Omics (2024) – Proteomics data analysis

FREDRIK.LEVANDER@IMMUN.LTH.SE

Goals of today

- Some knowledge about benefits of mzML standard MS data format
- Be familiar with concepts of:
  - Quality control of LC-MS data
  - Normalisation of quantitative protein data
  - Differential abundance analysis
  - Functional analysis / gene set level analyses

Mass spectrometry file formats

- Raw data files are in instrument vendor-specific binary formats.
  - Only accessible programatically through libraries provided by vendors
- Standard formats developed to allow for platform-independent access
  - Text-based formats that can be read and written on Windows, Mac, Linux
  - mzML: XML (eXtensible Markup Language) format representing everything (almost) in the raw data.

    https://doi.org/10.1074/mcp.R110.000133 , https://www.psidev.info/mzml
  - Mascot Generic Format (MGF). Simple text format for representation of MS or MS/MS spectra.

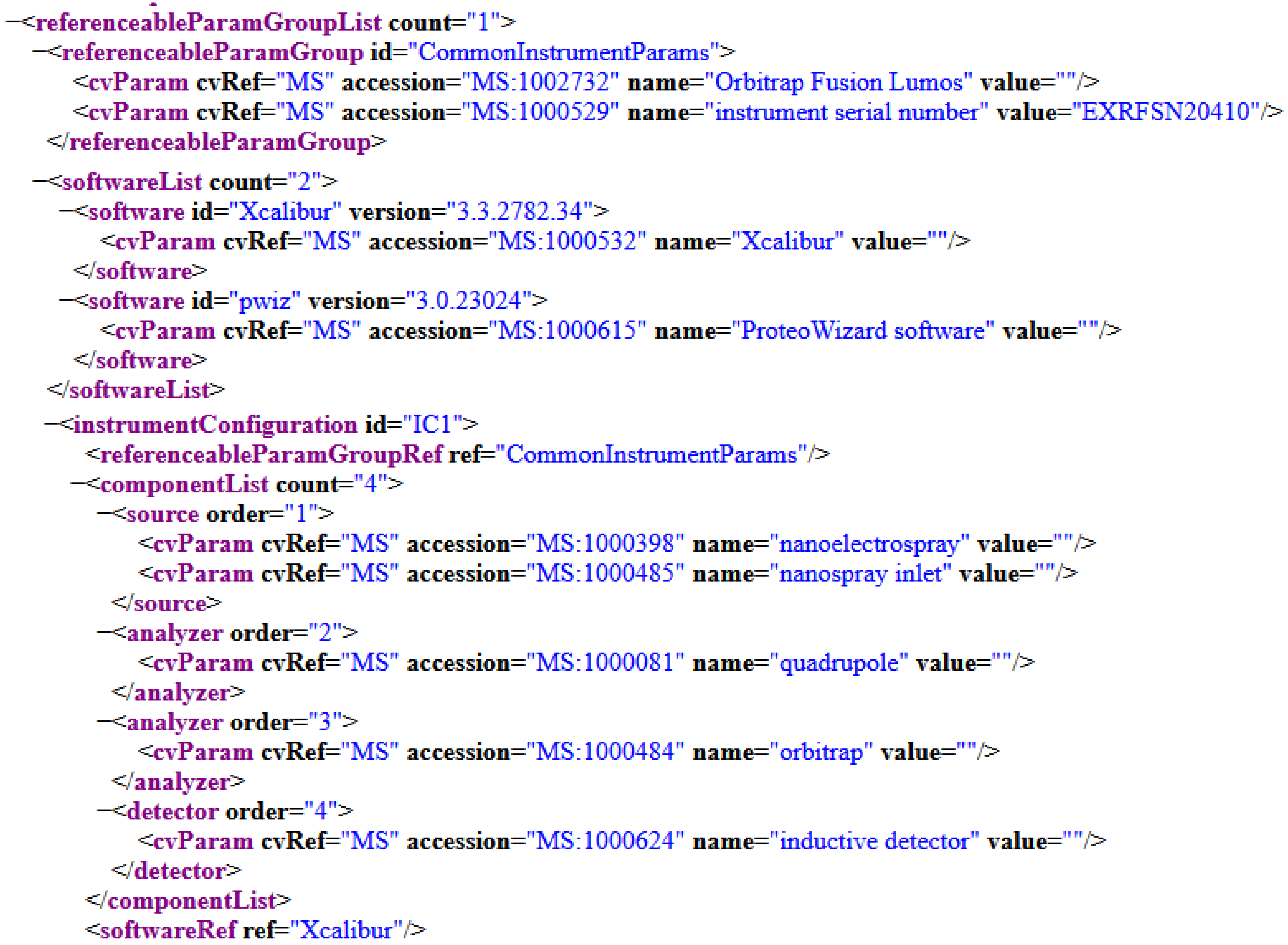
    https://www.matrixscience.com/help/data_file_help.html

Some advantages of mzML

- Metadata! Info about acquisition and processing of the spectra.
- Unique spectrum identifiers – enables tracking of an identification back to the raw data
- Spectrum settings like isolation windows etc.
- Access files on any platform. Several general libraries for accessing XML available in different programming languages

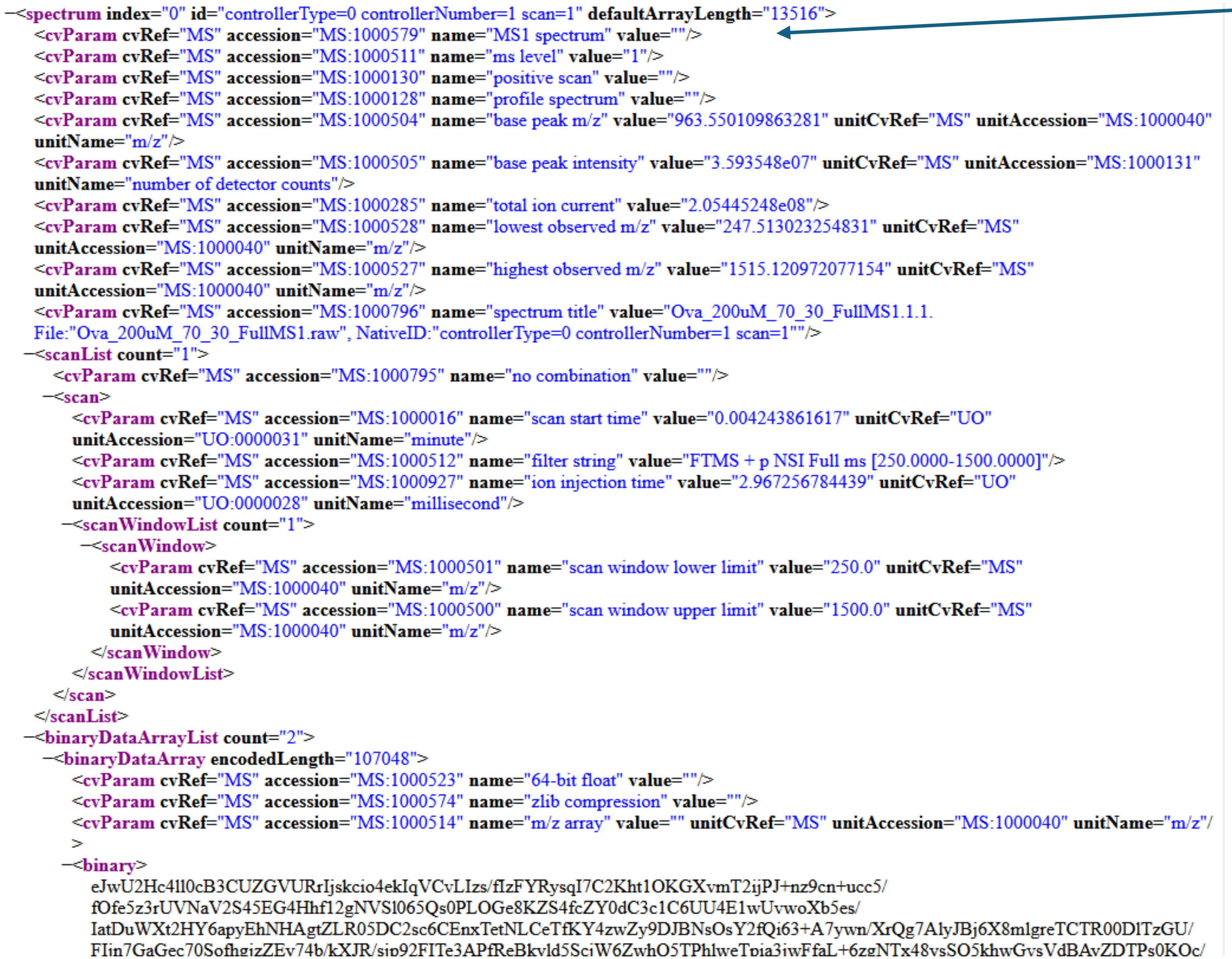
    - Limitations due to very large files

mzML extracts



```
-<referenceableParamGroupList count="1">
  -<referenceableParamGroup id="CommonInstrumentParams">
     <cvParam cvRef="MS" accession="MS:1002732" name="Orbitrap Fusion Lumos" value=""/>
     <cvParam cvRef="MS" accession="MS:1000529" name="instrument serial number" value="EXRFSN20410"/>
   </referenceableParamGroup>
  -<softwareList count="2">
    -<software id="Xcalibur" version="3.3.2782.34">
       <cvParam cvRef="MS" accession="MS:1000532" name="Xcalibur" value=""/>
     </software>
    -<software id="pwiz" version="3.0.23024">
       <cvParam cvRef="MS" accession="MS:1000615" name="ProteoWizard software" value=""/>
     </software>
   </softwareList>
   -<instrumentConfiguration id="IC1">
      <referenceableParamGroupRef ref="CommonInstrumentParams"/>
     -<componentList count="4">
       -<source order="1">
          <cvParam cvRef="MS" accession="MS:1000398" name="nanoelectrospray" value=""/>
          <cvParam cvRef="MS" accession="MS:1000485" name="nanospray inlet" value=""/>
        </source>
       -<analyzer order="2">
          <cvParam cvRef="MS" accession="MS:1000081" name="quadrupole" value=""/>
        </analyzer>
       -<analyzer order="3">
          <cvParam cvRef="MS" accession="MS:1000484" name="orbitrap" value=""/>
        </analyzer>
       -<detector order="4">
          <cvParam cvRef="MS" accession="MS:1000624" name="inductive detector" value=""/>
        </detector>
      </componentList>
      <softwareRef ref="Xcalibur"/>
```

mzML continued



```
−<spectrum index="0" id="controllerType=0 controllerNumber=1 scan=1" defaultArrayLength="13516">
   <cvParam cvRef="MS" accession="MS:1000579" name="MS1 spectrum" value=""/>
   <cvParam cvRef="MS" accession="MS:1000511" name="ms level" value="1"/>
   <cvParam cvRef="MS" accession="MS:1000130" name="positive scan" value=""/>
   <cvParam cvRef="MS" accession="MS:1000128" name="profile spectrum" value=""/>
   <cvParam cvRef="MS" accession="MS:1000504" name="base peak m/z" value="963.550109863281" unitCvRef="MS" unitAccession="MS:1000040"
   unitName="m/z"/>
   <cvParam cvRef="MS" accession="MS:1000505" name="base peak intensity" value="3.593548e07" unitCvRef="MS" unitAccession="MS:1000131"
   unitName="number of detector counts"/>
   <cvParam cvRef="MS" accession="MS:1000285" name="total ion current" value="2.05445248e08"/>
   <cvParam cvRef="MS" accession="MS:1000528" name="lowest observed m/z" value="247.513023254831" unitCvRef="MS"
   unitAccession="MS:1000040" unitName="m/z"/>
   <cvParam cvRef="MS" accession="MS:1000527" name="highest observed m/z" value="1515.120972077154" unitCvRef="MS"
   unitAccession="MS:1000040" unitName="m/z"/>
   <cvParam cvRef="MS" accession="MS:1000796" name="spectrum title" value="Ova_200uM_70_30_FullMS1.1.1.
   File:"Ova_200uM_70_30_FullMS1.raw", NativeID:"controllerType=0 controllerNumber=1 scan=1""/>
 −<scanList count="1">
     <cvParam cvRef="MS" accession="MS:1000795" name="no combination" value=""/>
   −<scan>
       <cvParam cvRef="MS" accession="MS:1000016" name="scan start time" value="0.004243861617" unitCvRef="UO"
       unitAccession="UO:0000031" unitName="minute"/>
       <cvParam cvRef="MS" accession="MS:1000512" name="filter string" value="FTMS + p NSI Full ms [250.0000-1500.0000]"/>
       <cvParam cvRef="MS" accession="MS:1000927" name="ion injection time" value="2.967256784439" unitCvRef="UO"
       unitAccession="UO:0000028" unitName="millisecond"/>
     −<scanWindowList count="1">
       −<scanWindow>
           <cvParam cvRef="MS" accession="MS:1000501" name="scan window lower limit" value="250.0" unitCvRef="MS"
           unitAccession="MS:1000040" unitName="m/z"/>
           <cvParam cvRef="MS" accession="MS:1000500" name="scan window upper limit" value="1500.0" unitCvRef="MS"
           unitAccession="MS:1000040" unitName="m/z"/>
         </scanWindow>
       </scanWindowList>
     </scan>
   </scanList>
 −<binaryDataArrayList count="2">
   −<binaryDataArray encodedLength="107048">
       <cvParam cvRef="MS" accession="MS:1000523" name="64-bit float" value=""/>
       <cvParam cvRef="MS" accession="MS:1000574" name="zlib compression" value=""/>
       <cvParam cvRef="MS" accession="MS:1000514" name="m/z array" value="" unitCvRef="MS" unitAccession="MS:1000040" unitName="m/z"/
       >
     −<binary>
         eJwU2Hc4ll0cB3CUZGVURrIjskcio4ekIqVCvLIzs/fIzFYRysqI7C2Kht1OKGXvmT2ijPJ+nz9cn+ucc5/
         fOfe5z3rUVNaV2S45EG4Hhf12gNVSl065Qs0PLOGe8KZS4fcZY0dC3c1C6UU4E1wUvwoXb5es/
         IatDuWXt2HY6apyEhNHAgtZLR05DC2sc6CEnxTetNLCeTfKY4zwZy9DJBNsOsY2fQi63+A7ywn/XrQg7AlyJBj6X8mlgreTCTR00DlTzGU/
         FIin7GaGec70SofhgizZEv74b/kXJR/sip92FITe3APfReBkvld5SciW6ZwhO5TPhlweTpia3iwFfaL+6zgNTx48vsSO5khwGvsVdBAvZDTPs0KOc/
```

Note that spectra will appear in the File in the order of acquistion.

For a DDA file this could be:

MS1
MS/MS of precursor 1
MS/MS of precursor 2
....
MS/MS of precursor N
MS1
MS/MS etc

mzML cons

- Verbose format with complex structure
- Spectra are encoded so not readable without decoding
- Large files (GBs). Compression helps to reduce file sizes, though.

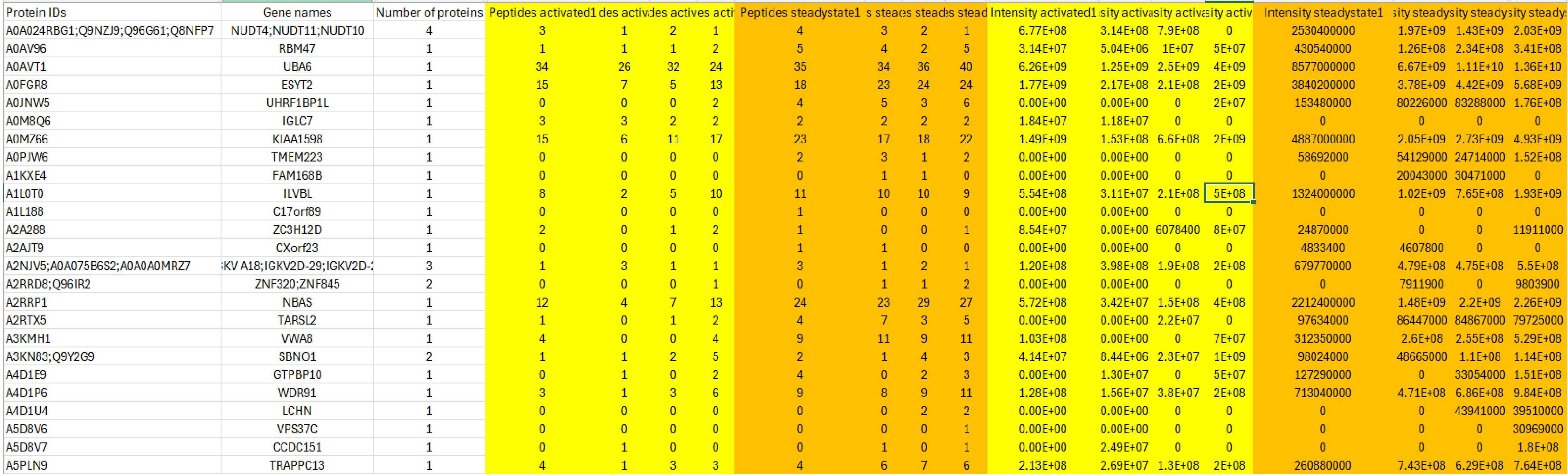
Results data formats

- Typically text tables with tab-separated data
  - Separate PSM-, peptide- and protein-level tables.
  - Can be imported to Excel, R, etc.
  - Usually each row represents a protein group / peptide / PSM
  - Sample scores and abundance values in columns
- Search engines may also report XML-based formats
- Standard formats mzIdentML and mzTab has not received widespread usage

  (https://psidev.info)

Example output from MaxQuant software

- Protein groups file (proteinGroups.txt)
- Comparison of 4 vs 4 samples (activated yellow, steady state orange)
- Note: Numbers of identified peptides differ between replicates, as do intensity values.
- Multiple accession numbers on some lines (protein groups).



Protein IDs	Gene names	Number of proteins	Peptides activated1	des activ	des active	s acti	Peptides steadystate1	s steac	s stead	s stead	Intensity activated1	sity activa	sity activa	sity activ	Intensity steadystate1	sity steady	sity steady	sity steady
A0A024RBG1;Q9NZJ9;Q96G61;Q8NFP7	NUDT4;NUDT11;NUDT10	4	3	1	2	1	4	3	2	1	6.77E+08	3.14E+08	7.9E+08	0	2530400000	1.97E+09	1.43E+09	2.03E+09
A0AV96	RBM47	1	1	1	1	2	5	4	2	5	3.14E+07	5.04E+06	1E+07	5E+07	430540000	1.26E+08	2.34E+08	3.41E+08
A0AVT1	UBA6	1	34	26	32	24	35	34	36	40	6.26E+09	1.25E+09	2.5E+09	4E+09	8577000000	6.67E+09	1.11E+10	1.36E+10
A0FGR8	ESYT2	1	15	7	5	13	18	23	24	24	1.77E+09	2.17E+08	2.1E+08	2E+09	3840200000	3.78E+09	4.42E+09	5.68E+09
A0JNW5	UHRF1BP1L	1	0	0	0	2	4	5	3	6	0.00E+00	0.00E+00	0	2E+07	153480000	80226000	83288000	1.76E+08
A0M8Q6	IGLC7	1	3	3	2	2	2	2	2	2	1.84E+07	1.18E+07	0	0	0	0	0	0
A0MZ66	KIAA1598	1	15	6	11	17	23	17	18	22	1.49E+09	1.53E+08	6.6E+08	2E+09	4887000000	2.05E+09	2.73E+09	4.93E+09
A0PJW6	TMEM223	1	0	0	0	0	2	3	1	2	0.00E+00	0.00E+00	0	0	58692000	54129000	24714000	1.52E+08
A1KXE4	FAM168B	1	0	0	0	0	0	1	1	0	0.00E+00	0.00E+00	0	0	0	20043000	30471000	0
A1L0T0	ILVBL	1	8	2	5	10	11	10	10	9	5.54E+08	3.11E+07	2.1E+08	5E+08	1324000000	1.02E+09	7.65E+08	1.93E+09
A1L188	C17orf89	1	0	0	0	0	1	0	0	0	0.00E+00	0.00E+00	0	0	0	0	0	0
A2A288	ZC3H12D	1	2	0	1	2	1	0	0	1	8.54E+07	0.00E+00	6078400	8E+07	24870000	0	0	11911000
A2AJT9	CXorf23	1	0	0	0	0	1	1	0	0	0.00E+00	0.00E+00	0	0	4833400	4607800	0	0
A2NJV5;A0A075B6S2;A0A0A0MRZ7	KV A18;IGKV2D-29;IGKV2D-	3	1	3	1	1	3	1	2	1	1.20E+08	3.98E+08	1.9E+08	2E+08	679770000	4.79E+08	4.75E+08	5.5E+08
A2RRD8;Q96IR2	ZNF320;ZNF845	2	0	0	0	1	0	1	1	2	0.00E+00	0.00E+00	0	0	0	7911900	0	9803900
A2RRP1	NBAS	1	12	4	7	13	24	23	29	27	5.72E+08	3.42E+07	1.5E+08	4E+08	2212400000	1.48E+09	2.2E+09	2.26E+09
A2RTX5	TARSL2	1	1	0	1	2	4	7	3	5	0.00E+00	0.00E+00	2.2E+07	0	97634000	86447000	84867000	79725000
A3KMH1	VWA8	1	4	0	0	4	9	11	9	11	1.03E+08	0.00E+00	0	7E+07	312350000	2.6E+08	2.55E+08	5.29E+08
A3KN83;Q9Y2G9	SBNO1	2	1	1	2	5	2	1	4	3	4.14E+07	8.44E+06	2.3E+07	1E+09	98024000	48665000	1.1E+08	1.14E+08
A4D1E9	GTPBP10	1	0	1	0	2	4	0	2	3	0.00E+00	1.30E+07	0	5E+07	127290000	0	33054000	1.51E+08
A4D1P6	WDR91	1	3	1	3	6	9	8	9	11	1.28E+08	1.56E+07	3.8E+07	2E+08	713040000	4.71E+08	6.86E+08	9.84E+08
A4D1U4	LCHN	1	0	0	0	0	0	0	2	2	0.00E+00	0.00E+00	0	0	0	0	43941000	39510000
A5D8V6	VPS37C	1	0	0	0	0	0	0	0	1	0.00E+00	0.00E+00	0	0	0	0	0	30969000
A5D8V7	CCDC151	1	0	1	0	0	0	1	0	1	0.00E+00	2.49E+07	0	0	0	0	0	1.8E+08
A5PLN9	TRAPPC13	1	4	1	3	3	4	6	7	6	2.13E+08	2.69E+07	1.3E+08	2E+08	260880000	7.43E+08	6.29E+08	7.64E+08

Normalisation

- Compensate for differences in sample processing, sample loading and ionisation efficiencies.
- After normalisation it should be easier to detect true biological differencs.

-> Little variation between technical replicates

Normalisation: assumptions

- Assume most proteins have similar levels in the different samples groups
- Total signal or median signal can be used to calculate normalisation factor and then scale protein abundance values

Protein	Sample 1	Sample 2
A	10	12
B	20	24
C	15	18
D	50	60
Total	95	114

Example

- Different amounts of same sample analysed:

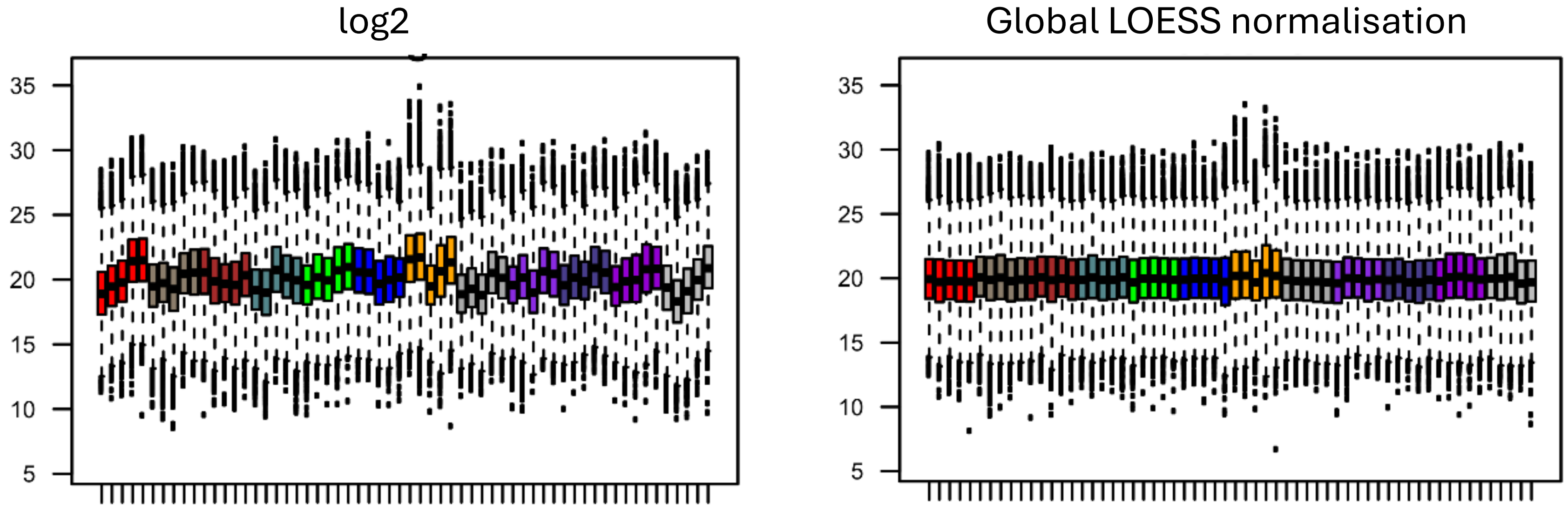
Protein	Sample 1	Sample 2
A	10	12
B	20	24
C	15	18
D	50	60
Total	95	114

Total intensity normalisation factors
Sample 1: 1.10
Sample 2: 0.92

Protein	Sample 1	Sample 2
A	11	11
B	22	22
C	16.5	16.5
D	55	55
Total	104.5	104.5

Generic problems for quantitative omics data:

Between sample normalization , batch effects removal

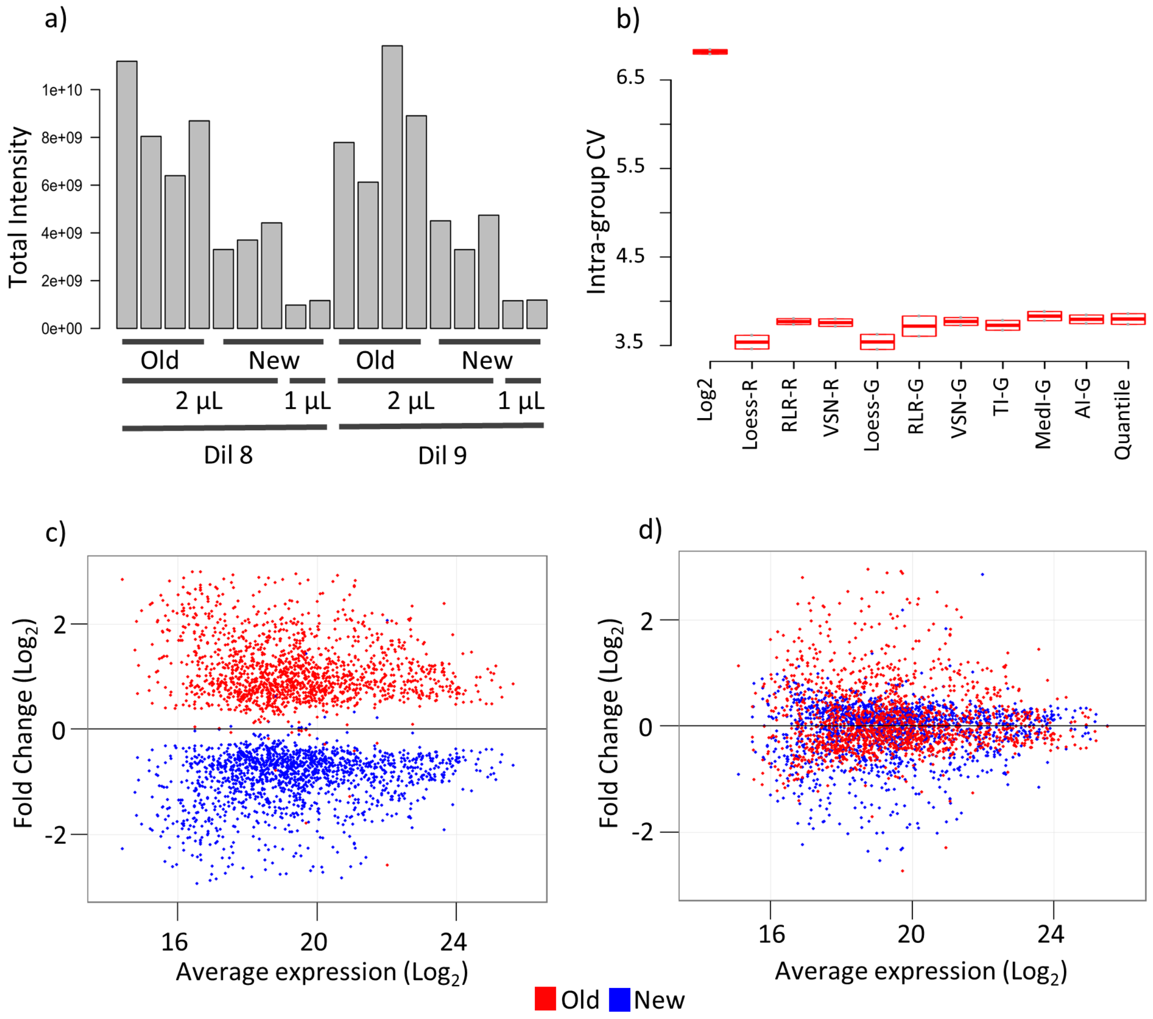


X axis represent different samples and Y axis the distribution of protein abundance values. Different colors represent different sample groups

Many normalisation methods exists. Median normalisation frequent, but more advanced like LOESS can be useful

Note: Regardless of whether normalisation is used or not the data should be log2 transformed before statistical examinations

Example: Data acquired years apart



Without normalization:
No significant changes found

With normalization:
66 true positives

(only 57 when using
old data only)

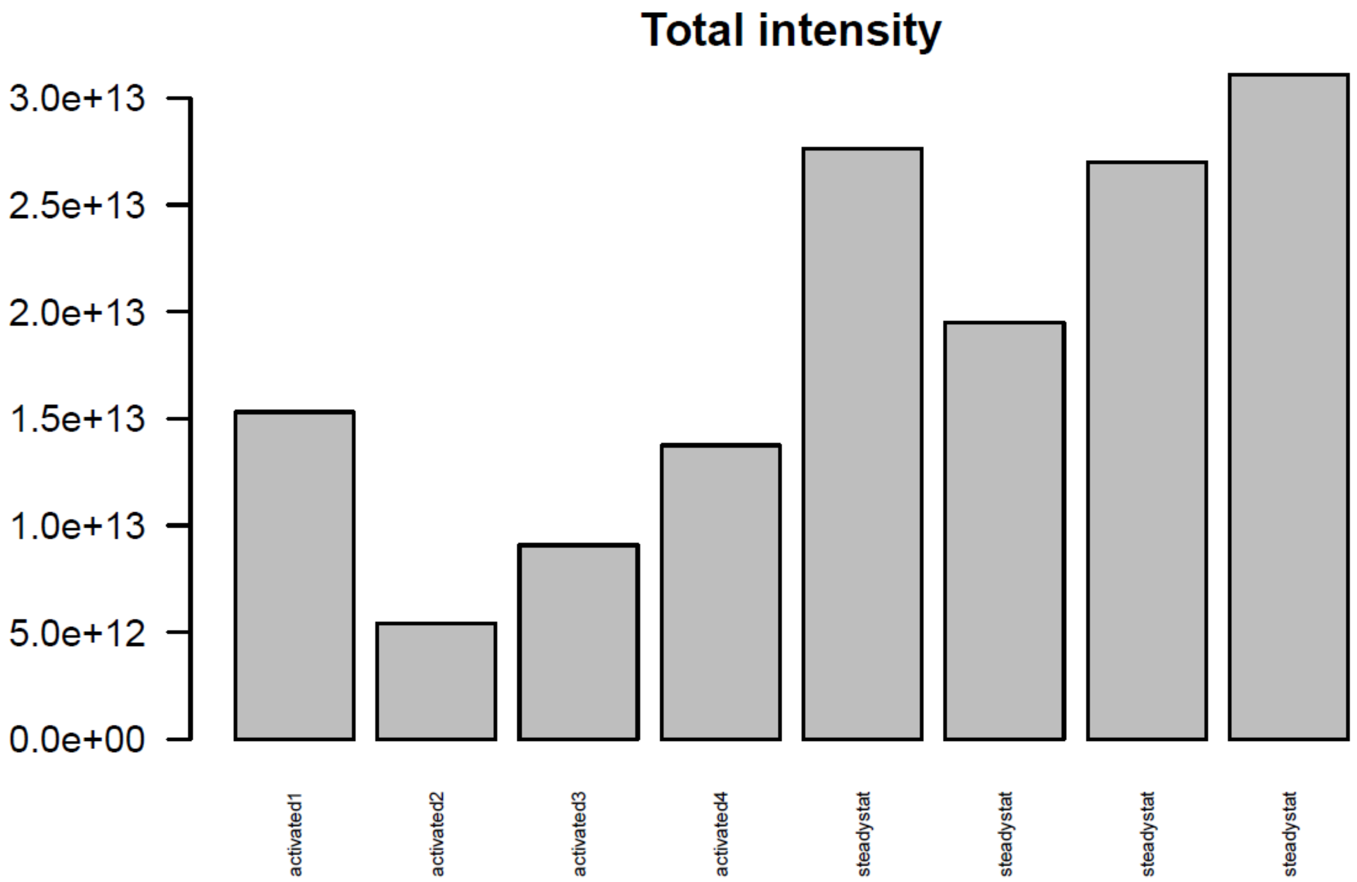
Sandin et al *Proteomics Clin Appl* 2015

Quality control of data

- Data distribution of different samples
- Total intensities of different samples
- Missing values in different samples
- Overview after dimensionality reduction:
  - PCA / MDS plots
  - Hierachical clustering / dendograms

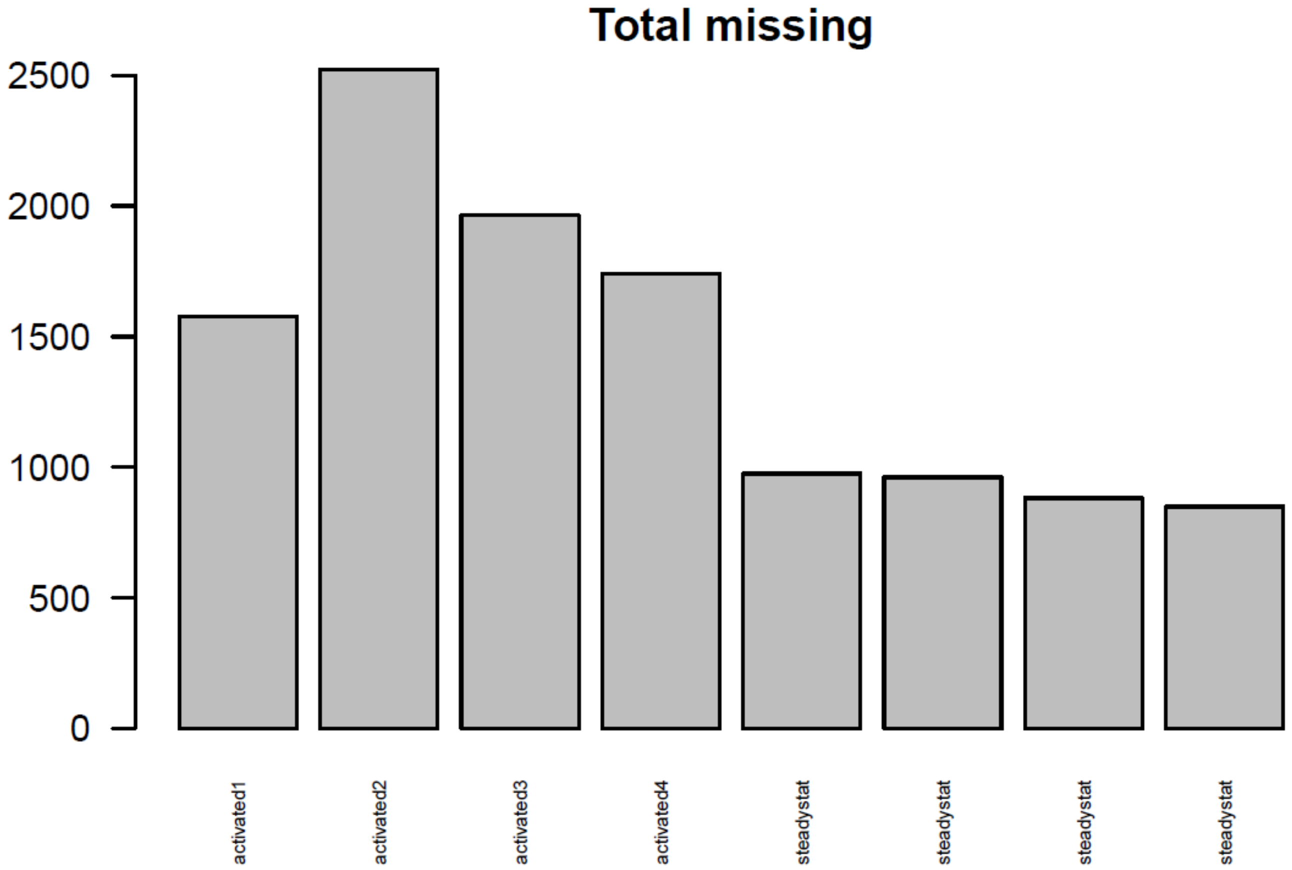
Outlier samples can be removed if thought to be outliers based on technical errors.

Total abundance of each sample



Normalisation likely needed!

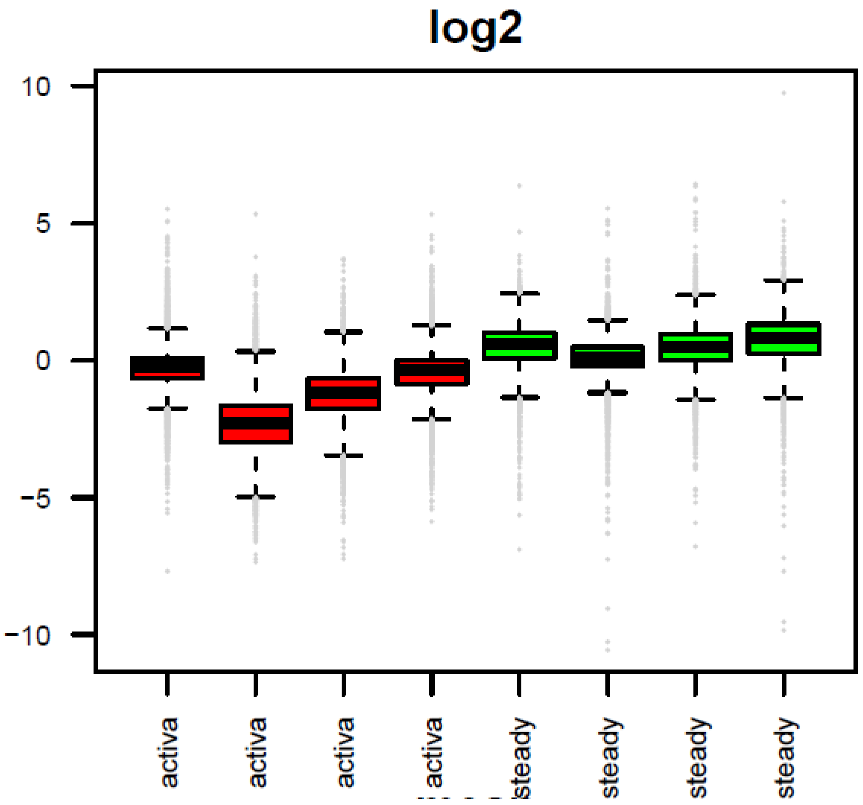
Missing values



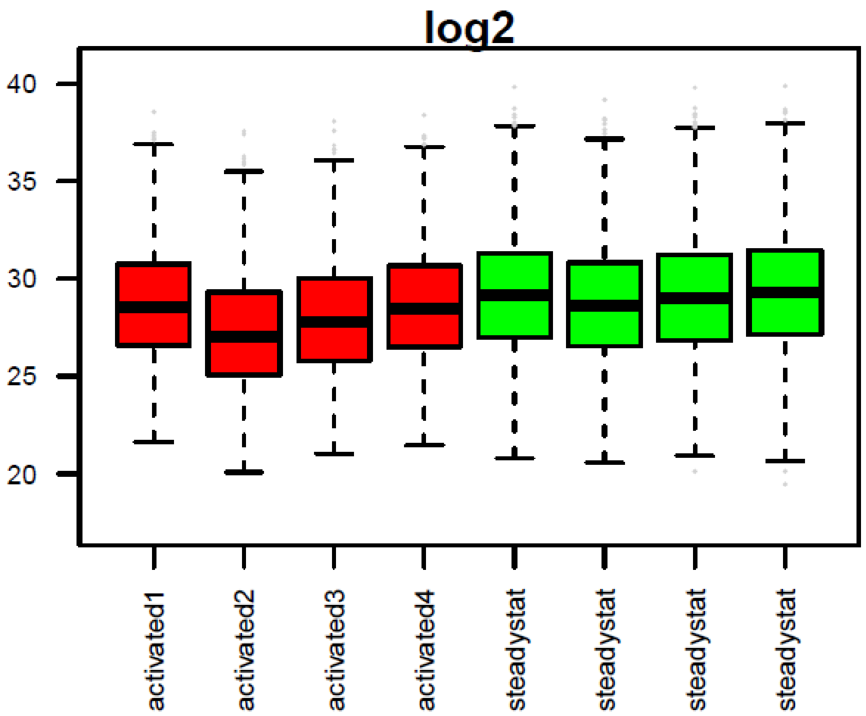
- Missing values differ between the groups (activated vs steady state).
- Is this biological or an artefact?

Data distribution

Relative log expression (RLE plot)
Showing abundance compare to median for all proteins



Box plot – abundance values for all detected proteins (Log2)



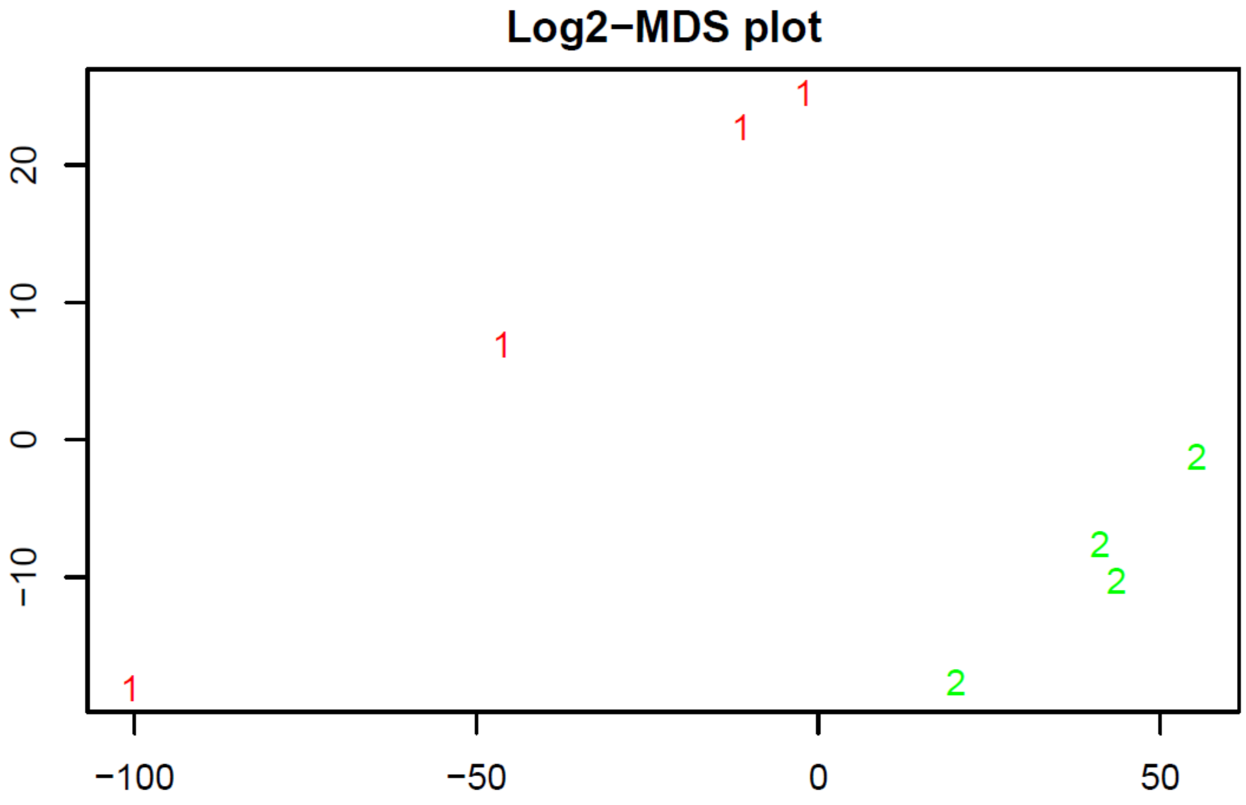
https://doi.org/10.1371/journal.pone.0191629

Clustering

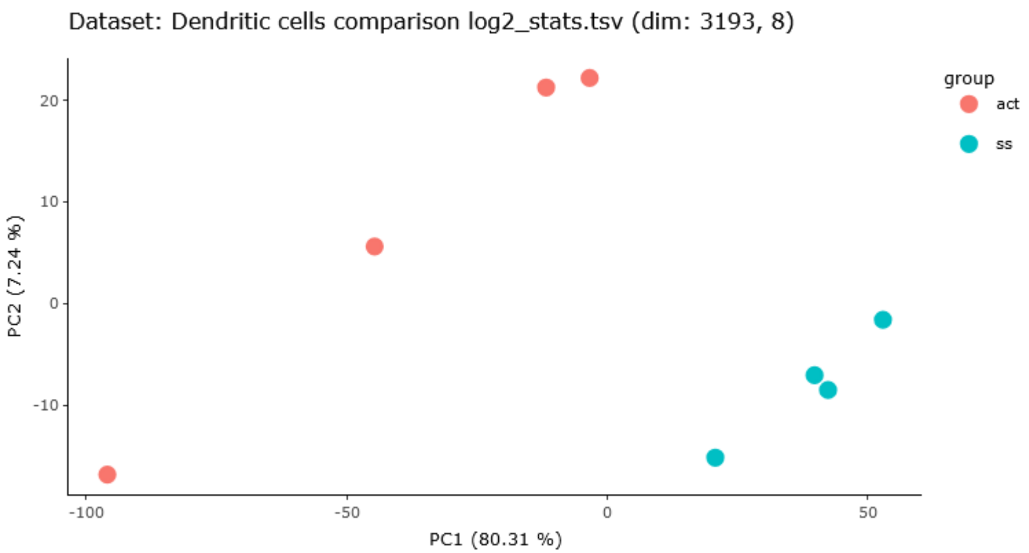
- PCA / MDS
- Hierarchical clustering

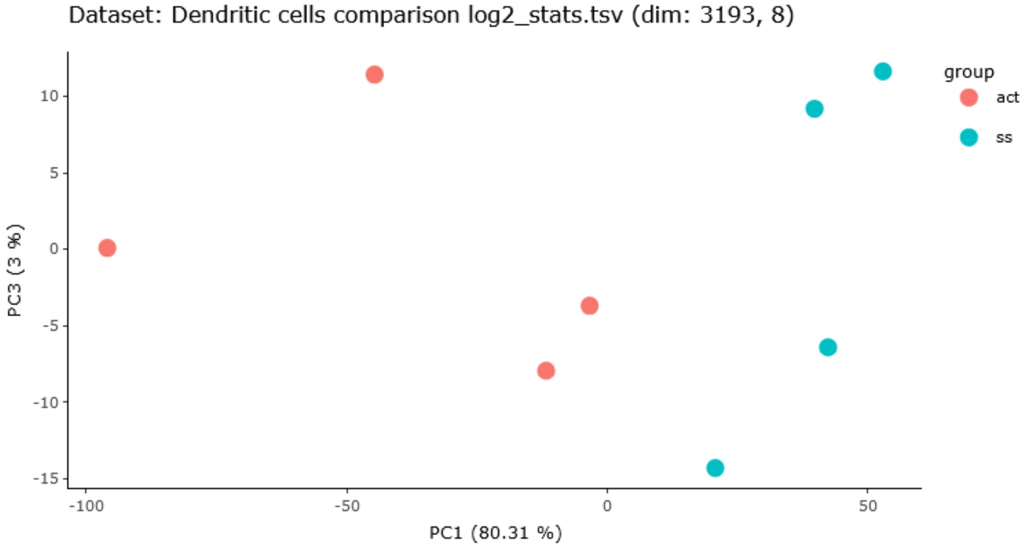
MDS (Multidimensional scaling)

- Plot taking into account sample similarities based on all measured variables
- Allows for identification of outlier samples and unexpected patterns



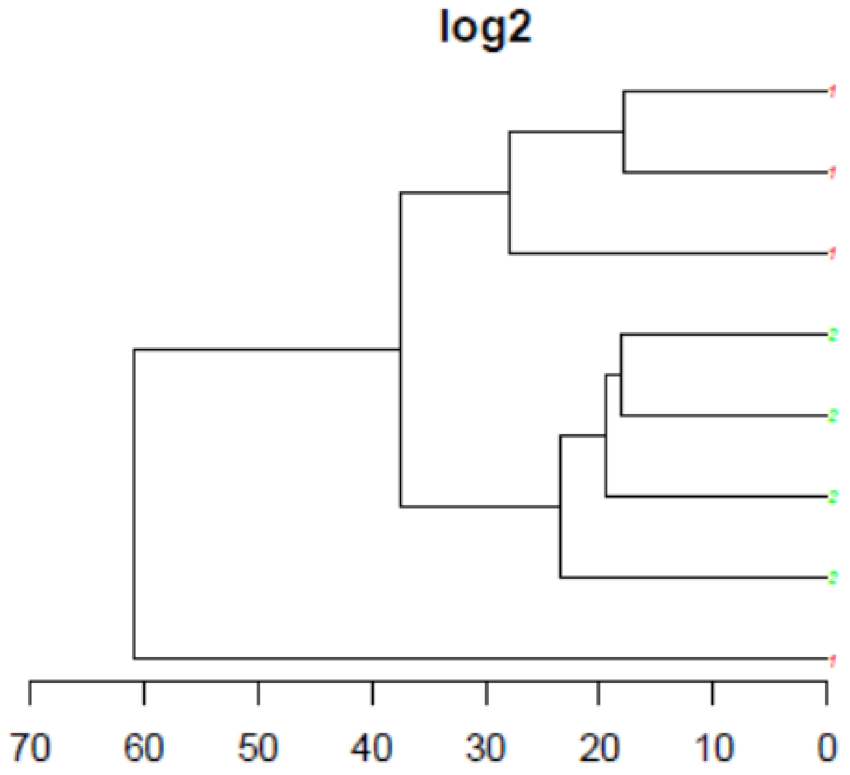
PCA (Principal Component Analysis)


Dataset: Dendritic cells comparison log2_stats.tsv (dim: 3193, 8)


Dataset: Dendritic cells comparison log2_stats.tsv (dim: 3193, 8)

PCA explained: https://doi.org/10.1038/nbt0308-303

Unsupervised clustering



Based on variables without missing values

Finding differences between groups

- How do we know if a protein differs between two groups of samples
- T-Test. Assumptions: Data normally distributed

  -> Log2 transformation of data to achieve this.

Multiple hypotesis correction

- P-value<0.05 (significant) in 1 out of 20 comparisons by random.
    - If calculating p-values for 5000 protein comparisons we can expect 250 significant results at p<0.05 by chance!
- Correct the p-value for the number of tests done and work with false discovery rates (FDRs)
    - Benjamini-Hochberg method used most frequently

      https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

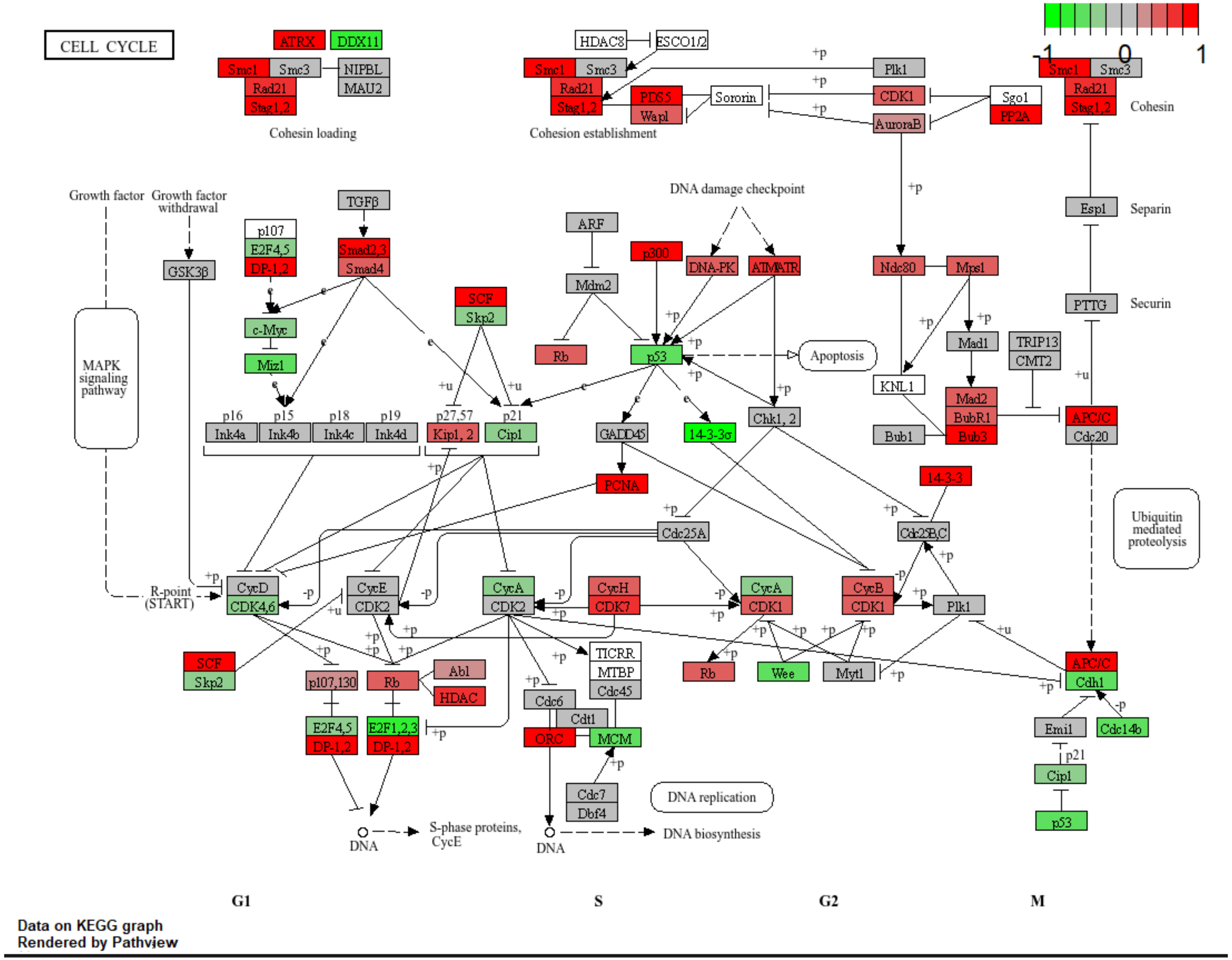
Functional analysis

- Mapping to biological pathways
- Pathway enrichment analysis
  - Are differentially abundant proteins mapping to specific pathways? Need cutoff
  - https://biostatsquid.com/pathway-enrichment-analysis-explained/
- Gene set enrichment analysis
  - Rank all proteins according to difference between groups. No significane cutoff.
  - https://biostatsquid.com/gene-set-enrichment-analysis/
- Network analysis

  https://string-db.org/
- Pathway analysis

  More sophisticated analyses including regulation and pathway fluxes etc.

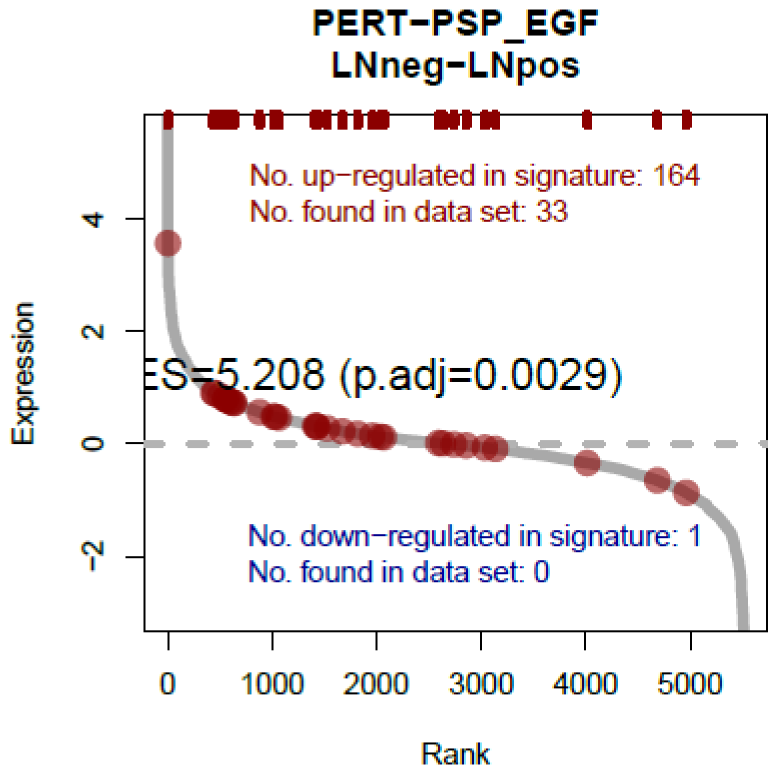
Mapping expression to KEGG pathway



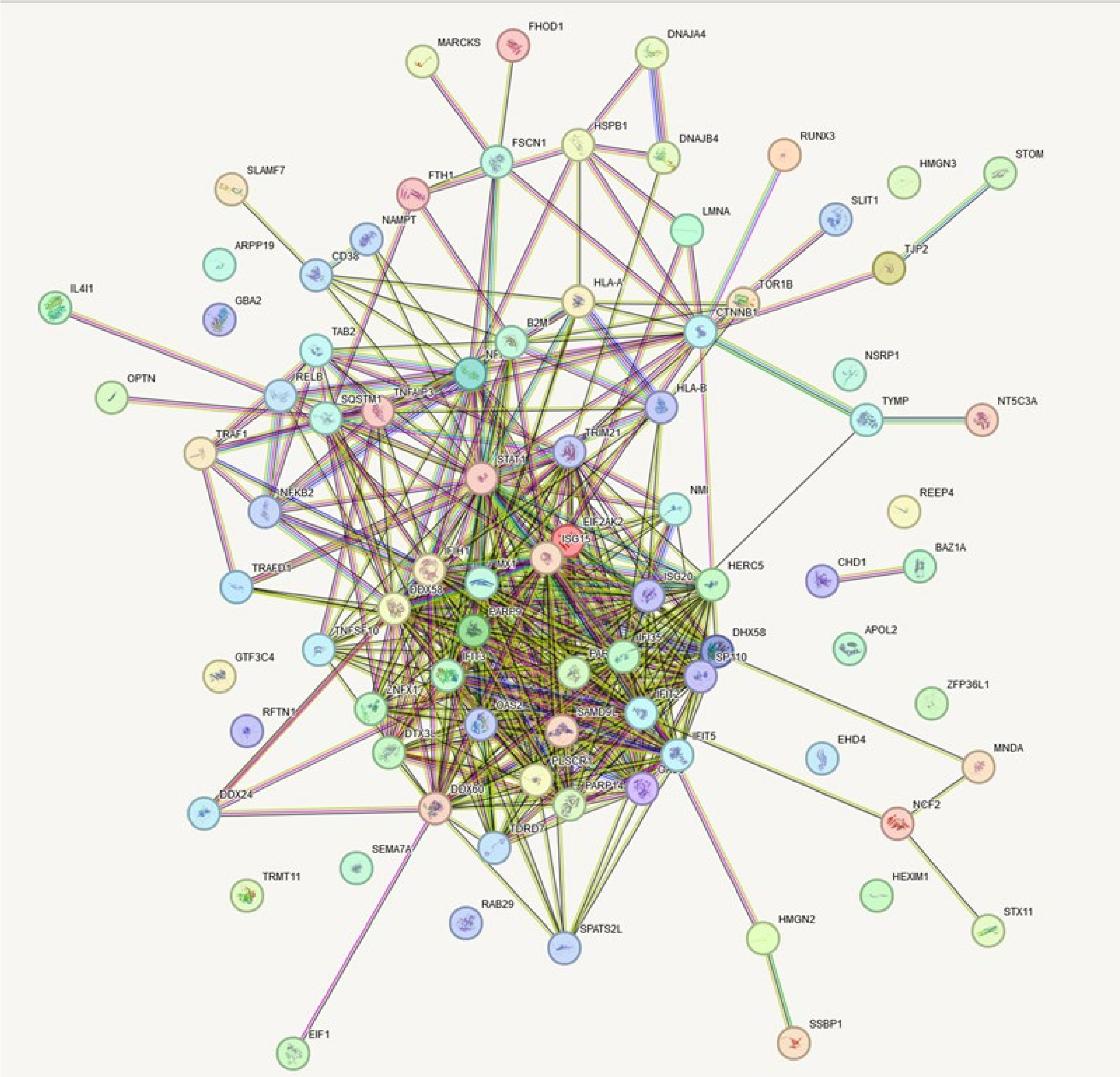
Example pathway analysed using GSEA

Genes in the gene set are overrepresented in
The beginning of the ranked list

Normalised Enrichment Score 5.2
Adjusted p-value 0.0029



STRING network - how are differentially abundant proteins connected?

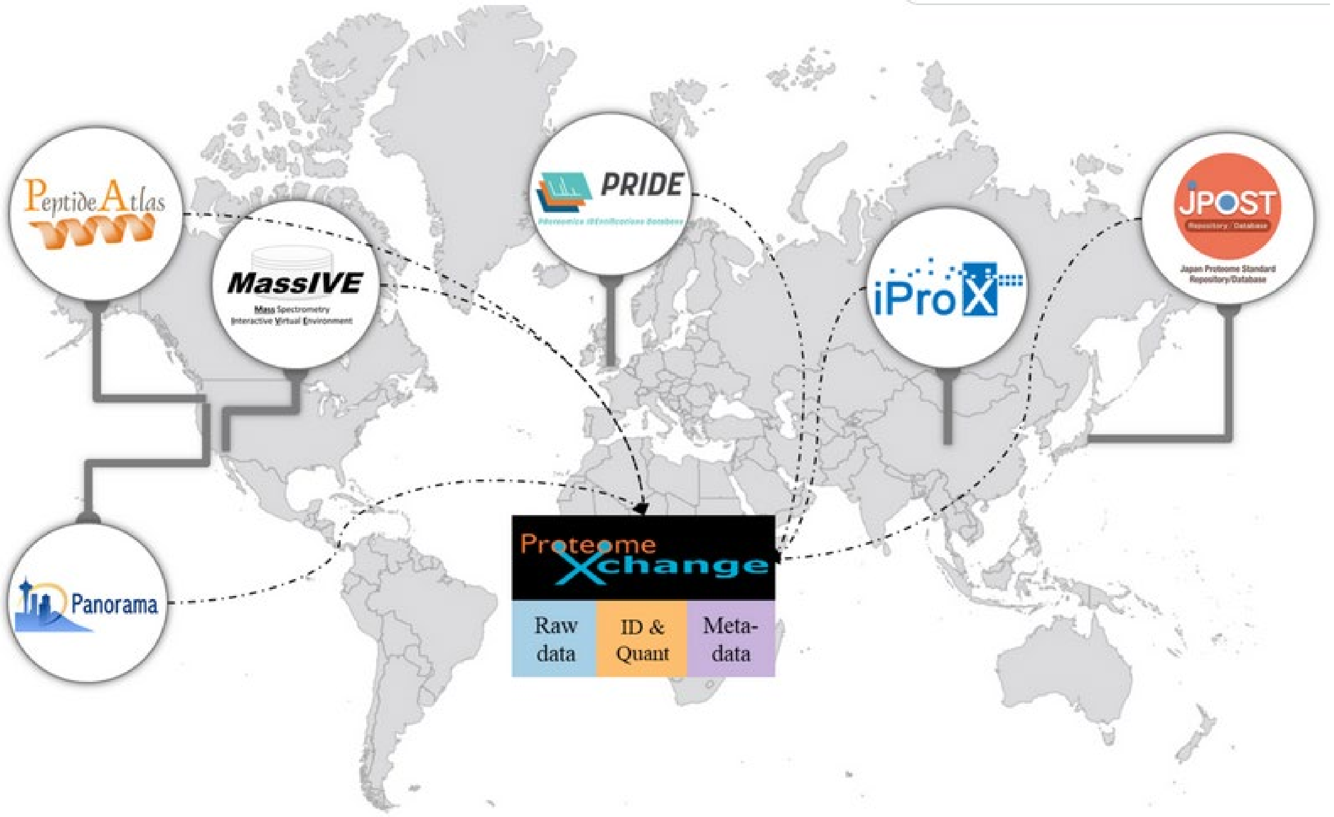


Pathway level analysis - caution

- Note that pathways and gene sets can be defined in different ways
- Depending on database annotations. Usually gene level annotations, even if studying proteins
- Experimental evidence can be of different quality
- Possible to map between species, but may be incorrect
- Protein interactions can also be defined in many different ways

Repositories with proteomics data

Possible to download proteomics data acquired by other researchers and reanalyse.
Deposition of research data at repositories before publishing.



Mission

The ProteomeXchange Consortium was established to provide globally coordinated standard data submission and dissemination pipelines involving the main proteomics repositories, and to encourage open data policies in the field.

Source: proteomexchange.org

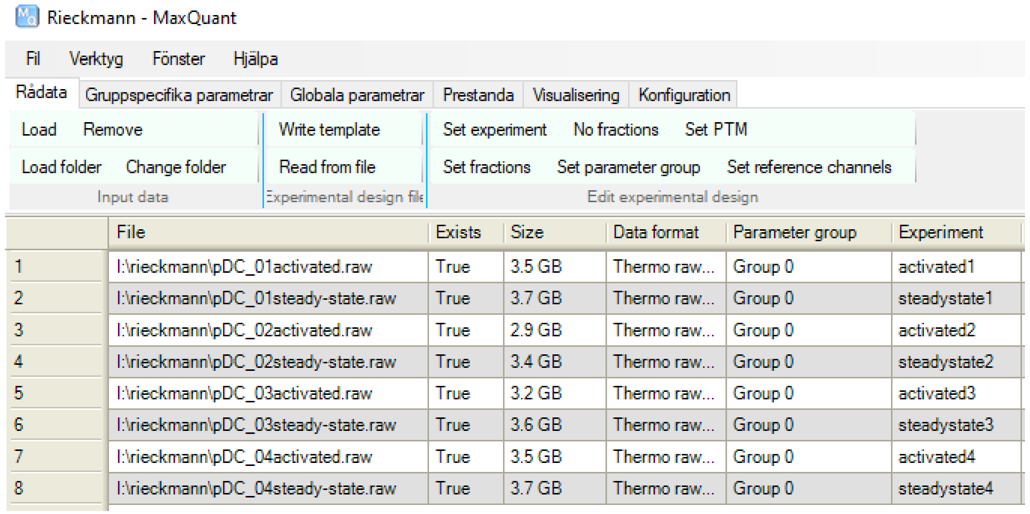
Example of data analysis

- Comparison of activated vs plasma-derived dendritic cells (pDCs)
- Dataset downloaded from ProteomeXchange (PXD004352) https://doi.org/10.1038/ni.3693
- Cells had been activated using LPS and R848.
- Four replicates of each
- Data acquired using LC-MS/MS (nano LC with 50 cm column, and Q Exactive HF mass spectrometer)

Processing through MaxQuant software - settings

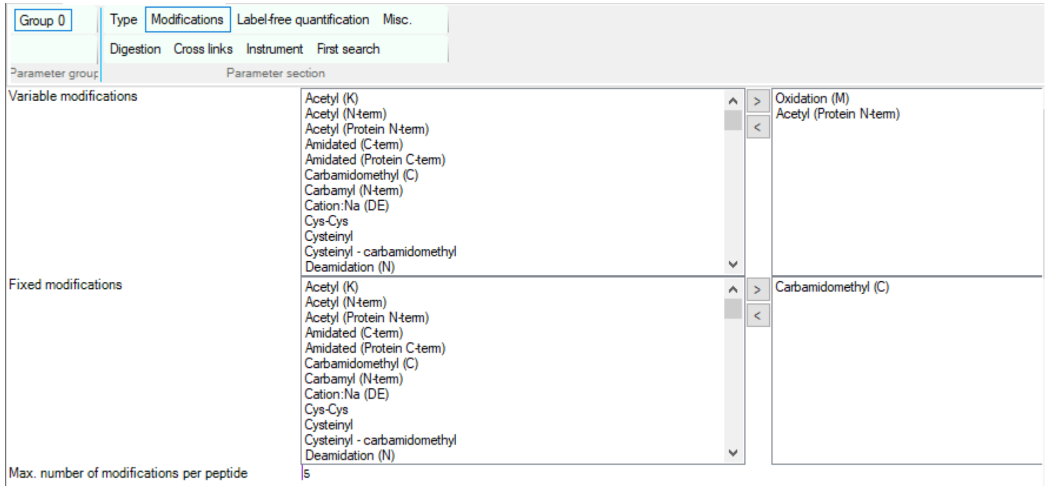
- Matching with UniProt human proteins (reviewed section = SwissProt)
- Variable modifications: Methionine oxidation, protein N-terminal acetylation
- Fixed modifications: carbamidomethylation of cysteins (from reduction and alkylation process)
- Trypsin digestion
- LFQ (Label free quantification) without match between runs
- 1% peptide level and protein level FDRs

Settings in graphical user interface for important parameters



	File	Exists	Size	Data format	Parameter group	Experiment
1	I:\rieckmann\pDC_01activated.raw	True	3.5 GB	Thermo raw...	Group 0	activated1
2	I:\rieckmann\pDC_01steady-state.raw	True	3.7 GB	Thermo raw...	Group 0	steadystate1
3	I:\rieckmann\pDC_02activated.raw	True	2.9 GB	Thermo raw...	Group 0	activated2
4	I:\rieckmann\pDC_02steady-state.raw	True	3.4 GB	Thermo raw...	Group 0	steadystate2
5	I:\rieckmann\pDC_03activated.raw	True	3.2 GB	Thermo raw...	Group 0	activated3
6	I:\rieckmann\pDC_03steady-state.raw	True	3.6 GB	Thermo raw...	Group 0	steadystate3
7	I:\rieckmann\pDC_04activated.raw	True	3.5 GB	Thermo raw...	Group 0	activated4
8	I:\rieckmann\pDC_04steady-state.raw	True	3.7 GB	Thermo raw...	Group 0	steadystate4

Files and sample (experiment) assignments



Variable modifications

Acetyl (K)
Acetyl (N-term)
Acetyl (Protein N-term)
Amidated (C-term)
Amidated (Protein C-term)
Carbamidomethyl (C)
Carbamyl (N-term)
Cation:Na (DE)
Cys-Cys
Cysteinyl
Cysteinyl - carbamidomethyl
Deamidation (N)

Oxidation (M)
Acetyl (Protein N-term)

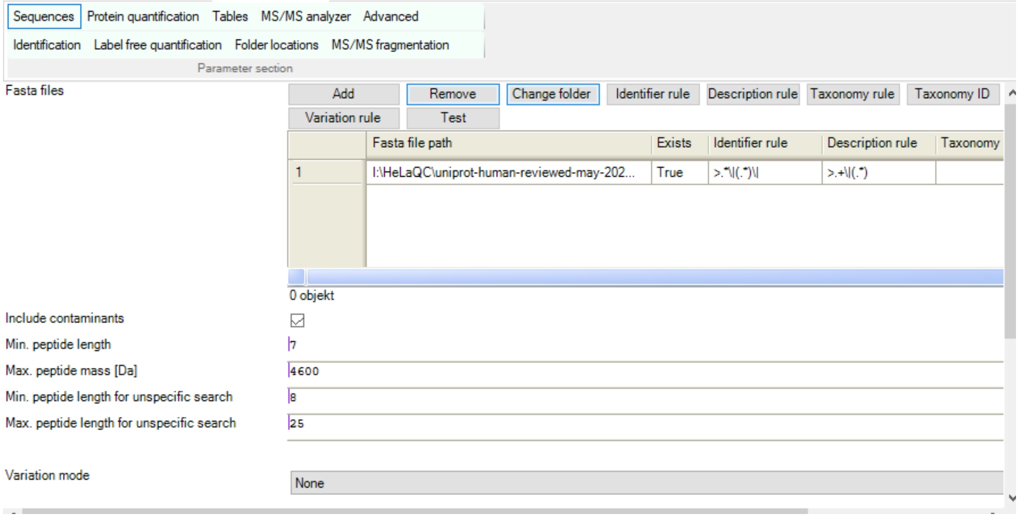
Fixed modifications

Acetyl (K)
Acetyl (N-term)
Acetyl (Protein N-term)
Amidated (C-term)
Amidated (Protein C-term)
Carbamidomethyl (C)
Carbamyl (N-term)
Cation:Na (DE)
Cys-Cys
Cysteinyl
Cysteinyl - carbamidomethyl
Deamidation (N)

Carbamidomethyl (C)

Max. number of modifications per peptide 5

Which amino acid modifications to consider



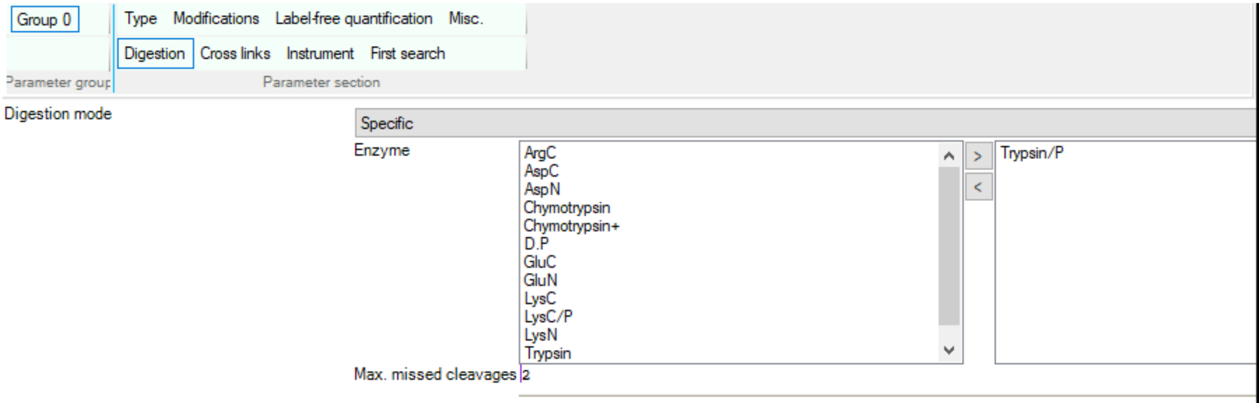
Fasta files

<table>
<tr><th></th><th>Fasta file path</th><th>Exists</th><th>Identifier rule</th><th>Description rule</th><th>Taxonomy</th></tr>
<tr><td>1</td><td>I:\HeLaQC\uniprot-human-reviewed-may-202...</td><td>True</td><td>>.*\|(.*)\|</td><td>>.+\|(.*)</td><td></td></tr>
</table>

0 objekt

Include contaminants	☑
Min. peptide length	7
Max. peptide mass [Da]	4600
Min. peptide length for unspecific search	8
Max. peptide length for unspecific search	25
Variation mode	None

Database to search.



Digestion mode Specific

Enzyme

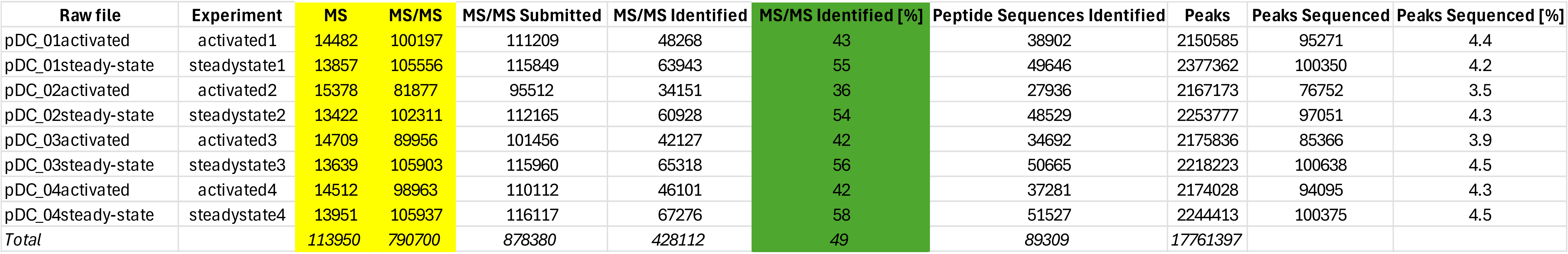
ArgC
AspC
AspN
Chymotrypsin
Chymotrypsin+
D.P
GluC
GluN
LysC
LysC/P
LysN
Trypsin

Trypsin/P

Max. missed cleavages 2

Which enzyme was used for digesting the proteins and max missed cleavages to consider

Results

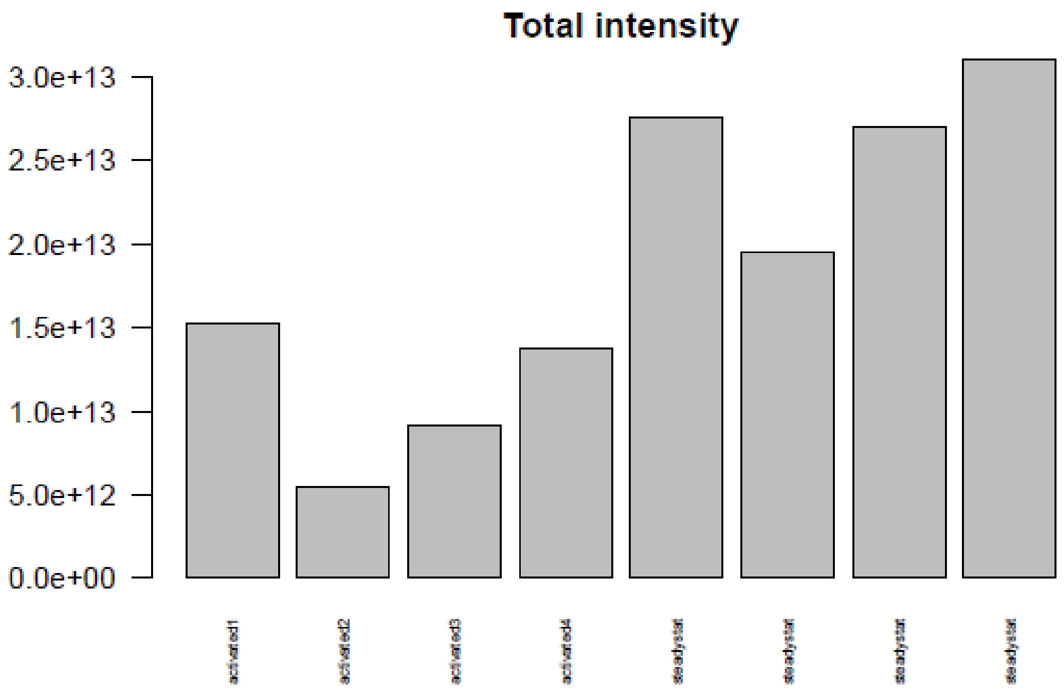


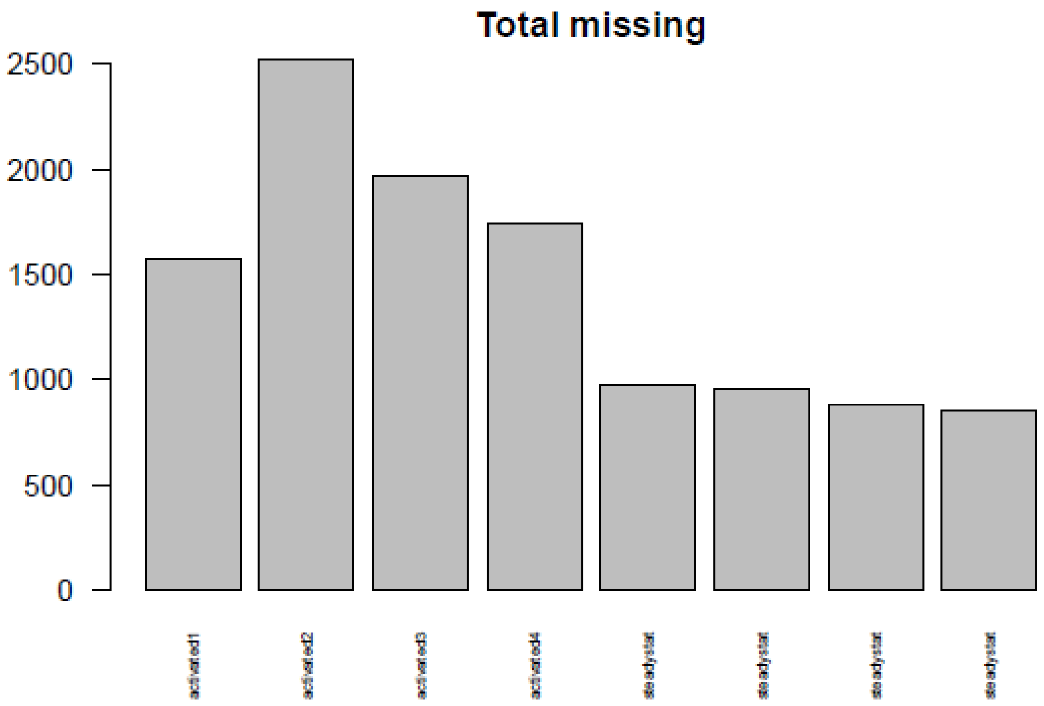
Raw file	Experiment	MS	MS/MS	MS/MS Submitted	MS/MS Identified	MS/MS Identified [%]	Peptide Sequences Identified	Peaks	Peaks Sequenced	Peaks Sequenced [%]
pDC_01activated	activated1	14482	100197	111209	48268	43	38902	2150585	95271	4.4
pDC_01steady-state	steadystate1	13857	105556	115849	63943	55	49646	2377362	100350	4.2
pDC_02activated	activated2	15378	81877	95512	34151	36	27936	2167173	76752	3.5
pDC_02steady-state	steadystate2	13422	102311	112165	60928	54	48529	2253777	97051	4.3
pDC_03activated	activated3	14709	89956	101456	42127	42	34692	2175836	85366	3.9
pDC_03steady-state	steadystate3	13639	105903	115960	65318	56	50665	2218223	100638	4.5
pDC_04activated	activated4	14512	98963	110112	46101	42	37281	2174028	94095	4.3
pDC_04steady-state	steadystate4	13951	105937	116117	67276	58	51527	2244413	100375	4.5
Total		*113950*	*790700*	*878380*	*428112*	*49*	*89309*	*17761397*		

Extract from summary.txt

Protein groups: 6879 rows in proteinGroups.txt table
84 decoy proteins (REV_) -> approximately 1% FDR
27 potential contaminant proteins (CON_)

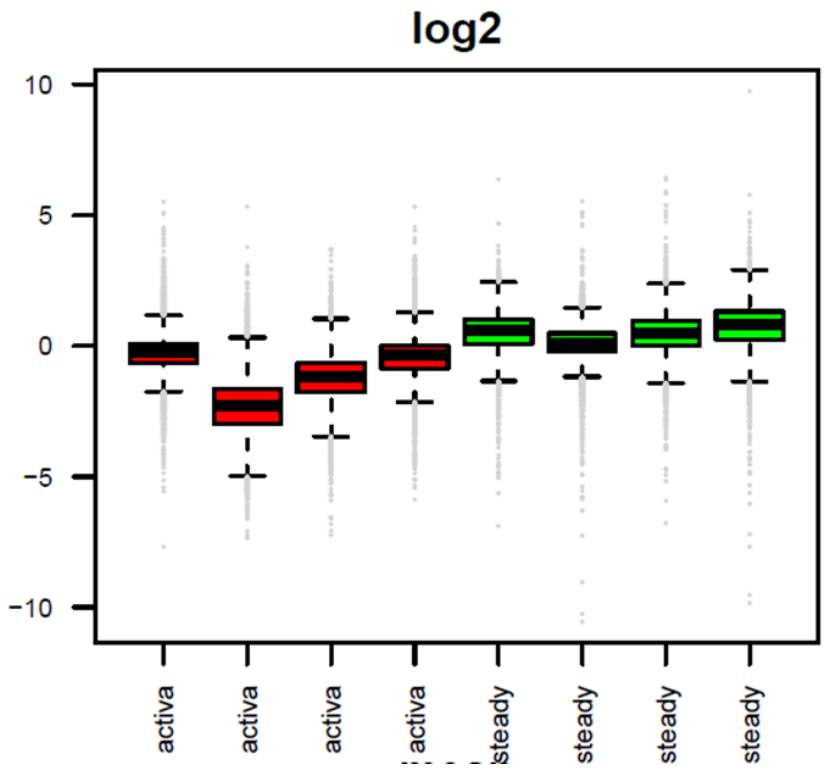
Need for normalisation?



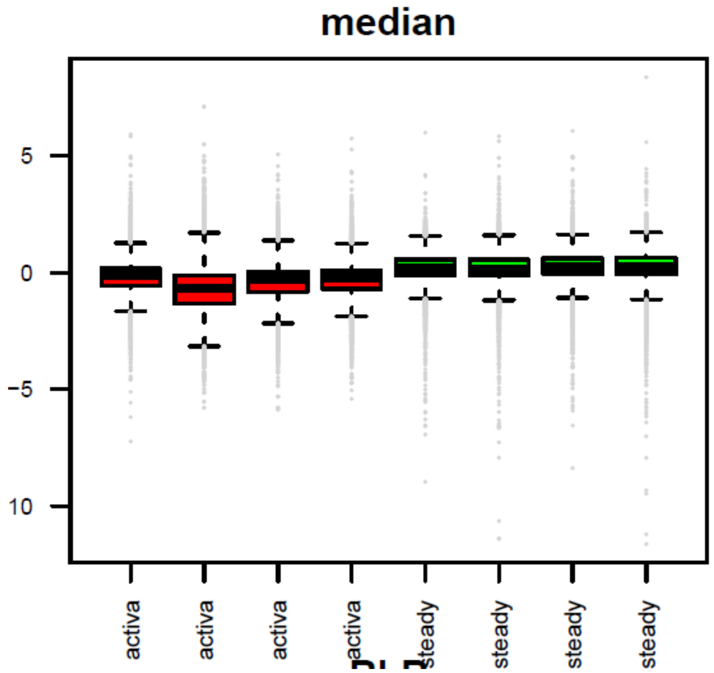


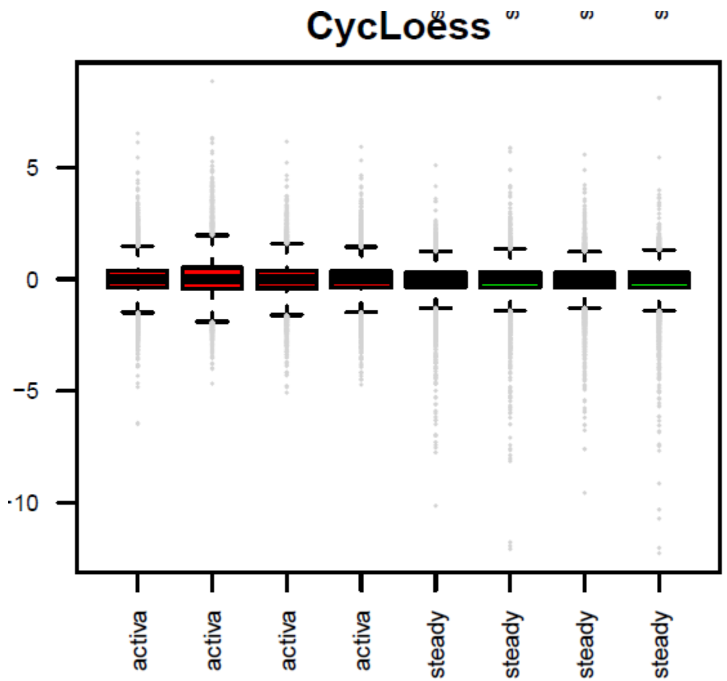
Total intensity of identified proteins varies a lot between samples, as does the number of missing values

RLE plots

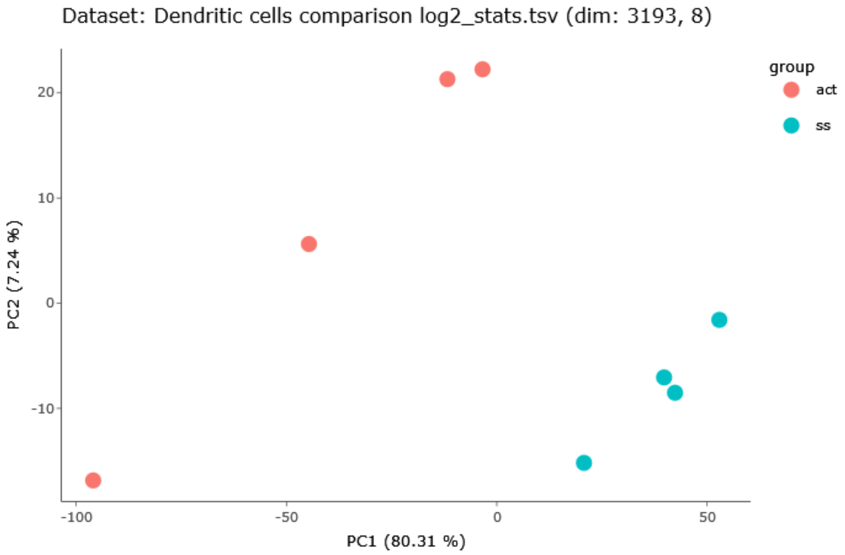


Normalisation

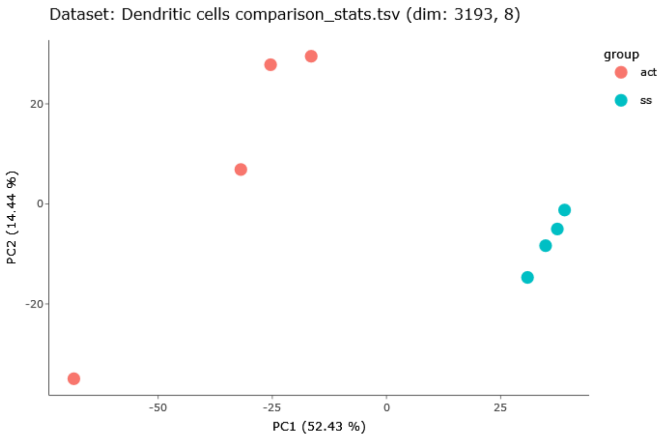




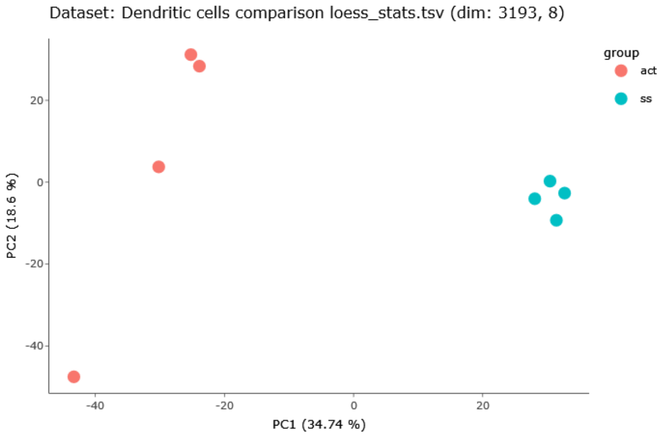
PCAs after different normalisations



No normalisation



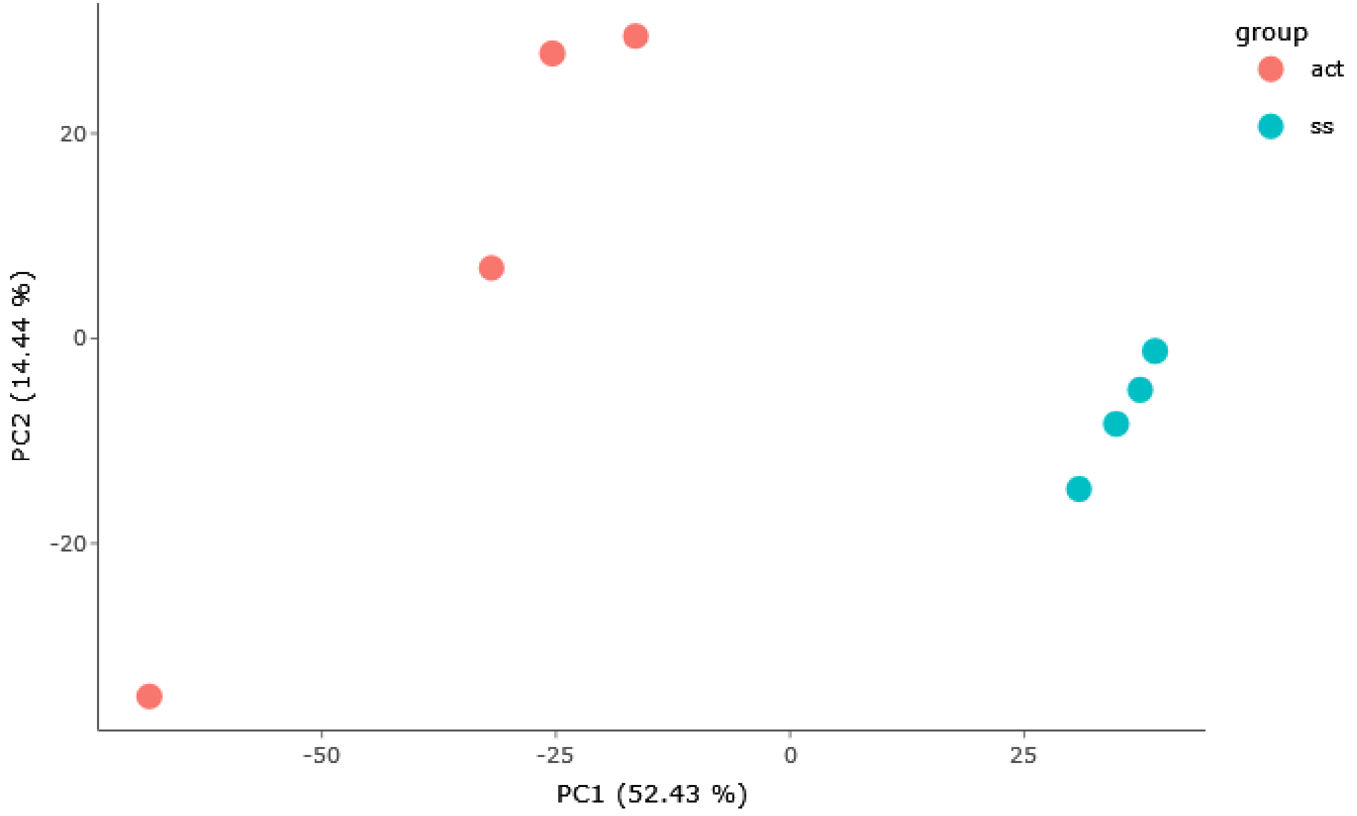
Median normalisation



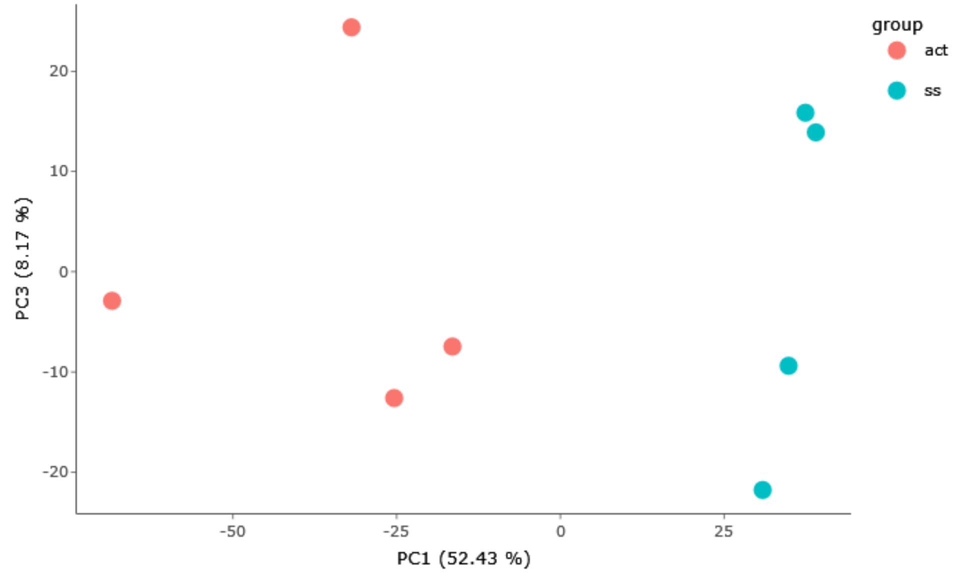
Cyclic LOESS normalisation

PCA after median normalisation

Dataset: Dendritic cells comparison_stats.tsv (dim: 3193, 8)

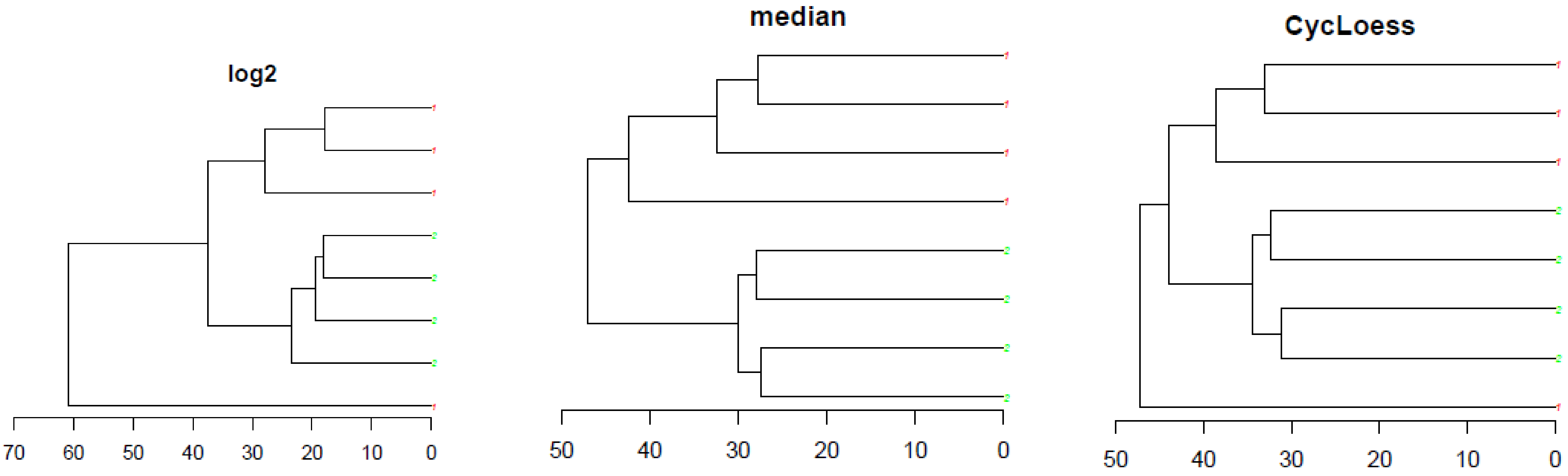


Dataset: Dendritic cells comparison_stats.tsv (dim: 3193, 8)



Good separation between groups in PC1, which represents >50 percent of the variation

Unsupervised clustering



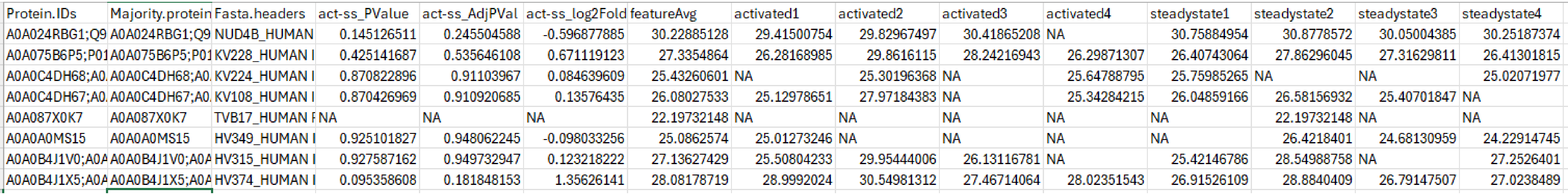
Differential abundance analysis

- Perform Empirical Bayes LIMMA test between groups using NormalyzerDE
- Calculates P-values, adjusted p-values and fold changes between the two groups (conditions)

Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Statistical Applications in Genetics and Molecular Biology Volume 3, Issue 1, Article 3. http://www.statsci.org/smyth/pubs/ebayes.pdf

Willforss et al (2004) J Proteome Res. 18, 2, 732–740 https://doi.org/10.1021/acs.jproteome.8b00523

Differential abundance – spreadsheet view



Protein.IDs	Majority.protein	Fasta.headers	act-ss_PValue	act-ss_AdjPVal	act-ss_log2Fold	featureAvg	activated1	activated2	activated3	activated4	steadystate1	steadystate2	steadystate3	steadystate4
A0A024RBG1;Q9	A0A024RBG1;Q9	NUD4B_HUMAN	0.145126511	0.245504588	-0.596877885	30.22885128	29.41500754	29.82967497	30.41865208	NA	30.75884954	30.8778572	30.05004385	30.25187374
A0A075B6P5;P01	A0A075B6P5;P01	KV228_HUMAN I	0.425141687	0.535646108	0.671119123	27.3354864	26.28168985	29.8616115	28.24216943	26.29871307	26.40743064	27.86296045	27.31629811	26.41301815
A0A0C4DH68;A0.	A0A0C4DH68;A0.	KV224_HUMAN I	0.870822896	0.91103967	0.084639609	25.43260601	NA	25.30196368	NA	25.64788795	25.75985265	NA	NA	25.02071977
A0A0C4DH67;A0.	A0A0C4DH67;A0.	KV108_HUMAN I	0.870426969	0.910920685	0.13576435	26.08027533	25.12978651	27.97184383	NA	25.34284215	26.04859166	26.58156932	25.40701847	NA
A0A087X0K7	A0A087X0K7	TVB17_HUMAN F	NA	NA	NA	22.19732148	NA	NA	NA	NA	NA	22.19732148	NA	NA
A0A0A0MS15	A0A0A0MS15	HV349_HUMAN I	0.925101827	0.948062245	-0.098033256	25.0862574	25.01273246	NA	NA	NA	NA	26.4218401	24.68130959	24.22914745
A0A0B4J1V0;A0A	A0A0B4J1V0;A0A	HV315_HUMAN I	0.927587162	0.949732947	0.123218222	27.13627429	25.50804233	29.95444006	26.13116781	NA	25.42146786	28.54988758	NA	27.2526401
A0A0B4J1X5;A0A	A0A0B4J1X5;A0A	HV374_HUMAN I	0.095358608	0.181848153	1.35626141	28.08178719	28.9992024	30.54981312	27.46714064	28.02351543	26.91526109	28.8840409	26.79147507	27.0238489

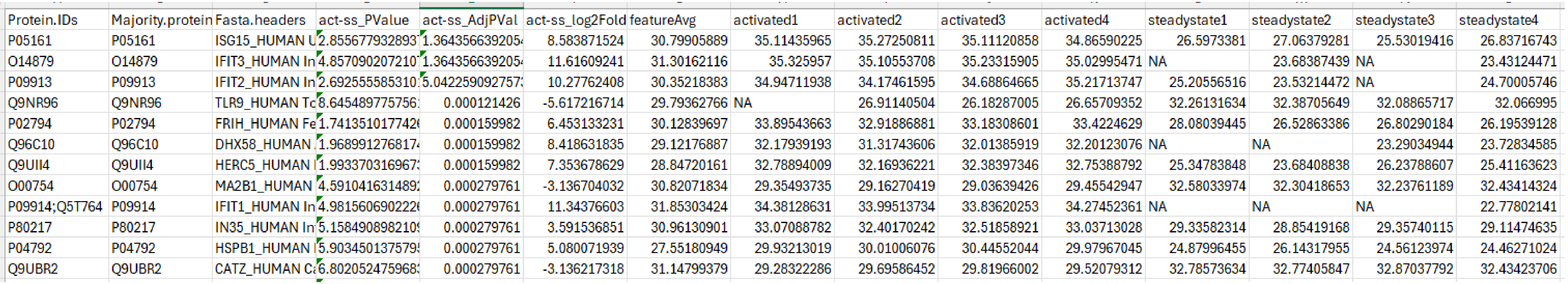
Note the columns:

act-ss_Pvalue: Unadjusted p-values for comparison between act and ss (activated vs steady state)

act-ss_AdjPVal: Adjusted p-values for the same comparison

act-ss_log2FoldChange: Log2 fold change for the comparison. Negative values downregulated!

Differential abundance – sorted on P-value



Protein.IDs	Majority.protein	Fasta.headers	act-ss_PValue	act-ss_AdjPVal	act-ss_log2Fold	featureAvg	activated1	activated2	activated3	activated4	steadystate1	steadystate2	steadystate3	steadystate4
P05161	P05161	ISG15_HUMAN U	2.855677932893	1.364356639205	8.583871524	30.79905889	35.11435965	35.27250811	35.11120858	34.86590225	26.5973381	27.06379281	25.53019416	26.83716743
O14879	O14879	IFIT3_HUMAN In	4.857090207210	1.364356639205	11.61609241	31.30162116	35.325957	35.10553708	35.23315905	35.02995471	NA	23.68387439	NA	23.43124471
P09913	P09913	IFIT2_HUMAN In	2.692555585310	5.042259092757	10.27762408	30.35218383	34.94711938	34.17461595	34.68864665	35.21713747	25.20556516	23.53214472	NA	24.70005746
Q9NR96	Q9NR96	TLR9_HUMAN To	8.645489775756	0.000121426	-5.617216714	29.79362766	NA	26.91140504	26.18287005	26.65709352	32.26131634	32.38705649	32.08865717	32.066995
P02794	P02794	FRIH_HUMAN Fe	1.741351017742	0.000159982	6.453133231	30.12839697	33.89543663	32.91886881	33.18308601	33.4224629	28.08039445	26.52863386	26.80290184	26.19539128
Q96C10	Q96C10	DHX58_HUMAN .	1.968991276817	0.000159982	8.418631835	29.12176887	32.17939193	31.31743606	32.01385919	32.20123076	NA	NA	23.29034944	23.72834585
Q9UII4	Q9UII4	HERC5_HUMAN I	1.993370316967	0.000159982	7.353678629	28.84720161	32.78894009	32.16936221	32.38397346	32.75388792	25.34783848	23.68408838	26.23788607	25.41163623
O00754	O00754	MA2B1_HUMAN	4.591041631489	0.000279761	-3.136704032	30.82071834	29.35493735	29.16270419	29.03639426	29.45542947	32.58033974	32.30418653	32.23761189	32.43414324
P09914;Q5T764	P09914	IFIT1_HUMAN In	4.981560690222	0.000279761	11.34376603	31.85303424	34.38128631	33.99513734	33.83620253	34.27452361	NA	NA	NA	22.77802141
P80217	P80217	IN35_HUMAN In	5.158490898210	0.000279761	3.591536851	30.96130901	33.07088782	32.40170242	32.51858921	33.03713028	29.33582314	28.85419168	29.35740115	29.11474635
P04792	P04792	HSPB1_HUMAN I	5.903450137579	0.000279761	5.080071939	27.55180949	29.93213019	30.01006076	30.44552044	29.97967045	24.87996455	26.14317955	24.56123974	24.46271024
Q9UBR2	Q9UBR2	CATZ_HUMAN Ca	6.802052475968	0.000279761	-3.136217318	31.14799379	29.28322286	29.69586452	29.81966002	29.52079312	32.78573634	32.77405847	32.87037792	32.43423706

Note the columns:

act-ss_Pvalue: Unadjusted p-values for comparison between act and ss (activated vs steady state)

act-ss_AdjPVal: Adjusted p-values for the same comparison

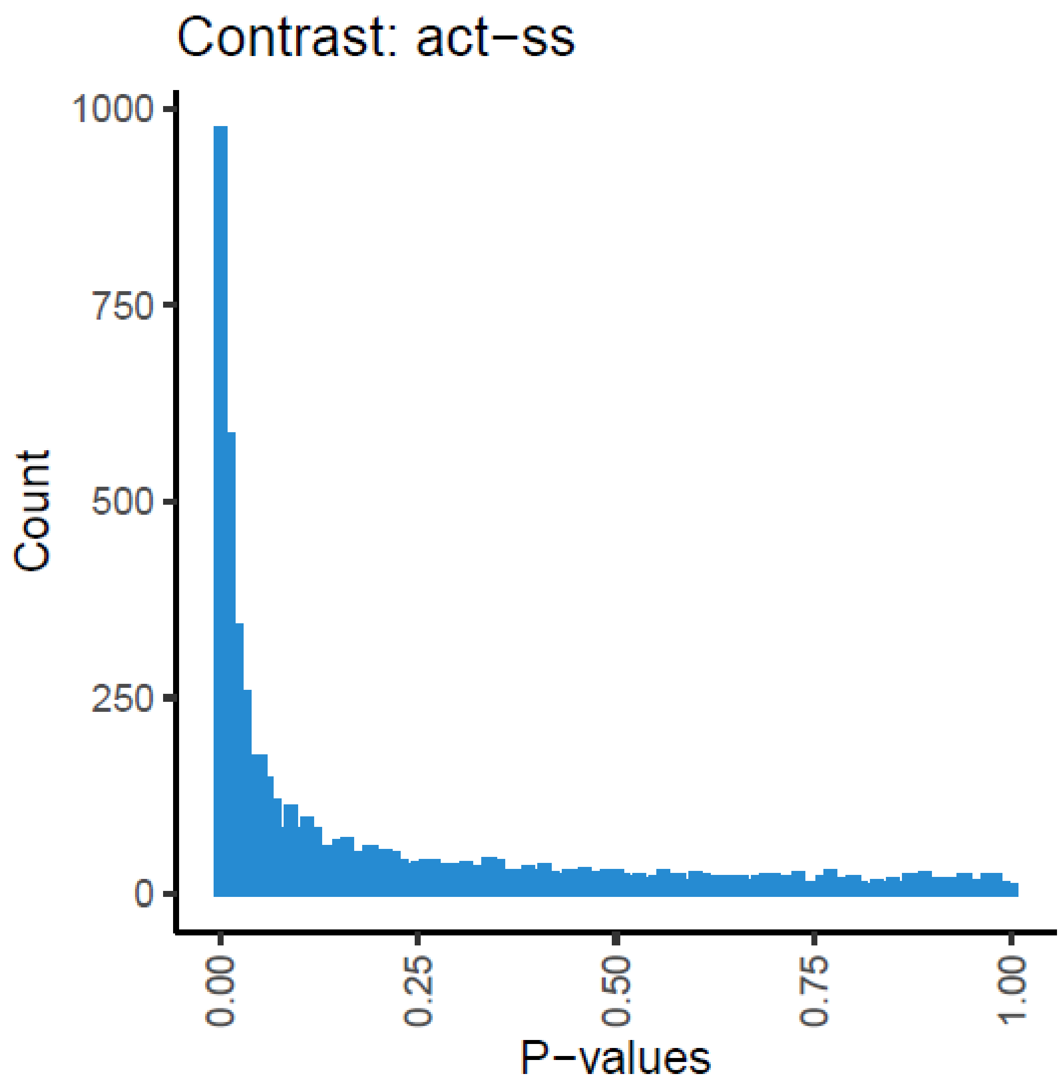
act-ss_log2FoldChange: Log2 fold change for the comparison. Negative values downregulated!

Filtering on AdjPVal < 0.05 to keep proteins that are differentially abundant at FDR<0.05

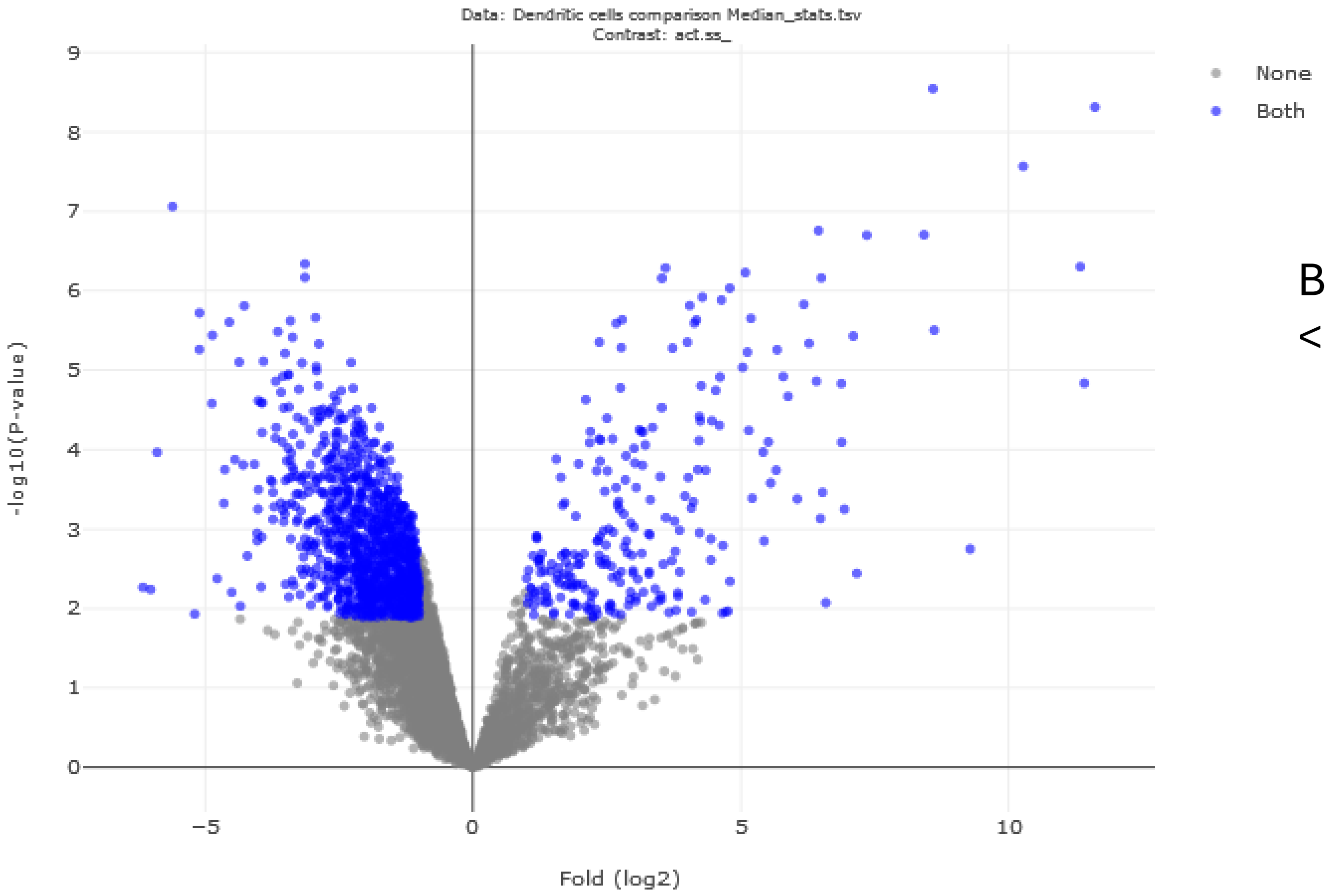
P-value histogram

- For well-behaved data p-values will be evenly distributed between 0 and 1 with a peak at 0 for the proteins that change between conditions

http://varianceexplained.org/statistics/interpreting-pvalue-histogram/

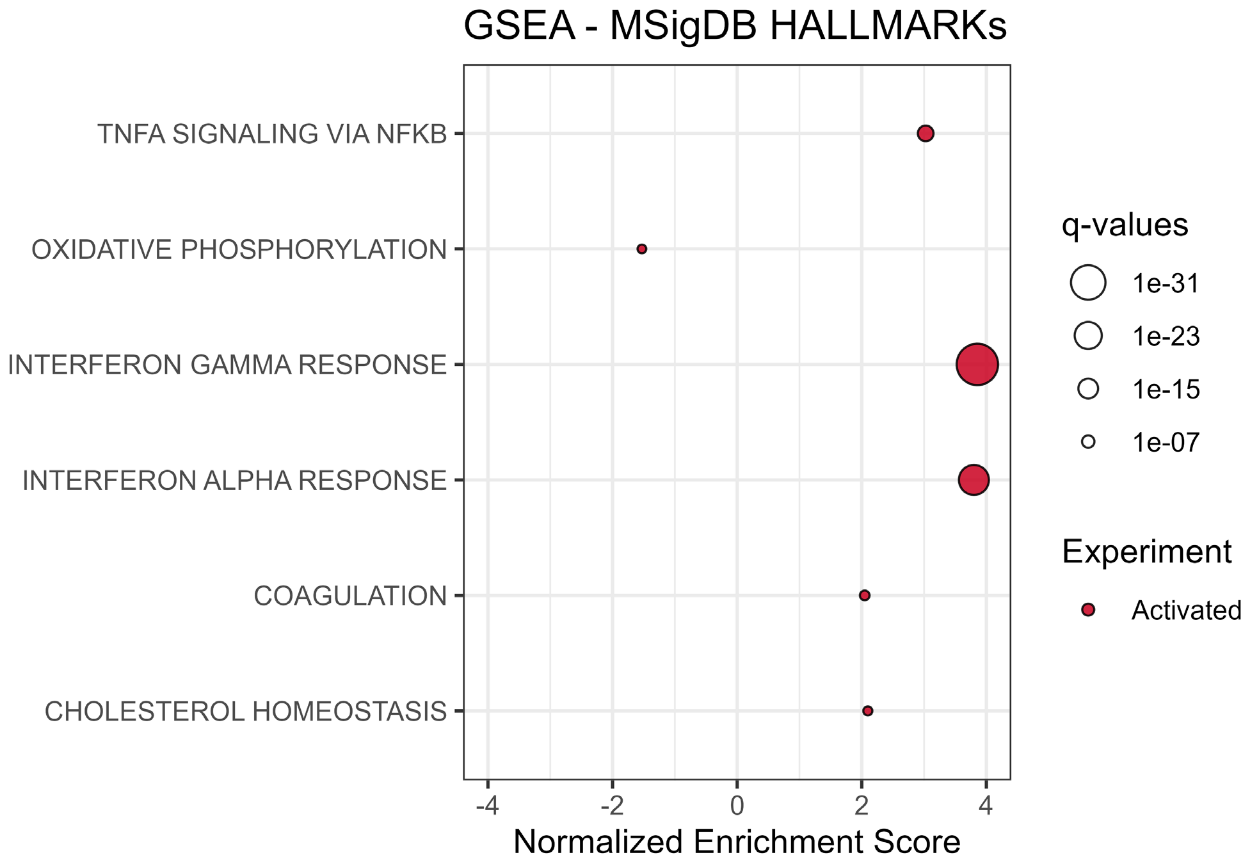


Differential abundance – volcano plots



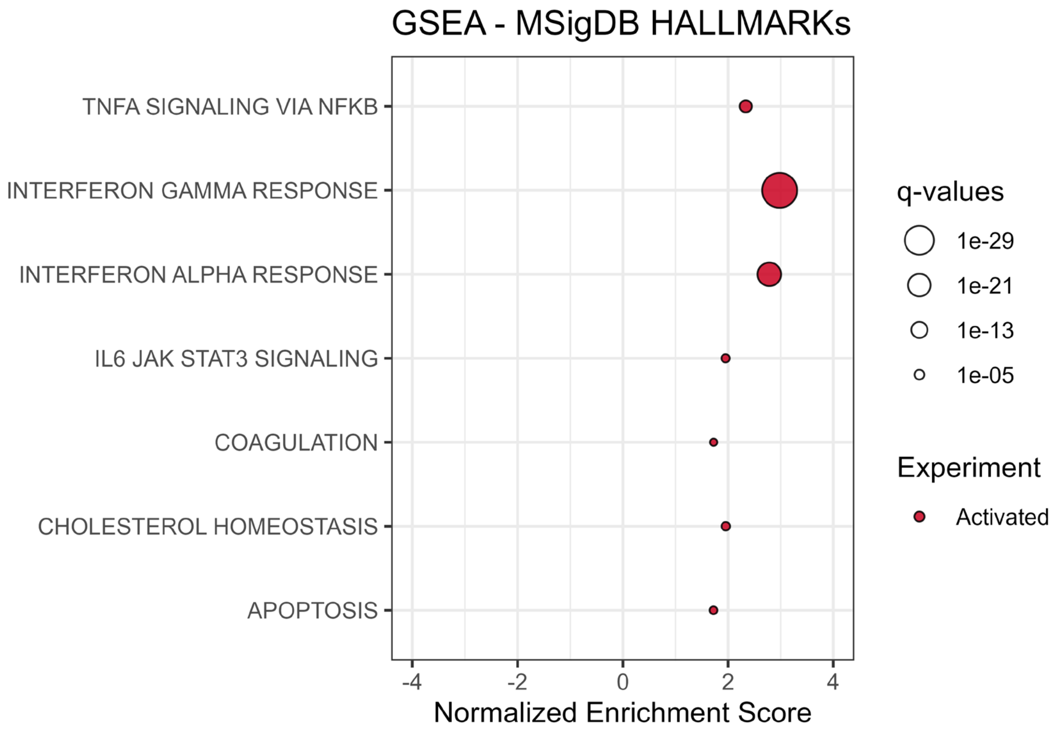
Blue dots: proteins significant at <5% FDR and log2FC>=1

GSEA

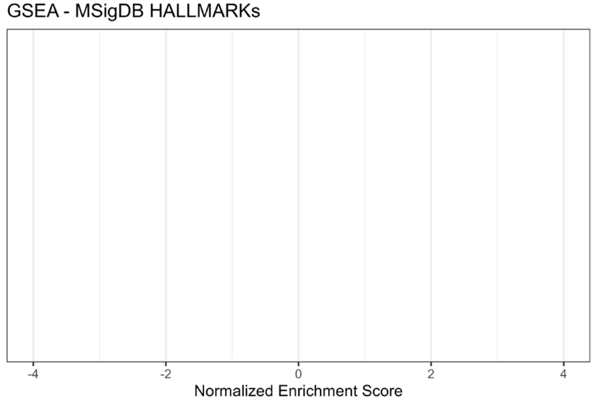


Median normalisation

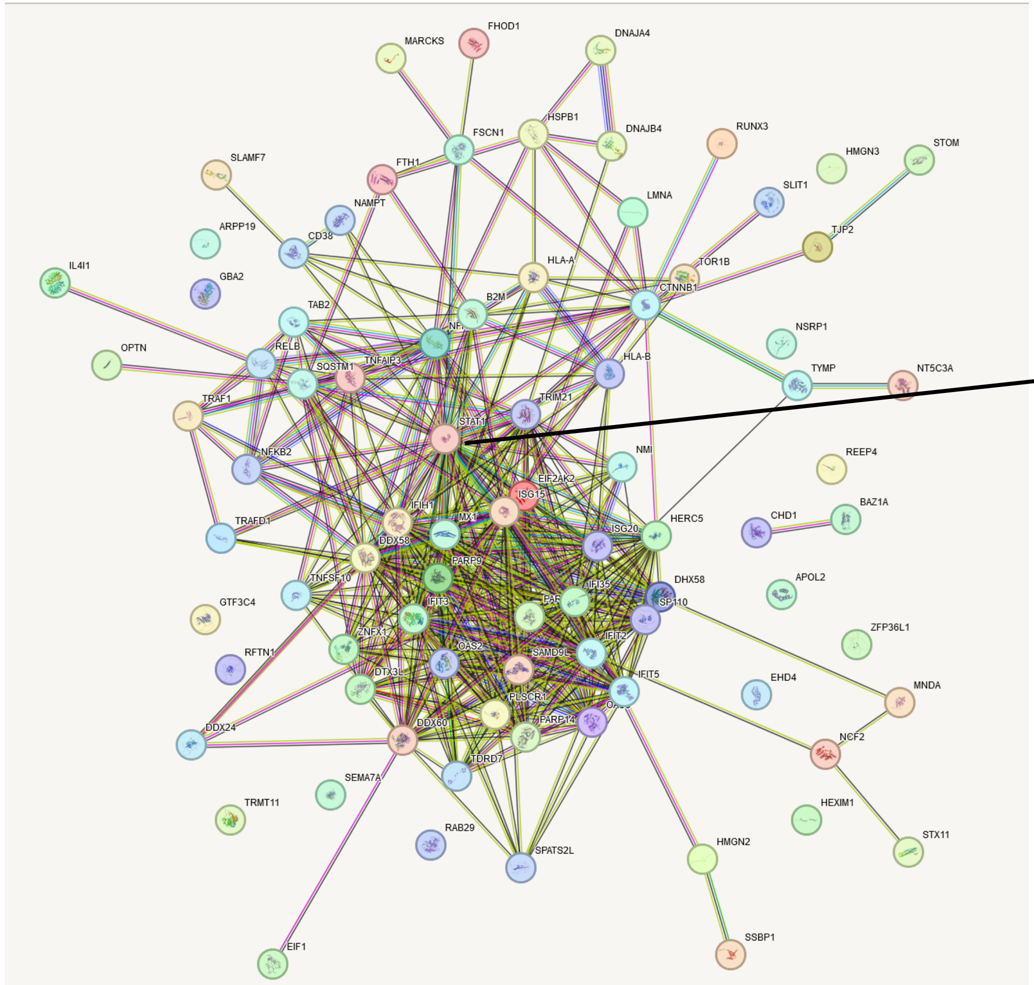
Cyclic loess normalisation



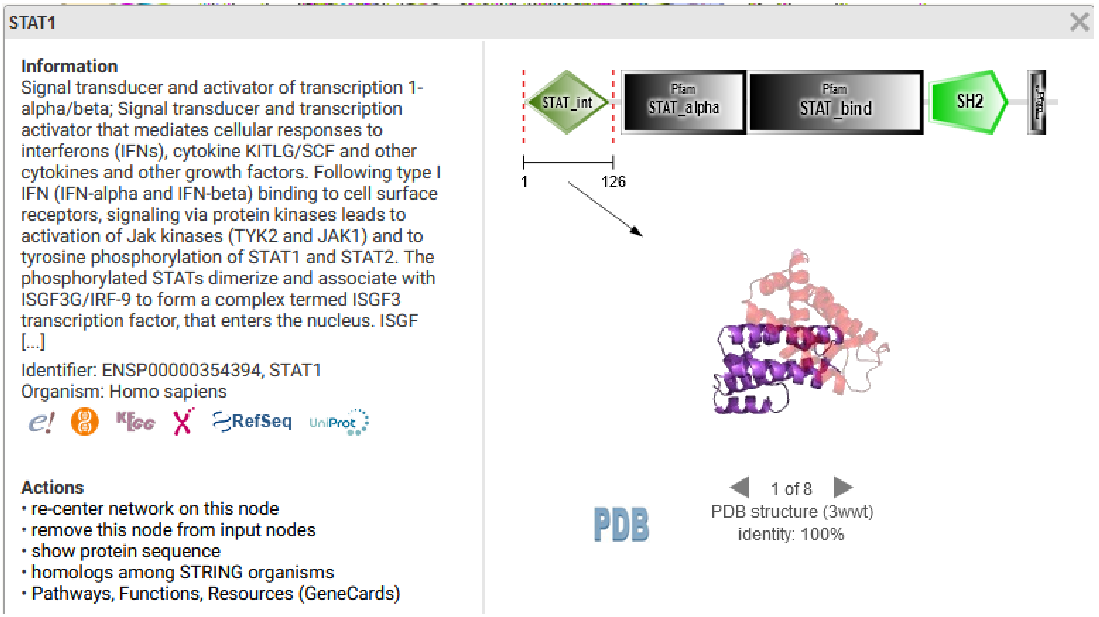
No normalisation



String network



Genes upregulated (FDR<0.01) after activation



STAT1

Information

Signal transducer and activator of transcription 1-alpha/beta; Signal transducer and transcription activator that mediates cellular responses to interferons (IFNs), cytokine KITLG/SCF and other cytokines and other growth factors. Following type I IFN (IFN-alpha and IFN-beta) binding to cell surface receptors, signaling via protein kinases leads to activation of Jak kinases (TYK2 and JAK1) and to tyrosine phosphorylation of STAT1 and STAT2. The phosphorylated STATs dimerize and associate with ISGF3G/IRF-9 to form a complex termed ISGF3 transcription factor, that enters the nucleus. ISGF [...]

Identifier: ENSP00000354394, STAT1

Organism: Homo sapiens

RefSeq UniProt

Actions

- re-center network on this node
- remove this node from input nodes
- show protein sequence
- homologs among STRING organisms
- Pathways, Functions, Resources (GeneCards)

PDB structure (3wwt)
identity: 100%