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.

It has been understood for a long time that cancer, a disease occurring in many distinct tissues of the body and giving rise to a wide range of presentations, is initiated and driven by the accumulation of mutations in a subset of a person's cells. Since the discovery of Immune Checkpoint Blockade (ICB)[1], there has been an explosion of interest in cancer therapies targeting immune response and ICB therapy is now widely used in clinical practice. ICB therapy works by targeting natural mechanisms (such as the proteins Cytotoxic T Lymphocyte Associated protein 4 (CTLA-4) and Programmed Death Ligand 1 (PD-L1)) to disengage the immune system. Inhibition of these *checkpoints* can promote a more aggressive anti-tumour immune response, and in some patients this leads to long-term remission . However, ICB therapy is not always effective and may have adverse side-effects, so determining which patients will benefit in advance of treatment is vital.

Exome-wide prognostic biomarkers for immunotherapy are now well-established – in particular, Tumour Mutation Burden (TMB) is used to predict response to immunotherapy. TMB is defined as the total number of non-synonymous mutations occurring throughout the tumour exome, and can be thought of as a proxy for how easily a tumour cell can be recognised as foreign by immune cells. However, the cost of measuring TMB using Whole Exome Sequencing (WES) currently prohibits its widespread use as standard-of-care. Sequencing

[1] For their work on ICB, James Allison and Tasuku Honjo received the 2018 Nobel Prize for Physiology/Medicine.

costs, both financial and in terms of the time taken for results to be returned, are especially problematic in situations where high-depth sequencing is required, such as when utilising blood-based Circulating Tumour DNA (ctDNA) from liquid biopsy samples. The same issues are encountered when measuring more recently proposed biomarkers such as Tumour Indel Burden (TIB), which counts the number of frameshift insertion and deletion mutations. There is, therefore, demand for cost-effective approaches to estimate these biomarkers.

In this paper we propose a novel, data-driven method for biomarker estimation, 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. Moreover, since our model allows variant type-dependent mutation rates, it can be adapted easily to predict other biomarkers, such as TIB. Finally, our method may also be used in combination with an existing targeted gene panel; we can estimate the biomarker directly from that panel, or first augment the panel and then construct an estimator.

Due to its emergence as a biomarker for immunotherapy in recent years, a variety of groups have considered methods for estimating TMB. A simple and common way to estimate TMB is via the proportion of mutated codons in a targeted region. Budczies et al. (2019) investigate how the accuracy of predictions made in this way are affected by the size of the targeted region, where mutations are assumed to occur at uniform rate throughout the genome. More recently Yao et al. (2020) modelled mutations as following a negative binomial distribution while allowing for gene-dependent rates, which are inferred by comparing nonsynonymous and synonymous mutation counts. In contrast, our method does not require data including synonymous mutations. Linear regression models have been used for both panel selection (Lyu et al., 2018) and for biomarker prediction (Guo et al., 2020). A review of some of the issues arising when dealing with targeted panel-based predictions of TMB biomarkers is given by Wu et al. (2019). Finally, we are unaware of any methods for estimating TIB from targeted gene panels.

References

J. Budczies, M. Allgäuer, K. Litchfield, E. Rempel, P. Christopoulos, D. Kazdal, V. Endris, M. Thomas, S. Fröhling, S. Peters, C. Swanton, P. Schirmacher, and A. Stenzinger. Optimizing panel-based tumor mutational burden (TMB) measurement. *Annals of On-*

cology: Official Journal of the European Society for Medical Oncology, 30(9):1496–1506, 2019. ISSN 1569-8041. doi: 10.1093/annonc/mdz205.

W. Guo, Y. Fu, L. Jin, K. Song, R. Yu, T. Li, L. Qi, Y. Gu, W. Zhao, and Z. Guo. An Exon Signature to Estimate the Tumor Mutational Burden of Right-sided Colon Cancer Patients. *Journal of Cancer*, 11(4):883–892, 2020. ISSN 1837-9664. doi: 10.7150/jca.34363.

G.-Y. Lyu, Y.-H. Yeh, Y.-C. Yeh, and Y.-C. Wang. Mutation load estimation model as a predictor of the response to cancer immunotherapy. *npj Genomic Medicine*, 3(1):1–9, Apr. 2018. ISSN 2056-7944. doi: 10.1038/s41525-018-0051-x. URL https://www.nature.com/articles/s41525-018-0051-x. Number: 1 Publisher: Nature Publishing Group.

H.-X. Wu, Z.-X. Wang, Q. Zhao, F. Wang, and R.-H. Xu. Designing gene panels for tumor mutational burden estimation: the need to shift from 'correlation' to 'accuracy'. *Journal for Immunotherapy of Cancer*, 7(1):206, 2019. ISSN 2051-1426. doi: 10.1186/s40425-019-0681-2.

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.