Supplementary Materials: 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

Table of Contents

Supplementary Methods

Introduction

This presentation of the Multiple Factorial Analysis (MFA) is mainly borrowed from Abdi et al. (2013) and Abdi and Williams (2010). MFA applies to the case when several (K) sets of variables describe a same set of I individuals (observations). MFA is part of the family of methods based on principal components analysis (PCA) and makes use of the basic notions of inertia and distance.

Inertia and distance

To start, let's consider a single data table $\mathbf{X} = \{x_{ij}\}_{I,J}$. The inertia of x_j, the j^{th} column of $\mathbf{X}$, is defined as the sum of its squared elements:

$$\gamma_j^2 = \sum_{i=1}^{I} x_{i,j}^2.$$

When the column is centered, γ_j^2 corresponds to the variance of x_j. The sum of all γ_j^2 across columns is denoted $\mathcal{I}$ and is called the *inertia* of the data table, or the total inertia.

The center of gravity of the rows, denoted **g**, is the vector of the means of each column of **X**. When **X** is column-centered (which is commonly the case), **g** is equal to the *J*-dimensional null vector. The Euclidean distance of the i^{th} observation to **g** is equal to

$$d_i^2 = \sum_{j=1}^{J} x_{i,j}^2.$$

The sum of all d_i^2 is equal to the total inertia $\mathcal{I}$ of the data table.

Principal components analysis

PCA computes new variables called *principal components*, which are obtained as linear combinations of the original variables, and is based on the concept of inertia. By design, the first principal component is required to have the largest possible inertia. The second component is then constructed to have the largest possible inertia, under the constraint of orthogonality with respect to the first component; the other components are computed in a similar manner. These new variables are called *factor scores*, which can be interpreted geometrically as the projections of the observations onto the principal components. The coefficients of the linear combinations used to compute the factor scores are called the *loadings*.

Specifically, **X** is factorized via a Singular Value Decomposition (SVD) as $\mathbf{X} = \mathbf{P}\,\Delta\,\mathbf{Q}^T$, where **P** is an $I \times L$ matrix of normalized left singular vectors (with L being the rank of **X**), **Q** the $J \times L$ matrix of normalized right singular vectors, and Δ the $L \times L$ diagonal matrix of the L ordered singular values ($\beta_{[1]}, \ldots, \beta_{[l]}, \ldots, \beta_{[L]}$), such that $\beta_{[1]} \geq \ldots \geq \beta_{[l]} \geq \ldots \geq \beta_{[L]}$. For convenience, we will denote the largest singular value of **X**, $\beta_{[1]}$, as λ. Matrices **P** and **Q** are orthonormal matrices (i.e., $\mathbf{Q}^T\mathbf{Q} = \mathbf{P}^T\mathbf{P} = \mathbf{I}$). The SVD is closely related to the canonical decomposition of $\mathbf{X}\mathbf{X}^T$, where **P** is the matrix of normalized eigenvectors of $\mathbf{X}\mathbf{X}^T$, **Q** is the matrix of normalized eigenvectors of $\mathbf{X}^T\mathbf{X}$, and the singular values are the square root of the eigenvalues of $\mathbf{X}\mathbf{X}^T$.

Loadings and factor scores

As **Q** is the matrix of loadings, the matrix of factor scores **F** is equal to:

$$\mathbf{F} = \mathbf{P}\Delta = \mathbf{X}\mathbf{Q}.$$

Deviation scores

We define the *deviation score* of an observation as the distance of its factor scores to the center of the cloud of points; that is, if we note $f_{i,l}$ the factor score of the i^{th} individual for the l^{th} component, the deviation score may be calculated as follows:

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

Note that this is equivalent to the Euclidean distance defined above, assuming **X** has been column-centered.

Multiple factor analysis

We now consider a set of *K* tables describing the same set of *I* individuals (observations). We denote the merged dataset **X** = (**X**$_{[1]}$,…,**X**$_{[k]}$,…**X**$_{[K]}$), where **X**$_{[k]}$ is the data matrix corresponding to the k^{th} set of variables, possibly preprocessed (generally centered and normalized). MFA comprises two main steps:

1. PCA of each table **X**$_{[k]}$.
   In this first step, each table **X**$_{[k]}$ is analyzed via a standard PCA. The largest singular value obtained from the SVD of **X**$_{[k]}$ is noted $\lambda_{[k]}$. We note **A** the *K* × *K* diagonal matrix whose general term **A**[*i*,*i*] is equal to $\lambda_{[i]}$.

   The MFA consists of a PCA of the merged table **X**, weighted so that the influence of each table **X**$_{[k]}$ is balanced.

2. PCA of the weighted table.
   We note **X*** the set of the merged reweighted **X***$_{[k]}$ matrices, obtained by dividing all the elements of **X**$_{[k]}$ by their largest singular value $\lambda_{[k]}$:

$$\mathbf{X}^* = \mathbf{X}\mathbf{A}^{-\mathbf{1}} = \left[\frac{\mathbf{X}_{[1]}}{\lambda_{[1]}}, \ldots, \frac{\mathbf{X}_{[k]}}{\lambda_{[k]}}, \ldots \frac{\mathbf{X}_{[K]}}{\lambda_{[K]}}\right].$$

   The MFA is then computed as the simple PCA of **X***, where the SVD of **X*** is given by **X*** = **P*Δ*Q*^{T}**, with **P*^{T}P*** = **Q*^{T}Q*** = **I**.

Loadings

Because the matrix **X*** concatenates *K* tables, with each of them respectively comprising $J_{[k]}$ variables, the matrix **Q*** of the left singular vectors can be partitioned in the same way as **X*.** Specifically, **Q*** can be expressed as a column block matrix:

$$\mathbf{Q}^* = \begin{bmatrix} \mathbf{Q}^*_{[1]} \\ \vdots \\ \mathbf{Q}^*_{[k]} \\ \vdots \\ \mathbf{Q}^*_{[K]} \end{bmatrix} = \left[\mathbf{Q}^{*\mathrm{T}}_{[1]}, \ldots, \mathbf{Q}^{*\mathrm{T}}_{[k]}, \ldots \mathbf{Q}^{*\mathrm{T}}_{[K]}\right]^{\mathrm{T}}.$$

The MFA loadings of the k^{th} table are then calculated as follows:

$$\mathbf{Q}_{[k]} = \lambda_{[k]}\mathbf{Q}^*_{[k]}.$$

Factor scores

The factor scores of the observations represent a compromise (i.e. a common representation) for the set of the *K* matrices. They are obtained by:

$$\mathbf{F} = \mathbf{P}^*\boldsymbol{\Delta}^* = \mathbf{X}\mathbf{A}^{-1}\mathbf{Q}.$$

Interestingly, this last equation can be rewritten as follows:

$$\mathbf{F} = \mathbf{X}\,\mathbf{A}^{-1}\mathbf{Q} = \left[\frac{\mathbf{X}_{[1]}}{\lambda_{[1]}}, \ldots, \frac{\mathbf{X}_{[k]}}{\lambda_{[k]}}, \ldots \frac{\mathbf{X}_{[K]}}{\lambda_{[K]}}\right]\begin{bmatrix}\mathbf{Q}_{[1]}\\ \vdots \\ \mathbf{Q}_{[k]} \\ \vdots \\ \mathbf{Q}_{[K]}\end{bmatrix} = \sum_{k=1}^{K} \frac{1}{\lambda_{[k]}}\mathbf{X}_{[k]}\mathbf{Q}_{[k]}.$$

We define the partial factor scores of the k^{th} table by $\mathbf{F}_{[k]} = K\ \frac{1}{\lambda_{[k]}}\ \mathbf{X}_{[k]}\mathbf{Q}_{[k]}$. The matrix of MFA factor scores **F** then corresponds to the average of all *K* partial factor scores:

$$\mathbf{F} = \frac{1}{K}\sum_{k=1}^{K}\mathbf{F}_{[k]}.$$

Deviation scores

As in a simple PCA, we define the *deviation score* of an observation as the distance of its factor scores to the center of the cloud of points; that is, if we denote $f_{i,l}$ the MFA factor score of the i^{th} individual for the l^{th} component:

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

Contribution of each table to the deviation scores

It may also be of interest to quantify the individual contributions of each data table to the overall deviation score; in particular, this can help identify which data tables have the greatest influence on an individual's deviation score. As above, we denote **F** the matrix of MFA factor scores and $\mathbf{F}_{[k]}$ the matrix of partial factor scores corresponding to the k^{th} table.
We first note that the overall deviation score d_i^2 can also be written as follows:

$$d_i^2 = \sum_{l=1}^{L} f_{i,l}^2 + \sum_{k=1}^{K}\sum_{l=1}^{L} f_{i,l}\,(f_{[k]i,l} - f_{i,l}).$$

Noting that

$$\sum_{k=1}^{K}\sum_{l=1}^{L} f_{i,l}\,(f_{[k]i,l} - f_{i,l}) = \sum_{k=1}^{K}\sum_{l=1}^{L} f_{i,l}^2 - (f_{[k]i,l}\,f_{i,l}) = 0,$$

the contribution of the k^{th} table to the deviation score for individual *i* is equal to

$$d_{i,k} = \frac{\sum_{l=1}^{L} f_{i,l}\,(f_{[k]i,l} - f_{i,l})}{\sum_{l=1}^{L} f_{i,l}^2}.$$

By construction, the contributions of the *K* tables to the overall deviation score sum to 0. In addition, per-table contributions can take on both negative and positive values according to the extent to which the table influences the deviation of the overall 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 tables which tend to be most similar to the global average.

Supplementary observations and variables

The results of the MFA can also be used to compute factor scores, loadings, and distances for new observations (rows) and variables (columns) that were not included in the original analysis. For simplicity, we focus on the use of the former in the following discussion, but analogous calculations can be performed for the latter. In particular, assuming that the supplementary rows are scaled in a manner comparable to the original rows of **X**, we can compute the factor scores $\mathbf{f}_{\text{sup}}$ for a supplementary row, which is represented by the *J*-dimensional vector $\mathbf{r}_{sup}^{T}$. The supplementary factor scores are then computed as $\mathbf{f}_{sup} = \mathbf{r}_{sup}^{T}\mathbf{A^{-1}Q}$, while the partial factor scores are obtained by :

$$\mathbf{f}_{\text{sup}[k]} = K\,\frac{\mathbf{1}}{\lambda_{[k]}}\mathbf{r}_{\text{sup}[k]}^{T}\mathbf{Q}.$$

Deviation scores

Similar to the case of observations contained in the original data table, the *deviation score* of a supplementary observation is the distance of its factor scores to the center of the cloud of points. That is, if $fsup_{i,l}$ represents the supplementary factor score of the i^{th} individual for the l^{th} component:

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

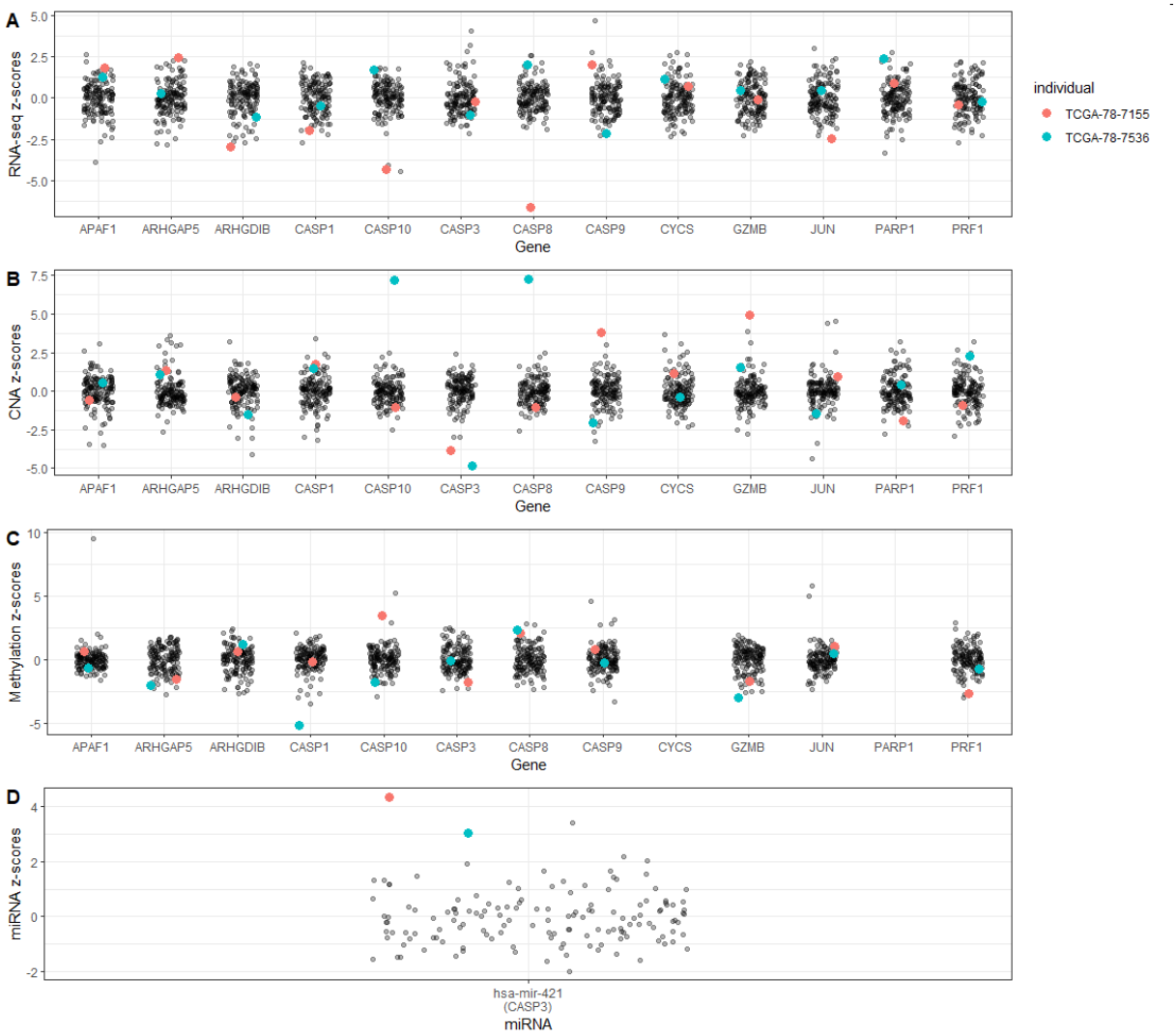
In addition, the contribution of each data table to this overall deviation score for a supplementary individual can be calculated in a manner analogous to that described above.

References

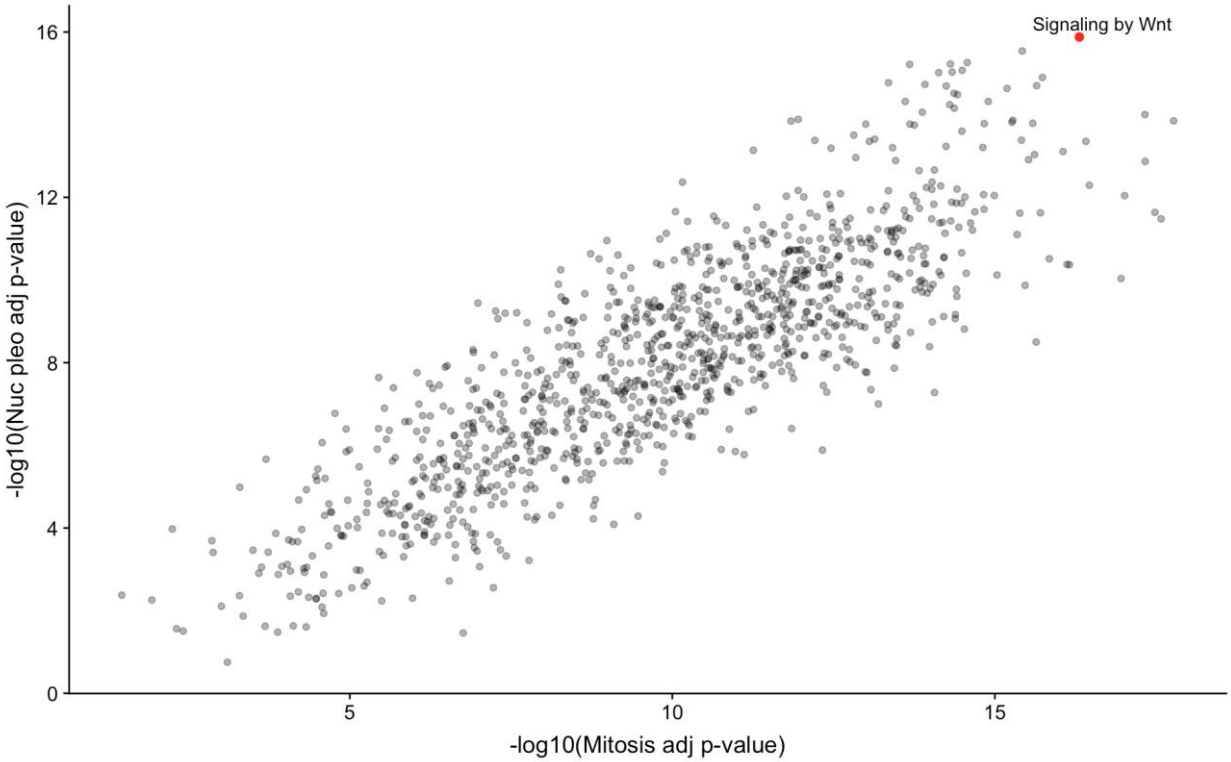
Abdi, H. and Williams, L.J. (2010) Principal component analysis. *WIREs Computational Statistics*, doi:10.1002/wics.101.

Abdi, H., Williams, L.J., and Valentin, D. (2013) Multiple factor analysis: principal component analysis for multitable and multiblock data sets. *WIREs Computational Statistics*, doi:10.1002/wics.1246.

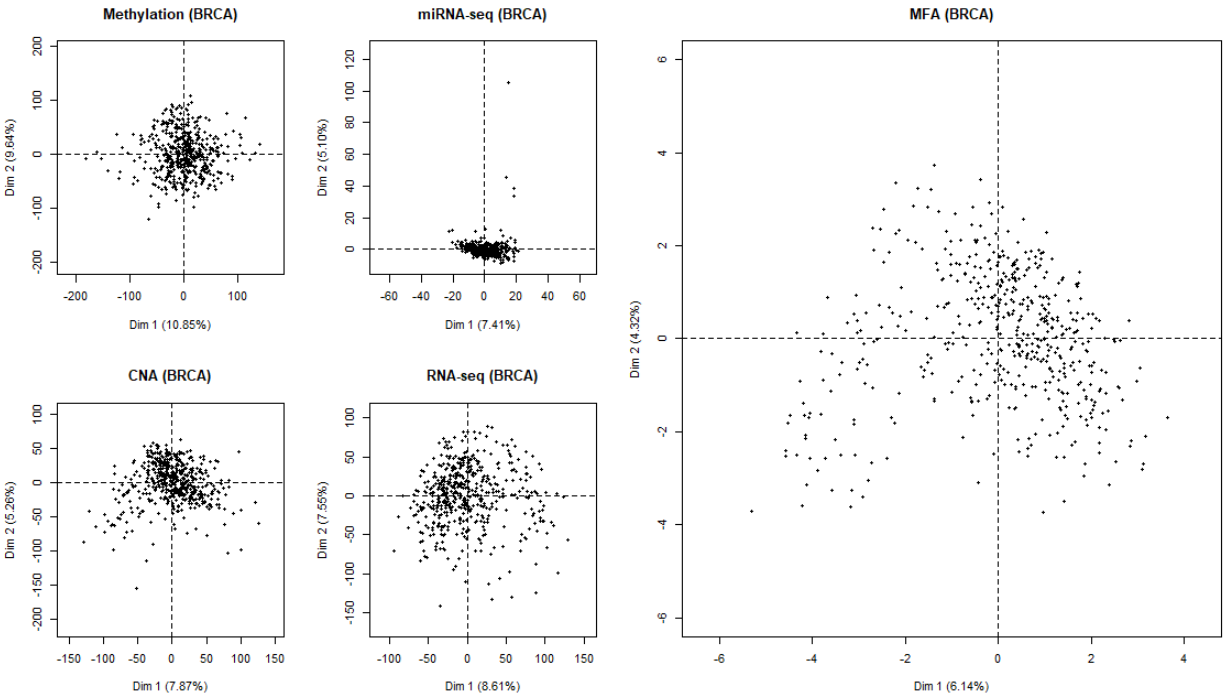
Supplementary Figures



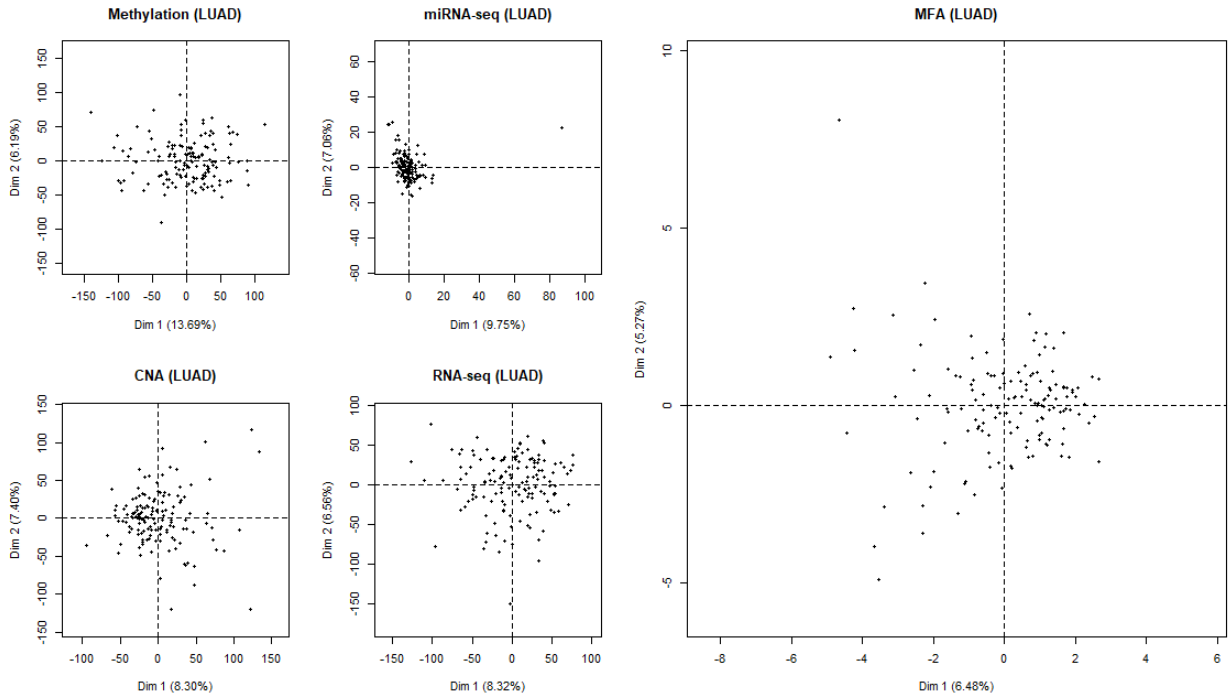
Supplementary Figure 1. Z-scores of RNA-seq, CNA, methylation, and miRNA-seq data for genes in the D4-GDI signaling pathway for individuals in the TCGA LUAD data (n = 144). Data corresponding to the two individuals with the largest overall pathway deviation scores, TCGA-78-7155 and TCGA-78-7536, are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



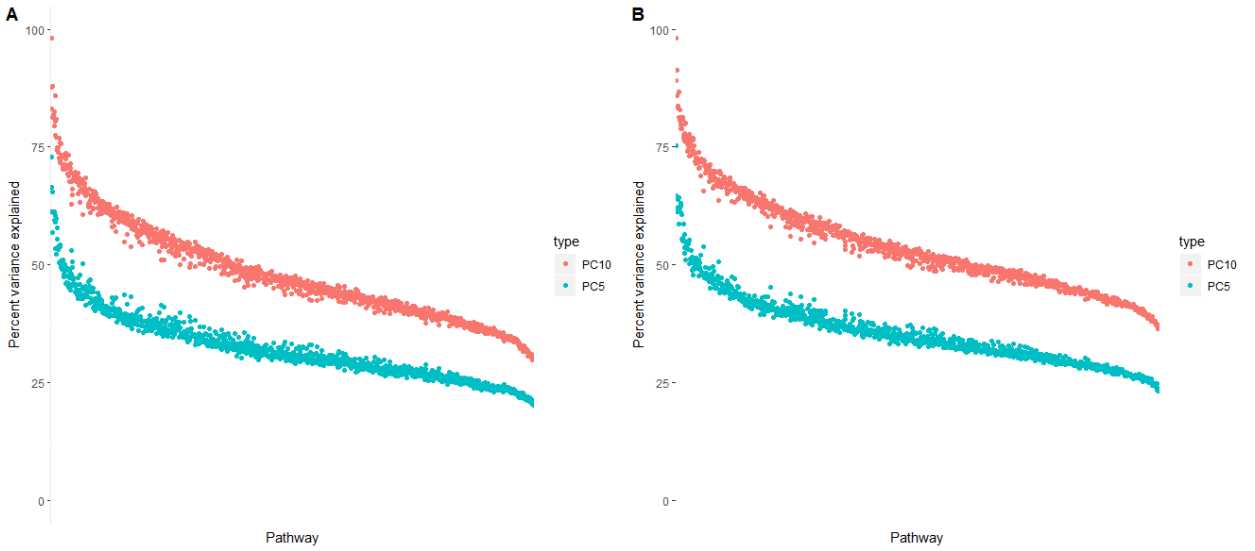
Supplementary Figure 2. Negative log10-transformed p-values from the ANOVA F-test of pathway deviation score versus mitosis and nuclear pleomorphism for each pathway among breast cancer individuals. The signaling by Wnt pathway is highlighted in red.



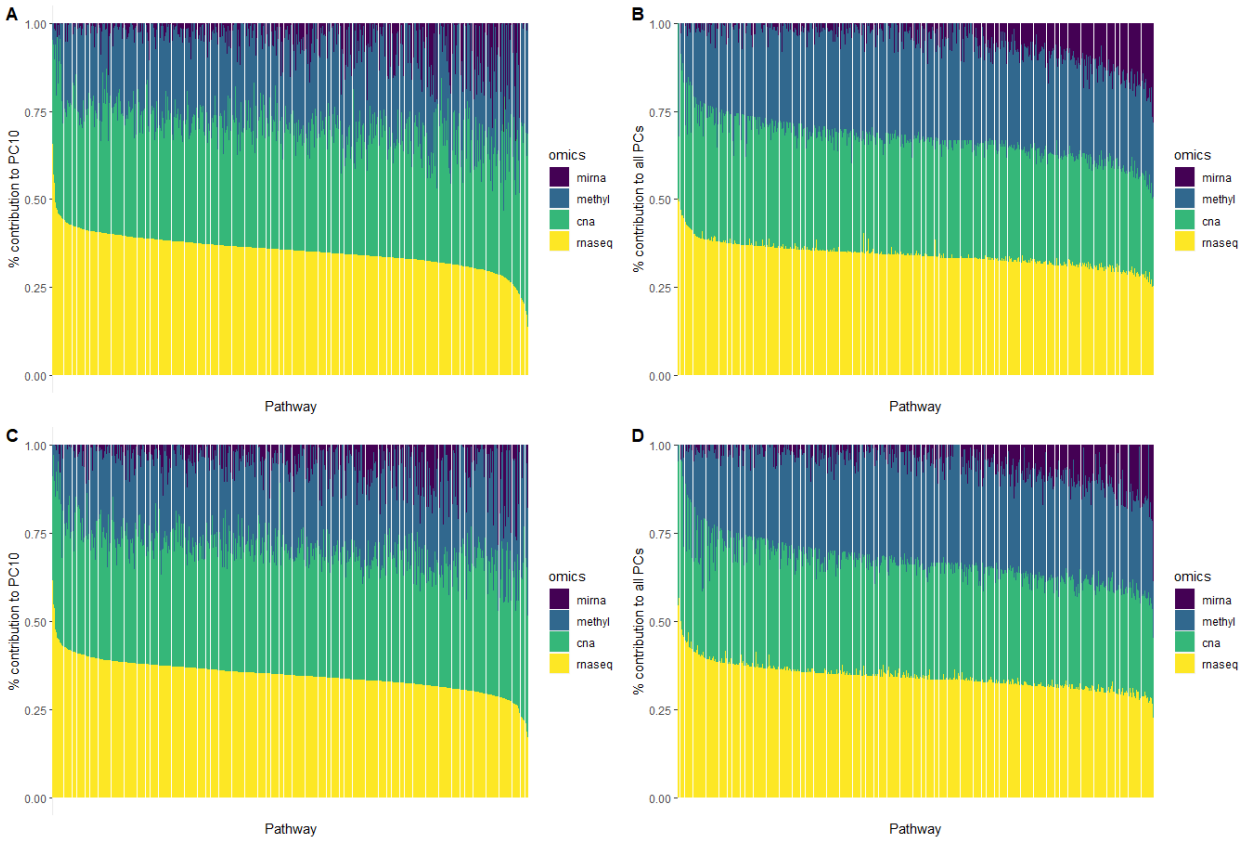
Supplementary Figure 3. Factor maps for the first two dimensions of a global transcriptome- and genome-wide PCA of the methylation, miRNA-seq, CNA, and RNA-seq data (left), as well as a global MFA of all four omics combined (right) for the TCGA BRCA data.



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 Tables

Histological measure	Stage	Number of individuals (*n* = 504)
Nuclear pleomorphism	I: Small regular nuclei II: Moderate increase in size III: Moderate to marked variation in size NA	47 207 162 87
Mitotic index	I: 0-5 per 10 high powered fields (HPF; low) II: 6-10 per 10 HPF (medium) III: >10 per 10 HPF (high) NA	225 89 101 89
Glandular/tubule formation	I: >75% (well differentiated) II: 10-75% (moderately differentiated) III: < 10% (poorly differentiated) NA	51 81 285 87

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

Pathway name	Genes
D4-GDI (GDP dissociation inhibitor) signaling pathway	APAF1, ARHGAP5, ARHGDIB, CASP1, CASP10, CASP3, CASP8, CASP9, CYCS, GZMB, JUN, PARP1, PRF1
NF-kB activation through FADD/RIP-1 pathway mediated by caspase-8 and -10	CASP10, CASP8, CHUK, DDX58, FADD, IFIH1, IKBKB, IKBKG, MAVS, RIPK1, RNF135, TRIM25
Class I PI3K signaling events mediated by Akt	AKT1, AKT2, AKT3, BAD, BCL2L1, CASP9, CDKN1A, CDKN1B, CHUK, FOXO1, FOXO3, FOXO4, GSK3A, GSK3B, HSP90AA1, KPNA1, MAP3K5, MAPKAP1, MLST8, MTOR, PDPK1, PRKACA, PRKDC, RAF1, RICTOR, SFN, SLC2A4, SRC, TBC1D4, YWHAB, YWHAE, YWHAG, YWHAH, YWHAQ, YWHAZ
ATM signaling pathway	ABL1, ATM, BRCA1, CDKN1A, CHEK1, CHEK2, GADD45A, JUN, MAPK8, MDM2, MRE11A, NBN, NFKB1, NFKBIA, RAD50, RAD51, RBBP8, RELA, TP53, TP73
CARM1 and regulation of the estrogen receptor	BRCA1, CARM1, CCND1, CREBBP, EP300, ERCC3, ESR1, GRIP1, GTF2A1, GTF2E1, GTF2F1, HDAC1, HDAC10, HDAC11, HDAC2, HDAC3, HDAC4, HDAC5, HDAC6, HDAC7, HDAC8, HDAC9, HIST2H3C, MED1, MEF2C, NCOR2, NR0B1, NRIP1, PELP1, PHB2, POLR2A, PPARGC1A, SPEN, SRA1, TBP
Homologous recombination repair of replication- independent double-strand breaks	ATM, BRCA1, BRCA2, BRIP1, H2AFX, LIG1, MDC1, MRE11A, NBN, RAD50, RAD51, RAD52, RPA1, RPA2, RPA3, TP53BP1
Role of BRCA1, BRCA2, and ATR in cancer susceptibility	ATM, ATR, BRCA1, BRCA2, CHEK1, CHEK2, FANCC, FANCD2, FANCE, FANCF, FANCG, HUS1, MRE11A, NBN, RAD1, RAD17, RAD50, RAD51, RAD9A, TP53, TREX1
CD40L signaling pathway	CD40, CD40LG, CHUK, DUSP1, IKBKAP, IKBKB, IKBKG, MAP3K1, MAP3K14, NFKB1, NFKBIA, RELA, TNFAIP3, TRAF3, TRAF6
Induction of apoptosis through DR4 and DR4/5 death receptors	APAF1, BCL2, BID, BIRC2, BIRC3, CASP10, CASP3, CASP6, CASP7, CASP8, CASP9, CFLAR, CHUK, CYCS, DFFA, DFFB, FADD, GAS2, LMNA, MAP3K14, NFKB1, NFKBIA, RELA, RIPK1, SPTAN1, TNFRSF10A, TNFRSF10B, TNFRSF25, TNFSF10, TNFSF12, TRADD, TRAF2, XIAP
Cell cycle: G1/S check point	ABL1, ATM, ATR, CCNA1, CCND1, CCNE1, CDC25A, CDK1, CDK2, CDK4, CDK6, CDKN1A, CDKN1B, CDKN2A, CDKN2B, DHFR, E2F1, GSK3B, HDAC1, RB1, SKP2, SMAD3, SMAD4, TFDP1, TGFB1, TGFB2, TGFB3, TP53
Double stranded RNA induced gene expression	CHUK, DNAJC3, EIF2AK2, EIF2S1, EIF2S2, MAP3K14, NFKB1, NFKBIA, RELA, TP53
Signaling events mediated by HDAC class III	ACSS1, ACSS2, BAX, CDKN1A, CREBBP, EP300, FHL2, FOXO1, FOXO3, FOXO4, HDAC4, HIST1H1E, HOXA10, KAT2B, MEF2D, MYOD1, PPARGC1A, SIRT1, SIRT2, SIRT3, SIRT7, TP53, TUBA1B, TUBB2A, XRCC6
HIV-1 Nef: Negative effector of Fas and TNF-alpha	APAF1, BAG4, BCL2, BID, BIRC3, CASP2, CASP3, CASP6, CASP7, CASP8, CASP9, CD247, CFLAR, CHUK, CRADD, CYCS, DAXX, DFFA, DFFB, FADD, FAS, FASLG, MAP2K7, MAP3K14, MAP3K5, MAPK8, NFKB1, NFKBIA, RELA, RIPK1, TNF, TNFRSF1A, TRADD, TRAF1, TRAF2
Regulation of telomerase	ABL1, ACD, AKT1, ATM, BLM, CCND1, CDKN1B, DKC1, E2F1, EGF, EGFR, ESR1, FOS, HDAC1, HDAC2, HNRNPC, HSP90AA1, HUS1, IFNAR2, IFNG, IL2, IRF1, JUN, MAPK1, MAPK3, MAX, MRE11A, MTOR, MXD1, MYC, NBN, NCL, NFKB1, NR2F2, PARP2, PIF1, PINX1, POT1, PTGES3, RAD1, RAD50, RAD9A, RBBP4, RBBP7, RPS6KB1, SAP18, SAP30, SIN3A, SIN3B, SMAD3, SMG5, SMG6, SP1, SP3, TERF1, TERF2, TERF2IP, TERT, TGFB1, TINF2, TNKS, UBE3A, WRN, WT1, XRCC5, XRCC6, YWHAE, ZNFX1

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.

Pathway name	Genes
Signaling by Wnt	APC, AXIN1, BTRC, CSNK1A1, CTNNB1, CUL1, FAM123B, FRAT1, FRAT2, PPP2CA, PPP2CB, PPP2R1A, PPP2R1B, PPP2R5A, PPP2R5B, PPP2R5C, PPP2R5D, PPP2R5E, PSMA1, PSMA2, PSMA3, PSMA4, PSMA5, PSMA6, PSMA7, PSMA8, PSMB1, PSMB10, PSMB2, PSMB3, PSMB4, PSMB5, PSMB6, PSMB7, PSMB8, PSMB9, PSMC1, PSMC2, PSMC3, PSMC4, PSMC5, PSMC6, PSMD1, PSMD10, PSMD11, PSMD12, PSMD13, PSMD14, PSMD2, PSMD3, PSMD4, PSMD5, PSMD6, PSMD7, PSMD8, PSMD9, PSME1, PSME2, PSME4, PSMF1, RPS27A, SKP1, UBA52
Apoptotic execution phase	ACIN1, ADD1, APC, BCAP31, BIRC2, BMX, CASP3, CASP6, CASP7, CASP8, CDH1, CTNNB1, DBNL, DFFA, DFFB, DNM1L, DSG1, DSG2, DSG3, DSP, FNTA, GAS2, GSN, H1F0, HIST1H1A, HIST1H1B, HIST1H1C, HIST1H1D, HIST1H1E, HMGB1, HMGB2, KPNA1, KPNB1, LMNA, LMNB1, LOC647859, MAPT, MST4, OCLN, PAK2, PKP1, PLEC, PRKCD, PRKCQ, PTK2, ROCK1, SATB1, SPTAN1, STK24, TJP1, TJP2, VIM
APC/C:Cdh1 mediated degradation of Cdc20 and other APC/C:Cdh1 targeted proteins in late mitosis/early G1	ANAPC1, ANAPC10, ANAPC11, ANAPC2, ANAPC4, ANAPC5, ANAPC7, AURKA, AURKB, CDC16, CDC20, CDC23, CDC26, CDC27, PLK1, PSMA1, PSMA2, PSMA3, PSMA4, PSMA5, PSMA6, PSMA7, PSMA8, PSMB1, PSMB10, PSMB2, PSMB3, PSMB4, PSMB5, PSMB6, PSMB7, PSMB8, PSMB9, PSMC1, PSMC2, PSMC3, PSMC4, PSMC5, PSMC6, PSMD1, PSMD10, PSMD11, PSMD12, PSMD13, PSMD14, PSMD2, PSMD3, PSMD4, PSMD5, PSMD6, PSMD7, PSMD8, PSMD9, PSME1, PSME2, PSME4, PSMF1, PTTG1, RPS27A, SKP2, UBA52, UBE2C, UBE2D1, UBE2E1
Genes involved in Beta-catenin phosphorylation cascade	APC, AXIN1, CSNK1A1, CTNNB1, FAM123B, FRAT1, FRAT2, PPP2CA, PPP2CB, PPP2R1A, PPP2R1B, PPP2R5A, PPP2R5B, PPP2R5C, PPP2R5D, PPP2R5E
Autodegradation of Cdh1 by Cdh1:APC/C	ANAPC1, ANAPC10, ANAPC11, ANAPC2, ANAPC4, ANAPC5, ANAPC7, CDC16, CDC23, CDC26, CDC27, PSMA1, PSMA2, PSMA3, PSMA4, PSMA5, PSMA6, PSMA7, PSMB1, PSMB10, PSMB2, PSMB3, PSMB4, PSMB5, PSMB6, PSMB7, PSMB8, PSMB9, PSMC1, PSMC2, PSMC3, PSMC4, PSMC5, PSMC6, PSMD1, PSMD10, PSMD11, PSMD12, PSMD13, PSMD14, PSMD2, PSMD3, PSMD4, PSMD5, PSMD6, PSMD7, PSMD8, PSMD9, PSME1, PSME2, PSMF1, RPS27A, UBA52, UBE2C, UBE2D1, UBE2E1
Genes involved in M/G1 transition	CDC45, CDC6, CDC7, CDK2, CDT1, DBF4, E2F1, E2F2, E2F3, GMNN, MCM10, MCM2, MCM3, MCM4, MCM5, MCM6, MCM7, MCM8, POLA1, POLA2, POLE, POLE2, PRIM1, PRIM2, PSMA1, PSMA2, PSMA3, PSMA4, PSMA5, PSMA6, PSMA7, PSMA8, PSMB1, PSMB10, PSMB2, PSMB3, PSMB4, PSMB5, PSMB6, PSMB7, PSMB8, PSMB9, PSMC1, PSMC2, PSMC3, PSMC4, PSMC5, PSMC6, PSMD1, PSMD10, PSMD11, PSMD12, PSMD13, PSMD14, PSMD2, PSMD3, PSMD4, PSMD5, PSMD6, PSMD7, PSMD8, PSMD9, PSME1, PSME2, PSME4, PSMF1, RPA1, RPA2, RPA3, RPA4, RPS27A, UBA52
Regulation of the Fanconi anemia pathway	ATM, ATR, FANCD2, RPS27A, UBA52, USP1, ZBTB32
Apoptotic cleavage of cellular proteins	ACIN1, ADD1, APC, BCAP31, BIRC2, BMX, CASP3, CASP6, CASP7, CASP8, CDH1, CTNNB1, DBNL, DSG1, DSG2, DSG3, DSP, FNTA, GAS2, GSN, LMNA, LMNB1, LOC647859, MAPT, MST4, OCLN, PKP1, PLEC, PRKCD, PRKCQ, PTK2, ROCK1, SATB1, SPTAN1, STK24, TJP1, TJP2, VIM
Apoptosis	ACIN1, ADD1, AKT1, APAF1, APC, APPL1, ARHGAP10, BAD, BAK1, BAX, BBC3, BCAP31, BCL2, BCL2L1, BCL2L11, BID, BIRC2, BMF, BMX, CASP10, CASP3, CASP6, CASP7, CASP8, CASP9, CDH1, CFLAR, CTNNB1, CYCS, DAPK1, DAPK2, DAPK3, DBNL, DCC, DFFA, DFFB, DIABLO, DNM1L, DSG1, DSG2, DSG3, DSP, DYNLL1, DYNLL2, E2F1, FADD, FAS, FASLG, FNTA, GAS2, GSN, GZMB, H1F0, HIST1H1A, HIST1H1B, HIST1H1C, HIST1H1D, HIST1H1E, HMGB1, HMGB2, KPNA1, KPNB1, LMNA, LMNB1, LOC647859, MAGED1, MAPK8, MAPT, MST4, NMT1, OCLN, PAK2, PKP1, PLEC, PMAIP1, PPP3R1, PRKCD, PRKCQ, PSMA1, PSMA2, PSMA3, PSMA4, PSMA5, PSMA6, PSMA7, PSMA8, PSMB1, PSMB10, PSMB2, PSMB3, PSMB4, PSMB5, PSMB6, PSMB7, PSMB8, PSMB9, PSMC1, PSMC2, PSMC3, PSMC4, PSMC5, PSMC6, PSMD1, PSMD10, PSMD11, PSMD12, PSMD13, PSMD14, PSMD2, PSMD3, PSMD4, PSMD5, PSMD6, PSMD7, PSMD8, PSMD9, PSME1, PSME2, PSME4, PSMF1, PTK2, RIPK1, ROCK1, RPS27A, SATB1, SPTAN1, STK24, TFDP1, TJP1, TJP2, TNF, TNFRSF10B, TNFRSF1A, TNFSF10, TP53, TRADD, TRAF2, UBA52, UNC5A, UNC5B, VIM, XIAP, YWHAB
ER-phagosome pathway	B2M, CALR, HLA-A, HLA-B, HLA-C, HLA-F, HLA-G, PDIA3, PSMA1, PSMA2, PSMA3, PSMA4, PSMA5, PSMA6, PSMA7, PSMA8, PSMB1, PSMB10, PSMB2, PSMB3, PSMB4, PSMB5, PSMB6, PSMB7, PSMB8, PSMB9, PSMC1, PSMC2, PSMC3, PSMC4, PSMC5, PSMC6, PSMD1, PSMD10, PSMD11, PSMD12, PSMD13, PSMD14, PSMD2, PSMD3, PSMD4, PSMD5, PSMD6, PSMD7, PSMD8, PSMD9, PSME1, PSME2, PSME4, PSMF1, RPS27A, SEC61A1, SEC61A2, SEC61B, SEC61G, TAP1, TAP2, UBA52

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.