OXFORD

Briefings in Bioinformatics, 2024, **25(5)**, bbae399
https://doi.org/10.1093/bib/bbae399
Problem Solving Protocol

Unveiling hidden connections in omics data *via* pyPARAGON: an integrative hybrid approach for disease network construction

Muslum Kaan Arici[1] and Nurcan Tuncbag[2,3,4,*]

[1]Graduate School of Informatics, Middle East Technical University, Ankara 06800, Turkey
[2]Chemical and Biological Engineering, College of Engineering, Koc University, Istanbul 34450, Turkey
[3]School of Medicine, Koc University, Istanbul 34450, Turkey
[4]Koc University Research Center for Translational Medicine (KUTTAM), Koc University, Istanbul 34450, Turkey

*Corresponding author. Nurcan Tuncbag, Department of Chemical and Biological Engineering, College of Engineering, Koc University, Rumelifeneri Yolu, Sariyer 34450 Istanbul, Turkey. Tel.: +90-212-338-0925, E-mail: ntuncbag@ku.edu.tr

Abstract

Network inference or reconstruction algorithms play an integral role in successfully analyzing and identifying causal relationships between omics hits for detecting dysregulated and altered signaling components in various contexts, encompassing disease states and drug perturbations. However, accurate representation of signaling networks and identification of context-specific interactions within sparse omics datasets in complex interactomes pose significant challenges in integrative approaches. To address these challenges, we present pyPARAGON (PAgeRAnk-flux on Graphlet-guided network for multi-Omic data integratioN), a novel tool that combines network propagation with graphlets. pyPARAGON enhances accuracy and minimizes the inclusion of nonspecific interactions in signaling networks by utilizing network rather than relying on pairwise connections among proteins. Through comprehensive evaluations on benchmark signaling pathways, we demonstrate that pyPARAGON outperforms state-of-the-art approaches in node propagation and edge inference. Furthermore, pyPARAGON exhibits promising performance in discovering cancer driver networks. Notably, we demonstrate its utility in network-based stratification of patient tumors by integrating phosphoproteomic data from 105 breast cancer tumors with the interactome and demonstrating tumor-specific signaling pathways. Overall, pyPARAGON is a novel tool for analyzing and integrating multi-omic data in the context of signaling networks. pyPARAGON is available at https://github.com/netlab-ku/pyPARAGON.

Keywords: network reconstruction; graphlets; data integration; interactome

Introduction

Omics technologies provide a multidimensional view of the cell's functional mechanism, context-specific alterations in diseases or drug perturbations, and biological processes [1, 2]. As the omics data accumulate, integrating them accurately and translating them into interpretable knowledge remains challenging due to data sparsity, missing data points, and computational complexity [3–5]. Omic hits are sparsely connected in a reference interactome and carry noise from high-throughput outcomes [6, 7].

Recent methods utilizing learning- and network-based algorithms are on the rise to overcome these challenges and decode causal relations between omic entities [8–11]. Learning-based methods efficiently integrate multi-omic data to extract interpretable annotations such as pathways, reactions, and processes [12–14]. Also, network-based algorithms, including shortest paths [15], Steiner trees/forests [16, 17], and random walks [18], have been frequently used to construct specific networks by propagating omic hits [19, 20]. Network-based methods can uncover the most relevant interactions between a given set of proteins/genes by either inferring from a reference protein–protein interaction (PPI) network or reconstructing them [1, 21, 22]. These reference networks integrate numerous databases and datasets, disregarding experimental context across diverse cell types and states [23]. Thus, the network inference methods may suffer from false positive interactions. However, these methods eventually obtain a network model, which may represent the alterations in disease models or the effects of drug treatments with the help of topological and statistical features [24–29]. The benefit of using global and local network features (e.g. degree distribution, clustering coefficients) for propagation or inference [30, 31] is limited when this type of sparse data is elaborated [32, 33]. Therefore, the frequent subgraphs, known as network motifs in biological networks such as metabolic [34], regulatory [35, 36], and cellular signaling networks [37], can provide a more comprehensive insight into their functional impact in complex cellular networks [38]. Specific network motifs function in rewiring signaling cascades and regulating cellular signaling and information processing, including feedback and feedforward loops, which entail signaling adaptations [39], cell lineage [40], and cell dynamics and functions in tissue [41]. Small connected, non-isomorphic subgraphs, called *graphlets*, are over-represented in the reference interactome and associated with specific functions [42, 43]. Graphlet statistics solve several complex problems in this context, such as the comparison of biological networks,

Received: December 22, 2023. **Revised:** June 26, 2024. **Accepted:** August 7, 2024

delineating the functional organization of networks, discovering functionally related genes, regulatory interactions, and parameter tuning for network-based approaches [12, 32, 33, 44–47]. Another challenge is the presence of highly connected and multifunctional proteins, particularly hub proteins, which can bring nonspecific interactions to the resulting network models. Therefore, using network motifs, graphlets, or revealing modules can improve the context-specific aspects of the models [1, 25, 48].

In this study, we hypothesize that the utilization of network motifs, *in lieu* of pairwise connections among proteins, may provide a more accurate representation of signaling networks and mitigate the inclusion of nonspecific interactions. Therefore, we present pyPARAGON (PAgeRAnk-flux on Graphlet-guided network for multi-Omic data integratioN) that combines network propagation with graphlets to construct context-specific networks. We found that graphlets filter out nonspecific interactions and mitigate the dominance of highly connected nodes, thereby trimming the reference interactome. pyPARAGON, as a hybrid method, performed better than the selected state-of-the-art methods in the reconstruction of known cancer signaling pathways. We demonstrated the utility of pyPARAGON in patient stratification using a breast cancer dataset comprising 105 tumors and associated phosphoproteomic data. Our analysis unveiled tumor-specific signaling pathways for each patient group.

Methods

Overview of pyPARAGON as a hybrid network inference framework

Hybrid approaches can be more effective than relying on a single method alone when integrating different types of omic data [17]. The accuracy of reconstructed networks is highly dependent on the reference interactome quality [49, 50]. On one hand, including interactions with low confidence scores may lead to the identification of false positive proteins and interactions. On the other hand, highly connected proteins (i.e. hubs) may dominate the final network and obscure context-specific connections of proteins/genes. pyPARAGON copes with these challenges in two independent steps. First, graphlet search mitigates the dominance of hub nodes. Graphlets are small, connected subgraphs with a specific pattern of edges and are similar to network motifs representing recurring patterns [42, 43]. Additionally, pyPARAGON calculates the flux value by multiplying a node's propagation score with the confidence score of its interaction and normalizing it with its degree. In this way, pyPARAGON prioritizes the high confidence scores and associated nodes while penalizing the highly connected nodes. pyPARAGON has three steps (Fig. 1A): (i) graphlet-guided network (GGN) construction; (ii) propagation and edge scoring *via* the Personalized PageRank (PPR) algorithm and flux calculation; (iii) preserving the edges in GGN with high scores and filtering out the rest.

In general, state-of-the-art methods use an immediate edge between two nodes in the reference network and node-based features (e.g. degree, betweenness, closeness, and eigenvector centralities). The GGN construction step of pyPARAGON goes further by following an unsupervised approach to identify a core region in the reference interactome by combining significantly frequent graphlets composed of 2-, 3-, and 4-nodes (Fig. 1B). In omics-based network construction, direct connections between the genes/proteins of interest are often sparse, and intermediate nodes are required to connect them and form a coherent network structure. Thus, we constrained that graphlets having more than two nodes may have an intermediate node. Intermediate nodes are the ones that have the highest connections to the seed nodes (initial nodes) in the corresponding graphlet (Fig. S1A).

In the second step, the PPR algorithm propagates signals from seed nodes across the reference interactome. Node weights after propagation, their degrees, and edge confidence scores are combined in a single function to calculate edge fluxes [51]. If the reference interactome is an unweighted graph, pyPARAGON sets a default score of 1.0 for all edges. Similarly, if seed nodes do not have weights, pyPARAGON assigns them a default value of 1.0. In this function, the degree component penalizes highly connected proteins that are nonspecifically present in the resulting subnetworks. In the final step, we map edges with flux scores to GGN to obtain a context-specific network (Fig. 1C). To simplify biological interpretation, pyPARAGON additionally uncovers modules, corresponding to network communities, which function in specific biological processes or pathways (Fig. 1D). Based on network topology, the Louvain community detection method divides inferred subnetworks into small modules [52]. Then, using a hypergeometric test, pyPARAGON discovers context-specific annotations [53]. In this way, we reveal not only hidden connections between initial nodes but also significant context-specific modules.

Network inference via PageRank-flux on graphlet-guided network

We used 2-, 3-, and 4-node-graphlets (G_0, G_1, G_2, …, G_8, shown in Fig. S1A), which are small non-isomorphic subgraphs. An isomorphism of graphlets between two subgraphs, $X(V_X, E_X)$ and $Y(V_Y, E_Y)$, is defined with bijections between V_X and V_Y [42]. We searched the graphlets for an intermediate node in one of the highest-degree orbits and seed nodes in the remaining orbits. The reference network is $R(V_R, E_R, c(e))$, where V_R, E_R, and $c(e)$ are node set, undirected edge set, and their confidence scores, respectively. Similarly, we calculated the frequencies of graphlets in 100 permuted networks, recruiting the same seed node set. To prepare permutated networks, we randomly swapped two edges between four different nodes so that the network topology and the number of interactions of the reference interactome could be used for statistical analysis [21]. We compared the graphlet frequencies in the reference and permuted networks with a z-test ($P < 0.05$, z-score > 1.65). The union of significant graphlets constructs the GGN, $G(V_G, E_G)$, where $G \subseteq R$.

The PPR algorithm calculates the probability of being at each node at a particular step in the reference networks according to Equation (1).

$$p_{(t+1)}(y) = \frac{1-\lambda}{N} + \lambda \sum_{y=x_i}^{x_N} \frac{p_t(x_i)}{deg\ (x_i)} \quad (1)$$

where $p(y)$ represents the probability of being at node y in the network at a particular time step t and λ is the damping factor. x_i represents each neighbor of y. $deg(x_i)$ is the degree of node x_i and N is the number of nodes [54, 55].

We modified the combined score formula described by Rubel and Ritz and introduced directed flux scores accordingly [51]. When combining two directional flux scores, we assigned the minimum flux score to the edge, instead of multiplying both directional flux scores. The scores are calculated for both directions ($f_{u \to t}$ and $f_{t \to u}$) by using Equations (2 and 3), where $u, t \in V_R$, and e is the edge between u and t, respectively. The negative

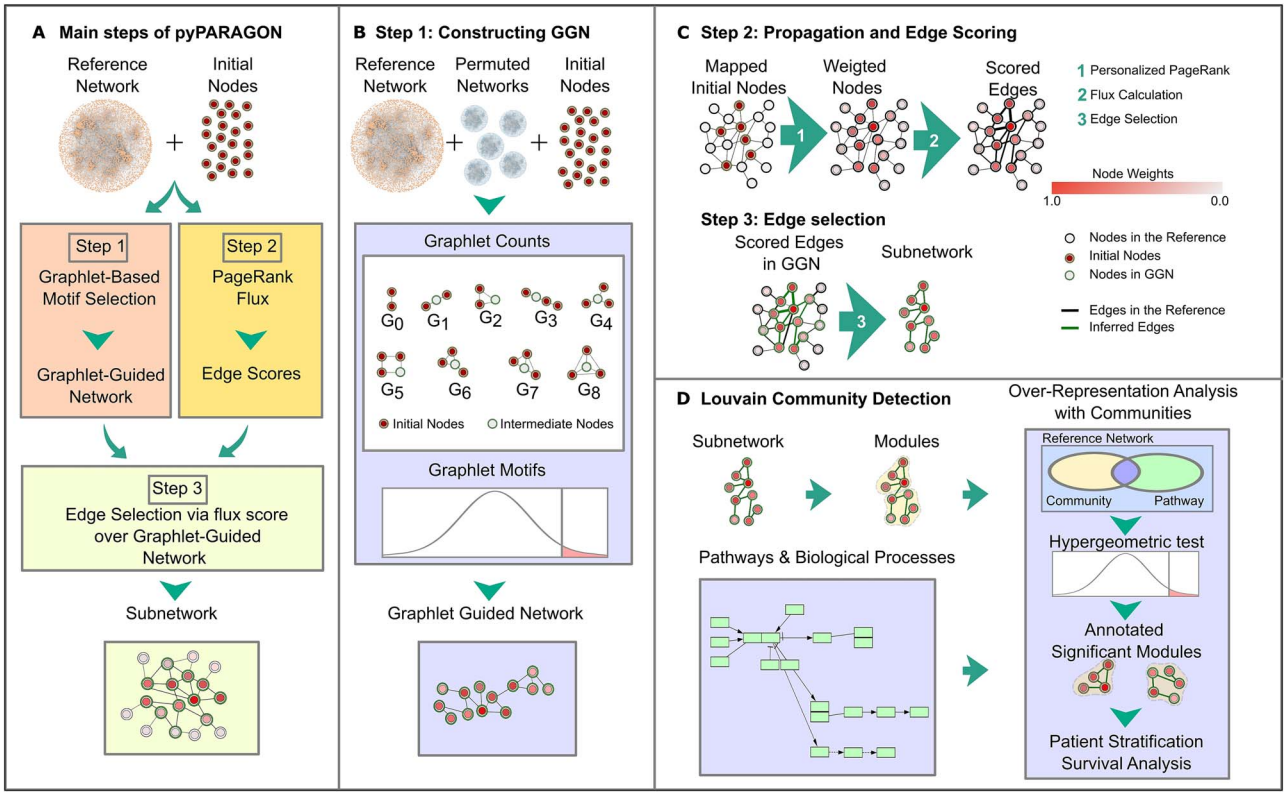

Figure 1. The overview of pyPARAGON. (A) pyPARAGON uses a reference network and a set of initial nodes (seed nodes) as the input. pyPARAGON has three steps: (i) GGN construction; (ii) edge scoring with PPR flux calculation; (iii) subnetwork inference using edge scores and GGN. (B) We investigated nine non-isomorphic graphlets (G_0–G_8) composed of 2, 3, and 4 nodes in the reference network and its 100 permuted networks. Except for G_0, each graphlet covers at least two seed nodes (red circles) and one intermediate node (white circles) that connects the seeds in the center of the orbit. We conducted a z-test to compare the frequency of graphlets in the reference and permuted networks. The union of significantly frequent graphlets constructs GGN. (C) By random walking from weighted initial nodes in the reference network, the PPR algorithm assigns weight to each node during propagation. Then, computed edge fluxes were used as the edge scores in the reference interactome. In the edge selection step, high-scoring ones in GGN construct the final subnetwork. (D) pyPARAGON employs the Louvain community detection method, based on network topology, to divide the inferred network into functional units. Significant biological processes and pathways in each module were found by using a hypergeometric test.

logarithm of minimum flux scores is used as a final edge score ($f(e)$) defined in Equation (4).

$$f_{u\to t}(u,t) = \frac{p(u).c(e)}{deg(u)} \quad (2)$$

$$f_{t\to u}(t,u) = \frac{p(t).c(e)}{deg(t)} \quad (3)$$

$$f(e) = -log\ \left(min\left(f_{u\to t}(u,t), f_{t\to u}(t,u)\right)\right) \quad (4)$$

We weighted the edge set of GGN, $G(V_G, E_G)$, with $f(e)$ where e_1, e_2, e_3, …, e_j, … $e_n \in E_G$, $1 \le j \le n$ and $f(e_{j-1}) > f(e_j) > f(e_{j+1})$. The total flux scores (F) in GGN are calculated as formulated in Equation (5).

$$F = \sum_{i=1}^{n} f(e_i) \quad (5)$$

Let τ $(0 \le \tau \le 1)$ represent the scaling factor describing the threshold percentage of F. We selected the edges by summing flux scores up to τxF (Equation (6)). In this way, we infer the context-specific network $C(V_C, E_C)$, where $E_C \subseteq E_G$ and $V_C \subseteq V_G$.

$$\tau xF = \sum_{i=1}^{j} f(e_i), 1 \le j \le n \quad (6)$$

Results

Network trimming via graphlets improves network inference

We used NetPath [56] as the benchmark dataset to reconstruct curated signaling pathways and assess the performance of pyPARAGON. In general, the performance of the methods is evaluated based on topological features, coverage of predicted nodes, and edges. As a result of screening all graphlets across the reference interactomes, we found G_2, G_5, G_6, G_7, and G_8 to be the most frequent graphlets (Fig. S1A). The frequency of direct interactions between input nodes (represented with G_0) is insignificant in the reference interactome; however, the direct interactions in a graphlet with at least three nodes are significant. For example, the direct interaction of seed nodes in G_2 gets more important in the presence of an intermediate node interacting with G_0. As to our observation, significant graphlets having at least one intermediate node to connect seeds provide more precision compared to including direct interactions between two seeds (i.e. G_0) in GGN.

Each available interactome has a specific evaluation and scoring scheme to integrate PPIs from different resources [49]. In this study, we used ConsensusPathDB [57], HIPPIE v2.2, and HIPPIE v2.3 [58], which have different topological features (Supplementary Methods, Table S1). The constructed GGN by pyPARAGON is a subnetwork of the reference interactome. When we separately compared the original interactomes and trimmed interactomes

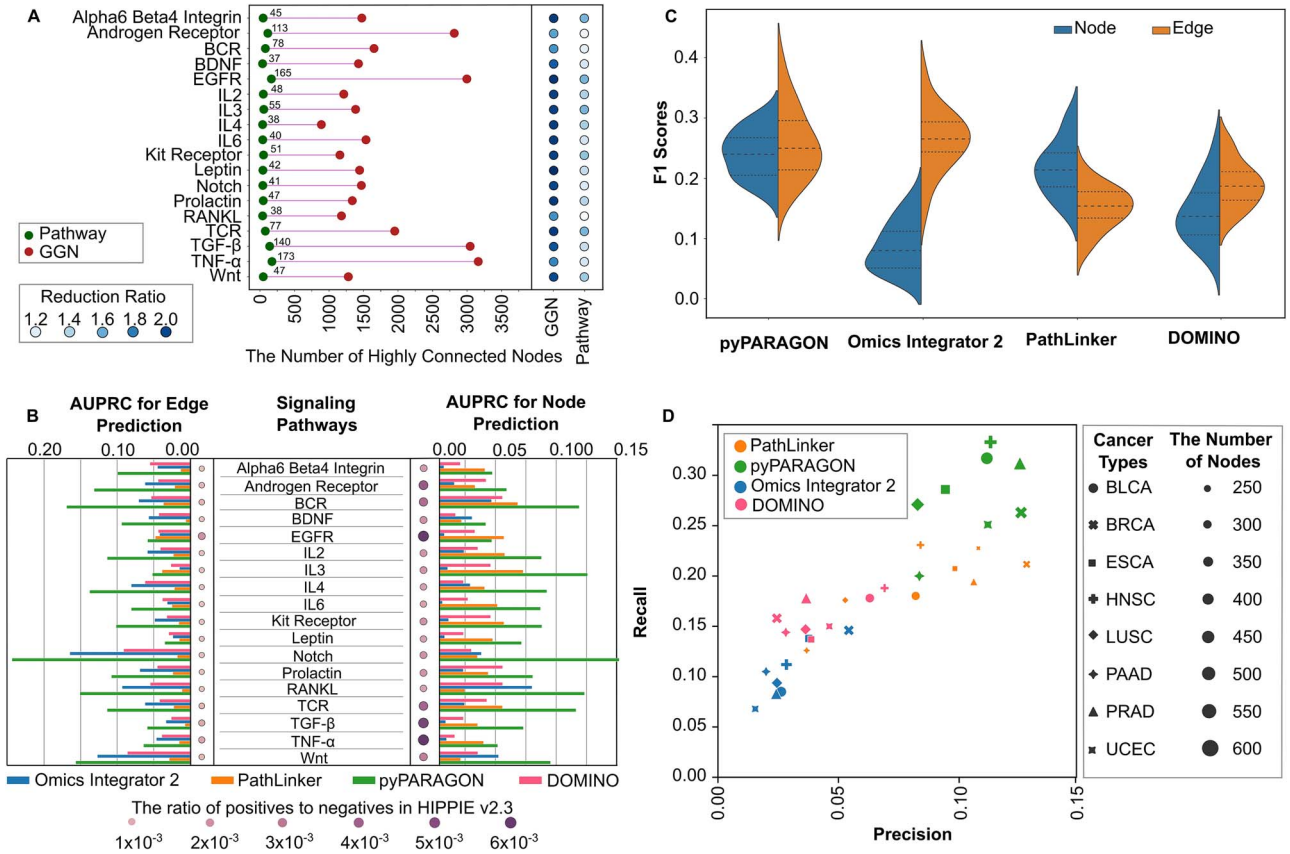

Figure 2. GGN trims reference interactome by removing some highly connected nodes and their non-specific interactions. (A) Highly connected nodes (3887) are defined with degrees within the top 20% of all proteins in HIPPIE interactome. On the left side, the presence of these nodes in GNNs and reconstructed pathways is shown for each signaling pathway (red and green dots, respectively). On the right side, the reduction ratio (RR) separately represents the decrease in the interaction number of highly connected nodes for GGN and pathways. (B) AUPRC of each tool (blue = OI2, orange = PL, red = DOMINO, and green = pyPARAGON; left panel for edge prediction and right panel for node prediction performance) in each pathway reconstruction. The ratio of positives to negatives in HIPPIE interactome is scaled at 10^{-3}, which demonstrates the sparsity of target nodes and edges in the reference network. (C) Distribution of F1-scores for each tool across 18 pathways is shown for node (blue) and edge (orange) predictions. (D) Performance evaluation in cancer-specific networks for eight distinct cancer types. Marker size represents network sizes, while recall and precision scores are shown on the x-axis and y-axis. The recall score represents the ratio of correctly predicted cancer driver genes in cancer-specific networks to the total number of drivers.

via GGN construction, we observed that their similarities significantly increase when GGNs are used (Fig. S1B). Another advantage of GGN construction is attenuating the dominance of the highly connected nodes with degrees within the top 20% of all nodes in the reference network [59]. Notably, highly connected proteins have numerous functions in the cellular processes known from prior knowledge and interactions in reference networks [60]. pyPARAGON puts a constraint on graphlets in that seed nodes must be connected via an intermediate node. The constructed GGN eventually consists of the topologically most important part of the reference interactome. Thus, nodes that are not present in any graphlet and not linked to the seed nodes, are trimmed. With this approach, we eliminate the most highly connected nodes (3887) in the reference network, HIPPIE v2.3 (Fig. 2A). The remaining highly connected nodes within GGN lost a large number of interactions that are not related to the given context (Fig. S1C). Also, the final GGNs preserve the properties of a scale-free network (Fig. S1D) [57], which characterize biological networks where the distribution of node degrees follows a power-law distribution [61, 62]. However, based on their degree exponent ($\gamma_{HIPPIE} = 1.45$ $R^2_{HIPPIE} = 0.88$, and $\gamma_{GGNs} = 2.40$, $R^2_{GGNs} = 0.84$), GGNs have stronger features of a scale-free network [23, 63, 64]. Scale-free networks are robust to the random loss of nodes, defined as error tolerance, and fragile to targeted worst-case attacks [65].

We compared the performance of pyPARAGON with three selected state-of-the-art tools, PathLinker [15], Omics Integrator 2 [16], and DOMINO [17]. PL computes multiple shortest paths between seed nodes and selects the highly-scored interactions by maximizing their own path scores to get rid of the noise. Omics Integrator 2 (OI2) and DOMINO solve the prize-collecting Steiner Forest problem. DOMINO statically selects the most relevant interactions and then solves the prize-collecting Steiner tree problem. These four approaches are compared based on their performance in inferring curated signaling pathways in NetPath. Since there is no definitive benchmark or ground truth for assessing tool performance, we relied on propagated nodes and predicted edges as evaluation criteria. Since the performance of tools is highly dependent on parameter sets, we inferred signaling pathways by applying various parameter sets in a grid for each network inference tool (Supplementary Methods). Then, we measured performance using the area under the precision-recall curve (AUPRC) to demonstrate how well each pathway's nodes and edges were recovered in the predicted networks. Bias toward hub proteins in the reference interactomes is a challenge in signaling pathway reconstruction that has been considered in all (pyPARAGON, OI2, PL, and DOMINO). Our analysis showed that pyPARAGON outperformed these tools at both the node and edge levels for inferring signaling pathways in all pathways of NetPath

(Fig. 2B). Furthermore, the proportion of positive and negative instances, based on both nodes and edges, indicated that our target nodes and edges are extremely scarce inside the reference network (Table S2).

Performance comparison of pyPARAGON with others was done in two directions: (i) node propagation, (ii) edge inference. We used the F1 score to compare them because it simultaneously represents precision and recall in one metric. The overall results show that pyPARAGON and PL are better at propagation, while pyPARAGON and OI2 are better at network inference (Fig. 2C). Due to the usage of significant modules in reference networks, DOMINO runs in a balance to propagate nodes and predict interactions. These modules are defined based on annotations in Gene Ontology. However, missing annotations in reference networks and databases may lead to low-performance scores. Thus, DOMINO exhibits lower F1 scores and AUPRC than pyPARAGON. Highly connected reference networks decreased the propagation ability of OI2 while providing more robust interactions than PL. On the other hand, PL propagated the seed node set more robustly due to considering multiple shortest paths but introducing many false positive interactions. Many seed nodes have a tendency to be connected by hub nodes as shortcuts due to biological networks being scale-free. Thus, multiple-shortest paths and random walk-based approaches may include more false positive interactions [19, 66]. However, penalizing highly connected nodes, e.g. the calculation of PageRank flux normalized the score in pyPARAGON or degree-dependent negative prizing in OI2, reduces false positive edges and improves F1-score in edge prediction.

Cancer driver genes provide a selective growth benefit and enhance cancer development via harboring specific mutations. Therefore, predicting and prioritizing genes likely to play a crucial role in oncogenesis are important tasks. We next utilized pyPARAGON to construct cancer network models to test its performance in detecting driver genes. The most frequently mutated genes in eight cancer types are utilized as seed nodes [67, 68]. We compared the nodes in the reconstructed networks with the known driver genes in IntOGen database [69] (Supplementary Methods), listed in pyPARAGON/Supplementaries folder on GitHub. Because we use 5-fold cross-validation, for each fold, we filtered out the common proteins between the seed list and known drivers and then reconstructed cancer type-specific networks with pyPARAGON, PL, DOMINO, and OI2.

Cancer-type-specific networks include both driver gene nodes and the intermediate nodes. However, not all cancer driver mutations, genes, and functionalities are known in the available datasets. Consequently, the accuracy of predicting driver genes in the absence of ground truth is the reason of low performance metrics, particularly in precision scores. As shown in Fig. 2D, the reconstructed network by pyPARAGON finds more driver genes and mostly achieves higher recall and precision than other methods in all cancer types (Supplementary Table S3). In PL-generated networks, precision scores are in general close to pyPARAGON. They are better than pyPARAGON for ESCA and BRCA. PL recruits the multiple shortest paths. Thus, intermediate nodes corresponded more to highly connected genes than specific driver genes with default parameters. In pyPARAGON, we use the PageRank algorithm to propagate seed nodes to the neighbors in the reference interactome, which helps obtain more candidate drivers. The prize-collecting Steiner tree algorithm terminates propagation at the seed nodes, which results in fewer driver genes being recovered in networks inferred by OI2 and DOMINO. In large reference networks, highly connected nodes generate network shortcuts instead of using signal cascades or motifs. Overall, pyPARAGON performs significantly better in cancer driver network prediction and can be further elaborated for tumor- or patient-specific network construction and network similarity-based comparisons.

Tumor-specific network inference unveil hidden commonalities across patients

We employed pyPARAGON to construct the specific networks for 105 breast cancer patients' tumors [68], where the seed nodes are significant phosphoproteins, as detailed in Supplementary Methods. It is important to note that pyPARAGON is also applicable to pan-cancer datasets. We consider the modules as functional subunits of networks that participate individually or jointly in context-specific molecular processes (Supplementary Methods). pyPARAGON uses hypergeometric tests to identify these active modules that are significantly over-represented in specific biological processes (Fig. S2). Figure 3 shows an example tumor-specific network composed of active modules that are significantly associated with KEGG pathways. All modules of the tumor-specific network are visualized and demonstrated in Fig. S3. Similarly, we identified active modules annotated with biological processes and then calculated the cosine similarities between patient-specific networks. Eventually, patient tumors are clustered into four groups (Fig. 4A, Supplementary Methods). Table S4 lists the 20 most common biological functions for each cluster. We uncovered critical biological processes in at least two clusters (Fig. 4B). In patient cluster-1, the most frequently associated biological process is the ubiquitin-dependent protein catabolic process, where several transcription factors (TFs) and enzymes are present. Ubiquitination (one of the post-translational modifications) is a multistep enzymatic process involved in the regulation of cancer metabolism [70]. The patients in cluster-2 frequently share the mitotic cytokinesis process. Cytokinesis defects increase chromosomal instability, vast genomic alteration, and point mutations, provoking intratumoral heterogeneity [71, 72]. The patient similarity network (Fig. S4) shows that only five patients in cluster-2 have higher similarity scores than 0.5 due to heterogeneity. Interestingly, we found that the nervous system development (NSD) process was the most frequent biological process in cluster-3. Breast cancer is the second most common cause of central nervous system metastasis after lung cancer [73]. In our datasets, just two patients had metastases. We found both patients with the NSD process in cluster-3. In cluster-4, the regulation process of actin cytoskeleton organization is significantly enriched which is relevant to cancer initiation, metastasis, and therapeutic responses. Rho GTPases, a family of the Ras GTPase superfamily, play a key role in this regulation [74].

Survival and KEGG pathway over-representation analysis revealed distinct molecular variations among clusters through tumor-specific signaling pathways. The Kaplan–Meier analysis and the log-rank test of the overall survival of patient clusters [75] showed that patients in cluster-4 have a significantly worse survival probability than cluster-1 (Fig. 4C, Fig. S5). Followingly, we annotated active modules with KEGG pathways to figure out their over-representation in these clusters (Fig. 4D) [76]. Cell cycle and PI3K/Akt signaling pathways are the most frequent pathways in the clusters, except cluster-2. Their presence in tumors in cluster-1 is more frequent than in cluster-4. Critical protein complexes in DNA replication, repair mechanisms, and mitosis; Cyclin-dependent kinases (CDKs) regulate the cell cycle pathway [77]. Dysregulation of CDKs in breast cancer mediates changes in cell cycle progression, driving uncontrolled cell proliferation [78]. Additionally, CDKs mediate crosstalk between PI3K/Akt and cell

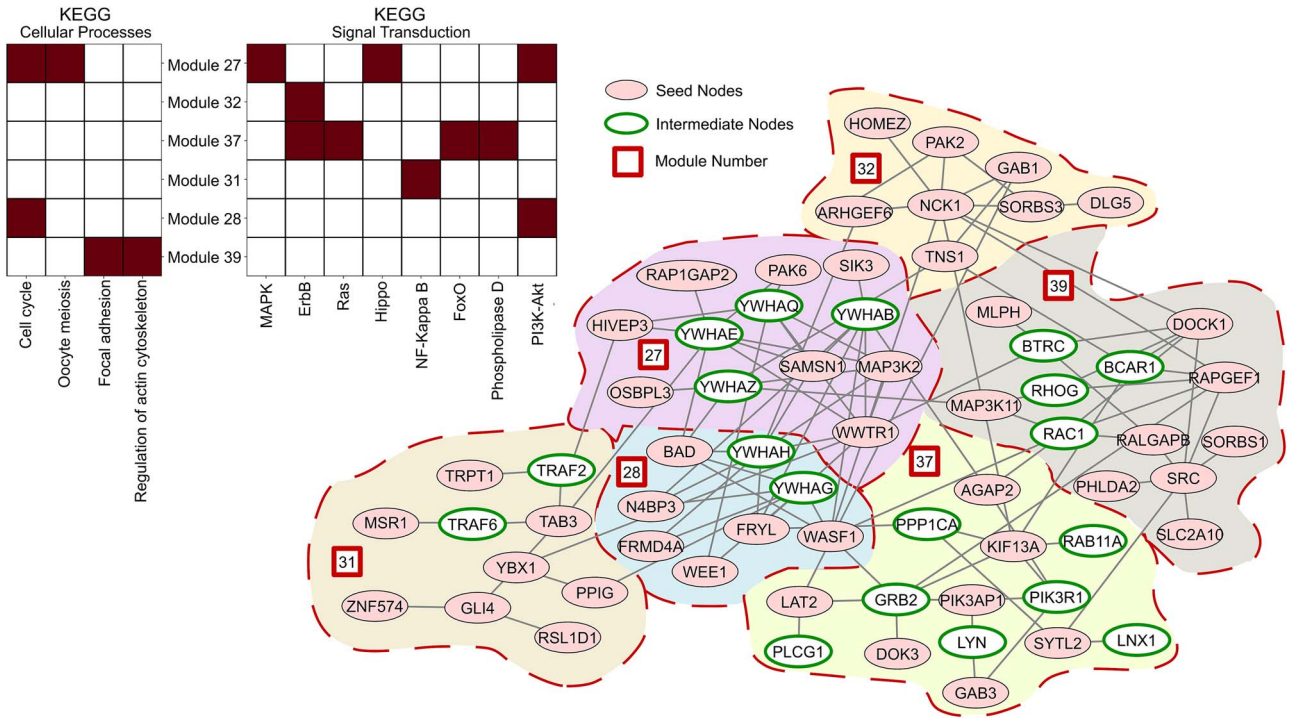

Figure 3. An example active module in tumor-specific network constructed by pyPARAGON (TCGA-A8-A079). Significantly phosphorylated proteins were used as the initial (seed) node-set (colored pink), and intermediate nodes predicted by pyPARAGON are in green circles. Active modules, bordered with dashed red lines and numbered within red boxes, are associated with at least one significantly overrepresented KEGG pathway. The pathways belonging to cellular processes and signal transduction are shown in the top left panel.

cycle pathways [79]. Thus, CDKs and their regulators have become prominent targets for drug development [80]. We identified 23 TFs from TRRUST database [81] regulating CDKs in tumor-specific networks (Table S5). Ninety drugs, authorized by the FDA in the Therapeutic Target Database (TTD) [82], target eight of these TFs (Table S6). On the other hand, the activation and inactivation in the components of the Hippo signaling pathway lead to drug resistance through rewiring in cell cycle cascades [83, 84]. The focal adhesion and Ras signaling pathways are significantly more frequent in cluster-4. The Ras signaling pathway is one of the key pathways for drug resistance owing to the bypassing of drug action mechanisms in the signaling network [85, 86]. In Fig. 4E, we demonstrated the module associated with the Ras-signaling pathway, where pyPARAGON linked phosphoproteins with intermediate nodes, including KRAS, NRAS, HRAS, RHOA, and RHOD. Next, we extracted 8297 drugs, 330 drug targets from TTD [82], and active modules were linked to 161 pathways that are found significantly enriched in 105 breast cancer patient-specific networks (on GitHub at pyPARAGON/Supplementaries/). An example of drugs connected to the active modules of patient A2-A0YD is given in Fig. S6. Adagrasib (MRTX849) and Sotorasib specifically target the Ras signaling-associated module. Both drugs are novel $KRAS^{G12C}$ inhibitors approved by the FDA [85, 87].

Discussion

In this work, we present pyPARAGON as a network-based multi-omic data integration tool. pyPARAGON simultaneously utilizes the most frequent graphlets covering omic hits and network propagation to construct context-specific networks. Network inference algorithms encounter challenges arising from sparse data and the complexity associated with the growing number of interactions within reference networks and potential false positives in the inferred subnetwork. In our study, by employing pyPARAGON, we mitigated the impact of highly connected nodes in the reference networks. pyPARAGON eliminates the interactions based on the calculated edge fluxes. Thus, the reference interactome is not prefiltered based on a confidence threshold. Additionally, pyPARAGON preserves scale-free properties inherent in biological networks in the constructed GGNs. We leveraged the PageRank flux calculation for edge prioritization and integrated GGNs to successfully construct context-specific networks. Additionally, driver networks that are inferred by pyPARAGON encompassed more precise and higher number of cancer drivers.

Although network inference algorithms can infer context-specific networks from large reference databases and experimental data, these networks prevent complete biological interpretations. Thus, module identification is crucial for gaining biological interpretations from network knowledge. Independent of the network inference, pyPARAGON is able to identify functional modules and their corresponding annotations. In tumor-specific networks, we integrated modules and different types of annotations, such as biological processes in GOA, pathways in KEGG, and drug knowledge in TTP. Additionally, we statistically explored interpretable biological knowledge in modules to disentangle tumor-specific pathways. These annotations can be valuable in revealing commonalities and differences across patients and drug perturbations.

Different molecular aberrations representing the context can induce identical disease outcomes [88–90]. pyPARAGON was developed as a general-purpose framework for integrating a given list of proteins/genes or other biological entities from any omic resources and an interactome for various contexts. In this study, we used omics data from CPTAC breast cancer samples [68] to infer context-specific networks with annotated modules. Patients were clustered based on the network-based similarity

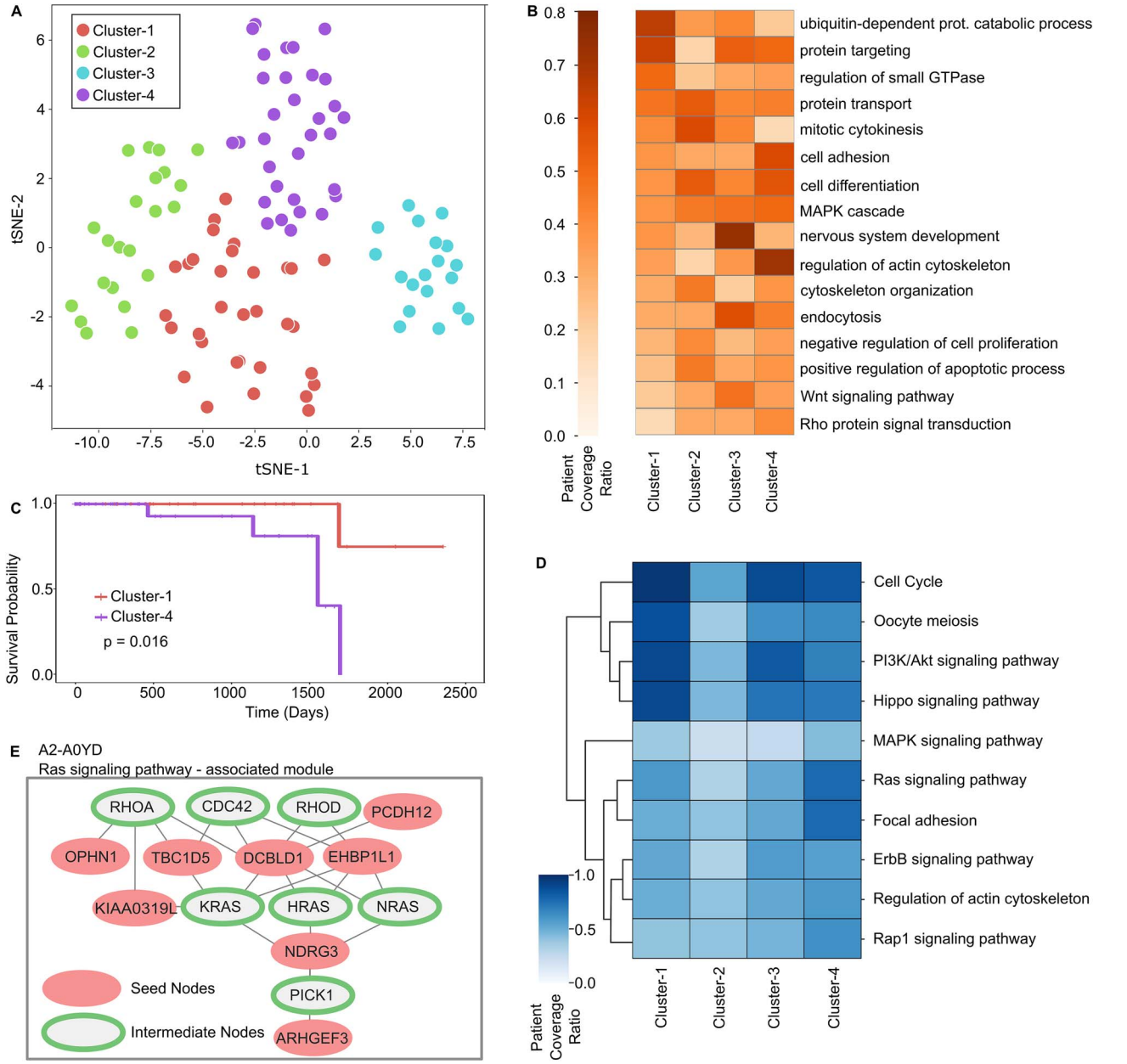

Figure 4. Stratification of tumors and associated biological processes with patient clusters. (A) 105 breast cancer tumors are stratified into four clusters based on their similarity of significant biological processes in their network modules: Cluster-1 (32 patients), cluster-2 (22 patients), cluster-3 (19 patients), and cluster-4 (32 patients). (B) Heatmap of patient coverage ratio for each cluster and significant process pairs. A biological process is included in the heatmap if it is enriched in at least two clusters. The patient coverage ratio represents the ratio of patients having the enriched biological process in the corresponding clusters. (C) Kaplan–Meier analysis shows the survival probabilities of cluster-1 (red) and cluster-4 (purple). The log-rank test ($P < 0.05$) statistically demonstrated that patients in cluster-4 have a significantly worse survival probability than cluster-1. (D) Heatmap shows significantly enriched KEGG pathways in active modules. (E) The example module of the A2-A0YD-specific network corresponding to the Ras signaling pathway is shown where seed nodes are red and intermediate nodes are green.

between the overrepresented biological processes identified with functional modules. We show that active modules with the same driver genes mediate various biological processes or pathways. Thus, pyPARAGON potentially enhances the identification of hidden functional commonalities beyond the common edges and nodes across context-specific networks.

pyPARAGON provides advantages for multi-omics data integration strategies as well. For instance, significantly expressed genes that are identified from transcriptomic data, can be used for identifying significantly active TFs and providing as a part of input seed sets. For the same example, the list of mutated proteins and significant TFs can be used together to form a seed node list. Similarly, enzymes or substrates associated with metabolomic hits can be extracted to be used as inputs, if available. Proteomics or phosphoproteomics hits can be used directly as the seed node set. In another independent case study, we determined differential TFs between cancer and autism spectrum disorders by identifying their common target pathways, using transcriptomics hits and frequent mutations [91]. Differential TFs in disease-specific networks demonstrated how rewired signaling mechanisms alter disease phenotypes [91]. All these case studies proved that pyPARAGON is capable of integrating omic datasets via

networks. pyPARAGON can be used to integrate various datasets, including the data from Pan-Cancer Atlas [92], the Cancer Cell Line Encyclopedia (CCLE) [93], the Genomics of Drugs Sensitivity in Cancer (GDSC) [94], and the LINCS [95], to reveal new biological insights in complex diseases, and drug perturbations.

Recent network inference methods, such as the SWEET tool [23], aim to construct sample-specific networks for individual samples [96–98]. pyPARAGON is highly modular, and the users can provide a custom reference interactome as input. For example, an aggregated interactome generated by SWEET to cover all phenotypic alterations across the entire sample set can be given as the reference. Then, pyPARAGON can infer a final subnetwork for each sample.

Despite the success of integrative approaches, including pyPARAGON, there is still significant potential for further enhancement.. Notably, network-based methods strongly depend on the features and coverage of reference networks [99]. As a result of incomplete knowledge in large reference interactomes, protein complexes tend to form more topological modules than metabolic pathways [100, 101]. Thus, generic biological processes, such as transcription and replication, can be found more frequently in inferred networks. Thus, biological interpretations of context-specific networks are challenging through causal relations, modules, and biological processes. Additionally, some network-based methods cannot handle the alternative copies of individual hits e.g. various protein isoforms and different post-translational modifications of a protein. Despite delivering more specific functions, this information is generalized and potentially lost in the network.

Extended integrations in reference networks and highly connected nodes have become a prominent challenge in recent network inference tools based on belief propagation [102, 103], random walks [18, 104], the prize-collecting Steiner Forest [16, 105, 106], heat diffusion [107, 108], and shortest path algorithms [15, 109]. Here, graphlets were deployed in our approaches for network trimming. In pathway reconstruction and the inference of context-specific networks, we compared our method with three popular tools: PL, OI2, and DOMINO. Hub proteins may dominate the inferred network with unrelated interactions. The prize-collecting Steiner Forest algorithm penalizes hubs based on the number of interactions. Similarly, the flux calculation in pyPARAGON is a countermeasure against the curse of hubs beyond scoring interactions. OI2 and pyPARAGON work better at predicting interactions. Regarding the identification of associated genes, our tool outperformed the other tools. In the PL algorithm, highly connected nodes further diminish the shortest paths between seed nodes. OI2 early terminates the propagation of the seed nodes in a large reference network. However, the PageRank algorithm in pyPARAGON propagates the seed nodes before network inference, independent of GGN. Thus, pyPARAGON optimizes the inference of interactions and the propagation of seed nodes in the network.

In conclusion, we released pyPARAGON as a novel tool, which infers context-specific networks by using graphlets and network propagation. pyPARAGON can infer a network from the omic datasets and potentially predict context-specific biomarkers, drugs, and therapeutic targets. For downstream analysis, communities in the network can potentially identify mechanistic molecular relations in complex and rare diseases. Here, pyPARAGAON integrated bulk omic data for static tumor-specific network models. The next version of pyPARAGON will be an extension that incorporates omic data at the single-cell level to elucidate cell-type specific interactions.

Key Points

- pyPARAGON combines graphlets with network propagation using the PPR algorithm. This is followed by interaction selection based on edge flux calculation, effectively addressing challenges in network modeling such as the inclusion of false positive proteins/genes and interactions, as well as accounting for the dominance of hubs and obscure context-specific relationships.
- pyPARAGON is an open-source method that offers easy accessibility and can be run in local environments. This feature provides a significant advantage for research groups interested in omic data integration.
- In constructing cancer signaling pathways and identifying cancer driver networks, pyPARAGON outperforms other state-of-the-art approaches in terms of node propagation and edge inference.
- We found that network trimming through graphlets plays a crucial role in improving the performance of network inference.
- pyPARAGON can construct tumor-specific networks, revealing hidden commonalities across tumors.

Acknowledgements

The results shown here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

Supplementary data

Supplementary data are available at *Briefings in Bioinformatics* online.

Conflict of interest: None declared.

Funding

N.T. was supported by the National Leader Researchers Program of The Scientific and Technological Research Council of Türkiye (TÜBİTAK) under the project number 121C292. M.K.A was supported by the TÜBİTAK 2211-A National Graduate Scholarship Program.

Data availability

The results shown here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. The input data and source codes to reproduce the results are available at https://github.com/netlab-ku/pyPARAGON. The list of cancer driver genes was retrieved from Integrative Onco Genomics (intOGen): https://www.intogen.org. The reference PPI networks, HIPPIE (v2.2 and v2.3), and ConsensusPathDB were retrieved from https://cbdm-01.zdv.uni-mainz.de/~mschaefer/hippie/download.php and http://cpdb.molgen.mpg.de, respectively.

References

1. Liu A, Trairatphisan P, Gjerga E. *et al.* From expression footprints to causal pathways: contextualizing large signaling

networks with CARNIVAL. *NPJ Syst Biol Appl* 2019;**5**:40. https://doi.org/10.1038/s41540-019-0118-z.
2. Ross KE, Huang H, Ren J. *et al.* IPTMnet: integrative bioinformatics for studying PTM networks. *Methods Mol Biol* 2017;**1558**:333–53. https://doi.org/10.1007/978-1-4939-6783-4_16.
3. Heumos L, Schaar AC, Lance C. *et al.* Best practices for single-cell analysis across modalities. *Nat Rev Genet* 2023;**24**:550–572. https://doi.org/10.1038/s41576-023-00586-w1-23.
4. Boehm KM, Khosravi P, Vanguri R. *et al.* Harnessing multimodal data integration to advance precision oncology. *Nat Rev Cancer* 2022;**22**:114–26. https://doi.org/10.1038/s41568-021-00408-3.
5. Rautenstrauch P, Vlot AHC, Saran S. *et al.* Intricacies of single-cell multi-omics data integration. *Trends Genet* 2022;**38**:128–39. https://doi.org/10.1016/j.tig.2021.08.012.
6. Aalto A, Viitasaari L, Ilmonen P. *et al.* Gene regulatory network inference from sparsely sampled noisy data. *Nat Commun* 2020;**11**:3493. https://doi.org/10.1038/s41467-020-17217-1.
7. Ren M, Pokrovsky A, Yang B. *et al.* SBNet: sparse blocks network for fast inference. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, UT, USA; 2018, 8711–20. https://doi.org/10.1109/CVPR.2018.00908.
8. Demirel HC, Arici MK, Tuncbag N. Computational approaches leveraging integrated connections of multi-omic data toward clinical applications. *Mol Omics* 2022;**18**:7–18. https://doi.org/10.1039/D1MO00158B.
9. Tong L, Mitchel J, Chatlin K. *et al.* Deep learning based feature-level integration of multi-omics data for breast cancer patients survival analysis. *BMC Med Inform Decis Mak* 2020;**20**:225. https://doi.org/10.1186/s12911-020-01225-8.
10. Kim SY, Jeong H-H, Kim J. *et al.* Robust pathway-based multi-omics data integration using directed random walks for survival prediction in multiple cancer studies. *Biol Direct* 2019;**14**:8. https://doi.org/10.1186/s13062-019-0239-8.
11. Duan R, Gao L, Gao Y. *et al.* Evaluation and comparison of multi-omics data integration methods for cancer subtyping. *PLoS Comput Biol* 2021;**17**:e1009224. https://doi.org/10.1371/journal.pcbi.1009224.
12. Malod-Dognin N, Petschnigg J, Windels SFL. *et al.* Towards a data-integrated cell. *Nat Commun* 2019;**10**:805. https://doi.org/10.1038/s41467-019-08797-8.
13. Shah HA, Liu J, Yang Z. *et al.* DeepRF: a deep learning method for predicting metabolic pathways in organisms based on annotated genomes. *Comput Biol Med* 2022;**147**:105756. https://doi.org/10.1016/j.compbiomed.2022.105756.
14. Costello Z, Martin HG. A machine learning approach to predict metabolic pathway dynamics from time-series multi-omics data. *NPJ Syst Biol Appl* 2018;**4**:19. https://doi.org/10.1038/s41540-018-0054-3.
15. Ritz A, Poirel CL, Tegge AN. *et al.* Pathways on demand: automated reconstruction of human signaling networks. *NPJ Syst Biol Appl* 2016;**2**:1–9. https://doi.org/10.1038/npjsba.2016.2.
16. Tuncbag N, Gosline SJC, Kedaigle A. *et al.* Network-based interpretation of diverse high-throughput datasets through the omics integrator software package. *PLoS Comput Biol* 2016;**20**:12:e1004879. https://doi.org/10.1371/journal.pcbi.1004879.
17. Levi H, Elkon R, Shamir R. DOMINO: a network-based active module identification algorithm with reduced rate of false calls. *Mol Syst Biol* 2021;**17**:e9593. https://doi.org/10.15252/msb.20209593.
18. Jagtap S, Çelikkanat A, Pirayre A. *et al.* BraneMF: integration of biological networks for functional analysis of proteins. *Bioinformatics* 2022;**38**:5383–9. https://doi.org/10.1093/bioinformatics/btac691.
19. Cowen L, Ideker T, Raphael BJ. *et al.* Network propagation: a universal amplifier of genetic associations. *Nat Rev Genet* 2017;**18**:551–62. https://doi.org/10.1038/nrg.2017.38.
20. Di Nanni N, Bersanelli M, Milanesi L. *et al.* Network diffusion promotes the integrative analysis of multiple omics. *Front Genet* 2020;**11**:106. https://doi.org/10.3389/fgene.2020.00106.
21. Reyna MA, Leiserson MDM, Raphael BJ. Hierarchical HotNet: identifying hierarchies of altered subnetworks. *Bioinformatics* 2018;**34**:i972–80. https://doi.org/10.1093/bioinformatics/bty613.
22. Silverman EK, Schmidt HHHW, Anastasiadou E. *et al.* Molecular networks in network medicine: development and applications. *Wiley Interdiscip Rev Syst Biol Med* 2020;**12**:1489. https://doi.org/10.1002/wsbm.1489.
23. Chen H-H, Hsueh C-W, Lee C-H. *et al.* SWEET: a single-sample network inference method for deciphering individual features in disease. *Brief Bioinform* 2023;**24**:bbad032. https://doi.org/10.1093/bib/bbad032.
24. Luna A, Siper MC, Korkut A. *et al.* Analyzing causal relationships in proteomic profiles using CausalPath. *STAR Protoc* 2021;**2**:100955. https://doi.org/10.1016/j.xpro.2021.100955.
25. Dugourd A, Kuppe C, Sciacovelli M. *et al.* Causal integration of multi-omics data with prior knowledge to generate mechanistic hypotheses. *Mol Syst Biol* 2021;**17**:e9730. https://doi.org/10.15252/msb.20209730.
26. Levitsky LI, Ivanov MV, Lobas AA. *et al.* IdentiPy: an extensible search engine for protein identification in shotgun proteomics. *J Proteome Res* 2018;**17**:2249–55. https://doi.org/10.1021/acs.jproteome.7b00640.
27. Unsal-Beyge S, Tuncbag N. Functional stratification of cancer drugs through integrated network similarity. *NPJ Syst Biol Appl* 2022;**8**:11. https://doi.org/10.1038/s41540-022-00219-8.
28. Nussinov R, Zhang M, Maloney R. *et al.* Mechanism of activation and the rewired network: new drug design concepts. *Med Res Rev* 2022;**42**:770–99. https://doi.org/10.1002/med.21863.
29. Dincer C, Kaya T, Keskin O. *et al.* 3D spatial organization and network-guided comparison of mutation profiles in Glioblastoma reveals similarities across patients. *PLoS Comput Biol* 2019;**15**:e1006789. https://doi.org/10.1371/journal.pcbi.1006789.
30. Hristov BH, Chazelle B, Singh M. uKIN combines new and prior information with guided network propagation to accurately identify disease genes. *Cell Systems* 2020;**10**:470–479.e3. https://doi.org/10.1016/j.cels.2020.05.008.
31. Ogris C, Hu Y, Arloth J. *et al.* Versatile knowledge guided network inference method for prioritizing key regulatory factors in multi-omics data. *Sci Rep* 2021;**11**:6806. https://doi.org/10.1038/s41598-021-85544-4.
32. Yaveroğlu ÖN, Malod-Dognin N, Davis D. *et al.* Revealing the hidden language of complex networks. *Sci Rep* 2014;**4**:4547. https://doi.org/10.1038/srep04547.
33. Wong SWH, Cercone N, Jurisica I. Comparative network analysis via differential graphlet communities. *Proteomics* 2015;**15**:608–17. https://doi.org/10.1002/pmic.201400233.
34. Sarajlić A, Malod-Dognin N, Yaveroğlu ÖN. *et al.* Graphlet-based characterization of directed networks. *Sci Rep* 2016;**6**:35098. https://doi.org/10.1038/srep35098.
35. Pereira L, Kratsios P, Serrano-Saiz E., *et al.* A cellular and regulatory map of the cholinergic nervous system of C. elegans. *Elife*; 2015:**4**:e12432.

36. Aibar S, González-Blas CB, Moerman T. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat Methods* 2017;**14**:1083–6. https://doi.org/10.1038/nmeth.4463.
37. Hu L, Zhang J, Pan X. *et al.* HiSCF: leveraging higher-order structures for clustering analysis in biological networks. *Bioinformatics* 2021;**37**:542–50. https://doi.org/10.1093/bioinformatics/btaa775.
38. Alon U. Network motifs: theory and experimental approaches. *Nat Rev Genet* 2007;**8**:450–61. https://doi.org/10.1038/nrg2102.
39. Adler M, Alon U. Fold-change detection in biological systems. *Curr Opin Syst Biol* 2018;**8**:81–9. https://doi.org/10.1016/j.coisb.2017.12.005.
40. Adler M, Korem Kohanim Y, Tendler A. *et al.* Continuum of gene-expression profiles provides spatial division of labor within a differentiated cell type. *Cell Syst* 2019;**8**:43–52.e5. https://doi.org/10.1016/j.cels.2018.12.008.
41. Mayer S, Milo T, Isaacson A. *et al.* The tumor microenvironment shows a hierarchy of cell-cell interactions dominated by fibroblasts. *Nat Commun* 2023;**14**:5810. https://doi.org/10.1038/s41467-023-41518-w.
42. Przulj N. Biological network comparison using graphlet degree distribution. *Bioinformatics* 2007;**23**:e177–83. https://doi.org/10.1093/bioinformatics/btl301.
43. Martin AJM, Dominguez C, Contreras-Riquelme S. *et al.* Graphlet based metrics for the comparison of gene regulatory networks. *PloS One* 2016;**11**:e0163497. https://doi.org/10.1371/journal.pone.0163497.
44. Windels SFL, Malod-Dognin N, Pržulj N. Graphlet Laplacians for topology-function and topology-disease relationships. *Bioinformatics* 2019;**35**:5226–34. https://doi.org/10.1093/bioinformatics/btz455.
45. Li Q, Milenkovic T. Supervised prediction of aging-related genes from a context-specific protein interaction subnetwork. *IEEE/ACM Trans Comput Biol Bioinform* 2022;**19**:2484–98. https://doi.org/10.1109/TCBB.2021.3076961.
46. Zhang L, Liu T, Chen H. *et al.* Predicting lncRNA-miRNA interactions based on interactome network and graphlet interaction. *Genomics* 2021;**113**:874–80. https://doi.org/10.1016/j.ygeno.2021.02.002.
47. Magnano CS, Gitter A. Automating parameter selection to avoid implausible biological pathway models. *npj Syst Biol Appl* 2021;**12**:7. https://doi.org/10.1038/s41540-020-00167-1.
48. Babur Ö, Luna A, Korkut A., *et al.* Causal interactions from proteomic profiles: molecular data meets pathway knowledge. *Patterns (N Y)* 2021;**2**:100257. https://doi.org/10.1016/j.patter.2021.100257.
49. Arici MK, Tuncbag N. Performance assessment of the network reconstruction approaches on various interactomes. *Front Mol Biosci* 2021;**8**:666705. https://doi.org/10.3389/fmolb.2021.666705.
50. Huang JK, Carlin DE, Yu MK. *et al.* Systematic evaluation of molecular networks for discovery of disease genes. *Cell Syst* 2018;**6**:484–495.e5. https://doi.org/10.1016/j.cels.2018.03.001.
51. Rubel T, Ritz A. Augmenting signaling pathway reconstructions. In: *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, Association for Computing Machinery, NY, USA*; 2020. https://doi.org/10.1145/3388440.3412411.
52. Blondel VD, Guillaume J-L, Lambiotte R. *et al.* Fast unfolding of communities in large networks. *J Stat Mech* 2008;**2008**:P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008.
53. Boyle EI, Weng S, Gollub J. *et al.* GO::TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics* 2004;**20**:3710–5. https://doi.org/10.1093/bioinformatics/bth456.
54. Langville AN, Meyer CD. A survey of eigenvector methods for web information retrieval. *SIAM Rev.* 2005;**47**:135–161. https://doi.org/10.1137/S0036144503424786.
55. Page L, Brin S, Motwani R. *et al. The PageRank Citation Ranking: Bringing Order to the Web*, Technical Report SIDL-WP-1999-0120, Stanford Digital Library Technologies Project 1998.
56. Kandasamy K, Sujatha Mohan S, Raju R. *et al.* NetPath: a public resource of curated signal transduction pathways. *Genome Biol* 2010;**11**:R3. https://doi.org/10.1186/gb-2010-11-1-r3.
57. Kamburov A, Herwig R. ConsensusPathDB 2022: molecular interactions update as a resource for network biology. *Nucleic Acids Res* 2022;**50**:D587–95. https://doi.org/10.1093/nar/gkab1128.
58. Alanis-Lobato G, Andrade-Navarro MA, Schaefer MH. HIPPIE v2.0: enhancing meaningfulness and reliability of protein-protein interaction networks. *Nucleic Acids Res* 2017;45:D408. https://doi.org/10.1093/nar/gkw985.
59. Jin G, Zhang S, Zhang X-S. *et al.* Hubs with network motifs organize modularity dynamically in the protein-protein interaction network of yeast. *PloS One* 2007;**2**:e1207. https://doi.org/10.1371/journal.pone.0001207.
60. Hu G, Wu Z, Uversky VN. *et al.* Functional analysis of human hub proteins and their interactors involved in the intrinsic disorder-enriched interactions. *Int J Mol Sci* 2017;**18**(12):2761. https://doi.org/10.3390/ijms18122761.
61. Panni S, Lovering RC, Porras P. *et al.* Non-coding RNA regulatory networks. *Biochim Biophys Acta Gene Regul Mech* 2020;**1863**:194417. https://doi.org/10.1016/j.bbagrm.2019.194417.
62. Khanin R, Wit E. How scale-free are biological networks. *J Comput Biol* 2006;**13**:810–8. https://doi.org/10.1089/cmb.2006.13.810.
63. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol* 2005;**4**:17. https://doi.org/10.2202/1544-6115.1128.
64. Barabási A-L, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nat Rev Genet* 2004;**5**:101–13. https://doi.org/10.1038/nrg1272.
65. Albert R, Jeong H, Barabasi AL. Error and attack tolerance of complex networks. *Nature* 2000;**406**:378–82. https://doi.org/10.1038/35019019.
66. Charmpi K, Chokkalingam M, Johnen R. *et al.* Optimizing network propagation for multi-omics data integration. *PLoS Comput Biol* 2021;**17**:e1009161. https://doi.org/10.1371/journal.pcbi.1009161.
67. Ellrott K, Bailey MH, Saksena G. *et al.* Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. *Cell Syst* 2018;**6**:271–281.e7. https://doi.org/10.1016/j.cels.2018.03.002.
68. Mertins P, Mani DR, Ruggles KV. *et al.* Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* 2016;**534**:55–62. https://doi.org/10.1038/nature18003.
69. Martamp F, Muiamp F, Deu-Pons J. *et al.* A compendium of mutational cancer driver genes. *Nat Rev Cancer* 2020;**20**:555–72. https://doi.org/10.1038/s41568-020-0290-x.
70. Cruz L, Soares P, Correia M. Ubiquitin-specific proteases: players in cancer cellular processes. *Pharmaceuticals* 2021;**14**:848. https://doi.org/10.3390/ph14090848.
71. Lens SMA, Medema RH. Cytokinesis defects and cancer. *Nat Rev Cancer* 2019;**19**:32–45. https://doi.org/10.1038/s41568-018-0084-6.

72. Li J, Dallmayer M, Kirchner T. *et al.* PRC1: linking cytokinesis, chromosomal instability, and cancer evolution. *Trends Cancer* 2018;**4**:59–73. https://doi.org/10.1016/j.trecan.2017.11.002.
73. Ben-Zion Berliner M, Yerushalmi R, Lavie I. *et al.* Central nervous system metastases in breast cancer: the impact of age on patterns of development and outcome. *Breast Cancer Res Treat* 2021;**185**:423–32. https://doi.org/10.1007/s10549-020-05959-x.
74. Haga RB, Ridley AJ. Rho GTPases: regulation and roles in cancer cell biology. *Small GTPases* 2016;**7**:207–21. https://doi.org/10.1080/21541248.2016.1232583.
75. Dudley WN, Wickham R, Coombs N. An introduction to survival statistics: Kaplan-Meier analysis. *J Adv Pract Oncol* 2016;**7**:91–100. https://doi.org/10.6004/jadpro.2016.7.1.8.
76. Kanehisa M, Furumichi M, Sato Y. *et al.* KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res* 2023;**51**:D587–92. https://doi.org/10.1093/nar/gkac963.
77. Wang Q, Bode AM, Zhang T. Targeting CDK1 in cancer: mechanisms and implications. *NPJ Precis Oncol* 2023;**7**:1–14. https://doi.org/10.1007/s12032-022-01748-2.
78. Ding L, Cao J, Lin W. *et al.* The roles of cyclin-dependent kinases in cell-cycle progression and therapeutic strategies in human breast cancer. *Int J Mol Sci* 2020;**21**:1960. https://doi.org/10.3390/ijms21061960.
79. Maidarti M, Anderson RA, Telfer EE. Crosstalk between PTEN/PI3K/Akt signalling and DNA damage in the oocyte: implications for primordial follicle activation, oocyte quality and ageing. *Cells* 2020;**9**:200. https://doi.org/10.3390/cells9010200.
80. Tang W, Lin C, Yu Q. *et al.* Novel medicinal chemistry strategies targeting CDK5 for drug discovery. *J Med Chem* 2023;**66**:7140–7161. https://doi.org/10.1021/acs.jmedchem.3c00566.
81. Han H, Cho J-W, Lee S. *et al.* TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Res* 2018;**46**:D380–6. https://doi.org/10.1093/nar/gkx1013.
82. Zhou Y, Zhang Y, Zhao D. *et al.* TTD: therapeutic target database describing target druggability information. *Nucleic Acids Res* 2024;**52**:1465–77. https://doi.org/10.1093/nar/gkad751.
83. Nussinov R, Tsai C-J, Jang H. *et al.* Oncogenic KRAS signaling and YAP1/β-catenin: similar cell cycle control in tumor initiation. *Semin Cell Dev Biol* 2016;**58**:79–85. https://doi.org/10.1016/j.semcdb.2016.04.001.
84. Zeng R, Dong J. The Hippo signaling pathway in drug resistance in cancer. *Cancer* 2021;**13**:318. https://doi.org/10.3390/cancers13020318.
85. Hallin J, Engstrom LD, Hargis L. *et al.* The KRAS inhibitor MRTX849 provides insight toward therapeutic susceptibility of KRAS-mutant cancers in mouse models and patients. *Cancer Discov* 2020;**10**:54–71. https://doi.org/10.1158/2159-8290.CD-19-1167.
86. Healy FM, Prior IA, MacEwan DJ. The importance of Ras in drug resistance in cancer. *Br J Pharmacol* 2022;**179**:2844–67. https://doi.org/10.1111/bph.15420.
87. Zhang SS, Nagasaka M. Spotlight on Sotorasib (AMG 510) for positive non-small cell lung cancer. *Lung Cancer* 2021;**12**:115–22. https://doi.org/10.2147/LCTT.S334623.
88. Peng J, Zhou Y, Wang K. Multiplex gene and phenotype network to characterize shared genetic pathways of epilepsy and autism. *Sci Rep* 2021;**11**:952. https://doi.org/10.1038/s41598-020-78654-y.
89. Riller Q, Rieux-Laucat F. RASopathies: from germline mutations to somatic and multigenic diseases. *Biom J* 2021;**44**:422–32. https://doi.org/10.1016/j.bj.2021.06.004.
90. Muñoz-Maldonado C, Zimmer Y, Medová M. A comparative analysis of individual RAS mutations in cancer biology. *Front Oncol* 2019;**9**:1088. https://doi.org/10.3389/fonc.2019.01088.
91. Yavuz BR, Arici MK, Demirel HC. *et al.* Neurodevelopmental disorders and cancer networks share pathways, but differ in mechanisms, signaling strength, and outcome. *NPJ Genom Med* 2023;**8**:37. https://doi.org/10.1038/s41525-023-00377-6.
92. Weinstein JN, Collisson EA, Mills GB. *et al.* The cancer genome atlas Pan-cancer analysis project. *Nat Genet* 2013;**45**:1113–20. https://doi.org/10.1038/ng.2764.
93. Barretina J, Caponigro G, Stransky N. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 2012;**483**:603–7. https://doi.org/10.1038/nature11003.
94. Yang W, Soares J, Greninger P. *et al.* Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res* 2013;**41**:D955–61. https://doi.org/10.1093/nar/gks1111.
95. Subramanian A, Narayan R, Corsello SM. *et al.* A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* 2017;**171**:1437–1452.e17. https://doi.org/10.1016/j.cell.2017.10.049.
96. Huttlin EL, Bruckner RJ, Navarrete-Perea J. *et al.* Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. *Cell* 2021;**184**:3022–3040.e28. https://doi.org/10.1016/j.cell.2021.04.011.
97. Ranzoni AM, Tangherloni A, Berest I. *et al.* Integrative single-cell RNA-seq and ATAC-seq analysis of human developmental hematopoiesis. *Cell Stem Cell* 2021;**28**:472–487.e7. https://doi.org/10.1016/j.stem.2020.11.015.
98. Quake SR. A decade of molecular cell atlases. *Trends Genet* 2022;**38**:805–10. https://doi.org/10.1016/j.tig.2022.01.004.
99. Kang Y, Xu Y, Wang X. *et al.* HN-PPISP: a hybrid network based on MLP-Mixer for protein-protein interaction site prediction. *Brief Bioinform* 2023;**24**:bbac480. https://doi.org/10.1093/bib/bbac480.
100. Mosca E, Bersanelli M, Matteuzzi T. *et al.* Characterization and comparison of gene-centered human interactomes. *Brief Bioinform* 2021;**22**:BBAB153. https://doi.org/10.1093/bib/bbab153.
101. Cheng F, Zhao J, Wang Y. *et al.* Comprehensive characterization of protein-protein interactions perturbed by disease mutations. *Nat Genet* 2021;**53**:342–53. https://doi.org/10.1038/s41588-020-00774-y.
102. Kirkley A, Cantwell GT, Newman MEJ. Belief propagation for networks with loops. *Sci Adv* 2021;**7**:17. https://doi.org/10.1126/sciadv.abf1211.
103. Korkut A, Wang W, Demir E. *et al.* Perturbation biology nominates upstream-downstream drug combinations in RAF inhibitor resistant melanoma cells. *Elife* 2015;**4**:e04640. https://doi.org/10.7554/eLife.04640.
104. Ietswaart R, Gyori BM, Bachman JA. *et al.* GeneWalk identifies relevant gene functions for a biological context using network representation learning. *Genome Biol* 2021;**22**:55.
105. Sychev ZE, Hu A, DiMaio TA. *et al.* Integrated systems biology analysis of KSHV latent infection reveals viral induction and reliance on peroxisome mediated lipid metabolism. *PLoS Pathog* 2017;**13**:e1006256. https://doi.org/10.1371/journal.ppat.1006256.
106. Dinstag G, Shamir R. PRODIGY: personalized prioritization of driver genes. *Bioinformatics* 2020;**36**:1831–9. https://doi.org/10.1093/bioinformatics/btz815.
107. Kuenzi BM, Ideker T. A census of pathway maps in cancer systems biology. *Nat Rev Cancer* 2020;**20**:233–46. https://doi.org/10.1038/s41568-020-0240-7.

108. Leiserson MDM, Vandin F, Wu HT. *et al.* Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat Genet* 2015;**47**: 106–14. https://doi.org/10.1038/ng.3168.

109. Licata L, Lo Surdo P, Iannuccelli M., *et al.* SIGNOR 2.0, the SIGnaling Network Open Resource 2.0: 2019 update. *Nucleic Acids Res* **2020**; 48:D504–10 https://doi.org/10.1093/nar/gkz949