ANALYSIS
https://doi.org/10.1038/s41588-018-0339-x

nature genetics

Corrected: Publisher Correction

Landscape of B cell immunity and related immune evasion in human cancers

Xihao Hu[1,12], Jian Zhang[2,12], Jin Wang[1,3], Jingxin Fu[1,3], Taiwen Li[4], Xiaoqi Zheng[5], Binbin Wang[1,3], Shengqing Gu[1], Peng Jiang[1], Jingyu Fan[1,3], Xiaomin Ying[2], Jing Zhang[3], Michael C. Carroll[6], Kai W. Wucherpfennig[7], Nir Hacohen[8], Fan Zhang[3,9], Peng Zhang[3,9], Jun S. Liu[10*], Bo Li[11*] and X. Shirley Liu[1,3,10*]

Tumor-infiltrating B cells are an important component in the microenvironment but have unclear anti-tumor effects. We enhanced our previous computational algorithm TRUST to extract the B cell immunoglobulin hypervariable regions from bulk tumor RNA-sequencing data. TRUST assembled more than 30 million complementarity-determining region 3 sequences of the B cell heavy chain (IgH) from The Cancer Genome Atlas. Widespread B cell clonal expansions and immunoglobulin subclass switch events were observed in diverse human cancers. Prevalent somatic copy number alterations in the *MICA* and *MICB* genes related to antibody-dependent cell-mediated cytotoxicity were identified in tumors with elevated B cell activity. The IgG3–1 subclass switch interacts with B cell–receptor affinity maturation and defects in the antibody-dependent cell-mediated cytotoxicity pathway. Comprehensive pancancer analyses of tumor-infiltrating B cell–receptor repertoires identified novel tumor immune evasion mechanisms through genetic alterations. The IgH sequences identified here are potentially useful resources for future development of immunotherapies.

B cells are a key component of adaptive immunity with diverse functions including antibody production[1,2], antigen presentation[3], and cellular cytotoxicity[4]. Infiltrating B cells have been frequently observed in multiple tumor tissues[5–7], yet their reported effects on patient outcome have been inconsistent[5,8–11]. It remains unclear what roles B cells play in the antitumor humoral response and how cancer cells interact with infiltrating B cells.

The B cell immunoglobulin heavy chain (IgH) consists of a hypervariable complementarity-determining region 3 (CDR3), which is critical in antigen recognition[12]. After binding a foreign antigen, B cells undergo proliferation, class switch recombination (CSR), and somatic hypermutations (SHMs) and produce high-affinity antibodies to eliminate the antigen[13,14]. Therefore, characterization of the tumor-infiltrating B cell immunoglobulin repertoire is critical to understanding B cell immunity in tumors. Efforts have been made to study the B cell repertoire by using either targeted deep sequencing (B cell–receptor repertoire sequencing (BCR-seq))[15–17] or unselected RNA-sequencing (RNA-seq) data[18,19] in both human and mouse models to understand the etiology of autoimmune diseases[20] or cancers[21,22]. However, a systematic investigation of tumor-infiltrating B cell repertoires using large cohorts of diverse cancer types is still lacking to elucidate the functional effect of tumor B cell immunity and identify potential therapeutic opportunities.

Previously, we developed an ultrasensitive de novo assembler, TRUST, to call the T cell–receptor hypervariable CDR3 sequences by using bulk tumor RNA-seq data[23,24]. In this work, we enhanced TRUST to assemble the B cell IgH CDR3 sequences from bulk RNA-seq data and applied it to study the infiltrating B cell IgH repertoire in The Cancer Genome Atlas (TCGA) cohorts. A subset of B cells with a defined signature of CSR emerged in our analysis and showed promising anti-tumor effects. We observed potential mechanisms of anti-tumor B cell responses and tumor evasion to B cell attack. These results help elucidate the functional effects of antibody-mediated cell cytotoxicity in antitumor immune responses and reveal promising opportunities in developing future immunotherapies.

Results

De novo assembly of IgH hypervariable sequence. We modified TRUST, a computational algorithm that we previously developed to detect T cell–receptor hypervariable CDR3 sequences, to assemble the CDR3 regions of tumor-infiltrating B cell IgH from unselected tissue or tumor RNA-seq data (see Methods). To systematically evaluate the performance of TRUST, we applied in silico simulations to produce artificially recombined and hypermutated immunoglobulin transcripts. The enhanced TRUST achieved high sensitivity and perfect precision at very low sequence coverage (0.1×) (Supplementary Fig. 1a), thus suggesting its suitability for detecting IgH hypervariable sequences from tumor RNA-seq data. In addition, we performed BCR-seq on six tumors to further evaluate the BCR clones that TRUST assembled from RNA-seq on the same tumors. We found that TRUST robustly recovered expanded B cells through highly sensitive and precise calling of abundant BCR clones (Fig. 1a),

[1]Department of Data Sciences, Dana-Farber Cancer Institute and Harvard T.H. Chan School of Public Health, Boston, MA, USA. [2]Center for Computational Biology, Beijing Institute of Basic Medical Sciences, Beijing, China. [3]Shanghai Key Laboratory of Tuberculosis, Clinical Translational Research Center, Shanghai Pulmonary Hospital, School of Life Sciences and Technology, Tongji University, Shanghai, China. [4]State Key Laboratory of Oral Diseases, West China Hospital of Stomatology, Sichuan University, Chengdu, China. [5]Department of Mathematics, Shanghai Normal University, Shanghai, China. [6]Program in Cellular and Molecular Medicine, Boston Children's Hospital, Boston, MA, USA. [7]Department of Cancer Immunology and Virology, Dana-Farber Cancer Institute, Boston, MA, USA. [8]Massachusetts General Hospital Cancer Center, Harvard Medical School, Boston, MA, USA. [9]Department of Thoracic Surgery, Shanghai Pulmonary Hospital, Tongji University School of Medicine, Shanghai, China. [10]Department of Statistics, Harvard University, Cambridge, MA, USA. [11]Lyda Hill Department of Bioinformatics, UT Southwestern Medical Center, Dallas, TX, USA. [12]These authors contributed equally: Xihao Hu, Jian Zhang. *e-mail: jliu@stat.harvard.edu; bo.li@utsouthwestern.edu; xsliu@jimmy.harvard.edu

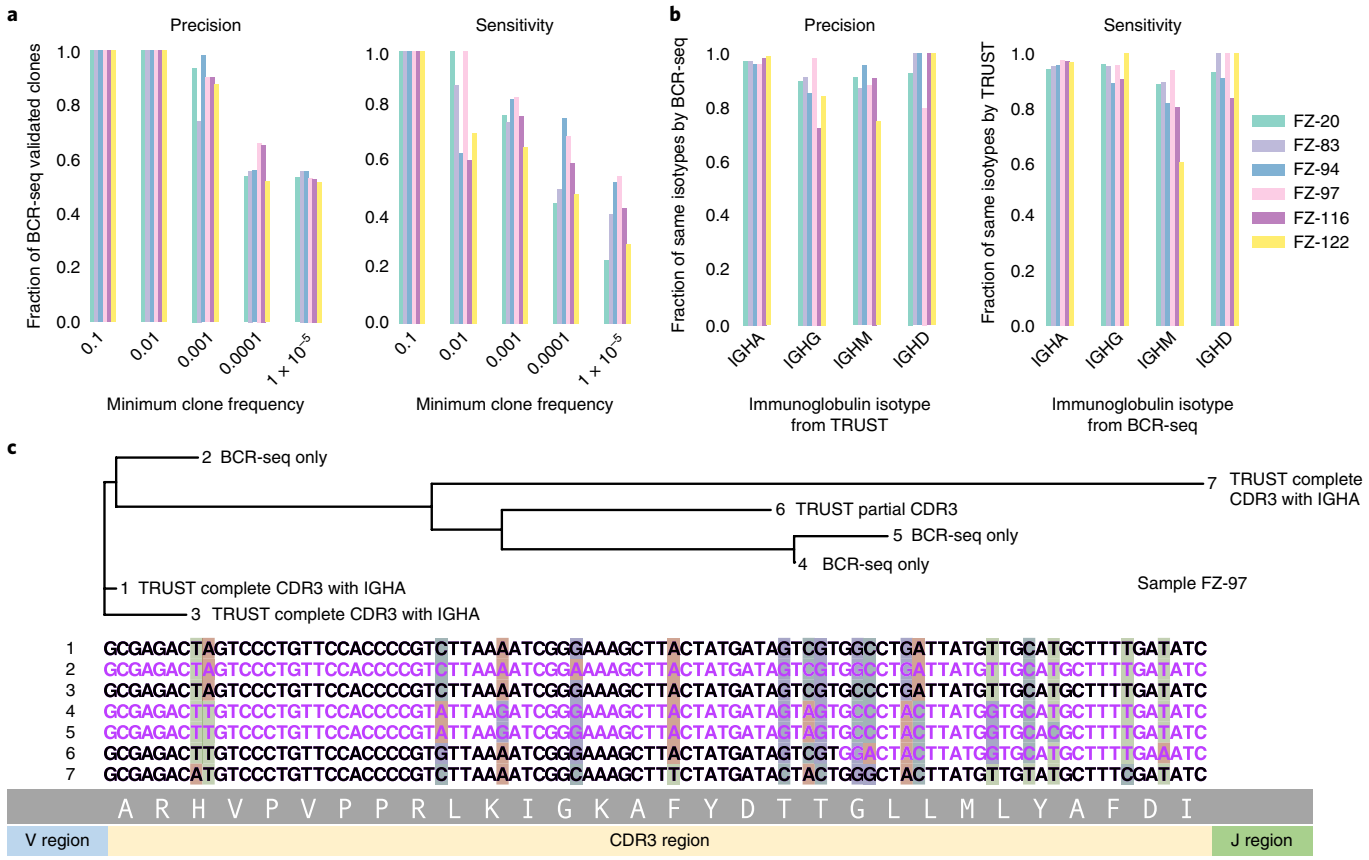

Fig. 1 | TRUST performance on tumor samples with matched BCR-seq data. a, Evaluation of the TRUST-reported CDR3s under different cutoffs on the minimum clonal frequency. Precision is the fraction of TRUST-called CDR3s validated by BCR-seq, and sensitivity is the fraction of BCR-seq CDR3s called by TRUST. **b**, Evaluation of the TRUST-reported immunoglobulin isotypes with the same CDR3 in BCR-seq. Precision is the fraction of TRUST-called isotypes validated by BCR-seq, and sensitivity is the fraction of BCR-seq isotypes called by TRUST. **c**, An example of B cell cluster where three sequences (1, 3, 7) were identified by TRUST with the same immunoglobulin isotype as in BCR-seq, and TRUST recovered a partial CDR3 for sequence number 6. Bases in BCR-seq clones but missed by TRUST are highlighted in violet.

with consistent clonal frequency estimations (Supplementary Fig. 1b) and high specificity in calling individual-specific clones (Supplementary Fig. 1c). Moreover, TRUST and BCR-seq agreed on most of the immunoglobulin isotype annotations (Fig. 1b), thus allowing us to investigate CSR events in expanded B cells by using TCGA data. Although some of the TRUST assemblies were partial CDR3 sequences, they still contained sufficient information to reconstruct B cell clusters (Fig. 1c).

Isotypes and SHMs of B cells in TCGA samples. TRUST assembled a total of 30.8 million CDR3 sequences from 9,025 TCGA RNA-seq samples across 32 cancer types with variable sequencing depths and read lengths (Supplementary Table 1). The number of assemblies was highest in lung squamous cell carcinoma (median 5,799 CDR3s per sample), a result consistent with the estimated leukocyte fractions[1]. A total of 20.6 million assemblies were assigned to known immunoglobulin classes by using paired-end reads or assembled constant regions (Methods). Of these, IgG was the most abundant (60%) class, followed by IgA (35%) and IgM (4%). The average length of IgH hypervariable sequences was 14.7 amino acids, and different immunoglobulin classes had similar length distribution and sequence motifs (Supplementary Fig. 2a). These results agree with those[25] from immunoglobulin sequences in the IMGT database[26]. Immunoglobulin class abundance varied across different cancer types (Fig. 2a), and IgG was the dominant class in thyroid, testicular, and skin cancers. In contrast, IgA was the largest fraction in kidney, pancreatic, and colorectal cancers, a finding consistent with the high secretion levels of IgA in mucous membrane and glands[27].

We called SHMs if two CDR3 sequences differed by only one nucleotide. The resulting SHMs were enriched on the third codon position of CDR3s across different lengths, thus suggesting that our SHM calls were unlikely to arise from sequencing errors (Fig. 2b). Indeed, 85% of the 5.2 million SHM calls in the CDR3s were synonymous, thus indicating strong selection pressures during affinity maturation (Supplementary Fig. 2b). Nearly half of the SHMs were transitions within pyrimidines or purines (Fig. 2c), results similar to previous observations on the whole BCR heavy chain[28]. We next performed 96-triplet mutation context analysis (Fig. 2e) on cases in which we were able to infer the mutation directions (Methods). The strongest mutation signature was the ACT to ATT triplet, a finding consistent with the (W)RCY motif of activation-induced cytidine deaminase (AICDA or AID)[29]. The second highest signature was a GC-rich motif, which is not specific to mutation types and might arise from different DNA repair pathways and BCR affinity selection[30]. Furthermore, we observed the highest SHM in IgG antibodies (Fig. 2d), in agreement with findings reported in healthy individuals[31,32]. Finally, the SHM rate (Methods) was positively correlated with the expression of *AID* across TCGA samples ($\rho = 0.2$, $P < 10^{-40}$, Supplementary Fig. 2c), thus supporting the role of AID in introducing SHMs in tumor-infiltrating B cells[33].

Detection of clonal expansion of tumor-infiltrating B cells. SHM and CSR are signatures of B cell clonal expansion on antigenic

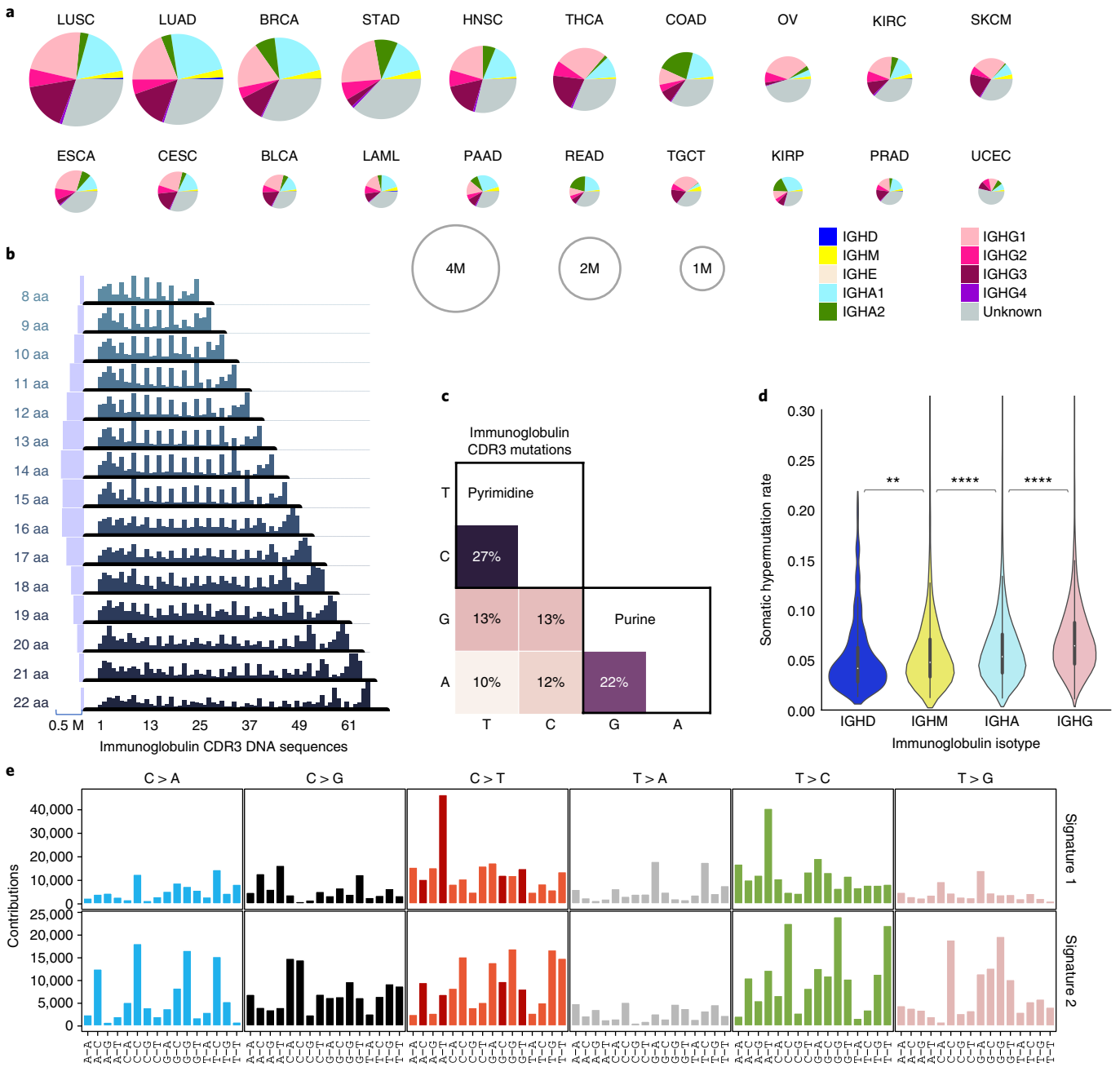

Fig. 2 | Immunoglobulin isotypes and SHMs of the tumor-infiltrating B cell IgH repertoire. a, Distribution of nine immunoglobulin isotypes across cancer types. Fractions of each immunoglobulin isotype were calculated as described in the Methods. The top 20 cancer types with most IgH CDR3 assemblies are displayed. Circle size is proportional to the total number of CDR3 assemblies in each cancer type. M, millions. **b**, Distribution of SHMs in the CDR3 region were grouped by CDR3 length, as indicated by the number of amino acids (aa). The number of mutation events is shown by the size of horizontal bars, where CDR3 with 14 aa had the largest number of mutations (~0.5 million). **c**, Proportion of nucleotide pairs ($n = 3{,}808{,}402$) before and after SHMs. The most likely mutations were within the pyrimidine (C/T) and purine (A/G) groups. **d**, Distributions of SHM rates within each immunoglobulin isotype. The SHM rates were calculated from CDR3 pairs with the same immunoglobulin isotype in each patient sample. Statistical significances were evaluated with Wilcoxon rank-sum test (two sided $^{****}P < 1\times10^{-4}$; $^{**}P < 0.01$). The boxes in violin plots show the lower, median, and upper quartiles of values, and the curves show the density of values (sample sizes and *P* values are shown in Supplementary Table 3). **e**, The spectrum of base changes identified in the CDR3 region with the mutation type displayed in the title. The *x* axis is the adjacent 3′ and 5′ bases of the mutation, and the *y* axis shows the contribution of the mutation context. The (W)RCY motif of AID (R = A/G, Y = T/C) is highlighted as dark red.

recognition[34]. To detect B cell clonal expansion, we developed a multiple local sequence alignment approach and optimized the parameters (Methods) by controlling the false discovery rate (FDR) and maximizing the number of clusters (Supplementary Fig. 3a–c). For example, in the experimental validation sample FZ-97, we obtained seven highly similar CDR3s from BCR-seq, of which four were detected by TRUST with one partial CDR3 sequence (number 6 in Fig. 1c). Using this alignment approach, we detected a total of 434,106 B cell lineage clusters across 5,866 TCGA tumors. Most (54.5%) of the clusters with unique immunoglobulin annotation were assigned IgG, a finding that might be related to a previous observation that tumor-infiltrating B cells generally express

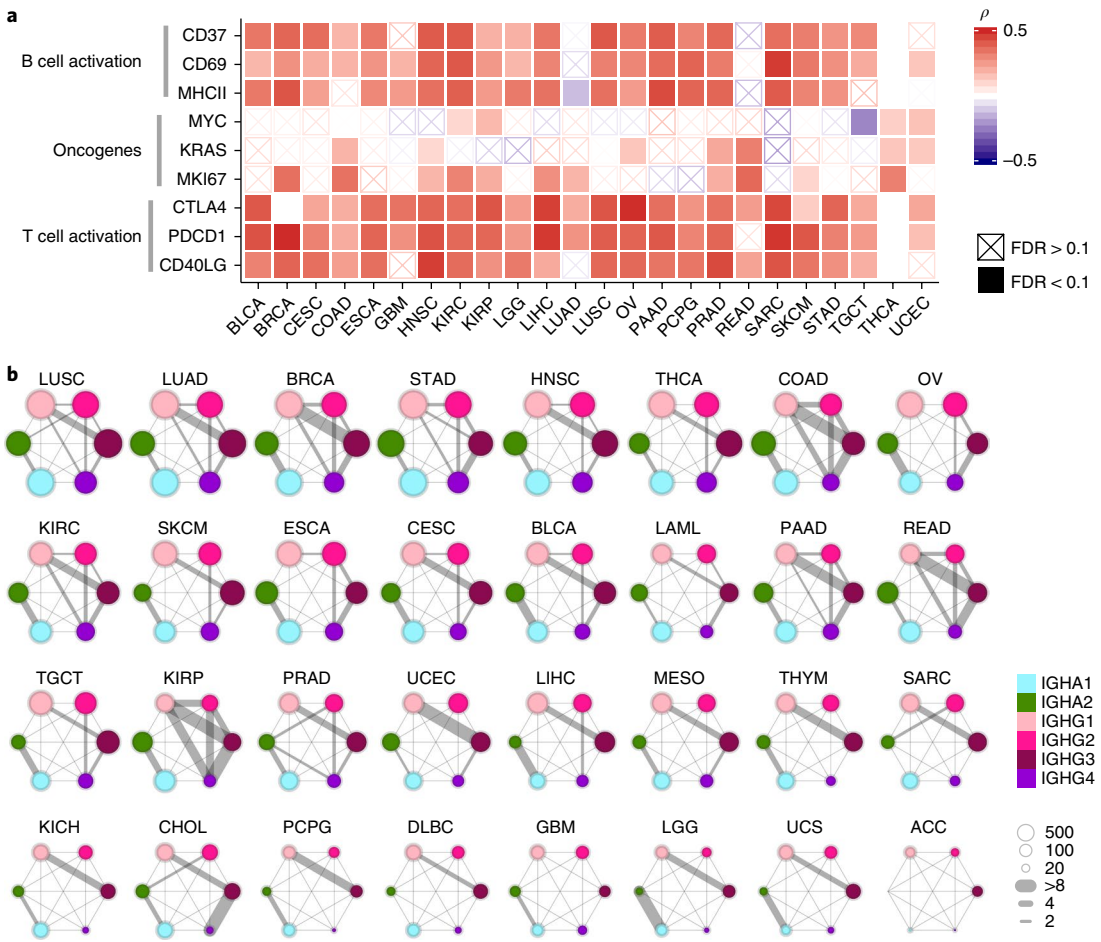

Fig. 3 | Immunoglobulin isotype subclass switches are associated with B cell and T cell activation. a, Correlations of immunoglobulin switch levels and highly expressed genes in immune cells and malignant cells with corrections on the tumor purity (Methods). The partial Spearman's correlations and two-sided *P* values were from 6,430 samples (sample sizes in Supplementary Table 3). Immunoglobulin switch levels are the number of B cell clusters with any type of switches divided by the total number of unique CDR3s in a patient sample. **b**, Visualization of immunoglobulin isotype co-occurrence within B cell clusters by cancer type. Circle size represents the number of clusters carrying a given immunoglobulin isotype. Lines connecting two circles indicate the enrichment level of observing switches in the two corresponding immunoglobulin subclasses. The enrichment level is the ratio of observed and expected switches if immunoglobulin isotypes are assumed to be independently distributed among B cell clusters. Cancer types were sorted by the total number of CDR3s.

IgG with evidence of antigen-driven expansion[5]. In addition, the complete CDR3 sequences from expanded B cell clones were significantly longer than the nonexpanded clones (Supplementary Fig. 3d), thus suggesting potentially higher structural complexity of the expanded BCRs.

In a subset of 311,173 B cell clusters distributed among 5,328 samples, we identified the coexistence of more than one immunoglobulin subclass (Methods), providing evidence of CSR during B cell clonal expansion. The normalized number of B cell clusters had a much higher positive correlation with B cell and T cell activation markers than with oncogene expression in most tumor types, thereby suggesting that the expanded B cell clusters might recognize tumor antigens (Fig. 3a). We observed coexistence of multiple immunoglobulin classes or subclasses in 296,820 clusters, a result suggesting subsequent CSRs following the initial CSR event. This finding is consistent with a previous report that B cells may undergo a programmed order of IgG subclass switch recombination (sCSR) in an immune response[35]. We examined the potential sCSR events for the IgA and IgG isotypes (Fig. 3b), and observed a total of 153,476 IgG3 to IgG1 (IgG3–1) sCSR events in which the same CDR3 cluster contained both IgG1 and IgG3 isotypes. IgG3–1 sCSR was the most abundant and enriched sCSR type in multiple cancers, especially in breast, kidney, endometrial, and colorectal cancers. The TRUST BCR results for TCGA, including the number of BCRs and IgG3–1 sCSR events, are summarized in Supplementary Table 2.

We next compared the signatures of B cell repertoires between tumor and adjacent normal tissues to evaluate whether clonal expansion and sCSR events might be enriched in tumors. B cell infiltration levels, as estimated from the Tumor Immune Estimation Resource (TIMER)[36], did not show consistent differences between tumors and adjacent normal tissues across different cancer types. However, B cell diversity, as defined by the number of CDR3s per thousand BCR reads[23], was significantly lower in tumor samples for most cancer types, a result indicating that B cells were more clonal in tumors than in adjacent normal tissues (Fig. 4). We observed <10% overlap of shared B cell clonotypes between matched tumor and normal tissue samples (Supplementary Fig. 3e). Additionally, in most cancer types, tumor samples were enriched for more B cell clusters with IgG CSRs or IgG3–1 sCSRs rather than with IgA CSRs. Furthermore, individuals with higher levels of B cell IgG3–1 switches had significantly better clinical outcomes in melanoma, ovarian cancer, and thyroid cancer (Supplementary Fig. 4). In contrast, kidney tumors with high IgG3–1 switches had poorer

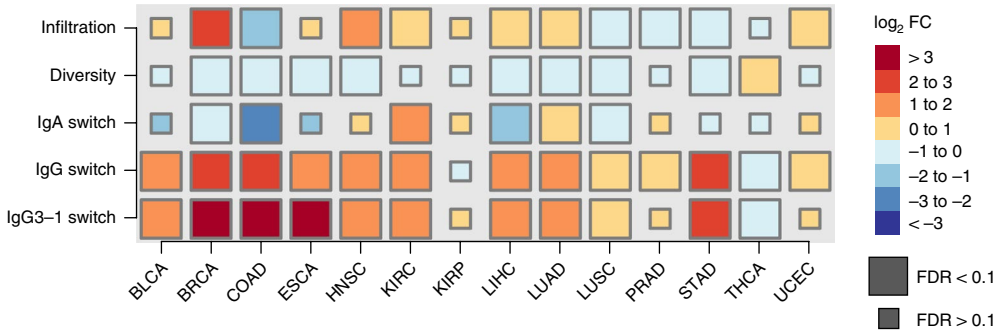

Fig. 4 | Features of tumor-infiltrating B cell repertoire compared between tumor and adjacent normal tissues. Multiple aspects of infiltrating B cells are displayed on the heat map, including B cell–infiltration levels, B cell diversity, and levels of different immunoglobulin switches. B cell immunoglobulin switch levels were calculated as the number of B cell clusters with the immunoglobulin switch divided by the total number of unique CDR3s (Methods). Cancer types with more than 100 samples for tumor and more than10 samples for adjacent normal tissues are presented (numbers in Supplementary Table 3). Log fold change (log FC) was calculated as the $\log_2$ ratio of the average values between tumor and adjacent normal tissues. Statistical significance was evaluated with Student's *t* test and two-sided *P* values. FDR correction was performed with the Benjamini–Hochberg procedure for testing on multiple cancer types.

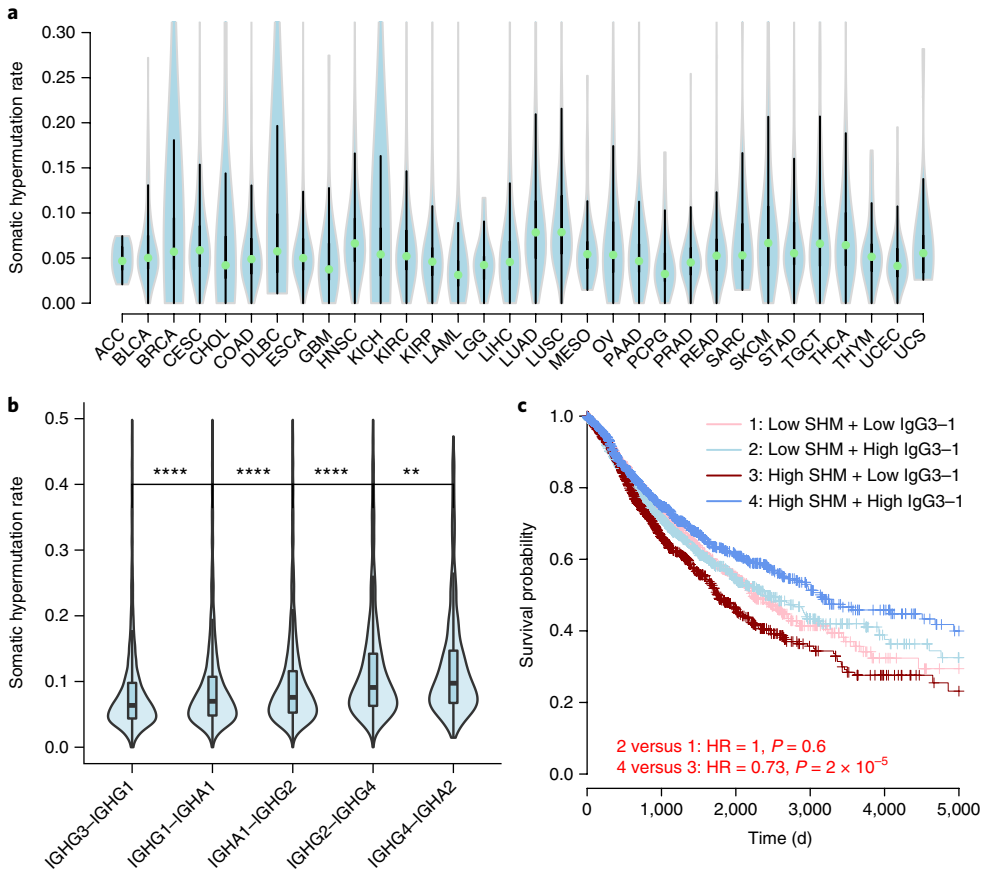

Fig. 5 | Interaction between the IgG3–1 switch and SHMs. a, Violin plots showing the distribution of SHM rates across cancer types. The points show the median values, lines extend to 1.5 times the interquartile range, and the curves show the density of values (sample sizes in Supplementary Table 3). **b**, Violin plot showing the SHM rates for samples with different immunoglobulin subclass switches. From left to right, the order of switches follows the germline gene positions, and so BCRs with a subclass switch on the right might arise from other subclass switch events on the left. Statistical significance was evaluated with the Wilcoxon rank-sum test (two-sided $^{****}P < 1 \times 10^{-4}$; $^{**}P < 0.01$). The boxes in violin plots show the lower, median, and upper quartiles of values, and the curves show the density of values (sample sizes and *P* values in Supplementary Table 3). **c**, Kaplan–Meier curves showing the interaction between SHM rate groups and IgG3–1 switch groups. SHM high ($n = 2{,}338$) or low ($n = 2{,}341$) were split by the median SHM rate in all patient samples. IgG3–1 switch high or low were split by the median IgG3–1 switch level in each SHM rate group. Statistical significance comparing different groups was evaluated with multivariate Cox regression corrected for tumor purity and patient age at diagnosis. HR, hazard ratio.

clinical outcomes, in agreement with the intriguing finding of poor prognosis of patients with kidney renal cell carcinoma with high lymphocyte infiltration[37]. These results suggest that the level of clonally expanded B cells with IgG3–1 sCSRs, instead of the total B cell infiltration level, is related to the B cell-mediated antitumor humoral response.

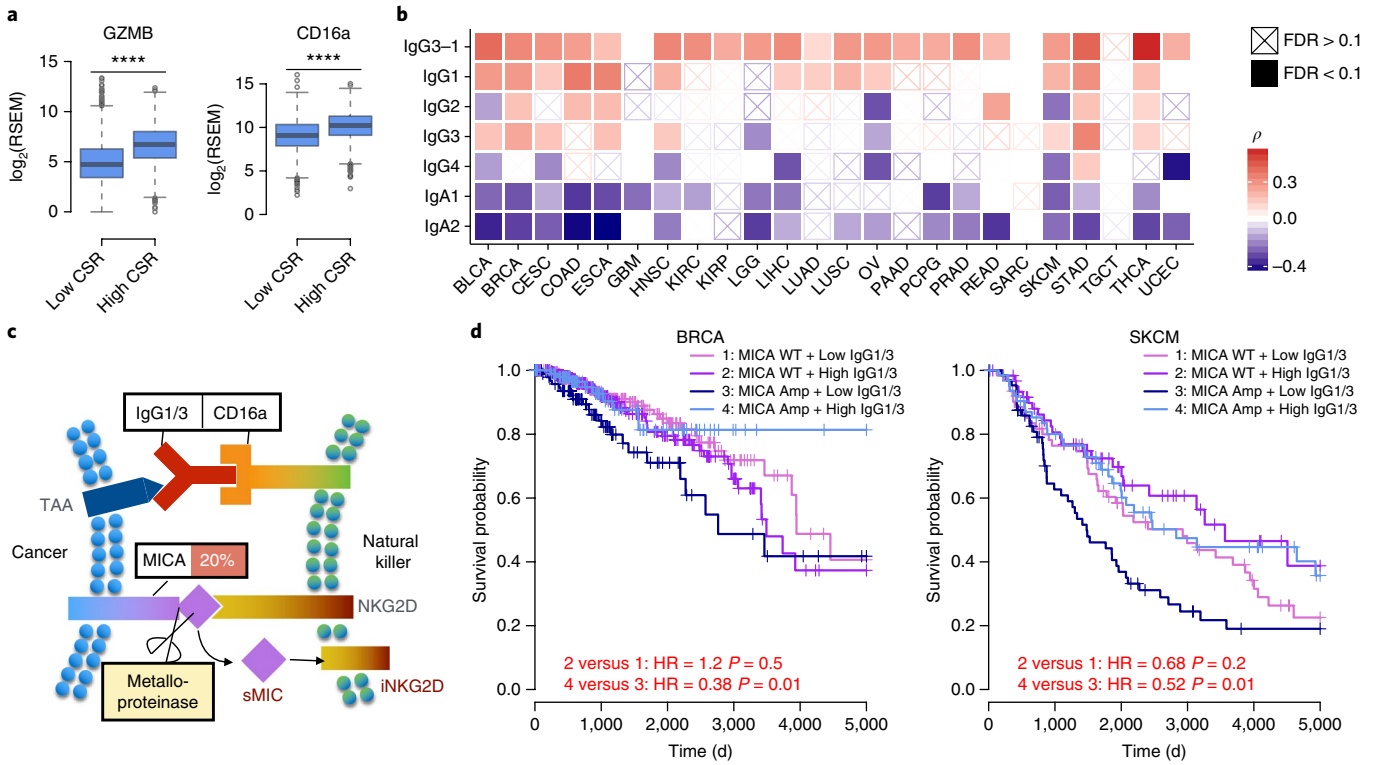

Fig. 6 | Potential immune evasion of ADCC through MIC shedding. a, NK cell activity, visualized by box plots of GZMB and CD16a. Both genes were expressed at higher levels in tumors with a high level of isotype switches (CSR). Statistical significance was evaluated with Wilcoxon rank-sum test on 4,273 low-CSR samples and 4,270 high-CSR samples. The two-sided *P* values were 6×10^{-295} and 5×10^{-156} for *GZMB* and *FCGR3A*, respectively, and are labeled with four asterisks. **b**, Heat map showing the associations between the fractions of IgG isotypes and NK cell Fc receptor *FCGR3A* (CD16a) expression across multiple cancers. Cancer type was selected on the basis of at least 100 patient samples with no missing values (sample sizes in Supplementary Table 3). Each entry in the heat map represents the partial Spearman's correlation corrected for tumor purity (Methods). FDR correction was performed on two-sided *P* values with the Benjamini–Hochberg procedure for testing on multiple cancer types. **c**, NK-cell inactivation through MIC shedding. Interaction between IgG1/3 Fc ligand and CD16a triggers NK response. *MICA* amplification was observed in 20% of tumors, and metalloproteinase overexpression may lead to the production of soluble MIC in these tumors, thus resulting in internalized NKG2D (iNKG2D) and inactivated NK cells. **d**, Interaction between *MICA* amplification and IgG1/3 levels, visualized by Kaplan–Meier curves for breast invasive carcinoma (BRCA) and skin cutaneous melanoma (SKCM). Tumor samples were categorized into four groups according to the presence of *MICA* amplification and IgG1/3 levels. MICA Amp refers to tumors with *MICA* amplification, and MICA WT refers to the remaining. High- or low-IgG1/3 groups were split by the median number of IgG1 or IgG3 B cell clusters divided by the total number of unique CDR3s in each MICA group. There were 579 MICA WT samples and 211 MICA Amp samples for BRCA, and 126 MICA WT samples and 130 MICA Amp samples for SKCM. Hazard ratios and statistical significance were evaluated with Cox proportional hazard regressions corrected for tumor purity and patient age.

Interaction between SHMs and the IgG3–1 switch. Because SHM is an important marker of B cell clonal evolution and affinity maturation[38], we studied the relationship between SHM and the IgG3–1 switch. B cell SHM rates were high in lung, head, neck, and skin cancers, and low in leukemia and brain cancers (Fig. 5a), possibly because of the immunoglobulin isotype composition in different cancer types (Fig. 2a). Because both SHM and sCSR depend on AID[39], we expected to observe a positive relationship between the steps of immunoglobulin switches and SHM rate. Indeed, we observed progressively higher SHM rates in samples with later rounds of sCSR events (Fig. 5b), thus suggesting that B cells undergo multiple rounds of sCSR events and gain additional SHMs during the process.

We next investigated the clinical relevance of interactions between the SHM rate and CSR events. Splitting TCGA samples based on the median SHM rate (5.7%), we observed a significant benefit of high IgG3–1 switches on survival in patients with high SHM rates, whereas the IgG3–1 level was not associated with survival in low-SHM samples (Fig. 5c). The same trend, although less significant because of the smaller sample sizes, was also observed in liver cancer and melanoma (Supplementary Fig. 5). The survival benefit in patients with high SHM rates and IgG3–1 sCSR suggests the role of SHM in generating a BCR repertoire with high binding affinity to the exposed tumor antigens[28].

Immune evasion of antibody-dependent cell-mediated cytotoxicity (ADCC) through MHC class I–related chain molecule (MIC) shedding. Tumors must evade immune attacks to survive and progress[40]. Tumors are well understood to evade T cell attack by expressing checkpoint ligands such as PD-L1/L2 or by acquiring defects in antigen presentation[41] or interferon-γ signaling pathways[42]. In contrast, how tumors evade B cell immunity remains largely uncharacterized. B cell IgG antibody is able to trigger downstream signaling that eliminates the affected cells; one of the most important ways in which this elimination is accomplished is through the ADCC pathway[43]. ADCC involves the recruitment of additional effector cells expressing receptors that bind the antibody Fc ligand[44]. In tumors, the ADCC pathway commonly activates natural killer (NK) cells in the microenvironment[45], which in turn carry out cell-mediated cytotoxicity. NK cells express a potent Fcγ receptor, CD16a (FcγRIIIA), whose binding to IgG Fc ligand triggers the release of cytolytic enzymes, such as granzyme B (GZMB)[46]. We observed significantly higher expression levels of *GZMB* and *FCGR3A* (which

encodes CD16a) in tumors with more CSR events (Fig. 6a), thus suggesting elevated NK cytotoxicity in tumors with B cell response. Among the four subclasses of IgG antibodies, the strongest binding to CD16a were IgG1 and IgG3 (ref. [47]). Consistently with this observation, we observed positive correlations between the IgG3–1 switch level and *FCGR3A* expression across almost all tumor types (Fig. 6b). One possible explanation for these associations and the prevalent IgG3–1 sCSR events is that tumor-infiltrating B cells and NK cells work synergistically in the microenvironment to exert antitumor responses.

Tumors under ADCC may develop evasive mechanisms against NK-cell attack. An established evasion pathway is through the shedding of endogenous MICA and MICB[48]. Metalloproteinase catalyzes the shedding of the MIC ectodomain, thereby producing soluble MIC, which binds the activation receptor NKG2D on NK cells[49] and results in internalization of NKG2D and decreased NK-cell activity[50]. Therefore, MIC shedding has been established as a mechanism to evade NK cell immunosurveillance[50] (Fig. 6c). Interestingly, *MICA* and *MICB* are in the same loci, and 20% of TCGA tumors showed *MICA* and *MICB* amplification (Fig. 6c), a frequency significantly higher than that in the germline variances (Supplementary Fig. 6a). *MICA*-amplified tumors had significantly higher expression of *ADAM17* and *MMP14* (Supplementary Fig. 6b), two metalloproteinases known to catalyze MIC shedding[48,51], as well as higher IgG1/3 B cell levels (Supplementary Fig. 6c). Furthermore, whereas *MICA* amplification was generally associated with poorer outcomes in cancer patients (Supplementary Fig. 6d–f), its clinical relevance further depended on the level of IgG1/3 B cells. Specifically, in tumors with *MICA* amplification, the presence of high levels of IgG1/3 B cells was associated with significantly better survival in breast cancer and melanoma (FDR <0.1, Fig. 6d). In contrast, the IgG1/3 level did not influence survival for tumors without *MICA* amplification. These results suggest complex interactions between B cell-mediated immune response and tumor ADCC-pathway defects[52].

Discussion

High levels of tumor-infiltrating B cells have been observed in many human cancers[6]. However, the functional effects of infiltrating B cells have been inconsistent in previous studies[5]. Future development of B cell-based therapies requires an improved understanding of tumor interactions with infiltrating B cells. In this study, we analyzed large cohorts of TCGA tumor RNA-seq data across 32 cancer types and generated a large data set of tumor-infiltrating B cell IgH hypervariable sequences. Using SHMs as lineage markers, we identified widespread B cell clonal expansions and immunoglobulin subclass switches. The prevalent IgG3–1 sCSR may reflect the selective pressure of B cells involved in ADCC.

Spontaneous antibody-dependent cell-mediated cytotoxicity in tumors is poorly characterized because of the challenge in detecting ADCC in vivo. In this study, we observed frequent *MICA* and *MICB* amplifications coupled with increased expression of metalloproteinases, thus suggesting the existence of soluble MIC in the tumor microenvironment. Together with *MICA* amplification's negative clinical effects and its notable co-occurrence with IgG1 and IgG3 B cells, we made orthogonal observations that NK-cell-mediated ADCC might be functional in the antitumor response. Complex interactions were observed between *MICA* amplification and the IgG-subclass switch (Fig. 6d). It is possible that the presence of the IgG subclass switch in *MICA*-amplified tumors is an indicator of effective B cell–mediated immune attack. Tumors with high B cell activity but an intact ADCC pathway might have compromised B cell function or might have developed other evasive mechanisms. An alternative explanation is that *MICA* amplification and MIC shedding in tumors induces the clonal expansion of B cells producing the anti-MICA IgG autoantibodies[53]. In this scenario, the anti-MICA IgG autoantibodies might prevent NKG2D internalization caused by MIC shedding, thus allowing antitumor ADCC and leading to survival benefits. Autoantibodies with similar functions have recently been reported in specific tumor types. Some patients with breast cancer can produce autoantibodies against the overexpressed oncogene *HER2* and gain survival benefits[54]. In gastric cancer, tumor-infiltrating B cells produce antibodies targeting sulfated glycosaminoglycans on the cellular surface[22]. Therefore, it is possible that overexpression of some of the membrane proteins, abnormal glycosylation, or lipoproteins are potential autoantibody targets. Future work is needed to elucidate the role of tumor-infiltrating IgG1-G3-expressing B cells in the context of immunotherapy.

In summary, we performed comprehensive pancancer analyses on tumor-infiltrating B cell repertoires. Our study gained statistical power by using large human cancer cohorts but is still limited by the heterogeneous treatments that different cancer patients received. Our observation regarding SHM rate might indicate B cell–receptor affinity maturation during tumor development, but this result awaits future experimental validation. Another limitation of this work is that, because of the use of bulk tissue data, it was impossible to distinguish different subtypes of infiltrating B cells, although this limitation did not affect our detection of B cell clonal expansion and subclass switches. The immune evasive mechanisms reported in our study may advance understanding of the complex interactions between tumor and infiltrating B cells. As the cost of tumor RNA-seq continues to decrease, our approach could be adopted to examine the ever-growing volume of tumor RNA-seq data to discover and refine hypotheses on tumor humoral immunity. The findings from this work have potential clinical utility in B cell–related cancer immunotherapies.

URLs. Cancer Genomics Hub, https://cghub.ucsc.edu/; TCGA data portal, https://portal.gdc.cancer.gov/legacy-archive/; GDAC Firehose, https://gdac.broadinstitute.org/; simNGS, http://www.ebi.ac.uk/goldman-srv/simNGS/.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41588-018-0339-x.

Received: 8 September 2017; Accepted: 18 December 2018;
Published online: 11 February 2019

References

1. Thorsson, V. et al. The immune landscape of cancer. *Immunity* **48**, 812–830.e14 (2018).
2. Nutt, S. L., Hodgkin, P. D., Tarlinton, D. M. & Corcoran, L. M. The generation of antibody-secreting plasma cells. *Nat. Rev. Immunol.* **15**, 160–171 (2015).
3. Raposo, G. et al. B lymphocytes secrete antigen-presenting vesicles. *J. Exp. Med.* **183**, 1161–1172 (1996).
4. Hagn, M. et al. Human B cells differentiate into granzyme B-secreting cytotoxic B lymphocytes upon incomplete T-cell help. *Immunol. Cell Biol.* **90**, 457–467 (2012).
5. Nelson, B. H. CD20$^+$ B cells: the other tumor-infiltrating lymphocytes. *J. Immunol.* **185**, 4977–4982 (2010).
6. Linnebacher, M. & Maletzki, C. Tumor-infiltrating B cells: the ignored players in tumor immunology. *Oncoimmunology* **1**, 1186–1188 (2012).
7. Nielsen, J. S. & Nelson, B. H. Tumor-infiltrating B cells and T cells: working together to promote patient survival. *Oncoimmunology* **1**, 1623–1625 (2012).
8. Al-Shibli, K. I. et al. Prognostic effect of epithelial and stromal lymphocyte infiltration in non-small cell lung cancer. *Clin. Cancer Res.* **14**, 5220–5227 (2008).
9. Coronella-Wood, J. A. & Hersh, E. M. Naturally occurring B-cell responses to breast cancer. *Cancer Immunol. Immunother.* **52**, 715–738 (2003).

10. Milne, K. et al. Systematic analysis of immune infiltrates in high-grade serous ovarian cancer reveals CD20, FoxP3 and TIA-1 as positive prognostic factors. *PLoS ONE* **4**, e6412 (2009).
11. Liu, X. S. & Mardis, E. R. Applications of immunogenomics to cancer. *Cell* **168**, 600–612 (2017).
12. Xu, J. L. & Davis, M. M. Diversity in the CDR3 region of V(H) is sufficient for most antibody specificities. *Immunity* **13**, 37–45 (2000).
13. Stavnezer, J., Guikema, J. E. & Schrader, C. E. Mechanism and regulation of class switch recombination. *Annu. Rev. Immunol.* **26**, 261–292 (2008).
14. Li, Z., Woo, C. J., Iglesias-Ussel, M. D., Ronai, D. & Scharff, M. D. The generation of antibody diversity through somatic hypermutation and class switch recombination. *Genes Dev.* **18**, 1–11 (2004).
15. Gadala-Maria, D., Yaari, G., Uduman, M. & Kleinstein, S. H. Automated analysis of high-throughput B-cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles. *Proc. Natl Acad. Sci. USA* **112**, E862–E870 (2015).
16. Lin, S. G. et al. Highly sensitive and unbiased approach for elucidating antibody repertoires. *Proc. Natl Acad. Sci. USA* **113**, 7846–7851 (2016).
17. Yaari, G. & Kleinstein, S. H. Practical guidelines for B-cell receptor repertoire sequencing analysis. *Genome Med.* **7**, 121 (2015).
18. Blachly, J. S. et al. Immunoglobulin transcript sequence and somatic hypermutation computation from unselected RNA-seq reads in chronic lymphocytic leukemia. *Proc. Natl Acad. Sci. USA* **112**, 4322–4327 (2015).
19. Mose, L. E. et al. Assembly-based inference of B-cell receptor repertoires from short read RNA sequencing data with V'DJer. *Bioinformatics* **32**, 3729–3734 (2016).
20. Liu, S. et al. Direct measurement of B-cell receptor repertoire's composition and variation in systemic lupus erythematosus. *Genes Immun.* **18**, 22–27 (2017).
21. Kurtz, D. M. et al. Noninvasive monitoring of diffuse large B-cell lymphoma by immunoglobulin high-throughput sequencing. *Blood* **125**, 3679–3687 (2015).
22. Katoh, H. et al. Immunogenetic profiling for gastric cancers identifies sulfated glycosaminoglycans as major and functional B cell antigens in human malignancies. *Cell Rep.* **20**, 1073–1087 (2017).
23. Li, B. et al. Landscape of tumor-infiltrating T cell repertoire of human cancers. *Nat. Genet.* **48**, 725–732 (2016).
24. Li, B. et al. Ultrasensitive detection of TCR hypervariable region in solid-tissue RNA-seq data. *Nat. Genet.* **49**, 482–483 (2017).
25. Shi, B. et al. Comparative analysis of human and mouse immunoglobulin variable heavy regions from IMGT/LIGM-DB with IMGT/HighV-QUEST. *Theor. Biol. Med. Model.* **11**, 30 (2014).
26. Lefranc, M. P. et al. IMGT(R), the international ImMunoGeneTics information system(R) 25 years on. *Nucleic Acids Res.* **43**, D413–D422 (2015).
27. Schroeder, H. W. Jr & Cavacini, L. Structure and function of immunoglobulins. *J. Allergy Clin. Immunol.* **125**, S41–S52 (2010).
28. Di Noia, J. M. & Neuberger, M. S. Molecular mechanisms of antibody somatic hypermutation. *Annu. Rev. Biochem.* **76**, 1–22 (2007).
29. Rogozin, I. B. & Kolchanov, N. A. Somatic hypermutagenesis in immunoglobulin genes. II. Influence of neighbouring base sequences on mutagenesis. *Biochim. Biophys. Acta* **1171**, 11–18 (1992).
30. Keim, C., Kazadi, D., Rothschild, G. & Basu, U. Regulation of AID, the B-cell genome mutator. *Genes Dev.* **27**, 1–17 (2013).
31. Krishnamurty, A. T. et al. Somatically hypermutated *Plasmodium*-specific IgM(+) memory B cells are rapid, plastic, early responders upon malaria rechallenge. *Immunity* **45**, 402–414 (2016).
32. Kitaura, K. et al. Different somatic hypermutation levels among antibody subclasses disclosed by a new next-generation sequencing-based antibody repertoire analysis. *Front. Immunol.* **8**, 389 (2017).
33. Odegard, V. H. & Schatz, D. G. Targeting of somatic hypermutation. *Nat. Rev. Immunol.* **6**, 573–583 (2006).
34. LeBien, T. W. & Tedder, T. F. B lymphocytes: how they develop and function. *Blood* **112**, 1570–1580 (2008).
35. Jackson, K. J., Wang, Y. & Collins, A. M. Human immunoglobulin classes and subclasses show variability in VDJ gene mutation levels. *Immunol. Cell Biol.* **92**, 729–733 (2014).
36. Li, B. et al. Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. *Genome Biol.* **17**, 174 (2016).
37. Geissler, K. et al. Immune signature of tumor infiltrating immune cells in renal cancer. *Oncoimmunology* **4**, e985082 (2015).
38. William, J., Euler, C., Christensen, S. & Shlomchik, M. J. Evolution of autoantibody responses via somatic hypermutation outside of germinal centers. *Science* **297**, 2066–2070 (2002).
39. Hwang, J. K., Alt, F. W. & Yeap, L. S. Related mechanisms of antibody somatic hypermutation and class switch recombination. *Microbiol. Spectr.* **3**, MDNA3-0037-2014 (2015).
40. Schreiber, R. D., Old, L. J. & Smyth, M. J. Cancer immunoediting: integrating immunity's roles in cancer suppression and promotion. *Science* **331**, 1565–1570 (2011).
41. Shukla, S. A. et al. Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. *Nat. Biotechnol.* **33**, 1152–1158 (2015).
42. Zaretsky, J. M. et al. Mutations associated with acquired resistance to PD-1 blockade in melanoma. *N. Engl. J. Med.* **375**, 819–829 (2016).
43. Iannello, A. & Ahmad, A. Role of antibody-dependent cell-mediated cytotoxicity in the efficacy of therapeutic anti-cancer monoclonal antibodies. *Cancer Metastasis Rev.* **24**, 487–499 (2005).
44. Nimmerjahn, F. & Ravetch, J. V. Fcgamma receptors as regulators of immune responses. *Nat. Rev. Immunol.* **8**, 34–47 (2008).
45. Waldhauer, I. & Steinle, A. NK cells and cancer immunosurveillance. *Oncogene* **27**, 5932–5943 (2008).
46. Smyth, M. J. et al. Activation of NK cell cytotoxicity. *Mol. Immunol.* **42**, 501–510 (2005).
47. Bruhns, P. et al. Specificity and affinity of human Fcgamma receptors and their polymorphic variants for human IgG subclasses. *Blood* **113**, 3716–3725 (2009).
48. Waldhauer, I. et al. Tumor-associated MICA is shed by ADAM proteases. *Cancer Res.* **68**, 6368–6376 (2008).
49. Raulet, D. H. Roles of the NKG2D immunoreceptor and its ligands. *Nat. Rev. Immunol.* **3**, 781–790 (2003).
50. Doubrovina, E. S. et al. Evasion from NK cell immunity by MHC class I chain-related molecules expressing colon adenocarcinoma. *J. Immunol.* **171**, 6891–6899 (2003).
51. Liu, G., Atteridge, C. L., Wang, X., Lundgren, A. D. & Wu, J. D. The membrane type matrix metalloproteinase MMP14 mediates constitutive shedding of MHC class I chain-related molecule A independent of A disintegrin and metalloproteinases. *J. Immunol.* **184**, 3346–3350 (2010).
52. Zhang, J., Basher, F. & Wu, J. D. NKG2D ligands in tumor immunity: two sides of a coin. *Front. Immunol.* **6**, 97 (2015).
53. Jinushi, M., Hodi, F. S. & Dranoff, G. Therapy-induced antibodies to MHC class I chain-related protein A antagonize immune suppression and stimulate antitumor cytotoxicity. *Proc. Natl Acad. Sci. USA* **103**, 9190–9195 (2006).
54. Tabuchi, Y. et al. Protective effect of naturally occurring anti-HER2 autoantibodies on breast cancer. *Breast Cancer Res. Treat.* **157**, 55–63 (2016).

Acknowledgements

We thank F. Alt, M. Shipp, and G. Freeman for helpful discussions during manuscript preparation, and T. Wu for suggesting the Single Instruction Multiple Data (SIMD) acceleration. We also acknowledge the following funding sources for supporting this work: NCI, grants U01 CA226196 (X.S.L.) and U24 CA224316 (X.S.L.), Chinese Scholarship Council Funding (Jian Zhang), CPRIT RR170079 (B.L.), the Breast Cancer Research Foundation (X.S.L.), and the National Natural Science Foundation of China, 81702701 (T.L.).

Author contributions

B.L. conceived this project, modified TRUST codes to assemble B cell CDR3 sequences, and analyzed TCGA data. X.H. and Jian Zhang accelerated the TRUST codes, performed the validation analyses, and tested the clustering method. J.W., J. Fu, and T.L. helped with figure generation and data analyses, and B.W. and J. Fan helped to generate intermediate data. P.Z. provided tumor samples, and S.G., F.Z., and X.Z. performed the BCR-seq and RNA-seq validation experiments. P.J., X.Y., Jing Zhang, M.C.C., K.W.W., and N.H. provided expertise in immune evasive mechanisms and data analysis. J.S.L. and X.S.L. supervised the study and wrote the manuscript with X.H. and B.L. All coauthors contributed to manuscript preparation and research progress discussion.

Competing interests

X.S.L. is a cofounder and board member of GV20 Oncotherapy, SAB of 3DMed Care, consultant for Genentech, and stock holder of BMY, TMO, WBA, ABT, ABBV, and JNJ. J.S.L. is a cofounder and board member of Beijing Neoantigen Biotechnology Co. Ltd, and shareholder of BGNE, DVAX, CELG, NKTR, CRSP, and EDIT. K.W.W. serves on the SAB of Nextech, TCR2, and T-scan. He is also a consultant for Novartis. His laboratory has sponsored research agreements with Novartis, BMS, and Astellas. N.H. is a founder and SAB member of Neon Therapeutics and SAB member of IFM Therapeutics.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0339-x.

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

Correspondence and requests for materials should be addressed to J.S.L., B.L. or X.S.L.

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

Methods

TRUST-method modification and performance evaluation. The TCR CDR3 de novo assembly workflows using single- or paired-end unselected RNA-seq data have been previously described[24]. In that work, we used a predefined list of motifs to search for the TCR variable or joining gene conserved regions. To allow TRUST to analyze the B cell IgH CDR3 region, we made the following modifications: (i) we downloaded IgH variable and joining DNA and amino acid reference sequences in fasta format from the International Immunogenetics Information System[26]; (ii) we included genomics locations (hg19) of 124 IGHV genes, 9 IGHJ genes, and 9 IGH constant genes in the TRUST search space to extract the mapped reads from these loci; (iii) we included an annotation step for the possible V, J, and C gene segments in the assembled CDR3 sequences; (iv) we added -B option in the source code to instruct TRUST to analyze BCR CDR3 regions.

We systematically evaluated the performance of modified TRUST by using in silico simulations. Specifically, 5,000 random full-length IgH sequences were generated by following the biological processes of VDJ recombination as previously described[24]. Throughout the IgH DNA sequence, we randomly added single-nucleotide changes at a rate of 0.05 per base to mimic SHMs. We then applied simNGS to construct in silico DNA fragment libraries and sample paired-end short reads with statistical-error models matching the Illumina sequencers. We fixed the read length to be 50 nt (from 200-nt fragments) and simulated libraries at different coverages, including 0.1×, 0.5×, and 1×. According to our previous estimation[23], 0.1× coverage corresponds to an RNA-seq data library size of 500 million reads, with 1% B cell infiltration in the tumor tissue. Therefore, a library size higher than 1× is unrealistically high at the current sequencing cost.

For each parameter setting, we repeated 20 simulations and aligned the fastq reads to the hg19 reference genome by using Tophat2[25], merging aligned reads and unmapped reads into one BAM file. Modified TRUST was then applied to each BAM file to assemble the CDR3 sequences. The results were then compared with the original 5,000 reference sequences to estimate precision and sensitivity. Partial CDR3 sequences shorter than 15 nt were excluded from downstream analysis. Because simNGS introduces sequencing errors, for each CDR3 assembly, we allowed one mismatch in its DNA sequence to the simulated truth.

We also performed RNA-seq and BCR-seq on tumor samples from six patients with early lung cancer to validate TRUST performance. Tumor samples were collected from Shanghai Pulmonary Hospital and procedures for this study were approved by the Ethics Committee of Shanghai Pulmonary Hospital. All patients provided written informed consent for sample collection and data analyses. For paired RNA-seq and BCR-seq data, we allowed one mismatch in CDR3s to validate TRUST-assembled clones. In both in silico simulation and experimental validation, precision was defined as the fraction of TRUST-called CDR3s validated by BCR-seq, and sensitivity was defined as the fraction of BCR-seq CDR3s called by TRUST. The above criteria were also applied in our previous analysis to evaluate the performance of TRUST on calling T cell receptor CDR3 regions[24]. We also evaluated the TRUST-reported immunoglobulin isotypes by using BCR-seq for matched CDR3s. To evaluate immunoglobulin isotype inference, we focused on the CDR3s called by both BCR-seq and TRUST from RNA-seq, then defined precision as the fraction of TRUST-called isotypes validated by BCR-seq and sensitivity as the fraction of BCR-seq isotypes called by TRUST.

Data preparation and preprocessing. RNA-seq data of 10,818 samples in BAM format were downloaded from the Cancer Genomics Hub in May 2016. The RNA-seq reads were previously aligned to the hg19 human reference genome with MapSplice[56]. RSEM gene expression data, GISTIC annotations of copy number alterations, level 3 somatic mutation profiles, and clinical annotations were downloaded from GDAC Firehose. The GISTIC2.0 (ref. [57]) annotation of CNAs had been previously binned into −2, −1, 0, 1, and 2, representing total copy loss, hemizygous deletion, euploidy, copy number gain, and high fold amplification. In the analysis of somatic copy number alterations (SCNAs), we grouped the −2 and −1 segments together and referred to them as deletions, and grouped 1 and 2 together and referred to them as amplifications. Tumor purity of 9,849 samples was obtained from our previous work[36].

Modified TRUST was applied to all of the RNA-seq samples with the following command:

```
python TRUST.py -f sample_name.bam -a -B
```

Samples with zero assemblies were excluded from downstream analysis, and 9,025 samples remained. For each sample, we parsed the fasta file information lines reported from TRUST analysis, keeping the following fields: TCGA identifier, disease abbreviation, tissue type, CDR3 amino acid sequence, estimated sample library size, CDR3 DNA sequence, and assigned immunoglobulin constant genes. Disease abbreviations followed TCGA naming conventions. Tissue type includes primary tumor (TP), adjacent normal (NT), metastatic tumor (TM), and recurrent tumor (TR). The IgH CDR3 region was defined as the region within the last C in the variable gene sequence and the W in the joining gene motif WGXG, with both C and W excluded. TRUST reports variable, joining, and constant genes on the basis of mapped reads linked to the CDR3 assembly and the similarity to the germline genes[24], but only CDR3 DNA sequences were used to uniquely define a B cell clone to avoid the complication of mapping to unknown variable or joining gene alleles.

Identification of B cell clusters and CSR. In this work, we developed a computationally efficient approach to identify B cell CDR3 clusters on the basis of local sequence alignment. For each sample, we first extracted all of the unique complete CDR3 sequences according to the co-occurrence of the variable gene motif YYC and joining gene motif WGXG. For samples that did not contain any complete sequence, no cluster was reported. For each complete CDR3 sequence, we extracted an octamer starting from the first position in the CDR3 as a motif. For each unique motif, we collected all of the CDR3 amino acid sequences, both partial and complete, containing the motif. When searching for matches, we allowed a one-letter mismatch. For example, motif RDMW**R**VGW was considered the same as RDMW**I**VGW. This approach provided the flexibility of detecting amino acid changes incurred from nonsilent mutations yet maintained low computational complexity. The motif-containing sequences constituted a B cell cluster. Clusters with fewer than three sequences were discarded.

We chose the octamer and starting position 1 on the basis of a systematic search of a wide range of parameters. The goal of the clustering optimization was to identify more clusters while minimizing the number of incorrectly clustered pairs. To estimate these two metrics, we randomly selected two samples from different patients 500 times, and, in each run, we clustered all of the CDR3s and computed FDR and the number of clusters. FDR is the fraction of clustered CDR3 pairs from two different patients over all of the pairs in clusters, because the possibility of having the same CDR3 in unrelated individuals is extremely low (10^{-3}–10^{-5})[58]. The cluster ratio is the number of clusters normalized by the median value of tested cases. After testing *k*-mer sizes from 6 to 15 and start positions from 1 to 9, we visualized the median value of FDR and the cluster ratio in heat maps (Supplementary Fig. 3a,b). To balance the contribution of FDR and the number of clusters, we computed the harmonic average of 1-FDR and cluster ratio as the final score for selecting the parameters (Supplementary Fig. 3c). The best solution was clustering octamers starting from the first position in the CDR3 region, thus yielding an FDR of 0.007.

For each cluster, we performed multiple local sequence alignment with ClustalW[59] implemented in the R package msa[60]. Aligned sequences had the same length, with gaps filled by the character '-'. To study the relationship between different sequences in a cluster, we calculated the number of mismatches between each pair of sequences and used this number as a distance measure. A mismatch was counted when neither sequence was '-', and the two sequences differed at this base. The pairwise distance matrix was then used to plot the neighbor-joining tree in Fig. 1c, which was implemented in the R package ape[61].

In the isotype analysis, we took advantage of paired-end sequencing data and assigned different IgH isotypes according to the mapping locations. For example, if one mate from a read pair was unmapped and contained CDR3 sequence, whereas the other mate was properly mapped to the IgM region, then the CDR3 assembly was assigned the IgM isotype. If the mate read was not mapped to a known constant region, we tried to infer the isotype from the sequence after the CDR3 region. If we were still unable to recover the isotype, the CDR3 assembly was assigned as an unknown isotype and excluded from the class-switch analysis. Sequence assembly errors can sometimes join reads from different transcripts into one CDR3 assembly and result in more than one immunoglobulin isotype assignment for a single CDR3 call. We excluded these sequences in our evaluation of class-switch events to decrease false-positive calls. Each CSR event identified in this work was supported by at least two CDR3 assemblies unambiguously assigned to different immunoglobulin isotypes. We normalized the number of CSR events by dividing the number of unique CDR3s, and samples with less than ten unique CDR3s were excluded from the analyses.

Analysis of SHMs. SHMs were defined as mismatches in B cell clusters. We used the BayesNMF function in the SignatureAnalyzer package to decompose the mutation count matrix[62]. We considered each mutation triplet, for example ACT to ATT, corresponding to one type of SHM in which the middle base was mutated from C to T. To infer the mutation direction, we used only CDR3 pairs with different immunoglobulin isotypes, in which we considered the CDR3 with a closer constant region on the genome to be the original sequence. Although this assumption would be violated if both CDR3s were mutated from a common ancestor along different paths, we did observe an enrichment in the correct direction after aggregating the SHMs. We counted the number of SHMs for each 96-triplet and for each immunoglobulin isotype of the original CDR3 to construct the mutation count matrix. We decomposed the matrix into two dimensions and selected the mutation signature matrix with the highest likelihood out of 1,000 separate optimizations.

To avoid overestimation of the SHM rate as a result of the aggregated mutations during B cell clonal expansion, we counted only mutations for two sequences with only one nucleotide mismatch. Then SHM rate per sample was the SHM count divided by the total number of assembled CDR3 bases, thereby avoiding the bias of unknown mutations outside partial CDR3 assembles. The high-SHM group was defined as samples with SHM rate greater than or equal to the median SHM rate, and the remaining samples were assigned to the low-SHM group.

The effect of sequencing errors on the SHM estimation should be low. Approximately 96% of the TCGA RNA-seq BAM files in this study had an average quality score greater than 30. According to the manufacturer's definition, a quality score of 30 corresponds to a 0.1% error rate, which is a magnitude lower than the ~5% SHM rate that we observed in B cell CDR3s.

Statistical analysis. Partial Spearman's rank correlation was used to verify the associations between gene expression and B cell features, including SHM rate and the number of B cell clusters normalized by the total number of unique CDR3s. Correlation coefficients and FDR-adjusted *P* values of compared values, controlling for the tumor purity, are shown in heat maps for tumor types with at least 100 samples (Figs. 3a and 6b). The fold changes between tumor and adjacent normal tissues (Fig. 4) were calculated by the average values in two groups and tested with Student's *t* test. We excluded the tumor types with <100 tumor samples or <10 normal samples. Survival analyses were visualized with Kaplan–Meier curves, and the statistical significance was estimated with Cox proportional hazard regression corrected for patient age and tumor purity. Survival analyses, including those shown in Figs. 5c and 6c, were conducted by using all of the samples with BCR sequence, tumor purity, and clinical annotation data available. Per-cancer-type survival analyses were performed only for cancer types with at least ten patients in all compared groups. Key findings of interactions between immune evasive SCNAs and B cell sCSRs were confirmed with Cox regression, corrected for age at diagnosis and tumor purity on samples. All survival analyses were truncated at 5,000 d, to exclude potential nonrelevant long-term survival or death incidence. Other sample comparisons, including metalloproteinase expression between *MICA*-amplified and unamplified tumors, and IgG1/3 levels between tumors with evasive SCNAs and those without, were performed with Wilcoxon rank-sum tests. We reported two-sided *P* values for Student's *t* tests and Wilcoxon rank-sum tests. Multiple hypothesis correction is performed, and *P* values adjusted through the Benjamini–Hochberg procedure are reported. All statistical tests and survival curves were implemented with R statistical programming language[63].

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

Code availability

Modified TRUST applicable to BCR IgH CDR3 assemblies and supporting files are available at https://bitbucket.org/liulab/trust/. Code for performance evaluation has been deposited at https://bitbucket.org/liulab/ng-bcr-validate/.

Data availability

The data sets generated during the study are available from FireCloud and with dbGap permission to retrieve restricted TCGA data. The RNA-seq data set generated for validating TRUST performance is available in the SRA repository (PRJNA492301), and the matched iRepertoire data are available at https://bitbucket.org/liulab/ng-bcr-validate/src/master/iRep/.

References

55. Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
56. Wang, K. et al. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.* **38**, e178 (2010).
57. Mermel, C. H. et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).
58. DeWitt, W. S. et al. A public database of memory and naive B-cell receptor sequences. *PLoS ONE* **11**, e0160853 (2016).
59. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680 (1994).
60. Bodenhofer, U., Bonatesta, E., Horejs-Kainrath, C. & Hochreiter, S. msa: an R package for multiple sequence alignment. *Bioinformatics* **31**, 3997–3999 (2015).
61. Paradis, E., Claude, J. & Strimmer, K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290 (2004).
62. Kim, J. et al. Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nat. Genet.* **48**, 600–606 (2016).
63. R v. 3.5.1 (The R Project for Statistical Computing, 2018).

nature research

Corresponding author(s): X Shirley Liu, Bo Li, Jun S Liu

Reporting Summary

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

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. 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
☐	☒	An indication of 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 statistics 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
☒	☐	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	The Cancer Genome Atlas (TCGA) RNA-seq data in BAM format were downloaded from Cancer Genomics Hub in May 2016.
Data analysis	RNA-seq reads were mapped by STAR (v2.5). TRUST applicable to BCR IgH CDR3 assemblies are available at https://bitbucket.org/liulab/trust/. BCR data analyses were performed in R (v3.3) with these packages: ggplot2 (v3.0.0), ggpubr (v0.1.8), plotrix (v3.7-4), seqinr (v3.4-5), ape (v5.2), msa (v1.11.0), scales (v1.0.0), reshape2 (v1.4.3), fields (v9.6), tidyr (v0.8.1), dplyr (v0.7.7), and ppcor (v1.1).

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/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

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

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

Modified TRUST applicable to BCR IgH CDR3 assemblies and supporting files are available at https://bitbucket.org/liulab/trust/. The datasets generated during the

study are available from the corresponding author on request and with permission of The Cancer Genome Atlas (TCGA) access. The RNA-seq dataset generated for validating TRUST performance are available in the SRA repository (PRJNA492301), and source codes are deposited to https://bitbucket.org/liulab/ng-bcr-validate.

Field-specific reporting

Please select 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/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

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

Sample size	All available The Cancer Genome Atlas (TCGA) RNA-seq data were included in the study. The sample size for the pan-cancer analyses is sufficient because there are more than 9,000 samples with BCR data. Cancer types with small sample sizes were ignored in the tumor-type specific analyses.
Data exclusions	RNA-seq samples were excluded if no BCR CDR3s were found.
Replication	Not relevant because all The Cancer Genome Atlas (TCGA) data were used.
Randomization	Not relevant because all samples were included in the study.
Blinding	Not relevant because The Cancer Genome Atlas (TCGA) data were collected by others.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants

Methods

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

Human research participants

Policy information about studies involving human research participants

Population characteristics	Patients with early lung cancer in China
Recruitment	No selection