HumanBase Documentation

Release 1.0

User Guide

Apr 25, 2018

Contents

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.

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.

Help topics

3.1 Citations

3.1.1 Tissue-specific networks

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.

3.1.2 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.

3.1.3 Tissue gene 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

3.2 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)

3.2.1 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.

3.2.2 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 purturbations 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

3.2.3 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.

3.3 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 Net-WAS better identifies genes associated with hypertension than either GWAS or tissue-naive NetWAS.

3.3.1 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).

3.3.2 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

3.3.3 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
........
```

3.3.4 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)

Additional GWAS

3.4 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.

3.4.1 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.
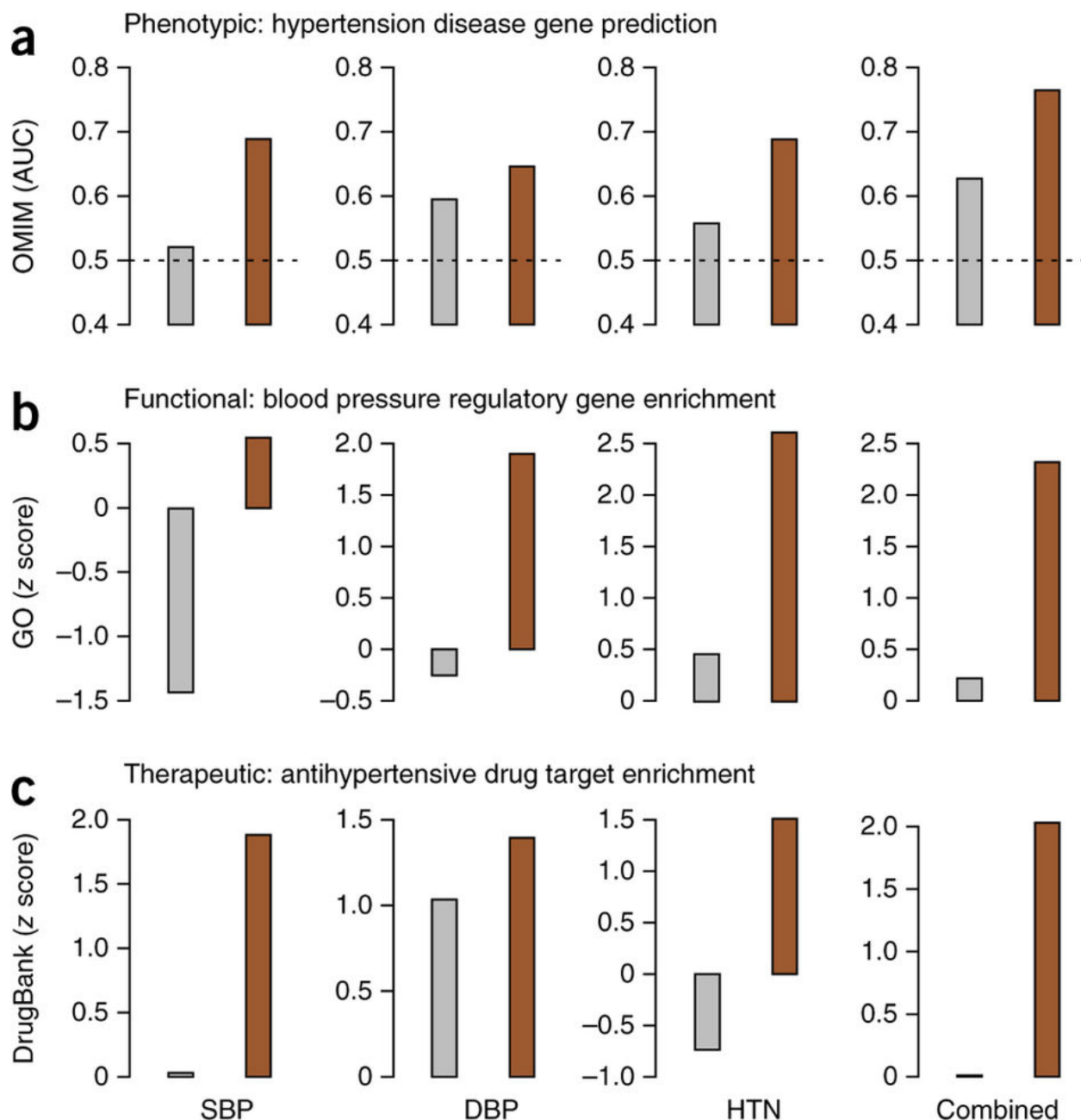
Fig. 3.1: Genes ranked using GWAS (gray) and genes reprioritized using NetWAS (brown) were assessed for correspondence to genes known to be associated with hypertension phenotypes, regulatory processes and therapeutics. We compared individual (systolic blood pressure, SBP; diastolic blood pressure, DBP; 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.
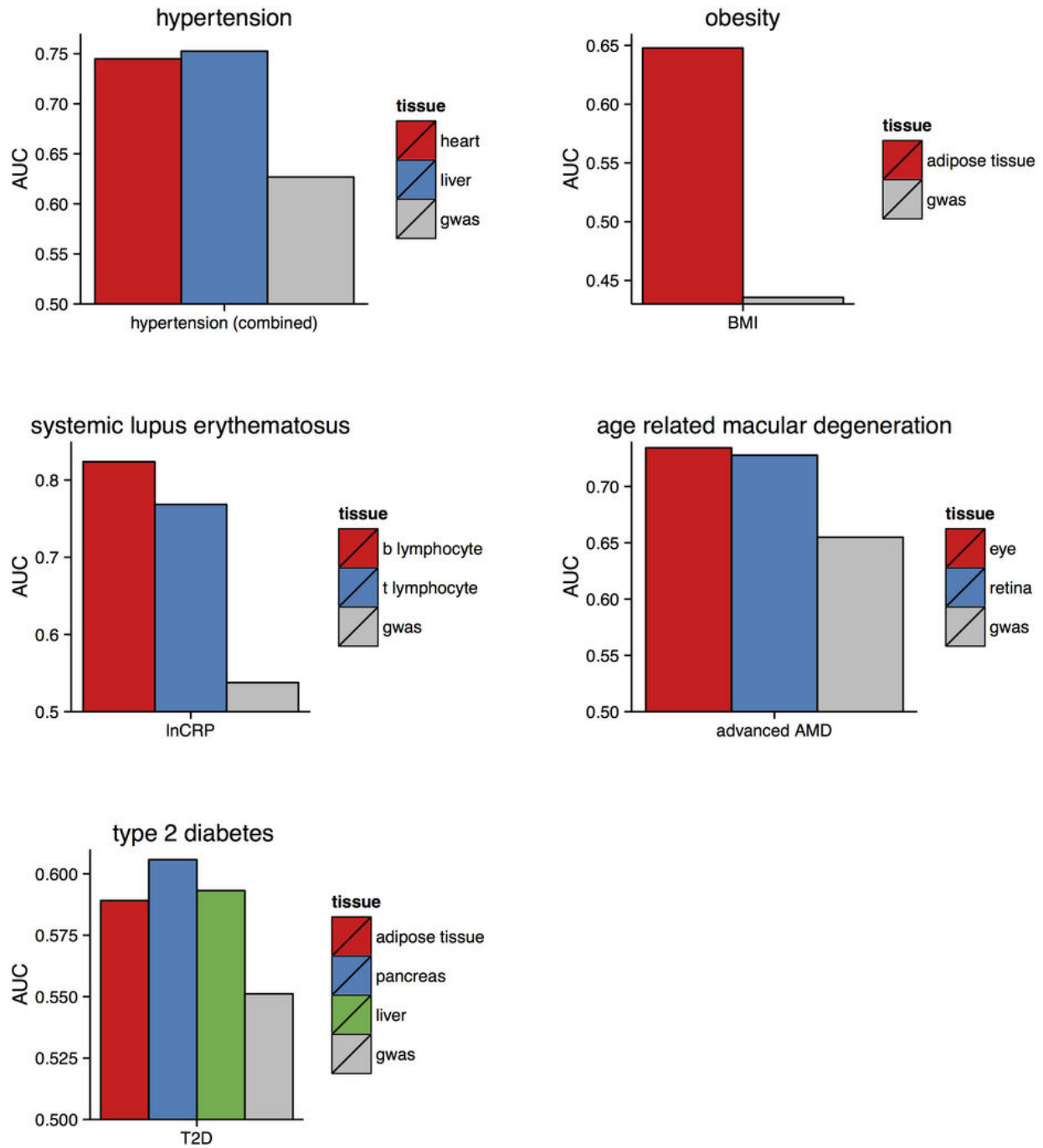
Fig. 3.2: Each bar shows the performance of NetWAS reprioritization as measured by the area under the curve (AUC) of documented disease associations with the disease specified in the label above the plot. The horizontal axis shows relevant networks (colored bars) and GWAS alone (gray bars), and the horizontal axis label describes the GWAS phenotype from which associations were obtained.

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.

3.4.2 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 proinflammatory 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.

3.5 Getting Started

3.6 Who are we?

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

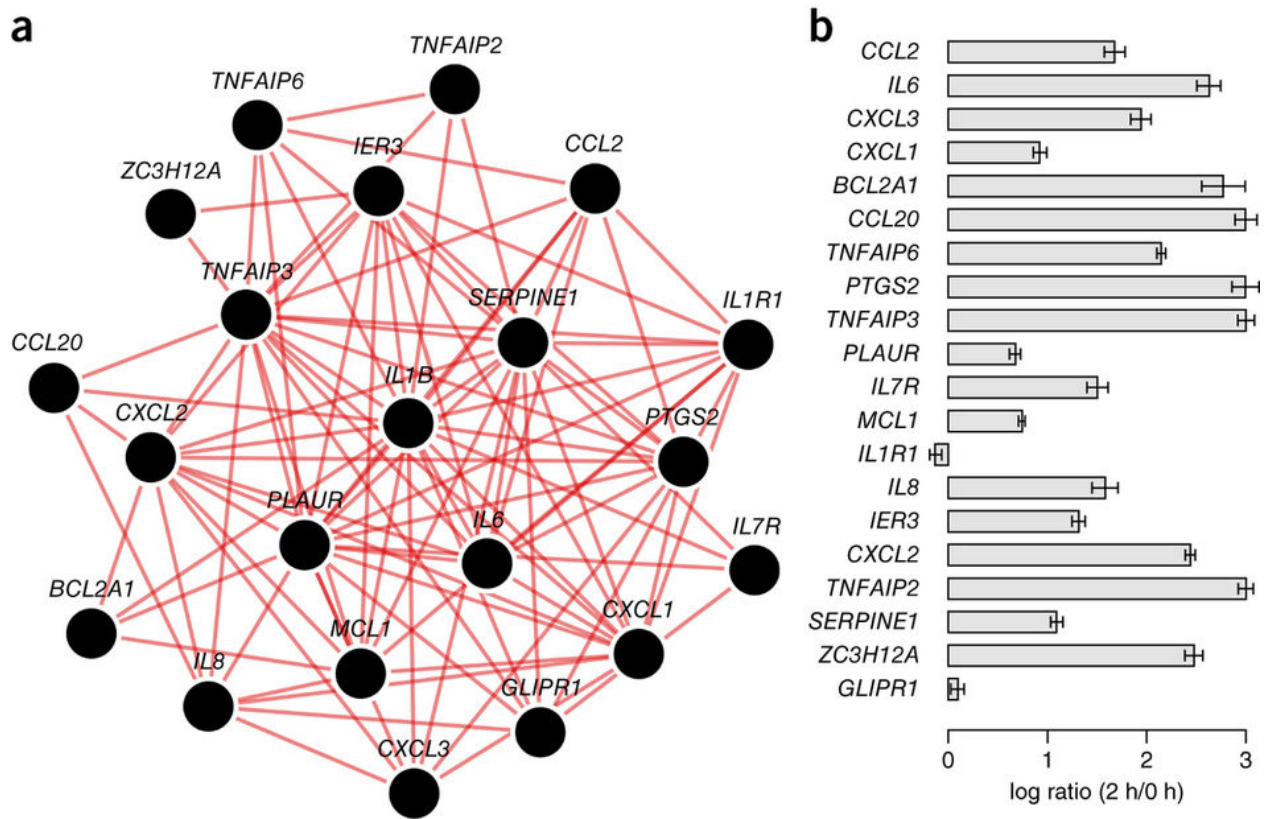Fig. 3.3: We anticipated that the genes most tightly connected to IL1B in the blood vessel network would be among those responding to IL-1β stimulation in blood vessel cells (a) The 20 genes most tightly connected to IL1B in the blood vessel network are shown. These genes are predicted to respond to 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).