Annals of Oncology 25: 1065–1072, 2014
doi:10.1093/annonc/mdu056
Published online 20 February 2014

Prediagnostic transcriptomic markers of Chronic lymphocytic leukemia reveal perturbations 10 years before diagnosis

M. Chadeau-Hyam[1,†*], R. C. H. Vermeulen[2,3,†], D. G. A. J. Hebels[4,†], R. Castagné[1], G. Campanella[1], L. Portengen[2], R. S. Kelly[1], I. A. Bergdahl[5,6], B. Melin[7], G. Hallmans[5], D. Palli[8], V. Krogh[9], R. Tumino[10], C. Sacerdote[11], S. Panico[12], T. M. C. M. de Kok[4], M. T. Smith[13], J. C. S. Kleinjans[4], P. Vineis[1,11] & S. A. Kyrtopoulos[14], on behalf of the EnviroGenoMarkers project consortium[‡]

[1]MRC-HPA Centre for Environment and Health, Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, Norfolk Place, London, UK; [2]Institute for Risk Assessment Sciences, Utrecht University, Utrecht; [3]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht; [4]Department of Toxicogenomics, Maastricht University, Maastricht, The Netherlands; Departments of [5]Biobank Research and the Department of Public Health and Clinical Medicine, Nutritional Research; [6]Occupational and Environmental Medicine, Department of Public Health and Clinical Medicine; [7]Radiation Sciences, Oncology, Umeå University, Umeå, Sweden; [8]Molecular and Nutritional Epidemiology Unit, ISPO Cancer Research and Prevention Institute , Florence; [9]Nutritional Epidemiology Unit, National Cancer Institute, Milan; [10]Ragusa Cancer Registry Azienda Ospedaliera 'Civile M.P. Arezzo', Ragusa; [11]Molecular and Genetic Epidemiology, HuGeF, Human Genetics Foundation, Torino; [12]Department of Social and Preventive Medicine, Universita Federico II, Naples, Italy; [13]Environmental Health Sciences, School of Public Health, University of California, Berkeley, Berkeley, USA; [14]National Hellenic Research Foundation, Institute of Biology, Medicinal Chemistry and Biotechnology, Athens, Greece

Received 19 November 2013; revised 19 January 2014; accepted 5 February 2014

Background: B-cell lymphomas are a diverse group of hematological neoplasms with differential etiology and clinical trajectories. Increased insights in the etiology and the discovery of prediagnostic markers have the potential to improve the clinical course of these neoplasms.

Methods: We investigated in a prospective study global gene expression in peripheral blood mononuclear cells of 263 incident B-cell lymphoma cases, diagnosed between 1 and 17 years after blood sample collection, and 439 controls, nested within two European cohorts.

Results: Our analyses identified only transcriptomic markers for specific lymphoma subtypes; few markers of multiple myeloma ($N = 3$), and 745 differentially expressed genes in relation to future risk of chronic lymphocytic leukemia (CLL). The strongest of these associations were consistently found in both cohorts and were related to (B-) cell signaling networks and immune system regulation pathways. CLL markers exhibited very high predictive abilities of disease onset even in cases diagnosed more than 10 years after blood collection.

Conclusions: This is the first investigation on blood cell global gene expression and future risk of B-cell lymphomas. We mainly identified genes in relation to future risk of CLL that are involved in biological pathways, which appear to be mechanistically involved in CLL pathogenesis**.** Many but not all of the top hits we identified have been reported previously in studies based on tumor tissues, therefore suggesting that a mixture of preclinical and early disease markers can be detected several years before CLL clinical diagnosis.

Key words: epidemiology, lymphoma, chronic lymphocytic leukemia, mRNA analyses, prospective cohort

*Correspondence to: Dr Marc Chadeau-Hyam, Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, St Mary's Campus, Norfolk Place, London W2 1PG, UK. Tel: +44-20-7594-1637; Fax: +44-20-7402-2150; E-mail: m.chadeau@imperial.ac.uk

[†]MCH, RCHV, and DGAJH Contributed equally to this work.
[‡]See Appendix for Additional Consortium Members.

introduction

Non-Hodgkin lymphomas (NHLs) are a heterogeneous collection of lymphoproliferative B- and T-cell malignancies [1], among which B-cell lymphomas [follicular lymphoma (FM), diffuse large B-cell lymphoma (DLBCL), chronic lymphocytic leukemia (CLL), and multiple myeloma (MM)] are the most common [2]. Relatively little is known about NHL etiology [3] and limited prediagnostic markers are currently known. Global gene expression investigation of various chronic conditions, including cancers [4–7], has led to the discovery of molecular signatures predictive of future risk, survival, or response to treatment. To our knowledge to date, no genome-wide gene expression studies have been published using peripheral blood mononuclear cells (PBMCs) to assess the future risk of lymphoma and potentially identify transcriptomic profiles reflecting both the early stages of the disease as well as the influence of risk factors on lymphomagenesis [8]. Further-more, exploration of the transcriptomic profiles as a function of time to diagnosis (TtD) could potentially provide information on gene trajectories involved in disease development and result in predictive sets of disease (progression) markers [9].

Within the EnviroGenomarkers project (www.envirogenomarkers.net), we conducted a global gene expression study using PBMCs from B-cell lymphoma cases collected 1–17 years before disease diagnosis, and healthy controls nested within two large European prospective cohorts.

materials and methods

envirogenomarkers data

The EnviroGenomarkers study was approved by the committees on research ethics at the relevant institutions. It includes participants from the Italian component of the European Prospective Investigation into Cancer and Nutrition [10] (EPIC-Italy, $N = 47\,749$ volunteers aged 35–70) and the Northern Sweden Health and Disease Study (NSHDS, $N = 95\,000$ healthy individuals aged 40–60) [11]. Anthropometric measurements, lifestyle factors, and a blood sample were collected at recruitment (EPIC-Italy 1993–1998; NSHDS 1990–2006). Incident B-cell lymphoma cases were identified through local Cancer Registries (loss to follow-up <2%) and occurred between 1 and 17 years after recruitment. Cases were classified into subtypes according to the SEER ICD-0-3 morphology [12].

biosamples and genome-wide expression profiles

We recently demonstrated that high-quality RNA can be obtained from stored PBMC samples from the EPIC-Italy and NSHDS cohorts [13]. We also showed that samples not cold-stored within 2 h after blood collection had significantly different expression profiles than fresh samples, and therefore only PBMC samples that had been placed in cold storage within 2 h after blood collection were included in the current study. Gene expression profiles were acquired using the Agilent 4 × 44K human whole genome microarray platform.

We analyzed a total of 281 B-cell lymphoma cases and 281 controls matched on sex, age (±2.5 years), center, fasting status, and date of blood collection (±6 months) in two analytical phases. In addition to the lymphoma study, the EnviroGenoMarkers project also comprises 100 breast cancer (BC) case–control pairs (corresponding to 87 and 93 successfully analyzed cases and controls, respectively), which were added in our B-cell lymphoma analyses as unmatched controls in order to maximize statistical power. The final numbers of successfully analyzed samples were 263 B-cell lymphoma cases and 439 controls (supplementary Table S1, available at *Annals of Oncology* online). Technical performance and quality of the microarrays was assessed according to a protocol described previously [13].

statistical analysis

As proposed before [14], we developed a linear-mixed model controlling for potential technically induced noise (nuisance variation) and investigated the relationship between the expression level of each probe and the disease outcome (see supplementary Section S1, available at *Annals of Oncology* online). The general formulation of our mixed model for a given probe defines its expression level observed in participant i (Y^i) as follows:

$$Y^i \sim \alpha + \beta_1 X^i + \beta_2 FE^i + u^{A^i} + \in^i,$$

where α is the intercept of the model, ε^i is the residual error, and X^i is the outcome of interest, a binary variable indicating if individual i is a B-cell lymphoma case or not. The resulting regression estimate β_1 can be expressed as the fold-change (f) by $f = 2^{\beta}$. FE^i is a vector of fixed effect observations for individual i and corresponding regression coefficients are compiled in the vector β_2. Fixed effect covariates included the matching criteria (age, gender and country), the experimental phase (1 or 2), a set of a priori potential confounders as observed in previous analyses of lymphoma within the EPIC cohort [15, 16]: body mass index (BMI, continuous), education (5 classes), physical activity (4 classes), smoking at enrollment (3 classes), and alcohol consumption at enrollment (continuous), and a binary variable indicating if the participant was a BC case or not. Nuisance variation was modeled through a random intercept model where u^{A^i} represents the shift associated to A^i, the value of the random effect variable(s) A observed for individual i. The dates of the three main steps of sample processing were used as random effect variables: RNA isolation, hybridization, and dye labeling. Model was fitted, using the R-statistical package *lme*4, on all 29 662 probes separately, and we accounted for multiple testing using a stringent Bonferroni correction, setting the family-wise error rate (FWER) to 5%. Analyses were (i) carried out on the full population and (ii) stratified by major histological subtypes, and a series of sensitivity analyses were performed.

Transcripts identified by the genome-wide screen were further investigated through gene-enrichment analyses (see supplementary Section S2, available at *Annals of Oncology* online), and gene trajectories linking expression level and TtD were investigated using both linear and generalized additive models.

results

In supplementary Table S2, available at *Annals of Oncology* online, the characteristics of the study population with respect to the main demographic covariates are summarized. Among the study participants, cases (number of successfully analyzed samples) included CLL ($n = 39$), DLBCL ($n = 41$), FL ($n = 38$), MM ($n = 72$), other B-cell lymphomas ($n = 69$), and four unspecified B-cell lymphoma. The distribution of B-cell lymphoma cases by histological subtype, cohort, and TtD is summarized in supplementary Table S3, available at *Annals of Oncology* online.

genome-wide transcriptomic profiles

The linear-mixed model fitted to all B-cell lymphoma cases and controls revealed nine significant associations at a Bonferroni 5%

Table 1. Strongest associations between expression level and NHL

Agilent ID	Full population		Lymphoma subtype								Symbol
			CLL ($N = 39$)		DLBL ($N = 41$)		FL ($N = 38$)		MM ($N = 72$)		
	f^a	*P*-value	f^a	*P*-value	f^a	*P*-value	f^a	*P*-value	f^a	*P*-value	
A_23_P26854	1.79	2.65E−10	24.25	6.06E−60	1.15	2.92E−01	1.19	1.95E−01	1.06	5.46E−01	ARHGAP44
A_23_P500400	1.59	6.39E−10	16.45	3.71E−81	0.97	7.40E−01	1.26	1.64E−02	0.96	5.02E−01	ABCA6
A_23_P210581	0.76	2.94E−08	0.71	7.29E−04	0.91	2.97E−01	0.84	4.67E−02	0.68	8.06E−08	KCNG1
A_23_P145889	1.25	3.24E−07	3.27	5.55E−36	1.01	1.00E−01	1.12	9.73E−02	0.97	5.84E−01	CDK14
A_24_P29733	1.27	7.06E−07	3.89	3.74E−41	1.01	8.45E−01	1.20	1.90E−02	0.95	1.00E+00	CDK14
A_23_P130158	1.53	1.02E−06	14.32	1.17E−46	1.00	6.00E−01	1.19	1.76E−01	0.93	3.79E−01	WNT3
A_23_P384127	1.22	1.08E−06	2.20	3.11E−18	1.04	5.51E−01	1.11	1.61E−01	1.06	3.17E−01	–
A_32_P44394	1.24	1.20E−06	2.66	2.26E−25	1.12	1.63E−01	1.16	5.62E−02	1.05	4.08E−01	AIM2
A_23_P419213	1.21	1.44E−06	2.28	8.45E−24	1.13	4.89E−02	1.19	3.40E−03	0.99	9.25E−01	KIAA1407

Probes declared significant and listed in the table were identified using a Bonferroni-corrected per-test significance level ensuring a FWER control at 5%. Corresponding *P*-values and effect size estimates obtained for subtype-specific analyses, only considering cases of a single subtype at a time and keeping all controls (regardless of the subtype of their matched case) are also given.

[a] Fold-change (f) is derived from the regression coefficient estimate (β) by the mixed model: $f = 2^{\beta}$.

FWER level (Table 1). Analyses by B-cell lymphoma subtype showed that eight of the nine probes show highly significant *P*-values exclusively in the CLL-specific analysis, while the remaining probe seems to be mainly driven by MM. Consistently, when CLL cases are excluded from the population, the *P*-values for these nine probes dramatically increase (supplementary Table S4, available at *Annals of Oncology* online) while, as expected, the MM-driven candidate remains (but more weakly) associated to disease status. Our data do not support the presence of a common signal associated with all B-cell lymphoma subtypes.

Subtype-specific analyses showed numerous associations for CLL ($N = 745$ at Bonferroni FWER 5%) and the 60 strongest signals are reported in Table 2, a. Other subtypes did not provide any realistic candidate signals, with the exception of MM for which we found a few ($N = 3$) weaker candidates with moderate effect sizes (Table 2, b). Subsequent analyses were therefore limited to CLL.

CLL-specific transcriptomic signals

High levels of correlation and strong clustering were observed among the 745 CLL-specific probes (supplementary Figure S3A, available at *Annals of Oncology* online). Consistently, the scree plot from the principal component analysis shows that only 25 components are necessary to explain 80% of the variance within the data (supplementary Figure S3B, available at *Annals of Oncology* online). While the two first principal components only explained <45% of the variance, they are able to clearly discriminate more than 65% of the CLL cases from controls (supplementary Figure S3C, available at *Annals of Oncology* online).

As illustrated in supplementary Figure S4, available at *Annals of Oncology* online, the vast majority of the CLL-specific signals show gene upregulation in cases, with the 20 strongest associations showing up to 25-fold upregulation. A few signals show downregulation in cases, but their association (*P*-values) tends to be weaker.

The 745 CLL-specific probes are spread across all chromosomes (Figure. 1A), and show a consistent overexpression pattern in cases regardless of the chromosome they relate to (Figure 1B). However, a cluster of three very strong signals ($P < 10^{-40}$) emerges in chromosome 17 with large effect size (fold-change >13.9) (Figure 1C).

The predictive ability of the CLL-specific signals was assessed by running a stepwise logistic regression procedure described in supplementary Section S2, available at *Annals of Oncology* online. Results (Figure 2) show excellent predictive performances of the model, even when a single probe is used to predict disease status (the maximum AUC found for a univariate model was based on probe A_23_P500400—gene ABCA6—and was over 90%). As expected, predictive ability improves with the number of probes included in the model and ranges between 89% and 96% for models including 20 probes. Potential for overfitting was assessed and ruled out from a cross-validation procedure (supplementary Section S2, available at *Annals of Oncology* online).

Additional robustness analyses showed that the inclusion over BC cases and controls yielded increased power without introducing a bias, and showed that the strongest findings were detected in both cohorts (supplementary Section S3, available at *Annals of Oncology* online).

biological interpretation of the findings: gene-enrichment analysis

Based on the consistency of the findings across cohorts, insights into the underlying biological process were sought by running gene-enrichment analyses on the 745 CLL-specific markers from the full population (supplementary Section S4, available at *Annals of Oncology* online). The results are summarized in Table 3 and show over 30 significantly enriched pathways and gene ontology terms. The identified pathways all relate to proliferation, differentiation, activation, and regulation of B cells, the

Table 2. Summary of the subtype-specific analyses

Rank[a]	Agilent ID	f^b	*P*-value	Gene	Rank[a]	Agilent ID	f^b	*P*-value	Gene	Rank[a]	Agilent ID	f^b	*P*-value	Gene
(a) CLL-specific analysis														
1	A_23_P500400	16.42	3.7E−81	ABCA6	21	A_23_P201211	5.41	3.0E−33	FCRL5	41	A_23_P147578	2.29	3.5E−26	–
2	A_32_P53234	5.26	4.1E−60	–	22	A_24_P376848	4.11	1.4E−32	FCRL5	42	A_32_P116989	2.20	3.6E−26	ZCCHC18
3	A_23_P26854	24.26	6.1E−60	ARHGAP44	23	A_23_P310931	2.29	2.1E−32	CNR2	43	A_23_P76402	2.10	4.8E−26	TCTN1
4	A_23_P130158	14.37	1.2E−46	WNT3	24	A_23_P46039	3.40	4.4E−32	FCRLA	44	A_23_P39067	2.36	4.9E−26	SPIB
5	A_23_P27332	3.82	3.0E−44	TCF4	25	A_23_P164773	3.39	8.1E−32	FCER2	45	A_23_P116533	2.13	7.1E−26	SWAP70
6	A_23_P131024	5.28	1.0E−43	ZBTB32	26	A_23_P160751	3.89	3.2E−31	FCRL2	46	A_23_P253321	2.35	7.5E−26	PNOC
7	A_24_P691826	5.60	2.3E−43	–	27	A_24_P402588	2.20	3.5E−31	BCL11A	47	A_23_P85269	2.98	8.5E−26	TTN
8	A_24_P29733	3.89	3.7E−41	CDK14	28	A_23_P163697	2.71	4.6E−31	SYT17	48	A_23_P259393	1.85	1.3E−25	SFMBT1
9	A_23_P67529	3.18	7.7E−41	KCNN4	29	A_23_P132378	3.35	2.0E−30	CELSR1	49	A_32_P44394	2.65	2.3E−25	AIM2
10	A_24_P931428	3.69	4.0E−37	TCF4	30	A_23_P17269	2.08	3.3E−30	CCDC88A	50	A_23_P342131	1.81	3.5E−25	CYBASC3
11	A_32_P108156	3.68	9.7E−37	MIR155HG	31	A_23_P45786	2.30	5.7E−30	COL9A2	51	A_24_P662636	3.70	8.3E−25	–
12	A_23_P145889	3.26	5.6E−36	CDK14	32	A_23_P102113	2.10	7.5E−30	WNT10A	52	A_23_P8961	2.39	9.7E−25	IL7
13	A_23_P20427	2.42	6.5E−36	RHOBTB2	33	A_23_P370830	3.92	6.3E−28	KLHL14	53	A_23_P113572	2.77	1.2E−24	CD19
14	A_32_P48054	2.86	9.2E−36	CNR2	34	A_23_P21758	2.81	1.1E−27	ADAM28	54	A_32_P107029	2.14	2.5E−24	NAPSA
15	A_23_P85250	3.01	1.5E−35	CD24	35	A_24_P184803	3.81	1.3E−27	COCH	55	A_23_P4551	2.10	6.1E−24	SETBP1
16	A_23_P156907	4.38	5.6E−35	SOBP	36	A_24_P54390	2.81	3.1E−27	RASGRP3	56	A_23_P419213	2.28	8.4E−24	KIAA1407
17	A_24_P319647	3.37	2.4E−34	FCRL2	37	A_23_P31725	3.09	4.4E−27	BLK	57	A_32_P73507	2.85	1.3E−23	CHDH
18	A_32_P49854	2.08	2.4E−34	–	38	A_24_P410605	2.96	1.4E−26	ROR1	58	A_23_P7185	2.54	4.2E−23	STAP1
19	A_23_P56553	2.59	3.1E−34	METTL8	39	A_23_P40108	3.76	2.3E−26	COL9A3	59	A_32_P72067	2.81	4.4E−23	–
20	A_23_P124335	2.90	8.2E−34	–	40	A_23_P30736	2.13	3.4E−26	HLA-DOB	60	A_23_P91764	2.17	5.6E−23	TNFRSF13C
(b) MM-specific analysis														
1	A_24_P139620	0.99	2.9E−11	USP21										
2	A_23_P210581	0.68	8.1E−08	KCNG1										
3	A_32_P8813	0.67	4.22E−07	LOC283663										

(a) lists the 60 strongest associations found for the CLL-specific analysis and (b) the three significant associations found for MM-specific analyses.

[a] Rank: Probes are ordered with respect to their estimated strength of association with the disease status.

[b] Fold-change (f) is derived from the regression coefficient estimate (β) by the mixed model: $f = 2^{\beta}$.

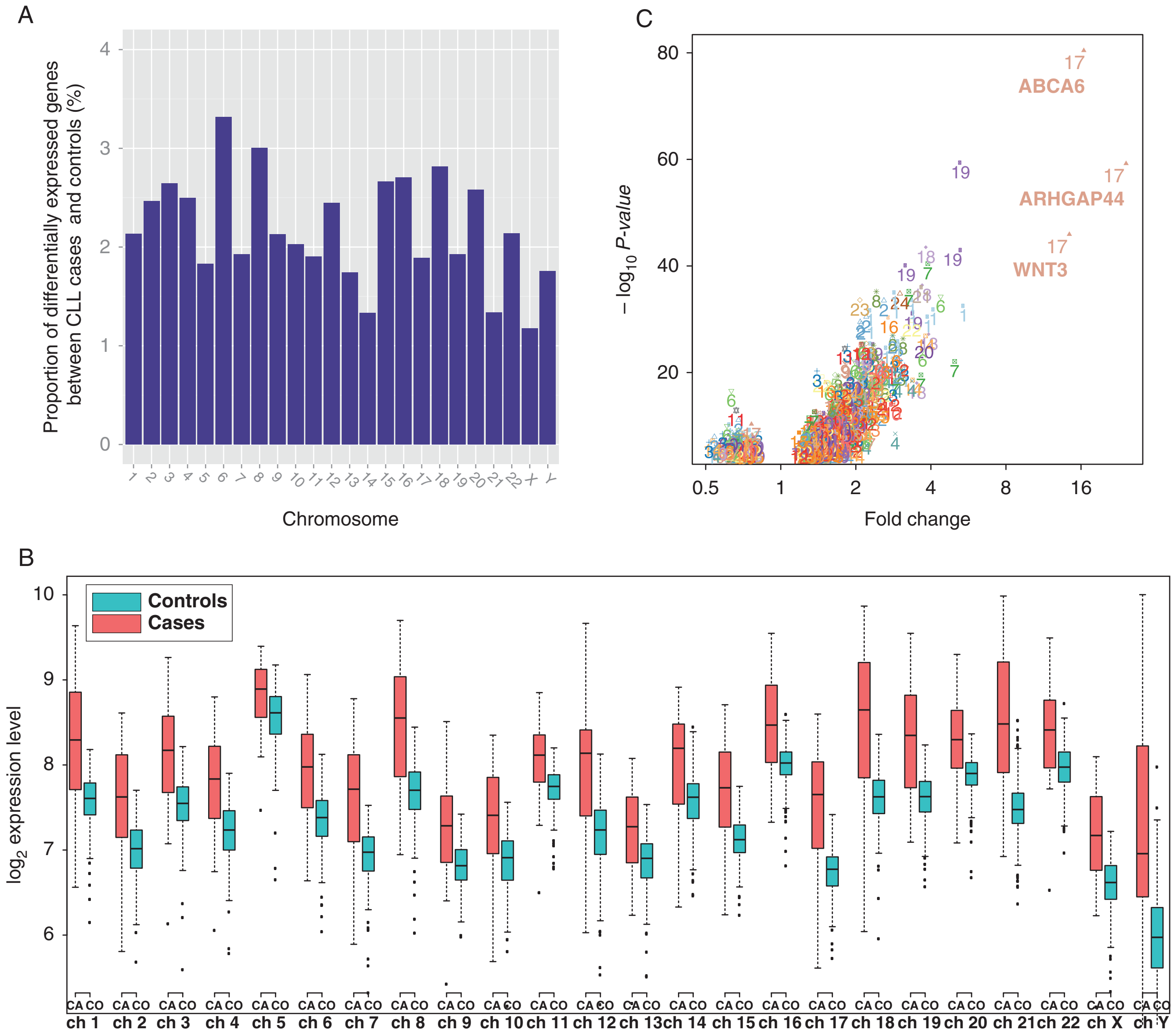

Figure 1. Physical repartition of the genes whose expression is measured by the 745 CLL-specific candidates. The per-chromosome proportion of significant probes (Figure 2A) is calculated from the 739 probes whose chromosome is annotated over the total number of probes assayed per chromosome. Figure 2B summarizes the expression levels in cases and controls for the probes relating to each of the 24 chromosomes (total 739 in which the chromosome is annotated). Figure 2C displays for each probe (labeled and colored accordingly to the chromosome they belong to) the *P*-value measuring the association with the disease status as a function of their effect size estimate (fold-change).

Pleckstrin homology domain (intracellular cell signaling) and immune system regulation.

relationship between CLL-specific transcriptomic markers and time to diagnosis

The results presented above suggest the existence of gene expression signals strongly related to future risk of CLL and present in blood several years before diagnosis. In order to evaluate prediagnostic/preclinical nature of these signals, we ran our CLL-specific analyses on cases enrolled less or more than 6 years before disease onset. Supplementary Figure S7, available at *Annals of Oncology* online clearly shows a large overlap between candidates significant in both TtD strata and in the pooled analysis ($n = 245$). Additional stratification of TtD shows that based only on the six CLL cases diagnosed more than 10 years after enrollment, 47 of the 50 strongest and 68 of the 100 strongest associations found in the full population are still observed.

We also investigated the temporal evolution of expression of the main signals observed among CLL cases only. For the 10 strongest transcriptomic signals, we observed a consistent upregulation while approaching diagnosis (supplementary Figure S8, available at *Annals of Oncology* online). Furthermore, we observed stronger effect sizes (absolute values of the slope with TtD) for the strongest signals with both TtD (supplementary Figure S9A, available at *Annals of Oncology* online) and CLL

Table 3. Summary of the results of the gene-enrichment analyses

Database	Term	Count	*P*-value	Fold enrichment	Bonferroni 5%
GOTERM_BP_FAT	GO:0051249~regulation of lymphocyte activation	24	1.6E−11	5.85	3.6E−08
GOTERM_BP_FAT	GO:0002694~regulation of leukocyte activation	25	2.9E−11	5.43	6.2E−08
GOTERM_BP_FAT	GO:0046649~lymphocyte activation	27	4.3E−11	4.89	9.4E−08
GOTERM_BP_FAT	GO:0050865~regulation of cell activation	25	8.9E−11	5.15	1.9E−07
GOTERM_BP_FAT	GO:0045321~leukocyte activation	28	6.9E−10	4.17	1.5E−06
GOTERM_BP_FAT	GO:0050670~regulation of lymphocyte proliferation	17	8.1E−10	7.39	1.8E−06
GOTERM_BP_FAT	GO:0070663~regulation of leukocyte proliferation	17	9.7E−10	7.30	2.1E−06
GOTERM_BP_FAT	GO:0032944~regulation of mononuclear cell proliferation	17	9.7E−10	7.30	2.1E−06
GOTERM_BP_FAT	GO:0050671~positive regulation of lymphocyte proliferation	14	2.3E−09	9.18	4.9E−06
GOTERM_BP_FAT	GO:0032946~positive regulation of mononuclear cell proliferation	14	2.9E−09	9.02	6.3E−06
GOTERM_BP_FAT	GO:0070665~positive regulation of leukocyte proliferation	14	2.9E−09	9.02	6.3E−06
GOTERM_BP_FAT	GO:0030098~lymphocyte differentiation	18	2.9E−09	6.30	6.4E−06
GOTERM_BP_FAT	GO:0050863~regulation of T cell activation	19	3.2E−09	5.86	7.0E−06
GOTERM_BP_FAT	GO:0001775~cell activation	29	7.0E−09	3.65	1.5E−05
GOTERM_BP_FAT	GO:0051251~positive regulation of lymphocyte activation	17	8.7E−09	6.32	1.9E−05
SP_PIR_KEYWORDS	B-cell	8	1.2E−08	23.53	4.6E−06
GOTERM_BP_FAT	GO:0002521~leukocyte differentiation	19	2.0E−08	5.23	4.4E−05
GOTERM_BP_FAT	GO:0002696~positive regulation of leukocyte activation	17	3.2E−08	5.79	7.0E−05
GOTERM_BP_FAT	GO:0030888~regulation of B cell proliferation	10	3.7E−08	12.88	8.1E−05
GOTERM_BP_FAT	GO:0002684~positive regulation of immune system process	25	4.5E−08	3.79	9.9E−05
GOTERM_BP_FAT	GO:0050867~positive regulation of cell activation	17	6.3E−08	5.52	1.4E−04
GOTERM_BP_FAT	GO:0050864~regulation of B cell activation	12	1.1E−07	8.49	2.4E−04
GOTERM_BP_FAT	GO:0050870~positive regulation of T cell activation	14	1.4E−07	6.65	3.1E−04
GOTERM_BP_FAT	GO:0002520~immune system development	26	1.9E−07	3.40	4.2E−04
INTERPRO	IPR011993:Pleckstrin homology-type	27	2.1E−07	3.30	1.7E−04
GOTERM_BP_FAT	GO:0048534~hemopoietic or lymphoid organ development	25	2.4E−07	3.47	5.3E−04
GOTERM_BP_FAT	GO:0042110~T cell activation	17	3.8E−07	4.87	8.2E−04
INTERPRO	IPR001849:Pleckstrin homology	25	5.3E−07	3.34	4.1E−04
KEGG_PATHWAY	hsa05340:Primary immunodeficiency	9	4.0E−06	9.08	4.6E−04
KEGG_PATHWAY	hsa04662:B cell receptor signaling pathway	12	6.2E−06	5.65	7.1E−04
GOTERM_BP_FAT	GO:0051249~regulation of lymphocyte activation	24	1.6E−11	5.85	3.6E−08
GOTERM_BP_FAT	GO:0002694~regulation of leukocyte activation	25	2.9E−11	5.43	6.2E−08

Pathways found significantly enriched are reported on the basis of their Bonferroni 5% adjusted *P*-values. Gene-enrichment analyses are based on the 745 CLL-specific candidates found for the full population.

status (supplementary Figure S9B, available at *Annals of Oncology* online). This suggests an overall tendency for an increase in intensity of CLL-specific signals closer to diagnosis.

discussion

As expected from the biological heterogeneity of B-cell lymphomas [17], our results do not support the existence of genes whose change in expression is common to the pathogenesis of all or multiple histological subtypes of NHL. Instead, and despite the limited number of CLL cases available, our analyses led to the identification of several strong signals associated with prospective CLL risk (more than 10 years before diagnosis). These include ABCA6, ARHGAP44, Wnt3, TCF4, ZBTB32, CDK14, KCNN4, and TCF4, which showed (except for Wnt3), consistency across both cohorts studied. While variation in the proportion of different subtypes of normal leukocytes may have contributed to these transcriptomic signals, it is unlikely to have been differential by disease status, and by histological subtypes. The substantial overexpression (up to 25 fold) in cases and the trend toward increased expression while approaching diagnosis suggest that the CLL-related signals reflect, at least partly, markers of disease progression arising from subpopulations of cells in which disease initiation has occurred long before diagnosis. This is further supported by the fact that some of the strongest associations we found (e.g. ARHGAP44, ABCA6, and WNT3) are strongly upregulated in CLL malignant cells [18, 19].

Most cases of CLL are believed to be preceded by monoclonal B-cell lymphocytosis, a hematological condition commonly found in normal subjects, increases with age and which evolves to CLL at a low rate (1%–2% per year) [20, 21], raising the possibility that the CLL-related profile we have observed may arise, at least to some extent, in CLL-like MBL cells. Some support for this possibility comes from the inclusion in the latter profile of a number of genes related to Wnt signaling (e.g. Wnt3, Wnt10A,

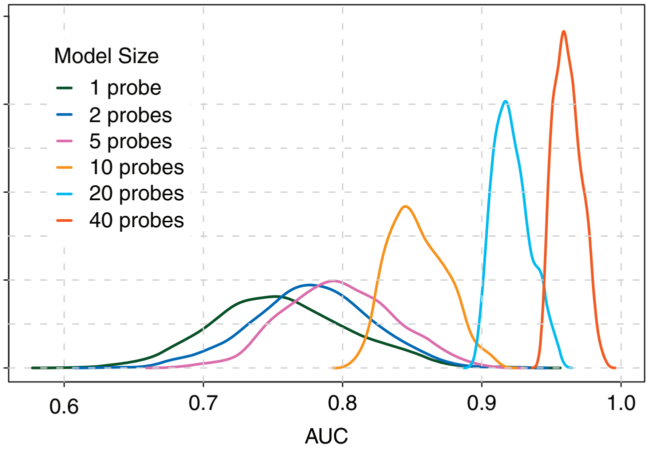

Figure 2. Quantitative assessment of the predictive abilities of the CLL candidate probes. Combinations of probes were selected based on a stepwise procedure including one probe at a time in a logistic model. At each step of the iterative procedure, an additional probe was added to the model such that it maximized the gain in AUC for the resulting ROC curve compared with the probe combination retained at the previous step. The plot presents the density estimate of the AUC at different steps of the algorithm (models containing 1, 5, 10, 20, and 40 probes).

ARHGAP44, TCF4, CDK14, and ZBTB32) which has been recently reported to be activated in MBL [22]. Furthermore, of 20 genes reported as being differentially expressed in CLL-like MBL [22], 4 (PRKCB, PAG1, TCL1A, ROR1) fall among the CLL-related genes identified in the current study. On the other hand, the MAPKinase and protein kinase A pathways, reported to be activated in MBL cells, were not among those indicated by our CLL-related profile. As such, the identified profile seems to be only in part drive by MBL. This is strengthened by the observation that the identified CLL-transcriptomic profile predicts more than 80% of the cases, whereas only ∼5%–10% of subjects of the MBL phenotype would be expected to progress to CLL over the 6-year average follow-up period of our study. Taken together, these observations are compatible with the possibility that the CLL-related differential expression profile detected is due to clones of malignant or premalignant cells, including MBL cells, present at low concentrations in our blood samples several years before clinical onset, and which evolve toward CLL via specific transcriptomic signals. This may not be surprising as, for most patients, CLL is indolent and progresses slowly, and it may take years for clinical symptoms to arise.

The most common chromosomal abnormalities in CLL, using conventional and molecular cytogenetics, are trisomy 12, del(13)(q14), del(11)(q22–23), del(17)(p13), and del(6)(q21) [23]. We did not find any strong evidence of chromosome specificity for our signals, except possibly for chromosomes 17, 18, and 19.

Due to the heterogeneity of NHL pathologies, and despite its reasonable size, our study was not sufficiently large to enable the in-depth investigation of signals associated with histological types other than CLL. The strongest associations we have identified were almost exclusively associated with CLL, but we cannot exclude the possibility that, with greater statistical power, transcripts specific for other subtypes would be identified.

In conclusion, from our agnostic search, several transcriptomics signals have been found to be associated with CLL risk in preclinical blood samples taken many years before actual diagnosis. The identified transcripts point toward an important contribution of B-cell signaling, and B-cell activation and proliferation in the etiology of CLL.

funding

This work was supported by the 7th European Framework Programme (FP7) (Grant Agreement 226756, Envirogenomarkers, to SAK; 308610, Exposomics, to PV; 266198, ECNIS2, to KR), the Fondation pour la Recherche Medicale (FRM, Grant number SPE20120523823, to RC), and by the county of Västerbotten.

disclosure

The authors have declared no conflicts of interest.

references

1. Khan AE, Gallo V, Linseisen J et al. Diabetes and the risk of non-Hodgkin's lymphoma and multiple myeloma in the European Prospective Investigation into Cancer and Nutrition. Haematologica 2008; 93: 842–850.
2. Clarke CA, Glaser SL, Dorfman RF et al. Expert review of non-Hodgkin's lymphomas in a population-based cancer registry: reliability of diagnosis and subtype classifications. Cancer Epidemiol Biomarkers Prev 2004; 13: 138–143.
3. Nieters A, Rohrmann S, Becker N et al. Smoking and lymphoma risk in the European Prospective Investigation into Cancer and nutrition. Am J Epidemiol 2008; 167: 1081–1089.
4. Levy H, Wang X, Kaldunski M et al. Transcriptional signatures as a disease-specific and predictive inflammatory biomarker for type 1 diabetes. Genes Immun 2012; 13: 593–604.
5. Alizadeh AA, Eisen MB, Davis RE et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 2000; 403: 503–511.
6. Fekete T, Raso E, Pete I et al. Meta-analysis of gene expression profiles associated with histological classification and survival in 829 ovarian cancer samples. Int J Cancer 2012; 131: 95–105.
7. van 't Veer LJ, Dai HY, van de Vijver MJ et al. Expression profiling predicts outcome in breast cancer. Breast Cancer Res 2003; 5: 57–58.
8. Vineis P, Perera F. Molecular epidemiology and biomarkers in etiologic cancer research: the new in light of the old. Cancer Epidemiol Biomarkers Prev 2007; 16: 1954–1965.
9. Lund E, Plancade S. Transcriptional output in a prospective design conditionally on follow-up and exposure: the multistage model of cancer. Int J Mol Epidemiol Genet 2012; 3: 107–114.
10. Palli D, Berrino F, Vineis P et al. A molecular epidemiology project on diet and cancer: the EPIC-Italy prospective study. Design and baseline characteristics of participants. Tumori 2003; 89: 586–593.
11. Hallmans G, Agren A, Johansson G et al. Cardiovascular disease and diabetes in the Northern Sweden Health and Disease Study Cohort—evaluation of risk factors and their interactions. Scand J Public Health Suppl 2003; 61: 18–24.
12. Fritz A, Percy C, Jack A et al. International Classification of Diseases for Oncology, 3rd edition (ICD-O-3). Geneva, Switzerland: World Health Organization 2000.
13. Hebels DGAJ, Georgiadis P, Keun HC et al. Performance in omics analyses of blood samples in long-term storage: opportunities for the exploitation of existing biobanks in environmental health research. Environ Health Perspect 2013; 121: 480–487.
14. McHale CM, Zhang LP, Lan Q et al. Global gene expression profiling of a population exposed to a range of benzene levels. Environ Health Perspect 2011; 119: 628–634.

15. Franceschi S, Lise M, Trepo C et al. Infection with hepatitis B and C viruses and risk of lymphoid malignancies in the European Prospective Investigation into Cancer and nutrition (EPIC). Cancer Epidemiol Biomarkers Prev 2011; 20: 208–214.
16. van Veldhoven CM, Khan AE, Teucher B et al. Physical activity and lymphoid neoplasms in the European Prospective Investigation into Cancer and nutrition (EPIC). Eur J Cancer 2011; 47: 748–760.
17. Morton LM, Turner JJ, Cerhan JR et al. Pro-posed classification of lymphoid neoplasms for epidemiologic research from the Pathology Working Group of the International Lymphoma Epidemiology Consortium (InterLymph). Blood 2007; 110: 695–708.
18. Jelinek DF, Tschumper RC, Stolovitzky GA et al. Identification of a global gene expression signature of B-chronic lymphocytic leukemia. Mol Cancer Res 2003; 1: 346–361.
19. Mahadevan D, Choi J, Cooke L et al. Gene expression and serum cytokine profiling of low stage CLL identify WNT/PCP, Flt-3L/Flt-3 and CXCL9/CXCR3 as regulators of cell proliferation, survival and migration. Hum Genomics Proteomics 2009; 2009: 453634.
20. Landgren O, Albitar M, Ma WL et al. B-Cell clones as early markers for chronic lymphocytic leukemia. N Engl J Med 2009; 360: 659–667.
21. Mowery YM, Lanasa MC. Clinical aspects of monoclonal B-cell lymphocytosis. Cancer Control 2012; 19: 8–17.
22. Lanasa MC, Allgood SD, Slager SL et al. Immunophenotypic and gene expression analysis of monoclonal B-cell lymphocytosis shows biologic characteristics associated with good prognosis CLL. Leukemia 2011; 25: 1459–1466.
23. López C, Delgado J, Costa D et al. Clonal evolution in chronic lymphocytic leukemia: analysis of correlations with IGHV mutational status, NOTCH1 mutations and clinical significance. Genes Chromosomes Cancer 2013; 52: 920–927.

appendix

Additional Consortium Members:

P. Georgiadis[14], M. Botsivali[14], C. Papadopoulou[14], A. Chatziioannou[14], I. Valavanis[14], R. Gottschalk[4], D. van Leeuwen[4], L. Timmermans[4], H. C. Keun[15], T. J. Athersuch[15], P. Lenner[7], B. Bendinelli[8], E. G. Stephanou[16], A. Myridakis[16], M. Kogevinas[16,17], F. Saberi-Hosnijeh[2], L. Fazzo[18], M. de Santis[18], P. Comba[18], H. Kiviranta[19], P. Rantakokko[19], R. Airaksinen[19], P. Ruokojarvi[19], M. S. Gilthorpe[20], S. Fleming[20], T. Fleming[20], Y.-K. Tu[20,21], B. Jonsson[22], T. Lundh[22], K.-L. Chien[21], W. J. Chen[21], W.-C. Lee[21], C. K. Hsiao[21], P.-H. Kuo[21], H. Hung[21], S.-F. Liao[21]

[15]Computational and Systems Medicine, Department of Surgery and Cancer, Faculty of Medicine, Imperial College London, London, UK; [16]Department of Chemistry, University of Crete, Voutes-Heraclion, Greece; [17]Centre for Research in Environmental Epidemiology (CREAL), Barcelona, Spain; [18]Istituto Superiore di Sanità, Rome, Italy; [19]Department of Environmental Health, National Public Health Institute Kuopio, Finland; [20]Leeds Institute of Genetics, Health and Therapeutics, Faculty of Medicine and Health, University of Leeds, Leeds, UK; [21]Institute of Epidemiology and Preventive Medicine, College of Public Health, National Taiwan University, Taipei, Taiwan; [22]Division of Occupational and Environmental Medicine, Lund University, Lund, Finland