Using DNA sequencing data to quantify T cell fraction and therapy response

https://doi.org/10.1038/s41586-021-03894-5

Received: 21 December 2020

Accepted: 10 August 2021

Published online: 08 September 2021

Robert Bentham[1,2,52], Kevin Litchfield[2,3,52], Thomas B. K. Watkins[4,52], Emilia L. Lim[2,4], Rachel Rosenthal[4], Carlos Martínez-Ruiz[1,2], Crispin T. Hiley[2,4], Maise Al Bakir[4], Roberto Salgado[5,6], David A. Moore[2,7,8], Mariam Jamal-Hanjani[2,8,9], TRACERx Consortium*, Charles Swanton[2,4,8] & Nicholas McGranahan[1,2] ✉

The immune microenvironment influences tumour evolution and can be both prognostic and predict response to immunotherapy[1,2]. However, measurements of tumour infiltrating lymphocytes (TILs) are limited by a shortage of appropriate data. Whole-exome sequencing (WES) of DNA is frequently performed to calculate tumour mutational burden and identify actionable mutations. Here we develop T cell exome TREC tool (T cell ExTRECT), a method for estimation of T cell fraction from WES samples using a signal from T cell receptor excision circle (TREC) loss during V(D)J recombination of the T cell receptor-α gene (*TCRA* (also known as *TRA*)). *TCRA* T cell fraction correlates with orthogonal TIL estimates and is agnostic to sample type. Blood *TCRA* T cell fraction is higher in females than in males and correlates with both tumour immune infiltrate and presence of bacterial sequencing reads. Tumour *TCRA* T cell fraction is prognostic in lung adenocarcinoma. Using a meta-analysis of tumours treated with immunotherapy, we show that tumour *TCRA* T cell fraction predicts immunotherapy response, providing value beyond measuring tumour mutational burden. Applying T cell ExTRECT to a multi-sample pan-cancer cohort reveals a high diversity of the degree of immune infiltration within tumours. Subclonal loss of 12q24.31–32, encompassing *SPPL3*, is associated with reduced *TCRA* T cell fraction. T cell ExTRECT provides a cost-effective technique to characterize immune infiltrate alongside somatic changes.

Checkpoint inhibitors (CPIs) have emerged as revolutionary cancer treatments, acting to release the brakes on the immune system[3,4]. The clinical response to CPI therapy, however, is not universal[5] and is principally governed by the presence of an immune stimulus, such as neoantigens, and an immune response, mediated by T cells[2]. Although neoantigens can be predicted from WES[1], T cell quantification has to date required additional biological material, time and expertise, adding to the cost of immunotherapy.

Here we propose a method for the estimation of the T cell fraction present in a WES sample. This method makes use of a signal based on somatic copy number from V(D)J recombination and the loss of TRECs. We explore the underlying features that predict T cell infiltration in tumours and blood and evaluate determinants of immune heterogeneity within tumours. Finally, we demonstrate that our estimated T cell fraction can be used as a predictor of clinical response to CPI therapy.

Results

Inferring T cell fraction from WES data

T cell diversity, which is required for immune system recognition of foreign antigens, is a product of V(D)J recombination, in which segments within the T cell receptor genes recombine. The α-chain of the T cell receptor is encoded by the *TCRA* gene, and the result of V(D)J recombination is the excision of unselected gene segments from *TCRA* as TRECs—*TCRA* thus undergoes a deletion event within T cells.

Tools to infer cancer somatic copy number alteration (SCNA)[6–9] rely on the read-depth ratio (RDR), reflecting the log of the ratio of reads between the tumour sample and its matched control (for example, buffy coat in a centrifuged blood sample). Deviation in the RDR from zero is assumed to reflect a tumour SCNA. However, this assumption does not hold for *TCRA*; a deviation in the RDR may reflect T cell-specific deletion events, and SCNA tools may thus erroneously infer tumour SCNA. Indeed, in the TRACERx100 cohort, multiple SCNAs within *TCRA* were inferred in 165 out of 327 tumour samples (Extended Data Fig. 1a). The RDR had the highest deviation from zero within segments frequently included within TRECs (Extended Data Fig. 1b, c). This suggests that most detected SCNAs within *TCRA* reflect a signal of relative T cell content rather than cancer SCNAs.

To exploit this signal to quantify T cell content we developed T cell ExTRECT, which uses a modified RDR within *TCRA* to directly quantify T cell infiltrate in WES samples (Fig. 1a), referred to as the *TCRA* T cell fraction. Unlike scores derived from RNA-sequencing (RNA-seq)

[1] Cancer Genome Evolution Research Group, Cancer Research UK Lung Cancer Centre of Excellence, University College London Cancer Institute, London, UK. [2] Cancer Research UK Lung Cancer Centre of Excellence, University College London Cancer Institute, London, UK. [3] The Tumour Immunogenomics and Immunosurveillance Lab, Cancer Research UK Lung Cancer Centre of Excellence, University College London Cancer Institute, London, UK. [4] Cancer Evolution and Genome Instability Laboratory, The Francis Crick Institute, London, UK. [5] Department of Pathology, GZA-ZNA, Antwerp, Belgium. [6] Division of Research, Peter MacCallum Cancer Centre, University of Melbourne, Melbourne, Victoria, Australia. [7] Department of Cellular Pathology, University College London Hospitals, London, UK. [8] Department of Medical Oncology, University College London Hospitals, London, UK. [9] Cancer Metastasis Lab, University College London Cancer Institute, London, UK. [52] These authors contributed equally: Robert Bentham, Kevin Litchfield, Thomas B. K. Watkins. *A list of authors and their affiliations appears at the end of the paper. ✉e-mail: nicholas.mcgranahan.10@ucl.ac.uk

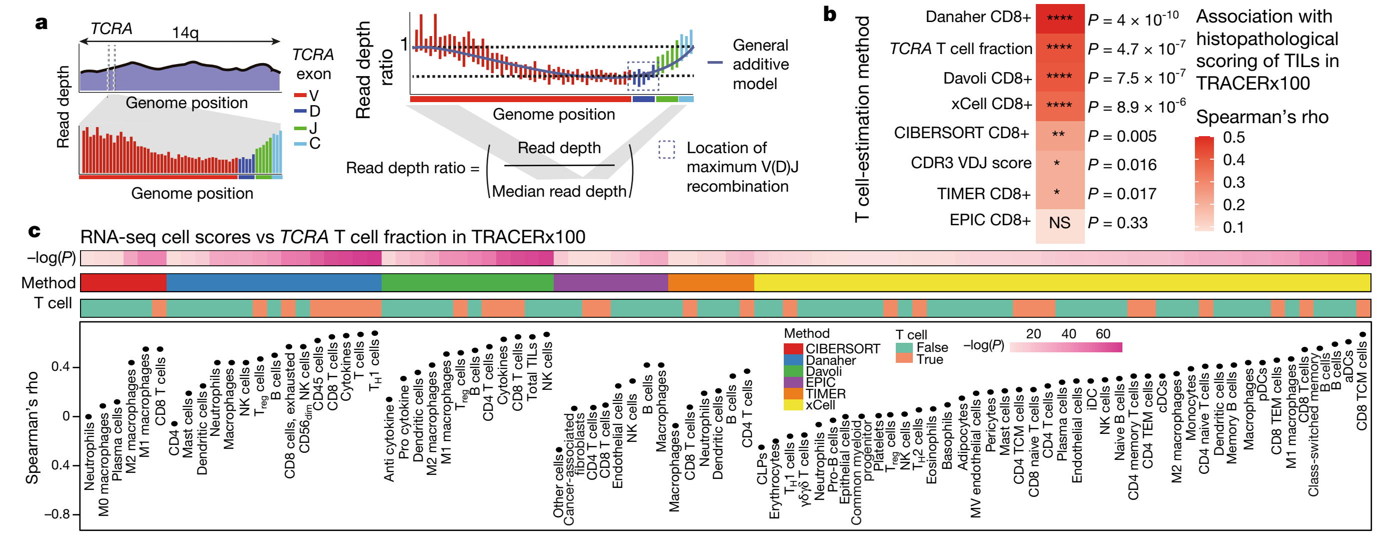


Fig. 1 | Overview and validation of T cell ExTRECT. a, Overview of how a V(D)J recombination signal is identified from read depth within *TCRA* in T cell fraction calculation. **b**, Association between histopathology TIL scores and measures of CD8+ T cell content from either RNA-seq (Danaher[14], Davoli[15], EPIC[19], TIMER[17], CIBERSORT[18] and xCell[16]) or DNA (T cell ExTRECT and CDR3 V(D)J score). NS, not significant. **c**, Association between *TCRA* T cell fraction with RNA-based scores for immune cell types (Danaher[14], Davoli[15], EPIC[19], TIMER[17], CIBERSORT[18] and xCell[16]), ordered by strength of association (Spearman's rho) with *TCRA* T cell fraction. aDCs, activated dendritic cells; cDCs, conventional dendritic cells; CLPs, common lymphoid progenitors; CMPs, common myeloid progenitors; iDCs, immature dendritic cells; NK, natural killer; pDCs, plasmacytoid dendritic cells; TCM cells, central memory T cells; TEM cells, effector memory T cells; T_H1, T helper 1; T_H2, T helper 2.

data, the *TCRA* T cell fraction represents a direct quantification of the proportion of T cells within a sample. *TCRA* T cell fraction was not dependent on whether samples were freshly frozen or formalin-fixed paraffin-embedded (FFPE) (Methods, Extended Data Fig. 1d, e). Thus, T cell ExTRECT can be applied to any WES sample, thereby enabling analysis of the T cell fraction in both tumour and blood samples.

Validation of *TCRA* T cell fraction

To evaluate the accuracy of T cell ExTRECT, we used five orthogonal approaches.

First, to assess the ability to accurately determine the presence or absence of T cells within a sample, we used WES data from cell lines originating from T cell lymphoma (JURKAT, PEER and HPB-ALL) and 14 colorectal cancer cell lines derived from HCT116 with varying degrees of genomic complexity[10,11]. All HCT116 cell lines had a calculated fraction of 0. Conversely, the three cell lines derived from T cell lymphoma had scores close to 1 (between 0.95 and 0.96) (Extended Data Fig. 1f).

Second, we used an alternative DNA-based method of inferring immune content[12], based on the number of reads that align to the complementarity-determining region 3 (CDR3) following V(D)J recombination (Methods, 'Calculation of CDR3 V(D)J score'). In the TRACERx100[13] cohort (Extended Data Fig. 1g) we observed a significant positive correlation between *TCRA* T cell fraction and the CDR3 V(D)J score ($\rho = 0.36$, $P = 1.4 \times 10^{-13}$; Extended Data Fig. 1h). However, the CDR3 V(D)J score was constrained by sequencing depth; the number of reads aligning to the CDR3 region was typically very low (1st quartile, 0; median, 2; mean, 2.335; 3rd quartile, 3; maximum, 14).

Third, we simulated next generation sequencing data with a range of T cell fractions (Extended Data Fig. 2a–d). We observed a highly significant relationship between simulated and calculated T cell fraction ($\rho = 0.99986$, $P < 2.2 \times 10^{-16}$; Extended Data Fig. 2b). Using downsampling and simulations, we found that the *TCRA* T cell fraction estimates remained consistent at coverage above and including 30× ($\rho = 0.84$, $P = 1.4 \times 10^{-14}$) (Extended Data Fig. 2e, f). By contrast, the results from the CDR3 method were heavily skewed by sequencing coverage—when selecting the five samples with the highest CDR3 coverage and downsampling to 50×, only one sample with 3 or more CDR3 reads was detected (Extended Data Fig. 2g).

Fourth, to further confirm the accuracy of the *TCRA* T cell fraction for quantifying T cells, we evaluated its association with histopathology-derived TIL scores from samples stained with haematoxylin and eosin. Selecting the subset of tumour samples with both RNA-seq data and histopathology-derived TIL scores (147 samples), we evaluated how the *TCRA* T cell fraction, CDR3 V(D)J score, and six RNA-seq-based immune measures for CD8+ cells (Danaher[14], Davoli[15], xCell[16], TIMER[17], CIBERSORT[18] and EPIC[19]) compared with histopathology-derived TIL scores (Fig. 1b). The Danaher CD8+ score had the strongest association ($\rho = 0.49$), followed by the *TCRA* T cell fraction ($\rho = 0.41$), Davoli ($\rho = 0.4$), xCell ($\rho = 0.36$), CIBERSORT ($\rho = 0.23$), TIMER ($\rho = 0.2$), CDR3 V(D)J score ($\rho = 0.2$) and EPIC ($\rho = 0.082$).

Finally, the *TCRA* T cell fraction from WES was compared directly with RNA-seq methods, and was found to have a significant positive relationship with multiple immune scores[1,14–19] with the strongest associations being found for T cell-related scores (Fig. 1c).

Determinants of T cell content in blood

We next explored the key determinants of T cell immune infiltrate in matched control blood WES samples.

Within the TRACERx100[13] cohort, the blood *TCRA* T cell fraction was significantly higher in females than males ($P = 0.0057$, effect size (ES) = 0.28; Fig. 2a) and we observed a trend for higher blood T cell fraction in patients with lung squamous cell carcinoma (LUSC) compared with those with lung adenocarcinoma (LUAD) ($P = 0.066$, ES = 0.19; Extended Data Fig. 3a). We also observed a significant positive relationship between the blood *TCRA* T cell fraction and matched tumour *TCRA* T cell fraction ($\rho = 0.42$, $P = 1.7 \times 10^{-5}$; Fig. 2a). These data suggest that tumour immune infiltrate may influence T cell levels in circulating blood or vice versa. We observed broadly consistent results from LUAD and LUSC The Cancer Genome Atlas (TCGA) cohorts[20,21] (Extended Data Fig. 3b, c).

To further examine the determinants of blood T cell fraction, we explored WES samples derived from both blood and physiologically normal oesophagus epithelial (PNE) tissue from a previous study[22]. Although blood samples exhibited a wide range of *TCRA* T cell fraction levels, the majority of PNE tissue had no detectable T cell infiltration (Extended Data Fig. 3d, e). Dividing the PNE samples by presence of

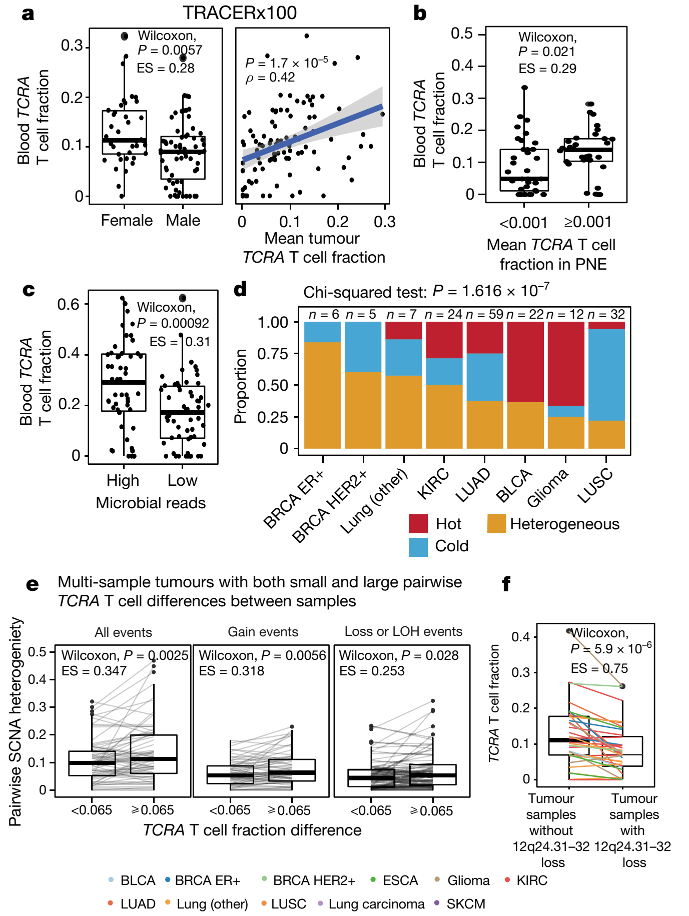


Fig. 2 | Determinants of T cell fraction. a, Predictors of blood *TCRA* T cell fraction in the TRACERx100 cohort. **b**, Association of *TCRA* T cell fraction in PNE with blood *TCRA* T cell fraction. **c**, Microbial reads from Kraken analysis versus blood *TCRA* T cell fraction (n = 111). **d**, Proportion of uniformly immune-hot, uniformly immune-cold or heterogeneous tumours (Methods). KIRC, renal clear cell carcinoma. **e**, Multi-sample tumours (n = 76) with heterogeneous immune infiltrate defined as having both a pair of samples with pairwise *TCRA* T cell fraction difference less than 0.065 and another with pairwise difference greater than or equal to 0.065, versus pairwise SCNA heterogeneity score (Methods). The threshold 0.065 is the mean of all pairwise differences between samples. **f**, *TCRA* T cell fraction difference between samples with or without subclonal loss of 12q24.31–32. All Wilcoxon tests are two-sided and box plots represent lower quartile, median and upper quartile, unpaired data uses the Wilcoxon rank-sum (Mann–Whitney U) test and paired data uses the Wilcoxon signed-rank test. ESCA, oesophegal carcinoma; SKCM, skin cutaneous melanoma.

T cell infiltration revealed a significant association with blood *TCRA* T cell fraction (P = 0.021, ES = 0.29; Fig. 2b). Therefore, similar to tumour samples, high levels of T cell infiltration in normal tissue may influence or be influenced by the *TCRA* T cell fraction observed in blood. In a linear model predicting T cell fraction in blood, only the infiltration level in normal tissue was significant (Extended Data Fig. 3f); no genomic factors, such as normal tissue mutation burden or driver mutation status were predictive of T cell infiltration in PNE tissue (Extended Data Fig. 3g).

Viral or bacterial infections could also influence T cell levels in blood. To explore this we obtained normalized estimates for the abundance of microbial reads from blood and tumour samples from the LUAD and LUSC TCGA cohorts[23]. Blood samples with elevated microbial reads (above the median of 6.81) had significantly higher blood *TCRA* T cell fractions (P = 0.00092, ES = 0.31, Wilcoxon test; Fig. 2c,). No corresponding association was identified in tumour samples (Extended Data Fig. 3h). No specific virus or bacteria were associated with the blood *TCRA* T cell fraction. In tumour samples, significant associations for bacteria of the genus *Williamsia* in LUAD (ρ = −0.17, P = 0.00011, false discovery rate (FDR) P = 0.013) and *Paeniclostridium* in LUSC (ρ = −0.2, P = 0.00013, FDR P = 0.015) were observed (Extended Data Fig. 3i–k). Both had higher normalized log counts per million (logCPM) values when *TCRA* T cell fraction was lower, suggesting that they may be opportunistic species exploiting an immune-cold tumour microenvironment.

Determinants of tumour T cell content

Next, we investigated factors influencing T cell infiltrate in tumour tissue. We used our recently published pan-cancer cohort of multi-sample data[24] to investigate both the extent and possible genomic basis for immune infiltrate heterogeneity. In total, we evaluated T cell infiltrate in 731 samples from 178 tumours of 12 cancer types (Extended Data Fig. 4a, b).

We classified each multi-sample tumour as uniformly hot (all samples ≥ 0.11, the mean *TCRA* T cell fraction), uniformly cold (all samples < 0.11) or heterogeneous. There was a significant difference in the proportion of these categories by cancer type (chi-squared test: $P = 1.62 \times 10^{-7}$; Fig. 2d,) with oestrogen receptor-positive (ER+) breast cancer (BRCA) tumours being the most heterogeneous (83%) and LUSC tumours being the least heterogeneous (22%). Clear differences in the prevalence and heterogeneity of immune infiltrate were observed across cancer types; for instance, while bladder cancer (BLCA) and LUAD showed similar numbers of heterogeneous tumours (36% versus 37%), about 64% of BLCA tumours were uniformly immune-hot and 0% were uniformly immune-cold, whereas in LUAD, 37% of tumours were uniformly immune-cold and 25% were uniformly immune-hot. This suggests that for certain cancer types there is a highly localized immune infiltrate, which can be subject to considerable sampling bias.

Next, we examined the relationship between SCNAs and immune diversity. We restricted the analysis to tumours for which there were at least three samples and a heterogeneous mixture of T cell infiltrate. Pairwise SCNA heterogeneity between any two samples was calculated as the sum of the proportion of the genome with unique SCNAs in either sample. Pairs of tumour samples with a large disparity in the *TCRA* T cell fraction (greater than or equal to the mean of all pairwise distances, 0.065) were associated with a larger differences in SCNA heterogeneity compared with matched-sample pairs with low *TCRA* T cell fraction heterogeneity (all events: P = 0.0025, ES = 0.347; gain events: P = 0.0056, ES = 0.318; loss or loss of heterozygosity (LOH) events: P = 0.028, ES = 0.253, n = 76; Fig. 2e).

To explore whether any specific subclonal SCNAs were associated with immune depletion or activation, we identified cytobands that were subclonally lost or gained in more than 30 tumours in the pan-cancer multi-sample cohort (Extended Data Fig. 4c) and investigated whether specific SCNAs were associated with changes in *TCRA* T cell fraction. Subclonal loss of 12q24.31–32 was found to be significantly associated with decreased *TCRA* T cell fraction ($P = 5.9 \times 10^{-6}$, ES = 0.75; Fig. 2f).

RNA-seq analysis of the TRACERx100 cohort identified *SPPL3* as exhibiting the most significant differential expression between samples with and without subclonal 12q24.31–32 loss (Extended Data Fig. 4d). The absence of *SPPL3* has been found to augment B3GNT5 enzyme activity, which upregulates cell surface glycosphingolipids that, in turn, impede class I HLA function and diminish CD8+ T cell activation[25]. Thus, these data suggest that subclonal loss of 12q24.31, encompassing *SPPL3*, may be selected in tumour evolution across cancer types (occurring in 18.7% of tumours within the cohort) as a mechanism of immune evasion.

T cell fraction is prognostic in LUAD

To explore the clinical utility of T cell ExTRECT, we considered whether the *TCRA* T cell fraction was prognostic in the TRACERx100 non-small cell lung cancer (NSCLC) cohort[13]. We categorized tumour samples as either hot or cold depending on whether the *TCRA* T cell fraction was greater than or equal to the mean in the cohort (0.081). In LUAD, we observed that patients harbouring an elevated number of immune-cold

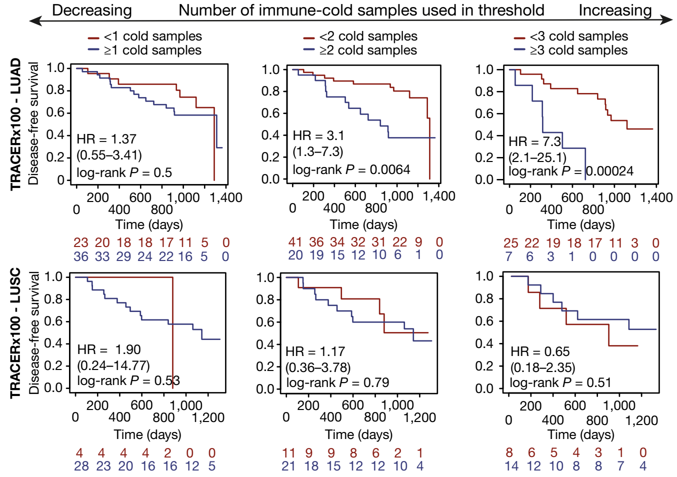


Fig. 3 | Prognostic value of *TCRA* T cell fraction for LUAD but not for LUSC. TRACERx100 multi-sample LUAD (top) and LUSC (bottom) Kaplan–Meier curves according to the number of immune-cold samples in the tumour. Immune-hot and immune-cold samples are defined using the mean of all tumour samples (0.08095) as threshold. Patients in Kaplan-Meier analyses were restricted to those with total samples greater than the number of immune-cold samples used to define the threshold. The numbers of surviving patients are shown below the graphs.

tumour samples were associated with significantly inferior prognosis (LUAD: ≥2 immune-cold samples, hazard ratio (HR) = 3.1, P = 0.0063, log-rank test; LUAD: ≥3 immune-cold samples, HR = 7.3, P = 0.00024, log-rank test; Fig. 3). By contrast, in patients with LUSC, there was no significant difference in survival. Using the median (0.074) as a threshold for immune hot or cold samples yielded similar results (Extended Data Fig. 5a). These results are consistent with previous analysis based on TIL scores inferred from computational pathology on the TRACERx100 cohort[26]. An association between the high *TCRA* T cell fraction and positive outcome was also observed in the TCGA LUAD cohort (overall survival: HR = 0.61, P = 0.0043; progression free survival: HR = 0.67, P = 0.016; Extended Data Fig. 5b), but not in the TCGA LUSC cohort (Extended Data Fig. 5c). A range of possible thresholds yielded similar results (Extended Data Fig. 5d).

Consistent with the importance of the tumour sample with the lowest immune infiltrate[26], the minimum, but not the maximum or mean, *TCRA* T cell fraction across tumour samples was prognostic in the TRACERx100 cohort. Other continuous measures, such as a *TCRA* T cell fraction divergence between tumour sample score (LUAD: HR = 2.2, P = 0.023, log-rank test; Extended Data Fig. 5d) and a model combining both the minimum and maximum scores (LUAD and LUSC: minimum HR = 0.5, P = 0.005, maximum HR = 1.5, P = 0.061; LUAD: minimum HR = 0.36, P = 0.016, maximum HR = 2.52, P = 0.029; Extended Data Fig. 5e) reached significance, suggesting that there is added predictive potential when considering the heterogeneity of the *TCRA* T cell fraction.

T cell fraction and response to CPIs

To further explore the clinical utility of T cell ExTRECT, we evaluated its ability to predict clinical response to CPIs. The CPI1000+ cohort[2] consists of 1,070 CPI-treated tumours receiving either anti-CTLA-4, anti-PD-L1 or anti-PD-1 therapy across eight main cancer types (Extended Data Fig. 6a, b). A responder was defined as a patient showing a complete or partial response, whereas a non-responder was defined as a patient exhibiting stable or progressive disease, as determined on the basis of imaging by RECIST criteria[27].

Consistent with the importance of T cells in influencing the response to CPIs, we observed a significantly higher ($P = 2.3 \times 10^{-7}$, ES = 0.17; Fig. 4a) tumour *TCRA* T cell fraction in responders. Similarly, immune-cold tumours (tumours with *TCRA* T cell fraction below 0.018, the median *TCRA* T cell fraction) were significantly enriched for non-responders (odds ratio (OR) = 2.14, $P = 1.33 \times 10^{-6}$, Fisher's exact test; Fig. 4b).

Separating the cohort by the medians for both clonal tumour mutational burden (TMB) and *TCRA* T cell fraction revealed that the association between *TCRA* T cell fraction and clinical response was independent of clonal TMB (Fig. 4b).

To evaluate the utility of T cell ExTRECT in comparison to RNA-seq-based measurements, we selected all studies with at least 10 samples from a cancer type with both RNA-seq and *TCRA* T cell fractions for univariate meta-analyses (557 patients across 7 studies and 5 cancer types; Fig. 4c). *TCRA* T cell fraction (OR = 1.39, P = 0.00858), clonal TMB (OR = 1.59, $P = 6.021 \times 10^{-5}$) and *CD8A* expression (OR = 1.45, P = 0.0004479) were all significantly associated with the response to CPIs.

To assess whether tumour *TCRA* T cell fraction improves prediction of the response to CPIs beyond clonal TMB and to a greater extent than *CD8A* expression, we evaluated different linear models (Extended Data Fig. 6c). Only the clonal TMB plus *TCRA* model was significantly better compared with clonal TMB alone (P = 0.0028, receiver operating characteristic test; general linear model (GLM): clonal TMB plus *TCRA*, area under the curve (AUC) = 0.68; GLM: clonal TMB, AUC = 0.62). When examining the significance of the variables in all models, *TCRA* T cell fraction was more significant than *CD8A* (GLM: clonal TMB plus *TCRA*, $P = 4.62 \times 10^{-5}$; GLM: clonal TMB plus *CD8A*, P = 0.000431), and when combined into a multivariable model, *TCRA* T cell fraction remained significant, but *CD8A* expression did not (*TCRA*, P = 0.00601; *CD8A*, P = 0.06246).

Finally, we assessed the predictive potential of the *TCRA* T cell fraction in a combined NSCLC CPI cohort (Extended Data Fig. 6d, e) lacking any RNA-seq immune measures. In univariate analyses (Fig. 4d), *TCRA* T cell fraction (OR = 1.44, P = 0.0071) and blood *TCRA* T cell fraction (OR = 1.39, P = 0.015) were significantly associated with response to CPI. The tumour *TCRA* T cell fraction had OR values greater than one in two out of three cohorts, whereas the blood *TCRA* T cell fraction had OR values greater than one in all three cohorts.

Together, these results suggest that the *TCRA* T cell fraction can be used as a substitute for RNA-seq measures of CD8+ infiltrate and that *TCRA* T cell fraction estimation adds prognostic value to TMB estimates.

Discussion

In summary, we present T cell ExTRECT, a method by which DNA WES can be harnessed to study the immune microenvironment. T cell ExTRECT provides an accurate estimate of immune infiltrate that shows clinical utility. We find that tumour *TCRA* T cell fraction is prognostic in LUAD and validate this in the TCGA LUAD cohort. Relatedly, we find the *TCRA* T cell fraction is associated with response to CPI in a pan-cancer cohort and improves upon the predictive value of clonal TMB. T cell ExTRECT enables the T cell fraction to be calculated in any WES sample. Using this information, we demonstrate that the T cell fraction in blood is heterogeneous, associated with microbial infections and significantly higher in females than males in TRACERx100 data from patients with NSCLC, consistent with previous findings[28,29]. Our analysis of blood samples in the lung CPI cohort revealed that the blood *TCRA* T cell fraction is predictive of the response to immunotherapy.

The T cell ExTRECT method has limitations. Although the tool quantifies the proportion of T cells in a sample, it cannot distinguish neoantigen-reactive T cells from bystander T cells, and is unable to detect clonotypes. Further, T cell ExTRECT loses fidelity at sequencing depths of less than 30×. Nevertheless, this relatively low sequencing depth means that it should be applicable to most DNA-sequencing datasets. T cell ExTRECT has so far been optimised only for WES, but further work will extend the method to whole-genome sequencing and to other species, including widely studied model organisms. T cell

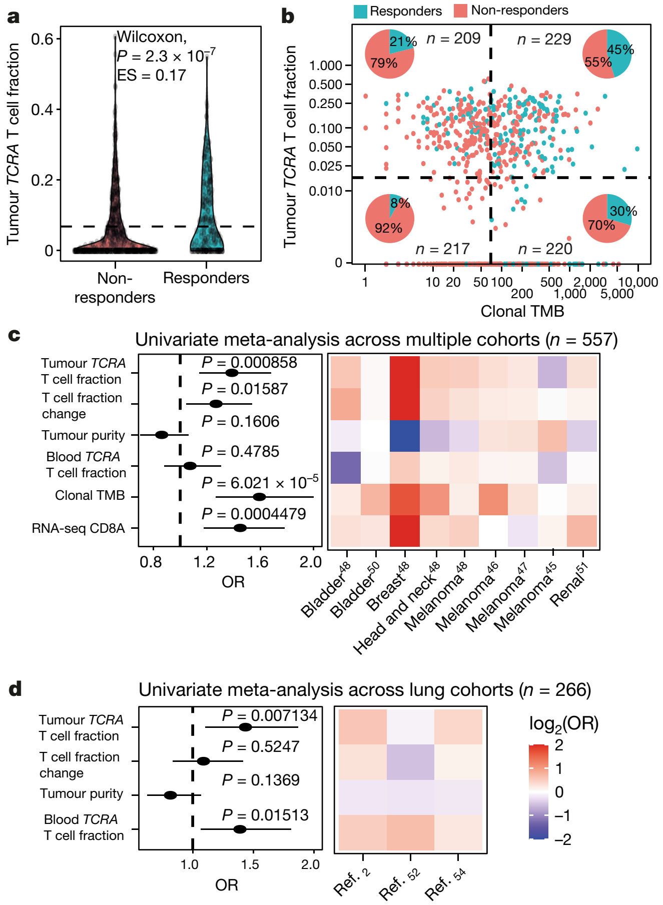


Fig. 4 | *TCRA* T cell fraction is predictive of survival and response to immunotherapy. a, Violin plot showing the tumour *TCRA* T cell fractions for non-responders versus responders across the CPI1000+ cohort. The dotted black line shows mean *TCRA* T cell fraction (0.067). **b**, Tumour *TCRA* T cell fraction versus clonal TMB. Dashed lines divide cohort into four quadrants with high and low clonal TMB, and immune-hot and immune-cold tumours, separated by the median values. Inset pie charts indicate the percentage of patients demonstrating CPI responses. **c**, Univariate meta-analysis of predictors of CPI responses across multiple cohorts with at least ten patients with each cancer type and both DNA and RNA-seq data. Left, forest plot of OR values from different clinical factors with associated *P*-values for predictive value of response. Right, heat map of OR values across individual studies from the CPI1000+ dataset, focusing on cohorts with both RNA-seq and *TCRA* T cell fraction. **d**, Univariate meta-analysis across three CPI lung datasets with DNA data but no RNA-seq data. Wilcoxon in boxplots refers to Wilcoxon rank-sum (Mann-Whitney *U*) test.

ExTRECT has clear applications in the immuno-oncological exploration of tumour samples, however it could also be used in a wider clinical setting, such as screening for severe combined immunodeficiency disease in newborns[30].

In summary, T cell ExTRECT could have important applications in both basic and translational research by providing a cost-effective technique to characterize immune infiltrate alongside somatic changes without the need for RNA-seq analysis.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-021-03894-5.

1. Rosenthal, R. et al. Neoantigen-directed immune escape in lung cancer evolution. *Nature* **567**, 479–485 (2019).
2. Litchfield, K. et al. Meta-analysis of tumor- and T cell-intrinsic mechanisms of sensitization to checkpoint inhibition. *Cell* **184**, 596–614 (2021).
3. Robert, C. et al. Ipilimumab plus dacarbazine for previously untreated metastatic melanoma. *N. Engl. J. Med.* **364**, 2517–2526 (2011).
4. Schadendorf, D. et al. Pooled analysis of long-term survival data from phase II and phase III trials of ipilimumab in unresectable or metastatic melanoma. *J. Clin. Oncol.* **33**, 1889–1894 (2015).
5. Goodman, A. M. et al. Tumor mutational burden as an independent predictor of response to immunotherapy in diverse cancers. *Mol. Cancer Ther.* **16**, 2598–2608 (2017).
6. Van Loo, P. et al. Allele-specific copy number analysis of tumors. *Proc. Natl Acad. Sci. USA* **107**, 16910–16915 (2010).
7. Favero, F. et al. Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann. Oncol.* **26**, 64–70 (2015).
8. Shen, R. & Seshan, V. *FACETS: Fraction and Allele-Specific Copy Number Estimates from Tumor Sequencing, Dept. Epidemiology and Biostatistics Working Paper Series* Vol. 1 No. 50 (Memorial Sloan-Kettering Cancer Center, 2015).
9. Carter, S. L. et al. Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30**, 413–421 (2012).
10. López, S. et al. Interplay between whole-genome doubling and the accumulation of deleterious alterations in cancer evolution. *Nat. Genet.* **52**, 283–293 (2020).
11. Ghandi, M. et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* **569**, 503–508 (2019).
12. Levy, E. et al. Immune DNA signature of T-cell infiltration in breast tumor exomes. *Sci. Rep.* **6**, 30064 (2016).
13. Jamal-Hanjani, M. et al. Tracking the evolution of non–small-cell lung cancer. *N. Engl. J. Med.* **376**, 2109–2121 (2017).
14. Danaher, P. et al. Pan-cancer adaptive immune resistance as defined by the tumor inflammation signature (TIS): results from The Cancer Genome Atlas (TCGA). *J. Immunother. Cancer* **6**, 1–17 (2018).
15. Davoli, T., Uno, H., Wooten, E. C. & Elledge, S. J. Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. *Science* **355**, eaaf8399 (2017).
16. Aran, D., Hu, Z. & Butte, A. J. xCell: Digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol.* **18**, 1–14 (2017).
17. Li, T. et al. TIMER: A web server for comprehensive analysis of tumor-infiltrating immune cells. *Cancer Res.* **77**, e108–e110 (2017).
18. Newman, A. M. et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 (2015).
19. Racle, J., de Jonge, K., Baumgaertner, P., Speiser, D. E. & Gfeller, D. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *eLife* **6**, e26476 (2017).
20. Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519–525 (2012).
21. Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550 (2014).
22. Yokoyama, A. et al. Age-related remodelling of oesophageal epithelia by mutated cancer drivers. *Nature* **565**, 312–317 (2019).
23. Poore, G. D. et al. Microbiome analyses of blood and tissues suggest cancer diagnostic approach. *Nature* **579**, 567–574 (2020).
24. Watkins, T. B. K. et al. Pervasive chromosomal instability and karyotype order in tumour evolution. *Nature* **587**, 126–132 (2020).
25. Jongsma, M. L. M. et al. The SPPL3-defined glycosphingolipid repertoire orchestrates HLA class I-mediated immune responses. *Immunity* **54**, 132-150.e9 (2021).
26. AbdulJabbar, K. et al. Geospatial immune variability illuminates differential evolution of lung adenocarcinoma. *Nat. Med.* **26**, 1054–1062 (2020).
27. Schwartz, L. H. et al. RECIST 1.1—update and clarification: from the RECIST committee. *Eur. J. Cancer* **62**, 132–137 (2016).
28. Conforti, F. et al. Sex-based dimorphism of anticancer immune response and molecular mechanisms of immune evasion. *Clin. Cancer Res.* **27**, https://doi.org/10.1158/1078-0432.CCR-21-0136 (2021).
29. Capone, I., Marchetti, P., Ascierto, P. A., Malorni, W. & Gabriele, L. Sexual dimorphism of immune responses: a new perspective in cancer immunotherapy. *Front. Immunol.* **9**, 552 (2018).
30. van der Spek, J., Groenwold, R. H. H., van der Burg, M. & van Montfrans, J. M. TREC based newborn screening for severe combined immunodeficiency disease: a systematic review. *J. Clin. Immunol.* **35**, 416–430 (2015).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

TRACERx Consortium

Charles Swanton[2,4,8], Mariam Jamal-Hanjani[2,8,9], Nicholas McGranahan[1,2], Carlos Martínez-Ruiz[1,2], Robert Bentham[1,2,52], Kevin Litchfield[2,3,52], Emilia L. Lim[2,4], Crispin T. Hiley[2,4], David A. Moore[2,7,8], Thomas B. K. Watkins[4,52], Rachel Rosenthal[4], Maise Al Bakir[4], Roberto Salgado[5,6], Nicolai J. Birkbak[10], Mickael Escudero[10], Aengus Stewart[10], Andrew Rowan[10], Jacki Goldman[10], Peter Van Loo[10], Richard Kevin Stone[10], Tamara Denner[10], Emma Nye[10], Sophia Ward[10], Stefan Boeing[10], Maria Greco[10], Jerome Nicod[10], Clare Puttick[10], Katey Enfield[10], Emma Colliver[10], Brittany Campbell[10], Alexander M. Frankell[10], Daniel Cook[10], Mihaela Angelova[10], Alastair Magness[10], Chris Bailey[10], Antonia Toncheva[10], Krijn Dijkstra[10], Judit Kisistok[10], Mateo Sokac[10], Oriol Pich[10], Jonas Demeulemeester[10], Elizabeth Larose Cadieux[10], Carla Castignani[10], Krupa Thakkar[10], Hongchang Fu[10],

Takahiro Karasaki[10,11], Othman Al-Sawaf[10,11], Mark S. Hill[10,12], Christopher Abbosh[11], Yin Wu[11], Selvaraju Veeriah[11], Robert E. Hynds[11], Andrew Georgiou[11], Mariana Werner Sunderland[11], James L. Reading[11], Sergio A. Quezada[11], Karl S. Peggs[11], Teresa Marafioti[11], John A. Hartley[11], Helen L. Lowe[11], Leah Ensell[11], Victoria Spanswick[11], Angeliki Karamani[11], Dhruva Biswas[11], Stephan Beck[11], Olga Chervova[11], Miljana Tanic[11], Ariana Huebner[11], Michelle Dietzen[11], James R. M. Black[11], Cristina Naceur-Lombardelli[11], Mita Afroza Akther[11], Haoran Zhai[11], Nnennaya Kanu[11], Simranpreet Summan[11], Francisco Gimeno-Valiente[11], Kezhong Chen[11], Elizabeth Manzano[11], Supreet Kaur Bola[11], Ehsan Ghorani[11], Marc Robert de Massy[11], Elena Hoxha[11], Emine Hatipoglu[11], Benny Chain[11], David R. Pearce[11], Javier Herrero[11], Simone Zaccaria[11], Jason Lester[13], Fiona Morgan[14], Malgorzata Kornaszewska[14], Richard Attanoos[14], Haydn Adams[14], Helen Davies[14], Jacqui A. Shaw[15], Joan Riley[15], Lindsay Primrose[15], Dean Fennell[15,16], Apostolos Nakas[16], Sridhar Rathinam[16], Rachel Plummer[16], Rebecca Boyles[16], Mohamad Tufail[16], Amrita Bajaj[16], Jan Brozik[16], Keng Ang[16], Mohammed Fiyaz Chowdhry[16], William Monteiro[17], Hilary Marshall[17], Alan Dawson[18], Sara Busacca[18], Domenic Marrone[18], Claire Smith[18], Girija Anand[19], Sajid Khan[19], Gillian Price[20], Mohammed Khalil[20], Keith Kerr[20], Shirley Richardson[20], Heather Cheyne[20], Joy Miller[20], Keith Buchan[20], Mahendran Chetty[20], Sylvie Dubois-Marshall[20], Sara Lock[21], Kayleigh Gilbert[21], Babu Naidu[22], Gerald Langman[22], Hollie Bancroft[22], Salma Kadiri[22], Gary Middleton[22], Madava Djearaman[22], Aya Osman[22], Helen Shackleford[22], Akshay Patel[22], Angela Leek[23], Nicola Totten[23], Jack Davies Hodgkinson[23], Jane Rogan[23], Katrina Moore[23], Rachael Waddington[23], Jane Rogan[23], Raffaele Califano[24], Rajesh Shah[24], Piotr Krysiak[24], Kendadai Rammohan[24], Eustace Fontaine[24], Richard Booton[24], Matthew Evison[24], Stuart Moss[24], Juliette Novasio[24], Leena Joseph[24], Paul Bishop[24], Anshuman Chaturvedi[24], Helen Doran[24], Felice Granato[24], Vijay Joshi[24], Elaine Smith[24], Angeles Montero[24], Philip Crosbie[24,25,26], Fiona Blackhall[26,27], Lynsey Priest[26,27], Matthew G. Krebs[26,27], Caroline Dive[26,28], Dominic G. Rothwell[26,28], Alastair Kerr[26,28], Elaine Kilgour[26,28], Katie Baker[27], Mathew Carter[27], Colin R. Lindsay[27], Fabio Gomes[27], Jonathan Tugwood[28], Jackie Pierce[28], Alexandra Clipson[28], Roland Schwarz[29,30], Tom L. Kaufmann[31,32], Matthew Huska[29], Zoltan Szallasi[33], Istvan Csabai[34], Miklos Diossy[34], Hugo Aerts[35,36], Charles Fekete[36], Gary Royle[37], Catarina Veiga[37], Marcin Skrzypski[38], David Lawrence[12], Martin Hayward[12], Nikolaos Panagiotopoulos[12], Robert George[12], Davide Patrini[12], Mary Falzon[12], Elaine Borg[12], Reena Khiroya[12], Asia Ahmed[12], Magali Taylor[12], Junaid Choudhary[12], Sam M. Janes[12], Martin Forster[12], Tanya Ahmad[12], Siow Ming Lee[12], Neal Navani[12], Dionysis Papadatos-Pastos[12], Marco Scarci[12], Pat Gorman[12], Elisa Bertoja[12], Robert C. M. Stephens[12], Emilie Martinoni Hoogenboom[12], James W. Holding[12], Steve Bandula[12], Ricky Thakrar[12], Radhi Anand[12], Kayalvizhi Selvaraju[12], James Wilson[12], Sonya Hessey[12], Paul Ashford[12], Mansi Shah[12], Marcos Vasquez Duran[12], Mairead MacKenzie[39], Maggie Wilcox[39], Allan Hackshaw[40], Yenting Ngai[40], Abigail Sharp[40], Cristina Rodrigues[40], Oliver Pressey[40], Sean Smith[40], Nicole Gower[40], Harjot Kaur Dhanda[40], Kitty Chan[40], Sonal Chakraborty[40], Christian Ottensmeier[41], Serena Chee[41], Benjamin Johnson[41], Aiman Alzetani[41], Judith Cave[41], Lydia Scarlett[41], Emily Shaw[41], Eric Lim[42], Paulo De Sousa[42], Simon Jordan[42], Alexandra Rice[42], Hilgardt Raubenheimer[42], Harshil Bhayani[42], Morag Hamilton[42], Lyn Ambrose[42], Anand Devaraj[42], Hema Chavan[42], Sofina Begum[42], Silviu I. Buderi[42], Daniel Kaniu[42], Mpho Malima[42], Sarah Booth[42], Andrew G. Nicholson[42], Nadia Fernandes[42], Christopher Deeley[42], Pratibha Shah[42], Chiara Proli[42], Kelvin Lau[43], Michael Sheaff[43], Peter Schmid[43], Louise Lim[43], John Conibear[43], Madeleine Hewish[44], Sarah Danson[45], Jonathan Bury[45], John Edwards[45], Jennifer Hill[45], Sue Matthews[45], Yota Kitsanta[45], Jagan Rao[45], Sara Tenconi[45], Laura Socci[45], Kim Suvarna[45], Faith Kibutu[45], Patricia Fisher[45], Robin Young[45], Joann Barker[45], Fiona Taylor[45], Kirsty Lloyd[45], Michael Shackcloth[46], Julius Asante-Siaw[46], John Gosney[47], Teresa Light[48], Tracey Horey[48], Peter Russell[48], Dionysis Papadatos-Pastos[48], Kevin G. Blyth[49], Craig Dick[49], Andrew Kidd[49], Alan Kirk[50], Mo Asif[50], John Butler[50], Rocco Bilancia[50], Nikos Kostoulas[50], Mathew Thomas[50] & Gareth A. Wilson[51]

[10]The Francis Crick Institute, London, UK. [11]University College London Cancer Institute, London, UK. [12]University College London Hospitals, London, UK. [13]Swansea Bay University Health Board, Swansea, UK. [14]Cardiff and Vale University Health Board, Cardiff, UK. [15]Cancer Research Centre, University of Leicester, Leicester, UK. [16]Leicester University Hospitals, Leicester, UK. [17]National Institute for Health Research Leicester Respiratory Biomedical Research Unit, Leicester, UK. [18]University of Leicester, Leicester, UK. [19]Barnet & Chase Farm Hospitals, Barnet, UK. [20]Aberdeen Royal Infirmary, Aberdeen, UK. [21]The Whittington Hospital NHS Trust, London, UK. [22]University Hospital Birmingham NHS Foundation Trust, Birmingham, UK. [23]Manchester Cancer Research Centre Biobank, Manchester, UK. [24]Wythenshawe Hospital, Manchester University NHS Foundation Trust, Manchester, UK. [25]Division of Infection, Immunity and Respiratory Medicine, University of Manchester, Manchester, UK. [26]Cancer Research UK Lung Cancer Centre of Excellence, University of Manchester, Manchester, UK. [27]Christie NHS Foundation Trust, Manchester, United Kingdom. [28]Cancer Research UK Manchester Institute, University of Manchester, Manchester, UK. [29]Berlin Institute for Medical Systems Biology, Max Delbrueck Center for Molecular Medicine, Berlin, Germany. [30]German Cancer Consortium (DKTK), partner site Berlin, Berlin, Germany. [31]Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC), Berlin, Germany. [32]BIFOLD, Berlin Institute for the Foundations of Learning and Data, Berlin, Germany. [33]Danish Cancer Society Research Center, Copenhagen, Denmark. [34]Department of Physics of Complex Systems, ELTE Eötvös Loránd University, Budapest, Hungary. [35]Artificial Intelligence in Medicine (AIM) Program, Mass General Brigham, Harvard Medical School, Boston, MA, USA. [36]Radiology and Nuclear Medicine, CARIM & GROW, Maastricht University, Maastricht, The Netherlands. [37]Department of Medical Physics and Bioengineering, University College London Cancer Institute, London, UK. [38]Department of Oncology and Radiotherapy, Medical University of Gdańsk, Gdańsk, Poland. [39]Independent Cancer Patients Voice, London, UK. [40]Cancer Research UK and UCL Cancer Trials Centre, London, UK. [41]University Hospital Southampton NHS Foundation Trust, Southampton, UK. [42]Royal Brompton and Harefield NHS Foundation Trust, London, UK. [43]Barts Health NHS Trust, London, UK. [44]Ashford and St Peter's Hospitals NHS Foundation Trust, Chertsey, UK. [45]Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield, UK. [46]Liverpool Heart and Chest Hospital NHS Foundation Trust, Liverpool, UK. [47]Royal Liverpool University Hospital, Liverpool, UK. [48]The Princess Alexandra Hospital NHS Trust, Harlow, UK. [49]NHS Greater Glasgow and Clyde, Glasgow, UK. [50]Golden Jubilee National Hospital, Clydebank, UK. [51]Achilles Therapeutics UK Limited, London, UK.

Methods

A full description of the T cell ExTRECT method is given in Supplementary Information.

Statistics

All statistical tests were performed in R 3.6.1. No statistical methods were used to predetermine sample size. Tests involving correlations were done using stat_cor from the R package ggpubr (v0.4.0) with the Spearman's method. Tests involving comparisons of distributions were done using stat_compare_means using wilcox.test using either the unpaired option, performing a Wilcoxon rank-sum (Mann–Whitney *U*) test, or a paired Wilcoxon signed-rank test. Effect sizes for the corresponding Wilcoxon tests were measured using the wilcox_effsize function from the rstatix package (v0.6.0). Hazard ratios and *P* values were calculated with the survival package (v3.2–3) for both Kaplan–Meier curves and the Cox proportional hazard model. For all statistical tests, the number of data points included are plotted or annotated in the corresponding figure. Plotting and analysis in R also made use of the ggplot2 (v3.3.3), dplyr (v1.0.4), tidyr (v1.1.1), gridExtra (v2.3) and gtable (v0.3.0) packages.

Fresh frozen versus FFPE samples

To test that the *TCRA* T cell fraction was reliable and consistent for both fresh frozen and formalin-fixed paraffin-embedded (FFPE) samples, the non-GC-corrected *TCRA* T cell fractions were calculated for six different studies within the CPI1000+ cohort. Three of these studies used WES derived from FFPE tissues ($n = 460$), while the other three utilised WES samples derived from fresh frozen tissue ($n = 357$).

Fitting a linear model to predict *TCRA* T cell fraction by histology and FFPE status (Extended Data Fig. 1i) revealed that cancer type was the main driver of this significance, with FFPE status not being significant. Additionally, for melanoma and bladder tumours that had FFPE and fresh frozen WES samples, no significant difference was found (Extended Data Fig. 1f). This led us to conclude that whether a WES sample is derived from fresh frozen or FFPE tissue does not significantly affect the values of the *TCRA* T cell fraction calculated by T cell ExTRECT.

Calculation of CDR3 V(D)J scores

The procedure outlined in Levy et al.[12] was followed to calculate the CDR3 V(D)J scores. First reads aligning to *TCRB* (hg19:chr7:142000817-142510993) and unaligned reads were extracted with samtools. This resulting bam file was converted to fastq using bedtools and then the tool IMSEQ (v1.1.0)[31] was used on the resulting output to identify V(D)J recombinant reads aligning to the CDR3 region, the number of aligned reads was then normalized by the total number of reads in the original bam file (as measured by samtools flagstat) to create the CDR3 V(D)J scores.

Kraken TCGA analysis

Pre-processed microbiome data output from the Kraken[32] analysis performed by Poore et al.[23] was downloaded from ftp://ftp.microbio.me/pub/cancer_microbiome_analysis/.

To create the high and low Kraken microbiome groups for both the blood and tumour samples, the file Kraken-TCGA-Voom-SNM-Most-Stringent-Filtering-Data.csv containing normalised logCPM values was downloaded. For each sample, the rows were summed giving an overall 'microbiome' score. The samples were then divided into high and low groups based on the median of this score.

To investigate the role of any individual microbial species in influencing *TCRA* T cell fraction, a reduced list of the species from the Kraken-TCGA-Voom-SNM-Most-Stringent-Filtering-Data.csv file was selected by removing all species with less than 1,000 total raw reads in the TCGA LUAD and LUSC cohort as called from the raw data file Kraken-TCGA-Raw-Data-17625-Samples.csv. This left a total of 59 microbial species that were individually tested for association with *TCRA* T cell fraction using Spearman's correlation for both LUAD and LUSC blood and tumour samples.

Patients from TRACERx100

The first 100 patients prospectively analysed by the NSCLC TRACERx study (https://clinicaltrials.gov/ct2/show/NCT01888601, approved by an independent research ethics committee, 13/LO/1546) were used in this study. This is identical to the 100-patient cohort originally described in Jamal-Hanjani et al.[13].

In brief, informed consent was a mandatory requirement for entry into the TRACERx study. This NSCLC cohort consisted of 68 males and 32 females with a median age of 68. Finally, the cohort was predominantly made up of early-stage tumours (Ia (26), Ib (36), IIa (13), IIb (11), IIIa (13) and IIIb (1)) and 28 patients also had adjuvant therapy.

TRACERx100 WES and RNA-seq samples

Both WES (aligned to the hg19 sequence) and RNA-seq samples were obtained from the TRACERx study for the first 100 patients; the method for processing these samples is as previously described[13]. Notably, for the WES samples, exome capture was performed using a custom version of Agilent Human All Exome V5 kit according to the manufacturer's instructions.

TCGA LUAD and LUSC cohorts

Aligned BAM files (hg38 sequence) from the TCGA LUAD and LUSC cohorts were downloaded from the genomic data commons (dataset ID: phs000178.v10.p8). Sample purity and ploidy calls were generated from ASCAT (v2.4.2) from a previous analysis of the TCGA data[33]. In short, Affymetrix SNP6 profiles from paired tumour-normal samples (dataset ID: phs000178.v10.p8) were processed by PennCNV libraries[34] to obtain BAFs and log ratios which were GC-corrected before being processed with ASCAT[6].

Cancer cell line data

The non-T cell derived colorectal cancer cell lines HCT116 were sequenced with Illumina HiSeq 2500 and aligned with bwa mem using hg19 as described in López et al.[10]. The T cell-derived cell lines were from the dataset described in Ghandi et al.[11] and downloaded from the Sequence Read Archive (SRA) under accession number PRJNA523380. Cell lines derived from T cells were chosen ensuring that any cell line derived from precursor T cell acute lymphoblastic leukemia were excluded as these have not undergone V(D)J recombination. This process led to WES data from three cell lines being chosen: JURKAT, HPB-ALL and PEER.

Owing to the difficulty of running ASCAT without matching germline samples, the naive *TCRA* T cell fraction was used for all cell line work.

Multi-sample tumour cohort of patients

The multi-sample pan-cancer cohort (Extended data Fig. 4b) was created by combining the TRACERx cohort with a subset of the cohort presented recently by Watkins et al.[24]. Tumours were included if they had at least two samples sequenced in the primary tumour for which it was possible to calculate the *TCRA* T cell fraction using T cell ExTRECT. The final cohort therefore consisted of a multi-sample primary tumour dataset with the addition of any metastasis samples that were also sequenced for these patients.

Besides TRACERx100 the following datasets were combined into the final multi-sample pan-cancer cohort: (1) Brastianos et al.[35]—a cohort focused on studying brain metastasis originating from different histologies, only tumours with multi-sample primary samples from this cohort were included; (2) Gerlinger et al.[36,37]—a multi-sample primary cohort of patients with KIRC; (3) Harbst et al.[38]—a multi-region primary cohort of patients with SKCM; (4) Lamy et al.[39]—a multi-sample primary cohort of patients with BLCA; (5) Savas et al.[40]—a multi-sample cohort of patients with ER+ and triple-negative BRCA (BRCA ER+ and TNBC); (6) Suzuki et al.[41]—a multi-sample primary cohort of glioma; (7) Turajlic et al[42]—a multi-sample primary cohort of KIRC; and (8) Messaoudene et al.[43]—a multi-sample primary cohort of patients with HER2+ and ER+ BRCA.

Selection of samples for multi-sample sequencing in different datasets

In all of the multi-sample cohorts samples were selected though by different methods (see associated publications) with two main criteria in mind, first that tumour content be maximized at the expense of stromal content in order to assure good quality mutation and copy-number analysis for the main goal of the genomic analysis, and second, that each sample represents a physically separate and distinct part of the tumour. In cases where these were not at separate sites, different measures were used. In the TRACERx100 cohort, for example, samples sequenced were a minimum of 3 mm apart.

Identification of gain, loss and LOH events in a pan-cancer multi-sample cohort

Analysis of whole-exome sequencing was performed as described[13]. Copy-number segmentation, tumour purity and ploidy for each sample were estimated using ASCAT[6] as described previously[13]. These data were used as input to a multi-sample SCNA-estimation approach to produce genome-wide estimates of the presence of loss of heterozygosity as well as loss, neutral, gain and amplification copy-number states relative to sample ploidy. The log ratio values present in each copy-number segment with ≥5 log ratio values in all samples of a tumour were examined relative to three sample-ploidy-adjusted log ratio thresholds using one-tailed t-tests with a $P < 0.01$ threshold. These log ratio thresholds were equivalent to <log2[1.5/2] for losses, >log2[2.5/2] for gains in a diploid tumour. Any segment not classified as a loss or gain were classed as neutral. For each segment, these relative to ploidy definitions were combined with loss of heterozygosity detection across all samples from a single tumour.

Pairwise subclonal SCNA scores

To calculate pairwise subclonal SCNA measures, the classifications outlined in the previous methods section were used to create three groups of pairwise subclonal SCNA scores. First, we considered any segment affected by any of gain or loss relative to ploidy or LOH as aberrant and compared each pair of samples from a single patient's disease, classifying aberrant areas as clonal if aberrant in both samples or subclonal if aberrant in only one sample. This same process was repeated for gains relative to ploidy alone and then losses relative to ploidy and LOH considered together.

Cytoband-level SCNA analysis

To enable comparisons across tumours, segments were mapped to hg19 cytobands. If multiple segments mapped to a cytoband, the SCNA status (gain or loss relative to ploidy) of the segment with the largest overlap with the cytoband was chosen.

For the SCNA gain-and-loss analysis, cytoband level events were selected if they occurred subclonally across the entire cohort more than 30 times. Bands passing this threshold within the same region (for example, all cytobands on 1p36) were then grouped together. A Wilcoxon paired test was used to assess whether the tumour regions within a single patient with the subclonal SCNA events had a significant difference in *TCRA* T cell fraction to those regions without the event.

Selection of multi-sample tumours with heterogeneous immune infiltration

To be included a tumour had to have at least 3 samples sequenced and meet the following two requirements: (1) have a pair of regions with a large change in immune infiltration, defined as having ≥0.065 difference in *TCRA* T cell fraction, and (2) have a pair of samples with a small or no change in immune infiltration, defined as having <0.065 difference in *TCRA* T cell fraction. An example of a tumour matching this requirement would be one with three samples R1, R2 and R3 with *TCRA* T cell fractions of 0.01, 0.01 and 0.2, respectively. The R1–R2 pair has a difference in *TCRA* T cell fraction of 0, while the R1–R3 and R2–R3 pairs would both have a large difference of 0.19. Within the multi-sample tumour cohort, 76 patients matched these criteria.

RNA-seq differential gene-expression analysis for patients with subclonal 12q24.31–32 loss

Differential gene-expression analysis was performed on the TRACERx100 patients with RNA-seq data showing subclonal 12q24.31–32 loss. First, using R 4.0.0, the edgeR package (version 3.32.1) was used for sample-specific trimmed mean of the M-values (TMM) normalization; any genes with low expression were then filtered out using the standard edgeR filtering method before using the Limma–Voom method from the limma R package (version 3.46.0) to calculate the Voom fit and obtain P-values for the gene-expression differences. The comparison controlled for patient and histology as blocking factors and P-values were FDR-corrected for multiple testing. Results were then visualised with the R EnhancedVolcano package (version 1.8.0).

CPI1000+ meta-analysis of cohorts

The CPI1000+ cohort is fully described in Litchfield et al.[2] and contains the following datasets: (1) Snyder et al.[44], an advanced melanoma anti-CTLA-4-treated cohort; (2) Van Allen et al.[45], an advanced melanoma anti-CTLA-4-treated cohort; (3) Hugo et al.[46], an advanced melanoma anti-PD-1-treated cohort; (4) Riaz et al.[47], an advanced melanoma anti-PD-1-treated cohort; (5) Cristescu et al.[48], an advanced melanoma anti-PD-1-treated cohort; (6) Cristescu et al.[48], an advanced head and neck cancer anti-PD-1-treated cohort; (7) Cristescu et al.[48] 'all other tumour types' cohort (from KEYNOTE-028 and KEYNOTE-012 studies), treated with anti-PD-1; (8) Snyder et al.[49], a metastatic urothelial cancer anti-PD-L1-treated cohort; (9) Mariathasan et al.[50], a metastatic urothelial cancer anti-PD-L1-treated cohort; (10) McDermott et al.[51], a metastatic renal cell carcinoma anti-PD-L1-treated cohort; (11) Rizvi et al.[52], a NSCLC anti-PD-1-treated cohort; (12) Hellman et al., a cohort of NSCLC samples treated with anti-PD-1 used by Litchfield et al.[2]; (13) Le et al.[53],a colorectal cancer cohort treated with anti-PD-1 therapy.

Of these studies, Snyder et al.[49] was excluded from the analysis owing to extremely poor coverage within the *TCRA* gene. Additionally, 55 patients were either on treatment at the time of the biopsy or had prior treatment with CPIs and were thus removed from the analysis. All samples were aligned to hg19 using bwa mem (v0.7.15) with purity and SCNA data calculated using ASCAT as described in Litchfield et al.[2].

Notably, 953 out of 1,070 samples (89%) had WES data and 888 out of 1,070 (83%) had sufficient purity and coverage to enable copy number estimation, enabling the *TCRA* T cell fractions to be calculated. Some 643 out of 1,070 (60%) of these samples had matched RNA-seq data, allowing orthogonal assessment of T cell fractions.

For an extension to this dataset, a NSCLC anti-PD-1 treated cohort[54] was added for a specific NSCLC analysis. In this cohort mutations were called as either clonal or subclonal using PyClone, as described by Litchfield et al.[2].

Orthogonal immune measures

RNA-seq signatures. We used the method of Danaher et al.[12] as our primary method for estimating T cell content from RNA-seq measures, as it has previously been demonstrated that this is most strongly correlated to TIL scores calculated in TRACERx[1]. Other RNA-seq signatures tested against the *TCRA* T cell fractions were the Davoli method[15], xCell[16], TIMER[17] and EPIC[19] and CIBERSORT[18].

Histopathology-derived TIL scores. TILs were estimated as previously described[1] from histopathology slides using internationally established guidelines developed by the International Immuno-Oncology Biomarker Working Group[55]. In brief, the relative proportion of stromal area to tumour area was determined from a pathology slide of a given tumour sample. TILs were reported for the stromal compartment (percent stromal TILs). The denominator used to determine the percentage of stromal TILs was the area of stromal tissue (that is, the area occupied by mononuclear inflammatory cells over the total intratumoral stromal area) rather than the number of stromal cells (that is, the fraction of total stromal nuclei that represent mononuclear inflammatory cell nuclei). This method has been demonstrated to be reproducible among trained pathologists[56]. An inter-person concordance test was performed, and this demonstrated high reproducibility. The International Immuno-Oncology Biomarker Working Group has developed a freely available training tool to train pathologists for optimal TIL assessment on haematoxylin–eosin slides (www.tilsincancer.org).

Univariate and multivariable model for CPI response. For the univariate model, an adapted procedure[2] was followed, with the main difference being that only samples with complete data (RNA-seq for *CD8A*, clonal TMB and *TCRA* T cell fraction) were included. The univariate model meta-analysis was conducted using R package meta (version 4.13-0). The multivariable model was created with general linear models using the function glm from the stats R package using default values. The R package ROCR (version 1.0-11) was used for the ROC curve analysis.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The RNA-seq data, WES data and histopathology-derived TIL scores (in each case from the TRACERx study) generated, used or analysed during this study are not publicly available and restrictions apply to the availability of these data. Such RNA-seq, WES data and histopathology-derived TIL scores are available through the Cancer Research UK and University College London Cancer Trials Centre (ctc.tracerx@ucl.ac.uk) for academic non-commercial research purposes upon reasonable request, and subject to review of a project proposal that will be evaluated by a TRACERx data access committee, entering into an appropriate data access agreement and subject to any applicable ethical approvals. Details of all other datasets obtained from third parties used in this study can be found in Extended Data Table 1. Clinical trial information (if applicable) is also available in the associated publications described in Extended Data Table 1.

Code availability

The code used to produce *TCRA* T cell fraction scores is available for academic non-commercial research purposes at https://github.com/McGranahanLab/TcellExTRECT. All other code used in the analysis and to produce figures is available at https://github.com/McGranahanLab/T-cell-ExTRECT-figure-code-2021.

31. Kuchenbecker, L. et al. IMSEQ-A fast and error aware approach to immunogenetic sequence analysis. *Bioinformatics* **31**, 2963–2971 (2015).

32. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).

33. Middleton, G. et al. The National Lung Matrix Trial of personalized therapy in lung cancer. *Nature* **583**, 807–812 (2020).

34. Wang, K. et al. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **17**, 1665–1674 (2007).

35. Brastianos, P. K. et al. Genomic characterization of brain metastases reveals branched evolution and potential therapeutic targets. *Cancer Discov.* **5**, 1164–1177 (2015).

36. Gerlinger, M. et al. Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nat. Genet.* **46**, 225–233 (2014).

37. Gerlinger, M. et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* **366**, 883–892 (2012).

38. Harbst, K. et al. Multiregion whole-exome sequencing uncovers the genetic evolution and mutational heterogeneity of early-stage metastatic melanoma. *Cancer Res.* **76**, 4765–4774 (2016).

39. Lamy, P. et al. Paired exome analysis reveals clonal evolution and potential therapeutic targets in urothelial carcinoma. *Cancer Res.* **76**, 5894–5906 (2016).

40. Savas, P. et al. The subclonal architecture of metastatic breast cancer: results from a prospective community-based rapid autopsy program "CASCADE". *PLoS Med.* **13**, e1002204 (2016).

41. Suzuki, H. et al. Mutational landscape and clonal architecture in grade II and III gliomas. *Nat. Genet.* **47**, 458–468 (2015).

42. Turajlic, S. et al. Deterministic evolutionary trajectories influence primary tumor growth: TRACERx renal. *Cell* **173**, 595-610.e11 (2018).

43. Messaoudene, M. et al. T-cell bispecific antibodies in node-positive breast cancer: novel therapeutic avenue for MHC class I loss variants. *Ann. Oncol.* **30**, 934–944 (2019).

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

45. Van Allen, E. M. et al. Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science* **352**, 207–212 (2016).

46. Hugo, W. et al. Genomic and transcriptomic features of response to anti-PD-1 therapy in metastatic melanoma. *Cell* **165**, 35–44 (2016).

47. Riaz, N. et al. Tumor and microenvironment evolution during immunotherapy with nivolumab. *Cell* **171**, 934-949.e15 (2017).

48. Cristescu, R. et al. Pan-tumor genomic biomarkers for PD-1 checkpoint blockade-based immunotherapy. *Science* **362**, eaar3593 (2018).

49. Snyder, A. et al. Contribution of systemic and somatic factors to clinical response and resistance to PD-L1 blockade in urothelial cancer: an exploratory multi-omic analysis. *PLoS Med.* **14**, 1–24 (2017).

50. Mariathasan, S. et al. TGFβ attenuates tumour response to PD-L1 blockade by contributing to exclusion of T cells. *Nature* **554**, 544–548 (2018).

51. McDermott, D. F. 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).

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

53. Le, D. T. et al. PD-1 blockade in tumors with mismatch repair deficiency. *N. Engl. J. Med.* **372**, 2509–2520 (2015).

54. Shim, J. H. et al. HLA-corrected tumor mutation burden and homologous recombination deficiency for the prediction of response to PD-(L)1 blockade in advanced non-small-cell lung cancer patients. *Ann. Oncol.* **31**, 902–911 (2020).

55. Hendry, S. et al. Assessing tumor-infiltrating lymphocytes in solid tumors. *Adv. Anat. Pathol.* **24**, 235–251 (2017).

56. Denkert, C. et al. Standardized evaluation of tumor-infiltrating lymphocytes in breast cancer: Results of the ring studies of the international immuno-oncology biomarker working group. *Mod. Pathol.* **29**, 1155–1164 (2016).

Acknowledgements R.B. is supported by the NIHR BRC at University College London Hospitals. K.L. is funded by the UK Medical Research Council (MR/P014712/1 and MR/V033077/1), Rosetrees Trust and Cotswold Trust (A2437) and Cancer Research UK (C69256/A30194). T.B.K.W. is supported by the Francis Crick Institute, which receives its core funding from Cancer Research UK (FC001169), the UK Medical Research Council (FC001169) and the Wellcome Trust (FC001169) as well as the Marie Curie ITN Project PLOIDYNET (FP7-PEOPLE-2013, 607722), Breast Cancer Research Foundation (BCRF), Royal Society Research Professorships Enhancement Award (RP/EA/180007) and the Foulkes Foundation. E.L.L. receives funding from NovoNordisk Foundation (ID 16584). R.R. is supported by Royal Society Research Professorships Enhancement Award (RP/EA/180007). C.M.-R. is supported by Rosetrees. C.T.H. is supported by the NIHR BRC at University College London Hospitals. M.J.-H. has received funding from Cancer Research UK, National Institute for Health Research, Rosetrees Trust, UKI NETs and NIHR University College London Hospitals Biomedical Research Centre. N.M. is a Sir Henry Dale Fellow, jointly funded by the Wellcome Trust and the Royal Society (Grant Number 211179/Z/18/Z) and also receives funding from Cancer Research UK, Rosetrees and the NIHR BRC at University College London Hospitals and the CRUK University College London Experimental Cancer Medicine Centre. C.S. is a Royal Society Napier Research Professor. His work was supported by the Francis Crick Institute, which receives its core funding from Cancer Research UK (FC001169), the UK Medical Research Council (FC001169) and the Wellcome Trust (FC001169). C.S. is funded by Cancer Research UK (TRACERx, PEACE and CRUK Cancer Immunotherapy Catalyst Network), Cancer Research UK Lung Cancer Centre of Excellence, the Rosetrees Trust, Butterfield and Stoneygate Trusts, NovoNordisk Foundation (ID16584), Royal Society Research Professorships Enhancement Award (RP/EA/180007), the NIHR BRC at University College London Hospitals, the CRUK-UCL Centre, Experimental Cancer Medicine Centre and the Breast Cancer Research Foundation (BCRF). This research is supported by a Stand Up To Cancer-LUNGevity-American Lung Association Lung Cancer Interception Dream Team Translational Research Grant (SU2C-AACR-DT23-17). Stand Up To Cancer is a program of the Entertainment Industry Foundation. Research grants are administered by the American Association for Cancer Research, the Scientific Partner of SU2C. C.S. also receives funding from the European Research Council (ERC) under the European Union's Seventh Framework Programme (FP7/2007-2013) Consolidator Grant (FP7-THESEUS-617844), European Commission ITN (FP7-PloidyNet 607722), an ERC Advanced Grant (PROTEUS) from the European Research Council under the European Union's Horizon 2020 research and innovation programme (835297) and Chromavision from the European Union's Horizon 2020 research and innovation programme (665233). The TRACERx study (Clinicaltrials.gov no: NCT01888601) is sponsored by University College London (UCL/12/0279) and has been approved by an independent Research Ethics Committee (13/LO/1546). TRACERx is funded by Cancer Research UK (C11496/A17786) and coordinated through the Cancer Research UK and UCL Cancer Trials Centre. The results published here are based in part on data generated by The Cancer Genome Atlas pilot project established by the NCI and the National Human Genome Research Institute. The data were retrieved through the database of Genotypes and Phenotypes (dbGaP) authorization (accession number phs000178.v9.p8). Information about TCGA and the constituent investigators and institutions of the TCGA research network can be found at http://cancergenome.nih.gov/. This project was enabled through access to the MRC eMedLab Medical Bioinformatics infrastructure, supported by the Medical Research Council (MR/L016311/1). In particular, we acknowledge the support of the High-Performance Computing at the Francis Crick Institute as well as the UCL Department of Computer Science Cluster and the support team. In addition, this work was supported by the CRUK City of London Centre Award (C7893/A26233). We thank all investigators, funders and industry partners that supported the generation of the data within this study, as well as patients for their participation. Specifically, we thank Merck & Co, Genentech and Bristol-Myers Squibb for generating the industry datasets used in this study, and E. Van Allen, L. Diaz, T. A. Chan, L. A. Garraway, R. S. Lo, D. F. Bajorin, D. Schadendorf, T. Powles, S.-H. Lee, A. Ribas and S. Ogawa for academic datasets. This work has been funded by Merck and Co. We gratefully acknowledge the patients and their families, the investigators and site personnel who participated in the KEYNOTE-001, -006, -012 and -028 studies.

Author contributions R.B. helped conceive the study, designed and conducted the bioinformatic analysis, and wrote the manuscript. K.L. curated the CPI1000+ cohort used in the study and provided considerable bioinformatic support on its analysis. T.B.K.W. provided considerable bioinformatic support on the analysis of the multi-sample pan-cancer cohort and helped conceive the study and write the manuscript. T.B.K.W. and E.L.L jointly curated the multi-sample pan-cancer cohort used in the study. R.R. and C.M.-R. provided considerable bioinformatic support in the transcriptomic analysis performed in the study, providing RNA-seq immune score metrics and assisting with the RNA-seq gene-expression analysis respectively. R.S., M.A.B., D.A.M. and C.T.H. jointly analysed histopathology-derived TIL estimates. M.J.-H. designed study protocols and helped to analyse patient clinical characteristics. C.S. helped provide study supervision and helped direct the avenues of bioinformatics analysis and also gave feedback on the manuscript. N.M. conceived and supervised the study and helped write the manuscript.

Competing interests D.A.M. reports speaker fees from AstraZeneca. M.A.B. has consulted for Achilles Therapeutics. R.R. has consulted for and has stock options in Achilles Therapeutics. K.L. has a patent on indel burden and CPI response pending and speaker fees from Roche tissue diagnostics, research funding from CRUK TDL/Ono/LifeArc alliance, and a consulting role with Monopteros Therapeutics. C.T.H. has received speaker fees from AstraZeneca. M.J.-H. is a member of the Scientific Advisory Board and Steering Committee for Achilles Therapeutics. N.M. has stock options in and has consulted for Achilles Therapeutics and holds a European patent in determining HLA LOH (PCT/GB2018/052004). C.S. acknowledges grant support from Pfizer, AstraZeneca, Bristol Myers Squibb, Roche-Ventana, Boehringer-Ingelheim, Archer Dx Inc. (collaboration in minimal residual disease sequencing technologies) and Ono Pharmaceutical; is an AstraZeneca Advisory Board Member and Chief Investigator for the MeRmaiD1 clinical trial; has consulted for Amgen, Pfizer, Novartis, GlaxoSmithKline, MSD, Bristol Myers Squibb, AstraZeneca, Illumina, Genentech, Roche-Ventana, GRAIL, Medicxi, Bicycle Therapeutics, Metabomed and the Sarah Cannon Research Institute; has stock options in Apogen Biotechnologies, Epic Bioscience and GRAIL; and has stock options and is co-founder of Achilles Therapeutics. C.S. holds patents relating to assay technology to detect tumour recurrence (PCT/GB2017/053289); to targeting neoantigens (PCT/EP2016/059401), identifying patent response to immune checkpoint blockade (PCT/EP2016/071471), determining HLA LOH (PCT/GB2018/052004), predicting survival rates of patients with cancer (PCT/GB2020/050221), to treating cancer by targeting Insertion/deletion mutations (PCT/GB2018/051893); identifying insertion/deletion mutation targets (PCT/GB2018/051892); methods for lung cancer detection (PCT/US2017/028013); and identifying responders to cancer treatment (PCT/GB2018/051912).

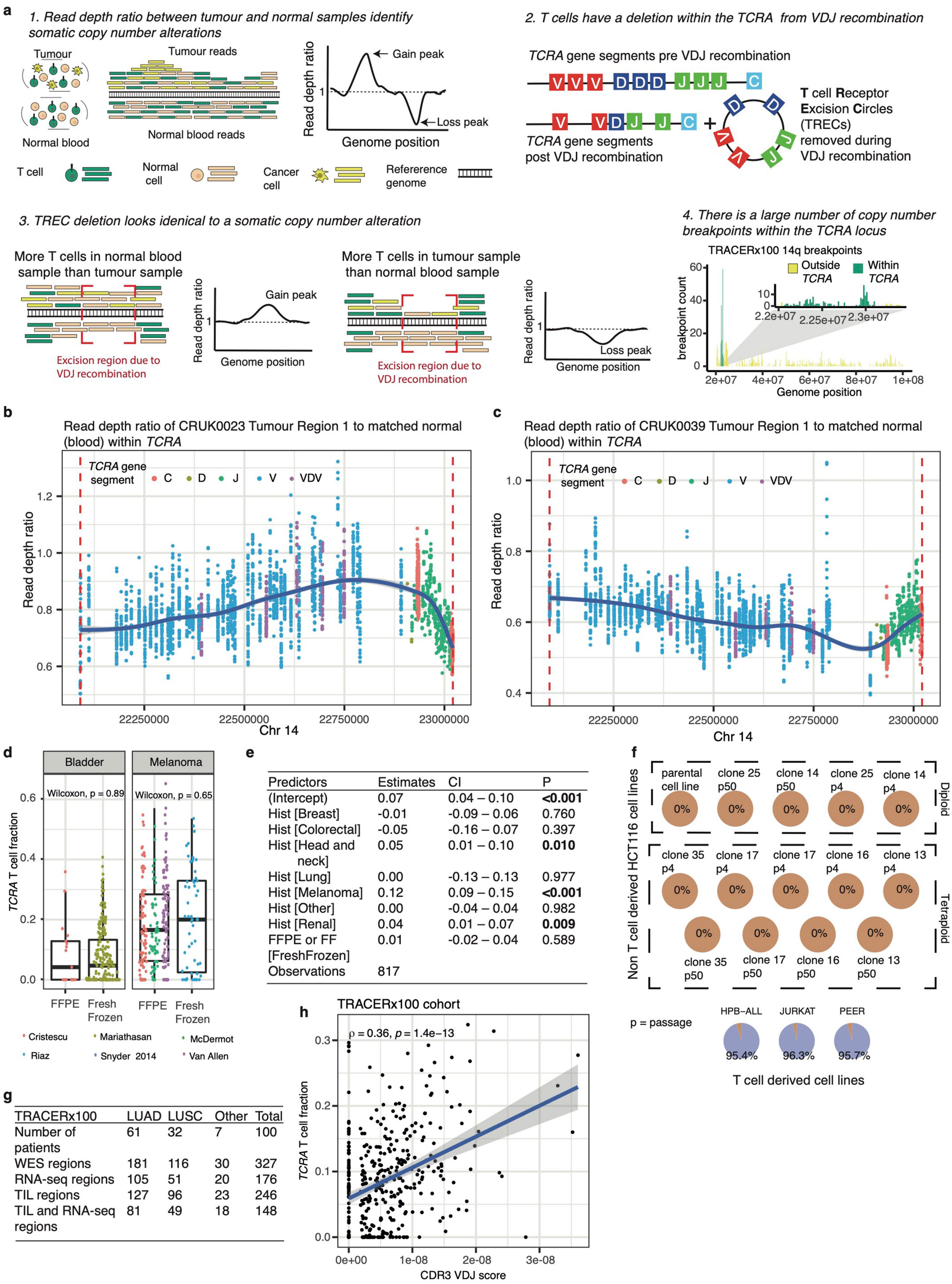
Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-021-03894-5.

Correspondence and requests for materials should be addressed to Nicholas McGranahan.

Peer review information *Nature* thanks Florian Markowetz and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints.

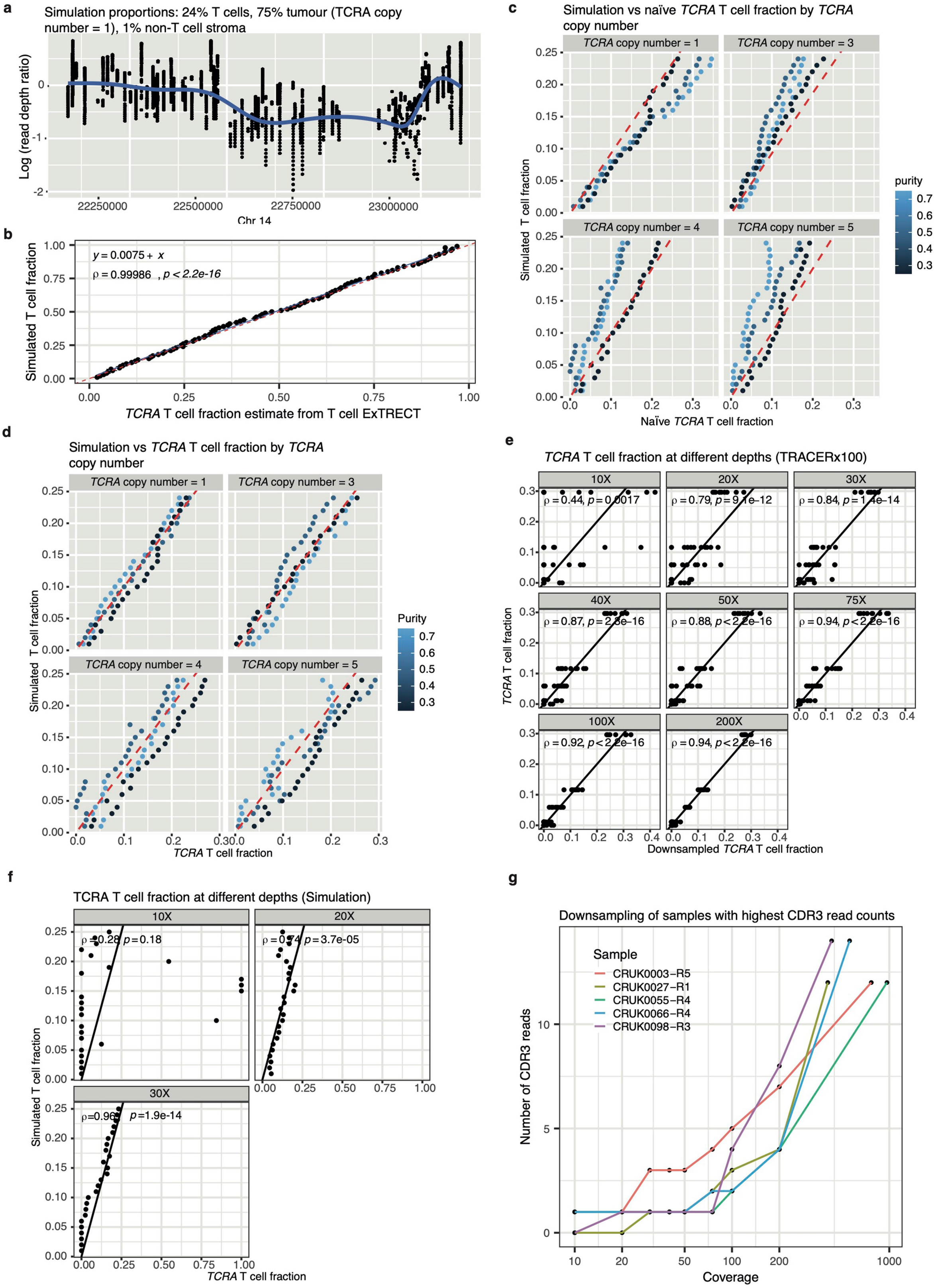


Predictors	Estimates	CI	P
(Intercept)	0.07	0.04 – 0.10	**<0.001**
Hist [Breast]	-0.01	-0.09 – 0.06	0.760
Hist [Colorectal]	-0.05	-0.16 – 0.07	0.397
Hist [Head and neck]	0.05	0.01 – 0.10	**0.010**
Hist [Lung]	0.00	-0.13 – 0.13	0.977
Hist [Melanoma]	0.12	0.09 – 0.15	**<0.001**
Hist [Other]	0.00	-0.04 – 0.04	0.982
Hist [Renal]	0.04	0.01 – 0.07	**0.009**
FFPE or FF [FreshFrozen]	0.01	-0.02 – 0.04	0.589
Observations	817		

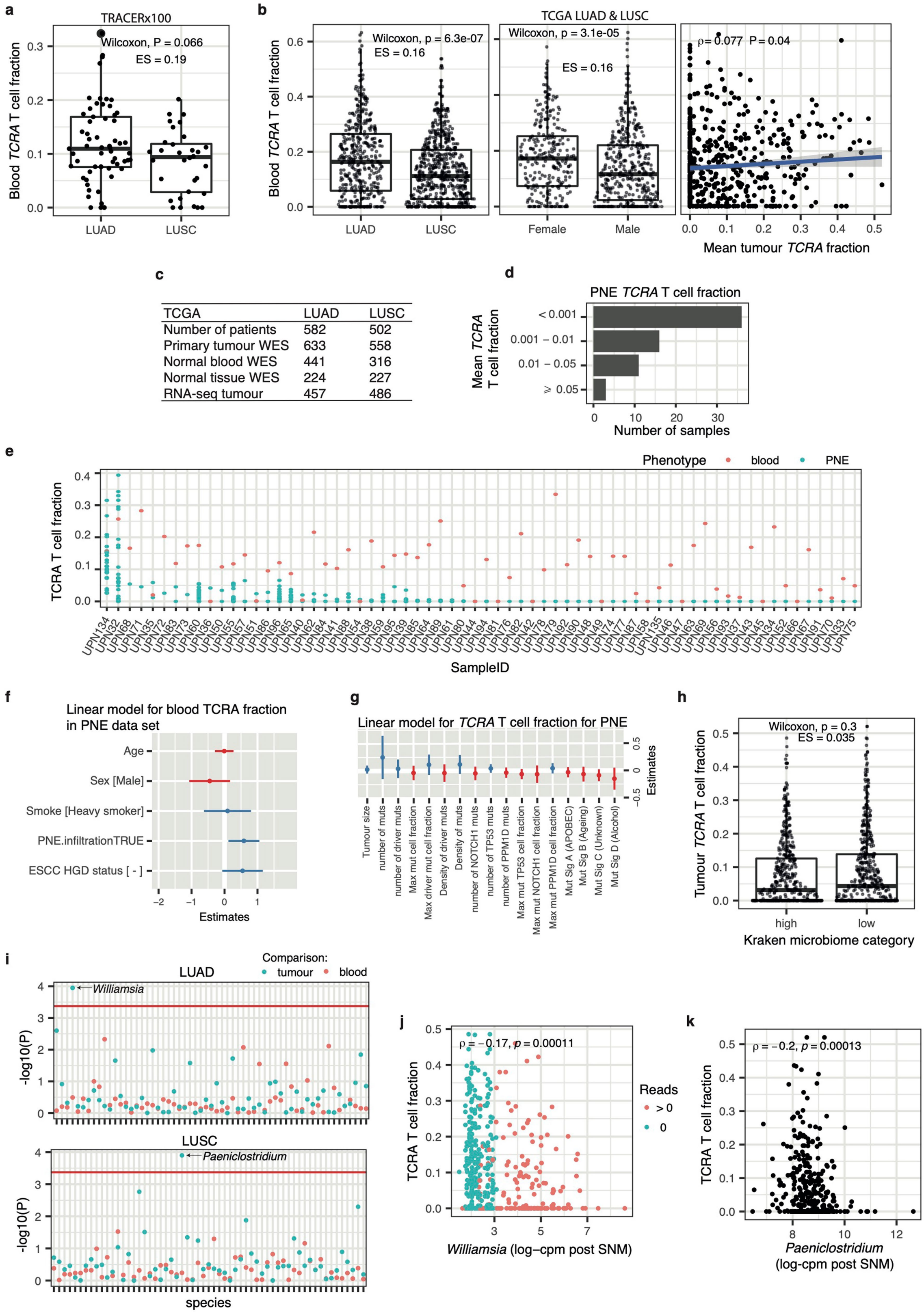
TRACERx100	LUAD	LUSC	Other	Total
Number of patients	61	32	7	100
WES regions	181	116	30	327
RNA-seq regions	105	51	20	176
TIL regions	127	96	23	246
TIL and RNA-seq regions	81	49	18	148

Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Overview and validation of T cell ExTRECT. a, Outline of quantification of the *TCRA* T cell fraction utilising V(D)J recombination and TRECs. *top:* Schematic demonstrating how RDR signals are used to detect SCNA gain or loss events in a standard tumour and matched control sample analysis. In this analysis cells consist of three distinct cell types: tumour cells, T cells and all other stromal cells. *bottom:* Schematic of how this same process works when focussing on the *TCRA* gene in relation to V(D)J recombination and TRECs, the lower right panel indicates an increased number of breakpoints detected in the TRACERx100 dataset within the *TCRA* gene relative to surrounding areas of 14q, suggesting that the TREC signal is captured. **b**, **c**, Plots showing examples of RDR in two TRACERx100 samples demonstrating either increased levels of T cell content in blood compared to matched tumour (**b**) or increased levels of T cell content in tumour compared to matched blood (**c**). VDV segments refer to variable segments in both the TCRα and TCRδ locus. **d**, *TCRA* T cell fraction (non-GC corrected) value for FFPE and fresh frozen samples for bladder and melanoma tumours within the CPI1000+ cohort (bladder: n = 228, melanoma: n = 297, two sided Wilcoxon rank-sum (Mann-Whitney U) test used, boxplot shows lower quartile, median and upper quartile values). **e**, Summary of linear model for prediction of non-GC corrected *TCRA* T cell fraction from histology and FFPE sample status within the CPI cohort. **f**, Pie charts of calculated *TCRA* T cell fraction from WES of either T cell-derived cell lines or non-T cell derived cell lines, all HCT116 cell lines had calculated fractions < 1 e-15. **g**, Overview of samples in the TRACERx100 cohort. **e**, Association of the CDR3 V(D)J read score based on the iDNA method to *TCRA* T cell fraction in TRACERx100, error bands represent the 95% confidence interval of the fitted linear model.



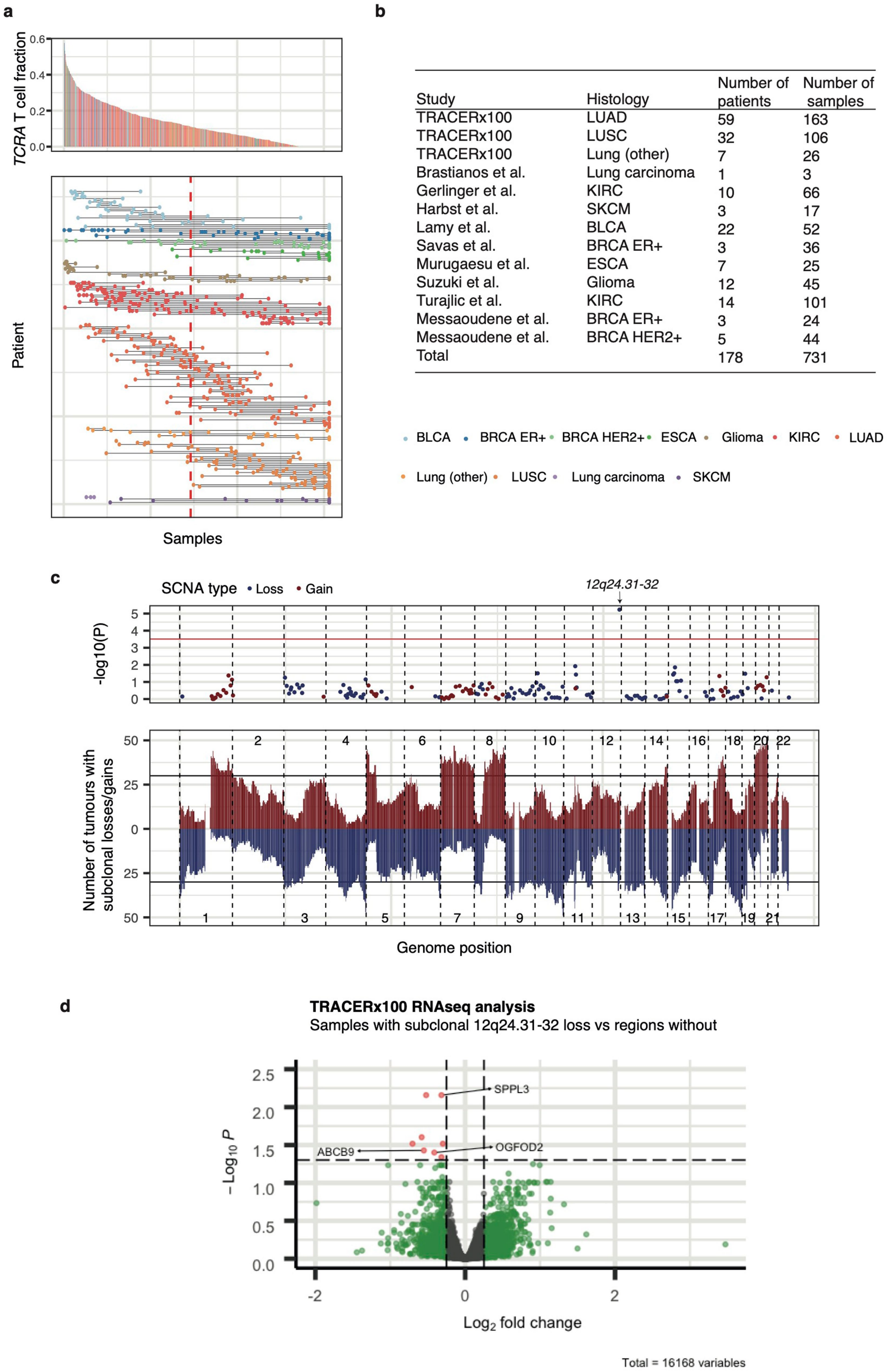
Extended Data Fig. 2 | Accuracy of *TCRA* T cell fraction by copy number and depth. a, Simulated log RDR from a sample consisting of 24% T cells, 75% tumour, and 1% non-T cell stroma (*TCRA* copy number = 1). **b**, Calculated *TCRA* T cell fraction versus actual T cell fraction value for simulated data **c**, Difference between calculated naive T cell fraction and actual fraction for range of tumour purities and local tumour copy number states at the *TCRA* locus. **d**, Difference between *TCRA* T cell fraction and actual fraction for a range of local tumour copy number to the *TCRA* locus and tumour purities. **e**,. Downsampling of 5 TRACERx100 samples to different depths. **f**, Downsampling of simulated data to different depth levels. **g**, Downsampling of the 5 TRACERx100 samples that with the highest CDR3 read counts to different depths and the resulting CDR3 read counts.



TCGA	LUAD	LUSC
Number of patients	582	502
Primary tumour WES	633	558
Normal blood WES	441	316
Normal tissue WES	224	227
RNA-seq tumour	457	486

Extended Data Fig. 3 | See next page for caption.

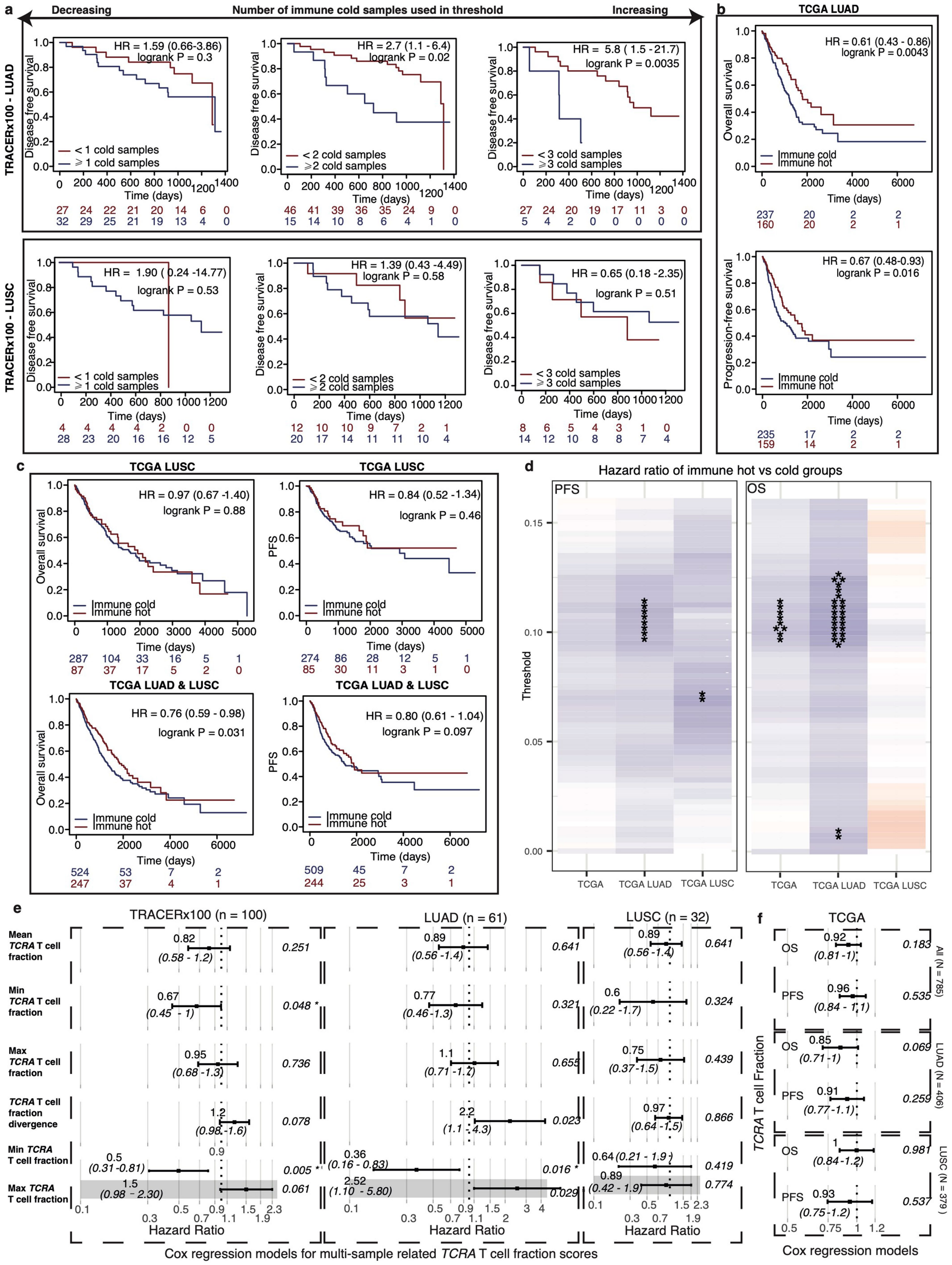
Extended Data Fig. 3 | Extended analysis on determinants of *TCRA* T cell fraction. **a**, Association of blood *TCRA* T cell fraction to histology in TRACERx100 (n = 93 LUAD and LUSC patients). **b**, Predictors of blood *TCRA* T cell fraction in TCGA LUAD and LUSC cohort (left panel: n = 1017, middle panel: n = 976, right panel: n = 714). **c**, Overview of samples in the TCGA LUAD and LUSC cohort. **d**, Summary of mean *TCRA* T cell fraction in PNE cohort. **e**, Overview plot of PNE cohort containing multi-sample microdissected tissue paired with normal blood samples. **f**, Summary of linear model for predicting blood *TCRA* T cell fraction, PNE infiltration defined as *TCRA* T cell fraction > 0.001, ESCC = Oesophageal squamous cell carcinoma, HGD = high grade dysplasia. **g**, Linear model for *TCRA* T cell fraction in PNE samples from genomic factors. **h**, Association of microbial reads from Kraken with *TCRA* T cell fraction in tumour samples (n = 880). **i**, -Log10 p-values for 59 microbial species tested for association with *TCRA* T cell fraction in blood and tumour sample in LUAD and LUSC. Red line represents the significance threshold at P = 0.000423. **j**, The significant hit *Willamsia* in LUAD tumours, red dots represent samples where reads were detected while blue represent samples with no reads detected (n = 501). **k**, The significant hit *Paeniclostridium* in LUSC tumours (n = 379). All Wilcoxon tests refer to Wilcoxon rank-sum (Mann-Whitney U) tests and are two sided. Boxplots represent lower quartile, median and upper quartile.



Study	Histology	Number of patients	Number of samples
TRACERx100	LUAD	59	163
TRACERx100	LUSC	32	106
TRACERx100	Lung (other)	7	26
Brastianos et al.	Lung carcinoma	1	3
Gerlinger et al.	KIRC	10	66
Harbst et al.	SKCM	3	17
Lamy et al.	BLCA	22	52
Savas et al.	BRCA ER+	3	36
Murugaesu et al.	ESCA	7	25
Suzuki et al.	Glioma	12	45
Turajlic et al.	KIRC	14	101
Messaoudene et al.	BRCA ER+	3	24
Messaoudene et al.	BRCA HER2+	5	44
Total		178	731

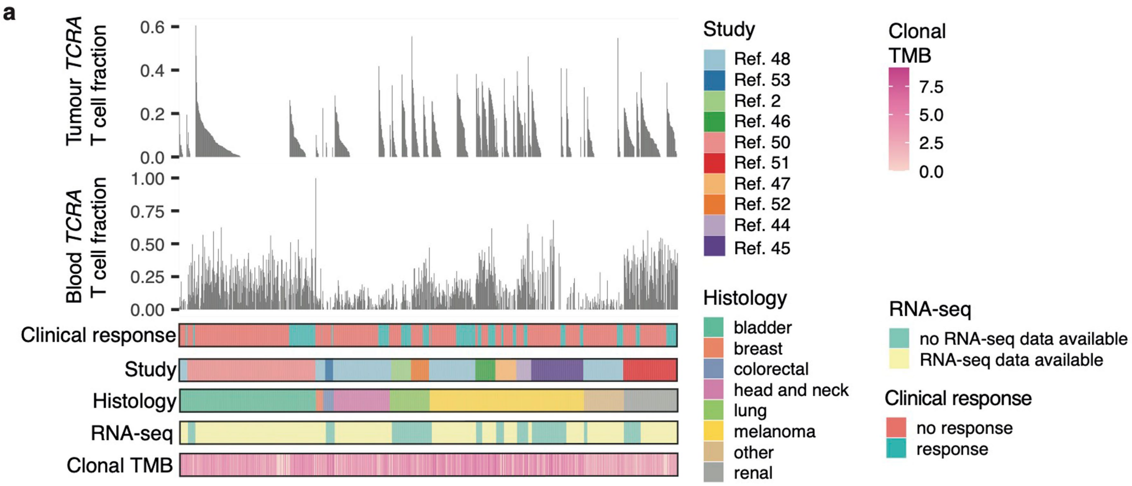
Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Subclonal SCNAs and T cell infiltration. a, Overview of immune heterogeneity across multi-sample pan-cancer cohort with tumour samples ranked by *TCRA* T cell fraction, *upper panel:* histogram of entire cohort, *lower panel:* tumour sample grouped by patients with solid horizontal lines joining regions from the same patient, each line includes 2 or more tumour region and dashed red line is at the mean *TCRA* T cell fraction in the cohort (0.11). **b**, Overview of patients in the multi-sample pan-cancer cohort. **c**, Lower panel: number of tumours in pan-cancer multi-sample cohort with subclonal gains (dark red) or losses (dark blue) across the genome, horizontal lines signify the samples which have more than 30 tumours (Methods) with subclonal gains or losses. *Upper panel*: - log10(p-value) of the 160 cytoband regions tested for association between *TCRA* T cell fraction and subclonal gains (dark red points) or losses (dark blue points). Red horizontal line marks significance threshold, only one region is significant, a loss event on chromosome 12q24.31-32. **d**, Volcano plot for the RNA-seq analysis in the TRACERx100 cohort between samples with 12q24.31-32 loss and samples without, genes within the locus are labeled, dotted lines at fold change of 0.25 and adjusted P = 0.05.



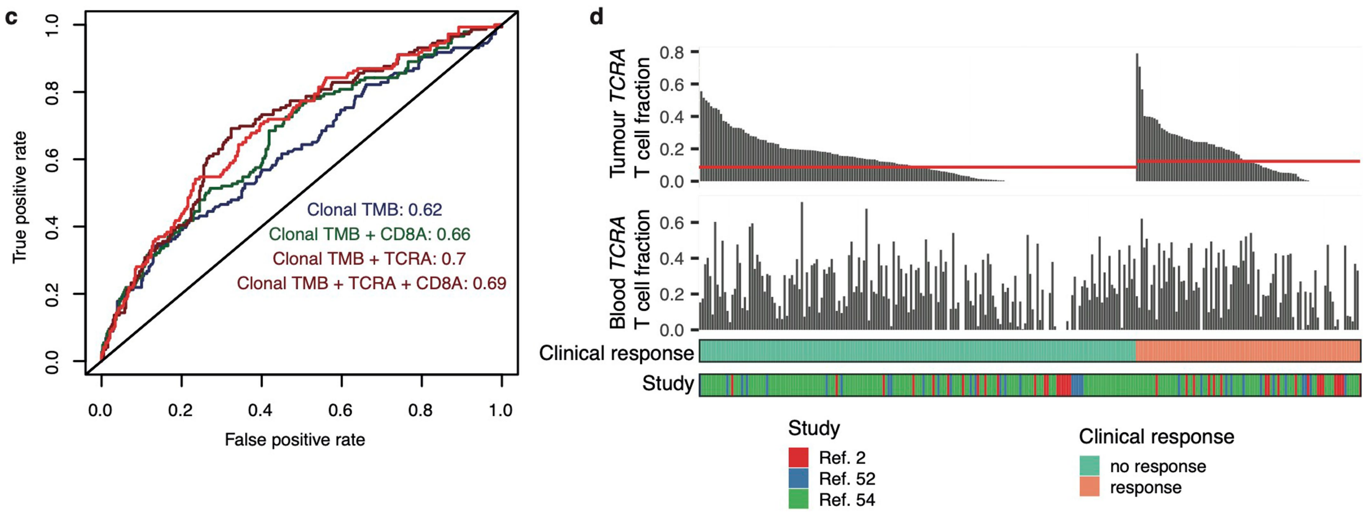
Extended Data Fig 5 | See next page for caption.

Extended Data Fig 5 | Association of *TCRA* T cell fraction with prognosis. **a**, Kaplan-Meier curves for the multi-sample TRACERx100 cohort for LUAD (top) and LUSC (bottom) divided by the number of cold samples in the tumour. Immune-hot and immune-cold samples were defined by using the median of all the tumour samples (0.0736) as a threshold. In each Kaplan-Meier curve the included patients were restricted to those with total samples greater than the number of immune-cold samples used in defining the threshold. **b**, Kaplan-Meier curves for overall and progression free survival in the TCGA LUAD cohort, dividing the cohort into immune-hot and immune-cold groups using the mean of the TCGA LUAD cohort (0.109) as a threshold. **c**, Kaplan-Meier curves for the TCGA LUSC, and TCGA LUAD & LUSC cohorts for overall and progression free survival using the mean of the TCGA LUAD cohort (0.109) as a threshold for distinguishing hot and cold tumours. **d**, Log2(Hazard ratios) from Kaplan-Meier plots for the TCGA separating the tumour samples into immune-hot and immune-cold based on different thresholds from 0 to 0.16 in steps of 0.0025 for overall and progression free survival. **e**, Hazard ratios of separate Cox regression models relating disease free survival to different multi-sample measures related to the *TCRA* T cell fraction in the entire TRACERx100 cohort as well as the LUAD and LUSC patients separately. *TCRA* divergence score is defined as the maximum divided by the upper 95% confidence interval of the minimum. **f**, Hazard ratios of separate Cox regression models for *TCRA* T cell fraction for the TCGA LUAD and LUSC cohort for both overall survival (OS) and progression free survival (PFS).



b

CPI1000+	Melanoma	Bladder	Head and neck	Renal	NSCLC	Other	Colorectal	Breast	Total
WES	295	261	107	104	76	76	20	14	953 (89%)
RNA-seq	184	348	105	72	4	70	5	14	802 (74%)
WES + RNA-seq	178	247	105	72	4	70	5	14	695 (65%)
Patients	309	363	107	105	76	76	20	14	1070



e

Study	Number of patients
Ref. 52.	34
Ref. 2	37
Ref. 54	195
Total	266

Extended Data Fig 6 | Overview of CPI1000+ cohort. a, Cohort overview of the CPI1000+ dataset. **b**, Overview of samples in the CPI1000+ cohort excluding Snyder et al.[49] and those with prior CPI treatment. **c**, ROC plot of GLM models for predicting CPI response (blue: clonal TMB, red: clonal TMB + *TCRA* T cell fraction, green: clonal TMB + *CD8A* expression). **d**, Cohort overview of the CPI lung dataset, red lines in upper panel reflect the median *TCRA* T cell fraction in patients with (0.10) or without (0.0070) a response to CPI, note that Tumour *TCRA* T cell fraction particularly in non-responders is often zero. **e**, Overview of patients in the CPI Lung cohort.

Extended Data Table 1 | Original source publications

Paper	PMID	Cohort
Yokoyama et al. 2020	30602793	PNE cohort
Harbst et al. 2016	27216186	Multi-sample pan-cancer cohort
Lamy et al. 2016	27488526	Multi-sample pan-cancer cohort
Brastianos et al. 2015	26410082	Multi-sample pan-cancer cohort
Gerlinger et al. 2012	22397650	Multi-sample pan-cancer cohort
Gerlinger et al. 2014	24487277	Multi-sample pan-cancer cohort
Savas et al. 2016	28027312	Multi-sample pan-cancer cohort
Suzuki et al. 2015	25848751	Multi-sample pan-cancer cohort
Turajlic et al. 2018	29656894	Multi-sample pan-cancer cohort
Messaoudene et al. 2019	30924846	Multi-sample pan-cancer cohort
Snyder et al. 2014	25409260	CPI1000+ cohort
Van Allen et al. 2015	26359337	CPI1000+ cohort
Hugo et al. 2016	26997480	CPI1000+ cohort
Riaz et al. 2017	29033130	CPI1000+ cohort
Cristescu et al. 2018	30309915	CPI1000+ cohort
Snyder et al. 2017	28552987	CPI1000+ cohort
Mariathasan et al. 2018	29443960	CPI1000+ cohort
McDermott et al. 2018	29867230	CPI1000+ cohort
Rizvi et al. 2015	25765070	CPI1000+ cohort (lung)
Le et al. 2015	26028255	CPI1000+ cohort
Litchfield et al. 2021	33508232	CPI1000+ cohort (lung)
Shim et al. 2020	32320754	CPI1000+ cohort (lung)

Original source publications (excluding TRACERx studies) containing the sequencing data used in either the multi-sample pan-cancer cohort, PNE cohort or the CPI1000+ cohort. Studies including lung cancer patients used in the lung CPI cohort are noted.

nature research

Corresponding author(s): Nicholas McGranahan

Last updated by author(s): Jul 30, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a | Confirmed

□ ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement

□ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

□ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.

□ ☒ A description of all covariates tested

□ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

□ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

□ ☒ For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.

☒ □ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

☒ □ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

□ ☒ Estimates of effect sizes (e.g. Cohen's d, Pearson's r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used to collect data
Data analysis	R (version 3.6.1) R (version 4.0.0 was used for the RNA-seq differential gene expression analysis) samtools (version 1.3.1) ART Illumina (version 2.5.8) MASCoTE (https://github.com/raphael-group/mascote) Picard tools (version 1.107) bwa mem (v0.7.15) R packages used in version 3.6.1: ggpubr (version 0.4.0) rstatix (version 0.6.0) survival (version 3.2.3) ggplot2 (version 3.3.3) dplyr (version 1.0.4) tidyr (version 1.1.1) gridExtra (version 2.3) gtable (version 0.3.0) meta (version 4.13.0) ROCR (version 1.0.11) EnhancedVolcano (version 1.8.0) gratia (version 0.5.1)

R packages used in version 4.0.0:
limma (version 3.46.0)
edgeR (version 3.32.1)

The code used to produce TCRA T cell fraction scores is available for academic non-commercial research purposes upon reasonable request.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

TRACERx sequencing datasets used in this paper are described in previous studies (see PMID:28445112 and PMID:30894752). Details of all datasets obtained from third parties used in this study (see Extended Table 1) are as follows:

- Yokoyama et al. 2020 (PMID:30602793)
- Harbst et al. 2016 (PMID:27216186)
- Lamy et al. 2016 (PMID:27488526)
- Brastianos et al. 2015 (PMID:26410082)
- Gerlinger et al. 2012 (PMID:22397650)
- Gerlinger et al. 2014 (PMID:24487277)
- Savas et al. 2016 (PMID:28027312)
- Suzuki et al. 2015 (PMID:25848751)
- Turajlic et al. 2018 (PMID:29656894)
- Messaoudene et al. 2019 (PMID:30924846)
- Snyder et al. 2014 (PMID:25409260)
- Van Allen et al. 2015 (PMID:26359337)
- Hugo et al. 2016 (PMID:26997480)
- Riaz et al. 2017 (PMID:29033130)
- Cristescu et al. 2018 (PMID:30309915)
- Snyder et al. 2017 (PMID:28552987)
- Mariathasan et al. 2018 (PMID:29443960)
- McDermott et al. 2018 (PMID:29867230)
- Rizvi et al. 2015 (PMID:25765070)
- Le et al. 2015 (PMID:26028255)
- Litchfield et al. 2021 (PMID:33508232)
- Shim et al. 2020 (PMID:32320754)

The RNA-seq data and Whole exome sequencing (WES) data (in each case from the TRACERx study) generated, used or analysed during this study are not publicly available and restrictions apply to the availability of these data. Such RNAseq and WES data are available through the Cancer Research UK & University College London Cancer Trials Centre (ctc.tracerx@ucl.ac.uk) for academic non-commercial research purposes upon reasonable request, and subject to review of a project proposal that will be evaluated by a TRACERx data access committee, entering into an appropriate data access agreement and subject to any applicable ethical approvals.

Clinical trial information (if applicable) is also available within the associated publications described both above and in Extended data table 1.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences □ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	All analysis was done on pre-existing data sets and as such no statistical methods were used to predetermine sample size.
Data exclusions	Samples with insufficient purity or unreliable copy number profiles were excluded from the analysis. The multi-sample pan cancer cohort was designed to focus on multi-sample primary and not metastatic tumours and thus was restricted to tumours with multiple regions sampled from the primary tumour. Patients that had already received CPIs or were on treatment at the time of the tumour sampling were excluded from the CPI1000+ cohort due to any possible confounding effects this would have on our analysis.

Replication

This study was on pre-existing data sets and hence findings were not replicated

Randomization

No randomization or permutation analysis was performed in this study, samples were split based on either categorical data or threshold values e.g. for the TCRA T cell fraction.

Blinding

Blinding was not applicable in this study, all data was from pre-existing data and there was no control and treatment arms involved

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	□ Antibodies
☒	□ Eukaryotic cell lines
☒	□ Palaeontology and archaeology
☒	□ Animals and other organisms
□	☒ Human research participants
□	☒ Clinical data
☒	□ Dual use research of concern

Methods

n/a	Involved in the study
☒	□ ChIP-seq
☒	□ Flow cytometry
☒	□ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics

There were 68 male and 32 female non-small cell lung cancer patients in the TRACERx study, with a median age of 68. The cohort is predominantly early-stage: Ia(26), Ib(36), IIa(13), IIb(11), IIIa(13), IIIb(1). Seventy-two had no adjuvant treatment and 28 had adjuvant therapy.
Patients were recruited into TRACERx according to the following eligibility criteria (taken from the study protocol).
Inclusion criteria:
-Written Informed consent
-Patients ≥18 years of age, with early stage I-IIIA disease who are eligible for primary surgery
-Histopathologically confirmed NSCLC, or a strong suspicion of cancer on lung imaging necessitating surgery (e.g. diagnosis determined from frozen section in theatre)
-Primary surgery in keeping with NICE guidelines planned (see section 9.3)
-Agreement to be followed up in a specialist centre
-Performance status 0 or 1
-Suspected tumour at least 15mm in diameter on pre-operative imaging
Exclusion criteria:
-Any other current malignancy or malignancy diagnosed or relapsed within the past 5 years (other than non-melanomatous skin
cancer, stage 0 melanoma in situ, and in situ cervical cancer)
-Psychological condition that would preclude informed consent
-Treatment with neo-adjuvant therapy for current lung malignancy deemed necessary
-Adjuvant therapy other than platinum-based chemotherapy and/or radiotherapy
-Known Human Immunodeficiency Virus (HIV), Hepatitis B Virus (HBV), Hepatitis C Virus (HCV) or syphilis infection.
-Sufficient tissue, i.e. a minimum of two tumour regions, is unlikely to be obtained for the study based on pre-operative imaging
Patient ineligibility following registration:
-There is insufficient tissue
-The patient is unable to comply with protocol requirements
-There is a change in histology from NSCLC following surgery, or NSCLC is not confirmed during or after surgery.
-Change in staging to IIIB/IV following surgery
-The operative criteria are not met (e.g. incomplete resection with macroscopic residual tumours (R2)); see section 9.3 for a list
of accepted surgical procedures. Patients with microscopic residual tumours (R1) are eligible and should remain in the study
-Adjuvant therapy other than platinum-based chemotherapy and/or radiotherapy is administered.

Recruitment

Patients seen with a new diagnosis of lung cancer in lung cancer units across the United Kingdom, according to the eligibility criteria above, were recruited. No selection bias has been identified to date.
All patient tumor regions with RIN scores > 5 were used for RNA-sequencing and analyzed in this study.
All patients were assigned a study ID that was known to the patient. These were subsequently converted to linked study Ids such
that the patients could not identify themselves in study publications. All human samples, tissue and blood, were linked to the study ID and barcoded such that they were anonymised and tracked on a centralised database overseen by the study sponsor only.
Informed consent for entry into the TRACERx study was mandatory and obtained from every patient.

Ethics oversight	The TRACERx study (Clinicaltrials.gov no: NCT01888601) is sponsored by University College London (UCL/12/0279) and has been approved by an independent Research Ethics Committee (13/LO/1546)

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	NCT01888601
Study protocol	The study protocol is available at NEJM.org linked to Jamal-Hanjani et al NEJM 2017 (PMID: 28445112)
Data collection	Patients seen with a new diagnosis of lung cancer in lung cancer units across the United Kingdom, according to the eligibility criteria outlined in the study protocol, were recruited. No selection bias has been identified to date. All patient tumor regions with RIN scores > 5 were used for RNA-sequencing and analyzed in this study. All patients were assigned a study ID that was known to the patient. These were subsequently converted to linked study Ids such that the patients could not identify themselves in study publications. All human samples, tissue and blood, were linked to the study ID and barcoded such that they were anonymised and tracked on a centralised database overseen by the study sponsor only. Informed consent for entry into the TRACERx study was mandatory and obtained from every patient
Outcomes	The outcome measures of the TRACERx trial are intratumour heterogeneity, disease-free survival, and overall survival.