Bioinformatics, 36(3), 2020, 676–681
doi: 10.1093/bioinformatics/btz661
Advance Access Publication Date: 26 August 2019
Original Paper

Genome analysis

Partition: a surjective mapping approach for dimensionality reduction

Joshua Millstein[1],*, Francesca Battaglin[2,3], Malcolm Barrett[1], Shu Cao[1], Wu Zhang[2], Sebastian Stintzing[4], Volker Heinemann[5] and Heinz-Josef Lenz[2]

[1]Department of Preventive Medicine and [2]Department of Medicine, Division of Medical Oncology, Norris Comprehensive Cancer Center, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA, [3]Clinical and Experimental Oncology Department, Medical Oncology Unit 1, Veneto Institute of Oncology IOV-IRCCS, Padua 35128, Italy, [4]Medical Department, Division of Oncology and Hematology, Charité Universitaetsmedizin Berlin, Berlin 10117, Germany and [5]Department of Medicine III, University Hospital Munich, Munich 80336, Germany

*To whom correspondence should be addressed.

Associate Editor: Russell Schwartz

Received on March 14, 2019; revised on May 22, 2019; editorial decision on August 16, 2019; accepted on August 20, 2019

Abstract

Motivation: Large amounts of information generated by genomic technologies are accompanied by statistical and computational challenges due to redundancy, badly behaved data and *noise*. Dimensionality reduction (DR) methods have been developed to mitigate these challenges. However, many approaches are not scalable to large dimensions or result in excessive information loss.

Results: The proposed approach partitions data into subsets of related features and summarizes each into one and only one new feature, thus defining a surjective mapping. A constraint on information loss determines the size of the reduced dataset. Simulation studies demonstrate that when multiple related features are associated with a response, this approach can substantially increase the number of true associations detected as compared to principal components analysis, non-negative matrix factorization or no DR. This increase in true discoveries is explained both by a reduced multiple-testing challenge and a reduction in extraneous noise. In an application to real data collected from metastatic colorectal cancer tumors, more associations between gene expression features and progression free survival and response to treatment were detected in the reduced than in the full untransformed dataset.

Availability and implementation: Freely available R package from CRAN, https://cran.r-project.org/package=partition.

Contact: joshua.millstein@usc.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Advancements in genomic technologies for the measurement of DNA variation, epigenome, proteome, transcriptome, microbiome, metabolome, etc., have led to decreasing costs and vastly increasing amounts of information collected from individual tissue samples (Karczewski and Snyder, 2018). However, in addition to new opportunities, the ability to capture and store massive amounts of information has brought new challenges, such as data storage, computational resources, large amounts of *noise* or unrelated data and highly dependent or redundant data (Malod-Dognin *et al.*, 2018; Wang *et al.*, 2016).

Algorithms have been developed to reduce the number of features in a data preparation step (Biswas *et al.*, 2008; Byrd and Segre, 2015; Kouchaki *et al.*, 2018) without discarding excessive amounts of relevant information (Wang *et al.*, 2010). For example, *feature selection* approaches attempt to filter unimportant features, leaving the selected features untransformed (Chandrashekar and Sahin, 2014; Fisher and Mehta, 2015). Feature extraction strategies transform data, extracting a reduced number of features. An example is principal components analysis (PCA), where components that capture a sufficient amount of variance are carried forward for the subsequent analyses. *Supervised* methods incorporate information from the disease response to assist in DR. However, if the selection process is not accounted for in subsequent analyses, a computationally burdensome process, it can bias results (Smialowski *et al.*, 2010). Therefore, here we focus on *unsupervised* approaches.

A statistical challenge of high dimensional data is the multiple testing or signal-to-noise ratio problem. Unrelated information makes detecting associations relating genomic features to clinical responses, more difficult (Fan *et al.*, 2014). This dynamic affects both conventional statistical approaches and machine learning. When testing multiple hypotheses, statistical significance depends on a *P*-value threshold, commonly determined by Bonferroni or Benjamini and Hochberg (BH) corrections, both of which are increasingly stringent with the number of features. Even with approaches such as cross validation, the presence of unrelated features decreases the ability to detect relationships of interest (Khalid *et al.*, 2014). By reducing redundancy and thus dimensionality, it may be possible to reduce noise (Khalid *et al.*, 2014), and thereby improve the signal-to-noise ratio.

As described above, the potential benefits of DR include reduced computational demands, reduced multiple-testing burden, reduced noise, and better-behaved data. However, few existing methods meet several important criteria, (i) interpretable mapping between original and reduced features, (ii) user-specified constraint on information loss and (iii) scalability to high dimensions. Criterion (ii) insures that strong associations involving original features will be detectable in reduced features given a sufficiently stringent constraint. The proposed method is a hybrid of feature selection and feature extraction, where some features are transformed while others are carried forward without modification.

We apply the approach to identify gene expression features associated with treatment response in colorectal cancer tumors from patients treated with first-line chemotherapy plus cetuximab in the FIRE-3 phase III clinical trial (Heinemann *et al.*, 2014).

2 Materials and methods

DR approaches are typically designed for going from high to low dimensions, discarding most of the information in the process. We propose a method, Partition, for reducing dimensionality by a small or moderate amount, summarizing dependent features into a smaller number of less dependent features. Below we describe the general framework and algorithm, we assess performance in simulated data comparing to standard approaches, and then we describe the application to real data from the FIRE-3 clinical trial.

2.1 Partition framework and algorithm

The framework is a partitioning of input features, where a function is applied to features within each subset to summarize them into a single new extracted feature, constrained to satisfy a maximum information loss criterion. That is, each new feature must capture at a minimum a specified proportion of information contained in the feature subset. The mapping between input and reduced features is surjective in that every reduced feature corresponds to one or more input features and every input feature yields one and only one feature in the reduced dataset (Fig. 1). The algorithm to find the partition given the information loss constraint has two objectives, (i) minimize the number of subsets, and (ii) minimize information loss, conditional on the minimum number of subsets.

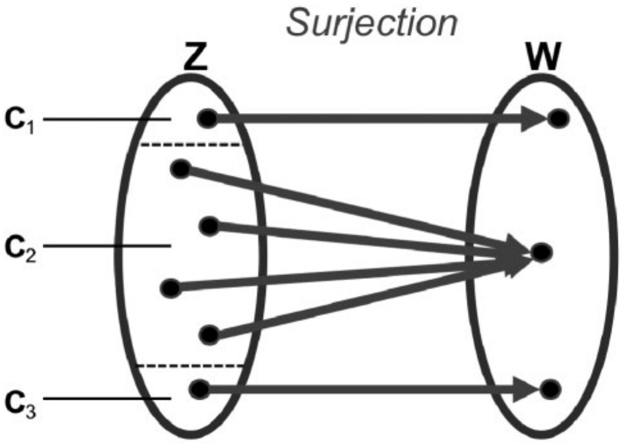

Fig. 1. Depiction of a surjective mapping in which elements of the first set map to one and only one element of the second set. Here the surjective mapping is defined by a partition that divides the full set of features, Z, into the subsets, C_j. In the proposed framework, features in sets C_1 and C_3 would be carried forward to the reduced set without modification, whereas a single new reduced feature would be generated by summarizing all features in subset C_2

More formally, let Z denote a set of m variables representing features with dependencies. A partition, C, groups variables in Z into subsets, C_q, such that $\cup_q C_q = Z$ and $C_q \cap C_r = \emptyset,\ \forall\ q \neq r$. We define a function $g(\cdot)$ that accepts as arguments variables in Z corresponding to some subset C_q and yields a single summary variable W_q, so that $g(C_q) = W_q$. Here we implement $g(\cdot)$ using the arithmetic mean. We define a second function $h(\cdot)$, that estimates the proportion of information in C_q captured by W_q, that is, $h(C_q, W_q) = \nu_q$, where ν_q is the estimate of the proportion of information captured. We implement $h(\cdot)$ with the intraclass correlation coefficient (ICC), however, there are other options for both $g(\cdot)$ and $h(\cdot)$ (Supplementary Data). There is a single parameter, ν^*, specified by the investigator that determines the minimum acceptable amount of information captured for any partition, $\nu_q \geq \nu^*\ \forall\ q$. Thus, the DR procedure yields a new dataset W, with $k \leq m$ features.

This framework yields a surjective mapping between Z and W, which helps with interpretation, because unlike methods such as PCA, factor analysis, or NMF, if a feature in W is identified as related to a response of interest, the exact subset of features in Z that are implicated is known with no ambiguity. Another important aspect of this framework is the information loss constraint, ν^*. This constraint allows the investigator to let the extent of dependences among features on a local level guide the extent to which the number of features is reduced.

We propose an agglomerative nearest neighbor algorithm for finding C. It is computationally efficient and deterministic, because it does not require a random process. The steps are:

1. Initialize $W: = Z$. As the algorithm progresses, each W_j will be generated from one or more features in Z. We will keep track of this mapping with a function $A(\cdot)$. Thus initially, $A(W_j) = Z_j\ \forall\ j$.
2. Propose a new partition (subset of features), $C'_q = A(\text{argmin}(d(W_j, W_l)_{(i)}\ \forall\ j \neq l))$. The prime symbol in C'_q indicates that this is a proposed but not yet accepted partition, the function $d(\cdot)$ measures dissimilarity between features W_j and W_l and is defined as $1 - r$, where r is Pearson's correlation coefficient. Initially, dissimilarity is computed for all pairs of features. The subscript (i) denotes the ith smallest value, where i is initialized to 1, that is, we begin by considering the two nearest (most similar) features. The subset of Z that corresponds to the ith nearest features in W is proposed as a new agglomerated partition.
3. Compute $\nu'_q = h(C'_q,\ g(C'_q))$. If $\nu'_q \geq \nu^*$ then $C_q = C'_q$, $W_q = g(C'_q)$ and $i = 1$, else $i = i + 1$. That is, accept the proposed partition if the information loss constraint is satisfied.
4. If $i \leq B$, return to step 2, else end algorithm. Note that if W has changed, dissimilarities must be recomputed in order to proceed with step 2.

The parameter B default is approximately $0.2m$. We've found this setting to be adequate under the conditions that we have explored. We have implemented Partition as the *partition* R package, which incorporates C++ via the Rcpp package to speed computation.

2.2 *In silico* comparison of approaches

In computer simulated data, we compared the performance of Partition with three other approaches, PCA, non-negative matrix factorization (NMF) (Brunet *et al.*, 2004) and k-means. For comparison purposes the top k principle components were used as the reduced set, where the reduced dimension, k, was determined by Partition. k was defined similarly for NMF. However, the k-means algorithm was applied in an approach related to the Partition concept. k was iteratively decremented from m to find the smallest k, where $g(\cdot) \equiv$ arithmetic mean, $h(\cdot)\ \equiv$ the intraclass correlation coefficient (ICC) and C is defined by the k-means solution.

Data were simulated under several scenarios to determine how DR affects the ability to detect associations between a set of Gaussian features and a smaller set of dependent Gaussian responses. Dependencies among molecular features may indicate a shared role in a biological process or pathway, as observed in gene expression *modules*, groups of coexpressed genes and summarized as eigengenes (Langfelder and Horvath, 2007). Relationships between these pathways and sample traits can result in correlations between the respective eigengenes and the traits (Bailey *et al.*, 2016; Chen *et al.*, 2016; Peters *et al.*, 2017). We simulate this type of relation by generating response features as linear combinations of block correlated Gaussian features with random error. In other simulations, responses are linearly dependent on single Gaussian features either within or external to block correlated features.

In the first two scenarios, A and B, response associated features (predictors) were simulated as block diagonal Gaussian variables with correlations within in a block randomly distributed as U(0.2, 0.6). For each dataset, 20 such blocks were generated, where block size was randomly distributed as discrete U(3, 15), scenario A, or U(16, 30), scenario B. A single response feature was generated for each block as a linear combination with coefficients randomly distributed as U(0.02, 0.05), plus random error distributed as N(0, 1). Correlation and effect size ranges were chosen for the dynamic range of statistical power to detect associations between the predictors and responses for the approaches assessed. To simulate noise, an additional 20 correlated blocks of predictors were simulated as above but with block size distributed as discrete U(1, 15). Also, 20 standard Gaussian variables were included as noise in the set of responses, that is, 20 additional response variables. 100 replicate datasets were simulated according to the above, with 200 observations in each.

For scenario C, all predictors were statistically independent of each other, block size equal to one and coefficients were randomly distributed as U(2.0, 2.5). Noise features were simulated as in scenario A of the main text but with all correlations equal to zero.

Scenario D was similar to A, but response associated predictors were statistically independent of each other (block size equal to one) and coefficients were randomly distributed as U(.2, .25). Correlations among features within noise blocks were distributed as U(0.2, 0.6), with block size distributed as discrete U(1, 15).

For each simulated dataset described above, we applied the four DR approaches with a series of 20 increasingly lenient information loss parameters in order to generate progressively extreme reductions in dimensionality. We then conducted tests of Pearson's correlation coefficients to identify associations between predictors and responses, accounting for multiple tests with the BH false discovery rate (FDR) approach. DR approaches were evaluated based on the mean number of true discoveries across replicate datasets. A true discovery was defined as a rejected null hypothesis when the coefficient for the association between an input predictor and response was not equal to zero. To avoid complications resulting from dependent tests, we counted a maximum of one true discovery for each response feature. Thus, the maximum number of true discoveries for a given DR approach applied to a given dataset was equal to the number of truly associated response features. It should be noted that here, the number of true discoveries is conceptually related to statistical power, which is the probability that the null hypothesis will be rejected when the null hypothesis is in fact false. In these simulations, the null hypothesis is linear independence between a reduced feature and a response, which was only the case when all corresponding input features were linearly independent of the response. Consequently, in this setting more true discoveries implies greater statistical power and vice versa.

2.3 Application to mRNA expression in tumors of metastatic colorectal cancer patients

Formalin-fixed paraffin embedded (FFPE) tumor tissue was collected at surgery from 101 patients enrolled in the FOLFIRI + cetuximab arm of FIRE-3, an open-label randomized phase III clinical trial (Heinemann *et al.*, 2014). All patients, aged 18–75, had stage IV colorectal cancer, were KRAS exon 2 wild-type, and had Eastern Cooperative Oncology Group (ECOG) performance status less than or equal to 2. Patients were followed an assessed for objective response (OR), progression-free survival (PFS) and overall survival (OS). FFPE tumor samples were profiled for the expression of 2548 genes using the HTG EdgeSeq Oncology Biomarker Panel. Normalization was performed according the Anders and Huber (2010) approach (Anders and Huber, 2010) implemented in the DESeq R package. Quality control assessment for removal of low-quality samples was conducted according to the approach advocated by HTG, greater than two standard deviations difference between observed mean of *ANT* control probes and the expected mean based on multiple sample types (plasma, serum, FFPE, Brain RNA, PAXgene, cell lines).

The Partition approach was used to reduce the dimensionality of the gene expression dataset for a series of five increasingly lenient information loss constraints. Each dataset was then analyzed to identify gene expression features associated with patient outcomes. Cox proportional hazards regression was used for the survival outcomes. To identify genes differentially expressed with respect to objective response (OR), defined as complete response (CR) or partial response (PR) versus stable disease (SD) or progressive disease (PR), a quasi-likelihood negative binomial generalized log-linear model with empirical Bayes quasi-likelihood F-tests (Lund *et al.*, 2012) was implemented using the edgeR R package. Because this approach is designed for count data, we recomputed the reduced features from each partition subset by summing counts rather than using the arithmetic mean. Multiple testing was accounted for by applying BH FDR. As a sensitivity analysis, logistic regression was also applied, treating OR as a binary outcome. Patient characteristics considered for inclusion as adjustment covariates included age, sex, BMI, liver-limited metastatic disease, ECOG and BRAF, as well as the first two PCs of control probes included on the EdgeSeq panel. For each outcome, a backward stepwise approach with alpha equal to 0.05 was applied to determine the final set of covariates. The final sets of adjustment covariates were (i) the two PCs for OR, (ii) age, BRAF and the first control PC for PFS and (iii) ECOG for OS.

3 Results

3.1 *In silico* comparison of approaches

Applying Partition and k-means to large-scale testing settings in simulated data as described in scenarios A and B resulted in substantial gains (Fig. 2A and B) in the number of true discoveries. These gains were likely due to a decrease in the multiple-testing burden and an increase in statistical power of tests under the alternative hypothesis when multiple correlated features are jointly associated with the response (Supplementary Fig. S1). Also, the gains are not explained by differences in observed FDR, which was generally well controlled and similar to results from the untranformed input datasets (Supplementary Fig. S4). Further, the number of true discoveries tended to increase with more extreme reductions in dimensionality. In contrast, the more conventional approaches, PCA and NMF, yielded very few discoveries and performed much worse than no DR under all of these scenarios. It's important to note that PCA is not plotted across the full range of reductions in dimension due to the constraint inherent in the method that the number of PCs cannot be larger than the smaller of the dimensions of features, m, or of samples, n. The k-means approach performed the best across the entire range for scenario A (Fig. 2A). However, when cluster sizes were increased to U(16, 30) the ICC approach slightly out-performed k-means over most of the domain (Fig. 2B). When all predictors were statistically independent of each other, all approaches resulted in a loss in true discoveries as compared to no DR (Fig. 2C). However, for Partition and k-means, the loss was roughly proportional to the amount of reduction, implying that even under this challenging scenario, the risk of a substantial negative impact to statistical power is small when the proportion of information discarded is small. When responses were solely dependent on single features that were independent of other predictors (scenario D) but blocks of dependent predictors were present, both k-means and partition out-performed no DR when the reduction was not overly extreme (Fig. 2D), suggesting that gains in true discoveries can be achieved solely by alleviating the multiple testing burden. Partition substantially out-performed k-means under this scenario. When responses were dependent on single feature effects within correlated

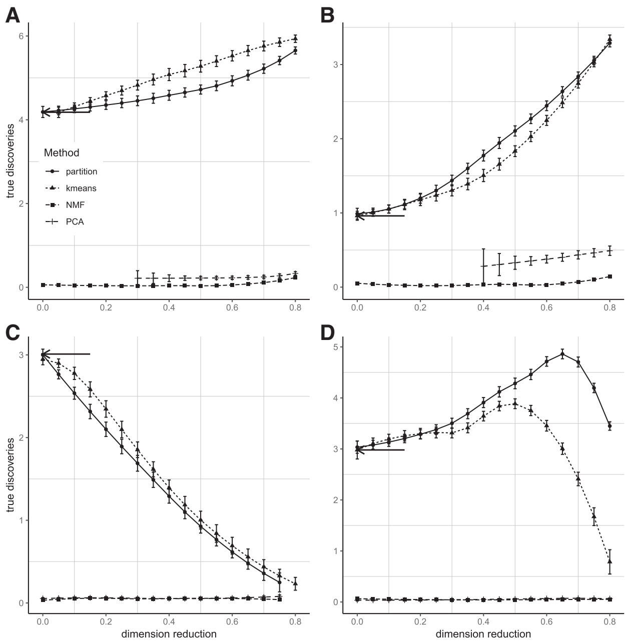

Fig. 2. Frequency of true discoveries, defined by associations between predictors and response, identified by DR approaches. Datasets were analyzed after applying a series of increasingly permissive information loss constraints, leading to increased dimension reduction, x-axis. Arrows indicate the number of true discoveries detected with the untransformed input data. (**A**) Block correlated Gaussian predictors with block sizes 3–15. Responses were dependent on linear combinations of the block correlated predictors. Block correlated noise features were simulated in block sizes 1–15. (**B**) Block correlated predictors as in A but block sizes were 16–30 rather than 3–15. (**C**) All predictors were statistically independent of each other, block size equal to one. (**D**) Block correlated predictors as in A but responses were dependent on other independent single Gaussian features

blocks, results were similar to scenario C but here the drop-off in discoveries with reduction was more gradual and Partition performed similarly to k-means (Supplementary Fig. S3).

Performance of Partition was also strong when other functions were used for g(·) and h(·), such as the first principal component, g(·), variance explained by the first principal component, h(·), minimum r-squared, h(·) and the standardized mutual information (Butte and Kohane, 2000), g(·), (Supplementary Figs S2 and S3). There are multiple aspects of performance that affect scalability and distinguish Partition from other approaches. The inability of PCA and NMF to identify meaningful numbers of true discoveries under these conditions limits their relevance when considering scalability, however, the k-means approach did perform competitively. When we assessed computational time of k-means versus Partition in the context of dimensionality reduction for predictors simulated in scenario A, we found that Partition far out-performed k-means both in terms of computation time and increase in time with increased numbers of input features (Supplementary Fig. S5). Also, the relation between dimensionality and computation time was much more stable for Partition than k-means. Just as concerning, the relation between the number of input features and the proportion of dimensionality reduction was not stable but decreased inversely with the number of input features (Supplementary Fig. S6). The implication is that the larger the dimensionality of the input dataset, the harder it is for k-means to satisfy the information loss constraint for all subsets. In contrast, with Partition there was no apparent dependency between the number of input features and proportion of dimensionality reduction, which was relatively stable regardless of the dimensionality of the input dataset.

3.2 Application to mRNA expression in tumors metastatic colorectal cancer patients

Extensive dependencies were observed among gene expression features in the heatmap and dendrogram (Supplementary Fig. S7), but the high density of lines causes display problems and limits scalability. The Partition approach leads to an easily interpretable graph, a *consolidation plot*, displaying the frequencies of size classes of partitions for a series of increasingly stringent information loss constraints (Fig. 3). The consolidation plot is informative at any scale. Depicted on the left of Figure 3, small partitions of features are summarized into reduced features while retaining at least 80% variance in the data (a maximum of 20 percent information loss). Moving right, as the information loss constraint becomes more permissive, partitions tend to include larger numbers of features and consequently, the number of features, *k*, in the reduced dataset decreases. By randomly permuting all gene expression features with respect to each other, we determined that if there were complete independence between features, consolidation would not happen unless the information loss constraint was relaxed to approximately 50% or more. The contrast between complete independence and the observed reduction, which has approximately 100 consolidated features with a maximum information loss of only 20%, indicates pervasive dependencies among gene expression features.

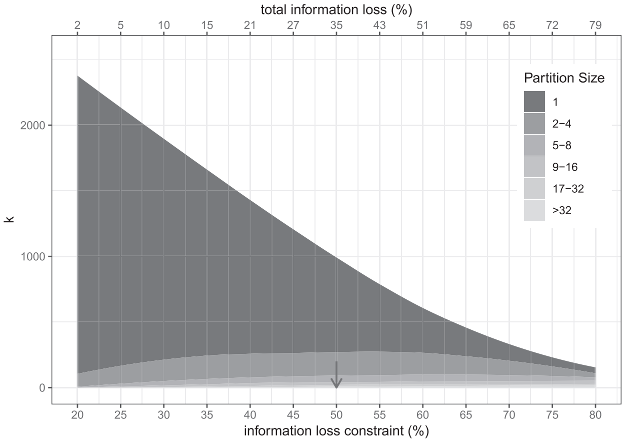

Fig. 3. Consolidation plot. k indicates the number of features in the reduced dataset and the frequencies of size classes by shade, of partitions for a series of increasingly stringent information loss constraints, lower x-axis. Shade indicates the size range of partition subsets. The upper x-axis shows the total loss of information from reduction, summing over partition subsets. The grey arrow indicates the minimum information loss constraint required for input features to be consolidated into reduced features. Note that total information loss is always less than the constraint, which is the maximum information lost to any partition subset

For association analysis, DR was conducted using Partition with 5 information loss constraints, the proportion of variance captured set to 40, 50, 60, 70, 80% and no DR (100%). The reductions from the original 2548 features yielded new datasets with 558, 943, 1359, 1834 and 2295 features, respectively. Discoveries were identified at four FDR levels, 0.05, 0.10, 0.15 and 0.20.

No discoveries were made at any of the specified FDR levels for the OS outcome. The analysis of PFS yielded one discovery at the 0.05 level for reduced datasets 949, 1359 and 1834, but not the full or 558 (Fig. 4). The feature identified at the conventional 0.05 level was a single gene, *SORBS1* (Supplementary Table S1), not a consolidated feature, implying that in this case low power due to multiple testing explains why it was not identified by the analysis of the full dataset. At the 0.10 FDR level, only the reduced datasets 949 and 1359 resulted in two discoveries, yielding the additional gene, *SLCO1B1*. The reduced datasets also generated more discoveries at the higher FDR levels of 0.15 and 0.20.

Analysis of the 1355 dataset yielded six genes differentially expressed (DE) for OR at the 0.05 FDR level, the largest number of discoveries at the conventional 0.05 threshold (Fig. 5). In contrast, analysis of the full dataset, yielded four genes. Analysis of the 1355 dataset identified the most DE genes at all four FDR levels, and in fact as many or more DE genes were identified by the reduced datasets as compared to the full dataset, at all FDR levels. Unlike the PFS results, extracted gene expression features summarized from multiple

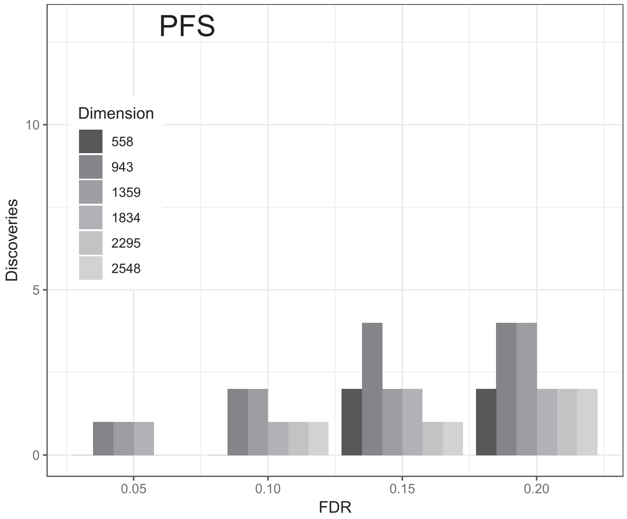

Fig. 4. Frequency of discoveries from the analysis of associations between gene expression and progression-free survival (PFS) in metastatic colorectal cancer patients from the FIRE-3 clinical trial using the full dataset, 2548 expression features and 5 reduced datasets generated by applying the Partition method. Discoveries are shown for a series of four increasingly lenient FDR thresholds

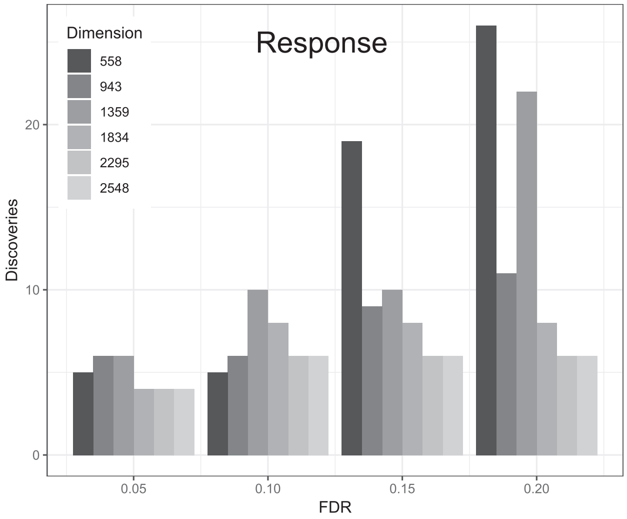

Fig. 5. Results from the differential gene expression analysis using edgeR across two categories of objective response (OR), complete response or partial response versus stable disease or progressed disease, in metastatic colorectal cancer patients from the FIRE-3 clinical trial using the full dataset, 2548 expression features, and 5 reduced datasets generated by applying the Partition method. Discoveries are shown for a series of four increasingly lenient FDR thresholds

original feature were identified as DE for OR (Supplementary Table S2). Interestingly, these tended to involve single genes identified in the full dataset, supporting the idea that summarizing these features does not necessarily reduce power to detect them.

There was suggestive evidence that some genes identified as associated with OR are also associated with survival (Supplementary Table S3). For example, expression of *MIA* was nominally associated with both PFS and OS (PFS, HR = 1.43, $P = 0.0061$; OS, HR = 1.43, $P = 0.0043$) in Cox models after adjusting for covariates as described in Methods, and KM plots support this supposition (Supplementary Fig. S6). In the sensitivity analyses, modeling OR as a binary outcome using logistic regression and adjusting for control probes, *SNCA* was the most significant gene ($P = 1.61 \times 10^{-4}$). *SNCA* was also suggestively associated with OS in Cox models after adjusting for covariates, and HR estimates were similar for PFS and OS (PFS, HR = 1.19, $P = 0.16$; OS, HR = 1.22, $P = 0.091$). Additionally, logrank *P*-values (unadjusted for covariates) for tests of differences in expression quintiles were nominally significant for PFS ($P = 0.0008$) and suggestive for OS ($P = 0.052$) (Supplementary Fig. S8).

4 Discussion

Here we propose Partition, an unsupervised DR method that limits information loss for every subset of features that is transformed into a new feature, consequently preserving or increasing under a variety of conditions the ability to detect true associations between features in the reduced dataset and external features such as clinical responses. We show that when associations involved a relatively limited number of features, this approach far out-performs widely used methods such as PCA and NMF. Another advantage of Partition is an unambiguous mapping between features in the full dataset and those in the reduced dataset, which substantially improves interpretability. In addition, Partition is deterministic, unlike conventional k-means, thus it is easier to achieve repeatability of analytic results.

The information loss constraint allows the extent of DR to depend on the magnitude of dependencies among features. The intent is to reduce dimensionality as much as possible while limiting loss of information for every input feature, according to the risk tolerance of the investigator. If the information loss constraint is stringent, only very similar features are consolidated, resulting in a similar number of true discoveries detected from the reduced versus the full dataset, which was shown here in simulation results. In contrast, PCA and NMF exhibited very poor ability to identify true associations under conditions explored here even when the dimensionality of the reduced dataset was very close to that of the input. Two general scenarios were identified with substantial gains in the numbers of true discoveries when Partition was applied, (1) multiple dependent features jointly associated with a response, and (2) independent features associated with a response along with block correlated features not associated. Both scenarios seem likely to occur in multi-omic studies. Scenario 1 corresponds to eigengene or metagene effects, whereas scenario 2 represents other likely conditions that include groups of dependent features independent of other individual features that are associated with responses. Though use of Pearson's correlation coefficient and the ICC in the Partition algorithm suggests that it is best suited for linear dependencies, when non-linear relationships are suspected, non-linear techniques such as mutual information and Spearman correlation coefficient could be substituted.

Multiple aspects of the Partition method make it scalable to large numbers of input features in comparison to the other DR approaches considered. First, we have shown that when only a small proportion of input features are truly associated with a response, approaches such as PCA and NMF that transform all input features into all reduced features tend to fail, whereas Partition performs well. Second, we demonstrated that our Partition software is much faster than the k-means approach and computation time increases more slowly with increasing numbers of input features. Third, computation time was much more stable for Partition than k-means, which varied substantially from dataset to dataset, even when the underlying structure and dimensionality were similar. Fourth, given a certain dependency structure in the input data and a specified information loss constraint, we would prefer a DR approach that is does not vary in the proportion of dimensionality reduction as the number of input features changes. We showed that Partition is very stable, whereas the k-means approach yields a larger proportion as the input number increases. The implication is that information loss is not well constrained across clusters, leading to the risk of some input features not being well represented in the reduced dataset. Fifth, *a priori* knowledge regarding dependencies among features, e.g. CpG DNA methylation or tumor copy number features ordered along a chromosome, where dependencies decrease with distance, could enable substantial increases in scalability. This may be achieved by placing a constraint on subset membership to those features residing within a threshold distance known *a priori*. It is also may be possible to apply a data-driven approach to divide the distance matrix into supersets that could be used to parallelize the algorithm. We plan to pursue these questions in future research.

Notably, applying Partition resulted in the identification of genes with previous evidence of a role in colorectal cancer biology or response to chemotherapy treatment. For example, *SLCO1B1* encodes for a liver-specific solute carrier responsible for hepatic drug uptake and is involved in irinotecan active metabolite SN-38 hepatic disposition in patients treated with FOLFIRI chemotherapy (Nozawa *et al.*, 2004).

Furthermore, applying Partition for analysis of OR resulted in a larger number of discoveries in the reduced datasets as compared to the full dataset, identifying gene expression features summarized from multiple input features as differentially expressed for OR. Interestingly, each of these genes (i.e. *PIAS2*, *PLA2G2A*, *SMAD7*, *ANPEP*, *LRP6*) have been reported to play a role in cancer and were identified in previous reports as associated with colorectal cancer risk or prognosis (Cormier *et al.*, 1997; Huang *et al.*, 2016; Sanz *et al.*, 2015; MacPhee *et al.*, 1995; Rabellino *et al.*, 2017; Yao *et al.*, 2017). Of note, *ANPEP* and *LRP6* cluster together in our analysis, as well as *PIAS2*, *PLA2G2A* and *SMAD7*, thus suggesting underlying interconnections between these genes or related pathways which may warrant further investigation.

Finally, the colorectal cancer findings generated an intriguing research hypothesis involving *MIA* and *SNCA*. The former encodes for a protein that inhibits growth of melanoma cells in vitro as well as some other neuroectodermal tumors, including gliomas (Bosserhoff and Buettner, 2002). In contrast, alpha-synuclein (SNCA) is a protein whose aberrant aggregation in central nervous system neurons, following pathogenic mutations in the *SNCA* gene, leads to Parkinson's Disease development (Deng *et al.*, 2018). Gene expression of both genes emerged in our analysis as associated with treatment outcomes and survival, opening new perspectives on the role of pathways related to neuronal differentiation and neurodegenerative diseases in colorectal cancer, which find a supporting rationale in evidence in previous literature (Feng *et al.*, 2015; Li *et al.*, 2015).

In conclusion, Partition provides a powerful and efficient information-based approach to reduce redundancy in large-scale datasets, thereby reducing the computational requirements of subsequent analyses without introducing bias. It may be possible to learn the optimal information-based dimensionality reduction in various contexts even when the extent and strength of dependencies vary from dataset to dataset. We expect future research to provide guidance on this topic, driven by analyses of data collected from multiple studies.

Acknowledgements

We thank patients of the FIRE-3 study for their participation. Computation for the work described in this paper was supported by the University of Southern California's Center for High-Performance Computing (hpcc.usc.edu).

Funding

This work was partly supported by the National Cancer Institute (grant numbers P30CA014089 and P01CA196569), the Gloria Borges WunderGlo Foundation-the Wunder Project, Dhont Family Foundation, San Pedro Peninsula Cancer Guild, Daniel Butler Research Fund and Call to Cure Fund.

Conflict of Interest: none declared.

References

Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.

Bailey,P. *et al.* (2016) Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature*, **531**, 47–52.

Biswas,S. *et al.* (2008) Mapping gene expression quantitative trait loci by singular value decomposition and independent component analysis. *BMC Bioinformatics*, **9**, 244.

Bosserhoff,A.K. and Buettner,R. (2002) Expression, function and clinical relevance of MIA (melanoma inhibitory activity). *Histol. Histopathol.*, **17**, 289–300.

Brunet,J.P. *et al.* (2004) Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci. USA*, **101**, 4164–4169.

Butte,A.J. and Kohane,I.S. (2000) Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac. Symp. Biocomput.*, 418–429.

Byrd,A.L. and Segre,J.A. (2015) Integrating host gene expression and the microbiome to explore disease pathogenesis. *Genome Biol.*, **16**, 70.

Chandrashekar,G. and Sahin,F. (2014) A survey on feature selection methods. *Comput. Electrical Eng.*, **40**, 16–28.

Chen,X. *et al.* (2016) Integrated analysis of long non-coding RNAs in human colorectal cancer. *Oncotarget*, **7**, 23897–23908.

Cormier,R.T. *et al.* (1997) Secretory phospholipase Pla2g2a confers resistance to intestinal tumorigenesis. *Nat. Genet.*, **17**, 88.

Deng,H. *et al.* (2018) The genetics of Parkinson disease. *Ageing Res. Rev.*, **42**, 72–85.

Fan,J. *et al.* (2014) Challenges of big data analysis. *Natl. Sci. Rev.*, **1**, 293–314.

Feng,D.D. *et al.* (2015) The associations between Parkinson's disease and cancer: the plot thickens. *Transl. Neurodegen.*, **4**, 20.

Fisher,C.K. and Mehta,P. (2015) Bayesian feature selection for high-dimensional linear regression via the Ising approximation with applications to genomics. *Bioinformatics*, **31**, 1754–1761.

Heinemann,V. *et al.* (2014) FOLFIRI plus cetuximab versus FOLFIRI plus bevacizumab as first-line treatment for patients with metastatic colorectal cancer (FIRE-3): a randomised, open-label, phase 3 trial. *Lancet Oncol.*, **15**, 1065–1075.

Huang,Y. *et al.* (2016) SMAD7 polymorphisms and colorectal cancer risk: a meta-analysis of case-control studies. *Oncotarget*, **7**, 75561.

Karczewski,K.J. and Snyder,M.P. (2018) Integrative omics for health and disease. *Nat. Rev. Genet.*, **19**, 299–310.

Khalid,S. *et al.* (2014) A survey of feature selection and feature extraction techniques in machine learning. In: Science and Information Conference (SAI), pp. 372–378. IEEE.

Kouchaki,S. *et al.* (2018) Application of machine learning techniques to tuberculosis drug resistance analysis. *Bioinformatics*, **35**, 2276–2282.

Langfelder,P. and Horvath,S. (2007) Eigengene networks for studying the relationships between co-expression modules. *BMC Syst. Biol.*, **1**, 54.

Li,W.H. *et al.* (2015) Detection of SNCA and FBN1 methylation in the stool as a biomarker for colorectal cancer. *Dis. Markers*, **2015**, 657570.

Lund,S.P. *et al.* (2012) Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates. *Stat. Appl. Genet. Mol. Biol.*, **11**, 1–42.

MacPhee,M. *et al.* (1995) The secretory phospholipase A2 gene is a candidate for the Mom1 locus, a major modifier of ApcMin-induced intestinal neoplasia. *Cell*, **81**, 957–966.

Malod-Dognin,N. *et al.* (2018) Precision medicined – a promising, yet challenging road lies ahead. *Curr. Opin. Syst. Biol.*, **7**, 1–7.

Nozawa,T. *et al.* (2004) Role of organic anion transporter OATP1B1 (OATP-C) in hepatic uptake of irinotecan and its active metabolite, 7-ethyl-10-hydroxycamptothecin: in vitro evidence and effect of single nucleotide polymorphisms. *Drug Metabol. Disposition Biol. Fate Chem.*, **33**, 434–439.

Peters,A.A. *et al.* (2017) Oncosis and apoptosis induction by activation of an overexpressed ion channel in breast cancer cells. *Oncogene*, **36**, 6490–6500.

Rabellino,A. *et al.* (2017) The role of PIAS SUMO E3-ligases in cancer. *Cancer Res.*, **77**, 1542–1547.

Sanz,B. *et al.* (2015) Aminopeptidase N activity predicts 5-year survival in colorectal cancer patients. *J. Investig. Med.*, **63**, 740–746.

Smialowski,P. *et al.* (2010) Pitfalls of supervised feature selection. *Bioinformatics*, **26**, 440–443.

Wang,L. *et al.* (2016) Feature selection methods for big data bioinformatics: a survey from the search perspective. *Methods*, **111**, 21–31.

Wang,Y. *et al.* (2010) High-dimensional pattern regression using machine learning: from medical images to continuous clinical variables. *Neuroimage*, **50**, 1519–1535.

Yao,Q. *et al.* (2017) LRP6 promotes invasion and metastasis of colorectal cancer through cytoskeleton dynamics. *Oncotarget*, **8**, 109632–109645.