analysys

May 21, 2025

1 Log2 transformation to the data

According to my searches, DIA-Skyline intensity values are:

- Right-skewed: Intensities span several orders of magnitude.
- Heteroscedastic: High-intensity peptides show larger variance.
- Non-normal: Violates logistic regression assumptions (even though it's more robust than linear regression).

Then, a log2 transformation of the data should make it more suitable to apply logistic regression and meet its assumptions.

2 Normalization of the data

Even with targeted methods like DIA:

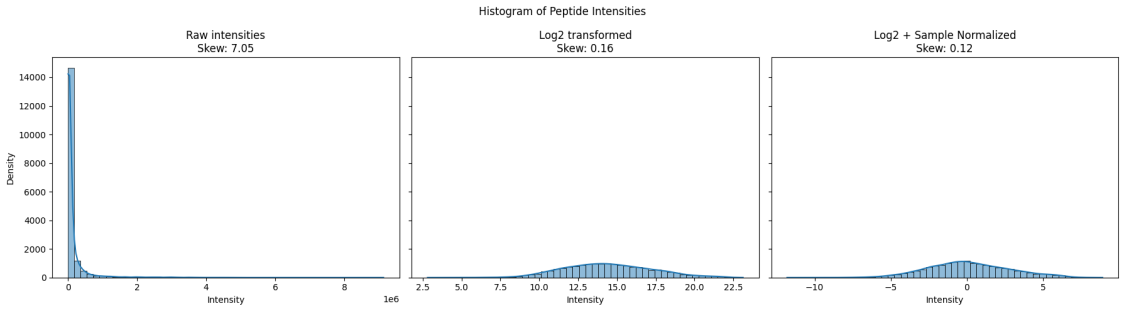
- Sample injection volumes may vary slightly.
- LC-MS performance can drift over time (batch effects).
- Total ion current per run is not constant.

As a result, raw intensities for the same peptide may differ not due to biology, but due to technical effects, making sample-to-sample comparisons misleading.

Thus, sample-wise normalization should reduce technical bias, increase comparability of peptide profiles and improves model robustness.

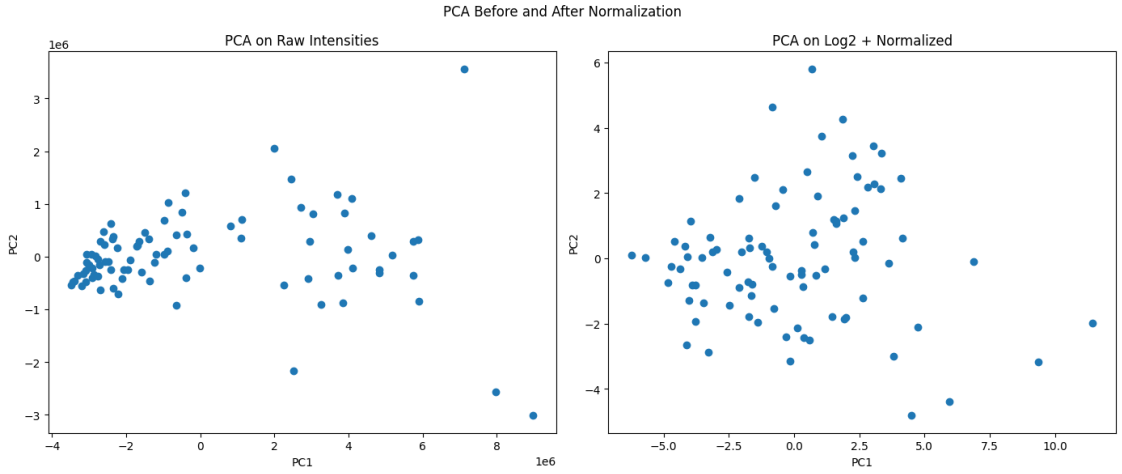
3 Checking the skeweness of the data

If our log2 transformation has worked well, the data should be more normally distributed and not skewed, otherwise it is not meeting the logistic regression assumptions.



4 Checking the removal of the technical bias through normalization

Each point of the following PCA is a sample. PCA projects high-dimensional intensity profiles (many peptides) into 2D, the distance between the points approximates the similarity of samples' peptide expression profiles.



Before normalization we see set of closer points and points spreading around (in the PC1 axis mostly).

This typically suggests that:

- Some samples have much higher total intensity (high injection volume, LC bias, etc.).
- These dominate the first principal component (PC1), pulling points away.
- The tighter cluster are samples with similar total intensity, while the spread ones are outliers due to technical variation — not biology.

These variations may reflect technical artifacts, not underlying biology.

After log2 + median normalization the points are sparse, with no agglomeration.These suggests:

- The major technical bias (total intensity shift across samples) has been removed.
- The PCA space shows more balanced spread: samples now differ by subtler differences in peptide profiles rather than being dominated by signal strength.

This **suggests** that we have ve eliminated global shifts and can now analyze true signal.

5 Model computing

Now, we can use the log2+normalized data in order to compute the logistic models.

Pathology	1	0
count	48	36

We have set benign and hyperplasia samples as controls and EC samples as cancers. We have 48 cancers and 36 controls.

6 Results summary

As a summary for the results, we're showing the classification performance for the peptides of the WomEC selected biomarkers. The full results table for individual biomarkers, along with the ROC curves plots and coordinates, can be found in the results folder within this analysis.

Also, as a clarification, the Sensitivity and Specificity values show in the following tables are those maximizing the sum of both metrics. For a full detail of the ROC coordinates find the corresponding model within the results folder.

Biomarker_1	AUC	Sensitivity	Specificity	NPV	PPV
MVP_ELPPGVEELLNK	1.00000	1.00000	1.00000	1.00000	1.00000
PKM_APIIAVTR	0.82060	0.75000	0.86111	0.72093	0.87805
PKM_NTGIICTIGPASR	0.81771	0.66667	0.94444	0.68000	0.94118
CLIC1_LAALNPESNTAGLDIFAK	0.77267	0.67391	0.82857	0.65909	0.83784
AGRN_LELGIGPGAATR	0.69108	0.59091	0.77778	0.53846	0.81250
CLIC1_NSNPALNDNLEK	0.68981	0.83333	0.47222	0.68000	0.67797
MVP_LAQDPFPLYPGEVLEK	0.66034	0.78049	0.55556	0.62500	0.72727
MMP9_SLGPALLLLQK	0.60426	0.85106	0.50000	0.68182	0.72727
MPO_IANVFTNAFR	0.58775	0.22917	0.93939	0.45588	0.84615
MPO_VVLEGGIDPILR	0.58498	0.89130	0.36364	0.70588	0.66129

As can be observed, MPV show perfect results for one of its peptides, but the other one has a poor performance. PKM (KPYM) peptides in the other hand show consistend results amongst them with an AUC around 0.82.