1 Cohort bias detection

Before we perform cohort bias removal we seek to quantify the presence of such bias. We have two general approaches: distribution based and pairwise similarity.

Distribution based:

- Wasserstein metrics
- Unsupervised non-parametric statistical significance test: Mann-Whitney U, Kolmogorov-Smirnof
- Supervised non-parametric statistical significance test: FDR-ANOVA

Two options for the application: 1. compare cohorts per feature (or reduced dimension) (columnwise) 2. compare cohorts per patients over the features (or reduced dimensions) (rowwise)

Pairwise similarity:

- Kullback-Leibler divergence
- Distance metrics/correlation

Here, in general we only have on option for the application which is to compare the cohorts per inter-cohort patient-pair.

When comparing cohorts we have choose to compare each cohort with eachother, or we can compare each cohort with the overall distribution (minus that specific cohort). Classification based:

- separation of biological classes by batch identities

Variation based:

- relations between in-group variance, out-group variance and between-group variances, see Hicks

2 Cohort bias removal

The following bias removal methods are applied

- RNA expression data: L/S adjustment & cohort based QN
- Methylation data: 1. cohort correction using ComBat & cohort based QN. 2. SmoothedQN (color), 4. SubsetQN (type)/SubsetQN (islands)

We apply the cohort bias removal to the measurement cohorts. These cohorts indicate measurement batches and the cohort bias removal reduces any bias that is seemingly related to the cohorts. Arguably we have to apply the bias removal, per cohort, per phenotypical cluster, otherwise the applicability of the cohort bias removal hinges on the degree of stratification of the phenotypes. This is however prohibited by the sparsity of the data. The ComBat method uses a combination of L/S normalisation/scaling and empirical Bayes to assess the bias that is introduced by the cohort. As a reference we apply L/S, and cohort-wise QN.
We use the same cohort-bias correction for both the RNA expression data and the methylation data.

Results are evaluated using:

- distribution of the log10 of the p-values (K-S, each cohort compared to the bulk), for the FDR we use the current cohort versus the rest as the label
- distribution of median deviation
- distribution of mean, max, min
- distribution of correlation values between PCA1, PCA2, PCA3
- plots of (PCA1, PCA2, PCA3), colored by cohort and by target.
- plots of (UMAP1, UMAP2, UMAP3), colored by cohort and by target.
- clustering of (sample, sample) similarity (HDBSCAN, AP, MC)
- differential expression

2.1 Batch wise normalisation

Location and scale adjustment (L/S):

$$\text{Standard} \quad \mathbf{x}_k^* = \frac{\mathbf{x}_k - \overline{\mathbf{x}}_k}{\sigma_k} + \overline{\mathbf{x}}_k, \quad \forall k \in \mathcal{C} \tag{1}$$

In literate this approach might be referred to as *standardisation*.

From Wang et al. [**?**]: Quantile normalisation replaces the signal intensity of a probe with the mean intensity of the probes that have the same rank from all studied arrays, and thus makes the distribution of probe intensities from each array the same. We will perform this normalisation on all samples, and per cohort.

Methods: QN (R, (methy)lumi), SQN (subset quantile normalisation)(R, wateRmelon), SWAN (subset-quantile within array normalisation)(R, minfi), BMIQ (beta-mixture quantile normalisation)(R, wateRmelon) Smoothed-QN QN followed by BMIQ

From Wang et al. [**?**]: Quantile normalisation replaces the signal intensity of a probe with the mean intensity of the probes that have the same rank from all studied arrays, and thus makes the distribution of probe intensities from each array the same. We will perform this normalisation on all samples, and per cohort. Other methods are : peak-based correction (PBC), implemented R (wateRmelon/ima/nimbl).
ComBat, Bayesian based → use 1library, part of Bioconductor's sva package.
BEclear, part of Bioconductor's BEclear package.
Functional normalisation, part of Bioconductor's minfi package.
Alternatively: Concordant bias detection, MANCIE, combining CNV data with expression data.

2.2 Empirical Bayes

3 Measurement group bias correction

To get rid of bias introduced by demographic variations within the cohorts we ideally have a large independent data set that relates genetic expression data to a wide range of demographic categories, such that research into demographic dependency of genetic measurement data is structurally open sourced and applied as common bench marks, see e.g. Viñuela et al[**?**].

References