Supporting Information

HyperQuant- a computational pipeline for higher order multiplexed quantitative proteomics

*Suruchi Aggarwal[1,2,3], Ajay Kumar[1], Shilpa Jamwal[1], Mukul Kumar Midha[1,4], Narayan Chandra Talukdar[2,3], Amit Kumar Yadav[1]**

[1]Translational Health Science and Technology Institute, NCR Biotech Science Cluster, 3rd Milestone, Faridabad-Gurgaon Expressway, Faridabad-121001, Haryana, INDIA

[2] Division of Life Sciences, Institute of Advanced Study in Science and Technology, Vigyan Path, Paschim Boragaon, Garchuk, Guwahati, Assam 781035, India

[3] Department of Molecular Biology and Biotechnology, Cotton University, Panbazar, Guwahati, Assam 781001, India

[4] Current address: Institute for Systems Biology, Seattle, 98109, WA, United States of America

Supplementary Figure Legends

Figure S1: The delayed protein summarization after replicate combination rescues protein identifications (also see supplementary note 2). (A) The Venn diagram represents the increase in number of identifications in replicate combination before protein summarization (combined) vs after summarization (separate). This effectively rescued 537 proteins by delaying protein summarization in combined. The comparison of two methods for Light (B), Medium (C) and Heavy (D) label shows no label bias. We segregated the proteins identified in light medium and heavy separately and found 624, 616 and 601 proteins rescued for individual labels respectively.

Figure S2: The flowchart demonstrates the steps taken to calculate ratios for a single protein expressed in four conditions (hypothetical HOM experiment with two SILAC states and two time-points labeled with iTRAQ reporters 113 and 114) as a representative example. All the steps represented here are performed and calculated for all proteins to get their corresponding ratios.

Figure S3: The ratio calculation method for the BONPlex study to depict the use of SILAC and iTRAQ labels for quantitative dimensions – iTRAQ ratios for temporal changes and SILAC ratios for strain specific changes as shown.

Figure S4: Comparison of NSS proteins. (A) Set 1 control against H37Ra and H37Rv, (B) Set 2 control against BND433 and JAL2287, and (C) Controls combined (from both set 1 and 2) against Avirulent (H37Ra) and combined Virulent (H37Rv, BND433 and JAL2287) strains.

Figure S5: Heat-maps depicting the strain-specific expression levels of NSS proteins in various infection conditions as compared to uninfected control in the respective time window. (A) NSS proteins in Ra and Rv infection (49 proteins), (B) NSS proteins in BND and JAL infection (34 proteins).

Figure S6: Heat-maps depicting the temporal expression levels of NSS proteins in various infection conditions but absent in control. The iTRAQ ratios are calculated based on first time-point. (A) Ra and Rv specific NSS proteins (52 proteins). (B) BND and JAL specific NSS proteins (40 proteins).

Figure S7: Temporal Expression of 16 NSS proteins common to all infections. Their levels are depicted in infections by- Ra (A), Rv (B), BND (C), JAL (D). The proteins outside grey zone are considered significantly under or overexpressed. The x-axis represents the time-points in which the proteins were quantitated and the y-axis represents the log2 fold change of the proteins as compared to uninfected control. The color represents the UniProt id of the proteins identified.

Figure S8: The overview of NSS proteins involved in various immune response pathways through Reactome analysis. The NSS proteins involved in major pathways is represented as heat maps with two types of information. Each row in a heat map represents one protein with 18 measured values. The first six represent iTRAQ ratios for uninfected controls with respect to the first time point. However, the next 12 represent the expression for a strain at a given time point with respect to the corresponding uninfected control. The base image is taken from reactome analysis

(https://reactome.org). The table in the centre shows the number of proteins involved in the respective immune pathways for the infected macrophages (but not control).

Supplementary Table Legends

Table S1: The table of keywords representing the number of proteins found for each keyword. Details of protein names can be found in Supplementary Table S3 (sheet- named keywords)

Table S2: Tables containing Set1 and Set2 protein quantitation values as obtained in the BONPlex experiment (provided as "**Supplementary table S2.xlsx"**)

Table S3: Tables containing the keyword search results, pathways, GO annotations of Set1 & Set2, and t-test results from Set1 & Set2 (provided as "**Supplementary table S3.xlsx"**).

Supplementary Notes

Supplementary Note 1: Replicate Combination

Supplementary Note 2: Outlier Removal

Supplementary Note 3: Protein summarization for quantitation

Supplementary Note 4: Statistical analysis BONPlex data

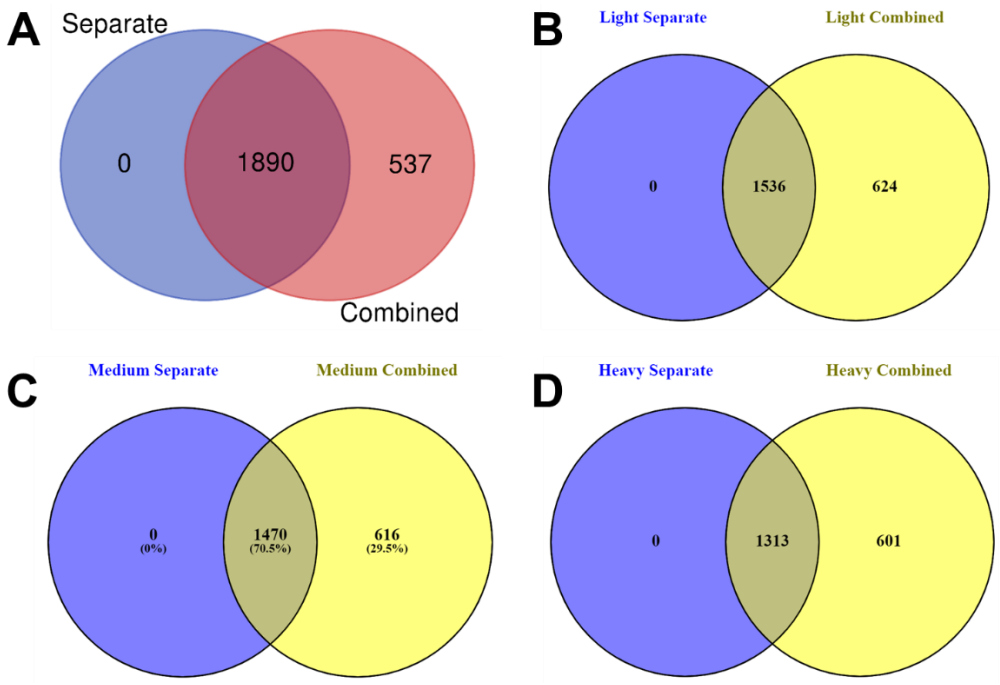

Figure S1: The delayed protein summarization after replicate combination rescues protein identifications (also see supplementary note 2). (A) The Venn diagram represents the increase in number of identifications in replicate combination before protein summarization (combined) vs after summarization (separate). This effectively rescued 537 proteins by delaying protein summarization in combined. The comparison of two methods for Light (B), Medium (C) and Heavy (D) label shows no label bias. We segregated the proteins identified in light medium and heavy separately and found 624, 616 and 601 proteins rescued for individual labels respectively.

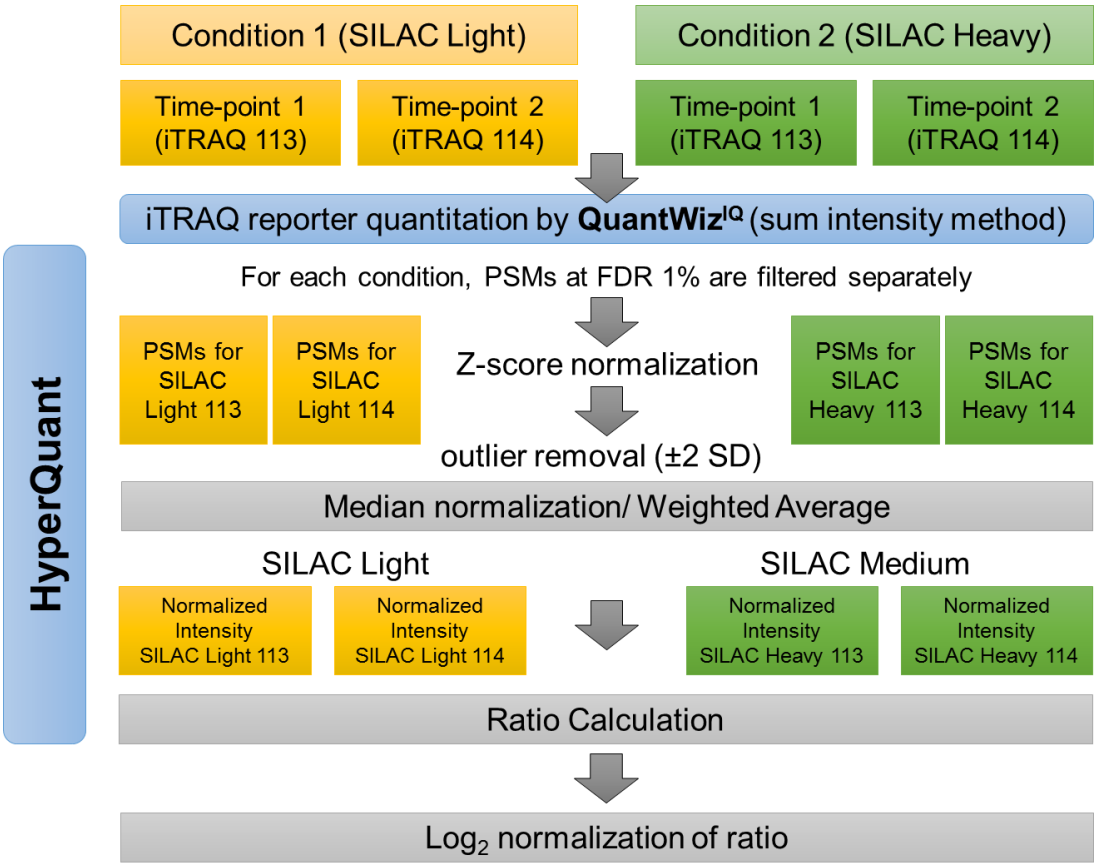

Figure S2: The flowchart demonstrates the steps taken to calculate ratios for a single protein expressed in four conditions (hypothetical HOM experiment with two SILAC states and two time-points labeled with iTRAQ reporters 113 and 114) as a representative example. All the steps represented here are performed and calculated for all proteins to get their corresponding ratios.

	SILAC(Light)						SILAC(Medium)						SILAC(Heavy)					
	Uninfected						Avirulent (H37Ra)						Virulent (H37Rv)					
iTRAQ	U1	U2	U3	U4	U5	U6	A1	A2	A3	A4	A5	A6	V1	V2	V3	V4	V5	V6
	6h	10h	14h	18h	22h	26h	6h	10h	14h	18h	22h	26h	6h	10h	14h	18h	22h	26h
	113	114	115	116	117	118	113	114	115	116	117	118	113	114	115	116	117	118

Temporal changes

Strain-specific changes

114/113 114/113 114/113 113/113 113/113

Figure S3: The ratio calculation method for the BONPlex study to depict the use of SILAC and iTRAQ labels for quantitative dimensions – iTRAQ ratios for temporal changes and SILAC ratios for strain specific changes as shown.

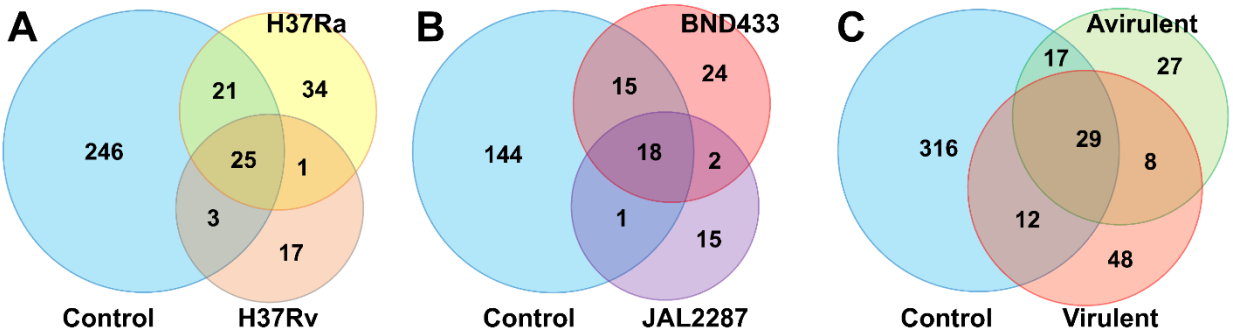

Figure S4: Comparison of NSS proteins. (A) Set 1 control against H37Ra and H37Rv, (B) Set 2 control against BND433 and JAL2287, and (C) Controls combined (from both set 1 and 2) against Avirulent (H37Ra) and combined Virulent (H37Rv, BND433 and JAL2287) strains.

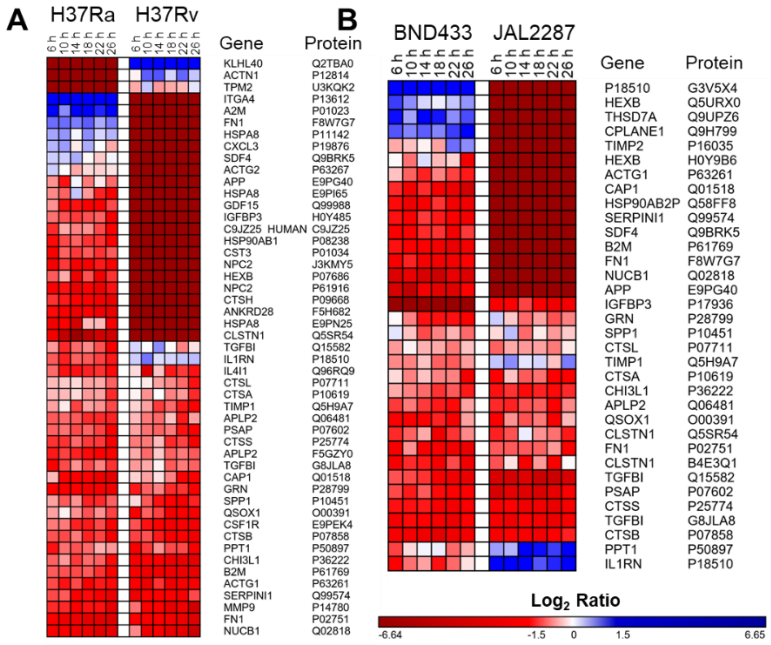

Figure S5: Heat-maps depicting the strain-specific expression levels of NSS proteins in various infection conditions as compared to uninfected control in the respective time window. (A) NSS proteins in Ra and Rv infection (49 proteins), (B) NSS proteins in BND and JAL infection (34 proteins).

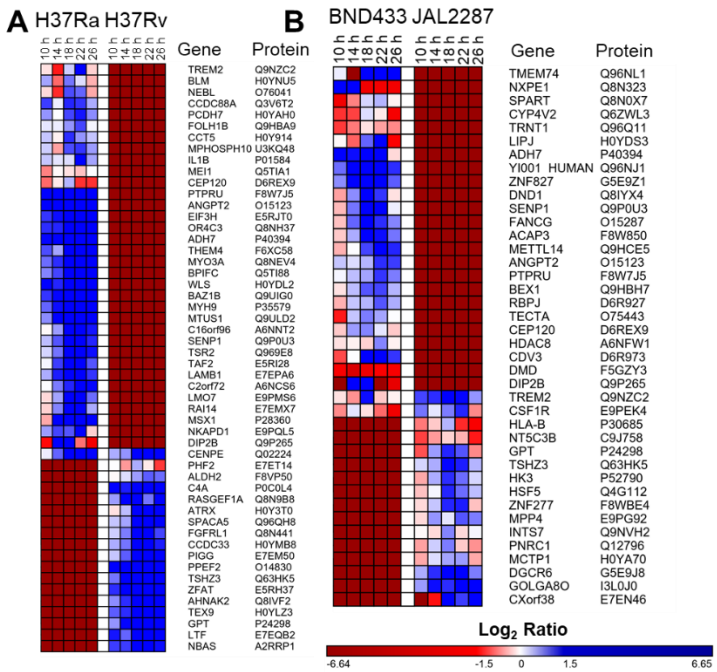

Figure S6: Heat-maps depicting the temporal expression levels of NSS proteins in various infection conditions but absent in control. The iTRAQ ratios are calculated based on first time-point. (A) Ra and Rv specific NSS proteins (52 proteins). (B) BND and JAL specific NSS proteins (40 proteins).

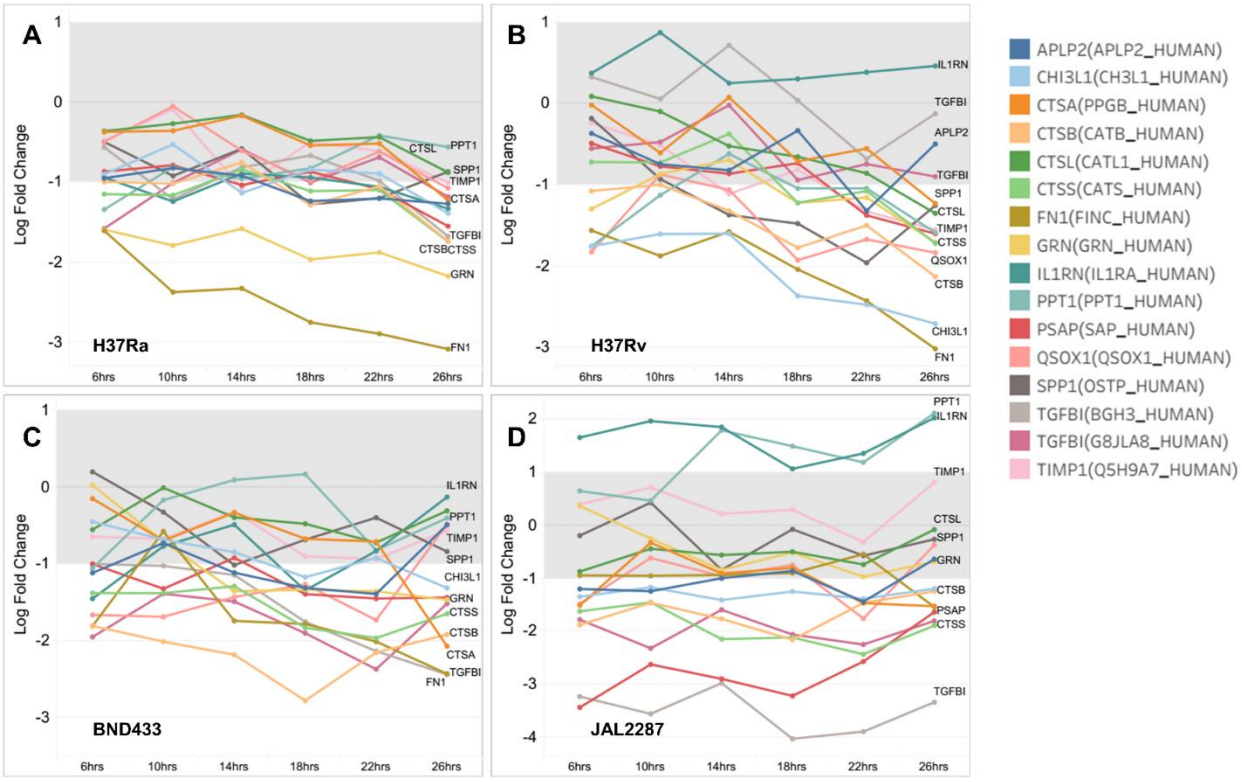

Figure S7 : Temporal Expression of 16 NSS proteins common to all infections. Their levels are depicted in infections by- Ra (A), Rv (B), BND (C), JAL (D). The proteins outside grey zone are considered significantly under or overexpressed. The x-axis represents the time-points in which the proteins were quantitated and the y-axis represents the log_2 fold change of the proteins as compared to uninfected control. The color represents the UniProt id of the proteins identified.

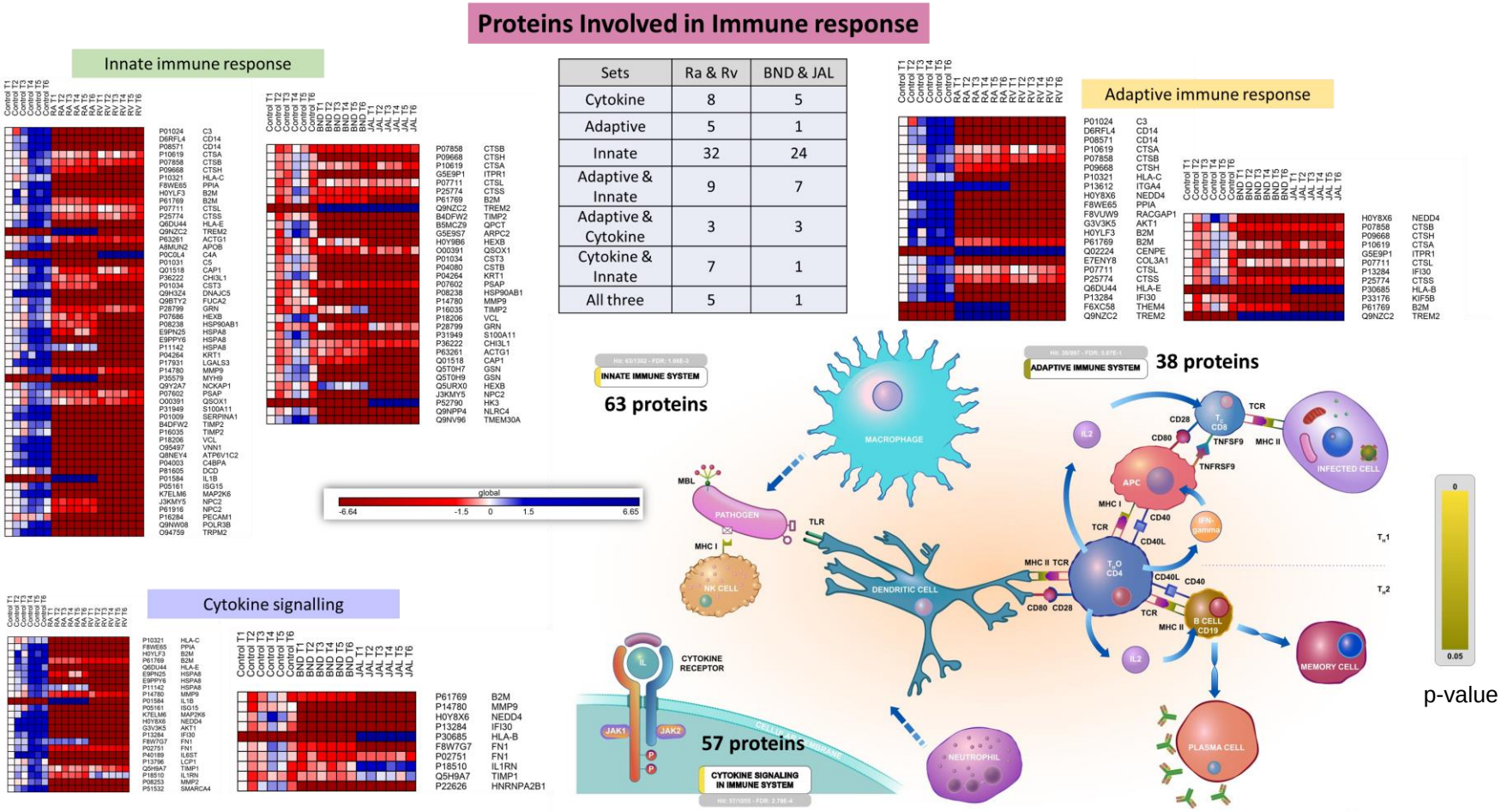

Figure S8: The overview of NSS proteins involved in various immune response pathways through Reactome analysis. The NSS proteins involved in major pathways is represented as heat maps with two types of information. Each row in a heat map represents one protein with 18 measured values. The first six represent iTRAQ ratios for uninfected controls with respect to the first time point. However, the next 12 represent the expression for a strain at a given time point with respect to the corresponding uninfected control. The cartoon image (bottom right panel[1], https://reactome.org/content/detail/R-HSA-168256) represents the proteins mapped to immune pathways by reactome[2]. The significantly enriched proteins in the three immune categories- innate immune system, adaptive immune system and cytokine signaling in immune system are shown, with their statistical values. The Reactome table in the centre shows the number of proteins involved in the respective immune pathways for the infected macrophages (but not control).

Table S1: The table of keywords representing the number of proteins found for each keyword. Details of protein names can be found in Supplementary Table 3 (sheet- named keywords)

Keywords searched	Total Proteins found	Control 1	Ra	Rv	Control 2	BND	JAL
antibacterial humoral response	6	4	1	2	2	1	0
caspases	3	1	0	0	2	0	0
cell proliferation	78	52	18	8	28	11	6
chemokine	13	9	4	1	5	2	2
collagen	38	30	13	8	18	9	7
cytokine	43	38	14	10	14	9	6
defense response	21	11	5	5	10	4	2
Extracellular matrix	65	51	17	12	33	14	10
Growth Factors	28	19	5	7	12	2	3
Immune response	55	38	12	10	22	9	7
Inflammatory Response	48	35	11	8	19	8	8
Laminin	11	10	3	2	2	1	1
Macrophage	32	26	12	7	11	7	6
Matrix metalloproteinase	6	6	2	2	3	2	1
Proteases	14	13	6	4	9	4	3
Regulation of cytokine production	21	17	8	8	8	6	4
Secreted	109	79	30	22	48	21	14
Secretory	50	34	10	9	21	6	5
Signal	275	179	62	36	108	41	24
Signalling	23	18	5	2	9	3	2
Tissue Remodelling	29	19	4	3	11	2	1

Supplementary Note 1: Replicate Combination

Apart from mapping ID and Quant results, the HyperQuant tool also allows for semantic protein replicate combination. For replicate combinations, the PSMs that pass the filter criteria of ≤1%FDR are all collated together as one pool for every protein. Subsequently, for each protein, the assigned PSMs are then segregated into the distinct label combinations. For example for a triple SILAC (Light, Medium, Heavy as L/M/H) with 8-plex iTRAQ (reporters of 113,114,115,116,117,118,119 and 121), the combinations are as follows-

L_{113}, L_{114}, L_{115}, L_{116}, L_{117}, L_{118}, L_{119} and L_{121}

M_{113}, M_{114}, M_{115}, M_{116}, M_{117}, M_{118}, M_{119} and M_{121}

H_{113}, H_{114}, H_{115}, H_{116}, H_{117}, H_{118}, H_{119} and H_{121}

Where the first term describes the SILAC label and second term (subscript) describes the iTRAQ reporter used. For every such label combination ($3 \times 8 = 24$ combinations shown in this hypothetical example experiment), all PSMs are taken together and outliers are removed (see Supplementary Note 2). Further, the protein quantitation is summarized (see Supplementary Note 3) based on area or ratio as chosen by the user for each label separately.

After protein summarization, the user gets a single value representing the central tendency of quantitation value from all the replicates. There are two ways a protein value can be summarized- either summarizing individual replicates before combination, or after replicate combination. If proteins are summarized in individual replicates first, the *one-hit-wonder removal* method will reduce the number of proteins. Only the proteins identified with two or more peptides will make it to the final output list. On the other hand, combining spectra level evidence for proteins from all replicates at the same time can be corroborative evidence for their presence and can make the consistently identified proteins (even with low peptide/spectral counts) more trustworthy. Such proteins can be rescued by delaying the protein summarization step until all replicate level information is combined.

To show its effect, we conducted a simple exercise to compare how many proteins are identified if we summarize proteins from all replicates versus individual replicates. Using 18plex data from Dephoure et al[3], we processed the 20 fractions as replicates once separately, and secondly combined as per our delayed summarization method. We observed that for all labels combined 537 proteins were rescued, if combined strategy was used (supplementary figure 1A). To check if these were affected only by light labels, we segregated the proteins identified in light medium and heavy separately and found 624, 616 and 601 proteins rescued for individual labels respectively (supplementary figure 1B, C and D). The similar number of rescued proteins denotes that there was no specific label bias towards abundant light labels.

Supplementary Note 2: Outlier Removal

The quantitation values for each label combination are processed separately (Supplementary Figure S2). For example, the intensity values for all the PSMs with 113-iTRAQ reporter of SILAC light are used to calculate the standard deviation (SD), and the values outside ±2 SD range are discarded as outliers. This step ensures that extreme values will be removed. The details of outlier removal in an iterative manner, are as follows-

1) Removal of "NA" values where there were no areas or ratios measured.
2) Removal of "#BT" (below threshold values) if "Area" instead of ratio is chosen for reporting. The default value is an arbitrary cutoff of 20 or user defined value.
3) Removal of extreme values (provided as 100 or 0.01) for PSMs ratios as these denote extremes of abundance ratio. Such values may skew the final calculations and are thus removed prior to statistical filtering.
4) For the remaining values, we assume proper measurements have been made and are treated as repeated observations of the quantitation of the particular protein. Now, we calculate a z-score for every value as follows –

$$zscore = \frac{x_i - \overline{x}}{SD}$$

Where *x* is the measured value, *zscore* is the standard score, $\overline{x}$ is the mean and *SD* is the standard deviation.

The *zscore* cutoff values beyond ±2 SD have been traditionally used for outlier removal as an easy, efficient and fast method for discarding unreliable values. We also use the same cutoff of ±2 SD for removing such outliers. Values that remain are used for further calculations.

NB: Please note that the outlier removal method DOES NOT apply to ratios when weighted average method is used.

Supplementary Note 3: Protein summarization for quantitation

There are three protein summarization methods available in HyperQuant, and one of these following can be chosen by the user –

1) Average method:

This method will summarize the quantitation values based on a simple average of the available values. This is applicable to ratios as well as normalized intensities (or area).

2) Median method:

This method will summarize the quantitation values based on median value from the available values. This is also applicable to ratios as well as normalized intensities (or area).

3) Weighted average method:

This method takes the peptide intensities as weights to calculate the average ratio. This is NOT applicable to normalized intensities (or area). The ratios are calculated as –

$$Protein\ Ratio = \frac{\sum_{i=1}^{N} peptide\ ratio_i \times peptide\ weight_i}{\sum_{i=1}^{N} peptide\ weight_i}$$

Where N is the number of peptides, peptide weights are -

$$peptide\ weight = \frac{1}{\%\ error\ factor}$$

Where % error factor is calculated as:

$$\%\ error\ factor = \sqrt{\left(\frac{reporter\ intensityB}{ErrorB}\right)^2 + \left(\frac{reporter\ intensityA}{ErrorA}\right)^2}$$

Where reporter intensityB and reporter intensityA represent reporter intensities of iTRAQ labels as calculated by QuantWiz[IQ].

ErrorB and ErrorA here simply represent the errors for the corresponding iTRAQ labels calculated as:

$$Error = \sqrt{reporter\ intensity}$$

Supplementary Note 4: Statistical analysis BONPlex data

The last three time-points of untreated vs infected cells were used for a simple t-test based statistical analysis to test if this HOM experiment has value even in the absence of biological replicates. The value of t-test will also help in statistical assessment of upregulated or downregulated proteins found using the ±1log FC. Although most of the secretome was shut down in infection conditions, the proteins only quantitated in uninfected cells or infected cells were automatically considered to be of significance. The rest of the proteins, which had, values in uninfected or anyone infected (46 for Ra, 28 for Rv, 33 for BND and 19 for JAL) were used to calculate t-test (supplementary table 3, sheets named “Set1 ttest” and “Set2 ttest”). Although this is just a proxy for statistical test considering the last three time-points as biological replicates, this data supports the previous findings using a simple fold-change cut-off. While we do not advocate study designs without replicates, this study was aimed solely at demonstrating HOM capability using the HyperQuant tool. Using this little statistical exercise, we highlight that a truly HOM experiment can still find meaningful results.

References

(1) de Bono, B., Gillespie, ME, Luo, F, Ouwehand, W.H. Image for "Immune System (Homo sapiens)". In *Reactome, release 58*; R-HSA-168256, Ed.; Reactome, release 66: Reactome 2006-03-30.

(2) Joshi-Tope, G.; Gillespie, M.; Vastrik, I.; D'Eustachio, P.; Schmidt, E.; de Bono, B.; Jassal, B.; Gopinath, G. R.; Wu, G. R.; Matthews, L.; Lewis, S.; Birney, E.; Stein, L. Reactome: a knowledgebase of biological pathways. *Nucleic acids research* **2005**, *33*, D428-432.

(3) Dephoure, N.; Gygi, S. P. Hyperplexing: a method for higher-order multiplexed quantitative proteomics provides a map of the dynamic response to rapamycin in yeast. *Science signaling* **2012**, *5*, rs2.