Similarity network fusion for the integration of multi-omics and microbiomes in respiratory disease

Jayanth Kumar Narayana [1], Micheál Mac Aogáin [1], Nur A'tikah Binte Mohamed Ali [1], Krasimira Tsaneva-Atanasova [2] and Sanjay H. Chotirmall [1]

[1] Lee Kong Chian School of Medicine, Nanyang Technological University Singapore, Singapore. [2] Dept of Mathematics, College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, UK.

Corresponding author: Sanjay H. Chotirmall (schotirmall@ntu.edu.sg)

Shareable abstract (@ERSpublications)
Similarity network fusion (SNF) is increasingly employed for multi-omics and microbiome data integration and assists patient endotyping. This Methods article describes its performance and explores current and future applications in respiratory medicine. https://bit.ly/3gtoYq9

Cite this article as: Narayana JK, Mac Aogáin M, Ali NABM, *et al*. Similarity network fusion for the integration of multi-omics and microbiomes in respiratory disease. *Eur Respir J* 2021; 58: 2101016 [DOI: 10.1183/13993003.01016-2021].

Received: 8 April 2021
Accepted: 4 June 2021

Introduction

Advances in platform technologies facilitate the design of large-scale "multi-omic" studies that encompass genomic, transcriptomic, proteomic, epigenomic, metabolomic and microbiomic components, each representing different views of a single biological specimen [1]. While useful, this is analogous to the "Flatland" *jeu d'esprit*, where the same reality (*i.e.* a sphere of constant diameter) is subject to different interpretations (*i.e.* circles of varying diameter) depending on one's point of view (from various two-dimensional cross sections). Although each -omics approach has value, they can be even more useful if holistically modelled through appropriate integration. While "mono-omic" analysis has been extremely beneficial, from a systems medicine perspective, this may fail to capture the emergent properties of an individual system and hence may yield limited understanding of non-linear and dynamic features, all of which are increasingly evident in the pathogenesis of respiratory disease [1]. There is clearly a growing need for a more holistic "all in" integration methodology that leverages each distinct -omic dataset derived from multi-omic studies (figure 1). Although several integrative methodologies are available (*e.g.* mixOmics, Anvi'o and integrOmics), similarity network fusion (SNF) has emerged as an appropriate, applicable and robust method in respiratory disease [2–4].

How does similarity network fusion work?

SNF requires three steps for implementation: creation of a similarity network based on individual -omic datasets; fusion of multiple similarity networks; followed by analysis of the integrated networks (figure 1).

Similarity network creation

In any data integration methodology, data standardisation is of paramount importance. This crucial step allows meaningful comparison between the different -omic datasets. In SNF, creation of a similarity network is used as a method for standardisation between different -omic datasets, for instance, for each mono-omic dataset, a network is created with individual patients as nodes and edges representing the value or magnitude of the similarity (measure) between patients (nodes), given that particular dataset (figure 1). Measures that are used to quantify similarity between patients largely depend upon the inherent properties of the data type. Therefore, theoretically appropriate and biologically meaningful similarity measures should be selected for each respective dataset [4–6]. For example, in work from our group, Bray–Curtis similarity, a metric reflecting taxonomic diversity, is applied to microbiome data [4]. In some cases, appropriate measures may not exist, in which case the similarity measure can be generated from distance measures (*e.g.* Euclidean distance) using a similarity kernel function, such as an exponential similarity kernel, which uses an exponential rule to assign exponentially decreasing values to distant patients and *vice versa* [2]. In specific cases where differing similarity measures are needed for different -omic datasets,

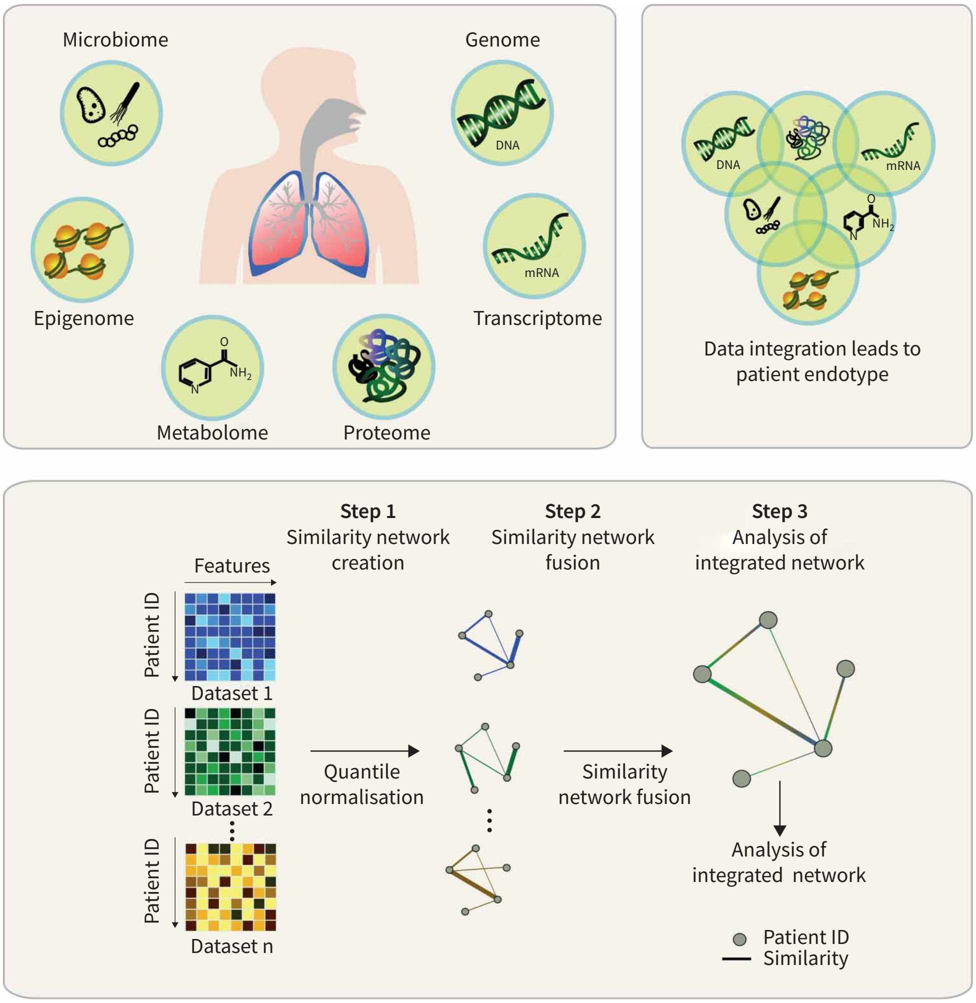

FIGURE 1 The concept of "integrated" multi-omics analysis: mono-omics-based profiling offers distinct views of a single clinical specimen and/or disease that, following integration, may improve patient endotyping. Similarity network fusion (SNF) can achieve this integration by creating similarity networks for each respective -omic dataset using an appropriate similarity measure. This is followed by quantile normalisation to normalise similarity values between datasets (if -omic specific similarity measures are used) followed by SNF to merge similarity networks that result in an integrated patient network. Further downstream analysis of the integrated patient network can then be performed to identify clinically relevant subgroups or clusters.

i.e. for integrating microbiomes using Bray–Curtis similarity (range: 0 to 1) with transcriptomes using Spearman correlation (range: −1 to 1), a potential false weighting of networks can be overcome by key normalisation steps, such as quantile normalisation, which makes the distributions from different similarity measures identical in statistical properties and hence rigorously comparable [7]. As such, the SNF approach is durable and can be adapted to suit a wide range of experimental datasets of diverse origin with the potential for inclusion of host -omic profiles (*e.g.* human transcriptomics, proteomics and/or metabolomics). This has particular relevance in the setting of chronic respiratory disease, where persistent bacterial infection may be accompanied by allergic manifestations of fungal origin in concert with a dysregulated host immune response [8], or, where microbiome profiles at disparate anatomical sites may functionally converge, contributing to pathology *via* the lung–gut axis [9, 10]. In such cases, an integrated analytical approach, encompassing multiple -omic datasets, is more likely to capture the ensuing complex and dynamic host–microbe interactions.

Similarity network fusion

SNF is used to combine multiple patient similarity networks (based on different -omics platforms) into a holistic network of patient relationships. It achieves this by decomposing the similarity network of a particular -omic dataset into two networks, one capturing the overall network structure *i.e.* similarity of a patient to all other patients, and the other capturing local network structure *i.e.* similarity of a patient to its "K"-most similar patients, where "K" may be tuned to an optimal value given the dataset [4]. Wang *et al.* [2] suggest setting "K" to the number of expected clusters or, if this is unknown, to N/10, where N represents the total patient number. Such an approach, capturing local structure from different -omic datasets, allows SNF to eliminate "noise". The decomposed networks (derived from different datasets) are then iteratively fused, a process best conceptualised as the diffusion of similarity information through common edges between the different patient similarity networks. In the resulting "final" network, an edge is said to have increased similarity if it is supported by the majority of -omic datasets and a decreased similarity if it is not.

Analysis of the integrated network

Typically, two forms of secondary analytical approach may be pursued given the multidimensional clinical data structure: supervised or unsupervised. Supervised analysis uses prior knowledge or "domain expertise" to model relationships between data "features", for example microbiome profiles, and "labels", the clinical patient phenotype. Conversely, unsupervised analysis attempts to elucidate "patterns", drawing inference from data "features". Importantly, since integrated SNF networks retain only similarity information between patients in the input space, implementation of supervised analysis is not straightforward and needs modification. Examples of this include integrative network fusion, which implements feature-ranked SNF within a machine learning framework, or SNF-NN, which implements deep learning alongside SNF [11, 12]. Alternatively, semi-supervised or unsupervised analysis can be pursued. Semi-supervised analysis employs combinations of labelled and unlabelled data in unsupervised fashion to predict unlabelled data in a supervised manner. One important semi-supervised algorithm is label propagation, a method leveraging on nodes connected by heavy edges, *i.e.* similar patients tend to have comparable labels. This graph algorithm assigns labels to previously unlabelled nodes by "propagating labels" of previously labelled nodes through associated edges. This is used to predict patient labels from the integrated network, given the known labels [13]. When practically applied, an integrated patient network derived from multi-omic endophenotypic profiles may be used to identify and risk stratify patients at higher risk of clinical deterioration.

Integrated patient networks may also be interrogated by similarity graph clustering algorithms, of which spectral clustering is most widely used [14]. This unsupervised clustering algorithm embeds the nodes, *i.e.* patients of a network, into a lower-dimensional space, preserving patient similarities essential for clustering but losing the original feature space (defining patient characteristics) for direct interpretation, and then "clusters" patients in this space with a preferred clustering method (*e.g.* k-means). By this approach, multi-omic integration by SNF, followed by spectral clustering, can demonstrably identify biologically meaningful subgroups, such as rhinovirus bronchiolitis endotypes [15].

Survival models, including Cox regression, which are used to predict survival time based on multiple risk factors, may also be used with SNF-integrated patient networks to improve survival analyses. Here, SNF-integrated networks can be leveraged as an additional input, to predict similar survival scores for similar patients (based on their integrated similarity scores) [2].

The current state of similarity network fusion in respiratory medicine

SNF has been successfully implemented in the integration of multi-omic data from the same group of patients using distinct molecular profiling methods or the assessment of distinct anatomical locations to better understand and characterise respiratory disease [10, 15]. SNF can identify "high-risk" bronchiectasis, while multi-omics analysis in COPD illustrates that integrating multiple -omics through SNF allows for a more accurate classification of a COPD diagnosis, even with small groups of individuals, over and above that achievable by "mono-omic" approaches [4, 16]. Work from our group further illustrates the added precision from multi-biome integration, *i.e.* bacteria, viruses and fungi, by SNF in bronchiectasis [4]. SNF is further valued as a potential integrative approach to endotype severe asthma and is currently being used to better understand sub-phenotypes of severe asthma by combining -omics datasets, including transcriptomics, proteomics, lipidomics and metabolomics [15, 17–19].

The advantages and limitations of similarity network fusion

Integrating multiple -omic datasets by SNF accumulates more information and therefore improves cluster precision and accuracy [4]. This process inherently down-weights "noise" if given enough -omic datasets and provides power to detect rarer subgroups from relatively small cohorts [16]. The number of variables or heterogeneity of the dataset does not influence SNF workflows as the similarity patient network is constructed based on each -omic dataset before integration. Missing data is also tolerated, provided that an appropriate similarity metric is implemented.

SNF assumes equal weights to different -omic datasets, and this may not be biologically appropriate since not all datasets characterise or represent the underlying disease pathology to the same extent. It is not always straightforward to identify a biologically relevant similarity measure for a given dataset, and the similarities calculated using Euclidean distances may not capture the topology of the data space and hence true biological similarities. SNF solely utilises similarity information between patients of an integrated network and clustering on this network offers no information on what combination of biological features cluster individuals. While clustering the integrated patient networks groups patients from clinically relevant disease-subtypes, it does not provide direct ways of delineating what combination of biological features, used during integration, define these subtypes. However, this can be achieved by secondary analysis such as normalised mutual information [2], or predictive modelling followed by model explainability techniques [20].

While modifications to SNF, including integrative network fusion (INF), affinity network fusion (ANF), similarity kernel fusion (SKF), association-signal-annotation boosted similarity network fusion (ab-SNF), robust similarity network fusion (RSNF), local scaling similarity network fusion (Ls-SNF) and weighted similarity network fusion (wSNF), can all improve the various limitations of SNF, no single integrative approach is best, and each has to be considered in terms of "best-use case" and its own inherent advantages and limitations [4, 11, 21–25]. It is also important to note that SNF-derived associations are still inferred, and experimental manipulations are required to definitively confirm causation. To date, SNF has been most widely applied in respiratory medicine, however, given the emerging disease heterogeneity in chronic respiratory disease and the increasing complexity of data obtained from a single clinical specimen, it is imperative that consensus and standardisation is reached for these methodologies. The development of accessible and reproducible software such as https://integrative-microbiomics.ntu.edu.sg/ will further assist in harnessing the value of multi-omic technologies for better translation in respiratory disease [4].

Conclusion

SNF and its associated methods are emerging as a key approach to integrate multi-omic and microbiome datasets in respiratory medicine [1, 4, 17–19, 26]. Given its advantages, we expect its use to further increase over the next decade to facilitate an improved understanding of respiratory disease pathogenesis and its associated patient endophenotypes, potentially offering a method for the application of precision medicine into clinical care.

Conflict of interest: None declared.

Support statement: This research is supported by the Singapore Ministry of Health's National Medical Research Council under its Clinician-Scientist Individual Research Grant (MOH-000141) (S.H. Chotirmall) and the NTU Integrated Medical, Biological and Environmental Life Sciences (NIMBELS; grant NIM/03/2018) (S.H. Chotirmall). K. Tsaneva-Atanasova gratefully acknowledges the financial support of the EPSRC *via* grant EP/T017856/1. Funding information for this article has been deposited with the Crossref Funder Registry.

References

1 Noell G, Faner R, Agustí A. From systems biology to P4 medicine: applications in respiratory medicine. *Eur Respir Rev* 2018; 27: 170110.

2 Wang B, Mezlini AM, Demir F, *et al.* Similarity network fusion for aggregating data types on a genomic scale. *Nat Methods* 2014; 11: 333–337.

3 Graw S, Chappell K, Washam CL, *et al.* Multi-omics data integration considerations and study design for biological systems and disease. *Mol Omics* 2021; 17: 170–185.

4 Mac Aogáin M, Narayana JK, Tiew PY, *et al.* Integrative microbiomics in bronchiectasis exacerbations. *Nat Med* 2021; 27: 688–699.

5 Fan C, Lei X, Pan Y. Prioritizing CircRNA-disease associations with convolutional neural network based on multiple similarity feature fusion. *Front Genet* 2020; 11: 540751.

6 Liu D, Ma Y, Jiang X, *et al.* Predicting virus-host association by Kernelized logistic matrix factorization and similarity network fusion. *BMC bioinformatics* 2019; 20: Suppl. 16, 594.

7 Bolstad BM, Irizarry RA, Astrand M, *et al.* A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 2003; 19: 185–193.

8 Mac Aogáin M, Tiew PY, Lim AYH, *et al.* Distinct 'Immunoallertypes' of disease and high frequencies of sensitization in non-cystic fibrosis bronchiectasis. *Am J Respir Crit Care Med* 2019; 199: 842–853.

9 Budden KF, Shukla SD, Rehman SF, *et al.* Functional effects of the microbiota in chronic respiratory disease. *Lancet Respir Med* 2019; 7: 907–920.

10 Narayana JK, Fransiskus Xaverius I, Oriano M, *et al.* Microbial dysregulation of the 'Lung-Gut' axis in high-risk bronchiectasis. D10 role of microbiome and bacteriophages in pulmonary infections. *Am Thorac Soc* 2021; 203: A1223.

11 Chierici M, Bussola N, Marcolini A, *et al.* Integrative network fusion: a multi-omics approach in molecular profiling. *Front Oncol* 2020; 10: 1065.

12 Jarada TN, Rokne JG, Alhajj R. SNF-NN: computational method to predict drug-disease interactions using similarity network fusion and neural networks. *BMC Bioinformatics* 2021; 22: 28.

13 Zhu X, Ghahramani Z. Learning from Labeled and Unlabeled Data with Label Propagation. Technical Report CMU-CALD-02–107. Pittsburgh, Carnegie Mellon University, 2002.

14 Chuanyi F, Huandong C, Jieqing X. Spectral Clustering and its Research Progress. Proceedings of the 2011 Seventh International Conference on Computational Intelligence and Security. IEEE Computer Society, 2011; pp. 1367–1369.

15 Raita Y, Camargo CA, Bochkov YA, *et al.* Integrated-omics endotyping of infants with rhinovirus bronchiolitis and risk of childhood asthma. *J Allergy Clin Immunol* 2020; 147: 2108–2117.

16 Li CX, Wheelock CE, Sköld CM, *et al.* Integration of multi-omics datasets enables molecular classification of COPD. *Eur Respir J* 2018; 51: 1701930.

17 Tyler SR, Bunyavanich S. Leveraging -omics for asthma endotyping. *J Allergy Clin Immunol* 2019; 144: 13–23.

18 De Meulder B, Lefaudeux D, Bigler J, *et al.* The first U-BIOPRED blood handprint of severe asthma. *Eur Respir J* 2015; 46: Suppl. 59, PA4889.

19 De Meulder B, Tching Chi Yen R, *et al.* U-BIOPRED accessible handprint: combining omics platforms to identify stable asthma subphenotypes. *Eur Respir J* 2018; 52: Suppl. 62, OA3578.

20 Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: a review of machine learning interpretability methods. *Entropy (Basel)* 2020; 23: 18.

21 Ma T, Zhang A. Affinity network fusion and semi-supervised learning for cancer patient clustering. *Methods* 2018; 145: 16–24.

22 Jiang L, Xiao Y, Ding Y, *et al.* Discovering cancer subtypes *via* an accurate fusion strategy on multiple profile data. *Front Genet* 2019; 10: 20.

23 Ruan P, Wang Y, Shen R, *et al.* Using association signal annotations to boost similarity network fusion. *Bioinformatics* 2019; 35: 3718–3726.

24 Zhang Y, Hu X, Jiang X. Multi-view clustering of microbiome samples by robust similarity network fusion and spectral clustering. *IEEE/ACM Trans Comput Biol Bioinform* 2017; 14: 264–271.

25 Duan X, Wang K, Ke J, *et al.* Multiomics-based colorectal cancer molecular subtyping using local scaling network fusion. *J Comput Biol* 2020; 27: 1295–1302.

26 Hurgobin B, de Jong E, Bosco A. Insights into respiratory disease through bioinformatics. *Respirology* 2018; 23: 1117–1126.