Linking the urban exposome to obesity via proteome in the UK Biobank

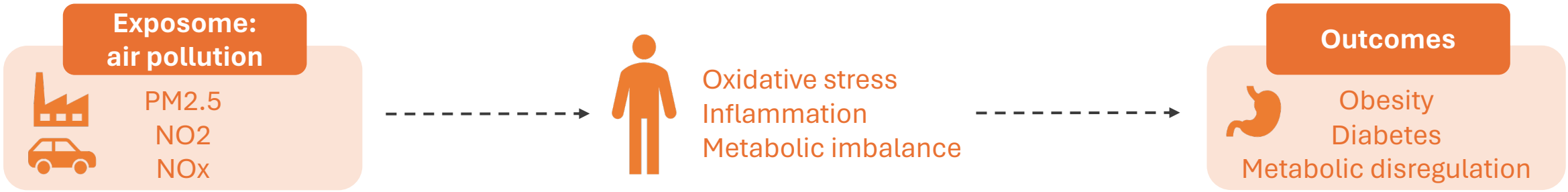
2nd May 2024

Computational Epidemiology

Group 1

Background

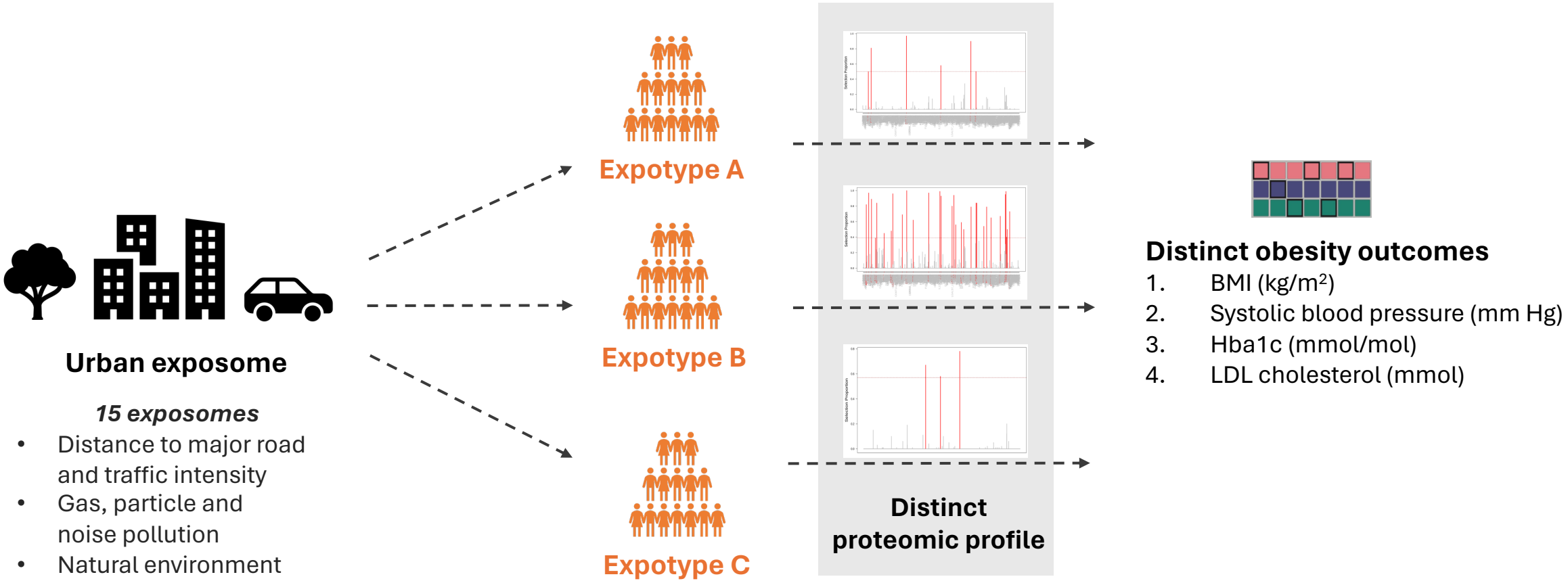
- In 2022, 2.5 billion (43%) adults were overweight and 890 million (16%) adults were obese
- By 2030, more than 80% of Europe's population will live in an urban environment
- Health disparities and air pollution are exacerbated within cities
- Need to define populations at risk → exposome profiling
- The impact of the **urban exposome** on obesity has emerged in recent epidemiological literature
- Notably, obesity has been linked to air pollution via potential mechanisms including inflammation, oxidative stress, metabolic imbalance



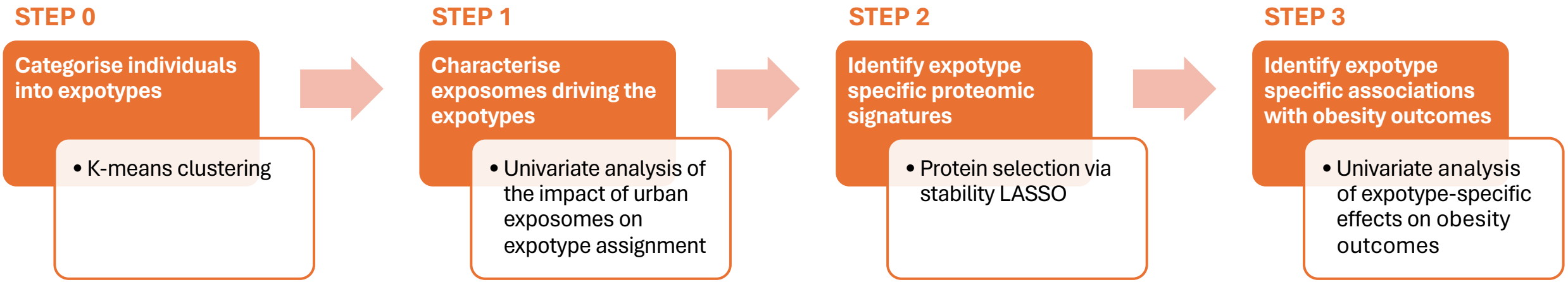
To adequately address health disparities in obesity, we require a system-level understanding of the urban exposome, how it is biologically mediated, and the risk of obesity it confers

From urban expotypes to obesity via proteome

To accommodate the complexity of the urban exposome we aim to identify **urban expotypes**. Further, we aim to characterise these expotypes, investigate their biological embodiment via the proteome, and the subsequent effect on obesity in the **UK Biobank**.



Statistical analysis workflow



Data processing

- **Urban exposome data :**
  - 2010 data with instance 0
  - *Average 24-hour sound level of noise pollution* computed as an average of day, evening and night measurements
  - *Exclusion*: individuals with more than 7 missing covariates

- **Proteome data :**
  - *Exclusion*: individuals with over 50% missing protein data
  - Imputation:
    - Partially missing values – **k-NN**
    - Measurements below the limit of detection – **quantile regression**

Final dataset (steps 0,1 and 2) : **51,879** individuals, **1,343** proteins and **15** urban exposomes

---

- **Obesity data :**
  - Covariates used as proxy for obesity : BMI, hba 1c, LDL cholesterol and systolic blood pressure
  - Additional exclusion of individuals with missing obesity covariates

Final dataset (step 3) : **43,572** individuals and **4** obesity outcomes

Results

STEP 0

Expotype identification using k-means clustering

- Use all 15 exposomes to do clustering
- Calibrate the number of clusters (*k*) from 2 to 20
- Best **k =4**

Cluster	Number of individuals
Expotype 0	3386
Expotype 1	27878
Expotype 2	17437
Expotype 3	3178

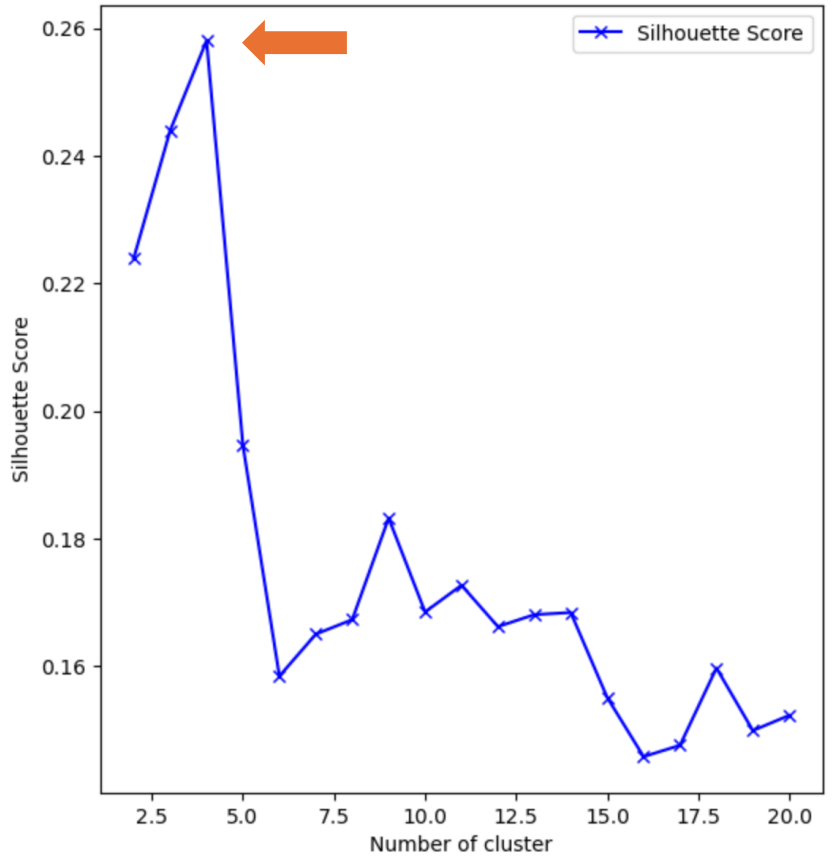


Table 1 – Participant distribution

Variables		Expotype 0	Expotype 1	Expotype 2	Expotype 3
Gender	Male	1604(47.4%)	12770(45.8%)	8015(46%)	1502(47.3%)
	Female	1782(52.6%)	15108(54.2%)	9422(54%)	1676(52.7%)
Age		59.1 (8.11)	58.4 (8.35)	59.4 (7.98)	58 (8.41)
Obesity variables	BMI	27.7 (4.73)	27.5 (4.88)	27.3 (4.63)	27.8 (5.1)
	HBA1C	36.5 (8.58)	36.4 (7.16)	36.1 (6.5)	36.4 (7.19)
	LDL	3.5 (0.87)	3.51 (0.88)	3.58 (0.88)	3.48 (0.87)
	BP	140 (19.1)	139 (19.6)	141 (19.8)	139 (20.1)

The value in each cell is counts(percentage) or mean(standard deviation)

STEP 1

Characterisation of exposomes driving expotypes

- **For each exposome and each expotype**
  - Outcome: 1 vs Rest expotype (e.g. Expotype 0 vs Not Expotype 0)
  - **Logistic regression : Logit(p) = Exposome + €**
  - Using Bonferroni (FWER) multiple tests correction, adjusted p-value threshold for significance is 0.0033 (0.05/15)

STEP 1

Characterisation of exposomes driving expotypes

exposome	OR			
	Expotype_0	Expotype_1	Expotype_2	Expotype_3
Particulate matter air pollution PM2.5 Absorbance	4.75	5.53	0.00	109.59
Particulate matter air pollution PM10	3.67	1.05	0.30	1.92
Particulate matter air pollution PM2.5		2.68	0.06	3.54
Water percentage buffer 1000 m	1.04	0.89	1.08	1.04
Greenspace percentage buffer 1000m	1.01	0.90	1.14	0.97
Natural environment percentage buffer 1000m	1.01	0.92	1.12	0.98
Traffic intensity on the nearest major road	1.00	1.00	1.00	1.00
Traffic intensity on the nearest road	1.00	1.00	1.00	1.00
Distance to coast	0.99	1.00	1.01	
Average 24-hour sound level of noise pollution		0.91	0.92	1.65
Nitrogen oxides air pollution	0.98	1.05	0.81	1.18
Domestic garden percentage buffer 1000m	0.98	1.16	0.86	0.98
Nitrogen dioxide air pollution	0.97	1.17	0.59	1.31
Inverse distance to the nearest road		1.69	0.39	2.67
Inverse distance to the nearest major road	0.00	0.00	0.00	>10000

Non-significant exposomes are not shown

- Majority of exposomes are significant for all expotypes
- Different expotypes have different characteristics

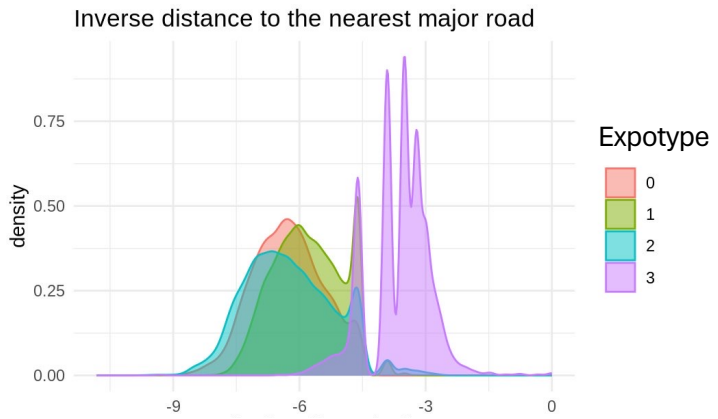
Positive OR
(Conferring risk)

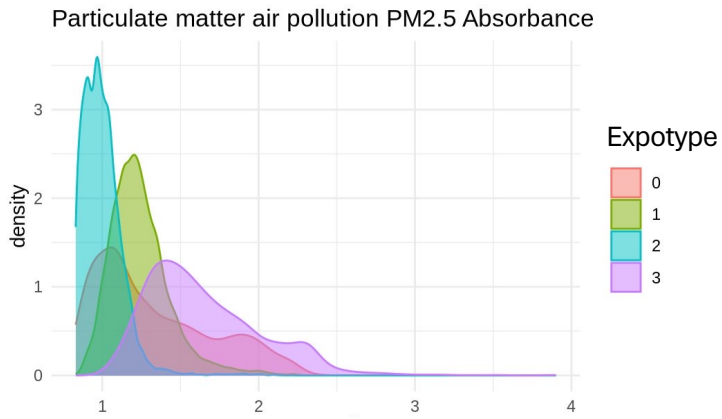
Negative OR
(Protective)

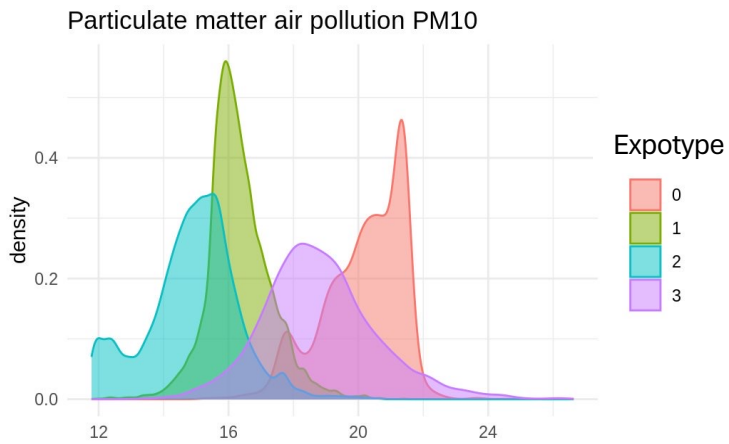
STEP 1

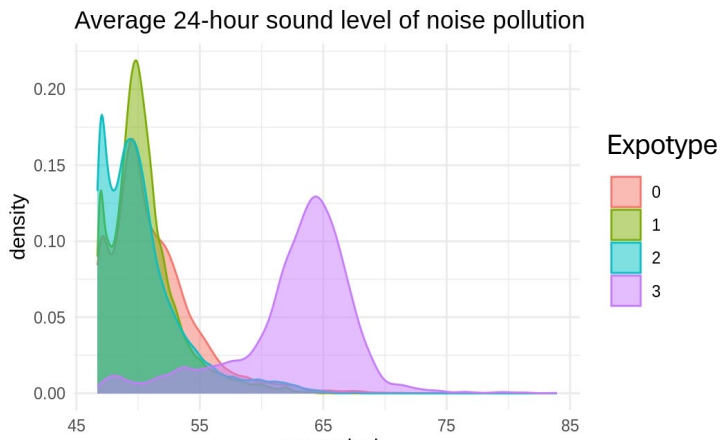
Distribution of exposomes driving expotypes

Important exposures i.e. higher absolute beta





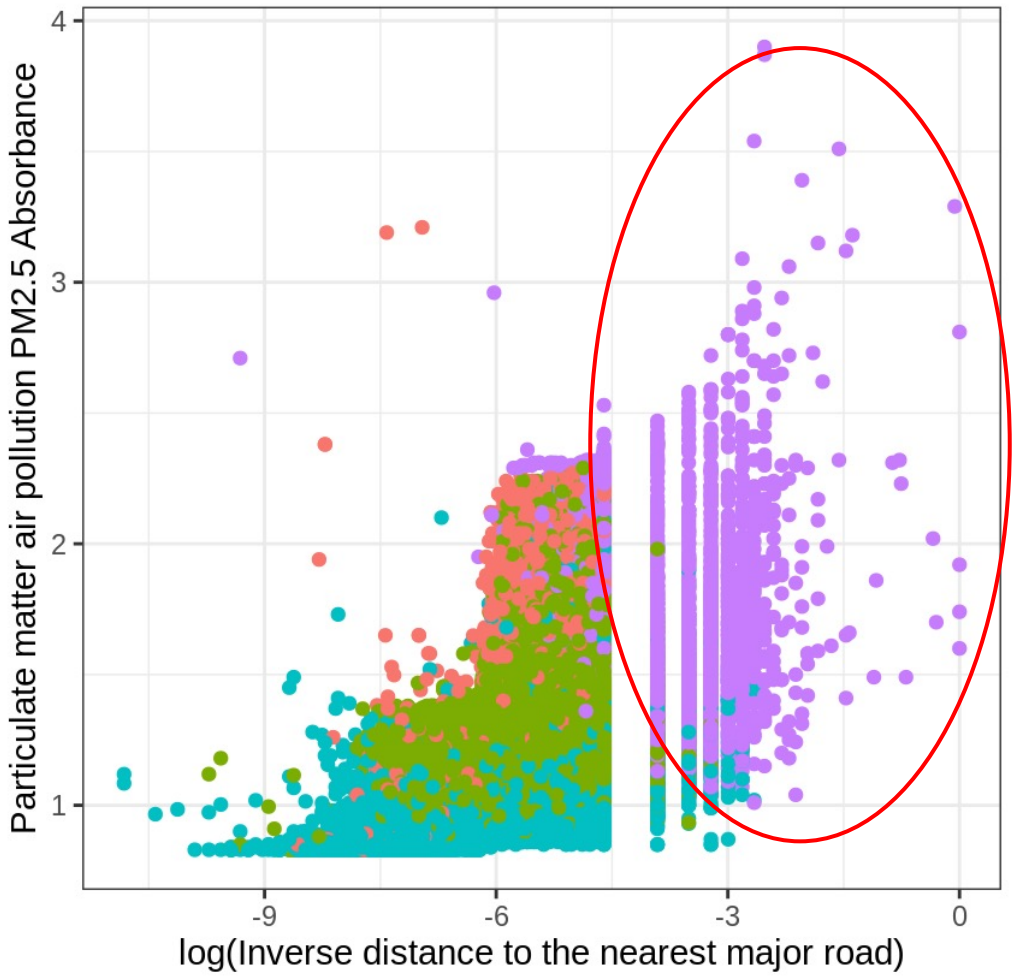




STEP 1

Distribution of exposomes driving expotypes

Important exposures



Combinations of exposomes in higher dimensions (15 dims) are capturing the difference between expotypes

STEP 1

Overall Expotype characteristics

Important exposures

Expotype 0	Expotype 1	Expotype 2	Expotype 3
Polluted environment	**Less natural and polluted environment**	**Healthier environment**	**Noisy and high polluted environment**
• Higher in both PM10 and PM2.5 absorbance • Much far away from the major road	• Higher in PM2.5 and PM2.5 absorbance • Slightly close to the road • Much far away from the major road • Slightly lower percentage in water, green space and natural environment	• Much lower in all PM2.5 and PM10 • Much far away from the major road • Slightly higher percentage in water, green space and natural environment	• Much higher PM2.5 absorbance • Much close from the nearest major road • Higher exposure to noise pollution

STEP 2

Identify expotype specific proteomic signatures

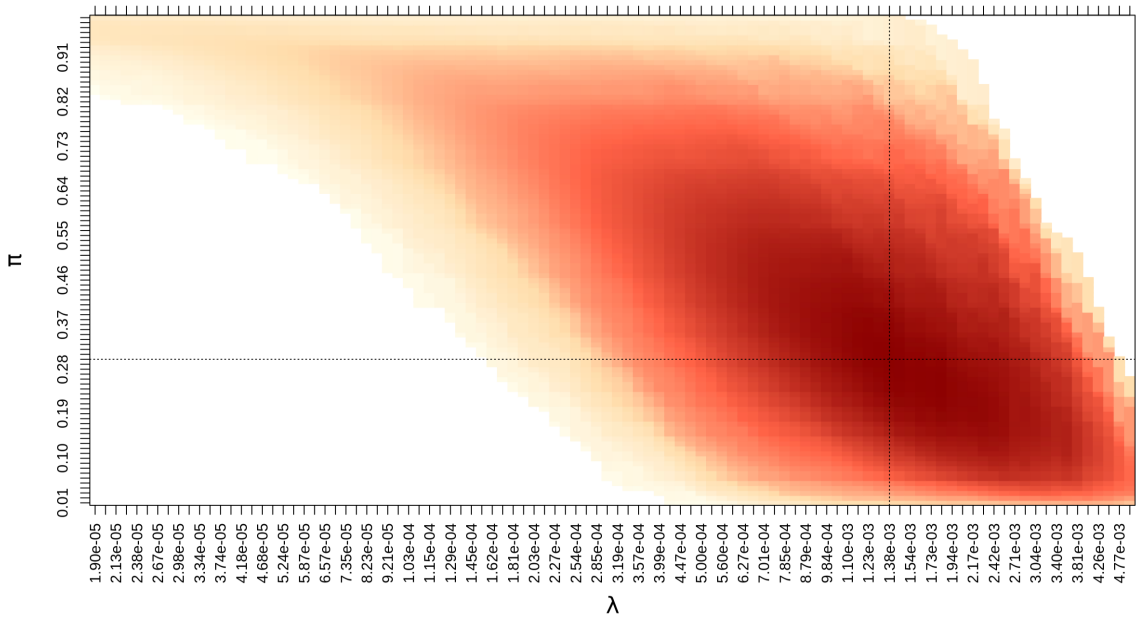
- Methodology for each expotype :
  - **A - Protein selection via stability selection**
  - **B - Fitting a logistic regression with stably selected proteins :** Logit(p) = Selected proteins + ϵ

- Sex and age as confounders

STEP 2 - A

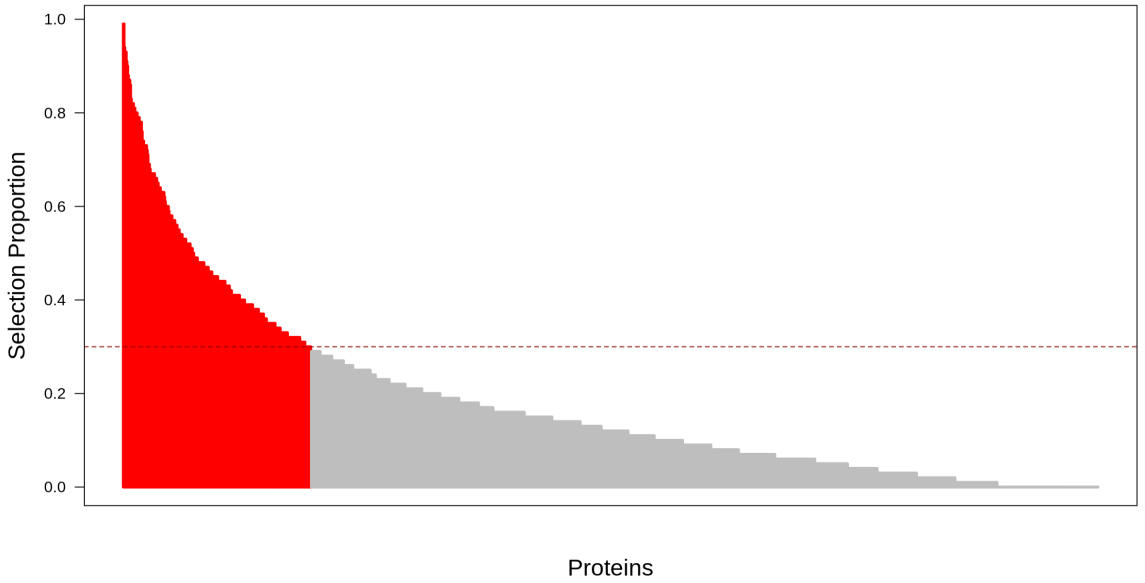
Protein selection via stability LASSO (Expotype 0)

Step 1: Calibration of the penalty parameter (λ) and selection proportion threshold (π)



Calibrated penalty parameter (λ) : 0.00138
Calibrated proportion selection threshold (π): 0.3

Step 2: Selection of proteins with selection proportion above π over the 100 models fitted with calibrated λ



254 selected proteins (red)

STEP 2 - A

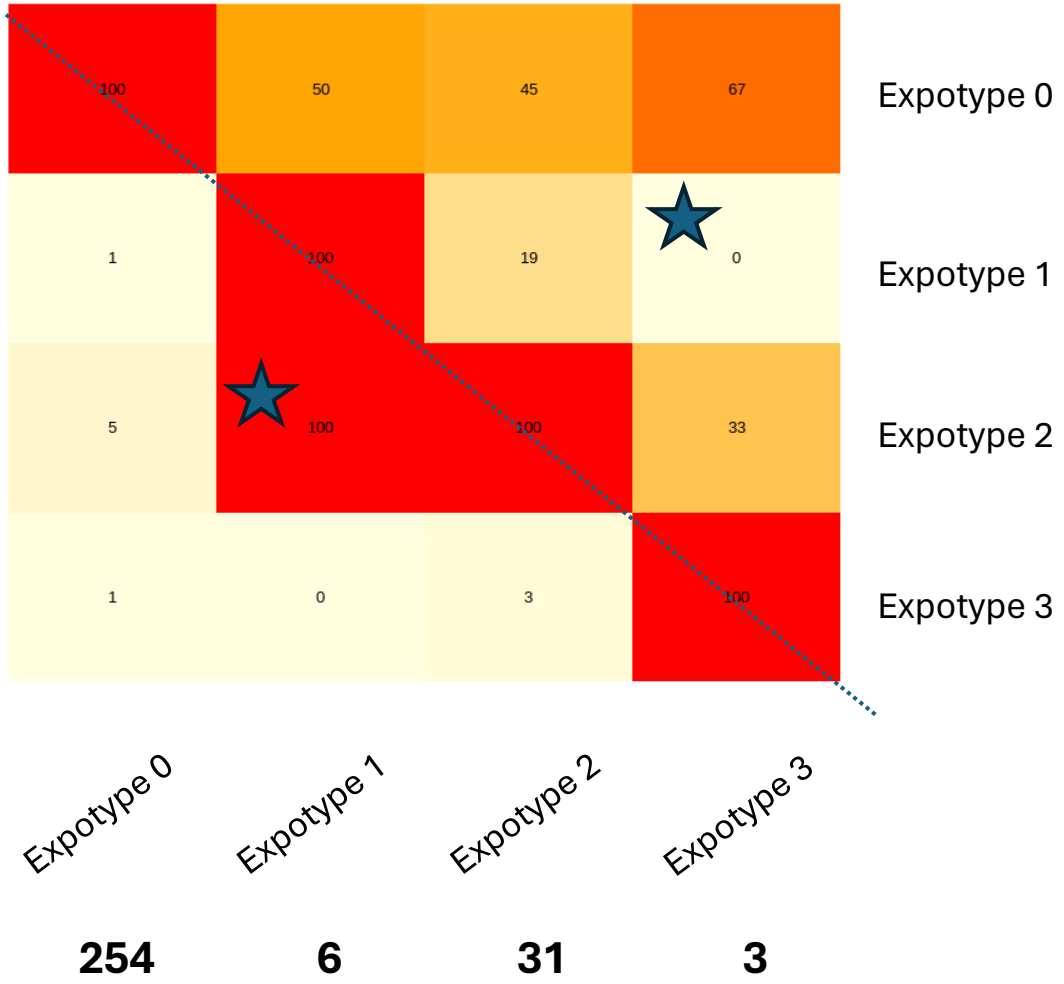
Results - Protein selection via stability LASSO

Model	Penalty parameter (λ)	Selection proportion threshold (π)	Number of stably selected proteins	Most stably selected proteins
Expotype 0 vs Rest	0.00138	**0.3**	**254**	**CD99L2** (0.99) **CHMP1A** (0.99) **IL22RA1** (0.94)
Expotype 1 vs Rest	**0.01737**	0.5	**6**	**CXCL14** (0.97) **OMG** (0.90) **APEX1** (0.81)
Expotype 2 vs Rest	0.01193	0.39	**31**	**CXCL14** (1.00) **IL15** (0.99) **TFF1** (0.99)
Expotype 3 vs Rest	0.00625	0.57	**3**	**MEPE** (0.78) **FRZB** (0.67) **IL15** (0.58)

STEP 2 - A

Percentage of commonly selected proteins

- **Expotype 1 shares all its selected proteins with Expotype 2** and 50% of its selected proteins with Expotype 0
- **Expotype 3 shares no protein with Expotype 1**
- Expotype 2 shares 45% of its selected proteins with Expotype 0
- Expotype 3 shares 67% of its selected proteins with Expotype 0

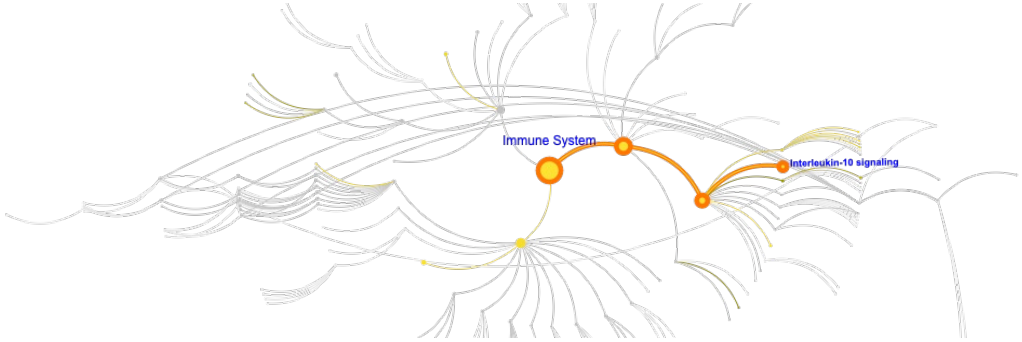


STEP 2 - A

Pathway analysis using *reactome* – Expotypes 0 and 2

Expotype 0 – **Interleukin-10 signaling**

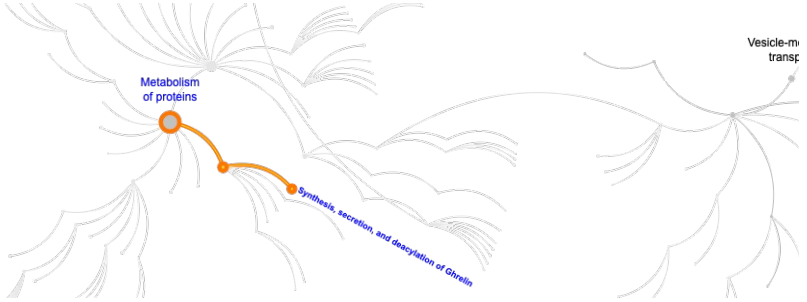
IL10 plays a critical role in limiting inflammatory responses. Dysregulation of IL10 is linked with susceptibility to numerous infectious and autoimmune diseases.



Pathway name	Entities found	Entities Total	Entities ratio	Entities pValue	Entities FDR	Reactions found
Interleukin-10 signaling	19	86	0.006	3.31E-14	2.44E-11	*12*
Immune System	92	2,663	0.171	5.58E-8	1.81E-5	*319*
Signaling by Interleukins	36	658	0.042	7.41E-8	1.81E-5	*121*
Cytokine Signaling in Immune system	44	1,099	0.071	1.12E-5	2.06E-3	*151*

Expotype 2 – **Synthesis, secretion and deacylation of Ghrelin**

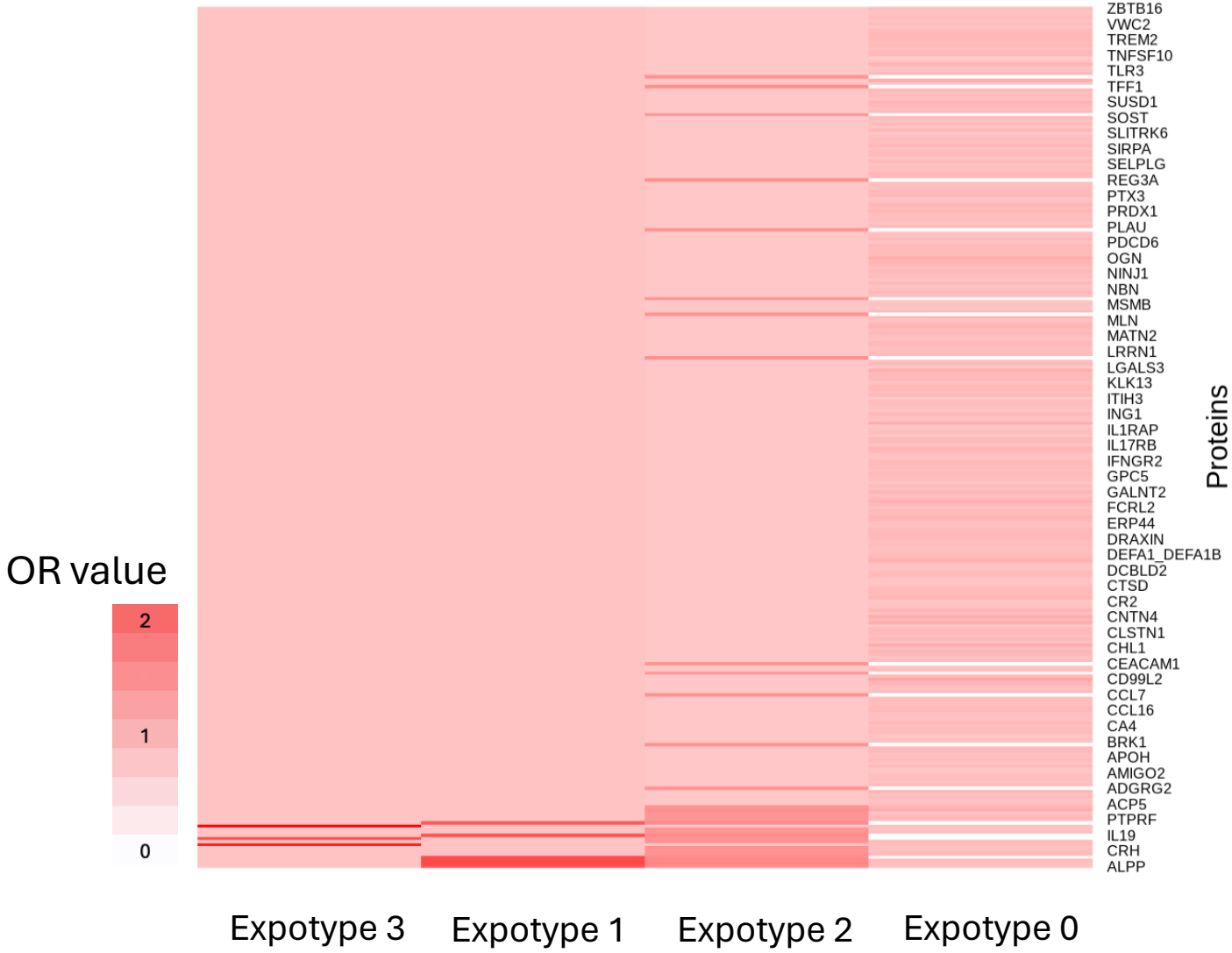
Ghrelin levels increase between mealtimes and decrease when your stomach is full. People who have obesity often have low ghrelin levels.



Pathway name	Entities found	Entities Total	Entities ratio	Entities pValue	Entities FDR	Reactions found
Synthesis, secretion, and deacylation of Ghrelin	4	26	0.002	7.49E-7	1.09E-4	*8*
Regulation of MITF-M dependent genes involved in invasion	2	7	0	1.64E-4	1.18E-2	*2*
Peptide hormone metabolism	4	129	0.008	3.73E-4	1.79E-2	*8*
MECP2 regulates transcription of neuronal ligands	2	13	0.001	5.6E-4	1.88E-2	*2*

STEP 2 - B

Logistic regression on selected proteins



STEP 2 - B

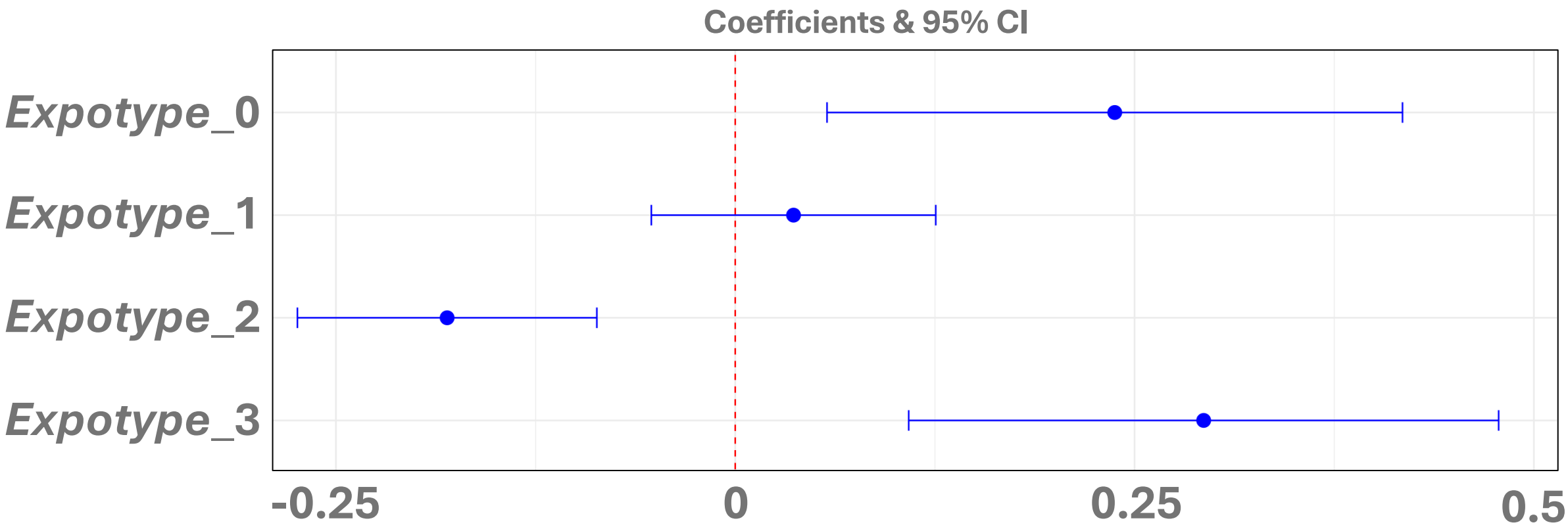
Logistic regression on selected proteins

Model	# selected proteins	Highest effect proteins		Beta value
Expotype 0 vs Rest	**254**	Top 3 positively associated	CD99L2 ITGAV IDS	1.694229 1.417556 1.284426
		Top 3 negatively associated	CNTN4 CLEC14A COLEC12	0.80435150 0.79503964 0.73901779
Expotype 1 vs Rest	**6**	Top 3 positively associated	IL19 APEX1 PKLR	2.1178038 1.0900836 1.0855737
		Top 3 negatively associated	OMG CXCL14	0.9900674 0.8830428
Expotype 2 vs Rest	**31**	Top 3 positively associated	**CEACAM1** PTPRF CXCL14	**1.252669** 1.209309 1.160688
		Top 3 negatively associated	LPO SPINK6 IL15	0.9311492 0.9214034 0.8619870
Expotype 3 vs Rest	**3**	Top 3 positively associated	IL15	1.3251678
		Top 3 negatively associated	FRZB MEPE	0.7353870 0.6850272

STEP 3

Expotype cluster association with obesity | BMI

Linear model = BMI ~ Expotype_x + Age + Sex + Household income



STEP 3

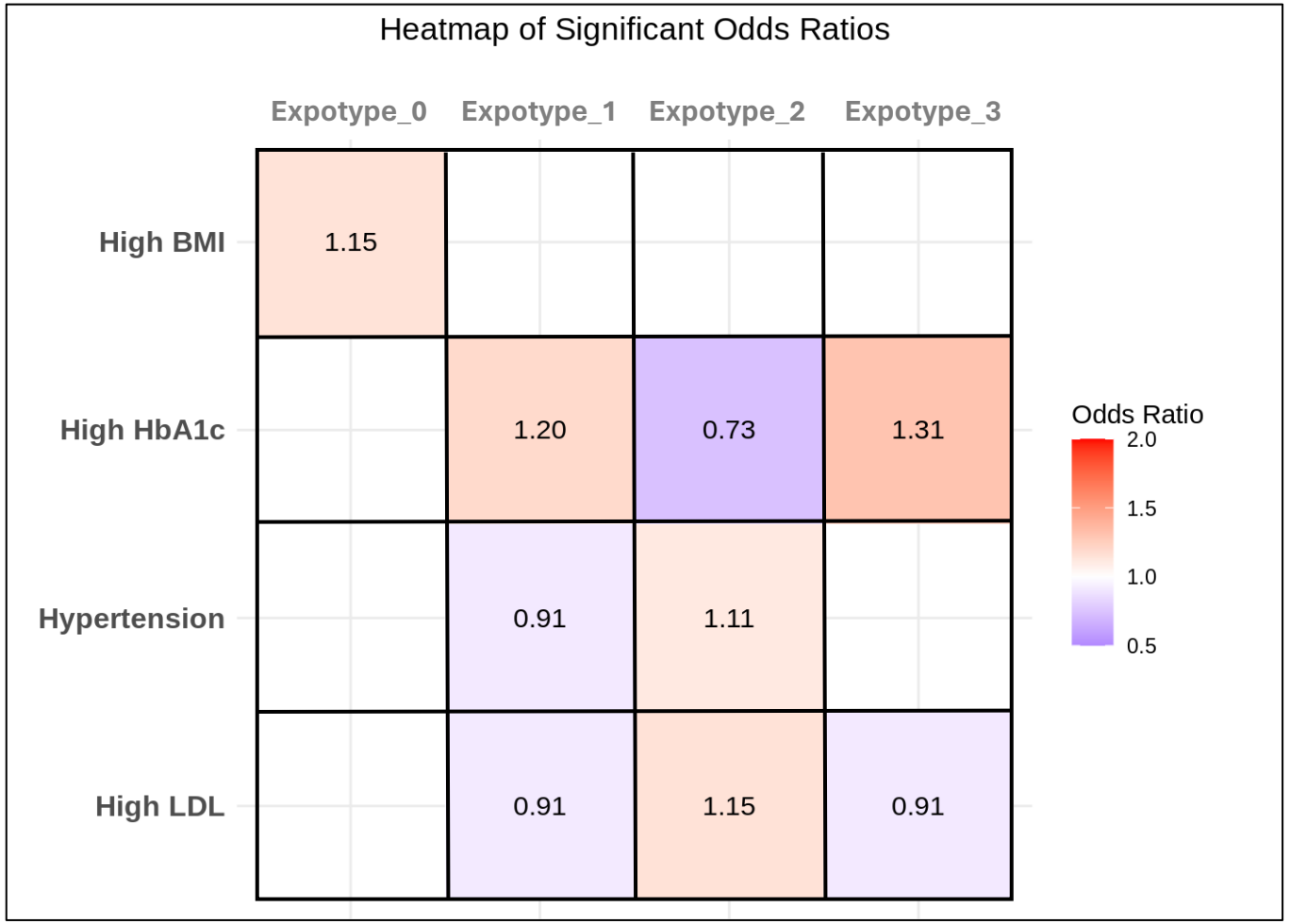
Logistic model = Outcome ~ Expotype_x + Age + Sex + Household income

Obese: BMI > 25

Diabetic: HbA1c > 42 mmol/mol

Hypertensive: sysBP > 140mmHG

Hyperlipidemic: LDL > 3 mmol



Results summary & conclusion

	Expotype 0	Expotype 1	Expotype 2	Expotype 3
Pollution & traffic				
Candidate pathways	**IL-10 Signalling** (inflammation regulation)	**IL-19 Signalling** (adipocyte homeostasis)	Ghrelin (hunger hormone) **IL-15 Signalling** (inflammation regulation)	**IL-15 Signalling** (inflammation regulation)
Obesity & diabetes				

Limitations and future work

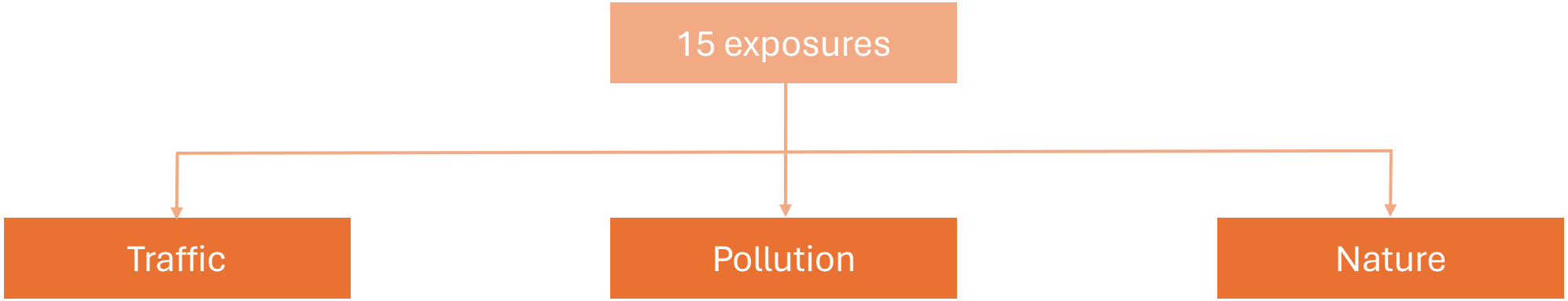
- Average high age of UKBB and focusing on smaller set with protein data
- Small percentage of missing data (< 7%) in obesity analysis > possible risk of bias if missing data is not MCAR.
- Relevance of smoking as confounder
- Imbalance in selected proteins
- Future direction: improving clustering (e.g. bigger and more inclusive study population & pollution relevant predictors), optimising proteome profiles, and detailed functional and pathway analysis

Select references

- Congdon, P. Obesity and Urban Environments. *Int J Environ Res Public Health*. 2019 Feb; 16(3): 464. doi: 10.3390/ijerph16030464.
- World Health Organisation. Obesity and Overweight. 2024 Mar. Available from: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight.
- Shi et al. Exposure to Outdoor and Indoor Air Pollution and Risk of Overweight and Obesity Across Different Life Periods: A Review. *Ecotoxicology and Environmental Safety.* 2022 Sep; 242. doi: 10.1016/j.ecoenv.2022.113893.
- Deng et al. The Effect of Urbanization on Air Pollution Damage. Journal of the Association of Environmental and Resource Economists. 2021; 8(5).

Appendix

Urban Exposome in the UK Biobank



1. Inverse distance to the nearest major road
2. Inverse distance to the nearest road
3. Traffic intensity on the nearest major road
4. Traffic intensity on the nearest road

1. Nitrogen dioxide air pollution
2. Nitrogen oxides air pollution
3. Particulate matter air pollution PM2.5
4. Particulate matter air pollution PM2.5 absorbance
5. Particulate matter air pollution PM10
6. Average 24-hour sound level of noise pollution

1. Natural environment percentage buffer 1000m
2. Greenspace percentage buffer 1000m
3. Domestic garden percentage buffer 1000m
4. Water percentage buffer 1000 m
5. Distance to coast

STEP 1

Characterisation of exposures driving expotypes

Cluster 0

exposome	OR	95% CI		p value	if significant
Particulate matter air pollution PM2.5 Absorbance	4.8E+00	4.3E+00	5.3E+00	1.5E-193	TRUE
Particulate matter air pollution PM10	3.7E+00	3.6E+00	3.8E+00	0.0E+00	TRUE
Particulate matter air pollution PM2.5	1.0E+00	1.0E+00	1.1E+00	4.3E-03	FALSE
Water percentage buffer 1000 m	1.0E+00	1.0E+00	1.0E+00	1.9E-09	TRUE
Greenspace percentage buffer 1000m	1.0E+00	1.0E+00	1.0E+00	2.2E-49	TRUE
Natural environment percentage buffer 1000m	1.0E+00	1.0E+00	1.0E+00	1.2E-39	TRUE
Traffic intensity on the nearest major road	1.0E+00	1.0E+00	1.0E+00	0.0E+00	TRUE
Traffic intensity on the nearest road	1.0E+00	1.0E+00	1.0E+00	2.6E-12	TRUE
Distance to coast	9.9E-01	9.9E-01	1.0E+00	7.7E-20	TRUE
Average 24-hour sound level of noise pollution	9.9E-01	9.9E-01	1.0E+00	1.4E-01	FALSE
Nitrogen oxides air pollution	9.8E-01	9.8E-01	9.9E-01	1.1E-36	TRUE
Domestic garden percentage buffer 1000m	9.8E-01	9.8E-01	9.8E-01	2.1E-37	TRUE
Nitrogen dioxide air pollution	9.7E-01	9.6E-01	9.7E-01	2.0E-36	TRUE
Inverse distance to the nearest road	7.1E-01	4.2E-01	1.2E+00	2.0E-01	FALSE
Inverse distance to the nearest major road	3.3E-61	4.2E-67	1.7E-55	1.2E-92	TRUE

Green shade is OR above 1 and grey shade is OR below 1. Red font represents insignificant results with the Bonferroni correction 0.0033

STEP 1

Characterisation of exposures driving expotypes

Cluster 1

exposome	OR	95% CI		p value	if significant
Particulate matter air pollution PM2.5 Absorbance	5.5E+00	5.1E+00	6.0E+00	0.0E+00	TRUE
Particulate matter air pollution PM2.5	2.7E+00	2.6E+00	2.7E+00	0.0E+00	TRUE
Inverse distance to the nearest road	1.7E+00	1.3E+00	2.2E+00	1.8E-05	TRUE
Nitrogen dioxide air pollution	1.2E+00	1.2E+00	1.2E+00	0.0E+00	TRUE
Domestic garden percentage buffer 1000m	1.2E+00	1.2E+00	1.2E+00	0.0E+00	TRUE
Particulate matter air pollution PM10	1.1E+00	1.0E+00	1.1E+00	2.8E-25	TRUE
Nitrogen oxides air pollution	1.0E+00	1.0E+00	1.0E+00	0.0E+00	TRUE
Traffic intensity on the nearest major road	1.0E+00	1.0E+00	1.0E+00	8.2E-262	TRUE
Traffic intensity on the nearest road	1.0E+00	1.0E+00	1.0E+00	1.7E-269	TRUE
Distance to coast	1.0E+00	1.0E+00	1.0E+00	2.4E-23	TRUE
Natural environment percentage buffer 1000m	9.2E-01	9.1E-01	9.2E-01	0.0E+00	TRUE
Average 24-hour sound level of noise pollution	9.1E-01	9.0E-01	9.1E-01	0.0E+00	TRUE
Greenspace percentage buffer 1000m	9.0E-01	9.0E-01	9.0E-01	0.0E+00	TRUE
Water percentage buffer 1000 m	8.9E-01	8.8E-01	9.0E-01	9.3E-121	TRUE
Inverse distance to the nearest major road	1.5E-19	1.1E-20	1.9E-18	1.2E-240	TRUE

Green shade is OR above 1 and grey shade is OR below 1. Red font represents insignificant results with the Bonferroni correction 0.0033

STEP 1

Characterisation of exposures driving expotypes

Cluster 2

exposome	OR	95% CI		p value	if significant
Greenspace percentage buffer 1000m	1.1E+00	1.1E+00	1.1E+00	0.0E+00	TRUE
Natural environment percentage buffer 1000m	1.1E+00	1.1E+00	1.1E+00	0.0E+00	TRUE
Water percentage buffer 1000 m	1.1E+00	1.1E+00	1.1E+00	3.1E-74	TRUE
Distance to coast	1.0E+00	1.0E+00	1.0E+00	2.1E-60	TRUE
Traffic intensity on the nearest road	1.0E+00	1.0E+00	1.0E+00	7.4E-61	TRUE
Traffic intensity on the nearest major road	1.0E+00	1.0E+00	1.0E+00	0.0E+00	TRUE
Average 24-hour sound level of noise pollution	9.2E-01	9.2E-01	9.3E-01	3.1E-205	TRUE
Domestic garden percentage buffer 1000m	8.6E-01	8.6E-01	8.6E-01	0.0E+00	TRUE
Nitrogen oxides air pollution	8.1E-01	8.0E-01	8.1E-01	0.0E+00	TRUE
Nitrogen dioxide air pollution	5.9E-01	5.8E-01	6.0E-01	0.0E+00	TRUE
Inverse distance to the nearest road	3.9E-01	2.9E-01	5.2E-01	8.1E-11	TRUE
Particulate matter air pollution PM10	3.0E-01	2.9E-01	3.0E-01	0.0E+00	TRUE
Particulate matter air pollution PM2.5	5.6E-02	5.3E-02	5.9E-02	0.0E+00	TRUE
Particulate matter air pollution PM2.5 Absorbance	2.1E-05	1.7E-05	2.5E-05	0.0E+00	TRUE
Inverse distance to the nearest major road	1.0E-29	1.7E-31	5.3E-28	1.1E-231	TRUE

Green shade is OR above 1 and grey shade is OR below 1. Red font represents insignificant results with the Bonferroni correction 0.0033

STEP 1

Characterisation of exposures driving expotypes

Cluster 3

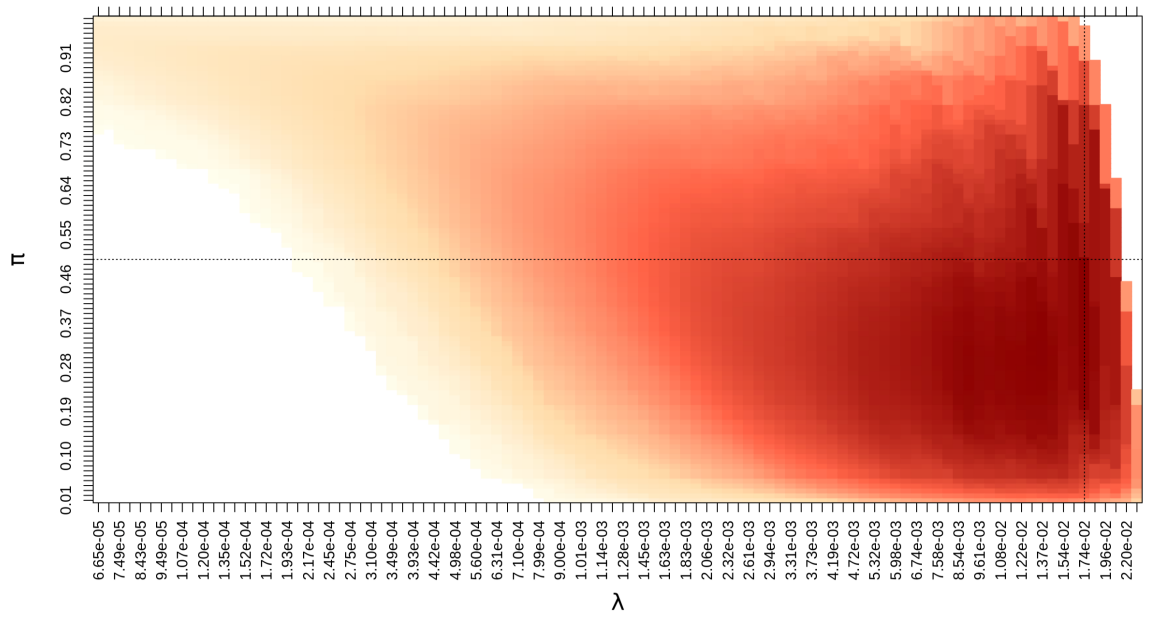
exposome	OR	95% CI		p value	if significant
Inverse distance to the nearest major road	2.5E+101	4.8E+98	1.5E+104	0.0E+00	TRUE
Particulate matter air pollution PM2.5 Absorbance	1.1E+02	9.7E+01	1.2E+02	0.0E+00	TRUE
Particulate matter air pollution PM2.5	3.5E+00	3.4E+00	3.7E+00	0.0E+00	TRUE
Inverse distance to the nearest road	2.7E+00	1.8E+00	3.8E+00	1.4E-07	TRUE
Particulate matter air pollution PM10	1.9E+00	1.9E+00	2.0E+00	0.0E+00	TRUE
Average 24-hour sound level of noise pollution	1.6E+00	1.6E+00	1.7E+00	0.0E+00	TRUE
Nitrogen dioxide air pollution	1.3E+00	1.3E+00	1.3E+00	0.0E+00	TRUE
Nitrogen oxides air pollution	1.2E+00	1.2E+00	1.2E+00	0.0E+00	TRUE
Water percentage buffer 1000 m	1.0E+00	1.0E+00	1.1E+00	3.3E-10	TRUE
Traffic intensity on the nearest road	1.0E+00	1.0E+00	1.0E+00	0.0E+00	TRUE
Traffic intensity on the nearest major road	1.0E+00	1.0E+00	1.0E+00	1.3E-05	TRUE
Distance to coast	1.0E+00	1.0E+00	1.0E+00	2.6E-02	FALSE
Domestic garden percentage buffer 1000m	9.8E-01	9.8E-01	9.8E-01	4.0E-36	TRUE
Natural environment percentage buffer 1000m	9.8E-01	9.8E-01	9.8E-01	8.7E-151	TRUE
Greenspace percentage buffer 1000m	9.7E-01	9.7E-01	9.8E-01	8.7E-141	TRUE

Green shade is OR above 1 and grey shade is OR below 1. Red font represents insignificant results with the Bonferroni correction 0.0033

STEP 2

Protein selection via stability LASSO (Expotype 1)

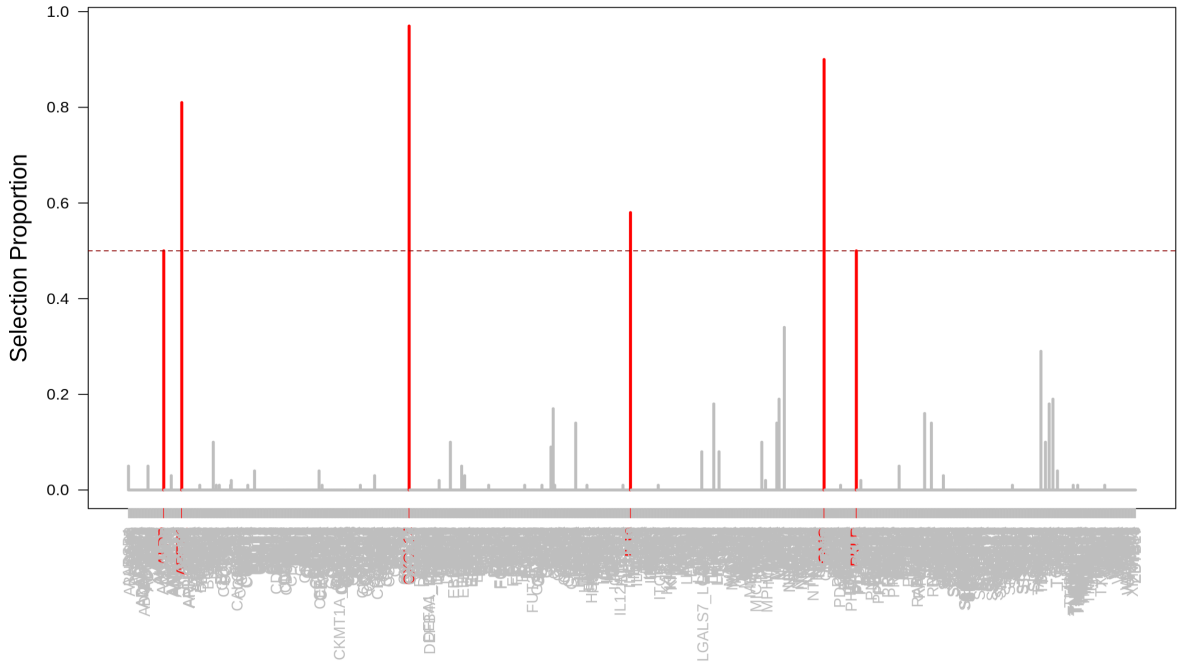
Step 1: Calibration of the penalty parameter (λ) and selection proportion threshold (π)



Calibrated penalty parameter (λ): 0.01737038

Calibrated proportion selection threshold (π): 0.5

Step 2: Selection of proteins with selection proportion above π over the 100 models fitted with calibrated λ

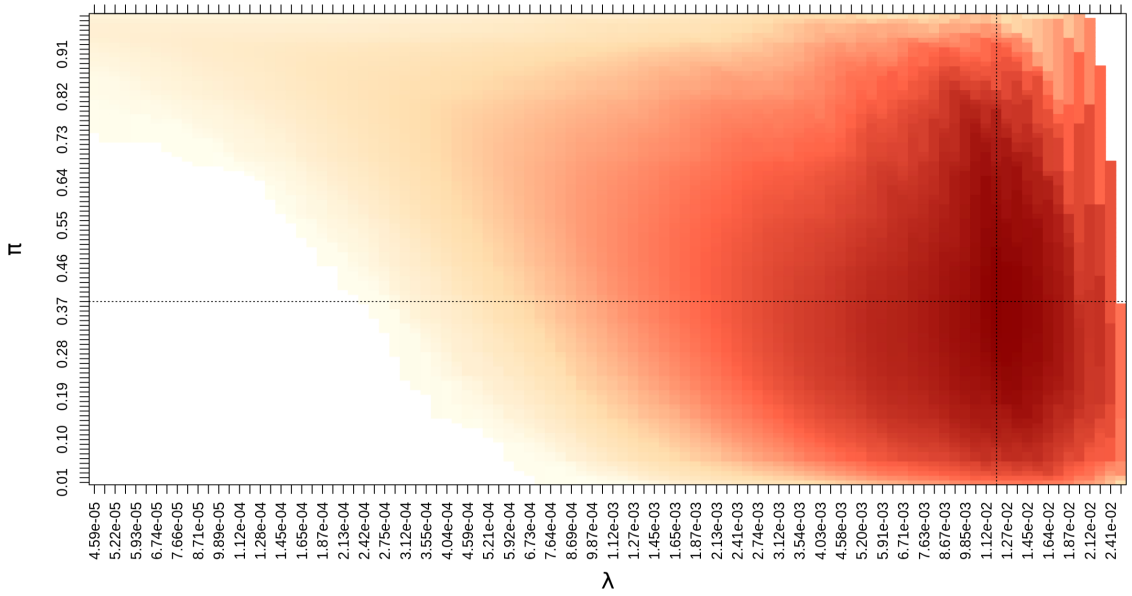


6 selected proteins (red)

STEP 2

Protein selection via stability LASSO (Expotype 2)

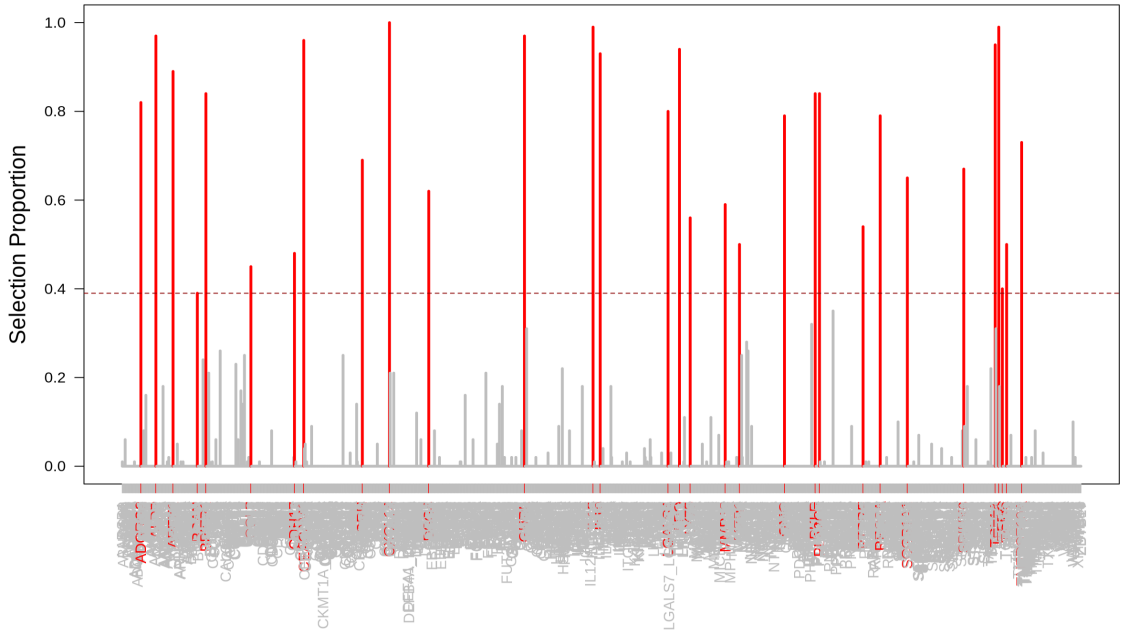
Step 1: Calibration of the penalty parameter (λ) and selection proportion threshold (π)



Calibrated penalty parameter (λ): 0.01193215

Calibrated proportion selection threshold (π): 0.39

Step 2: Selection of proteins with selection proportion above π over the 100 models fitted with calibrated λ

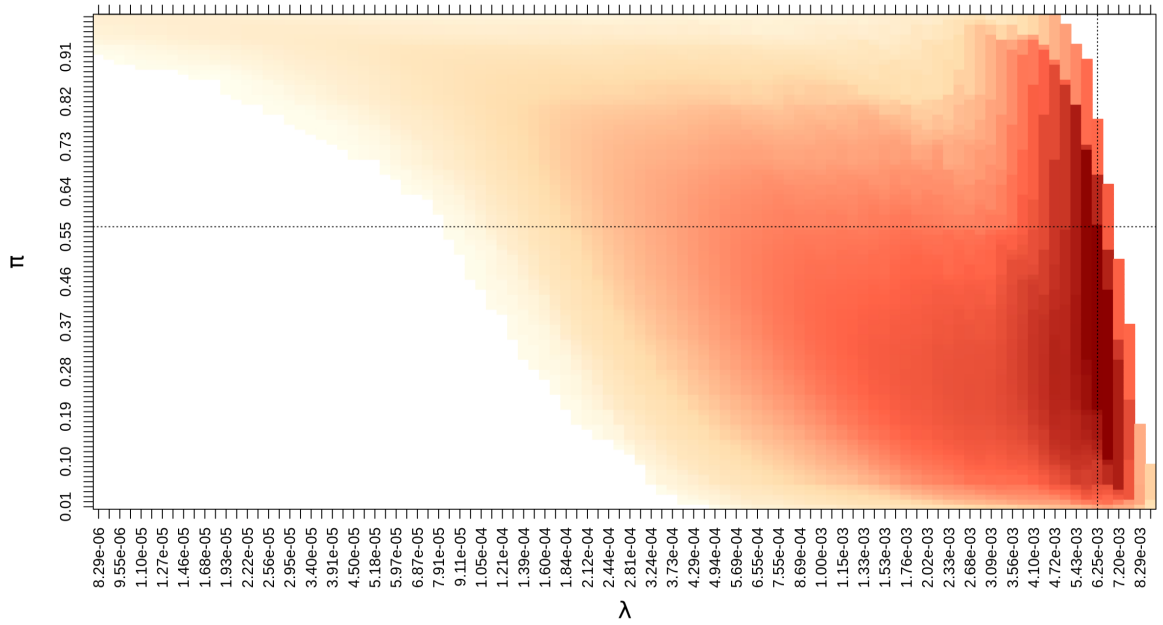


31 selected proteins (red)

STEP 2

Protein selection via stability LASSO (Expotype 3)

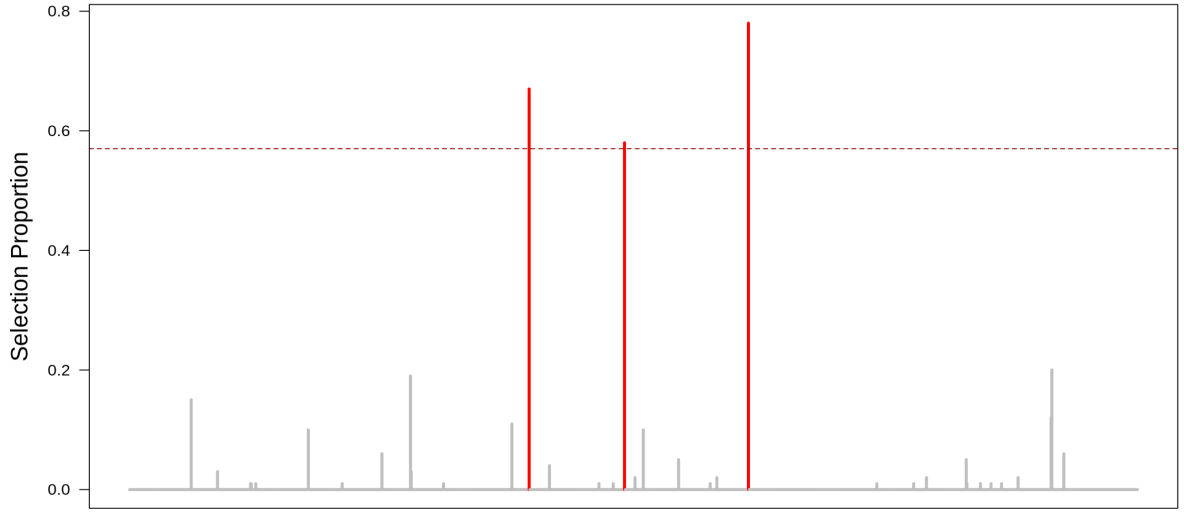
Step 1: Calibration of the penalty parameter (λ) and selection proportion threshold (π)



Calibrated penalty parameter (λ): 0.006252482

Calibrated proportion selection threshold (π): 0.57

Step 2: Selection of proteins with selection proportion above π over the 100 models fitted with calibrated λ

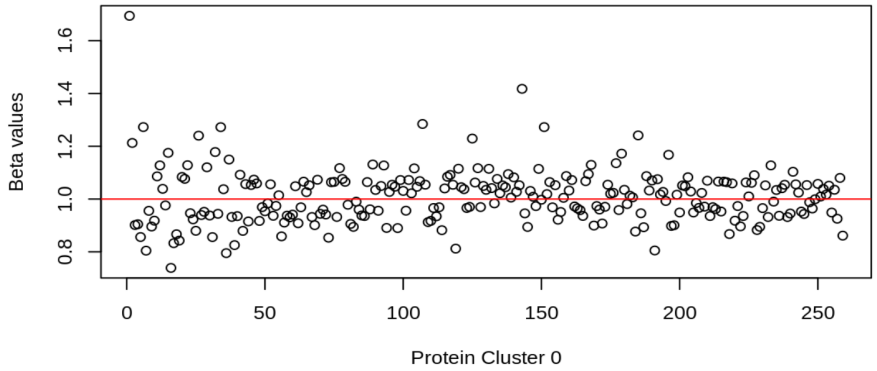


3 selected proteins (red)

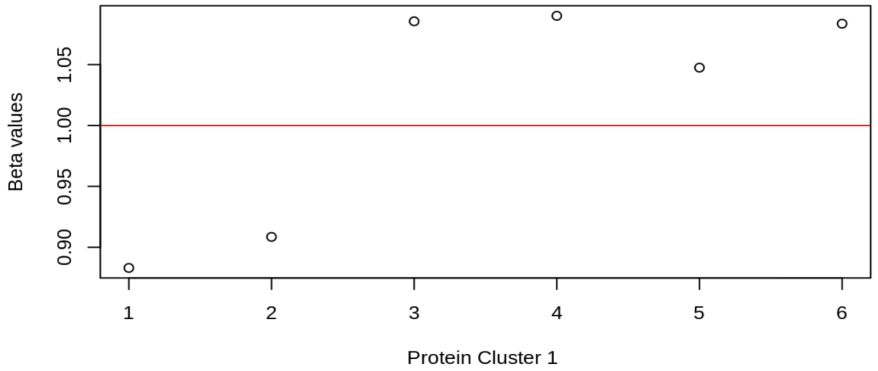
STEP 2 - B

Logistic regression on selected proteins

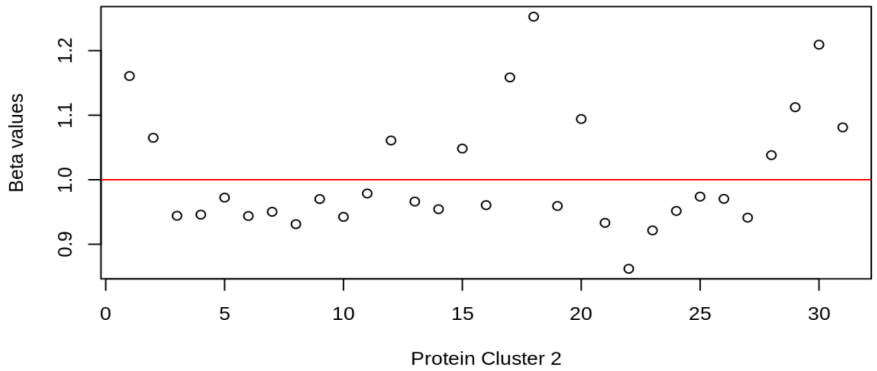
Expotype 0 vs all



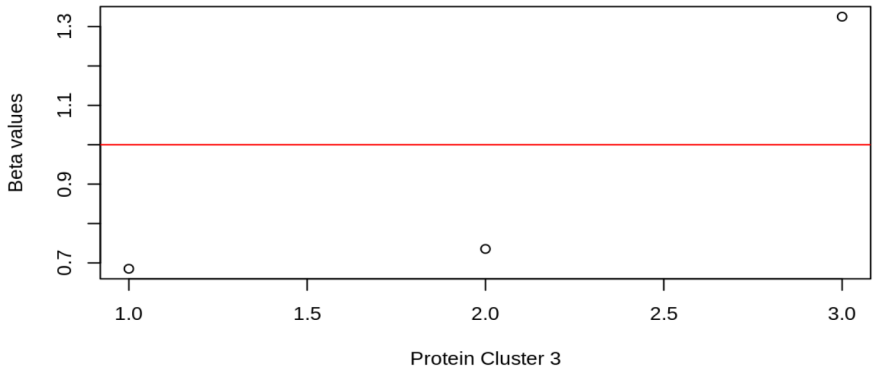
Expotype1 vs all



Expotype 2 vs all



Expotype 3 vs all



STEP 2 - A

Protein selection via stability LASSO

Number of times a protein was selected across the 4 clusters	4	3	2	1	0
Number of proteins selected	0 protein	3 proteins	17 proteins	256 proteins	1067 proteins

276 proteins stably selected in total

Missing obesity-related data by Expotype/clusters