Individualized multi-omic pathway deviation scores using multiple factor analysis

Andrea Rau[1,2], Regina Manansala[2], Michael J. Flister[3],
Hallgeir Rui[3], Florence Jaffrézic[1], Denis Laloë[1*], Paul L. Auer[2*]

[1] GABI, INRA, AgroParisTech, Universite Paris-Saclay, 78350, Jouy-en-Josas, France
[2] Joseph J. Zilber School of Public Health, University of Wisconsin-Milwaukee, Milwaukee, WI 53201, USA
[3] Department of Pathology, Medical College of Wisconsin, Milwaukee, WI 53226, USA
[*] Corresponding authors: denis.laloe@inra.fr, pauer@uwm.edu

ABSTRACT

Malignant progression of normal tissue is typically driven by complex networks of somatic changes, including genetic mutations, copy number aberrations, epigenetic changes, and transcriptional reprogramming. To delineate aberrant multi-omic tumor features that correlate with clinical outcomes, we present a novel pathway-centric tool based on the multiple factor analysis framework called *padma*. Using a multi-omic consensus representation, *padma* quantifies and characterizes individualized pathway-specific multi-omic deviations and their underlying drivers, with respect to the sampled population. We demonstrate the utility of *padma* to correlate patient outcomes with complex genetic, epigenetic, and transcriptomic perturbations in clinically actionable pathways in breast and lung cancer.

Keywords: Multi-omic data, multiple factor analysis, pathways, cancer genomics

BACKGROUND

Large sets of patient-matched multi-omics data have become widely available for large-scale human health studies in recent years, with notable examples including the The Cancer Genome Atlas (TCGA)[1] and Trans-omics for Precision Medicine (TOPMed) program. The increasing emergence of multi-omic data has in turn led to a renewed interest in multivariate, multi-table approaches[2] to account for interdependencies within and across data types[3]. In such large-scale multi-level data, there is often limited or incomplete *a priori* knowledge of relevant phenotype groups for comparisons, and a primary goal may be to identify subsets of individuals that share common molecular characteristics, design therapies in the context of personalized medicine, or identify relevant biological pathways for follow-up. With these goals in mind, many multivariate approaches have the advantage of being unsupervised, using matched or partially matched omics data across genes, obviating the need for predefined groups for comparison as in the framework of standard differential analyses. A variety of such approaches has been proposed in recent years. For example, Multi-omics Factor Analysis (MOFA) uses group factor analysis to infer sets of hidden factors that capture biological and technical variability for downstream use in sample clustering, data imputation, and sample outlier detection[4].

In multi-omic integrative analyses, an intuitive first approach is to consider a gene-centric analysis, as we previously proposed in the *EDGE in TCGA* tool[5]. Expanding such analyses to the pathway-level is also of great interest, as it can lead to improved biological interpretability as well as reduced or condensed gene lists to facilitate the generation of relevant hypotheses. In particular, our goal is to define a method that quantifies an individual's deviation from a sample average, at the pathway-level, while simultaneously accounting for multiple layers of molecular information. Several related approaches for pathway-specific single-sample analyses have been proposed in recent years[6–8]. For example, PARADIGM[7] is a widely used approach based on structured probabilistic factor graphs to prioritize relevant pathways involved in cancer progression as well as identify patient-specific alterations; both pathway structures and multi-omic relationships are hard-coded directly in the model, but it requires a discretization of the data and is now a closed-source software, making extensions and application to other gene sets difficult. Pathway relevance ranking[9] integrates binarized tumor-related omics data into a comprehensive network representation of genes, patient samples, and prior knowledge to calculate the relevance of a given pathway to a set of individuals. A pathway-centric supervised principal component-based analysis implemented in *pathwayPCA*[10] performs gene selection and estimates latent variables for association testing with respect to binary, continuous, and survival outcomes within each set of omics data independently. Pathifier[6] instead seeks to calculate a personal pathway deregulation score (PDS), based on the distance of a single individual from the median reference sample on a principal curve; this principal curve approach is analogous to a nonlinear principal components analysis (PCA), but can be applied only to a single-omic dataset (e.g., gene expression). For both PARADIGM and Pathifier, clusters of scores across pathways are shown to correlate with clinically relevant clustering of patients.

Here, we extend the basic philosophy of the Pathifier approach to multi-omics data, using an innovative application of a Multiple Factor Analysis (MFA), to quantify individualized pathway deviation scores. In particular, we propose an approach called *padma* (“PAthway Deviation scores using Multiple factor Analysis”) to characterize individuals with aberrant multi-omic profiles for a given pathway of interest and to quantify this deviation with respect to the sampled population using a multi-omic consensus representation. We further investigate the following succession of questions. In which pathways are high deviation scores strongly associated with measures of poor prognosis? For such pathways, which specific individuals are characterized by the most highly aberrant multi-omic profile? And for such individuals, which specific genes and omics drive large pathway deviation scores? By providing graphical and numerical outputs to address these questions, *padma* represents both an approach for generating hypotheses as well as an exploratory data analysis tool for identifying individuals and genes/omics of potential interest for a given pathway.

There is already some precedent for using MFA to integrate multi-omic data, although existing approaches differ from that proposed here. For instance, de Tayrac et al. suggested using MFA for paired CGH array and microarray data, superimposed with functional gene ontology terms, to highlight common structures and provide graphical outputs to better understand the relationships between omics[11]. In addition, *padma* shares some similarities with a recently proposed integrative multi-omics unsupervised gene set analysis called *mogsa*, which is similarly based on a MFA[12]. By calculating an integrated multi-omics enrichment score for a given gene set with respect to the full gene list, *mogsa* identifies gene sets driven by features that explain a large proportion of the global correlated information among different omics. In addition, these integrated enrichment scores can be decomposed by omic and used to identify differentially expressed gene sets or reveal biological pathways with correlated profiles across multiple complex data sets. However, the fundamental difference in the two approaches is that *mogsa* evaluates pathway-specific enrichment with respect to the entire set of genes, while *padma* instead focuses on identifying and quantifying pathway-specific multi-omic deviations between each individual and the sampled population.

RESULTS AND DISCUSSION

Description of the approach

Pathway-centric multiple factor analysis for multi-omic data

MFA represents an extension of principal component analysis for the case where multiple quantitative data tables are to be simultaneously analyzed [13–16]. As such, MFA is a dimension reduction method that decomposes the set of features from a given gene set into a lower dimension space. In particular, the MFA approach weights each table individually to ensure that tables with more features or those on a different scale do not dominate the analysis; all features within a given table are given the same weight. These weights are chosen such that the first eigenvalue of a PCA performed on each weighted table is equal to 1, ensuring that all tables play an equal role in the global multi-table analysis. According to the desired focus of the analysis, data can be structured either with molecular assays (e.g., RNA-seq, methylation, miRNA-seq, copy number alterations) as tables (and genes as features within omics), or with genes as tables (and molecular assays as features within genes). The MFA weights balance the contributions of each omic or of each gene, respectively. In this work, we focus on the latter strategy in order to allow different omics to contribute to a varying degree depending on the chosen pathway. In addition, we note that because the MFA is performed on standardized features, simple differences in scale between omics (e.g., RNA-seq log-normalized counts versus methylation logit-transformed beta values) do not impact the analysis.

More precisely, consider a pathway or gene set composed of *p* genes (Figure 1A), each of which is measured using up to *k* molecular assays (e.g., RNA-seq, methylation, miRNA-seq, copy number alterations), contained in the set of gene-specific matrices $X_1 ,\dots, X_p$ that have the same *n* matched individuals (rows) and $j_1 ,\dots, j_p$ potentially unmatched variables (columns) in each, where $j_g \in \{1,\dots,k\}$ for each gene $g = 1,\dots,p$. Because only the observations and not the variables are matched across data tables, genes may be represented by potentially different subset of omics data (e.g., only expression data for one gene, and expression and methylation data for another).

In the first step, these data tables are generally standardized (i.e., centered and scaled). Next, an individual PCA is performed using singular value decomposition for each gene table X_g, and its largest singular value λ_g^1 is calculated (Figure 1B). Then, all features in each gene table $X_g$are weighted by $\frac{1}{\lambda_g^1}$, and a global PCA is performed using a singular value decomposition on the concatenated set of weighted standardized tables, $X^* = \left[\frac{X_1}{\lambda_1^1}, \ldots, \frac{X_p}{\lambda_p^1}\right]$ (Figure 1C). This yields a matrix of components (i.e., latent variables) in the observation and variable space. Optionally, an independent set of supplementary individuals (or supplementary variables) can then be projected onto the original representation; this is performed by centering and scaling variables for the supplementary individuals (or individuals for the supplementary variables, respectively) to the same scale as for the reference individuals, and projecting these rescaled variables into the reference PCA space. Note that in the related *mogsa* approach, supplementary binary variables representing gene membership in gene sets are projected onto a transcriptome-wide multiple factor analysis to calculate gene set scores[12].

The MFA thus provides a consensus across-gene representation of the individuals for a given pathway, and the global PCA performed on the weighted gene tables decomposes the consensus variance into orthogonal variables (i.e., principal components) that are ordered by the proportion of variance explained by each. The coordinates of each individual on these components, also referred to as factor scores, can be used to produce factor maps to represent individuals in this consensus space such that smaller distances reflect greater similarities among individuals. In addition, *partial factor scores*, which represent the position of individuals in the consensus for a given gene, can also be represented in the consensus factor map; the average of partial factor scores across all dimensions and genes for a given individual corresponds to the factor score

(Figure 1D). A more thorough discussion of the MFA, as well as its relationship to a PCA, may be found in the Supplementary Methods.

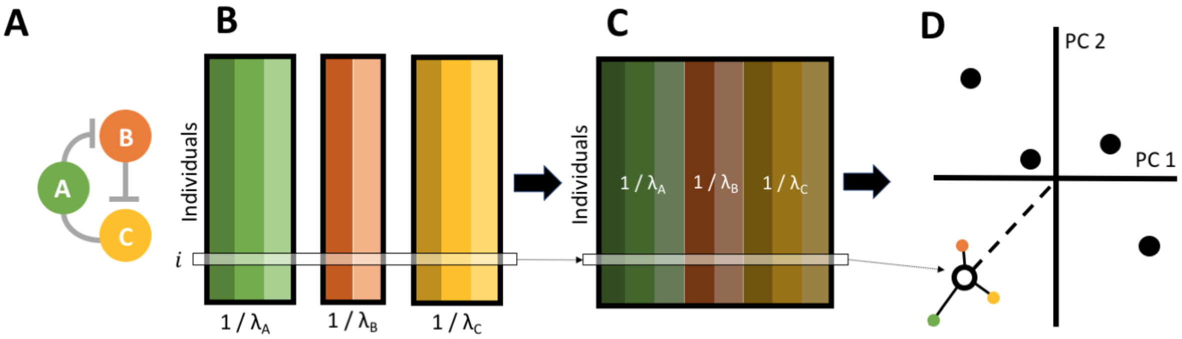

Figure 1. Illustration of the *padma* approach for calculating individualized multi-omic pathway deviation scores. (A-B) For a given pathway, matched multi-omic measures for each gene are assembled, with individuals in rows. Note that genes may be assayed for varying types of data (e.g., measurements for one gene may be available for expression, methylation, and copy number alterations, while another may only have measurements available for expression and methylation). (C) Using a Multiple Factor Analysis, each gene table is weighted by its *largest singular value*, and per-gene weighted tables are combined into a global table, which in turn is analyzed using a Principal Component Analysis. (D) Finally, each individual *i* is projected onto the consensus pathway representation; the individualized pathway deviation score is then quantified as the distance of this individual from the average individual. These scores can be further decomposed into parts attributed to each gene in the pathway.

Individualized pathway deviation scores

In the consensus space obtained from the MFA, the origin represents the “average” pathway behavior across genes, omics, and individuals; individuals that are projected to increasingly distant points in the factor map represent those with increasingly aberrant values, with respect to this average, for one or more of the omics measures for one or more genes in the pathway. To quantify these aberrant individuals, we propose an individualized pathway deviation scored_i based on the multidimensional Euclidean distance of the MFA component loadings for each individual to the origin:

$$d_i^2 = \sum_{l=1}^{L} f_{i,l}^2,$$

where $f_{i,l}$corresponds to the MFA factor score of individual *i* in component *l*, and *L* corresponds to the rank of X^*. Note that this corresponds to the weighted Euclidean distance of the scaled multi-omic data (for the genes in a given pathway) of each individual to the origin. These individualized pathway deviation scores are thus nonnegative, where smaller values represent individuals for whom the average multi-omic pathway variation is close to the average, while larger scores represent individuals with increasingly aberrant multi-omic pathway variation with respect to the average. An individual with a large pathway deviation score is thus characterized by one or more genes, with one or more omic measures, that explain a large proportion of the global correlated information across the full pathway.

Note that the full set of components is used for this deviation calculation, rather than subsetting to an optimal number of components; we remark that due to their small variance relative to lower dimensions, components from larger dimensions contribute relatively little to the overall pathway deviation scores. Finally, to facilitate comparisons of scores calculated for pathways of differing sizes (e.g., the number of genes), deviation scores with respect to the origin are normalized for the pathway size.

Decomposition of individualized pathway deviation scores into per-gene contributions

In order to quantify the role played by each gene for each individual, we decomposed the individualized pathway deviation scores into gene-level contributions. Recall that the average of partial factor scores across all MFA dimensions corresponds to each individual's factor score. We define the gene-level deviation for a given individual as follows:

$$d_{i,g} = \frac{\sum_{l=1}^{L} f_{i,l}(f_{i,l,g} - f_{i,})}{\sum_{l=1}^{L} f_{i,l}^{2}},$$

where as before $f_{i,l}$corresponds to the MFA factor score of individual *i* in component *l*, *L* corresponds to the rank of X^{*}, and $f_{i,l,g}$corresponds to the MFA partial factor score of individual *i* in gene *g* in component *l*. Note that by construction, the contributions of all pathway genes to the overall deviation score sum to 0. In particular, per-gene contributions can take on both negative and positive values according to the extent to which the gene influences the deviation of the overall pathway score from the origin (i.e., the global center of gravity across individuals); large positive values correspond to tables with a large influence on the overall deviation of an individual, while large negative values correspond to genes that tend to be most similar to the global average. In the following, we additionally scale these per-gene scores by the inverse overall pathway score to highlight genes with highly atypical multi-omic measures both with respect to other genes in the pathway and with respect to individuals in the population.

Quantifying percent contribution of omics to pathway-centric multiple factor analysis

The richness of MFA outputs also includes various decompositions of the total variance (that is, the sum of the variances of each individual MFA component) of the multi-omic data for a given pathway. Similarly to a standard PCA, the percent contribution of each axis of the MFA can be calculated as the ratio between the variance of the corresponding MFA component and the total variance; by construction, the fraction of explained variance explained decreases as the MFA dimension increases. Similarly, the percent contribution to the inertia of each axis for a given omic, gene, or individual can be quantified as the ratio between the inertia of its respective partial projection in the consensus space and the inertia of the full data projection for that axis. These per-gene, per-omic, and per-individual contributions can be quantified for a subset of components (e.g., the first ten dimensions) or for the entire set of components; here, as we calculate individualized pathway deviation scores using the full set of dimensions, we also calculated a weighted per-omic contribution, which corresponds to the average contribution across all dimensions, weighted by the corresponding eigenvalue.

Application

TCGA data acquisition and pre-processing

We illustrate the utility of *padma* on data from two cancer types with sufficiently large multi-omic sample sizes in the TCGA database: invasive breast carcinoma (BRCA), which was chosen as individuals have previously been classified into one of five molecular subtypes [17] (Luminal A, Luminal B, Her2+, Basal, and Normal-like), as well as lung adenocarcinoma (LUAD), which was chosen for its high recorded mortality.

The *padma* approach integrates multi-omic data by mapping omics measures to genes in a given pathway. Although this assignment of values to genes is straightforward for RNA-seq, CNA, and methylation data, a definitive mapping of miRNA-to-gene relationships does not exist, as miRNAs can each potentially target multiple genes. Many methods and databases based on text-mining or bioinformatics-driven approaches exist to predict miRNA-target pairs [18]. Here, we make use of the curated miR-target interaction (MTI) predictions in miRTarBase (version 7.0)[19], using only exact matches for miRNA IDs and target gene symbols and predictions with the “Functional MTI” support type. Although the TCGA data used here have been filtered to include only those genes for which expression measurements are available, there are cases where missing values are recorded in other omics datasets (e.g., when no methylation probe was available in the promoter region of a gene, or when no predicted MTIs were identified) or where a given feature has little or no variance across individuals. In this analysis, features for a given omics dataset were removed from the analysis only if missing values are recorded for all individuals or if the feature has minimal variance across all individuals (defined here as $< 10^{-5}$ after scaling). After running *padma*, we remark that the first ten MFA dimensions represent a large proportion of the total multi-omic variance across pathways for both cancers (Supplementary Figure 5; BRCA median = 46.1%, LUAD median = 51.9%).

As a measure of patient prognosis, we focused on two different metrics. First, we used the standardized and curated clinical data included in the TCGA Pan-Cancer Clinical Resource (TCGA-CDR)[20] to identify the progression-free interval (PFI). The PFI corresponds to the period from the date of diagnosis until the date of the first occurrence of a new tumor event (e.g., locoregional recurrence, distant metastasis) and typically has a shorter minimum follow-up time than measures such as overall survival. In the BRCA data, a total of 72 uncensored and 434 censored events were recorded (median PFI time of 792 and 915 days, respectively); among LUAD individuals, a total of 65 uncensored and 79 censored events were recorded (median PFI time of 439 and 683 days, respectively). Second, we used the histological grade for breast cancer, which is an established cancer hallmark of cellular de-differentiation and poor prognosis[21] (downloaded from http://legacy.dx.ai/tcga_breast on March 7, 2019). Tumors are typically graded by pathologists on a scale of 1 (well-differentiated), 2 (moderately differentiated), or 3 (poorly differentiated) based on three different measures, including nuclear pleomorphism, glandular/tubule formation, and mitotic index, where higher grades correspond to faster-growing cancers that are more likely to spread (Supplementary Table 1).

Large deviation scores for relevant oncogenic pathways are associated with survival in lung cancer

The first question we address is the prioritization of pathways that are associated with a given phenotype of interest. After processing the TCGA data and assembling the collection of gene sets, we sought to identify a subset of pathways for which deviation scores were significantly associated with patient outcome, as measured by PFI. To focus on pathways with the largest potential signal (i.e., those for which a small number of individuals have very large deviation scores relative to the remaining individuals) we consider only those with the most highly positively skewed distribution of deviation scores. For each of the top 5% of pathways (n = 57) ranked according to their Pearson's moment coefficient of skewness, we fit a Cox proportional hazards (PH) model for the PFI on the pathway deviation score. Using the Benjamini-Hochberg[22] adjusted p-values from a likelihood ratio test (FDR < 5%), we identified 14 pathways with deviation scores that were significantly associated with the progression-free interval in lung cancer (Table 1; see Supplementary Table 2 for the full gene lists in each pathway); for all of these, higher pathways scores corresponded to a worse survival outcome. Note that the filtering on skewness of the pathway scores is performed completely independently of the survival phenotype, ensuring that the downstream survival analysis is not biased [23]. Of note, while candidates within the majority deviated pathways (Table 1) have been univariately associated with patient outcome (e.g., cell cycle, DNA repair, and apoptosis [24,25]), the *padma* TCGA analysis is unique in its ability to extend these associations across multiple gene patient-specific perturbations within a pathway at the genomic and transcriptomic RNA levels.

The detection of several pathways related to DNA repair (ATM, Homologous DNA repair, BRCA1/2-ATR; Table1), as well as cell cycle and apoptosis related pathways, prompted us to consider is whether these pathway deviation scores are simply acting as proxies for the tumor mutational burden (i.e., the total number of nonsynonymous mutations) for each individual. To investigate this, we estimated the mutational burden for each individual by counting the number of somatic nonsynonymous mutations in a set of cancer-specific driver genes (*n*=183 and *n*=181 genes in breast and lung cancer, respectively) identified by IntOGen[26]. After adding a constant of 1 to these counts and log-transforming them, we fit a linear model to evaluate their association with the pathway deviation scores; after correcting p-values from the Wald test statistic for multiple testing (FDR < 10%), no pathways were found to be associated with the mutational burden. In addition, when repeating the Cox PH model described above including the log-mutational burden as an additional covariate, adjusted p-values were generally similar to previous values, and the top six pathways remained significant at a significance threshold of 5%. This suggests that the biological signal contained in the pathway deviation scores is indeed independent of that linked to mutational burden.

Pathway name	**Pathway database**	**Adj. p-value**	**Hazard ratio**	**# of genes**
D4-GDI (GDP dissociation inhibitor) signaling pathway	Biocarta	0.0111	1.2692	13
NF-kB activation through FADD/RIP-1 pathway mediated by caspase-8 and -10	Reactome	0.0111	1.2839	12
Class I PI3K signaling events mediated by Akt	PID	0.0251	1.1700	35
ATM signaling pathway	Biocarta	0.0265	1.1644	20
CARM1 and regulation of the estrogen receptor	Biocarta	0.0265	1.1426	35
Homologous recombination repair of replication-independent double-strand breaks	Reactome	0.0265	1.2432	16
Role of BRCA1, BRCA2, and ATR in cancer susceptibility	Biocarta	0.0467	1.1823	21

CD40L signaling pathway	Biocarta	0.0467	1.1880	15
Induction of apoptosis through DR4 and DR4/5 death receptors	Biocarta	0.0467	1.1208	33
Cell cycle: G1/S check point	Biocarta	0.0467	1.1263	28
Double stranded RNA induced gene expression	Biocarta	0.0467	1.2007	10
Signaling events mediated by HDAC class III	PID	0.0467	1.1543	25
HIV-1 Nef: Negative effector of Fas and TNF-alpha	PID	0.0467	1.1268	35
Regulation of telomerase	PID	0.0467	1.0950	68

Table 1. Pathways whose deviation scores are significantly correlated with progression-free interval in lung cancer. Hazard ratios and adjusted p-values correspond to a Cox PH model for pathway deviation alone, with FDR < 5%. The number of genes for each pathway corresponds to the number of genes with expression quantified by RNA-seq in the TCGA data.

Padma identifies individualized aberrations in the D4-GDP dissociation inhibitor signaling pathway in lung cancer

To illustrate the full range of results provided by *padma*, we focus in particular on the results for the *D4-GDP dissociation inhibitor (GDI) signaling* pathway. D4-GDI is a negative regulator of the ras-related Rho Family of GTPases, and it has been suggested that it may promote breast cancer cell proliferation and invasiveness [27,28]. The D4-GDI signaling pathway is made up of 13 genes; RNA-seq, methylation, and CNA measures are available for all 13 genes, with the exception of CYCS and PARP1, for which no methylation probes were measured the promoter region. In addition, miRNA-seq data were included for one predicted target pair: hsa-mir-421 → CASP3. Over the 13 genes in the pathway, 130 of the 144 individuals had no nonsynonymous mutations, while 13 and 1 individuals had 1 or 3 such mutations; ARHGAP5 and CASP3 were most often characterized by mutations (3 individuals affected for each). Notably, although the D4-GDI pathway has been previously implicated in breast cancer aggressiveness [27,28], this is to our knowledge the first evidence suggesting that D4-GDI pathway might play a similar role in promoting lung cancer.

Using the multi-omic data available for the D4-GDI signaling pathway, we can use the outputs of *padma* to better understand the individualized drivers of multi-omic variation. In particular, it is possible to quantify both gene-specific deviation scores as well as an overall pathway deviation score for each individual, respectively based on the set of partial or full MFA components. We first visualize the scaled gene-specific deviation scores for the top and bottom decile of individuals, according to their overall pathway deviation score (Figure 2); these groups thus correspond to the individuals that are least and most similar to the average individual within the population. We remark that the 10% of individuals with the most aberrant overall scores for the D4-GDI signaling pathway, who also had a high 1- and 5-year mortality rate, are those that also tend to have large aberrant (i.e., red in the heatmap) scaled gene-specific deviation scores for one or more genes. For example, the two individuals with the largest overall scores, TCGA-78-7536 and TCGA-78-7155 (12.79 and 12.31, respectively), both had large scaled gene-specific scores for CASP3 (12.93 and 17.05, respectively), CASP1 (27.80 and 10.85, respectively), and CASP8 (29.72 and 22.61, respectively). While a subset of five individuals from the top decile were all characterized by high deviation scores for JUN (TCGA-64-5775, TCGA-55-6972, TCGA-50-5051, TCGA-44-6779, TCGA-49-4488), several other genes appear to have relatively small deviation scores for all individuals plotted here (e.g., PRF1, PARP1). In addition, we remark the presence of highly individualized gene-specific aberrations (e.g., APAF1 in individual TCGA-55-7725).

To provide an intuitive link between these gene-specific deviation scores with the original batch-corrected multi-omics data that were input into *padma*, we further focus on the three genes (CASP1, CASP3, and CASP8) for which large deviation scores were observed for the two highly aberrant individuals (TCGA-78-7536 and TCGA-78-7155) in the D4-GDI signaling pathway. We plot boxplots of the Z-scores for each available omic for the three genes across all 144 individuals with lung cancer (Figure 3), specifically highlighting the two aforementioned individuals; full plots of all 13 genes in the pathway are included in Supplementary Figure 1. This plot reveals that both individuals are indeed notable for their overexpression, with respect to the other individuals, of miRNA hsa-mir-421 (Figure 3D), which is predicted to target CASP3; in coherence with this, both individuals had weaker CASP3 expression than average (although we note that its expression was not particularly extreme with respect to the full sample). Individual TCGA-78-7536 appears to have a hypomethylated CASP1 promoter, but a significantly higher number of copies of CASP8, while individual TCGA-78-7155 is characterized by a large underexpression of CASP8 with respect to other individuals. Both individuals appear to have deletions of CASP3, and hypermethylated CASP8 promoters. This seems to indicate that, although the large overall pathway deviations for these two individuals share some common etiologies, each also exhibit unique characteristics.

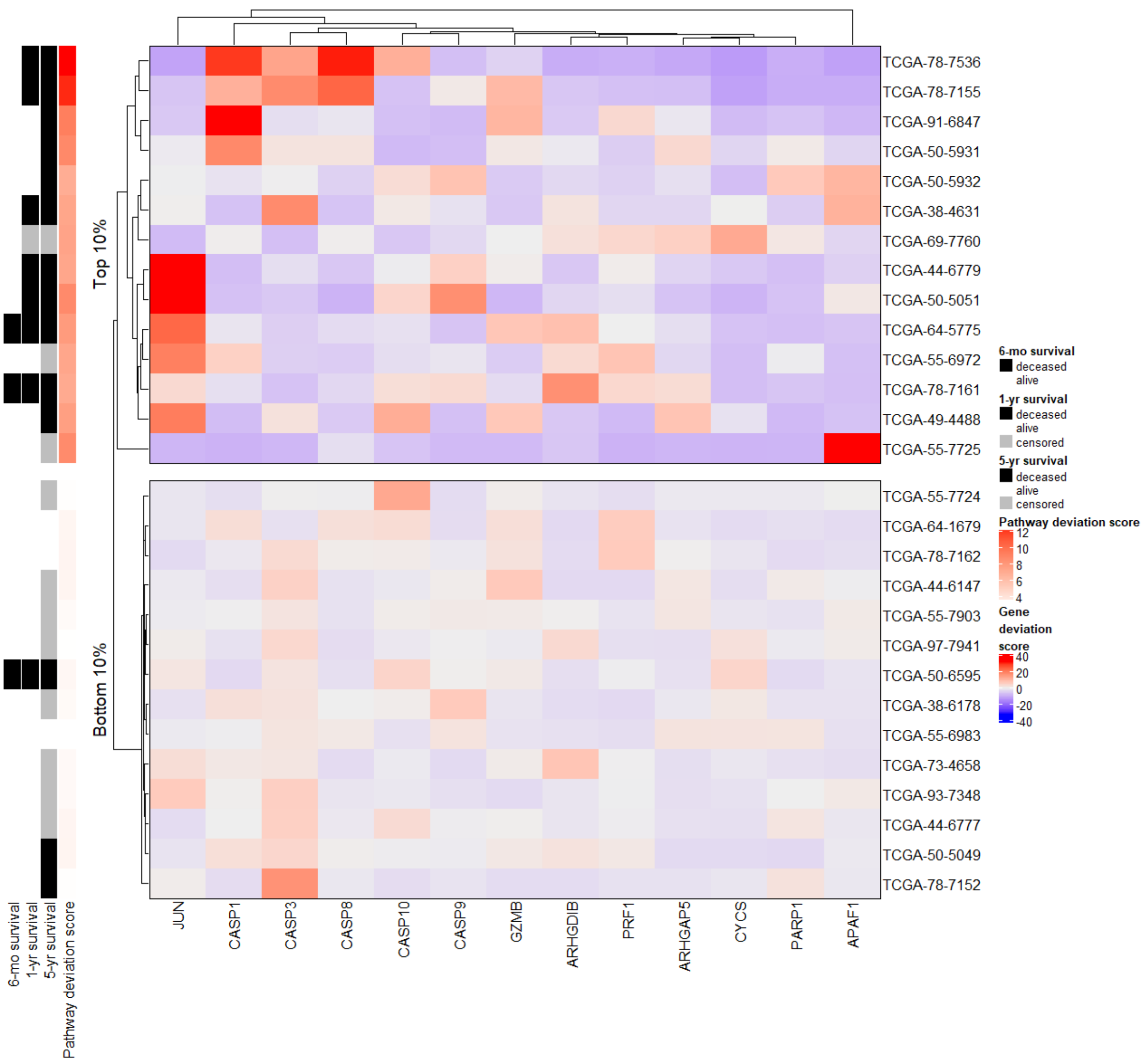

Figure 2. Scaled per-gene deviation scores for the D4-GDI signaling pathway for individuals corresponding to the top and bottom decile of overall pathway deviation scores. Red scores correspond to highly aberrant gene scores with respect to each individual's global score, while blue indicates gene scores close to the overall population average. Annotations on the left indicate the 6-month, 1-year, and 5-year survival status (deceased, alive, or censored) and overall pathway deviation score for each individual. Genes and individuals within each sub-plot are hierarchically clustered using the Euclidean distance and complete linkage.

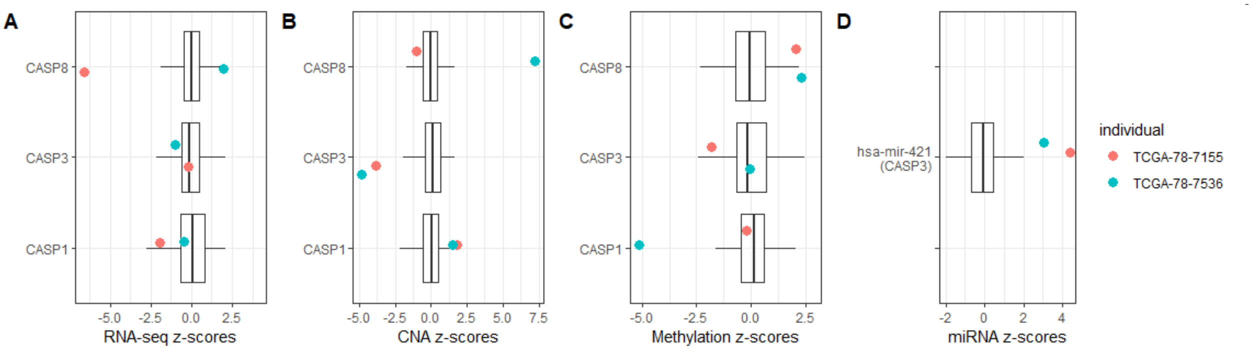

Figure 3. Boxplots of Z-scores of gene expression (A), copy number alterations (B), methylation (C), and miRNA expression (D) for all individuals with lung cancer, with the 3 genes (CASP1, CASP3, CASP8) and one miRNA (hsa-mir-421, predicted to target CASP3) of interest in the D4-GDI signaling pathway. The two individuals with the largest pathway deviation score (TCGA-78-7155, TCGA-78-7536) are highlighted in red and turquoise, respectively.

As overall pathway deviation scores represent the multi-dimensional average of these gene-specific deviation scores, a deeper investigation into them can also provide useful insight for a given pathway. We first note that the distribution of deviation scores for the D4-GDI signaling pathway (Figure 4A) is highly skewed, with a handful of individuals (e.g., TCGA-78-7536, TCGA-78-7155, TCGA-91-6847, TCGA-50-5931, TCGA-50-5051, and TCGA-66-7725) characterized by particularly large scores with respect to the remaining individuals. The individual with the most aberrant score for this pathway, TCGA-78-7536, had a single pathway-specific somatic mutation in the CASP1 gene, and a total of 7 cancer-specific driver gene mutations (corresponding to the 80th percentile of individuals considered here). Although these pathway deviation scores are calculated across all dimensions of the MFA, it can also be useful to represent individuals in the first few components of the consensus MFA space (Figure 4B); the farther away an individual is from the origin over multiple MFA dimensions, the larger the corresponding pathway deviation score. In this case, we see that TCGA-78-7536 is a large positive and negative outlier in the second (9.55% total variance explained), and third (8.07% total variance explained) MFA components, respectively, although less so in the first component (11.97% total variance explained). In addition, we note that RNA-seq is the major driver of the first MFA dimension (54.38% contribution), while promoter methylation and copy number alterations take a larger role in the second and third dimensions (42.29% and 59.18% contribution, respectively). miRNA expression appears to play a fairly minor role in the MFA, with its maximum contribution (21.14%) occurring at only the 16th dimension.

When examining the partial factor maps for this individual over the first three MFA dimensions (Figures 4C-D), we note the large contribution of CASP3 (axis 1), CASP10 (axis 2), CASP1 and CASP 8 (axis 3), as evidenced by their distance from the origin in these dimensions. Overall, this is coherent with the previous gene-level analyses (Figure 2), where hypomethylation in CASP1 and large copy number gains for CASP3 and CASP8 with respect to the population were identified for this individual. Other individuals with large overall deviation scores (e.g., TCGA-50-5931) are not obvious outliers in the first two MFA dimensions, reflecting the fact that additional dimensions play a more important role for them. Taken together, the individualized gene-specific and overall pathway deviation scores output by *padma* provide complementary and interesting exploratory insight into atypical multi-omic profiles for a given pathway of interest (here, the D4-GDI signaling pathway in lung cancer).

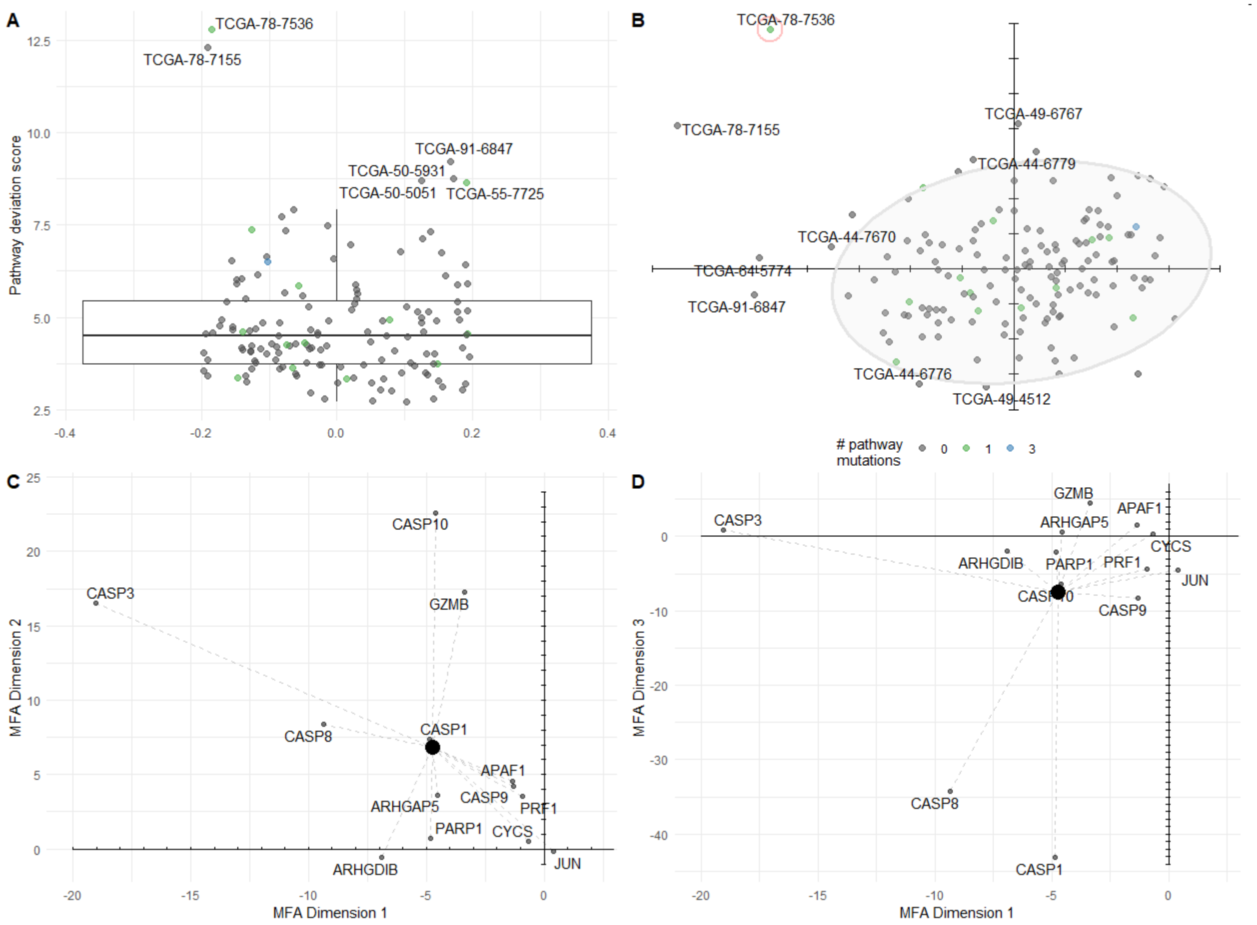

Figure 4. (A) Distribution of pathway deviation scores for the D4-GDI signaling pathway in lung cancer; individuals with unusually large scores are labeled. (B) Factor map, representing the first two components of the MFA for the D4-GDI signaling pathway in lung cancer, with normal confidence ellipse superimposed. Individuals with extreme values in each plot are labeled with their barcode identifiers and colored by the number of pathway-specific nonsynonymous mutations. For the individual circled in red, TCGA-78-7536, a partial factor map representing the first MFA components 1 and 2 is plotted in (C), and MFA components 1 and 3 in (D). The large black dot represents the individual's overall pathway deviation score, as plotted in panel (B) for the first two axes, and gene-specific scores are joined to this point with dotted lines.

Pathway deviation scores globally recapitulate histological grade in breast cancer

For some cancers, additional clinical phenotypes beyond survival information may be of particular interest; to illustrate the use of *padma* in such a case, we focus on histological grade for breast cancer. To quantify whether pathway deviation scores tend to be associated with histological grade in breast cancer, we performed a one-way ANOVA on the three measures that comprise histological grade for each of the 1136 pathways. Based on the Benjamini-Hochberg[22] adjusted p-values from an F-test (FDR < 5%), all (1136) or nearly all (1135) pathways were found to have deviation scores that are significantly correlated with mitotic index and nuclear pleomorphism. Intriguingly, no pathways were found to be associated with degree of glandular/tubule formation; this may in part be due to the large proportion of individuals identified as grade III (poorly differentiated) for this measure (n = 285). The rankings of pathways based on mitotic index and nuclear pleomorphism were generally in agreement (Supplementary Figure 2). In all but two cases, higher deviation pathway scores corresponded to the higher grades for these two measures, corresponding to more aggressive tumors; the two exceptions were the *Presynaptic nicotinic acetylcholine recepto*r and *Highly calcium permeable postsynaptic nicotinic acetylcholine receptor* pathways (both from Reactome), for which the largest pathway deviation scores were associated with grade II, rather than grade III, of the mitotic index.

To prioritize pathways among this list, we calculated the rank product of the individual rankings by p-value for mitosis and nuclear pleomorphism; the top 10 pathways according to this joint ranking are shown in Table 2 (see Supplementary Table 3 for the full gene lists in each pathway). The *signaling by Wnt* pathway, which is made up of 63 genes, had the highest combined ranking for these two histological measures. Of this set of genes, all had RNA-seq, methylation, and CNA measures available, with the exception of FAM123B and PSMD10 (no CNA measures with nonzero variance) and PSMB1 to PSMB10, PSMC2, PSMC3, PSMC5, PSMC6, PSME1, and

PSME2 (no promoter methylation measures). miRNA-seq data were included for only two predicted target pairs: hsa-mir-375 →CTNNB1 and hsa-mir-320a →CTNNB1. Over the 63 genes in the pathway, 453 individuals had no nonsynonymous mutations, while 39, 6, 3, 2, and 1 individuals had 1, 2, 3, 4, or 5 such mutations; APC, PSMD1, and FAM123B were most often characterized by mutations (10, 7, and 7 individuals affected, respectively).

Pathway name	Pathway database	Combined ranking	# of genes
Signaling by Wnt	Reactome	3.16	63
Apoptotic execution phase	Reactome	5.00	52
APC/C:Cdh1 mediated degradation of Cdc20 and other APC/C:Cdh1 targeted proteins in late mitosis/early G1	Reactome	6.78	64
Genes involved in Beta-catenin phosphorylation cascade	Reactome	10.49	16
Autodegradation of Cdh1 by Cdh1:APC/C	Reactome	10.95	56
Genes involved in M/G1 transition	Reactome	11.62	72
Regulation of the Fanconi anemia pathway	Reactome	13.93	7
Apoptotic cleavage of cellular proteins	Reactome	14.14	38
Apoptosis	Reactome	14.28	143
ER-phagosome pathway	Reactome	15.62	58

Table 2. Pathways whose deviation scores are significantly correlated with measures of histological grade (mitosis, nuclear pleomorphism) in breast cancer. Adjusted p-values after Benjamini-Hochberg correction were $< 3.31 \times 10^{-12}$ for all pathways presented in the table. Combined ranks correspond to the rank product of the individual rankings from mitosis and nuclear pleomorphism, and the number of genes for each pathway corresponds to the number of genes with expression quantified by RNA-seq in the TCGA data.

Similarly to the distribution of D4-GDI pathway scores in lung adenocarcinomas, a small number of breast cancer patients are characterized by highly aberrant scores in the signaling by Wnt pathway, including TCGA-BH-A1FM, TCGA-E9-A22G, and TCGA-EW-A1PH, and the number of pathway-specific nonsynonymous somatic mutations does not appear to be related to this score. The associated factor map on the first two dimensions of the MFA (Figure 5A) clearly captures relevant biological structure from the data, as evidenced by the quasi-separation of individuals in different intrinsic inferred molecular subtypes (AIMS). Notably, individuals with Basal and Luminal A breast cancer are clearly separated in the first two dimensions and tend to respectively have positive and negative loadings in the first dimension of the MFA; Luminal B and Normal-like subtypes largely overlap with the Luminal A subtype for this pathway, while Her2 is located intermediate to the Luminal and Basal subtypes, as could be anticipated due to the equal prevalence of Her2 amplification in both Luminal and Basal subtypes. Similar relevant biological signal can be seen when considering a larger spectrum of pathways (Figure 5C). In particular, individuals with the Basal and Luminal B subtypes tend to have much more highly variant deviation scores across all pathways, whereas Luminal A and Normal-like subtypes are generally much less variant.

When examining the percent contribution of each omic to the axes of the MFA for the Wnt signaling pathway (Figure 5B), we remark the preponderant contribution of gene expression to the first component (84.40%), while variability in the second component is largely driven by both gene expression and copy numbers (45.66 and 35.37%, respectively). The large role played by RNA-seq here is coherent with the definition of the AIMS subtypes themselves, which are defined on the basis of gene expression. On average, after weighting by the eigenvalue of each component, gene expression and copy number alterations were found to have similar contributions to the overall variation (36.6%, 35.4%, respectively), while methylation played a less important role (26.8%). For this pathway, as for most others we studied (Supplementary Figure 6), miRNA expression contributed relatively little to the overall variation (1.2%).

Taken together, these results illustrate that the *padma* approach, which is used in an unsupervised manner on multi-omic cancer data for a given pathway, is able to recapitulate known sample structure in the form of intrinsic tumor subtypes as well as relevant prognostic factors such as histological grade.

CONCLUSIONS

Unsupervised dimension reduction approaches (such as PCA) have been widely used in genetics and genomics for many years, both to identify sample structure and batch effects[29] and to visualize overall variation in large data[30]. Here, we present a generalization of this approach to multi-omic data for investigating biological variation at the pathway-level by aggregating across genes, omic-type, and individuals. Compared to single-omics approaches (for instance, running a PCA on RNA-seq data alone), *padma* accommodates multiple omics-sources which, for some sample sets and pathways, account for more than 50% of the overall variation (Figure 5B). Using MFA to partition variance, we construct a clinically relevant pathway disruption score that correlates with survival outcomes in lung cancer patients, and histological grade in breast cancer patients.

Our MFA-based approach allows investigators to (a) identify overall sources of variation (such as batch effects); (b) prioritize high variance pathways defined by variability across subjects; (c) identify aberrant observations (i.e., individuals) within a given pathway; and (d) identify the genes and omics sources that drive these aberrant observations. For large, multi-omic data such as TCGA, *padma* allows investigators to summarize overall variation and assist in generating hypotheses for more targeted analyses and follow-up studies. As a case in point, we identified two lung cancer patients with aberrant multi-omic profiles at three *CASP* genes. With access to the tumor samples and more fine-grained clinical data, future molecular experiments could help to clarify the role (if any) that these genes play in contributing to lung cancer mortality.

There are a number of natural extensions and alternative formulations to our MFA-based approach. If comparisons between sets of individuals (e.g., healthy vs. disease) are of interest, the MFA can be based on one set of samples (e.g., healthy, or a "reference set"), and the other set of samples (e.g., diseased, or a "supplementary set") can be projected onto this original representation. This is accomplished by centering and scaling supplementary individuals to the same scale as the reference individuals, and projecting these rescaled variables into the reference MFA space. In this setting, the interpretation of pathway deviation scores would no longer correspond to the identification of "aberrant" individuals compared to an overall average, but rather individuals that are most different from the reference set (e.g., the most "diseased" as compared to a healthy reference); this strategy would be similar in spirit to the individualized pathway aberrance score (iPAS) approach, which proposed using accumulated (unmatched) normal samples as a reference set[31]. There is also no reason to limit this approach to pathways, as the analysis could be performed just once, genome-wide (accordingly, inferences would no longer be applicable to specific pathways). Here, we have structured the data with genes representing data tables and omics representing columns within each table. Alternatively, the data could be re-weighted by having omics represented as data tables and genes as columns within each, similar to de Tayrac et al. (2009)[11]. Extensions to our work could include incorporating the hierarchical structure of genes within pathways, or relatedness structure among samples. In principle, other types of omics that do not map to genes or pathways (e.g., genotypes on single nucleotide polymorphisms) could also be incorporated. Finally, though we illustrate the use of *padma* for cancer genomics data, we anticipate that it will be broadly useful to other multi-omic applications in human health or agriculture.

MATERIALS AND METHODS

TCGA data acquisition and pre-processing

The multi-omic TCGA data were downloaded and processed as described in Rau et al. (2019)[5]. Briefly, using *TCGA2STAT*[32] we downloaded processed TCGA Level 3 data from the Broad Institute Genome Data Analysis Center (GDAC) Firehose on March 18, 2017 for individuals of self-reported European ancestry for whom gene expression, methylation, copy number alterations (CNA), microRNA (miRNA) abundance, and somatic mutation data were all available; this ancestry filter was applied to minimize population-specific variance and focus on the group with the largest available sample size. In addition, two individuals from the BRCA dataset (TCGA-E9-A245, TCGA-BH-A1ES) were identified as outliers with consistently extreme deviation scores across multiple pathways and were removed from the remainder of the analyses; the final sample sizes were thus n=504 and n=144 individuals for the BRCA and LUAD datasets, respectively.

Per-gene normalized expression estimates were calculated using RSEM[33]. Methylation was quantified using the maximally variant probe from the Illumina Infinium Human Methylation450 BeadChip located within ±1500bp of the transcription start site, and representative probe beta measures were transformed to the logit scale. Somatic CNAs were called by comparing Affymetrix 6.0 probe intensities from normal (i.e., non-cancer tissue) and cancer tissue, and genome segments were aggregated to gene-level measures by *TCGA2STAT* and *CNTools*. Individuals were classified as carriers or noncarriers of a nonsynonymous somatic mutation for each gene using *TCGA2STAT*. Normalized miRNA abundance was quantified as Reads per million microRNA mapped (RPMMM) values. RNA-seq and miRNA-seq quantifications were TMM-normalized[34], converted to counts per million (CPM), and log2-transformed. Only genes with available RNA-seq expression measures were retained for the remainder of the analysis, corresponding to 20,501 and 19,971 genes for BRCA and LUAD, respectively. Finally, batch effects have been shown to have a strong impact on the analysis of high-throughput data in general[29] and for the TCGA data specifically[35]. As specific sample plates have been shown to represent significant batch effects in previous analyses[36], each processed omic (with the exception of somatic mutation data) was individually batch adjusted for each cancer to correct for plate-specific effects using `removeBatchEffects` in limmal[37]. Plots of the first two components from a transcriptome-wide and genome-wide single-omics PCA and multi-omics MFA for the batch-corrected data are included in Supplementary Figures 3 and 4.

Choice of curated pathway collection

We consider the pathways included in the MSigDB canonical pathways curated gene set catalog[38], which includes genes whose products are involved in metabolic and signaling pathways reported in curated public databases. We specifically use the "C2 curated gene sets" catalog from MSigDB v5.2 available at http://bioinf.wehi.edu.au/software/MSigDB/ as described in the *limma* Bioconductor package[37]. We focus in particular on a collection of 1322 gene sets from public databases, including Biocarta, Pathway Interaction Database[39], Reactome[40]; Sigma Aldrich, Signaling Gateway, Signal Transduction Knowledge Environment, and the Matrisome Project[41], the smallest and largest of which were respectively made up of 6 and 478 genes (median size 29 genes). For the subsequent *padma* analysis, we excluded gene sets for which fewer than 3 genes mapped to quantified features in the TCGA gene expression data, corresponding to a total of 1136 gene sets.

Padma R software package

The proposed method described above has been implemented in an open-source R package called *padma*, freely available on GitHub. *Padma* notably makes use *FactoMineR*[3,15] to run the MFA; heatmaps in the following results were produced using *ComplexHeatmap*[42]. All of the analyses in this paper were performed using R v3.5.1.

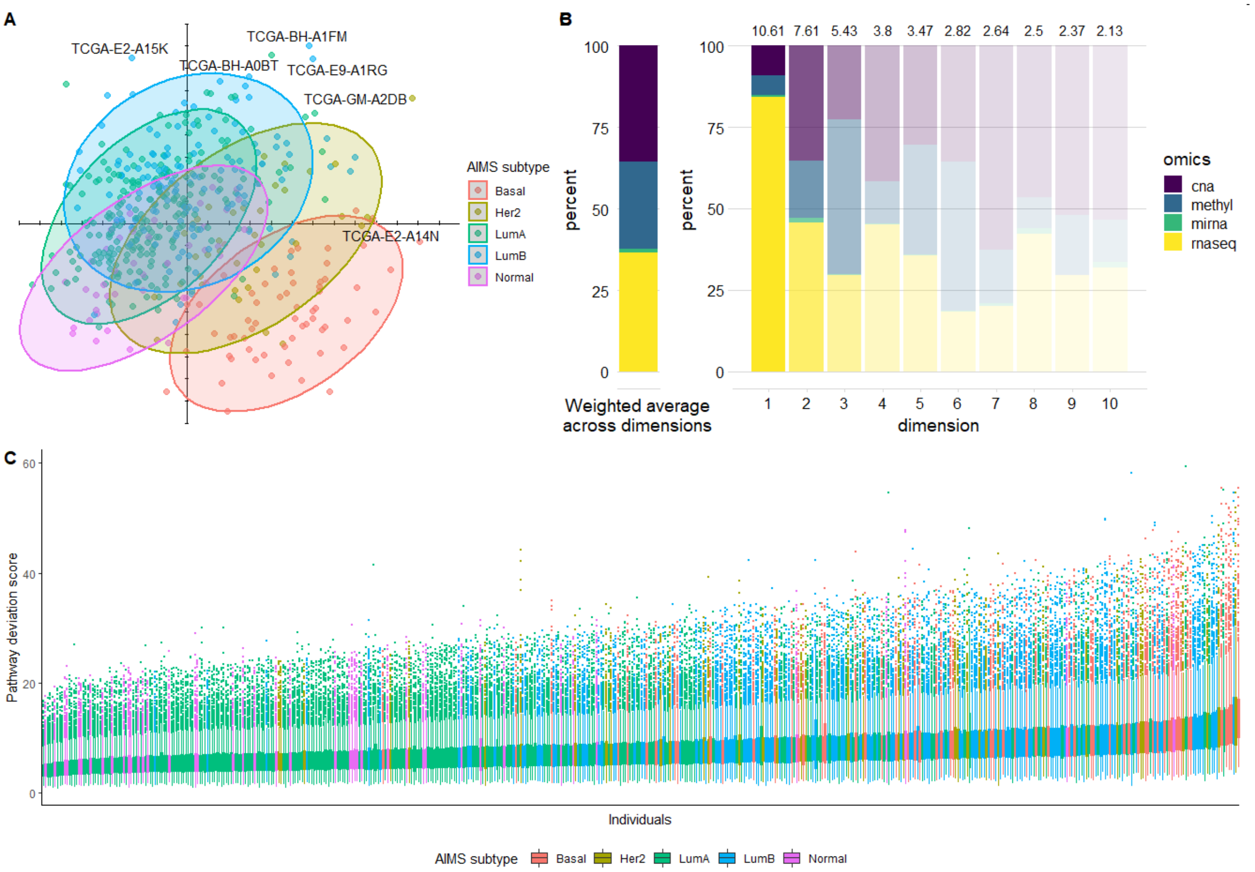

Figure 5. (A) Factor map of individuals, representing the first two components of the MFA, for the Wnt signaling pathway in breast cancer, with normal confidence ellipses superimposed for the five AIMS subtypes. B) Weighted overall percent contribution per omic (left) and for each of the first 10 MFA components (right) for the Wnt signaling pathway, with colors faded according to the percent variance explained for each (represented in text above each bar). (C) Distribution of pathway deviation scores for each individual in the breast cancer data, with individuals colored according to their AIMS subtype.

DECLARATIONS

Ethics approval and consent to participate: Not applicable.

Consent for publication: Not applicable.

Availability of data and materials: The TCGA data analysed in the current study were retrieved and pre-processed as described in the Methods section and in Rau et al. (2018)[5]; in particular, all associated scripts can be found at https://github.com/andreamrau/EDGE-in-TCGA (https://doi.org/10.5281/zenodo.3524080). All R scripts used to generate the results in this work may be found at https://github.com/andreamrau/RMFRJLA_2019, and the associated *padma* R package may be found at https://github.com/andreamrau/padma.

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

Funding: Dr. Andrea Rau was supported by the AgreenSkills+ fellowship program, which received funding from the EU's Seventh Framework Program under grant agreement FP7-60939 (AgreenSkills+ contract).

Authors' contributions: AR conceived and designed the study, wrote the *padma* R package, analyzed the data, and drafted the manuscript. RM analyzed the data and contributed to the R package development. MJF and HR interpreted results and contributed to study design. FJ contributed to the study conception and writing of the manuscript. DL supervised the study conception and method implementation and drafted the manuscript and supplementary materials. PLA conceived and designed the study and drafted the manuscript. All authors read and approved the final manuscript.

Acknowledgements: Not applicable.

REFERENCES

1. The Cancer Genome Atlas Research Network *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
2. Meng, C. *et al.* Dimension reduction techniques for the integrative analysis of multi-omics data. *Brief. Bioinform.* **17**, 628–641 (2016).
3. Husson, F., Lê, S. & Pagès, J. *Exploratory multivariate analysis by example using R*. (CRC Press, 2017).
4. Argelaguet, R. *et al.* Multi- Omics Factor Analysis—a framework for unsupervised integration of multi- omics data sets. *Mol. Syst. Biol.* **14**, (2018).
5. Rau, A., Flister, M., Rui, H. & Auer, P. L. Exploring drivers of gene expression in the Cancer Genome Atlas. *Bioinformatics* (2018) doi:10.1093/bioinformatics/bty551.
6. Drier, Y., Sheffer, M. & Domany, E. Pathway-based personalized analysis of cancer. *Proc. Natl. Acad. Sci.* **110**, 6388–6393 (2013).
7. Vaske, C. J. *et al.* Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* **26**, i237–i245 (2010).

9. Verbeke, L. P. C. *et al.* Pathway Relevance Ranking for Tumor Samples through Network-Based Data Integration. *PLOS ONE* **10**, e0133503 (2015).

10. Odom, G. J. *et al. pathwayPCA: an R package for integrative pathway analysis with modern PCA methodology and gene selection*. http://biorxiv.org/lookup/doi/10.1101/615435 (2019) doi:10.1101/615435.

11. de Tayrac, M., Le, S., Aubry, M., Mosser, J. & Husson, F. Simultaneous analysis of distinct Omics data sets with integration of biological knowledge: Multiple Factor Analysis approach. *BMC Genomics* **10**, 32 (2009).

12. Meng, C. *et al.* MOGSA: Integrative Single Sample Gene-set Analysis of Multiple Omics

Data. *Mol. Cell. Proteomics* **18**, S153–S168 (2019).

13. Escofier, B. & Pagès, J. *Analyses factorielles simples et multiples: objectifs, méthodes et interprétation*. (2014).

14. Pagès, J. *Multiple factor analysis by example using R*. (CRC Press, Taylor & Francis Group, 2015).

15. Lê, S., Josse, J. & Husson, F. **FactoMineR** : An *R* Package for Multivariate Analysis. *J. Stat. Softw.* **25**, (2008).

16. Abdi, H., Williams, L. J. & Valentin, D. Multiple factor analysis: principal component analysis for multitable and multiblock data sets: Multiple factor analysis. *Wiley Interdiscip. Rev. Comput. Stat.* **5**, 149–179 (2013).

17. Paquet, E. R. & Hallett, M. T. Absolute assignment of breast cancer intrinsic molecular subtype. *J. Natl. Cancer Inst.* **107**, 357 (2015).

18. Riffo-Campos, Á., Riquelme, I. & Brebi-Mieville, P. Tools for Sequence-Based miRNA Target Prediction: What to Choose? *Int. J. Mol. Sci.* **17**, 1987 (2016).

19. Chou, C.-H. *et al.* miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic Acids Res.* **46**, D296–D302 (2018).

20. Liu, J. *et al.* An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell* **173**, 400-416.e11 (2018).

21. Heng, Y. J. *et al.* The molecular basis of breast cancer pathological phenotypes: Molecular basis of breast cancer pathological phenotypes. *J. Pathol.* **241**, 375–391 (2017).

22. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).

23. Bourgon, R., Gentleman, R. & Huber, W. Independent filtering increases detection power for high-throughput experiments. *Proc. Natl. Acad. Sci.* **107**, 9546–9551 (2010).

24. Bosken, C. H., Wei, Q., Amos, C. I. & Spitz, M. R. An analysis of DNA repair as a determinant of survival in patients with non-small-cell lung cancer. *J. Natl. Cancer Inst.* **94**,

1091–1099 (2002).

25. Singhal, S., Vachani, A., Antin-Ozerkis, D., Kaiser, L. R. & Albelda, S. M. Prognostic implications of cell cycle, apoptosis, and angiogenesis biomarkers in non-small cell lung cancer: a review. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **11**, 3974–3986 (2005).

26. Gonzalez-Perez, A. *et al.* IntOGen-mutations identifies cancer drivers across tumor types. *Nat. Methods* **10**, 1081–1082 (2013).

27. Zhang, Y., Rivera Rosado, L. A., Moon, S. Y. & Zhang, B. Silencing of D4-GDI inhibits growth and invasive behavior in MDA-MB-231 cells by activation of Rac-dependent p38 and JNK signaling. *J. Biol. Chem.* **284**, 12956–12965 (2009).

28. Zhang, Y. & Zhang, B. D4-GDI, a Rho GTPase regulator, promotes breast cancer cell invasiveness. *Cancer Res.* **66**, 5592–5598 (2006).

29. Leek, J. T. *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **11**, 733–739 (2010).

30. Novembre, J. *et al.* Genes mirror geography within Europe. *Nature* **456**, 98–101 (2008).

31. Ahn, T., Lee, E., Huh, N. & Park, T. Personalized identification of altered pathways in cancer using accumulated normal tissue data. *Bioinformatics* **30**, i422–i429 (2014).

32. Wan, Y.-W., Allen, G. I. & Liu, Z. TCGA2STAT: simple TCGA data access for integrated statistical analysis in R. *Bioinforma. Oxf. Engl.* **32**, 952–954 (2016).

33. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).

34. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).

35. Akulenko, R., Merl, M. & Helms, V. BEclear: Batch Effect Detection and Adjustment in DNA Methylation Data. *PLOS ONE* **11**, e0159921 (2016).

36. *MBatch: TCGA Batch Effects Viewer*. (2019).

37. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing

and microarray studies. *Nucleic Acids Res.* **43**, e47–e47 (2015).

38. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).

39. Schaefer, C. F. *et al.* PID: the Pathway Interaction Database. *Nucleic Acids Res.* **37**, D674–D679 (2009).

40. Fabregat, A. *et al.* The Reactome Pathway Knowledgebase. *Nucleic Acids Res.* **46**, D649–D655 (2018).

41. Naba, A. *et al.* The Matrisome: *In Silico* Definition and *In Vivo* Characterization by Proteomics of Normal and Tumor Extracellular Matrices. *Mol. Cell. Proteomics* **11**, M111.014647 (2012).

42. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).

Supplementary Materials

Supplementary Figure 1. Z-scores of RNA-seq, CNA, methylation, and miRNA-seq data for genes in the D4-GDI signaling pathway for individuals in the TCGA LUAD data (n = 144). Data corresponding to the two individuals with the largest overall pathway deviation scores, TCGA-78-7155 and TCGA-78-7536, are highlighted in red and blue.

Supplementary Figure 2. Negative log10-transformed p-values from the ANOVA F-test of pathway deviation score versus mitosis and nuclear pleomorphism for each pathway among breast cancer individuals. The signaling by Wnt pathway is highlighted in red.

Supplementary Figure 3. Factor maps for the first two dimensions of a global transcriptome- and genome-wide PCA of the methylation, miRNA-seq, CNA, and RNA-seq data (left), as well as a global MFA of all four omics combined (right) for the TCGA BRCA data.

Supplementary Figure 4. Factor maps for the first two dimensions of a global transcriptome- and genome-wide PCA of the methylation, miRNA-seq, CNA, and RNA-seq data (left), as well as a global MFA of all four omics combined (right) for the TCGA BRCA data.

Supplementary Figure 5. Percent variance explained by the first 5 (blue) or 10 (red) components of the MFA for each pathway for the TCGA BRCA (A) and LUAD (B) data.

Supplementary Figure 6. Average percent contribution to the MFA of each omic (miRNA-seq, methylation, CNA, RNA-seq) for each pathway. (A) Per-omic average contribution across the first

10 MFA components for TCGA BRCA. (B) Per-omic average contribution across all MFA components for TCGA BRCA. (C) Per-omic average contribution across the first 10 MFA components for TCGA LUAD. (D) Per-omic average contribution across all MFA components for TCGA LUAD.

Supplementary Table 1. Sample size for each histological measure for the *n* = 504 breast cancer patients.

Supplementary Table 2. Full gene lists for pathways in Table 1. Genes correspond to those with expression quantified by RNA-seq in the TCGA data.

Supplementary Table 3. Full gene lists for pathways in Table 2. Genes correspond to those with expression quantified by RNA-seq in the TCGA data.