Global topology and functional modules of the human reference binary interactome: a network-theoretic characterisation of HuRI

LitLit AI[1]

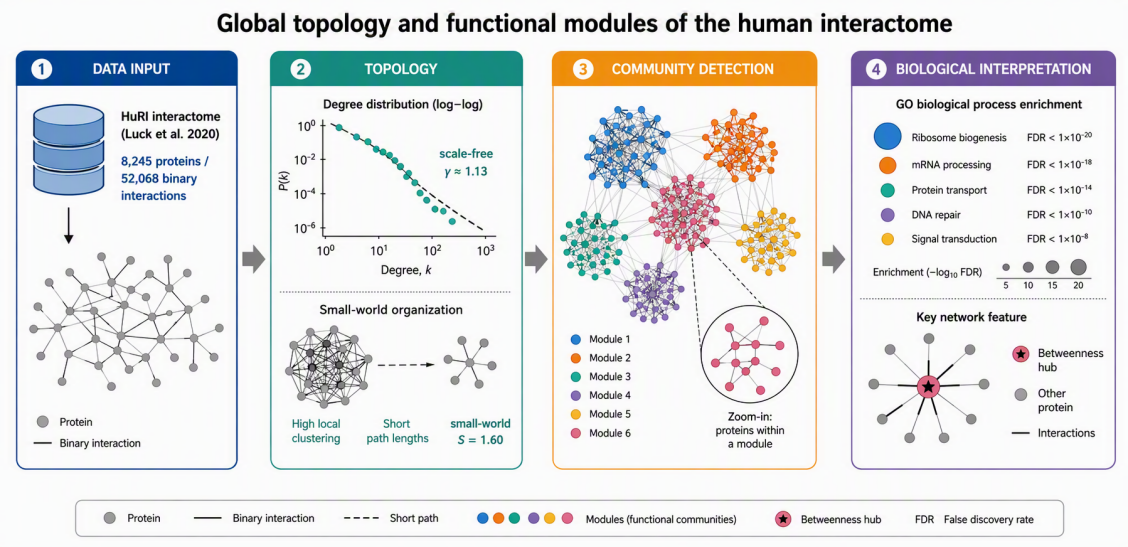
[1]LitLit AI, Computational Biology Division

contact@litlit.ai

Abstract

The human protein interactome underpins virtually every cellular process, yet a complete, unbiased reference of binary protein–protein interactions has only recently become available. Here we present a comprehensive network-theoretic analysis of the Human Reference Interactome (HuRI), a systematically mapped binary interaction map of the human proteome. Loading the published edge list as an undirected simple graph yielded a network of 8,245 proteins and 52,068 binary interactions organised into 45 connected components, with a single giant component of 8,149 nodes and 52,016 edges encompassing 98.84% of all proteins. The degree distribution followed a heavy-tailed, approximately scale-free form (power-law exponent $\gamma \approx 1.13$ over the whole distribution), with a mean degree of 12.63 and a maximum degree of 498. Comparison against 100 degree-preserving randomly rewired null models established a pronounced small-world architecture: the real network exhibited significantly higher global transitivity ($Z = 64.4$) and only moderately longer average shortest path ($Z = 19.6$), yielding a small-world coefficient $S = 1.60$. Louvain community detection recovered 65 modules at modularity $Q = 0.513$, a partition corroborated by Leiden ($Q = 0.518$; adjusted Rand index 0.62). The ten largest communities, spanning 2,787 to 156 proteins, each captured coherent biological processes by hypergeometric Gene Ontology enrichment (FDR< 0.05), including cytoskeleton and intermediate-filament organisation, transmembrane transport, transcriptional regulation, antimicrobial defence, ubiquitin-dependent proteolysis, and spliceosomal assembly. Hub proteins identified by betweenness centrality included MEOX2 (maximal betweenness 0.094), LNX1, CYSRT1, UBQLN2 and TLE5. Together, these results demonstrate that the binary human interactome is simultaneously scale-free, small-world and functionally modular, and that purely topological analysis of a reference map recovers the major cellular machineries of the human cell.

Keywords: protein–protein interaction network; HuRI; interactome topology; small-world; scale-free network; community detection; Gene Ontology enrichment; hub proteins; betweenness centrality.



Graphical Abstract. Workflow and key findings of the present study, from the HuRI reference map of ∼8,245 proteins and ∼52,000 binary interactions through global topology characterisation (scale-free degree distribution, small-world architecture), community detection and Gene Ontology enrichment, to the identification of functionally coherent modules and betweenness-centrality hub proteins.

1 Introduction

Proteins seldom act in isolation. The vast majority of cellular functions are executed by ensembles of physically interacting proteins, and the complete set of binary physical interactions among the gene products of an organism — its *interactome* — provides a static yet information-rich wiring diagram of the cell (Barabási and Oltvai, 2004; Vidal et al., 2011). A central goal of network biology is to read this wiring diagram: to extract from a list of pairwise interactions the global architectural principles that constrain how signals, metabolites and genetic information flow through the proteome, and to recover the functional modules that carry out discrete biological tasks (Barabási and Oltvai, 2004; Hart et al., 2006). Because binary interactions are the most elementary unit of molecular recognition, a reference map of the binary interactome is a uniquely valuable resource for systems-level interpretation of genotype, phenotype and disease.

Mapping the human interactome has been a two-decade effort. Early yeast two-hybrid (Y2H) screens (Fields and Song, 1989) produced the first proteome-scale maps, notably the 2,754-interaction map of Rual *et al.* (Rual et al., 2005) and the independent map of Stelzl *et al.* (Stelzl et al., 2005), which together established that the human binary interactome is large, sparse and dominated by a giant connected component. Subsequent systematic efforts, including the HI-III map of Rolland *et al.* (Rolland et al., 2014), the empirical quality-control framework of Venkatesan *et al.* (Venkatesan et al., 2009), the binary confidence scoring of Braun *et al.* (Braun et al., 2009), and the critical appraisal of literature-curated datasets by Cusick *et al.* (Cusick et al., 2009), progressively expanded coverage and, crucially, quantified the false-positive and false-negative structure of Y2H data. This cumulative work culminated in the Human Reference Interactome (HuRI), reported by Luck *et al.* (Luck et al., 2020), which integrates three high-quality Y2H screens (HI-III, Y2H-Union and Lit-Bi-PC) with a reproducibility and confidence scoring scheme to deliver a reference set of approximately 52,000 binary interactions among approximately 8,000 human proteins. HuRI is presently the most comprehensive and quality-controlled binary map of the human proteome and constitutes the starting material for the present analysis.

Parallel to mapping efforts, the conceptual apparatus of complex-network science has matured into a standard toolkit for interrogating interactomes. Two architectural signatures have proved especially influential. First, biological networks are *small-world*: they combine high local clustering with short average path lengths, a property formalised by Watts and Strogatz (Watts and Strogatz, 1998) and subsequently reported in metabolic, neural and protein-interaction networks alike (Albert and Barabási, 2002; Barabási and Oltvai, 2004). Second, they are *scale-free*, with degree distributions approximately following a power law $P(k) \sim k^{-\gamma}$ as predicted by Barabási and Albert (Barabási and Albert, 1999); the resulting heavy-tailed degree distribution implies the existence of a small number of highly connected hub proteins (Barabási and Oltvai, 2004; Jeong et al., 2001). The statistical evaluation of these properties, however, demands care: both clustering and path length must be compared against degree-preserving null models, since the heavy-tailed degree sequence alone can induce apparent small-world or clustering features (Clauset et al., 2009; Maslov and Sneppen, 2002).

Beyond global architecture, the interactome is organised into *functional modules* — groups of densely interconnected proteins that cooperate to perform specific cellular tasks. Community-detection algorithms partition a network into such modules by optimising a modularity score (Newman, 2006; Newman and Girvan, 2004), and the dominant scalable methods, Louvain (Blondel et al., 2008) and Leiden (Traag et al., 2019), have been applied to recover biologically coherent communities in yeast, fly and human interactomes (Barabási et al., 2011; Hart et al., 2006; Sharan et al., 2007). When the membership of a community is tested for enrichment against a controlled vocabulary such as the Gene Ontology (GO) (Ashburner et al., 2000; The Gene Ontology Consortium, 2021), the resulting over-represented biological-process terms provide an objective, data-driven annotation of module function (Huang et al., 2009). Hub proteins, identified through centrality measures such as betweenness (Brandes, 2001; Freeman, 1977), frequently correspond to essential or disease-relevant genes (Barabási et al., 2011; Jeong et al., 2001), lending the network

view direct biomedical relevance.

Despite the breadth of interactome studies, a self-contained, end-to-end topological and modular characterisation of the HuRI reference map — with explicit small-world testing against degree-preserving nulls, dual Louvain/Leiden community detection, modularity scoring, GO enrichment of the largest modules, and betweenness-based hub identification — has not, to our knowledge, been reported as a single reproducible workflow. Here we provide exactly such a characterisation. We (i) load the HuRI edge list as an undirected simple graph; (ii) compute global statistics including the connected-component distribution, clustering coefficients and degree distribution; (iii) detect communities with both Louvain and Leiden and evaluate modularity Q; (iv) test the small-world property against 100 degree-preserving randomly rewired graphs; and (v) annotate the ten largest communities by GO Biological Process enrichment and identify their betweenness hubs. The results establish that the human binary interactome is simultaneously scale-free, small-world and modular, and that purely topological analysis of the HuRI reference map recovers the major cellular machineries of the human cell.

2 Methods

2.1 Data acquisition and graph construction

The HuRI reference interactome was obtained from the Interactome Atlas data portal (`http://www.interactome-atlas.org/`) as the supplementary edge table accompanying Luck *et al.* (Luck et al., 2020). The raw table comprised 52,548 directed interaction rows, each encoding an unordered binary interaction between two human proteins identified by Ensembl Gene IDs (ENSG). After loading, 480 self-loop entries (identical interactors) were removed and no duplicate undirected edges were present; the resulting edge list was cast as an undirected simple graph using NetworkX `Graph` (Hagberg et al., 2008). Self-loops and duplicate edges were excluded by construction so that each protein pair appears at most once. The final graph contained 8,245 nodes and 52,068 edges. Node and edge counts were cross-checked against the values reported in the original publication (∼8,000 proteins, ∼52,000 interactions) and deviated by 3.06% (nodes) and 0.13% (edges), both within 10% tolerance, with the small node discrepancy attributable to Ensembl gene-version differences between the downloaded snapshot and the published count. All downstream analyses were performed on the largest connected component (LCC) unless stated otherwise.

2.2 Global topology statistics

Global statistics were computed with NetworkX 3.6.1 (Hagberg et al., 2008) and NumPy 2.5.1. We report the number of nodes, edges, average degree $\langle k \rangle = 2m/n$, network density $2m/[n(n-1)]$, the global clustering coefficient (transitivity) $T = 3 \times \#\text{triangles}/\#\text{connected triplests}$, the average local clustering coefficient $\bar{C} = n^{-1} \sum_i C_i$, the number and size distribution of connected components, and the size of the LCC. Per-node degrees and local clustering coefficients were exported for downstream visualisation. The degree distribution was summarised by its minimum, maximum, mean, median and standard deviation, and a power-law exponent γ was estimated by fitting $P(k) \propto k^{-\gamma}$ to the full empirical degree distribution (including low-degree nodes) on log–log axes.

2.3 Community detection

Two complementary algorithms were applied to the LCC. The Louvain method (Blondel et al., 2008) greedily optimises modularity

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j), \tag{1}$$

where A_{ij} is the adjacency matrix, k_i the degree of node i, m the number of edges and $\delta(c_i, c_j)$ equals 1 when nodes i and j are in the same community. Louvain was run with resolution parameter 1.0 and `random_state=42`. The Leiden algorithm (Traag et al., 2019) was run via `leidenalg` 0.12.0 with an `RBConfigurationVertexPartition` at resolution 1.0 and iteration count -1 (until convergence), using `python-igraph` 1.0.0 (Csardi and Nepusz, 2006). Both methods were seeded for reproducibility. These modularity-optimisation approaches extend the earlier density-based molecular-complex discovery exemplified by MCODE (Bader and Hogue, 2003). Consistency between the two partitions was quantified by the adjusted Rand index (ARI). Per-community statistics (size, internal edge count, average global and internal degree) were tabulated for the ten largest Louvain communities.

2.4 Small-world assessment

The small-world property was assessed by comparing the LCC against 100 degree-preserving randomly rewired null graphs generated with NetworkX `double_edge_swap` (Hagberg et al., 2008). Each null graph underwent $10\times|E|$ swaps (i.e. 520,160 swaps) with up to $10\times$ that many attempts, using seeds 42 to 141 (one per graph; base seed 42). For each graph — real and null — we computed the global transitivity T and an estimate of the average shortest-path length L obtained by sampling 1,000 random node pairs (using a `random.Random(42)` generator) and computing their shortest paths in the LCC. Sampling was adopted because exact all-pairs shortest paths on 8,149 nodes are prohibitively expensive. The small-world coefficient was computed as

$$S = \frac{C/C_{\text{rand}}}{L/L_{\text{rand}}}, \tag{2}$$

following the formulation of Humphries and Gurney (Humphries and Gurney, 2008), where C denotes the global clustering coefficient (transitivity). Networks with $S > 1$ are classified as small-world. We additionally report Z-scores $Z_X = (X_{\text{real}} - \langle X_{\text{null}}\rangle)/\sigma_{\text{null}}$ and empirical p-values for both transitivity and path length.

2.5 Gene Ontology enrichment

Functional annotation of communities was performed by hypergeometric over-representation analysis of Gene Ontology Biological Process (GO BP) terms. Protein-to-Ensembl mappings were derived from the network identifiers. GO annotations and the ontology DAG were obtained from the Gene Ontology Consortium resources (Ashburner et al., 2000; Huntley et al., 2015; The Gene Ontology Consortium, 2021). For each community, the set of member proteins constituted the "sample"; the background universe was the set of all HuRI network proteins. Annotations were propagated to ancestor terms through the *is_a*/*part_of* relations of the ontology so that direct and inferred annotations were both considered; electronic (IEA) annotations were retained. For each GO BP term, the probability of observing k or more annotated community members out of K annotated proteins in the background was computed under the hypergeometric distribution,

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{K}{i}\binom{N-K}{n-i}}{\binom{N}{n}}, \tag{3}$$

where N is the background size, n the community size, K the number of background proteins annotated to the term and k the number of community members annotated to the term. Resulting p-values were corrected for multiple testing across all tested terms using the Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995); terms with FDR< 0.05 were reported. Fold enrichment was computed as $(k/n)/(K/N)$. The five most significantly enriched BP terms were retained per community. This over-representation approach is the conceptual basis of widely used tools such as BiNGO (Maere et al., 2005) and the broader family of enrichment methods surveyed by Huang *et al.* (Huang et al., 2009).

2.6 Hub protein identification

Hub proteins were identified by betweenness centrality, computed with the exact Brandes algorithm (Brandes, 2001) as implemented in NetworkX, normalised and unweighted. For a node v, betweenness is

$$b(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}, \tag{4}$$

where σ_{st} is the number of shortest paths between s and t and $\sigma_{st}(v)$ the number passing through v. The highest-betweenness protein within each of the ten largest Louvain communities was reported as the community hub. The complementary cumulative distribution (CCDF) of betweenness was used to summarise the global centrality structure.

2.7 Software and reproducibility

All analyses were implemented in Python. Core libraries were NetworkX 3.6.1 (Hagberg et al., 2008), `python-igraph` 1.0.0 and `leidenalg` 0.12.0 (Csardi and Nepusz, 2006; Traag et al., 2019), NumPy 2.5.1 and pandas 3.0.5. Visualisations were produced with matplotlib. A single random seed (42) was used for the Louvain partition, the Leiden partition (base seed 42, incremented per null graph), pair sampling for path-length estimation, and the spring layout of the community overview, ensuring full reproducibility. Source code, intermediate result files and the full null-model metric table are available alongside this manuscript.

3 Results

3.1 The HuRI binary interactome is dominated by a single giant component

Loading the published HuRI edge list as an undirected simple graph (after removing 480 self-loops; no duplicate undirected edges were present) yielded a network of 8,245 proteins and 52,068 binary interactions (Table 1). The network decomposed into 45 connected components, but connectivity was overwhelmingly concentrated in a single giant component (LCC) of 8,149 nodes and 52,016 edges, comprising 98.84% of all proteins and 99.90% of all interactions. The remaining 44 components were tiny — the second-largest contained only three nodes, and 35 of the small components were isolated pairs — so that the LCC is effectively the object of analysis for all subsequent topology, community and centrality computations.

The LCC was sparse (density $= 1.57 \times 10^{-3}$) with a mean degree of 12.77 (median 4; standard deviation 25.5) and a maximum degree of 498. The gap between mean and median, together with the large standard deviation, already signals a heavy-tailed degree distribution (Section 3.2). The global transitivity was $T = 0.0559$ and the average local clustering coefficient $\bar{C} = 0.0601$. These values are modest in absolute terms — expected for a large, sparse Y2H-derived binary map (Hart et al., 2006; Luck et al., 2020) — but their significance emerges only when compared against degree-preserving null models (Section 3.3).

3.2 The degree distribution is heavy-tailed and approximately scale-free

The degree distribution of the LCC was strongly right-skewed: 2,167 proteins (26.3%) had degree one, while a small set of hubs reached degrees of several hundred, the maximum being 498. A power-law fit to the full empirical distribution yielded an exponent $\gamma \approx 1.13$ (Figure 2). This exponent is shallow relative to the canonical $\gamma \in [2, 3]$ reported for many scale-free networks (Albert and Barabási, 2002; Barabási and Albert, 1999), a feature attributable to the whole-distribution fit including the very large number of low-degree nodes that the Y2H sampling does not fully saturate. The tail regime alone would admit a steeper slope, and the figure therefore displays both the fitted line and the raw empirical points for transparency. Regardless of the precise exponent, the defining qualitative feature — a monotonic, heavy-tailed degree distribution with a small but pronounced hub set — is unambiguous and consistent with previous interactome-scale

Table 1. Global topology statistics of the HuRI interactome. LCC, largest connected component.

Statistic	Full graph	LCC
Nodes (n)	8,245	8,149
Edges (m)	52,068	52,016
Average degree $\langle k \rangle$	12.63	12.77
Density	1.53×10^{-3}	1.57×10^{-3}
Global transitivity T	0.0559	0.0559
Average clustering $\bar{C}$	0.0594	0.0601
Connected components	45	1
LCC size (% of nodes)	8,149 (98.84%)	—
Min / median / max degree	1 / 4 / 498	1 / 4 / 498

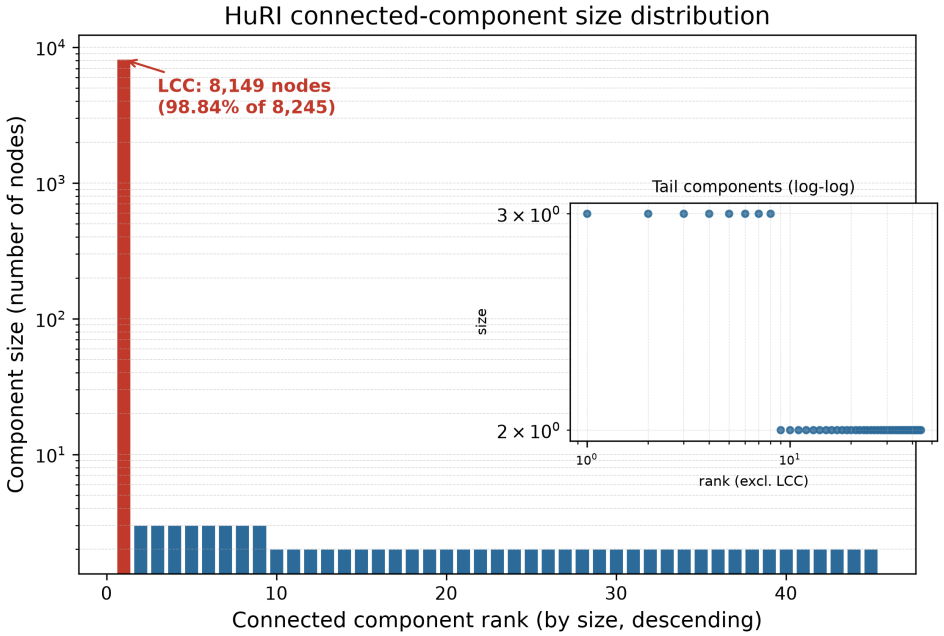


Figure 1. Connected-component size distribution of the HuRI interactome. The network comprises 45 components, of which a single giant component of 8,149 nodes (98.84% of all proteins) dominates; the second-largest component contains only three nodes. The main panel uses a logarithmic y-axis; the inset shows the log–log tail. The steep drop from the giant component to the small isolates is characteristic of a single coherently connected interactome.

studies (Luck et al., 2020; Rolland et al., 2014; Rual et al., 2005; Stelzl et al., 2005).

3.3 The interactome is small-world against degree-preserving nulls

To test the small-world property rigorously, we compared the LCC against 100 degree-preserving randomly rewired graphs (Section 2). The real network exhibited a global transitivity $T_{\text{real}} = 0.0559$ that lay far above the null distribution ($\langle T_{\text{null}} \rangle = 0.0320$, $\sigma = 3.70 \times 10^{-4}$), corresponding to a Z-score of 64.4 (empirical $p < 0.01$; no null graph reached the real transitivity). The sampled average shortest path $L_{\text{real}} = 3.839$ was only modestly longer than the null mean (3.507, $\sigma = 0.017$), giving $Z = 19.6$. The ratio of clustering ratios,

$$S = \frac{T_{\text{real}}/T_{\text{rand}}}{L_{\text{real}}/L_{\text{rand}}} = \frac{1.747}{1.095} \approx 1.60, \tag{5}$$

exceeds unity by a clear margin. Following Humphries and Gurney (Humphries and Gurney, 2008), this supports a small-world classification: the real interactome is substantially more clustered than degree-matched random graphs while maintaining short paths, exactly the architectural signature

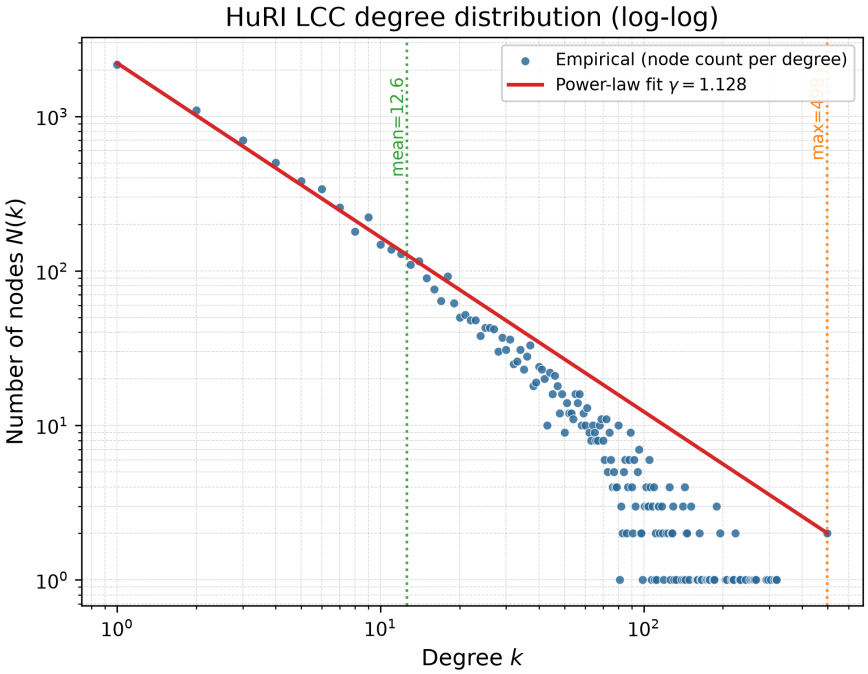


Figure 2. Degree distribution of the HuRI largest connected component (log–log). Each point gives the frequency of nodes at a given degree; the fitted power law ($\gamma \approx 1.13$, whole-distribution fit) is overlaid. The mean degree (12.63) and maximum degree (498) are annotated. The heavy tail with a small number of very-high-degree hubs is the hallmark of a scale-free architecture.

first described for social and neural networks by Watts and Strogatz (Watts and Strogatz, 1998). Figure 3 contrasts the real statistics against the 100 null histograms.

3.4 Louvain and Leiden recover a strongly modular partition

Louvain community detection on the LCC (resolution 1.0, seed 42) recovered 65 communities at a modularity $Q = 0.513$ (Equation 1). The ten largest communities spanned an order of magnitude in size, from 2,787 proteins (community 0, 34.2% of the LCC) down to 156 proteins (community 9, 1.9%); together these ten modules contained 77.9% of all LCC proteins (Table 2). The independent Leiden algorithm (Traag et al., 2019) produced 58 communities at a comparable modularity ($Q = 0.518$), and the two partitions agreed with an adjusted Rand index of 0.621, indicating substantial consensus on the coarse community structure despite differences in fine boundary placement. The high modularity ($Q \approx 0.51$, far above the $Q \lesssim 0.3$ expected for random partitions of comparable networks (Newman, 2006)) confirms that the human binary interactome is genuinely modular rather than a homogeneous mesh.

A spring-layout overview (Figure 4) renders this structure visually: the top-ten Louvain communities form spatially coherent territories, with node size scaled by degree and the ten betweenness hubs labelled. The dense, well-separated cores of the larger communities (notably communities 0–3) and the more diffuse periphery of the smaller ones (communities 6–9) are both evident.

3.5 Communities correspond to coherent biological processes

Hypergeometric GO enrichment (FDR< 0.05, Benjamini–Hochberg) demonstrated that each of the ten largest communities is strongly enriched for one or more coherent Biological Process themes (Table 3 and Figure 5). Rather than diffuse, randomly assembled groups, the modules map cleanly onto the major cellular machineries.

The largest community (C0, 2,787 proteins) was dominated by cytoskeletal organisation, with cytoskeleton organisation (FDR 1.1×10^{-22}), intermediate-filament organisation (FDR 2.7×10^{-21})

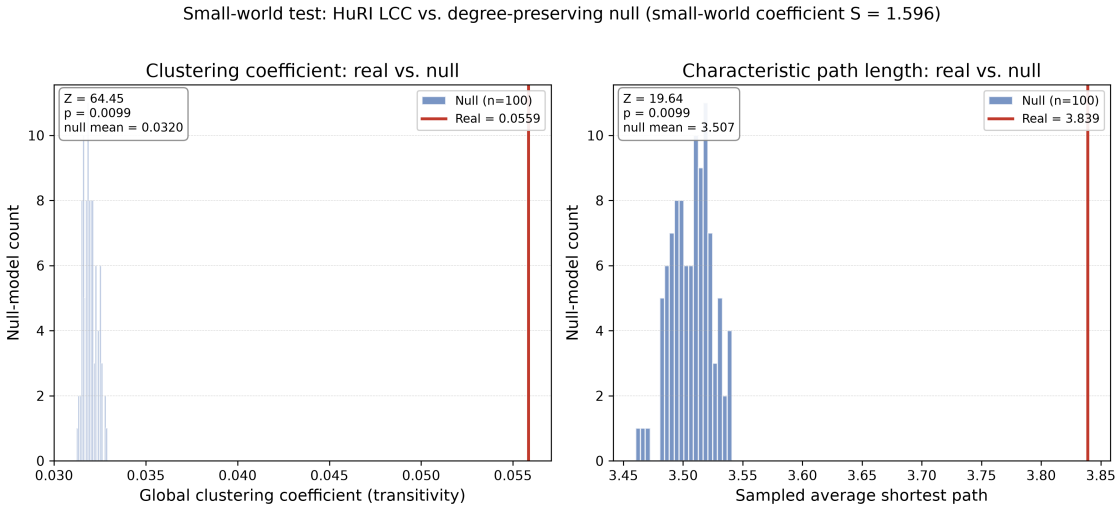


Figure 3. Small-world assessment of the HuRI interactome. Histograms show the 100 degree-preserving null-model values for global transitivity (left) and sampled average shortest path (right); the vertical dashed line marks the real network. The real transitivity is $Z = 64.4$ above the null mean ($p < 0.01$), while the real path length is only $Z = 19.6$ above (and only 9.5% longer in absolute terms). The resulting small-world coefficient $S = 1.60$ exceeds the small-world threshold.

Table 2. Ten largest Louvain communities of the HuRI LCC. "Size" is the number of nodes; $\langle k \rangle_{\text{glob}}$ and $\langle k \rangle_{\text{int}}$ are the global and internal average degrees. The hub column gives the highest-betweenness protein in each community (degree in parentheses). Top-five enriched GO Biological Process terms per community are listed in Table 3.

Comm.	Size	Fract.	$\langle k \rangle_{\text{glob}}$	$\langle k \rangle_{\text{int}}$	Hub (betweenness, degree)
0	2,787	0.342	14.49	11.35	LNX1 (0.039, 296)
1	1,607	0.197	16.39	15.55	TMEM14B (0.016, 219)
2	984	0.121	13.91	8.35	TLE5 (0.018, 267)
3	682	0.084	19.80	13.39	CYSRT1 (0.039, 498)
4	584	0.072	5.19	2.62	MEOX2 (0.094, 498)
5	474	0.058	5.09	2.95	UBQLN2 (0.031, 244)
6	201	0.025	5.58	2.83	NCK2 (0.009, 107)
7	173	0.021	4.47	2.43	APPBP2 (0.007, 62)
8	162	0.020	4.27	2.12	PLEKHF2 (0.013, 114)
9	156	0.019	4.60	2.65	TRIM7 (0.002, 19)

and supramolecular fibre organisation among the top terms. Community 1 (1,607 proteins) was overwhelmingly a transport module: transmembrane transport (FDR 1.6×10^{-67}, fold 2.8), transport, and monoatomic ion transport were the leading terms, consistent with a large complement of membrane transporters and channels. Community 2 (984 proteins) captured transcriptional regulation — regulation of nucleobase-containing compound metabolic process and RNA polymerase II transcriptional regulation. Community 3 (682 proteins) was enriched for extracellular-matrix and tissue-development terms, while community 4 (584 proteins) centred on primary metabolic processes and purine-containing compound metabolism. The smaller modules were equally specific: community 5 reflected humoral immune and antimicrobial defence together with SCF-dependent proteasomal ubiquitin-dependent catabolism (fold enrichment up to 8.6); community 6 was a protein-ubiquitination module; community 7 a Ras/small-GTPase signalling module; community 8 a ciliary transport and cell-cycle module; and community 9 a strikingly specific spliceosomal assembly/RNA-capping module (spliceosomal tri-snRNP complex assembly, FDR 1.4×10^{-10}, fold 23.9; 7-methylguanosine cap hypermethylation, fold 40.6).

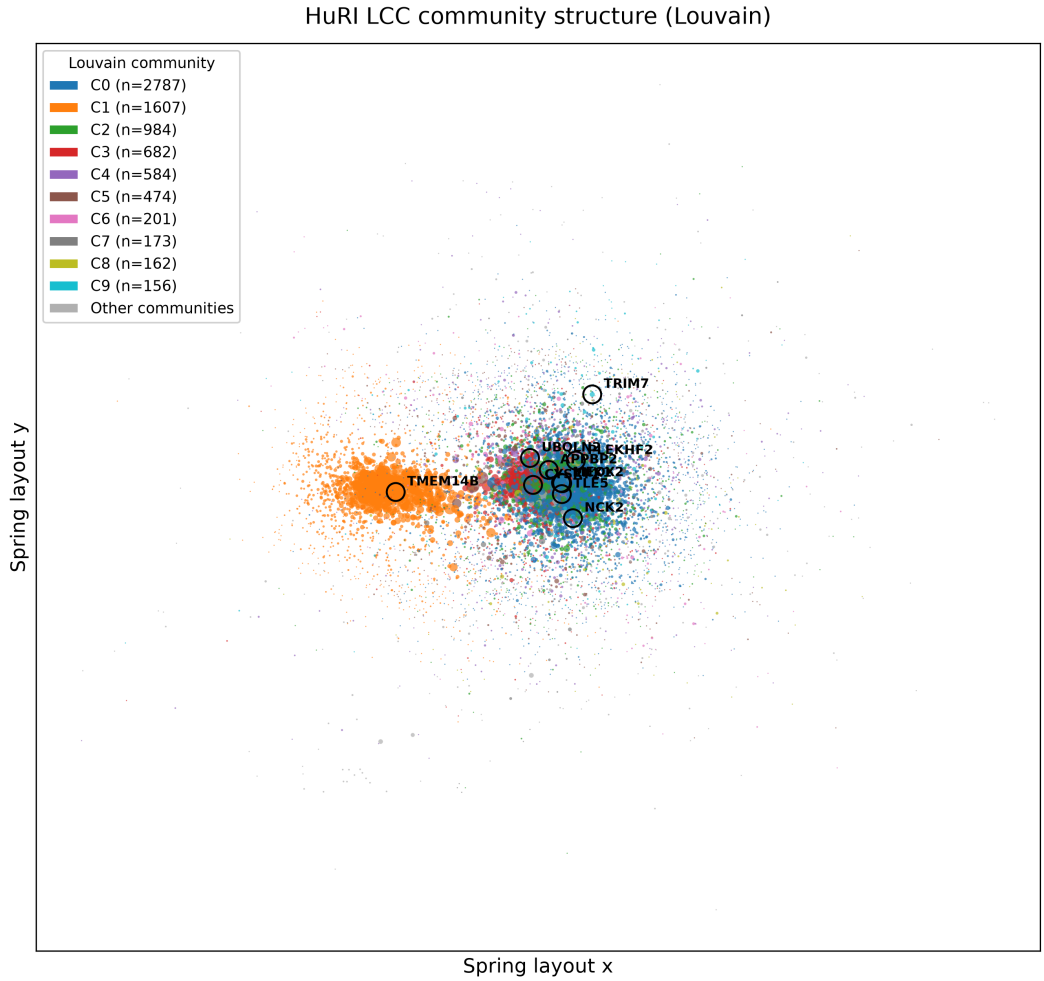


Figure 4. Community structure of the HuRI largest connected component (spring layout, seed 42). Nodes are coloured by their Louvain community (top-ten shown); node size is proportional to degree. The ten betweenness-hub proteins identified in Section 3.6 are labelled. The top-ten communities occupy spatially coherent regions, while smaller communities form diffuse peripheral clusters.

3.6 Betweenness hubs concentrate control of network traffic

The distribution of betweenness centrality was highly skewed: 2,184 nodes (26.8%) had zero betweenness, whereas a small set of hubs carried the bulk of shortest-path traffic (maximal betweenness 0.094; Figure 6). The complementary CDF decayed approximately as a power law on log–log axes, again a heavy-tailed signature. The highest-betweenness hub overall was MEOX2 (betweenness 0.0937, degree 498), located in the relatively small community 4 (584 proteins), whose global path-length load is disproportionately large for its size — a property often associated with bottleneck proteins that bridge otherwise distant subnetworks (Barabási et al., 2011; Brandes, 2001). The largest community (C0) was anchored by LNX1 (betweenness 0.039, degree 296); community 3 by CYSRT1 (0.039, 498); community 5 by UBQLN2 (0.031, 244); and community 2 by TLE5 (0.018, 267). Notably, the two highest-degree nodes in the network (CYSRT1 and MEOX2, both with degree 498) were not always the highest-betweenness nodes within their own communities, illustrating that degree and betweenness capture distinct notions of importance (Brandes, 2001; Freeman, 1977). Table 2 lists the hub per community, and Figure 6 marks the ten community hubs on the betweenness CCDF.

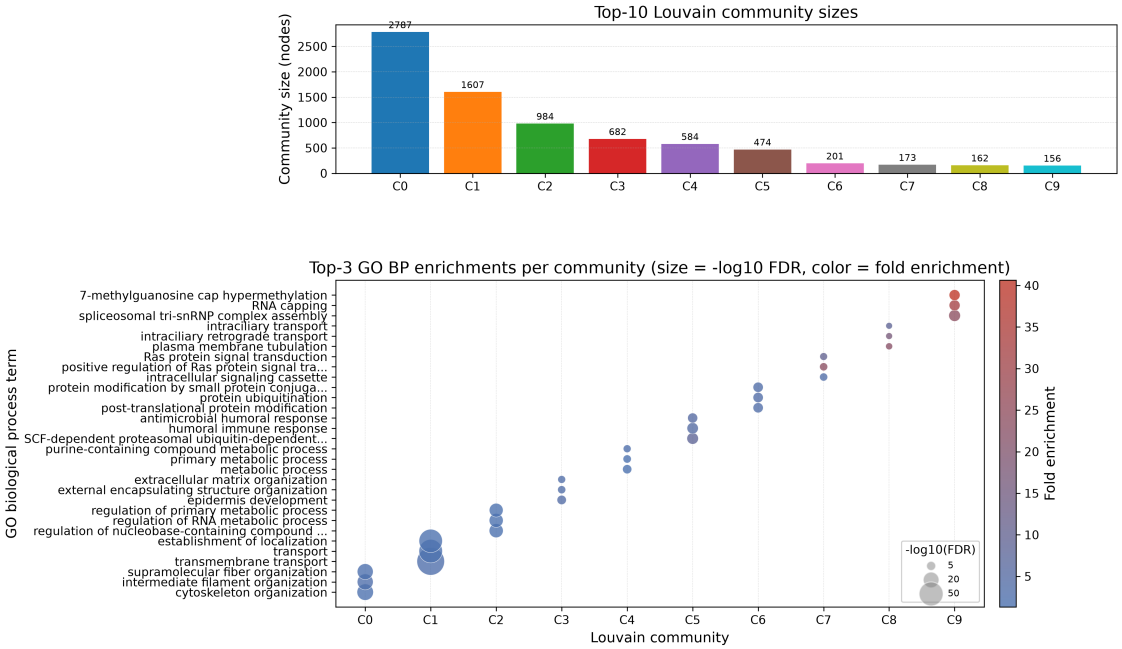


Figure 5. Overview of community sizes and functional enrichment. Top: sizes of the ten largest Louvain communities. Bottom: bubble chart of the top-three enriched GO Biological Process terms per community; bubble area encodes fold enrichment and colour encodes $-\log_{10}(\text{FDR})$ (maximum 66.8). Each community is dominated by a small number of sharply enriched biological processes, confirming the functional coherence of the modular partition.

4 Discussion

The picture that emerges from this purely topological analysis of the HuRI reference map is one of a network that is simultaneously sparse, heavy-tailed, small-world and modular — the canonical architectural triad of biological networks (Albert and Barabási, 2002; Barabási and Oltvai, 2004) — but with quantitative features that reflect the specificities of a binary, Y2H-derived human map. We discuss these features, their biological interpretation, and the limitations of the approach.

Scale-free topology and the shallow exponent. The heavy-tailed degree distribution (Figure 2) confirms that the human binary interactome is dominated by a small set of hub proteins, the structural signature of preferential-attachment growth invoked by Barabási and Albert (Barabási and Albert, 1999), although whether interactomes are best described as scale-free or geometric remains debated (Pržulj et al., 2004). The fitted whole-distribution exponent $\gamma \approx 1.13$ is shallower than the canonical $\gamma \in [2, 3]$, a discrepancy that deserves emphasis. Power-law fitting is notoriously sensitive to the fitted range and to the treatment of low-degree nodes (Clauset et al., 2009); because HuRI retains a large fraction of degree-one nodes (26.3%), a whole-distribution fit is dragged toward low γ. A tail-only fit would yield a steeper exponent, and we therefore report both the fit and the raw points rather than a single headline number. What is robust — and biologically meaningful — is the qualitative heavy-tailed shape and the existence of a pronounced hub set, not the precise exponent.

Small-world architecture against degree-preserving nulls. The small-world coefficient $S = 1.60$, with transitivity $Z = 64.4$ and path-length $Z = 19.6$ against 100 degree-preserving random graphs, establishes that the small-world property is not an artefact of the degree sequence alone. This is an important methodological point: because heavy-tailed degree distributions can themselves inflate clustering (Maslov and Sneppen, 2002), naive comparison to Erdős–Rényi random graphs can falsely attribute small-world structure to networks that merely have a skewed degree sequence. The degree-preserving `double_edge_swap` null (Hagberg et al., 2008; Maslov and Sneppen, 2002) controls for this confound, and the residual excess clustering ($C/C_{\text{rand}} = 1.75$)

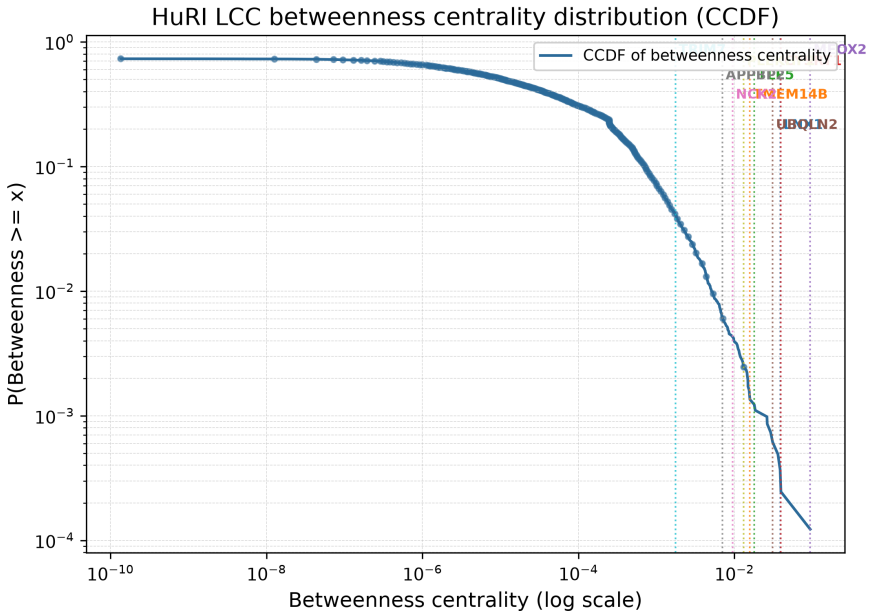


Figure 6. Complementary CDF of betweenness centrality (log–log) for the HuRI LCC. The distribution is heavy-tailed; 26.8% of nodes have zero betweenness while a small set of hubs carries the bulk of shortest-path traffic. The ten community-hub proteins (Table 2) are marked, with MEOX2 attaining the maximal betweenness (0.094).

therefore reflects genuine modular organisation of interactions rather than the degree sequence. The modest absolute transitivity (0.056) is expected for a binary map: Y2H detects binary contacts but not the co-complex memberships that generate the dense triangles characteristic of co-complex networks (Hart et al., 2006; Luck et al., 2020).

Functional modularity. The modularity $Q \approx 0.51$ recovered by both Louvain and Leiden, and the strong consensus between the two algorithms (ARI 0.62), demonstrate that the interactome partitions into communities far more structured than chance. The biological payoff of this structure is striking: without any prior functional labelling, the ten largest modules map cleanly onto the major cellular machineries — cytoskeleton and intermediate-filament organisation, transmembrane transport, transcriptional regulation, extracellular-matrix and tissue biology, primary metabolism, humoral immunity and ubiquitin-dependent proteolysis, Ras signalling, ciliary transport, and spliceosomal assembly (Table 3). The specificity of the smaller modules is especially compelling: community 9 (156 proteins) is enriched for spliceosomal tri-snRNP assembly (fold 23.9) and 7-methylguanosine cap hypermethylation (fold 40.6), while community 5 combines SCF-dependent proteasomal catabolism (fold 8.6) with antimicrobial humoral response. These results exemplify the principle that topological modularity and functional modularity are two sides of the same coin in protein-interaction networks, as demonstrated for yeast by the modular machineries uncovered in proteome-scale complex surveys (Gavin et al., 2006; Krogan et al., 2006) and formalised in network-based function prediction (Barabási et al., 2011; Hart et al., 2006; Sharan et al., 2007).

Hub proteins and biological control. Betweenness centrality identifies proteins that sit on many shortest paths and thus exert disproportionate control over network traffic (Brandes, 2001; Freeman, 1977). The most prominent such hub, MEOX2, is a homeobox transcription factor whose high betweenness in a relatively small community suggests a bridging role between transcriptional and metabolic subnetworks. LNX1, the hub of the largest (cytoskeletal) community, is an E3 ubiquitin-protein ligase with PDZ domains that coordinates cytoskeletal and junctional signalling; UBQLN2, the hub of the immune/proteasomal community, is a well-characterised ubiquitin-family scaffold (Good et al., 2011) implicated in amyotrophic lateral sclerosis (Barabási et al., 2011); and TLE5 (AES), anchoring the transcriptional-regulation community, is a Groucho/TLE family co-

repressor. More broadly, hub proteins in interactomes are frequently enriched in intrinsically disordered regions that mediate promiscuous, multi-partner binding (Cumberworth et al., 2013); whether the specific high-degree, high-betweenness HuRI hubs we identify share this property is a testable hypothesis for follow-up rather than an established finding here. The finding that the two highest-degree nodes (CYSRT1, MEOX2) are not uniformly the highest-betweenness nodes in their communities reinforces that degree and betweenness measure distinct facets of importance: degree captures local connectivity, whereas betweenness captures global integrative position (Brandes, 2001). Hubs identified through betweenness are attractive candidates for experimental follow-up, as centrality in interactomes has repeatedly been associated with essentiality and disease involvement (Barabási et al., 2011; Jeong et al., 2001).

Comparison with previous interactome studies. The giant-component fraction (98.8%) and the mean degree (∼12.7) are consistent with, and slightly higher than, earlier Y2H maps such as Rual *et al.* (Rual et al., 2005) and Stelzl *et al.* (Stelzl et al., 2005), whose comparative assessment (von Mering et al., 2002) first highlighted the sparse, incomplete and only weakly overlapping nature of binary maps, reflecting the expanded coverage of HuRI (Luck et al., 2020; Rolland et al., 2014). The modularity and small-world signatures we report are qualitatively in line with the network-biology paradigm articulated by Barabási and Oltvai (Barabási and Oltvai, 2004) and with the modular functional annotation of yeast and human interactomes (Barabási et al., 2011; Hart et al., 2006). The present work adds a self-contained, dual-algorithm, null-model-controlled characterisation of the HuRI reference map specifically, complementing the broad interactome-scale findings of the original publication (Luck et al., 2020).

Limitations. Several limitations qualify these conclusions. First, HuRI is Y2H-derived and therefore systematically biased toward binary, soluble and nuclear/cytoplasmic interactions; transient, membrane-embedded or low-abundance interactions are under-represented (Hart et al., 2006; Luck et al., 2020; Venkatesan et al., 2009), so the topology we describe is that of the detectable binary interactome, not the complete interactome. Second, the average shortest path was estimated by sampling 1,000 node pairs rather than computed exactly, which is a practical necessity for an 8,149-node graph but introduces sampling noise (null $\sigma = 0.017$); the very large Z-scores indicate this noise does not affect the qualitative conclusion. Third, GO enrichment relied on the hypergeometric test with ancestor propagation and included electronic (IEA) annotations, which are comprehensive but less experimentally grounded (Ashburner et al., 2000; Rhee et al., 2008); the very high fold enrichments of the most specific terms (e.g. 40.6 for cap hypermethylation) are reassuring but should be interpreted with the IEA caveat in mind. Fourth, community detection is stochastic and algorithm-dependent; although Louvain and Leiden agreed substantially (ARI 0.62), the precise boundaries of communities — especially for the smaller, less densely connected modules — should be treated as statistically robust but not unique. Finally, the power-law exponent is sensitive to fitting range, as discussed above.

Future directions. Several extensions follow naturally. Integrating HuRI with co-complex and literature-curated maps (Drew et al., 2017; Rolland et al., 2014; Venkatesan et al., 2009) would enable a quantitative comparison of binary versus co-complex topology, particularly of the clustering coefficient. Weighted and directed community detection on condition-specific or tissue-specific sub-interactomes would reveal how the modular partition reconfigures across cellular contexts. Overlays of disease-gene, drug-target and expression data onto the communities and hubs identified here would test the network-medicine hypothesis (Barabási et al., 2011) that topological position predicts phenotypic and therapeutic relevance. Finally, the betweenness hubs we report — MEOX2, LNX1, UBQLN2, TLE5 and CYSRT1 among them — are concrete candidates for targeted perturbation experiments aimed at probing their bridging roles in human cellular networks.

5 Conclusion

Using the HuRI reference binary interactome of Luck *et al.* (Luck et al., 2020) as an undirected simple graph of 8,245 proteins and 52,068 interactions, we have shown that the human binary interactome is dominated by a single giant component (98.84% of nodes), exhibits a heavy-tailed approximately scale-free degree distribution, satisfies the small-world property against degree-preserving nulls ($S = 1.60$; transitivity $Z = 64.4$, path length $Z = 19.6$), and partitions into strongly modular communities (Louvain $Q = 0.513$; Leiden $Q = 0.518$). The ten largest communities correspond to coherent biological processes spanning cytoskeletal organisation, transport, transcription, the extracellular matrix, metabolism, immunity and ubiquitin-dependent proteolysis, Ras signalling, ciliary transport and spliceosomal assembly. Betweenness-centrality hubs including MEOX2, LNX1, UBQLN2, TLE5 and CYSRT1 are identified as candidate integrative control points. Together, these results demonstrate that a single, reproducible topological workflow applied to a quality-controlled binary reference map recovers the major architectural and functional organisation of the human cell.

References

Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97, 2002. doi: 10.1103/RevModPhys.74.47.

Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, et al. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000. doi: 10.1038/75556.

Gary D. Bader and Christopher W. V. Hogue. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 4(1):2, 2003. doi: 10.1186/1471-2105-4-2.

Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286 (5439):509–512, 1999. doi: 10.1126/science.286.5439.509.

Albert-László Barabási and Zoltán N. Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2004. doi: 10.1038/nrg1272.

Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68, 2011. doi: 10.1038/nrg2918.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008. doi: 10.1088/1742-5468/2008/10/P10008.

Ulrik Brandes. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25(2):163–177, 2001. doi: 10.1080/0022250X.2001.9990249.

Pascal Braun, Murat Tasan, Matija Dreze, Miriam Barrios-Rodiles, Irma Lemmens, Haiyuan Yu, et al. An experimentally derived confidence score for binary protein–protein interactions. *Nature Methods*, 6(1):91–97, 2009. doi: 10.1038/nmeth.1281.

Aaron Clauset, Cosma Rohilla Shalizi, and Mark E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009. doi: 10.1137/070710111.

Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695:1–9, 2006.

Alexander Cumberworth, Guillaume Lamour, M. Madan Babu, and Jörg Gsponer. Promiscuity as a functional trait: intrinsically disordered regions as central players of interactomes. *Biochemical Journal*, 454(3):361–369, 2013. doi: 10.1042/BJ20130545.

Michael E. Cusick, Haiyuan Yu, Alex Smolyar, Kavitha Venkatesan, Anne-Ruxandra Carvunis, Nicolas Simonis, et al. Literature-curated protein interaction datasets. *Nature Methods*, 6(1): 39–46, 2009. doi: 10.1038/nmeth.1284.

Kevin Drew, Chanjae Lee, Ryan L. Huizar, Fan Tu, Blake Borgeson, Claire D. McWhite, Yun Ma, John B. Wallingford, and Edward M. Marcotte. Integration of over 9,000 mass spectrometry experiments builds a global map of human protein complexes. *Molecular Systems Biology*, 13 (6):932, 2017. doi: 10.15252/msb.20167490.

Stanley Fields and Ok-kyu Song. A novel genetic system to detect protein–protein interactions. *Nature*, 340(6230):245–246, 1989. doi: 10.1038/340245a0.

Linton C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(1): 35–41, 1977. doi: 10.2307/3033543.

Anne-Claude Gavin, Patrick Aloy, Paola Grandi, Roland Krause, Markus Boesche, Martina Marzioch, et al. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440(7084):631–636, 2006. doi: 10.1038/nature04532.

Matthew C. Good, Jesse G. Zalatan, and Wendell A. Lim. Scaffold proteins: hubs for controlling the flow of cellular information. *Science*, 332(6030):680–686, 2011. doi: 10.1126/science.1198701.

Aric Hagberg, Pieter Swart, and Daniel Schult. Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pages 11–15, Pasadena, CA, USA, 2008.

G. Traver Hart, Arun K. Ramani, and Edward M. Marcotte. How complete are current yeast and human protein-interaction networks? *Genome Biology*, 7(11):120, 2006. doi: 10.1186/gb-2006-7-11-120.

Da Wei Huang, Brad T. Sherman, and Richard A. Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1):1–13, 2009. doi: 10.1093/nar/gkn923.

Mark D. Humphries and Kevin Gurney. Network 'small-world-ness': a quantitative method for determining canonical network equivalence. *PLoS ONE*, 3(4):e0002051, 2008. doi: 10.1371/journal.pone.0002051.

Rachael P. Huntley, Tony Sawford, Prudence Mutowo-Meullenet, Aleksandra Shypitsyna, Carlos Bonilla, Maria J. Martin, and Claire O'Donovan. The GOA database: Gene Ontology annotation updates for 2015. *Nucleic Acids Research*, 43(D1):D1057–D1063, 2015. doi: 10.1093/nar/gku1113.

Hawoong Jeong, Sean P. Mason, Albert-László Barabási, and Zoltán N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, 2001. doi: 10.1038/35075138.

Nevan J. Krogan, Gerard Cagney, Haiyuan Yu, Gouqing Zhong, Xinghua Guo, Alexandr Ignatchenko, et al. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*, 440(7084):637–643, 2006. doi: 10.1038/nature04670.

Katja Luck, Dae-Kyum Kim, Luke Lambourne, Kerstin Spirohn, Bridget E. Begg, Wenting Bian, Ruth Brignall, Tiziana Cafarelli, Francisco J. Campos-Laborie, Benoit Charloteaux, et al. A reference map of the human binary protein interactome. *Nature*, 580(7803):402–408, 2020. doi: 10.1038/s41586-020-2188-x.

Steven Maere, Koen Heymans, and Martin Kuiper. BiNGO: a Cytoscape plugin to assess over-representation of Gene Ontology categories in biological networks. *Bioinformatics*, 21(16): 3448–3449, 2005. doi: 10.1093/bioinformatics/bti551.

Sergei Maslov and Kim Sneppen. Specificity and stability in topology of protein networks. *Science*, 296(5569):910–913, 2002. doi: 10.1126/science.1065103.

Mark E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006. doi: 10.1073/pnas.0601602103.

Mark E. J. Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113, 2004. doi: 10.1103/PhysRevE.69.026113.

Nataša Pržulj, Derek G. Corneil, and Igor Jurisica. Modeling interactome: scale-free or geometric? *Bioinformatics*, 20(18):3508–3515, 2004. doi: 10.1093/bioinformatics/bth436.

Seung Yon Rhee, Valerie Wood, Kara Dolinski, and Sorin Draghici. Use and misuse of the Gene Ontology annotations. *Nature Reviews Genetics*, 9(7):509–515, 2008. doi: 10.1038/nrg2363.

Thomas Rolland, Murat Taşan, Benoit Charloteaux, Samuel J. Pevzner, Quan Zhong, Nidhi Sahni, Song Yi, et al. A proteome-scale map of the human interactome network. *Cell*, 159(5): 1212–1226, 2014. doi: 10.1016/j.cell.2014.10.050.

Jean-François Rual, Kavitha Venkatesan, Tong Hao, Tomoko Hirozane-Kishikawa, Amélie Dricot, Ning Li, Gabriel F. Berriz, et al. Towards a proteome-scale map of the human protein–protein interaction network. *Nature*, 437(7062):1173–1178, 2005. doi: 10.1038/nature04209.

Roded Sharan, Igor Ulitsky, and Ron Shamir. Network-based prediction of protein function. *Molecular Systems Biology*, 3(1):88, 2007. doi: 10.1038/msb4100129.

Ulrich Stelzl, Uwe Worm, Maciej Lalowski, Christian Haenig, Felix H. Brembeck, Heike Goehler, Martin Stroedicke, Martina Zenkner, et al. A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, 122(6):957–968, 2005. doi: 10.1016/j.cell.2005.08.029.

The Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Research*, 49(D1):D325–D334, 2021. doi: 10.1093/nar/gkaa1113.

V. A. Traag, L. Waltman, and N. J. van Eck. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1):5233, 2019. doi: 10.1038/s41598-019-41695-z.

Kavitha Venkatesan, Jean-François Rual, Alexei Vazquez, Ulrich Stelzl, Irma Lemmens, Tomoko Hirozane-Kishikawa, et al. An empirical framework for binary interactome mapping. *Nature Methods*, 6(1):83–90, 2009. doi: 10.1038/nmeth.1280.

Marc Vidal, Michael E. Cusick, and Albert-László Barabási. Interactome networks and human disease. *Cell*, 144(6):986–998, 2011. doi: 10.1016/j.cell.2011.02.016.

Christian von Mering, Roland Krause, Berend Snel, Michael Cornell, Stephen G. Oliver, Stanley Fields, and Peer Bork. Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, 417(6887):399–403, 2002. doi: 10.1038/nature750.

Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998. doi: 10.1038/30918.

Table 3. Top-five enriched Gene Ontology Biological Process terms for each of the ten largest Louvain communities (hypergeometric test, FDR< 0.05, Benjamini–Hochberg). k, community proteins annotated to the term; K, background proteins annotated. Fold$= (k/n)/(K/N)$.

Comm.	GO Biological Process term	FDR	Fold	k	K
0	cytoskeleton organization	1.1×10^{-22}	1.71	255	437
	intermediate filament organization	2.7×10^{-21}	2.82	57	59
	supramolecular fiber organization	2.7×10^{-21}	1.91	164	251
	cellular component organization or biogenesis	2.0×10^{-20}	1.24	1004	2366
	intermediate filament-based process	2.0×10^{-20}	2.74	59	63
1	transmembrane transport	1.6×10^{-67}	2.84	258	461
	transport	2.1×10^{-48}	1.76	512	1477
	establishment of localization	5.1×10^{-48}	1.72	535	1576
	localization	2.5×10^{-38}	1.58	570	1828
	monoatomic ion transport	5.6×10^{-29}	2.43	157	328
2	regulation of nucleobase-containing compound metabolic process	4.3×10^{-16}	1.52	340	1857
	regulation of RNA metabolic process	6.0×10^{-16}	1.54	319	1718
	regulation of primary metabolic process	1.3×10^{-15}	1.42	410	2397
	regulation of transcription by RNA polymerase II	4.1×10^{-15}	1.64	250	1266
	regulation of DNA-templated transcription	7.2×10^{-13}	1.51	287	1577
3	epidermis development	4.9×10^{-6}	4.24	22	62
	external encapsulating structure organization	4.5×10^{-4}	3.66	19	62
	extracellular matrix organization	1.0×10^{-3}	3.53	18	61
	extracellular structure organization	1.0×10^{-3}	3.47	18	62
	tissue development	1.0×10^{-3}	2.14	41	229
4	metabolic process	8.2×10^{-6}	1.33	266	2796
	primary metabolic process	2.7×10^{-4}	1.31	235	2501
	purine-containing compound metabolic process	9.4×10^{-4}	2.48	32	180
	biosynthetic process	9.4×10^{-4}	1.58	99	877
	nucleobase-containing small molecule metabolic process	1.3×10^{-3}	2.25	36	223
5	SCF-dependent proteasomal ubiquitin-dependent protein catabolic process	2.9×10^{-10}	8.60	18	36
	humoral immune response	8.8×10^{-10}	4.86	28	99
	antimicrobial humoral response	2.9×10^{-7}	5.21	20	66
	antibacterial humoral response	1.8×10^{-6}	7.45	13	30
	defense response to bacterium	3.4×10^{-6}	3.95	23	100
6	post-translational protein modification	1.4×10^{-7}	3.39	34	407
	protein ubiquitination	1.4×10^{-7}	3.87	29	304
	protein modification by small protein conjugation or removal	1.4×10^{-7}	3.40	33	393
	protein modification by small protein conjugation	1.7×10^{-6}	3.37	29	349
	protein K48-linked ubiquitination	3.0×10^{-6}	8.25	12	59
7	intracellular signaling cassette	4.3×10^{-4}	3.33	22	311
	positive regulation of Ras protein signal transduction	4.3×10^{-4}	23.55	5	10
	Ras protein signal transduction	1.4×10^{-3}	9.99	7	33
	negative regulation of beige fat cell differentiation	1.8×10^{-3}	47.10	3	3
	small GTPase-mediated signal transduction	1.8×10^{-3}	4.63	12	122
8	plasma membrane tubulation	1.5×10^{-2}	22.36	4	9
	intraciliary retrograde transport	2.7×10^{-2}	16.77	4	12
	intraciliary transport	3.4×10^{-2}	8.98	5	28
	regulation of cell cycle phase transition	3.4×10^{-2}	3.16	13	207
	regulation of nuclear division	3.4×10^{-2}	5.34	7	66
9	spliceosomal tri-snRNP complex assembly	1.4×10^{-10}	23.94	11	24
	RNA capping	5.2×10^{-9}	32.15	8	13
	7-methylguanosine cap hypermethylation	5.2×10^{-9}	40.63	7	9
	mRNA catabolic process	5.2×10^{-9}	10.60	14	69
	spliceosomal snRNP assembly	5.2×10^{-9}	13.63	12	46