Decoding of persistent multiscale structures in complex biological networks

Fan Zheng[1,3,*], She Zhang[2,3], Christopher Churas[1], Dexter Pratt[1], Ivet Bahar[2] and Trey Ideker[1,*]

[1]Division of Genetics, Department of Medicine, University of California, San Diego, CA 92093, USA.

[2]Department of Computational and Systems Biology, School of Medicine, University of Pittsburgh, Pittsburgh, PA 15123, USA.

[3]These authors contributed equally to this work

*To whom correspondence should be addressed.

Abstract

Networks of genes, proteins, and cells exhibit significant multiscale organization, with distinct communities appearing at different spatial resolutions. Here, we apply the concept of 'persistent homology' to identify network communities that persist within defined scale ranges, yielding a hierarchy of robust structures in data. Application to mouse single-cell transcriptomes significantly expands the catalog of cell types identified by current tools, while analysis of SARS-COV-2 networks suggests pro-viral hijacking of WNT.

Significant patterns in data often become apparent only when looking at the right scale. For example, single-cell RNA sequencing data can be clustered coarsely to identify large groupings of cells (e.g. general germ layers), or analyzed more sharply to identify compact cell communities delineating specific subtypes (e.g. pancreas islet β-cells, thymus epithelium)[1]. Likewise, protein-protein interaction networks can inform cellular structures spanning a wide range of spatial dimensions, from protein dimers (e.g. leucine zippers) to larger complexes of dozens or hundreds of subunits (e.g. proteasome, nuclear pore) to entire organelles (e.g. centriole, mitochondria)[2]. Many different approaches have been devised or applied to detect structures in biological data, including standard clustering, network community detection, and low-dimensional data projection[3–5], some of which can be tuned for sensitivity to objects of a certain size or scale (so-called 'resolution parameters')[6,7]. Even tunable algorithms, however, face the dilemma that the particular scale(s) at which the significant biological structures arise are typically unknown in advance.

Guidelines for detecting robust patterns across scales come from the field of topological data analysis, which studies the geometric "shape" of data using tools from algebraic topology and pure mathematics[8]. A fundamental concept in this field is "persistent homology"[9], the idea that the core structures intrinsic to a dataset are those that persist across different scales. Recently, this concept has begun to be applied to biological networks[10,11]. Here, we sought to integrate concepts from persistent homology with existing algorithms for network community detection, resulting in a fast and practical multiscale approach we call the Hierarchical community Decoding Framework (HiDeF).

HiDeF works in the three modular phases to analyze the structure of a biological network (**Methods**). The first phase detects network communities, which are identified continually as the spatial resolution is scanned, producing a comprehensive pool of candidates across all scales of

analysis (**Fig. 1a**). Second, candidate communities arising at different resolutions are pairwise aligned to identify those that have been redundantly identified and are thus *persistent* (**Fig. 1b**). Third, persistent communities are analyzed to identify cases where a community is fully or partially contained within another (typically larger) community, resulting in a hierarchical assembly of nested and overlapping biological structures embedded in the network data (**Fig. 1c**). HiDeF is implemented as a Python package and can be accessed interactively in the Cytoscape network analysis and visualization environment[12] (**Code and data availability**).

We first explored the idea of measuring community persistence *via* analysis of synthetic networks[13] in which communities were simulated and embedded at two different scales (**Supplementary Fig. 1a**; **Methods**). Notably, the communities determined to be most persistent by HiDeF were found to accurately recapitulate the simulated communities at the two scales (**Supplementary Fig. 1b-g**). In contrast, applying community detection algorithms at a fixed resolution had limited capability to capture both scales of simulated structures simultaneously (**Supplementary Fig. 2; Methods**).

We next evaluated whether persistent community detection improves the characterization of cell types. We applied HiDeF to detect robust nested communities within cell-cell similarity networks based on the mRNA expression profiles of 100,605 single cells gathered across the organs and tissues of mice (obtained from the *Tabula Muris* project[14]; **Methods**). These cells had been annotated with a controlled vocabulary of cell types from the Cell Ontology (CO)[15], *via* analyses of cell-type-specific expression markers[14]. We used groups of cells sharing the same annotations to define a panel of 81 reference cell types and measured the degree to which each reference cell type could be recapitulated by a HiDeF community of cells (**Methods**). We compared these results to TooManyCells[16] and Conos[17], two recently developed methods that generate nested communities of single cells in divisive and agglomerative manners, respectively

(**Methods**). Reference cell types tended to better match communities generated by HiDeF than those of other approaches, with 65% (54/81) having a highly overlapping community (Jaccard index > 0.5) in the HiDeF hierarchy (**Fig. 2a,b**). This favorable performance was observed consistently when adjusting HiDeF parameters to formulate a simple hierarchy, containing only the strongest structures, or a more complex hierarchy including additional communities that are less persistent but still significant (**Fig. 2c**).

The top-level communities in the HiDeF hierarchy corresponded to broad cell lineages such as “T cell”, “B cell”, and “epidermal cell”. Finer-grained communities mapped to more specific known subtypes (**Fig. 2d**) or, more frequently, putative new subtypes within a lineage. For example, “epidermal cell” was split into two distinct epidermal tissue locations, skin and tongue; further splits suggested the presence of still more specific uncharacterized cell types (**Fig. 2e**). HiDeF communities also captured known cell types that were not apparent from 2D visual embeddings (**Supplementary Fig. 3a,b**), and also suggested new cell-type combinations. For example, astrocytes were joined with two communities of neuronal cells to create a distinct cell type not observed in the hierarchies of TooManyCells[16], Conos[17], or a two-dimensional data projection with UMAP[18] (**Fig. 2f**, **Supplementary Fig. 3c**). This community may correspond to the grouping of a presynaptic neuron, postsynaptic neuron, and a surrounding astrocyte within a so-called “tripartite synapse”.

Next, we applied HiDeF to analyze protein-protein interaction networks, with the goal of characterizing protein complexes and higher-order protein assemblies spanning spatial scales. We benchmarked this task by the agreement between HiDeF communities and the Gene Ontology (GO)[19], a database that manually assigns proteins to cellular components, processes, or functions based on curation of literature (**Methods**). Application to protein-protein interaction networks from budding yeast and human found that HiDeF captured knowledge in GO more significantly than

previous pipelines proposed for this task, including the NeXO approach to hierarchical community detection[20] and standard hierarchical clustering of pairwise protein distances calculated by two recent approaches[21,22] (**Fig. 3a,b**; **Supplementary Figs. 4,5**; **Methods**).

We also applied HiDeF to analyze a collection of 27 human protein interaction networks[23,24]. We found significant differences in the distributions of community sizes across these networks, loosely correlating with the different measurement approaches used to generate each network. For example, BioPlex 2.0, a network characterizing biophysical protein-protein interactions by affinity-purification mass-spectrometry (AP-MS)[25], was dominated by small communities of 10-50 proteins, whereas a network based on mRNA coexpression[26] tended towards larger-scale communities of >50 proteins. In the middle of this spectrum, the STRING network, which integrated biophysical protein interactions and gene co-expression with a variety of other features[27], contained both small and large communities (**Fig. 3c**). In consistent with the observation above, the hierarchy of BioPlex has a relatively shallow shape in comparison to that of STRING, in which communities across many scales formed a deep hierarchy (**Fig. 3d,e**). In contrast to clustering frameworks, HiDeF recognizes when a community is contained by multiple parent communities, which in the context of protein-protein networks suggests that the community participates in diverse (pleiotropic) biological functions. For example, a community corresponding to the MAPK (ERK) pathway participated in multiple larger communities, including RAS and RSK pathways, sodium channels, and actin capping, consistent with the central roles of MAPK signaling in these distinct biological processes[28] (**Fig. 3f**). The hierarchies of protein communities identified from each of these networks have been made available as a resource in the NDEx database[29] (**Code and data availability**).

To explore multiscale data analysis in the context of an urgent public health issue, we considered a recent application of AP-MS that characterized interactions between the 27

SARS-COV-2 viral subunits and 332 human host proteins[30]. We used network propagation to select a subnetwork of the BioPlex 3.0 human protein interactome[31] proximal to these 332 proteins (1948 proteins and 22,835 interactions) and applied HiDeF to identify its community structure (**Methods**). Among the 252 persistent communities identified (**Supplementary Fig. 6**), we noted one consisting of human Transducin-Like Enhancer (TLE) family proteins, TLE1, TLE3, and TLE5, which interacted with SARS-COV2 Nsp13, a highly conserved RNA synthesis protein in corona and other nidoviruses (**Fig. 3g**)[32]. TLE proteins are well-known inhibitors of the Wnt signaling pathway[33]. Inhibition of WNT, in turn, has been shown to reduce coronavirus replication[34] and recently proposed as a COVID-19 treatment[35]. If interactions between Nsp13 and TLE proteins can be shown to facilitate activation of WNT, TLEs may be of potential interest as drug targets.

Community persistence provides a basic metric for distilling biological structure from data, which can be tuned to select only the strongest structures or to include weaker patterns that are less persistent but still significant. This concept applies to diverse biological subfields, as demonstrated here for single cell transcriptomics and protein interaction mapping. While these subfields currently employ very different analysis tools which largely evolve separately, it is perhaps high time to seek out core concepts and broader fundamentals around which to unify some of the ongoing development efforts. To that effect, the methods explored here have wide applicability to analyze the multiscale organization of many other biological systems, including those related to chromosome organization, the microbiome and the brain.

Acknowledgments

This work has been supported by the NIH grants to T.I. (R01 HG009979, U54 CA209891) and I.B. (P41 GM103712, P01 DK096990). We are grateful for the helpful discussions with Drs. Jianzhu Ma, Karen Mei, and Daniel Carlin.

Author Contributions

F.Z. designed the study and developed the conceptual ideas. F.Z. and S.Z. implemented the main algorithm. F.Z. and D.P collected the input data and conducted analysis. S.Z. made the Python package. C.C. developed the server that provides the Cytoscape integration. F.Z. and T.I. wrote the manuscript with suggestions from S.Z., D.P. and I.B.

Competing Interests

T.I. is cofounder of Data4Cure, is on the Scientific Advisory Board, and has an equity interest. T.I. is on the Scientific Advisory Board of Ideaya BioSciences and has an equity interest. The terms of these arrangements have been reviewed and approved by the University of California San Diego, in accordance with its conflict of interest policies.

References

1. Regev, A. et al. *Elife* **6**, (2017).
2. Harold, F.M. *Microbiol. Mol. Biol. Rev.* **69**, 544–564 (2005).
3. Jain, A.K., Murty, M.N. & Flynn, P.J. *ACM Comput. Surv.* **31**, 264–323 (1999).
4. Fortunato, S. & Hric, D. *Phys. Rep.* **659**, 1–44 (2016).
5. Maaten, L. van der & Hinton, G. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
6. Arenas, A., Fernández, A. & Gómez, S. *New J. Phys.* **10**, 053039 (2008).
7. Traag, V.A., Krings, G. & Van Dooren, P. *Sci. Rep.* **3**, 2930 (2013).
8. Carlsson, G. *Bull. Am. Math. Soc.* **46**, 255–308 (2009).
9. Edelsbrunner, H. & Harer, J. *Contemp. Math.* **453**, 257–282 (2008).
10. Dabaghian, Y., Mémoli, F., Frank, L. & Carlsson, G. *PLoS Comput. Biol.* **8**, e1002581 (2012).
11. Petri, G. et al. *J. R. Soc. Interface* **11**, 20140873 (2014).
12. Shannon, P. et al. *Genome Res.* **13**, 2498–2504 (2003).
13. Lancichinetti, A., Fortunato, S. & Radicchi, F. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **78**, 046110 (2008).
14. Tabula Muris Consortium et al. *Nature* **562**, 367–372 (2018).
15. Diehl, A.D. et al. *J. Biomed. Semantics* **7**, 44 (2016).
16. Schwartz, G.W. et al. *Nat. Methods* (2020).
17. Barkas, N. et al. *Nat. Methods* (2019).
18. Becht, E. et al. *Nat. Biotechnol.* (2018).
19. Ashburner, M. et al. *Nat. Genet.* **25**, 25–29 (2000).
20. Dutkowski, J. et al. *Nat. Biotechnol.* **31**, 38–45 (2013).
21. Cho, H., Berger, B. & Peng, J. *Cell Syst* **3**, 540–548.e5 (2016).
22. Cao, M. et al. *PLoS One* **8**, e76339 (2013).
23. Huang, J.K. et al. *Cell Syst* **6**, 484–495.e5 (2018).
24. Choobdar, S. et al. *Nat. Methods* **16**, 843–852 (2019).
25. Huttlin, E.L. et al. *Nature* **545**, 505–509 (2017).
26. Subramanian, A. et al. *Cell* **171**, 1437–1452.e17 (2017).
27. Szklarczyk, D. et al. *Nucleic Acids Res.* **47**, D607–D613 (2019).
28. Cargnello, M. & Roux, P.P. *Microbiology and Molecular Biology Reviews* **76**, 496–496 (2012).
29. Pratt, D. et al. *Cancer Res.* **77**, e58–e61 (2017).
30. Gordon, D.E. et al. *Nature* (2020).
31. Huttlin, E.L. et al. *bioRxiv* 2020.01.19.905109 (2020).doi:10.1101/2020.01.19.905109

32. Snijder, E.J., Decroly, E. & Ziebuhr, J. *Coronaviruses* 59–126 (2016).
33. Chodaparambil, J.V. et al. *EMBO J.* **33**, 719–731 (2014).
34. Wu, C.-J. et al. *Antimicrob. Agents Chemother.* **48**, 2693–2696 (2004).
35. Xu, J., Shi, P.-Y., Li, H. & Zhou, J. *ACS Infect Dis* (2020).

ONLINE CONTENT

Methods

Overview of the approach

Consider an undirected network graph G, representing a set of biological *objects* (vertices) and a set of *similarity relations* between these objects (edges). Examples of interest include networks of cells, where edges represent pairwise cell-cell similarity in transcriptional profiles characterized by single-cell RNA-seq, or networks of proteins, where edges represent pairwise protein-protein biophysical interactions. We seek to group these objects into *communities* (subsets of objects) that appear at different scales and identify approximate *containment relationships* among these communities, so as to obtain a hierarchical representation of the network structure. The workflow is implemented in three phases. Phase I identifies communities in G at each of a series of spatial resolutions γ. Phase II identifies which of these communities are *persistent* by way of a *pan-resolution community graph* G_C, in which vertices represent communities, including those identified at each resolution, and each edge links pairs of similar communities arising at different resolutions. Persistent communities correspond to large components in G_C. Phase III constructs a final hierarchical structure H that represents containment and partial containment relationships (directed edges) among the persistent communities (vertices).

Phase I: Pan-resolution community detection

Community detection methods generally seek to maximize a quantity known as the *network modularity*, as a function of community assignment of all objects[36]. A *resolution parameter* integrated into the modularity function can be used to tune the scale of the communities identified[7,37,38], with larger/smaller scale communities having more/fewer vertices on average (**Fig. 1a**). Of the several types of resolution parameter that have been proposed, we adopted that of the Reichardt-Bornholdt configuration model[37], which defines the generalized modularity as:

$$Q(\vec{G},\gamma) = \sum_{ij}(A_{ij} - \gamma\frac{k_i k_j}{2m})\delta(i,j)$$

where $\vec{G}$ defines a mapping from objects in G to community labels; k_i is the degree of vertex i; m is the total number of edges in G; γ is the resolution parameter; $\delta(i,j)$ indicates that vertices i and j are assigned to the same community by $\vec{G}$; and A is the adjacency matrix of G. To determine $\vec{G}$ we use the extended Louvain algorithm implemented in the Python package louvain-igraph (http://github.com/vtraag/louvain-igraph; version 0.6.1). Values of γ are sampled logarithmically between lower and upper bounds γ_{min} and γ_{max} at a minimum density such that for all γ there exist at least 10 nearby γ' satisfying:

$$|\log_{10}(\gamma') - \log_{10}(\gamma)| < 0.1$$

Two γ values satisfying the above formula are defined as γ*-proximal*. The sampling step, which was practically set to 0.1 to sufficiently capture the interesting structures in the data; it is conceptually similar to the Nyquist sampling frequency in signal processing[39]. We used γ_{min} = 0.001, which we found always resulted in the theoretical minimum number of communities, equal to the number of connected components in G. We used γ_{max} = 25 for single-cell data (**Fig. 2**, **Supplementary Fig. 3**) and γ_{max} = 50 for simulated networks (**Supplementary Figs. 1,2**) and protein interaction networks (**Fig. 3**, **Supplementary Figs. 4-6**). Performing Louvain community detection at each γ over this defined progression of values resulted in a set of communities $\vec{G}$ at each γ.

Phase II: Identification of persistent communities

To identify persistent communities, we define the pairwise similarity between any two communities a and b as the Jaccard similarity of their sets of objects, $s(a)$ and $s(b)$:

$$J(a,b) = \frac{s(a) \cap s(b)}{s(a) \cup s(b)}$$

Pairwise community similarity is computed only for pairs of communities identified at two different γ-proximal resolution values, as communities within a resolution do not overlap. To represent these similarities, we define a *pan-resolution community graph* G_C, in which vertices are communities identified at any resolution and edges connect pairs of similar communities having $J(a, b) > \tau$. Each component of G_C defines a family of similar communities spanning resolutions, for which the *persistence* can be naturally defined by the number of distinct γ values covered by the component. For each component in G_C larger than a persistence threshold χ, the biological objects participating in more than p% of communities represented by the vertices of that component define a *persistent community.*

Phase III: A hierarchy of nested and overlapping communities

We initialize a hierarchical structure represented by H, a directed acyclic graph (DAG) in which each vertex represents a persistent community. A *root vertex* is added to represent the community of all objects. The containment relationship between two vertices, v and w, is quantified by the *containment index* (*CI*):

$$CI(v,w) = \frac{|s(v) \cap s(w)|}{|s(w)|}$$

which measures the fraction of objects in w shared with v. An edge is added from v to w in H if $CI(v,w)$ is larger than a threshold σ (w is σ*-contained* by v). Since $J(v, w) < \tau$ for all v, w (a property established by the procedure for connecting similar communities in phase II), setting $\sigma \geq 2\tau/(1+\tau)$ guarantees H to be acyclic. In practice we used a relaxed threshold $\sigma = \tau$, which we found generally maintains the acyclic property but includes additional containment relations. In

the (in our experience rare) event that cycles are generated in H, i.e. $CI(v,w) \geq \tau$ and $CI(w,v) \geq \tau$, we add a new community to H, the union of v and w, and remove v and w from H.

Finally, redundant relations are removed by obtaining a transitive reduction[40] of H, which represents the hierarchy returned by HiDeF describing the organization of communities. The biological objects assigned to each community are expanded to include all objects assigned to its descendants. Throughout this study, we used the parameters τ = 0.75, χ = 5, p = 75. Generally, we observed that the conclusions drawn in this study were robust to this choice of parameters (**Supplementary Fig. 5**).

Simulated benchmark networks

Simulated network data were generated using the Lancichinetti–Fortunato–Radicchi (LFR) method[13] (**Supplementary Figs. 1,2**). We used an available implementation (LFR benchmark graphs package 5 at http://www.santofortunato.net/resources) to generate benchmark networks with two levels of embedded communities, a coarse-grained (macro) level and a fine-grained (micro) level. Within each level, a vertex was exclusively assigned to one community. Two parameters, μ_c and μ_f, were used to define the fractions of edges violating the simulated community structures at the two levels. All other edges were restricted to occur between vertices assigned to the same community (**Supplementary Fig. 1a**). We fixed other parameters of the LFR method to values explored by previous studies[7]. In particular, N = 1000 (number of vertices), k = 10 (average degree), *maxk* = 40 (maximum degree), *minc* = 5 (minimum number of vertices for a micro-community), *maxc* = 20 (maximum number of vertices for a micro-community), *minC* = 50 (minimum number of vertices for a macro-community), *maxC* = 100 (maximum number of vertices for a macro-community), t_1 = 2 (minus exponent for the degree sequence), t_2 = 1 (minus exponent for the community size distribution). The numbers of coarse-grained communities and fine-grained communities in each simulated network were approximately bounded by *minC*, *maxC*, *minc* and *maxc* (10-20 and 50-200, respectively), and the sizes of communities within each level were set to be close to each other (as t_2 = 1).

Some community detection algorithms include iterations of local optimization and vertex aggregation, a process that, like HiDeF, also defines a hierarchy of communities, albeit as a tree rather than a DAG. We demonstrated that without scanning multiple resolutions, this process alone was insufficient to detect the simulated communities at all scales (**Supplementary Fig. 2**). We used Louvain and Infomap[41,42], which have stable implementations and have shown strong performance in previous community detection studies[43]. For Louvain, we optimized the standard Newman-Girvan modularity (equivalent to $\gamma = 1$, see above) using the implementation at http://github.com/vtraag/louvain-igraph. For Infomap, we used the version 1.0.0-beta.47, and set 'Markov time' (the 'resolution parameter' of Infomap) to 1 and other parameters to default. In general, these settings generated trees with two levels of communities. Note that Infomap sometimes determined that the input network was non-hierarchical, in which cases the coarse- and fine-grained communities were identical by definition.

Single-cell RNA-seq data

Mouse single-cell RNA-seq data (**Fig. 2**) were obtained from the *Tabula Muris* project[14], (https://tabula-muris.ds.czbiohub.org/), which contains two datasets generated with different experimental methods of separating bulk tissues into individual cells: FACS and microfluidic droplet. We applied HiDeF to the shared nearest neighbor graph parsed from the data files (R objects) provided in that study, focusing analysis on the FACS dataset. The same analysis of the droplet dataset supported the same conclusion (data not shown). Application of HiDeF to this dataset generated a hierarchy of 273 communities (**Fig. 2d**). ScanPy 1.4.5[44] was used to create tSNE or UMAP embeddings and associated two-dimensional visualizations[18] as baselines for comparison (**Fig. 2e,f**; **Supplementary Fig. 3a,b**). Through previous analysis of the single-cell RNA data, all cells in these datasets had been annotated with matching cell-type classes in the Cell Ontology (CO)[15]. Before comparing these annotations with the communities detected by HiDeF, we expanded the set of annotations of each cell according to the CO structure, to ensure the set also included all of the ancestor cell types of the type that was annotated. For example, CO has the relationship "[keratinocyte] (is_a) [epidermal_cell]", and thus all cells annotated as "keratinocyte" are also annotated as "epidermal cell". The CO was obtained from http://www.obofoundry.org/ontology/cl.html and processed by the Data Driven Ontology Toolkit (DDOT)[45] retaining "is_a" relationships only.

We compared HiDeF to TooManyCells[16] and Conos[17] as baseline methods. The former is a divisive method which iteratively applies bipartite spectral clustering to the cell population until the modularity of the partition is below a threshold; the latter uses the Walktrap algorithm to agglomeratively construct the cell-type hierarchy[46]. TooManyCells (version 0.2.2.0) was run with the parameter *"*min-modularity*"* set to 0.025 as recommended in the original paper[16]*,* with other settings set to default. This process generated dendrograms (binary trees) with 463 communities. The Walktrap algorithm was run from the Conos package (version 1.2.1) with the parameter "step" set to 20 as recommended in the original paper[17], yielding a dendogram. The *greedyModularityCut* method in the Conos package was used to select *N* fusions in the original dendrogram, resulting in a reduced dendrogram with 2*N*+1 communities (including *N* internal and *N+1* leaf nodes). Here we used *N* = 125, generating a hierarchy with 251 communities (**Fig. 2c**).

The communities in each hierarchy were ranked to analyze the relationships between cell-type recovery and model complexity (**Fig. 2c**). HiDeF communities were ranked by their persistence; Conos and TooManyCells were ranked according to the modularity scores those methods associate with each branch-point in their dendrograms. Conos/Walktrap uses a score based on the gain of modularity in merging two communities, whereas TooManyCells uses the modularity of each binary partition.

Protein-protein interaction networks

We obtained a total of 27 human protein interaction networks gathered previously by survey studies[23,24], along with one integrated network from budding yeast (*S. cerevisiae*) that had been used in a previous community detection pipeline, NeXO[20]. This collection contained two versions of the STRING interaction database, with the second removing edges from text mining (labeled

STRING-t versus STRING, respectively; **Fig. 3**). Benchmark experiments for the recovery of the Gene Ontology (GO) were performed with STRING and the yeast network (**Fig. 3a,b**, **Supplementary Fig. 4**). The reference GO for yeast proteins was obtained from http://nexo.ucsd.edu/. A reference GO for human proteins was created by the DDOT package[45].

HiDeF was directly applied to all of the above benchmark networks. The NeXO communities were obtained from http://nexo.ucsd.edu/, with a robustness score assigned to each community. To benchmark communities created by hierarchical clustering, we first calculated two versions of pairwise protein distances (HC.1 and HC.2; **Fig. 3a,b; Supplementary Fig. 4**) using Mashup and DSD[21,22]. Mashup was used to embed each protein as a vector, with 500 and 800 dimensions for yeast and human, as recommended in the original paper[21]. A pairwise distance was computed for each pair of proteins as the cosine distance between the two vectors. DSD generates pairwise distances by default. Given these pairwise distances, UPGMA clustering was applied to generate binary hierarchical trees. Following the procedure given in the NeXO and Mashup papers[20,21] communities with <4 proteins were discarded.

Since all methods had slight differences in the resulting number of communities, communities from each method were sorted in decreasing order of score, enabling comparison of results across the same numbers of top-ranked communities. HiDeF communities were scored by persistence. NeXO communities were scored by the robustness value assigned to each community in the original paper[20]. To score each community *c* of hierarchical clustering (branch in the dendrogram), a one-way Mann-Whitney U-test was used to test for significant differences between two sets of protein pairwise distances: (set 1) all pairs consisting of a protein in *c* and a protein in the sibling community of *c*; (set 2) all pairs consisting of a protein in each of the two children communities of *c*. The communities were sorted by the one-sided p-value of significance that distances in set 1 are greater than those in set 2.

Analysis of SARS-COV-2 viral-human protein network

332 human proteins identified to interact with SARS-COV-2 viral protein subunits were obtained from a recent study[30]. This list was expanded to include additional human proteins connected to two or more of the 332 virus-interacting human proteins in the BioPlex 3.0 network[31]. These operations resulted in a network of 1948 proteins and 22,835 interactions. HiDeF was applied to this network with the same parameter settings as for other protein-protein interaction networks (see previous **Methods** sections), and enrichment analysis was performed via g:Profiler[47] (**Fig. 3f; Supplementary Fig. 6**).

Code and data availability

HiDeF is available through CDAPS (Community Detection APplication and Service), which enables simultaneous visualization of the hierarchical model and the underlying network data and is integrated with the Cytoscape visualization and analysis environment. The Cytoscape App can be downloaded at: http://apps.cytoscape.org/apps/cycommunitydetection.

HiDeF is separately available as a Python package: https://github.com/fanzheng10/HiDeF.

The hierarchical models generated in this study can be obtained as a network collection within the Network Data Exchange (NDEx) database[29]: http://www.ndexbio.org/#/networkset/9460f0d2-ac0b-11ea-aaef-0ac135e8bacf. These models include the hierarchy of murine cell types (**Fig. 2**), the hierarchies of yeast and human protein communities identified through protein network analysis (**Fig. 3**), and the hierarchy of human protein complexes targeted by SARS-COV2 (**Supplementary Fig. 6**).

Online references

36. Newman, M.E.J. & Girvan, M. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **69**, 026113 (2004).
37. Reichardt, J. & Bornholdt, S. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **74**, 016110 (2006).
38. Jeub, L.G.S., Sporns, O. & Fortunato, S. *Sci. Rep.* **8**, 3259 (2018).
39. Oshana, R. *DSP for Embedded and Real-Time Systems* 1–14 (2012).
40. Aho, A.V., Garey, M.R. & Ullman, J.D. *SIAM J. Comput.* **1**, 131–137 (1972).
41. Blondel, V.D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. *J. Stat. Mech.* **2008**, P10008 (2008).
42. Rosvall, M. & Bergstrom, C.T. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 1118–1123 (2008).
43. Yang, Z., Algesheimer, R. & Tessone, C.J. *Sci. Rep.* **6**, 30750 (2016).
44. Wolf, F.A., Angerer, P. & Theis, F.J. *Genome Biol.* **19**, 15 (2018).
45. Yu, M.K. et al. *Cell Syst* **8**, 267–273.e3 (2019).
46. Pons, P. & Latapy, M. *J. Graph Algorithms Appl.* **10**, 191–218 (2006).
47. Raudvere, U. et al. *Nucleic Acids Res.* **47**, W191–W198 (2019).
48. Fabregat, A. et al. *Nucleic Acids Res.* **46**, D649–D655 (2018).

Figures

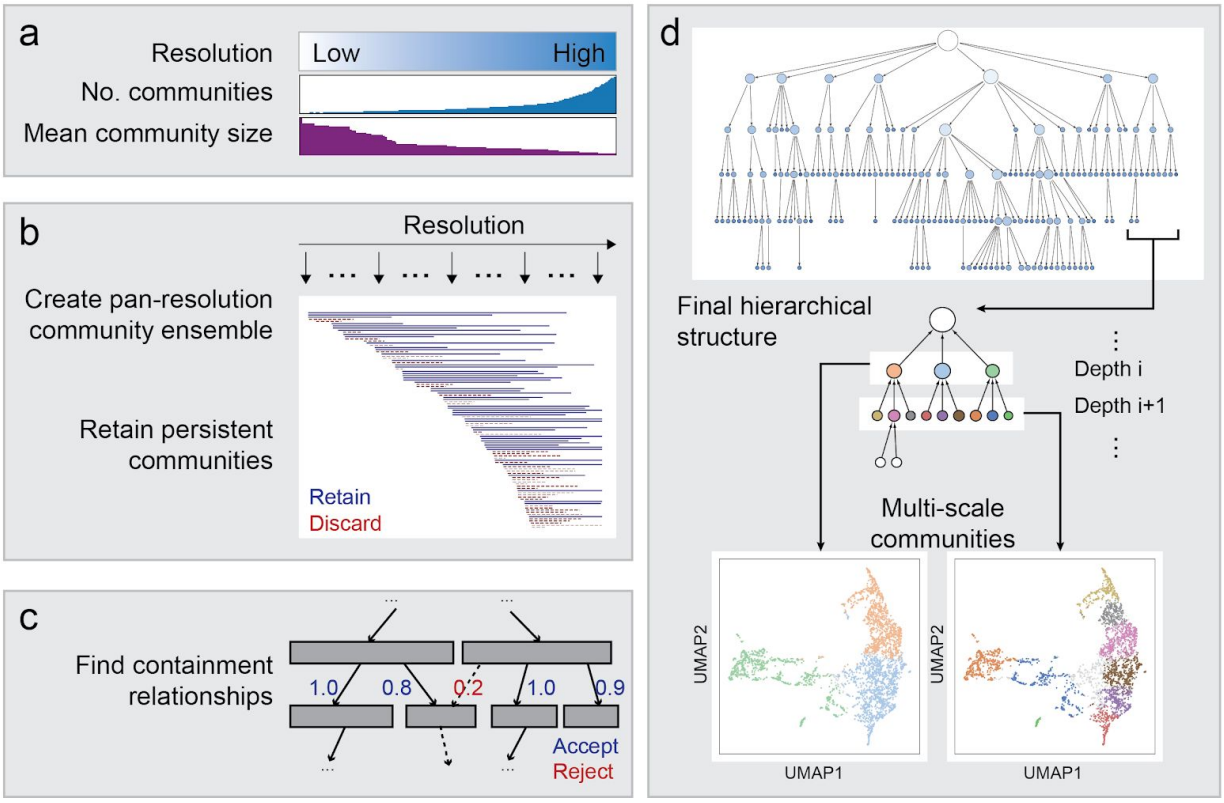


Fig. 1. Identification of persistent communities with HiDeF. a, Increasing the spatial resolution generally increases the number of communities and decreases the average community size. **b,** Pan-resolution community detection yields a candidate pool of communities. Communities that are robustly identified across a wide range of resolutions are considered persistent and retained. **c,** Set containment analysis is used to define the relationships between communities, leading to **d,** the final hierarchical model, in which vertices of increasing depths from the root represent communities of increasingly high resolutions.

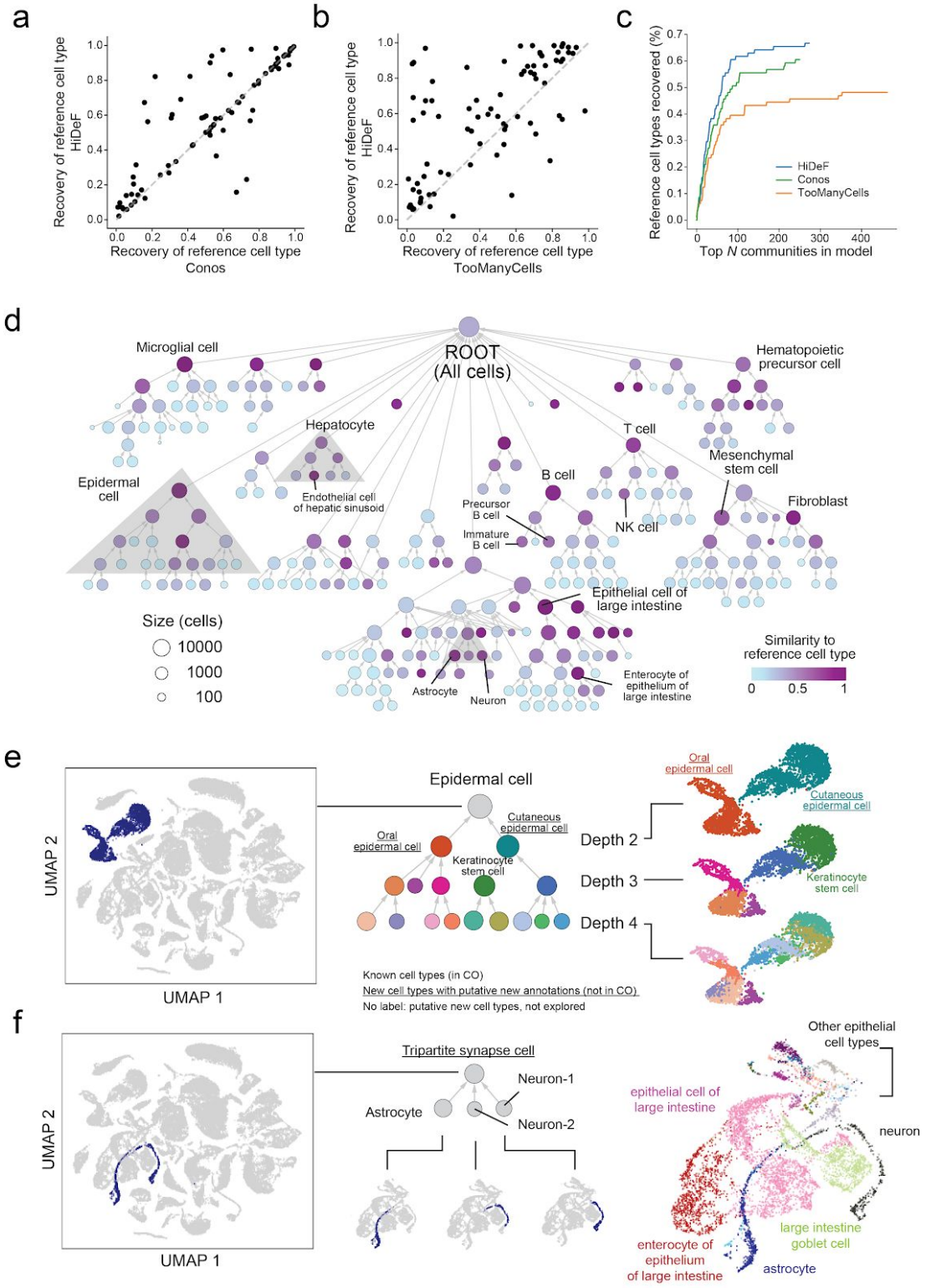

a
Recovery of reference cell type HiDeF
Recovery of reference cell type Conos
b
Recovery of reference cell type HiDeF
Recovery of reference cell type TooManyCells
c
Reference cell types recovered (%)
Top N communities in model
HiDeF
Conos
TooManyCells
d
ROOT (All cells)
Microglial cell
Hematopoietic precursor cell
Hepatocyte
Endothelial cell of hepatic sinusoid
Epidermal cell
T cell
B cell
Precursor B cell
Immature B cell
NK cell
Mesenchymal stem cell
Fibroblast
Epithelial cell of large intestine
Enterocyte of epithelium of large intestine
Astrocyte
Neuron
Size (cells)
10000
1000
100
Similarity to reference cell type
0
0.5
1
e
UMAP 2
UMAP 1
Epidermal cell
Cutaneous epidermal cell
Oral epidermal cell
Keratinocyte stem cell
Depth 2
Depth 3
Depth 4
Known cell types (in CO)
New cell types with putative new annotations (not in CO)
No label: putative new cell types, not explored
f
UMAP 2
UMAP 1
Tripartite synapse cell
Neuron-1
Astrocyte
Neuron-2
Other epithelial cell types
epithelial cell of large intestine
neuron
large intestine goblet cell
enterocyte of epithelium of large intestine
astrocyte

Fig. 2. A hierarchy of mammalian cell types from single-cell transcriptomes. a-b, Recovery of reference cell types by HiDeF (y axis) in comparison to Conos[17] or TooManyCells[16] (x axis of panels a or b, respectively). For each reference cell type (points), the extent of recovery is measured as the maximum Jaccard similarity of the set of reference cells with those of any detected community. **c,** Percentages of reference cell types recovered (Jaccard similarity > 0.5) among the top *N* ranked cell communities. Communities are ranked in the descending order of score for each community detection tool (**Methods**). **d,** Hierarchy of 273 putative mouse cell types identified by HiDeF. Vertices are cell communities, with color gradient indicating the extent of the optimal match (Jaccard similarity) to a reference cell type. Selected matches to reference cell types are labeled. Gray regions indicate sub-hierarchies (epidermal cells, astrocytes/neurons, and hepatocytes) related to subsequent panels and other figures (**Supplementary Fig. 3**). **e**, Epidermal cell communities. Left: UMAP 2D projection of all cells, with epidermal cells highlighted in dark blue. Middle: Sub-hierarchy of epidermal cell communities as determined by HiDeF. Right: Correspondence between the UMAP projection and the sub-hierarchy, with colors marking the same cell populations across the two representations. **f,** Astrocyte and neuron communities. Left: UMAP 2D projection of all cells, with astrocytes and neurons highlighted in dark blue. Middle: Sub-hierarchy of astrocyte and neuron communities as determined by HiDeF. Cells in the three small communities are highlighted in the below UMAP projections. Right: Broader UMAP context with cells colored and labeled as per the original *Tabula Muris* analysis[14].

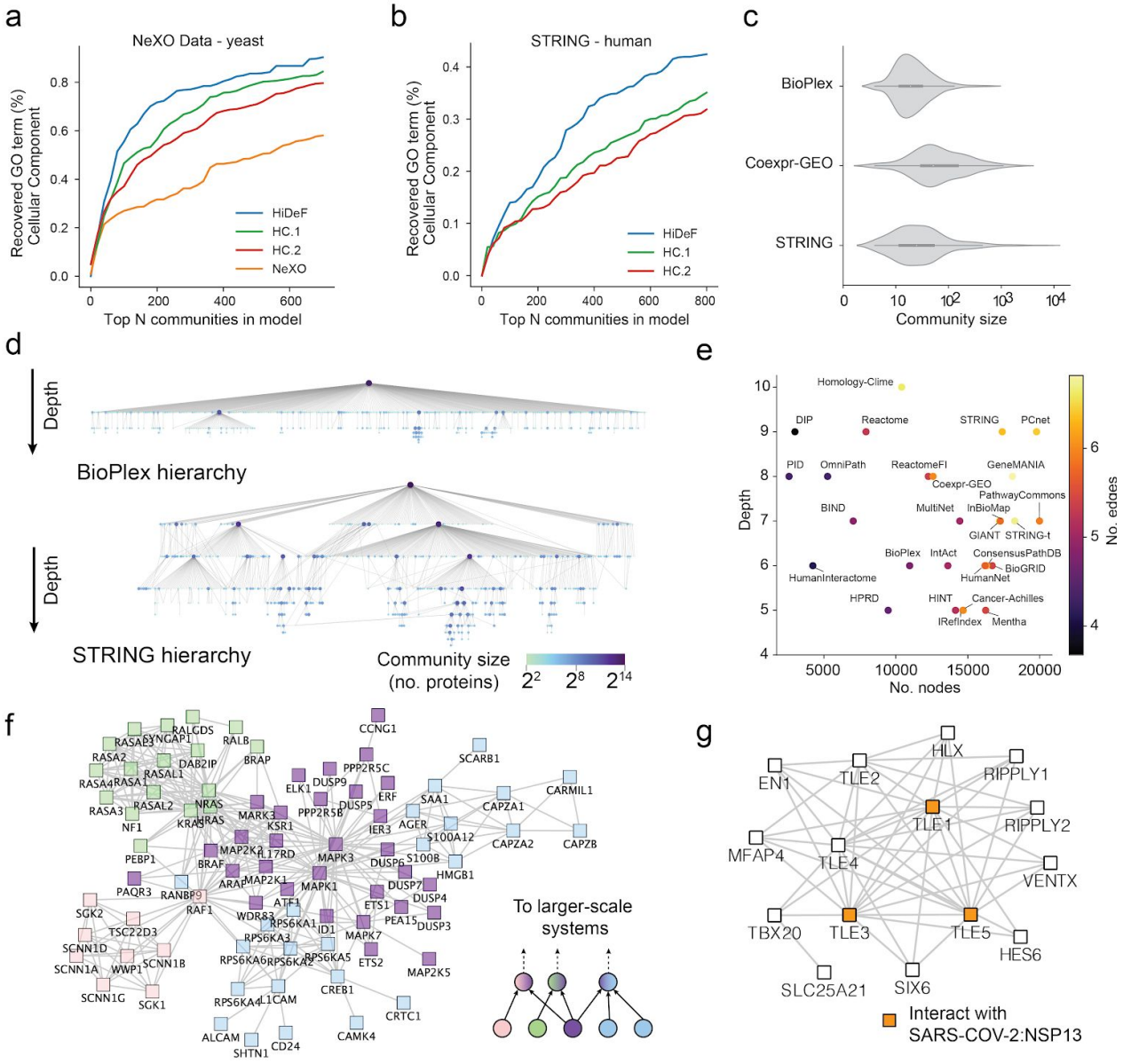
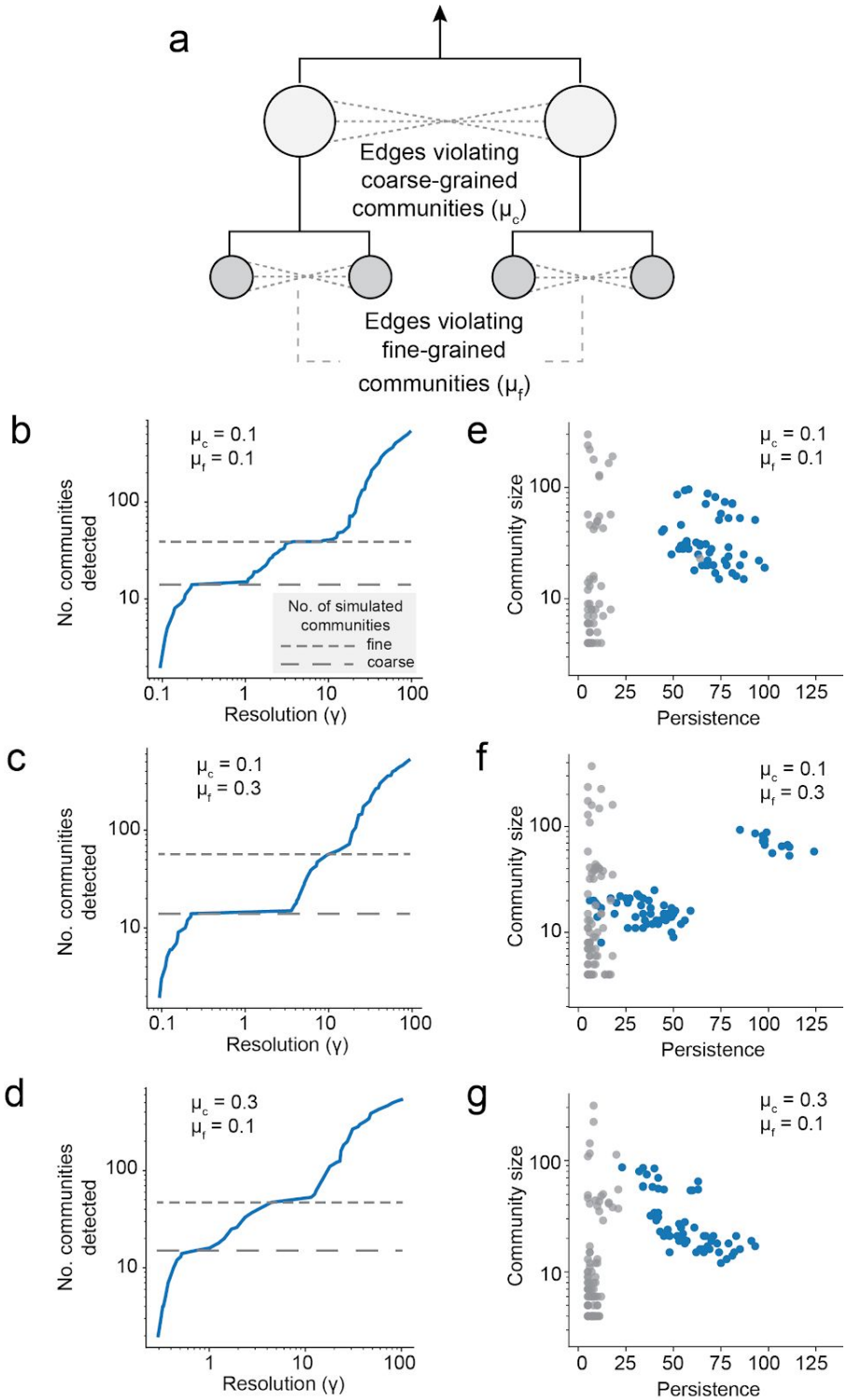
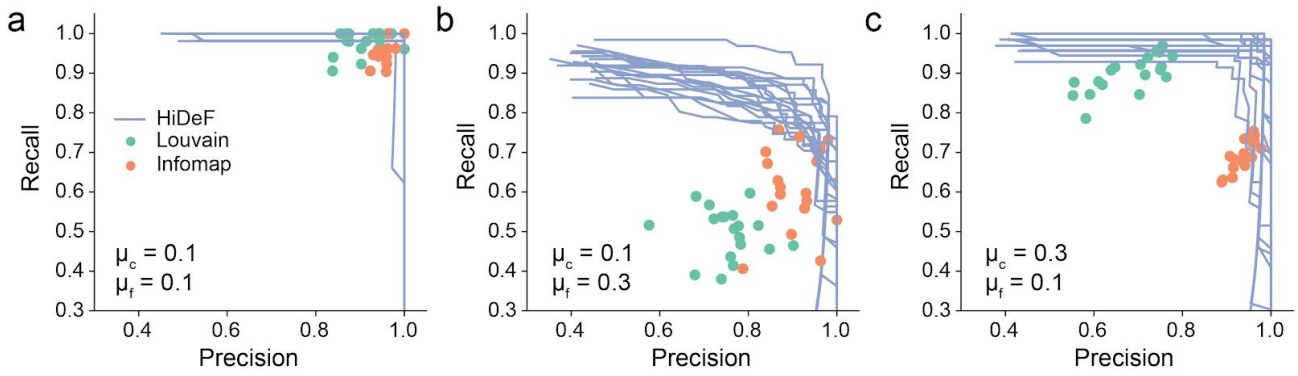

a
NeXO Data - yeast
Recovered GO term (%) Cellular Component
Top N communities in model
HiDeF
HC.1
HC.2
NeXO
b
STRING - human
Recovered GO term (%) Cellular Component
Top N communities in model
HiDeF
HC.1
HC.2
c
BioPlex
Coexpr-GEO
STRING
Community size
d
Depth
BioPlex hierarchy
Depth
STRING hierarchy
Community size (no. proteins)
e
Homology-Clime
DIP
Reactome
STRING
PCnet
PID
OmniPath
ReactomeFI
GeneMANIA
Coexpr-GEO
PathwayCommons
BIND
MultiNet
InBioMap
GIANT
STRING-t
BioPlex
IntAct
ConsensusPathDB
BioGRID
HumanInteractome
HumanNet
HPRD
HINT
Cancer-Achilles
IRefIndex
Mentha
Depth
No. nodes
No. edges
f
To larger-scale systems
g
HLX
EN1
TLE2
RIPPLY1
TLE1
RIPPLY2
MFAP4
TLE4
VENTX
TBX20
TLE3
TLE5
HES6
SLC25A21
SIX6
Interact with SARS-COV-2:NSP13

Fig. 3. Hierarchical community structure of protein networks. a-b, Percent recovery of cellular components documented in GO by four community detection methods (colored traces) versus number of top communities examined. A community is said to recover a cellular component if Jaccard similarity > 0.1 for the two sets of proteins. Communities ranked in descending order of score for each community detection tool (**Methods**). A yeast network[20] and the human STRING network[27] were used as the inputs of **a** and **b**, respectively. HC.1 and HC.2 represent UPGMA hierarchical clustering of pairwise distances generated by the Mashup and DSD tools[21,22], respectively. **c,** Distributions of community sizes (x-axis, number of proteins) for three human protein networks: BioPlex 2.0[25], Coexpr-GEO[26], and STRING[27]. **d,** Community hierarchies identified for BioPlex 2.0 (upper) or STRING (lower) databases. Vertex sizes and colors indicate the number of proteins in the corresponding communities. **e,** Twenty-seven public databases of protein-protein interaction networks were analyzed by HiDeF and profiled by the maximum depths of their resulting hierarchies (y axis), which do not correlate with their total sizes (numbers of proteins, x axis; numbers of edges, color bar). **f,** Convergence of communities into multiple super-systems. A community of mitogen-activated protein kinases and dual-specificity phosphatases (purple, center) participates in three distinct larger communities involving separate functions related to RAS pathways (green), sodium channels (pink), and acting capping (blue). The corresponding hierarchical relationships of these communities are depicted at lower right. The source network is Reactome[48]. **g,** A community of interacting human proteins targeted by the SARS-COV-2 viral protein Nsp13 (**Methods**). Direct interactors of Nsp13 (TLE1, TLE3, TLE5) are shown in orange.

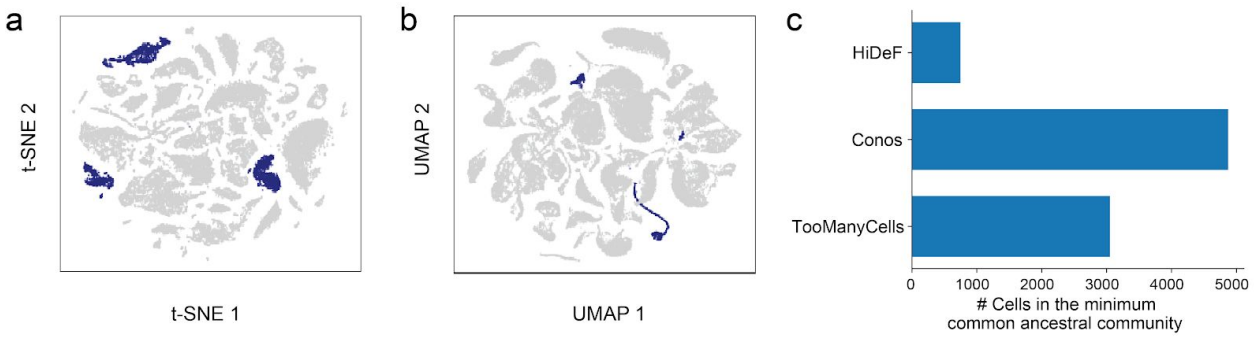
Supplementary Figures



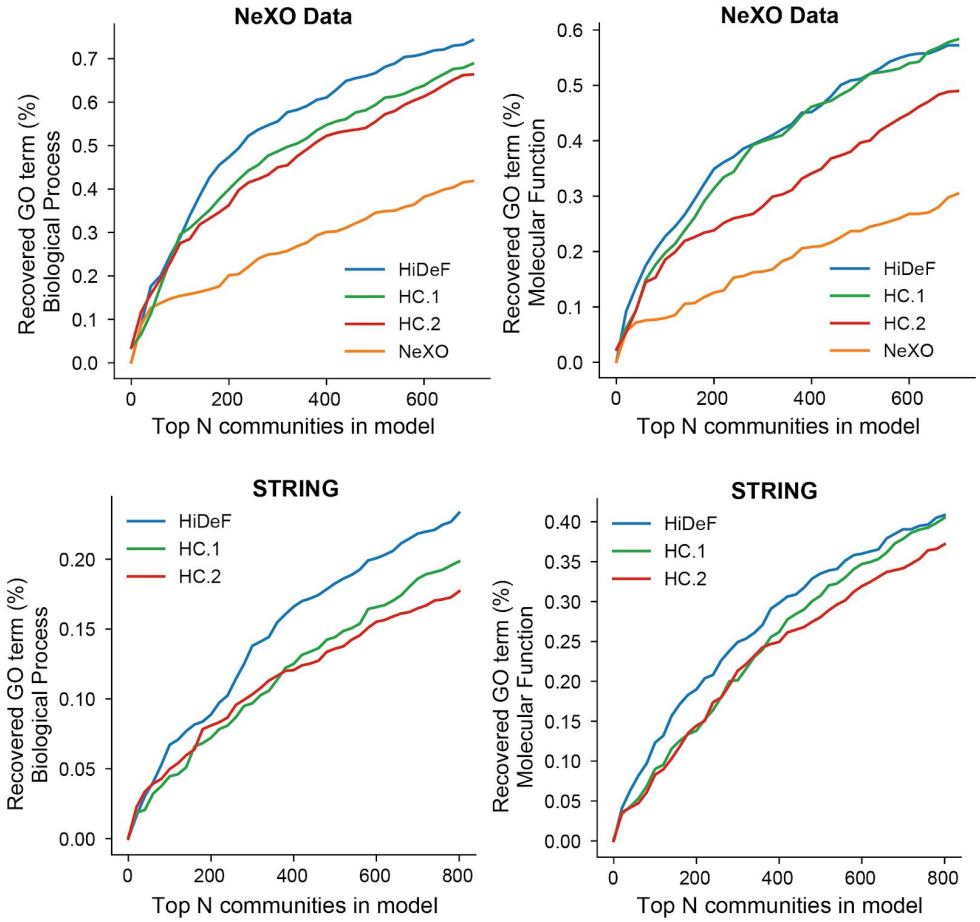
Supplementary Fig. 1. Exploring simulated networks. a, The LFR generative model was used to simulate networks with 1000 vertices and average degree 10 (**Methods**). The simulation included two layers of communities, "coarse" (10-20 communities, 50-100 vertices per community) and "fine" (25-200 communities, 5-40 vertices per community), with each fine community nested within a coarse community. Two "mixing parameters" μ_c and μ_f controlled the amount of noise, by setting the fraction of edges violating the coarse and fine community structures, respectively. **b-d,** HiDeF analysis of three simulated networks created with different mixing parameters: low balanced noise (b); increased noise in fine communities (c); and increased noise in coarse communities (d). Each plot shows the number of identified communities (y axis) as the resolution is progressively scanned (x axis). The number of communities increases with the resolution parameter, with plateaus matching the actual numbers of coarse and fine communities in the simulated network (dashed lines). Note that the sizes of the plateaus (i.e. the extent of community "persistence", see text) are affected by the mixing parameters. **e-g**, Companion plots to panels (b-d). Points represent identified communities, delineated by size (y axis) and persistence (x axis). Blue/gray point colors indicate a match/non-match to a true community in the simulated network (Jaccard similarity > 0.75). Note that when noise is low (e), the highest persistence communities correctly recover simulated communities with near-perfect accuracy, e.g. for persistence threshold >20.



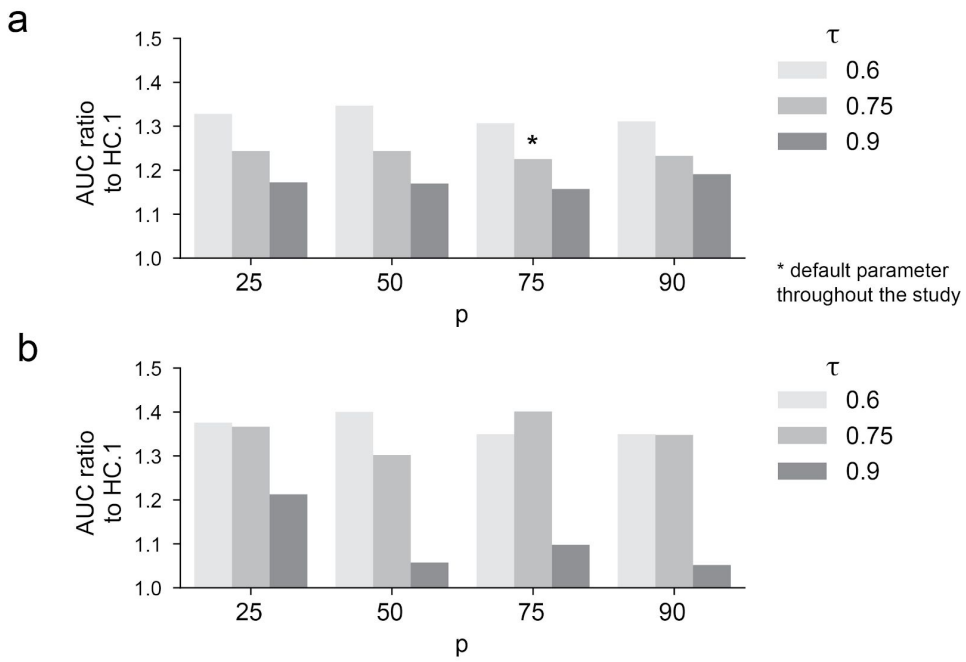
Supplementary Fig. 2. Comparison of methods in recovery of simulated communities. HiDeF is compared with the Louvain and Infomap algorithms[41,42], with Louvain and Infomap fixed at their default single resolutions (**Methods**). The three precision-recall plots (**a-c**) compare the performance of the three algorithms in recovering simulated communities at different settings of the coarse/fine mixing parameters (see Supplementary Fig. 1). The communities returned by HiDeF are ordered by persistence to produce a precision-recall curve of community recovery (blue, Jaccard similarity > 0.75), whereas Louvain and Infomap generate results with fixed precision and recall (green and orange points, respectively). The results for 20 simulated networks are overlaid on the same plots.



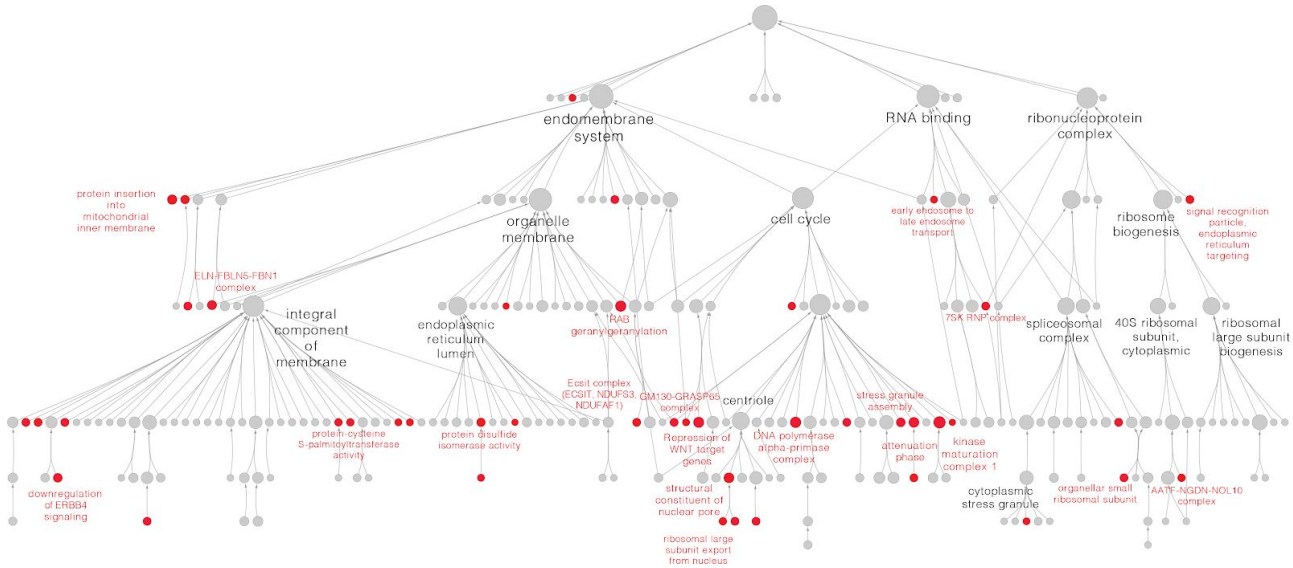
Supplementary Fig. 3. Example cell types captured by HiDeF but not by other approaches. a, t-SNE projection of all cells, with the epidermal cell type highlighted (blue). **b,** UMAP projection of all cells, with the hepatocyte cell type highlighted (blue). **c**. Distances between astrocyte and neuron communities in the cell-type hierarchies generated by HiDeF, Conos, or TooManyCells. HiDeF identifies a specific super-community joining both cell types (<1000 cells), whereas such a specific community is not identified by Conos and TooManyCells.



Supplementary Fig. 4. Recovery of GO terms from community detection in protein networks. HiDeF and alternative methods were applied to build a hierarchy of protein communities from analysis of an integrated protein interaction network for budding yeast (Top: NeXO) or human (Bottom: STRING). The hierarchy of each method (colors) is scored by its recovery of GO terms (Jaccard similarity > 0.1; Left: Biological Process; Right: Molecular Function) as a function of the number of top-scoring protein communities examined. HC, Hierarchical Clustering based on either of two protein pairwise distance functions (Mashup and DSD)[21,22].



Supplementary Fig. 5. Robustness of GO term recovery to the choice of parameters. a, Using the performance analysis depicted in Fig. 3b, the Area Under Curve (AUC) was computed for different sets of HiDeF parameters (p, τ). This AUC was compared to that of HC.1 (hierarchical clustering of pairwise distances generated by Mashup[21]) to generate an equal number of communities (**Methods**). Note the ratio HiDeF AUC / HC.1 AUC is always higher than 1, indicating that the favorable performance of HiDeF is robust to parameter choice. As per Fig. 3b, the analysis was undertaken using the STRING network and the GO Cellular Component branch. **b**, Similar analysis with subsampling of network edges (in which a random 10% of network edges are removed prior to community detection at each resolution).



Supplementary Fig. 6. Hierarchy of communities of human proteins interacting with SARS-COV-2. A hierarchy generated by HiDeF (**Methods**) contains 252 communities of 1948 human proteins (communities are vertices, containment relations are edges, similar to Fig. 3d). Communities colored red are enriched (odds ratio > 1.5) for the 332 human proteins interacting with viral proteins of SARS-COV-2. Selected communities are labeled by gene set enrichment function provided in CDAPS (**Code and data availability**).