Final report of the dataset PXD045844

Tamara González, Linda Id, Tyko Runeberg

Dataset selection

The dataset was selected from PRIDE database using FileZilla to identify suitable datasets. Our criteria for dataset selection included label-free quantification (LFQ) and the presence of .mzID files. To locate such datasets, the following regular expression was used in FileZilla while browsing ftp://ftp.pride.ebi.ac.uk/pride/data/archive/2024: (?i)(?=.*mzid)(?=.*LFQ)

Background of the study

Prevalence of neurological and psychiatric disorders are shown to differ between sexes. Women show a higher prevalence of Alzheimer's disease, post-traumatic stress disorder, and mood and anxiety disorders, whereas men are more frequently diagnosed with schizophrenia and attention deficit hyperactivity disorder (ADHD). These disparities point to underlying biological differences in brain function. The locus coeruleus (LC), a key noradrenergic nucleus, exhibits structural and gene expression differences between sexes and may play a central role in mediating these effects. For example, animal studies have reported that females have a larger LC volume, a greater number of neurons, and longer, more complex dendritic structures compared to males. Despite recent advances in identifying sex differences, the intrinsic functional and molecular characteristics of male and female LC neurons remain poorly understood. Understanding these sex-dependent differences in noradrenergic neuron function and composition could contribute to the development of more effective, sex-specific treatments for both psychiatric and neurodegenerative disorders.

Methods

To investigate the properties of LC noradrenergic neurons and their sex-specific differences, this study conducted a single-cell proteomic analysis of individual LC neuron somas. The neuron somas were obtained from a genetically engineered mouse model, which enabled LC neuron identification through fluorescence microscopy. In total, a subset of four neurons, one from each mouse representing both sexes, was selected for single-cell proteomics. Digestion of the proteomics samples was done by using trypsin and Lys-C. Chromatographic separation of the peptides was done with DNV PepMap Neo column and samples were analysed on a LC-MS/MS system consisting of Orbitrap Eclipse Mass Spectrometer and Vanquish Neo nano-UPLC system. Data were collected in top speed data-dependent mode.

Results

The study compares electrophysiological properties of LC neurons between sexes. Key findings include higher membrane capacitance in female neurons, which may indicate increased cellular maturity. In contrast, male LC neurons exhibited more frequent spontaneous action potential firing, suggesting differences in intrinsic excitability.

In terms of proteomic profiling, a total of 728 proteins were identified by matching MS/MS spectra to the UniProt mouse protein FASTA database. Among these, 24 proteins were uniquely expressed in female samples, whereas 159 were unique to males. To ensure confidence in the results, 415 proteins meeting high-confidence criteria, defined as passing the Protein FDR Validator node and having at least two peptide spectral matches (PSMs), were included in downstream quantitative analysis. This analysis revealed that transcription regulators, enzymes, disease-related proteins, and signalling pathway components were generally expressed at lower levels in female LC neurons compared to males. Statistical significance was determined using a p-value threshold of 0.05 and a fold change (FC) cutoff of 1.0.

OUR RESULTS

Dataset acquisition and quality assessment

During dataset processing, we observed that no decoy hits remained in the .mzID file, indicating data had been filtered already. Additionally, 99% of peptide spectral matches (PSMs) were of rank 1, suggesting high-confidence identifications. From the .mzID file, we identified 2,030 peptides and 728 proteins, which is consistent with the numbers reported in the original publication. The total number of MS/MS spectral scans identified in our analysis was 8,296 (to be confirmed).

	.mzID file from paper
Number of decoy hits	0
Score distribution	NA
PSM rank	Ranks: 1) 9180, 2) 60, 3) 1
Identified peptides	2030
Identified proteins	728
Razor proteins	606
Protein groups	650

Table 1. Summary of .mzID file

QFeature aggregation

MaxQuant processed the eight raw files in parallel, matching features across them and outputting a combined quantification and identification table in a single evidence.txt file. To generate Q features, we transform this file into a wide format. During this process, we observed duplicate intensity values for the same sample and peptide; for each duplicate, we retain only the entry with the maximum intensity.
We applied two filters to the data:

1. Contaminants and decoys were removed to align with the protein FDR validation strategy used in the referenced paper.
2. We retained only peptides supported by at least two PSMs in at least one sample. Note this was applied only for the strictly filtered dataset, as described in the paper.

The dataset contained many missing (NA) and zero values. To create a minimal and tidy dataset suitable for downstream analysis, we summarized the maximum intensity per peptide per sample instead of using mean or median values. Each peptide sequence was directly linked to its leading razor protein.
The resulting Q feature matrix contains peptides (rows) by sample identities (columns), and the associated metadata includes file names and corresponding gender information. This structure supports accurate peptide-to-protein mapping for subsequent analyses.

Unique and shared proteins by groups

In the original study, 752 proteins were identified initially. In comparison, our dataset yielded a maximum of 354 proteins, which is less than half the original count. For a meaningful comparison, we proceeded with a strictly filtered dataset, retaining 178 proteins, while the paper reported 415 after similar filtration.
Despite the difference in absolute numbers, the relative patterns between male and female samples were consistent with the original findings. Notably, our results also show that male samples exhibit more uniquely expressed proteins than female samples (Figure 1; Table 2).

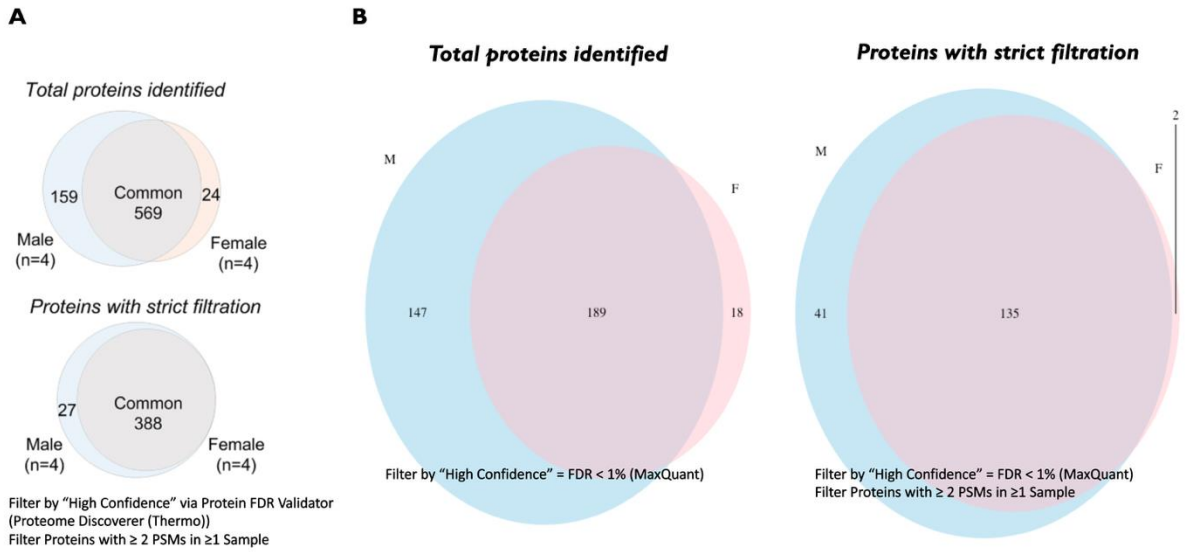

Figure 1. A. Original study results. B. Our Venn diagrams showing protein identification before (354 proteins, left) and after (178 proteins, right) strict filtration.

	Original results		**Our results: evidence.txt file (from MaxQuant)**	
Sex	Identified proteins	Filtering	Identified proteins	Filtering
F	593	Loose	207	Loose
M	728	Loose	336	Loose
F+M	752	Loose	354	Loose
F	388	Strict	137	Strict
M	415	Strict	176	Strict
F+M	415	Strict	178	Strict

Table 2. Summary of proteins identified per group

Normalization and imputation

According to the original paper, raw peptide abundances were normalized against the total peptide abundance in each sample. However, MaxLFQ (used by MaxQuant) already performs its own normalization and intensity estimation at the protein level, providing normalized relative protein abundances by design.
After applying the same normalization criteria, we assessed its effect by plotting the mean intensity versus standard deviation for each peptide. The results showed no noticeable difference before and after normalization, indicating that the variance (standard deviation) is largely independent of the mean. This suggests that the data was already appropriately normalized by MaxLFQ (Figure 2.

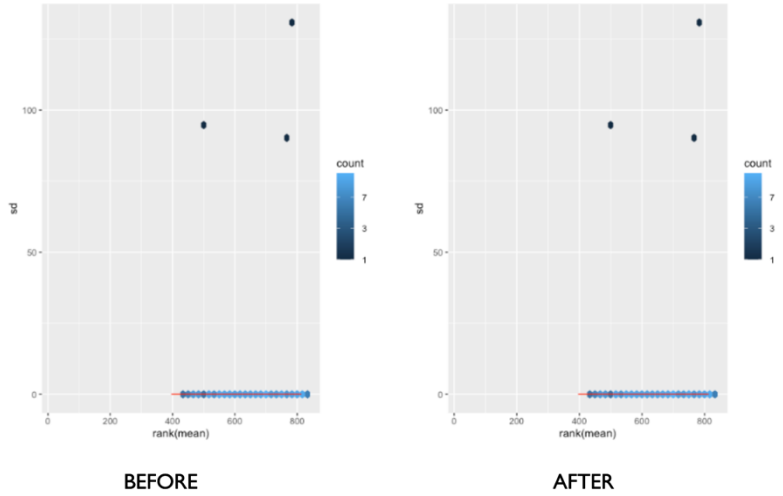

Figure 2. Mean intensity vs. standard deviation (sd) of peptide abundances before and after normalization. No significant differences are observed, indicating that MaxLFQ normalization effectively stabilizes variance across peptides.

We applied the same imputation strategy used in the original study: missing values were replaced with a small constant equal to one-fifth of the minimum abundance of the corresponding proteins. Our dataset contained a high number of missing (NA) and zero values prior to imputation. To assess the distribution of protein intensities after imputation, we generated a density plot. The plot confirmed a right-skewed distribution with a long tail, indicating that high-intensity proteins are relatively rare. And while imputation reduced the number of missing values, zero values remained abundant.

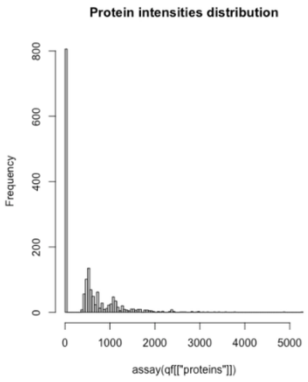

Figure 3. Density plot of protein intensities after imputation. The distribution remains right-skewed, with high-intensity proteins being less frequent. Zero values persist in the dataset, highlighting the prevalence of low or undetected signals.

Distribution of protein expression profiles

To compare our results with those reported in the paper, we performed principal component analysis (PCA). The first two components, PC1 and PC2, explain 47.8% of the total

variability. There is some separation between females (F) and males (M) along PC1, suggesting sex-specific variation in protein expression. One sample located in the upper region of PC2 may be an outlier or represent biological or technical variation not shared by other samples. Notably, this PCA plot visually supports the conclusion that sex influences protein expression patterns, consistent with findings from the original study (Figure 4).

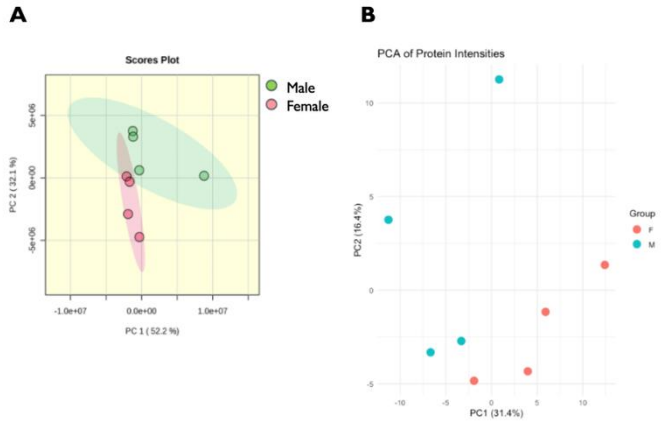

Figure 4. Principal component analysis (PCA) of protein expression. A. Adapted from original results (labelling error). Our PCA shows partial sex separation on PC1; one sample deviates on PC2.

Heatmap Analysis

We generated a heatmap to visualize protein expression patterns across samples. In the original study, a clear separation between male and female samples was observed. In our dataset, when plotting all proteins, this separation is less distinct (Figure 5).
To enhance interpretability, we zoomed in on the top 50 most differentially expressed proteins (based on log2 fold change) (Figure 5). This focused view shows a somewhat clearer pattern of sex-specific expression, but inconsistencies remain: one female sample clusters more closely with males, and one male sample does not fully align with the rest of the male group. These inconsistencies may reflect sample-specific variation, technical noise, or issues in protein detection and normalization.

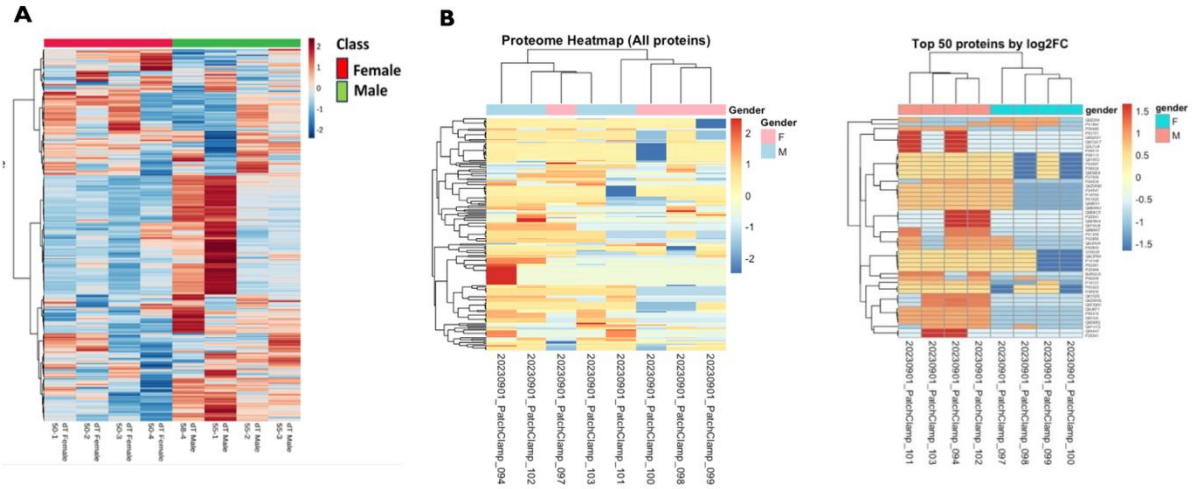

Figure 5. Heatmap of protein expression across male and female samples. A. Original results. B. Left: All proteins heat map show limited separation between sexes. B. Right: Top 50 most differentially expressed proteins shows clearer clustering, though one male and one female sample deviate from group patterns.

Volcano plot analysis and data processing concerns

Although our earlier analyses showed results similar to the original study, the volcano plot comparison reveals noticeable discrepancies. The original publication identified multiple proteins enriched in males compared to females, none of these among our top hits. Instead, our volcano plot highlights only four proteins, with differing functions, for example Q9JKF1 (Ras GTPase-activating-like protein) associates with calmodulin and may promote neurite outgrowth. These four proteins display extreme values, with a log2 fold change of around -30 and a -log10 p-value of 6, indicating an unrealistically strong enrichment in males. This raises concerns about potential errors in our data processing or filtering (Figure 6).

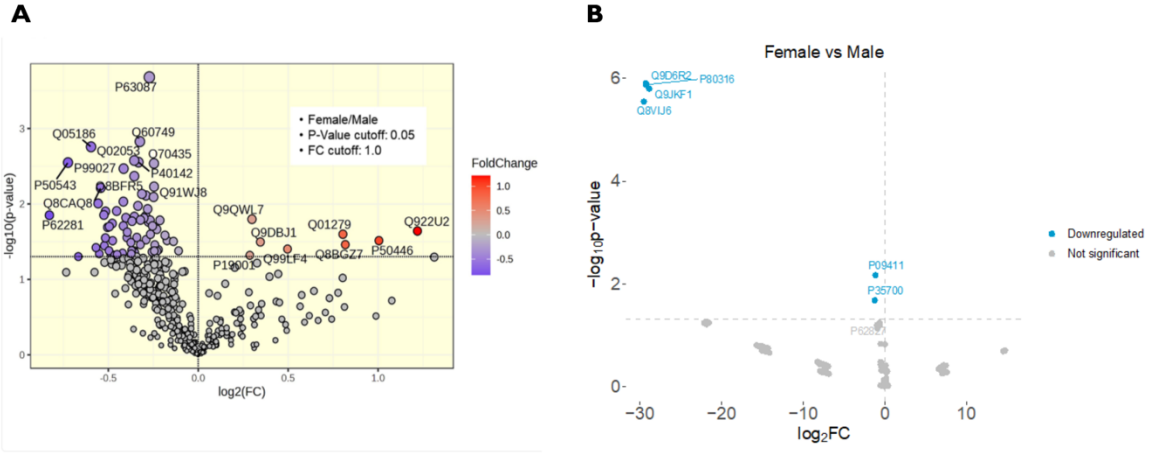

Figure 6. Volcano plot comparing protein expression between male and female samples. A. Original study results showing multiple proteins enriched in males. B. Our results show four significantly enriched proteins ($log_2FC \approx -30$, $-log_{10}P \approx 6$), likely exaggerated due to imputation bias and limited detection, highlighting potential processing issues.

Several factors may contribute to these differences. First, we used MaxQuant for data analysis, whereas the original study employed Proteome Discoverer, which may affect protein identification and intensity normalization. Additionally, the distribution of points in the volcano plot differs: the original study's proteins are dispersed, while our plot shows clustering, suggesting many proteins share similar fold changes or p-values. This pattern may stem from the high number of missing (NA) and zero values in our dataset, which after imputation were replaced with very low values, potentially biasing the results. This imputation approach might artificially inflate differences, especially if a protein is detected in only one out of eight samples, with the remaining values imputed as low intensities. Such bias could lead to misleadingly large fold changes in the volcano plot.
Overall, these observations suggest that our current approach may not adequately address missing values or filtering criteria. A more stringent filtering strategy, such as requiring proteins to be detected in at least two samples, might reduce noise and produce more biologically realistic results.
The large differences observed in the volcano plot likely stem from differences in protein identification, intensity normalization, and handling of missing data between our MaxQuant-based analysis and the original study. This highlights the need to carefully reconsider the data processing pipeline to improve the reliability of our results.

Conclusions

The original study reported clear sex-related differences in LC neuron activity and protein expression. While our analysis captures some of these patterns, such as partial sex separation in PCA and a few differentially expressed proteins, it does not replicate the full extent of the original findings. The lower number of identified proteins and inconsistencies in differential expression likely stem from differences in data processing, filtering, and imputation, particularly due to the use of MaxQuant.

References

Lee J, Wang ZM, Messi ML, Milligan C, Furdui CM, Delbono O. Sex differences in single neuron function and proteomics profiles examined by patch-clamp and mass spectrometry in the locus coeruleus of the adult mouse. Acta Physiol (Oxf). 2024 Apr;240(4):e14123.