Tensor Decomposition for Sub-typing of Complex Diseases based on Clinical and Genomic Data

Diana Diaz
Dep. of Computer Science
Wayne State University
Detroit, USA
dmd@wayne.edu

Aliccia Bollig-Fischer
Dep. of Oncology, School of Medicine
Wayne State University and
Karmanos Cancer Institute, Detroit, USA
bollig@karmanos.org

Alexander Kotov
Dep. of Computer Science
Wayne State University
Detroit, USA
kotov@wayne.edu

Abstract—**It has long been understood that stratification of patients into fine-grained cohorts is a foundation of accurate diagnosis and effective treatment of complex diseases, such as cancer. Nevertheless, cancer therapies still fail or cause unnecessary suffering to many patients, which suggests that our current understanding of cancer sub-types needs to be refined. In this paper, we propose CLIGEN, a novel computational pipeline for high-throughput data-driven stratification of patients with a complex disease into cohorts corresponding to multi-modal disease sub-types based on clinical and genomic data. We applied CLIGEN to discover breast cancer sub-types based on the clinical and genomic data of 503 patients with breast ductal carcinoma in the Cancer Genome Atlas (TCGA). Quantitative and qualitative evaluation of the breast cancer sub-types discovered by CLIGEN indicate that they are biologically meaningful and correlate with clinical outcomes, such as patient survival time.**

Index Terms—**disease sub-typing, tensor factorization, somatic mutations, breast cancer, complex diseases**

I. Introduction

It is now a well-accepted fact that each type of cancer is a collection of distinct genetic diseases characterized by multiple dysregulations at different levels of a biological system. Substantial degree of heterogeneity significantly complicates accurate diagnosis and effective treatment of cancers. Although it has long been understood that patient stratification into fine-grained cohorts that correspond to disease sub-types and identification of clinically actionable markers characteristic of these cohorts are the central tenets of effective treatment of complex diseases, such as cancer, many cancer patients are still misdiagnosed [1] and, as a result, either receive unnecessary surgery or chemotherapy [2] or do not receive the needed treatment. Therefore, **cancer diagnosis and treatment can greatly benefit from computational methods for fine-grained and comprehensive cancer sub-typing**.

Remarkable advances in next-generation sequencing coupled with widespread adoption of electronic health records by healthcare delivery systems have enabled collection of unprecedented amounts of clinical and genomic patient data. However, despite the recent progress in computational methods to analyze clinical or genomic data in isolation, neither of these types of data can capture all aspects of pathogenesis of complex diseases, such as cancer [3]. In particular, past attempts at cancer patient stratification and treatment selection based on clinical [4], genomic [5], [6] or transcriptomic [7]–[9] data alone have yielded only modest success thus far. Since cancer development and progression are influenced by multiple factors, including germ-line or somatic tumor genetics, overall patient health as well as patient demographics [10], it is natural to assume that cancer sub-types should account for these modalities of patient data. However, there has been *relatively little research on computational methods for sub-typing of complex diseases based on clinical and genomic data*.

To address this limitation, we propose **CLIGEN**, a computational pipeline for fully-unsupervised sub-typing of complex diseases by *integrating micro- (genomic) and macro-level (demographic and clinical) patient data*. In the case of cancer, *CLIGEN* takes as input both clinical and genomic data of a given population of cancer patients, which includes demographic attributes of patients along with somatic mutation profiles and clinical properties of their tumors, and identifies the patient cohorts within a given population that share demographic attributes, somatic gene mutations and clinical properties of tumors. We hypothesize that cancer patient cohorts discovered by *CLIGEN* from genomic and clinical cancer patient data: i) correspond to entirely novel or refined existing cancer sub-types characterized by both molecular and clinical markers ii) allow to shed additional light on complex interactions between clinical outcomes, such as survival time, and patient demographics, molecular aberrations and clinical properties of tumors.

II. Materials

The dataset for experiments in this work was constructed based on the patient data from The Cancer Genome Atlas - Genomic Data Commons Data Portal[a] (TCGA-GDC) [11]. Specifically, we used somatic mutation (non-silent mutation from the whole exome sequencing level 3) profiles and clinical data of patients with breast ductal carcinoma. Out of 825 breast cancer patients in TCGA, we considered only the patients for whom both somatic mutation and clinical data were available (507 patients) and also discarded the patients with fewer than 10 somatic mutations (4 patients). The resulting dataset consists of the somatic mutation profiles over 11,996 genes and

[a] downloaded from Bioportal on April 9th, 2017

978-1-7281-1867-3/19/$31.00 ©2019 IEEE

70 values and value ranges of 11 discrete and dichotomized continuous clinical variables of 503 patients.

Somatic Mutations. The TCGA somatic mutation table aggregates information about mutations in breast cancer patient tumors and consists of 37 columns and 34032 rows. A row in this table corresponds to a mutation in the gene designated in the column "*Hugo Symbol*" for the tumor sample in the column "*Tumor Sample Barcode*". Additional details regarding the organization of TCGA data are available in [11]. We constructed patient mutation profiles as binary vectors with 11,996 dimensions, in which a bit is set, if the patient's gene corresponding to that dimension in the vector harbors at least one non-silent mutation (i.e. missense mutation, nonsense mutation, non-stop mutation, frameshift mutation, in-frame insertion or in-frame deletion).

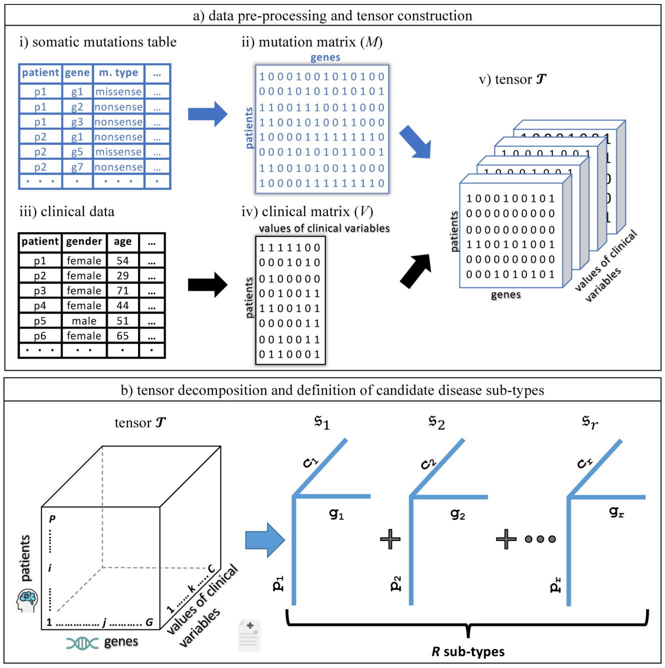
Clinical and Demographic Variables. The 70 values and value ranges (further referred to as the values of clinical variables) were obtained from 11 discrete and continuous clinical and demographic variables in TCGA, which include the age of cancer diagnosis (dichotomized into 10 value ranges), gender (2 values), estrogen receptor (ER) status (3 values), progesterone receptor (PR) status (3 values), human epidermal growth factor receptor 2 (HER2) final status (3 values), American Joint Committee on Cancer (AJCC) breast cancer stage (9 values), AJCC coded tumor stage (2 values), AJCC coded lymph node stage (2 values), AJCC coded metastasis stage (2 values), immunohistochemistry expression level (dichotomized into 29 value ranges) and PAM50 profile (5 values). Detailed description of these variables can be found in [12].

III. Method

Figure 1 provides a graphical overview of *CLIGEN*[b], the proposed pipeline for unsupervised sub-typing of complex diseases. *CLIGEN* consists of three stages: i) data pre-processing ii) tensor construction and iii) non-negative decomposition of the constructed tensor to derive *multi-modal disease sub-types*. Each of these stages is discussed in detail below.

Data pre-processing. The first stage of *CLIGEN* illustrated in Figure 1.a involves pre-processing the input to create a *combined multi-dimensional representation* of genomic and clinical patient data for subsequent analysis. Given the input mutation table, *CLIGEN* constructs a binary mutation matrix $\mathbf{M}$ with patients as rows and genes as columns. A value of the cell $\mathbf{M}_{ij}$ of matrix $\mathbf{M}$ is set to 1, if the ith patient has a mutation in the jth gene and to 0, otherwise. Continuous clinical variables, such as the age of cancer diagnosis, are discretized into intervals. The values of clinical variables for all patients are represented as a binary clinical matrix $\mathbf{V}$ with patients as rows and values of clinical variables as columns. A value of the cell $\mathbf{V}_{ik}$ of matrix $\mathbf{V}$ is set to 1, if the ith patient has the kth value of clinical variables.

Tensor construction. Matrices $\mathbf{M}$ and $\mathbf{V}$ are combined to create a three-dimensional binary tensor (i.e. multidimensional



patient	gene	m. type	...
p1	g1	missense	...
p1	g2	nonsense	...
p1	g3	nonsense	...
p2	g1	nonsense	...
p2	g5	missense	...
p2	g7	nonsense	...
. . .	.	. . .	.

patient	gender	age	...
p1	female	54	...
p2	female	29	...
p3	female	71	...
p4	female	44	...
p5	male	51	...
p6	female	65	...
. . .	. .	.	.

Fig. 1. Stages of the *CLIGEN* pipeline: a) data pre-processing and construction of three-dimensional tensor τ b) deriving multi-modal sub-types via CP decomposition of tensor τ.

array) $\tau \in \mathbb{R}^{P \times G \times C}$, which captures interactions between somatic mutations and clinical variables. The first mode of tensor τ corresponds to P patients in the population, while the other two modes correspond to G distinct genes and C distinct values of clinical variables (Figure 1.b). A value of the cell t_{ijk} of tensor τ is set to 1, if the ith patient has at least one mutation in the jth gene and the kth value of clinical variables and to 0, otherwise.

Tensor decomposition. *CLIGEN* utilizes Canonical Decomposition (CANDECOMP) and Parallel Factor Analysis (PARAFAC) or CP tensor factorization [13] to identify disease sub-types as groups of latent factors in τ. CP decomposition approximates τ with $\hat{\tau}$, a linear combination of rank-one tensors. Formally:

$$\tau \approx \hat{\tau} = [\![\lambda, \mathbf{P}, \mathbf{G}, \mathbf{C}]\!] = \sum_{r=1}^{R} \lambda_r \cdot \mathfrak{s}_r = \sum_{r=1}^{R} \lambda_r \cdot \mathbf{p}_r \circ \mathbf{g}_r \circ \mathbf{c}_r \quad (1)$$

where R is the number of rank-one tensors $\mathfrak{s}_r$ that τ is decomposed into, $\lambda_r \in \mathbb{R}$ is the weight of the rth rank-one tensor. Each $\mathfrak{s}_r$ is an outer product ($\circ$) of patient $\mathbf{p}_r \in \mathbb{R}^P$, gene $\mathbf{g}_r \in \mathbb{R}^G$ and clinical $\mathbf{c}_r \in \mathbb{R}^C$ latent factors. Patient, gene and clinical latent factors that correspond to each rank-one tensor can be thought of as clusters of patients with frequently co-occurring somatic gene mutations and clinical variables. Latent factors for all rank-one tensors can be grouped into the columns of the patient $\mathbf{P}$, gene $\mathbf{G}$ and clinical $\mathbf{C}$ factor matrices. CP decomposition of τ is obtained by solving the following optimization problem:

$$\min_{\hat{\tau}} \|\tau - \hat{\tau}\|_{\mathcal{F}} \quad (2)$$

[b] source code is publicly available at *https://github.com/datad/CLIGEN*

aimed at finding the best approximation of each element t_{ijk} of the original tensor τ from the latent factors corresponding to rank-one tensors as follows:

$$t_{ijk} \approx \sum_{r=1}^{R} \lambda_r p_{ir} g_{jr} c_{kr} \qquad (3)$$

Uniqueness of the optimal solution to the above optimization problem is an important property of CP decomposition [13]. Molecular and clinical markers of disease sub-types are derived from the gene and clinical latent factors associated with each rank-one tensor obtained by CP decomposition of τ. Each element of a gene and clinical latent factor can be interpreted as a degree of specificity of a particular gene or a clinical variable to the corresponding disease sub-type. Each element of a patient latent factor can be interpreted as a membership proportion of a particular patient in the corresponding disease sub-type.

IV. Results

We performed both qualitative and quantitative evaluation of breast cancer sub-types identified by *CLIGEN* based on the genomic and clinical data of TCGA breast cancer patients.

A. *Quantitative evaluation*

Quantitative evaluation was conducted for the task of cancer patient survival prognosis, which is important for personalizing cancer treatment [14]. Specifically, we compared the Cox proportional hazards models that use the following predictors of patient survival: **M1)** patient membership proportions in *multi-modal sub-types* of breast cancer discovered by *CLIGEN*, which correspond to rows in the patient factor matrix **P**; **M2)** patient membership proportions in *molecular phenotypes* of breast cancer, which correspond to rows in the patient factor matrix **P** obtained through non-negative factorization of the somatic mutation matrix as $\mathbf{M} = \mathbf{PG}$; **M3)** patient membership proportions in *clinical phenotypes* of breast cancer, which correspond to rows in the patient factor matrix **P** obtained through non-negative factorization of the clinical matrix as $\mathbf{V} = \mathbf{PC}$; **M4)** random patient membership proportions in a given number of breast cancer sub-types. In the first experiment, we compared the accuracy of the Cox models using each of the above predictors for survival prognosis of breast cancer patients, while in the second experiment, we compared the goodness of fit of these models.

1) Accuracy of survival prognosis: In the first experiment, we compared the area under the ROC curve (AUC) for the models M1-M4 using randomized 10-fold cross validation. The Cox models were estimated using the data in the training splits and evaluated using the data in the testing splits. The plot of AUC values for models M1-M4 micro-averaged over splits by varying the number of the most prevalent cancer sub-types we well as molecular and clinical phenotypes is shown in Figure 2. Two major conclusions can be drawn from this figure.

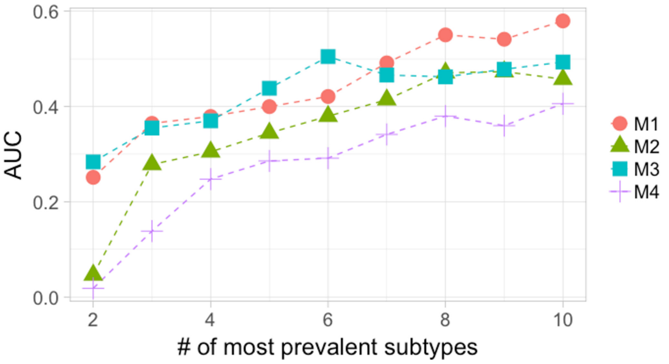


Fig. 2. AUC of Cox models for breast cancer patient survival time prediction that utilize patient membership proportions in most prevalent cancer sub-types discovered by CLIGEN (M1), molecular (M2) and clinical (M3) cancer phenotypes and random patient membership (M4).

First, the Cox regression model that utilizes patient membership proportions in cancer sub-types obtained by *CLIGEN* (M1) is consistently more accurate at predicting patient survival time than the Cox model that uses patient membership proportions in molecular (M2) and clinical (M3) phenotypes obtained through NMF, which indicates the importance of taking into account both clinical and genomic data when determining cancer sub-types. In particular, the Cox model utilizing patient membership proportions in multi-modal sub-types as predictors achieved the highest AUC of 0.5796, when 10 most prevalent sub-types were used, whereas the Cox model utilizing patient molecular phenotype membership proportions as predictors achieved the highest AUC of 0.4731, when 9 most prevalent phenotypes were used and the Cox model utilizing patient clinical phenotype membership proportions as predictors achieved the highest AUC of 0.5047, when 6 most prevalent phenotypes were used.

Second, the Cox models utilizing patient membership proportions in the top-*k* most prevalent sub-types derived by *CLIGEN* as well as molecular and clinical phenotypes derived by NMF are all more accurate at predicting patient survival time than the baseline Cox model utilizing random patient membership proportions in the same number of cancer sub-types (AUC = 0.4056).

2) Model goodness-of-fit: In the second experiment, we compared the goodness of fit of the models M1-M4 estimated on the entire TCGA dataset. The p-values of Log-rank and Wald tests of these models are summarized in Table I. Both tests indicate that patient membership proportions in the sub-

Model	**Log-rank test**	**Wald test**
M1	8.327e-15s	0.000082
M2	0.315	0.3772
M3	0.0007	0.0124
M4	0.5060	0.5509

TABLE I
P-values of Log-rank and Wald tests of M1, M2, M3 and M4 Cox proportional hazard models.

types derived by *CLIGEN* are more statistically significant predictors of breast cancer patient survival than membership proportions in breast cancer clinical phenotypes, which in turn are more statistically significant predictors than random patient membership proportions and membership proportions in molecular phenotypes. Kaplan-Meier survival plots for the 4 most prevalent breast cancer sub-types obtained by *CLIGEN* and NMF of mutation and clinical matrices are shown in Figure 3. As follows from Figure 3, breast cancer patient cohorts that correspond to the 4 most prevalent sub-types obtained using *CLIGEN* are more distinct in terms of survival dynamics ($p = 0.0493$) than patient cohorts that correspond to the 4 most prevalent molecular ($p = 0.241$) and clinical ($p = 0.2073$) phenotypes.

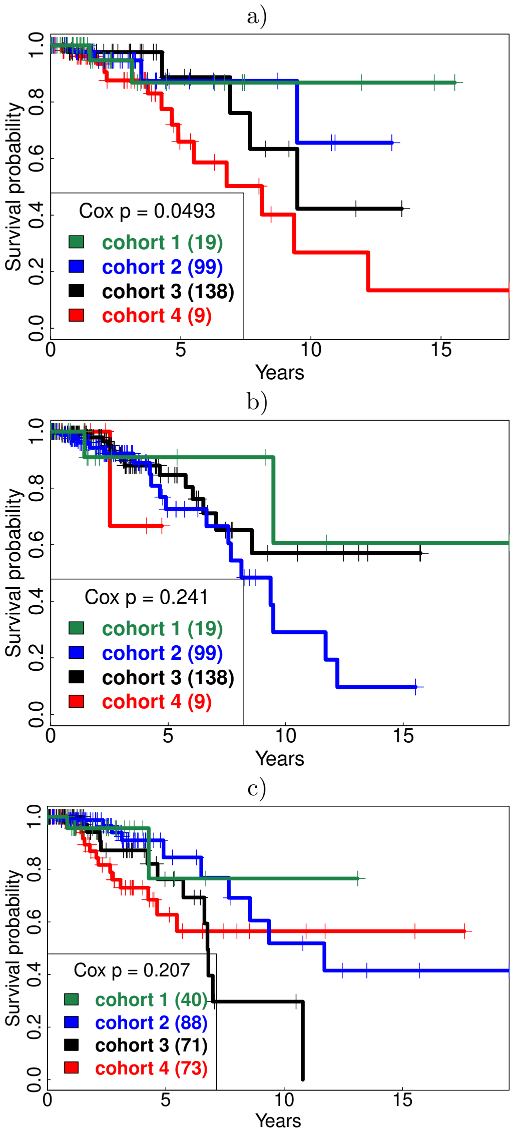


Fig. 3. Kaplan-Meier survival plots for the four most prevalent: a) sub-types obtained using *CLIGEN*, b) NMF-based molecular phenotypes, c) NMF-based clinical phenotypes.

B. *Qualitative evaluation*

An oncologist performed qualitative evaluation of breast cancer sub-types identified by *CLIGEN* through CP decomposition of the input tensor into 10 rank-one tensors, since this decomposition results in the most accurate prediction of cancer survival. An enrichment analysis of the list of genes associated with each subtype was performed using the Ingenuity Systems Upstream Analysis tool [15].

Each breast cancer sub-type obtained by *CLIGEN* was analyzed and compared with the known breast cancer subtypes obtained by clustering gene expression data [7]. This analysis revealed that one of the *CLIGEN* sub-types corresponds to a small cohort of patients with high mutation load. A detailed investigation of the molecular markers associated with this sub-type indicted that these markers correspond to mutations in the tumor suppressor genes (BRCA1, BRCA2, TP53, PTEN, RB1) that participate in DNA repair, which suggests that the high mutation load may be associated with a mutation in a DNA repair gene pathway(s). As follows from Figure 4, for each sample, these mutations were mutually exclusive. Further investigation of these genes can elucidate the biological process(es) underlying these mutations.

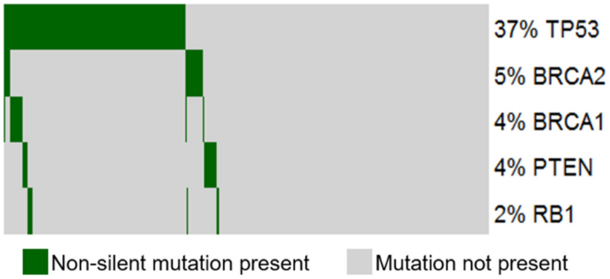


Fig. 4. Mutual exclusivity across the characteristic genes of the CLIGEN sub-type with high mutation load.

Another *CLIGEN* sub-type corresponds to a sub-type of the triple negative breast cancer (TNBC), which is a defined by the lack of ER, PR and HER2 expression. Further analysis of putative cancer driver genes, potentially activated or inactivated by the mutations associated with this TNBC-related *CLIGEN* sub-type resulted in significant enrichment of genes with a role in signaling networks that promote the function of cancer stem-like cells (CSCs), i.e., downstream of transcription factor TWIST1, and alternative mRNA splicing, i.e., downstream of serine and arginine-rich splicing factor SRSF2. CSCs are identified in patient TNBC tumors as a fraction of self-renewing, tumor-initiating cancer cells that also give rise to drug resistance and metastatic recurrence [16], [17]. Alternative mRNA splicing has also been implicated in maintaining and generating CSCs [18].

The other two *CLIGEN* sub-types of breast cancer refine progesterone receptor and estrogen receptor alpha-positive (ER+) breast cancer, which is responsive to anti-ER therapies, and the known sub-type of breast cancer, which is driven by over-expression of the epidermal growth factor receptor oncogene (HER2) and responsive to HER2-targeted inhibitors.

V. Discussion

Methods for integrative high-throughput analysis of genomic and clinical data face a common challenge of dealing with large volumes of data. By utilizing sparse representations and inexpensive linear algebra operations, tensor factorization methods effectively address this challenge. Successful application of tensor decomposition in different domains led to further research into efficient optimization methods for tensor decomposition [19], which makes tensor decomposition the method of choice for high-throughput cancer sub-typing.

Since tensor factorization methods are parametric, selecting the optimal number of rank-one tensor components for CP decomposition (i.e. model order estimation) is an important practical aspect of *CLIGEN*. Too few components typically result in general sub-type definitions, which may combine several actual disease sub-types. Too many components typically result in specific sub-type definitions, which may split the actual cancer sub-types. It is important to point out that, in terms of the number of model parameters, CP decomposition, which assumes that the number of components is the same per each tensor mode, has an advantage over Tucker tensor decomposition, which requires specifying the number of components per each mode. While it is known that the number of components that minimizes the reconstruction error of the original tensor from its components is equal to its rank [13], finding tensor rank is an NP-complete problem. Even if the rank of a tensor is known, the number of components that minimizes the reconstruction error may not result in the best accuracy for a particular task, such as patient survival time prognosis. Therefore, the optimal number of components is typically determined using heuristics, such as cross-validation [20] (as was done in this work) or hierarchical Bayesian approach [21], if a suitable prior can be defined.

VI. Conclusion

In this paper, we introduced *CLIGEN*, a novel computational pipeline for unsupervised sub-typing of complex diseases based on non-negative decomposition of a binary tensor combining clinical and somatic mutation patient data. Qualitative and quantitative evaluation of the sub-types discovered by *CLIGEN* for breast cancer indicates that representation of clinical and genomic patient data as a binary tensor and its subsequent non-negative decomposition is an efficient computational approach to high-throughput sub-typing of complex diseases for precision medicine. *CLIGEN* was not only able to refine the known breast cancer sub-types, but also elucidate new characteristics of a complex breast cancer sub-type (triple negative), which provides an opportunity for further research to define new cancer sub-types. We also demonstrated that patient membership proportions in breast cancer sub-types discovered by *CLIGEN* are more effective predictors of survival time than patient membership proportions in data-driven molecular and clinical phenotypes of breast cancer.

References

[1] Laura J Esserman, Ian M Thompson, and Brian Reid. Overdiagnosis and overtreatment in cancer: an opportunity for improvement. *The Journal of the American Medical Association*, 310(8):797–798, 2013.

[2] Archie Bleyer and H Gilbert Welch. Effect of three decades of screening mammography on breast-cancer incidence. *New England Journal of Medicine*, 367(21):1998–2005, 2012.

[3] Peter N. Robinson. Deep phenotyping for precision medicine. *Human Mutation*, 33(5):777–780, 2012.

[4] Alan S Coates, Eric P Winer, Aron Goldhirsch, Richard D Gelber, Michael Gnant, M Piccart-Gebhart, Beat Thürlimann, H-J Senn, Panel Members, Fabrice André, et al. Tailoring therapies improving the management of early breast cancer: St gallen international expert consensus on the primary therapy of early breast cancer 2015. *Annals of Oncology*, 26(8):1533–1546, 2015.

[5] Matan Hofree, John P Shen, Hannah Carter, Andrew Gross, and Trey Ideker. Network-based stratification of tumor mutations. *Nature Methods*, 10(11):1108–1115, 2013.

[6] Marine Le Morvan, Andrei Zinovyev, and Jean-Philippe Vert. NetNorM: Capturing cancer-relevant information in somatic exome mutation data with gene networks for cancer stratification and prognosis. *PLoS Computational Biology*, 13(6):e1005573, 2017.

[7] Charles M Perou, Therese Sørlie, Michael B Eisen, Matt Van De Rijn, Stefanie S Jeffrey, Christian A Rees, Jonathan R Pollack, Douglas T Ross, Hilde Johnsen, Lars A Akslen, et al. Molecular portraits of human breast tumours. *Nature*, 406(6797):747–752, 2000.

[8] Diana Diaz, Michele Donato, Tin Nguyen, and Sorin Draghici. MicroRNA-augmented pathways (mirAP) and their applications to pathway analysis and disease subtyping. In *Pacific Symposium on Biocomputing*. World Scientific, 390–401, 2017.

[9] Diana Diaz, Tin Nguyen, and Sorin Draghici. A systems biology approach for unsupervised clustering of high-dimensional data. In *Machine Learning, Optimization, and Big Data*, L N in Computer Science. Springer, Cham, 193–203, 2016.

[10] Aliccia Bollig-Fischer. How the future of clinical cancer diagnostics can contribute to overcoming race-associated cancer disparities. *Expert Review of Molecular Diagnostics*, 16(12):1233–1235, 2016.

[11] Cancer Genome Atlas Research Network and et al. Comprehensive molecular portraits of human breast tumours. *Nature*, 1:61—-70, 2012.

[12] TCGA Cancer Genome Atlas Research Network and et al. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113—-1120, 2013.

[13] Tamara G. Kolda and Brett W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009.

[14] S Saria and A Goldenberg. Subtyping: What it is and its role in precision medicine. *IEEE Intelligent Systems*, 30(4):70–75, 2015.

[15] Andreas Kramer, Jeff Green, Jack Pollard, and Stuart Tugendreich. Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics*, 30(4):523–530, 2014.

[16] Bhuvanesh Dave, Vivek Mittal, Nicholas M Tan, and Jenny C Chang. Epithelial-mesenchymal transition, cancer stem cells and treatment resistance. *Breast Cancer Research*, 14(202), 2012.

[17] Huiping Liu, Manishkumar R Patel, Jennifer A Prescher, and et al. Cancer stem cells from human breast tumors are involved in spontaneous metastases in orthotopic mouse models. *Proceedings of the National Academy of Sciences of USA*, 107(42):18115–18120, 2010.

[18] Bin Bao, Cristina Mitrea, Priyanga Wijesinghe, and et al. Treating triple negative breast cancer cells with erlotinib plus a select antioxidant overcomes drug resistance by targeting cancer cell heterogeneity. *Scientific Reports*, 7(44125), 2017.

[19] Evrim Acar, Daniel M. Dunlavy, and Tamara G. Kolda. A scalable optimization approach for fitting canonical tensor decompositions. *Journal of Chemometrics*, 25(2):67–86, 2011.

[20] R. Bro, K. Kjeldahl, A. K. Smilde, and H. Kiers. Cross-validation of component models: A critical look at current methods. *Analytical and Bioanalytical Chemistry*, 390(5):1241–1251, 2008.

[21] Morten Mørup and Lars Kai Hansen. Automatic relevance determination for multi-way models. *Journal of Chemometrics*, 23(7–8):352–363, 2009.