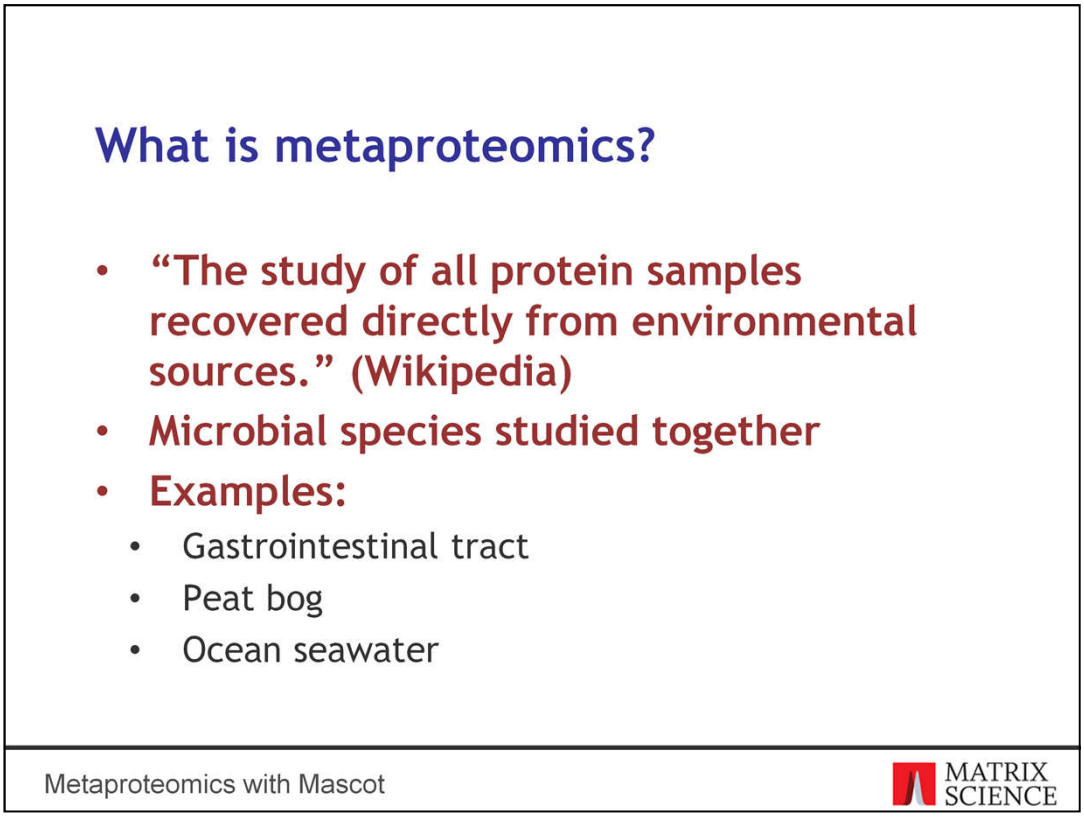
Metaproteomics with Mascot

Ville Koskinen
Matrix Science

MATRIX SCIENCE

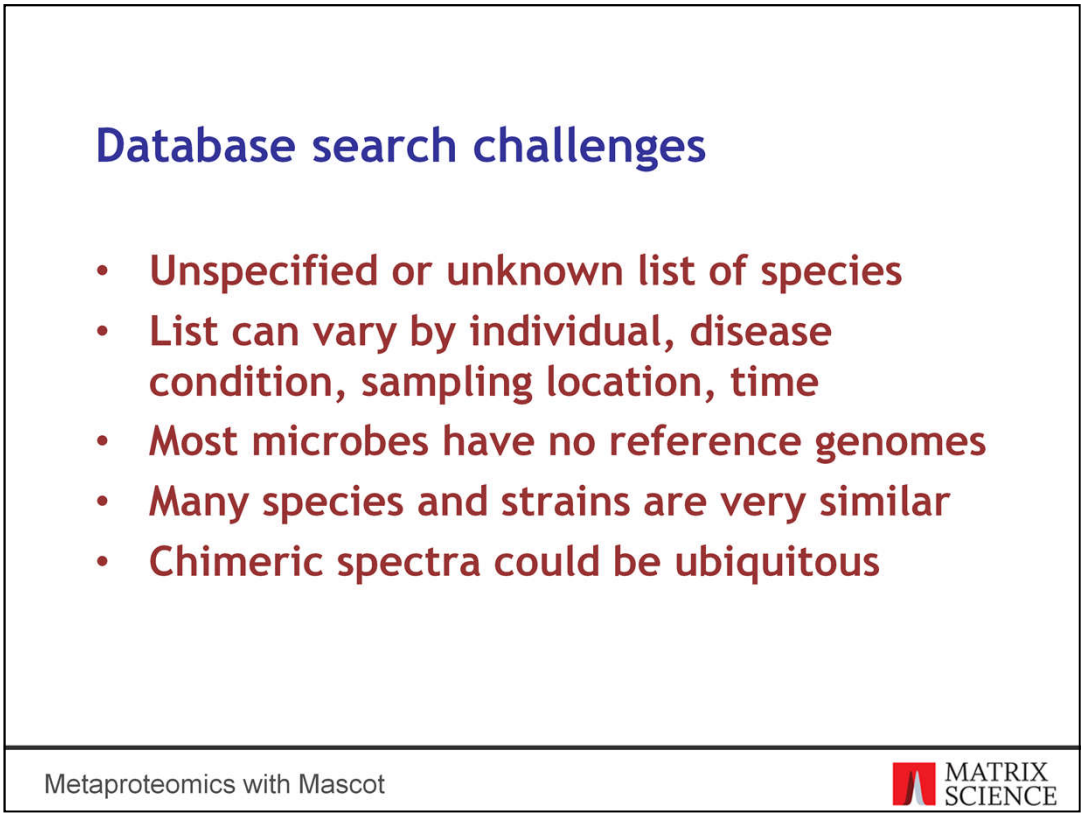


Metaproteomics has a few different definitions. This one is from Wikipedia: the study of all protein samples recovered directly from environmental sources. An example experiment is characterising microbial function and gene expression in the gastrointestinal tract. For example, you may be interested in which bacteria are present and what roles they play in food digestion.

Another example is studying carbon metabolism in peat bogs. These kinds of experiments compare microbial activity at different depths in the bog and at different times.

Yet another example is characterising the microbial content of ocean seawater.

Common to all of these is that you cannot isolate individual microbial species from their environment. The species have to be studied together. This is in stark contrast to single-species studies or model organisms.



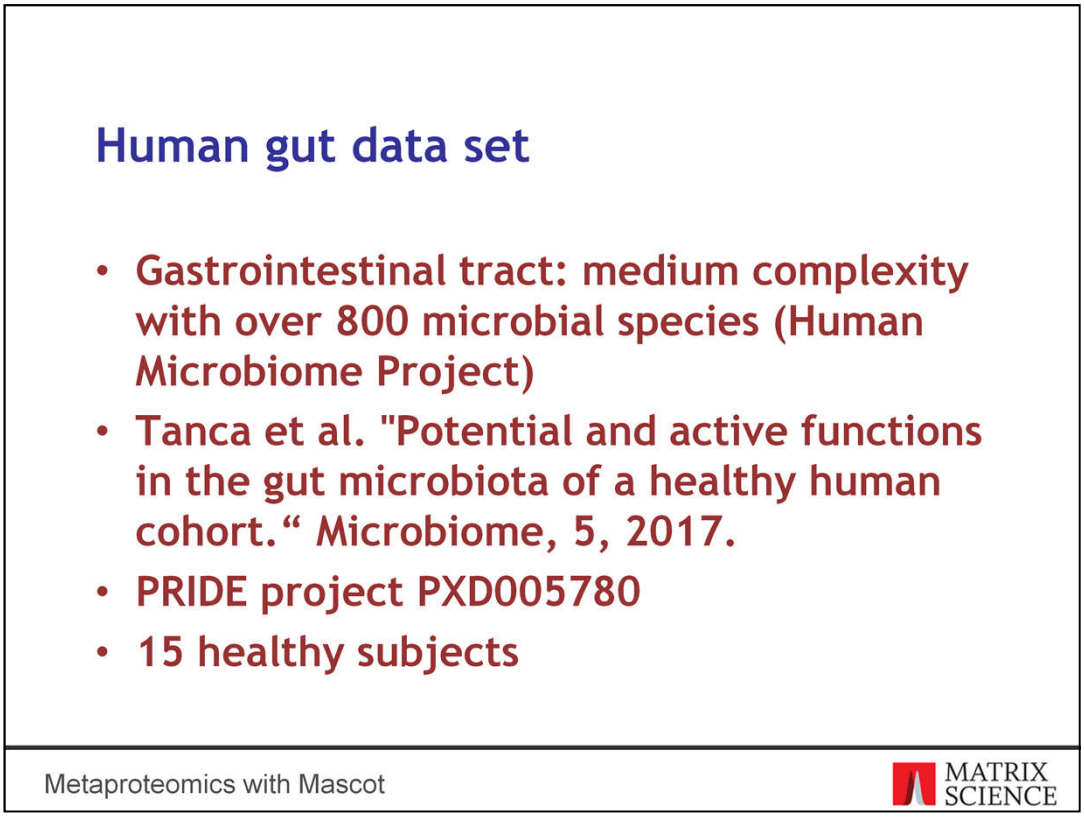
The first and the most important step in any metaproteomics experiment is reliable protein identification, and that's the focus of my presentation today. Database searching is the best technology for the task, but there are several unique challenges.

First of all, there may be no way to give a definitive list of all microbial species in your sample, even in principle. Consider gut bacteria: the microbial species could vary by individual, by diet, by disease condition or by time. Or consider peat bogs: the species could vary by sampling location or depth, temperature, season, soil acidity and many other factors. Discovering and characterising these differences are typically one of the goals in a metaproteomics experiment. But from a database searching perspective, it means your search space has fuzzy, badly defined boundaries.

Another major difficulty is that most microbes have no reference genomes. Of course, more and more genomes are being sequenced all the time, but if you're entering a brand new area, proteomics will at first be very difficult. You can't identify the unidentifiable.

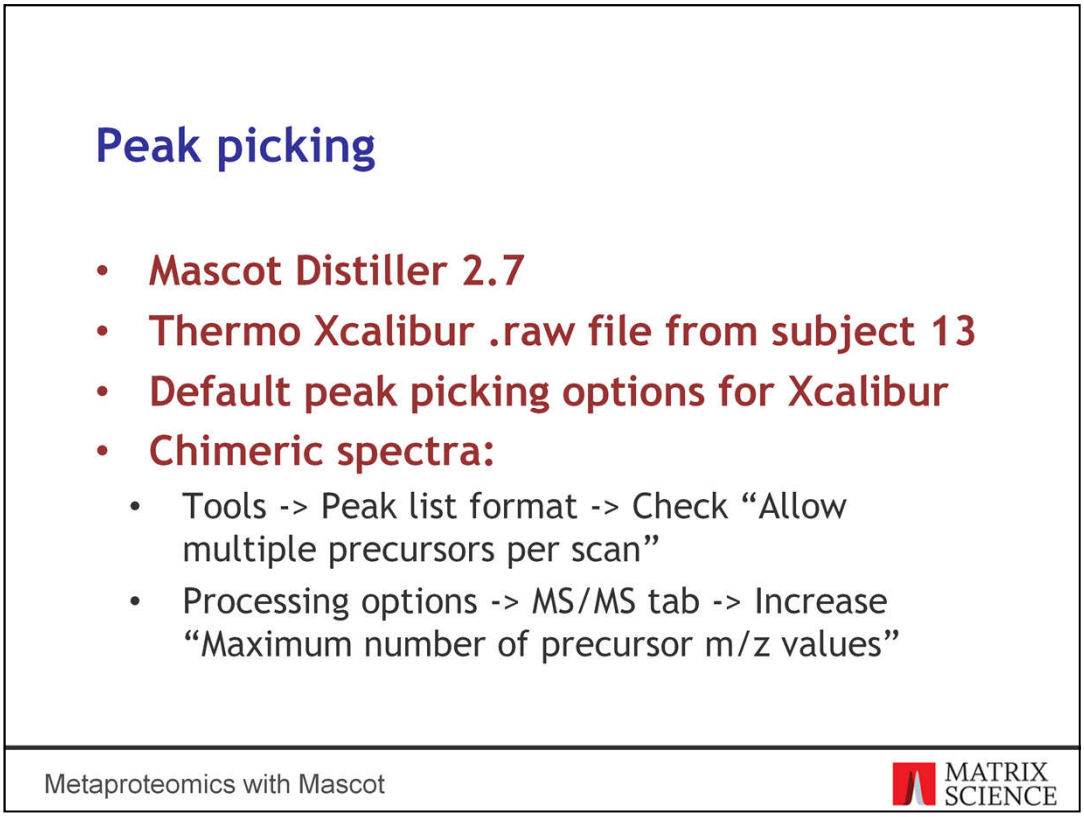
Adding to the ambiguity is that many bacterial species and strains are very similar to one another. It's both a blessing and a curse. The advantage is, you may be able to identify proteins by using the genome of a related species, even if the true genome is unavailable. On the other hand, the more homologous genomes you add to the search space, the harder it becomes to classify any given protein hit.

Lastly, chimeric spectra could be much more common in metaproteomics samples than in single species studies. This is because there is bound to be lots of peptide sequence homology between species as the number of species increases, which means increasing numbers of precursors within mass tolerance.



I’ll illustrate the database search challenges and ambiguities with a human gut data set, although the points apply to all metaproteomics experiments. The gastrointestinal tract could be considered a medium complexity microbial environment. The Human Microbiome Project is sequencing and cataloguing reference genomes from multiple human body sites. In May 2019, the project catalog has 823 genomes to be sequenced in the gastrointestinal tract, and about half of the species have at least a draft genome available.

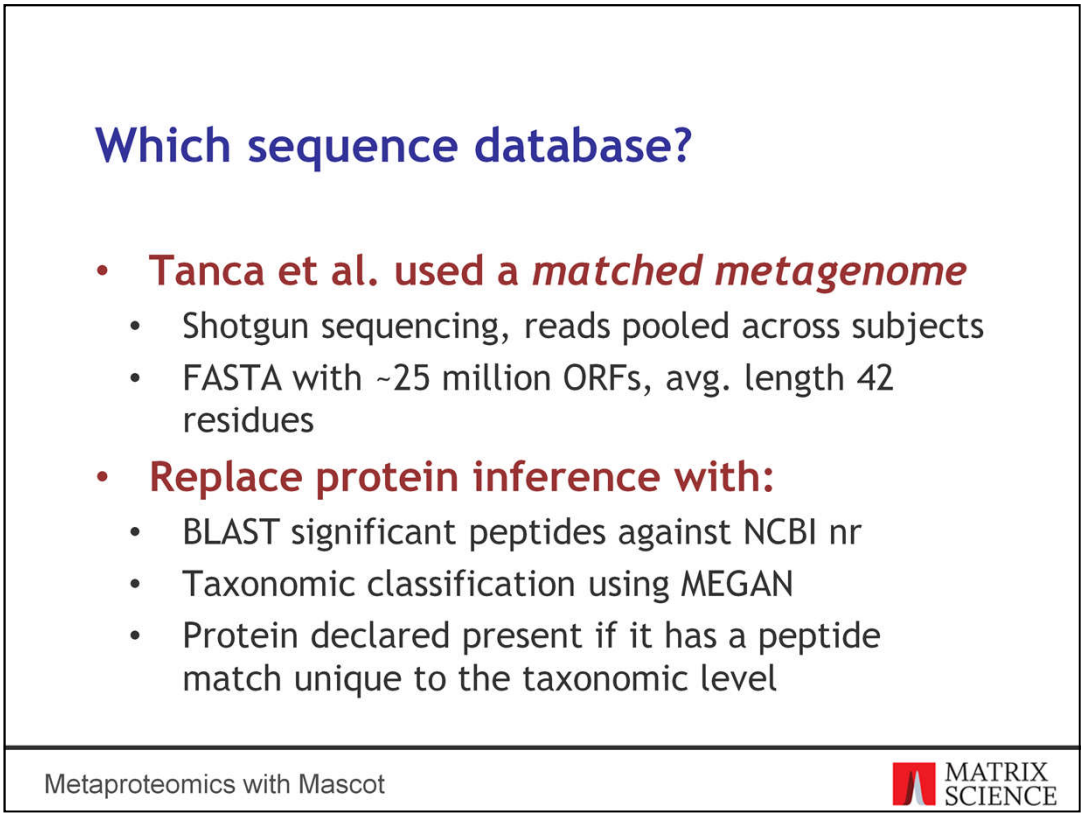
I’ve chosen a gut microbiota study available in the PRIDE repository. The study contains 15 healthy subjects from which stool samples were collected. The study goal is to compare the differences in functional potential at gene level and actually expressed functions at protein level. Of course, protein function analysis is strongly dependent on reliable protein identification.



Peak picking was done in Mascot Distiller 2.7. The raw data is in Thermo Xcalibur format from an LTQ Orbitrap Velos instrument. I used the default peak picking options, and I selected one raw file arbitrarily, from subject number 13.

I mentioned chimeric spectra as a potential problem. The reasoning is, when you're adding more and more genomes, there's bound to be increasing numbers of homologous peptide sequences. Some of these will have very close masses, even if precursor tolerance is tight.

Enable chimeric spectra detection by checking this box under Tools and Peak list format. You also need to increase the maximum number of precursor m/z values in the MS/MS processing options. A reasonable maximum is 5. If you set it too high, Distiller will likely just find different charge states of the same precursor.

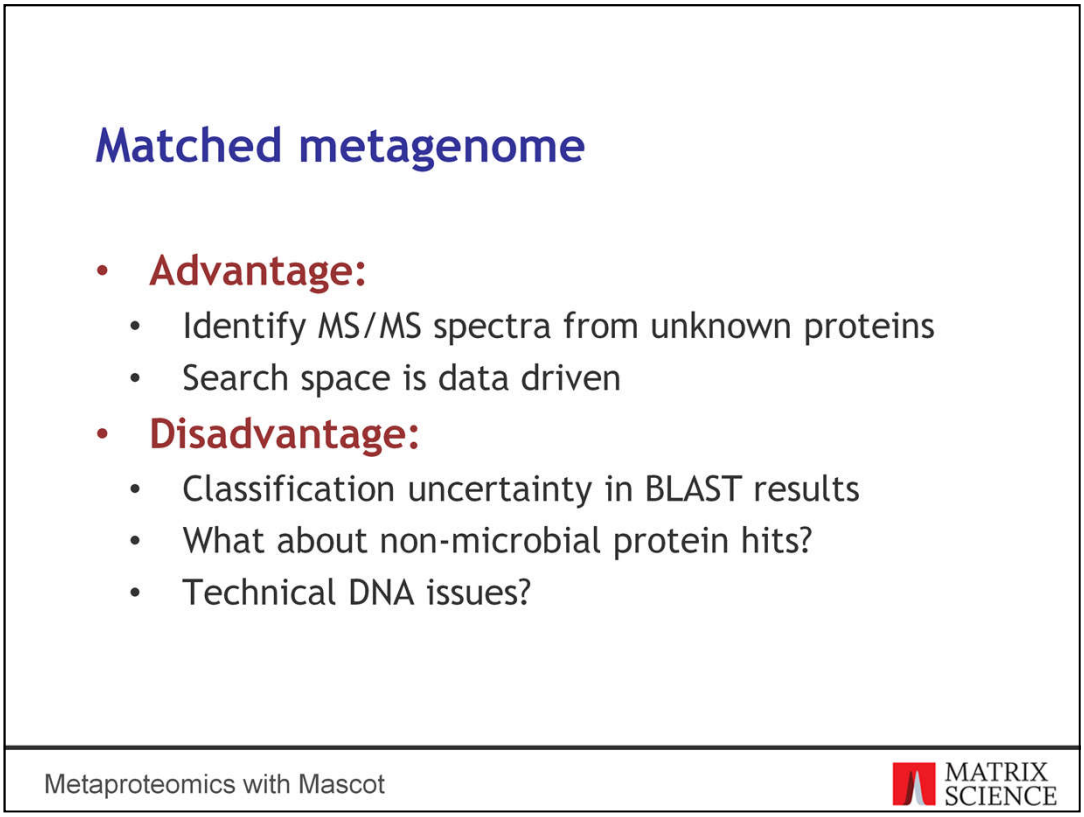


Now, let's talk about sequence databases. It's vital to spend time on database design and do it right. Biological conclusions will depend on it. Metaproteomics experiments are even more sensitive to this than single-species studies, because the full list of target species is rarely known with certainty.

Tanca et al. followed a fairly common practice, which is to create a matched metagenome. This means shotgun sequencing the sample, possibly assembling into contigs and translating into proteins. The reads were pooled across subjects to account for individual variation. The FASTA file contains about 25 million open reading frames. The authors decided not to assemble the reads into contigs, although this would normally make annotation easier.

Next, search the MS/MS spectra against the matched metagenome. Take all the significant peptide matches and BLAST them against NCBI nr to map onto proteins. The final step is to use a metagenome annotator software called MEGAN. The software classifies the BLAST results using a lowest common ancestor algorithm. For example, if a sequence is shared between species in the same genus, it's assigned to the genus level. If it's shared between several genera, it's assigned to the family level, and so on until no ambiguity is left.

The matched metagenome approach has some advantages, but it comes at a cost.

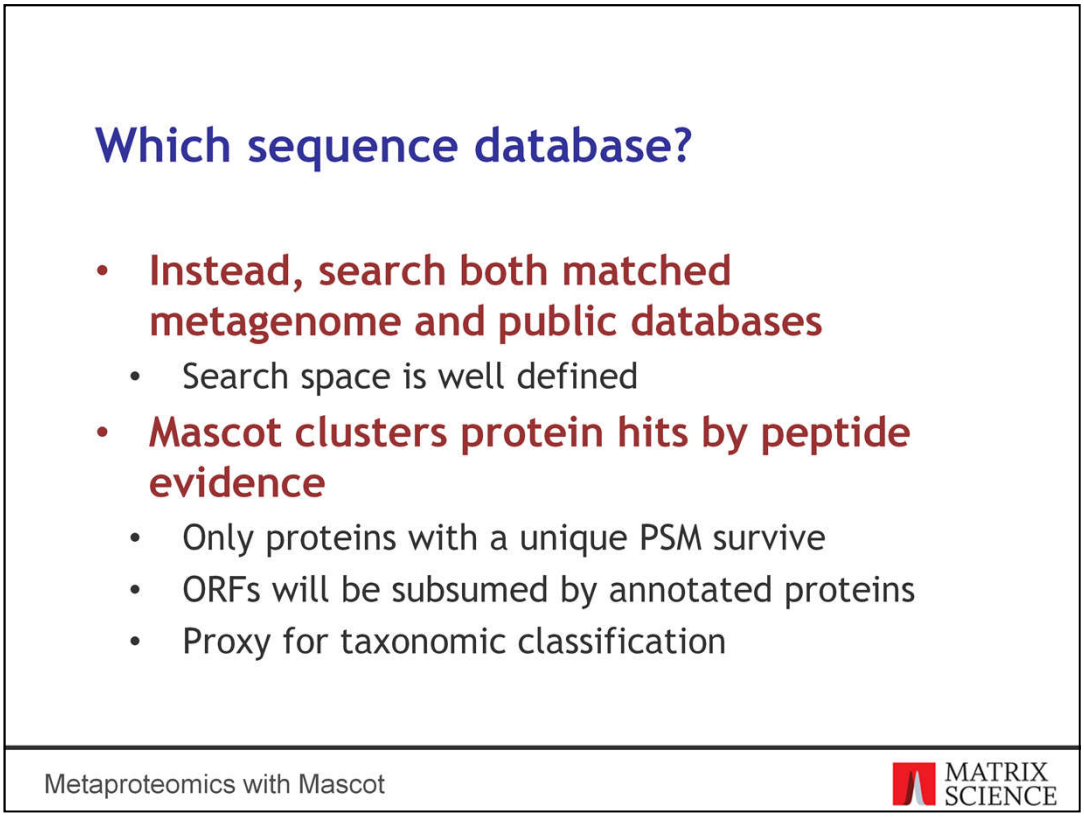


In theory, the matched metagenome maximises the number of identified MS/MS spectra. This is because the sequence database is built directly from the sample, so it contains all observable sequences and nothing extra. If a spectrum fails to match, it's probably not generated by a peptide. There's also no need to agonise over species selection; the taxonomic classification is driven by data.

There are some obvious downsides. A BLAST search of a peptide could bring up a dozen perfect matches, all with similar and highly significant E-values. Often only the top hit is chosen as the correct one and the others are ignored. This ambiguity is rarely present in downstream analysis.

Annotation relies on public sequence databases. You might get more peptide matches, but if those peptides fail to map to a protein sequence with reasonable homology, you haven't gained any additional information. If they map to non-microbial proteins, what should you do? Maybe keep them if they map to human proteins and filter out otherwise. Whatever you do, you can't avoid making taxonomic decisions.

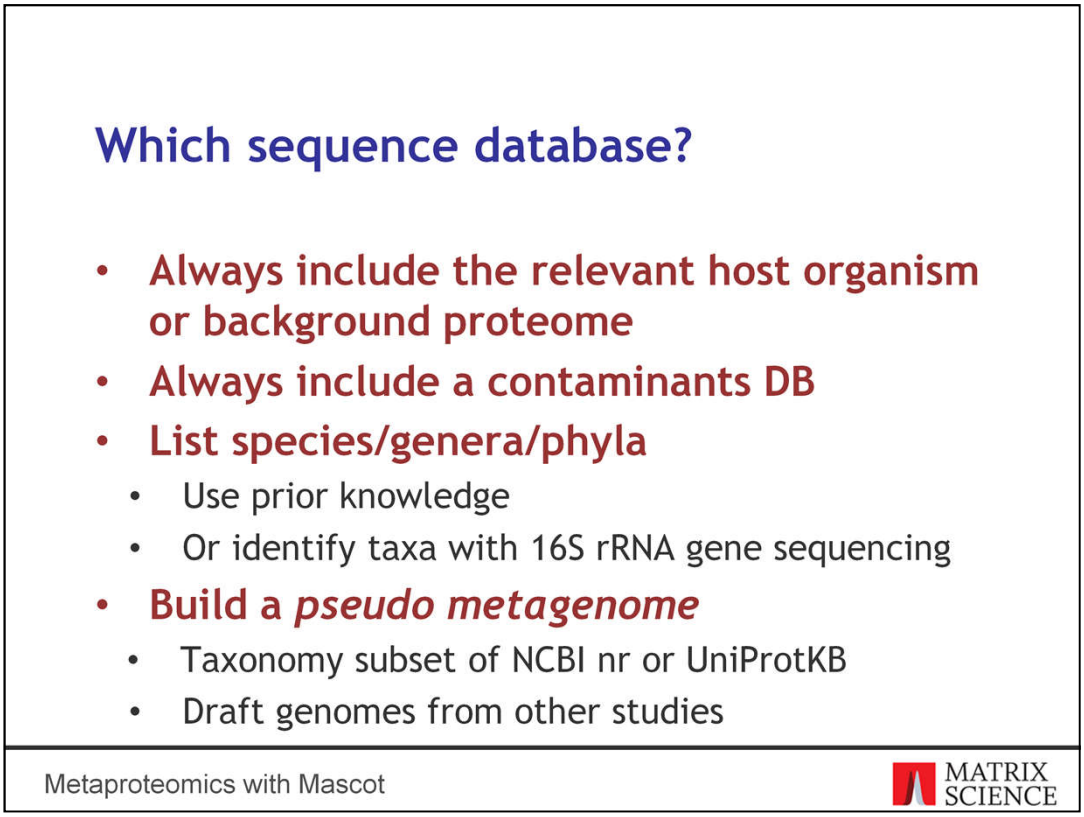
Finally, there may be technical issues in DNA extraction or sequence assembly that prevent creating a complete search space. Sequencing has a low but non-negligible error rate, and it's not guaranteed to provide 100% coverage.



Here's a better strategy. Define the search space as fully as you can using publicly available databases, based on prior knowledge about the microbial environment. Augment the search space with the matched metagenome. Public databases will give matches to peptides shotgun sequencing failed to predict, while the matched metagenome will give matches to previously unknown peptides. This way the search space is well defined.

The other reason for a combined search is to let Mascot protein inference take care of the ambiguity. Mascot clusters protein hits based on shared and non-shared peptide matches. Only proteins with at least one unique, significant peptide match survive. Any ORF that is a subsequence of a database protein is very likely to be subsumed by it, which will provide an explanation for most ORFs. ORFs with unique PSMs may survive clustering if the public databases are not complete.

Protein clustering is a proxy for taxonomic classification. I'll show you an example in a few slides.



Here's how you can create a suitable database.

Always include the host organism. In this case, we have a human gut sample, so include the UniProt human proteome. You have to account for non-microbial proteins and enzymes that are caught up in the stool sample as it passes through the digestive system. If you don't, MS/MS spectra of human peptides could remain unidentified at all or worse, they could be misidentified as bacterial.

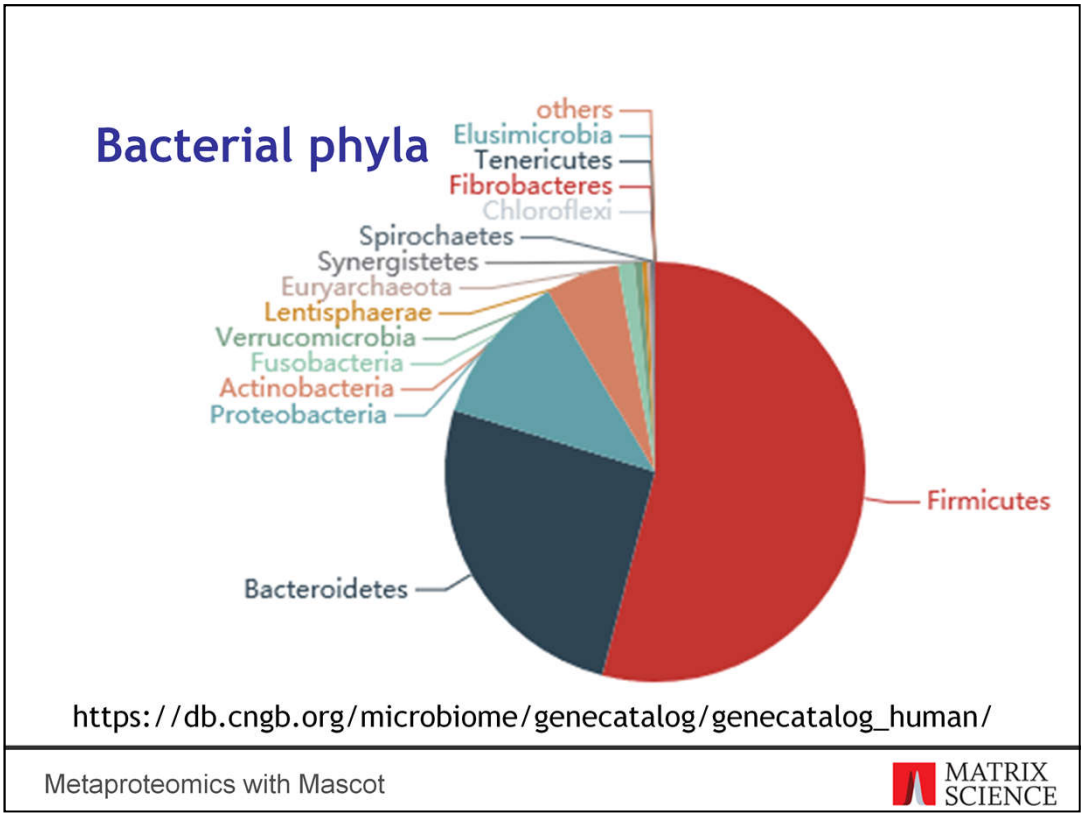
Always include a contaminants database. It needs to contain your enzyme and the usual suspects like keratins. You probably shouldn't include the contents of the last meal the subject had when you consider stool samples. However, if your sample was collected directly from the small intestine and contains partially digested matter, you should include those in the search space.

Of course, in a peat bog or ocean seawater sample, the microbes aren't embedded in another organism, so there's no relevant background proteome. But you still need to account for known contaminant proteins.

Next, define the target search space by making a so-called pseudo metagenome. There are several ways of doing this. You could identify taxa with 16S rRNA gene sequencing, and use it to create a taxonomy subset of NCBI nr or UniProt. Some or even most of the identified species could be poorly represented, in which case you should go higher up in

taxonomy to genus or phylum level.

An alternative is to build on the work of others. If you're studying a well-characterised environment, like the human gut, there could already be draft genomes for most of the microbial species. There are ocean metaproteomics repositories with the same goal.



This is a pie chart of the bacterial phyla identified in the human gut. The graph comes from the China National GeneBank website, based on an integrated catalogue of reference genes. The 14 phyla shown here cover nearly all the bacterial genes. There is also a genus composition chart, but a third of the genera are in the category “others”.

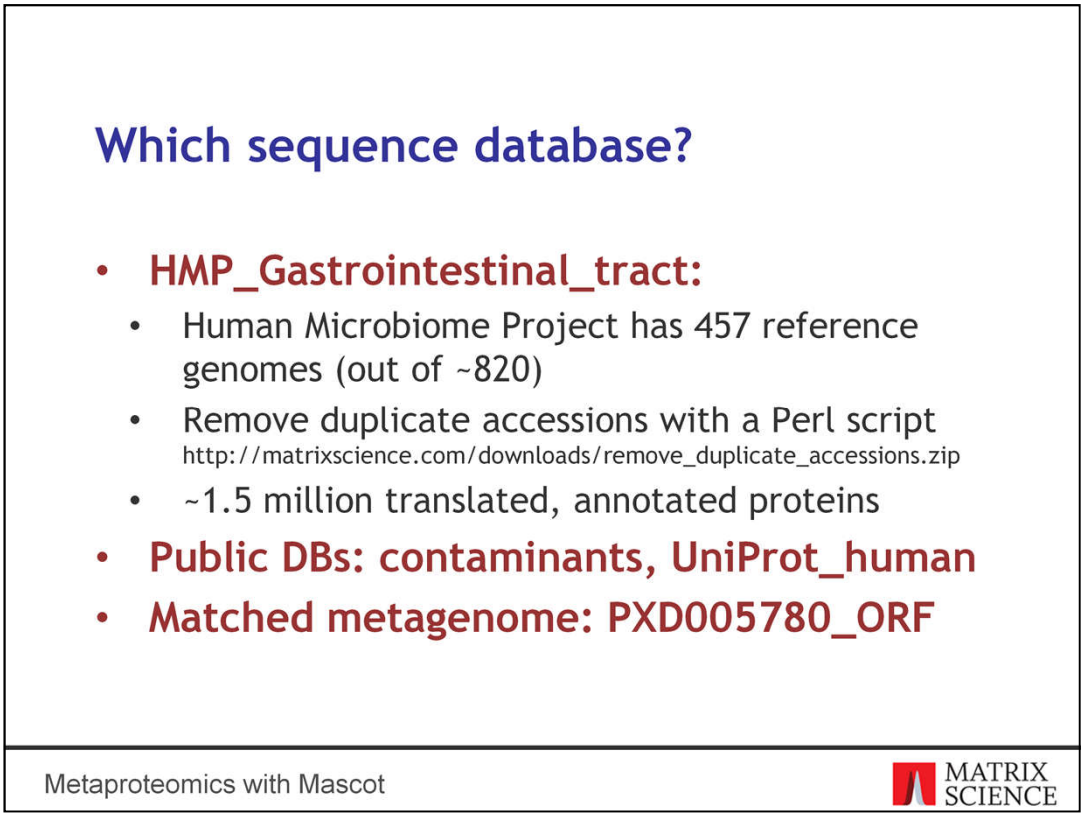
Bacterial phyla in human gut

Phylum	UniProtKB sequences	Phylum	UniProtKB sequences
Proteobacteria	50,527,470	Spirochaetes	930,291
Firmicutes	18,969,713	Tenericutes	264,456
Actinobacteria	17,030,783	Fusobacteria	248,945
Bacteroidetes	8,905,014	Fibrobacteres	102,884
Euryarchaeota	2,268,852	Elusimicrobia	150,312
Chloroflexi	1,210,494	Lentisphaerae	103,265
Verrucomicrobia	1,061,507	Synergistetes	82,921

- FASTA would be 40.5 GB
- Only use as last resort

UniProtKB contains millions of reviewed and unreviewed sequences for the 14 phyla. The total size of the FASTA file, if you download it, is about 40.5 GB. This is a vast search space, and it's not far off from searching the whole of NCBI nr. You could make the database more specific by including sequences at genus level, which is some hours of work.

I'm only showing these counts as an example of what you may need to do if there are no alternatives.



For the present data set, I selected the 457 available gastrointenstinal tract genomes from the Human Microbiome Project. Some of the genomes are at draft stage, and obviously not all of the known genomes have been sequenced yet. It might not be the perfect representation of all human guts, because these will vary by population, diet and so on. But as you'll see shortly, it's far better than the matched metagenome alone.

I had to write a small Perl script to remove duplicate accessions. The final FASTA file has almost 1.5 million annotated protein sequences. If you face a similar problem, you can download the script from our website.

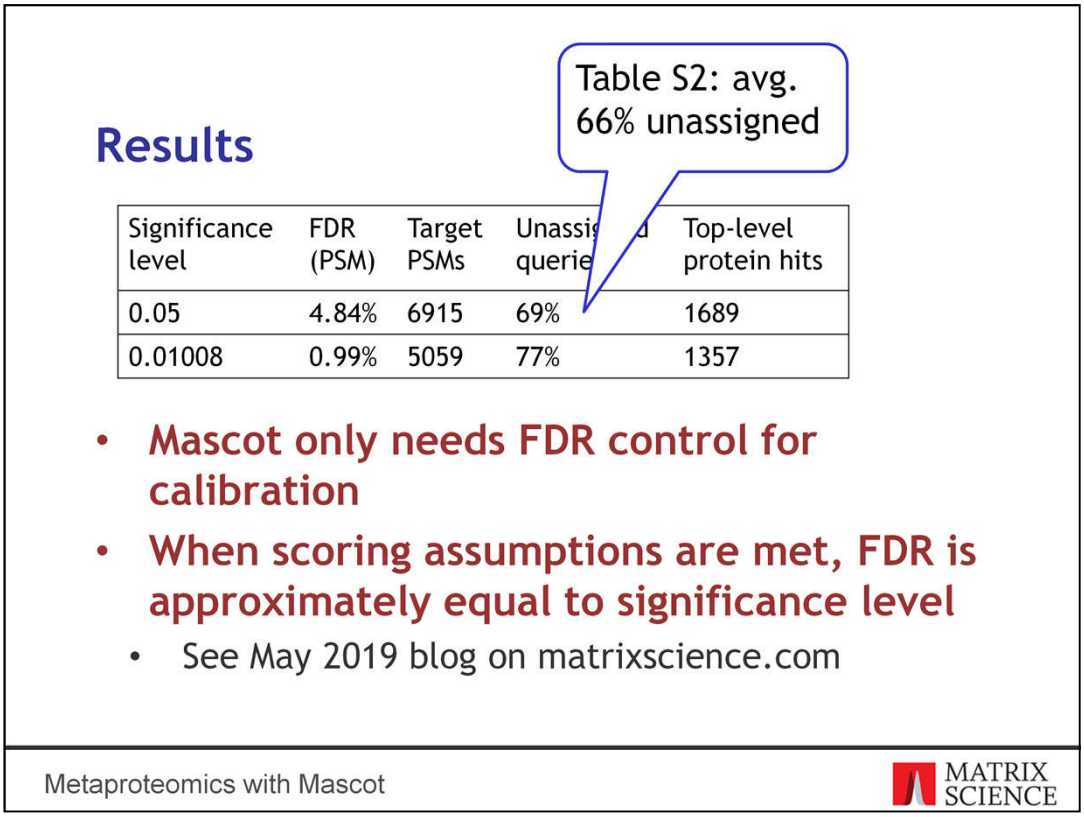
I also selected the standard contaminants database and UniProt human proteome, and I set up the matched metagenome of ORFs in Database Manager.

Other search parameters

- Fixed mods: Carbamidomethyl (C)
- Variable mods: Oxidation (M)
  - Leave out all but essential variable mods
- Precursor tol 10ppm, fragment tol 0.02Da
- Instrument: ESI-TRAP (LTQ Orbitrap Velos)
- Check Decoy box for FDR

Here are the other search parameters. Leave out all but essential variable modifications. You should only choose modifications that are abundant in most of the proteins in most of the species.

Precursor and fragment tolerance are the same the authors used.



Once the search is finished, false discovery rate at the default significance level is about 5%. Adjusting FDR to 1% gives 5059 significant matches and over 1300 top-level protein hits. The proportion of unassigned queries is typical for shotgun proteomics data sets. In fact, if you look at table S2 in the supplementary materials of the original study, the average proportion of unassigned is 66% across the 15 subjects.

Is the FDR estimate accurate?

Mascot scoring is based on a statistical model. FDR control is only needed for calibrating the model if the search space and the data set violate the model assumptions. In large searches, when all the model assumptions are valid, the FDR is approximately equal to the significance level. As you can see here, the estimated FDR closely tracks the significance level, so it seems reasonable to accept the FDR estimates as accurate.

If you're interested in the statistics, have a look at the May 2019 blog article on our website.

Results

Significance level	FDR (PSM)	Target PSMs	Unassigned queries	Top-level protein hits
0.05	4.84%	6915	69%	1689
0.01008	0.99%	5059	77%	1357

- **Chimeric spectra?**

Spectra	17,440
Queries	22,188
Average precursors per spectrum	1.27

Metaproteomics with Mascot

MATRIX SCIENCE

There are 17,440 MS/MS spectra in this raw file.

Are there many chimeric spectra? Here's a quick comparison of counts of MS/MS spectra and counts of queries. If a spectrum has multiple precursors, Mascot splits them into separate queries. In this search, the average number of precursors per spectrum is 1.27. So, the majority of spectra are not chimeric, but there are enough to warrant enabling chimeric spectrum detection.

Results

Significance level	FDR (PSM)	Target PSMs	Unassigned queries	Top-level protein hits
0.05	4.84%	6915	69%	1689
0.01008	0.99%	5059	77%	1357

- **Search PXD005780_ORF only?**

Significance level	FDR	Target PSMs	Unassigned queries	Top-level ORFs
0.05	7.36%	3833	83%	2165
0.006178	0.99%	2832	87%	1630

Metaproteomics with Mascot

MATRIX SCIENCE

What would the results look like if we had searched only the matched metagenome?

Recall that the "ORF" database contains about 25 million short open reading frames. No matter what decoy algorithm you use, it's unlikely the decoy database will provide an unbiased estimate of false positives. Sure enough, the FDR estimates behave a bit more erratically. The proportion of unassigned queries is much higher as well, so it's not worth spending time on analysing the search in detail.

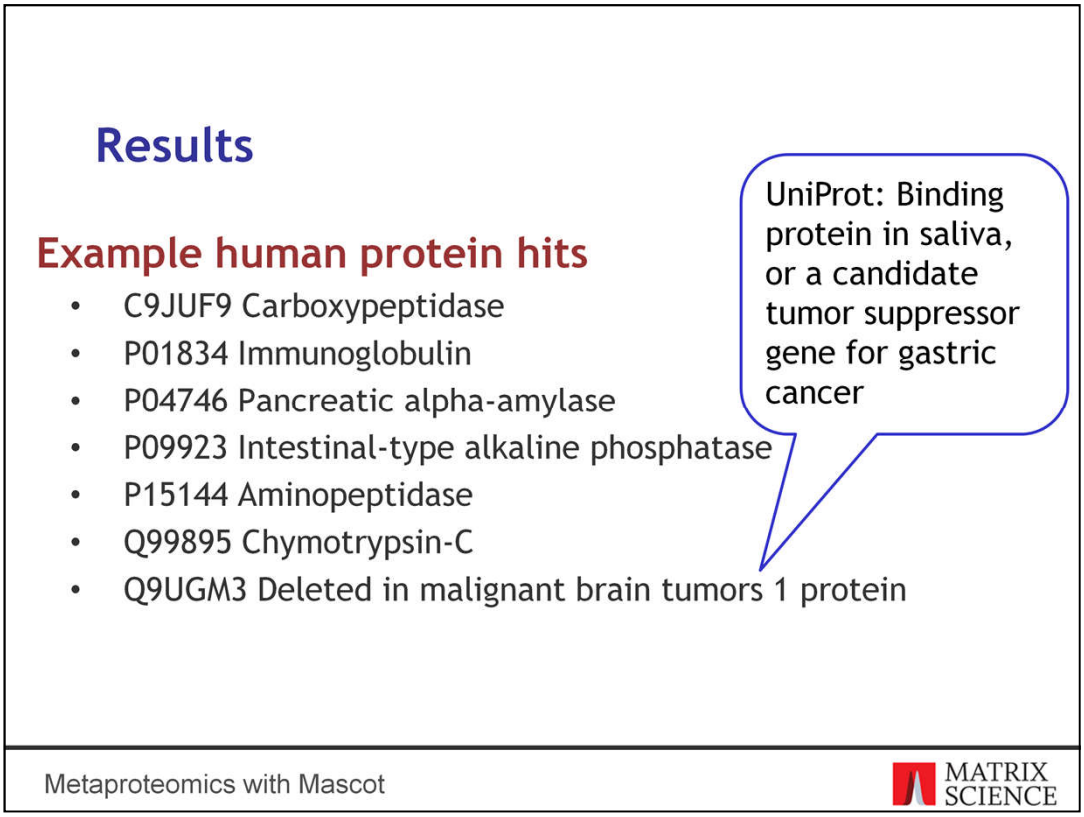
Results

Significance level	FDR (PSM)	Target PSMs	Unassigned queries	Top-level protein hits
0.01008	0.99%	5059	77%	1357

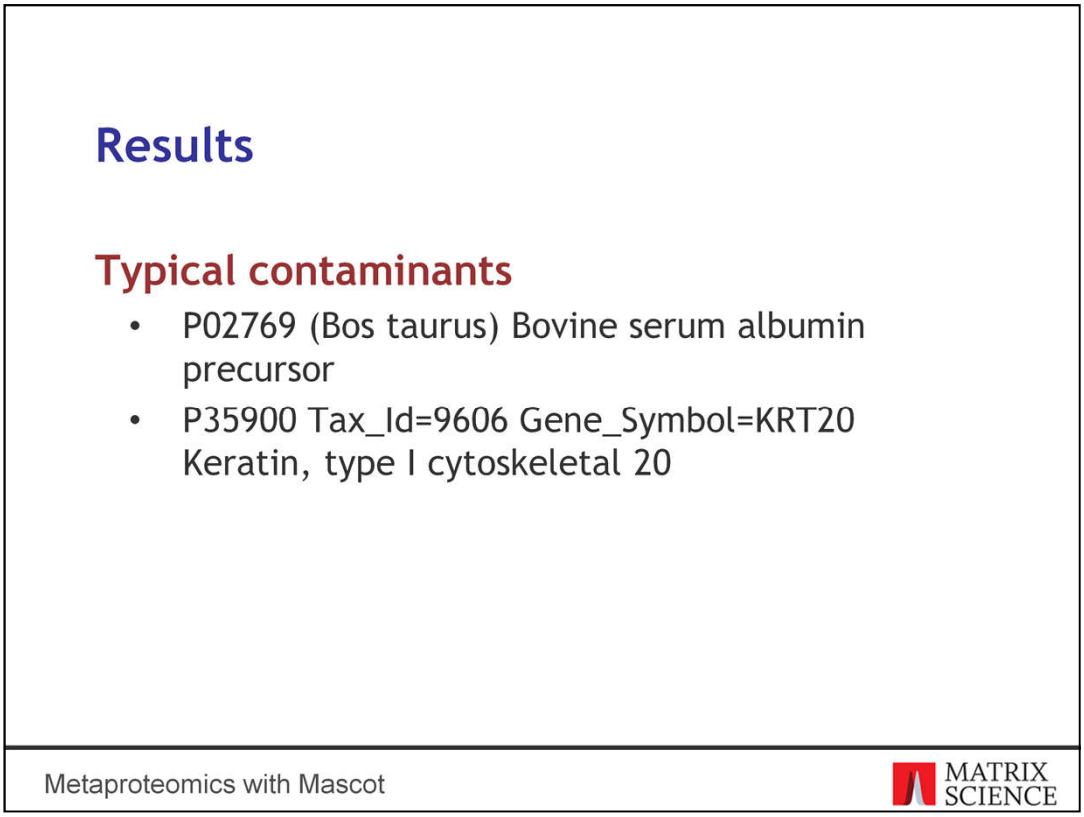
Database	Database entries	Top-level protein hits	All protein hits	Samesets and subsets
human	93,606	171	717	546 (76%)
contaminants	247	9	22	13 (59%)
gut proteome	1,468,777	416	9085	8669 (95%)
ORFs	25,328,860	761	57337	56576 (98.7%)

Let's concentrate on the search results at 1% FDR. This gives a confident set of peptide matches.

Here's a breakdown of the protein hits by database. The matched metagenome has the largest number of entries, so it's not surprising it gets the most protein hits. What should also not come as a surprise is the high number of human proteins identified. Let's look at some examples.



I can't list all 171 here but they all follow the same pattern. Namely, all of them can reasonably be expected to be present in the digestive system. The last one, Q9UGM3, seems a bit odd at first, so I looked it up on uniprot.org. It has many potential functions, such as a binding protein in saliva or a candidate tumor suppressor gene for gastric cancer.



The contaminant hits are the expected ones, like BSA and human keratin.

Results

Significance level	FDR (PSM)	Target PSMs	Unassigned queries	Top-level protein hits
0.01008	0.99%	5059	77%	1357

- **HMP_Gastrointestinal_tract is ~50% complete**

Database	PSMs	Peptide sequences	Unique peptide sequences
gut proteome	2547	1689	1038
ORFs	1271 (33%)	980 (37%)	873 (46%)

Now, let's concentrate on the bacterial proteins. The search has 5059 PSMs in total. Of these, 2547 appear only in bacteria in the gut proteome, and 1271 appear only in the open reading frames. In other words, 33% of peptide matches are there because we included the matched metagenome in the search space. It is clearly bringing in more matches. The story is the same if you look at counts of matched peptide sequences and counts of unique sequences.

Recall that the Human Microbiome Project FASTA file contains about half of the genomes expected to be present in the human gut. The counts of matches are perfectly in line with that. As more genomes are sequenced, the proportion of matches unique to the matched metagenome will reduce.

Results

Significance level	FDR (PSM)	Target PSMs	Unassigned queries	Top-level protein hits
0.01008	0.99%	5059	77%	1357

Database	Database entries	Top-level protein hits	All protein hits	Samesets and subsets
human	93,606	171	717	546 (76%)
contaminants	247	9	22	13 (59%)
gut proteome	1,468,777	416	9085	8669 (95%)
ORFs	25,328,860	761	57337	56576 (98.7%)

Metaproteomics with Mascot

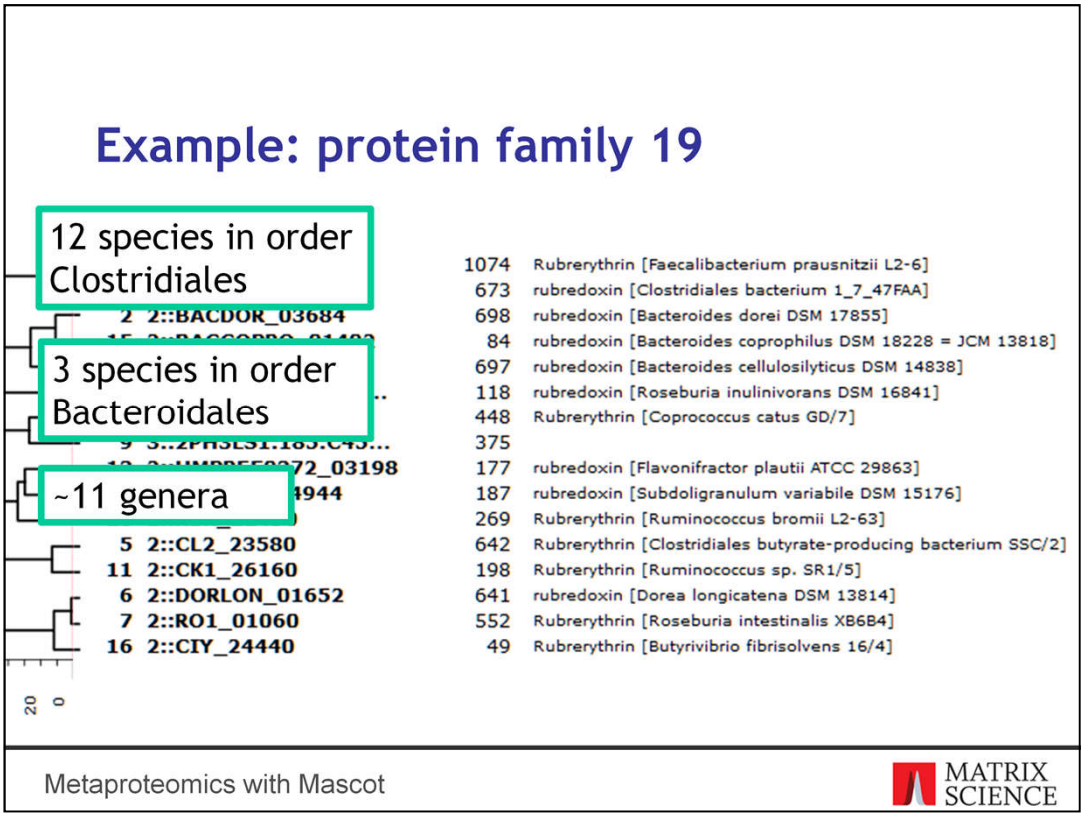
MATRIX SCIENCE

Coming back to the breakdown of protein counts, the last two columns illustrate the effectiveness of Mascot's protein inference. The top-level protein hits are the ones with at least one unique peptide match. Sameset and subset proteins are hits for which there is no unique evidence.

First of all, the vast majority of ORFs are clustered with microbial proteins. A few percent appear as subsets of human proteins. This is not really a surprise, because clustering is a strong function of entry length. The average ORF length in this database is 42 residues, while most database entries in the gut proteome are much longer.

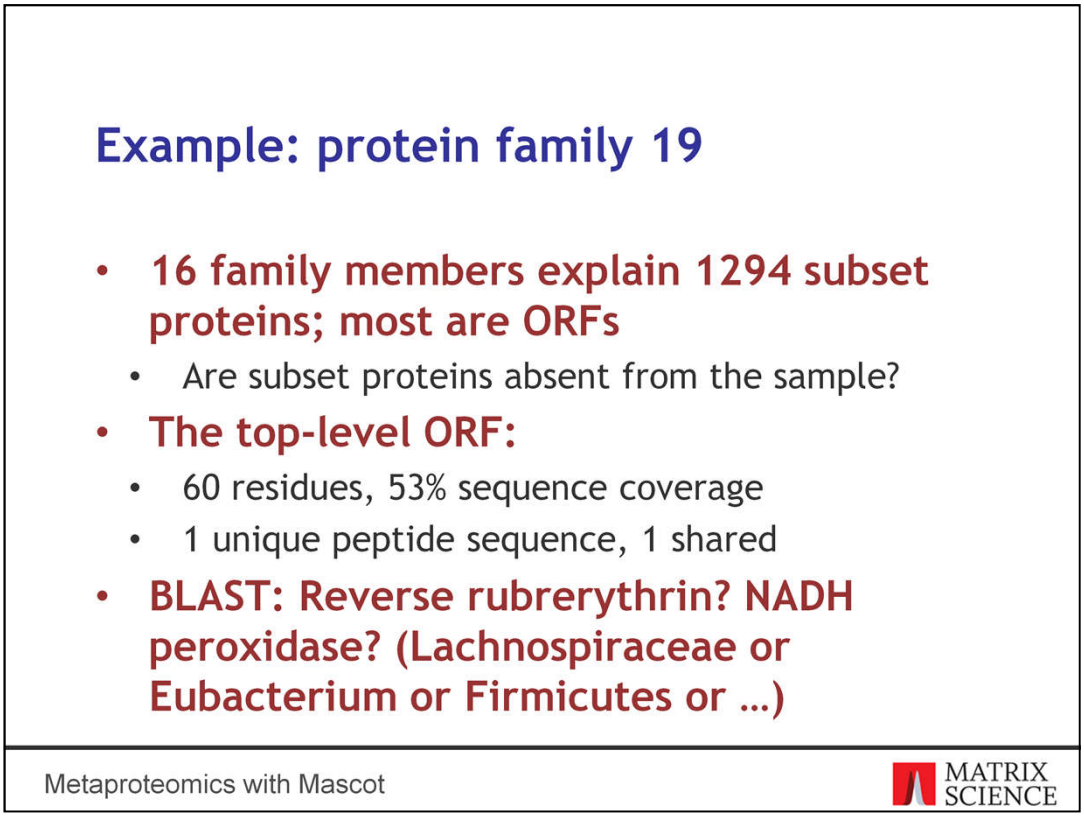
Relatively few ORFs with unique peptide evidence survive. Those that do signify gaps in the public sequence databases: there's clear evidence that the sample contains proteins that haven't been sequenced yet.

Next, look at the protein hits in the gut proteome. The vast majority are again sameset and subset proteins. This means there is a lot of ambiguity at species level, which prevents unambiguous species determination in many cases.



This is a representative example of protein clustering. Protein family 19 has matches to rubrerythrin and rubredoxin across a range of species. Twelve are in the same taxonomic order (Clostridiales), three are in another (Bacteroidales) and then there's one ORF in the middle. The species represent eleven different genera, and two of the species have no more specific classification than at order level.

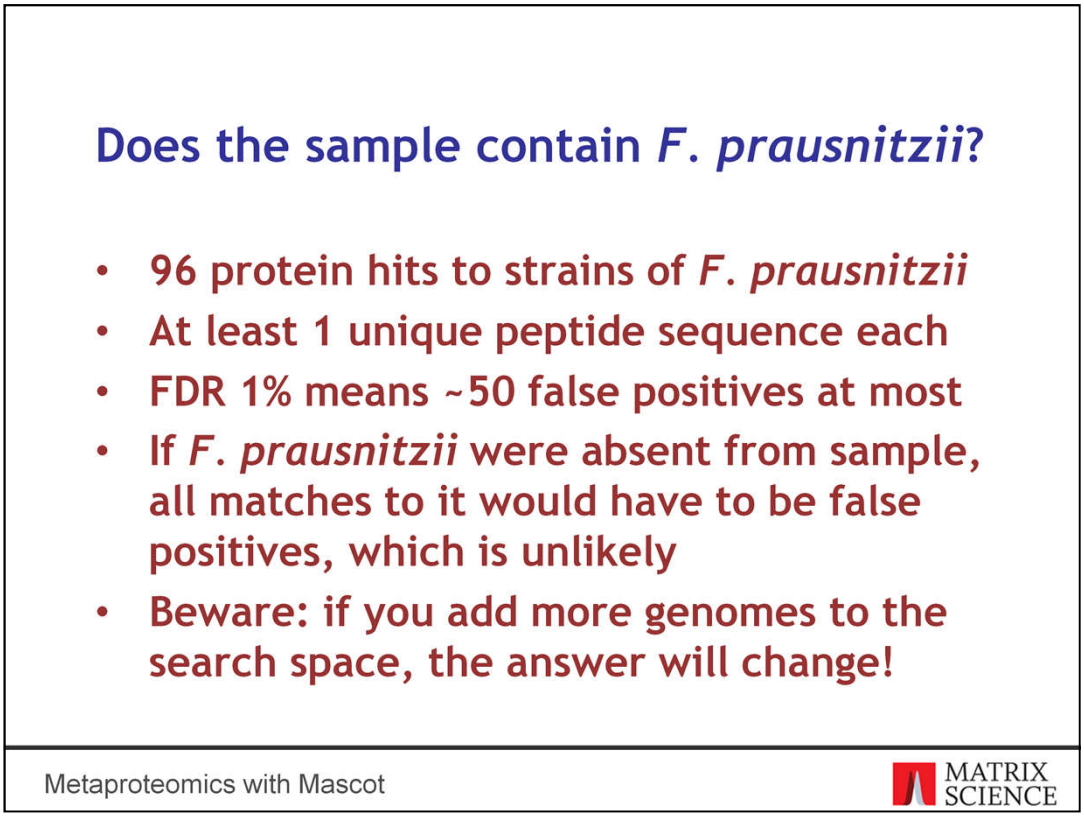
Mascot clusters the proteins together because they share some peptide matches. However, every top-level protein hit in this list has at least one unique peptide match, so there's evidence for each protein in the sample. There's indirect evidence for the presence of the 15 bacterial species due to the unique peptides.



The flip side of protein clustering is that proteins with no unique matches are hidden away as sameset or subset proteins. The 16 family members explain 1294 subset proteins. Most of them are ORFs, which is as expected: when the search space contains the right protein sequences, the ORFs are explained by them. But some of them are microbial proteins, which just happen to not have unique peptide evidence.

Can you say with certainty that subset proteins are not present in the sample? No. If you were to remove the top-level proteins from the database, some of the subset proteins could gain unique peptide evidence and be reported as top-level proteins. Conversely, if you add more proteins to the database, some of the top-level proteins could lose their unique evidence. This is again why it's so important to define the databases clearly before you start the search.

The top-level ORF hit in this family is 60 residues long and has one unique peptide sequence. Out of curiosity, I did a BLAST search to see what it could be. The top hits were to reverse rubrerythrin and NADH peroxidase, and you can choose among a number of different bacterial species depending on which one you believe is the most likely. There is some lachnospiraceae in this protein family already, but the protein sequence is also found in eubacterium and other firmicutes.

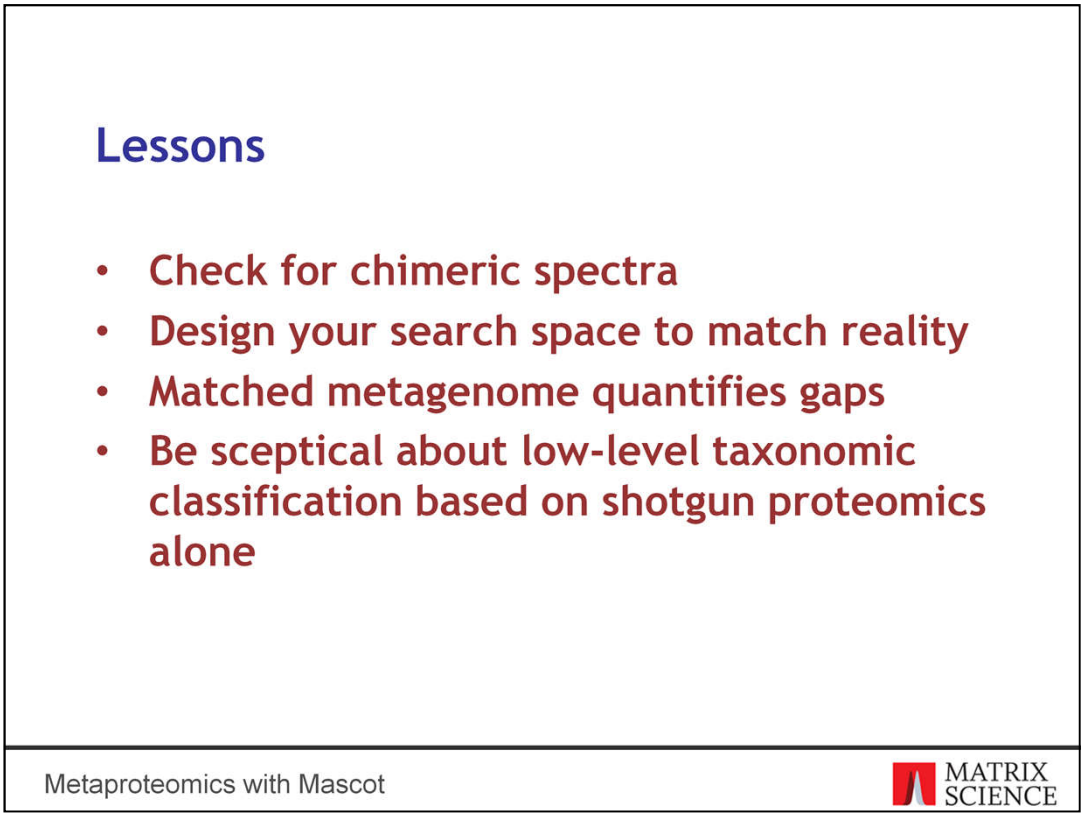


Let's ask a different question. Does the sample contain F. prausnitzii? There are 96 top-level protein hits to different strains of F. prausnitzii, and each one has at least one unique peptide sequence. We've set the FDR to 1%, which means that in the whole search, there are around 50 false positives.

If F. prausnitzii were absent from the sample, all of the 96 unique peptide matches would have to be false positives. It doesn't seem likely given the FDR, so you can confidently say the sample contains F. prausnitzii. The reasoning will be more complicated if you pose the same question about multiple species simultaneously.

Now, it's very important to keep in mind that the answer will change as you add more genomes to the search space. Some of the peptides unique to F. prausnitzii proteins might no longer be unique because they are shared between species, or the peptide matches cease to be statistically significant because the search space is bigger.

We don't really have anything in Mascot that gives you an overview of discovered species, so it's something we will need to look into.



To summarise, searching metaproteomic data with Mascot is possible. You can make it a success by paying attention to these details.

Check for the existence of chimeric spectra. You can turn on the option in Distiller and see how many more queries you get. You could also look at how many spectra get a match to the secondary or tertiary precursor, although at the moment this requires writing a script.

Design your search space to match the list of species known to be present in the sample. You may need to go higher up to phylum level if there are few sequences available for those species. Try to keep the search space as small as possible.

A matched metagenome is useful in quantifying the gaps in public sequence databases. On its own it's not enough.

Finally, shotgun proteomics has its limits when it comes to low-level taxonomic classification. It's sensitive not only to which peptides can be detected but also to the design of your sequence database.