SimpliFi: a GPU-driven data-to-meaning analytics engine bringing omics understanding to all

Jim Palmeri,
Darryl J.C. Pappin,
John P. Wilson*
ProtiFi, LLC, Farmingdale, NY
*To whom correspondence should be addressed: john@protifi.com.

PROTIFI Innovative omics solutions

I. Introduction

We often employ omics analyses like proteomics and metabolomics to try to understand different states: what makes this healthy and that diseased? The results of such techniques then beg the question “What’s different?”. Commonly applied statistical tests are parametric: t-tests, ANOVA, etc. and predicated on a Gaussian distribution of data, which mass spec data frequently do not follow.

To prevent invalid conclusions from flowing from false assumptions, we developed a SimpliFi, a novel data to meaning engine that uses exclusively nonparametric statistics that assumes only that data model themselves. It determines p-values and fold-changes as a function of biological variation, number of samples and observations and measurement error. We employ resampling to generate confidence intervals for all values, including p-values, to give researchers insight into often-ignored uncertainties of statistical difference; routines run on efficient parallel GPU architecture, reducing analysis times from days to minutes. Crucially, SimpliFi accounts for increased data variance at low or high intensities.

SimpliFi’s user interface is intuitive and user-friendly even for new-to-omics users. All intensive computation is done on SimpliFi servers and users view their results on any browser; results can be shared by simply sending a URL. To our knowledge, SimpliFi is the first platform to combine unbiased statistics, the efficiency of GPU-based algorithms, and an interactive and intuitive user interface to allow anyone to understand their data.

II. Approach

Statistical approach

1. Two means are indistinguishable without further knowledge of their scatter = distribution.
2. Biological replicates yield biological empirical distributions which give p-values in comparisons between groups.
3. Bootstrapping handles all distributions and yields confidence intervals for p-values and fold changes.
4. Measurement error is a function of measurement intensity and must be controlled for
5. Condition p-values represent “the chance the observed difference between conditions could occur due only to intrinsic variability within ‘same’.” Alternatively, they are “FDRs of different.”
6. This computational framework is applicable to all kinds of biological entities (protein, peptide, metabolite, etc.), and allows unified analyses of multiomics data.

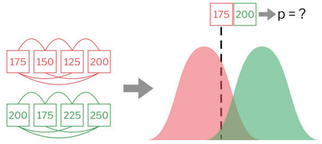

Figure 1 Illustration of SimpliFi p-values being obtained by empirical distribution.

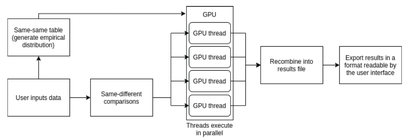

Figure 2 Flowchart depicting GPU parallel processing scheme for SimpliFi p-values.

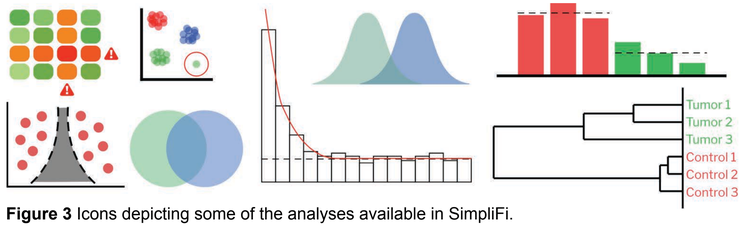

Figure 3 Icons depicting some of the analyses available in SimpliFi.

III. Method and results

1) Statistical tests: isobaric ratios and MS data are not Gaussian

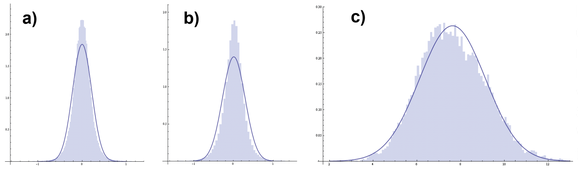

Figure 4 iTRAQ reporter ion ratios for logs of a) same:same, b) same:different or c) raw protein counts fail the following normality tests: Anderson-Darling, Cramér-von Mises, Jarque-Bera ALM, Kolmogorov-Smirnov, Kuiper, Pearson X^2, Shapiro-Wilk, Watson U^2. Maximum $p < 10^{-49}$, it is unlikely these data are normally distributed.

2) MS signal variability is strongly a function of intensity

Empirical fold-change distributions at different intensities. Relative density of log2 fold changes of biological replicates within one condition (same:same) for given intensity regions. Middle intensities (e.g. 100,000), show less scatter and thus narrower distributions with more confident p-values. However, distributions at the extremes of intensity often exhibit more scatter and thus wider distributions. SimpliFi p-values account for this and report decreased confidence that meaningful differences exist between conditions with intensity regions of increased scatter.

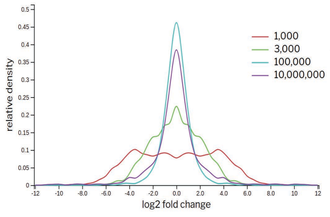

Figure 5 Relative densities of log2 fold changes in windows around four intensity values. Low and high intensities show a higher frequency of higher fold changes than intensities of observation in the middle of the data.

Intensity	1,000	3,000	10,000	30,000	100,000	300,000	1,000,000	3,000,000	10,000,000
fc = 2	0.857	0.635	0.394	0.272	0.298	0.324	0.324	0.409	0.466
fc = 4	0.672	0.352	0.115	0.063	0.073	0.107	0.121	0.161	0.238
fc = 8	0.470	0.162	0.031	0.018	0.020	0.036	0.055	0.070	0.129

Table 1 Table of single-comparison p-values for this data set. For example, if the pair (100k, 200k) was observed in this data set, its p-value would be 0.298.

3) “Agreeing” replicates increase certainty of observed difference

With an increasing number of biological replicates, if observed changes between states are in the same direction, p-values become more certain; observations of different directions, or inclusion of fewer replicates, have the opposite effect.

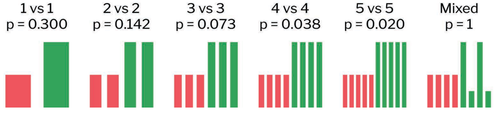

Figure 6 P-value effect of the number, direction of replicates.

Samples	6x6	7x7	8x8	9x9	10x10
p-value	0.011	0.006	0.003	0.002	0.001

Table 2 Additional p-values resulting from n x n equal comparisons when one comparison has a p-value of 0.300.

4) For non-Gaussian biological data, the t-test is insufficient

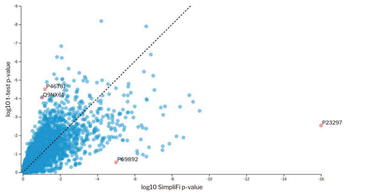

Figure 7 SimpliFi p-values versus t-test p-values. SimpliFi and t-test p-values can often differ by several orders of magnitude due to multiple factors. See Fig. 8 at right.

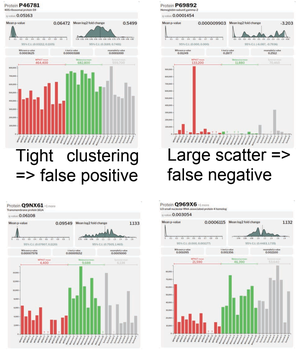

Figure 8 Limitations of t-tests. False positives and negatives result from undersampling of variability or outliers. The bottom panels show the effect of intensity. Despite the same fold change and relatively tight clustering, Q9NX61 observed in the 20k – 50k intensity range is highly significant but Q9NX61 at low intensity (4k - 10k) is not significant. Accounting for intensity drastically reduces certainty.

IV. Screenshots

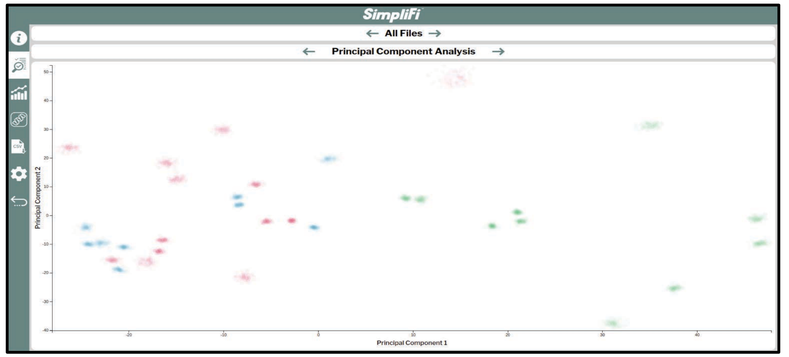

Figure 11 Principal component analysis (PCA) screen, part of the QC step in SimpliFi.

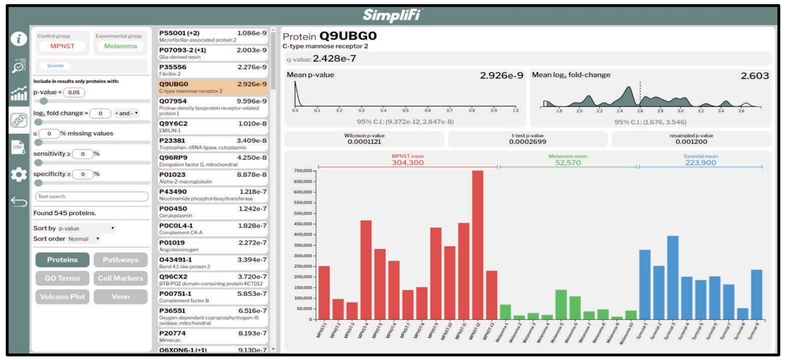

Figure 12 Protein view in SimpliFi, which displays all computed statistics for a single protein.

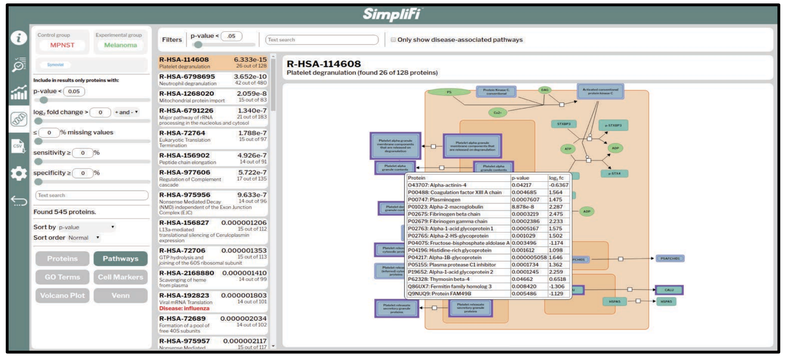

Figure 13 Pathway view in SimpliFi. If a user hovers over a pathway node, a list of proteins in that node is displayed.