Identification of pre-diagnostic proteomics markers contributing to the future risk of lung cancer

Sonia Dagnino, **Barbara Bodinier**, Marc Chadeau-Hyam

8^{th} October, 2020

Imperial College London

Data overview

- Prospective case-control study of lung cancer on 648 participants from EPIC Italy (N=382) and NOWAC (N=266, Women only)

- 92 circulatory inflammatory proteins measured with the Olink platform
- Protein levels provided are $\log_2$-transformed (Olink's NPX unit where an increase of 1 NPX corresponds to the doubling of protein concentration)
- Replicated measurements (N=2) for 56 EPIC Italy participants
- Quality control: 16 controls (same sample measured 16 times)

  $\Rightarrow$ A total of 720 measurements (704 samples and 16 controls) over 8 plates (N=90 samples each)

- In the same participants (NOWAC only, N=222): transcriptomics measurements (N=11,610 probes)

Participant characteristics

	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: Female	435 (67.1)	266 (100.0)	169 (100.0)	0 (0.0)
Body Mass Index	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)
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 lung cancer	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	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)
Quality Control: Warning	15 (2.3)	3 (1.1)	1 (0.6)	11 (5.2)
Storage time (years)	14.91 (4.94)	9.03 (0.98)	18.97 (1.17)	18.74 (1.23)
Gel status: poor quality	21 (10.7)	-	4 (5.2)	17 (14.2)

$\Rightarrow$ Differences in smoking habits between Men and Women in EPIC

Participant characteristics

	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: Female	435 (67.1)	266 (100.0)	169 (100.0)	0 (0.0)
Body Mass Index	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)
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 lung cancer	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	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)
Quality Control: Warning	15 (2.3)	3 (1.1)	1 (0.6)	11 (5.2)
Storage time (years)	14.91 (4.94)	9.03 (0.98)	18.97 (1.17)	18.74 (1.23)
Gel status: poor quality	21 (10.7)	-	4 (5.2)	17 (14.2)

$\Rightarrow$ Differences in smoking duration between NOWAC and EPIC Women

Participant characteristics

	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: Female	435 (67.1)	266 (100.0)	169 (100.0)	0 (0.0)
Body Mass Index	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)
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 lung cancer	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	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)
Quality Control: Warning	15 (2.3)	3 (1.1)	1 (0.6)	11 (5.2)
Storage time (years)	14.91 (4.94)	9.03 (0.98)	18.97 (1.17)	18.74 (1.23)
Gel status: poor quality	21 (10.7)	-	4 (5.2)	17 (14.2)

$\Rightarrow$ Differences in time to diagnosis between NOWAC and EPIC

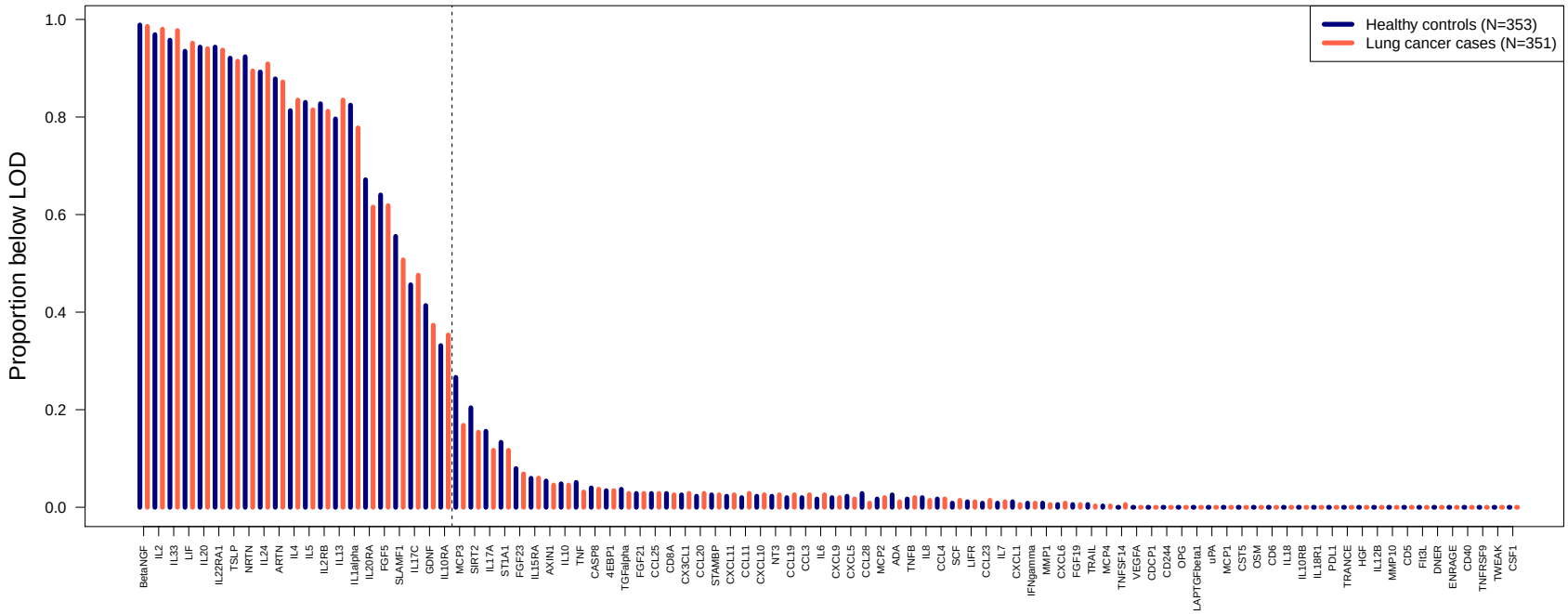
Participant characteristics

	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: Female	435 (67.1)	266 (100.0)	169 (100.0)	0 (0.0)
Body Mass Index	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)
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 lung cancer	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	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)
Quality Control: Warning	15 (2.3)	3 (1.1)	1 (0.6)	11 (5.2)
Storage time (years)	14.91 (4.94)	9.03 (0.98)	18.97 (1.17)	18.74 (1.23)
Gel status: poor quality	21 (10.7)	-	4 (5.2)	17 (14.2)

$\Rightarrow$ Differences in storage time between NOWAC and EPIC

Below detection limit/missing data by disease status

- Proportion of measurements below the limit of detection in cases (N=351 samples from 323 participants) and controls (N=353 samples from 325 participants) separately



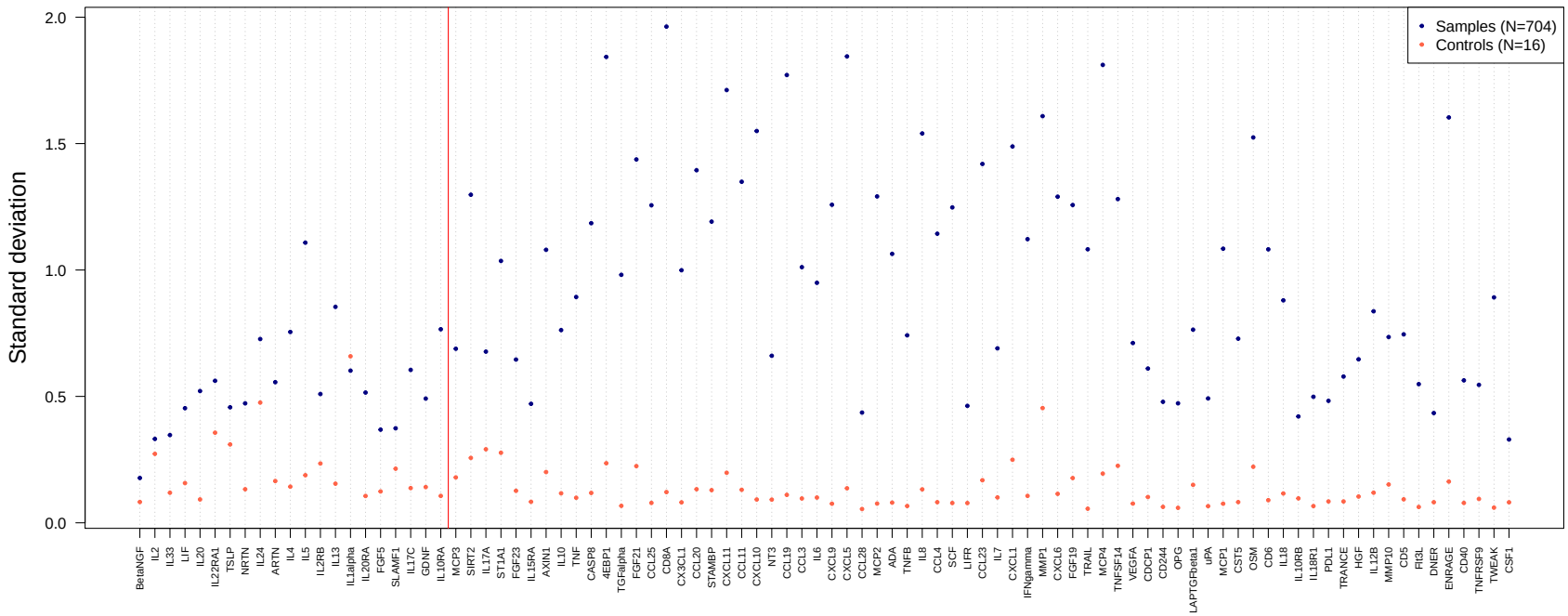
$\Rightarrow$ 21 proteins with more than 30% of measurements below the detection limit

$\Rightarrow$ Additionally, 3 missing values for sample 103590 (warning in QC)

$\Rightarrow$ Very small differences in proportions of missing values between cases and controls

Quality check: control measurements

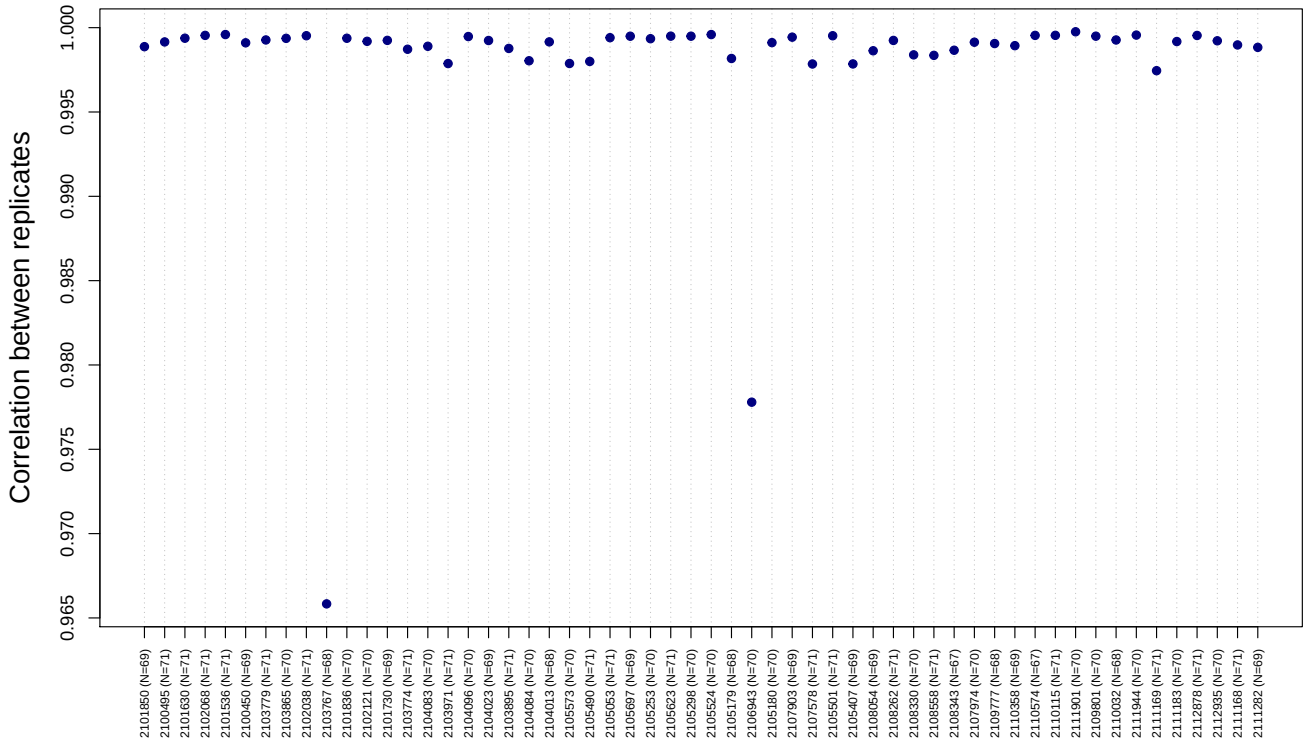
- Standard deviation between samples (dark blue) and controls (red)
- Protein names are ordered by decreasing proportion of samples below the LOD



⇒ More variability between the actual samples than between the controls for all proteins except IL1alpha ($\sim$ 80% of data below the detection limit)

Repeated measurements

- Pearson's correlation between replicates (N=56 samples measured twice) on the 71 proteins with less than 30% of missing values



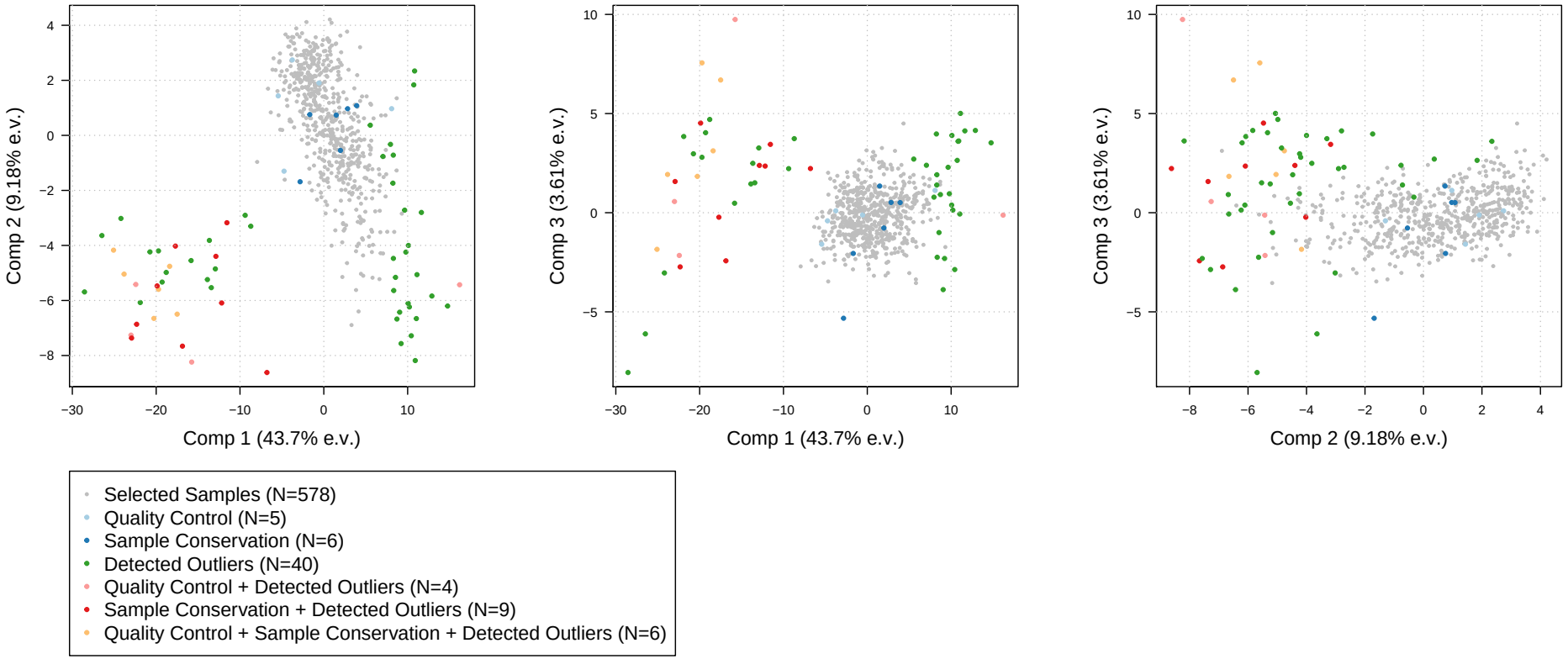
⇒ High consistency between both measurements (correlations above 0.96)

⇒ The average value is taken for replicated measurements

⇒ Remaining missing values are imputed using QRILC for left-censored data

Quality control and outlier detection

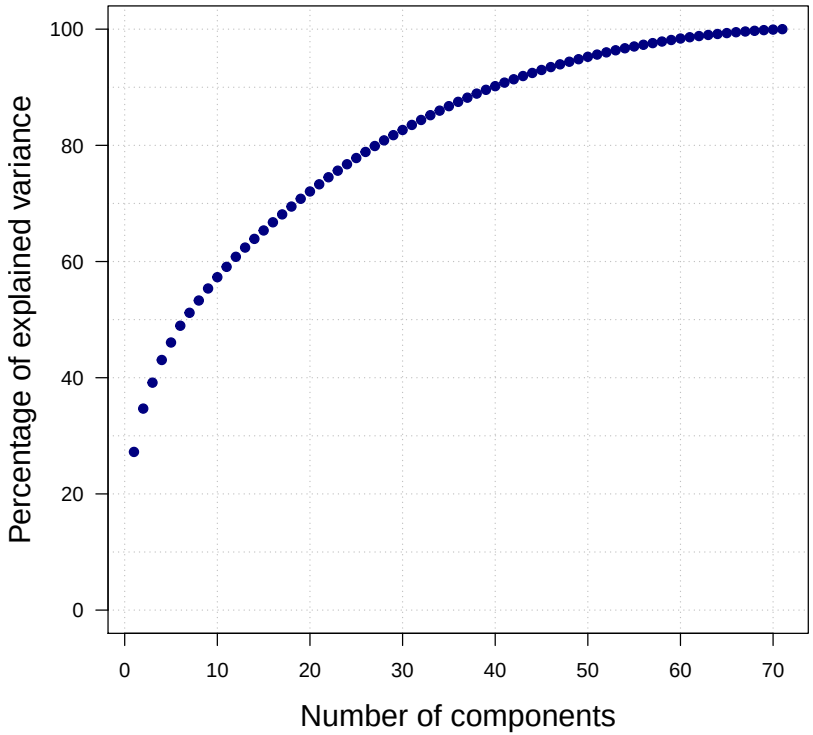
- Principal Component Analysis: score plots along the first 3 PCs
- Detection of outliers along the first 5 PCs using the multivariate distance-based algorithm implemented in the R package mvoutlier

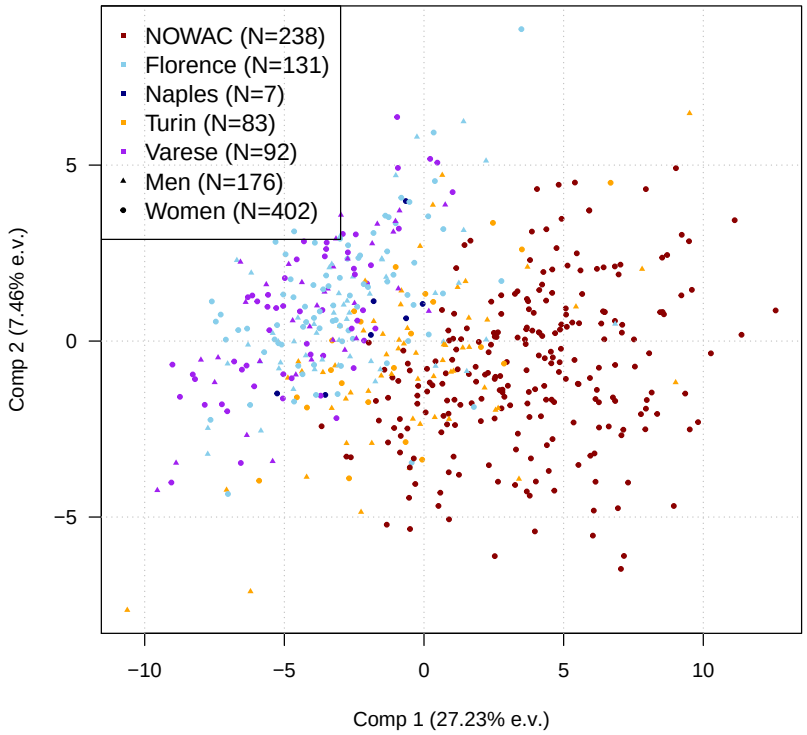


$\Rightarrow$ All 70 samples with poor quality data are removed for further analyses

Principal component analysis after exclusions

- PCA on the imputed data after exclusion of participants (N=578)





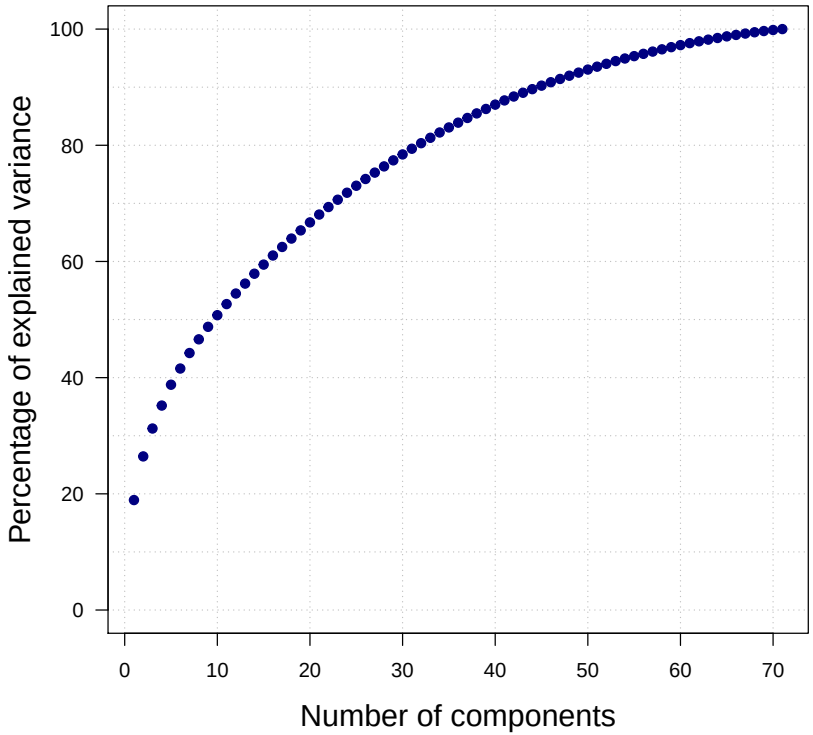
$\Rightarrow$ The first PC explains 27% of the variance

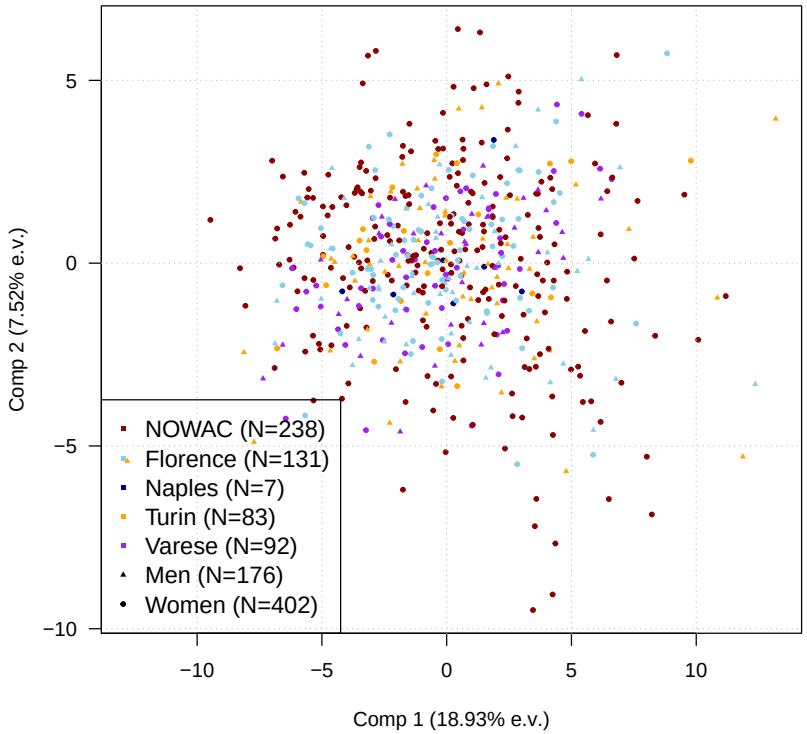
$\Rightarrow$ Score plots show strong differences between EPIC centres and NOWAC

$\Rightarrow$ Need to account for heterogeneity between study/cohorts in the models

Principal component analysis after exclusions

- Denoising of the data by extracting the residuals from a linear mixed model with protein levels (outcome) against plate and centre as random effects
- PCA on the residuals (denoised data)





$\Rightarrow$ The first PC explains 19% of the variance

$\Rightarrow$ Linear mixed models were successful in removing centre effects

Analytical plan

1. Investigating the associations with lung cancer in Women only
2. Accounting for the effects of smoking using packyears and smoking status
3. Sensitivity analyses: stratification by cohort and median time to diagnosis
4. Validation in Men
5. Accounting for joint effects of the proteins in multivariate analyses
6. Stratification on main histological subtypes
7. Validation in external cohorts (EPIC, NSHDS)
8. Evaluation of the complementarity between proteins and packyears in lung cancer status discrimination using ROC curves
9. Exploration of the functional role of lung cancer-related proteins via OMICs integration

Univariate analyses in Women

- Univariate logistic models with lung cancer status/packyears (outcome) against protein levels (predictor) and adjusted on age and BMI

	Lung cancer				Packyears	
	Base model		Adjusted on packyears			
	β	p-value	β	p-value	β	p-value
CDCP1	0.67	**5.49e-09**	0.46	**3.09e-04**	0.23	**2.45e-08**
SCF	-0.46	**1.02e-05**	-0.24	3.94e-02	-0.23	**5.25e-09**
HGF	0.35	**6.82e-04**	0.18	1.19e-01	0.20	**8.93e-07**
IL6	0.36	**7.63e-04**	0.24	3.27e-02	0.11	**6.91e-03**
OSM	0.33	**1.09e-03**	0.21	5.98e-02	0.13	**2.04e-03**
MCP1	0.31	**2.12e-03**	0.20	6.62e-02	0.15	**2.01e-04**
IL8	0.26	**3.84e-03**	0.26	1.16e-02	0.07	6.92e-02
VEGFA	0.28	**5.39e-03**	0.19	9.00e-02	0.13	**1.33e-03**
TWEAK	-0.27	**6.47e-03**	-0.08	4.94e-01	-0.15	**1.78e-04**
IL12B	-0.28	**6.65e-03**	-0.09	4.33e-01	-0.19	**4.95e-06**
CD6	0.26	**7.08e-03**	0.14	1.99e-01	0.14	**2.87e-04**
CD5	0.27	**7.41e-03**	0.15	1.72e-01	0.12	**3.82e-03**

$\Rightarrow$ 12 significant associations with lung cancer after FDR correction

$\Rightarrow$ 11/12 are also associated with packyears

$\Rightarrow$ Overall attenuation of the strength of the association with lung cancer upon adjustment on packyears, only CDCP1 survives adjustment

Investigating the effects on smoking

- Stratified analyses by smoking status

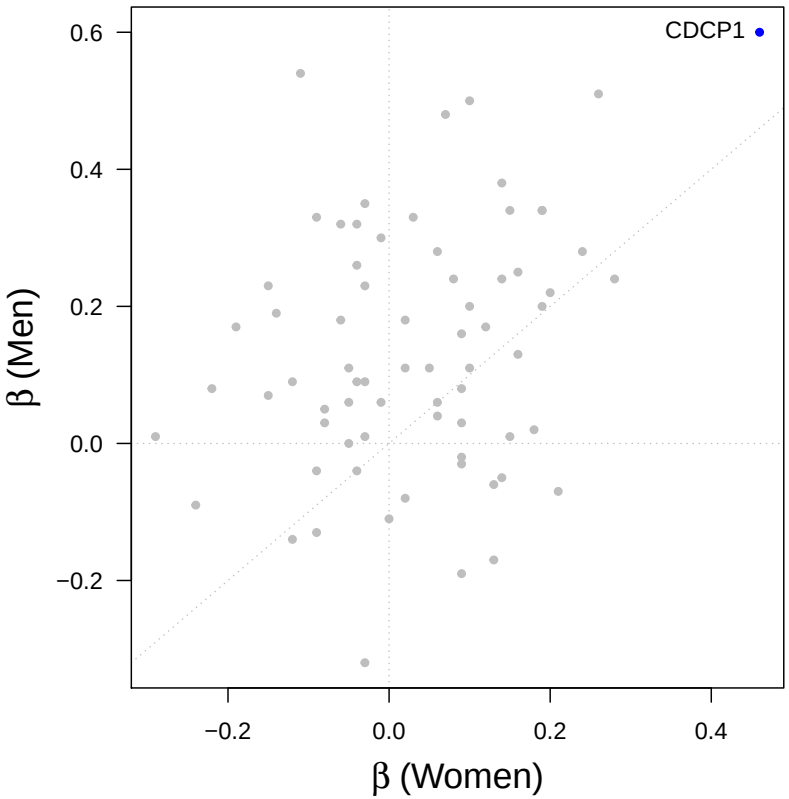
	All Women				Never smoking Women		Current smoking Women			
	Base model (N=397)		Adjusted on packyears (N=388)		Base model (N=132)		Base model (N=169)		Adjusted on packyears (N=163)	
	β	p-value	β	p-value	β	p-value	β	p-value	β	p-value
CDCP1	0.67	**5.49e-09**	0.46	**3.09e-04**	0.40	7.78e-02	0.80	**4.91e-05**	0.80	**2.00e-04**
SCF	-0.46	**1.02e-05**	-0.24	3.94e-02	0.01	9.68e-01	-0.70	**5.31e-04**	-0.64	**1.91e-03**
HGF	0.35	**6.82e-04**	0.18	1.19e-01	0.12	6.26e-01	0.28	8.13e-02	0.26	1.20e-01
IL6	0.36	**7.63e-04**	0.24	3.27e-02	0.25	2.23e-01	0.90	**1.34e-04**	0.81	**6.07e-04**
OSM	0.33	**1.09e-03**	0.21	5.98e-02	0.20	3.46e-01	0.35	3.94e-02	0.35	4.70e-02
MCP1	0.31	**2.12e-03**	0.20	6.62e-02	0.21	3.63e-01	0.37	1.25e-02	0.34	2.83e-02
IL8	0.26	**3.84e-03**	0.26	1.16e-02	0.49	1.69e-02	0.14	2.86e-01	0.15	2.68e-01
VEGFA	0.28	**5.39e-03**	0.19	9.00e-02	0.17	4.56e-01	0.20	1.98e-01	0.23	1.48e-01
TWEAK	-0.27	**6.47e-03**	-0.08	4.94e-01	0.08	6.92e-01	-0.26	1.44e-01	-0.21	2.86e-01
IL12B	-0.28	**6.65e-03**	-0.09	4.33e-01	0.14	5.62e-01	-0.14	4.50e-01	-0.13	4.84e-01
CD6	0.26	**7.08e-03**	0.14	1.99e-01	0.12	5.08e-01	0.16	3.18e-01	0.15	3.79e-01
CD5	0.27	**7.41e-03**	0.15	1.72e-01	0.02	9.27e-01	0.25	1.40e-01	0.26	1.27e-01

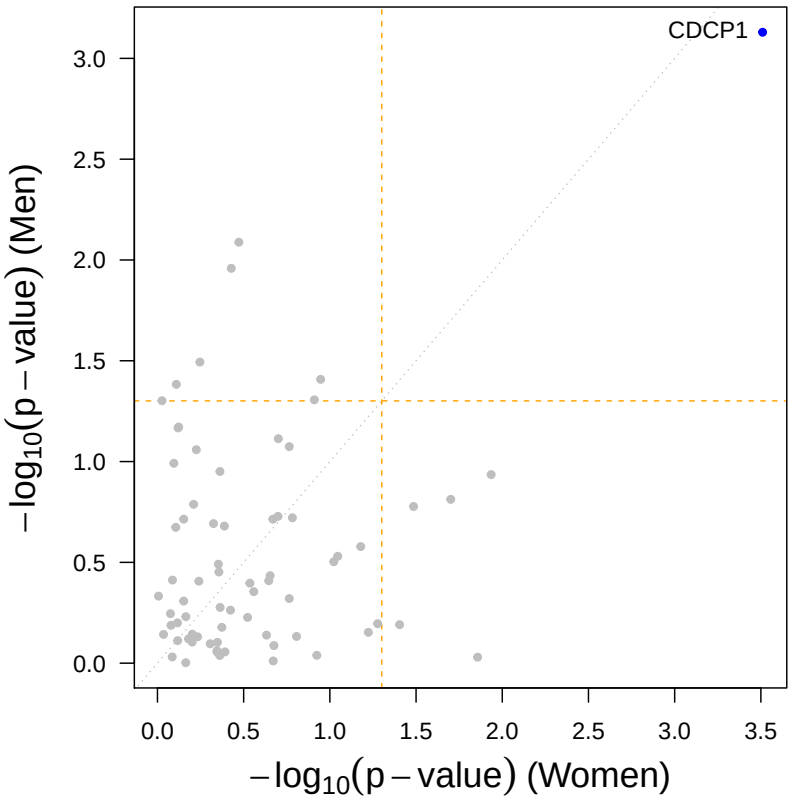
$\Rightarrow$ No significant association in never smoking Women

$\Rightarrow$ CDCP1, SCF and IL6 are significantly associated with future lung cancer status in current smoking Women

Validation in Men

- Analyses adjusted on packyears are conducted in Women (N=388) and Men (N=173) separately

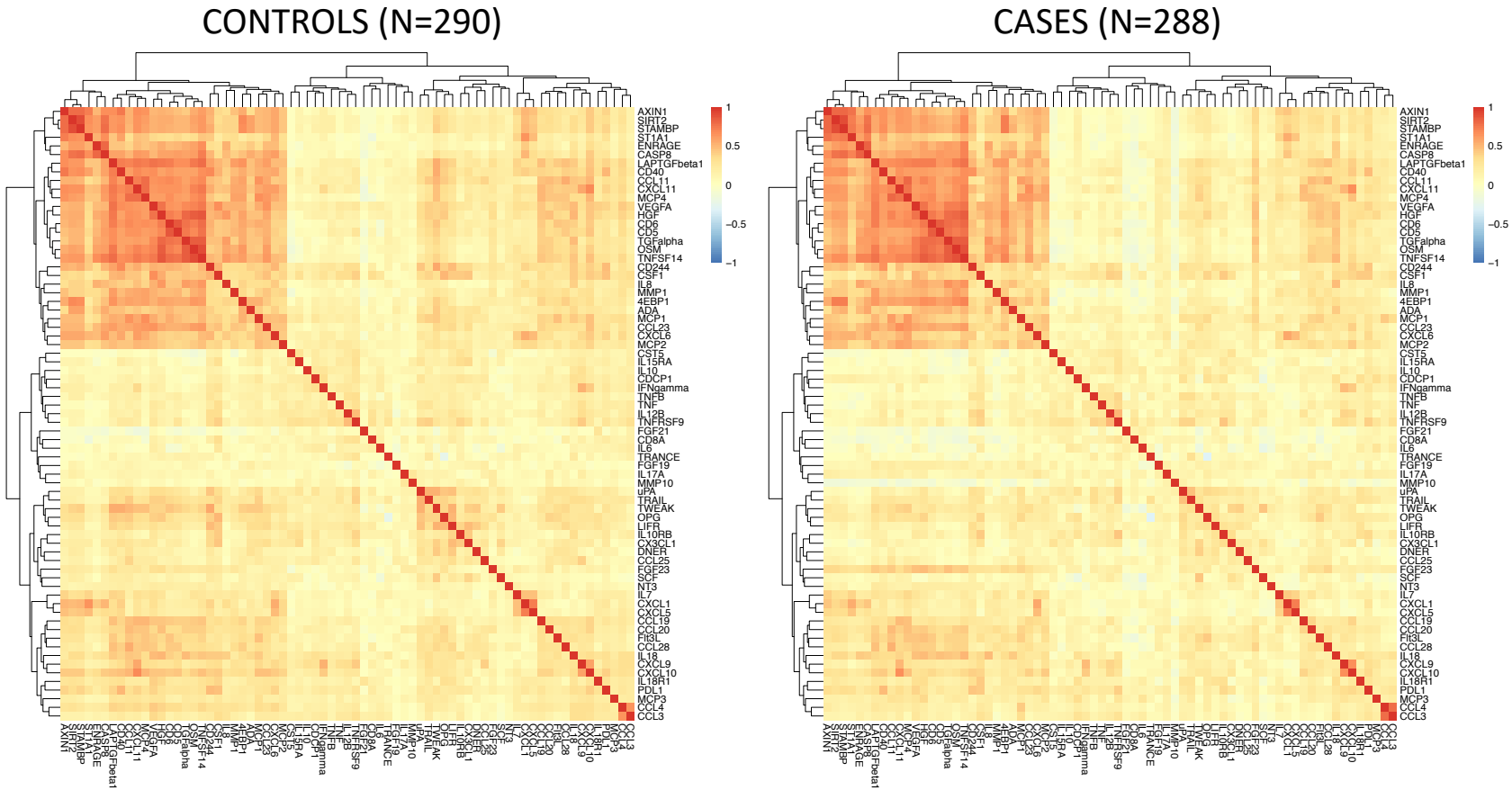




$\Rightarrow$ CDCP1 is nominally significant in Men

Pairwise correlations between proteins

- Heatmap of Pearson's correlations between the imputed levels of the 71 inflammatory proteins in controls (left) and cases (right)
- Hierarchical clustering performed in healthy controls

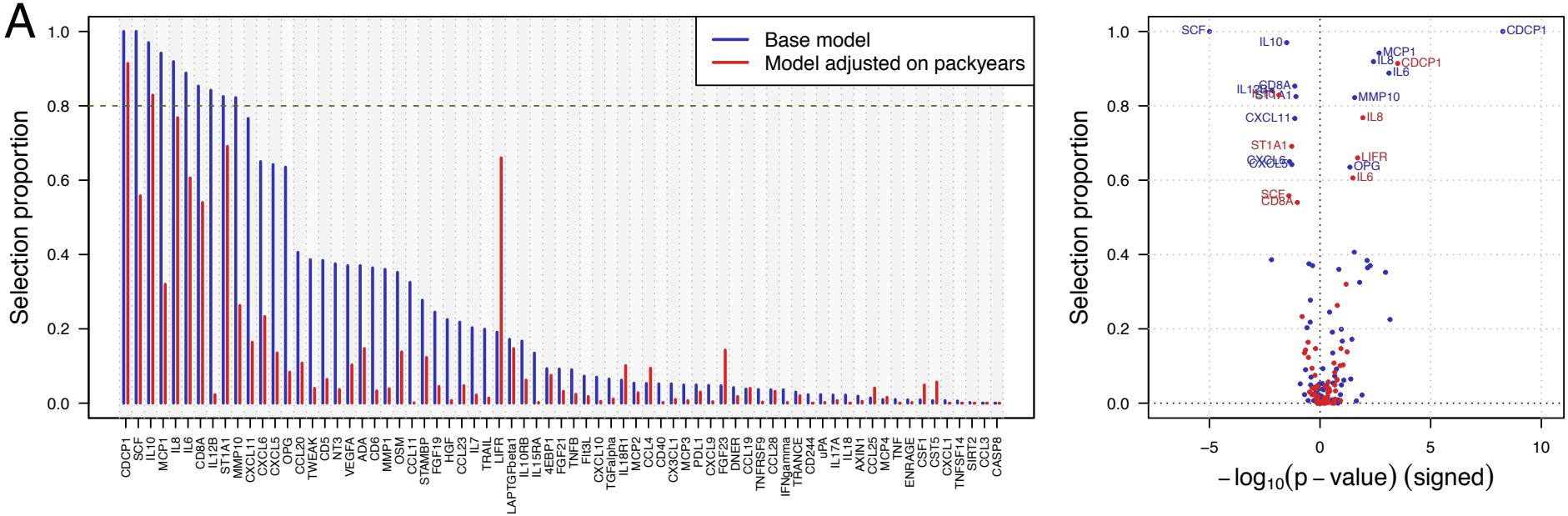


$\Rightarrow$ Overall, similar correlation patterns in cases and controls

$\Rightarrow$ Some correlated proteins, need to account for this in the models

Multivariate analyses

- Logistic-LASSO on lung cancer against all proteins, adjusted on age and BMI
- Variable selection used in combination with stability analyses (1,000 iterations on subsamples of 80% of the data) to derive selection proportions in the base and further adjusted on packyears models

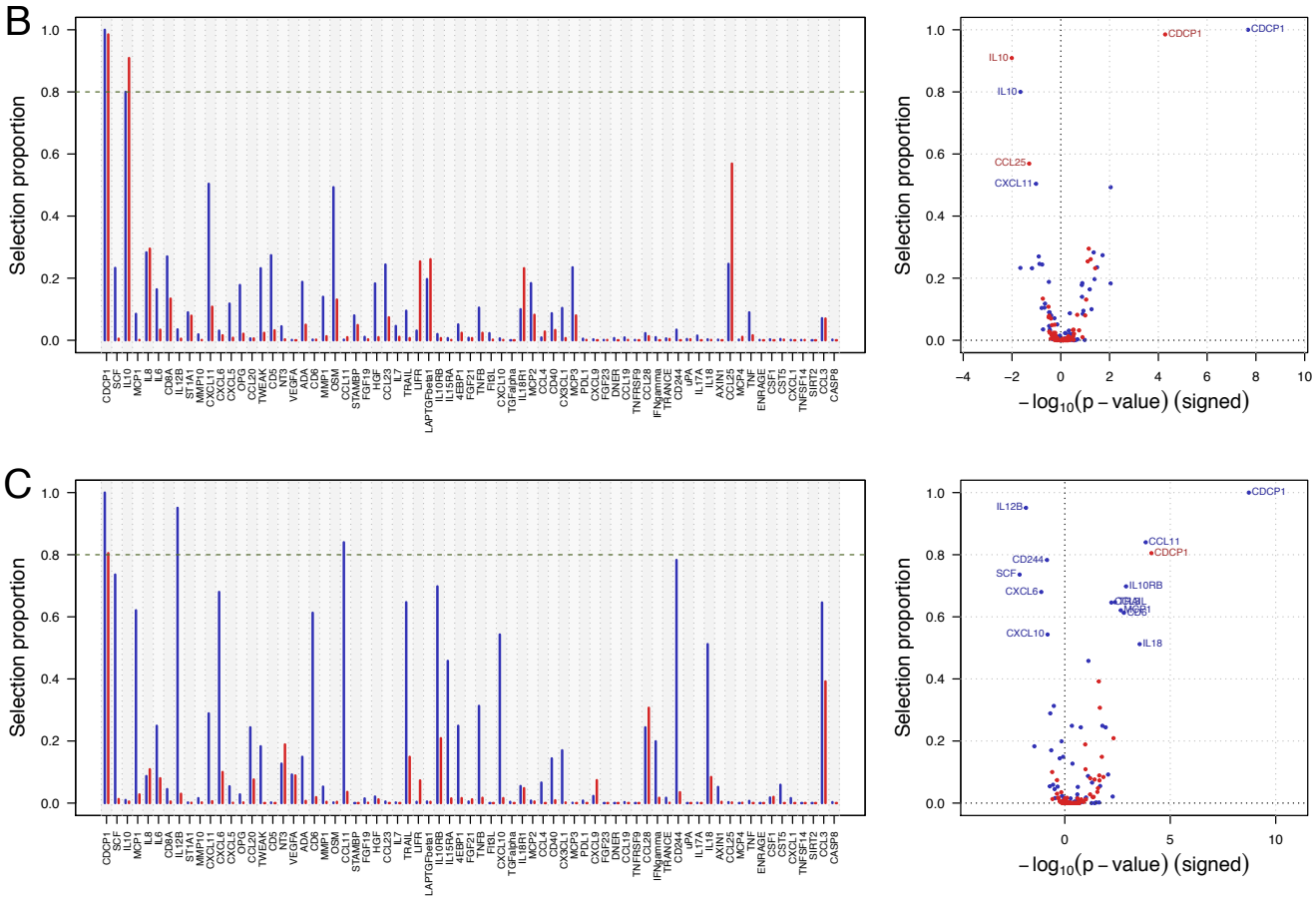


$\Rightarrow$ Good consistency with univariate results

$\Rightarrow$ CDCP1 and IL10 in the model adjusted on smoking

Analyses by histological subtypes

- Stability analyses of the logistic-LASSO stratified by subtype: adenocarcinoma (N=91 cases) and small-cell carcinoma (N=32 cases)



$\Rightarrow$ CDCP1 is highly selected but re-ordering of the other signals suggesting heterogeneity between the subtypes

Validation of CDCP1 in external cohorts

- Validation using data from two external cohorts: EPIC (Netherlands, UK, Germany, Spain) and NSHDS (Northern Sweden Health and Disease Study)
- Logistic models adjusted on age and BMI (base model), and further adjusted on packyears or smoking status

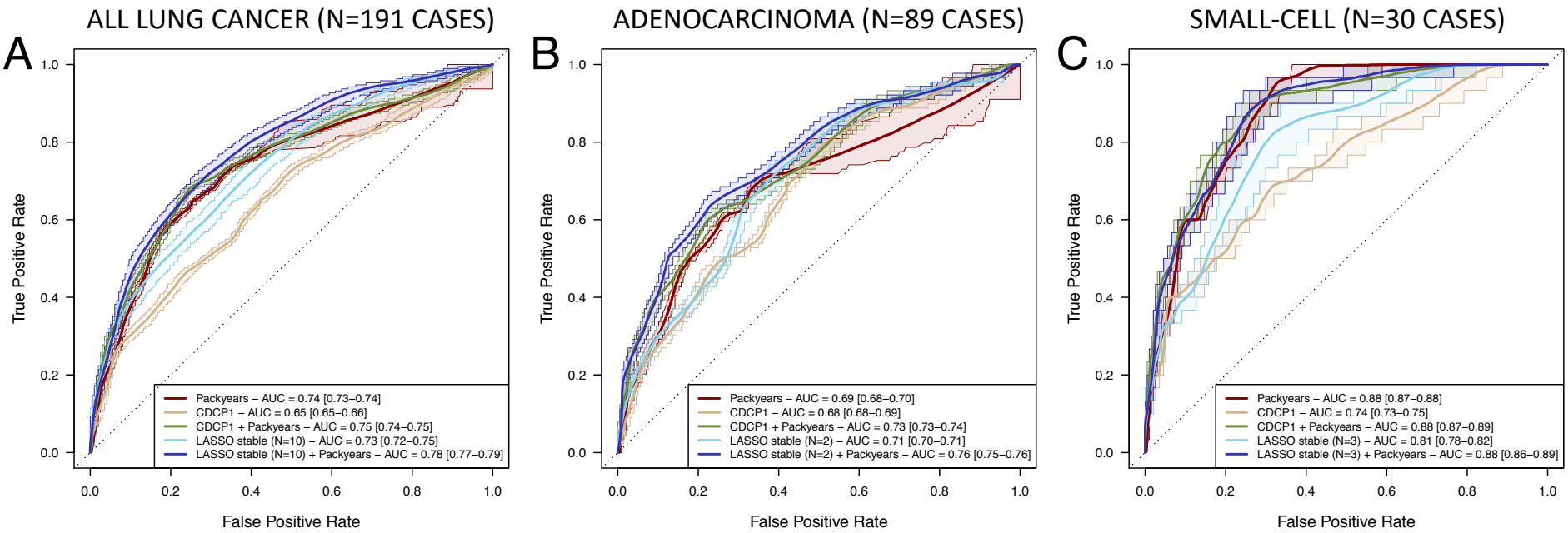
	Pooled			Adenocarcinoma			Small-cell carcinoma		
	N	Sign	p-value	N	Sign	p-value	N	Sign	p-value
Full Population									
Base model	225/225	+	8.76e-06	71/71	+	2.24e-04	38/38	+	1.45e-01
Adjusted on smoking status	225/225	+	8.16e-06	71/71	+	2.28e-04	38/38	+	1.47e-01
Adjusted on packyears	161/155	+	1.66e-03	45/44	+	1.42e-02	31/30	+	2.33e-01
Women									
Base model	86/86	+	4.11e-04	36/35	+	6.26e-03	13/13	+	8.69e-02
Adjusted on smoking status	86/86	+	4.02e-04	36/35	+	5.80e-03	13/13	+	9.33e-02
Adjusted on packyears	52/51	+	2.32e-02	20/20	+	1.83e-01	9/9	+	5.19e-01
Men									
Base model	139/139	+	1.81e-03	35/35	+	1.18e-02	25/25	+	4.32e-01
Adjusted on smoking status	139/139	+	8.40e-03	35/35	+	1.26e-02	25/25	+	4.37e-01
Adjusted on packyears	109/104	+	2.32e-02	25/24	+	3.08e-02	22/21	+	3.25e-01

$\Rightarrow$ Associations with all lung cancer survive adjustment on smoking

$\Rightarrow$ Significant associations with adenocarcinoma despite small sample size

Quantifying the amount of disease-relevant information

- ROC analyses with packyears, CDCP1 and LASSO-selected proteins to quantify the amount of disease-relevant information brought by proteins



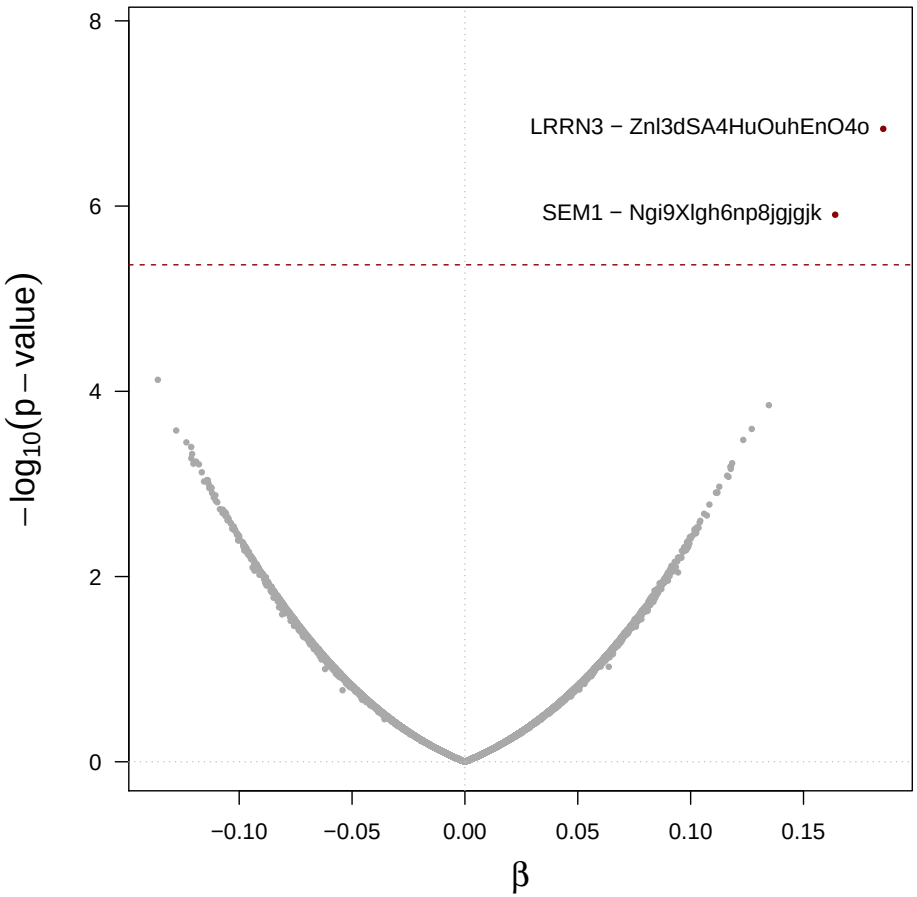
$\Rightarrow$ CDCP1 alone yields an AUC of 0.65 (all LC), 0.68 (adenocarcinoma) and 0.74 (small-cell carcinoma)

$\Rightarrow$ Increase from 0.69 to 0.73 with CDCP1 on top of packyears (adenocarcinoma)

$\Rightarrow$ Moderate additional information with more proteins (N=10, AUC going from 0.75 to 0.78)

Integration of CDCP1 with transcriptomics

- Univariate analyses of CDCP1 against all N=11,610 transcripts measured in the same participants (N=222, NOWAC)



⇒ Significant association of CDCP1 with LRRN3 (marker of tobacco smoking) and SEM1

Exploration of the functional role of CDCP1

- 11,610 transcripts measured in the same participants (N=222, NOWAC)
- Transcript nuIDs could be linked to 11,485 unique gene symbols
- 10,656 of these gene symbols could be identified on the Panther Database

- Functional annotation based on classifications from knowledgebases:
  - Biological Process
  - Reactome

$\Rightarrow$ Each classification provides different levels of grouping of the genes (classification in groups and sub-groups of genes)

$\Rightarrow$ One gene can belong to different groups within each classification

Associations between CDCP1 and Reactome pathways

- Grouping of the transcripts by Reactome pathways

  $\Rightarrow$ 1,545 functional groups involving 6,401 unique genes (some of these pathways are made of just one gene)

- Summary of each group using PCA: all components explaining more than 5% of the variance of the group are kept

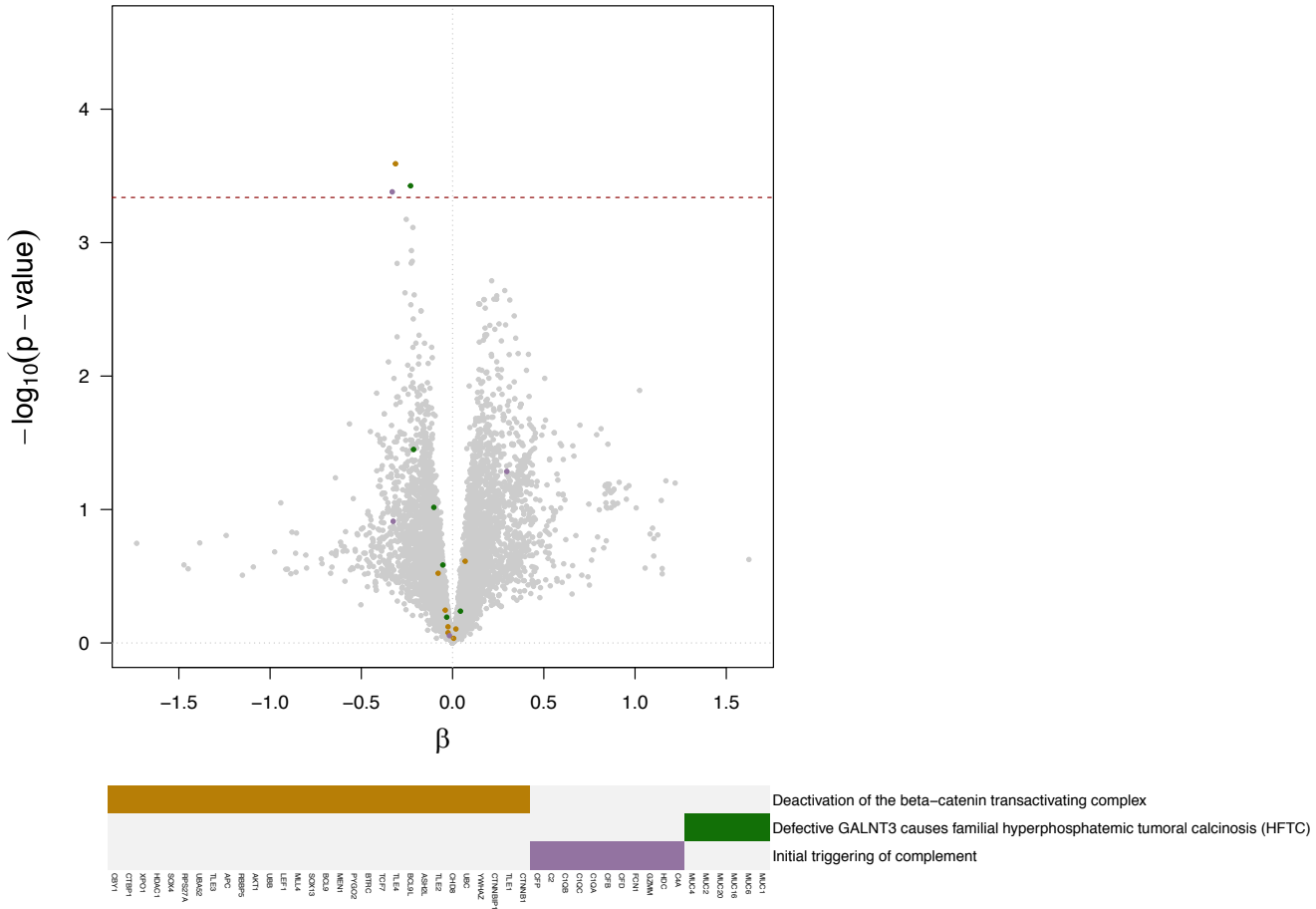
  $\Rightarrow$ 8,043 PC scores summarising the pathways (new dataset)

  $\Rightarrow$ Number of PCs per pathway ranging between 1 and 10

- Univariate regressions of CDCP1 against PC scores summarising the groups
- Estimation of the Effective Number of Test with the number of PCs to explain 90% of the variance over the 8,043 scores (ENT=109)

Associations between CDCP1 and Reactome pathways

- Univariate regressions of CDCP1 against PC scores summarising Reactome pathways



⇒ Identification of 3 groups including 6-30 transcripts that were not detected in univariate analyses

Associations between CDCP1 and Biological Processes

- Grouping of the transcripts by Biological Processes

  $\Rightarrow$ 3,600 functional groups involving 9,826 unique genes (some of these pathways are made of just one gene)

- Summary of each group using PCA: all components explaining more than 5% of the variance of the group are kept

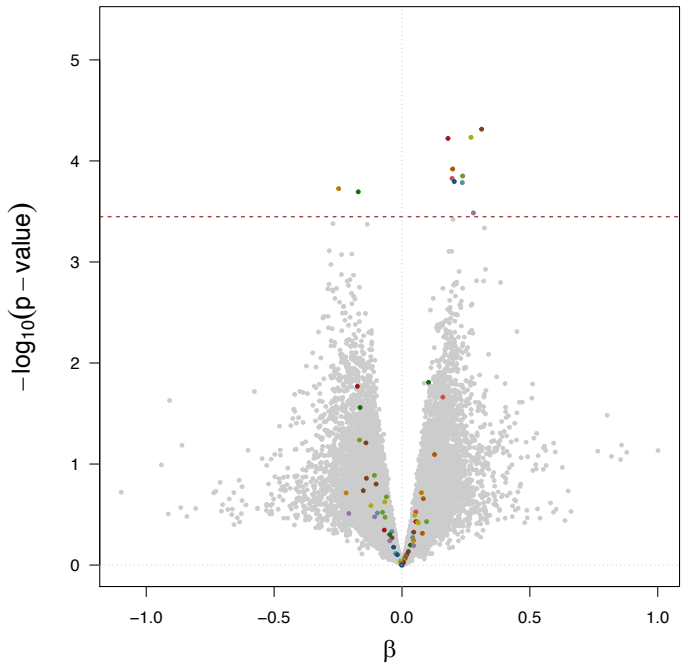
  $\Rightarrow$ 20,974 PC scores summarising the pathways (new dataset)

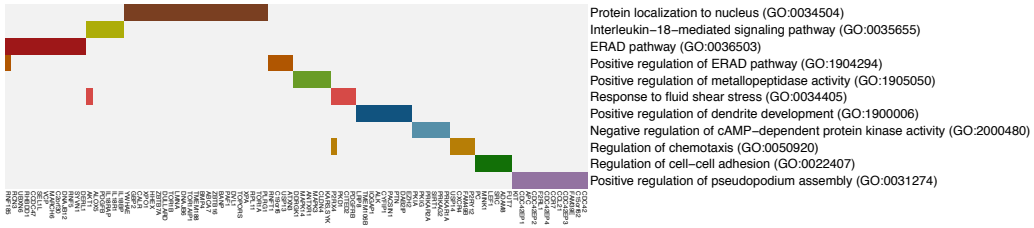
  $\Rightarrow$ Number of PCs per pathway ranging between 1 and 10

- Univariate regressions of CDCP1 against PC scores summarising the groups
- Estimation of the Effective Number of Test with the number of PCs to explain 90% of the variance over the 20,974 scores (ENT=140)

Associations between CDCP1 and Biological Processes

- Univariate regressions of CDCP1 against PC scores summarising Biological Processes





⇒ Identification of 13 groups including protein localization to nucleus, regulation of cell-cell adhesion and regulation of chemotaxis

Conclusions and perspectives

- Analyses of a panel of circulating inflammatory proteins in association with the future risk of lung cancer in two prospective cohorts

  $\Rightarrow$ Identification of robust associations between CDCP1 and all lung cancer and adenocarcinoma

  $\Rightarrow$ Associations hold in stratified analyses by cohort (EPIC/NOWAC) or median time-to-diagnosis

  $\Rightarrow$ Validation of these findings in two independent cohorts

  $\Rightarrow$ Moderate gain in AUC on top of packyears (0.04 for adenocarcinoma)

  $\Rightarrow$ Limited gain when considering joint effects of inflammatory proteins

- CDCP1 (CUB domain containing protein 1): transmembrane noncatalytic receptor involved in the loss of anchorage in epithelial cells during mitosis

  $\Rightarrow$ Previously found associated with higher proliferation and poor prognosis in lung cancer

Conclusions and perspectives

- Using transcriptomics measurements in the same individuals, integration of CDCP1 with transcript levels to gain insight into its functional role
- Functional grouping of the transcripts based on the Reactome and Biological Processes knowledgebases
- Identification of associations between CDCP1 and summarised pathways

  $\Rightarrow$ β-catenin transactivating complex: linked to a range of cancers, implicated in tumour development

  $\Rightarrow$ cell-cell adhesion: CDCP1 disruption associated with interference in EGF/EGFR (Epidermal growth factor receptor) induced cell migration
- Metabolomics measurements in the same participants: to be integrated with proteins and transcripts to explore joint effects at multiple molecular levels

Acknowledgements

- Dr Sonia Dagnino
- Prof Marc Chadeau-Hyam (PI)
- Prof Roel Vermeulen
- Dr Florence Guida
- Dr Karl Smith-Byrne
- Dr Mattias Johansson
- Dr Therese Haugdahl Nost
- Prof Torkjel Sandanger

Funding: Cancer Research - UK 'Mechanomics' PRC project grant (Grant PRC 22184 to MC-H)