1 Preprocessing

RNA expression:

- (between-array normalisation) L/S adjustment: no inclusion of covariants
- (between-array normalisation) ComBat (L/S plus empirical Bayes): inclusion of covariants using empirical Bayes

Methylation:

- (between-array normalisation) ComBat, BEclear
- (probewise normalisation) Infinium I/II correction, color-bias adjustment, background correction

The TCGA methylation data contains pre-calculated Beta-values, where it is unknown if dye bias and background noise is removed. To apply quantile normalisation we should convert the Beta-values to Methylation values. We then apply QN per type, for methylated and unmethylated, after which we re-establish the Beta-values. The so-called M-values are determined as follows:

$$M = log2\left(\frac{\beta}{1-\beta}\right), \quad \beta = \frac{2^M}{2^M+1}, \tag{1}$$

where the methylation is also a direct representation of the amount of methylated versus unmethylated probes according to

$$M = log2\left(\frac{max(\gamma_{meth}, 0) + \alpha}{max(\gamma_{unmeth}, 0) + \alpha}\right), \tag{2}$$

where γ is the probe intensity and α (usually set to 1) is some offset to prevent spurious behavior in case of small γ.

Benefits of using M-values as opposed to Beta-values are

- more homoscedastic than Beta-value distributions, i.e. more suitable for application of techniques like ANOVA to compare the variances between samples, see e.g. Du et al.[**?**]
-

2 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
- Bhattacharyya coefficient

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
- ANOVA and Kruskall-Wallis to check for significantly different distributions.

Is between-array normalisation even appropriate given the large variation of biological groups between the cohorts? According to Dedeurwaerder et al.[**?**] when looking differential expressions, the loss in signal is not justified by any benefit that a between-array normalisation can incur. From the same paper it is concluded that RNA expression correction methods cannot be used as-is to methylation data. The problem being that the different infinium types I and II and the colors within type I are measured on different channels. Also, according to Dedeurwaerder et al. batch effects can generate artifacts that only affect a subset of probes and which thus cannot be corrected with global correction methods. A noted exception is the ComBat[**?**] method as demonstrated in Leek et al.[**?**] and Sun et al.[**?**].

Within-array normalisation may be justified by the notion that the genetic data may differ over the target variable in terms of the relative importance of type I/II red/green probe values. According to Dedeurwaerder et al.[**?**] a Type I/II correction offers the most significant improvement. Hicks and Irizarry[**?**] also question the use of within-array normalisation depending on the nature of the global/between group variation.

In subsequent papers new methods (other than ComBat) are introduced that do preserve differential expression whilst succesfully removing unwanted bias, methods such as functional normalisation and BEclear.

One obvious way to avoid the need for probewise normalisation is too separate the datasets per distinct group: Type I color red, Type I color green and Type II, standardize per dataset, subsequently apply a cohort-bias removal method such as ComBat or BEclear.

Only apply probewise normalisation to sample groups with little variation (if at all)!

We can use R-quantro to verify? Preliminaries:

- what is the common difference in β-value between genes?
- how are the biological groups distributed over the cohorts/samples
- how different are the samples from eachother (K-S, MW-U)
- what are distances between the distributions and between the batches: Mallow's distance (1st Wasserstein metric)
- what is distribution of the $\Delta\beta$'s over the target groups (cancer-type in this case)
- what is per cohort the distribution of median $beta's$ and $\Delta\beta$
- how are the PC's clustered per batch/target?

3 Cohort bias removal

The following bias removal methods are applied

- RNA expression data: L/S adjustment & cohort based QN

- Methylation data: ComBat with the covariates gender, age and cigarette consumption

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

The basic observation we should be able to make is the following: prior to cohort-bias correction the cohort-based clusters should be distinctly seperated, and the target based clusters should be distinctly seperated as well. After the CBC the cohort-based clusters should be significantly more similar.

For the patient-based clustering we should see an increasing seperation of the different patient groups after the CBC based on the different target values.

3.1 Batch wise normalisation

Location and scale adjustment (L/S):

$$\text{Standard} \quad \mathbf{x}_k^* = \frac{\mathbf{x}_k - \overline{\mathbf{x}}_k}{\sigma_k} + \overline{\mathbf{x}}_k, \quad \forall k \in \mathcal{C} \tag{3}$$

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

ComBat, Bayesian based → use 1ibrary, part of Bioconductor's sva package. ComBat is a supervised batch effect removal method that requires the explicit input of batches and covariants.
BEclear, K-S to detect bias-affected batches followed by matrix factorisation techniques to replace suspected batch affected genes in those batches. Downside of BEclear is that it does not take care of the co-variances, upside is that it only applies batch correction to the genes/probes that seem to have a batch effect.

Alternatively: Concordant bias detection, MANCIE, combining CNV data with expression data.

To reduce the effect of collinearity we remove all samples that are correlated more than 99% with any other sample, we also remove all NaN probe's.

How are the targets distributed over the batches? How do the phenotypical covariants vary within the cohorts and between the cohorts?

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[**?**].

3.2 Measurement group bias correction

Methods: QN (R, (methy)lumi), SQN (subset quantile normalisation)(R, wateRmelon), SmoothedQN, SWAN (subset-quantile within array normalisation)(R, minfi), BMIQ (beta-mixture quantile normalisation)(R, wateRmelon) Smoothed-QN QN followed by BMIQ BEclear, part of Bioconductor's BEclear package.
Functional normalisation, part of Bioconductor's minfi package, function: preprocessFunnorm. An unsupervised normalisation method that uses negative control probes. This method is explicitly designed for 450k methylation data although the idea of negative control probes can be generalised.

peak-based correction (PBC), implemented R (wateRmelon/ima/nimbl).

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 are explicitly interested in variance between the target groups, hence we are fine with probewise bias as long as it is roughly stratified over the target groups. As an alternative to probewise normalisation it is wiser to simply split the dataset in seperate datasets per probewise group.

4 Methylation plus RNA expression

We have about 60.000 RNA expression values, and about 400.000 methylation values, per sample. To do a full correlation scan of all combinations we need to perform $60.000 \times 400.000 \times 1000$ computations, or more specifically, we need to perform $60.000 \times 400.000 = 24\,10^9$ in-products on vectors with length $\propto 1000$.

To make this tractable we can be selective in the gene's by considering the target variable at hand (say the type of cancer) and only select the gene's or probe values that seperate the target variables the best based on some non-parametric distribution comparison such as Kolmogorov-Smirnof or Mann-Whitney U, or we apply a dimension reduction on both data sets and only directly compare the top components per datasets.

The caveat with all these approaches is the bias we introduce by considering only the individually strong components per data set, The only immediate approach at hand to find the strong combinations of components is a brute-force approach? Another approach we might try is relatively straightforward: we simply append the components to eachother and apply a dimension reduction based on the variance (PCA) or the separation (LDA) after we can reconstruct what components co-occur in the reduced dimensions.

$(1000, 60.000), (1000, 400.000) \rightarrow (1000, 460.000)$ The most biased but biologically sensible approach is the paired combination of methylation and RNA expression data based on the affected genes.

5 Gene/probe distribution comparisons per target

targets are: cancer type, healthy/cancerous tissue, survived/diseased, etc.

References