AIMS: Absolute Assignment of Breast Cancer Intrinsic Molecular Subtype

Eric R. Paquet (eric.r.paquet@gmail.com), Michael T. Hallett (michael.t.hallett@mcgill.ca)\textsuperscript{1}

\textsuperscript{1}Department of Biochemistry, Breast cancer informatics, McGill University, Montreal, Canada

October 24, 2023

Contents

1 Introduction 2

2 Case Study: Assigning breast cancer subtype to a dataset of breast cancer microarray data 3

3 Session Info 8
1 Introduction

Technologies to measure gene expression in a massively parallel fashion have facilitated a more complete molecular understanding of breast cancer (BC) and a deeper appreciation for the heterogeneity of this disease[1, 2, 3]. Although the subtypes defined by Estrogen Receptor (ER) and Human Epidermal growth factor Receptor 2 (Her2) status have long been recognized as distinct forms of the disease, early genomic studies emphasized the magnitude of the differences between the subtypes at the molecular level [4, 5]. Unbiased bioinformatic analyses of expression profiles provided the so-called intrinsic subtyping scheme, consisting of the Luminal A (LumA), Luminal B (LumB) and Normal-like (NormL) subtypes enriched for ER+ tumors, the Her2-enriched (Her2E) subtype containing many Her2+ tumors and the Basal-like (BasalL) subtype enriched for ER-/Her2- tumors [1, 2]. The subtypes have differing clinicopathological attributes, prognostic characteristics, and treatment options, providing sufficient clinical utility as to support their inclusion within international guidelines for treatment[6]. In fact, the ProsignaTM PAM50-based risk of recurrence (ROR) score generated from the intrinsic subtypes has recently received FDA approved for clinical use [7]. Several alternative omics-based subtyping schemes promise similar utility[8, 9, 10, 11].

Subtyping tools built in the manner of PAM50 have severe shortcomings[12, 13, 14]. The application of normalization procedures and gene-centering techniques are necessary to remove batch effects and insure comparability between expression levels of genes across patients[12, 13]. However, different normalization procedures have been shown to lead to different subtype classifications for patients[12, 13, 14, 15]. In lieu of absolute estimations of mRNA copy number per gene per individual, gene-centering techniques are used to adjust expression measurements to represent the change in abundance of a specific mRNA species relative to the cohort of patients. However, these steps make such methods sensitive to the composition of patients in the dataset. Both differences in the ratio of ER+ to ER- samples and differences in the frequency of each intrinsic subtype within a dataset have been shown to influence subtype assignments[12]. It remains unknown as to the degree of imbalance necessary to cause such “subtype instability”.

We present a bioinformatics approach entitled Absolute Intrinsic Molecular Subtyping AIMS for estimating patient subtype that circumvents these shortcomings. The method does not require a panel of gene expression samples. As such, it is the first method that can accurately assign subtype to a single patient whilst being insensitive to changes in normalization procedures or the relative frequencies of ER+ tumors, of subtypes or of other clinicopathological patient attributes. The stability and accuracy of AIMS is explored for the intrinsic subtyping scheme across several datasets generated via microarrays or RNA-Seq. More detail about AIMS could be found in Paquet et al. (in review at the Journal of the National Cancer Institute).

The AIMS package is providing the necessary functions to assign the five intrinsic subtypes of breast cancer to either a single gene expression experiment or to a dataset of gene expression data [16].
2 Case Study: Assigning breast cancer subtype to a dataset of breast cancer microarray data

We first need to load the package \textit{AIMS} and our example dataset. In this case study we will use a fraction of the McGill dataset describe in the paper and also \textit{breastCancerVDX}.

\begin{verbatim}
> library(AIMS)
> data(mcgillExample)
\end{verbatim}

To get breast cancer subtypes for a dataset we need to provide the expression values that have not been gene centered. It means all the expression values will be positive. In the case of a two-colors array the user should select only the channel that contains the tumor sample (usually the Cy5 channel).

In the previous code we have shown the size of the expression matrix for the McGill example as well as the first expression values and the characters vector of Entrez ids. \textit{AIMS} require the use of Entrez ids to prevent any confusion related to unstable gene symbols.

\begin{verbatim}
> mcgill.subtypes <- applyAIMS(mcgillExample$D, 
+ mcgillExample$EntrezID)
> names(mcgill.subtypes)

[1] "cl"       "prob"      "all.probs"   "rules.matrix"
[5] "data.used" "EntrezID.used"

> head(mcgill.subtypes$cl)

  20 1 "Normal"
  2 "LumA"
  3 "LumB"
  4 "Her2"
  5 "LumA"
  6 "LumB"

> head(mcgill.subtypes$prob)

     20          1 0.7784899
     1 1.0000000 0.9929477
     4 0.9722196 0.9998480
     5 0.9533438
applyAIMS is the function used to assign AIMS to both single sample and dataset of gene expression data. The first parameter should be a numerical matrix composed of positive-only values. Rows represent genes and columns samples. The second argument represents the EntrezIds corresponding to genes in the first parameter. AIMS will deal with duplicated EntrezId so you should leave them there.

applyAIMS will return a list of arguments. cl represents the molecular assignment in (Basal, Her2, LumA (Luminal A), LumB (Luminal B) and Normal). The variable prob corresponds to a matrix with five columns and number of rows corresponding to the number of samples in D. This matrix contains the posterior probabilities returned from the Naive Bayes classifier.

```r
> mcgill.first.sample.subtype <- applyAIMS(mcgillExample$D[,1,drop=FALSE],
+ mcgillExample$EntrezID)
> names(mcgill.first.sample.subtype)
[1] "cl"      "prob"     "all.probs" "rules.matrix"
[5] "data.used" "EntrezID.used"
> head(mcgill.first.sample.subtype$cl)
  20
  1 "Normal"
> head(mcgill.first.sample.subtype$prob)
  20
  1 0.7784899
> table(mcgill.first.sample.subtype$cl)

Normal
  1

This is the same example as before except now we are assigning subtype to only one sample.
Here we are assigning AIMS subtypes on the vdx dataset. We are using (hgu133a.db) to obtain EntrezIds corresponding to the probes of the HG-133A array. AIMS has been designed to be applicable on this platform.
References


3 Session Info

- R version 4.3.1 (2023-06-16), x86_64-pc-linux-gnu
- **Locale:** LC_CTYPE=en_US.UTF-8, LC_NUMERIC=C, LC_TIME=en_GB, LC_COLLATE=C, LC_MONETARY=en_US.UTF-8, LC_MESSAGES=en_US.UTF-8, LC_PAPER=en_US.UTF-8, LC_NAME=C, LC_ADDRESS=C, LC_TELEPHONE=C, LC_MEASUREMENT=en_US.UTF-8, LC_IDENTIFICATION=C
- **Time zone:** America/New_York
- **TZcode source:** system (glibc)
- **Running under:** Ubuntu 22.04.3 LTS
- **Matrix products:** default
- **BLAS:** /home/biocbuild/bbs-3.18-bioc/R/lib/libRblas.so
- **LAPACK:** /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.10.0
- **Base packages:** base, datasets, grDevices, graphics, methods, stats, stats4, utils
- **Other packages:** AIMS 1.34.0, AnnotationDbi 1.64.0, Biobase 2.62.0, BiocGenerics 0.48.0, IRanges 2.36.0, S4Vectors 0.40.0, breastCancerVDX 1.39.0, e1071 1.7-13, hgu133a.db 3.13.0, org.Hs.eg.db 3.18.0
- **Loaded via a namespace (and not attached):** Biostrings 2.70.0, DBI 1.1.3, GenomeInfoDb 1.38.0, GenomeInfoDbData 1.2.11, KEGGREST 1.42.0, R6 2.5.1, RCurl 1.98-1.12, RSQLite 2.3.1, XVector 0.42.0, bit 4.0.5, bit64 4.0.5, bitops 1.0-7, blob 1.2.4, cachem 1.0.8, class 7.3-22, cli 3.6.1, compiler 4.3.1, crayon 1.5.2, fastmap 1.1.1, http 1.4.7, memoise 2.0.1, pkgconfig 2.0.3, png 0.1-8, proxy 0.4-27, rlang 1.1.1, tools 4.3.1, vctrs 0.6.4, zlibbioc 1.48.0