Quantifying T- and B-cell immune receptor distribution diversity to uncover their clinical relevance in clear cell renal cell carcinoma

AUTHORS

Meghan C. Ferrall-Fairbanks, PhD[1,2], Nicholas Chakiryan, MD[3], Boris I. Chobrutskiy[4], Youngchul Kim, PhD[5], Jamie K. Teer, PhD[5], Anders Berglund, PhD[5], James Mulé, PhD[6], Michelle Fournier[3], Erin Siegel, PhD[7], Shayan Falasiri[4], Juan F. Arturo[4] Esther N. Katende[3], George Blanck, PhD[4,6], Brandon J. Manley, MD[3, *] Philipp M. Altrock, PhD[1,8,9, *]

AFFILIATIONS

[1] Department of Integrated Mathematical Oncology, Moffitt Cancer Center, Tampa, FL 33612, USA
[2] J. Crayton Pruitt Family Department of Biomedical Engineering, University of Florida, Gainesville, FL 32611, USA
[3] Department of Genitourinary Oncology, H Lee Moffitt Cancer Center, Tampa, FL, 33612, USA
[4] Department of Molecular Medicine, Morsani College of Medicine, University of South Florida, Tampa, FL, 33612, USA
[5] Department of Biostatistics and Bioinformatics, H Lee Moffitt Cancer Center, Tampa, FL, 33612, USA
[6] Department of Immuno Oncology, H Lee Moffitt Cancer Center, Tampa, FL, 33612, USA
[7] Department of Cancer Epidemiology, H Lee Moffitt Cancer Center, Tampa, FL, 33612, USA
[8] Department of Malignant Hematology, Moffitt Cancer Center, Tampa, FL 33612, USA
[9] Department of Blood and Marrow Transplant and Cellular Immunotherapy, Moffitt Cancer Center, Tampa, FL 33612, USA

KEYWORDS

clear cell renal cell carcinoma, immune receptor heterogeneity, immune infiltration, V(D)J recombination, T-cell receptor CDR3 diversity

ABBREVIATIONS

CDR3 = complementarity determining region-3
GDI = generalized diversity index
IP = inflection point
KM = Kaplan-Meier survival curve
OS = overall survival
RFS = recurrence free survival
TCR = T-cell receptor
TRA = T-cell receptor alpha
TRB = T-cell receptor beta
TRG = T-cell receptor gamma
TRD = T-cell receptor delta
BCR = B-cell receptor
IGH = B-cell immunoglobulin heavy chain
IGK = B-cell immunoglobulin kappa light chain
IGL = B-cell immunoglobulin lambda light chain

ABSTRACT

Immune-modulating systemic therapies are often used to treat metastatic clear cell renal cell carcinoma (ccRCC). Used alone, sequence-based biomarkers neither accurately capture patient dynamics nor the tumor-immune microenvironment. To better understand the *tumor ecology* of this immune microenvironment, we used complementarity determining region-3 (CDR3) sequence recovery counts of the tumor infiltrating lymphocytes, and quantified tumor infiltration by sequences recovered from patient tumors by applying a generalized diversity index (GDI) to CDR3 sequence distributions across two distinct ccRCC cohorts. GDI can be understood as a curve over a continuum of diversity scales and allows sensitive characterization of distributions to capture richness, evenness, and subsampling uncertainty, along with other important metrics. For example, richness quantifies the total unique sequence count, while evenness quantifies similarities across sequence frequencies. We observed significant differences in receptor sequence diversity across gender and race. Further, our analysis revealed that patients with larger and more clinically aggressive tumors had increased richness of recovered tumoral CDR3 sequences, specifically in those from T-cell receptor alpha and B-cell immunoglobulin lambda light chain. We identified a novel and robust measure of distribution evenness, using GDI's inflection point (IP). High IP values associated with improved overall survival, suggesting that normal-like sequence distributions lead to better outcomes. Our results propose a new quantitative tool that can be used to better characterize patient-level differences related to immune cell infiltration and, can be used to identify unique characteristics of tumor-infiltrating lymphocyte heterogeneity in ccRCC and other malignancies.

INTRODUCTION

Renal cell carcinoma ranks seventh and tenth among the most commonly diagnosed cancers among men and women in the US, respectively, accounting for 3.8% of all cancer cases and 2.5% of all cancer deaths (*1*). The most common type of renal cell carcinoma is clear cell renal cell carcinoma (ccRCC). Historically, metastatic ccRCC has been one of the first malignancies successfully treated with immune-modulating systemic therapy, using interleukin-2 and interferon-α (*2*). Immune checkpoint inhibitors (ICI) such as nivolumab, ipilimumab, pembrolizumab, and avelumab, have emerged as the first line therapy for metastatic ccRCC, typically administered in combination with each other or with a targeted therapeutic agent (*3*). The arrival of immune checkpoint inhibitors has precipitated a tremendous research effort aiming to accurately characterize the tumor-immune microenvironment and explore potential biomarkers to predict ICI response, for which robust markers have been largely elusive. The vast majority of investigations have focused on tumor-centric variables including somatic mutations and gene expression. Fewer studies have been focused on host factors that contribute to the microenvironment or focused on how differences among these factors may affect clinical or therapeutic outcomes.

Across tumor types, response to ICI has been correlated with higher frequencies of somatic mutations that are believed to give rise to tumor-specific neoantigens, and to stimulate a robust antitumor immune response (*4-6*). In contrast, analyses of renal cell carcinomas have demonstrated a relatively low frequency of somatic mutations, yet very high levels of immune cell infiltration. These findings suggest that a high mutational burden is not solely responsible for inducing immune infiltration in ccRCC (*7-11*).

Additionally, recent work has demonstrated that CD8+ T cell infiltration alone does not predict response to ICI, suggesting that refinements in characterizing immune cell populations are needed to understand the microenvironment and the biology underlying ICI response (*12*).

To initiate an anti-tumor immune response, tumor-specific neoantigens first require recognition by a T- or B-cell receptor (TCR, BCR) on a tumor infiltrating lymphocyte. The tumor infiltrating lymphocyte, complementarity determining region-3 (CDR3) is a highly variable region in the TCR/BCR that provides a complementary binding surface for antigens and largely determines the antigen specificity of the receptor. Investigations have shown the promise of CDR3 features as prognostic biomarkers for several malignancies (*13-16*), using sequencing and bioinformatic pipelines to recover single reads that represent the CDR3 amino acid sequence. These reads can be quantified as total recovered reads, potentially as a primary metric of immune infiltration (*14-16*). However, total count-based measures of CDR3 variability are unlikely to reflect the underlying complex biology of host adaptive immune response with the same accuracy as other measures of receptor diversity (*17-19*).

Here, we leverage properties of a generalized diversity index (GDI) (*20-22*), a measure applied in ecology and evolution, to quantify CDR3 diversity, and assess whether this diversity is associated with important clinicopathologic outcomes in ccRCC. GDI is evaluated as a continuous function along a range of order of diversity (q) values (*20-22*). At low values of q (low-q GDI), the index is a measure of distribution richness, i.e. the count of distinct types, sequences or clones, while the value at q=1 is related to Shannon's diversity index (*23*). At high values of q (high-q GDI), the index approaches a

measure of evenness or dominance, i.e., focusing on the dominant clone or sequence and its frequency. Here, we applied these diversity metrics to ccRCC samples, and assessed the properties of GDI in ccRCC tumors and which may serve as possible prognostic markers. We assessed tumor infiltrating lymphocyte TCR and BCR CDR3 diversity across the entire range of q, and for isolated values that have direct statistical interpretations. We analyzed two independent cohorts of ccRCC patients with bulk RNA-sequencing samples; the Moffitt Total Cancer Care (TCC) cohort (*24*), and the Clinical Proteomic Tumor Analysis Consortium 3 (CPTAC-3) cohort (*25*).

RESULTS

General diversity index (GDI) quantifies tumor infiltrating lymphocyte receptor subtype diversity in the TCC cohort

For each patient, we measured individual receptor CDR3 diversities across the 7 human adaptive immune receptor genes (TRA, TRB, TRG, TRD, IGH, IGK, IGL), as well as common combinations of these receptor subtypes (TRA+TRB, TRG+TRD, IGH+IGK+IGL, along with all 7 together, denoted at TRs+IGs). In the TCC cohort (n = 105), CDR3 sequences were recovered from bulk RNA-sequencing of patient tumor tissue. GDI was then calculated for each subtype and group of subtypes (**Fig. 1A-B** and **Supplemental Figs. 1-2**). The Moffitt TCC cohort of ccRCC patients represented a cohort of clinically high-risk and advanced patients. Over two-thirds of the cohort contains patients with pathologic stages 3 or 4, including 6% of patients with highly aggressive sarcomatoid histology (**Table 1**). To quantify the GDI, we generated a continuum of

diversity measures D(q) for each patient across values of the order of diversity, q. Then, we were able to compare clinical variables at specific point estimates of the continuum of diversity measures, as shown in **Fig. 1C**. We compared immune receptor subtype diversity across patients, and found that the point estimates ΔD diversity, high-q diversity, and inflection point (IP) of the GDI curve (see Methods) were unique summary metrics. The value of ΔD summarizes richness (total number of unique sequences) of receptor subtypes in a patient sample. High-q diversity focuses on the dominance (frequency of largest sequence) of a receptor subtype and de-emphasizes a rare receptor subtype. IP is a measure of receptor subtype evenness; with high IP values indicating an overall more level distribution, largely independent of receptor subtype richness (*26-28*).

Immune receptor subtype richness is associated with important pathological features in the TCC cohort

Across individual receptor subtypes, TRA and IGL receptor diversity consistently showed increased richness (in **Fig. 2** exemplified with ΔD diversity comparisons) in tumors with larger diameters, higher grade, sarcomatoid status, and tumors from the left side. TRA receptor diversity split the Moffitt TCC cohort at the median of ΔD diversity. Of these, patients with ΔD values above the cohort median had a mean largest diameter size of 6.1 cm, compared to those with below the median who had a mean largest diameter size of 7.6 cm (p-value 0.0287, **Fig. 2A i**). This same trend was reflected in IGL receptor diversity with the high ΔD diversity group (p-value 0.0195, **Fig. 2B i**).

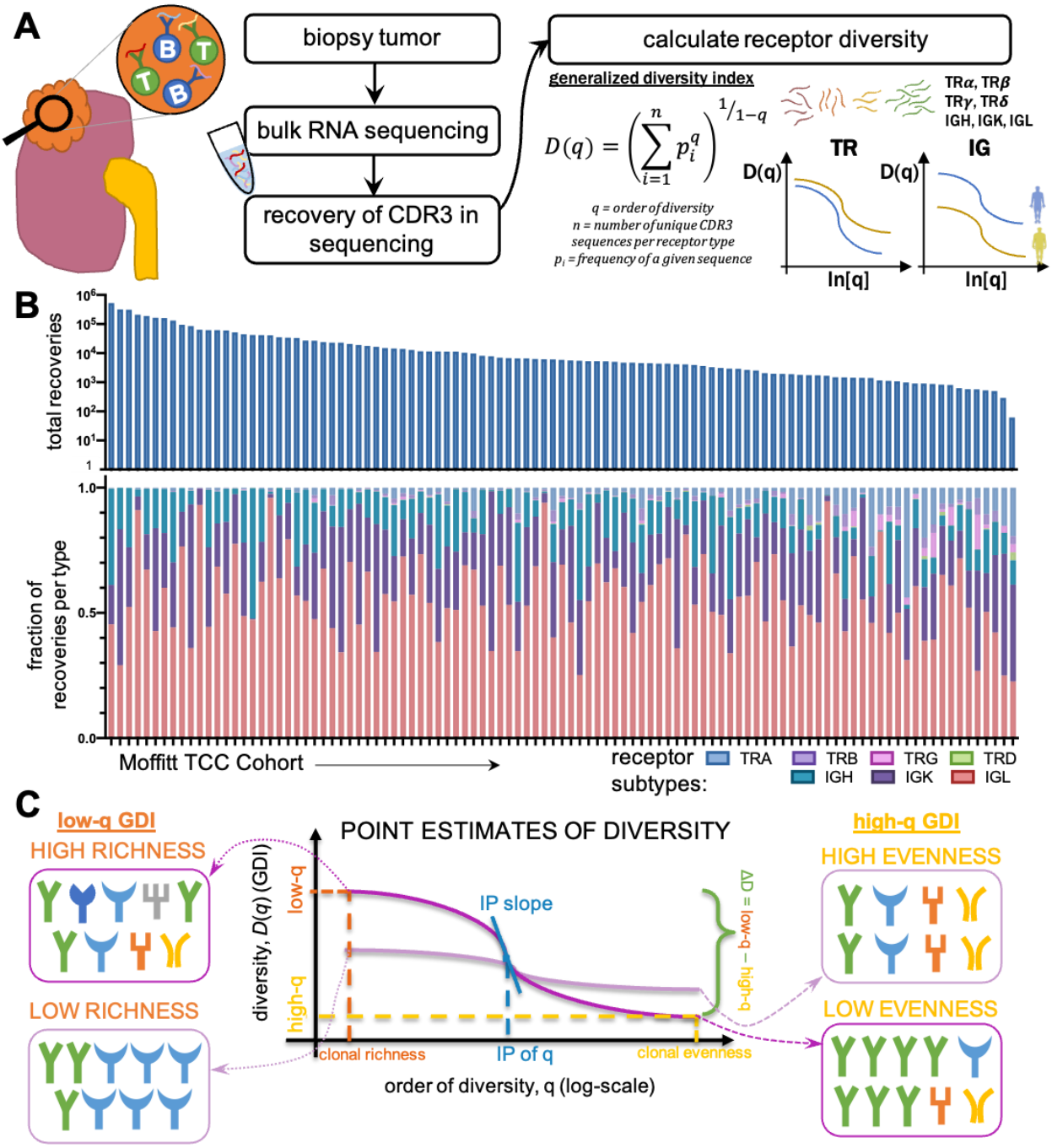


Figure 1: Tumor infiltrating lymphocyte receptor diversity as a marker in clear cell renal cell carcinoma. (A) Overall workflow schematic of calculating tumor infiltrating lymphocyte (TIL) diversity across a cohort of patients. Patient tumors undergo bulk RNA sequencing and then CDR3 sequences from TCR and BCR receptors are recovered. Then for each patient, CDR3 sequences are segregated by receptor class (TRA, TRB, TRG, TRD, IGH, IGK, and IGL) and then patient frequencies across the CDR3 landscape per receptor are calculated and used to quantify the individual patient's receptor diversity using the generalized diversity index from ecology. (B) Receptor recovery distributions across the seven major receptors types in the Moffitt cohort consented under the Total Cancer Care (TCC) protocol. (C) Patient diversity curves can be distilled down to five point-estimates of diversity: low q (q=0.02), high q (q=100), ΔD (D(0.01)–D(100)), inflection point (IP), and inflection point slope.

TRA receptor ΔD diversity in CDR3 amino acid sequences recovered from tumors with left laterality had an average diversity score 2.3-fold higher compared to those with right laterality tumors (**Fig. 2A ii**, p-value 0.0097), which was also reflected in IGL receptor ΔD diversity (**Fig. 2B ii**, p-value 0.0445). In addition, patients with high tumor grade had increased TRA receptor ΔD diversity (**Fig. 2A iii**, p-value 0.0227), which was also demonstrated in IGL receptor ΔD diversity (**Fig. 2B iii**, p-value 0.0459).

Overall, Moffitt TCC cohort tumors that were identified with sarcomatoid histology had increased overall lymphocyte receptor diversity compared to those individuals who did not have sarcomatoid histology (demonstrated in **Fig. 2A & B iv**; p-value 0.0430 in TRA and p-value 0.0152 in IGL). This trend for increased diversity in sarcomatoid carcinoma tumors was statistically significant in all combinations except for TRG receptor diversity, which was one of the rarest CDR3 receptor subtypes recovered.

Our observations of increased lymphocyte receptor richness in larger diameter tumors, higher grade tumors, left laterality tumors, and sarcomatoid carcinomas were also discovered in other receptor subtypes, as well as in the combinations (**Supplemental Fig. 3-4** and **Supplemental File 1** for all comparisons).

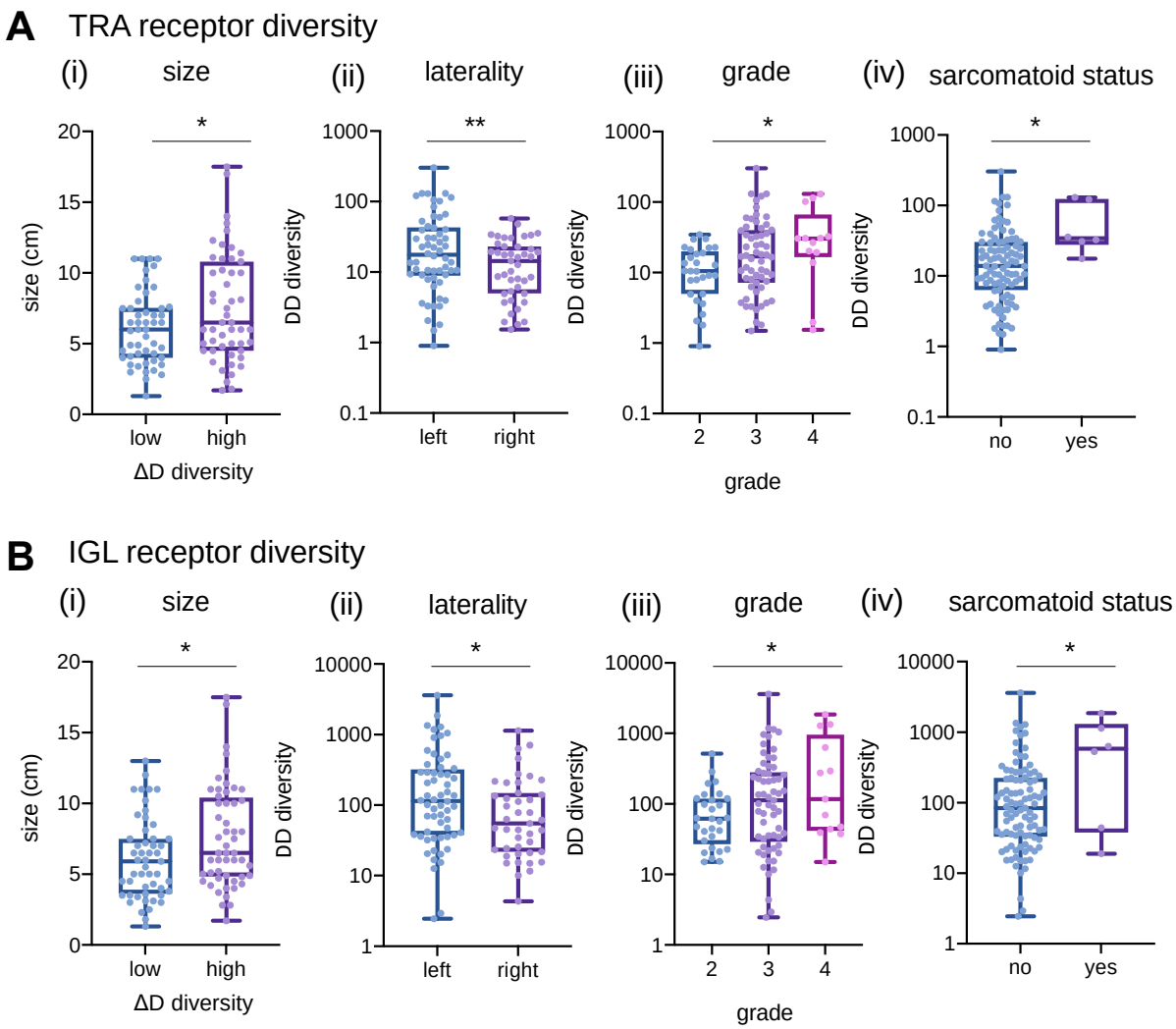


Figure 2: Patients with tumors that are larger in diameter, higher grade, left laterality and sarcomatoid status have increased diversity in TRA and IGL receptors. (A) TRA receptor CDR3 sequence ΔD diversity across the Moffitt TCC Cohort have increased diversity in (i) larger diameter tumors (low diversity had mean diameter of 6.1 cm and high diversity had a mean diameter for 7.6cm; p-value: 0.0287), (ii) left laterality tumors (score of 37.55 vs 16.39; p-value: 0.0097), (iii) with higher grade tumors (mean score from grade 2 was 12.96, mean score from grade 3 was 33.09, and mean score from grade 4 was 42.62; p-value: 0.0227), and (v) in patients with sarcomatoid status (sarcomatoid status evaluated as yes (at least 5%) or no, in TRA no had a mean ΔD diversity score of 26.47 vs yes with a mean score of 61.32; p-value 0.0430). (B) IGL receptor CDR3 sequence ΔD diversity showed the same trends as TRA receptor CDR3 sequence diversity for (i) size, (ii) laterality (score of 331.1 vs 141.3; p-value: 0.0445), (iii) grade (mean score from grade 2 was 93.99, mean score from grade 3 was 281.4, and mean score from grade 4 was 465.5; p-value: 0.0459), and (iv) sarcomatoid status (no had a mean score of 223.9 vs yes with a mean score of 704.4; p-value: 0.0152). Unpaired t-tests were used to compare two group data and ANOVA was used to compare grade, three group data. p-value significance represented by * < 0.05, ** < 0.01, *** < 0.001

Independent validation of GDI metrics

To validate the findings of increased diversity with poor pathological features we first calculated xCell scores (*29*) for patients in the Moffitt TCC cohort to confirm that the detected recoveries came from tumor-infiltrating lymphocytes. TRA receptor total and unique recoveries had the highest Spearman correlation scores with T-cell subtype xCell score and were less correlated with the B-cell subtype xCell score (**Fig. 3A** and **Supplemental Fig. 5-7**). Furthermore, the Immune Score strongly correlated positively with total and unique recoveries, compared to the Microenvironment Score and Stroma Score, which both showed weaker Spearman correlation coefficients.

Once we confirmed that the diversity scores we measured were attributable to the immune cell infiltration in the tumor with the xCell scores, we sought to independently validate our findings with a replicative study, using the ccRCC CPTAC-3 cohort (n=110). The CDR3 recovery landscape of CPTAC-3 and Moffitt TCC differed slightly. CPTAC-3 had more recoveries from T-cells; B-cells accounted for an average of only 78.96%. However, we did not identify any significant trend of differences between the proportion of T-cell and B-cell recovered based on stage (**Supplemental Figs. 8-9**). Laterality and sarcomatoid status could not be evaluated, as these are not available in CPTAC-3.

We confirmed the observation that higher grade and larger size of tumors are associated with increases in TRA and IGL receptor ΔD diversity in the CPTAC-3 cohort (**Fig. 3B** compared to **Supplemental Fig. 10,** and **Supplemental File 2** for all comparisons)—which independently validated that rich, but rather uneven receptor subtype distribution is associated with larger and high-grade tumors. TRA receptor ΔD

diversity was significantly different between low and high grade (p-value: 0.0212, **Fig. 3B i**) and pT stages (p-value: 0.0368, **Fig. 3B iii**). In both TRA and IGL receptor distributions, the ΔD diversity score showed that large tumors (described pathologically as tumors with the largest diameter 7 cm or greater) have increased receptor subtype diversity compared to smaller tumors (those with largest diameters below 7 cm). Of note, CPTAC-3 has significantly different distribution of tumor sizes, with many small tumors (largest diameter < 7cm), compared to the Moffitt TCC cohort (**Fig. 3B ii** and **v**, **Supplemental Fig. 11**).

In addition to tumor sample differences, CPTAC-3 had a subset of 75 patients with matched normal tissue samples (matched normal tissue recovery landscape described in **Supplemental Fig. 12**). Lymphocyte receptor richness was increased in tumor samples compared to normal tissue, which was observed across all receptor subtypes and combinations (**Fig. 4A**, **Supplemental Fig. 13**, and **Supplemental File 2**). In the TRs+IGs combination, tumor tissues had, on average, at least 2.6-fold increase in richness of CDR3 sequences recovered, compared to the matched patient's normal tissue (mean score of 144.0 in normal tissue compared to a mean score of 377.1 in matched tumor). Furthermore, sequence dominance (measured by low values of high-q diversity) was decreased in all receptor subtypes with the exception of the IGL receptor and IGH+IGK+IGL receptor combination (**Fig. 4B**, **Supplemental Fig. 13**, and **Supplemental File 2**). In the TRs+IGs combination, the tumor tissue had, on average, at least 1.3-fold increase in high-q diversity compared to normal tissue, which indicates that the most abundance CDR3 sequence in the sample was about 2% lower, thus less dominant, in the tumor (mean score of 14.27 in the normal tissue compared to a mean score of 19.17 in the tumor tissue). Analyzing the inflection point q metric of evenness in

TRs+IGs combinations in the normal-tumor matched patients showed that normal samples have an almost 10% mean increase evenness compared to their matched tumor samples (**Fig. 4C**; p-value 0.0857). This data supported the hypothesis that normal tissues are expected to show very even CDR3 sequence distributions, and that better outcomes are to be expected in tumors that appear more normal in this context.

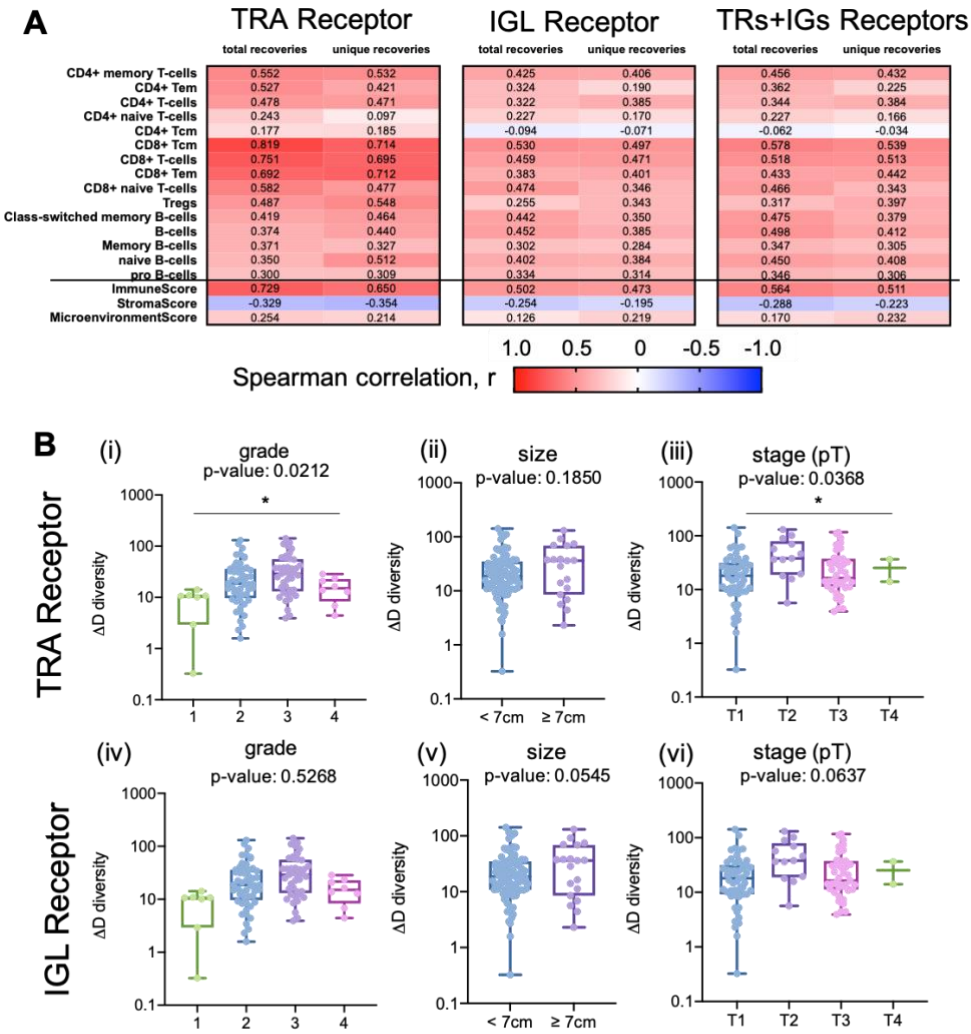
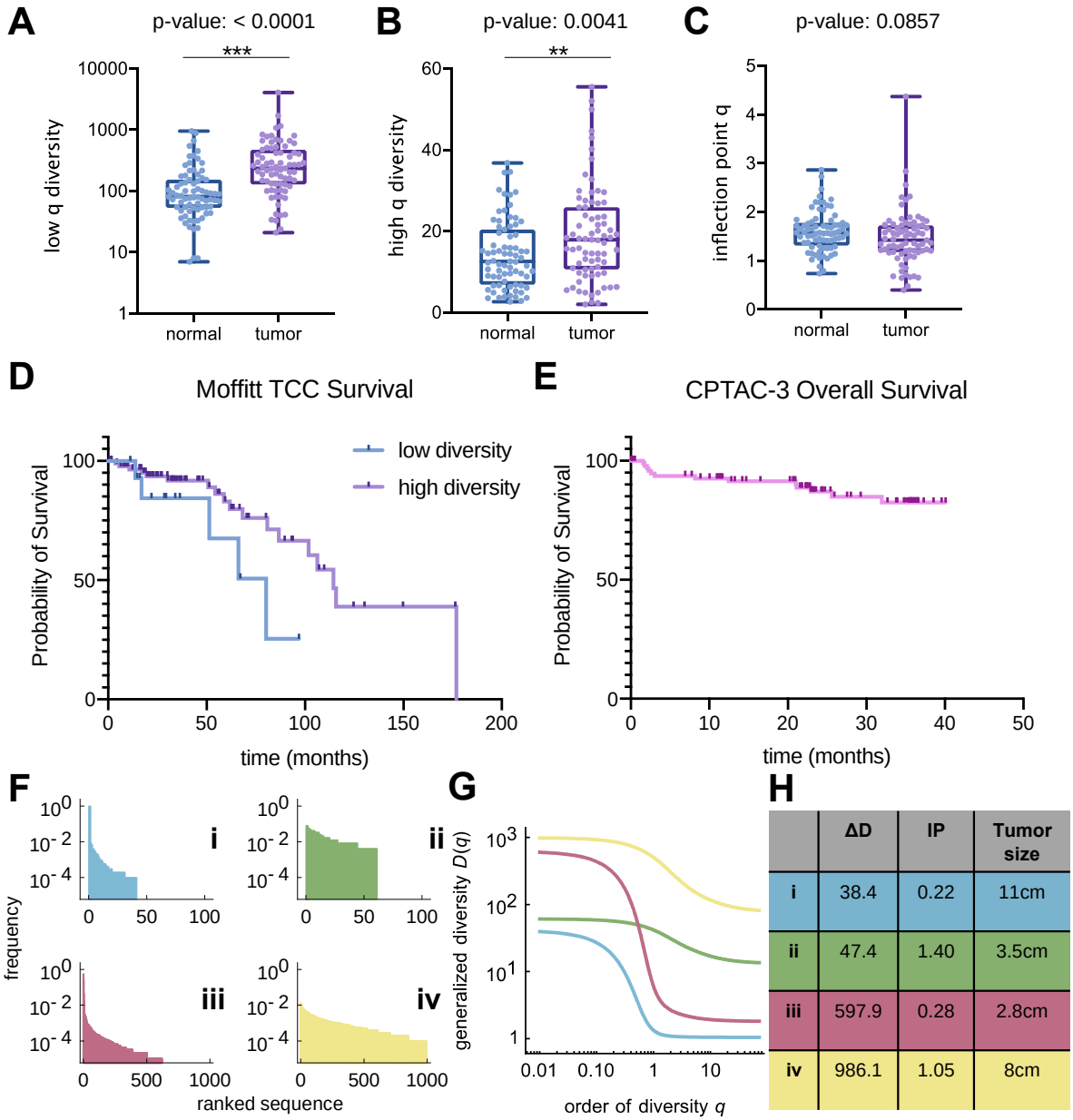


Figure 3: CDR3 sequence diversity trends were validated using xCell scores and a secondary RCC CPTAC3 cohort. (A) Spearman correlation coefficient, r, was calculated between the total and unique number of TRA, IGL, and total (TRs+IGs) recoveries and the xCell scores for various T-cell and B-cell subtypes, Immune Score, Stroma Score, and Microenvironment Score. (B) Grade, largest diameter size, and stage (pT) trends in TRA receptor and IGL receptor ΔD diversity in the CPTAC-3 cohort. Unpaired t-tests were used to compare two group data and ANOVA was used to compare grade, three group data. p-value significance represented by * < 0.05, ** < 0.01, *** < 0.001

Immune receptor subtype evenness, measured by IP, is associated with survival

We first analyzed associations between immune receptors and overall survival in the Moffitt TCC Cohort. Each of our chosen diversity metrics (Methods) was compared for each of the 7 immune receptor types. We used the maximally selected rank statistics (maxstat) approach (Methods), with a cut-point that yielded a maximal survival difference, together with a multivariate Cox regression analysis. We found larger TRA sequence distribution IP, which measures distribution evenness, was significantly associated with longer overall survival (**Fig. 4D**). The IP value of the cut-point in the cohort that yielded a maximal survival difference was 0.826. Using this optimal cut-point resulted in a hazard ratio of 0.526 (Log-rank p-value: 0.049, Cox p-value: 0.036) with the low IP group (IP <0.826, n=15) having a median overall survival of 80 months, and the high diversity group (IP >0.826, n=88) having a median overall survival of 115 months. This trend could not be confirmed with the CPTAC-3 cohort due to the lack of survival information, i.e., overall survival data information was censored for 85 of the 98 CPTAC-3 cases (**Fig. 4E**).

To further illustrate the utility of our diversity metrics for receptor subtype heterogeneity estimation, we show in **Fig. 4F-H** four Moffitt TCC cohort examples of characteristic differences in richness versus evenness space. In both examples of low evenness, dominance (low high-q diversity) is high (**Fig. 4G**). Interestingly, the survival difference unveiled by IP distribution evenness comparison does not necessarily coincide with tumor size, as size rather correlated with richness (ΔD, **Fig. 4H**). Taken together, these findings highlight the ability of these metrics to assess immune sequence distribution heterogeneity via GDI-derived point estimates.



	ΔD	IP	Tumor size
i	38.4	0.22	11cm
ii	47.4	1.40	3.5cm
iii	597.9	0.28	2.8cm
iv	986.1	1.05	8cm

Figure 4: Novel measures of diversity and overall survival. Tumor samples have increased (A) TRs+IGs (all receptor combination) species richness and (B) evenness of CDR3 receptor sequences compared to patient-matched normal tissue. However, tumor samples have (C) reduced TRs+IGs inflection point q diversity compared to normal tissue (mean of normal 1.602 vs mean of tumor 1.465; p-value 0.0857). (D) Individuals in the Moffitt TCC Cohort with larger TRA distribution inflection point (IP, see Methods) had significantly improved overall survival (hazard ratio: 0.526, Cox p-value: 0.036) with a median survival of 115 months compared to those with lower IP. (E) CPTAC-3 Cohort overall survival could not confirm the Moffitt TCC survival trend because over 85% of patients were censored. (F,G,H) Four example patients of the Moffitt TCC cohort with fundamentally different characteristics. While ΔD assesses mainly sequence richness, the inflection IP, clearly visible in (G), is a robust measure of evenness: high IP-distributions are more even.

Demographic and aberrant splicing differences in tumor-infiltrating lymphocyte receptor diversity

In addition to capturing differences in tumor pathology and survival, the GDI was also able to discriminate patients based on demographic differences. B-cell receptor high-q diversity is an estimate of dominance by the most abundant sequence: the lower this value, the more dominant the most abundant sequence. High-q diversity demonstrated differences in immune infiltration of white versus non-white ccRCC patients in the TCC cohort (**Supplemental Fig. 14**). IGH and IGL receptor high-q diversity alone, as well as the IGH+IGK+IGL and total recovery (TRs+IGs) showed at least a 2.1-fold increase in non-white patients compared to white patients (**Supplemental Fig. 14**; p-values < 0.01). This trend was also observed in the IGK receptor high-q diversity, but only with a 1.6-fold increase in non-white individuals (**Supplemental Fig. 14**; p-value 0.0561).

Furthermore, diversity metrics from T-cell receptors TRG and TRD discriminate patients based on gender differences in the TCC cohort. TRG and TRD receptor inflection point q was at least 1.7-fold higher in female patients compared to male patients (TRG: score of 2.503 in females vs. a score of 1.442 in males, p-value: 0.0033; TRD: score of 2.883 in females vs. a score of 1.454 in males, p-value: 0.0003; **Supplemental Fig. 15**). Moreover, when high-q diversity (dominance of the most abundant sequence) was compared between female and male patients with respect to TRG and TRD diversity, in both the Moffitt TCC and CPTAC-3 cohorts, female patients had increased high-q diversity compared to male patients (**Supplemental Fig. 16**).

Finally, we evaluated association of GDI with a recently described aberrant EGFR splice variant in ccRCC (*30*). The Moffitt TCC Cohort of patients were profiled for the presence of this EGFR variant (reported as % EGFR variant), and we found a positive correlation between the percentage of tumors expressing the EGFR variant and the evenness (inflection point q) scores from B-cell receptors (**Supplemental Fig. 17**). This correlation was most significant in IGL receptor diversity, with a Spearman correlation coefficient of r=0.3430 (p-value of 0.004, **Supplemental Fig. 17 & 18**), indicating more even (high IP) distributions have higher proportion of cells with that EGFR variant. This positive correlation between IP and percentage of EGFR variant was not as strong nor significant in the T-cell receptors (**Supplemental Fig. 18**).

DISCUSSION

Here, we demonstrate how a GDI, often used in ecology and evolution (*26, 31, 32*), can be applied to quantify tumor infiltrating lymphocyte receptor subtype diversity in ccRCC. We identified point estimates of this index that are associated with important differences in patient demographics, tumor pathology, and survival. These metrics can help objectively characterize host differences in immune receptor subtypes in ccRCC patients. These novel objective metrics have the ability to provide insight into underlying tumor and host immune relationships by defining differences within and across patients. We used bulk sequencing data from ccRCC tumors to better understand these differences in patient immune receptor subtypes and these metrics can be replicated in other similar cohorts with available sequencing data. These host diversity metrics maybe

especially helpful in elucidating the ideal tumor microenvironment for response to immunotherapy agents in patients with metastatic ccRCC (*33*).

Our findings suggest that individuals with more advanced disease have increased richness in tumor recovered CDR3 sequences. In the Moffitt TCC Cohort, there were statistically significant differences in TRA and IGL diversity with increased richness in tumors with larger diameter and higher grade. Furthermore, we identified tumors with sarcomatoid carcinoma pathology that represent a rare and aggressive histology showed significant increases in different diversity metrics and receptor subtypes, in particular increased CDR3 sequence richness. The immune receptor profiles of these tumors are particularly interesting since sarcomatoid histology has also been associated with very favorable response to checkpoint inhibitors. We postulate that a similar immune receptor profile in other patient tumor may portend a favorable response to checkpoint inhibition. Also, we found a significant increase in richness in the left sided tumors. This difference may explain some of the host related factors associated with left-sided tumors having a poorer clinical outcome than right-sided tumors (*34*). Many of these associations were able to be validated in similar comparisons with the CPTAC-3 cohort studies.

In the TCC Cohort, B-cell receptors with IGL had the most recoveries and TRA had the most T-cell CDR3 sequences recovered (**Supplemental Fig. 1**). It should be noted that point-estimates of diversity or heterogeneity are only comparable within the subtype of interest, within a cohort. However, trends can be compared between point-estimates and patient cohorts. Furthermore, different cohorts based on their clinical context may show differences in immune cell infiltrates, even when looking at simple cell marker differences between T-cells and B-cells. Interestingly, B-cell CDR3 recoveries

dominated in the TCC Cohort, accounting for an average of 93.41% (ranging from at least 53.24% to 99.95%). This contrasted with the CPTAC-3 renal cell carcinoma cohort, which contains generally less aggressive tumors, with only 40.9% of the cohorts respectively comprised of stage 3 or 4 tumors.

Our results demonstrated that increased richness was indicative of larger and more advanced ccRCC tumors which may be related to principle underlying tumor biology. Evenness, as measured by the inflection point q, segregated patients based on survival. In a cross-validation of cut-point analysis, patients with a higher TRA evenness had a significantly improved overall survival compared to individuals with lower TRA evenness. These results indicate that patients' TRA evenness, not richness, may be a possible prognostic biomarker and could possibly have direct therapeutic consequences for response to systemic agents that elicit their effect in the tumor microenvironment (*33*). Furthermore, this evenness metric could be extended to other solid tumor types as a quantitative metric of host contributions to immune infiltration as a possible result of tumor evolution. Such characterizations might become especially interesting in the context of terminally exhausted CD8+ T cells, which were recently shown to be enriched in advanced renal cell carcinoma, interacting with M2-like tumor-associated macrophages leading to immune dysfunction and poorer prognosis (*35*).

Interestingly, we identified demographic differences based on the CDR3 diversity of B-cell receptors, both individually and in combination, showing increased high-q diversity. This amounted to decreased dominance of the most abundant sequence, in non-white individuals compared to white individuals in the Moffitt TCC cohort. However, in this cohort, 88.5% of the patients were white (**Table 1**) and we were unable to confirm

these results in the CPTAC-3 cohort due to missing data. Nevertheless, previously found race-related differences in BCR pathway activation in African Americans compared to European Americans lends support to our finding in differences in BCR dominance/clonality diversity (*36*). Furthermore, high-q diversity/sequence dominance may reveal gender-based differences in the ccRCC microenvironment. We found that females had higher clonality compared to male patients, which also persist in the CPTAC-3 cohort. These demographic differences need to be further investigated, but our results suggest underlying race and gender differences in the heterogeneity of ccRCC microenvironments as revealed by differences in tumor immune infiltration.

Biodiversity has historically been summarized into: alpha-diversity, which measures a single community's diversity; beta-diversity, which quantifies the relative change of species between communities; and gamma diversity, which measures the total diversity in ecology (*37*). Diversity at the individual receptor subtype-level relates closest to alpha-diversity, which is then compared across the patients in a cohort. In a beta-diversity context, we see that many of the trends hold true between the Moffitt TCC and CPTAC-3 cohorts. The most commonly used diversity measures applied to cancer systems have been Shannon and Simpson indices (*38, 39*), which are special cases of the GDI at intermediate values of the parameter *q* (*40, 41*). Our analysis found that it is at the extremes of the continuum of diversity measures (low-q and high-q values) that we can segregate patients in clinically meaningful ways (for example, **Fig. 2** vs. **Supplemental Fig. 4**). In this sense, novel properties of the GDI that are discussed here may allow a more nuanced, and thus more clinically comprehensive characterization of

sequence heterogeneity. These novel objective diversity scales could have important applications for other systems in which tumor heterogeneity with its ecological and evolutionary impact are assessed quantitatively.

Table 1: Clinical and demographic summary of cohorts.

Variables	TCC ccRCC patients, no. (%) (n=105)	CPTAC-3 ccRCC patients, no. (%) (n=110)
Sex		
Female	35 (33.3)	30 (27.3)
Male	70 (66.7)	80 (72.7)
Race		
Asian Indian or Pakistani	2 (1.9)	1 (0.01)
Black	3 (2.9)	1 (0.01)
Other	7 (6.7)	-
White	93 (88.5)	61 (55.5)
Not Reported	-	47 (42.7)
Tumor laterality		
Left	62 (59.0)	NA
Right	43 (41.0)	NA
Surgery type		
Partial nephrectomy	22 (21.0)	NA
Radical nephrectomy	65 (61.9)	NA
Radical nephrectomy with thrombectomy	18 (17.1)	NA
Fuhrman nuclear grade		
1	0	7 (6.4)
2	30 (28.6)	53 (48.2)
3	62 (59.0)	41 (37.3)
4	13 (12.4)	9 (8.2)
pT		
T1	31 (29.5)	52 (47.3)
T2	3 (2.9)	13 (11.8)
T3+T4	70 (66.7)	45 (40.9)
Not Reported	1 (1.0)	-
pN		
N0	27 (25.7)	16 (14.5)
N1	5 (4.8)	4 (3.6)
NX	73 (69.5)	89 (81.9)
pM		
M1	23 (21.9)	NA
MX	82 (78.1)	NA
Sarcomatoid Status		
No	99 (94.3)	NA
Yes	6 (5.7)	NA
Vitality		
Alive	82 (78.1)	96 (87.3)
Dead	23 (21.9)	14 (12.7)
Age (yr)	*at surgery*	*at diagnosis*
Median (range)	65 (36 - 87)	60 (30 - 89)
Pathological tumor size (cm)		
Median (range)	6.0 (1.3 - 17.5)	6.4 (1.0 - 16.0)
% EGFR splice variant alpha		
Median (range)	1.01 (0.00 - 41.33)	NA

NA = no available; ccRCC = clear cell renal cell carcinoma; TCC = Total Cancer Care Protocol

Different point-estimates of the generalized diversity give unique information about the tumor. Increased richness in TRA and IGL diversity informs the size and aggressiveness of a tumor, while dominance of the most abundant sequence segregates patients based on prognosis. We identified a novel measure of evenness among immune receptor subtypes that could accurately classify patients' overall survival. We also found important differences in receptor subtype contributions based on patient demographics such as race and gender. Using these diversity metrics, we identify a new statistical approach to stratify ccRCC patients-based differences in immune infiltration diversity.

METHODS

Clinical samples

Following institutional review board approval (H. Lee Moffitt Cancer Center's Total Cancer Care protocol MCC# 14690; Advarra IRB Pro00014441), we retrospectively obtained clinicopathological and bulk RNA-sequencing patient data from electronic medical records, where all patients had provided consent under the institutional Total Cancer Care (TCC) Protocol. RNA was prepared using the Illumina TruSeq RNA Access kit (now called TruSeq RNA Exome), which was an exome capture of cDNA. Captured material was sequenced on an Illumina HiSeq 4000 according to manufacturer protocols. **Table 1** shows a summary of the clinical information obtained from individuals in the Moffitt TCC Cohort. Relevant clinical and pathological outcomes available from the Moffitt TCC Cohort, including ranges of percentage of tumor with EGFR spice variant alpha are recorded in **Table 1**.

To further investigates the trends identified in point-estimates of diversity from T-cell receptor (TCR) and B-cell immunoglobulin recoveries the Moffitt TCC patient bulk RNA-sequencing, we validated trends identified in the Moffitt TCC Cohort analysis with complementary analysis with the RNA-sequencing from the CPTAC-3 Cohort. **Table 1** shows a summary of the clinical information obtained from individuals in the CPTAC-3 cohort. Relevant clinical and pathological outcomes, aligning with outcomes available in the Moffitt TCC Cohort, that are available in CPTAC-3 Cohort are reported in **Table 1**.

Recovery of immune receptor V(D)J recombination reads from bulk RNA-sequencing

Recovery of immune receptor V(D)J recombination reads was performed in two steps. First, RNAseq binary alignment map (BAM) files were searched, as a straight string search, for 10-mer nucleotide sequences representing the 3' ends of every human V-gene and 5' end of every human J-gene, for all seven immune receptors. Next, the resulting reads were aligned to reference V and J regions obtained from the International Immunogenetics Information System (IMGT). The quantitative parameters for the pairwise alignment were: (i) nucleotide match, + 5, (ii) mismatch, − 10, (iii) opening gap, − 10, and (iv) extending gap, − 10. The threshold for a V or J gene segment match was a score of ≥ 65. To ensure V and J read fidelity, only reads with a 20 nucleotide or greater match length for both V and J regions, and within the 20 nucleotides, a > 90% nucleotide match fidelity for both V and J regions were considered as matches. Additionally, and a productive CDR3 domain, defined as an in-frame junction without stop codons, was required for recombination read identification. Code for the method described above can

be obtained at: https://github.com/bchobrut-USF/vdj under “Code Package A”. See also https://hub.docker.com/r/bchobrut/vdj for a container version of the code with a readme file.

Generalized diversity index for patient quantifying CDR3 receptor diversity

The generalized diversity index (GDI) can be viewed as a continuous, non-increasing function over a range of values described by the parameter *q*, called order of diversity. This parameter allows a consideration of multiple scales of diversity simultaneously or in combination. GDI is often used in ecology (*20*) and was more recently introduced to quantify intratumor heterogeneity and evolution (*26-28*). Formally, GDI is calculated as:

$$D(q) = \left(\sum_{i=1}^{n} p_i^q\right)^{\frac{1}{1-q}}$$

where $D(q)$ is called the diversity index at the given *order of diversity* q, n is the number of unique CDR3 sequences recovered across the entire cohort, and p_i is the relative proportion of i-th CDR3 sequence. We typically evaluated the diversity score, $D(q)$, for q between 0.01 and 100 numerically for each patient, for each of the receptor and immunoglobulin types individually, as well as in biologically meaningful groups (TRA+TRB together, TRG+TRD together, and IGH+IGK+IGL together). Varying the value of *q* represents interpolating between richness and evenness: richness is weighted more

at low values of q, and evenness is weighted more at higher values of q. High-q GDI scales inversely with dominance or clonality.

Point-estimates derived from patient's GDI, D(q), can be used for easy of comparison of sequence distributions across patients and cohorts. These point-estimates of interest include (i) low-q diversity ($D(0.01)$), which describes the epitope richness of the patients, (ii) high-q diversity ($D(100)$), which describes dominance of the "main driving" epitope, and (iii) ΔD, which measures the difference between low-q and high-q diversity ($D(0.01) - D(100)$). Furthermore, when visualizing the continuum of diversity measures *D(q)* with q in log-scaling, the continuum of diversity measures appears to have an inflection point, corresponding to a scale of diversity where small changes in the key parameter *q* can have large impact: the higher this value, the more even we expect a distribution to be, as the inflection point tends to infinity for perfectly even distributions (corresponding to *n* sequences all at frequency 1/*n*). Thus, two additional point-estimates of interest that we used are (iv) the value *q* at which an inflection point occurs (IP), and (v) the slope at the inflection point (denoted as IP slope). All code for calculating CDR3 diversity and its summary metrics has been implemented in Julia (version 1.4.0) and documented in the publicly available package OncoDiversity.jl.

To determine the impact of all five point-estimates of diversity, we ran a correlation analysis and determined that we could reduce our five point-estimates of diversity down to three metrics for comparison across receptor groups and patients. The Spearman correlation coefficients were calculated between point-estimate metrics and comparisons between low-q diversity, ΔD diversity, and the inflection point slope all had significant and

very strong Spearman correlation coefficients of 0.98 or greater, so moving forward, we just focused on one of those metrics as a measure of species richness diversity (**Supplemental Fig. 19**). There was not a strong correlation between high-q diversity and the inflection point q metrics and either metric compared to any of the three species richness diversity metrics (low-q diversity, ΔD diversity, and inflection point slope), so we continue to look at the high-q diversity and inflection point q separately and in additional to the single species richness measure.

Assessment of clinical and survival associations with CDR3 features

Clinical associations were evaluated for recoveries in TRA, TRB, TRG, TRD, IGH, IGK, and IGL separately as well as in combinations of TRA+TRB, TRG+TRD, and IGH+IGK+IGL. After point estimates of diversity were calculated for each patient and each receptor subtype/combination, the clinical parameter values were assessed to identify if CDR3 receptor diversity could discriminated ccRCC patients based on relevant clinical and pathological outcomes, as well as the percentage of tumor with EGFR spice variant alpha. Largest diameter size and age were the only two continuous variables, which were evaluated by dividing the cohort into above and below the median of the diversity and compared with unpaired t-tests. All other categorial data types were divided by categories and the diversity metric was compared across the categories using unpaired t-test for 2 categories and ANOVA for 3 or more categories.

Survival correlations for the above combinations were performed by separating the cohort into above and below the median based on point-estimates of the generalized

diversity index. In addition, the maximally selected rank statistical analysis was performed to estimate an optimal cutpoint in the quantitative point estimates as a binary classification rule regarding overall survival time (*42*). The Kaplan-Meier (KM) curve method was used to calculate survival probability and log-rank test was used to compare the above (high diversity) and below (low diversity) groups. Graph-Pad Prism software (version 8) and R version 3.6.1 were used for computing statistical comparisons and outputting figures.

xCell Scores

From bulk RNA-sequencing for each patient, xCell scores were calculated for various T-cell and B-cell subtypes as well as the Immune Score, Stroma Score, and Microenvironment Score. Then for each xCell score calculated a Spearman correlation was calculated for the total and unique recoveries identified for TRA receptor, IGL receptor, and aggregate combination (TRs+IGs).

Patients in the Moffitt TCC cohort had previously undergone bulk RNA sequencing of macro-dissected tumor samples using the TruSeq RNA Exome kit (Illumina) for 50 million 100–base pair paired-end reads. RNA sequence reads were aligned to the human reference genome in a splice-aware fashion using Spliced Transcripts Alignment to a Reference (STAR) (*43*), allowing for accurate alignments of sequences across introns. Aligned sequences were assigned to exons using the HTseq package (*44*) to generate initial counts by region. Normalization, expression modeling, and difference testing were performed using DESeq2 (*45*). For the CPTAC cohorts, detailed methodology regarding RNA sequencing can be found at its source web page(*25*).

RNA sequencing data was analyzed for cell-type enrichment using the xCell bioinformatic pipeline (*25*). xCell uses a compendium of validated gene expression signatures for 64 individual cell-types derived from thousands of expression profiles. Single-sample gene set enrichment analysis scores were adjusted for spillover compensation to generate an adjusted enrichment score for each cell type within the specimen, which is referred to as the xCell score. xCell scores were generated for each of the 64 cell-types for each ccRCC tumor specimen.

FUNDING

United States Army Medical Research Acquisition Activity Department of Defense (KC180139 to B.J.M). P.M.A. was supported by an American Cancer Society Moffitt IRG award, the Richard O. Jacobson Foundation (Evolutionary Therapy Center of Excellence at Moffitt Cancer Center), and the William G. 'Bill' Bankhead Jr and David Coley Cancer Research Program (20B06).

AUTHOR CONTRIBUTIONS

MCF: conceptualization, data curation, formal analysis, investigation, methodology, software, validation, visualization, writing.

NC: data curation, critical review.

BIC: data curation, adaptive immune receptor recombination read extractions from the Moffitt TCC cohort RNAseq files.

YK: statistical analysis, validation, critical review.

JKT: logistic support, data curation, statistical analysis, critical review.

AB: logistic support, data curation, statistical analysis, critical review.

JM: investigation, critical review.

MF: logistic support, data curation.

ES: logistic support, data curation.

ENK: logistic support, data curation.

GB: adaptive immune receptor recombination read extractions from the Moffitt TCC cohort RNAseq files, critical review, study supervision.

BJM: conceptualization, investigation, methodology, validation, critical review, writing, study supervision.

PMA: conceptualization, investigation, methodology, software, validation, critical review, writing, study supervision.

COMPETING INTERESTS

No potential conflicts of interest to report: MCF, NC, BIC, YK, JKT, AB, MF, ES, ENK, GB, BJM, PMA. JM has ownership interest in Fulgent Genetics, Inc., Aleta Biotherapeutics, Inc., Cold Genesys, Inc., Myst Pharma, Inc., and Tailored Therapeutics, Inc., and is a consultant/advisory board member for ONCoPEP, Inc., Cold Genesys, Inc., Morphogenesis, Inc., Mersana Therapeutics, Inc., GammaDelta Therapeutics, Ltd., Myst Pharma, Inc., Tailored Therapeutics, Inc., Verseau Therapeutics, Inc., Iovance Biotherapeutics, Inc., Vault Pharma, Inc., Noble Life Sciences Partners, Fulgent Genetics, Inc., UbiVac, LLC, Vycellix, Inc., and Aleta Biotherapeutics, Inc.

DATA AND CODE AVAILABILITY

Code for VDJ epitope recoveries from patient sequencing data (BAM files) is publicly available on Docker at https://hub.docker.com/r/bchobrut/vdj and GitHub at https://github.com/bchobrut/vdj_recovery.

Code for calculating patient CDR3 diversity from CDR3 recoveries is publicly available at: https://github.com/mcfefa/OncoDiversity.jl.

REFERENCES

1. K. A. Cronin *et al.*, Annual Report to the Nation on the Status of Cancer, part I: National cancer statistics. *Cancer* **124**, 2785-2800 (2018).

2. R. J. Motzer, P. Russo, Systemic therapy for renal cell carcinoma. *J Urol* **163**, 408-417 (2000).

3. B. Escudier, Combination Therapy as First-Line Treatment in Metastatic Renal-Cell Carcinoma. *N Engl J Med* **380**, 1176-1178 (2019).

4. A. Snyder *et al.*, Genetic basis for clinical response to CTLA-4 blockade in melanoma. *N Engl J Med* **371**, 2189-2199 (2014).

5. N. A. Rizvi *et al.*, Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science* **348**, 124-128 (2015).

6. M. Yarchoan, A. Hopkins, E. M. Jaffee, Tumor Mutational Burden and Response Rate to PD-1 Inhibition. *N Engl J Med* **377**, 2500-2501 (2017).

7. S. Maleki Vareki, High and low mutational burden tumors versus immunologically hot and cold tumors and response to immune checkpoint inhibitors. *J Immunother Cancer* **6**, 157 (2018).

8. K. Yoshihara *et al.*, Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat Commun* **4**, 2612 (2013).

9. D. Miao *et al.*, Genomic correlates of response to immune checkpoint therapies in clear cell renal cell carcinoma. *Science* **359**, 801-806 (2018).

10. D. F. McDermott *et al.*, Clinical activity and molecular correlates of response to atezolizumab alone or in combination with bevacizumab versus sunitinib in renal cell carcinoma. *Nat Med* **24**, 749-757 (2018).

11. D. A. Braun *et al.*, Interplay of somatic alterations and immune infiltration modulates response to PD-1 blockade in advanced clear cell renal cell carcinoma. *Nat Med* **26**, 909-918 (2020).

12. A. Hajiran *et al.*, Reconnaissance of tumor immune microenvironment spatial heterogeneity in metastatic renal cell carcinoma and correlation with immunotherapy response. *Clin Exp Immunol* **204**, 96-106 (2021).

13. B. I. Chobrutskiy, S. Zaman, A. Diviney, M. M. Mihyu, G. Blanck, T-cell receptor-alpha CDR3 domain chemical features correlate with survival rates in bladder cancer. *J Cancer Res Clin Oncol* **145**, 615-623 (2019).

14. A. M. Roca, B. I. Chobrutskiy, B. M. Callahan, G. Blanck, T-cell receptor V and J usage paired with specific HLA alleles associates with distinct cervical cancer survival rates. *Hum Immunol* **80**, 237-242 (2019).

15. B. I. Chobrutskiy, S. Zaman, W. L. Tong, A. Diviney, G. Blanck, Recovery of T-cell receptor V(D)J recombination reads from lower grade glioma exome files correlates with reduced survival and advanced cancer grade. *J Neurooncol* **140**, 697-704 (2018).

16. B. M. Callahan *et al.*, T-cell receptor-beta V and J usage, in combination with particular HLA class I and class II alleles, correlates with cancer survival patterns. *Cancer Immunol Immunother* **67**, 885-892 (2018).

17. V. I. Zarnitsyna, B. D. Evavold, L. N. Schoettle, J. N. Blattman, R. Antia, Estimating the diversity, completeness, and cross-reactivity of the T cell repertoire. *Front Immunol* **4**, 485 (2013).

18. J. Nikolich-Zugich, M. K. Slifka, I. Messaoudi, The many important facets of T-cell repertoire diversity. *Nat Rev Immunol* **4**, 123-132 (2004).

19. J. Glanville *et al.*, Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proc Natl Acad Sci U S A* **106**, 20216-20221 (2009).

20. L. Jost, Entropy and diversity. *Oikos* **113**, 363-375 (2006).

21. M. O. Hill, Diversity and Evenness: A Unifying Notation and Its Consequences. *Ecology* **54**, 427-432 (1973).

22. R. H. MacArthur, Patterns of Species Diversity. *Biological Review Cambridge Philosophical Society* **40**, 510-533 (1965).

23. J. Kaplinsky, R. Arnaout, Robust estimates of overall immune-repertoire diversity from high-throughput measurements on samples. *Nat Commun* **7**, 11881 (2016).

24. Moffitt Cancer Center, Total Cancer Care. https://moffitt.org/research-science/total-cancer-care/.

25. NCI Office of Cancer Clinical Proteomics Research, Clinical Proteomic Tumor Analysis Consortium (CPTAC). https://cptac-data-portal.georgetown.edu/cptac/public?scope=Phase+III.

26. D. Miroshnychenko *et al.*, Spontaneous Cell Fusions as a Mechanism of Parasexual Recombination in Tumor Cell Populations. *Nature Ecology and Evolution* **5**, 379–391 (2021).

27. M. C. Ferrall-Fairbanks, P. M. Altrock, in *Translational Bioinformatics for Therapeutic Development: Methods and Protocols,* J. Markowitz, Ed. (Springer US, 2020).

28. M. C. Ferrall-Fairbanks, M. Ball, E. Padron, P. M. Altrock, Leveraging Single-Cell RNA Sequencing Experiments to Model Intratumor Heterogeneity. *JCO Clinical Cancer Informatics* ***3***, 1-10 (2019).

29. D. Aran, Z. Hu, A. J. Butte, xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol* **18**, 220 (2017).

30. S. Zaman *et al.*, Aberrant Epidermal Growth Factor Receptor RNA Splice Products Are Among the Most Frequent Somatic Alterations in Clear Cell Renal Cell Carcinoma and Are Associated with a Poor Response to Immunotherapy. *Eur Urol Focus*, (2019).

31. H. Tuomisto, A consistent terminology for quantifying species diversity? Yes, it does exist. *Oecologia* **164**, 853-860 (2010).

32. R. Cazzolla Gatti, N. Amoroso, A. Monaco, Estimating and comparing biodiversity with a single universal metric. *Ecological Modelling* **424**, 109020 (2020).

33. C. M. Diaz-Montero, B. I. Rini, J. H. Finke, The immunology of renal cell carcinoma. *Nat Rev Nephrol*, (2020).

34. S. Guo *et al.*, Prognostic significance of laterality in renal cell carcinoma: A population-based study from the surveillance, epidemiology, and end results (SEER) database. *Cancer Med* **8**, 5629-5637 (2019).

35. D. A. Braun *et al.*, Progressive immune dysfunction with advancing disease stage in renal cell carcinoma. *Cancer Cell*, (2021).

36. D. M. Longo *et al.*, Racial differences in B cell receptor signaling pathway activation. *J Transl Med* **10**, 113 (2012).

37. L. Jost, Partitioning diversity into independent alpha and beta components. *Ecology* **88**, 2427-2439 (2007).

38. V. Almendro *et al.*, Inference of tumor evolution during chemotherapy by computational modeling and in situ analysis of genetic and phenotypic cellular diversity. *Cell Rep* **6**, 514-527 (2014).

39. A. Marusyk *et al.*, Non-cell-autonomous driving of tumour growth supports sub-clonal heterogeneity. *Nature* **514**, 54-58 (2014).

40. C. E. Shannon, A Mathematical Theory of Communication. *Bell Syst Tech J* **27**, 379-423 (1948).

41. E. H. Simpson, Measurement of Diversity. *Nature* **163**, 688-688 (1949).

42. T. Hothorn, B. Lausen, On the exact distribution of maximally selected rank statistics. *Computational Statistics & Data Analysis* **43**, 121-137 (2003).

43. A. Dobin *et al.*, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).

44. S. Anders, P. T. Pyl, W. Huber, HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166-169 (2015).

45. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).

ACKNOWLEDGEMENTS

This work was supported in part by the Biostatistics and Bioinformatics Shared Resource at the H. Lee Moffitt Cancer Center & Research Institute, a National Cancer Institute–designated Comprehensive Cancer Center (P30-CA076292). Total Cancer Care (TCC) Protocol at Moffitt Cancer Center, which was enabled in part by the generous support of the DeBartolo Family. Some data used in this publication were generated by the National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC), available to which author via dbGaP project approval number 6757. Editorial assistance was provided by the Moffitt Cancer Center's Scientific Editing Department by Dr. Paul Fletcher and Daley Drucker (no compensation was given beyond their regular salaries). We also thank Gregory J. Kimmel, PhD for valuable comments.