Bioinformatics, 33(18), 2017, 2951–2953
doi: 10.1093/bioinformatics/btx297
Advance Access Publication Date: 4 May 2017
Applications Note

OXFORD

Gene expression

SinCHet: a MATLAB toolbox for single cell heterogeneity analysis in cancer

Jiannong Li[1], Inna Smalley[2], Michael J. Schell[1], Keiran S. M. Smalley[2] and Y. Ann Chen[1,*]

[1]Department of Biostatistics and Bioinformatics and [2]Departments of Tumor Biology and Cutaneous Oncology, H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL 33612, USA

*To whom correspondence should be addressed.

Associate Editor: Bonnie Berger

Received on January 4, 2017; revised on April 11, 2017; editorial decision on April 28, 2017; accepted on May 3, 2017

Abstract

Summary: Single-cell technologies allow characterization of transcriptomes and epigenomes for individual cells under different conditions and provide unprecedented resolution for researchers to investigate cellular heterogeneity in cancer. The SinCHet (**Sin**gle **C**ell **Het**erogeneity) toolbox is developed in MATLAB and has a graphical user interface (GUI) for visualization and user interaction. It analyzes both continuous (e.g. mRNA expression) and binary omics data (e.g. discretized methylation data). The toolbox does not only quantify cellular heterogeneity using Shannon Profile (SP) at different clonal resolutions but also detects heterogeneity differences using a D statistic between two populations. It is defined as the area under the Profile of Shannon Difference (PSD). This flexible tool provides a default clonal resolution using the change point of PSD detected by multivariate adaptive regression splines model; it also allows user-defined clonal resolutions for further investigation. This tool provides insights into emerging or disappearing clones between conditions, and enables the prioritization of biomarkers for follow-up experiments based on heterogeneity or marker differences between and/or within cell populations.

Availability and implementation: The SinCHet software is freely available for non-profit academic use. The source code, example datasets, and the compiled package are available at http://labpages2.moffitt.org/chen/software/.

Contact: ann.chen@moffitt.org

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Tumor heterogeneity between and within tumors plays a critical role in tumor aggression and the development of drug resistance. Understanding and characterizing clonal heterogeneity enables us to gain insights into the progression of cancer and guide the effective therapeutic strategies (Marusyk and Polyak, 2010). New high-throughput single-cell technologies provide unprecedented resolution for researchers to explore cellular heterogeneity in cancer (Tirosh *et al.*, 2016). However, these technologies pose new challenges in data analysis and interpretation. Currently, there are several single cell analysis tools available such as SCATT, TSCAN, SPADE, vi-SNE (Amir el *et al.*, 2013; Anchang *et al.*, 2016; Ji and Ji, 2016; Mitra *et al.*, 2016). Each tool has its own strengths and limitations (Supplementary Table S1). There is only a limited number of tools available for quantifying cellular heterogeneity and comparing heterogeneities quantitatively between populations, and identifying markers based on heterogeneity. Therefore, we have developed the SinCHet toolbox, in MATLAB with a GUI for visualization and user interaction, originally for cancer research but with the potential to be used for any single cell research. The toolbox has four parts (Supplementary Fig. S1): (1) imports continuous or categorical omics data and output the figures and results for review; (2) performs exploratory analyses including hierarchical cluster analyses and principal components analyses (PCA); (3) Estimates the clonal

heterogeneity using Shannon Profile (SP), provides a Profile of Shannon Differences (PSD) to characterize the heterogeneity differences between populations and a novel D statistic to quantitatively compare heterogeneities between populations; (4) Prioritizes biomarkers based on between and/or within group cellular heterogeneity.

2 Materials and methods

2.1 Clonal richness and heterogeneity estimation

We assume proper normalization has been performed prior to using the tool. Although data normalization is beyond the scope of this application, some considerations for data pre-processing and normalization are discussed in the supplemental materials. Currently, the tool is developed in a two-group comparison setting.

Hierarchical cluster analyses with different linkage methods were performed to cluster cells into phenotypic clonal groups, referred to as clones in this application note, based on the similarities of the input dataset (Supplementary Figs. S2A and S3A). Cophenetic correlation coefficient (Sokal, 1962) is used to choose the default linkage method. PCA analyses are available to visualize the relationships and patterns of the samples (Supplementary Figs. S2B and S3B). Clonal richness, i.e. counts of clones and Shannon index (Hernandez-Walls and Trujillo-Ortiz, 2010; Southwood and Henderson, 2000) are used to quantify clonal diversity and heterogeneity (Supplementary equation S1).

SinCHet provides Shannon Profile (SP) under each condition by evaluating the heterogeneity using Shannon index at different heights along the dendrogram (Supplementary Figs. S2C and S3C). PSD, the profile of the differences of Shannon index calculated along the same X-axis as SP, is used characterize the heterogeneity differences between two conditions (or populations; Supplementary Figs. S2D and S3D). A novel D statistic is then defined as the area under the PSD, or equivalently, the differences of the areas under the SPs between two groups (Supplementary equation S2). We have shown that this D statistic is empirically robust to choice of different linkage methods for hierarchical cluster analyses (the Supplementary Material results). Permutation is used to evaluate its statistical significance (Supplementary equation S3). To identify the number of existing clones under each condition, Multivariate adaptive regression splines (MARS) model (Friedman, 1991; Jekabsons, 2016) is used to detect the change points in PSD (Supplementary Figs. S2D and S3D). The higher the clone numbers, the fewer the cells there are in each clone, which will reduce the statistical power for comparison. So, minimum of change points determined by MARS is chosen as the default to provide the clonal snap shot (Supplementary Figs. S2E–F and S3E–F) which provides the information on clonal compositions and biomarker analyses (Supplementary Figs. S2G–H and S3G–H). The SinCHet toolbox also allows the user-defined number of clones along the profile for exploration accordingly.

2.2 Biomarkers prioritization

The within- and between- population comparisons are performed and results are all saved for further investigation (Supplementary Tables S2 and S3). Each comparison could have its own biological significance and results of the top-ranked markers could be visualized individually (Supplementary Fig. S1 and Supplementary Figs. S2H and S3H). Given the large amount of information generated by each single cell experiment, a composite score, Generalized Fisher Product Score (GF), is devised to summarize the overall difference between- and within-population comparisons and to prioritize biomarkers for further investigation.

For categorical data, GF is aggregation of evidence from three separate Fisher's exact tests for each biomarker (e.g. methylation site).

$$X_i^2 = -2\sum_j^3 \ln(p_{ij}) \quad (1)$$

where p_{ij} is the P value from Fisher's exact test for the ith biomarker and jth comparison (when $j = 1$, it is the dominant clone comparison between groups and $j = 2$ or 3, the tests are comparisons between dominant clone and the remaining minor clones within each population).

When the biomarker is a continuous variable, three rank sum tests are performed to compare the difference of the expression levels. Markers with large differences are often desired by the researchers for validation experiments, therefore, fold change (FC) for each of the three comparisons are also incorporated in the GF score:

$$GF_i = X_i^2 + \sum_j^3 \left|ln\left(FC_{ij}\right)\right| \quad (2)$$

FC_{ij}, is the FC for biomarker i at jth comparison as described above.

3 Results

We applied the SinCHet toolbox to published single-cell expression and methylation datasets (Cheow *et al.*, 2016). Data processing procedures were summarized in the supplement. For the gene expression dataset, the toolbox identified that the heterogeneity is higher in the *EGFR*-mutant lung cancer tumors than the wild type group ($D = -63.8$, $P < 0.001$, Supplementary Fig. S2C). This was supported by a previous report (Bai *et al.*, 2013). Nine clones were identified by SinCHet using the default setting (Supplementary Fig. S2E and F). The dominant clone from each group identified by SinCHet, i.e. Clone 1 from the the wild type tumors and Clone 2 from the mutant tumors, were in general agreement with the two clusters identified in the original paper. Additional clonal heterogeneity was characterized by SinCHet, with 7 additional clones identified (Supplementary Results). Furthermore, SinCHet was not only able to identify the same reported top genes (e.g. MUC1, SFTPC and KRT7; which differed significantly between *EGFR*-mutant and wild-type tumors) using the GF score but also was able to identify novel markers such as CD44, MT2A within each subpopulation (Supplementary Table S2). For the methylation dataset, SinCHet enabled the identification of the significantly hypermethylated loci HOXA9, PROM1 and PAX3 as shown by the paper reported in *EGFR*-mutant cells. In addition, the SinCHet top-ranked hypermethylated loci PAX5, SOX9 and SPINT1 found in subpopulations within *EGFR*-mutant cells might infer that some of the subpopulations could acquire stochastic epigenetic aberrations during tumor evolution as discussed in the original paper (Supplementary Table S3).

SinCHet can quantify cellular heterogeneity and identify novel candidate biomarkers, considering heterogeneity both between- and/or within groups. It provides unique insights into emerging or disappearing clones at different clonal resolutions between cell populations in different contexts. It could be easily applied to compare heterogeneity between groups with versus without mutations or before versus after acquired drug resistance. It could be also applied to

quantify heterogeneity during the course of cancer treatment, potentially changing the face of cancer therapeutic strategies in the future.

Acknowledgement

We thank Ms. Paula Price for providing exemplary administrative support.

Funding

This work was supported by Moffitt Skin SPORE grant (P50CA168536), R21CA198550, R21CA216756 and Cancer Center Support Grant P30CA076292 to the H. Lee Moffitt Comprehensive Cancer Center and Research Institute.

Conflict of Interest: none declared.

References

Amir el,A.D. *et al*. (2013) viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat. Biotechnol.*, **31**, 545–552.

Anchang,B. *et al*. (2016) Visualization and cellular hierarchy inference of single-cell data using SPADE. *Nat. Protocols*, **11**, 1264–1279.

Bai,H. *et al*. (2013) Detection and clinical significance of intratumoral EGFR mutational heterogeneity in Chinese patients with advanced non-small cell lung cancer. *PloS One*, **8**, e54170.

Cheow,L.F. *et al*. (2016) Single-cell multimodal profiling reveals cellular epigenetic heterogeneity. *Nat. Methods*, **13**, 833–836.

Friedman,J.H. (1991) Multivariate adaptive regression splines. *Ann. Statis.*, Institute of Mathematical Statistics, p 1–67.

Hernandez-Walls,R. and Trujillo-Ortiz,A. (2010) index_SaW:Shannon-Wiener Index. A MATLAB file.

Jekabsons,G. (2016) ARESLab: Adaptive Regression Splines toolbox for Matlab/Octave.

Ji,Z. and Ji,H. (2016) TSCAN: pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Res.*, **44**, e117.

Marusyk,A. and Polyak,K. (2010) Tumor heterogeneity: causes and consequences. *Biochim. Et Biophys. Acta*, **1805**, 105–117.

Mitra,A.K. *et al*. (2016) Single-cell analysis of targeted transcriptome predicts drug sensitivity of single cells within human myeloma tumors. *Leukemia*, **30**, 1094–1102.

Sokal,R.R. and Rohlf,F.J. (1962) The comparison of dendrograms by objective methods. *Taxon*, **11**, 33–40.

Southwood,T.R.E. and Henderson,P.A. (2000) *Ecological Methods*. Blackwell Science, Malden, MA.

Tirosh,I. *et al*. (2016) Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science*, **352**, 189–196.