RESEARCH

Scope and limitations of yeast as a model organism for studying human tissue-specific pathways

Shahin Mohammadi[1*], Baharak Saberidokht[1], Shankar Subramaniam[2] and Ananth Grama[1]

*Correspondence:
mohammadi@purdue.edu
[1]Department of Computer Sciences, Purdue University, 47907 West Lafayette, USA
Full list of author information is available at the end of the article

Abstract

Background: Budding yeast, S. cerevisiae, has been used extensively as a model organism for studying cellular processes in evolutionarily distant species, including humans. However, different human tissues, while inheriting a similar genetic code, exhibit distinct anatomical and physiological properties. Specific biochemical processes and associated biomolecules that differentiate various tissues are not completely understood, neither is the extent to which a unicellular organism, such as yeast, can be used to model these processes within each tissue.

Results: We present a novel framework to systematically quantify the suitability of yeast as a model organism for different human tissues. To this end, we develop a computational method for dissecting the global human interactome into tissue-specific cellular networks. By individually aligning these networks with the yeast interactome, we simultaneously partition the functional space of human genes, and their corresponding pathways, based on their conservation both across species and among different tissues. Finally, we couple our framework with a novel statistical model to assess the conservation of tissue-specific pathways and infer the overall similarity of each tissue with yeast. We further study each of these subspaces in detail, and shed light on their unique biological roles in the human tissues.

Conclusions: Our framework provides a novel tool that can be used to assess the suitability of the yeast model for studying tissue-specific physiology and pathophysiology in humans. Many complex disorders are driven by a coupling of housekeeping (universally expressed in all tissues) and tissue-selective (expressed only in specific tissues) dysregulated pathways. While tissue-selective genes are significantly associated with the onset and development of a number of tissue-specific pathologies, we show that the human-specific subset has even higher association. Consequently, they provide excellent candidates as drug targets for therapeutic interventions.

Keywords: human tissue-specific pathways; protein-protein interaction (PPI); network alignment; baker's yeast

Background

Budding yeast, *S. cerevisiae*, is widely used as an experimental system, due to its ease of manipulation in both haploid and diploid forms, and rapid growth compared to animal models. Coupled with the continuous development of new experimental methodologies for manipulating various aspects of its cellular machinery, it has served as the primary model organism for molecular and systems biology[1]. Moti-

vated by the availability of its full genome in 1996 as the first eukaryotic organism to be sequenced[2], an array of functional genomics tools emerged, including a comprehensive collection of yeast deletion mutants[3, 4], genome-wide over-expression libraries[5], and green fluorescent protein (GFP)-tagged yeast strains[6, 7]. The maturity of yeast's genetic and molecular toolbox has, in turn, positioned it as the primary platform for development of many high-throughput technologies, including transcriptome [8, 9, 10], proteome [11], and metabolome [12, 13] screens. These -*omic* datasets, all originally developed in yeast, aim to capture dynamic snapshots of the state of biomolecules during cellular activities. With the advent of "systems modeling", a diverse set of methods have been devised to assay the interactions, both physical and functional, among different active entities in the cell, including protein-protein[14, 15, 16], protein-DNA[17, 18], and genetic[19, 20, 21] interactions. These interactions, also referred to as the *interactome*, embody a complex network of functional pathways that closely work together to modulate the cellular machinery. Comparative analysis of these pathways relies on network alignment methods, much the same way as sequence matching and alignments are used for individual genes and proteins. Network alignments use both the homology of genes, as well as their underlying interactions, to project functional pathways across different species[22, 23, 24, 25]. These methods have been previously applied to detection of ortholog proteins, projection of functional pathways, and construction of phylogenetic trees.

Yeast and humans share a significant fraction of their functional pathways that control key aspects of eukaryotic cell biology, including the cell cycle [26], metabolism[27], programmed cell death[28, 29], protein folding, quality control and degradation[30], vesicular transport[31], and many key signaling pathways, such as mitogen-activated protein kinase (MAPK)[32, 33], target of rapamycin (TOR)[34], and insulin/IGF-I[35] signaling pathways. In the majority of cases, yeast has been the model organism in which these pathways were originally identified and studied. These conserved biochemical pathways drive cellular growth, division, trafficking, stress-response, and secretion, among others, all of which are known to be associated with various human pathologies. This explains the significant role for yeast as a model organism for human disorders[36, 37, 38]. Yeast has contributed to our understanding of cancers[39, 40, 41] and neurodegenerative disorders[42, 43, 44]. Having both chronological aging (amount of time cells survive in post-mitotic state) and replicative aging (number of times a cell can divide before senescence occurs), yeast is also used extensively as a model organism in aging research. It has contributed to the identification of, arguably, more human aging genes than any other model organism[45].

Depending on the conservation of the underlying pathways, there are two main approaches to studying them in yeast. It has been estimated that, out of 2,271 known disease-associated genes, 526 genes ($\sim$ 23%) have a close ortholog in the yeast genome, spanning 1 out of every 10 yeast genes[46]. For these orthologous pairs of disease-associated genes, we can directly increase the gene dosage of the endogenous yeast protein by using overexpression plasmids, or decrease it, through either gene knockout or knockdown experiments, in order to study gain- or loss-of-function phenotypes, respectively. A key challenge in phenotypic screens is that

disrupting genes, even when they have close molecular functions, can result in characteristically different organism-level phenotypes. *Phenologs*, defined as phenotypes that are related by the orthology of their associated genes, have been proposed to address this specific problem[47]. A recent example of such an approach is the successful identification of a highly conserved regulatory complex implicated in human leukemia[48]. This complex, named COMPASS (Complex of Proteins Associated with Set1), was originally identified by studying protein interactions of the yeast Set1 protein, which is the ortholog of the human mixed-lineage leukemia (MLL) gene, and years later was shown to be conserved from yeast to fruit flies to humans. On the other hand, if the disease-associated gene(s) in humans does not have close orthologs in yeast, heterologous expression of the human disease-gene in yeast, also referred to as *"humanized yeast"*, can be used to uncover conserved protein interactions and their context, to shed light on the molecular mechanisms of disease development and progression. For the majority of disease-genes with known yeast orthologs, heterologous expression of the mammalian gene is functional in yeast and can compensate for the loss-of-function phenotype in yeast deletion strains[1]. This approach has already been used to construct humanized yeast model cells to study cancers[39], apoptosis-related diseases[49], mitochondrial disorders[50], and neurodegenerative diseases[43]. Perhaps one of the more encouraging examples is the very recent discovery of a new compound, N-aryl benzimidazole (NAB), which strongly protects cells from α-synuclein toxicity in the humanized yeast model of Parkinson's disease[51]. In a follow-up study, they tested an analog of the NAB compound in the induced pluripotent stem (iPS) cells generated from the neuron samples of Parkinson's patients with α-synuclein mutations. They observed that the same compound can reverse the toxic effects of α-synuclein aggregation in neuron cells[52]. Using this combined phenotypic screening, instead of the traditional target-based approach, they were not only able to discover a key compound targeting similar conserved pathways in yeast and humans, but also uncover the molecular network that alleviates the toxic effects of α-synuclein. These humanized yeast models have also been used to study human genetic variations[53].

Various successful instances of target identification, drug discovery, and disease network reconstruction using humanized yeast models have established its role as a model system for studying human disorders. When coupled with more physiologically relevant model organisms to cross-validate predictions, yeast can provide a simple yet powerful first-line tool for large-scale genetic and chemical screening[41, 43]. However, as a unicellular model organism, yeast fails to capture organism-level phenotypes that emerge from inter-cellular interactions. Perhaps, more importantly, it is unclear how effectively it can capture tissue-specific elements that make a tissue uniquely susceptible to disease. All human tissues inherit the same genetic code, but they exhibit unique functional and anatomical characteristics. Similar sets of molecular perturbations can cause different tissue-specific pathologies given the network context in which the perturbation takes place. For example, disruption of energy metabolism can contribute to the development of neurodegenerative disorders, such as Alzheimer's, in the nervous system, while causing cardiomyopathies in muscle tissues[54]. These context-dependent phenotypes are driven by genes that are specifically or preferentially expressed in one or a set of biologically relevant tissue types,

also known as *tissue-specific* and *tissue-selective* genes, respectively. Disease genes, and their corresponding protein complexes, have significant tendencies to selectively express in tissues where defects cause pathology[55, 56]. How tissue-selective pathways drive tissue-specific physiology and pathophysiology is not completely understood; neither is the extent to which we can use yeast as an effective model organism to study these pathways.

We propose a quantitative framework to assess the scope and limitations of yeast as a model organism for studying human tissue-specific pathways. Our framework is grounded in a novel statistical model for effectively assessing the similarity of each tissue with yeast, considering both expressed genes and their underlying physical interactions as a part of functional pathways. To understand the organization of human tissues, we present a computational approach for partitioning the functional space of human proteins and their interactions based on their conservation both across species and among different tissues. Using this methodology, we identify a set of *core genes*, defined as the subset of the most conserved housekeeping genes between humans and yeast. These core genes are not only responsible for many of the fundamental cellular processes, including translation, protein targeting, ribosome biogenesis, and mRNA degradation, but also show significant enrichment in terms of viral infectious pathways. On the other hand, human-specific housekeeping genes are primarily involved in cell-to-cell communication and anatomical structure development, with the exception of mitochondrial complex I, which is also human-specific. Next, we identify comprehensive sets of tissue-selective functions that contribute the most to the computed overall similarity of each tissue with yeast. These conserved, tissue-selective pathways provide a comprehensive catalog for which yeast can be used as an effective model organism. Conversely, human-specific, tissue-selective genes show the highest correlation with tissue-specific pathologies and their functional enrichment resembles highly specific pathways that drive normal physiology of tissues.

Comparative analysis of yeast and human tissues to construct conserved and non-conserved functional tissue-specific networks can be used to elucidate molecular/ functional mechanisms underlying dysfunction. Moreover, it sheds light on the suitability of the yeast model for the specific tissue/ pathology. In cases where suitability of yeast can be established, through conservation of tissue-specific pathways in yeast, it can serve as an experimental model for further investigations of new biomarkers, as well as pharmacological and genetic interventions.

Results and discussion

In this section, we present our comparative framework for investigating the scope and limitations of yeast as a model organism for studying tissue-specific biology in humans. Figure 1 illustrates the high-level summary of our study design. We start by aligning each of the human tissue-specific networks with the yeast interactome. We couple the alignment module with a novel statistical model to assess the significance of each alignment and use it to infer the respective similarity/ dissimilarity of human tissue-specific networks with their corresponding counterparts in yeast. Using a network of tissue-tissue similarities computed using their transcriptional profile, we show that our network alignment p-values are consistent with groupings derived

from transcriptional signatures. We use this network of tissue similarities to identify four major groups of tissues/ cell-types. These groups; representing brain tissues, blood cells, ganglion tissues, and testis-related tissues; are further used to identify tissue-selective genes that are active within each group compared to the rest of tissues.

We partition both housekeeping and tissue-selective subsets of human genes separately into the conserved and human-specific subsections. We provide extensive validation for the selective genes with respect to blood cells and brain tissues. Figure 2 illustrates the overall partitioning of the genes and their relative subsets. We provide an in-depth analysis of each of these subsets, and show that while conserved subsets provide the *safe zone* for which yeast can be used as an ideal model organism, the human-specific subset can shed light on the *shadowed subspace* of the human interactome in yeast. This subset can provide future directions for constructing humanized yeast models.

Aligning yeast interactome with human tissue-specific networks

The *global* human interactome represents a static snapshot of potential physical interactions that *can* occur between pairs of proteins. However, it does not provide any information regarding the spatiotemporal characteristics of the actual protein interactions. These interactions have to be complemented with a dynamic *context*, such as expression measurements, to help interpret cellular rewiring under different conditions.

[57] overlaid the mRNA expression level of each transcript (transcriptome) in different human tissues[58] on top of the *global* human interactome, integrated from 21 PPI databases, and constructed a set of 79 reference tissue-specific networks. We adopt these networks and align each one of them separately to the yeast interactome that we constructed from the BioGRID database.

In order to compare these human tissue-specific networks with the yeast interactome, considering both the sequence similarity of proteins and the topology of their interactions, we employ a recently proposed sparse network alignment method, based on the Belief Propagation (BP) approach. This method is described in the Materials and methods section[59].

Genes, and their corresponding proteins, do not function in isolation; they form a complex network of interactions among coupled biochemical pathways in order to perform their role(s) in modulating cellular machinery. Moreover, each protein may be involved in multiple pathways to perform a diverse set of functions. Using a network alignment approach to project these pathways across species allows us to not only consider their first-order dynamics, through co-expression of homologous protein pairs, but also the context in which they are expressed.

To construct the state space of potential homologous pairs, we align all protein sequences in human and yeast and pre-filter hits with sequence similarity E-values greater than 10. For genes with multiple protein isoforms we only store the most significant hit. Using these sequence-level homologies, we construct a matrix L that encodes pairwise sequence similarities between yeast and human proteins. Entries in matrix L can be viewed as edge weights for a bipartite graph connecting human genes on one side, and the yeast genes, on the other side. We use this matrix to

restrict the search space of the BP network alignment method (please see Supplementary Methods for details on E-value normalization and Materials and Methods section for BP alignment method).

Parameters α and $\beta(= 1 - \alpha)$ control the relative weight of sequence similarity (scaled by α) as compared to topological conservation (scaled by β) in the BP network alignment. Using a set of preliminary simulations aligning the global human interactome with its tissue-specific sub-networks, for which we have the *true* alignment, with various choices of α in the range of 0.1 to 0.9, we identify the choices of $\alpha = \frac{1}{6}$ and $\beta = \frac{5}{6}$ to perform the best in our experiments. We use the same set of parameters to align each tissue-specific network with the yeast interactome, as it provides a balanced contribution from sequence similarities and the number of conserved edges. The final set of all alignments is available for download as Additional file .

Investigating roles of housekeeping genes and their conservation across species

Housekeeping genes comprise a subset of human genes that are universally expressed across all tissues and are responsible for maintaining core cellular functions needed by all tissues, including translation, RNA processing, intracellular transport, and energy metabolism[60, 61, 62]. These genes are under stronger selective pressure, compared to tissue-specific genes, and evolve more slowly[63]. As such, we expect to see a higher level of conservation among human housekeeping genes compared with yeast genes. We refer to the most conserved subset of housekeeping genes between humans and yeast, computed using network alignment of tissues-specific networks with the yeast network, as the *core genes*.

We identify a gene as housekeeping if it is expressed in *all* 79 tissues. We identify a total of 1,540 genes that constitute the shared section of human tissue-specific networks. These genes, while having similar set of interactions among each other, are connected differently to the set of tissue-selective genes.

Using the alignment partners of all housekeeping genes in the yeast interactome, we construct an alignment consistency table of size $1,540 \times 79$, which summarizes the network alignments over the shared subsection of tissue-specific networks. Then, we use the majority voting method to classify housekeeping genes as *core*, which are conserved in yeast, *human-specific*, which are consistently unaligned across human tissues, and *unclassified*, for which we do not have enough evidence to classify it as either one of the former cases.

Network alignments are noisy and contain both false-positive (defined as aligned pairs that are not functionally related), as well as false-negatives (pairs of functional orthologs that are missed in the alignment). These errors can come from different sources, including gene expression data (node errors), interactome (edge errors), or the alignment procedure (mapping errors). We propose a method based on majority voting across different alignments to (partially) account for these errors. Given a set of network alignments, we consider a pair of entities consistently aligned (either matched or unmatched) if they are consistent in at least $100 * \tau\%$ of alignments in the set. The parameter τ, called the *consensus rate*, determines the level of accepted disagreement among different alignments. A higher value of consensus rate increases the precision of the method at the cost of decreased sensitivity. In order to select

the optimal consensus rate parameter, we tried values in range $[0.5 - 1.0]$ with increments of $\frac{1}{2}$. We identified the parameter choice of $\tau = 0.9$, equivalent to 90% agreement among aligned tissues, to perform the best in classifying human-specific and conserved genes, while keeping the sets well-separated. Using this approach, we were able to tri-partition 1,540 housekeeping genes into 595 conserved, 441 human-specific, and 504 unclassified genes, respectively. The complete list of these genes is available for download as Additional file .

In order to investigate the conserved sub-network of core genes, we construct their alignment graph as the Kronecker product of the subgraph induced by core genes in the human interactome and its corresponding aligned subgraph in yeast. Conserved edges in this network correspond to interologs, i.e., orthologous pairs of interacting proteins between yeast and human[64]. The final alignment graph of the core housekeeping genes is available for download as Additional file .

Figure 3 shows the largest connected component of this constructed alignment graph. We applied the MCODE[65] network clustering algorithm on this graph to identify highly interconnected regions corresponding to putative protein complexes. We identified five main clusters, which are color-coded on the alignment graph, and are shown separately on the adjacent panels. Ribosome is the largest, central cluster identified in the alignment graph of core genes, and together with proteasome and spliceosome, constitutes the three most conserved complexes in the alignment graph. This complex is heavily interconnected to the eIFs, to modulate eukaryotic translation initiation, as well as proteasome, which controls protein degradation. Collectively, these complexes regulate protein turnover and maintain a balance between synthesis, maturation, and degradation of cellular proteins.

In order to further analyze the functional roles of these housekeeping genes, we use the g:Profiler[66] R package to identify highly over-represented terms. Among functional classes, we focus on the gene ontology (GO) biological processes, excluding electronic annotations, KEGG pathways, and CORUM protein complexes to provide a diverse set of functional roles. We use the Benjamini-Hochberg procedure to control for false-discovery rate (FDR), with p-value threshold of $\alpha = 0.05$, and eliminate all enriched terms with more than 500 genes to prune overly generic terms. Using this procedure, we identify enriched functional terms for both core and human-specific subsets of housekeeping genes. The complete list of enriched functions for different classes of housekeeping genes is available for download as Additional file .

We manually group the most significant terms (p-value $\leq 10^{-10}$) in core genes, which results in five main functional classes, namely ribosome biogenesis, translation, protein targeting, RNA splicing, and mRNA surveillance. First, we observe a one-to-one mapping between enriched terms and identified putative complexes corresponding to translation initiation (p-value $= 7.1 * 10^{-17}$) and ribosome (p-value $= 5.97 * 10^{-11}$). In addition, translation termination and elongation are also enriched with decreasing levels of significance. Moreover, these processes are tightly linked to SRP-dependent co-translational protein targeting (p-value $= 2.7 * 10^{-15}$). This, in turn, suggests protein synthesis as one of the most conserved aspects of eukaryotic cells. Next, we note that both mRNA splicing (p-value $= 7.04 * 10^{-10}$) and nonsense-mediated decay (p-value $= 4.66 * 10^{-16}$) are also enriched among the

most significant functional terms, which supports our earlier hypothesis related to the role of splicesome in the alignment graph of core genes. Finally, we find that the most significant functional term, as well as a few other related terms, are involved in viral infection, which suggests that (a subset of the) core genes provides a *viral gateway* to mammalian cells. This can be explained in light of two facts: i) viral organisms rely on the host machinery for their key cellular functions, and ii) housekeeping genes are more ancient compared to tissue-selective genes, and core genes provide the most conserved subset of these housekeeping genes. As such, these genes may contain more conserved protein interaction domains and be structurally more "familiar" as interacting partners for the viral proteins and provide ideal candidates for predicting host-pathogen protein interactions.

Next, we perform a similar procedure for the human-specific housekeeping genes. This subset, unlike core genes, is mostly enriched with terms related to anatomical structure development and proximal cell-to-cell communication (paracrine signaling), with the exception of complex I of the electron transport chain, which is the strongest identified term. This NADH-quinone oxidoreductase is the largest of the five enzyme complexes in the respiratory chain of mammalian cells. However, this complex is not present in yeast cells and has been replaced with a single subunit NADH dehydrogenase encoded by gene NDI1. Impairment of complex I has been associated with various human disorders, including Parkinson's and Huntington's disease. Transfecting complex I-defective cells with yeast NDI1 as a therapeutic agent has been proposed as a successful approach to rescue complex I defects[67, 68]. This technique, also known as *NDI1 therapy*, opens up whole new ways in which yeast can contribute to the research and development on human diseases: not only yeast can be used as a model organism, but also can provide candidates that can be used for gene therapy in mammalian cells.

A key observation here is that the human-specific subset of housekeeping genes is not only associated with fewer functional terms, but is also less significantly associated with these terms. This effect can be attributed to two factors. First, we note that some of the genes predicted to be human-specific might be an artifact of the method. For example, the belief propagation (BP) method enforces sequence similarity as a necessary, but not sufficient, condition for a pair of genes to be aligned, which means that any human gene with no sequence similarity to yeast genes will not be aligned, resulting in genes being artificially classified as human-specific. Second, and more importantly, a majority of functional annotations for human genes are initially attributed in other species, specially yeast, and transferred across ortholog groups. Based on our construction, human-specific genes are defined as the subset of housekeeping genes with no orthology with yeast. As such, it can be expected that these genes span the *shadowed subspace* of the functional space of human genes that is under-annotated.

Quantifying similarity of human tissues with yeast

Housekeeping genes are shared across all human tissues and cell types. They provide a conserved set of functions that are fundamental to cellular homeostasis. However, these genes do not provide direct insight into how different tissues utilize these key functions to exhibit their dynamic, tissue-specific characteristics. To assess the

similarity of each tissue with yeast, we propose a novel statistical model, called *tissue-specific random model (TRAM)*, which takes into account the ubiquitous nature of housekeeping genes and mimics the topological structure of tissue-specific networks (please see Materials and Methods section for the details of the random model).

We use the alignment score of each tissue-yeast pair as the objective function. To asses the significance of each alignment score, we use a Monte Carlo simulation method to sample from the underlying probability distribution of alignment scores.

For each tissue-specific network, we sample $k_{\mathcal{R}} = 10,000$ pseudo-random tissues of the same size from TRAM, separately align them with the yeast interactome, and compute the number of conserved edges and sequence similarity of aligned protein pairs as alignment statistics, in order to compute the empirical p-values. For each network alignment, we compute a *topological*, a *homological* (sequence-based), and a *mixed* (alignment score) p-value. Additionally, we use cases in which alignment quality is significantly better in the original tissue alignment, both in terms of sequence and topology, to quantify an *upper bound* on the alignment p-values. Conversely, cases in which both of these measures are improved in the random samples can be used to define a *lower bound* on the alignment p-value. The final table of alignment p-values is available for download as Additional file .

First, we note that all tissues with significant *mixed* p-values also have both significant topological and homological (sequence-based) p-values. For a majority of tissues with insignificant *mixed* p-values, we still observe significant homological, but insignificant topological p-values. We summarize the most and the least similar tissues to yeast by applying a threshold of $\alpha_l = \alpha_u = 10^{-2}$ to the p-value upper and lower bounds, respectively. There are a total of 23/79 tissues that have $\Delta_{\mathcal{R}} \leq 10^{-2}$ (p-value upper bound), listed in Table 1, which show the most significant similarity to the yeast interactome. Among these, blood cells show coherent high significance, with not even a single instance from 10,000 samples having either the alignment weight or the overlap of the random sample exceeding the original alignment. Similarly, blood/immune cell lines consistently show significant alignment p-values. Male reproductive tissues also show a strong similarity to yeast cells. Conversely, there are 19/79 tissues with $10^{-2} < \delta_{\mathcal{R}}$ (p-value lower bound), which show the least significant similarity to yeast. Among these tissues, listed in Table 2, ganglion tissues consistently show the least similarity to yeast. An interesting observation is that tissues and cell types at either end of the table (either the most or the least similar) usually have very high confidence values, i.e., both their topology and homology p-values are consistent.

Identifying groups of coherent tissues

Next, we investigate the correlation between the similarity of human tissues among each other and how it relates to their corresponding alignment p-values with yeast in order to better understand the transitivity of this relationship. We expect that similar tissues should exhibit consistent alignment p-values, resulting in groups of homogenous tissues with coherent alignments scores.

To this end, we first construct a network of tissue-tissue similarities (TTSN) using the global transcriptome of human tissues from the GNF Gene Atlas, including

44,775 human transcripts covering both known, as well as predicted and poorly characterized genes. For each pair of tissues/ cell types, we compute a similarity score using the Pearson correlation of their transcriptional signatures and use the 90th percentile of similarity scores to select the most similar pairs. We annotate each node in the TTSN with its corresponding alignment p-value as a measure of similarity with the yeast interactome. This meta-analysis allows us to investigate how linear measurements of gene/protein activity project to the space of protein interactions, in order to re-wire the underlying interactome in each human tissue.

Figure 4 presents the final network. In this network, each node represents a human tissue/cell type and each weighted edge illustrates the extent of overall transcriptional similarity between pairs of tissues. This network is filtered to include only tissue pairs with the highest overlap with each other. In order to assign color to each node, we use z-score normalization on the log-transformed alignment mixed p-values. Green and red nodes correspond to the highly positive and highly negative range of z-scores, which represent similar and dissimilar tissues to yeast, respectively.

Preliminary analysis of this network indicates that the alignment p-value of tissues highly correlates with their overall transcriptional overlap. Furthermore, these high-level interactions coincide with each other and fall within distinct *groups* with consistent patterns. We manually identified four such groups and separately annotated them in the network. These groups correspond to brain tissue, blood cells, ganglion tissues, and testis tissues. Among these groups, blood cells and testis tissues exhibit consistent similarity with yeast, whereas brain and ganglion tissues bear consistent dissimilarity.

The existence of homogenous group of tissues with consistent similarity with yeast suggests an underlying conserved machinery in these clusters. This raises the question of what is consistently aligned within each tissue group and how it relates to the computed alignment p-values? We address this question, and relate it to the onset of tissue-specific pathologies in the remaining subsections.

Dissecting tissue-selective genes with respect to their conservation

In this subsection, we investigate the subset of non-housekeeping genes in each homogenous group of human tissues and partition them into sets of genes, and their corresponding pathways that are either conserved in yeast or are human-specific. Next, we analyze how these pathways contribute to the overall similarity/ dissimilarity of human tissues with yeast.

Figure 5 presents the probability density function for the membership distribution of non-housekeeping genes in different human tissues. The observed bi-modal distribution suggests that most non-housekeeping genes are either expressed in a very few selected tissues or in the majority of human tissues. We use this to partition the set of expressed non-housekeeping genes, with the goal of identifying genes that are selectively expressed in each group of human tissues.

We start with all *expressed non-housekeeping genes* in each tissue group, i.e., genes that are expressed in *at least* one of the tissue members. Next, in order to identify the subset of expressed genes that are *selectively* expressed in each group, we use the *tissue-selectivity* p-*value* of each gene. In this formulation, a gene is

identified as selectively expressed if it is expressed in a significantly higher number of tissues in the given group than randomly selected tissue subsets of the same size (see Materials and Methods section for details). Figure 6 illustrates the distribution of tissue-selectivity p-values of expressed genes with respect to four major groups in Figure 4. Each of these plots exhibit a bi-modal characteristic similar to the membership distribution function in Figure 5. This can be explained by the fact that membership distribution is a mixture distribution, with the underlying components being the same distribution for the subset of genes that are expressed in different tissue groups. We use critical points of the p-value distributions to threshold for tissue-selective genes. The motivation behind our choice is that these points provide shifts in the underlying distribution, from tissue-selective to ubiquitous genes. Given the bi-modal characteristics of these distributions, they all have three critical points, the first of which we use as our cutoff point. This provides highest precision for declared tissue-selective genes, but lower recall than the other two choices.

Having identified the subset of tissue-selective genes with respect to each tissue group, we use the majority voting scheme to tri-partition these sets based on their alignment consistency with yeast. Similar to the procedure we used to tri-partition housekeeping genes, we tried different choices of consensus rate parameter from 50% to 100% with increments of 5%. The percent of unclassified genes decreases linearly with the consensus rate, while relative portions of human-specific/ conserved genes remain the same. We chose 90% for our final results, as it leaves the least number of genes as unclassified, while keeping human-specific and conserved genes well-separated. The set of all tissue-specific genes is available for download as Additional file .

Table 3 presents the number of expressed genes, selectively expressed genes, and the percent of tissue-selective genes that are conserved, human-specific, or unclassified within each group of tissues. There is a similar relationship between the ratio of conserved/human-specific genes within each group of tissues and their alignment p-values, suggesting that alignment p-values are highly correlated with the conservation of tissue-selective genes and their corresponding pathways. Figure 7 illustrates the relative sizes of each subset of genes identified in this study.

Conserved genes and their corresponding pathways comprise the functional subspace in which we can use yeast as a suitable model organism to study tissue-specific physiology and pathophysiology. On the other hand, human-specific genes provide a complementary set that can be used to construct *tissue-engineered* humanized yeast models. They also provide promising candidates for tissue-specific gene therapies in a similar fashion to NDI1 therapy, in cases where an alternative functional mechanism can be found in yeast. To further investigate these subsets, we focus on blood cells and brain tissues, which illustrate the clearest separation between their tissue-selective and conserved genes in their TSS distribution, and subject them to more in depth functional analysis in next subsections.

Elucidating functional roles of the brain and blood selective genes

We use g:ProfileR on both human-specific and conserved genes to identify their enriched functions. The complete list of enriched functions is available for download as Additional file . These two subsets share many common terms, due to the

underlying prior of both being subsets of tissue-selective genes. To comparatively analyze these functions and rank them based on their human-specificity, we use the log of p-value ratios between human-specific and conserved genes to filter terms that are at least within 2-fold enrichment. We focus on GO biological processes, KEGG pathways, and CORUM protein complexes and remove all genesets with more than 500 genes to filter for overly generic terms. Finally, to group these terms together and provide a visual representation of the functional space of genes, we use EnrichmentMap (EM)[69], a recent Cytoscape[70] plug-in, to construct a network (map) of the enriched terms. We use the log ratio of p-values to color each node in the graph. Figures 8 and 9 illustrate the final enrichment map of unique human-specific and conserved blood-selective and brain-selective functions, respectively.

Conserved blood-selective functions, shown in Figure 8 (A), are primarily enriched with terms related to DNA replication, cellular growth, and preparing cell for cell-cycle. Among these terms, DNA replication-is tightly linked to both DNA repair and telomere maintenance related terms. Telomere maintenance, specially via telomerase enzyme, is one of the cellular functions that is known to be conserved in yeast, but only active in a selected subset of differentiated human tissues and cell types, including hematopoietic stem cells and male reproductive tissues [71]. Functional terms involved in DNA conformation changes, including condensin complex, as well as cell cycle phase transition, specifically from G1 to S phases, are two other groups of conserved functional terms that are highly conserved from yeast to human. On the other hand, human-specific blood-selective functions, shown in Figure 8 (B), are mainly involved in lymphocyte proliferation and activation. Terms in these two groups are also tightly related to each other and form a larger cluster together. In addition, cytokine production and T-cell mediated cytotoxicity also exhibit human-specific, blood-selective characteristics. This is partially expected, as these functions are highly specialized immune-cell functions that are evolved particularly in humans to ensure his survival in less-favorable conditions.

Figure 9 (A) shows the functional space of conserved brain-selective functions. Many of these terms correspond to various aspects of brain development, including olfactory bulb, telencephalon, pallium, and cerebral cortex development, as well as the regulatory circuit that controls nervous system development. Considering the unicellular nature of yeast, the exact mechanisms in which orthologs of these pathways modulate yeast cellular machinery is less studied. An in-depth analysis to identify matching phenologs can help us use yeast to study various disorders related to brain development. Another functional aspect that exhibits high conservation is the mTOR complex 2. The target of rapamycin (TOR) signaling is a highly conserved pathway, which forms two structurally distinct protein complexes, mTORC1 and mTORC2. The former complex has a central role in nutrient-sensing and cell growth, and as such, has been used extensively to study calorie restriction (CR) mediated lifespan extension. On the other hand, mTORC2 has been recently proposed to modulate consolidation of long-term memory[72]. Cholesterol biosynthesis and transport is another conserved functional aspect that differs significantly from other human tissues. As the most cholesterol-rich organ in the body, expression of genes corresponding to lipoprotein receptors and apolipoproteins is tightly regulated among different brain cells and plays an important role in normal brain

development. Dysregulation of these metabolic pathways is implicated in various neurological disorders, such as Alzheimer's disease[73].Finally, microtubular structure and tubulin polymerization also shows significant conservation and is known to play a key role in brain development[74]. These cytoskeletal proteins have recently been associated with brain-specific pathologies, including epilepsy[75].

Finally, we study human-specific brain functions, which are shown in Figure 9 (B). One of the key functional aspects in this group is the semaphorin-plexin signaling pathway. This pathway was initially characterized based on its role in the anatomical structure maturation of the brain, specifically via the repulsive axon guidance, but later was found to be essential for morphogenesis of a wide range of organ systems, including sensory organs and bone development[76]. Another human-specific signaling pathway identified in brain is the glutamate receptor signaling pathway, which also cross-talks with circadian entrainment, as well as neuron-neuron transmission. This pathway plays a critical role in neural plasticity, neural development and neurodegeneration[77]. It has also been associated with both chronic brain diseases, such as schizophrenia, as well as neurodegenerative disorders, such as Alzheimer's disease[78].

Both conserved and human-specific genes play important roles in tissue-specific pathologies. In addition, these genes, which are enriched with regulatory and signaling functions, cross-talk with housekeeping genes to control cellular response to various factors. As such, a complete picture of disease onset, development, and progression can only be achieved from a systems point of view. From this perspective, we study not only the genes (or their states) that are frequently altered in disease, but also the underlying tissue-specific and housekeeping pathways in which they interact to exhibit the observed phenotype(s). In the next subsection, we further investigate this hypothesis. We study the potential of different subsets of the identified tissue-selective genes for predicting tissue-specific pathologies.

Assessing the significance of tissue-specific pathologies among conserved and human-specific tissue-selective genes

To further study the predictive power of tissue-selective genes for human pathologies, we use the *genetic association database (GAD)* disease annotations as our gold standard[79]. This database collects gene-disease associations from genetic association studies. Additionally, each disease has been assigned to one of the 19 different disease classes in GAD database. We use DAVID functional annotation tool for disease enrichment analysis of tissue-selective genes[80].

First, we seek to identify which disease classes are significantly enriched among each set of tissue-selective genes. Table 4 shows the disease classes enriched in each group of brain and blood selective genes. Conserved blood-selective genes are predominantly enriched with cancers, whereas human-specific blood-selective genes are mainly associated with immune disorders. This can be linked to our previous results indicating that conserved subset is mainly involved in regulating growth, DNA replication, and cell cycle, whereas human-specific genes are primarily involved in lymphocyte proliferation and activation. Conversely, brain-selective genes show higher similarities in terms of disease classes that they can predict. Both conserved and human-specific brain-selective genes can predict psychiatric disorders,

but human-specific subset seems to be a more accurate predictor. On the other hand, neurological disorders are only enriched in human-specific subset of brain-selective genes, whereas disorders classified as pharmacogenomic and chemdependency show higher enrichment in conserved genes.

To summarize the specific disorders that are enriched in each subset of brain-selective genes, we integrate all identified diseases and rank them based on their enrichment p-value, if it is only enriched in one set, or their most significant p-value, if it is enriched in both sets. Table 5 shows the top 10 disease terms enriched in either human-specific or conserved brain-selective genes. In majority of cases, human-specific genes are more significantly associated with brain-specific pathologies than conserved genes. In addition, there are unique disorders, such as schizophrenia, bi-polar disorder, and seizures, that are only enriched among human-specific genes.

In conclusion, both conserved and human-specific subsets of tissue-selective genes are significantly associated with different human disorders. However, the human-specific subset shows higher association with tissue-specific pathologies. To this end, they guide us to appropriate molecular constructs (gene insertions) in yeast to explore molecular/functional mechanisms that cause tissue-specific dysfunction. Such mechanisms can be tested in humans, and if validated, yeast can serve as an experimental model for further investigations of biomarkers and pharmacological and genetic interventions.

Conclusions

In this study, we demonstrated a novel methodology for aligning tissue-specific interaction networks with the yeast interactome and assess their statistical significance. We demonstrated that these alignments can be used to dissect tissue-specific networks into their core component and tissue-specific components. Tissue specific components were used for multiple purposes: (i) by showing that a number of pathologies manifest themselves in dysregulated genes in the tissue-specific group, we motivate exploration of these genes as particularly suitable candidates as drug targets; (ii) by quantifying the alignment of tissue-specific components with yeast, we quantify the suitability of yeast as a model organism for studying corresponding disease/ phenotype; (iii) in cases where there is (statistically) insignificant alignment, it is still possible to use yeast as a model organism, if the dysregulated pathways are aligned; and (iv) in cases where none of these conditions hold, our alignments provide mechanisms for assessing the feasibility of different molecular constructs (gene insertions) for creating more appropriate, tissue-specific, humanized yeast models.

Materials and methods

Datasets

Protein-protein interaction (PPI) networks

We adopted human tissue-specific networks from *Bossi et al.*[57]. They integrated protein-protein interactions from 21 different databases to create the whole human interactome consisting of 80,922 interactions among 10,229 proteins. Then, they extracted the set of expressed genes in each tissue from GNF Gene Atlas and used it to construct the tissue-specific networks, defined as the vertex-induced subgraphs

of the entire interactome with respect to the nodes corresponding to the expressed genes in each tissue.

Additionally, we obtained the yeast interactome from the BioGRID[81] database, update 2011[82], version 3.1.94, by extracting all physical interactions, excluding interspecies and self interactions. This resulted in a total of 130,483 (76,282 non-redundant) physical interactions among 5,799 functional elements in yeast (both RNA and protein). Next, we downloaded the list of annotated CDS entries from the Saccharomyces Genome Database (SGD)[83] and restricted interactions to the set of pairs where both endpoints represent a protein-coding sequence, i.e., protein-protein interactions. The final network consists of 71,905 interactions between 5,326 proteins in yeast and is available for download as Additional file .

Protein sequence similarities between yeast and humans

We downloaded the protein sequences for yeast and humans in FASTA format from Ensembl database, release 69, on Oct 2012. These datasets are based on the GRCh37 and EF4 reference genomes, each of which contain 101,075 and 6,692 protein sequences for *H. Sapiens* and *S. Cerevisiae*, respectively. Each human gene in this dataset has, on average, 4.49 gene products (proteins). We identified and masked low-complexity regions in protein sequences using *pseg* program[84]. The *ssearch36* tool, from *FASTA*[85] version 36, was then used to compute the local sequence alignment of the protein pairs using the Smith-Waterman algorithm[86]. We used this tool with the BLOSUM50 scoring matrix to compute sequence similarity of protein pairs in humans and yeast. All sequences with E-values less than or equal to 10 are recorded as possible matches, which results in a total of 664,769 hits between yeast and human proteins. For genes with multiple protein isoforms, coming from alternatively spliced variants of the same gene, we only record the most significant hit. The final dataset contains 162,981 pairs of similar protein-coding genes, and is available for download as Additional file .

Sparse network alignment using belief propagation

Analogous to the sequence alignment problem, which aims to discover conserved genomic regions across different species, network alignment is motivated by the need for extracting shared functional pathways that govern cellular machinery in different organisms. The network alignment problem in its abstract form can be formulated as an optimization problem with the goal of identifying an optimal mapping between the nodes of the input networks, which maximizes both sequence similarity of aligned proteins and conservation of their underlying interactions. At the core of every alignment method are two key components: i) a scoring function and ii) an efficient search strategy to find the optimal alignment. The scoring function is usually designed to favor the alignment of similar nodes, while simultaneously accounting for the number of conserved interactions between the pair of aligned nodes. Biologically speaking, this translates to identifying functional *orthologs* and *interologs*, respectively.

Given a pair of biological networks, $\mathcal{G} = (\mathcal{V}_\mathcal{G}, \mathcal{E}_\mathcal{G})$ and $\mathcal{H} = (\mathcal{V}_\mathcal{H}, \mathcal{E}_\mathcal{H})$, with $n_\mathcal{G} = |V_G|$ and $n_\mathcal{H} = |V_H|$ vertices, respectively, we can represent the similarity of vertex pairs between these two networks using a weighted bipartite graph

$\mathcal{L} = (\mathcal{V}_\mathcal{G} * \mathcal{V}_\mathcal{H}, \mathcal{E}_\mathcal{L}, \boldsymbol{w})$, where $\boldsymbol{w} : \mathcal{E}_\mathcal{L} \rightarrow \mathcal{R}$ is a weight function defined over edges of $\mathcal{L}$. We will denote mapping between vertices $v_i \in \mathcal{V}_\mathcal{G}$ and $v_{i'} \in \mathcal{V}_\mathcal{H}$ with (i, i') and ii', interchangeably. Let us encode the edge conservations using matrix $\mathbf{S}$, where $\mathbf{S}(ii', jj') = 1$, iff alignment of $v_i \rightarrow v_{i'}$ together with $v_j \rightarrow v_{j'}$ will conserve an edge between graphs $\mathcal{G}$ and $\mathcal{H}$, and $\mathbf{S}(ii', jj') = 0$, otherwise. Then, the network alignment problem can be formally represented using the following integer quadratic program:

$$\max_{\boldsymbol{x}} \quad (\alpha \boldsymbol{w}^T \boldsymbol{x} + \frac{\beta}{2} \boldsymbol{x}^\mathbf{T} \mathbf{S} \boldsymbol{x}) \tag{1}$$

$$\text{Subject to:} \quad \begin{cases} \mathbf{C}\boldsymbol{x} \leq \mathbf{1}_{n_\mathcal{G} * n_\mathcal{H}} & \text{Matching constraints;} \\ x_{ii'} \in \{0, 1\}, & \text{Integer constraint.} \end{cases}$$

Here, $\mathbf{C}$ and $\boldsymbol{w}$ are the incidence matrix and edge weights of the graph $\mathcal{L}$, respectively, whereas $\boldsymbol{x}$ is the matching indicator vector. Vector $\boldsymbol{w}$, which encodes the *prior* knowledge of node-to-node similarity between the input pair of networks, defines the search space of *potential orthologs* and can be computed using sequence, structural, or functional similarity of the proteins corresponding to node pairs. In this study, we chose sequence similarity of aligned protein sequences to assign edge weights in the bipartite graph defined by $\mathcal{L}$. When $\mathcal{L}$ is a complete bipartite graph, i.e. each pair of vertices between $\mathcal{G}$ and $\mathcal{H}$ represents a viable ortholog candidate, we will have $\mathbf{S} = \mathcal{G} \otimes \mathcal{H}$. However, *Bayati et al.*[59] recently proposed an efficient method, based on the message passing algorithm, for cases where $\mathcal{L}$ is sparse, i.e., $|\mathcal{E}_\mathcal{L}| << n_\mathcal{G} * n_\mathcal{H}$, by restricting the search space to the subset of promising candidates that are provided by $\mathcal{E}_\mathcal{L}$. We will use this algorithm throughout this paper for solving the network alignment problem.

Tissue-specific random model (TRAM) for generating pseudo-random tissues

Let us denote the global human interactome by $\mathcal{G} = (\mathcal{V}_\mathcal{G}, \mathcal{E}_\mathcal{G})$, and each tissue-specific network by $\mathcal{T} = (\mathcal{V}_\mathcal{T}, \mathcal{E}_\mathcal{T})$, respectively. Using this notation, we have $n_\mathcal{T} = |\mathcal{V}_\mathcal{T}|$, $\mathcal{V}_\mathcal{T} \subset \mathcal{V}_\mathcal{G}$, and $\mathcal{E}_\mathcal{T} \subset \mathcal{E}_\mathcal{G}$ is the subset of all edges from $\mathcal{G}$ that connect vertices in $\mathcal{V}_\mathcal{T}$, i.e., $\mathcal{T}$ is the vertex-induced subgraph of $\mathcal{G}$ under $\mathcal{V}_\mathcal{T}$. This is the formal description of the model used by Bossi *et al.*[57] to construct human tissue-specific networks. Using this construction model, we note that every tissue-specific network inherits a shared core of interactions among housekeeping genes that are universally expressed to maintain basic cellular functions. Let us denote this subset of genes by $\mathcal{V}_\mathcal{U} \subset \mathcal{V}_\mathcal{T}$, having $n_\mathcal{U} = |\mathcal{V}_\mathcal{U}|$ members, and the corresponding induced core sub-graph using $\mathcal{U} = (\mathcal{V}_\mathcal{U}, \mathcal{E}_\mathcal{U})$.

In this setting, we propose a new random model to explicitly mimic the topology of tissue-specific networks. Formally, given each human tissue-specific network, we seed an ensemble of *pseudo-random tissues* denoted by $\boldsymbol{\mathcal{R}}_\mathcal{T} = \mathcal{G}(\mathcal{V}_\boldsymbol{\mathcal{R}}, \mathcal{E}_\boldsymbol{\mathcal{R}})$, in which every instance shares two main characteristics from the original network: (i) the total number of vertices, (ii) the shared core of housekeeping interactions. To summarize, our random graph sampling scheme is as follows: first, we initialize the vertex set $\mathcal{V}_\boldsymbol{\mathcal{R}}$ using $\mathcal{V}_\mathcal{U}$, which includes $n_\mathcal{U}$ housekeeping genes. Next, to ensure that the newly generated random instance has the same number of vertices as the seed network, we

sample $n_{\mathcal{T}} - n_{\mathcal{U}}$ vertices without replacement from the remaining vertices, $\mathcal{V}_{\mathcal{G}} \setminus \mathcal{V}_{\mathcal{U}}$. Finally, we construct the random graph as the vertex induced sub-graph of the global human interactome imposed by $\mathcal{V}_{\boldsymbol{\mathcal{R}}}$.

It can be noted that our random model not only provides a pseudo-random network seeded on each tissue-specific network, but also provides a node-to-node similarity score between the newly generated graph and the yeast interactome. This is a critical component of our framework, which distinguishes it from other *random graph models*, such as Erdos-Renyi, network growth, or preferential attachment. The only other effort to combine topology with the node-to-node similarity score is proposed by Sahraeian *et al.*[87], which fits a gamma distribution over the the known pairs of ortholog/ non-orthologs proteins in three species (according to their KEGG pathways), and uses the fitted distribution to sample new sequence similarity scores. However, this model does not benefit from the structural knowledge of the tissue-specific networks. Moreover, its sequence similarity generation model loosely fits the observed data and does not provide a fine-tuned model to assess the significance of tissue-specific alignments. Our model, one the other hand, is grounded in the same construction model as the original tissue-specific networks, and provides enough selectivity to distinguish similarity/ dissimilarity of aligned networks with yeast and to assign an empirical p-value to each alignment.

Significance of network alignments

For each optimal alignment of a human tissue-specific network with yeast, given by its indicator variable $\boldsymbol{x}$, we quantify the overall sequence similarity of aligned proteins with the matching score of the alignment, $\hat{w} = \boldsymbol{w}^T\boldsymbol{x}$, and the total number of conserved edges by the alignment overlap, $\hat{o} = \frac{1}{2}\boldsymbol{x}^T\mathbf{S}\boldsymbol{x}$. These measures can be used to rank different network alignments. However, without a proper reference to compare with, it is almost impossible to interpret these values in a statistical sense. To address this issue, we sample an ensemble of $k_{\boldsymbol{\mathcal{R}}}$ random networks from the *tissue-specific random model (TRAM)*, independently align each instance to the yeast interactome, and empirically compute a *topological*, a *homological* (sequence-based), and a *mixed* alignment p-value for each alignment using Monte-Carlo simulation.

Let $\hat{w}_{\boldsymbol{\mathcal{R}}}$ and $\hat{o}_{\boldsymbol{\mathcal{R}}}$ be the random vectors representing the weight and overlap of aligning random tissues with yeast, respectively. First, we define individual p-values for the conservation of network topology and sequence homology. Let us denote by $k_P^{(\hat{w})}$ and $k_P^{(\hat{o})}$ the number of random samples that have weight and overlap greater than or equal to the original alignment, respectively. Then, we can define the following p-values:

$$p - val_{\mathbf{homolgy}} = \frac{k_P^{(\hat{w})}}{k_{\boldsymbol{\mathcal{R}}}} \tag{2}$$

$$p - val_{\mathbf{topology}} = \frac{k_P^{(\hat{o})}}{k_{\boldsymbol{\mathcal{R}}}} \tag{3}$$

Before we define the *mixed* p-value, we define an upper bound and a lower bound on the p-value that is independent of the mixing function. For cases where both $\hat{o} \leq \hat{o}_{\boldsymbol{\mathcal{R}}}(i)$ and $\hat{w} \leq \hat{w}_{\boldsymbol{\mathcal{R}}}(i)$, for $1 \leq i \leq k_{\boldsymbol{\mathcal{R}}}$, we can report that the random

alignment is at least as good as the original alignment. Conversely, if both $\hat{o}_{\boldsymbol{\mathcal{R}}}(i) < \hat{o}$ and $\hat{w}_{\boldsymbol{\mathcal{R}}}(i) < \hat{w}$, we can assert that the original alignment outperforms the random alignment. Let us denote the number of such cases by k_P and k_N, respectively. Using this formulation, we can compute the following bounds on the mixed p-value of the alignment:

$$\delta_{\boldsymbol{\mathcal{R}}} = \frac{k_P}{k_{\boldsymbol{\mathcal{R}}}} \leq \text{alignment p-value} \leq 1 - \frac{k_N}{k_{\boldsymbol{\mathcal{R}}}} = \Delta_{\boldsymbol{\mathcal{R}}} \tag{4}$$

We can use these bounds to estimate the similarity of each tissue-specific network to the yeast interactome. Tissues where the upper-bound of the alignment p-value is smaller than a given threshold α_u are considered similar to yeast, while tissues with the lower-bound larger than α_l are considered dissimilar. We note that the cases with contradictory results for weight and overlap, either if $\hat{o}_{\boldsymbol{\mathcal{R}}}(i) < \hat{o}$ and $\hat{w} < \hat{w}_{\boldsymbol{\mathcal{R}}}(i)$, or $\hat{o} < \hat{o}_{\boldsymbol{\mathcal{R}}}(i)$ and $\hat{w}_{\boldsymbol{\mathcal{R}}}(i) < \hat{w}$, are not straightforward to interpret. To quantify this ambiguity, we define the confidence of a p-value interval as $\frac{k_N + k_P}{k_{\boldsymbol{\mathcal{R}}}}$. Finally, we define a mixed p-value based on the mixing function of the network alignment. Let us define a new random variable $o\hat{w}_{\boldsymbol{\mathcal{R}}} = \alpha * \hat{o}_{\boldsymbol{\mathcal{R}}} + \beta * \hat{w}_{\boldsymbol{\mathcal{R}}}$. Finally, we define the mixed p-value as:

$$p - value = Prob(\alpha * \hat{o} + \beta * \hat{w} \leq o\hat{w}_{\boldsymbol{\mathcal{R}}}) \tag{5}$$

Differential expression of genes with respect to a group of tissues

Given a homogenous group of human tissues/cell types, we first identify all *expressed genes* in the group, i.e., all non-housekeeping genes that are expressed in *at least* one of the tissue members. Next, in order to identify the subset of expressed genes that are *selectively* expressed, we use a *hypergeometric* random model. A gene is identified as selectively expressed if it is expressed in significantly higher number of tissues in the given group than randomly selected tissue subsets of the same size. Let N and n denote the total number of tissues in this study and the subset of tissues in the given group, respectively. Moreover, let us represent by c_N the number of all tissues in which a given gene is expressed, whereas c_n similarly represents the number of tissues in the given group that the gene is expressed. Finally, let the random variable X be the number of tissues in which the gene is expressed, if we randomly select subsets of tissues of similar size. Using this formulation, we can define the *tissue-selectivity* p-*value* of each expressed gene in the given group as follows:

$$\begin{aligned} p\text{-value}(X = c_n) &= Prob(c_n \leq X) \\ &= HGT(c_n | N, n, c_N) \\ &= \sum_{x=c_n}^{min(c_N, n)} \frac{C(c_N, x) C(N - c_N, n - x)}{C(N, n)} \end{aligned} \tag{6}$$

In order to partition genes into *selective* and *ubiquitous* genesets, we derive the tissue-selectivity p-value distribution of all expressed non-housekeeping genes in the given group. We use the Gaussian kernel to smooth this distribution and then find the critical points of the smoothed density function to threshold for tissue-selective genes. The motivation behind our choice is that these points provide shifts in the underlying distribution, from tissue-selective to ubiquitous genes. Given the bi-modal characteristic of the distribution, it has three expected critical points. We use the first of these points as our cutoff point. This provides highest precision for declared tissue-selective genes, but lower recall than the other two choices.

Conservation of genesets based on the majority voting rule

Given a set of genes that are selectively expressed in a homogenous group of tissues/cell types, we are interested in tri-partitioning them into either *conserved*, *human-specific*, or *unclassified* genes. *Conserved genes* are the subset of tissue-selective genes that are consistently aligned in majority of aligned tissues in the given group. Conversely, *human-specific genes* are the subset of tissue-selective genes that are consistently unaligned in majority of tissues in the given group. Finally, *unclassified genes* are the subset of tissue-selective genes for which we do not have enough evidence to classify them as either conserved or human-specific.

The key data-structure we use to tri-partition genesets is the *alignment consistency table*. Let C be a group of homogenous tissues with $n = |C|$. Furthermore, let $g_C^{\mathbf{TS}}$ represent the set of tissue-selective genes with respect to C, such that $k_C^{\mathbf{TS}} = |g_C^{\mathbf{TS}}|$. The alignment consistency table is a table of size $k_C^{\mathbf{TS}} \times n$, represented by $\mathcal{T}_C^{\mathbf{TS}}$, in which $\mathcal{T}_C^{\mathbf{TS}}(i, j)$ is the aligned yeast partner of i^{th} tissue selective gene under the network alignment of j^{th} tissue in C, or $'-'$ (gap), if it is unaligned. We find the most common partner for each tissue-selective gene and use a *consensus rate*, represented by τ, to summarize each rows of the alignment consistency table. If a gene is consistently aligned to the same yeast partner in at least $\tau * n$ tissues in C, we declare it as conserved. Similarly, if it is unaligned in at least $\tau * n$ tissues in C, we classify it as human-specific. If neither one of these conditions hold, we report it as unclassified.

Availability of supporting data

The data sets supporting the results of this article are included within the article and its additional files.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

SM proposed the initial idea of research, conceived of the study, and prepared the manuscript. SM and BS designed and implemented most of methods and performed the experiments. SS helped with the experimental design, as well as analyzing and interpreting the biological implications of the results. AG provided guidance relative to the theoretical and practical aspects of the methods, and design of proper statistical model(s) to validate the results. All authors participated in designing the structure and organization of final manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work is supported by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-0939370, and by NSF grants DBI 0835677 and 0641037.

Author details

[1]Department of Computer Sciences, Purdue University, 47907 West Lafayette, USA. [2]Department of Bioengineering, University of California at San Diego, 9500 Gilman Drive, 92093 La Jolla, USA.

References

1. Botstein, D., Fink, G.R.: Yeast: an experimental organism for 21st Century biology. Genetics **189**(3), 695–704 (2011). doi:10.1534/genetics.111.130765
2. Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y., Philippsen, P., Tettelin, H., Oliver, S.G.: Life with 6000 genes. Science (New York, N.Y.) **274**(5287), 546–5637 (1996)
3. Giaever, G., Chu, A.M., Ni, L., Connelly, C., Riles, L., Véronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., André, B., Arkin, A.P., Astromoff, A., El-Bakkoury, M., Bangham, R., Benito, R., Brachat, S., Campanaro, S., Curtiss, M., Davis, K., Deutschbauer, A., Entian, K.-D., Flaherty, P., Foury, F., Garfinkel, D.J., Gerstein, M., Gotte, D., Güldener, U., Hegemann, J.H., Hempel, S., Herman, Z., Jaramillo, D.F., Kelly, D.E., Kelly, S.L., Kötter, P., LaBonte, D., Lamb, D.C., Lan, N., Liang, H., Liao, H., Liu, L., Luo, C., Lussier, M., Mao, R., Menard, P., Ooi, S.L., Revuelta, J.L., Roberts, C.J., Rose, M., Ross-Macdonald, P., Scherens, B., Schimmack, G., Shafer, B., Shoemaker, D.D., Sookhai-Mahadeo, S., Storms, R.K., Strathern, J.N., Valle, G., Voet, M., Volckaert, G., Wang, C.-y., Ward, T.R., Wilhelmy, J., Winzeler, E.A., Yang, Y., Yen, G., Youngman, E., Yu, K., Bussey, H., Boeke, J.D., Snyder, M., Philippsen, P., Davis, R.W., Johnston, M.: Functional profiling of the Saccharomyces cerevisiae genome. Nature **418**(6896), 387–91 (2002). doi:10.1038/nature00935
4. Winzeler, E.A., Shoemaker, D.D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J.D., Bussey, H., Chu, A.M., Connelly, C., Davis, K., Dietrich, F., Dow, S.W., El Bakkoury, M., Foury, F., Friend, S.H., Gentalen, E., Giaever, G., Hegemann, J.H., Jones, T., Laub, M., Liao, H., Liebundguth, N., Lockhart, D.J., Lucau-Danila, A., Lussier, M., M'Rabet, N., Menard, P., Mittmann, M., Pai, C., Rebischung, C., Revuelta, J.L., Riles, L., Roberts, C.J., Ross-MacDonald, P., Scherens, B., Snyder, M., Sookhai-Mahadeo, S., Storms, R.K., Véronneau, S., Voet, M., Volckaert, G., Ward, T.R., Wysocki, R., Yen, G.S., Yu, K., Zimmermann, K., Philippsen, P., Johnston, M., Davis, R.W.: Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. Science (New York, N.Y.) **285**(5429), 901–6 (1999)
5. Jones, G.M., Stalker, J., Humphray, S., West, A., Cox, T., Rogers, J., Dunham, I., Prelich, G.: A systematic library for comprehensive overexpression screens in Saccharomyces cerevisiae. Nature methods **5**(3), 239–41 (2008). doi:10.1038/nmeth.1181
6. Huh, W.-K., Falvo, J.V., Gerke, L.C., Carroll, A.S., Howson, R.W., Weissman, J.S., O'Shea, E.K.: Global analysis of protein localization in budding yeast. Nature **425**(6959), 686–91 (2003). doi:10.1038/nature02026
7. Ghaemmaghami, S., Huh, W.-K., Bower, K., Howson, R.W., Belle, A., Dephoure, N., O'Shea, E.K., Weissman, J.S.: Global analysis of protein expression in yeast. Nature **425**(6959), 737–41 (2003). doi:10.1038/nature02046
8. Lashkari, D.A., DeRisi, J.L., McCusker, J.H., Namath, A.F., Gentile, C., Hwang, S.Y., Brown, P.O., Davis, R.W.: Yeast microarrays for genome wide parallel genetic and gene expression analysis. Proceedings of the National Academy of Sciences of the United States of America **94**(24), 13057–62 (1997)
9. DeRisi, J.L., Iyer, V.R., Brown, P.O.: Exploring the metabolic and genetic control of gene expression on a genomic scale. Science (New York, N.Y.) **278**(5338), 680–6 (1997)
10. Cho, R.J., Campbell, M.J., Winzeler, E.A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T.G., Gabrielian, A.E., Landsman, D., Lockhart, D.J., Davis, R.W.: A genome-wide transcriptional analysis of the mitotic cell cycle. Molecular cell **2**(1), 65–73 (1998)
11. Zhu, H., Bilgin, M., Bangham, R., Hall, D., Casamayor, A., Bertone, P., Lan, N., Jansen, R., Bidlingmaier, S., Houfek, T., Mitchell, T., Miller, P., Dean, R.A., Gerstein, M., Snyder, M.: Global analysis of protein activities using proteome chips. Science (New York, N.Y.) **293**(5537), 2101–5 (2001). doi:10.1126/science.1062191
12. Villas-Bôas, S.G., Moxley, J.F., Akesson, M., Stephanopoulos, G., Nielsen, J.: High-throughput metabolic state analysis: the missing link in integrated functional genomics of yeasts. The Biochemical journal **388**(Pt 2), 669–77 (2005). doi:10.1042/BJ20041162
13. Jewett, M.C., Hofmann, G., Nielsen, J.: Fungal metabolite analysis in genomics and phenomics. Current opinion in biotechnology **17**(2), 191–7 (2006). doi:10.1016/j.copbio.2006.02.001
14. Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., Sakaki, Y.: A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proceedings of the National Academy of Sciences of the United States of America **98**(8), 4569–74 (2001). doi:10.1073/pnas.061034498
15. Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., Rothberg, J.M.: A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature **403**(6770), 623–7 (2000). doi:10.1038/35001009
16. Krogan, N.J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A.P., Punna, T., Peregrín-Alvarez, J.M., Shales, M., Zhang, X., Davey, M., Robinson, M.D., Paccanaro, A., Bray, J.E., Sheung, A., Beattie, B., Richards, D.P., Canadien, V., Lalev, A., Mena, F., Wong, P., Starostine, A., Canete, M.M., Vlasblom, J., Wu, S., Orsi, C., Collins, S.R., Chandran, S., Haw, R., Rilstone, J.J., Gandi, K., Thompson, N.J., Musso, G., St Onge, P., Ghanny, S., Lam, M.H.Y., Butland, G., Altaf-Ul, A.M., Kanaya, S., Shilatifard, A., O'Shea, E., Weissman, J.S., Ingles, C.J., Hughes, T.R., Parkinson, J., Gerstein, M., Wodak, S.J., Emili, A., Greenblatt, J.F.: Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature **440**(7084), 637–43 (2006). doi:10.1038/nature04670
17. Lieb, J.D., Liu, X., Botstein, D., Brown, P.O.: Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. Nature genetics **28**(4), 327–34 (2001). doi:10.1038/ng569
18. Iyer, V.R., Horak, C.E., Scafe, C.S., Botstein, D., Snyder, M., Brown, P.O.: Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. Nature **409**(6819), 533–8 (2001). doi:10.1038/35054095
19. Tong, A.H., Evangelista, M., Parsons, A.B., Xu, H., Bader, G.D., Pagé, N., Robinson, M., Raghibizadeh, S., Hogue, C.W., Bussey, H., Andrews, B., Tyers, M., Boone, C.: Systematic genetic analysis with ordered arrays of yeast deletion mutants. Science (New York, N.Y.) **294**(5550), 2364–8 (2001). doi:10.1126/science.1065810
20. Tong, A.H.Y., Lesage, G., Bader, G.D., Ding, H., Xu, H., Xin, X., Young, J., Berriz, G.F., Brost, R.L., Chang, M., Chen, Y., Cheng, X., Chua, G., Friesen, H., Goldberg, D.S., Haynes, J., Humphries, C., He, G., Hussein, S.,

Ke, L., Krogan, N., Li, Z., Levinson, J.N., Lu, H., Ménard, P., Munyana, C., Parsons, A.B., Ryan, O., Tonikian, R., Roberts, T., Sdicu, A.-M., Shapiro, J., Sheikh, B., Suter, B., Wong, S.L., Zhang, L.V., Zhu, H., Burd, C.G., Munro, S., Sander, C., Rine, J., Greenblatt, J., Peter, M., Bretscher, A., Bell, G., Roth, F.P., Brown, G.W., Andrews, B., Bussey, H., Boone, C.: Global mapping of the yeast genetic interaction network. Science (New York, N.Y.) **303**(5659), 808–13 (2004). doi:10.1126/science.1091317

21. Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E.D., Sevier, C.S., Ding, H., Koh, J.L.Y., Toufighi, K., Mostafavi, S., Prinz, J., St Onge, R.P., VanderSluis, B., Makhnevych, T., Vizeacoumar, F.J., Alizadeh, S., Bahr, S., Brost, R.L., Chen, Y., Cokol, M., Deshpande, R., Li, Z., Lin, Z.-Y., Liang, W., Marback, M., Paw, J., San Luis, B.-J., Shuteriqi, E., Tong, A.H.Y., van Dyk, N., Wallace, I.M., Whitney, J.A., Weirauch, M.T., Zhong, G., Zhu, H., Houry, W.A., Brudno, M., Ragibizadeh, S., Papp, B., Pál, C., Roth, F.P., Giaever, G., Nislow, C., Troyanskaya, O.G., Bussey, H., Bader, G.D., Gingras, A.-C., Morris, Q.D., Kim, P.M., Kaiser, C.A., Myers, C.L., Andrews, B.J., Boone, C.: The genetic landscape of a cell. Science (New York, N.Y.) **327**(5964), 425–31 (2010). doi:10.1126/science.1180823
22. Sharan, R., Suthram, S., Kelley, R.M., Kuhn, T., McCuine, S., Uetz, P., Sittler, T., Karp, R.M., Ideker, T.: Conserved patterns of protein interaction in multiple species. Proceedings of the National Academy of Sciences of the United States of America **102**(6), 1974–9 (2005). doi:10.1073/pnas.0409522102
23. Koyutürk, M., Kim, Y., Topkara, U., Subramaniam, S., Szpankowski, W., Grama, A.: Pairwise alignment of protein interaction networks. Journal of computational biology : a journal of computational molecular cell biology **13**(2), 182–99 (2006). doi:10.1089/cmb.2006.13.182
24. Singh, R., Xu, J., Berger, B.: Global alignment of multiple protein interaction networks with application to functional orthology detection. Proceedings of the National Academy of Sciences of the United States of America **105**(35), 12763–8 (2008). doi:10.1073/pnas.0806627105
25. Kuchaiev, O., Przulj, N.: Integrative network alignment reveals large regions of global network similarity in yeast and human. Bioinformatics (Oxford, England) **27**(10), 1390–6 (2011). doi:10.1093/bioinformatics/btr127
26. Hartwell, L.H.: Nobel Lecture. Yeast and cancer. Bioscience reports **22**(3-4), 373–94 (2002)
27. Petranovic, D., Tyo, K., Vemuri, G.N., Nielsen, J.: Prospects of yeast systems biology for human health: integrating lipid, protein and energy metabolism. FEMS yeast research **10**(8), 1046–59 (2010). doi:10.1111/j.1567-1364.2010.00689.x
28. Munoz, A.J., Wanichthanarak, K., Meza, E., Petranovic, D.: Systems biology of yeast cell death. FEMS yeast research **12**(2), 249–65 (2012). doi:10.1111/j.1567-1364.2011.00781.x
29. Carmona-Gutierrez, D., Ruckenstuhl, C., Bauer, M.A., Eisenberg, T., Büttner, S., Madeo, F.: Cell death in yeast: growing applications of a dying buddy. Cell death and differentiation **17**(5), 733–4 (2010). doi:10.1038/cdd.2010.10
30. Brodsky, J.L., Skach, W.R.: Protein folding and quality control in the endoplasmic reticulum: Recent lessons from yeast and mammalian cell systems. Current opinion in cell biology **23**(4), 464–75 (2011). doi:10.1016/j.ceb.2011.05.004
31. Bonifacino, J.S., Glick, B.S.: The mechanisms of vesicle budding and fusion. Cell **116**(2), 153–66 (2004)
32. Widmann, C., Gibson, S., Jarpe, M.B., Johnson, G.L.: Mitogen-activated protein kinase: conservation of a three-kinase module from yeast to human. Physiological reviews **79**(1), 143–80 (1999)
33. Chen, R.E., Thorner, J.: Function and regulation in MAPK signaling pathways: lessons learned from the yeast Saccharomyces cerevisiae. Biochimica et biophysica acta **1773**(8), 1311–40 (2007). doi:10.1016/j.bbamcr.2007.05.003
34. De Virgilio, C., Loewith, R.: The TOR signalling network from yeast to man. The international journal of biochemistry & cell biology **38**(9), 1476–81 (2006). doi:10.1016/j.biocel.2006.02.013
35. Barbieri, M., Bonafè, M., Franceschi, C., Paolisso, G.: Insulin/IGF-I-signaling pathway: an evolutionarily conserved mechanism of longevity from yeast to humans. American journal of physiology. Endocrinology and metabolism **285**(5), 1064–71 (2003). doi:10.1152/ajpendo.00296.2003
36. Smith, M.G., Snyder, M.: Yeast as a model for human disease. Current protocols in human genetics / editorial board, Jonathan L. Haines ... [et al.] **Chapter 15**, 15–6 (2006). doi:10.1002/0471142905.hg1506s48
37. Perocchi, F., Mancera, E., Steinmetz, L.M.: Systematic screens for human disease genes, from yeast to human and back. Molecular bioSystems **4**(1), 18–29 (2008). doi:10.1039/b709494a
38. Petranovic, D., Nielsen, J.: Can yeast systems biology contribute to the understanding of human disease? Trends in biotechnology **26**(11), 584–90 (2008). doi:10.1016/j.tibtech.2008.07.008
39. Guaragnella, N., Palermo, V., Galli, A., Moro, L., Mazzoni, C., Giannattasio, S.: The expanding role of yeast in cancer research and diagnosis: insights into the function of the oncosuppressors p53 and BRCA1/2. FEMS yeast research (2013). doi:10.1111/1567-1364.12094
40. Pereira, C., Coutinho, I., Soares, J., Bessa, C., Leão, M., Saraiva, L.: New insights into cancer-related proteins provided by the yeast model. The FEBS journal **279**(5), 697–712 (2012). doi:10.1111/j.1742-4658.2012.08477.x
41. Pereira, C., Coutinho, I., Soares, J., Bessa, C., Leão, M., Saraiva, L.: New insights into cancer-related proteins provided by the yeast model. The FEBS journal **279**(5), 697–712 (2012). doi:10.1111/j.1742-4658.2012.08477.x
42. Khurana, V., Lindquist, S.: Modelling neurodegeneration in Saccharomyces cerevisiae: why cook with baker's yeast? Nature reviews. Neuroscience **11**(6), 436–49 (2010). doi:10.1038/nrn2809
43. Pereira, C., Bessa, C., Soares, J., Leão, M., Saraiva, L.: Contribution of yeast models to neurodegeneration research. Journal of biomedicine & biotechnology **2012**, 941232 (2012). doi:10.1155/2012/941232
44. Tenreiro, S., Munder, M.C., Alberti, S., Outeiro, T.F.: Harnessing the power of yeast to unravel the molecular basis of neurodegeneration. Journal of neurochemistry **127**(4), 438–52 (2013). doi:10.1111/jnc.12271
45. Longo, V.D., Shadel, G.S., Kaeberlein, M., Kennedy, B.: Replicative and chronological aging in Saccharomyces cerevisiae. Cell metabolism **16**(1), 18–31 (2012). doi:10.1016/j.cmet.2012.06.002
46. Forslund, K., Schreiber, F., Thanintorn, N., Sonnhammer, E.L.L.: OrthoDisease: tracking disease gene orthologs across 100 species. Briefings in bioinformatics **12**(5), 463–73 (2011). doi:10.1093/bib/bbr024

47. McGary, K.L., Park, T.J., Woods, J.O., Cha, H.J., Wallingford, J.B., Marcotte, E.M.: Systematic discovery of nonobvious human disease models through orthologous phenotypes. Proceedings of the National Academy of Sciences of the United States of America **107**(14), 6544–9 (2010). doi:10.1073/pnas.0910200107
48. Takahashi, Y.-h., Westfield, G.H., Oleskie, A.N., Trievel, R.C., Shilatifard, A., Skiniotis, G.: Structural analysis of the core COMPASS family of histone H3K4 methylases from yeast to human. Proceedings of the National Academy of Sciences of the United States of America **108**(51), 20526–31 (2011). doi:10.1073/pnas.1109360108
49. Clapp, C., Portt, L., Khoury, C., Sheibani, S., Eid, R., Greenwood, M., Vali, H., Mandato, C.A., Greenwood, M.T.: Untangling the Roles of Anti-Apoptosis in Regulating Programmed Cell Death using Humanized Yeast Cells. Frontiers in oncology **2**, 59 (2012). doi:10.3389/fonc.2012.00059
50. Qian, Y., Kachroo, A.H., Yellman, C.M., Marcotte, E.M., Johnson, K.A.: Yeast Cells Expressing the Human Mitochondrial DNA Polymerase Reveal Correlations between Polymerase Fidelity and Human Disease Progression. The Journal of biological chemistry **289**(9), 5970–85 (2014). doi:10.1074/jbc.M113.526418
51. Tardiff, D.F., Jui, N.T., Khurana, V., Tambe, M.A., Thompson, M.L., Chung, C.Y., Kamadurai, H.B., Kim, H.T., Lancaster, A.K., Caldwell, K.A., Caldwell, G.A., Rochet, J.-C., Buchwald, S.L., Lindquist, S.: Yeast reveal a "druggable" Rsp5/Nedd4 network that ameliorates α-synuclein toxicity in neurons. Science (New York, N.Y.) **342**(6161), 979–83 (2013). doi:10.1126/science.1245321
52. Chung, C.Y., Khurana, V., Auluck, P.K., Tardiff, D.F., Mazzulli, J.R., Soldner, F., Baru, V., Lou, Y., Freyzon, Y., Cho, S., Mungenast, A.E., Muffat, J., Mitalipova, M., Pluth, M.D., Jui, N.T., Schüle, B., Lippard, S.J., Tsai, L.-H., Krainc, D., Buchwald, S.L., Jaenisch, R., Lindquist, S.: Identification and rescue of α-synuclein toxicity in Parkinson patient-derived neurons. Science (New York, N.Y.) **342**(6161), 983–7 (2013). doi:10.1126/science.1245296
53. Dunham, M.J., Fowler, D.M.: Contemporary, yeast-based approaches to understanding human genetic variation. Current opinion in genetics & development **23**(6), 658–64 (2013). doi:10.1016/j.gde.2013.10.001
54. Bier, E., Mcginnis, W.: Model Organisms in the Study of Development and Disease. In: Inborn Errors of Development: The Molecular Basis of Clinical Disorders of Morphogenesis, (2008)
55. Goh, K.-I., Cusick, M.E., Valle, D., Childs, B., Vidal, M., Barabási, A.-L.: The human disease network. Proceedings of the National Academy of Sciences of the United States of America **104**(21), 8685–90 (2007). doi:10.1073/pnas.0701361104
56. Lage, K., Hansen, N.T., Karlberg, E.O., Eklund, A.C., Roque, F.S., Donahoe, P.K., Szallasi, Z., Jensen, T.S.t., Brunak, S.r.: A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. Proceedings of the National Academy of Sciences of the United States of America **105**(52), 20870–5 (2008). doi:10.1073/pnas.0810772105
57. Bossi, A., Lehner, B.: Tissue specificity and the human protein interaction network. Molecular systems biology **5**, 260 (2009). doi:10.1038/msb.2009.17
58. Su, A.I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K.A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., Cooke, M.P., Walker, J.R., Hogenesch, J.B.: 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**(16), 6062–7 (2004). doi:10.1073/pnas.0400782101
59. Bayati, M., Gleich, D.F., Saberi, A., Wang, Y.: Message-Passing Algorithms for Sparse Network Alignment. ACM Trans. Knowl. Discov. Data **7**(1), 3–1331 (2013). doi:10.1145/2435209.2435212
60. Dezso, Z., Nikolsky, Y., Sviridov, E., Shi, W., Serebriyskaya, T., Dosymbekov, D., Bugrim, A., Rakhmatulin, E., Brennan, R.J., Guryanov, A., Li, K., Blake, J., Samaha, R.R., Nikolskaya, T.: A comprehensive functional analysis of tissue specificity of human gene expression. BMC biology **6**, 49 (2008). doi:10.1186/1741-7007-6-49
61. Chang, C.-W., Cheng, W.-C., Chen, C.-R., Shu, W.-Y., Tsai, M.-L., Huang, C.-L., Hsu, I.C.: Identification of human housekeeping genes and tissue-selective genes by microarray meta-analysis. PloS one **6**(7), 22859 (2011). doi:10.1371/journal.pone.0022859
62. Souiai, O., Becker, E., Prieto, C., Benkahla, A., De las Rivas, J., Brun, C.: Functional integrative levels in the human interactome recapitulate organ organization. PloS one **6**(7), 22051 (2011). doi:10.1371/journal.pone.0022051
63. Zhang, L., Li, W.-H.: Mammalian housekeeping genes evolve more slowly than tissue-specific genes. Molecular biology and evolution **21**(2), 236–9 (2004). doi:10.1093/molbev/msh010
64. Yu, H., Luscombe, N.M., Lu, H.X., Zhu, X., Xia, Y., Han, J.-D.J., Bertin, N., Chung, S., Vidal, M., Gerstein, M.: Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. Genome research **14**(6), 1107–18 (2004). doi:10.1101/gr.1774904
65. Bader, G.D., Hogue, C.W.V.: An automated method for finding molecular complexes in large protein interaction networks. BMC bioinformatics **4**, 2 (2003)
66. Reimand, J., Arak, T., Vilo, J.: g:Profiler–a web server for functional interpretation of gene lists (2011 update). Nucleic acids research **39**(Web Server issue), 307–15 (2011). doi:10.1093/nar/gkr378
67. Yagi, T., Seo, B.B., Nakamaru-Ogiso, E., Marella, M., Barber-Singh, J., Yamashita, T., Kao, M.-C., Matsuno-Yagi, A.: Can a single subunit yeast NADH dehydrogenase (Ndi1) remedy diseases caused by respiratory complex I defects? Rejuvenation research **9**(2), 191–7 (2006). doi:10.1089/rej.2006.9.191
68. Marella, M., Seo, B.B., Yagi, T., Matsuno-Yagi, A.: Parkinson's disease and mitochondrial complex I: a perspective on the Ndi1 therapy. Journal of bioenergetics and biomembranes **41**(6), 493–7 (2009). doi:10.1007/s10863-009-9249-z
69. Merico, D., Isserlin, R., Stueker, O., Emili, A., Bader, G.D.: Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. PloS one **5**(11), 13984 (2010). doi:10.1371/journal.pone.0013984
70. Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.-L., Ideker, T.: Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics (Oxford, England) **27**(3), 431–2 (2011). doi:10.1093/bioinformatics/btq675
71. Lin, J., Epel, E., Cheon, J., Kroenke, C., Sinclair, E., Bigos, M., Wolkowitz, O., Mellon, S., Blackburn, E.:

Analyses and comparisons of telomerase activity and telomere length in human T and B cells: insights for epidemiology of telomere maintenance. Journal of immunological methods **352**(1-2), 71–80 (2010). doi:10.1016/j.jim.2009.09.012
72. Huang, W., Zhu, P.J., Zhang, S., Zhou, H., Stoica, L., Galiano, M., Krnjević, K., Roman, G., Costa-Mattioli, M.: mTORC2 controls actin polymerization required for consolidation of long-term memory. Nature neuroscience **16**(4), 441–8 (2013). doi:10.1038/nn.3351
73. Orth, M., Bellosta, S.: Cholesterol: its regulation and role in central nervous system disorders. Cholesterol **2012**, 292598 (2012). doi:10.1155/2012/292598
74. Tucker, R.P.: The roles of microtubule-associated proteins in brain morphogenesis: a review. Brain research. Brain research reviews **15**(2), 101–20 (1990)
75. Kandratavicius, L., Monteiro, M.R., Hallak, J.E., Carlotti, C.G., Assirati, J.A., Leite, J.P.: Microtubule-associated proteins in mesial temporal lobe epilepsy with and without psychiatric comorbidities and their relation with granular cell layer dispersion. BioMed research international **2013**, 960126 (2013). doi:10.1155/2013/960126
76. Zhou, Y., Gunput, R.-A.F., Pasterkamp, R.J.: Semaphorin signaling: progress made and promises ahead. Trends in biochemical sciences **33**(4), 161–70 (2008). doi:10.1016/j.tibs.2008.01.006
77. Nakanishi, S., Nakajima, Y., Masu, M., Ueda, Y., Nakahara, K., Watanabe, D., Yamaguchi, S., Kawabata, S., Okada, M.: Glutamate receptors: brain function and signal transduction. Brain research. Brain research reviews **26**(2-3), 230–5 (1998)
78. Willard, S.S., Koochekpour, S.: Glutamate, glutamate receptors, and downstream signaling pathways. International journal of biological sciences **9**(9), 948–59 (2013). doi:10.7150/ijbs.6426
79. Becker, K.G., Barnes, K.C., Bright, T.J., Wang, S.A.: The genetic association database. Nature genetics **36**(5), 431–2 (2004). doi:10.1038/ng0504-431
80. Huang, D.W., Sherman, B.T., Lempicki, R.A.: Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols **4**(1), 44–57 (2009). doi:10.1038/nprot.2008.211
81. Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., Tyers, M.: BioGRID: a general repository for interaction datasets. Nucleic acids research **34**(Database issue), 535–9 (2006). doi:10.1093/nar/gkj109
82. Stark, C., Breitkreutz, B.-J., Chatr-Aryamontri, A., Boucher, L., Oughtred, R., Livstone, M.S., Nixon, J., Van Auken, K., Wang, X., Shi, X., Reguly, T., Rust, J.M., Winter, A., Dolinski, K., Tyers, M.: The BioGRID Interaction Database: 2011 update. Nucleic acids research **39**(Database issue), 698–704 (2011). doi:10.1093/nar/gkq1116
83. Cherry, J.M., Hong, E.L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E.T., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S.R., Fisk, D.G., Hirschman, J.E., Hitz, B.C., Karra, K., Krieger, C.J., Miyasato, S.R., Nash, R.S., Park, J., Skrzypek, M.S., Simison, M., Weng, S., Wong, E.D.: Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic acids research **40**(Database issue), 700–5 (2012). doi:10.1093/nar/gkr1029
84. Wootton, J.C., Federhen, S.: Statistics of local complexity in amino acid sequences and sequence databases. Computers & Chemistry **17**(2), 149–163 (1993). doi:10.1016/0097-8485(93)85006-x
85. Pearson, W.R., Lipman, D.J.: Improved tools for biological sequence analysis. Proc. Natl. Acad. Sci. **85**, 2444–2448 (1988)
86. Smith, T.F., Waterman, M.S.: Identification of common molecular subsequences. Journal of molecular biology **147**(1), 195–7 (1981)
87. Sahraeian, S.M.E., Yoon, B.-J.: A Network Synthesis Model for Generating Protein Interaction Network Families (2012). doi:10.1371/journal.pone.0041474

Figures

Figure 1 Workflow summary. Main components of the analysis framework proposed in this paper. Each intermediate processing step is further discussed in details in separate subsections.

Figure 2 Functional classification of human genes. A high-level summary of gene classification performed in this study.

Figure 3 Alignment graph of core human genes. Conserved edges in the alignment graph of core housekeeping genes, which correspond to the "interologs," i.e. orthologous pairs of interacting proteins between yeast and human. Five main protein clusters, identified as dense regions of interaction in the alignment graph, are marked accordingly and annotated with their dominant functional annotation as follows: **A** Ribosome, **B** Processing of capped intron-containing pre-mRNA, **C** Proteasome, **D** vATPase, **E** Cap-dependent translation initiation.

Figure 4 Projection of alignment p-values on the network of tissue-tissue similarities. Each node represents a human tissue and edges represent the overall transcriptional similarity among them. Color intensity of nodes represents the similarity/dissimilarity of each tissue to yeast interactome, with colors green and red corresponding to similar and dissimilar tissues, respectively. Group of similar tissues with coherent *p*-values are marked and annotated in the network, accordingly.

Figure 5 Membership distribution of non-housekeeping genes in human tissues. Number of tissues in which non-housekeeping genes are expressed in is smoothed using normal kernel density to estimated the pdf function. The observed bi-modal distribution suggests that most non-housekeeping genes are either expressed in a very few selected tissues or in the majority of human tissues.

Figure 6 Distribution of tissue-selectivity p-values in different tissue groups. (**A**) Brain tissues, (**B**) Blood cells, (**C**) Ganglion tissues, (**D**) Testis tissues. Each plot resembles the same bi-modal distribution as the gene-tissue membership density, with blood cells and brain tissues presenting the most clear separation of tissue-selective genes. The critical points of each distribution function, where the derivative of pdf function is approximately zero, is marked on each plot. These points provide optimal cutoff points for the tissue-selectivity *p*-values as they mark the points of shift in the underlying distribution function.

Figure 7 Summary of gene classifications in this study. Housekeeping and tissue-selective genes, in four main groups of human tissues, are classified into three main classes based on their conservation in yeast.

Figure 8 Enrichment map of unique blood-selective functions.

Figure 9 Enrichment map of unique brain-selective functions.

Tables

Name	pval lower bound	overall pval	pval upper bound	confidence
Myeloid Cells	< 1.00e-04	< 1.00e-04	< 1.00e-04	1
Monocytes	< 1.00e-04	< 1.00e-04	< 1.00e-04	1
Dentritic Cells	< 1.00e-04	< 1.00e-04	< 1.00e-04	1
NK Cells	< 1.00e-04	< 1.00e-04	< 1.00e-04	1
T-Helper Cells	< 1.00e-04	< 1.00e-04	< 1.00e-04	1
Cytotoxic T-Cells	< 1.00e-04	< 1.00e-04	< 1.00e-04	1
B-Cells	< 1.00e-04	< 1.00e-04	< 1.00e-04	1
Endothelial	< 1.00e-04	< 1.00e-04	< 1.00e-04	1
Hematopoietic Stem Cells	< 1.00e-04	< 1.00e-04	< 1.00e-04	1
MOLT-4	< 1.00e-04	< 1.00e-04	< 1.00e-04	1
B Lymphoblasts	< 1.00e-04	< 1.00e-04	< 1.00e-04	1
HL-60	< 1.00e-04	< 1.00e-04	< 1.00e-04	1
K-562	< 1.00e-04	< 1.00e-04	< 1.00e-04	1
Early Erythroid	< 1.00e-04	< 1.00e-04	< 1.00e-04	1
Bronchial Epithelial Cells	< 1.00e-04	< 1.00e-04	0.0002	0.9998
Colorectal Adenocarcinoma	< 1.00e-04	< 1.00e-04	0.0004	0.9996
Daudi	< 1.00e-04	< 1.00e-04	0.0009	0.9991
Testis Seminiferous Tubule	< 1.00e-04	< 1.00e-04	0.0012	0.9988
Smooth Muscle	< 1.00e-04	< 1.00e-04	0.0016	0.9984
Blood (Whole)	< 1.00e-04	< 1.00e-04	0.0053	0.9947
Thymus	< 1.00e-04	0.0001	0.0062	0.9938
Testis Interstitial	< 1.00e-04	0.0004	0.0086	0.9914

Table 1 Tissues with the most significant similarity to the yeast interactome

Name	pval lower bound	overall pval	pval upper bound	confidence
Trigeminal Ganglion	0.9947	0.9994	1	0.9947
Superior Cervical Ganglion	0.9847	0.9991	1	0.9847
Ciliary Ganglion	0.9407	0.9813	0.9964	0.9443
Atrioventricular Node	0.8746	0.9792	0.9921	0.8825
Skin	0.8355	0.9297	0.9809	0.8546
Heart	0.7934	0.9585	0.9815	0.8119
Appendix	0.7596	0.9371	0.973	0.7866
Dorsal Root Ganglion	0.7065	0.933	0.9717	0.7348
Skeletal Muscle	0.3994	0.5902	0.7866	0.6128
Uterus Corpus	0.233	0.7736	0.8769	0.3561
Lung	0.0771	0.3853	0.5544	0.5227
Pons	0.0674	0.5201	0.6983	0.3691
Salivary Gland	0.0639	0.3449	0.5173	0.5466
Liver	0.0600	0.6857	0.8519	0.2081
Ovary	0.0388	0.2735	0.4481	0.5907
Trachea	0.0259	0.2376	0.4146	0.6113
Globus Pallidus	0.0206	0.2471	0.4336	0.587
Cerebellum	0.0127	0.1950	0.3783	0.6344

Table 2 Tissues with the least significant similarity to the yeast interactome

Cluster name	# expressed genes	# TS genes	# CG (%)	# HS (%)	# unclassified (%)
Brain Tissues	5936	891	273 (30.64 %)	401 (45.01 %)	217 (24.35 %)
Blood Cells	6092	1093	460 (42.09 %)	385 (35.22 %)	248 (22.69 %)
Testis Tissues	5358	328	119 (36.28 %)	126 (38.41 %)	83 (25.30 %)
Ganglion Tissues	5278	274	76 (27.74 %)	136 (49.64 %)	62 (22.63 %)

Table 3 Summary of tissue-selective gene partitioning CG: Conserved gene, HS: Human-specific gene

Table 4 Enriched disease classes of tissue-selective genes

	Conserved genes		Human-specific genes	
	Disease class	*p*-value	Disease class	*p*-value
Blood cells	Cancer	$9.29 * 10^{-4}$	Immune	$1.19 * 10^{-5}$
Brain tissues	Psych	$3.59 * 10^{-4}$	Psych	$5.70 * 10^{-8}$
	Chemdependency	$2.60 * 10^{-3}$	Neurological	$2.97 * 10^{-2}$
	Pharmacogenomic	$9.74 * 10^{-2}$		

Table 5 Comparative analysis of brain-specific pathologies Top 10 Enriched disorders were identified based on the GAD annotations for conserved and human-specific genes in the brain.

Disorder	Conserved genes	Human-specific genes
schizophrenia	0.008573	8.4905E-06
autism	0.048288	0.00077448
dementia	0.0014356	-
schizophrenia; schizoaffective disorder; bipolar disorder	-	0.0021433
myocardial infarct; cholesterol, HDL; triglycerides; atherosclerosis, coronary; macular degeneration; colorectal cancer	0.0051617	-
epilepsy	0.071562	0.0064716
seizures	-	0.020381
bipolar disorder	0.048288	0.022016
attention deficit disorder conduct disorder oppositional defiant disorder	0.032444	0.023865

Additional Files

Additional file 1 — network alignments
Compressed (*.zip) file containing individual tissue-specific alignments.

Additional file 2 — HK genes
List of housekeeping genes and their classifications into conserved, human-specific, and unclassified subsets

Additional file 3 — Core gene alignment
Alignment graph of core housekeeping genes

Additional file 4 — HK Enrichment
Functional enrichment analysis of different subsets of HK genes

Additional file 5 — Alignment statistics
Alignment statistics for each tissue alignment

Additional file 6 — TS genes
Tissue-selective genesets and their respective classifications for brain tissues, blood cells, testis tissues, and ganglion tissues

Additional file 7 — TS Enrichment
Functional analysis of different subsets of tissue-selective genes

Additional file 8 — PPI Nets
Protein-protein interaction networks used as input in this study.

Additional file 9 — Sequence similarities

Sequence similarity between yeast and human proteins.