INTRODUCTION TO MSSTATS

Olga Vitek

College of Science
College of Computer and Information Science

Northeastern University
Example: A label-free experiment

Question: which proteins change in abundance?

A typical analysis workflow (also in MSstats)

**Experimental design**
- QC and normalization
  - Statistical modeling
    - Model-based conclusions
      - Experimental design

**Stresses or conditions**
- Time course
- Multiple groups
- Paired design

**Problem statement**
- Screening experiment
- Confirmation experiment

A typical analysis workflow (also in MSstats)

**Experimental design**

**Stresses or conditions**
- Time course
- Multiple groups
- Paired design

**Problem statement**
- Screening experiment
- Confirmation experiment

**QC and normalization**

**Statistical modeling**

**Model-based conclusions**

**Experimental design**

---

**Figure 1:** Study of breast cancer cell lines. Two cultures from two breast cancer cell lines (MCF7, Hs578T).

**Figure 3:** Exploratory data analysis by colored boxes. The study had a combination of a time course and a group comparison design.

---

**Table:**

<table>
<thead>
<tr>
<th></th>
<th>normoxia</th>
<th>normoxia</th>
</tr>
</thead>
<tbody>
<tr>
<td>low invasive</td>
<td></td>
<td>high invasive</td>
</tr>
</tbody>
</table>

**Graphs:**

- Log Abundance vs. time for normoxia and hypoxia conditions for different treatment times (6 and 24 hours).

---

A typical analysis workflow (also in MSstats)

Experimental design

QC and normalization

Statistical modeling

Model-based conclusions

Experimental design

Stresses or conditions
- Time course
- Multiple groups
- Paired design

Problem statement
- Screening experiment
- Confirmation experiment

T. Clough et al. BMC Bioinformatics, 2012
A typical analysis workflow (also in MSstats)

- Experimental design
  - Stresses or conditions
    - Time course
    - Multiple groups
    - Paired design
  - Problem statement
    - Screening experiment
    - Confirmation experiment
- QC and normalization
- Statistical modeling
- Model-based conclusions
- Experimental design
A typical analysis workflow (also in MSstats)

**Experimental design**

**Data preparation**
- Formatting
- Visualization
- Remove artifacts
- Normalization

**QC and normalization**

**Statistical modeling**

**Model-based conclusions**

**Experimental design**

\[ \text{log}_2(\text{intensity}) \text{ of all transitions in the run} \]

C.-Y. Chang et al. *MCP*, 2012
A typical analysis workflow (also in MSstats)

Experimental design

QC and normalization

Statistical modeling

Summarize all protein features in a statistical model

- Systematic variation
- Random variation

Verify the assumptions!

\[
\log(\text{peak intensity}) = \mu_{i11} + F_i + C_j + (F \times C)_{ij} + S(C)_{k(j)} + \varepsilon_{ijkl}
\]
Finding differentially abundant proteins

Simple example: one protein, one feature per protein, label-free

Log(abundance) of a feature in a run

Table 1: Outcomes of testing m null hypotheses H0: ‘status quo’, no change in abundance, \( \hat{G}_1 - \hat{G}_0 = 0 \)
Ha: change in abundance, \( \hat{G}_1 - \hat{G}_0 \neq 0 \)

\[
\text{observed } t = \frac{\hat{G}_1 - \hat{G}_0}{\sqrt{\text{Estimate of variation}}}
\]
Linear mixed models describe Normal distributions

**Multiple conditions allow us to better learn the extent of variation**

\[
\text{log(peak intensity)} = \text{reference abundance} + \text{LC-MS feature} + \text{condition} + \text{feature \times condition interaction} + \text{biol. replicate} + \text{Random meas. error}
\]

\[
y_{ijkl} = \mu_{111} + F_i + C_j + (F \times C)_{ij} + S(C)_{k(j)} + \epsilon_{ijkl}
\]

where

\[
F_1 = C_1 = (F \times C)_{i1} = (F \times C)_{1j} = 0
\]

expanded scope of biological replication: \( S(C)_{k(j)} \overset{iid}{\sim} N\left(0, \sigma^2_S\right) \)
Labeled reference peptides help separate the biological and the technological variation.

**Label-based SRM workflow**

**Analysis of heavy/light peak pairs**

**Table of quantified peaks**

<table>
<thead>
<tr>
<th>Run M</th>
<th>Group 1</th>
<th>Group 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject J</td>
<td>Subject J</td>
<td>Subject J</td>
</tr>
</tbody>
</table>

Legend: **Label** | **Feature**: Transition/Peptide | **Group** | **Run** | **Subject**
A full linear mixed model for an experiment with labeled reference peptides

Example: ovarian cancer dataset

\[
\begin{align*}
\text{observed log2(int of peak)} &= \text{overall mean} + \text{group or time} + \text{subject} + \text{feature} + \text{run} + \text{group by feature} + \text{run by feature} + \text{random error} \\
\end{align*}
\]

\[
y_{ijklm} = \mu + G_i + S(G)_{j(i)} + F_{kl} + R_m + (G \times F)^*_{ikl} + (R \times F)^*_{klm} + \varepsilon_{ijklm} \\
\]

Fixed/Random

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>overall</td>
<td>mean</td>
<td>group</td>
<td>or time</td>
<td>subject</td>
<td>feature</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>F/R</td>
<td>F</td>
<td>F</td>
<td>F/R</td>
</tr>
</tbody>
</table>

Model log2(int) instead of ratios light/heavy

‘Run’ pairs endogenous and reference transitions from a same run
Deviations from independence or from constant variance are often mistaken for deviations from Normality.
A typical analysis workflow (also in MSstats)

1. Experimental design
2. QC and normalization
3. Statistical modeling
4. Model-based group comparisons
   - Quantify the uncertainty
   - Adjust p-values to control FDR

Relative protein quantification
- In one sample
- In one condition

-log10(FDR-adjusted p-value) vs. log2 FC

C.-Y. Chang et al. MCP, 2012
A typical analysis workflow (also in MSstats)

- Experimental design
- QC and normalization
- Statistical modeling
- Model-based conclusions

**Model-based group comparisons**
- Quantify the uncertainty
- Adjust p-values to control FDR

**Relative protein quantification**
- In one sample
- In one condition

**Heatmap:**
- All proteins, several comparisons

- Color Key
  - Blue: no significant change in abundance
  - Red: significant up-regulation
  - Blue: no significant change in abundance
  - Black: significant down-regulation

- Heatmap of results of testing proteins for differential abundance.

- Table of adjusted p-values and fold change cutoffs for T7-T1, T3-T1, T5-T1 conditions.
A typical analysis workflow (also in MSstats)

1. **Experimental design**
   - Model-based average log abundance

2. **QC and normalization**
   - Low nm 6 hrs
   - Low nm 24 hrs
   - Low hyp 6 hrs
   - Low hyp 24 hrs
   - High nm 6 hrs
   - High nm 24 hrs
   - High hyp 6 hrs
   - High hyp 24 hrs

3. **Statistical modeling**
   - Model-based estimated log-abundance

4. **Model-based group comparisons**
   - Quantify the uncertainty
   - Adjust p-values to control FDR

5. **Relative protein quantification**
   - In one sample
   - In one condition

---

**Week**

0 4 5 6 9 10 13 14 16 17 e 03 06 08

---

Model-based estimated log-abundance
**Model-based conclusions**

*Comparisons between conditions are estimated by linear combinations of model terms*

\[
\log(\text{peak intensity}) = \text{Expected reference abundance} + \text{LC-MS + condition feature} + \text{feature \times condition interaction} + \text{biol. replicate} + \text{Random meas. error}
\]

\[
y_{ijkl} = \mu_{111} + F_i + C_j + (F \times C)_{ij} + S(C)_{k(j)} + \varepsilon_{ijkl}
\]

**Quantity of interest:**

\[
H_0 : L = \bar{\mu}_{[\text{high, nm, 6}]} - \bar{\mu}_{[\text{low, nm, 6}]} = 0
\]

**Model-based estimate and test statistic:**

\[
\hat{L} = \hat{C}_{[\text{high, nm, 6}]} + \frac{1}{I} \sum_{i=1}^{I} (F \times C)_{i,[\text{high, nm, 6}]} + \frac{1}{K} \sum_{k=1}^{K} S(C)_{k([\text{high, nm, 6}])}
- \left( \hat{C}_{[\text{low, nm, 6}]} + \frac{1}{I} \sum_{i=1}^{I} (F \times C)_{i,[\text{low, nm, 6}]} + \frac{1}{K} \sum_{k=1}^{K} S(C)_{k([\text{low, nm, 6}])} \right)
\]

\[
t = \frac{\hat{L}}{SE(L)} \sim \text{Student distribution}
\]

**In balanced datasets:**

\[
\hat{L} = \bar{Y}_{[\text{high, nm, 6}]} - \bar{Y}_{[\text{low, nm, 6}]}
\]

\[
t = \frac{\hat{L}}{\sqrt{\frac{\hat{\sigma}^2}{IR \sigma^2_{\text{Error}}}}} \sim \text{Student}_{IJK(L-1)+(I-1)(K-1)} \text{ distribution}
\]

*Model-based S/N*
A typical analysis workflow (also in MSstats)

1. Experimental design
2. QC and normalization
3. Statistical modeling
4. Model-based conclusions

Use the dataset to improve:
- Subject selection: matching
- Resource allocation: blocking
- Calculation of sample size

Experimental design

Desired fold change

# biological replicates per group

- FDR: 1% of changes
- FDR: 50% of changes
- Single feature

mass/charge
Linear mixed effects models are required to calculate the sample size and the power

Statistical power = \( P(\text{detect change}) = 0.8 \)

Sample size = 3

**Need to know in advance:**

- \( q \) - the False Discovery Rate
- \( m_0/m_1 \) - anticipated ratio of unchanging features
- \( \beta \) - statistical power (i.e. probability of a true positive discovery)
- \( \Delta \) - anticipated (log-) fold change
- \( \sigma^2_{\text{Indiv}} \) and \( \sigma^2_{\text{Error}} \) - anticipated variance