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)[1]

[1]*Department of Biochemistry, Breast cancer informatics, McGill University, Montreal, Canada*

April 15, 2025

Contents

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 *AIMS* and our example dataset. In this case study we will use a fraction of the McGill dataset describe in the paper and also *breastCancerVDX*.

```
> library(AIMS)
> data(mcgillExample)
```

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. *AIMS* require the use of Entrez ids to prevent any confusion related to unstable gene symbols.

```
> 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
2 1.0000000
3 0.9929477
4 0.9722196
5 0.9998480
6 0.9533438
```

```
> table(mcgill.subtypes$cl)

 Basal   Her2   LumA   LumB Normal
    52     65     98     54     52
```

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.

```
> 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.

```
> library(breastCancerVDX)
> library(hgu133a.db)
> data(vdx)
> hgu133a.entrez <- as.character(as.list(hgu133aENTREZID)[featureNames(vdx)])
> vdx.subtypes <- applyAIMS(vdx,
+                           hgu133a.entrez)
> names(vdx.subtypes)

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

> head(vdx.subtypes$cl)

      20
VDX_3 "Normal"
VDX_5 "Her2"
VDX_6 "Basal"
VDX_7 "Basal"
VDX_8 "LumB"
VDX_9 "Her2"

> head(vdx.subtypes$prob)

             20
VDX_3 1.0000000
VDX_5 0.6131510
VDX_6 1.0000000
VDX_7 1.0000000
VDX_8 1.0000000
VDX_9 0.5376401

> table(vdx.subtypes$cl)

 Basal   Her2   LumA   LumB Normal
    97     67     67     86     27
```

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

[1] Perou CM, Sorlie T, Eisen MB, et al. Molecular portraits of human breast tumours. Nature 2000;406(6797):747-52.

[2] Sorlie T, Perou CM, Tibshirani R, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A 2001;98(19):10869-74.

[3] Weigelt B, Baehner FL, Reis-Filho JS. The contribution of gene expression profiling to breast cancer classification, prognostication and prediction: a retrospective of the last decade. J Pathol 2010;220(2):263-80.

[4] Gruvberger S, Ringner M, Chen Y, et al. Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. Cancer Res 2001;61(16):5979-84.

[5] Pusztai L, Ayers M, Stec J, et al. Gene expression profiles obtained from fine-needle aspirations of breast cancer reliably identify routine prognostic markers and reveal large-scale molecular differences between estrogen-negative and estrogen-positive tumors. Clin Cancer Res 2003;9(7):2406-15.

[6] Goldhirsch A, Wood WC, Coates AS, et al. Strategies for subtypes–dealing with the diversity of breast cancer: highlights of the St. Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2011. Ann Oncol 2011;22(8):1736-47.

[7] Harbeck N, Sotlar K, Wuerstlein R, et al. Molecular and protein markers for clinical decision making in breast cancer: Today and tomorrow. Cancer Treat Rev 2013; 10.1016/j.ctrv.2013.09.014.

[8] Guedj M, Marisa L, de Reynies A, et al. A refined molecular taxonomy of breast cancer. Oncogene 2012;31(9):1196-206.

[9] Lehmann BD, Bauer JA, Chen X, et al. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. J Clin Invest 2011;121(7):2750-67.

[10] Jonsson G, Staaf J, Vallon-Christersson J, et al. Genomic subtypes of breast cancer identified by array-comparative genomic hybridization display distinct molecular and clinical characteristics. Breast Cancer Res 2010;12(3):R42.

[11] Curtis C, Shah SP, Chin SF, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature 2012;486(7403):346-52.

[12] Lusa L, McShane LM, Reid JF, et al. Challenges in projecting clustering results across gene expression-profiling datasets. J Natl Cancer Inst 2007;99(22):1715-23.

[13] Sorlie T, Borgan E, Myhre S, et al. The importance of gene-centring microarray data. Lancet Oncol 2010;11(8):719-20; author reply 720-1.

[14] Weigelt B, Mackay A, A'Hern R, et al. Breast cancer molecular profiling with single sample predictors: a retrospective analysis. Lancet Oncol 2010;11(4):339-49.

[15] Perou CM, Parker JS, Prat A, et al. Clinical implementation of the intrinsic subtypes of breast cancer. Lancet Oncol 2010;11(8):718-9; author reply 720-1.

[16] Parker JS, Mullins M, Cheang MCU, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. Journal of clinical oncology 2009;27:1160-7.

3 Session Info

- R version 4.5.0 RC (2025-04-04 r88126), `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 24.04.2 LTS`
- Matrix products: default
- BLAS: `/home/biocbuild/bbs-3.21-bioc/R/lib/libRblas.so`
- LAPACK: `/usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.12.0`
- Base packages: base, datasets, grDevices, graphics, methods, stats, stats4, utils
- Other packages: AIMS 1.40.0, AnnotationDbi 1.70.0, Biobase 2.68.0, BiocGenerics 0.54.0, IRanges 2.42.0, S4Vectors 0.46.0, breastCancerVDX 1.45.0, e1071 1.7-16, generics 0.1.3, hgu133a.db 3.13.0, org.Hs.eg.db 3.21.0
- Loaded via a namespace (and not attached): Biostrings 2.76.0, DBI 1.2.3, GenomeInfoDb 1.44.0, GenomeInfoDbData 1.2.14, KEGGREST 1.48.0, R6 2.6.1, RSQLite 2.3.9, UCSC.utils 1.4.0, XVector 0.48.0, bit 4.6.0, bit64 4.6.0-1, blob 1.2.4, cachem 1.1.0, class 7.3-23, cli 3.6.4, compiler 4.5.0, crayon 1.5.3, fastmap 1.2.0, httr 1.4.7, jsonlite 2.0.0, memoise 2.0.1, pkgconfig 2.0.3, png 0.1-8, proxy 0.4-27, rlang 1.1.6, tools 4.5.0, vctrs 0.6.5