

# A combined LC-MS and LC-MS/MS inferential framework for reliable and reproducible discovery of candidate biomarkers.

Olga Vittek<sup>1</sup>, Andrew Garbutt<sup>1</sup>, Mi-Youn Brusniak<sup>1</sup>, David Campbell<sup>1</sup>, James Eddes<sup>1</sup>, Simon Letarte<sup>1</sup>, Daniel Martin<sup>1</sup>, Julian Watts<sup>1</sup>, Ning Zhang<sup>1</sup>, and Ruedi Aebersold<sup>1,2</sup>.

<sup>1</sup>Institute for Systems Biology, 1441 North 34th Street, Seattle WA 98103, USA and <sup>2</sup>Institute of Molecular Systems Biology, ETH Zurich HPT E 78 CH-8093 Zurich, Switzerland

We introduce **Corra**: An LC-MS based framework for determination of differentially abundant peptides and proteins that:

- Assigns MS/MS-derived sequence identities to LC-MS features
- Determines differentially abundant features
- Combines features to determine changes in protein abundance
- Outputs a list of candidate features or proteins with a controlled proportion of false positives

## Introduction

We are developing a computational and statistical framework for the discovery of candidate peptide/protein biomarkers for both diagnosis of human disease and monitoring of therapeutic response. The Corra framework benefits from combining the sensitivity of LC-MS data with the sequence identification specificity of LC-MS/MS.

The framework (1) identifies, (2) annotates, (3) aligns and (4) quantifies LC-MS features. We utilize existing ISB computational tools in combination with the development of new tools specific for the needs of the Corra project workflow. Rigorous statistical analyses are performed to confidently detect differentially abundant LC-MS features.

Knowing the sequence identities of some of the LC-MS features plays an important role for data analysis. Firstly, they help with the alignment of LC-MS data and allow us to validate workflow performance. Secondly, we can infer changes in protein abundances by combining information from LC-MS features sharing the same protein identity.

## Data sets

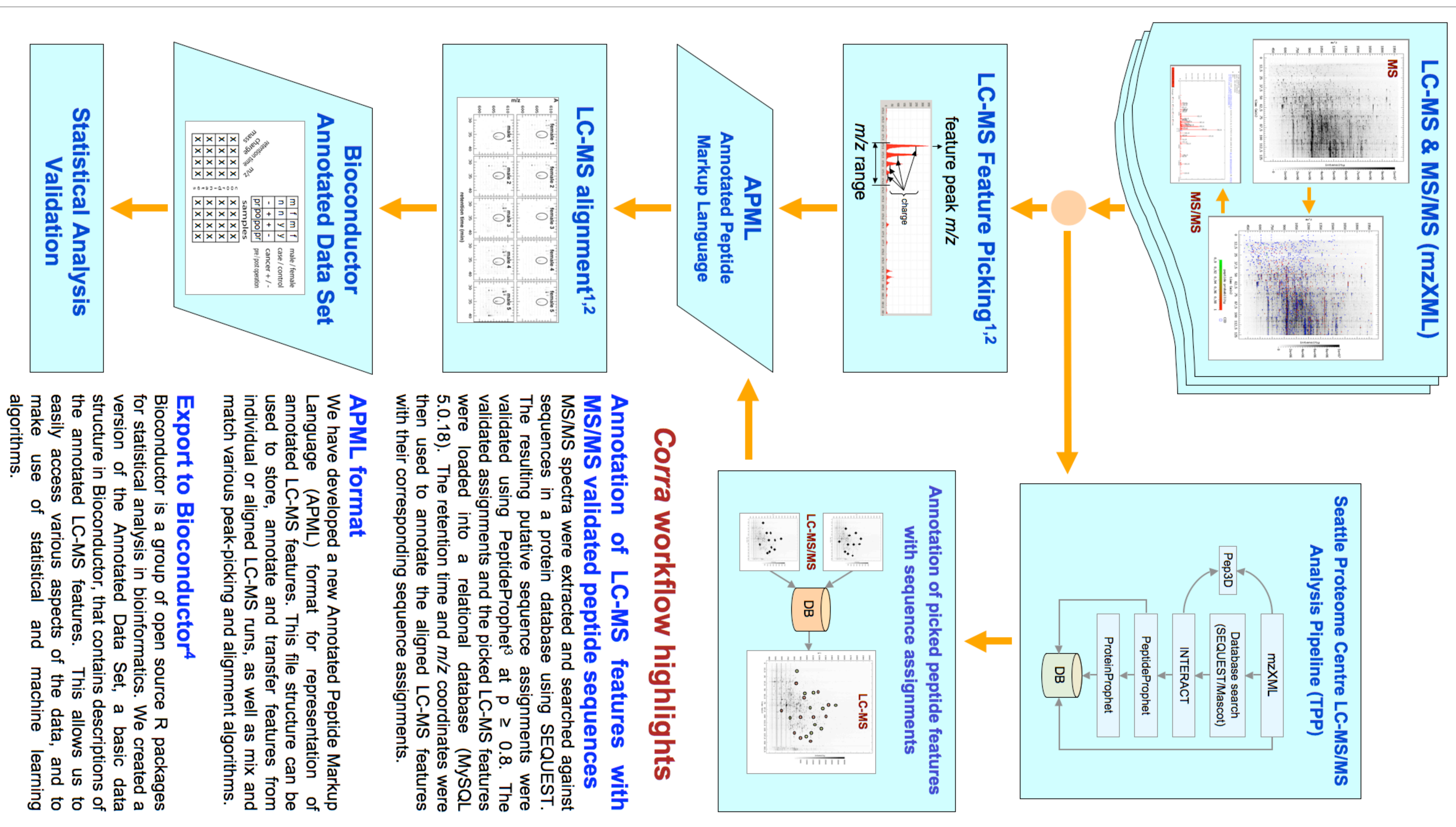
	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
AMHO	400	12.5	25	50	100	200
CAH2_BOVIN	200	400	12.5	25	50	100
CYC_HORSE	100	200	400	12.5	25	50
LYSC_CHICK	50	100	200	400	12.5	25
ADH1_YEAST	25	50	100	200	400	12.5
ADREDA	12.5	25	50	100	200	400

**Fig. 1** We validate the performance of the workflow using a spike-in experiment. Numbers in the table are concentrations in fmol/ml of proteins spiked in a Latin Square design into a complex biological mixture (peptides isolated from human serum). An advantage of the design is that it allows to study multiple fold changes in a balanced manner, using multiple proteins and with a limited number of runs. The samples were analyzed in LC-MS/MS mode (QSTAR ESI-Q-TOF).

	Diabetes		Control	
	Male	Female	Male	Female
Feature 1	P1	P2	P1	P2
Feature 2	1 2 3	1 2 3	1 2 3	1 2 3
Feature 3	x x x	x x x	x x x	x x x

**Fig. 2** A small pilot study of human type 2 diabetes allows us to understand the performance of the workflow on real data, and to assess the relative importance of individual variation, variation due to sample preparations, and run-to-run variation. The samples were analyzed in LC-MS/MS mode (QSTAR ESI-Q-TOF).

## Methods: Corra computational workflow



## Corra workflow highlights

**Annotation of LC-MS features with MS/MS validated peptide sequences**  
MS/MS spectra were extracted and searched against sequences in a protein database using SEQUEST. The resulting putative sequence assignments were validated using PeptideProphet<sup>3</sup> at  $p \geq 0.8$ . The validated assignments and the picked LC-MS features were loaded into a relational database (MySQL 5.0.18). The retention time and  $m/z$  coordinates were then used to annotate the aligned LC-MS features with their corresponding sequence assignments.

### APML format

We have developed a new Annotated Peptide Markup Language (APML) format for representation of annotated LC-MS features. This file structure can be used to store, annotate and transfer features from individual or aligned LC-MS runs, as well as mix and match various peak-picking and alignment algorithms.

### Export to Bioconductor<sup>4</sup>

Bioconductor is a group of open source R packages for statistical analysis in bioinformatics. We created a version of the Annotated Data Set, a basic data structure in Bioconductor, that contains descriptions of the annotated LC-MS features. This allows us to easily access various aspects of the data, and to make use of statistical and machine learning algorithms.

## Methods: statistical analysis

Differentially abundant LC-MS features are detected using a linear mixed model (Limma<sup>5</sup> package in Bioconductor<sup>4</sup>). Each feature is characterized by its fold change and its variance. A moderated T-statistic ranks the features by evidence for differential abundance.

Differentially abundant proteins inferred with a meta-analysis<sup>6</sup> hierarchical linear model:

*Change in abundance = true change + between-feature variation + within-feature variation.*

We replace the moderated T-statistics of features with shared protein identities with a single combined value.

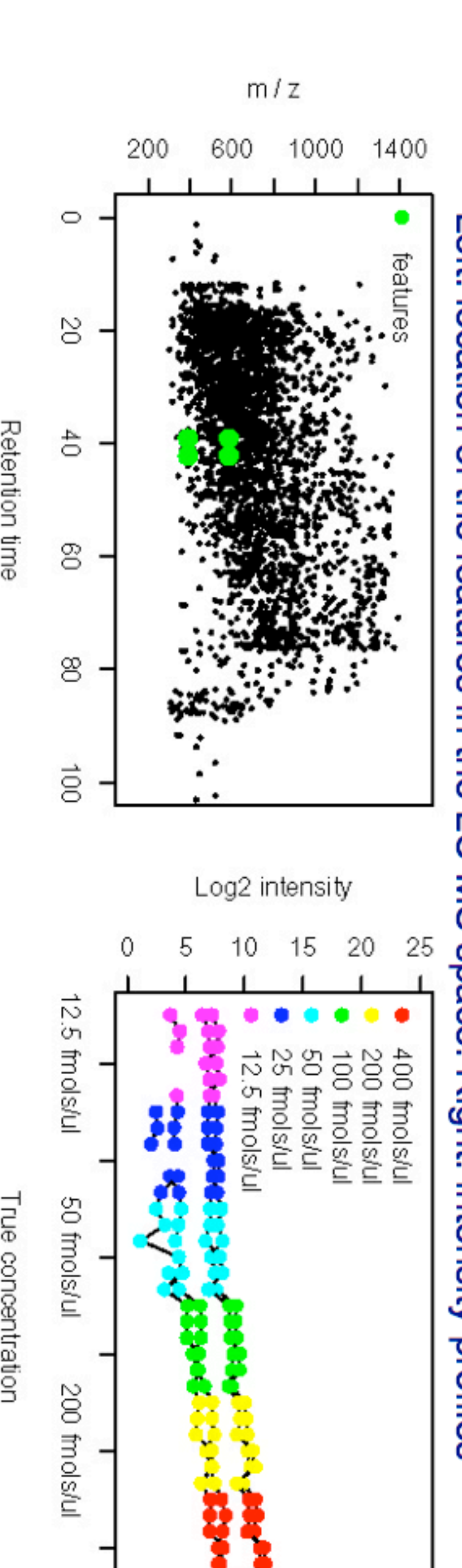
False discovery rate T-statistic threshold is estimated by multiple permutations of group labels in the data set. For each threshold, the error rate is estimated as:

*Average No. of T-statistics above threshold in permutations*

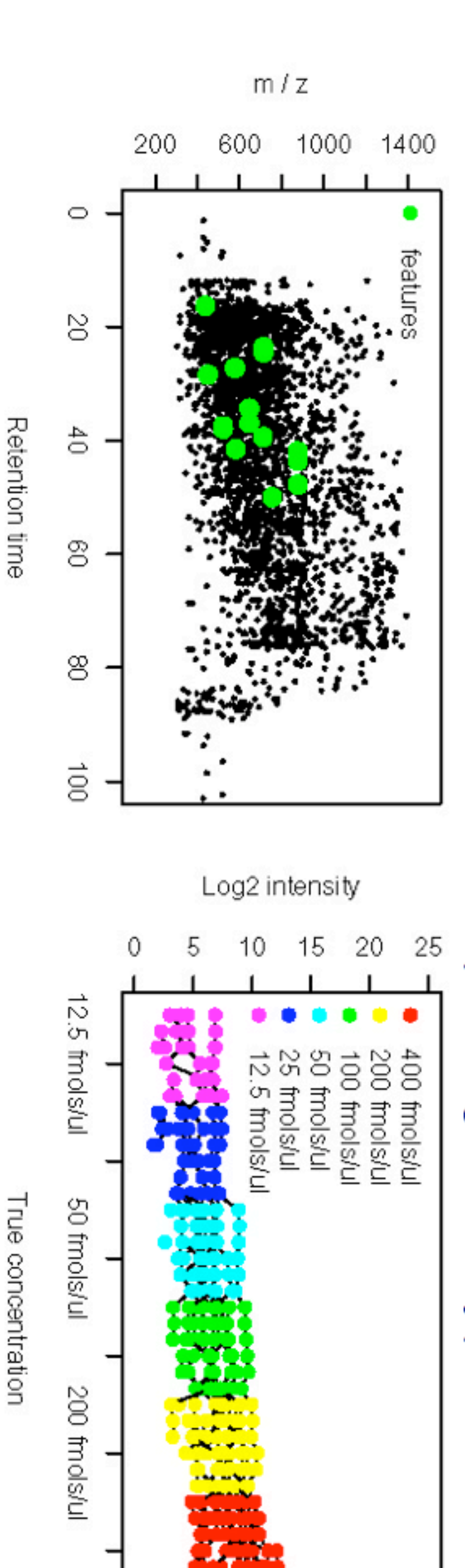
*No. of T-statistics above threshold*

## Results: Latin square

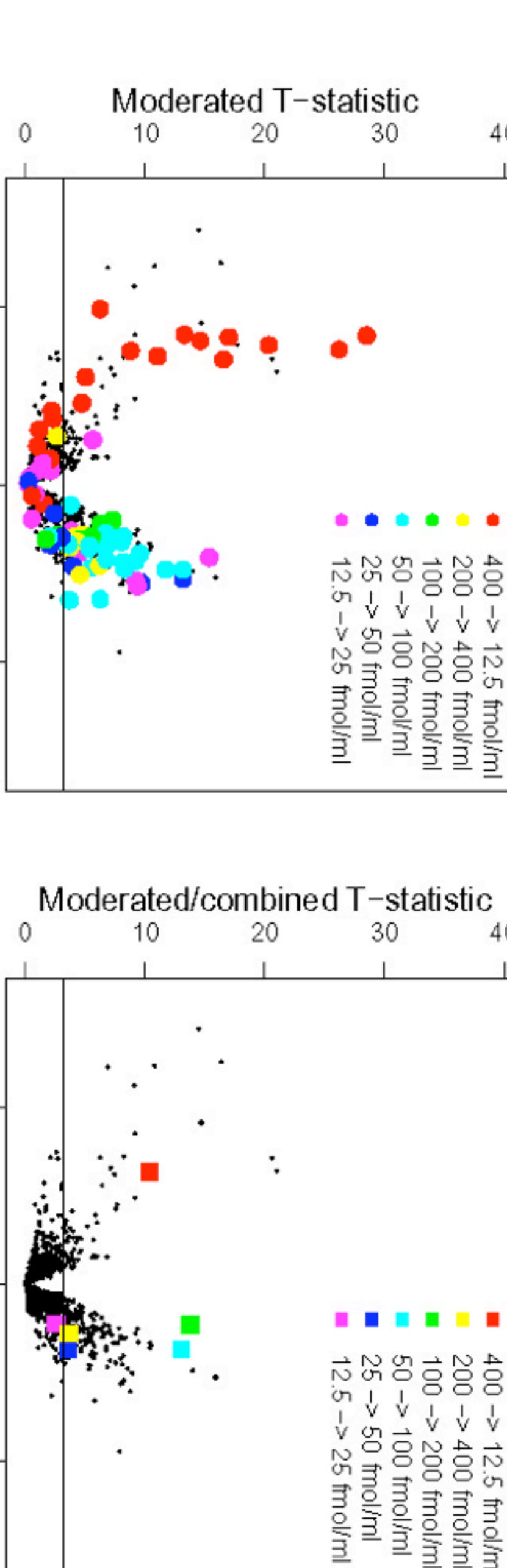
4 features from a same peptide: TGNPLHGFGR



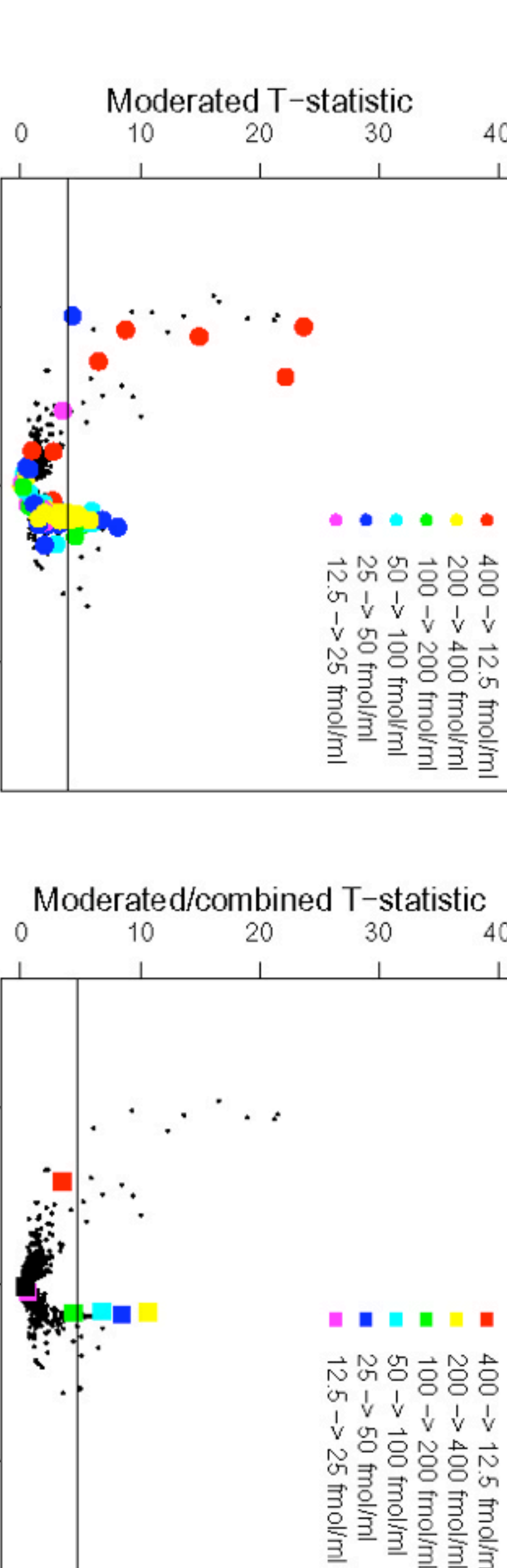
15 features from same protein: LYSC\_CHICK



Sample 1 vs. sample 6. Left: individual features. Right: features combined into proteins

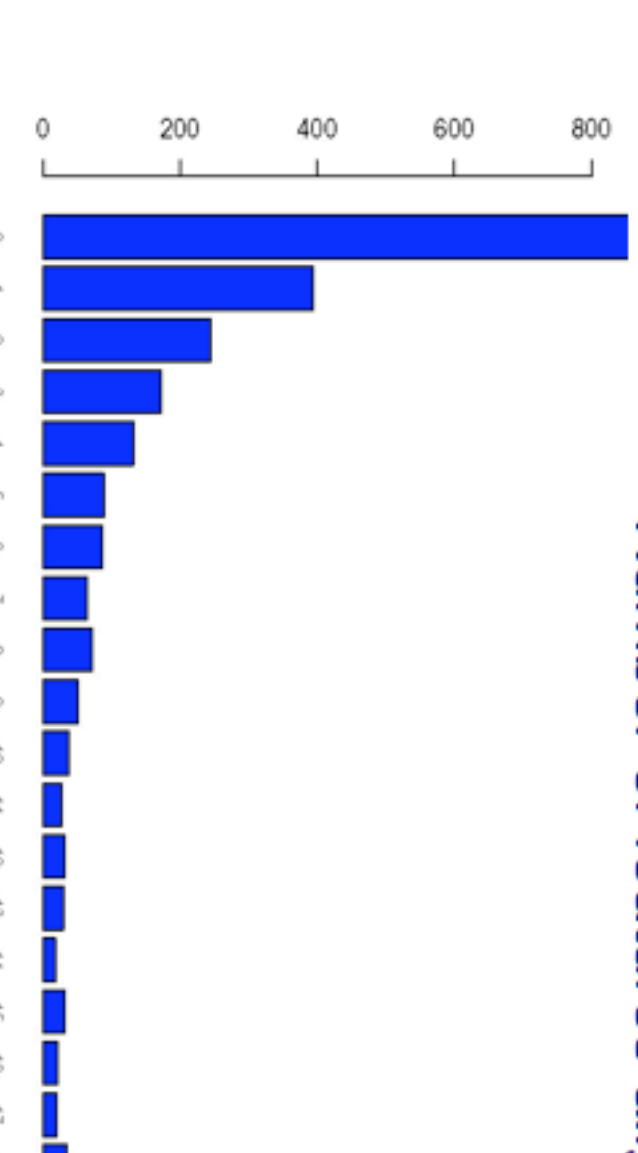


Sample 3 vs. sample 2. Left: individual features. Right: features combined into proteins



## Annotations of LC-MS features

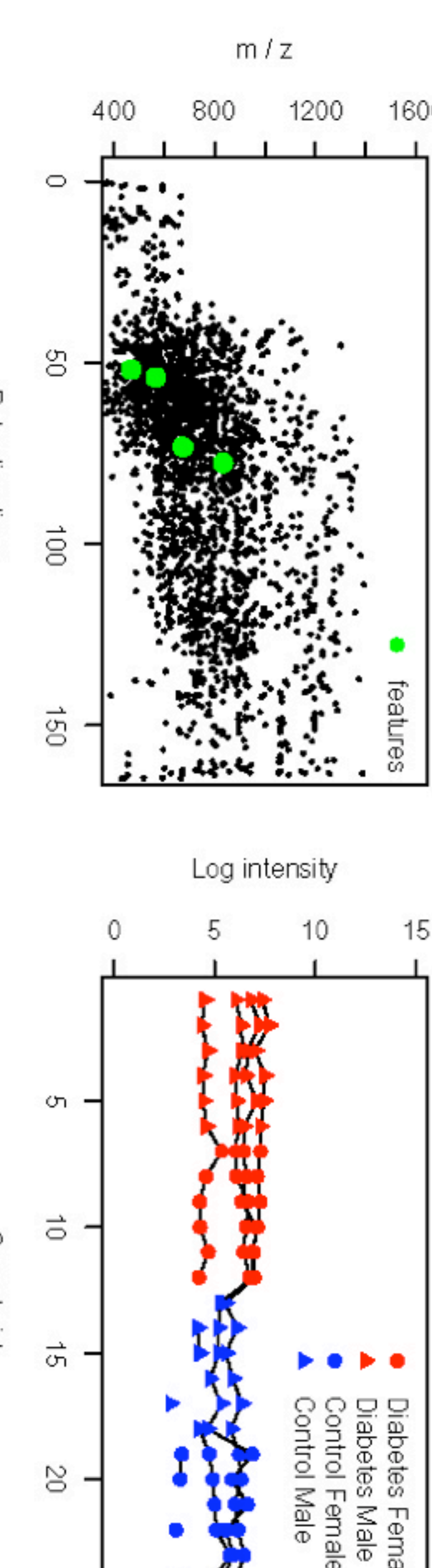
Number of features aligning across multiple LC-MS runs



**Fig. 3** Extent of LC-MS/MS sequence annotations: x-axis: No. of runs in which a feature has sequence annotation, y-axis: count of aligned features. No filtering applied to MS/MS identifications.

## Results: diabetes pilot study

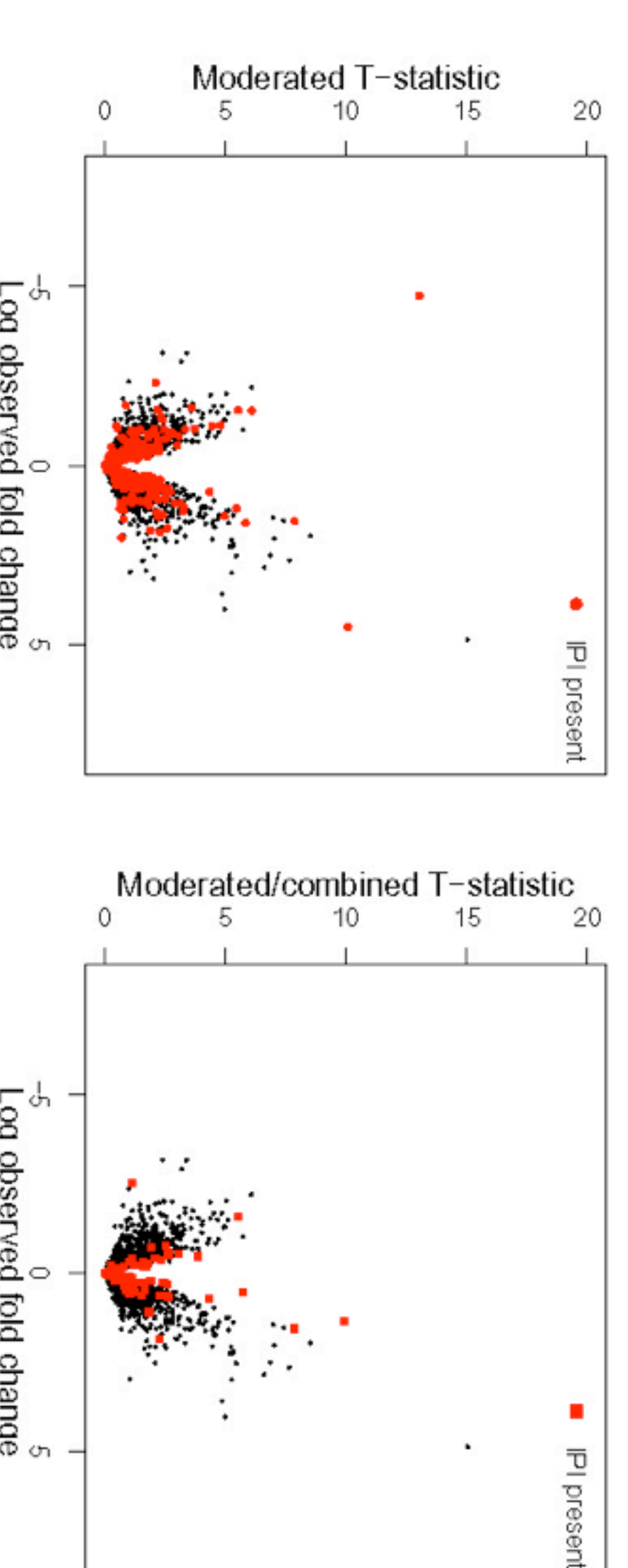
4 features from a same protein: IP100023673



Left: location of the features in the LC-MS space. Right: intensity profiles

**Fig. 4** Intensity profiles of features with shared sequence or protein identities. Results for the Latin square show that there can be a discrepancy in intensities between features from the same peptide. Some of the discrepancy can be explained by "trailing features", i.e. features eluting across a range of retention times. However, most features are informative of the fold change. Thus fold changes can be combined using a hierarchical model to infer the change in protein abundance (see Fig. 5).

Diabetes vs. Control. Left: individual features. Right: features combined into proteins



**Fig. 5** "Volcano" plots illustrating detection of differential abundance. x-axis: observed log fold change in feature intensity. y-axis: T-statistics of differential expression obtained for individual features, or combined across features from a same protein. Colored dots are features with protein identities. Results for the Latin Square show that the workflow is capable of detecting two-fold changes in abundance starting from a relatively low baseline, but the results may be inconsistent across features. Consistency is greatly improved when combining features with shared protein identities. Results for the diabetes pilot study will be validated using a larger sample set

## Conclusions

Combination of LC-MS and LC-MS/MS experimental information allows us to better understand both sample content and the nature of the LC-MS space. This information can be used to validate and improve upon the quantitative characteristics of a non-labelling workflow.

The proposed experimental and Corra analytical framework is a step towards reliable and reproducible discovery of candidate biomarkers for both diagnosis and therapeutic treatments.

## References

1. U. et al. (2005) A software suite for the generation and comparison of peptide arrays from sets of data collected by liquid chromatography-mass spectrometry. *Mol. Cell. Proteomics*, 4, 1328-40.
2. Huang D. et al. (2006) MSAnalyzer: A computational framework for comparative proteomics based on LC-MS analysis of peptides. *Submitted*
3. Keller, A. et al. (2005) A uniform proteomics MS/MS analysis platform utilizing open XML. *Nucleoms. Molecular Systems Biology*
4. Gentlemen et al. (2005) Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Springer, New York.
5. Smyth, C.K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, 3.
6. Choi et al. (2003). Combining multiple microarray studies and modeling intensity variation. *Bioinformatics*, 19, 184-190.