








# Predictability of B cell clonal persistence and immunosurveillance in breast cancer

Received: 20 February 2024

Accepted: 15 March 2024

Published online: 2 May 2024

 Check for updates

Stephen-John Sammut <sup>1,2</sup>✉, Jacob D. Galson<sup>3</sup>, Ralph Minter<sup>3</sup>, Bo Sun <sup>4,5</sup>, Suet-Feung Chin <sup>6</sup>, Leticia De Mattos-Arruda <sup>7,8</sup>, Donna K. Finch<sup>3</sup>, Sebastian Schätzle<sup>3</sup>, Jorge Dias<sup>3</sup>, Oscar M. Rueda<sup>9</sup>, Joan Seoane <sup>10</sup>, Jane Osbourn<sup>3</sup>, Carlos Caldas <sup>11</sup>✉ & Rachael J. M. Bashford-Rogers <sup>4,12,13</sup>✉

B cells and T cells are important components of the adaptive immune system and mediate anticancer immunity. The T cell landscape in cancer is well characterized, but the contribution of B cells to anticancer immunosurveillance is less well explored. Here we show an integrative analysis of the B cell and T cell receptor repertoire from individuals with metastatic breast cancer and individuals with early breast cancer during neoadjuvant therapy. Using immune receptor, RNA and whole-exome sequencing, we show that both B cell and T cell responses seem to coevolve with the metastatic cancer genomes and mirror tumor mutational and neoantigen architecture. B cell clones associated with metastatic immunosurveillance and temporal persistence were more expanded and distinct from site-specific clones. B cell clonal immunosurveillance and temporal persistence are predictable from the clonal structure, with higher-centrality B cell antigen receptors more likely to be detected across multiple metastases or across time. This predictability was generalizable across other immune-mediated disorders. This work lays a foundation for prioritizing antibody sequences for therapeutic targeting in cancer.

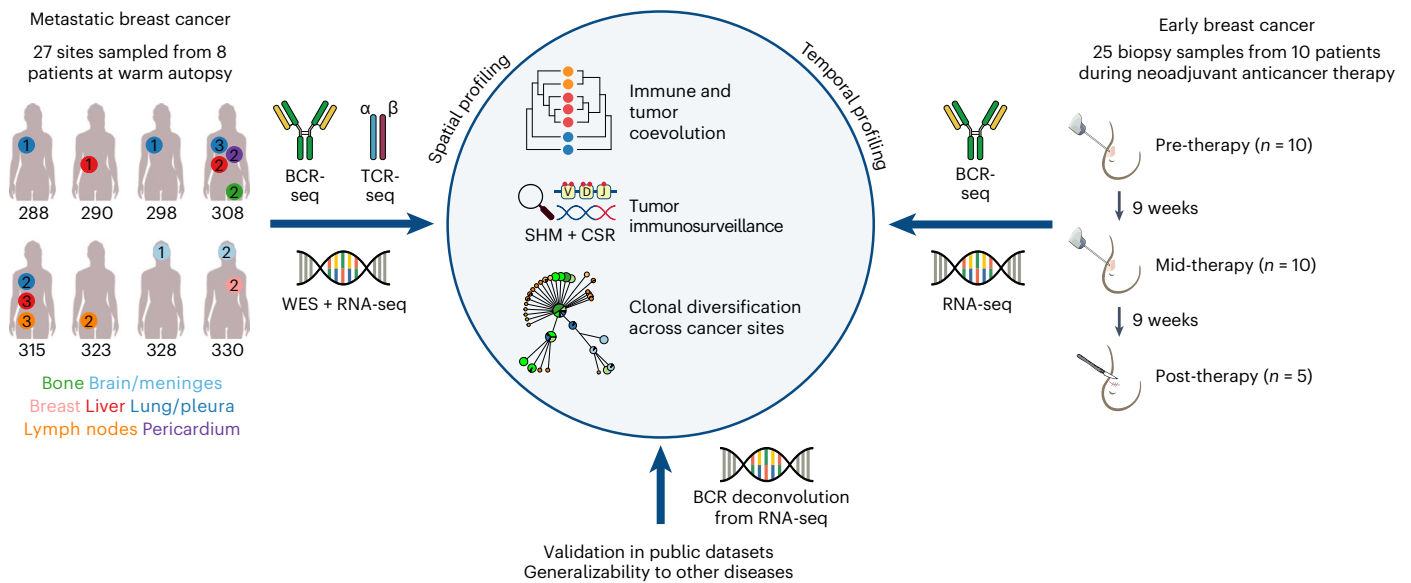
The mechanisms by which tumors evade immune control are critical to developing better targeted immunotherapies. B and T cells play an important role in anticancer immunity<sup>1,2</sup>. However, while the T cell immune response to cancer and its therapeutic manipulation is well characterized, the B cell contribution to antitumor immunity remains less well studied.

B cells contribute to antitumor responses by binding tumor antigens via their B cell antigen receptor (BCR) and presenting these to follicular helper T cells, by antibody secretion and by cytokine signaling

to other cells. Tumor-infiltrating B cells are associated with improved clinical outcomes<sup>3–6</sup> and response to chemotherapy and immunotherapy<sup>7,8</sup>, and the persistence of plasma antitumor antibodies and tumor-associated tertiary lymphoid structures (TLSs) associate with improved survival<sup>4,9</sup>.

B and T cell clones selectively expand following antigen recognition by their BCR and T cell antigen receptor (TCR), respectively. These receptors are generated through DNA recombination and have the potential to recognize a vast array of antigens. On encountering

<sup>1</sup>Breast Cancer Now Toby Robins Research Centre, The Institute of Cancer Research, London, UK. <sup>2</sup>The Royal Marsden Hospital NHS Foundation Trust, London, UK. <sup>3</sup>Alchemab Therapeutics, Whittlesford, UK. <sup>4</sup>Wellcome Centre for Human Genetics, Oxford, UK. <sup>5</sup>Nuffield Department of Clinical Neuroscience, University of Oxford, Oxford, UK. <sup>6</sup>Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK. <sup>7</sup>IrsiCaixa, Germans Trias i Pujol University Hospital, Badalona, Spain. <sup>8</sup>Germans Trias i Pujol Research Institute (IGTP), Badalona, Spain. <sup>9</sup>MRC Biostatistics Unit, University of Cambridge, Cambridge, UK. <sup>10</sup>Vall d'Hebron Institute of Oncology (VHIO), Vall d'Hebron University Hospital, Institutió Catalana de Recerca i Estudis Avançats (ICREA), Universitat Autònoma de Barcelona (UAB), CIBERONC, Barcelona, Spain. <sup>11</sup>School of Clinical Medicine, University of Cambridge, Cambridge, UK. <sup>12</sup>Department of Biochemistry, University of Oxford, Oxford, UK. <sup>13</sup>Oxford Cancer Centre, Oxford, UK. ✉e-mail: [stephen-john.sammut@icr.ac.uk](mailto:stephen-john.sammut@icr.ac.uk); [cc234@cam.ac.uk](mailto:cc234@cam.ac.uk); [rachael.bashford-rogers@bioch.ox.ac.uk](mailto:rachael.bashford-rogers@bioch.ox.ac.uk)



**Fig. 1 | Description of breast cancer cohorts and overview of study design.** Schematic of the sampling, data collection and analysis of the breast cancer cohorts in this study. Female silhouette is from the public domain diagrams of the human body at [https://commons.wikimedia.org/wiki/Human\\_body\\_diagrams](https://commons.wikimedia.org/wiki/Human_body_diagrams). WES, whole-exome sequencing.

antigen, B cells can be stimulated to proliferate and further diversify their BCR sequences via class switching and somatic hypermutation (SHM) resulting in high-affinity B cell responses<sup>10</sup>. Previous studies in breast cancer have shown significant heterogeneity in tumor-infiltrating B cell subpopulations, significant levels of SHM and clonal expansion, and local differentiation of infiltrated memory B cells<sup>11,12</sup>. Indeed, some studies have shown that tumor-infiltrating B cells can have antitumor BCR specificities, such as anti-HER2 autoantibodies in breast cancer<sup>13,14</sup>.

The immune system can monitor, recognize and destroy transformed cells or pathogens, a concept termed immunosurveillance<sup>15</sup>. Immunosurveillance is responsible for shaping the tumor molecular landscape and is key to the effectiveness of anticancer therapies. However, despite the potential impact of B cells in antitumor responses and patient survival, the nature of B cell immunosurveillance during systemic anticancer therapy and across metastatic sites in breast cancer is unknown.

Here, we perform a comprehensive analysis of breast cancer immunosurveillance in metastatic and early breast cancer. By integrating BCR, TCR, DNA and RNA-sequencing (RNA-seq) data from a multisite metastatic cohort, and during neoadjuvant therapy in an early disease cohort, we tracked and characterized clones that were temporally persistent throughout therapy and across metastatic sites (spatio-migratory mapping). Using this data, we aimed to uncover three key features of B cell clonal temporal persistence and immunosurveillance. Firstly, to determine whether the intra-tumoral B cell response across metastases is correlated with the tumor genomic landscape and T cell response, in keeping with the immunoeediting hypothesis. Secondly, to determine the nature of B cell immunosurveillance between metastatic sites and throughout anticancer therapy. Lastly, we sought to identify what key B cell clonal features predict immunosurveillance and temporal persistence for future therapeutic exploration.

## Results

### Multi-platform metastatic tumor profiling

We performed BCR repertoire sequencing on 27 metastatic tumor biopsy samples obtained through warm autopsies of eight participants with therapy-resistant metastatic breast cancer to identify

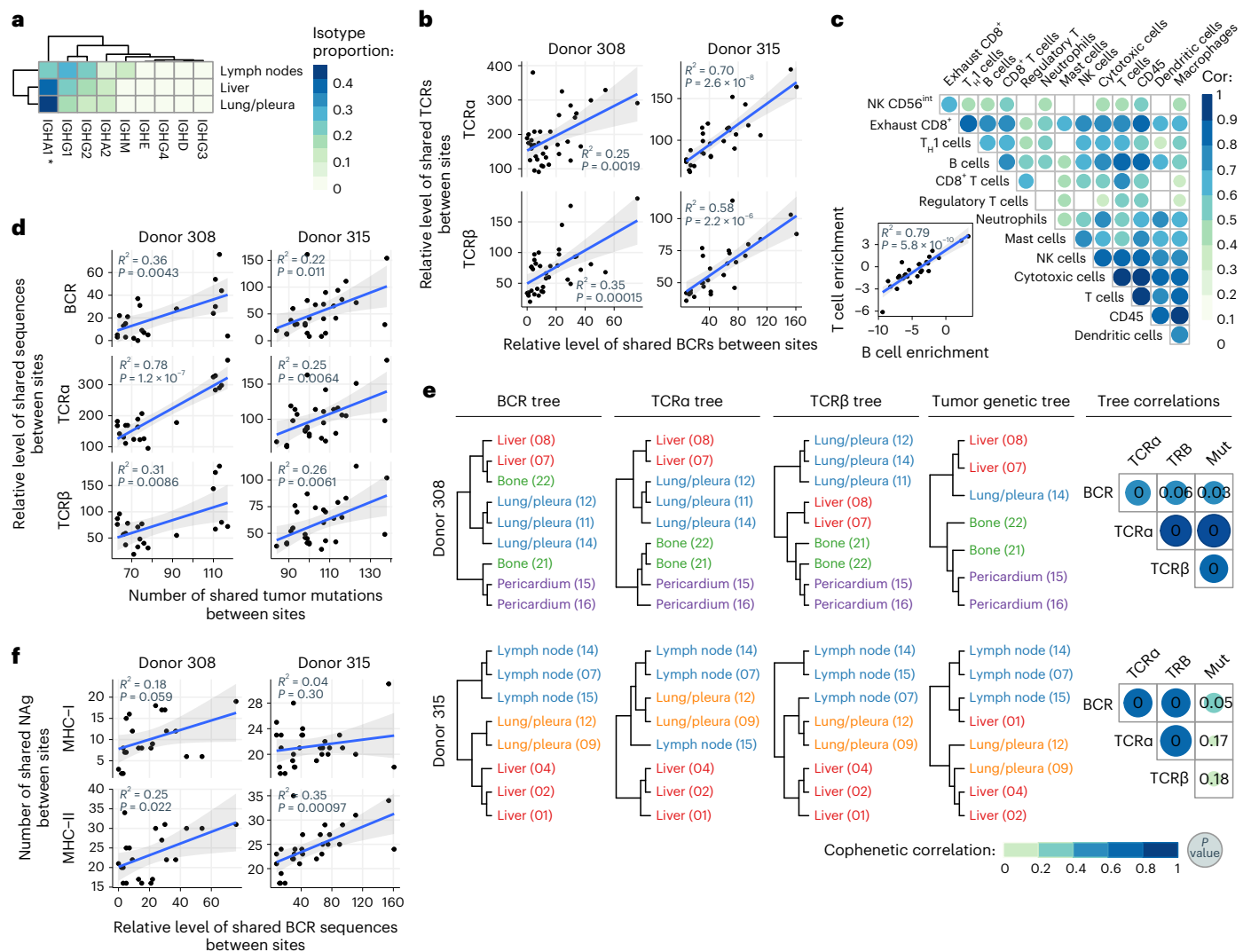
B cell clonality, isotype usages and clonal diversification across the metastases (Fig. 1 and Supplementary Table 1). The mean yield of unique BCRs for each metastatic site after filtering was 9,332 (range, 701–80,409; Extended Data Fig. 1a and Methods). The genomic, transcriptomic and TCR repertoires of these metastatic tumors have been previously reported<sup>16</sup>.

Significant BCR isotype usage variations were observed across metastatic sites, with liver and lung/pleura dominated by IgA1 (Fig. 2a and Extended Data Fig. 1b). The distribution of BCR isotypes across metastatic sites was distinct from that observed in healthy normal tissues using deconvolution of bulk RNA-seq data from the Genotype-Tissue Expression (GTEx) Consortium atlas<sup>17</sup> (Extended Data Fig. 1c and Methods). Additionally, there was a higher expression of both IGH and TCR genes in metastatic tumor tissues compared to normal tissues (Extended Data Fig. 1d). Together, these data suggest that the BCR and TCR patterns observed were the result of tumor-associated responses rather than reflecting healthy tissue heterogeneity.

### B cell and T cell clonal structures are correlated

B cell and T cell clones are defined by cells sharing related BCR or TCR VDJ rearrangements. We used the Jaccard index to quantify the degree of clonal sharing of the VDJ regions of the BCR, TCR $\alpha$  and TCR $\beta$  clones between sites (Extended Data Fig. 1e,f), revealing that BCR and TCR repertoires were distinct between each participant, in keeping with previous studies<sup>18</sup>. A low degree of BCR and TCR VDJ sequence sharing, which may occur by chance at low frequencies<sup>19</sup>, was observed between different participants, while high levels of BCR and TCR VDJ sharing were only observed in the metastases from the same participant.

We next compared the clonal structures across metastatic sites in the two participants in which BCR and TCR sequencing data were available for four or more sites (participants 308 and 315). TCR $\alpha$  and TCR $\beta$  clonal structures were correlated across metastases (strong correlation in participant 315 and, to a lesser degree, but also significant, in participant 308; Extended Data Fig. 2a), in keeping with the common origin of these receptors. BCR clonal structures across metastatic sites were also correlated with TCR $\alpha$  and TCR $\beta$  clonal structures, indicating shared factors driving B cell and T cell infiltration and selection (Fig. 2b). This was confirmed by deconvoluting tumor immune microenvironment composition and activity<sup>20</sup>



**Fig. 2 | Site-specific B cell infiltration correlates with T cell infiltration and tumor genomic landscape.** **a**, Mean BCR isotype usage across metastatic sites with more than two samples (lymph nodes  $n = 5$ , liver  $n = 6$ , lung/pleura  $n = 7$ ; IGH1  $**P = 0.042$ ).  $P$  values calculated using Kruskal–Wallis test and adjusted for multiple comparisons. **b**, Scatterplots showing the relative level of sharing of TCR $\alpha/\beta$  VDJ sequences between pairwise metastatic site comparisons. **c**, Correlation between tumor immune microenvironment components deconvoluted from bulk RNA-seq data using Danaher gene sets. Inset, scatterplot showing relationship between T cell and B cell enrichment.  $P$  value and  $R^2$  obtained from linear regression. NK, natural killer cell, T<sub>H</sub>1, type 1 helper T cells. Data from all sites ( $n = 27$ ) from all participants

are shown. **d**, Scatterplots showing number of shared BCR and TCR $\alpha/\beta$  VDJ sequences and tumor mutations between pairwise metastatic site comparisons. **e**, Clonal similarity trees for BCR, TCR $\alpha/\beta$  VDJ sequences and mutational phylogenetic trees for participants 308 and 315. Inter-tree correlations shown on the left. One-sided  $P$  values derived from permutation tests shown within correlation circles. Trees have arbitrary units for branch lengths. **f**, Scatterplots showing number of shared BCR VDJ sequences and predicted MHC class I and II neoantigens (NAg) between pairwise metastatic site comparisons. **b, d, f**, BCR and TCR sequences were downsampled.  $P$  value and  $R^2$  obtained from linear regression analysis. The shaded area, in gray, represents the 95% confidence interval. Inter-sample comparisons: participant 308,  $n = 36$ ; participant 315,  $n = 28$ .

from the bulk RNA-seq data using the Danaher gene sets<sup>21</sup> and MCP-counter<sup>22</sup>, which showed that both the abundance ( $R^2 = 0.79$ , Fig. 2c and Extended Data Fig. 2b) and activation ( $R^2 = 0.65$ ; Extended Data Fig. 2c) of tumor-infiltrating B cells and T cells were strongly correlated. B cell and T cell enrichment was also significantly associated with the expression of a TLS signature (Extended Data Fig. 2d)<sup>23</sup>, in keeping with observations that coordination between BCR and TCR repertoires occurs within these structures<sup>24</sup>. This relationship was also observed in The Cancer Genome Atlas (TCGA) early breast cancer cohort (Extended Data Fig. 2e).

In summary, B cell and T cell infiltration, clonality and activation are significantly correlated across metastases, providing evidence that B cell and T cell responses are coordinated across metastatic sites in each individual breast cancer participant.

### Adaptive immune and tumor genomic coevolution

We previously showed that T cell responses, assessed by TCR sequencing, appear to coevolve with the metastatic tumor genomes<sup>16</sup>. This prompted us to investigate whether a similar association would be observed for the B cell response. In the two participants for which more than four metastases were sequenced, B cell and T cell clonal compositions mirrored the tumor mutational landscape, with significant associations observed between the number of shared TCRs, BCRs and somatic mutations across metastatic sites ( $R^2$  range, 0.22–0.78,  $P \leq 0.011$ ; Fig. 2d).

To confirm this, unsupervised VDJ BCR and TCR Jaccard phylogenetic trees segregated metastases by organ, with consistent clustering patterns between BCR, TCR $\alpha$  and TCR $\beta$  chains (Fig. 2e). Similar tree structures were observed when tumor mutational phylogenies were



constructed from the whole-exome sequencing data (Fig. 2e). The BCR and TCR tree structures in both participants were significantly correlated when analyzed using the cophenetic statistic, with similar but weaker correlations observed when these were compared to the tumor mutational phylogenetic trees (Fig. 2e), providing further evidence that the tumor and the adaptive immune response coevolve. Finally, maps of B cell clonal structure across metastatic sites, generated through quantifying the degree of clonal sharing of the BCR clonotypes between sites (Extended Data Fig. 2f,g), confirmed that there was clonal overlap between most sites within an individual, but the levels were highly variable between sites.

We subsequently characterized correlations between predicted major histocompatibility complex (MHC) class I and II neoantigens and BCR and TCR clonal structure. There was a significant correlation between BCR clonal structure and shared MHC class II-predicted neoantigens ( $R^2$  range, 0.25–0.35,  $P < 0.022$ ; Fig. 2f) but not MHC class I-predicted neoantigens. Similar observations were made with TCR clonal structure (Extended Data Fig. 2h,i), suggesting that B cell and T cell clonal structures significantly mirror tumor MHC class II-predicted neoantigen architecture.

In summary, each individual metastasis has a unique BCR and TCR clonal architecture. However, more similar BCR and TCR repertoires exist between metastases sharing similar mutational landscapes, suggesting coevolution between tumors and B cell and T cell responses across metastases.

### Persistence and immunosurveillance of intra-tumoral B cells

We performed BCR repertoire sequencing on an early breast cancer cohort comprising ten participants with sequential tumor biopsy samples obtained during neoadjuvant therapy (25 serial samples:  $n = 10$  before therapy,  $n = 10$  after 9 weeks of therapy and  $n = 5$  on completion of therapy). We obtained a mean yield of 8,132 unique BCRs per biopsy after filtering (range, 762–15,493; Extended Data Fig. 3a and Supplementary Table 2). Each participant harbored distinct BCR repertoires (Extended Data Fig. 3b). Together with the metastatic dataset, this allowed an exploration of the spatial and temporal nature of tumor-infiltrating B cells.

B cell clones present at multiple time points during treatment (temporally persistent clones) or at multiple sites (immunosurveillance clones) were significantly enlarged compared with private clones, with BCR clone size correlating with both the number of time points and metastatic sites in which BCR clones were observed (Fig. 3a;  $P < 2.2 \times 10^{-16}$ , ordinal regression over the mean percentage clone size within each participant averaged over all sites observed), suggestive of immune surveillance by activated B cell clones. This directly shows that larger clones per site are associated with temporal persistence and immunosurveillance, rather than just a larger number of BCRs detected across all sites. Similarly, by classifying BCR clones as stem, clade or private depending on whether they were present in all, some or one tumor sample from the same participant, respectively, we observed that immunosurveillance and temporally persistent clones were significantly enlarged (stem > clade > private,  $P < 2.2 \times 10^{-16}$ , ordinal regression; Extended Data Fig. 4a).

Tumor-infiltrating BCRs were classified into four clone classes (A–D; Fig. 3a and Methods) based on whether they were (1) expanded or unexpanded within the tumor microenvironment, and (2) private to one site or shared between time points (temporally persistent) or multiple metastatic sites (immunosurveillance). There was no significant enrichment of BCR sequences with known binding to viral or bacterial antigens in these four clonal categories, indicating that these were not enriched for established systemic responses to non-cancer antigens and, therefore, did not just represent re-expansions of non-tumor-specific B cell clones (Extended Data Fig. 4b, Supplementary Table 3 and Methods). Expanded temporally persistent clones (clone class B) comprised the majority of tumor-infiltrating BCR

sequences throughout the course of therapy in early breast cancer (Fig. 3b). Likewise, expanded immunosurveillance clones (clone class B) comprised the majority of tumor-infiltrating BCR sequences in metastatic disease (Fig. 3b). These clones were also present at higher proportions in liver and lung/pleura metastases compared to private expanded clones (class A), suggesting that they are highly activated in these sites. Interestingly, within lymph node metastases, there was no significant difference between class A and B clone proportions, suggesting that a large fraction of activated B cell clones in lymph nodes are resident and not undergoing immunosurveillance.

### Antigen experience of migratory and persistent clones

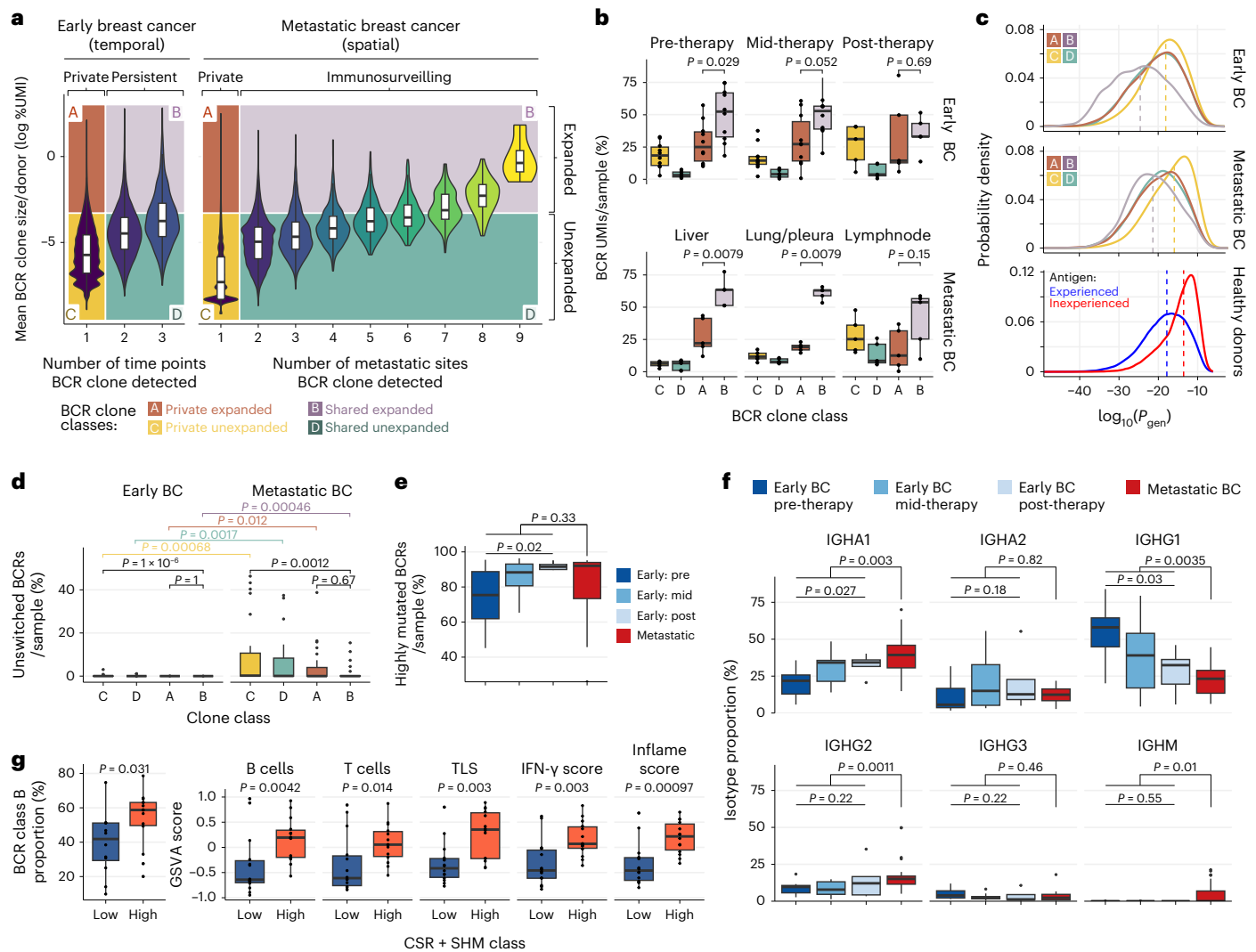
We next investigated whether the nature of shared B cell clones (clone classes B and D) was significantly distinct from private clones (clone classes A and C) based on BCR repertoire features. We calculated BCR CDR3 probability of generation ( $P_{\text{gen}}$ ) as a result of VDJ recombination (that is, the likelihood of being generated by chance rather than being individual specific) using OLGA<sup>25</sup>. We observed that clone class C (private unexpanded clones) had the highest probability of generation by chance, and the distribution was comparable to naive or antigen-inexperienced B cells from healthy peripheral blood mononuclear cells (Fig. 3c)<sup>26</sup>. The other clonal groups (B, C and D) had higher probabilities of BCR amino acid sequences resembling antigen-experienced BCRs, with the majority of these sequences being mutated and class switched, with clone class B (expanded and immunosurveillance) having the lowest  $P_{\text{gen}}$  scores. This suggests that the expanded immunosurveillance and temporally persistent clones are both selected on the basis of their BCR sequence and that these are likely to be participant-specific clones and from antigen-experienced B cells.

On encountering antigen, BCR sequences may diversify further via SHM, which introduces point mutations into the BCR, and class-switch recombination (CSR), which changes BCR isotype, to generate finely tuned humoral responses<sup>10</sup>. Measuring SHM and CSR between the different clone classes and by disease stage yielded three key observations. Firstly, expanded immunosurveillance clones (clone class B) had greater overall levels of class-switched BCRs (that is, lower levels of unswitched BCRs; Fig. 3d) compared to unexpanded private clones (clone class C). Furthermore, B cells infiltrating early tumors had lower levels of unswitched (IGHM/D) BCRs compared to metastasis-infiltrating B cells (Fig. 3d and Extended Data Fig. 4c). Secondly, the levels of SHM of tumor-infiltrating B cells varied by clone class (Extended Data Fig. 4d) and increased during treatment in early breast cancer, but this trend was reversed in the metastasis-infiltrating B cells (Fig. 3e and Extended Data Fig. 4e). These differences were driven by a higher proportion of low SHM BCRs and a lower proportion of high SHM BCRs in the metastasis-infiltrating B cells (Extended Data Fig. 4d). The association observed here of reduced SHM and CSR in metastasis-infiltrating B cells compared to B cells infiltrating the primary tumor site in early breast cancer is supported by the reduced expression levels of *AICDA*, which encodes a key enzyme associated with these processes (Extended Data Fig. 4f). Thirdly, the isotype usage proportions varied by clone class (Extended Data Fig. 4g) and varied with disease course, with IGHAI increasing with time and IGHG1 decreasing with time (Fig. 3f), with this trend driven by clonal class B BCRs (Extended Data Fig. 4g). This is supported by the higher expression of IgA isotype switching and the lower expression of IgG isotype switching signatures in metastatic samples (Extended Data Fig. 4h).

Furthermore, tumors with high levels of both BCR SHM and class switching were associated with significantly higher levels of class B clonal B cells, as well as higher levels of B cell and T cell infiltration, TLS score, interferon gamma (IFN- $\gamma$ ) score and inflammation scores (Fig. 3g). The effect observed in the tumor was much more pronounced compared to that seen in healthy tissues (Extended Data Fig. 4i).

In summary, these data suggest that temporally persistent and immunosurveillance clones are significantly distinct from private clones





**Fig. 3 | Immunosurveillance and persistent clones are enlarged and distinct from private clones.** **a**, Violin plots of mean BCR clone sizes per sample across sampling time points in early breast cancer ( $n = 94,495$  unique BCR clones) and across number of sites in metastatic breast cancer ( $n = 155,451$  unique BCR clones). BCR clones classified as private expanded (class A,  $n = 10,507$  clones), shared expanded (class B,  $n = 6,358$  clones), private unexpanded (class C,  $n = 217,093$  clones) and shared unexpanded (class D,  $n = 15,988$  clones). **b**, Box plots showing percentage of expanded BCRs in early breast cancer (pre-therapy:  $n = 10$ , mid-therapy:  $n = 10$ , post-therapy:  $n = 5$ ) and metastatic breast cancer (liver:  $n = 5$ , lymph node:  $n = 5$ , lung:  $n = 5$ ) samples. **c**, Distribution of BCR CDR3  $P_{gen}$  scores in the BCR clone classes in the early breast cancer and metastatic datasets. Distribution of BCR CDR3  $P_{gen}$  scores in a comparative healthy peripheral blood mononuclear cell dataset across antigen-experienced (CSR and SHM) and antigen-inexperienced (no CSR and SHM) BCRs. **d**, Box plots showing the percentage of unswitched BCRs (IgD/IgM) per sample across the four BCR clone classes. **e**, Box plot showing the percentage of highly mutated

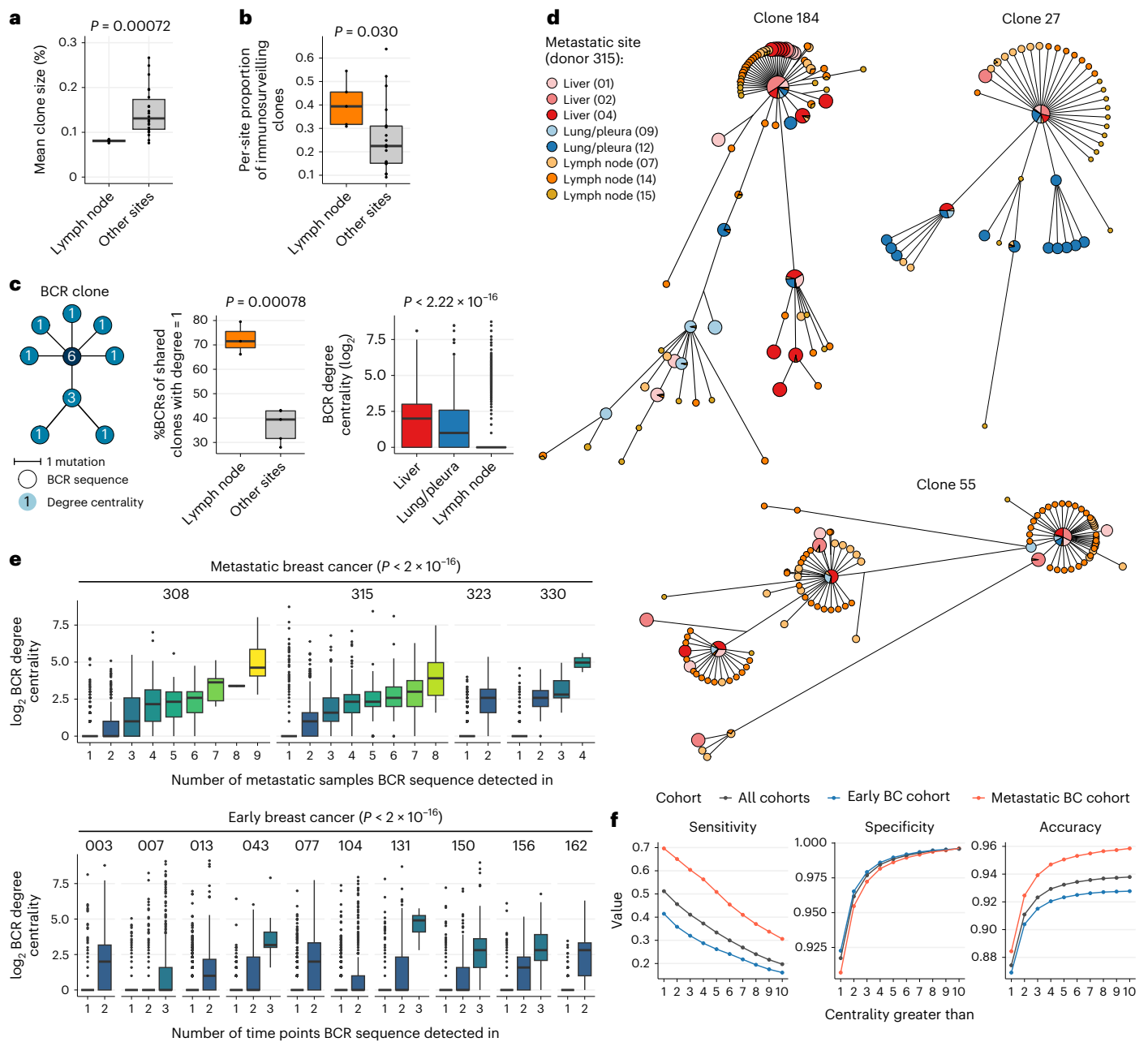
BCRs in early and metastatic breast cancer samples. **f**, Box plots showing the percentage isotype usage in early and metastatic breast cancer samples. **g**, Box plots showing the distribution of the proportion of BCR class B clones, and signature scores of B cell, T cell, TLS, IFN- $\gamma$  and T cell inflamed deconvoluted from bulk RNA-seq data between samples with low and high levels of SHM and CSR (high, >50th percentile (SHM, CSR); low,  $\leq$ 50th percentile (SHM, CSR)). **a, c, d, g**, (Data (left) from participants with more than one tumor site sampled shown (early breast cancer cohort: all participants ( $n = 10$ ), metastatic breast cancer cohort participants: 308, 315, 323 and 330). **e, f, g**, Data (right) from all participants ( $n = 18$ ) and all samples ( $n = 52$ ). **e, f**,  $P$  early versus late breast cancer: Wilcoxon rank-sum tests. All  $P$  values are two sided. **b, d, e, g**, Wilcoxon rank-sum tests. All  $P$  values are two sided. **a, b, d–g**, The box bounds denote the interquartile range, the line indicates the median, and whiskers indicate maximum of 1.5 times the interquartile range beyond the box. Individual data points are shown as dots.

by being clonally expanded and antigen experienced, rather than being naive B cells, in agreement with previous studies<sup>11</sup>. Higher levels of CSR and SHM are associated with higher levels of B cell and T cell infiltration and TLS scores, suggesting that the tumor microenvironment drives these differences.

### BCR centrality reveals sites of clonal diversification

To determine whether B cell clonal diversification occurred within each metastasis or was localized to specific anatomical locations, per-sample BCR clonal expansion and diversification measures were

calculated<sup>26</sup>. Lymph nodes had significantly lower levels of clonal unevenness, thus by extension, higher levels of clonal diversity (measured by the normalized mean clone size index (Fig. 4a) and Shannon and Gini indices (Extended Data Fig. 5a),  $P < 0.05$  with effect sizes  $>1.33$ ). However, there was a greater abundance of expanded clones in lymph node than non-lymph node sites (Extended Data Fig. 5b), indicating that there are more B cell clonal expansions in the lymph nodes, and only some clones are overrepresented in the non-lymph node sites. Additionally, lymph nodes had a higher proportion of unique BCRs from immunosurveillance clones compared to other sites (Fig. 4b). Together,



**Fig. 4 | Higher BCR centrality describes clonal structure and predicts B cell immunosurveillance and persistence.** **a**, Box plot showing mean BCR clone size in lymph nodes versus other sites ( $n = 5$  lymph node,  $n = 22$  other sites). **b**, Box plot showing per-site proportion of immunosurveillance clones in lymph nodes versus other sites ( $n = 5$  lymph node,  $n = 22$  other sites). **c**, Left, schematic of degree centrality as applied to BCR sequences within a BCR clone. Middle, box plots showing percentage of BCRs per sample with a degree centrality of 1 (that is, no progeny) in lymph nodes ( $n = 3$ ) compared to other metastatic sites ( $n = 5$ ) in participant 315. Right, box plots showing distribution of BCR degree centrality across different metastatic sites sampled in participant 315. **d**, BCR VDJ network plots showing three examples of expanded immunosurveillance clones shared between multiple metastatic sites in participant 315. These networks are based

on maximum parsimony trees calculated from BCR sequence alignments. **e**, Box plots showing association between BCR degree centrality and the number of metastatic sites and therapy time points in which the BCR is observed.  $P$  values calculated using two-sided analysis of variance. **f**, Profile plots showing changes in sensitivity, specificity and accuracy at identifying immunosurveillance BCRs at different degree centrality thresholds in all samples, early breast cancer samples and metastatic breast cancer samples. **a–c**, Data from four participants with more than one metastatic site sampled (308, 315, 323 and 330) used. **a–c**, Wilcoxon rank-sum tests. All  $P$  values are two sided. **a–c**, The box bounds denote the interquartile range divided by the median, with the whiskers extending to a maximum of 1.5 times the interquartile range beyond the box. Individual data points are shown as dots.

this suggests that clonal diversification predominantly occurs within lymph nodes, and these are a main source of immunosurveillance B cells in metastatic breast cancer.

Next, we derived the BCR phylogenetic degree centrality, representing the number of edges connected to each BCR node in the network (Fig. 4c). This allowed us to distinguish between BCRs derived

from B cells that underwent subsequent clonal diversification and were progenitors to many other BCR variants (high centrality) from those derived from B cells that did not undergo subsequent clonal diversification and were not progenitors to further BCR variants (unitary centrality). The majority of lymph node BCRs had a degree centrality of one compared to other metastatic sites (Fig. 4c), indicating that

lymph nodes are key sites of clonal diversification where many exploratory variants are generated. Conversely, non-lymph node metastatic sites had a higher proportion of BCRs with degree centrality greater than one, indicating that these BCRs are predominantly variants of expanded clones under significant selection (that is, non-exploratory variants). These data suggest that higher levels of clonal diversification occur in lymph nodes, with high-centrality BCRs more likely to migrate to non-lymph node sites than low-centrality BCRs. There is minimal additional diversification in non-lymph node metastatic sites, and these typically do not undergo immunosurveillance to other sites.

### High BCR centrality of immunosurveillance and persistent BCRs

We next investigated B cell clonal relationships across sites to determine whether all members of expanded immunosurveillance clones (clone class B) underwent active metastatic immunosurveillance, or whether migration was restricted to a predictable subset of BCRs within each clone. BCRs from expanded clones ( $\geq 10$  BCRs) were aligned and phylogenetic trees estimated to determine their lineage relationships. These were then represented as non-cyclic networks (Fig. 4c), with nodes representing unique BCRs and edges representing SHM between related BCRs. Visual representations of BCR clonal phylogenetic trees (Fig. 4d and Extended Data Fig. 5c) demonstrate this trend, with highly central BCRs shared between multiple sites and BCRs with a centrality of one typically observed in single sites.

Furthermore, BCR degree centrality was also strongly correlated with both (a) the number of metastatic sites in which the BCR was observed ( $P < 2.2 \times 10^{-16}$ ; Fig. 4e), indicating that a small proportion of variants per activated clone, which are typically more central within the clone, perform immunosurveillance across multiple metastatic sites, and (b) the number of time points in which the BCR was observed ( $P < 2.2 \times 10^{-16}$ ; Fig. 4e), indicating that a small proportion of high-centrality BCR variants per activated clone are temporally persistent. This increased BCR degree centrality was not associated with systemic responses against noncancerous antigen (Extended Data Fig. 5d). BCR degree centrality was independent of BCR SHM level (Extended Data Fig. 5e), showing that immunosurveillance BCRs are not necessarily the most mutated versions of these clones, but rather represent local optima of the clonal response to its antigen. Finally, BCR degree centrality also correlated significantly with BCR frequency (Extended Data Fig. 5f;  $P < 2.2 \times 10^{-16}$ ) in addition to immunosurveillance and clonal persistence. Together, this points to BCR clonal structure as a predictor of B cell activation, expansion and migratory potential.

We lastly determined whether BCR degree centrality would have sufficient power to predict immunosurveillance and clonal persistence. Indeed, degree centrality was highly predictive of BCR immunosurveillance status and clonal persistence, with a degree classification threshold greater than two resulting in an immunosurveillance and clonal persistence BCR identification accuracy greater than 80% (Fig. 4f and Extended Data Fig. 5g), which was robust to sequencing depth (Extended Data Fig. 5h).

The same association was observed in two independent breast cancer datasets<sup>27,28</sup> (BCR data obtained following the deconvolution of bulk RNA-seq data; Extended Data Fig. 5i). We also observed that this trend of higher BCR centrality correlating with immunosurveillance is generalizable to noncancerous disease states, including autoimmunity (diabetes mellitus<sup>29</sup> and multiple sclerosis<sup>30</sup>; Extended Data Fig. 5i).

In summary, BCRs diversify predominantly in the lymph nodes and only a small selection of B cells expressing these clonal BCR variants are able to perform immunosurveillance across other sites or are temporally persistent. Higher-centrality BCRs are more likely to be seen across a larger number of sites. These immunosurveillance and temporally persistent BCRs can be predicted from their centrality with respect to the overall clonal structure.

## Discussion

Anticancer immunosurveillance by B cells and T cells plays a central role in sculpting malignant clones, and disruption of this process is a hallmark of cancer<sup>31</sup>. A central finding of our study was that it appears that both arms of the adaptive immune response coevolve in a correlated fashion, suggesting common drivers of immune cell infiltration, selection and clonal expansion across metastatic sites. These adaptive immunity B cell and T cell clonal structures also correlate with the tumor mutational phylogenetic landscape, providing further support in favor of the immunoeediting hypothesis<sup>32</sup>, where failure of the immune system to eliminate malignant cell populations results in a phase of equilibrium, in which the immune system limits but cannot eradicate the tumor, resulting in selection pressures that drive tumor evolution toward a state of reduced immunogenicity.

Mutated peptides can be presented on both MHC class I and class II molecules. MHC class II molecules are primarily expressed on professional antigen-presenting cells such as dendritic cells, B cells and macrophages, and predominantly present exogenously derived peptide antigens to CD4<sup>+</sup> T cells<sup>33</sup>. Indeed, B cells use a specialized MHC class II presentation to internalize and process BCR-bound antigen for presentation to CD4<sup>+</sup> T cells, which has been shown to influence the fate of both B and T cells<sup>34,35</sup>. The majority of intra-tumoral B cells have been shown to be non-antibody-secreting cells, but rather have a naive or memory phenotype with surface BCR<sup>11,36,37</sup>. The significant correlation between shared BCR sequences and MHC II, but not MHC I, supports the notion that B cells play a role in presenting antigen to T cells through BCR-dependent mechanisms<sup>35</sup>. Even though our data are unable to distinguish between CD4<sup>+</sup> and CD8<sup>+</sup> TCRs, they strongly support the hypothesis that tumor MHC class II neoantigens may be important in coordinating tumor-specific B and T cell responses, as the tumor MHC class II neoantigen landscape correlated with both B and T clonal structures. In keeping with this observation, MHC class II neoantigens have been recently shown to predict outcomes in HER2-negative breast cancer<sup>38</sup> and associate with tumor-infiltrating lymphocytes and interferon signaling<sup>39</sup>.

The nature of B cells and T cells migrating between the tumor and draining lymph nodes is important for mounting effective antitumor immune responses, for TLS formation and for establishing long-term systemic memory, which are strongly associated with outcome<sup>40</sup>. However, despite the potential impact of B cells in antitumor responses and participant survival, the nature of B cell immunosurveillance across metastatic sites is unknown. Here we show that the majority of intra-tumoral B cells are temporally persistent and undergo tumor immunosurveillance across sites. These immunosurveillance and temporally persistent B cell clones are antigen experienced and isotype usages vary with disease stage. While some of these measures do not show a high correlation and causality remains unexplored, this is in line with previous studies showing a need for a diverse antibody repertoire for early neoplastic cell recognition and the critical role B cells play in anticancer immunity<sup>41,42</sup>.

Finally, we show that not all BCRs from expanded shared clones perform immunosurveillance. We have generated a pipeline that uses network graph theory to predict which BCR sequences within an immunosurveillance BCR clone perform cross-site immunosurveillance. These B cells tend to have higher BCR degree centrality but do not have the highest level of SHM within the clone. Therefore, these are likely to represent local optima of the B cell clonal response to its antigen. Furthermore, we show that BCR degree centrality can be used to predict BCR clonal persistence and demonstrate its generalizability across other breast cancer datasets and non-cancer datasets. While the concept of BCR degree centrality has been used to describe B cell population distributions<sup>43</sup>, we show functional differences between low-centrality and high-centrality B cell clonal variants for the prioritization of specific BCRs. Indeed, this study shows functional and BCR-dependent associations with B cell immunosurveillance and clonal



persistence. While these findings are primarily observational, hence the significance in the broader context of cancer immune response and participant outcomes is mostly correlative, they potentially lay the foundation for expediting the discovery of tumor-specific or persistent B cell clones. Given these findings, we hypothesize that this can be used to develop personalized antibody-based therapies based on BCR network degree centrality.

## 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/s41590-024-01821-0>.

## References

- Hiam-Galvez, K. J., Allen, B. M. & Spitzer, M. H. Systemic immunity in cancer. *Nat. Rev. Cancer* **21**, 345–359 (2021).
- Tao, H. et al. Antitumor effector B cells directly kill tumor cells via the Fas/FasL pathway and are regulated by IL-10. *Eur. J. Immunol.* **45**, 999–1009 (2015).
- Bindea, G. et al. Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. *Immunity* **39**, 782–795 (2013).
- Garaud, S. et al. Tumor infiltrating B-cells signal functional humoral immune responses in breast cancer. *JCI Insight* **5**, 129641 (2019).
- Petitprez, F. et al. B cells are associated with survival and immunotherapy response in sarcoma. *Nature* **577**, 556–560 (2020).
- Alberts, E., Wall, I., Calado, D. P. & Grigoriadis, A. Immune crosstalk between lymph nodes and breast carcinomas, with a focus on B Cells. *Front. Mol. Biosci.* **8**, 673051 (2021).
- Helmink, B. A. et al. B cells and tertiary lymphoid structures promote immunotherapy response. *Nature* **577**, 549–555 (2020).
- Sammut, S. -J. et al. Multi-omic machine learning predictor of breast cancer therapy response. *Nature* **601**, 623–629 (2022).
- Norton, N. et al. Generation of HER2-specific antibody immunity during trastuzumab adjuvant therapy associates with reduced relapse in resected HER2 breast cancer. *Breast Cancer Res.* **20**, 52 (2018).
- Victoria, G. D. & Nussenzweig, M. C. Germinal centers. *Annu. Rev. Immunol.* **30**, 429–457 (2012).
- Hu, Q. et al. Atlas of breast cancer infiltrated B-lymphocytes revealed by paired single-cell RNA-sequencing and antigen receptor profiling. *Nat. Commun.* **12**, 2186 (2021).
- Aizik, L. et al. Antibody repertoire analysis of tumor-infiltrating B cells reveals distinct signatures and distributions across tissues. *Front. Immunol.* **12**, 705381 (2021).
- Tabuchi, Y. et al. Protective effect of naturally occurring anti-HER2 autoantibodies on breast cancer. *Breast Cancer Res. Treat.* **157**, 55–63 (2016).
- Bushey, R. T. et al. A therapeutic antibody for cancer, derived from single human B cells. *Cell Rep.* **15**, 1505–1513 (2016).
- Dunn, G. P., Bruce, A. T., Ikeda, H., Old, L. J. & Schreiber, R. D. Cancer immunoeediting: from immunosurveillance to tumor escape. *Nat. Immunol.* **3**, 991–998 (2002).
- De Mattos-Arruda, L. et al. The genomic and immune landscapes of lethal metastatic breast cancer. *Cell Rep.* **27**, 2690–2708 (2019).
- GTE Consortium. The GTE Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
- Dupic, T. et al. Immune fingerprinting through repertoire similarity. *PLoS Genet.* **17**, e1009301 (2021).
- Trück, J. et al. Identification of antigen-specific B cell receptor sequences using public repertoire analysis. *J. Immunol.* **194**, 252–261 (2015).
- Rooney, M. S., Shukla, S. A., Wu, C. J., Getz, G. & Hacohen, N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell* **160**, 48–61 (2015).
- Danaher, P. et al. Gene expression markers of tumor infiltrating leukocytes. *J. Immunother. Cancer* **5**, 18 (2017).
- Becht, E. et al. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol.* **17**, 218 (2016).
- Cabrita, R. et al. Tertiary lymphoid structures improve immunotherapy and survival in melanoma. *Nature* **577**, 561–565 (2020).
- Sautès-Fridman, C., Petitprez, F., Calderaro, J. & Fridman, W. H. Tertiary lymphoid structures in the era of cancer immunotherapy. *Nat. Rev. Cancer* **19**, 307–325 (2019).
- Sethna, Z., Elhanati, Y., Callan, C. G., Walczak, A. M. & Mora, T. OLGA: fast computation of generation probabilities of B- and T-cell receptor amino acid sequences and motifs. *Bioinformatics* **35**, 2974–2981 (2019).
- Bashford-Rogers, R. J. M. et al. Analysis of the B cell receptor repertoire in six immune-mediated diseases. *Nature* **574**, 122–126 (2019).
- Priestley, P. et al. Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature* **575**, 210–216 (2019).
- Siegel, M. B. et al. Integrated RNA and DNA sequencing reveals early drivers of metastatic breast cancer. *J. Clin. Invest.* **128**, 1371–1383 (2018).
- Seay, H. R. et al. Tissue distribution and clonal diversity of the T and B cell repertoire in type 1 diabetes. *JCI Insight* **1**, e88242 (2016).
- Stern, J. N. H. et al. B cells populating the multiple sclerosis brain mature in the draining cervical lymph nodes. *Sci. Transl. Med.* **6**, 248ra107 (2014).
- Hanahan, D. Hallmarks of cancer: new dimensions. *Cancer Discov.* **12**, 31–46 (2022).
- Dunn, G. P., Old, L. J. & Schreiber, R. D. The three Es of cancer immunoeediting. *Annu. Rev. Immunol.* **22**, 329–360 (2004).
- Axelrod, M. L., Cook, R. S., Johnson, D. B. & Balko, J. M. Biological consequences of MHC-II expression by tumor cells in cancer. *Clin. Cancer Res.* **25**, 2392–2402 (2019).
- McShane, A. N. & Malinova, D. The ins and outs of antigen uptake in B cells. *Front. Immunol.* **13**, 892169 (2022).
- Adler, L. N. et al. The other function: class II-restricted antigen presentation by B cells. *Front. Immunol.* **8**, 319 (2017).
- Zou, Y. et al. The single-cell landscape of intratumoral heterogeneity and the immunosuppressive microenvironment in liver and brain metastases of breast cancer. *Adv. Sci.* **10**, 2203699 (2023).
- Chen, Y. et al. Single-cell sequencing and bulk RNA data reveal the tumor microenvironment infiltration characteristics of disulfidptosis related genes in breast cancer. *J. Cancer Res. Clin. Oncol.* <https://doi.org/10.1007/s00432-023-05109-y> (2023).
- Gonzalez-Ericsson, P. I. et al. Tumor-specific major histocompatibility-II expression predicts benefit to anti-PD-1/L1 therapy in patients with HER2-negative primary breast cancer. *Clin. Cancer Res.* **27**, 5299–5306 (2021).
- Park, I. A. et al. Expression of the MHC class II in triple-negative breast cancer is associated with tumor-infiltrating lymphocytes and interferon signaling. *PLoS ONE* **12**, e0182786 (2017).
- Yanguas, A. et al. ICAM-1-LFA-1-dependent CD8<sup>+</sup> T-lymphocyte aggregation in tumor tissue prevents recirculation to draining lymph nodes. *Front. Immunol.* **9**, 2084 (2018).

41. Rawat, K., Tewari, A. & Jakubzick, C. V. A critical role for B cells in cancer immune surveillance. Preprint at *bioRxiv* <https://doi.org/10.1101/2020.09.19.304790> (2020).
42. Rawat, K., Tewari, A., Morrisson, M. J., Wager, T. D. & Jakubzick, C. V. Redefining innate natural antibodies as important contributors to anti-tumor immunity. *eLife* **10**, e69713 (2021).
43. Miho, E., Roškar, R., Greiff, V. & Reddy, S. T. Large-scale network analysis reveals the sequence space architecture of antibody repertoires. *Nat. Commun.* **10**, 1321 (2019).

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

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024

## Methods

### Study population

Eight participants with metastatic breast cancer enrolled within the Vall d'Hebron Institute of Oncology (VHIO) Warm Autopsy Program were included within this study. Ethical approval from the institutional review board of the Vall d'Hebron University Hospital (Barcelona, Spain) was obtained for the use of biospecimens with linked pseudo-anonymized clinical data. The ten participants with primary invasive early breast cancer included in this study were enrolled in the TransNEO study at Cambridge University Hospitals NHS Foundation Trust. Appropriate ethical approval from the institutional review board (research ethics ref.12/EE/0484) was obtained for the use of biospecimens with linked pseudo-anonymized clinical data. All participants provided informed consent for sample collection, and all participants consented to the publication of research results. Full details regarding sample collection, DNA and RNA extraction, library preparation and sequencing have been published elsewhere<sup>8,16</sup>. No statistical methods were used to predetermine sample sizes, but our sample sizes are similar to those reported in previous publications<sup>16</sup>. When performing statistical testing, we assessed whether the data met the assumptions of the tests used.

### DNA somatic mutation calling and neoantigen prediction

Somatic mutations (Fig. 2d) and predicted HLA class I neoantigens (Fig. 2f) were identified from whole-exome sequencing data and tumor phylogenetic trees (Fig. 2e) were generated using OncoNEM<sup>44</sup>, as previously described<sup>16</sup>. MHC class II allele genotyping was performed on the normal tissue DNA sequencing data using HLA-HD<sup>45</sup> (version 1.4) using default parameters. MHC class II neoantigens were predicted from the whole-exome mutation data using mixMHC2pred<sup>46</sup> (version 1.2) and putative candidates with a percentage rank cutoff of 2% were retained (Fig. 2f).

### TME composition and activity deconvolution from bulk RNA-seq

RNA-seq data from the early and metastatic breast cancer cohorts were processed as previously described<sup>8,16</sup>. Briefly, FASTQ files were aligned to the GRCh37 assembly of the human genome using STAR<sup>47</sup> (version 2.5.2b) in two-pass mode and counting of reads aligned over exonic features performed using HTSeq<sup>48</sup> (version 0.6.1p1) in read strand-aware union overlap resolution mode.

Immune cell enrichment was performed using MCPcounter<sup>22</sup> (version 1.2.0), using as input normalized log-transformed RNA-seq expression data (Extended Data Fig. 2b), and enrichment over 14 cell types using 60 genes<sup>21</sup> (Figs. 2c and 3g and Extended Data Figs. 2c, d and 4i). In Extended Data Fig. 2e, published TCGA B cell and T cell enrichment scores are shown<sup>21</sup>. Correlations between tumor micro-environment components shown in Fig. 2c and Extended Data Fig. 2b were generated using the *cor* function in the base R stats package and visualized using the *corrplot* package (version 0.92). The TLS gene signature (*CCL19*, *CCL21*, *CXCL13*, *CCR7*, *CXCR5*, *SELL*, *LAMP3*)<sup>23</sup> shown in Fig. 3g and Extended Data Figs. 2d, e and 4i was calculated using gene-set enrichment analysis. TCGA TLS enrichment scores (Extended Data Fig. 2e) were obtained using FPKM normalized counts provided by TCGA (Genomic Data Commons data release 37.0).

The cytolytic activity score<sup>20</sup> (CYT; Extended Data Fig. 2c) was computed as the geometric mean of *GZMA* and *PRFI* expression (TPM, 0.01 offset). The T cell inflamed score<sup>49</sup> (Fig. 3g and Extended Data Figs. 2c and 4i) was computed using the GSVA<sup>50</sup> R package (version 1.38.2) using as input the log-normalized expression of 18 inflammatory genes (*TIGIT*, *CD27*, *CD8A*, *PDCD1LG2*, *LAG3*, *CD274*, *CXCR6*, *CMKLR1*, *NKG7*, *CCL5*, *PSMB10*, *IDO1*, *CXCL9*, *HLA-DQA1*, *CD276*, *STAT1*, *HLA-DRB1* and *HLA-E*), while the interferon- $\gamma$  score<sup>49</sup> (Fig. 3g and Extended Data Figs. 2c and 4i) was computed using gene-set variation analysis of six genes (*IFNG*, *STAT1*, *IDO1*, *CXCL10*, *CXCL9* and *HLA-DRA*). The B cell activation

score shown in Extended Data Fig. 2c was computed using GSVA on the MSigDB<sup>31</sup> (version 7.3) C5 Gene Ontology Biological Processes POSITIVE\_REGULATION\_OF\_B\_CELL\_ACTIVATION (GO:0050871) gene set, using as input the log<sub>2</sub> TPM expression, with 0.01 offset.

### Healthy tissue GTEx isotype analysis

In the healthy tissue BCR isotype analysis shown in Extended Data Fig. 1c, normalized gene counts (TPM) were downloaded from the GTEx<sup>17</sup> consortium website (version 8, <https://gtexportal.org/home/datasets>) and the expression of IGH isotypes retained. Expression data were available for 3,905 samples from organ sites sampled within this study (GTEx *n*: brain = 2,642, breast = 459, liver = 226, lung = 578). In Extended Data Fig. 1c, the heat map shows the proportion of isotype TPM expression per organ site. In Extended Data Fig. 1d, the median z-score scaled expression of BCR isotypes is shown. The expression values of *CD3D*, *CD3G*, *CD3E* and *CD247*, which encode for the four different parts of the CD3 complex, were summed to calculate TCR expression. In Extended Data Fig. 4i, samples with high expression of unswitched transcripts were defined as those with a >50th percentile expression of IGH/IGHM genes, while those with low expression of unswitched transcripts were defined as those with a  $\leq$ 50th percentile expression of IGH/IGHM.

### BCR library preparation and sequencing

BCR libraries were prepared from RNA samples extracted from 27 metastatic sites and 25 primary breast tumors. BCR variable heavy domains were first amplified using a protocol we have previously described<sup>52</sup>. Briefly, RNA was reverse transcribed to cDNA using a mixture of IgA/IgD/IgE/IgG/IgM isotype specific primers, incorporating 15 nucleotide unique molecular identifiers (UMIs). The resulting cDNA was used as a template for PCR amplification using a set of six FRI-specific forward primers including sample-specific barcode sequences (seven nucleotides) along with a reverse primer specific to the reverse transcription primer. For three of the replicate libraries, a modified primer set was used where the sample-specific barcode was instead incorporated into the reverse transcription primers after the UMI.

BCR variable heavy domain amplicons (~450 bp) were quantified by TapeStation (Beckman Coulter) and subjected to gel purification. Dual-indexed sequencing adapters (KAPA) were ligated onto  $\leq$ 500 ng of amplicon per sample using the HyperPrep library construction kit (KAPA). The adaptor-ligated libraries were finally PCR amplified (initial denaturation at 95 °C for 1 min, for 2–83 cycles at 98 °C for 15 s, 60 °C for 30 s, 72 °C for 30 s and a final extension at 72 °C for 1 min). The libraries were sequenced on an Illumina MiSeq using the 2 × 300-bp chemistry.

### BCR-sequencing processing

Raw BCR-sequencing reads were processed for analysis using the Immcantation framework, using previously described parameters (docker container v3.0.0)<sup>52,53</sup>. Briefly, paired-end reads were joined based on a minimum overlap of 20 nucleotides, and a maximum error of 0.2, and reads with a mean Phred score below 20 were removed. Primer regions, including UMIs and sample barcodes, were then identified within each read, and trimmed. Together, the sample barcode, UMI, and constant region primer were used to assign molecular groupings for each read. Within each grouping, *usearch*<sup>54</sup> was used to subdivide the grouping, with a cutoff of 80% nucleotide identity, to account for randomly overlapping UMIs. Each of the resulting groupings is assumed to represent reads arising from a single RNA. Reads within each grouping were then aligned, and a consensus sequence determined. To remove low-level noise, molecular groupings with two or fewer sequences contributing to the UMI consensus were filtered out (Supplementary Table 1). Duplicate reads were then collapsed into a single processed sequence. *IgBlast*<sup>55</sup> (version 1.14.0) was used to annotate the processed sequences, and unproductive sequences were removed. Sequence data from replicate libraries were then pooled for analysis.



### BCR clonotype assembly

Annotation of TCR and BCR sequences were performed using IMG/HighV-QUEST<sup>56</sup> (version 1.8.5) and clonotype assembly performed using MRDARCY<sup>57</sup>, which was run using default parameters. B cell clones are groups of B cells from an individual that derive from the same pre-B cell, and thus have identical BCR sequences or BCR sequences related by SHM. Computationally, BCRs from clonal B cells can be clustered together via network generation using a previously described pipeline<sup>26</sup>. Briefly, each vertex represents a unique sequence, and the relative vertex size is proportional to the number of identical reads. Edges join vertices that differ by single-nucleotide non-indel differences and clusters are collections of related, connected vertices. A clone (cluster) refers to a group of clonally related B cells, each containing BCRs with identical CDR3 regions and IGHV gene use, or differing by single point mutations, such as through SHM. Likewise, a T cell clone (cluster) refers to a group of related T cells arising from the same pre-T cell, each containing TCRs with identical CDR3 regions and TCRV gene usage.

### BCR CDR3 overlap with reference pathogen antibody libraries

A reference antibody database with known binding to viral or bacterial antigen was constructed from existing public databases: the structural antibody database<sup>58</sup>, abYsis human antibody database<sup>59</sup> and the immune epitope database<sup>60</sup>. Antibody sequences corresponding to synthetic fusion proteins and animal-derived BCRs were excluded.

After preprocessing, 5,800 antibody sequences reacting to antigens were retained, including those derived from human immunodeficiency virus-1 ( $n = 3,525$ ), *Clostridium tetani* ( $n = 817$ ), influenza A ( $n = 486$ ), vaccinia virus ( $n = 92$ ), hepatitis C virus ( $n = 80$ ), *Streptococcus pneumoniae* ( $n = 59$ ), *Staphylococcus aureus* ( $n = 38$ ) and human betaherpesvirus 5 ( $n = 32$ ) were used for downstream analysis (Supplementary Table 2).

To determine potential matches, we screened the cancer CDR3 amino acid sequences to the reference antibody database, allowing for up to three amino acid mismatches by fuzzy string matching via a custom Python script. The proportions of BCRs/sample associated with known binding to viral or bacterial antigen across clone classes (Extended Data Fig. 4b) and degree centrality (Extended Data Fig. 5d) were calculated to show that the observations made were not secondary to established systemic responses to non-cancer antigens.

### TCR library preparation and sequencing

TCR-sequencing library preparation, sequencing and repertoire identification and network analysis performed by us have been described previously<sup>16</sup>. Briefly, MiSeq libraries were prepared using the same protocol as for the BCR libraries. Raw MiSeq reads were filtered for base quality, primer and constant region trimming, annotation and clustering using the same protocol as for the BCR libraries but using TCR as the chain parameter.

### Clonal overlap between metastatic sites

In Fig. 2 and Extended Data Fig. 2, the clonal repertoire analyses for participants 308 and 315 that were dependent on sequencing depth were generated by subsampling each sample to 90% of the number of unique VDJ sequences present in the sample with the lowest depth (unique VDJ subsampling thresholds: participant 308:  $n = 980$  (BCR), 4,657 (TCR $\alpha$ ), 2,620 (TCR $\beta$ ); participant 315:  $n = 1,524$  (BCR), 3,199 (TCR $\alpha$ ), 2,535 (TCR $\beta$ ). Throughout the paper, we have used the term ‘relative level’ to indicate that the analyses were performed using subsampled data.

In Fig. 2b,d,f and Extended Data Fig. 2a,f,g, the relative level of shared BCR/TCR VDJ sequences was computed by calculating the number of shared VDJ sequences between different metastatic sites in 10,000 subsampling operations and then computing the median of the number of overlaps across iterations. In Fig. 2e, the median Jaccard coefficient of shared VDJ sequences derived in the same 10,000 subsampling operations was used to generate BCR and TCR similarity

matrices, from which hierarchical clustering was performed to generate the BCR and TCR clonal similarity trees via the `hclust` function in R using the `ward.D2` agglomeration method. In the spatio-migratory maps of B cell clonal migration shown in Extended Data Fig. 2f,g, the clonal repertoire analyses for participants 308 and 315 were generated by calculating the median number of shared BCR clones across the same 10,000 subsampling operations.

### Clonal overlap correlations with tumor genomic landscape

The tumor phylogenetic trees were generated using OncoNEM<sup>44</sup>, as previously described by us<sup>16</sup>. The `hclust` (hierarchical clustering) function in the base R stats package was used to compute the BCR, TCR and genomic trees using the `ward.D2` agglomeration method. The comparison of the `hclust` objects was done using the `cophenetic` correlation, using the `cor_cophenetic` function from the `dendextend` package (version 1.15.2)<sup>61</sup>. A permutation test was used to calculate correlation one-sided  $P$  values, where the tree labels were randomly shuffled for 100 permutations, while keeping the tree topologies constant. The comparison of the BCR and TCR Jaccard clustering trees with the genetic trees was done by using the `cophenetic` definition for edge-weighted trees. In this version of the `cophenetic`, the distance between each pair of nodes is the sum of the weights of edges along the path connecting these pairs of nodes.

### BCR and TCR clonotype classification

In all participants with more than one tumor sampled (metastatic breast cancer cohort participants: 308, 315, 323, 330; early breast cancer cohort: all participants; Fig. 1), the clone proportion per sample was calculated by dividing the number of UMIs from each clone identified using MRDARCY with the total number of UMIs present in the sample. BCR clones were classified as stem, clade or private depending on whether they were observed in all, some or a single sample from the same participant, respectively (Extended Data Fig. 4a). Stem and clade clones were considered to be immunosurveillance given that they were present in more than one metastatic sample from a single participant.

We further refined the stem, clade and private clone classification by taking into account clone size (percentage of UMIs) to identify clonal expansion. We fitted a Gaussian mixture model to the log percentage UMI values of all BCR sequences of all early and metastatic breast cancer samples using the `MClust` (version 5.4.9)<sup>62</sup> R package to identify an overall BCR clone size cutoff threshold for expanded versus unexpanded clones. This threshold was set to ensure representation of all four clonal classes in all samples and that the expanded clones represented less than 10% of the total repertoire. Using this threshold, BCR clones were classified into four categories: (A) private and expanded, (B) shared and expanded, (C) private and unexpanded and (D) shared and unexpanded (Fig. 3a). Clones where clone size was above the cutoff threshold in some sites (that is, expanded) and below the threshold in others (that is, unexpanded) were classified as expanded.

### CDR3 probability of generation analysis

We calculated BCR CDR3  $P_{\text{gen}}$  as a result of VDJ recombination with OLGA<sup>25</sup> version 1.2.4 using as input the default human B cell heavy chain model and the amino acid CDR3 sequence of each BCR (Fig. 3c). In Fig. 3c,  $P_{\text{gen}}$  scores derived from BCR-sequencing data obtained from the peripheral blood mononuclear cells from a published healthy participant<sup>26</sup> are shown. Antigen-experienced BCRs were defined as those that were class switched (IgA, IgE, IgG) and had more than four somatic mutations. Antigen-inexperienced BCRs were defined as non-class-switched BCRs (IgD and IgM) with four or fewer mutations.

### Isotype usages and SHM across BCR clone classes

In Fig. 3d, the number of UMIs in each clone per IGH isotype were counted for each sample and summarized by summing the UMI counts by clone class (A, B, C, D) for each isotype/sample, resulting

in 192 sample/clone class combinations (48 samples  $\times$  4 BCR clone classes) for all 9 BCR isotypes. The total proportion of unswitched BCRs comprised the sum of the proportion of IgD and IgM UMI BCRs. In Fig. 3f, the total proportion of each IGH isotype across sequential samples obtained during therapy in early breast cancer and metastatic samples is shown. Statistical comparisons between early breast cancer time points were performed using an ordinal logistic regression to identify whether there was a monotonic association between IGH isotype proportion and time point. Statistical comparisons between early and metastatic breast cancer samples were performed using Wilcoxon rank-sum tests. In Extended Data Fig. 4g, the data plotted in Fig. 3f are subset across the four BCR clone classes.

BCRs were classified into four SHM categories (no, low, high and very high SHM) using the `normalmixEM` function from the `mixtools` R package (version 2.0.0), providing as input the log SHM count. The thresholds used were 0–1 mutation, 1–10 mutations, 11–33 mutations and >33 mutations for the no, low, high and very high SHM categories, respectively. In Extended Data Fig. 4d, the proportion of BCRs for each of the four SHM classes per sample is shown across the four BCR clone classes. In Fig. 3e, highly mutated BCRs were defined as those BCRs classified as having high and very high SHM counts. Statistical comparisons between early breast cancer time points were performed using an ordinal logistic regression to identify whether there was a monotonic association between the percentage of highly mutated BCRs and time point. Statistical comparisons between early and metastatic breast cancer samples were performed using Wilcoxon rank-sum tests.

In the analyses shown in Fig. 3g, all samples from all participants were used. The sample isotype usage was calculated by summing the total number of BCR UMIs per isotype per sample and then dividing this by the total number of UMIs within the sample, as described previously. The total proportion of unswitched BCR comprised the sum of the proportion of IgD and IgM BCRs. The mean sample BCR mutation count was calculated by first calculating the mean SHM per clone per sample, and then calculating the mean SHM per sample (so that larger clones are not overrepresented). Samples with high SHM and CSR were defined as those with a >50th percentile SHM and CSR, while those with low SHM and CSR were defined as those with a  $\leq$ 50th percentile SHM and CSR (Fig. 3g). In Fig. 3g, data from participants with more than one tumor site sampled are shown (early breast cancer cohort: all participants ( $n = 10$ ), metastatic breast cancer cohort participants: 308, 315, 323, 330), as classification into the four clonal groups required the sampling of more than one site/participant. In Fig. 3g, all samples from all participants are shown.

### BCR clonal expansion and diversification

We calculated BCR clonal expansion by first subsampling each tumor's BCR-sequencing data to 90% of the number of unique UMIs present in the sample with the lowest depth and summing the total number of UMIs associated with each unique BCR VDJ sequence. The Gini, Shannon index and mean clone sizes were calculated using the `ineq` R package (version 0.2–13), the `posterior` R package (version 1.4.1) and custom code, respectively. The mean of 1,000 iterations was used to calculate the final clonal expansion metrics (Fig. 4a and Extended Data Fig. 5a).

To calculate the per-site proportion of immunosurveillance clones (Fig. 4b), the total number of unique VDJ sequences per clone across all samples was calculated, and clones that were present in more than one site and had at least four unique VDJs in at least one metastatic site retained. The proportion of each of these clones across all samples was then calculated by dividing the total number of VDJs per clone per sample by the sum of the number of VDJs for that clone in all samples. The mean of these clone proportions per site was then calculated (Fig. 4b). In Extended Data Fig. 5b, the percentage of clones per sample that had at least four unique VDJ sequences were calculated.

### BCR clonal network analysis

Network clustering of BCR clones was performed using MRDARCY<sup>57</sup> in participants with more than one site sampled (metastatic breast cancer cohort: participants 308, 315, 323 and 330, early breast cancer cohort: all 10 participants). BCRs were clustered using a sequence identity threshold of 0.95, and clones that were present in a minimum of two tumor samples for each participant and had a minimum of ten unique BCR sequences were retained (number of clones retained in metastatic dataset: participant 308 = 204; participant 315 = 733; participant 323 = 85; participant 330 = 23).

For each BCR clone, the ends of the multiple sequence alignment were trimmed until 95% of all BCR sequences had an aligned nucleotide at the end of the sequence, with a minimum trimmed length of 80 nucleotides required for network clustering to be performed. A distance matrix was subsequently constructed for all sequences per clone, identical BCR sequences grouped together into clusters, and the abundance of these clusters across metastatic sites was calculated by dividing the total number of UMIs present in the cluster by the total number of UMIs in the sample being analyzed. BCR clone network diagrams were generated by computing the pairwise Hamming distances between sequences using the `phangorn`<sup>63</sup> R package (version 2.7.1), followed by neighbor-joining tree estimation and phylogenetic tree construction and optimization using the `pml` and `optim.pml` functions in `phangorn` (Fig. 4c,d and Extended Data Fig. 5c).

To calculate the degree of a BCR sequence, a minimum spanning tree was calculated on the Hamming distance matrix using the `mst` function in the `ape`<sup>64</sup> R package (version 5.6), which was then converted into an undirected graph using the `graph_from_adjacency_matrix` function in the `igraph` R package (version 1.2.10). The degree centrality was then computed using the `degree` function in `igraph` (Fig. 4c–f and Extended Data Figs. 5c–h).

We validated in our network clustering findings in four independent datasets (Extended Data Fig. 5i). Two metastatic breast cancer datasets (from the Hartwig Medical Foundation (HMF)<sup>27</sup> and the Rapid Autopsy tumor Donation program (RAP) at the UNC at Chapel Hill<sup>28</sup>) were identified and TRUST4 (ref. 65) was used to reconstruct the BCR immune receptor repertoires from the RNA-seq data, which were then processed using MRDARCY. Sixteen participants in the HMF dataset had breast tumor RNA-seq data for more than one metastatic deposit and clonotype assembly, and intra-participant comparison was only possible in one participant (participant ID: HMFN\_0320), which had a higher coverage (1,085 BCRs identified in one sample and 1,757 in another). Similarly, clonotype assembly and intra-participant comparison were possible in one participant in the RAP dataset (participant ID: 828433). BCR-sequencing data for diabetes<sup>29</sup> and a multiple sclerosis<sup>30</sup> datasets were downloaded from the iReceptor gateway<sup>66</sup> and processed using MRDARCY. Eight participants in the diabetes dataset and three participants in the multiple sclerosis dataset had multisite BCR-sequencing data for which clonotype assembly and intra-participant comparisons were possible.

We have created and uploaded an R framework hosted at <https://github.com/sjslab/BCR-ImmunoSurveillance> to generate network clustering of BCR clones and compute the centrality analyses from BCR repertoire data derived from BCR sequencing, as well as BCR repertoire data obtained from bulk RNA-seq data.

To determine the predictability of immunosurveillance clones based on BCR degree in the early and metastatic breast cancer cohorts (Fig. 4f and Extended Data Fig. 5g,h), we calculated the sensitivity, specificity and accuracy of a classification that categorizes BCRs as immunosurveillance or not based on a series of degree cutoffs (>1, >2 >10). Model performance metrics were generated using the `confusionMatrix` function in the `caret` (version 6.0–90) R package.

### Reporting summary

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

## Data availability

Sequence data (aligned to the GRCh37 of the human genome) have been deposited in the European Genome-phenome Archive (EGA), which is hosted by the EBI and the CRG, under the accession codes [EGAS00001002703](https://ega-archive.org/studies/EGAS00001002703) (tumor DNA and RNA) and <https://ega-archive.org/studies/EGAS50000000241> (BCR-sequencing data). Example processed data are available at <https://github.com/sjslab/BCR-Immunosurveillance/>.

## Code availability

Our R framework for the network clustering and centrality analyses of BCR repertoire data derived from BCR-sequencing or bulk RNA-seq data is made available to accelerate the identification of potential immunosurveillance and clonally persistent antibodies (<https://github.com/sjslab/BCR-Immunosurveillance>). The R source code used to run the analyses and generate the figures shown in this paper is also available at this repository.

## References

44. Ross, E. M. & Markowitz, F. OncoNEM: inferring tumor evolution from single-cell sequencing data. *Genome Biol.* **17**, 69 (2016).
45. Kawaguchi, S., Higasa, K., Shimizu, M., Yamada, R. & Matsuda, F. HLA-HD: an accurate HLA typing algorithm for next-generation sequencing data. *Hum. Mutat.* **38**, 788–797 (2017).
46. Racle, J. et al. Robust prediction of HLA class II epitopes by deep motif deconvolution of immunopeptidomes. *Nat. Biotechnol.* **37**, 1283–1286 (2019).
47. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
48. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
49. Ayers, M. et al. IFN- $\gamma$ -related mRNA profile predicts clinical response to PD-1 blockade. *J. Clin. Invest.* **127**, 2930–2940 (2017).
50. Hänzelmann, S., Castelo, R. & Guinney, J. GSEA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* **14**, 7 (2013).
51. Liberzon, A. et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).
52. Galson, J. D. et al. Deep sequencing of B cell receptor repertoires from COVID-19 patients reveals strong convergent immune signatures. *Front. Immunol.* **11**, 605170 (2020).
53. Vander Heiden, J. A. et al. pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. *Bioinformatics* **30**, 1930–1932 (2014).
54. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
55. Ye, J., Ma, N., Madden, T. L. & Ostell, J. M. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res.* **41**, W34–W40 (2013).
56. Lefranc, M. -P. et al. IMGT, the international ImMunoGeneTics information system. *Nucleic Acids Res.* **37**, D1006–D1012 (2009).
57. Bashford-Rogers, R. J. M. et al. Eye on the B-ALL: B-cell receptor repertoires reveal persistence of numerous B-lymphoblastic leukemia subclones from diagnosis to relapse. *Leukemia* **30**, 2312–2321 (2016).
58. Dunbar, J. et al. SAbDab: the structural antibody database. *Nucleic Acids Res.* **42**, D1140–D1146 (2014).
59. Swindells, M. B. et al. abYsis: integrated antibody sequence and structure-management, analysis, and prediction. *J. Mol. Biol.* **429**, 356–364 (2017).
60. Vita, R. et al. The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res.* **47**, D339–D343 (2019).
61. Galili, T. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics* **31**, 3718–3720 (2015).
62. Scrucca, L., Fop, M., Murphy, T. B. & Raftery, A. E. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *R J.* **8**, 289–317 (2016).
63. Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592–593 (2011).
64. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).
65. Song, L. et al. TRUST4: immune repertoire reconstruction from bulk and single-cell RNA-seq data. *Nat. Methods* **18**, 627–630 (2021).
66. Corrie, B. D. et al. iReceptor: a platform for querying and analyzing antibody/B-cell and T-cell receptor repertoire data across federated repositories. *Immunol. Rev.* **284**, 24–41 (2018).

## Acknowledgements

The authors thank the late Nir Friedman at the Weizmann Institute for the many scientific discussions that contributed to the central idea that rooted this work. S.-J.S. was supported by a Whitney Wood Scholarship awarded by the Royal College of Physicians (United Kingdom). C.C. was supported by funding from CRUK (grant numbers A17197, A27657 and A29580), an NIHR Senior Investigator Award (grant number NF-SI-0515-10090), and a European Research Council Advanced Award (grant number 694620). R.J.M.B.-R. was supported by the Wellcome Trust and University of Oxford. O.M.R. was supported by the NIHR Cambridge Biomedical Research Centre (BRC-1215-20014) and the Medical Research Council (UK; MC\_UU\_00002/16). B.S. is supported by an NIHR Academic Clinical Lectureship (CL-2021-13-002), Academy of Medical Sciences (SGL028\1074) and The British Medical Association Vera Down Award. We thank Breast Cancer Now for funding this work as part of Programme Funding to the Breast Cancer Now Toby Robins Research Centre. We thank the Asociación Española contra el Cáncer, Cellex foundation, and the clinical team at the Breast Cancer Unit of Vall d'Hebron University Hospital/Institute of Oncology and the Cambridge Breast Cancer Research Unit for facilitating the collection and processing of biological samples. We are very grateful for the generosity of all the participants that donated samples for analysis.

## Author contributions

S.-J.S., R.J.M.B.-R. and C.C. conceived the study, led data analysis and wrote the paper. Sample collections were led by L.D.M.-A., J.S. and S.-J.S. BCR library preparation was performed by S.S. and J.D. with input from J.O., R.M. and D.K.F. J.D.G. contributed to the sequence processing, BCR sequence preprocessing and alignment. S.-J.S. and R.J.M.B.-R. performed the analyses. S.-F.C. helped with genomic and transcriptomic data generation and data analysis. O.M.R. provided statistical oversight. B.S. created the pathogen antibody database. All authors read and approved the manuscript.

## Competing interests

R.J.M.B.-R. is a cofounder of Alchemab Therapeutics and consultant for Alchemab Therapeutics, GSK, Roche, EnaraBio and UCB. S.S., J.D., J.O., R.M., D.K.F. and J.D.G. are, or were, employed by Alchemab Therapeutics. J.S. is a cofounder of Mosaic Biomedicals and has ownership interests in Mosaic Biomedicals and Northern Biologics. J.S. received grant/research support from Mosaic Biomedicals, Northern Biologics, Roche/Glycart, Hoffmann-La Roche and AstraZeneca. The other authors declare no competing interests.

## Additional information

**Extended data** is available for this paper at <https://doi.org/10.1038/s41590-024-01821-0>.

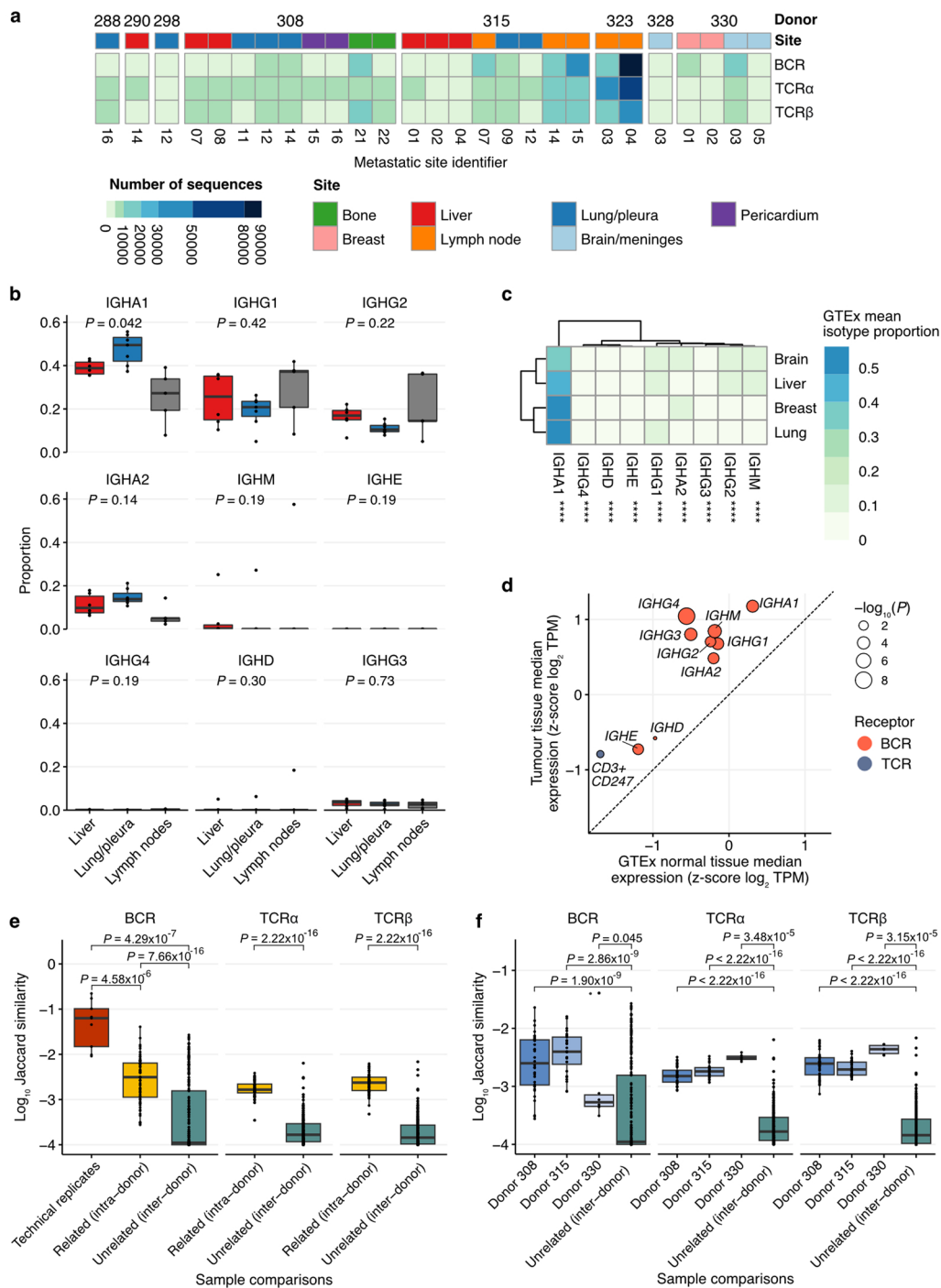


**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41590-024-01821-0>.

**Correspondence and requests for materials** should be addressed to Stephen-John Sammut, Carlos Caldas or Rachael J. M. Bashford-Rogers.

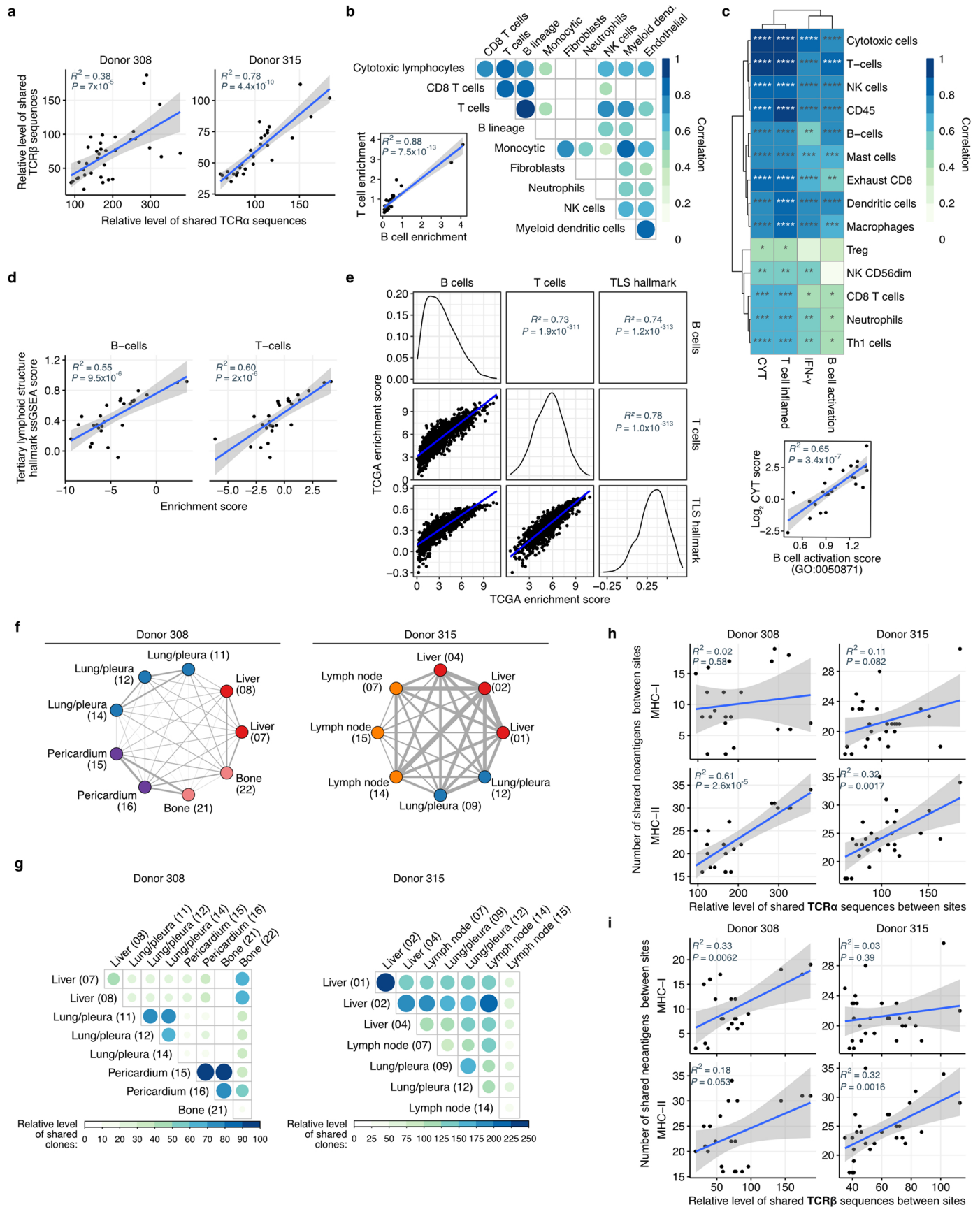
**Peer review information** *Nature Immunology* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editor: N. Bernard, in collaboration with the *Nature Immunology* team.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).



**Extended Data Fig. 1 | BCR sequencing quality-control statistics. a)** Heatmap showing number of unique BCR and TCR sequences identified across 27 metastatic sites sampled from 8 patients. **b)** Box plots showing distribution of BCR isotype usages by metastatic site ( $n = \text{liver: 6, lymph nodes: 5, lung/pleura: 7}$ ).  $P$  values were calculated using Kruskal-Wallis test and adjusted for multiple comparisons. **c)** Heatmap showing BCRIGH isotype usage across  $n = 3,905$  healthy tissue samples in GTEx ( $****P < 6 \times 10^{-16}$ ).  $P$  values were calculated using Kruskal-Wallis test and adjusted for multiple comparisons. **d)** Scatter plot showing expression of BCR isotypes and CD3/CD247 in tumour versus healthy tissues in GTEx. **e)** Box plots

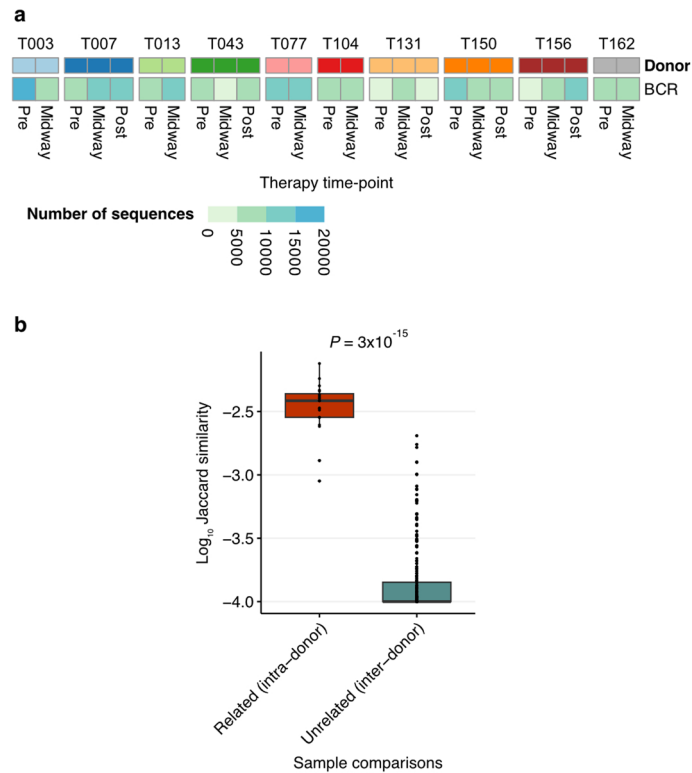
showing the  $\log_{10}$  Jaccard BCR and TCR similarity between technical replicates ( $n = 9$  comparisons), related samples derived from the same patient ( $n = 71$  comparisons) and unrelated samples ( $n = 208$  comparisons). **f)** Box plots showing the  $\log_{10}$  Jaccard BCR and TCR similarity between samples obtained from patients 308 ( $n = 36$  comparisons), 315 ( $n = 28$  comparisons) and 330 ( $n = 6$  comparisons), as well as unrelated samples ( $n = 208$  comparisons). **d–f)** Wilcoxon rank sum tests, all  $P$  values two-sided. **b, e, f)** The box bounds the interquartile range divided by the median, with the whiskers extending to a maximum of 1.5 times the interquartile range beyond the box. Individual data points shown as dots.



Extended Data Fig. 2 | See next page for caption.



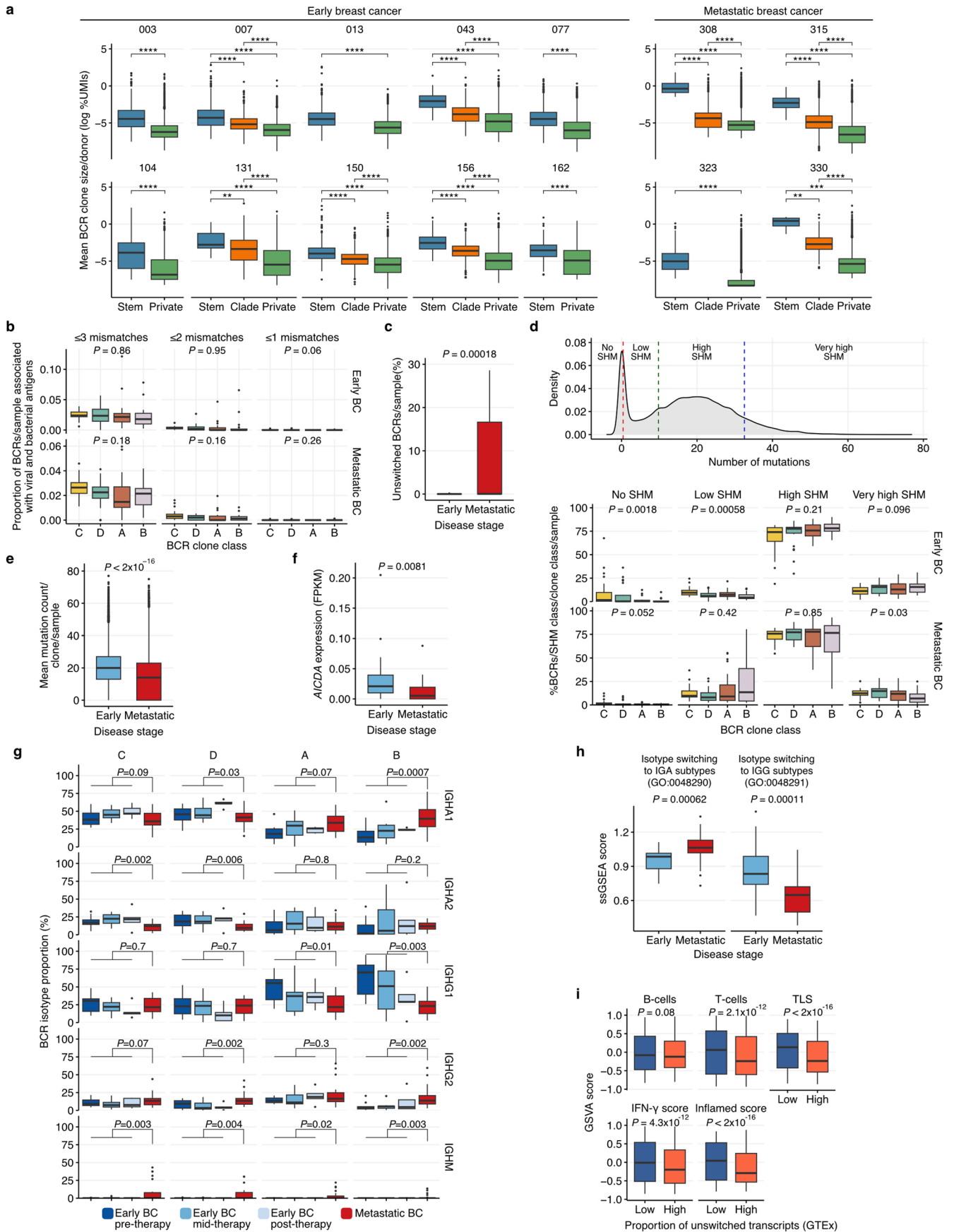
**Extended Data Fig. 2 | Adaptive immune and tumour co-evolution.** **a)** Scatter plots showing relationship between shared TCR $\alpha$  and TCR $\beta$  VDJ sequences across sampled metastatic sites (inter-sample comparisons: patient 308  $n = 36$ , patient 315:  $n = 28$ ). TCR sequences down sampled.  $P$  value and  $R^2$  obtained from linear regression analysis. **b)** Correlation plot showing relationship between tumour immune microenvironment components deconvoluted from the bulk RNA-Seq data using MCPcounter. Inset: scatter plot showing relationship between T and B cell enrichment.  $P$  value and  $R^2$  obtained from linear regression. **c) Top:** heatmap showing Pearson's correlations between tumour immune microenvironment composition (obtained using Danaher gene sets) and activity. Enrichment scores obtained using bulk RNA-Seq data. \*\*\*\* $P < 0.0001$ , \*\*\* $P < 0.001$ , \*\* $P < 0.01$ , \* $P < 0.05$ . **Bottom:** scatter plot showing correlation between B cell activation and cytolytic activity (CYT).  $P$  value and  $R^2$  obtained from linear regression analysis. **d)** Scatter plots showing relationship between B and T cell enrichment and expression of a tertiary lymphoid structure gene set.  $P$  value and  $R^2$  obtained from linear regression analysis. **e)** Relationship between B cell, T cell and tertiary lymphoid structure (TLS) hallmark signature in the TCGA breast cancer cohort ( $n = 1083$  tumours).  $P$  value and  $R^2$  obtained from linear regression analysis. **f)** Spatio-migratory map of B cell clonal migration between metastatic sites. Edge width proportional to relative number of shared BCR clones between sites. **g)** Heatmaps showing relative number of shared BCR clones between sites. **h)** Scatter plots showing relationship between shared TCR $\alpha$  VDJ sequences and predicted MHC class I and II neoantigens across pairwise metastatic sites. **i)** Scatter plots showing relationship between shared TCR $\beta$  VDJ sequences and predicted MHC class I and II neoantigens across pairwise metastatic sites. **b–d)** Data from all sites ( $n = 27$ ) from all patients used. **a–d, h, i)** The shaded area, in grey, represents the 95% confidence interval. **a, f–i)** Data from two patients with more than four metastatic sites sampled used (308, 315). **h, i)** TCR sequences were downsampled.  $P$  value and  $R^2$  obtained from linear regression analysis. Inter-sample comparisons: patient 308  $n = 36$ ; patient 315:  $n = 28$ .



**Extended Data Fig. 3 | Early breast cancer cohort BCR sequencing metrics.**

**a)** Heatmap showing number of unique BCR sequences identified across 25 tumour biopsies sampled from 10 patients. **b)** Box plot showing the log<sub>10</sub> Jaccard BCR similarity between related samples derived from the same patient

( $n = 20$  comparisons) and unrelated samples ( $n = 280$  comparisons). Wilcoxon rank sum test,  $P$  value two-sided. The box bounds the interquartile range divided by the median, with the whiskers extending to a maximum of 1.5 times the interquartile range beyond the box. Individual data points shown as dots.

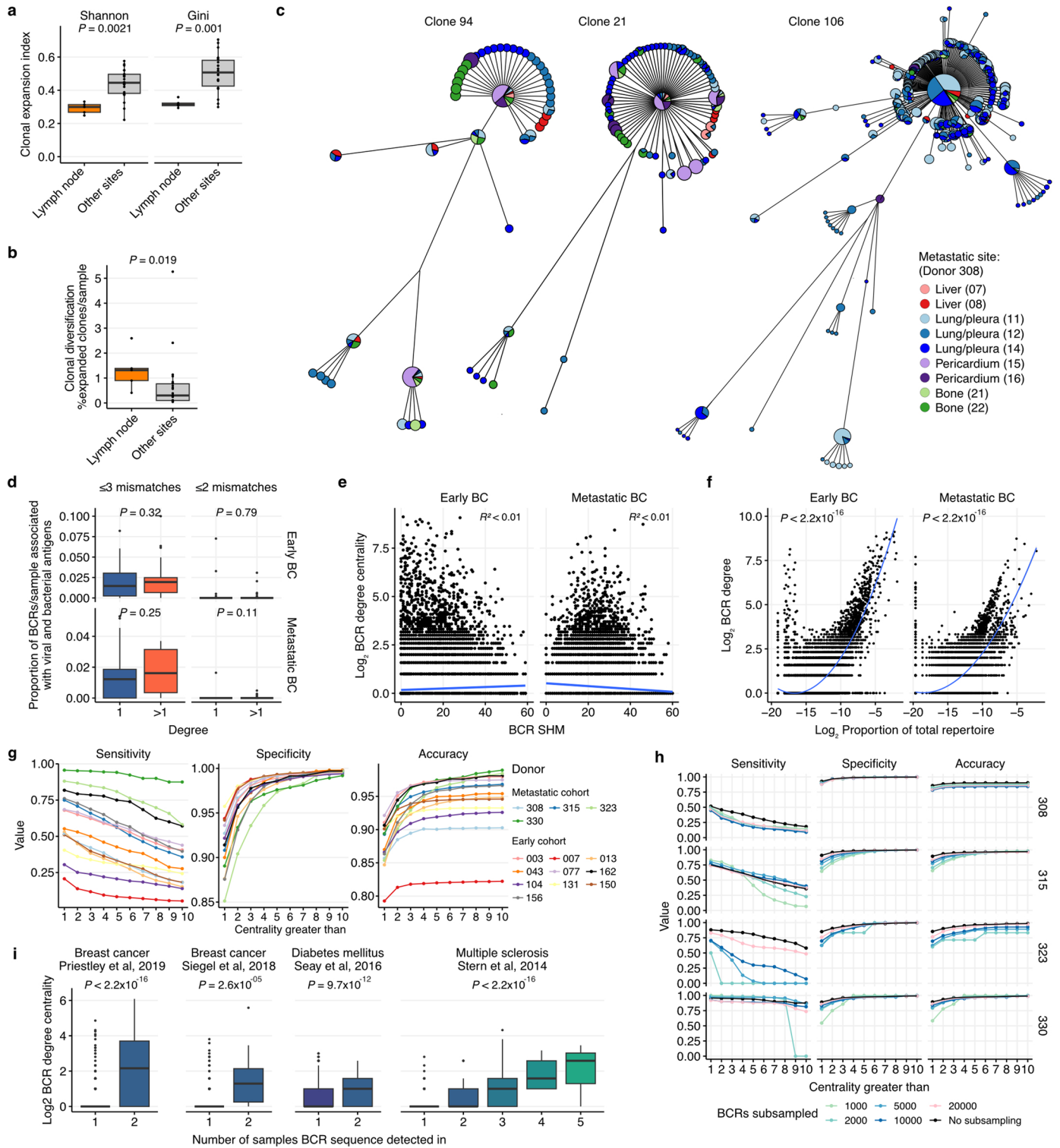


Extended Data Fig. 4 | See next page for caption.



**Extended Data Fig. 4 | BCR clone classification analyses.** **a)** Boxplots showing distribution of BCR clone sizes in stem, clade, and private BCR clones across early (n = 4487 stem, 4569 clade, 85439 private unique BCR clones) and metastatic cancer (n = 1268 stem, 12022 clade, 142161 private unique BCR clones) cohorts. \*\*\*\* $P < 4.9 \times 10^{-5}$ , \*\*\* $P = 0.00099$ , \*\* $P = 0.002$ , two-sided Wilcoxon Rank sum test and adjusted for multiple comparisons. **b)** Boxplots showing proportion of unique BCRs/sample with CDR3 sequences matching a reference antibody database with known binding to viral or bacterial antigens with  $\leq 3$ ,  $\leq 2$  and  $\leq 1$  CDR3 mismatches, across the four BCR clone classes.  $P$  values calculated using two-sided analysis of variance (ANOVA). **c)** Boxplot showing the percentage of unswitched BCRs per sample in early and metastatic breast cancer cohorts.  $P$  values calculated two-sided Wilcoxon Rank sum test. **d) Top:** density plot showing distribution of BCRSHM and thresholds used to classify BCRs in four SHM classes. **Bottom:** Boxplots showing the distribution of BCRs proportions in four SHM classes across the four BCR clone classes in early and metastatic breast cancer cohorts.  $P$  values calculated using Kruskal-Wallis test. **e)** Boxplot showing the mean mutation count per BCR clone per sample in early and

metastatic breast cancer cohorts. **f)** Boxplot showing expression of *AICDA* in early and metastatic breast cancer cohorts. **g)** Boxplots showing % isotype usage in early and metastatic breast cancer samples across the four BCR clone classes. Wilcoxon rank sum tests,  $P$  values adjusted and two-sided. **h)** Boxplots showing distribution of enrichment scores of two RNA isotype switching signatures. **i)** Boxplots showing distribution of immune microenvironment cell type scores and activation signatures across n = 3,905 healthy tissue samples in GTEx. Samples with high expression of unswitched transcripts were defined as those with a >50th percentile expression of IGHD/M genes. **a, b, d, g)** Data from patients with more than one tumour site sampled shown (early breast cancer cohort: n = 10 patients, 25 samples; metastatic breast cancer cohort patients: 308, 315, 323, 330, n = 23 metastatic cancer samples) used. **c, e, f, h)** Data from all patients used (early breast cancer cohort: n = 10 patients, 25 samples; metastatic breast cancer cohort n = 8 patients, 27 samples) used. **a–i)** The box bounds the interquartile range divided by the median, with the whiskers extending to a maximum of 1.5 times the interquartile range beyond the box. Individual data points shown as dots. **a, c, e–i)** Wilcoxon rank sum tests, all  $P$  values two-sided.



Extended Data Fig. 5 | See next page for caption.

**Extended Data Fig. 5 | BCR degree centrality analyses.** **a)** Boxplots showing clonal expansion as measured by the Shannon and Gini indices in lymph nodes versus other sites. **b)** Boxplots showing clonal diversity, as measured by the percentage of expanded BCR clones ( $\geq 4$  unique VDJs) in lymph nodes versus other sites. **c)** BCR VDJ network plots showing three expanded immunosurveillance clones shared between multiple sites in patient 308. **d)** Boxplots showing proportion of unique BCRs/sample with CDR3 sequences matching a reference antibody database with known binding to viral or bacterial antigens with  $\leq 3$  and,  $\leq 2$  CDR3 mismatches, across BCRs with degree centrality = 1 and  $> 1$ . **e)** Scatter plots showing lack of association between BCR degree centrality and BCR SHM in early and metastatic breast cancer cohorts. **f)** Scatter plots showing association between BCR degree centrality and proportion of total repertoire in early and metastatic breast cancer cohorts. Two-sided *P* value derived from polynomial regression. **g)** Profile plots showing changes in sensitivity, specificity, and accuracy at identifying immunosurveillance

BCRs at different degree centrality thresholds in early and metastatic breast cancer cohorts. **h)** Profile plots showing changes in sensitivity, specificity and accuracy at identifying immunosurveillance BCRs at different centrality thresholds in four patients with metastatic breast cancer at five subsampling depths (1000, 2000, 5000, 10000 and 20000). **i)** Boxplots showing association between BCR degree centrality and the number of sites in which the BCR is observed in four external datasets (number of cases: *n* = 1 Priestley et al, *n* = 1 Siegel et al, *n* = 8 Seay et al, *n* = 3 Stern et al). Kruskal-Wallis tests, all *P* values two-sided. **a, b)** Data from four patients with more than one metastatic site sampled (308, 315, 323, 330) used (*n* = 5 lymph node, *n* = 22 other sites). **d–f)** Data from four patients with more than one metastatic site sampled (patients: 308, 315, 323, 330) and all patients (*n* = 10) with early breast cancer. **a, b, d, i)** The box bounds the interquartile range divided by the median, with the whiskers extending to a maximum of 1.5 times the interquartile range beyond the box. Individual data points shown as dots. **a, b, d)** Wilcoxon rank sum tests, all *P* values two-sided.



## 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 Clinical data was collected in Microsoft Excel (as part of the Office 365 suite).

Data analysis

List of software used:

- HLA-HD: version 1.4
- HTSeq: version 0.6.1p1
- IgBlast: version 1.14.0
- IMGT/HighV-QUEST: version 1.8.5
- Immcantation framework: docker container v3.0.0
- mixMHC2pred: version 1.2
- MRDARCY <https://github.com/Bashford-Rogers-lab/MRDARCY>
- OLGA: version 1.2.4 <https://github.com/statbiophys/OLGA>
- OncoNEM [https://bitbucket.org/edith\\_ross/onconem/src/master/](https://bitbucket.org/edith_ross/onconem/src/master/)
- STAR: version 2.5.2b
- TRUST4: version 1.0.11

R version 4.1.2 and associated packages:

- ape: version 5.6
- caret: version 6.0-90
- corrplot: version 0.92
- dendextend: version 1.15.2

- GSVA: version 1.38.2
- igraph: version 1.2.10
- MClust: version 5.4.9
- MCPcounter: version 1.2.0
- phangorn: version 2.7.1

Python version 3.10.1

The code to recreate analyses described in this manuscript has been uploaded to <https://github.com/sjslab/BCR-Immunosurveillance>

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](#)

Sequence data (aligned to the GRCh37 of the human genome) have been deposited at the European Genome-phenome Archive (EGA), which is hosted by the EBI and the CRG, under accession number EGAS00001002703 (Tumour DNA and RNA) and EGA00002343328 (BCR sequencing data). Once approval from the data access committee is secured processed data will be provided through direct communication with corresponding authors. Example processed data is available at <https://github.com/sjslab/BCR-Immunosurveillance>.

## Human research participants

Policy information about [studies involving human research participants and Sex and Gender in Research](#).

Reporting on sex and gender	<a href="#">Eight women with metastatic breast cancer and ten women with early breast cancer undergoing neoadjuvant therapy were analysed in this study.</a>
Population characteristics	<p>Eight women analysed in this study had lethal metastatic breast cancer: their detailed clinical information, including treatments, have been previously published: <a href="https://doi.org/10.1016/j.celrep.2019.04.098">https://doi.org/10.1016/j.celrep.2019.04.098</a></p> <p>Ten women analysed in this study had early breast cancer and were treated with preoperative therapies: their detailed clinical information, including treatments, have been previously published: <a href="https://doi.org/10.1038/s41586-021-04278-5">https://doi.org/10.1038/s41586-021-04278-5</a></p>
Recruitment	Eight patients with metastatic breast cancer who underwent post-mortem warm autopsies were included in this study. All patients were enrolled as part of the Vall d'Hebron Institute of Oncology (VHIO) Warm Autopsy Program. All ten women with early breast cancer were enrolled to the TransNEO study at Cambridge University Hospitals NHS Foundation Trust.
Ethics oversight	<p>Metastatic breast cancer cohort: research autopsies were performed under VHIO Warm Autopsy Program protocols approved by the institutional review board (IRB) of Vall d'Hebron University Hospital (Barcelona, Spain).</p> <p>Early breast cancer cohort: East of England Research Ethics Committee: 12/EE/0484.</p>

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](https://www.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	8 women with lethal metastatic breast cancer and 10 women with early breast cancer were recruited to this study. No statistical methods were used to pre-determine sample sizes but our sample sizes are similar to those reported in previous publications.
Data exclusions	No data was excluded. BCR sequencing was performed in cases which had remaining RNA extracted.
Replication	For two of the metastatic tumour samples, two replicate BCR libraries were created, and for three of the metastatic samples, three replicate BCR libraries were created. High levels of BCR VDJ sharing were observed in technical replicates. All attempts at replication were successful as shown in Extended Data Figure 1e.
Randomization	Randomization not applicable - all cases were treated with standard of care therapy regimens.
Blinding	Blinding not applicable - no group allocations.

## 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	Involvement in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

### Methods

n/a	Involvement in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging