Introduction

Trisomy 21, aka Down Syndrome, includes among its constellation of symptoms, intellectual disability and learning problems. Identifying protein expression differences that identify differences in learning in Down Syndrome brains is an early step in identifying potential targets for treatments for the cognitive effects of trisomy 21. A pharmaceuticals company or academic research lab that is doing basic research aimed at finding potential new treatments for intellectual disabilities can use this to identify new candidate drugs to test.

The problem: identify proteins whose expression in mouse brains is specific to Down Syndrome and the experience of a learning task, particularly for interactions between disease and learning conditions.

The data for this project comes from a study investigating the drug memantine in a mouse model of Down Syndrome (DS). Memantine is a NMDA receptor antagonist used to treat severe Alzheimer's. In spite of activity in a mouse model of DS, it was not found to be helpful in older adults with DS in subsequent clinical studies. The data is from the UCI Machine Learning Repository: http://archive.ics.uci.edu/ml/datasets/Mice+Protein+Expression. Being from the UCI collection, it is freely available and fairly clean.

This data set has 3 independent variables and 76 dependent variables.
The 3 independent variables are:

Genotype: mice are either wildtype or a genetically engineered model of Down Syndrome
Behavior: whether they were exposed to a learning task or a control experience
Treatment: whether they received memantine

The dependent variables are the levels of 76 proteins or protein modifications measured in cell nuclei in the mouse cerebral cortex. This is a subset of a larger data set which included membrane-bound and cytosolic fractions as well as the nuclear fraction and also included all three cellular fractions from the hippocampus.

The proteins were chosen for measurement based on known or hypothesized relationships to DS and/or learning and memory; they are:

- Genes on the human Chromosome 21 (and therefore with extra copies in DS)
- Proteins related to Alzheimer's Disease
- MTOR pathway
- MAPK pathway
- Glutamate receptors and proteins they interact with
- proteins involved in apoptosis or inflammation
- immediate early gene proteins (markers of neuronal activation)
- histone modifications

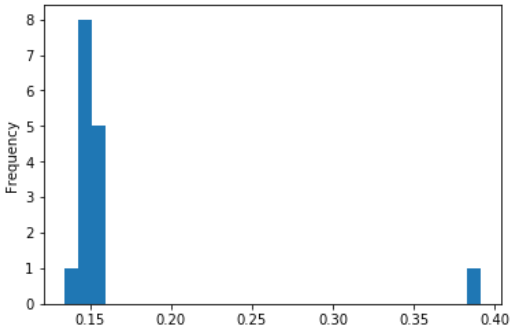
Overall there are 8 groups with 7-10 biological replicates (i.e. mice) per group and 15 technical replicates (measurements) per mouse per protein/modification. Ultimately, this is a supervised categorization problem.

Data cleaning

The numerical portion of the data set was 1080 rows by 77 columns, for a total of 83,160 data points. Of those, 1396 data points were missing (null). I found that two of the columns ('pS6_N' and 'ARC_N') have duplicate values, so I merged these columns into one and relabeled it with both protein names. Thus the data set contains 76 proteins/modifications, not 77, and 82,080 total data points (80,684 non-null).

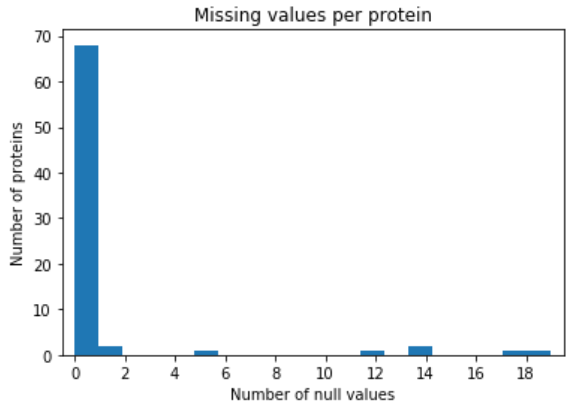
Missing data or outliers can be missing or outlying technical replicates or entire missing/outlying biological replicates. Given the relatively small sample sizes, I chose not to attempt any removal/interpolation of outliers/missing data among biological replicates. However, I felt that extreme outliers among technical replicates was likely to reflect technical error. I defined an outlier as 3 standard deviations away from the mean within that mouse x protein. Only 70 data points (out of 82,080) met this criterion.

Example of the data for a single mouse and single protein, showing one extreme outlier:



I replaced each of these outliers with null values. I then collapsed the data set by mice by taking the means (and also standard deviations) of each set of technical replicates. For all subsequent analysis, I used this collapsed data set, which consists of 76 columns (proteins or protein modifications) and 72 rows (individual mice).

This dataset contains a total of 84 missing values (1.5% of the 76 x 72 = 5472 expected values). These are not evenly distributed; for 68/76 proteins, no data is missing.

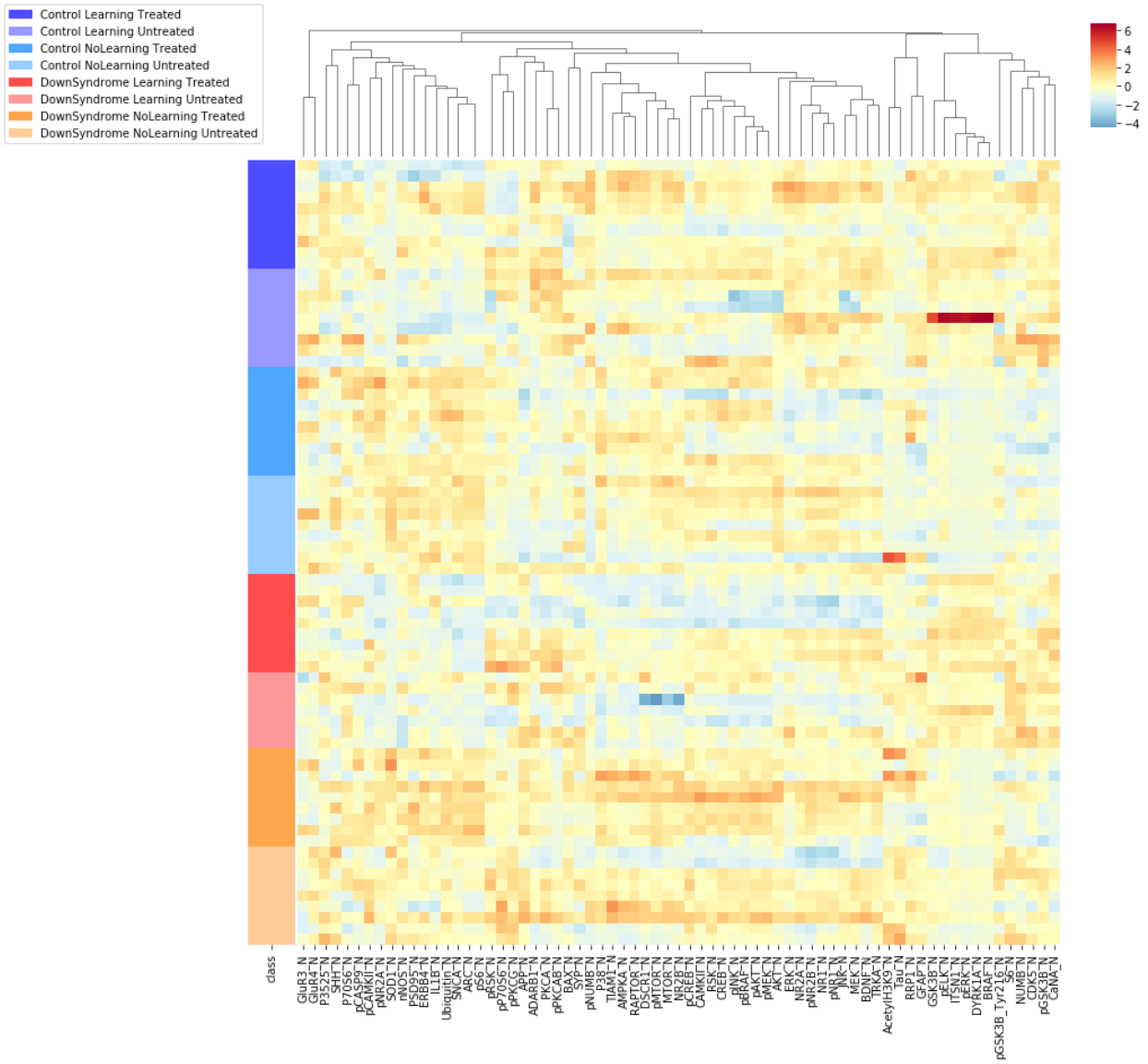


No group is entirely missing data for a protein; the minimum sample size per group is 5.

Exploratory analysis

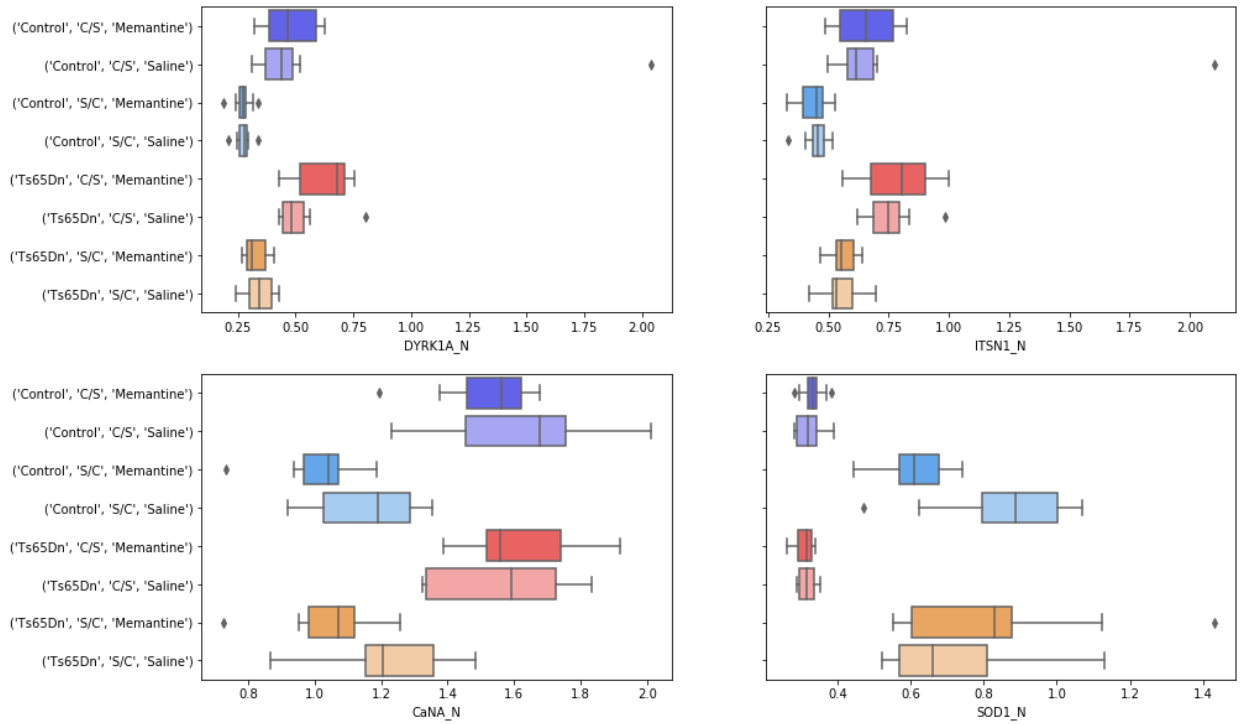
Which proteins' levels in the nucleus rise or fall with learning and/or differ between wildtype mouse brains and those of the Down Syndrome model? Which ones are affected by learning? Can I find interactions between genotype and learning that suggest ways learning differs between wildtype and Down Syndrome mics?

By clustering the proteins and keeping the mouse groups together, I can see that there is an obvious effect of the learning condition on the levels of many proteins:



(I can also see here that there are a few mice with unusually high or low levels of some proteins, levels that may skew analysis especially based on means.) Note that proteins with missing data points (8/76 proteins) are not compatible with this cluster map and are left out.

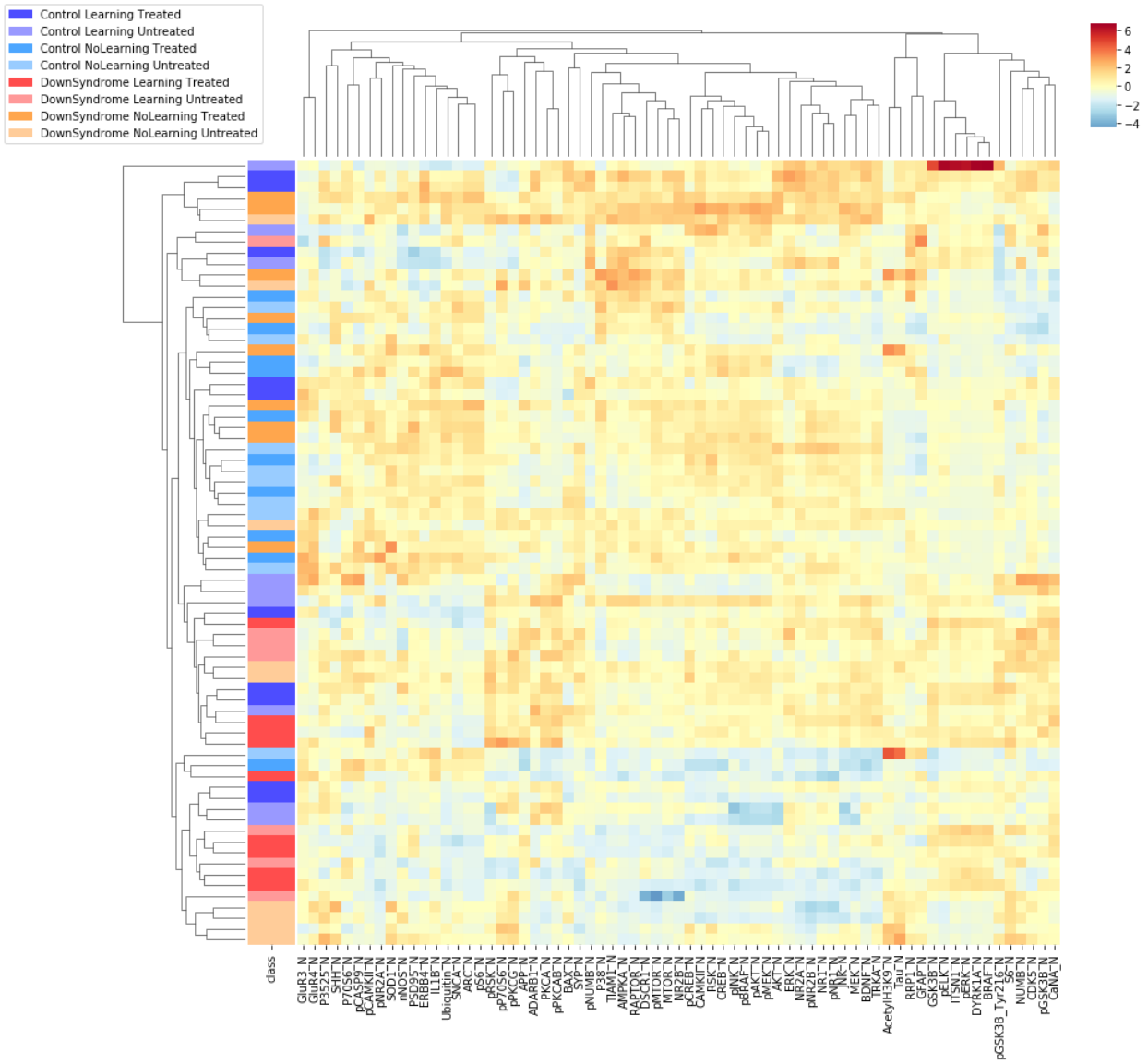
As an example, here are 3 proteins whose levels are up in mice that experienced learning and one that is down (“C/S” represents the learning condition):



Interestingly, the genes for 3 of these 4 proteins (all except CaNA aka PPP3CA) are found on Chromosome 21. Yet their nuclear levels are similar between wildtype and mutant mice and change primarily with genotype, not learning.

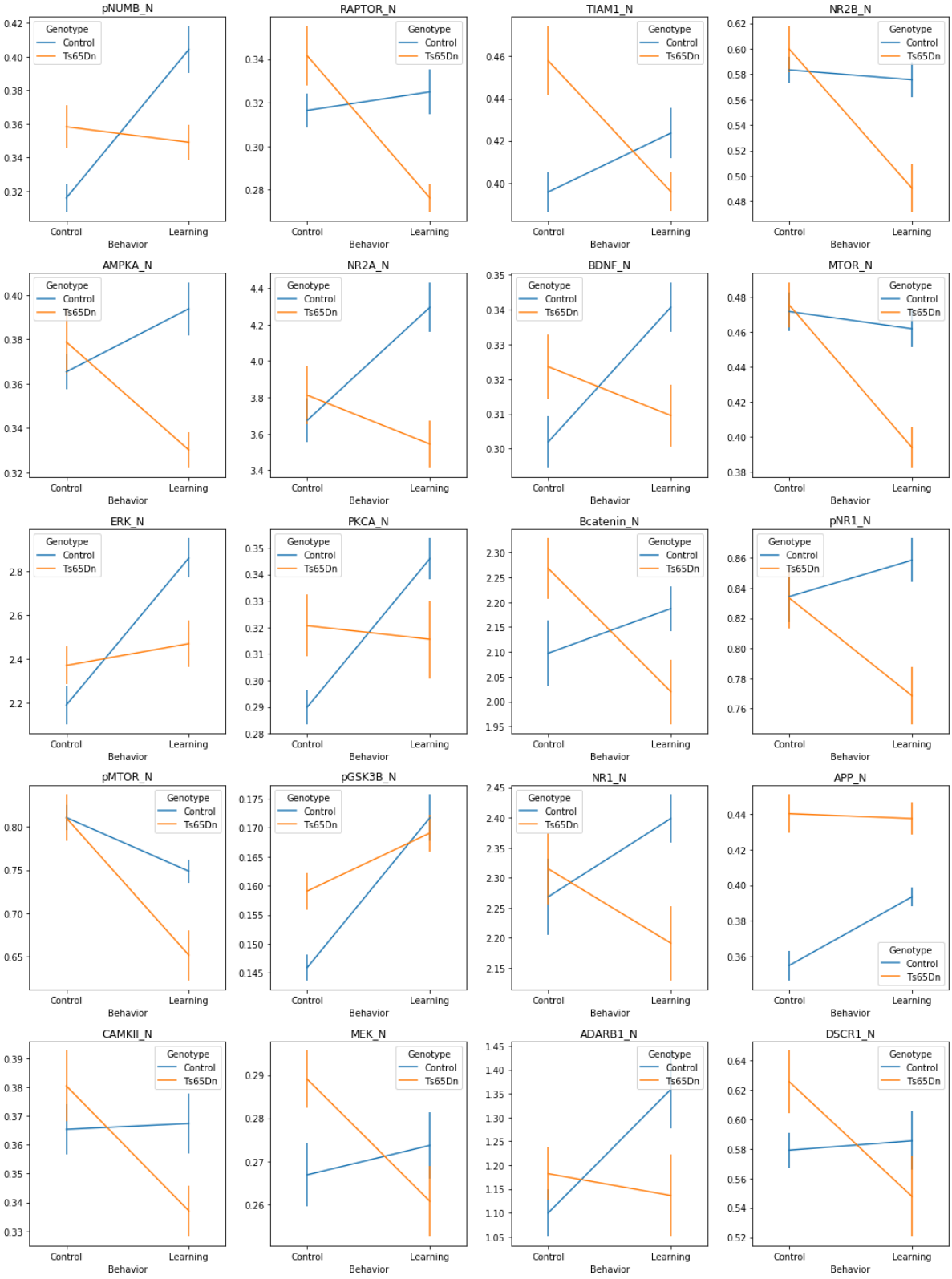
In my initial clustering, I found many proteins that appeared to distinguish the learning condition from non-learning, but I didn't see visually obvious clusters of proteins that distinguished mice based on genotype or treatment.

When I cluster both proteins and mice, there is a cluster that mostly consists of wildtype/no learning mice (light blues), with some members of the DS model/no learning/memantine-treated (bright orange) group. That is the clearest cluster, though DS model/learning tends to cluster with DS model/no learning/untreated, suggesting a specific effect of memantine on DS mice not experiencing learning. In spite of the effects of the learning condition on many individual proteins, mice from the wildtype/learning groups are scattered in multiple clusters.



I am most interested in the interaction between genotype and behavior. That is, how does a mouse with Down Syndrome respond to a learning task differently than a wildtype mouse? This is not an effect that is visually obvious from the cluster analysis.
I performed 3-way ANOVAs for each protein to identify effects of each variable and their interactions. As was visible in the data story, many proteins were affected by the behavior paradigm (learning vs no learning), while few were affected by either genotype (down syndrome model) or the drug memantine.

To find proteins that represent the difference between learning in the DS model and learning in wildtype mice, I sorted the proteins by the p-value of the interaction between genotype and behavior. Below are interaction plots for the 20 proteins/modifications sorted with the lowest uncorrected p-values for this interaction:


pNUMB_N
RAPTOR_N
TIAM1_N
NR2B_N
AMPKA_N
NR2A_N
BDNF_N
MTOR_N
ERK_N
PKCA_N
Bcatenin_N
pNR1_N
pMTOR_N
pGSK3B_N
NR1_N
APP_N
CAMKII_N
MEK_N
ADARB1_N
DSCR1_N
Genotype
Control
Ts65Dn
Behavior
Control
Learning

When corrected for multiple comparisons using the Benjamini-Hochberg procedure with a False Discovery Rate of 0.05, 11 proteins or modifications showed significant interaction between genotype and behavior:

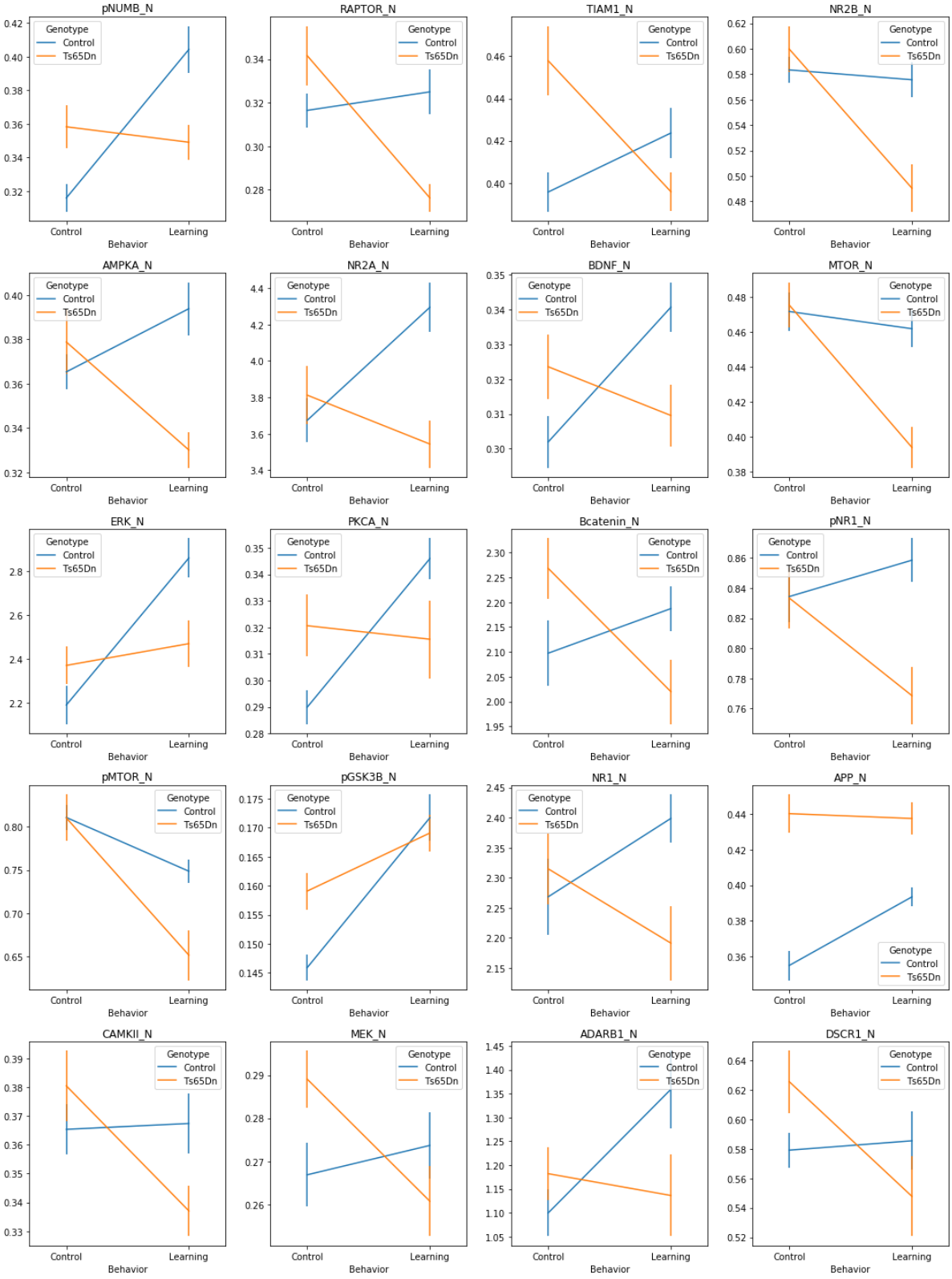
Protein or modification	**Full name**	**Change**
pNUMB	phosphorylated NUMB endocytic adaptor protein	Increases with learning in wildtype from below mutant levels to above
RAPTOR	regulatory associated protein of MTOR complex 1	Decreases with learning in mutant
TIAM1	T-cell lymphoma invasion and metastasis-inducing protein 1	High baseline in mutant, decreases with learning in mutant
NR2B	glutamate ionotropic receptor NMDA type subunit 2B	Decreases with learning in mutant
AMPKA	5'-AMP-activated protein kinase catalytic subunit alpha	Decreases with learning in mutant
NR2A	glutamate ionotropic receptor NMDA type subunit 2A	Increases with learning in wildtype
BDNF	brain-derived neurotrophic factor	Increases with learning in wildtype
MTOR	mammalian target of rapamycin	Decreases with learning in mutant
ERK	mitogen-activated protein kinase 1	Increases with learning in wildtype
PKCA	protein kinase C alpha	Increases with learning in wildtype
Bcatenin	catenin beta 1	Decreases with learning in mutant

Some notes on specific results

While I can speculate on the possible functional reasons for changes in specific protein/modification levels, many of these proteins are principally known for their function as membrane-bound or cytosolic proteins. Therefore, it will not always be obvious what changes in nuclear levels mean for the overall function of the cells.

Interestingly, this list includes 2 subunits of the mTORC1 complex (mTOR and RAPTOR) and 2 subunits of NMDA receptors (NR2A and NR2B). Both mTOR and RAPTOR decreased in the mutant in the learning condition, while levels did not change much in the wildtype. The functional receptor consists of 2 NR1 subunits, 1 NR2, and 1 NR3. Each of these subunits have variants which determine the properties of the assembled receptor. Higher NR2B has been found to lead to better learning via long term potentiation (LTP), and NR2B was decreased with learning specifically in the DS mice. Meanwhile, NR2A increased with learning only in wildtype mice. NR1 and phosphorylated NR1 also showed interaction effects with less stringent correction of p-values.

The only protein in this subset that is on the human chromosome 21 is TIAM1, which was higher in the mutant than in the wildtype at baseline but decreased in the learning condition. TIAM1 is involved in response to BDNF (which increased with learning in the wildtype). BDNF signaling is important for NMDA receptor trafficking and LTP. At least in migrating neurons, the phosphorylation of NUMB (also increased with learning in the wildtype) is an important component of BDNF-mediated cell polarization.


pNUMB_N
RAPTOR_N
TIAM1_N
NR2B_N
AMPKA_N
NR2A_N
BDNF_N
MTOR_N
ERK_N
PKCA_N
Bcatenin_N
pNR1_N
pMTOR_N
pGSK3B_N
NR1_N
APP_N
CAMKII_N
MEK_N
ADARB1_N
DSCR1_N
Genotype
Control
Ts65Dn
Behavior
Control
Learning

While I can speculate on the possible functional reasons for changes in specific protein/modification levels, many of these proteins are principally known for their function as membrane-bound or cytosolic proteins. Therefore, it will not always be obvious what changes in nuclear levels mean for the overall function of the cells.