Breast cancer is an umbrella term for encompassing multiple different subtypes, with each subtype having a different biological and clinical features, progression pattern, treatment response and prognosis. Stratification of breast cancer is an important objective, which allows clinicians to better evaluate patient risk and select effective therapeutic strategies based on the subtype in question (Dai et al., 2015).

Earlier, breast cancer stratification was conventionally performed using a combination of immunohistochemistry (IHC) detection for cellular markers as well as anatomical features such as tumour node size, number of metastases detected, etc. However, information from gene expression profiling by microarrays became an increasingly important means of breast cancer classification (Dai et al., 2015).

It is a well-known issue that there is little consensus in data mined between different omics studies, including microarray gene expression studies, where lack of reproducibility has been brought up as an issue (Draghici et al., 2006; Shi et al., 2008; Sweeney et al., 2017). There could be multiple reasons for this lack of reproducibility. Firstly, the diversity of available microarray platforms with different manufacturing strategies, probe sequences, versions and measurement methods (for example, one channel versus two-channel microarrays), makes it difficult to compare differentially-expressed genes between studies using different platforms. Secondly, technical variation generated during the data generation and collection process introduces unwanted noise into the data, compromising the accuracy, reproducibility and generalizability of the results obtained from that data. An example of such variation are batch effects (Goh et al., 2017). Since such technical variability is dataset-specific, it leads to different datasets being affected by batch effects to different degrees, leading to lack of consistency in the obtained results. Furthermore, even though batch-effect correction algorithms such as ComBat (Johnson et al., 2007) and surrogate variable analysis (SVA) (Leek & Storey, 2007) exist, if they are incorrectly applied, or if used on datasets with unbalanced distribution of batches across classes, they might inadvertently remove relevant variation or add irrelevant variation into the dataset and thereby compromise reproducibility. Thirdly, the properties of certain commonly-used pre-processing methods may introduce biases into the data. For example, the often used robust multichip average (RMA) method of pre-processing data has been shown to dilute biological signal during the quantile normalization (Kim et al., 2014; Zhao et al., 2020) stage as well as during its summarization stage via median polish (Kim et al., 2014).

A fourth reason for lack of reproducibility among different datasets is that compositional differences in the underlying subpopulations comprising the datasets may cause differences in the results generated by each dataset. For example, imagine that we are investigating differentially-expressed genes (DEGs) between diseased and healthy populations. Suppose that the disease under study has two distinct subtypes. Also imagine that the diseased population in dataset 1 contained 70% samples from subtype 1 and 20% samples from subtype 2, while that in dataset 2 contained 20% from subtype 1 and 80% from subtype 2. In this case, the DEGs from dataset 1 would be dominated by those differentiating subtype 1 samples from the healthy samples, while those from dataset 2 would be dominated by those differentiating subtype 2 samples from dataset 2. If the two subtypes are sufficiently different in terms of gene expression profiles and underlying pathological pathways, then there would be very little overlap between the DEG lists obtained from the two datasets. Compositional differences may also arise from factors such as age, gender or race.

There have been several methods proposed to rescue reproducibility from all these forces adversely affecting it. For instance, to overcome the problem of erasure of biological signal due to global quantile normalization, Zhao et al. (2020)

# Bibliography

Dai, X., Li, T., Bai, Z., Yang, Y., Liu, X., Zhan, J., & Shi, B. (2015). Breast cancer intrinsic subtype classification, clinical use and future trends. *American Journal of Cancer Research*, *5*(10), 2929–2943. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4656721/

Draghici, S., Khatri, P., Eklund, A. C., & Szallasi, Z. (2006). Reliability and reproducibility issues in DNA microarray measurements. *Trends in Genetics : TIG*, *22*(2), 101–109. https://doi.org/10.1016/j.tig.2005.12.005

Goh, W. W. B., Wang, W., & Wong, L. (2017). Why Batch Effects Matter in Omics Data, and How to Avoid Them. *Trends in Biotechnology*, *35*(6), 498–507. https://doi.org/10.1016/j.tibtech.2017.02.012

Huang, S., Murphy, L., & Xu, W. (2018). Genes and functions from breast cancer signatures. *BMC Cancer*, *18*(1), 473. https://doi.org/10.1186/s12885-018-4388-4

Johnson, W. E., Li, C., & Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, *8*(1), 118–127. https://doi.org/10.1093/biostatistics/kxj037

Kim, C. S., Hwang, S., & Zhang, S.-D. (2014). RMA with quantile normalization mixes biological signals between different sample groups in microarray data analysis. *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 139–143. https://doi.org/10.1109/BIBM.2014.6999142

Leek, J. T., & Storey, J. D. (2007). Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PLoS Genetics*, *3*(9), e161. https://doi.org/10.1371/journal.pgen.0030161

Shi, L., Jones, W. D., Jensen, R. V., Harris, S. C., Perkins, R. G., Goodsaid, F. M., Guo, L., Croner, L. J., Boysen, C., Fang, H., Qian, F., Amur, S., Bao, W., Barbacioru, C. C., Bertholet, V., Cao, X. M., Chu, T.-M., Collins, P. J., Fan, X., … Tong, W. (2008). The balance of reproducibility, sensitivity, and specificity of lists of differentially expressed genes in microarray studies. *BMC Bioinformatics*, *9*(Suppl 9), S10. https://doi.org/10.1186/1471-2105-9-S9-S10

Sweeney, T. E., Haynes, W. A., Vallania, F., Ioannidis, J. P., & Khatri, P. (2017). Methods to increase reproducibility in differential gene expression via meta-analysis. *Nucleic Acids Research*, *45*(1), e1. https://doi.org/10.1093/nar/gkw797

Zhao, Y., Wong, L., & Goh, W. W. B. (2020). How to do quantile normalization correctly for gene expression data analyses. *Scientific Reports*, *10*(1), Article 1. https://doi.org/10.1038/s41598-020-72664-6