Reference	rsID	Alternatuve Allele	Allele Frequency	Asian Allele Frequence	American Allele Frequence	Afica Allele Frequence	European Allele Frequence
chr15:78894339	G	rs1051730	A	0.19	0.03	0.25	0.09	0.36

SNP information in NHGRI ⤓

Position	StrongestSNP-Risk Allele	Risk Allele Frequence	Reported Genes	Mapped Genes	Upstream Distance	Downstream Distance	Context	p-Value	OR(Beta)	95% CI	Disease
chr15:78894339	rs1051730-G	0.66	CHRNA5,CHRNA3	CHRNA3			cds-synon	2.00E-66	0.08		Smoking behavior
chr15:78894339	rs1051730-G	0.65	CHRNA3	CHRNA3			cds-synon	3.00E-73	1.02		Smoking behavior
chr15:78894339	rs1051730-A	0.34	NR	CHRNA3			cds-synon	2.00E-69	0.8		Smoking behavior
chr15:78894339	rs1051730-T	0.35	CHRNA3, CHRNA5	CHRNA3			cds-synon	2.00E-51	1.31		Lung adenocarcinoma
chr15:78894339	rs1051730-A	NR	Intergenic	CHRNA3			cds-synon	1.00E-15	1.35		Lung cancer
chr15:78894339	rs1051730-T	0.35	CHRNA3,CHRNA5,CHRNB4	CHRNA3			cds-synon	6.00E-20	0.1		Nicotine dependence

chr1:8021000~8022000

Download Downstream & Upstream ⤓

Downstream Genes ⤓

ERRFI1 [-] (42464) Gene Card NCBI SNP Info	ERRFI1-IT1 [-] (44074) Gene Card NCBI SNP Info	RP11-431K24.1 [+] (64798) Gene Card NCBI SNP Info	CTA-215D11.4 [-] (95305) Gene Card NCBI SNP Info	RP11-431K24.2 [+] (227884) Gene Card NCBI SNP Info
RP11-431K24.3 [+] (246732) Gene Card NCBI SNP Info	RP11-431K24.4 [-] (256250) Gene Card NCBI SNP Info	SLC45A1 [+] (355886) Gene Card NCBI SNP Info	RERE [-] (390457) Gene Card NCBI SNP Info	RP5-1115A15.1 [+] (462705) Gene Card NCBI SNP Info

Upstream Genes ⤓

TNFRSF9 [-] (20074) Gene Card NCBI SNP Info	UTS2 [-] (107428) Gene Card NCBI SNP Info	PER3 [+] (115763) Gene Card NCBI SNP Info	RP3-467L1.4 [-] (133598) Gene Card NCBI SNP Info	VAMP3 [+] (179508) Gene Card NCBI SNP Info
CAMTA1 [+] (193084) Gene Card NCBI SNP Info	RP4-549F15.1 [-] (519386) Gene Card NCBI SNP Info	RP3-453P22.2 [-] (571186) Gene Card NCBI SNP Info	CAMTA1-IT1 [+] (590669) Gene Card NCBI SNP Info	RP11-334N17.1 [-] (946661) Gene Card NCBI SNP Info

SNP information in population ⤓

Position	Reference	rsID	Alternatuve Allele	Allele Frequency	Asian Allele Frequence	American Allele Frequence	Afica Allele Frequence	European Allele Frequence
chr1:8021075	T	rs145704349	C	0.0027	?	?	0.01	?
chr1:8021305	G	rs141961688	T	0.0037	0.01	?	?	?

Fig 4. Peak annotations. We assume that users have aimed on some specific regions and this function is best to help them to explore local information of these regions. User could define region(s) of interest (A), number of neighbour gene, or distance to the given region in the query page. Gene information including gene name, strand, distance to region of query, and links of annotations will be shown (B). Users could expand links (C) and the detail gene information will be shown on the webpage or download them in EXCEL CSV format (D). SNP information in population and disease are also provided (E).

doi:10.1371/journal.pone.0148321.g004

non-small cell lung cancer because inactivation of *SMARCA4* could promote cancer aggressiveness by altering chromatin organization [31]. In *e*PIANNO, this genomic variant was annotated as the *SMARCA4* binding site by one experiment using cellline Helas3.

Example 2: rs35675666(chr1: 8021973;G>T) and *NFKB*

In 2011, a genomic variant, rs35675666, was identified to be associated with inflammatory bowel disease (IBD), including Crohn's disease and ulcerative colitis [32]. We use "Search transcription factors" function of *e*PIANNO to explore any potential transcription factor may bind on it. First, we put "chr1:8021973" in the query box (Fig 2A), select "ENCODE_standford" as the data source (Fig 2B), and search "Transcription factor or DNA-binding protein" CHiP type (Fig 2C). We found the binding regions of 32 transcription factors exactly flank this position (Fig 2D). Among them, 3 transcription factors, NFKB, HSF1, and HNF4A had been reported to be linked to IBD [33–35].

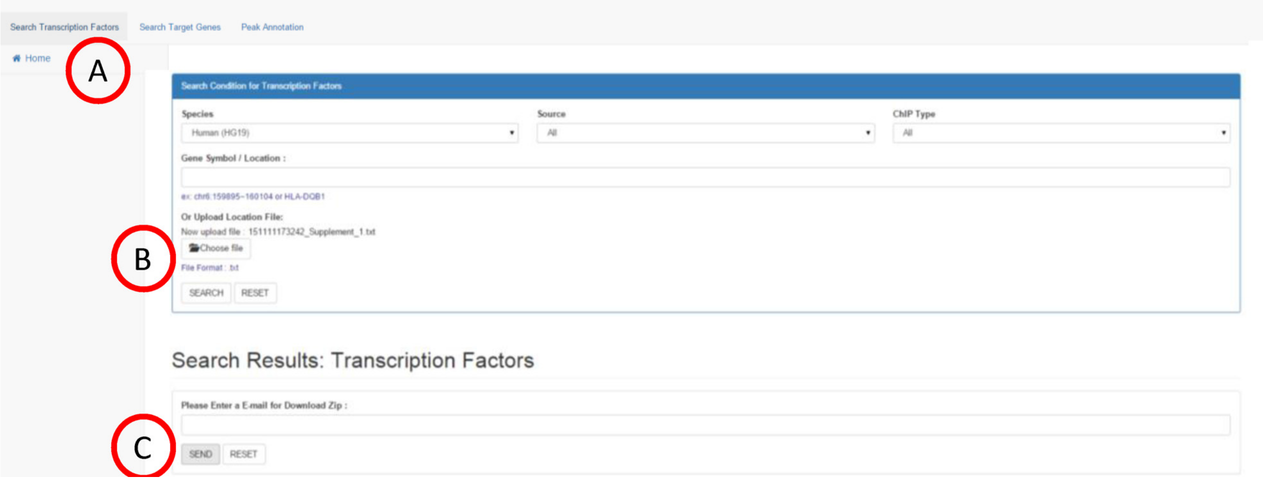
Alternations of the TF binding region may cause the expression dysregulation of the downstream gene. If a genomic variant occurred in the TF binding region was found to be associated with some diseases, the expression dysregulation of the downstream gene may also be associated with such diseases. As a demonstration, we used the "peak Annotation" function in the *e*PIANNO and search "chr1:8021973" for 20 nearby genes both upstream and downstream (Fig 4A). We found three genes locate nearby rs35675666 in 30K upstream or downstream: *DDX11L1*, *WASH7P*, and *TNFRSF9* (Fig 4C and 4D). DDX11L1 is a lncRNA with unknown function and *WASH7P* is a pseudogene, we turned to the gene *TNFRSF9* (chr1: 7975931~8003225; minus strand) which located 17K upstream of rs35675666. In 2012, this gene was reported as one candidate gene associated with IBD [36]. In 2014, another independent study of human intestinal T cell, which combined results of transcriptome and genomic variants to infer IBD associated features, showed that rs35675666 and expression of both *TNFRSF9* and *HNF4A* were associated with IBD [37].

Example 3: Batch query in *e*PIANNO

*e*PIANNO allowed users to upload their own peak data in TXT format or paste their peak regions in the query region and help them to add annotation. For example, (Fig 5A) user can choose "Search Transcription Factors", (Fig 5B) upload the sample peak file by click "SEARCH" (S1 Text), (Fig 5C) fill your email and click "SEND". After *e*PIANNO finishing this job, one compressed file will send email for user to download it. In this compressed file, each query region was saved in separated EXCEL files, respectively. In each EXCEL file, users could find the disease-associated SNPs, population SNP information, and peak data information. Currently, *e*PIANNO allowed maximum 300 query regions in each batch and each query region range within 100 kilo base pairs.

Discussion

The main goal of *e*PIANNO is to help users to explore the associations between protein-binding event, disease-associated genomic variants, and information of general populations by annotating the ChIP-seq datasets. Currently, most ChIP-seq web services or depositories, such as CR Cistrome, ChIPBase, CistromeFinder, hmChIP, SwissRegulon, and CistromeMap, provided annotations and gene symbol query function. However, they did not provide batch query or region query functions. On the other hand, GWAS web services, such as GWAS Central and GWASdb, provided user-friendly interfaces for users to query disease-associated SNPs with batch query or region query functions. However, they did not provide the linkage between



Search Condition for Transcription Factors

Species: Human (HG19)
Source: All
ChIP Type: All

Gene Symbol / Location :

ex: chr6:159895~160104 or HLA-DQB1

Or Upload Location File:
Now upload file : 151111173242_Supplement_1.txt
File Format : txt

Search Results: Transcription Factors

Please Enter a E-mail for Download Zip :

Fig 5. Batch query in ePIANNO. User could upload their queries in a text file instead of pasting on the website. For example, (A) user can choose "Search Transcription Factors", (B) upload the sample peak file (S1 Text), (C) fill your email, and then click "SEND". After *e*PIANNO finishing this job, one downloadable file will be shown on the page.

doi:10.1371/journal.pone.0148321.g005

SNPs, peak information of protein-DNA binding, or information of general populations (1000 Genomes Project). As a consequence, *e*PIANNO was built to fill this gap.

Furthermore, *e*PIANNO allowed users to upload their own peak data in TXT format or paste their peak regions in the query region and help them to add annotation.

SNPs represent the most common type of genomic variation. SNPs in the coding region are most intensively studied since they are relatively easy interpretable with well-annotated protein-coding sequences of human genome. Some SNPs localized to the binding sites of various transcription factors may alter the regulation of gene transcription and exert functional effects. *e*PIANNO integrated the experiment results of ChIP-seq (hmChIP, ENCODE, and Roadmap Epigenomics), SNP information of populations (1000 Genomes Project), disease association information of GWAS (NHGRI), and gene annotation (ENCODE annotation). Users could not only search the relationship between TFs and target genes, but also explore the relevant genome variants and diseases. If users have their own genomic regions of interest, *e*PIANNO provides batch query methods, currently not provided by UCSC Genome Browser, (separated by semi-colon or another line) to explore genes and SNPs located in upstream or downstream with ENCODE gene annotations.

GWAS Catalog of NHGRI is the most comprehensive disease-SNP association database of the public domain. The 1000 Genomes Project contains most comprehensive genomic variants of populations worldwide. Genomic variants may contribute to binding affinity or other properties of protein-DNA interaction. In addition, some genomic variants are associated with known diseases. We demonstrated how users could use functions of *e*PIANNO webserver to explore useful information about TF related genomic variants. The example of SNP rs1051730 showed that DNA-binding proteins targeting a disease-associated genomic variant were also likely to be disease-associated genes. The example of SNP rs35675666 illustrated how *e*PIANNO helped establish the connection between the genomic variant of upstream TF-binding region, the associated disease and the TF expression. We believe that if users have their own disease-associated SNPs, may be from their own study, *e*PIANNO would be helpful to find out the potential relationship between SNPs and disease by annotating their roles at the epigenomic level.

Previous studies had used ChIP-seq data of ENCODE, GWASs Catalogs of NHGRI, and 1000 Genomes Project dataset or HAPMAP III dataset to investigate the relationships between SNPs, disease, and protein-DNA binding effects[18–20]. Results of these studies showed high quality statistical analysis and promising results. However, comprehensive information of ChIP-seq data (ENCODE, ROADMAP epigenomics, and hmChIP) and genomic variation (1000 Genomes Project) was still needed updated. In addition, previous studies showed that combinations of genomic variants and ChIP-seq data did provide valuable information for researchers. However, researchers without computational abilities cannot explore associations between their interesting TFs, binding sites, or genomic variations by themselves. Hence, a web based server with user friendly interface, easy input format, and output function would be useful for researchers without computational skills to obtain valuable information. As a consequence, we build this server *e*PIANNO with most update ChIP-seq datasets, the GWASs Catalog dataset, and annotations for users to query and annotate regions of their interest.

Genomic variants in the protein-DNA binding region might directly affect the binding strength of TFs or mode of chromatin modification at the transcription level. In addition, effects of the genomic variants might further reach the post-transcription level if the expressions of miRNA or lncRNA are affected. Although we did not give such complicate case of indirect effects of genomic variants in our examples, *e*PIANNO might help user to find out such case especially when they have their own protein-DNA binding regions of their own interest.

In *e*PIANNO, data of 1000 Genomes Project came from historical healthy subjects. However, data of NHGRI came mostly from various diseases and ChIP-seq data were generated from cells lines with various experimental conditions. As a consequence, combining data of annotations and ChIP-seq had its own limitations. Although *e*PIANNO helped users to explore potential associations between disease, protein-binding information, and their own regions of interest, these associations may not broadly applied in all cell types or in all experimental conditions. Results had to be evaluated by users' aims and hypothesis of their studies. For an example, if lung cancer was considered in NHGRI database, cell line of ChIP-seq experiments was better to be lung cell lines. It may provide a more reasonable association between a SNP and a protein-binding site.

Supporting Information

S1 Text. Supporting text.
(TXT)

Acknowledgments

We especially thank that Integrated Core Facility for Functional Genomics (C5), Pharmacogenomics Lab (TR6), and Mathematics in Biology Group of Institute of Statistical Science AS helped us to test website tools and share their user experiences to improve our work.

Author Contributions

Conceived and designed the experiments: HYC. Performed the experiments: CHL BCH CLC. Analyzed the data: CHL YCH CSC YHC. Wrote the paper: CHL KCL HYC. Design the index system to improving query efficiency: YHH SSY YCL.

References

1. Kouzarides T. Chromatin Modifications and Their Function. Cell. 2007; 128(4):693–705. PMID: 17320507
2. Pan Y, Tsai C-J, Ma B, Nussinov R. Mechanisms of transcription factor selectivity. Trends in Genetics. 2010; 26(2):75–83. doi: 10.1016/j.tig.2009.12.003 PMID: 20074831
3. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-Wide Mapping of in Vivo Protein-DNA Interactions. Science. 2007; 316(5830):1497–502. doi: 10.1126/science.1141319 PMID: 17540862
4. Blecher-Gonen R, Barnett-Itzhaki Z, Jaitin D, Amann-Zalcenstein D, Lara-Astiaso D, Amit I. High-throughput chromatin immunoprecipitation for genome-wide mapping of in vivo protein-DNA interactions and epigenomic states. Nat Protocols. 2013; 8(3):539–54. doi: 10.1038/nprot.2013.023 PMID: 23429716
5. Wang Q, Huang J, Sun H, Liu J, Wang J, Wang Q, et al. CR Cistrome: a ChIP-Seq database for chromatin regulators and histone modification linkages in human and mouse. Nucleic Acids Research. 2013. doi: 10.1093/nar/gkt1151
6. Yang J-H, Li J-H, Jiang S, Zhou H, Qu L-H. ChIPBase: a database for decoding the transcriptional regulation of long non-coding RNA and microRNA genes from ChIP-Seq data. Nucleic Acids Research. 2012. doi: 10.1093/nar/gks1060
7. Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF, et al. NCBI GEO: archive for functional genomics data sets—10 years on. Nucleic Acids Research. 2011; 39(suppl 1):D1005–D10. doi: 10.1093/nar/gkq1184
8. Coordinators NR. Database resources of the National Center for Biotechnology Information. Nucleic Acids Research. 2014; 42(D1):D7–D17. doi: 10.1093/nar/gkt1146
9. Lachmann A, Xu H, Krishnan J, Berger SI, Mazloom AR, Ma'ayan A. ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. Bioinformatics. 2010; 26(19):2438–44. doi: 10.1093/bioinformatics/btq466 PMID: 20709693
10. Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, Greven MC, et al. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Research. 2012; 22(9):1798–812. doi: 10.1101/gr.139105.112 PMID: 22955990
11. Sun H, Qin B, Liu T, Wang Q, Liu J, Wang J, et al. CistromeFinder for ChIP-seq and DNase-seq data reuse. Bioinformatics. 2013; 29(10):1352–4. doi: 10.1093/bioinformatics/btt135 PMID: 23508969
12. Qin B, Zhou M, Ge Y, Taing L, Liu T, Wang Q, et al. CistromeMap: a knowledgebase and web server for ChIP-Seq and DNase-Seq studies in mouse and human. Bioinformatics. 2012; 28(10):1411–2. doi: 10.1093/bioinformatics/bts157 PMID: 22495751
13. Chen L, Wu G, Ji H. hmChIP: a database and web server for exploring publicly available human and mouse ChIP-seq and ChIP-chip data. Bioinformatics. 2011; 27(10):1447–8. doi: 10.1093/bioinformatics/btr156 PMID: 21450710
14. Pachkov M, Erb I, Molina N, van Nimwegen E. SwissRegulon: a database of genome-wide annotations of regulatory sites. Nucleic Acids Research. 2007; 35(suppl 1):D127–D31. doi: 10.1093/nar/gkl857
15. Tang Q, Chen Y, Meyer C, Geistlinger T, Lupien M, Wang Q, et al. A Comprehensive View of Nuclear Receptor Cancer Cistromes. Cancer Research. 2011; 71(22):6940–7. doi: 10.1158/0008-5472.can-11-2091 PMID: 21940749
16. Lanz RB, Jericevic Z, Zuercher WJ, Watkins C, Steffen DL, Margolis R, et al. Nuclear Receptor Signaling Atlas (www.nursa.org): hyperlinking the nuclear receptor signaling community. Nucleic Acids Research. 2006; 34(suppl 1):D221–D6. doi: 10.1093/nar/gkj029
17. Fingerman IM, McDaniel L, Zhang X, Ratzat W, Hassan T, Jiang Z, et al. NCBI Epigenomics: a new public resource for exploring epigenomic data sets. Nucleic Acids Research. 2011; 39(suppl 1):D908–D12. doi: 10.1093/nar/gkq1146
18. Huang D, Ovcharenko I. Identifying causal regulatory SNPs in ChIP-seq enhancers. Nucleic Acids Research. 2014. doi: 10.1093/nar/gku1318
19. Bryzgalov LO, Antontseva EV, Matveeva MY, Shilov AG, Kashina EV, Mordvinov VA, et al. Detection of Regulatory SNPs in Human Genome Using ChIP-seq ENCODE Data. PLoS ONE. 2013; 8(10): e78833. doi: 10.1371/journal.pone.0078833 PMID: 24205329
20. Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M. Linking disease associations with regulatory information in the human genome. Genome Research. 2012; 22(9):1748–59. doi: 10.1101/gr.136127.111 PMID: 22955986
21. Consortium TGP. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012; 491(7422):56–65. doi: 10.1038/nature11632 PMID: 23128226

22. Consortium TGP. A map of human genome variation from population-scale sequencing. Nature. 2010; 467(7319):1061–73. doi: 10.1038/nature09534 PMID: 20981092

23. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Research. 2014; 42(D1):D1001–D6. doi: 10.1093/nar/gkt1229

24. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: The reference human genome annotation for The ENCODE Project. Genome Research. 2012; 22(9):1760–74. doi: 10.1101/gr.135350.111 PMID: 22955987

25. McKay JD, Hung RJ, Gaborieau V, Boffetta P, Chabrier A, Byrnes G, et al. Lung cancer susceptibility locus at 5p15.33. Nat Genet. 2008; 40(12):1404–6. doi: 10.1038/ng.254 PMID: 18978790

26. Landi MT, Chatterjee N, Yu K, Goldin LR, Goldstein AM, Rotunno M, et al. A Genome-wide Association Study of Lung Cancer Identifies a Region of Chromosome 5p15 Associated with Risk for Adenocarcinoma. The American Journal of Human Genetics. 2009; 85(5):679–91. doi: 10.1016/j.ajhg.2009.09.012 PMID: 19836008

27. Wu S, Kasim V, Kano MR, Tanaka S, Ohba S, Miura Y, et al. Transcription Factor YY1 Contributes to Tumor Growth by Stabilizing Hypoxia Factor HIF-1α in a p53-Independent Manner. Cancer Research. 2013; 73(6):1787–99. doi: 10.1158/0008-5472.can-12-0366 PMID: 23328582

28. Wang C-C, Tsai M-F, Hong T-M, Chang G-C, Chen C-Y, Yang W-M, et al. The transcriptional factor YY1 upregulates the novel invasion suppressor HLJ1 expression and inhibits cancer cell invasion. Oncogene. 2005; 24(25):4081–93. PMID: 15782117

29. Wang C-C, Chen JJ, Yang P-C. Multifunctional transcription factor YY1: a therapeutic target in human cancer? Expert Opinion on Therapeutic Targets. 2006; 10(2):253–66. doi: 10.1517/14728222.10.2.253 PMID: 16548774.

30. Wang C-C, Tsai M-F, Dai T-H, Hong T-M, Chan W-K, Chen JJW, et al. Synergistic Activation of the Tumor Suppressor, HLJ1, by the Transcription Factors YY1 and Activator Protein 1. Cancer Research. 2007; 67(10):4816–26. doi: 10.1158/0008-5472.can-07-0504 PMID: 17510411

31. Orvis T, Hepperla A, Walter V, Song S, Simon JM, Parker JS, et al. BRG1/SMARCA4 inactivation promotes non-small cell lung cancer aggressiveness by altering chromatin organization. Cancer Research. 2014. doi: 10.1158/0008-5472.can-14-0061

32. Anderson CA, Boucher G, Lees CW, Franke A, D'Amato M, Taylor KD, et al. Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. Nat Genet. 2011; 43(3):246–52. doi: 10.1038/ng.764 PMID: 21297633

33. Darsigny M, Babeu J-P, Dupuis A-A, Furth EE, Seidman EG, Lévy É, et al. Loss of Hepatocyte-Nuclear-Factor-4α Affects Colonic Ion Transport and Causes Chronic Inflammation Resembling Inflammatory Bowel Disease in Mice. PLoS ONE. 2009; 4(10):e7609. doi: 10.1371/journal.pone.0007609 PMID: 19898610

34. Atreya I, Atreya R, Neurath MF. NF-κB in inflammatory bowel disease. Journal of Internal Medicine. 2008; 263(6):591–6. doi: 10.1111/j.1365-2796.2008.01953.x PMID: 18479258

35. Tanaka K-I, Mizushima T. Protective role of HSF1 and HSP70 against gastrointestinal diseases. International Journal of Hyperthermia. 2009; 25(8):668–76. doi: 10.3109/02656730903213366 PMID: 20021227.

36. Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature. 2012; 491(7422):119–24. doi: 10.1038/nature11582 PMID: 23128233

37. Raine T, Liu JZ, Anderson CA, Parkes M, Kaser A. Generation of primary human intestinal T cell transcriptomes reveals differential expression at genetic risk loci for immune-mediated disease. Gut. 2014. doi: 10.1136/gutjnl-2013-306657