Hands-on MaxQuant Analysis: Comprehensive Guide

Raw File Processing

- Load RAW files from mass spec
- Automatic peak detection
- Retention time alignment

Parameter Settings

- Enzyme: Trypsin/P
- Fixed/variable modifications
- FDR thresholds (1% peptide/protein)

Perseus Downstream

- Statistical analysis platform
- Filtering and normalization
- Differential expression analysis

Quality Assessment

- Check identification rates
- Review mass error distributions
- Evaluate quantification reproducibility

1 Raw File Processing

Raw file processing is the initial and crucial step in MaxQuant analysis where mass spectrometry data is imported, processed, and prepared for protein identification. This stage involves converting vendor-specific RAW files into a format that MaxQuant can analyze, extracting peptide features, and aligning data across multiple runs.

Raw File Processing Workflow

MS RAW Files (.raw, .wiff, .d)

↓

1 File Import & Conversion

Load vendor-specific files and convert to internal format

2 Peak Detection

Identify MS1 peaks and extract isotope patterns

3 3D Peak Assembly

Combine m/z, retention time, and intensity information

4 Retention Time Alignment

Align features across multiple runs for accurate quantification

↓

Processed Features Ready for Database Search

Key Processing Steps

1. Peak Detection Algorithm:

MaxQuant uses a sophisticated 3D peak detection algorithm that identifies peptide features based on their mass-to-charge ratio (m/z), retention time, and signal intensity. The algorithm recognizes isotope patterns and distinguishes true peptide signals from noise.

2. Retention Time Alignment:

This critical step ensures that the same peptide appearing in different LC-MS runs is correctly matched. MaxQuant builds a nonlinear alignment model that accounts for retention time shifts between runs, enabling accurate "match between runs" (MBR) functionality.

> **Pro Tip:** Enable "Match Between Runs" (MBR) feature to increase protein identification by transferring identifications from one run to another based on accurate mass and retention time alignment. This typically increases identifications by 30-50%.

Critical Considerations:

- ✓ Ensure RAW files are from the same instrument type for optimal alignment
- ✓ Check that gradient conditions are consistent across all runs
- ✓ Verify that file sizes are reasonable (corrupted files will cause processing errors)

✓ Monitor memory usage for large datasets (>50 files may require >32GB RAM)

2 Parameter Settings

Proper parameter configuration is essential for accurate protein identification and quantification. MaxQuant offers extensive customization options that must be carefully selected based on your experimental design, sample preparation method, and instrument type.

Key Parameter Categories

Enzyme Settings

Digestion specificity
Missed cleavages
Cleavage rules

Modifications

Fixed modifications
Variable modifications
Max modifications per peptide

Search Parameters

Quantification

Precursor mass tolerance
Fragment mass tolerance
Database selection

Label-free quantification
SILAC labels
iBAQ calculation

Essential Parameter Configuration

Parameter	Recommended Setting	Purpose
Enzyme	Trypsin/P	Cleaves after K/R, including before proline
Max Missed Cleavages	2	Allows incomplete digestion events
Fixed Modification	Carbamidomethyl (C)	Alkylation of cysteine residues
Variable Modifications	Oxidation (M), Acetyl (Protein N-term)	Common biological/chemical modifications
Max Modifications	5	Limits search space while capturing real modifications
Main Search Tolerance	4.5 ppm	Precursor mass accuracy for high-resolution MS
MS/MS Tolerance	20 ppm (Orbitrap) / 0.5 Da (Ion trap)	Fragment ion mass accuracy
PSM FDR	0.01 (1%)	Peptide spectrum match false discovery rate
Protein FDR	0.01 (1%)	Protein identification false discovery rate

Quantification Methods

Label-Free Quantification (LFQ):

MaxQuant's LFQ algorithm normalizes peptide intensities across samples and calculates protein abundances without requiring isotopic labeling. It uses the "MaxLFQ" algorithm which:

- ✓ Performs pairwise ratio comparison between samples
- ✓ Applies median normalization to account for loading differences
- ✓ Requires at least 2 ratio counts for quantification
- ✓ Handles missing values through advanced imputation strategies

> **Important:** Always include at least 3 biological replicates per condition for robust statistical analysis. Enable "Match Between Runs" to reduce missing values and improve quantification accuracy.

SILAC (Stable Isotope Labeling by Amino acids in Cell culture):

For SILAC experiments, configure the multiplicity settings to match your labeling scheme (Lys0/Arg0 for light, Lys4/Arg6 or Lys8/Arg10 for heavy labels). MaxQuant automatically identifies and quantifies peptide pairs.

---

3 Perseus Downstream Analysis

Perseus is a comprehensive statistical analysis platform designed specifically for processing MaxQuant output. It provides a wide range of tools for data filtering, normalization, statistical testing, and visualization, making it ideal for proteomic data interpretation.

Perseus Analysis Pipeline

MaxQuant proteinGroups.txt

↓

1. **Data Import & Annotation**
   Load protein intensities and define experimental groups

2. **Filtering & Transformation**
   Remove contaminants, reverse sequences, and apply log2 transformation

3. **Normalization**
   Median or Z-score normalization across samples

4. **Imputation**
   Handle missing values using appropriate methods

5. **Statistical Testing**
   T-test, ANOVA, or other appropriate tests

6 **Visualization & Export**

Create volcano plots, heatmaps, PCA plots

↓

Biological Insights & Publication-Ready Figures

Key Analysis Steps

1. Data Filtering:

Essential first step to ensure data quality:

- ✓ Remove proteins identified "Only by site" (modified peptides without unmodified evidence)
- ✓ Remove reverse database hits (false positives from decoy search)
- ✓ Remove potential contaminants (keratins, trypsin, BSA)
- ✓ Filter based on valid values (e.g., require protein to be present in at least 70% of samples in at least one group)

2. Log2 Transformation:

Transform intensity values to log2 scale to:

- ✓ Normalize the distribution of protein abundances

- ✓ Make the data more suitable for statistical tests assuming normal distribution
- ✓ Linearize fold-change relationships (2-fold up = +1, 2-fold down = -1)

3. Missing Value Imputation:

Perseus offers multiple imputation strategies:

Method	Use Case	Description
From Normal Distribution	MNAR (Missing Not At Random)	Imputes low values for proteins below detection limit
Replace by NaN	Keep missing values	No imputation, use tests that handle missing data
k-Nearest Neighbor	MAR (Missing At Random)	Estimates values based on similar proteins

Statistical Analysis

Two-Sample T-test:

For comparing two experimental groups (e.g., control vs. treatment):

- ✓ Use permutation-based FDR correction (recommended FDR < 0.05)
- ✓ Set S0 parameter (typically 0.1) to avoid over-emphasizing small fold changes
- ✓ Consider both statistical significance (p-value) and biological significance (fold change)

Volcano Plot Interpretation: Proteins in the upper left/right quadrants are both statistically significant (high -log10 p-value) and show substantial fold change. These are your primary candidates for biological interpretation.

Multi-Group Analysis (ANOVA):

For experiments with more than two groups:

- ✓ Identify proteins that vary significantly across groups
- ✓ Follow up with post-hoc tests for pairwise comparisons
- ✓ Use hierarchical clustering to identify expression patterns

Enrichment Analysis

Perseus integrates with annotation databases to perform functional enrichment analysis:

- ✓ Gene Ontology (GO) term enrichment - identify overrepresented biological processes
- ✓ KEGG pathway analysis - map proteins to metabolic and signaling pathways
- ✓ Protein domain enrichment - detect structural motifs
- ✓ Keyword enrichment - text mining of protein annotations

4 Quality Assessment

Quality assessment is crucial for ensuring the reliability and reproducibility of proteomic experiments. Systematic evaluation of multiple quality metrics helps identify technical issues, validate experimental procedures, and ensure that results are publication-ready.

Quality Control Metrics Dashboard

📊 Identification Metrics

MS/MS Count: 25,000-50,000

Peptides: 15,000-30,000 | Proteins: 3,000-6,000

Target: Consistent across replicates (CV < 10%)

🎯 Mass Accuracy

Precursor: < 5 ppm | Fragment: < 20 ppm

Distribution: Normal

Check: Histogram should be centered at 0 ppm

📈 Quantification Quality

CV < 20% (technical) | Correlation r > 0.9

Missing < 30%

Goal: High reproducibility between replicates

✅ Data Completeness

Sequence Coverage | Peptide Count/Protein

Unique Peptides

Ideal: ≥2 unique peptides per protein

Critical Quality Checks

1. Identification Rate Assessment:

Metric	Good Quality	Poor Quality	Action if Poor
MS/MS Identified (%)	> 50%	< 30%	Check sample prep, LC gradient, MS parameters
Protein Groups	3,000-6,000	< 1,000	Verify database, check sample complexity

Metric	Good Quality	Poor Quality	Action if Poor
Peptides per Protein	> 5 (median)	< 2 (median)	Improve digestion, increase gradient time
Unique Peptides	> 70%	< 50%	May indicate high protein redundancy

2. Mass Error Distribution:

The mass error histogram is one of the most important QC plots. It shows the distribution of mass deviations between observed and theoretical peptide masses.

What to Look For:

- ✓ Symmetric, bell-shaped distribution centered at 0 ppm
- ✓ Standard deviation < 3 ppm for high-resolution instruments
- ✓ No systematic shifts (would indicate calibration issues)
- ✓ Minimal outliers beyond ±10 ppm

Red Flag: If the mass error distribution is asymmetric or shows multiple peaks, this indicates serious calibration problems or incorrect search parameters. Re-run with recalibration enabled or check instrument calibration.

Quantification Reproducibility

Coefficient of Variation (CV) Analysis:

Calculate CV for all proteins across technical or biological replicates:

- ✓ Technical replicates: CV < 15% (excellent), < 20% (acceptable)
- ✓ Biological replicates: CV < 30% (good), < 50% (acceptable)
- ✓ Plot CV distribution as histogram - should be skewed toward low values
- ✓ High CV proteins may reflect true biological variability or technical noise

Correlation Analysis:

Create scatter plots comparing protein intensities between replicates:

- ✓ Technical replicates: Pearson r > 0.95
- ✓ Biological replicates: Pearson r > 0.85
- ✓ Points should cluster around diagonal with minimal scatter
- ✓ Check for outliers that may represent interesting biology or technical errors

Sample Quality Indicators

Principal Component Analysis (PCA):

PCA provides a global view of sample relationships and can reveal:

- ✓ Clustering of biological replicates (should group together)
- ✓ Separation between experimental conditions

- ✓ Potential batch effects or outlier samples
- ✓ Overall data structure and variance composition

Intensity Distribution:

Examine box plots of log2 protein intensities across all samples:

- ✓ All samples should show similar median intensities
- ✓ Similar distribution shape and range
- ✓ Major differences may indicate loading issues or technical problems
- ✓ Use normalization if systematic differences are observed

Final QC Checklist

Before Publishing: Ensure all of the following criteria are met:

- ✓ ✓ Identification rates are consistent across all samples (< 10% variation)
- ✓ ✓ Mass accuracy is excellent (< 5 ppm standard deviation)
- ✓ ✓ Technical replicates show high correlation (r > 0.95)
- ✓ ✓ Biological replicates cluster in PCA space
- ✓ ✓ No significant batch effects detected

- ✓ ✓ FDR thresholds are appropriate (1% at both PSM and protein level)
- ✓ ✓ Sufficient peptides per protein for confident identification (≥2 unique)
- ✓ ✓ Missing value patterns are reasonable and have been properly handled
- ✓ ✓ Quantification CVs are within acceptable ranges
- ✓ ✓ All quality control plots have been reviewed and archived

Remember:

Quality assessment is not just a box to check at the end of analysis. It should be an ongoing process that guides parameter optimization and identifies issues early. Invest time in thorough QC to ensure your results are reproducible, reliable, and publication-ready. Poor quality data cannot be rescued by sophisticated statistics!