A comprehensive simulation analysis of TMB estimation by targeted oncology panels

Dan Li[1], Binsheng Gong[1], Yifan Zhang[1] , Joshua Xu[1*]

[1] National Center for Toxicological Research, U.S. Food and Drug Administration, Jefferson, Arkansas 72079, United States

Abstract

Tumor mutational burden (TMB), when at a high level, is an emerging indicative factor of sensitivity to immune checkpoint inhibitors. Previous studies showed that, the more affordable and accurate oncopanels can be utilized to measure TMB as a substitute for whole exome sequencing (WES). However, additional processes such as hotspot mutations exclusion and TMB adjustment are usually required to deal with the effect of the limited panel sizes. A comprehensive and quantitative investigation of the effective factors is needed for accurate TMB estimation by oncopanels. In this study, we evaluated the TMB measured by oncopanels based on TCGA-WES annotated mutations. Seven oncopanels plus one union panel were investigated. Then, 10,000 panels with sizes from 0.2 million bases to 3 million bases were simulated and the distribution of the TMB variance were described. We also assessed and compared the panel TMB in some high confidence genomic regions. We demonstrate that the absolute differences between panels and TCGA-WES TMB are roughly consistent along TMB levels. The main factor when measuring TMB using oncopanels is the panel size. The assessment of 10,000 simulated panels indicated that the TMB variance increases dramatically when the panels are under 0.6MB. Quantitatively, we observed that the Root Mean Square Deviation approximately equals to **5*(panel-size-in-MB)^-1/2**. We fixed regression models for each simulated panels. The distribution of the slops and intercepts can be used to assess the performance of panels. This study revealed the quantitative relation between TMB variance and panel size. The large number of simulations can predict the performance of a real-world oncopanel for TMB evaluation.

a

Panel	Size in CDS
AGL	2,635,700
UNI6	1,344,477
BRP	1,043,782
MSK	896,895
QGN	764,916
ILM	406,619
IDT	409,996
IGT	299,005

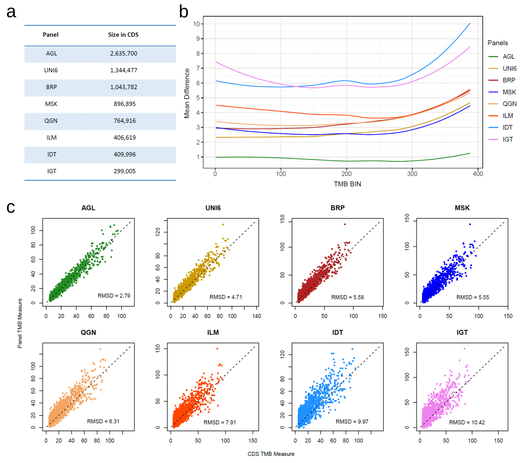


Figure 1. TMB evaluation by oncopanels compared with TCGA CDS region. **a**. Evaluated region sizes of eight oncopanels. **b**. Mean difference of TMBs by each panels and TCGA CDS. Samples with TCGA TMB between 5 and 30 were sorted and grouped (by 10 samples). **c**. Scatter plot of TMBs by panels and TCGA CDS. Samples with TCGA TMB between 5 and 100 were illustrated. RSMD was calculated within TMB bin 5~30.

Simulation of 10,000 Panels

The intersection regions of TCGA CDS and 9,717 COSMIC CGC genes (Tier1: 557, Tier2: 139, Other: 9021) were used to simulate panels. To mimic the real oncopanels, we modified the probability of random selection to make the simulated panels cover more tier1 and tier2 COSMIC genes. In total of 10,000 panels were simulated. The TMBs of the TCGA mutations in each of the panels were calculated and the RMSD was measured.

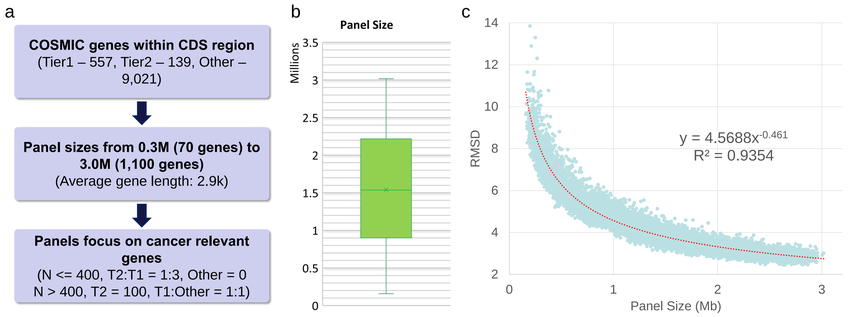
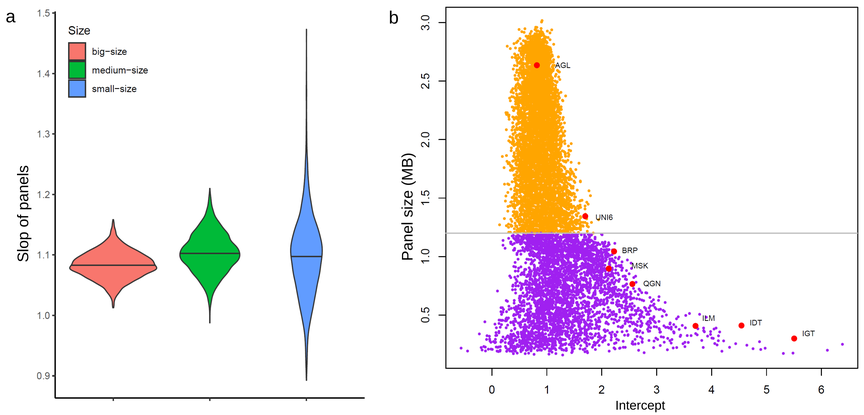


Figure 2. Simulation and TMB measurement of 10,000 panels. a. The procedure of simulating the panels. **b.** The distribution of the panel sizes of the simulated panels. **c.** RMSD evaluation between TCGA and simulated panels. Larger panels tend to have better TMB evaluation compare with TCGA WES.

Performance assessment of individual panels



$$TMB_{Panel.i} = \alpha_i \times TMB_{CDS.i} + \beta_i \quad (i = 1, 2, 3, \ldots, 10000)$$

Figure 3. Assessment of the correlation between TMB evaluation and panel size by regression analysis. a. Distribution of the slops of three group of panels with various sizes. Small group, panel size < 1M; medium group, 1M <= panel size < 2M; big group consisted with panels over 2M. b. The scatter plot of intercepts fitted by regression and panel sizes. The intercepts of the eight oncopanels were added (red). c. Comparison of the MCCs calculated with original and adjust TMBs.

TMB - Tumor mutational burden
Total number of somatic mutations present in a tumor specimen. Measured within a specific genomic region (CDS, targeted region). Usually calculated by per million bases.

RMSD - Root Mean Square Deviation

$$RMSD = sqrt(\frac{\sum_{i=1}^{N}(Oncopanel.TMB_i - TCGA.TMB_i)^2}{N})$$

$$N = \#\ samples\ in\ TMB\ bin\ 5{\sim}30$$

CTR – Consensus Targeted Regions
A high confidence genomic region that covered by multiple WES panels with the low complexity regions were excluded. Our previous study showed that this region could report somatic variant calls with high sensitivity and low false positive rate.

COSMIC CGC Genes
COSMIC stands for Catalogue Of Somatic Mutations In Cancer which is a database holding details on millions of mutations.
The Cancer Gene Census (CGC) is an ongoing effort to catalogue those genes which contain mutations that have been causally implicated in cancer and explain how dysfunction of these genes drives cancer.

Conclusions

By simulating 10,000 oncopanels we assessed the variance between the TMB estimations by whole exon region and the panel regions. The observations indicated that panel size is the main factor that affects the TMB estimation by oncopanels and 1 million bases or over in size could provide reliable results. The relationship between the TMB measure and panel size can be described by a radical equation: 5*(panel-size-in-MB)^-1/2.
Small panels tend to overestimate the TMB values by 1.1 times plus a constant value around 1. Our simulation analysis provides a measure to predict the performance of the TMB estimation by oncopanels using TCGA WES mutation annotation data in clinical practice.

Acknowledgements

This study is part of the SEquencing Quality Control Phase II (SEQC2) project.