HumanBase Documentation

Release 1.0

User Guide

Aug 09, 2023

Contents

CHAPTER 1

About

HumanBase is a “one stop shop” for biological and biomedical researchers interested in data-driven predictions of gene expression, function, regulation, and interactions in human, particularly in the context of specific cell types/tissues, development, and human disease.

Data-driven integrative analyses are especially powerful because they reach beyond “known biological knowledge” represented in the biological literature to identify novel associations that are not biased toward well-studied areas of biomedical research. Thus, carefully designed algorithms can drive the development of experimentally testable hypotheses, enabling deeper understanding of basic biology at the molecular level, pathophysiology, and paving the way to therapy and drug development.

CHAPTER 2

Example use case

A researcher who studies the role of the immune system and inflammation in chronic kidney disease wants to identify candidate genes for these disorders. Unfortunately, as with most specific disease contexts outside of cancer, few datasets are available for these diseases, none are focused on the role of inflammation or the immune system, and no dataset is specific to her cell-lineage of interest. Even identifying which genes are expressed in the cell type relevant to glomerular disease (podocytes) is currently impossible as this cell lineage cannot be isolated for high-throughput experiments in human.

Using HumanBase, she will be able to examine data-driven predictions of genes expressed in the podocyte cells and analyze predicted functional and mechanistic networks specific to the kidney glomerulus. She could also provide the system with a list of relevant GWAS or family-based study results and the system will reprioritize these results based on the relevant functional maps. She will be able to iteratively refine this analysis by limiting the data used in the integration only to kidney datasets or by integrating her own data in the analysis.

CHAPTER 3

Licensing

All data in HumanBase are freely available under a CC-BY 4.0 license. Please give appropriate credit, provide a link to the license, and indicate if changes were made.

CHAPTER 4

Who are we?

HumanBase is actively developed by the Genomics group at the Flatiron Institute .

CHAPTER 5

Help topics

5.1 Getting Started

5.2 Functional Networks

In order to leverage the vast collections of raw, noisy genomic data, they must be integrated, summarized, and presented in a biologically informative manner. We provide a means of mining tens of thousands of whole-genome experiments by way of functional interaction networks. Each interaction network represents a body of data, probabilistically weighted and integrated, focused on a particular biological question. These questions can include, for example, the function of a gene, the relationship between two pathways, or the processes disrupted in a genetic disorder. (Huttenhower, et. al 2008)

5.2.1 Method

Briefly, functional integration relies on the construction of process-specific functional relationship networks. These are interaction networks in which each node represents a gene, each edge a functional relationship, and an edge between two genes is probabilistically weighted based on experimental evidence relating to those genes. We integrate evidence from many data sets, with each data set weighted in a process-specific manner.

One naïve Bayesian classifier is trained per biological area of interest (e.g. a tissue, or a specific biological process), using the appropriate gold standard for the biological context in addition to one global process-unaware classifier trained using the complete gold standard. Each classifier f consisted of a class node predicting the binary presence or absence of a functional relationship (FR) between two genes and n nodes conditioned on FR, each representing the value of a data set Dk.

Parameter regularization is performed as described in Steck and Jaakkola (2002) using mutual information between data sets to estimate a strength of prior belief for each data set. While a large amount of shared information does not guarantee a redundant data set, since the same subset of information could be shared many times, it provides a valuable quantitative estimate of data set uniqueness.

5.2.2 Genomics data types

We collected and integrated 987 genome-scale data sets encompassing approximately 38,000 conditions from an estimated 14,000 publications including both expression and interaction measurements.

- Gene co-expression: All gene expression data sets are from NCBI's Gene Expression Omnibus (GEO). Genes with more than 30% of values missing were removed, and remaining missing values were imputed using ten nearest neighbors. Non-log-transformed data sets were log transformed. Expression measurements were summarized to Entrez identifiers, and duplicate identifiers were merged. The Pearson correlation was calculated for each gene pair, normalized with Fisher's z transform, mean subtracted and divided by the standard deviation.
- Protein-interaction: Interaction data are collected from BioGRID, IntAct, MINT, and MIPS.
- TF regulation: To estimate shared transcription factor regulation between genes, we collected binding motifs from JASPAR. Genes were scored for the presence of transcription factor binding sites using the MEME software suite. Motif matches were treated as binary scores (present if P < 0.001). The final score for each gene pair was obtained by calculating the Pearson correlation between the motif association vectors for the genes.
- MSigDB perturbations and miRNA: Chemical and genetic perturbation (c2:CGP) and microRNA target (c3:MIR) profiles were downloaded from the Molecular Signatures Database (MSigDB). Each gene pair's score was the sum of shared profiles weighted by the specificity of each profile

5.2.3 Evidence

The "evidence" for an edge is measured as the contribution or "influence" of each dataset on the posterior classification probability. Each dataset contribution is calculated as the posterior probability of a functional relationship given only that dataset, minus the prior probablility.

Contribution of dataset D to an edge functional relationship prediction (FR):

```
contribution(D) = P(FR | D) - P(FR)
```

Note that the contributions will not sum to 1.0, as each contribution is measured separately. Generally, individual gene expression datasets will not contribute much to the posterior probability but cumulatively can make a significant contribution.

5.3 Tissue-specific Networks

The precise actions of genes are frequently dependent on their tissue context, and human diseases result from the disordered interplay of tissue- and cell lineage–specific processes. These factors combine to make the understanding of tissue-specific gene functions, disease pathophysiology and gene-disease associations particularly challenging.

5.3.1 Functional interactions

HumanBase builds genome-scale functional maps of human tissues by integrating a collection of data sets covering thousands of experiments contained in more than 14,000 distinct publications. To integrate these data, we automatically assess each data set for its relevance to each of 144 tissue- and cell lineage–specific functional contexts. The resulting functional maps provide a detailed portrait of protein function and interactions in specific human tissues and cell lineages ranging from B lymphocytes to the renal glomerulus and the whole brain. This approach allows us to profile the specialized function of genes in a high-throughput manner, even in tissues and cell lineages for which no or few tissue-specific data exist.

These maps can answer biological questions that are specific to a single gene in a single tissue. For example, we have used these maps for the gene IL1B (encoding interleukin (IL)-1β) in the blood vessel network, where it has a key role in inflammation, to predict lineage-specific responses to IL-1β stimulation, which we experimentally confirmed.

5.3.2 Examples

IL1B in blood vessel

We examined and experimentally verified the tissue-specific molecular response of blood vessel cells to stimulation by IL-1β (IL1B), a pro-inflammatory cytokine. We anticipated that the genes most tightly connected to IL1B in the blood vessel network would be among those responding to IL-1β stimulation in blood vessel cells. We tested this hypothesis by profiling the gene expression of human aortic smooth muscle cells (HASMCs; the predominant cell type in blood vessels) stimulated with IL-1β.

Examination of the genes whose expression was significantly upregulated at 2 h after stimulation showed that 18 of the 20 IL1B network neighbors were among the top 500 most upregulated genes in the experiment (P = 2.07 × 1023). The blood vessel network was the most accurate tissue network in predicting this experimental outcome; none of the other 143 tissue-specific networks or the tissue-naive network performed as well when evaluated by each network's ability to predict the result of IL-1β stimulation on the cells.

Fig. 1: We anticipated that the genes most tightly connected to IL1B in the blood vessel network would be among those responding to IL-1β stimulation in blood vessel cells (a) The 20 genes most tightly connected to IL1B in the blood vessel network are shown. These genes are predicted to respond to IL-1β stimulation in blood vessel. (b) The bar plot shows the differential expression levels of the 20 IL1B neighbors measured in a microarray experiment at 0 h and 2 h after IL-1β stimulation in HASMCs, which constitute most of the blood vessel. Each bar represents the gene's log ratio of mean expression at 2 h over its mean expression at 0 h. Error bars represent regularized pooled standard errors estimated by LIMMA (n = 4 biological replicates). Eighteen of the 20 IL1B network neighbors (labeled in bold) were found to be among the most significantly differentially expressed genes at 2 h relative to 0 h (P = 1.95 × 1023).

5.4 Functional module detection

HumanBase applies community detection to find cohesive gene clusters from a provided gene list and a selected relevant tissue. Genes within a cluster share local network neighborhoods and together form a cohesive, specific functional module. Module detection enables systematic association of genes - even functionally uncharacterized genes - to specific processes and phenotypes represented in the detected modules. Functional modules are identified with tissue-specific networks, which predict gene interactions from massive data collections. Thus the discovered modules potentially capture higher-order tissue-specific function.

5.4.1 Method

The approach[1] is based on shared k-nearest-neighbors (SKNN) and the Louvain community-finding algorithm to cluster the user-selected tissue network into distinct modules of tightly connected genes. The SKNN-based strategy has the advantages of alleviating the effect of high-degree genes and accentuating local network structure by connecting genes that are likely to be functionally clustered together.

This technique proceeds as follows:

(i) First, we create a subset of the user-selected network containing only the user-provided genes and all the edges between them. Given the resulting graph G with V nodes (user-provided genes) and E edges, with each edge between genes i and j associated with a weight p_{ij},

(ii) Calculate a new weight for the edge between each pair of nodes i and j that is equal to the number of k nearest neighbors (based on the original weights p_{ij}) shared by i and j;

(iii) Choose the top 5% of the edges based on the new edge weights, and apply a graph clustering algorithm.

This approach has two key desirable characteristics:

(i) Choosing the highest k values instead of all edges deemphasizes high-degree ‘hub’ nodes and brings equal attention to highly specific edges between low-degree nodes;

(ii) Emphasizing local network-structure by connecting nodes that share a number of local neighbors automatically links genes that are highly likely to be part of the same cluster.

We use a dynamic `k = min(50, 0.2 * |V|)` to obtain the shared-nearest-neighbor tissue-specific network and apply the Louvain algorithm to cluster this network into distinct modules. To stabilize clustering across different runs of the Louvain algorithm, we run the algorithm 100 times and calculate cluster comembership scores for each pair of genes that was equal to the fraction of times (out of 100) the pair was assigned to the same cluster. Genes are assigned to clusters where their comembership score 0.9.

Resulting modules are then tested for functional enrichment using genes annotated to Gene Ontology biological process terms. Representative processes and pathways enriched within each cluster are presented alongside of the cluster with their resulting Q value. The Q value of each term associated to the modules is calculated using one-sided Fisher’s exact tests and Benjamini–Hochberg corrections to correct for multiple tests.

1. Krishnan A*, Zhang R*, Yao V, Theesfeld CL, Wong AK, Tadych A, Volfovsky N, Packer A, Lash A, Troyanskaya OG.(2016) Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. Nature Neuroscience.

5.5 NetWAS - Network-wide Association Study

Tissue-specific networks provide a new means to generate hypotheses related to the molecular basis of human disease. We developed an approach, termed network-wide association study (NetWAS). In NetWAS, the statistical associations from a standard GWAS guide the analysis of functional networks. This reprioritization method is driven by discovery and does not depend on prior disease knowledge. NetWAS, in conjunction with tissue-specific networks, effectively reprioritizes statistical associations from distinct GWAS to identify disease-associated genes, and tissue-specific NetWAS better identifies genes associated with hypertension than either GWAS or tissue-naive NetWAS.

5.5.1 Method

NetWAS trains a support vector machine classifier using nominally significant ($P < 0.01$) genes as positive examples and 10,000 randomly selected non-significant (P 0.01) genes as negatives. The classifier is constructed using a tissue network relevant to a disease (e.g. kidney for hypertension), where the features of the classifier are the edge weights of the labeled examples to all the genes in the network. Genes are re-ranked using their distance from the hyperplane, which represent a network-based prioritization of a GWAS, termed NetWAS.

To calculate per-gene P values for a GWAS, we suggest the versatile gene-based association study (VEGAS) system.

We have performed and evaluated NetWAS on six GWAS: C-reactive protein levels (lnCRP), type 2 diabetes (T2D), body mass index (BMI), hypertension (ht), alzheimer’s (adni) and advanced age-related macular degeneration (advanced AMD).

5.5.2 GWAS File

NetWAS requires as input a GWAS result file, with per-gene p-values. We suggest the versatile gene-based association study (VEGAS) system for calculating gene p-values, but we also support forge and pseq formats.

- VEGAS: versatile gene-based association study
- FORGE: multivariate calculation of gene-wide p-values from Genome-Wide Association Studies Authors and Affiliations
- PLINK/SEQ: a library for the analysis of genetic variation data

  Note that the expected format is from the output of Gene/group-based association tests

5.5.3 NetWAS Results

When a NetWAS analysis finishes, a result file will be emailed to the provided address and/or can be accessed at a given URL. An example file is show below:

```
##################################################################################
# HumanBase NetWAS Analysis Results
#
# Job id:      d7732f19-916d-4458-97b5-936b8d6345cb
# Job title:
# Email:
# Created:     2017-08-21 17:07:33 EDT
# GWAS file:   bmi-2012.out.txt
# GWAS format: vegas
# Tissue:      adipose_tissue
# P-value:     0.01
#
# Result file format:
#
# Column 1) Gene symbol
# Column 2) Training label: 1 (+, nominally significant p-value)
#                          -1 (-, not nominally significant p-value)
#                           0 (not used in training)
# Column 3) NetWAS Score: Distance from the SVM separating hyperplane. Positive scores
# are in the positive direction (more like nominally significant), negative scores
# are in the negative direction (more like non-significant)
##################################################################################
# NetWAS citation:
# Greene CS*, Krishnan A*, Wong AK*, Ricciotti E, Zelaya RA, Himmelstein DS, Zhang
# R, Hartmann BM, Zaslavsky E, Sealfon SC, Chasman DI, FitzGerald GA, Dolinski K,
# Grosser T, Troyanskaya OG. (2015). Understanding multicellular function and
# disease with human tissue-specific networks. Nature Genetics. 10.1038/ng.3259w.
##################################################################################
KRT6B  -1      0.561327
EMP1   -1      0.541169
ZBTB41 -1      0.503238
PNPLA8 -1      0.454396
ITGB4  -1      0.440985
........
```

5.5.4 Examples

Hypertension GWAS

Hypertension is a major cardiovascular risk factor and a complex trait involving a large number of genetic variants. We converted SNP-level association statistics into gene-level statistics for each of three recorded phenotypes—diastolic blood pressure (DBP), systolic blood pressure (SBP) and hypertension. Using the tissue-specific network for kidney, a tissue that has a central role in blood pressure control, NetWAS constructed a classifier that identified tissue-specific network connectivity patterns associated with the phenotype of interest. Genes annotated to hypertension phenotypes in the Online Mendelian Inheritance in Man (OMIM) database were more highly ranked by this classifier than by the initial GWAS. (citation)

Fig. 2: Genes ranked using GWAS (gray) and genes reprioritized using NetWAS (brown) were assessed for correspondence to genes known to be associated with hypertension phenotypes, regulatory processes and therapeutics. We compared individual (systolic blood pressure, SBP; diastolic blood pressure, DBP; hypertension, HTN) as well as combined hypertension endpoints. (a) Gene rankings were compared to OMIM-annotated hypertension genes using AUC. The AUC for the tissue-specific NetWAS is consistently higher than that for the original GWAS for all hypertension endpoints. Merging the network-based predictions for the three hypertension-related endpoints into a combined phenotype results in the best performance (AUC = 0.77; original GWAS AUC = 0.62; the dashed line at 0.5 denotes the AUC of a baseline random predictor). (b,c) Gene rankings were also assessed for enrichment of genes involved in the regulation of blood pressure (GO) (b) and targets of antihypertensive drugs (DrugBank) (c). The top NetWAS results were significantly enriched for genes involved in blood pressure regulation as well as for genes that are targets of antihypertensive drugs. Enrichment was calculated as a z score (Online Methods), with higher scores indicating a greater shift from the expected ranking toward the top of the list. In nearly all cases, the NetWAS ranking was both significantly enriched with the respective gene sets (z score > 1.645 P value < 0.05) and more enriched than in the original GWAS ranking.

Additional GWAS

Fig. 3: Each bar shows the performance of NetWAS reprioritization as measured by the area under the curve (AUC) of documented disease associations with the disease specified in the label above the plot. The horizontal axis shows relevant networks (colored bars) and GWAS alone (gray bars), and the horizontal axis label describes the GWAS phenotype from which associations were obtained.

5.6 Sei

5.6.1 Introduction

Sei is a deep-learning-based framework for systematically predicting sequence regulatory activities and applying sequence information to understand human genetics data. Sei provides a global map from any sequence to regulatory activities, as represented by 40 sequence classes. Each sequence class integrates predictions for 21,907 chromatin profiles (transcription factor, histone marks, and chromatin accessibility profiles across a wide range of cell types) from the underlying Sei deep learning model. You can also find the Sei code repository here or read about our manuscript here.

Sequence class-level variant effects are computed by comparing the predictions for the reference and the alternative alleles. A positive score indicates an increase in sequence class activity by the alternative allele and vice versa. Sequence class-level scores are computed by projecting the 21,907 chromatin profile predictions for the sequence to the unit vector that represents each sequence class.

For older DeepSEA models see: *DeepSEA (Beluga)* (2019)

5.6.2 Input

File formats

We support three types of input: vcf, fasta, bed. If you want to predict effects of noncoding variants, use vcf format input. If you want to predict chromatin feature probabilities for DNA sequences, use fasta format. If you want to specify sequences from the human reference genome (GRCh37/hg19), you can use bed format. See below for a quick introduction:

VCF format is used for specifying a genomic variant. A minimal example is `chr1 109817590 - G T` (if you want to copy cover this text as input, you will need to change spaces to tabs). The five columns are chromosome, position, name, reference allele, and alternative allele. Currently, the genome position needs to be in GRCh37/hg19

Fasta format input should include sequences of 4096bp length each. If a sequence is longer than 4096bp, only the center 4096bp will be used.

Bed format provides another way to specify sequences in human reference genome (hg19). The bed input should specify 4096bp-length regions. A minimal example is `chr1 109817091 109821186`. The three columns are chromosome, start position, and end position.

Genome coordinates

We support only `GRCh37/hg19` genome coordinates. You can use LiftOver to convert your coordinates to the correct version.

Large submissions

We recommend using the web server if you have <10,000 variants or sequences. You will experience degraded performance when submitting a larger set of sequences. In those instances, we suggest that you split the set into multiple <10,000 submissions, or run the standalone version on your local machine, or contact our group directly.

5.6.3 Output

Sequence classes

The Sei framework predicts 40 sequence class scores, covering a wide range of regulatory activities such as cell-type-specific enhancers and promoters, as well as 21,907 chromatin profiles for any DNA sequence.

To help interpretation, we grouped sequence classes into groups including P (Promoter), E (Enhancer), CTCF (CTCF-cohesin binding), TF (TF binding), PC (Polycomb-repressed), HET (Heterochromatin), TN (Transcription), and L (Low Signal) sequence classes. Please refer to our manuscript for a more detailed description of the sequence classes.

Note: sequence class predictions are only available for vcf inputs.

```
| Sequence class label |            Sequence class name | Rank by size | Group |
|---------------------:|-------------------------------:|-------------:|------:|
|                  PC1 |     Polycomb / Heterochromatin |            0 |   PC  |
|                   L1 |                     Low signal |            1 |    L  |
|                  TN1 |                  Transcription |            2 |   TN  |
|                  TN2 |                  Transcription |            3 |   TN  |
|                   L2 |                     Low signal |            4 |    L  |
|                   E1 |                      Stem cell |            5 |    E  |
|                   E2 |                   Multi-tissue |            6 |    E  |
|                   E3 |             Brain / Melanocyte |            7 |    E  |
|                   L3 |                     Low signal |            8 |    L  |
|                   E4 |                   Multi-tissue |            9 |    E  |
```

(continues on next page)

(continued from previous page)

```
|                 TF1  |                    NANOG / FOXA1  |           10 |   TF  |
|                 HET1 |                  Heterochromatin  |           11 |  HET  |
|                  E5  |                      B-cell-like  |           12 |    E  |
|                  E6  |                  Weak epithelial  |           13 |    E  |
|                 TF2  |                            CEBPB  |           14 |   TF  |
|                 PC2  |                    Weak Polycomb  |           15 |   PC  |
|                  E7  |            Monocyte / Macrophage  |           16 |    E  |
|                  E8  |                Weak multi-tissue  |           17 |    E  |
|                  L4  |                       Low signal  |           18 |    L  |
|                 TF3  |                FOXA1 / AR / ESR1  |           19 |   TF  |
|                 PC3  |                         Polycomb  |           20 |   PC  |
|                 TN3  |                    Transcription  |           21 |   TN  |
|                  L5  |                       Low signal  |           22 |    L  |
|                 HET2 |                  Heterochromatin  |           23 |  HET  |
|                  L6  |                       Low signal  |           24 |    L  |
|                   P  |                         Promoter  |           25 |    P  |
|                  E9  |                Liver / Intestine  |           26 |    E  |
|                 CTCF |                     CTCF-Cohesin  |           27 |  CTCF |
|                 TN4  |                    Transcription  |           28 |   TN  |
|                 HET3 |                  Heterochromatin  |           29 |  HET  |
|                 E10  |                            Brain  |           30 |    E  |
|                 TF4  |                             OTX2  |           31 |   TF  |
|                 HET4 |                  Heterochromatin  |           32 |  HET  |
|                  L7  |                       Low signal  |           33 |    L  |
|                 PC4  | Polycomb / Bivalent stem cell Enh |           34 |   PC  |
|                 HET5 |                       Centromere  |           35 |  HET  |
|                 E11  |                           T-cell  |           36 |    E  |
|                 TF5  |                               AR  |           37 |   TF  |
|                 E12  |                Erythroblast-like  |           38 |    E  |
|                 HET6 |                       Centromere  |           39 |   HET |
```

Regulatory feature scores

- **diffs**: The difference between the the predicted probability of the reference allele and the alternative allele for a regulatory feature ($p_{alt} - p_{ref}$).

5.7 DeepSEA (Beluga)

5.7.1 Introduction

DeepSEA is a deep learning-based algorithmic framework for predicting the chromatin effects of sequence alterations with single nucleotide sensitivity. DeepSEA can accurately predict the epigenetic state of a sequence, including transcription factors binding, DNase I sensitivities and histone marks in multiple cell types, and further utilize this capability to predict the chromatin effects of sequence variants and prioritize regulatory variants.

The 2019 version of DeepSEA, nicknamed '**Beluga**', can predict **2002** chromatin features. Beluga is described in:

Jian Zhou, Chandra L. Theesfeld, Kevin Yao, Kathleen M. Chen, Aaron K. Wong, and Olga G. Troyanskaya, **Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk**. Nature Genetics (2018).

To determine if certain features (ie. transcription factors, marks, or cell types) are present/accounted for in the model, refer to the supplemental feature table which has all the profiles used to train DeepSEA.

DeepSEA is originally described in the following manuscript:

Jian Zhou, Olga G. Troyanskaya. **Predicting the Effects of Noncoding Variants with Deep learning-based Sequence Model.** Nature Methods (2015).

To determine if certain features (ie. transcription factors, marks, or cell types) are present/accounted for in the model, refer to the supplemental feature table which has all the profiles used to train DeepSEA.

5.7.2 Input

DeepSEA predicts genomic variant effects on a wide range of chromatin features at the variant position (Transcription factors binding, DNase I hypersensitive sites, and histone marks in multiple human cell types). DeepSEA can also be utilized for predicting chromatin features for any DNA sequence.

File formats

We support three types of input: vcf, fasta, bed. If you want to predict effects of noncoding variants, use vcf format input. If you want to predict chromatin feature probabilities for DNA sequences, use fasta format. If you want to specify sequences from the human reference genome (GRCh37/hg19), you can use bed format. See below for a quick introduction:

VCF format is used for specifying a genomic variant. A minimal example is `chr1 109817590 - G T` (if you want to copy cover this text as input, you will need to change spaces to tabs). The five columns are chromosome, position, name, reference allele, and alternative allele.

Fasta format input should include sequences of 2000bp length each. If a sequence is longer than 2000bp, only the center 2000bp will be used. A minimal example is

```
>TestSequence
TGGGATTACAGGCGTGAGCCACCGCGCCCGGCCCATTGTACCATTCTTAT
GCCTTTGCGTCCTCATAGCTTAGCTCCCGTATATCAGTGAGAACATACTA
TGTTTGGTTTTCCATACCCGAGTTACTTCACTTAGAATAATAGTCTCCAA
TTTCATCCAGGTCAGTGCAAATGCGTTAATTCGTTCCTTTTATGGCTGAG
TAGTATTCCATCATATATATATACTACAGTTTCTTTATCCACTCGTAAAT
TGATGGGCATTTGTGTTGGAACACTTCTCCACTGCTGGTGGGAATGTAAA
TTAGTGCAGCCACTATGGATAACAGTGTGGAGATTTGTTAAAGAACTAAA
ACTAGAACTACCATTTGATCCAGCAATCCCACTACTGGGTATCTACCCAG
AAGAAAAGAAGTCATTATTTGAAAAAGATACTTGCACGGGCATGTTTATA
GCAGCACAATTCACAATTGTAGTTGTATTTCTTTAAGCGTGTCTTTTCAA
TATCTCTCATGTTTCTGGTATAGATGGTATATATGTTAATCTTGTTCCTG
AGGTCTGTTTTTTATTTTTGTCATTAAAGTGGGAATTAAATAGTTTTGTA
GTGCATATAAATTAAAGAAAAAGTTCACATAAGCATATTTGCCAATCATC
TCAAAATGCTATATTCTCCTTCACGGTTTTGAAAATAATTCAGGGTTTTC
TCTTCCTCATTGCTTTCCCACCAACTGACAGTATTATTTTCTTAGTCATT
TTACTGACCTTTGAAATTACTCCTTTGAGGTCTTCTAAAAAATTTTATGG
GCTCTGCTGCTTTTTGGTGGCCTCCTTGTATCATTTATTCTATTACAGGA
CGACTTACAAAAGGAAGCACATAAATTGACCCATATACATATCCTATCAT
TGGGGAGTTTCTGTGCAAATGTTATTTATTGGAAGCTATTACTAAGAATT
GTAAGAAAAATAATTGGTATTGATGCAGCTAGTATGGTTCCTGTAATTAT
CGTACTCAGCCACGTAAATCATAGCTATATGTAGCCAAAGATCCATGAAC
AAAATTTCCAGTAACATCATTATAATTCAAAAGGCAGACTTTCAGAACCA
GACAGACTTGAATTTAAATTCTAGCTTTACCACACATGAATTTAACCTTG
TGGAAGGTTAACCTATCTAAACTCATGTTTCTTCATTGGTAGCTGATAAA
ATTAAGGATCATGTATATAACCACCTAGTAGAGTTGTTTAAGAAACTGTT
AGAATTCCATAAATTGTTAGTATTAATGAGTTTTTGTTGGACATGTGTTA
GGCTAGGCCACTCCTTGACCTTCATAGAGGTATGGATTATGACACAAATT
```

(continues on next page)

(continued from previous page)

```
CTAAACTGTAGGTAGGCATGGCTTTGTAGCAAGTATTAAAATAGTAAATA
TTTTATTTTTATAAGATAAATGTAAACCTTTTAAAAGTTTCATTACATTT
GTATTTATGAAATATCATCCTATATCAACTATAGAGAGAAGATCGCAAGA
AGGCAGTGGCAGCAGAGGCTCCAGTTAGGAGGCTACTAGTCCAAATACAT
TGCGATAAAAACTTGGCAAAAGGTGCTGGTAGTCTGATGAAATAAAGTAG
ATAAATTTTAGAGGTATTTATAAAATAATTAAAGAATATTCAATAATAGG
AGATATATTACCCAATAGAGTGGAGATTCAAAGATAACTCCGAAAGTTTT
TTGCTAAAGCAACATTTGGCTGTGCTATCATTTACTAAGAAAGACAACAA
GAGAGTAAAATCAAGTTTGAGGATGAAGTGAATTTATTCCTTTTTGATTG
ATACATAATTGACATGTAATAAAACCCACAATGTTAAGAGTTCGGTTTGA
TGTGCTTGACTATTTTAGGCACTGGTGTTATCACAACACAAGACAACAGA
TAGGACATTCTCAGAAAATTTTTTCATGTCCCTTTCCAGTCAGTTTCAAG
CCTTCTTTCCATGCAATAATTTTCTCACTTTGCCATTCTAGTAGGTGTGA
```

Bed format provides another way to specify sequences in human reference genome (hg19). The bed input should specify 2000bp-length regions. A minimal example is `chr1 109817091 109819090`. The three columns are chromosome, start position, and end position.

Genome coordinates

We support only `GRCh37/hg19` genome coordinates. You can use LiftOver to convert your coordinates to the correct version.

Large submissions

We recommend using the web server if you have <10,000 variants or sequences. You will experience degraded performance when submitting a larger set of sequences. In those instances, we suggest that you split the set into multiple <10,000 submissions, or run the standalone version on your local machine, or contact our group directly.

5.7.3 Output

Regulatory feature scores

- **diffs**: The difference between the the predicted probability of the reference allele and the alternative allele for a regulatory feature ($p_{alt} - p_{ref}$).
- **e-value**: E-value is defined as the expected proportion of SNPs with a larger predicted effect. We calculate an ‘e-value’ based on the empirical distribution of that feature’s effect ($abs(p_{alt} - p_{ref})$) among gnomAD variants. For example, a feature e-value of 0.01 indicates that only 1% of gnomAD variants have a larger predicted effect.
- **z-score**: A scaled score where the feature diff score ($p_{alt} - p_{ref}$) is divided by the root mean square of the feature diff score across gnomAD variants. Note that this is “sign-preserving”, i.e. a negative z-score indicates that a mutation **decreases** the probability of a regulatory feature.

Variant scores

- **Disease Impact Score (DIS)**: DIS is calculated by training a logistic regression model that prioritizes likely disease-associated mutations on the basis of the predicted transcriptional or post-transcriptional regulatory effects of these mutations (See Zhou et. al, 2019). The predicted DIS probabilities are then converted into ‘DIS e-values', computed based on the empirical distributions of predicted effects for gnomAD variants. The final DIS score is:

$$-log10(DISevalue_{feature})$$

- **Mean -log e-value (MLE)**: For each predicted regulatory feature effect ($abs(p_{alt} - p_{ref})$) of a variant, we calculate a 'feature e-value' based on the empirical distribution of that feature's effects among gnomAD variants (see above Regulatory feature scores: e-value). The MLE score of a variant is

$$\sum -log10(evalue_{feature})/N$$

5.7.4 In-silico mutagenesis

Perform "In silico saturated mutagenesis" (ISM) analysis to discover informative sequence features within any sequence. Specifically, it performs computational mutation scanning to assess the effect of mutating every base of the input sequence on chromatin feature predictions. This method for context-specific sequence feature extraction takes advantage of DeepSEA's ability to utilize flanking context sequences information.

Note that ISM only accepts a sequence (FASTA file) as input.

ISM outputs effects for each of three possible substitutions of all 2000 bases, across all chromatin features.

5.8 ExPecto

5.8.1 Introduction

ExPecto is a framework for ab initio sequence-based prediction of mutation gene expression effects and disease risks. With this web interface, we provide an explorer of tissue-specific expression effect predictions. The current release contains all single nucleotide substitutions within 1kb to the representative TSS of a gene and all 1000 Genomes variants that passed a minimum predicted effect threshold (>0.3 log fold-change in any tissue).

The code for predicting expression effects for human genome variants and training new expression models is available at this github repository.

The ExPecto framework is described in the following manuscript:

Jian Zhou, Chandra L. Theesfeld, Kevin Yao, Kathleen M. Chen, Aaron K. Wong, and Olga G. Troyanskaya, Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk, Nature Genetics, 2018

5.8.2 Download

Predicted expression effects

This is the bulk download link of all mutation predictions.

Variation potential directionality scores

Variation potential of a gene in a tissue or cell-type can reflect the evolutionary constraint on its expression level. Specifically, we compute the variation potential directionality score as the sum of all directional mutation effects within 1kb to TSS. A negative variation potential indicates active expression and constraint toward higher expression level, and vice versa. The sum of absolute mutation effects, or the magnitudes, is predictive of tissue/condition-specificity of a gene. The variation potential directionality scores and the inferred evolution constraint probabilities can be downloaded here.

The full prediction of all 140 million mutations can be downloaded here (~125G).

5.8.3 Method Details

ExPecto uses exponential basis function-based linear models upon deep convolutional network model of chromatin effects. ExPecto predicts expression levels directly from sequence and is capable of predicting effects of sequence variations.

For detailed procedures of the prediction, the chromatin predictions were computed from DeepSEA "Beluga" per 200bp bin, and 200 bins centered at TSS (40kb region) were used as input to predict expression effects. To reduce the dimensionality for ExPecto model training, the predicted chromatin spatial patterns were summarized to spatial features by 10 exponential basis functions. The summarized spatial features and gene expression levels were used to train regularized linear models for the final step of the prediction. The representative TSSes are selected based on FANTOM CAGE data.

We also propose a path toward ab initio disease risk prediction through combining the prediction of expression effects and the estimation of evolution constraints on expression levels. For example, mutations predicted to have strong negative expression effects on a positively constrained gene are predicted to be deleterious. We estimate evolutionary constraints through systematic profiling of potential mutation effects through in silico mutagenesis. As proof-of-principle we showed that this approach can predict the disease alleles from both curated HGMD disease mutation data and disease GWASes.

5.9 Citations

5.9.1 Tissue-specific networks, NetWAS

Greene CS*, Krishnan A*, Wong AK*, Ricciotti E, Zelaya RA, Himmelstein DS, Zhang R, Hartmann BM, Zaslavsky E, Sealfon SC, Chasman DI, FitzGerald GA, Dolinski K, Grosser T, Troyanskaya OG. (2015). Understanding multicellular function and disease with human tissue-specific networks. Nature Genetics. 10.1038/ng.3259w.

5.9.2 Variant effect predictions (ExPecto)

Zhou J, Theesfeld CL, Yao K, Chen KM, Wong AK, and Troyanskaya OG. (2018) Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk, Nature Genetics.

5.9.3 Autism gene predictions

Krishnan A*, Zhang R*, Yao V, Theesfeld CL, Wong AK, Tadych A, Volfovsky N, Packer A, Lash A, Troyanskaya OG.(2016) Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. Nature Neuroscience.

5.9.4 Tissue-expression predictions

W. Ju#, C.S. Greene#, F. Eichinger, V. Nair, J.B. Hodgin, M. Bitzer, Y. Lee, Q. Zhu, M. Kehata, M. Li, M.P. Rastaldi, C.D. Cohen, O.G. Troyanskaya*, and M. Kretzler*. Defining cell type specificity at the transcriptional level in human disease. Genome Research. 23:1862-1873. 2013