Using network inference to discover molecular pathways underlying cytokine synergism and age-related neurodegeneration

by

Bryce Hwang

S.B., M.I.T. (2018)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Computer Science and Molecular Biology

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2018

Author .....
Department of Electrical Engineering and Computer Science
May 25, 2018

Certified by .....
Ernest Fraenkel
Professor
Thesis Supervisor

Accepted by .....
Katrina LaCurts
Chair, Master of Engineering Thesis Committee

Using network inference to discover molecular pathways underlying cytokine synergism and age-related neurodegeneration

by

Bryce Hwang

Submitted to the Department of Electrical Engineering and Computer Science
on May 25, 2018, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Computer Science and Molecular Biology

Abstract

New high-throughput "omic" methods can help shed light on molecular pathways underpinning diseases ranging from cancers to neurodegenerative disorders. However, effectively integrating information across these diverse data types is challenging. Network modeling approaches can help bridge this gap. In particular, the Prize-Collecting Steiner Forest approach (PCSF) is a network modeling method that provides high-confidence subnetworks of physically interacting molecules by integrating diverse "omics" data with prior knowledge from protein-protein interaction networks (PPIs). However, PCSF is sensitive to initial parameterization and generating biological hypotheses from the resulting subnetworks can often be difficult. This study increases the interpretability of subnetwork solutions generated PCSF by studying the effect of varying PCSF free parameters and adding annotations for subcellular localization. The PCSF approach is then used to elucidate pathways underlying synergy between cytokines, pro-inflammatory molecules that mediate diverse biological phenomena ranging from anti-viral immunity to autoimmune disorders like inflammatory bowel disease (IBD). In addition, PCSF approach is applied in a cross-species context to integrate information from Drosophila models for neurodegeneration and human Alzheimer's Disease (AD) patients to investigate proximal conserved mechanisms of age-related neurodegeneration.

Thesis Supervisor: Ernest Fraenkel
Title: Professor

Acknowledgments

First and foremost, I would like to thank Professor Fraenkel for his unfailing guidance over the past two years. I am immensely grateful for how he has challenged and inspired me to become a much better scientist.

Thank you also to my numerous collaborators, without whom this project could not have happened. For the cytokine project, I would especially like to thank Kenneth Nally at the University of Cork for coordinating the project and providing invaluable insights. Thank you also to Jerzy Woznicki, who performed many difficult experiments to acquire much of the data used in this analysis, and to Paolo, Armel, Francsico, and Cristina for their work on the genetic screens and proteomics data. For the Alzheimer's project, thank you to Mel Feany and her lab for the *Drosophila* data and Clemens Scherzer, Xianjung Dong, Zhixiang, and the rest of the Scherzer lab for the human data. I would also like to thank Mel not just for her tireless assistance with the project, but also for her insights into a career combining medicine and basic science.

To all the wonderful friends in the Fraenkel lab: Natasha, Tobi, Alex, Brook, Miriam, Johnny, Divya, and Max - I can't thank you all enough for your constant support both in research and in life. Shout out especially to our cluster for the countless interesting discussions about both science and beyond. And of course, thank you to all my friends at MIT and beyond for always being there to lend an ear or a hand.

Finally, my deepest gratitude to my parents, to my brother Brendan, and to Genna for their endless love and support - I couldn't have done any of this without you all.

Author

Bryce Hwang

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The past two decades have seen an explosion of "omics", techniques that assess a global set of molecules **(Table 1.1)**. However, while each of these data modalities can provide some insight into disease etiology, single datasets are inherently limited to correlative, but not causal analyses. Therefore, it is critically important to develop approaches to integrate information across a variety of "omics" datasets in order to develop causal models of disease mechanism.

Networks methods provide a powerful technique for integrating together a variety of "omics" datasets. In essence, these techniques treat interactions between molecules in a cell as a graph. In the most general models, the molecular players in a cell (i.e. proteins, mRNA transcripts, metabolites) can be modeled as nodes in a graph, and the interactions between these cells (i.e. between protein and metabolite or between transcription factor and gene regulated) can be modeled as edges. However, with tens of thousands of nodes and hundreds of thousands of edges the graph of these interactions can quickly grow into intractable "hairballs" that are exceedingly difficult to interpret. To solve this problem, network inference approaches apply well-known graph theory problems like multi-commodity flow and minimum cost flow to biological problems [1, 2]. These approaches leverage previous knowledge from protein-protein interaction networks (PPIs) to reveal physical pathways linking "omics" data through known and novel pathways. Resulting networks can then clustered to discover groups of molecules that function coherently, including molecules not originally detected in the input data.

Type of "omics" data	**Molecules profiled**	**Associated technologies**	**Example use cases**
Genomics	DNA	DNA sequencing	Identify genetic variants associated with disease through genome wide association studies
Epigenomics	DNA and histones	ATAC-seq ChIP-seq Bisulfite-seq	Identify changes in DNA methylation and histone modification that regulate gene expression
Transcriptomics	RNA	RNA-seq Microarrays	Quantify expression of protein-encoding genes and identification of mRNA splice variants
Proteomics	Proteins	Mass spectrometry (MS)	Quantify relative protein expression and covalent modifications to proteins relevant to cell signaling and protein regulation.
Metabolomics	Small molecules	GC-MS HPLC	Quantify small molecules that reflect underlying metabolic function and disrupted enzymatic pathways

Table 1.1: Description of different "omics" data.

Thus, these approaches can identify novel pathways not detected in standard pathway analyses by integrating previous knowledge from high-throughput experiments. This paradigm can be applied to a wide range of big-data biological questions, ranging from discovering signaling pathways underlying cancers to elucidating conserved mechanisms of neurodegeneration [3, 4].

In particular, an approach based on solving the Prize-Collecting Steiner Forest problem (PCSF) can extract biologically relevant pathways from multi-omics data by generating high-confidence subnetworks of physically interacting molecules [5, 6, 7].[1] This method avoids over-reliance on hub proteins, takes into account the reliability of interactions in the starting PPI, and determines the robustness of the each node based on uncertainty in the data and the network. Moreover, it has been implemented for public use as the OmicsIntegrator package [6]. However, PCSF is highly sensitive to the initial parameters chosen. Moreover, it can often be tedious to formulate biological hypotheses based on the outputted subnetworks.

In this study, I aim to investigate and make improvements to the PCSF approach, then use this technique to approach two different biological problems: discovering pathways relevant to synergism between pro-inflammatory molecules implicated in inflammatory bowel disease (IBD) and uncovering hidden genes and pathways related to the age-related neurodegeneration of Alzheimer's disease.

Chapter one will introduce the PCSF approach, discuss problems with the PCSF approach, and introduce biological problems that have been solved by extensions of the PCSF approach.

Chapter two will focus on the investigation of free parameters in the PCSF problem. It will also discuss some improvements to the output of a popular package that solves the PCSF problem for biological networks, OmicsIntegrator [6].

Chapter three will discuss a multi-omics study that uncovered molecular pathways underlying non-additive effects when treating human adenocarcinoma cells with

[1]For convenience, in this study, "PCSF" will refer to both the graph theory problem and to the approach based on solving the Prize-Collecting Steiner Forest problem to extract biologically relevant pathways.

two pro-inflammatory molecules, TNFα and IFNγ.[2] In particular, it highlights the flexibility of network inference approaches like PCSF in uncovering hidden pathways using epigenetic, transcriptomic, proteomic, phosphoproteomic, and genetic screen data.

Chapter four will discuss an application of the PCSF approach in a cross-species context to integrate information from *Drosophila* metabolomics, proteomics, and genetic screens with human RNA-seq data to investigate proximal conserved mechanisms of neurodegeneration underlying Alzheimer's Disease (AD).[3]

1.1 Summary of PCSF approach

The PCSF problem is formally defined on an undirected graph $G = (V, E)$. Nodes are labeled with prizes, p and edges are labeled with non-negative costs, c **(Figure 1-1a)**. $G = (V, E)$ is then transformed to the graph $H = (V', E')$ as follows:

1. Copy all nodes and edges in G into H.
2. Add a dummy node, v_d to H
3. For all $v \in V$ where $p_v > 0$, add an edge with cost ω between v and the dummy node v_d.
4. For all $e \in E$, rescale the edge cost to $c'_e = c_e + \gamma d(v_i) * d(v_j)$, where $d(v_i)$ and $d(v_j)$ are the degrees of the two nodes connected by edge e and γ is a parameter penalizing nodes with many neighbors **(Figure 1-1b)**.

The goal is then to identify a connected subgraph of H, $H' = (V'', E'')$, that maximizes the objective function:

$$\beta \sum_{v \in V''} p_v - \sum_{e \in E''} c_e \tag{1.1}$$

[2]TNFα and IFNγ are both members of a more general class of pro-inflammatory molecules known as cytokines. In this study, "cytokine synergy" and "cytokine synergism" will specifically refer to synergy between signaling with TNFα and IFNγ.

[3]In this study, *Drosophila* will specifically refer to *Drosophila melanogaster*, the common fruit fly often used as a model organism.

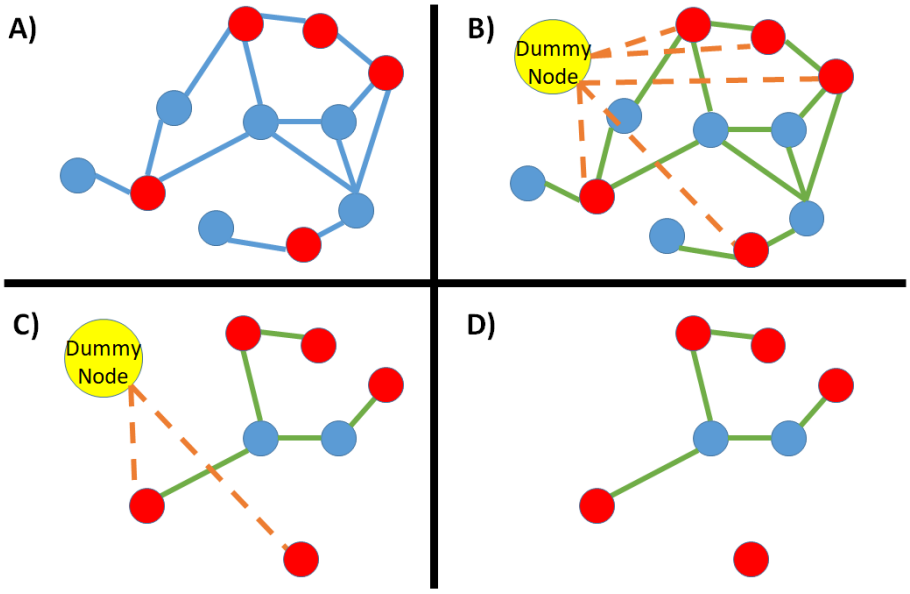

Figure 1-1: Overview of PCSF approach. **A)** Graph, $G = (V, E)$, with the circles as nodes, V, and blue lines as edges, E. Terminals are nodes with $p > 0$ and are shown in red. **B)** Transformed graph, $H = (V', E')$. The yellow dummy node is connected to terminals with the dotted edges with weight ω. The edges are rescaled with $c'_e = c_e + \gamma d(v_i) * d(v_j)$ and are now shown in green. **C)** Tree solution to PCST that maximizes the objective above. The dummy node is still attached. **D)** The final subnetwork consists of two connected components once the dummy node and its adjacent edges are removed.

where β is a scaling for the prizes. The dummy node, v_d and all edges adjacent to the dummy node are then removed to produce a final subnetwork **(Figure 1-1c, 1-1d)**. In the resulting subgraph $H' = (V'', E'')$, the subset of nodes with positive-weight prizes are "terminals" and the subset of nodes with zero-weight prizes are termed "Steiner" nodes [6].

This approach can be used to study biological networks by treating a PPI such as iRefWeb as a graph, G [8]. Nodes, V, represent proteins and edges, E, represent interactions between proteins. Edge costs, c, representing the inverse confidence of the interaction. The prizes, p, can be assigned to the nodes of interest based on proteomic, phosphoproteomic, transcriptomic, epigenomic, or other biological data. Solving the PCSF problem is APX-hard, meaning that finding an exact solution is NP-hard, but approximate solutions can be found in polynomial-time [9]. Multiple algorithms have been implemented to approximate the PCSF problem; the most recent iteration is a fast heuristic approximation, OI2, based on the graph-structured sparsity approach described by Hegde, Indyk, and Schmidt [7, 6, 10]. However, there have been no comprehensive studies on how the parameters γ, β, and ω influences the ability of the PCSF approach to uncover true biologically relevant pathways.

1.2 Applying PCSF framework to novel biological problems

The PCSF framework has been applied to a variety of biological problems. One previous work used the frequency of genetic events (SNPs, indels, and CNVs) in genes as prizes for different subtypes for glioblastoma, selecting highly mutated genes as prizes. The PCSF subnetworks were then used to infer molecular pathways that drive different subtypes of glioblastoma [11]. Another approach used a Bayesian approach to assign m/z peaks to metabolites. Edges were then inferred between metabolites and a protein-protein interaction network. The resulting approach integrated proteomic and metabolomic data to probe how the metabolism of various lipids were dysregulated in

a Huntington disease model [5]. Another extension of the PCSF approach has been used to integrate data from yeast and humans in order to discover genes underlying Parkinson's disease [4]. Yet another extension of the PCSF approach has been used to impute signaling pathways shared across individual patients [7].

However, to date the PCSF framework has been applied a disease/control conditions to reconstruct single perturbations of cellular pathways. However, the PCSF approach has not yet been applied to reconstruct molecular pathways underlying non-additive effects between distinct signaling pathways. Moreover, while the PCSF approach was successfully extended in integrating yeast and human data, this study was heavily reliant on the injection of edges from the yeast interactome to the human interactome. In this project, I extended the PCSF approach to reconstruct non-additive molecular pathways between two distinct pro-inflammatory molecules relevant to inflammatory bowel disease and cancer. Next, I extend the PCSF to a cross-species context that is not reliant on the injection of additional edges. Instead, I directly mapped *Drosophila* proteins, genes, and metabolites to their human homologs, then integrated these data with human RNA-seq data to infer molecular pathways underlying conserved pathways of neurodegeneration in Alzheimer's disease.

Chapter 2

Improving PCSF parameter selection and subnetwork interpretability

The PCSF approach outlined in chapter one is a flexible approach for inferring biologically relevant subnetworks using multi-omic data coupled with high confidence protein-protein interaction networks. However, the PCSF solution generated using a popular implementation, OmicsIntegrator2 (OI2), is sensitive to the choice of free parameters, β, ω, γ, as well as the distribution of prize weights. Since β, ω, and γ are not linearly separable, it is difficult to tune these parameters sequentially. For example, **Figures 2-1a** and **2-1b** are both derived from the same set of prizes. While figure 2-1a does capture a large portion of the input data, it also selects many intermediate nodes and pathways, making biological interpretation exceedingly difficult.[1] By contrast, figure 2-1b selects only a small portion of the input nodes; however, the selected intermediate nodes are poorly connected to proteins identified from our biological evidence. In both these cases, poor choices of PCSF free parameters adversely affect the formulation of biological hypotheses about pathways that explain the data.

[1] "Intermediate" nodes and "Steiner" nodes will be used interchangeably in this study.

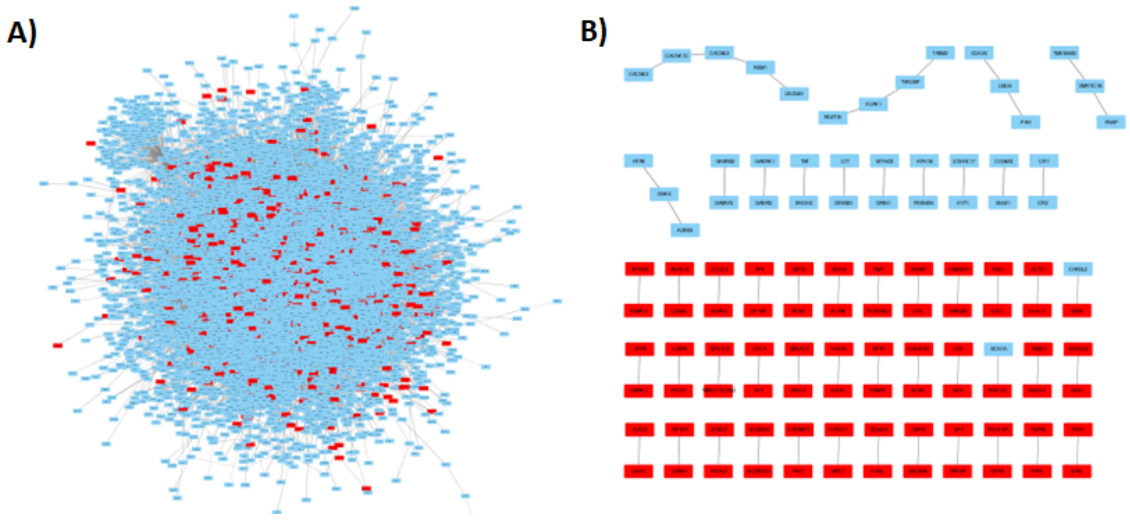

Figure 2-1: Two examples of PCSF subnetworks generated from the same prize set (based on a list of genes associated with Alzheimer's Disease). Red nodes signify terminal nodes ($p_v > 0$) and blue nodes signify Steiner nodes ($p_v = 0$). **(A)** is a "hairball" subnetwork constructed using high values for β and ω and a low value for γ that captures a large portion of the input data, but that also selects a variety of intermediate nodes. **(B)** is a disjoint subnetwork constructed using low values for β and ω and a high value for γ that does a poor job of capturing interactions between input data and intermediate nodes.

2.1 Previous approaches to PCSF parameterization

Previous approaches to PCSF parameter selection have focused on minimizing the difference between various aspects of the degree distribution of Steiner nodes and terminal nodes, and exploring a "representative" distribution of parameter space **(Figure 2-2a,b)**. The weakness of the former approach is that it assumes that Steiner nodes will possess similar property to terminal nodes. This can often be misleading when considering the underlying biology. For example, in many disease conditions, perturbation of master regulators lead to severe phenotypes that can lead to cell death. Single datasets such as proteomics or transcriptomics can often miss these master regulators. However, the PCSF approach *should* recover these proteins; however, making the assumption that the degree distribution of Steiner and terminal nodes should be similar will bias networks chosen against finding these master regulators.

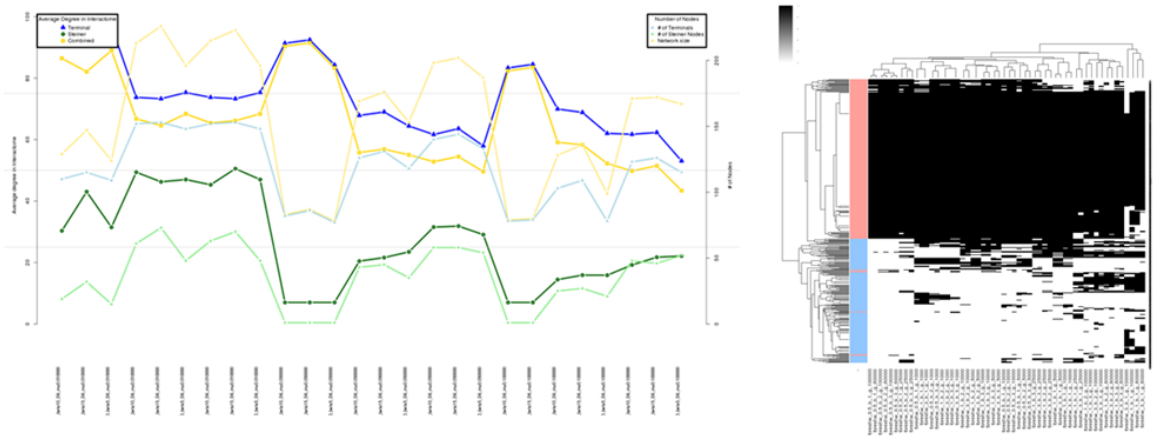

Figure 2-2: Plots used for PCSF parameter selection. **(A)** This plot of Steiner and terminal node degree distribution is used to qualitatively pick parameters. Sets of parameters are on the x-axis and the number of nodes is on the y-axis. Darker lines indicate cardinality, and lighter lines indicate average degree. Blue represents terminal nodes, green represents steiner nodes, and yellow represents all nodes. **(B)** Heatmap showing the presence or absence of nodes (rows) in subnetworks across different parameters (columns). Both parameters and nodes are clustered by hierarchical clustering. Parameter sets which include a "reasonable" subset of terminals and "sufficient" numbers of steiner nodes are chosen for further analysis.

Similarly, attempting to find sets of parameters that appear to capture a "reasonable" number of Steiner nodes while retaining the majority of terminal nodes is a highly biased endeavor. Moreover, both these approaches have only been applied to datasets in which the underlying true positives are unknown, making quantitative claims about the accuracy of these approaches difficult. These approaches could easily be picking up many false positive terminal nodes and excluding true positive Steiner nodes.

2.2 Using synthetic datasets to assess PCSF parameterization

To evaluate the effect of changing PCSF, one hundred synthetic datasets were constructed to systematically assess the role of PCSF free parameters on PCSF performance (dataset construction is detailed in **Appendix B**). In brief, previously annotated genetic confidence scores from the OpenTargets database for ten diseases were used as prize weights **(Figure 2-3 and Figure A-1)** [12]. For each disease's

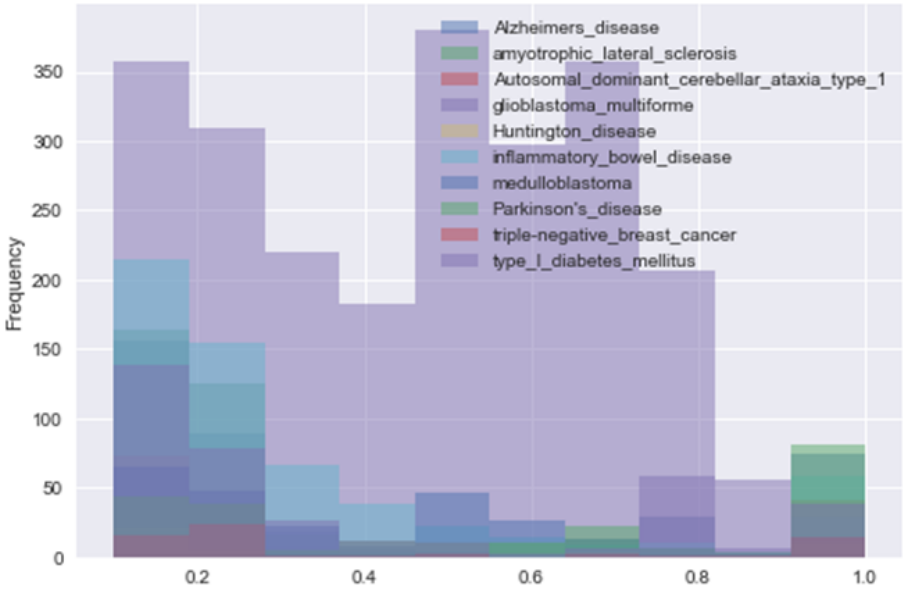

Figure 2-3: Overlaid histograms of the distribution of genetic confidence scores from OpenTargets for each disease used in creating synthetic datasets. Genetic confidence scores of less than 0.1 were excluded in creating the prizes for synthetic datasets and are excluded in this graph.

genes, 100 genes were sampled from the prize distribution and 100 interactome degree-matched and prize weight matched nodes were added as noise. For each synthetic dataset, the PCSF algorithm was run varying parameters β, ω, γ independently. For each parameter set, 100 prize randomizations were run to assess specificity, and 100 noise-edge randomizations were performed to assess sensitivity. Next precision, recall, and AUC were calculated between the consensus subnetwork solution against the reference set of true genetic associations for each disease using a variety of robustness and specificity thresholds.

These results suggest that precision and recall alone are not good ways of assessing PCSF performance. High precision networks generally included very few nodes, while high recall nodes included large amounts of noise **(Figure 2-4a,b)**. Taking the area under the curve (AUC) of the receiver-operating curve (ROC) similarly selected for many small networks, suggesting a strong bias for high precision networks that are generally not permissive enough to allow for the generation of novel biological hypotheses **(Figure 2-4c)**. The best metric seemed to the Jaccard index, which

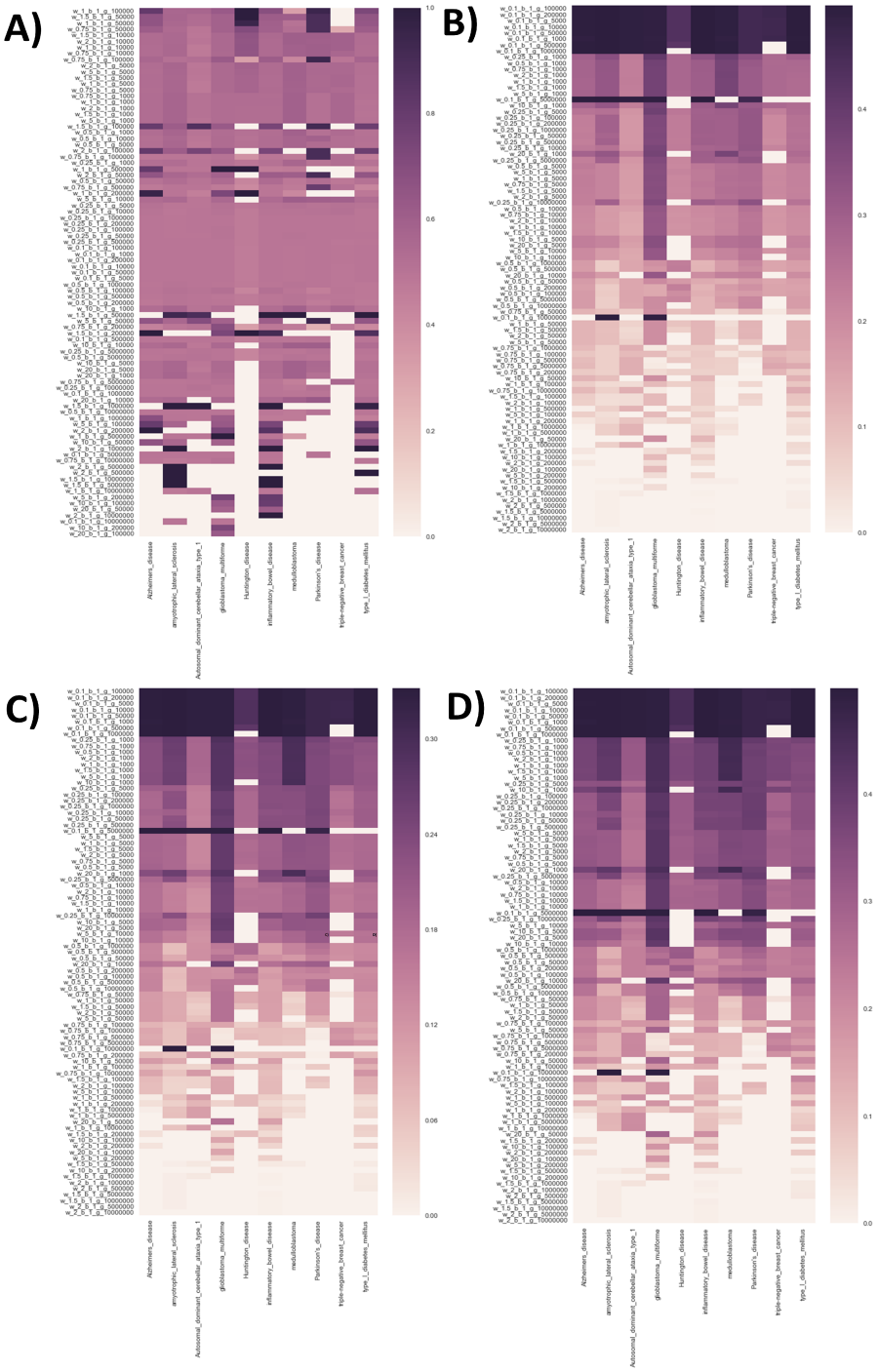

Figure 2-4: Subnetworks constructed with a variety of free parameters for PCSF using synthetic datasets described in Appendix B. Precision, recall, AUC and Jaccard scores were calculated for each subnetwork between the genetics hits for each disease from OpenTargets and the nodes present in the inferred subnetworks. Parameter sets were then ordered by their average score for each calculated value. For each parameter set (x-axis), the calculated (color) is plotted for each parameter set (y-axis). The metrics calculated are **(A)** precision **(B)** recall **(C)** AUC, **(D)** Jaccard score.

measures the ratio of the cardinality of the intersection over the union of the two sets. The Jaccard index generally selected for meaningful biological networks. In practice, network robustness thresholds of between 0.6 and 0.8 yielded the best results **(Figure A-2)**. Overall, this analysis suggested that different parameter sets perform differently for different prize distributions. Therefore, the best way of selecting a good parameter set for any given prize distribution is to construct a set of degree-matched synthetic datasets with noise in a similar fashion to the datasets described above. Next, one should vary the PCSF free parameters over a large range, run a PCSF solver once, and calculate the average Jaccard index between the nodes in the inferred subnetwork and known true positives.[2] Finally, one should run 100 randomizations for both sensitivity and specificity for a smaller set of parameters on the synthetic datasets, then calculate Jaccard scores between inferred subnetworks and known true positives. The parameter sets with the top average Jaccard scores should then be used as the PCSF free parameters. This approach provides a more unbiased approach to parameter selection for PCSF than current heuristics.

2.3 Adding subcellular annotations to improve PCSF subnetwork interpretability

Even after choosing a reasonable set of parameters, networks can still often be difficult to interpret. This prevents users of the PCSF approach from rapid forming of biological hypotheses for further validation. Therefore, in order to help users better understand networks, I helped automate the annotation of subcellular localization in PCSF outputs. Starting with the COMPARTMENTS database, a directed graph weighting scheme was used to bin proteins into their most probable subcellular localization **(Figure 2-5)** [13]. These annotations were then included into a web application for visualization.[3] As evidenced in the contrast between the networks shown in **Figure**

[2]Recommended parameters to choose for the OmicsIntegrator2 implementation are: ω between 0.1 to 10, β between 0.1 and 10, γ between 10 to 10^8.

[3]This visualization application developed by a member of the Fraenkel lab works also as a stand alone application. It can be accessed at interactome.info.

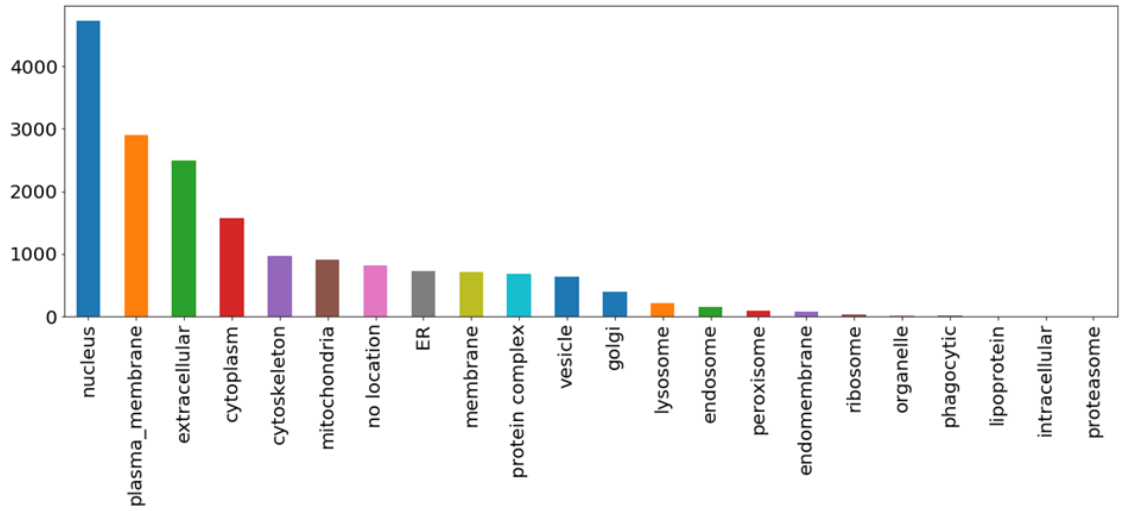

Figure 2-5: Evidence for subcellular location was collected using the knowledge and experiments in the COMPARTMENTS database, which aggregates evidence for subcellular locations based on expert curation of the literature and antibody-tagging experiments. A directed graph evidence-weighting scheme was then used to bin proteins in their most probable subcellular localization. A barchart of the number of proteins predicted to localize to each subcellular compartment is then plotted here.

2-6a and **Figure 2-6b**, adding subcellular annotations greatly increasing the ease of generating biological hypotheses using the PCSF approach.

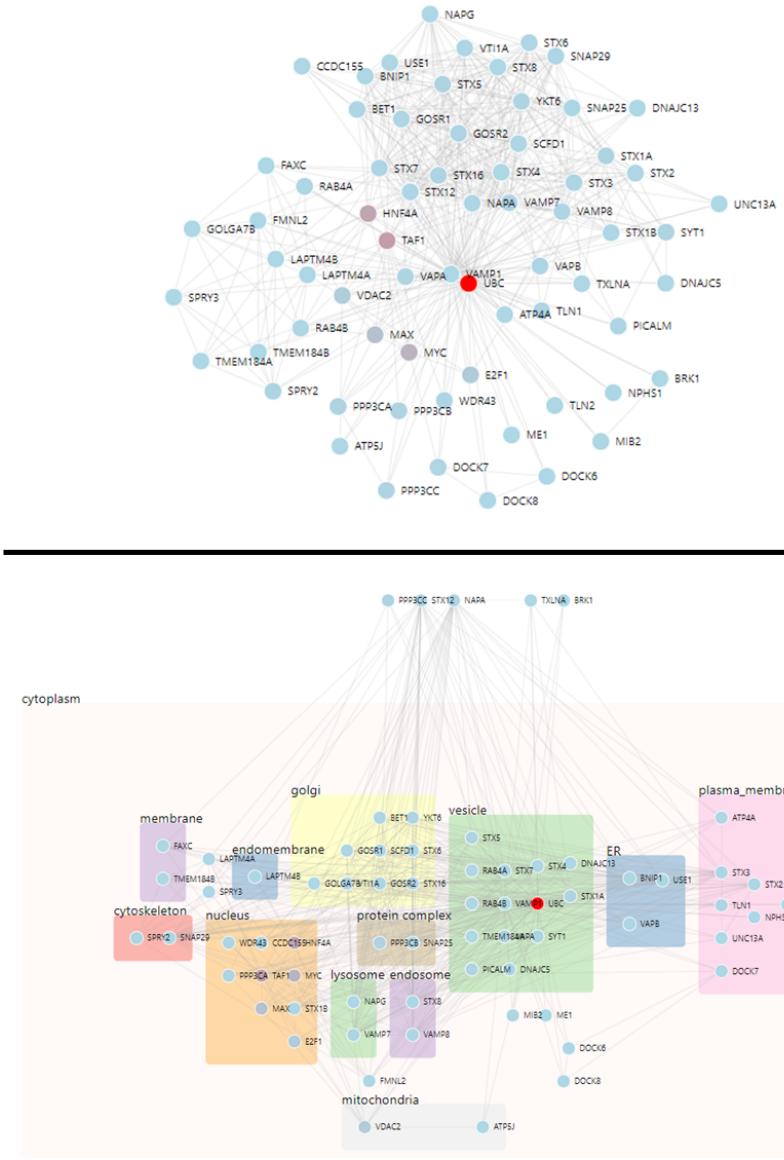

Figure 2-6: All proteins that were reported to interact with Vesicle Associated Membrane Protein (VAMP1) in the iRef database were plotted using interactome.info [8]. Edges between these neighbors of VAMP1 are retained. **(Top)** The resultant network is plotted using force-directed graph visualization. **(Bottom)** The same network is clustered according to subcellular locations as predicted in Figure 2-5.

Chapter 3

Using PCSF to elucidate pathways underlying synergism between TNFα and IFNγ

Cytokines are extracellular molecular regulators that mediate immune cell recruitment and complex intracellular signaling control mechanisms underlying inflammation [14]. For example, interferon gamma (IFNγ) and tumor necrosis factor alpha (TNFα) are both cytokines that play diverse functions in inflammation and immunity [14]. IFNγ is produced by both adaptive immune cells like $CD4^+$ and $CD8^+$ T-cells and innate immune cells like natural killer cells. This pathway has both antiviral activity, as well as immunomodulatory functions, such as promotion of T_{reg} cell differentiation and macrophage priming [15]. The TNFα pathway plays a distinct but complementary role to the IFNγ pathway. TNFα is also produced by macrophages and has context dependent anti-viral and anti-tumoral effects. This pathway also possess pleiotropic functions in homeostasis and immunopathogenicity, opposes cell proliferation, and signals for cell death [16]. Unsurprisingly, mixtures of cytokines have been shown to exhibit non-additive, "synergistic" responses that are more nuanced than the simple summation of single cytokine responses [17].[1] In particular, joint signaling with both

[1] In the rest of this study "synergy" and "joint signaling" or "joint treatment" will specifically describe synergy between IFNγ and TNFα

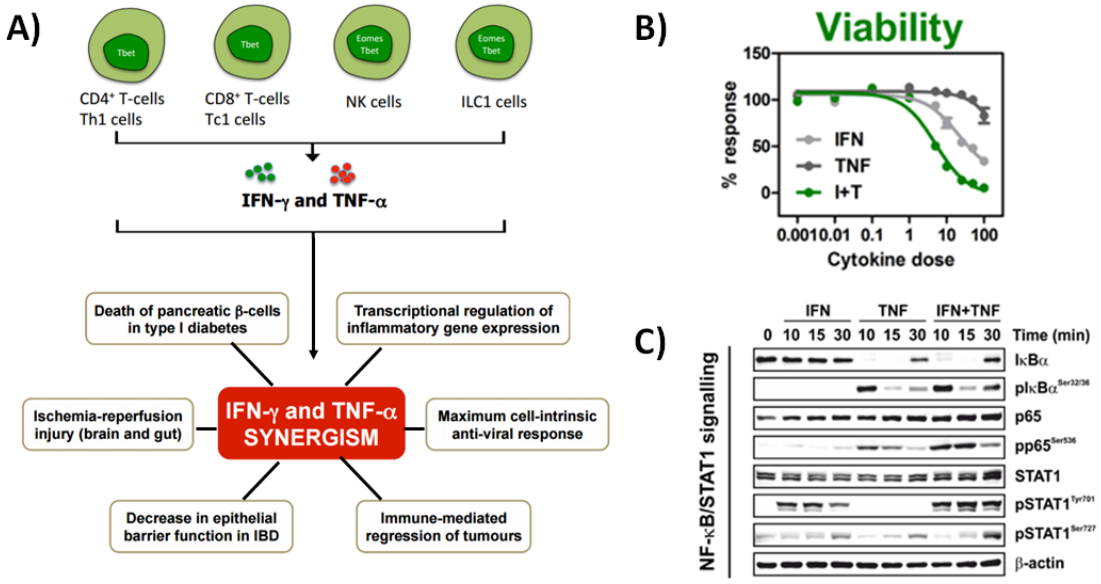

Figure 3-1: Non-additive effects after joint signaling with IFNγ and TNFα. **(A)** A summary of the cell types that secrete IFNγ and TNFα and the pleitropic synergistic downstream effects. **(B)** Dose response curve of cytokine treatment in human adenocarcinoma cells. The doses are in mM, and I+T is the joint signaling condition. Cells were assessed for viability to Cell Titer Glow (Promega) forty-eight hours after treatment. **(C)** Western blot of phosphorylation of various proteins in the NF-$\kappa\beta$ and JAK-STAT signaling cascades in response to cytokine treatment (all doses 10 ng/mL). β-actin is the control. *(Data courtesy of Ken Nally)*

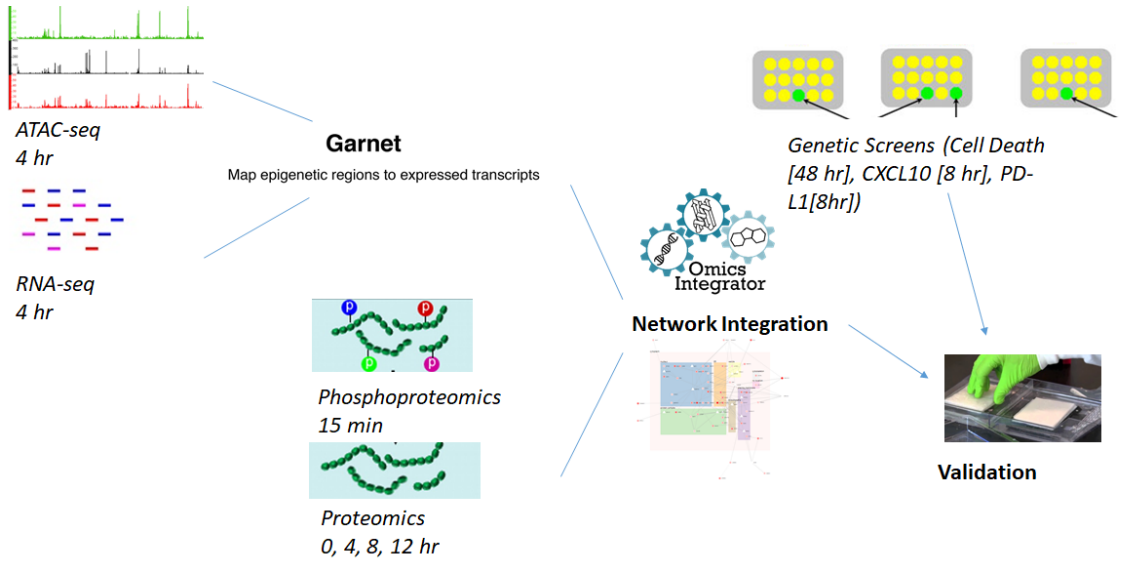

Figure 3-2: Overview of datasets, timepoints, and conditions used in the cytokine synergy study.

IFNγ and TNFα has been shown to mediate diverse synergistic responses ranging from decreased epithelial barrier function in IBD to immune-mediated regression of tumours **(Figure 3-1a)** [18, 19].

One leading hypothesis for this synergy is IFNγ priming, where signaling via IFNγ prepares the cell for a more pronounced response to other immune factors. TNFα [20]. While this hypothesis does explain some cellular responses to synergistic signaling, such as apoptosis, it fails to explain the diverse events happening at the molecular level **(Figure 3-1b)** [14]. Previous work has shown that the JAK-STAT and NF-$\kappa\beta$ pathways are involved in cytokine synergism **(Figure 3-1c)**. In turn, this likely triggers downstream transcriptional and epigenetic changes that mediate diverse cellular phenomena [21]. However, the precise genes and proteins in involved in these additional regulatory mechanisms are poorly understood [22]. In order to bridge this gap, our collaborators in the laboratory of Prof. Ken Nally from University of Cork in Ireland have collected a rich dataset, including phosphoproteomic, transcriptomic, epigenomic, and genetic screen data for synergistic signaling between IFNγ and TNFα in human adenocarcinoma cells **(Figure 3-2)**.

In this study, the diverse datasets described in figure 3-2 were integrated to elucidate

Genetic Screen Type	Top genes
Cell death (48 hours)	*JAK2, JAK1, APC, IFNGR2, STAT1, TNFRSF1A, LAP3, ARPC1A, IFNGR1, SYNGR2, IRF1, PINX1, WDR61, GALE, SUPT3H1*
CXCL10 production (8 hours)	*POMP, PSMD1, PSMC2, SNRPD3, PSMB5, PSMD14, SF3B5, SNRPB, POLR2A, PSMC3, SKIIP, CDCA5, PSMD6, SF3A3, SNRPF, POLR2I, UBC, SF3B3, SF3B2, MADD, PLK1, RPL6*
PD-L1 expression (8 hours)	*SF3B5, SF3B2, RPS2, PLK1, SNRPD3, PSMB7, SF3B1, NLGN4X, DDX48, KIAA1604, PSMB5, NUP205, BAT1, AQR, CDCA5, POLR2A, RPL37A, RPS3A, PSMD14, PSMD8, RPL1, SUPT6H*

Table 3.1: Top significant genes from RNAi rescue screen for 1) cell death 48 hours after stimulation with IFNγ and TNFα 2) CXCL10 production 8 hours after stimulation with IFNγ and TNFα 3) PD-L1 cell-surface expression 8 hours after stimulation with IFNγ and TNFα.

the specific molecular pathways underlying non-additive effects from joint TNFα and IFNγ signaling. These analyses not only discovered novel pathways, but also highlight the efficacy of the PCSF approach discussed in chapter one and incorporates some of the improvements to PCSF discussed in chapter two.

3.1 Genetic screens reveal upstream master regulators for cytokine synergy

Genome wide RNAi screens were performed in order to discover genes central to the disease relevant phenotypes characteristic of synergistic cytokine signaling. Previous data showed that synergistic signaling with IFNγ and TNFα lead to a much more pronounced cell death phenotype relevant in a variety of autoimmune disorders (Figure 3-1b) [23]. Another central process regulated by co-stimulation of cytokines was the production of CXCL10 and PD-L1 [24, 25]. CXCL10 controls the recruitment of regulatory T cells to a cell's local environment, while PD-L1 suppresses immune function and the activity of cytotoxic T cells [24, 25]. Numerous previous studies

have shown that the inhibition of PD-L1 can enhance immune function, serving as the basis of many immunotherapies [26]. By contrast, inhibition of CXCL10 can suppress inflammation in a tissue, which could potentially be important for treating inflammatory bowel disease. In other words, CXCL10 and PD-L1 can be thought of as tuning the immune response in a tissue: over-expression of CXCL10 and decreased expression of PD-L1 leads to aberrant inflammation, while decreased expression of CXCL10 and over-expression of PD-L1 can lead to cancers. Previous data also showed that CXCL10 and PD-L1 are expressed at high levels when stimulating with both IFNγ and TNFα, but not with either cytokine alone **(Figure 3-6b)**. Therefore, in order to profile these disease relevant phenomena, we performed genetic screens for cell death at 48 hours, and the production of CXCL10 and PD-L1 as proxies for immune system activation at 8 hours.

The top hits for the screen for cell death rescue include the Janus kinases, the IFNγ receptor, and the TNFα receptor, which are all involved in cytokine signaling (Table 3.1) [27]. As expected, the knockout of any of these genes prevented the activation of both pathways simultaneously, blocking synergistic effects. More surprisingly, another top hit was *ARPC1A*, a member of the Arp2/3 complex involved in actin polymerization. Some previous studies have indicated connections between actin dynamics and cytokine signaling [28]. GO enrichment analysis similarly implicated both components of the IFNγ and TNFα pathways, however, these provided little insight into the exact genes underlying this phenomenon.

The CXCL10 and PD-L1 screen recovered many components of the TNFα but few in the IFNγ signaling pathway. In particular, many components of the NF-κB and Wnt signaling pathways were uncovered in both these screens (Table 3.1). More specifically, the enrichments for the NF-κB and Wnt signaling pathways were driven by strong signal from the proteosome, a complex that degrades proteins that is activated by TNFα signaling [29]. The only components of the IFNγ signaling pathway that were recovered were *JAK2* and *STAT1*, suggesting the synergistic phenotype was mediated through STAT1's role as a transcription factor. Finally, both the screens were heavily enriched for a variety of transcriptional processes (FDR $< 10^{-27}$ for

both), further reinforcing that a lot of these synergistic phenomena were regulated by epigenetic changes, which is consistent with findings from previous literature [21, 30].

One disadvantage of these screens was that they recovered many genes that were likely specific to one pathway or the other, but did not necessarily recover genes that were common to both pathways. Unfortunately, since signaling with neither cytokine alone was sufficient to induce cell death, CXCL10 expression, or PD-L1 expression, it was impossible to tease apart proteins that were core to the *interaction* between the two pathways. To this end, we decided to profile the signaling dynamics, epigenetic changes, and transcriptional changes that occurred after individual and joint TNFα and IFNγ treatment.

3.2 Co-stimulation with IFNγ and TNFα activates distinct phosphorylation signaling cascades

The genetic screen indicated that profiling signaling cascades activated by IFNγ and TNFα was necessary to identify proteins important to the interaction between cytokine signaling. To this end, serine/threonine phosphorylation, tyrosine phosphorylation, ubiquitination, and general phosphorylation (IMAC) data were collected from human adenocarcinoma cells fifteen minutes after individual and joint stimulation with IFNγ and TNFα. Each of these four assays provided complementary information about signaling dynamics shortly after signaling. Additionally, there was strikingly little overlap between the genetic screen and the phosphoproteomics, indiciating that this assay was picking up complementary information **(Figure A-3)**.

Treatment with each cytokine stimulated known components of their respective pathways, but did not lead to activation of further downstream receptors. For example, stimulation of TNFα lead to activation of components of the NF-κ B signaling pathway (IKBKG, NFKB1), MAPK pathway (MAPK14, MAPKB, MAP2K7), and necroptosis (RIPK1), all of which have been previously described to be part of the TNFα signaling pathway (**Figure A-4**) [31]. As expected, TNFα did not lead to the activation of any

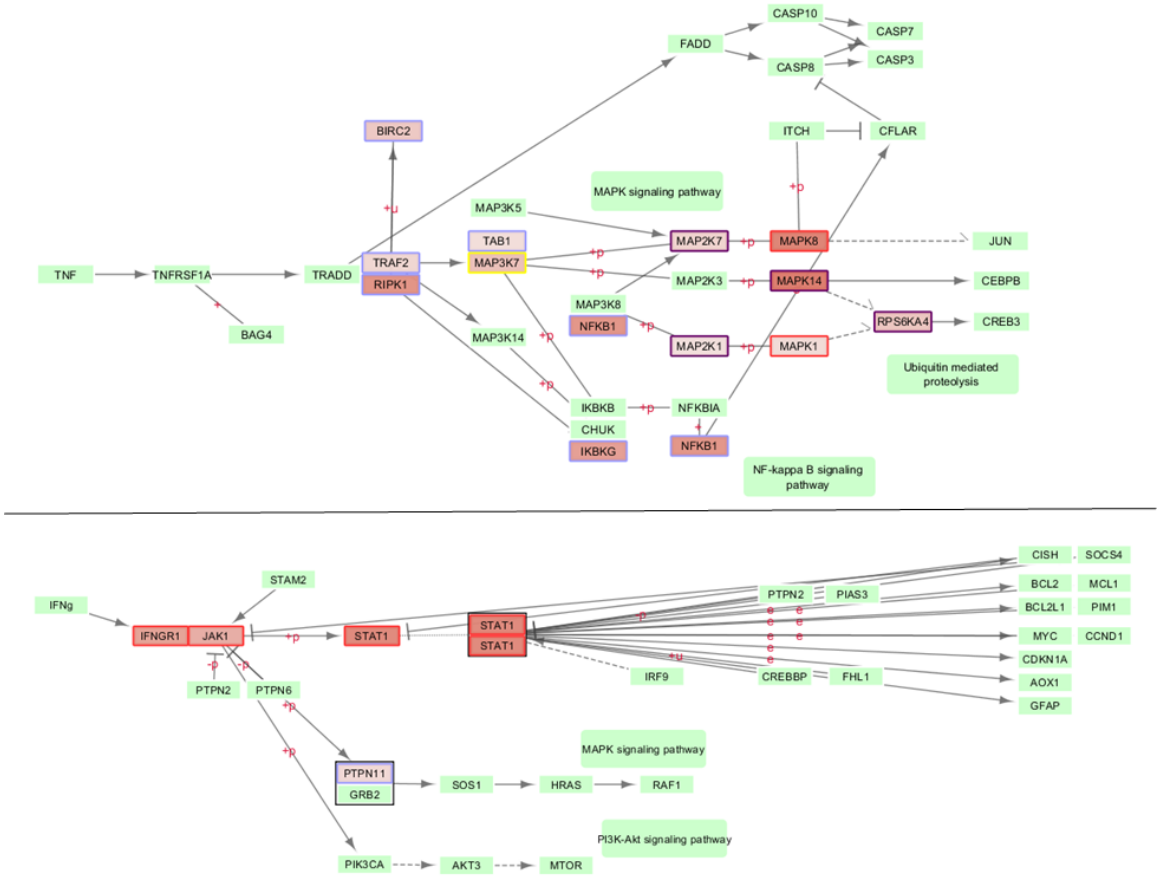

Figure 3-3: Protein modifications in the KEGG pathways for TNFα and IFNγ after joint cytokine treatment. The maximum fold change across the four assay types was then taken, and is indicated by the border color (yellow = Ser/Thr, dark purple = IMAC, red = Tyr, light purple = ubiquitination). The proteins in red showed a fold change greater than two after joint TNFα and IFNγ treatment. **(Top)** The KEGG pathway for TNFα signaling. **(Bottom)** A modified KEGG pathway for IFNγ signaling.

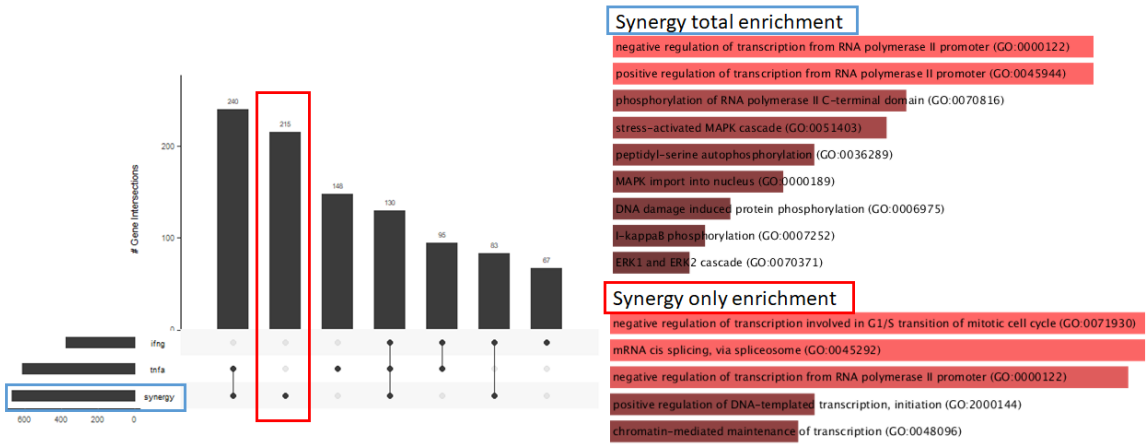

Figure 3-4: GO enrichments for protein modifications after TNFα and IFNγ treatment. **(Left)** Overlapping set visualization between protein modifications in each condition. The small barchart shows the number of proteins which had a modification that had more than a two-fold change after fifteen minutes of joint cytokine signaling in any of the four assays. The main barchart shows the overlap between each set. **(Top right)** Gene ontology enrichments performed on the synergy protein against a background of all proteins had more than a two-fold change after fifteen minutes in any of the four assays for any treatment condition. **(Bottom Right)** Gene ontology enrichment performed on proteins modified only in the synergistic condition against the same background as above.

part of the IFNγ signaling pathway except for a weak activation of JAK1. Similarly, treatment with IFNγ lead to the stimulation of the JAK-STAT pathway, which is the primary pathway known to be activated by IFNγ. In addition, some components of NF-κB pathway were also activated, indicating that downstream targets of the NFκB pathway could potentially be the mediators of the interaction between the two pathways **Figure A-5** [32].

Treatment with both IFNγ and TNFα jointly lead to the the activation of known components both pathways. In the IFNγ pathway, joint cytokine treatment lead to large upregulation of tyrosine phosphorylation in the JAK-STAT pathway (JAK is a tyrosine kinase). Similarly, the TNFα pathway was upregulated in essentially the same fashion in the individual treatment condition, including upregulation of the the NFκB, MAPK, and necroptosis pathways **(Figure 3-3)**

Next, novel proteins involved in the synergistic response were identified by performing gene ontology enrichment on proteins identified as differentially modified in

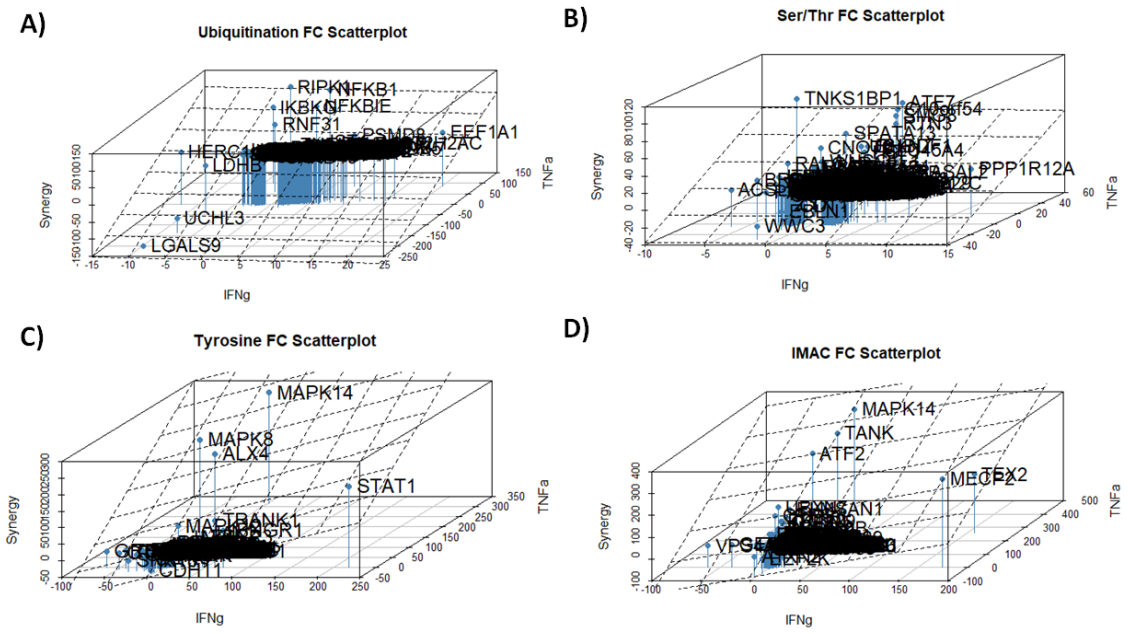

Figure 3-5: Multiple linear regression was used to predict the fold change of protein modifications after synergistic cytokine treatment using the fold change individual protein modification responses after TNFα and IFNγ treatment. The analysis was performed separately for **(A)** ubiquitination **(B)** ser/thr phosphorylation **(C)** tyrosine phosphorylation **(D)** general phosphorylation (IMAC).

the synergistic condition across any of the orthogonal signaling assays. This analysis returned enrichments associated with both the TNFα and IFNγ signaling pathways such as activation of the MAPK cascade and NFκB pathway **(Figure 3-4)**. However, subsetting only genes that were activated in the synergistic response, there was little enrichment for anything other than general biological processes like transcription and mRNA splicing (Figure 3-4).

This previous analysis indicated that at the fifteen minute timepoint, there were

Experiment Type	**IFNg regression coefficient**	**TNFa regression coefficient**
Ubiquitination	-0.41	0.64
Ser/Thr Phosphorylation	-0.1	0.95
Tyr Phosphorylation	0.67	1.04
IMAC (Phosphorylation)	0.75	0.82

Table 3.2: Regression coefficients for the multiple linear regression described in figure 3-5.

no annotated pathways that were activated in any of our orthogonal assays that were due to *crosstalk* between IFNγ and TNFα signaling. However, this analysis failed to account for how the fold changes in these proteins modifications were influenced by either TNFα or IFNγ signaling. Therefore, we performed a multiple linear regression for fold changes of each type of modification, using the individual treatment conditions to predict the joint signaling condition. The regression are plotted in **Figure 3-5** and the regression coefficients are reported in **Table 3.2**. The ubiquitination and serine/threonine phosphorylation were largely driven by TNFα signaling with regression coefficients for TNFα being 0.64 and 0.95 respectively. Some of the largest outliers in the ubiquitination regression analysis were components of the NF-κB pathway (RIPK1, IKB, NFKB1) suggesting that, that this pathway's activity was enhanced in a non-additive fashion in the synergistic condition (Figure 3-5 a). The largest outlier in the serine/threonine phosphorylation regression analysis was the protein TNKS1BP1, an ankyrin binding protein that plays a role in cytoskeletal signaling (Figure 3-5b). Combined with the identification of Arp2/3 in the genetic screen assay, this highlights the critical role of the cytoskeleton scaffolding in cytokine synergy [33]. On the other hand the tyrosine and general phosphorylation (IMAC) assays saw contributions from both IFNγ and TNFα. The top hits in these pathways were components of the MAPK (MAPK8, MAPK14, ATF2), NFκB (TANK) and JAK-STAT pathways (STAT1). Taken together, the regression analysis suggested that while the pathways being activated in synergistic signaling were the same as in both individual cytokine signaling conditions, there were non-additive effects, especially in the activation of the MAPK and NF-κB pathways.

Some weaknesses of this assay were that proteomics were not collected from the same timepoints, meaning that some of the fold changes in phosphorylation could occur due to changes in protein copy number. However, considering that even four hours after signaling with TNFα and IFNγ, there were very few significantly differential proteins, most of the phosphoproteomic signal most likely was due to actual changes in the phosphorylation state of proteins present before cytokine treatment. Another weakness of this assay was that it only provided a single snapshot of global phosphorylation

state, so it is possible that this analysis missed transient early signaling events or integrative signaling events at later timepoints. Nonetheless, this assay demonstrated that at early timepoints, synergy between the two pathways was largely mediated by the *strength* of the response of the activation of the respective pathways. However, the low overlap between different data types also demonstrate the need for integrated network analysis to use information not directly captured in any individual dataset. Moreover, the observation that many of the proteins and pathways identified act through transcriptional means suggested the need to profile the transcriptional and epigenetic states of the cell after joint cytokine treatment.

3.3 Epigenetic changes mediate synergy between IFNγ and TNFα

IFNγ and TNFα both lead to epigenetic changes that affect the response of treated cells to further signals. For example, in mice, IFNγ has been shown to induce histone acetylation in the TNFα and Nos2 loci in macrophages, eventually leading to the development of colitogenic macrophages [34]. IFNγ has also been shown to increase the occupancy of transient transcription factors like STAT1 and IRF1 in macrophages, priming the chromatin environment and augmenting gene transcription in response to subsequent stimulation of toll-like receptors. [21] More generally, IFNγ has been shown to alter the expression of a variety of histone-modifying enzymes, while inhibiting a subset of these enzymes decreases macrophage response to IFNγ [35].

Similarly, TNFα has been shown to modulate levels of a member of the polycomb repression complex, EZH2, which in turn as an epigenetic brake to modulate TNFα functions in colitis [36]. Previous work has also showed that TNFα is regulated epigenetically, and that IFNγ leads to transcription factors binding to TNFα promoters and enhancers, preparing cells for a greater response to subsequent TNFα signaling [21, 30]. In the assay for changes in phosphorylation in response to TNFα and IFNγ described above, many of the top enrichments were also for transcription-related

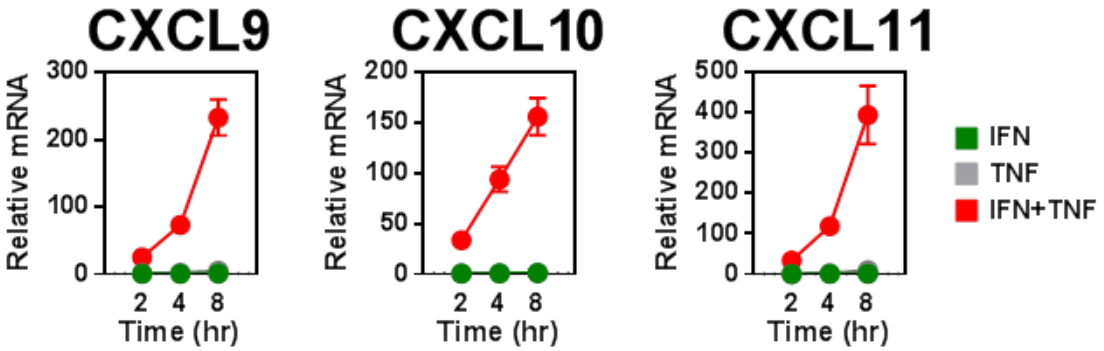

Figure 3-6: mRNA levels of the chemokines CXCL9, CXCL10, and CXCL11 after TNFα and IFNγ stimulation at 10 ng/mL. The mRNA levels after stimulation were determined relative to an untreated control. *Data courtesy of Jerzy Woznicki.*

processes. However, little previous work has focused on the epigenetic effects of IFNγ on non-immune cells, nor the interaction between epigenetic changes caused by both signaling simultaneously with TNFα and IFNγ.

Since previous literature has described sites becoming more accessible and occupied by transcription factors in response to individual signaling by IFNγ, we decided to probe for differentially open chromatin using an Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq). In short, ATAC-seq uses the Tn5 transposase to identify accessible regions of DNA that are more likely to be transcribed [37]. However, ATAC-seq requires almost one thousand times less starting material than comparable methods like DNAse hypersensitivity [37]. Using the protocol previously described by Buenrostro et al., we performed ATAC-seq on human adenocarcinoma cells four hours after individual and joint stimulation with TNFα and IFNγ at 1- ng/mL in two biological replicates [37]. We chose the cell lines and cytokine concentrations to be consistent with our other orthogonal assays. We performed the assays at the four hour timepoint, because we determined that the transcription of a variety of C-X-C motif chemokines whose expression are known to be induced by interferons increased at the four hour timepoint **Figure 3-6** [24]. We then aligned the reads from the assay, performed peak calling and quality control, then found differential peaks (full protocol details in **Appendix B**).

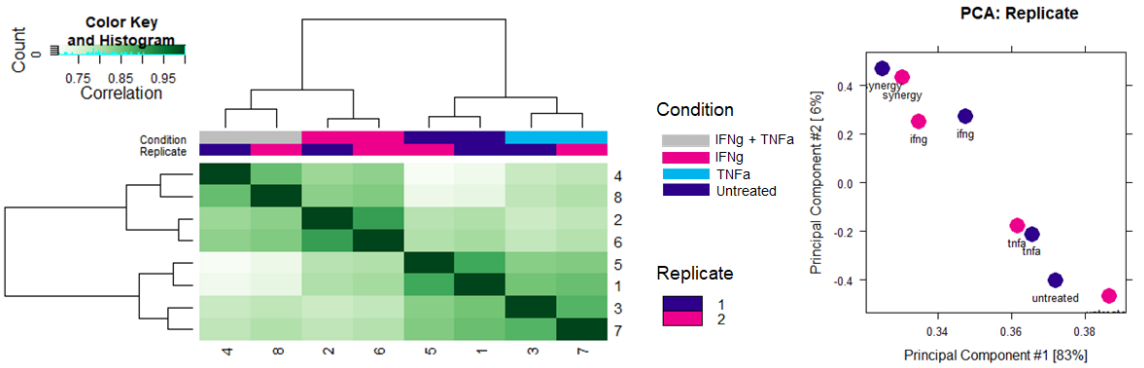

Figure 3-7: Hierarchical clustering and PCA of ATAC-seq peaks. **(Left)** ATAC-seq peaks were hierarchically clustered. IFNγ and synergistic cytokine signaling conditions cluster together, while TNFα and untreated control conditions cluster together. **(Right)** Principal components analysis on the frequency of reads in ATAC-seq consensus peaks.

Overall, the data was quite high quality. We show representative outputs for per-base quality scores from FastQC in **Figure A-6a**. Similarly, the fraction of reads mapping to peaks stayed consistently around twenty percent, which is quite high **(Figure A-6b)**. There were some issues with PCR-overamplification and overrepresentation of certain sequences on the ends of reads, but these were resolved computationally (**Figure A-7**). The reads were also distributed evenly across all the chromosomes, with a representative examples shown in **Figure A-8**. There were also significant overlaps in the majority of peaks **(Figure A-9)**. Hierarchical clustering revealed that biological replicates clustered together. Unsupervised approaches further confirmed the quality of the data. Performing a principal components analysis similarly demonstrated that biological replicates clustered closely together (**Figure 3-7**).

Next, differential peaks were calculated using DiffBind. Looking at the distribution of differential peaks, we see that in both TNFα and IFNγ samples, most differential peaks are upregulated in the cytokine stimulated condition (**Figure 3-8**). This is consistent with previous observations that stimulation with these cytokines leads to increased binding of transcription factors and transcription, which requires more regions of DNA to be open [21, 30]. However, there are many more sites in the synergy vs. untreated conditions (2490) than in either IFNγ (1425) or TNFα (192) alone, indicating some non-additive effects of joint cytokine treatment. However, most of the

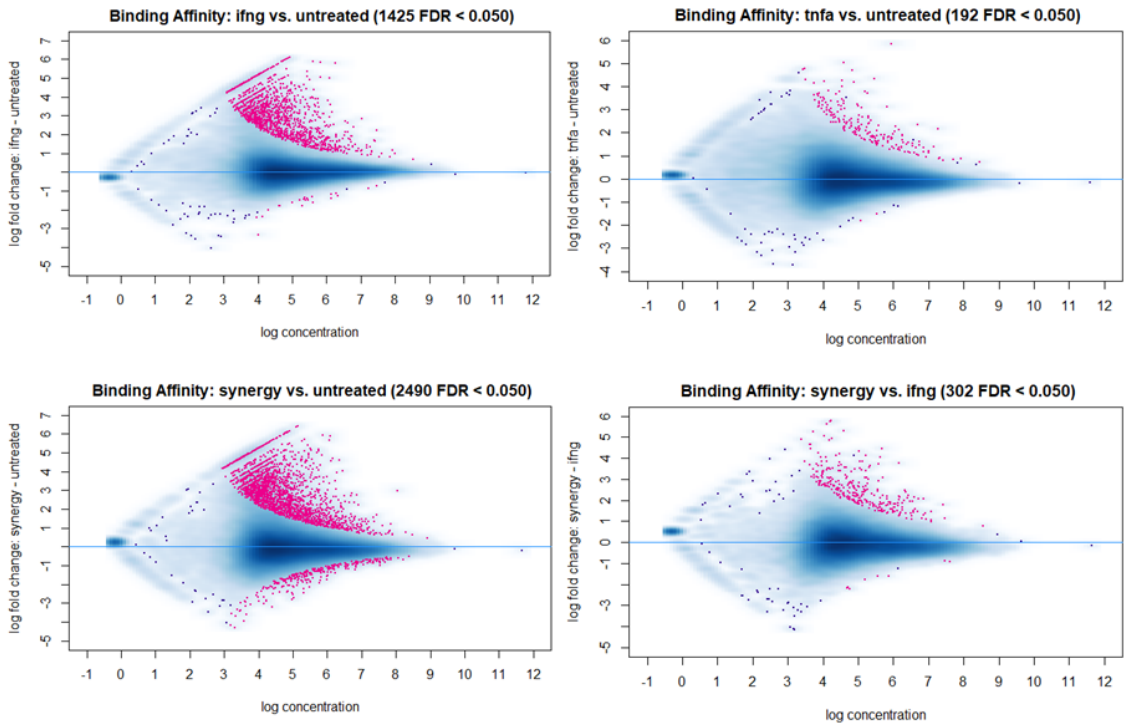

Figure 3-8: Log of the mean average read count of each peak plotted against the log ratio of the peaks of each condition. Differential peaks with a FDR < 0.05 are plotted in red. **(Top Left)** IFNγ vs. untreated. **(Top Right)** TNFα vs. untreated. **(Bottom Left)** IFNγ + TNFα against untreated. **Bottom Right** IFNγ + TNFα treated against IFNγ only.

signal seems to be driven by IFNγ alone, with only 302 differential peaks between the synergistically treated vs. IFNγ only conditions (Figure 3-8).

Differential peaks were then assigned to previously annotated genomic features from UCSC hg19 (**Figure 3-9**). These revealed that more than half of the peaks for all the conditions fell in distal intergenic features, which was consistent with previous ChIP-seq experiments performed using similar cytokine stimulated conditions (data not shown). Next, peaks within 3 KB of a transcription start site (TSS) were assigned to their nearest gene. Enrichment analysis was then performed on this gene set. As expected, this analysis revealed IFNγ signaling related processes like interferon and anti-viral responses were enriched in the IFNγ signaled condition, and TNFα signaling related processes like NF$\kappa\beta$ signaling were enriched in the TNFα signaled condition (**Figure A-10a,b**). For the synergistically signaled vs. untreated differential peaks,

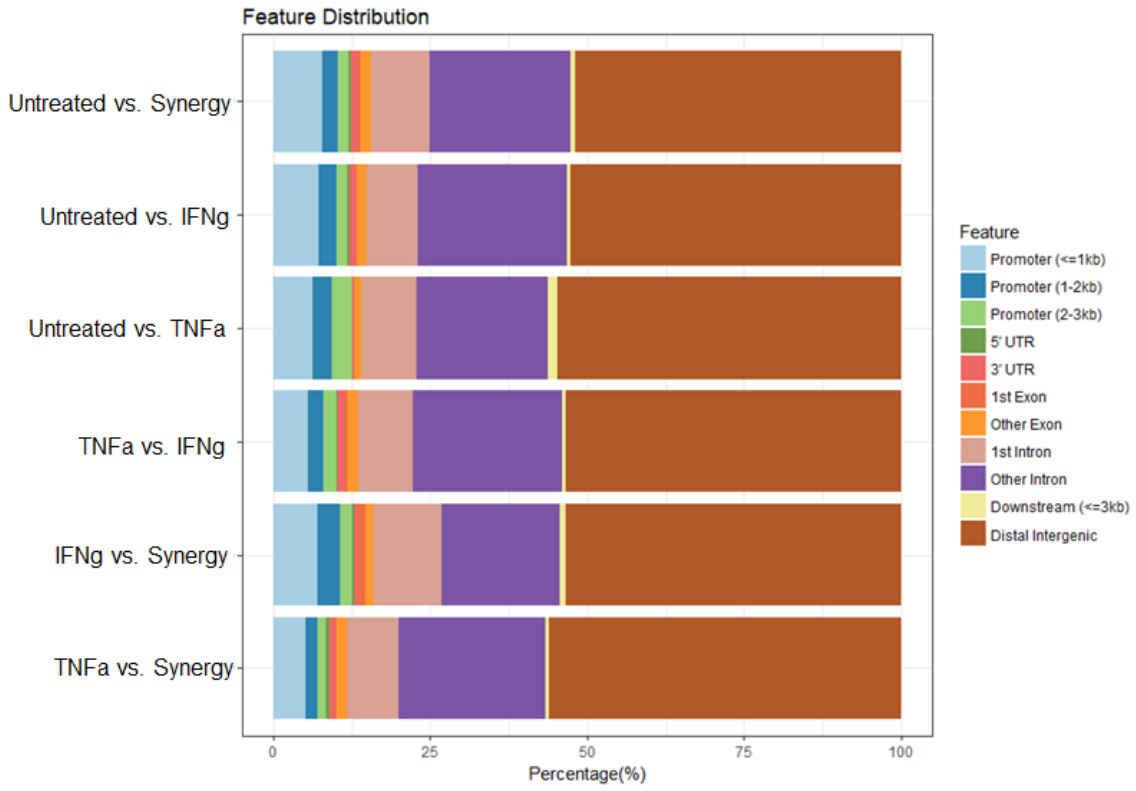

Figure 3-9: Percentage of the most probable epigenetic feature for each differential ATAC-seq peaks. Each barchart represents a different set of differential peaks (ex: untreated vs. IFNγ + TNFα stimulated).

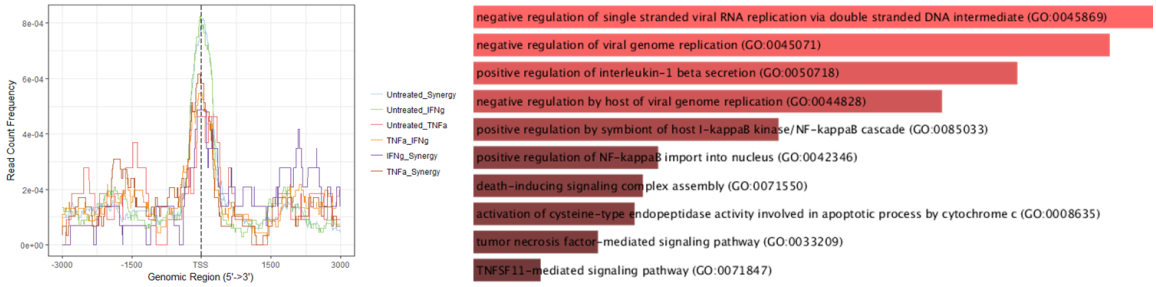

Figure 3-10: (Left) Distance from promoters within 3 kilobases of a transcription start site for each set of differential peaks. **(Right)** GO enrichment of genes with differential peaks within 3 kilobases of their transcription start site for the synergistically signalled vs. untreated condition.

the genes near a TSS were enriched for both TNFα and IFNγ processes (**Figure 3-10**). The signal for the synergistically signaled condition seemed to be driven more by IFNγ signaling, with comparisons between IFNγ and synergistic conditions revealing only weak enrichments for general biological processes, while comparisons between TNFα and synergistic conditions revealed some enrichments for TNFα only processes **(Figure A-10c,d)**. Similarly, looking at genes near differential peaks only in the synergy vs. untreated condition revealed very few significant enrichments (all FDR $>$ 0.3) **(Figure A-11)**. Both these pieces of evidence suggested that instead of directly looking for peaks that were only differential in the IFNγ condition, it was important to find genes that were activated in both the IFNγ and synergistic conditions that could be moderated by TNFα signaling.

One such differential peak was a peak in the promoter region of ZBP1, a protein that binds the Z-isoform of DNA that also plays an important role in anti-viral responses [38]. There were many reads in the IFNγ and synergistically treated conditions but not in the untreated or TNFα treated conditions, suggesting this gene was selectively induced by IFNγ signaling **(Figure 3-11)**. Previous studies have shown that ZBP1 can induce necroptosis in a RIPK3 dependent fashion [39]. In the phosphoproteomic data, RIPK1, an inhibitor of RIPK3, was highly ubiquitinated in the TNFα (106.1 fold change) and synergistic (126.8 fold change) conditions, but not the IFNγ condition (1.5 fold change). Previous studies have shown that RIPK1 is inhibited by high ubiquitination [40]. This suggests a potential ZBP1 mediated synergistic response

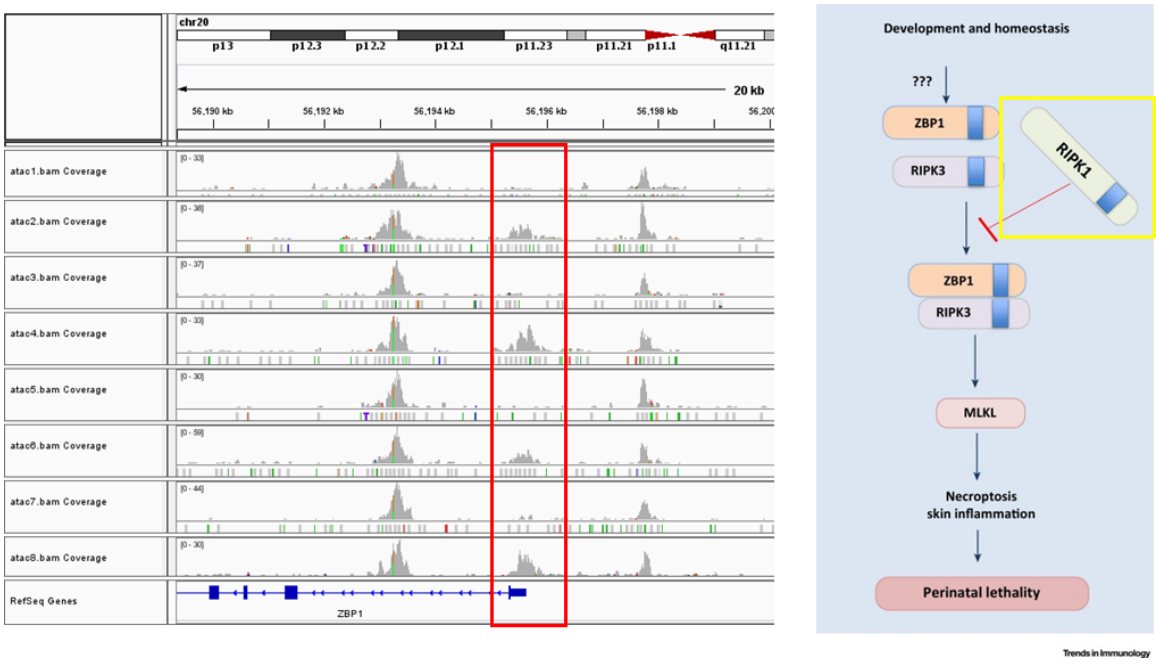

Figure 3-11: (Left) ATAC-seq peaks around the transcription start site of ZBP1. The first and fifth rows are untreated controls, the second and sixth rows are treated with IFNγ only, the third and seventh rows are treated with TNFα only, and the fourth and eighth rows are treated with both IFNγ and TNFα. **(Right)** Pathway for ZBP1 highlighting the protein RIPK1 in yellow. *(reproduced from Kurikose et al.)*[38]

between IFNγ and TNFα. After stimulation by both cytokines, TNFα signaling leads to ubiquitination and downregulation of the activity of RIPK1, which releases the RIPK1 inhibition on the binding of ZBP1 and RIPK3. This in turn could lead to the synergistic cell death that was observed in figure 3-1b. In order to validate this hypothesis, we are currently working with our collaborators to validate the synergistic interaction between ZBP1 and RIPK3.

3.4 TNFα and IFNγ synergistically induce changes in protein expression

Previous data has showed that there is poor correlation between mRNA expression and protein levels [41]. Moreover, the early protein modification dataset showed strong activation of the NF-κB pathway which has been shown to lead to activation of the

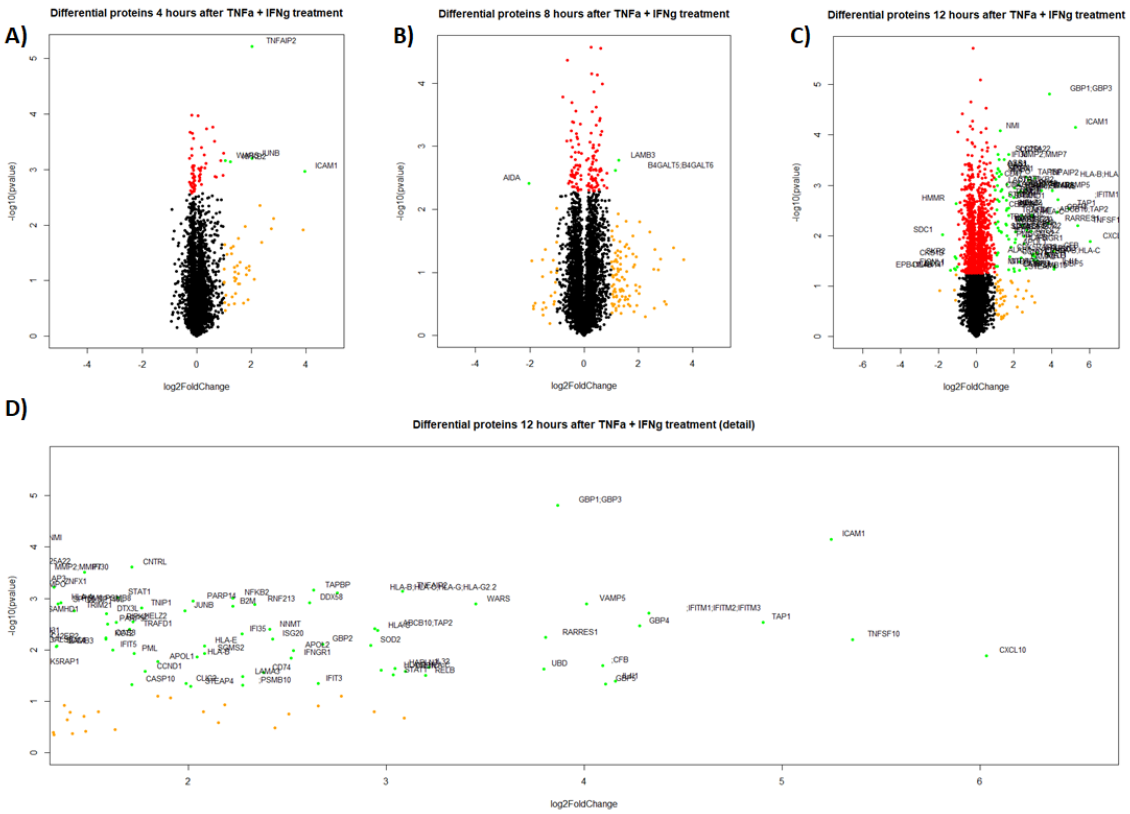

Figure 3-12: (Top) Volcano plot of differential proteins 4, 8, 12 hours after IFNγ + TNFα treatment **(Bottom)** Detail of volcano plot at 12 hour timepoint showing only proteins with a fold change greater than 2.

proteasome [31]. Therefore, we next sought to find out which proteins are actually differentially translated as well as degraded. Although previous literature has shown that many IFNγ proteins are only translated between 12-24 hours after treatment, we noted from the genetic screen that there were already pronounced differences in immune marker proteins by the 8 hour timepoint [42]. Therefore, we decided to profile the relatively early timepoints of 4 hours, 8 hours, and 12 hours after treatment with both TNFα and IFNγ.

There were few differential proteins at the 4 and 8 hour timepoints (5 and 3 proteins with FDR < 0.1) **(Figure 3-12a, b)**. Some of these proteins were known early response genes to TNFα and IFNγ (JUNB/ TNFAIP2 and ICAM1 respectively). There were many more differential proteins at the 12 hour timepoint (137) **(Figure 3-12c)**. Some of the top hits are protein products known to be strongly upregulated

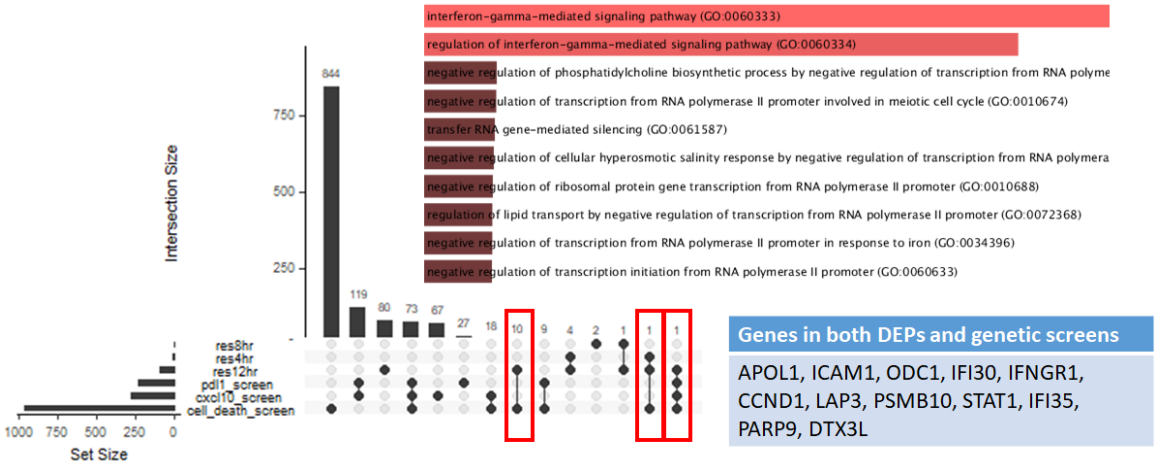

Figure 3-13: (Left) Overlapping set visualization between DEPs and genetic hits. The proteins in the boxed overlaps are listed in the table. (Right) Enrichments for the overlap between DEPs and genetic screen hits.

following type 2 interferon signaling including CXCL10 and ICAM1. More generally, the differentially expressed proteins at 12 hours were enriched for IFNγ related processes (FDR $< 10^{-21}$) **(Figure A-12)**.

In order to see how the proteins identified in this assay overlapped with master regulators, differentially expressed proteins were overlapped with genetic hits for cell death, CXCL10 production, and PD-L1 expression. The resulting set of 12 genes were heavily enriched for IFNγ related processes (FDR $< 10^{-9}$) **(Figure 3-13)**. Interestingly, these genes were poorly enriched for the TNFα pathway, suggesting that many of the crucial late events involved in cytokine synergy are mediated through the IFNγ pathway. This further suggests that TNFα plays a role in enhancing the strength of IFNγ signaling.

In order to discover novel processes involved in the interaction between the two pathways in the unbiased way, the PCSF algorithm was applied to differentially expressed proteins. This network analysis identified the apoptotic pathway of the TNFα pathways as a crucial component of cytokine synergy. Therefore, even though both the enrichment analysis and overlap with genetic screen failed to identify components of the TNFα pathway, our unbiased approach uncovered proteins like FADD and TRADD that are associated with TNFα signaling. Moreover, this approach suggests

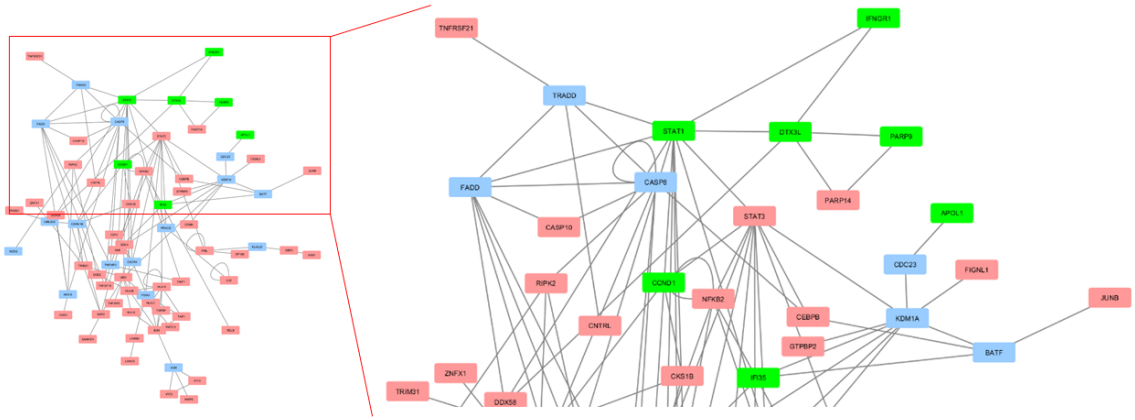

Figure 3-14: (Left) PCSF subnetwork generated using OmicsIntegrator2. Prizes were the log fold changes of differential proteins, parameters were (w = 1, b = 1, g = 3000), and the starting interactome used was iRef14 [8]. **(Right)** Detail of subnetwork highlighting the upstream parts of the network.

that the transcription factor STAT1 interacted with downstream components of the TNFα pathway to generate the synergistic phenotypes. In order to leverage the network approach further, we used the PCSF approach to integrate phosphoproteomic and genetic screen data.

3.5 Integrating phosphoproteomic and genetic screen data reveals known and novel genes and pathways underlying cytokine synergism

All of the previous single dataset analysis has indicated the need to integrate data across a variety of orthogonal assays in order to reconstruct the pathways underlying cytokine synergy. In order to do so, OmicsIntegrator2 was used to solve the PCSF problem to generate subnetworks for further analysis.

Integrating different datasets demonstrates the capacity of network modeling to more fully capture the biology of cytokine synergism. The top hits from each data type were assigned prizes based on the degree of confidence in each hit. For example, for the phosphoproteomic data, the top synergy score across the different phosphorylation

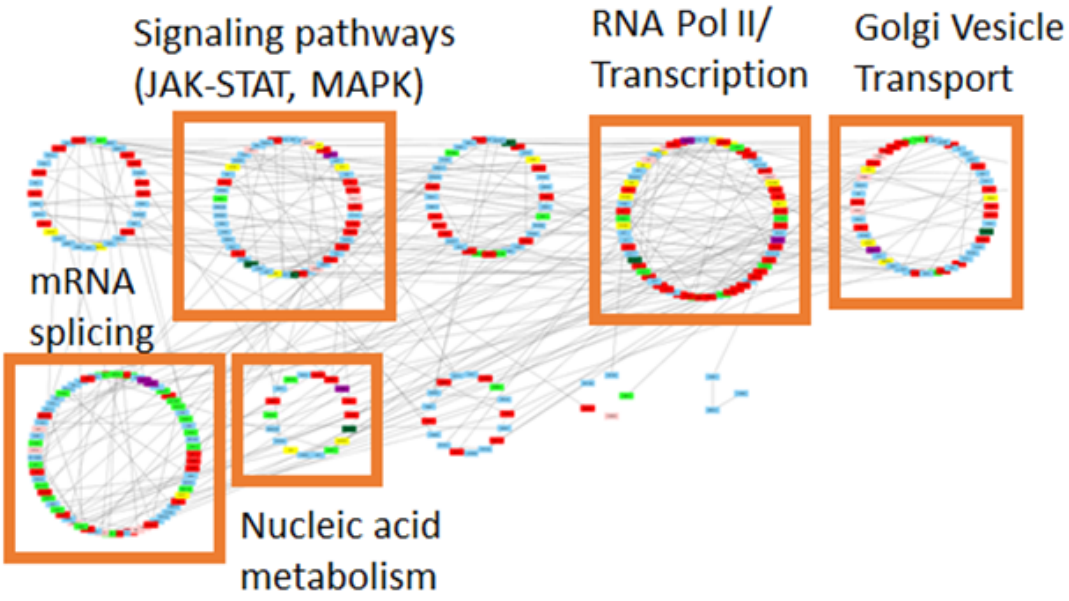

Figure 3-15: Integrated network analysis with PCSF using prizes from the phosphoproteomic (magnitude of fold change in any protein modification dataset > 2) and genetic screen (magnitude of robust $Z > 2$) The resulting subnetwork was Louvain clustered and labelled GO enrichments were performed with BiNGO [43]. Yellow nodes represent CXCL10 only prizes, dark green represents PD-L1 only prizes, red nodes represent cell death only prizes, dark blue represents phosphorylation only prizes, purple represents cell death and phosphorylation prizes, lime green represents CXCL10 and PD-L1 prizes, and light blue represents Steiner nodes.

datasets was used as the prize input. For the genetic screens, all proteins with a robust Z-score of greater than two were treated as hits and were assigned prize weights proportional to the Z-score. The list of hits were then appended together taking the maximum across datasets with the sum of prize weights of each data set being normalized to be equal. These hits were used as the prizes for PCSF, which was solved with OI2. The parameters of β, ω, and γ were determined using the qualitative heuristics described in chapter two. Specificity and robustness of resulting networks was calculated by 100 randomization trials. The union of all nodes and edges with specificity of less than 0.2 and robustness of greater than 0.8 were used to create the final subnetwork. Community detection on the ensuing network was performed using Louvain clustering, and gene ontology enrichments were calculated for each cluster using BiNGO **(Figure 3-15)** [43].

These analyses revealed known clusters such as JAK-STAT signaling and mRNA processing, as well as novel clusters like the RNA Pol II/transcription cluster. This data corroborates well with previous literature suggesting the involvement of epigenetic changes in cytokine synergism, but also reveals previously unknown proteins involved in this process like SMAD and BMP4 [21]. Thus, this analysis demonstrates the power of integrated network analysis in generating new biological hypotheses. However, the inability to resolve the exact changes in the RNA Pol II/transcription cluster point to the need to integrate additional transcription factor data derived from RNA-seq and ATAC-seq to pinpoint how epigenetic and transcriptional changes influence cytokine synergism.

3.6 Future work

Collecting and analyzing RNA-seq data will allow for the prediction of transcription factors. In turn, this will allow for more meaningful network analyses to link changes in protein signaling to changes in transcription. In addition, there are planned experiments for ChIP-seq, which will also help predict transcription factors that are imperative for further network analyses.

Other potential avenues for further work are using different network modeling algorithms. In particular, using multi-commodity flows could help outline synergistic pathways in a directed graph, while using the multi-PCSF approach could help prioritize shared pathways between IFNγ and TNFα signaling.

Once integrated networks have been generated using the datasets outlined above, synergistic proteins will be prioritized for further biological validation. Specifically one hundred genetic targets will be assayed using a CRISPR-Cas9 knockout study assaying changes in chemokine production, resistant to viral infection, and transcription factor localization.

Chapter 4

Discovering conserved pathways of age-related neurodegeneration across *Drosophila* and human AD patients using the PCSF approach

Alzheimer's disease (AD) is a chronic disease with an estimate prevalence of ten to thirty percent in people over 65 [46]. On the cellular level, Alzheimer's disease is associated with buildup of insoluble forms of extracellular amyloid-beta (Aβ) in plaques and aggregation of tau in neurofibrillary tangles [46]. The importance of Aβ and tau is corroborated by evidence that mutations in the amyloid precursor protein (APP) and in the gene encoding tau protein (TAU) can cause progressive neurodegeneration [47, 48]. Although the precise nature of the interaction between Aβ and tau is not understood, the amyloid hypothesis suggests that Aβ works upstream of tau to promote neurodegeneration **(Figure 4-1)** [49]. Despite these insights, Alzheimer's disease is not a highly penetrant Mendelian disease [50]. Indeed, a variety of risk factors including age, hypertension, depression, smoking, diabetes mellitus, and obesity have been shown to linked with AD [51]. The most prominent of these risk factors is age, with AD morbidity increasing 100-fold between 45 and 80 years

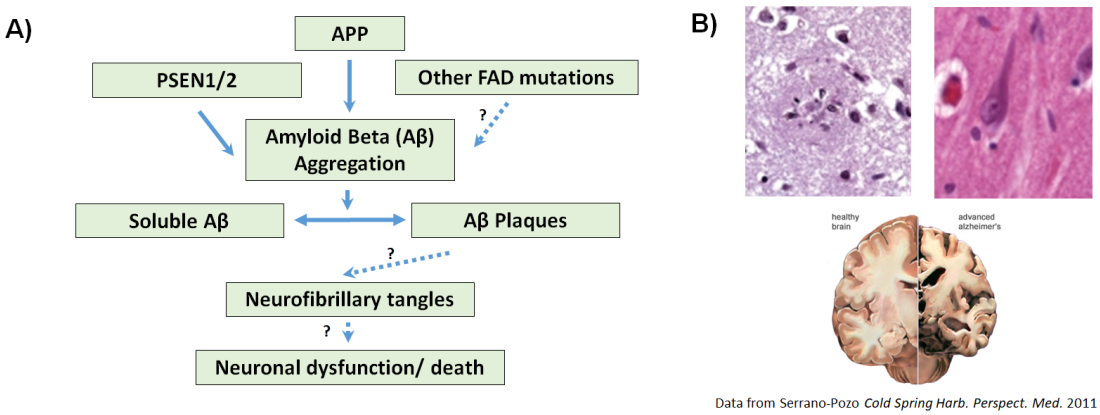

Figure 4-1: **(A)** Outline of the amyloid-beta hypothesis of Alzheimer's disease progression. This posits that mutations in presinilin 1 (*PSEN1*), presinilin 2 (*PSEN2*), amyloid precursor protein (*APP*) and other mutations involved in familial Alzheimer's disease (FAD) lead to aberrant processing of amyloid beta protein. In turns this leads to aggregation of soluble Aβ in extracellular spaces into amyloid-β plaques. These Aβ plaques indirectly leads to the formation of neurofibrillary tangles consisting of aggregates of hyperphosphorylated tau protein through processes that are poorly understood. In turn, neurofibrillary tangles ultimately lead to neuronal dysfunction/ death that contribute to the dementia characteristic of Alzheimer's disease [44]. **(B)** *(Top)* H&E stains for Aβ plaques and tau tangles, that eventually lead to *(bottom)* neuronal death, brain vacuole formation, and dementia [45].

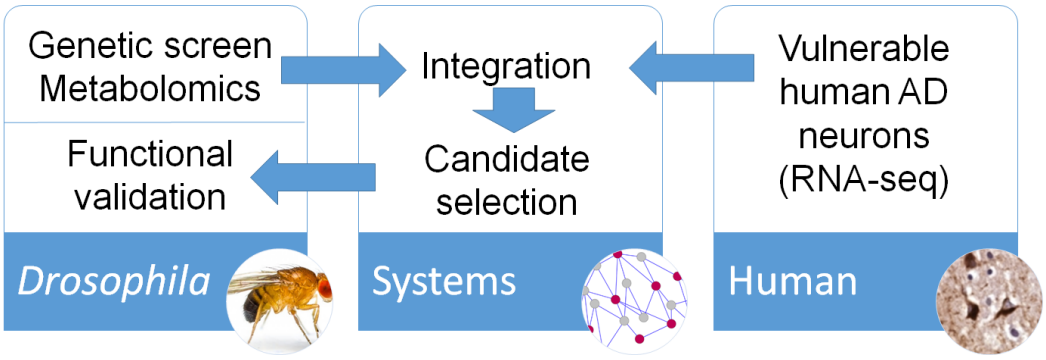

Figure 4-2: Outline of our integrative approach to discover conserved pathways of neurodegeneration underlying AD. A screen for genes that rescued brain mass in *Drosophila* models for AD and metabolomics on whole fly heads were performed in *Drosophila*. RNA-seq was collected from pyramidal neurons from layer III/IV of the temporal cortex, which have previously been shown to be preferentially affected in Alzheimer's disease. Both of these datasets were integrated together using the OmicsIntegrator approach and were used to select candidates for further functional validation in *Drosophila*.

of age [52]. However, the cellular mechanisms connecting aging with AD is not well understood, and in-depth genetic studies of the mechanisms controlling age-related neurodegeneration using model organisms like *Drosophila* have not yet been reported.

In order to address this gap, the labs of Prof. Mel Feany, Dr. Clemens Scherzer, and Prof. Fraenkel are collaborating to link proximal conserved mechanisms of age-related neurodegeneration from *Drosophila* studies to cell type specific expression quantitative trait loci (eQTLs) in human Alzheimer's patients using a systems approach. In this study, I used the data generated from the labs of Prof. Feany and Dr. Scherzer as input to the PCSF algorithm to discover hidden genes and pathways linked to AD (see **Appendix B** for specific details on the data). To do so I investigated changes in metabolites and proteomics in a *Drosophila* model for AD. Separately, I analyzed how gene expression was perturbed in a neuronal subpopulation that is affected preferentially by AD. Next, I mapped information from both of these datasets onto a protein-protein interaction network to discern pathways dysregulated in AD. This approach will not only reveal mechanisms of age-related neurodegeneration that promote AD, but also highlights the power of the PCSF approach to integrate data from both human and *Drosophila* to discover hidden pathways underlying neurodegenerative

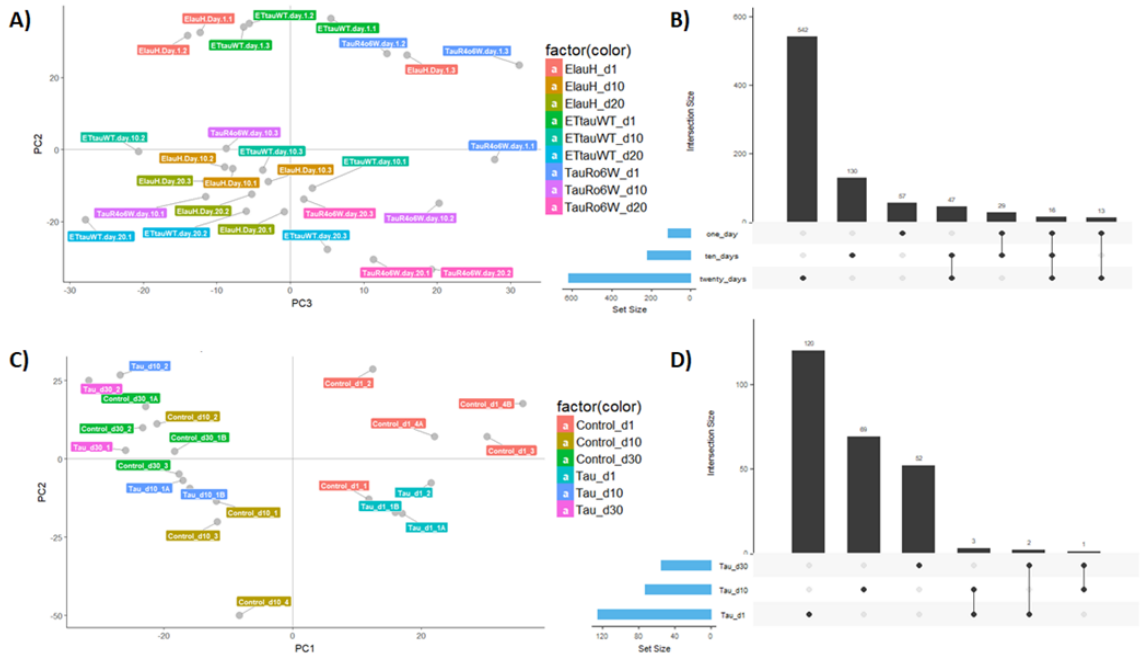

Figure 4-3: Reanalysis of proteomic data from the Emory *Drosophila* study and RNA expression data from Scherzer *et al.* 2003.[53] **(A)** Principal components analysis of proteomics from the Emory *Drosophila* study. The first and second principal components, explaining 41% of the variation, are plotted here. **(B)** Intersections between differentially expressed proteins in WT and TauRo6W tau mutant flies at 1 day, 10 day, and 20 day old. The blue barchart shows the cardinality of each set, and the black barchart shows the number of proteins for each subset as indicated by the black dots below each bar. **(C)** Principal components analysis of RNA expression data from Scherzer *et al.* 2003. The first and second principal components, explaining 68% of the variation, are plotted here. **(D)** Intersections between differentially expressed genes in WT and TauRo6W tau mutant flies at 1 day, 10 day, and 30 day old. The blue barchart shows the cardinality of each set, and the black barchart shows the number of genes for each subset as indicated by the black dots below each bar.

diseases **(Figure 4-2)**.

4.1 Previous data from *Drosophila* tauopathy models show increased differential signal at later timepoints

In order to connect findings from *Drosophila* genetic screens to Alzheimer's pathogenesis, more detailed profiling of phosphoproteomics, proteomics, and RNA-seq is

necessary. Previous data for proteomics and RNA expression are available; however, further profiling of phosphoproteomics is still necessary [53]. In particular, due to financial and practical experimental constraints, it is important to profile phosphoproteomics in *Drosophila* models for neurodegeneration at only timepoints with sufficient signal. To identify the best timepoint for more in depth profiling, I examined proteomic and RNA expression data from two previous studies to determine that later timepoints were best for further data collection [53].

First, I analyzed proteomics data from two *Drosophila* models of tau neurotoxicity relevant to Alzheimer's disease and related tauopathies: *Abeta*, a line that overexpressed Aβ protein, and *tau*, a line that overexpressed humanized mutant tau protein **(Figure 4-3a,b)** [54]. A principal components analysis showed clear segregation of samples by age, reinforcing the need to carefully choose timepoints for follow up analysis. In general, samples overexpressing Tau also segregated from control samples as expected. However, there was significant technical variation, even in pooled samples (Figure 4-3a). The data in Figure 4-3b indicated that later timepoints had more differential proteins.

Similarly, principal components analysis of the RNA expression data from a previous study showed segregation both by genotype and age **(Figure 4-3c)** [53]. However, there was also significant technical variation. In the RNA-seq data, there were more differential proteins at an earlier timepoint; however, this effect was far less pronounced **(Figure 4-3d)**. Taking both these types of data into account, further profiling of *Drosophila* phosphoproteomics was done at a later timepoint (10 days) and with more technical replicates to account for the large technical variation seen in previous studies.

4.2 *Drosophila* genetic screen for induction of age-dependent neurodegeneration is enriched for Alzheimer's related phenotypes

Drosophila models have increasingly been used to study human neurodegenerative diseases. An unbiased approach to understanding the mechanisms underlying neurodegeneration is to perform genetic screens for the maintenance of neuronal viability without regard to previous annotations of gene function. This unbiased approach can then allow the identification of novel genes. Many previous *Drosophila* screens have focused on phototaxis defects and abnormalities in retinal function; however, these approaches do not directly assess neurodegeneration. Therefore, in this study, our collaborators performed a forward genetic screen of 2,304 RNAi lines to directly examine RNAi knockout animals for neurodegeneration after a period of growth from histological samples **(Figure 4-4)**. In particular, our collaborators examined defects that lead to vacuole formation in brains, since the formation of vacuoles have been previously linked to neurodegeneration in *Drosophila*.

Examples of some genes identified from this unbiased approach are shown in **(Figure 4-4b)**. In these cases, the knockout of heat-shock protein (*Hsf*), a nuclear matrix protein (*Chmp1*), and MAPK pathway protein (*Tao1*) all lead to vacuole formation, aggregates of ubiquitinated proteins, and neurofibrillary tangles (Figure 4-4). Other genes with human homologs relevant to AD included the amyloid precursor protein *APP* and a component of the γ-secretase complex, *PSEN1*, both of which have been previously shown to be critical to Alzheimer's disease. Of the 202 *Drosophila* genes identified to cause neurodegeneration, 61 were also genes previously implicated in Alzheimer's disease (p-value $< 10^{-30}$) [12]. Next, we were interested in examining the proteins identified from a pathway level in order to see which *Drosophila* neurodegeneration pathways were also relevant to AD. Some of the top enriched gene ontology terms included Iκ phosphorylation (p-value $< 10^{-7}$), negative regulation of Wnt signaling (p-value $< 10^{-6}$), and positive regulation of RNA polymerase II

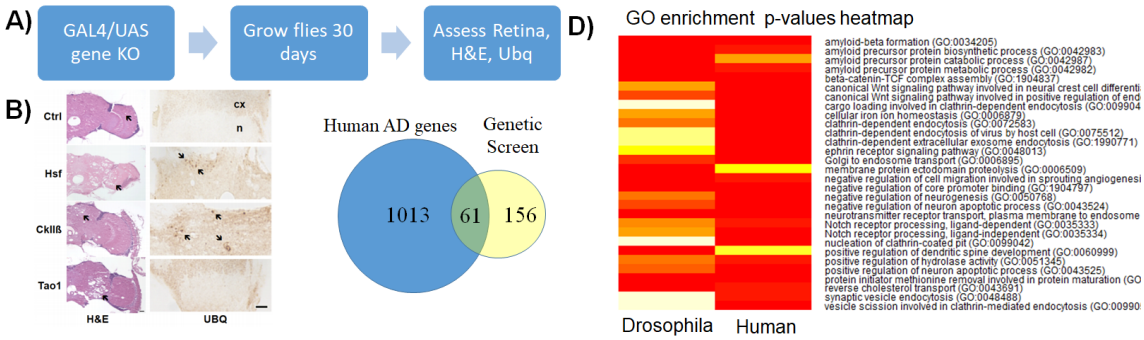

Figure 4-4: **(A)** Outline of genetic screen in for rescue in AD *Drosophila* models. RNAi lines were created using the GAL4/UAS system. Flies were then grown for 30 days and assessed for retinal dysfunction, vacuole formation, and amyloid-β and tau aggregates using HE staining and antibody staining for ubiquitination. **(B)** Example of vacuole formation and tau aggregates in H&E staining, and amyloid-β plaques assessed using antibody staining for ubiquitination in control and three knockout lines. **(C)** Overlap between human AD genes determined from genetic hits compiled in the OpenTargets database, and the human homolog of *Drosophila* genes identified in the genetic screen. **(D)** Heatmap of GO enrichments for pathways identified in both *Drosophila* and human data, with darker red colors representing higher enrichments.

(p-value $< 10^{-7}$). These enrichments overlapped with many of Alzheimer's relevant gene ontologies, again demonstrating the strength of this approach in revealing genes relevant to AD **(Figure 4-4 d)**. Performing a further GO enrichment for genes found in the genetic screen but not in previous annotations for AD was then used to determine genes and pathways that are potentially novel. Some of the top enrichments included JUN phosphorylation (p-value $< 10^{-6}$), Iκ phosphorylation (p-value $< 10^{-6}$), and phosphorylation of RNA polymerase II (p-value $< 10^{-6}$), which have all been previously implicated in Alzheimer's disease [55, 56].

4.3 *Drosophila* metabolomics show dysregulation of lipid metabolism

In order to provide a complementary view of AD disease-related neurodegeneration pathways, we investigated changes in the abundance of lipids, polar metabolites, and non-polar metabolites in tau and Aβ transgenic flies. These analyses revealed

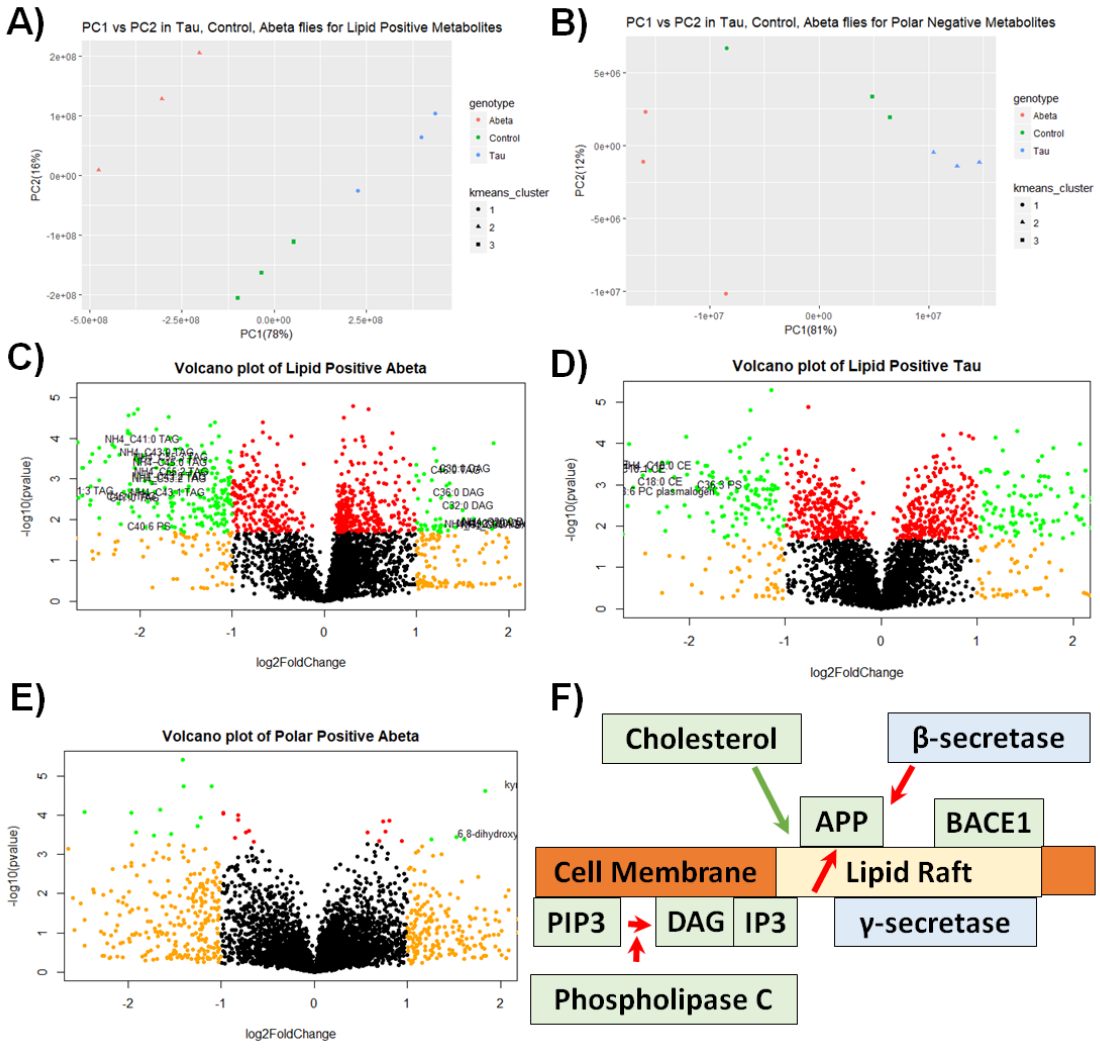

Figure 4-5: Two previously published models of Alzheimer's disease, a humanized Tau model and an Aβ overexpression model, as well as a control fly line were grown for ten days. Approximately 40 whole fly heads were then collected in triple biological replicates for each genotype, and untargeted positively and negatively charged polar and non-polar metabolites were assessed using mass spectrometry. **(A)** First two principal components of lipid positive metabolites were plotted. K-means analysis (k=3) was also performed to cluster the samples in an unbiased manner. **(B)** First two principal components of polar negative metabolites were plotted. K-means analysis was also performed to cluster the samples in an unbiased fashion. **(C)** Log fold changes vs. -log(p-values) were plotted for metabolites in lipid positive Aβ flies compared to control. Metabolites with fold changes with magnitude greater than two are shown in orange, metabolites with FDR < 0.1 are shown in red, and metabolites with fold changes with magnitude greater than two and FDR < 0.1 are shown in green, with annotated metabolites labelled. **(D)** (C) for Tau flies **(E)** (C) for polar positive metabolites. **(F)** Outline of lipid metabolites relevant to AD. Cholesterol and cholesterol esters upregulate APP, γ-secretase, and BACE1 recruitment to lipid rafts, leading to APP cleavage and increased soluble Aβ protein. Phospholipase C also cleaves PIP2 into DAG and IP3 [57].

fundamental dysregulation of lipid metabolism in AD, with many changes being consistent with previous findings. Both the lipid and polar metabolites were generally able to cluster the flies by genotype quite well **(Figure 4-5a,b)** both in principal components space and through an unsupervised k-means approach. This indicates not only that Alzheimer's disease leads to a very different metabolic profile from controls, but that Aβ and tau lead to distinct changes in metabolism. More specifically, the lipid negative and polar negative metabolites did not have annotated metabolites with both significant fold changes and p-values **(Figure A-13)**.

In Aβ flies, diacylglycerides (DAGs) are generally upregulated, which is consistent with the hypothesis that beta-amyloid plaques upregulated phospholipase C, which in turn breaks down PIP to IP and DAG. By contrast, DAGs were not significantly changed in Tau transgenic flies. Decanoate was also down in Abeta flies only, with decanoate acting as a non-competitive AMPA receptor antagonist [58].

Interestingly, sphingolipids were not significantly different between disease models and controls. Ceramide and sphingosine lipid levels were similar between Abeta and control flies (FDR = 0.976, 0.217, as well as between tau and control flies (FDR = 0.873, 0.689). However, ceramide and sphingosine have been shown to be involved in lipid rafts [59, 60], with lipid rafts being associated with pathogenic APP processing and BACE1 localization [61, 62]. In particular, ceramide levels have previously been found in elevated levels in AD, and also regulates BACE1-mediated processing of APP [63, 64]. In our genetic screen, the Drosophila homolog of *PSEN1* was isolated, *PSEN1* being a component of the gamma secretase complex that is associated with lipid rafts. This suggests the possibility that Alzheimer's disease may not lead to the production of more components of lipid rafts, but may lead to the recruitment of AD-related proteins to existing lipid rafts. However, without more detailed targeted metabolomics data and experiments testing lipid localization, we cannot prove this hypothesis.

One of the biggest surprises were that revealed that C18:1 CE and C18:0 CE were down only in Tau flies compared to control (FDR = 0.0280, 0.0286 and log2(fold change) = -2.84 and -1.92). In addition, the NH4 adducts for these esters, with (FDR

= 0.0254, 0.0152 and log2(fold change) of -2.02 and -1.74), suggesting that these reflected a true effect. This is surprising given that cholesterol depletion decreases the association of BACE1 with lipid rafts, which correlates with decreased amyloidogenic processing of APP [62]. Indeed pharmacological inhibition of ACAT1, an enzyme that forms cholesterol esters from cholesterol, has been shown to lead to the reduction of both amyloid-beta and cholesteryl esters [65, 66, 67]. The overall known components of this pathway that were identified in either the genetic screen or through this assay are summarized in **(Figure 4-5f)**.

One of the polar metabolites that were identified in this assay was kynurenic acid, a neuroactivate derivative of tryptophan was has previously been implicated in other neurodegenerative disorders like Huntington's disease [68]. It has also been previously implicated in preventing the aggregation of Aβ and has been shown to protect against Aβ toxicity in *C. elegans* [68]. Therefore, it might be interesting to follow up upon for further analyses.

However, the weakness of this approach was that it did not fully leverage the untargeted metabolites identified in this assay. Moreover, it did not exploit the connections between metabolites through enzymatic reactions. In order to fully leverage these data, we will assay proteomics and phosphoproteomics in order to perform integrative network analyses. Finally, it is important to note that these experiments were conducted in *Drosophila* models of AD, and that these fly models may not fully capture AD biology. Therefore, comparing these results to data from human patients is also extremely important for validation.

4.4 RNA-seq from temporal cortex neurons in human AD patients highlight the role of cellular energetics in Alzheimer's disease

The genetic screen and metabolomics collected in the previous sections provided a perspective on conserved mechanisms of neurodegeneration. However, in order to

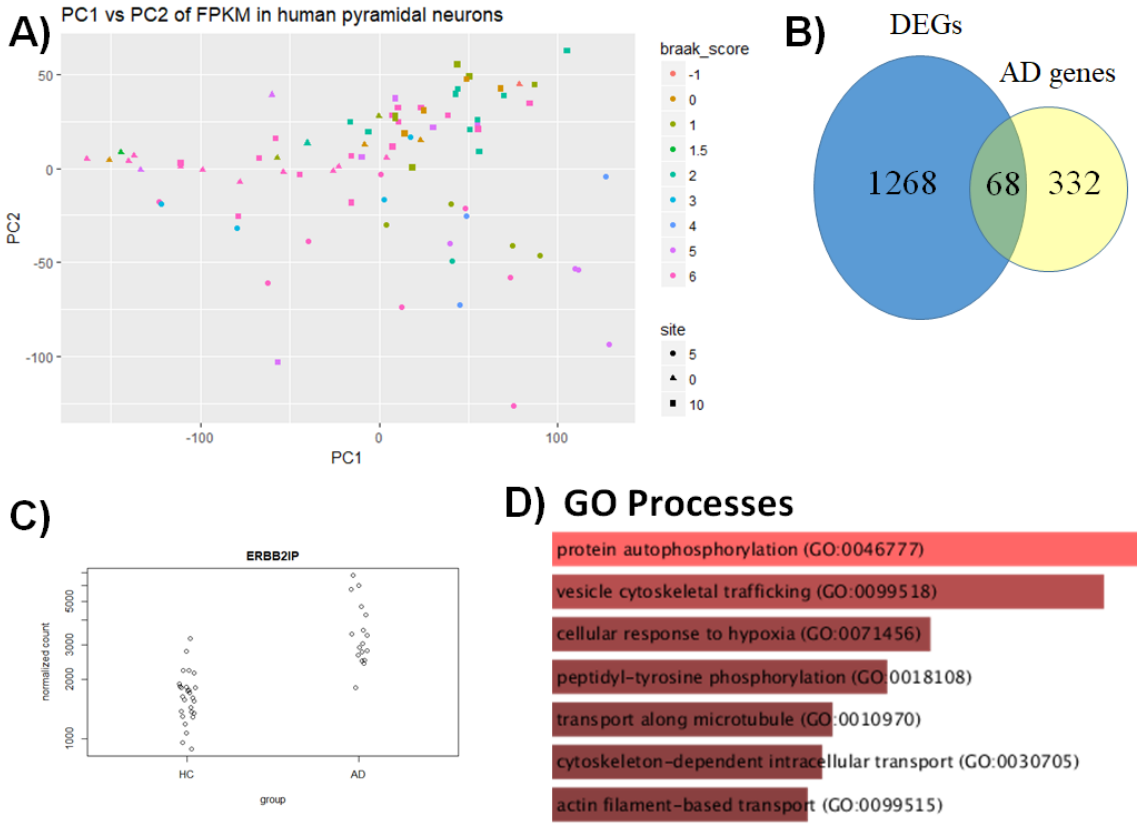

Figure 4-6: RNA-seq was collected from pyramidal neurons of the temporal cortex of 86 Alzheimer's patients and healthy controls. The data was then FPKM normalized for principal components and clustering analysis, while the raw counts were used to determined differential gene expression. **(A)** PC1 vs PC2 of RNA-seq data. Patient Braak scores which represent the formation of tau tangles are used to color nodes, with -1 indicating no staging information available. The shape of the nodes represent the batches in which the data were collected. **(B)** Overlap between differential expressed genes from this analysis and previously annotated AD genes derived from the OpenTargets database. **(C)** Dotplot of gene expression of the top differential genes, ERBB2IP. **(D)** Gene ontology enrichments for differentially expressed genes.

assess the functional implications of these conserved mechanisms of neurodegeneration in humans, we need to determine which of these pathways are linked to AD. The most compelling line of evidence would be genetic evidence linking genetic alterations to certain pathways. However, current GWAS studies both have revealed broad association peaks that span dozens of genes and that contain many variants in non-coding regions [69]. Therefore, it is important to link these genetic changes to functional changes in RNA expression through eQTL analysis. Unfortunately, current studies collect eQTL data from brain homogenates, which are extremely noisy. Not only are there a variety of non-neural cells present in these analyses, but also the presence of a variety of different neurons, only a subset of which contain disease relevant signal [70]. Moreover, changes due to disease can lead to changes in the number of glial cells, further washing out true signal [71].

AD shows a preference for certain brain regions and neuron types [72]. For example, previous data has shown that pyramidal neurons of the middle temporal gyrus are preferentially affected in AD [72]. In order to assess the functional changes only in vulnerable neurons, our collaborators performed RNA-sequencing on pyramidal neurons of layer V/VI of the temporal gyrus from 86 AD patients and age-matched controls.

A principal components analysis revealed that RNA-seq data does *not* separate AD patients from controls **(Figure 4-6a)**. Similarly, the variance stabilized FPKM values do not cluster well by diagnosis, batch, or sex **(Figure A-14 and A-15)**. This demonstrates that RNA-seq does not provide evidence for radical changes in cell state in AD, but instead capturing more subtle dysregulation of specific pathways. This is largely consistent with the findings of previous studies; for example, bulk RNA-seq collected from the temporal cortex of AD patients in the Mayo study similarly showed poor segregation by genotype **(Figure A-16)**.

Next, differentially expressed genes (DEGs) were identified using DESeq2. Of the 400 genes with the highest genetic association scores with AD from the OpenTargets database, 68 of these were DEGs (p-value $< 10^{-12}$) **(Figure 4-6b)**. By contrast, DEGs in a whole brain AD study only contained 23 of the 400 genes with the highest

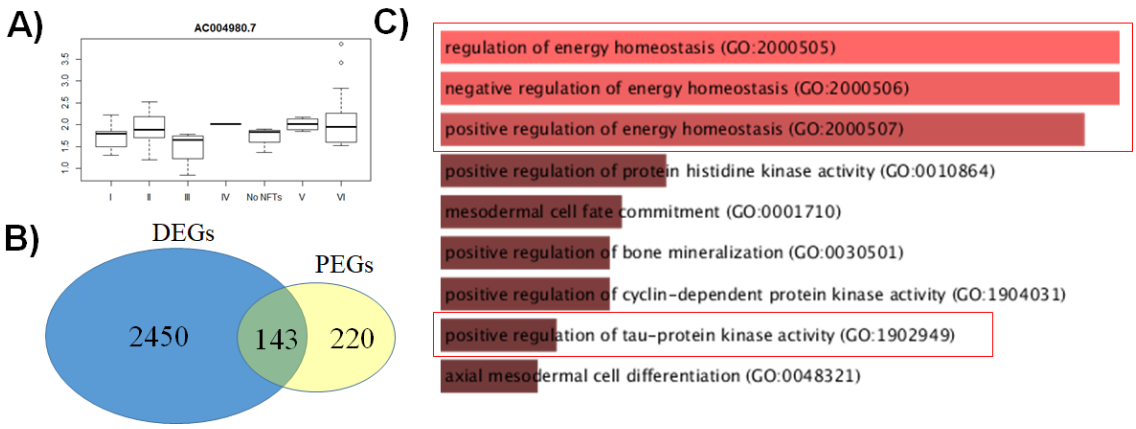

Figure 4-7: Ordinal logistic regression was performed using gene expression data against Braak scores of patients and controls. Braak scores, which measure tau tangle formation, were grouped into low (Braak stages 0, I, II), medium (Braak stages III, IV) and high (Braak stages V, VI) groups. Genes with regression coefficients with robust Z magnitude > 2 were labelled as phenotypically associated genes (PEGs). **(A)** Boxplot of expression of a lincRNA which varies by Braak stage. **(B)** Venn diagram of PEGs and DEGs from RNA-seq. **(C)** Gene ontology enrichemtns of PEGs, with top enrichments indicating regulation of energy homeostasis and tau-protein kinase activity.

genetic association scores with AD from the OpenTargets database, highlighting the efficacy of the laser capture approach [73]. An example of a gene with differential expression between AD patients and control, ERBB2IP, is shown in **(Figure 4-6c)**. ERBB2IP is a protein that interacts with ERBB2, a tyrosine kinase mutated in a variety of cancers, and with the Ras/Raf pathways, potentially indicating changes in cell proliferation and cell state [74]. This is confirmed by enrichments for dysregulation of tyrosine phosphorylation and protein autophoshorylation, suggesting a dysregulation in cellular signaling in AD. However, this differential expression analysis relies solely on case/control, and fails to leverage the detailed clinicopathological information available for these samples.

In order to leverage the physician graded clinical markers for tau phosphorylation, we identified genes whose expression correlated with Braak staging, which measures the formation of tau protein [75]. For example, a gene with expression that varied across stages is shown in **Figure 4-7a**. To this end, we performed ordinal linear regression between Braak stages and gene expression to identify phenotype associated genes

(PEGs). This approach has previously been shown to be efficacious at identifying PEGs in Huntington's disease [76]. There was significant overlap between differentially expressed genes (p-value $<10^{-7}$) **(Figure 4-7b)**. Performing gene ontolgoy enrichment analysis on this data identified relevant pathways affected in PEGs. Ordinal regression was able to better recapitulate pathways previously implicated in Alzheimer's disease, especially the up-regulation of tau-protein kinase activity. However, one of strongest signals in the PEGs was for the dysregulation of energy homeostasis. There have been very few previous studies on the relationship between cellular energetics and AD, so this could represent a promising area for further work [77].

4.5 Network analysis of a *Drosophila* genetic screen of neurodegeneration reveals known and novel AD genes and pathways

Next, integrated network analyses were performed to connect mechanisms of neurodegeneration in *Drosophila* to AD pathogenesis in humans. In order to explore the efficacy of such an network approach, I used PCSF to assess pathways involved in neurodegeneration using the forward *Drosophila* screen described previously.

In order to do so, I tried mapping all genes identified in the forward *Drosophila* genetic screen to their human homologs using the Homologene tool from NCBI [78]. I then assigned a prize of one to all *Drosophila* genes that had a human homolog. Then, I ran the PCSF algorithm using the iRefWeb interactome, using the heuristics described in chapter two to identify parameters, ran 100 randomization each for robustness and specificity, and took the union across all networks for robust (> 0.8) and specific nodes(< 0.2). I then performed Louvain clustering and GO enrichment with BiNGO [43].

This analysis revealed some processes previously identified in neurodegenerative diseases, including protein trafficking and degradation, DNA damage and repair, and synaptic neurotransmission **(Figure 4-8)**. Moreover, this analysis identified some

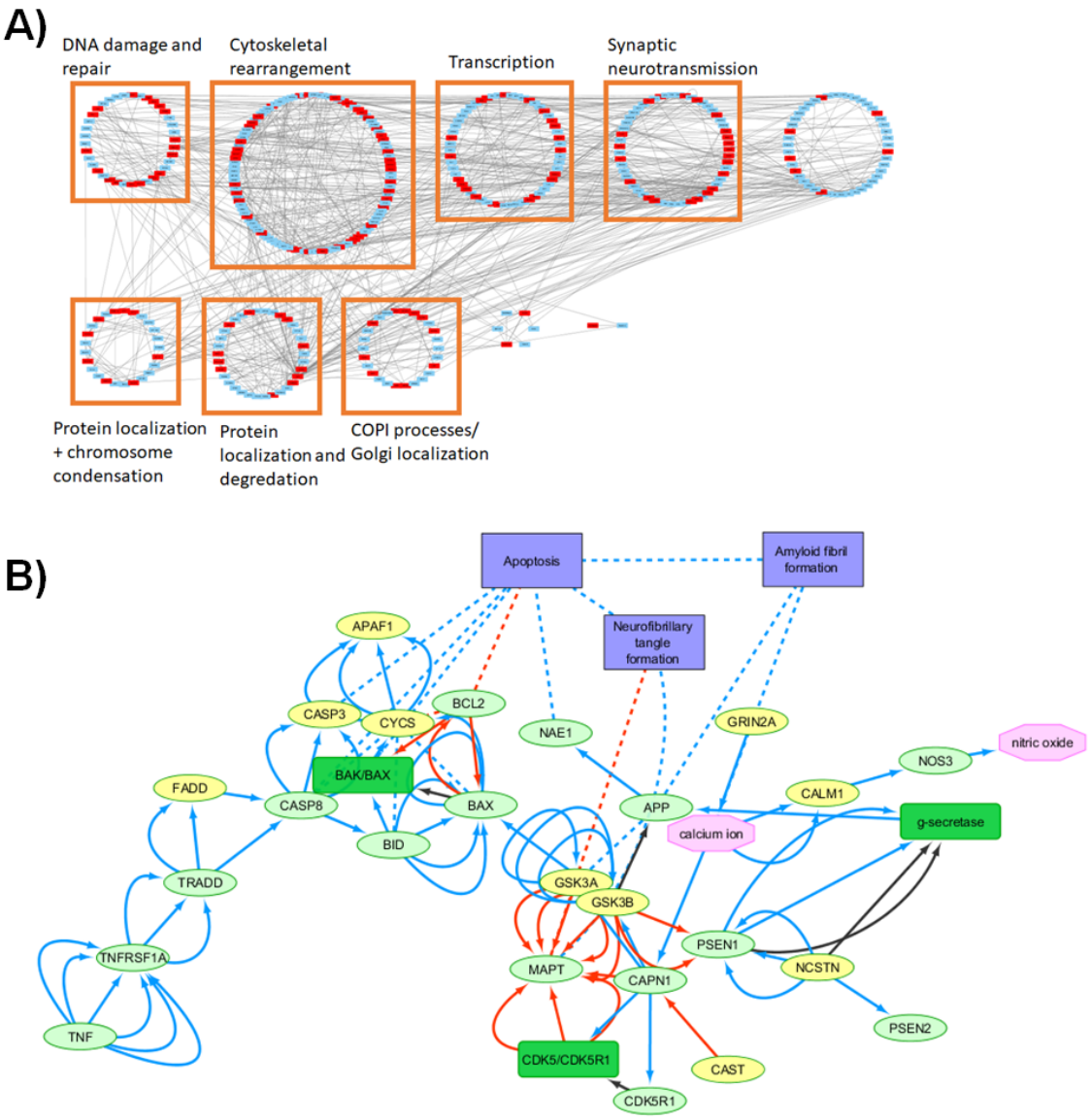

Figure 4-8: (A) Integrated network analysis of *Drosophila* genetic screen using PCSF. Prizes were derived from the forward screen and the iRefWeb interactome was used as the starting PPI. This representation shows the top enrichments in selected Louvain clusters. Terminals are shown in red and Steiner nodes shown in blue. **(B)** The network from (A) was mapped onto a previously annotated network for Alzheimer's disease relevant proteins. Nodes in light green were identified in (A), while nodes in yellow were not. Metabolites are shown in pink, general cellular processes dysregulated in AD are shown in purple, and complexes are shown in dark green. Edges between nodes represent literature curated molecular interactions, with blue edges representing upregulation, red edges representing downregulation, black edges reprsenting changes in localization, and dotted edges representing indirect edges.

novel subnetworks such as cytoskeletal rearrangement, highlighting the potential of this approach.

Next, I mapped these data onto known existing protein networks for AD from the Signor database **(Figure 4-8b)** [79]. Although only three of the proteins identified in our screen were present in this network (APP, PSEN1, and TRADD), the PCSF approach was able to identify 16/26 of the compounds in the Alzheimer's specific network. In particular, this approach suggests that dysregulation of apoptotic pathways and processing of APP were important components of the neurodegeneration phenotype.

4.6 Conclusions and future work

4.6.1 Conclusions

This study has recapitulated many known genes and pathways implicated in Alzheimer's disease, as well as revealed novel genes and metabolites. It also highlighted the power of the integrated network analysis in the discover of hidden pathways relevant to neurodegenerative disease.

In particular, this study identified two promising metabolite targets relevant to AD. Kynurenic acid is an amino acid derivative that has been shown to be therapeutically relevant in other neurodegenerative diseases and relevant to AD in *C. Elegans* [68]. Another interesting target are cholesterol esters, which were downregulated in transgenic flies expressing high levels of humanized tau, suggesting some kind of immunoprotective feedback loop at early timepoints. The human RNA-seq data also supports the role of cellular energetics in AD, suggesting another promising avenue for future validation work. Finally, the power of the PCSF approach in highlighted in its ability to discover many known components of Alzheimer's disease pathway despite our genetic screen only identifying three homologs of known AD related proteins.

4.6.2 Future work

The incorporation of human genetic information, as well as proteomic and phosphoproteomic data from *Drosophila* models for Alzheimer's will allow us to confirm the findings from our current analysis. and to better integrate information across species.

Whole genome data from the AD patients and controls whose RNA-seq data were profiled will allow for the identification of expression quantitative trait loci (eQTL). This in turn will allow for the prioritization of genetic variants that have functional implications in vulnerable neurons. Moreover, genotyping will allow for the segregation of patients by APOE status, which has previously been shown to affect the progression of AD.

Collecting proteomic and phosphoproteomic data in *Drosophila* will allow for the reconstruction of signaling pathways persistently disrupted in neurodegeneration. This will provide complementary information by highlighting affected pathways, non of whose individual components are the master regulators identified in genetic screen. Moreover, these findings can be compared to the public proteomics datasets in humans, such as the Banner Brain and Body study, to discover conserved pathways of proteomic dysregulation in neurodegeneration **(Figure A-17)** [80].

Finally, collecting phosphoproteomic and proteomic data will allow the creation of higher quality integrated network analyses. One can first use the PIUMet approach to map m/z peaks from the untargeted metabolomics, then integrate this data with *Drosophila* proteomic data using the PCSF approach. This can provide insight into *hidden* conserved mechanisms of neurodegeneration that emerge from orthogonal assays. These data can then be compared to evidence from the genetic hits in both the *Drosophila* unbiased genetic screen and the eQTL data in humans to prioritize targets relevant to the progression of neurodegeneration in human AD patients.

Appendix A

Supplemental Figures

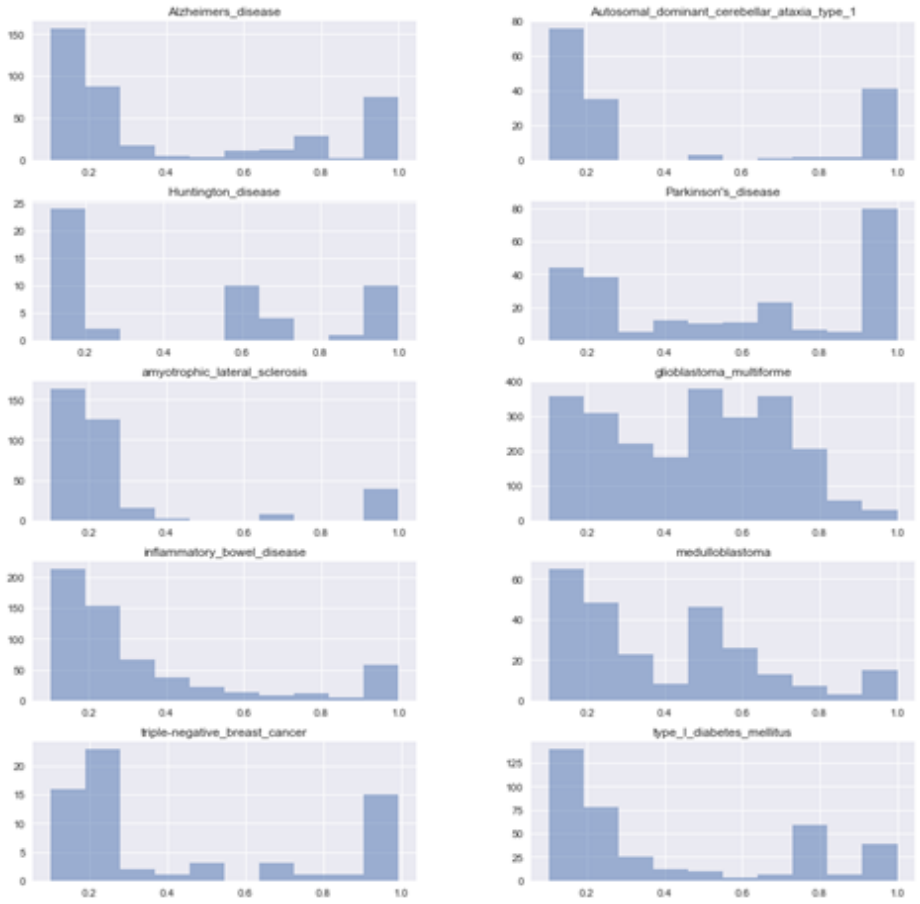

Figure A-1: Histograms of the distribution of genetic confidence scores for each disease from OpenTargets used in creating synthetic datasets. Genetic confidence scores of less than 0.1 were excluded in creating the prizes for synthetic datasets and are excluded in this graph. This separated view clearly highlights the variety of prize distributions encountered in real data.

Figure A-2: Subnetworks constructed with a variety of free parameters for PCSF using synthetic datasets described in B. The Jaccard score was then calculated between a reference of true positive for genetic hits for the disease and with the nodes present in these inferred subnetworks. Parameter sets were ordered by their average Jaccard score. For each parameter and robustness threshold (x-axis), the Jaccard score (color) is plotted for each parameter set (y-axis).

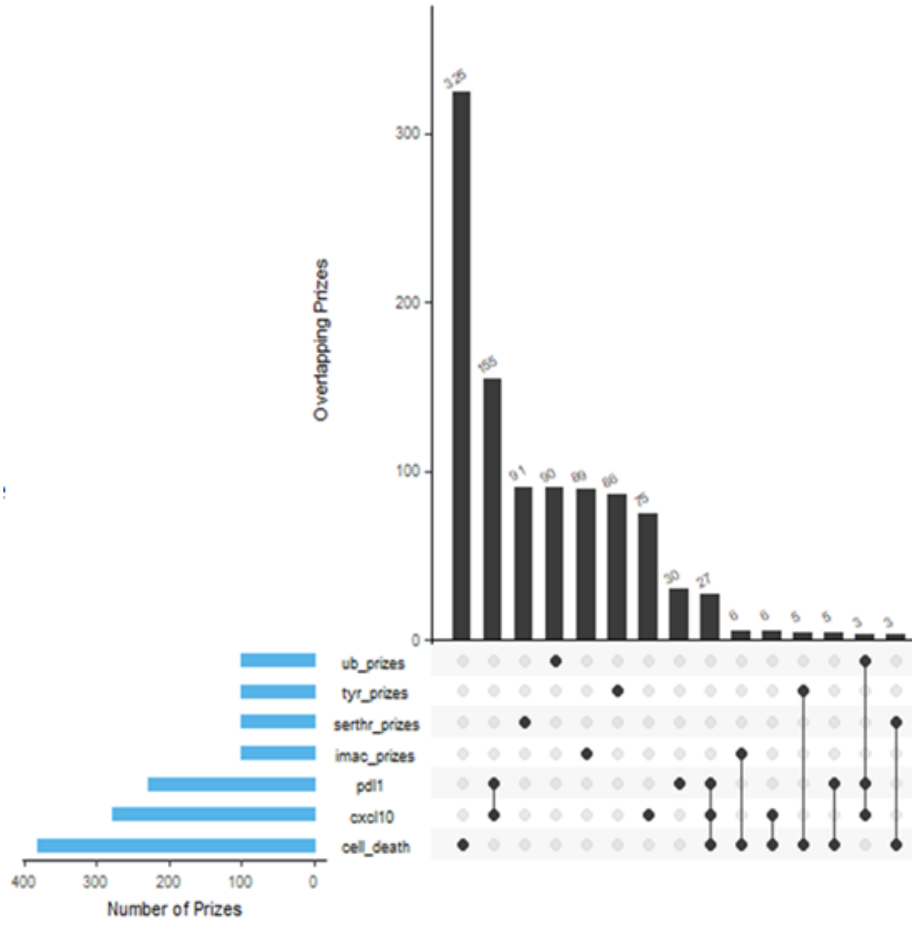

Figure A-3: Overlapping set visualization between different treatment conditions. The smaller barchart in blue shows the number of top hits for each assay (the top 100 hits were used for protein modification assay [mapped to their respective genes]), while genes with a Z-score higher than 2 for the genetic screen were retained. The main barchart shows the overlap between these different assays.

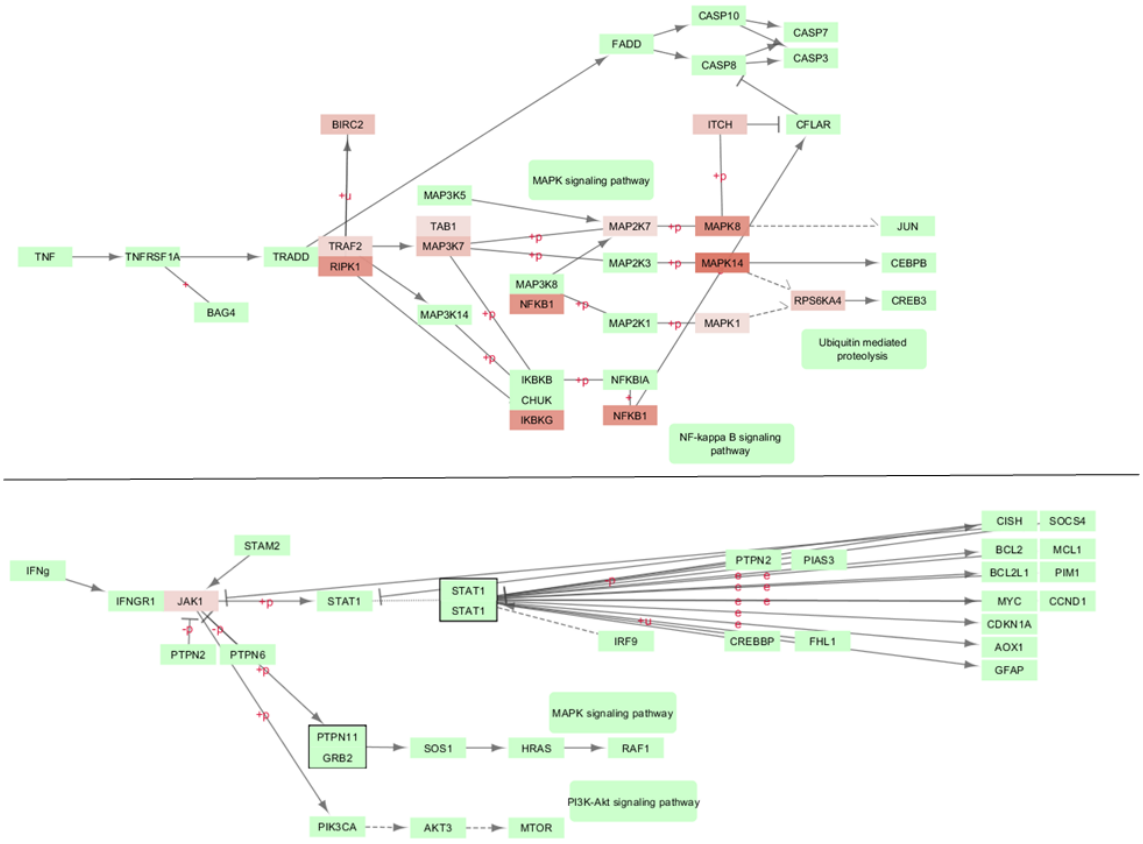

Figure A-4: Protein modifications in the KEGG pathways for TNFα and IFNγ after TNFα treatment. The proteins in red showed a fold change greater than two after TNFα treatment. The top is the KEGG pathway for TNFα signaling and the bottom is a modified KEGG pathway for IFNγ signaling.

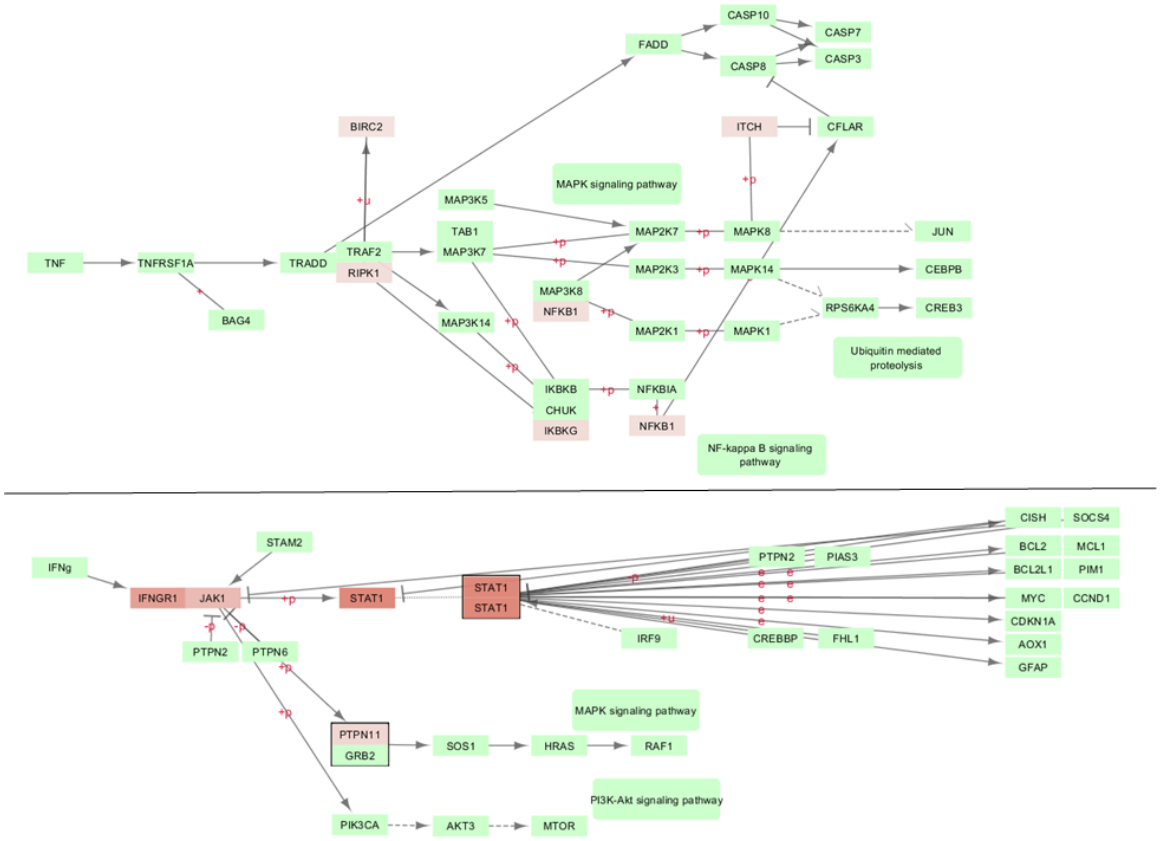

Figure A-5: Protein modifications in the KEGG pathways for TNFα and IFNγ after IFNγ treatment. The proteins in red showed a fold change greater than two after TNFα treatment. The top is the KEGG pathway for TNFα signaling and the bottom is a modified KEGG pathway for IFNγ signaling.

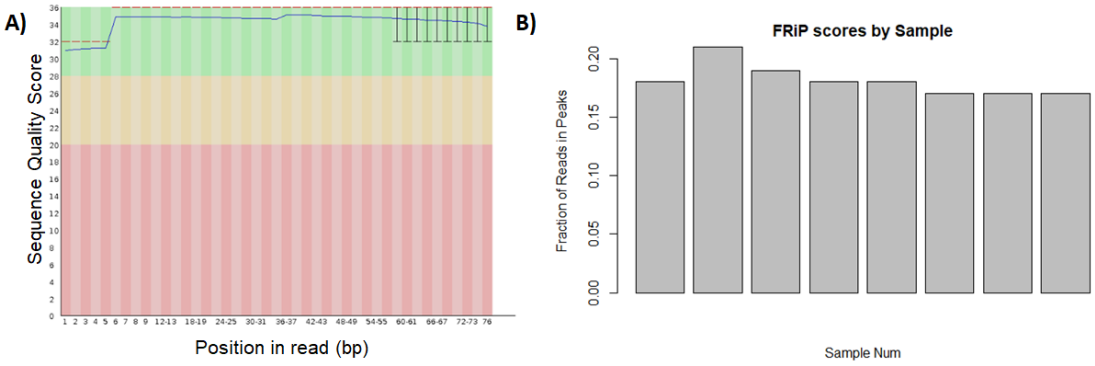

Figure A-6: ATAC-seq overall quality control metrics. **A)** Average sequence quality score by position in read. Sequences with quality scores in the green are considered high quality. **B)** Fraction of reads mapping to peaks (FRiP) for each sample. In general, samples with FRiP scores above 0.1 are considered high quality samples.

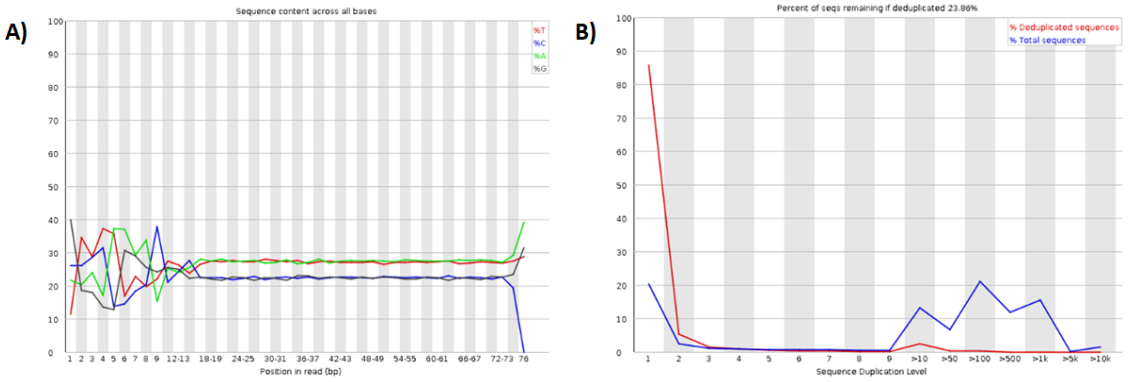

Figure A-7: ATAC-seq read quality control metrics. **A)** Percentage of each nucleobase by position in reads. Note for the first twelve bases, and for the last two bases, the distribution of nucleobases is not uniform, indicating the presence of over-represented sequences. **B)** Percentage of reads by number of repeated reads. The blue is before and the red is after de-duplication. Note that after de-duplication of reads, the vast majority of reads map uniquely, which is what we expect.

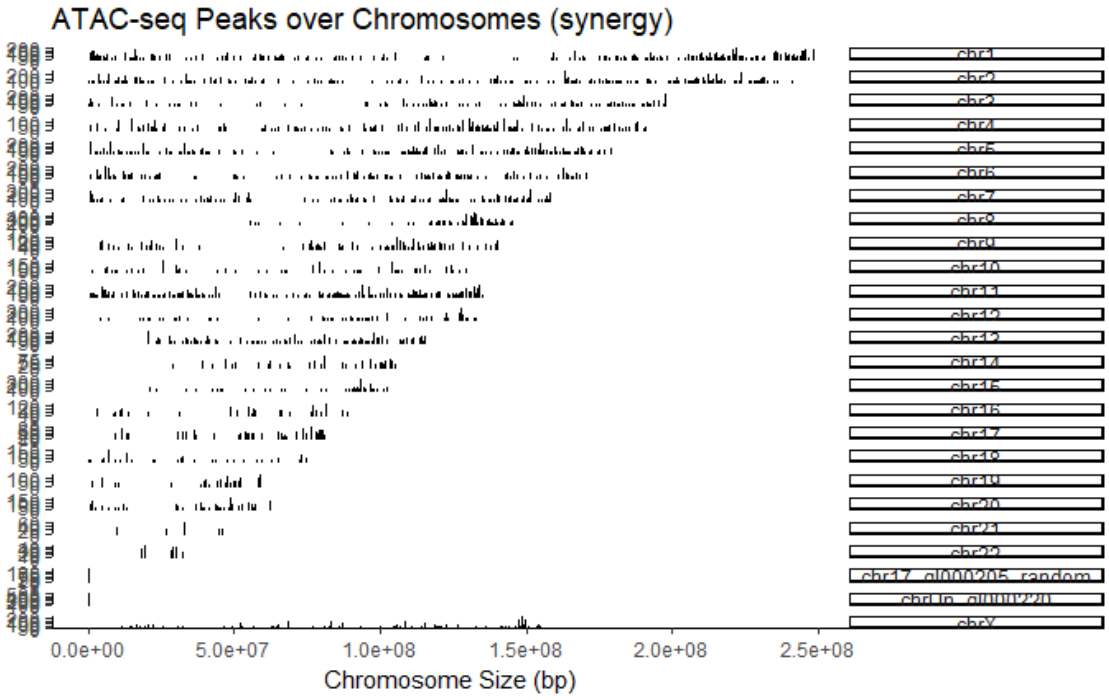

Figure A-8: ATAC-seq read distribution across chromosomes. Reads are represented as peaks in each chromosome, and the chromosomes are arranged in order of length.

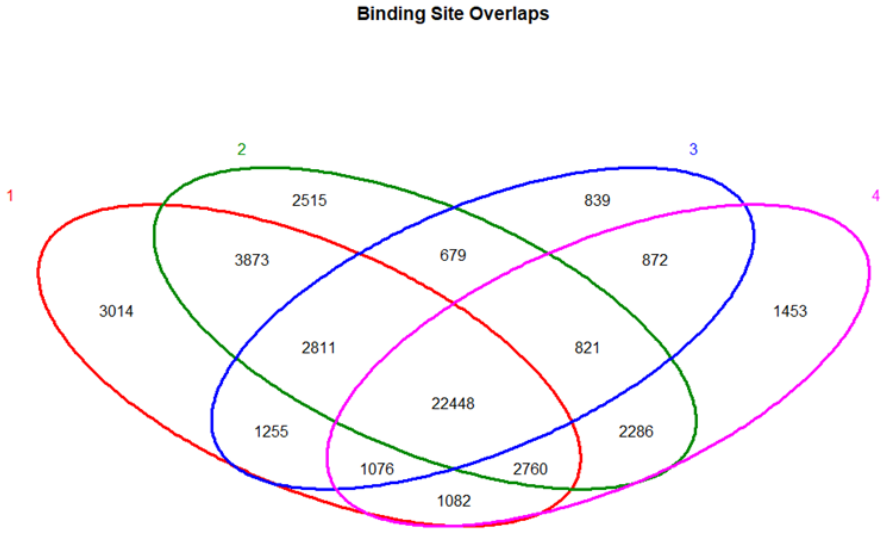

Figure A-9: Venn diagram of overlap between ATAC-seq peaks for each sample found in the consensus peakset. 1 is untreated, 2 is IFNγ treated, 3 is TNFα treated, and 4 is IFNγ + TNFα treated.

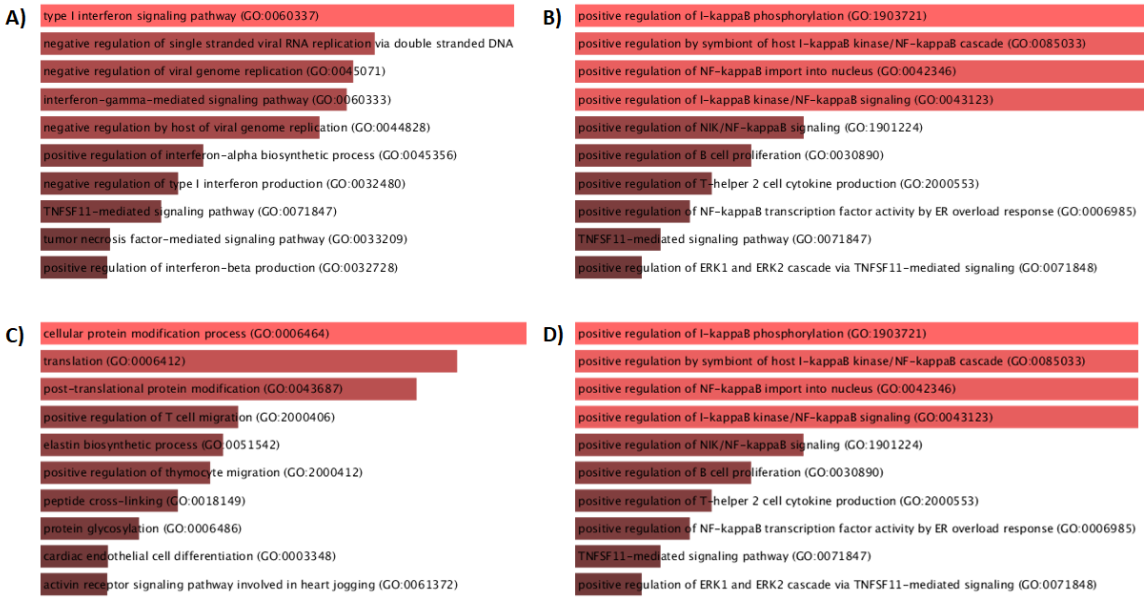

Figure A-10: GO enrichments for differential ATAC-seq peaks near transcriptional start sites.

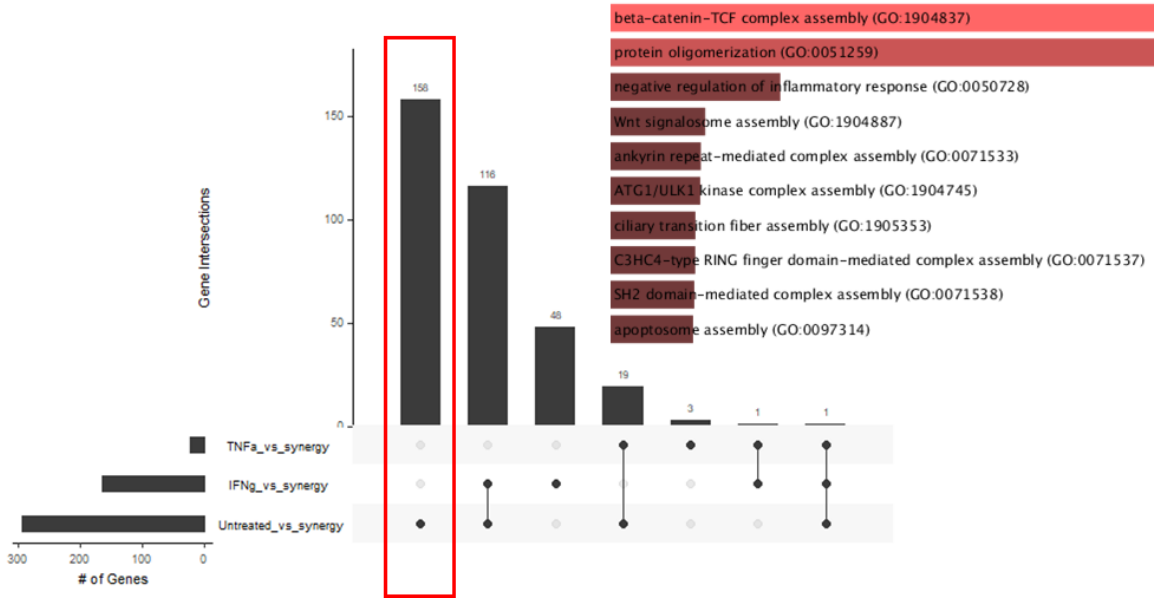

Figure A-11: (Left) Overlapping set visualization between different treatment conditions. The smaller barchart shows the number of genes with peaks near their TSS for each differential peakset, while the main barchart shows the various overlaps between genes near TSS for each differential peakset. The boxed condition represents genes with peaks near their TSS only in the synergistically signalled condition. (Right) GO enrichments genes for the genes boxed in the (left) plot.

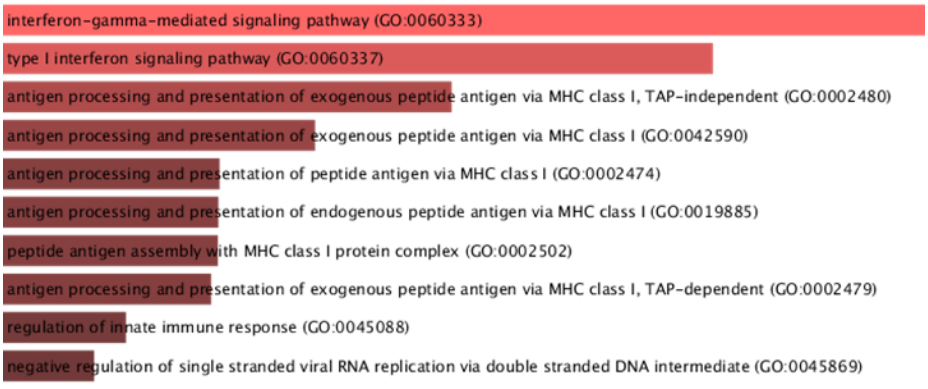

Figure A-12: Gene ontology enrichments for differential proteins (FC $>$ 2 and FDR $<$ 0.2) at 12 hours after treatment with IFNγ and TNFα.

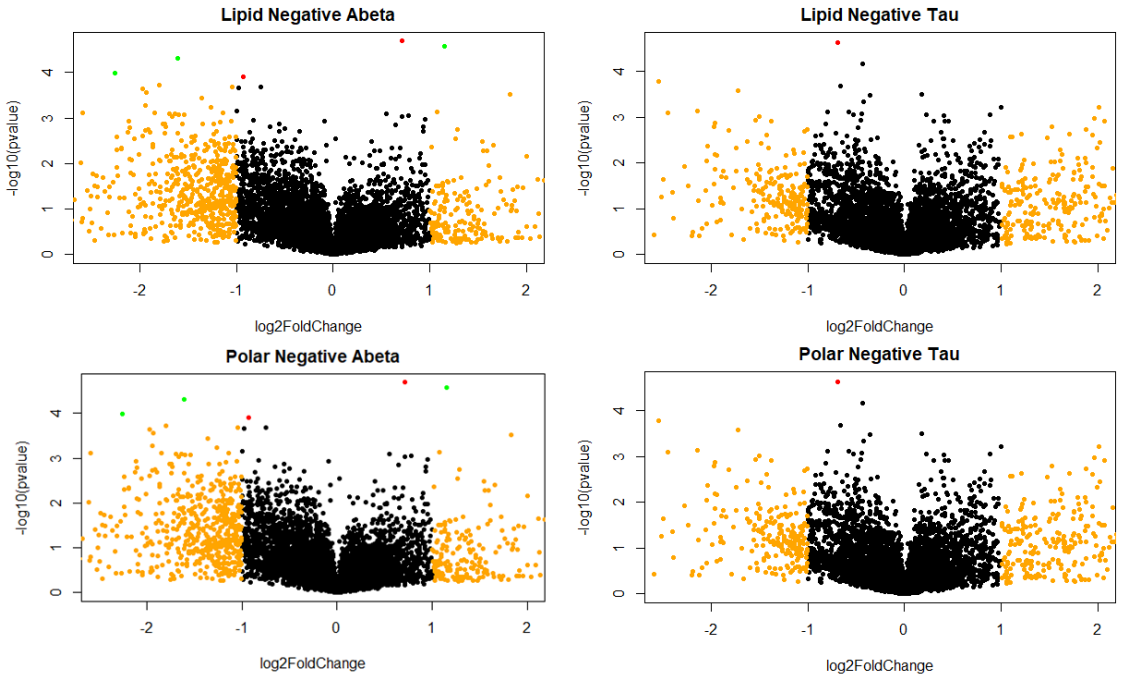

Figure A-13: Volcano plots for negatively charged metabolites in transgenic *Drosophila* models of Alzheimer's Disease. Two transgenic *Drosophila* fly lines and a control line were sacrificed at ten days in three biological replicates, each with approximately forty fly heads. The log 2 fold change between each transgenic line and control is plotted on the x-axis and negative log p-values are plotted on the y-axis. Black dots are m/z peaks with FDR > 0.1 and fold change < 2, yellow dots are m/z peaks with FDR > 0.1 and fold change > 2, red dots are m/z peaks with FDR < 0.1 and fold change < 2, and green dots are m/z peaks with FDR < 0.1 and fold change < 1 (Top left) Aβ transgenic flies in lipid negative mode (Top Right) Tau transgenic flies in lipid negative mode, (Bottom left) Aβ transgenic flies in polar negative mode (Bottom right) Tau transgenic flies in polar negative mode.

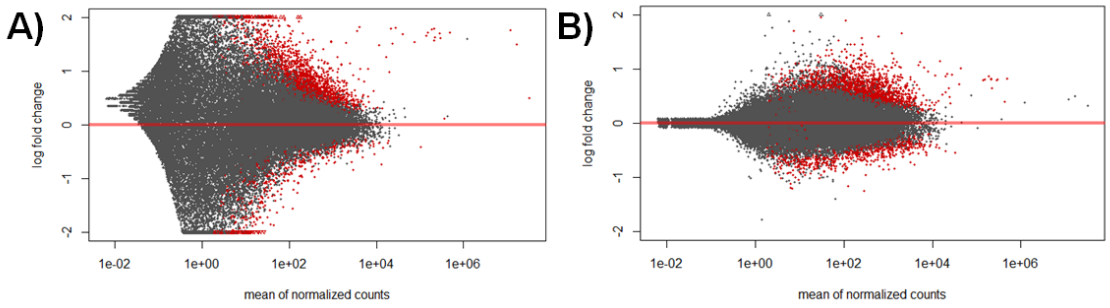

Figure A-14: Mean of normalized reads plotted against log fold change (A) before normalization (B) after variance stabilizing transform normalization.

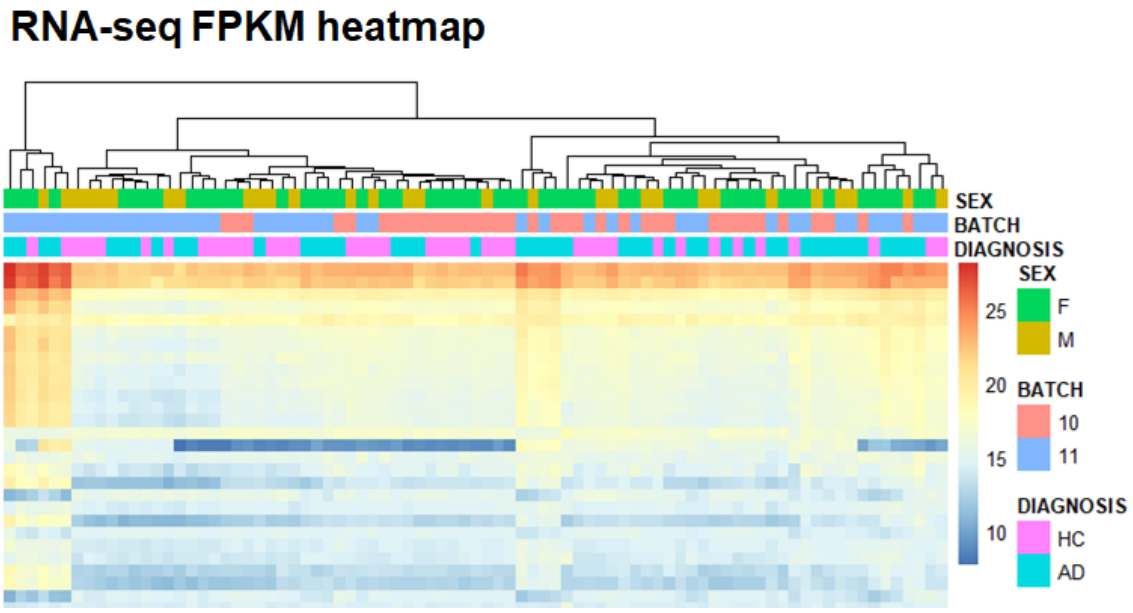

Figure A-15: Heatmap of FPKM of genes with high expression in temporal cortex pyramidal neurons from Alzheimer's patients and controls. Data collection is described in B. Samples are hierarchically clustered. *XIST* has been excluded due to causing a batch effect separating male from female patients.

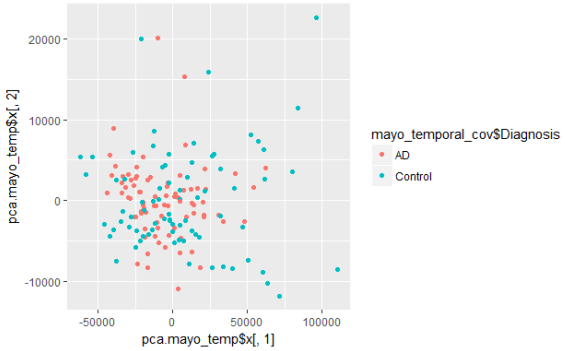

Figure A-16: PC1 vs. PC2 of RNA-seq FPKM values from temporal cortex neurons of Mayo study [73]. Genes have been scaled to mean 0. AD patients are shown in red, and control patients are shown in blue.

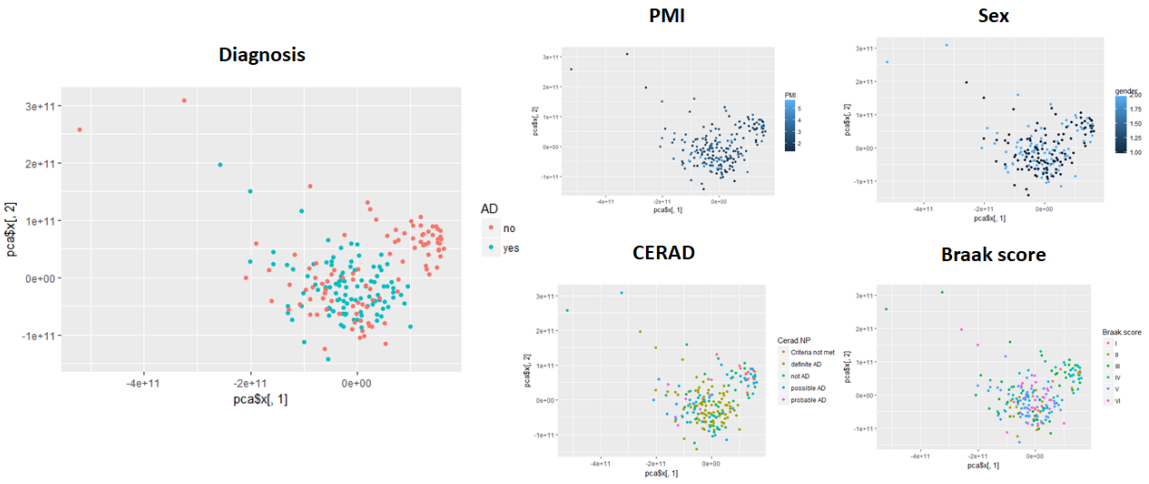

Figure A-17: PC1 vs. PC2 of proteomic expression values from post-mortem tissue of Alzheimer's disease patients and healthy controls of the Banner Brain and Body study [80]. The points have been colored by diagnosis, post mortem interval (PMI), sex, a score for cognitive progression of AD (CERAD), and a score for neurofibrillary tangle formation in AD (Braak score) [75].

Appendix B

Methods

B.1 Constructing and evaluating synthetic datasets

B.1.1 Constructing synthetic datasets

Genetic association scores for Alzheimer's disease, autosomal dominant cerebellar ataxia type 1, Huntington's disease, Parkinson's disease, ALS, glioblastoma multiforme, inflammatory bowel disorder, medulloblastoma, triple negative breast cancer, and type 1 diabetes mellitus were downloaded from OpenTargets [12]. All genes with genetic association scores above 0.1 were considered true positive, and all other genes present in the iRef14 protein-protein interactome were considered negatives [8]. One hundred true positives genes were sampled ten times from each disease. To inject noise, another one hundred non-associated genes were added to each dataset in a degree-matched fashion and were assigned scores to match those of the selected genes. Thus, the final synthetic datasets consisted of ten datasets for each of the ten diseases mentioned above, each with two hundred prizes (100 true positives and 100 noise).

B.1.2 Evaluating synthetic datasets

Subnetwork solutions were calculated for each of the 100 synthetic dataset using OmicsIntegrator [6]. Each synthetic dataset was tested against 250 parameters sets (ω =0.25, 0.5, 1, 2, 5, γ =1000, 2500, 5000, 10000, 25000, 50000, 100000, 250000,

500000, 1000000, β =0.5, 1, 2, 5, 10. One hundred noisy edge randomizations were conducted for each parameter set to evaluate robustness, and one hundred degree-matched randomizations were conducted for each parameter set to evaluate specificity. The nodes in the averaged subnetwork across the solution for each parameter set were used as positives, genes with a genetic association score higher than 0.1 for each disease were used as a true positive, and the rest of the genes in the iRef14 interactome were used as true negatives [8]. The average precision, recall, AUC, and Jaccard score (intersection/union) were then calculated for each parameter set for each dataset for the robustness thresholds 0, 0.2, 0.4, 0.6, 0.8.

B.2 Adding subcellular annotations to PCSF output

The databases for knowledge and experiments "channels" were downloaded from the COMPARTMENTS databases [13]. Terms were then mapped to broad subcellular locations such as 'mitochondria', 'cytoskeleton', 'plasma membrane', and 'nucleus'. Next, for each gene, a score was determined for each cellular compartment by taking:

$$\sum_{t \in T} 2^{s_t} * \delta(t) \tag{B.1}$$

where T is the list of terms for each gene, ranging between 1 and 5, s_t is the score associated with each term, and $\delta(t)$ is an indicator variable for if the term is the subcellular compartment being scored. The most probable subcellular location was then determined for each gene. These annotations were assessed for quality by checking the concordance with physical evidence from antibody staining in the Human Protein Atlas [81]. The full code is available as part of the OmicsIntegrator2 GitHub repository located here: `https://github.com/fraenkel-lab/OmicsIntegrator2/blob/master/subcellular/`.

B.3 Datasets and methods for cytokine synergism study

B.3.1 Datasets

The cytokine synergism data for this project were generously provided by Prof. Nally's group from the University of Cork in Ireland. The datasets included phosphoproteomic, ATAC-seq, proteomic, Affymetrix, and genome wide RNA interference screen data. All data were collected from HT29 cells, a human colon adenocarcinoma model cell line.

Phosphoproteomic data

Cells were divided into three biological replicates, then further divided into four groups: IFNγ stimulated (10 ng/mL), TNFα stimulated (10 ng/mL), and IFNγ + TNFα co-stimulated (10ng/ mL each). Each of these groups were then purified and protease digested 15 minutes post treatment. Modified peptide enrichment were then performed with Fe-IMAC enrichment, ubiquitin remnant K-GG motify antibody (#3925), serine-threonine antibody mix (#25801), and phosphotyrosine pY-1000 motify antibody (#8954). LC-MS/MS analysis was performed using Q-Exactive and SEQUEST was used to identify the peptides. A five percent false positive rate was then used to filter the results. The final dataset contained all peptide modifications that passed this threshold, along with their intensity.

ATAC-seq data

Cells were divided into two biological replicates further divided into four treatment groups: IFNγ stimulated (10 ng/mL), TNFα stimulated (10 ng/mL), and IFNγ + TNFα co-stimulated (10 ng/mL each). Four hours after treatment, ATAC-seq data were collected as described in Buenrostro *et al.* [37]. Paired-end reads were then sequenced using Illumina sequencing at a depth of 40 million reads. Qualtiy control was performed using FastQC, sequences were trimmed using Trimmomatic using

a base quality score of 15, and aligned to the hg19 genome using BowTie2 [82, **?**]. BAMs were then deduplicated, then normalized to reads per kilobase per million using bamCoverage [83]. MACS2 was then used to call peaks [84].

Genetic screens for cell-death, CXCL10, and PD-L1:

Cells were assessed for transfectability with RNAimax (Life Tech.) The Genome-wide ON TARGET-Plus pooled siRNA library (Dharmacon) was used to target 18301 genes. Cells were reverse-transfected with the library at 10nM final siRNA concentration. Forty-eight hours after transfection, all cells except positive and negative controls were treated with IFNγ+TNFα at 10 ng/mL each. Eight hours after cytokine addition, CXCL10 and PD-L1 expression were assayed using antibodies and quantitative fluorescent microscopy. Forty-eight hours after cytokine addition, cell death was assessed using Cell Titer Glow (Promega). The data were normalized using GeneData software, and robust Z-scores were calculated for each gene.

Proteomics

Cells were three biological replicates of each of the following groups: IFNγ stimulated (10 ng/mL), TNFα stimulated (10 ng/mL), and IFNγ + TNFα co-stimulated (10ng/mL each). At four, eight, and twelve hours after stimulation, total protein were collected and quantified using mass spectrometry. The experiments and sequencing were conducted by DC Biosciences and the samples were processed according to their standard protocol.

B.3.2 Methods

For phosphoproteomics, fold-change was calculated between control and treated samples, and t-tests were used to calculate p-values. Multiple linear regression was performed by using the protein fold change of TNFα treated cells and IFNγ treated cells to predict protein fold change in co-stimulated cells. DiffBind was used to calculate differential ATAC-seq peaks, and ChIPseeker was used to assign annotations

to peaks [85, 86]. The Integrative Genomics Viewer was used to visualize peaks [87]. Enrichr was used to calculate gene ontology enrichments [88].

Network analyses were performed using OmicsIntegrator [6]. Parameters were selected using the heuristics discussed in chapter two of this study. One hundred noisy edge randomizations and one hundred degree matched randomizations were performed to assess robustness and specificity respectively. A consensus network consisting of nodes and edges with robust > 0.8 and specificity < 0.2 were used to construct the final subnetwork. Cytoscape was used to visualize networks. Resulting subnetworks were also clustered by subcellular location and community cluster (Louvain clustering). BinGO was also used to calculate the enrichment of Louvain clusters [43].

B.4 Datasets and methods for Alzheimer's Disease study

B.4.1 Datasets

The Alzheimer's datasets were generated by the lab of Mel Feany from the Department of Pathology at Harvard Medical School (*Drosophila* data) and Clemens Scherzer from the Department of Neurology at Harvard Medical School (RNA-seq data).

Drosophila forward genetic screen

2304 transgenic RNAi lines were constructed as part of the Transgenic RNAi Resource Project (TRIP) [89]. These lines were crossed to an *elav-GAL4;UAS-Dcr* line and aged to 30 days. The brains were then fixed in formalin, sections taken at 4μm thickness, and assessed for neurodegeneration by looking for vacuoles, a common presentation of neurodegeneration in *Drosophila* [90, 91].

Drosophila metabolomics

Two previously published models of Alzheimer's disease, a humanized Tau model and an Aβ over-expression model, as well as a control fly line were grown for ten days

[90, 60]. Approximately 40 whole fly heads were then collected in triple biological replicates for each genotype, and untargeted positively and negatively charged polar and non-polar metabolites were assessed using mass spectrometry. Samples were collected at the Broad Institute in collaboration with Dr. Clary Clish.

Human data

Several cell types, including pyramidal neurons from layer V/VI in the middle temporal gyrus, giant Betz pyramidal neurons from the motor cortex, and dopamine neurons from the substantia nigra were laser captured and profiled with RNA-seq. These cells were derived from 83 AD patients and healthy controls. Data from human AD patients were also analyzed from the Mayo clinic study and the Banner Brain and Body project [73, 80].

B.4.2 Methods

Gene ontology enrichments were performed using Enrichr [88]. T-tests for control against disease genotype were used to calculate p-values and comparison of control against each disease genotype were used to calculate fod changes for the *Drosophila* metabolomics. K-means and PCA were performed on the *Drosophila* metabolomics samples to assess clustering. DESeq2 was used to perform differential expression analysis for the human RNA-seq data [92]. Ordinal linear regression was performed by first grouping samples into low (Braak stages 0, I, II), medium (Braak stages III, IV) and high (V, VI) groups [75]. Gene expression was then regressed against these modified Braak stages using ordinal regression using the procedure outlined in Pirhaji *et al.* [76].

Bibliography

[1] Gosline *et al.* SAMNetWeb: identifying condition-specific networks linking signaling and transcription. *Bioinformatics*, pages 1124–1126, 2015.

[2] Lan *et al.* ResponseNet: revealing signaling and regulatory networks linking genetic and transcriptomic screening data. *Nucleic Acids Res*, 2011.

[3] Tuncbag *et al.* Network modeling identifies patient-specific pathways in glioblastoma. *Scientific Reports*, 2016.

[4] Khurana *et al.* Genome-scale networks link neurodegenerative disease genes to α-synuclein through specific molecular pathways. *Cell Syst.*, 2017.

[5] Pirhaji *et al.* Revealing disease-associated pathways by network integration of untreated metabolomics. *Nat. Methods*, pages 770–776, 2016.

[6] Tuncbag *et al.* Network-based interpretation of diverse high-throughput datasets through the OmicsIntegrator software package. *PLoS Comput. Biol.*, 2016.

[7] Tuncbag *et al.* Simultaneous reconstruction of multiple signaling pathways via the prize-collecting steiner forest problem. *J. Comput. Biol.*, pages 124–136, 2013.

[8] Turner *et al.* iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. *Database*, 2010.

[9] Hajiagahyi *et al.* Prize-collecting Steiner network problems. *ACM Trans. on Algorithms.(IPCO 14)*, pages 71–84, 2010.

[10] Hegde *et al.* A nearly-linear time framework for graph-structured sparsity. *ICML-15*, pages 928–937, 2015.

[11] Northcott *et al.* The whole-genome landscape of medulloblastoma subtypes. *Nature*, pages 311–317, 2017.

[12] Koscielny *et al.* Open Targets: a platform for therapeutic target identification and validation. *Nucleic Acids Res.*, pages 985–994, 2017.

[13] Binder *et al.* COMPARTMENTS: unification and visualization of protein subcellular localization evidence. *Database*, 2014.

[14] Turner *et al.* Cytokines and chemokines: At the crossroads of cell signalling and inflammatory disease. *BBA*, pages 2563–2582, 2014.

[15] Gao *et al.* Loss of IFN-γ pathway genes in tumor cells as a mechanism of resistance to anti-CTLA-4 therapy. *Cell*, 167:397–404, 2016.

[16] Wong *et al.* Tumor necrosis factors alpha and beta inhibit virus replication and synergize with interferons. *Nature*, 323:819–822, 1986.

[17] Bartree *et al.* Cytokine synergy: an underappreciated contributor to innate anti-viral immunity. *Cytokine*, 63:237–40, 2013.

[18] Müller-Hermelink *et al.* TNFR1 signaling and IFN-γ signaling determine whether T cells induce tumor dormancy or promote multistage carcinogenesis. *Cancer Cell*, 13:507–518, 2008.

[19] Wang *et al.* Interferon-gamma and tumor necrosis factor-alpha synergize to induce intestinal epithelial barrier dysfunction by up-regulating myosin light chain kinase expression. *Am. J. Pathol.*, 166:409–419, 2005.

[20] Jablonska *et al.* Priming effects of GM-CSF, IFN-gamma and TNF-alpha on human neutrophil inflammatory cytokine production. *Melanoma Research*, 12:123–128, 2002.

[21] Qiao *et al.* Synergistic activation of inflammatory cytokine genes by interferon-γ-induced chromatin remodeling and toll-like receptor signaling. *Immunity*, 39:454–69, 2013.

[22] Williamson *et al.* Human tumor necrosis factor produced by human B-cell lines synergistic cytotoxic interactions with human interferons. *Proc. Natl. Acad. Sci. USA.*, 80:5397–5401, 1983.

[23] Eguchi *et al.* Apoptosis in autoimmune diseases. *Intern Med.*, 40:275–84, 2000.

[24] Liu *et al.* The emerging role of CXCL10 in cancer. *Oncology Letters*, 2:583–589, 2011.

[25] Wang *et al.* Inflammatory cytokines IL-17 and TNF-α up-regulate PD-L1 expression in human prostate and colon cancer cells. *Immunol Lett.*, pages 7–14, 2017.

[26] Alsaab *et al.* PD-1 and PD-L1 checkpoint signaling inhibition for cancer immunotherapy: Mechanism, combinations, and clinical outcome. *Front Pharmacol.*

[27] Mitchell *et al.* Signal transducer and activator of transcription (STAT) signalling and T-cell lymphomas. *Immunology*, 114:301–312, 2005.

[28] Uruno *et al.* Haematopoietic lineage cell-specific protein 1 (HS1) promotes actin-related protein. *Biochem. J.*, 371:485–493, 2003.

[29] Sohn *et al.* The proteasome is required for rapid initiation of death receptor-induced apoptosis. *Mol. Cell. Bio.*, 2006.

[30] Sullivan *et al.* Epigenetic regulation of tumor necrosis factor alpha. *Mol Cell Biol*, 27, 2007.

[31] Parameswaran *et al.* Tumor necrosis factor-α signaling in macrophages. *Crit Rev Eukaryot Gene Expr*, 2011.

[32] Zaidi *et al.* The two faces of interferon-γ in cancer. *Clinical Cancer Res.*, 2011.

[33] Ohishi *et al.* Tankyrase-binding protein TNKS1BP1 regulates actin cytoskeleton rearrangement and cancer cell invasion. *Cancer Res.*, 77:2328–2338, 2017.

[34] Nakanishi *et al.* IFN-γ-dependent epigenetic regulation instructs colitogenic monocyte/macrophage lineage differentiation in vivo. *Mucosal Immunol.*, 2017.

[35] Yıldırım-Buharalıoğlu *et al.* Regulation of epigenetic modifiers, including KDM6B, by interferon-γ and interleukin-4 in human macrophages. *Front Immunol.*, 92, 2017.

[36] Liu *et al.* Epithelial EZH2 serves as an epigenetic determinant in experimental colitis by inhibiting TNFα-mediated inflammation and apoptosis. *PNAS*, 114, 2017.

[37] Buenrostro *et al.* ATAC-seq: A method for assaying chromatin accessibility genome-wide. *Curr. Protocols*, pages 1–9, 2015.

[38] Kuriakose *et al.* ZBP1: Innate sensor regulating cell death and inflammation. *Trends in Immunology*, 39:123–134, 2017.

[39] Pasparakis *et al.* Necroptosis and its role in inflammation. *Nature*, 517:311–320, 2015.

[40] Annibaldi *et al.* Ubiquitin-mediated regulation of RIPK1 kinase activity independent of IKK and MK2. *Molecular Cell*, 69:566–580, 2018.

[41] Gygi *et al.* Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Bio.*, 19:1720–1730, 1999.

[42] Hall *et al.* Precise probes of type II interferon activity define the origin of interferon signatures in target tissues in rheumatic diseases. *PNAS*, 109, 2012.

[43] Maere *et al.* BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Nucleic Acids Res.*, pages 3448–3449, 2005.

[44] Karran *et al.* The amyloid cascade hypothesis for Alzheimer's disease: an appraisal for the development of therapeutics. *Nat. Rev. Drug Discovery*, pages 698–712, 2011.

[45] Serrano-Pozo *et al.* Neuropathological alterations in Alzheimer's disease. *Cold Spring Harb Perspect Med.*, 2011.

[46] Masters *et al.* Alzheimer's disease. *Nat. Rev. Dis. Prim.*, 2015.

[47] Goate *et al.* Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. *Nature*, pages 704–706, 1991.

[48] Hutton *et al.* Association of missense and 5'-splice-site mutations in tau with the inherited dementia FTDP-17. *Nature*, pages 702–705, 1998.

[49] Hardy *et al.* The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics. *Science*, pages 353–356, 2002.

[50] Medway *et al.* The genetics of Alzheimer's disease; putting flesh on the bones. *Neuropathol. Appl. Neurobiol.*, pages 97–105, 2014.

[51] Norton *et al.* Potential for primary prevention of Alzheimer's disease: an analysis of population-based data. *Lancet Neurol.*, pages 788–794, 2014.

[52] Milman *et al.* Dissecting the mechanisms underlying unusually successful human health span and life span. *Cold Spring Harbor Perspectives in Medicine*, 2016.

[53] Scherzer *et al.* Gene expression changes presage neurodegeneration in a Drosophila model of Parkinson's disease. *Hum. Mol. Gen.*, pages 2457–2666, 2003.

[54] Wittman *et al.* Tauopathy in Drosophila: neurodegeneration without neurofibrillary tangles. *Science*, pages 711–714, 2001.

[55] Thakur *et al.* c-Jun phosphorylation in Alzheimer disease. *Neurosci. Res.*, 2007.

[56] Jiang *et al.* Genetic deletion of TNFRII gene enhances the Alzheimer-like pathology in an APP transgenic mouse model via reduction of phosphorylated I$\kappa\beta\alpha$. *Hum. Mol. Genetics*, 23:4906–4918, 2014.

[57] DiPaolo *et al.* Linking lipids to Alzheimer's disease: cholesterol and beyond. *Nat. Rev. Neurosci.*, 12:284–296, 2011.

[58] Oliveira *et al.* Phospholipase D in brain function and Alzheimer's disease. *Biochim. Biophys. Acta*, pages 899–905, 2010.

[59] Hannun *et al.* Principles of bioactive lipid signalling: lessons from sphingolipids. *Nature Rev. Mol. Cell Biol.*, pages 139–150, 2008.

[60] de Chaves *et al.* Sphingolipids and gangliosides of the nervous system in membrane function and dysfunction. *FEBS Lett.*, page 1748–1759, 2010.

[61] Hartmann *et al.* Alzheimer's disease: the lipid connection. *J. Neurochem.*, pages 159–170, 2007.

[62] Riddell *et al.* Compartmentalization of β-secretase (Asp2) into low-buoyant density, noncaveolar lipid rafts. *Curr. Biol.*, page 1288–1293, 2001.

[63] He *et al.* Deregulation of sphingolipid metabolism in Alzheimer's disease. *Neurobiol. Aging*, pages 398–408, 2008.

[64] Puglielli *et al.* Ceramide stabilizes beta-site amyloid precursor protein-cleaving enzyme 1 and promotes amyloid beta-peptide biogenesis. *JBC*, page 19777–19783, 2003.

[65] Puglielli *et al.* Acyl-coenzyme A: cholesterol acyltransferase modulates the generation of the amyloid β-peptide. *Nature Cell Biol.*, pages 905–912, 2001.

[66] Bhattacharyya *et al.* ACAT inhibition and amyloid *beta* reduction. *Biochim. Biophys. Acta*, pages 960–965, 2010.

[67] Hutter-Paier *et al.* The ACAT inhibitor CP-113,818 markedly reduces amyloid pathology in a mouse model of Alzheimer's disease. *Neuron*, pages 227–238, 2004.

[68] Kampesan *et al.* The kynurenine pathway modulates neurodegeneration in a Drosophila model of Huntington's disease. *Curr. Biology*, 2011.

[69] Lambert *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.*, pages 1452–1458, 2013.

[70] Montgomery *et al.* From expression QTLs to personalized transcriptomics. *Nat. Rev. Genet.*, pages 277–282, 2011.

[71] Franzen *et al.* Cardiometabolic risk loci share downstream cis- and trans-gene regulation across tissues and diseases. *Science*, pages 827–830, 2016.

[72] Arnold *et al.* The topographical and neuroanatomical distribution of neurofibrillary tangles and neuritic plaques in the cerebral cortex of patients with Alzheimer's disease. *Cereb. Cortex*, pages 103–116, 1991.

[73] Bennett *et al.* Overview and findings from the Rush Memory and Aging Project. *Curr. Alzheimer Res.*, pages 646–663, 2013.

[74] Huang *et al.* Erbin loss promotes cancer cell proliferation through feedback activation of Akt-Skp2-p27 signaling. *BBRC*, 2015.

[75] Braak *et al.* Staging of Alzheimer disease-associated neurofibrillary pathology using paraffin sections and immunocytochemistry. *Acta Neuropathol.*, pages 389–404, 2006.

[76] Pirhaji *et al.* Identifying therapeutic targets by combining transcriptional data with ordinal clinical measurements. *Nature Comm.*, 2017.

[77] Zhang *et al.* Altered brain energetics induces mitochondrial fission arrest in Alzheimer's disease. *Sci. Rep.*, 2016.

[78] NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, pages 12–17, 2017.

[79] Perfetto *et al.* SIGNOR: a database of causal relationships between biological entities. *Nucleic Acids Research*, pages D548–D554, 2015.

[80] Beach *et al.* The Sun Health Research Institute Brain Donation Program: description and experience, 1987-2007. *Cell. Tissue Bank*, pages 229–245, 2008.

[81] Fagerberg *et al.* Tissue-based map of the human proteome. *Science*, 2015.

[82] Bolger *et al.* Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, pages 2114–2120, 2014.

[83] Ramirez *et al.* deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res*, 2014.

[84] Zhang *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, 2008.

[85] Stark *et al.* DiffBind: differential binding analysis of ChIP-Seq peak data. 2011.

[86] Yu *et al.* ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics*, pages 2382–2383, 2015.

[87] Robinson *et al.* Integrative Genomics Viewer. *Nat. Biotechnol.*, pages 24–26, 2012.

[88] Kuleshov *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.*, 2016.

[89] Ni *et al.* A drosophila resource of transgenic rnai lines for neurogenetics. *Genetics*, pages 1089–1100, 2009.

[90] Wittman *et al.* Tauopathy in drosophila: neurodegeneration without neurofibrillary tangles. *Science*, pages 711–714, 2001.

[91] Heisenberg *et al.* Isolation of anatomical brain mutants of Drosophila by histological means. *Naturforsch*, pages 143–147.

[92] Love *et al.* Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 2014.