A Pan-Cancer and Polygenic Bayesian Hierarchical Model for Effect of Somatic Mutations on Survival

Sarah Samorodnitsky[1], Katherine A. Hoadley[2], Eric F. Lock[1]
[1]Division of Biostatistics, University of Minnesota; [2]Department of Genetics, Computational Medicine Program, Lineberger Comprehensive Care Center, University of North Carolina at Chapel Hill

SCHOOL OF PUBLIC HEALTH
UNIVERSITY OF MINNESOTA

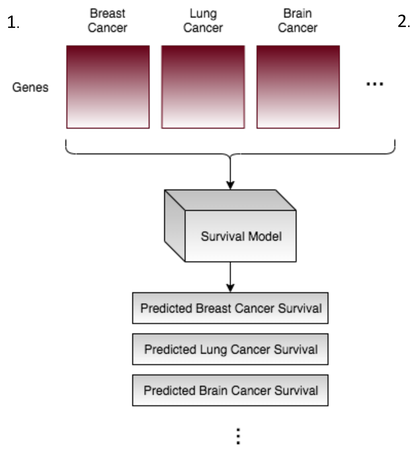
INTRODUCTION

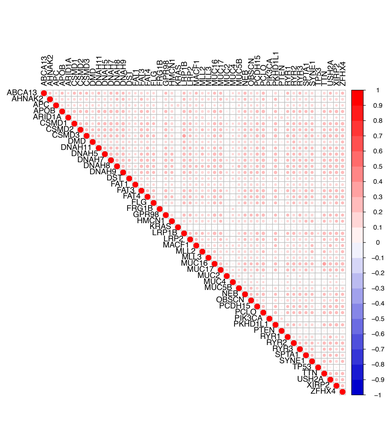
- The Cancer Genome Atlas (TCGA): molecular data for 33 types of cancer, over 10,000 patients.[1]
- 2013: TCGA began Pan-Cancer Analysis Project for study of themes consistent across cancer types.
  - Motivation: Cancers from same tissue often distinct while cancers from different tissues often similar.[2]
  - Results: discovery of shared somatic mutations across cancer types.[3]
- Our goal: build a pan-cancer model for overall patient survival based on age and somatic mutation status.[4]
  - Assess which genes most predictive of survival.

MOTIVATION

- Somatic mutation data: Broad Institute GDAC Firehose
- Clinical data: TCGA Clinical Data Resource.
  - Removed cancers with survival rates > 90%
  - Selected top 50 highest mutated genes on average.
  - Total: 5698 patients from 27 cancer types, 50 genes.
- Modeling approach: Bayesian hierarchical model
  - Effect of each predictor allowed to vary by cancer type.
  - Allows "borrowing" of information across cancers
  - Gibbs sampling approach to infer posterior
- Model selection: Compared normal, log-normal, exponential, Weibull model frameworks
  - Selected best model based on cross-validation of posterior predictive likelihood
- Forward selection procedure to assess marginal contribution of each gene on prediction.

FIGURES 1 & 2: MODEL SCHEMATIC AND CORRELATION PLOT





Fig, 1: Model schematic shows data matrix for cancer types being used in a survival model for prediction of survival for several cancers simultaneously.
Fig. 2: Correlation plot for mutation status of all 50 genes across 27 cancer types.

MODEL

- y_{ij} = (possibly censored) survival time for patient j, cancer type $i, j = 1, \dots, n_i, i = 1, \dots, 27$
- Four parametric survival models:

$$y_{ij} \sim \text{Normal}(\lambda_{ij}, \sigma^2) \qquad y_{ij} \sim \text{Log-Normal}(\lambda_{ij}, \sigma^2)$$
$$y_{ij} \sim \text{Exponential}\left(\frac{1}{\lambda_{ij}}\right) \qquad y_{ij} \sim \text{Weibull}(\alpha, \frac{1}{\lambda_{ij}})$$

- where $\lambda_{ij} = \beta_{i0} + \beta_{i1}x_{ij1} + \beta_{i2}x_{ij2} + \dots + \beta_{i51}x_{ij51}$
  - x_{ijp} = centered age if $p = 1$, mutation status for gene $p - 1, p = 2, \dots, 51$.
- Effect of each covariate, β_{ip}, varies by cancer type.
  - $\beta_{ip} \sim \text{Normal}(\tilde{\beta}_p, \lambda_p^2), p = 0, \dots, 51$
  - $\tilde{\beta}_p \sim \text{Normal}(0, 10000^2)$
  - $\lambda_p^2 \sim \text{Inverse-Gamma}(0.01, 0.01)$
- Normal and log-normal models: survival time variance $\sigma^2 \sim \text{Inverse-Gamma}(0.01, 0.01)$
- Weibull model: shape parameter $\alpha \sim \text{Uniform}(0, 5)$

RESULTS

- For pan-cancer modeling, log-normal model fit data best.
  - Modeling cancers individually: model cannot converge, small sample size leading to inference problems.
- TP53 (mutation rate: 38.3%) and FAT4 (mutation rate: 8.8%) together most predictive of patient survival.
- If TP53 excluded, APOB (mutation rate: 7.7%) most predictive.
- Overall negative effect of TP53 mutation on survival, seemingly positive effect of FAT4 mutation on survival.

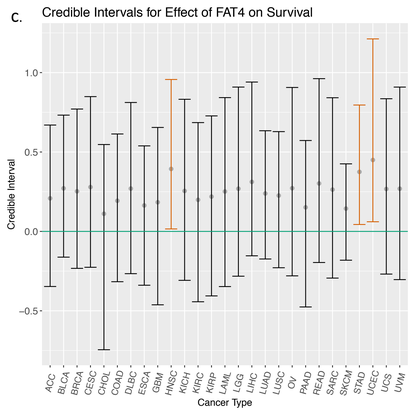
a.

Covariates In Model (TP53 Included)	Mean Log Posterior Likelihood
Age, No Genes	-5014.895
Age, TP53	-1009.630
Age, TP53, FAT4	-1009.135
Age, TP53, FAT4, DNAH5	-1009.179

b.

Covariates In Model (TP53 Excluded)	Mean Log Posterior Likelihood
Age, No Genes	-5014.895
Age, APOB	-1011.321
Age, APOB, ARID1A	-1011.694

c.



d.

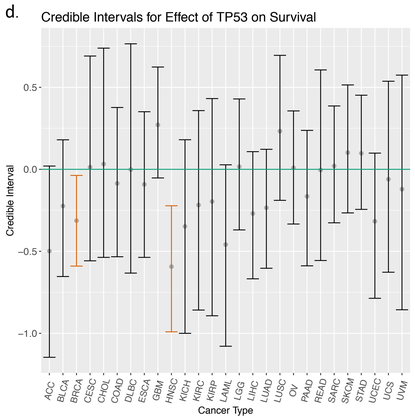
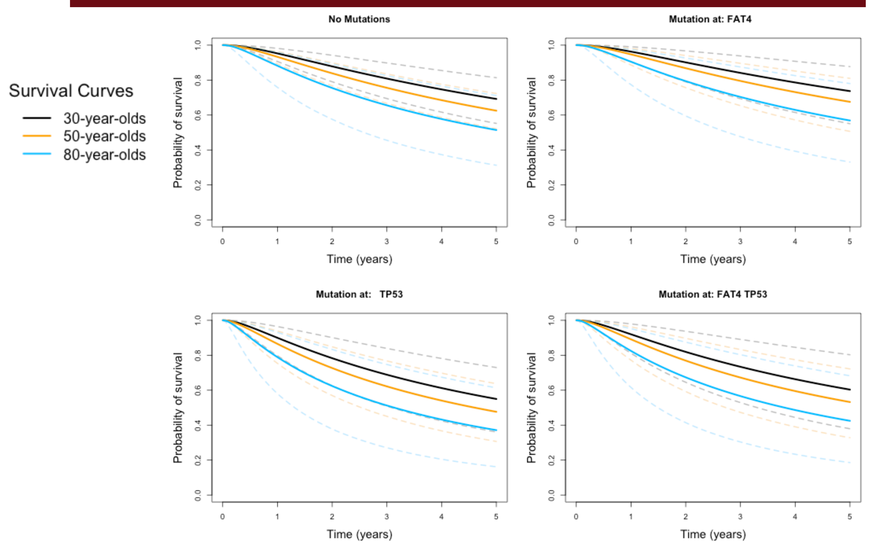


Table (a) summarizes the posterior likelihoods for forward selection when TP53 is considered as a covariate. Table (b) summarizes the results from forward selection when TP53 is not considered as a covariate. Figures (c) and (d) show the credible interval estimates for the effects of a mutation at TP53 and FAT4 on survival by cancer type. Orange-highlighted intervals are entirely above or below 0, which is highlighted in green.

FIGURE 2. PREDICTED SURVIVAL CURVES FOR PATIENTS WITH ACC



Predicted survival curves for patients with Adrenocortical Carcinoma (ACC) based on different mutation combinations. Black lines refer to 30-year-old patients, orange lines refer to 50-year-old patients, and blue lines refer to 80-year-old patients.

CONCLUSIONS

- Our modeling approach – with several genes and uninformative priors - only worked if cancers were modeled jointly
- TP53 and FAT4 were together the most predictive of survival
  - TP53 dominated in predictive strength, all genes afterwards contributed less
- Positive credible intervals for FAT4 may be an artifact of FAT4 mutations being comparably less deleterious to TP53 mutations
- Future work: choose specific sets of cancers that are known to be related; choose gene subset differently ;include interactions in model.

REFERENCES

[1]Hutter C and Zenklusen JC. The Cancer Genome Atlas: creating lasting value beyond its data. *Cell* 2018; 173(2): 283–285.
[2]Weinstein JN, Collisson EA, Mills GB et al. The cancer genome atlas pan-cancer analysis project. *Nature genetics* 2013; 45(10): 1113–1120.
[3]Kandoth C, McLellan MD, Vandin F et al. Mutational landscape and significance across 12 major cancer types. *Nature* 2013; 502(7471): 333.
[4]Samorodnitsky, S., Hoadley, K. A., & Lock, E. F. (2020). A Pan-Cancer and Polygenic Bayesian Hierarchical Model for the Effect of Somatic Mutations on Survival. *Cancer Informatics*. https://doi.org/10.1177/1176935120907399

FUNDING

This work was supported by the National Institutes of Health (NIH) National Cancer Institute (NCI) grant R21CA231214-01.

CONTACT INFORMATION

Sarah Samorodnitsky
PhD Student
samor007@umn.edu
(607) 280 - 2673

Learn more: sph.umn.edu