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`, which utilizes a negative binomial, generalized linear model (NB GLM) framework. 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 built on 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 NB distribution. A linear relationship between the observed and theoretical values is goodness-of-fit indicator. 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` $\sim$ `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} Spatial 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 linear mixed models (LMMs) is in their ability to account for complex sources of variation in an experimental design.

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.

Huang *et al.*, describe a linear mixed model framework for statistical inference in mass spectrometry experiments. A generalized TMT proteomics experiment consists of the analysis of `m = 1` ... `M` concatensions of isobarically labeled samples or `Mixtures`. Within a mixture, each TMT channel is dedicated to the analysis of `c = 1` ... `C` individual biological or treatment `Conditions` prepared from `b = 1` ... `B` biological replicates or `Subjects`. A single mixture may be profiled in `t = 1` ... `T` technical replicate mass spectrometry runs.

Equation (EQ) is a mixed-effects model which describes the response, protein abundance Y_{mcbt}, in an experiment composed of `M` mixtures, `T` technical replicates of mixture, `C` conditions, and `B` biological subjects.

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

$$Mixture_m \overset{iid}{\sim} N(0, \sigma^2_M) \quad (2)$$

$$
\begin{aligned}
TechRep(Mixture)_{t(m)} &\overset{iid}{\sim} N(0, \sigma_T^2) \\
\sum_{c=1}^{C} Condition_c &= 0 \\
Subject_{mcb} &\overset{iid}{\sim} N(0, \sigma_S^2) \\
\epsilon mtcb &\overset{iid}{\sim} N(0, \sigma^2)
\end{aligned}
\tag{3}
$$

The model's constraints, **Equation**, 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 (σ^2), and is assumed to be independent and identically distributed.

If a component of the model is not estimable, it is removed. Thus, if there is no technical replication of Mixture, the model is reduced to:

$$
Y_{mcbt} = \mu + Mixture_m + Condition_c + \epsilon_{mcb} \tag{4}
$$

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

We prepared 7 `BioFractions` from 'Control' and SWIP^{P1019R} 'Mutant' mice. Thus, in our experiment, the fixed effect term `Condition` represents the interaction of `Genotype` and `BioFraction` is the 14 unique combinations of 7 subcellular `BioFractions` prepared from 'Control' and 'Mutant' mice. **Figure** shows the proportion of variance attributable to major covariates in our TMT experiment for each protein.

In our experimental design, we made seven repeated measurments 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. 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`. Our TMT proteomics experimental design is summarized in **Figure**.

Model based testing of differential abundance between pairs of conditions is assessed through contrast of conditioned means estimated by fitting the parameters of the model by REML. We obtain $\hat{\beta}$, σ^2, and $\hat{V}$ from the fitted model. The degrees of freedom are determined by the Satterthwaite approximation. Given a contrast vector defining a comparison between coefficients in the model we use the model estimates to evaluate a T-statistic for the comparison given by the formula (REF):

$$t = \frac{l^T * \hat{\beta}}{\sqrt{lx\sigma^2 x\hat{V}xl^T}} \tag{5}$$

Where m^T is the a vector of sum 1 which indicates the positive and negative coefficients of the comparison. σ^2 is the error from **Equation**. l^T is a vector specifying a contrast between positive and negative coefficients in the model. Together, the denominator, $\sqrt{l * \sigma^2 * \hat{V} * l^T}$, is the standard error of the contrast.

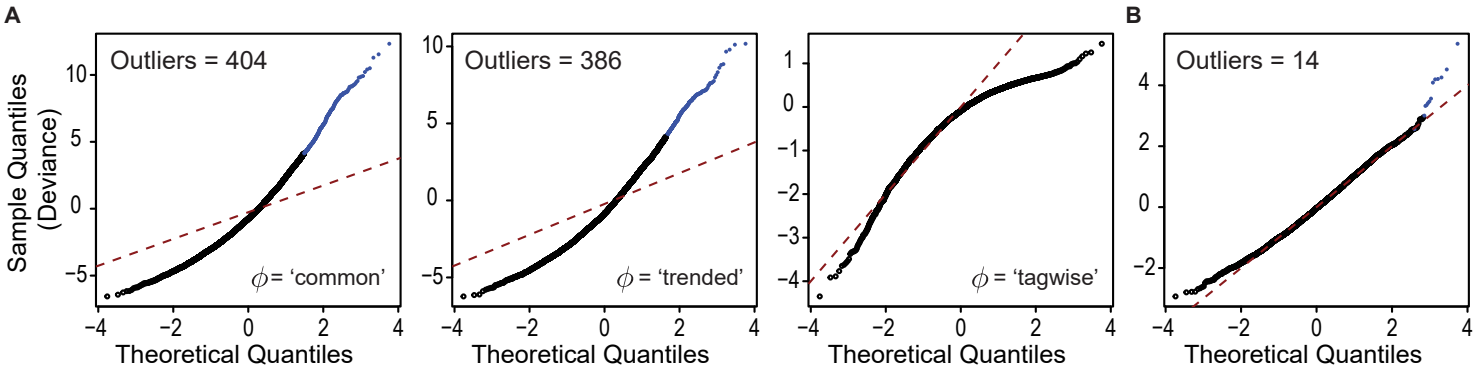

Figure 1. Goodness-of-fit of `edgeR` (A), and `MSstatsTMT` (B) statistical approaches. The overall adequacy of the linear models fit to the data were assessed by plotting the residual deviance for all proteins as a quantile-quantile plot (McCarthy *et al.*, (2012)). **(A)** For analysis with `edgeR`, The normalized protein data from `MSstatsTMT` were fit with a negative binomal generalized linear model (NBGLM) of the form: `Abundance` ~ `Mixture + Condition`. Where `Mixure` is an additive blocking factor that accounts for variablity between experiments. The NB framework used by edgeR utilizes a dispersion parameter to account for mean-variance relationships in the data. The dispersion parameter can take several forms. `edgeR` supports three dispersion models: 'common', 'trended', and 'tagwise'. However, when using `edgeR's` robust quasi-likelihood test methods, only global (i.e. 'common' or 'trended') dispersion metrics are appropriate (see `edgeR::glmQLFit's` documentation). We plot the protein-wise deviance from the data fit withe ach of the disperions parameters. Protein-wise deviance statistics were transformed to normality and plotted against theoretical normal quantiles using the `edgeR::gof` function. **(B)** For analysis with `MSstatsTMT`, the normalized protein data were fit with a linear mixed-effects model (LMM) of the form: `Abundance` ~ `0 + Condition + (1|Mixture)`. Where `Mixture` represents the random effect of `Mixture`. The residual deviance and degrees of freedom were extracted from the fitted models, z-score normalized, and plotted as in (A). Proteins with a significantly poor fit are indicated as outliers in blue (Holm-adjusted P-value < 0.05).

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16
Mix1	WT-5K	WT-9K	WT-12K	WT-15K	WT-30K	WT-79K	WT-129K	QC1	MUT-5K	MUT-9K	MUT-12K	MUT-15K	MUT-30K	MUT-79K	MUT-129K	QC2
Mix2	WT-5K	WT-9K	WT-12K	WT-15K	WT-30K	WT-79K	WT-129K	QC1	MUT-5K	MUT-9K	MUT-12K	MUT-15K	MUT-30K	MUT-79K	MUT-129K	QC2
Mix3	WT-5K	WT-9K	WT-12K	WT-15K	WT-30K	WT-79K	WT-129K	QC1	MUT-5K	MUT-9K	MUT-12K	MUT-15K	MUT-30K	MUT-79K	MUT-129K	QC2

Figure 2. Experimental Design. We performed three 16-plex TMT experiments. Each TMT mixture is a concatenation of 16 labeled samples. In each experiment we analyzed 7 subcellular `BioFractions` prepared from the brain of a 'Control' or 'Mutant' mouse. In all we analyzed 3 `Subjects` from each Condition. Each `Mixture` includes two `Channels` dedicated to the analysis of a common quality control sample.

	F4	F5	F6	F7	F8	F9	F10	F4	F5	F6	F7	F8	F9	F10	
L1	-1	0	0	0	0	0	0	+1	0	0	0	0	0	0	Mutant.F4-Control.F4
L2	0	-1	0	0	0	0	0	0	+1	0	0	0	0	0	Mutant.F4-Control.F4
L3	0	0	-1	0	0	0	0	0	0	+1	0	0	0	0	Mutant.F4-Control.F4
L4	0	0	0	-1	0	0	0	0	0	0	+1	0	0	0	Mutant.F4-Control.F4
L5	0	0	0	0	-1	0	0	0	0	0	0	+1	0	0	Mutant.F4-Control.F4
L6	0	0	0	0	0	-1	0	0	0	0	0	0	+1	0	Mutant.F4-Control.F4
L7	0	0	0	0	0	0	-1	0	0	0	0	0	0	+1	Mutant.F4-Control.F4
L8	-1/7	-1/7	-1/7	-1/7	-1/7	-1/7	-1/7	+1/7	+1/7	+1/7	+1/7	+1/7	+1/7	+1/7	Mutant.F4-Control.F4

Figure 3. Statistical Comparisons. We assessed two types of contrasts. Each row of the matrix specifies a contrast between positive and negative coefficients in the mixed effects model fit to each protein. Contrasts1-7 are 'intra-BioFraction' contrasts that specify the pairwise comparisons of Control and Mutant groups for a single fraction. In Contrast 8 we compare 'Mutant-Control' and asses the overall difference of 'Control' and 'Mutant' conditions. Each contrast is a vector of sum 1.