Updates from ELSA: proteomics data quality control and study applications

Dr Jessica Gong

University College London

Jessica.gong@ucl.ac.uk

Talk overview

Part 1: Quality control pipeline

Part 2: Examples of study applications

Quality control pipeline

The protocol follows a six-step process:

1. Selection of samples and Olink panels
2. Data pre-processing by Olink
3. Bridging of Neuro Exploratory (NEX) panel
4. Importing data and removing Olink control and bridging samples
5. Removing data with quality control (QC) warnings or assay warnings
6. Outlier sample detection and removal
7. Removing proteins with ≥50% below limit of detection (LOD)

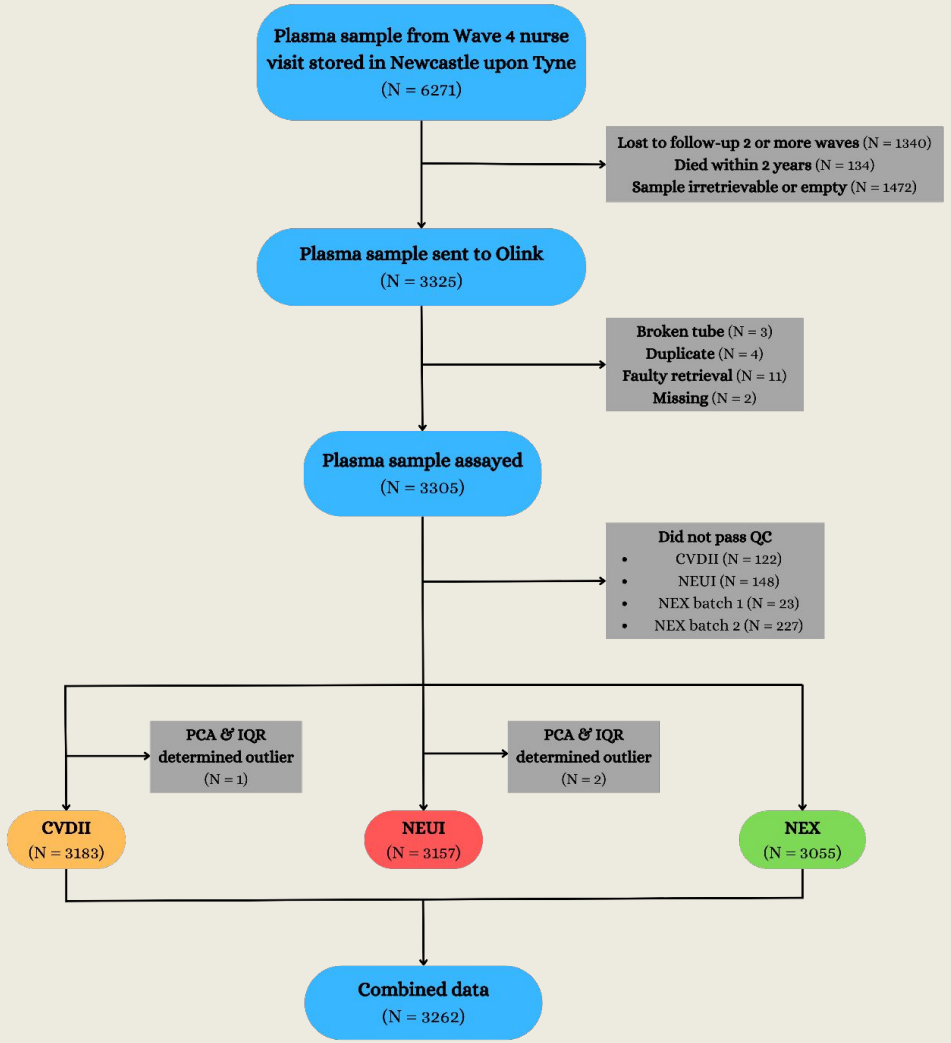
1. Selection of samples and Olink panels

- The English Longitudinal Study of Ageing (ELSA) cohort
  - **Aged 50+** in England
  - Started **2002–03**, followed up every **two years**

- Blood sample collection in ELSA
  - Collected in **wave 4** in **2008-09** during nurse visits
  - Samples collected in people's homes, then mailed to the lab in Newcastle stored at **-80°C**
  - No information on the time of the day blood sample was collected

- Proteomics data curation in ELSA
  - Specific focus on proteomics signatures associated with **dementia** and **cognitive decline**
  - **276** proteins across **three** Olink Target 96 panels (Cardiovascular disease II, Neurology I, Neuro Exploratory (analysed in two batches))

1. Selection of samples and Olink panels



QC, quality control; PCA, principal component analysis; IQR, interquartile range; CVDII, cardiovascular disease 2 Olink panel; NEU, neurology Olink panel; NEX, neurology exploratory Olink panel.

2. Data pre-processing by Olink

- Data in **normalized protein expression (NPX)** values, Olink's arbitrary unit on $\log_2$ scale
- Calculated from Cycle threshold (Ct) values and **data pre-processing (normalization)** is performed to minimize both intra- and inter-assay variation
- NPX data allows users to identify **changes for individual protein levels** across their sample set, and then use this data to **establish protein signatures**

3. Bridging for samples for NEX panel

NEX panel was analysed in two phases, bridging was performed. **16 random samples** were used for bridging, as recommended by Olink. Randomization was done using 'OlinkAnalyze'.

Dataset	n.Row (# of rows)	n.Ind (# of individuals)	n.Olinkid (# of proteins)	n.Plate (# of plates)	n.npx.na (# of missing NPX)
NEX_DATA_1	42872	466	92	6	184
NEX_DATA_2	270664	2942	92	38	319

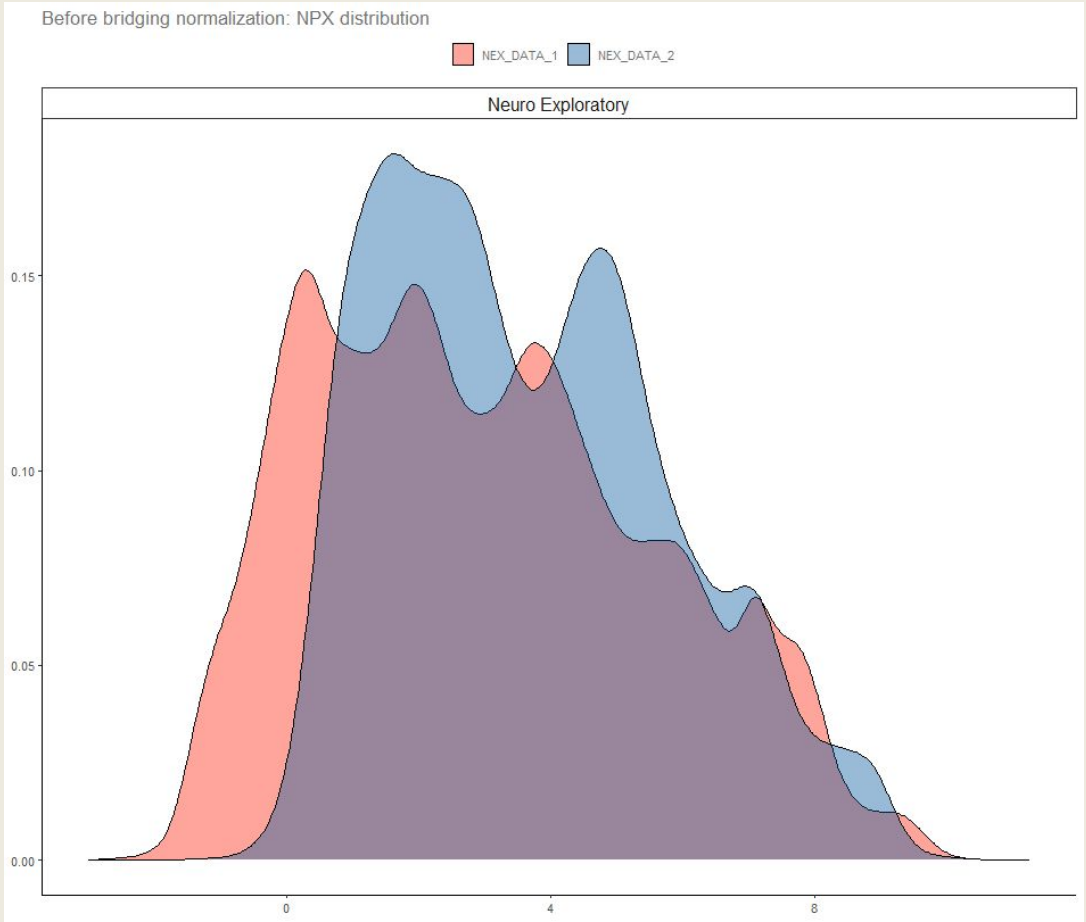
↓ Combining the two batches

Dataset	n.row	n.ind	n.olinkid	n.plate	n.npx.na
NEX_DATA	313536	3382	92	44	503

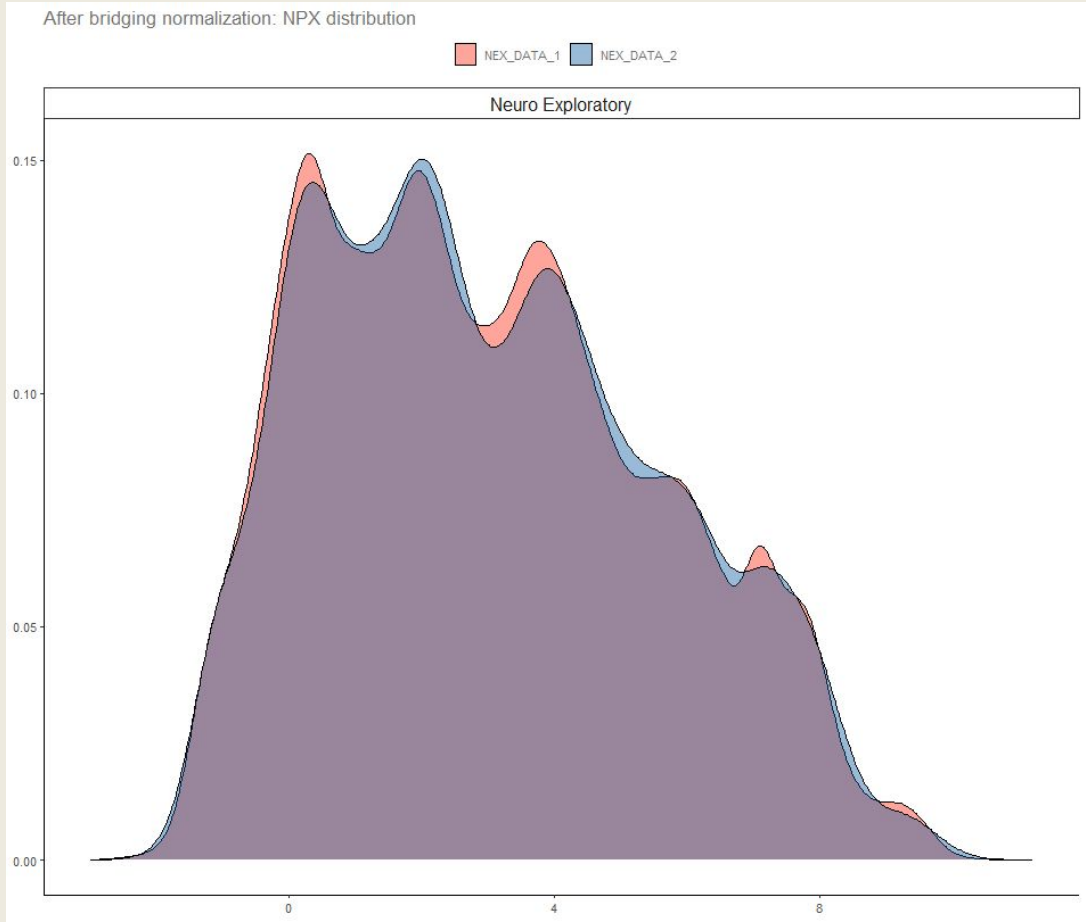
3. Bridging for samples for NEX panel

Density plot before and after normalization for NEX panel

Before bridging



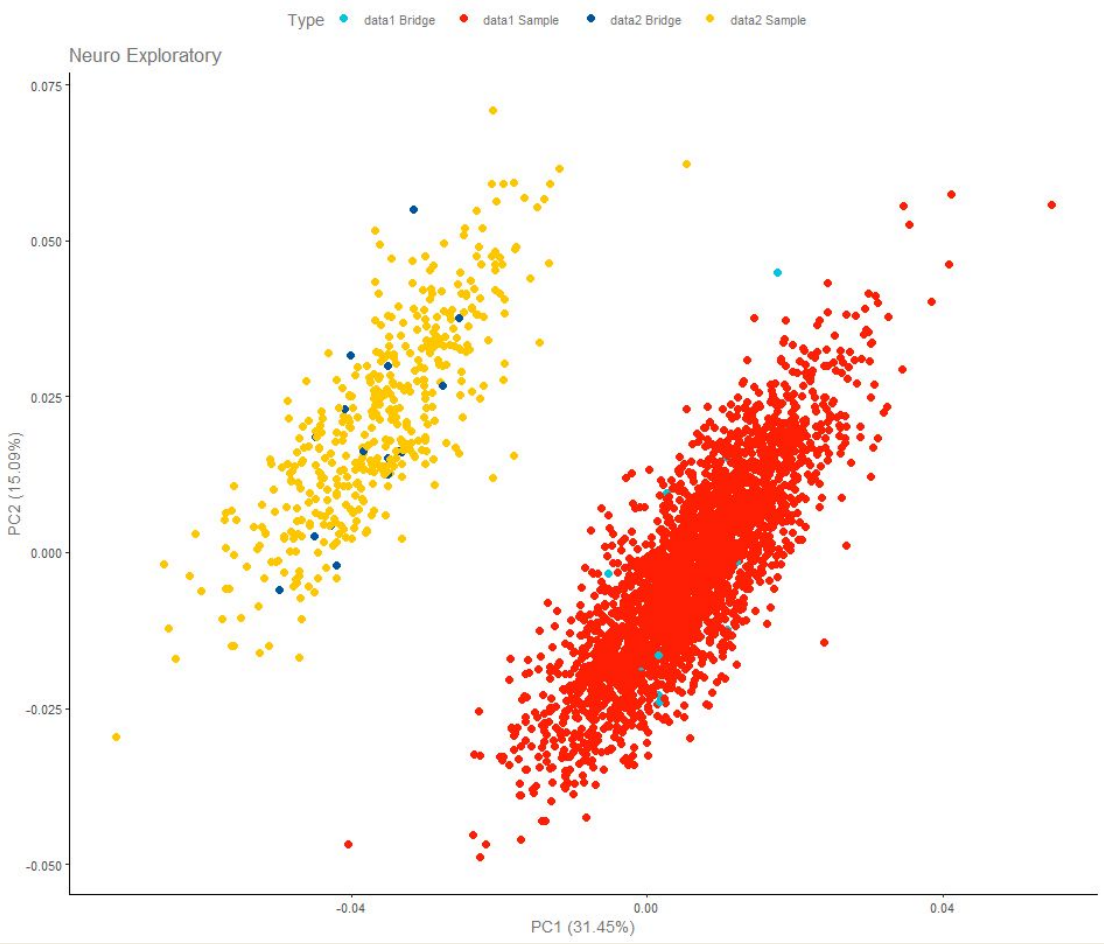
After bridging



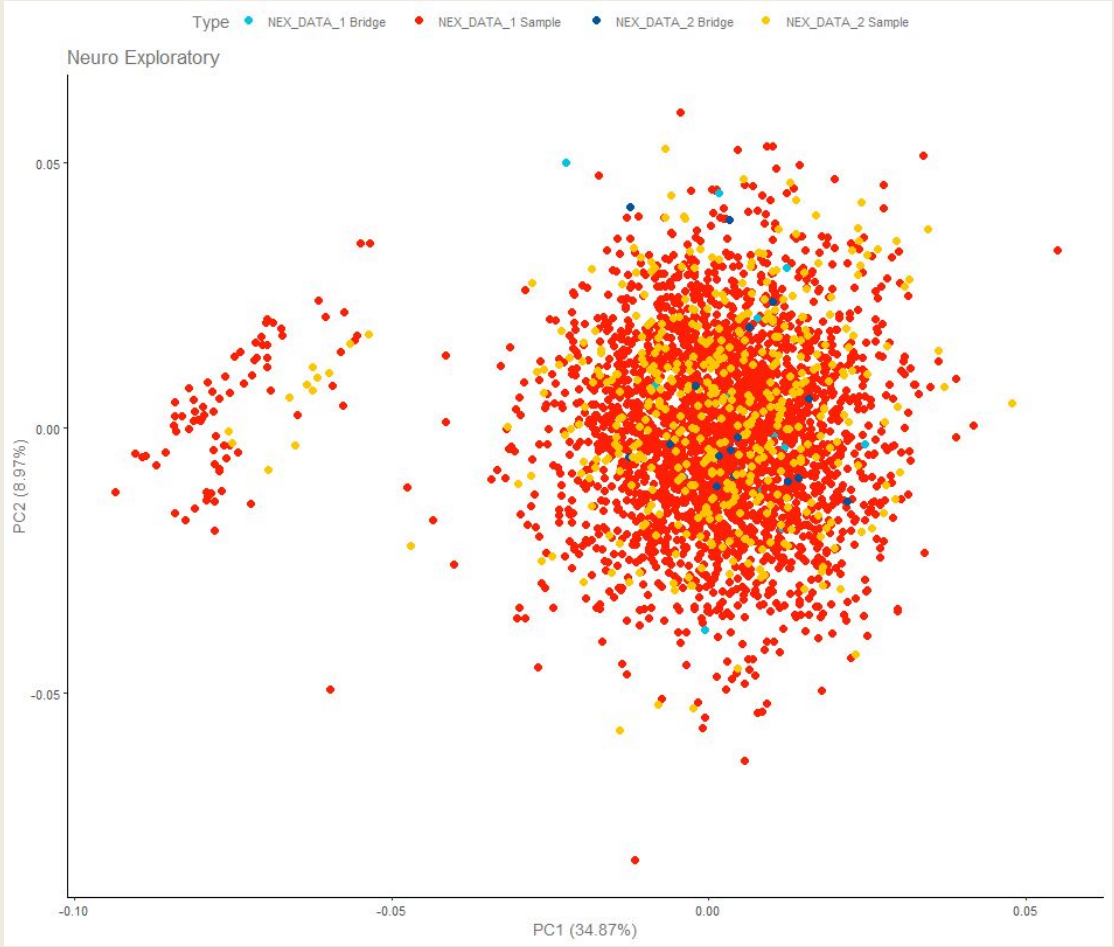
3. Bridging for samples for NEX panel

Principal component analysis (PCA) plot to visualise the samples in two datasets before and after bridging.

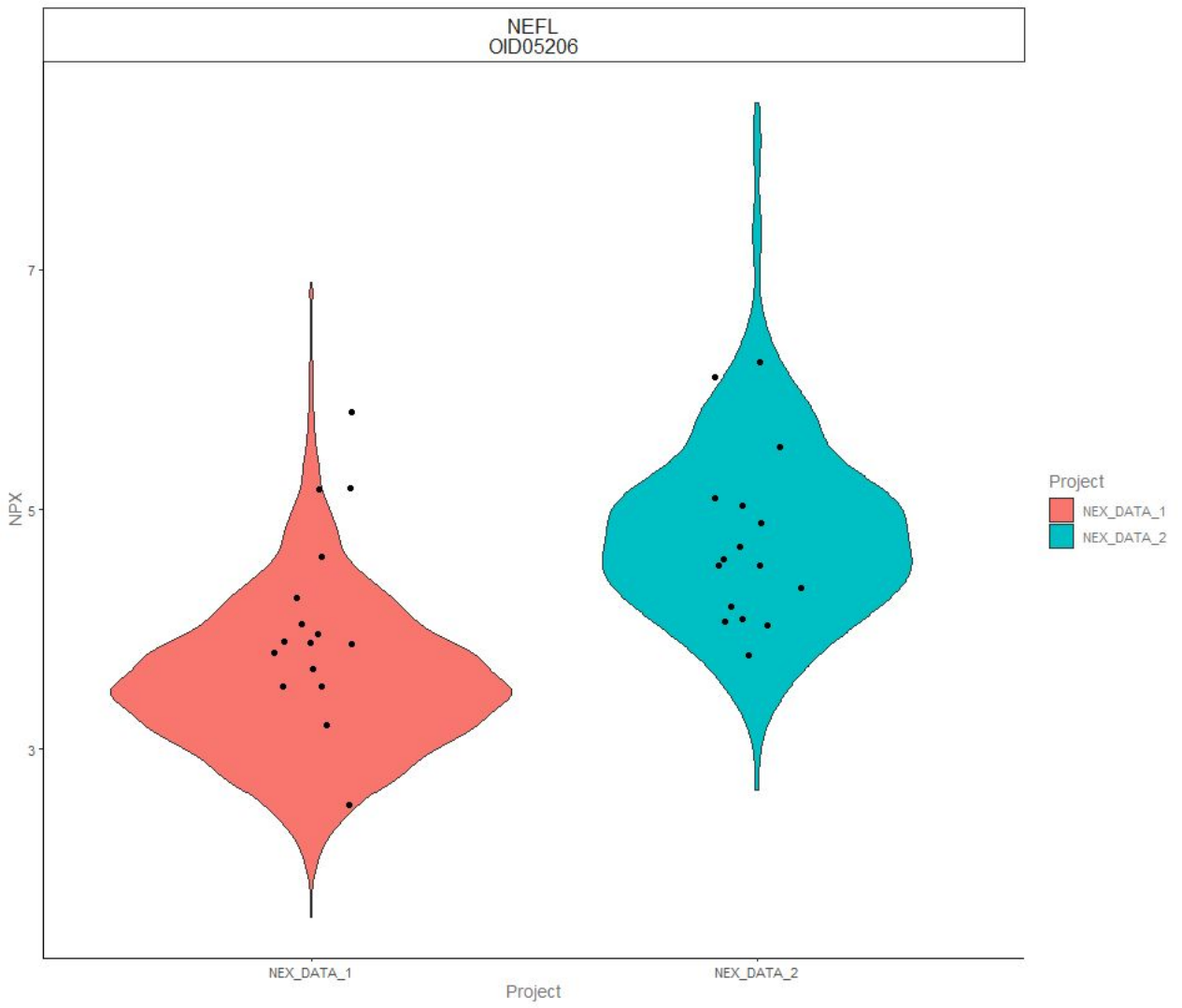
Before bridging



After bridging



3. Bridging for samples for NEX panel



4. Importing data and removing Olink control and bridging samples

- We began by importing the Olink™ Target 96 dataset for each of the three panels:

- Cardiovascular II (hereby referred to as '**CVDII_DATA**')
- Neurology I (hereby referred to as '**NEUI_DATA**')
- Neurology exploratory (hereby referred to as '**NEX_DATA**') – After bridging.
  - Neurology exploratory phase 1 (hereby referred to as '**NEX_DATA_1**')
  - Neurology exploratory phase 2 (hereby referred to as '**NEX_DATA_2**')

4. Importing data and removing Olink control and bridging samples

Dataset	n.row	n.ind	n.olinkid	n.plate	n.npx.na
CVDII_DATA	311052	3381	92	38	375
NEUI_DATA	311052	3381	92	38	1103
NEX_DATA	313536	3382	92	44	503
NEX_DATA_1	42872	466	92	6	184
NEX_DATA_2	270664	2942	92	38	319

↓ Removing controls and bridging samples in NEX data

Dataset	n.row	n.ind	n.olinkid	n.plate	n.npx.na
CVDII_DATA	304060	3305	92	38	375
NEUI_DATA	304060	3305	92	38	1101
NEX_DATA	303968	3304	92	44	500

5. Olink internal quality control

Four internal controls are added to each sample to **monitor the quality of assay performance**, as well as the **quality of individual samples**.

The internal quality control (QC) is performed in **two** steps:

1. Each sample plate is evaluated on the **standard deviation of the internal controls**. This should be below **0.2 NPX**. Only data from sample plate that pass this quality control will be reported.
2. The **quality of each sample** is assessed by **evaluating the deviation from the median value** of the controls for each individual sample. Samples that **deviate less than 0.3 NPX** from the **median** pass the quality control.

5. Olink internal quality control

Samples that did not pass the QC are indicated in columns named **"QC Warning"**

Dataset characteristics after removing data with QC Warnings.

Dataset	n.row	n.ind	n.olinkid	n.plate	n.npx.na
CVDII_DATA	292836	3183	92	38	6
NEUI_DATA	290444	3157	92	38	78
NEX_DATA	281060	3055	92	44	36

6. Outlier sample detection and removal

IQR-median and PCA analyses are used to detect and visualise outliers.

IQR-median quality control plot in ‘OlinkAnalyze’

The olink_qc_plot function generates a facet plot per Panel plotting interquartile range (IQR) vs. median for all samples. This checks if any samples tend to be classified as outliers.

6. Outlier sample detection and removal

PCA qc plot using 'OlinkAnalyze'

- Generates PCA projection of all samples from NPX data along two principal components (Default PC2 vs PC1) coloured by the variable and including the percentage of explained variance.
- By default, the values scaled and cantered in the PCA.

Based on these analyses, we then removed:

1) Samples with a PC1 or PC2 value more than five standard deviations from the mean, and
2) samples with median concentrations (NPX values) across proteins that were more than five standard deviations from the mean median, or those with IQR (NPX) across proteins that were great than five standard deviations from the mean IQR.
3) The removal of outliers is panel specific.

6. Outlier sample detection and removal

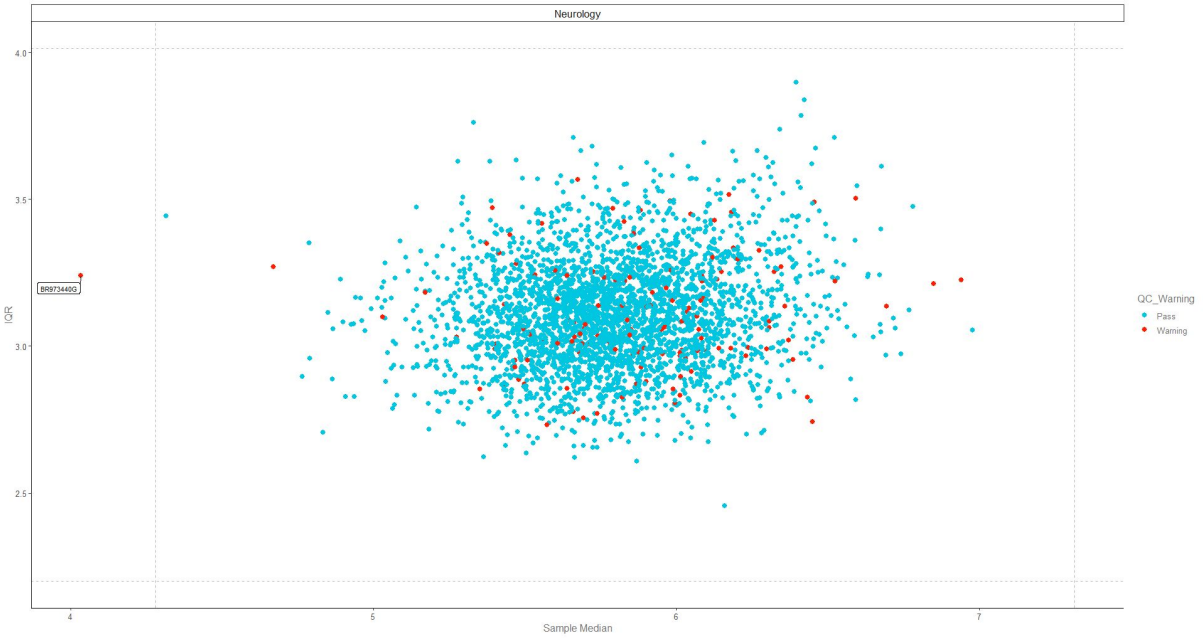
IQR-median and PCA analyses are used to detect and visualise outliers.

IQR-median plot for CVD II panel.

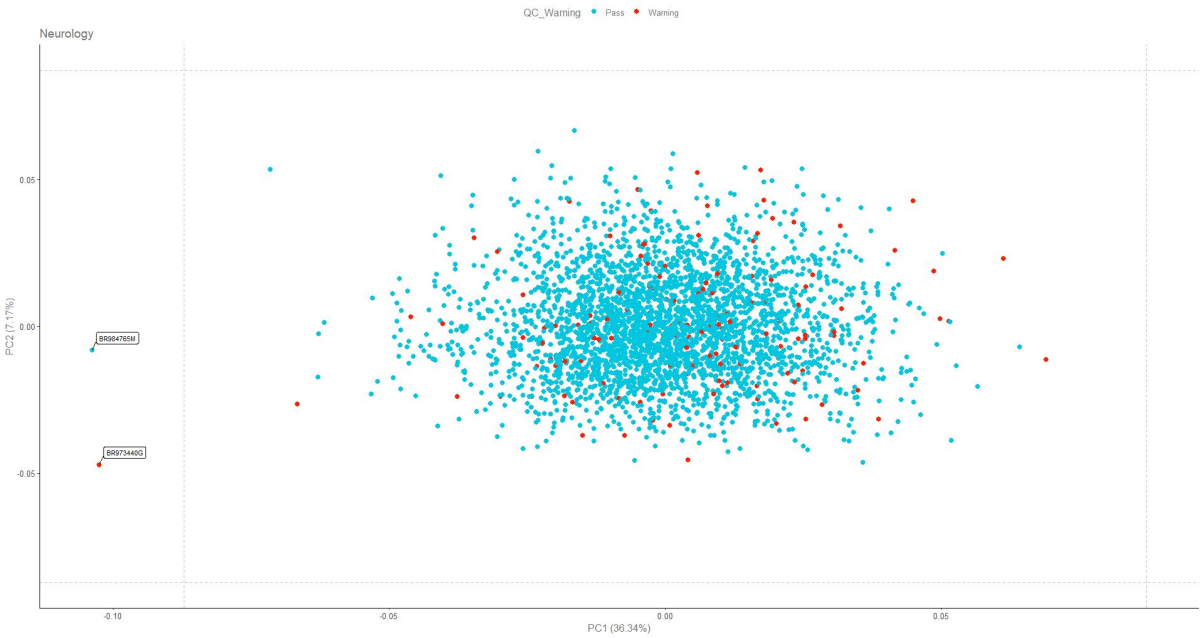
PCA plot for CVD II panel.

6. Outlier sample detection and removal

IQR-median plot for NEU I panel.

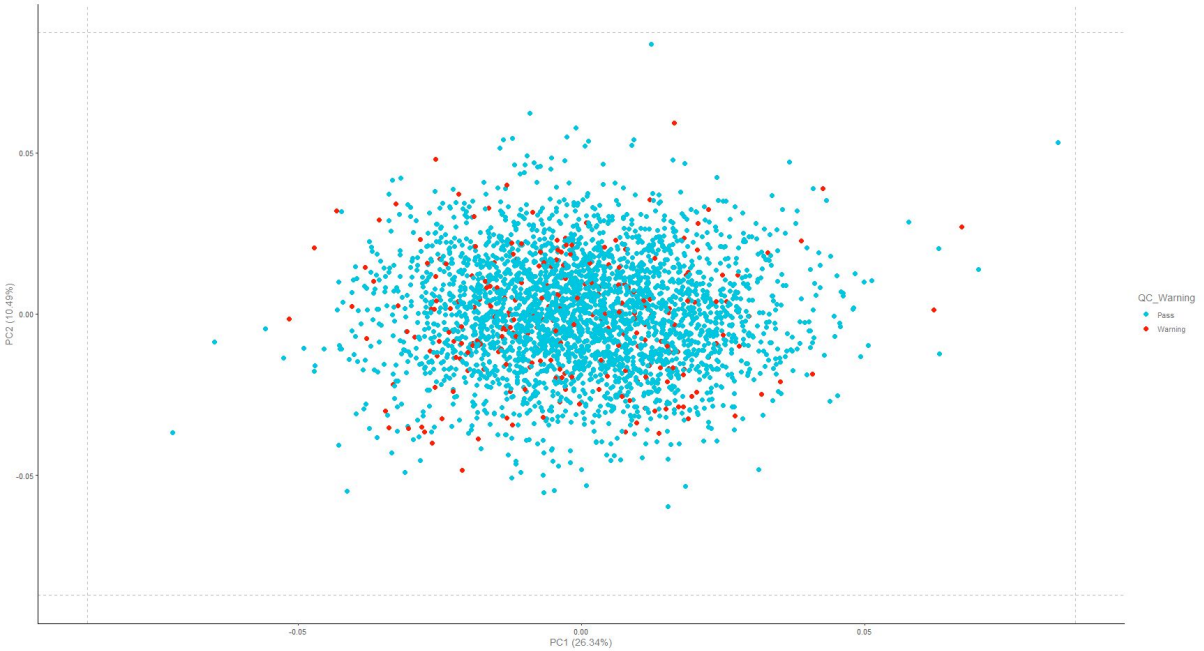


PCA plot for NEU I panel.

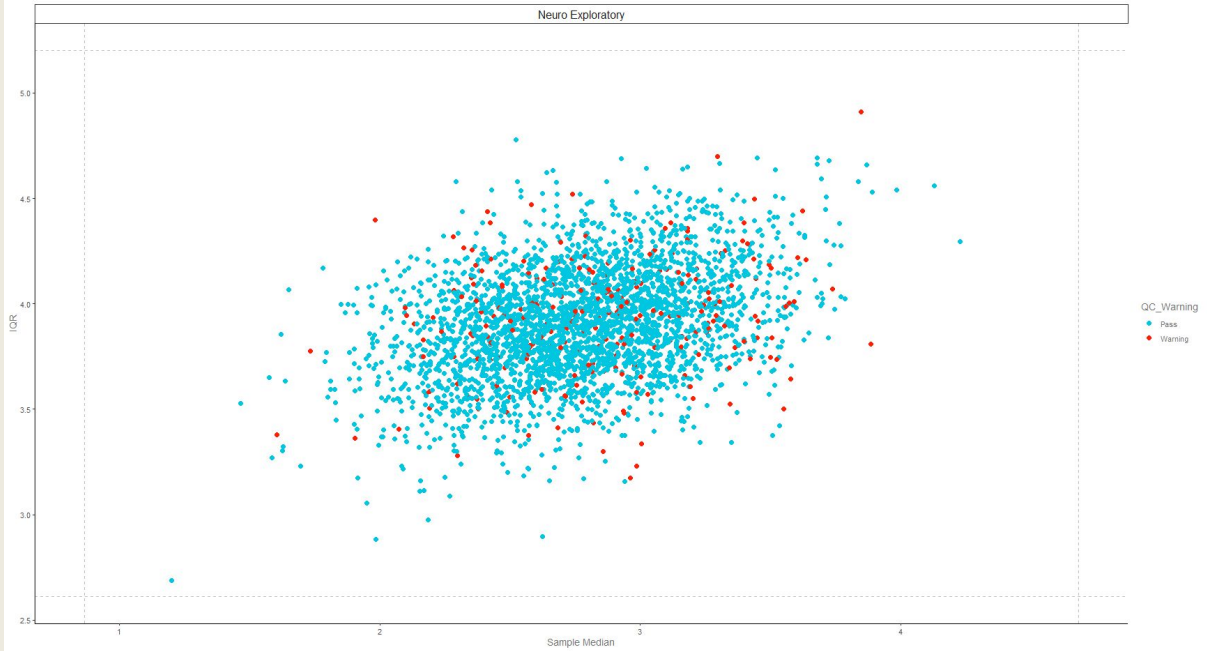


6. Outlier sample detection and removal

IQR-median plot for NEX panel.



PCA plot for NEX panel.



6. Outlier sample detection and removal

Panel	Outlier sample #
CVDII_DATA	1
NEUI_DATA	2
NEX_DATA	0

↓ Removing outliers

Dataset	n.row	n.ind	n.olinkid	n.plate	n.npx.na
CVDII_DATA	292744	3182	92	38	6
NEUI_DATA	290352	3156	92	38	78
NEX_DATA	281060	3055	92	44	36

↓ Combing all three panels

Dataset	n.row	n.olinkid
Combined data	3262	276

7. Dealing with the limit of detection (LOD)

- Limit of detection (LOD) is calculated separately for each Olink assay and sample plate.
- The LOD is based on the background, estimated from negative controls included on every plate, plus three standard deviations.
- The standard deviation is assay specific and estimated during product validation for every panel.
- As with all affinity-based assays, data from the Olink platform has a S-curve (sigmoid) relationship with the true protein concentration in a sample.
- Data below LOD have a higher risk to be in the non-linear phase of the S-curve meaning that 1 NPX difference may not correspond to 2x protein concentration in this region.

7. Dealing with the limit of detection (LOD)

CVDII_DATA	NEUI_DATA	NEX_DATA
Protein name (% below LOD)	**Protein name (% below LOD)**	**Protein name (% below LOD)**
BNP (57%)	MAPT (89%)	HSP90B1 (56.9%)
	CADM3 (81.5%)	NXPH1 (96.3%)
	Beta-NGF (98%)	IKZF2 (72.1%)
		EPHA10 (84.2%)

7. Dealing with the limit of detection (LOD)

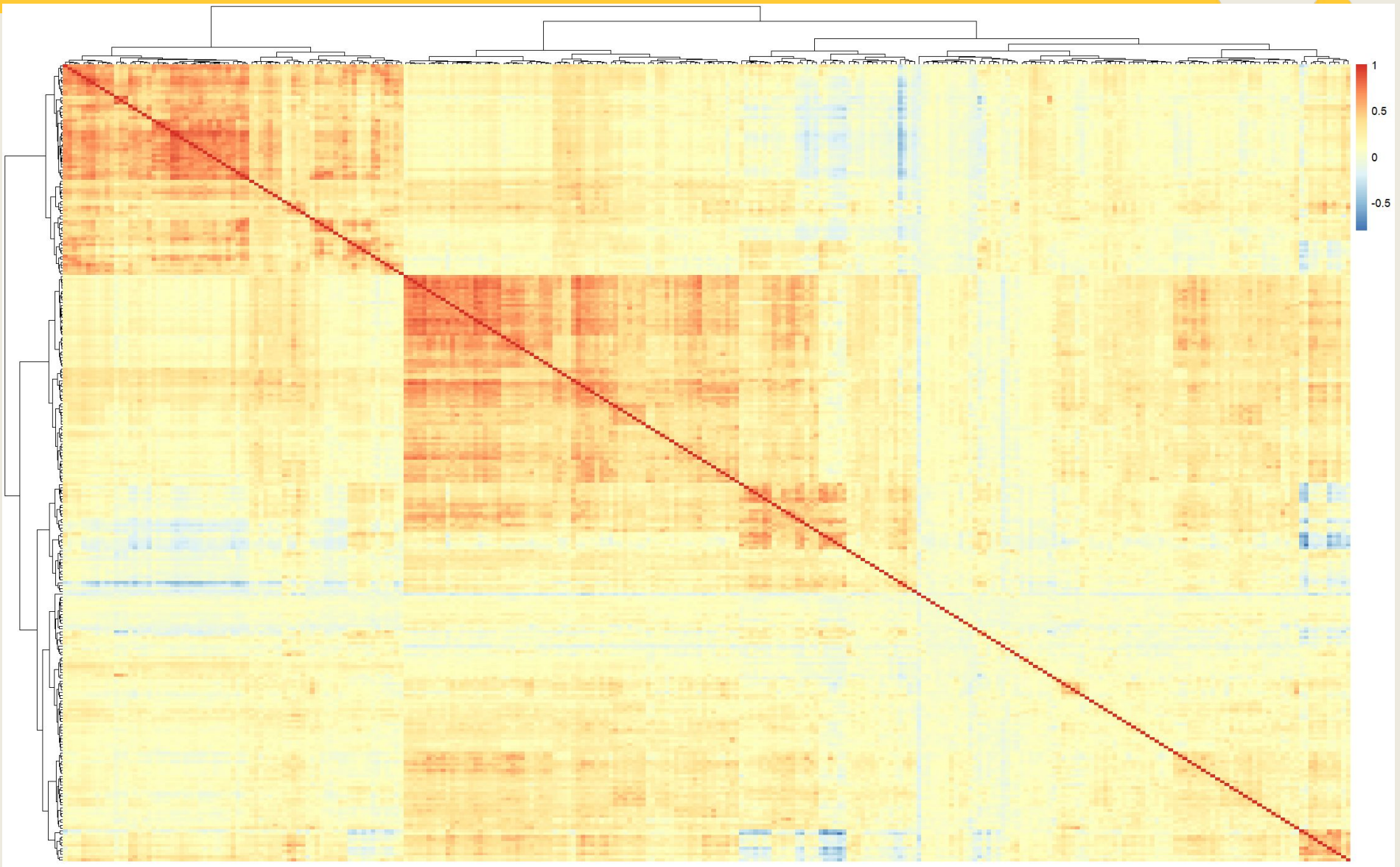
Strategies for handling data below LOD in data analysis (proposed by Olink):

1. Use actual data below LOD
2. Replace data below LOD
3. Impute data below LOD

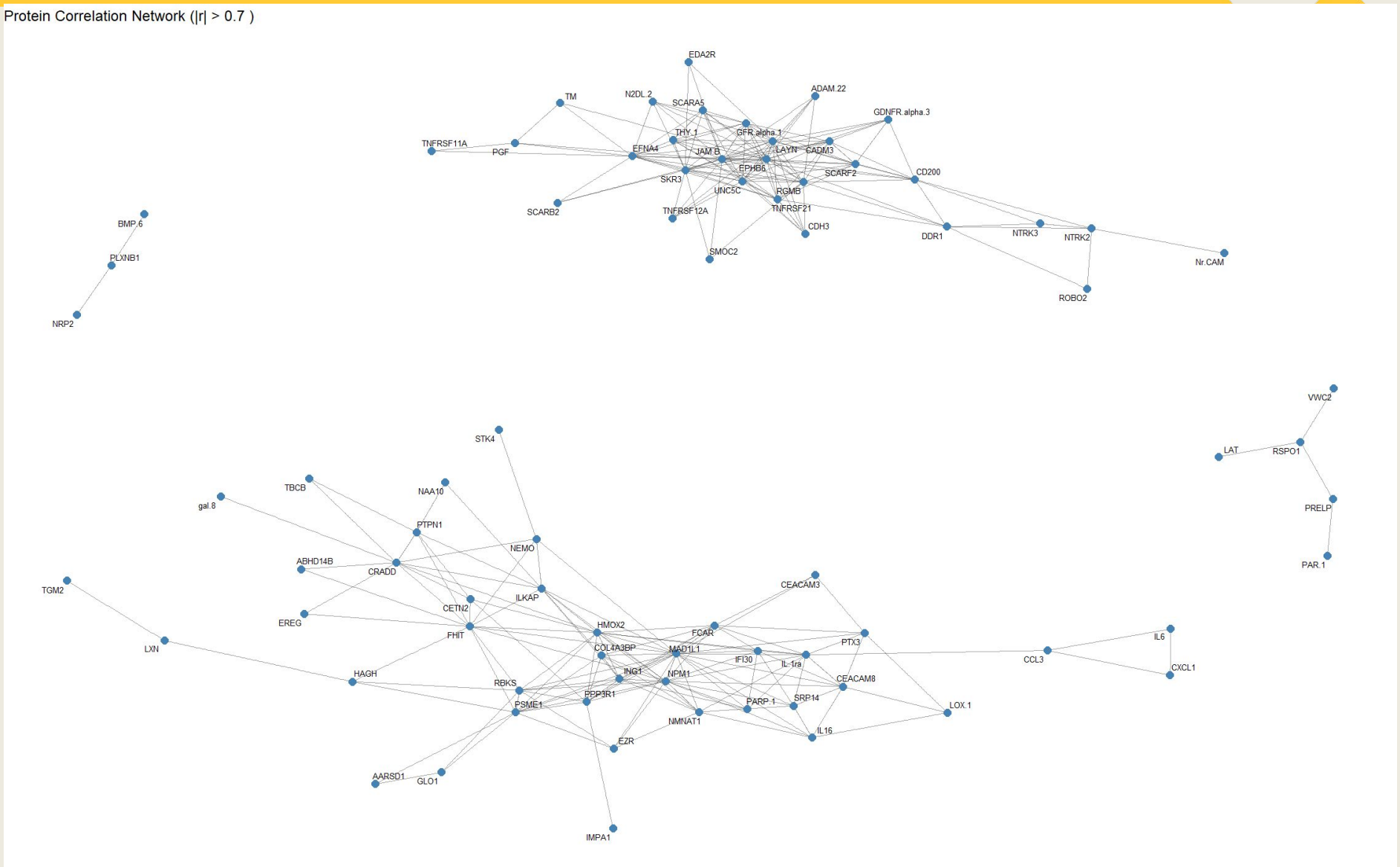
Olink recommends to not remove data points below LOD, as some of the most distinct biomarkers may be low in some groups analysed but high in other groups

Taking a more liberal approach, the exclusion of proteins with ≥50% below LOD may not be necessary What we did in ELSA

Correlations between proteins

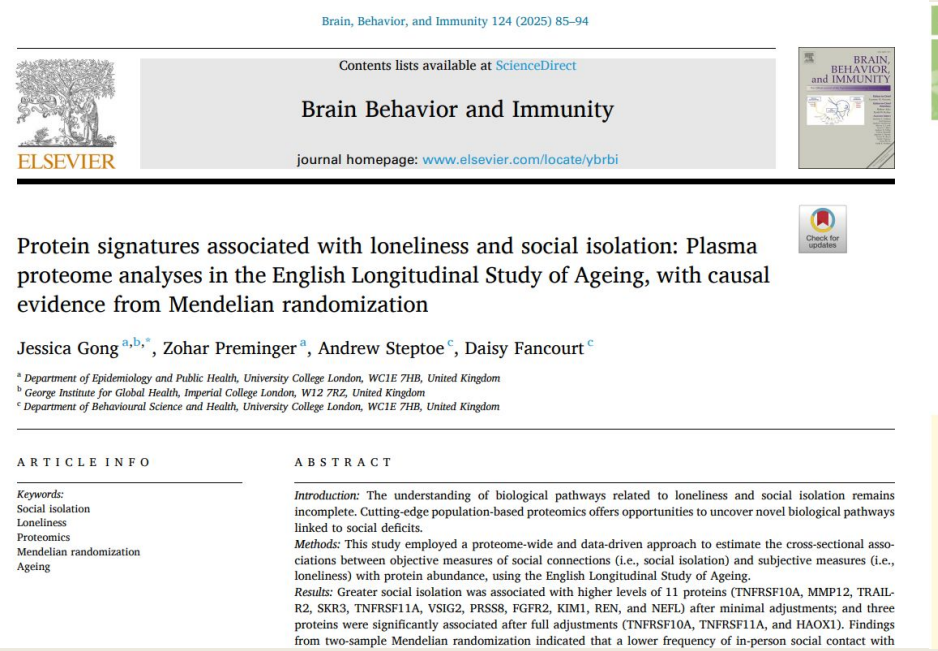


Correlations between proteins (highly correlated |r|>0.7)



Study applications

Social isolation and loneliness



Brain, Behavior, and Immunity 124 (2025) 85–94

Contents lists available at ScienceDirect

Brain Behavior and Immunity

journal homepage: www.elsevier.com/locate/ybrbi

Protein signatures associated with loneliness and social isolation: Plasma proteome analyses in the English Longitudinal Study of Ageing, with causal evidence from Mendelian randomization

Jessica Gong [a,b,*], Zohar Preminger [a], Andrew Steptoe [c], Daisy Fancourt [c]

[a] Department of Epidemiology and Public Health, University College London, WC1E 7HB, United Kingdom
[b] George Institute for Global Health, Imperial College London, W12 7RZ, United Kingdom
[c] Department of Behavioural Science and Health, University College London, WC1E 7HB, United Kingdom

ARTICLE INFO

Keywords:
Social isolation
Loneliness
Proteomics
Mendelian randomization
Ageing

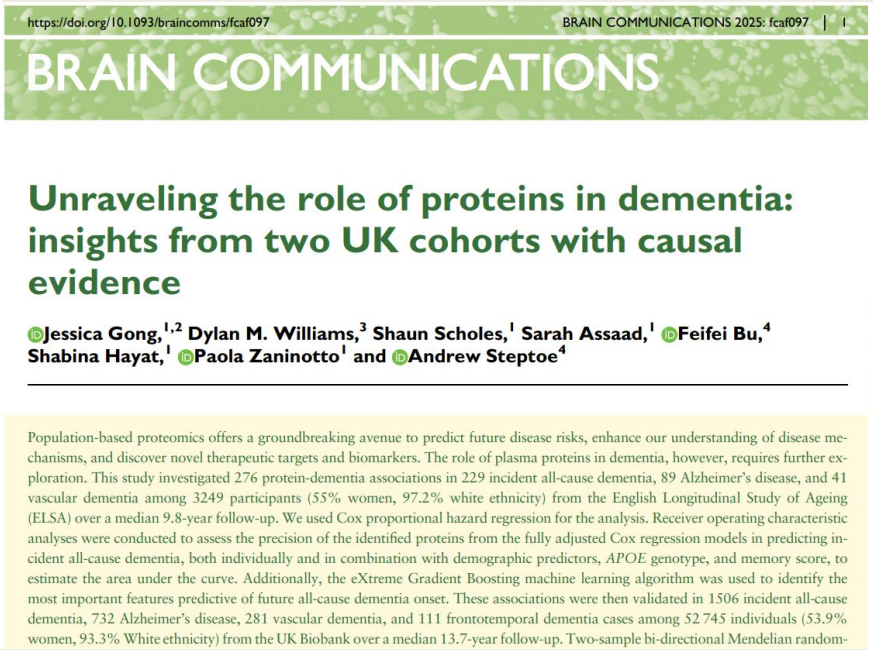
ABSTRACT

Introduction: The understanding of biological pathways related to loneliness and social isolation remains incomplete. Cutting-edge population-based proteomics offers opportunities to uncover novel biological pathways linked to social deficits.

Methods: This study employed a proteome-wide and data-driven approach to estimate the cross-sectional associations between objective measures of social connections (i.e., social isolation) and subjective measures (i.e., loneliness) with protein abundance, using the English Longitudinal Study of Ageing.

Results: Greater social isolation was associated with higher levels of 11 proteins (TNFRSF10A, MMP12, TRAIL-R2, SKR3, TNFRSF11A, VSIG2, PRSS8, FGFR2, KIM1, REN, and NEFL) after minimal adjustments; and three proteins were significantly associated after full adjustments (TNFRSF10A, TNFRSF11A, and HAOX1). Findings from two-sample Mendelian randomization indicated that a lower frequency of in-person social contact with

Incident dementia (all-cause dementia, Alzheimer's disease, vascular dementia)



https://doi.org/10.1093/braincomms/fcaf097 BRAIN COMMUNICATIONS 2025: fcaf097 | 1

BRAIN COMMUNICATIONS

Unraveling the role of proteins in dementia: insights from two UK cohorts with causal evidence

Jessica Gong,[1,2] Dylan M. Williams,[3] Shaun Scholes,[1] Sarah Assaad,[1] Feifei Bu,[4] Shabina Hayat,[1] Paola Zaninotto[1] and Andrew Steptoe[4]

Population-based proteomics offers a groundbreaking avenue to predict future disease risks, enhance our understanding of disease mechanisms, and discover novel therapeutic targets and biomarkers. The role of plasma proteins in dementia, however, requires further exploration. This study investigated 276 protein-dementia associations in 229 incident all-cause dementia, 89 Alzheimer's disease, and 41 vascular dementia among 3249 participants (55% women, 97.2% white ethnicity) from the English Longitudinal Study of Ageing (ELSA) over a median 9.8-year follow-up. We used Cox proportional hazard regression for the analysis. Receiver operating characteristic analyses were conducted to assess the precision of the identified proteins from the fully adjusted Cox regression models in predicting incident all-cause dementia, both individually and in combination with demographic predictors, *APOE* genotype, and memory score, to estimate the area under the curve. Additionally, the eXtreme Gradient Boosting machine learning algorithm was used to identify the most important features predictive of future all-cause dementia onset. These associations were then validated in 1506 incident all-cause dementia, 732 Alzheimer's disease, 281 vascular dementia, and 111 frontotemporal dementia cases among 52 745 individuals (53.9% women, 93.3% White ethnicity) from the UK Biobank over a median 13.7-year follow-up. Two-sample bi-directional Mendelian random-

Psychological wellbeing (eudaimonic wellbeing, hedonic wellbeing, evaluative wellbeing, depressive symptoms)

medRxiv THE PREPRINT SERVER FOR HEALTH SCIENCES

Cold Spring Harbor Laboratory BMJ Yale

Follow this preprint

Associations between plasma proteins and psychological wellbeing: evidence from over 20 years of the English Longitudinal Study of Ageing

Jessica Gong, Shaun Scholes, Steven Cole, Paola Zaninotto, Andrew Steptoe

doi: https://doi.org/10.1101/2025.02.05.25321715

This article is a preprint and has not been peer-reviewed [what does this mean?]. It reports new medical research that has yet to be evaluated and so should *not* be used to guide clinical practice.

Abstract Full Text Info/History Metrics Preview PDF

ABSTRACT

A deeper understanding of the molecular processes involved in psychological wellbeing in older adults is essential for advancing knowledge of underlying biological mechanisms. Leveraging proteomics data from 3,262 older adults (mean age=63.5 years, 55% female) of the English Longitudinal Study of Ageing (ELSA), we investigated

+ cognitive decline x1, sugar rationing x1