Mass Spectrometry Lemierre - Working document

Gustav Torisson

Descripton of raw data

- Raw data consists of a wide dataframe with 654 rows with protein names and 30 columns (1 with protein names and 29 samples.
- These were grouped into 8 with Lemierre syndrome ("LS"), 15 other sepsis ("Sepsis"), 3 Deep vein thrombosis ("DVT") and 3 other septic thromboses ("Other_thrombosis").
- There was one measurement per sample.
- Each data point represents an intensity measure from Mass pectrometry.

Initial data management

- Proteins with several names, separated with semicolon(;) were renamed to only the first name (left of semicolon)
- Datapoints labelled as "Filtered" were re-labelled as NA.
- Data was converted to "numeric" format, as it was "character" in Excel
- Datapoints labeeled as Nan (Not a Number) were also labelled as NA

Log2 transformation

- all measurements were log2 transformed

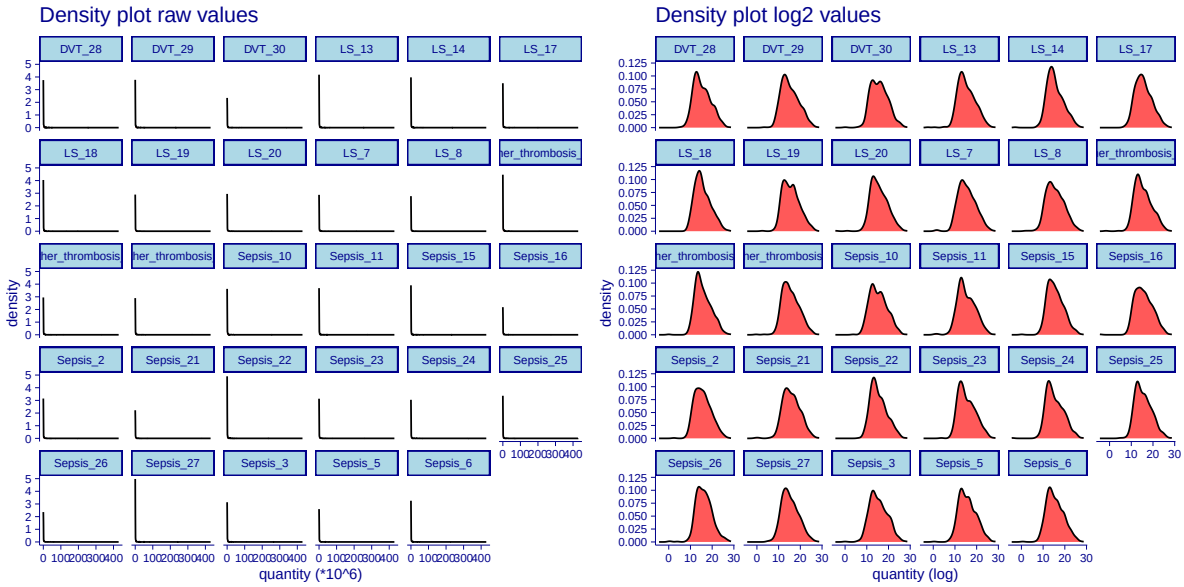


Figure 1: Density plots of raw and log2 transformed values

Filtering

- Of 654 proteins, 329 (50.3%) had complete data, in all 29 samples
- 200 proteins (30.6%) were missing in > 9 (30% of all) samples.
- These were filtered, leaving 454 proteins

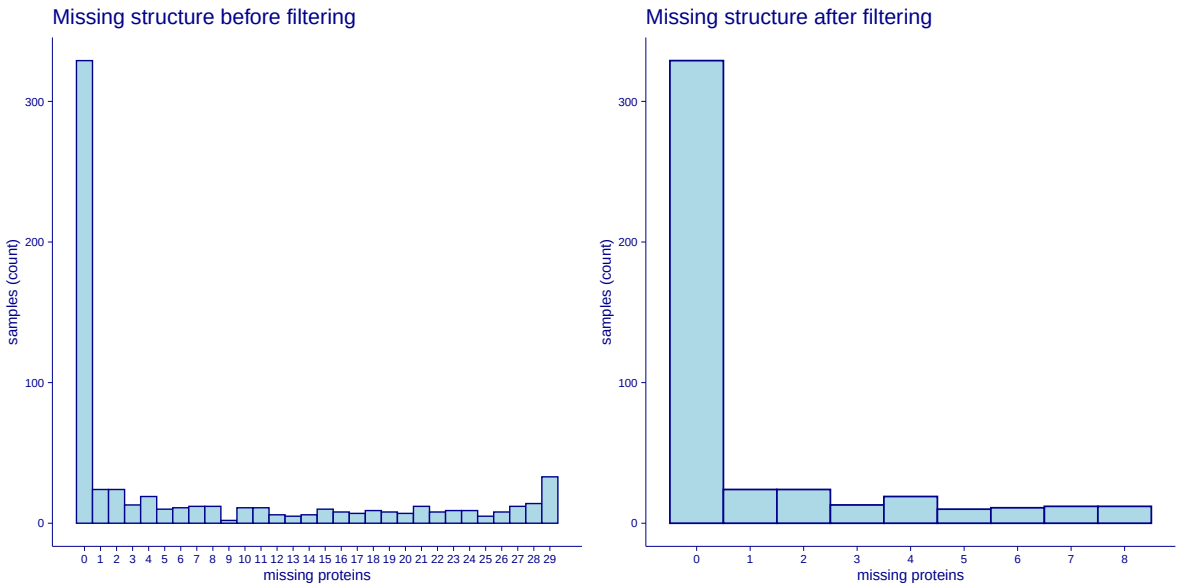


Figure 2: Missing structure before and after filterina

Normalisation

- all values were normalised by sample by subtracting the sample median from the Log2 intensity values

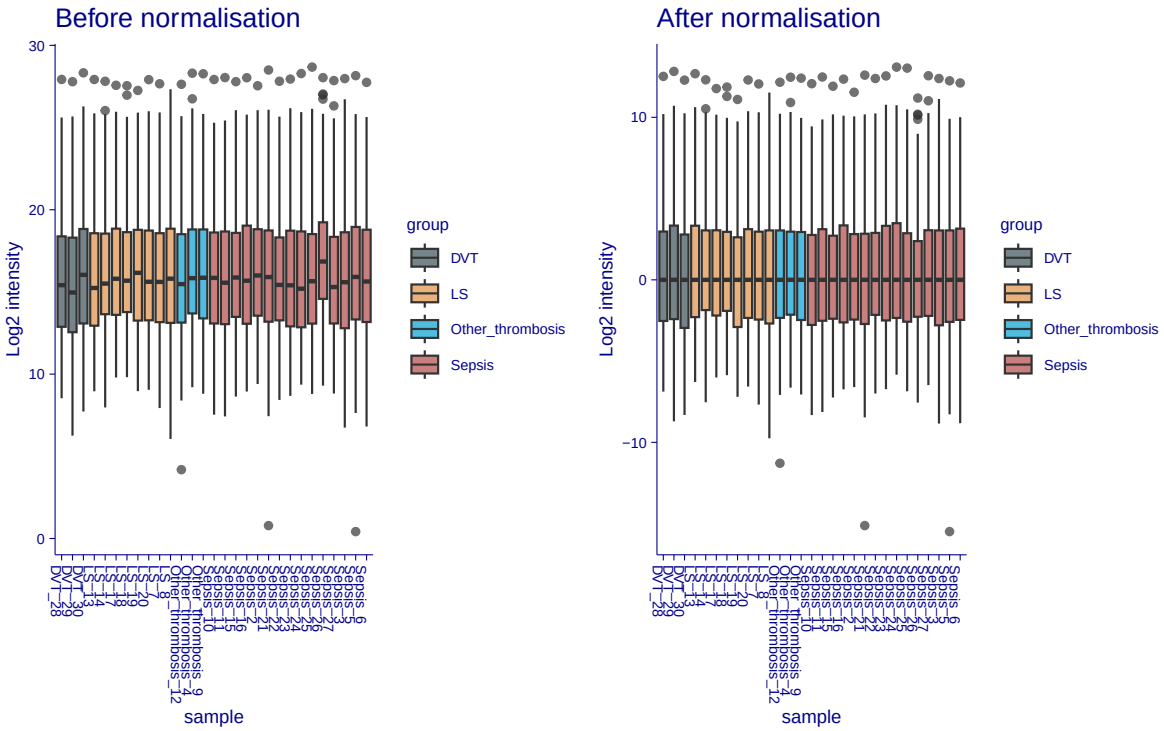


Figure 3: Before and after by-sample normalisation.

Missing per sample and group

- there were most missing values in the DVT group, but differences were small
- All NAs (both "Filtered" and NaN) were considered to represent low intensities and to represent MNAR (Correct???)

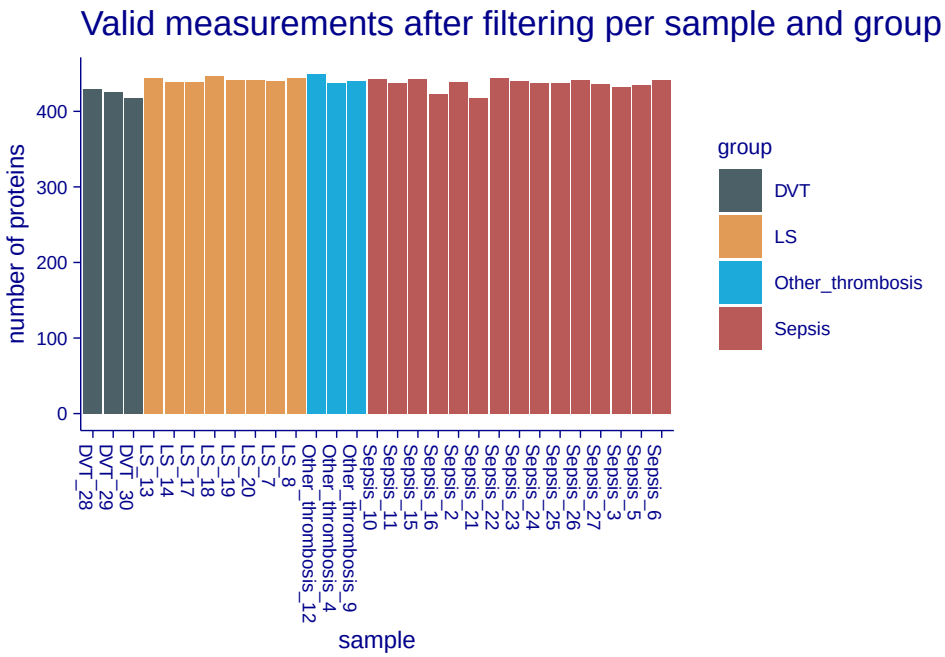


Figure 4: Valid measurements after filtering per sample and group

Imputation

- NAs were imputed using single imputation, assuming MNAR
- For each sample, the sample mean and sample sd were determined
- Then imputations were performed, using a random draw from a Gaussian distribution
- The mean for imputations was donwshifted with -1.8 sample sd and the width 0.3 * sample sd

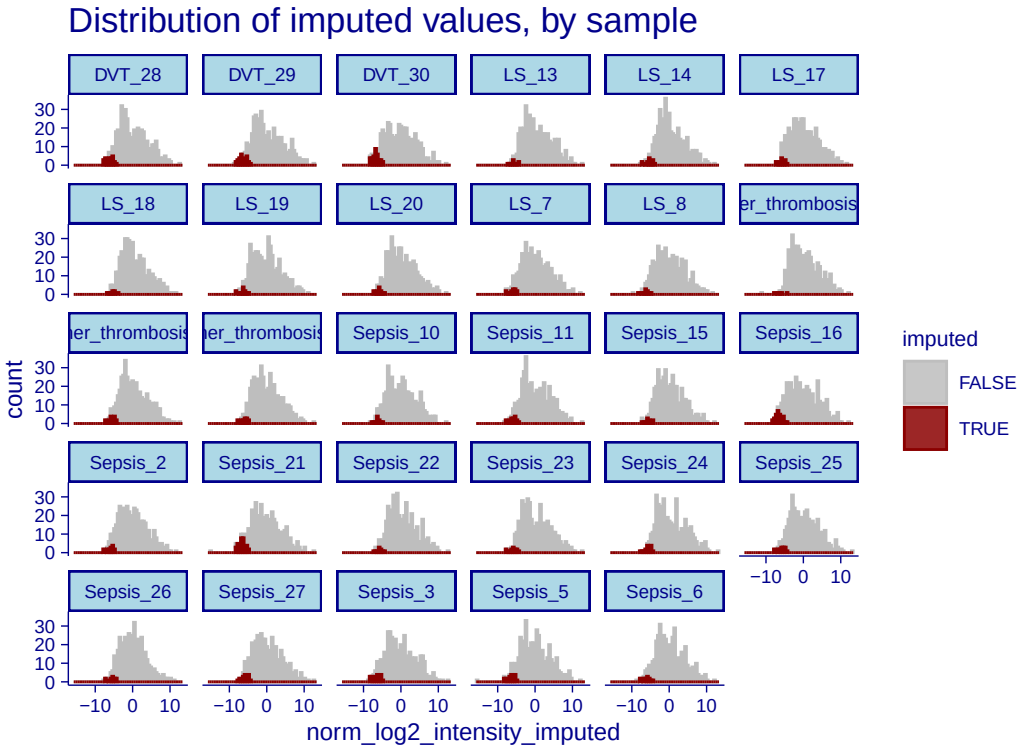


Figure 5: Distribution of imputed values.

Differential expression

- 454 t-tests (students t-test) were performed, one for each protein, between LS and Sepsis groups
- Results are presented as:
  - Log2FC = mean(log(LS)) - mean(log(Sepsis))
  - p values from t-test
  - q values using Benjamini-Hochberg corrections
  - FC(Fold change) = 2^Log2FC
  - Values with Log2FC ± 1.0 and q value < 0.05 were considered significant
  - number of imputed NAs (in LS and Sepsis groups)

Table 1: Differential expression between LS and Sepsis.

protein	log2FC	pval	qval	FC	sign	nas
P07988	2.803815	0.00026	0.0119	6.98	+	7
P80188	2.577285	0.00240	0.0404	5.97	+	4
P05362	2.429348	0.00006	0.0077	5.39	+	0
P01833	2.229912	0.00197	0.0358	4.69	+	0
Q8NBJ4	2.211645	0.00014	0.0094	4.63	+	1
P22897	2.155904	0.00037	0.0141	4.46	+	3
Q8WWZ8	2.010152	0.00109	0.0274	4.03	+	1
Q8N6C8	1.993165	0.00019	0.0094	3.98	+	3
P59665	1.859841	0.00073	0.0238	3.63	+	0
Q9Y6R7	1.797364	0.00002	0.0050	3.48	+	0
O43493	1.738423	0.00121	0.0274	3.34	+	3
P13987	1.726551	0.00019	0.0094	3.31	+	0
P13796	1.523030	0.00000	0.0009	2.87	+	0
P18065	1.456903	0.00103	0.0274	2.75	+	0
P01011	1.228792	0.00033	0.0136	2.34	+	0
P16070	1.209563	0.00007	0.0077	2.31	+	0
P08637	1.205137	0.00217	0.0380	2.31	+	0
P02763	1.192744	0.00041	0.0144	2.29	+	0
P19652	1.119510	0.00016	0.0094	2.17	+	0
P02649	1.004638	0.00080	0.0243	2.01	+	0
P02647	-1.097774	0.00102	0.0274	0.47	+	0
P80108	-1.765261	0.00009	0.0084	0.29	+	0
A0A0C4DH33	-2.345859	0.00295	0.0479	0.20	+	5

Including LS vs VTE

- the proteins that were differentially expressed for LS vs Sepsis were also evaluated for differences between LS and VTE.

Table 2: LS vs Sepsis (left, FC1, q1 and sign1) and LS vs VTE (right, FC2, q2, sign2)

protein	FC1	q1	sign1	FC2	q2	sign2
P07988	6.98	0.0119	+	2.27	0.4992	-
P80188	5.97	0.0404	+	20.49	0.0338	+
P05362	5.39	0.0077	+	11.72	0.0157	+
P01833	4.69	0.0358	+	3.23	0.1478	-
Q8NBJ4	4.63	0.0094	+	19.58	0.0205	+
P22897	4.46	0.0141	+	37.11	0.0010	+
Q8WWZ8	4.03	0.0274	+	2.97	0.1137	-
Q8N6C8	3.98	0.0094	+	8.89	0.0205	+
P59665	3.63	0.0238	+	8.07	0.0338	+
Q9Y6R7	3.48	0.0050	+	3.69	0.0396	+
O43493	3.34	0.0274	+	3.24	0.0807	-
P13987	3.31	0.0094	+	3.95	0.1087	-
P13796	2.87	0.0009	+	3.22	0.0224	+
P18065	2.75	0.0274	+	3.63	0.0977	-
P01011	2.34	0.0136	+	3.62	0.0305	+
P16070	2.31	0.0077	+	3.12	0.0403	+
P08637	2.31	0.0380	+	4.43	0.0072	+
P02763	2.29	0.0144	+	2.67	0.0487	+
P19652	2.17	0.0094	+	2.15	0.0905	-
P02649	2.01	0.0243	+	1.48	0.4023	-
P02647	0.47	0.0274	+	0.24	0.0224	+
P80108	0.29	0.0084	+	0.13	0.0157	+
A0A0C4DH33	0.20	0.0479	+	0.10	0.0949	-

Volcano plot

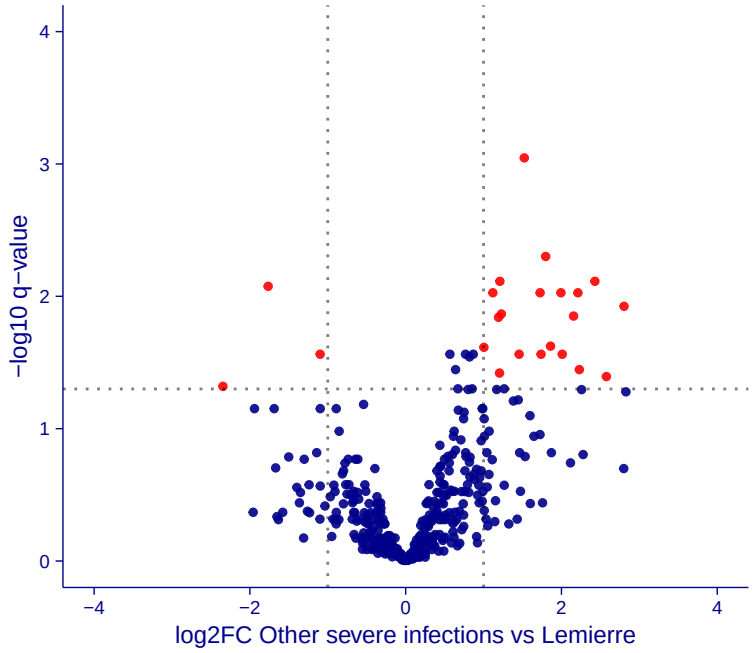
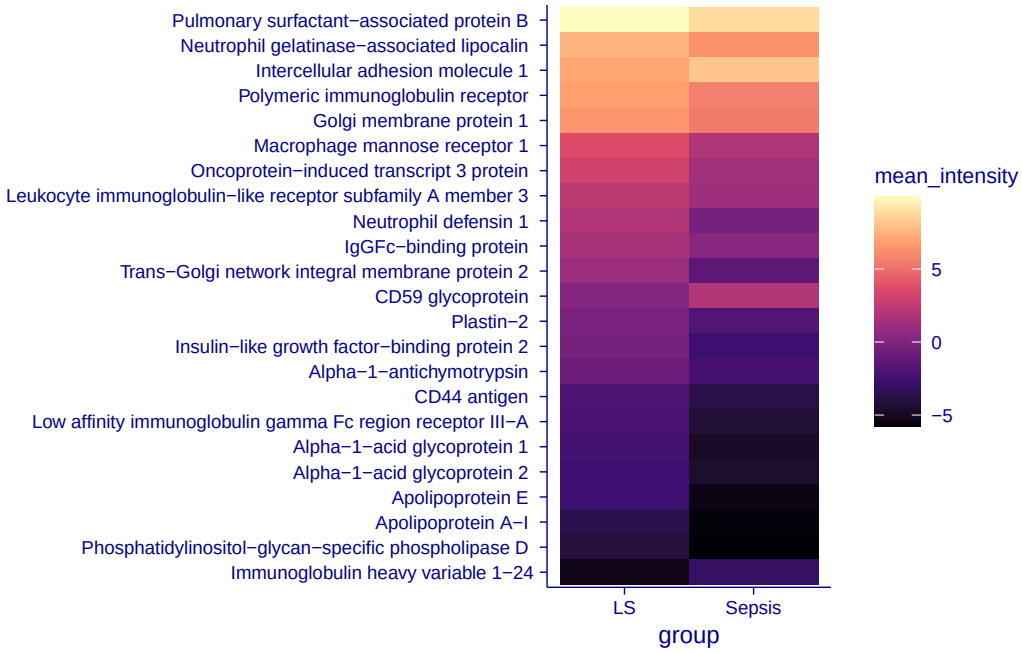


Figure 6: Volcano plot. Proteins that are differentially expressed between LS and Sepsis = red

Heatmap


Pulmonary surfactant−associated protein B
Neutrophil gelatinase−associated lipocalin
Intercellular adhesion molecule 1
Polymeric immunoglobulin receptor
Golgi membrane protein 1
Macrophage mannose receptor 1
Oncoprotein−induced transcript 3 protein
Leukocyte immunoglobulin−like receptor subfamily A member 3
Neutrophil defensin 1
IgGFc−binding protein
Trans−Golgi network integral membrane protein 2
CD59 glycoprotein
Plastin−2
Insulin−like growth factor−binding protein 2
Alpha−1−antichymotrypsin
CD44 antigen
Low affinity immunoglobulin gamma Fc region receptor III−A
Alpha−1−acid glycoprotein 1
Alpha−1−acid glycoprotein 2
Apolipoprotein E
Apolipoprotein A−I
Phosphatidylinositol−glycan−specific phospholipase D
Immunoglobulin heavy variable 1−24
LS
Sepsis
group
mean_intensity
5
0
−5

KOMMENTARER TILL DAVID

- Jag fick 23 protein jmfrt med 25 i manuset. det är 3 av dina som inte är med (mer om det nedan) och sen har det tillkommit ett (A0A0C4DH33), detta har qval precis < 0.05 och 5 imputationer så rätt känsligt för variationer i imputationen
- Jag gjorde volcano plotten med -log10 q värde istället för pvärde (känns mer rätt eftersom man kan sätta strecket på 1.3 för -log10), därför de skiljer sig på y-axeln.
- Nedan finns en tabell med de 25 proteiner som var signifikanta i din körning, de är sorterade i bokstavsordning
- 3 st är inte med här men var med i manuset, det är
  - P05556 - som verkar imputationskänslig, varierar väldigt mkt
  - P07998 och Q08380 - som ligger väldigt nära 0.05 och därför blir känsliga för variation

protein	log2FC	pval	qval	FC	sign	nas
O43493	1.738423	0.00121	0.0274	3.34	+	3
P01011	1.228792	0.00033	0.0136	2.34	+	0
P01833	2.229912	0.00197	0.0358	4.69	+	0
P02647	-1.097774	0.00102	0.0274	0.47	+	0
P02649	1.004638	0.00080	0.0243	2.01	+	0
P02763	1.192744	0.00041	0.0144	2.29	+	0
P05362	2.429348	0.00006	0.0077	5.39	+	0
P05556	-1.239634	0.06336	0.2656	0.42	-	5
P07988	2.803815	0.00026	0.0119	6.98	+	7
P07998	1.168512	0.00357	0.0506	2.25	-	0
P08637	1.205137	0.00217	0.0380	2.31	+	0
P13796	1.523030	0.00000	0.0009	2.87	+	0
P13987	1.726551	0.00019	0.0094	3.31	+	0
P16070	1.209563	0.00007	0.0077	2.31	+	0
P18065	1.456903	0.00103	0.0274	2.75	+	0
P19652	1.119510	0.00016	0.0094	2.17	+	0
P22897	2.155904	0.00037	0.0141	4.46	+	3
P59665	1.859841	0.00073	0.0238	3.63	+	0
P80108	-1.765261	0.00009	0.0084	0.29	+	0
P80188	2.577285	0.00240	0.0404	5.97	+	4
Q08380	1.264859	0.00341	0.0500	2.40	-	0
Q8N6C8	1.993165	0.00019	0.0094	3.98	+	3
Q8NBJ4	2.211645	0.00014	0.0094	4.63	+	1
Q8WWZ8	2.010152	0.00109	0.0274	4.03	+	1
Q9Y6R7	1.797364	0.00002	0.0050	3.48	+	0