**Re-analysis of data from PXD003164**

**Background:**

Data is from: Bayram, H.L., Claydon, A.J., Brownridge, P.J., Hurst, J.L., Mileham, A., Stockley, P., Beynon, R.J. and Hammond, D.E., 2016. Cross-species proteomics in analysis of mammalian sperm proteins. *Journal of Proteomics*, *135*, pp.38-50. (<https://www.sciencedirect.com/science/article/abs/pii/S1874391915302268>)

Sperm samples were collected from six rodent species (12 samples) and 13? ungulate species (18 samples). Collected samples were processed similarly across species for proteomic analysis. The tryptic digests were run in a single shot 90-minute method on an LTQ Orbitrap Velos instrument. The data deposited at PRIDE (PXD003164) was downloaded for analysis with the PAW pipeline (<https://github.com/pwilmart/PAW_pipeline>).

**Re-analysis details:**

The PAW pipeline is freely available to process results from Comet database searches. The pipeline uses open source tools MSConvert (<http://proteowizard.sourceforge.net/>) and Comet (<http://comet-ms.sourceforge.net/>) for upstream steps. The downstream steps (target/decoy PSM filtering, protein inference and grouping, and quantitative summarization) are done with Python scripts.

A key aspect of a multiple species study is getting species-specific protein FASTA files. To have consistent FASTA files, the canonical reference proteomes available at UniProt (<https://www.uniprot.org/downloads>) via FTP were used. The “UniProt\_reference\_proteome\_manager.py” script (from <https://github.com/pwilmart/fasta_utilities>) was used to fetch the FASTA files for the Comet searches. FASTA files for human, cow, deer, horse, mouse, pig, rat, and sheep were used as peptide/protein identification screens for the 30 samples. These canonical FASTA files are built using ortholog relationships to have roughly a one gene produces one protein structure. They have similar numbers of protein sequences (around 20K).

Other than varying the protein FASTA files, the other search setting and post processing were done similarly. Wide parent ion tolerance searches of fully- and semi-tryptic peptides with fixed alkylation of Cys and variable oxidation of Met residues were used. The target/decoy method was used to accept PSMs at a 1% FDR. Protein inference required two peptides per protein. An additional protein grouping step was used. The numbers of PSMs at 1% FDR was found to be more relevant for the species screening.

**Sample key determination:**

There was a sample key file in the PRIDE archive. It did not seem to match up with the samples and species described in the publication. Private communication with Dr. Rob Beynon suggested an alternative key (associated with gels). The different FASTA files could be used to screen the samples and see which key was more plausible. Assuming there would be more PSMs identified when the correct species (or at least a closer species) FASTA file was used, it became clear that the key in the PRIDE archive was incorrect and that the alternative key was the right one.

**Some data patterns:**

Human was used as a related species that would be reasonably close to rodents and ungulates to establish a base line for numbers of identified PSMs and proteins. One way to compare the 8 FASTA files for each sample would be how much gain each species database made in PSMs or proteins. This could be done as percent gains or as fold-changes. Those comparisons would be more normalized. Alternatively, a simple starting comparison is just the raw number of identified PSMs. This will vary more sample-to-sample due to sperm quality (road kill samples are not expected to control for variability like lab strain animals). Numbers of proteins identified could also have been compared; however, those numbers are moderated and make the changes more muted.

The species/samples were put into eight groups: cows, deer/oryx, pig, sheep, mouse/vole, rat, zebra, and squirrel.



**Figure 1.** Cow and cow-like species. The y-axis is the number of PSMs identified at 1% FDR. The x-axis has the 5 samples blocked by FASTA file. Sample key is on the bottom.

Cow and closely related species have maximal PSMs for the bovine database. Sheep PSM numbers are also high, indicating cow and sheep similarity. The other FASTA files all have similar, lower PSM numbers. Sample-to-sample PSM numbers are variable. The pattern, however, is very similar for all FASTA files except bovine. Orange and grey bars are at similar numbers with the bovine FASTA file, and the gold bar (the domestic cow sample?) is higher than the light blue bar.



**Figure 2.** Deer and oryx species. The y-axis is the number of PSMs identified at 1% FDR. The x-axis has the 5 samples blocked by FASTA file. Sample key is at the bottom.

There is one canonical FASTA file for the European red deer available. We see that bovine and sheep FASTA files resulted in more identified PSMs than the deer FASTA. The deer and oryx species may be close to cows. It appears that the deer FASTA did a little better than human and rodent databases. The Alfred’s Deer sample did seem to benefit more from using the deer FASTA file than the other 4 samples. The gain from the deer FASTA file seems minimal.



**Figure 3.** Sheep samples. Axes are similar to previous plots.

Sample-to-sample consistency of identified PSMs is better for the sheep samples. We see that the sheep FASTA file maximizes the identifications. The similarity of sheep and cow is also apparent with the bovine FASTA file giving considerably more identifications than the other 6 FASTA files. Those other six FASTA files give very similar numbers of identified PSMs.



**Figure 4.** Pig (and closely related species) samples. Axes are similar to previous plots.

The pig FASTA file greatly increases the number of identified PSMs compared to all of the other species. The other 7 FASTA files (species) are fairly similar in PSM numbers. This hints at what happens when related species protein databases are used as a proxy for the species of interest. We need to be thinking at the tryptic peptide level for sequence similarity arguments. There are probably a core set of peptides that are homologous across the species (including pig). A more detailed analysis of the identified peptide sets would need to be done to verify this. The increase in PSMs for the pig FASTA file are the unique peptides that can only be identified when using the correct species database.

How do we get identical tryptic peptides across species? There can be highly conserved motifs that persist across many species (zinc-finger proteins are an example). Those are relatively rare, though. (Enzyme binding pockets, for example, may be conserved via specific residues that are usually not in the same peptide). Most identical peptides will come from homologous proteins where only some resides differ, giving a mix of identical and non-identical (although some could be very similar) peptides. When we have the correct protein sequences in the FASTA file, then we can recover the non-identical peptides and we see large gains in PSMs.

What happens at the protein level? Proteins are inferred from the peptides using the FASTA file protein sequences as scaffolds. It is possible to have the identical peptides in a related species (which are only a subset of the true protein’s peptides) be inferred as a related protein rather than the true ortholog. Species may create functional specialized proteomes (like sperm) using slightly different “parts lists”. One species may use a protein that is not used in a related species for the specialized proteome. A similar protein may be used instead that has enough homology to be incorrectly inferred. Comparing inferred proteins between species can be tricky in situations like this. There can be artifactual “scrambling”, even for abundant proteins.



**Figure 5.** Mouse/vole species. Axes are similar to previous plots.

The three woodmice samples vary a lot in their numbers of PSMs. The vole samples are more consistent. We do see some modest increases in identified PSMs for the mouse FASTA file. The rat FASTA file is nearly the same for PSM numbers as mouse. The rodent databases do not increase the PSM number as much compared to the other species as we saw above.



**Figure 6.** Rat samples. Axes are similar to previous plots.

The rat samples are quite consistent sample-to-sample. We see a very different pattern compared to the mouse-like species. We have a large gain in PSMs when using the rat FASTA file. There is a smaller gain for the mouse FASTA file compared to the other 6 species. The other 6 are similar (except for deer). We see more gains using the rodent databases for the rat sperm peptides than we had for the mouse/vole sperm peptides. We also have an asymmetry where rat sperm seems to differ more compared to mouse/vole sperm than vice versa. It might be okay to assume that the mouse and rat FASTA files are of high quality given the frequent use of these lab animals. This could suggest some biological differences between mouse/vole and rat sperm.



**Figure 7.** The zebra sample. Axes are similar to previous plots.

We do not have a zebra FASTA file, so horse was used. We see a factor of 2 more PSMs with the horse FASTA file compared to the others. This pattern is similar to what we saw with pig.



**Figure 8.** The squirrel sample. Axes are similar to previous plots.

The squirrel sample does not see gains with any of the 8 FASTA files used. There is a 13-lined ground squirrel canonical database that could be tried.

**Summary:**

Using a span of 8 canonical reference proteomes, the 30 samples could be characterized by species well enough to resolve the sample key question. The deer FASTA file seemed to result in lower numbers of identifications in most of the plots above. We also saw minimal gains when using the deer database for the deer-like samples. This was in contrast to pig or rat. This raises the question of the quality of the deer database.

Some evolutionary distances can also be seen. Cows and sheep seem to be quite similar. Horse and pig seem to have similar, larger distances to the other FASTA file species. There seems to be something different about rat and mouse. It is not clear if the differences in PSM patterns are due to the FASTA files themselves or the peptides in the samples. We do have lab mice and rats, so some additional experiments could be done to figure this out.

-Phil Wilmarth, OHSU, 20201017