



# Multi-cohort high-dimensional proteomics reveals early risk markers for lymphoid cancer subtypes

Received: 5 June 2025

Accepted: 19 September 2025

Published online: 28 October 2025

 Check for updates

P. Martijn Kolijn<sup>1,2,3</sup>, Karl Smith-Byrne<sup>4</sup>, Vernon Burk<sup>5</sup>, Vivian Viallon<sup>6</sup>, Matthew A. Lee<sup>6,7</sup>, Keren Papier<sup>4</sup>, Ziqiao Wang<sup>8</sup>, Anton W. Langerak<sup>3</sup>, Florentin Späth<sup>9</sup>, Arjan Diepstra<sup>10</sup>, Christina M. Lill<sup>11,12</sup>, Raul Zamora-Ros<sup>13</sup>, Alessandra Macciotta<sup>14</sup>, Amaia Aizpurua<sup>15,16</sup>, Rosario Tumino<sup>17</sup>, Nilanjan Chatterjee<sup>8</sup>, Ruth C. Travis<sup>4</sup>, Marc J. Gunter<sup>6,18</sup>, Elizabeth A. Platz<sup>5</sup>, Elio Riboli<sup>18</sup>, James McKay<sup>6</sup> & Roel C. H. Vermeulen<sup>1,2</sup> 

This study aims to investigate the early stages of lymphoid malignancy pathogenesis and identify pre-diagnostic proteomic markers for lymphoma. Using the SomaScan-7K platform, we analyze 6412 unique plasma proteins in a case-cohort study nested within the European Prospective Investigation into Cancer and Nutrition (EPIC) cohort, comprising 4565 participants (484 incident lymphoid malignancy cases, median follow-up 9 years). We identify over 500 unique protein-lymphoid malignancy associations. Enriched pathways include viral protein interactions, cytokine signaling, B-cell receptor signaling, and NF- $\kappa$ B activation, reflecting key mechanisms in lymphoma pathogenesis. Cross-cohort validation of the top 20 FDR-significant proteins reveals concordant nominal significance for 70%–95% of the associations in the UK Biobank (Olink) and ARIC (SomaScan) studies. Time-stratified analyses reveals that a subset of these protein-lymphoma associations is evident over a decade before diagnosis. These findings highlight the potential of circulating proteomic markers in risk stratification, early diagnosis, and targeted prevention strategies for lymphoid malignancies.

Lymphoid malignancies comprise a heterogeneous group of cancers varying in etiology, incidence, and survival<sup>1–8</sup>. Notably, several lymphoid malignancy subtypes are preceded by precursor conditions detectable years before overt malignancy. Examples of precursor conditions include: monoclonal B cell lymphocytosis (MBL) preceding chronic lymphocytic leukemia (CLL), circulating t(14;18)-positive B-cells preceding follicular lymphoma (FL), non-IgM isotype monoclonal gammopathy of undetermined significance (MGUS) preceding multiple myeloma (MM,) and IgM isotype MGUS preceding Waldenström macroglobulinemia<sup>9–13</sup>. Studying the early development of lymphoid neoplasms holds the potential to reveal key insights into the etiology of these precursor conditions

and the biological drivers contributing to progression to overt malignancy.

Circulating proteins play an essential role in both the early pathogenesis of lymphoid malignancies and the immunological response to these neoplasms<sup>14,15</sup>. Proteomic studies have identified prediagnostic changes in cytokines and B-cell activation markers, with longitudinal analyses confirming alterations in markers such as sCD23 (FCER2), sCD27, sCD30, and CXCL13 up to two decades before lymphoma diagnosis<sup>16–28</sup>. Several large studies, including the UK Biobank and Uppsala-Umeå Comprehensive Cancer Consortium (U-CAN), have employed high-throughput proteomics, analyzing panels ranging from 1463 to 2963 proteins<sup>29,30</sup>. These studies identified hundreds of

protein-cancer associations and reproduced established lymphoma-protein associations, such as soluble B-cell maturation antigen (sBCMA) with MM, sCD23 with CLL, and CXCL13 with diffuse large B-cell lymphoma (DLBCL). However, these studies were constrained by the relatively limited number of proteins assessed (1463–2963) and the absence of cross-cohort validation across different proteomic platforms.

In this work, we conduct high-throughput proteomic profiling of 6412 circulating proteins using the SomaScan 7K Assay<sup>®</sup> in a case-cohort study nested within the European Prospective Investigation into Cancer and Nutrition (EPIC) cohort and perform validation analyses in the UK Biobank (Olink platform) and the Atherosclerosis Risk in Communities (ARIC) study (SomaScan-5K assay).

## Results

Baseline characteristics of the 4565 EPIC participants are shown in Table 1 and Supplementary Table 1. The 484 incident lymphoid malignancy cases were diagnosed a median of 9 years (range 0.1–19 years) after blood collection. Compared to the sub-cohort, participants who later developed lymphoid malignancy were on average older and more likely to be male, consistent with previous reports (Table 1)<sup>31</sup>.

### Proteomic Associations with B-cell lymphoma

In our fully-adjusted models for B-cell lymphoma (BCL,  $n = 330$ ), we observed 157 unique protein-BCL associations (173 aptamers), of which 139 were associated with an increased risk and 18 were associated with a decreased risk of BCL (Fig. 1B, Supplementary Data 1). BCL development was associated with increased plasma levels of members of several major immunomodulatory protein families, including the FC-receptor family, the semaphorin family, the TNF-receptor and TNF-ligand superfamily, the leukocyte immunoglobulin-like receptor family, the interleukins, and a selection of chemokines promoting migration of activated lymphocytes (Fig. 1B, Supplementary Data 1). Pathway enrichment analysis for the 157 proteins associated with BCL revealed an enrichment of proteins associated with processes such as cytokine and chemokine signaling, B-cell receptor (BCR) signaling, the NF- $\kappa$ B signaling pathway, hematopoietic cell lineage, antigen processing and presentation, and N-glycan biosynthesis (Fig. 1C, Supplementary Data 3).

### Time-stratified analyses for B-cell lymphoma

In time-stratified analyses of 151 BCL cases with an early blood sample who were diagnosed more than 10 years after blood draw, we observed 20 significant protein-BCL associations (Supplementary Fig. 3A, Supplementary Data 2). Early markers for BCL development included proteins associated with BCR signaling and the Fc receptor family, such as FCMR and sCD23. For 106 BCL cases with blood samples collected between 5 to 10 years before diagnosis, we also observed 29 significant

protein-BCL associations (Supplementary Fig. 3B, Supplementary Data 2). The top associated proteins included CXCL13, FDCSP, SELL, and several members of the Fc-receptor and semaphorin families. For the 73 BCL cases with blood samples collected less than 5 years prior to diagnosis, we observed 201 significant protein-BCL associations (Supplementary Fig. 3C, Supplementary Data 2). Pathway analysis revealed an enrichment of the cytokine, chemokine, BCR, and NF- $\kappa$ B signaling pathways (Supplementary Fig. 3D, Supplementary Data 3). In total, nine proteins showed consistent associations with BCL across all time intervals from >10 years to <5 years before diagnosis (Supplementary Fig. 3E). Trajectory analyses for the top 10 proteins associated with BCL ranked by FDR-adjusted P-value revealed increasing hazard ratios for FCMR, sCD23, FCRL1, FCRL3, CXCL13, CD72, SEMA7A and SEMA4A in blood samples drawn within 5 to 10 years before diagnosis, compared to participants with blood samples drawn over 10 years before diagnosis (Fig. 2A). The hazard ratios for FCMR, sCD23, FCRL3, CXCL13, CD72, SEMA7A and SEMA4A then reached a plateau, while the hazard ratio for FCRL1 continued to increase linearly among individuals with a blood draw within 5 years of BCL diagnosis. Hazard ratios for SLAMF6 and CD28 remained relatively stable across all time intervals from >10 years to <5 years before diagnosis.

### B-cell malignancy subtype specific analyses for CLL, DLBCL, FL and MM

We additionally performed separate analyses for the most common B-cell malignancy subtypes: CLL ( $n = 80$ ), DLBCL ( $n = 80$ ), FL ( $n = 51$ ), and MM ( $n = 116$ ) (Supplementary Data 1). We observed significant heterogeneity among the top 10 associated proteins for each subtype (Fig. 3). While some proteins, such as sCD23, CD28, CD72, and FCRL3, were associated across multiple B-cell malignancy subtypes, others, such as sBCMA, SLAMF7, and IGSF3, were specifically associated with a particular B-cell malignancy subtype. Strikingly, all the top proteins associated with B-cell lymphoma development (sCD23, FCMR, FCRL1, FCRL3, SELL, SEMA7A, and SEMA4A) were most strongly associated with CLL development, suggesting CLL was a major driver of the top protein associations observed in the grouped analysis.

### CLL

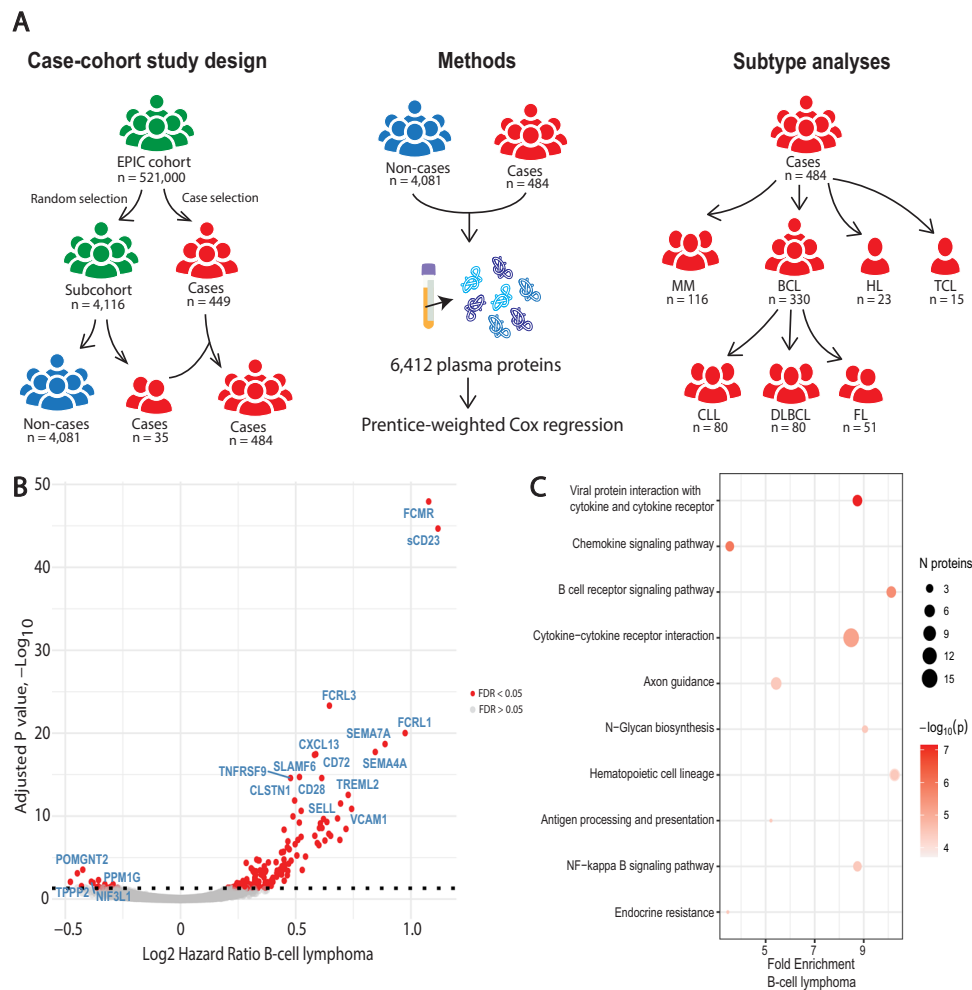
We observed 589 protein-CLL associations (617 aptamers), of which 544 were associated with an increased risk and 45 were associated with a decreased CLL risk (Supplementary Fig. 4A, Supplementary Data 1). CLL development was associated with enrichment of proteins associated with the spliceosome, ribosome, base excision repair, viral carcinogenesis, p53 signaling and the polycomb repressive complex (Supplementary Fig. 4D, Supplementary Fig. 5D). Time-stratified analyses revealed an exponential increase in the number of FDR-significant CLL-protein associations among participants with a blood

**Table 1 | Descriptive table**

	Total cases N = 484	BCL N = 330	MM N = 116	HL N = 23	TCL N = 15	Sub-cohort N = 4116
BMI, mean (Q1–Q3)	26.9 (23.8–29.1)	26.8 (23.8–28.9)	27.2 (24.2–29.8)	26.2 (25.5–28.11)	26.5 (23.5–28.9)	26.9 (23.9–29.5)
Age in years, mean (Q1–Q3)	55.4 (49.6–61.6)	55.5 (49.8–62.0)	55.1 (51.7–61.3)	49.1 (43.7–55.2)	55.5 (47.8–62.9)	51.4 (44.8–57.9)
Years blood draw to diagnosis, mean (Q1–Q3)	8.8 (5.2–12.7)	8.9 (5.3–12.9)	9.0 (4.9–12.6)	6.1 (2.5–8.4)	9.3 (5.8–13.5)	-
Sex N (%)						
Male	233 (48.2)	153 (46.4)	56 (48.3)	17 (74)	7 (46.7)	1573 (38.2)
Female	250 (51.8)	177 (53.6)	60 (51.7)	6 (26)	8 (53.3)	2543 (61.8)
Smoking status N (%)						
Non-smoker	235 (48.7)	168 (50.9)	53 (45.7)	6 (26.1)	8 (53.3)	2098 (51)
Former smoker	149 (30.8)	101 (30.6)	38 (32.8)	7 (30.4)	3 (20)	984 (23.9)
Current smoker	99 (20.5)	61 (18.5)	25 (21.5)	10 (43.5)	4 (26.7)	1034 (25.1)

BCL B-cell lymphoma, MM multiple myeloma, HL Hodgkin lymphoma, TCL T-cell lymphoma, BMI body mass index, Q quantile.

"Years blood draw to diagnosis" refers to the number of years between the moment of blood draw and lymphoid malignancy diagnosis. BMI and age are listed at recruitment.



**Fig. 1 | Protein associations with B-cell lymphoma. A** Graphical abstract of the study describing the study design, experimental approach and subtype analyses. For further details on the subtype grouping see Supplementary Table 2. **B** Volcano plot of the hazards ratio and FDR-adjusted P-value as determined through two-sided Prentice-weighted Cox regression models for the risk of BCL for all aptamers included in the study (330 BCL cases and 4088 non-cases were included in this

analysis). **C** Pathway enrichment analysis results for all 157 FDR-significant hits for BCL, top 10 pathways ranked by P-value are shown. Pathway enrichment analyses are performed via one-sided hypergeometric testing. P-values are adjusted for multiple comparisons using the Bonferroni method. For further details on the pathway analysis results, see Supplementary Data 3. Source data are provided as a Source Data file.

draw collected less than 5 years prior to diagnosis (1094 proteins), compared to CLL cases with a blood draw collected between 5 to 10 years prior to diagnosis (41 proteins) and CLL cases with a blood draw collected between over 10 years prior to diagnosis (6 proteins, Supplementary Fig. 5ABC). The only protein significantly associated across all time intervals from >10 years to <5 years before CLL diagnosis was FCMR (Supplementary Fig. 5E). Trajectory analyses for the top 10 proteins associated with CLL ranked by FDR-adjusted P-value revealed increasing hazard ratios for FCMR, sCD23, SELL and IGSF3 in blood samples drawn within 5 to 10 years before diagnosis, compared to participants with blood samples drawn over 10 years before diagnosis (Fig. 2B). The hazard ratio for FCMR, FCRL3, SLAMF6, CD72, SEMA4A, CLSTN1, SIGLEC6 increased sharply among participants with a blood sample drawn within 5 years before diagnosis.

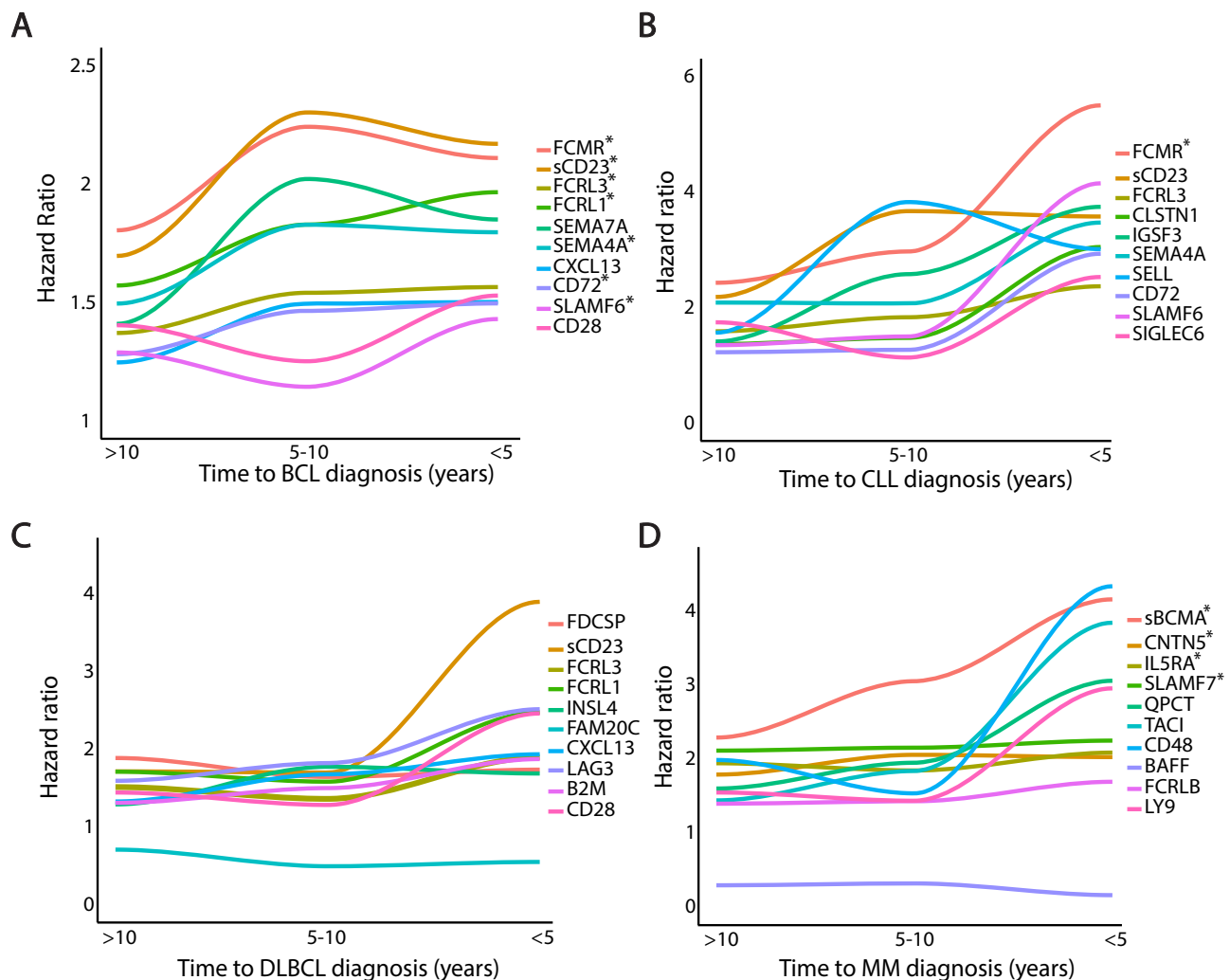
### DLBCL

We observed 34 protein-DLBCL associations (36 aptamers), of which 33 were associated with an increased risk and 1 was associated with a decreased DLBCL risk (Supplementary Fig. 4B, Supplementary Data 1). DLBCL development was associated with PD-1 signaling, T-cell receptor signaling, homologous recombination and antigen processing and presentation (Supplementary Fig. 4E). Time-stratified analyses

revealed an increase in the number of FDR-significant DLBCL-protein associations among participants with a blood draw collected less than 5 years prior to diagnosis (48 proteins), compared to DLBCL cases with a blood draw collected between 5 to 10 years prior to diagnosis (18 proteins) and DLBCL cases with a blood draw collected between over 10 years prior to diagnosis (11 proteins, Supplementary Fig. 6ABC). No proteins were consistently associated across all time intervals from >10 years to <5 years before DLBCL diagnosis (Supplementary Fig. 6E). Trajectory analyses for the top 10 proteins associated with DLBCL ranked by FDR-adjusted P-value revealed relatively stable hazard ratios for all proteins until 5 years before DLBCL diagnosis (Fig. 2C). Among individuals with a blood draw less than 5 years before DLBCL diagnosis, we observed an increasing hazard ratio for sCD23, LAG3, FCRL1 and CD28.

### FL

We observed 20 protein-FL associations (21 aptamers), of which 18 were associated with an increased risk and 2 were associated with a decreased FL risk (Supplementary Fig. 4C, Supplementary Data 1). FL development was associated with an enrichment of proteins associated with the hematopoietic cell lineage, T cell receptor signaling pathway, and the phospholipase D signaling pathway (Supplementary



**Fig. 2 | Trajectory analysis for B-cell malignancy subtypes.** Trajectory analysis of the hazard ratio of the top 10 proteins ranked by two-sided FDR-adjusted  $p$ -value. Proteins marked with an asterisk are significantly associated with the lymphoid malignancy subtype across all three time bins. **A** Cases were divided in 3 bins (151 BCL cases diagnosed over 10 years after blood collection, 106 BCL cases diagnosed between 10 to 5 years of blood collection, 73 BCL cases diagnosed within 5 years of blood collection), each bin was compared to 4088 non-cases using Prentice-weighted Cox regression. For further details on the time-stratified analyses for BCL, see Supplementary Fig. 3. **B** Cases were divided in 3 bins (31 CLL cases diagnosed over 10 years after blood collection, 29 CLL cases diagnosed between 10 to 5 years of blood collection, 20 CLL cases diagnosed within 5 years of blood collection), each bin was compared to 4109 non-cases using Prentice-weighted Cox regression.

For further details on the time-stratified analyses for CLL, see Supplementary Fig. 5. **C** Cases were divided in 3 bins (45 DLBCL cases diagnosed over 10 years after blood collection, 21 DLBCL cases diagnosed between 10 to 5 years of blood collection, 14 DLBCL cases diagnosed within 5 years of blood collection), each bin was compared to 4111 non-cases using Prentice-weighted Cox regression. For further details on the time-stratified analyses for DLBCL, see Supplementary Fig. 6. **D** Cases were divided in 3 bins (53 MM cases diagnosed over 10 years after blood collection, 33 MM cases diagnosed between 10 to 5 years of blood collection, 30 MM cases diagnosed within 5 years of blood collection), each bin was compared to 4111 non-cases using Prentice-weighted Cox regression. For further details on the time-stratified analyses for MM, see Supplementary Fig. 8. Source data are provided as a Source Data file.

Fig. 4F). No time-stratified analyses were performed for FL due to the limited sample size for this subgroup ( $n = 51$ ).

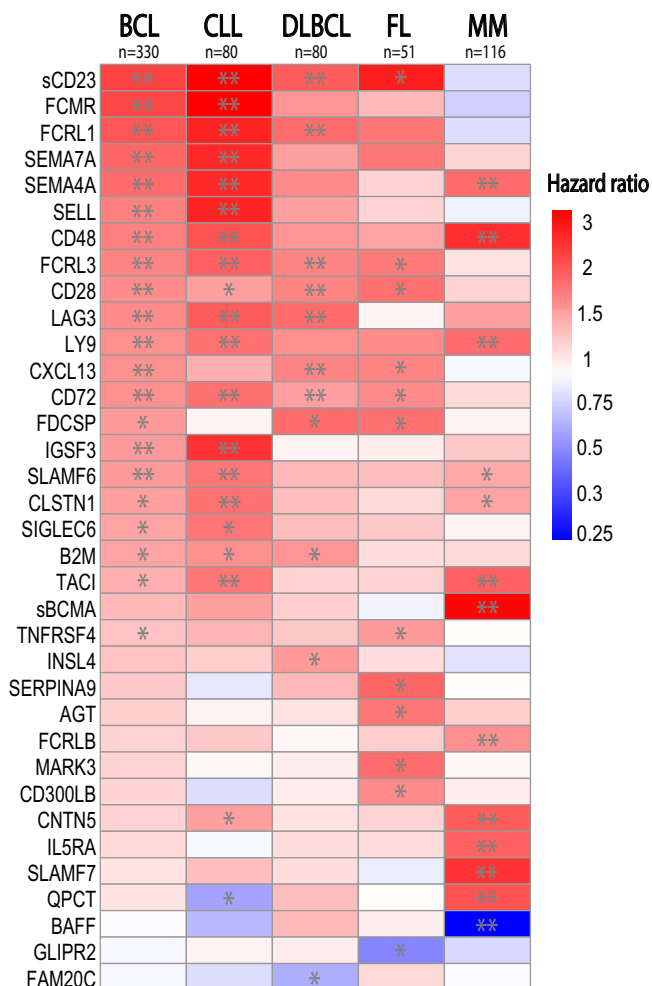
### Overlap between B-cell lymphoma subtypes

Overlap of the protein associations with CLL, DLBCL, and FL was limited to sCD23, CD28, FCRL3, and CD72 (Fig. 3, Supplementary Fig. 4G). All 3 entities were associated with BCR signaling and cytokine signaling (Supplementary Fig. 4DEF, Supplementary Data 3). Interestingly, the overlap between DLBCL and FL included several proteins involved in germinal center signaling, namely: CXCL13, follicular dendritic cell-secreted protein (FDCSP) and CCL21 (Supplementary Fig. 4G).

### MM

We observed associations prior to MM diagnosis ( $n = 116$ ) for 30 unique proteins (32 aptamers), of which 26 were associated with an

increased risk and 4 were associated with a decreased MM risk (Supplementary Fig. 7A, Supplementary Data 1). The top hits included plasma cell activation markers previously associated with MM, such as sBCMA (TNFRSF17), SLAMF7, TACI (TNFRSF13B), and BAFF (TNFSF13B, inverse association)<sup>32–36</sup>. We also observed associations for several proteins that have only very recently been linked to MM development, namely contactin-5 (CNTN5), IL5RA, and QPCT (Figs. 2D, 3)<sup>29,30</sup>. Pathway analysis revealed an enrichment of proteins associated with antibody production and plasma cell activation, JAK-STAT signaling and cytokine signaling in the years prior to MM diagnosis (Supplementary Fig. 7B). Overlap between protein associations with MM and BCL was limited to 8 proteins (Supplementary Fig. 7D). Time-stratified analyses revealed an increase in the number of FDR-significant MM-protein associations among participants with a blood draw collected less than 5 years prior to diagnosis (95 proteins),



**Fig. 3 | B-cell malignancy subtype analyses.** In this heatmap, each row represents the hazard ratio for a protein resulting from two-sided Prentice-weighted Cox regression for the outcome listed for each column. FDR-significant associations are marked with an asterisk. Externally validated associations in the ARIC and UK Biobank cohorts are indicated with two asterisks. Only the top 10 associated proteins sorted by *p*-value are shown for each comparison. For each column, the number of lymphoid malignancy cases listed at the top of the column was compared to the full subcohort ( $N = 4116$ ) through Prentice-weighted Cox regression. Color intensities are based on a log<sub>2</sub>-transformed scale of the hazard ratio (with blue representing a hazard ratio below 1 and red representing a hazard ratio above 1). Untransformed hazard ratio values are shown in the legend to facilitate interpretation. Source data are provided as a Source Data file.

compared to MM cases with a blood draw collected between 5 to 10 years prior to diagnosis (14 proteins) and MM cases with a blood draw collected between over 10 years prior to diagnosis (7 proteins, Supplementary Fig. 8, Supplementary Data 1). sBCMA, CNTN5, IL5RA and SLAMF7 were consistently associated across all time intervals from >10 years to <5 years before MM diagnosis (Supplementary Fig. 8E). Trajectory analyses for the top 10 proteins associated with MM ranked by FDR-adjusted *P*-value revealed a linear increase in hazard ratio over time for sBCMA (Fig. 2D). We additionally observed an increasing hazard ratio for CD48, QPCT, SLAMF7 and LY9 among participants with a blood draw less than 5 years before MM diagnosis.

### Germinal center-derived B-cell lymphoma and non-germinal center-derived B-cell lymphoma

The observed protein-BCL associations included several proteins associated with the germinal center reaction (FDCSP, CXCL13, CCL21).

To further investigate which proteins are associated specifically with germinal center-derived lymphoma, we performed separate analyses for germinal center-derived B-cell lymphoma (DLBCL, FL, and Burkitt lymphoma,  $n = 132$ ) and non-germinal center-derived B-cell lymphoma cases (CLL, mantle cell lymphoma, lymphoplasmacytic lymphoma, marginal zone lymphoma, hairy cell leukemia, and primary effusion lymphoma,  $n = 134$ , Fig. 4A, Supplementary Table 2)<sup>37</sup>.

In our fully-adjusted models, we observed 28 unique proteins associated with an increased risk of germinal center-derived B-cell lymphoma (Supplementary Fig. 9AB, Supplementary Data 1). Top proteins specifically associated with germinal center-derived B-cell lymphoma risk included LSAMP, FDCSP, SERPINA9, CCL21, and CD40LG (Fig. 4B). Among germinal center-derived B-cell lymphomas, only FDCSP showed consistent associations across all pre-diagnostic time intervals (Supplementary Fig. 10ABCE).

In contrast, non-germinal center-derived B-cell lymphoma development was associated with 187 unique protein associations, of which 175 were associated with an increased risk and 13 were associated with a decreased risk of non-germinal center-derived B-cell lymphoma (Supplementary Fig. 9CD, Supplementary Data 1). Top proteins specifically associated with non-germinal center-derived B-cell lymphoma risk included FCMR, SPOCK2, IGSF3, and SEMA4A (Fig. 4B). Time-stratified analyses for non-germinal center-derived B-cell lymphoma revealed 4 proteins (FCMR, FCRL3, SEMA4A, and SLAMF6) associated with non-germinal center-derived B-cell lymphoma from early development (>10 years before diagnosis) until less than 5 years before diagnosis (Supplementary Fig. 11ABCE).

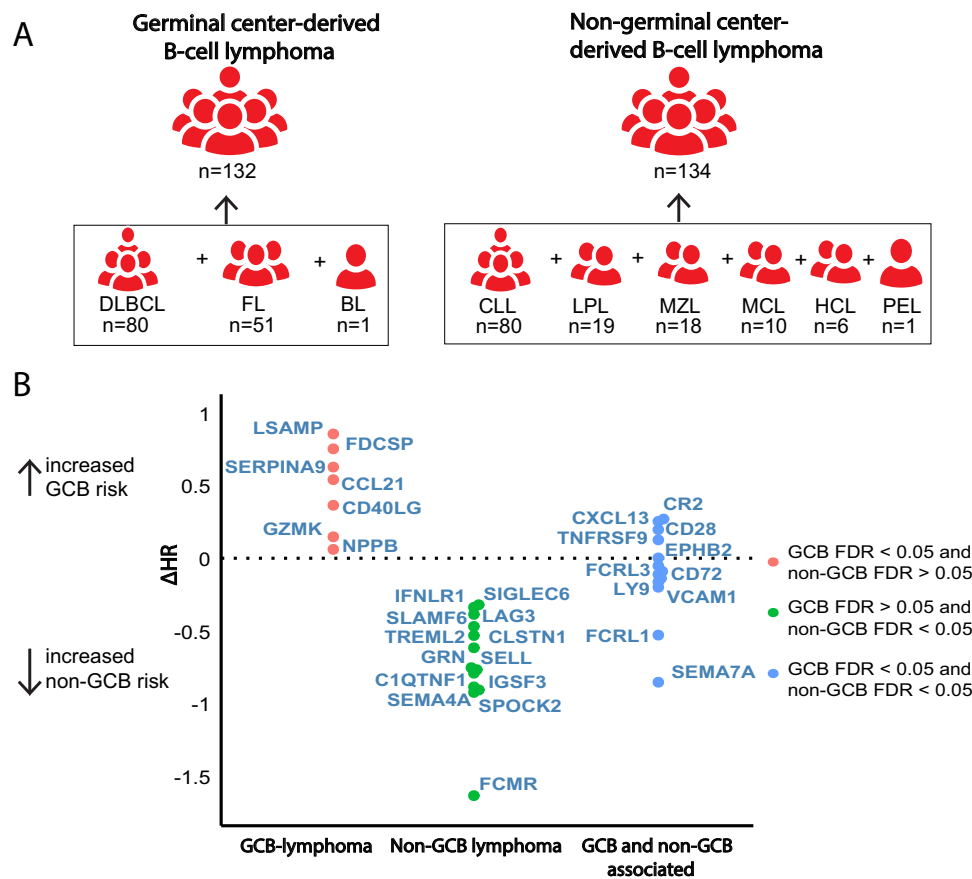
### Subtype analysis for Hodgkin lymphoma and T-cell lymphoma

Due to the rarity of T-cell lymphoma (TCL,  $n = 15$ ) and HL cases ( $n = 23$ ) in our cohort, statistical power was very limited for these sub-analyses. We observed 26 protein-HL associations (Supplementary Fig. 7CD). Overlap between protein associations with HL and BCL was limited to 4 proteins (CD28, CXCL13, FAM20B, and IL-18BP, Supplementary Fig. 7D). We observed 49 unique protein-TCL associations (Supplementary Fig. 12A, Supplementary Data 1). Pathway analysis revealed an enrichment of proteins associated in processes such as cell adhesion, growth hormone, cell cycle, transcriptional dysregulation, and the p53 signaling pathway (Supplementary Fig. 12B, Supplementary Data 3). Notably, only two proteins (CD14 and IGF1) overlapped between BCL and TCL, suggesting largely distinct proteomic risk profiles (Supplementary Fig. 12C).

### Cross-cohort comparison with ARIC and the UK Biobank

We compared our top 20 protein-cancer associations by FDR-adjusted *P*-value within EPIC with nominal protein-cancer associations observed in the ARIC and UK Biobank cohorts (Fig. 5). The ARIC cohort had the longest follow up, but a smaller sample size in terms of included cases (Supplementary Fig. 13). The UK-biobank cohort had the largest overall sample size and the shortest follow up (Supplementary Fig. 13). The EPIC cohort included the largest number of included lymphoid malignancy cases and had an intermediate follow up length compared to the other two cohorts (Supplementary Fig. 13). The EPIC (6412 proteins) and ARIC cohorts (4712 proteins) were measured using SomaScan technology, while the UK biobank cohort (3072 proteins) was measured using the Olink Explore assay.

The top 20 FDR significant proteins within EPIC showed high concordance in ARIC for BCL (95%, HR  $r = 0.74$ ), CLL (90%, HR  $r = 0.68$ ), and MM (85%, HR  $r = 0.82$ ) (Fig. 5A, B, Supplementary Fig. 14A and Supplementary Data 4). The top 20 FDR significant proteins within EPIC showed high concordance in the UK Biobank in terms of nominal significance for BCL (95%,  $r = n.s.$ ), DLBCL (70%,  $r = n.s.$ ), and MM (80%, HR  $r = 0.85$ ) (Fig. 5C, D, Supplementary Fig. 14B and Supplementary Data 4). However, no significant correlation was observed in size of the HR between EPIC and UK Biobank for the top 20 overlapping proteins



**Fig. 4 | Protein associations with germinal center-derived B-cell lymphoma and non-germinal center-derived B-cell lymphoma. A** Overview of grouping for germinal center-derived B-cell lymphoma and non-germinal center-derived B-cell lymphoma groups. Prentice-weighted Cox regression models for the risk of GCB-lymphoma included 4106 non-cases and the 132 germinal center-derived B-cell lymphoma cases. Prentice-weighted Cox regression models for the risk of non-GCB lymphoma included 4101 non-cases and the 134 non-germinal center-derived B-cell lymphoma cases. **B** Comparison of the change in hazard ratio between the models for germinal center-derived B-cell lymphoma risk and non-germinal center-derived B-cell lymphoma risk. Proteins were divided into 3 groups, only

FDR-significantly associated with germinal center-derived B-cell lymphoma risk (red), only FDR-significantly associated with non-germinal center-derived B-cell lymphoma risk (green) or FDR-significantly associated with both groups (blue). The top 20 proteins (ranked by FDR-adjusted P-value) associated with germinal center-derived B-cell lymphoma risk and the top 20 proteins associated with non-germinal center-derived B-cell lymphoma risk were included in the figure. For further details on the analyses relating to germinal center B-cell lymphoma and non-germinal center B-cell lymphoma, see Supplementary Figs. 9–11. Source data are provided as a Source Data file.

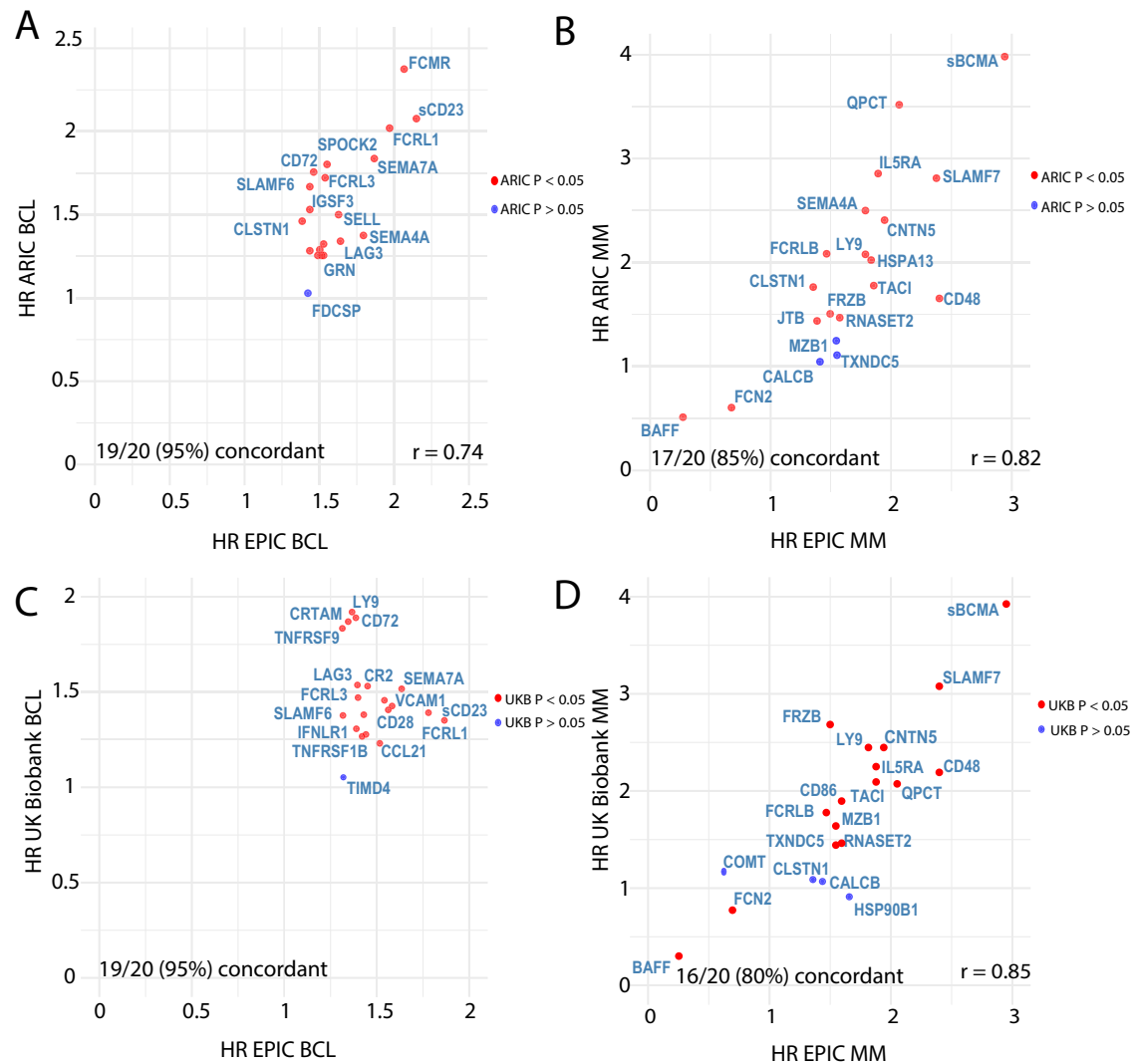
associated with BCL and DLBCL (Fig. 5C, Supplementary Fig. 14B). Top proteins that were consistently associated with BCL across the EPIC, ARIC and UK Biobank cohorts included: sCD23, FCRL1, FCRL3, SEMA7A, CD72, CXCL13 and SLAMF6 (Table 2, Supplementary Data 4). FCMR and SEMA4A were consistently associated with BCL across the EPIC and ARIC cohorts and were not measured in the UK Biobank cohort. Top proteins consistently associated with MM across the EPIC, ARIC, and UK Biobank cohorts included: sBCMA, CNTN5, ILSRA, CD48, SLAMF7, TACI, BAFF, QPCT, FCRLB, and LY9 (Table 2, Supplementary Data 4).

## Discussion

In this large-scale prospective study evaluating 6412 circulating proteins in 4565 EPIC participants, we identified key proteomic markers linked to early lymphoid malignancy pathogenesis (e.g., sBCMA, CXCL13, sCD23, CD28, CD72, FCRL1, FCRL3, SEMA4A, and SEMA7A) and reveal several promising protein markers associated specifically with germinal center-derived B-cell lymphoma (e.g. FDCSP, CCL21, CD40LG, and SERPINA9) and non-germinal center-derived B-cell lymphoma subtypes (e.g. FCMR and SELL). Validation in the UK Biobank and ARIC cohorts revealed high concordance among the top 20 FDR-significant proteins. Our observations extend previously

published data describing elevated levels of B-cell activation markers and cytokines prior to lymphoma diagnosis<sup>21,22,25,28,33,38</sup>.

Predictably, not all of the identified prediagnostic protein-lymphoma associations were novel. In particular, sCD23 serum levels have been previously linked to early development of BCL, especially of CLL<sup>27</sup>. CXCL13 serum levels have been previously linked to DLBCL development<sup>27</sup>. sBCMA and SLAMF7 plasma levels have been previously linked to MGUS and progression to MM<sup>32,33,36</sup>. Additionally, some of the other identified candidate proteins overlap with markers of active disease. A recent study in the Uppsala-Umeå Comprehensive Cancer Consortium (U-CAN) biobank measured 1463 proteins in treatment-naïve diagnostic plasma samples from 48 CLL, 55 DLBCL, and 38 MM patients using Olink Explore PEA technology<sup>29</sup>. Notable proteins associated with active disease included SLAMF7, MZB1, QPCT, BAFF and CNTN5 for MM, FCRL1, FCRL3, IGSF3, SEMA7A, SIGLEC6, TCL1A and sCD23 for CLL and SERPINA9 and CXCL13 for DLBCL. Similarly, FCMR and SELL levels are increased in the plasma of CLL patients<sup>39,40</sup>. CD48 expression is upregulated in MGUS and MM cells and soluble CD48 is increased in the serum of patients with leukemia or lymphoma<sup>41,42</sup>. LY9 (CD229) serum levels are increased in MM patients at diagnosis<sup>43</sup>. Key prediagnostic plasma markers for lymphoid malignancy development identified in our current study



**Fig. 5 | Cross-cohort comparison of top 20 overlapping proteins ranked by FDR-adjusted P-value with the UK Biobank (Olink) and ARIC (SomaScan) cohorts.** Two-sided FDR-adjusted P-values for EPIC were obtained using Prentice-weighted Cox regression. P-values for ARIC and the UK biobank were obtained using two-sided Cox proportional hazards regression without multiple comparisons correction. **A** Hazard ratio of top 20 FDR-significant protein-associations in EPIC plotted against the hazard ratios observed for these proteins in ARIC for B-cell lymphoma (including CLL). Correlation coefficient in the bottom right corner indicates the significant positive correlation found for the hazard ratios across the studies. Each dot represents a single protein. Red dots reached FDR significance in EPIC and nominal significance in either the UK Biobank or ARIC. Blue dots are significant only in EPIC. **B** Hazard ratio of the top 20 FDR-significant protein-

associations in EPIC plotted against the hazard ratios observed for these proteins in ARIC for MM. Correlation coefficient in the bottom right corner indicates the significant positive correlation found for the hazard ratios across the studies. **C** Hazard ratio of top 20 FDR-significant protein-associations in EPIC plotted against the hazard ratios observed for these proteins in the UK biobank for BCL (not including CLL). No significant correlation was observed between the hazard ratios. **D** Hazard ratio of top 20 FDR-significant protein-associations in EPIC plotted against the hazard ratios observed for these proteins in the UK Biobank for MM. Pearson correlation coefficient in the bottom right corner indicates the significant positive correlation found for the hazard ratios across the studies. Exact  $p$ -values can be obtained from Supplementary Data 4. Source data are provided as a Source Data file.

include: CD72, CD28, and SLAMF6 for CLL, FDCSP, CD72, CD28, LAG3, and INSL4 for DLBCL, FDCSP, CD72, and CD28 for FL, TACI, and FCRLB for MM. Notably, several of these markers have previously been described to be highly expressed on the cell surface of the malignant cells<sup>44–46</sup>. These findings indicate that there is considerable overlap between key prediagnostic lymphoid malignancy markers and markers of overt disease, suggesting that early disease biology may be driven by similar molecular pathways as clinically apparent malignancy.

Growth and survival of malignant B cells is sustained through constitutive activation of normal B cell signaling pathways<sup>15,47</sup>. Malignant B-cells exploit these pathways through a combination of gain-of-function mutations that activate downstream signaling mediators, loss-of-function mutations that inactivate negative signaling regulators and autocrine receptor activation. Our study revealed an

enrichment of circulating proteins associated with several of these B-cell signaling pathways, including BCR signaling, cytokine signaling, epigenetic dysregulation, JAK kinase pathways and NF- $\kappa$ B activation<sup>15</sup>. Increased plasma levels of proteins associated with BCR signaling are consistent with the central role of the BCR as a survival signal during lymphomagenesis, both in response to various (auto)antigens and through antigen-independent signaling following crosslinking of the BCR on the cell surface<sup>48,49</sup>. Similarly, constitutive NF- $\kappa$ B activation results in continuous lymphocyte proliferation and survival, a critical pathogenic factor in lymphoma<sup>47</sup>. The top pathway associated with BCL development involved altered cytokine signaling due to viral protein interaction, consistent with the essential role of viruses such as Epstein-Barr virus in the pathogenesis of many BCL subtypes<sup>50,51</sup>. Germinal center signaling through CD40L, CCL21, FDCSP and CXCL13

**Table 2 | Comparison of top protein-BCL and protein-MM associations across EPIC, ARIC and UK biobank cohorts**

Protein	Cancer	HR EPIC	FDR EPIC	HR ARIC	P-value ARIC	HR UKB	P-value UKB
FCMR	BCL	2.1 [1.9–2.3]	<0.0001	2.4 [1.9–2.9]	<0.0001	NM	NM
sCD23	BCL	2.1 [1.9–2.4]	<0.0001	2.1 [1.7–2.6]	<0.0001	1.4 [1.1–1.6]	<0.0001
FCRL3	BCL	1.5 [1.4–1.7]	<0.0001	1.7 [1.4–2.1]	<0.0001	1.5 [1.3–1.8]	<0.0001
FCRL1	BCL	2.0 [1.8–2.2]	<0.0001	2.0 [1.6–2.5]	<0.0001	1.4 [1.1–1.7]	<0.0001
SEMA7A	BCL	1.8 [1.7–2.1]	<0.0001	1.8 [1.5–2.3]	<0.0001	1.5 [1.3–1.7]	<0.0001
SEMA4A	BCL	1.8 [1.6–2.0]	<0.0001	1.4 [1.1–1.7]	<0.01	NM	NM
CXCL13	BCL	1.5 [1.4–1.6]	<0.0001	1.3 [1.0–1.6]	0.04	1.5 [1.4–1.6]	<0.0001
CD72	BCL	1.5 [1.4–1.6]	<0.0001	1.7 [1.4–2.1]	<0.0001	1.9 [1.6–2.2]	<0.0001
SLAMF6	BCL	1.4 [1.3–1.5]	<0.0001	1.7 [1.3–2.0]	<0.0001	1.4 [1.2–1.6]	<0.0001
CD28	BCL	1.5 [1.4–1.7]	<0.0001	NM	NM	1.4 [1.3–1.6]	<0.0001
sBCMA	MM	2.9 [2.5–3.5]	<0.0001	4.0 [2.9–5.5]	<0.0001	4.0 [3.4–4.6]	<0.0001
CNTN5	MM	1.9 [1.7–2.2]	<0.0001	2.4 [1.8–3.2]	<0.0001	2.4 [2.1–2.8]	<0.0001
IL5RA	MM	1.9 [1.6–2.2]	<0.0001	2.8 [2.1–3.8]	<0.0001	2.3 [2.0–2.5]	<0.0001
CD48	MM	2.4 [2.0–2.9]	<0.0001	1.7 [1.2–2.2]	<0.001	2.2 [1.9–2.5]	<0.0001
TACI	MM	1.9 [1.6–2.2]	<0.0001	1.8 [1.3–2.4]	<0.0001	2.1 [2.0–2.2]	<0.0001
SLAMF7	MM	2.4 [2.0–2.8]	<0.0001	2.8 [2.1–3.8]	<0.0001	3.1 [2.7–3.5]	<0.0001
BAFF	MM	0.3 [0.2–0.3]	<0.0001	0.5 [0.4–0.7]	<0.0001	0.3 [0.3–0.4]	<0.0001
QPCT	MM	2.1 [1.8–2.4]	<0.0001	3.5 [2.6–4.7]	<0.0001	2.1 [1.9–2.3]	<0.0001
FCRLB	MM	1.5 [1.3–1.7]	<0.0001	2.1 [1.6–2.7]	<0.0001	1.8 [1.7–1.9]	<0.0001
LY9	MM	1.8 [1.6–2.1]	<0.0001	2.1 [1.6–2.8]	<0.0001	2.4 [2.1–2.8]	<0.0001

BCL B-cell lymphoma, MM multiple myeloma, HR hazard ratio, FDR false discovery rate, NM not measured.

Table includes the top 10 proteins ranked by FDR-adjusted p-value in EPIC. Two-sided FDR-adjusted P-values for EPIC were obtained using Prentice-weighted Cox regression. P-values for ARIC and the UK biobank were obtained using two-sided Cox proportional hazards regression without multiple comparisons correction. The 95% confidence interval of the HR is shown in brackets. For all other significant protein associations in ARIC and UK biobank, and exact p-values for the proteins shown in Table 2, see Supplementary Data 4.

appeared to profoundly shape the plasma proteome of germinal center-derived compared to non-germinal center-derived B-cell lymphoma subtypes<sup>52–56</sup>.

Our pathway enrichment analysis highlights the biological heterogeneity between lymphoid malignancy subtypes, each showing subtly distinct plasma proteomic profiles. For example, the proteomic profile observed prior to MM diagnosis was strongly distinct from the other lymphoid malignancy subtypes, with the top proteins reflecting the altered survival signals that plasma cells require compared to other B-cell subtypes (e.g. sBCMA, SLAMF7, TACI and BAFF). In contrast, the proteomic profile associated with CLL development was marked by increased levels associated with the spliceosome, base excision repair and NF- $\kappa$ B signaling and particular proteins such as FCMR, SELL and IGSF3. Moreover, the number of protein associations varied across subtypes, with more localized subtypes, such as FL and DLBCL, yielding significantly fewer protein associations than leukemic forms, like CLL. Nonetheless, shared features were observed across subtypes, notably involving cytokine and BCR signaling pathways. These observations suggest that lymphoid malignancies are not a homogeneous group but should be considered as distinct entities, each with its own molecular pathogenesis.

Cross-cohort comparison of our results with the ARIC study (SomaScan) and the UK Biobank (Olink) revealed concordant associations for 70%–95% of the top 20 proteins identified in EPIC ranked by FDR significance, indicating strong agreement between the cohorts for these top hits. Furthermore, we observed a positive correlation in the observed HRs for the proteins in ARIC and EPIC. In the UK Biobank, we only observed a significant correlation in the HRs for the top 20 proteins associated with MM, not for BCL and DLBCL. Observed discrepancies for certain proteins may reflect biological heterogeneity, differences in subtype distribution and follow up duration between the cohorts, or technical variation across proteomic platforms. Indeed, previous studies comparing the Olink and SomaScan platforms observed only a moderate positive correlation for overlapping

proteins ( $r = 0.3–0.4$ )<sup>57–61</sup>. Protein families consistently associated with lymphoid malignancy development across all three cohorts included members of the Fc-receptor family, the semaphorins and the TNF superfamily<sup>30</sup>.

Importantly, the extended indolent period prior to lymphoid malignancy blurs the distinction between early risk markers and indicators of subclinical disease<sup>30</sup>. It currently remains unclear at what stage irreversible malignant conversion occurs<sup>13</sup>. Even risk markers detected over a decade prior to lymphoid malignancy diagnosis may still be attributed to reverse causality. One potential strategy to evaluate the impact of reverse causality would be to pair proteomic plasma measurements with an early cancer detection test, e.g. cell-free (cf) DNA methylation, to identify occult malignant populations<sup>62–64</sup>. This approach would allow for the enhanced differentiation between those individuals who are still in a premalignant inflammatory state and those individuals with occult malignant disease, facilitating preventative intervention. However, cfDNA-based early detection tests generally have lower sensitivity for indolent and very early-stage cancers, as these tumors shed relatively little cfDNA into circulation<sup>64</sup>. Therefore, the utility of such tests may be limited for certain lymphoid malignancy subtypes.

The clinical pathway for applying the identified biomarkers remains to be defined. For MM, clinical utility could consist of early treatment with Lenalidomide at the high-risk smoldering MM stage, which has been shown to improve outcomes for MM patients<sup>65,66</sup>. Similar results might be potentially achieved in the future for aggressive lymphoma subtypes, such as DLBCL and MCL. For other, more indolent lymphoid malignancy subtypes, the path to clinical utility is more challenging. Early identification of individuals with CLL and FL at increased risk of transformation to a more aggressive B-cell lymphoma may provide additional opportunities for early therapeutic intervention<sup>67,68</sup>. However, the significant side-effects of chemotherapy and targeted therapeutic options make them an unattractive treatment option for early-stage disease.

Ideally, prevention strategies would rely on interventions with a more favorable risk–benefit profile. These could include targeting chronic inflammation, which is thought to play a key role in lymphoma pathogenesis<sup>69</sup>, reducing environmental exposures associated with increased risk (e.g. benzene, pesticides, viral infections)<sup>70–73</sup>, and promoting protective lifestyle changes such as improvements in diet and physical activity<sup>74–78</sup>. Such public health interventions have the added advantage of benefiting multiple disease areas, further strengthening their risk–benefit profile. Ultimately, detection of premalignant inflammatory states and occult disease is a crucial first step towards the development of effective and safe prevention and early intervention strategies.

Strengths of the current study include its large sample size, the long follow-up period prior to disease onset, the large quantity and breadth of unique proteins measured, the cross-cohort validation of top associations across cohorts, and the ability to evaluate marker associations across the major lymphoma subtypes. Limitations of our current study include a lack of detailed data on clinical characteristics and disease outcome after lymphoid malignancy diagnosis, and the absence of repeated samples to study individual trajectories. A limitation in our analyses studying germinal center-derived B-cell lymphoma is that we did not have sufficiently detailed descriptive data available to divide the DLBCL cases in Germinal Center B-Cell-Like DLBCL and Activated B-Cell-Like DLBCL. Potentially, the observed associations could be strengthened by enhanced stratification based on B-cell biology.

Our findings highlight several processes relevant to the early stages of lymphoid malignancy development, including epigenetic alterations, Fc-receptor signaling, BCR signaling, and germinal center signaling. Understanding which cells produce the proteins we detect in the circulation will require single-cell technology, ideally supplemented by spatial transcriptomics to relate our observations back to tissue architecture. For example, research into HL has implicated an essential role for rosetting (i.e. clusters of surrounding) CD4<sup>+</sup> T-cells in Hodgkin tumor cell survival<sup>79</sup>. These rosetting T-cells express CXCL13, CD28 and TNFRSF18, proteins that were elevated in plasma years prior to HL diagnosis in the current study<sup>80,81</sup>. Additionally, longitudinal dynamics of early disease markers may hold potential to capture key transitions during early disease development that predict progression to overt lymphoma and aid in clinical risk stratification.

The predominance of cell surface proteins among those associated with lymphoid malignancy risk raises questions about whether these proteins are actively secreted or reflect shedding from malignant or pre-malignant cells. While the merit of a protein risk factor is not necessarily affected by its origin, understanding the mechanism by which circulating protein levels are altered would aid interpretation of fluctuations in protein levels over time to lymphoid malignancy diagnosis.

In summary, our study provides insights into the molecular mechanisms underlying lymphoid malignancy. The proteins identified, many of which are cell surface markers, represent promising targets for biomarker-driven strategies for early detection, interception, and prevention. Additionally, these proteins may provide crucial insights into lymphoma etiology. Future research leveraging multi-omics approaches, single-cell analyses, and spatial transcriptomics will be critical to fully unravel the origins and progression of these diseases. By bridging early molecular alterations to overt malignancy, these findings lay the groundwork for earlier detection and more precise intervention strategies in lymphoid cancers.

## Methods

### EPIC participants

EPIC is a prospective cohort study of approximately 521,000 participants (aged 35–70 years) recruited between 1992 and 2000 in 23 centers located in 10 European countries<sup>82</sup>. Among the participants,

~70% were women and blood samples were collected from ~75%. All participants provided informed consent. In EPIC, incident first primary cancer cases (excluding non-melanoma skin cancers) were identified through a combination of center-specific methods, including health insurance records, cancer and pathology registries and active follow-up through study participants and their next-of-kin. Follow-up for each participant and event of interest began upon inclusion in the study and ended upon the occurrence of the event, loss to follow-up, or the last date of ascertainment, whichever came first. Cancer endpoints were defined as the first incident cancer diagnosis, coded using the 10th revision of the WHO's International Statistical Classification of Diseases (ICD-10). In the current study, blood samples from a total of 4565 individuals recruited in the UK, the Netherlands, Spain, and Italy underwent proteomic analysis by Somalogic using the SomaScan 7k Assay.

### Study design

For the current study, we used a case-cohort study design (Fig. 1A). We initially included 449 lymphoid malignancy cases from the total EPIC cohort (~500,000 individuals) in the study. For the controls, we selected a random subcohort ( $n=4116$ ) from the total EPIC cohort (~500,000 individuals). Selection into the sub-cohort was conducted without selecting on future case status, meaning that some participants ( $n=35$ ) diagnosed with a lymphoid malignancy subsequently were included by chance. The subcohort consists of 3715 cancer-free individuals, 366 individuals with a solid cancer diagnosis (e.g., breast or colorectal cancer), and 35 individuals with a lymphoid malignancy diagnosis (Supplementary Fig. 1A). Therefore, we included a total of 484 lymphoid malignancy cases and 4081 non-cases (Fig. 1A, Supplementary Fig. 1B).

In each of the analyses described below, we included the lymphoid malignancy subset of interest and compared it to the full sub-cohort through Prentice-weighted Cox regression. We analyzed BCL ( $n=330$ ), TCL ( $n=15$ ) and MM ( $n=116$ ) separately, as MM is a disease of the plasma cells, a strongly differentiated B-cell subset. Additionally, subtype-specific analyses were performed for CLL ( $n=80$ ), DLBCL ( $n=80$ ) and FL ( $n=51$ ), the three most prevalent subtypes of BCL, as well as for Hodgkin lymphoma (HL,  $n=23$ ). We then divided the BCL cases in germinal center B-cell lymphoma (encompassing DLBCL, FL and Burkitt lymphoma,  $n=132$ ) and non-germinal center B-cell lymphoma cases (encompassing CLL, mantle cell lymphoma, lymphoplasmacytic lymphoma, marginal zone lymphoma, hairy cell leukemia and primary effusion lymphoma,  $n=134$ ) and ran separate analyses for each group (Fig. 3B). A graphical overview of the analytical strategy is shown in Supplementary Fig. 1B and a full overview of the lymphoid malignancy cases by subtype and by group is provided in Supplementary Table 2.

Power calculation for the study was performed using the power-EpiCont.default function from the powerSurvEpi package (0.1.5). For our main analyses for BCL ( $n=330$ ), we estimate power for a HR of 1.5 for our Cox regression models to be approximately 99.8%, using a  $p$ -value cutoff of 0.00008 (0.05/6,412). For our secondary analyses for MM ( $n=116$ ), DLBCL ( $n=80$ ), CLL ( $n=80$ ) and FL ( $n=51$ ), we estimate power for a HR of 2.1 for our Cox regression models to be at least 80%, using a  $p$ -value cutoff of 0.00008 (0.05/6,412). Subtype analyses for TCL ( $n=15$ ) and HL ( $n=23$ ) were underpowered to detect effects with a HR lower than 4.

### Proteomic measurement, processing and quality control

Plasma samples from all 4565 participants underwent high throughput proteomic profiling by Somalogic using the SomaScan 7k assay. Briefly, the SomaScan platform utilizes modified nucleotides (Slow Off-rate Modified Aptamers), which bind directly to specific proteins. The aptamers are tagged with a fluorophore, allowing for quantification in relative fluorescent units (RFUs) using a DNA microarray<sup>83</sup>. For some

targets, the SomaScan 7k panel includes aptamers that bind to isoforms of the same protein or can bind to the same protein at different sites, extending the detectable range of proteoforms. The SomaScan 7k Assay uses 7596 aptamers to measure 6432 proteins (UniProt IDs). Measurements were performed blinded to case-status.

In our main analysis, we used RFUs normalized by Somalogic through the following steps: hybridization normalization, intra-plate median normalization, plate scaling and calibration, and adaptive normalization to a population reference. The normalized RFUs were then  $\log_{10}$ -transformed to reduce skewness. Samples detected as outliers using PCA and a local outlier factor using a Tukey rule modified to account for skewness and multiple testing were excluded<sup>84,85</sup>. We corrected the measurements for each aptamer for plate effects estimated in linear mixed effect models adjusted for center, age, sex, BMI, smoking status, and incidence of cancer, to preserve possible biological variation due to these factors<sup>86</sup>. Finally, measurements of each aptamer were centered and scaled so that their mean and standard deviation were respectively 0 and 1 in the sub-cohort. Log-transformed relative abundance of each aptamer was capped at greater or less than 5 standard deviations from the mean. Original deidentified data is available through the International Agency for Research on Cancer's Scientific IT Platform. Interested parties may contact [epic@iarc.who.int](mailto:epic@iarc.who.int).

### Reproducibility

For every 96-well plate that the SomaScan Assay was ran on, eleven wells were allocated for control samples used to control for batch effects and to estimate the accuracy, precision, and buffer background levels of the assay. The eleven control wells consisted of five pooled Calibrator Control replicates, three pooled Quality Control (QC) replicates, and three buffer (no protein) replicates. Calibrator and QC replicates were created by pooling plasma samples from healthy EPIC participants. Twelve Hybridization Control SOMAmer reagents not exposed to sample proteins were added during the SOMAmer reagent elution step to control for readout variability. The median coefficient of variation (%CV) for the aptamers included in this study was 5% (Q1–Q3 = 4%–7%). An overview of the CVs for all aptamers included in this study can be found in Supplementary Data 5.

Samples detected as outliers using PCA and a local outlier factor using a Tukey rule modified to account for skewness and multiple testing were excluded. In total, 9 samples were identified as outliers and excluded. Exclusion criteria were determined following pre-established guidelines as provided by Somalogic.

### Statistical analysis

We estimated hazard ratios (HRs) and 95% confidence intervals (CI) for the major lymphoid malignancy subsets separately using Prentice-weighted Cox regression models as appropriate for analyzing case-cohort data with age as the underlying time variable. The minimally adjusted models were stratified by age group at recruitment (5-year categories), the center at which the blood sample was collected, and sex. Multivariable-adjusted models were additionally adjusted for smoking status and BMI, two known risk factors for lymphoid malignancy subtypes. Changes in hazard ratio in the fully adjusted vs the minimally adjusted model are shown in Supplementary Fig. 2. A Benjamin-Hochberg false discovery rate (FDR) control was applied for multiple testing (FDR-adjusted  $P < 0.05$ ). Time-stratified analyses were employed for lymphoid malignancy subtypes that included at least 80 cases, meaning BCL, MM, CLL and DLBCL, to explore dynamics in protein-lymphoid malignancy associations over time to lymphoid malignancy diagnosis. For the time-stratified analyses, we divided the cohort into bins based on time to diagnosis (> 10 years, 5–10 years, and <5 years) prior to running Prentice-weighted Cox regression models. All correlation coefficients were determined using Pearson correlation.

### Software

All analyses were conducted using R version 4.1.2. We used the following R packages: survival (3.5.5), tidyr (1.3.1), tidyverse (2.0.0), dplyr (1.1.4), ggplot2 (3.5.2), pheatmap (1.0.12), ggbeeswarm (0.7.2), SomaDataIO (6.1.0), ggpubr (0.6.0), ggrepel (0.9.6), haven (2.5.5), devtools (2.4.5) and Epi (2.60).

### Pathway Enrichment analysis

To identify whether prospectively associated proteins were associated with specific molecular pathways we performed gene-set enrichment analysis (GSEA) for all proteins identified in the Cox proportional hazards model with an FDR-adjusted  $P$ -value  $< 0.05$ . GSEA was performed using an active-subnetwork-oriented approach, implemented with pathfindR (version 2.4.1), which maps each gene onto a protein-protein interaction network to better account for gene interaction information, with each sub-network identified using a greedy algorithm with 10 iterations<sup>87</sup>. We used the BioGrid interaction database (version 4.4.232) for our protein-protein interaction network as it had the largest overlap with the protein-coding genes mapped to our aptamers and used all 580 currently available KEGG Pathway Database gene-sets for the pathway enrichment analysis. Pathway enrichment analyses are performed via one-sided hypergeometric testing.  $P$ -values are adjusted for multiple comparisons using the Bonferroni method. All significantly enriched KEGG pathways ( $P < 0.05$ ) were included in Supplementary Data 3.

### Cross-cohort comparison with ARIC and UK Biobank

In the visit 2 of the Atherosclerosis Risk in Communities (ARIC) study, 4712 plasma proteins were measured using the SomaScan 5K Assay (4953 aptamers passing quality control) in 9478 middle-aged adults (age  $56.9 \pm 5.7$  years) at risk for a first primary cancer and who provided appropriate consent<sup>88,89</sup>. Among them, 95 B-cell lymphoma (of which 40 were CLL) and 49 MM first primary cases were ascertained primarily by state cancer registry linkage over a maximum follow up of 25 years (Supplementary Fig. 13)<sup>90,91</sup>. We evaluated concordance of the nominal significance within ARIC for the top 20 FDR-significant hits within EPIC for B-cell lymphoma, CLL, and MM. Cox proportional hazards regression models were harmonized with EPIC and adjusted for age, race/field center, sex, smoking status, and BMI. In a recent UK Biobank study, 3072 proteins were measured in plasma in 54,306 middle-aged adults using Olink technology, including 206 lymphoma (of which 89 DLBCL), 130 leukemia (C91–C95), and 96 MM cases<sup>30</sup>. We evaluated concordance of the nominal significance within the UK Biobank for the top 20 FDR-significant hits within EPIC for total lymphoma (excluding CLL), MM, and DLBCL. In the UK Biobank analysis, CLL was grouped under leukemia (including myeloid leukemias) rather than BCL, preventing a direct comparison.

### Ethical approval

The EPIC study was conducted in accordance with the Declaration of Helsinki. The study was approved by the local ethical committees in participating countries and the International Agency for Research on Cancer (IARC) ethical committee (EPIC25-34). All participants provided written informed consent for data collection and storage, as well as individual follow-up before study entry. Approval for the ARIC study was received from the Institutional Review Board at each study center (Johns Hopkins Medicine IRB-3, IRB00311861). All participants provided informed consent. The UK Biobank study was approved by the National Information Governance Board for Health and Social Care and the National Health Service Northwest Multicenter Research Ethics Committee (06/MRE08/65).

## Reporting summary

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

## Data availability

Original deidentified data are available through the International Agency for Research on Cancer's Scientific IT Platform. EPIC-Europe data are only available for the purpose of non-commercial research exclusively related to health and chronic diseases. Projects must be driven by a general interest objective, with an intention to disseminate Research Results through peer-reviewed publications and to share any Derived Data to the wider scientific community for future research. As a general principle, any research to be performed with the use of/access to EPIC-Europe Resources must comply with internationally recognized ethical standards and must be ethically and scientifically reviewed and approved by an appropriate independent board or committee. All EPIC-Europe projects require approval by the IARC Ethics committee. The described restrictions were put in place to respect and protect the privacy and consent of EPIC participants. Pre-processing and analytical scripts are available on request to allow the replication of findings by researchers with EPIC data access. Interested parties may contact [epic@iarc.who.int](mailto:epic@iarc.who.int) or [r.c.h.vermeulen@uu.nl](mailto:r.c.h.vermeulen@uu.nl). Researchers will receive a prompt reply to their data access request. Data access will be provided for a maximum of 5 years. For further information regarding EPIC data access, see <https://epic.iarc.fr/access/>. The pre-existing data access policy for the ARIC cohort study specifies that research data requests can be submitted to the ARIC steering committee; these will be promptly reviewed for confidentiality or intellectual property restrictions and will not unreasonably be refused. Individual level patient or protein data may further be restricted by consent, confidentiality or privacy laws/considerations. These policies apply to both clinical and proteomic data. Data access will be provided for a maximum of 5 years. For information on how to access available data and study protocols, see [www2.csc.unc.edu/aric/](http://www2.csc.unc.edu/aric/). ARIC data are also available via BioLINCC (controlled access database). The UK Biobank Resource is available to all bona fide researchers for all types of health-related research that is in the public interest, without preferential or exclusive access for any person (<https://www.ukbiobank.ac.uk/register-apply/>). All researchers, whether in universities, charities, government agencies, or commercial companies, and whether based in the UK or abroad, will be subject to the same application process and approval criteria. The specific objective of the UK Biobank Access Procedures is to maximize access to the samples and data, while ensuring that such access and usage is consistent with the undertakings given to the Participants and the wider public interest (including complying with the prevailing law and upholding respect for human rights). The consent of each Participant (to take part in UK Biobank) remains the cornerstone of UK Biobank's activities. Researchers will receive a prompt reply to their data access request. Data access will be provided for a minimum of 3 years and a maximum of 6 years. This research has been conducted using the UK Biobank Resource under Application Number 67506. Further information is available from the corresponding author upon request. Source data are provided with this paper.

## References

- Hallek, M. & Al-Sawaf, O. Chronic lymphocytic leukemia: 2022 update on diagnostic and therapeutic procedures. *Am. J. Hematol.* **96**, 1679–1705 (2021).
- Hunter, Z. R. et al. The genomic landscape of Waldenström macroglobulinemia is characterized by highly recurring MYD88 and WHIM-like CXCR4 mutations, and small somatic deletions associated with B-cell lymphomagenesis. *Blood* **123**, 1637–1646 (2014).
- Krijgsman, O. et al. Dissecting the gray zone between follicular lymphoma and marginal zone lymphoma using morphological and genetic features. *Haematologica* **98**, 1921–1929 (2013).
- Alaggio, R. et al. The 5th edition of the World Health Organization Classification of haematolymphoid tumours: lymphoid neoplasms. *Leukemia* **36**, 1720–1748 (2022).
- Seifert, M., Scholtysik, R. & Kuppers, R. Origin and pathogenesis of B cell lymphomas. *Methods Mol. Biol.* **1956**, 1–33 (2019).
- Salaverria, I. et al. Translocations activating IRF4 identify a subtype of germinal center-derived B-cell lymphoma affecting predominantly children and young adults. *Blood* **118**, 139–147 (2011).
- Odejide, O. et al. A targeted mutational landscape of angioimmunoblastic T-cell lymphoma. *Blood* **123**, 1293–1296 (2014).
- Maura, F. et al. Genomic landscape and chronological reconstruction of driver events in multiple myeloma. *Nat. Commun.* **10**, 3835 (2019).
- Kolijn, P. M. et al. High-risk subtypes of chronic lymphocytic leukemia are detectable as early as 16 years prior to diagnosis. *Blood* **139**, 1557–1563 (2022).
- Roulland, S. et al. t(14;18) Translocation: a predictive blood biomarker for follicular lymphoma. *J. Clin. Oncol.* **32**, 1347–1355 (2014).
- Landgren, O. et al. Monoclonal gammopathy of undetermined significance (MGUS) consistently precedes multiple myeloma: a prospective study. *Blood* **113**, 5412–5417 (2009).
- Kyle, R. A. et al. Long-term follow-up of monoclonal gammopathy of undetermined significance. *N. Engl. J. Med.* **378**, 241–249 (2018).
- Kolijn, P. M. et al. Genetic drivers in the natural history of chronic lymphocytic leukemia development as early as 16 years before diagnosis. *Blood* **142**, 1399–1403 (2023).
- Fanok, M. H. et al. Role of dysregulated cytokine signaling and bacterial triggers in the pathogenesis of cutaneous T-cell lymphoma. *J. Invest. Dermatol.* **138**, 1116–1125 (2018).
- Shaffer, A. L. 3rd, Young, R. M. & Staudt, L. M. Pathogenesis of human B cell lymphomas. *Annu. Rev. Immunol.* **30**, 565–610 (2012).
- Breen, E. C. et al. Non-Hodgkin's B cell lymphoma in persons with acquired immunodeficiency syndrome is associated with increased serum levels of IL10, or the TL10 promoter-592 C/C genotype. *Clin. Immunol.* **109**, 119–129 (2003).
- Breen, E. C. et al. Elevated serum soluble CD30 precedes the development of AIDS-associated non-Hodgkin's B cell lymphoma. *Tumour Biol.* **27**, 187–194 (2006).
- Breen, E. C. et al. The development of AIDS-associated Burkitt's/Small noncleaved cell lymphoma is preceded by elevated serum levels of interleukin 6. *Clin. Immunol.* **92**, 293–299 (1999).
- Yawetz, S., Cumberland, W. G., Vandermeiden, M. & Martinezmaza, O. Elevated serum levels of soluble Cd23 (Scd23) precede the appearance of acquired immunodeficiency syndrome-associated non-hodgkins-lymphoma. *Blood* **85**, 1843–1849 (1995).
- Widney, D. et al. Aberrant expression of CD27 and soluble CD27 (sCD27) in HIV infection and in AIDS-associated lymphoma. *Clin. Immunol.* **93**, 114–123 (1999).
- Purdue, M. P. et al. A prospective study of 67 serum immune and inflammation markers and risk of non-Hodgkin lymphoma. *Blood* **122**, 951–957 (2013).
- Purdue, M. P. et al. Prediagnostic serum levels of cytokines and other immune markers and risk of non-hodgkin lymphoma. *Cancer Res.* **71**, 4898–4907 (2011).
- Purdue, M. P. et al. A prospective study of serum soluble CD30 concentration and risk of non-Hodgkin lymphoma. *Blood* **114**, 2730–2732 (2009).
- De Roos, A. J. et al. Markers of B-cell activation in relation to risk of non-Hodgkin lymphoma. *Cancer Res.* **72**, 4733–4743 (2012).

25. Purdue, M. P. et al. Elevated serum sCD23 and sCD30 up to two decades prior to diagnosis are associated with increased risk of non-Hodgkin lymphoma. *Leukemia* **29**, 1429–1431 (2015).
26. Purdue, M. P. et al. Prediagnostic serum sCD27 and sCD30 in serial samples and risks of non-Hodgkin lymphoma subtypes. *Int. J. Cancer* **146**, 3312–3319 (2020).
27. Spath, F. et al. Biomarker dynamics in B-cell lymphoma: a longitudinal prospective study of plasma samples up to 25 years before diagnosis. *Cancer Res.* **77**, 1408–1415 (2017).
28. Purdue, M. P. et al. Circulating sCD27 and sCD30 in pre-diagnostic samples collected fifteen years apart and future non-Hodgkin lymphoma risk. *Int. J. Cancer* **144**, 1780–1785 (2019).
29. Alvez, M. B. et al. Next-generation pan-cancer blood proteome profiling using proximity extension assay. *Nat. Commun.* **14**, 4308 (2023).
30. Papier, K. et al. Identifying proteomic risk factors for cancer using prospective and exome analyses of 1463 circulating proteins and risk of 19 cancers in the UK Biobank. *Nat. Commun.* **15**, 4010 (2024).
31. Lee, J. S., Bracci, P. M. & Holly, E. A. Non-Hodgkin lymphoma in women: reproductive factors and exogenous hormone use. *Am. J. Epidemiol.* **168**, 278–288 (2008).
32. O'Connor, B. P. et al. BCMA is essential for the survival of long-lived bone marrow plasma cells. *J. Exp. Med.* **199**, 91–97 (2004).
33. Sanchez, E. et al. Serum B-cell maturation antigen is elevated in multiple myeloma and correlates with disease status and survival. *Br. J. Haematol.* **158**, 727–738 (2012).
34. Ghermezi, M. et al. Serum B-cell maturation antigen: a novel biomarker to predict outcomes for multiple myeloma patients. *Haematologica* **102**, 785–795 (2017).
35. Ishibashi, M. et al. Clinical impact of serum soluble SLAMF7 in multiple myeloma. *Oncotarget* **9**, 34784–34793 (2018).
36. Novak, A. J. et al. Expression of BCMA, TACI, and BAFF-R in multiple myeloma: a mechanism for growth and survival. *Blood* **103**, 689–694 (2004).
37. Mlynarczyk, C., Fontan, L. & Melnick, A. Germinal center-derived lymphomas: the darkest side of humoral immunity. *Immunol. Rev.* **288**, 214–239 (2019).
38. Saberi Hosnijeh, F. et al. Mediating effect of soluble B-cell activation immune markers on the association between anthropometric and lifestyle factors and lymphoma development. *Sci. Rep.* **10**, 13814 (2020).
39. Li, F. J. et al. Enhanced levels of both the membrane-bound and soluble forms of IgM Fc receptor (FcμR) in patients with chronic lymphocytic leukemia. *Blood* **118**, 4902–4909 (2011).
40. Beke Debreceni, I. et al. L-selectin expression is influenced by phosphatase activity in chronic lymphocytic leukemia. *Cytom. B Clin. Cytom.* **96**, 149–157 (2019).
41. Boiarsky, R. et al. Single-cell characterization of myeloma and its precursor conditions reveals transcriptional signatures of early tumorigenesis. *Nat. Commun.* **13**, 7040 (2022).
42. Smith, G. M., Biggs, J., Norris, B., Anderson-Stewart, P. & Ward, R. Detection of a soluble form of the leukocyte surface antigen CD48 in plasma and its elevation in patients with lymphoid leukemias and arthritis. *J. Clin. Immunol.* **17**, 502–509 (1997).
43. Roncador, G. et al. CD229 (Ly9) a novel biomarker for B-cell malignancies and multiple myeloma. *Cancers* **14**, 2154 (2022).
44. Garand, R., Robillard, N. & Bataille, R. Cd72 is constantly expressed in chronic lymphocytic-leukemia and other B-cell lymphoproliferative disorders. *Leuk. Res.* **18**, 651–652 (1994).
45. Yigit, B. et al. A combination of an anti-SLAMF6 antibody and ibrutinib efficiently abrogates expansion of chronic lymphocytic leukemia cells. *Oncotarget* **7**, 26346–26360 (2016).
46. Keane, C. et al. LAG3: a novel immune checkpoint expressed by multiple lymphocyte subsets in diffuse large B-cell lymphoma. *Blood Adv.* **4**, 1367–1377 (2020).
47. Jost, P. J. & Ruland, J. Aberrant NF-κB signaling in lymphoma: mechanisms, consequences, and therapeutic implications. *Blood* **109**, 2700–2707 (2007).
48. Maity, P. C. et al. IGLV3-21\*01 is an inherited risk factor for CLL through the acquisition of a single-point mutation enabling autonomous BCR signaling. *Proc. Natl Acad. Sci. USA* **117**, 4320–4327 (2020).
49. Eken, J. A. Antigen-independent, autonomous B cell receptor signaling drives activated B cell DLBCL. *J. Exp. Med.* **221**, e20230941 (2024).
50. Shannon-Lowe, C., Rickinson, A. B. & Bell, A. I. Epstein-Barr virus-associated lymphomas. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **372**, 20160271 (2017).
51. Vockerodt, M. et al. The Epstein-Barr virus and the pathogenesis of lymphoma. *J. Pathol.* **235**, 312–322 (2015).
52. Basso, K. et al. Tracking CD40 signaling during germinal center development. *Blood* **104**, 4088–4096 (2004).
53. Marshall, A. J. et al. FDC-SP, a novel secreted protein expressed by follicular dendritic cells. *J. Immunol.* **169**, 2381–2389 (2002).
54. Allen, C. D. C., Okada, T., Tang, H. L. & Cyster, J. G. Imaging of germinal center selection events during affinity maturation. *Science* **315**, 528–531 (2007).
55. Stebegg, M. et al. Regulation of the germinal center response. *Front. Immunol.* **9**, 2469 (2018).
56. Cremasco, V. et al. B cell homeostasis and follicle confines are governed by fibroblastic reticular cells. *Nat. Immunol.* **15**, 973–981 (2014).
57. Wang, B. et al. Comparative studies of 2168 plasma proteins measured by two affinity-based platforms in 4000 Chinese adults. *Nat. Commun.* **16**, 1869 (2025).
58. Haslam, D. E. et al. Stability and reproducibility of proteomic profiles in epidemiological studies: comparing the Olink and SOMAscan platforms. *Proteomics* **22**, e2100170 (2022).
59. Raffield, L. M. et al. Comparison of proteomic assessment methods in multiple cohort studies. *Proteomics* **20**, e1900278 (2020).
60. Pietzner, M. et al. Synergistic insights into human health from aptamer- and antibody-based proteomic profiling. *Nat. Commun.* **12**, 6822 (2021).
61. Katz, D. H. et al. Proteomic profiling platforms head to head: leveraging genetics and clinical traits to compare aptamer- and antibody-based methods. *Sci. Adv.* **8**, eabm5164 (2022).
62. Corcoran, R. B. & Chabner, B. A. Application of cell-free DNA analysis to cancer treatment. *N. Engl. J. Med.* **379**, 1754–1765 (2018).
63. Schrag, D. et al. Blood-based tests for multicancer early detection (PATHFINDER): a prospective cohort study. *Lancet* **402**, 1251–1260 (2023).
64. Klein, E. A. et al. Clinical validation of a targeted methylation-based multi-cancer early detection test using an independent validation set. *Ann. Oncol.* **32**, 1167–1177 (2021).
65. Lonial, S. et al. Randomized trial of lenalidomide versus observation in smoldering multiple myeloma. *J. Clin. Oncol.* **38**, 1126–1137 (2020).
66. Rajkumar, S. V., Kumar, S., Lonial, S. & Mateos, M. V. Smoldering multiple myeloma current treatment algorithms. *Blood Cancer J.* **12**, 129 (2022).
67. Jacobsen, E. Follicular lymphoma: 2023 update on diagnosis and management. *Am. J. Hematol.* **97**, 1638–1651 (2022).
68. van der Straten, L., Hengeveld, P. J., Kater, A. P., Langerak, A. W. & Levin, M. D. Treatment approaches to chronic lymphocytic leukemia with high-risk molecular features. *Front. Oncol.* **11**, 780085 (2021).

69. Kolijn, P. M. & Langerak, A. W. Immune dysregulation as a leading principle for lymphoma development in diverse immunological backgrounds. *Immunol. Lett.* **263**, 46–59 (2023).
70. Seidler, A. et al. Solvent exposure and malignant lymphoma: a population-based case-control study in Germany. *J. Occup. Med. Toxicol.* **2**, 2 (2007).
71. O'Connor, S. R., Farmer, P. B. & Lauder, I. Benzene and non-Hodgkin's lymphoma. *J. Pathol.* **189**, 448–453 (1999).
72. Pearce, N. & McLean, D. Agricultural exposures and non-Hodgkin's lymphoma. *Scand. J. Work Environ. Health* **31**, 18–25 (2005).
73. Andreotti, G. et al. Glyphosate use and cancer incidence in the agricultural health study. *J. Natl Cancer Inst.* **110**, 509–516 (2018).
74. Bradbury, K. E., Appleby, P. N. & Key, T. J. Fruit, vegetable, and fiber intake in relation to cancer risk: findings from the European Prospective Investigation into Cancer and Nutrition (EPIC). *Am. J. Clin. Nutr.* **100**, 394S–398S (2014).
75. Skibola, C. F. Obesity, diet, and risk of non-Hodgkin lymphoma. *Cancer Epidemiol. Biomark. Prev.* **16**, 392–395 (2007).
76. Polesel, J. et al. Linoleic acid, vitamin D and other nutrient intakes in the risk of non-Hodgkin lymphoma: an Italian case-control study. *Ann. Oncol.* **17**, 713–718 (2006).
77. Donaldson, M. S. Nutrition and cancer: a review of the evidence for an anti-cancer diet. *Nutr. J.* **3**, 19 (2004).
78. Boyle, T. et al. Physical activity and the risk of non-Hodgkin lymphoma subtypes: a pooled analysis. *Int J. Cancer* **152**, 396–407 (2023).
79. Hartmann, S. et al. Tumour cell characteristics and microenvironment composition correspond to clinical presentation in newly diagnosed nodular lymphocyte-predominant Hodgkin lymphoma. *Br. J. Haematol.* **199**, 382–391 (2022).
80. Aoki, T. et al. Single cell profiling reveals unique CXCL13 positive T cell subsets in the tumor microenvironment of lymphocyte rich classic hodgkin lymphoma. *Blood* **136**, e2105822118 (2020).
81. Ysebaert, L. et al. Lymphoma heterogeneity unraveled by single-cell transcriptomics. *Front. Immunol.* **12**, 597651 (2021).
82. Riboli, E. et al. European prospective investigation into cancer and nutrition (EPIC): study populations and data collection. *Public Health Nutr.* **5**, 1113–1124 (2002).
83. Rohloff, J. C. et al. Nucleic acid ligands with protein-like side chains: modified aptamers and their use as diagnostic and therapeutic agents. *Mol. Ther. Nucleic Acids* **3**, e201 (2014).
84. Breunig, M. M., Kriegel, H. P., Ng, R. T. & Sander, J. LOF: identifying density-based local outliers. *Sigmod Rec.* **29**, 93–104 (2000).
85. Hubert, M. & Vandervieren, E. An adjusted boxplot for skewed distributions. *Comput. Stat. Data Anal.* **52**, 5186–5201 (2008).
86. Viallon, V. et al. A new pipeline for the normalization and pooling of metabolomics data. *Metabolites* **11**, 631 (2021).
87. Ulgen, E., Ozisik, O. & Sezerman, O. U. PathfindR: an R package for comprehensive identification of enriched pathways in omics data through active subnetworks. *Front. Genet.* **10**, 858 (2019).
88. Wright, J. D. et al. The ARIC (Atherosclerosis Risk In Communities) study JACC focus seminar 3/8. *J. Am. Coll. Cardiol.* **77**, 2939–2959 (2021).
89. Walker, K. A. et al. Large-scale plasma proteomic analysis identifies proteins and pathways associated with dementia risk. *Nat. Aging* **1**, 473–489 (2021).
90. Shah, A. M. et al. Large scale plasma proteomics identifies novel proteins and protein networks associated with heart failure development. *Nat. Commun.* **15**, 528 (2024).
91. Joshi, C. E. et al. Enhancing the infrastructure of the Atherosclerosis Risk in Communities (ARIC) study for cancer epidemiology research: ARIC cancer. *Cancer Epidemiol. Biomark. Prev.* **27**, 295–305 (2018).

## Acknowledgments

We thank all EPIC participants for donating the samples that enabled our research. The authors thank the National Institute for Public Health and the Environment (RIVM), Bilthoven, the Netherlands, for their contribution and ongoing support to the EPIC Study. We wish to express our gratitude to the UK Biobank participants and those involved in building the resource. The coordination of EPIC-Europe is financially supported by the International Agency for Research on Cancer (IARC), the Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, and the NIHR Imperial Biomedical Research Center (BRC). The national cohorts are supported by: Associazione Italiana per la Ricerca sul Cancro-AIRC-Italy, Italian Ministry of Health, Italian Ministry of University and Research (MUR), Compagnia di San Paolo (Italy); Dutch Ministry of Public Health, Welfare and Sports (VWS), the Netherlands Organization for Health Research and Development (ZonMW), World Cancer Research Fund (WCRF), (The Netherlands); Health Research Fund (FIS) - Instituto de Salud Carlos III (ISCIII), Regional Governments of Andalucía, Asturias, Basque Country, Murcia and Navarra, and the Catalan Institute of Oncology - ICO (Spain); Cancer Research UK (C864/A14136 to EPIC-Norfolk; C8221/A29017 to EPIC-Oxford), Medical Research Council (MR/N003284/1, MC-UU\_12015/1 and MC\_UU\_00006/1 to EPIC-Norfolk (DOI 10.22025/2019.10.105.00004); MR/Y013662/1 to EPIC-Oxford) (United Kingdom). Previous support has come from the “Europe against Cancer” Program of the European Commission (DG SANCO). SomaScan data were generated under Master Research Agreement, 14th December 2021, between Imperial College London and SomaLogic Inc. SomaLogic was not involved in analyzing or interpreting the data, or in writing or submitting the manuscript for publication. The authors thank the staff and participants of the ARIC study for their important contributions. Cancer data were provided by the Maryland Cancer Registry, Center for Cancer Prevention and Control, Maryland Department of Health, with funding from the State of Maryland and the Maryland Cigarette Restitution Fund. The collection and availability of cancer registry data are also supported by the Cooperative Agreement NU58DP007114, funded by the Centers for Disease Control and Prevention. The Atherosclerosis Risk in Communities study has been funded in whole or in part with Federal funds from the National Heart, Lung, and Blood Institute, National Institutes of Health, Department of Health and Human Services, under Contract nos. (75N92022D00001, 75N92022D00002, 75N92022D00003, 75N92022D00004, 75N92022D00005). Studies on cancer in ARIC are also supported by the National Cancer Institute (U01 CA164975). SomaLogic Inc. conducted the SomaScan assays in exchange for use of ARIC data. This work was also supported in part by NIH/NHLBI grant R01 HL134320. This manuscripts contents are solely the responsibility of the authors and do not necessarily represent the official views of the Centers for Disease Control and Prevention, the Department of Health and Human Services, National Institutes of Health or the International Agency for Research on Cancer/World Health Organization.

## Author contributions

V.V. performed QC analyses and batch correction, P.M.K., K.S.B., V.B., V.V., M.L. and K.P. contributed to data analysis, all authors contributed to interpretation of results, P.M.K. and R.C.H.V. wrote the manuscript, K.S.B., V.B., V.V., M.L., K.P., Z.W., A.W.L., F.S., A.D., C.M.L., R.Z.R., A.M., A.A., R.T., N.C., R.C.T., E.R., M.J.G., E.A.P. and J.M. critically reviewed and edited the manuscript, E. R., M.J.G., J.M. and R.C.H.V. designed and supervised the study.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-025-64534-4>.

**Correspondence** and requests for materials should be addressed to Roel C. H. Vermeulen.

**Peer review information** *Nature Communications* thanks Maja Ludvigsen, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

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

**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-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, 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 you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. 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-nc-nd/4.0/>.

© The Author(s) 2025

<sup>1</sup>Division of Environmental Epidemiology and Veterinary Public Health, Institute for Risk Assessment Sciences, Utrecht University, Utrecht, The Netherlands. <sup>2</sup>Julius Global Health, the Julius Center for Health Sciences and Primary Care, University Medical Center, Utrecht, The Netherlands. <sup>3</sup>Department of Immunology, Erasmus MC, Rotterdam, the Netherlands. <sup>4</sup>Cancer Epidemiology Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK. <sup>5</sup>Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA. <sup>6</sup>International Agency for Research on Cancer (IARC) - World Health Organization, Lyon, France. <sup>7</sup>Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK. <sup>8</sup>Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA. <sup>9</sup>Department of Diagnostics and Intervention, Umeå University, Umeå, Sweden. <sup>10</sup>Department of Pathology and Medical Biology, University of Groningen, University Medical Center Groningen, Groningen, Netherlands. <sup>11</sup>Institute of Epidemiology and Social Medicine, University of Münster, Münster, Germany. <sup>12</sup>Ageing Epidemiology Research Unit, School of Public Health, Imperial College, London, UK. <sup>13</sup>Unit of Nutrition and Cancer, Cancer Epidemiology Research Program, Catalan Institute of Oncology (ICO), Bellvitge Biomedical Research Institute (IDIBELL), Barcelona, Spain. <sup>14</sup>Centre for Biostatistics, Epidemiology and Public Health (C-BEPH), Department of Clinical and Biological Sciences, University of Turin, Orbassano, Italy. <sup>15</sup>Sub-Directorate for Public Health and Addictions of Gipuzkoa, Donostia, Spain. <sup>16</sup>Biodonostia Health Research Institute, Epidemiology of Chronic and Communicable Diseases Group, San Sebastián, Spain. <sup>17</sup>Hyblean Association for Epidemiology Research, AIRE-ONLUS Ragusa, Italy. <sup>18</sup>Cancer Epidemiology and Prevention Research Unit, School of Public Health, Imperial College London, London, UK. ✉ e-mail: [r.c.h.vermeulen@uu.nl](mailto:r.c.h.vermeulen@uu.nl)