A Statistical Framework to Model Somatic Mutations in Cancer

Siming Zhao, Xin He and Matthew Stephens

Department of Human Genetics, University of Chicago

Abstract

We proposed a Bayesian statistical approach to model somatic mutations identified in cancer sequencing studies. This approach will help us to understand the underlying mechanisms of tumorigenesis as well as pinpoint new genetic drivers for cancers.

- We developed a novel model of background mutation rates
- We used multiple features to predict positive selection in driver genes
- Different models for Tumor Suppressor Genes (TSGs)/Oncogenes (OCGs)
- Our method is flexible and has more power than current mainstream methods.

Introduction

- Cancer is a disease driven by acquired genetic mutations (somatic mutations).

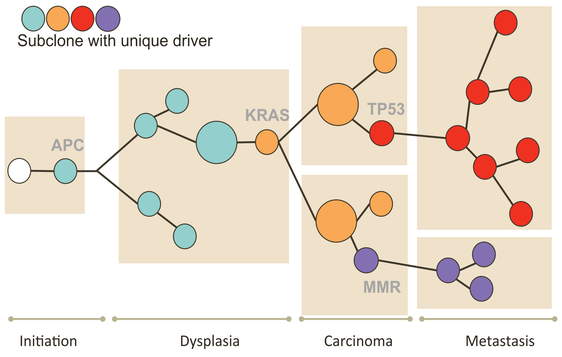

Figure 1: Mutational process underlying carcinogenesis

- Representive TSGs and OCGs with mutations identified in cancer sequencing studies [1].

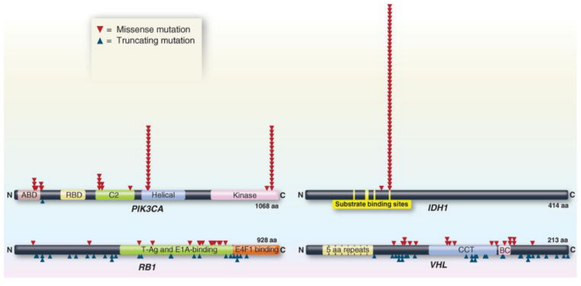

Figure 2: Representive TSGs and OCGs

Data

Somatic mutation lists for 26 types of tumors were collected from the Cancer Genome Atlas (TCGA). These datasets provide somatic single nucleotide changes for sequenced coding regions.

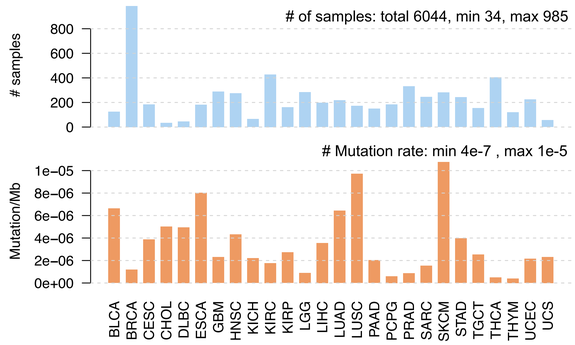

Figure 3: Somatic mutation profile across 26 TCGA cohorts

Model

At each genomic position i, we counted the number of mutations with type t across all tumors in this study, denoted by Y_{it}.

$$Y_{it} \sim_{ind} Poisson(\phi_{it}) \quad (1)$$

$$\phi_{it} = \mu_{it}\lambda_{g(i)}\gamma_{it} \quad (2)$$

The parameter ϕ_{it} denotes the rate at which "type t" mutations occur at position i. The parameter μ_{it} denotes the background mutation rate at i, t determined by mutational features of the mutation:

$$\log \mu_{it} = \beta^b_{t0} + \mathbf{x}^b_{it}\beta^b \quad (3)$$

The parameter $\lambda_{g(i)}$ adjusts for gene level background mutation rate:

$$\lambda_g \sim_{ind} Gamma(\alpha, \alpha) \quad (4)$$

The parameter γ_{it} denotes the increased mutation rate attributes to the selection advantage introduced by the mutation i, t. It is determined by:

$$\log \gamma_{it} = \beta^f_0 + \mathbf{x}^f_{it}\beta^f \quad (5)$$

We propose three models based on the roles of the mutation in cancer. In Model $M_0(\alpha, \lambda_{g(i)}, \beta^b_{t0}, \beta^b, \beta^f_0, \beta^f)$, the mutations come from a non cancer gene; Model M_1, the mutations come from TSGs; Model M_2, the mutations come from OCGs.

Fitted model parameters

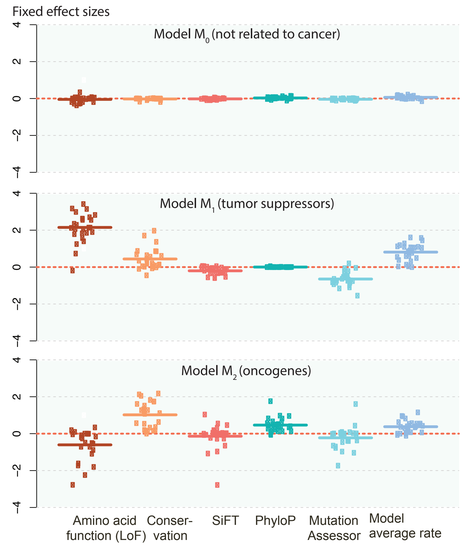

Figure 4: Effect sizes of 6 functional features(β^f, β^f_0) for 3 models

Results 1

Our method identified more known cancer genes in top ranked genes compared to the results generated by MutsigCV[2], which is the mainstream method in the field.

Threshold	Top 20	Top 100
Our method	240	477
MutsigCV	201	346

Table 1: Total number of known cancer genes in top ranked genes

We could also control false positives by setting up a FDR cut off of 0.1. At this threshold, we identified more novel cancer genes than MutsigCV.

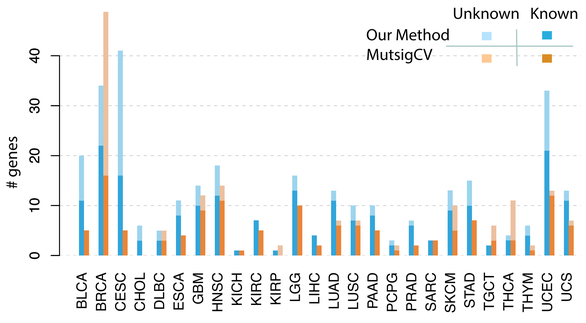

Figure 5: Known and unknown cancer genes identified at FDR 0.1

Results 2

Of the newly identified cancer genes, we found some new gene families recurrently mutated. The following table listed two families, N(alpha)-acetyltransferase family and protocadherin family.

Tumor	Gene	Percent	#nonsilent
CESC	NAA15	3.3%	6
ESCA	NAA16	2.7%	5
HNSC	NAA25	2.5%	7
UCEC	NAA30	1.8%	4
LUAD	PCDHA2	5.9%	10
BRCA	PCDHB7	1.3%	13
PRAD	PCDH18	2.4%	8

Table 2: Recurrently mutated novel gene families

Conclusion

More power. our method has a greater power to identify cancer genes across a wide range of tumors.

Novel cancer genes. Some new families of genes appeared significant in our lists and they provide potential new therapeutic targets for cancer.

Ongoing work

We are trying to incorporate more features in our model:

- model with spatial clustering for oncogenes
- DNA structural variation
- tissue type specific annotations

References

[1] Bert et al Vogelstein.
Cancer Genome Landscapes.
Science, 339(6127):1546 LP – 1558, mar 2013.

[2] Michael S et al Lawrence.
Mutational heterogeneity in cancer and the search for new cancer-associated genes.
Nature, 499(7457):214–8, jul 2013.

Acknowledgements

We want to thank the Research Computing Center(RCC) at University of Chicago for providing wonderful computing resources for this project.