

# SCIENTIFIC REPORTS



OPEN

## Tissue Specificity of Human Disease Module

Maksim Kitsak<sup>1,2</sup>, Amitabh Sharma<sup>1,2,3</sup>, Jörg Menche<sup>1,2,4,5</sup>, Emre Guneş<sup>1,2,6</sup>, Susan Dina Ghiassian<sup>1,2,7</sup>, Joseph Loscalzo<sup>8</sup> & Albert-László Barabási<sup>1,2,4,7,8,9</sup>

Received: 16 May 2016

Accepted: 20 September 2016

Published: 17 October 2016

Genes carrying mutations associated with genetic diseases are present in all human cells; yet, clinical manifestations of genetic diseases are usually highly tissue-specific. Although some disease genes are expressed only in selected tissues, the expression patterns of disease genes alone cannot explain the observed tissue specificity of human diseases. Here we hypothesize that for a disease to manifest itself in a particular tissue, a whole functional subnetwork of genes (disease module) needs to be expressed in that tissue. Driven by this hypothesis, we conducted a systematic study of the expression patterns of disease genes within the human interactome. We find that genes expressed in a specific tissue tend to be localized in the same neighborhood of the interactome. By contrast, genes expressed in different tissues are segregated in distinct network neighborhoods. Most important, we show that it is the integrity and the completeness of the expression of the disease module that determines disease manifestation in selected tissues. This approach allows us to construct a disease-tissue network that confirms known and predicts unexpected disease-tissue associations.

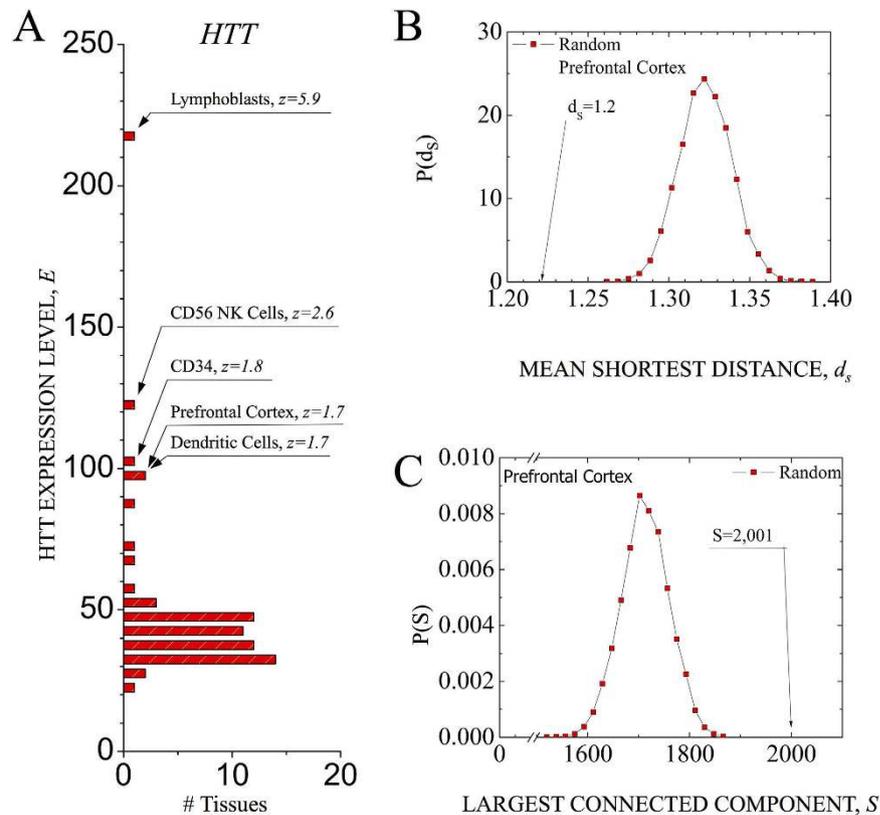
The rapidly increasing knowledge of the role of genetic variants in human disease raises an important question: Why do pathologic variants that exist in the genome of every organ or tissue affect primarily specific tissues, like asthma affecting the lung or schizophrenia the brain? After all, the mutated genes are present in all cells. The accepted answer holds that genes associated with asthma may be expressed only in the lung and those associated with schizophrenia may show elevated expression only in the brain, eliminating the impact of disease-causing variants in other tissues. This assumption is, however, typically not valid: many disease-associated genes are expressed in multiple tissues, most of which do not show pathophysiological manifestations of the disease or of any functional abnormality<sup>1,2</sup>. Consider, for example, Huntington's disease, a neurodegenerative genetic disorder caused by a mutation of the *HTT* gene (excessive repeats of the trinucleotide CAG). As shown in Fig. 1A, we find that *HTT* is significantly expressed in CD34 T cells ( $z_E^* = 1.83$ ), CD56 NK cells ( $z_E^* = 2.63$ ), and X721 B lymphoblasts ( $z_E^* = 5.99$ ), prompting us to ask again, why do the mutations of *HTT* not cause pathophysiological changes in these non-neural tissues?

The driving hypothesis of this paper is that the expression of genes carrying the disease-associated mutation in a particular tissue or organ does not fully explain the tissue-specificity of the disease. Instead, it is the integrity of the cellular subnetwork induced by all disease-associated genes that determines the manifestation of a disease in the tissue. Throughout the text we refer to this cellular subnetwork as a *disease module*<sup>3–8</sup>.

Thus, to understand disease manifestations or pathophenotypes, it is not sufficient to focus narrowly on disease genes and their expression patterns; we must also determine the presence/absence of the disease module, i.e., the tissue-dependent subnetwork, whose breakdown may be responsible for the disease.

This hypothesis is supported by several recent findings. First, disease-associated genes are more likely to exhibit tissue-specific expression than non-disease-associated genes<sup>2,9–11</sup>. The integration of gene expression, disease manifestation, molecular network connectivity, and tissue specificity data leads to better predictions of

<sup>1</sup>Center for Complex Networks Research and Department of Physics, Northeastern University, 110 Forsyth Street, 111 Dana Research Center, Boston, MA 02115, USA. <sup>2</sup>Center for Cancer Systems Biology (CCSB) and Department of Cancer Biology, Dana-Farber Cancer Institute, 450 Brookline Ave., 02215 Boston, USA. <sup>3</sup>Channing Division of Network Medicine, Harvard Medical School, 181 Longwood Avenue, Boston, MA 02115, USA. <sup>4</sup>Center for Network Science, Central European University, Nador u. 9, 1051 Budapest, Hungary. <sup>5</sup>CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences, Lazarettgasse 14, AKH BT 25.3, A-1090 Vienna, Austria. <sup>6</sup>Institute for Research in Biomedicine, Carrer de Baldri Reixac, 08028 Barcelona, Spain. <sup>7</sup>DZZOM Inc., 111 Huntington Avenue, Boston, MA 02119, USA. <sup>8</sup>Department of Medicine, Brigham and Womens Hospital, Harvard Medical School, 75 Francis Street, Boston, MA 02115, USA. <sup>9</sup>Department of Theoretical Physics, Budapest University of Technology and Economics, H1111, Budapest, Hungary. Correspondence and requests for materials should be addressed to A.-L.B. (barabasi@gmail.com)



**Figure 1. Network-based localization of expressed genes.** (A) The distribution of expression levels of the *HTT* gene across 64 non-diseased tissues. (B) The distribution of the mean shortest distance  $d_s$  for randomly chosen sets of genes in the human interactome. The number of randomly selected genes in each set is chosen to match the number of genes expressed in prefrontal cortex at  $z_E \geq 1.0$ . Note that the observed mean shortest distance for genes expressed in prefrontal cortex,  $d_{\text{prefr.cort}} = 1.22$ , is significantly lower than the random expectation. (C) The distribution of the connected components size,  $P(S)$ , for randomly chosen sets of genes in the human interactome. The number of randomly selected genes in each set is chosen to match the number of genes expressed in prefrontal cortex at  $z_E \geq 1.0$ . Note that the size of the connected component for genes expressed in prefrontal cortex,  $S_{\text{prefr.cort}} = 2,001$ , is significantly greater than the random expectation.

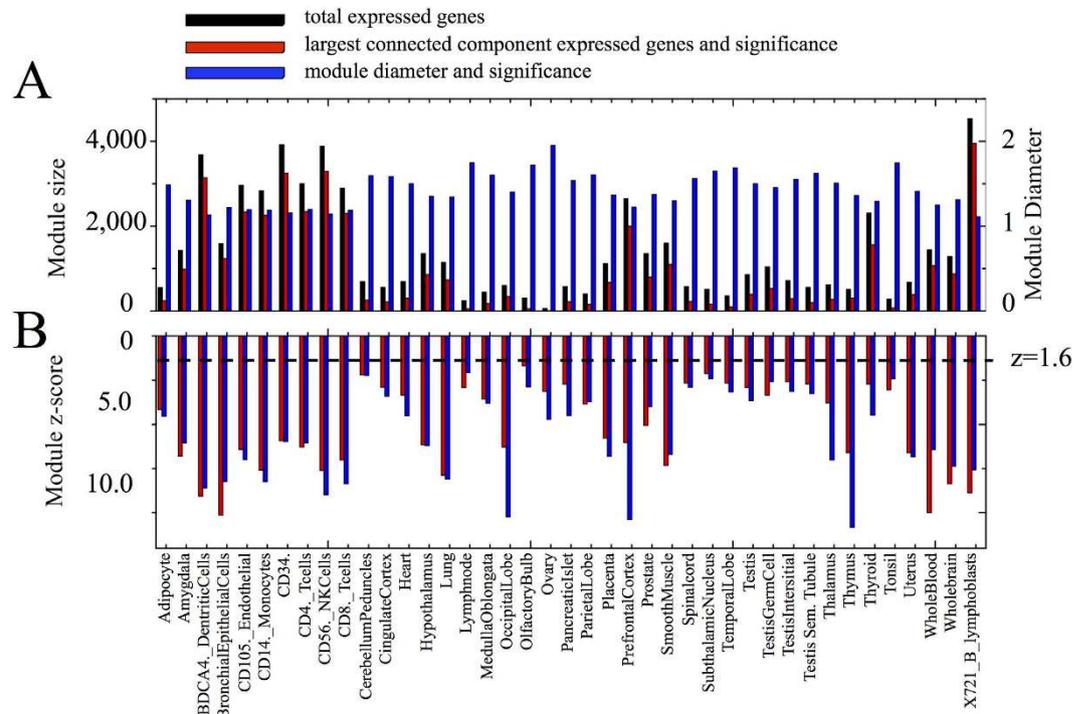
novel disease-gene candidates than any of these elements alone<sup>9</sup>. Similarly, a tissue-specific interactome considerably improves disease gene prioritization compared to prioritization derived from a generic interactome<sup>12</sup>. Our hypothesis is further supported by recent evidence that disease-causing genes tend to have elevated transcript level and increased number of tissues-specific protein interactions in their disease tissue<sup>2</sup>. Similarly, tissue-specific networks were found to accurately predict lineage-specific responses to perturbations, allowing the development of software packages that return the tissue-specific interaction landscape of selected genes<sup>1</sup>.

To test the interplay between the disease module and tissues, we analyze the expression patterns of disease genes and genes in their network vicinity in the human interactome<sup>3,4,13,14</sup>. We show that (i) genes expressed in a specific tissue are localized in the same neighborhood of the interactome, being in each other's close network-based proximity; (ii) genes expressed in different tissues are segregated in distinct network neighborhoods; and (iii) the integrity of the expression of the disease module determines disease manifestations in selected tissues. These findings offer an accurate and self-consistent formalization of tissue-specificity of human disease modules<sup>15</sup>, offering the graph-theoretical underpinning of relating diseases to specific tissues, consistent with the numerous empirical and experimental studies<sup>1,2,9–12</sup>.

## Results

**Tissue-Specific Expression Patterns in the Interactome.** In our analysis we use experimentally documented molecular interactions as compiled by Menche *et al.*<sup>15</sup>. The resulting interactome contains 141,296 physical interactions between 13,460 proteins, including protein-protein and regulatory interactions, and metabolic pathway and kinase-substrate interactions. To identify the tissue-dependent patterns of expression of each gene, we use global gene expression data for 64 non-diseased tissues from the GNF Atlas<sup>16</sup>. To combine the interactome with the gene expression data we consider molecular interactions only among 10,434 proteins that were annotated to probe ids (see Supplementary Section II).

Further, we consider a protein-coding gene to be expressed in a tissue if its expression level in this tissue is significantly higher than in other tissues. We use  $z$ -scores to quantify expression significance and define the



**Figure 2. Localization statistics for 41 tissues where a significant fraction of expressed genes forms a connected subgraph.** (A) The total number of expressed genes (black), the number of genes constituting the largest connected component (module size, red), and the mean shortest distance,  $d_s$  (module diameter, blue), calculated for 41 tissues where a significant fraction of expressed genes form a connected subgraph. (B) The significance of the observed largest connected components (red) and the significance of the mean shortest distance (blue). The horizontal dashed line corresponds to  $z = 1.6$ .

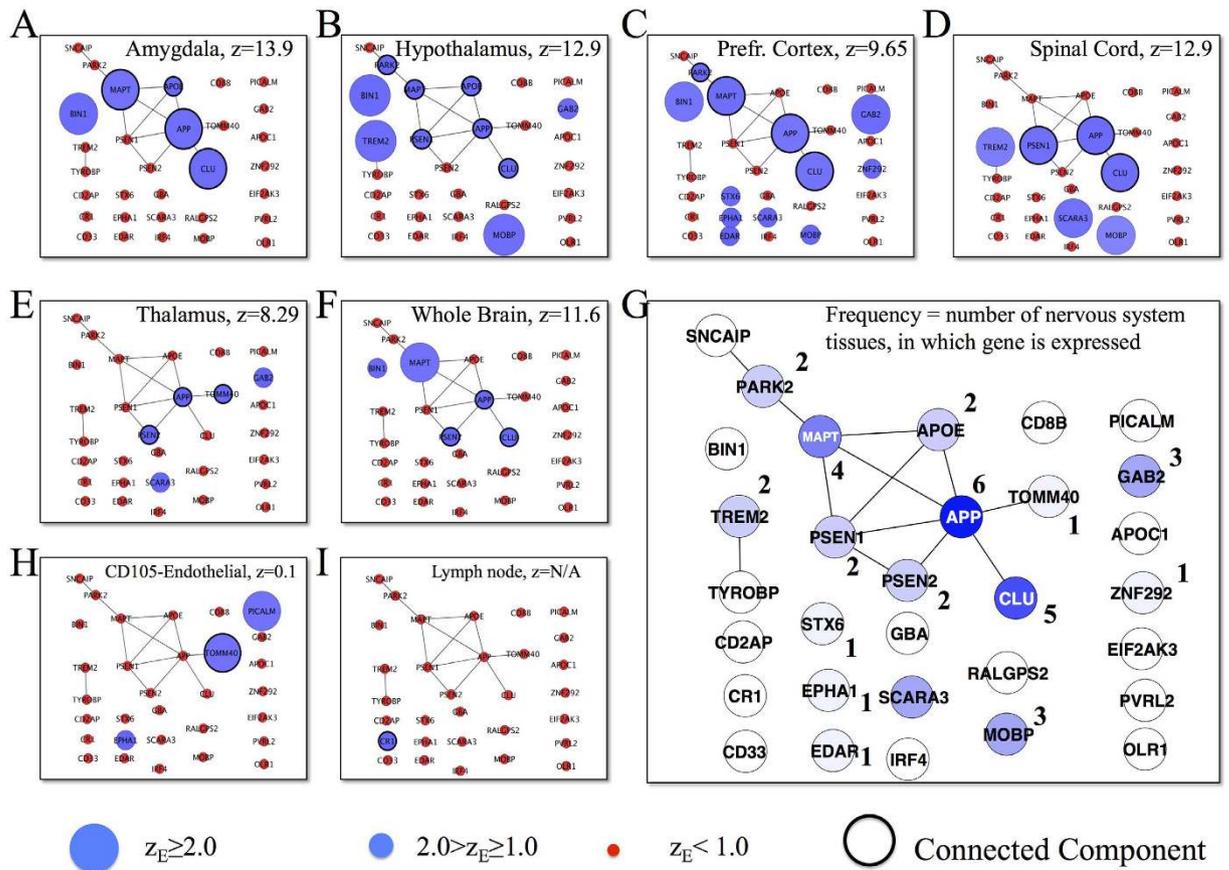
significance level at  $z_E^* = 1.0$  (see Methods and Fig. S1). At this threshold, on average only 10% of genes are expressed in a tissue, ranging from 0.6% for ovary to 44% for X721 B lymphoblasts ( Fig. 2A). We also repeated our calculations for a more stringent threshold of  $z_E^* = 2.0$ , with similar outcomes. We next identified all diseases defined by MeSH that have at least 20 associated genes in the current OMIM and GWAS database<sup>17,18</sup>, and then filtered out overly-inclusive diseases, yielding a list of 70 diseases (see Supplementary Section II).

To illustrate the impact of tissue-specific gene expression patterns on the interactome, consider prefrontal cortex tissue, in which only 2, 644 of the 10, 434 expressed genes are expressed with significance  $z_E \geq 1.0$ . We first asked: are the expressed genes distributed randomly in the interactome or do they tend to agglomerate in some well-defined network neighborhood? We relied on two network-based measures to address this question. (i) We measured the network-based mean shortest distance<sup>15</sup> among 2, 644 expressed genes, finding  $\langle d_{\text{prefr.cort}} \rangle = 1.22$ , significantly smaller than the random expectation,  $\langle d_{\text{prefr.cort}}^{\text{random}} \rangle = 1.33$  ( $p$  value =  $8.9 \times 10^{-17}$ ,  $z$  score =  $-11.0$ ) (Fig. 1B). (ii) We next found that 75% of the expressed genes form a single connected subgraph (connected component in graph theory<sup>19</sup>) ( $S_{\text{prefr.cort}} = 2, 001$ , Fig. 1C). This value is significantly greater than the random expectation,  $S_{\text{prefr.cort}}^{\text{random}} = 1, 715$  ( $p$  value =  $8.7 \times 10^{-10}$ ,  $z$  score = 6.02) (Fig. 1C) (see Supplementary Section III).

Taken together, these results show that the genes expressed in the prefrontal cortex are localized in the same neighborhood of the interactome, and that a very significant fraction of them forms a single connected subgraph. This pattern is not unique to the prefrontal cortex. Indeed, we find that a significant fraction of expressed genes forms a single connected subgraph in 41 of the 64 tissues considered here (see Figs 2, S2 and S3), indicating that each tissue has a characteristic interactome neighborhood in which its proteins agglomerate.

Given that on average only 10% of all genes are expressed in any specific tissue, we next asked: are the genes expressed in different tissues localized in the same or different network neighborhoods? Consider, for example, the hypothalamus and the lung, with  $n_{\text{hypothalamus}} = 1, 354$  and  $n_{\text{lung}} = 1, 141$  genes expressed at  $z_E \geq 1.0$ . The number of genes simultaneously expressed in these tissues is  $n_{\text{hl}} = 84$ , significantly lower than the random expectation,  $n_{\text{hl}}^{\text{rand}} = 148$  (Jaccard index (see Supplementary Section II),  $J_{\text{hl}} = 0.035$ ,  $J_{\text{hl}}^{\text{random}} = 0.063$ ,  $p$  value =  $4.5 \times 10^{-9}$ ). We also find that the network-based mean shortest distance between gene pairs expressed in the two tissues,  $d_{\text{AB}} = 1.30$ , is significantly larger than random expectation  $d_{\text{AB}}^{\text{random}} = 1.26$  ( $p$  value = 0.034,  $z$  score = 1.82) (see Supplementary Section IV). Taken together, these results indicate that the genes expressed in the hypothalamus and the lung are distinct, and that their proteins are located in different interactome neighborhoods.

In contrast, genes expressed in the hypothalamus and prefrontal cortex exhibit both significant overlap and network-based colocalization:  $n_{\text{hypothalamus}} = 1, 354$  and  $n_{\text{prefr.cortex}} = 2, 644$ , with overlap of  $n = 992$ , significantly greater than the random expectation,  $n^{\text{rand}} = 343$  ( $J_{\text{hl}} = 0.33$ ,  $J_{\text{hl}}^{\text{random}} = 0.12$ ,  $p$  value =  $2.4 \times 10^{-14}$ ). In addition, the



**Figure 3. Tissue specificity of tauopathies.** (A–F,H,I) Expression levels of genes related to tauopathies in (A) amygdala, (B) hypothalamus, (C) prefrontal cortex, (D) spinal cord, (E) thalamus, (F) whole brain, (H) CD105-positive endothelial cells, and (I) lymph node. Node sizes are chosen to reflect the expression significance of the corresponding genes. Blue nodes correspond to genes with expression significance  $z_E \geq 1.0$ , while red nodes correspond to genes with  $z_E < 1.0$ . (G), A putative disease modules for tauopathies. The numbers show the number of specific nervous system tissues in which the corresponding gene is expressed.

mean shortest distance between gene expressed in the two tissues is  $d_{AB} = 0.69$ , significantly smaller than  $d_{AB}^{\text{random}} = 1.13$  ( $p$  value  $< 10^{-152}$ ,  $z$  score  $= -26.4$ ) (see Supplementary Section II and Fig. S4).

We systematically measured the network-based mean shortest distances  $d_{AB}$  between gene sets expressed at  $z_E \geq 1.0$  for all 2,016 possible tissue pairs. For 1,415 pairs, the network-based separation is significantly different from the random expectation. Of these, 851 tissue pairs are closer than expected by chance, while 564 pairs show a statistically significant separation. The obtained 1,415 separations allowed us to build a dendrogram for the hierarchy of tissue clusters (see Fig. S5), predicting three major clusters, one of which consists of several subclusters and contains brain regions, and the remaining contain immune system cells and reproductive tissues.

Taken together, we find that genes expressed in a particular tissue are not scattered randomly throughout the interactome, but tend to form a well-localized connected subnetwork. Subnetworks corresponding to similar tissues tend to overlap, but pathologically distinct tissue pairs tend to agglomerate in different neighborhoods of the interactome. Hence, we can divide the interactome into tissue-specific neighborhoods, a partition that has direct implications for our understanding of diseases, as we next demonstrate.

**Disease-Tissue Associations.** To illustrate how a disease manifests itself in a particular tissue, consider tauopathies, a class of neurodegenerative diseases associated with the pathological aggregation of  $\tau$  protein within neurons. We find that several genes associated with tauopathies, such as the *MAPT* gene (involved in production of  $\tau$  proteins), are expressed in many tissues, indicating that gene expression alone is a poor predictor of the tissue-specificity of the disease. According to our hypothesis, tauopathies manifest mainly in the brain because the subnetwork of proteins supporting the relevant molecular mechanisms for the disease are expressed integrally only in brain tissue.

In support of this hypothesis, we find that the expressed tauopathies genes connect to each other to form a statistically significant connected component only in six tissues, all of which belong to the nervous system (Fig. 3A–F): amygdala, hypothalamus, prefrontal cortex, spinal cord, thalamus, and whole brain. For comparison we also show the subnetwork of tauopathies-related genes in endothelial cells and lymph nodes, where the expressed disease genes do not form a significant connected component, a pattern characteristic of the remaining 60 tissues (Fig. 3H,I).

Combining these observations, we arrive at a putative disease module for tauopathies consisting of nine genes that form a connected subgraph ( Fig. 3G). Three of these, *MAPT*, *APP*, and *CLU*, are simultaneously expressed in three of the six brain regions (amygdala, prefrontal cortex, hypothalamus). A key role is played by the amyloid precursor protein *APP*, which not only is the central hub of the module, but also is expressed in all six brain tissues. This network-based conclusion is supported by previous experimental evidence. Indeed, the *APP* on chromosome 21q21.2 is the first causative gene identified for early onset Alzheimer's disease<sup>20</sup>, a disorder that belongs to the class of tauopathies. Furthermore, *APP* and *MAPT* are known to form soluble complexes that may promote the self aggregation of *APP* into the insoluble forms observed in Alzheimer's disease<sup>21</sup>. In the nervous system, the expression of *CLU* (clusterin) is elevated in neuropathological conditions, such as Alzheimer's disease, where *CLU* co-precipitates with *APP*, suggesting a physiological interaction<sup>22</sup>. Finally, recent large genome-wide association studies have identified loci not previously associated with the disease at the *CLU* (also known as *APOJ*) gene (rs11136000,  $p$  value =  $1.4 \times 10^{-9}$ )<sup>23</sup>. Since *TOMM40* is located on chromosome 19, closely adjacent to *APOE*, another gene known to be associated with Alzheimer's disease<sup>24</sup>, other investigations have suggested that the statistically significant correlation of *TOMM40* with Alzheimer's is due to linkage disequilibrium. By contrast, we find *APOE* in Alzheimer's disease and tauopathies in the connected subgraph and not *TOMM40*, which indicates a greater mechanistic relevance of *APOE* for the disease than *TOMM40*.

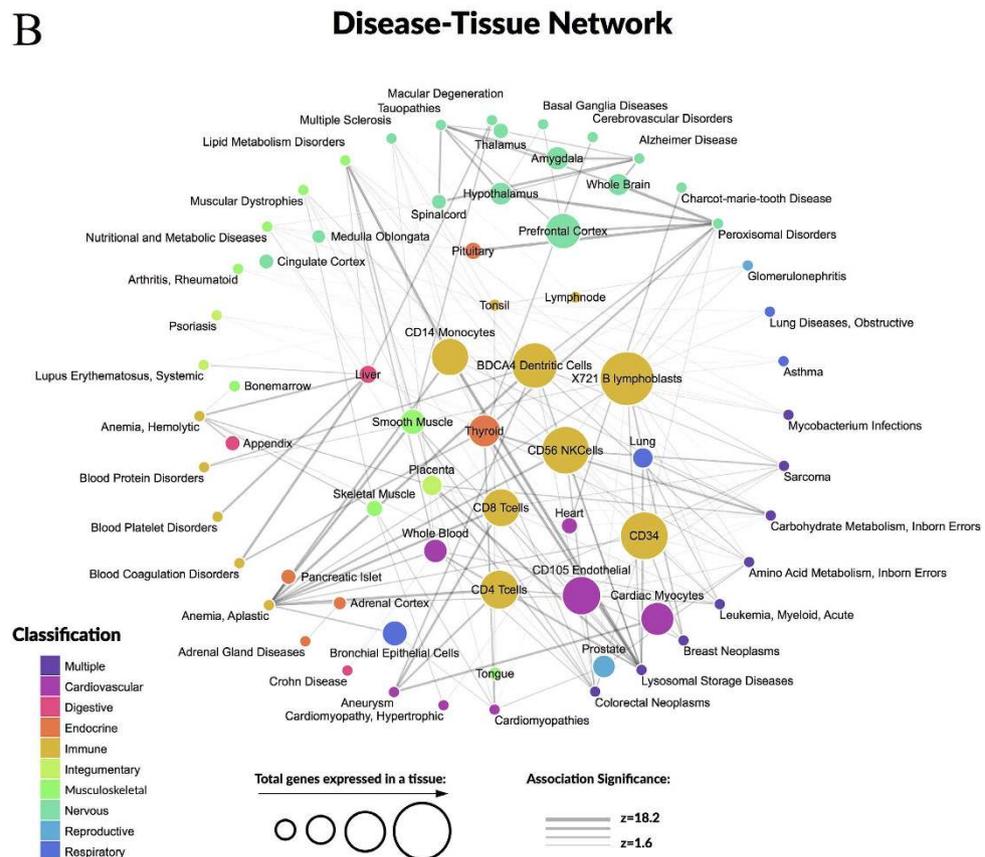
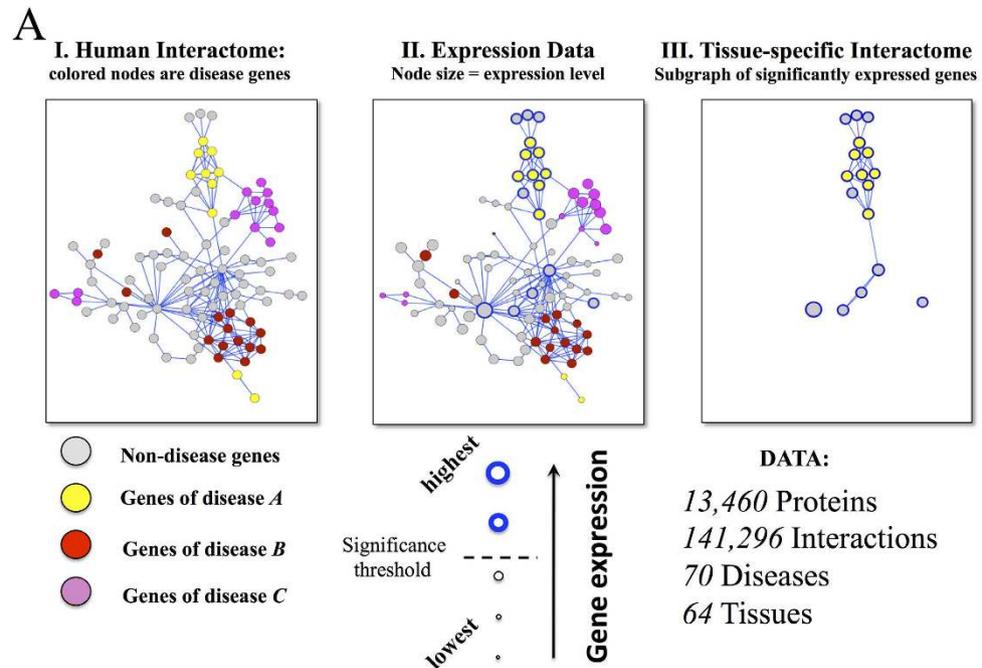
This example supports our hypothesis that a disease manifests itself only in tissues in which the expressed disease genes form a disease module, i.e., a statistically significant connected subgraph. It also expands earlier findings of elevated expression and protein interaction in specific tissues<sup>1</sup>, indicating that one needs to inspect the integrity of the full disease module to establish the tissue-specificity of a given disease. Building on this hypothesis, we identified all diseases that form a significant subgraph in a particular tissue, following the procedure depicted in Fig. 4A. This strategy allows us to build a disease-tissue bipartite network that links 70 diseases to 64 tissues via 187 links (see Fig. 4B, Supplementary Section V, and Fig. S6). While some diseases manifest in as many as 12 tissues, on average, each of the diseases included in the network manifests in 2.4 tissues. The tissues with the largest number of expressed diseases are BDCA4 dendritic cells (18 diseases), X721 B lymphoblast cells (16 diseases), and CD56 NK cells (12 diseases), appearing as the hubs of the disease-tissue network. For 35 diseases, we did not find a statistically significant module in any tissue, and of the 64 tissues included in this analysis, 29 tissues did not have any disease associated with them. Plausible reasons for these latter two findings are the incompleteness of the interactome<sup>25–28</sup> and the limited number of known disease genes.

In addition to a large number of expected disease-tissue associations (e.g., tauopathies and brain tissues), we also find a number of less obvious disease-tissue associations, such as macular degeneration and liver, or lipid metabolism disorders and CD14 monocytes. Next, we discuss several examples to illustrate the predictive power of the developed map and to validate some of the predicted associations.

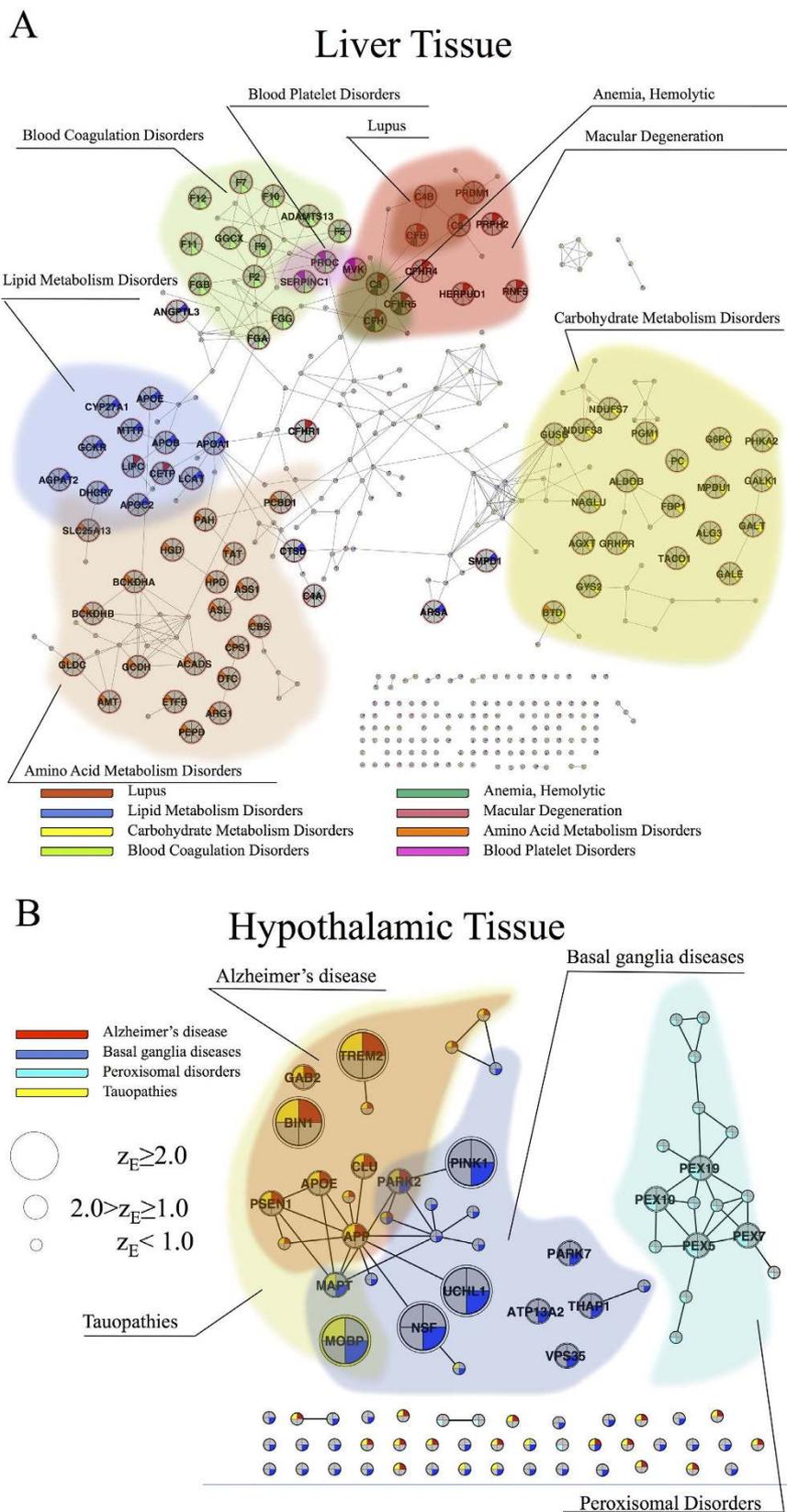
**Liver.** Eight diseases have a statistically significant module in liver tissue. Some, like blood coagulation disorders or certain types of anemia, are expected. Indeed, of the 40 genes related to blood coagulation disorders, 14 are expressed in liver and 12 of the expressed genes form a single connected subgraph ( $p$  value <  $10^{-53}$ ,  $z$  score = 15.4) (see Fig. 5A). Five of them, *F5*, *F7*, *F9*, *F10*, and *F11*, are parts of the extrinsic and intrinsic blood clotting pathways. Others, like macular degeneration, are somewhat more surprising. Interestingly, seven genes associated with macular degeneration (*CFH*, *C3*, *C2*, *CFHR5*, *CFB* and *CFHR4*, *CFHR1*) also form a statistically significant module in liver ( $p$  value <  $10^{-26}$ ,  $z$  score = 10.85, see Fig. 5A). Yet, while macular degeneration is an ophthalmologic disorder, there are plausible molecular reasons for this association with the network of expressed genes in liver. Indeed, genes in the complement pathway, including complement factor *H* (*CFH*), *C2/CF*, and *C3*, are known to be associated with age-related macular degeneration (AMD) (see Fig. 5A). The complement system in the blood of mammals comprises more than 30 proteins that are primarily synthesized in the liver and that circulate in their inactive forms. In addition, *CFH*, a major circulating protein, is mainly produced in liver<sup>29</sup>, and livers of patients with AMD are more likely to produce an abnormal form of *CFH*, which is thought to increase inflammation in the eye.

**Hypothalamus.** In the hypothalamus, only three brain-related diseases form significant modules: tauopathies ( $S = 6$ ,  $p$  value <  $10^{-37}$ ,  $z$  score = 12.9), Alzheimer's disease ( $S = 4$ ,  $p$  value <  $10^{-17}$ ,  $z$  score = 8.6), and basal ganglia diseases ( $S = 3$ ,  $p$  value =  $1.8 \times 10^{-7}$ ,  $z$  score = 5.1) (see Fig. 5B). We also find a significant module for peroxisomal disorders ( $S = 4$ ,  $p$  value <  $10^{-65}$ ,  $z$  score = 17.1), a class of conditions caused by defects in peroxisome function. Neurological dysfunction is a prominent feature of most peroxisomal disorders<sup>30</sup>. Peroxisomal disorders are divided into two groups, the peroxisome biogenesis disorders (PBDs) and the peroxisome single-enzyme peroxisome disorders (PSEDs). The connected component of peroxisomal disorders expressed in the hypothalamus consists of *PEX19*, *PEX10*, *PEX6*, and *PEX7*. *PEX7* encodes peroxisome targeting signal receptor, while *PEX19* is proposed to be essential for the proper localization and stability of peroxisomal membrane proteins. At the same time, *PEX6* is required for membrane fusion in an early step of peroxisome biogenesis<sup>31</sup>. While tauopathies and basal ganglia diseases have overlapping modules, peroxisome disorders are separated from them. Moreover, a correlation has been found between the level of peroxisome proliferation in hypothalamus and protection from *APP* (amyloid)-associated neurodegeneration, a process that is linked to Alzheimer's disease<sup>32</sup>. Current protein-protein interaction maps are estimated to cover 10–15% of all potential interactions<sup>26</sup>. Hence, the isolation of peroxisome disorders from tauopathies and basal ganglia diseases may be a reflection of this limited interactome coverage.

Taken together, the obtained disease-tissue bipartite network supports our hypothesis that disease modules are tissue-specific. To validate this disease-tissue network we examined diseases associated with liver and hypothalamus. While most of the observed associations, including blood coagulation disorders and anemia in liver, Alzheimer disease, Basal ganglia diseases in hypothalamus are well known, we also identified less known disease



**Figure 4. Disease-tissue bipartite network.** (A) From left to right: I. Human interactome is compiled and disease-related genes are identified. II. Gene expression data are mapped onto the human interactome. The expression level is reflected in the node size. III. A tissue-specific interactome is constructed as a subgraph of the human interactome consisting of genes expressed with significance  $z_E \geq 1.0$ . Disease *A* has significantly connected module within the tissue-specific interactome, while diseases *B* and *C* do not. (B) Disease-tissue bipartite network. Tissues are placed within the circle while diseases are positioned along the circumference. Nodes are colored according to tissue classification. The sizes of tissue nodes are proportional to the total number of genes expressed in them. The widths of connecting links correspond to the significance of the association. For higher resolution see Fig. S6.



**Figure 5. Diseases associated with liver and hypothalamus.** (A) Diseases associated with the liver-specific interactome. (B) Diseases associated with the hypothalamus-specific interactome. Node sizes depend on the expression significance of the corresponding genes. We use pie-chart coloring to reflect disease-gene associations. To improve the readability of diagrams, we grouped genes related to the same disease into modules and highlighted each module accordingly. We excluded disease genes associated with nutritional and metabolic diseases from both panels owing to the large number of associated genes.

tissue associations of macular degeneration in liver, and peroxisomal disorder in hypothalamus. Our finding also provides new routes to analyze tissue-specific disease modules, helping highlight specific disease genes and the molecular mechanisms mediating the pathobiological relationships between diseases.

**Tissue-specific disease modules.** Genes associated with the same disease are known to have similar biological characteristics<sup>33–35</sup>. Yet, the precise mechanistic role at the molecular level remains unknown for most disease genes. This prompts us to ask: Could tissue specificity help filter out genes that may have only limited mechanistic relevance to a disease? To answer this question, we start by inspecting the functional annotations of the disease genes according to the three Gene Ontology categories<sup>36</sup>: biological processes (bp), molecular function (mf), and cellular component (cc).

Functional similarity of two genes is defined as the similarity between the sets of their GO annotations<sup>37</sup>. To define functional similarity of two genes we use Simpson and Jaccard similarity coefficients (see Methods). We report the results for the Simpson coefficient in the main text; results obtained with Jaccard similarity are qualitatively similar and are reported in Supplementary Section IID. The functional similarity of a group of genes is then given by the average pairwise functional similarity. For example, for the 40 genes associated with blood coagulation disorders (BCD), the average pairwise biological process similarity is  $bp(BCD) = 0.36$ , the average molecular function similarity is  $mf(BCD) = 0.48$ , and the average cellular component similarity is  $cc(BCD) = 0.77$  (Supplementary Section IID). Yet, only 14 of the 40 BCD genes are expressed in liver, and these 14 have higher functional similarity values in all three categories:  $bp(BCD, liver) = 0.44$ ,  $mf(BCD, liver) = 0.63$ , and  $cc(BCD, liver) = 0.91$ , suggesting that the disease genes excluded by tissue specificity are those whose functional relatedness is smaller.

We, therefore, hypothesize that tissue specificity acts as a “cleansing” or filtering procedure, automatically eliminating the potentially false positive BCD genes (*Hypothesis A*). Alternatively, the observed increase in the functional similarity of disease genes could occur solely due to the fact that tissue-specific genes are functionally more uniform (*Hypothesis B*).

To test the validity of *Hypothesis B* we measured tissue-specific GO similarities for all disease-tissue pairs and calculated their *deviations* from GO similarities measured for all disease genes,  $\Delta bp(d, t) = bp(d, t) - bp(d)$ . Here  $bp(d, t)$  is the biological process similarity of genes associated with disease  $d$  that are expressed in tissue  $t$ , and  $bp(d)$  is the biological process similarity of all genes of disease  $d$ . We also define  $\Delta mf(d, t)$  and  $\Delta cc(d, t)$  in a similar way, using molecular function and cellular component categories, respectively (see Supplementary Section IID). Positive values of  $\Delta bp(d, t)$  indicate that tissue-specific disease genes are more similar to each other than all genes associated with the disease are similar to each other. For instance,  $\Delta bp(BCD, liver) = 0.08$ ,  $\Delta mf(BCD, liver) = 0.15$ , and  $\Delta cc(BCD, liver) = 0.14$ .

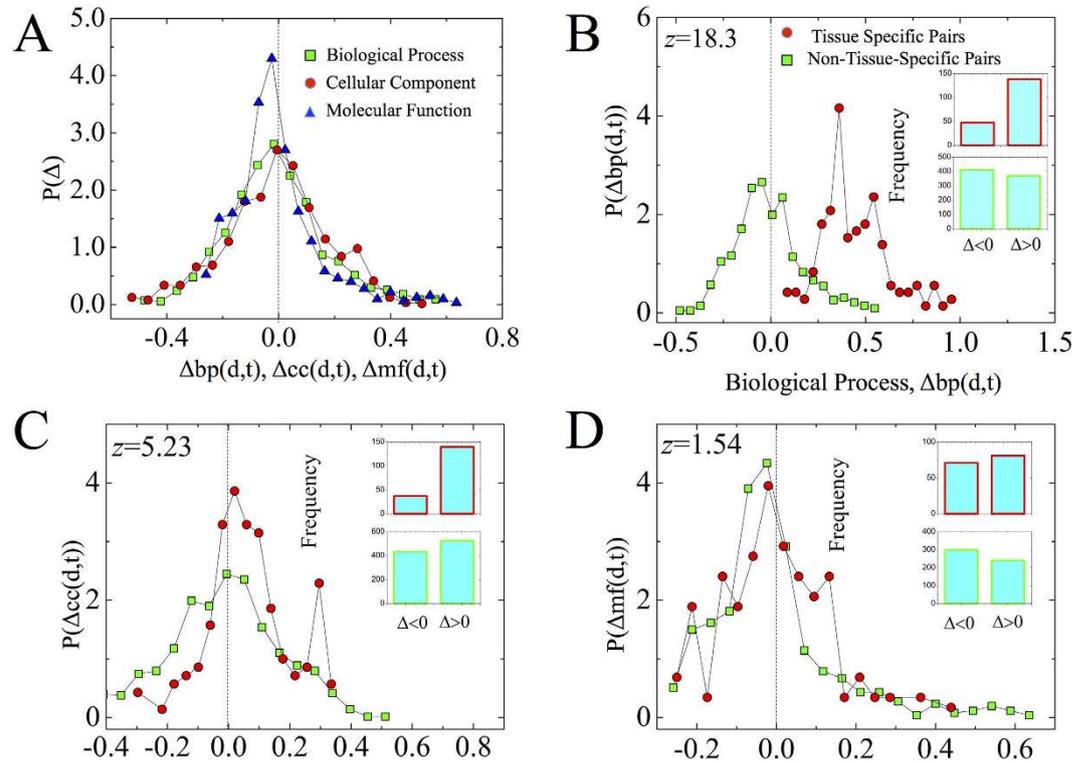
We first calculated the distributions of the deviations  $\Delta bp(d, t)$ ,  $\Delta mf(d, t)$ , and  $\Delta cc(d, t)$  for all possible disease-tissue pairs (Fig. 6A). For all GO categories, we have nearly equal numbers of disease-tissue pairs with positive and negative deviations, which results in median values close to zero (Fig. 6A). This finding indicates that the functional similarity of disease genes  $\Delta bp(d, t)$  may increase or decrease depending on the disease-tissue pair ( $d, t$ ) considered. Note that if hypothesis B were true, deviations would be positive for all disease-tissue pairs, including those diseases that have no significant module in the tissue. Hence, the results of Fig. 6A rule out *Hypothesis B*.

To test *Hypothesis A*, the expectation that disease genes are functionally more uniform only in tissues in which they have a significant connected component, we analyzed the distributions of  $\Delta bp(d, t)$ ,  $\Delta cc(d, t)$ , and  $\Delta mf(d, t)$  separately for disease-tissue pairs in which disease genes form a significant connected component (set CC) and in which they do not (set  $N - CC$ ). Even though the two distributions look visually similar (Fig. 6B–D), the Mann-Whitney U test performed to compare the distributions indicates that they are statistically different ( $p_{bp}$  value  $< 10^{-74}$ ). Of all disease-tissue pairs in CC, 70% have  $\Delta bp(d, t) > 0$ , and 30% have  $\Delta bp(d, t) < 0$  (see top inset of Fig. 6B). In the case of  $N - CC$ , we observe 39% of disease-tissue pairs with  $\Delta bp(d, t) > 0$  and 61% with  $\Delta bp(d, t) < 0$  (bottom inset of Fig. 6B). We, therefore, find that in 70% of tissues the disease genes become functionally more uniform in disease-tissue combinations for which the disease has a significant connected component. We obtained similar results for the cellular component category: ( $p_{cc}$  value  $= 1.7 \times 10^{-7}$ ), 79% of disease-tissue pairs have  $\Delta cc(d, t) > 0$ , while 21% have  $\Delta cc(d, t) < 0$  (top and bottom insets of Fig. 6C). In the case of  $N - CC$  we observe that 55% of disease-tissue pairs have  $\Delta cc(d, t) > 0$ , while 45% have  $\Delta cc(d, t) < 0$  (top and bottom insets of Fig. 6C). These results did not hold for the molecular functions GO categories (Fig. 6D): ( $p_{mf}$  value  $= 0.17$ ). To test the robustness of our results we repeated calculations using the Jaccard coefficient as the similarity measure. Obtained results are qualitatively similar to those obtained with the Simpson coefficient (see Supplementary Section IID).

Taken together, we find that disease genes expressed in a tissue in which a given disease has a significant connected component are functionally more uniform. Consequently, reducing the disease module to a specific tissue acts as a natural filtering procedure, automatically eliminating the functionally and potentially mechanistically less relevant disease genes.

To demonstrate how tissue-specificity can be used to filter out less relevant disease gene associations we consider genes associated with macular degeneration and arthritis. Figure 7A depicts the correlation between the significance of GWAS associations and gene expression levels in liver for macular degeneration. Of the 15 genes expressed at  $z_E < 1.0$ , 9 are characterized by relatively low GWAS significance ( $-\text{Log}(p) < 6$ ). Of the 11 genes expressed at  $z_E \geq 1.0$ , 7 have GWAS significance of ( $-\text{Log}(p) > 20$ ) and 4 have GWAS significance of ( $7 < -\text{Log}(p) < 12$ ) (Fig. 7B).

To understand better the liver-specific interactions among the expressed genes we construct the GWAS-based subnetwork of macular degeneration genes (Fig. 7B). We find that 7 genes with GWAS significance  $-\text{Log}(p) > 20$  (C2, C3, CFB, CFH, CFHR1, CFHR4 and CFHR5) form a connected component, while the remaining 4 expressed



**Figure 6. Functional similarity of tissue-specific disease genes.** (A–D), Distributions of changes in GO similarities: (A), The distributions of  $\Delta bp(d, t)$ ,  $\Delta cc(d, t)$ , and  $\Delta mf(d, t)$  obtained for all disease-tissue pairs. Measured mean similarity values are, respectively,  $M(\Delta bp) = 7.6 \times 10^{-3}$ ,  $M(\Delta cc) = 0.034$ , and  $M(\Delta mf) = -6.3 \times 10^{-3}$ . (B),  $\Delta bp(d, t)$ , (C),  $\Delta cc(d, t)$ , and (D)  $\Delta mf(d, t)$  for linked disease-tissue pairs (red circles) and non-linked disease-tissue pairs (green squares). The insets in panels (B–D) compare the total number of  $\Delta > 0$  cases with the total number of  $\Delta < 0$  cases. Listed  $z$ -scores are the results of the Mann-Whitney U test applied to the comparison of GO similarities of linked disease-tissue pairs and non-linked disease tissue pairs.

genes (*LIPC*, *RNF5*, *CETP*, *HERPUD1*) with lower GWAS significance are not connected ( Fig. 7B). Taken together, Fig. 7 indicates that the tissue-specific disease module acts as an effective filter, aggregating the diseases genes with high GWAS significance and keeping less relevant genes apart from the disease module.

The analysis of arthritis-related genes expressed in BDCA4 cells is performed in Supplementary Section VI. The examples of macular degeneration and arthritis demonstrate the possibility of using tissue-specificity of disease modules to filter out less relevant disease genes. Yet, at this time, we can not systematically explore tissue-specific filtering effects on all diseases owing to the limited knowledge of GWAS disease-gene associations and the incompleteness of the human interactome.

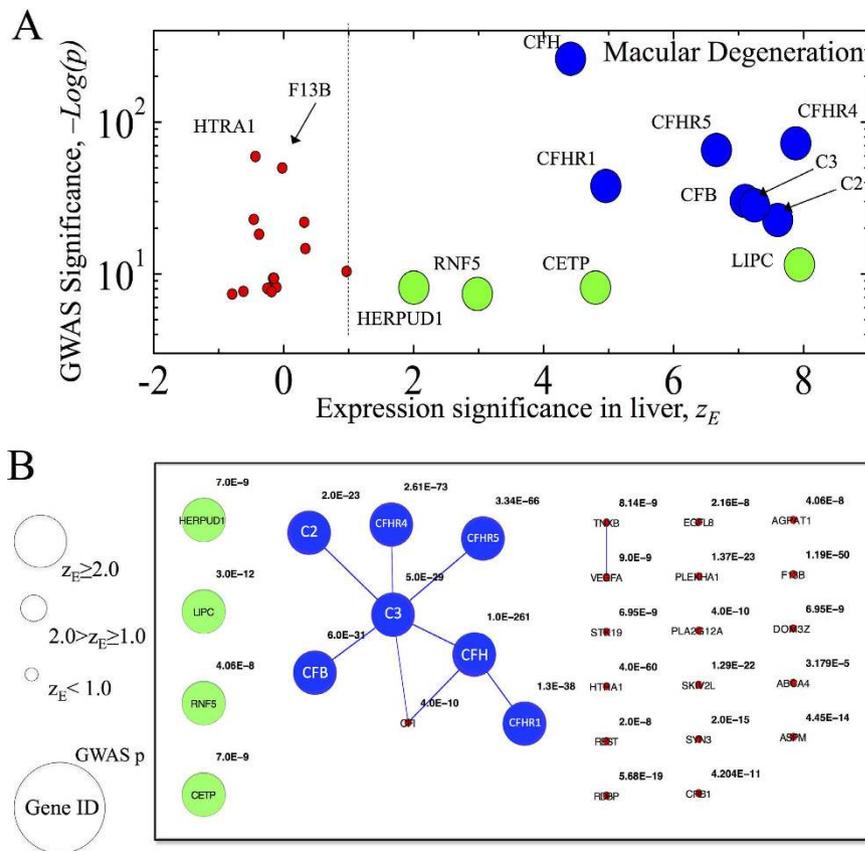
## Discussion

In this paper we offer a disease module-based approach towards understanding the tissue specificity of human diseases. We showed that for a disease to manifest itself in a particular tissue, a statistically significant functional *subnetwork* of genes associated with the disease needs to be expressed in that tissue. This approach lead to the construction of the disease-tissue network that offers a predictive map of the statistically significant disease-tissue associations. This approach allowed us to examine known disease-tissue relationships and to predict newly definable disease-tissue associations. We also showed that expressed disease genes tend to be more functionally similar if the disease manifests itself in that tissue. These observations can be used as an additional test of the relevance of individual genes to disease.

Throughout this paper, we used gene expression to define tissue-specific interactomes and diseases modules. We considered a gene expressed in a particular tissue when its expression level exceeds the significance threshold of  $z_E^* = 1.0$ . To probe the robustness of our key results, we repeated the analysis for different values of  $z_E^*$  (Supplementary Section V, Figs S4 and S7), finding that:

(i) The localization of disease genes in a particular network-based neighborhood of the human interactome is significant for a wide range of  $z_E^*$  values. The higher  $z_E^*$ , the stronger is the localization, as measured by the mean shortest distance among expressed genes.

(ii) The separations between the subnetworks of expressed genes corresponding to dissimilar tissues increases as  $z_E^*$  increases.



**Figure 7. Tissue-specific filtering.** To demonstrate the filtering effect of tissue-specificity we consider genes associated with macular degeneration. Node sizes correspond to gene expression significance in liver tissue. Blue nodes correspond to genes in the connected component of the macular degeneration with expression significance of  $z_E > 1.0$ . Green nodes correspond to genes with  $z_E > 1.0$  which are not part of the macular degeneration connected component. Red nodes correspond to genes with  $z_E < 1.0$ . (A), The scatter plot of GWAS association significance as a function of gene expression significance in liver. (B), Macular degeneration genes and links among them. Numeric values correspond to GWAS association significance. The most relevant macular degeneration genes form the connected component of the macular degeneration. Note that there are two genes with  $z_E < 1.0$  that have high GWAS significance scores: *HTRA1* with  $p$ -value of  $p = 4 \times 10^{-60}$  ( $-\text{Log}(p) = 59$ ) and *F13B* with  $p = 1.210 \times 10^{-50}$  ( $-\text{Log}(p) = 50$ ). This observation can be explained as follows. The locus of  $p = 4 \times 10^{-60}$  is close to two the genes *ARMS2* and *HTRA1*. Genes expressed in the liver are from complementary and lipid metabolism pathways. At the same time, the *ARMS2* is from inflammatory pathway<sup>38</sup>. *HTRA1* is not expressed in the liver as per NCBI UniGene. We also note that the apparent association for *CFHR2-5* and *F13B* could be due to their proximity to *CFH*<sup>39</sup>. This, again, supports our observation that *F13B* is not expressed in the liver.

(iii) A large portion of the identified disease-tissue associations are significant even in the case of a more stringent threshold of  $z_E^* = 2.0$ . The bipartite disease-tissue network obtained for  $z_E^* = 2.0$  is an almost exact subset of the original network obtained for  $z_E^* = 1.0$  (Supplementary Section V and Fig. S7).

Finally, as we demonstrated in Figs 7 and S8, combining the tissue-specific modules with the interactome allowed us to filter out the less relevant disease genes for arthritis and macular degeneration. Consequently, our findings could considerably improve the predictive power of the interactome for diseases, allowing the construction of more accurate disease modules.

## Methods

**Significance of Gene Expression.** We use expression data from 64 non-diseased tissues. We convert probe ids to gene ids using the *U133* annotation platform. From 13,460 proteins, only 10,434 proteins were annotated to probe ids, the basis of our subsequent analysis.

To quantify the expression significance of gene  $i$  in tissue  $t$ , we calculate the average expression ( $E(i)$ ) and the standard deviation of a gene's expression across all considered tissues  $\sigma_E(i)$ . The significance of gene expression in tissue  $t$  is defined as

$$z_E(i, t) = (E(i, t) - \langle E(i) \rangle) / \sigma_E(i). \quad (1)$$

**Tissue Specificity of Disease.** To test if disease  $d$  has a significant connected component in tissue  $t$ , we first compile the tissue-specific interactome that consists of genes with expression significance  $z_E \geq 1.0$  in tissue  $t$  and interactions between them. Next, we map genes associated with disease  $d$  onto the tissue-specific interactome and then measure the size of the largest connected component  $S$  and the total number of disease genes  $S_{total}$  expressed in tissue  $t$  (Fig. 3A).

To test the significance of the observed disease module we assume that disease genes do not preferentially interact in the tissue-specific interactome. With this null hypothesis, we select  $S_{total}$  genes randomly in the tissue-specific interactome and determine the resulting size of the largest connected component  $S^{rand}$ . We repeat the same procedure 1,000 times to obtain the distribution  $P^{rand}(S)$ . Assuming the normality of  $P^{rand}(S)$ , we use the  $z$ -score to compute the significance of the real data with the threshold of  $z \geq 1.6$  for modules to be larger than expected by chance.

**Gene Ontology Similarity of Disease Modules.** The Gene Ontology database<sup>36</sup> annotates genes and their products with specific molecular functions (mf), biological processes (bf), and cellular components (cc). The molecular function similarity of two genes,  $i$  and  $j$ , is defined as the Simpson similarity of their molecular function terms,  $S_{mf}(i)$  and  $S_{mf}(j)$ :

$$mf_{ij} = \frac{|S_{mf}(i) \cap S_{mf}(j)|}{\min(|S_{mf}(i)|, |S_{mf}(j)|)} \quad (2)$$

We define the molecular function similarity of a disease  $d$ , as the average pairwise molecular function similarity of genes related to this disease:

$$mf_{full}(d) = \frac{2}{N(d)(N(d)-1)} \sum_{i \neq j} mf_{ij}, \quad (3)$$

where  $N(d)$  is the number of genes associated with diseases  $d$  and the summation runs through all pairs of genes associated with disease  $d$ . Similarly, we define the tissue-specific molecular function similarity of disease  $d$  as

$$mf(d, t) = \frac{2}{N(d, t)(N(d, t)-1)} \sum_{i \neq j} mf_{ij}, \quad (4)$$

where  $N(d, t)$  is the number of genes associated with disease  $d$  that are expressed in tissue  $t$ , and the summation runs through all pairs of disease-related genes expressed in tissue  $t$ . Biological process and cellular component similarities are defined similarly using, respectively, biological process or cellular component annotation terms.

## References

- Greene, C. S. *et al.* Understanding multicellular function and disease with human tissue-specific networks. *Nature genetics* (2015).
- Barshir, R., Shwartz, O., Smoly, I. Y. & Yeger-Lotem, E. Comparative analysis of human tissue interactomes reveals factors leading to tissue-specific manifestation of hereditary diseases. *PLoS computational biology* **10**, e1003632 (2014).
- Barabási, A.-L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics* **12**, 56–68 (2011).
- Ideker, T. & Sharan, R. Protein networks in disease. *Genome research* **18**, 644–652 (2008).
- Zanzoni, A., Soler-López, M. & Aloy, P. A network medicine approach to human disease. *FEBS letters* **583**, 1759–1765 (2009).
- Feldman, I., Rzhetsky, A. & Vitkup, D. Network properties of genes harboring inherited disease mutations. *Proceedings of the National Academy of Sciences* **105**, 4323–4328 (2008).
- Baranzini, S. E. *et al.* Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. *Human molecular genetics* **18**, 2078–2090 (2009).
- Wuchty, S. *et al.* Gene pathways and subnetworks distinguish between major glioma subtypes and elucidate potential underlying biology. *Journal of biomedical informatics* **43**, 945–952 (2010).
- Reverter, A., Ingham, A. & Dalrymple, B. P. Mining tissue specificity, gene connectivity and disease association to reveal a set of genes that modify the action of disease causing genes. *BioData Min* **1**, 8 (2008).
- Lage, K. *et al.* A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proceedings of the National Academy of Sciences* **105**, 20870–20875 (2008).
- Goh, K.-I. *et al.* The human disease network. *Proceedings of the National Academy of Sciences* **104**, 8685–8690 (2007).
- Magger, O., Waldman, Y. Y., Rupp, E. & Sharan, R. Enhancing the prioritization of disease-causing genes through tissue specific protein interaction networks. *PLoS computational biology* **8**, e1002690 (2012).
- Rual, J.-F. *et al.* Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173–1178 (2005).
- Stelzl, U. *et al.* A human protein-protein interaction network: a resource for annotating the proteome. *Cell* **122**, 957–968 (2005).
- Menche, J. *et al.* Uncovering disease-disease relationships through the incomplete interactome. *Science* **347** (2015).
- Su, A. I. *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 6062–6067 (2004).
- Mottaz, A., Yip, Y. L., Ruch, P. & Veuthey, A.-L. Mapping proteins to disease terminologies: from uniprot to mesh. *BMC bioinformatics* **9**, S3 (2008).
- Ramos, E. M. *et al.* Phenotype-genotype integrator (phegeni): synthesizing genome-wide association study (gwas) data with existing genomic resources. *European Journal of Human Genetics* **22**, 144–147 (2014).
- Caldarelli, G. *Scale-free networks: complex webs in nature and technology* (Oxford University Press, 2007).
- Kang, J. *et al.* The precursor of alzheimer's disease amyloid a4 protein resembles a cell-surface receptor. *Nature* **325**, 733–736 (1987).
- Guo, J.-P., Arai, T., Miklossy, J. & McGeer, P. L. A $\beta$  and tau form soluble complexes that may promote self aggregation of both into the insoluble forms observed in alzheimers disease. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 1953–1958 (2006).
- Boggs, L. N. *et al.* Clusterin (apo j) protects against *in vitro* amyloid- $\beta$  (1-40) neurotoxicity. *Journal of neurochemistry* **67**, 1324–1327 (1996).

23. Harold, D. *et al.* Genome-wide association study identifies variants at *CLU* and *PICALM* associated with Alzheimer's disease. *Nature genetics* **41**, 1088–1093 (2009).
24. Lyall, D. M. *et al.* Alzheimer's disease susceptibility genes *APOE* and *TOMM40*, and brain white matter integrity in the Lothian Birth Cohort 1936. *Neurobiology of Aging* **35**, 1513–e25 (2014).
25. Mosca, R., Pons, T., Céol, A., Valencia, A. & Aloy, P. Towards a detailed atlas of protein-protein interactions. *Current Opinion in Structural Biology* **23**, 929–940 (2013).
26. Rolland, T. *et al.* A proteome-scale map of the human interactome network. *Cell* **159**, 1212–1226 (2014).
27. Hart, G. T., Ramani, A. K. & Marcotte, E. M. How complete are current yeast and human protein-interaction networks. *Genome Biol* **7**, 120 (2006).
28. Mestres, J., Gregori-Puigjané, E., Valverde, S. & Sole, R. V. Data completeness—the Achilles heel of drug-target networks. *Nature biotechnology* **26**, 983–984 (2008).
29. Friese, M. *et al.* Different regulation of factor H and fHL-1/reconectin by inflammatory mediators and expression of the two proteins in rheumatoid arthritis (RA). *Clinical & Experimental Immunology* **121**, 406–415 (2000).
30. Powers, J. M. & Moser, H. W. Peroxisomal disorders: genotype, phenotype, major neuropathologic lesions, and pathogenesis. *Brain Pathology* **8**, 101–120 (1998).
31. Hashiguchi, N. *et al.* Peroxisomes are formed from complex membrane structures in *inx6*-deficient CHO cells upon genetic complementation. *Molecular Biology of the Cell* **13**, 711–722 (2002).
32. Santos, M. J. *et al.* Peroxisomal proliferation protects from  $\beta$ -amyloid neurodegeneration. *Journal of Biological Chemistry* **280**, 41057–41068 (2005).
33. Jimenez-Sanchez, G., Childs, B. & Valle, D. Human disease genes. *Nature* **409**, 853–855 (2001).
34. Karlin, S., Brocchieri, L., Bergman, A., Mrázek, J. & Gentles, A. J. Amino acid runs in eukaryotic proteomes and disease associations. *Proceedings of the National Academy of Sciences* **99**, 333–338 (2002).
35. López-Bigas, N. & Ouzounis, C. A. Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Research* **32**, 3108–3114 (2004).
36. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nature genetics* **25**, 25–29 (2000).
37. Bass, J. I. F. *et al.* Using networks to measure similarity between genes: association index selection. *Nature Methods* **10**, 1169–1176 (2013).
38. Parmeggiani, F. *et al.* Mechanism of inflammation in age-related macular degeneration. *Mediators of Inflammation* **2012** (2012).
39. Scheetz, T. E. *et al.* A genome-wide association study for primary open angle glaucoma and macular degeneration reveals novel loci. *PLoS One* **8**, e58657 (2013).

## Acknowledgements

We thank M. Vidal, A. Kamburov and R. Wang for many discussions of the manuscript; and K. Albrecht for assistance in figure design. This work was supported by NIH grants P50-HG004233, U01-HG001715, and U01-HG007690 from NHGRI, P50-GM107618 from NIGMS and PO1-HL083069, R37-HL061795, RC2-HL101543, and U01-HL108630 from NHLBI.

## Author Contributions

M.K., A.S., J.L. and A.-L.B. designed research; M.K. performed research and analyzed data; M.K., A.S. and A.-L.B. wrote the manuscript; M.K., A.S., J.M., E.G., S.D.G., J.L., and A.-L.B. discussed the results and reviewed the manuscript; A.-L.B. supervised the project.

## Additional Information

**Supplementary information** accompanies this paper at <http://www.nature.com/srep>

**Competing financial interests:** The authors declare no competing financial interests.

**How to cite this article:** Kitsak, M. *et al.* Tissue Specificity of Human Disease Module. *Sci. Rep.* **6**, 35241; doi: 10.1038/srep35241 (2016).



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

© The Author(s) 2016