Supplemental Material:

Combined statistical analyses of peptide intensities and peptide occurrences improves identification of significant peptides from MS-based proteomics data

Bobbie-Jo M. Webb-Robertson, Lee Ann McCue, Melissa M. Matzke, Jon Jacobs, Tom Metz, Susan M. Varnum, Katrina Waters and Joel G. Pounds

Pacific Northwest National Laboratory, P.O. BOX 999, Richland, WA 99352

Material and Methods

G-Test Simulated Datasets. The biological effects are not known *a priori* for experimental datasets and thus the general methodology of the G-Test cannot be fully demonstrated. To systematically evaluate the sensitivity and specificity of the IMD-ANOVA approach, we simulated datasets such that the censored peptides were known *a priori* and the confounding issue of differential measured abundance was removed. Two datasets were simulated to approximate the size and distributional characteristics of the experimental datasets described in the main manuscript (BALF and Plasma); we did not attempt to simulate the unknown interdependencies between peptides. These datasets were simulated for the sole purpose of demonstrating the utility of G-test under a qualitative difference model and are not reflective of real proteomics data. The results demonstrate that ANOVA-based analyses in the presence of missing data do not identify censored peptides as significant beyond what is expected by chance, a key motivating factor of the G-test approach.

To fully evaluate the capability of a statistical test to identify data associated with group membership that are censored (significant based on missing data), i.e., calculate the true positives and false discovery rates of qualitative differences due to missing data, we needed datasets without the quantitative effect of differential observed abundances. Therefore, we used the following assumptions and simulated peptide abundances in the following way. Within a treatment group, for a single peptide the Log10 transformed intensities are generally assumed to be normally distributed. Thus, the simulated intensities for a given peptide *j* can be sampled from a normal distribution with mean (μ_j) and a standard deviation (σ_j), where each peptide has a single unique mean and standard deviation (SD) under the assumption that there is no between group effect. We used two experimental datasets to compute the mean and SD of each peptide within each group (ignoring missing data), requiring at least 4 or 6 peptides with observed values in the BALF and Plasma datasets, respectively, to compute the mean and SD values. This approach yielded a total of 3199 and 6870 computed means and standard deviations from the two respective datasets. Supplemental Figure S1 shows a density plot of these means and SDs fit to a Generalized Extreme Value (GEV) distribution. The standard error on the parameter estimates is less than 0.019 for the means and less than 0.009 for the SDs. The mean and SD of each simulated peptide was then generated by sampling from these GEV distributions.

Simulated Dataset 1 (simBALF). The first simulated dataset was designed to approximate the size of the BALF experimental dataset. There were 4 groups of sizes 12, 8, 8 and 8 samples and therefore, under the assumption of complete data with no quantitative effects, each peptide had 36 simulated intensity values sampled from the same distribution. In total, 36 intensity values were simulated for 4000 'peptides'. Specifically, for a single peptide *j*, a mean μ_j was randomly sampled from the GEV(ξ = 0.0424, σ=0.9435,μ=-0.1617) and a standard deviation σ_j was randomly sampled from the GEV(ξ = -0.0609, σ=0.2662,μ=0.6436). The intensities for peptide *j* in the 36 samples were then obtained by sampling from a normal distribution, $N(\mu_k, \sigma_k)$. The dataset was iteratively simulated in its entirety for the full 4000 peptides in all 36 samples. Data were then removed based on a probability value from the first 990 peptides (~25% of the complete dataset) to imitate censored peptides, where the probability-based selection allowed for random interjection of missing data as well. Specifically, there are 33 peptide subsets each with a total of *m* missing values where *m* ranged from 4 to 33 (total of 33x30 = 990 peptides). In particular, for peptide *j* (x_{ij}), where *i* represents the samples, n_k is the number of samples in group *k*, and the number of values to be removed is defined as *m*, the data was removed from the first 990 peptides with the following algorithm:

```
Initialize j = 0
For iteration 1 to 33 {
     For $m$ = 4 to 33 {
          Set $j = j + 1$
          Initialize $count$ = 0
          While count < $m$ {
               1) Select sample $i$ at random from 1 to 36.
               2) Set $x_{ij}$ = NaN and set $count$ = 1 (for peptide $j$)
```

```
            3) Compute the number of observed values per group; $n_{Ok}$ given the current information for peptide $j$.
            4) Sample the group from which the missing value will be removed based on $\frac{n_k / n_{Ok}}{\sum_a n_a / n_{Oa}}$, thus groups with more missing data will have a higher probability of being selected
            5) Select sample $i$ at random from $(x_{ij})^k$, the samples associated with the $k$-th group
                a. If all $(x_{ij})^k$ = NaN, return to Step 4
                b. If $x_{ij}$ = NaN, return to Step 5
            6) Set $x_{ij}$ = NaN
            7) Set $count = count + 1$
        }
    }
}
```

Globally, this procedure results in 18315 missing values, or 12.72% of the dataset removed. Data were then removed completely at random from the remaining 3010 peptides until a defined global percentage of the data were missing. The second phase or random removal is not applied to the first 990 peptides since the probability-based sampling allowed random removal. For example, 15% global missing data is a total of 21600 missing values from the dataset (36 x 4000). Since 18315 values were already missing, an additional 3285 peptide intensity values were removed at random from the 3010 peptides with complete data, creating dataset 15% (D_{15}).

The same process was repeated for 20%, which required that a total of 10485 additional intensities be removed (*D_{20}*). In this manner a total of 8 datasets are created with 15 to 50% missing data in increments of 5% to represent various ranges of missing data.

Simulated Dataset 2 (simPlasma). The simPlasma dataset was designed to approximate the size and characteristics of the Plasma experimental dataset, and thus consisted of 2 groups of sample sizes 13 and 14, where the 27 intensity values were simulated for a total of 6000 peptides. Specifically, for a peptide *k*, a mean μ_k was randomly sampled from the GEV(ξ = -0.1356, σ=1.0519,μ=-0.5388) and a standard deviation σ_k was randomly sampled from the GEV(ξ = 0.1019, σ=0.1685,μ=0.3709). The complete dataset was simulated, and data non-randomly removed in the same manner as described for the first simulated dataset, however in this case the number of samples with a missing value ranged from 4 to 24 per peptide and ~10% (609) of the peptides had missing data removed in a censored manner, resulting in a base *simPlasma* dataset with ~5.3% missing data. In addition, the proportions for sampling the group from which a value was removed was modified to increase the probability of selecting peptide intensities from the same group, by changing Step 4 of the simulation algorithm to sample based on probabilities $\dfrac{{}^{n_k}\!/_{n_{Ok}}\left(1+n_k-n_{Ok}\right)}{\sum_a {}^{n_a}\!/_{n_{Oa}}\left(1+n_a-n_{Oa}\right)}$. This code modification increased the likelihood that more peptides were removed from the an individual group. For example, suppose there are three groups of size 8 and there have been 6, 4 and 0 observations removed from each group, respectively, by the algorithm. Using the sampling proportion from the *simBALF* dataset, $\left({}^{n_k}\!/_{n_{Ok}}\right)$, the next peptide would be removed from each group with probability 0.571, 0.286, and

0.143, however, with ${}^{n_k}\!/_{n_{Ok}}\left(1+n_k-n_{Ok}\right)$ these probabilities change to 0.718, 0.256 and 0.026, increasing the likelihood of selecting a peptide from the first group dramatically. Finally, data were randomly removed from the 5391 remaining peptides with complete data to create 8 datasets with total missing data of 15 to 50%, as described for *simBALF*.

Results and Discussion

Analysis of qualitative differences in simulated datasets. Real proteomics datasets have unknown quantitative (peptide abundance values) and qualitative (missing data, the censored peptides) differences, and therefore cannot be used to evaluate the accuracy and specificity of the G-test, nor compare G-test performance to that of ANOVA-based approaches. Our *simBALF* and *simPlasma* datasets have no defined quantitative differences, and only the defined subsets (990 and 609 peptides, respectively) have potential qualitative differences based on a non-random sampling algorithm. To introduce the defined qualitative differences, we used a non-random sampling algorithm to remove data, and additional data were removed from each dataset in an entirely random manner (see Methods). The simulated datasets actually each consist of many simulations: for both *simBALF* and *simPlasma*, there are eight levels of missing data (15 to 50%, in 5% increments), and at each level we generated 100 independent simulations. This approach provided the datasets necessary to rigorously assess the true positive (TPR) and false positive (FPR) rates of identification of data with qualitative differences. That is, a single dataset with a defined global fraction of missing data would give a single point estimate of a TPR and FPR, whereas with 100 datasets at each level, we could calculate error bars on the TPR and FPR. For each simulation and analysis, the peptides that were significant at a p-value ≤ 0.05 were used to derive the TPR and FPR. In this way, we evaluated ANOVA and the G-test on their ability to identify qualitative differences, i.e., the censored simulated data.

ANOVA is based on a comparison of variance between treatment groups to within treatment groups. Figure S2 shows that when using a simple ANOVA (or a t-test), the false positive rate (FPR) was ~5% for the simulated datasets, and the true positive rate (TPR) was consistently less than 5%, independent of the total amount of missing data. This analysis demonstrates that, as expected, ANOVA-based analyses in the presence of missing data do not identify censored peptides as significant beyond what is expected by chance because there are not sufficient data to accurately compute the necessary estimates of means and variances.

ANOVA fails to identify censored peptides; however, a common strategy in proteomics research is to impute the missing values prior to statistical analysis. A simple, common strategy to impute missing peptide intensity values is to insert, for each given peptide with missing values, a simple constant such as ½ of the minimum observed abundance of that peptide [6]. This quasi-LOD imputation assumes that the missing values are due to analytical sensitivity. More advanced approaches, such as K-nearest neighbors (KNN) or singular value decomposition (SVD), assume a correlative structure in the data, so that peptides with similar intensity patterns can be used to infer the intensities of the missing peptide values. Our simulated datasets were designed so that the occurrence and intensity of each peptide was independent from the other peptides, and thus KNN and SVD approaches are inappropriate for these data, in fact yielding worse results than ANOVA (data not shown). This data structure is a major caveat of simulated data, however our simulated datasets achieve the purpose of testing the assumptions of ANOVA in the presence of missing data. KNN was evaluated on the experimental datasets described in the paper.

Because the simulated datasets do have not quantitative differences beyond that expected by random chance (validated in Figure S2), they were used determine the TPR and FPR of censored peptide identification. For ANOVA-based approaches with imputed data, and the G-test described in Methods. Specifically, for each of the 100 simulations for *simBALF* and *simPlasma* at a pre-defined level of missing data, three statistical tests were employed to identify peptides with qualitative differences between the experimental groups when missing values were imputed with the LOD. The first was the basic ANOVA approach. The second was a non-parametric version of ANOVA (Kruskal-Wallis test), which is most appropriate when the underlying assumptions of ANOVA are not met. In particular, imputation with a constant will typically invalidate the normality assumption of ANOVA, making Kruskal-Wallis more appropriate[22]. Lastly, the G-test was used, which gives a significance value associated with the null hypothesis that the data are missing at random across treatment groups for the peptide. For the G-test, if a peptide is observed in all samples, the p-value for the G-test would be 1; the larger the p-value, the more likely the data are simply missing in a random manner with respect to the experimental groups. The TPR and FPR were computed from peptides with a p-value ≤ 0.05, and the averages of the TPRs and FPRs achieved across the 100 repeated simulations are shown in Figure S3. Similar trends were observed for the two simulated datasets, however higher TPR are achieved for simulated dataset 2, as the likelihood of repeated sampling from the same group was higher (see Methods).

The results in Figure S3 show that, on the simulated datasets, all the statistical tests exhibited a relatively constant TPR across the levels of missing data (from 15 to 50%), and that the G-test identifies more true positive peptides as significant (TPR) than ANOVA or Kruskal-Wallis with a LOD imputation. Importantly, the G-test also exhibits an equivalent or a smaller FPR than either ANOVA-based method, and the FDR defined by the G-test is sensitive to the percentage of missing data. At low fractions of missing data, less than 30%, the G-test had extraordinary low FPRs in comparison to ANOVA and Kruskal-Wallis. This is because with low levels of random missing data among the ‘non-censored’ peptides, the G-test will find nearly zero false positives; since it is specifically designed to test the null hypothesis that the number of missing observations is independent of the groups. Yet, ANOVA and Kruskal-Wallis will still find ~5% of these false positives significant by random chance because of quantitative differential abundance. However, above ~30% total missing data, the FPR for the G-test was relatively stable and similar to the FPR estimated for the ANOVA and Kruskal-Wallis analyses. Overall, at all levels of missing data, the G-test out-performs ANOVA-based methods at identifying true censored data from the simulated values.