PLOS ONE

RESEARCH ARTICLE

Network based stratification of major cancers by integrating somatic mutation and gene expression data

Zongzhen He, Junying Zhang*, Xiguo Yuan, Zhaowen Liu, Baobao Liu, Shouheng Tuo, Yajun Liu

School of Computer Science and Technology, Xidian University, Xi'an, PR China

* jyzhang@mail.xidian.edu.cn

Citation: He Z, Zhang J, Yuan X, Liu Z, Liu B, Tuo S, et al. (2017) Network based stratification of major cancers by integrating somatic mutation and gene expression data. PLoS ONE 12(5): e0177662. https://doi.org/10.1371/journal.pone.0177662

Editor: Enrique Hernandez-Lemus, Instituto Nacional de Medicina Genomica, MEXICO

Received: December 13, 2016

Accepted: May 1, 2017

Published: May 16, 2017

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by the Natural Science Foundation of China under Grants 61571341, and 61201312, and the Natural Science Foundation of Shaanxi Province in China under Grants 2016JM6047 and 2015JM6275. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

The stratification of cancer into subtypes that are significantly associated with clinical outcomes is beneficial for targeted prognosis and treatment. In this study, we integrated somatic mutation and gene expression data to identify clusters of patients. In contrast to previous studies, we constructed cancer-type-specific significant co-expression networks (SCNs) rather than using a fixed gene network across all cancers, such as the network-based stratification (NBS) method, which ignores cancer heterogeneity. For each type of cancer, the gene expression data were used to construct the SCN network, while the gene somatic mutation data were mapped onto the network, propagated, and used for further clustering. For the clustering, we adopted an improved network-regularized non-negative matrix factorization (netNMF) (netNMF_HC) for a more precise classification. We applied our method to various datasets, including ovarian cancer (OV), lung adenocarcinoma (LUAD) and uterine corpus endometrial carcinoma (UCEC) cohorts derived from the TCGA (The Cancer Genome Atlas) project. Based on the results, we evaluated the performance of our method to identify survival-relevant subtypes and further compared it to the NBS method, which adopts priori networks and netNMF algorithm. The proposed algorithm outperformed the NBS method in identifying informative cancer subtypes that were significantly associated with clinical outcomes in most cancer types we studied. In particular, our method identified survival-associated UCEC subtypes that were not identified by the NBS method. Our analysis indicated valid subtyping of patient could be applied by mutation data with cancer-type-specific SCNs and netNMF_HC for individual cancers because of specific cancer co-expression patterns and more precise clustering.

Introduction

Cancer is a heterogeneous disease which is formed by various subtypes. In an organ, different tumour subtypes is a reflection of certain molecular oncogenic processes and different clinical outcomes, which means these subtypes are supposed be regarded as different kinds of cancers in treatment design [1]. As cancer genomic, transcriptomic and epigenomic information is

Competing interests: The authors have declared that no competing interests exist.

becoming increasingly available, the stratification of tumours into valid clinic subtypes according to molecular data is crucial for guiding better treatment and prognosis. Due to the growth of mass high-throughput omics data, standard unsupervised clustering methods are used to cluster samples [2–4], such as non-negative matrix factorization (NMF) and hierarchical clustering. Computational methods can be used to identify tumour subtypes which have different survival rates, tumour levels or stage, histological types, and responses of treatment.

Key point of previous studies is utilizing messenger RNA (mRNA) expression data [2–3,5] to successfully group patients based on similarities in gene expression into clinically relevant phenotypes. While somatic mutations can make mutated genes lose function and offer more clinical guidance [6–8], these mutations are not universal phenomenon among patients [9–10]; thus, it is impossible to directly measure the similarity among tumours using mutated genes. The most advanced state-of-the-art integrative method for cancer analysis, network-based stratification (NBS), considers the sparsity of mutations by searching for mutational consistencies at the network level rather than at the individual gene level. This method can be used to identify patients' subgroups with similar molecular-network patterns by propagating mutation labels on a prior gene interaction network.

As the molecular network can be regarded as an obvious sign for interactions and relationships between molecules, it is adopted for the biological discovery of complex diseases [11–12]. NBS and current studies based on NBS all utilized the same prior networks across cancers [13–16]. However, we recognize that the regulation of gene expression levels among genes might be related to the type of cancer, and different cancers correspond to different regulations. Yang Y et al. constructed 4 cancer-type-specific co-expression networks (CNs) and revealed that, the hub genes which can be found in specific cancer networks are merely slightly overlap [17]. Thus, for a distinct cancer, the mutations should be mapped onto a network that was derived specifically from that cancer. A CN is constructed based on the correlations among the quantitative gene expression levels in tumours, and the CN is presented as a graph in which the nodes correspond to genes, and the edges correspond to co-expressions among the genes. A CN contains more precise information regarding the connectivity among genes in individual cancer types than prior networks. CNs have been shown to be helpful for describing the pair wise relationships among gene transcripts [18–19]. Genes with similar or correlated expression patterns might contribute to the same regulatory function, and gene co-expression patterns in a CN may lead to more insightful discoveries regarding the underlying regulatory mechanisms [20–21].

Thus, we propose a method based on NBS for the stratification of cancer by constructing a gene network for each cancer. In this study, we integrated somatic mutation data and gene expression data. For each cancer, the gene expression data were used to construct a CN, and the somatic mutations in the genes were mapped into the network and propagated, which was useful for further clustering.

Furthermore, network-regularized NMF (netNMF) clustering using the NBS method has been shown to be better than the standard NMF method. netNMF first maps the samples with smoothed mutations into a lower k-dimensional feature vector space using netNMF matrix factorization, which constrains the genes with respect to the gene interaction network; then, the sample category is determined by the column with the largest value among the k feature vectors. This clustering is reasonable because it is possible to select the class that has the largest weight concerning the most relevant feature vector [7], but the class cannot be easily estimated if there are two similarly large values. The class of the samples may be more precisely identified by clustering the factorized low rank feature vectors of the samples using a clustering method, such as hierarchical clustering.

In this study, we propose an improved stratification method based on NBS through combining gene expression data and somatic mutation. First, we constructed a significant co-expression network (SCN) using gene expression profiles for each cancer type. Then, for patients, we projected the profile of mutation onto the cancer-type-specific SCN network. Network propagation was applied to diffuse the effect of mutation over its network neighbourhood. Finally, the matrix of the 'network-smoothed' was stratified into different subtypes with numbers ranging from 2 to 8 via the netNMF_HC clustering method. The effectiveness of our method was evaluated based on the relevance of our subtypes and clinical outcomes and compared with NBS, which used prior gene networks and the netNMF clustering method. The results showed that our method outperformed NBS for the three cancers and identified the survival-relevant subtypes of uterine endometrial carcinoma (UCEC), which were not identified by NBS. The workflow of our method is described in Fig 1.

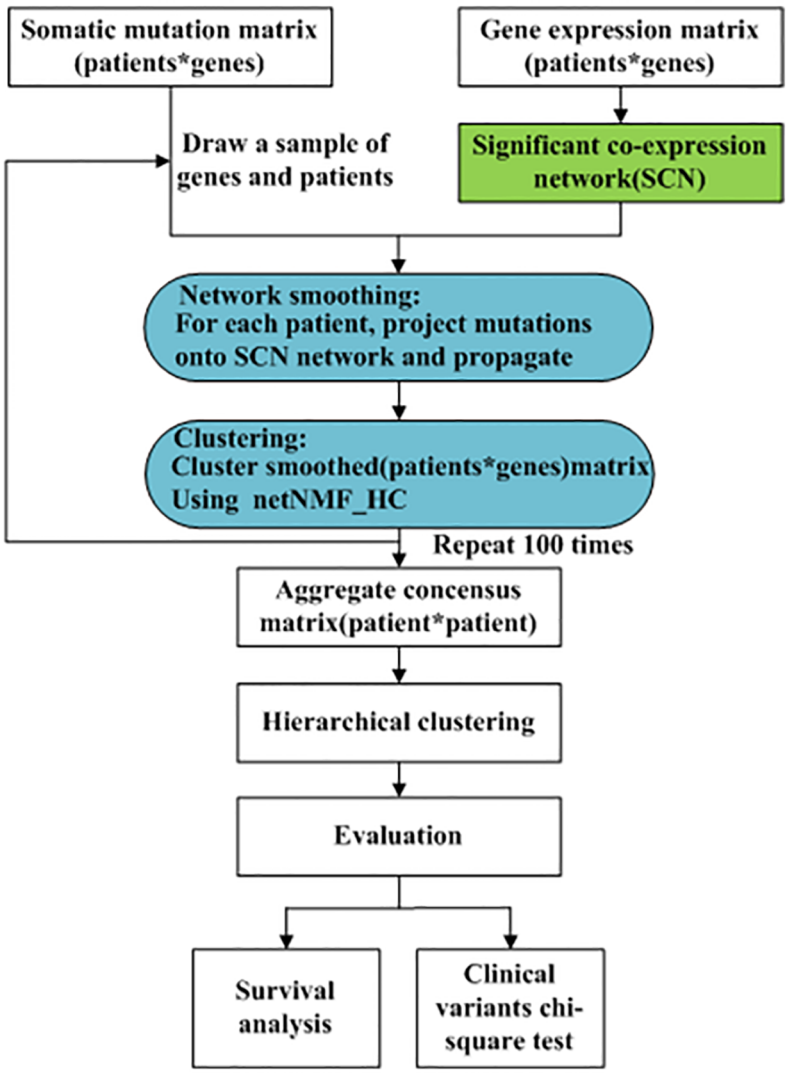

Fig 1. Workflow of our method.

https://doi.org/10.1371/journal.pone.0177662.g001

Materials and methods

Data and pre-processing

TCGA somatic mutation data and mRNA expression data. For comparing with NBS fairly, the somatic mutation profiles for serous ovarian cancer (OV), UCEC, and lung adenocarcinoma (LUAD) from TCGA and two prior gene-gene interaction networks STRING and Humannet were collected from materials of Hofree et al [4]. The RNAseqv2 gene expression profiles for the three cancers were downloaded from the Synapse database (https://www.synapse.org/#!Synapse:syn300013) in which the expression levels are normalized using MAS5. The expression data were used to construct an SCN for further analysis. As shown in Table 1, the ovarian carcinoma dataset contained somatic mutations in 9,850 genes from 356 samples and gene expression profiles for 20,534 genes from 430 samples. The LUAD dataset contained somatic mutations in 15,967 genes from 381samples and gene expression profiles for 20,199 genes from 576 samples. The UCEC dataset contained somatic mutations in 17,968 genes from 248 samples and gene expression profiles for 20,531 genes from 381 samples.

First, the mutation matrix F was generated based on samples with somatic mutations. F is binary as follows: if any gene mutates (a single nucleotide base change or an insertion or deletion of bases) in a certain sample relative to the germ line, the mutation is marked the number 1; otherwise, 0 is assigned. The expression matrix E is a real matrix, and each of its entries indicates a normalized given gene expression in a given sample. In all matrices, the samples and genes are represented by the rows and columns. We filtered samples with fewer than 10 mutations in mutation matrix F and genes with 0 expression in all samples in the gene expression matrix E. The clinical data, including survival, stage and grade of the three cancers, were gotten from the Synapse database (https://www.synapse.org/#!Synapse:syn300013) and were applied in evaluation of the relevance of the identified subtypes and clinical outcomes.

Gene interaction networks. The patient mutations were projected onto the gene interaction networks as follows: NBS utilizes STRING v.9 [22] and HumanNet v.1 [23], and our method employs an SCN that was constructed by gene expression profiles.

Method

Construction of the SCN. For each cancer type, an SCN was constructed to represent the significant correlations between a pair of genes without a prior interaction network. First, the absolute Spearman's rank correlations and the corresponding p-values of these correlations were calculated for all gene pairs based on the expression profile matrix of each cancer type. The original co-expression network (CN) was constructed based on the absolute correlations as the weight of the edges. Then, the expression of gene pairs was considered significantly correlated if the q-value (Bonferroni corrected p-value) of their correlation was smaller than 0.05. The SCN was obtained by filtering the edges with correlations whose corresponding q-values were greater than 0.05. Thus, the SCN is an unweighted undirected network in which each vertex denotes a gene used in our work. Spearman's rank correlation was chosen as a measure to

Table 1. Sample sizes, somatic mutations and gene expression profiles for three cancers.

	Somatic mutation data		Gene expression profiles	
	Sample size	Number of genes	Sample size	Number of genes
OV	356	9,850	430	20,534
LUAD	381	15,967	576	20,199
UCEC	248	17,968	381	20,531

https://doi.org/10.1371/journal.pone.0177662.t001

evaluate the relevance of two genes because it can detect nonlinear relationships, and it has been verified to have better performance than other measures, such as Euclidean distances and Pearson's correlations, in measuring the similarity between genes [24]. Spearman's rank correlations and the corresponding p-values were calculated using the function corr() in MATLAB.

Finally, three cancer-type-specific SCNs were obtained, in which two genes are connected if they are estimated to be significantly co-expressed.

Network smoothing. For each patient of each cancer, we mapped the mutation profile onto the constructed cancer-type-specific SCN. Network propagation [25] was used to propagate the mutation signal among networks. The key is to spread the mutation information of every gene to its neighbours iteratively until a stable state is achieved. The algorithm used is as follows:

Step1: Construct the degree-normalized matrix $W' = D^{-1/2}WD^{-1/2}$, where D is a diagonal matrix, and its columns sums W on the diagonal; W is the adjacency matrix of the SCN network; and the diagonal elements of W are set to zero. The normalized matrix W' is utilized for the following smooth process.

Step2: Iterate $F_{t+1} = \alpha F_t W' + (1-\alpha)F_0$ until convergence is achieved (the matrix norm of $F_{t+1} - F_t < 1\times10^{-7}$), where F_0 is the patient-by-gene somatic mutation matrix, and the parameter α is a tuning parameter in (0, 1) that governs the relative amount of the information from the gene's neighbours and its initial mutation information. It should be noted that self-reinforcement should be avoided because the diagonal elements of the adjacency matrix W are set to zero at the beginning. The optimal value of α is set as 0.7and is represented in NBS [4]. Each row in F_t represents the smoothed mutation of genes in a sample after it is influenced by its neighbours.

Step3: For F_t, quantile normalization is regarded as the guarantee for patient to follow the same smoothed mutation profile distribution. F was applied to show the transpose of the final normalized and smoothed mutation matrix.

Improved netNMF_HC. NMF is a matrix factorization algorithm which can resolve a matrix into two lower rank non-negative matrices [26]. netNMF can be regarded as an evolution of NMF which rules NMF to keep the gene interaction network structure. The objective is producing two non-negative matrices, W and H, to minimize the following function using an iterative method [27]:

$$\min_{W,H>0} \|F - WH\|_F^2 + \lambda \mathrm{trace}(W^t L W) \quad (1)$$

Where $\|.\|_F^2$ denotes the matrix Frobenius norm, W and H are decompositions of the smoothed m by n matrix F. W is an m by k basis matrix or "metagenes", and H is a k by n coefficient matrix. The reduced dimension is controlled by the value k and k is set 2~8.

L is the graph Laplacian of the p-nearest-neighbour network derived from the original weighted gene co-expression network (CN). If v_i and v_j are two connected vertices in the original network, the weight of edge w_{ij} of the p-nearest neighbour network w is as follows:

$$w_{ij} = \begin{cases} 1, & \text{if } v_i \in N_p(v_j) \text{ or } v_j \in N_p(v_i) \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

where $N_p(v_i)$ indicates the set of p-nearest neighbours of v_i. The graph Laplacian of w is $L = D(w) - w$, where D is a diagonal matrix with the sums of a column (or row as w is

symmetrical) of w $D_{ii} = \Sigma_j w_{ij}$ on the diagonal line. We set the number of nearest neighbours $p = 11$, and the regularization parameter λ was set to 200, which are on the same scale in NBS [4]. The function was run iteratively until it converged ($||F–WH|| < 1\times10^{-3}$).

Finally, in previous works, the class of the samples was obtained from the coefficient matrix H. H is a k by n coefficient matrix, indicating n samples with k feature vectors. For a sample n, class $c_n = \arg\max_k(H_{kn})$; therefore, the sample belongs to the column number of the feature vector with the largest value.

It is not precise enough to use the netNMF method for determining the category of n samples; for example, if two vector values are almost the same, it may be difficult to determine the final class. Therefore, we propose the improved method netNMF_HC, which considers the k by n coefficient matrix H from netNMF to be a low dimension feature space of the patients, and then, H is utilized to group samples by hierarchical clustering (HC) to obtain the class of the patients.

Consensus clustering. Robust clustering was achieved by applying consensus clustering [28] to produce the final subtypes. Precisely, 80% of the patients and genes were sampled randomly for network smoothing and netNMF_HC was used to perform the clustering. The process was repeated 100 times. The results of the 100 clustering made up the patient-patient similarity matrix. The matrix recorded the frequency of the sampling of each pair of patients and the rates at which the pairs were clustered in same group among all replicates. According to the similarity matrix, hierarchical clustering with average linkage can be produced.

Clinical analysis. The analysis of survival was generated by the R "survival" package. Kaplan-Meier survival curves of subtypes and log-rank test p-values were utilized to evaluate the association between the subtypes and patient 10 year survival time. Pearson's chi-squared test was aimed to evaluate the relationship between the subtypes and tumour grade or stage.

Identification of differentially mutated genes in subgroups. The significantly mutated genes of each subtype relative to the remaining subtypes were identified using the significance analysis of microarrays (SAM) method [29] with the network-smoothed mutation data. The q-values were estimated by SAM with the Wilcoxon-rank statistic and 1,000 permutations. Then, the differentially mutated genes with q-values$<$0.05 for each subtype were selected for further analysis.

We performed a biological processes and pathway enrichment analysis for the significantly differentially mutated genes in each subtype using DAVID 6.8 (https://david.ncifcrf.gov/). Only enriched annotation terms whose q-values were lower than 0.05 were retained.

Results

We tested our method and the NBS method in ovarian, uterine and lung cancer. In our method, the somatic mutations in these cancer types were propagated onto an SCN, and then, the smoothed mutation profiles were clustered with consensus clustering based on netNMF_HC. In NBS, the somatic mutations in these cancer types were mapped onto the gene interaction networks STRING and Humannet, and then, the smoothed mutation profiles were clustered with consensus clustering based on netNMF. To determine whether the SCN network or the improved clustering method netNMF_HC contribute to our method, we also performed experiments in which only changing the network or the clustering method. The clustering outcomes (cluster number k = 2~8) for each kind of cancer are shown in the S1 File. Finally, different outcomes were observed for a given cancer type and a number of clusters k.

Table 2. Critical relationship between subtypes and survival in 3 tumour types using NBS and improved methods.

Survival p-value	Cluster number k	NBS		Only changing network	Only changing clustering method		Our method
		STRING +netNMF	Humannet +netNMF	SCN+ netNMF	STRING +netNMF_HC	Humannet +netNMF_HC	SCN +NetNMF_HC
OV	4	0.058222	0.040828	0.007334	0.00792	0.003891	7.70E-07
LUAD	6	0.177624	0.014901	0.033739	0.076693	0.07448	0.024237
UCEC	3	0.260764	0.913596	0.080808	0.248289	0.491311	0.000314

https://doi.org/10.1371/journal.pone.0177662.t002

Clinical analysis

The relationship between subtypes and 10 year survival was investigated in the first time. All three cancer subtypes derived from our method were significantly associated with survival in certain cluster numbers (log-rank test p-value < 0.05) as shown in Table 2 and Fig 2. Compared with NBS, when only the network was changed, the cancer-specific SCN performed better than prior network STRING in discovering clinically relevant subtypes for OV and LUAD; when only changing the clustering method, the improved clustering method netNMF_HC performed better than netNMF for OV. When both were changed simultaneously, our method performed better than the NBS method for all three cancers. Additionally, using our method (SCN + netNMF_HC), the OV, LUAD and UCEC samples were divided into 4, 6 and 3 clusters respectively having most significant association with the survival time, each cluster was independent and differed in survival (Fig 3), while the NBS method using the STRING network was less effective (p-values were not significant). Especially, the survival-relevant UCEC subtypes could not be obtained with NBS based on the STRING or Humannet network, which is consistent with a previous work [4]. We then measured the relationship between the subtypes and the tumour grade or stage. Among three kinds of cancer, there was no relationship between the clusters and the tumour grade or stage, except for UCEC.

Overall, the different networks applied in the work influenced the stratification results. The SCN network performed better than the prior networks STRING and Humannet in most of the studied cancer types. Although the clustering method improvement did not contribute much, the combination of SCN network and netNMF_HC clustering method achieved better performance than NBS for all three cancers.

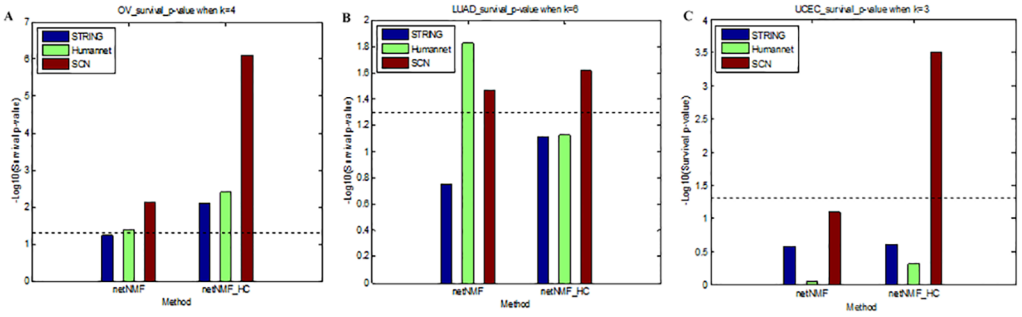

Fig 2. Survival p-values in three cancers with distinct methods. (A) Significance with -Log10(p-value) association between 10 year survival and subtypes obtained by distinct combination of networks (STRING (blue), Humannet(green) and Significant co-expression network(SCN)(red)) and clustering methods (netNMF and netNMF_HC) for ovarian cancer (OV) with cluster number k = 4. (B) for lung cancer(LUAD) with cluster number k = 6. (C) for uterine cancer(UCEC) with cluster number k = 3. Dashed lines represent the -log10 (P = 0.05) threshold.

https://doi.org/10.1371/journal.pone.0177662.g002

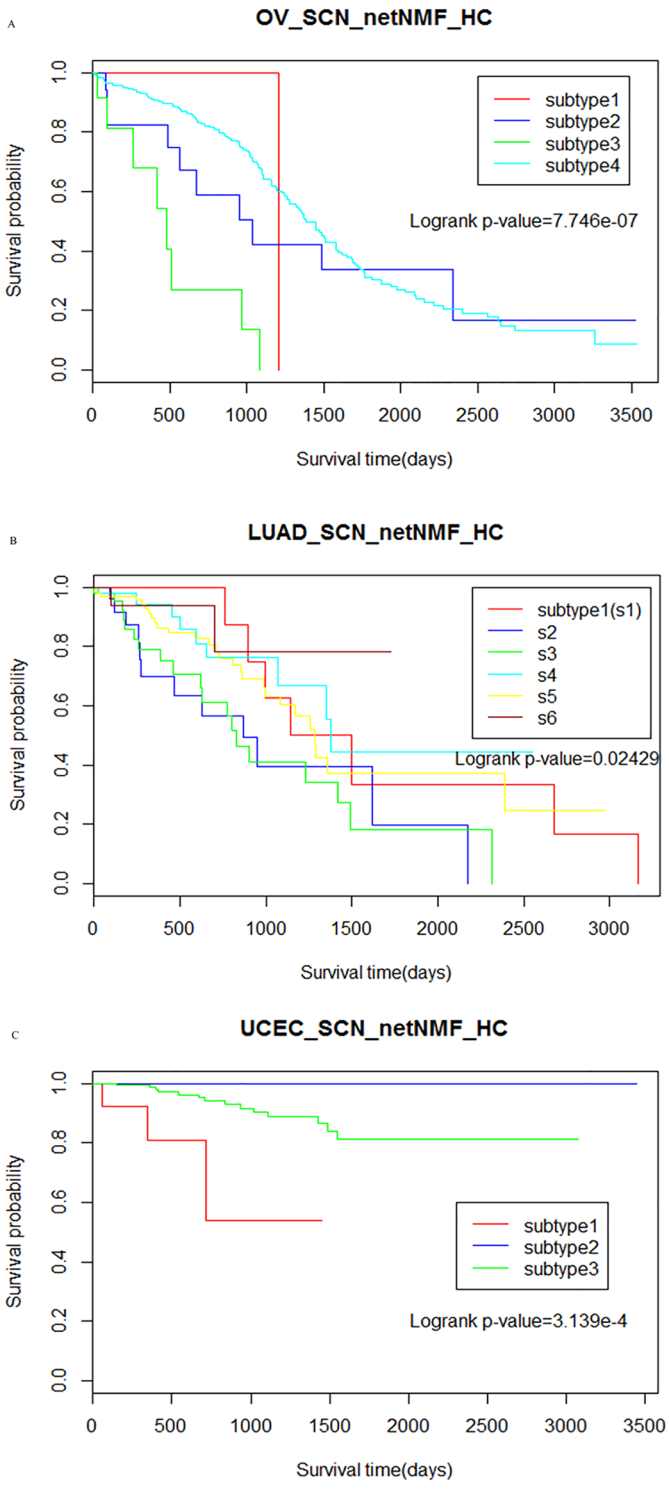

Fig 3. Survival curves of three cancers with our method. (A) Kaplan-Meier survival plots of subtypes obtained by our method that combining the significant co-expression network(SCN) and clustering method netNMF_HC for ovarian cancer (OV) with cluster number k = 4. (B) for lung cancer(LUAD) with cluster number k = 6. (C) for uterine cancer(UCEC) with cluster number k = 3.

https://doi.org/10.1371/journal.pone.0177662.g003

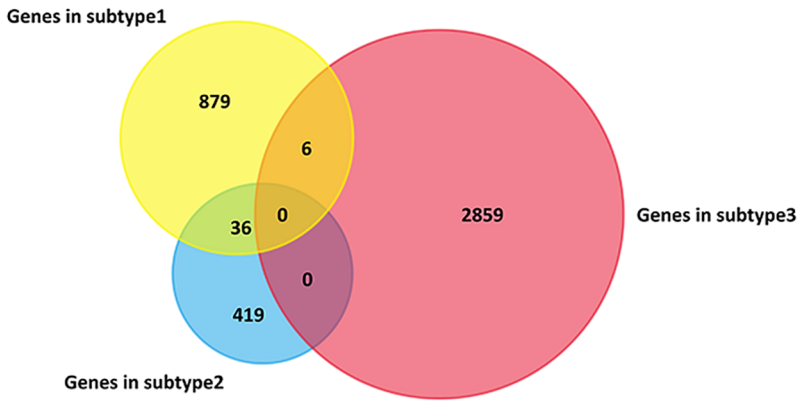

Fig 4. Overlap of genes distinguishing the three subtypes of UCEC.

https://doi.org/10.1371/journal.pone.0177662.g004

Identifying differentially mutated genes in subgroups

We further identified the significantly differentially mutated genes in each subtype of UCEC for instance. The overlap of these gene sets is very few, which means these genes are specific to certain subtype (Fig 4). In addition, the enriched biological processes and pathways of these genes are also distinct for different subtypes (Fig 5).

Mutation pattern analysis

As shown in Fig 3C, three UCEC subtypes were obtained. Cluster 1 (red) has the worst survival and cluster 2 (blue) has the best survival. The mutation patterns (before the network smoothing) of the three UCEC subtypes which are predictive of survival was analysed. As shown in Fig 6, the three subtypes have different mutation schemas, and cluster 3 harboured more mutations than the other clusters. Both *PIK3CA* and *PTEN* alterations have been reported to have strong relationships with UCEC [30].

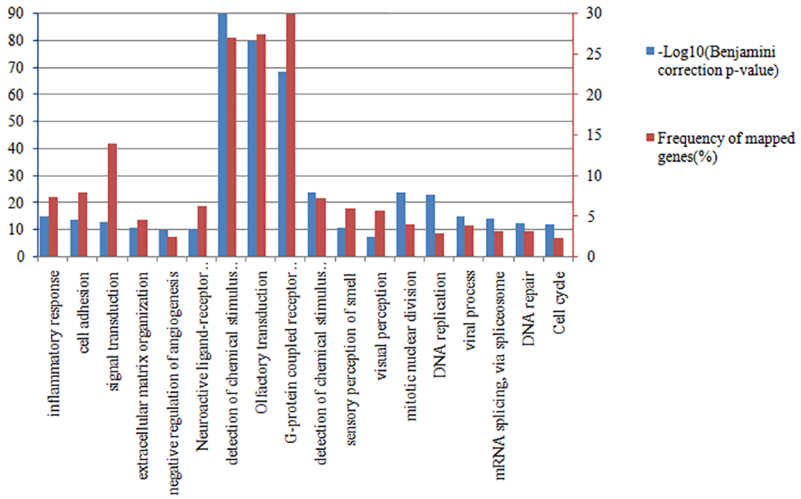

Fig 5. Top enriched biological processes and pathways of the differentially mutated genes in each subtype of UCEC. Bars indicate the significance with -log10(Benjamini correction p-value) (blue) and the frequency of the mapped genes (red) of the corresponding function.

https://doi.org/10.1371/journal.pone.0177662.g005

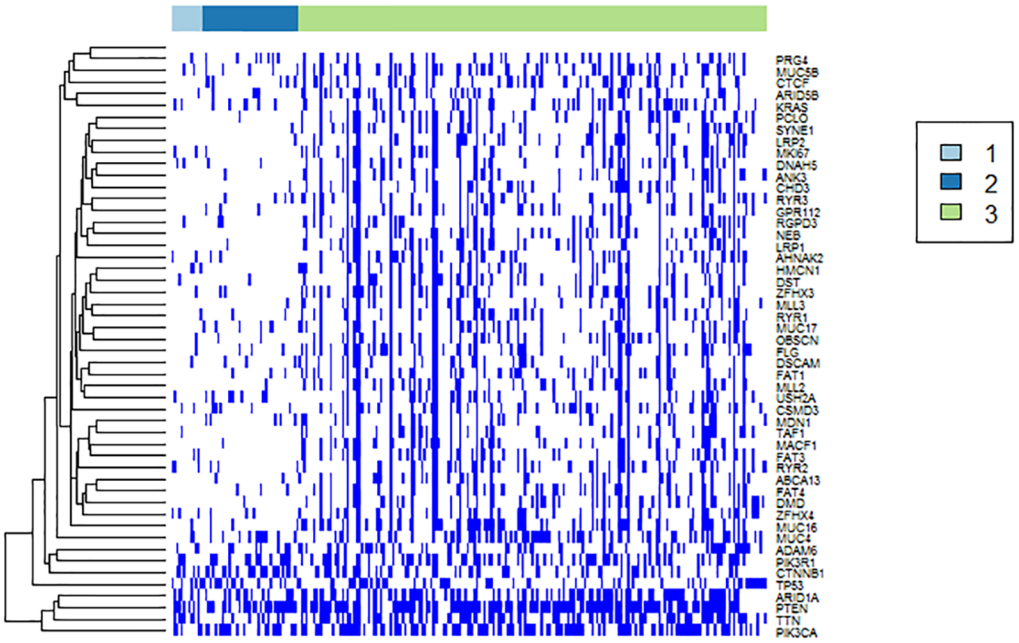

Fig 6. Somatic mutation patterns in three UCEC subtypes. Each row indicates a highly frequent gene, and each column denotes a sample. Dark colours in the figure indicate that a mutation occurred in a gene in a sample.

https://doi.org/10.1371/journal.pone.0177662.g006

Discussion

Exome and whole genome sequencing have provided a large amount of genomic and transcriptome data. These data enable the stratification of patients into clinically relevant subtypes, making molecular-driven diagnoses and therapy feasible. Network-based methods have integrated mutations and prior gene interaction networks to identify the clinically relevant subtypes. In this study, we showed that due to tumour heterogeneity, combination of a cancer-type-specific SCN and improved clustering method for each cancer type can achieve a superior stratification compared to using the prior fixed gene network for all cancers and often also has a better predictive performance of survival. This finding indicated that the cancer-type-specific gene SCNs can offer useful individual cancer biological knowledge for effective subtyping.

Clinically relevant tumour subtypes of some cancer types may be driven by various mechanisms, such as copy number aberration or methylation, besides somatic mutations and gene expression levels. Integrating multiple types of molecular data to discover truly predictive subtypes is essential for the future. Based on our analysis, the top 50 genes that are mutated frequently across tumour samples (generally called mutation drivers) can distinguish the UCEC subtypes by the mutation patterns. Driver genes may be beneficial for stratification because they are signatures that can capture the differences among the subtypes. In conclusion, clinically relevant subtyping performance may be further improved by integrating clinical driver genes obtained from integrated molecular data and cancer-specific networks.

Supporting information

S1 File. Significant p-value of association between subtypes and survival for three cancers.
(PDF)

Acknowledgments

This work was supported by the Natural Science Foundation of China under Grants 61571341, and 61201312, and the Natural Science Foundation of Shaanxi Province in China under Grants 2016JM6047 and 2015JM6275.

Author Contributions

Conceptualization: ZH JZ.

Data curation: ZH.

Formal analysis: ZH.

Funding acquisition: JZ.

Investigation: ZH.

Methodology: ZH.

Project administration: ZH JZ.

Resources: JZ.

Software: ZH.

Supervision: JZ.

Writing – original draft: ZH.

Writing – review & editing: ZH JZ XY ZL BL ST YL.

References

1. Liu Z, Zhang XS, Zhang S. Breast tumor subgroups reveal diverse clinical prognostic power. Scientific Reports. 2014; 4(2):4002.
2. Network TCGA. Integrated genomic analyses of ovarian carcinoma. Nature. 2011; 474(7419):609–15.
3. Reis-Filho JS, Pusztai L. Breast Cancer 2 Gene expression profiling in breast cancer: classification, prognostication, and prediction. 2011; 378(9805):1812–23.
4. Hofree M, Shen JP, Carter H, Gross A, Ideker T. Network-based stratification of tumor mutations. Nature Methods. 2013; 10(11).
5. Tothill RW, Tinker AV, George J, Brown R, Fox SB, Lade S, et al. Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. Clinical Cancer Research An Official Journal of the American Association for Cancer Research. 2008; 14(16):5198–208. https://doi.org/10.1158/1078-0432.CCR-08-0196 PMID: 18698038
6. Chmielecki J, Pietanza MC, Aftab D, Shen R, Zhao Z, Chen X, et al. EGFR mutant lung adenocarcinomas treated first-line with the novel EGFR inhibitor, XL647, can subsequently retain moderate sensitivity to erlotinib. Journal of Thoracic Oncology Official Publication of the International Association for the Study of Lung Cancer. 2012; 7(7):434–42.
7. Olivier M, Taniere P. Somatic mutations in cancer prognosis and prediction: lessons from TP53 and EGFR genes. Current Opinion in Oncology. 2011; 23(1):88–92. https://doi.org/10.1097/CCO.0b013e3283412dfa PMID: 21045690
8. Pirazzoli V, Nebhan C, Song X, Wurtz A, Walther Z, Cai G, et al. Acquired resistance of EGFR-mutant lung adenocarcinomas to afatinib plus cetuximab is associated with activation of mTORC1. Cell Reports. 2014; 7(4):999–1008. https://doi.org/10.1016/j.celrep.2014.04.014 PMID: 24813888
9. Lawrence MS, Stojanov P, Mermel CH, Garraway LA, Golub TR, Meyerson M, et al. Discovery and saturation analysis of cancer genes across 21 tumor types. Nature. 2014; 505(7484):495–501. https://doi.org/10.1038/nature12912 PMID: 24390350
10. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer genes. Nature. 2015; 499(7457):214–8.

11. Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell. 2014; 158 (4):929–44. https://doi.org/10.1016/j.cell.2014.06.049 PMID: 25109877
12. Ideker T, Sharan R. Protein networks in disease. Genome Res. Genome Research. 2008; 18(4):644–52. https://doi.org/10.1101/gr.071852.107 PMID: 18381899
13. Xue Z, Yang H, Zhao S, Yu S, Li B. Network-based stratification analysis of 13 major cancer types using mutations in panels of cancer genes. BMC Genomics. 2015; 16(7):1–8.
14. Liu Z, Zhang S. Tumor characterization and stratification by integrated molecular profiles reveals essential pan-cancer features. BMC Genomics. 2015; 16(1):1–12.
15. Ahmadi AA, Qian X. Tumor stratification by a novel graph-regularized bi-clique finding algorithm. Computational Biology & Chemistry. 2015; 57(C):3–11.
16. Van TL, Leeuwen MV, Fierro AC, Maeyer DD, Eynden JVD, Verbeke L, et al. Simultaneous discovery of cancer subtypes and subtype features by molecular data integration. Bioinformatics. 2016; 32(17).
17. Sun P, Cha BR, Dong UA. Automatic Multi-document Summarization Based on Clustering and Nonnegative Matrix Factorization. Iete Technical Review. 2010; 27(2):167–78.
18. Chuang CL, Jen CH, Chen CM, Shieh GS. A pattern recognition approach to infer time-lagged genetic interactions. Bioinformatics. 2008; 24(9):1183–90. https://doi.org/10.1093/bioinformatics/btn098 PMID: 18337258
19. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008; 9(1):1–13.
20. Yang Y, Han L, Yuan Y, Li J, Hei N, Liang H. Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. Nature Communications. 2014; 5(1).
21. Carlson MR, Zhang B, Fang Z, Mischel PS, Horvath S, Nelson SF. Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. BMC Genomics. 2006; 7(1):1–15.
22. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, et al. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Research. 2011; 39(suppl_1):D561–D8.
23. Lee I, Blom UM, Wang PI, Shim JE, Marcotte EM. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome Research. 2011; 21(7):1109–21. https://doi.org/10.1101/gr.118992.110 PMID: 21536720
24. Jayaswal V, Lutherborrow M, Ma D D F, et al. Identification of microrna with regulatory potential using a matched microrna-mrna time-course data. Nucleic Acids Research. 2009; 37(8):e60. https://doi.org/10.1093/nar/gkp153 PMID: 19295134
25. Zhou D, Bousquet O, Lal TN, Weston J, Olkopf BS. Learning with Local and Global Consistency. Advances in Neural Information Processing Systems. 2004; 16(4):321–8.
26. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature. 1999; 401(6755):788–91. https://doi.org/10.1038/44565 PMID: 10548103
27. Cai D, He X, Wu X, Han J. Non-negative Matrix Factorization on Manifold. 2008:63–72.
28. Monti S, Tamayo P, Mesirov J, Golub T. Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. Machine Learning. 2003; 52(1):91–118.
29. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences. 2001; 98(9):5116–21.
30. Jhawer M, Goel S, Wilson AJ, Montagna C, Ling YH, Byun DS, et al. PIK3CA mutation/PTEN expression status predicts response of colon cancer cells to the epidermal growth factor receptor inhibitor cetuximab. Cancer Research. 2008; 68(6):1953–61. https://doi.org/10.1158/0008-5472.CAN-07-5659 PMID: 18339877