A novel Bayesian model for assessing intratumor heterogeneity of tumor infiltrating leukocytes with multi-region gene expression sequencing

Peng Yang [1,2] Shawna M. Hubert [3] P. Andrew Futreal [4] Xingzhi Song [4] Jianhua Zhang [4] J. Jack Lee [2] Ignacio Wistuba [5] Jianjun Zhang [3,4] Ying Yuan [2] Ziyi Li [2]

[1]Department of Statistics, Rice University [2]Department of Biostatistics [3]Thoracic Head Neck Medical Oncology [4]Genomic Medicine [5]Translational Molecular Pathology, The University of Texas MD Anderson Cancer Center

Introduction

Intratumor heterogeneity (ITH) of tumor-infiltrated leukocytes (TILs) is an important phenomenon of cancer biology with potentially profound clinical impacts. **Multi-region gene expression data** provide a promising opportunity that allows for explorations of TILs and their ITH for each subject described as follows:

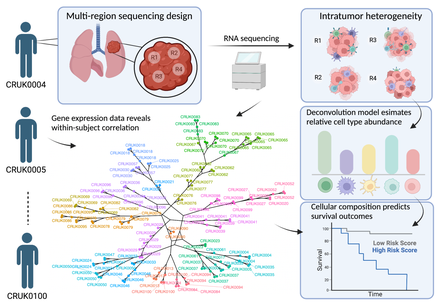

Figure 1. Multi-region sequencing study design.

ICeITH model is designed to address the challenges of assessing ITH using multi-region RNA-seq data, as they can **reveal differences in gene expression and immune cell infiltration between different regions of the same tumor.**

Method

Considering a total of K immune cell types are of interest here, the observed gene expression Y_{sg} can be decomposed as:

$$log(Y_{sg}) = \sum_k h_{sk} W_{sgk} + \epsilon_{sg}, \text{ for } s \in I_i, \tag{1}$$

where h_{sk} is the unobserved cellular abundance from cell type k in sample s, and W_{sgk} is a three-dimensional tensor that stands for the hidden expression profiles of gene g in sample s from cell type k. ϵ_{sg} is the error term that follows a normal distribution.

To characterize intratumor heterogeneity within the same patient subject, a **hierarchical Bayesian approach** is taken in two steps. First, each patient is allowed to have their own pure cell type profile parameters μ_{igk} per gene and per cell type. This is expressed as:

$$\log(W_{sgk}) \overset{\text{i.i.d}}{\sim} N(\mu_{igk}, \frac{1}{\lambda_{gk}}), \text{ for } s \in I_i.$$

Second, we use Dirichlet distribution to model cell-type-specific parameter, h_{sk}. For sample $s \in I_i$, we have

$$h_{s1}, \ldots, h_{sK} \sim \text{Dir}(C_i \boldsymbol{\pi}), \tag{2}$$

where (1) $\boldsymbol{\pi}$ is a K by 1 vector pooled across all samples with $\sum_k \pi_k = 1$ to represent the global cellular composition, and (2) C_i is a patient-specific parameter that controls the variability of the cellular composition across samples within each patient, thereby revealing ITH.

The Fenton-Wilkinson (FW) approximation [1] is used to approximate equation (1) by another log-normal distribution as follows:

$$\log(Y_{sg}) = N(\log(\sum_k h_{sk} W_{sgk}), \frac{1}{\lambda_{sg}}), \text{ for } s \in I_i,\ i = 1, \ldots, n.$$

Prior specifications

To incorporate prior knowledge from existing reference profiles, we use conjugate prior distributions for the patient-specific mean expression parameter μ_{igk}'s and variability λ_{gk}:

$$\mu_{igk} \overset{\text{i.i.d}}{\sim} N(\mu_{gk}, \frac{1}{\rho_{gk}\lambda_{gk}}), \quad \lambda_{gk} \sim \text{Gamma}(\alpha_{gk}, \beta_{gk}),$$

where μ_{gk}, $\alpha_{g,k}$ and $\beta_{g,k}$ are determined by the mean and variance from the reference matrix.

Optimization

We use the **Collapsed Variational Bayesian (CVB)** method to optimize the ICeITH model. To perform a CVB method, we first marginalize over the hidden random variables μ_{igk}'s and λ_{gk}'s,

$$\prod_{i=1}^{N} \prod_{s \in I_i} P(\log(W_{sgk}) | \mu_{gk}, \rho_{gk}, \alpha_{gk}, \beta_{gk})$$
$$= \int_\lambda \int_\mu \prod_{i=1}^{N} \prod_{s \in I_i} p(\log(W_{sgk}) | \mu_{igk}, \lambda_{gk}) \times p(\mu_{igk} | \mu_{gk}, \lambda_{gk}, \rho_{gk}) \times p(\lambda_{gk} | \alpha_{gk}, \beta_{gk})\, d\mu_{1gk} \cdots d\mu_{Ngk}\, d\lambda_{gk}$$

After we integrate out the latent variables, we introduce the following variational distributions on the remaining unobserved variables, that is W_{sgk} and h_{sk},

$$Q(\log(W_{sgk})) \sim N(\gamma_{igk}, \tau^2_{gk}), \quad Q(h_{s1}, \ldots, h_{sK}) \sim \text{Dir}(\xi_{s1}, \ldots, \xi_{sK}),$$

where $Q(\cdot)$ denotes the variational distribution aims to approximate the true posterior density and $\{\gamma_{igk}\}_{i,g,k}$, $\{\tau_{gk}\}_{g,k}$, and $\{\xi_{sk}\}_{s,k}$ are variational parameters that seek to be optimized. In total, there are $(N \times G \times K) + (G \times K) + (\sum_{i=1}^{N} I_i \times K)$ parameters to estimate.

The final **objective function** is based on the evidence lower bound, derived as follows:

$$\begin{aligned} \log(P(Y)) &\geq \underbrace{E_{Q(W,H)}\{\log \frac{P(Y, W, H|\boldsymbol{\theta})}{Q(W, H)}\}}_{\text{ELBO}} \\ &\geq \underbrace{E_Q\{\log P(Y|W, H, \lambda)\}}_{\text{a}} + \underbrace{E_Q\{\log P(W|\mu, \rho, \alpha, \beta)\}}_{\text{b}} + \underbrace{E_Q\{\log P(H|C, \pi)\}}_{\text{c}} \\ &\quad - \underbrace{E_Q\{\log P(W|\gamma, \tau)\}}_{\text{d}} - \underbrace{E_Q\{\log P(H|\xi)\}}_{\text{e}}, \end{aligned} \tag{3}$$

where $Z = (W, H)$ and $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\rho}, \boldsymbol{\mu}, \boldsymbol{\lambda})$ denote the unobserved variables and hyperparameters, respectively. We apply Limited-memory BFGS (Broyden–Fletcher–Goldfarb–Shanno) to iteratively maximize the objective function defined in equation 3 and its gradient with respect to variational parameters has been derived to speed the optimization.

To obtain the **intratumor heterogeneity parameter** C_i in equation 2 for a specific subject i, we compute the first and second central moment for $s \in I_i$ as follows,

$$E[h_{s,k}] = \frac{C_i \hat{\pi}_k}{C_i \sum_c \hat{\pi}_c} = \hat{\pi}_k, \quad Var[h_{s,k}] = \frac{\hat{\pi}_k(1 - \hat{\pi}_k)}{C_i + 1}. \tag{4}$$

Then, the total variance across samples within subject i is $\sum_k Var[h_{s,k}] = \sum_k \frac{\hat{\pi}_k(1-\hat{\pi}_k)}{C_i+1}$, which the ITH score can be calculated as follow:

$$\hat{C}_i = f_C(\hat{\pi}_k, \hat{h}_{sk}) = \frac{\sum_k \hat{\pi}_k(1 - \hat{\pi}_k)}{\sum_k var(h_{sk})} - 1. \tag{5}$$

This equation estimates the degree of ITH within each patient. In particular, the numerator in equation 5 represents the expected variance of the cell type proportions across all samples, while the denominator represents the actual variance within each subject.

Simulation study

We conduct extensive simulation studies to evaluate the performance and robustness of our proposed method, ICeITH, as well as benchmark it against CIBERSORT [3] and EPIC [4].

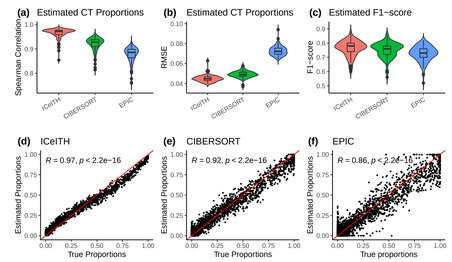

Figure 2. Results of analyzing the simulation data in the setting with four cell types and randomly generated sample numbers (3-5) per patient.

The results show the ICeITH provides the most accurate estimation of cell-type-specific proportions and a better ability to dichotomize low vs high intratumor heterogeneity groups.

Real data application on TRACERx

We apply the ICeITH model to analyze RNA-seq data in patients from the TRACERx cohort [2]. The multi-region RNA-seq data are available in 45 patients, and it results in 140 tumors in total. We seek to associate the survival outcome with intratumor heterogeneity scores estimated from our proposed method.

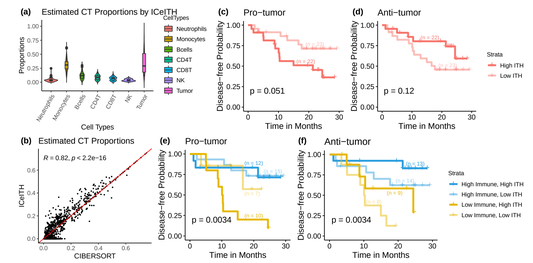

Figure 3. Results of analyzing the TRACERx data.

ICeITH is capable of classifying patients into different risk groups according to the ITH estimation of targeted TILs that shape the either pro- or anti-tumor processes.

References

[1] Lawrence Fenton. The sum of log-normal probability distributions in scatter transmission systems. *IRE Transactions on communications systems*, 8(1):57–67, 1960.

[2] Mariam Jamal-Hanjani, Gareth A Wilson, Nicholas McGranahan, Nicolai J Birkbak, Thomas BK Watkins, Selvaraju Veeriah, Seema Shafi, Diana H Johnson, Richard Mitter, Rachel Rosenthal, et al. Tracking the evolution of non–small-cell lung cancer. *New England Journal of Medicine*, 376(22):2109–2121, 2017.

[3] Aaron M Newman, Chih Long Liu, Michael R Green, Andrew J Gentles, Weiguo Feng, Yue Xu, Chuong D Hoang, Maximilian Diehn, and Ash A Alizadeh. Robust enumeration of cell subsets from tissue expression profiles. *Nature methods*, 12(5):453–457, 2015.

[4] Julien Racle and David Gfeller. Epic: a tool to estimate the proportions of different cell types from bulk gene expression data. In *Bioinformatics for Cancer Immunotherapy*, pages 233–248. Springer, 2020.