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

Jacob R. Bradley, Timothy I. Cannings

January 2021

MAXWELL INSTITUTE FOR MATHEMATICAL SCIENCES

THE UNIVERSITY of EDINBURGH

igmm
INSTITUTE OF GENETICS & MOLECULAR MEDICINE

Abstract

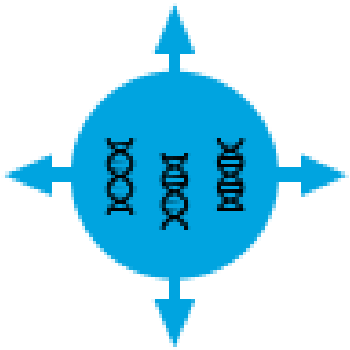
I. Exome-wide biomarkers such as tumour mutation burden (TMB) are useful predictors of response to immunotherapy.
II. While whole-exome sequencing directly measures TMB, its cost prevents it from being standard-of-care.
III. We develop a data-driven framework both for selecting targeted gene panels and for using them to intelligently estimate immunotherapy biomarkers.
IV. To do this, we utilise an exome-wide generative model of mutation, whose structure can be chosen to reflect biological assumptions.

Outline

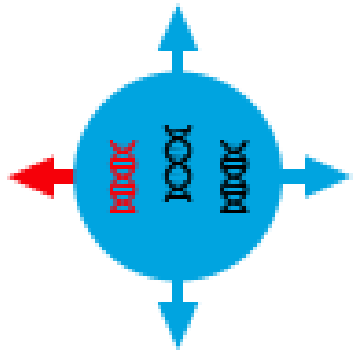
1. Biological/Clinical Background
   - Cancer and immunotherapy
   - Exome-wide biomarkers
   - Targeted gene panels
2. Generative/predictive modelling framework
   - Generative models: what and why?
   - From generative models to predictive models
3. Application to non-small cell lung cancer

Cancer is a disease of the genome

Cancer arises when DNA in cells changes (mutates).



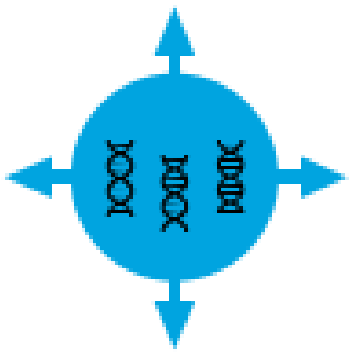
(a) Non-cancer cell: Normal DNA.



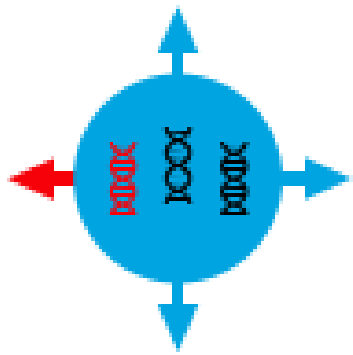
(b) Cancer cell: Mutated DNA.

Immunotherapy enables the immune system

Immunotherapy 'releases the brakes' on the immune system in order for it to attack tumours.



(a) Low-damage cell: **unrecognisable** to immune system.



(b) High-damage cell: **recognisable** to the immune system.

However, the immune system can only attack tumours that it recognises!

Tumour mutation burden stratifies patients for immunotherapy response

As a simple proxy for likelihood of immune response, we can use **tumour mutation burden**: the total count of non-synonymous mutations throughout the tumour exome.

Advantages:

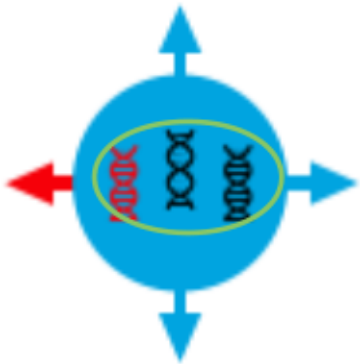
I. Calculated from DNA sequencing only (compatible with liquid biopsy).
II. Relevant across cancer types.

Disadvantages:

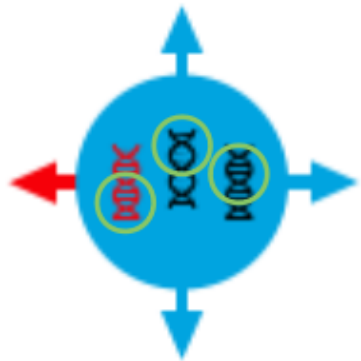
I. Requires Whole-Exome Sequencing (WES) to measure directly.
II. Fails to incorporate other features of the immune environment.

Targeted panels make genomic biomarkers viable

Rather than the entire exome/genome, targeted gene panels sequence a **subset** of genes.



(a) Whole-exome sequencing: **all genes** sequenced.



(b) Targeted panel sequencing: **subset** of genes sequenced.

Problem Statement

Aims:

1. Select a discrete set of genes to form a panel (of pre-specified cost).
2. Estimate TMB from panel sequencing data.

Mutations and biomarkers: some notation

We'll use i, g, and s to refer to:

1. A sample i (ranging from $i = 1$ to $i = N$),
2. A gene g (belonging to some index set G),
3. A variant type s (for example synonymous/non-synonymous).

Finally, we write:

4. M_{igs}, to refer to a mutation count of a particular sample, gene, and mutation type.

We then can define the TMB of the ith sample via:

$$T_i = \sum_{\substack{g \in G \\ s\text{: non-synonymous}}} M_{igs}.$$

What is a generative model of mutation?

A generative model of mutation attempts to capture the **underlying distribution** of variants across the entire exome/genome.

A simple example (Poisson):

$$M_{igs} \sim \text{Poisson}(\lambda_{gs}) \tag{1}$$

Why use generative models of mutation?

1. Utilise known biology to inform the structure of the model.
2. (Sometimes) make interpretable inferences.
3. Generate trust that we've thought about it!

Why use generative models of mutation?

I. Utilise known biology to inform the structure of the model.

Choice of distribution, exome organisation, parameter space.

II. (Sometimes) make interpretable inferences.

III. Generate trust that we've thought about it!

Why use generative models of mutation?

I. Utilise known biology to inform the structure of the model.

II. (Sometimes) make interpretable inferences.

Parameter estimates, post-model inference.

III. Generate trust that we've thought about it!

Generative models to predictive models

Aims:

1. Select a discrete set of genes to form a panel (of pre-specified cost).
2. A mechanism for estimating TMB from panel sequencing data.

Proposed mechanism: Linear estimator.

$$\hat{T} = \sum_{g,s} w_{gs} M_{igs} \tag{2}$$

Optimisation procedure:

$$\hat{w} := \arg\min_{w} \{\mathbb{E}[(\hat{T}_i - T_i)^2] + \lambda |w|\}, \tag{3}$$

where λ is a penalty specifying panel cost, and the expectation $\mathbb{E}$ is over the distribution learned by our generative model.

NSCLC: Poisson-based generative model fit

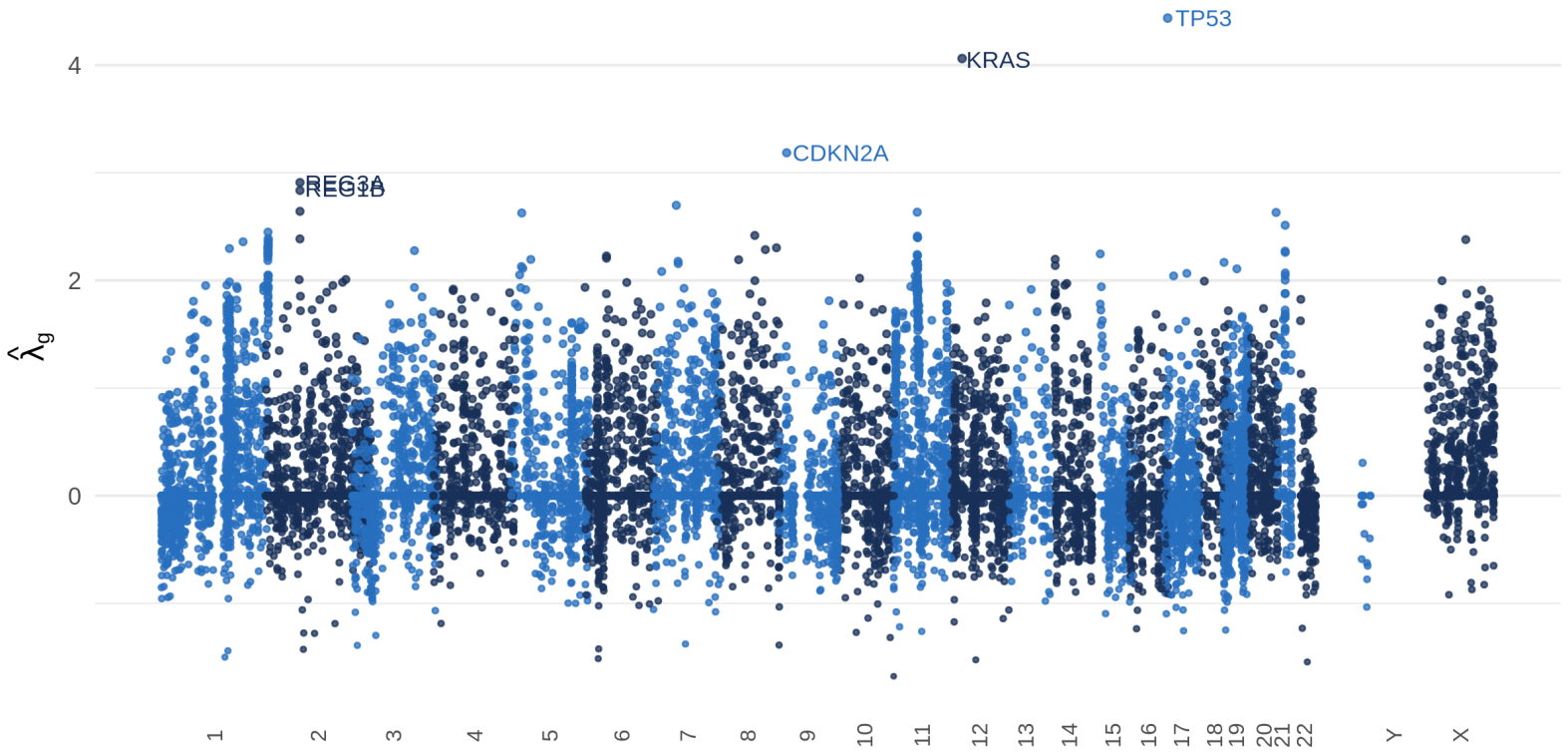


Figure: Generative model parameters.

NSCLC: Poisson-based generative model fit

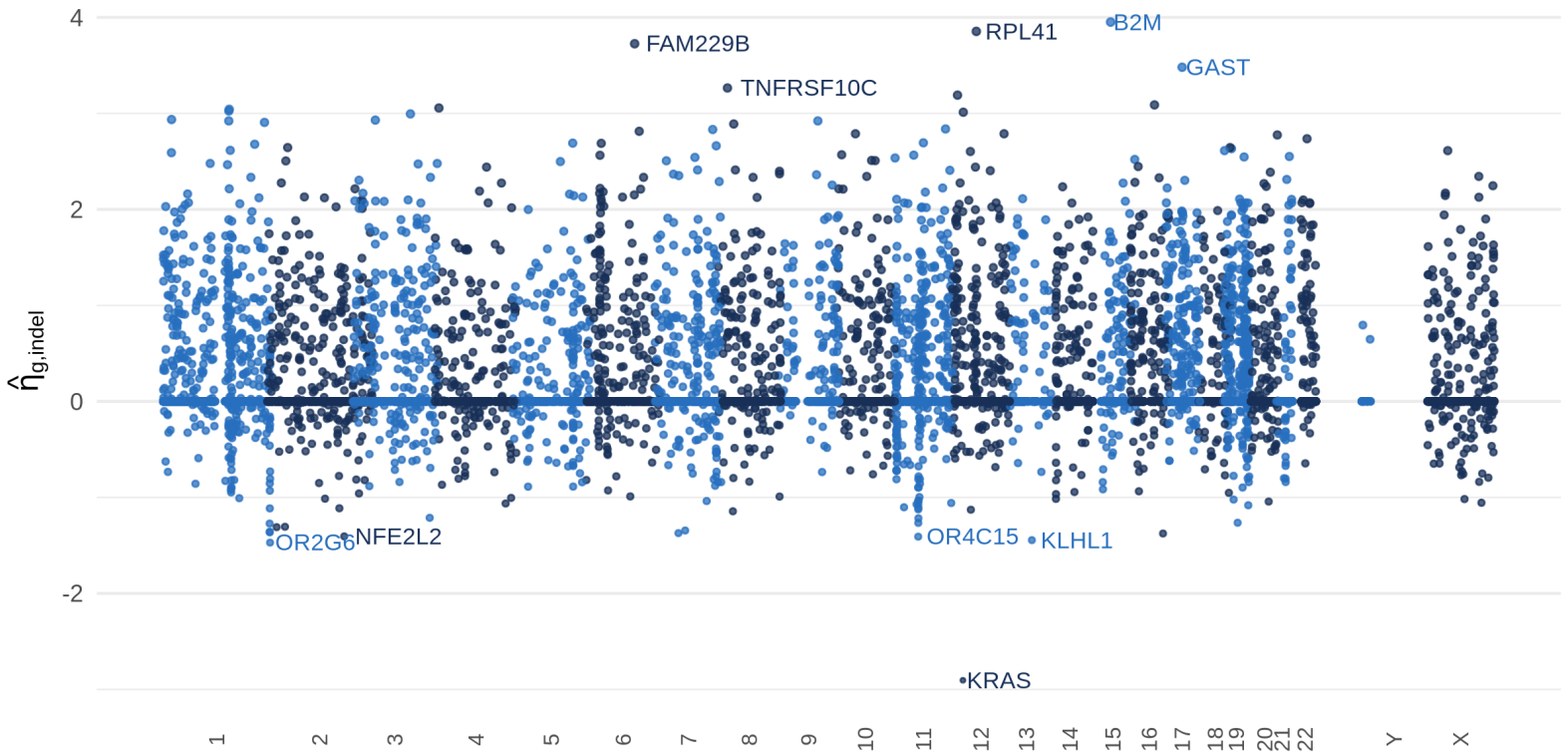


Figure: Generative model parameters (indel-specific).

NSCLC: Estimation performance

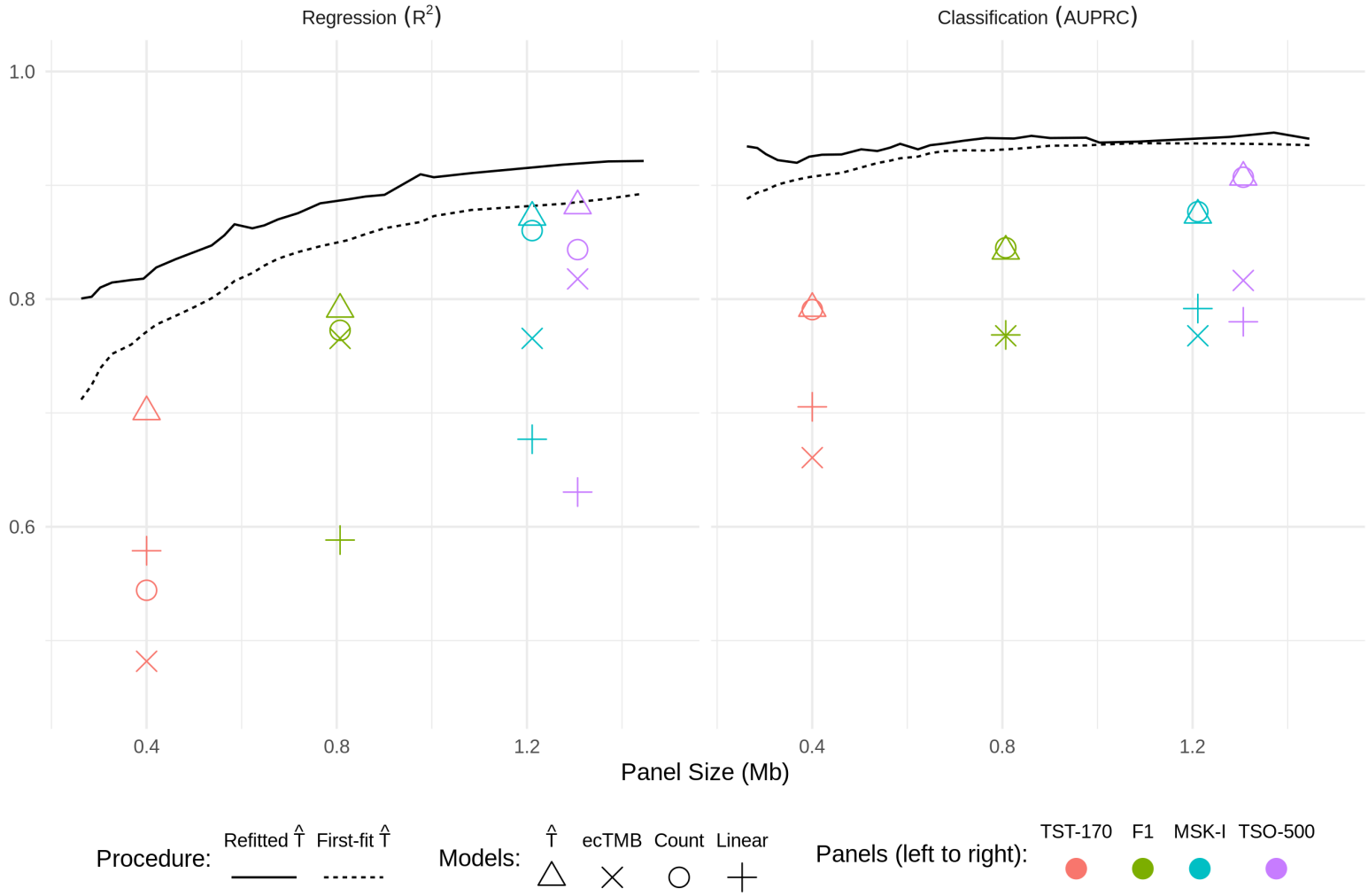


Figure: Predictive model performance: comparison with commercial panels and other predictive methods.