UKB – Olink Explore 1536 - Data Normalization Strategy

Samples from the study were divided into two sets: i) Set 1 – UKB; and ii) Set 2 – COVID; depending on the time point they were randomly selected from the UKB population. Samples were randomly assigned to 96-well plates, and fully randomized within plates. Each plate contained: i) 87 samples from set 1, set 2 or both; ii) 1 empty well, bridge or overlapping sample; iii) 2 Olink control samples used for quality control; iv) 3 Olink negative control samples used to compute the baseline assay level of each plate; and v) 3 Olink plate control samples used for normalization of protein expression. All Olink samples were placed in column 12 of plates while the remaining 87 samples + 1 empty well were randomized across columns 1-11 and rows A to H (Figure 1).
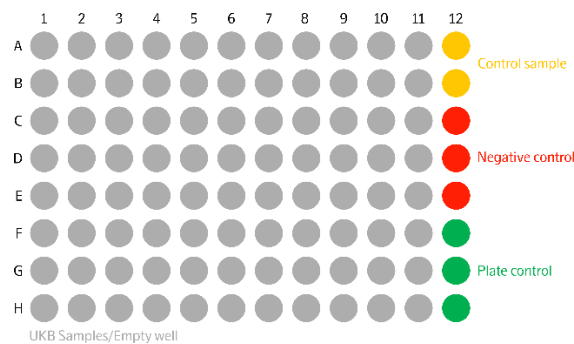

Figure 1: Layout of Olink plates with UKB samples. Each plate had one empty well between A1 and H11.

Plates were shipped in 8 batches (0-7) that consisted of different numbers of samples (Figure 2). Batches 0-6 contained exclusively samples from set 1, while batch 7 contained samples from both sets. A subset of 25 plates from batch 7 contained exclusively samples from set 1, while 94 of them contained samples from both sets.
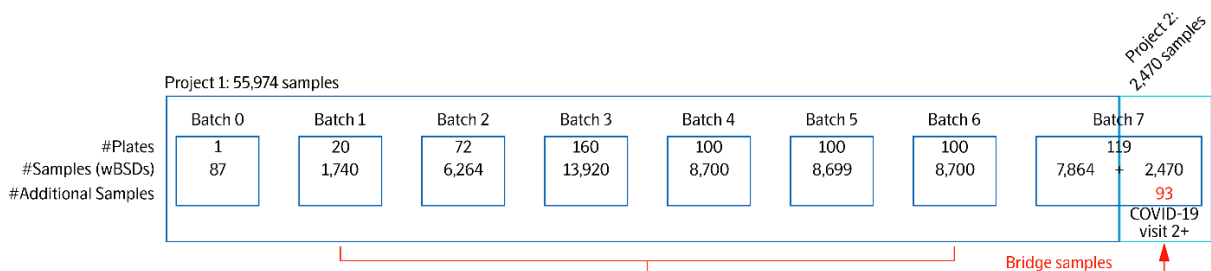

Figure 2: Schematic representation of the study design for UKB.

Calculation of Normalized Protein eXpression (NPX) values was performed stepwise as described below. Initially, we calculated the $\log_2$ ratio of counts of each assay of each sample to the counts of the extension control

$$ExtNPX_{i,j} = \log_2\left(\frac{Counts(Sample_j Assay_i)}{Counts(ExtensionControl_j)}\right)$$

, and the assay-specific median value of the plate controls was subtracted.

$$NPX_{i,j} = ExpNPX_{i,j} - median(ExtNPX(PlateControls_i))$$

This provided us with plate control (PC) normalized NPX values for both sets.

For samples belonging to set 1, we subtracted the assay-specific plate median NPX value, and we added the batch-specific median NPX value of each assay to account for effects within each batch.

$$NPX^{set_1}_{IntNorm_{i,j,k}} = NPX_{i,j,k} - plate_k median(NPX_i)$$

At this stage data was normalized within each batch. Next, we computed adjustment factors from the difference of the assay specific median NPX value of each batch to the reference batch (batch 1). Batch 1 was the first to run chronologically, underwent extensive investigation on randomization and assay performance, hence, it was selected as the reference batch. However, selection of any batch as reference would result in identical results.

$$adj\_factor^{set_1}_{batch_1-batch_{n_i}} = median\left(NPX_{batch_{1_i}}\right) - median(NPX_{batch_{n_i}})$$

Finally, adjustment factors from above were added to the NPX values of each batch of set 1.

$$NPX^{final_{set_1}}_{IntNorm_{i,j,n}} = NPX^{set_1}_{IntNorm_{i,j}} + adj\_factor^{set_1}_{batch\_1-batch\_n_i}$$

In summary, set 1 was normalized using a two-step approach of within-batch and across-batches intensity normalization.

Samples from set 2 were normalized using reference (bridge) samples that were shared between the two sets. All plates that contained at least one sample of set 2 were assigned a randomly selected sample from set 1. 93 samples with missing frequency <10% and representative of the NPX dynamic range were selected from batches 1-6 of set 1. These samples were assigned to the empty well of the 93 plates from batch 7 of set 2. In this case, adjustment factors were computed from the assay-specific median of the pair-wise differences between set 1 and set 2.

$$adj\_factor^{set_2}_{set_1-set_{2_i}} = median\left(NPX_{set_{1_{i,j}}} - NPX_{set_{2_{i,j}}}\right)$$

Adjustment factors were added to the NPX values of set 2.

$$NPX^{final_{set_2}}_{BridgeNorm_{i,j,n}} = NPX^{set_2}_{i,j} + adj\_factor^{set_2}_{set_1-set_{2_i}}$$

Intensity normalized NPX values for set 1 and bridge normalized NPX values for set 2 were the final set of NPX values that was used for the downstream analysis.