PLOS ONE

RESEARCH ARTICLE

PEREGRINE: A genome-wide prediction of enhancer to gene relationships supported by experimental evidence

Caitlin Mills[1], **Anushya Muruganujan**[2], **Dustin Ebert**[2], **Crystal N. Marconett**[3,4,5], **Juan Pablo Lewinger**[1], **Paul D. Thomas**[2], **Huaiyu Mi**[2]*

1 Division of Biostatistics, Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA, United States of America, **2** Division of Bioinformatics, Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA, United States of America, **3** Department of Surgery, Keck School of Medicine, University of Southern California, Los Angeles, CA, United States of America, **4** Department of Biochemistry and Molecular Medicine, Keck School of Medicine USC, Los Angeles, CA, United States of America, **5** Norris Cancer Center, Keck School of Medicine USC, Los Angeles, CA, United States of America

* huaiyumi@usc.edu

Citation: Mills C, Muruganujan A, Ebert D, Marconett CN, Lewinger JP, Thomas PD, et al. (2020) PEREGRINE: A genome-wide prediction of enhancer to gene relationships supported by experimental evidence. PLoS ONE 15(12): e0243791. https://doi.org/10.1371/journal.pone.0243791

Editor: Ludmila Prokunina-Olsson, National Cancer Institute, UNITED STATES

Received: September 3, 2020

Accepted: November 25, 2020

Published: December 15, 2020

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0243791

Data Availability Statement: All data from the PEREGRINE database are freely available through

Abstract

Enhancers are powerful and versatile agents of cell-type specific gene regulation, which are thought to play key roles in human disease. Enhancers are short DNA elements that function primarily as clusters of transcription factor binding sites that are spatially coordinated to regulate expression of one or more specific target genes. These regulatory connections between enhancers and target genes can therefore be characterized as enhancer-gene links that can affect development, disease, and homeostatic cellular processes. Despite their implication in disease and the establishment of cell identity during development, most enhancer-gene links remain unknown. Here we introduce a new, publicly accessible database of predicted enhancer-gene links, PEREGRINE. The PEREGRINE human enhancer-gene links interactive web interface incorporates publicly available experimental data from ChIA-PET, eQTL, and Hi-C assays across 78 cell and tissue types to link 449,627 enhancers to 17,643 protein-coding genes. These enhancer-gene links are made available through the new Enhancer module of the PANTHER database and website where the user may easily access the evidence for each enhancer-gene link, as well as query by target gene and enhancer location.

Introduction

Enhancers are short regulatory DNA elements that regulate their target genes in a tissue-and cell-type specific manner [1]. They may be located locally or distally from their target genes; they may even be located within the intronic regions of a target gene. Enhancers are found upstream and downstream from their target genes and often function regardless of orientation and are identifiable by the presence of post-translational modifications to the histone tails of nucleosomes that organize the DNA within the enhancer, specifically the acetyl moiety

interactive interface or bulk download at https://www.peregrineproj.org.

Funding: HM, AM, DE, JPL, PDT, CM P01CA196569 National Institute of Health https://www.nih.gov/ 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.

attached to lysine 27 of histone H3 (H3K27Ac) as well as mono-methylation of histone H3 on lysine 4 (H3K4me1) [2,3]. It is estimated that up to one million enhancers in the human genome regulate the roughly 20,000 protein-coding genes through a complex network of many-to-many relationships where each enhancer may regulate multiple genes and each gene may be subject to regulation by multiple enhancers [4,5]. This transcriptional regulation has been demonstrated to occur through transcription factor binding of enhancers which then assist in the recruitment of the Mediator complex and additional transcriptional machinery to the promoter of the target gene [6]. Enhancers play an important role in establishing cell-specific identities during differentiation and development through the regulation of gene transcription and in the maintenance of cellular homeostasis [7,8]. When dysregulation occurs, disease may follow as a result. Indeed, enhancers have been implicated in many human diseases, including various cancers [8,9].

Despite the importance of enhancer-gene links, relatively little is known about which target genes are regulated by specific enhancers. Years of painstaking benchwork and analyses have been performed at the individual laboratory level to discover many high-confidence enhancer-gene links in many cell types [10]. Although these experimentally validated enhancer-gene links are highly valuable, they represent only a small percentage of the total enhancer-gene links acting within in the human organism and are not available in any central location for researchers to access, making it difficult to wield the full power of this data in any large scale analysis [11]. Higher throughput methods for predicting enhancer-gene links offer a desirable alternative for characterizing this vast regulatory network. High throughput experimental methods such as promoter capture genome-wide chromosome conformation capture (PCHi-C) [12] and genome-wide screens using Clustered Regulatory Interspersed Short Palindromic Repeat (CRISPR) mediated deletions have helped toward this aim. However, these methods are still relatively new and therefore results are not yet widely available in a range of cell types. Many research groups lack the equipment or knowledge of computational methodology to perform or analyze these experiments, presenting a barrier to gaining this information in new cell types. Computational prediction methods have yielded some enhancer-gene link databases which do not require new experimental data and instead utilize the existing publicly available data to generate predicted enhancer-gene links [1,13–17]. However, these databases often do not provide certain important information related to these enhancer-gene links to the end user. It is also not always clear what specific types of experimental evidence in which cell types supports each enhancer-gene link in the download data, an important consideration for many researchers [13]. Frequently, predicted enhancer-gene links are only available through individual webpages for each enhancer-gene link or gene [13,14,17]. Many websites lack an up to date and complete bulk data download file to allow overall analysis of the enhancer-gene link data with minimal processing [13–15,17].

Here we present the PEREGRINE (**P**redicted by **E**xperimental **R**esults: **E**nhancer-**G**ene **R**elationships **I**llustrated by a **N**exus of **E**vidence) enhancer-gene links (www.peregrineproj.org), which can be queried via the PANTHER [18,19] website. These PEREGRINE links represent a comprehensive set of enhancer-gene links with accompanying experimental evidence available via bulk download (www.peregrineproj.org), and also searchable by variants and putative target gene(s) of interest (www.pantherdb.org). Furthermore, since PANTHER is a comprehensive resource for gene function which provides the most up-to-date functional annotations to genes, including Gene Ontology [20,21], Reactome [22] and PANTHER Pathways [23], these enhancer-gene links provide functional implications to the non-coding regions in the genome.

To generate reliable enhancer-gene links, we have assembled a set of reliable enhancers from sources well-known and trusted by the scientific community that contained minimally

redundant enhancer information. Enhancer data was accordingly gathered from ENCODE [4], Ensembl [24], FANTOM [25], and VISTA [26] to generate a list of putative enhancers which were then linked to protein-coding genes using publicly available experimental data from Hi-C, ChIA-PET, and eQTL experiments from ENCODE and GTEx [27]. By incorporating several types of experiments that provide information on different aspects of enhancer activity and function, the PEREGRINE enhancer-gene links represent an amalgam of information gathered by examining the various characteristics of the enhancer-gene regulatory dynamic.

Although enhancers often act on their target genes from distant regions of the primary DNA sequence, the probability of a regulatory relationship is thought to drop considerably with increasing genomic distance between the enhancer and the gene [13]. A megabase of separation is considered to be a practical upper bound on most enhancer-gene links [11,28]. To this end, topologically associated domain (TAD) data was incorporated from Hi-C experiments. TADs are spatial subdivisions of chromatin where physical contacts within the TAD are much more likely than contact between DNA elements located within separate TADs [11]. TADs are typically a few hundred kilobases to a few megabases of contiguous DNA which has been shown to consist of regions folded within themselves into local compartments where a higher number of chromatin contacts take place [28]. It is believed that the majority of enhancer-gene links are found within the same TAD, which are fairly stable across cell types and remain largely stable throughout development [29–31].

Enhancers usually bind to multiple transcription factors [6]. Indeed, the fully characterized enhancers are known to upregulate their target genes through the recruitment of transcriptional machinery. In order to do this, they must enter into close proximity of their target gene's promoter [32–35]. Active enhancers engaged in upregulation of their target genes are typically associated with the presence of H3K27ac marks as well as the presence of RNA polymerase II, the polymerase responsible for mRNA transcription of genes in humans [36]. In order to assay this function of enhancer activity on specific target genes, data was incorporated from an experiment that could detect genome-wide looping interactions at high resolution. Chromatin Interaction Analysis by Paired-End Tag sequencing (ChIA-PET) is an assay that is capable of assessing chromatin interaction frequency through the targeting of DNA regions that are bound to a specific protein of interest [37]. The protein of interest (in this case RNA polymerase II) is pulled down from cross-linked fragmented chimeric DNA fragments with an antibody which is specific to the protein of interest [37]. Sequencing is then performed on the paired fragments to enable the investigator to examine which DNA regions were interacting with which DNA regions via binding of the protein of interest [37]. ChIA-PET data were examined to ascertain which enhancers were localized to which promoters in the presence of RNA polymerase II.

An expression quantitative trait locus (eQTL) is a locus that "explains a fraction of the genetic variance of a gene expression phenotype. Standard eQTL analysis involves a direction association test between markers of genetic variation with gene expression levels typically measured in tens or hundreds of individuals [38]." Oftentimes eQTL are located within an exon and may result in a nonsynonymous mutation in the gene product, but eQTL occur outside of exons and beyond the gene body as well [39,40]. Indeed, we used many statistically significant eQTL that mapped to enhancers to link these regulatory elements to their putative target genes.

Many factors must be considered when seeking to link enhancers to their target genes. They should be observed in spatial proximity to the promoter of their target gene when it is bound by transcriptional machinery, which ChIA-PET targeting RNA Polymerase II can assay. Hi-C data can be informative when enhancers and promoters are located within the

same TAD, and especially when they are thought to be interacting within the same TAD, sometimes referred to as the within-TAD contacts called as hierarchical TADs [28]. Additionally, when an enhancer contains a variant that has been shown to explain part of the variation of gene expression, it implicates the enhancer in a transcriptional regulatory relationship with the gene. Taken together, all of these data predict many potential enhancer-gene links with varying amounts of evidence across many cell types. By providing the details of these predicted enhancer-gene links in a cell-type specific manner via the PANTHER website, we introduce a useful and easily accessible network of well-supported enhancer-gene regulatory links in 78 cell and tissue types. By presenting these enhancer-gene links within the existing framework of PANTHER, not only will we be able to extend our understanding of function for those enhancers, but also users will be able to browse them in the context of function and existing gene pathways.

Materials and methods

Gathering the enhancer set

Enhancer data was gathered from reliable and highly utilized sources. Enhancer coordinate data from four sources comprised the final enhancer set. These sources include: ENCODE's catalog of candidate Cis-Regulatory Elements, VISTA, Ensembl, and FANTOM (URLs available in S1 File). Tissue-specific information on the enhancers was not included, but rather a list of enhancers defined by genomic location was the end product. These enhancer coordinates were the result of experiments performed across various tissues by their respective sources, which differed somewhat according to each enhancer source. We did not filter out any enhancers from these sources, instead opting to utilize their differing enhancer calling pipelines simultaneously to maximize the chance of capturing the highest number of enhancers. Each enhancer was assigned a numeric ID except for enhancers taken from ENCODE's cCRE which retained their original alphanumeric IDs assigned by that project. See Tables 1–3 for a brief summary of enhancer information from these sources. Enhancer overlap for comparison purposes only was determined using bedtools [41] using the minimum threshold of 1bp to determine overlap between two elements unless otherwise stated.

Constructing the enhancer-gene links

ChIA-PET. ChIA-PET experimental data were taken from ENCODE's data matrix for cell types K562, MCF7, and HCT116 and processed with the MANGO [42] pipeline to obtain pairs of interacting regions with p-values for each pair. The protein target for the ChIA-PET experiments was RNA Polymerase II. Bedtools [41] (an open source software package comprised of multiple tools for comparing and exploring genomic datasets) was then used to screen each region for PEREGRINE enhancers and protein-coding genes. If at least 50% of the enhancer overlapped with the ChIA-PET region, or at least 50% of the ChIA-PET region

Table 1. Summary of the PEREGRINE enhancer set by source of enhancers and average length of enhancers in base pairs.

Enhancer Source (Number of enhancers)	Average length (base pairs)
ENCODE cCRE (991,173)	423
Ensembl (28,239)	662
FANTOM (65,423)	281
VISTA (959)	2037
Total combined (1,085,794)	**422**

https://doi.org/10.1371/journal.pone.0243791.t001

Table 2. Enhancer overlap between sources.

	ENCODE	Ensembl	FANTOM	VISTA
ENCODE	991,173 (100%)	36,284 (3.6%)	42,457 (4.3%)	1,916 (0.2%)
Ensembl	24,294 (86.0%)	28,239 (100%)	5,234 (18.5%)	92 (0.3%)
FANTOM	40,378 (61.7%)	5,699 (8.7%)	65,423 (100%)	203 (0.3%)
VISTA	763 (79.6%)	75 (7.8%)	144 (15.0%)	959 (100%)

Numbers in the cells represent the number of enhancers from the sources in each row that were found to overlap with enhancers from the sources in each column. Percentages are of the total number of enhancers from the source listed for each row.

https://doi.org/10.1371/journal.pone.0243791.t002

overlapped with the enhancer, the ChIA-PET region was considered to contain the enhancer. Then the enhancer-containing region's interaction partner was screened for the promoter of a gene. For purposes of this analysis, the gene's transcription start site and the preceding 600bp of its promoter were considered to be the promoter. If at least 50% of a gene's promoter overlapped with the ChIA-PET region, or at least 50% of the ChIA-PET region overlapped with the gene's promoter, the region was considered to contain a gene capable of upregulation by the enhancer in its ChIA-PET interacting partner region. Thus, the pairs of interacting ChIA-PET regions containing an enhancer and a promoter according to these parameters were recorded as enhancer-gene links if the ChIA-PET interaction achieved significance at the $\alpha = 0.05$ level.

Expression Quantitative Trait Loci (eQTL). eQTL data were downloaded from GTEx for all 48 available tissues if they were statistically significant ($p<0.05$). Any eQTL located within the exons of the gene they were associated with were excluded from analysis. Only protein-coding genes were considered for this analysis. Bedtools intersect was then used to map eQTL to enhancers. If an eQTL was located within an enhancer, it was considered linked to the gene influenced by the eQTL. Individual eQTL were recorded with tissue type, eQTL, p-value, enhancer, and gene.

Hierarchical topologically associated domains. Analyzed Hi-C data was downloaded from PSYCHIC[28] for the 9 available cell types. Regions were provided that interacted with the promoter of the listed gene with a FDR of <0.01. Bedtools intersect was then used to map PEREGRINE enhancers to these regions. If at least 90% of an enhancer overlapped with one of these regions, it was recorded as linked to the gene PSYCHIC reported as physically interacting with the region. The cell type and FDR were also recorded for each enhancer-gene link.

Topologically associated domains. Topologically associated domain (TAD) boundary data were downloaded from ENCODE's Hi-C experiments for 19 cell types. Bedtools intersect was then used to screen each region for PEREGRINE enhancers and protein-coding genes. If at least 90% of the enhancer overlapped with the TAD, the TAD was considered to contain the enhancer. Then the TAD was screened for the promoter of a gene. For purposes of this analysis, the gene's transcription start site and the preceding 600bp of its promoter were considered

Table 3. Summary of the PEREGRINE enhancer-gene links dataset. These enhancer-gene links were taken from datasets across 78 tissues.

Enhancer-Gene Links By Assay	Number of Links Generated From Each Assay
ChIA-PET	11,402
eQTL	435,973
TAD	855,976
Hierarchical TAD	491,346
Total Enhancer-Gene Links	890,403

https://doi.org/10.1371/journal.pone.0243791.t003

to be the promoter. If at least 90% of a gene's promoter overlapped with the TAD, the TAD was considered to contain a gene capable of upregulation by the enhancer within the same TAD. Thus, all enhancers contained within a TAD were linked to all of the genes with promoters located in the same TAD. These enhancer-gene links were only recorded for enhancer-gene links already generated from another assay. This ensured that enhancer-gene links generated only due to the enhancer and the gene being located within the same TAD (a relatively weaker form of supporting evidence likely to include a disproportionately large amount of false enhancer-gene links) were not recorded.

Integrating the PEREGRINE enhancer-gene link data into PANTHER for interactive online access

The enhancer-gene link data is indexed and in an Apache Solr [43] database. The stored data contains gene, enhancer (ID and coordinates), assay (tissue, score), and source information. PANTHER retrieves the data from the Solr DB through requests by gene, enhancer, coordinate to a python Flask REST [44] API server that then communicates with Solr to return the results.

The PANTHER website primarily handles genomic data and its attributes. The Enhancer REST API is used to retrieve additional information about enhancers that have been mapped to genes. There are three areas where the enhancer REST API is utilized:

1. It is queried via SNP i.e. chromosome with start and end position to return list of associated enhancers and genes. This feature is used when mapping VCF data.
2. It is queried via gene identifier after SNP data has been mapped to genes, to determine enhancers associated with genes. The same functionality is used to retrieve information about enhancers for a single gene when displaying gene detail information.
3. It is queried via enhancer ID to determine additional details about an enhancer as well as the list of genes that it enhances.

Statistical analysis

We performed enrichment analysis using Gene Ontology[20] Biological Processes if there were any gene pathways enriched by having more or less than the expected number of linked enhancers per gene under the null hypothesis of the Mann Whitney U test that the two samples come from the same distribution via the PANTHER web interface [45].

Data availability

The PEREGRINE enhancer-gene links are available at www.peregrineproj.org, and can be queried via the PANTHER website (www.pantherdb.org). The GitHub repository for this work, which includes the URLs to all the source data as well as scripts to generate the final dataset, is available at https://github.com/USCbiostats/PEREGRINE_enhancer_gene_links.

Results and discussion

The PEREGRINE enhancer set

The PEREGRINE enhancer set consists of 1,085,794 enhancers from ENCODE's catalog of candidate Cis-Regulatory Elements, Ensembl, FANTOM, and VISTA with an average length of 422 bp (Table 1). A total of 991,173 non-overlapping enhancers were collected from ENCODE with an average length of 423 bp. Another 28,239 non-overlapping enhancers were collected from Ensembl with an average length of 662 bp. Additionally, 65,423 non-

overlapping enhancers were collected from FANTOM with an average length of 281 bp. Finally, 959 enhancers were collected from VISTA with an average length of 2,037 bp. Seven of these overlapped with another enhancer within the VISTA set. Although the enhancers from each source were almost perfectly non-overlapping among enhancers from the same source, there was some overlap between enhancers from different sources. Overlap in this context was calculated at the minimum threshold of 1 bp. 157,539 PEREGRINE enhancers overlapped with at least one other enhancer to form 1,002,071 non-overlapping enhancer regions (426 million base pairs total) from 1,085,794 total enhancers, accounting for ~13% of the human genome. These overlapping enhancers represent 14.5% of the PEREGRINE enhancer set. For a breakdown of enhancer overlap by source, see Table 2.

Characterizing genome-wide enhancer-gene link relationships in PEREGRINE

Altogether, there were 890,403 enhancer-gene links generated from ChIA-PET, eQTL, hierarchical TAD, and linear TAD data across 78 cell and tissue types (Fig 1). These enhancer-gene links linked 449,627 enhancers representing nearly 181 million bp (~6% of the genome) to 17,643 genes. Across all enhancer-gene link data, each enhancer was linked to an average of 2 genes (1.98) and each gene was linked to an average of 50 enhancers (50.47). These averages are about what might be expected for roughly one million enhancers regulating roughly 20,000 protein-coding genes. Histograms of the number of enhancers per gene and the number of genes per enhancer are provided in Fig 2 as well as Table 4 giving the cumulative percentages for each frequency value. Unsurprisingly, the value with greatest density in both histograms is 1. Indeed, most enhancers (56%) had only one putative target gene, and 96% of enhancers had 5 putative target genes or less. The highest number of putative target genes per enhancer was 34, but this accounted for only one enhancer. Enhancers with 10 or more putative target genes accounted for less than 1% of the total enhancers linked to genes in this analysis. When

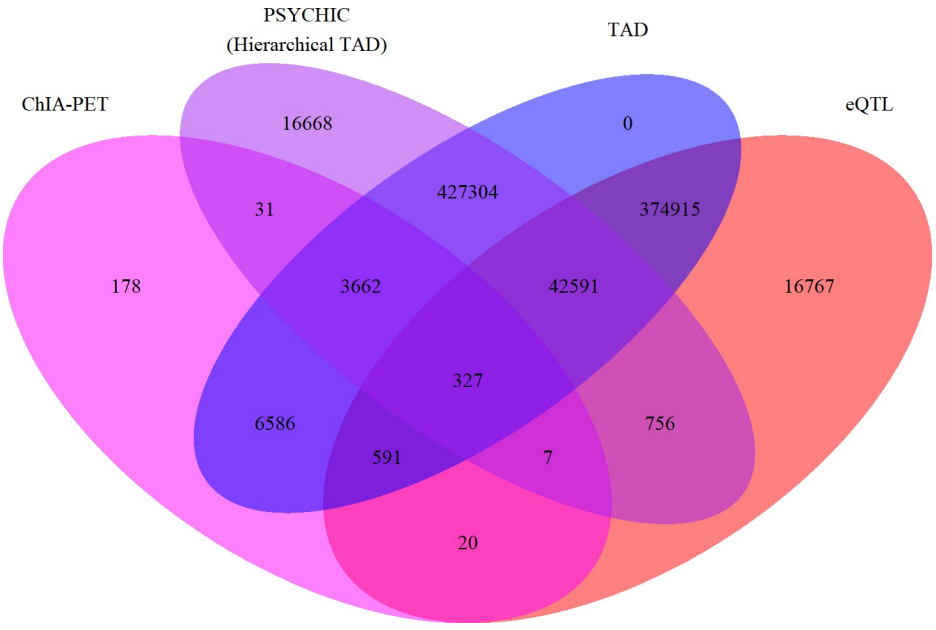
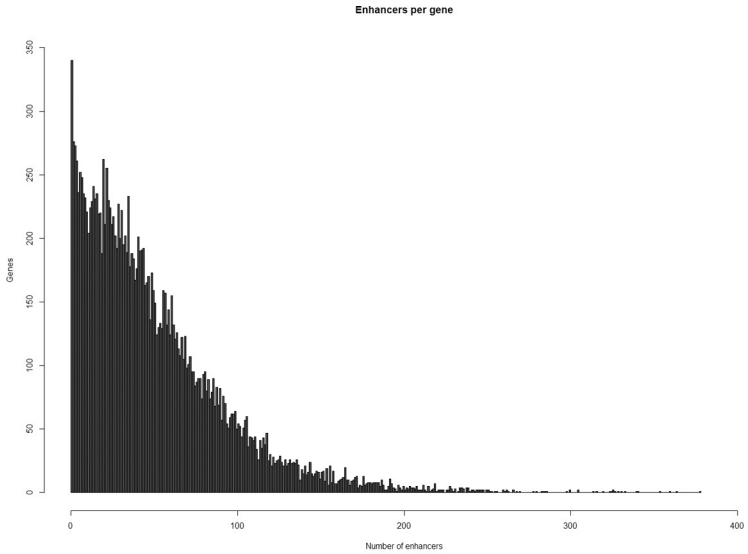
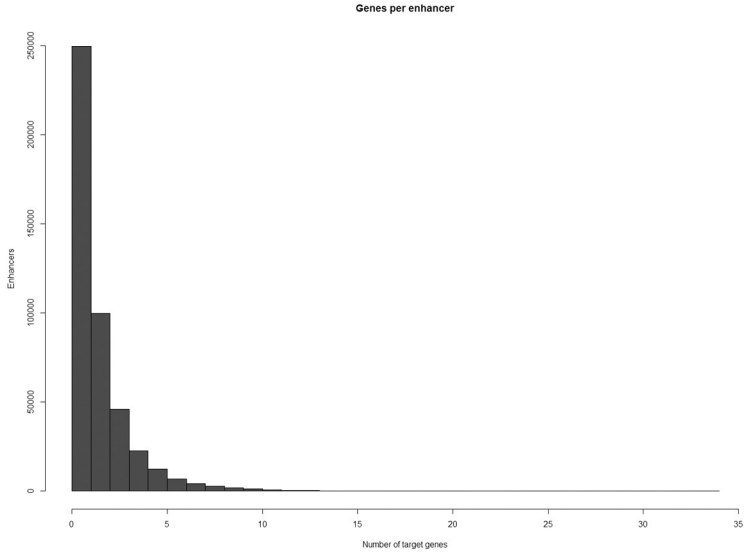


Fig 1. The number of enhancer-gene links found in each assay. This Venn diagram (not to scale) shows the number of enhancer gene-links found in each assay and each combination of assays.

https://doi.org/10.1371/journal.pone.0243791.g001


Enhancers per gene
Genes
0
50
100
150
200
250
300
350
0
100
200
300
400
Number of enhancers


Genes per enhancer
Enhancers
0
50000
100000
150000
200000
250000
0
5
10
15
20
25
30
35
Number of target genes

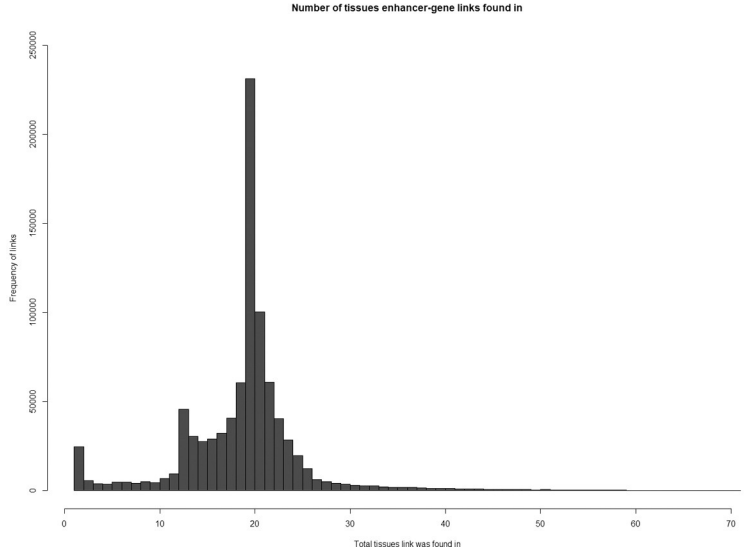

Number of tissues enhancer-gene links found in
Frequency of links
0
50000
100000
150000
200000
250000
0
10
20
30
40
50
60
70
Total tissues link was found in

Fig 2. Distributions of enhancer-gene links. A. The distribution of the number of putative target genes for each enhancer. Each bar represents the quantity of enhancers with each given value of linked genes. **B.** The distribution of the number of enhancers linked to each gene. Each bar represents the quantity of genes with each given value of linked enhancers. **C.** The distribution of the number of total tissues each enhancer-gene link was found in. Each bar represents the quantity of enhancer-gene links with each given value of tissues giving supporting evidence of the link.

https://doi.org/10.1371/journal.pone.0243791.g002

stratified by tissue and cell type, the average number of putative target genes per enhancer was largely the same, with a range of 1.00–1.95 putative target genes per enhancer from each tissue and a median of 1.31 average putative target genes per enhancer in each tissue. When

Table 4. Distribution of enhancers with each number of target genes.

Number of target genes	Number of Enhancers	Cumulative Percentage of Enhancers
1	249,820	0.56
2	99,919	0.78
3	46,024	0.88
4	22,620	0.93
5	12,404	0.96
6	6,851	0.97
7	4,097	0.98
8	2,610	0.99
9	1,757	0.99
10	1,179	0.99
11	800	>0.99
12	490	>0.99
13	289	>0.99
14	183	>0.99
15	154	>0.99
16	104	>0.99
17	59	>0.99
18	30	>0.99
19	55	>0.99
20	17	>0.99
21	35	>0.99
22	16	>0.99
23	31	>0.99
24	22	>0.99
25	15	>0.99
26	11	>0.99
27	13	>0.99
28	5	>0.99
29	6	>0.99
30	3	>0.99
31	2	>0.99
32	2	>0.99
33	3	>0.99
34	1	1.0

The most target genes an enhancer was found to have was 34, with the least amount being 1. Over 56% of enhancers linked to genes in this analysis were found to only have a single target gene. About 97% of enhancers here were found to have six or less target genes.

https://doi.org/10.1371/journal.pone.0243791.t004

Table 5. Mean number of genes linked per enhancer by assay.

Assay	Mean Putative Target Genes per Enhancer
ChIA-PET	1.17
eQTL	2.02
Linear TAD	1.96
Hierarchical TAD	1.56

Each row lists the mean number of genes that were linked to each enhancer in the assay listed.

https://doi.org/10.1371/journal.pone.0243791.t005

enhancer-gene links were stratified by assay (Table 5), the lowest average putative target genes per enhancer was found in ChIA-PET data (1.17 putative target genes per enhancer), and the highest average from eQTL data (2.02 putative target genes per enhancer).

Although the mean number of enhancers linked to each gene was 50.46 across the entire dataset, most genes had 40 or less linked enhancers (51%). However, the top 12% of genes had 100 or more linked enhancers, and a single gene (*ERI1*) was linked to 378 enhancers. Interestingly, *ERI1* is an evolutionarily conserved exoribonuclease involved in the regulation of diverse types of RNA to function as an important modulator of epigenetic gene expression [46]. When stratified by tissue and cell type, the average number of enhancers linked to each gene was lower in many tissues, with a range of 4.40–50.59 enhancers linked to each gene from each tissue and a median of 13.03 average enhancers linked to each gene in each tissue. The average across all data remains higher due to the outsized number of enhancer-gene links found in the tissues with the highest mean enhancers linked to each gene (Fig 2C). When enhancer-gene links were stratified by assay (Table 6), the lowest average enhancers linked to each gene was found in ChIA-PET data (5.40 enhancers linked to each gene), and the highest average from linear TAD data (50.88 enhancers linked to each gene). This is likely due to the fact that linking enhancers and genes together based on being located within the same TAD is the least discriminate way to link an enhancer to a gene of all methods used in PEREGRINE.

GO Biological Processes that were statistically significantly enriched for having fewer enhancers per gene than expected included several processes related to immune function. We also observed clustering of these genes and their linked enhancers in the genome. It has been previously shown that clustering occurs with genes related to immune function [47], which is thought to potentially be due to coregulation by shared enhancers [48]. Such a phenomenon could account for why these groups of genes were linked to less enhancers on average than others. We also observed several other clustered sets of genes among the groups of genes linked to fewer enhancers than expected—the olfactory receptors on chromosomes 14 and 17 and the taste receptors on chromosomes 7 and 12 (Fig 3). These clustered genes and their nearby enhancers linked by PEREGRINE may be good candidates for further research on the relationship between clustered genes and their potential coregulators.

Table 6. Mean number of enhancers linked per gene by assay.

Assay	Mean Enhancers per Gene
ChIA-PET	5.40
eQTL	26.97
Linear TAD	50.88
Hierarchical TAD	38.30

Each row lists the mean number of enhancers that were linked to each gene in the assay listed.

https://doi.org/10.1371/journal.pone.0243791.t006

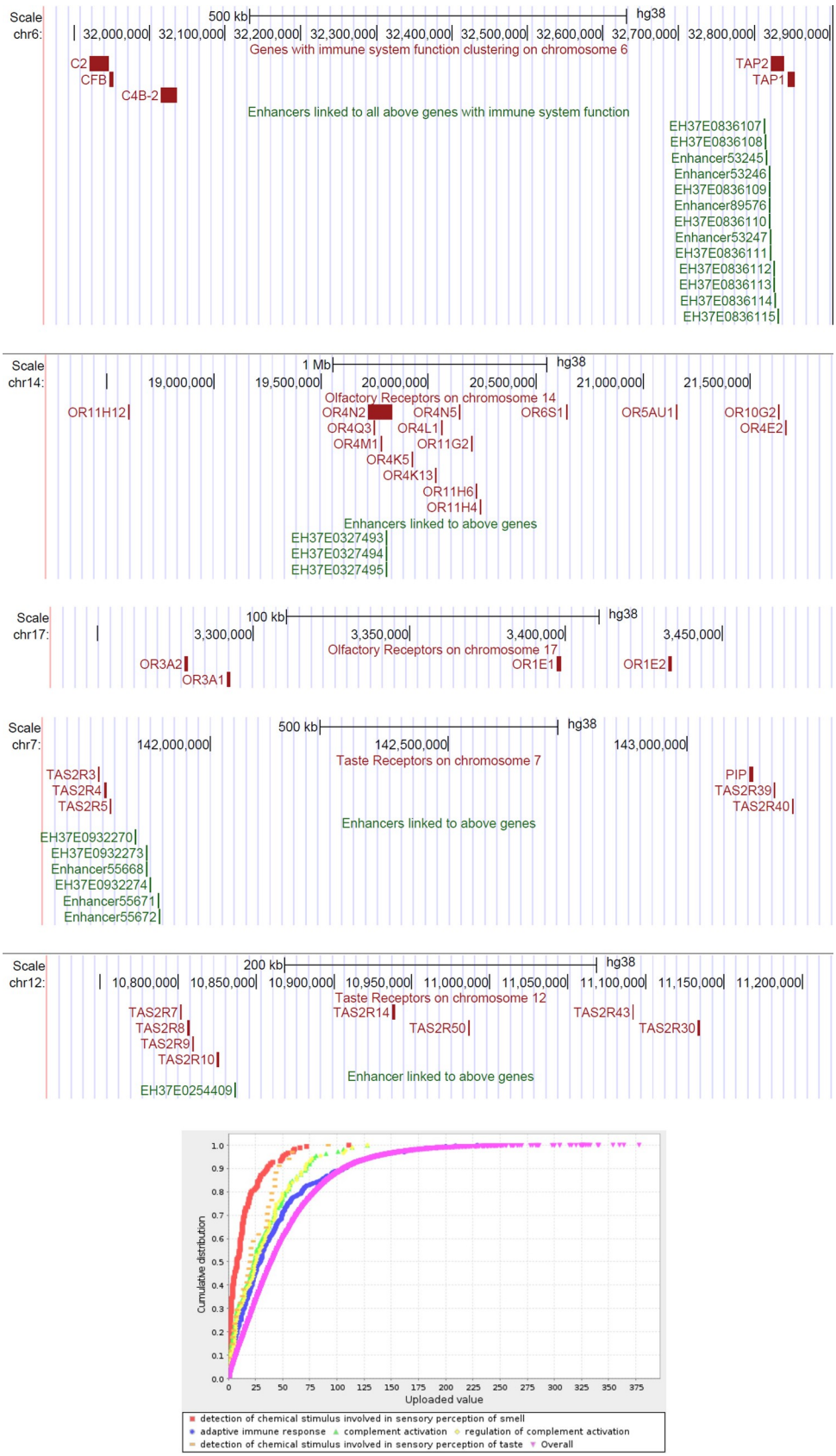

Scale
chr6:
500 kb
hg38
32,000,000
32,100,000
32,200,000
32,300,000
32,400,000
32,500,000
32,600,000
32,700,000
32,800,000
32,900,000
Genes with immune system function clustering on chromosome 6
C2
CFB
C4B-2
TAP2
TAP1
Enhancers linked to all above genes with immune system function
EH37E0836107
EH37E0836108
Enhancer53245
Enhancer53246
EH37E0836109
Enhancer89576
EH37E0836110
Enhancer53247
EH37E0836111
EH37E0836112
EH37E0836113
EH37E0836114
EH37E0836115
Scale
chr14:
1 Mb
hg38
19,000,000
19,500,000
20,000,000
20,500,000
21,000,000
21,500,000
Olfactory Receptors on chromosome 14
OR11H12
OR4N2
OR4N5
OR6S1
OR5AU1
OR10G2
OR4E2
OR4Q3
OR4L1
OR4M1
OR11G2
OR4K5
OR4K13
OR11H6
OR11H4
Enhancers linked to above genes
EH37E0327493
EH37E0327494
EH37E0327495
Scale
chr17:
100 kb
hg38
3,300,000
3,350,000
3,400,000
3,450,000
Olfactory Receptors on chromosome 17
OR3A2
OR3A1
OR1E1
OR1E2
Scale
chr7:
500 kb
hg38
142,000,000
142,500,000
143,000,000
Taste Receptors on chromosome 7
TAS2R3
TAS2R4
TAS2R5
PIP
TAS2R39
TAS2R40
Enhancers linked to above genes
EH37E0932270
EH37E0932273
Enhancer55668
EH37E0932274
Enhancer55671
Enhancer55672
Scale
chr12:
200 kb
hg38
10,800,000
10,850,000
10,900,000
10,950,000
11,000,000
11,050,000
11,100,000
11,150,000
11,200,000
Taste Receptors on chromosome 12
TAS2R7
TAS2R8
TAS2R9
TAS2R10
TAS2R14
TAS2R50
TAS2R43
TAS2R30
Enhancer linked to above genes
EH37E0254409
Cumulative distribution
Uploaded value
detection of chemical stimulus involved in sensory perception of smell
adaptive immune response
complement activation
regulation of complement activation
detection of chemical stimulus involved in sensory perception of taste
Overall

Fig 3. Clustering of genes enriched among PEREGRINE enhancer-gene links as being linked to fewer enhancers than expected. A. Genes with function related to immune system processes shown to cluster. **B.** Olfactory receptors on chromosome 14 annotated with as part of the GO biological process "detection of chemical stimulus involved in sensory perception of smell" GO:0050911 ($p<1.0E-10$). **C.** Olfactory receptors on chromosome 17 annotated with as part of the GO biological process "detection of chemical stimulus involved in sensory perception of smell" GO:0050911 ($p<1.0E-10$). **D.** Taste receptors on chromosome 7 annotated with as part of the GO biological process "detection of chemical stimulus involved in sensory perception of taste" GO:0050912 ($p<1.0E-10$). **E.** Taste receptors on chromosome 12 annotated with as part of the GO biological process "detection of chemical stimulus involved in sensory perception of taste" GO:0050912 ($p<1.0E-10$) **F.** Enrichment analysis for GO biological processes based on number of enhancers linked to each gene.

https://doi.org/10.1371/journal.pone.0243791.g003

Out of all 890,402 enhancer-gene links spanning 78 tissue and cell types, each link was found in an average of 19.30 tissues. Two enhancer-gene links were found in 71 tissue and cell types, which was the maximum for any enhancer-gene link in PEREGRINE. In order to examine the patterns between these data more easily, the number of tissues the enhancer-gene links were found in was binned in groups of 5. The enhancer-gene links found in 71 tissues were binned with the enhancer-gene links found in 66–70 tissues. A chi-squared test of association was performed between the binned total tissues enhancer-gene links were found in and the number of assays they were supported by, indicating that the number of assays the enhancer-gene links were supported by was related to the number of tissues the enhancer-genes links were found in ($p<2.2$ e-16). The Pearson correlation coefficient between these two variables was 0.39 ($p<2.2$ e-16), indicating that the more tissues an enhancer-gene link was found in, the more likely it was to be supported by multiple assays.

Utilizing the PEREGRINE data in PANTHER

A user can access the data in two ways. The first is to retrieve the enhancers linked to genes of interest. A single gene or a list of genes may be uploaded and viewed by selecting "Functional Classification viewed in gene list." This will generate a gene list page with annotations to the genes (Fig 4). Each gene is displayed with a list of its linked enhancers. Each enhancer will be a hyperlink to the enhancer detail page (Fig 5A) for that enhancer. The enhancer detail page contains information on the experimental evidence (assay, p-value, and eQTL ID if applicable) supporting each enhancer-gene link that that enhancer is involved in. The user can also click the gene identifier hyperlink to go to the gene detail page (Fig 5B). Click the "View enhancers" link to view details of all enhancers linked to this gene. The second way is to map the genetic variants to enhancers and retrieve a list of genes that are regulated by the enhancers. The user can upload a VCF file on the home page, select the VCF File as list type and check the Search Enhancer Data box.

The PANTHER website (www.pantherdb.org) allows various methods for querying the PEREGRINE predicted enhancer-gene links. On the homepage, the user may upload a VCF file with SNPs of interest and return any enhancers or genes the rsIDs map to. For example, a user may be interested in rs143969848, a rare single-nucleotide variant found in 5.4% of suspected Lynch syndrome [49] (the most common type of hereditary colon cancer [50]) patients. PANTHER maps this rsID to enhancer EH37E0652188 (Fig 5), which PEREGRINE links to just three genes (*LRRFIP2*, *MLH1*, and *EPM2AIP1*). All three are associated with Lynch syndrome in ClinVar [51], and two of the genes (*MLH1* and *EPM2AIP1*) are overlapping, with *MLH1* on the plus strand and *EPM2AIP1* on the minus strand. The shorter *EPM2AIP1* gene (8.3kb) is completely within the coordinates of the much larger *MLH1* (57.6 kb), though on the opposite strand. *LRRFIP2* is located just 1.7kb away from *MLH1*, on the minus strand. *MLH1* is a tumor suppressor gene involved in DNA mismatch repair. It is involved in the pathogenesis of Lynch syndrome as well as endometrial and colorectal carcinomas. The *LRRFIP2*

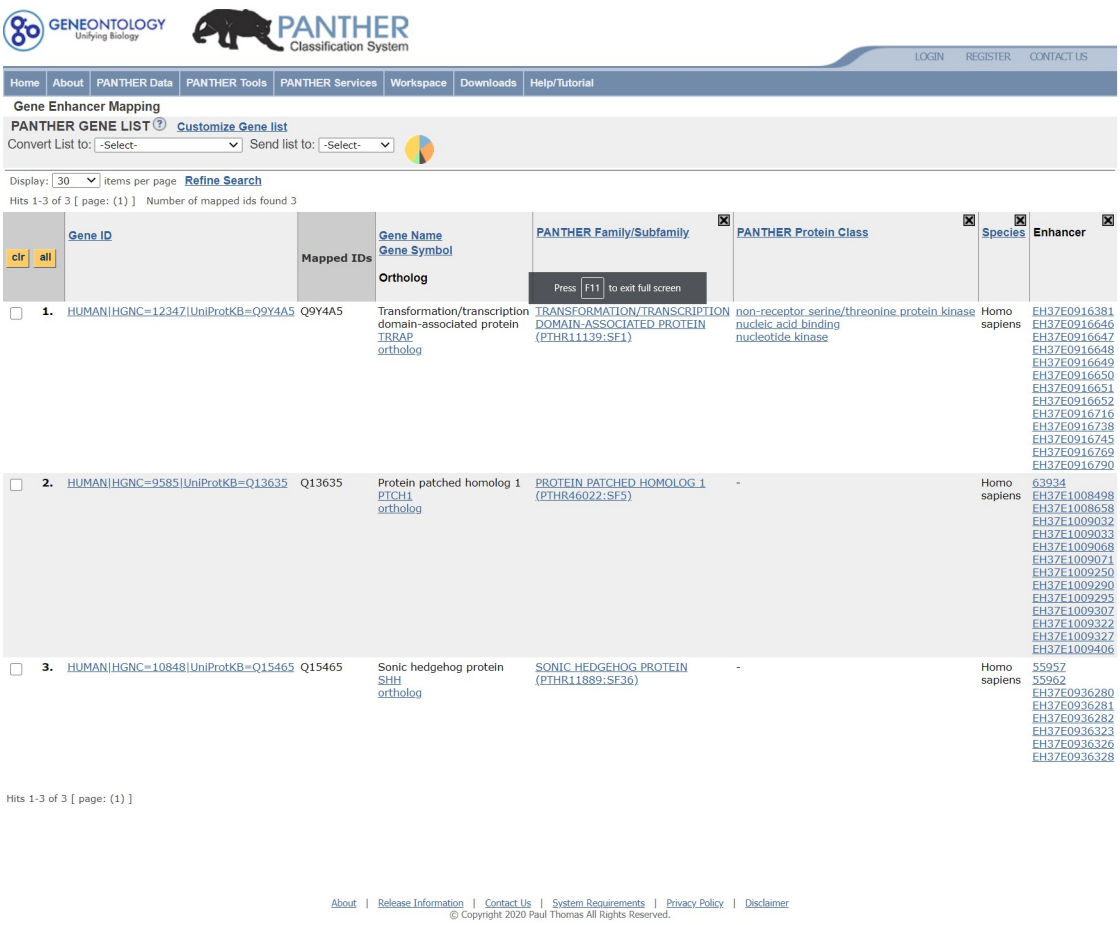


Fig 4. Viewing enhancer-gene link information on a gene list. The PANTHER website is able to take a list of genes from the user and provide a list of enhancers associated with each gene presented as a hyperlink to more information about the supporting evidence for each enhancer-gene link as well as its cell and tissue type. (Screenshot of the PANTHER website[45] published under CC BY license with permission from the original copyright holder).

https://doi.org/10.1371/journal.pone.0243791.g004

protein product binds to the cytosolic tail of TLR4, resulting in activation of nuclear factor kappa B signaling. Dysregulation of the nuclear factor kappa B signaling is a common event in many cancer types which contributes to tumor initiation and progression by driving expression of pro-proliferative/anti-apoptotic genes [50]. High expression of NF-κB has also been significantly associated with late stage colorectal cancer [52]. The function of the protein encoded by *EPM2AIP1* is not known. Therefore, it seems that *MLH1* and *LRRFIP2* and their connection to the EH37E0652188 enhancer warrant further investigation. While little is known about the distal regulation of *LRRFIP2*, Liu et al [49] recently showed that a 1.8kb region located 35kb upstream of *MLH1* interacted with the *MLH1* promoter, displayed enhancer function in luciferase reporter assays, and statistically significantly altered the expression of *MLH1* using CRISPR-Cas9-mediated deletion of endogenous regions. This region also includes a CTCF-binding motif, which has been shown to disrupt enhancer activity in SW620 colorectal carcinoma cells [49]. Also within this region lies the entire 770bp enhancer EH37E0652188, the enhancer that PEREGRINE predicted as regulating *MLH1*.

The SNP rs2144300 has been statistically significantly associated with HDL cholesterol levels in humans [53]. PANTHER maps this rsID to enhancer EH37E0145522, which

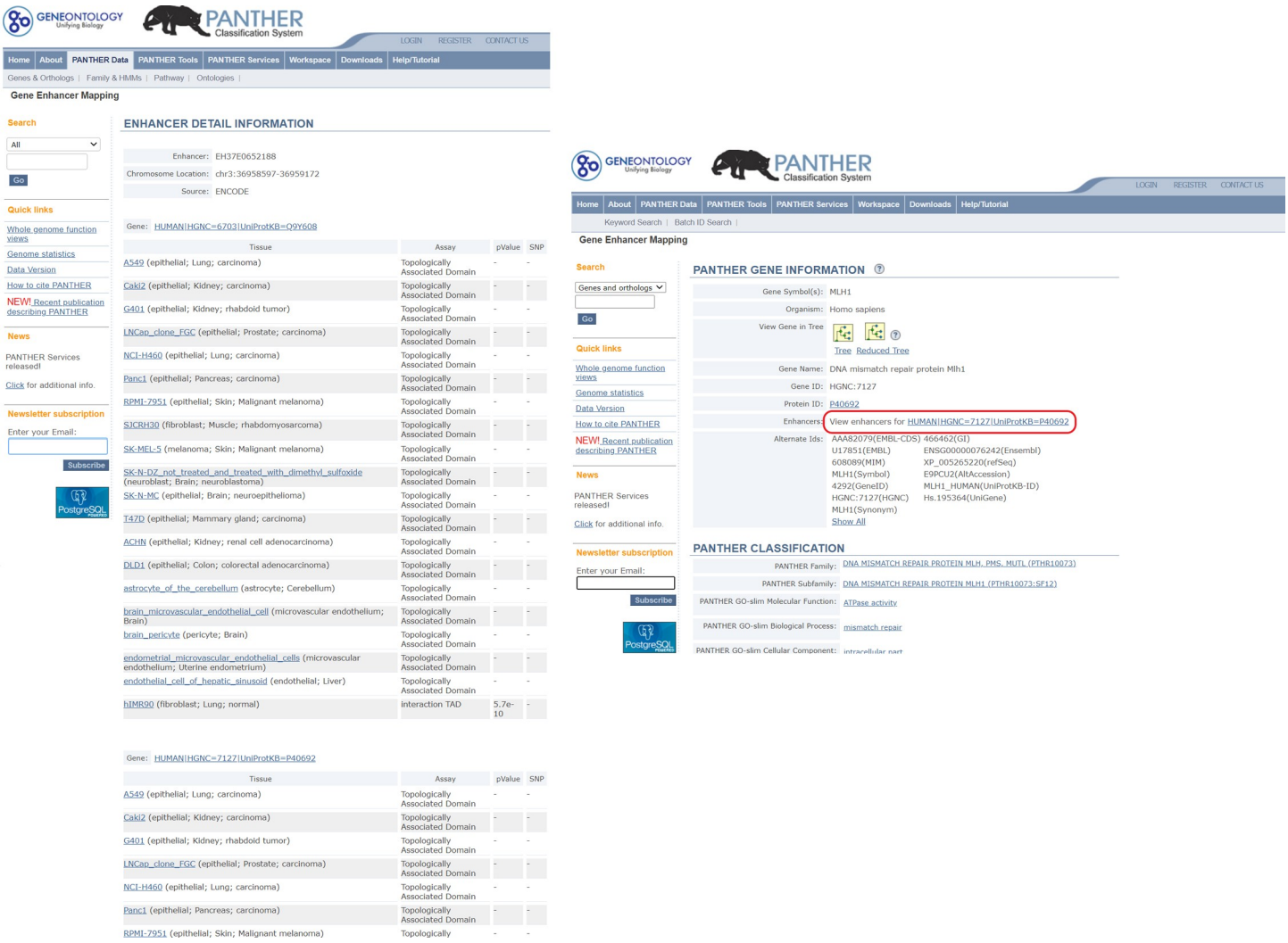


Fig 5. The schema of PEREGRINE within PANTHER. This schema shows what information will be made available within the PANTHER website. **A.** Gene detail page. Each gene in PANTHER has a gene detail page which now includes an Enhancers section (circled in red) with a link to view all enhancers associated with that gene by PEREGRINE. **B.** Enhancer detail page. Each enhancer has an enhancer detail page in PANTHER with the enhancer's ID, genomic location, original source, and detailed information on the experimental evidence used by PERGERINE to link that enhancer each of its associated genes. (Screenshots of the PANTHER website[45] published under CC BY license with permission from the original copyright holder).

https://doi.org/10.1371/journal.pone.0243791.g005

PEREGRINE links to just one gene, *GALNT2* (S1 Fig). This variant is found within the first intron of *GALNT2*, a gene strongly associated to HDL cholesterol levels [54]. Roman et al [55] explored the SNPs at this locus to reveal a 780-bp segment containing rs4846913, rs2144300, and rs6143660 that displayed allelic differences in regulatory enhancer activity in luciferase assays. They also showed differential CEBPB binding to rs4846913 using electrophoretic mobility shift assays which they confirmed occurred in a native chromatin context with ChIP assays in two liver cancer cell lines. Allelic-expression-imbalance assays performed with RNA from primary human hepatocyte samples and expression-quantitative-trait-locus (eQTL) data confirmed that these SNPs are associated with increased *GALNT2* expression. They proposed that at minimum, rs4846913 and rs2281721 play key roles in influencing *GALNT2* expression at this locus. Cavalli et al [54] showed that rs4846913 and the neighboring rs2144300 displayed allele specific enhancer activity and proposed that events occurring at these SNPs influence the

transcription levels of *GALNT2*. All three SNPs (rs4846913, rs2144300, and rs6143660) in the 780-bp segment validated by Roman et al fall within EH37E0145522, which is only 813bp long. The other SNP, rs2281721, did not map to any enhancers in the PEREGRINE set.

The colorectal cancer risk-associated variant rs2238126 is located within an intron of *ETV6*, an ETS family transcription factor. PANTHER maps this variant to a single enhancer, EH37E0254775. Wang et al [56] showed that the G allele of rs2238126 reduces the binding affinity of MAX, a transcription factor thought to enhance transcription of *ETV6*, resulting in significantly lower mRNA levels of *ETV6*. They proposed that *ETV6* gene expression is regulated by the SNP rs2238126 and that the rs2238126 G allele is associated with an increased risk of colorectal cancer because of decreased transcription factor MAX binding, resulting in downregulating *ETV6* expression. They also tested a putative enhancer region centering rs2238126 (1kb in length) for enhancer activity using luciferase assays in HCT116 and SW480 cells and found that the A allele of rs2238126 conferred statistically significantly higher luciferase expression in both cell types as compared to the G allele and as compared to the vector with no enhancer region. Though rs2238126 mapped only to enhancer EH37E0254775 (647 bp), this longer putative enhancer region also mapped to enhancer 12808 (292 bp), which overlaps almost completely with EH37E0254775.

Previously, a user could upload variants of interest in a VCF file to the PANTHER homepage for analysis, and PANTHER would call the variants to their nearest genes according to a gene flanking region distance set by the user. This means that if a user was looking for variants within an enhancer that could have a regulatory relationship with a gene, those enhancer variants would need to be within the flanking region of that gene to appear in the PANTHER gene list output as associated with that gene. Using the PEREGRINE data integrated into PANTHER, the user can now select a search for variants called to the PEREGRINE enhancer regions and obtain a gene list of all of the genes associated with those variant-containing enhancers. This is especially important when considering that of the PEREGRINE enhancer-gene links, only 12% involve an enhancer that is within 20kb of its putative target gene. Indeed, 42% of the PEREGRINE enhancer-gene links comprise a gene and an enhancer that are at least 100kb apart, thus greatly expanding the user's ability to examine putative regulatory variants using the PANTHER framework.

Discussion

Enhancers are vital regulatory elements that increase transcription of their target genes many times over. They play a key role in development and are implicated in many common diseases, including many cancers. Determining which genes are the target genes of specific enhancers is key to informing to what extent and how enhancers contribute to disease pathology. Although significant efforts have been made to successfully elucidate enhancer-gene links at the bench, these experimental findings represent only a small fraction of all enhancer-gene links. Additionally, these results are not automatically deposited into any central repository of known enhancer-gene links. Thus, utilizing these data on a large scale is laborious. A high throughput method of predicting enhancer-gene links with good ability is desirable. High throughput experimental methods have helped in this direction, with good ability to predict enhancer-gene contacts in the cell types the experiments are conducted in. However, these methods are relatively new and therefore data is not yet widely available in a large range of cell types. Additionally, a bench laboratory is necessary to perform these experiments in new cell types, which is a limiting factor for many analytical groups. Some computationally predicted enhancer-gene link databases have been developed, which do not rely on new experimental data and instead use publicly available data to compile predicted enhancer-gene links. However, these

methods are limited in their accessibility to the scientific community. Most do not offer an up to date bulk downloadable option for all of the data to be examined *en masse*, but instead only make the complete data for each enhancer-gene link available to the end user via individual webpages for each enhancer, gene, or enhancer-gene link. There are also varying degrees of the amount of information available regarding what specific evidence in which cell types supports each enhancer-gene link, which may be of great interest to the end user. The PEREGRINE enhancer-gene links, made available via the PANTHER website, represent a comprehensive set of enhancer-gene links with accompanying experimental evidence available via bulk download, and also searchable by genomic region and putative target gene(s) of interest.

In order to assay the enhancer and promoter binding of the transcription machinery that enhancers are known to recruit to their target genes, ChIA-PET data was used to identify pairs of regions containing enhancers and promoters bound to the target protein RNA Polymerase II. Enhancers are often located within the same topologically associated domain as their target genes, so Hi-C data was used to link enhancers to genes within the same topologically associated domain, but these links were only recorded if they supported an enhancer-gene link that was already found in another experiment. This was done in order to reduce the number of false positives likely to be incurred by linking every enhancer to every gene within the same topologically associated domain. Since enhancers are thought to function through achieving close proximity with their target genes' promoters, enhancer-gene links were also taken from Hi-C data where the hierarchy of contacts within topologically associated domains was captured. Enhancers were screened for eQTL and linked to any gene which showed statistically significant differences in expression due to that eQTL. Together, these data assay the characteristics of enhancer regulation of genes in 78 cell and tissue types to yield 890,402 enhancer-gene links. On average, each gene was linked to 50 enhancers while each enhancer was linked to 2 putative target genes. Enhancer-gene links which were found in many tissues were more likely to be supported by more assays ($p<2.2e-16$) according to a chi-squared test of association. The Pearson correlation coefficient between the number of assays supporting an enhancer-gene link and the number of tissues and cell types that link was found in was 0.39 ($p<2.2e-16$), indicating that enhancer-gene links supported by more assays are more likely to be found within a wider range of cell and tissue types in these data.

Although each gene was linked to an average of 50 enhancers using the PEREGRINE enhancer set, there is some overlap among the PEREGRINE enhancers (Table 2). In order to determine how much this might be influencing the average enhancers linked to each gene, analysis was redone using only the enhancers taken from ENCODE, as these are all mutually exclusive enhancer elements and also account for over 91% of the PEREGRINE enhancer set. Restricting only to the ENCODE enhancers, each gene was linked to an average of 45 enhancers. Thus, the modest overlap between enhancers in the PEREGRINE set does not dramatically change the average number of enhancers linked to each gene.

Two enhancer-gene link prediction databases employing strategies most similar to those used in PEREGRINE, GeneHancer and HACER, were compared to the PEREGRINE enhancer-gene link database. An analysis of PEREGRINE enhancers compared to the enhancer sets from HACER and GeneHancer provided in their Data Download sections show that PEREGRINE enhancers are more comprehensive than either of these sets (Fig 6). Of the 1,085,794 enhancers in the PEREGRINE set, only 268,811 (24.8%) overlap with the enhancers in the HACER set. In contrast, of the 1,685,398 HACER enhancers, 1,644,428 (97.6%) have overlap with PEREGRINE enhancers. Of the 1,085,794 enhancers in the PEREGRINE set, 461,202 (42.5%) overlap with the enhancers in the GeneHancer set. Conversely, of the 217,695 GeneHancer enhancers that successfully converted to hg19, 180,743 (83.0%) have overlap with

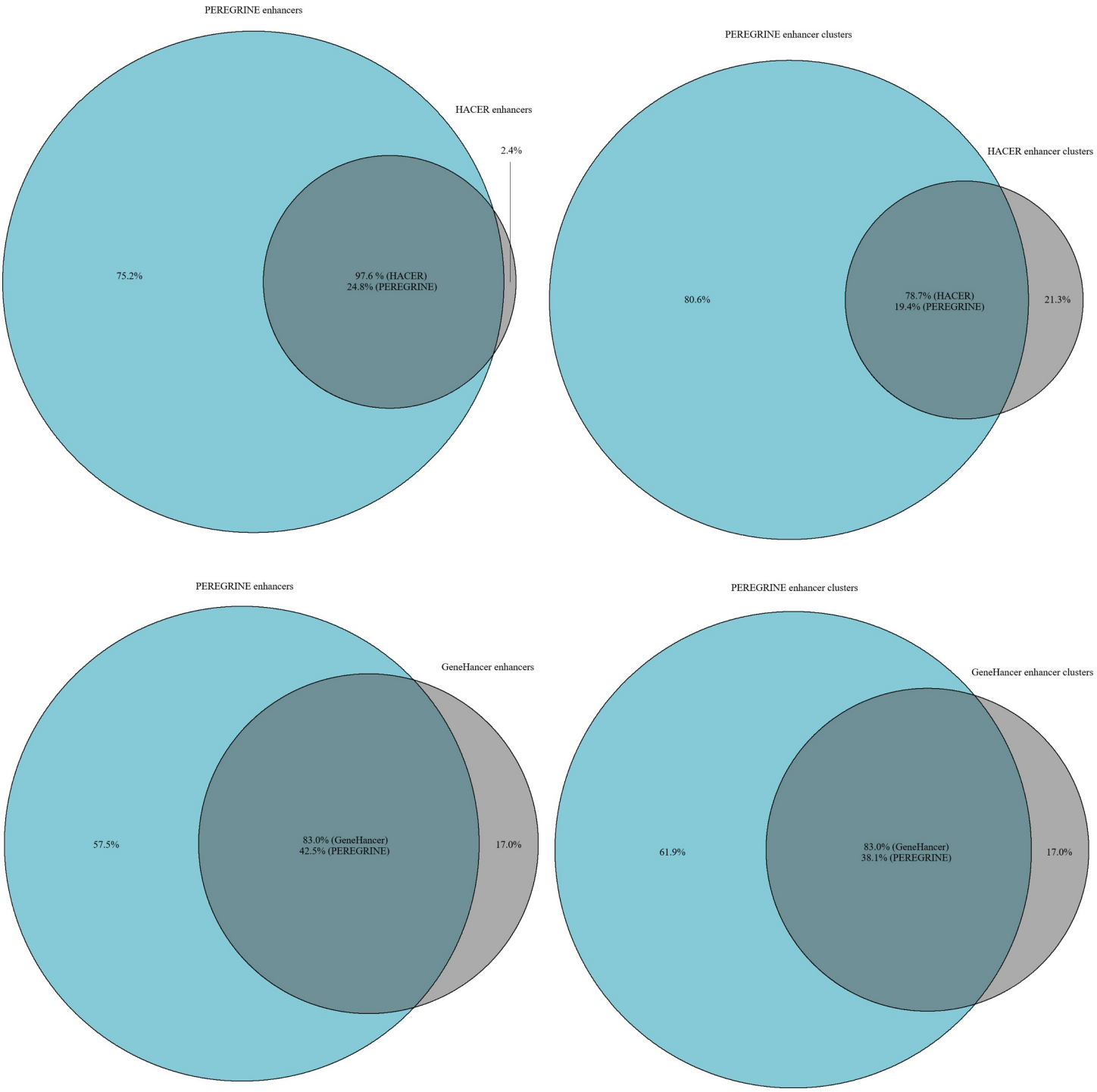


Fig 6. Comparison of PEREGRINE enhancers with HACER and GeneHancer enhancers. Percentages are in terms of the enhancer set that the percentage labels are in. **a.** HACER and PEREGRINE enhancer overlap in terms of percentages of numbers of enhancers from each source. **b.** HACER and PEREGRINE enhancer clusters overlapped in terms of percentages of numbers of clusters from each source. **c.** GeneHancer and PEREGRINE enhancer overlap in terms of percentages of numbers of enhancers from each source. **d.** GeneHancer and PEREGRINE enhancer clusters overlapped in terms of percentages of numbers of clusters from each source.

https://doi.org/10.1371/journal.pone.0243791.g006

PEREGRINE enhancers. Even when analyses require that overlap between two enhancers require at least one of the elements to overlap with the other at least 50%, these numbers

remain stable (HACER shifts from 24.8% and 97.6% to 23.7% and 97.4% respectively; GeneHancer shifts from 42.5% and 83.0% to 40.0% and 81.2% respectively). The reason for such overlap is most likely due to PEREGRINE using some enhancer data from sources that are common among both databases (e.g. VISTA, FANTOM). However, the PEREGRINE enhancer set is larger and not well captured by either HACER or GeneHancer enhancer sets in part because enhancer data from various sources with mostly non-overlapping enhancer coordinates was utilized to make up the PEREGRINE enhancer set. In fact, the enhancers in the PEREGRINE exhibit less internal overlap than the enhancers in the HACER set.

Using the cluster and merge commands from the bedtools suite to output clusters of partially overlapping elements, 1,085,794 PEREGRINE enhancers clustered to 1,002,071 non-overlapping enhancer regions (PEREGRINE clusters). Performing the same analysis with HACER enhancers resulted in 104,078 clusters of non-overlapping enhancer regions from 1,685,398 enhancers. This high instance of overlapping among enhancers in the HACER set is in part due to the fact that HACER reports cell-type specificity for each enhancer and names each enhancer from each cell type uniquely, even if there is very high overlap between them (which possibly indicates that these elements sometimes refer to the same enhancer in two different cell lines). These likely refer to the same enhancer region, but the data from each cell line gives slightly different coordinates for the enhancer region. GeneHancer enhancers were almost perfectly unique, resulting in non-overlapping clusters nearly identical to their enhancers with an average length of 1,572 bp.

Another interesting difference between PEREGRINE enhancers compared to HACER and GeneHancer enhancers relates to the average length of the enhancers in each set. The average length of HACER and GeneHancer enhancers are much longer than the average length of PEREGRINE enhancers. The average length of PEREGRINE enhancers is 422 bp. The average length of PEREGRINE clusters is 434 bp. The average length of HACER enhancers is 713 bp. The average length of HACER clusters is 3,440 bp. The average length of GeneHancer enhancers is 1,572 bp. This indicates that PEREGRINE enhancers are more closely scaled to the length that most enhancers are thought to be, which is closer to hundreds of base pairs than to thousands.

These databases, including the PEREGRINE enhancer-gene links, are limited by their lack of statistical validation. Due to the lack of a gold standard database of enhancer-gene links for new predictions to be judged by, it is nearly impossible to statistically validate predicted enhancer-gene links which have been generated across many assays and cell types. Although there are multiple instances of experimentally validated enhancer-gene links from years of benchwork being captured by prediction databases, little is known about the magnitude of how many spurious enhancer-gene links are included along with the legitimate ones in these sets of predictions. Future work will be focused on the generation of a statistically validated enhancer-gene link score. Such a score would allow researchers to see which enhancer-gene links are reported with the highest confidence, which would be a valuable addition to the PEREGRINE enhancer-gene links.

Conclusions

Enhancers are specialized regions of the genome that control target gene expression levels. They can occur at great distances from their target gene, and loop in complicated structures to accomplish this. Determining which enhancers interact with target genes is a question the field has been trying to address for several years, and many experimental techniques to connect them have significant drawbacks in computational difficulty, feasibility, or reproducibility. Here, we have incorporated publicly available enhancer data from ENCODE, Ensembl,

FANTOM and VISTA, and experimental data from ChIA-PET, eQTL, and Hi-C assays across 78 cell and tissue types to generate an enhancer-gene link database called PEREGRINE. The database provides links between 449,627 enhancers and 17,643 protein-coding genes. The data have been incorporated into the PANTHER Classification System (www.pantherdb.org) for gene and variant search, and are available for download at the PEREGRINE website (www.peregrineproj.org). This tool will allow biologists to leverage this compendium of enhancer-gene link knowledge to answer fundamental questions about development, disease, and homeostatic cellular regulation.

Supporting information

S1 Fig. The enhancer-gene link between EH37E0145522 and *GALNT2*. a. Gene detail page. b. Enhancer detail page.
(TIF)

S1 File. URLs for data downloads. Download URLs for enhancer sets from original sources.
(PDF)

Acknowledgments

The authors thank Drs. Graham Casey, David Conti, Ite Offringa and Kimberly Siegmund for helpful discussion. We thank Tremayne Mushayahama and Laurent-Philippe Albou for the help on the PEREGRINE website.

Author Contributions

Conceptualization: Huaiyu Mi.

Formal analysis: Caitlin Mills.

Funding acquisition: Huaiyu Mi.

Investigation: Caitlin Mills, Huaiyu Mi.

Methodology: Caitlin Mills.

Resources: Paul D. Thomas.

Software: Caitlin Mills, Anushya Muruganujan, Dustin Ebert.

Supervision: Huaiyu Mi.

Validation: Caitlin Mills, Huaiyu Mi.

Writing – original draft: Caitlin Mills.

Writing – review & editing: Crystal N. Marconett, Juan Pablo Lewinger, Paul D. Thomas, Huaiyu Mi.

References

1. Gao T, He B, Liu S, Zhu H, Tan K, Qian J. EnhancerAtlas: a resource for enhancer annotation and analysis in 105 human cell/tissue types. Bioinformatics. 2016; 32(23):3543–3551. https://doi.org/10.1093/bioinformatics/btw495 PMID: 27515742
2. Ong CT, Corces VG. Enhancer function: new insights into the regulation of tissue-specific gene expression. Nat Rev Genet. 2011; 12(4):283–293. https://doi.org/10.1038/nrg2957 PMID: 21358745
3. Banerji J, Rusconi S, Schaffner W. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. Cell. 1981; 27(2 Pt 1):299–308. https://doi.org/10.1016/0092-8674(81)90413-x PMID: 6277502

4. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489(7414):57–74. https://doi.org/10.1038/nature11247 PMID: 22955616

5. The ENCODE Project Consortium. A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol. 2011; 9(4):e1001046.

6. Yao L, Berman BP, Farnham PJ. Demystifying the secret mission of enhancers: linking distal regulatory elements to target genes. Crit Rev Biochem Mol Biol. 2015; 50(6):550–573. https://doi.org/10.3109/10409238.2015.1087961 PMID: 26446758

7. Heintzman ND, Hon GC, Hawkins RD, et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature. 2009; 459(7243):108–112. https://doi.org/10.1038/nature07829 PMID: 19295514

8. Hnisz D, Abraham BJ, Lee TI, et al. Super-enhancers in the control of cell identity and disease. Cell. 2013; 155(4):934–947. https://doi.org/10.1016/j.cell.2013.09.053 PMID: 24119843

9. Corradin O, Scacheri PC. Enhancer variants: evaluating functions in common disease. Genome Med. 2014; 6(10):85. https://doi.org/10.1186/s13073-014-0085-3 PMID: 25473424

10. Gasperini M, Tome JM, Shendure J. Towards a comprehensive catalogue of validated and target-linked human enhancers. Nat Rev Genet. 2020. https://doi.org/10.1038/s41576-019-0209-0 PMID: 31988385

11. Visel A, Rubin EM, Pennacchio LA. Genomic views of distant-acting enhancers. Nature. 2009; 461(7261):199–205. https://doi.org/10.1038/nature08451 PMID: 19741700

12. Schoenfelder S, Javierre BM, Furlan-Magaril M, Wingett SW, Fraser P. Promoter Capture Hi-C: High-resolution, Genome-wide Profiling of Promoter Interactions. J Vis Exp. 2018(136). https://doi.org/10.3791/57320 PMID: 30010637

13. Fishilevich S, Nudel R, Rappaport N, et al. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database (Oxford). 2017;2017. https://doi.org/10.1093/database/bax028 PMID: 28605766

14. Wang J, Dai X, Berry LD, Cogan JD, Liu Q, Shyr Y. HACER: an atlas of human active enhancers to interpret regulatory variants. Nucleic Acids Res. 2019; 47(D1):D106–D112. https://doi.org/10.1093/nar/gky864 PMID: 30247654

15. Wang Z, Zhang Q, Zhang W, et al. HEDD: Human Enhancer Disease Database. Nucleic Acids Res. 2018; 46(D1):D113–D120. https://doi.org/10.1093/nar/gkx988 PMID: 29077884

16. Gao T, Qian J. EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. Nucleic Acids Res. 2020; 48(D1):D58–D64. https://doi.org/10.1093/nar/gkz980 PMID: 31740966

17. Jiang Y, Qian F, Bai X, et al. SEdb: a comprehensive human super-enhancer database. Nucleic Acids Res. 2019; 47(D1):D235–D243. https://doi.org/10.1093/nar/gky1025 PMID: 30371817

18. Mi H, Muruganujan A, Casagrande JT, Thomas PD. Large-scale gene function analysis with the PANTHER classification system. Nat Protoc. 2013; 8(8):1551–1566. https://doi.org/10.1038/nprot.2013.092 PMID: 23868073

19. Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Res. 2019; 47(D1):D419–D426. https://doi.org/10.1093/nar/gky1038 PMID: 30407594

20. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000; 25(1):25–29. https://doi.org/10.1038/75556 PMID: 10802651

21. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res. 2019; 47(D1):D330–D338. https://doi.org/10.1093/nar/gky1055 PMID: 30395331

22. Jassal B, Matthews L, Viteri G, et al. The reactome pathway knowledgebase. Nucleic Acids Res. 2020; 48(D1):D498–D503. https://doi.org/10.1093/nar/gkz1031 PMID: 31691815

23. Mi H, Thomas P. PANTHER pathway: an ontology-based pathway database coupled with data analysis tools. Methods Mol Biol. 2009; 563:123–140. https://doi.org/10.1007/978-1-60761-175-2_7 PMID: 19597783

24. Zerbino DR, Achuthan P, Akanni W, et al. Ensembl 2018. Nucleic Acids Res. 2018; 46(D1):D754–D761. https://doi.org/10.1093/nar/gkx1098 PMID: 29155950

25. Fraser J, Ferrai C, Chiariello AM, et al. Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. Mol Syst Biol. 2015; 11(12):852. https://doi.org/10.15252/msb.20156492 PMID: 26700852

26. Visel A, Minovitsky S, Dubchak I, Pennacchio LA. VISTA Enhancer Browser—a database of tissue-specific human enhancers. Nucleic Acids Res. 2007; 35(Database issue):D88–92. https://doi.org/10.1093/nar/gkl822 PMID: 17130149

27. Consortium G. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013; 45(6):580–585. https://doi.org/10.1038/ng.2653 PMID: 23715323

28. Ron G, Globerson Y, Moran D, Kaplan T. Promoter-enhancer interactions identified from Hi-C data using probabilistic models and hierarchical topological domains. Nat Commun. 2017; 8(1):2237. https://doi.org/10.1038/s41467-017-02386-3 PMID: 29269730

29. Ghavi-Helm Y, Klein FA, Pakozdi T, et al. Enhancer loops appear stable during development and are associated with paused polymerase. Nature. 2014; 512(7512):96–100. https://doi.org/10.1038/nature13417 PMID: 25043061

30. Dixon JR, Selvaraj S, Yue F, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012; 485(7398):376–380. https://doi.org/10.1038/nature11082 PMID: 22495300

31. Rao SS, Huntley MH, Durand NC, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014; 159(7):1665–1680. https://doi.org/10.1016/j.cell.2014.11.021 PMID: 25497547

32. Marsman J, Horsfield JA. Long distance relationships: enhancer-promoter communication and dynamic gene transcription. Biochim Biophys Acta. 2012; 1819(11–12):1217–1227. https://doi.org/10.1016/j.bbagrm.2012.10.008 PMID: 23124110

33. Blackwood EM, Kadonaga JT. Going the distance: a current view of enhancer action. Science. 1998; 281(5373):60–63. https://doi.org/10.1126/science.281.5373.60 PMID: 9679020

34. Bulger M, Groudine M. Looping versus linking: toward a model for long-distance gene activation. Genes Dev. 1999; 13(19):2465–2477. https://doi.org/10.1101/gad.13.19.2465 PMID: 10521391

35. de Laat W, Klous P, Kooren J, et al. Three-dimensional organization of gene expression in erythroid cells. Curr Top Dev Biol. 2008; 82:117–139. https://doi.org/10.1016/S0070-2153(07)00005-1 PMID: 18282519

36. Attema JL, Bert AG, Lim YY, et al. Identification of an enhancer that increases miR-200b~200a~429 gene expression in breast cancer cells. PLoS One. 2013; 8(9):e75517. https://doi.org/10.1371/journal.pone.0075517 PMID: 24086551

37. Fullwood MJ, Ruan Y. ChIP-based methods for the identification of long-range chromatin interactions. J Cell Biochem. 2009; 107(1):30–39. https://doi.org/10.1002/jcb.22116 PMID: 19247990

38. Nica AC, Dermitzakis ET. Expression quantitative trait loci: present and future. Philos Trans R Soc Lond B Biol Sci. 2013; 368(1620):20120362. https://doi.org/10.1098/rstb.2012.0362 PMID: 23650636

39. Ward LD, Kellis M. Interpreting noncoding genetic variation in complex traits and human disease. Nat Biotechnol. 2012; 30(11):1095–1106. https://doi.org/10.1038/nbt.2422 PMID: 23138309

40. Maurano MT, Humbert R, Rynes E, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012; 337(6099):1190–1195. https://doi.org/10.1126/science.1222794 PMID: 22955828

41. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26(6):841–842. https://doi.org/10.1093/bioinformatics/btq033 PMID: 20110278

42. Phanstiel DH, Boyle AP, Heidari N, Snyder MP. Mango: a bias-correcting ChIA-PET analysis pipeline. Bioinformatics. 2015; 31(19):3092–3098. https://doi.org/10.1093/bioinformatics/btv336 PMID: 26034063

43. Apache Solr. 2020; https://lucene.apache.org/solr/.

44. Burke K, Conroy K, Horn R, Stratton F, Binet G. Flask-RESTful. 2020; https://flask-restful.readthedocs.io/en/latest/.

45. Mi H, Muruganujan A, Huang X, et al. Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). Nat Protoc. 2019; 14(3):703–721. https://doi.org/10.1038/s41596-019-0128-8 PMID: 30804569

46. Thomas MF, L'Etoile ND, Ansel KM. Eri1: a conserved enzyme at the crossroads of multiple RNA-processing pathways. Trends Genet. 2014; 30(7):298–307. https://doi.org/10.1016/j.tig.2014.05.003 PMID: 24929628

47. Makino T, McLysaght A. Interacting gene clusters and the evolution of the vertebrate immune system. Mol Biol Evol. 2008; 25(9):1855–1862. https://doi.org/10.1093/molbev/msn137 PMID: 18573844

48. West AG, Gaszner M, Felsenfeld G. Insulators: many functions, many mechanisms. Genes Dev. 2002; 16(3):271–288. https://doi.org/10.1101/gad.954702 PMID: 11825869

49. Liu Q, Thoms JAI, Nunez AC, et al. Disruption of a -35 kb Enhancer Impairs CTCF Binding and. Clin Cancer Res. 2018; 24(18):4602–4611. https://doi.org/10.1158/1078-0432.CCR-17-3678 PMID: 29898989

50. Chen J, Stark LA. Aspirin Prevention of Colorectal Cancer: Focus on NF-κB Signalling and the Nucleolus. Biomedicines. 2017; 5(3). https://doi.org/10.3390/biomedicines5030043 PMID: 28718829

51. Landrum MJ, Lee JM, Riley GR, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014; 42(Database issue):D980–985. https://doi.org/10.1093/nar/gkt1113 PMID: 24234437

52. Wu D, Wu P, Zhao L, et al. NF-κB Expression and Outcomes in Solid Tumors: A Systematic Review and Meta-Analysis. Medicine (Baltimore). 2015; 94(40):e1687. https://doi.org/10.1097/MD.0000000000001687 PMID: 26448015

53. Murray A, Cluett C, Bandinelli S, et al. Common lipid-altering gene variants are associated with therapeutic intervention thresholds of lipid levels in older people. Eur Heart J. 2009; 30(14):1711–1719. https://doi.org/10.1093/eurheartj/ehp161 PMID: 19435741

54. Cavalli M, Pan G, Nord H, Wadelius C. Looking beyond GWAS: allele-specific transcription factor binding drives the association of GALNT2 to HDL-C plasma levels. Lipids Health Dis. 2016; 15:18. https://doi.org/10.1186/s12944-016-0183-x PMID: 26817450

55. Roman TS, Marvelle AF, Fogarty MP, et al. Multiple Hepatic Regulatory Variants at the GALNT2 GWAS Locus Associated with High-Density Lipoprotein Cholesterol. Am J Hum Genet. 2015; 97(6):801–815. https://doi.org/10.1016/j.ajhg.2015.10.016 PMID: 26637976

56. Wang M, Gu D, Du M, et al. Common genetic variation in ETV6 is associated with colorectal cancer susceptibility. Nat Commun. 2016; 7:11478. https://doi.org/10.1038/ncomms11478 PMID: 27145994