Available online at www.sciencedirect.com

ScienceDirect

journal homepage: http://www.elsevier.com/locate/euprot

# Detecting significant changes in protein abundance

Kai Kammers[a], Robert N. Cole[b], Calvin Tiengwe[c,d], Ingo Ruczinski[a,*]

[a] Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
[b] Mass Spectrometry and Proteomics Core Facility, Johns Hopkins University School of Medicine, Baltimore, MD, USA
[c] Department of Cell Biology, Johns Hopkins University School of Medicine, Baltimore, MD, USA
[d] Department of Microbiology and Immunology, School of Medicine and Biomedical Sciences, University at Buffalo, Buffalo, NY, USA

ARTICLE INFO

*Article history:*
Available online 25 February 2015

*Keywords:*
Empirical Bayes
Inference
Protein abundance

ABSTRACT

We review and demonstrate how an empirical Bayes method, shrinking a protein's sample variance towards a pooled estimate, leads to far more powerful and stable inference to detect significant changes in protein abundance compared to ordinary t-tests. Using examples from isobaric mass labelled proteomic experiments we show how to analyze data from multiple experiments simultaneously, and discuss the effects of missing data on the inference. We also present easy to use open source software for normalization of mass spectrometry data and inference based on moderated test statistics.



## 1. Introduction

Detecting significant changes in protein abundance is a fundamental task in mass-spectrometry based experiments when trying to compare treated to untreated cells, wildtypes to mutants, or samples from diseased to non-diseased subjects. The statistical inference for proteomic data in these settings is usually based on standard 2-sample t-tests, comparing the measured relative or absolute abundances for each peptide or protein across the conditions of interest. However, sample sizes are often small, sometimes as small as 4 or 8 samples total, which result in great uncertainty in the sample variability estimates. Since these estimates are used in the test statistics to assess the statistical significance of the observed fold change, proteins exhibiting a large fold change are often declared non-significant because of a large sample variance, while at the same time small observed fold changes might be declared statistically significant, because of a small sample variance.

Additional methods to assess biological and technical sources of variability have been proposed [1–6], including methods to analyze data from multiple experiments simultaneously. For case–control iTRAQ experiments, Oberg et al. [7] and Hill et al. [8] extended a linear mixed effects approach originally proposed by Kerr and Churchill [9,10] as analysis of variance for gene expression studies. This mixed model adjusts for potential differences due to channel effects, loading, mixing, and sample handling. The parameter of interest in the model is the interaction between protein and group status, with a statistically significant result indicating differential expression (abundances) between cases and controls. One of the noteworthy features of this approach is that it simultaneously estimates protein relative abundance and assesses differential expression, albeit with substantial computational cost due to the numerical complexity of optimizing the likelihood and estimating a rather large number of parameters. Herbrich et al. [11] demonstrated that estimating protein abundances using median sweeps reduces computational cost substantially, and is as efficient yet more robust than

* *Corresponding author*. Tel.: +1 4106147840.
E-mail address: ingo@jhu.edu (I. Ruczinski).
http://dx.doi.org/10.1016/j.euprot.2015.02.002

protein abundance estimation procedures based on linear mixed effects models.

An implicit assumption in the approach of Oberg et al. [7] and Hill et al. [8] is that the biological variability is the same for all proteins identified and quantified. Though *"all models are wrong, but some are useful"* [12], incorrect model assumptions can lead to a loss in power even if no bias is incurred. This was for example observed in gene expression studies when LIMMA ("Linear Models for Microarray Data") [13] was introduced as an empirical Bayes approach that specifically allowed for a realistic distribution of biological variances, compared to the models of Kerr and Churchill [9,10], which assumed constant variability. The statistical trick in LIMMA is to use the full data to shrink the observed sample variances towards a pooled estimate. This results in far more stable and powerful inference compared to ordinary *t*-tests particularly when the number of samples is small [13], yet still allows for a distribution of variances. LIMMA arguably is the contemporary analytical standard for gene expression experiments, as evidenced by over 6000 citations in the last ten years (http://scholar.google.com). LIMMA has also been sporadically used in the context of proteomic experiments [14–19], but is far from being regarded as the analytical standard. This is surprising since proteomic experiments often have somewhat small sample sizes, and for those the potential gains of an empirical Bayes procedure are highest. One possible explanation for this phenomenon (besides being originally developed for a different genomic application) might be that LIMMA has been implemented as a Bioconductor package in the language R, a statistical environment the proteomics community only recently started to embrace [20–26].

In this manuscript we use examples from quantitative proteomic experiments using isobaric mass tags to demonstrate how better results in case–control studies can be achieved by using the LIMMA moderated test statistics. We show how to analyze data from multiple experiments simultaneously, and discuss the effects of missing data on the inference. We give sufficient detail for the statistically inclined reader to understand what happens "under the hood" of this empirical Bayes approach, and also present easy to use open source software for the practitioner to carry out the normalization of these mass spectrometry data, and to readily obtain the inference from moderated test statistics.

## 2. Materials and methods

### 2.1. Sample description

The data stem from two *Trypanosoma brucei* transgenic cell lines overexpressing either *TbHslV*-wild type or *TbHslV*-mutant protease. The *T. brucei* mitochondrion contains a proteasome-like ATP-dependent protease named *TbHslVU* that plays a critical role in regulating the timing of mitochondrial DNA replication [27]. Previous experiments suggested that *TbHslVU* controls the timing of kDNA synthesis by degrading "positive regulator of replication" [27,28]. To search for *TbHslVU* substrates its catalytically active subunit (denoted as *TbHslV-wt*) and its catalytically dead mutant (denoted as *TbHslV-mt*) were fused to the tandem affinity purification (TAP) tag. TAP-tagged *TbHslV-wt* or *TbHslV-mt* overexpressing cell lines were generated and the overexpressed proteins were purified using a TAP protocol adapted from Ringpis [29]. *TbHslV-wt* and *TbHslV-mt* were performed in four independent biological replicates.

Quantitative mass spectrometry was used to identify proteins that are associated with overexpressed and purified *TbHslV-mt* but not with *TbHslV-wt* treated similarly; since the latter binds and degrades its substrates. Proteins were digested with trypsin, labelled using the eight-plex iTRAQ isobaric mass tags (ABSciex) and analyzed using tandem mass spectrometry on an LTQ Velos Orbitrap interfaced with an Eksigent 2D NanoLC as previously described [11,30,31], except mass tagged peptides were fractionated by basic reverse phase chromatography [32]. Peptides were identified using Proteome Discoverer v1.4 (Thermo Scientific, San Jose, CA) and Mascot v2.2 (Matrix Sciences). Software defaults were used to control the false discovery rate (FDR) and only peptides spectra with less than 1% FDR and less than 30% isolation interference were included in analysis.

Protein $\log_2$ relative abundances were estimated using the method of Herbrich et al. [11]. In this procedure, a logarithmic transformation of the reporter ion intensities is employed since systematic effects and variance components are usually assumed to be additive on this scale [7,8]. The $\log_2$ reporter ion intensities for each spectrum are "median-polished" by subtracting the spectrum median $\log_2$ intensity from the observed $\log_2$ intensities. The relative abundance estimate for a particular protein is calculated as the median of these residuals, from all reporter ion intensity spectra belonging to this protein. Corrections for differences in amounts of material loaded in the channels and sample processing are carried out by subtracting the channel median from the relative abundance estimate, normalizing all channels to have median zero.

### 2.2. Statistical inference

#### 2.2.1. Two group comparisons

To detect differentially expressed proteins in a balanced proteomic experiment with $n$ cases ($\log_2$ relative abundances $X_{1p}$, …, $X_{np}$ for protein $p$) and $n$ controls ($\log_2$ relative abundances $Y_{1p}$, …, $Y_{np}$), inference is typically based on a 2-sample *t*-test for each protein $p$, with test statistic

$$t_p = \frac{\text{estimated log fold change}}{\text{estimated standard error}} = \frac{\overline{X}_p - \overline{Y}_p}{s_p\sqrt{2/n}}, \quad (1)$$

where $\overline{X}_p$ and $\overline{Y}_p$ are the group mean $\log_2$ relative abundances, and

$$s_p = \sqrt{\frac{\sum_i (X_{ip} - \overline{X}_p)^2 + \sum_i (Y_{ip} - \overline{Y}_p)^2}{2n-2}} \quad (2)$$

is the within-group sample standard deviation. For each protein, a $p$-value is then calculated referring the test statistic $t_p$ to a *t*-distribution with $d_p = 2 \times n - 2$ degrees of freedom as null distribution. For the above the $\log_2$ relative abundances are assumed to be normally distributed with equal variance in each group, although *t*-tests are robust to departures from the