Supplementary information:

Prospective identification of elevated circulating CDCP1 in patients years before onset of lung cancer.

Sonia Dagnino[1+], Barbara Bodinier[1+], Florence Guida[2], Karl Smith-Byrne[2], Dusan Petrovic[1,3,4], Matthew D. Whitaker[1], Therese Haugdahl Nøst[5], Claudia Agnoli[6], Domenico Palli[7], Carlotta Sacerdote[8], Salvatore Panico[9], Rosario Tumino[10], Matthias B. Schulze[11,12], Mikael Johansson[13], Pekka Keski-Rahkonen[2], Augustin Scalbert[2], Paolo Vineis[1,14], Mattias Johansson[2], Torkjel M. Sandanger[5], Roel C.H. Vermeulen[1,15], Marc Chadeau-Hyam[1,15,*]

1. MRC Centre for Environment and Health, Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, United Kingdom
2. International Agency for Research on Cancer (IARC), Lyon, France
3. Department of Epidemiology and Health Systems (DESS), University Center for General Medicine and Public Health (UNISANTE), Lausanne, Switzerland
4. Department and Division of Primary Care Medicine, University Hospital of Geneva, Geneva, Switzerland
5. Department of Community Medicine, UiT- The Arctic University of Norway, Tromsø, Norway
6. Epidemiology and Prevention Unit, Fondazione IRCCS Istituto Nazionale dei Tumori, Milano.
7. Cancer Risk Factors and Life-Style Epidemiology Unit, Institute for Cancer Research, Prevention and Clinical Network - ISPRO, Florence, Italy
8. Unit of Cancer Epidemiology, Città della Salute e della Scienza University-Hospital, Turin Italy
9. Department of Clinical Medicine and Surgery, Federico II University, Naples, Italy
10. Cancer Registry and Histopathology Department, Provincial Health Authority (ASP) Ragusa, Italy
11. Department of Molecular Epidemiology, German Institute of Human Nutrition Potsdam-Rehbruecke, Nuthetal, Germany
12. Institute of Nutritional Science, University of Potsdam, Nuthetal, Germany
13. Department of Radiation Sciences, Oncology, Umeå University, Umeå, Sweden
14. Italian Institute of Technology, Genova, Italy
15. Division of Environmental Epidemiology, Institute for Risk Assessment Sciences, Utrecht University, Utrecht, The Netherlands

[+] These authors contributed equally to this work

*Corresponding author:
Correspondence and requests for materials should be addressed to
Professor Marc Chadeau-Hyam (email: m.chadeau@imperial.ac.uk)

SI Table 1. **Participants characteristics in the full population and by gender and cohort**. We report the mean (standard deviation) for continuous variables and the count (percentage) for categorical variables. Results are reported for the full population and separated by cohort and gender.

	Full population (N=648)	NOWAC (N=266)	EPIC Women (N=169)	EPIC Men (N=213)
Age at sample (years)	55.29 (5.58)	56.49 (4.01)	53.17 (7.49)	55.48 (4.99)
Gender				
Male	213 (32.9%)	0 (0.0%)	0 (0.0%)	213 (100.0%)
Female	435 (67.1%)	266 (100.0%)	169 (100.0%)	0 (0.0%)
Body Mass Index (kg/m^2)	25.58 (3.86)	24.90 (3.53)	25.56 (4.91)	26.41 (3.05)
Centre				
NOWAC	266 (41.0%)	266 (100.0%)	0 (0.0%)	0 (0.0%)
Florence	132 (20.4%)	0 (0.0%)	79 (46.7%)	53 (24.9%)
Naples	8 (1.2%)	0 (0.0%)	8 (4.7%)	0 (0.0%)
Ragusa	28 (4.3%)	0 (0.0%)	0 (0.0%)	28 (13.1%)
Turin	122 (18.8%)	0 (0.0%)	26 (15.4%)	96 (45.1%)
Varese	92 (14.2%)	0 (0.0%)	56 (33.1%)	36 (16.9%)
Lung cancer status				
Case	323 (49.8%)	133 (50.0%)	84 (49.7%)	106 (49.8%)
Control	325 (50.2%)	133 (50.0%)	85 (50.3%)	107 (50.2%)
Time to diagnosis (years)	5.80 (3.59)	3.81 (2.02)	7.24 (3.69)	7.17 (3.87)
Subtype				
Adenocarcinoma	142 (44.0%)	63 (47.4%)	37 (44.0%)	42 (39.6%)
Large-cell carcinoma	42 (13.0%)	6 (4.5%)	13 (15.5%)	23 (21.7%)
Small-cell carcinoma	46 (14.2%)	26 (19.5%)	10 (11.9%)	10 (9.4%)
Squamous-cell carcinoma	50 (15.5%)	19 (14.3%)	11 (13.1%)	20 (18.9%)
Other	43 (13.3%)	19 (14.3%)	13 (15.5%)	11 (10.4%)
Smoking status				
Never	181 (28.1%)	70 (26.3%)	69 (40.8%)	42 (20.0%)
Former	199 (30.9%)	73 (27.4%)	35 (20.7%)	91 (43.3%)
Current	265 (41.1%)	123 (46.2%)	65 (38.5%)	77 (36.7%)
Smoking intensity	9.56 (8.75)	7.79 (6.12)	6.87 (8.66)	13.80 (9.89)
Packyears	14.36 (14.59)	11.44 (10.75)	9.72 (13.00)	21.64 (17.03)
Time since quitting smoking (years)	5.42 (8.77)	4.96 (9.49)	4.06 (7.44)	6.75 (8.51)
Smoking duration (years)	22.04 (16.83)	24.29 (17.91)	15.31 (15.41)	24.64 (15.08)
Cumulative Smoking Index	0.94 (0.93)	0.56 (0.43)	0.93 (1.03)	1.42 (1.08)
Storage time (years)	14.91 (4.94)	9.03 (0.98)	18.97 (1.17)	18.74 (1.23)
Quality Control				
Warning	15 (2.3%)	3 (1.1%)	1 (0.6%)	11 (5.2%)
Passed	633 (97.7%)	263 (98.9%)	168 (99.4%)	202 (94.8%)
Sample quality				
Poor quality	21 (10.7%)	-	4 (5.2%)	17 (14.2%)
Good quality	627 (89.3%)	-	165 (94.8%)	196 (85.8%)

SI Table 2. **Participants characteristics in the full validation set and by cohort**. We report the mean (standard deviation) for continuous variables and the count (percentage) for categorical variables. Results are reported for the full population and separated by cohort.

	Full population (N=450)	EPIC (N=322)	NSHDS (N=128)
Age at sample (years)	58.11 (8.26)	58.59 (9.01)	56.89 (5.81)
Gender			
Male	276 (61.3%)	214 (66.5%)	64 (50.0%)
Female	174 (38.7%)	108 (33.5%)	64 (50.0%)
Body Mass Index (kg/m^2)	25.91 (3.83)	25.92 (3.85)	25.90 (3.83)
Lung cancer status			
Case	225 (50%)	161 (50.0%)	64 (50.0%)
Control	225 (50%)	161 (50.0%)	64 (50.0%)
Time to diagnosis (years)	1.79 (1.05)	1.81 (1.12)	1.74 (0.87)
Subtype			
Adenocarcinoma	71 (31.5%)	48 (29.8%)	23 (35.9%)
Large-cell carcinoma	29 (12.9%)	29 (18.0%)	0 (0.0%)
Small-cell carcinoma	38 (14.2%)	38 (23.6%)	6 (9.4%)
Squamous-cell carcinoma	41 (18.2%)	28 (17.4%)	13 (20.3%)
Other	46 (20.4%)	24 (14.9%)	22 (34.4.5%)
Smoking status			
Never	0 (0%)	0 (0%)	0 (0%)
Former	154 (34.2%)	102 (31.7%)	52 (40.6%)
Current	296 (65.8%)	220 (68.3%)	76 (59.4%)

SI Table 3. Regression models with packyears (linear model) or smoking status (logistic models for never vs. current or never vs. former) as the outcome and individual protein levels as predictor. Centre and plate effects are removed from the data by taking the residuals from linear mixed models with protein levels as the outcome and centre and plate as random intercepts. Models are adjusted on age and BMI. The odds ratios are expressed as the risk change for a one standard deviation increase in the protein levels. The p-values are derived from likelihood ratio tests comparing the fit of the model with to that of the model without the protein levels in the set of predictors. Results are presented for the twelve proteins found associated with lung cancer in the base model on all Women after Benjamini-Hochberg correction for multiple testing.

	Packyears (N=388)		Smoking status			
			Never vs. former (N=228)		Never vs. current (N=301)	
	β	p-value	OR	p-value	OR	p-value
CDCP1	0.23	**2.45e-08**	1.34	3.85e-02	1.82	**4.86e-06**
SCF	-0.23	**5.25e-09**	0.75	4.37e-02	0.47	**1.97e-08**
HGF	0.20	**8.93e-07**	0.98	8.68e-01	2.16	**3.21e-08**
IL6	0.11	**6.91e-03**	1.20	1.90e-01	1.47	**3.80e-03**
OSM	0.13	**2.04e-03**	0.92	5.76e-01	1.98	**2.01e-07**
MCP1	0.15	**2.01e-04**	0.94	6.61e-01	1.38	**7.38e-03**
IL8	0.07	6.92e-02	0.99	9.16e-01	1.36	**1.54e-02**
VEGFA	0.13	**1.33e-03**	0.94	6.44e-01	1.68	**5.25e-05**
TWEAK	-0.15	**1.78e-04**	0.82	1.41e-01	0.55	**2.42e-06**
IL12B	-0.19	**4.95e-06**	0.98	8.67e-01	0.38	**1.91e-12**
CD6	0.14	**2.87e-04**	1.06	6.53e-01	1.62	**1.01e-04**
CD5	0.12	**3.82e-03**	1.04	7.89e-01	1.69	**3.51e-05**

SI Table 4. Logistic regression models with future disease status as the outcome and individual protein levels as predictor in Women. Results are presented by smoking status. Centre and plate effects are removed from the data by taking the residuals from linear mixed models with protein levels as the outcome and centre and plate as random intercepts. Models are adjusted on age and BMI (base model). For populations including smokers, models further adjusted on packyears are also reported. The odds ratios are expressed as the risk change for a one standard deviation increase in the protein levels. The p-values of association with future disease status are derived from likelihood ratio tests comparing the fit of the model with to that of the model without protein levels in the set of predictors. Results are presented for the twelve proteins found associated with lung cancer in the base model on all Women after Benjamini-Hochberg correction for multiple testing. The numbers of cases and controls used in each model are reported.

	All Women				Never smoking Women		Current smoking Women			
	Base model (196/201)		Adjusted on packyears (191/197)		Base model (32/100)		Base model (54/115)		Adjusted on packyears (52/111)	
	OR	p-value	OR	p-value	OR	p-value	OR	p-value	OR	p-value
CDCP1	1.94	**5.49e-09**	1.58	**3.09e-04**	1.46	7.78e-02	2.16	**4.91e-05**	2.16	**2.00e-04**
SCF	0.62	**1.02e-05**	0.78	3.94e-02	1.01	9.68e-01	0.50	**5.31e-04**	0.53	**1.91e-03**
HGF	1.43	**6.82e-04**	1.20	1.19e-01	1.12	6.26e-01	1.36	8.13e-02	1.33	1.20e-01
IL6	1.46	**7.63e-04**	1.28	3.27e-02	1.27	2.23e-01	2.44	**1.34e-04**	2.14	**6.07e-04**
OSM	1.41	**1.09e-03**	1.25	5.98e-02	1.22	3.46e-01	1.44	3.94e-02	1.45	4.70e-02
MCP1	1.38	**2.12e-03**	1.24	6.62e-02	1.21	3.63e-01	1.54	1.25e-02	1.49	2.83e-02
IL8	1.35	**3.84e-03**	1.34	1.16e-02	1.61	1.69e-02	1.21	2.86e-01	1.23	2.68e-01
VEGFA	1.33	**5.39e-03**	1.22	9.00e-02	1.18	4.56e-01	1.24	1.98e-01	1.29	1.48e-01
TWEAK	0.76	**6.47e-03**	0.92	4.94e-01	1.09	6.92e-01	0.77	1.44e-01	0.82	2.86e-01
IL12B	0.75	**6.65e-03**	0.91	4.33e-01	1.13	5.62e-01	0.88	4.50e-01	0.88	4.84e-01
CD6	1.32	**7.08e-03**	1.16	1.99e-01	1.15	5.08e-01	1.18	3.18e-01	1.17	3.79e-01
CD5	1.32	**7.41e-03**	1.17	1.72e-01	1.02	9.27e-01	1.30	1.40e-01	1.33	1.27e-01

SI Table 5. Logistic regression models with future disease status as the outcome and individual protein levels as predictor in Men. Models are adjusted on age and BMI (A). Centre and plate effects are removed from the data by taking the residuals from linear mixed models with protein levels as the outcome and centre and plate as random intercepts. Models further adjusted on packyears are also reported (B). The odds ratios are expressed as the risk change for a one standard deviation increase in the protein levels. The p-values of association with future disease status are derived from likelihood ratio tests comparing the fit of the model with to that of the model without protein levels in the set of predictors. Results are presented for the twelve proteins found associated with lung cancer in the base model on all Women after Benjamini-Hochberg correction for multiple testing. The numbers of cases and controls used in each model are reported.

	Lung cancer				Adenocarcinoma				Small-cell carcinoma				Large-cell carcinoma				Squamous-cell carcinoma			
	Base model (88/88)		Adjusted on packyears (85/88)		Base model (36/88)		Adjusted on packyears (36/88)		Base model (8/88)		Adjusted on packyears (8/88)		Base model (19/88)		Adjusted on packyears (19/88)		Base model (18/88)		Adjusted on packyears (16/88)	
	OR	p-value	OR	p-value	OR	p-value	OR	p-value	OR	p-value	OR	p-value	OR	p-value	OR	p-value	OR	p-value	OR	p-value
CDCP1	1.68	1.51e-03	1.86	7.42e-04	1.42	8.62e-02	1.48	7.02e-02	3.82	9.64e-04	3.91	1.58e-03	2.74	6.71e-04	2.86	8.80e-04	1.24	4.29e-01	1.26	4.46e-01
SCF	0.64	5.39e-03	0.92	6.44e-01	0.65	3.29e-02	0.84	4.63e-01	0.66	2.76e-01	0.83	6.71e-01	0.54	2.09e-02	0.68	2.14e-01	0.53	1.50e-02	0.70	2.93e-01
HGF	1.33	7.87e-02	1.02	9.14e-01	0.98	9.09e-01	0.79	3.44e-01	1.55	2.37e-01	1.44	3.81e-01	1.97	1.10e-02	1.65	9.26e-02	1.41	2.10e-01	1.29	4.69e-01
IL6	1.47	1.82e-02	1.29	1.67e-01	1.30	1.84e-01	1.19	4.31e-01	1.38	3.46e-01	1.21	6.40e-01	2.08	2.52e-03	1.89	1.73e-02	1.09	7.51e-01	0.86	6.69e-01
OSM	1.13	4.16e-01	0.93	7.03e-01	0.85	4.53e-01	0.79	2.92e-01	2.20	2.83e-02	2.10	4.23e-02	1.44	1.48e-01	1.28	3.54e-01	1.19	4.96e-01	1.23	4.92e-01
MCP1	1.29	1.05e-01	1.21	2.64e-01	0.88	5.13e-01	0.93	7.36e-01	1.95	3.69e-02	2.04	2.95e-02	1.87	2.01e-02	1.58	9.63e-02	1.19	5.08e-01	1.37	3.13e-01
IL8	1.28	1.12e-01	1.31	1.16e-01	1.10	6.54e-01	1.21	3.60e-01	1.42	2.85e-01	1.50	2.56e-01	1.48	1.11e-01	1.39	2.13e-01	1.25	3.64e-01	1.27	4.55e-01
VEGFA	1.19	2.64e-01	1.20	2.95e-01	1.05	8.08e-01	1.19	4.33e-01	0.90	7.92e-01	0.98	9.69e-01	1.90	1.74e-02	1.91	2.71e-02	1.28	3.37e-01	1.71	9.81e-02
TWEAK	0.93	6.44e-01	1.05	8.01e-01	0.85	4.19e-01	0.97	8.88e-01	0.70	3.31e-01	0.80	5.53e-01	1.39	2.35e-01	1.48	1.79e-01	0.90	6.93e-01	0.95	8.80e-01
IL12B	0.74	5.13e-02	0.89	5.29e-01	0.76	1.93e-01	0.89	6.08e-01	0.48	5.21e-02	0.52	8.52e-02	0.97	9.10e-01	1.10	7.35e-01	0.51	1.77e-02	0.66	1.92e-01
CD6	1.42	3.19e-02	1.40	7.70e-02	1.41	1.19e-01	1.50	8.63e-02	1.85	7.59e-02	1.80	9.44e-02	1.63	5.97e-02	1.58	1.06e-01	1.21	4.76e-01	1.40	2.73e-01
CD5	1.57	5.65e-03	1.37	8.43e-02	1.41	1.00e-01	1.37	1.48e-01	2.51	1.27e-02	2.37	2.54e-02	1.84	2.13e-02	1.68	7.22e-02	1.41	2.01e-01	1.58	1.58e-01

SI Table 6. Univariate logistic regression models with future disease status as the outcome and individual protein levels as predictor in the full population. Models are adjusted on age, gender and BMI (A). Centre and plate effects are removed from the data by taking the residuals from linear mixed models with protein levels as the outcome and centre and plate as random intercepts. Models further adjusted on packyears are also reported (B). The odds ratios are expressed as the risk change for a one standard deviation increase in the protein levels. The p-values of association with future disease status are derived from likelihood ratio tests comparing the fit of the model with to that of the model without protein levels in the set of predictors. Results are presented for pooled lung cancer and by histological subtype for proteins found associated with at least one of the subtypes considered after Benjamini-Hochberg correction for multiple testing. The numbers of cases and controls used in each model are reported.

	Lung cancer				Adenocarcinoma				Small-cell carcinoma				Large-cell carcinoma				Squamous-cell carcinoma			
	Base model (284/289)		Adjusted on packyears (276/285)		Base model (127/289)		Adjusted on packyears (125/285)		Base model (40/289)		Adjusted on packyears (38/285)		Base model (36/289)		Adjusted on packyears (36/285)		Base model (44/289)		Adjusted on packyears (42/285)	
	OR	p-value	OR	p-value	OR	p-value	OR	p-value	OR	p-value	OR	p-value	OR	p-value	OR	p-value	OR	p-value	OR	p-value
CDCP1	1.83	**6.29e-11**	1.65	**8.18e-07**	1.90	**2.52e-08**	1.71	**7.45e-06**	3.69	**8.03e-12**	3.02	**3.02e-07**	1.67	6.78e-03	1.56	2.05e-02	1.44	3.33e-02	1.42	5.35e-02
SCF	0.63	**2.23e-07**	0.82	4.34e-02	0.72	2.78e-03	0.88	2.67e-01	0.62	**3.88e-03**	0.78	2.06e-01	0.58	2.51e-03	0.69	5.40e-02	0.49	**3.03e-06**	0.57	1.43e-03
IL6	1.47	**3.27e-05**	1.30	6.69e-03	1.28	1.78e-02	1.19	1.26e-01	1.44	1.06e-02	1.37	6.38e-02	1.60	1.46e-03	1.55	5.06e-03	1.34	4.07e-02	1.35	6.87e-02
HGF	1.39	**2.01e-04**	1.14	1.84e-01	1.27	2.64e-02	1.12	3.40e-01	1.61	**2.70e-03**	1.38	6.64e-02	1.41	5.35e-02	1.24	2.58e-01	1.48	1.11e-02	1.32	1.09e-01
CD5	1.36	**3.70e-04**	1.19	6.33e-02	1.33	7.33e-03	1.22	8.67e-02	1.68	**2.15e-03**	1.55	2.82e-02	1.42	5.18e-02	1.30	1.64e-01	1.28	1.29e-01	1.26	1.99e-01
MCP1	1.35	**4.88e-04**	1.24	2.59e-02	1.14	2.41e-01	1.10	4.26e-01	1.86	**2.82e-04**	1.68	5.02e-03	1.73	3.65e-03	1.57	1.78e-02	1.38	5.90e-02	1.37	8.36e-02
IL12B	0.75	**8.77e-04**	0.90	2.89e-01	0.82	7.81e-02	0.97	8.10e-01	0.59	**2.13e-03**	0.71	7.60e-02	0.85	3.66e-01	0.95	7.95e-01	0.55	**5.45e-04**	0.68	4.19e-02
CCL11	1.33	**1.03e-03**	1.08	4.16e-01	1.12	2.84e-01	1.01	9.64e-01	2.56	**9.11e-07**	2.19	**4.90e-04**	1.61	1.03e-02	1.43	6.46e-02	1.55	9.52e-03	1.41	7.38e-02
MMP10	1.33	**1.10e-03**	1.22	3.71e-02	1.18	1.42e-01	1.10	4.12e-01	1.28	1.59e-01	1.11	6.02e-01	1.87	8.80e-04	1.64	9.28e-03	1.28	1.42e-01	1.05	7.67e-01
IL8	1.32	**1.32e-03**	1.32	4.09e-03	1.23	4.36e-02	1.24	4.83e-02	1.33	4.86e-02	1.34	6.24e-02	1.24	2.43e-01	1.20	3.16e-01	1.44	1.39e-02	1.28	1.05e-01
CD6	1.31	**1.44e-03**	1.17	8.96e-02	1.16	1.55e-01	1.07	5.64e-01	1.81	**3.97e-04**	1.66	1.10e-02	1.27	1.89e-01	1.17	4.10e-01	1.47	1.95e-02	1.40	7.07e-02
OSM	1.32	**1.47e-03**	1.15	1.42e-01	1.22	6.40e-02	1.11	3.58e-01	1.62	**3.98e-03**	1.57	1.57e-02	1.15	4.35e-01	1.08	6.99e-01	1.50	1.12e-02	1.44	3.56e-02
VEGFA	1.29	**3.08e-03**	1.22	3.44e-02	1.14	2.26e-01	1.14	2.66e-01	1.47	2.11e-02	1.44	4.85e-02	1.30	1.57e-01	1.31	1.53e-01	1.46	1.80e-02	1.58	8.76e-03
TNFSF14	1.29	**3.12e-03**	1.13	1.94e-01	1.16	1.65e-01	1.08	5.12e-01	1.47	1.55e-02	1.38	7.05e-02	1.17	3.91e-01	1.08	6.79e-01	1.52	7.35e-03	1.44	3.19e-02
LAPTGFbeta1	1.29	**3.20e-03**	1.26	1.43e-02	1.27	2.80e-02	1.29	2.34e-02	1.27	1.69e-01	1.26	1.99e-01	1.32	1.12e-01	1.27	1.64e-01	1.43	2.51e-02	1.52	1.21e-02
IL18R1	1.26	**8.45e-03**	1.27	1.60e-02	1.31	1.64e-02	1.38	8.50e-03	1.35	8.79e-02	1.35	1.47e-01	1.34	1.17e-01	1.31	1.55e-01	1.10	5.53e-01	1.20	3.27e-01
FGF21	1.24	**1.09e-02**	1.22	4.19e-02	1.20	8.97e-02	1.20	1.19e-01	1.18	3.45e-01	1.23	2.87e-01	1.34	1.21e-01	1.32	1.50e-01	1.41	3.85e-02	1.38	7.35e-02
TRAIL	1.21	2.41e-02	1.12	2.13e-01	1.25	3.75e-02	1.20	1.21e-01	1.68	**3.43e-03**	1.69	1.00e-02	1.17	4.03e-01	1.15	4.71e-01	0.97	8.56e-01	1.06	7.27e-01
IL18	1.19	4.22e-02	1.07	4.62e-01	1.13	2.61e-01	1.04	7.49e-01	2.05	**3.96e-05**	1.74	6.25e-03	0.84	3.63e-01	0.79	2.24e-01	1.08	6.55e-01	1.02	8.96e-01

SI Table 7. Logistic regression models with future disease status as the outcome and individual protein levels as predictor in the full population. Results are presented by smoking status. Centre and plate effects are removed from the data by taking the residuals from linear mixed models with protein levels as the outcome and centre and plate as random intercepts. Models are adjusted on age, BMI and gender (base model). For populations including smokers, models further adjusted on packyears are also reported. The odds ratios are expressed as the risk change for a one standard deviation increase in the protein levels. The p-values of association with future disease status are derived from likelihood ratio tests comparing the fit of the model with to that of the model without protein levels in the set of predictors. Results are presented for the twelve proteins found associated with lung cancer in the base model on the full population after Benjamini-Hochberg correction for multiple testing. The numbers of cases and controls used in each model are reported.

	Full population				Never smokers		Current smokers			
	Base model (284/289)		Adjusted on packyears (276/285)		Base model (24/130)		Base model (163/72)		Adjusted on packyears (159/70)	
	OR	p-value	OR	p-value	OR	p-value	OR	p-value	OR	p-value
CDCP1	1.83	**6.29e-11**	1.65	**8.18e-07**	1.44	6.42e-02	2.30	**4.99e-07**	2.26	**5.91e-06**
SCF	0.63	**2.23e-07**	0.82	4.34e-02	1.02	9.27e-01	0.60	**1.58e-03**	0.66	1.27e-02
IL6	1.47	**3.27e-05**	1.30	6.69e-03	1.23	2.70e-01	2.10	**1.22e-04**	1.86	**8.63e-04**
HGF	1.39	**2.01e-04**	1.14	1.84e-01	1.02	9.32e-01	1.31	7.12e-02	1.26	1.46e-01
CD5	1.36	**3.70e-04**	1.19	6.33e-02	0.99	9.60e-01	1.52	7.98e-03	1.50	1.46e-02
MCP1	1.35	**4.88e-04**	1.24	2.59e-02	1.10	6.22e-01	1.51	5.61e-03	1.45	1.85e-02
IL12B	0.75	**8.77e-04**	0.90	2.89e-01	1.08	7.16e-01	0.99	9.65e-01	1.02	8.92e-01
CCL11	1.33	**1.03e-03**	1.08	4.16e-01	0.85	4.29e-01	1.37	3.36e-02	1.25	1.68e-01
MMP10	1.33	**1.10e-03**	1.22	3.71e-02	1.01	9.70e-01	1.08	6.06e-01	1.03	8.35e-01
IL8	1.32	**1.32e-03**	1.32	4.09e-03	1.48	3.13e-02	1.19	2.41e-01	1.21	2.30e-01
CD6	1.31	**1.44e-03**	1.17	8.96e-02	1.12	5.40e-01	1.29	8.95e-02	1.26	1.36e-01
OSM	1.32	**1.47e-03**	1.15	1.42e-01	1.10	6.50e-01	1.33	5.69e-02	1.28	1.13e-01
VEGFA	1.29	**3.08e-03**	1.22	3.44e-02	1.05	8.07e-01	1.37	3.55e-02	1.44	2.02e-02
TNFSF14	1.29	**3.12e-03**	1.13	1.94e-01	1.10	6.25e-01	1.28	1.03e-01	1.24	1.68e-01
LAPTGFbeta1	1.29	**3.20e-03**	1.26	1.43e-02	1.14	5.12e-01	1.31	6.79e-02	1.49	1.39e-02
IL18R1	1.26	**8.45e-03**	1.27	1.60e-02	1.11	6.18e-01	1.40	2.35e-02	1.53	8.95e-03
FGF21	1.24	**1.09e-02**	1.22	4.19e-02	1.01	9.57e-01	1.23	1.71e-01	1.30	9.99e-02

SI Table 8. Logistic regression models with individual levels of CDCP1 as the outcome and future disease status as predictor in EPIC (validation cohort). Models are adjusted on age and BMI as fixed effects and plate ID and centre as random intercepts (base model). The models are adjusted for smoking status, the only smoking exposure variable available for all (N=450) participants, or packyears for the (N=316) participants for whom this information was available. The p-values of association with future disease status are derived from likelihood ratio tests comparing the fit of the model with to that of the model without disease status in the set of predictors. The numbers of cases and controls in each sample are reported. Analyses are conducted in the full population (A), in Women (B) and in Men (C).

A. Full population

	Pooled			Adenocarcinoma			Small-cell carcinoma			Large-cell carcinoma			Squamous-cell carcinoma		
	N	Sign	p-value	N	Sign	p-value	N	Sign	p-value	N	Sign	p-value	N	Sign	p-value
Base model	225/225	+	8.76e-06	71/71	+	2.24e-04	38/38	+	1.45e-01	29/29	+	5.16e-01	41/41	+	1.44e-02
Adjusted on smoking status	225/225	+	8.16e-06	71/71	+	2.28e-04	38/38	+	1.47e-01	29/29	+	5.31e-01	41/41	+	1.54e-02
Adjusted on packyears	161/155	+	1.66e-03	45/44	+	1.42e-02	31/30	+	2.33e-01	28/27	+	4.76e-01	27/27	+	7.52e-02

B. Women only

	Pooled			Adenocarcinoma			Small-cell carcinoma			Large-cell carcinoma			Squamous-cell carcinoma		
	N	Sign	p-value	N	Sign	p-value	N	Sign	p-value	N	Sign	p-value	N	Sign	p-value
Base model	86/86	+	4.11e-04	36/35	+	6.26e-03	13/13	+	8.69e-02	11/11	+	9.11e-01	11/11	+	2.94e-02
Adjusted on smoking status	86/86	+	4.02e-04	36/35	+	5.80e-03	13/13	+	9.33e-02	11/11	+	9.29e-01	11/11	+	7.20e-02
Adjusted on packyears	52/51	+	2.32e-02	20/20	+	1.83e-01	9/9	+	5.19e-01	10/10	+	7.28e-01	6/6	+	2.81e-02

C. Men only

	Pooled			Adenocarcinoma			Small-cell carcinoma			Large-cell carcinoma			Squamous-cell carcinoma		
	N	Sign	p-value	N	Sign	p-value	N	Sign	p-value	N	Sign	p-value	N	Sign	p-value
Base model	139/139	+	1.81e-03	35/35	+	1.18e-02	25/25	+	4.32e-01	18/18	+	6.33e-01	30/30	+	6.88e-02
Adjusted on smoking status	139/139	+	8.40e-03	35/35	+	1.26e-02	25/25	+	4.37e-01	18/18	+	6.51e-01	30/30	+	6.76e-02
Adjusted on packyears	109/104	+	2.32e-02	25/24	+	3.08e-02	22/21	+	3.25e-01	18/17	+	7.64e-01	21/21	+	2.18e-01

SI Table 9. Logistic regression models with future lung cancer status as the outcome and individual protein levels as predictor by time-to-diagnosis sub groups. Centre and plate effects are removed from the data by taking the residuals from linear mixed models with protein levels as the outcome and centre and plate as random intercepts. Models are adjusted on age and BMI (base model). Models further adjusted on packyears are also reported. The odds ratios are expressed as the risk change for a one standard deviation increase in the protein levels. The p-values of association with future disease status are derived from likelihood ratio tests comparing the fit of the model with to that of the model without protein levels in the set of predictors. Results are presented for all Women and comparing the cases with a time-to-diagnosis above or below the median (4.9 years) to all controls. Results are presented for the twelve proteins found associated with lung cancer in the base model on all Women after Benjamini-Hochberg correction for multiple testing. The numbers of cases and controls used in each model are reported.

	All Women				Time-to-diagnosis below 4.9 years				Time-to-diagnosis above 4.9 years			
	Base model (196/201)		Adjusted on packyears (191/197)		Base model (96/201)		Adjusted on packyears (94/197)		Base model (100/201)		Adjusted on packyears (97/197)	
	OR	p-value	OR	p-value	OR	p-value	OR	p-value	OR	p-value	OR	p-value
CDCP1	1.94	**5.49e-09**	1.58	**3.09e-04**	1.67	**1.39e-04**	1.35	5.45e-02	2.28	**4.18e-09**	1.91	**1.67e-05**
SCF	0.62	**1.02e-05**	0.78	3.94e-02	0.64	**3.75e-04**	0.80	1.22e-01	0.63	**2.18e-04**	0.75	4.44e-02
HGF	1.43	**6.82e-04**	1.20	1.19e-01	1.43	4.74e-03	1.19	2.20e-01	1.38	9.72e-03	1.23	1.24e-01
IL6	1.46	**7.63e-04**	1.28	3.27e-02	1.45	3.28e-03	1.23	1.29e-01	1.33	2.12e-02	1.30	5.03e-02
OSM	1.41	**1.09e-03**	1.25	5.98e-02	1.53	**9.08e-04**	1.35	3.49e-02	1.27	6.21e-02	1.16	3.00e-01
MCP1	1.38	**2.12e-03**	1.24	6.62e-02	1.29	4.82e-02	1.14	3.61e-01	1.49	**2.03e-03**	1.39	1.86e-02
IL8	1.35	**3.84e-03**	1.34	1.16e-02	1.40	6.14e-03	1.31	4.38e-02	1.28	4.45e-02	1.30	5.18e-02
VEGFA	1.33	**5.39e-03**	1.22	9.00e-02	1.36	1.51e-02	1.26	9.33e-02	1.29	4.31e-02	1.22	1.52e-01
TWEAK	0.76	**6.47e-03**	0.92	4.94e-01	0.74	1.69e-02	0.92	5.66e-01	0.77	3.83e-02	0.93	5.95e-01
IL12B	0.75	**6.65e-03**	0.91	4.33e-01	0.73	1.27e-02	0.92	5.53e-01	0.78	5.65e-02	0.92	5.67e-01
CD6	1.32	**7.08e-03**	1.16	1.99e-01	1.28	5.04e-02	1.11	4.75e-01	1.38	1.10e-02	1.24	1.21e-01
CD5	1.32	**7.41e-03**	1.17	1.72e-01	1.21	1.23e-01	1.05	7.50e-01	1.45	3.14e-03	1.35	3.10e-02

SI Table 10. Logistic regression models with future lung cancer status as the outcome and individual protein levels as predictor by cohort. Centre and plate effects are removed from the data by taking the residuals from linear mixed models with protein levels as the outcome and centre and plate as random intercepts. Models are adjusted on age and BMI. Models further adjusted on packyears are also reported. The odds ratios are expressed as the risk change for a one standard deviation increase in the protein levels. The p-values of association with future disease status are derived from likelihood ratio tests comparing the fit of the model with to that of the model without protein levels in the set of predictors. Results are presented for all Women and by cohort. We only list proteins found at least once associated to lung cancer status after Benjamini-Hochberg correction for multiple-testing. The numbers of cases and controls used in each model are reported.

	All Women				NOWAC Women				EPIC Women			
	Base model (196/201)		Adjusted on packyears (191/197)		Base model (114/119)		Adjusted on packyears (112/116)		Base model (82/82)		Adjusted on packyears (79/81)	
	OR	p-value	OR	p-value	OR	p-value	OR	p-value	OR	p-value	OR	p-value
CDCP1	1.94	**5.49e-09**	1.58	**3.09e-04**	1.73	**8.04e-05**	1.43	2.26e-02	2.49	**6.44e-06**	2.05	2.22e-03
SCF	0.62	**1.02e-05**	0.78	3.94e-02	0.66	2.81e-03	0.83	2.23e-01	0.58	**1.94e-03**	0.71	9.60e-02
HGF	1.43	**6.82e-04**	1.20	1.19e-01	1.33	3.63e-02	1.15	3.50e-01	1.82	**6.51e-04**	1.48	5.08e-02
IL6	1.46	**7.63e-04**	1.28	3.27e-02	1.25	1.08e-01	1.18	2.50e-01	2.19	**9.21e-05**	1.67	1.23e-02
OSM	1.41	**1.09e-03**	1.25	5.98e-02	1.46	5.83e-03	1.33	6.21e-02	1.37	5.92e-02	1.16	4.15e-01
MCP1	1.38	**2.12e-03**	1.24	6.62e-02	1.08	5.81e-01	1.03	8.20e-01	2.69	**3.31e-07**	2.16	**4.41e-04**
IL8	1.35	**3.84e-03**	1.34	1.16e-02	1.44	8.81e-03	1.48	1.19e-02	1.37	7.34e-02	1.19	3.38e-01
VEGFA	1.33	**5.39e-03**	1.22	9.00e-02	1.20	1.67e-01	1.15	3.32e-01	1.71	**1.79e-03**	1.44	6.27e-02
TWEAK	0.76	**6.47e-03**	0.92	4.94e-01	0.65	1.96e-03	0.80	1.44e-01	0.94	6.91e-01	1.14	4.74e-01
IL12B	0.75	**6.65e-03**	0.91	4.33e-01	0.77	6.72e-02	0.96	8.00e-01	0.77	9.92e-02	0.87	4.36e-01
CD6	1.32	**7.08e-03**	1.16	1.99e-01	1.43	8.53e-03	1.20	2.40e-01	1.13	4.30e-01	1.11	5.60e-01
CD5	1.32	**7.41e-03**	1.17	1.72e-01	1.35	2.42e-02	1.21	1.97e-01	1.28	1.26e-01	1.11	5.58e-01
IL18	1.29	1.26e-02	1.15	2.26e-01	1.12	3.82e-01	1.02	9.15e-01	1.76	**8.04e-04**	1.56	1.91e-02
MMP10	1.26	2.72e-02	1.18	1.65e-01	1.09	5.19e-01	0.99	9.35e-01	1.61	**6.26e-03**	1.51	3.24e-02
CCL20	1.25	2.82e-02	1.14	2.33e-01	1.07	6.32e-01	1.00	9.83e-01	1.57	**6.32e-03**	1.43	4.87e-02
uPA	1.17	1.19e-01	1.09	4.73e-01	0.97	8.19e-01	0.94	6.50e-01	1.58	**6.15e-03**	1.40	7.97e-02
LIFR	1.12	2.75e-01	1.31	1.99e-02	0.86	2.59e-01	1.06	6.87e-01	1.72	**3.62e-03**	1.97	1.80e-03

SI Table 11. Scores of the Reactome pathways and Biological Processes significantly associated with CDCP1 after correction for multiple testing using the Effective Number of Tests (ENT=109 for Reactome and 140 for Biological Processes). The number of transcripts in the functional group, number of Principal Components explaining more than 5% of the group's variance, Principal Component order and corresponding percentage of explained variance (e.v.) are reported, along with the absolute effect size and p-value.

A.

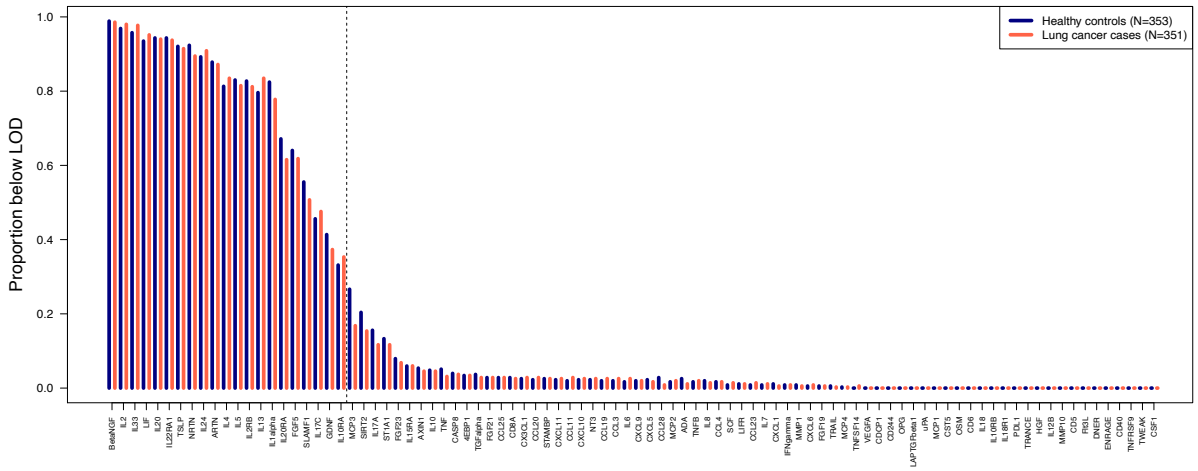
Pathway	Number of transcripts	Number of PCs	PC index	Percentage e.v.	$\lvert\beta\rvert$	p-value
Initial triggering of complement	11	8	2	13.05%	0.31	2.56e-04
Defective GALNT3 causes familial hyperphosphatemic tumoral calcinosis (HFTC)	6	6	4	13.67%	0.23	3.75e-04
Deactivation of the beta-catenin transactivating complex	30	4	4	5.73%	0.33	4.17e-04

B.

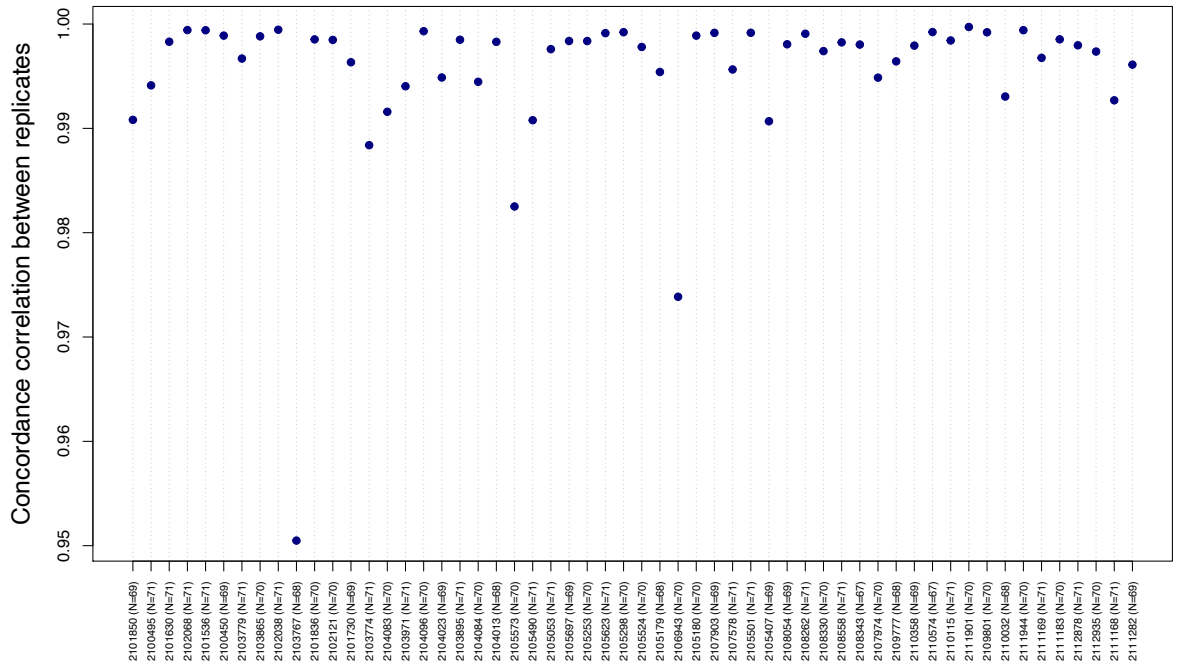
Pathway	Number of transcripts	Number of PCs	PC index	Percentage e.v.	$\lvert\beta\rvert$	p-value
Positive regulation of pseudopodium assembly (GO:0031274)	12	9	3	9.68%	0.31	4.85e-05
Regulation of cell-cell adhesion (GO:0022407)	6	6	3	14.85%	0.27	5.84e-05
Regulation of chemotaxis (GO:0050920)	5	5	5	7.95%	0.18	5.99e-05
Negative regulation of cAMP-dependent protein kinase activity (GO:2000480)	6	5	5	8.73%	0.20	1.20e-04
Positive regulation of dendrite development (GO:1900006)	9	9	6	8.51%	0.24	1.41e-04
Response to fluid shear stress (GO:0034405)	5	5	4	10.5%	0.20	1.49e-04
Positive regulation of metallopeptidase activity (GO:1905050)	6	5	5	11.54%	0.20	1.60e-04
Positive regulation of ERAD pathway (GO:1904294)	5	5	4	15.4%	0.24	1.64e-04
ERAD pathway (GO:0036503)	13	6	5	6.64%	0.25	1.88e-04
Interleukin-18-mediated signaling pathway (GO:0035655)	6	6	5	6.93%	0.17	2.02e-04
Protein localization to nucleus (GO:0034504)	23	5	5	5.15%	0.28	3.27e-04

SI Figure 1. Technical repeatability A. The proportion of samples with levels below the limit of detection (LoD) for each of the 92 proteins is reported for cases (N=351, in red) and controls (N=353, in blue) separately. The 21 proteins with more than 30% of levels below the LoD over the 704 samples are excluded from further analyses. **B.** Lin's concordance correlations between measured protein levels for the 56 samples with replicated measurements. The number of proteins used to compute the correlation for each sample is written in brackets (and ranged from 67 to 71).

A.

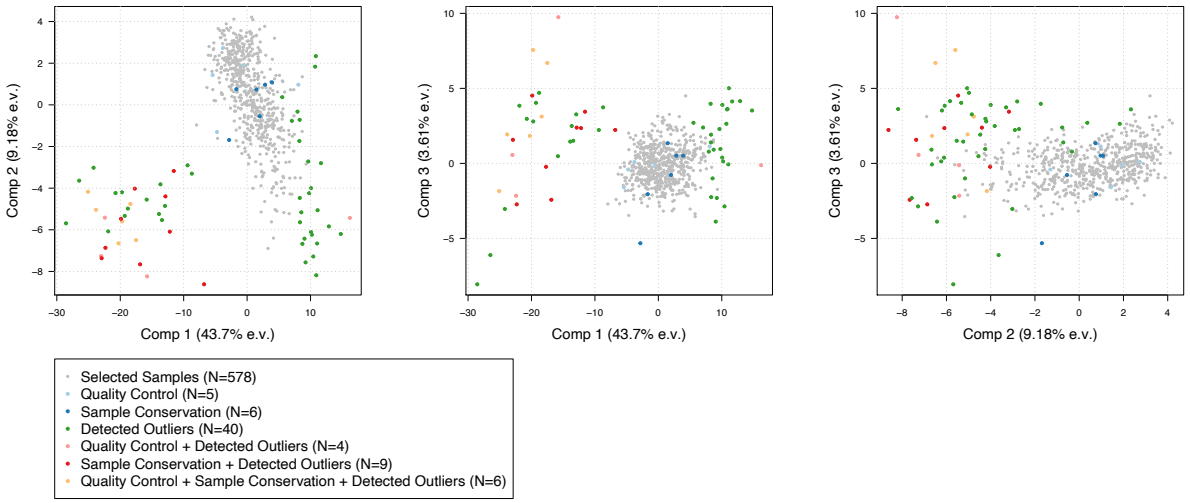


B.

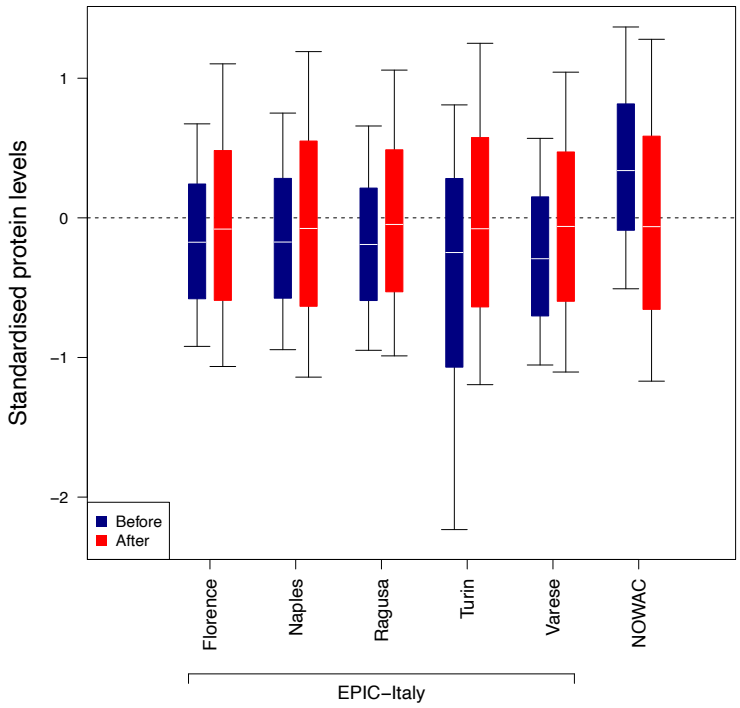


SI Figure 2. Outlier detection and visualisation of the effects of the data preparation. Score plots along the first three PCs of a Principal Component Analysis applied on the 71 proteins measured in the 648 participants are presented (A). We highlight the samples that did not pass the quality control carried out by Olink (N=15, in light blue, pink and orange), the samples with conservation issues due to a default in sample vials (N=21, in dark blue, red and orange), and outliers as detected by the multivariate outlier detection algorithm implemented in the R package mvoutlier applied on the first five PCs (explaining more than 62% of the variance) with an outlier boundary of 0.05 (N=59, in green, pink, red and orange). We represent samples that were both detected as outliers and (i) did not pass quality control (N=4, in pink), (ii) had sample conservation issues (N=9, in red), and (iii) did not pass quality control and had sample conservation issues (N=6, in orange). Boxplots showing the distribution of standardised protein levels by centre (B) and plate (C) before (N=648, dark blue) and after (N=578, red) the exclusion of participants and the denoising by extracting the residuals from linear mixed models with random intercepts on centre and plate.

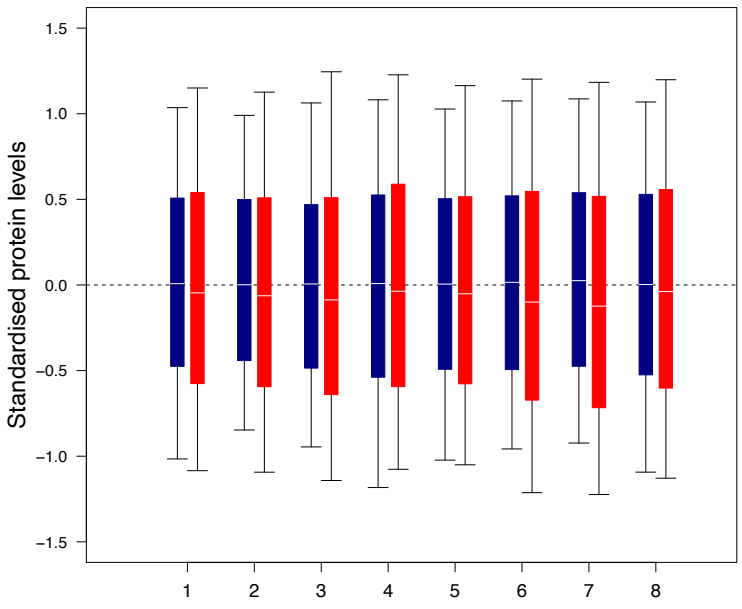
A.



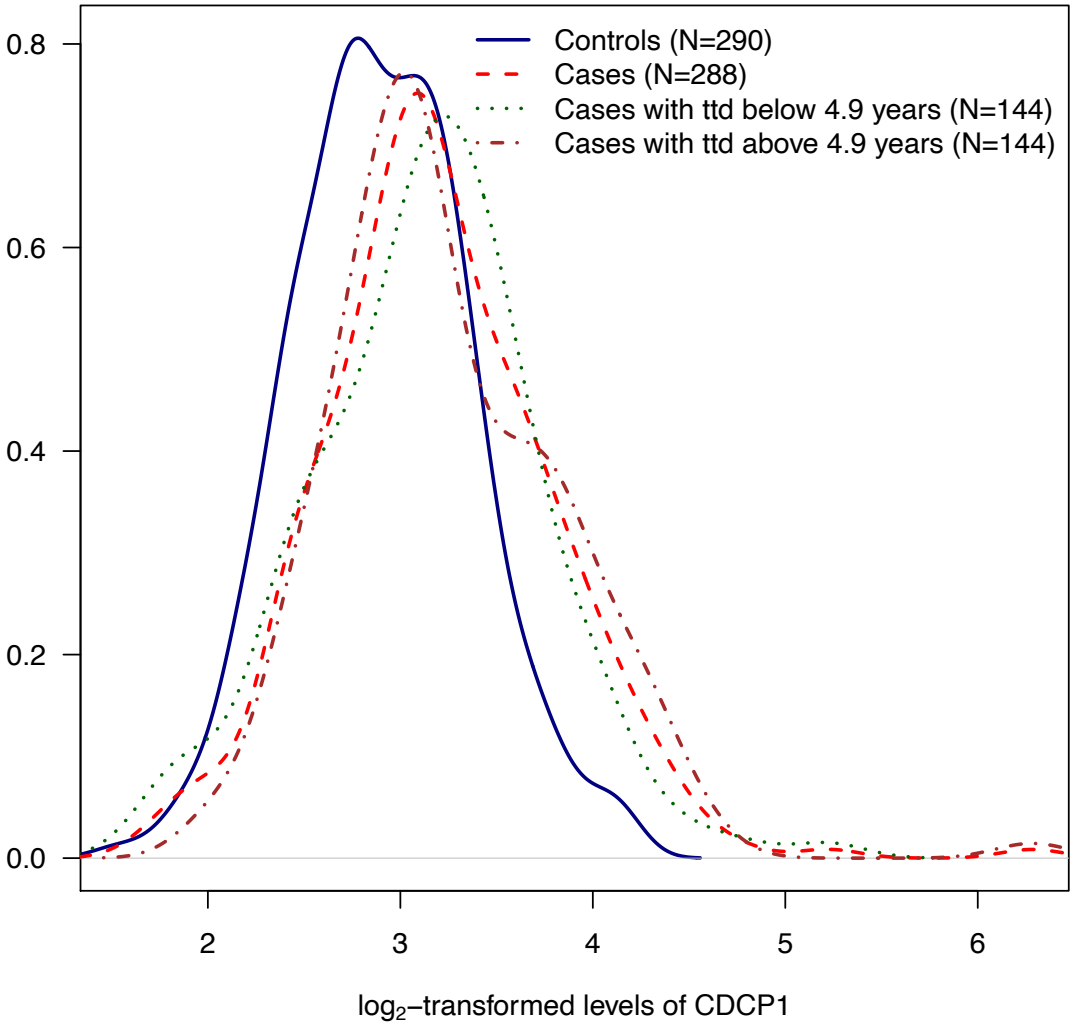
B.


Standardised protein levels
1
0
−1
−2
Before
After
Florence
Naples
Ragusa
Turin
Varese
NOWAC
EPIC−Italy

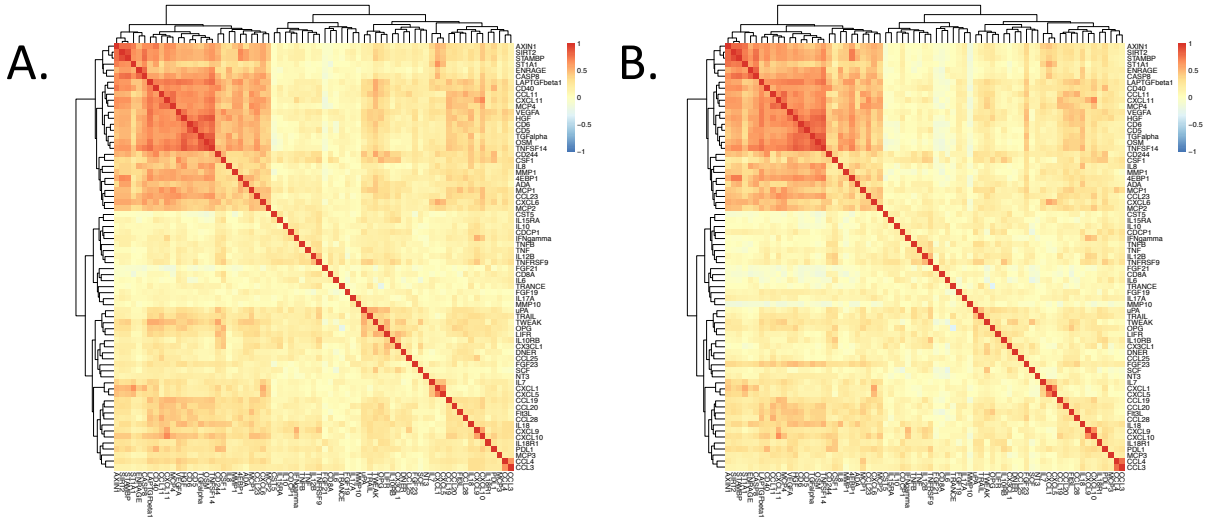
C.


Standardised protein levels
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
1
2
3
4
5
6
7
8

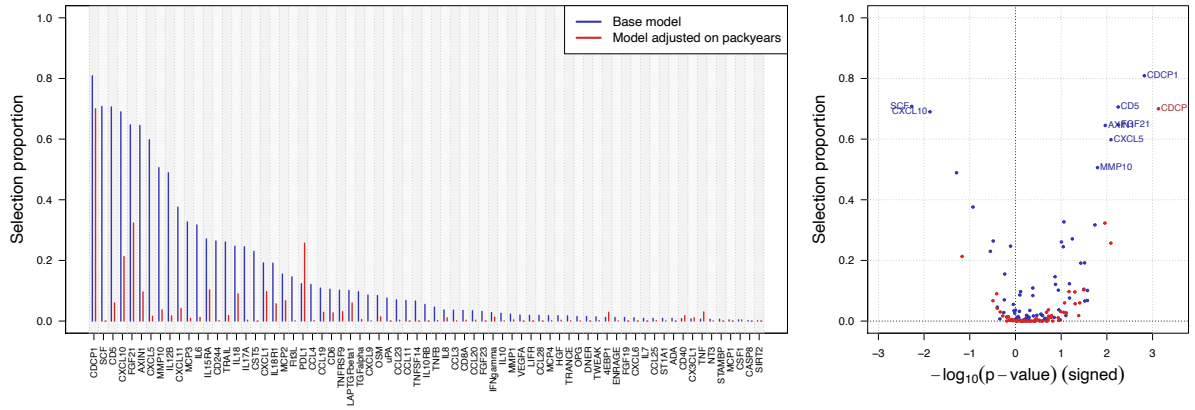
SI Figure 3. Distribution of the circulating levels of CDCP1 in all (N=288) lung cancer cases (in red, dashed line), in (N=144) cases diagnosed less than 4.9 years after enrolment (in green, dotted line), in (N=144) cases diagnosed more than 4.9 years after enrolment (in brown, dashed-dotted line), and in (N=290) healthy controls (in blue, plain line).



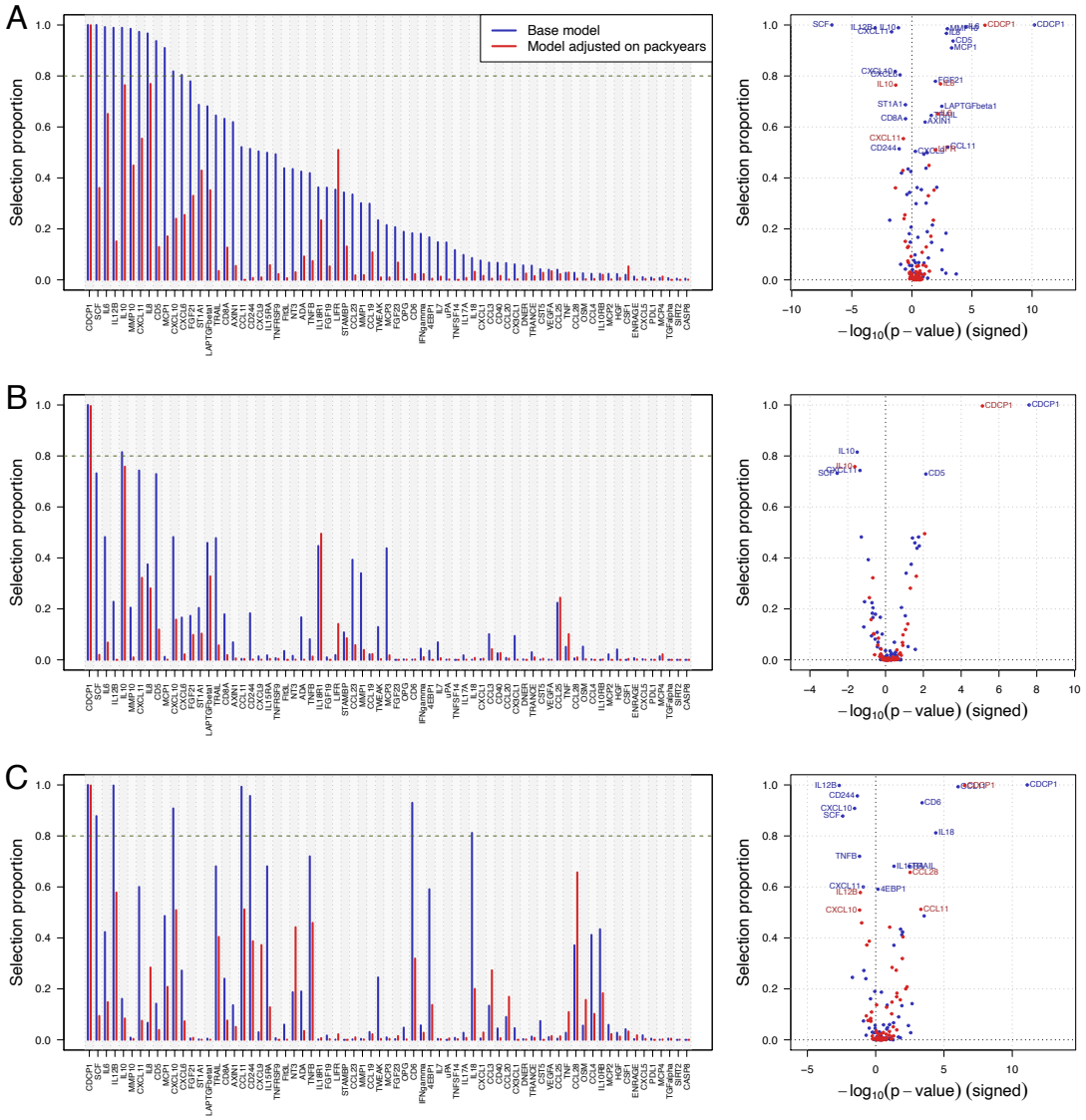
SI Figure 4. Heatmap of Pearson's correlations between the 71 protein levels in controls (A, N=290) and cases (B, N=288) separately.



SI Figure 5. Stability analyses of the logistic-LASSO models investigating the association between the 71 inflammatory proteins and future lung cancer status in Men. Centre and plate effects are removed from the data by taking the residuals from linear mixed models with protein levels as the outcome and centre and plate as random intercepts. The LASSO models are adjusted on age, BMI (in blue, base model) and further adjusted on packyears (in red, model adjusted on packyears) by incorporating these covariates in the model without penalising them. The penalty parameter (lambda) of the models is calibrated at each subsampling iteration using M-fold cross-validation (M=10) to minimise model deviance. Selection proportion of individual proteins are computed over 1,000 random sub-samples of 80% of the sample size (left panel). Selection proportions in LASSO models are compared to the strength of association in univariate models, as measured by their p-value (right panel). Analyses are conducted in participants with complete data on age, BMI and pack years.



SI Figure 6. Stability analyses of the logistic-LASSO models investigating the association between the 71 inflammatory proteins and future lung cancer status in the full population. Centre and plate effects are removed from the data by taking the residuals from preliminary linear mixed models with protein levels as the outcome and centre and plate as random intercepts. The LASSO models are adjusted on age, gender, BMI (in blue, base model) and further adjusted on pack years (in red, model adjusted on pack years) by incorporating these covariates in the model without penalisation. Selection proportion of individual proteins are computed over 1,000 random sub-samples of 80% of the sample size and ensuring that the proportion of cases and controls is kept constant in each subsample (left panel). The penalty parameter (lambda) of the models is calibrated at each subsampling iteration using M-fold cross-validation (M=10) to minimise model deviance. Selection proportions in LASSO models are compared to the strength of association in univariate models, as measured by their p-value (right panel). Analyses are conducted in participants with complete data on age, BMI and pack years. Results are presented for all lung cancer cases (A, N=276 cases) and by subtype (B: Adenocarcinoma, N=125 cases, C: small-cell carcinoma, N=38 cases) compared to the 285 healthy controls.



SI Figure 7. Cross-omics analysis. Volcano plot of the results from linear mixed models regressing CDCP1 (as the outcome) against the measured levels of 11,610 transcripts available in Women from the NOWAC study (N=222). The models are adjusted on technical confounders using plate as a random intercept and on age and BMI as fixed effects. The transcripts showing a association at an FDR controlled (Benjamini Hochberg) level of 0.05 are coloured in dark red.