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-iBioID** and **SWIP-TMT** proteomics datasets. Our previous statistical approach was dependent upon the R package `edgeR` to evaluate differential protein abundance. `edgeR` utilizes a negative binomial generalized linear model (NB GLM) framework, originally developed for analysis of read counts data generated in RNA-seq transcriptomics experiments. 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 in the dataset 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 the observed mean-variance relationship in proteomics and transcriptomics data.

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 with `MSstatsTMT`, 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 our TMT data fit with the NB GLM model. These plots emphasize the overall lack of fit of proteomics data fit by the `edgeR` model.

Given our experimental design, `MSstatsTMT` fits an analagous LMM: `Abundance ~ Condition + (1|Mixture)`. The QQ plot in Figure indicates that the data are well described by `MSstatsTMT's` LMM framework, which does not depend upon the negative binomial assumption.

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 lies in their flexibility. In mixed models, the response variable is taken to be a function of both fixed and mixed effects. Using LMMs we can untagle the variance attributable to the biological effect we are interested in from the experimental and biological covariates which which mask this response.

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 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. (Bates) As such, mixed-effect parameters 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 experi-

ments. 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.

Equation is a mixed-effects formula which describes a general TMT experiment composed of `M` mixtures, `T` technical replicates of mixture, `C` conditions, and `B` biological subjects. The abundance of a given protein, Y_{mcbt}, is then:

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

The model's constraints distinguish fixed and random components of variation in the response.

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

`Mixture` is a mixed effect and represents the variation between TMT mixtures which is assumed to be random and normally distributed (iid). `TechRep(Mixture)` represents random variation between replicate mass spectrometry runs of a same mixture. The term `Subject` cooresponds to each unique 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.

If a component of the model is not estimable, then 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

`MSstatsTMT` performs protein-wise comparisons between `Conditions` of biological `Subjects` by a 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.

Model based testing of differential abundance between pairs of conditions is done by comparing the estimates obtained from the fit LMM. We are interested in testing the null hypothesis $H0 : l^T\beta = 0$. Kutzenova *et al.,* derive a test statistic for such contrasts (Kutzenova2017):

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

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 (Kutzenova2017). 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. P-values for the protein-wise tests are adjusted using the Benjamnii hochberg method.

SWIP-TMT Proteomics Experimental Design

In our experiment, the fixed-effect term `Condition` represents the 14 combinations of `Genotype` and `BioFraction` obtained from subcellular fractionation of the brains of 'Control' and SWIP^{P1019R} 'Mutant' mice. We refer to these as `BioFractions` to distinguish them from a MS `Fraction`. Our TMT proteomics experimental design is summarized in **Figure**.

In our experiment, each TMT mixture contains seven repeated measurements made from each biological `Subject`. 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. Me choose to account for variability of `Mixture` under the assumption that the effect of this experimental batch effect is greater than the variance attributable to the random variability inherint in making repeated measurements of each subject. Thus we omit the term `Subject`. The reduced model is then the same as equation (EQ) when `Condition` is the interaction of `Genotype:BioFraction`.

Protein level comparisions

Using `MSstatsTMT` we assesed comparisons at two levels:

- 'intra-BioFraction' contrasts
- 'Mutant-Control' contrast

'Intra-BioFraction' comparisons are the 7 pairwise compairisons of Control and Mutant protein abundance for each subcellular fractions. We also assessed differential abundance for the overall comparison between 'Control' and 'Mutant' groups. Each of these contrast is represented by a vector, l^T, which specifies a comparison between coefficients in the LMM. Figure (FIG) illustrates a matrix defining all 8 contrasts.

`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.

Following data preprocessing, summarization, and normalization, statistical inference by `MSstatsTMT` can be summarized in two steps:

- Fit each protein with the appropriate LMM, and then
- given the fitted model, assess a contrast of interest.

At the core of the model fitting step is the R package `lme4` which implements mixed-effects models with its function `lmer`. The package `lmerTest` extends `lme4's` functionality and enables the computation of Sattertwaite degrees of freedom.

```r
# load dependencies
library(dplyr)
library(data.table)

#library(SwipProteomics)

# load the data
data(swip)
data(msstats_prot)

# formula to be fit to WASHC4, aka SWIP:
fx <- formula("Abundance ~ 0 + Genotype:BioFraction + (1|Mixture)")

# fit the LMM
fm <- lmerTest::lmer(fx, msstats_prot %>% filter(Protein == swip))

# examine the model's summary
summary(fm, ddf = "Satterthwaite")
```

Coefficient	Estimate	Std. Error	df	t value	p value
Mutant:F4	5.404300	0.121126	17.30594	44.61718	2.59e-19
Control:F4	6.710959	0.121126	17.30594	55.40477	6.26e-21
Mutant:F5	5.567441	0.121126	17.30594	45.96405	1.56e-19
Control:F5	6.945583	0.121126	17.30594	57.34180	3.47e-21
Mutant:F6	5.640188	0.121126	17.30594	46.56463	1.24e-19
Control:F6	7.240081	0.121126	17.30594	59.77313	1.7e-21
Mutant:F7	5.631680	0.121126	17.30594	46.49440	1.28e-19
Control:F7	7.321074	0.121126	17.30594	60.44180	1.4e-21
Mutant:F8	5.492772	0.121126	17.30594	45.34759	1.96e-19
Control:F8	7.129632	0.121126	17.30594	58.86129	2.21e-21
Mutant:F9	5.781022	0.121126	17.30594	47.72734	8.15e-20
Control:F9	6.954472	0.121126	17.30594	57.41518	3.39e-21
Mutant:F10	5.784403	0.121126	17.30594	47.75525	8.07e-20
Control:F10	7.618697	0.121126	17.30594	62.89894	7.04e-22

The model's estimates, β, represent our best estimate of the mean protein abundance in the 14 conditions of `Genotype:BioFraction`. To illustrate a comparison, we define a contrast comparing 'Mutant:F7' and 'Control:F7'. The function `lmerTestContrast` performs model-based comparisons of conditions defined by a contrast matrix. While the work done by this function is the same as the work done internally by `MSstatsTMT's groupComparisonsTMT` function, its strength lies in its flexibility.

```
# create a contrast
coeff <- lme4::fixef(fm)
contrast7 <- setNames(rep(0,length(coeff)), nm = names(coeff))
contrast7["GenotypeMutant:BioFractionF7"] <- +1 # positive coeff
contrast7["GenotypeControl:BioFractionF7"] <- -1 # negative coeff

# evaluate contrast
lmerTestContrast(fm, contrast7)
```

Contrast	log2FC	percentControl	SE	Tstatistic	Pvalue	DF
Mutant:F7-Control:F7	-1.689	0.31	0.151	-11.153	2.09e-11	26

Provided the correct contrast, we easily assess the overall comparison between 'Mutant' and 'Control' groups:

```
# use convenience function to contruct a contrast
lT <- getContrast(fm, "Mutant","Control")

# assess the comparison
lmerTestContrast(fm, lT)
```

```
df <- lmerTestContrast(fm, lT)
df <- df %>% mutate(Contrast = 'Mutant-Control')
df$SE <- round(df$SE,3)
df$log2FC <- round(df$log2FC,3)
df$percentControl <- round(df$percentControl,3)
df$Tstatistic <- round(df$Tstatistic,3)
df$Pvalue <- formatC(df$Pvalue,digits=3)
df$isSingular <- NULL
df %>% unique() %>% knitr::kable()
```

Contrast	log2FC	percentControl	SE	Tstatistic	Pvalue	DF
Mutant-Control	-1.517	0.349	0.057	-26.496	2.42e-20	26

Goodness-of-fit

It is useful to consider the goodness-of-fit of our LMM. A straight forward measure of the quality of a mixed model is Nagagawa's coefficient of determination. Nakagawa's conditional R^2 is interpreted as the total variance explained by a LMM (R^2_{total}). The marginal R^2 is interpreted as the variance explained by the LMM's fixed effects (R^2_{fixed}).

We implment Nakagawa's coeffficient of determination using the `r.squaredGLMM.merMo` function forked from the `MuMin` package.

```
# assess gof with Nakagawa coefficient of determination
r.squaredGLMM.merMod(fm) %>% knitr::kable()
```

R2m	R2c
0.9353344	0.949433

We can see the total variation explained by the model, $R_c^2 is 0.949$. The variance explained by fixed effects, `Genotype:BioFraction`, equates to 0.935 ($R^2 m$). A vast majority of the variance is attributable to the fixed effects. Only about 1.5% of the remaing variance is attributable to mixed effects and the residual variance.

Module-level analysis

We wish to perform inference at the level of protein modules. These groups of covarying proteins represent hypothesized biological niches defined by proteins that localized together in subcellular space.

Effects are fixed if they are interesting in themselves or random if there is interest in the underlying population. Searle, Casella, and McCulloch (1992, Section 1.4) explore this distinction in depth. morover as the sample does not We are interested in the overall affect on the group of proteins and not the Mixed effects represent a subset of the exhaust the population. A strenght of the LMM approach applied to is the partial pooling which strengthens the power of the statistical test. Here we hypothesize that covarying protein represetnt groups of proteins that are a part of a larger groups with a common mean effect. Proteins within a module represnt correlated observations with we model as a mixed effect. We take the stance that Protein is a random effect in that we are primarily interested in makein inference about the overall distribution responses for a module rather than within between the particular sublevels of a module.

We model protein groups or modules by adding the mixed effect term `Protein`. The protein constituents of a module.

$$Y_{mcbt} = \mu + Mixture_m + Condition_c + Protein_p + \epsilon_{mcb} \tag{5}$$