Comparison of unsupervised machine-learning methods to identify metabolomic signatures in patients with localized breast cancer

Jocelyn Gal a,1,⇑, Caroline Bailleux b,1, David Chardin c,d,1, Thierry Pourcher d, Julia Gilhodes e, Lun Jing d, Jean-Marie Guigonis d, Jean-Marc Ferrero b, Gerard Milano f, Baharia Mograbig, Patrick Brest g, Yann Chateaua, Olivier Humberta,d, Emmanuel Chamoreya

a University Côte d'Azur, Epidemiology and Biostatistics Department, Centre Antoine Lacassagne, Nice F-06189, France
b University Côte d'Azur, Medical Oncology Department Centre Antoine Lacassagne, Nice F-06189, France
c University Côte d'Azur, Nuclear Medicine Department, Centre Antoine Lacassagne, Nice F-06189, France
d University Côte d'Azur, Commissariat à L'Energie Atomique, Institut de Biosciences et Biotechnologies d'Aix-Marseille, Laboratory Transporters in Imaging and Radiotherapy in Oncology, Faculty of Medicine, Nice F-06100, France
e University Côte d'Azur, International Laboratory of Cellular and Molecular Biology, Department of Biostatistics, Institut Claudius Regaud, IUCT-O Toulouse, France
f University Côte d'Azur, Centre Antoine Lacassagne, Oncopharmacology Unit, Nice F-06189, France
g University Côte d'Azur, CNRS UMR7284, INSERM U1081, IRCAN TEAM4 Centre Antoine Lacassagne FHU-Oncoage, Nice F-06189, France

Article info

Article history:
Received 11 February 2020
Received in revised form 15 May 2020
Accepted 16 May 2020
Available online 3 June 2020

Keywords:
Unsupervised machine learning
Metabolomics
Breast neoplasms
Computer simulation

Abstract

Genomics and transcriptomics have led to the widely-used molecular classification of breast cancer (BC). However, heterogeneous biological behaviors persist within breast cancer subtypes. Metabolomics is a rapidly-expanding field of study dedicated to cellular metabolisms affected by the environment. The aim of this study was to compare metabolomic signatures of BC obtained by 5 different unsupervised machine learning (ML) methods. Fifty-two consecutive patients with BC with an indication for adjuvant chemotherapy between 2013 and 2016 were retrospectively included. We performed metabolomic profiling of tumor resection samples using liquid chromatography-mass spectrometry. Here, four hundred forty-nine identified metabolites were selected for further analysis. Clusters obtained using 5 unsupervised ML methods (PCA k-means, sparse k-means, spectral clustering, SIMLR and k-sparse) were compared in terms of clinical and biological characteristics. With an optimal partitioning parameter k = 3, the five methods identified three prognosis groups of patients (favorable, intermediate, unfavorable) with different clinical and biological profiles. SIMLR and K-sparse methods were the most effective techniques in terms of clustering.

In-silico survival analysis revealed a significant difference for 5-year predicted OS between the 3 clusters. Further pathway analysis using the 449 selected metabolites showed significant differences in amino acid and glucose metabolism between BC histologic subtypes. Our results provide proof-of-concept for the use of unsupervised ML metabolomics enabling stratification and personalized management of BC patients. The design of novel computational methods incorporating ML and bioinformatics techniques should make available tools particularly suited to improving the outcome of cancer treatment and reducing cancer-related mortalities.

© 2020 The Authors. Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Breast cancer (BC) is the most common type of cancer in women worldwide and the second leading cause of cancer-associated deaths [1]. The treatment strategy may be guided by two classifications indicating the aggressiveness of the tumor. The anatomic-clinical classification is based on age, TNM, histological factors (histological grade, Ki-67) as well as on hormonal-receptor status and Her-2 expression. The molecular classification resulting from genomic [2], transcriptomic [3] and proteomic [4] analyses introduced the concept of luminal A, luminal B, Her-2 and basal-like BC [5–7]. This latter classification from Perou and Sorlie was assessed using unsupervised analyses [6,8]. Efforts have been made to develop multivariate prognostic models such as, AdjuvantOnline®,
Several genomic prognostic markers have been described for BC such as OncotypeDX™, Prosigna®, MammaPrint®, Endopredict® Genomic grade index® and BC Index® [19]. Two markers are commercially available and are increasingly used in clinical practice (21-gene recurrence score OncotypeDX® and 70-gene prognostic signature MammaPrint®). However, heterogeneity persists in biological features within BC subtypes, thus highlighting the need to improve the taxonomy [20]. This heterogeneity may be related to specific combinations of genetic, pathological and environmental factors leading to specific metabolic alterations and interactions [21,22].

Metabolomics is a new and growing field dedicated to the study of metabolism at overall level that promises to provide new insights into disease mechanisms and drug effects. Indeed, metabolomics may offer a complementary approach to genomics and could be used to better understand the influence of the environment on tumor phenotype [23]. Two distinct approaches characterize metabolomics: a targeted approach aimed at quantifying as accurately as possible a limited number of predefined metabolites of interest [24] and an untargeted approach aimed at measuring, without any a priori, as many metabolites as possible in a sample [25,26]. As with other omics approaches, metabolomics generates high-dimensional data. The processing of these data can be done by applying supervised or unsupervised machine learning (ML) algorithms that are increasingly used for medical diagnosis and therapeutic strategy guidance [27–29]. Unsupervised ML, in which no a priori class label information is given to guide the algorithm [30], seems a suitable alternative to analyze these data and address the problem of BC heterogeneity [6]. The aim of this study was to compare metabolomic signatures of BC obtained using five different unsupervised ML methods. To evaluate the consistency of our results, the clusters obtained by unsupervised ML methods were compared with patients’ clinical characteristics and identified metabolic pathways.

2. Material and methods

2.1. Patients

This is a retrospective cohort study based on data and samples from 52 patients already available in the Centre Antoine Lacassagne tumor bank and collected during routine practice between 2013 and 2016. Patient tumor characteristics were: clinical stages of disease, hormone receptor status, Her-2 status and tumor grade. Follow-up data were retrospectively extracted from our facility’s database sources [34]. The impact score was determined by the relative pathway topological effect of the metabolites, and -log(p) was used as the enrichment score, reflecting the probability of the pathway being identified at random; the number of “hits” was the actual number of matched metabolites in the pathway. For the selection of the most relevant pathways, we applied the following criteria: Impact >0, FDR < 0.25 and p < 0.05 [35].

A Venn diagram (http://bioinformatics.psb.ugent.be/webtools/Venn/) was used to display all possible logical relations between the metabolites or pathways identified by the clustering methods. Differences between clusters regarding the most active metabolites were plotted using boxplots.

2.3. Clustering algorithms

Five unsupervised clustering methods were selected and compared: Principal Component Analysis (PCA) k-means, Sparse k-means, Single-cell Interpretation via Multi-kernel LeaRning (SIMLR), K-sparse and Spectral clustering. Many clustering approaches exist, among which two of the most popular are K-means and spectral clustering [36]. PCA k-means and Sparse k-means are two well established, K-means based methods frequently used in computational. SIMLR and K-sparse are two recently developed k-means based methods of particular interest for omics data. These methods use different dimension reduction steps with k-means. In order to apply these five unsupervised clustering methods, the optimal number of clusters was determined in advance using five criteria: gap [37], silhouette [38,39], Davies-Bouldin [40], Calinski-Harabasz [41] and SIMLR method [42]. PCA k-means clustering, combines PCA to reduce the number of dimensions of a dataset and the k-means method to minimize the intra-cluster variance for a chosen number of k clusters [43–45]. Spectral clustering [46,47] is based on graph theory. It consists of identifying dense regions in a multidimensional dataset, i.e. observations that can form a non-convex set but are close to each other. Sparse k-means clustering was developed in 2010 by Witten and Tibshirani [8]. This method is based on a Least Absolute Shrinkage and Selection Operator (LASSO) approach [48] and combines the LASSO approach and the k-means method which simultaneously find the clusters and select features. SIMLR clustering [42] was developed to analyze scRNA-seq data. This method searches for appropriate cell-to-cell similarity metrics to perform dimension reduction and clustering. In multiple-kernel learning frameworks, this
method may be especially beneficial for data containing no identifiable clusters. K-sparse clustering [49] is an algorithm combining dimension reduction and relevant feature selection using a constraint in L1-norm rather than a lasso-type penalty to select the features. The performance of an unsupervised clustering method is measured by its ability to partition data. Partitioning is considered optimal when it minimizes the average distance between patients within a cluster (homogeneity) and maximizes cluster distances 2 by 2 (separability). The performances of the five methods were compared using the silhouettes index (SI) [39]. The SI ranges between –1 and 1 and assesses whether a patient belongs to the “right” cluster. The closer the index is to 1, the more satisfactory the assignment of a patient to a cluster. The t-SNE method was used for data visualization [50]. Processing times were obtained on a computer using an i5 processor (3.1 GHz).

2.4. Clinical evaluation

The relevance of the discovered clusters was assessed by comparing the clinical and survival characteristics between clusters using 2 or Fisher’s exact tests for categorical data, analysis of variance or Mann-Whitney’s test for continuous variables and log-rank test for censored data. Overall survival (OS) was defined as the time between diagnosis and death due to any cause. Specific survival (SS) was determined by the time between diagnosis and death due to BC. Recurrence-Free Survival (RFS) was defined as the time between diagnosis and the first recurrence (local, regional and metastasis). Patients showing no event (death or recurrence) or lost to follow-up were censored at the date of their last contact. OS, SS, and RFS were estimated using the Kaplan-Meier method. Median follow-up with a 95% confidence interval was calculated for censored data. The difference between clusters regard-

| Clinical characteristic | No. of patients |
|-------------------------|-----------------|
| Age (median min - max)  | 63.2 (37–88)    |
| Histology type          |                 |
| Invasive ductal carcinoma | 48              |
| Invasive lobular carcinoma | 3              |
| Microinvasive carcinoma | 1               |
| Tumor stage             |                 |
| T1                      | 21              |
| T2                      | 24              |
| T3                      | 7               |
| Axillary lymph node status |             |
| N0                      | 28              |
| N+                      | 24              |
| Metastasis              |                 |
| M0                      | 50              |
| M1                      | 2               |
| Tumor phenotype         |                 |
| Her2                    | 12              |
| Luminal                 | 25              |
| Triple-Negative         | 15              |
| Adjuvant Chemotherapy   | 13              |
| Adjuvant Radiotherapy   | 9               |
| Adjuvant Hormonotherapy |                |

2.5. Prediction for 5- and 10-year overall and specific survival

Web-based prognostication PREDICT tool (https://breast.predict.nhs.uk/tool) [9,10,53] was used to estimate predicted OS (pOS) and predicted SS (pSS) at 5 and 10 years, based on several patient and tumor characteristics. For each patient, ten characteristics were entered manually: age at diagnosis, menopausal status, estrogen receptor status, Her-2 status, Ki-67 status, tumor stage, histological grade, mode of detection, number of positive nodes and presence of micrometastases. PREDICT tool can be used to estimate expected overall survival at 5 years and 10 years in the absence of available survival data due to short follow-up. If information was missing for detection, bisphosphonate therapy or menopausal status, patients were not included but the “unknown” category was used. Only one patient was excluded because of missing tumor grade data. A 1000 resamples bootstrap was used to estimate the 95% confidence interval.

3. Results

3.1. Patient characteristics

Tumor and treatment features of the 52 patients were described in Table 1. Median age was 63 years (range: 37–88). The main histological type was invasive ductal carcinoma (92%), and the main tumor stages were T1 (40.5%) and T2 (46%). Twenty-four patients (46%) presented axillary lymph node invasion. Two patients (4%) were oligometastatic at diagnosis. Forty-three percent of patients had histological grade II tumors and 47% had grade III tumors. Half of the patients had negative hormone receptor status (48%) and 24% of patients had Her-2 over-expression. Median follow–up was 48.5 months (95%CI [43–54.5]). Twenty-one patients presented a recurrence: 4 local recurrences (7.5%), 6 regional recurrences (11.5%) and 11 metastatic recurrences (21%). Three-year OS was 90% [82–99], 3-year SS was 92% [85–100] and 3-year RFS was 82% [72–93] (Supplementary Fig. 2). Median OS, SS, and RFS were not reached.

3.2. Clustering results

3.2.1. Estimated number of clusters

Using four methods (Gap statistic, Calinski-Harabasz, Silhouette and SIMLR criterion), the optimal number of clusters was equal to three (k = 3) (Supplementary Fig. 3). Only for Davies-Bouldin criterion, the optimal number of clusters was equal to four (k = 4). It
seems reasonable, therefore, to conclude that the optimal number of clusters is equal to 3.

3.2.2. Patient distribution

Three clusters were identified with each of the five clustering methods. (Fig. 1). In terms of processing times, PCA k-means was the fastest and K-sparse was the longest (Supplementary Table 1). SIMLR and k-sparse methods were the most discriminants with an average silhouette value of 0.85 and 0.91, respectively (Fig. 2). Seventy-three percent of patients (38/52) were ranked in the same clusters by the five methods, 17.5% of patients (9/52) were classified in the same clusters by 4 methods and 9.5% of patients (5/52) were classified in the same clusters by 3 methods.

3.2.3. Comparison of clinical characteristics between clusters

As shown in Table 2, the 5 methods revealed significant inter-cluster differences. Patients in cluster 3 had mainly unfavorable prognostic factors: tumor stage T2/T3, histological grade III, high mitotic score and triple-negative phenotype. In contrast, patients in cluster 1 had mainly favorable prognosis factors: tumor stage T1, histological grade I/II, lower mitotic score and luminal phenotype, whereas patients in cluster 2 constitute an intermediate group presenting both good and poor prognostic factors. Clusters defined by PCA k-means were significantly different for 5 characteristics: tumor stage, mitosis, tumor phenotype, Her-2 status and luminal. Clusters defined by Spectral Clustering were significantly different for 6 characteristics: tumor stage, histological grade, mitosis, Ki67, tumor phenotype and luminal. Clusters defined by Sparse k-means were significantly different for 4 characteristics: histological grade, tumor phenotype, Her-2 status and luminal. Clusters defined by SIMLR were significantly different for 6 characteristics: tumor stage, histological grade, mitosis, Ki67, tumor phenotype and luminal. Clusters defined by K-Sparse were significantly different for 6 characteristics: tumor stage, histological grade, mitosis, Ki67, tumor phenotype and luminal. From a strictly clinical point of view, Spectral clustering, SIMLR and K-sparse are the 3 most discriminating methods. Indeed, for these 3 methods, six prognostic factors (tumor stage, histological grade, mitosis score, Ki-67, tumor phenotype and luminal) were distributed significantly different between the 3 clusters.

3.2.4. Comparison of survival and predicted survival between clusters

None of the methods created clusters showing significant differences for OS, SS or RFS. Analysis of patients' simulated survival data.
using PREDICT tool are presented in Table 3 and show a predicted survival gradient for clusters obtained with the 5 methods for OS and SS. There were significant differences for 5-year pOS between clusters obtained with K-sparse (p = 0.021), Sparse K-means (p = 0.049), Spectral and clustering (p = 0.021). The five methods showed a significant difference for 5-year pSS between clusters. In terms of 10-year pOS, there were no significant differences between clusters obtained by any of the 5 methods. In contrast, for 10-year pSS, the 5 methods showed significant differences between clusters. Patients in cluster 3 clearly showed the poorest predicted survival.

### 3.2.5. Comparison of the most impactful metabolites according to the five methods

To relate the impact of 449 metabolites to cluster construction, we ranked these metabolites extracted from each of the five methods based on their functional contributions to outputs. With this approach, we classified the relative impact of metabolites on cluster construction and on the identification of metabolic signatures. The highest-ranked metabolites were those that provided relevant information to the signature versus those that provided redundant information or no information. Among a total of 449 metabolites, 116 (26%) were selected by K-sparse clustering and 69 (15%) by Sparse K-means clustering. As for the three other methods, which don't select sparse features, the number of metabolites remained equal to 449. The 50 most effective metabolites identified by the five methods are presented in Supplementary Table 2. Furthermore, a comparison of the top 50 metabolites in each of the 5 methods is presented using a Venn diagram (Fig. 3). Two metabolites were shared by the 5 methods (Creatine, L-Proline), 9 were shared by 4 methods (Betaine, Glutathione, Humulinic Acid A, Isoleucyl-Methionine, L-Carnitine, L-Methionine, L-Phenylalanine Triethanolamine, Alnustone), 28 were shared by 3 methods and 38 were shared by 2 methods (Table 4).

### 3.2.6. Comparison between 5 methods of identified metabolic pathways

For a better understanding of metabolic dysregulation among BC subtypes, pathway analysis was performed. Identification of all the metabolic pathways highlighted by each of the 5 methods as shown in Supplementary Table 3. The most relevant pathways for each of the 5 methods are shown in Table 5. Sparse K-means identified only one statistically significant pathways, “cysteine and methionine metabolism”, involved in amino acid metabolism. K-Sparse identified 3 different pathways: “glycerolipid metabolism”, “Starch and sucrose metabolism” involved in carbohydrates metabolic pathway and “Aminoacyl-tRNA biosynthesis” involved in translation pathway. Spectral clustering identified 17 pathways, the 3 most important being “Glycine, serine and threonine metabolism”, “Alanine, aspartate and glutamate metabolism” and “Histidine metabolism and glutathione metabolism” involved in amino acid metabolic pathway. PCA K-
Table 2
Clinical comparison of 52 patients between clusters.

| Clinical characteristic | PCA-K-means | Spectral Clustering | Sparse K-means | SIMLR | K-Sparse |
|-------------------------|-------------|---------------------|---------------|-------|---------|
|                         | C1 (N = 21) | C1 (N = 10)        | C1 (N = 12)   | C1 (N = 24) | C1 (N = 12) |
| Age *                  | 62.7 (15.2) | 64.8 (16.4)        | 62.5 (16.5)   | 62.8 (15.3) | 0.8      |
|                         | 64.1 (15.4) | 60.5 (17.2)        | 63 (14.9)     | 64.3 (14.1) | 0.85     |
|                         | 64.3 (14.3) | 64.9 (16.1)        | 61.4 (15.6)   | 61.4 (16.3) | 0.755    |
|                         | 61.4 (16.3) | 61.4 (16.3)        | 62.5 (16.5)   | 62.5 (15.3) | 0.827    |
| Histology type         |             |                     |               |       |         |
| Ductal carcinoma       | 19 (90.5)   | 17 (89.5)          | 20 (95.2)     | 21 (87.5) | 0.392    |
|                         | 17 (89.5)   | 11 (91.7)          | 20.5 (95.2)   | 15 (88.2) | 0.106    |
| Lobular carcinoma      | 2 (9.5)     | 2 (10.5)           | 8 (3.3)       | 3 (12.5)  | 0.752    |
|                         | 2 (10.5)    | 1 (8.3)            | 0 (0)         | 2 (11.8)  | 0.392    |
| Microinvasive carcinoma| 0 (0)       | 0 (0)              | 1 (4.8)       | 0 (0)     | 0 (0)    |
| Tumor stage            |             |                     |               |       |         |
| T1                     | 14 (66.7)   | 12 (63.2)          | 22 (11)       | 14 (58.3) | 0.005    |
|                         | 7 (33.3)    | 7 (36.8)           | 7 (33.3)      | 7 (41.7)  | 0.006    |
| Auxillary lymph node   | 1 (0.8)     | 1 (0.8)            | 0.075         | 0.075   | 0.005    |
| N0                     | 14 (66.7)   | 14 (73.7)          | 16 (25.0)     | 16 (25.0) | 0.075    |
|                         | 7 (33.3)    | 6 (36.3)           | 11 (15.6)     | 11 (24.5) | 0.378    |
| Metastasis             | 0.667       | 1                 | 0.497         | 1       |         |
| M0                     | 23 (90.5)   | 18 (94.7)          | 19 (95.5)     | 17 (100) | 0.109    |
|                         | 1 (5.3)     | 1 (5.3)            | 1 (5.3)       | 1 (5.3)  | 0.025    |
| Histological grade     |             |                     |               |       | 0.008    |
| I/II                   | 13 (61.9)   | 12 (63.2)          | 15 (25.0)     | 16 (25.0) | 0.024    |
|                         | 8 (38.1)    | 7 (36.8)           | 7 (33.3)      | 7 (41.7)  | 0.016    |
| Mitosis                |             |                     |               |       | 0.004    |
| 1                      | 11 (52.4)   | 10 (52.6)          | 11 (45.8)     | 10 (58.8) | 0.024    |
|                         | 3 (14.3)    | 3 (15.8)           | 4 (16.7)      | 2 (11.8)  | 0.026    |
| Ki67 *                 | 25 (100)    | 41.1 (33.2)        | 30.0 (20.2)   | 38.3 (31) | 0.024    |
|                         | 6 (25.0)    | 9 (42.9)           | 27.2 (18.7)   | 32.8 (27.2) | 0.012 |
| Tumour phenotype       |             |                     |               |       | 0.006    |
| Her-2 over-expressed    | 1 (4.8)     | 1 (5.3)            | 2 (8.3)       | 1 (5.9)  | 0.024    |
| Luminal                | 28 (25.0)   | 38 (31)            | 20 (15.5)     | 13 (22.7) | 0.012    |
| Triple-Negative        | 6 (25.0)    | 9 (42.9)           | 27.2 (18.7)   | 32.8 (27.2) | 0.012 |
|                   | 20 (100)    | 27 (66.7)          | 18 (66.7)     | 21 (70.6) | 0.196    |
| Hormonal receptors status | 0.178     | 0.075              | 0.112         | 0.071   | 0.075    |
| Negative               | 7 (33.3)    | 6 (31.6)           | 6 (31.6)      | 5 (29.4)  | 0.097    |
| Positive               | 16 (66.7)   | 10 (58.8)          | 7 (39.5)      | 7 (39.5)  | 0.097    |
| Her-2 status           | 0.028       | 0.061              | 0.115         | 0.061   | 0.028    |
| Non-over-expressed      | 20 (95.2)   | 18 (94.7)          | 22 (91.7)     | 16 (94.1) | 0.028    |
| Over-expressed          | 1 (4.8)     | 1 (5.3)            | 2 (8.3)       | 1 (5.9)  | 0.104    |
| Triple-Negative status | 0.272       | 0.104              | 0.087         | 0.104   | 0.027    |
| No                     | 15 (71.4)   | 14 (73.7)          | 18 (75.0)     | 14 (73.7) | 0.047    |
| Yes                    | 6 (28.6)    | 5 (26.3)           | 6 (25.0)      | 5 (26.3)  | 0.014    |
| Luminal                | 0.047       | 0.104              | 0.015         | 0.014   | 0.047    |
| No                     | 7 (33.3)    | 6 (31.6)           | 8 (33.3)      | 5 (29.4)  | 0.047    |
| Yes                    | 14 (66.7)   | 13 (68.4)          | 12 (70.6)     | 13 (68.4) | 0.52     |
| Adjuvant               | 0.52        | 0.423              | 0.459         | 0.459   | 0.423    |
| Chemotherapy           |             |                     |               |       |         |
| No                     | 7 (33.3)    | 7 (36.8)           | 7 (36.8)      | 6 (35.3)  | 0.52     |
| Yes                    | 14 (85.7)   | 12 (63.2)          | 12 (63.2)     | 11 (64.7) | 0.561    |
| Adjuvant Radiotherapy  | 0.561       | 0.140              | 1             | 1       | 0.083    |
| No                     | 3 (14.3)    | 3 (15.8)           | 3 (12.5)      | 3 (17.6)  | 0.561    |
| Yes                    | 18 (85.7)   | 16 (84.2)          | 21 (87.5)     | 14 (82.4) | 0.561    |

C1: cluster 1; C2: cluster 2; C3: cluster 3; *: mean (sd) or median (min, max).
means identified 10 pathways the 3 most important of which are “Alanine, aspartate and glutamate metabolism” involved in amino acid metabolic pathway, “Pyruvate metabolism” involved in carbohydrates metabolic/glucose oxidation pathway and “Citrate cycle (TCA cycle)” involved in energy metabolic pathway.

Finally, with 30 identified pathways, SIMLR is the method that identified the most metabolic pathways. Of these, the 3 most important highlighted metabolic pathways are “arginine and proline metabolism”, “glycine, serine and threonine metabolism” and “alanine, aspartate and glutamate metabolism”, involved in amino acid metabolic pathways. The Venn diagram (Fig. 4) shows the overlap of pathways detected by the five methods. Amino acid metabolism appeared to be the most frequently modified pathway. Enrichment and pathway analyses also showed modifications in glucose metabolism. From the biological point of view, SIMLR and spectral clustering are the two methods that identified the most relevant metabolic pathways.

3.2.7. Comparison of intensity of metabolites between the 5 methods

Among amino acid and glucose metabolisms, fourteen related metabolites were selected as potential biomarkers in BC [54–57]. As shown in Supplementary Fig. 4, the intensities of these 14 metabolites were compared between the 3 clusters for each of the 5 methods. The intensity of Uridine diphosphate (UDP) glucose, Guanine, L-Glutamine, L-Glutamic acid, L-Isoleucine, L-Proline, L-Methionine, L-Phenylalanine, Pyruvic acid, Spermine, Glutathione, Creatine, L-Carnitine and L-Acetylcarnitine were statistically significant between at least one of the clusters. The five methods agree that cluster 3 patients have low levels of Creatine, L-acetylcarnitine, L-Glutamic acid and high levels of Guanine, L-Isoleucine, L-Phenylalanine, Pyruvic acid and Spermine (Fig. 5). These metabolite levels seem to be predictive of poor prognosis [57–59].

4. Discussion

4.1. From a machine learning perspective

To the best of our knowledge, this proof-of-concept study is the first to compare different unsupervised ML methods to identify metabolomics-based prognostic signatures in BC. Analyses were performed intentionally without any prior clinical or biological assumptions. Clinical and biological interpretations were performed only after cluster identification. The objective of our study was to compare different unsupervised ML algorithms for feature selection from untargeted metabolomic data and to evaluate the capacity of these methods to select relevant features for further use in prediction models. This study did not seek to highlight significant differences but rather to assess how unsupervised methods might behave with high-dimension metabolic data and to open up new perspectives in the particularly active domain of BC.
phenotype predictors. We demonstrated that the K-sparse and SIMLR methods have a higher clustering performance compared with the three other popular unsupervised ML methods in detecting groups of patients with BC using metabolomic data. Interestingly, even though the spectral method is a little less clinically efficient than the k-sparse and SIMLR methods, it identified relevant metabolic pathways.

Our study suffers from various limitations, namely the relatively small number of patients and the monocentric and retrospective nature of the study. Besides, our results could not be validated on an external cohort. The clustering performances were assessed only by internal validation based on silhouette value. Indeed, we could not compare the labels obtained from our classification with the true labels to calculate the accuracy of the classification since the true labels were unknown.

Other unsupervised ML methods such as model-based clustering, bi-clustering and deep learning may be of value in this analysis and should be further explored. Yet it is worth noting that, even

Table 4
Table indicating which metabolites are in each intersection or are unique to a certain list.

| Clustering Methods | Nbr | Metabolites |
|--------------------|-----|-------------|
| 5 K-Sparse PCA K-means SIMLR Sparse K-means Spectral clustering | 2 | Creatine; L-Proline; |
| 4 K-Sparse SIMLR Sparse K-means Spectral clustering K-Sparse PCA K-means SIMLR Sparse K-means Spectral clustering K-Sparse PCA K-means SIMLR Sparse K-means Spectral clustering | 1 | Triethanolamine; |
| 3 K-Sparse SIMLR Sparse K-means K-Sparse PCA K-means SIMLR Sparse K-means | 20 | Aminoadipic acid; Methylmalonic acid; 1b-Furanoeudesm-4(15)-en-1-ol acetate; Glycerophosphocholine; Lidocaine; Adenosine monophosphate; 2-Methyl-3-ketovaleric acid; Lecoumarin; p-Cresol sulfate; 2-Methylbutyrylcarnitine; Methoxsalen; Citramalic acid; Hypoxanthine; N-Acetylcarnitine; Ethyl acetate; Guanine; L-Glutamic acid; Uridine 5'-monophosphate; N1,N2-Diacetylserine; 5-Aminimidazole ribonucleotide |
| 3 K-Sparse SIMLR Sparse K-means | 2 | L-Carnitine; Betaine; |
| 2 K-Sparse Sparse K-means K-Sparse PCA K-means SIMLR Spectral clustering | 3 | L-Methionine; L-Phenylalanine |
| 1 K-Sparse PCA K-means Sparse K-means Spectral clustering | 10 | Prolylhydroxyproline; Guanidinoacetic acid; Histamine; PC-M6; (&)-Histidine; N-Acetyl-L-aspartic acid; 3-Mercaptophe nyl hexanone; Trimethylamine N-oxide; Pantothetic acid; Fluoxetin |
### Table 5
List of significant relevant pathways identified by 5 methods.

**K-Sparse method**

| Clusters Comparison | Interaction metabolite | Pathway Name                                      | Total Cmpd | Match Status | Raw P | -log(p) | Impact |
|---------------------|------------------------|---------------------------------------------------|------------|--------------|-------|---------|--------|
| C1 vs C3            | UDP – glucose          | Starch and sucrose metabolism                    | 50         | 1            | 0,0107| 4,5388 | 0,1390 |
|                     | UDP – glucose          | Amino sugar and nucleotide sugar metabolism       | 88         | 1            | 0,0107| 4,5388 | 0,0928 |
|                     | UDP - glucose; Glyceric acid | Glycolipid metabolism                            | 32         | 2            | 0,0153| 4,1831 | 0,0206 |

**SIMLR method**

| Clusters Comparison | Interaction metabolite | Pathway Name                                      | Total Cmpd | Match Status | Raw P | -log(p) | Impact |
|---------------------|------------------------|---------------------------------------------------|------------|--------------|-------|---------|--------|
| C1 VS C2            | Glutathione; Oxidized glutathione; Glycine; β-Glutamic acid; Pyroglutamic acid; Spermidine; Ornithine; Putrescine; Spermine; Cadaverine; Aminopropylcadaverine; Ascorbic acid | Glutathione metabolism                             | 38         | 12           | 0     | 12,826  | 0,3628 |
|                     | Ascorbic acid; Uridine diphosphate glucose; Pyruvic acid; D-Glucuronic acid 1-phosphate; Oxoglutaric acid; l-Tryptophan; N-Acetylsertotonin; 5-Hydroxyindoleacetic acid; 2-Aminomuconic acid semialdehyde; 3-Hydroxyanthranilic acid; l-Kynurenine; Acetyl-N-formyl-5-methoxykynurenamine; Isophenoxazine; S- Methylthioadenosine; N-Formyl-l-methionine; l-Homocysteine; l-Methionine; Glutathione; Phosphoserine; 3-Sulfinoalanine; l-Aspartyl-4-phosphate; Pyruvic acid; | Ascorbate and adarlate metabolism                   | 45         | 5            | 0     | 12,469  | 0,1338 |
|                     | l-Glutamine; Glucosamine 6-Phosphate; Uridine diphosphate-N-acetylglucosamine; Creatine; D-Glucose; D-Xylose; l-Threonine; O-Phosphohomoserine; l-Aspartyl-4-phosphate; Glycolic acid; Pyruvic acid; l-Glutamine; Ornitiline; Citrulline; l-Arginine; β-Glutamic acid; 4-Guanidinobutanoic acid; N2-Succinyl-l-ornithine; Putrescine; Spermidine; N-Acetylputrescine; O-phosphohomoserine; Cytidine monophosphate N-acetylneuraminic acid; D-Glucose; D-Xylose; l-Argnine; l-Proline; Hydroxyproline; Guanidoacetic acid; Creatine; Oxoglutaric acid; Oxalosuccinic acid; Pyruvic acid; | Arginine and proline metabolism                     | 77         | 19           | 0,0053| 5,238   | 0,6514 |
|                     | D-Xylose; Uridine diphosphate glucose; D-Glucuronic acid 1-phosphate; Pyruvic acid; | Citrate cycle (TCA cycle)                          | 20         | 3            | 0,0075| 4,8991 | 0,176  |
|                     | 2-Hydroxyethanesulfonate; Pyruvic acid; 3-Sulfinoalanine; | Pentose and glucuronate interconversions           | 53         | 4            | 0,0076| 4,8821 | 0,0394 |
|                     | Glycic acid; Betaine; Guanidoacetic acid; Dimethylglycine; Glycine; Phosphoserine; l-Threonine; O-Phosphohomoserine; l-Aspartyl-4-phosphate; Creatine; Glycolic acid; Pyruvic acid; d-Tryptophan; | Taurine and hypotaurine metabolism                | 20         | 3            | 0,0154| 4,1754 | 0,0324 |
|                     | Uridine diphosphate glucose; D-Glucuronic acid 1-phosphate; N-Acetyl-D-Glucosamine 6-Phosphate; Uridine diphosphate-N-acetylglucosamine; Cytidine monophosphate N-acetylneuraminic acid; D-Glucose; D-Xylose; Formiminoglutamic acid; l-Glutamic acid; Urocanic acid; l-Histidine; Histamine; D-Erythro-imidazole-glycerol-phosphate; Ergothioneine; Hydantoin-5-propionic acid; Imidazole acetyl-phosphate; Oxoglutaric acid; Pyruloxamine; Oxoglutaric acid; 3-Hydroxy-2-methylpyridine-4,5-dicarboxylate; Pyruvic acid; | Amino sugar and nucleotide sugar metabolism         | 48         | 13           | 0,018 | 4,0154 | 0,46986 |
|                     | Vitamin B6 metabolism  | Histidine metabolism                              | 32         | 4            | 0,0412| 3,1893 | 0,3705 |
|                     | vitamin B6 metabolism  | Vitamin B6 metabolism                              | 32         | 4            | 0,0412| 3,1898 | 0,0773 |
| C1 VS C3            | Formiminoglutamic acid; l-Glutamic acid; Urocanic acid; Histidine; Histamine; D-Erythro-imidazole-glycerol-phosphate; Ergothioneine; Hydantoin-5-propionic acid; Imidazole acetyl-phosphate; Oxoglutaric acid; Phenylpyruvic acid; l-Phenylalanine; l-Tyrosine; 3-Dehydroquinicinate; l-Tryptophan; | Histidine metabolism                               | 44         | 10           | 0,0139| 4,2752 | 0,3705 |
|                     | l-Tryptophan; N-Acetylsertotonin; 5-Hydroxyindoleacetic acid; 2-Aminomuconic acid semialdehyde; 3-Hydroxyanthranilic acid; l-Kynurenine; Acetyl-N-formyl-5-methoxykynurenamine; Isophenoxazine; S- Methylthioadenosine; N-Formyl-l-methionine; l-Homocysteine; l-Methionine; Glutathione; Phosphoserine; 3-Sulfinoalanine; l-Aspartyl-4-phosphate; Pyruvic acid; | Phenylalanine, tyrosine and tryptophan biosynthesis | 27         | 5            | 0,0189| 3,9687 | 0,099  |
|                     | | Tryptophan metabolism                             | 79         | 8            | 0     | 16,409 | 0,2741 |
| C2 VS C3            | Glutathione; Oxidized glutathione; Glycine; β-Glutamic acid; Pyroglutamic acid; Spermidine; Ornithine; Putrescine; Spermine; Cadaverine; Aminopropylcadaverine; Ascorbic acid; Ascorbic acid; Uridine diphosphate glucose; Pyruvic acid; D-Glucuronic acid 1-phosphate; Oxoglutaric acid S- Methylthioadenosine; N-Formyl-l-methionine; l-Homocysteine; l-Methionine; Glutathione; Phosphoserine; 3-Sulfinoalanine; l-Aspartyl-4-phosphate; | Ascorbate and adarlate metabolism                   | 45         | 5            | 0     | 13,096 | 0,1383 |
|                     | | Cysteine and methionine metabolism               | 56         | 9            | 0,0001| 9,8548 | 0,2509 |

(continued on next page)
### SIMLR method

| Clusters | Interaction metabolite | Pathway Name | Total Cmpd | Match Status | P Value | -log(p) | Impact |
|----------|------------------------|--------------|------------|--------------|---------|---------|--------|
|          | phosphate; Pyruvic acid; Phenylpyruvic acid; \(-\)-Phenylalanine; \(-\)-Tyrosine; 3-Dehydroquinase; \(-\)-Tryptophan; | metabolism | 27 | 5 | 0.0001 | 8.9814 | 0.099 |
|          | \(-\)-Histidine; \(-\)-Phenylalanine; \(-\)-Arginine; \(-\)-Glutamine; Glycine; \(-\)-Methionine; \(-\)-Lysine; \(-\)-Isocitric acid; \(-\)-Threonine; \(-\)-Tryptophan; \(-\)-Tyrosine; \(-\)-Proline; \(-\)-Glutamic acid; Phosphoserine; Glyoxylate; Oxoglutarate acid; N-Formyl-\(-\)-methylamine; Glycolic acid; | Aminoacyl-RNA biosynthesis | 75 | 14 | 0.0002 | 8.758 | 0.1127 |
|          | \(-\)-Histidine; \(-\)-Phenylalanine; \(-\)-Arginine; \(-\)-Glutamine; Glycine; \(-\)-Methionine; \(-\)-Lysine; \(-\)-Isocitric acid; \(-\)-Threonine; \(-\)-Tryptophan; \(-\)-Tyrosine; \(-\)-Proline; \(-\)-Glutamic acid; Phosphoserine; Glyoxylate; Oxoglutarate acid; N-Formyl-\(-\)-methylamine; | Glyoxylate and dicarboxylate metabolism | 50 | 6 | 0.0004 | 7.7271 | 0.268 |
|          | \(-\)-Glutamine; Phosphoribosylformylglycinamide; Cyclic AMP; Adenosine monophosphate; Adenosine; Inosine; Adenine; Hypoxanthine; Guanine; Uric acid; 5-Hydroxyisourate; Guanosine; Adenosine diphosphate ribose; 5-Aminomimidazole ribonucleotide; Glycine; Adenosine; | Purine metabolism | 92 | 17 | 0.0007 | 7.306 | 0.2048 |
|          | Malonic acid; Beta-Alanine; Spermine; Spermidine; Dihydouracil; Pantothentic acid; Uracil; \(-\)-Histidine | beta-Alanine metabolism | 28 | 8 | 0.0012 | 6.7568 | 0.3577 |
|          | Uracil 5'-monophosphate; \(-\)-Glutamine; Dihydouracil; Cytidine monophosphate; Cytidine; Cytosine; Uracil; Dihydrothymine; Uridine diphosphate glucose; Malonic acid; Adenosine; Adenosine 3',5'-diphosphate; Malonic acid; Baeta-Alanine; | Pyrimidine metabolism | 60 | 13 | 0.0014 | 6.5817 | 0.2756 |
|          | N-Acetyl-l-aspartic acid; Pyruvic acid; Ureidosuccinic acid; Oxoglutaric acid; | Citrate cycle (TCA cycle) | 56 | 2 | 0.0007 | 6.0789 | 0.2736 |
|          | l-Phenylalanine; Phenylpyruvic acid; Benzoic acid; Hippuric acid; Pyruvic acid; \(-\)-Tyrosine; \(-\)-Glutamic acid; l-Arginine; l-Aspartate; Ethylmalonic acid; | Taurine and hypotaurine metabolism | 44 | 14 | 0.0002 | 8.758 | 0.1127 |
|          | l-Phenylalanine; Phenylpyruvic acid; Benzoic acid; Hippuric acid; Pyruvic acid; \(-\)-Tyrosine; \(-\)-Glutamic acid; l-Arginine; l-Aspartate; Ethylmalonic acid; | Taurine and hypotaurine metabolism | 48 | 18 | 0.0002 | 8.758 | 0.1127 |
|          | l-Phenylalanine; Phenylpyruvic acid; Benzoic acid; Hippuric acid; Pyruvic acid; \(-\)-Tyrosine; \(-\)-Glutamic acid; l-Arginine; l-Aspartate; Ethylmalonic acid; | Taurine and hypotaurine metabolism | 52 | 20 | 0.0002 | 8.758 | 0.1127 |
|          | l-Phenylalanine; Phenylpyruvic acid; Benzoic acid; Hippuric acid; Pyruvic acid; \(-\)-Tyrosine; \(-\)-Glutamic acid; l-Arginine; l-Aspartate; Ethylmalonic acid; | Taurine and hypotaurine metabolism | 56 | 2 | 0.0007 | 6.0789 | 0.2736 |
|          | l-Phenylalanine; Phenylpyruvic acid; Benzoic acid; Hippuric acid; Pyruvic acid; \(-\)-Tyrosine; \(-\)-Glutamic acid; l-Arginine; l-Aspartate; Ethylmalonic acid; | Taurine and hypotaurine metabolism | 56 | 2 | 0.0007 | 6.0789 | 0.2736 |
|          | l-Phenylalanine; Phenylpyruvic acid; Benzoic acid; Hippuric acid; Pyruvic acid; \(-\)-Tyrosine; \(-\)-Glutamic acid; l-Arginine; l-Aspartate; Ethylmalonic acid; | Taurine and hypotaurine metabolism | 60 | 13 | 0.0014 | 6.5817 | 0.2756 |
|          | l-Phenylalanine; Phenylpyruvic acid; Benzoic acid; Hippuric acid; Pyruvic acid; \(-\)-Tyrosine; \(-\)-Glutamic acid; l-Arginine; l-Aspartate; Ethylmalonic acid; | Taurine and hypotaurine metabolism | 60 | 13 | 0.0014 | 6.5817 | 0.2756 |

### Sparse K-means method

| Clusters | Interaction metabolite | Pathway Name | Total Cmpd | Match Status | Raw p | -log(p) | Impact |
|----------|------------------------|--------------|------------|--------------|-------|---------|--------|
| C1 VS C2 | l-Methionine; Glutathione | Cysteine and methionine metabolism | 56 | 2 | 0.0007 | 6.0789 | 0.2736 |
| C1 VS C3 | l-Methionine; Glutathione | Cysteine and methionine metabolism | 56 | 2 | 0.0007 | 6.0789 | 0.2736 |

### Spectral clustering method

| Clusters | Interaction metabolite | Pathway Name | Total Cmpd | Match Status | Raw p | -log(p) | Impact |
|----------|------------------------|--------------|------------|--------------|-------|---------|--------|
| C1 VS C3 | l-Methionine; Glutathione | Nicotinate and nicotinamide metabolism | 44 | 5 | 0.0024 | 6.0206 | 0.0712 |
| C1 VS C3 | l-Methionine; Glutathione | Nicotinate and nicotinamide metabolism | 48 | 13 | 0.0040 | 5.5100 | 0.4699 |
though deep learning methods are of particular interest in many fields, they necessitate a very large number of patients to be efficiently trained and may therefore not be suitable for small metabolomics datasets obtained on real-life patients, such as the one we have used. While obtaining imaging or clinical data concerning several thousands of patients seems achievable, obtaining metabolomics data for that many patients is currently much more complicated. Furthermore, even though some efforts are being made to tackle this issue [60], it is currently impossible to understand which features are responsible for the outcome when using deep-learning clustering techniques. It would therefore be impossible to understand the metabolic differences underlying different patient clusters if deep-learning clustering was used.

These considerations raise important questions: in the future, on what basis should decisions be made? On results from a single patient clusters if deep learning clustering was used.
findings we have highlighted, it seems that decisions should be taken collegially, i.e. based on the results of a set of methods, as at multidisciplinary consultation meetings involving health professionals from different disciplines and whose skills are essential to take decisions ensuring patients the best possible care according to the state of the science.

4.2. From a clinical perspective

From a clinical point of view, the methods were able to highlight three distinct groups of patients with different clinical profiles. Patients identified in cluster 1 may be considered to have the best prognosis, patients in cluster 2 an intermediate prognosis, while patients in cluster 3 may be considered to have the worst prognosis. The results in Table 2 show that the tumors of patients in cluster 1 were predominantly non-invasive and non-proliferative, whereas the tumors of cluster 3 patients were mainly invasive and proliferative. Tumors in cluster 2 were rather invasive but not proliferative, hence the intermediate prognosis. We hypothesize that these patients would have an intermediate (atypical) biological profile, which is why the methods are discordant.

We further evidence heterogeneity within the triple-negative BC subpopulation with most of the patients classified in cluster 3. However, a third of the triple-negative patients were in cluster 1. Recent molecular profiling studies of triple-negative BC using parallel sequencing and other “omics” technologies have also uncovered an unexpectedly high level of heterogeneity as well as a number of common features [61,62].

In addition, no significant difference between clusters could be demonstrated in terms of age, histologic type, lymph node involvement, metastasis or survival (OS, SS or RFS). Indeed, with a median follow-up of only 48.5 months, this duration is insufficient to demonstrate a significant difference in terms of OS, SS, or RFS. Nevertheless, it is quite easy to predict that patients in cluster 3 have the highest risk of progression and that, conversely, patients in cluster 1 have the lowest risk of progression. To confirm this intuition and try to reduce this short follow-up limitation, we analyzed simulated survival data obtained with the PREDICT tool. With a 5-year pOS rate at around 75% for cluster 1, 70% for cluster 2 and 60% for cluster 3, in-silico analyses have demonstrated their high potential value [28,63,64] and confirmed that patients in cluster 3 have a poorer prognosis [65,66]. One limitation of our study could be the

Fig. 4. Venn diagram of pathways that were in common or unique to the five clustering methods.
Fig. 5. Boxplot of the 8 metabolites extracted from 5 ML methods.
ences between the BC subtypes of three metabolic pathways (i.e. cer development [72] and are emerging hallmarks of cancers of amino acid metabolism are well-known key events during can-
molecular stratification of BC to metabolomic profiles. Indeed,
our results suggest that dysregulation of metabolic pathways
exists between BC subtypes and that a particular amino acid profile
may result from our method of tumor selection. We decided
to analyze frozen samples available in our biobank. Obviously,
hormonal-receptor negative, triple-negative, Her-2-positive tumors are more often frozen and stored for further molecular test-
ing and inclusion in clinical trials. In the present study, it is inter-
esting to note that the five methods classified 73% of the patients in
the same cluster. Among the 27% of patients classified differently
by at least one of the methods, 9.5% of patients were classified heterogeneously by the five methods. Indeed, for each of these 5 patients, three methods classified them in one cluster and 2 others in another cluster without any connection between the types of
methods used. Moreover, it is interesting to note that the different methods classified patients, on the one hand, in either the good prognostic cluster or the intermediate prognostic cluster or, on
the other, in either the intermediate prognostic cluster or the poor
prognostic cluster, but never in the good prognostic cluster or the poor prognostic cluster. A clinical analysis of these 5 patients
showed that they had atypical clinical profiles, probably due to
particular biological profiles. These atypical profiles would explain
why no classification consensus could be highlighted. Overall, ML
methods must remain a decision-making tool for the clinician,
especially in cases where patients have particular clinical and bio-
logical characteristics. To avoid possible medical errors, the final
responsibility for the decision lies with the clinician [71].

Finally, the initial clinical objective of this study was to define a
metabolomic signature to refine the current classification and help
the clinician in his chemotherapy prescription. This paper is the
result of methodological research analyzing the best ML methods
to develop this new tool. The patients selected were therefore
patients eligible for adjuvant chemotherapy. An analysis of the
metastatic population could help define a specific signature of
metastatic status and/or a signature associated to survival. How-
ever, the use of biopsy faces two practical difficulties: 1) the intra-
tumoral and inter-site heterogeneity that could be overcome through
the analysis of blood or urine samples; and 2) the amount of material available once the pathologic analyses essential for
patient management have been performed. Metabolomic analysis
on paraffin slides could facilitate access to specimens and limit
the amount of material required.

4.3. From a biological perspective

From a physiological point-of-view, this study extends the molecular stratification of BC to metabolomic profiles. Indeed,
our results suggest that dysregulation of metabolic pathways
exists between BC subtypes and that a particular amino acid profile
characterizes the different BC histologic subtypes. Dysregulations of
amino acid metabolism are well-known key events during can-
cer development [72] and are emerging hallmarks of cancers [73,74]. Amino acids serve not only as building blocks in protein
synthesis but also as energy sources favoring cancer cell prolifera-
tion and growth [75]. Of interest, we identified significant differ-
ences between the BC subtypes of three metabolic pathways (i.e.
Glycolysis and lactate production, Glutaminolysis, and amino acid)
that play a pivotal role in BC growth [76,77]. Using the five meth-
ods, we consistently found that patients in cluster 3 showed higher
levels of Guanine, L-Isoleucine, L Methionine, L-Phenylalanine,
Pyruvic acid, Spermine and low levels of Creatine, L-
Acetylcarnitine and L-Glutamic acid. Our results suggested that
these metabolites could be candidate biomarker predictors of
poorer prognosis [78–82]. All these results are consistent with
the literature [57,83–86].

Given the exploratory nature of our study, we decided to use an
FDR rate of 0.25 as a threshold in order to identify relevant candidate
pathways (https://software.broadinstitute.org/cancer/software/gsea/wiki/index.php/FAQ).

A validation of these pathways, during a study whose main
objective will be to evaluate the usefulness of our metabolomics
signatures for decision-making, will need to be established with
the use of a lower False Discovery Rate or Family Wise Error Rate
(<0.05).

Indeed, to meet the biosynthetic needs associated with rapid pro-
liferation, cancer cells must increase the import of nutrients. Two
main metabolites are essential for biosynthesis and survival in
mammalian cells, and particularly in cancer cells: glucose [87] and
The increased glucose uptake in tumors compared to other healthy and non-proliferative tissues was first described
more than 90 years ago by Otto Warburg [89]. Glucose is the primary
energy source of all cells because of its involvement in many pro-
cesses such as glycolysis or the Krebs cycle [90] in mitochondria.
Unlike healthy cells that adapt to available substrates (glucose/fatty
acids/proteins), some tumor cells are addicted to glucose. The other
important point is that, once metabolized, tumor cells will prefer
lactic fermentation to the Krebs cycle.

Lastly, the precise etiology of BC is still unknown even though
some genetic, epigenetic and environmental factors have been
identified [91]. It has been conclusively demonstrated that cancer
cell metabolism is heavily influenced by microenvironmental fac-
tors, including nutrient availability. Sullivan and coworkers [92]
found that diet affects local nutrient availability. This effect can
lead to substantial changes in the metabolism of tumor cells,
thereby modifying the response of these cells to drugs targeting
metabolism. Drugs capable of inhibiting tumor proliferation may
then become ineffective. Therefore, knowledge of microenvironmental
nutrient levels is essential to a better understanding of
tumor metabolism.

Outcomes for cancer patients vary greatly. The classification of
BC into subtypes has been was defined in the literature on the basis
of molecular characterization of proteomics (single omic). This has
helped improve prognosis and personalized treatment. These con-
siderations have motivated efforts to produce large amounts of
multi-omic data such as TCGA [93] and ICGC [94]. However, cur-
rent algorithms still face challenges and need to integrate omic
data [95–98]. Defining BC subtypes using multi-omic data could
help to better understand some of the dark areas that still persist
in the field of tumor mechanisms in order to offer even more per-
sonalized treatments.

5. Conclusion

In the era of personalized medicine, OMICS science (genomics,
transcriptomics, proteomics, and metabolomics) must contribute
to the quest for cancer-specific biomarkers. The present study
argues in favor of further research in this domain. Metabolomics
is emerging as a relevant and promising tool for the classification
of BC to enable more precise diagnosis [54,99–101]. Even though
it is less accurate than the targeted approach, untargeted metabo-
lomics nevertheless permits identification and quantification of a
vast number of major metabolites. Thus, this approach presents a particular interest in the search for new candidate biomarkers [102–104] and could be applied in everyday medical practice given that the cost and duration of metabolomic analyses are relatively low. However, due to the retrospective design of our study and the small number of patients recruited, our results need to be validated in a larger cohort and in the context of a prospective clinical trial.

Funding

The authors declare no competing financial interests.

Credit authorship contribution statement

Jocelyn Gal: Methodology, Formal analysis, Writing - original draft. Caroline Bailleux: Writing - original draft. David Chardin: Software, Writing - original draft. Thierry Pochier: Conceptualization, Writing - review & editing. Julia Gilhodes: . Lun Jing:. Jean-Marie Guigonis: Methodology, Writing - review & editing. Jean-Marc Ferrero: Data curation. Gerard Milano: Writing - review & editing. Baharia Mograbi: Writing - review & editing. Patrick Brest: Writing - review & editing. Yann Chateau: Olivier Humbert: Conceptualization, Writing - review & editing. Emmanuel Chamorey: Supervision, Methodology, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors acknowledge support from the Centre Antoine Lacassagne, TlRO Unit, University Côte d’Azur and the Departmental Council of the Alpes Maritimes, France. The authors sincerely thank Mrs. Claire Della Vedova for her help in developing the figures.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2020.05.021.

References

[1] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2017. CA Cancer J Clin 2017;67:7–30.
[2] Perou CM, Jeffrey SS, van de Rijn M, et al. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. Proc Natl Acad Sci USA 1999;96:9212–7.
[3] Lockhart DJ, Winzeler EA. Genomics, gene expression and DNA arrays. Nature 2000;405:827–36.
[4] Pandey A, Mann M. Proteomics to study genes and genomes. Nature 2000;405:837–46.
[5] Perou CM, Sorlie T, Eisen MB, et al. Molecular portraits of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A 2000;97:5566–70.
[6] 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 2000;97:5566–70.
[7] Sorlie T, Tibshirani R, Parker J, et al. Repeated observation of breast cancer subtypes in independent gene expression data sets. Proc Natl Acad Sci USA 2003;100:8418–23.
[8] Witten DM, Tibshirani R. A framework for feature selection in clustering. J Am Stat Assoc 2010;105:713–26.
[9] Candido Dos Reis FJ, Wishart GC, Dicks EM, et al. An updated PREDICT breast cancer prognostication and treatment benefit prediction model with independent validation. Breast Cancer Res 2017;19:58.
[10] Wishart GC, Azzoato EM, Greenberg DC, et al. PREDICT: a new UK prognostic model that predicts survival following surgery for invasive breast cancer. Breast Cancer Res 2010;12:R1.
[11] Ross JS, Multigene predictors in early-stage breast cancer: moving in or moving out?. Expert Rev Mol Diagn 2008;8:129–35.
[12] Ross JS, Hatziis C, Symmans WF, et al. Commercialized multigene predictors of clinical outcome for breast cancer. Oncologist 2008;13:477–92.
[13] Buyse M, Loi S, van’t Veer L, et al. Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. J Natl Cancer Inst 2006;98:1183–92.
[14] Gao Y, DePinho RA, Ernst M, Voussden K. Cancer research: past, present and future. Nat Rev Cancer 2011;11:749–54.
[15] Ehmann F, Caneva L, Prasad K, et al. Pharmacogenomic information in drug labels: European Medicines Agency perspective. Pharmacogenomics J 2015:15:201–10.
[16] McShane LM, Polley MY. Development of omics-based clinical tests for prognosis and therapy selection: the challenge of achieving statistical robustness and clinical utility. Clin Trials 2013;10:853–65.
[17] von de Vijver MJ, He YD, van’t Veer LJ, et al. A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 2002;347:1999–2009.
[18] Wang Y, Klijn JG, Zhang Y, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet 2005;365:671–9.
[19] Wesołowski R, Ramaswamy B. Gene expression profiling: changing face of breast cancer classification and management. Gene Exp 2011;15:105–15.
[20] Marusyk A, Almendro V, Polyak K. Intra-tumour heterogeneity: a looking glass for cancer?. Nat Rev Cancer 2012;12:324–36.
[21] Hsu PP, Sabatini DM. Cancer cell metabolism: Warburg and beyond. Cell 2008;134:703–7.
[22] McClérian J, King MC. Genetic heterogeneity in human cell. Cell 2008;134:210–7.
[23] Cannon WB. The wisdom of the body. 2nd ed. Oxford, England: Norton & Co.; 1939.
[24] Roberts LD, Souza AL, Gerszten RE, Clish CB. Targeted metabolomics. Curr Opin Mol Biol 2012. Chapter 30: Unit 30.32.31-24.
[25] Schrimpe-Rutledge AC, Cordreau SG, Sherrod SD, McLean JA. Untargeted metabolomics strategies-challenges and emerging directions. J Am Soc Mass Spectrom 2010;27:1897–905.
[26] Vinayavethun N, Saghatelian A. Untargeted metabolomics. Curr Protoc Mol Biol 2010. Chapter 30: Unit 30.31.31-24.
[27] Camacho DM, Collins KM, Powers RK, et al. Next-generation machine learning for biological networks. Cell 2018;173:1581–92.
[28] Gal J, Milano G, Ferrero JM, et al. Optimizing drug development in oncology by clinical trial simulation: why and how? Brief Bioinform 2017.
[29] Yu MK, Ma J, Fisher J, et al. Visible machine learning for biomedicine. Cell 2018;173:1562–5.
[30] Jordan MI, Mitchell TM. Machine learning; trends, perspectives, and prospects. Science 2015;349:255–60.
[31] Tang P, Tse GM. Immunohistochemical surrogates for molecular classification of breast carcinomas. Arch Pathol Lab Med 2016;140:806–14.
[32] Katamala M, Orecis M. Processing methods for differential analysis of LC/MS profile data. BMC Bioinf 2005:6;179.
[33] Prusak T, Castillo S, Villar-Briones A, Orecis M. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinf 2010:11;395.
[34] Xia J, Mandral R, Sinelnikov IV, et al. MetaAnalyzeL 2.0—a comprehensive server for metabolomic data analysis. Nucleic Acids Res 2012;40:W127–133.
[35] Tusar RA, Wang C, Zhou Y, Speed TP. Gene set enrichment analysis made simple. Stat Methods Med Res 2009;18:565–75.
[36] Saxena A, Prasad M, Gupta A, et al. A review of clustering techniques and developments. Neurocomputing 2017;267:664–81.
[37] Tibshirani R, Walter G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. J Royal Stat Soc: Series B (Statistical Methodol) 2001;63:411–23.
[38] Kaufman L, Rousseuw PJ. Finding groups in data: an introduction to cluster analysis. John Wiley & Sons.; 2009.
[39] Rousseuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 1987;20:53–65.
[40] Davies DL, Bouldin DW. A cluster separation measure. IEEE Trans Pattern Anal Mach Intell 1979:224–7.
[41] Caliski T, Harbasz J. A dendrite method for cluster analysis. Commun Stat-Theory Methods 1974:3:1–27.
[42] Begg B, Zhu J, Pierson E, et al. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. Nat Methods 2017:14:414–6.
[43] Arthur D, Vassilvitskii S. k-means++: The advantages of careful seeding. In: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics; 2007. p. 1027–35.
[44] Lloyd S. Least squares quantization in PCM. IEEE Trans. Inform. Theory 1982;28(2):129–37. http://dx.doi.org/10.1109/TIT.1982.1056485.
[45] Steinhass H. Sur la division des corps materiels en parties. Bull. Acad. Polon. Sci., Cl. III 1956;V:801–4.
[46] Ng AY, Jordan MI, Weiss Y. Analysis and an algorithm. In: Advances in neural information processing systems. On spectral clustering; 2002. p. 849–56.
[47] Von Luxburg U. A tutorial on spectral clustering. Stat Comput 2007;17:395–416.
Tibshirani R. Regression shrinkage and selection via the lasso. J Roy Stat Soc: Ser B (Methodol) 1996;58:267–88.

Gilet C, Deprez M, Caillau J-B, Barlaud M. Clustering with feature selection using alternating minimization. Application to computational biology. arXiv preprint arXiv:1711.02974 2017.

Lvd Maaten, Hinton G. Visualizing data using t-SNE. J Mach Learn Res 2008;9:2579–605.

Team RCR. A language and environment for statistical. Computing 2013.

Witten DM, Tibshirani R. sparcI: Perform sparse hierarchical clustering and sparse k-means clustering. R package version 2013;1.

Wishart GC, Bajdik CD, Azzoato EM, et al. A population-based validation of the prognostic model PREDICT for early breast cancer. Eur J Surg Oncol 2011;37:411–7.

Beger RD. A review of applications of metabolomics in cancer. Metabolites 2013;3:552–74.

Gunther UL. Metabolomics biomarkers for breast cancer. Pathobiology 2015;82:153–65.

McCarty AE, Vignoli A, Biganzoli L, et al. Metabolomics in breast cancer: a decade in review. Cancer Treat Rev 2018;67:88–96.

Silva C, Perestrello R, Silva P, et al. Breast cancer metabolomics: from analytical platforms to multivariate data analysis. A Review. Metabolites 2019;9.

Asagio VM, Alvarado LZ, Shanaiah N, et al. Early detection of recurrent breast cancer using metabolite profiling. Cancer Res 2010;70:8309–18.

Cardoso MR, Santos JC, Ribeiro ML, et al. A Metabolomic approach to predict breast cancer behavior and chemotherapy response. Int J Mol Sci 2018;19.

Karim MR, Beyan O, Zappa A, et al. Deep learning-based clustering approaches for bioinformatics. Brief Bioinform 2020.

Bianchini G, Ballo JM, Mayer IA, et al. Triple-negative breast cancer: challenges and opportunities of a heterogeneous disease. Nat Rev Clin Oncol 2016;13:674–90.

Mills MN, Yang QG, Oliver DE, et al. Metabolic heterogeneity of triple negative breast cancer: A national cancer centre database analysis. Eur J Cancer 2018;89:48–58.

Belkacemi Y, Hanna NE, Besnard C, et al. Local and regional breast cancer recurrences: salvage therapy options in the new era of molecular subtypes. Front Oncol 2018;8:112.

Buonaguro FM, Caposio P, Tornesello ML, et al. Cancer diagnostic and predictive biomarkers 2018. Biomed Res Int 2019;2019:3879015.

Ponde NF, Zardavas D, Piccart M. Progress in adjuvant systemic therapy for breast cancer. Nat Rev Clin Oncol 2018;15:515–30.

Senkus E, Kyriakides S, Ohno S, et al. Primary breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol 2015;26(Suppl 5):v8–30.

Assi HA, Khoury KE, Dhouk H, et al. Epidemiology and prognosis of breast cancer in young women. J Thorac Dis 2013;5(Suppl 1):52–8.

Wang K, Zhuo GG, Shi Y, et al. Long-term survival differences between T1–2 invasive lobular breast cancer and corresponding ductal carcinoma after breast-conserving surgery: A propensity-scored matched longitudinal cohort study. Clin Breast Cancer 2019;19:e101–15.

Wasif N, Maggard MA, Ko CY, Giuliano AE. Invasive lobular vs. ductal breast cancer: a stage-matched comparison of outcomes. Ann Surg Oncol 2018;25:50–1.

Yersal O, Barutca S. Biological subtypes of breast cancer: Prognostic and therapeutic implications. World J Clin Oncol 2014;5:412–24.

Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019;25:44–56.

Pavlova NN, Thompson CB. The emerging hallmarks of cancer metabolism. Cell Metab 2016;23:27–47.

Hainaut P, Ploymth A. Targeting the hallmarks of cancer: towards a rational approach to next-generation cancer therapy. Curr Opin Oncol 2013;25:50–1.

Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell 2011;144:646–74.

Li Z, Zhang H. Reprogramming of glucose, fatty acid and amino acid metabolism for cancer progression. Cell Mol Life Sci 2016;73:377–92.

DeBerardinis RJ, Chandel NS. Fundamentals of cancer metabolism. Sci Adv 2016;2:e1600200.

Hauskaa TH, Euceda LR, Giskegeorgadj G, Bathein TF. Metabolic portraits of breast cancer by HR MAS MR spectroscopy of intact tissue samples. Metabolites 2017;7.

Jion H, Kim JT, Lee E, et al. Methionine deprivation suppresses triple-negative breast cancer metastasis in vitro and in vivo. Oncotarget 2016;7:67223–34.

Melone MA, Valentino A, Margarucci S, et al. The carnitine system and cancer metabolic plasticity. Cell Death Dis 2019;9:228.

Thomas TJ, Thomas T. Cellular and animal model studies on the growth inhibitory effects of polyamine analogues on breast cancer. Med Sci (Basel) 2018;6.

Xiao F, Wang C, Yin H, et al. Leucine deprivation inhibits proliferation and induces apoptosis of human breast cancer cells via fatty acid synthase. Oncotarget 2016;7:63679–89.

Zuo Y, Ufu A, Chang JT, Frost JA. Contributions of the Rhoa guanine nucleotide exchange factor Nef1 to polyoma middle T antigen-mediated mammary gland tumorigenesis and metastasis. Breast Cancer Res 2018;20:41.

Lecuyer L, Dalle C, Lyan B, et al. Plasma metabolic signatures associated with long-term breast cancer risk in the SU.VI.MAX prospective cohort. Cancer Epidemiol Biomarkers Prev 2019.

Oikari S, Ketunen T, Taininen S, et al. UDP-sugar accumulation drives hyaluronan synthesis in breast cancer. Matrix Biol 2018;67:63–74.

Pan H, Xia K, Zhou W, et al. Low serum creatine kinase levels in breast cancer patients: a case-control study. PLoS One 2013;8:e62112.

Phannasil P, Ansari IH, El Azzouny M, et al. Mass spectrometry analysis shows the biosynthetic pathways supported by pyruvate carboxylase in highly invasive breast cancer cells. Biochim Biophys Acta Mol Basis Dis 2017;1863:357–51.

Mason EF, Rathmell JC. Cell metabolism: an essential link between cell growth and apoptosis. Biochim Biophys Acta 2011;1813:645–54.

Hensley CT, Wasif N, Ai D, DeBerardinis RJ. Glutamine and cancer: cell biology, physiology, and clinical opportunities. J Clin Invest 2013;123:3678–84.

Warburg O, Wind F, Ngelele E. The metabolism of tumors in the body. J Gen Physiol 1927;8:519–30.

Anderson NM, Mucka P, Kern JC, Feng H. The emerging role and targetability of the TCA cycle in cancer metabolism. Protein Cell 2018;9:216–37.

Fernandez MF, Reina-Perez I, Astorga JM, et al. Breast Cancer and Its Relationship with the Microbiota. Int J Environ Res Public Health 2018;15.

Sullivan MR, Danai LV, Lewis CA, et al. Quantification of microenvironmental metabolites in murine cancers reveals determinants of tumor nutrient availability. Elife 2019;8.

Cancer Genome Atlas Research N. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 2008;455:1061–8.

Zhang J, Baran J, Crouser E, et al. International Cancer Genome Consortium Data Portal—a one-stop shop for cancer genomics data. Database (Oxford) 2011;2011:bar026.

Mitra S, Saha S. A multiojective multi-view cluster ensemble technique: Application in patient subclassification. PLoS One 2019;14:e0216904.

Ramazzotti D, Lal A, Wang B, et al. Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival. Nat Commun 2018;9:4453.

Rappoport N, Shamir R. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. Nucleic Acids Res 2018;46:10546–62.

Wu C, Zhou F, Ren J, et al. A selective review of multi-level omics data integration using variable selection. High Throughput 2019;8.

Armitage EG, Barbas C. Metabolomics in cancer biomarker discovery: current trends and future perspectives. J Pharm Biomed Anal 2014;87:1–11.

Bennett DA, Waters MD. Applying biomarker research. Environ Health Perspect 2000;108:907–10.

Vermeersch KA, Styczynski MP. Applications of metabolomics in cancer research. J Carcinog 2013;12:9.

Jacob M, Lopata AL, Daisouki M, Abdel Rahman AM. Metabolomics toward personalized medicine. Mass Spectrom Rev 2017.

Trivedi DK, Hollywood KA, Goodacre R. Metabolomics for the masses: The future of metabolomics in a personalized world. New Horiz Transl Med 2017;3:294–305.

Wasif DS. Emerging applications of metabolomics in drug discovery and precision medicine. Nat Rev Drug Discov 2016;15:473–84.