Discovering the Hidden Value Within Data: How Public Proteomics Data Can Reveal Biomarkers Your Volcano Plot Missed

Introduction to the problem:

Proteins are arguably the most important frontline biological unit during pathogen and drug responses, and tumorogenesis. From designing diagnostic tests, discovering biomarkers and drug targets to understanding the underlying causes of disease, accurate analysis of protein dynamics in patient samples is critical.

Differential abundance analysis is a simple and effective means of visualising changes in protein dynamics between two groups of samples. However, it is important to be aware of the unique challenges posed by proteomics data, especially for analysts used to operating in the more established 'omics disciplines.

Typically, differential abundance analysis involves a volcano plot of proteins shared between two groups of biological replicates, after missing data points have been filled with zeros or mean values. However, data-loss or incorporation of numerical artifacts can occur surprisingly easily, even in a simple comparative analysis, unless strategies for imputing missing values and normalising data are not carefully considered.

Here, we show how it's possible to minimise information loss and explore the full depth of your data, uncovering critically interesting data points that might easily be overlooked, whilst introducing good practices for imputing missing values so that subtle patterns within the data are preserved.

The data:

In this case study, we analyse a public ovarian cancer dataset (PXD045417, Bateman et al. 2004 https://pmc.ncbi.nlm.nih.gov/articles/PMC10937683/). This is a high-quality, deep-coverage, isobarically-labelled DDA (data-dependent analysis) dataset comprising three different tissue categories: Tumor (70 samples), Stroma (30 samples) and a tumor-stromal mixture (70 samples).

We also take healthy ovarian proteomics data from Ouni et al. 2018 (https://pubmed.ncbi.nlm.nih.gov/29475978/), Di Meo et al. 2020 (https://pubmed.ncbi.nlm.nih.gov/33107741/), Wilhelm et al. 2014 (https://pubmed.ncbi.nlm.nih.gov/24870543/), Wang et al. 2019 (https://pubmed.ncbi.nlm.nih.gov/30777892/) and Kim et al. 2014 (https://pubmed.ncbi.nlm.nih.gov/24870542/).

Initial data preparation:

Although the authors used labelled data, normalised to internal standards, here we demonstrate techniques that require un-normalised data. We therefore have manually calculated protein intensities from the peptide

spectral match level. We also filter proteins against a collection of common contaminants (https://pubmed.ncbi.nlm.nih.gov/23921808/) and delete some values (see below).

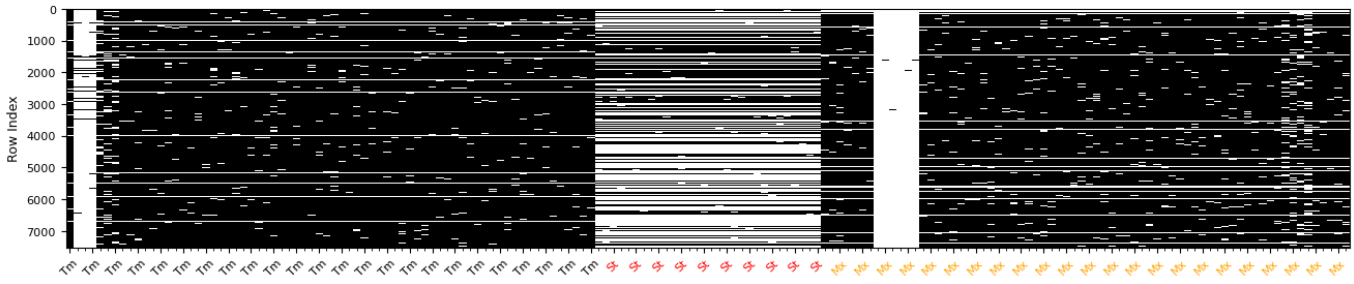
Identifying the different categories of missing values present in the data

Missing values can be grouped into categories, depending on the likely reason why the data is missing. In Figure 1, marked in white, we can see four categories of missing value across the three tissue types:

1. Proteins only present in certain tissue batches
2. Protein missing-not-at-random (MNAR) where most values are missing
3. Proteins missing at random (MAR)
4. Samples with a missing value pattern somewhere between MNAR and MAR

Category 1 is a natural feature of this data, whilst categories 2 - 4 have been created by the intentional deletion of some datapoints, for the purposes of demonstrating how to handle these different types of missing values.

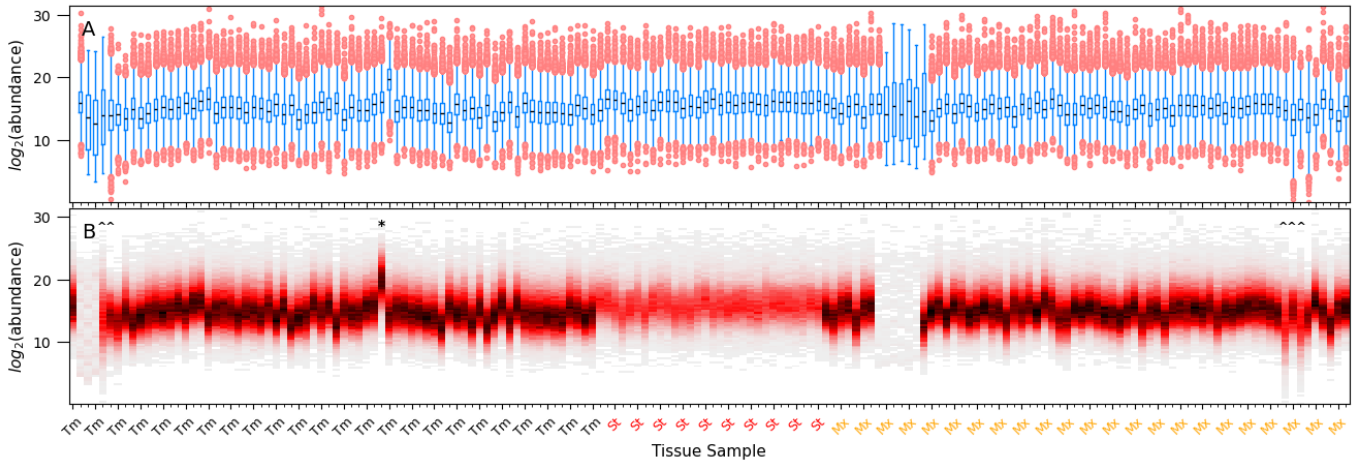
Figure 1 Location of missing values (white) within and between tissue datasets



Simply imputing or zero-filling across the rows for each category of missing data would introduce artifacts, so we'll need a sensitive and bespoke approach.

Category 2 missing values are straightforward; with so much data missing, something must have gone wrong during sampling or processing. The only posible fate for this data is to leave it out. For Category 3, the pattern of missing values really does look random. It's therefore likely safe to impute these values using the protein means from within each sample group. Categories 1 and 4 are less straightforward. We can gain more insight into this data by plotting the distributions of proteins intensity, using boxplots (Figure 2A) and density histograms (Figure 2B).

Figure 2 Tissue Sample Value Distributions as boxplots (top) or density histograms (bottom)



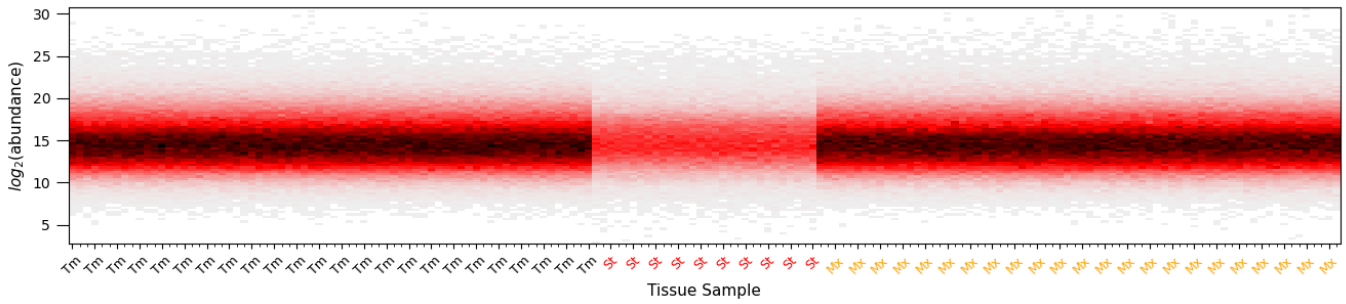
Category 4 missing values (marked '^' in Figure 2) have a slightly lower mean intensity and wider distribution than other Tumor or Mixed samples. Injection of a sample into the mass spectrometer at below-average protein concentration is one possible reason for this. Low-abundance proteins are more likely to generate missing values, so these may be otherwise typical samples where the contribution of missing values from low-abundance proteins has been exaggerated by an overall lower sample concentration. If we first scale intensities so they are consistent with their Tumor or Mixed counterparts, we can then impute using mean values, as per Category 3.

Is it noteworthy that not all outliers have missing values; some samples have been artificially adjusted to mimic injection at an above-average concentration (marked '*' in Figure 2). This effect can easily be eliminated during normalisation in Figure 3.

A genuine pattern in this data is the low overall intensity of the Stromal tissue samples (Figure 2B). The overall low intensity of the Stromal samples explains in part why there are fewer shared proteins with Tumor and Mixed samples (Category 1 missing values), in addition to it simply being a different tissue. Overall low intensity can arise from initial low proteins yields, poor enzymatic digestion or signal suppression from within the instrument, or from a few highly abundant sample proteins.

After carefully imputing values where possible, all three tissue groups are scaled to a single mean and variance (Figure 3) to facilitate comparisons between groups.

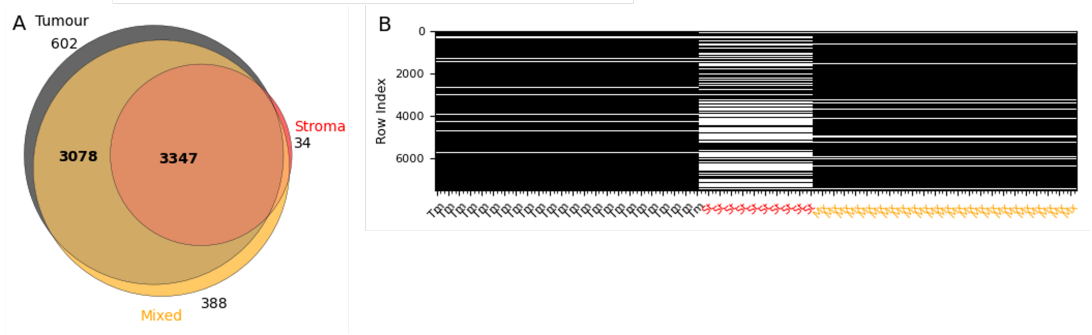
Figure 3 Density histogram of protein intensity distributions after centering and scaling Tumour, Stroma and Mixed tissue datasets



Comparisons between imbalaced groups

We still haven't properly addressed the Category 1 missing values. Making comparisons between datasets of different sizes and composition is a common problem for biologists. Figure 4 illustrates the scale of the problem and the impact it will have on data loss.

Figure 4 Shared and unique proteins between tissue types



Typically, a protein must be present in both groups to be captured by a comparative analysis. As show in Figure 4, any comparison involving the Stroma would mean losing almost half the protein information available. This 'lost' data contains interesting biomarkers; the 34 proteins unqiue to the Stroma included P15090, P01889, P06312 and P01037, all of which are under investigation as serum biomarkers fro ovarian cancer, as well as other potential prognostic biomarkers. Conversely, proteins absent from the Stroma are also interesting as these can be specific to the tumour microenvironment.

This dataset doesn't have a comprehensive non-cancerous control proteome, but if we build one from publicly available data then we can compare each tissue group to this reference. This will partially address the problem of Category 1 missing values; as long as all the experimental proteins are present in the healthy reference, we will be able to see if their abundance is affected by ovarian cancer.

Bringing in an external dataset for comparison

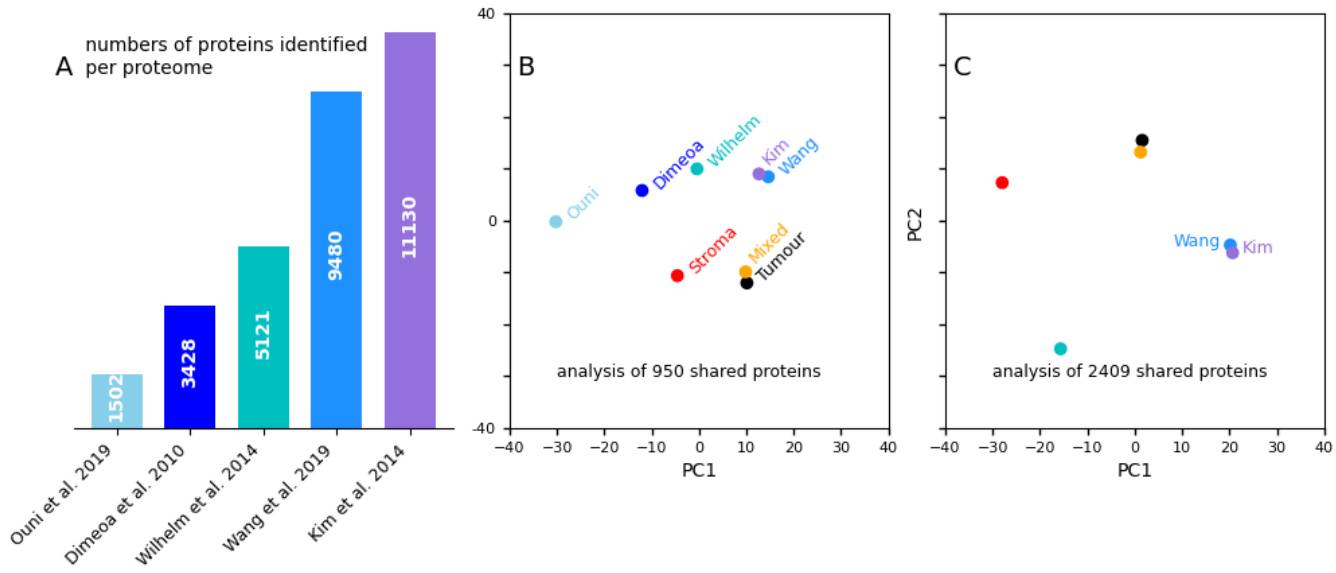
There are a few key steps to be aware of when making comparisons to other datasets. Firstly, it's necessary to have information at the peptide level, so the protein intensities can be calculated and corrected as we did in the 'preparatory steps' for this dataset. If peptide-level information hasn't been upload to public data repositories, we can download the instrument files and re-search these against the current release of the human proteome.

The experimental data is a few years old, so it's been searched against a slightly older version of the human proteome. Protein accessions are updated, merged and split regularly. We need to update the experimental accessions so we are comparing like for like. Sometimes, one old accession becomes updated to multiple new ones. In these cases we must remember that all new proteins are equally valid if this accession becomes a protein-of-interest.

After updating the accessions, we select the proteomes that are the most comparable to the experimental data. The three most comprehensive proteomes (Figure 5A) appear most similar to the experimental data,

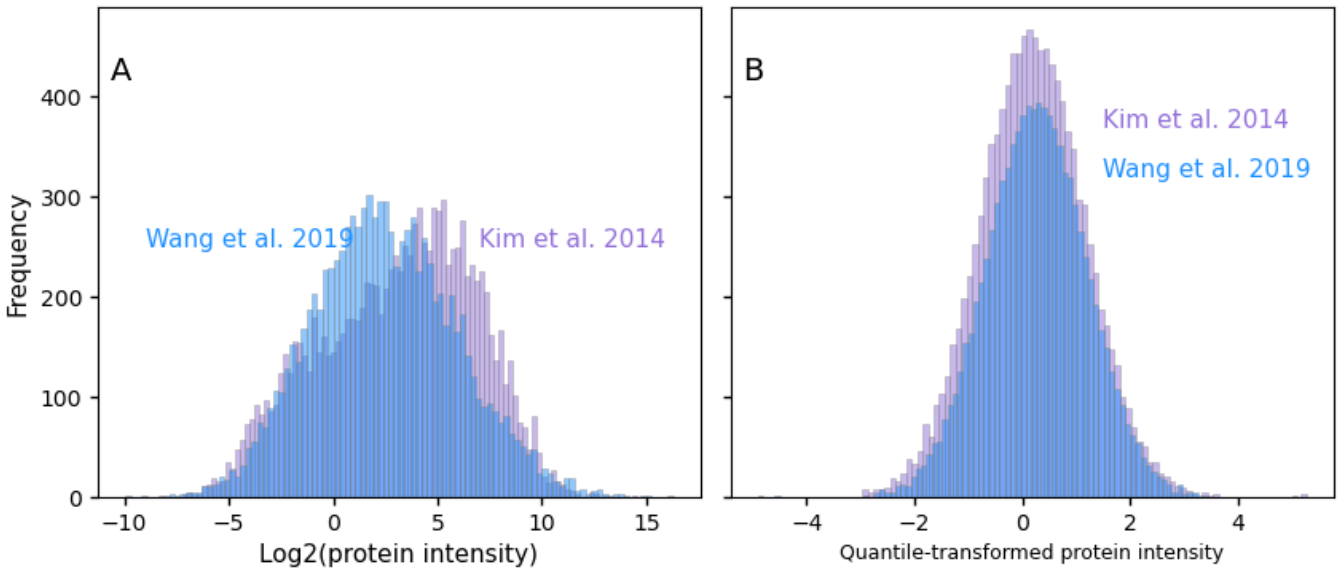
when comparing shared proteins (Figure 5B). Eliminating the two least similar proteomes increases the number of comparable proteins (Figure 5C). Kim et al. 2014 and Wang et al. 2019 appear to be the most similar to the data in question, so we will build our 'healthy reference' from these two.

Figure 5 Similarity of healthy Ovarian proteomes to experimental data, by protein numbers and PCA



In neither healthy proteome are the log-transformed protein intensities normally distributed (Figure 6A). We need them to have the same distribution as the experimental data against which they are being compared, so we apply a quantile transformation (Figure 6B).

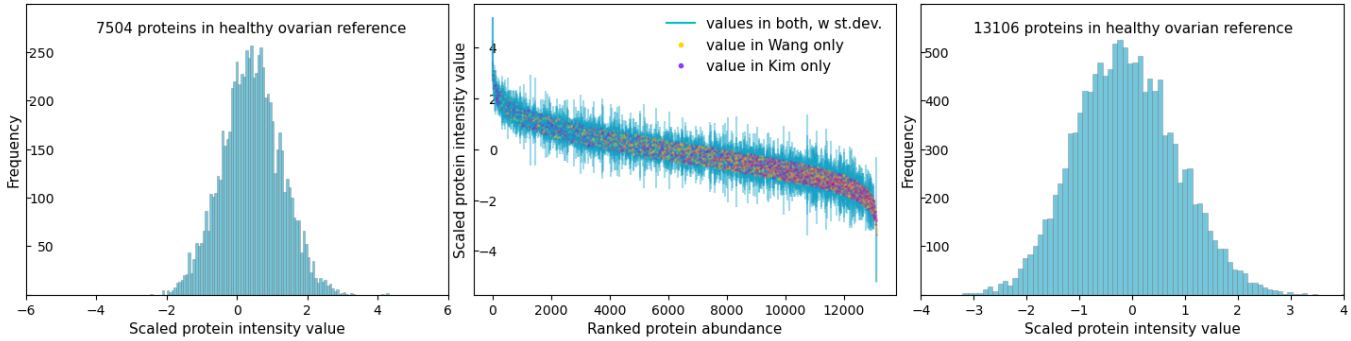
Figure 6 Distribution of protein intensities before and after quantile transformation



Averaging the normalised values of proteins shared between the reference proteomes yields 7416 proteins (Fig. 7A) - but the total number of unique proteins is almost 13,000. We want as many proteins as possible in our healthy reference, so that we minimise data loss is comparisons with the experimental data. We could simply use single values for proteins present in only one reference proteome - but we can't calculate the standard deviation of a single value, and we need standard deviations for a comparative analysis.

Overlaying the position of missing values on the union of these proteomes shows that missing proteins are much more frequent, and have higher standard deviation, at lower abundances (Fig. 7B). A reasonable approximation is, therefore, to take the single value as the mean but penalise these values with a high standard deviation e.g. 80th percentile of all standard deviations for shared proteins.

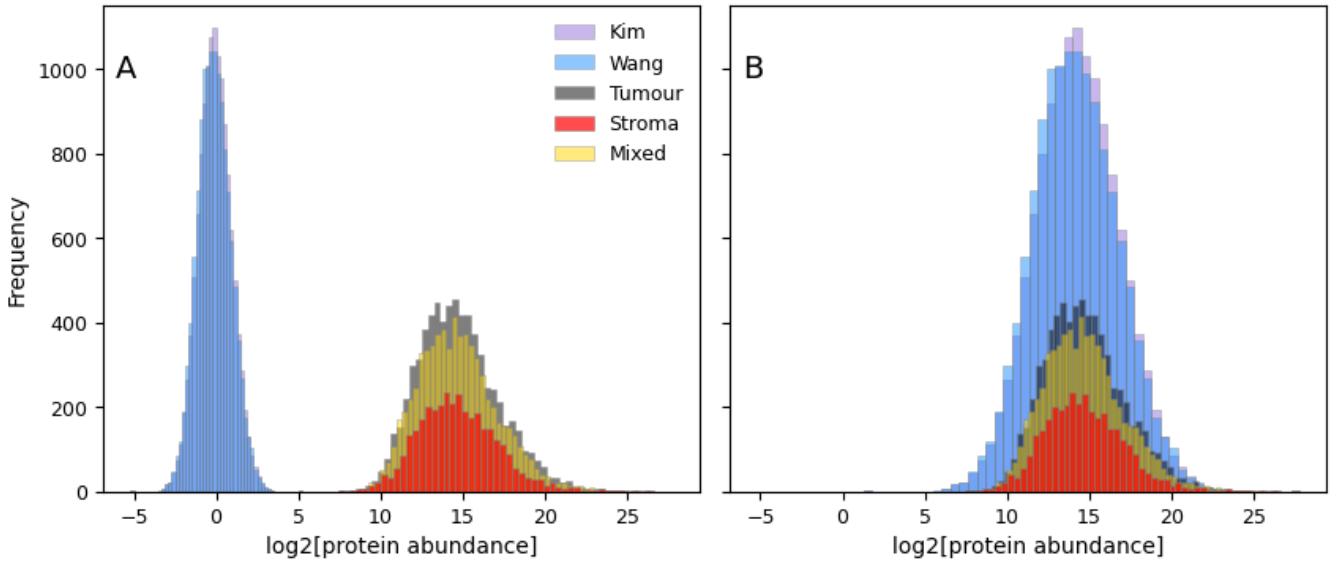
Figure 7 Numbers and distributions of protein intensities in combined healthy proteome before and after compensating for single-value proteins



This yields a normally distributed histogram of almost 13,000 proteins (Fig. 7C), enough to prevent excessive data loss during comparisons. We will note which proteins received artifical standard deviations, so any interesting biomarkers we find later can be assessed for confidence in more detail.

The final task is to scale the healthy ovarian reference so it's directly comparable to the experimental data, which is achieved by setting the mean and variance of the whole reference set to be mean and variance of its shared proteins in the experimental datasets (Figure 8).

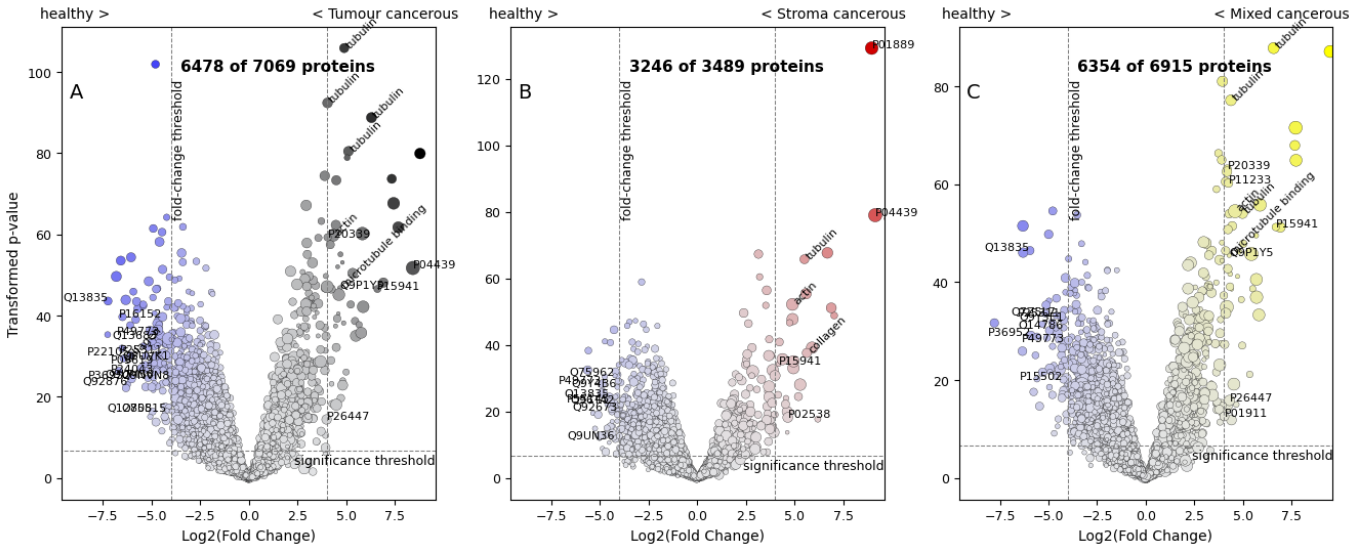
Figure 8 Scaling the distribution of healthy protein abundances to experimental data



We can now perform a differential abundance analysis between each of the tissue types and the healthy reference (Figure 9). If we set p=0.01 (transformed p-value 6.64) and our fold change threshold to 8-fold (Log2[3]), we capture a lot of proteins with significant fold changes - maybe more than is practicaly useful. Taking advantage of the Uniprot API, we can quickly pull out proteins by keyword. Searching on ['cancer' | 'tumor' | 'metasta*' | 'cell migration' | 'cell proliferation' | 'ovar*' | 'apoptosis'] quickly triages the significant proteins for known cancer-associated proteins.

The resulting picture is one where proteins associated with immune responses (HLA-A [P04439], HLA-DRB1 [P01911]) or with tumour progression (RAB5A [P02239], MUC1 [P15941], S100A4 [P26447], RALA [P11233]) show increased abundance, and proteins involved in tumour suppression and regulation of cell migration/proliferation show decreased abundance. Notably, we see an increase in cytoskeletal proteins in tumour-containing tissue, which is a characteristic of ovarian cancer. So too is increased collagen but, interestingly, this is only observed in the Stroma. Collagen is part of the extracellular matrix, changes significantly in cancer progression and is used as a serum biomarker for ovarian cancer.

Figure 9 Differential abundance analysis Patient vs Healthy, visualising shared proteins



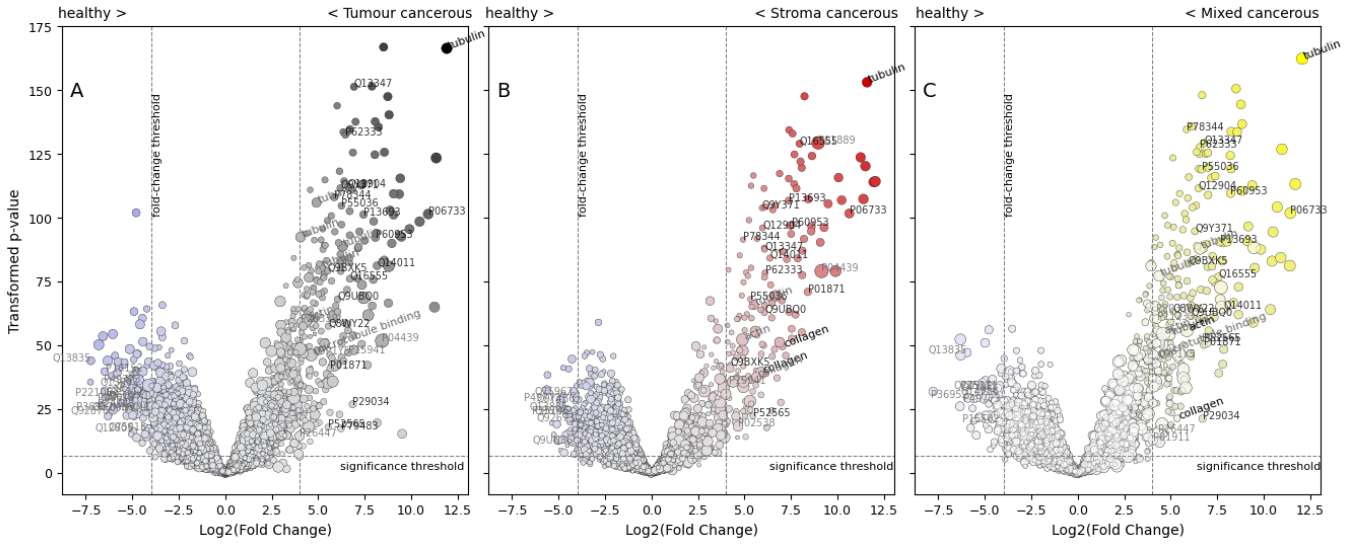
Had we not brought in the healthy data, the stromal samples would have formed the control instead. Re-running the comparison using only proteins shared between either Tumour or Mixed tissue and the Stromal tissue yielded only 51 significant proteins, of which 5 were labelled in the keyword search. This compares to 249 significant proteins, of which 41 were labelled, when the healthy reference was included, plus a contextual setting for the Stromal tissue of cancer patients.

Filling in values for comparisons missing between experimental and healthy reference datasets

Note that for each comparison in Figure 9, a few hundred proteins are still missing. These are proteins absent in the healthy reference. It would be interesting to include these in the visualisation, as the unique presence of a protein in the experimental data implies that protein is present by necessity, especially when the experimental dataset is appreciably smaller than the healthy reference.

Often a missing value means a protein is below the limit of detection for the mass spectrometer. Filling in the missing reference proteins with a low-but-realistic value, for example the 10th centile value for protein intensity, and penalising them with the 80th centile standard deviation as before, allows missing comparisons to be added to the volcano plot. Once again, we keep careful note of any protein with partially artifical data and color them differently.

Figure 10 Differential abundance analysis Patient vs Healthy, visualising experimental proteins absent from the healthy reference

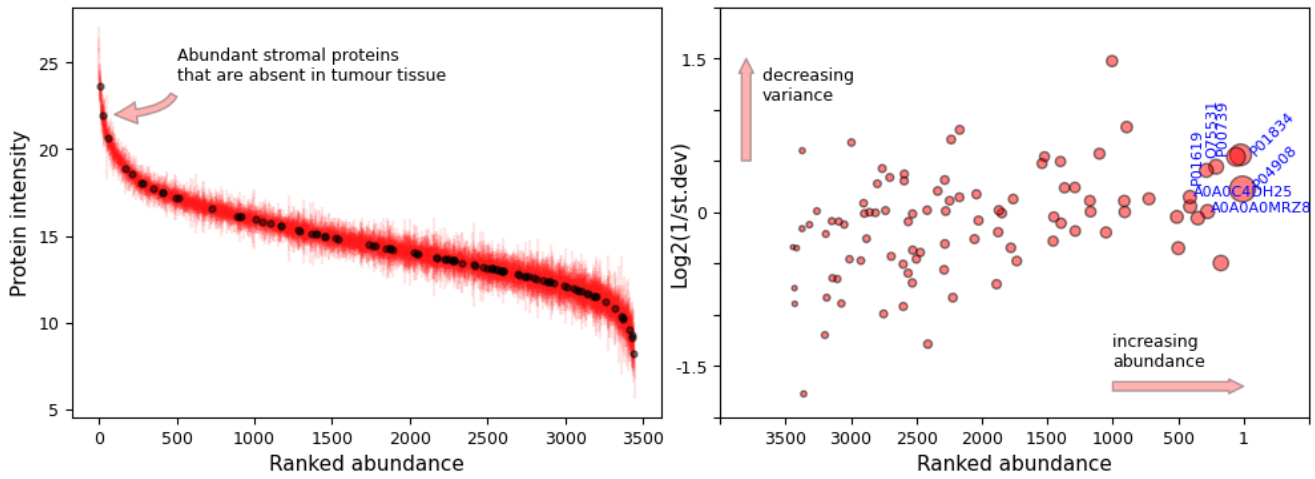


Previously-discovered significant proteins are whited-out and labelled in grey (Figure 10). Using the same keyword search in the Uniprot API shows that this strategy has yielded many more significatly abundant proteins likely associated with cancer. Of the newly-observed, 8 are currently under investigation as early diagnostic markers (Q16555 [DPYSL2] and Q12904 [AIMP1] and P62333 [PSMC6], P29034 [S100A2], P06733 [ENO1], P01871 [IGHN], P13693 [TPT1], P78344 [EIF4G2]) and 10 as prognostic markers in various cancers. Again, increased collagen appears to be associated with Stromal, rather than Tumour, tissue.

Filling in values for comparisons missing between two experimental datasets

The previous section demonstrated the impact of filling in missing data in the healthy reference data on the assumption that missing proteins were close to, or below, the mass spectrometer's detection limit. This is a reasonable assumtion when the proportion of missing proteins relatively low but if this approach were extended to Stromal vs Tumour tissue, over 3000 protein values would need to be filled. For very unbalanced datasets, a more direct visualisation can be used. In Figure 11, we highlight uniquely Stromal proteins within the Stromal data (Figure 11A) and highlight which of those unqiue proteins show the greatest abundance and lowest variance (Figure 11B). This reveals the presence of proteins (P00739 (HPR), O75531 (BANF1), O14950 (H2A) and A0A0A0MRZ8 (CALM3)) currently under investigation as less-invasive biomarkers in ovarian and/or cervical cancers.

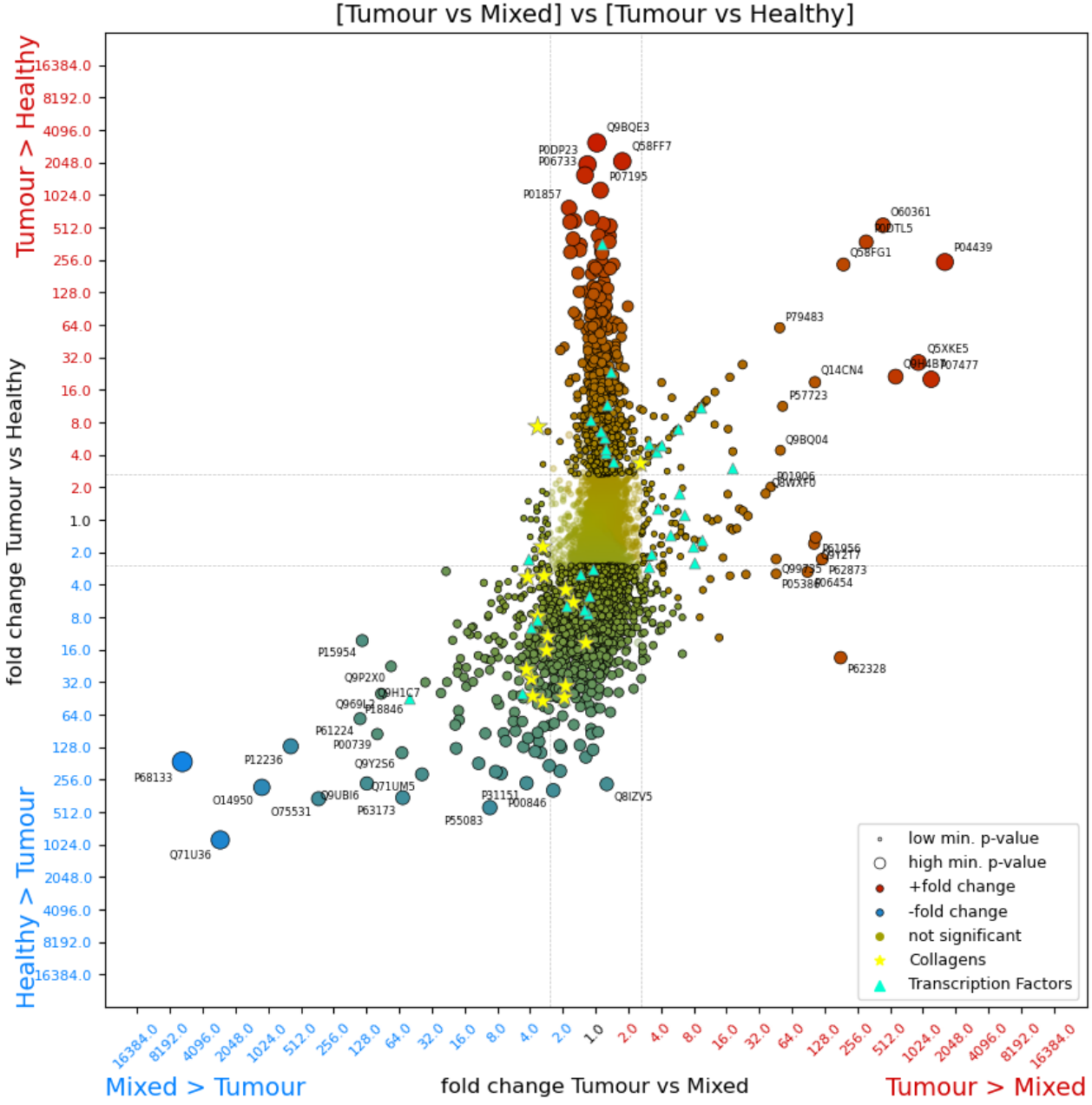
Figure 11 Alternative visualisation of proteins unique to Stromal tissues compared to Tumour tissues



A new view on Differential Abundance analysis - combinations of combinations:

Differential abundance analysis works well when two samples categories are being compared, but more complex comparisons are often necessary. By comparing the fold changes (FCs) of pairwise comparisons, up to four categories can be simultaneously analysed. In the below FoldChange-FoldChange plot (Figure 12), we separate proteins in tumour vs healthy vs mixed tissue proteomes (for clarity, only proteins at the extremeties have been labelled).

Figure 12 Fold changes in one pair-wise comparison plotted against fold changes in another



The separation axes are arranged such that the further up and/or right a protein appears, the more tumour-associated it it. Those far right but low down are tumour-associated but also very abundant in healthy tissue e.g. Thymosin beta-4 (P62328). Those in at the middle top are equally associated with tumour tissue and mixed stromal-tumour tissue. The bottom left corner shows proteins are are abundant in both healthy ovarian tissue and mixed stromal-tumour tissue but not specific to the tumour microenvironment. Therefore, although there is a 'tumour-tumour' corner, there is no 'healthy-healthy' corner; all proteins are present in cancerous tissue to some degree. Conseqently, the upper left corner is largely empty as it is difficult for proteins to be simultaneously more abundant in tumours compared to healthy tissue AND increased in stromal-tumour tissue compared to tumours (Figure 12).

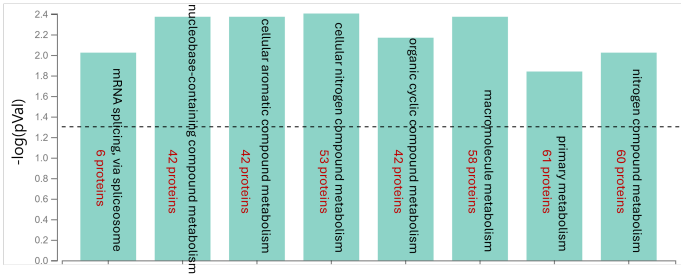
We can highlight proteins-of-interest on the plot such as collagens (yellow stars) and transcription factors (cyan triangles). Note the gap between transcription factors associated with tumour specific tissue (towards upper right) vs. stromal-tumour and healthy tissue (towards lower left), implying spatial specificity in tumour

development. Note how the yellow stars tend towards the left, as they are greater in the stromal-tumour mix compared to tumour tissue. The lower left corner is enriched in Stromal proteins involved in cell attachment, cell migration and the extracellular matrix, so could be a useful source of diagnoistic and prognostic markers in addition to collagen.

The upper-right corner of Figure 12 contains a deeply fascinating array of proteins. These are the most specific to the tumour microenvironment and tend to be low- or very low-abundance under healthy conditions. Aligned on the diagonal are proteins present only in tumour samples. Most proteins locate to the nucleoplasm and are likely involved in malignant reprogramming. A few components of the mitochondrial electron transport chain, secretory/trafficking pathways, immune evasion and cell release hint at other indispensible steps at the tumour core.

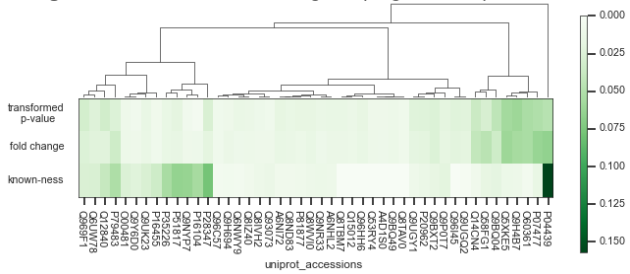
We can capture the general profile of altered protein expression in the ovarian tumour environment by conducting a GO-enrichment analysis of middle-top proteins in Figure 12. GO-compass enables fold-changes from differential abundance analysis to be incorporated into agglomerated GO term analyses. Results show how synthesis- and energy-requiring metabolic processes are significantly enhanced compared to the healthy ovarian proteome (Fig. 13).

Figure 13 Significantly increased GO-terms in GO-Compass analysis of top-middle proteins in Figure 13.



For researchers looking for their next target proteins there's a lot of candidates to choose from. We can visualise the current state of 'known-ness' of each potential target using information from the Unknomne database (Unknome). Researchers may be interested in well-researched targets or want to pursue a higher-risk, novel target. The clustermap in Figure 14 shows proteins from the top-right corner of Figure 12 clustered by their transformed p-value, fold-change and degree of known-ness. The clusters towards the right of Figure 15 would be a good starting point for people interested in little-known targets showing sizeable, consistent fold changes in tumour tissue.

Figure 14 Hierarchical clustering of top-right corner proteins



Summary and next steps

In this article we've demonstrated good practices for filling missing values and scaling datasets. These are essential steps before conducting any form of differential abundance analysis. We've shown how to evaluate

data losses during comparisons of unequal datasets and we've shown how, by a careful process of selection, normalisation and scaling, we can bring in external data to mitigate these losses. This has revealed hundreds more proteins that would otherwise have been hidden from view, many of which have known diagnostic, prognostic and therapeutic importance. By performing a multi-way comparison of healthy, tumorous and mixed stromal-tumour tissue, we can connect spatial location with altered abundance of certain proteins.

This has given us far greater functional insight into an ovarian proteomics dataset that we would ever have achieved using simple volcano plots of shared proteins. It also sets up a framework for wider comparisons, such as between breast and ovarian cancer datasets. With appropriate data-handling in place, there is a wealth of functional insights to be gained from public proteomics datasets, which can hugely enhance the conclusions drawn from de novo experimental data.