Jim Palmeri[1]; Darryl J.C. Pappin[2]; John P. Wilson[1]

# Higher and higher N: SimpliFi for the masses now takes masses of samples

SimpliFi™

[1]ProtiFi, LLC, Farmingdale, NY
[2]Cold Spring Harbor laboratory, Cold Spring Harbor, NY

Large cohorts? No worries; they're great!

Ever-increasing amounts of omics data can limit our ability to understand those data and reach actionable decisions. Adding to the challenge, studies continue to grow both in the number of analyzed samples and in the kinds of omics analyses. Previously we developed SimpliFi, a cloud-based, browser-driven data-to-meaning engine with an intuitive, user-friendly accessible to users of all experience levels. We have now recoded both the backend, speeding calculations by orders of magnitude, as well as the frontend, where summary displays now enable human understanding of information present even in large (>100s) sets of samples. SimpliFi accepts and integrates data from all omics analyses and allows results to easily be shared, explored or published by simply sending a URL.
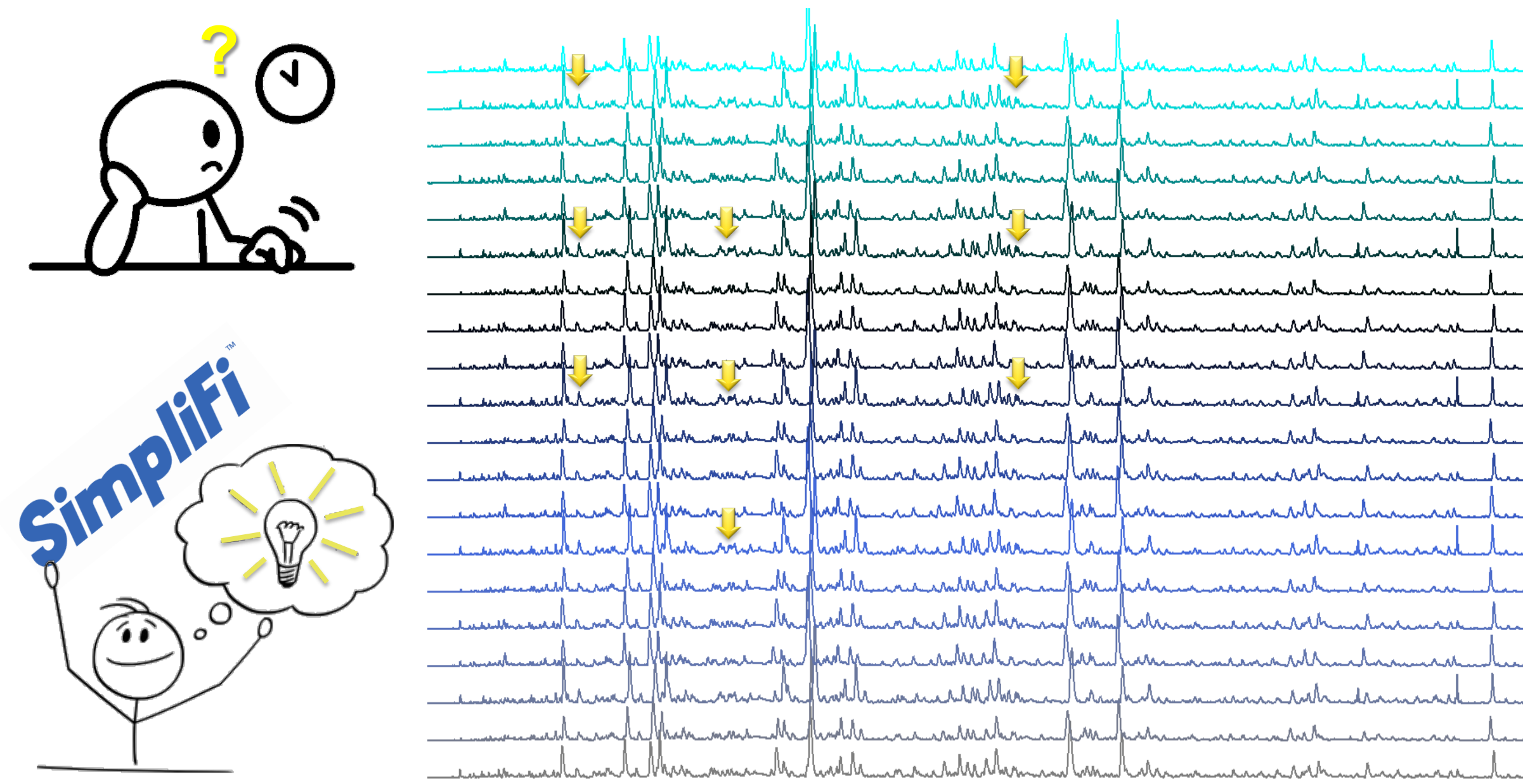
Data analysis (including quality control, visualization, candidate selection and systems biological conclusions) of larger cohorts is not easy and time consuming especially for non-experts. Wrong decision about statistical methods can have fatal consequences. SimpliFi is a software which provides tailored statistical analysis for omics datasets. It is easy to use and intuitive and can be used by both experts and non-experts.

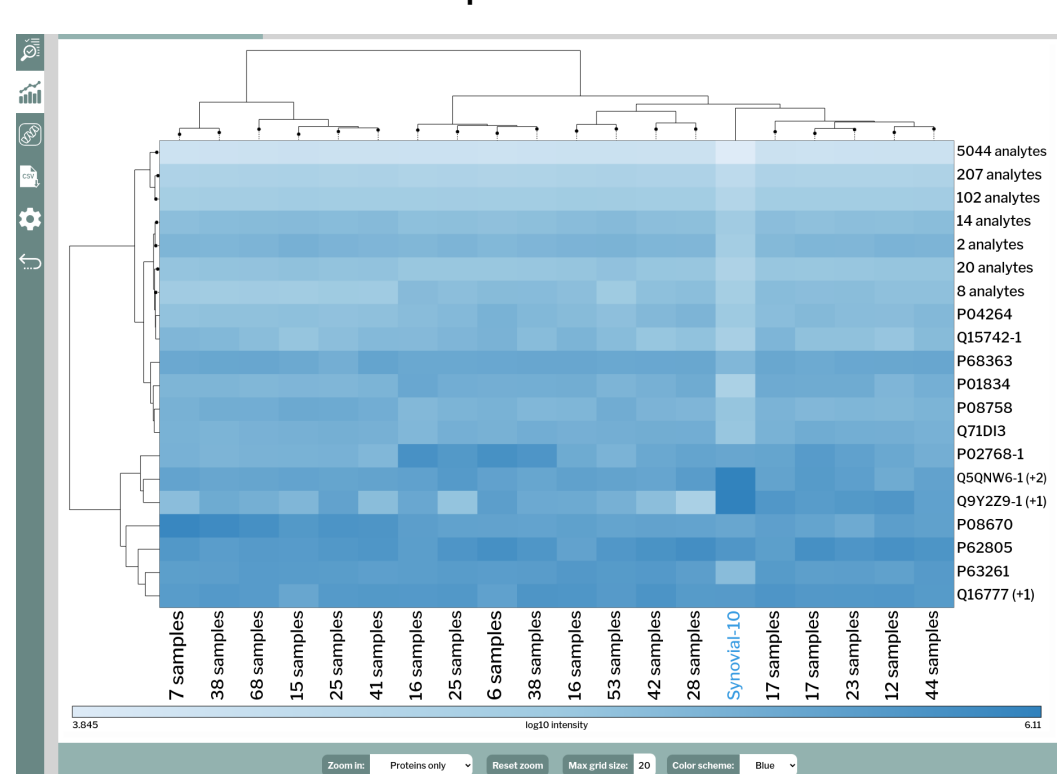## QC and visualization for large N

Quality control (QC) steps are essential for every analytics workflow. In case of large sample cohorts it is even more important to control for batch effects and address reproducibility. SimpliFi's new features are adapted to visualize even large datasets and to interactively inform the user about irregularities.
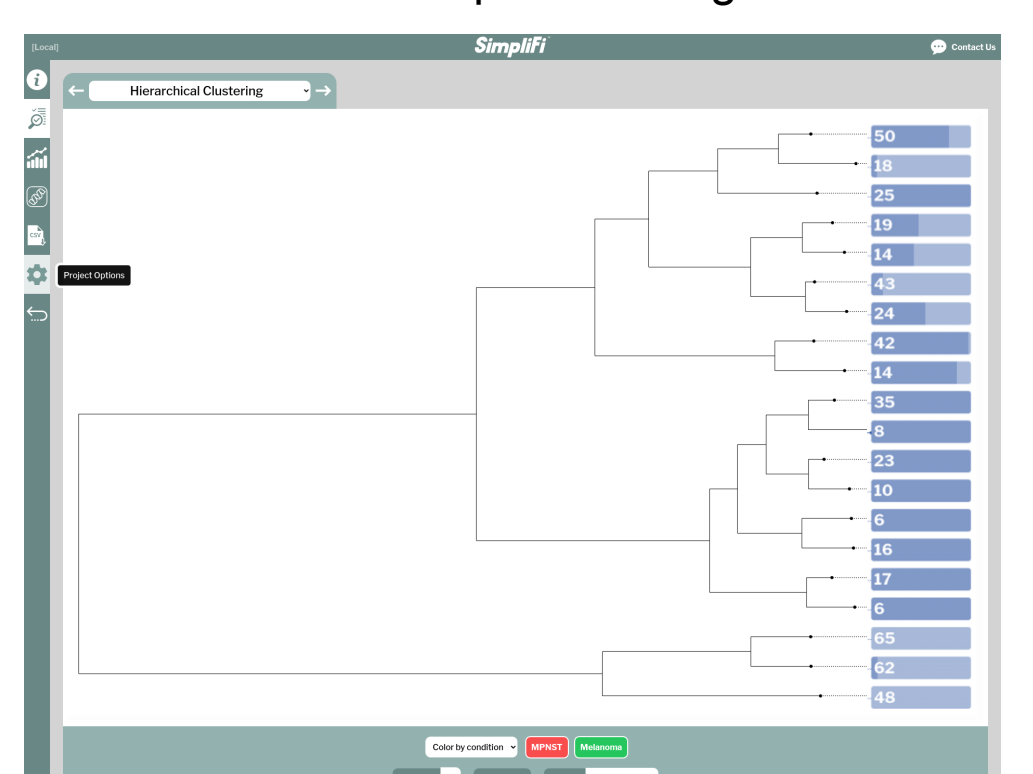

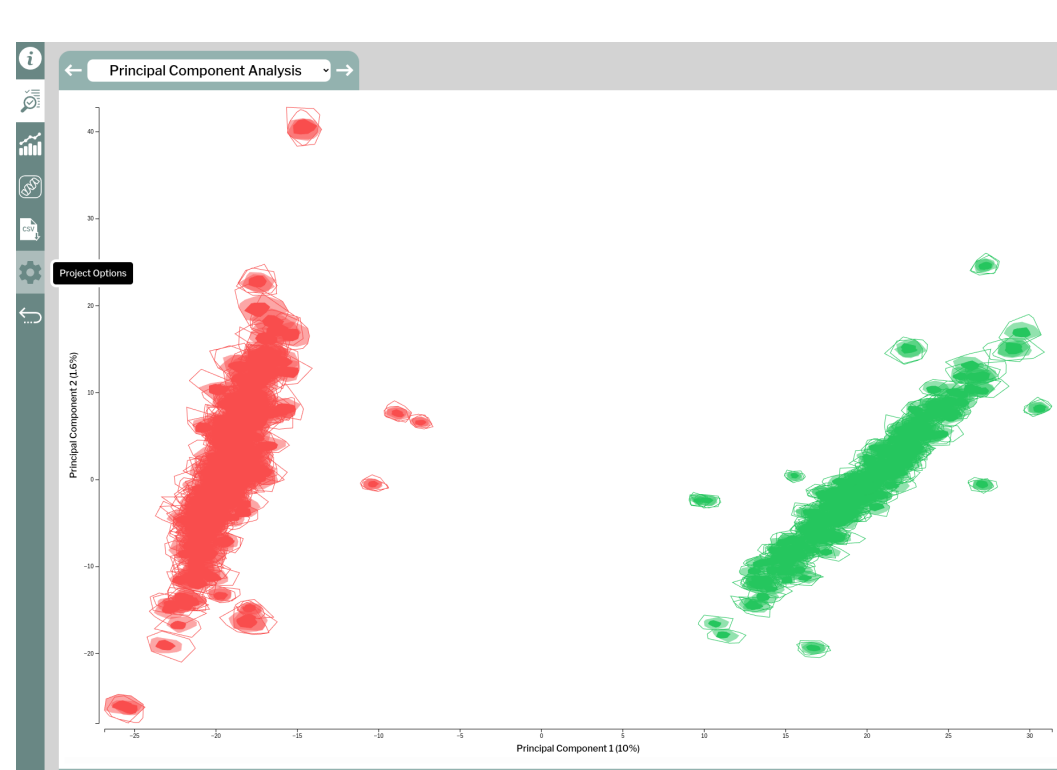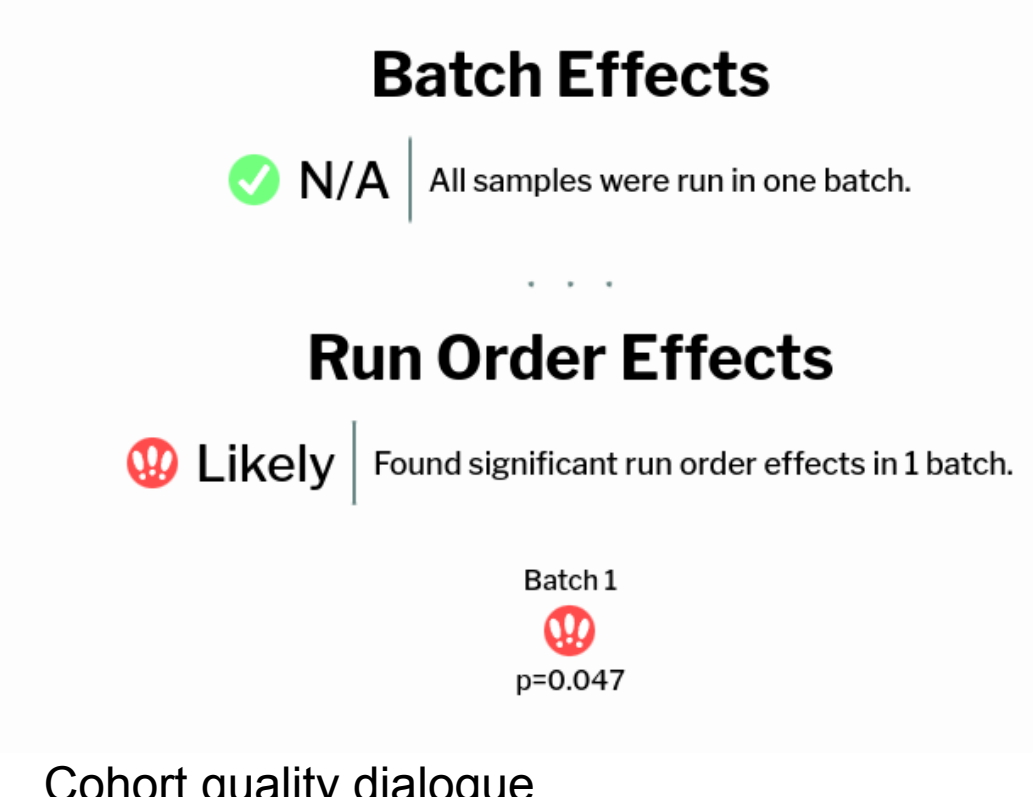Pearson correlation plots

Pearson correlation plots for large datasets

Hierarchical clustering of smaller subset

Hierarchical clustering of large sample datasets

Principal component analysis (PCA)

Batch Effects
✓ N/A   All samples were run in one batch.

Run Order Effects
⚠ Likely   Found significant run order effects in 1 batch.

Batch 1
p=0.047

Cohort quality dialogue
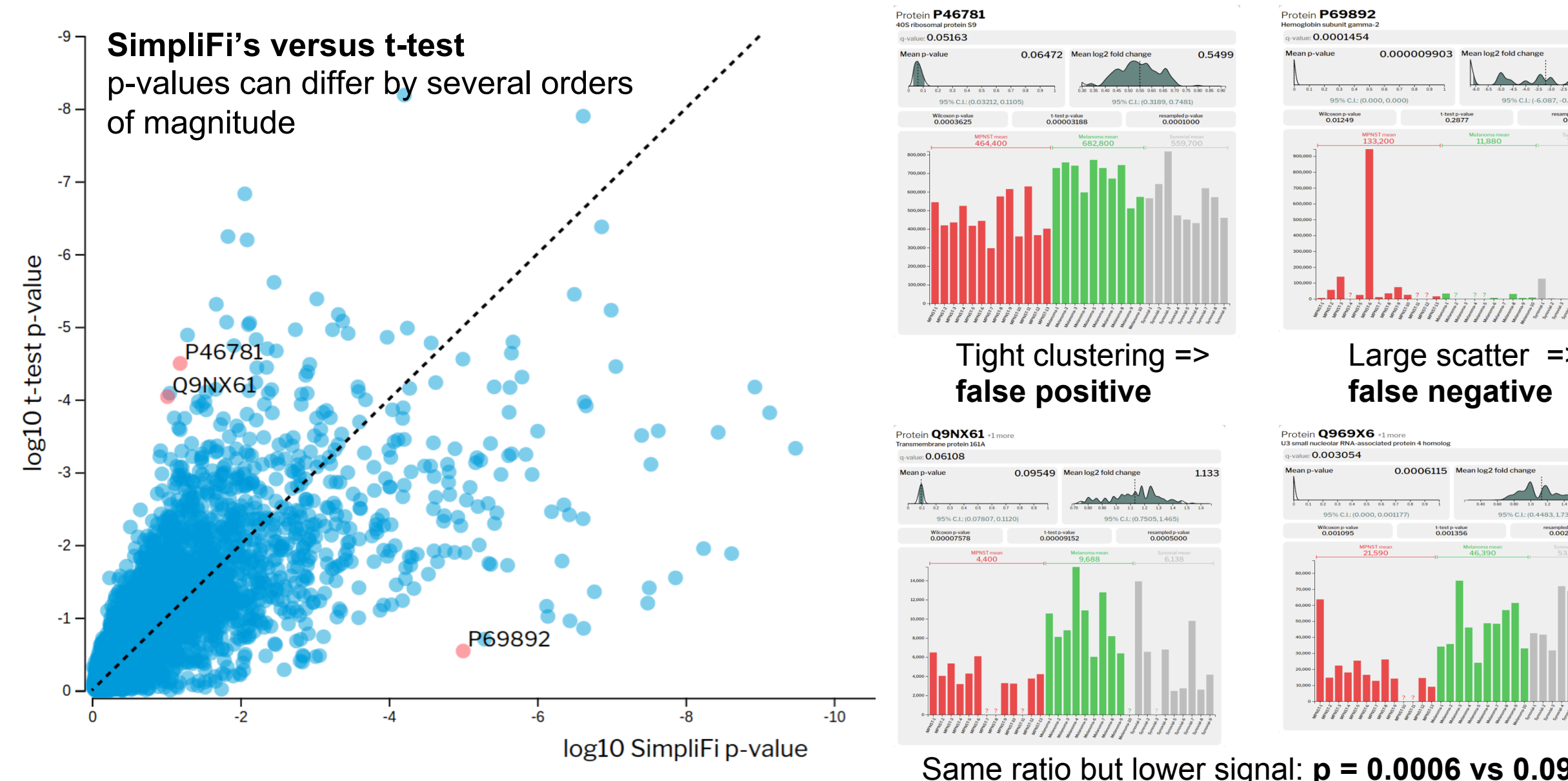
Statistical correctness

SimpliFi models biology using nonparametric statistics in which sample replicates define their own distributions; such distributions always fit. These distributions are frequently non-normal (insert below) and thus do not satisfy the assumptions of many classic statistics tools, easily resulting in wrong conclusions. SimpliFi automatically accounts for increased data variance at low or high intensities: at low intensity, low stochastic sampling results in expectedly high variation, while at high intensity, effects like saturation become apparent. Importantly, p-values and fold-changes are always reported with confidence intervals.
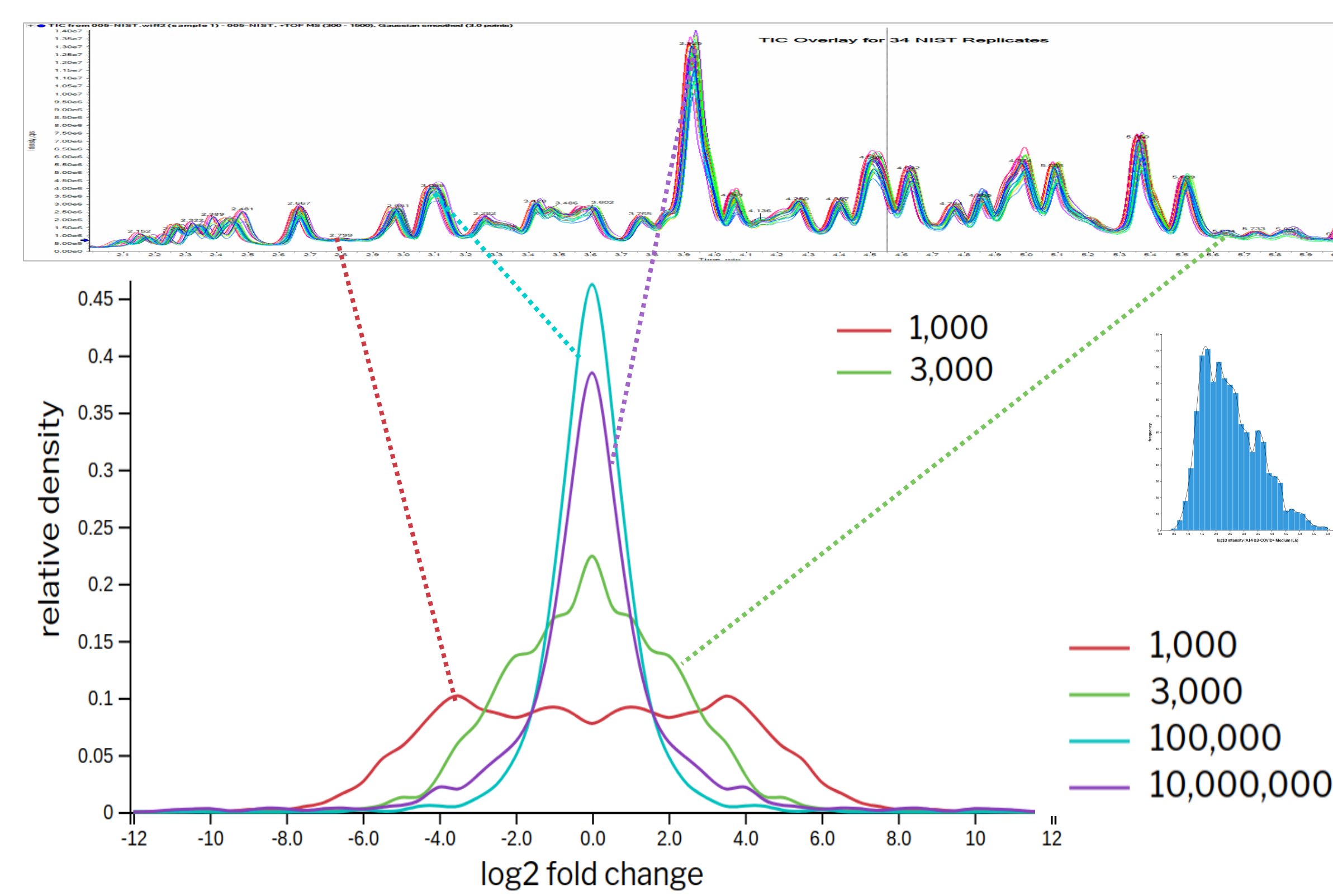
Effect of intensity of observation on observed variability expressed as fold-changes between replicas of the same condition.. In LC-MS-based proteomics, variance if a function of feature intensity wherein low intensity measurements are subject to large amounts of stochastic sampling error and high intensity measurements encounter effects such as saturation. SimpliFi accounts for the individual variance of every feature as a function of intensity.


| 1 vs 1 | 2 vs 2 | 3 vs 3 | 4 vs 4 | 5 vs 5 | Mixed |
| p = 0.300 | p = 0.142 | p = 0.073 | p = 0.038 | p = 0.020 | p = 1 |
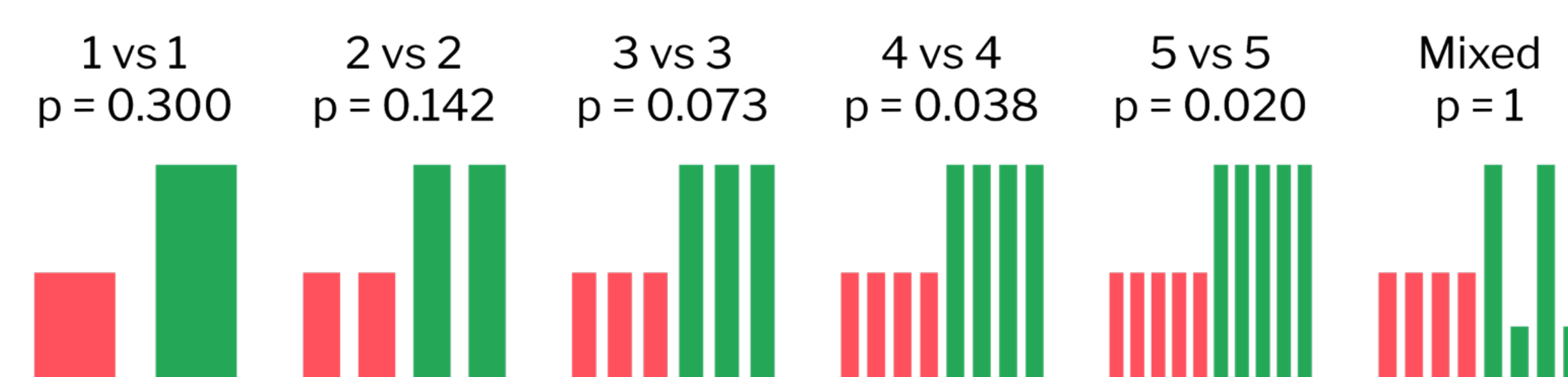
Effect of number of replicates that agree, or don't. With an increasing number of biological replicates, if observed changes between states are in the same direction, p-values become more certain; observations of different directions, or inclusion of fewer replicates, have the opposite effect. Certainty increases with increased numbers of agreeing biological replicates and p-values become more certain if observed changes between states are in the same direction. In contrast, observations of different directions or fewer replicate numbers, have the opposite effect.


SimpliFi's versus t-test
p-values can differ by several orders of magnitude

Tight clustering => **false positive**

Large scatter => **false negative**

Same ratio but lower signal: **p = 0.0006 vs 0.096!**

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, 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.

## Interactive plotting for understanding

Data can be explored and visualized with multiple interactive tools including volcano plots, distribution plots, heat maps, etc. SimpliFi's on-the-fly response produces a machine-human interface that facilitates human intuition guiding data exploration. Users of all skill levels can take deep dives into the data and share projects via a simple URL.
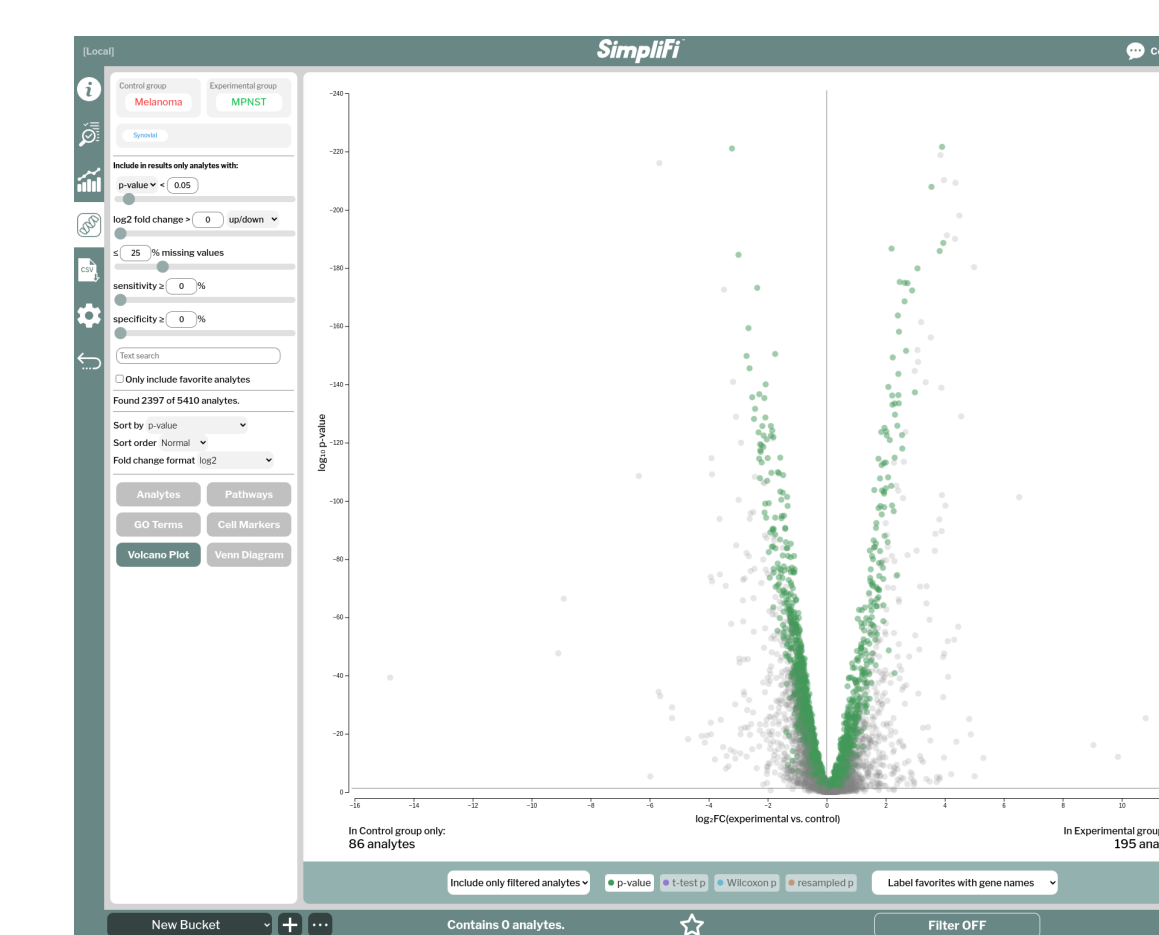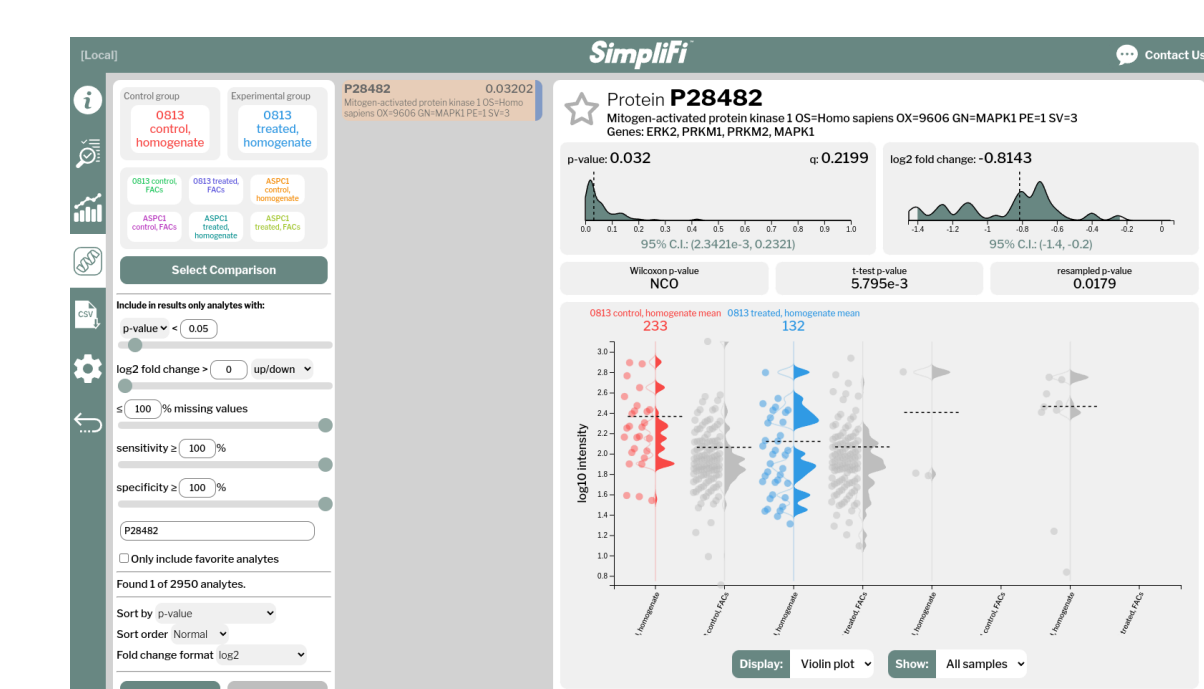

Interactive volcano plots for comparison of conditions
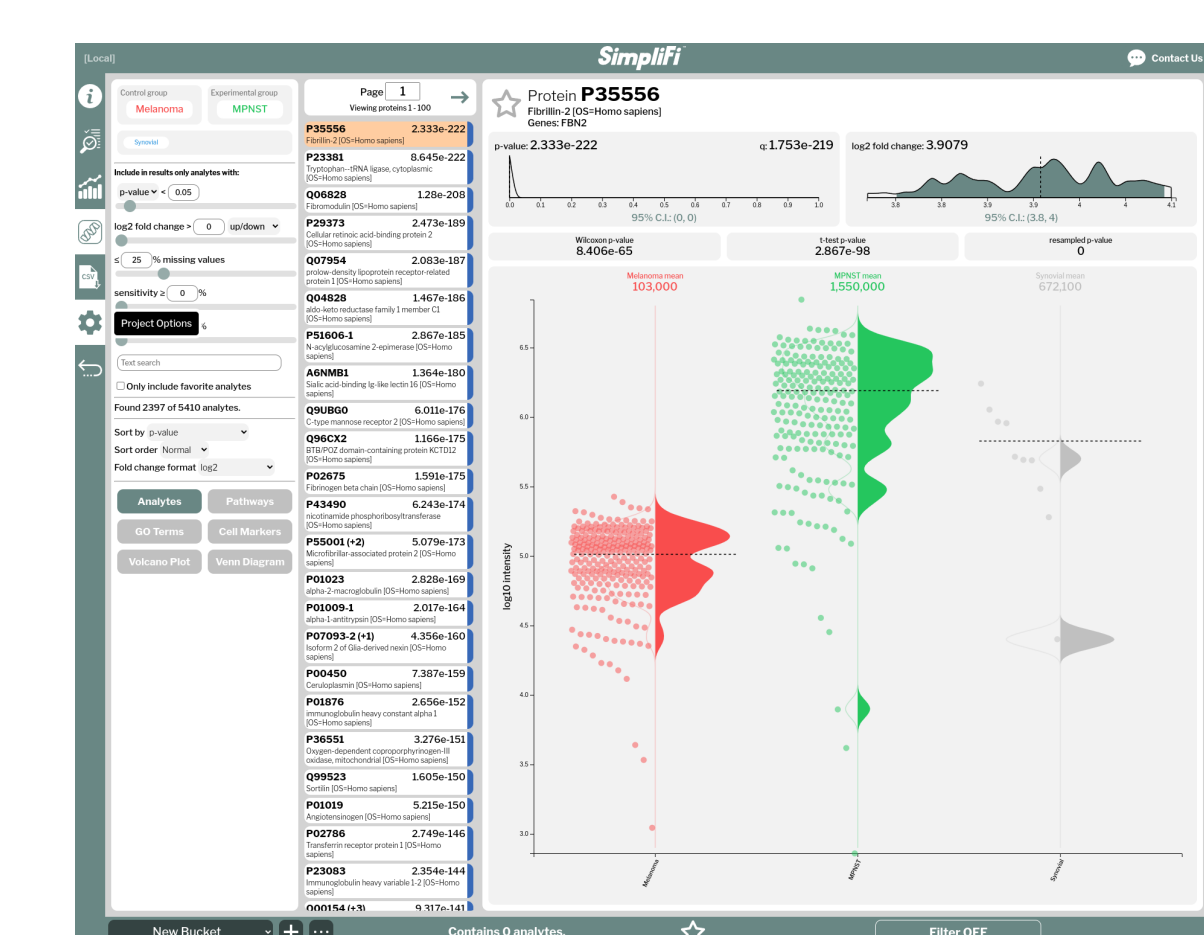
A click on each protein displays exact information
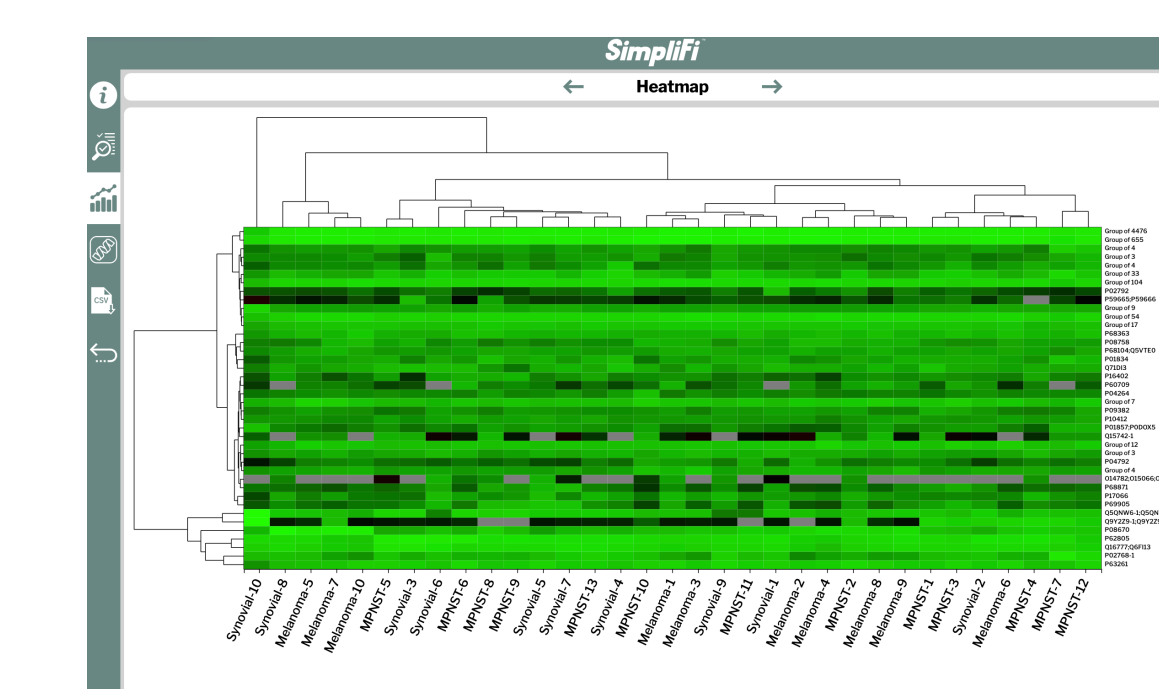
Heatmaps help finding regulation patterns

Easily compare conditions in multiple comparisons.

Data-to-meaning via reactome integration

Understanding a dataset requires understanding of regulation within the biological systems. SimpliFi provides tools to map data to pathways and analyze cellular compartments and biological functions. SimpliFi uses the reactome pathway database to quickly understand biological effects.
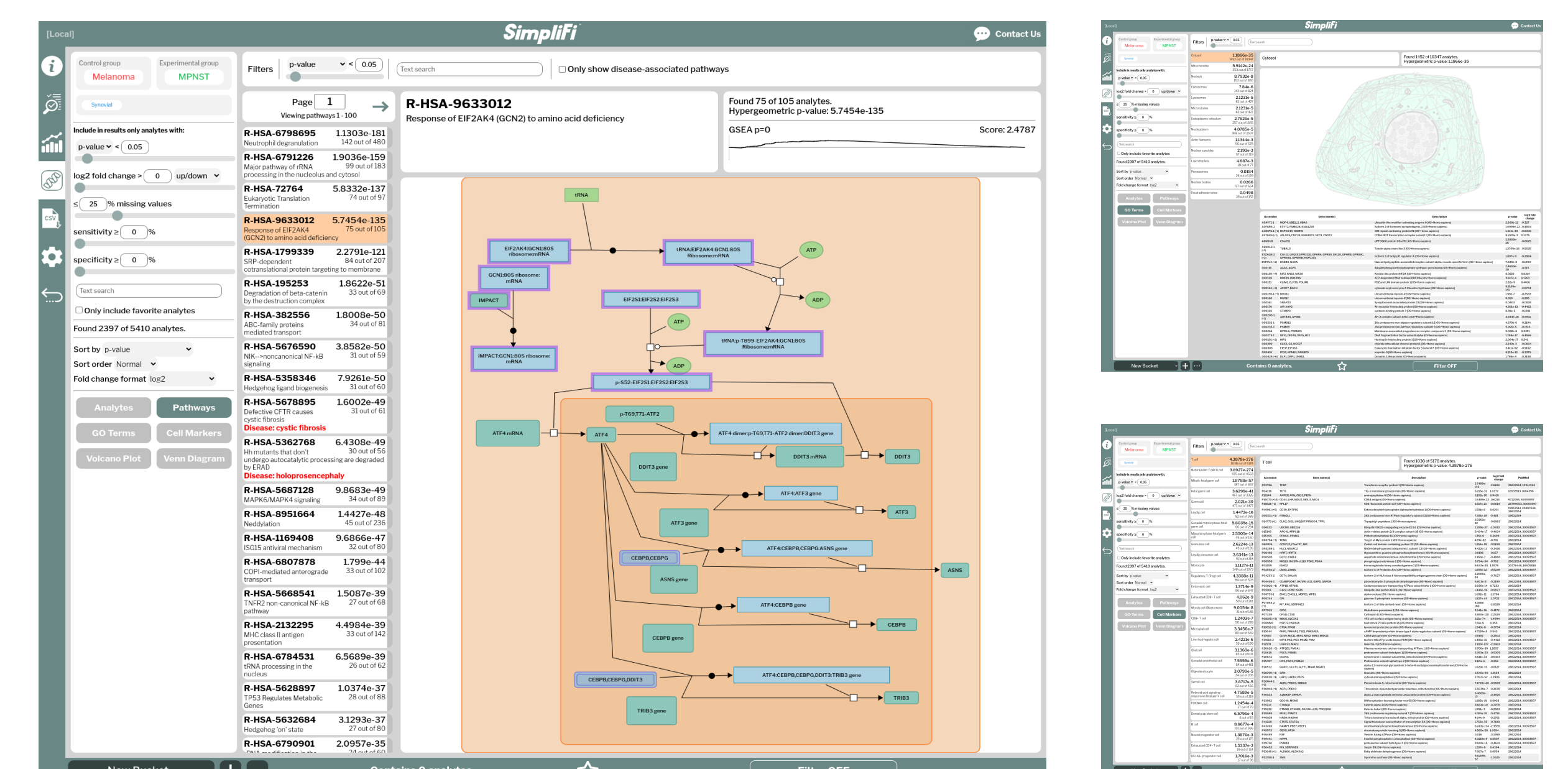
Pathway- and cellular compartments maps in conjunction with GO annotation allow for deep exploration of proteomics, metabolomics, lipidomics, transkriptomics and genomics datasets

In Summary: SimpliFi your data!

- Straightforward analysis of any size omics datasets for all experience levels
- Non-parametric statistics defined by the data themselves yield statistically correct p-values confidence intervals for all features
- Quality control tools ensure identification of suspect samples
- Interactive visualization tools allow rapid exploration of datasets to quickly derive meaning from data
- Projects can be conveniently shared via URLs with any associate rendering it perfect for joint collaborations.