Mathematical modelling of molecular heterogeneity identifies novel markers and subpopulations in complex tumors

Lulu Chen, Niya Wang, Robert Clarke, Zhen Zhang, Yue Wang

Abstract— Intratumor heterogeneity, as both a major confounding factor and an underexploited information source, is widely implicated as a key driver of drug resistance. While a handful of reports have demonstrated the potential of supervised methods to deconvolute intratumor heterogeneity, these approaches require a priori information on the marker genes or composition of known subpopulations. To address the critical problem of the absence of validated marker genes for many (including novel) subpopulations, we developed convex analysis of mixtures (CAM), a fully unsupervised deconvolution method, for identifying marker genes and subpopulations directly from original mixed molecular expressions.

I. INTRODUCTION

Drug resistance is a major barrier to increasing overall survival rates, whether this resistance is present at the time of presentation or emerges after treatment response. Multiscale intratumor heterogeneity (ITH) is characteristic of complex tumors and is widely implicated as a key driver of drug resistance. Precisely how ITH develops within and across scales is unclear.

ITH, arising from multiple subpopulations within a tumor, is both a major confounding factor in studying individual subpopulations and an underexploited information source for characterizing complex tumors [1]. ITH is multiscale, i.e., evident at many levels (e.g., genetic, transcriptomic, proteomic, metabolic, cellular, phenotypic). Complex tumors can be characterized by the identity, composition, and expression profile of possibly unknown subpopulations [2], where subpopulations are often defined by marker genes (genes whose expressions are exclusively enriched in a particular subpopulation [3, 4]). While many reports have demonstrated the potential of supervised methods to deconvolute ITH, these approaches almost exclusively require a priori information on the known subpopulations [2, 3]. Such supervised methods (including cell sorting) have difficulty dissecting ITH that are subtle, condition-specific or previously unknown [1, 5].

To address the critical problem of the absence of validated marker genes for many (including novel) subpopulations, we recently developed a fully unsupervised computational method (convex analysis of mixtures – CAM) [6, 7] that can identify subpopulation-specific marker genes directly from the original mixed expressions - a nontrivial task. CAM requires no prior information on the number, identity, or composition

of the subpopulations present in mixed samples [8], and does not require the presence of pure subpopulations in sample space [9, 10]. CAM dissects heterogeneity into molecularly distinct subpopulations by leveraging the advantages of both tissue-wide and single-cell approaches [10, 11]. Discerning differences among single cells can provide valuable information about inter-cellular heterogeneity but may lose critical information of cell-cell interactions and is prone to cell-cycle/state confounders; while sample-wide measures give a detailed picture of the average state of a cell population but at the cost of losing information about inter-subpopulation heterogeneity.

II. RESULTS

A. Validation of CAM on real dataset

CAM's ability for unsupervised data deconvolution has been validated on several real benchmark gene expression datasets (GSE19830, GSE11058) [6]. More recently, CAM analysis of proteomic data from left anterior descending coronary artery (LAD) and abdominal aorta (AA) specimens from a large cohort identified 58 informative proteins of normal (NL), fatty streak (FS), and fibrous plaque (FP) burden levels with 48 proteins verified in an independent validation cohort [12].

B. Application of CAM to model ITH in breast cancer

To address ITH remodeling in the context of endocrine resistance, the MCF-7 variants LCC1 cells (estrogen independent; antiestrogen sensitive) and LCC9 cells (estrogen independent; antiestrogen resistant) are mixed in different ratios, and treated with drug ICI or vehicle. Hierarchical clustering of their iTRAC proteomic data shows that the molecular profiles of 1:1 and 5:1 (LCC1:LCC9) ICI-treated and 1:1 untreated mixtures cluster with LCC9 (Fig. 3a). Untreated 5:1 mixtures are close to untreated LCC1. Data from the mixtures were deconvoluted in silico (unsupervised) by CAM into two subclone signals (C1, C2; Fig. 3b). When combined with data from pure cells, PCA shows that C1 signals from untreated 1:1 and 5:1 are close to LCC1, and the C2 signals from ICI treated and untreated mixtures are close to LCC9. C1 signals from the treated 1:1 and 5:1 mixtures form a separate cluster distant from LCC1 and LCC9 cells, implying that resistant cells and sensitive cells alter each other's molecular signatures. These reflect the importance studies into the in vivo tumor microenvironment.

Research supported by the National Institutes of Health under Grants CA160036-05S1, CA184902, and HL111362.

Lulu Chen, Niya Wang, Yue Wang are with the Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Arlington, VA 22203, USA (e-mail: luluchen@vt.edu, wangny@vt.edu, yuewang@vt.edu).

Robert Clarke is with Lombardi Comprehensive Cancer Center, Georgetown University, Washington, DC 20057, USA (e-mail: clarker@georgetown.edu).

Zhen Zhang is with the Department of Pathology, Johns Hopkins University, Baltimore, MD 21231, USA (e-mail: zzhang7@jhmi.edu).

- [2] S. S. Shen-Orr, R. Tibshirani, P. Khatri, D. L. Bodian, F. Staedtler, N. M. Perry, *et al.*, "Cell type-specific gene expression differences in complex tissues," *Nat Methods*, vol. 7, pp. 287-9, Apr 2010.
- [3] A. Kuhn, D. Thu, H. J. Waldvogel, R. L. Faull, and R. Luthi-Carter, "Population-specific expression analysis (PSEA) reveals molecular changes in diseased brain," *Nat Methods*, vol. 8, pp. 945-7, Nov 2011.
- [4] G. Yu, H. Li, S. Ha, I. M. Shih, R. Clarke, E. P. Hoffman, *et al.*, "PUGSVM: a caBIGTM analytical tool for multiclass gene selection and predictive classification," *Bioinformatics*, vol. 27, pp. 736-738, 2011.
- [5] R. Gaujoux and C. Seoighe, "Semi-supervised Nonnegative Matrix Factorization for gene expression deconvolution: A case study," *Infection, Genetics and Evolution*, vol. 12, pp. 913-921, 7// 2012.
- [6] N. Wang, E. P. Hoffman, L. Chen, L. Chen, Z. Zhang, C. Liu, *et al.*, "Mathematical modelling of transcriptional heterogeneity identifies novel markers and subpopulations in complex tissues," *Sci Rep*, vol. 6, p. 18909, Jan 07 2016.
- [7] N. Wang, F. Meng, L. Chen, S. Madhavan, R. Clarke, E. P. Hoffman, *et al.*, "The CAM software for nonnegative blind source separation in R-Java," *J. Mach. Learn. Res.*, vol. 14, pp. 2899-2903, 2013.
- [8] N. S. Zuckerman, Y. Noam, A. J. Goldsmith, and P. P. Lee, "A self-directed method for cell-type identification and separation of gene expression microarrays," *PLoS Comput Biol*, vol. 9, p. e1003189, 2013.
- [9] Y. Hart, H. Sheftel, J. Hausser, P. Szekely, N. B. Ben-Moshe, Y. Korem, *et al.*, "Inferring biological tasks using Pareto analysis of high-dimensional data," *Nat Meth*, vol. 12, pp. 233-235, 03//print 2015.
- [10] R. Schwartz and S. Shackney, "Applying unmixing to gene expression data for tumor phylogeny inference," *BMC Bioinformatics*, vol. 11, p. 42, 2010.
- [11] F. Buettner, K. N. Natarajan, F. P. Casale, V. Proserpio, A. Scialdone, F. J. Theis, *et al.*, "Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells," *Nat Biotech*, vol. 33, pp. 155-160, 02//print 2015.
- [12] D. Herrington, C. Mao, S. Parker, Z. Fu, G. Yu, L. Chen, *et al.*, "Proteomic Architecture of Human Coronary and Aortic Atherosclerosis," *bioRxiv*, 2017.
- [13] Y. Zhong and Z. Liu, "Gene expression deconvolution in linear space," *Nat Methods*, vol. 9, pp. 8-9; author reply 9, Jan 2012.
- [14] B. J. Frey and D. Dueck, "Clustering by Passing Messages Between Data Points," *Science*, vol. 315, pp. 972-976, 2007.
- [15] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, pp. 387-392, 1985.