Supplementary Methods

Genetic Disruption of WASHC4 Drives Endo-lysosomal Dysfunction and Cognitive-Movement Impairments in Mice and Humans

Jamie Courtland[1]*, Tyler W. A. Bradshaw[1]*, Greg Waitt[2], Erik J. Soderblom[2,3], Tricia Ho[2], Anna Rajab[4], Ricardo Vancini[5], Il Hwan Kim[2†], Ting Huang[6], Olga Vitek[6], Scott H. Soderling[3]

[1]Department of Neurobiology, Duke University School of Medicine, Durham, NC 27710, USA; [2]Proteomics and Metabolomics Shared Resource, Duke University School of Medicine, Durham, NC 27710, USA; [3]Department of Cell Biology, Duke University School of Medicine, Durham, NC 27710, USA; [4]Burjeel Hospital, VPS Healthcare, Muscat, Oman; [5]Department of Pathology, Duke University School of Medicine, Durham, NC 27710, USA; [6]Khoury College of Computer Sciences, Northeaster University, Boston, MA 02115, USA

Author coorespondence:
jlc123@duke.edu (JC); tyler.w.bradshaw@duke.edu (TWAB); greg.waitt@duke.edu (GW); erik.soderblom@duke.edu (EJB); tricia.ho@duke.edu (TH); drannarajab@gmail.com (DR); ricardo.vancini@duke.edu (RV); ikim9@uthsc.edu (IK); huang.tin@northeastern.edu (TH); o.vitek@northeastern.edu (OV); scott.soderling@duke.edu (SHS)

*These authors contributed equally to this work.

Present address:
†Department of Anatomy and Neurobiology, University of Tennessee Health Science Center, Memphis, TN 38163, USA

Abstract

In the review of this manuscript, significant concerns were raised by the reviewers about the validity of our statistal approach to perform protein- and module-level inference from our **WASH-BioID** and **SWIP-TMT** proteomics datasets. Our previous statistical approach relied upon the R package `edgeR` to evaluate differential protein abundance. `edgeR` utilizes a negative binomial, generalized linear model (NB GLM) framework, fitting each protein with a NB GLM. Previously, we failed to fully consider the validity of the NB GLM model used by `edgeR` for proteomics data. In response to this critique, we explore the goodness-of-fit of the NB GLM model for our `SWIP-TMT` data, and find evidence of a lack-of-fit. Thus, we revised our statistical approach and reanlyzed our data making use of the recently published tool `MSstatsTMT`. `MSstatsTMT` uses a linear mixed model (LMM) framework to model major sources of variation in a proteomics experiment. We extend the LMM framework used by `MSstatsTMT` to re-evaluate both protein- and module-level statistical comparisions. Despite evidence of a lack-of-fit for the NB GLM method used by `edgeR`, we find that the inferences we derived from our previous analysis are largely preserved in our reanalysis using `MSstatsTMT`.

Lack-of-fit of the Negative Binomial Model

Our previous approach is summarized as the 'Sum + IRS' method by Huang *et al.* (REF). Following protein summarization and Internal Reference Scaling (IRS) normalization, we applied `edgeR` to assess differential abundance of individual proteins and protein-groups or modules. We drew precidence for the use of `edgeR` from previous work by Plubell and Khan, *et al.* (REFS) who describe IRS normalization and the use of `edgeR` for statistical testing in TMT mass spectrometry experiments. We failed however, to consider the overall adequacy of the NB GLM model for our TMT proteomics data.

Statisitical inference in `edgeR` is performed for each gene or protein using a negative binomial framework in which the data are assumed to be adequately described by a NB distribution parameterized by a dispersion parameter, ϕ. Practically, the dispersion parameter accounts for mean-variance relationships observed in proteomics and transcriptomics data. `edgeR` employs empirical Bayes methods that allow for the estimation of feature-specific (i.e. gene or protein) biological variation, even for experiments with small numbers of biological replicates, as is common in transcriptomics and proteomics experiments. This empirical Bayes strategy is a strength of the `edgeR` approach as it reduces the uncertainty of the estimates and improves testing power.

As signal intensity in protein mass spectrometry is fundamentally related to the number of ions generated from a ioninized, fragmented protein, we incorrectly infered that TMT mass spectrometry data can be modeled as negative binomial count data. Based on this assumption, we justified the use of `edgeR`. Here we reconsider the overall adequacy of the `edgeR` NB GLM model for TMT mass spectrometry data.

To evaluate the overall adequacy of the `edgeR` model, we plot the residual protein deviance statistics of all proteins against their theoretical, normal quantiles in a quantile-quantile (QQ) plot **Figure**. The QQ plot addresses the question of how similar the observed data are to the theoretical distribution given by NB GLM fit. A linear relationship between the observed and theoretical values is an indicator of goodness-of-fit. Deviation from this linear trend is evidence of a lack-of-fit.

Following protein summarization and normalization, the data were fit with a simple NB GLM of the form `Abundance ~ Mixture + Condition` using `edgeR's glmFit` function which fits a NB GLM model to each protein or gene (the sub-subplot summaries) in the data. The dispersion parameter ϕ can take several forms, and `edgeR` supports three different dispersion metrics: 'common', 'trended',

and 'tagwise'. **Figure** illustrates the divergence of the observed deviance statistics from the theoretical distribution for data fit with the NB GLM model. These plots emphasize the overall lack of fit of proteomics data fit by the `edgeR` model.

Reanalysis of SWIP^{P1019R} TMT Proteomics

Of note, most tools for analysis of protein mass spectromety data are derived from tools originally developed for analysis of genomics and transcriptomics data. An exception to this norm is `MSstatsTMT`, an extension of `MSstats` for analysis of TMT proteomics experiments.

`MSstatsTMT` utilizes a linear mixed-model framework. The strength of LMMs is their flexibility. LMMs model what we are interested in as a function of the complex variation in an experimental design. Using LMMs we untagle whate we are intested in from experimental and biological covariates which mask that response. In mixed models, the response variable is taken to be a function of both fixed and mixed effects. If the set of possible levels of the covariate is fixed and reproducible then the factor is modeled as a fixed-effect parameter. In contrast, if the levels of an observation reflect a random sampling of the set of all possible levels then the covariate is modeled as a random effect. Random or mixed-effects represent categorical variables that reflect experimental or observational "units" in the data set. Mixed-effect parameters thus account for the variation occurring among all of the lower level units of a particular upper level unit in the data. For this reason, mixed models may also be refered to as heirarchical models.

Tandem mass tag, or TMT reagents enable the combination and simultaneous quantifiaction of muliple biological samples by mass spectrometry. Currently commercially available reagents are capable of labeling up to 16 protein preparations which are then analyzed together in a single mass spectrometry run. Peptides labeled with TMT tags are distinuisable from each other due to the unique reporter ions generated by the TMT tag which is used for relative quantification. In a TMT experiment, ionized features are matched to peptides, these peptide spectrum matches (PSM), for all unique TMT channels are analyzed simultaneously as a single precursor. Quantification of all biological conditions is thus achived within a single MS run in which all features for a protein are quantified simultaneously.

Huang *et al.* created `MSstatsTMT`, an R package for data normalization and hypothesis testing in multiplex TMT proteomics experiments. They outline a common vocabulary for describing the experimental design of TMT MS experiments. A TMT experiment consists of the analysis of `m = 1` ... `M` concatensions of

isobarically labeled samples or `Mixtures`. This mixture is then analyzed by the mass spectrometer in a mass spectrometry `Run`. This mixture is often fractionated into multiple liquid chromotography `Fractions` to decrease sample complexity, and thereby increase the depth of proteome coverage. Within a mixture, each of the unique TMT channels is dedicated to the analysis of `c = 1` ... `C` individual biological or treatment `Conditions`. There may then be `b = 1` or more `B` biological replicates or `Subjects`. Finally, a sigle TMT mixture may be repeatedly analyzed in `t = 1` ... `T` technical replicate mass spectrometry runs.

Protein-wise comparisons between `Conditions` of `Subjects` is performed by contrast of conditioned means obtained from fitting the data with a linear mixed-effects model expressing the major sources of variation in the experimental design. **Equation** is a mixed-effects model which describes protein abundance, the response Y_{mcbt}, in an experiment composed of `M` mixtures, `T` technical replicates of mixture, `C` conditions, and `B` biological subjects.

$$Y_{mcbt} = \mu + Mixture_m + TechRep(Mixture)_{m(t)} + Condition_c + Subject_b + \epsilon_{mcbt} \quad (1)$$

$$\begin{gathered} Mixture_m \overset{iid}{\sim} N(0, \sigma^2_M) TechRep(Mixture)_{t(m)} \overset{iid}{\sim} N(0, \sigma^2_T) \\ \sum_{c=1}^{C} Condition_c = 0 \\ Subject_{mcb} \overset{iid}{\sim} N(0, \sigma^2_S) \\ \epsilon mtcb \overset{iid}{\sim} N(0, \sigma^2) \end{gathered} \quad (2)$$

The model's constraints distinguish fixed and random components of variation in the response. `Mixture` is a mixed effect and represents the variation between TMT mixtures which is assumed to be random and normally distributed. `TechRep(Mixture)` represents random variation between replicate mass spectrometry runs of a same mixture. The term `Subject` cooresponds to each biological replicate and represents biological variation among the levels of the fixed effect term `Condition`. The term ϵ_{mtcb}, is a random-effect representing both biological and technical variation, quantifying any remaining error, assumed to be be independent and identically distributed.

If a component of the model is not estimable, it is removed. For example, if there is no technical replication of mixture (T=0), the model is reduced to:

$$Y_{mcbt} = \mu + Mixture_m + Condition_c + \epsilon_{mcb} \quad (3)$$

In the reduced model, biological variation among individual subjects is captured by the term `Condition` and is thus ommited.

Test Statistic

Model based testing of differential abundance between pairs of conditions is obtained by comparing the estimates of coefficients of interest. We are interested in tested the null hypothesis of equality of conditioned means $H0 : l^T\beta = 0$ (lmerTest, Kutzenova2017). Kutzenova et al derive a test statistic for such contrasts as:

$$t = \frac{l^T\hat{\beta}}{\sqrt{l\sigma^2\hat{V}l^T}} \tag{4}$$

We obtain the model estimates $\hat{\beta}$, error σ^2, and variance-covariance matrix $\hat{V}$ from the fitted model. Together $\sigma^2 * \hat{V}$ is the scaled variance-covariance matrix describing the error estimates of the models mixed-effect parameters. Given l^T, a vector of sum 1 specifying the positive and negative coefficients of the comparison, the numerator of the equation is then the fold change of a given comparison, and together the denominator represents the standard error of the contrast.

The degrees of freedom for the contrast are derived using the Satterthwaite moment of approximation method as previously described by Kutzenova et al. Finally, given the t-statistic, which is assumed to follow an approximate chi^2 distribution, and the degrees of freedom, a p-value is calculated for the test statistic. P-values for the protein-wise tests are adjusted using the Benjamnii hochberg method.

SWIP-TMT Proteomics Experimental Design

We prepared 7 `BioFractions` from 'Control' and SWIP^{P1019R} 'Mutant' mice. Thus in our experiment, the fixed effect term `Condition` of Equation 1 represents the interaction of `Genotype` and `BioFraction` and is the 14 unique combinations of 7 subcellular `BioFractions` prepared from 'Control' and 'Mutant' mice. Our TMT proteomics experimental design is summarized in **Figure**.

Note that in our experiment each TMT mixture contains seven repeated measurments made from each biological `Subject`. We refer to the subcellular fractions obtained by differential centrifugation of the brain as `BioFractions` to distinguish them from MS `Fractions` in the `MSstatsTMT` nomenclator. To account for this

source of intra-Subject variability, we should include the random-effect term `Subject` representing the random error within a subject. However, in our design `Mixture` is confounded with the term `Subject` – in each mixture we analyzed all `BioFractions` from a single Control and Mutant mouse. Thus we can choose to account for the effect of `Mixture` or `Subject`, but not both. Under the assumption that the effect of TMT `Mixture` is greater than the variance attributable to the repeated measurements of each subject, we omit the term `Subject`. The reduced model is equivalent to equation (EQ) when condition is `Genotype:BioFraction`.

Figure shows the proportion of variance attributable to major covariates in our TMT experiment for each protein.
Given the LMM describing protein abundance as a respones of the experiments major sources of variaiton, we assessed two different types of contrasts:

1. 'intra-BioFraction' comparisons for the difference between Mutant and Control conditions within a biological fraction.
2. 'Mutant-Control' contrast for the overall difference between Mutant and Control conditions.

The vector l^T is a contrast vector specifying each comparison between the negative coefficient, e.g. 'GenotypeControl.BioFractionF4' and positive coefficient, e.g. 'GenotypeMutant.BioFractionF4'. Figure (FIG) shows these two types of contrast. Contrasts 1-7 specify 'intra-BioFraction' comparisons for each of the subcellular fractions. Contrast 8 specifies the overall 'Mutant-Control' comparison.

Protein-level comparisons

`MSstatsTMT` attempts to automatically parse the experimental design and fit an appropriate LMM for the experimental design. In order to understand and extend the function of `MSstatsTMT`, we We extracted `MSstatsTMT's` core model-fitting and statistical testing steps and illustrate them here.