Executable network models of integrated multi-omics data

**Mukta G. Palshikar1, Xiaojun Min2, Alexander Crystal2, Jiayue Meng2, Shannon Hilchey3, Martin Zand3,4, Juilee Thakar1,4,5,6,7\***

1 Biophysics, Structural, and Computational Biology Program, University of Rochester School of Medicine and Dentistry, Rochester, USA. 2University of Rochester College of Arts, Sciences and Engineering, Rochester, USA. 3Division of Nephrology, University of Rochester School of Medicine and Dentistry, Rochester, USA. 4University of Rochester Clinical & Translational Science Institute, Rochester, USA. 5Department of Biostatistics and Computational Biology, University of Rochester School of Medicine and Dentistry, Rochester, USA. 6Department of Microbiology and Immunology, University of Rochester School of Medicine and Dentistry, Rochester, USA. 7Department of Biomedical Genetics, University of Rochester School of Medicine and Dentistry, Rochester, USA.

# Abstract

* Integrate 3 multi-omics datasets from the same system to create multi-layered networks for which Boolean rules can be inferred.
* **Network-based integration of multiomics data suggests a mechanism for HIF-1α-mediated chemotaxis in B cells.**

**Manuscript outline:**

1. Multi-omics network modeling and pathway enrichment analysis with mBONITA
   1. Outline methods for data integration, rule inference, node modulation scores, and pathway analysis (Figure 1)
2. mBONITA identifies mechanisms of hypoxia-mediated chemotaxis in RAMOS B cells (pathway analysis with mBONITA)
   1. Supplement – transcriptomics data analysis (Supplementary Figure 1)
   2. Correlation between omics datasets (Figure 2A and B)
   3. Pathway analysis with mBONITA on multiomics data (Figure 2C )
3. Pathway-based prioritization of genes in a signaling network with mBONITA
   1. Node importance score: show a case study of a LSP1/HIF1A-centric signaling network
4. Benchmarking of mBONITA
   1. Rule inference: Supplement - show that mBONITA identifies a smaller rule set from combined omics data than from individual datasets
   2. Pathway analysis:
      1. mBONITA identifies more significant pathways than:
         1. PaintOmics:
         2. LeapR:
         3. ActivePathways:
         4. CAMERA:
         5. BONITA (TBD)

Show pathways in supplement figures & tables

* + 1. mBONITA identifies different node importance scores from BONITA:
       1. Supplement – low correlations between node importance score from single omics and multi-omics data (ie, a comparison to mBONITA)

# Introduction

molecular signaling is complex and has transcriptional and post transcription, Protein and genes integration is important

Modern sequencing technology allows biologists to study conditions at different molecular levels by generating ‘omics’ datasets such as transcriptomics (mRNA levels), proteomics (protein levels) and phosphoproteomics (phosphoprotein levels). These datasets are extraordinarily rich and allow sophisticated inferences about molecular signaling individually. In particular, pathway enrichment analysis allows the identification of modulated biological processes by two main classes of methods 1, 2 – over-representation analysis and functional class scoring, or topology-based pathway enrichment analysis. However, technical and biological variability between these layered datasets present challenges for integrative computational analyses in general and pathway analysis in particular.

Discrete-state modeling characterizes network topologies with Boolean rules or gates that define signal flow through the network. These networks can be simulated either synchronously or asynchronously to identify limit cycles or attractors that correspond to network-specific states/phenotypes. We have recently published two algorithms that infer regulatory rules for prior knowledge networks (PKNs) from omics data 3, 4. These inferred regulatory rules are used to perturb and simulate networks *in silico* to calculate the influence of nodes over signaling through the network. This perturbation-based scores are combined with expression data to perform pathway analysis. However, both these methods rely on information from a single omics training dataset to perform both rule inference and pathway analysis.

Here, we present a method multi-omics Boolean Omics Network Invariant Time Analysis (mBONITA) to (a) use multiple layers of omics data to improve inference of regulatory rules, (b) use expression from all these layers to calculate node importance scores and identify condition-related subnetworks and (c) perform pathway analysis that incorporates information from multiple molecular layers. We demonstrate the utility of this algorithm on a multi-omics dataset from RAMOS B cells grown under hypoxic conditions and treated with cyclosporine A (CyA). This dataset consists of three layers – transcriptomics, proteomics, and phosphoproteomics. We have previously published the analysis and validation of the proteomics 5 and phosphoproteomics dataset 6. Here wepresent the conventional analysis of the transcriptomics dataset in this manuscript. These datasets have limited correlation in accordance with previous studies. We show that our method can effectively use this multi-omics dataset in combination with PKNs from KEGG 7 and WikiPathways 8, 9 to (a) identify an accurate and reliable regulatory rule set for PKNs, (b) use this improved rule set to calculate node modulation scores that incorporate all available expression information and the network topology and (c) use these modulation scores to identify modulated pathways that suggest a mechanism for hypoxia-mediated chemotaxis in B cells. We also demonstrate this on a multi-omics dataset from persons with renal cancer 10. We used mBONITA to identify pathways that are significantly dysregulated in the three contrasts in our multi-omics dataset, including pathways that are not significantly dysregulated in individual datasets. We compare these pathways to those identified by our competitor algorithms PaintOmics4 11, 12, CAMERA 13 in combination with Fisher’s method of p-value combination as suggested in ReactomeGSA 14, LeapR 15, multiGSEA 16 and ActivePathways 17, and show that mBONITA identifies the most relevant pathways to these conditions. We use mBONITA to calculate node modulation scores for a large signaling network describing the HIF1A-mediated signaling in B cells and identify that are highly modulated in this condition. We show that the genes identified by mBONITA show improved condition specificity and contain strong candidates for experimental validation.

In this manner, we show here that mBONITA uses a Boolean rule-based network propagation method that integrates multiple sources of omics/perturbation data to prioritize nodes in an interaction network, can identify subnetworks of highly modulated genes, and can perform pathway analysis using these multiple sources of omics data to present a complete picture of modulated signaling in the condition under study.

# Results

## Multi-omics network modeling and pathway enrichment analysis with mBONITA

multiomics -Boolean Omics Network Invariant-Time Analysis (mBONITA) extends our previous approaches to Boolean modeling and pathway analysis with omics datasets 18, 19. mBONITA uses trascriptomics, proteomics and phophoproteomics data to learn logic rules for prior knowledge networks (PKNs). Briefly, the BONITA algorithm uses omics datasets in combination with prior knowledge networks from sources such as KEGG 7 and WikiNetworks 9, 20 to infer Boolean rules which define signal flow through these biological networks. These Boolean models allow *in silico* perturbations and simulation of the networks to calculate node-wise importance scores that quantify a node’s influence in a network. These node importance scores are used in concert with variance and condition-specific fold changes from the training dataset to calculate a metric of pathway modulation and hence perform pathway analysis. We have reimplemented the BONITA Python tool in Python3, resulting in significant upgrades in speed, and use this updated algorithm as a basis for the mBONITA module. mBONITA is a three-step process that requires four inputs (Figure 1): (1) prior knowledge networks in graphml format, defining the topology of the signaling network(s) for which Boolean rules are to be inferred, (2) a matrix of gene/protein expression values from the multi-omics datasets under consideration, (3) a design matrix specifying the treatment for each sample in the training dataset and (4) a contrast matrix. We have only tested mBONITA in scenarios where conditions are matched across all omics datasets. In the first step, pathways are downloaded from KEGG using the KEGG API if desired, and prepared for rule inference. In the second step, Boolean rule inference is performed with a combination of a genetic algorithm and a local search as described previously. In the third and final step, *node importance scores* are calculated for each network by in-silico knock-ins and knock-outs. These node importance scores are topology-specific, not condition-specific. These scores are weighted by the fold-changes for each contrast from each dataset, and by the standard deviation of the gene across each dataset, and by the strength of the evidence for that gene across all datasets, to calculate a *node modulation score*. This means that each gene in the pathway is assigned an *evidence score* from 1 to the number of omics datasets, depending on the number of omics datasets in which it has a measured non-zero abundance. A *pathway modulation score* is calculated by summing up the node modulation scores for nodes in the pathway. A *p-value* is calculated by generating a distribution of the pathway modulation scores by resampling fold-changes, standard deviations, and evidence scores from the dataset. In a typical mBONITA analysis, these steps are automatically performed for all KEGG pathways that overlap with the training dataset. The outputs of this analysis are a table of p-values for each pathway in each contrast, graphml files annotated with fold-changes and importance scores, ready to be imported into network visualization software such as Cytoscape or Gephi, and tables of node modulation scores for each combination of pathway and contrast.

## mBONITA identifies mechanisms of hypoxia-mediated chemotaxis in RAMOS B cells

We used mBONITA to perform an integrative pathway analysis of three omics datasets generated from RAMOS B cells grown under hypoxic and normoxic conditions, in combination with treatment with the calcineurin inhibitor cyclosporine A (CyA) (Supplementary Table 1). CyA modulates oxygen-dependent chemotaxis in human B cells via the transcription factor HIF1A. Transcriptomics, proteomics, and phosphoproteomics were measured to XXXX. Previous analysis of proteomic and phosphoproteomic levels shows modulation of cytoskeletal rearrangement 21, 22. However, differentially regulated pathways identified from proteomic and phosphoproteomic analysis were not different at at the transcriptomic level (Supplementary Figure 1A-C). These discrepancies suggested that an integrative pathway analysis that took all three levels of available omics data into account would provide more insights into the mechanisms underlying the observed phenotypes of modulated cell migration. Datasets were processed as described in the Methods. We considered only conditions that were profiled in all three datasets (Supplementary Table 1). Preliminary analysis showed that there were significant differences in the number of molecular entities profiled in the three datasets. Only 1505 genes were profiled in all three datasets out of a total of 17846 profiled genes (Figure 1A). In addition, the measured abundances of these 1505 genes had a low Spearman correlation across datasets even when separated by condition, in line with previous studies (move to discussion), ranging from x-y, pvalue < z (insert values from Jupyter notebook here) (Figure 1B).

We then performed pathway analysis with mBONITA on the combined omics datasets and identified pathways modulated in three contrasts (Figure 1C, Supplementary File 1). mBONITA identified that the HIF1-A signaling pathway is dysregulated between samples grown at 1% oxygen without CyA and those grown at 1% oxygen with CyA, and the chemokine signaling pathway is dysregulated between samples grown at 1% oxygen without CyA and those grown at 1% oxygen with CyA. Other modulated pathways include the progesterone-mediated oocyte maturation, oocyte meiosis and breast cancer pathways, all of which include many nodes linked to the MAPK signaling pathway. The long-term depression pathway, which includes components of the calcium signaling pathway, is also dysregulated between samples grown at 1% and 19% oxygen without cyclosporine. This small list of pathways identified by mBONITA is highly interpretable and specific to the condition under study.

## Pathway-based prioritization of genes in a signaling network with mBONITA

We used BONITA’s rule determination algorithm BONITA-RD to identify Boolean rules for all three datasets as well as an integrated dataset as described in the Methods. The average size of the equivalent rule set (ERS) for nodes with in-degree >= 3, which is used as a proxy for BONITA-RD’s ability to narrow down the state space of Boolean rules, was significantly different for the same networks between different datasets (t-test, adjusted p-value < 0.05) (Supplementary Figure 2A). We found that BONITA-RD’s rule inference algorithm inferred smaller (and hence more high-confidence) rule sets when omics datasets were combined to form a single training dataset (Supplementary Figure 2A). We also calculated node importance scores using BONITA using each training dataset and found that importance scores were highly correlated between datasets (Supplementary Figure 2B, Supplementary Table 3). We note that node importance scores are independent of dataset-specific fold changes and are dependent solely on network topology and inferred Boolean rules. Both these observations underscore the difference in signaling information that can be obtained from these molecular layers and the importance of a meaningful combination of abundance measures to learn node importance scores.

We demonstrate the effectiveness and interpretability of mBONITA’s node modulation score on a previously-described custom network describing the HIF1A-mediated response of B cells to hypoxia and treatment with CyA 6 (Supplementary Table 2). (Results pending – experiment still running, placeholder figure on slide #7).

***Corresponds to Figure 2***

**Outline:**

Rule inference – mBONITA identifies a small set of rules from combined omics data

Node importance scores – mBONITA prioritizes genes according to their influence over signal flow

Case study – LSP1/HIF1A centric network, node importance scores,

## Benchmarking of pathway analysis with mBONITA

We compared mBONITA to six other pathway analysis methods as described in the Methods – ActivePathways 17, CAMERA in combination with Fisher’s method of p-value combination as suggested by the authors of ReactomeGSA 13, 14, PaintOmics4 11, 12, leapR 15, multiGSEA 16, and the original BONITA pathway analysis in combination with Fisher’s method of p-value combination 18. Complete pathway analysis results with each of these methods are presented in the Supplementary Data (Supplementary Figures 4-9, Supplementary Files 2-7). Across all three tested contrasts, the original BONITA pathway analysis returned the largest number of significantly modulated pathways (Benjamini-Hochberg adjusted p-value < 0.05) (Figure 4A-C). Most of these pathways were only identified by the original BONITA pathway analysis and represented a large fraction of the total number of KEGG pathways, suggesting that these results are non-specific. LeapR did not identify any modulated pathways. ActivePathways, PaintOmics, CAMERA, and mBONITA all identified a moderate number of significantly modulated pathways (1 - 8) across all contrasts, suggesting greater specificity for these methods (Figure 4 A-C). Across all contrasts, mBONITA identifies nine significantly modulated pathways of which two are also identified by BONITA in the 1% O2, CyA+ vs 1% O2, CyA- contrast (Figure 4B, Figure 2C). Out of 13 key KEGG pathways known to be involved in the mechanism of the HIF1A-mediated chemotactic response of human B cells to oxygen gradients and treatment with CyA (Supplementary Table 2), mBONITA correctly identifies the HIF1-signaling pathway as being modulated in the contrast 19% O2,CyA- vs 1% O2,CyA- and is the only method to identify the chemokine signaling pathway as being modulated in the contrast 1%O2,CyA+ vs 1%O2,CyA- (Figure 4D).

# Materials and Methods

## Transcriptomics data collection and analysis

RAMOS cells were maintained in a 37 degree Celsius, 5%CO2, humidified incubator in cR10 media (RPMI 1640 media supplemented with 10% heat inactivated fetal bovine serum (FBS), 50 U/mL Penicillin, 50 ug/mL Streptomycin and 50 uM 2-Mercaptoethanol). RAMOS cells, in triplicate, were treated with either 0 or 1 ug/mL cyclosporine A (CyA) and incubated at either 19% oxygen (traditional tissue culture) or 1% oxygen for 24 hours. After incubation with CyA at the indicated oxygen conditions, cells were harvested by centrifugation and washed 3X with phosphate buffered saline (PBS). RNA was extracted from the resultant cell pellets using TRIzol™ Plus RNA Purification Kits according to the manufacturer's recommendations (Invitrogen). Single-end RNA-sequencing was performed on the Illumina NextSeq 550. Raw data was formatted using bcltofastq-2.19.0. Sequence reads were trimmed for adaptor sequence/low-quality sequence using Trimmomatic-0.36. Trimmed sequence reads were mapped to Reference Genome hg38/GencodeV28 using STAR\_2.6.0c. Read quantification was performed using featureCounts from the R package subread version 1.34.7 using genome assembly GRCh38.p12.

Differentially expressed (DE) genes were identified using DESeq2 23. The R package ‘ashr’ was used for log fold change shrinkage 24. Genes with a Benjamini-Hochberg adjusted p-value < 0.05 and an absolute log2-fold change > 0.5 were identified as being DE. Heatmaps were prepared using ComplexHeatmap 25. Over-representation analysis of DE genes was performed with the R package clusterprofiler, using gene sets of canonical KEGG pathways from the MSigDB database 7, 26, 27. Gene sets were identified as being over-represented if the un-adjusted p-value was less than 0.05.

## Data processing for pathway analysis

Proteomics and phosphoproteomics data were collected and processed as described in 21 and 6 respectively. We retained only samples from the experimental conditions represented in all three datasets (Supplementary Table 1). In the case of the proteomics and phosphoproteomics datasets, we mapped protein names to gene names using Entrez and retained these gene names for downstream analysis, for consistency between datasets. We discarded observations for genes whose median value was 0. All data was log2(x+1)-transformed.

## Pathway analysis with mBONITA

We reimplemented our previously published algorithm BONITA 18 in Python3, resulting in a significant improvement in speed. We used this implementation of BONITA to infer Boolean rules individually for the three multi-omics datasets and for an integrated dataset comprising samples for conditions that were profiled in all three datasets (Supplementary Table 1). For each of these experiments, we used all KEGG networks with an overlap of 5 or more genes with the training dataset. *Node impact scores (Ig)*, which quantify the effect of individual nodes *g* over signal flow through a network, were calculated by *in silico* perturbation of networks as previously described 18. We used this node impact score in concert with dataset *d*-specific fold changes (*Qgd*), standard deviation (*std(Vgd*)), and *evidence scores* (*Eg*, Eqn. 1) to calculate *node modulation scores (Mg,* Eqn. 2*)* and hence an overall *pathway modulation score (Mp, Eqn.* 3*)*. A p-value was calculated by generating a distribution of pathway modulation scores by resampling dataset-specific fold changes, standard deviations, and evidence scores from the training dataset.

|  |  |  |
| --- | --- | --- |
|  |  | (1) |

where is the evidence score for a gene, is the number of multi-omics datasets, and is the measured abundance value of gene in dataset .

|  |  |  |
| --- | --- | --- |
|  |  | (2) |

where is the modulation score for a gene, and is the standard deviation of the measured abundance value of gene in dataset .

|  |  |  |
| --- | --- | --- |
|  |  | (3) |

where is the pathway modulation score for pathway and is the number of genes in the pathway .

## Comparison to other methods

Parameters for other packages here:

multiGSEA: used DE genes from limma

PaintOmics: relevant genes defined as those with foldchange (check this) > 4th quantile

LeapR:

ActivePathways:

CAMERA:

BONITA:

**Other multi-omics pathway analysis methods:**

1. **DONE - reactomeGSA** – applies CAMERA + Reactome to each dataset separately, returns individual p-values, correlations between datatypes. Adapted this method – used CAMERA + KEGG on each dataset separately, used a p-value combination method (Stouffer’s) to get an overall p-value per pathway.
2. **DONE - paintOmics** (<https://doi.org/10.1093/nar/gkac352>, <https://doi.org/10.1093/nar/gky466>): Fisher’s exact test for each dataset, joint pathway enrichment p-value for all omics data is computed by applying either Fisher combined probability test or Stouffer’s method. Other interesting pathway visualization features.
3. **DONE - ActivePathways** ([doi:10.1038/s41467-019-13983-9](https://doi.org/10.1038/s41467-019-13983-9)): “From a matrix of p-values, ActivePathways creates a ranked gene list where genes are prioritized based on their combined significance of in the series of omics datasets provided in the input matrix. The ranked gene list includes the most significant genes first. ActivePathways then performs a ranked hypergeometric test to determine if a pathway (i.e., a gene set with a common functional annotation) is enriched in the ranked gene list, by performing a series of hypergeometric tests (also known as Fisher’s exact tests). In each such test, a larger set of genes from the top of the ranked gene list is considered. At the end of the series, the ranked hypergeometric test returns the top most significant p-value from the series, corresponding to the point in the ranked gene list where the pathway enrichment reached the greatest significance of enrichment. This approach is useful when the genes in our ranked gene list have varying signals of biological importance in the input omics datasets, as the test identifies the top subset of genes that are the most relevant to the enrichment of the pathway.”
4. Tried a pca-based method PathwayPCA (<https://doi.org/10.1002/pmic.201900409>) and couldn’t get the R package to work. Update – this was my error, it’s not really a pathway analysis method and only tests association of pathways with specific categorical phenotypes, no way to test a contrast.
5. **DONE - leapR** (<https://doi.org/10.1021/acs.jproteome.0c00963>): “Many applications of enrichment compare one group of samples (case) against another group (control) with the goal of identifying pathways that have significantly different abundance in this comparison. The leapR package accomplishes this in the enrichment\_comparison (see [Figure 1](https://pubs.acs.org/doi/10.1021/acs.jproteome.0c00963) and [Table 1](https://pubs.acs.org/doi/10.1021/acs.jproteome.0c00963)) using a *t* test in which the overall abundance of the pathway members is summarized in distributions for the case and control groups and then compared. Output from this analysis will yield *p*-values for each input pathway that indicate the significance of enrichment. Examining the mean abundance from each condition will provide an idea of the effect size and the direction of enrichment—that is, is the pathway more abundant in the case or control condition? A small effect size can still yield very significant *p*-values, but these kinds of results must be treated with caution.”
6. **NOT SURE WHETHER TO INCLUDE - MGSEA** (<https://doi.org/10.1186/s12859-019-2716-6>): adaptation of GSEA to multiomics data. Haven’t looked at this code yet, it is just a supplementary R file.
   1. Not to be confused with multiGSEA (<https://doi.org/10.1186/s12859-020-03910-x>) which just applies GSEA to each sample individually and uses a pvalue combination method or with MOGSA (<https://doi.org/10.1074/mcp.TIR118.001251>) which is a single-sample method

## Data and software availability statements

The transcriptomics dataset described in this manuscript has been deposited to NCBI-GEO with the accession number insert accession number here and is available at insert URL here. The mass spectrometry phosphoproteomics and proteomics datasets are available at the ProteomeXchange Consortium partner repository PRIDE with the dataset identifiers PXD036167 and insert accession number respectively. The source code, documentation, and tutorials for the BONITA3 Python tool and the mBONITA pathway analysis module are freely available at <https://github.com/Thakar-Lab/BONITA-Python3>.

# Discussion

# Figures and Tables

## Main figures:

1. Figure 1: moBONITA integrates information from multiple omics datasets to learn a consensus set of logic rules for simulation and perturbation of prior knowledge networks
2. Figure 2: mBONITA identifies mechanisms of hypoxia-mediated chemotaxis from a multi-omics datasets from RAMOS B cells grown under three conditions (pathway analysis with mBONITA) (A) 1505 genes were profiled in all three omics datasets (median log2-abundance > 0) (B) The multi-omics datasets showed low inter-dataset correlations. Distinct experimental conditions are indicated by colors and shapes as shown in the legend. (C )Pathways known to be involved in the hypoxia-mediated response to CyA, Only pathways identified as significant from a combined dataset by mBONITA are shown. Pathways are defined as differentially regulated if the Benjamini-Hochberg corrected p-value < 0.05.
3. Figure 3: Pathway-based prioritization of genes in a signaling network with mBONITA. Node importance score: show a case study of a LSP1/HIF1A-centric signaling network (TO BE DONE). (A) Network figure (B) Heatmap of node modulation scores. This is a placeholder/draft figure showing node modulation scores for each dataset/contrast combination for the B cell receptor signaling network. NB – this is just IS \* RA, need to multiply by std.dev as well.
4. Figure 4: Benchmarking of mBONITA. Numbers of differentially regulated KEGG pathways identified from combination multi-omics data by tested methods in three contrasts (A) 19%O2,CyA- vs 1%O2,CyA- (B) 1%O2,CyA+ vs 1%O2,CyA- (C ) 19%O2,CyA- vs 1%O2,CyA+ (D). Pathways known to be involved in the hypoxia-mediated response to CyA, Only pathways identified as significant from a combined dataset by at least one method are shown. Pathways are defined as differentially regulated if the Benjamini-Hochberg corrected p-value < 0.05.

## Supplementary Materials:

1. Supplementary Table 1 Experimental conditions in the three datasets from RAMOS B cells. Conditions that are in all datasets are highlighted in red.
2. Supplementary Figure 1: Transcriptomics analysis of RAMOS B cells grown under three conditions. (A) Numbers of differentially expressed genes identified by DESeq2 in all three contrasts (absolute log2-fold change > 0.5 and Bonferroni-adjusted p-value < 0.05) (B) z-scored RPM values of DE genes identified in all/any contrast. Experimental conditions are indicated by colors as shown in the legend. (C ) Over-representation analysis of DE genes in all three contrasts (unadjusted p-value < 0.05). Complete tables of DE genes and over-represented pathways may be found in the Supplementary Data.
3. Supplementary Table 2: KEGG Pathways involved in the HIF1A-mediated response of B cells to hypoxia and CyA
4. Supplementary Figure 2: Rule inference from all three datasets – (A) Rule set sizes, (B) Importance scores (Spearman correlations between 0.5 and 0.8, p << 0.01). See Supplementary Table 3 for all correlation coefficients.
5. Supplementary Table 3: Spearman correlation between importance scores
6. Supplementary Figure 3: Comparison of mBONITA-PA to BONITA-PA Numbers of differentially regulated pathways identified from combination multi-omics data by mBONITA in three contrasts (A) 19%O2,CyA- vs 1%O2,CyA- (B) 1%O2,CyA+ vs 1%O2,CyA- (C ) 19%O2,CyA- vs 1%O2,CyA+. Pathways are defined as differentially regulated if the Benjamini-Hochberg corrected p-value is < 0.05.
7. Supplementary Figure 4: Pathway analysis with Bonita. All p-values are Bonferroni-corrected and are < 0.01. The top 10 pathways with the lowest p-values are shown. A complete table of significantly modulated pathways may be found in the Supplementary Data. (a) Proteomics (top 10 pathways with the lowest p-values are shown) (b) Phosphoproteomics (top 10 pathways with the lowest p-values are shown) (c) Transcriptomics (top 4 pathways with the lowest p-values are shown) (d) TO BE ADDED. Multiomics network. Contrasts are color-coded as shown in the legend.
8. Supplementary File 1: mBONITA-PA results Excel workbook - pvalues\_concatenated\_20220816
9. Supplementary File 2: PaintOmics results (paintomics\_allResults.csv)
10. Supplementary File 3: leapR results (leapR\_allResults.csv)
11. Supplementary File 4: CAMERA results (camera\_allResults.csv)
12. Supplementary File 5: ActivePathways results (activePathways\_allResults.csv)
13. Supplementary File 6: BONITA results
14. Supplementary Figure 5: paintOMICS results
15. Supplementary Figure 6: leapR results
16. Supplementary Figure 7: ActivePathways
17. Supplementary Figure 8: CAMERA + Fisher results

# Article and Author Information

## Author affiliations

NB: Journal of Proteome Research requires the current affiliations of authors as well as the affiliation when the work was performed.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name | Email | Current affiliation | Past affiliation | ORCID |
| Xiaojun Min | xmin2@u.rochester.edu | Global Technology Analyst at Bank of America | Department of Computer Science and Department of Statistics at UofR | 0000-0002-0470-2558 |
| Alexander Crystal | acrystal@u.rochester.edu | Khoury College of Computer Sciences, Northeastern University | Goergen Institute for Data Science at UofR | 0000-0002-9414-5782 |
| Jiayue Meng | joyce.meng0417@outlook.com | - | Goergen Institute for Data Science, Department of Computer Science and Department of Mathematics at UofR | 0000-0002-0577-2566 |

## Authors' contributions

*Please see* [*https://casrai.org/credit/*](https://casrai.org/credit/) *for a list of possible author roles in the CRediT system.*

**MGP:** Conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing (original draft), writing (review & editing), visualization; **XM:** Formal analysis; **AC:** Formal analysis; **SH:** Resources, writing (review & editing); **MZ:** Resources, writing (review & editing); **JT:** Conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing (original draft), writing (review & editing), visualization, supervision, project administration, funding acquisition

## Competing interests

The authors declare that they have no competing interests.

## Funding

The study was supported by U.S. National Institutes of Health. MGP is supported by R01 AI134058. JT is supported by UM1 AI069511, P30 AI078498, R01 AI134058 and R21 AI136668. (*Insert funding information for other authors*)

## Acknowledgements

We would like to thank all past and present members of the Thakar Lab and the Zand Lab at the University of Rochester, and the Jun Qu Lab at State University of New York at Buffalo for helpful discussions. The Center for Integrated Research Computing at the University of Rochester provided high-performance computing resources.

# References

1. Ma, J.; Shojaie, A.; Michailidis, G., A comparative study of topology-based pathway enrichment analysis methods. *BMC Bioinformatics* **2019,** *20* (1), 546.

2. Ihnatova, I.; Popovici, V.; Budinska, E., A critical comparison of topology-based pathway analysis methods. *PLoS One* **2018,** *13* (1), e0191154.

3. Palshikar, M. G.; Palli, R.; Tyrell, A.; Maggirwar, S.; Schifitto, G.; Singh, M. V.; Thakar, J., Executable models of pathways built using single-cell RNA seq data reveal immune signaling dysregulations in people living with HIV and atherosclerosis. *medRxiv* **2022**, 2022.03.07.22271522.

4. Palli, R.; Palshikar, M. G.; Thakar, J., Executable pathway analysis using ensemble discrete-state modeling for large-scale data. *Plos Computational Biology* **2019,** *15* (9).

5. Hilchey, S. P.; Palshikar, M. G.; Emo, J. A.; Li, D.; Garigen, J.; Wang, J.; Mendelson, E. S.; Cipolla, V.; Thakar, J.; Zand, M. S., Cyclosporine a directly affects human and mouse b cell migration in vitro by disrupting a hIF-1 αdependent, o(2) sensing, molecular switch. *BMC Immunol* **2020,** *21* (1), 13.

6. Hilchey, S. P.; Palshikar, M. G.; Shen, S.; Rasam, S.; Mendelson, E. S.; Emo, J. A.; Thakar, J.; Qu, J.; Zand, M. S., LSP1 Attenuates Human B Cell Migration at Physiological Oxygen Levels, as Revealed by Phosphoproteomics Analysis. *In preparation* **2022**.

7. Kanehisa, M.; Furumichi, M.; Sato, Y.; Ishiguro-Watanabe, M.; Tanabe, M., KEGG: integrating viruses and cellular organisms. *Nucleic acids research* **2021,** *49* (D1), D545-D551.

8. Martens, M.; Ammar, A.; Riutta, A.; Waagmeester, A.; Slenter, D. N.; Hanspers, K.; R, A. M.; Digles, D.; Lopes, E. N.; Ehrhart, F.; Dupuis, L. J.; Winckers, L. A.; Coort, S. L.; Willighagen, E. L.; Evelo, C. T.; Pico, A. R.; Kutmon, M., WikiPathways: connecting communities. *Nucleic acids research* **2021,** *49* (D1), D613-D621.

9. Palshikar, M. G.; Hilchey, S. P.; Zand, M. S.; Thakar, J., WikiNetworks: translating manually created biological pathways for topological analysis. *Bioinformatics* **2021**, btab699.

10. Dugourd, A.; Kuppe, C.; Sciacovelli, M.; Gjerga, E.; Gabor, A.; Emdal, K. B.; Vieira, V.; Bekker-Jensen, D. B.; Kranz, J.; Bindels, E. M. J.; Costa, A. S. H.; Sousa, A.; Beltrao, P.; Rocha, M.; Olsen, J. V.; Frezza, C.; Kramann, R.; Saez-Rodriguez, J., Causal integration of multi-omics data with prior knowledge to generate mechanistic hypotheses. *Mol Syst Biol* **2021,** *17* (1), e9730.

11. Hernández-de-Diego, R.; Tarazona, S.; Martínez-Mira, C.; Balzano-Nogueira, L.; Furió-Tarí, P.; Pappas, G. J., Jr.; Conesa, A., PaintOmics 3: a web resource for the pathway analysis and visualization of multi-omics data. *Nucleic acids research* **2018,** *46* (W1), W503-W509.

12. Liu, T.; Salguero, P.; Petek, M.; Martinez-Mira, C.; Balzano-Nogueira, L.; Ramšak, Ž.; McIntyre, L.; Gruden, K.; Tarazona, S.; Conesa, A., PaintOmics 4: new tools for the integrative analysis of multi-omics datasets supported by multiple pathway databases. *Nucleic acids research* **2022,** *50* (W1), W551-W559.

13. Wu, D.; Smyth, G. K., Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic acids research* **2012,** *40* (17), e133.

14. Griss, J.; Viteri, G.; Sidiropoulos, K.; Nguyen, V.; Fabregat, A.; Hermjakob, H., ReactomeGSA - Efficient Multi-Omics Comparative Pathway Analysis. *Mol Cell Proteomics* **2020,** *19* (12), 2115-2125.

15. Danna, V.; Mitchell, H.; Anderson, L.; Godinez, I.; Gosline, S. J. C.; Teeguarden, J.; McDermott, J. E., leapR: An R Package for Multiomic Pathway Analysis. *Journal of Proteome Research* **2021,** *20* (4), 2116-2121.

16. Canzler, S.; Hackermüller, J., multiGSEA: a GSEA-based pathway enrichment analysis for multi-omics data. *BMC Bioinformatics* **2020,** *21* (1), 561.

17. Paczkowska, M.; Barenboim, J.; Sintupisut, N.; Fox, N. S.; Zhu, H.; Abd-Rabbo, D.; Mee, M. W.; Boutros, P. C.; Abascal, F.; Amin, S. B.; Bader, G. D.; Beroukhim, R.; Bertl, J.; Boroevich, K. A.; Brunak, S.; Campbell, P. J.; Carlevaro-Fita, J.; Chakravarty, D.; Chan, C. W. Y.; Chen, K.; Choi, J. K.; Deu-Pons, J.; Dhingra, P.; Diamanti, K.; Feuerbach, L.; Fink, J. L.; Fonseca, N. A.; Frigola, J.; Gambacorti-Passerini, C.; Garsed, D. W.; Gerstein, M.; Getz, G.; Gonzalez-Perez, A.; Guo, Q.; Gut, I. G.; Haan, D.; Hamilton, M. P.; Haradhvala, N. J.; Harmanci, A. O.; Helmy, M.; Herrmann, C.; Hess, J. M.; Hobolth, A.; Hodzic, E.; Hong, C.; Hornshøj, H.; Isaev, K.; Izarzugaza, J. M. G.; Johnson, R.; Johnson, T. A.; Juul, M.; Juul, R. I.; Kahles, A.; Kahraman, A.; Kellis, M.; Khurana, E.; Kim, J.; Kim, J. K.; Kim, Y.; Komorowski, J.; Korbel, J. O.; Kumar, S.; Lanzós, A.; Lawrence, M. S.; Lee, D.; Lehmann, K.-V.; Li, S.; Li, X.; Lin, Z.; Liu, E. M.; Lochovsky, L.; Lou, S.; Madsen, T.; Marchal, K.; Martincorena, I.; Martinez-Fundichely, A.; Maruvka, Y. E.; McGillivray, P. D.; Meyerson, W.; Muiños, F.; Mularoni, L.; Nakagawa, H.; Nielsen, M. M.; Park, K.; Park, K.; Pedersen, J. S.; Pich, O.; Pons, T.; Pulido-Tamayo, S.; Raphael, B. J.; Reyes-Salazar, I.; Reyna, M. A.; Rheinbay, E.; Rubin, M. A.; Rubio-Perez, C.; Sabarinathan, R.; Sahinalp, S. C.; Saksena, G.; Salichos, L.; Sander, C.; Schumacher, S. E.; Shackleton, M.; Shapira, O.; Shen, C.; Shrestha, R.; Shuai, S.; Sidiropoulos, N.; Sieverling, L.; Sinnott-Armstrong, N.; Stein, L. D.; Stuart, J. M.; Tamborero, D.; Tiao, G.; Tsunoda, T.; Umer, H. M.; Uusküla-Reimand, L.; Valencia, A.; Vazquez, M.; Verbeke, L. P. C.; Wadelius, C.; Wadi, L.; Wang, J.; Warrell, J.; Waszak, S. M.; Weischenfeldt, J.; Wheeler, D. A.; Wu, G.; Yu, J.; Zhang, J.; Zhang, X.; Zhang, Y.; Zhao, Z.; Zou, L.; von Mering, C.; Reimand, J.; Drivers, P.; Functional Interpretation Working, G.; Consortium, P., Integrative pathway enrichment analysis of multivariate omics data. *Nature Communications* **2020,** *11* (1), 735.

18. Palli, R.; Palshikar, M. G.; Thakar, J., Executable pathway analysis using ensemble discrete-state modeling for large-scale data. *PLoS computational biology* **2019,** *15* (9), e1007317.

19. Palshikar, M. G.; Min, X.; Crystal, A.; Meng, J.; Hilchey, S. P.; Zand, M.; Thakar, J., Executable network models of integrated multi-omics data. *In preparation* **2022**.

20. Slenter, D. N.; Kutmon, M.; Hanspers, K.; Riutta, A.; Windsor, J.; Nunes, N.; Melius, J.; Cirillo, E.; Coort, S. L.; Digles, D.; Ehrhart, F.; Giesbertz, P.; Kalafati, M.; Martens, M.; Miller, R.; Nishida, K.; Rieswijk, L.; Waagmeester, A.; Eijssen, L. M. T.; Evelo, C. T.; Pico, A. R.; Willighagen, E. L., WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic acids research* **2018,** *46* (D1), D661-D667.

21. Hilchey, S. P.; Palshikar, M. G.; Li, D.; Garigen, J.; Cipolla, V.; Thakar, J.; Zand, M. S., Cyclosporine A Directly Affects Human and Mouse B cell Migration <em>in vitro</em> by Disrupting a HIF-1<em>α</em> Dependent, O<sub>2</sub> Sensing, Molecular Switch. *bioRxiv* **2019**, 622514.

22. Hilchey, S. P.; Palshikar, M. G.; Mendelson, E. S.; Shen, S.; Rasam, S.; Emo, J. A.; Qu, J.; Thakar, J.; Zand, M. S., Cyclosporine A Modulates LSP1 Protein Levels in Human B Cells to Attenuate B Cell Migration at Low O2 Levels. *Life* **2022,** *12* (8), 1284.

23. Love, M. I.; Huber, W.; Anders, S., Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **2014,** *15* (12), 550.

24. Stephens, M., False discovery rates: a new deal. *Biostatistics* **2017,** *18* (2), 275-294.

25. Gu, Z.; Eils, R.; Schlesner, M., Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **2016,** *32* (18), 2847-9.

26. Wu, T.; Hu, E.; Xu, S.; Chen, M.; Guo, P.; Dai, Z.; Feng, T.; Zhou, L.; Tang, W.; Zhan, L.; Fu, X.; Liu, S.; Bo, X.; Yu, G., clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (N Y)* **2021,** *2* (3), 100141.

27. Liberzon, A.; Subramanian, A.; Pinchback, R.; Thorvaldsdottir, H.; Tamayo, P.; Mesirov, J. P., Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **2011,** *27* (12), 1739-40.

# Supplementary Data

## Transcriptomics data suggests mechanisms of HIF1A-mediated chemotaxis disrupted by CyA

Differentially expressed (DE) genes and over-represented KEGG gene sets were identified as described in the Methods. The 24 genes that were DE between the `19% O2, CyA-` and `1% O2, CyA-` conditions (i.e., the DE genes driven solely by the response to hypoxia) were enriched for gene sets related to glucose metabolism, as expected (Figure 1 A-C). These genes included LSP1, EGLN1, BNIP3/BNIP3L, and ALDOC, which were upregulated in cells grown at 1% oxygen. All these genes are well-known regulators of cellular responses to cells to hypoxia (*many citations here*). BNIP3/BNIP3L and EGLN1 are direct downstream targets of HIF1A. While the over-representation analysis of the 49 genes that were DE between the `1% O2, CyA+` and `1% O2, CyA-` conditions (i.e., the DE genes driven solely by the response to CyA under physiological hypoxic conditions) did not identify gene sets directly linked to either hypoxia response or immunosuppression (Figure 1C).

VEGFA , PHGDH (<https://aacrjournals.org/cancerres/article/77/22/6321/622947/PHGDH-as-a-Key-Enzyme-for-Serine-Biosynthesis-in>, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5286373/), PSAT1 (regulated by ATF4/5)

ADA - <https://pubmed.ncbi.nlm.nih.gov/16670267/>

PIM2 - <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0088301>, <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7783746/>

However, these DE genes included ATF4/5, SESN2 (HIF1A-induced, <http://www.aginganddisease.org/EN/10.14336/AD.2019.0320>),

## Multi-omics datasets showed low inter-dataset consistency

We profiled expression of mRNA, proteins, and phosphorylated proteins from RAMOS B cells grown under different oxygenation conditions and treatment with CyA and CXCL12, and retained only those conditions that were profiled in all three omics datasets for downstream analysis (Table 1). After processing and filtering the datasets as described in the Methods, we found that only 1505 genes were profiled in all three datasets (Figure 1A). The transcriptomic and proteomic data for these 1505 genes were moderately well correlated (Spearman correlation = 0.58, p-value < 0.01). However, the phosphoproteomics data showed low correlation to both the transcriptomics and proteomics data (Spearman correlation = 0.15 and 0.17 respectively, p-value < 0.01) (Figure 1B). We next attempted to find whether the intra-dataset correlations were consistent between datasets, i.e., whether the same signaling processes could be identified in all three datasets. We constructed a correlation network for each dataset as described in the Methods.

The overlap between genes that were highly correlated between pairs of datasets was similarly low (Figure 2). We constructed a network comprised of all KEGG networks, performed network propagation using TIEDIE, and identified modules of genes that were significantly differentially expressed between pairs of conditions. A representative comparison between B cells grown at 1% oxygen with CyA and without CyA is shown in Figure 1.