MOVIE - Supplementary Materials

Sean D. McCabe[1], Dan-Yu Lin[1], and Michael I. Love [1*]

May 10, 2019

1 Supplementary Methods

1.1 Data Pre-processing

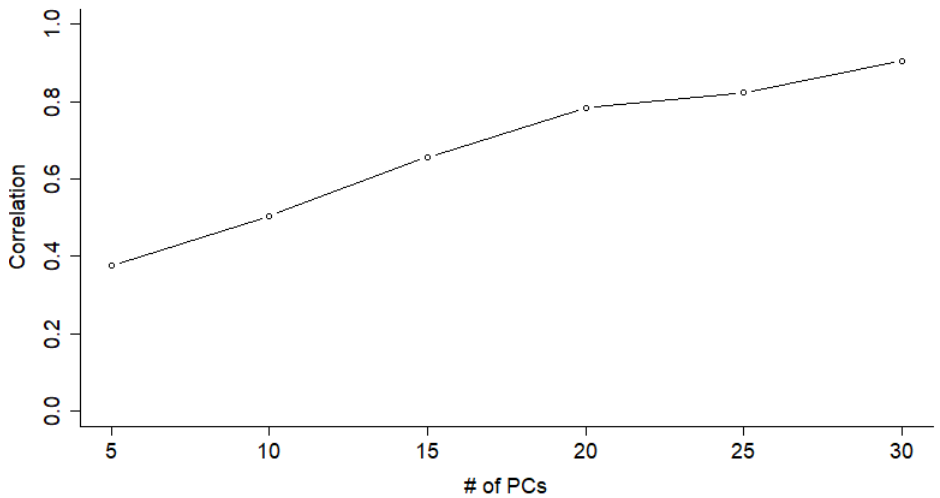
Data from The Cancer Genome Atlas (TCGA) [3] was collected for 558 breast cancer samples using Copy Number Variation (CNV), RNA expression, and micro RNA expression. CNV was summarized for 216 genes; RNA expression was measured for 12,434 genes; and miRNA expression was measured for 305 genes. Examination of the scree plot for each data type resulted in selecting AJIVE initial signal values of 25, 25, and 50 components for CNV, RNA, and miRNA, respectively.

RNA expression, DNase, and protein expression were collected for Yoruban lymphoblastoid cell lines from Li, et al (2016). To avoid dealing with issues of missing data, only samples that had all three data types observed were used in the analysis (n=55). Two samples were identified as outliers in the DNase data by examing the first two principal components; these were subsequently removed from the analysis. DNase was measured for 699,906 genes; RNA expression was measured for 13,967 genes; and protein was measured expression for 4,375 genes. The protein expression data had a high number of missing values, and thus, only genes that were observed for all samples were included in the analysis (2,435). To accommodate the large differences in dimensionality of the data types, the DNase dataset was reduced to the top 5000 most variable genes, as recommended by MOFA [20]. For consistency across methods, the reduced DNase dataset was used for all analyses. Examination of the scree plot for each data type resulted in the selection of AJIVE initial signal values of 3, 3, and 2 components for DNase, RNA, and protein expression, respectively.

2 Supplementary Figures

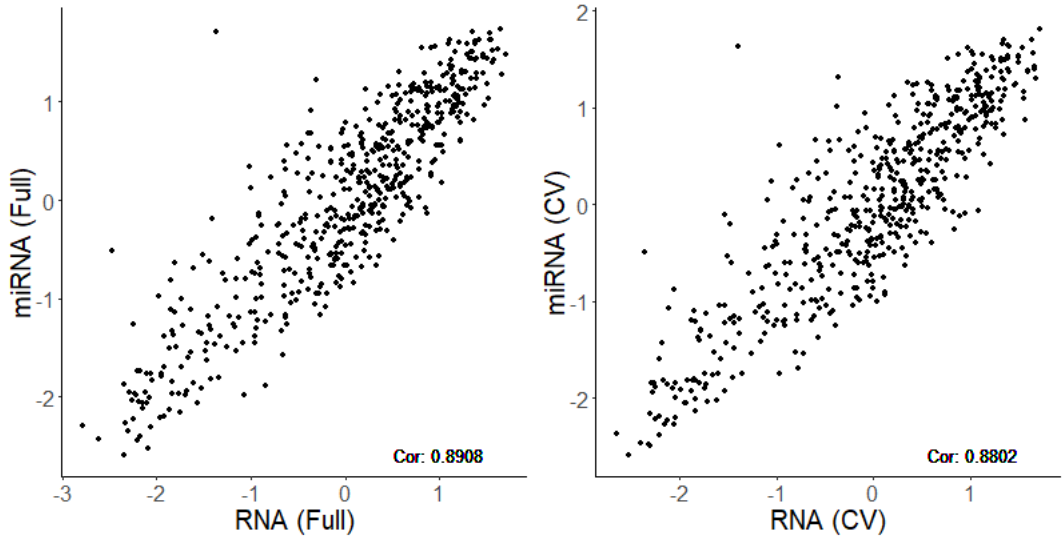
2.1 Motivation

Supplementary Figure 1: CCA on PCs of null Gaussian data returns correlations as high as 0.9, for datasets of sizes that are typical genomics (e.g. 100 samples, 5000 genes).

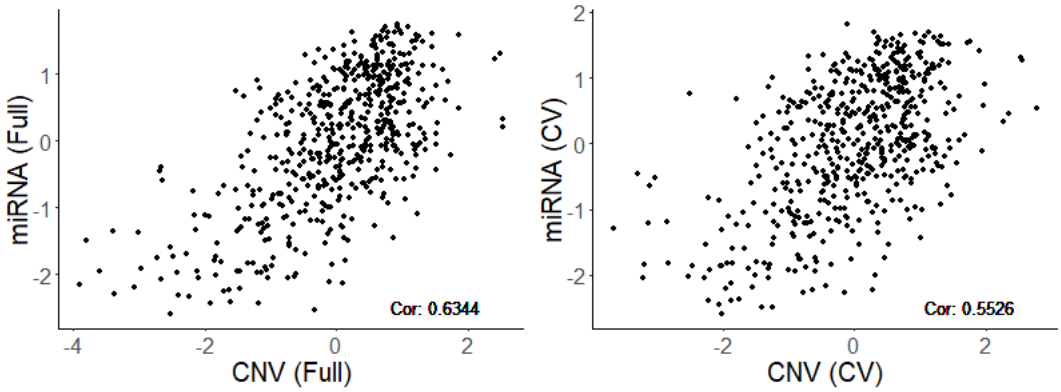


2.2 Large Sample Size - TCGA Data

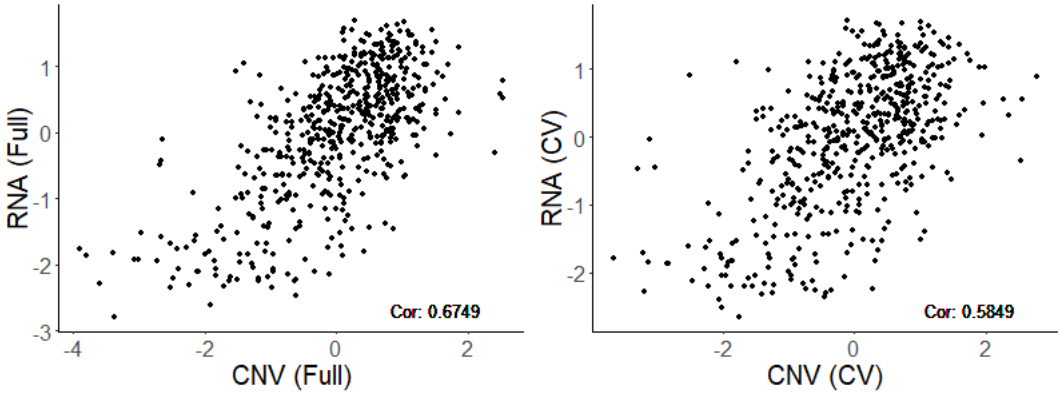
Supplementary Figure 2: Sparse mCCA with experimental data: Side-by-side contribution plots for gene expression versus miRNA using Sparse mCCA.



Supplementary Figure 3: Sparse mCCA with experimental data: Side-by-side contribution plots for CNV versus miRNA using Sparse mCCA.

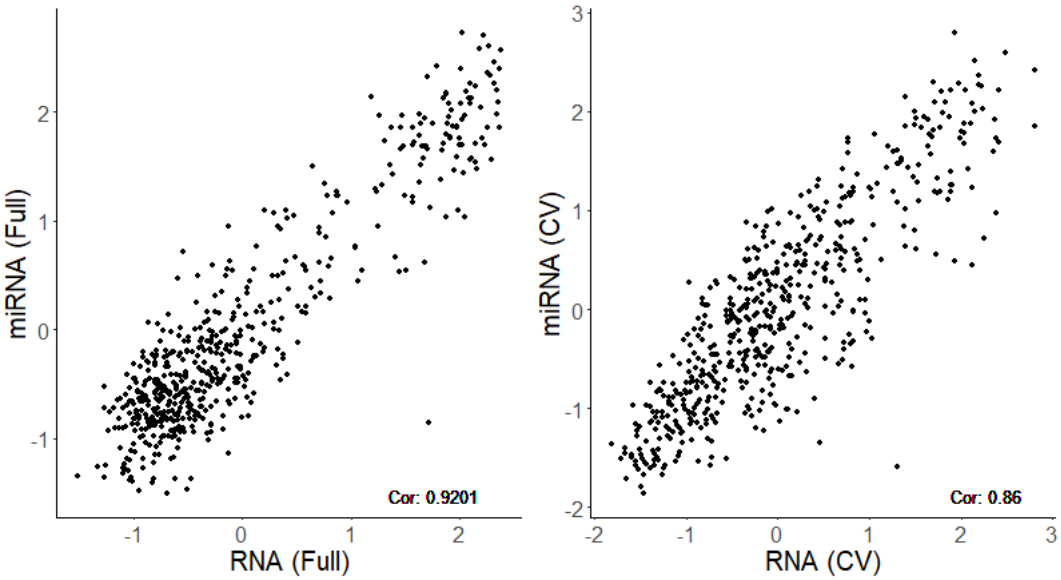


Supplementary Figure 4: Sparse mCCA with experimental data: Side-by-side contribution plots for CNV versus gene expression using Sparse mCCA.

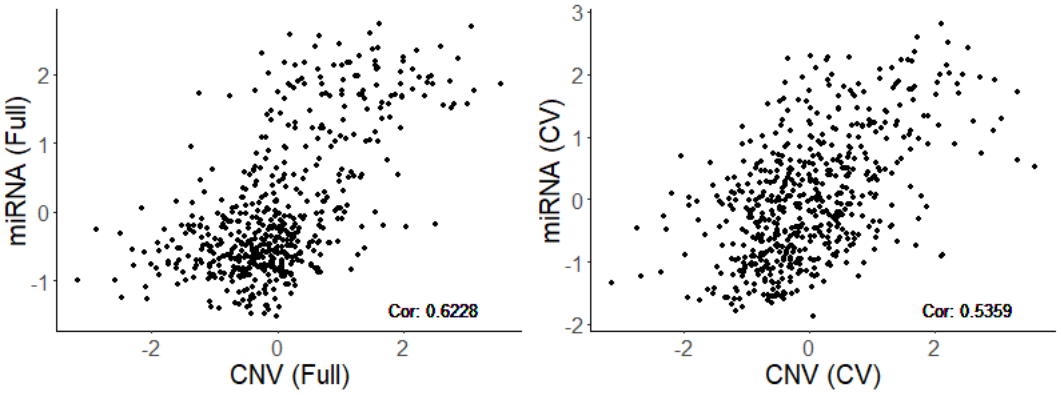


re

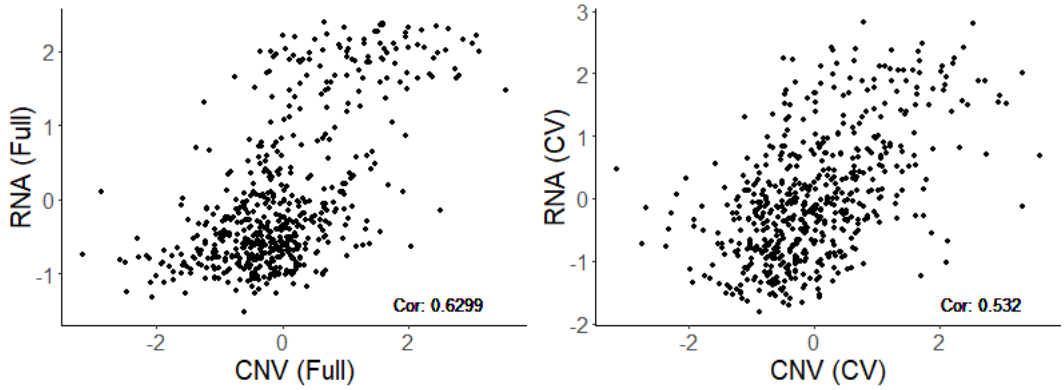
Supplementary Figure 5: MOFA with experimental data: Side-by-side contribution plots for gene expression versus miRNA using MOFA.



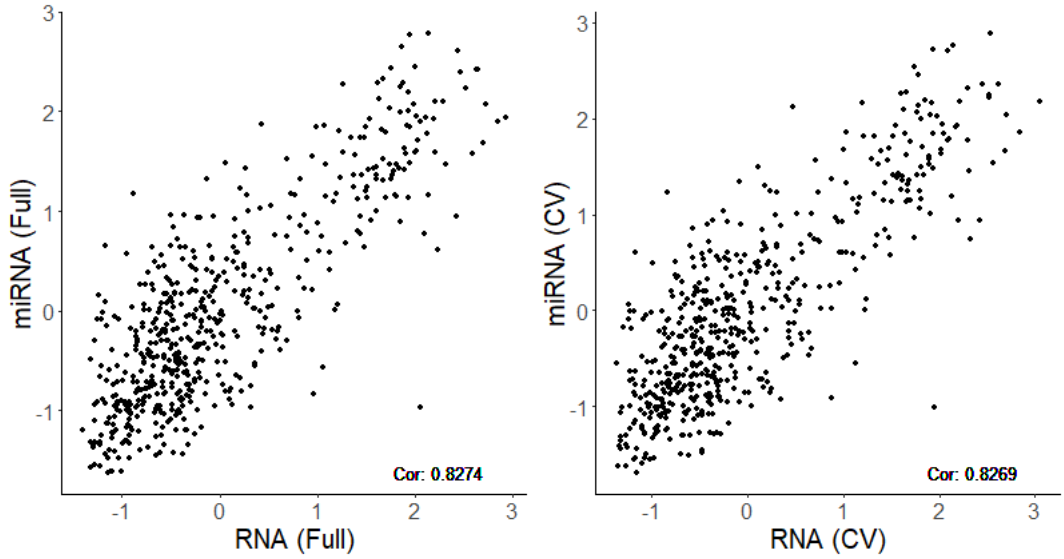
Supplementary Figure 6: MOFA with experimental data: Side-by-side contribution plots for CNV versus miRNA in experimental data using MOFA.



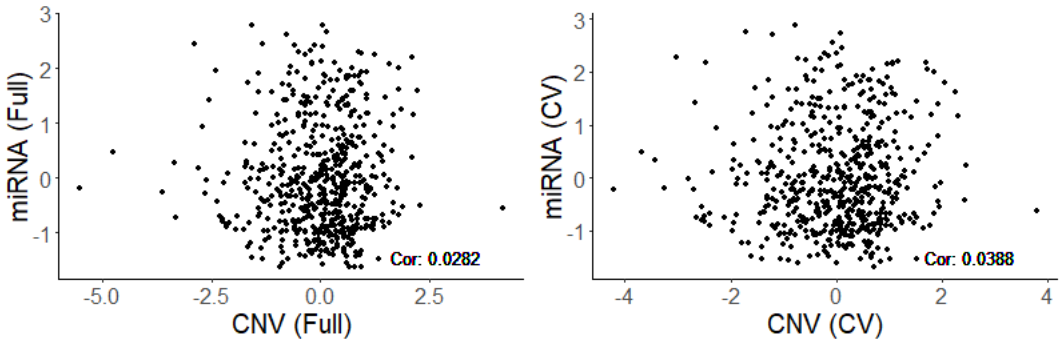
Supplementary Figure 7: MOFA with experimental data: Side-by-side contribution plots for CNV versus gene expression using MOFA.



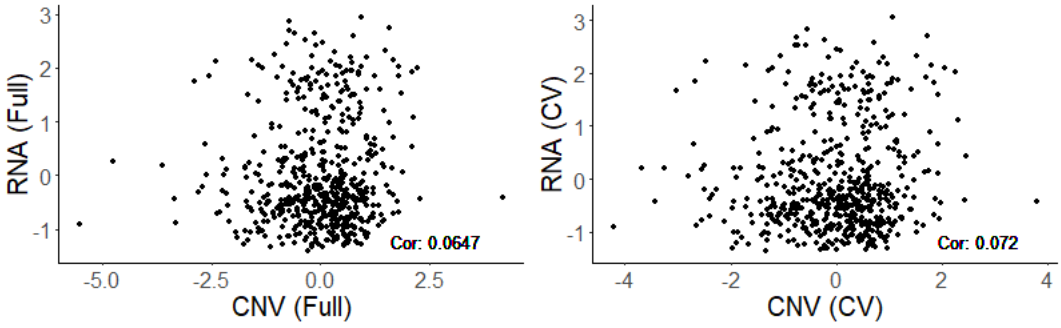
Supplementary Figure 8: AJIVE with experimental data: Side-by-side contribution plots for gene expression versus miRNA using AJIVE.



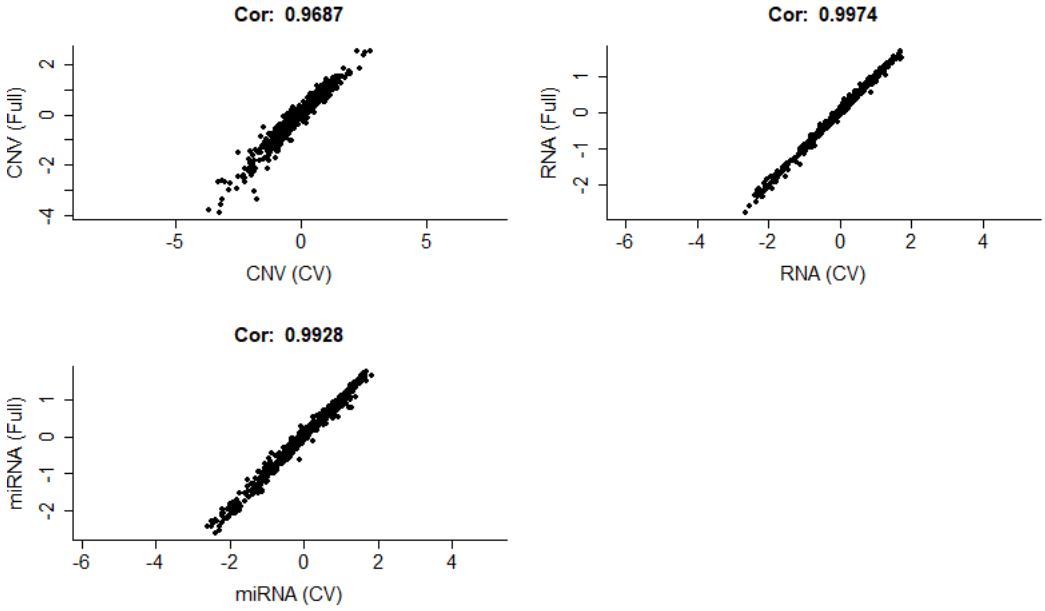
Supplementary Figure 9: AJIVE with experimental data: Side-by-side contribution plots for CNV versus micro RNA using AJIVE.



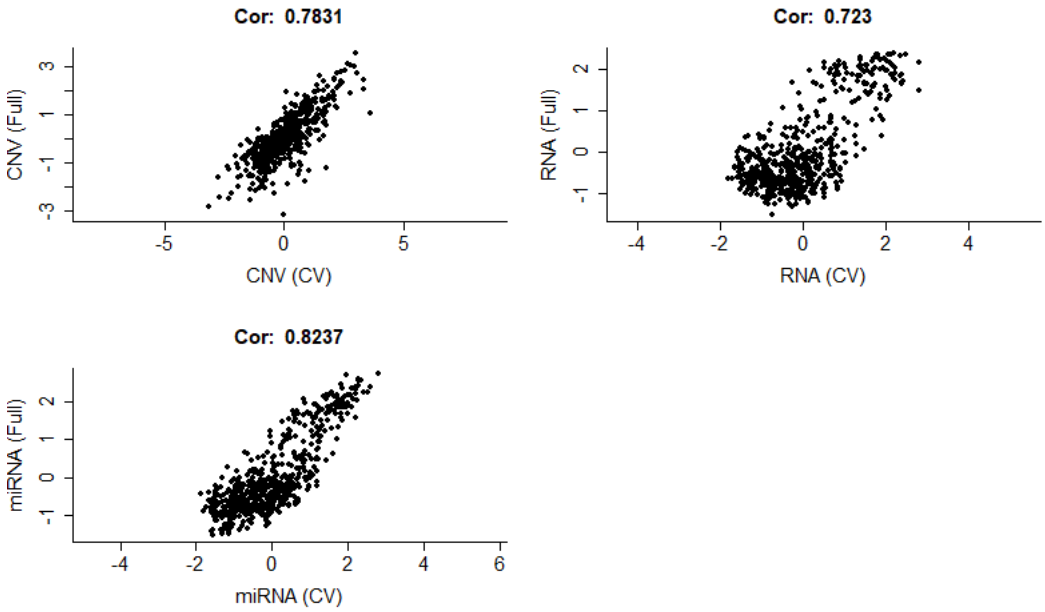
Supplementary Figure 10: AJIVE with experimental data: Side-by-side contribution plots for CNV versus gene expression using AJIVE.



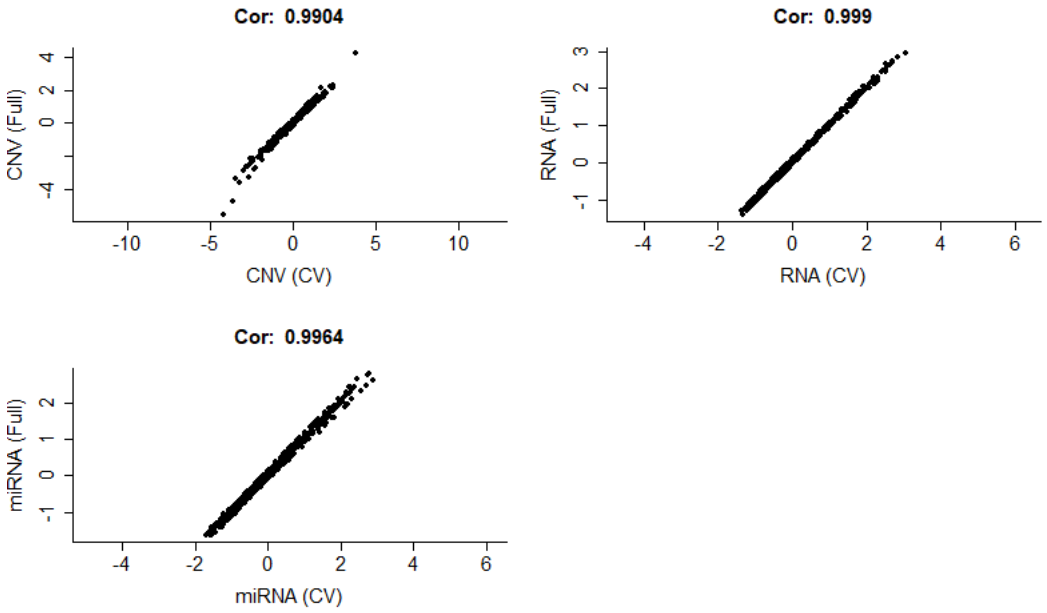
Supplementary Figure 11: Sparse mCCA with experimental data: Comparison plots for CNV, gene expression, and miRNA using Sparse mCCA.



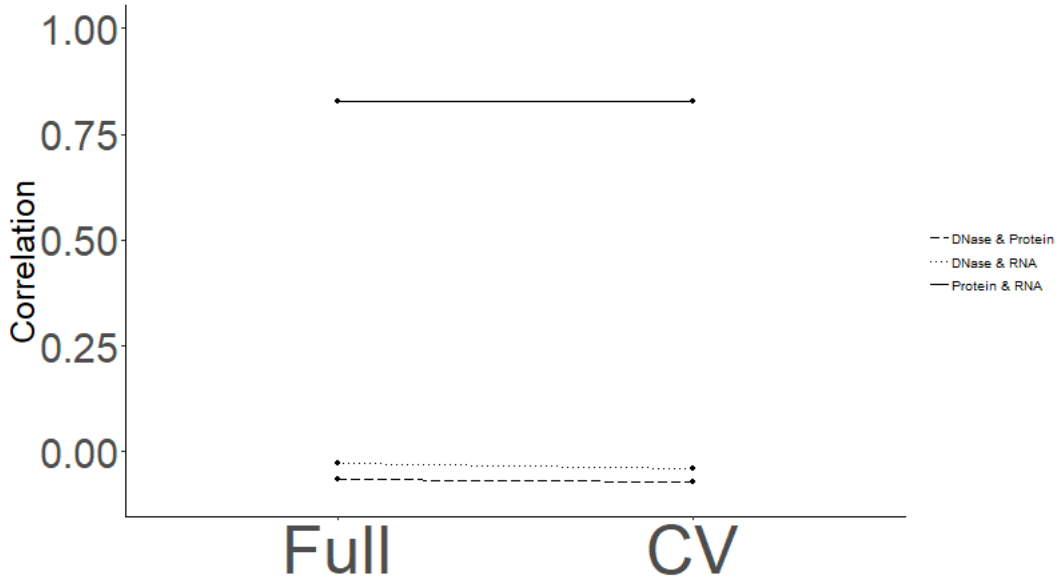
Supplementary Figure 12: MOFA with experimental data: Comparison plots for CNV, gene expression, and miRNA using MOFA.



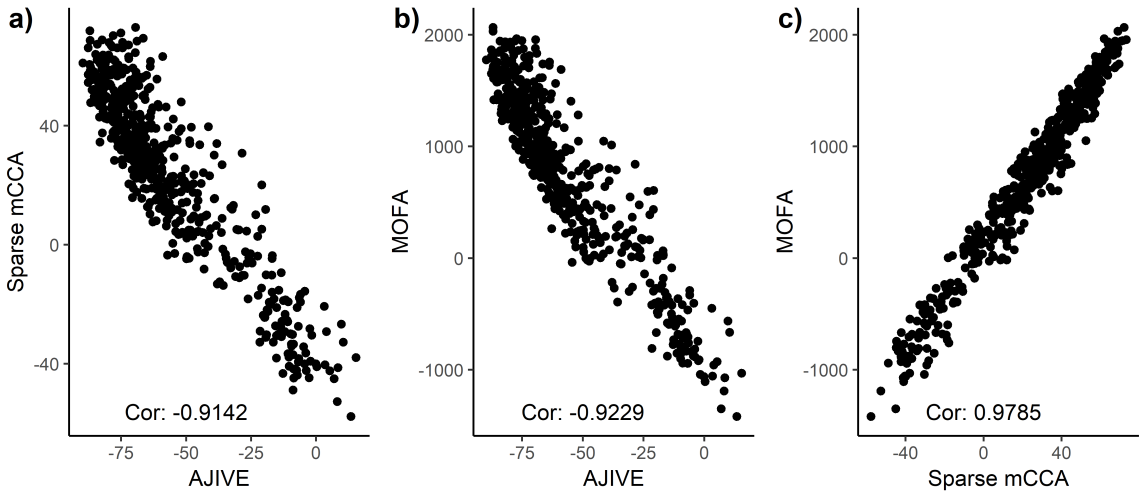
Supplementary Figure 13: AJIVE with experimental data: Comparison plots for CNV, gene expression, and miRNA using AJIVE.



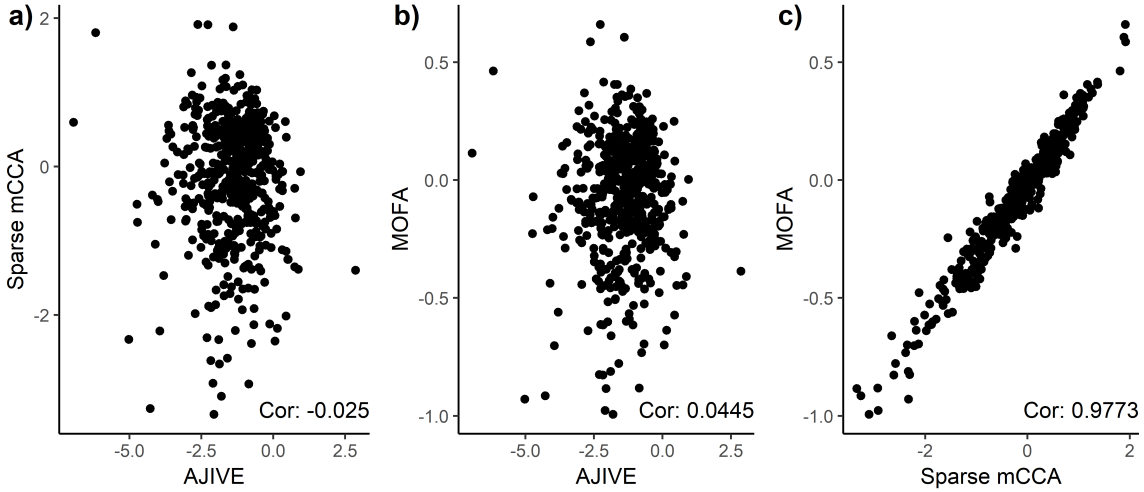
Supplementary Figure 14: Contribution correlations for AJIVE for each pair of data sets in an analysis with increased ranks for CNV.



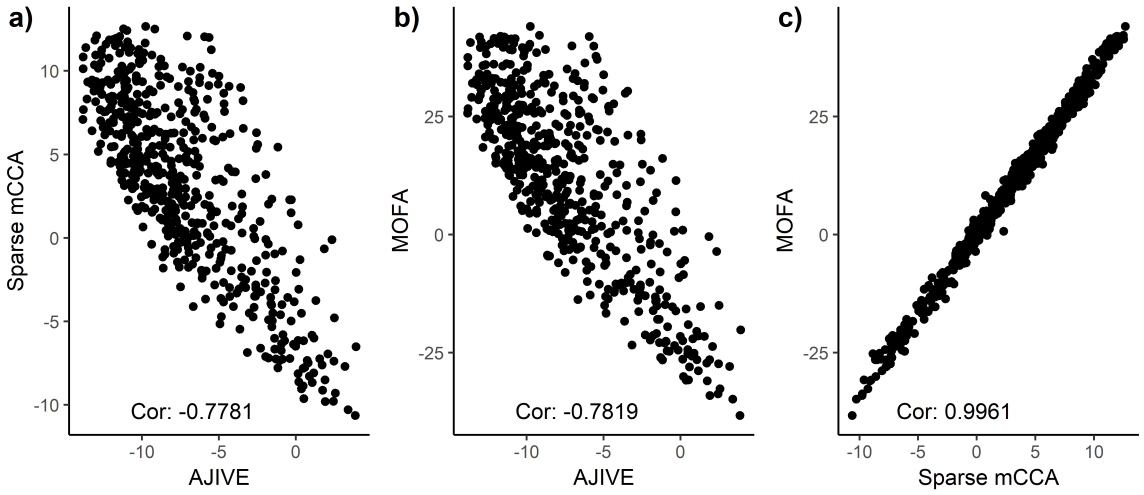
Supplementary Figure 15: Pairwise plots of RNA contributions for **a)** Sparse mCCA vs. AJIVE, **b)** MOFA vs. AJIVE, and **c)** MOFA vs. Sparse mCCA



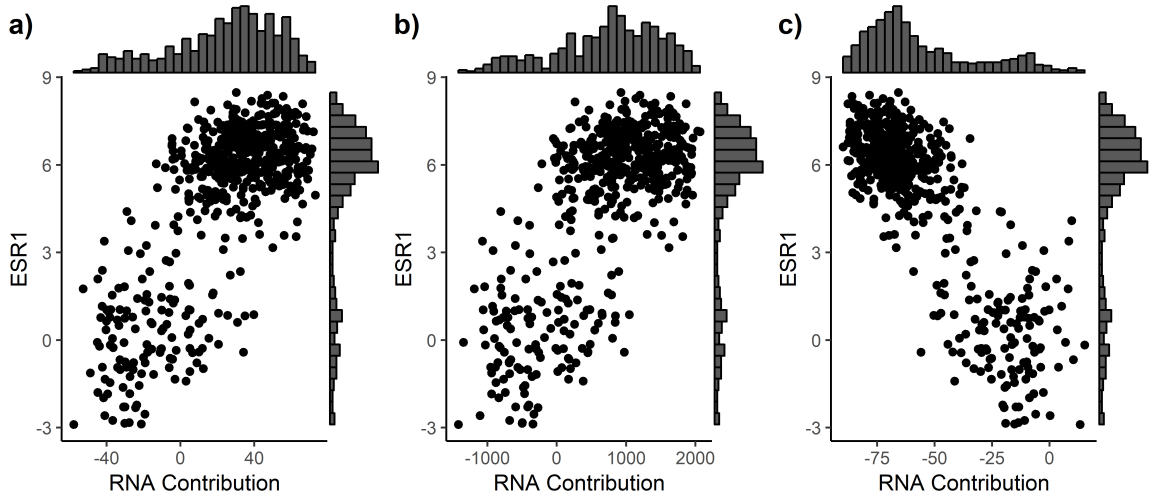
Supplementary Figure 16: Pairwise plots of CNV contributions for **a)** Sparse mCCA vs. AJIVE, **b)** MOFA vs. AJIVE, and **c)** MOFA vs. Sparse mCCA



Supplementary Figure 17: Pairwise plots of miRNA contributions for **a)** Sparse mCCA vs. AJIVE, **b)** MOFA vs. AJIVE, and **c)** MOFA vs. Sparse mCCA

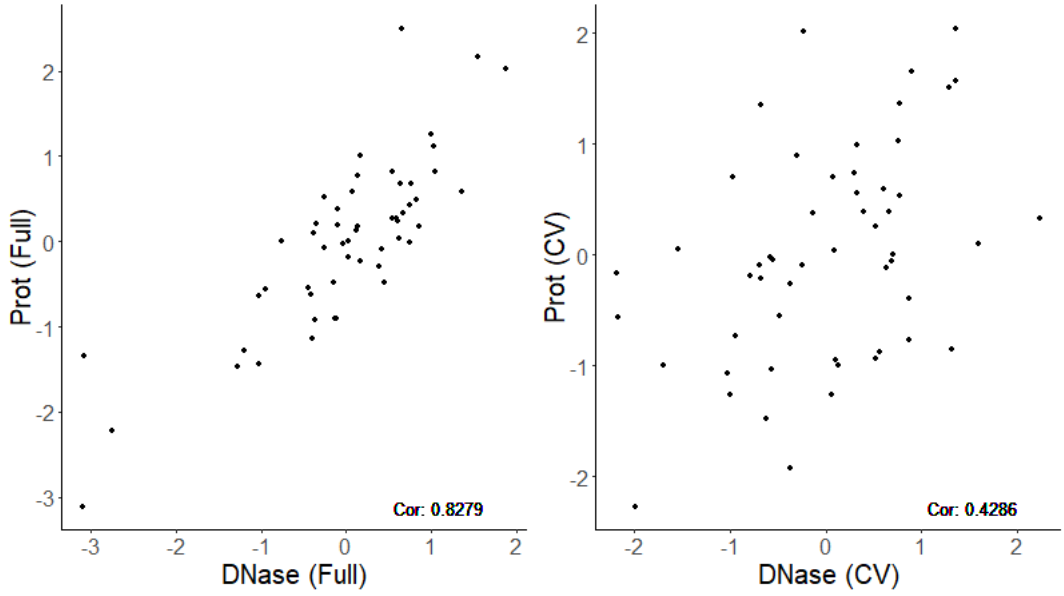


Supplementary Figure 18: Plot of expression of ESR1 gene against the RNA contribution for **a)** Sparse mCCA, **b)** MOFA, and **c)** AJIVE.

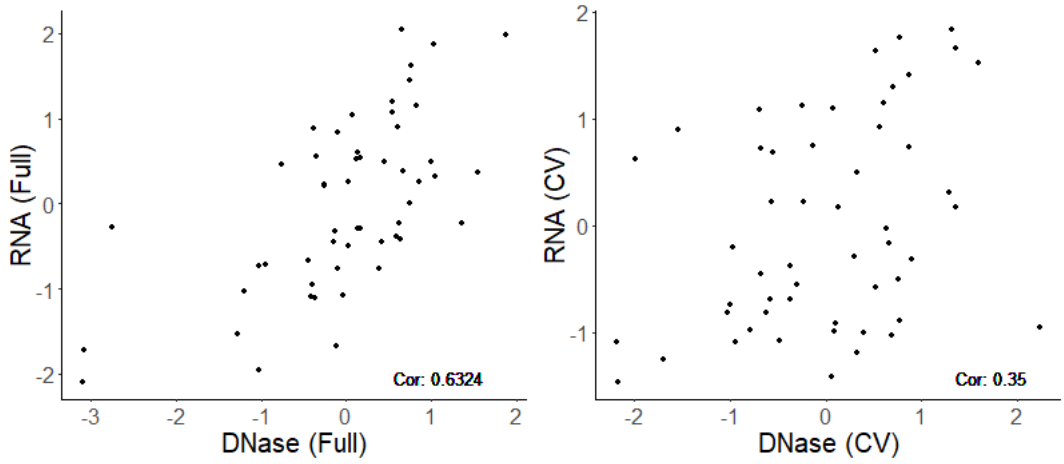


2.3 Small Sample Size - Li et al, 2016

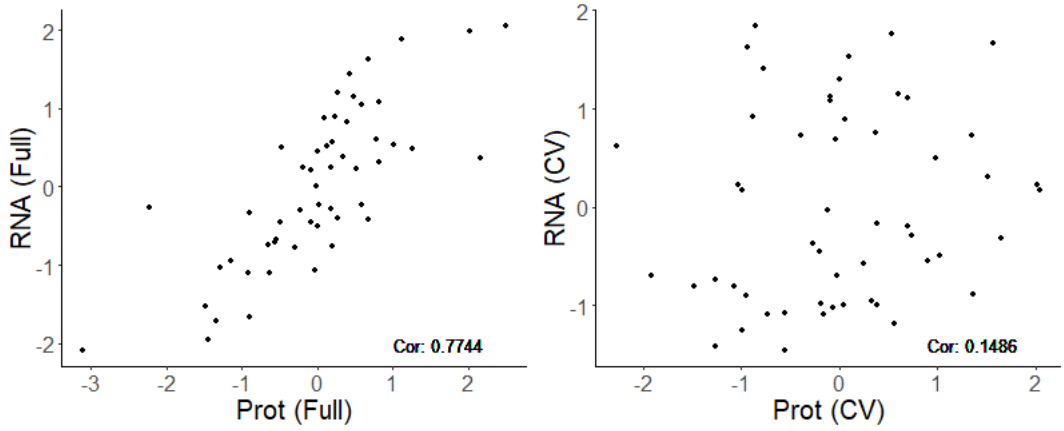
Supplementary Figure 19: MCCA with small-sample dataset: Contribution plots for DNase and Protein expression using MCCA.



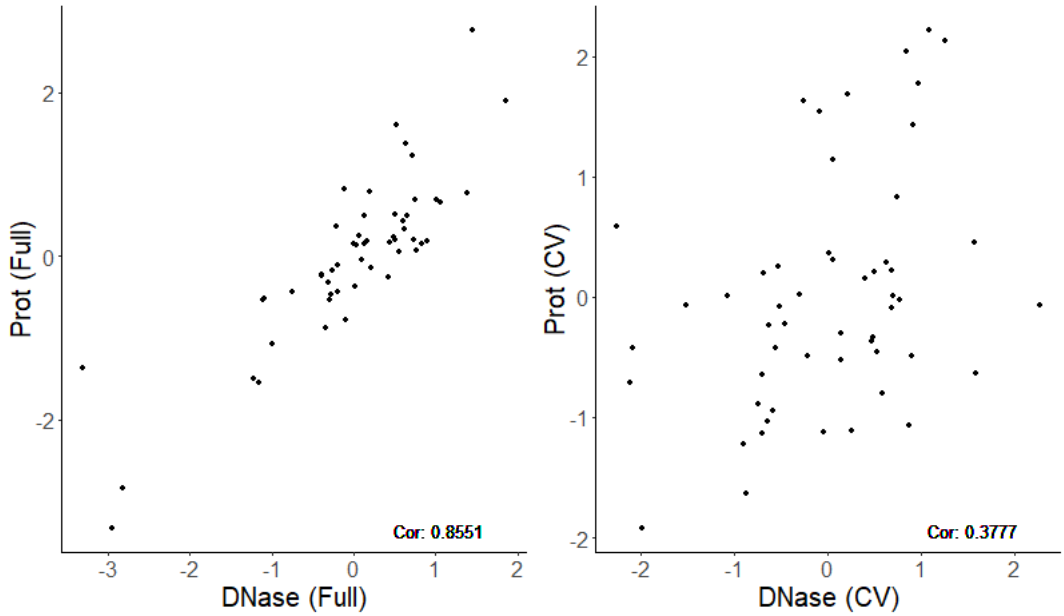
Supplementary Figure 20: MCCA with small-sample dataset: Contribution plots for DNase and gene expression using MCCA.



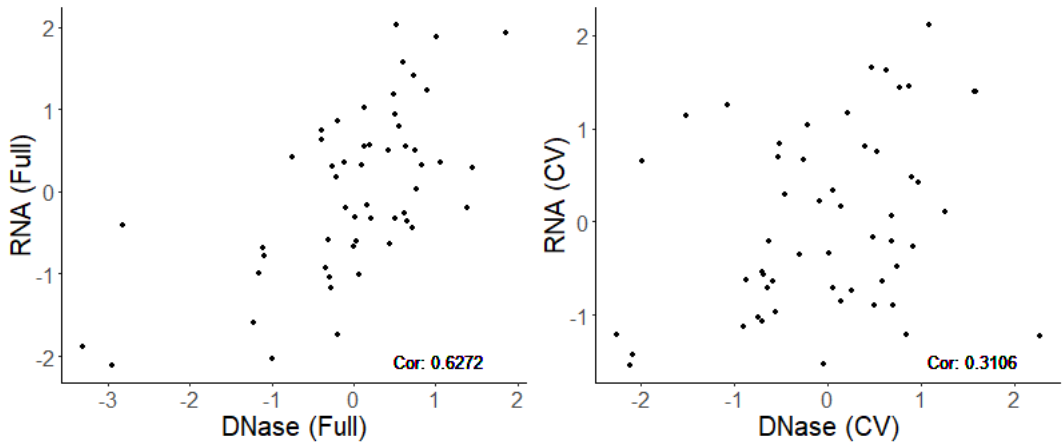
Supplementary Figure 21: MCCA with small-sample dataset: Contribution plots for protein expression and gene expression using MCCA.



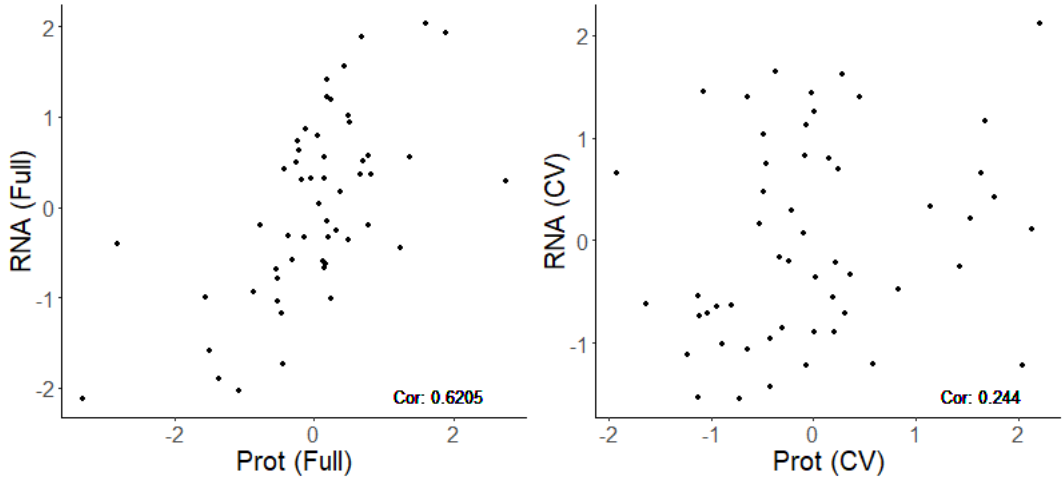
Supplementary Figure 22: MOFA with small-sample dataset: Contribution plots for DNase and protein expression using MOFA.



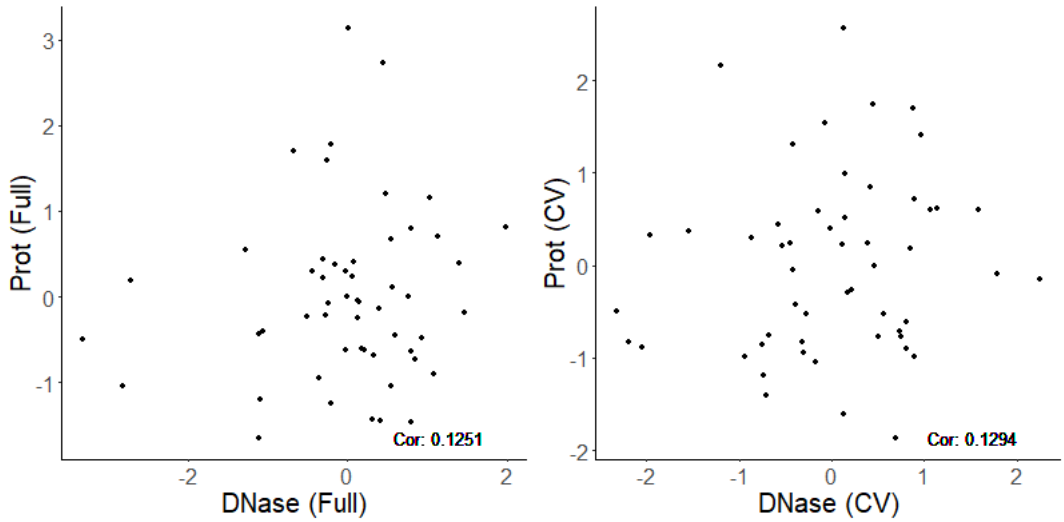
Supplementary Figure 23: MOFA with small-sample dataset: Contribution plots for DNase and gene expression using MOFA.



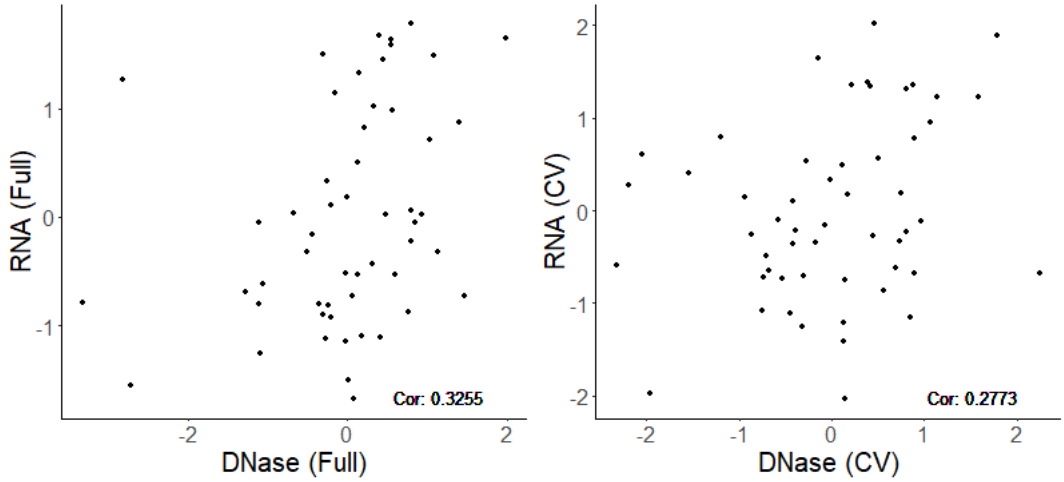
Supplementary Figure 24: MOFA with small-sample dataset: Contribution plots for protein expression and gene expression using MOFA.



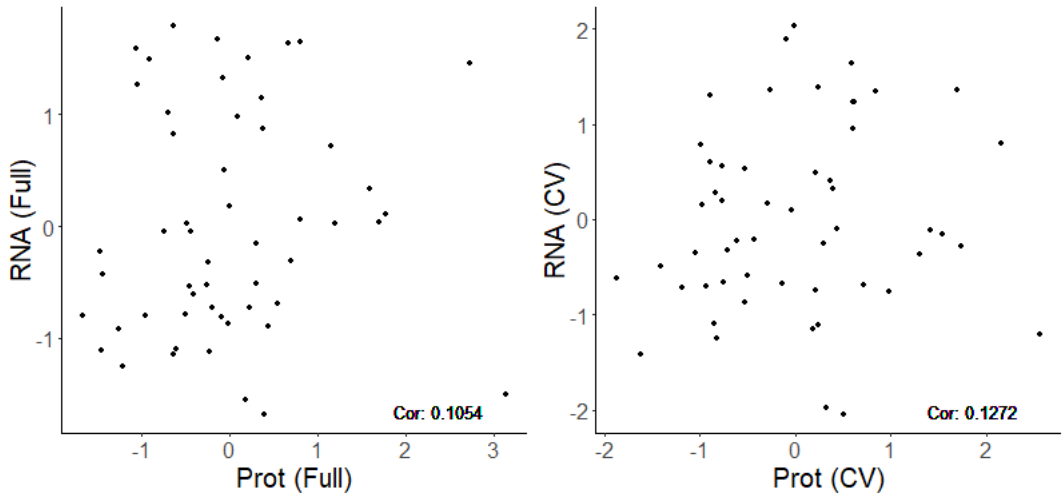
Supplementary Figure 25: AJIVE with small-sample dataset: Contribution plots for DNase and protein expression using AJIVE.



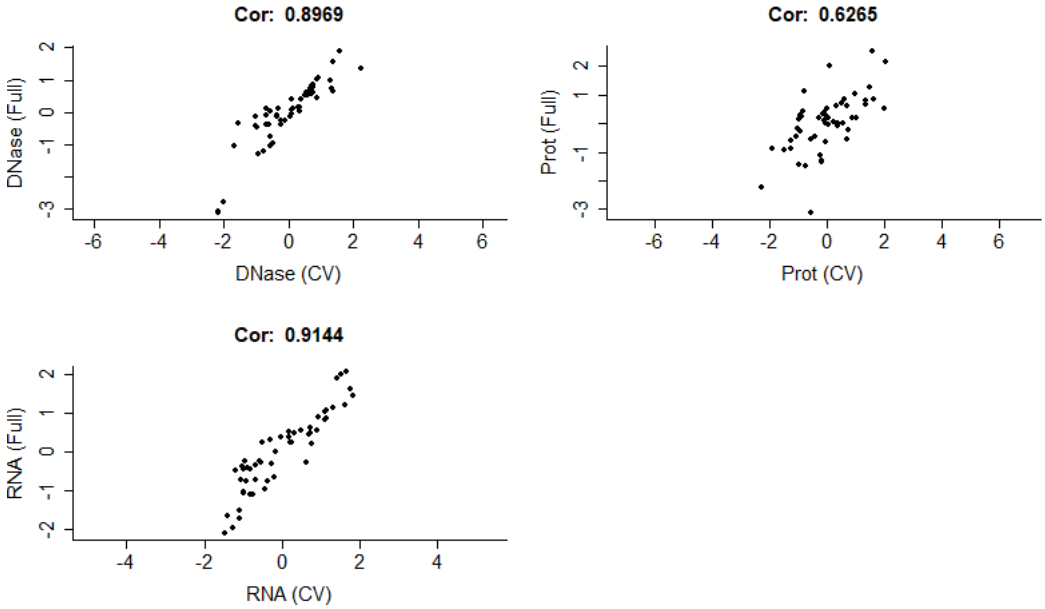
Supplementary Figure 26: AJIVE with small-sample dataset: Contribution plots for DNase and gene expression using AJIVE.



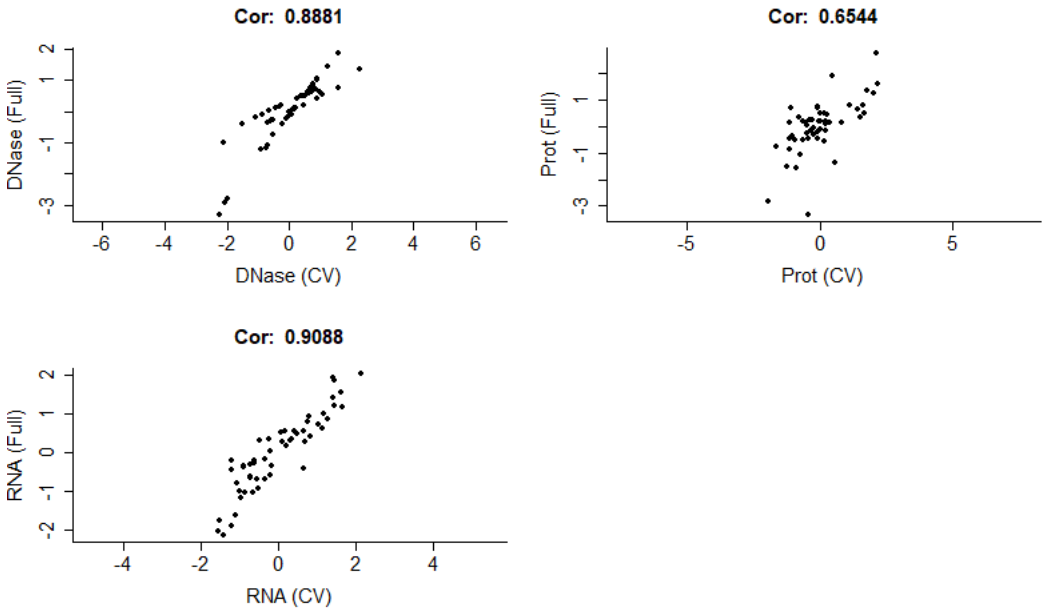
Supplementary Figure 27: AJIVE with small-sample dataset: Contribution plots for protein expression and gene expression using AJIVE.



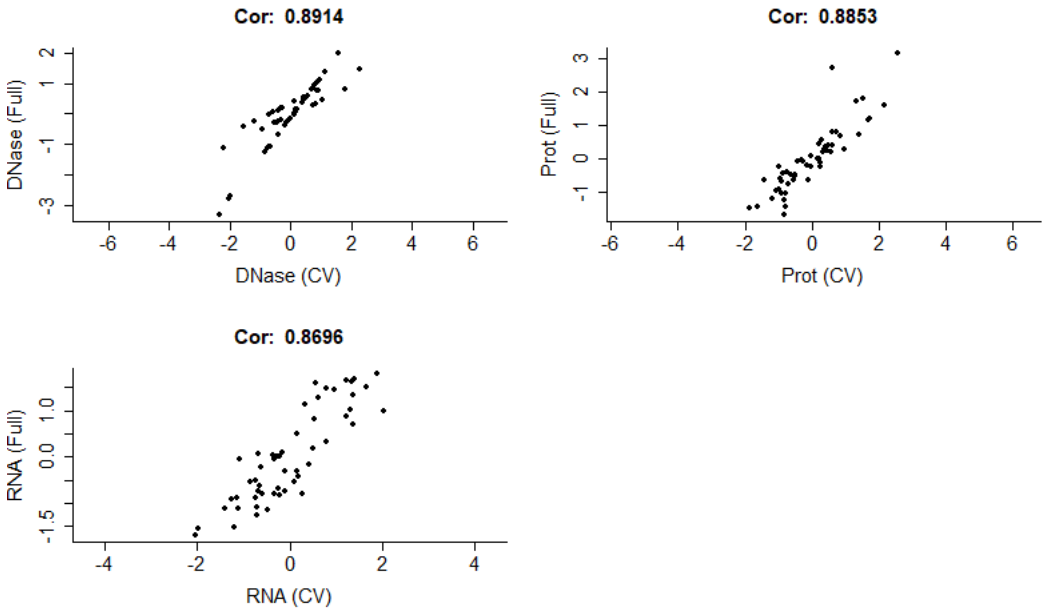
Supplementary Figure 28: MCCA with small-sample dataset: Comparison plots for DNase, protein expression, and gene expression using AJIVE.



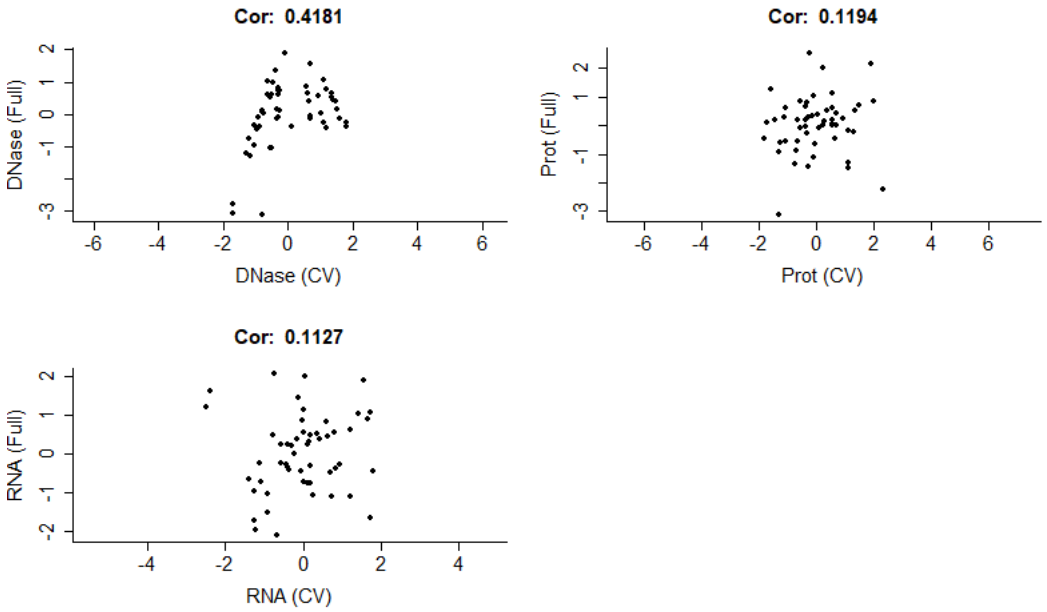
Supplementary Figure 29: MOFA with small-sample dataset: Comparison plots for DNase, protein expression, and gene expression using AJIVE.



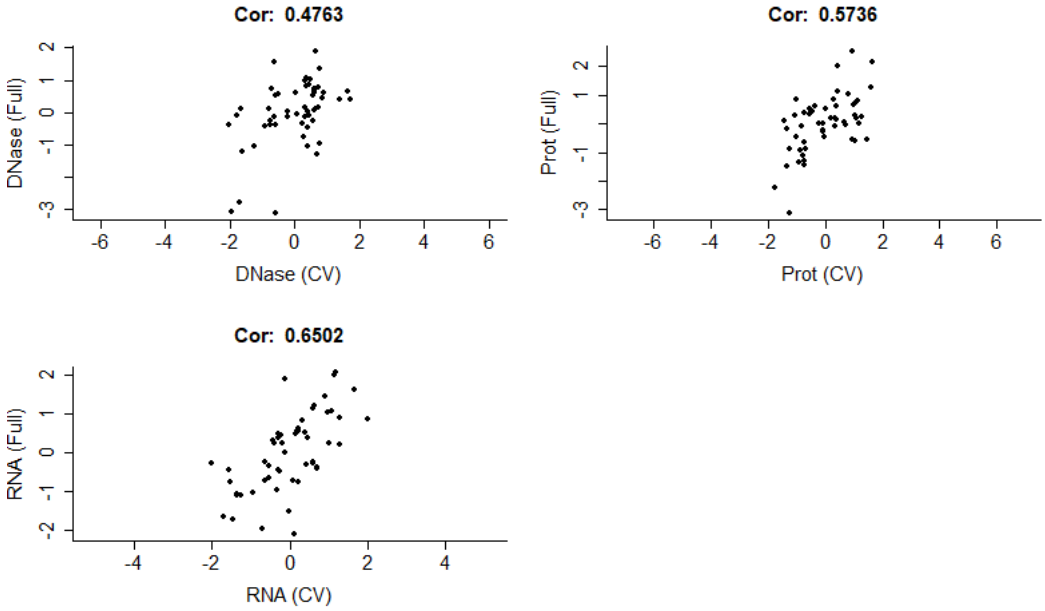
Supplementary Figure 30: AJIVE with small-sample dataset: Comparison plots for DNase, protein expression, and gene expression using AJIVE.



Supplementary Figure 31: Comparison plots of contributions from Sparse mCCA for the 3 fold analysis.

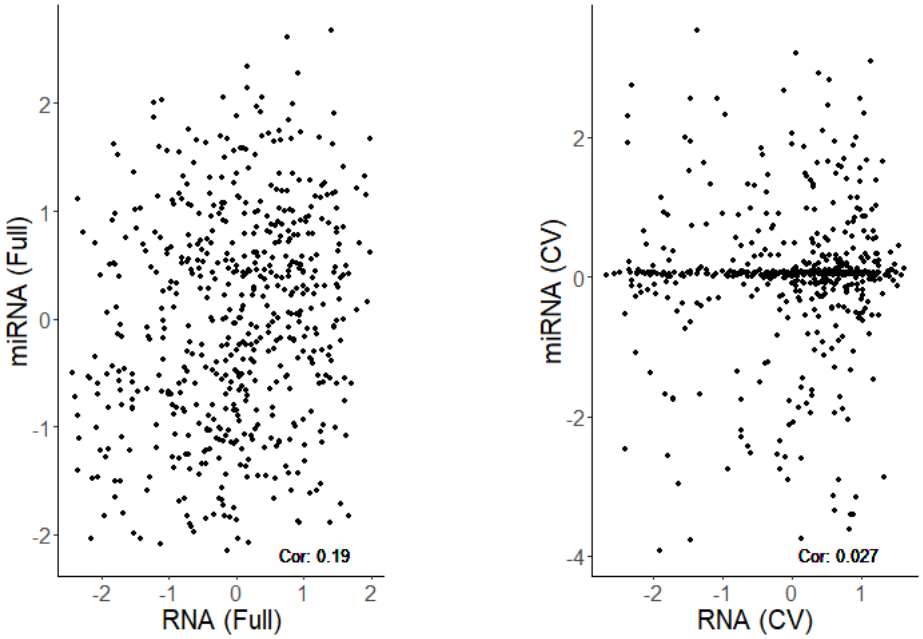


Supplementary Figure 32: Comparison plots of contributions from Sparse mCCA for the 10 fold analysis.

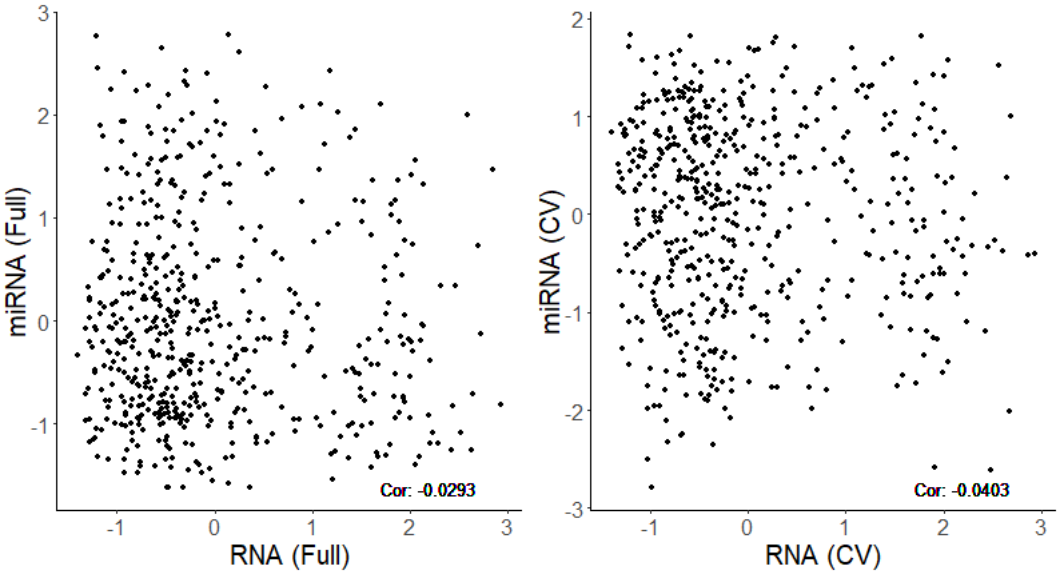


2.4 Permuted Null Data

Supplementary Figure 33: MOFA with null data: Side-by-side contribution plots for miRNA versus gene expression using MOFA.



Supplementary Figure 34: AJIVE with null data: Side-by-side contribution plots for miRNA versus gene expression using AJIVE.



References

[1] Shen, R., Olshen, A.B., and Ladanyi, M. (2010). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, **26(2), 292-293**.

[2] Wang, B., Mezlini, A.M., Demir, F., et al. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods*, **11, 333-337 doi: 10.1038/nmeth.2810**

[3] Wong, K.Y., Fan, C., Tanioka, M., et al. (2019). I-Boost: an integrative boosting approach for predicting survival time with multiple genomics platforms. *Genome Biology*, **20:52**.

[4] Witten, D.M., and Tibshirani, R.J. (2009). Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical Applications in Genetics and Molecular Biology*, **8 (1), 28**.

[5] Feng, Q., Jiang, M., Hannig, J., and Marron, J.S. (2018). Angle-based joint and individual variation explained. *aRxiv*, **1704.02060v3**.

[6] Argelaguet, R., Velten, B., Arnol, D. et al. (2017). Multi-Omics Factor Analysis - a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, **14, e8124**.

[7] Hotelling, H. (1936). Relations between two sets of variates. *Biometrika,* **28, 321-377**.

[8] Meng, C., Zeleznik, O.A., Thallinger, G.G., et al. (2016). Dimension reduction techniques for the integrative analysis of multi-omics data. *Briefings in Bioinformatics*,**17(4), 628-641.**

[9] Pucher, B.M., Zeleznik, O.A., and Thallinger, G.G. (2018). Comparison and evaluation of integrative methods for the analysis of multilevel omics data: a study based on simulated and experimental cancer data. *Briefings in Bioinformatics,* **1-11.**

[10] Tini, G., Marchetti, L., Priami, C., and Scott-Boyer, M.P. (2017). Multi-omics integration—a comparison of unsupervised clustering methodologies. *Briefings in Bioinformatics*, **bbx167, https://doi.org/10.1093/bib/bbx167.**

[11] Soneson, C., Lilljebjörn, H., Fioretos, T., and Fontes, M. (2010). Integrative analysis of gene expression and copy number alterations using canonical correlation analysis. *BMC Bioinformatics*, **11, 191.**

[12] Brown, B.C., Bray, N.L., and Pachter, L. (2018). Expression reflects population structure. *bioRxiv*, **10.1101/364448**.

[13] Fertig, E.J., Ren, Q., Cheng, H., et al. (2012). Gene expression signatures modulated by epidermal growth factor receptor activation and their relationship to cetuximab resistance in head and neck squamous cell carcinoma. *BMC Genomics*, **13, 160**.

[14] Broad Institute TCGA Genome Data Analysis Center. Analysis-ready standardized TCGA data from Broad GDAC Firehose 2016_01_28 run.

[15] Li, Y.I., van de Geinj, B., Raj, A., et al. (2016). RNA splicing is a primary link between genetic variation and disease. *Science*, **352(6285, :600-604**.

[16] Holst, F., Stahl, P.R., Ruiz, C., et al. (2007). Estrogen receptor alpha (ESR1) gene amplification is frequent in breast cancer. *Nature Genetics*, **39**, 655-660.

[17] Lê Cao, K.A., Martin, P.G.P., Robert-Granié, C., et al. (2009). Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinformatics*, **10:34**

[18] Meng, C., Kuster, B., Culhane, A.C., et al. (2014). A multivariate approach to the integration of multi-omics datasets. *BMC Bioinformatics*, **15:162**

[19] Rohart, F., Gautier, B., Sing, A., et al. (2017). mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Computational Biology* , **13(11)**

[20] Argelaguet, R., Mohammed, H., Clark, S., et al. (2018). Single cell multi-omics profiling reveals a hierarchical epigenetic landscape during mammalian germ layer specification. *bioRxiv*, **519207; doi: https://doi.org/10.1101/519207** .

[21] ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489(7414):57-74**

[22] Roadmap Epigenomics Consortium. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, **518**,317-330.