Similarity network fusion for aggregating data types on a genomic scale

Bo Wang[1,5], Aziz M Mezlini[1,2], Feyyaz Demir[1,2], Marc Fiume[2], Zhuowen Tu[3], Michael Brudno[1,2], Benjamin Haibe-Kains[4,5] & Anna Goldenberg[1,2]

Recent technologies have made it cost-effective to collect diverse types of genome-wide data. Computational methods are needed to combine these data to create a comprehensive view of a given disease or a biological process. Similarity network fusion (SNF) solves this problem by constructing networks of samples (e.g., patients) for each available data type and then efficiently fusing these into one network that represents the full spectrum of underlying data. For example, to create a comprehensive view of a disease given a cohort of patients, SNF computes and fuses patient similarity networks obtained from each of their data types separately, taking advantage of the complementarity in the data. We used SNF to combine mRNA expression, DNA methylation and microRNA (miRNA) expression data for five cancer data sets. SNF substantially outperforms single data type analysis and established integrative approaches when identifying cancer subtypes and is effective for predicting survival.

Rapidly evolving technologies are making it progressively easier to collect multiple and diverse genome-scale data sets to address clinical and biological questions. For example, large-scale efforts by The Cancer Genome Atlas (TCGA) have already amassed genome, transcriptome and epigenome information for over 20 cancers from thousands of patients. The availability of such a wealth of data makes integrative methods essential for capturing the heterogeneity of biological processes and phenotypes, leading to, for example, the identification of homogeneous subtypes in breast cancer. Data-integration methods need to overcome at least three computational challenges: (i) the small number of samples compared to the large number of measurements; (ii) the differences in scale, collection bias and noise in each data set, and (iii) the complementary nature of the information provided by different types of data. Current integration approaches have yet to address all of these challenges together[1–4].
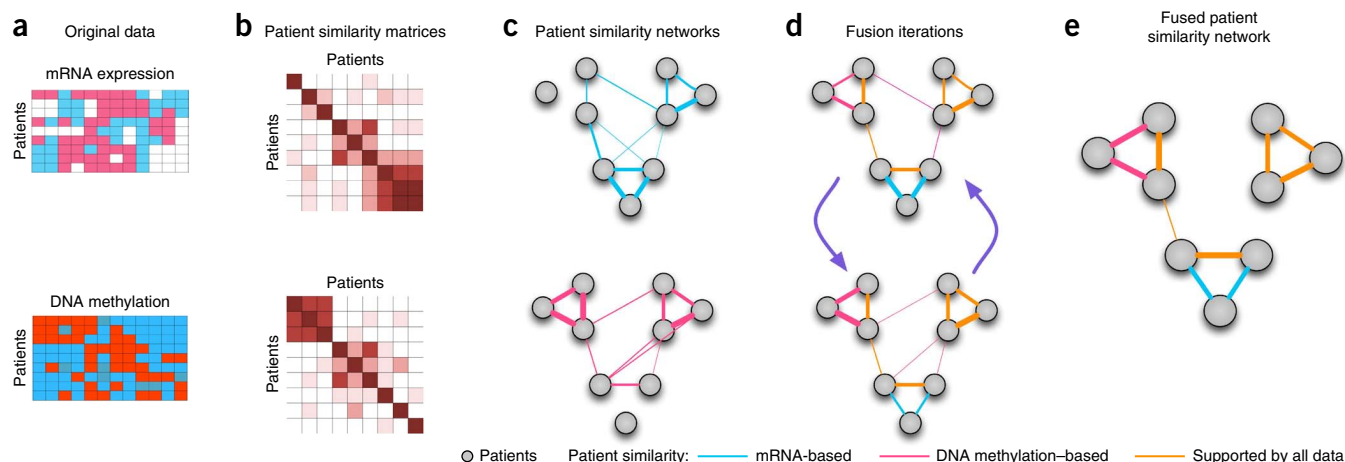
The simplest way to combine biological data is to concatenate normalized measurements from various biological domains, such as mRNA expression and DNA methylation, for each sample. Unfortunately, concatenation further dilutes the already low signal-to-noise ratio in each data type. To avoid this, a common strategy is to analyze each data type independently[2,3,5,6] before combining data. However, such independent analyses often lead to inconsistent conclusions that are hard to integrate. Another approach to increase signal is to preselect a set of important genes from each data source and use Consensus Clustering[1] to combine the data[3]. However, preselecting genes leads to a biased analysis,

and focusing only on common patterns can miss valuable complementary information. One recent machine-learning approach, iCluster[7], uses a joint latent variable model for integrative clustering. Though powerful, iCluster and related machine-learning approaches[4] do not scale to the full spectrum of available measurements, making the methods sensitive to the gene preselection step.

Our SNF approach is distinct in that it uses networks of samples as a basis for integration. For example, when combining data from patient samples, SNF creates a patient network. Although networks of individuals have been extensively studied in other contexts, most notably in social science[8] or in relation to disease[9], to our knowledge patient-similarity networks have not been used specifically for integrating biological data. SNF consists of two main steps: construction of a sample-similarity network for each data type and integration of these networks into a single similarity network using a nonlinear combination method.

The fused network captures both shared and complementary information from different data sources (**Supplementary Results** and **Supplementary Figs. 1–3**), offering insight into how informative each data type is to the observed similarity between samples. Because it is based on networks of samples, SNF can derive useful information even from a small number of samples, is robust to noise and data heterogeneity, and scales to a large number of genes. In addition to integrating data, our fused networks can efficiently identify subtypes among existing samples by clustering and predict labels for new samples based on the constructed network

Figure 1 | Illustrative example of SNF steps. (**a**) Example representation of mRNA expression and DNA methylation data sets for the same cohort of patients. (**b**) Patient-by-patient similarity matrices for each data type. (**c**) Patient-by-patient similarity networks, equivalent to the patient-by-patient data. Patients are represented by nodes and patients' pairwise similarities are represented by edges. (**d**) Network fusion by SNF iteratively updates each of the networks with information from the other networks, making them more similar with each step. (**e**) The iterative network fusion results in convergence to the final fused network. Edge color indicates which data type has contributed to the given similarity.

(Online Methods, **Supplementary Note 1** and **Supplementary Fig. 4**). Combining diverse data types from five different human cancers, we demonstrated that SNF yields coherent, clinically relevant patient subtypes and improves on the performance of popular integrative approaches and a network-based approach that uses individual data types. The SNF software easily scales to multiple genome-wide data types with tens of thousands of measurements and is freely available as **Supplementary Software** and at http://compbio.cs.toronto.edu/SNF/.

RESULTS
Method overview
Given two or more types of data for the same set of samples (e.g., patients), SNF first creates a network for each data type and then fuses these into one similarity network. The initial step is to use a similarity measure for each pair of samples to construct a sample-by-sample similarity matrix for each available data type (**Fig. 1a,b**). The matrix is equivalent to a similarity network where nodes are samples (e.g., patients) and the weighted edges represent pairwise sample similarities (**Fig. 1c**). Both matrices and networks are effective visual representations: similarity matrices help identify global patterns (clusters), whereas networks emphasize the detailed similarity patterns and the types of data that support each edge.

The network-fusion step (**Fig. 1d**) uses a nonlinear method based on message-passing theory[10] that iteratively updates every network, making it more similar to the others with every iteration. After a few iterations, SNF converges to a single network (**Fig. 1e**). The empirical convergence for a variety of data sets is shown on **Supplementary Figures 5–7**. The method is robust to a variety of the hyperparameter settings (Online Methods and **Supplementary Figs. 8–10**). The advantage of our integrative procedure is that weak similarities (low-weight edges) disappear, helping to reduce the noise (**Fig. 2** and **Supplementary Fig. 2**), and strong similarities (high-weight edges) present in one or more networks are added to the others. Additionally, low-weight edges supported by all networks are retained depending on how tightly connected their neighborhoods are across networks. Such nonlinearity allows SNF to make full use of a network's local structure, integrating common as well as complementary information across networks.

A case study: glioblastoma multiforme
Multiple integrative approaches have been applied to understand the heterogeneity and identify the subtypes of glioblastoma multiforme (GBM), an aggressive adult brain tumor. Depending on the type of data used, these integrative analyses often lead to different conclusions. For example, one analysis that had combined expression and copy-number-variant data had identified two subtypes[11], but a later analysis[2], driven primarily by expression data, had identified four subtypes, which does not agree with the previous findings. A recent DNA methylation–based approach had identified three subtypes: one characterized by a somatic mutation in *IDH1* (ref. 12) and two others roughly corresponding to the subtypes identified in ref. 2. Though methylation data had been used for the analysis in ref. 2, the IDH subtype had not been identified because of the expression data–driven subtyping analysis.

We used SNF to fuse three data types for 215 patients with GBM: DNA methylation (1,491 genes), mRNA expression (12,042 genes) and miRNA expression (534 miRNAs). As expected, networks built using a single data type yielded very different patterns supports of patient similarity. For example, DNA methylation strongly supports connectivity in the smallest patient cluster (**Fig. 2a**), whereas mRNA expression supports similarity in the medium-sized cluster (**Fig. 2b**). DNA methylation and mRNA expression suggest relatively strong intercluster similarity (**Fig. 2a,b**), though the exact patterns are different between those data types. It is difficult to discern patterns in the patient-similarity network based on miRNA data alone (**Fig. 2c**). The fused network gives a much clearer picture of clustering in our set of patients with GBM, illustrated by the tightness of connectivity within clusters and relatively few edges between clusters (**Fig. 2d**).

We unified the results of several previous GBM analyses as well as identified new and potentially interesting associations. For example, our smallest cluster (subtype 3) corresponds to the previously identified IDH subtype[12] consisting of younger patients with a substantially more favorable prognosis. All patients with

Figure 2 | Patient similarities for each of the data types independently compared to SNF fused similarity. (**a**–**d**) Patient-to-patient similarities for 215 patients with GBM represented by similarity matrices and patient networks, where nodes represent patients, edge thickness reflects the strength of the similarity, and node size represents survival. Clusters are coded in grayscale (subtypes 1–3) and arranged according to the subtypes revealed through spectral clustering of the combined patient network. The clustering representation is preserved for all four networks to facilitate visual comparison. DNA methylation (**a**), mRNA expression (**b**), miRNA expression (**c**) and SNF-combined similarity matrix and network (**d**; see **Supplementary Fig. 11** for more information about network edges).
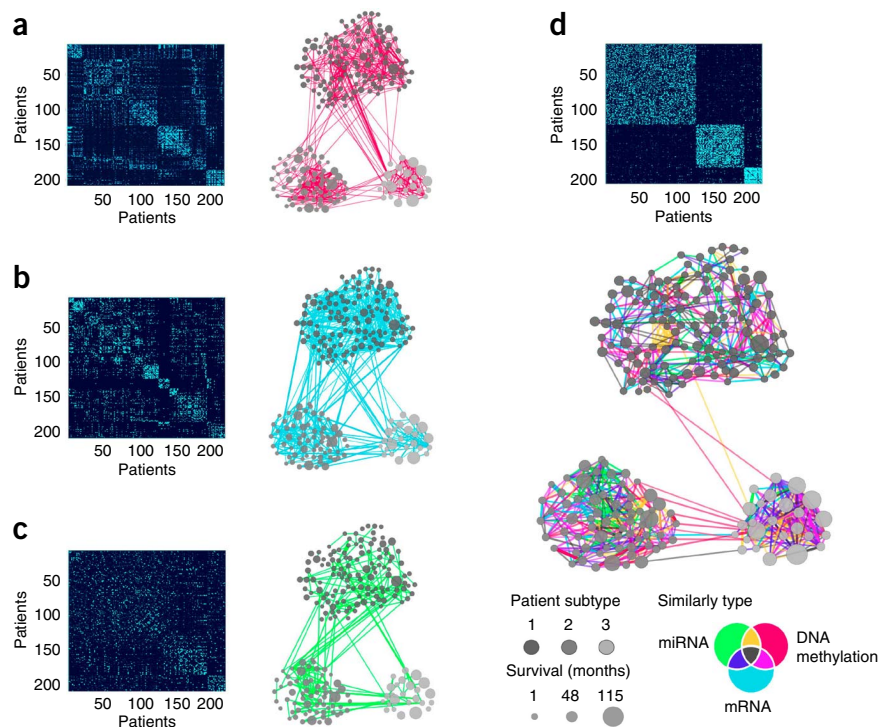
an *IDH1* mutation for whom the information was available (*n* = 14 patients, Fisher exact test $P = 4.87 \times 10^{-11}$) belong to this cluster. Subtype 1 patients (hazards ratio (HR) = 0.278, Cox log-rank test $P = 0.001$; **Supplementary Fig. 11c**) had a favorable response to temozolomide (TMZ), a drug commonly used to treat GBM (**Supplementary Figs. 11** and **12**). One of the reasons for the lack of such an effect in subtype 2 could be its significant association with *CTSD* overexpression ($P < 0.001$, Bonferroni-corrected), which has been found to prevent the effect of TMZ *in vitro*[13] (**Supplementary Results**).

Our network analysis goes beyond subtyping. Each edge in the fused network is colored by the data type(s) that contributed to the given similarity. A multicolor cluster means that no single data type or combination support patient similarity across GBM. We found that most edges were supported by at least two data types: 49.5% of all patient similarities (edges) were due to two data types, 17.2% were supported by all three data types and the remaining 33.3% of the edges were supported by only one data type, with strong enough similarity that those edges remained prominent in the fused network (**Supplementary Fig. 13**). The GBM analysis highlights three important features of our network-based integrative approach: (i) the ability to detect common as well as complementary signals (**Fig. 2d** and **Supplementary Fig. 1**); (ii) the ability to reduce noise by aggregating across multiple types of data (**Fig. 2d**, and **Supplementary Figs. 2** and **3**); and (iii) insight into the relative importance of each data source for determining patient similarity, thus refining our understanding of the heterogeneity within each subtype (**Fig. 2d** and **Supplementary Fig. 14**).

Evaluating SNF across a wide spectrum of cancers

In addition to the GBM analysis, we applied SNF to four other cancer profiles by TCGA: breast invasive carcinoma (BIC), kidney renal clear cell carcinoma (KRCCC), lung squamous cell carcinoma (LSCC) and colon adenocarcinoma (COAD). The DNA methylation, mRNA and miRNA expression data for these cancers vary in sample size (from 92 for COAD to 215 for GBM) and number of measurements (from 534 miRNAs in GBM to 27,578 methylated genes in LSCC and COAD) as well as heterogeneity[3,5,6] (**Supplementary Data** and **Supplementary Table 1**).

We evaluated SNF performance by identifying subtypes in each of these cancers. We report three commonly used measures: (i) *P* value in Cox log-rank test to evaluate the significance of the difference in survival profiles between subtypes[14]; (ii) silhouette score[15], a measure of cluster coherence, to evaluate whether patients are more similar within or across subtypes; and (iii) algorithm running time to evaluate scalability (**Supplementary Note 2**). We used spectral clustering (Online Methods) on the patient network to identify homogeneous cancer subtypes. We compared SNF to iCluster[7] and the concatenation of the three types of data (**Supplementary Note 3**).

We first compared data integration to the use of individual data types separately across the five cancers. We obtained patient clusters for individual data types by building a patient-similarity network and clustering it using spectral clustering (same as for SNF). Except for a few cases, single data type analysis did not lead to significantly different survival profiles, but networks fused by SNF had significant differences in survival among subtypes in all five cancers (**Table 1**). Note that the added fusion step is the only difference between the single and fused analyses. Spatial embedding of the subtypes from the fused network for each cancer showed very clear separation between clusters (**Supplementary Fig. 15**).

One major limitation of current integrative methods such as iCluster is the need for *a priori* gene selection. Although SNF

Table 1 | SNF-based analysis versus individual data types

Cancer type	mRNA expression	DNA methylation	miRNA	SNF
GBM (3 clusters)	0.54	0.11	0.21	2.0×10^{-4}
BIC (5 clusters)	0.03	0.05	0.30	1.1×10^{-3}
KRCCC (3 clusters)	0.20	0.61	0.17	2.9×10^{-2}
LSCC (4 clusters)	0.06	0.26	0.46	2.0×10^{-2}
COAD (3 clusters)	0.18	0.04	0.46	8.8×10^{-4}

Analysis using Cox log-rank test *P* values.
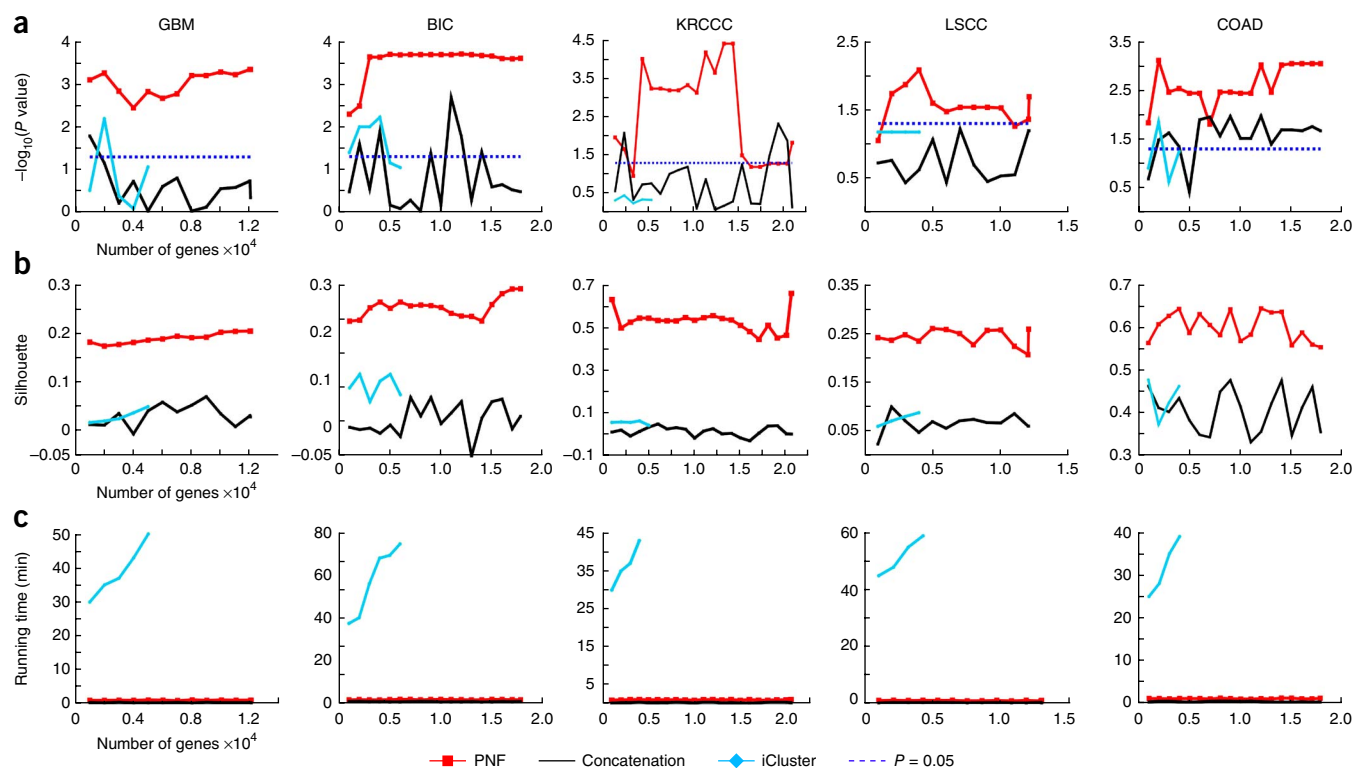
Figure 3 | Comparison of the SNF approach to iCluster and concatenation. (**a–c**) Cox log-rank test *P* value for survival analysis (**a**), silhouette score representing the coherence of clusters (**b**) and running time (**c**) for each of the indicated cancers. Number of preselected genes is shown on *x* axes.

does not require preselection, for comparison we report the performance of all three methods as a function of the number of preselected genes, ordering genes by significance for differential expression between tumor and healthy tissue using the significance analysis of microarrays (SAM) test[16] (**Supplementary Note 2**). SNF achieved significance in survival analysis across the spectrum of preselected genes (**Fig. 3a**) and resulted in substantially more coherent clusters according to the silhouette score (**Fig. 3b**). Comparative performance across cancers showed that in GBM and BIC, Cox survival *P* values were very stable with respect to the number of preselected genes. There is more fluctuation in survival *P* values for KRCCC and LSCC. This is explained by the fact that both KRCCC and LSCC have at least one subtype with very few patients (**Supplementary Fig. 15**), making the *P* values very sensitive to any change in clustering. This is a common problem of rare disease subtypes; the silhouette score in this case is a better indicator of clustering stability.

iCluster achieved significance for a small number of genes but was very sensitive to gene preselection. The performance of concatenation was even less predictable, though it was substantially faster, as illustrated by the running-time analysis (**Fig. 3c**). The computational complexity of the concatenation approach was equivalent to the time needed to run hierarchical clustering. Running time for SNF was only marginally higher than for concatenation. iCluster performance scaled exponentially in the number of genes, which explained the necessity for gene preselection.

From subtype-based to network-based outcome prediction

We showed that clustering patient networks derived from multiple data types using SNF performs as well or better than the state-of-the-art subtyping methods applied to survival analysis (**Supplementary Figs. 16–19**). On the harder task of survival risk prediction, we also found that subtyping was inferior to a true network-based approach (**Supplementary Fig. 20**).

We used the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) breast cancer data set[17] to validate our network-based prediction. The METABRIC data set consists of a discovery cohort (997 patients) and validation cohort (995 patients). We performed a PAM50 analysis (a standard breast cancer signature), iCluster analysis (InterClust[17]) and SNF analysis with five clusters (chosen by our model selection criterion) and ten clusters (for comparative purposes) (**Table 2**). The published[17] significance value obtained using iCluster on the validation set is lower than both the iCluster-based discovery cohort *P* value and the validation *P* value obtained using SNF, suggesting the potential for overfitting by iCluster. The concordance index (CI) is a continuous and robust accuracy measure to assess the prognostic

Table 2 | METABRIC survival analysis and prediction

	PAM50 (5 clusters)	iCluster (10 clusters)	SNF (5 clusters)	SNF (10 clusters)	Network
P value discovery cohort	3.0×10^{-9}	1.2×10^{-14}	6.10×10^{-11}	3.31×10^{-12}	–
P value validation cohort	1.7×10^{-9}	2.9×10^{-11}	5.12×10^{-13}	7.86×10^{-12}	–
CI discovery cohort	0.560	0.621	0.638	0.638	0.720
CI validation cohort	0.551	0.605	0.633	0.633	0.706

Comparison of SNF and alternative methods on survival analysis (Cox log-rank test *P* value) and risk of death prediction in the METABRIC data (CI, concordance index).

value of risk prediction models (**Supplementary Note 2**). CI for SNF was higher (better) than CI for PAM50 and iCluster on both discovery and validation cohorts for both five and ten clusters (**Table 2**). The CI values were relatively similar for all compared methods, indicating that subtype-based analyses have certain limitations.

We developed a network-based prediction approach that takes advantage of the whole network of patients rather than just individual clusters. Specifically, our network-based approach uses the fused network to constrain the Cox regression model to predict similar survival values for biologically similar patients (Online Methods and **Supplementary Results**). The network-based approach resulted in over 10% improvement in CI without any parameter tuning (**Table 2**). This network-based CI prediction on the validation cohort ranks in the top 20 of 1,400 models designed specifically for this task[18]. As we used the same network to assess CI for subtyping survival analysis and network-based survival analysis, we attribute the improvement in our results to the incorporation of richer information contained in the network.

DISCUSSION

We propose the SNF to integrate data in the space of samples (e.g., patients) rather than measurements (e.g., genes). Using SNF we constructed patient networks and combined mRNA expression, DNA methylation and miRNA expression data to identify subtypes with differential survival profiles. SNF also has many other applications. In the clinical domain, patient networks allow integration of very different kinds of measurements, such as microbiome and metabolomics data, questionnaires and functional magnetic resonance imaging, together with genomic, clinical and demographic data, as long as the data can be used to identify similarity between patients (Online Methods). Although some of these data types have been combined previously, our method enables their combination into a single comprehensive network that yields precise manifolds of diseases.

SNF can help answer questions that require combining multiple types or sources of data for the same set of objects or subjects, not just humans. For example, combining transcriptomic, epigenetic and genetic data for different tomato strains helps to visualize how biological similarity relates to the phenotype of interest, such as tomato sweetness. SNF can also integrate various gene-interaction data such as physical interactions, coexpression and colocalization data. In another context, it can improve the reliability and remove the experimental bias in constructing gene coexpression networks by integrating tissue-specific gene-expression data from a variety of experiments.

One important advantage of our approach is that it goes beyond current subtyping strategies to capture continuous phenotypes. Our analysis of cancers shows that although there are broad categories of patients (subtypes), the reality is more complex. Capturing variability in similarity and underlying biology via similarity networks moves us closer to the clinic of the future[19]. We believe that our fused networks will ultimately pave the way for much more refined representation and understanding of diseases, phenotypes and other biological phenomena.

METHODS

Methods and any associated references are available in the online version of the paper.

1. Monti, S., Tamayo, P., Mesirov, J. & Golub, T. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Mach. Learn.* **52**, 91–118 (2003).
2. Verhaak, R.G.W. *et al.* Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* **17**, 98–110 (2010).
3. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
4. Kirk, P., Griffin, J.E., Savage, R.S., Ghahramani, Z. & Wild, D.L. Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics* **28**, 3290–3297 (2012).
5. Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519–525 (2012).
6. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).
7. Shen, R., Olshen, A.B. & Ladanyi, M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* **25**, 2906–2912 (2009).
8. Goldenberg, A., Zheng, A.X., Fienberg, S.E. & Airoldi, E.M. A survey of statistical network models. *Foundations and Trends in Machine Learning.* **2**, 129–233 (2010).
9. Barabási, A.-L. Network medicine -from obesity to the 'diseasome. *N. Engl. J. Med.* **357**, 404–407 (2007).
10. Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (Morgan Kaufmann, 1988).
11. Nigro, J.M. *et al.* Integrated array-comparative genomic hybridization and expression array profiles identify clinically relevant molecular subtypes of glioblastoma. *Cancer Res.* **65**, 1678–1686 (2005).
12. Sturm, D. *et al.* Hotspot mutations in H3F3A and IDH1 define distinct epigenetic and biological subgroups of glioblastoma. *Cancer Cell* **22**, 425–437 (2012).
13. Sun, S. *et al.* Protein alterations associated with temozolomide resistance in subclones of human glioblastoma cell lines. *J. Neurooncol.* **107**, 89–100 (2012).
14. Hosmer Jr, D.W., Lemeshow, S. & May, S. *Applied Survival Analysis: Regression Modeling of Time to Event Data* (Wiley, 2011).
15. Rousseeuw, P. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).
16. Tusher, V.G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA* **98**, 5116–5121 (2001).
17. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
18. Margolin, A.A. *et al.* Systematic analysis of challenge-driven improvements in molecular prognostic models for breast cancer. *Sci. Transl. Med.* **5**, 181 (2013).
19. Friend, S.H. & Ideker, T. Point: Are we prepared for the future doctor visit? *Nat. Biotechnol.* **29**, 215–218 (2011).

ONLINE METHODS

Experimental details. We used data from five different cancer types available from the TCGA website: GBM, BIC, LSCC, KRCCC and COAD. For each of these tumor types, we downloaded TCGA-curated level 3 data sets containing gene expression, miRNA expression and DNA methylation information. TCGA repository contains multiple platforms for each data type. We always chose the platform corresponding to the largest number of available individuals and describing both tumor samples and controls whenever possible. For expression data, we used the Broad Institute HT-HG-U133A platform in GBM and LSCC, the UNC-Agilent-G4502A-07 platform in BIC and COAD and the UNC-Illumina-Hiseq-RNASeq platform in KRCCC. For miRNA expression data, we used the BCGSC-Illumina-Hiseq-miRNAseq platform in BIC, the UNC-miRNA-8X15K platform in GBM and the BCGSC-Illumina-GA-miRNAseq in LSCC, KRCCC and COAD. Finally, for the methylation data we used the JHU-USC-Illumina-DNA-Methylation platform in GBM, the JHU-USC-Human-Methylation-27 platform for BIC, LSCC, KRCCC and COAD. For all these tumor types, we also downloaded patients' clinical information including the overall survival data.

We also used METABRIC data set to evaluate the effectiveness of survival prediction with network regularization. METABRIC data set consists of two cohorts: discovery (997 patients) and validation (995 patients). For each of these patients, matched DNA and RNA were extracted from each primary tumor specimen and subjected to copy-number and genotype analysis on the Affymetrix SNP 6.0 platform and transcriptional profiling on the Illumina HT-12 v3 platform (Illumina-Human-WG-v3). We used the normalized data available from the European Genome-Phenome Archive (EGA, http://www.ebi.ac.uk/ega/). High-quality follow up clinical data including information on disease-free survival were also available for both cohorts. As a preprocessing step, we mapped copy-number variations to genes using the PennCNV package[20].

Before applying our SNF, we performed three steps of preprocessing: outlier removal, missing-data imputation and normalization. If a patient had more than 20% missing data in a certain data type, we did not consider this patient. Similarly, if a certain biological feature (for example, mRNA expression) had more than 20% of missing values across patients, we filtered out this feature. Also, for missing data, we used K nearest neighbor (KNN) imputation[21], where the number of neighbors is the same with K value used in our method (see below); therefore we do not have any free parameters. Last, before constructing the patient network, we performed the following normalization:

$$\tilde{f} = \frac{f - E(f)}{\sqrt{\text{Var}(f)}},$$

where f is any biological feature, $\tilde{f}$ is the corresponding feature after normalization, $E(f)$ and $Var(f)$ represent the empirical mean and variance of f, respectively.

Evaluation metrics. We used several metrics for evaluation and comparison of our method to existing approaches. In the real-cancer data, we use three metrics, as ground truth was not known. First, we use silhouette[15] to measure the homogeneity of the subtypes. For each patient i, let $a(i)$ denote the average dissimilarity to all other patients within the same subtype and $b(i)$

denote the lowest average dissimilarity to all other patients in different subtypes. The value of silhouette for patient i was defined as $s(i) = (b(i) - a(i))/(\max a(i), b(i))$. The mean value of silhouette for all the patients was then used as a measure of how tightly grouped all the data in the cluster are. If silhouette value was close to 1, then it means the data were appropriately clustered.

We also used P value for log-rank test of survival separation in Cox regression model[14]. P value measures the significance in the difference of survival profiles between subtypes. In our test, we set 0.05 to be the threshold of the significance. The lower the P value was, the less likely it was that such differential survival was observed by chance, i.e., the more significantly different the survival profiles was between subtypes. For most cancers, we used days to the last follow-up and the vital status to perform the log-rank test for survival analysis. However, for COAD, we used the consensus of the days to last known alive together with the last follow-up as a proxy because there were a lot of missing values in the data for days to last follow up. We used running time (in minutes) to compare the scalability of each method.

Similarity network fusion. Suppose we have n samples (for example, patients) and m measurements (for example, mRNA gene expression). We will use the patient network example throughout this section for clarity though the method has broad applicability as discussed above. A patient similarity network is represented as a graph $G = (V, E)$. The vertices V correspond to the patients $\{x_1, x_2, \cdots, x_n\}$ and the edges E are weighted by how similar the patients are. Edge weights are represented by an $n \times n$ similarity matrix $\mathbf{W}$ with $\mathbf{W}(i, j)$ indicating the similarity between patients x_i and x_j and are computed as follows. We denote $\rho(x_i, x_j)$ as the Euclidean distance between patients x_i and x_j. We then use a scaled exponential similarity kernel to determine the weight of the edge:

$$\mathbf{W}(i, j) = \exp\left(-\frac{\rho^2(x_i, x_j)}{\mu \varepsilon_{i, j}}\right) \quad (1)$$

where μ is a hyperparameter that can be empirically set and $\varepsilon_{i,j}$ is used to eliminate the scaling problem. Here we define

$$\varepsilon_{i, j} = \frac{\text{mean}(\rho(x_i, N_i)) + \text{mean}(\rho(x_j, N_j)) + \rho(x_i, x_j)}{3}$$

where mean $(\rho(x_i N_i))$ is the average value of the distances between x_i and each of its neighbors. We recommend setting μ in the range of [0.3, 0.8]. Note that while this distance measure is suitable for continuous variables, we propose to use chi-squared distance for discrete variables and agreement-based measure for binary variables.

To compute the fused matrix from multiple types of measurements, we define a full and sparse kernel on the vertex set V. The full kernel is a normalized weight matrix $\mathbf{P} = \mathbf{D}^{-1}\mathbf{W}$, where $\mathbf{D}$ is the diagonal matrix whose entries $\mathbf{D}(i, i) = \Sigma_j\mathbf{W}(i, j)$, so that $\Sigma_j P(i, j) = 1$. However, this normalization may suffer from numerical instability since it involves self-similarities on the diagonal entries of $\mathbf{W}$. One way to perform a better normalization is as follows:

$$\mathbf{P}(i, j) = \begin{cases} \dfrac{\mathbf{W}(i, j)}{2\Sigma_{k \neq i}\mathbf{W}(i, k)}, & j \neq i \\ 1/2, & j = i \end{cases} \quad (2)$$

This normalization will be free of the scale of self-similarity in the diagonal entries and $\Sigma_j \mathbf{P}(i,j) = 1$ still holds.

Let N_i represent a set of x_i's neighbors including x_i in G. Given a graph, G, we use K nearest neighbors (KNN) to measure local affinity as:

$$\mathbf{S}(i,j) = \begin{cases} \dfrac{\mathbf{W}(i,j)}{\Sigma_{k \in N_i} \mathbf{W}(i,k)}, & j \in N_i \\ 0 & \text{otherwise} \end{cases} \qquad (3)$$

This operation sets the similarities between non-neighboring points (in terms of the pairwise similarity values) to zero. Essentially we make the assumption that local similarities (high values) are more reliable than remote ones; and we thus assign similarities to non-neighbors through graph diffusion on the network. This is a mild assumption widely adopted by other manifold learning algorithms. Note that $\mathbf{P}$ carries the full information about the similarity of each patient to all others whereas $\mathbf{S}$ only encodes the similarity to the K most similar patients for each patient. Our algorithm always starts from $\mathbf{P}$ as the initial state using $\mathbf{S}$ as the kernel matrix in the fusion process for both capacity of capturing local structure of graphs and computational efficiency.

Given m different data types, we can construct similarity matrices $\mathbf{W}^{(v)}$ using equation (1) for the vth view, $v = 1, 2, ..., $ m. $P^{(v)}$ and $\mathbf{S}^{(v)}$ are obtained from equations (2) and (3), respectively. Below we introduce our network fusion process given a set of networks.

Let us first consider the case when we have two data types, i.e., $m = 2$. We calculate the status matrices $\mathbf{P}^{(1)}$ and $\mathbf{P}^{(2)}$ as in equation (2) from two input similarity matrices; then the kernel matrices $\mathbf{S}^{(1)}$ and $\mathbf{S}^{(2)}$ are obtained as in equation (3).

Let $\mathbf{P}_{t=0}^{(1)} = \mathbf{P}^{(1)}$ and $\mathbf{P}_{t=0}^{(2)} = \mathbf{P}^{(2)}$ represent the initial two status matrices at $t = 0$. The key step of SNF is to iteratively update similarity matrix corresponding to each of the data types as follows:

$$\mathbf{P}_{t+1}^{(1)} = \mathbf{S}^{(1)} \times \mathbf{P}_t^{(2)} \times (\mathbf{S}^{(1)})^T \qquad (4)$$

$$\mathbf{P}_{t+1}^{(2)} = \mathbf{S}^{(2)} \times \mathbf{P}_t^{(1)} \times (\mathbf{S}^{(2)})^T \qquad (5)$$

where $\mathbf{P}_{t+1}^{(1)}$ is the status matrix of the first data type after t iterations. $\mathbf{P}_{t+1}^{(2)}$ is the similarity matrix for the second data type. This procedure updates the status matrices each time generating two parallel interchanging diffusion processes. After t steps, the overall status matrix is computed as

$$\mathbf{P}^{(c)} = \frac{\mathbf{P}_t^{(1)} + \mathbf{P}_t^{(2)}}{2}.$$

Since $\mathbf{S}$ is a KNN graph of $\mathbf{P}$, which can reduce some noise between instances, our SNF is robust to the noise in similarity measures.

Another way to think of the updating rule (4) is

$$\mathbf{P}_{t+1}^{(1)}(i,j) = \sum_{k \in N_i} \sum_{l \in N_j} \mathbf{S}^{(1)}(i,k) \times \mathbf{S}^{(1)}(j,l) \times \mathbf{P}_t^{(2)}(k,l) \qquad (6)$$

(the same for $\mathbf{P}_{t+1}^{(2)}$). Note N_i represents the neighborhood of x_i. We can see that similarity information is only propagated through the common neighborhood. This renders SNF robust to noise. An important observation is that if x_i and x_j have common neighbors in both similarity matrices, it is highly possible that they belong to the same cluster. Another essential fact our method benefits from is that even if x_i and x_j are not very similar

in one data type, their similarity can be expressed in another data type and this similarity information can be propagated through the fusion process.

After each iteration, we performed normalization on $\mathbf{P}_{t+1}^{(1)}$ and $\mathbf{P}_{t+1}^{(2)}$ as in equation (2). By performing the normalization, we (i) ensure that throughout SNF iterations a patient is always most similar to himself than to other patients; (ii) ensure that our final network is full rank, important for the classification and clustering applications of the final network. Finally, we have found that the use of such normalization leads to quicker convergence of SNF.

Finally, an extension to the case $m > 2$ follows equations (4) and (5):

$$\mathbf{P}^{(v)} = \mathbf{S}^{(v)} \times \left(\frac{\Sigma_{k \neq v} \mathbf{P}^{(k)}}{m-1} \right) \times (\mathbf{S}^{(v)})^T, v = 1, 2, \cdots, m \qquad (7)$$

The input to our algorithm can be feature vectors, pairwise distances, or pairwise similarities. The learned status matrix $P^{(c)}$ can then be used for retrieval, clustering and classification; in this work, we focus mostly on clustering and prediction.

SNF is inspired by the theoretical multiview learning framework developed for the computer vision and image processing applications[22] that is not directly applicable to biological data. SNF constructs networks of samples (for example, patients) by comparing samples' molecular (or phenotypic) profiles; fused networks are used for subtyping and label prediction distinguishing SNF from all the previously published research.

Network clustering (for example, for disease subtyping). Given n samples and m measurements we want to identify C clusters of samples, each of which corresponds to a (known or new) subtype. We associate each sample x_i with a label indicator vector $y_i \in \{0,1\}^C$ such that $y_i(k) = 1$ if sample x_i belongs to the kth cluster (subtype), otherwise $y_i(k) = 0$. So a partition matrix $\mathbf{Y} = \left(y_1^T; y_2^T; \cdots; y_n^T \right)$ is used to represent a clustering scheme.

Given the fused graph, in this work we used spectral clustering to obtain network clusters. Traditional state-of-the-art spectral methods[23], aim to minimize RatioCut[24], an objective function that effectively combines MinCut and equipartitioning, by solving the following optimization problem:

$$\min_{Q \in \mathbb{R}^{n \times C}} \text{Trace}(\mathbf{Q}^T \mathbf{L}^+ \mathbf{Q}) \qquad (8)$$
$$\text{s.t.} \mathbf{Q}^T \mathbf{Q} = \mathbf{I}$$

where $\mathbf{Q} = \mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1/2}$ is a scaled partition matrix, $\mathbf{L}^+$ denotes the normalized Laplacian matrix $\mathbf{L}^+ = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$ given the similarity matrix $\mathbf{W}$). Matrix $\mathbf{D}$ is a network degree matrix, with degrees of each node on the diagonal and off-diagonal elements set to 0). Spectral clustering is effective in capturing global structure of the graph[25].

Network-based survival risk prediction. With the fused network, we can perform tasks beyond disease subtyping. An example in this paper is survival prediction with network regularization. Cox model has been successfully applied to perform survival/risk prediction of given new patients. Given all the feature matrix $\mathbf{X}$, the risk of an event (death) at time t for the i-th patient is given by $h(t|\mathbf{X}) = h_0(t) \exp(\mathbf{X}^T z)$, where z is a vector of regression coefficients and $h_0(t)$ is the baseline hazard function. This regression

coefficient vector z is estimated by maximizing the Cox's log-partial likelihood:

$$lp(z) = \sum_{i=1}^{n} \delta_i \left(\mathbf{X}_i^T z - \log \left(\sum_{j \in R(t_i)} \exp \left(\mathbf{X}_j^T z \right) \right) \right) \qquad (9)$$

where n is the number of patients, t_i is the survival time for the i-th patient and $R(t_i)$ is the risk set at time t_i, i.e., the set of patients who still survived before t_i. $\delta_i(\cdot)$ is an indicator function whether the survival time is observed ($\delta_i = 1$) or censored ($\delta_i = 0$).

It is possible to improve survival prediction by incorporating additional information, such as gene interaction data[26] or patient similarity based constraints. To incorporate the network structure, similarity between either features or patients (or both) can be used as a regularizer. According to the hazard function of Cox's model, the relative risk between patient i and patient j is $\exp(\mathbf{X}_i^T z - \mathbf{X}_j^T z)$, therefore, a regularizer can be constructed as $(\mathbf{X}_i^T z - \mathbf{X}_j^T z)^2 w_{ij}$. To estimate z, we can use a modified likelihood expression as follows:

$$lp(z) = \sum_{i=1}^{n} \delta_i \left(X_i^T z - \log \left(\sum_{j \in R(t_i)} \exp(X_j^T z) \right) \right) - \lambda \sum_i \sum_j (X_i^T z - X_j^T z)^2 w_{ij} \qquad (10)$$

where λ is the regularizing coefficient. Newton optimization techniques are applied to solve this maximization problem.

Combining data types. SNF can be used to incorporate arbitrary types of discrete (binary or categorical) and continuous data. For integration of discrete data, we recommend the use of chi-squared distance as the similarity measure. Compatibility of data sources can be checked via normalized mutual information (NMI). If the patient similarity obtained from different data sources is completely discordant; NMI can help to clarify which data should and which should not be combined.

NMI帮助确定不同数据类型之间的兼容情况

20. Wang, K. *et al.* PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **17**, 1665–1674 (2007).
21. Troyanskaya, O. *et al.* Missing value estimation methods for DNA microarrays. *Bioinformatics* **17**, 520–525 (2001).
22. Wang, B., Jiang, J., Wang, W., Zhou, Z.-H. & Tu, Z. Unsupervised metric fusion by cross diffusion. in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* 2997–3004 (IEEE, 2012).
23. Ng, A.Y., Jordan, M.I. & Weiss, Y. On spectral clustering: analysis and an algorithm. *Adv. Neural Inf. Process. Syst.* **2**, 849–856 (2002).
24. Wei, Y.C. & Cheng, C.K. Towards efficient hierarchical designs by ratio cut partitioning. in *Proc. Int. Conf. Computer-Aided Design* 298–301 (ICCAD, 1989).
25. Luxburg, U. A tutorial on spectral clustering. *Stat. Comput.* **17**, 395–416 (2007).
26. Zhang, W. *et al.* Network-based survival analysis reveals subnetwork signatures for predicting outcomes of ovarian cancer treatment. *PLoS Comput. Biol.* **9**, e1002975 (2013).