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. Finally, another reason for lack of reproducibility is the inherent instability of the p-value metric used to select the DEG list (Halsey et al., 2015).

There have been several proposals to rescue reproducibility from all these forces adversely affecting it. For instance, to overcome the erasure of biological signal due to global quantile normalization in the presence of high class-effect proportion, Zhao et al. (2020) suggested performing quantile normalization in a class-specific manner, so as to preserve the inter-class differences. Sweeney et al. (2017) proposed pooling information from several independent studies by meta-analysis in order to obtain more reproducible lists of DEGs. Zhang et al. (2009) took into consideration correlation within DEG lists rather than just the degree of overlap between DEG lists as a metric for reproducibility. A rank-based normalization method, gene fuzzy scoring (GFS), was proposed to overcome the biases introduced by other normalization methods (Belorkar & Wong, 2016) which compromise reproducibility.

1. **METHODOLOGY**
   1. **Datasets**
      1. **GSE762752 microarray dataset from GEO**

Our main dataset is GSE76275 which is a compilation of microarray data conducted on the platform GPL570 with  Affymetrix Human Genome U133 Plus 2.0 Array. It is a superseries comprising 2 microarray submission, encompassing 198 TNBC tumours and 67 non-TNBC tumours. The dataset is described in the accompanying publication of Burstein et al.  (2015). In this study, the authors have identified and validated 4 subtypes : Luminal-AR (LAR), Mesenchymal (MES), Basal-Like Immune-Suppressed (BLIS), and Basal-Like Immune-Activated (BLIA). Metadata for this dataset was obtained from the GEOquery Bioconductor R package (Davis & Meltzer, 2007). Progesterone receptor (PR) status, estrogen receptor (ER) status, and HER2 amplification status were also provided in the metadata. Raw Affymetrix CEL files were downloaded from the corresponding accession page on GEO.

1. **GSE43358** **microarray dataset from GEO**

Our analyses on the main dataset were subsequently repeated on our validation dataset (GSE43358) comprising microarray-derived gene expression data from a smaller set of 57 samples utilising the same platform and chip. This dataset consisted of 17 TNBC samples and 40 non-TNBC samples. Similar to the original dataset, the metadata was obtained using GEOquery, and the raw CEL files were downloaded from the GEO accession page. PR status, ER status, and HER2 amplification status were also provided in the metadata for this dataset.

Figure 1 below describes our project workflow. In brief, for each dataset, the expression data from the raw CEL files was processed according to the selected pre-processing method, either class-specific quantile normalization (method 1) or GFS (method 2). More details regarding these pre-processing methods are given in the subsequent sections. Following pre-processing, the differential gene expression analysis was carried out. Limma was used for class-specific quantile-normalized data, while the Mann-Whitney U test was performed for the GFS-normalized data. The differentially expressed gene list was obtained for each pre-processing method, for each dataset. The DEG lists from the two datasets were intersected to obtain the percentage of overlapping genes (POG) as defined in (Zhang et al., 2009).

The RMA method consists of three steps: (1) Background subtraction, (2) Quantile normalization and (3) Summarization (Irizarry et al., 2003). In method 1, since quantile normalization was to be carried out separately for the two classes, the raw CEL files for the TNBC and non-TNBC samples were read in separately using the oligo package (Carvalho & Irizarry, 2010).

# Bibliography

Belorkar, A., & Wong, L. (2016). GFS: Fuzzy preprocessing for effective gene expression analysis. *BMC Bioinformatics*, *17*(Suppl 17), 540. https://doi.org/10.1186/s12859-016-1327-8

Carvalho, B. S., & Irizarry, R. A. (2010). A framework for oligonucleotide microarray preprocessing. *Bioinformatics*, *26*(19), 2363–2367. https://doi.org/10.1093/bioinformatics/btq431

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/

Davis, S., & Meltzer, P. S. (2007). GEOquery: A bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics (Oxford, England)*, *23*(14), 1846–1847. https://doi.org/10.1093/bioinformatics/btm254

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

Halsey, L. G., Curran-Everett, D., Vowler, S. L., & Drummond, G. B. (2015). The fickle P value generates irreproducible results. *Nature Methods*, *12*(3), Article 3. https://doi.org/10.1038/nmeth.3288

Irizarry, R. A., Hobbs, B., Collin, F., Beazer‐Barclay, Y. D., Antonellis, K. J., Scherf, U., & Speed, T. P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, *4*(2), 249–264. https://doi.org/10.1093/biostatistics/4.2.249

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

Zhang, M., Zhang, L., Zou, J., Yao, C., Xiao, H., Liu, Q., Wang, J., Wang, D., Wang, C., & Guo, Z. (2009). Evaluating reproducibility of differential expression discoveries in microarray studies by considering correlated molecular changes. *Bioinformatics*, *25*(13), 1662–1668. https://doi.org/10.1093/bioinformatics/btp295

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