UK Biobank

Olink proteomics data

Version 1.0

http://www.ukbiobank.ac.uk/

March 2023

Contents

1. Background

1.1. This document provides an explanation of the proteomics data available in UK Biobank.

1.2. Proteomic analysis of blood plasma samples was performed by Olink, using Proximity Extension Assay (PEA).

1.3. Funding was provided by the Pharma Proteomics Project consortium – a collaboration between UK Biobank and thirteen biopharmaceutical companies, organised to profile the plasma proteome within a cross-section of UK Biobank participants.

1.4. Proteomic biomarker data were produced using 55,000 samples from unique UK Biobank participant-visits, analysed over the period April 2021 – Feb 2022.

1.5. This covers up to 2923 unique assays per sample from the Olink Explore platform.

2. Proteomics data format

2.1. Data Portal – the Normalized protein expression (NPX) measure is outlined in UKB_Olink_Explore_1536_B0_to_B7_Data_Normalization_Strategy.pdf. Due to size and complexity, NPX results are not provided as individual Showcase fields, and are accessed via the Data Portal. The olink_data table lists the NPX value for each available protein per participant eid and instance.

2.2. Showcase fields – Plate and well position for each sample supplied are provided as instanced Showcase fields. There is additionally a field to indicate the number of NPX results available for each sample analysed, and is used to grant access to the proteomics portal table in 2.1.

2.3. Showcase resources – Assay-level results (e.g. the plate limit of detection for a protein, or QC indicators like batch or lot number) are provided as downloadable showcase resources. These are generic tab-separated datasets and are available via the resources section in Category 1839. An overview of each resource is provided in the next section of this document.

2.4. Encodings – an encoding index (encoding 143) can be used to link the protein ID in the NPX data to the UniProt text description of the protein. This lookup also allows for joining between the NPX results data and the QC or assay level data.

3. Resource datasets

3.1. olink_assay – provides the lookup between an assay, its respective UniProt ID, and the Olink Explore panel in which it is categorised.

3.2. olink_assay_version – provides the version number for each assay per panel lot number.

3.3. olink_batch_number – provides the shipment batch number for each plate ID, allowing for correction of potential batch processing effects.

3.4. olink_limit_of_detection – provides the instance-level limit of detection for each assay per shipment plate, allowing for filtering of sample results based on target protein detectability (e.g. result > LOD).

3.5. olink_panel_lot_number – provides the processing lot number per assay panel within each shipment batch.

3.6. olink_processing_start_date – provides the processing date for each shipment plate, broken down by assay panel.

4. Data preparation steps

4.1. The Olink data have been assessed and pre-filtered by the Pharma Proteomics Project consortium, details of which are evidenced in UKB_PPP_Phase_1_QC_dataset_companion_document.pdf.

4.2. An additional data reduction step was performed by UK Biobank for 3 validation assays which overlapped the 4 Explore 1,536 assay panels (Cardiometabolic, Inflammation, Neurology, and Oncology) and 3 validation assays which overlapped the 4 Expansion 1460 assay panels (Cardiometabolic II, Inflammation II, Neurology II and Oncology II)

4.3. The following versions of each assay were chosen based on highest target protein detectability, followed if necessary by the largest number of NPX data points exceeding the respective limit of detection:

4.3.1. Tumor necrosis factor (TNF, P01375) – Cardiometabolic

4.3.2. Interleukin-6 (IL6, P05231) – Oncology

4.3.3. Interleukin-8 (CXCL8, P10145) – Oncology

4.3.4. Indoleamine 2,3-dioxygenase 1 (IDO1, P14902) – Cardiometabolic II

4.3.5. Leiomodin-1 (LMOD1, P29536) – Neurology II

4.3.6. Protein scribble homolog (SCRIB, Q14160) – Cardiometabolic II

5. Example resources use case

5.1. Olink resources cannot be joined directly to the results data from within the Data Portal.

5.2. The resource datasets can be combined using a statistical software package to create a long format list of assay results per shipment plate.

5.3. Limit of detection for each assay per shipment plate can be retrieved using the olink_limit_of_detection.dat dataset. This can then be joined to the NPX data via the indexed assay name, instance index value, plus the plate ID obtained from field 30901. Results can be filtered for subsequent analysis if, for example, detectability was sufficient (result > LOD).

6. Proteomics datasets diagram