nature genetics

Article

https://doi.org/10.1038/s41588-025-02086-5

ImmuneLENS characterizes systemic immune dysregulation in aging and cancer

Received: 2 May 2023

Accepted: 10 January 2025

Published online: 18 February 2025

Check for updates

Robert Bentham[1,2], Thomas P. Jones[1,2], James R. M. Black[1,2,3], Carlos Martinez-Ruiz[1,2], Michelle Dietzen[1,2,3], Maria Litovchenko[1,2], Kerstin Thol[1,2], Thomas B. K. Watkins[4], Chris Bailey[3], Oriol Pich[3], Zhihui Zhang[5], Peter Van Loo[5,6], TRACERx Consortium*, Genomics England Consortium*, Charles Swanton[2,3,7] & Nicholas McGranahan[1,2] ✉

Recognition and elimination of pathogens and cancer cells depend on the adaptive immune system. Thus, accurate quantification of immune subsets is vital for precision medicine. We present immune lymphocyte estimation from nucleotide sequencing (ImmuneLENS), which estimates T cell and B cell fractions, class switching and clonotype diversity from whole-genome sequencing data at depths as low as 5× coverage. By applying ImmuneLENS to the 100,000 Genomes Project, we identify genes enriched with somatic mutations in T cell-rich tumors, significant sex-based differences in circulating T cell fraction and demonstrated that the circulating T cell fraction in patients with cancer is significantly lower than in healthy individuals. Low circulating B cell fraction was linked to increased cancer incidence. Finally, circulating T cell abundance was more prognostic of 5-year cancer survival than infiltrating T cells.

Measuring the quantity, quality and location of immune cells is vital to understanding their role and function. In cancer research, studies have focused on tumor-infiltrating lymphocytes and their roles in cancer evolution and immune evasion[1–5].

High tumor T cell infiltration is prognostic in cancer[6,7] and influences immunotherapy response[8]. However, B cell infiltration has been linked to both cancer promotion and inhibition[9]. Classifying tumor-infiltrating B cells into different lineages might elucidate their role in cancer, but such analysis without direct assays of B cell markers with flow cytometry, targeted B cell receptor repertoire sequencing or single-cell RNA sequencing (RNA-seq) is challenging.

Although circulating immune cell counts from routine blood tests have been associated with response to therapy and prognosis in cancer[10,11], a systematic exploration of their relative importance compared to tumor-infiltrating immune cells and their clinical correlates is lacking.

Here we present immune lymphocyte estimation from nucleotide sequencing (ImmuneLENS)—a method to quantify immune content from whole-genome sequencing (WGS) data and available at https://github.com/McGranahanLab/ImmuneLENS. In addition to T cell content, our method predicts B cell fraction and class switching from the *IGH* locus and provides estimates for both T cell receptor (TCR) and B cell receptor (BCR) diversity.

Results

Inferring T cell and B cell fractions from WGS data

WGS enables in-depth characterization of somatic mutations, structural variants and copy number alterations[12]. However, whether WGS can simultaneously be used for the estimation of lymphocyte infiltration has not been extensively evaluated.

To this end, we created ImmuneLENS. This tool significantly enhances and extends our previous method, T cell ExTRECT[6] (see

[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]Cancer Evolution and Genome Instability Laboratory, The Francis Crick Institute, London, UK. [4]Department of Pathology, Stanford University School of Medicine, Stanford, CA, USA. [5]Department of Genetics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [6]Department of Genomic Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [7]Department of Medical Oncology, University College London Hospitals, London, UK. *Lists of authors and their affiliations appear at the end of the paper. ✉e-mail: nicholas.mcgranahan.10@ucl.ac.uk

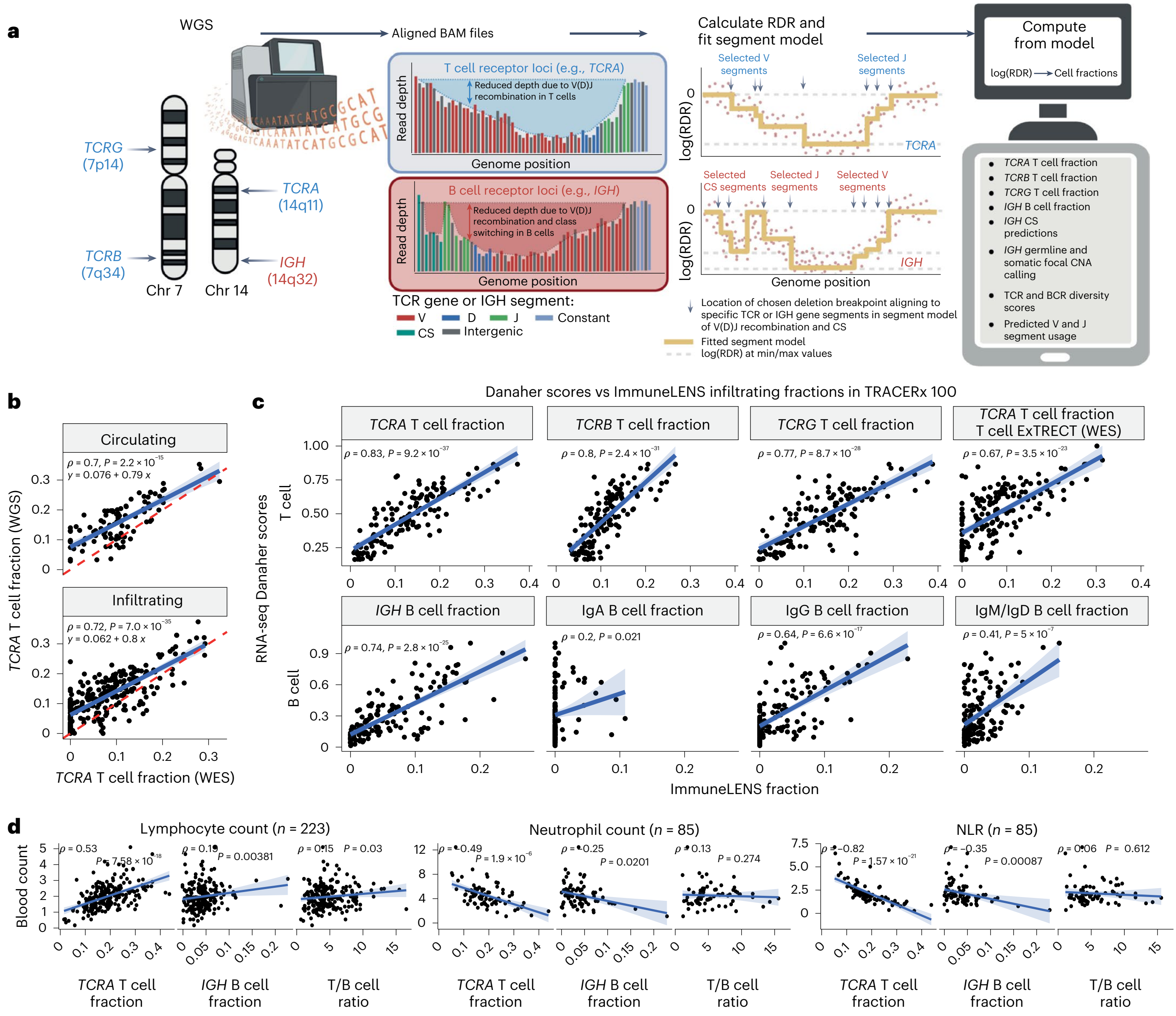


Fig. 1 | Overview of ImmuneLENS and validation. a, Overview of the ImmuneLENS method. The figure is created with BioRender.com. **b**, Scatter plot of *TCRA* T cell fractions calculated from TRACERx WGS versus WES data. The red dotted line represents $y = x$; the blue line shows the line of best fit with a light blue-shaded 95% confidence interval (CI). **c**, Scatter plots comparing ImmuneLENS fractions from TRACERx100 WGS and *TCRA* T cell fractions from T cell ExTRECT (TRACERx100 WES data) against T cell- and B cell-related Danaher scores from matched RNA-seq samples. The blue line represents the line of best fit with a light blue-shaded 95% CI. **d**, Correlation of *TCRA* T cell fraction, *IGH* B cell fraction and T/B cell ratio with date-matched blood count data (lymphocyte count, neutrophil count and NLR values) within the 100KGP cohort. The blue line represents the line of best fit with a light blue-shaded 95% CI. *P* values for Spearman's *ρ* were derived from a two-tailed *t* distribution using the correlation coefficient and sample size. CNA, copy number alteration; CS, class switching; RDR, read depth ratio.

Methods and Fig. 1a for an overview), which was designed for whole-exome sequencing (WES). We made three improvements to the method. First, we introduced a segment-based model, enabling precise breakpoint fitting based on the locations of individual V and J segments. Second, harnessing the predicted V and J segment usage, the model estimates both lymphocyte cell fractions and clonotype diversity. Finally, we introduced an estimate of B cell fraction and immunoglobulin (Ig) class switching from the *IGH* locus, enabling distinction between non-class-switched B cells (IgM/IgD) and class-switched B cells that produce IgA, IgG or IgE antibodies (Extended Data Fig. 1a and Supplementary Fig. 1). The B cell fraction model also incorporates an *IGH* locus germline copy number variant caller (Supplementary Fig. 2). Example output of ImmuneLENS is illustrated in Extended Data Fig. 1b,c.

WGS enables accurate measurement of T cell and B cell fraction

We first evaluated the accuracy of T cell fractions predicted by ImmuneLENS on TRACERx100 (ref. 13) lung cancer samples that had both matched WES and WGS ($n = 322$) or orthogonal RNA-seq data ($n = 126$).

WES and WGS *TCRA* T cell fractions showed positive correlations in both blood ($\rho = 0.70$, $P = 2.2 \times 10^{-15}$) and tumor samples ($\rho = 0.72$, $P = 7.0 \times 10^{-35}$; Fig. 1b). Notably, fewer samples exhibited no T cell infiltrate in WGS than WES (54 WES samples had $<10^{-4}$ T cell fraction compared to only two of the WGS samples), likely reflecting ImmuneLENS' increased sensitivity (Methods; Supplementary Fig. 3).

αβ T cells have recombined TCRα and TCRβ (encoded by *TCRB*) chains, whereas γδ T cells have TCRγ (encoded by *TCRG*) and TCRδ chains (encoded by *TCRD*). *TCRB* and *TCRG* can provide T cell fraction estimates independent of *TCRA*. The T cell fraction estimates of these distinct T cell classes were all positively correlated with each other, indicating that each independently measures T cell content ($\rho > 0.8$, $P < 10^{-70}$; Extended Data Fig. 1d). There are two important caveats to this result. First, *TCRB* T cell fraction was systematically smaller than *TCRA* (line of best fit: $y = 0.56\times + 0.042$), likely due to allelic exclusion[14]. Second, *TCRG*, which is typically expressed only in γδ T cells (1–5% of $CD3^+$ T cells[15]), strongly correlated with *TCRA*, suggesting αβ T cells commonly retain rearranged *TCRG* loci. Previous reports have shown that αβ T cells frequently rearrange their *TCRG* locus before committing to the αβ lineage[16]. Thus, *TCRG* appears to measure total T cell fraction rather than solely γδ T cells.

We further validated ImmuneLENS using TRACERx100 RNA-seq data ($n = 126$). *TCRA*, *TCRB* and *TCRG* T cell fractions strongly correlated with the Danaher T cell signature[17], previously shown to reflect T cell content[18] (*TCRA*: $\rho = 0.83$, $P = 9.2 \times 10^{-37}$; *TCRB*: $\rho = 0.8$, $P = 2.4 \times 10^{-31}$; *TCRG*: $\rho = 0.77$, $P = 8.7 \times 10^{-28}$; Fig. 1c). Consistent correlations were also observed with RNA-seq signatures from TIMER[19], CIBERSORT[20], xCell[20] and scores from ref. 21 (Extended Data Fig. 1e). Likewise, *IGH* B cell fraction strongly correlated with Danaher RNA-seq-based B cell score (Fig. 1c; $\rho = 0.74$, $P = 2.8 \times 10^{-25}$). However, we observed that correlation strength varied by class—IgG ($\rho = 0.64$, $P = 6.6 \times 10^{-17}$), IgM/IgD ($\rho = 0.41$, $P = 5 \times 10^{-7}$) and IgA ($\rho = 0.2$, $P = 0.021$). Thus, conceivably IgA B cells may be underrepresented in RNA-seq data or overestimated in DNA (Supplementary Figs. 4 and 5). These results aligned with other RNA-seq-based B cell signatures (Extended Data Fig. 1e). Samples with high WGS-inferred B cell fractions (>median) showed significant enrichment of B cell gene expression for all subsets, even IgA (IgG: adjusted $P = 4.4 \times 10^{-4}$; IgM/IgD: adjusted $P = 1.3 \times 10^{-3}$; IgA: adjusted $P = 2.4 \times 10^{-3}$; Extended Data Fig. 1f).

ImmuneLENS provided accurate T cell measurements at depths as low as 5× for *TCRA* ($R = 0.96$, $P = 5.4 \times 10^{-223}$), *TCRB* ($R = 0.61$, $P = 6.59 \times 10^{-43}$) and *TCRG* ($R = 0.72$, $P = 1.93 \times 10^{-67}$) as evidenced using downsampled data (Extended Data Fig. 2a–d). For B cell quantification, >10× was required for accurate germline copy number inference (Extended Data Fig. 2e). Similarly, correction for possible tumor copy number alterations was only accurate at depths >20× (Extended Data Fig. 2f). Consistent results were observed for matched high and low-coverage WGS data (Supplementary Fig. 6).

To further validate ImmuneLENS, we applied the tool to blood-derived WGS samples with date-matched blood count data from 441 participants of the 100,000 Genomes Project (100KGP; Methods). Circulating T cell fraction correlated positively with lymphocyte count ($\rho = 0.53$, $P = 7.6 \times 10^{-18}$) and negatively with both neutrophil count ($\rho = -0.49$, $P = 1.9 \times 10^{-6}$) and neutrophil-to-lymphocyte ratio (NLR) ($\rho = -0.82$, $P = 1.6 \times 10^{-21}$; Fig. 1d). These data suggest that *TCRA* T cell fraction serves as an NLR proxy. A weak negative correlation with albumin concentration was observed (Extended Data Fig. 3a; $\rho = -0.25$, $P = 2.1 \times 10^{-6}$). No significant associations were found with C-reactive protein or ferritin (Extended Data Fig. 3a), although weak negative correlations existed between white blood cell count and both T cell fraction ($\rho = -0.18$, $P = 0.0076$) and T/B cell ratio ($\rho = -0.17$, $P = 0.013$). Similar trends were observed for *IGH* B cell fraction (Fig. 1d). Notably, the T/B cell ratio correlated significantly with lymphocyte count ($\rho = 0.15$, $P = 0.03$) but not with neutrophil count ($\rho = 0.13$, $P = 0.27$) or NLR ($\rho = 0.06$, $P = 0.612$; Fig. 1d). Thus, the T/B cell ratio provides a measure of lymphocyte count, which is independent of neutrophil levels.

Investigating T cell receptor diversity from WGS data

The ability of ImmuneLENS to fit individual V and J segments allows TCR and BCR diversity analysis from WGS data (Extended Data Fig. 4a).

We assessed TCR diversity accuracy from WGS data using three methods. First, we compared ImmuneLENS output with matched TCR-sequencing (TCR-seq) data in TRACERx. ImmuneLENS' T cell receptor alpha variable (TRAV) segment proportions were significantly different between quartiles representing different levels of actual segment usage inferred from TCR-seq (Kruskal–Wallis, $P = 3.3 \times 10^{-31}$; Extended Data Fig. 4b). ImmuneLENS likely underestimates TCR diversity, illustrated by low TRAV segment usage predictions in Extended Data Fig. 4b. Second, we compared Shannon diversity scores from WGS-derived V segment usage (ImmuneLENS) with the calculated TCR repertoire diversity from RNA-seq data (using MiXCR[22]). A significant correlation was observed between DNA- and RNA-derived TCR diversity scores (Extended Data Fig. 4c; $\rho = 0.34$, $P = 7.4 \times 10^{-7}$). Third, using the Jensen–Shannon divergence, we found significantly more similar repertoires in samples from the same patient than from different patients, for both infiltrating–infiltrating and infiltrating–circulating comparisons (Wilcoxon rank-sum $P = 9.74 \times 10^{-92}$ and $P = 0.011$, respectively; Extended Data Fig. 4d).

These results therefore demonstrate that predicted TRAV segments enable the assessment of TCR diversity within samples and enable TCR repertoire comparisons across samples.

The immune landscape in the 100KGP cohort

Having validated the accuracy of ImmuneLENS, we next applied the tool to 90,232 WGS samples from the 100KGP cohort[23] (see Fig. 2a and Supplementary Data for a full clinical overview). This included 14,501 cancer samples across 33 distinct histologies (>100 participants each), with 13,870 having matched blood samples. In total, 631 cancer samples, including 538 with hematological cancers, lacked matched blood samples. Additionally, blood samples from 30,665 healthy individuals sequenced as relatives within the 100KGP rare disease cohort were analyzed. The remaining 30,565 WGS germline samples from the 100KGP rare disease cohort were excluded from our main analysis because they originated from either the rare disease probands or non-blood samples. We analyzed the following two measures: infiltrating immune cell fraction (from tumor WGS samples) and circulating immune cell fraction (from the buffy coat of blood WGS samples).

Significant differences in T cell fractions were observed between cancer types for both circulating and infiltrating T cells (Kruskal–Wallis, $P = 5.3 \times 10^{-209}$ and $P = 3.7 \times 10^{-490}$, respectively). Circulating T cell fractions were highest in patients with childhood cancer (median = 0.19) and lowest in patients with glioblastoma (median = 0.051). The low circulating T cell fraction in patients with glioblastoma may reflect steroid treatment, which increases circulating neutrophil levels[24]. Likewise, both infiltrating and circulating B cell fractions differed significantly across cancer types (Kruskal–Wallis, $P = 3.8 \times 10^{-291}$ and $P = 3.3 \times 10^{-42}$, respectively; Fig. 2). We found similar significant differences between cancer types for infiltrating T cells when adjusted by age and for the T/B cell ratio (Extended Data Fig. 5a). Additionally, consistent results in the Pan-Cancer Analysis of Whole Genomes (PCAWG)[12] and The Cancer Genome Atlas (TCGA) WGS and WES datasets were identified (Supplementary Fig. 7).

Tumor samples exhibited higher B cell content compared to blood samples (effect size = 0.181, $P = 9.6 \times 10^{-199}$; Extended Data Fig. 6a); conversely, T cell content was higher in blood samples than in tumors (effect size = 0.522, $P < 2.22 \times 10^{-308}$; Extended Data Fig. 6a). Circulating and infiltrating T cell fractions showed no clear correlation ($R = 0.03$, adjusted $P = 0.068$), except in colorectal adenocarcinoma ($R = 0.13$, adjusted $P = 6.9 \times 10^{-7}$; Extended Data Fig. 6b).

ImmuneLENS revealed differences in B cell class switching between infiltrating and circulating B cells across cancer types (Fig. 2a). Elevated tumor-infiltrating B cell levels compared to circulating were primarily due to an enrichment in class-switched IgA and IgG B cells (IgA: effect size = 0.21, $P = 6.2 \times 10^{-270}$; IgG: effect size = 0.046, $P = 4.1 \times 10^{-14}$; Extended Data Fig. 6a). IgM/IgD B cells showed a smaller but significant difference between circulating and infiltrating fractions (IgM/IgD: effect size = 0.023, $P = 1.3 \times 10^{-4}$; Extended Data Fig. 6a). This highlights key differences in the function and make-up of circulating and tumor-infiltrating B cells, underscoring the specialized roles of

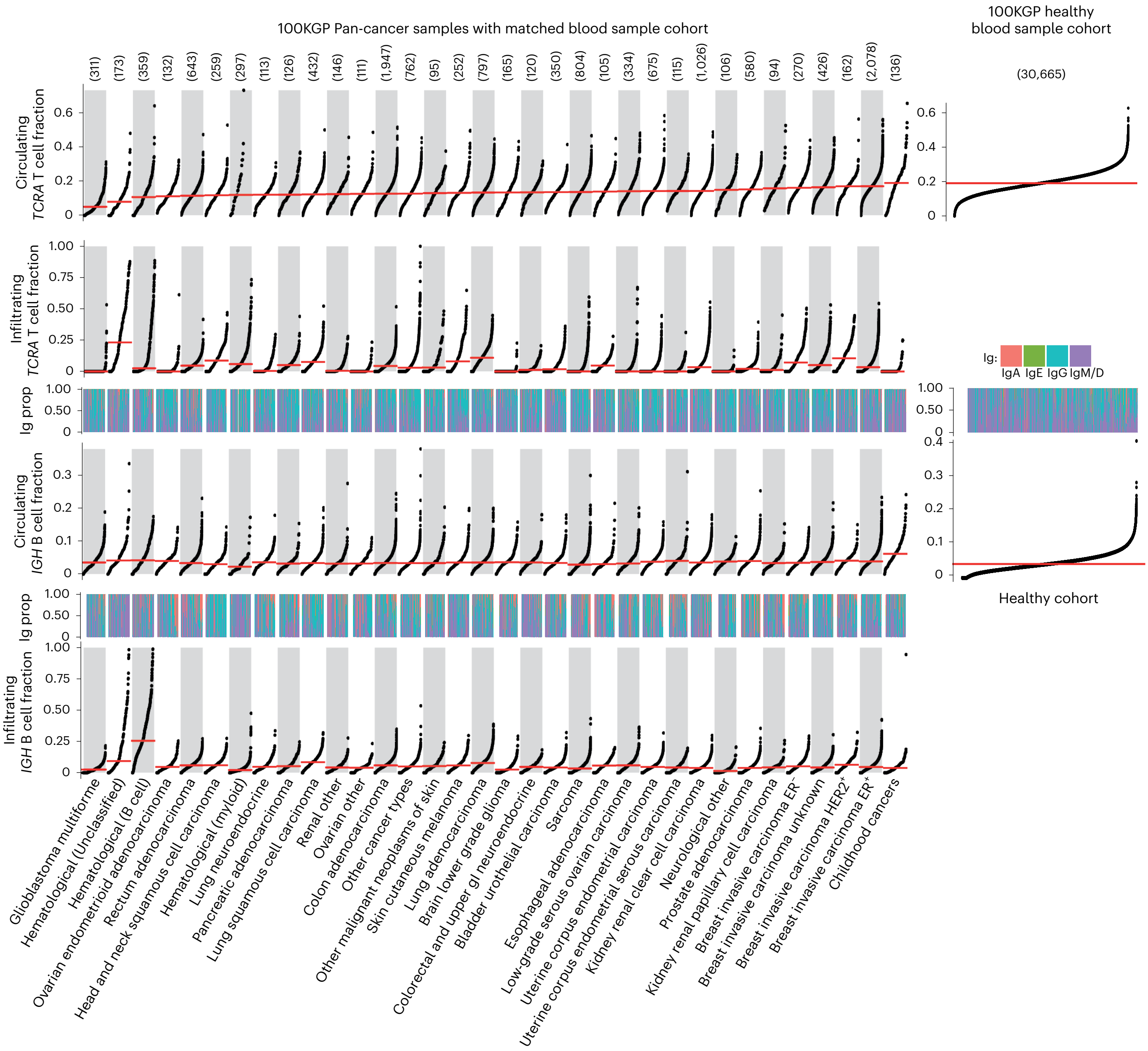


Fig. 2 | ImmuneLENS applied to 100KGP. The number of tumor samples per cancer histology is given above the plot. The panels represent snake plots for circulating and infiltrating *TCRA* T cell fractions and *IGH* B cell fractions, with each point representing a single blood or tumor sample. Above each *IGH* B cell fraction snake plot is a track, shown as a heatmap, displaying the proportion of different Ig B cells for each sample. Histology groups are arranged in ascending order based on the median circulating *TCRA* T cell fraction, and within each group, samples are sorted from lowest to highest value in each snake plot. Right, snake plots for the circulating T cell and B cell fractions within the 100KGP healthy cohort. No significant differences were identified (using ANOVA) in the proportions of B cell Ig status among cancer histology groups in either circulating or infiltrating samples. Horizontal red lines represent the median value per histology group. GI, gastrointestinal.

B cell subtypes. For instance, IgM antibodies in the circulatory system have an important role in activating the complement system, while in mucosal tissue, IgA antibodies are crucial for immune homeostasis. Consistent with these findings, we observed a significantly higher T/B cell ratio in blood compared to tumor samples (effect size = 0.61, $P < 2.22 \times 10^{-308}$; Extended Data Fig. 6a).

IGH B cell fractions, particularly IgG, correlated strongly between circulating and infiltrating in the majority of histologies (pan-cancer *IGH*: $R = 0.17$, adjusted $P = 2.7 \times 10^{-86}$; see full results in Supplementary Data and Extended Data Fig. 6b). Additionally, we found that TRAV segment usage differed significantly between circulating and infiltrating T cells and across cancer types, with *TRAV1-2* enriched in mucosal cancers. This may reflect mucosal-associated invariant T cells, a subset of T cells that recognize bacterial-produced metabolites that exclusively use the *TRAV1-2* segment[25] (Supplementary Data and Supplementary Fig. 8).

Determinants of circulating leukocyte fraction

Given the wide range of circulating immune fractions across both healthy participants and patients with cancer, we next sought to investigate the key determinants of leukocyte fraction.

Analysis of 100KGP participants in 5-year age brackets revealed declining T cell and B cell fractions with age in both healthy and cancer cohorts. For B cells, this effect was strongest for the IgM/IgD B cells, with the relative proportion of class-switched B cells increasing with age (Fig. 3a). Patients with cancer consistently exhibited lower circulating T cell and B cell fractions and higher class-switched B cell proportions compared to healthy individuals. Notably, on average a 40–45-year-old female patient with cancer has similar median circulating T cell fraction levels to a healthy >80-year-old female (0.161 versus 0.157). The decreased T/B cell ratio in the blood of patients with cancer suggests that this effect extends beyond a relative increase in neutrophils. Thus, in both healthy individuals and those with cancer, age is a key determinant of circulating immune fractions, and patients with cancer exhibit an inflated 'immunological age'.

Sex differences in circulating immune fractions were evident in both cancer and healthy cohorts. Female patients with cancer showed significantly higher T cell fractions than male patients across most age groups (adjusted $P < 0.001$ for all age groups >40 years; Fig. 3a and Supplementary Data). In the healthy cohort, significant (adjusted $P < 0.001$) sex differences were primarily observed in older age groups (>55 years; Fig. 3a). Likewise, the T/B cell ratio difference between sexes was more pronounced in people with cancer, particularly in the 65–69 (adjusted $P = 2.7 \times 10^{-5}$, effect size = 0.11) and 70–74 age groups (adjusted $P = 1.1 \times 10^{-4}$, effect size = 0.10), compared to the healthy cohort (significant at adjusted $P < 0.001$ only in the 60–64 age group, adjusted $P = 5.2 \times 10^{-4}$, effect size = 0.11). This suggests neutrophil count may contribute to T cell fraction differences in the healthy cohort. For the IgM/IgD B cell fraction in the healthy group, a sex-based switch was observed from age 55, with male individuals higher initially and female individuals after age 55. In the cancer cohort, female patients exhibited higher B cell fractions in age groups >55. These trends were generally consistent across individual cancer types (Supplementary Fig. 9 and Supplementary Data).

To assess whether circulating immune cell levels could predict future cancer incidence, we identified 301 participants within the 100KGP healthy cohort who developed cancer within 3 years following germline blood sequencing. Compared to a set of controls propensity-matched for age and sex, those diagnosed with cancer within 2 but not 3 years showed significantly lower IgM/IgD (2 years: $P = 0.006$, effect size = 0.14; 3 years: $P = 0.08$, effect size = 0.07) and higher Ig class-switched B cell fractions (2 years: $P = 0.02$, effect size = 0.12; 3 years: $P = 0.07$, effect size = 0.07; Fig. 3b). This suggests circulating immune fraction may serve as a potential cancer marker.

Association of genetic ancestry with lymphocyte fraction

Beyond the effects of age and sex, genetic ancestry may influence immune infiltrate[26] and leukocyte counts in blood[27]. While germline variants associated with the immune system may affect cancer outcomes[28], most relevant GWAS studies use samples from non-cancer patients.

The 100KGP participants were grouped into super-populations defined from the 1000 Genomes Project[29]. Significant differences in circulating T cell fractions were observed among genetically inferred ancestry groups in both healthy and cancer cohorts, with genetic African ancestry showing significantly higher immune fractions (Fig. 3c). However, no significant genetic ancestry-based differences were found in tumor-infiltrating T cell fractions (Fig. 3d).

We examined 1,635 SNPs known to influence circulating leukocyte traits[27] to evaluate whether these could explain differences in lymphocyte fractions between individuals (Methods and Supplementary Data). After accounting for linkage disequilibrium (LD), 15 SNPs were significantly associated with circulating T cell fraction in the healthy European cohort, but only one SNP in the European cancer cohort (Fig. 3e). The Duffy-negative SNP rs2814778, linked to neutropenia[30], was significant in the African healthy cohort but not in the African cancer cohort (Fig. 3e). Only 7 out of 15 significant SNPs in the healthy European cohort were also associated with immune cell levels (unadjusted $P < 0.05$) in the European cancer cohort (Extended Data Fig. 7a,b). This suggests that germline SNPs influencing T cell fraction differ between healthy and cancer contexts (Supplementary Fig. 10).

Quantifying selection pressure due to immune infiltration

We reasoned that the immune system may act as a potent selection pressure during cancer evolution. We, therefore, investigated whether mutations in known cancer genes were associated with tumor immune infiltrate.

Using a Poisson model[31], controlling for the background mutation rate, cancer type, sex and tumor purity, we identified seven genes significantly associated with infiltrating T cell fraction. Nonsynonymous mutations in *PIK3CA, MAP3K1, PTEN, CBFB* and *CDH1* were enriched in T cell-depleted tumors, while *MUC16, B2M* and *BAP1* mutations were enriched in T cell-replete tumors (Fig. 4a and Supplementary Data). Additionally, *MUC4* mutations were associated with IgM/IgD B cell-depleted tumors, while *KMT2C* mutations were linked to IgG B cell-enriched tumors (Extended Data Fig. 8a and Supplementary Data).

Disease-specific effects were also identified (Fig. 4b). For example, *TP53* nonsynonymous mutations were associated with increased T cell infiltrate in breast invasive carcinoma estrogen receptor-positive (BRCA ER$^+$) tumors (estimate = 2.56, adjusted $P = 6 \times 10^{-4}$), while *PIK3R1* nonsynonymous mutations were associated with T cell-enriched glioblastomas (estimate = 7.23, adjusted $P = 0.003$). Previous studies have linked *TP53* alterations with increased T cell infiltrate in breast cancer[32]. Many disease-specific effects were also detected for B cell subsets. *MUC4* nonsynonymous mutations were significantly associated with reduced IgM/IgD B cells in colon adenocarcinoma (estimate = −3.7, adjusted $P = 0.007$). *NOTCH1* nonsynonymous mutations were associated with class-switched and IgA B cells in BRCA ER$^+$ tumors (estimate = 10.6, adjusted $P = 0.003$), and nonsynonymous mutations in *DGRC8* were associated with enriched IgA B cells in lung adenocarcinoma (estimate = 18.7, adjusted $P = 0.04$). *MUC4* and *KMT2C* were both found to be significant in uterine corpus endometrial carcinoma, being associated with tumors enriched with IgA (*MUC4*: estimate = 3.58, adjusted $P = 0.049$; *KMT2C*: estimate = 4.80, adjusted $P = 0.01$) and depleted of IgG (*MUC4*: estimate = −4.85, adjusted $P = 8 \times 10^{-5}$; *KMT2C*: estimate = −3.82, $P = 0.007$, adjusted $P = 1$; Extended Data Fig. 8b and Supplementary Data).

These findings reveal diverse interactions between the tumor immune microenvironment and cancer cells. Somatic mutations may be selected in response to immune presence or promote immunosuppression, with effects varying by cancer type, histology and immune cell type.

Prognostic value of ImmuneLENS lymphocyte fraction

Infiltrating tumor lymphocytes[6,33,34] and circulating immune cells[10,35–38] have prognostic value in many cancer types. However, a direct comparison of both has proven challenging. We therefore used ImmuneLENS to compare the prognostic significance of circulating and tumor-infiltrating T cell and B cell fractions.

In the pan-cancer cohort, elevated circulating T cell fractions were strongly associated with improved overall survival (Fig. 5a; hazard ratio (HR) = 0.53, log-rank $P = 2.8 \times 10^{-73}$, split by the median). Infiltrating tumor T cells showed a significantly weaker association (Fig. 5a; HR = 0.86, log-rank $P = 3.2 \times 10^{-6}$; $P = 2.68 \times 10^{-25}$, z test on the log of the HR values). Elevated circulating *IGH* B cells (HR = 0.79, $P = 2 \times 10^{-12}$) and IgM/IgD B cells (HR = 0.76, $P = 4 \times 10^{-16}$) were associated with better prognosis while infiltrating B cells and class-switched circulating B cells were not significant. The circulating T/B cell ratio, which is independent of neutrophil levels, was also significantly prognostic (Fig. 5a; HR = 0.76, log-rank $P = 8.3 \times 10^{-14}$).

Circulating T cell fraction remained highly significant after adjusting for clinical factors in the pan-cancer cohort (Fig. 5b (HR = 0.76, $P = 2.6 \times 10^{-46}$) and Extended Data Fig. 9a) and within five individual cancer types after multiple hypothesis correction—head and neck

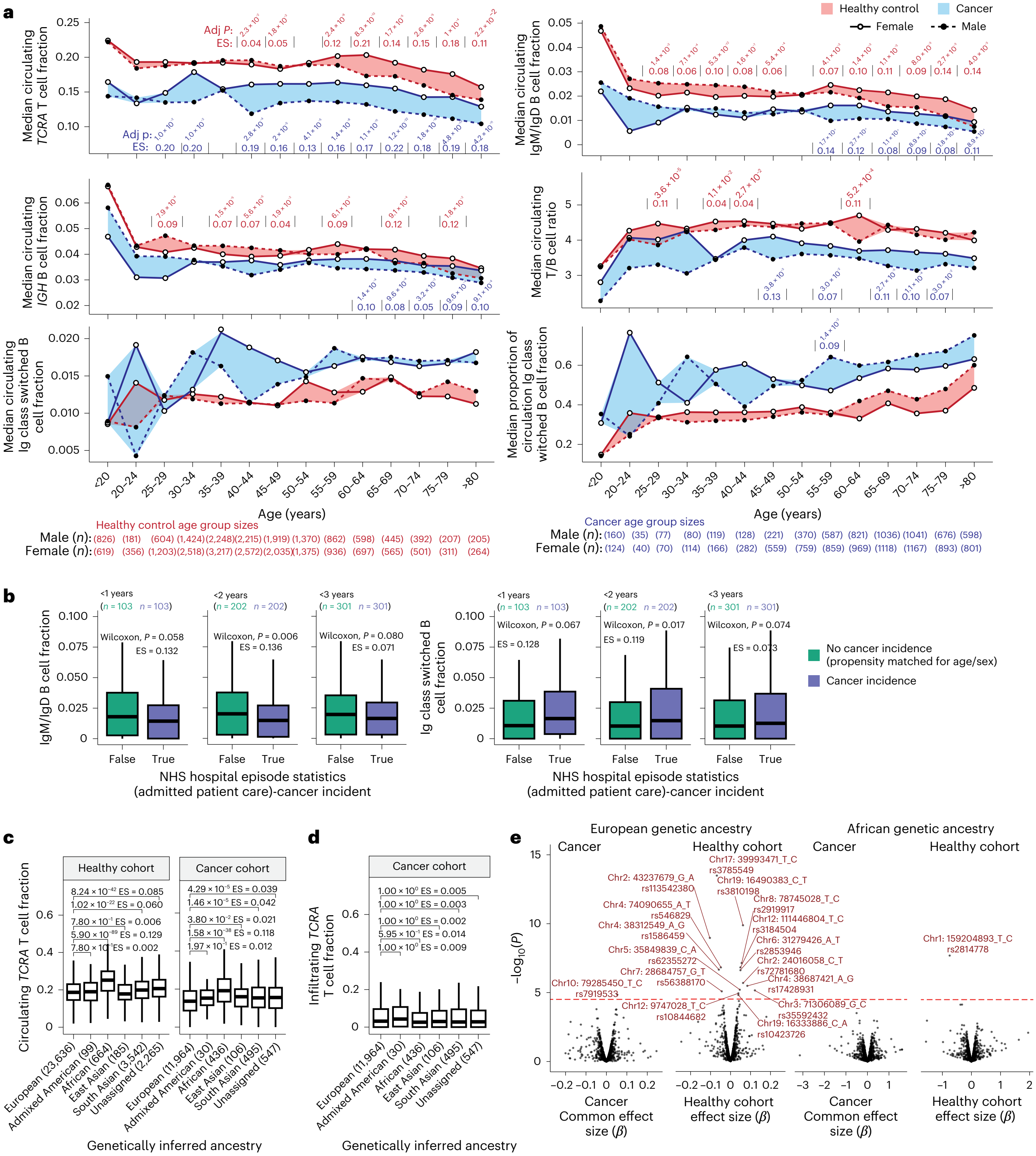


Fig. 3 | Disruption of circulating T cell fraction in patients with cancer. a, Ribbon plots of ImmuneLENS-related fractions in 5-year age brackets, split by the healthy control and cancer cohorts. The width of bands represents the extent of sexual dimorphism between male and female individuals, with significance assessed by two-sided Wilcoxon rank-sum tests within each age group and adjusted *P* values with effect size (ES) values shown. **b**, Boxplots of IgM/IgD and Ig class-switched B cell fractions from a subset of the healthy cohort with recorded cancer incidence post-WGS sequencing (from hospital episode statistics), compared to an age- and sex-matched propensity cohort of the same size. **c**, Boxplots of blood *TCRA* T cell fraction versus genetically inferred ancestry in the 100KGP healthy and cancer cohorts. **d**, Boxplots for tumor *TCRA* T cell fraction versus genetically inferred ancestry in the 100KGP cancer cohort. **e**, Volcano plots of known GWAS SNP associations with circulating *TCRA* T cell fraction. Multiple hypothesis adjustments were performed using the Benjamini–Hochberg method. Boxplots in **b**–**d** show the median and lower and upper quartiles, with whiskers extending to 1.5× interquartile range. Two-sided Wilcoxon rank-sum tests were used to assess the significance between groups in **b**–**d**. The *P* values in **e** are derived from PLINK software, which uses a linear regression model and performs a Wald test for each SNP. For the cancer cohort, this was done separately for each histology, and *P* values were combined using a meta-analysis with a common effects model.

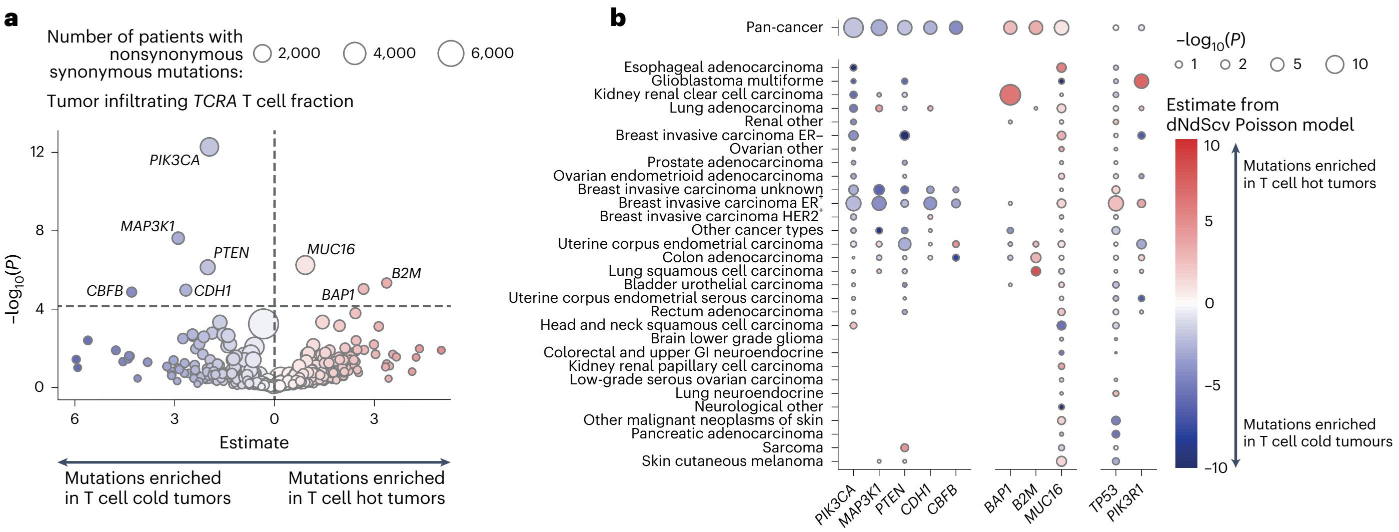


Fig. 4 | Association of selection with infiltrating T cell fraction. a, Volcano plot showing results from a Poisson model predicting observed nonsynonymous mutations as a function of *TCRA* T cell fraction and other covariates (age, tumor mutation burden, sex and disease type). The plot shows the estimates and $-\log_{10}(P)$ for the *TCRA* variable, highlighting genes where observed mutations significantly depend on T cell immune infiltration (hot tumors denote high levels of immune cell infiltration; cold tumors lack immune cell infiltration). Point size represents the number of patients in the cancer cohort with nonsynonymous mutations, excluding patients with hematological cancers. Genes tested were limited to known cancer drivers from the Cancer Gene Census[47]. **b**, Bubble plot showing the significance of *TCRA* infiltrating T cell fraction within a Poisson model applied to individual cancer types. In both plots (**a**,**b**), *P* values represent the significance of the *TCRA* T cell fraction term in the Poisson model and are calculated using a Wald test.

squamous cell carcinoma (HR = 0.70, adjusted $P = 4.68 \times 10^{-2}$), pancreatic adenocarcinoma (HR = 0.74, adjusted $P = 4.81 \times 10^{-2}$), colon adenocarcinoma (HR = 0.85, adjusted $P = 1.12 \times 10^{-2}$), lung adenocarcinoma (HR = 0.80, adjusted $P = 5.4 \times 10^{-3}$) and sarcoma (HR = 0.68, adjusted $P = 1.57 \times 10^{-4}$). Conversely, tumor-infiltrating T cell fraction was only significant in the pan-cancer cohort (HR = 0.93, adjusted $P = 3.99 \times 10^{-3}$) and B cell hematological cancers, where higher infiltration correlated with worse prognosis (HR = 1.24, adjusted $P = 3.99 \times 10^{-3}$). We observed broadly consistent results using TCGA data (Extended Data Fig. 10a,b). Moreover, we observed that the infiltrating T cell fraction was associated with significant heterogeneity by cancer type ($P = 0.02$, Cochran's *Q* test).

Circulating T/B cell ratio remained significantly associated with survival after controlling for clinical variables (HR = 0.90, adjusted $P = 3.4 \times 10^{-6}$), suggesting that both circulating T cells and neutrophils contribute to the prognostic value of circulating T cell fraction. A Cox proportional hazard (CoxPH) model including circulating T cell fraction, T/B cell ratio and their interaction revealed independent significance for both factors (*TCRA*: HR = 0.7, $P = 9 \times 10^{-41}$; T/B ratio: HR = 0.91, $P = 0.014$; interaction term: HR = 1.24, $P = 2 \times 10^{-8}$; Extended Data Fig. 9b). When stratifying patients by T/B cell ratio, circulating T cell fraction was prognostic in both the low (HR = 0.45, 95% confidence interval (CI): 0.40–0.51; Extended Data Fig. 9c) and high T/B cell ratio group (HR = 0.67, 95% CI: 0.60–0.75; Extended Data Fig. 9c), but was significantly more prognostic in the low group ($P = 2.4 \times 10^{-6}$, *z* test). When stratifying patients by circulating T cell fraction, higher T/B ratio correlated with better outcome in the low T cell group (HR = 0.82, $P = 2 \times 10^{-4}$; Extended Data Fig. 9d), while the opposite was true for the high T cell group (HR = 1.2, $P = 0.0013$; Extended Data Fig. 9d). Thus, this suggests that in patients with relatively low neutrophil levels in their blood, B cells confer an improved prognosis compared to T cells.

We next investigated sex-specific prognostic associations of circulating T cell fraction (Fig. 5b). In the pan-cancer context, both male (HR = 0.73, $P = 1.5 \times 10^{-27}$) and female (HR = 0.79, $P = 2.6 \times 10^{-20}$) individuals showed similar associations between high T cell fraction and improved prognosis. However, cancer-specific differences were evident. For example, higher T cell fraction was associated with a significantly better prognosis for females in bladder urothelial carcinoma (HR = 0.28, adjusted $P = 2.53 \times 10^{-3}$) and lung adenocarcinoma (HR = 0.72, adjusted $P = 2.53 \times 10^{-3}$). A random effects meta-analysis showed circulating T cell fraction had a uniform prognostic effect across cancer types ($I^2 = 4\%$, not significant). However, sex-specific analysis revealed significant heterogeneity (male: $I^2 = 46\%$, $P = 0.006$; female: $I^2 = 34\%$, $P = 0.03$).

Considered together, these findings highlight the clinical importance of circulating lymphocytes and the interplay between biological sex and immune activation in different cancers.

Discussion

We introduce ImmuneLENS, a method for inferring immune cell fractions from WGS data. Building upon our previous method T cell ExTRECT, ImmuneLENS not only provides more accurate T cell content inference but also expands functionality to measure B cell fraction, B cell class switching and T cell clonotype diversity.

To evaluate this approach, we applied it to the 100KGP cohort. In patients with cancer, circulating T cell and B cell fractions were reduced compared to healthy controls, and female individuals exhibited higher T cell fractions than male individuals, indicating significant sexual dimorphism. Although healthy controls also show sexual dimorphism, it appears mainly after age 55. Thus, ImmuneLENS quantified the sexual dimorphism in circulating immune cells, which is known to occur in both aging[39] and cancer[40].

In the healthy control cohort, we found significant associations between circulating T cell fractions and 15 SNPs at 11 genetic loci. This is consistent with recent work discussed in ref. 41, in which T cell ExTRECT[6] was used on WGS data from 207,000 individuals and 27 loci associated with circulating T cell fraction were identified. Our analysis also identified an SNP within *FOXP1* (rs35592432; $P = 7.3 \times 10^{-6}$) that was not reported in ref. 41, with the variant allele associated with an enrichment of T cells. We observed that more than half of SNPs (8/15) linked to circulating T cell fractions in healthy individuals were not linked within the cancer population. This discrepancy highlights the need for further research on how germline genetics influence immune composition in cancer.

https://doi.org/10.1038/s41588-025-02086-5

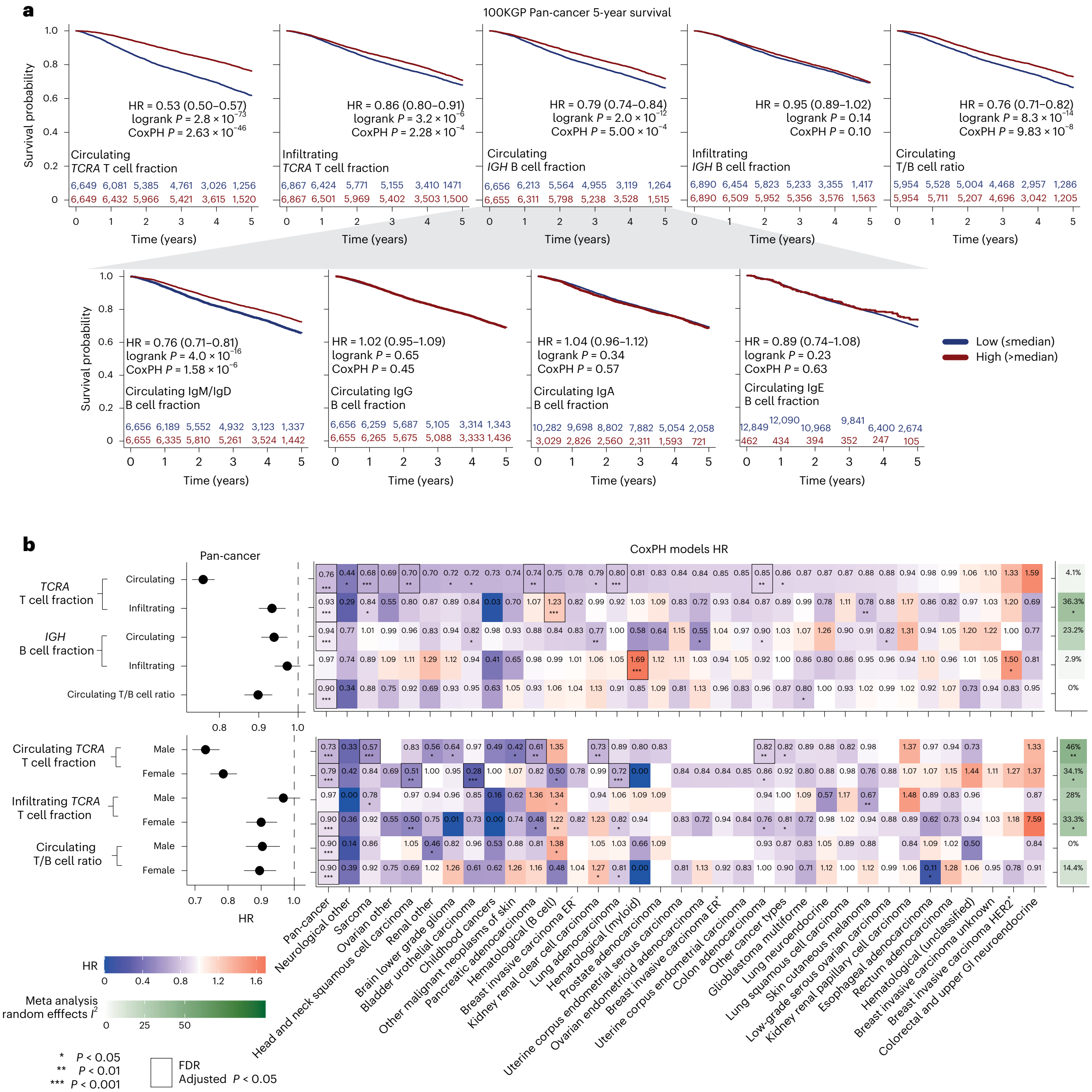


Fig. 5 | Prognostic value of ImmuneLENS lymphocyte fractions in 100KGP. **a**, Five-year survival Kaplan–Meier plots for the entire pan-cancer 100KGP cohort, stratified into high and low groups based on the median circulating or infiltrating *TCRA* T cell fractions and *IGH* B cell fractions. **b**, Results from CoxPH models for 13,872 participants within the 100KGP pan-cancer cohort with complete clinical annotation. The models account for the effects of age, sex, genetically inferred ancestry, pretreatment chemotherapy and cancer stage. Left, pan-cancer HRs with 95% CIs. Right, a heatmap of HRs for different cancer histologies, including the I^2 score from a meta-analysis using a random effects model across all histologies. Significance was calculated using Cochran's *Q* test. Multiple hypothesis adjustments were performed using the Benjamini–Hochberg method, applied by row. Individual *P* values were calculated using a two-sided Wald test within the Cox model. *$P < 0.05$, **$P < 0.01$ and ***$P < 0.001$.

We found that circulating T cell fractions in patients with cancer are more prognostic than tumor-infiltrating T cells. This relationship may reflect systemic inflammation from tumor growth[42]. The association of the T/B cell ratio with prognosis suggests lymphocyte depletion, not just increased neutrophils, contributes to this signal. Circulating lymphocyte fractions may indicate 'immunological age'[43], reflecting the diminishing ability of the immune system to suppress cancer. This hypothesis is supported by the reduction we observed in circulating IgM/IgD B cell fractions of healthy participants who later developed cancer; however, we found no similar association with T cell fractions.

Despite the technological advancement of ImmuneLENS, there are limitations. While ImmuneLENS can accurately estimate immune cell fractions at ≥5× WGS coverage, additional sequencing coverage is required for B cell germline copy number correction (>10×) and somatic copy number adjustment (>20×). Moreover, TCR repertoire analysis from WGS data lacks the accuracy to make it comparable to TCR-seq, and predicting exact clonotypes remains a challenge. We do not envisage this method replacing TCR-seq. Rather, it can provide orthogonal insight that was previously lacking in the TCR repertoire in samples with solely WGS. ImmuneLENS can accurately estimate B cell fractions and deconvolve the separate class-switched fractions. This has much potential outside the context of cancer, for instance, within autoimmunity research[44–46].

In summary, with the growing size of population-level WGS datasets, we have provided a tool that can accurately quantify lymphocyte fraction without the need for additional data collection. We hope ImmuneLENS will enable a deeper exploration of immune dysregulation.

Online content

Any methods, additional references, Nature Portfolio 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/s41588-025-02086-5.

References

1. Gonzalez, H., Hagerling, C. & Werb, Z. Roles of the immune system in cancer: from tumor initiation to metastatic progression. *Genes Dev.* **32**, 1267–1284 (2018).
2. Rosenthal, R., Swanton, C. & McGranahan, N. Understanding the impact of immune-mediated selection on lung cancer evolution. *Br. J. Cancer* **124**, 1615–1617 (2021).
3. McGranahan, N. et al. Allele-specific HLA loss and immune escape in lung cancer evolution. *Cell* **171**, 1259–1271 (2017).
4. Kloor, M., Michel, S. & von Knebel Doeberitz, M. Immune evasion of microsatellite unstable colorectal cancers. *Int. J. Cancer* **127**, 1001–1010 (2010).
5. Juneja, V. R. et al. PD-L1 on tumor cells is sufficient for immune evasion in immunogenic tumors and inhibits CD8 T cell cytotoxicity. *J. Exp. Med.* **214**, 895–904 (2017).
6. Bentham, R. et al. Using DNA sequencing data to quantify T cell fraction and therapy response. *Nature* **597**, 555–560 (2021).
7. Barnes, T. A. & Amir, E. HYPE or HOPE: the prognostic value of infiltrating immune cells in cancer. *Br. J. Cancer* **118**, e5 (2018).
8. Litchfield, K. et al. Meta-analysis of tumor- and T cell-intrinsic mechanisms of sensitization to checkpoint inhibition. *Cell* **184**, 596–614 (2021).
9. Fridman, W. H. et al. B cells and cancer: to B or not to B? *J. Exp. Med.* **218**, e20200851 (2021).
10. Ying, H.-Q. et al. The prognostic value of preoperative NLR, d-NLR, PLR and LMR for predicting clinical outcome in surgical colorectal cancer patients. *Med. Oncol.* **31**, 305 (2014).
11. Cedrés, S. et al. Neutrophil to lymphocyte ratio (NLR) as an indicator of poor prognosis in stage IV non-small cell lung cancer. *Clin. Transl. Oncol.* **14**, 864–869 (2012).
12. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
13. Jamal-Hanjani, M. et al. Tracking the evolution of non-small-cell lung cancer. *N. Engl. J. Med.* **376**, 2109–2121 (2017).
14. Levin-Klein, R. & Bergman, Y. Epigenetic regulation of monoallelic rearrangement (allelic exclusion) of antigen receptor genes. *Front. Immunol.* **5**, 625 (2014).
15. Pistoia, V. et al. Human γδ T-cells: from surface receptors to the therapy of high-risk leukemias. *Front. Immunol.* **9**, 984 (2018).
16. Kang, J., Baker, J. & Raulet, D. H. Evidence that productive rearrangements of TCRγ genes influence the commitment of progenitor cells to differentiate into αβ or γδ T cells. *Eur. J. Immunol.* **25**, 2706–2709 (1995).
17. Danaher, P. et al. Gene expression markers of tumor infiltrating leukocytes. *J. Immunother. Cancer* **5**, 18 (2017).
18. Rosenthal, R. et al. Neoantigen-directed immune escape in lung cancer evolution. *Nature* **567**, 479–485 (2019).
19. Li, T. et al. TIMER: a web server for comprehensive analysis of tumor-infiltrating immune cells. *Cancer Res.* **77**, e108–e110 (2017).
20. Newman, A. M. et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 (2015).
21. 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).
22. Bolotin, D. A. et al. MiXCR: software for comprehensive adaptive immunity profiling. *Nat. Methods* **12**, 380–381 (2015).
23. Caulfield, M. et al. National Genomic Research Library. *Figshare* https://doi.org/10.6084/m9.figshare.4530893.v7 (2017).
24. Liles, W. C., Dale, D. C. & Klebanoff, S. J. Glucocorticoids inhibit apoptosis of human neutrophils. *Blood* **86**, 3181–3188 (1995).
25. Godfrey, D. I., Koay, H.-F., McCluskey, J. & Gherardin, N. A. The biology and functional importance of MAIT cells. *Nat. Immunol.* **20**, 1110–1128 (2019).
26. Nédélec, Y. et al. Genetic ancestry and natural selection drive population differences in immune responses to pathogens. *Cell* **167**, 657–669 (2016).
27. Mikhaylova, A. V. et al. Whole-genome sequencing in diverse subjects identifies genetic correlates of leukocyte traits: the NHLBI TOPMed program. *Am. J. Hum. Genet.* **108**, 1836–1851 (2021).
28. Chao, B. N., Carrick, D. M., Filipski, K. K. & Nelson, S. A. Overview of research on germline genetic variation in immune genes and cancer outcomes. *Cancer Epidemiol. Biomarkers Prev.* **31**, 495–506 (2022).
29. 1000 Genomes Project Consortium. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
30. Reich, D. et al. Reduced neutrophil count in people of African descent is due to a regulatory variant in the Duffy antigen receptor for chemokines gene. *PLoS Genet.* **5**, e1000360 (2009).
31. Martincorena, I. et al. Universal patterns of selection in cancer and somatic tissues. *Cell* **173**, 1823 (2018).
32. Liu, Z. et al. TP53 mutations promote immunogenic activity in breast cancer. *J. Oncol.* **2019**, 5952836 (2019).
33. Nosho, K. et al. Tumour-infiltrating T-cell subsets, molecular changes in colorectal cancer, and prognosis: cohort study and literature review. *J. Pathol.* **222**, 350–366 (2010).
34. Hwang, W.-T., Adams, S. F., Tahirovic, E., Hagemann, I. S. & Coukos, G. Prognostic significance of tumor-infiltrating T cells in ovarian cancer: a meta-analysis. *Gynecol. Oncol.* **124**, 192–198 (2012).
35. Kumarasamy, C. et al. Prognostic significance of blood inflammatory biomarkers NLR, PLR, and LMR in cancer—a protocol for systematic review and meta-analysis. *Medicine (Baltimore)* **98**, e14834 (2019).
36. Ferrucci, P. F. et al. Baseline neutrophils and derived neutrophil-to-lymphocyte ratio: prognostic relevance in metastatic melanoma patients receiving ipilimumab. *Ann. Oncol.* **27**, 732–738 (2016).
37. Bagley, S. J. et al. Pretreatment neutrophil-to-lymphocyte ratio as a marker of outcomes in nivolumab-treated patients with advanced non-small-cell lung cancer. *Lung Cancer* **106**, 1–7 (2017).
38. Bartlett, E. K. et al. High neutrophil-to-lymphocyte ratio (NLR) is associated with treatment failure and death in patients who have melanoma treated with PD-1 inhibitor monotherapy. *Cancer* **126**, 76–85 (2020).

39. Márquez, E. J. et al. Sexual-dimorphism in human immune system aging. *Nat. Commun.* **11**, 751 (2020).
40. Clocchiatti, A., Cora, E., Zhang, Y. & Dotto, G. P. Sexual dimorphism in cancer. *Nat. Rev. Cancer* **16**, 330–339 (2016).
41. Poisner, H., Faucon, A., Cox, N. & Bick, A. G. Genetic determinants and phenotypic consequences of blood T-cell proportions in 207,000 diverse individuals. *Nat. Commun.* **15**, 6732 (2024).
42. Zahorec, R. Neutrophil-to-lymphocyte ratio, past, present and future perspectives. *Bratisl. Lek. Listy* **122**, 474–488 (2021).
43. Alpert, A. et al. A clinically meaningful metric of immune age derived from high-dimensional longitudinal monitoring. *Nat. Med.* **25**, 487–495 (2019).
44. Duddy, M. et al. Distinct effector cytokine profiles of memory and naive human B cell subsets and implication in multiple sclerosis. *J. Immunol.* **178**, 6092–6099 (2007).
45. Elsner, R. A. & Shlomchik, M. J. Germinal center and extrafollicular b cell responses in vaccination, immunity, and autoimmunity. *Immunity* **53**, 1136–1150 (2020).
46. Tipton, C. M. et al. Diversity, cellular origin and autoreactivity of antibody-secreting cell population expansions in acute systemic lupus erythematosus. *Nat. Immunol.* **16**, 755–765 (2015).
47. Sondka, Z. et al. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* **18**, 696–705 (2018).

TRACERx Consortium

Charles Swanton[2,3,7], Mariam Jamal-Hanjani[2,7,8], Carlos Martínez-Ruiz[1,2], Kerstin Thol[1,2], Maria Litovchenko[1,2], Nicholas McGranahan[1,2], Robert Bentham[1,2], Thomas P. Jones[1,2], Chris Bailey[3], Oriol Pich[3], Thomas B. K. Watkins[4], Peter Van Loo[5,6], James R. M. Black[2,3], Takahiro Karasaki[2,3,8,9], Abigail Bunkum[2,8,10], Sonya Hessey[2,8,10], Wing Kin Liu[2,8], Nicolai J. Birkbak[2,3,11,12,13], Alexander M. Frankell[2,3], Ariana Huebner[1,2,3], Clare Puttick[1,2,3], Crispin T. Hiley[2,3], David A. Moore[2,3,14], Dhruva Biswas[2,3,15], Emilia L. Lim[2,3], Kristiana Grigoriadis[1,2,3], Maise Al Bakir[2,3], Olivia Lucas[2,3,10,16], Roberto Vendramin[2,3,17], Sophia Ward[2,3,18], Sian Harries[2,3,18], Simone Zaccaria[2,10], Rija Zaidi[2,10], Lucrezia Patruno[2,10], Despoina Karagianni[2,19], Sergio A. Quezada[2,19], Supreet Kaur Bola[2,19], Martin D. Forster[2,7], Siow Ming Lee[2,7], Corentin Richard[2], Cristina Naceur-Lombardelli[2], Francisco Gimeno-Valiente[2], Krupa Thakkar[2], Monica Sivakumar[2], Nnennaya Kanu[2], Ieva Usaite[2], Sadegh Saghafinia[2], Selvaraju Veeriah[2], Sharon Vanloo[2], Antonia Toncheva[2], Paulina Prymas[2], Bushra Mussa[2], Michalina Magala[2], Elizabeth Keene[2], Michelle M. Leung[1,2,3], Gareth A. Wilson[3], Rachel Rosenthal[3], Andrew Rowan[3], Claudia Lee[3], Emma Colliver[3], Katey S. S. Enfield[3], Mihaela Angelova[3], Cian Murphy[3], Maria Zagorulya[3], Teresa Marafioti[14], Elaine Borg[14], Mary Falzon[14], Reena Khiroya[14], Yien Ning Sophia Wong[16,20], Emilie Martinoni Hoogenboom[16], Fleur Monk[16], James W. Holding[16], Junaid Choudhary[16], Kunal Bhakhri[16], Pat Gorman[16], Robert C. M. Stephens[16], Maria Chiara Pisciella[16], Steve Bandula[16], Jerome Nicod[18], Angela Dwornik[21], Angeliki Karamani[21], Benny Chain[21], David R. Pearce[21], Georgia Stavrou[21], Gerasimos-Theodoros Mastrokalos[21], Helen L. Lowe[21], James L. Reading[21], John A. Hartley[21], Kayalvizhi Selvaraju[21], Leah Ensell[21], Mansi Shah[21], Piotr Pawlik[21], Samuel Gamble[21], Seng Kuong Anakin Ung[21], Victoria Spanswick[21], Yin Wu[21], Jason F. Lester[22], Sean Dulloo[23,24], Dean A. Fennell[23,24], Amrita Bajaj[24], Apostolos Nakas[24], Azmina Sodha-Ramdeen[24], Mohamad Tufail[24], Molly Scotland[24], Rebecca Boyles[24], Sridhar Rathinam[24], Claire Wilson[25], Gurdeep Matharu[26], Jacqui A. Shaw[26], Ekaterini Boleti[27], Heather Cheyne[28], Mohammed Khalil[28], Shirley Richardson[28], Tracey Cruickshank[28], Gillian Price[29,30], Keith M. Kerr[30,31], Sarah Benafif[7,32], Dionysis Papadatos-Pastos[7], James Wilson[7], Tanya Ahmad[7], Jack French[32], Kayleigh Gilbert[32], Babu Naidu[33], Akshay J. Patel[34], Gary Middleton[35,36], Aya Osman[35], Mandeesh Sangha[35], Gerald Langman[35], Helen Shackleford[35], Madava Djearaman[35], Angela Leek[37], Jack Davies Hodgkinson[37], Nicola Totton[37], Philip Crosbie[38,39,40], Eustace Fontaine[38], Felice Granato[38], Juliette Novasio[38], Kendadai Rammohan[38], Leena Joseph[38], Paul Bishop[38], Vijay Joshi[38], Sara Waplington[38], Adam Atkin[38], Katherine D. Brown[40,41], Mathew Carter[40,41], Anshuman Chaturvedi[40,41], Pedro Oliveira[40,41], Colin R. Lindsay[40,42], Fiona H. Blackhall[40,42], Yvonne Summers[40,42], Matthew G. Krebs[42], Antonio Paiva-Correia[43], Jonathan Tugwood[40,44], Caroline Dive[40,44], Hugo J. W. L. Aerts[45,46,47], Roland F. Schwarz[48,49], Tom L. Kaufmann[49,50], Zoltan Szallasi[51,52,53], Miklos Diossy[51,52,54], Roberto Salgado[55,56], Jonas Demeulemeester[57,58,59], Carla Castignani[60,61], Stephan Beck[61], George Kassiotis[62,63], Imran Noorani[62,64,65], Clare E. Weeden[62], Eva Grönroos[62], Jacki Goldman[62], Mickael Escudero[62], Philip Hobson[62], Stefan Boeing[62], Tamara Denner[62], Vittorio Barbè[62], Wei-Ting Lu[62], William Hill[62], Yutaka Naito[62], Erik Sahai[62], Zoe Ramsden[62], Emma Nye[66], Richard Kevin Stone[66], Jayant K. Rane[3,21], Jeanette Kittel[2,8],

Kerstin Haase[2,8], Kexin Koh[2,8], Rachel Scott[2,8], Karl S. Peggs[67,68], Catarina Veiga[69], Gary Royle[70], Charles-Antoine Collins-Fekete[70], Francesco Fraioli[71], Paul Ashford[72], Arjun Nair[73,74], Alexander James Procter[73], Asia Ahmed[73], Magali N. Taylor[73], David Lawrence[75], Davide Patrini[75], Neal Navani[76,77], Ricky M. Thakrar[76,77], Sam M. Janes[77], Zoltan Kaplar[78,79], Allan Hackshaw[80], Camilla Pilotti[80], Rachel Leslie[80], Anne-Marie Hacker[80], Sean Smith[80], Aoife Walker[80], Anca Grapa[81], Hanyun Zhang[82], Khalid AbdulJabbar[83], Xiaoxi Pan[84], Yinyin Yuan[84], David Chuter[85], Mairead MacKenzie[85], Serena Chee[86], Patricia Georg[86], Aiman Alzetani[87], Judith Cave[88], Eric Lim[89,90], Andrew G. Nicholson[90,91], Paulo De Sousa[90], Simon Jordan[90], Alexandra Rice[90], Hilgardt Raubenheimer[90], Harshil Bhayani[90], Lyn Ambrose[90], Anand Devaraj[90], Hema Chavan[90], Sofina Begum[90], Silviu I. Buderi[90], Daniel Kaniu[90], Mpho Malima[90], Sarah Booth[90], Nadia Fernandes[90], Pratibha Shah[90], Chiara Proli[90], Madeleine Hewish[92,93], Sarah Danson[94,95], Michael J. Shackcloth[96], Lily Robinson[97], Peter Russell[97], Kevin G. Blyth[98,99,100], Andrew Kidd[101], Craig Dick[102], John Le Quesne[103,104,105], Alan Kirk[106], Mo Asif[106], Rocco Bilancia[106], Nikos Kostoulas[106], Jennifer Whiteley[106] & Mathew Thomas[106]

[8]Cancer Metastasis Laboratory, University College London Cancer Institute, London, UK. [9]Department of Thoracic Surgery, Respiratory Center, Toranomon Hospital, Tokyo, Japan. [10]Computational Cancer Genomics Research Group, University College London Cancer Institute, London, UK. [11]Department of Molecular Medicine, Aarhus University Hospital, Aarhus, Denmark. [12]Department of Clinical Medicine, Aarhus University, Aarhus, Denmark. [13]Bioinformatics Research Centre, Aarhus University, Aarhus, Denmark. [14]Department of Cellular Pathology, University College London Hospitals, London, UK. [15]Bill Lyons Informatics Centre, University College London Cancer Institute, London, UK. [16]University College London Hospitals, London, UK. [17]Tumour Immunogenomics and Immunosurveillance Laboratory, University College London Cancer Institute, London, UK. [18]Genomics Science Technology Platform, The Francis Crick Institute, London, UK. [19]Immune Regulation and Tumour Immunotherapy Group, Cancer Immunology Unit, Research Department of Haematology, University College London Cancer Institute, London, UK. [20]National Cancer Centre, Singapore City, Singapore. [21]University College London Cancer Institute, London, UK. [22]Singleton Hospital, Swansea Bay University Health Board, Swansea, UK. [23]University of Leicester, Leicester, UK. [24]University Hospitals of Leicester NHS Trust, Leicester, UK. [25]Leicester Medical School, University of Leicester, Leicester, UK. [26]Cancer Research Centre, University of Leicester, Leicester, UK. [27]Royal Free London NHS Foundation Trust, London, UK. [28]Aberdeen Royal Infirmary NHS Grampian, Aberdeen, UK. [29]Department of Medical Oncology, Aberdeen Royal Infirmary NHS Grampian, Aberdeen, UK. [30]University of Aberdeen, Aberdeen, UK. [31]Department of Pathology, Aberdeen Royal Infirmary NHS Grampian, Aberdeen, UK. [32]The Whittington Hospital NHS Trust, London, UK. [33]Birmingham Acute Care Research Group, Institute of Inflammation and Ageing, University of Birmingham, Birmingham, UK. [34]Guy's and St Thomas' NHS Foundation Trust, London, UK. [35]University Hospital Birmingham NHS Foundation Trust, Birmingham, UK. [36]Institute of Immunology and Immunotherapy, University of Birmingham, Birmingham, UK. [37]Manchester Cancer Research Centre Biobank, Manchester, UK. [38]Wythenshawe Hospital, Manchester University NHS Foundation Trust, Wythenshawe, UK. [39]Division of Infection, Immunity and Respiratory Medicine, University of Manchester, Manchester, UK. [40]Cancer Research UK Lung Cancer Centre of Excellence, University of Manchester, Manchester, UK. [41]The Christie NHS Foundation Trust, Manchester, UK. [42]Division of Cancer Sciences, The University of Manchester and The Christie NHS Foundation Trust, Manchester, UK. [43]Manchester University NHS Foundation Trust, Manchester, UK. [44]CRUK Manchester Institute Cancer Biomarker Centre, University of Manchester, Manchester, UK. [45]Artificial Intelligence in Medicine (AIM) Program, Mass General Brigham, Harvard Medical School, Boston, MA, USA. [46]Department of Radiation Oncology, Brigham and Women's Hospital, Dana–Farber Cancer Institute, Harvard Medical School, Boston, MA, USA. [47]Radiology and Nuclear Medicine, CARIM & GROW, Maastricht University, Maastricht, The Netherlands. [48]Institute for Computational Cancer Biology, Center for Integrated Oncology (CIO), Cancer Research Center Cologne Essen (CCCE), Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany. [49]Berlin Institute for the Foundations of Learning and Data (BIFOLD), Berlin, Germany. [50]Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC), Berlin, Germany. [51]Danish Cancer Institute, Copenhagen, Denmark. [52]Computational Health Informatics Program, Boston Children's Hospital, Boston, MA, USA. [53]Department of Bioinformatics, Semmelweis University, Budapest, Hungary. [54]Department of Physics of Complex Systems, ELTE Eötvös Loránd University, Budapest, Hungary. [55]Department of Pathology, ZAS Hospitals, Antwerp, Belgium. [56]Division of Research, Peter MacCallum Cancer Centre, Melbourne, Victoria, Australia. [57]Integrative Cancer Genomics Laboratory, VIB Center for Cancer Biology, Leuven, Belgium. [58]VIB Center for AI & Computational Biology, Leuven, Belgium. [59]Department of Oncology, KU Leuven, Leuven, Belgium. [60]Cancer Genomics Laboratory, The Francis Crick Institute, London, UK. [61]Medical Genomics, University College London Cancer Institute, London, UK. [62]The Francis Crick Institute, London, UK. [63]Department of Infectious Disease, Faculty of Medicine, Imperial College London, London, UK. [64]Department of Neurosurgery, National Hospital for Neurology and Neurosurgery, London, UK. [65]Institute of Neurology, University College London, London, UK. [66]Experimental Histopathology, The Francis Crick Institute, London, UK. [67]Department of Haematology, University College London Hospitals, London, UK. [68]Cancer Immunology Unit, Research Department of Haematology, University College London Cancer Institute, London, UK. [69]Centre for Medical Image Computing, Department of Medical Physics and Biomedical Engineering, London, UK. [70]Department of Medical Physics and Bioengineering, University College London Cancer Institute, London, UK. [71]Institute of Nuclear Medicine, Division of Medicine, University College London, London, UK. [72]Institute of Structural and Molecular Biology, University College London, London, UK. [73]Department of Radiology, University College London Hospitals, London, UK. [74]UCL Respiratory, Department of Medicine, University College London, London, UK. [75]Department of Thoracic Surgery, University College London Hospital NHS Trust, London, UK. [76]Department of Thoracic Medicine, University College London Hospitals, London, UK. [77]Lungs for Living Research Centre, UCL Respiratory, Department of Medicine, University College London, London, UK. [78]Integrated Radiology Department, North-Buda St John's Central Hospital, Budapest, Hungary. [79]Institute of Nuclear Medicine, University College London Hospitals, London, UK. [80]Cancer Research UK & UCL Cancer Trials Centre, London, UK. [81]The Institute of Cancer Research, London, UK. [82]Garvan Institute of Medical Research, Sydney, New South Wales, Australia. [83]Case45, London, UK. [84]The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [85]Independent Cancer Patient's Voice, London, UK. [86]University Hospital Southampton NHS Foundation Trust, Southampton, UK. [87]The NIHR Southampton Biomedical Research Centre, University Hospital Southampton NHS Foundation Trust, Southampton, UK. [88]Department of Oncology, University Hospital Southampton NHS Foundation Trust, Southampton, UK. [89]Academic Division of Thoracic Surgery, Imperial College London, London, UK. [90]Royal Brompton and Harefield Hospitals, Part of Guy's and St Thomas' NHS Foundation Trust, London, UK. [91]National Heart and Lung Institute, Imperial College, London, UK. [92]Royal Surrey Hospital, Royal Surrey Hospitals NHS Foundation Trust, Guildford, UK. [93]University of Surrey, Guildford, UK. [94]University of Sheffield, Sheffield, UK. [95]Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield, UK. [96]Liverpool Heart and Chest Hospital, Liverpool, UK.

[97]Princess Alexandra Hospital, The Princess Alexandra Hospital NHS Trust, Harlow, UK. [98]School of Cancer Sciences, University of Glasgow, Glasgow, UK. [99]Beatson Institute for Cancer Research, University of Glasgow, Glasgow, UK. [100]Queen Elizabeth University Hospital, Glasgow, UK. [101]Institute of Infection, Immunity & Inflammation, University of Glasgow, Glasgow, UK. [102]NHS Greater Glasgow and Clyde, Glasgow, UK. [103]Cancer Research UK Scotland Institute, Glasgow, UK. [104]Institute of Cancer Sciences, University of Glasgow, Glasgow, UK. [105]NHS Greater Glasgow and Clyde Pathology Department, Queen Elizabeth University Hospital, Glasgow, UK. [106]Golden Jubilee National Hospital, Clydebank, UK.

Genomics England Consortium

J. C. Ambrose[107], P. Arumugam[107], E. L. Baple[107], M. Bleda[107], F. Boardman-Pretty[107,108], J. M. Boissiere[107], C. R. Boustred[107], H. Brittain[107], M. J. Caulfield[107,108], G. C. Chan[107], C. E. H. Craig[107], L. C. Daugherty[107], A. de Burca[107], A. Devereau[107], G. Elgar[107,108], R. E. Foulger[107], T. Fowler[107], P. Furió-Tarí[107], J. M. Hackett[107], D. Halai[107], A. Hamblin[107], S. Henderson[107,108], J. E. Holman[107], T. J. P. Hubbard[107], K. Ibáñez[107,108], R. Jackson[107], L. J. Jones[107,108], D. Kasperaviciute[107,108], M. Kayikci[107], L. Lahnstein[107], L. Lawson[107], S. E. A. Leigh[107], I. U. S. Leong[107], F. J. Lopez[107], F. Maleady-Crowe[107], J. Mason[107], E. M. McDonagh[107,108], L. Moutsianas[107,108], M. Mueller[107,108], N. Murugaesu[107], A. C. Need[107,108], C. A. Odhams[107], C. Patch[107,108], D. Perez-Gil[107], D. Polychronopoulos[107], J. Pullinger[107], T. Rahim[107], A. Rendon[107], P. Riesgo-Ferreiro[107], T. Rogers[107], M. Ryten[107], K. Savage[107], K. Sawant[107], R. H. Scott[107], A. Siddiq[107], A. Sieghart[107], D. Smedley[107,108], K. R. Smith[107,108], A. Sosinsky[107,108], W. Spooner[107], H. E. Stevens[107], A. Stuckey[107], R. Sultana[107], E. R. A. Thomas[107,108], S. R. Thompson[107], C. Tregidgo[107], A. Tucci[107,108], E. Walsh[107], S. A. Watters[107], M. J. Welland[107], E. Williams[107], K. Witkowska[107,108], S. M. Wood[107,108] & M. Zarowiecki[107]

[107]Genomics England, London, UK. [108]William Harvey Research Institute, Queen Mary University of London, London, UK.

Methods

Statistical information

All statistical tests were performed in R 4.0.2. No statistical methods were used to predetermine the sample size. Tests involving correlations were done using stat_cor from the R package ggpubr (v0.6.0) with Spearman's method, except for situations when we directly tested if there exists a linear relationship between two variables for which Pearson correlation was used. 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 size values for the corresponding Wilcoxon tests were measured using the wilcox_effsize function from the rstatix package (v0.7.2). HR values and *P* values were calculated with the survival package (v3.1-12) for both Kaplan–Meier curves and the CoxPH model. HR values between different models were compared by using a *z* test on the log of the HR values. 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 (v3.4.1), dplyr (v1.1.0), tidyr (v1.3.0), gridExtra (v2.3), tidyverse (1.3.2), gtable (v0.3.2), scales (v1.2.1), lubridate (v1.9.2), survminer (0.4.9), survcomp (1.40.0), RColorBrewer (v1.1.3), GGgally (v2.1.2), ggforce (v0.4.1), TCellExTRECT (v1.0.1), MatchIt (v4.5.0) and dNdScv (v0.0.1.0) packages. *P* value adjustments were made using either the Holm–Bonferroni method or the false discovery rate (FDR)/Benjamini–Hochberg method, with the FDR method being used for exploratory analysis and those involving many tests.

TRACERx 100

The TRACERx study (Clinicaltrials.gov registration: NCT01888601) is sponsored by University College London (UCL/12/0279) and has been approved by an independent Research Ethics Committee (13/LO/1546). All TRACERx samples used in this sample have been previously described[13], and obtaining informed consent from each patient was a mandatory requirement for participation in the TRACERx study. 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[13,48]. For the WES samples, exome capture was performed using a custom version of the Agilent Human All Exome V5 kit according to the manufacturer's instructions.

TCR-seq TRACERx100 data used in this analysis have been previously published[49]; FASTQ data are deposited at the Sequence Read Archive (SRA) under accession code BioProject (PRJNA544699).

TRACERx100 DNA samples were sequenced and aligned to GRCh38 by Genomics England using the same Illumina sequencing pipeline as used in the 100KGP to produce WGS samples to a mean depth of 175× (median 223×). WGS coverage values from the *TCRA, TCRB, TCRG* and *IGH* loci for these samples were then extracted for use for ImmuneLENS using SAMtools (v.1.3.1) depth. No other WGS-derived data besides these coverage values was used from the TRACERx cohort for this analysis.

100KGP WGS cohort

The 100KGP was ethically approved by the East of England–Cambridge South Research Ethics Committee (Research Ethics Committee reference 14/EE/1112, Integrated Research Application System ID 166046). Participants were recruited from 13 National Health Service (NHS) Genomic Medicine Centres, and all provided written informed consent.

All WGS samples within the 100KGP cohort were sequenced by Illumina and carried out on behalf of Genomics England. The Illumina pipeline for all samples used in this analysis performed alignment with the Issac aligner to the GRCh38 reference. Full details can be found at https://re-docs.genomicsengland.co.uk/genomic_data. Tumor samples were sequenced to a median depth of 97.5×, while germline blood samples were sequenced to a median depth of 32.7× in the pan-cancer cohort and 39.7× in the rare disease cohort.

T cells and B cells were calculated for the entire 100KGP cohort (lung—data release v8 (28 November 2019), remaining pan-cancer data release v12 (06 May 2021) and rare disease v12 (07 May 2021)). In total, scores were calculated for 92,905 WGS BAM files.

Of these, 31,675 BAM files were part of the 100KGP cancer cohort, representing 16,294 cancer BAM files and 15,381 germline BAM files. Some participants had multiple tumor samples collected for WGS; due to the lack of annotation of the reason for multiple samples (for example, technical resequencing, representative of metastasis, multiple region sequenced or occurrence of a second primary tumor at a later time point), these were removed from the pan-cancer cohort, leading to a final cohort of 14,501 tumor WGS samples. Of the 14,501 tumor WGS samples of our cohort, 13,868 have a matched blood WGS sample.

For the rare disease cohort, scores for 61,230 BAMs in total were calculated. Limiting to samples taken from blood samples leads to 59,903 BAMs representing 29,238 samples taken from probands with a rare disease and 30,665 relatives of these probands. This cohort of 30,665 blood samples from relatives was taken as our healthy cohort to compare with the 13,868 blood samples from patients with cancer. Additionally, from this healthy cohort, propensity-matched cohorts for each cancer histology were created using the R package matchit, controlling for both age and sex.

T cell and B cell fractions were calculated with the WGS version of ImmuneLENS; adjustments for tumor purity were made using estimates from Genomics England and local copy number using CANVAS (v.1.3.1) calls produced by Genomics England for nearby genes to the V(D)J loci (*TCRA*–*OR10G3*; *TCRB*–*PRSS58*; *TCRG*–*STARD3NL*; *IGH*–*TMEM121*).

Cancer histology in terms of disease and disease subtype was curated by Genomics England and is as described in the cancer_analysis_table available to researchers within the Genomics England research environment. To be consistent with other pan-cancer analyses, particularly TCGA, we used the histology groups designed by Genomics England to align as closely as possible to the TCGA (https://re-docs.genomicsengland.co.uk/cancer_analysis_histology/). We, however, kept the Childhood Cancers group from Genomics England's own annotation of cancer disease type due to the effect of age on our analysis. We also split up the breast invasive carcinoma group by hormone receptor status where available, and for the hematologically derived cancers, we split them up by cell of origin to distinguish B cell- and T cell-derived cancers from those from myeloid cells. All cancer types with occurrences less than 100 cases in the total cohort were assigned to the other cancer type groups, for both ease of analysis and to avoid the risk of any personally identifying features.

TCGA pan-cancer data

T cell ExTRECT was applied to the pan-cancer TCGA WES dataset. For different TCGA cohorts, different exome capture kits were used, and exon quality control was undertaken to ensure consistent results across the entire TCGA cohort. In brief, for each capture kit, the median GC-corrected read depth ratio was calculated for each exon using the exonsTcellExTRECT function across all samples. Exons with low coverage (median read depth ratio < −0.5) were filtered from the capture kit BED file.

TCGA sample ancestry calls are the consensus of five genetic ancestry calling approaches described in ref. 50.

1000 Genome cohort

In total, 2,544 samples with matched high- and low-coverage CRAM files, along with their indexed CRAI files, were downloaded directly from the 1000 Genomes cohort server (ftp://ftp.1000genomes.ebi.ac.uk) using wget. The median depth of the high-coverage samples was 34×, and the median depth for the low-coverage samples was 1.25×. ImmuneLENS with SAMtools (v.1.3.1) was then used on these CRAM files to extract the coverage and then calculate T cell and B cell fractions. Processed RNA-seq data from the Geuvadis project for 465 lymphoblastoid cell lines from the 1000 Genomes were downloaded from https://www.ebi.ac.uk/gxa/experiments/E-GEUV-1/Downloads with RNA-seq analysis performed using R packages limma and edgeR.

PCAWG

Our analysis was restricted to the TCGA portion of the PCAWG that contained 539 WGS tumor-normal pairs from BAM files that had been realigned to hg38 at MD Anderson. Germline normal samples had a median depth = 37.3× and tumor samples median depth = 51.2×. ImmuneLENS/SAMtools (v.1.3.1) was used to extract coverage files for the *IGH* and *TCRA* loci, which were then used in the calculation of all T cell and B cell fractions using the ImmuneLENS R package.

Low-pass TCGA data

In total, 317 low-pass TCGA BAM files with a median depth of 4.95× that were from samples with corresponding PCAWG high-coverage WGS were downloaded using the TCGA GDC client; coverage values for each sample were downloaded from the GDC client API.

scRNA cohort and analysis

Processed scRNA data with associated metadata from a lung cancer dataset described in ref. 51 were used. Annotation of B cell subtypes was used, and class switching was determined from the expression of *IGH* class switch segments with cells with unclear annotation removed from the analysis.

Nested downsampling of WGS files

Nested downsampling was performed on WGS BAM files using SAMtools view (v.1.3.1) recursively with the following options:

samtools view -b -h --subsample FRAC - -subsample-seed SEED BAM CHR_LOC > OUT_BAM

The depth of the original BAM files was calculated with mosdepth (v.0.3.2) and then downsampled to 60×. After each downsampling, each output BAM was indexed using Picard (v.2.20.3) BuildBamIndex before being downsampled again with SAMtools view to create a set of nested downsampled BAMs with depths of 60, 50, 40, 30, 20, 10, 5, 2, 1, 0.5 and 0.1×. The values for FRAC were calculated to obtain these depths based on the depth of the original BAM file as calculated with mosdepth. SEED values were changed in each nested down sample to avoid using the same seed repeatedly.

The above-given procedure was done to generate downsampled BAMs for all of the TRACERx WGS samples for the *TCRA* (chr14: 21621904–22752132), *TCRB* (chr7: 142299011–142813287), *TCRG* (chr7: 38240024–38368055) and *IGH* (chr14: 105566277–106879844) loci. These downsampled BAMs were then used as input for the WGS version of ImmuneLENS to calculate the T cell and B cell fractions.

Differential gene expression analysis and gene set enrichment

We performed differential gene expression on TRACERx100 patients with RNA-seq data, separating the cohort into high or low groups based on either *IGH* B cell fraction or class-switching B cell fraction scores for *IGHG1, IGHA1* B cell fraction or nonclass-switched IgM/IgD B cell fraction. First, using R 4.0.0, the edgeR package (v.3.32.1) was used for the sample-specific trimmed mean of the *M* values 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 (v.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 visualized with the R EnhancedVolcano package (v.1.8.0). Gene set enrichment analysis was then performed using the fgsea R package, which uses (v.1.24.0) the MSigDB C8 genesets of cell type signatures (https://www.gsea-msigdb.org/gsea/msigdb/human/genesets.jsp?collection=C8). An adaptive multi-level split Monte Carlo scheme for *P* value estimation was used and full results are given in Supplementary Data.

Calculation of TCR/BCR diversity metrics

T cell diversity metrics from V or J segment usage were predicted by the ImmuneLENS model. For the Shannon diversity, we used the formula:

$$\text{Shannon diversity} = -\sum_i p_i ln(p_i),$$

where p_i represents the proportion of an individual V or J segment predicted from the model.

To compare between two samples with different predicted segment usage, we used the Jensen–Shannon divergence metric defined as follows:

$$P = \text{probability distribution of V or J segment usage in sample A.}$$

$$Q = \text{probability distribution of V or J segment usage in sample B.}$$

$$M = \frac{P+Q}{2}.$$

$D(P|Q) = \sum_i P_i \log(\frac{P_i}{Q_i})$, where P_i and Q_i represent the proportion of segment *i* used in sample A or B, respectively.

The Jensen–Shannon divergence (JSD) is then defined as follows:

$$\text{JSD}(P|Q) = \frac{1}{2}D(P|M) + \frac{1}{2}D(Q|M).$$

MiXCR[22] was used to call TCR clonotypes from TRACERx 100 RNA-seq data directly. From these calls, the Shannon entropy was then calculated using the proportions of the TCR clonotypes found in each sample.

TRUST4 (ref. 52) was also used to call BCR sequences and therefore infer class-switching proportions.

TRAV segment usage analysis

TRAV segment predictions from ImmuneLENS were generated for the entire 100KGP cohort. For analysis of differing TRAV segment usage, we restricted to samples with >0.05 total *TCRA* T cell fraction. To test for TRAV segment usage changes between different populations, we used propensity matching to control for *TCRA* T cell fraction and to ensure that the distribution of T cell fraction was the same between comparison populations and hence there would be no bias with higher T cell fraction cohorts associated with more diversity due to our increased power to detect TRAV segment usage. Once the comparison cohorts were created, we defined a TRAV segment as being selected within the ImmuneLENS model if it had a fraction >0.001 of T cells predicted to use the TRAV segment. For every segment, we then calculated the percentage of samples within the cohort that selected that segment >0.001 fraction. Differences in segment usage between cohorts were then assessed using a χ^2 test, and then *P* values were adjusted for multiple hypothesis testing.

100KGP genetic ancestry inference

Genetic ancestry inference was provided by Genomics England for the entire 100KGP cohort using ethnicities from the 1000 Genomes Project phase 3 as truth by first generating principal components and then projecting the 100KGP project onto them to identify the broad genetic ancestry super-category of each participant. Full details can be found at https://re-docs.genomicsengland.co.uk/ancestry_inference/.

100KGP date-matched blood count data

Blood count data were only available for a subset of the rare disease cohort within the 100KGP. We selected blood count data that were time-matched for the exact date of genomic sample collection resulting in data from 441 participants for which we had date-matched blood count and calculated T cell or B cell fractions. From this data, we further subsetted for participants with matched albumin count ($n = 361$), lymphocyte count ($n = 222$), neutrophil count ($n = 84$) and both neutrophil and lymphocyte count data ($n = 84$).

100KGP treatment data

Treatment data were extracted from 100KGP using the clinical data available in the Genomics England research environment from the 'cancer_systemic_anti_cancer_therapy' version 13 table.

dNdScv analysis of selection in protein-coding genes

We used dNdScv[31] to measure the expected number of nonsynonymous mutations for genes associated with cancer within the cancer gene census. We then adapted code from ref. 53, which used a Poisson model of observed nonsynonymous mutations using the neutral background expectation as calculated within dNdScv as an offset variable influenced by age or sex in normal bladder tissue. In our analysis, we also added tumor purity, tumor mutation burden and disease type as controlling variables. We added either infiltrating *TCRA* T cell fraction, *IGH* B cell fraction, total class-switched B cell fraction, total non-class-switched B cell fraction (IgM/IgD), IgA or IgG B cell fraction into the Poisson model and identified genes for which these variables were significant (model: observed mutations ~ offset(log(expected mutations)) + age + tumor mutation burden + sex + disease type + immune fraction). For the pan-cancer analysis, we selected significant genes after adjusting for multiple hypothesis testing, for the disease-type-specific models, we only tested genes that had ≥10 tumors containing nonsynonymous mutations in that disease type and did not include sex in the model for cancer types predominant in one sex.

Analysis of known SNPs associated with leukocyte traits using PLINK

In total, 1,962 SNPs known to be associated with leukocyte traits (either basophil, eosinophil, lymphocyte, monocyte, neutrophil counts or white blood count) were downloaded from the data released by ref. 27. Of these, only 1,635 SNPs were listed within the VCF files provided within the 100KGP above the 0.001 mean allele frequency threshold. PLINK was used to test for associations between these SNPs and *TCRA* T cell fraction, *IGH* B cell fraction, IgM/IgD B cell fraction, IgG B cell fraction, IgA B cell fraction or the T/B cell ratio. All T cell and B cell fractions were first transformed using the inverse normal transformation to ensure normality. Association tests were run separately in the different genetic ancestry groups (as defined by the 1000 Genomes Project). Our analysis focused on participants with genetically inferred European ancestry, as this was the largest ancestry group in both the healthy and pan-cancer cohorts and those with genetically inferred African ancestry, as this group showed the most significant difference in circulating T cell fraction compared to the European ancestry group. For the healthy and pan-cancer cohorts age, sex and the first ten genetic ancestry principal components (PCs) were used as covariates. PLINK (v1.9) was run separately on each cancer subtype using these covariates. PLINK was run with the following steps: (1) LD pruning of the tested SNPs using the option –indep-pairwise 500 5 0.5 to test 500 SNPs at a time, moving the window by five SNPs at each step and using an R^2 threshold of 0.5 to remove any high LD SNPs. (2) A cutoff on genotype frequencies using –geno 0.2 –maf 0.01 to remove any SNPs with a missing genotype rate >20% and a minor allele frequency >1%. (3) A Hardy–Weinberg filter using –hwe 0.000001 to specify the *P* value threshold. (4) A test for association of the phenotype using linear regression and the –linear option, as well as calculation of genotype frequencies using –freq. For the pan-cancer analysis, the output of each cohort from PLINK was then combined in a common effect meta-analysis using the metagen function from the R package meta, with the input being the treatment effect and its s.e. from each of the separate cancer histology runs of PLINK.

Survival analysis

Survival data were collated on the Genomics England research environment using available data for date of cancer diagnosis, death records from the Office of National Statistics and the latest follow-up times from the most recent records in the hospital episode statistics (release v16). The data then underwent additional quality control to exclude any patients with any conflicting data, such as follow-up times greater than 10 years resulting from multiple dates of diagnosis values from previous incidences of cancer. In total, 13,348 participants with both survival data and ImmuneLENS fractions in blood and 13,342 participants with both survival data and ImmuneLENS fractions in tumor tissue were available for analysis.

A meta-analysis using a random effects model was applied to the output of the HRs of the disease-specific CoxPH models using the R function metagen from the package meta (v.6.5-0). Thus, the I^2 values were calculated, and their significance was tested using Cochran's *Q* test.

Figure generation

Biorender.com aided in the generation of Fig. 1a, Extended Data Fig. 4a and Supplementary Fig. 3a.

Reporting summary

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

Data availability

The RNA-seq and WES data (in each case from the TRACERx study) used during this study are a subset of the TRACERx421 dataset and have been deposited at the European Genome–Phenome Archive, which is hosted by The European Bioinformatics Institute and the Centre for Genomic Regulation under the accession codes EGAS00001006517 (RNA-seq) and EGAS00001006494 (WES). Access is controlled by the TRACERx data access committee to ensure patient privacy and data confidentiality are protected while fostering impactful scientific discoveries. Details on how to apply for access are available on the linked pages. The data access committee aims to respond to requests within 1 week. For TCR-seq data used in this analysis, the FASTQ data are deposited at the SRA under accession code BioProject (PRJNA544699). Coverage files for the *TCRA*, *TCRB*, *TCRG* and *IGH* loci were generated from WGS TRACERx 100 samples. These coverage files used for the calculation of the T cell and B cell fractions are available at Zenodo (https://doi.org/10.5281/zenodo.7785803)[54] and were the only data derived from the TRACERx WGS analysis used within this paper.

WGS and phenotypic data from the 100KGP can be accessed by application to Genomics England following the procedure outlined at https://www.genomicsengland.co.uk/join-us for both academic and industry users. Genomics England restricts access to 100KGP data to bona fide researchers to protect the sensitive genomic data of its participants. For academic users, Genomics England aims to review all applications within ten working days, and access will be granted within two working days after confirmation of affiliation from the researcher's institution and completion of online governance training.

The 1000 Genome Data used are publicly available and can be accessed at https://www.internationalgenome.org/data. Calculated ImmuneLENS output including class-switching and polyclonal predictions for each LCL cell line included are available on Zenodo (https://doi.org/10.5281/zenodo.11093976)[55].

PCAWG data used in this study were obtained through our collaboration with MD Anderson. To gain access to the raw WGS samples of the TCGA portion of the PCAWG data used in this study, researchers need to apply to the TCGA data access committee via a database of Genotypes and Phenotypes (dbGaP; https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=login). Access is controlled due to respect and protection of the interests of research participants. The calculated ImmuneLENS output for these samples is available on Zenodo (https://doi.org/10.5281/zenodo.11093961)[56].

TCGA pilot project was established by the National Cancer Institute (NCI) and the National Human Genome Research Institute. The data were retrieved through the dbGaP authorization (accession 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/. To access TCGA WES and low-pass WGS data, researchers will need to apply to the TCGA Data Access Committee (DAC) via dbGaP (https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=login). Access is controlled due to respect and protection of the interests of research participants. The calculated T cell ExTRECT *TCRA* T cell fraction scores along with the ImmuneLENS output for the low-pass WGS samples used in this study are available at Zenodo (https://doi.org/10.5281/zenodo.7794867)[57].

Single-cell data used in this analysis are previously described[51] and available at https://cellxgene.cziscience.com/collections/edb893ee-4066-4128-9aec-5eb2b03f8287.

Code availability

The code used to produce T cell and B cell fraction scores will be made available for academic noncommercial research purposes as the R package ImmuneLENS, available for download and installation at https://github.com/McGranahanLab/ImmuneLENS.

All other code used in the analysis and necessary data to reproduce figures is available at Zenodo[58] (https://doi.org/10.5281/zenodo.14046632). 100KGP data cannot be exported outside the Genomics England research environment. All data and code to reproduce the 100KGP analysis, including T cell and B cell fractions, are available within the Genomics England research environment (see https://re-docs.genomicsengland.co.uk/access/ for information on using the research environment) within the folder '/re_gecip/shared_allGeCIPs/rbentham/ImmuneLENS_figure_code/'. Researchers can gain access to the Genomics England research environment and associated data following application to Genomics England following the procedure outlined at https://www.genomicsengland.co.uk/join-us.

References

48. Martínez-Ruiz, C. et al. Genomic-transcriptomic evolution in lung cancer and metastasis. *Nature* **616**, 543–552 (2023).
49. Joshi, K. et al. Spatial heterogeneity of the T cell receptor repertoire reflects the mutational landscape in lung cancer. *Nat. Med.* **25**, 1549–1559 (2019).
50. Carrot-Zhang, J. et al. Comprehensive analysis of genetic ancestry and its molecular correlates in cancer. *Cancer Cell* **37**, 639–654 (2020).
51. Salcher, S. et al. High-resolution single-cell atlas reveals diversity and plasticity of tissue-resident neutrophils in non-small cell lung cancer. *Cancer Cell* **40**, 1503–1520 (2022).
52. Song, L. et al. TRUST4: immune repertoire reconstruction from bulk and single-cell RNA-seq data. *Nat. Methods* **18**, 627–630 (2021).
53. Lawson, A. R. J. et al. Extensive heterogeneity in somatic mutation and selection in the human bladder. *Science* **370**, 75–82 (2020).
54. Bentham, R. Coverage files of TRACERx100 samples for input for ImmuneLENS (0.1). *Zenodo* https://doi.org/10.5281/zenodo.7785803 (2023).
55. Bentham, R. ImmuneLENS output for 1000 genomes cohort. *Zenodo* https://doi.org/10.5281/zenodo.11093976 (2024).
56. Bentham, R. ImmuneLENS output for PCAWG (TCGA subset). *Zenodo* https://doi.org/10.5281/zenodo.11093961 (2024).
57. Bentham, R. & Jones, T. T cell ExTRECT scores for TCGA. *Zenodo* https://doi.org/10.5281/zenodo.7794867 (2023).
58. Bentham, R. ImmuneLENS manuscript figure code. *Zenodo* https://doi.org/10.5281/zenodo.14046632 (2024).
59. Travaglini, K. J. et al. A molecular cell atlas of the human lung from single-cell RNA sequencing. *Nature* **587**, 619–625 (2020).

Acknowledgements

TRACERx is funded by Cancer Research UK (C11496/A17786) and coordinated through the Cancer Research UK and UCL Cancer Trials Centre, which has a core grant from Cancer Research UK (CRUK) (C444/A15953). We gratefully acknowledge the patients and relatives who participated in the TRACERx study. We thank all site personnel, investigators, funders and industry partners who supported the generation of the data within this study.

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).

This research was made possible through access to data in the National Genomic Research Library, which is managed by Genomics England (a wholly owned company of the Department of Health and Social Care). The National Genomic Research Library holds data provided by patients and collected by the NHS as part of their care and data collected as part of their participation in research. The National Genomic Research Library is funded by the National Institute for Health Research and NHS England. The Wellcome Trust, Cancer Research UK and the Medical Research Council have also funded research infrastructure.

R.B. is supported by the National Institute of Health and Care Research (NIHR) Biomedical Research Centre (BRC) at University College London Hospitals. T.P.J. is supported by the Cancer Research UK City of London Centre. J.R.M.B. is supported by Cancer Research UK. C.P. is supported by the Francis Crick Institute. C.M.-R. is supported by the Rosetrees (M630) and Wellcome trusts. M.D. is supported by CRUK and the Lung Cancer Centre of Excellence. M.L. is supported by the Cancer Research UK City of London Centre. K.T. is supported by the Radiation Research Unit at the Cancer Research UK City of London Centre Award (C7893/A28990). T.B.K.W. is supported by the Francis Crick Institute and by the Marie Curie ITN Project PLOIDYNET (FP7-PEOPLE-2013, 607722), the Breast Cancer Research Foundation (BCRF), a Royal Society Research Professorships Enhancement Award (RP/EA/180007) and the Foulkes Foundation. O.P. is supported by funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie Grant (101068227). C.S. is a Royal Society Napier Research Professor (RSRP\R\210001). His work is supported by the Francis Crick Institute, which receives its core funding from Cancer Research UK (CC2041), the UK Medical Research Council (CC2041) and the Wellcome Trust (CC2041). For Open Access, the author has applied a CC BY public copyright license to any author-accepted manuscript version arising from this submission. C.S. is funded by Cancer Research UK (TRACERx (C11496/A17786), PEACE (C416/A21999) and CRUK Cancer Immunotherapy Catalyst Network); Cancer Research UK Lung Cancer Centre of Excellence (C11496/A30025); the Rosetrees Trust, Butterfield and Stoneygate Trusts; Novo Nordisk Foundation (ID16584); Royal Society Professorship Enhancement Award (RP/EA/180007 and RF\ERE\231118); NIHR University College London Hospitals Biomedical Research Centre; the Cancer Research UK-University College London Centre; Experimental Cancer Medicine Centre; the BCRF (US; BCRF-22-157); Cancer Research UK Early Detection and Diagnosis Primer Award (grant EDDPMA-Nov21/100034); the Mark Foundation for

Cancer Research Aspire Award (grant 21-029-ASP) and ASPIRE Phase II Award (grant 23-034-ASP). C.S. is in receipt of an European Research Council (ERC) Advanced Grant (PROTEUS) from the European Research Council under the European Union's Horizon 2020 research and innovation program (grant 835297). P.V.L. is a CPRIT Scholar in Cancer Research and acknowledges CPRIT grant support (RR210006). N.M. was a Sir Henry Dale Fellow, jointly funded by the Wellcome Trust and the Royal Society (211179/Z/18/Z), and also receives funding from a CRUK Programme Foundation Award (DRCPFA-Nov23/100003) and a CRUK Biomarker Project Award (CRCBPA-Mar24/100001), the Rosetrees Trust, the NIHR BRC at University College London Hospitals and the CRUK University College London Experimental Cancer Medicine Centre.

Author contributions

R.B. helped conceive the study, designed and conducted the bioinformatic analysis, built and maintained the R packages T cell ExTRECT and ImmuneLENS used in this study and wrote the manuscript. T.P.J. set up and ran T cell ExTRECT on the WES samples within the TCGA cohort and assisted with the TCGA analysis. J.R.M.B. provided expertise and assisted with the analysis of the blood count data within the 100KGP cohort. C.M.R. assisted and provided expertise with the RNA-seq analysis on the TRACERx cohort. M.D., M.L. and K.T. all assisted with the processing and analysis of WGS and clinical data for the 100KGP cohort. K.T. helped analyze and collate the treatment-associated data of 100KGP patients with cancer. T.B.K.W. helped conceive the original study and its application to B cells. C.B. and O.P. assisted with the processing of the TRACERx WGS samples, and O.P. ran MiXCR and TRUST4 on the TRACERx data and provided assistance on the analysis of the TCR-seq TRACERx data. P.V.L. provided access to the WGS PCAWG WGS cohort, and Z.Z. ran ImmuneLENS on these WGS samples. T.P.J., J.R.M.B., C.M.R., M.L., K.T., T.B.K.W., C.B., P.V.L. and O.P. all provided feedback on the manuscript. C.S. helped provide study supervision, 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

N.M. and R.B. hold a European patent for determination of B cell fraction in mixed samples (PCT/EP2024/062999). N.M., R.B., T.B.K.W. and C.S. hold a European patent for determination of lymphocyte abundance in mixed samples (PCT/EP2022/070694). N.M. has stock options in and has consulted for Achilles Therapeutics and holds a European patent relating to targeting neoantigens (PCT/EP2016/059401), identifying patient response to immune checkpoint blockade (PCT/EP2016/071471), determining HLA LOH (PCT/GB2018/052004) and predicting survival rates of patients with cancer (PCT/GB2020/050221). C.S. acknowledges grant support from AstraZeneca, Boehringer-Ingelheim, Bristol Myers Squibb, Pfizer, Roche-Ventana, Invitae (previously Archer Dx Inc - collaboration in minimal residual disease sequencing technologies), Ono Pharmaceutical and Personalis. He is chief investigator for the AZ MeRmaiD 1 and 2 clinical trials and is the steering committee chair. He is also co-chief investigator of the NHS Galleri trial funded by GRAIL and a paid member of GRAIL's scientific advisory board (SAB). He receives consultant fees from Achilles Therapeutics (also SAB member), Bicycle Therapeutics (SAB member, and chair of clinical advisory group), Genentech, Medicxi, China Innovation Centre of Roche (CICoR) (formerly Roche Innovation Centre – Shanghai, Metabomed (until July 2022)), Relay Therapeutics (SAB member), Saga Diagnostics (SAB member), and the Sarah Cannon Research Institute. C.S. has received honoraria from Amgen, AstraZeneca, Bristol Myers Squibb, GlaxoSmithKline, Illumina, MSD, Novartis, Pfizer and Roche-Ventana. C.S. has previously held stock/options in GRAIL, and currently has stock/options Bicycle Therapeutics, Relay Therapeutics, and has stock and is co-founder of Achilles Therapeutics. C.S. declares a patent application for methods to lung cancer (PCT/US2017/028013); targeting neoantigens (PCT/EP2016/059401); identifying patent response to immune checkpoint blockade (PCT/EP2016/071471); methods for lung cancer detection (US20190106751A1); identifying patients who respond to cancer treatment (PCT/GB2018/051912); determining HLA LOH (PCT/GB2018/052004); predicting survival rates of patients with cancer (PCT/GB2020/050221), methods and systems for tumour monitoring (PCT/EP2022/077987), Analysis of HLA alleles transcriptional deregulation (PCT/EP2023/059039). C.S. is an inventor on a European patent application (PCT/GB2017/053289) relating to assay technology to detect tumour recurrence. This patent has been licensed to a commercial entity and under their terms of employment C.S is due a revenue share of any revenue generated from such license(s). The remaining authors declare no competing interests.

Additional information

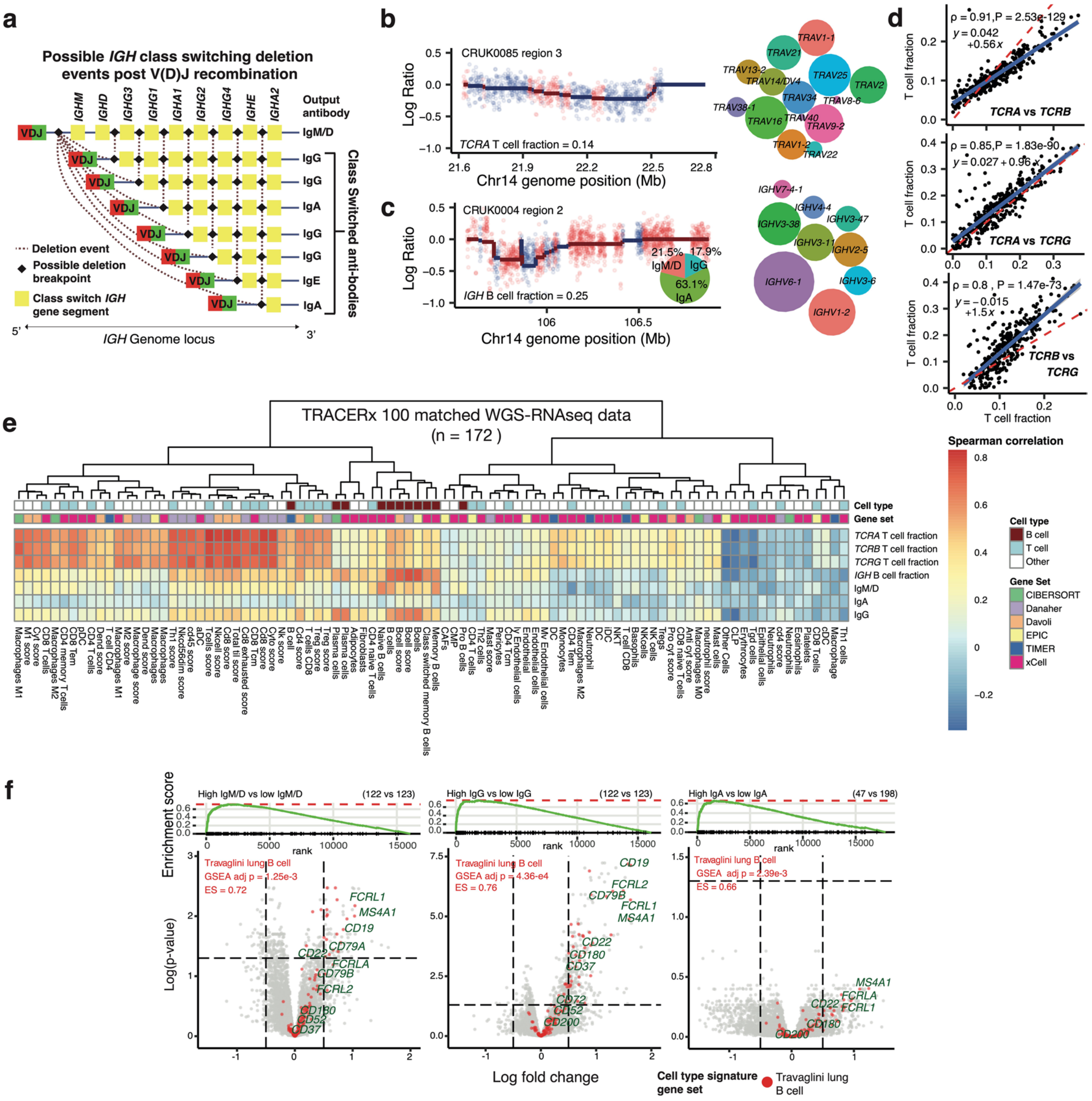
Extended data is available for this paper at https://doi.org/10.1038/s41588-025-02086-5.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-025-02086-5.

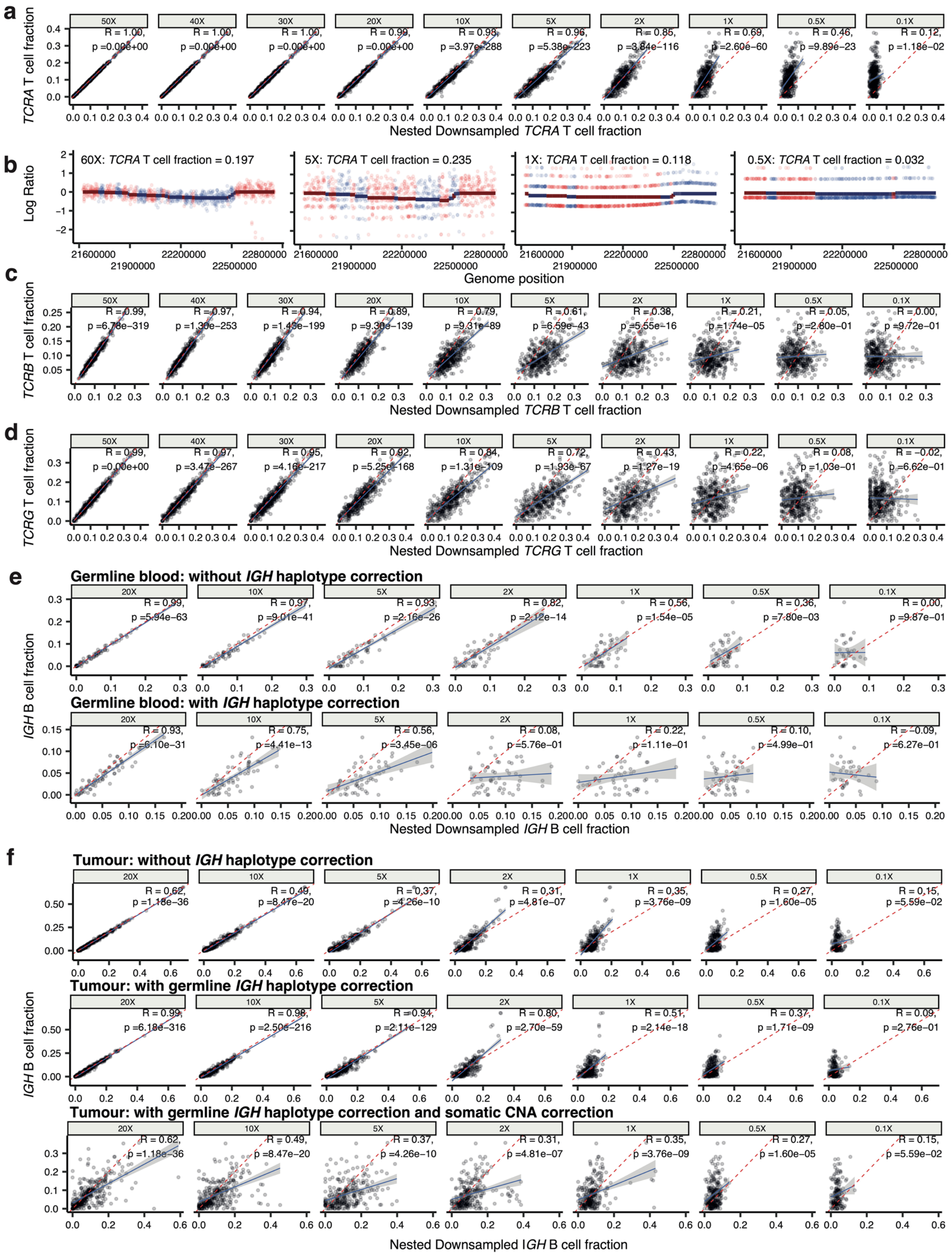
Correspondence and requests for materials should be addressed to Nicholas McGranahan.

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

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

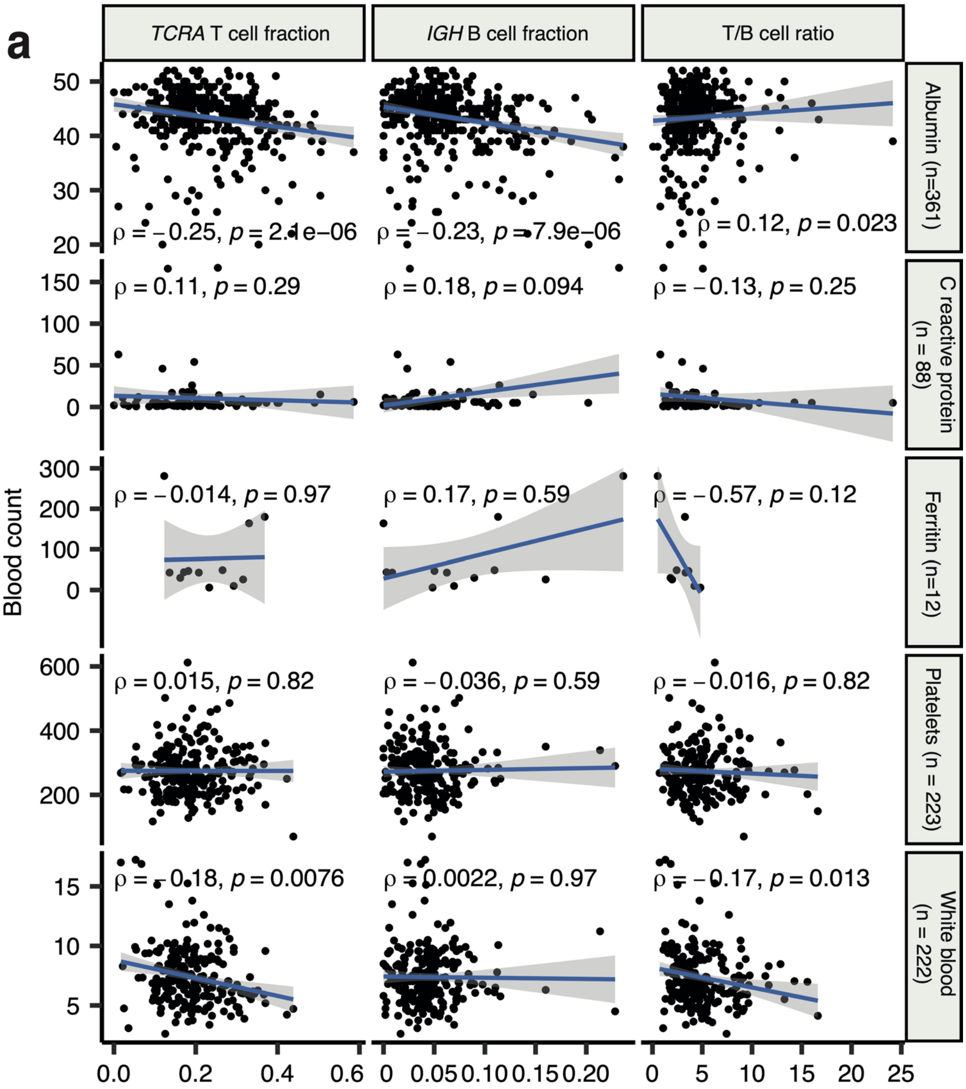


Extended Data Fig. 1 | Validation and description of ImmuneLENS. a, Diagram illustrating possible class-switching deletion events following VDJ recombination at the *IGH* locus, resulting in B cells producing different antibodies. **b**, Example ImmuneLENS output showing the *TCRA* locus for TRACERx sample CRUK0085 region 3. The log read depth ratio plot corresponds to a predicted T cell fraction of 0.14. Alternating colors represent changes in the 14 TRAV segments selected by the model, which are also depicted in the bubble plot (right). **c**, Example ImmuneLENS output showing the *IGH* locus for TRACERx sample CRUK0004 region 2, with a predicted B cell fraction of 0.25. IGHV segment usage and class-switching percentages are also displayed (right). **d**, Scatter plots showing the correlation between T cell fraction values calculated by ImmuneLENS from the *TCRA*, *TCRB* or *TCRG* loci. The blue line represents the line of best fit, and the gray region indicates the 95% confidence interval. **e**, Heatmap of ImmuneLENS' fractions compared to RNA-seq signatures for different cell types. **f**, Differential gene expression analysis (bottom) of TRACERx RNA-seq samples split into high and low groups based on median predicted non-class-switched IgM/IgD B cell fractions and class-switched IgA and IgG B cell fractions. Analysis and significance were assessed using limma–voom (Methods), accounting for multiple hypothesis testing. Red points represent genes within the Travaglini lung B cell gene signature[59]. P values were derived from a GSEA analysis (top) of all cell type signature genesets defined by MSigDB, with P-value estimation based on an adaptive multi-level split Monte Carlo scheme. The P values for Spearman's ρ in **d** were derived from a two-tailed t-distribution using the correlation coefficient and sample size.

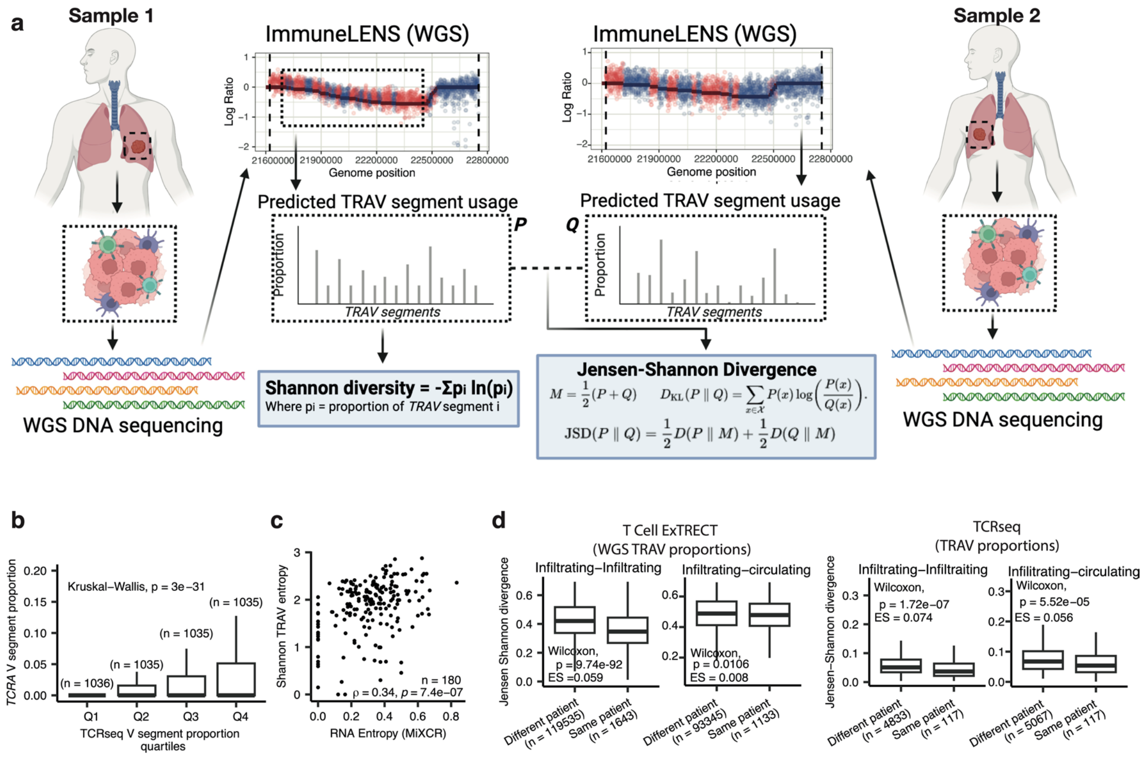


Extended Data Fig. 2 | Nested downsampling of TRACERx data. a, Correlation of T cell fraction at 60× coverage with samples downsampled to different coverage levels using nested downsampling. **b**, Example ImmuneLENS model outputs of the same sample at different downsampled coverage depths. **c**,**d**. Correlation of *TCRB* and *TCRG* T cell fractions at 60× coverage with samples downsampled to different coverage levels using nested downsampling. **e**, Correlation of circulating B cell fraction at 30× coverage with samples downsampled to different coverage levels using nested downsampling. Top: B cell fraction calculated without *IGH* haplotype correction. Bottom: B cell fraction calculated with *IGH* haplotype correction **f**, Correlation of infiltrating B cell fraction at 30× coverage with samples downsampled to different coverage levels using nested downsampling. Top: B cell fraction calculated without *IGH* haplotype correction. Middle: B cell fraction calculated with germline *IGH* haplotype correction. Bottom: B cell fraction calculated with both germline and somatic *IGH* haplotype correction. Throughout, blue lines represent the line of best fit, and gray regions indicate the 95% confidence interval. The P values for Pearson's R were derived from a two-tailed t-distribution using the correlation coefficient and sample size.

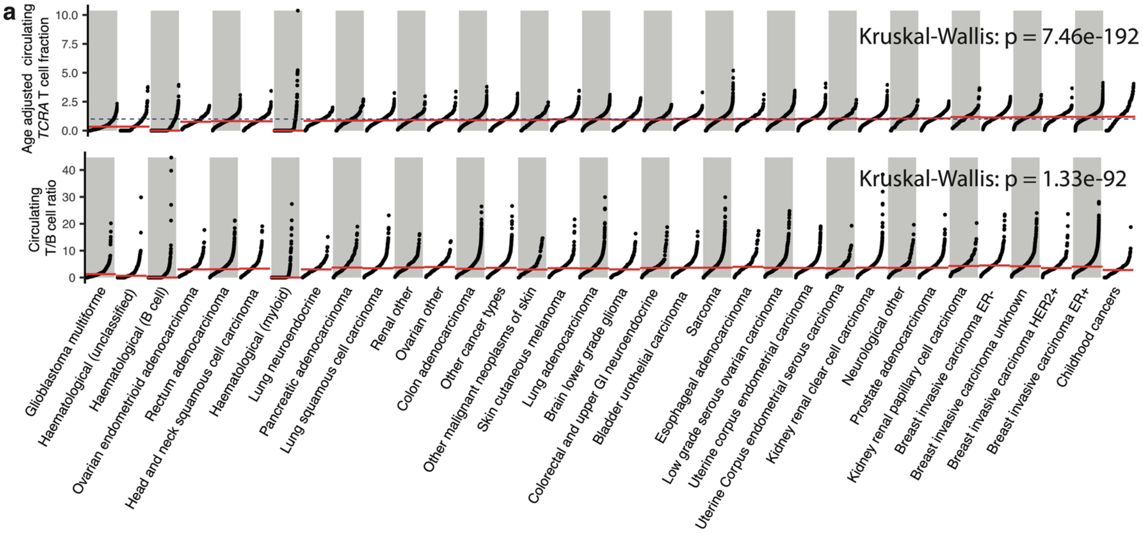
https://doi.org/10.1038/s41588-025-02086-5



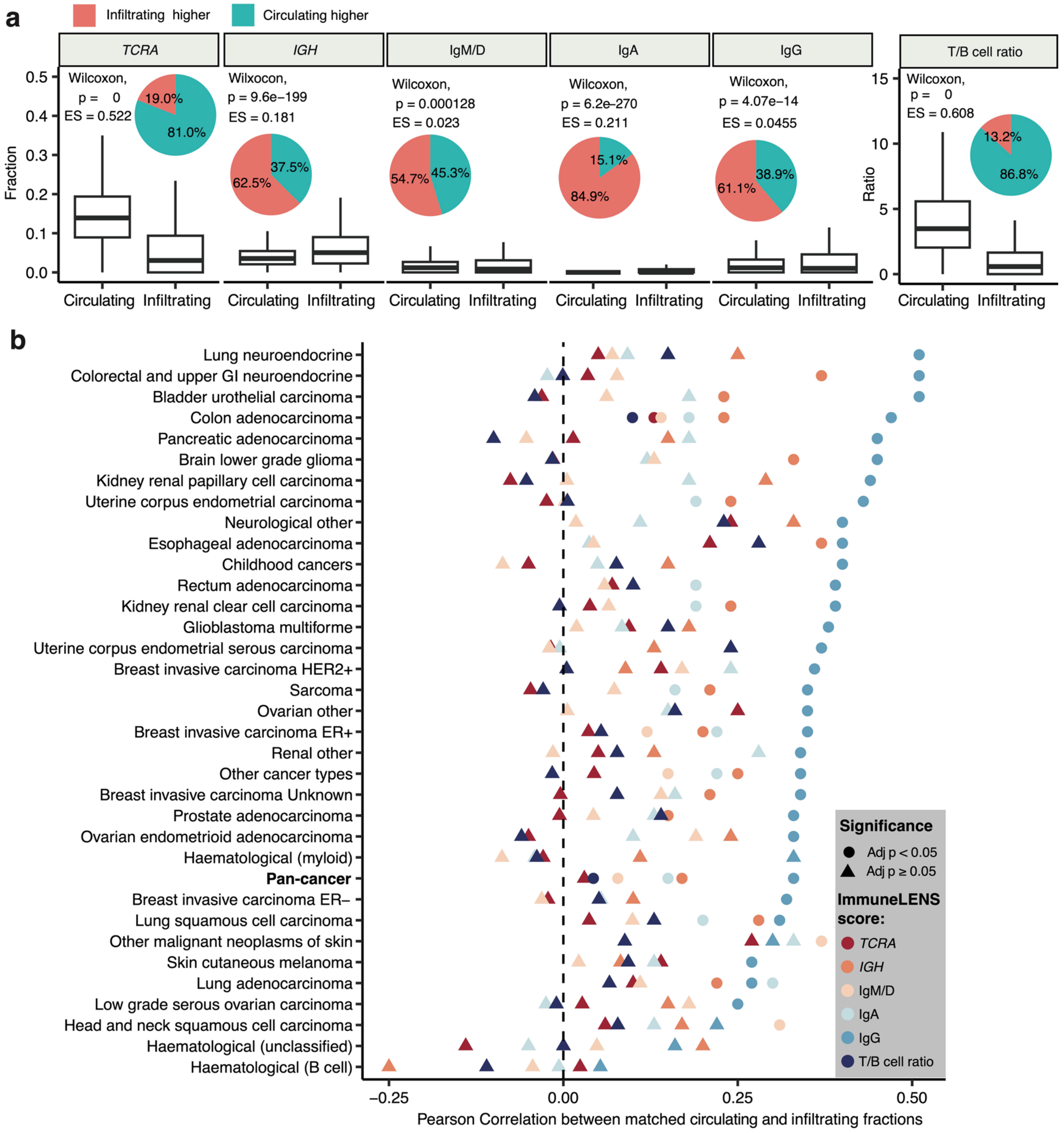
Extended Data Fig. 3 | Association of ImmuneLENS scores with blood count data. a. ImmuneLENS fractions versus date-matched blood count data for albumin, C-reactive protein, ferritin, platelets and white blood count. Blue lines represent the line of best fit with gray regions representing 95% confidence interval. P values for Spearman's ρ were derived from a two-tailed t-distribution using the correlation coefficient and sample size.



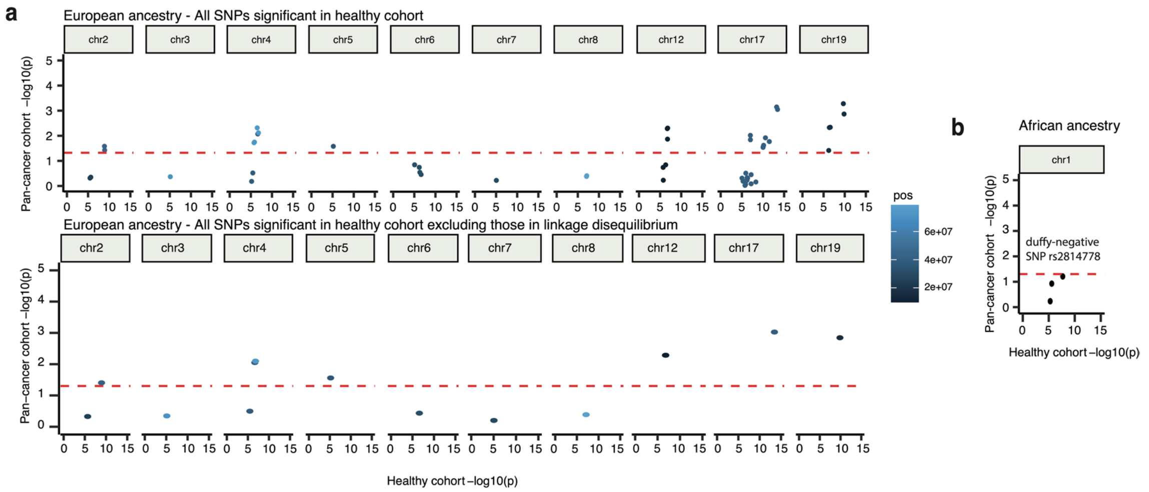
Extended Data Fig. 4 | Validation of TCR diversity metrics. a, Cartoon overview of the use of ImmuneLENS to calculate Shannon diversity values from TRAV segment usage as well as divergence metrics between two samples using the Jensen–Shannon divergence (JSD) from TRAV segment usage. The figure is created with BioRender.com. **b**, Proportion of TRAV segment usage in samples as measured from either TRACERx TCR-seq data (separated into 4 quartiles) or predicted by ImmuneLENS. **c**, Correlation of Shannon entropy measurements as measured by either TRAV segments predicted by ImmuneLENS or from MiXCR (RNA-seq). **d**, Jensen–Shannon divergence of samples either from different or same TRACERx patients, for tumor–tumor or tumor–blood sample comparisons measured by ImmuneLENS or TCR-seq from TRAV segment proportions, with significance assessed using a two-sided Wilcoxon rank-sum test. Boxplots in **b** and **d** show the median, lower and upper quartile and with whiskers extending to 1.5× the interquartile range above and below the interquartile range. P values for Spearman's ρ were derived from a two-tailed t-distribution using the correlation coefficient and sample size.



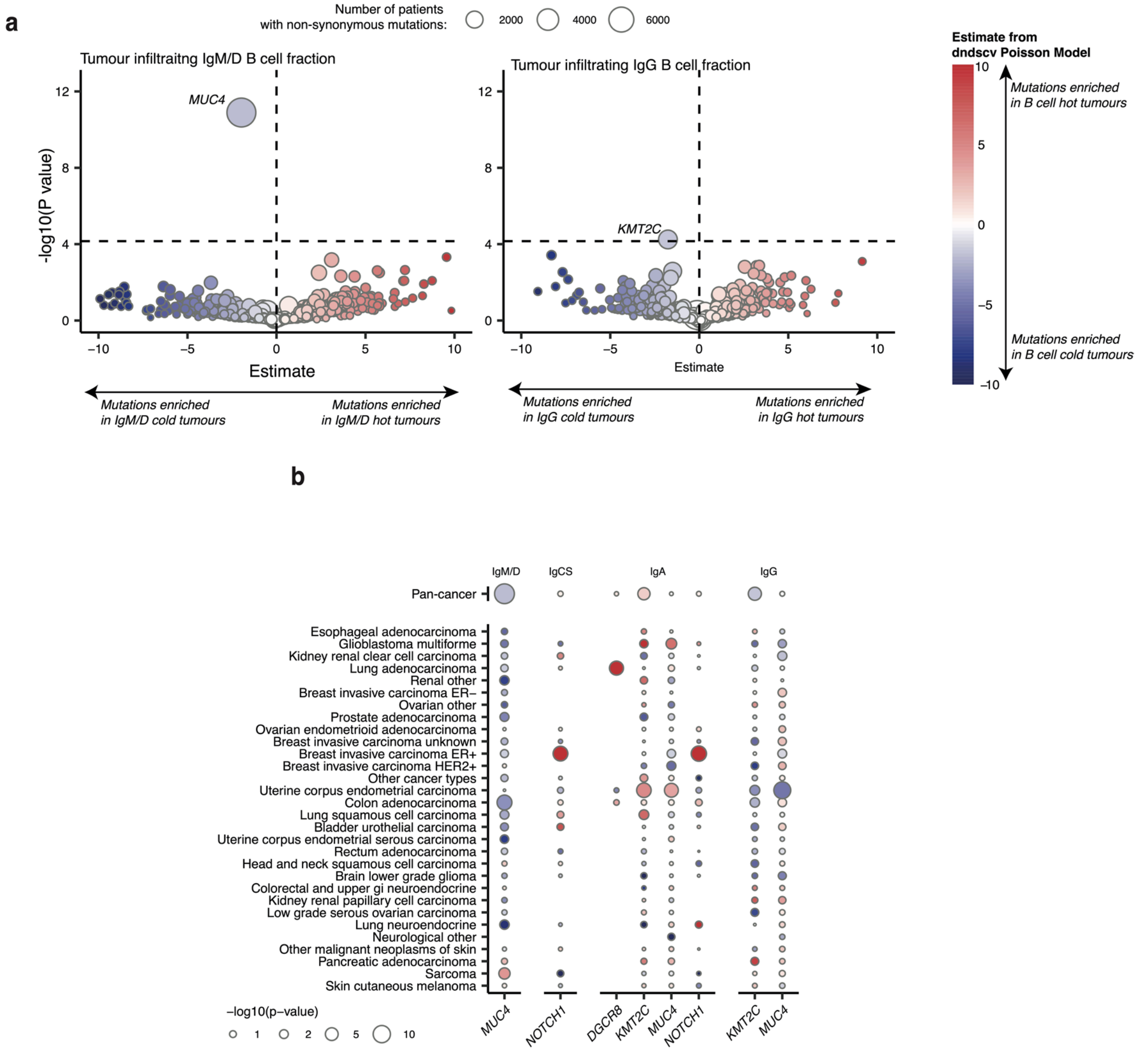
Extended Data Fig. 5 | Pan-cancer overview of IGH and T/B cell ratio. a. Overview snake plot of age-adjusted *TCRA* T cell fraction and circulating T/B cell ratio, with horizontal red line representing the median value per histology and dashed black line at y = 1 for the age-adjusted circulating *TCRA* T cell fraction to show the median value expected for the age distribution of that cohort.



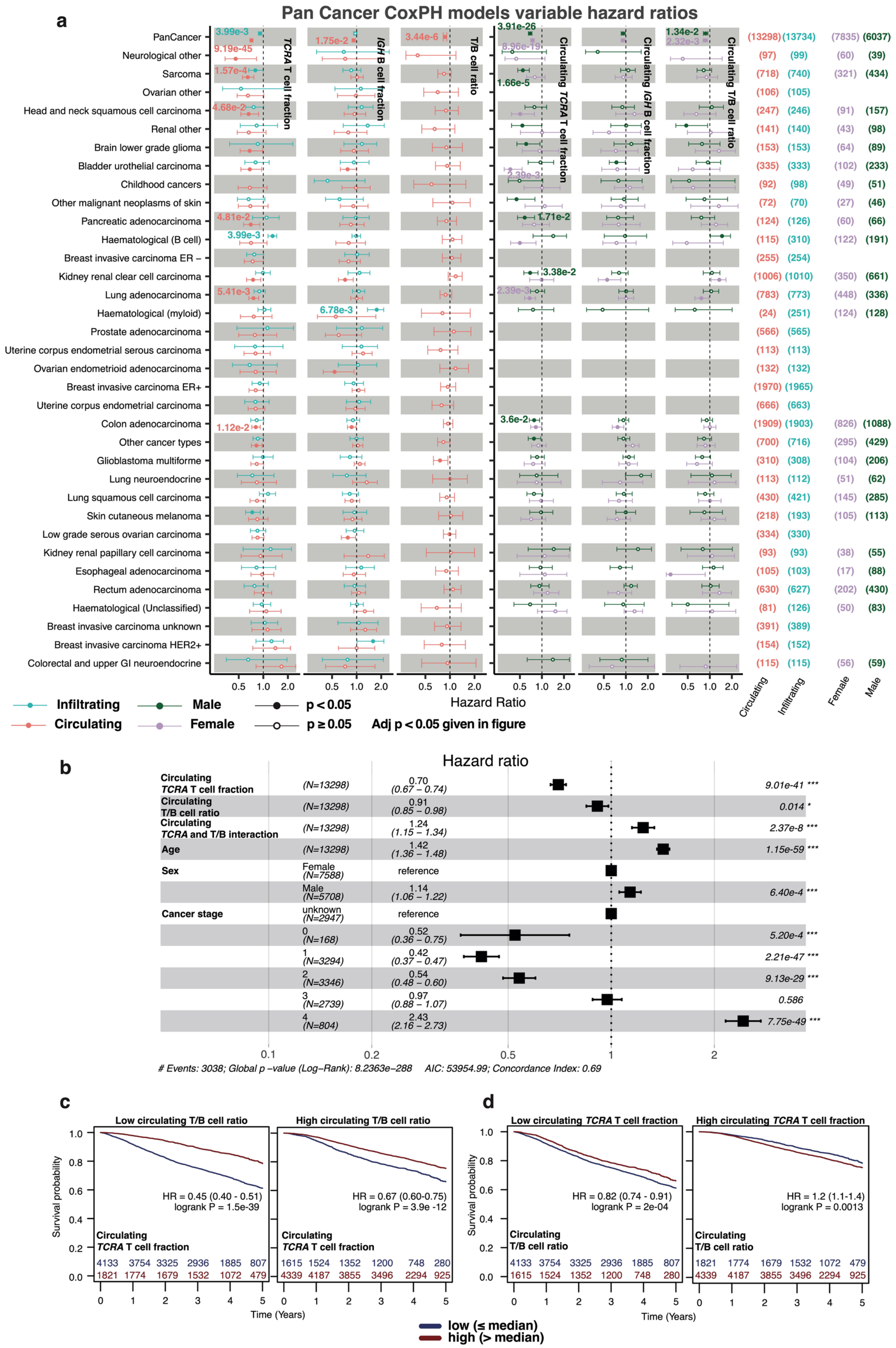
Extended Data Fig. 6 | Comparison of circulating and infiltrating fractions. **a**, Boxplots showing different fractions calculated in ImmuneLENS in circulating blood and infiltrating tumor samples; pie charts show the percentage of cases when infiltrating fractions is higher or lower than circulating with significance assessed using a two-sided Wilcoxon rank-sum test. P = 0, represents P values less than the limit of double-precision floating numbers in R, 2.22 × 10^{-308} **b**, Pearson correlation of infiltrating and circulating fractions within the 100KGP cohort. Dashed black line represents the Pearson correlation = 0 and separates positive and negative correlations. Multiple hypothesis adjustments were performed using the Holm–Bonferroni method. Boxplots in **a** show the median, lower and upper quartile and with whiskers extending to 1.5× interquartile range above and below the interquartile range. P values for Pearson's R values were derived from a two-tailed t-distribution using the correlation coefficient and sample size.



Extended Data Fig. 7 | Additional determinants of ImmuneLENS fractions. **a**,**b**, Significance of hit SNPs from PLINK analysis for circulating *TCRA* T cell fraction in healthy cohorts identified in European ancestry (**a**) and African ancestry (**b**) versus their significance in the cancer cohort, with SNPs colored by their position within the chromosome to distinguish between multiple significant loci. Dashed red line represents p = 0.05 (unadjusted) in the cancer cohort. The P values are derived from the PLINK software that uses a linear regression model and performs a Wald test for each SNP. For the cancer cohort, this was done separately for each histology, and the P values were combined using a meta-analysis with a common effects model.

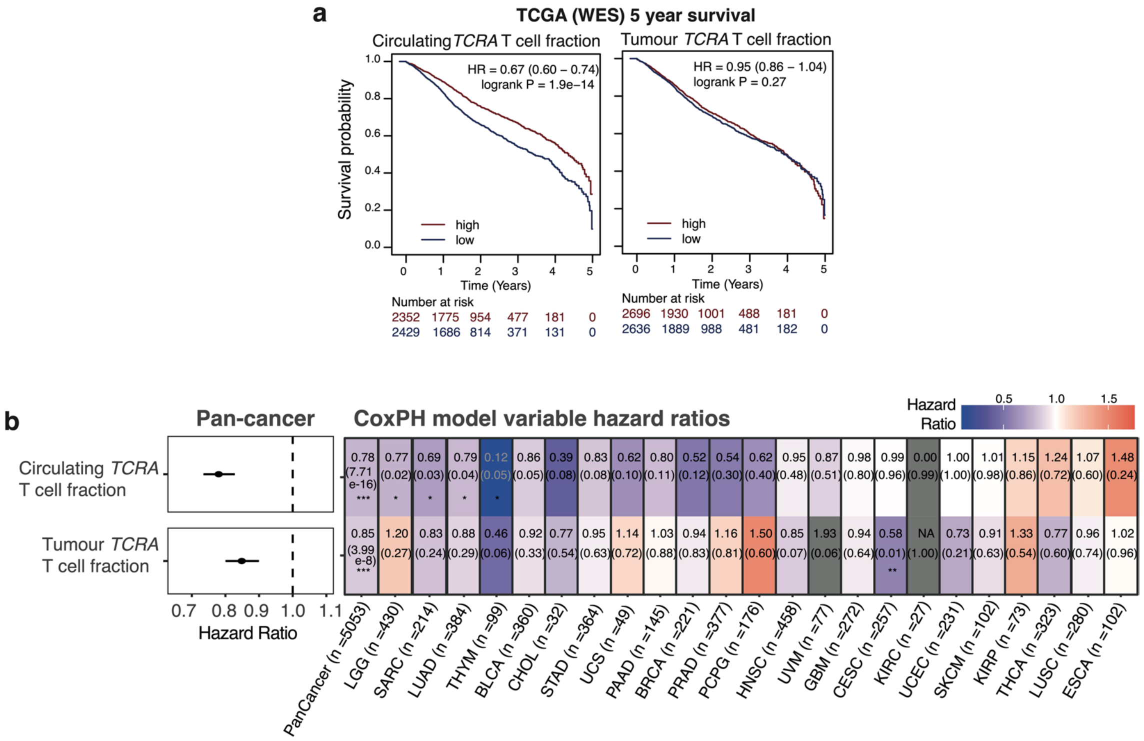


Extended Data Fig. 8 | Association of selection with infiltrating B cell fraction. a, Volcano plots for the significance of association of IgM/IgD and IgG infiltrating B cell fraction for selection of nonsynonymous genes from the cancer gene census as measured with Poisson model: observed mutations ~ offset(log(expected mutation)) + age + sex + purity + tumor mutation burden + disease type + infiltrating B cells, with expected mutations calculated by dNdScv, and run on the entire pan-cancer cohort excluding hematological and childhood cancers. Dashed black lines are at estimate = 0, and the FDR significance threshold of $-\log_{10}(P) = 4.16$. **b**, Bubble plot showing disease-type-specific significance from Poisson model for different infiltrating B cell fractions, with genes selected as those that are significant either at the pan-cancer level (*MUC4* in IgM/IgD and *KMT2C* in IgG) or within a single disease-type at an adjusted $P < 0.05$ with genes only tested if 10 or more patients had nonsynonymous mutations within that gene in that cancer type. P values in **a** and **b** represent the significance of the term for the *TCRA* T cell fraction variable in the Poisson model and are calculated using a Wald test.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Survival forest plots and circulating T cell fraction and T/B cell ratio interaction. a, Output from CoxPH models showing hazard ratio with 95% confidence intervals for 5-year survival controlled for age, sex, cancer stage and treatment before surgery. Dashed black line represents a hazard ratio of 1. **b**, Results from a CoxPH model for circulating TCRA T cell fraction and T/B cell ratio together with their interaction term from the 100KGP pan-cancer cohort showing hazard ratio with 95% confidence interval. Dashed black line represents a hazard ratio of 1. **c**,**d**, Kaplan–Meier curves for 5-year survival for 100KGP pan-cancer cohort separated into either low- or high-circulating T/B cell ratio or circulating TCRA T cell fraction based on the median values. Multiple hypothesis adjustments were performed using the Benjamini–Hochberg method with individual P values calculated using a two-sided Wald test within the Cox model.



Extended Data Fig. 10 | Survival analysis in TCGA. a, Five-year survival Kaplan–Meier curves for the TCGA data using the median value of *TCRA* T cell fraction as calculated by T cell ExTRECT to assign high and low groups. **b**. Left: hazard ratio associated with pan-cancer TCGA cohort with 95% confidence interval. Right: heatmap of hazard ratios for each individual cancer type with P values given in brackets. All P values were calculated using a two-sided Wald test within the Cox model. *, P < 0.05; **, P < 0.01; ***, P < 0.001. LGG, brain lower grade glioma; SARC, sarcoma; LUAD, lung adenocarcinoma; THYM, thymoma; BLCA, bladder urothelial carcinoma; CHOL, cholangiocarcinoma; STAD, stomach adenocarcinoma; UCS, uterine carcinosarcoma; PAAD, pancreatic adenocarcinoma; BRCA, breast invasive carcinoma; PRAD, prostate adenocarcinoma; PCPG, pheochromocytoma and paraganglioma; HNSC, head and neck squamous cell carcinoma; UVM, uveal melanoma; GBM, glioblastoma multiforme; CESC, cervical squamous cell carcinoma and endocervical adenocarcinoma; KIRC, kidney renal clear cell carcinoma; UCEC, uterine corpus endometrial carcinoma; SKCM, skin cutaneous melanoma; KIRP, kidney renal papillary cell carcinoma; LUSC, lung squamous cell carcinoma; ESCA, esophageal carcinoma.

nature portfolio

Corresponding author(s): Dr Nicholas McGranahan

Last updated by author(s): Nov 5, 2024

Reporting Summary

Nature Portfolio 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 Portfolio 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 4.0.2 samtools (1.3.1) PLINK (1.9) R packages used: tidyverse (1.3.2) ggplot2 (3.4.1) dplyr (1.1.0) tidyr (1.3.0) ggpubr (0.6.0) scales (1.2.1) rstatix (0.7.2) lubridate (1.9.2) survminer (0.4.9) survival (3.1-12) survcomp (1.40.0) RColorBrewer (1.1-3) gridExtra (2.3) gtable (0.3.2)

GGgally (2.1.2)
ggforce (0.4.1)
TCellExTRECT (1.0.1)
MatchIt (4.5.0)
dndscv (0.0.1.0)

The code to estimate T and B cell fractions, B cell class switching and diversity metrics was produced using the custom made ImmuneLENS R package which is available at https://github.com/McGranahanLab/ImmuneLENS.

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 Portfolio 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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

TRACERx100:
The RNA sequencing (RNA-seq) and whole exome sequencing (WES) data (in each case from the TRACERx study) used during this study is a subset of the TRACERx421 data set and have been deposited at the European Genome–phenome Archive (EGA), which is hosted by The European Bioinformatics Institute (EBI) and the Centre for Genomic Regulation (CRG) under the accession codes EGAS00001006517 (RNAseq), EGAS00001006494 (WES); access is controlled by the TRACERx data access committee. Details on how to apply for access are available at the linked pages. For TCRseq data used in this analysis the FASTQ data is deposited at the Short Read Archive (SRA) under accession code BioProject: PRJNA544699

Coverage files for the TCRA, TCRB, TCRG and IGH loci were generated from WGS TRACERx100 samples. These coverage files used for the calculation of the T and B cell fractions is available at zenodo.org (10.5281/zenodo.7785803) and were the only data derived from the TRACERx WGS analysis used within this paper.

100KGP:
WGS and phenotypic data from the 100,000 Genomes Project can be accessed by application to Genomics England following the procedure outlined at https://www.genomicsengland.co.uk/about-gecip/joining-research-community/.

1000 Genomes:
The 1000 Genome Data used is publicly available and can be accessed on https://www.internationalgenome.org/data. Calculated ImmuneLENS output including class switching and polyclonal predictions for each LCL cell line included are available on zendodo (10.5281/zenodo.1109397).
PCAWG:
PCAWG data used in this study was obtained through our collaboration with the MD Anderson. To gain access to the TCGA portion of the PCAWG data used in this study, researchers need to apply to the TCGA data access committee via dbGaP (https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=login). The calculated ImmuneLENS output for these samples is available on zenodo.org (10.5281/zenodo.1109396)

TCGA:
To access TCGA WES and low pass WGS data, researchers will need to apply to the TCGA Data Access Committee (DAC) via dbGaP (https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page=login).

The calculated T cell ExTRECT TCRA T cell fraction scores along with the ImmuneLENS output for the low pass WGS samples used in this study is available at zenodo.org (10.5281/zenodo.7794867).
scRNA datasets:
All data used in this analysis is described in Salcher et al50. available at https://cellxgene.cziscience.com/collections/edb893ee-4066-4128-9aec-5eb2b03f8287.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender

100KGP pan cancer cohort: Breakdown by phenotypic sex classification at birth as provided in the Genomics England Research Environment: Females (8197), Males (6304)

100KGP healthy cohort: Breakdown by phenotypic sex classification at birth as provided in the Genomics England Research Environment: Females (17169), Males (13496)

TRACERx: There were 68 male and 32 female non-small cell lung cancer patients in the TRACERx study

PCAWG: The subset of the PCAWG data used for the analysis contains 308 females and 231 males

Reporting on race, ethnicity, or other socially relevant groupings

Most probable ancestry within the 100KGP is provided in the Genomics England Research Environment based on five broad super-populations (see https://re-docs.genomicsengland.co.uk/ancestry_inference/) for African, Admixed American, East Asian, European and South Asian. An unassigned ancestry group was also used for participants with admixed ancestry where no probability for an individual super-population was above 80%.

The breakdown by ancestry within the 100KGP is as follows:
Healthy cohort:
African (664), Admixed American (99), East Asian (185), European (23636), South Asian (3542), Unassigned (2265), Unknown (274)
Pan cancer cohort:
African (447), Admixed American (32), East Asian (114), European (12489), South Asian (515), Unassigned (577), Unknown (327)

Population characteristics

100KGP:
Healthy cohort consists of 30,665 participants originating from the rare disease cohort arm of the 100KGP and representing the non-affected relatives.
Age breakdown (years): < 20 (1445), 20-24 (537), 25-29 (1807), 30-34 (3942), 35-39 (5465), 40-44 (4787), 45-49 (3954), 50-54 (2745), 55-59 (1798), 60-64 (1295), 65-69 (1010), 70-74 (893), 75-79 (518) ,>80 (469)

Pan cancer cohort consists of 14,501 participants covering 33 main cancer types with samples derived from both tumour tissue and matched germline.
Age breakdown (years): < 20 (284), 20-24 (75), 25-29 (147), 30-34 (194), 35-39 (285), 40-44 (410), 45-49 (780), 50-54 (1129), 55-59 (1446), 60-64 (1790), 65-69 (2154), 70-74 (2208), 75-79 (1569),>80 (1399)

TRACERx:
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

100KGP: Cases were recruited by referring clinicians through the National Health Service

TRACERx:
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.

PCAWG:
Patients were recruited by the participating centres following local protocols. Samples obtained had to meet criteria on amount of tumour DNA available, meaning that the cohort is potentially somewhat biased towards larger tumours. Otherwise, we anticipate no major recruitment biases.

Ethics oversight

100KGP:
The 100,000 Genomes project was approved by East of England–Cambridge Central Research Ethics Committee ref:20/

EE/0035.
TRACERx:
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)
PCAWG: The Ethics oversight for the PCAWG protocol was undertaken by the TCGA Program Office and the Ethics and Governance Committee of the ICGC. Each individual ICGC and TCGA project that contributed data to PCAWG had their own local
arrangements for ethics oversight and regulatory alignment.

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

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

No sample size calculations were performed, we analysed existing datasets, namely the TRACERx100 cohort and the 100KGP cohort.

The TRACERx100 data set was chosen as it was an existing cohort containing orthogonal immune data (RNAseq, TCRseq) for which ImmuneLENS could be easily validated. The 100KGP dataset was chosen due to being a large WGS cohort without any orthogonal immune related data for which insights on regulation of circulating and infiltrating immune cell fractions could be gained.

Additional validation was done on:
1) 1000 genome cohort, which was chosen due to containing WGS samples originating from B cell derived cell lines to specifically validate our B cell fractions.
2) PCAWG and TCGA cohorts data sets for validation of our pan cancer analysis results from the 100KGP on a separate data set.

Data exclusions

Within the 100KGP pan cancer cohort participants with multiple tumour samples were excluded due to lack of annotation of the reason for multiple samples (e.g. technical resequencing, representative of metastasis, multiple region sequenced or occurrence of a second primary tumour at a later time point) these were removed from the pan cancer cohort. All germline samples not derived from blood samples were also excluded from the analysis due to our focus on circulating immune fraction in this study.

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
☐	☒	Clinical data
☒	☐	Dual use research of concern
☒	☐	Plants

Methods

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

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	TRACERx100: NCT01888601 100KGP: N/A
Study protocol	TRACERx100: The study protocol is available at NEJM.org linked to Jamal-Hanjani et al NEJM 2017 (PMID: 28445112) 100KGP: Refer to Genomic England Limited website for information on data collection.
Data collection	TRACERx100: 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 100KGP: Refer to Genomic England Limited website for information on data collection.
Outcomes	TRACERx100: The outcome measures of the TRACERx trial are intratumour heterogeneity, disease-free survival, and overall survival. 100KGP: N/A