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

Peng Yang

Rice University

Aug 10, 2023

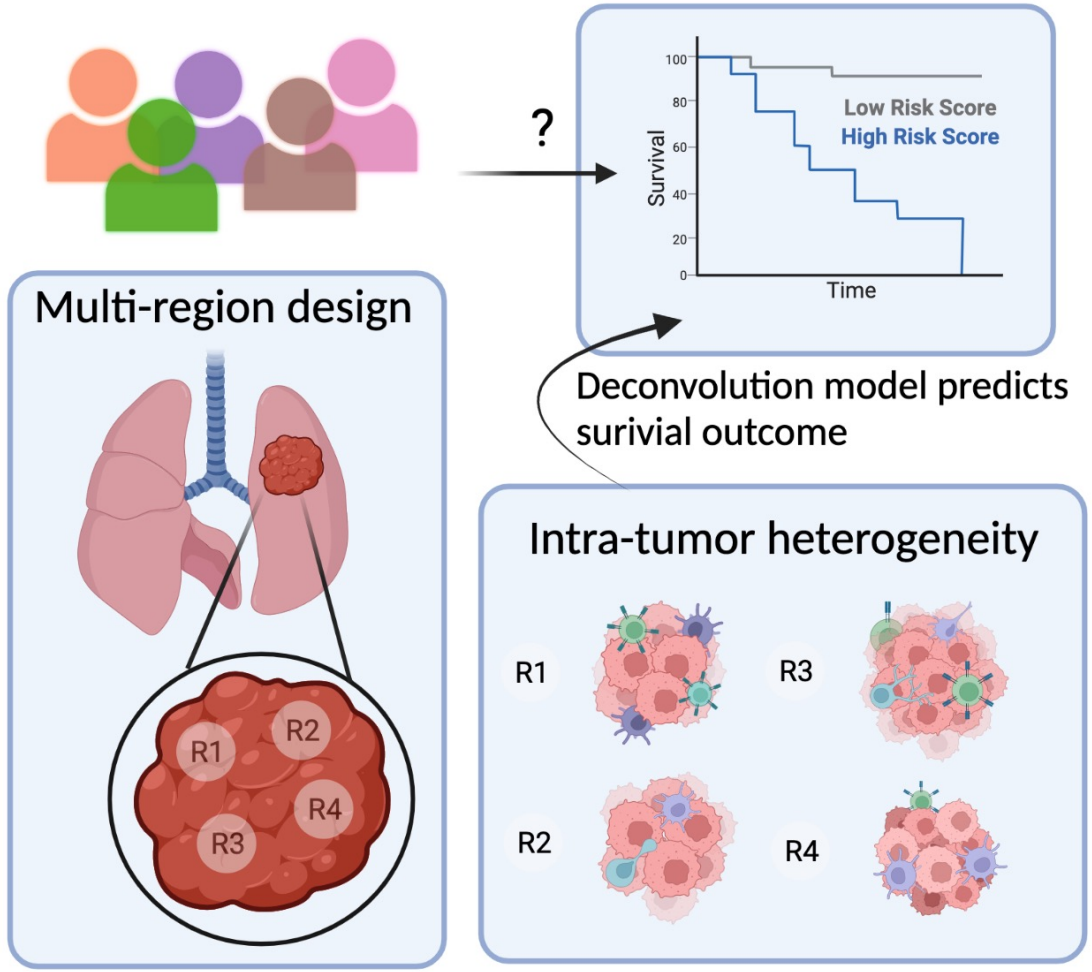
Contents

- Background
  - Background knowledge
  - Motivation
- Statistical method
  - Model development
  - Optimization
- Simulation study
  - Settings
  - Results interpretations
- Real data application
  - TRACERx study
- Conclusion

Background

Intratumor heterogeneity (ITH)

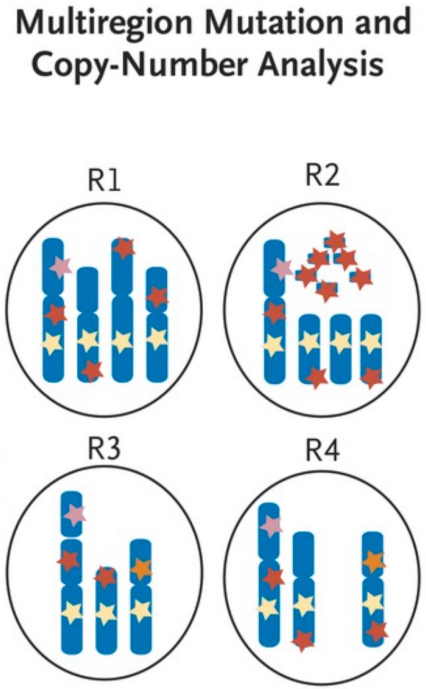
- ITH of tumor-infiltrated leukocytes (TILs) is an important phenomenon of cancer biology with potentially profound clinical impacts.
- Multi-region sequencing data provides a promising opportunity that allows the exploration of ITH, i.e.,
  - TRACERx study revealed the association of SCNA to patient prognosis (Jamal-Hanjani et al. 2017);
  - AbdulJabbar et al. 2020 studied differentiate highly immune-infiltrated tumor regions by the number of immune hot and cold tumors at a population level reveal patient survival.

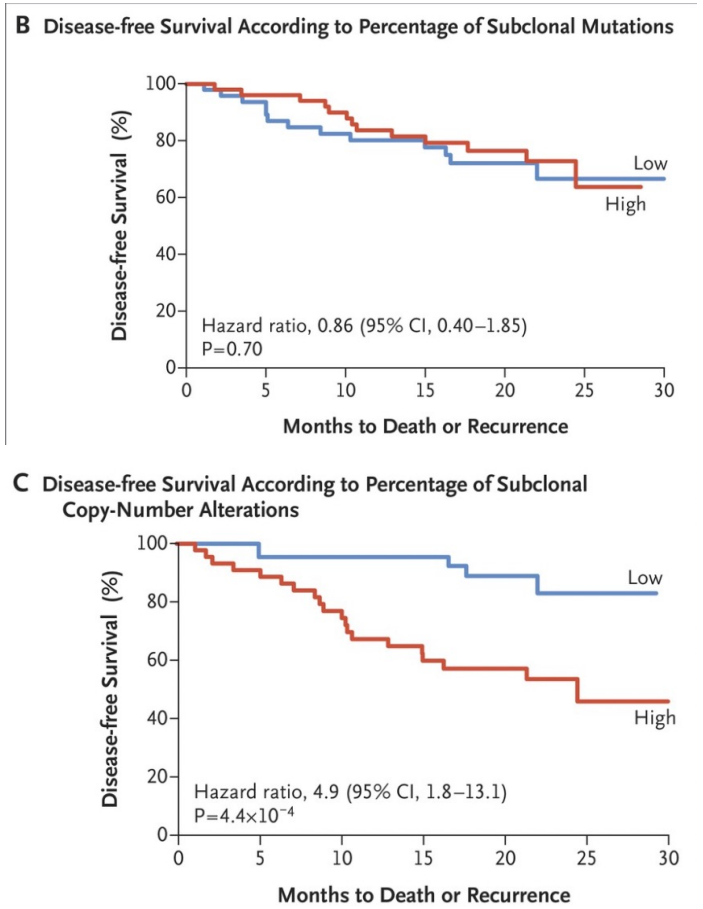


Background

Intratumor heterogeneity (ITH)

- ITH of tumor-infiltrated leukocytes (TILs) is an important phenomenon of cancer biology with potentially profound clinical impacts.
- Multi-region sequencing data provides a promising opportunity that allows the exploration of ITH, i.e.,
  - TRACERx study revealed the association of SCNA to patient prognosis (Jamal-Hanjani et al. 2017);
  - AbdulJabbar et al. 2020 studied differentiate highly immune-infiltrated tumor regions by the number of immune hot and cold tumors at a population level reveal patient survival.





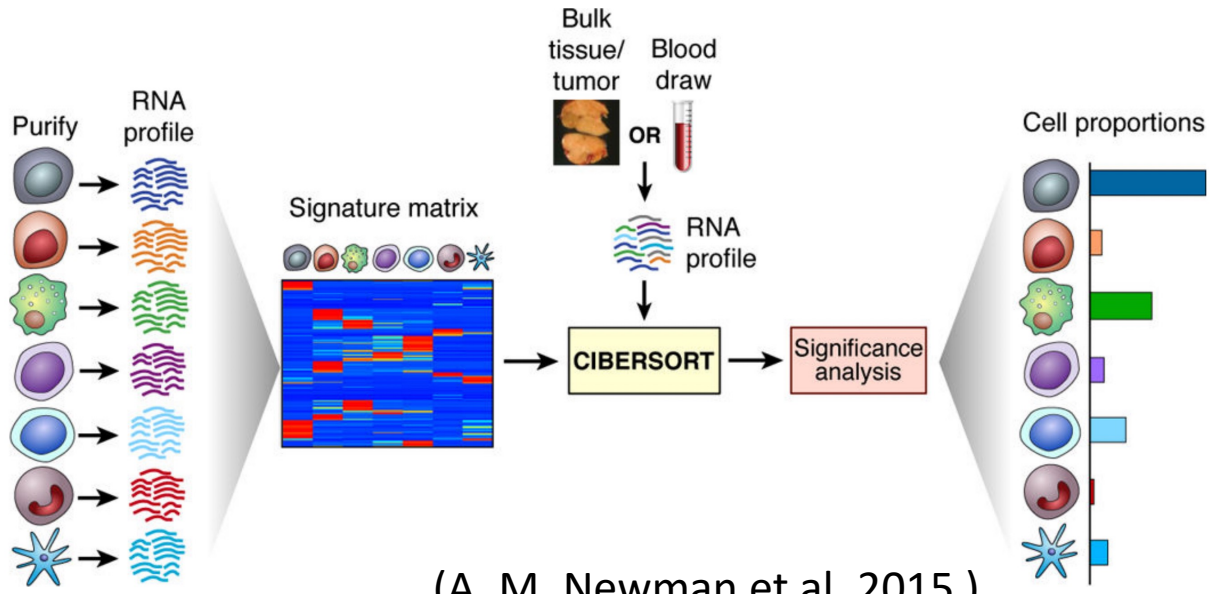
(Jamal-Hanjani et al. 2017)

Background

Motivation

- However, none of these studies have systematically studied the intratumor heterogeneities of TILs, which may provide valuable insights into cancer biology and personalized cancer treatment.

- Computational methods have developed to study the composition of the TIL in bulk gene expression data:
  - CIBERSORT applies linear support vector regressior (SVR) (A. M. Newman et al. 2015);
  - EPIC employs a weighted least squares and impose weights on informative genes.



(A. M. Newman et al. 2015)

- However, none of these methods developed for ITH and not suitable for multi-region design.

Background

Aim

- In this work, we aim to develop a computational method:
  - decompose mixed bulk gene expression data to estimate the relative immune cell abundance while accounting for the within-subject correlation.
  - assess the intratumor heterogeneity by the variability of cellular compositions from immune cells for each patient and seek its association with the survival outcomes.

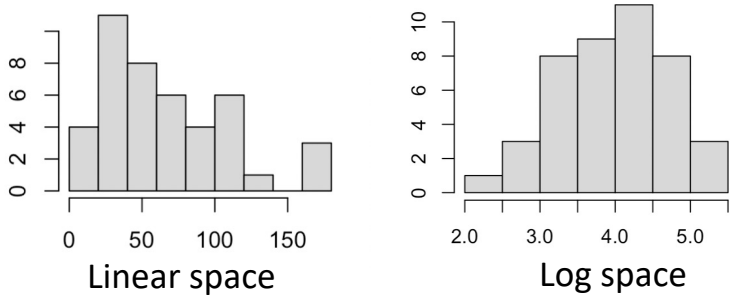
Statistical model

Reference profile construction

- Let X_{vgk} represents an observed cell-type-specific reference
  - v indicate the sample index belong to cell type k
  - g indicate the gene index
  - k indicate the cell type

$$\log(X_{vgk}) \sim N(\mu^r_{gk}, \frac{1}{\lambda^r_{gk}}),$$

- The observed gene expression is modeled as log-normal distribution
  - Estimated cell-type-specific mean expression $\hat{\mu}_{gk}$
  - Estimated cell-type-specific variability $\hat{\lambda}_{gk}$



	CT 1	CT 1	CT 2	CT 2	CT 3	CT 3
Gene 1	243	823	45	30	537	299
Gene 2	46	52	20	7	46	41
Gene 3	99	50	90	69	38	25
Gene 4	95	45	17	23	25	11
Gene 5	111	187	26	22	17	8
Gene 6	90	45	19	6	21	7
Gene 7	644	415	167	98	125	57
Gene 8	100	58	133	6	49	18
Gene 9	15	7	111	9	5	32
Gene 10	107	164	44	113	14	22
Gene 11	4582	3270	2104	5756	860	672
Gene 12	321	151	32	60	250	310
Gene 13	425	617	69	230	42	155

Statistical model

Mixed bulk data

- Consider a study of N patients. Let $I_i(|I_i| > 1)$ denote the tumor sample index for patient i.
- Let Y_{sg} represent the mixed gene expression from sample $s \, \epsilon \, I_i$

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

- h_{sk} is the unobserved cellular composition.
- W_{sgk} is a three-dimensional tensor that stands for the hidden.
- ϵ_{sg} is the error term that follow a normal distribution with mean 0 and variance $1/\lambda_{sg}$.

	Patient 1		Patient 2	
	Sample 1	Sample 2	Sample 3	Sample 4
Gene 1	83	134	134	249
Gene 2	22	62	78	32
Gene 3	48	171	145	24
Gene 4	50	28	53	33
Gene 5	118	67	60	164
Gene 6	21	74	36	21
Gene 7	80	219	148	144
Gene 8	28	45	49	29
Gene 9	6	12	19	19
Gene 10	47	179	49	135
Gene 11	3301	3670	643	5192
Gene 12	120	206	415	147
Gene 13	370	537	257	420

Statistical model

Mixed bulk data

- Consider a study of N patients. Let $I_i(|I_i| > 1)$ denote the tumor sample index for patient i.
- Let Y_{sg} represent the mixed gene expression from sample $s \in I_i$

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

$$\hat{\mu}_{gk}$$

- h_{sk} is the unobserved cellular composition.
- W_{sgk} is a three-dimensional tensor that stands for the hidden.
- ϵ_{sg} is the error term that follow a normal distribution with mean 0 and variance $1/\lambda_{sg}$.

	Patient 1		Patient 2	
	Sample 1	Sample 2	Sample 3	Sample 4
Gene 1	83	134	134	249
Gene 2	22	62	78	32
Gene 3	48	171	145	24
Gene 4	50	28	53	33
Gene 5	118	67	60	164
Gene 6	21	74	36	21
Gene 7	80	219	148	144
Gene 8	28	45	49	29
Gene 9	6	12	19	19
Gene 10	47	179	49	135
Gene 11	3301	3670	643	5192
Gene 12	120	206	415	147
Gene 13	370	537	257	420

Statistical model

Model hidden variables – cell-type-specific gene expression

- To characterize intratumor heterogeneity within the sample patient subject, a hierarchical Bayesian approach is taken in two steps:
  - First, we allow each patient to have his or her own pure cell type profile parameters μ_{igk}

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

where the hidden gene expression, W_{sgk}, also follows a log normal distribution.

We then center the patient-specific mean expression to cell-type-specific mean expression

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

  - ρ_{gk} controls how much information we borrow from the mean reference profile.
  - α_{gk} and β_{gk} determine the prior knowledge of the variability of gene expression.

Statistical model

Model hidden variables – cell-type-specific gene expression

- To characterize intratumor heterogeneity within the sample patient subject, a hierarchical Bayesian approach is taken in two steps:
  - First, we allow each patient to have his or her own pure cell type profile parameters μ_{igk}

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

where the hidden gene expression, W_{sgk}, also follows a log normal distribution.

We then center the patient-specific mean expression to cell-type-specific mean expression

$$\mu_{igk} \overset{\text{i.i.d}}{\sim} N(\overset{\hat{\mu}_{gk}}{\mu_{gk}}, \frac{1}{\rho_{gk}\lambda_{gk}}), \text{ for } i \in 1, \ldots, N, \quad \lambda_{gk} \sim \text{Gamma}(\overset{\hat{\lambda}_{gk}}{\alpha_{gk}, \beta_{gk}}).$$

  - ρ_{gk} controls how much information we borrow from the mean reference profile.
  - α_{gk} and β_{gk} determine the prior knowledge of the variability of gene expression.

Statistical model

Model hidden variables – relative cell type abundance

- Second, we use Dirichlet distribution to model the proportions of cell types for each sample

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

where

- $\boldsymbol{\pi}$ is a K by 1 vector pooled across all samples with $\sum_k \pi_k = 1$.
- C_i is a patient-specific parameter that controls the variability of the cellular composition across samples within each patient,
  - C_i tends to be **small**, it indicates a more **heterogeneity** cellular composition
  - C_i tends to be **large**, it indicates a more **homogeneous** cellular composition

Statistical model

Likelihood approximation

- Recall the observed gene expression Y_{sg} can be decomposed as:

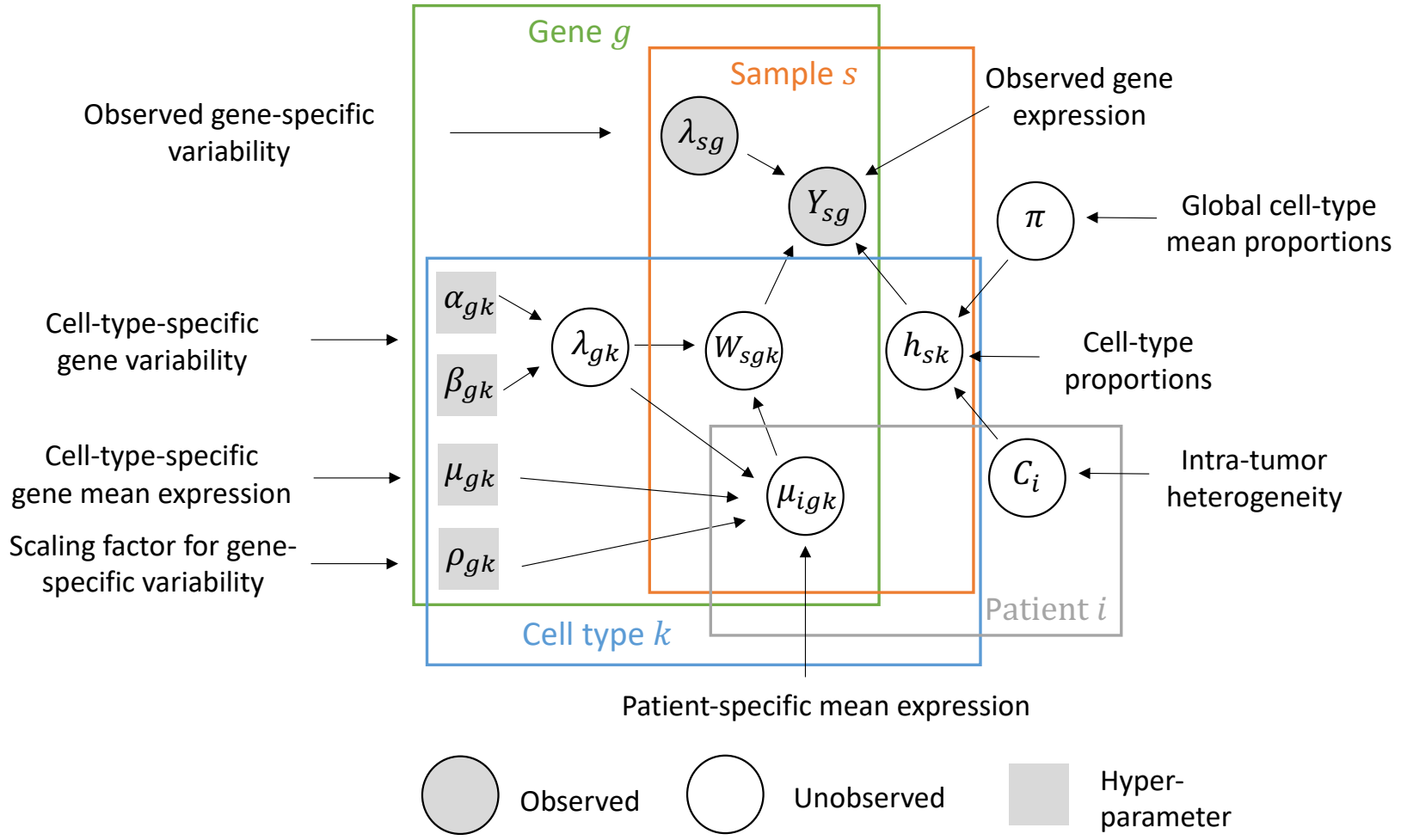
$$log(Y_{sg}) = \sum_k h_{sk} W_{sgk} + \epsilon_{sg}, \text{ for } s \in I_i,$$

- However, there is no closed form solution for a summation of independent log-normal distribution. We, therefore, adopt the Fenton-Wilkinson (FW) approximation (Fenton 1960) to approximate the likelihood by another log-normal distribution as follows:

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

Statistical model

Overview of model structure



Model optimization

Collapse over the hidden gene expression

- We adopt Collapsed Variational Bayesian (CVB) method to optimize the model.
  - Computationally more efficient than MCMC.
- To perform CVB method, we first marginalize over the hidden variables μ_{igk}'s and λ_{gk}'s

$$\prod_{i=1}^{N}\prod_{s\in I_i} P(\log(W_{sgk})|\mu_{gk},\rho_{gk},\alpha_{gk},\beta_{gk})$$

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

$$=\frac{\Gamma(\alpha_n)}{\Gamma(\alpha_{gk})}\frac{\beta_{gk}^{\alpha_{gk}}}{\beta_n^{\alpha_n}}\frac{\rho_{gk}^{N/2}}{\prod_{i=1}^{N}\sqrt{|I_i|+\rho_{gk}}}(2\pi)^{-\frac{\sum_i I_i}{2}},$$

where $\alpha_n=\frac{\sum_i I_i}{2}+\alpha_{gk}$, $\log(\bar{W}_{sgk})=\frac{1}{|I_i|}\sum_{s\in I_i}\log(W_{sgk})$, and

$$\beta_n=\beta_{gk}+\frac{1}{2}\sum_i\{\sum_{s\in I_i}(\log(W_{sgk})-\log(\bar{W}_{sgk}))^2+\frac{\rho_{gk}|I_i|(\log(\bar{W}_{sgk})-\mu_{gk})^2}{|I_i|+\rho_{gk}}\}.$$

Model optimization

Variational parameters

- To estimate the remaining hidden variables, we introduce the following variational distributions:

$$Q(\log(W_{sgk})|\gamma_{igk}, \tau_{gk}) \sim N(\gamma_{igk}, \tau_{gk}^2), \text{ for } s \in I_i,\ i = 1, 2, \ldots, N;\ g = 1, 2, \ldots, G;\ k = 1, 2, \ldots, K;$$

$$Q(h_{s1}, \ldots, h_{sK}|\xi_{s1}, \ldots, \xi_{sK}) \sim \text{Dir}(\xi_{s1}, \ldots, \xi_{sK}) \text{ for } s \in I_i;\ i = 1, 2, \ldots, N;\ k = 1, 2, \ldots, K,$$

where

- $Q(\cdot)$ denotes the variational distribution to approximate the posterior distribution.
- $\{\gamma_{igk}\}_{i,g,k}$, $\{\tau_{gk}\}_{g,k}$ and $\{\xi_{sk}\}_{s,k}$ are variational parameters to be optimized.
- In total, there are $(N \times G \times K) + (G \times K) + (S \times K)$ parameters to be estimated.

Model optimization

Derivation of the evidence lower bound

- Let $Z = (W, H)$ denote the unobserved variables of interests and $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\rho}, \boldsymbol{\mu}, \boldsymbol{\lambda})$ denote the hyper-parameters.

$$\log(P(Y)) \geq \underbrace{E_{Q(W,H)}\{\log\frac{P(Y,W,H|\boldsymbol{\theta})}{Q(W,H)}\}}_{\text{ELBO}}$$

$$\geq \int Q(W)Q(H)\log(\frac{P(Y|W,H,\boldsymbol{\theta})P(W|\boldsymbol{\theta})P(H)}{Q(W)Q(H)})dZ$$

$$\geq \underbrace{E_Q\{\log P(Y|W,H,\lambda)\}}_{\text{a}} + \underbrace{E_Q\{\log P(W|\mu,\rho,\alpha,\beta)\}}_{\text{b}} + \underbrace{E_Q\{\log P(H|C,\pi)\}}_{\text{c}}$$

$$- \underbrace{E_Q\{\log P(W|\gamma,\tau)\}}_{\text{d}} - \underbrace{E_Q\{\log P(H|\xi)\}}_{\text{e}}.$$

- We applied Limited-memory BFGS to maximize this objective function iteratively.
- The gradients with respect to variational parameters have been derived to speed the optimization.

Model optimization

Empirical solution through moments

- The relative cell type abundance can be estimated by variational parameters.
  - specifically for cell type k, $\hat{h}_{sk} = \frac{\xi_{sk}}{\sum_c \xi_{sc}}$
- The expectation of the fraction of cell type k can be obtained as $E[h_{.k}] = \frac{C_i \pi_k}{C_i \sum_c \pi_c} = \hat{\pi}_k$
- The intratumor heterogeneity score C_i for each patient can be computed through the first and second moments,

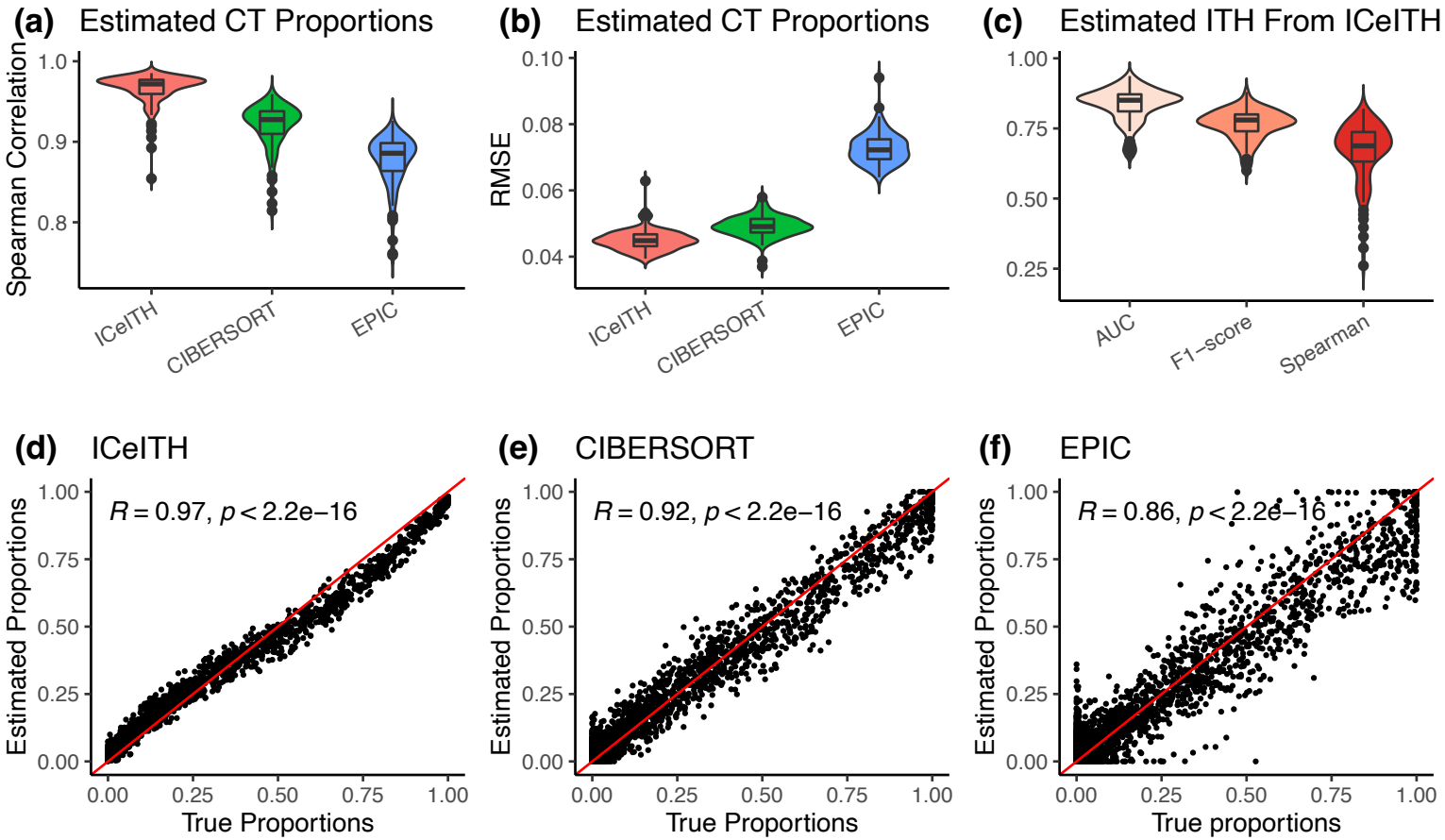
$$\hat{C}_i = f_C(\hat{\pi}_k, \hat{h}_{sk}) \stackrel{c}{=} \frac{\sum_k \hat{\pi}_k(1-\hat{\pi}_k)}{\sum_k var(h_{sk})}.$$

where $\sum_k var(h_{sk}) = \sum_k \frac{\hat{\pi}_k(1-\hat{\pi}_k)}{C_i+1}$

Simulation

Simulation settings and results

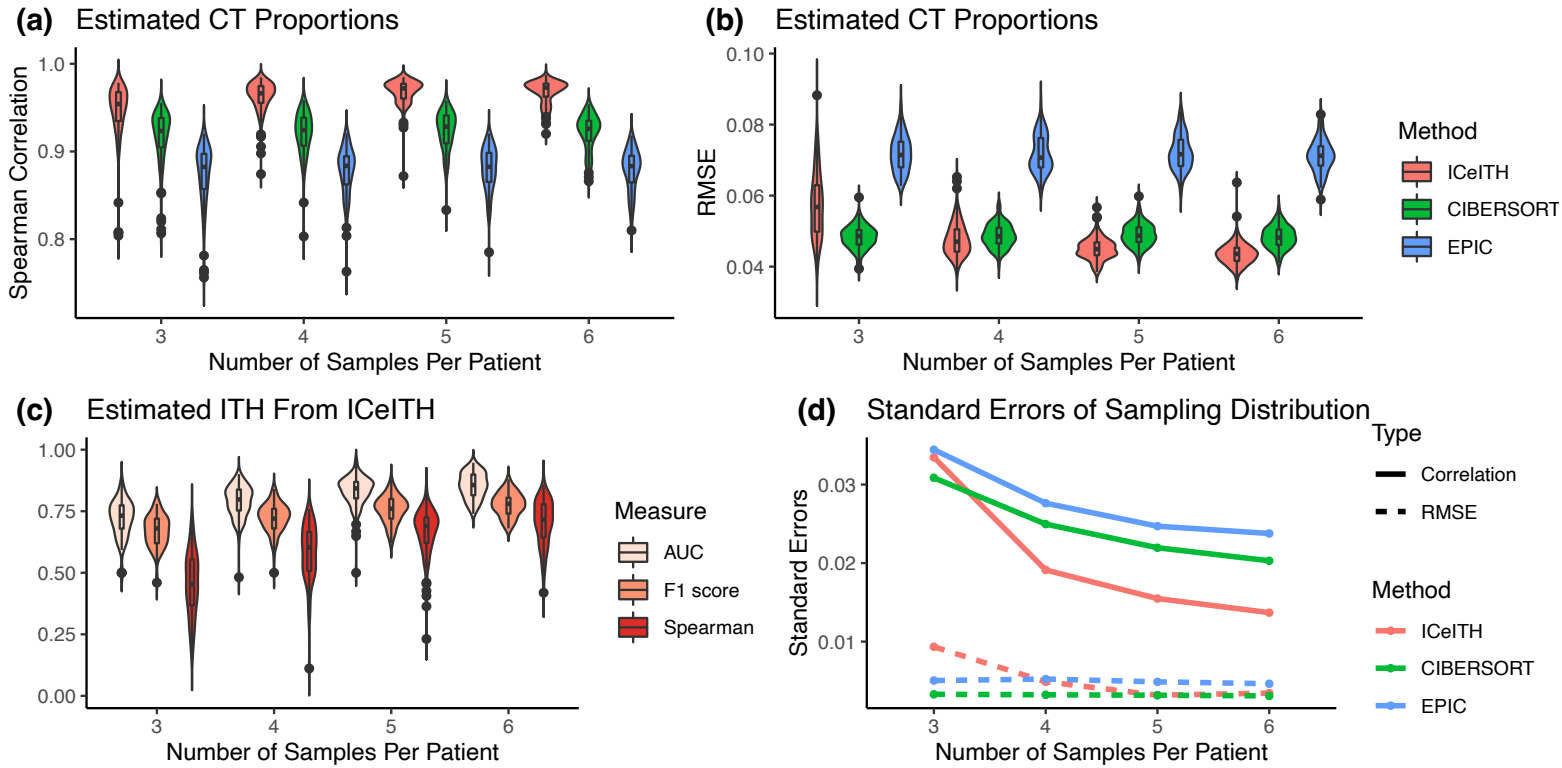
- We consider 100 patient, with 4 to 8 samples randomly assigned to each patient.
- Mixed gene expression with 500 genes and 4 cell types are generated.
- We estimated the cell types and benchmarked our results with CIBERSORT and EPIC.



Simulation

Sensitivity analysis and results

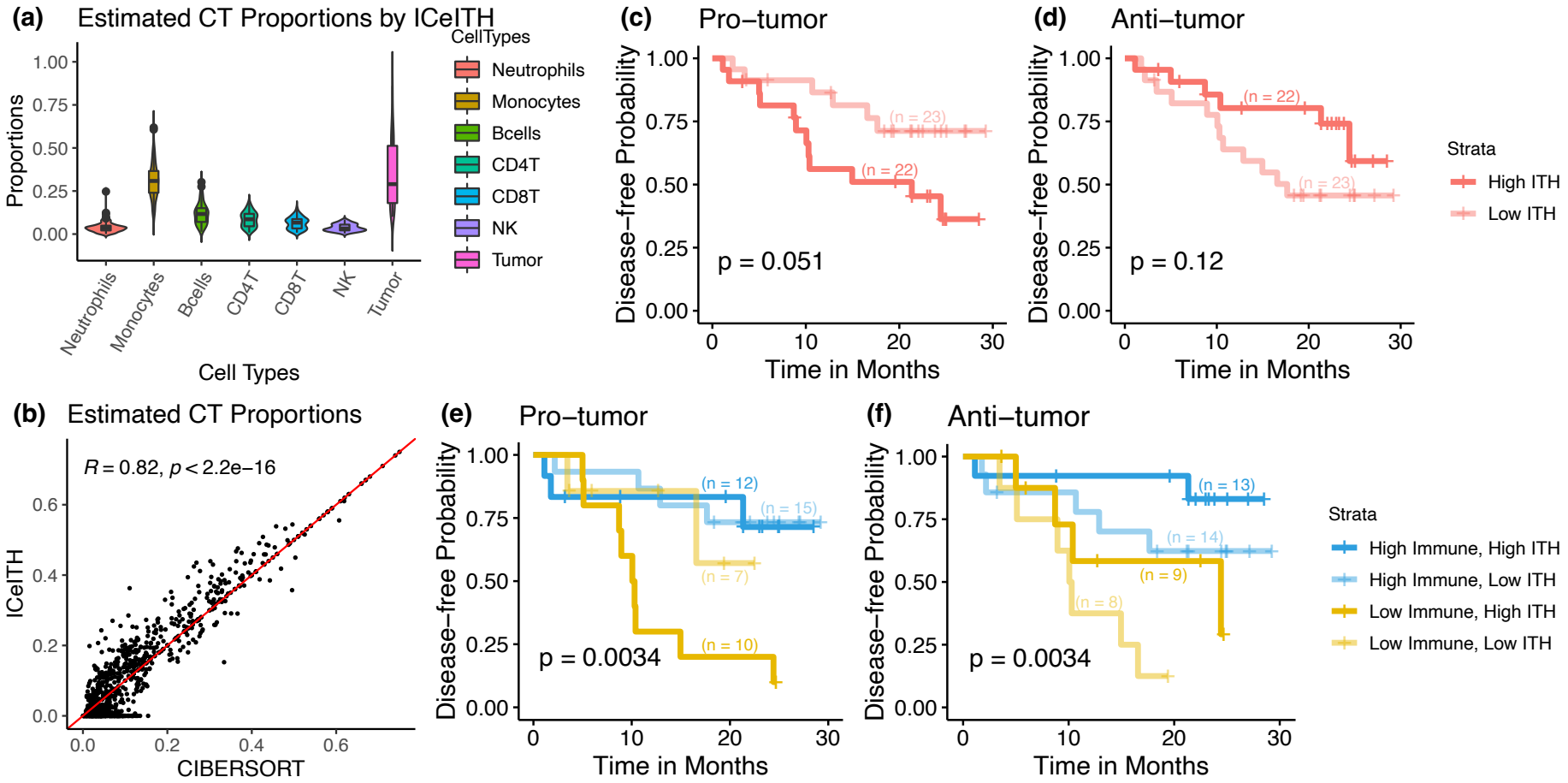
- We consider 100 patient, with fixed number of samples for each patient.
- Mixed gene expression with 500 genes and 4 cell types are generated.
- We estimated the cell types and benchmarked our results with CIBERSORT and EPIC.



Real data application

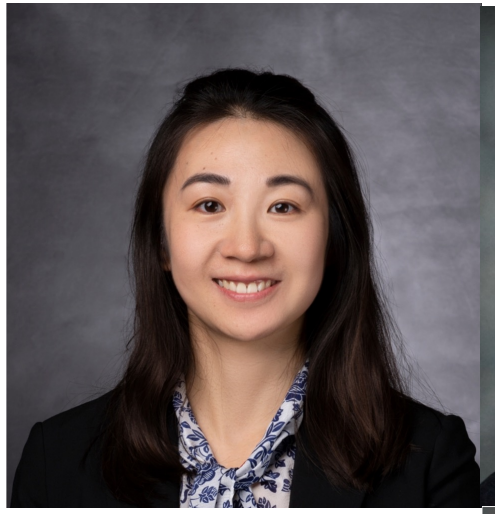
TRACERx Study

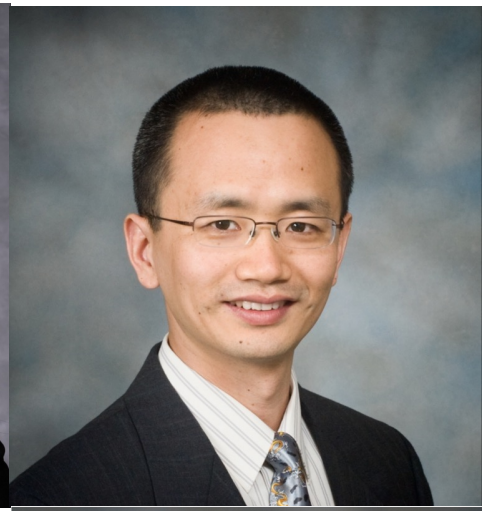
- The multi-region RNA-seq data are available in 45 patients with 140 samples in total.



Conclusion

- In this project, we proposed a Bayesian hierarchical model, ICeITH, to estimate the relative cell type abundance by leveraging the prior information while accounting for the within-subject correlations.
- ICeITH assesses the intratumor heterogeneity by quantifying the variability of the targeted cellular composition and reveals the association between heterogeneity of immune cells to the patient survival.
- We develop an efficient variational inference approach to the model estimation, and the method is available through a user-friendly R package on Github (https://github.com/pengyang0411/ICeITH).





THE UNIVERSITY OF TEXAS
MD Anderson Cancer Center
Making Cancer History®

Reference

- AbdulJabbar, Khalid et al. (2020). “Geospatial immune variability illuminates differential evolution of lung adenocarcinoma”. In: Nature Medicine 26.7, pp. 1054–1062.
- Andrade Barbosa, B ́arbara et al. (2021). “Bayesian log-normal deconvolution for enhanced in silico microdissection of bulk gene expression data”. In: Nature communications 12.1, pp. 1–13.
- Jamal-Hanjani, Mariam et al. (2017). “Tracking the evolution of non–small-cell lung can- cer”. In: New England Journal of Medicine 376.22, pp. 2109–2121.
- Newman, Aaron M et al. (2015). “Robust enumeration of cell subsets from tissue expression profiles”. In: Nature methods 12.5, pp. 453–457.
- Racle, Julien and David Gfeller (2020). “EPIC: a tool to estimate the proportions of dif- ferent cell types from bulk gene expression data”. In: Bioinformatics for Cancer Im- munotherapy. Springer, pp. 233–248.
- Wilson, Douglas R et al. (2020). “ICeD-T Provides Accurate Estimates of Immune Cell Abundance in Tumor Samples by Allowing for Aberrant Gene Expression Patterns”. In: Journal of the American Statistical Association 115.531, pp. 1055–1065.