RESEARCH

Consensus clustering applied to multi-omic disease subtyping

Galadriel Brière[1,2*], Élodie Darbo[1,3], Patricia Thébault[1†] and Raluca Uricaru[1†]

[*]Correspondence: marie-galadriel.briere@u-bordeaux.fr
[1]Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, UMR 5800, F-33400 Talence, France
Full list of author information is available at the end of the article
[†]Equal contributor

Abstract

Background: Facing the diversity of omic data and the difficulty of selecting one result over all those produced by several clustering methods, consensus strategies have the potential to reconcile multiple inputs and to produce robust results.

Results: Here, we introduce ClustOmics, a generic consensus clustering tool that we use in the context of cancer subtyping. ClustOmics relies on a non-relational graph database, which allows for the simultaneous integration of both multiple omic data and results from various clustering methods. This new tool conciliates input clusterings, regardless of their origin, their number, their size or their shape. ClustOmics implements an intuitive and flexible strategy, based upon the idea of *evidence accumulation clustering*. It computes co-occurrences of pairs of samples in input clusters and uses this score as a similarity measure to reorganise data into consensus clusters.

Conclusion: We applied ClustOmics to multi-omic disease subtyping on real TCGA cancer data from ten different cancer types. We showed that ClustOmics is robust to heterogeneous qualities of input partitions, smoothing and reconciling preliminary predictions into high quality consensus clusters, both from a computational and a biological point of view. In this regard, ClustOmics is not meant to compete with other integrative tools, but rather to make profit from their various predictions when no gold-standard metric is available to assess their significance.

Availability: ClustOmics' source code, released under MIT licence, as well as the results obtained on TCGA cancer data are available on Github: https://github.com/galadrielbriere/ClustOmics.

Keywords: Disease subtyping; Multi-omic data; Data integration; Consensus clustering

Background

Recent advances in biological data acquisition made it possible to measure a wide range of data. Polymorphism data, DNA methylation, RNA expression, Copy Number Variations and other "omic" data are now routinely observed and analysed. Each omic type has the potential to reveal different molecular mechanisms associated to a phenotype, and making use of all available omic data could decipher complex and multi-level molecular interactions. Though several integrative tools have been developed, all of them aiming to answer biological questions by using multiple available data sources, the issue of omic data integration is far from solved. Along with the issue of omic data heterogeneity and integration, scientists are also challenged

with the diversity of strategies and methods available to answer a same biological question, each approach having its own perks and benefits.

The question of cancer subtyping is particularly representative of this kind of issues. Disease subtyping aims at detecting subgroups of patients (samples) showing similar characteristics, by performing a clustering analysis. Even in the single-omic context such analysis can be challenging and numerous clustering strategies have been implemented and/or tested to this end: hierarchical clustering strategies, density-, distribution- or centroid-based strategies, supervised and unsupervised strategies, etc. As a matter of fact, the selection of a clustering method as well as of the optimal parameters to use, is generally tricky. Moreover, the various biological mechanisms that are involved may vary from one patient to another: each tumor is different and has its own characteristics, both in the tumor cells themselves and in their interaction with their environment. As these mechanisms are not circumscribed to a single molecular level, the detection of groups of patients showing similar characteristics across different omics is a key issue to enable personalised medicine, which aims to offer patients a treatment adapted to the characteristics of their tumors.

This motivated the development of new computational methods implementing different strategies to analyse simultaneously several omic datasets (for detailed reviews see [1, 2]). According to the classification proposed in [1], the *early integration* strategy consists in concatenating omic datasets in a large matrix and applying a clustering method conceived for single-omic data [3, 4]. On the other hand, *late integration* approaches first cluster each omic dataset independently and fuse single-omic clusterings into one multi-omic clustering [5, 6, 7]. Other approaches perform *intermediate integration*, fusing sample similarities across omics [8, 9, 10], using dimension reduction strategies [11, 12], or statistical modelling with Bayesian frameworks [13, 14, 15].

To tackle both issues mentioned above, multi-omic and multi-strategy integration, one may want to take multiple clustering results (using different data, methods and parameters) and fuse all of them into one *consensus clustering*. Such a consensus clustering could benefit from the complementary information carried by various omics data and capitalize upon each method strengths while fading their weaknesses. A manifest way to compute a consensus clustering would be to perform the intersection of the clustering results. However, the greater the number of clusterings to fuse, the smaller the intersection. Moreover, when clusterings show different numbers of clusters, the question of the intersection is not trivial. Therefore, the issue of consensus clustering requires further methodological developments.

In the context of multi-omic data integration, strategies that first cluster each omic dataset separately before fusing omic-specific partitions into a multi-omic clustering are referred as *late integration* strategies [5, 6, 7]. Late integration can then be considered as a specific type of *consensus clustering*, where input clusterings to be fused are generated from different datasets. The strength of such strategy is to theoretically allow the use of any single-omic clustering during the first step. Thus, algorithms that are dedicated to a specific omic data type can be used to generate trustworthy single-omic clusterings. However, in practice, the generation step of omic-specific-partitions is handled by the late integration clustering software and

the user does not have full control over the process, neither on the quality of the generated single-omic clusterings. [5, 15, 16].

To our knowledge, no solution has been specifically developed to integrate simultaneously, and in an exploratory manner, several omics and various clustering strategies, while this type of concurrent integration could reveal sharper patterns in the data.

Our Contribution

Here, we present ClustOmics, a new graph-based multi-method and multi-source consensus clustering strategy. ClustOmics can be used to fuse multiple input clustering results, obtained with existing clustering methods that were applied on diverse omic datasets, into one *consensus clustering*, whatever the number of input clusters, the number of individuals clustered, the omics and the methods used to generate the input clusterings.

We designed ClustOmics as an exploratory tool to investigate clustering results with the objective to increase the robustness of predictions, taking advantage of accumulating evidences. In order to allow the user to tackle a specific question and to explore relationship patterns within input clusterings and generated consensus, we store the data in a non-relational graph based database implemented with the Neo4j graph platform [17]. The use of a graph native database facilitates the storage, the query and the visualisation of heterogeneous data, hence allowing to develop a solution that is flexible to various integration strategies. Indeed, by fusing clusterings from different clustering methods, different data types, different experimental conditions, or several options at the same time, through the use of what we call *integration scenarios*, ClustOmics can address a wide range of biological questions.

To compute a consensus clustering from a set of input clusterings, two main strategies exist: object co-occurrence based approaches and median partition based approaches [18]. In the former strategy, consensus clustering is computed from a matrix counting co-occurrences of objects in the same clusters. The latter strategy focuses on finding a consensus clustering maximizing the similarity with the input partitions. Both strategies raise several non-trivial questions. The choice of a clustering algorithm and its tuning is not straightforward when working with a co-occurrence matrix. On the other hand, for the median partition based approach, the choice of a similarity measure is determinant. Nevertheless, for consensus clustering in a multi-omic, multi-method context, comparing co-occurrences of objects is more pertinent than comparing similarities between partitions.

The co-occurrence strategy implemented in ClustOmics (detailed in the following section) is based on *Evidence Accumulation Clustering* (EAC), first introduced by Fred A. and Jain A. [16]. The idea is to consider each partition as an independent evidence of data organisation and to combine them using a voting mechanism. Similarly to clustering methods that use a distance or similarity measure to compare objects, EAC considers the co-occurrences of pairs of objects in the same cluster as a vote for their association. The underlying assumption is that objects belonging to a *natural* cluster are more likely to be partitioned in the same groups for different data partitions. Thus, one can use the counting of the objects' co-occurrences in clusters as a pairwise similarity measure. This measure, summarizing the results

from the input clusterings, is a good indicator of the agreement between the partitions and allows to produce a new partitioning that can be qualified as consensus. Although computationally expensive, this strategy allows exploiting all clustering results, regardless of the number of clusters, their size and shape.

The original EAC strategy as proposed by Fred and Jain [16] uses input partitions obtained by running the K-means algorithm multiple times (≈ 200) with random initialisation of cluster centroids. From these partition results, a co-occurrence matrix is computed and a Minimum Spanning Tree algorithm is applied to find consensus clusters, by cutting weak links between objects at a threshold t defined by the user. The authors recommend that clusterings obtained for several values of t should be analysed. In ClustOmics, we developed a weighted modularization optimisation strategy to automatically select the best filtering threshold. Also, rather than generating input clusterings from running multiple times the same algorithm as proposed by Fred and Jain, here we make benefit from using various clustering strategies, each searching for different patterns and giving different insights to the data. This also allows the use of algorithms that are specialised for one omic.

ClustOmics was applied in the context of multi-omic cancer subtyping, with TCGA data from different cancer types and multiple omic datasets. Input clusterings were computed with several multi-omic clustering methods and then were fused in a consensus clustering. Further details on the strategy implemented in ClustOmics are given in section *Methods*. In order to assess the benefit of this novel method, we further explored the robustness of our consensus clustering with respect to the input clusterings, as well as its biological relevance (see section *Results* for details).

Methods

In this section we detail the four-phase strategy implemented in ClustOmics. We then describe the datasets and the metrics that were used to evaluate our new method. For the sake of simplicity, as in this paper ClustOmics was applied in the context of cancer subtyping, we will further refer to objects of interest as *Patients*. However, ClustOmics can be applied on different biological entities, like genes or cells for instance.

ClustOmics integration strategy

ClustOmics, which takes multiple clustering results as input, is organised in four main phases that are illustrated in Figure 1.

First, from a metadata file and from the input clusterings, a graph is instantiated, in which *Patients (P)* and *Clusters (C)* are depicted as nodes (step A). *Support Edges* and their associated labels are computed to summarize patient co-occurrences in the input clusters (step B). These two steps are performed only once for the input data (regardless of the integration scenario) in order to synthesise the entire information present in the input clusters.

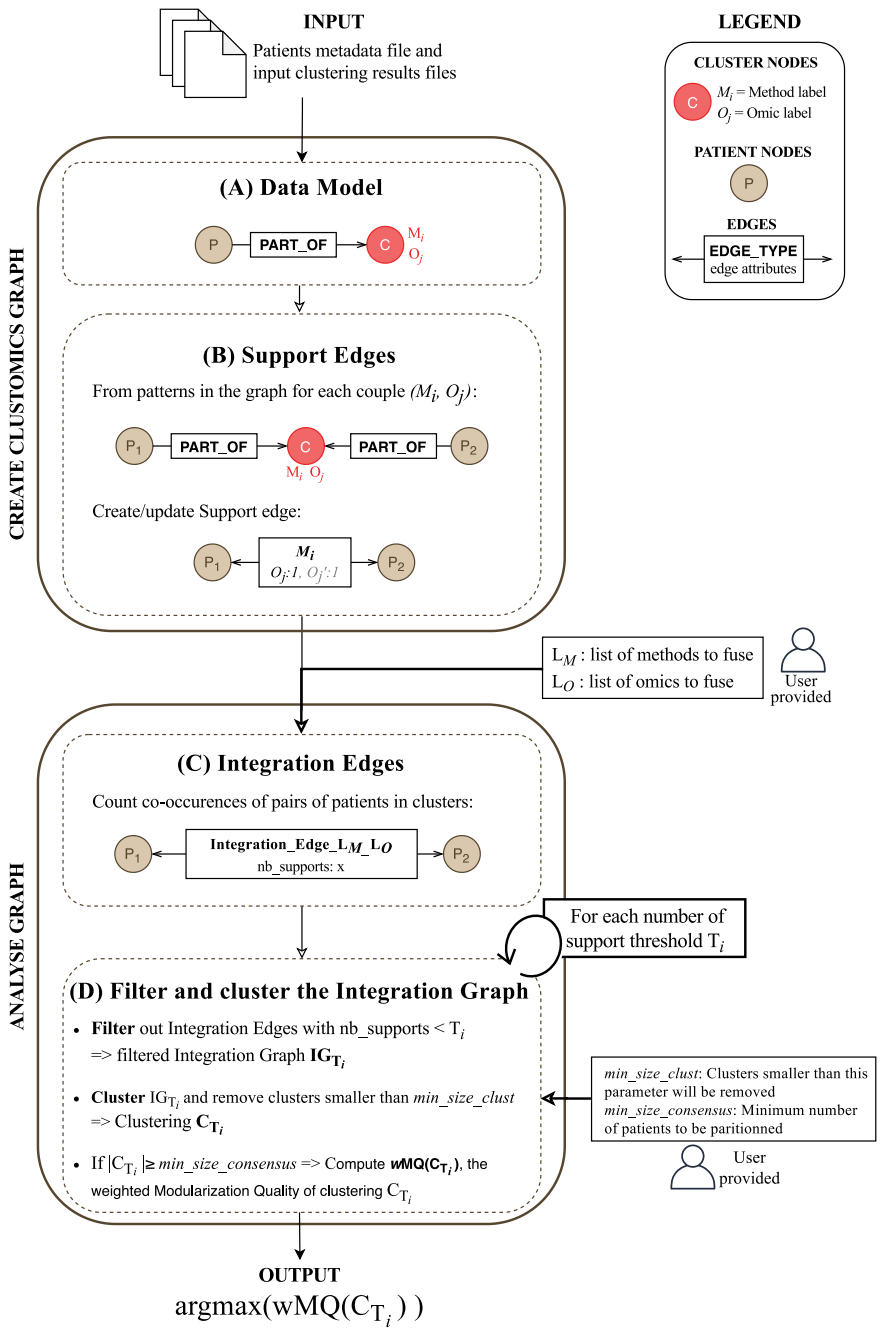
Next, given an integration scenario, meaning a list of omics and methods to integrate, the corresponding *Integration Graph (IG)* is computed (step C). This graph is meant to aggregate heterogeneous data into co-occurrence counts. An integration graph is specific to a given integration scenario. Indeed, when going from step

B to step C the origin of patient co-occurrences is lost and only the aggregated information needed to compute the consensus for this scenario is kept.

Finally, the integration graph is filtered and clustered to produce a consensus clustering for the given integration scenario (step D).

Note that, compared to steps A and B (computed once on the input data), step C is to be performed for each integration scenario being tested, and step D is to be tuned by the user. Below we detail each one of these steps.

Figure 1 An overview of the four-phase strategy implemented in ClustOmics.



(A) The data model.
In the context of cancer subtyping, the minimal version of ClustOmics' data model contains two types of nodes: *Patient* nodes and *Cluster* nodes with labels specifying the clustering method and the type of omic they were computed with. Each *Patient* node is linked to the different *Cluster* nodes to which it was assigned to (*PART_OF* relationship), as depicted in Figure 1-A.

(B) Compute Support Edges.
Each input clustering may differ in the way patients are grouped or in the number of clusters. In order to fuse different clusterings, we exploit the information of patients having been grouped together (co-clustered) by several clustering methods for one or several types of omics. On the basis of the *PART_OF* relationships, ClustOmics creates *Support Edges* among pairs of *Patient* nodes that summarize this information, *i.e.*, for each pair of patients and for each input clustering method $\mathcal{M}$, a *Support Edge* is added between the two *Patient* nodes if the method $\mathcal{M}$ co-clustered them together for at least one type of omic data (see Figure 1-B). *Support Edges* are relative to one input clustering method and store properties to indicate which omics support the patient association for the considered method. Thus a pair of patients can share several *Support Edges* and pairs of patients that were never co-clustered do not share any such edge.

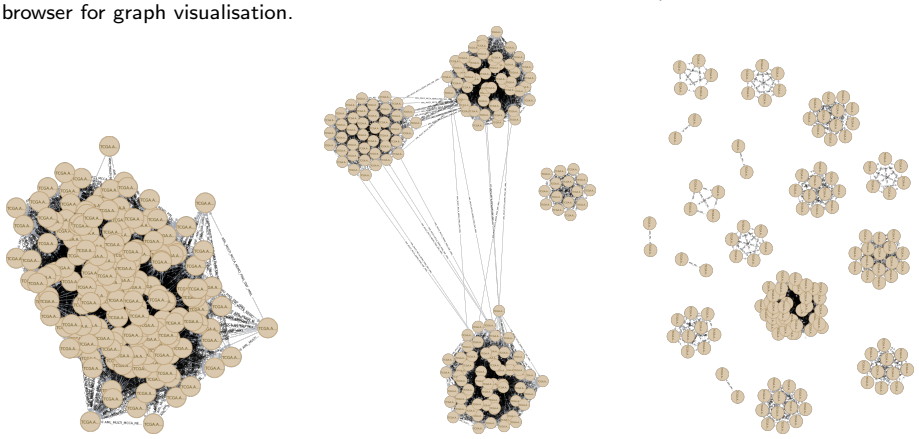
At the end of this step the graph is narrowed down as *Cluster* nodes are discarded, all the available information being summarised on the *Support Edges* and the way they connect the *Patient* nodes.

(C) Compute Integration Edges.
Support Edges are further used to compute an *Integration Graph*. Given an *Integration Scenario* (defined by a set of input clusterings to fuse), ClustOmics counts the number of input clusterings supporting a given association of patients (see Figure 1-C) and computes the so-called *number of supports*, an evidence accumulation count that can be seen as a similarity measure between individuals. The numbers of supports are reported on the *Integration Edges*. For a given integration scenario, a pair of patients may share at most one *Integration Edge*, summarizing the number of co-occurrences of the pair in the same clusters along the considered input clusterings. Thus a set of *Integration Edges* is specific to one integration scenario, defined by a set of omics and/or methods to fuse. This strategy allows to carry out different integration scenarios: integration of a subset of clustering methods, of omics, or both simultaneously. Two examples of integration scenarios are described in section *Results*.

(D) Filter and cluster the Integration Graph.
At this stage, ClustOmics has built an *Integration Graph*: the set of patient nodes are pairwise connected with integration edges summarizing the agreement of methods and omics to classify two patients in the same group. As each pair of nodes may have been clustered at least once in the same group over the set of omics and methods, the graph is generally complex and densely connected. However, as integration edges are weighted with the number of supports agreeing for an association, most robust

integration edges connecting patients that have co-clustered from multiple methods and omics can be distinguished from predictions that are not consistent across omics and methods. Hence, ClustOmics filters the graph according to the number of supports by removing non-consistent integration edges, and thus cleaning the noise in the integration data. The goal is to obtain a filtered graph foreshadowing *natural* clusters, meaning groups of patients that highly co-clustered with each others but poorly co-clustered with patients from other groups, and to cluster it into a consensus clustering. The choice of a threshold to filter integration edges is therefore determinant.

Figure 2 An integration graph filtered with increasing threshold values: 1, 3, 5 (the maximum number of supports for an integration edge being 5 in this example). Screenshots from Neo4j browser for graph visualisation.



(a) No filtering
197 nodes, 8122 edges

(b) $nb_supports \geq 3$
170 nodes, 3471 edges

(c) $nb_supports \geq 5$
148 nodes, 1244 edges

Figure 2 depicts the impact on the internal structure of the integration graph of an increasing number of supports filtering threshold. One can observe that two issues arise from this filtering process :

- First, increasing the threshold generates smaller and sparser graphs. Indeed, pairs of nodes that do not share any integration edge with a sufficient number of supports are removed, leading to a partial classification of the set of patients.
- The second issue is the loss of structure in the filtered integration graph: when filtering at a high threshold, the resulting graph can be too disconnected to be considered informative. For instance, the graph in Figure 2-c is composed of numerous small connected components, with nodes being organised in small groups of 2 or 3 patients. As clustering a graph with N connected components will result in at least N clusters of potentially even smaller sizes than those of the originating connected components, one can see that clustering of graphs alike the one presented in Figure 2-c will result in uninformative consensus clustering.

Therefore, getting a relevant classification requires finding the best compromise (thus the best support threshold) between reducing the graph to high confidence associations and preserving the structure of the data. For this, ClustOmics tests all

possible configurations by iteratively filtering the integration graph using increasing support thresholds. At each iteration, it uses state-of-the-art graph clustering methods to compute consensus clusterings for the corresponding filtered integration graph.

Moreover, in order to deal with the two issues described above and to keep merely informative results, the consensus clusterings go through an additional filtering step with the two following parameters:

- the *min_size_cluster* parameter indicates the minimum accepted size for a consensus cluster. When clustering the integration graph, all consensus clusters having a number of nodes below the *min_size_cluster* threshold are removed from the analysis.
- the *min_size_consensus* parameter indicates to what extent ClustOmics is allowed to discard nodes. ClustOmics will further consider only consensus clustering results composed of a sufficient total number of nodes.

Finally, a quality metric is computed for all consensus clusterings that were obtained after the filtering. The consensus clustering maximizing this metric is returned.

One must keep in mind that the resulting consensus clustering must be analysed with respect to the number of supports used to filter the graph prior to clustering. This number of supports indicates the level of agreement between the input clusterings and gives insight on the extent to which the resulting consensus clustering can be considered as being truly multi-omic.

Below we give insights on the graph clustering algorithms that are used to compute the consensus clusters, as well as on the quality metric employed for the identification of an optimal consensus clustering, the *weighted Modularization Quality*.

Graph clustering algorithms

ClustOmics filters the integration graph for each possible threshold on the number of supports and clusters it with two state-of-the-art, complementary graph clustering algorithms: the Louvain Community Detection (LCD) algorithm, based on modularity optimisation [19], and the Markov Clustering (MCL) algorithm, based on the simulation of stochastic flow in graphs [20].

Modularity optimisation is one the most popular strategy in graph clustering algorithms [21], while MCL/MCL-based methods have proven highly efficient in various biological network analysis (protein-protein interaction networks [22, 23], protein complex identification [24], detection of proteins families [25]). Moreover, it has been shown that modularity optimisation algorithms present a resolution issue [26, 27]: a tendency to fuse small clusters (even for those that are well defined and have few inter-connections) thus favoring the formation of bigger clusters than those computed by MCL [28]. Small clusters predicted by MCL can be an issue in ClustOmics case, as it removes clusters smaller than the user-defined *min_size_cluster* parameter, considering them to be non-informative. Hence, for each integration graph filtering threshold, ClustOmics generates two consensus clusterings: one computed with LCD, the other with MCL.

Selection of the best consensus clustering based on the Weighted MQ index
The *Modularization Quality* (MQ) was first defined by *Mancoridis et al.* in the context of software engineering [29]. Compared to the popular *Modularity* measure [30] optimised in graph clustering algorithms, which compares the distribution of edges with respect to a random graph with the same number of vertices and edges as the original graph, the *Modularization Quality* (MQ) evaluates the quality of a clustering as the difference between internal and external connectivity ratios. That is, the ratio between the number of connections observed within a given cluster and between two given clusters, and the maximum possible number of such edges. An optimal clustering for this measure should maximize the intra-connectivity ratio (every two nodes belonging to the same cluster share an edge) and minimize the inter-connectivity ratio (nodes classified in different clusters do not share edges). Indeed, in the context of consensus clustering based on evidence accumulation, it makes more sense to compare the distribution of the edges in the integration graph to the case where all nodes would have been partitioned in the same, optimal way in all input clusterings, *i.e.* a graph where all intra-cluster nodes are connected and all inter-cluster nodes are disconnected.

Moreover, we adapted the original MQ index for weighted, undirected graphs with no self loops (in our case a *Patient* node cannot share an integration edge with itself). We denote this adaptation of the Modularization Quality, as the *weighted Modularization Quality* (wMQ).

Let $G = (V, E)$ be a graph where V denotes the set of nodes and E the set of edges of G. Let $C = (C_1,\ldots, C_k)$ be a consensus clustering with k clusters and $|C_i|$ the number of nodes classified in cluster C_i. Let us also note $w(e)$ the weight of a given edge, $W(e_{ii})$ the sum of weights of the edges internal to C_i cluster (connecting vertices from C_i), $W(e_{ij})$ the sum of weights between clusters C_i and C_j (connecting a vertex from C_i to a vertex from C_j) and $max(w_{IG})$ the maximum weight observed on the edges of the given integration graph (the maximum possible number of supports, also corresponding to the number of input clusterings being fused). We therefore define the wMQ index computed for a consensus clustering C obtained on the integration graph IG as:

$$wMQ(IG, C) = \frac{1}{k \times max(w_{IG})} \sum_{i=1}^{k} \left(\frac{2 \times W(e_{ii})}{|C_i| \times (|C_i| - 1)} - \frac{1}{k-1} \sum_{j \neq i} \frac{W(e_{ij})}{|C_i| \times |C_j|} \right)$$

The first term of the sum corresponds to the weighted internal connectivity ratio for a cluster C_i. Indeed, the sum of the internal edges weights $W(e_{ii})$ is adjusted with the maximum possible value of the sum of the edges linking a set of $|C_i|$ nodes, which would be reached if all C_i nodes were connected with $max(w_{IG})$ weighted edges. Note that for an undirected and no self-loop graph, the maximum number of edges in a subgraph of $|C_i|$ nodes is $\binom{|C_i|}{2}$. Similarly, the second term of the sum represents the weighted external connectivity ratio of a cluster C_i, given by the sum of the weights of the edges linking a node from cluster C_i to a node belonging to a cluster C_j ($\neq C_i$).

The wMQ values range from -1 to 1, where a wMQ of -1 corresponds to the case where there is no intra-cluster edge and all inter-cluster pairs of vertices are

connected with edges of weight $max(w_{IG})$. A wMQ of 1 corresponds to the case where no inter-cluster vertices are connected and all pairs of intra-cluster vertices are connected with edges of weight $max(w_{IG})$. A high-standard consensus clustering should maximize this index.

ClustOmics computes the wMQ for the LCD and MCL consensus clusterings obtained with various numbers of supports thresholds and having passed the filtering steps, and returns the one that maximises this quality measure.

Datasets and tools used for computing input clusterings

We used ClustOmics to predict cancer subtypes from gene expression, micro-RNA expression and DNA methylation datasets available in *The Cancer Genome Atlas* (TCGA) [31]. Our case-study is based on the same datasets as in *Rappoport and Shamir*'s review on multi-omic clustering methods [1]. The data covers ten cancer types: Leukemia (AML), Breast (BIC), Colon (COAD), Glioblastoma (GBM), Kidney (KIRC), Liver (LIHC), Lung (LUSC), Ovarian (OV), Sarcoma (SARC) and Skin (SKCM). For each cancer type, from 197 and up to 1098 patients were measured for at least one of the three omics (expression, miRNA and methylation), of which 170 up to 621 patients were measured for all three. More details on missing data per cancer type are given in Table 1.

Table 1 Number of patients measured per omic for each cancer type. Total: Number of patients measured for at least one omic; Multi-omic: Patients measured for all of the 3 omics; Exp+Met: Patients measured for expression and methylation only; Exp+miRNA: Patients measured for expression and miRNA only; Met+miRNA: Patients measured for methylation and miRNA only; Exp: Patients measured for expression only ; Met: Patients measured for methylation only; miRNA: Patients measured for miRNA only.

Cancer	Total	Multi-omic	Exp+Met	Exp+miRNA	Met+miRNA	Exp	Met	miRNA
AML	197	170	0	3	15	0	9	0
BIC	1096	621	159	132	2	181	1	0
COAD	303	220	57	0	0	8	18	0
GBM	578	274	4	245	3	5	4	43
KIRC	534	183	135	71	0	144	1	0
LIHC	377	367	4	0	5	0	1	0
LUSC	501	341	29	1	0	130	0	0
OV	591	287	7	0	166	9	122	0
SARC	261	257	2	0	2	0	0	0
SKCM	368	351	16	1	0	0	0	0

To generate input clusterings to be fused by ClustOmics, we used 5 state-of-the-art integrative clustering tools, summarized in Table 2: PINS [5], SNF [8], NEMO [9], rMKL [10] and MultiCCA [11]. The first 4 tools can be used in a single-omic context as well as in multi-omic context, for which they were all designed. In addition, NEMO is the only one that can handle partial data. Each tool has been run with default parameters and following authors' recommendations.

Table 2 Methods used to compute input clusterings.

Software	Multi-omic context	Single-omic context	Handles missing data
PINS [5]	Yes	Yes	No
SNF [8]	Yes	Yes	No
NEMO [9]	Yes	Yes	Yes
rMKL [10]	Yes	Yes	No
MultiCCA [11]	Yes	No	No

Biological metrics

To explore the biological relevance of input clusterings and consensus clustering results, we computed the *overall survival rate* of patients. As cancer acuteness is proved to be related to its molecular subtype [32, 33, 34], we further investigated whether it was significantly different across clusters using the log-rank test. For each clustering, we computed the p-value of the log-rank test [35]. To increase the robustness of the results, we applied a bootstrapping strategy as described in [1], computing the log-rank test on randomly permuted data to derive an empirical p-value.

In addition, we performed an analysis of *clinical labels enrichment* in clusters, using 32 labels available from TCGA metadata (see Table 3). The idea is that patients affected by the same cancer subtype should also share, to a certain extent, the same clinical characteristics. The abundance of clinical labels in clusters and their statistical over-representation provide information on the biological robustness of clusterings. To perform this analysis, we used pan-cancer (e.g. age at diagnosis or pathologic stage of cancer) and cancer-specific clinical labels for each cancer type (e.g. presence of colon polyps for colon cancer, or smoking history for lung cancer). Clinical labels that were absent for more than half of the patients were removed from the analysis. We used the χ^2 test for independence for discrete parameters and the Kruskal-Wallis test for numeric parameters to assess the enrichment of the clinical labels in a cluster, by performing patient/cluster permutations as for the survival analysis.

Table 3 Pan-cancer and cancer-specific labels used for clinical label enrichment analysis. Pathological M, N and T labels refer to the TNM staging system, which describes the anatomical extent of tumor cancers [36].

Pan-cancer	Age at initial pathologic diagnosis, Gender, Pathologic M, Pathologic N, Pathologic T, Pathologic stage, Histological type, New neoplasm event type, Neoplasm histologic grade
AML	CALGB cytogenetics risk category, FAB morphology code
BIC	PAM50Call RNAseq, Estrogen receptor status, Progesterone receptor status, ER level cell percentage category, PR level cell percentage category
COAD	Presence of colon polyps, History of colon polyps
GBM	Prior glioma
KIRC	Hemoglobin result, Platelet qualitative result, Serum calcium result, White cell count result
LIHC	Adjacent hepatic tissue inflammation extent type, Albumin result specified value, Creatinine value, Fetoprotein outcome value, Fibrosis ishak score
LUSC	Tobacco smoking history, Number pack years smoked
OV	No supplementary clinical label
SARC	No supplementary clinical label
SKCM	Melanoma Clark level value, Melanoma ulceration indicator

One has to keep in mind that molecular data does not always explain survival or clinical differences between groups of samples. Therefore, in the following discussion of results we consider survival and clinical analysis as ways to interpret patterns captured by the various clustering results, and do not favor one metric over the other.

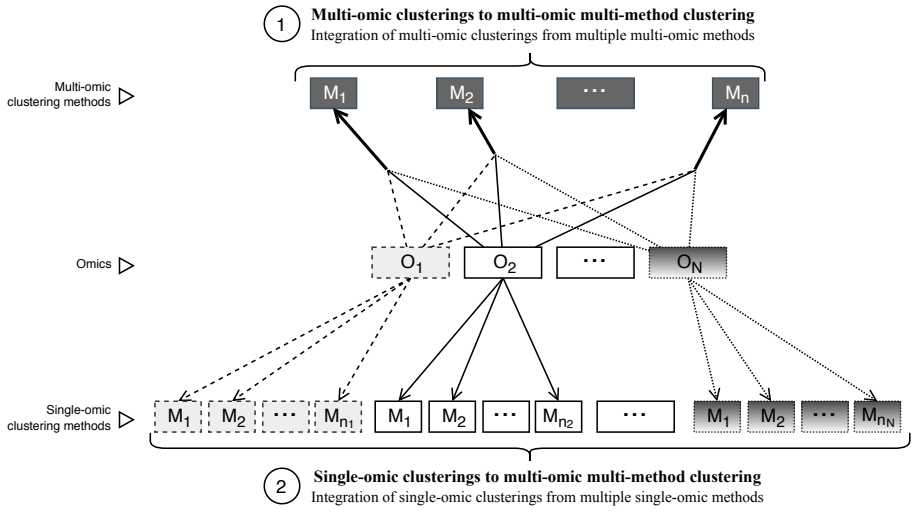
Results

Integration Scenarios

When performing consensus clustering in a multi-omic context, two main strategies can be implemented: an *a priori* solution computing consensus of omic-specific input clusterings, or an *a posteriori* computation of a consensus from multi-omic input clusterings. In order to better understand the perks and benefits of fusing omic data in a way or another, ClustOmics was tested in these two contexts with two integration scenarios.

First, we used ClustOmics to fuse multi-omic clusterings computed from existing integrative methods. In this scenario (Multi-to-Multi), the integration of omics is performed by various existing clustering tools and ClustOmics' integration computes a consensus result of all multi-omic clusterings produced. The second scenario (Single-to-Multi) involves both methods and omics integration, as only single-omic clusterings computed from various methods are fused into one consensus clustering. See Figure 3 for a visual representation of these two scenarios.

Figure 3 Two integration scenarios: Multi-to-Multi consensus clustering and Single-to-Multi consensus clustering. Arrows are dashed according to the omic each input clustering method consider.



In the sections below, in addition to describing the consensus multi-omic clusterings produced with ClustOmics for both integration scenarios, we explored their biological relevance based on the clinical and the survival metadata available for each patient.

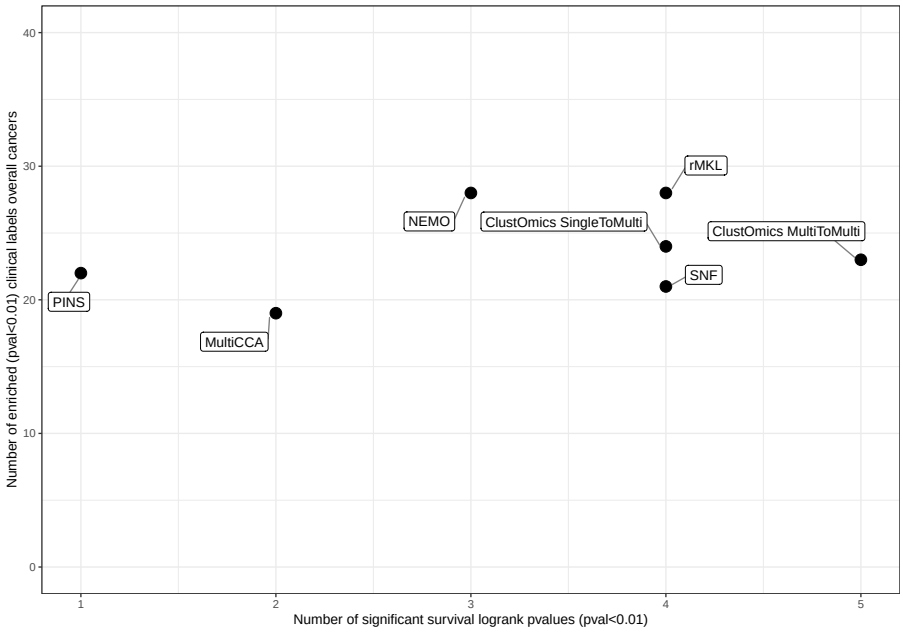
Results overview upon the ten cancer types analysed

We analysed the results produced by ClustOmics with respect to the quality of the input clusterings computed with five state-of-the-art clustering tools (see section *Datasets and tools used for computing input clusterings*), for the two scenarios described above. For this, we first computed multi and single-omic input clusterings. For the Multi-to-Multi scenario, multi-omic input clusterings were obtained using

existing multi-omic clustering tools. For the Single-to-Multi scenario, we used the same tools on each omic dataset independently, except for MultiCCA, which can only be used in a multi-omic context (see Table 2).

A survival and clinical label enrichment analysis was conducted on ClustOmics multi-omic consensus clusterings as well as on the single-omic and multi-omic input clusterings. An overview of the results upon the ten cancer types is displayed in Figure 4.

Figure 4 Overview of survival and clinical label enrichment results for the ten cancer types analysed. The x axis represents the number of significant survival p-values (< 0.01) found for each clustering overall the ten cancer types. The y axis represents the total number of significantly enriched clinical labels (p-values < 0.01), all cancer types included. In total, 79 enrichment p-values were computed from 32 distinct clinical labels.



In terms of clinical label enrichment in clusters, the results are relatively homogeneous, with 19 up to 28 clinical labels being enriched for all tools (over 79 enrichment p-values computed from 32 distinct clinical labels). The majority of clinical labels found enriched in ClustOmics consensus clusters were also found enriched for at least one input clustering, and clinical labels stably enriched in input clusterings were also found enriched in ClustOmics consensus clusterings.

On the other hand, the survival analysis results show high heterogeneity. The heterogeneity of input clustering qualities supports the idea of computing a consensus clustering, especially when no gold-standard metric or ground-truth data are available. In this sense, it is important to stress out that ClustOmics succeeded to compute biologically relevant consensus partitions from input clusterings of variable quality. Indeed, in the Multi-to-Multi case, ClustOmics managed to find five out of ten significant log-rank p-values, counterbalancing PINS and MultiCCA mitigated results, although they were part of the input clustering results used for this integration scenario. For the Single-to-Multi scenario, ClustOmics seems slightly less efficient, but still produced quality results for both biological metrics compared

to other multi-omic clusterings produced from state-of-the-art methods. Quality results for omic-specific input clusterings used for this scenario do not appear in Figure 4, but are further detailed in Section *Integration of single-omic clusterings*. Note that in the Single-to-Multi scenario, patients that were not measured for one or two of the three omics were also classified by ClustOmics, which is the only tool except NEMO capable of managing such partial data.

Detailed results for each integration scenario are given in the following sections. A focus on Single-to-Multi consensus clustering produced on the breast cancer dataset is also to be found below.

Integration of multi-omic clusterings (scenario Multi-to-Multi).

Input multi-omic clusterings were computed with the five multi-omic clustering methods presented above using default parameters and following authors' recommendations. The clusterings were produced using exclusively the multi-omic patients (those for which all three omics data were available), except for NEMO that handles missing data and for which all patients were considered.

ClustOmics was run with the *min_size_cluster* parameter arbitrarily set to 8 nodes for all cancer types, meaning that clusters of size below 8 were removed from the consensus clustering. We also set the *min_size_consensus* parameter to preserve at least 95% of the patients included in all input clusterings (that is 95% of the patients that were measured for the 3 omics).

Table 4 Scenario Multi-to-Multi - Total multi-omic (measured for the three omics) population size, number of patients effectively clustered by ClustOmics, number of supports used to filter the graph, number of consensus clusters generated by ClustOmics and average number of clusters in input clusterings.

Cancer	Multi-omic population size	# patients clustered	# supports threshold	# clusters in ClustOmics	Avg # clusters in input clusterings
AML	170	163	4	6	4.4
BIC	621	621	2	2	3.8
COAD	220	217	3	6	5.6
GBM	274	262	4	7	6.25
KIRC	183	176	2	2	5.6
LIHC	367	354	4	8	3.8
LUSC	341	341	3	3	4.4
OV	287	282	4	7	4
SARC	257	250	4	4	3.4
SKCM	351	350	3	5	4.8

Quantitative global measures on ClustOmics consensus clusterings are detailed in Table 4. Note that the maximum number of supports on an integration edge is bounded to 5, as five input clusterings (computed from five integrative clustering tools) were used for this integration scenario. For each cancer type, the integration graph was filtered with a number of supports ranging from 2 to 4, meaning that only pairs of patients clustered in the same multi-omic cluster by at least 2 to 4 clustering methods were taken into account to compute the consensus clustering.

ClustOmics succeeded to cluster all patients for two cancer types, BIC and LUSC. For the other cancer types, up to 13 patients (LIHC) were removed, either during the integration graph filtering stage or because the consensus cluster they were assigned to was too small (< 8 nodes).

When comparing input and consensus clusterings, there is a certain consistency between the average number of clusters in input partitions and the number of consensus clusters found by ClustOmics, which shows a coherent behaviour of ClustOmics. However, consensus clusterings of BIC and KIRC distinguish only two consensus clusters, while respectively 3.8 and 5.6 clusters appeared in the corresponding input clusterings, meaning that ClustOmics was not able to preserve all patterns observed in input clusterings. This resolution drop is due to the low number of supports (*i.e.* 2) that was found to be optimal with respect to the wMQ when filtering the integration graphs. This low filtering threshold indicates that input partitions for BIC and KIRC cancer types show low similarity, suggesting that input clustering algorithms revealed different omic patterns from the same datasets.

Figure 5 Survival analysis results for ClustOmics Multi-to-Multi consensus clustering and for each input multi-omic clustering. The horizontal dashed line indicates the threshold for significantly different survival (p-value ≤ 0.01). Boxplots were computed considering only input clusterings, then ClustOmics results were added to the plot.

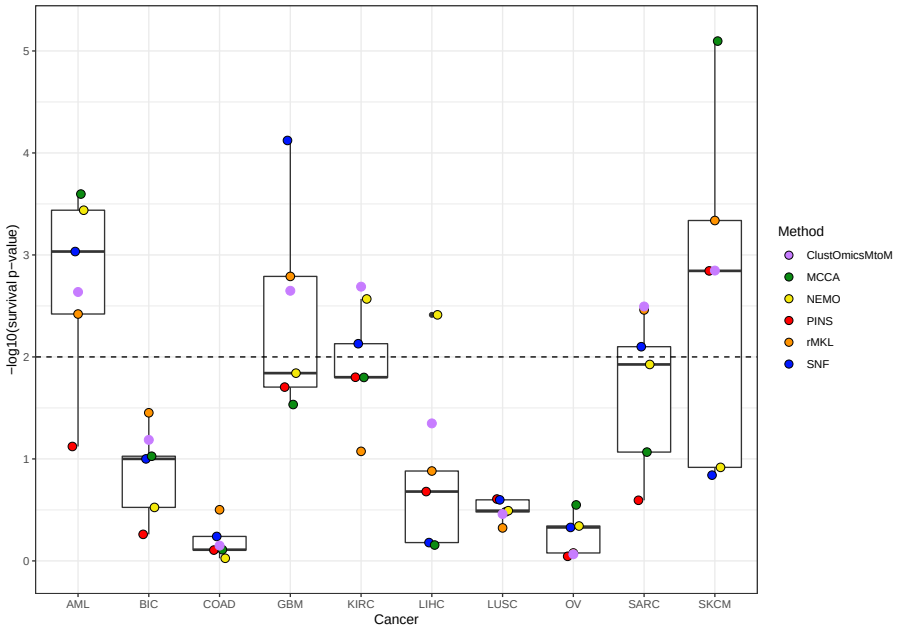


Figure 5 presents survival analysis results for the various multi-omic clusterings given as input to ClustOmics Multi-to-Multi, and for the resulting consensus clustering. When looking at the input clustering survival results, we can differentiate two cases:

- For AML, BIC, GBM, KIRC, LIHC, SARC and SKCM, input clusterings show a quite high heterogeneity in terms of survival quality
- For COAD, LUSC and OV cancer types, input clusterings show homogeneous survival quality

For the first group of cancer types, the heterogeneity of input clusterings survival qualities indicates how the choice of one clustering method can drastically impact the results. For these cancer types, the benefit of using ClustOmics is manifest, as it reconciles input partitions in consensus clusterings of survival quality higher than what was initially obtained. Indeed, except for AML, ClustOmics computed

consensus clusterings with survival qualities higher than the median of the input clusterings. Thus, given input clusterings of various quality, ClustOmics is able to extract the most stable patterns across input partitions, based on evidence accumulation.

When input partitions show homogeneous survival quality, ClustOmics gives similar results, which is an expected behaviour. Though in this case ClustOmics has less added value, it is important to note that computing consensus clustering does not result in aberrant classification in view of input clustering results.

Table 5 Clinical labels found enriched in Multi-to-Multi (MtoM) scenario consensus clusters, in Single-to-Multi (StoM) scenario consensus clusters, and for both scenarios.

Cancer	Scenarios	Enriched clinical labels
AML	Both	Age at initial pathologic diagnosis CALGB cytogenetics risk category Leukemia French–American–British Morphology Code
BIC	Both	Pathologic N, Pathologic Stage, Histological type, PAM50 call Estrogen receptor status, Progesterone receptor status
	StoM	Age at initial pathologic diagnosis Pathologic M, Pathologic T
COAD	MtoM	Histological type
	StoM	Age at initial pathologic diagnosis
GBM	MtoM	Age at initial pathologic diagnosis, New neoplasm event type
KIRC	MtoM	Platelet qualitative result, Serum calcium result
	StoM	Pathologic M, Pathologic T, Neoplasm histologic grade
LIHC	Both	Gender, Age at initial pathologic diagnosis, Fetoprotein outcome value
	MtoM	Neoplasm histologic grade
LUSC	MtoM	Histological type
	StoM	Gender, Age at initial pathologic diagnosis
OV	Both	None
SARC	Both	Gender, Age at initial pathologic diagnosis, Histological type
	MtoM	New neoplasm event type
SKCM	Both	None

Consensus clusterings were also investigated for clinical labels enriched in clusters (listed in Table 5). For AML for example, ClustOmics computed clusters enriched for the CALGB cytogenetics risk category, a risk classification based on the Cancer and Leukemia Group B clinical trial [37], and for the French-American-British (FAB) morphology code, a clinical classification for AML tumors [38]. Reassuringly, BIC consensus clustering was found enriched for the PAM50 classification, a widely used breast-cancer subtype predictor [39].

Integration of single-omic clusterings (scenario Single-to-Multi).

To assess ClustOmics performance when fusing simultaneously input clusterings computed from different omic data and with different clustering methods, we investigated a second integration scenario, combining single-omic clusterings produced independently on each omic dataset. Overall cancer consensus results for this scenario are displayed in Figure 4, but discussed in detail in this section.

Single-omic clusterings were computed using the same clustering tools as in the first integration scenario, except for the MultiCCA tool, as it can not be applied to single-omic data. PINS, SNF, rMKL and NEMO were run on each omic dataset independently, while considering all available patient data (including patients which were not measured for the three omics). This implies that considered patients are not necessarily the same across input clusterings and the total size of the population can be significantly bigger than the multi-omic population size (see Table 1). For instance, the multi-omic population size for BIC cancer is 621, while 1093 patients were measured for gene expression and appear in input clusterings for this omic.

Here, the maximum number of supports is 12 as the 4 clustering methods were run on 3 omic datasets for each cancer. Note that for this scenario the threshold on the number of supports used to filter the integration graph has great influence on the capacity of ClustOmics to produce consensus clusters across omics and on the interpretation of results. Indeed, the threshold has to be greater than or equal to 5 in order to ensure that all the conserved integration edges rely on an association consistent across at least two different omics (one omic being represented by 4 input clusterings). To ensure that all integration edges are built upon all 3 omics, the threshold must be 9 or higher. In fact, one should estimate an acceptable threshold depending on the experimental design and the biological question to address.

In our case, as we did not wish to bring any *a priori* on which omic should have a stronger impact on the results (indeed one omic data type could particularly well explain the disparities in molecular profiles of patients for a cancer type, but not for the others), we considered a number of supports of 5 to be sufficient to ensure that selected integration edges are either moderately consistent across the 3 omics, either strongly consistent in one omic. Together with the constraint to preserve at least 95% of the multi-omic population (*min_size_consensus* parameter), this gave a number of supports used to filter the integration graph ranging from 5 to 7.

Moreover, for all cancer types, we set the *min_size_cluster* parameter to 5% of the multi-omic population. This parameter was set to 8 nodes for the Multi-to-Multi integration scenario, but here we chose to increase the minimum size of a cluster as the total number of patients is higher in this integration scenario.

Table 6 Scenario Single-to-Multi - Total population size (and multi-omic population size), number of multi-omic patients and number of not multi-omic patients effectively clustered by ClustOmics (and percentage of the total population that were clustered), number of supports used to filter the graph, number of clusters generated by ClustOmics and average number of clusters in input clusterings.

Cancer	Total population size (multi-omic population size)	# multi-omic + other patients clustered (%total population size)	# supports threshold	# clusters ClustOmics	Avg # clusters input clusterings
AML	197 (170)	153 + 15 (85,28%)	6	7	7.58
BIC	1096 (621)	565 + 189 (68,80%)	7	6	4.33
COAD	303 (220)	220 + 56 (91,09%)	5	5	3.83
GBM	578 (274)	259 + 177 (75,43%)	7	9	4.58
KIRC	534 (183)	182 + 194 (70,41%)	6	8	4.58
LIHC	377 (367)	364 + 9 (98,94%)	5	3	5.33
LUSC	501 (341)	335 + 29 (72,65%)	5	3	5.33
OV	591 (287)	273 + 136 (69,20%)	7	10	3.83
SARC	261 (257)	257 + 4 (100,00%)	5	3	4.83
SKCM	368 (351)	325 + 15 (92,39%)	6	8	5.25

One of the major benefits of this integration scenario (besides the fact that single-omic clusterings are easier to compute) is its ability to cluster individuals that did not appear in all input partitions. Indeed, Table 6 shows that for all cancer types, ClustOmics succeeded to cluster patients that were not measured for all of the 3 omic types, which is not possible for most existing multi-omic clustering tools.

Input clusterings can show great similarity for a given omic. If this omic allows to differentiate groups of individuals in a clear-cut way, it will drive consensus clustering. However, if the omic is less relevant to partition patients, input clusterings are more likely to show different patient associations. Such clusterings add noise-like integration edges in the integration graph, with low number of supports on edges. Therefore, we expect each omic to have a different impact on the final consensus clustering. To evaluate the impact of omic-specific input clusterings on the consensus result, we computed the Adjusted Rand Index (ARI) [40], a measure of similarity between partitions. Each ClustOmics consensus clustering was compared to each of its input clusterings (see Figure 6-A). While ARI has been used to evaluate the quality of classifications compared to ground-truth data, here we use it to compare the similarity of various input clusterings with ClustOmics consensus, without considering any quality aspect. However, we relate ARI pairwise similarity between input clusterings and consensus clustering to survival results (see Figure 6-B) to interpret consensus clusterings survival quality.

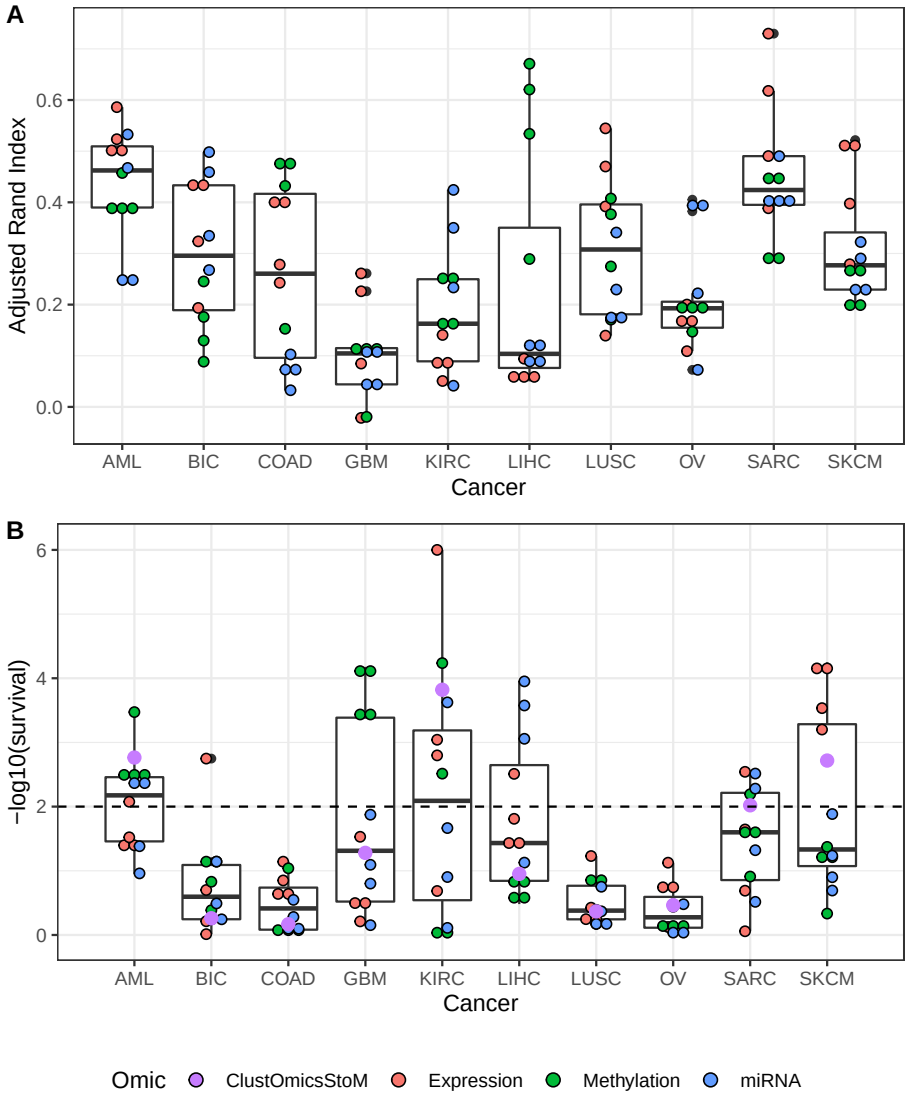
Interestingly, gene expression input clusterings have uneven impact on the consensus clustering across cancer types, miRNA clusterings show a higher impact on BIC and KIRC consensus clusterings and methylation clusterings drive the consensus on COAD and LIHC. SARC consensus clusterings show ARI values ranging from 0.4 up to 0.7 for at least one input clustering computed from each of the 3 omics datasets, suggesting similar associations at each molecular level.

Survival analysis for this integration scenario shows two groups of cancer types, as already noticed for the Multi-to-Multi scenario. For BIC, COAD, LUSC and OV, gene expression, methylation and miRNA input clusterings show homogeneous survival p-values. For these cancer types, ClustOmics computed a consensus clustering with similar quality scores. For the cancer datasets showing higher heterogeneity among input clusterings survival quality, ClustOmics found significant survival p-values for AML, KIRC, SARC and SKCM, despite the low quality clusterings that were given as input. Interestingly, for the LIHC dataset, ClustOmics clustering was highly influenced by methylation input clusterings (see Figure 6-A), while miRNA input clusterings were the most significant for the survival analysis (see Figure 6-B).

Clinical labels found enriched in consensus clusters are listed in Table 5. AML clusters were found enriched for both CALGB and FAB classifications, KIRC clusters for histologic grade and SARC clusters for histological type. While BIC consensus clustering did not show good survival-wise results, pathologic M, N and T labels (referring to the TNM classification of tumors [36]) were found enriched in clusters, as well as pathologic stage, histological type and PAM50 call.

In the following section we further explored the Single-to-Multi consensus clustering for BIC dataset.

Figure 6 Adjusted Rand Index of input clusterings relative to consensus multi-omic clustering (A) and Survival analysis results for ClustOmics Single-to-Multi consensus clustering and for each input omic-specific clustering (B). A:Each point corresponds to one input clustering, and is colored according to the omic data type used. Each omic dataset has been clustered using 4 different clustering tools (PINS, NEMO, SNF, rMKL), thus is represented by 4 input clusterings. **B**: The horizontal dashed line indicates the threshold for significantly different survival rate (p-value ≤ 0.01). Boxplots were computed considering only input clusterings, then ClustOmics results were added to the plot.



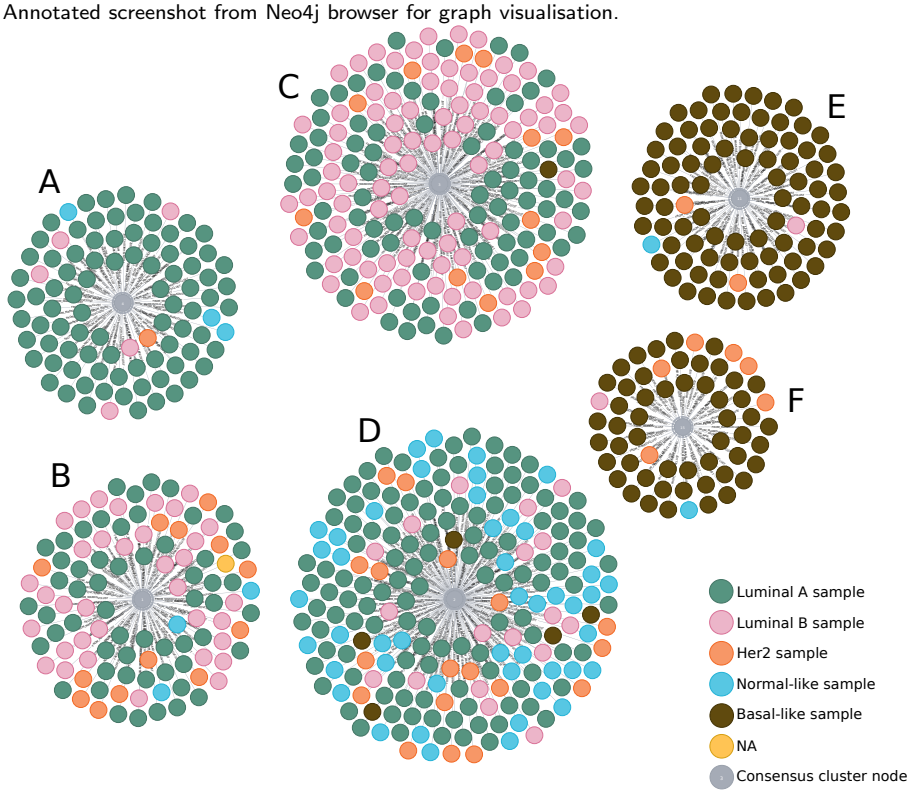
Study case: BIC Single-to-Multi consensus clustering

In this section we focus on the consensus clustering of the 12 single-omic clusterings for the BIC dataset and analyse these results in parallel to the PAM50 classification. As the PAM50 classification is computed from the expression of 50 specific genes, while in this work we capitalise on 3 different omics, a certain heterogeneity in the clusters when compared to the PAM50 prediction is expected. Moreover, this heterogeneity is to be further explored, as it could reveal subtypes that are not distinguishable when considering only PAM50 genes, but that are heterogeneous when integrating other data sources.

From the 1096 patients available in BIC dataset (of which only 621 patients are measured for the 3 omics), ClustOmics succeeded to classify 754 patients in consensus clusters of which: 565 were measured for the 3 omics, 108 for gene and miRNA expression only, 80 for gene expression and methylation only, and 1 was measured for miRNA expression and methylation only. These 754 individuals were partitioned in 6 consensus clusters with sizes ranging from 59 patients for the smallest one (the minimum allowed size for a cluster *min_size_cluster* being set to 5% of the multi-omic population, that is 31 individuals for BIC) and up to 212 patients for the biggest one.

As the PAM50 clinical labels were missing for 228 patients out of the 754, we applied the original classifier introduced by Parker *et al* [39] in order to call the missing labels. We further mapped the PAM50 calls to ClustOmics consensus clusters, and observed significant concordance, as depicted in Figure 7. Moreover, clinical labels enrichment analysis and additional tests applied to describe the clusters (see Table 7) show good mapping between ClustOmics clusters and key biological clinical labels, such as Estrogen and Progesterone Receptor status (ER/PR) and histological type of tumors.

Figure 7 BIC consensus clustering with patients colored according to the PAM50 prediction. Annotated screenshot from Neo4j browser for graph visualisation.



Indeed, consensus cluster labelled A in Figure 7 mostly gathers Luminal A samples, and consensus clusters E and F are mostly constituted of Basal samples. The other clusters show less concordance with the PAM50 classification, but specific PAM50 subtypes were also found enriched in these groups, luminal B samples more likely clustered in clusters B and C and luminal A samples in clusters A and D

(Table 7). This fuzzy mapping of PAM50 calls on consensus clusters is not surprising as it has been shown that separation of Luminal A and B samples was not reconstructed by RNA-seq unsupervised analysis [41]. Several studies also reported that the separation between Luminal subtypes was not consistent, suggesting that Luminal A and Luminal B samples may represent part of a continuum rather than distinct subgroups [41, 42, 43].

While consensus cluster D seems heterogeneous in terms of PAM50 labels, Invasive Lobular Carcinoma (ILC) histological subtype was found over-represented in the cluster, an histological subtype described in the literature as being detected at a more advanced stage than the other main histological subtype, the Invasive Ductal Carcinoma (IDC) [44]. Cluster D is also the only cluster to over-represent Stage III patients (Table 7).

Table 7 Over and under represented clinical labels within BIC consensus clusters. ER+/ER- and PR+/PR- respectively correspond to Estrogen Receptor status and Progesterone Receptor status, positive and negative. M, N, T stages refer to the TNM staging system.

Cluster	Over-represented labels	Under-represented labels
A	Luminal A ER+, PR+ T1, Stage I	Basal, Luminal B ER-,PR-
B	Luminal B ER+	Basal ER-
C	Luminal B ER+, PR+ M0 Infiltrating Ductal Carcinoma, Mucinous Carcinoma	Basal, Normal ER-,PR- MX Infiltrating Lobular Carcinoma
D	Luminal A, Normal ER+, PR+ MX, N3, T3, Stage III Infiltrating Lobular Carcinoma	Basal, Luminal B ER-,PR- M0 Infiltrating Ductal Carcinoma
E	Basal ER-, PR- Infiltrating Ductal Carcinoma, Medullary Carcinoma, Metaplastic Carcinoma	Luminal A, Luminal B ER+,PR+ Infiltrating Lobular Carcinoma
F	Basal N0 ER-, PR- Infiltrating Ductal Carcinoma, Invasive Ductal Carcinoma	Luminal A, Luminal B ER+,PR+

Furthermore, regarding the poor correlation between survival and ClustOmics consensus clustering for this dataset (see Figure 6-B), we further compared these results to the correlation between the PAM50 classification and survival. Thus, we conducted a survival analysis on the data from all the 1096 available patients with respect to their PAM50 classification and observed a significant survival p-value (0.04). However, when applying the same analysis on the reduced 754 patients set (the set of patients clustered by ClustOmics), we obtain a non-relevant p-value (0.26), as for ClustOmics consensus clustering survival analysis. This suggests a lack of data rather than a poor classification of patients in consensus clusters.

Discussion

The novel method we present in this paper deals with two key issues raised by the present context in biology and medicine and, in parallel, in bioinformatics. Indeed these domains witness an actual revolution in the acquisition of molecular data thus facing a flood of various types of omic data. The ultimate goal is to benefit from the diversity and complementarity of these omic data (DNA methylation, Copy Number Variations, polymorphism data, etc.), by analysing them simultaneously. However, multi-omics data integration is only a facet, as at the same time we face an outburst of biocomputational approaches meant to deal with this unprecedented variety and quantity of data, and the choice of a method or of the optimal parameters is generally challenging. In this paper both simultaneous integration of multiple omics and of various methods are tackled in an innovative manner through an original integration strategy based on consensus.

More particularly, in this work we address the cancer subtyping problem from a personalised medicine related perspective that is gaining increasing attention. In order to treat patients according to their disease profile, one should be able to distinguish between disease subtypes. These can be predicted from omic data (traditionally gene expression, but also methylation, miRNA, etc.) by performing patient clustering (hierarchical clustering, density-based clustering, distribution-based clustering, etc.). Our novel graph-based multi-integration method, can fuse multiple input clustering results (obtained with existing clustering methods on diverse omic datasets) into one consensus clustering, whatever the number of input clusters, the number of objects clustered, the omics and the methods used to generate the input clusterings.

To compute a consensus clustering, our method, implemented in a tool called ClustOmics, uses an intuitive strategy based on evidence accumulation. The evidence accumulation counts (*i.e.* the number of supports on the integration edges) make consensus clustering results easier to interpret, as they add insight to the extent to which the consensus clustering can be considered as being multi-source (issued from multiple omics) and to the overall agreement of input partitions. While in this paper we chose to favor the use of several multi-omic clustering methods for the generation of input clusterings, the evidence accumulation clustering strategy is highly flexible and can be used with omic-specific clustering methods. Moreover, by taking as input a high number of clusterings obtained with a same tool with varying parameters, the convergence of the consensus clustering, especially in a single-omic context, can be improved.

TCGA real datasets from 3 different omics and 10 cancer types were analysed with respect to two integration scenarios : (1) fusing multi-omic clusterings obtained with existing integrative clustering tools and (2) fusing omic-specific input clusterings. In both cases, ClustOmics succeeded to compute high quality multi-omic consensus clusterings with clusters showing different survival curves and enriched for clinical labels of interest, coherent with what could be found in the cancer literature. Moreover, results indicate that ClustOmics is robust to heterogeneous input clustering qualities, reconciling and smoothing the disparities of partitions.

Notably, our method is not meant to compete with existing single or integrative omics clustering methods but, on the contrary, it aims at capitalizing on the preliminary input predictions in order to increase their robustness by taking advantage

of accumulating evidence to reveal sharper patterns in the data. This selection of robust patterns across input partitions renders ClustOmics stable when facing heterogeneous input clusterings, and particularly useful when no gold-standard metric is available to assess the quality of the results. Hence, with a sufficient number of input clusterings, no prior analysis of the input is needed, given that low quality clusterings, likely to add noise to the integration graph, play a smaller part in the evidence accumulation. Omics for which the separation of samples is clear-cut will drive the consensus clustering while omics that do not show interesting patterns across samples will be faded via the integration graph filtering step. For the same reason, it is important to highlight that, as long as the signals in the available omics are strong, ClustOmics is able to cluster samples that do not appear in all omic datasets, making use of available data and addressing the issue of partial data.

Finally, though presented in a disease subtyping context, one should grasp that our method is not limited to this application case. Indeed taking partition results (instead of raw data) as input, makes it generic and adaptable to a wide range of biological questions, as one can use any kind of partitioning of the data, *e.g.* clinical labels, groups of genes of interest, etc., as an input clustering. A major strength of ClustOmics resides in its exploratory aspect, resulting on one side from a flexible intrinsic model that gives the user complete power on the integration scenario to investigate and, on the other side, from the use of the graph-oriented database Neo4j. All input data and metadata are stored in this kind of database, which may be easily queried and visualised by a non-specialist with the Neo4j browser.

Conclusion

Facing the diversity and heterogeneity of omic data and clustering strategies, one might want to make profit from all available data to compute a consensus clustering. ClustOmics is able to fuse any set of input clusterings into one robust consensus, which can easily be interpreted based on the number of supports evidence accumulation score. ClustOmics can be adapted to answer a wide range of biological questions. The use of integration scenarios allows users to explore various integration strategies, by adding or discarding data sources and/or clustering methods.

Implementation

ClustOmics is implemented based on the Neo4j graph database management system and uses APOC and Graph Data Science Neo4j libraries. Queries on the graph database are performed in Cypher, Neo4j's graph query language, and are encapsulated in Python scripts. To facilitate its use, ClustOmics can be run through the Snakemake workflow management system.

ClustOmics was tested on a desktop-computer with an Intel Xeon processor (2.70GHz, 62GB of RAM) running on Ubuntu 18.04. For the TCGA real datasets it was applied to, ClustOmics runtimes range from a few minutes for small datasets (AML, Multi-to-Multi scenario) up to 2 hours for the largest dataset (BIC, Single-to-Multi scenario), most of the computation time being consumed for the construction of the integration graph. The graph being stored in a Neo4j database, this step is only to be performed once for each integration scenario and parameters for graph filtering can be further set and tuned without re-computing the graph.

ClustOmics source code, released under MIT licence, as well as the results obtained on the 10 cancer types with the two integration scenarios described in this paper are available on Github: https://github.com/galadrielbriere/ClustOmics.

Competing interests
The authors declare that they have no competing interests.

Author's contributions
GB, PT and RU conceived the project and the method on which ClustOmics is based. GB developed the ClustOmics tool and performed the bioinformatics experiments. All authors participated at the analysis of the results. All authors wrote, read and approved the final manuscript.

Acknowledgements
This work was supported by the INS2I PEPS Blanc NEOMICS project and the LaBRI.

Author details
[1]Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, UMR 5800, F-33400 Talence, France. [2]Univ. Bordeaux, INRA, Bordeaux INP, NutriNeuro, UMR 1286, F-33000 Bordeaux, France. [3]Institut Bergonié, Univ. Bordeaux, INSERM U1218, F-33076 Bordeaux, France.

References
1. Rappoport, N., Shamir, R.: Multi-omic and multi-view clustering algorithms: review and cancer benchmark. Nucleic Acids Research **46**(20), 10546–10562 (2018). doi:10.1093/nar/gky889. Accessed 2019-11-21
2. Tini, G., Marchetti, L., Priami, C., Scott-Boyer, M.-P.: Multi-omics integration—a comparison of unsupervised clustering methodologies. Briefings in Bioinformatics **20**(4), 1269–1279 (2019). doi:10.1093/bib/bbx167. Accessed 2019-12-11
3. Wu, D., Wang, D., Zhang, M.Q., Gu, J.: Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: application to cancer molecular classification. BMC genomics **16**, 1022 (2015). doi:10.1186/s12864-015-2223-8
4. Wang, H., Nie, F., Huang, H.: Multi-view clustering and feature learning via structured sparsity. In: Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28. ICML'13, pp. 352–360. JMLR.org, Atlanta, GA, USA (2013)
5. Nguyen, T., Tagett, R., Diaz, D., Draghici, S.: A novel approach for data integration and disease subtyping. Genome Research **27**(12), 2025–2039 (2017). doi:10.1101/gr.215129.116. Accessed 2019-11-21
6. Hoadley, K.A., Yau, C., Wolf, D.M., Cherniack, A.D., Tamborero, D., Ng, S., Leiserson, M.D.M., Niu, B., McLellan, M.D., Uzunangelov, V., Zhang, J., Kandoth, C., Akbani, R., Shen, H., Omberg, L., Chu, A., Margolin, A.A., Veer, L.J.v., Lopez-Bigas, N., Laird, P.W., Raphael, B.J., Ding, L., Robertson, A.G., Byers, L.A., Mills, G.B., Weinstein, J.N., Waes, C.V., Chen, Z., Collisson, E.A., Network, T.C.G.A.R., Benz, C., Perou, C.M., Stuart, J.M.: Multi-platform analysis of 12 cancer types reveals molecular classification within and across tissues-of-origin. Cell **158**(4), 929 (2014). doi:10.1016/j.cell.2014.06.049. Publisher: NIH Public Access. Accessed 2020-09-21
7. Bruno, E., Marchand-Maillet, S.: Multiview clustering: A late fusion approach using latent models. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '09, pp. 736–737. Association for Computing Machinery, New York, NY, USA (2009). doi:10.1145/1571941.1572103. https://doi.org/10.1145/1571941.1572103
8. Wang, B., Mezlini, A.M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., Goldenberg, A.: Similarity network fusion for aggregating data types on a genomic scale. Nature Methods **11**(3), 333–337 (2014). doi:10.1038/nmeth.2810. Accessed 2019-11-21
9. Rappoport, N., Shamir, R.: NEMO: cancer subtyping by integration of partial multi-omic data. Bioinformatics **35**(18), 3348–3356 (2019). doi:10.1093/bioinformatics/btz058. Accessed 2019-11-25
10. Speicher, N.K., Pfeifer, N.: Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery. Bioinformatics **31**(12), 268–275 (2015). doi:10.1093/bioinformatics/btv244. Accessed 2019-11-25
11. Witten, D.M., Tibshirani, R.J.: Extensions of Sparse Canonical Correlation Analysis with Applications to Genomic Data. Statistical Applications in Genetics and Molecular Biology **8**(1), 1–27 (2009). doi:10.2202/1544-6115.1470. Accessed 2019-11-25
12. Lock, E.F., Hoadley, K.A., Marron, J.S., Nobel, A.B.: Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. The Annals of Applied Statistics **7**(1), 523–542 (2013). doi:10.1214/12-AOAS597. Accessed 2020-02-07
13. Gabasova, E., Reid, J., Wernisch, L.: Clusternomics: Integrative context-dependent clustering for heterogeneous datasets. PLOS Computational Biology **13**(10), 1005781 (2017). doi:10.1371/journal.pcbi.1005781. Publisher: Public Library of Science. Accessed 2020-09-22
14. Mo, Q., Shen, R., Guo, C., Vannucci, M., Chan, K.S., Hilsenbeck, S.G.: A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. Biostatistics (Oxford, England) **19**(1), 71–86 (2018). doi:10.1093/biostatistics/kxx017. Accessed 2020-02-07
15. Lock, E.F., Dunson, D.B.: Bayesian consensus clustering. Bioinformatics **29**(20), 2610–2616 (2013). doi:10.1093/bioinformatics/btt425. Accessed 2020-01-28
16. Fred, A.L.N., Jain, A.K.: Combining multiple clusterings using evidence accumulation. IEEE Transactions on Pattern Analysis and Machine Intelligence **27**(6), 835–850 (2005). doi:10.1109/TPAMI.2005.113. Accessed 2020-01-27

17. Neo4j Graph Platform – The Leader in Graph Databases. https://neo4j.com/ Accessed 2020-09-28
18. Vega-Pons, S., Ruiz-Shulcloper, J.: A survey of clustering ensemble algorithms. International Journal of Pattern Recognition and Artificial Intelligence **25**(03), 337–372 (2011). doi:10.1142/S0218001411008683. Publisher: World Scientific Publishing Co. Accessed 2020-09-22
19. Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment **2008**(10), 10008 (2008). doi:10.1088/1742-5468/2008/10/P10008. Accessed 2019-11-25
20. Van Dongen, S.M.: Graph clustering by flow simultion. PhD Thesis, University of Utrecht, Netherlands (May 2000)
21. Fortunato, S.: Community detection in graphs. Physics Reports **486**(3) (2010). doi:10.1016/j.physrep.2009.11.002. Accessed 2020-09-16
22. Brohée, S., van Helden, J.: Evaluation of clustering algorithms for protein-protein interaction networks. BMC Bioinformatics **7**(1), 488 (2006). doi:10.1186/1471-2105-7-488. Accessed 2020-02-20
23. Vlasblom, J., Wodak, S.J.: Markov clustering versus affinity propagation for the partitioning of protein interaction graphs. BMC Bioinformatics **10**(1), 99 (2009). doi:10.1186/1471-2105-10-99. Accessed 2020-02-20
24. Lei, X., Wang, F., Wu, F.-X., Zhang, A., Pedrycz, W.: Protein complex identification through Markov clustering with firefly algorithm on dynamic protein–protein interaction networks. Information Sciences **329**, 303–316 (2016). doi:10.1016/j.ins.2015.09.028. Accessed 2020-02-20
25. Enright, A.J., Van Dongen, S., Ouzounis, C.A.: An efficient algorithm for large-scale detection of protein families. Nucleic Acids Research **30**(7), 1575–1584 (2002). doi:10.1093/nar/30.7.1575. Accessed 2020-02-20
26. Lancichinetti, A., Fortunato, S.: Limits of modularity maximization in community detection. Physical Review E **84**(6), 066122 (2011). doi:10.1103/PhysRevE.84.066122. Accessed 2020-02-19
27. Fortunato, S., Barthélemy, M.: Resolution limit in community detection. Proceedings of the National Academy of Sciences **104**(1), 36–41 (2007). doi:10.1073/pnas.0605965104. Accessed 2020-02-19
28. Sardana, D., Bhatnagar, R.: Graph Clustering Using Mutual K-Nearest Neighbors. In: Active Media Technology. Lecture Notes in Computer Science, pp. 35–48. Springer, Cham (2014). doi:10.1007/978-3-319-09912-5_4
29. Mancoridis, S., Mitchell, B.S., Rorres, C., Chen, Y., Gansner, E.R.: Using automatic clustering to produce high-level system organizations of source code. In: Proceedings. 6th International Workshop on Program Comprehension. IWPC'98 (Cat. No.98TB100242), pp. 45–52. IEEE Comput. Soc, Ischia, Italy (1998). doi:10.1109/WPC.1998.693283. http://ieeexplore.ieee.org/document/693283/ Accessed 2020-01-20
30. Newman, M.E.J.: Modularity and community structure in networks. Proceedings of the National Academy of Sciences of the United States of America **103**(23), 8577–8582 (2006). doi:10.1073/pnas.0601602103
31. The Cancer Genome Atlas Research Network: Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature **455**(7216), 1061–1068 (2008). doi:10.1038/nature07385. Accessed 2019-11-25
32. Noone, A.-M., Cronin, K.A., Altekruse, S.F., Howlader, N., Lewis, D.R., Petkov, V.I., Penberthy, L.: Cancer Incidence and Survival Trends by Subtype Using Data from the Surveillance Epidemiology and End Results Program, 1992-2013. Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology **26**(4), 632–641 (2017). doi:10.1158/1055-9965.EPI-16-0520
33. Fallahpour, S., Navaneelan, T., De, P., Borgo, A.: Breast cancer survival by molecular subtype: a population-based analysis of cancer registry data. CMAJ Open **5**(3), 734–739 (2017). doi:10.9778/cmajo.20170030. Accessed 2020-02-10
34. Jones, J.C., Renfro, L.A., Al-Shamsi, H.O., Schrock, A.B., Rankin, A., Zhang, B.Y., Kasi, P.M., Voss, J.S., Leal, A.D., Sun, J., Ross, J., Ali, S.M., Hubbard, J.M., Kipp, B.R., McWilliams, R.R., Kopetz, S., Wolff, R.A., Grothey, A.: Non-V600BRAF Mutations Define a Clinically Distinct Molecular Subtype of Metastatic Colorectal Cancer. Journal of Clinical Oncology **35**(23), 2624–2630 (2017). doi:10.1200/JCO.2016.71.4394. Accessed 2020-02-10
35. Bland, J.M., Altman, D.G.: The logrank test. BMJ : British Medical Journal **328**(7447), 1073 (2004). Accessed 2020-09-23
36. Sobin, L.H., Gospodarowicz, M.K., Wittekind, C., International Union against Cancer (eds.): TNM Classification of Malignant Tumours, 7th ed. 2009 edn. Wiley-Blackwell, Chichester, West Sussex, UK ; Hoboken, NJ (2010)
37. Byrd, J.C., Mrózek, K., Dodge, R.K., Carroll, A.J., Edwards, C.G., Arthur, D.C., Pettenati, M.J., Patil, S.R., Rao, K.W., Watson, M.S., Koduru, P.R.K., Moore, J.O., Stone, R.M., Mayer, R.J., Feldman, E.J., Davey, F.R., Schiffer, C.A., Larson, R.A., Bloomfield, C.D., Cancer and Leukemia Group B (CALGB 8461): Pretreatment cytogenetic abnormalities are predictive of induction success, cumulative incidence of relapse, and overall survival in adult patients with de novo acute myeloid leukemia: results from Cancer and Leukemia Group B (CALGB 8461). Blood **100**(13), 4325–4336 (2002). doi:10.1182/blood-2002-03-0772
38. Bennett, J.M., Catovsky, D., Daniel, M.-T., Flandrin, G., Galton, D.a.G., Gralnick, H.R., Sultan, C.: Proposals for the Classification of the Acute Leukaemias French-American-British (FAB) Co-operative Group. British Journal of Haematology **33**(4), 451–458 (1976). doi:10.1111/j.1365-2141.1976.tb03563.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2141.1976.tb03563.x. Accessed 2020-07-21
39. Parker, J.S., Mullins, M., Cheang, M.C.U., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., Quackenbush, J.F., Stijleman, I.J., Palazzo, J., Marron, J.S., Nobel, A.B., Mardis, E., Nielsen, T.O., Ellis, M.J., Perou, C.M., Bernard, P.S.: Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. Journal of Clinical Oncology **27**(8), 1160–1167 (2009). doi:10.1200/JCO.2008.18.1370. Accessed 2020-07-21
40. Steinley, D.: Properties of the Hubert-Arabie Adjusted Rand Index. Psychological methods **9**, 386–96 (2004). doi:10.1037/1082-989X.9.3.386
41. Netanely, D., Avraham, A., Ben-Baruch, A., Evron, E., Shamir, R.: Expression and methylation patterns partition luminal-A breast tumors into distinct prognostic subgroups. Breast Cancer Research **18**(1), 74 (2016). doi:10.1186/s13058-016-0724-2. Accessed 2020-07-30

42. Alizart, M., Saunus, J., Cummings, M., Lakhani, S.R.: Molecular classification of breast carcinoma. Diagnostic Histopathology **18**(3), 97–103 (2012). doi:10.1016/j.mpdhp.2011.12.003. Accessed 2020-07-30
43. Weigelt, B., Mackay, A., A'hern, R., Natrajan, R., Tan, D.S., Dowsett, M., Ashworth, A., Reis-Filho, J.S.: Breast cancer molecular profiling with single sample predictors: a retrospective analysis. The Lancet Oncology **11**(4), 339–349 (2010). doi:10.1016/S1470-2045(10)70008-5. Accessed 2020-07-30
44. Du, T., Zhu, L., Levine, K.M., Tasdemir, N., Lee, A.V., Vignali, D.A.A., Houten, B.V., Tseng, G.C., Oesterreich, S.: Invasive lobular and ductal breast carcinoma differ in immune response, protein translation efficiency and metabolism. Scientific Reports **8**(1), 7205 (2018). doi:10.1038/s41598-018-25357-0. Number: 1 Publisher: Nature Publishing Group. Accessed 2020-07-24