Data-driven design of targeted gene panels for estimating immunotherapy biomarkers

Jacob R. Bradley and Timothy I. Cannings
School of Mathematics, University of Edinburgh

Abstract

We introduce a novel data-driven framework for the design of targeted gene panels for estimating exome-wide biomarkers in cancer immunotherapy. Our first goal is to develop a generative model for the profile of mutation across the exome, which allows for gene- and variant type-dependent mutation rates. Based on this model, we then propose a new procedure for estimating biomarkers such as Tumour Mutation Burden and Tumour Indel Burden. Our approach allows the practitioner to construct a targeted gene panel of a prespecified size, alongside an estimator that only depends on the selected genes, which facilitates cost-effective prediction. Alternatively, the practitioner may apply our method to make predictions based on an existing gene panel, or to augment a gene panel to a given size. We demonstrate the excellent performance of our proposal using an annotated mutation dataset from 1144 Non-Small Cell Lung Cancer patients.

Keywords: cancer, gene panel design, tumour indel burden, tumour mutation burden.

1 Overview

Tumour Mutation Burden (TMB) is an emerging genomic biomarker of response to immunotherapy in a variety of cancer types. In this paper we propose a novel, data-driven method for estimating TMB and other biomarkers, based on a generative model of how mutations arise in the tumour exome. More precisely, we model mutation counts as independent Poisson variables, where the mean number of mutations depends on the gene of origin and variant type, as well as the Background Mutation Rate (BMR) of the tumour. Due to the ultrahigh-dimensional nature of sequencing data, we use a regularisation penalty when estimating the model parameters, in order to reflect the fact that in many genes' mutations arise purely according to the BMR. In addition, this identifies a subset of genes that are mutated above or below the background rate. Our model facilitates the construction of a new estimator of TMB, based on a weighted linear combination of the number of mutations in each gene. The vector of weights is chosen to be sparse (i.e. have many entries equal to zero), so that our estimator of TMB may be calculated using only the mutation counts in a subset of genes. In particular, this allows for accurate estimation of TMB from a targeted gene panel, where the panel size (and therefore the cost) may be determined by the user. We demonstrate the excellent practical performance of our framework using a Non-Small Cell Lung Cancer (NSCLC) dataset, and include a comparison with the existing state-of-the-art data-driven approaches for estimating TMB – see the results higlights below.

Further contributions in our paper include the following: since our model allows variant type-dependent mutation rates, it can be adapted easily to predict other biomarkers, such as Tumour Indel Burden (TIB). Our method may also be used in combination with an existing targeted gene panel. In particular, we can estimate a biomarker directly from the panel, or first augment the panel and then construct an estimator. We also discuss a number of practical considerations in detail.

2 Results highlights

To demonstrate the performance of our proposal we make use of the NSCLC dataset produced by Campbell et al. (2016), which contains data from 1144 patient-derived tumours, and apply our method for estimating TMB. We compare our method with state-of-the-art estimators applied to commonly used gene panels. The three next-generation sequencing panels that we consider are chosen for their relevance to TMB. These are TST-170 (Heydt et al., 2018), Foundation One (Frampton et al., 2013) and MSK-IMPACT (Cheng et al., 2015). For each panel we use four different methods to predict TMB: a) Our refitted estimator applied to the panel P; b) Estimation and Classification of Tumour Mutation Burden (ecTMB): the procedure proposed by Yao et al. (2020); c) A count estimator; and d) A linear model. The latter three comprise existing methods for estimating TMB available to practitioners. The second (ecTMB), which is based on a negative binomial model, is the state-of-the-art.

Figure 1 highlights the potential utility of our proposal: First, for each of the three panels considered here, we see that our refitted estimator applied to the panel outperforms all existing approaches in terms of regression performance, and that for smaller panels we are able to improve regression accuracy even further by selecting a panel based on the training data. For instance, in comparison to predictions based on the TST-170 panel, our

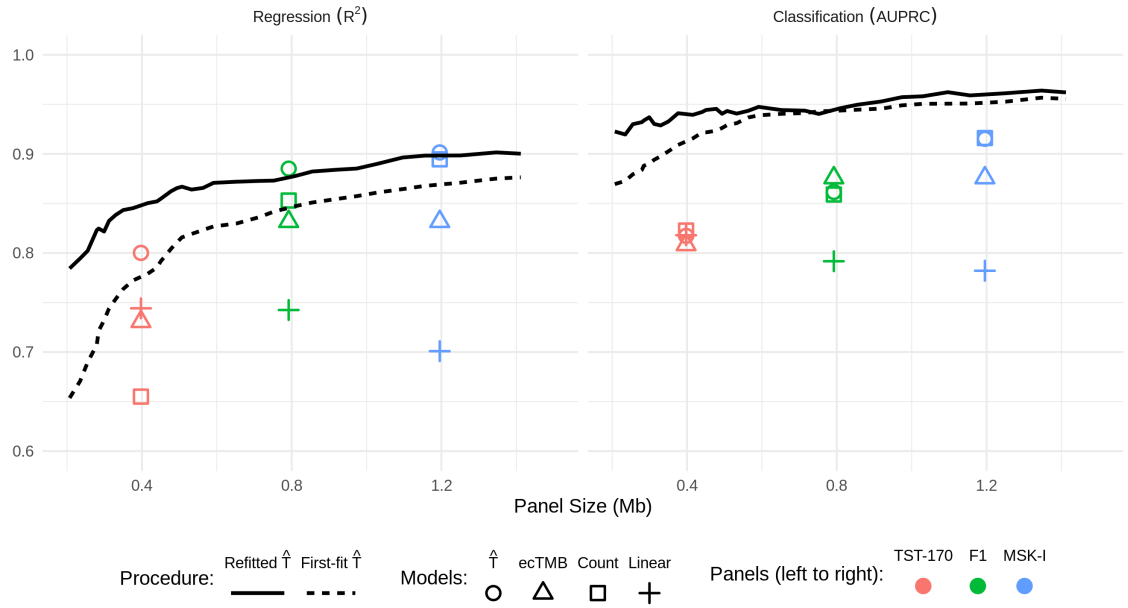

Figure 1: The performance of our TMB estimator in comparison to existing approaches. **Left**: R^2, **Right**: area under the precision-recall curve (AUPRC).

procedure with a selected panel of the same size (0.4Mb) achieves an R^2 of 0.85. The best available existing method based on the TST-170 panel, in this case the linear estimator, has an R^2 of 0.74. Moreover, data-driven selection of panels considerably increases the classification performance for the whole range of panel sizes considered. In particular, even for the smallest panel size shown in Figure 1 ($\sim$0.2Mb), the classification performance of our method outperforms the best existing methodology applied to the MSK-IMPACT panel, despite being almost a factor of six times smaller.

3 Conclusions

We introduce a new data-driven framework for designing targeted gene panels which allows for cost-effective estimation of exome-wide biomarkers, such as Tumour Mutation Burden and Tumour Indel Burden. Using the Non-Small Cell Lung Cancer dataset from Campbell et al. (2016), we demonstrate the excellent predictive performance of our proposal, and show that it outperforms the existing state-of-the-art procedures. Our framework can be applied to any tumour dataset containing annotated mutations, and we provide an `R` package (Bradley and Cannings, 2021) which implements the methodology.

References

J. R. Bradley and T. I. Cannings. Immune Checkpoint BioMarkers (R Package), Jan. 2021. URL https://github.com/cobrbra/ICBioMark.

J. D. Campbell, A. Alexandrov, J. Kim, J. Wala, A. H. Berger, C. S. Pedamallu, S. A. Shukla, G. Guo, A. N. Brooks, B. A. Murray, M. Imielinski, X. Hu, S. Ling, R. Akbani, M. Rosenberg, C. Cibulskis, A. Ramachandran, E. A. Collisson, D. J. Kwiatkowski, M. S.

Lawrence, J. N. Weinstein, R. G. W. Verhaak, C. J. Wu, P. S. Hammerman, A. D. Cherniack, G. Getz, Cancer Genome Atlas Research Network, M. N. Artyomov, R. Schreiber, R. Govindan, and M. Meyerson. Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nature Genetics*, 48(6):607–616, 2016. ISSN 1546-1718. doi: 10.1038/ng.3564.

D. T. Cheng, T. N. Mitchell, A. Zehir, R. H. Shah, R. Benayed, A. Syed, R. Chandramohan, Z. Y. Liu, H. H. Won, S. N. Scott, A. R. Brannon, C. O'Reilly, J. Sadowska, J. Casanova, A. Yannes, J. F. Hechtman, J. Yao, W. Song, D. S. Ross, A. Oultache, S. Dogan, L. Borsu, M. Hameed, K. Nafa, M. E. Arcila, M. Ladanyi, and M. F. Berger. Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT). *The Journal of Molecular Diagnostics : JMD*, 17(3):251–264, May 2015. ISSN 1525-1578. doi: 10.1016/j.jmoldx.2014.12.006. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5808190/.

G. M. Frampton, A. Fichtenholtz, G. A. Otto, K. Wang, S. R. Downing, J. He, M. Schnall-Levin, J. White, E. M. Sanford, P. An, J. Sun, F. Juhn, K. Brennan, K. Iwanik, A. Maillet, J. Buell, E. White, M. Zhao, S. Balasubramanian, S. Terzic, T. Richards, V. Banning, L. Garcia, K. Mahoney, Z. Zwirko, A. Donahue, H. Beltran, J. M. Mosquera, M. A. Rubin, S. Dogan, C. V. Hedvat, M. F. Berger, L. Pusztai, M. Lechner, C. Boshoff, M. Jarosz, C. Vietz, A. Parker, V. A. Miller, J. S. Ross, J. Curran, M. T. Cronin, P. J. Stephens, D. Lipson, and R. Yelensky. Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nature Biotechnology*, 31(11):1023–1031, Nov. 2013. ISSN 1546-1696. doi: 10.1038/nbt.2696.

C. Heydt, R. Pappesch, K. Stecker, J. Neumann, R. Buettner, and S. Merkelbach-Bruse. Evaluation of the TruSight Tumor 170 (TST170) assay and its value in clinical research. *Annals of Oncology*, 29:vi7–vi8, Sept. 2018. ISSN 0923-7534, 1569-8041. doi: 10.1093/annonc/mdy318.003. URL https://www.annalsofoncology.org/article/S0923-7534(19)32372-5/abstract.

L. Yao, Y. Fu, M. Mohiyuddin, and H. Y. K. Lam. ecTMB: a robust method to estimate and classify tumor mutational burden. *Scientific Reports*, 10(1):1–10, Mar. 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-61575-1. URL https://www.nature.com/articles/s41598-020-61575-1. Number: 1 Publisher: Nature Publishing Group.