Exploiting tumor RNA-sequencing data for prediction of immune checkpoint inhibition response

Frederick De Baene, Carolina Fierro, Emmanuel Rivière, Reindert Van Cauwenberghe, Pieter-Jan Van Dam, Jo Vandesompele, Elise Van Hoof, Pieter Mestdagh
Affiliations: CellCarta

POSTER NUMBER: 5182

CellCarta

INTRODUCTION

The rapid development and widespread adoption of immune checkpoint inhibitors (ICIs) as an immunotherapy, has transformed cancer treatment[1]. However, not all patients respond to ICIs, and it can be challenging to predict who will benefit from such treatment[2-3]. Biomarkers are critical in uncovering ideal candidates for ICI therapy, alternatives, or novel combinations[4].

- Several established features associated with ICI responses include tumor mutational burden (TMB), microsatellite instability (MSI), infiltrating lymphocytes (TILs), and immune gene expression signatures[4-8].
- Accurate clinical response prediction requires assessment of said features, however, the complexity of extracting and combining information from each can be time-consuming and costly, as they are typically measured using different omics techniques (Figure 1)[9-12].
- Moreover, individual evaluation of the features does not provide as strong of a therapy response prediction as would a combined evaluation. To address such challenges, we have developed a suite of data processing and analysis pipelines to extract the aforementioned features from the RNA sequencing (RNA-Seq) profile of tumor samples, thus, reducing the amount of omics data needed, and integrating it into a single model that improves prediction accuracy.

Figure 1. Summary of established features associated with ICI responses and the different techniques used to measure them. WES: whole exome sequencing, WGS: whole genome sequencing, RNA-Seq: RNA sequencing, qPCR: quantitative PCR, IHC: immunohistochemistry, TMB: tumor mutational burden, MSI: microsatellite instability, TIL: infiltrating lymphocytes.

METHODOLOGY

Technical Comparison of the RNA-Seq Bio-IT Models to Gold Standards

eTMB

We established a variant calling pipeline for tumor RNA-Seq data, whereby machine learning (ML) models were trained to identify somatic non-synonymous coding mutations from RNA-Seq data (Figure 2). The algorithm was trained using samples for varying tumor entities, where both RNA-Seq and WES data were available for the tumor and matching germline from The Cancer Genome Atlas (TGCA) database. TMB and eTMB values were assessed for multiple indications including lung adenocarcinoma, head and neck squamous cell carcinoma, and stomach adenocarcinoma (STAD). TMB values were determined by WES of tumor and germline samples, while eTMB values were determined based only on matching RNA-Seq from the same tumor samples (Figure 3). We demonstrated a high correlation between TMB and eTMB measurements in varying cancer types—STAD shown as an example (Figure 3). While both TMB (exome mutations) and eTMB (expressed exome mutations) are mutation measurements, a perfect correlation (r value of 1) should not be expected, as eTMB only considers expressed mutations. Since neoantigens are more directly involved in the immune recognition of cancer cells, eTMB may provide a more accurate measure of the immunogenicity of the tumor than TMB[13,14].

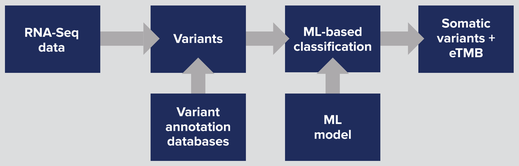

Figure 2. Dedicated variant calling pipeline with ML classifier to define somatic mutations and eTMB. Variant calling on RNA-Seq comes with several challenges including highly variable coverage (driven by expression abundance), variants resulting from RNA-editing and variants resulting from the random priming of RNA. Extensive variant annotation obtained from the variant calling and from different databases (i.e., VEP, dbSNP, gnomAD) was used to train an ML algorithm that selects true somatic mutations. ML: machine learning.

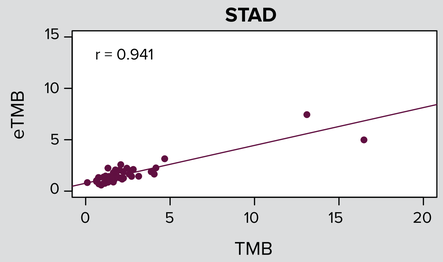

Figure 3. TMB and eTMB are correlated in STAD. TMB measured using WES of tumor and germline samples, and eTMB measured only from matching RNA-Seq of the same tumor samples for STAD (r = 0.941). R values determined via Pearson correlation.

MSI

Microsatellites are short DNA sequences composed of a central core of repeating units and non-repeating peripheral flanks[15]. DNA sequencing, or fragment length analysis of PCR products, is normally performed to assess the variation of these repeating units and thus determination of MSI status in a tumor sample[15]. We extracted eight different types of mononucleotide repeats from RNA-Seq reads and compared RNA repeat length to reference genome repeat length (Figure 4). By calculating the mean difference in repeat length between RNA and reference genome, we quantitatively measured microsatellite instability. Following, an ML algorithm was trained on the variation in repeat length for eight different repeat types to determine MSI status. We used TCGA RNA-Seq data from cancer types with a high frequency of microsatellite instability (STAD, COAD (colon adenocarcinoma), UCEC (uterine corpus endometrial carcinoma)) to assess the accuracy of RNA-Seq MSI predictions. For each of these tumor samples, the MSI status was determined based on the algorithm and defined in parallel using the gold standard pentaplex PCR method (data available in TCGA Program). MSI predictions using the RNA-Seq model were highly accurate in STAD and UCEC samples, and perfectly accurate for COAD samples (Figure 5).

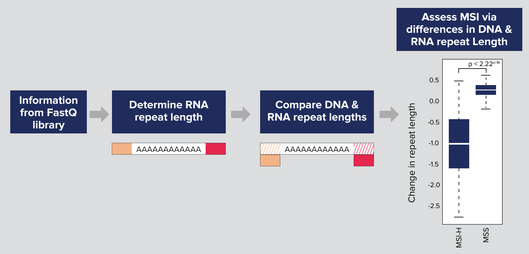

Figure 4. Defining MSI through repeat analysis of RNA-Seq reads. Mononucleotide repeats were extracted from RNA-Seq data, and repeat lengths were compared between the RNA and the reference genome. Differences in repeat length between RNA and DNA are significantly altered in MSI-H versus MSS samples (Mann-Whitney test, p < 2.22e-16). These repeat length differences are used to train an ML algorithm to identify MSI-H and MSS. MSI-H: microsatellite instability-high, MSS: microsatellite stable.

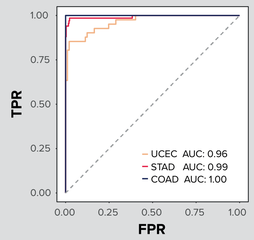

Figure 5. High accuracy of MSI status prediction from RNA-Seq data. Gold standard MSI status determined by PCR of tumor sample, and RNA-Seq MSI status determined from trained ML algorithm on difference in repeat length of the same tumor sample. Prediction performance is demonstrated by a ROC curve and area under the ROC curve. TPR: true positive rate, FPR: false positive rate, UCEC: uterine corpus endometrial carcinoma, COAD: colon adenocarcinoma, STAD: stomach adenocarcinoma, AUC: area under the curve.

RESULTS

Clinical Predictions from the RNA-Seq Bio-IT Model

eTMB and MSI

To further assess the clinical validity of the pipeline, eTMB and MSI scores were evaluated for gastric cancer patients utilizing data from a recent publication by Kim et al.[16]. The eTMB values were correlated to different patient response groups (Figure 6A and B). Patients were subsequently grouped into eTMB-high or eTMB-low categories, and the fraction of responders to treatment in each group was evaluated (Figure 6C). MSI status was similarly assessed; patients were placed into MSI-high (MSI-H) or microsatellite stable (MSS) groups (Figure 7).

Significant differences were found in both eTMB and MSI scores in responders compared to non-responders (Figures 6 and 7). Patients with eTMB-H and MSI-H tumors responded more effectively to treatment than their MSS and eTMB-L tumor counterparts. Ultimately, as determined by RNA-Seq data, both biomarkers were significantly associated to clinical therapy outcomes.

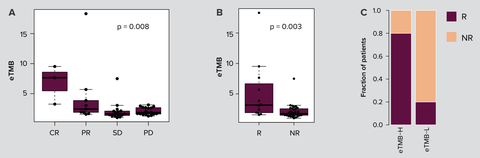

Figure 6. eTMB scores in gastric cancer samples demonstrate significant differences between patients grouped by A) best of response (Kruskall-Wallis, p = 0.008) and B) response (Mann-Whitney, p = 0.003). C) Higher fraction of responders in eTMB-H versus eTMB-L samples. CR: complete response, PR: partial response, S: stable disease, PD: progressive disease, R: responders (CR+PR), NR: non-responders (SD + PD), eTMB-H: eTMB-high, eTMB-L: eTMB-low.

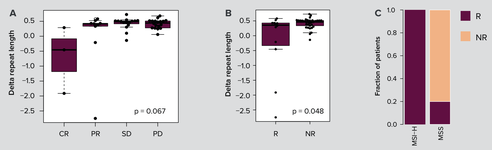

Figure 7. MSI scores in gastric cancer samples demonstrate non-significant difference between patients grouped by A) best of response (Kruskall-Wallis, p = 0.067) and a significant difference between patients grouped by B) response (Mann-Whitney, p = 0.048). C) Higher fraction of responders in MSI-H versus MSS samples. CR: complete response, PR: partial response, SD: stable disease, PD: progressive disease, R: responders (CR+PR), NR: non-responders (SD + PD), MSI-H: microsatellite instability-high, MSS: microsatellite stable.

TILs and immune gene signatures

We applied four different computational deconvolution algorithms to quantify TILs. Deconvolution algorithms enable the enumeration of individual cell types directly from a bulk RNA-Seq profile[17]. The output was integrated into a single score per cell type (abundance score). In addition, we assessed various inflammation gene expression signatures that were obtained from literature. Each signature was represented as a single score reflecting the overall abundance of the genes in the signature. Abundance scores for TILs (Figure 8) and immune gene expression signatures (Figure 9) were compared between responders and non-responders, revealing different cell types and signatures with significantly higher or lower abundance in responders compared to non-responders.

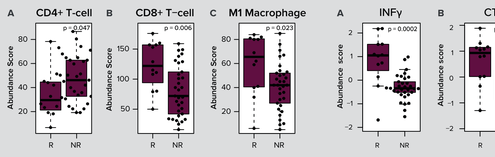

Figure 8. Significant differences in response scores for tumor infiltrating lymphocytes between responders versus non-responders (Mann-Whitney) A) CD4+ T-cell (p = 0.047), B) CD8+ T-cell (p = 0.006), and C) M1 macrophage (p = 0.023). R: responders, NR: non-responders.

Figure 9. Significant differences in response scores for immune gene signatures between responders and non-responders (Mann-Whitney) A) IFNγ (p = 0.0002) and B) CTL (p = 0.007). IFNγ: interferon gamma, CTL: cytotoxic T cell. R: responders, NR: non-responders.

Improved Clinical Predictions: Combining All 4 Features Into a Single Model

By integrating eTMB, MSI, TILs abundance scores, as well as inflammation gene signatures into a single logistic regression model, we were able to further improve response prediction for ICI compared to predictions based on individual features (Figure 10).

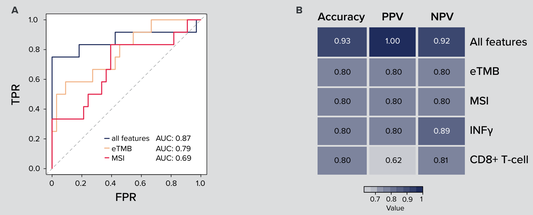

Figure 10. Improved response prediction when integrating all features compared to individual features. A) Logistic regression model incorporating all features (eTMB, MSI, CD8+ T-cell, M1 macrophage, IFNγ signature, and CTL signature; AUC = 0.87) vs. eTMB (AUC = 0.79) and MSI (AUC = 0.69). Higher AUC value for all features compared to eTMB or MSI scoring alone. Prediction performance is demonstrated by a ROC curve and area under the ROC curve (AUC). B) Performance metrics for each model. AUC: area under the ROC curve, TPR: true positive rate, FPR: false positive rate, PPV: positive predictive value, NPV: negative predictive value.

CONCLUSIONS

- The BioIT pipeline has been designed to use RNA-Seq and can be used to interpret gene expression using ideas of TMB, general gene expression, microsatellite instability and immune cell deconvolution analysis.
- A multiomics approach is typically required to assess treatment response, however, the developed RNA-Seq Bio-IT model can be utilized to evaluate all four key established features more efficiently and, through integration, more accurately predict responses to ICI therapy.

References

1. Immune Checkpoint Inhibitors - NCI. https://www.cancer.gov/about-cancer/treatment/types/immunotherapy/checkpoint-inhibitors.
2. McLoughlin, K. C. et al. Discov. Med. 26, 85-92 (2018).
3. Dall'Olio, F. G. et al. Nat. Rev. Clin. Oncol. 19, 75-90 (2021).
4. Bai, R. et al. Biomarkers Research 8 (2020).
5. Newman, J. et al. Cancer Med. J. 3, 17-21 (2020).
6. Eso, Y. et al. J. Gastroenterol 55, 15 (2020).
7. Linette, G. P. et al. Curr. Hematol. Malig. Rep. 14, 286 (2019).
8. Mallik, S. et al BMC Syst. Biol. 12, 21-29 (2018).
9. Park, S. et al. J. Clin. Oncol. 40, 1916-1928 (2022).
10. Eso, Y. et al. J. Gastroenterol 55, 15-26 (2019).
11. Endris, V. et al. Int. J. Cancer 144, 2303-2312 (2019).
12. Meléndez, B. et al. Lung Cancer Res. 7, 661 (2018).
13. Wang, Z. et al. Nat. Rev. Genet. 10, 57-63 (2009).
14. Szeto, C. et al. J. Clin. Oncol. 38, (2020).
15. Li, K. et al. J. Clin. Oncol. Cancer Cell Int. 20, 1-13 (2020).
16. Kim, S. T. et al. Nature Medicine 24, 1449-1458 (2018).
17. Zhang, Y. et al. Comput. Struct. Biotechnol. J. 21, 176-184 (2023).

Please send inquiries to
nathalie.bernard@cellcarta.com
and pieter.mestdagh@cellcarta.com
with **RNAseq BioIT** in email subject.