Computational and Structural Biotechnology Journal 20 (2022) 3764–3778

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Biological interacting units identified in human protein networks reveal tissue-functional diversification and its impact on disease

Marina L. García-Vaquero [a,b,*], Margarida Gama-Carvalho [a,1], Francisco R. Pinto [a,1], Javier De Las Rivas [b,1]

[a] University of Lisboa, Faculty of Sciences, BioISI – Biosystems & Integrative Sciences Institute, Campo Grande, C8 bdg, Lisboa 1749-016, Portugal
[b] Cancer Research Center (CiC-IBMCC, CSIC/USAL and IBSAL), Consejo Superior de Investigaciones Científicas (CSIC), University of Salamanca (USAL) and Instituto de Investigación Biomédica de Salamanca (IBSAL), Salamanca 37007, Spain

ARTICLE INFO

Article history:
Received 10 April 2022
Received in revised form 4 July 2022
Accepted 4 July 2022
Available online 15 July 2022

Keywords:
Biological function
Disease gene
Housekeeping gene
PPI network
Protein module
Tissue-specific gene

ABSTRACT

Protein-protein interactions (PPI) play an essential role in the biological processes that occur in the cell. Therefore, the dissection of PPI networks becomes decisive to model functional coordination and predict pathological de-regulation. Cellular networks are dynamic and proteins display varying roles depending on the tissue-interactomic context. Thus, the use of centrality measures in individual proteins fall short to dissect the functional properties of the cell. For this reason, there is a need for more comprehensive, relational, and context-specific ways to analyze the multiple actions of proteins in different cells and identify specific functional assemblies within global biomolecular networks. Under this framework, we define *Biological Interacting units* (BioInt-U) as groups of proteins that interact physically and are enriched in a common Gene Ontology. A search strategy was applied on 33 tissue-specific (TS) PPI networks to generate *BioInt* libraries associated with each particular human tissue. The cross-tissue comparison showed that housekeeping assemblies incorporate different proteins and exhibit distinct network properties depending on the tissue. Furthermore, disease genes (DGs) of tissue-associated pathologies preferentially accumulate in units in the expected tissues, which in turn were more central in the TS networks. Overall, the study reveals a tissue-specific functional diversification based on the identification of specific protein units and suggests vulnerabilities specific of each tissue network, which can be applied to refine protein-disease association methods.

1. Introduction

Cell physiology, defined as the ability to exert biological functions, emerges from the dynamic interactions in protein networks. Likewise, pathological manifestations arise from genetic alterations that result in protein interaction failure and network malfunction [4,40,44]. While great progress has been made towards the characterization of protein interactions (PPIs) and disease genes (DGs) [25,28], the relation between protein network connectivity and phenotypic manifestation is still poorly understood. The majority of diseases with restricted histological hallmarks are known to be triggered by DGs with wide tissue expression [19]. In that sense it still an open debate how mutations in housekeeping (ubiquitously expressed) genes can distinctively affect to the physiology only on certain tissues.

One fundamental reason for this knowledge gap is that biological networks are complex. Protein networks include large numbers of participating elements and these hold a large range of interchangeable partners. For instance, the complete human protein network available in the APID repository [2] included in April 2021 more than 17,000 proteins, with each one being able to interact with more than 30 partners on average (https://bioinfow.dep.usal.es/apid/). In fact, the combinatorial range of PPIs is an eminent force for tissue functional diversification [10,16]. The same protein may establish different interactions and exert varied functional roles depending on the context [12]. As a consequence, the pro-

Abbreviations: PPI, Protein-protein interaction; GO-BP, Gene Ontology biological process; TS, Tissue-specific; DG, Disease gene; DEg, Differentially expressed gene; HK, Housekeeping; BiU, BioInt unit; CO, CORUM complex; SS, Simpson's similarity; UB, Ubiquitous; TE, Tissue enriched.

* Corresponding author at: University of Lisboa, Faculty of Sciences, BioISI – Biosystems & Integrative Sciences Institute, Campo Grande, C8 bdg, Lisboa 1749-016, Portugal.

E-mail addresses: mlgarciavaquero@fc.ul.pt (M.L. García-Vaquero), mhcarvalho@fc.ul.pt (M. Gama-Carvalho), frpinto@fc.ul.pt (F.R. Pinto), jrivas@usal.es (J. De Las Rivas).

[1] Equal contribution to this article as senior authors.

https://doi.org/10.1016/j.csbj.2022.07.006

teins will be localized at different positions in the network depending on their active functional partners in the considered tissues. On this basis, one could argue the same protein might acquire varying topological properties across TS-networks that distinctively resonate in TS-physiology. Indeed, DGs do not locate at random positions in the PPI networks but tend to display more TS-PPI in the disease tissue than in the unaffected tissues [6]. This observation suggests the idea that different TS-network may have distinct vulnerable spots, and strongly supports that the characterization of topological properties underlying tissue functional diversity might be critical to understand the emergence of TS patho-phenotypes.

Protein-protein interactions are a strong indicative of functional collaboration. Network connectivity measures, such as clustering coefficient, degree and betweenness centrality, are well-established predictors of protein essentiality and so of potential vulnerabilities in cell physiology [4]. Based on these notions, a variety of PPI network-based strategies have been proposed to identify densely connected modules, recently reviewed by [24]. While these methods are valuable to predict functional collaboration and DG candidates, they are not suitable to characterize their topological context.

In this study we define and characterize *Biological Interacting units* (referred to as: BioInt-U), identified as biological modules found in PPI networks using tissue-specific mapping and topological interactomic analysis. In this way, BioInt units are found using a network-based framework to define topologically unbiased functional PPI consortia in multiple tissue-specific (TS) interactomes. These units represent an intermediate level of PPI functional coordination in TS networks, which allow the characterization of topological properties of normal and disease-targeted cell processes. A search for these BioInt units was performed within an extensive catalog of human tissues yielding 33 TS-BioInt libraries. Disease impact was assessed by mapping known disease genes (DGs) in BioInt libraries. The cross-tissue and cross-disease mapping revealed distinctive topological properties on the BioInt units, suggesting new explanatory insights into the occurrence of pathologies affecting specific tissues.

The benefits of using BioInt-U are illustrated, as an example study, by its integration with publicly available gene expression profiles (RNA-Seq) derived from patients affected by two diseases: psoriasis and pulmonary fibrosis. Our analysis revealed that proteins corresponding to differentially expressed transcripts/genes (DEg) collaborate in the same BioInt units in expected disease tissues. Furthermore, these BioInt units were involved in biological processes previously considered critical in the development of these diseases (fibrosis and psoriasis), providing new potential research targets or candidate proteins to be modulated in these diseases.

2. Results

2.1. Framework for dissecting functionally meaningful interactions: BioInt-U

The BioInt-U method was designed to identify groups of interacting proteins collaborating in the same biological processes (Fig. 1A), i.e., biologically interacting modules hereafter referred to as BioInt units. We first **(i)** reconstructed 33 tissue-specific (TS) networks by mapping TS transcripts/genes identified from TS RNA-Seq profiles in a tissue-naive PPI network. Next **(ii)** the TS networks were functionally characterized by evaluating the Gene Ontology Biological Process (GO-BP) annotated in the network; and **(iii)** by applying functional enrichment analysis of GO-BP terms. The enriched GO-BP terms were then used to dissect BioInt units. The BioInt units did only retain the proteins enriched in the GO-BP that, at the same time were physically interacting in the TS network **(iv)** (Fig. 1A). Assuming that some GO terms are very general and define too large and fuzzy functional groups, which are quite uninformative, we only considered BioInt units including <200 proteins. The use of BioInt-U in the TS networks returned 33 independent TS functional libraries, each including between 200 and 350 BioInt units (256 on average) (Fig. 1B), each including a mean of 103 proteins (Fig. 1C).

2.2. BioInt units more closely recover real tissue-specific functional modules

The recently published work by Skinnider and colleagues soundly demonstrates the benefits of mapping the *in vivo* interactome of mouse tissues using PCP-SILAM experiments [36]. The work expands ~2.5 times the former mouse interactome in an unbiased manner, identifies interactions for more than 360 "interactome orphan" proteins and characterizes tissue-specific interaction rewiring. As the authors stated, most widely used methods for PPI mapping are based on *in silico* experiments that lack cell or tissue-context. To compensate the current knowledge gap, most researchers opt to reconstruct context-specific interactome mapping gene expression data onto tissue-naive PPI data. On this basis, the authors evaluated the protein overlap between experimental PPI networks (PCP-SILAM) and predicted tissue-specific networks from protein expression data. Their analysis revealed a modest overlap between experimental and predicted networks, raising questions about whether or not the predictions drawn using BioInt-U method are accurate. BioInt-U method relies on predicted networks and identifies biologically interacting units defined as groups of interacting proteins sharing the same enriched GO-BP term. Interestingly, authors found that the PCP-SILAM approach tended to connect proteins with similar annotations. On this basis, we hypothesized that the groups of proteins in a PCP-SILAM network module are also likely to be together in one BioInt unit.

In order to answer this question, we evaluated the concordance between the experimental mouse tissue-specific networks with the predicted networks and the BioInt libraries (see methods). We used Wallace coefficient [29] to evaluate the clustering agreement between clusters in PCP-SILAM against clusters in APID or BioInt units. SILAM clusters revealed higher overlap with BioInt units than with clusters from APID (Supplementary Fig. X). The Wallace coefficient was highly variable depending on the clustering algorithm but its relative change from using APID to BioInt units was positive in almost all cases (21 out of 24 comparisons). Based on these results, we argue that BioInt-U outperforms the use of predicted networks based solely on gene expression data and therefore BioInt units more closely recover real tissue-specific functional modules.

2.3. TS BioInt libraries recap functional landscape of TS transcriptomes

An ultimate goal in constructing BioInt libraries is to dissect how the interactome is coordinated into TS functional consortia. On this basis, we corroborated that TS BioInt libraries recapitulate the functional landscape of TS transcriptomes resembling well-established biological properties.

We assessed TS transcriptome coverage at each step of the framework. A tissue-naive human PPI network retrieved from APID [2] was found to incorporate >90% of genes identified in each TS transcriptome [38] (see **(i)** in Fig. 1D). BioInt units are generated from Gene Ontology annotations, so the performance directly relies on the characterization state of the proteins. We found that >80% of proteins incorporated in TS networks are functionally annotated (see **(ii)** in Fig. 1D). Of note, the statistical functional

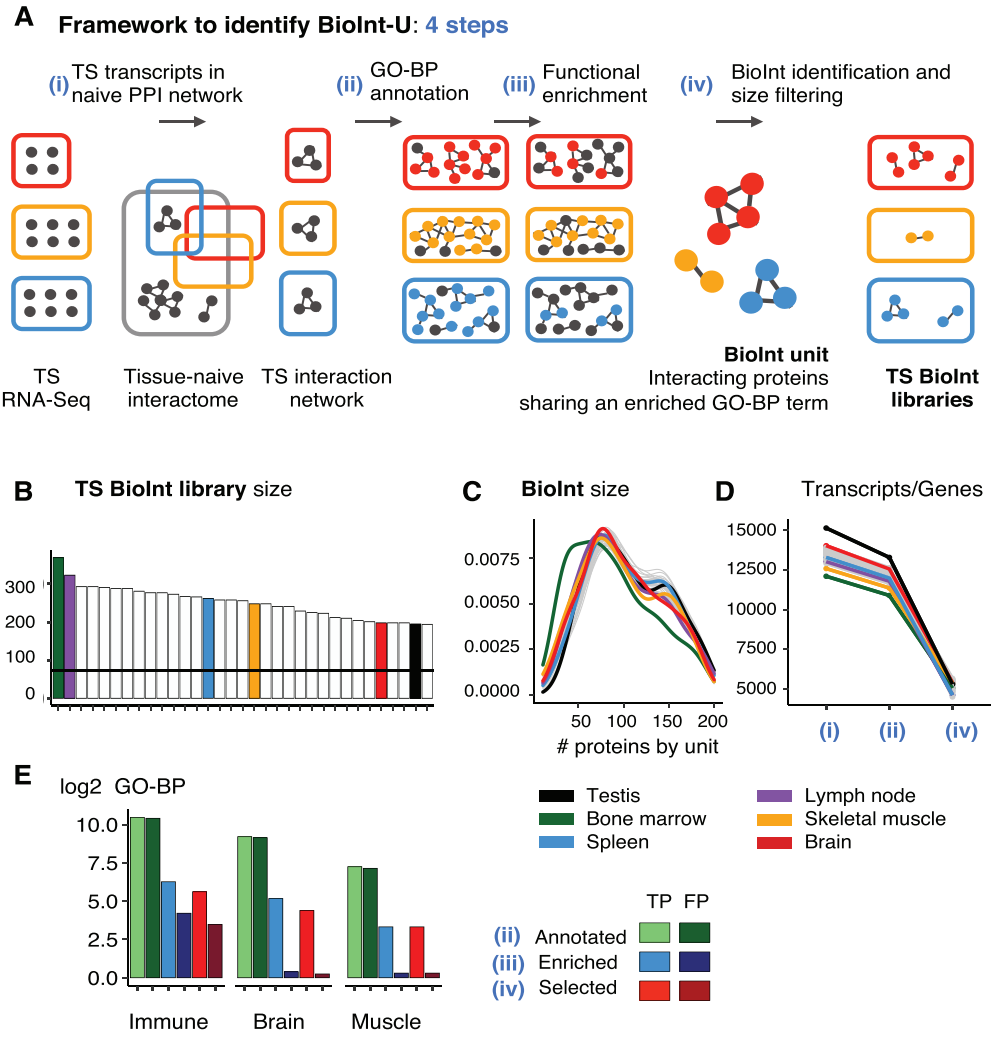

Fig. 1. BioInt-U framework performance overview. (A) Schematic illustration of BioInt-U framework**. (i)** First, we reconstruct tissue-specific (TS) protein-interaction (PPI) network by mapping TS RNA-seq profiles from 33 human tissue samples to tissue-naive PPI data. **(ii)** TS networks are functionally annotated using Gene-Ontology Biological Process (GO-BP) terms. **(iii)** TS networks are functionally enriched to keep only functions characteristic of each tissue compared to tissue-naive network. **(iv)** BioInt units are generated from the lists of enriched functions in each tissue context. The BioInt units are made by groups of proteins physically interacting and annotated by the same enriched GO-BP term. Only BioInt units including <200 proteins are selected to construct the 33 TS BioInt libraries. **(B) Bar plot** summarizing total number of BioInt units identified in each TS library. **(C) Density plot** describing the number of proteins incorporated in each BioInt unit in each TS library. **(D) Line plot** describing the transcript/gene recovery along the framework. X-axis points represent three of the steps defined in the framework (in panel **A**): **(i)** total transcripts/genes in TS RNA-Seq profiles; **(ii)** proteins in TS networks annotated with at least one GO-BP; and **(iv)** final number of proteins in the selected BioInt units. Step **(ii)** and **(iii)** returned very similar protein coverage (not shown for clarity). Colored bars and lines in panel **B**, **C** and **D** point to six illustrative specific libraries from tissues: testis, bone-marrow, spleen, lymph-node, muscle and brain. **(E) Bar plot** summarizing BioInt-U performance at identifying tissue-consistent BioInt units in three representative tissues: immune, brain, muscle. TP: Number of functions correctly annotated **(ii)**, enriched **(iii)** or selected **(iv)** in the expected tissues. FP: Number of tissue-specific functions assigned to other than the expected tissues.

enrichment did not affected TS transcriptome coverage (step **(iii)** omitted in Fig. 1D).

In order to minimize shallow functional terms, only BioInt units with <200 proteins were selected for the BioInt libraries. This filtering step discarded ~50% of the enriched GO-BP terms and reduced the TS transcriptome coverage down to 40% (see **(iv)** in Fig. 1D). Supplementary Table 1 summarizes the properties of each tissue set in the successive steps. Knowing that vague and general terms of GO-BP tend to be associated with many genes, it is likely that large BioInt units, including numerous genes, are not functionally very informative. Therefore, we interpret that the genes/proteins we missed by filtering by size were only annotated with shallow terms (i.e., superficial in the Gene Ontology and rather general), and so are assigned to still poorly defined functions.

Despite the sharp decrease, we confirmed in the forthcoming analysis that the filtering of large BioInt units did not exclude tissue-specific annotations. To assess the ability of BioInt-U to characterize tissue-specific (TS) and housekeeping (HK) processes, we classified the BioInt units into 22 broad "functional categories" (Supplementary Table 2). For three representative tissues, we calculated the number of these "functional categories" correctly enriched to the expected tissue (considered as true positive cases, TP) and the number of tissue-specific functions assigned to other than the expected tissues (false positive cases, FP) (Fig. 1E). We confirmed that the functional enrichment was crucial to discard FP annotations, especially in brain and muscle libraries (Fig. 1E, *blue bars*). Moreover, we confirmed that size filtering does not affect the selection of tissue-specific BioInt units (Fig. 1E, *red bars*).

2.4. BioInt units represent functional assemblies beyond molecular machines

Molecular machines are commonly defined as "assemblies of molecular components that are designed to perform machine-

like movements" [3]. The main components of many molecular machines are proteins/polypeptides, such as the proteasome, spliceosome, respiratory chain complexes, etc. Being that molecular machines lie at the center of every biological process, we expect BioInt units to incorporate them. To address this issue, we took advantage from the CORUM repository as gold standard of curated molecular machines [14], and evaluated the degree of overlap with TS-BioInt libraries. We first confirmed that ~90% of proteins involved in molecular machines are actually mapped in TS PPI networks and functionally annotated in GO-BPs (Fig. 2A). CORUM protein coverage dropped when filtering-out BioInt units with more than 200 proteins, indicating that a fraction of CORUM-annotated molecular machines were only incorporated in the largest BioInt units. It is noteworthy though that the decrease in coverage was less pronounced than when considering overall transcripts possibly indicating that CORUM complexes are also incorporated in smaller BioInt units (Fig. 2B). As expected, due to their central roles in cellular activity, we observed that proteins involved in molecular machines tend to be more ubiquitously expressed than proteins not identified as being part of any molecular machine in the CORUM repository (Wilcoxon Rank Sum test, p-value < 10^{-4}, Fig. 2C).

In order to assess the protein overlap between BioInt units and CORUM complexes (BiU-CO pairs), we first combined the TS-BioInt units into a unified library. The combination of all BioInt units identified along the 33 tissues returned a unified Bioint library consisting on 728 unique BioInt units including 7765 proteins overall. Due to the size imbalance between BioInt and CORUM complexes, the pair-wise overlap was addressed using Simpson's similarity (SS) index (see Methods). For each BioInt unit, we calculated the maximum SS index found with at least one CORUM complex (Fig. 2D). Next, we compared the percentage and total number of overlapping complexes at increasing SS index intervals (Fig. 2E and F). Lastly, we evaluated the SS index distribution along all pairs of complexes sharing at least one protein (Fig. 2G). We first confirmed that all BioInt units partially intersected with numerous molecular machines (SS index >0.25, Fig. 2D and E). Further, we found that more than half of the BioInt units can partially incorporate up to 50 molecular machines (SS index 0.25–0.50 Fig. 2F). Most notably, more than 60% of BioInt units displayed a SS index higher than 0.75 with at least 5 CORUM complexes in average (Fig. 2D and E). Fig. 2H illustrates several examples of BioInt units incorporating complete or close to complete molecular machines. We found that numerous CORUM molecular machines incorporated into single BioInt units were related to DNA and RNA metabolism. This is in good agreement with the fact that ribonucleic acid biogenesis and processing is exerted through successive biochemical processes that require the collaboration of multiple molecular machines. Overall, these results indicate that BioInt units can recapitulate how multiple molecular machines collaborate in more complex biological processes. Supplementary Table 3 provides the full results of the pairwise SS analysis between BioInt units and CORUM complexes in human. Supplementary Table 4 provides all properties and relevant information regarding the BioInt units generated from the analysis.

2.5. BioInt units are not redundant and recapitulate protein multifunctionality

We also addressed the overlap between BioInt units (BiU-BiU pairs) to evaluate functional redundancy and protein multifunctionality. We found that almost every BioInt unit slightly overlapped with at least one additional BioInt unit (SS index <0.25, Fig. 2E). Furthermore, ~57% out of the >10^6 possible BiU-BiU combinations shared at least one protein indicating that Bioint units frequently overlap (Fig. 2G). Notwithstanding, the SS index was consistently lower than the one observed for BiU-CO pairs. Likewise, the number of complexes overlapping with a SS index >0.25 was significantly lower than when considering the BiU-CO overlap (Wilcoxon Rank Sum test, Fig. 2E). The low but consistent overlap suggests that most proteins tend to be involved in varied functional consortia. Thus, the overlap analysis confirmed that BioInt libraries are not exceedingly redundant but rather recapitulate protein multifunctionality. Conversely, BioInt units incorporate complete or close to complete molecular machines characterizing the molecular activities at the center of biological processes (schematic interpretation in Fig. 2I).

2.6. The functional landscape of tissue-specific BioInt libraries is consistent with the characteristic functions of each tissue

We next investigated whether TS BioInt libraries recapitulate the functional landscape expected for each tissue. To do so, the BioInt units incorporated in each TS BioInt library were first assigned to 22 broad functional groups (see Methods). Then, we evaluated the distribution of these 22 functional classes along the 33 reference tissues. The analysis corroborated that TS processes such as muscle or neuron-related processes are distinctively enriched in the expected tissues (hyper-geometric test, p-value < 0.05, Fig. 3A). This is shown, for example, for: neuron and brain, mitosis and testis, or muscle and heart. Conversely, transversal processes as signaling, DNA, RNA or protein metabolism were consistently identified across all the tissues, corroborating that these BioInt units are actually reflecting housekeeping (HK) functions. Notwithstanding, multiple HK functional classes were significantly enriched in particular organs. Direct inspection of these cases, however, reveals striking agreement with known organ and tissue biology. Examples include the enrichment of signaling-related BioInt units in lung tissue, lipid metabolism processes enrichment in liver, or mitosis overrepresentation in testis. This result is in conformity with the general conception that different tissues rely more heavily on certain basal processes than others. Furthermore, the analysis revealed that the function types considered as HK can be divided in two subgroups based on their distribution across the tissues (Fig. 3B). The majority of BioInt units related to RNA, mitochondria, organelle trafficking, protein metabolism and localization were essentially detected across all tissues (blue plots in Fig. 3B). In contrast, many functional groups as signaling, mitosis or cytoskeleton incorporated BioInt units with mixed expression patterns (purple plots in Fig. 3B).

2.7. Dissection of BioInt units brings insight into the mechanisms underlying tissue functional diversity: Ubiquitous (UB) and non-ubiquitous proteins collaborate in HK and TE functions

Excluding the transcriptomic profiles of sexual tissues, we found that a large fraction of the gene transcripts (9,686 expressed genes) to be ubiquitously (UB) expressed across the TS transcriptomes. However, the distribution of BioInt units across tissues drew a notably distinct pattern when compared to transcript expression (Fig. 3C). We found 357 BioInt units annotated in <5 tissues (hereafter-called tissue enriched units, TEu) and 122 units annotated in more than 28 tissues (housekeeping units, HKu). While all TS networks incorporated ~70% of UB proteins on average, the percentage of housekeeping units dropped to 17.3% (Fig. 3C). These trends are likely justified by the observation that both HKu and TEu incorporated a mixed composite of UB and nonUB proteins (Fig. 3F). In particular, we found that TE units incorporated a large percentage of UB proteins and HK BioInt units also included a small fraction of nonUB proteins.

Being that different proteins can exert similar biochemical activities, we hypothesized the same HK functional unit might

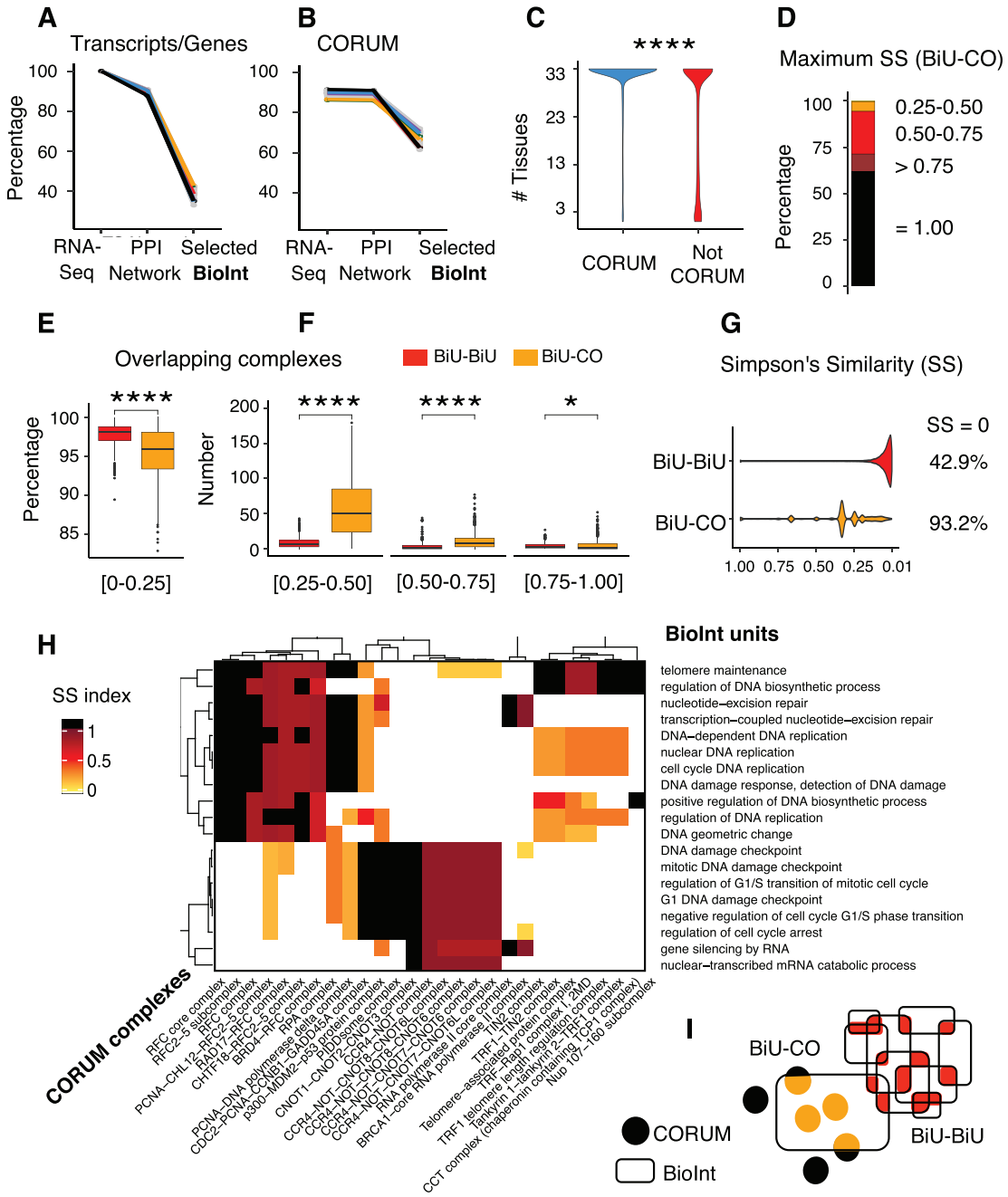

Fig. 2. Mapping of molecular machines from CORUM repository to BioInt units. Line plots describing mapping % of transcripts and proteins retrieved from each TS-RNA-Seq profile **(A)** or CORUM repository **(B)**, in the TS PPI networks annotated with at least one GO-BP, and in the selected BioInt units. **(C) Violin plot** comparing the tissue expression distribution of proteins identified in CORUM complexes (blue) and transcripts only identified in TS-networks (red). Wilcoxon Rank Sum test, p-value < 10^{-16}. **(D) Staked barplot** summarizing the maximum Simpson's similarity (SS) index found when mapping BioInt units to CORUM complexes. **Box plots** describing the % and total number of BioInt units sharing proteins – i.e., overlap – with other BioInt units (BiU-BiU) **(E)** and CORUM complexes (BiU-CO) at increasing SS index intervals **(F)**. Wilcoxon Rank Sum test; p-value **** <0.0005 and * <0.05. **(G) Violin plots** comparing SS distribution between BioInt units and CORUM complexes. The distributions do not include pair comparisons with no overlap (SS = 0). **(H) Heatmap** representing SS index between an illustrative subset of BioInt and CORUM complexes (BiU-CO). **(I) Schematic picture** of the predominant types of overlap found in the comparisons between BioInt units (BiU-BiU) and between units and complexes (BiU-CO). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

incorporate varying proteins depending on the tissue of context. Additionally, we sought to assess whether the percentage of protein variability in HKu could be associated to their functional roles. We calculated the heterogeneity of each HKu as the percentage of proteins found in common along all tissues (Fig. 3D). Protein variability analysis generated a bimodal density plot in which two major groups can be distinguished: **i)** heterogeneous HKu with more than 20% of protein variability and **ii)** highly consistent HKu with tissue variability below 10%. Similarly when evaluating BioInt unit distribution profile across tissues (Fig. 3B), we found that heterogeneous HK units are more frequently associated to functional classes with mixed expression patterns as signaling, mitosis or stress-related processes; while monotonous HK units are distinctively related to RNA, DNA or protein metabolism and localization**.** Nonetheless, these trends might indicate that functions considered as "mixed HK" are less characterized, at PPI

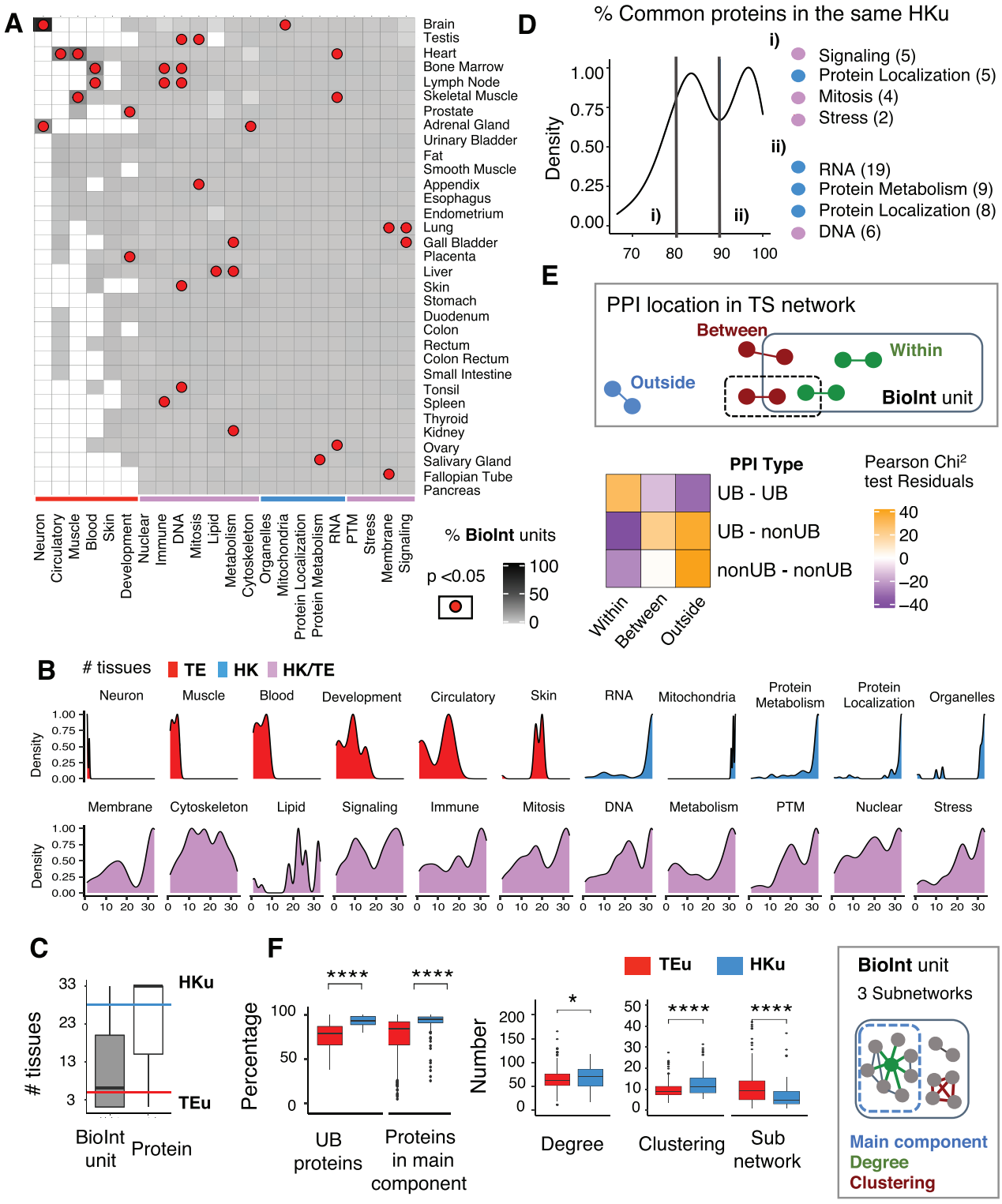

Fig. 3. Analysis and comparison of functional and topological features of BioInt units with distinct tissue distributions. (A) Heatmap representing the percentage of BioInt units associated to each functional category (columns) along the tissues. BioInt units were classified into 22 general functional categories by text mining key words in the BioInt unit description (see Methods). Red dots point statistically significant enrichments, hyper-geometric test; p-value < 0.05. Bottom column color divides function classes according to density distribution in panel B. **(B) Density plots** describing the tissue distribution of BioInt units assigned to each functional category. Density plots are filled in red, blue and purple to point functional classes with tissue enriched (**TE**), housekeeping (**HK**) and mixed **HK/TE** expression, respectively. **(C) Box plot** comparing tissue distribution of transcripts and BioInt units. Colored lines separate the TE and HK units identified in <4 tissues or in more than 28 tissues, respectively. **(D) Density plot** describing the percentage of common proteins identified for the same HK unit along the tissues. Vertical lines highlight the two classes of BioInt units according to protein expression heterogeneity. **(E) Heatmap** representing the statistical association between the type of PPI and their location along BioInt units (illustrated in top box). Color scale represents the Pearson's residuals obtained from Chi^2 test p-value < 10^{-4}. **(F) Box plots** comparing the network properties of HK units (**HKu**) and TE units (**TEu**) (illustrated in right box). Left to right: Percentage of UB proteins in each BioInt unit and total proteins in the main component (largest connected subnetwork) of the unit. Average protein degree, average protein global clustering coefficient in TS-network and number of subnetworks by BioInt unit. Wilcoxon Rank Sum test; p-value **** <10^{-4} and * <0.05. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and/or Gene Ontology level, than functions classified as consistent HK units. Either way, these results support the notions that, ubiquitous and non-ubiquitous proteins collaborate in TE or HK processes and some HK functions could acquire additional relevance in certain tissues.

2.8. Network characterization of ubiquitous and tissue-specific BioInt units

Protein localization at the global network can provide valuable information on the coding-gene evolutionary history and current functional essentiality. At the same time, the PPI location in the BioInt units can also indicate whether the protein interactions play a core role in a given function or, rather, coordinate complex functional mechanisms. On this basis, we evaluated the position of HK and TE units in the TS networks using standard network connectivity measures (right box in Fig. 3F). We found that HK units frequently incorporated more proteins and these predominantly located in the main component (i.e., the largest connected subnetwork of the BioInt unit, Fig. 3F). Moreover, the proteins collaborating in the HK units displayed significantly larger average degree (lager number of interactions per protein) and global clustering coefficients (larger interaction density in the protein neighborhood), indicating that HK units hold central positions in TS networks (with a significant difference according to the Wilcoxon tests, Fig. 3F).

To further explore the biological implications of the collaboration between UB and nonUB proteins, we addressed the frequency of homotypic (UB-UB or nonUB-nonUB) and heterotypic (UB-nonUB) PPI interactions at distinct locations in the TS networks: (1) outside units, (2) within the same unit or (3) between two units or one unit to outside (Fig. 3E). We applied the Chi-square test to evaluate the statistical association between the type of PPI and its location in the network (p-value < 10^{-4}) and used Pearson's residuals to describe the positive or negative association between the conditions. As expected given the ratio of UB and nonUB proteins along BioInt units, we found that UB-UB interactions are more frequently located within BioInt units while UB-nonUB hetero and nonUB-nonUB homotypic interactions are significantly located outside the Bioint units. Most notably, heterotypic interactions also appeared frequently connecting the BioInt units with proteins outside the units. Overall, these results indicate that HK units are central in the TS networks and further; UB-UB PPIs lie at the center of BioInt units. On the other hand, the heterotypic interactions between nonUB-UB proteins seem to be key to link the functions in the network.

2.9. Systematic mapping of disease genes (DG) in BioInt-U reveals potential large-scale topological vulnerabilities: DGs are widely expressed but accumulate in TEu

The preferential location of disease-associated genes (DGs) in the TS networks may bring valuable insights into sensitive points in network connectivity. To explore this, we collected 9,259 DG associations for 1,948 pathologies from the DisGeNET repository [28]. Our global transcriptome covered 86.8% of DGs and conversely, 43.5% of transcripts (expressed genes) were associated to at least to one disease (Fig. 4A). The DG coverage was barely affected when considering the proteins in the TS-networks but notably dropped in the selected BioInt libraries (Fig. 4B). This indicates once again that a fraction of DGs is only incorporated in units including more than 200 proteins. Furthermore, we found that more than 55% of total DGs were ubiquitously expressed and overall, displayed a broader expression profile than nonDG proteins (Wilcoxon Rank Sum test, p-value < 10^{-4}, Fig. 4C).

However, when evaluating the DG% by BioInt unit, we found DGs preferentially accumulate in TE units with lower percentage of UB proteins (Fig. 4D and E). In fact, the BioInt Units that incorporate the highest percentage of DGs are almost exclusively annotated in <5 tissues. Unexpectedly though, BioInt units accumulating >50% of DGs are more sparsely connected (i.e., exhibit an smaller main component and more subnetworks), but incorporate proteins with more central positions in the TS network (Fig. 4G). To evaluate whether the DGs tend to accumulate in any particular type of function, we took profit from the functional classification retrieved in Fig. 3B, and found that BioInt units including 50–80% DGs were implicated in most types of functions (column Y in Fig. 4F). However, BioInt units accumulating more than 80% of DGs were found to be more frequently related to TE or mixed HK/TE processes. Concordantly, the less targeted BioInt units were distinctively associated to HK biological processes.

2.10. Interaction of proteins encoded by DGs predominantly located between highly overlapping TEu

Having confirmed that tissue-enriched BioInt units (TEu) tend to incorporate more DGs, we next questioned whether the PPIs between DGs would present distinctive positions in the HK units or in the TE units, that might indicate topological vulnerabilities. We found that DG-DG interactions were more frequently located within BioInt units while nonDG-nonDG interactions were more frequent outside (Fig. 4H). More important, DG-DG interactions were also notably located between TE units. We also found that DGs tend to be found in BioInt units presenting a larger overlap when compared to nonDGs (Fig. 4I).

The connectivity properties of any protein directly depend on the range of PPI available at each TS network. This feature might be crucial to understand the variable impact the same DG can have in different tissues. We found that the variation (in terms of standard deviation) of betweenness and degree coefficients were significantly larger in DGs than other proteins not associated to any disease (Fig. 4I). This observation may provide critical insights into the mechanisms underlying TS disease-phenotypes linked to DGs with wide distribution.

2.11. Genes associated with TS diseases accumulate significantly in BioInt units characteristic of the target tissue

To further explore the mechanisms underlying the emergence of TS patho-phenotypes, we next proceed to evaluate the DG mapping at disease-specific and tissue-specific levels. From the 1,948 diseases annotated in DisGeNET, we identified 463 diseases unambiguously associated with 11 tissues (for example, "nefrotic failure" is a kidney dysfunction or "T-cell lymphoma" is associated to alterations in the immune system). The complete list of disease-tissue associations and DGs is available in Supplementary Table 5. It is reasonable to assume that the functions with most critical roles for a given tissue will accumulate more DGs found in the patient population. Likewise, it is also reasonable that a functional unit will only be efficient when a large fraction of its components are available in normal standards. Of note, the DG associations in DisGeNET do not only refer to causal mutations but also to biomarkers or de-regulated genes. Thus, it is plausible we could find several DGs simultaneously altered in the same patient. On this basis, we estimated the potential impact of each disease in the TS functions by addressing the overrepresentation of disease-specific DGs in each BioInt unit (hyper-geometric test, p-value < 0.05).

Most diseases exhibit tissue-specific phenotypes from which it follows that DGs should accumulate in certain tissues in particular (hereafter referred to as "tissue-consistent" impact). We have also

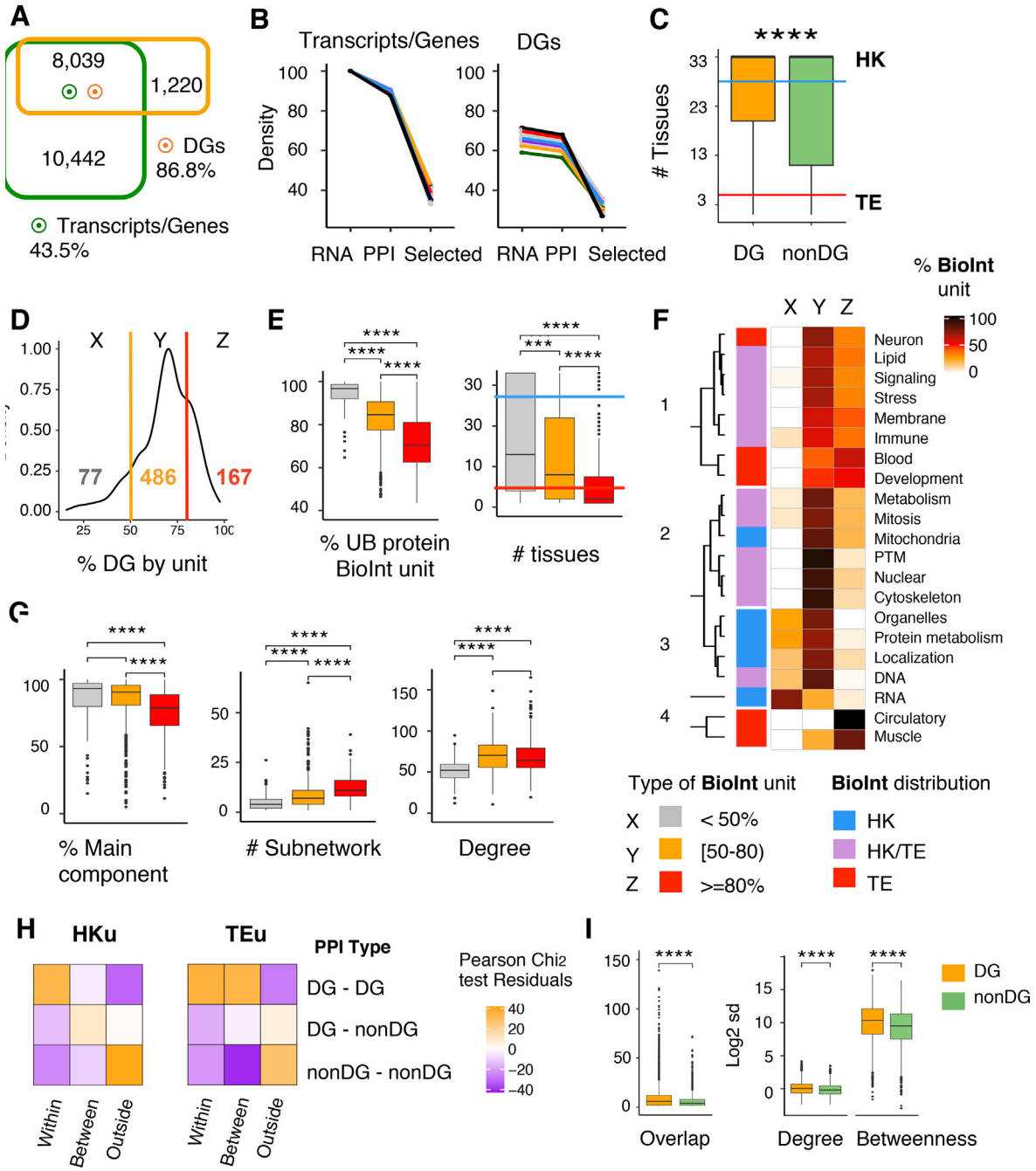

Fig. 4. Systematic mapping of DGs into TS BioInt libraries. (A) Venn diagram illustrating the overlap between our unified transcriptome and DGs from the DisGeNET repository. **(B) Line plots** showing the % of DGs and transcripts mapped to TS RNA-Seq profile, in the TS networks annotated with at least one GO-BP and in the selected BioInt units. **(C) Box plots** comparing tissue distribution of proteins encoded by nonDG (green) and DGs (orange) **(D) Density plot** describing the DG % identified in each BioInt unit overall. Vertical lines separate Bioint units in three groups based in DG %. **(E) Box plots** comparing the DG accumulation (X, Y and Z groups as defined in panel D) with the % of UB protein per unit tissue distribution. **(F) Heatmap** representing the broad functional classes (rows) assigned to the BioInt units including an increasing % of DGs (columns). Left dendogram and clusters result from a complete-linkage clustering using Euclidean distance. Left column summarizes the functional classes according to tissue expression patterns observed in Fig. 3B. Tissue-enriched (TE, red), ubiquitously-expressed (HK, blue) and mixed HK/TE functions (purple). **(G) Box plots** comparing the protein % incorporated in the largest subnetwork of the BioInt unit (main component), the number of subnetworks by BioInt unit, and average protein degree. **(H) Heatmaps** representing the statistical association between the type of PPI and their location along HK and TE units. Color scale represents the Pearson's residuals obtained from Chi^2 test p-value < 10^{-4}. **(I) Box plots** comparing the total overlap along BioInt units including or not including DGs and the standard deviation (sd) of degree and betweenness coefficients of DGs and nonDGs across TS networks (**C, E, G and I**) (Wilcoxon Rank Sum test; p-value **** <10^{-4} and *** < 10^{-3}). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

discussed in previous section that different cell types might specialize in certain functions. Based on this, we speculate the DGs of tissue-consistent diseases might accumulate in functional classes characteristic of given tissue physiology. The BioInt units enriched in tissue-consistent DGs were homogeneously related to almost all types of functions (Fig. 5A). Thus, to increase the analysis resolution, we only considered Bioint units enriched in DGs for at least 10 tissue-consistent diseases (corresponding to top 3rd Quartile) (Fig. 5B). We found that the BioInt units enriched in tissue-consistent DG lists are accordingly involved in functions specific

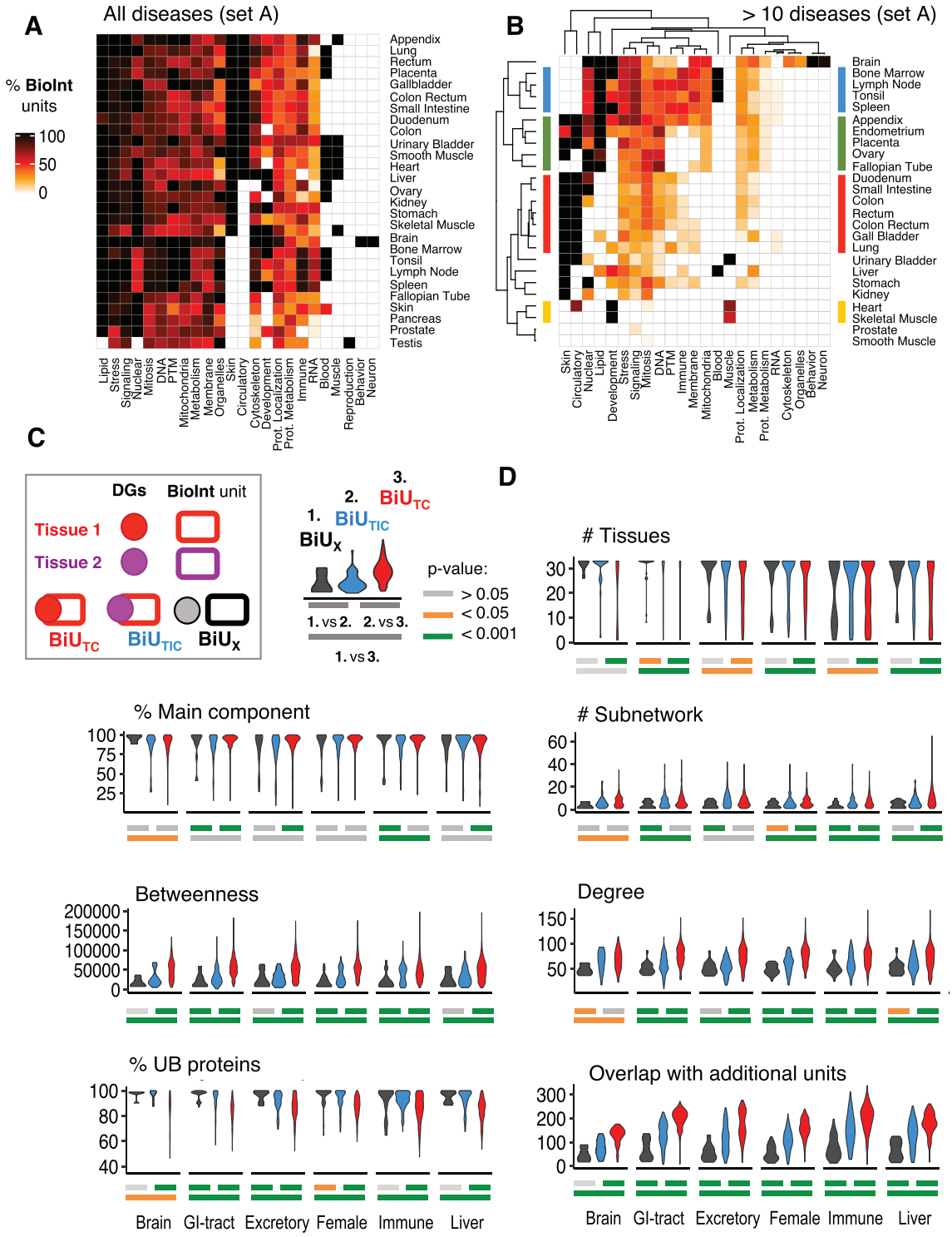

Fig. 5. Comparison of topological properties of BioInt units accumulating tissue-consistent DGs. Heatmaps representing the % of Bioint units by functional class enriched in at least one tissue-consistent disease **(A)** or in at least 10 tissue-consistent diseases **(B)**. Dendograms and clusters result from a complete-linkage clustering using Euclidean distance. **(C) Schematic picture** illustrating the types of BioInt unit evaluated in panel **D**. For each TS BioInt library we can distinguish: BioInt units enriched in DGs of tissue-consistent diseases (BiU_{DT}); BioInt units enriched in DGs of diseases not expected at the tissue (BiU_{D}); and BioInt units not significantly enriched by any DG set (BiU_{X}). **(D) Violin plots** comparing the network properties of these three types of BioInt units defined above. (Bottom bars indicate the statistical significance, Wilcoxon Rank Sum test p-value; grey > 0.05, orange < 0.05 and green < 0.001). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

to the tissues in consideration. This trend was most conspicuous in TE BioInt units related to immune, muscle or neuron functions, wich are predominantly enriched in DGs from tissue-congruent diseases. Likewise, the clustering analysis corroborated that tissues associated to the same broad histological groups (colored rows in Fig. 5B) significantly accumulate DGs in units involved in the same functional classes (columns in Fig. 5B). This analysis further enabled to discern that several function classes considered as HK processes were distinctively altered in different tissues. For example, DNA-related functions appeared to be more frequently altered

in female-organ diseases than in gastro-intestinal disorders. Likewise, stress and signaling-related functions were preferentially altered in immune system-related disorders.

2.12. BioInt units enriched in tissue-consistent DGs (BiU_{TC}) exhibit distinctive network properties

From the total 8,285 of BioInt units identified in the 25 TS libraries considered for DG mapping (Supplementary Table 5), 60% were significantly enriched in DGs of tissue-consistent pathologies (set named **BiU_{TC}** in the schematic representation in Fig. 5C). Nonetheless, a 35.8% of these BioInt units were also enriched in DGs associated to diseases specific to other tissues (referred to as tissue-inconsistent, and set **BiU_{TIC}** in Fig. 5C). This indicates that the enrichment in DGs is not sufficient to justify the emergence of the pathology. Thus, we speculated that the BioInt units that are really decisive to trigger pathomechanisms must hold distinctive properties in the network topology. To explore this hypothesis, the 25 TS libraries enriched in at least one tissue-consistent pathology were grouped in 11 major organ groups (Supplementary Table 5) to then compare the topological properties of: BioInt units enriched in tissue-consistent diseases (named BiU_{TC}); BioInt units enriched in tissue-inconsistent DGs (BiU_{TIC}); and BioInt units not enriched in any DGs (BiU_X, Fig. 5D).

The network analysis at disease-specific level revealed the same trends observed for the systematic DG mapping (Fig. 4). BioInt units accumulating DGs of tissue-consistent diseases (set BiU_{TC} in red in Fig. 5C and D) tend to be expressed in fewer tissues and incorporate less UB proteins than BioInt units not affected by any disease or enriched in tissue-inconsistent DGs (sets BiU_{TIC} and BiU_X in Fig. 5C,D). Furthermore, the comparative analysis revealed that the BioInt units in set BiU_{TC} frequently included proteins with higher degree and betweenness coefficients in the global TS PPI networks. Most remarkably, BioInt units enriched in DGs of tissue-consistent diseases displayed a larger overlap with additional functional units. In fact, the DGs assigned to tissue-consistent diseases are 1.5 times more frequently located at the intersection between BioInt units than proteins encoded by nonDG (Wilcoxon test, p-value < 10^{-4}). However, an unexpected observation is that the percentage of proteins in main component is similar but BioInt units in set BiU_{TC} exhibit a larger number of disconnected subnetworks according to our current map of protein interactions.

2.13. A case study: Mapping of differentially expressed genes to BioInt units predicts most vulnerable tissues and functions in pulmonary fibrosis and psoriasis

The dissection of the molecular mechanisms underlying complex diseases is still an open challenge. One of the most widely used strategies to investigate pathological conditions is the identification of differential expressed genes (DEg) in RNA-Seq profiles from patient-derived samples.

However, the most popular algorithms for DEg analysis assess the expression of each gene independently, thus DEg datasets frequently include a large number of transcripts/proteins disconnected from the PPI network. Likewise, gene expression is highly dynamic and so DEg datasets characterizing the same disease often give different profiles. All this makes the DEg data difficult to integrate and interpret. The integration of differential gene expression profiles with functional enrichment analysis in protein interaction networks has been recently proposed to assist in the prioritization of disease-relevant targets [27]. In a similar argumentative line and to test the analytical procedure presented in this work, we next illustrate how the mapping of disease-related DEg profiles into BioInt libraries can improve the prioritization of potential functional targets. We selected and analyzed two independent transcriptomic profiles characterizing gene expression changes in patients suffering from psoriasis and idiopathic pulmonary fibrosis [26,31]. Within these datasets, 91% of fibrosis-related and 83.5% of psoriasis-related DEg were mapped in our unified transcriptome dataset, respectively (Fig. 6A and B).

We next collected DGs associated with psoriasis and fibrosis in DisGeNET. Despite of the large number of DGs already associated to fibrosis (203), only 6.9% were found to be DEg in the transcriptome profile (Fig. 6A). On the other hand, we only identified 30 DGs associated to psoriasis and none was DEg (Fig. 6B). Similar to our previous analysis, we calculated the overrepresentation of DEg in each BioInt unit across all tissues (hyper-geometric test, p-value < 0.05) and selected the 25% most affected units. Interestingly, we found that the tissues including more functional units significantly enriched in DEg were precisely those in which the symptoms are commonly observed (Fig. 6C). Furthermore, the functional types accumulating highest percentage of BioInt units enriched in DEg were also related to functions suspected to be critical in the diseases (Fig. 6D and E). Finally, Fig. 6F and G summarize the functional signatures associated to the BioInt units enriched in DEg from fibrosis and psoriasis profiles in lung and spleen tissues, respectively. To simplify the analysis, we collapsed the BioInt units (dots) presenting a Wang's Semantic similarity coefficient >0.6 into functional clusters (top 10 largest clusters are arranged in Y axes). In particular, the BioInt units most targeted by fibrosis-related DEg in lung included membrane permeability, proteolysis and apoptosis signaling-related functions [35]. In the case of psoriasis, stress-related protein folding and degradation-regulatory pathways were consistently altered in immune-related organs [39]. The analysis confirmed that DE genes preferentially accumulate in biological processes already involved with fibrosis and psoriasis. Therefore, our analysis illustrates how BioInt units can provide additional insight into why these functions are more vulnerable and also suggest new DG candidates for further evaluation.

3. Discussion

The topological characterization of TS networks is crucial to dissect the mechanisms underlying tissue functional diversity and identify potential vulnerabilities, namely those related to genetic disorders. However, to our best knowledge, most investigations have focused on characterizing the topology of individual proteins and DGs without considering their functional context (recently reviewed by [19,24,43]. However, it should be noted that PPI networks are static representations of all the physically possible interactions, and these may not be always biologically meaningful. We advocate that the integration of proteins within their functional context can improve the assessment of network properties relevant for cell physiology. On this basis, we designed a network-based strategy to characterize functionally collaborating TS PPI consortia. We applied this framework on 33 human TS networks and conducted a systematic study of the topology patterns associated to distinct normal and pathological states. This analysis revealed how the topological properties of functional units may elucidate the mechanisms of TS functional diversity and deregulation (hypothesis illustration in Fig. 7).

As the very name implies, housekeeping (HK) functions are essential for the survival of any type of cell and are mostly exerted by ubiquitous (UB) proteins expressed in all tissues. Evolutionary selection has favored proteins involved in these functions and so UB proteins dominate TS network composition, accumulate more PPIs and locate at central positions in TS networks [6,7,11,23]. Beyond the characterization of individual proteins, the systematic analysis of TS BioInt libraries further supported an in-depth com-

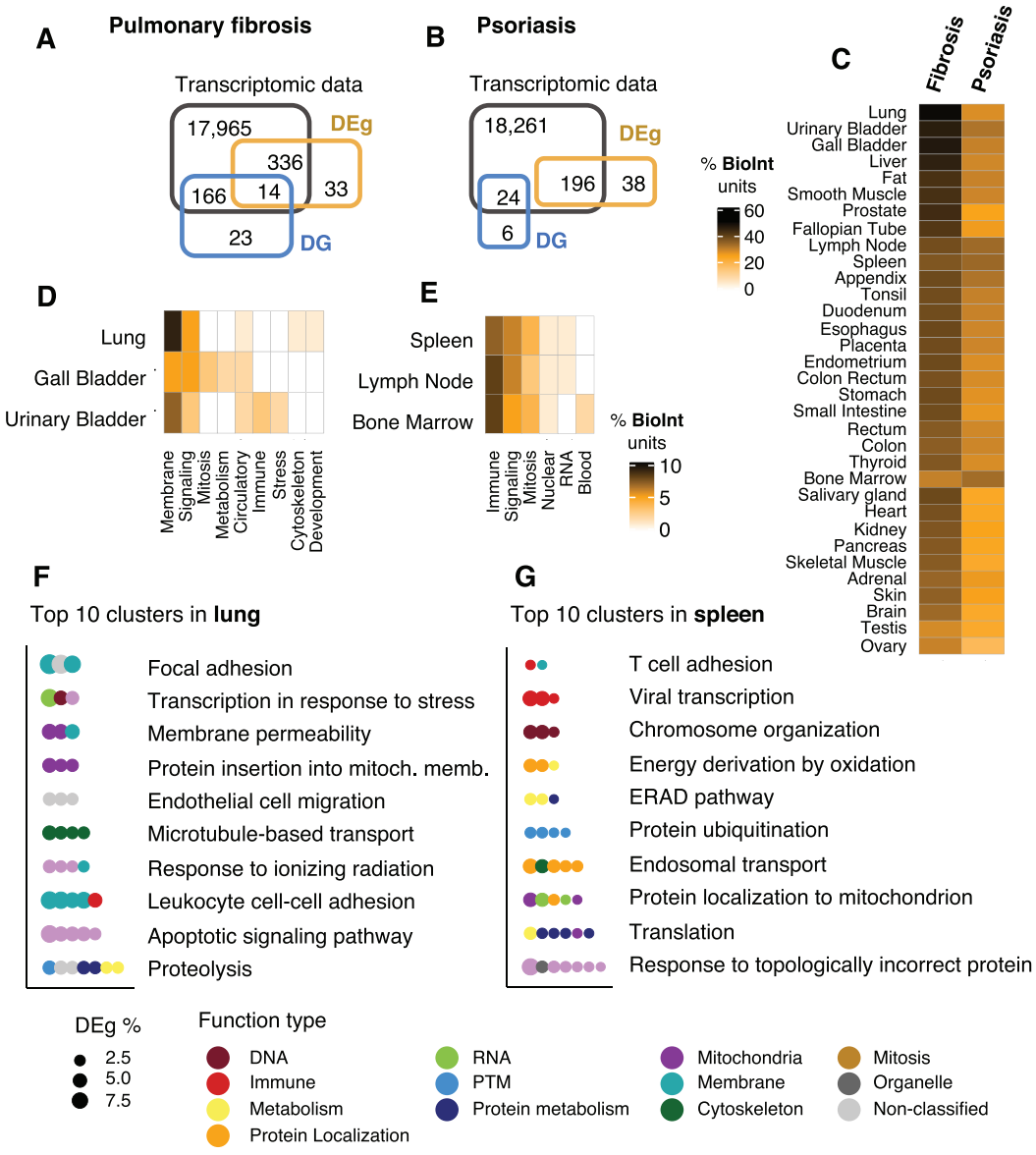

Fig. 6. Mapping of pulmonary fibrosis and psoriasis RNA-Seq gene expression profiles into TS BioInt libraries. (A, B) Venn diagrams summarizing the overlap between unified transcriptome, DGs collected from DisGeNET and differentially expressed genes (DEg) identified in RNA-Seq profiles of patients suffering from pulmonary fibrosis and psoriasis, respectively. **(C) Heatmap** representing the percentage of BioInt units enriched in DEg across TS libraries. **(D, E) Heatmaps** representing the types of functions accumulating most BioInt units enriched in DEg in the top 3 TS libraries in panel **C**. **(F, G) Dot plots** summarizing the BioInt units enriched in DEg identified in fibrosis and psoriasis profiles in lung and spleen tissues, respectively. Each dot represents a single BioInt unit and functions with >0.6 Wang semantic similarity are grouped in clusters (Y-axis). The figure only includes top 10 largest clusters. The text describes the parent GO-BP in common to the BioInt units grouped together. Dot color indicates the type of function (bottom legend) and size the % of DEg.

parison between HK and TE functions. We corroborated that HK units are related to core functions such as organelle trafficking, RNA or protein metabolism and are mostly made up of UB proteins with significantly larger degree and betweenness coefficients than proteins exclusively involved in TE functions. Most HK units included a small percentage of nonUB proteins that varied across the TS networks. In parallel, TE units also incorporated a large percentage of UB proteins (Fig. 7A). While the extensive re-use of UB proteins in TS functions is well described [7,8,30], the role of nonUB proteins in HK functions is less studied. Our analysis corroborated that UB–UB PPIs are frequently located within functional units highlighting their fundamental roles at the core of the biological processes (Fig. 7C). Conversely, we found that heterotypic nonUB–UB interactions preferentially connect functional units with other proteins outside the network. Our observations are in line with a recent investigation showing that cell-specific interactions link protein complexes in the TS interactome [20] and underscore that nonUB proteins are critical players in the coordination of both HK and TE functions.

It is reasonable to assume that the characterization of mechanisms underlying tissue functional diversity will bring insights into the events triggering TS diseases. The pioneer studies characterizing the DGs topology suggested that deleterious proteins tend to display TS expression [15,21]. Currently though, we find innumerous instances of UB proteins involved in diseases with tissue-restricted phenotypes. This indicates that TS protein expression is not sufficient to explain the emergence of TS diseases [19]. Barshir and colleagues found that DGs tend to display tissue-exclusive PPIs

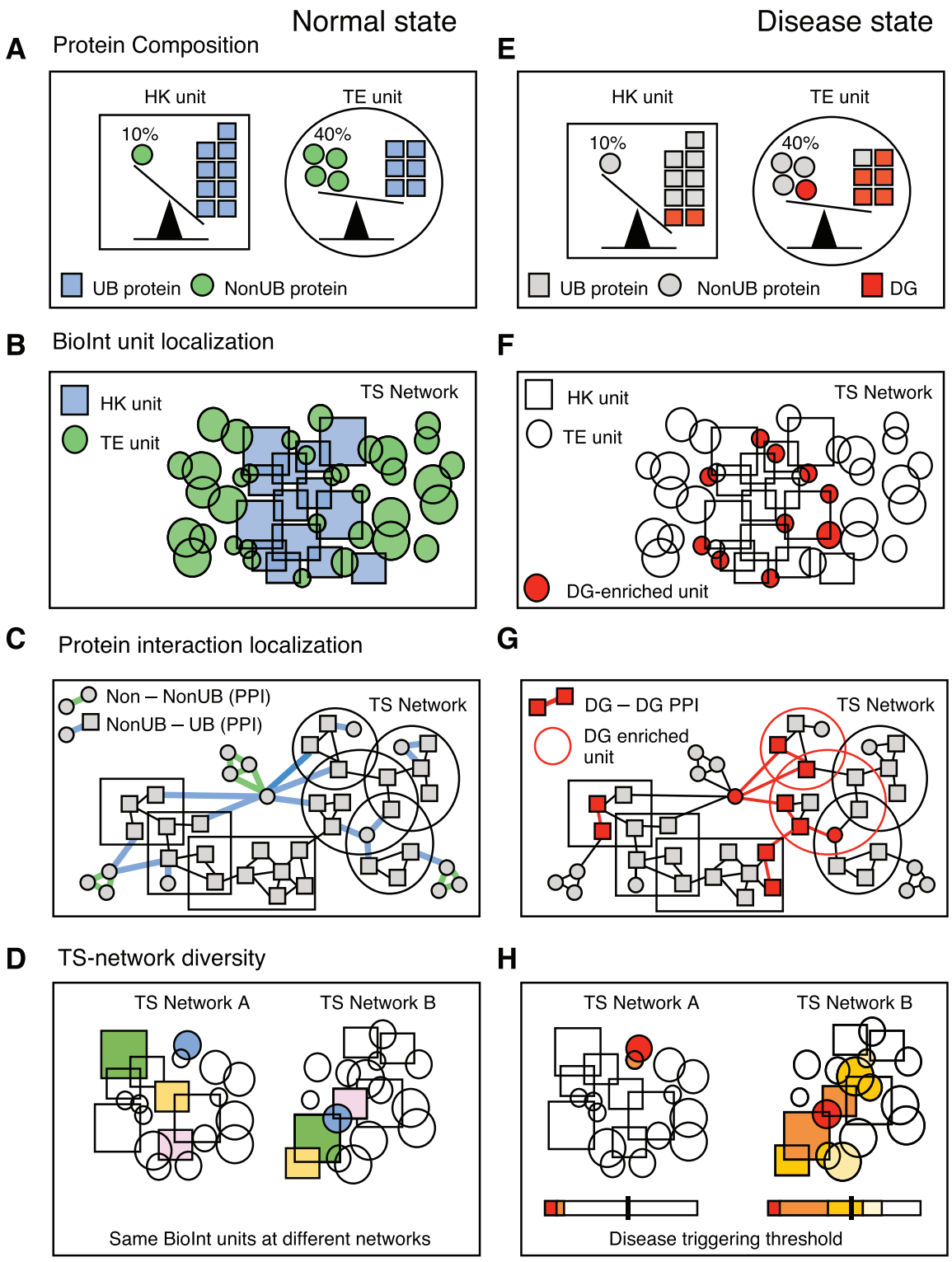

Fig. 7. Schematic drawings presenting the concepts and hypotheses proposed about the mechanisms underlying functional diversity and tissue vulnerability linked to TS protein networks. The main conclusions retrieved form this work can be applied in normal (left column) and disease (right column) conditions**. First column: (A)** HK units are mostly made of ubiquitous (UB) proteins while tissue-enriched (TE) units incorporate a balanced percentage of UB and nonUB proteins. **(B)** Housekeeping (HK) units include proteins with significantly larger degree and betweenness coefficients when comparing to proteins in TE functions. **(C)** Homotypic UB protein-interactions (PPIs) are frequently located within functional units, homotypic nonUB PPIs lay outside functional units and heterotypic interactions preferentially connect the functional units with other units or proteins outside in the tissue-specific (TS) network. **(D)** Proteins can establish different interactions according to their interactomic neighborhood and so the TS networks can show distinct topological rearrangements. **Second column: (E)** Disease genes (DGs) are more widely expressed than nonDGs but significantly accumulate in TE units. **(F)** TE units amassing most DGs also incorporate proteins with larger degree and overlap. **(G)** DG–DG interactions are frequently located within BioInt units. Particularly in TE units, DG–DG interactions do frequently connect TE units with additional proteins or units outside in the network. These observations suggest that the distinctive disease impact observed for TE units might be triggered by DGs with additional roles in functional coordination. **(H)** Overall, the TS connectivity patterns might be key to understand why the impact of UB-DGs could distinctively trigger the degeneration of particular tissues.

in the tissue where the disease is manifested [6]. Lee and colleagues reached a similar conclusion when exploring the topology of neuron-related TS networks and hypothesized this might be key to understand the high prevalence of neurological diseases [22]. The systematic mapping of DGs onto BioInt units corroborated that the transcript products of DGs tend to be more widely expressed than those coded by nonDGs (Fig. 7E). Interestingly though, DGs tended to accumulate in functional units annotated in fewer tissues and DG–DG interactions and were more frequently located at the interface of TE and HK functions or connecting TE functions to other proteins outside in the network (Fig. 7G). These results suggest that the impairment of TE functional coordination might be a key feature to spread TS homeostasis deregulation and overcome the threshold to trigger TS pathophenotypes (Fig. 7D and H).

A more thorough analysis of TS diseases revealed that DGs accumulate more frequently in functional units found in the disease-target tissues. This observation was particularly apparent in TS functions related to muscle, immune function or neuron physiology. Nonetheless, many DGs accumulated in BioInt units in tissues other than the expected, indicating that DG enrichment is not the only event accounting for disease manifestation. In our view, this observation illustrates why the functional characterization of DGs might fall short to understand pathological mechanisms. Proteins are multifunctional and collaborate both in HK and TE functions. In turn, proteins establish dynamic PPIs and, as suggested in this work, acquire varying relevance depending on their TS interactome context (Fig. 7D). In particular, our topological analysis reiterated that the functional units accumulating tissue-consistent DGs were actually TE BioInt units with significantly larger overlap with additional units (Fig. 7E). Although multifunctional proteins have been previously associated to pathological events, our analysis brings further evidence towards this from a TS functional perspective.

In parallel, we made the unexpected observation that BioInt units accumulating most DGs are more sparsely connected, while tending to incorporate significantly more central proteins in the TS network. Dynamic interactions are known to play critical roles in the regulation and coordination of protein function. However, high-throughput PPI detection techniques preferentially detect stable interactions and thus, are more likely to dismiss transient PPIs. Although caution must be taken until the advent of more sensitive technologies, this observation suggests that the most vulnerable functions tend to include numerous transient interactions not yet identified. Our conjecture is aligned with previous results indicating that biological and disease modules do not necessarily coincide with topological clusters [1,13,41]. If confirmed, this observation would question the pivotal role of clustering algorithms in the design of network-based methods for biomedical research.

To illustrate the benefits of the BioInt framework in a real case problem, we took advantage of two public transcriptome profiles from patients suffering from psoriasis and pulmonary fibrosis. The scarcity of already known psoriasis-causal genes together with the low overlap between fibrosis DGs and the corresponding DE transcriptome reflects the need for additional research bridging the molecular and patho-phenotypic observations. The analysis presented here demonstrates the ability of our method to independently identify the most afflicted tissues and functions and thus bring novel insights to refine DG prioritization methods.

The BioInt-U framework sets the stage for novel approaches to explore the functional relevance of TS topological properties. Nonetheless, it also has limitations. The identification of BioInt units relies on PPI and Gene Ontology datasets, which are known to be overfitted by proteins of significant research interest. Until a more comprehensive characterization of the interactome and functionome, our investigation is likely to underestimate poorly characterized players. On the other hand, the analysis exploits static networks and ignores cell-specific temporal information of the particular tissue. The integration of dynamic and quantitative expression data could surely benefit network-centered investigations. Notwithstanding, is worth recalling that the use of quantitative data would also increase the analytical complexity. To compensate the lack of spatiotemporal data, we enabled functional units to overlap. In this way, we could evaluate all the possible combinations of functional consortia.

Overall, the work presented here showcases the relevance of evaluating network topology from the functional perspective. The large-scale topological vulnerabilities inferred from our analysis could contribute to the refinement of network-based methods for DG candidate prioritization. Likewise, the evaluation of the topological context of DGs across tissues could facilitate the identification of the most critical drug targets while avoiding unpredicted off-targets.

4. Methods

4.1. Computational pipeline to define BioInt units

Reconstruction of TS networks. RNA-Seq datasets representing 33 major tissues and organs were retrieved from Uhlén and colleagues work [38]. The datasets were filtered to only evaluate transcripts expressed above 1 FPKM (Fragments Per Kilobase of transcript per Million). The dataset was TMM-normalized (Trimmed mean of M values) using the *limma* R package [33]. Biological replicates were combined calculating the average transcript expression. Next, human physical PPI data reported at least in two experiments was retrieved from the APID repository in April 2021 [2]. The tissue-naive PPI network was filtered to create a TS network including only interactions between proteins coded by transcripts expressed in each TS transcriptome. The TS networks were simplified to remove self-loops and isolated proteins using the *igraph* R package [9].

Functional enrichment of TS networks. The *GOfuncR* R package was used to functionally characterize TS networks in comparison to the unspecific network using Gene Ontology Biological Process (GO-BP), hyper-geometric test, FDR = 0.1 on 500 randomizations [17]. Functional enrichment was simplified into functional groups by collapsing terms with more than 0.9 Jaccard's similarity coefficient, defined as the number of common elements between two sets, divided by the union set size. When GO-BPs are collapsed, the new functional group includes all proteins associated to each term but is assigned to the functional description with fewest characters.

Generation of TS-BioInt libraries. The functional enrichment of TS-networks was used to identify the BioInt units, which consist of groups of proteins physically interacting and annotated under the same enriched GO-BP term. The inconsistencies across high-throughput PPI data and the constant PPI data growth in multiple repositories suggest the human interactomic data is still far from complete. Knowing this, we enabled BioInt units to be formed by non-connected subnetworks. The isolated clusters were discarded only when the main component (largest subnetwork) represented more than 90% of the total BioInt unit. On the other hand, proteins can display transient and varied PPIs. Additionally, most proteins are multifunctional and are frequently annotated with several GO-BP terms. In order to recapitulate protein multifunctionality and the network dynamics, we enabled proteins to be involved in several units simultaneously**.** The Bioint units were classified in 28 functional categories by performing a direct text mining of key words found in the description of functional units. The list of key terms is available at Supplementary Table 2. From the total

28, we selected the 22 functional categories associated with sufficient BioInt units. The network topological analysis was focused on betweenness, degree and clustering coefficient measures that were evaluated using the *igraph* R package [9].

4.1.1. Reconstruction of mouse predicted tissue-specific networks and BioInt units

We reconstructed predicted interactomes for four tissues, based on tissue-specific RNA-Seq profiles obtained from normal mouse samples (ArrayExpress E-MTAB-6081 [37] and tissue-naive mouse interactome (APID database accessed on May 2022, [2]). As suggested by the authors, transcripts expressed above 1.4 RPKM (corresponding to the top 3rd Quartile) were selected for the analysis. We were unable to reconstruct the thymus and lung networks because the RNA-Seq profiles lack the samples. Then, the BioInt libraries were defined from the predicted networks using the same parameters as for human. We applied six clustering methods available in igraph R package (walktrap, fast greedy, louvain, spinglass, infomap and, leading eigenvalue) in experimental (SILAM) and predicted (APID) networks. In the case of APID networks, only the 'spinglass' method was evaluated as it yielded clusterings with higher modularities for all the tissues, while in SILAM networks we found variable modularity coefficients. We used Wallace coefficient [29] to evaluate the clustering agreement between clusters in SILAM against clusters in APID or BioInt units (Supplementary Fig. 1).

4.2. CORUM protein complex intersection

The molecular machines described at the CORUM repository [14] were used as gold standard to evaluate the ability of BioInt-U method to identify already established protein functional complexes. Curated 'core' complexes were retrieved from CORUM database in March 2021. The dataset was filtered to only evaluate CORUM complexes including at least 3 distinct proteins. In order to assess the protein overlap between CORUM and BioInt-U, we first combined TS BioInt libraries into a unified version. BioInt units annotated to the same GO-BP term along different tissues were collapsed to include all TS proteins. The full description of unified and TS BioInt units is available in Supplementary Table 4. The average size of the BioInt units was >17 times larger than CORUM protein complexes. Due to the wide difference in size, the overlap analysis between CORUM complexes and BioInt units was performed applying Simpson's similarity (*SS*) coefficient, defined as the number of common elements between two sets, divided by the minimum set size (complete analysis available in Supplementary Table 3).

4.3. Disease-gene association

Disease gene (DG) associations were retrieved from the DisGeNET repository in December 2020 [28]. Disease references annotated as 'Symptom', 'Finding, 'Injury or poisoning' and 'Individual Behavior' were discarded. DGs with a confidence score lower than 0.1 were discarded. Only diseases including 10 to 200 genes were evaluated. Similar to functional enrichment, disease list was simplified by collapsing terms with more than 0.9 Jaccard's similarity coefficient. When diseases are collapsed, the new disease group includes all genes associated to each pathology but is assigned to the disease description with fewest characters. In order to evaluate the performance of the BioInt-U framework, we created a list of 463 diseases with 11 presumable tissue-specific phenotypes. To generate the TS disease list, we used the same text mining approach as for the functional classification of BioInt units. The DG list and disease classification is available in Supplementary Table 5.

4.4. Gene expression profiles from public repositories

Two independent RNA-Seq transcriptomic profiles characterizing gene expression changes in samples derived from patients affected with psoriasis (GSE166388) and idiophatic pulmonary fibrosis (GSE24206) were downloaded from the Gene Expression OmniBus [5]. Differential gene expression analysis was performed using the GEO2R tool available through the GEO platform. Transcripts with fold change (FC) values of $|\log_2 FC|>2$ and $|\log_2 FC|>1.5$ and p-value < 0.05 were selected as differentially expressed genes (DEg) in fibrosis and psoriasis datasets, respectively. The DEg datasets were mapped in the BioInt units to calculate the % of DEg by BioInt unit. Then, the BioInt units including DEg above the 3rd Quartile (0.9% and 1.3%) were considered the most potentially altered functional processes in fibrosis and psoriasis profiles, respectively.

4.5. BioInt-U method and output availability

All the analyses presented in this work were performed in R studio environment and figures were generated using *ggplot2* and *ComplexHeatmap* R packages [18,32,34,42]. The framework can be employed for other species and only requires PPI and TS transcriptomic data. The R functions necessary to generate additional BioInt units are available in Github repository **https://github.com/Gama-PintoLab/BioInt-U.**

CRediT authorship contribution statement

Marina L. García-Vaquero: Conceptualization, Investigation, Visualization, Data curation, Writing – original draft. **Margarida Gama-Carvalho:** Supervision, Writing – review & editing, Funding acquisition. **Francisco R. Pinto:** Supervision, Writing – review & editing, Resources. **Javier De Las Rivas:** Supervision, Writing – review & editing, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

Work in MGC and FRP research group is supported by UIDB/04046/2020 and UIDP/04046/2020 Centre grants from FCT, Portugal (to BioISI). MGV is a recipient of a fellowship from BioSys-PhD programme (Ref PD/BD/128109/2016) from FCT (Portugal). Work in JDR research group is supported by Instituto de Salud Carlos III (ISCiii, Spanish Ministerio de Salud) grants of reference AC14/00024 and PI18/00591 co-funded by the Fondo Europeo de Desarrollo Regional (FEDER). This work is part of an EU Joint Programme – Neurodegenerative Disease Research (JPND) project with the acronym 'Fly-SMALS'. The project is supported through the following funding organizations under the aegis of JPND – www.jpnd.eu: France, Agence Nationale de la Recherche; Germany, Bundesministerium für Bildung und Forschung (BMBF, FKZ) Portugal, Fundação para a Ciência e a Tecnologia and Spain, Instituto de Salud Carlos III (ISCIII).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.07.006.

References

[1] Agrawal M, Zitnik M, Leskovec J. Large-scale analysis of disease pathways in the human interactome. Pacific Symp. Biocomput. vol. 23. NIH Public Access. 2018. 111–22. https://doi.org/10.1142/9789813235533_0011.

[2] Di A-L, Campos-Laborie FJ, Gutiérrez MA, Lambourne L, Calderwood MA, Vidal M, et al. APID database: Redefining protein-protein interaction experimental evidences and binary interactomes. Database 2019;2019. https://doi.org/10.1093/database/baz005.

[3] Balzani VV, Credi A, Raymo FM, Stoddart JF. Artificial Molecular Machines. Angew Chem Int Ed Engl 2000;39:3348–91. https://doi.org/10.1002/1521-3773(20001002)39:19<3348::aid-anie3348>3.0.co;2-x.

[4] Barabási A, Gulbahce N, Loscalzo J. Network Medicine: A Network-based approach to human disease. Nat Rev Genet 2011;12:56–68. https://doi.org/10.1038/nrg2918.Network.

[5] Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update D995. Nucleic Acids Res 2013;41:D991. https://doi.org/10.1093/NAR/GKS1193.

[6] Barshir R, Shwartz O, Smoly IY, Yeger-Lotem E. Comparative Analysis of Human Tissue Interactomes Reveals Factors Leading to Tissue-Specific Manifestation of Hereditary Diseases. PLOS Comput Biol 2014;10:. https://doi.org/10.1371/JOURNAL.PCBI.1003632e1003632.

[7] Bossi A, Lehner B. Tissue specificity and the human protein interaction network. Mol Syst Biol 2009;5:260. https://doi.org/10.1038/msb.2009.17.

[8] Chapple CE, Robisson B, Spinelli L, Guien C, Becker E, Brun C. Extreme multifunctional proteins identified from a human protein interaction network. Nat Commun 2015;6:7412. https://doi.org/10.1038/ncomms8412.

[9] Csárdi G, Nepusz T. The igraph software package for complex network research. InterJournal. 2006. Complex Sy. 1695.

[10] Deeds E, Krivine J, Feret J, Danos V, Fontana W. Combinatorial complexity and compositional drift in protein interaction networks. PLoS ONE 2012;7. https://doi.org/10.1371/JOURNAL.PONE.0032032.

[11] Dezső Z, Nikolsky Y, Sviridov E, Shi W, Serebriyskaya T, Dosymbekov D, et al. A comprehensive functional analysis of tissue specificity of human gene expression. BMC Biol 2008;6:1–15. https://doi.org/10.1186/1741-7007-6-49.

[12] Espinosa-Cantú A, Cruz-Bonilla E, Noda-Garcia L, DeLuna A. Multiple Forms of Multifunctional Proteins in Health and Disease. Front Cell Dev Biol 2020;8:451. https://doi.org/10.3389/fcell.2020.00451.

[13] Ghiassian SD, Menche J, Barabási A-L. A DIseAse MOdule Detection (DIAMOnD) Algorithm Derived from a Systematic Analysis of Connectivity Patterns of Disease Proteins in the Human Interactome. PLOS Comput Biol 2015;11:. https://doi.org/10.1371/JOURNAL.PCBI.1004120e1004120.

[14] Giurgiu M, Reinhard J, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, et al. CORUM: the comprehensive resource of mammalian protein complexes-2019 D563. Nucleic Acids Res 2019;47:D559. https://doi.org/10.1093/NAR/GKY973.

[15] Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabási A-L. The human disease network. PNAS 2007;104:8685–90. https://doi.org/10.1073/pnas.0701361104.

[16] Greene CS, Krishnan A, Wong AK, Ricciotti E, Zelaya RA, Himmelstein DS, et al. Understanding multicellular function and disease with human tissue-specific networks. Nat Genet 2015;47:569–76. https://doi.org/10.1038/ng.3259.

[17] Grote S. GOfuncR: Gene ontology enrichment using FUNC. 2020.

[18] Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics 2016.

[19] Hekselman I, Yeger-Lotem E. Mechanisms of tissue and cell-type specificity in heritable traits and diseases. Nat Rev Genet 2020;21:137–50. https://doi.org/10.1038/s41576-019-0200-9.

[20] Huttlin E, Bruckner R, Navarrete-Perea J, Cannon J, Baltier K, Gebreab F, et al. Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. Cell 2021;184:3022–3040.e28. https://doi.org/10.1016/J.CELL.2021.04.011.

[21] Lage K, Hansen N, Karlberg E, Eklund A, Roque F, Donahoe P, et al. A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. Proc Natl Acad Sci U S A 2008;105:20870–5. https://doi.org/10.1073/PNAS.0810772105.

[22] Lee CE, Singleton KS, Wallin M, Faundez V. Rare Genetic Diseases. Nature's Experiments on Human Development IScience 2020;23. https://doi.org/10.1016/J.ISCI.2020.101123.

[23] Lin W-H, Liu W-C, Hwang M-J. Topological and organizational properties of the products of house-keeping and tissue-specific genes in protein-protein interaction networks. BMC Syst Biol 2009;3. https://doi.org/10.1186/1752-0509-3-32.

[24] Liu C, Ma Y, Zhao J, Nussinov R, Zhang YC, Cheng F, et al. Computational network biology: Data, models, and applications. Phys Rep 2020;846:1–66. https://doi.org/10.1016/j.physrep.2019.12.004.

[25] Low TY, Syafruddin SE, Mohtar MA, Vellaichamy A, Rahman A, NS, Pung YF, et al. Recent progress in mass spectrometry-based strategies for elucidating protein–protein interactions. Cell Mol Life Sci 2021;78:5325–39. https://doi.org/10.1007/s00018-021-03856-0.

[26] Meltzer EB, Barry WT, D'Amico TA, Davis RD, Lin SS, Onaitis MW, et al. Bayesian probit regression model for the diagnosis of pulmonary fibrosis: Proof-of-principle. BMC Med Genomics 2011;4. https://doi.org/10.1186/1755-8794-4-70.

[27] Nadeau R, Byvsheva A, Lavallée-Adam M. PIGNON: a protein–protein interaction-guided functional enrichment analysis for quantitative proteomics. BMC Bioinf 2021;22:1–22. https://doi.org/10.1186/s12859-021-04042-6.

[28] Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, Ronzano F, Centeno E, Sanz F, et al. The DisGeNET knowledge platform for disease genomics: 2019 update D855. Nucleic Acids Res 2020;48:D845. https://doi.org/10.1093/NAR/GKZ1021.

[29] Pinto FR, Melo-Cristino J, Ramirez M. A confidence interval for the wallace coefficient of concordance and its application to microbial typing methods. PLoS ONE 2008;3. https://doi.org/10.1371/journal.pone.0003696.

[30] Podder S, Mukhopadhyay P, Ghosh TC. Multifunctionality dominantly determines the rate of human housekeeping and tissue specific interacting protein evolution. Gene 2009;439:11–6. https://doi.org/10.1016/j.gene.2009.03.005.

[31] Qiu X, Zheng L, Liu X, Hong D, He M, Tang Z, et al. ULK1 Inhibition as a Targeted Therapeutic Strategy for Psoriasis by Regulating Keratinocytes and Their Crosstalk With Neutrophils. Front Immunol 2021;12:3096. https://doi.org/10.3389/fimmu.2021.714274.

[32] R Core Team. R: A Language and Environment for Statistical Computing. 2020.

[33] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015;43:e47. https://doi.org/10.1093/nar/gkv007.

[34] RStudio Team. RStudio: Integrated Development for R. Inc, Boston, MA 2016.

[35] Sharma P, Alizadeh J, Juarez M, Samali A, Halayko AJ, Kenyon NJ, et al. Autophagy, apoptosis, the unfolded protein response, and lung function in idiopathic pulmonary fibrosis. Cells 2021;10. https://doi.org/10.3390/cells10071642.

[36] Skinnider MA, Scott NE, Prudova A, Kerr CH, Stoynov N, Stacey RG, et al. An atlas of protein-protein interactions across mouse tissues. Cell 2021;184:4073–4089.e17. https://doi.org/10.1016/j.cell.2021.06.003.

[37] Söllner JF, Leparc G, Hildebrandt T, Klein H, Thomas L, Stupka E, et al. An RNA-Seq atlas of gene expression in mouse and rat normal tissues. Sci Data 2017;4. https://doi.org/10.1038/sdata.2017.185.

[38] Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Tissue-based map of the human proteome. Science (80). 2015. 347. https://doi.org/10.1126/science.1260419.

[39] Wang WM, Jin HZ. Heat shock proteins and psoriasis. Eur J Dermatology 2019;29:121–5. https://doi.org/10.1684/ejd.2019.3526.

[40] Wang X, Wei X, Thijssen B, Das J, Lipkin S, Yu H. Three-dimensional reconstruction of protein networks provides insight into human genetic disease. Nat Biotechnol 2012;30:159–64. https://doi.org/10.1038/NBT.2106.

[41] Wang Z, Zhang J. In search of the biological significance of modular structures in protein networks. PLoS Comput Biol 2007;3:1011–21. https://doi.org/10.1371/journal.pcbi.0030107.

[42] Wickham H. ggplot2: Elegant Graphics for Data Analysis. 2016.

[43] Yeger-Lotem E, Sharan R. Human protein interaction networks across tissues and diseases. Front Genet 2015;257. https://doi.org/10.3389/FGENE.2015.00257.

[44] Zhong Q, Simonis N, Li Q, Charloteaux B, Heuze F, Klitgord N, et al. Edgetic perturbation models of human inherited disorders. Mol Syst Biol 2009;5. https://doi.org/10.1038/MSB.2009.80.