A Tool for Quantitative Analysis of Proteomics Data

Ashoka D. Polpitiya[1], Jared Kirschner[1], Navdeep Jaitly[2], and Konstantinos Petritis[1]
[1]Center for Proteomics, Translational Genomics Research Institute, Phoenix, AZ, [2]Department of Computer Science, University of Toronto, Toronto, ON.

◆ Introduction

- **Proteomics** is "the study of proteins, how they're modified, when and where they're expressed, how they're involved in metabolic pathways and how they interact with one another" – *Mark Wilkins*
- Quantitative proteomics has become increasingly effective in understanding the biology and biomarkers for diseases.
- Issues related to quantitative proteomics[1]:
  - Systematic variations among technical and biological replicate measurements
  - Inference of protein abundances from the observed peptide abundances
  - Undetected peptides leading to "missing values"
  - Statistical comparison of sample groups
- **Inferno** is designed to address these issues featuring:
  - Normalization methods
  - Missing value imputation algorithms
  - Peptide to protein rollup methods
  - Statistical plots
  - Hypothesis testing schemes (unbalanced data, random effects)

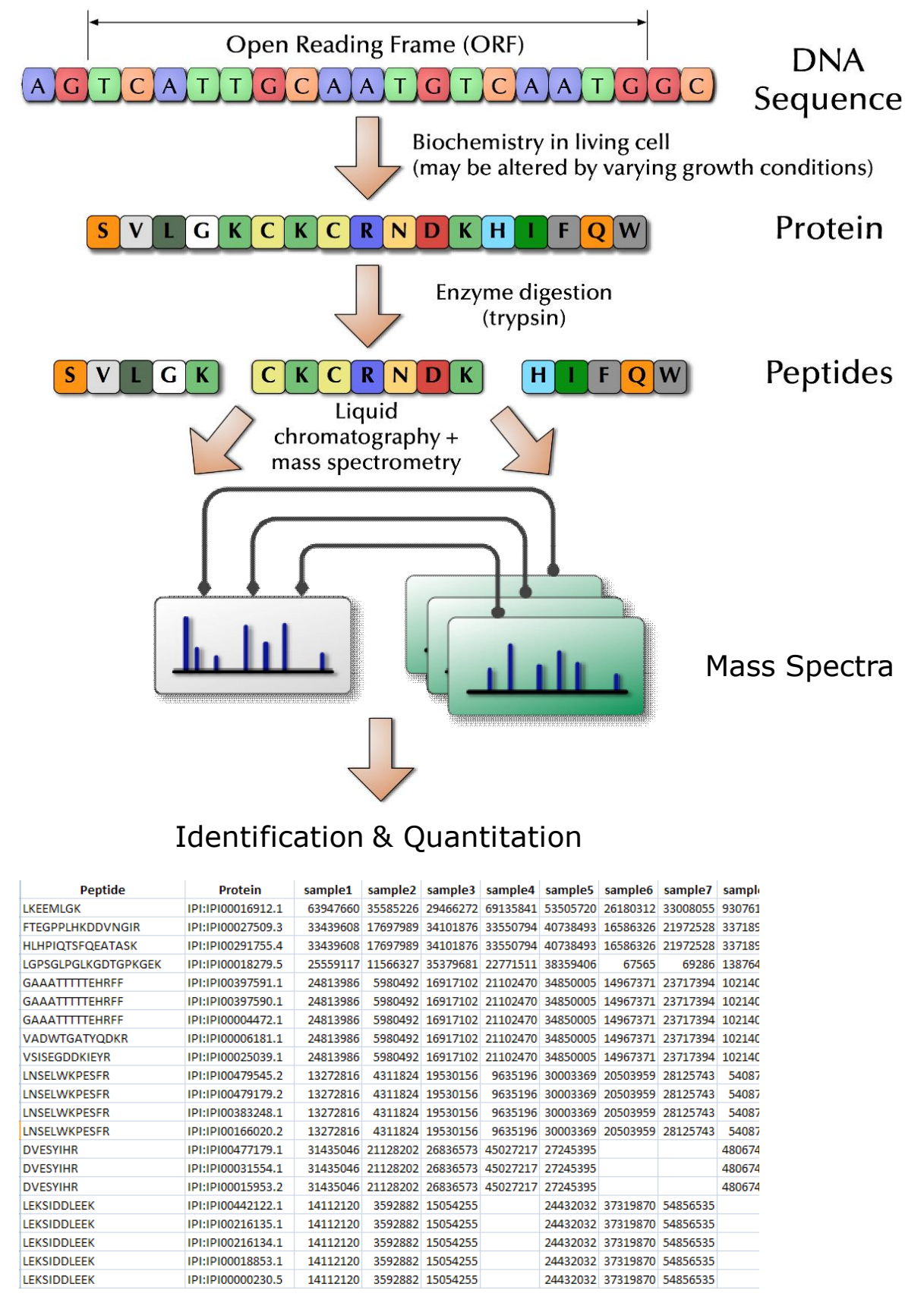


Figure 1. Proteomics Process

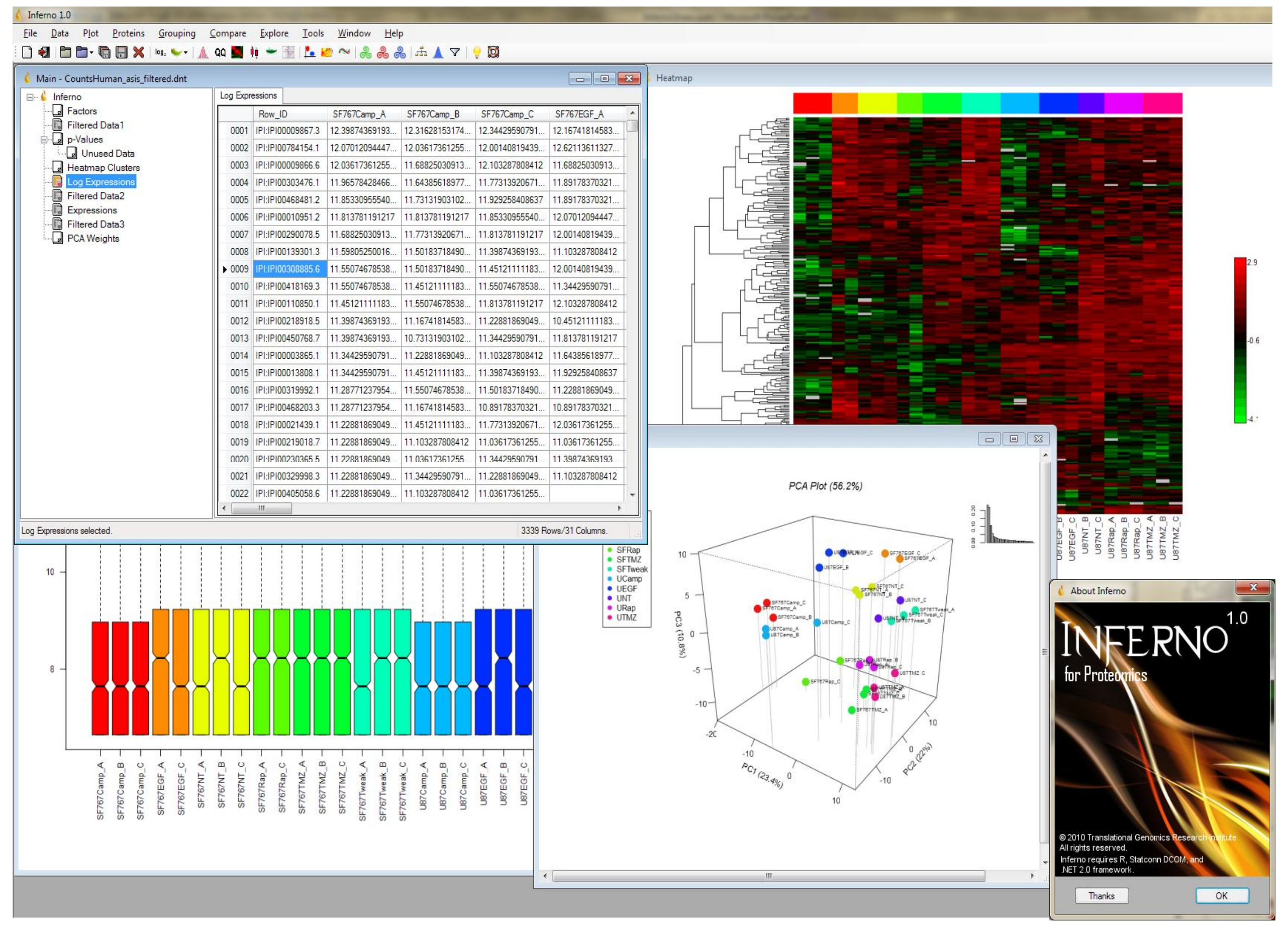
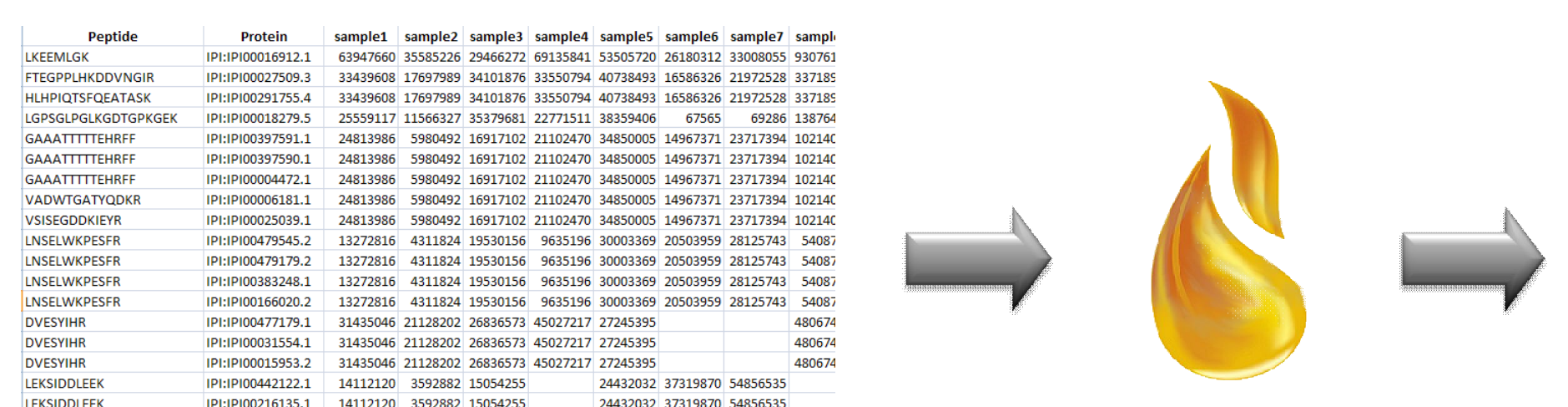


Figure 2. Inferno for Proteomics Software

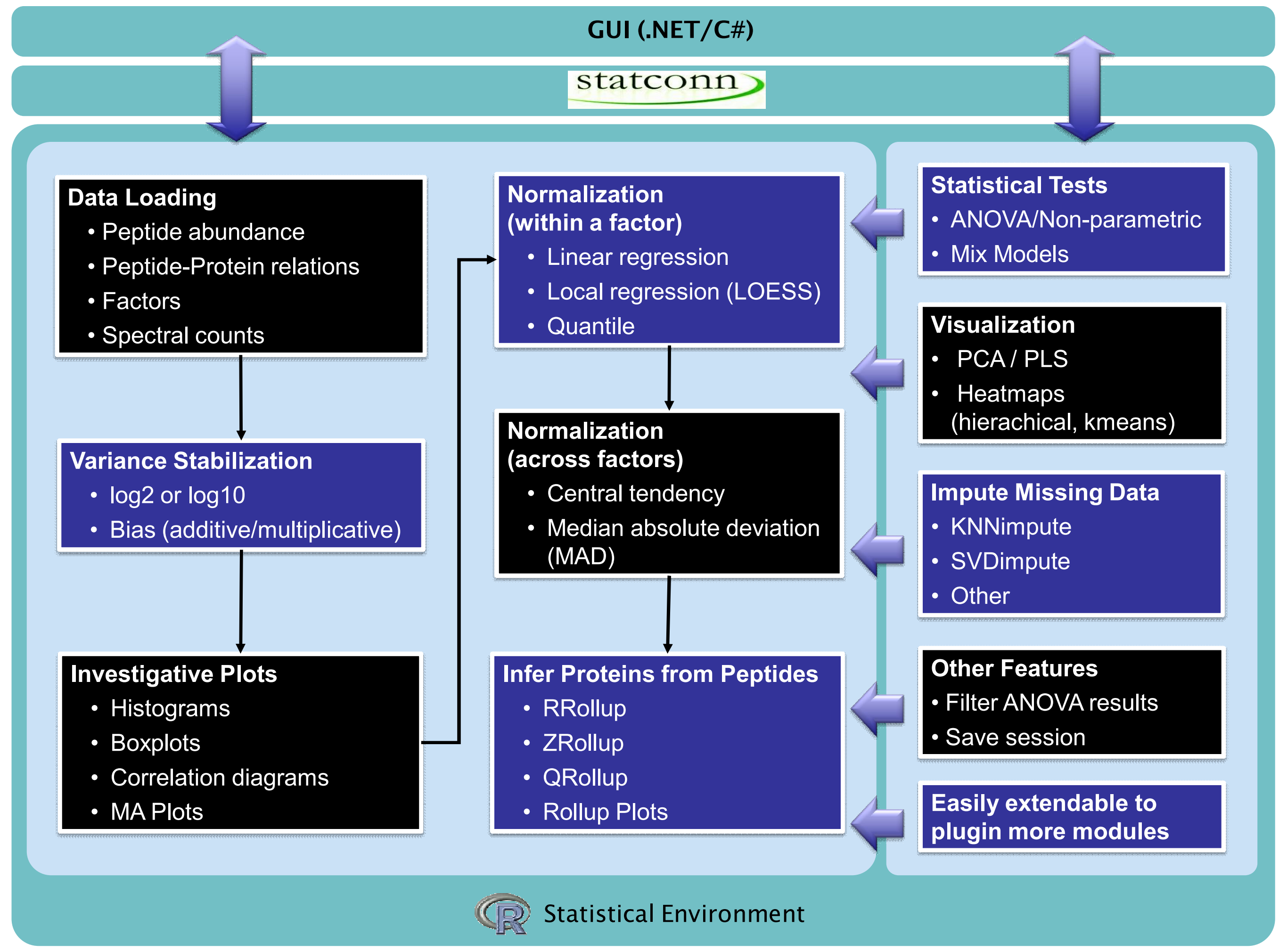
◆ Inferno Design and Features

Multiple samples grouped using factors

- Biological conditions
- Biological replicates
- Technical replicates



Factors capture the experimental design via fixed and random effects. This information is later used in normalization, imputation, and hypothesis testing methods in Inferno.



GUI (.NET/C#)

statconn

Data Loading
- Peptide abundance
- Peptide-Protein relations
- Factors
- Spectral counts

Variance Stabilization
- log2 or log10
- Bias (additive/multiplicative)

Investigative Plots
- Histograms
- Boxplots
- Correlation diagrams
- MA Plots

Normalization (within a factor)
- Linear regression
- Local regression (LOESS)
- Quantile

Normalization (across factors)
- Central tendency
- Median absolute deviation (MAD)

Infer Proteins from Peptides
- RRollup
- ZRollup
- QRollup
- Rollup Plots

Statistical Tests
- ANOVA/Non-parametric
- Mix Models

Visualization
- PCA / PLS
- Heatmaps (hierachical, kmeans)

Impute Missing Data
- KNNimpute
- SVDimpute
- Other

Other Features
- Filter ANOVA results
- Save session

Easily extendable to plugin more modules

Statistical Environment

Figure 3. Inferno Design and Analysis Flow

◆ Normalization Methods[2]

Normalization is done to minimize systematic variation.

- Robust linear regression
- Lowess method
- Quantile method
- Global intensity adjustment using Median Absolute Deviation (MAD)
- Central tendency adjustment

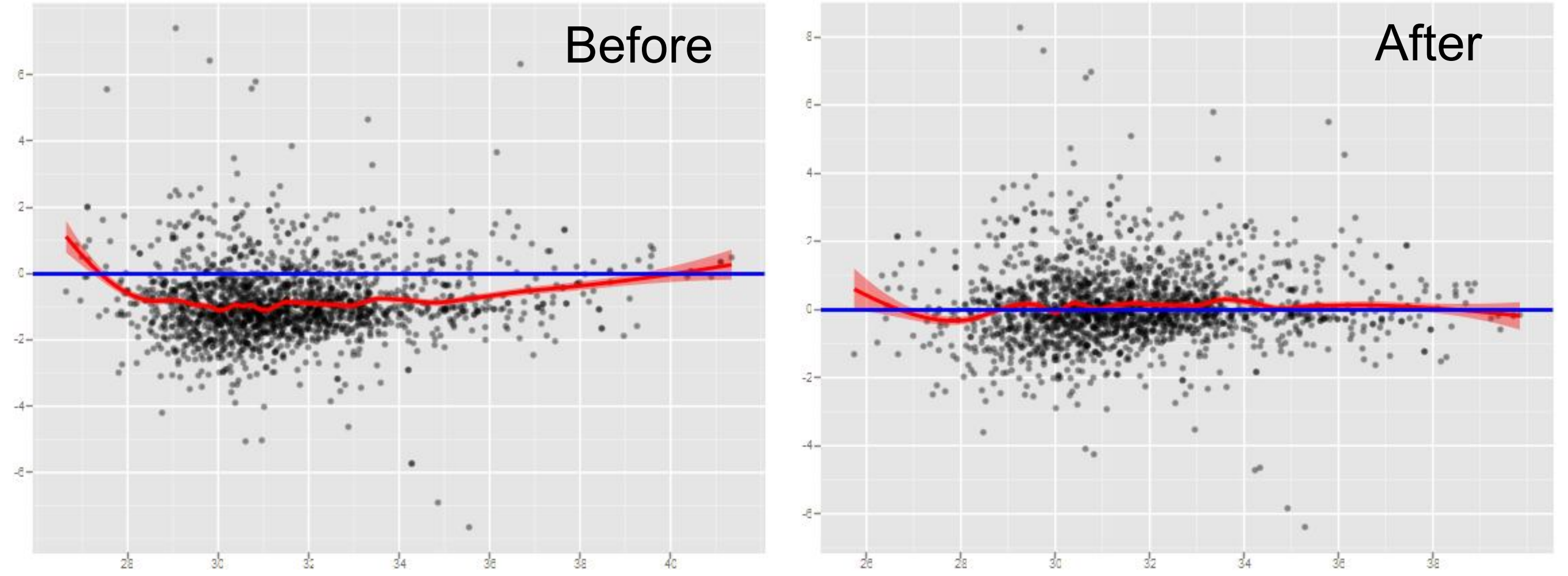


Figure 4. Normalization using LOWESS

◆ Protein Quantitation

- 156 proteins (ANOVA with an FDR of 5%)
- Hierarchical clustering groups related proteins

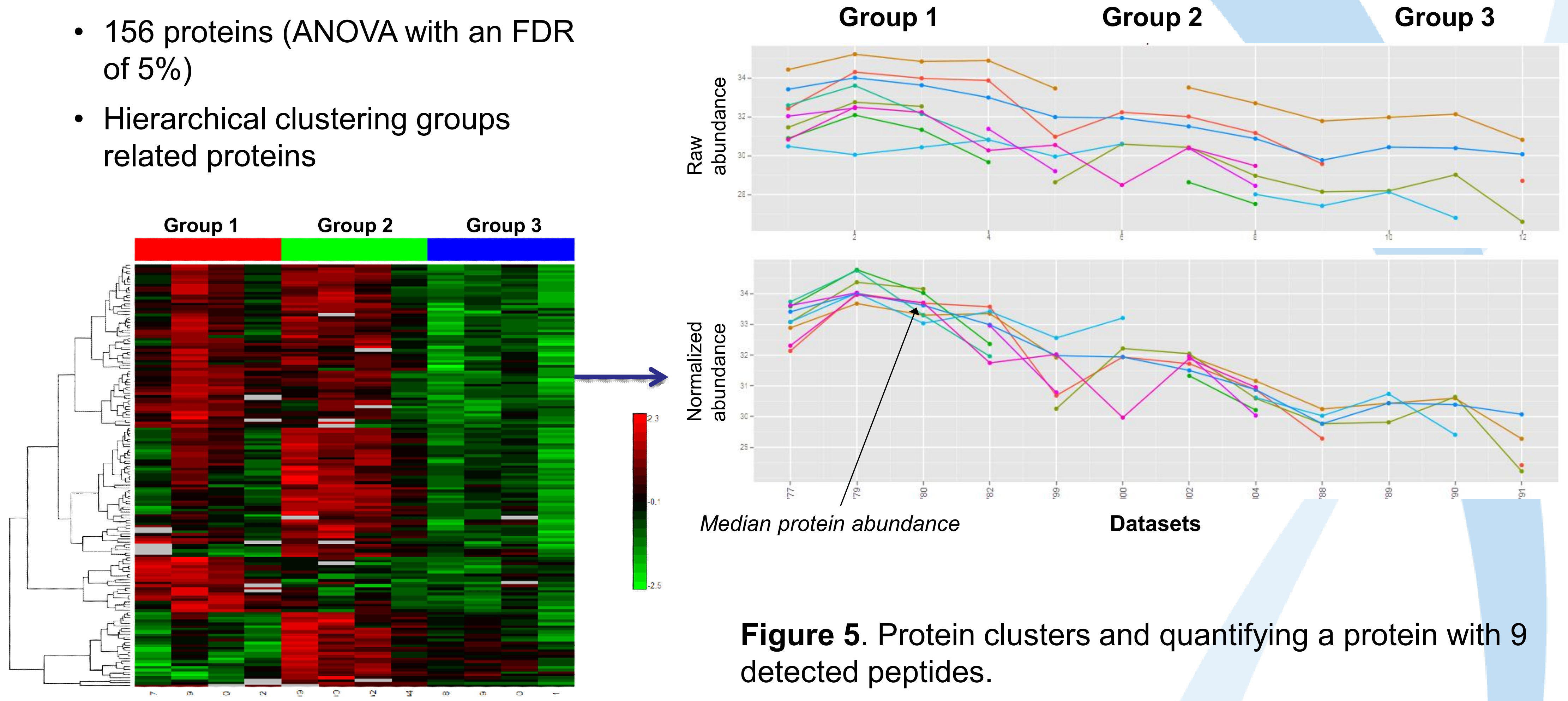
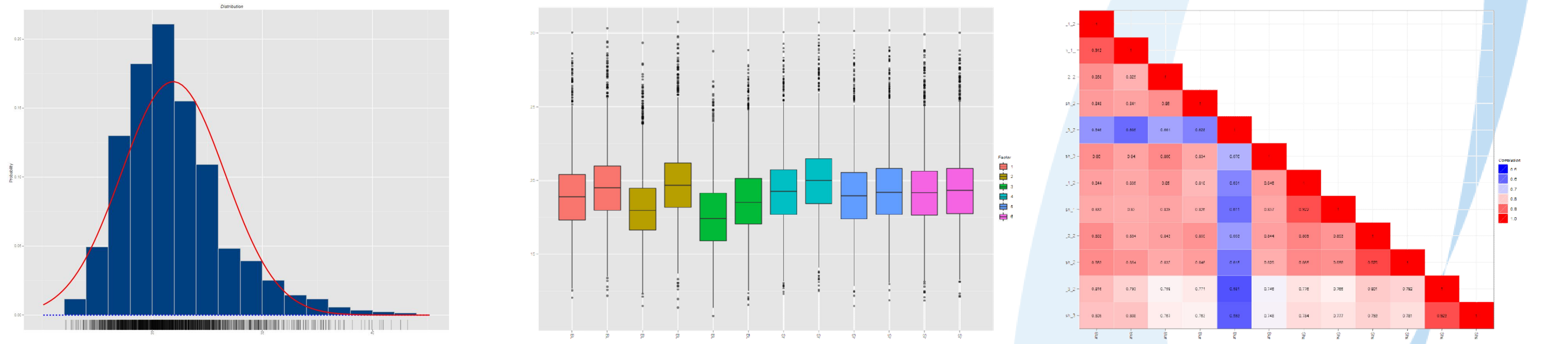
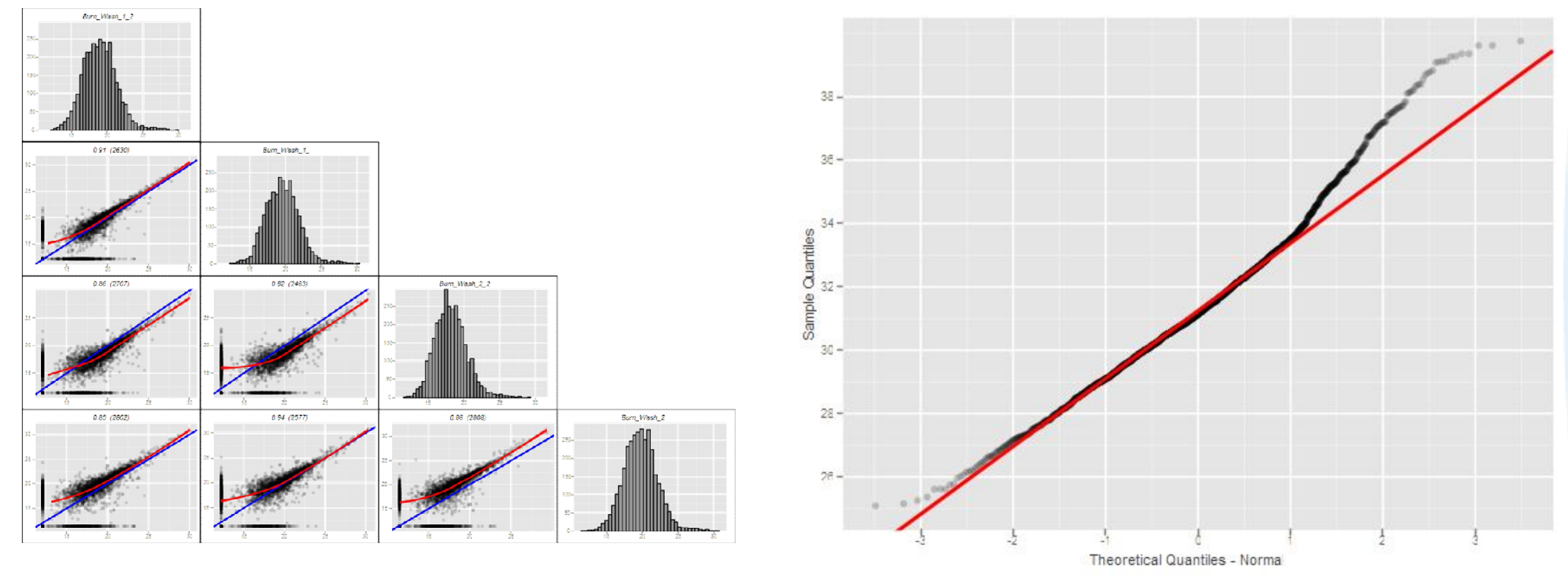


Figure 5. Protein clusters and quantifying a protein with 9 detected peptides.

◆ Statistical Plots





- Histograms
- BoxPlots
- Correlation diagrams
- QQ Plots
- …

◆ References

1. AD Polpitiya, et al. *Bioinformatics* 2008, 24: 1556-1558.
2. SJ Callister, et al. *J Proteome Res* 2006, 5: 277-286.

◆ Acknowledgements

Significant portions of the work were performed at Translational Genomics Research Institute in Phoenix, Arizona with the generous support from Virginia G. Piper Charitable Trust and the Flinn Foundation. Authors would also like to thank the Environmental Molecular Science Laboratory at PNNL in Richland, Washington for early portions of the work.

◆ Contact

Ashoka D. Polpitiya, D.Sc.
Center for Proteomics
Translational Genomics Research Institute
445 N. 5th St., Phoenix, AZ 85004
e-mail: ashoka@tgen.org

Inferno is available at http://inferno4proteomics.googlecode.com