

## Abstract

In 2005, the ABRF Proteomics Standards Research Group (sPRG) created a proteomics standard composed of 49 highly purified human proteins in an equimolar mixture. The sPRG conducted a blind study to assess the proteomics community's ability to determine the identities of the constituent proteins using their proteomics platforms of choice. The results of this study were presented at ABRF 2006 in Long Beach, CA. The study revealed the value of multiple independent analyses of this proteomics standard, and contributors were asked to voluntarily contribute their datasets for future public distribution. Approximately 30 datasets were submitted representing a wide variety of proteomics strategies and mass spectrometry platforms.

The sPRG Bioinformatics Committee (BIC) has analyzed these datasets in detail using a wide variety of informatics tools. Generating a definitive list of protein constituents was our initial goal. We present an assessment of the relative quality of the acquired data sets, independent of how they were analyzed by the submitter. The submitted proteins lists were also re-graded using our revised master protein list. The 2006 ABRF sPRG Proteomics Standard constitutes a valuable dataset for the evaluation of proteome informatics tools.

## Introduction

Questions of interest in examining 2006 sPRG data sets

- Which 'bonus proteins' reported by respondents last year are correct?

Careful protein inference analysis assumed to be key in answering this.

A 'best answer' for the sample would allow these data sets to be used for future informatics tools development, testing, etc.

- Can 'experts' in informatics get consistent results using different analysis tools?
- How similar are existing protein inference tools?
- Can we dissect acquisition and informatics sources of variation in the quality of answers from last year's sPRG study?

## Methods

- Raw data were available for 24 of the 78 labs who responded to the 2006 sPRG study.
- Format conversion utilities were used to create/convert peak lists to .mgf, mzData, and mzXML, allowing a total of 19 of the 24 sets to be searched.
- All BIC members searched the data sets using different software tools for peptide identification and protein inference.
- After initial explorations, a FASTA database was assembled for final searches as the human component of Swiss-Prot plus 39 non-human and 5 non-Swiss-Prot human candidates (15,681 total proteins).
- For final searches, each BIC reported a count of confident peptide sequences and protein probability or binary decision.
- These identification results were manually aligned at the protein level with careful attention to recognizing synonymous or indistinguishable protein accessions. Where multiple

For more information about this study, please visit

[www.abrf.org/sprg](http://www.abrf.org/sprg)



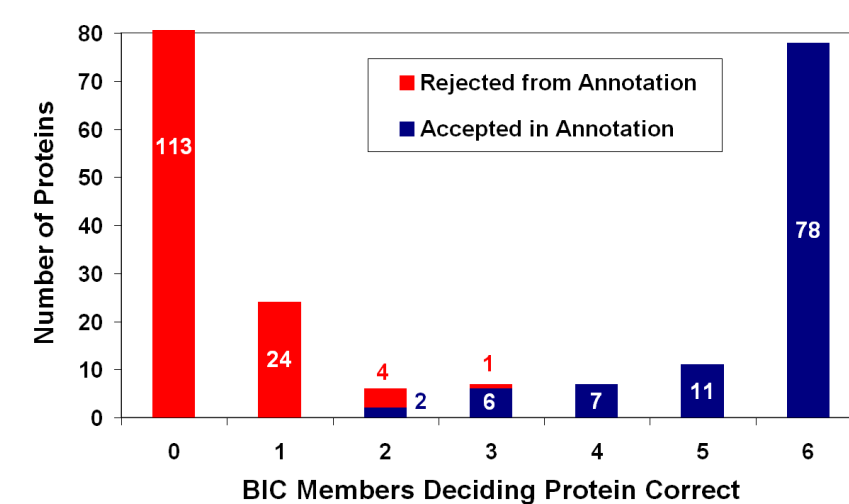
**Table 1. BIC Peptide Measures for Protein Identifications in Individual 2006 sPRG Study Data Sets**

- Columns:** 84 final searches on data sets from 19 labs. Searches are grouped by lab, where each color column corresponds to one member of the BIC, always in the same sequence.
- Rows:** Proteins identified including ambiguity among multiple accessions, grouped by category of protein and roughly sorted by decreasing certainty of identification within the category.
- Cells:** Indicate the peptide count metric reported by the BIC member. These differed in exact meaning among the members, thus they are scaled differently in some cases. The numbers are generally a count of confident distinct peptide sequences in most cases. The cell background formatting classifies the amount of reported peptide evidence into three broad groups:

□ ≥4    ■ 1-3    ■ <1

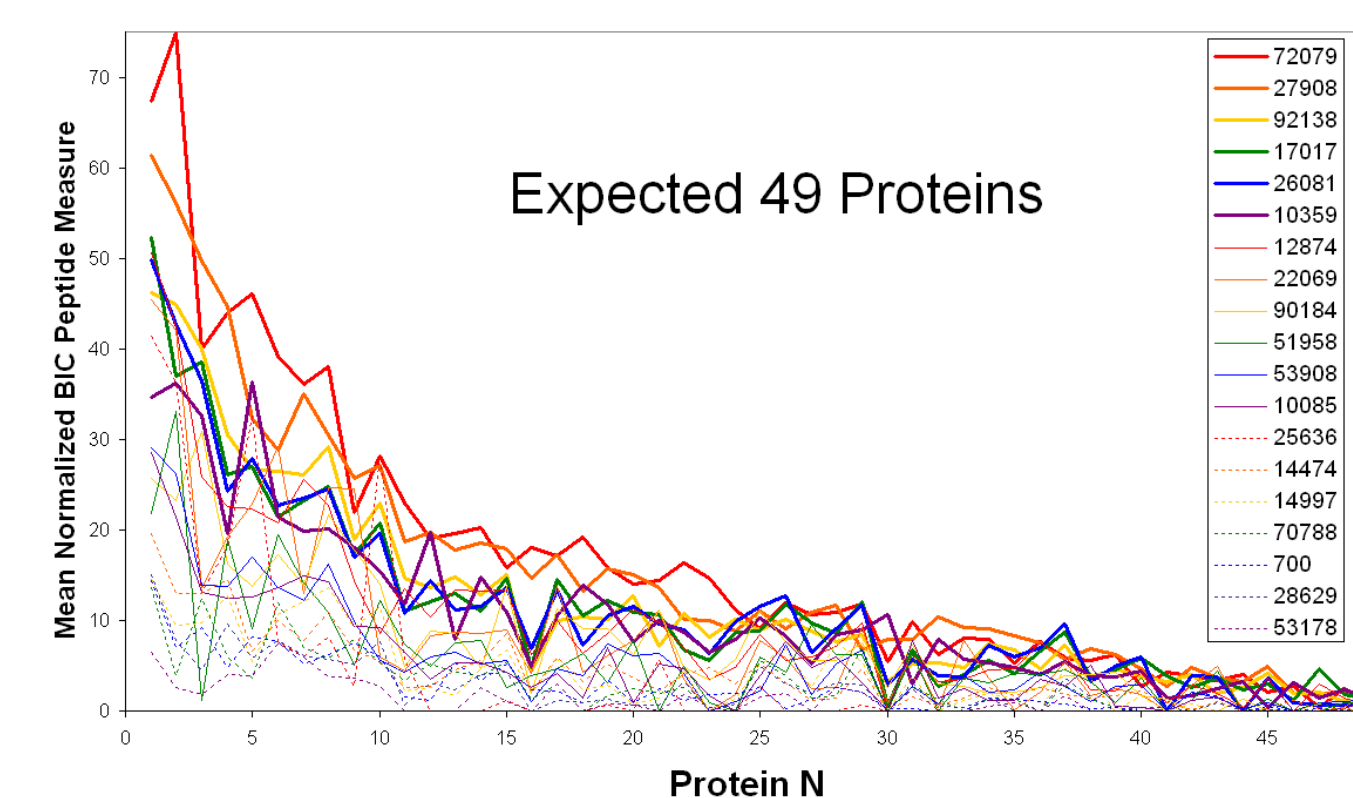
- Secondary isoform detections - Highlighted rows indicate putative detections of multiple isoforms. For example, many searches found evidence of simultaneous detection of the delta form of hemoglobin in addition to the expected beta form (protein N=29 and 29.1).
- Unintended proteins - These stand out very clearly as specific to certain labs. For example, the BIC strongly agreed that sheep proteins are detected in lab 51958 and trypanosome proteins are detected only in lab 28629.

**Figure 1. BIC Consensus on Protein Identifications**



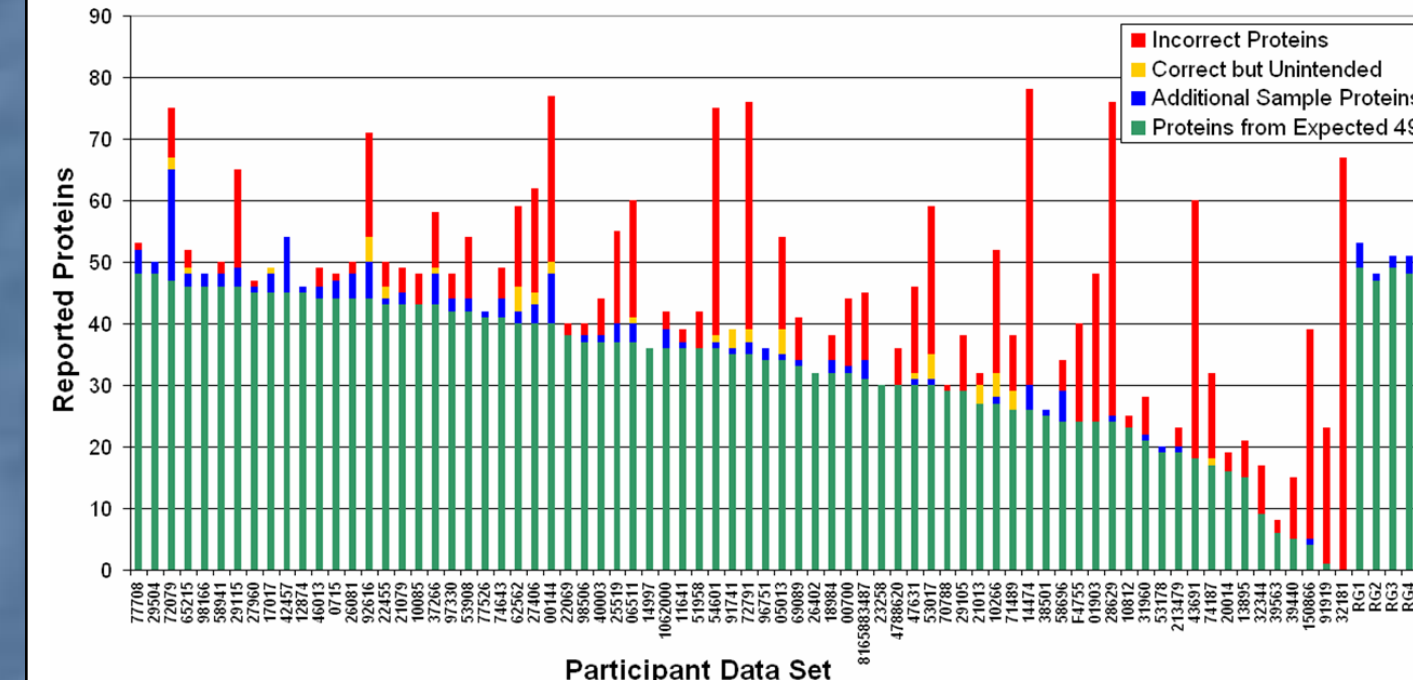
- Figure 1 shows a histogram of the degree of consensus among each BIC members' final decisions about which proteins were and were not detected.
- 78 proteins were declared by all 6 BIC members and 89 were declared by at least 5 or more.
- Only 15 proteins ultimately considered correct had less consensus than this - in some cases because not all BIC member searched the lab where a protein was implicated.

**Figure 2. Relative Quality of Acquired Data**



- Not everyone counted peptides the same way, so the data from each BIC member were normalized to the mean across the group.
- An average of normalized BIC measures for each protein was computed for each lab, shown here as trend lines.
- The relative integrals of these lines are an indication of the relative information content of the data acquired by each lab. The legend is sorted by this integral in descending order.

**Figure 3. Re-Grading of sPRG2006 Protein Lists**



- The picture changes slightly as some of last year's putative bonus proteins are confirmed.

## Conclusions

- A master protein list has been produced for the sPRG2006 study sample.
- Despite using very different tools, experts were able to get surprisingly consistent results.
- A key to recognizing this consistency was careful alignment of reported proteins preserving ambiguity among accession numbers where possible.
- By doing our own informatics analyses, we were able to order the participants' data sets by our assessment of the quality of the acquired data.
- Re-grading submitted protein lists with our master protein list allowed confirmation of additional correct proteins reported by last year's respondents.

## Acknowledgments

The BIC thanks Jayson Falkner for support in using the Tranche repository, the sPRG, and all sPRG2006 study respondents who contributed their raw data, making this BIC study possible.