Pacific Northwest
NATIONAL LABORATORY

PNNL-36661

# Proteomics Analysis of Human Contaminant Proteins

September 2024

Fanny Chu
Andy Lin
Daniel H. Lewis
Sarah Jenson
Robert W. Seymour
Eric Merkley
Karen Wahl

U.S. DEPARTMENT OF
ENERGY

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

## DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor Battelle Memorial Institute, nor any of their employees, makes **any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights**. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or Battelle Memorial Institute. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

PACIFIC NORTHWEST NATIONAL LABORATORY
*operated by*
BATTELLE
*for the*
UNITED STATES DEPARTMENT OF ENERGY
*under Contract DE-AC05-76RL01830*

**Printed in the United States of America**

**Available to DOE and DOE contractors from**
**the Office of Scientific and Technical Information,**
**P.O. Box 62, Oak Ridge, TN 37831-0062**
**www.osti.gov**
**ph: (865) 576-8401**
**fox: (865) 576-5728**
**email: reports@osti.gov**

**Available to the public from the National Technical Information Service**
**5301 Shawnee Rd., Alexandria, VA 22312**
**ph: (800) 553-NTIS (6847)**
**or (703) 605-6000**
**email: info@ntis.gov**
**Online ordering: http://www.ntis.gov**

PNNL-36661

# Proteomics Analysis of Human Contaminant Proteins

September 2024

Fanny Chu
Andy Lin
Daniel H. Lewis
Sarah Jenson
Robert W. Seymour
Eric Merkley
Karen Wahl

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99354

# Abstract

Complete characterization of unknowns via proteomics remains challenging. There exist regions of mass spectrometry-based proteomics data where empirical measurements are not attributed to peptides, and/or sequenced peptides from mass spectra are not attributed to any source. These uncharacterized regions are known as the “dark” proteome. Many proteomics tools rely on some *a priori* knowledge of sample composition; few tools allow for investigation of unknowns without relying on composition assumptions. Further, the potential low abundance of minor traces in these uncharacterized regions can make elucidation of the “dark” proteome challenging. Herein, we describe the development and evaluation of approaches to study the “dark” proteome and move towards an untargeted approach for more complete characterization, namely by studying minor human protein traces in non-human samples and combining that approach with non-human source organism identification without relying on assumptions.

Human protein markers, in the form of genetically variant peptides, have been extensively examined in a variety of human matrices, including blood, plasma, and hair, but have yet to be investigated in non-human samples, such as cell cultures, as human contaminant traces. Genetically variant peptides are those that are found in proteins carrying single nucleotide polymorphisms.

In this work, we aimed to (1) investigate the feasibility of detecting human contaminant genetically variant peptides (GVPs) in a diverse set of non-human organisms using public proteomics data and a computational pipeline, as well as to (2) develop a combined capability for untargeted source organism characterization and GVP detection. To our knowledge, this is the first report of applying these approaches towards a more complete proteomic characterization of unknowns.

We successfully demonstrate the feasibility of broad human contaminant GVP detection in proteomics data, develop a better understanding of GVP detectability, characterize the sample-to-sample variability in GVP detection, and identify a core set of GVPs that can potentially be used as markers indicative of the human contaminant traces portion of the “dark” proteome. Further, we developed and evaluated a combined pipeline, MARLOWE-GVP, that enables both untargeted source organism characterization and GVP detection. We show high accuracy of correct source organism characterization and high degree of similarity of human contaminant GVP detection compared to the conventional approach. Success on both these efforts have allowed us to advance our understanding and characterization of the “dark” proteome.

# Acknowledgments

The authors would like to thank Dr. Eva Brayfindley for insightful discussions on support vector machine (SVM) model development. This research was supported by the Open Call Initiative, under the Laboratory Directed Research and Development (LDRD) Program at Pacific Northwest National Laboratory (PNNL). PNNL is a multi-program national laboratory operated for the U.S. Department of Energy (DOE) by Battelle Memorial Institute under Contract No. DE-AC05-76RLO 1830.

# Acronyms and Abbreviations

DDA: data-dependent acquisition mass spectrometry

DIA: data-independent acquisition mass spectrometry

ENSEMBL: public database of genomes and protein sequences for known organisms

FASTA: text-based format for representing protein sequence information

FDR: false discovery rate

gnomAD: Genome Aggregation Database

GVP: genetically variant peptide

KAP: keratin-associated protein

KEGG: Kyoto Encyclopedia of Genes and Genomes

KRT: keratin

MS/MS: tandem mass spectrometry

PRIDE: public repository of mass spectrometry-based proteomics data

PSM: peptide-spectrum match

SNP: single nucleotide polymorphism

SVM: support vector machine

UniProtKB: public database of protein sequences for known organisms

## Contents

Abstract......ii
Acknowledgments......iii
Acronyms and Abbreviations......iv
Figures......vii
Tables......x
1.0 Introduction......1
1.1 Forensic Proteomics......1
1.2 Unknown Source Organism Identification......1
1.3 Aims......2
2.0 Human Contaminant Peptide Detection......3
2.1 Genetically Variant Peptide Detection Pipeline Development......3
2.1.1 Genetically Variant Peptide (GVP) Selection......4
2.1.2 GVP FASTA File Creation......4
2.1.3 Database Search Workflow......4
2.2 Detection from PRIDE Repository Datasets......5
2.2.1 Dataset Selection......5
2.2.2 Human Contaminant GVP Detection......6
2.2.3 Proteins and Genes Associated with GVPs......8
2.2.4 GVP Protein Assignment Ambiguity......10
2.2.5 GVP Alleles......11
2.2.6 Common GVPs......14
2.3 Detection from In-House Repository Datasets......17
2.3.1 Dataset Selection......18
2.3.2 GVP Detection from a Single Sample Preparer......20
2.3.3 Proteins and Genes Associated with GVPs......22
2.3.4 GVP Protein Assignment Ambiguity......24
2.3.5 GVP Alleles......24
2.3.6 Common GVPs......27
2.4 Comparison of GVP Detection Performance......30
2.5 Conclusions......34
3.0 MARLOWE-GVP Analysis......35
3.1 MARLOWE-GVP Pipeline Development......36
3.1.1 *De Novo* Peptide Sequencing......36
3.1.2 Source Organism Characterization with MARLOWE......37
3.1.3 Support Vector Machine Model Development for Organism Filtering......37

3.1.4 MARLOWE-Informed GVP FASTA File Creation and Database Search....39
3.2 Results of MARLOWE-GVP Analysis....39
3.2.1 MARLOWE Correct Organism Classification....39
3.2.2 SVM Organism Filtering....42
3.2.3 GVP Detection Performance....44
3.3 Conclusions....48
4.0 Concluding Remarks & Future Outlook....49
5.0 References....50

# Figures

Figure 1. Barplot of the distribution of the number of data-dependent acquisition mass spectrometry projects deposited into the ProteomeXchange/PRIDE repository by source organism, as of May 2023. ....6

Figure 2. Barplot displaying the number of unique GVP sequences detected from each PRIDE project (each represented by a single bar), grouped by source organism, at 1% peptide-level FDR control. ....7

Figure 3. Barplot displaying the number of unique GVP sequences detected from each PRIDE project (each represented by a single bar), grouped by source organism, at 5% peptide-level FDR control. ....8

Figure 4. Distribution of the number of SNPs contained in each detected human contaminant GVP, grouped by organism. The vast majority of detected sequences contain a single SNP....12

Figure 5. Barplot displaying the distribution of variant types among detected human contaminant GVPs, grouped by source organism. Variant types include those with single SNPs (ref = reference, alt = alternate) and for those GVPs containing multiple SNPs (multi-ref = multiple reference alleles, multi-alt = multiple alternate alleles, both = containing both reference and alternate alleles in some combination). Notably, the reference alleles are most prevalent in detected GVPs, across all organisms....13

Figure 6. Barplot displaying distribution of GVP variant type, by type of gene associated with GVP sequence. Gene types include genes resulting in keratins (KRT), keratin-associated proteins (KAP), and other structural proteins. Variant types include those with single SNPs (ref = reference, alt = alternate) and for those GVPs containing multiple SNPs (multi-ref = multiple reference alleles, multi-alt = multiple alternate alleles, both = containing both reference and alternate alleles in some combination)....14

Figure 7. Histogram displaying distribution of number of human contaminant GVPs most often detected from PRIDE projects for each source organism. ....15

Figure 8. Venn diagram showing GVP sequence overlap among datasets from different organisms. 22 GVPs were found to be detected in at least one PRIDE project from each organism....17

Figure 9. Barplot of the distribution of the number of data-dependent acquisition mass spectrometry datafiles deposited into the in-house repository by source organism, between 2015 and 2020. Datafiles are further associated with campaigns. ....19

Figure 10. Barplot of the number of campaigns and datafiles, grouped by organism, of samples prepared by each individual, between 2015 and 2020. Sample preparer 37 prepared the largest number of samples that exhibit the greatest organism diversity. ....20

Figure 11. Barplot displaying the number of unique GVP sequences detected from in-house repository campaigns (each represented by a single bar), grouped by source organism, at 1% peptide-level FDR control, where each sample within the campaign was prepared by sample preparer 37....21

Figure 12. Barplot displaying the number of unique GVP sequences detected from in-house repository campaigns (each represented by a single bar), grouped by source organism, at 5% peptide-level FDR control, where each sample within the campaign was prepared by sample preparer 37................................22

Figure 13. Histogram displaying the distribution of the number of SNPs contained in each detected human contaminant GVP, grouped by organism, across the in-house repository campaigns handled by sample preparer 37. Most GVPs contain only a single SNP. ....................................................................25

Figure 14. Barplot displaying the distribution of variant types among detected human contaminant GVPs, grouped by source organism, for campaigns from the in-house repository of samples handled by sample preparer 37. Variant types include those with single SNPs (ref = reference, alt = alternate) and for those GVPs containing multiple SNPs (multi-ref = multiple reference alleles, multi-alt = multiple alternate alleles, both = containing both reference and alternate alleles in some combination). Notably, the reference alleles are most prevalent in detected GVPs, across all organisms. ..........................................................................................26

Figure 15. Barplot displaying distribution of GVP variant type, by type of gene associated with GVP sequence, among in-house repository campaigns prepared by sample preparer 37. Gene types include genes resulting in keratins (KRT), keratin-associated proteins (KAP), and other structural proteins. Variant types include those with single SNPs (ref = reference, alt = alternate) and for those GVPs containing multiple SNPs (multi-ref = multiple reference alleles, multi-alt = multiple alternate alleles, both = containing both reference and alternate alleles in some combination)................27

Figure 16. Histogram displaying distribution of number of human contaminant GVPs most often detected from select in-house repository campaigns for each source organism, wherein each sample from these campaigns was handled by sample preparer 37...........................................................................28

Figure 17. Venn diagram showing GVP sequence overlap among campaigns from different organisms. Only a single human contaminant GVP sequence overlap among all four organisms. ....................................................................29

Figure 18. Venn diagram of human contaminant GVP overlap between GVPs detected in PRIDE projects and the in-house repository campaigns prepared by sample preparer 37. ..........................................................................................31

Figure 19. Venn diagram of human contaminant GVP overlap between GVPs detected in PRIDE *E. coli* (K-12 strain) projects and the in-house repository *E. coli* (BL21 strain) campaigns prepared by sample preparer 37. ...............................32

Figure 20. Venn diagram of human contaminant GVP overlap between GVPs detected in PRIDE *M. musculus* projects and the in-house repository *M. musculus* campaigns prepared by sample preparer 37. ....................................................33

Figure 21. Bioinformatics pipeline combining MARLOWE and GVP detection. SVM machine learning model-based filtering occurs following a list of potential source organisms produced by MARLOWE, and filtered results are then used to inform creation of the final FASTA file that combines GVP sequences of interest and source organism proteomes for database searching and GVP detection...........................................................................36

Figure 22. Histogram displaying the distribution of the number of organisms' proteomes included in the source organism FASTA file, as informed by SVM model filtering of MARLOWE's potential source organism lists, grouped by ground truth source organism, for downstream database search and GVP detection from select PRIDE projects. The vast majority of FASTA files created from this approach contain a single organism’s proteome, though up to a maximum of 28 organisms’ proteomes for a minority of FASTA files. ....................43

Figure 23. Comparison of the numbers of detected human contaminant GVPs per PRIDE project at 5% peptide-level FDR control, and grouped by ground truth organism, with the two bioinformatics pipeline (gvp-only = GVP detection only; marlowe-gvp = via MARLOWE-GVP pipeline). ....................44

Figure 24. Barplot displaying similarity of human contaminant GVP profiles for each PRIDE project and grouped by ground truth source organism, between the GVP detection-only and MARLOWE-GVP pipelines....................46

# Tables

Table 1. List of top 10 most frequently detected human contaminant GVPs across PRIDE datasets and genes and chromosomes from which they derive................9
Table 2. List of top 10 predominant genes associated with detected human contaminant GVPs. ....................................................................................................9
Table 3. List of GVPs that map to more than one human keratin.............................................10
Table 4. Most frequently detected GVPs across PRIDE projects for each source organism. ....................................................................................................16
Table 5. List of top 11 most frequently detected human contaminant GVPs across in-house repository campaigns prepared by sample preparer 37. ..........................22
Table 6. List of top 9 predominant genes associated with detected human contaminant GVPs across in-house repository campaigns prepared by sample preparer 37. ....................................................................................................23
Table 7. Most frequently detected GVPs across campaigns handled by sample preparer 37 for each source organism, where each GVP is observed in more than one organism. ....................................................................................................30
Table 8. Performance of each SVM model. ....................................................................................................38
Table 9. Number of datafiles, grouped by organism, from PRIDE projects on which conventional MARLOWE performance was evaluated .......................................39
Table 10. Conventional MARLOWE correct classification rate for each organism, at two different levels (top rank, any rank) ....................................................................................................40
Table 11. Conventional MARLOWE correct classification rate to the ground truth organism, organized by sample complexity. ....................................................................................................41
Table 12. GVP profile similarity between GVP detection-only and MARLOWE-GVP pipelines for PRIDE projects containing datafiles searched with the incorrect source organism ....................................................................................................47
Table 13. List of top 10 most frequently detected GVPs in PRIDE projects using the MARLOWE-GVP pipeline that are also detected from the GVP detection-only pipeline. ....................................................................................................47

# 1.0 Introduction

Characterization of unknowns via proteomics can be challenging. Several strategies and applications, depending on the question at hand, exist, but may have limitations. Detection and characterization of minor protein components in unknowns for complete proteomic characterization presents even more of a challenge. Often, uncharacterized regions of proteomics samples will remain, which is known as the “dark” proteome. In this work, we set our sights on better characterization and understanding of unknowns via proteomics, and develop computational capabilities that enable us to investigate the “dark” proteome.

## 1.1 Forensic Proteomics

While mass spectrometry-based proteomics is the premier tool for detecting and quantifying proteins in a sample, this technique has only recently been applied to address forensic questions. The application of proteomics to forensics (i.e., forensic proteomics) is most useful when analyzing samples where DNA is absent or degraded. For example, common samples such as hair, protein toxins, and red blood cells do not contain intact, genomic DNA. In addition, protein-based analysis is expected to allow for analysis of samples that have undergone a variety of storage or weathered conditions, whether intended or not, as proteins are typically more stable than DNA. Proteomics analysis has been successfully used for snake venom identification[1], protein toxin detection and identification[2-6], identification of body fluids at crime scenes[7-9], human individualization from hair[10-12] and bone[13], and species identification, including microbes[14-16].

One application of forensic proteomics that could potentially be leveraged to investigate the “dark” proteome is genetically variant peptide detection, which has primarily been performed in human proteomics samples[10, 17]. However, this approach has not been applied to non-human samples beyond an initial proof-of-concept reported in Chu and Lin (2024, preprint)[18], which entails detection of minor components potentially at low abundance, even though it is known that human protein traces could be left behind owing to sample handling. These minor protein traces, if present, could represent a part of the “dark” proteome of non-human samples that is currently not well-characterized.

## 1.2 Unknown Source Organism Identification

Determining the source organism of an unknown sample is often one of the first questions that is asked when a forensic or biodefense sample is obtained. While DNA-based analysis is often used to answer this question, proteomics analysis can provide additional or confirmatory information. In this section, we describe common strategies and existing tools for source organism identification of unknown proteomics samples.

One strategy for proteomic species identification relies on detecting peptides present in a sample using database search. In a database search, experimentally collected spectra are searched against a user-defined protein database that contains protein sequences that are expected to be present in the sample. This approach assumes that the composition of the sample is well-characterized, which may not be true for forensic samples. Two different approaches to database search can be implemented for species identification. The first relies on curating a database of suspected organisms in the search, and then relying on a list of organism-unique peptides to determine the taxonomic composition of the sample.[19] For example, if a detected peptide sequence was only found in *Escherichia coli* proteomes, then *E.*

*coli* can be said to be present in the sample. The second method utilizing database search makes no assumption on potential source organism by including proteomes of all known organisms into the database. An example method that utilizes this type of approach includes MiCId[20]. This method produces species identification, relying on statistics to describe the confidence of each identification.

One of the challenges with the first database search method is that it suffers from signal erosion. Signal erosion occurs when peptides that were originally unique to a single organism but becomes non-unique as additional genomes are sequenced.[21] As a result, new strategies for proteomic species identification are needed, that are not as strongly affected by signal erosion as more genomes are sequenced.

An alternative approach for proteomics species identification relies on *de novo* detection of peptides present in a sample. Instead of searching spectra against a user-defined database, evidence of peptides is directly derived from the spectra by looking directly at the distance between fragment peaks and comparing those distances to known masses of amino acids. This approach makes no assumptions on the source organism of unknown proteomics samples. Following *de novo* peptide detection, other tools to assign peptides to organisms are typically used (e.g., UniPept[22]). MetaNovo utilizes UniPept to assign *de novo* peptides to organisms[23].

However, *de novo* peptide sequencing is rarely accurate for the entire sequenced peptide; it usually yields a partially correct peptide sequence. This is because mass spectra rarely contain complete information on every amino acid within the associated peptide sequence. For this reason, the unknown organism characterization tool MARLOWE uses only the highly-confident regions of *de novo* peptide detections, thus maximizing the value and avoiding the limitations of *de novo* peptide sequencing. MARLOWE returns a ranked list of potential taxonomic contributors to a proteomics sample by performing confident peptide region assignments to organisms via protein inference as well as peptide strength[14] to avoid signal erosion issues[21]. MARLOWE has been demonstrated on a variety of samples without knowing their composition, which has utility in forensic science and metaproteomics.

## 1.3 Aims

Two efforts are outlined in this work to push the boundaries of proteomic characterization of unknowns. We aim to (1) characterize the feasibility and reproducibility of human contaminant GVP detection, that is, presence of minor traces, for broad application to non-human proteomics samples, and (2) demonstrate and evaluate a combined, untargeted capability for unknown source organism characterization via MARLOWE and human contaminant GVP detection. Success in achieving both aims will enable a more complete proteomic characterization of and further elucidation of the “dark” proteome in unknowns.

# 2.0 Human Contaminant Peptide Detection

During mass spectrometry-based proteomics data analysis, spectra are searched against a database of protein sequences that are expected to be present in the sample. This database typically consists of the reference proteomes of the set of organisms in the sample. In addition, a set of contaminant proteins, such as human keratins and trypsin, are appended to account for proteins that are artificially introduced into the sample during sample preparation. Recent work has been performed to create universal contaminant libraries for both data-dependent acquisition (DDA) and data-independent acquisition (DIA) mass spectrometry data[24] as well as affinity purification mass spectrometry[25]. In addition, there are legacy databases, such as CRAPome[26], that have been created but not updated in years.[24]

In addition to the development of contaminant proteins databases, there has been additional research identifying new classes of protein contaminants not already present in these databases. For example, recent work has shown that mass spectrometry can be used to detect contaminant human genetically variant peptides (GVPs) present in non-human samples[18]. GVPs are peptides that contain single amino acid polymorphisms that result from non-synonymous SNPs in protein coding regions of DNA.

In this work, we build on the previous effort[18] to detect human contaminant GVPs in non-human proteomics samples. Chiefly, we aim to (1) examine the feasibility of detecting these GVPs with a more diverse set of non-human organisms and across more available datasets and (2) examine the variability of human contaminant GVP detection when we expect the same GVP profiles in datasets prepared by the same individual. These efforts will expand our understanding and provide a more complete characterization of the “dark” proteome, that is, the unknown portions of a proteomics sample that have not been attributed to the source organism, of which we hypothesize that human contaminant GVPs are a fraction.

We find that human contaminant GVPs can be broadly detected in proteomics data, regardless of source organism, and that a subset of GVPs is frequently detected, albeit not strictly reproducible across all proteomics datafiles. Human contaminant GVP detection is still variable, even when we expect to detect the same GVPs across proteomics samples prepared by the same individual. This effort enabled a better understanding of which human contaminant GVPs could be reasonably detected, from which human protein sources, and expected variant type, and any effects of source organism on GVP detectability, thus advancing the characterization of the “dark” proteome.

## 2.1 Genetically Variant Peptide Detection Pipeline Development

Detection of genetically variant peptides in proteomics samples relies first on curation of a set of these variant peptides to then be incorporated into a database search. As our focus is on detection of GVPs from human contaminant proteins, we considered only human keratins and keratin-related proteins. Following the workflow outlined in Chu and Lin (2024, preprint)[18] with minor modifications, we generate a set of *in silico* trypsin-digested peptides, including variants (i.e., GVPs) from expected effects of SNPs, from human contaminant proteins of interest. These target peptides are then combined with the proteome of the ground truth source organism for database search of a sample’s mass spectrometry data, to detect peptides, and more importantly, GVPs. The sections below detail the development of this bioinformatics pipeline.

### 2.1.1 Genetically Variant Peptide (GVP) Selection

We curated genetically variant peptides derived from genes related to human keratins. First, we selected relevant genes, after surveying UniProtKB[27], and retrieved all resultant transcripts from ENSEMBL/Biomart[28]. Then, using ENSEMBL/Variant Effect Predictor[28], we surveyed all single nucleotide polymorphisms (SNPs) detected in these genes (with GRCh38 build of the reference genome), including only those SNPs with global minor allele frequencies ≥ 0.01 (as reported by the Genome Aggregation Database, gnomAD[29]). This includes multi-allelic SNPs, that is, SNPs that contain more than two alleles. This selection process yielded 430 SNPs.

### 2.1.2 GVP FASTA File Creation

The GVP FASTA file contains a combination of human keratin reference peptides (i.e., peptides that do not contain any SNP sites) and GVPs. GVPs were generated from the selected SNPs acting on the appropriate transcript using an in-house script. Here, we produce both the reference and mutated (also alternate) alleles on *in silico* trypsin-digested peptides resulting from transcripts to represent the effects of SNPs. Peptides were required to contain between 6 and 50 amino acids. Proline blocking was not considered. Effects of multiple SNPs acting on the same peptide were accounted for by permuting all combinations. Multi-allelic SNPs were also accounted for by permuting all alternate alleles.

Custom headers were then created for GVPs as unique identifiers, such that the SNP(s) included can be identified as either the reference or mutated variant following the database search, even if there are multiple SNPs present in the peptide. Gene name, transcript, and SNP information were included in the header to track each GVP. These GVPs were then combined into a GVP FASTA file with human reference keratins for database search. This GVP FASTA file, combined with a FASTA file containing common contaminant proteins as described by Frankenfield et al. (2022)[24], augmented the FASTA file containing the source organism's proteome (downloaded from UniProtKB[27]).

### 2.1.3 Database Search Workflow

We used a database search to detect peptides in proteomics runs. In this work, we used the Tide search engine[30] implemented within Crux[31]. Each mass spectrometry run was searched against a database containing the source organism proteome, human reference keratins, GVPs, and common contaminant proteins[24]. Additional details regarding database construction are described above. For all of these searches, all other parameters were set to their default values, except --compute-sp=T and --pin-output=T. Two different sets of post-translational modifications were used during database search of proteomics data from the two data sources. The first, for PRIDE projects, applied modifications that aligned with the reported modifications from the dataset source. Standard modifications include static cysteine carbamidomethylation, variable methionine oxidation (up to 5 modifications), and variable peptide N-terminal acetylation. Additional custom modifications to PRIDE projects, as needed, include glutamine deamidation and serine, threonine, and/or tyrosine phosphorylation. The second, for in-house repository datasets, used solely static cysteine carbamidomethylation. Following the database search, which was performed separately for each datafile within a PRIDE project or an in-house repository campaign, all peptide detections for datafiles within a project or campaign were combined and the false discovery rate was estimated using Percolator[32] within Crux. Results were filtered to not allow any matches with an FDR of more than 1% and 5%, respectively, at the peptide-level[33]. The GVPs that were identified at the end of this pipeline were then reported. Prior to the database search, raw mass spectrometry files were converted to mzML format using either MSConvert within the Proteowizard[34] suite or ThermoFileRawParser[35].

## 2.2 Detection from PRIDE Repository Datasets

### 2.2.1 Dataset Selection

We selected a subset of data-dependent acquisition mass spectrometry datasets within the external ProteomeXchange/PRIDE repository[36] for this work to represent a diversity of non-human organisms that are well-studied. To inform dataset selection, we obtained statistics for the number of projects—projects and datasets are henceforth used interchangeably—by organism in the repository, as of May 2023 (Figure 1). Figure 1 below displays the top 10 organisms most well-represented with datasets in the PRIDE repository. Given this information and our criteria above, we elected to use datasets from the following species: *Mus musculus*, *Saccharomyces cerevisiae*, *Arabidopsis thaliana*, *Escherichia coli* (K-12 strain), and *Drosophila melanogaster*.

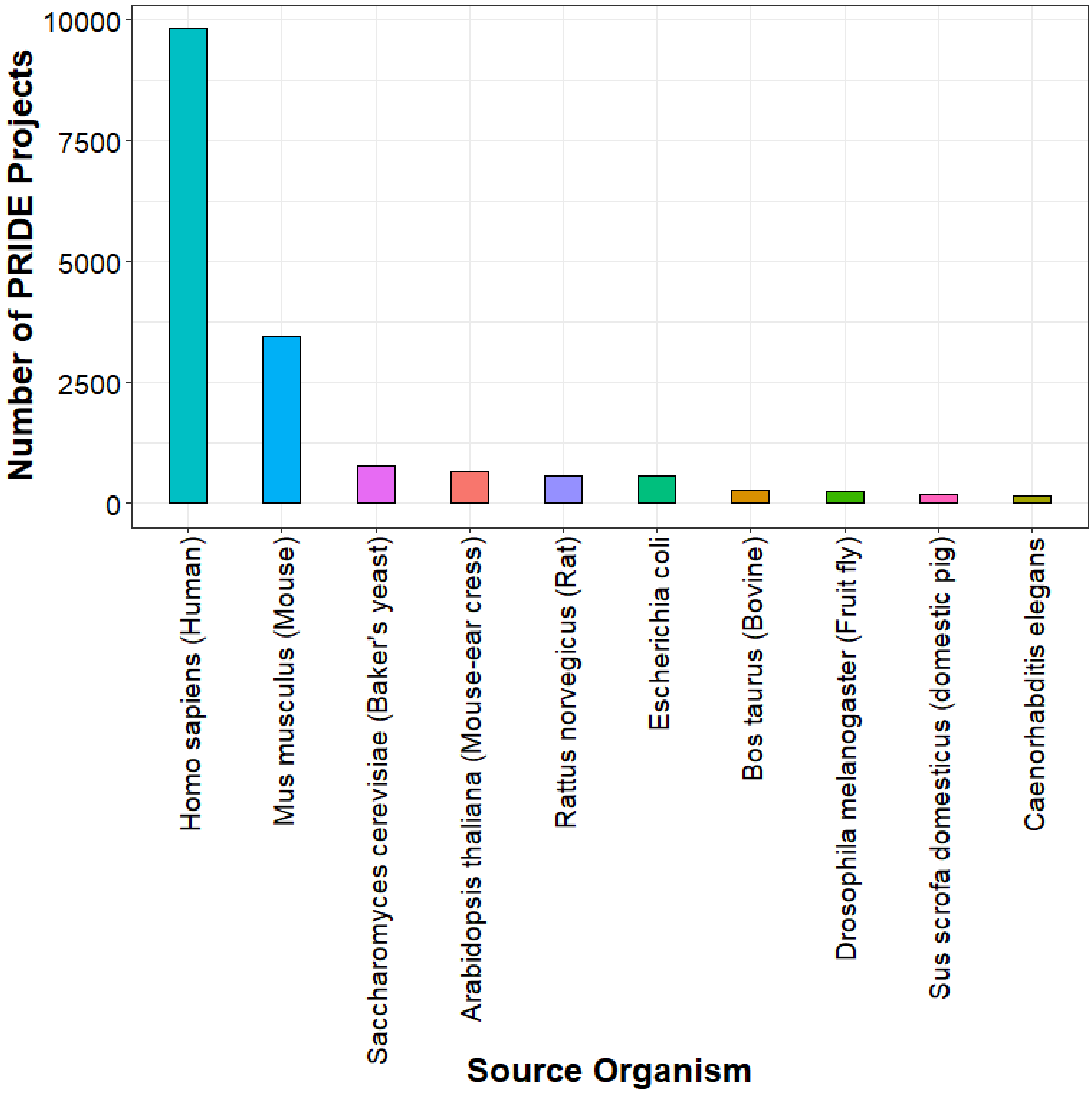

Figure 1. Barplot of the distribution of the number of data-dependent acquisition mass spectrometry projects deposited into the ProteomeXchange/PRIDE repository by source organism, as of May 2023.

We curated datasets and their metadata derived from the five organisms listed above. Of the 55 total projects selected (11 *E. coli*, 13 *D. melanogaster*, 11 *S. cerevisiae*, 10 *M. musculus*, 10 *A. thaliana*), the total number of viable datafiles per project ranged between 3 and 637; on average, 42 ± 91 (s.d.) datafiles.

### 2.2.2 Human Contaminant GVP Detection

We confirm the feasibility of human contaminant GVP detection that is broadly applicable to non-human organisms. Of the 55 DDA projects curated for database search and GVP detection, 45 projects were found to contain detectable GVPs at both 1% and 5% peptide-level FDR control. Figure 2 below displays the number of detected unique GVP sequences per dataset, by organism, at 1% peptide-level FDR control. Interestingly, GVP detections ranged widely among datasets, though *M. musculus* datasets tended towards having fewer detected GVPs (on average, 6 ± 4 (s.d.) GVPs) compared to other datasets from other organisms.

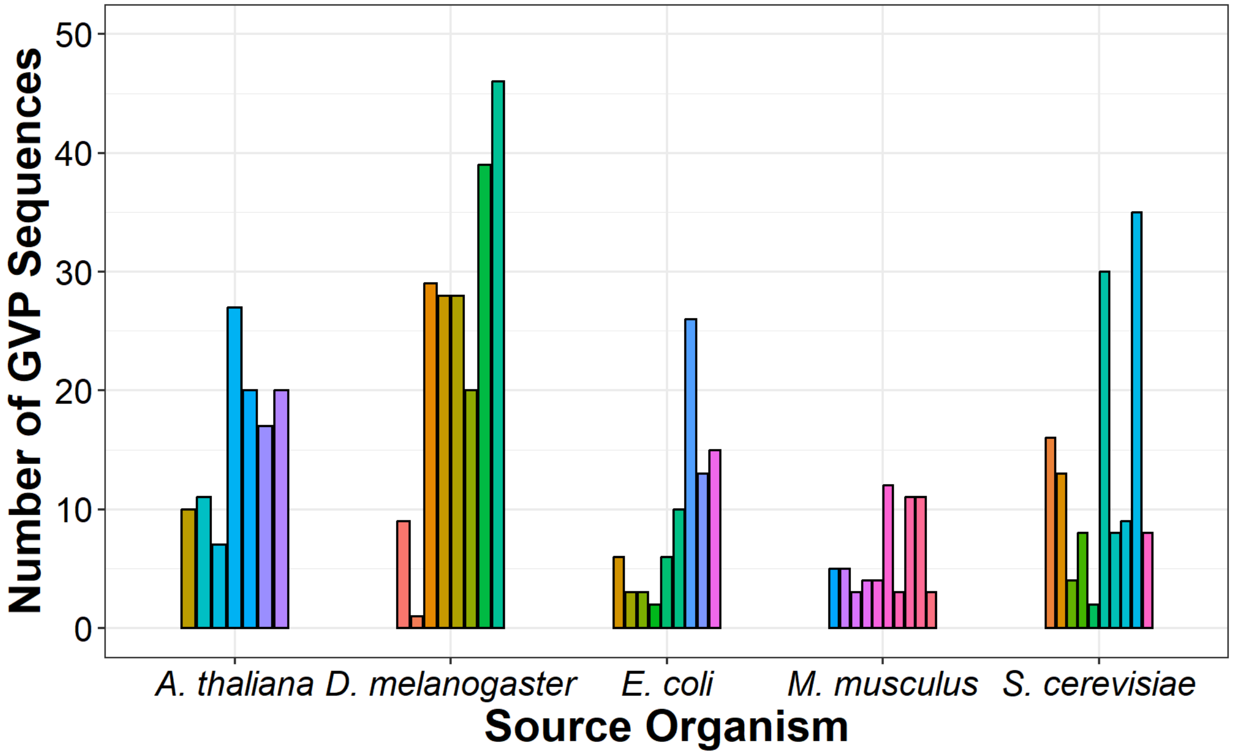

Figure 2. Barplot displaying the number of unique GVP sequences detected from each PRIDE project (each represented by a single bar), grouped by source organism, at 1% peptide-level FDR control.

As expected, we see a slight increase overall in the number of GVPs detected per dataset when applying 5% FDR control. The most obvious increase is observed in *M. musculus* datasets (on average, 17 ± 8 (s.d.) GVPs). To account for the range of datafiles per PRIDE project in which we detected GVPs, we normalized the number of detected unique GVP sequence to the number of respective datafiles per project. On average, we observe 2 ± 2 (s.d.) GVPs per datafile (median = 1, max = 8).

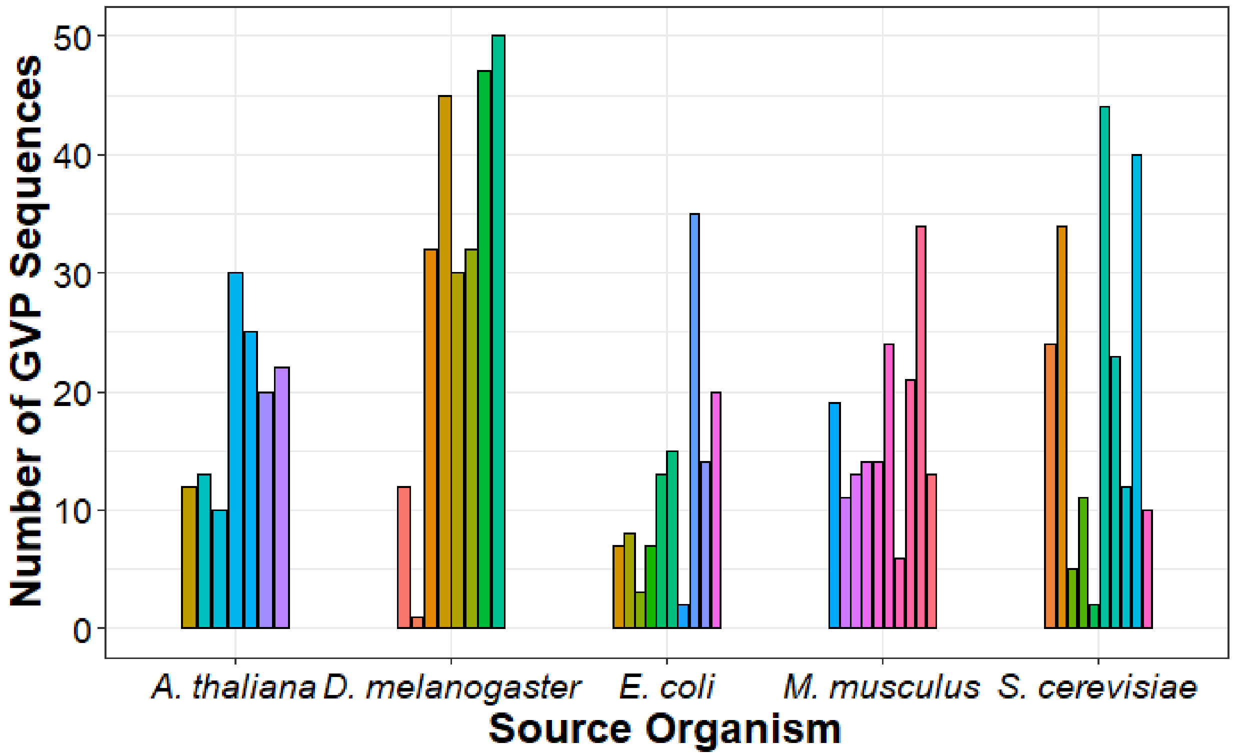

Figure 3. Barplot displaying the number of unique GVP sequences detected from each PRIDE project (each represented by a single bar), grouped by source organism, at 5% peptide-level FDR control.

The detected GVPs at 5% FDR were then examined further, described in the next sections, to address the following questions:

1. What proteins and genes do the detected human contaminant GVPs derive from?
2. Are GVP sequences unique to human contaminant proteins?
3. Which alleles are more likely to be detected: reference (typically the major allele) or alternate (typically the minor allele)?
4. Are there common GVPs among these organisms?

### 2.2.3 Proteins and Genes Associated with GVPs

Across the 45 PRIDE projects with human contaminant GVP detection, the top 10 most frequently detected human contaminant GVPs (out of 244 unique sequences) derive from keratins, and not surprisingly, most are cytoskeletal keratins and/or are enriched in skin. This is consistent with the conventional wisdom that contamination during sample handling likely derives from skin cells. Table 1 below displays the GVP detection frequency along with the gene names associated with the proteins that these human contaminant GVPs derive from. Also not unexpectedly, the predominant chromosomes associated with these genes are chromosomes 12 and 17, which contain the vast majority of keratin-related genes.

Table 1. List of top 10 most frequently detected human contaminant GVPs across PRIDE datasets and genes and chromosomes from which they derive

| GVP sequence | Detection frequency | Gene name | Chromosome |
|---|---|---|---|
| LAADDFR | 39 | KRT13 | 17 |
| VTMQNLNDR | 37 | KRT14 | 17 |
| AQYEEIAQR | 36 | KRT76 | 12 |
| FASFIDK | 36 | KRT75 | 12 |
| DYQELMNVK | 33 | KRT76 | 12 |
| LEQEIATYR | 30 | KRT14 | 17 |
| FLEQQNQVLETK | 28 | KRT74 | 12 |
| SLYGLGGSK | 24 | KRT6C | 12 |
| FLEQQNK | 22 | KRT6B | 12 |
| YQELQITAGR | 19 | KRT77 | 12 |

Of the 244 unique GVP sequences, the predominant proteins that contain these detected human contaminant GVPs are a mix of cytoskeletal keratins (e.g., KRT76) as well as structural proteins (e.g., FLG) (Table 2).

Table 2. List of top 10 predominant genes associated with detected human contaminant GVPs.

| Gene name | GVP detection frequency | Chromosome |
|---|---|---|
| FLG | 39 | 1 |
| DST | 13 | 6 |
| MICA | 9 | 6 |
| KRT72 | 7 | 12 |
| KRT76 | 7 | 12 |
| KRT78 | 7 | 12 |
| KRT13 | 6 | 17 |
| KRT37 | 6 | 17 |

| | | |
|---|---|---|
| KRT77 | 6 | 12 |
| KRT74 | 5 | 12 |

### 2.2.4 GVP Protein Assignment Ambiguity

Next, we examined protein assignment ambiguity of detected GVP sequences, that is, whether these sequences unambiguously belong to human proteins or can be found in proteins in other organisms (i.e., the source organism). Protein assignment ambiguity of detected human contaminant GVP sequences is important to establish, as any ambiguity in their protein assignment, particularly to other organisms, would not allow us to rule out the competing hypothesis that the detected GVP may instead be an unmodified peptide from a different protein source.

We find that some of the detected human contaminant GVPs are also assigned to mouse proteins (74 GVPs across all 879 unique sequence detections from 45 projects) and sheep proteins (17 GVPs across 45 projects), regardless of source organism. However, interestingly and fortuitously, with the exception of some GVPs mapping to mouse or sheep proteins, none of the detected contaminant GVPs mapped to the other source organisms (e.g., *E. coli*, *D. melanogaster*). Further, many of the ambiguous mappings to mouse and sheep proteins are to mouse and sheep keratins, respectively, which are known to be highly homologous to human keratins (e.g., Keratin, type II cytoskeletal 1b, from the Krt77 gene in mouse; Keratin, type II microfibrillar, component 7C in sheep) and are also known contaminants.

This observation indicates that any detection of human contaminant GVPs are much more likely to be from human contamination than attributed to the source organism.

Unsurprisingly, some ambiguity in human protein assignment exists for a few of the detected GVPs. Human keratins are notoriously well-conserved for its structural functions. Notably, 5 GVP sequences can be attributed to more than one human keratin (Table 3), though both protein sources in each ambiguous mapping derive from the same chromosome (e.g., Chromosome 12). Also of note is that most of these genes encode for hard (cuticular) keratins that are mostly found in hair and nails (with the exception of KRT40), as opposed to the cytoskeletal keratins reported in Table 1 that are associated with the most often detected human contaminant GVPs.

Table 3. List of GVPs that map to more than one human keratin.

| GVP | Detection frequency | Genes | Chromosome |
|---|---|---|---|
| DNAELENLIR | 8 | KRT31, KRT33B | 17 |
| DSLENTLTESEAR | 12 | KRT33B, KRT34 | 17 |
| EEINELNR | 5 | KRT81, KRT83 | 12 |

| | | | |
|---|---|---|---|
| LEGEINTYR | 3 | KRT40, KRT32 | 17 |
| SQYEALVETNR | 7 | KRT31, KRT34 | 17 |

### 2.2.5 GVP Alleles

Characterizing the variant type (also allele type) of each detected human contaminant GVP provides us with further insight into the SNPs that are associated with each GVP, along with an understanding of which SNPs' effects we are able to detect as GVPs in proteomics data. Given the limitations of data-dependent acquisition mass spectrometry, wherein only the most abundant peptides are likely to be detected, we expect that only a small fraction of possible human contaminant GVPs can be detected. Attribution of detected GVPs to the SNPs and variant types provides insight into proteotypic GVPs (that is, more likely to be detected in proteomics data owing to their ionization efficiency, which is a distinct issue from GVP detection as related to SNP population frequencies).

For this analysis, we excluded the GVPs that exhibit human protein assignment ambiguity (i.e., the GVPs represented in Table 3). Additionally, any GVPs that exhibited SNP assignment ambiguity were also removed. These GVPs with SNP assignment ambiguity come about from the same GVP peptide sequence mapping to different transcripts or protein isoforms for the same gene (and protein product), or mapping to different regions of the same protein, in which different combinations of SNPs act on different DNA regions but produce the same resultant GVP sequence.

First, we examined the number of SNPs contained in each GVP. Figure 4 below displays the distribution of the number of SNPs per detected human contaminant GVP, grouped by organism. As expected, the vast majority of detected GVPs contain a single SNP, followed by 2 SNPs, though interestingly, a handful of detected sequences contain 5 SNPs. Also of note is that the SNP distribution profile looks extremely similar across the 5 source organisms, indicating that we expect to detect similar types of GVPs (those containing single SNPs) agnostic to source organism.

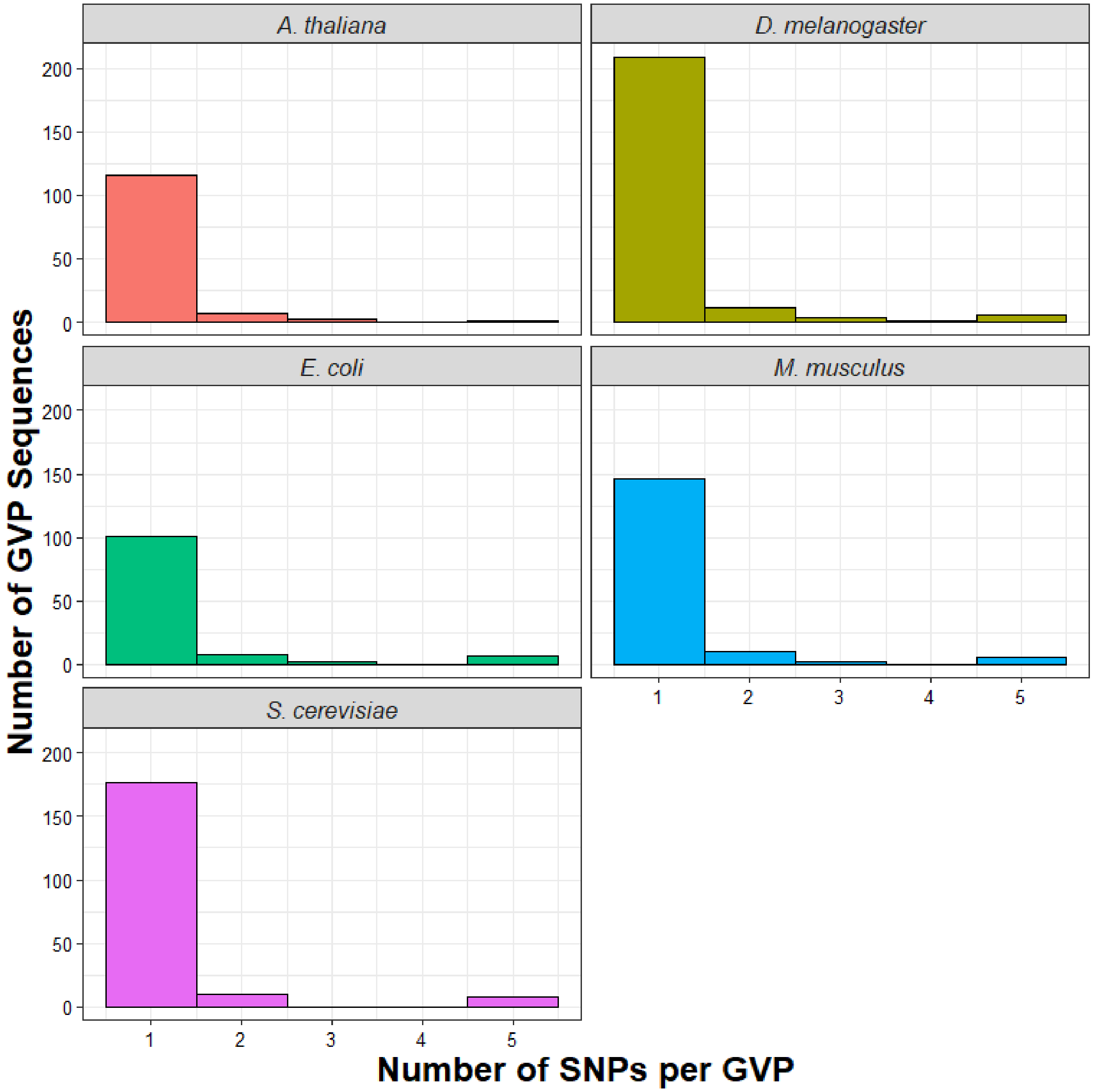

Figure 4. Distribution of the number of SNPs contained in each detected human contaminant GVP, grouped by organism. The vast majority of detected sequences contain a single SNP.

We then examined variant type (i.e., reference or alternate allele) for each detected GVP. Figure 5 below displays the distribution of variant type, grouped by source organism, with labels representing types for GVPs containing single and multiple SNPs, respectively. It is not surprising that the vast majority of detected human contaminant GVPs contain reference alleles of single SNPs. It is, however, interesting to note that the presence of alternate alleles from single SNPs is not insignificant (171 alternate alleles out of 833 GVP detections across all organisms). Surprisingly, for those GVPs containing multiple SNPs, a sizeable portion comes from having both the reference and alternate alleles, and those occur more frequently than having just multiple reference alleles.

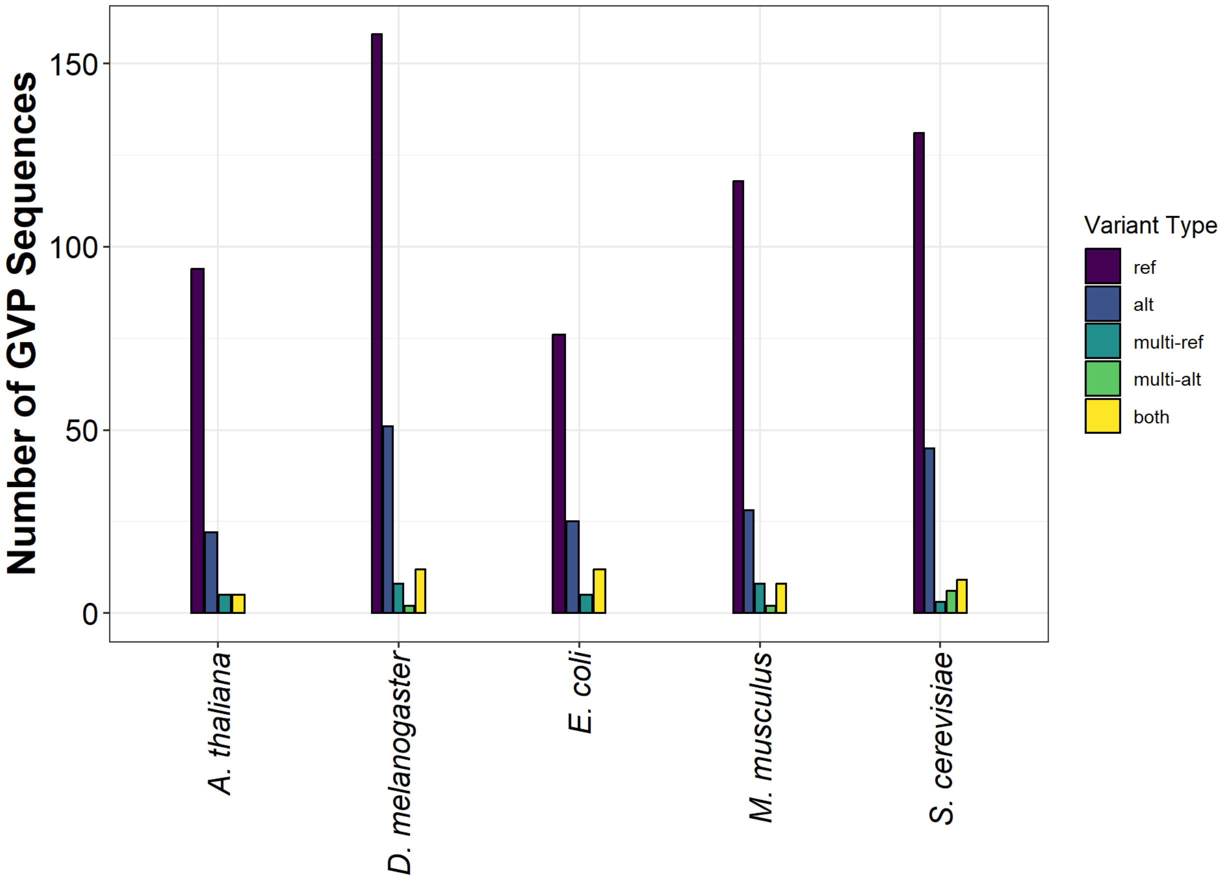

Figure 5. Barplot displaying the distribution of variant types among detected human contaminant GVPs, grouped by source organism. Variant types include those with single SNPs (ref = reference, alt = alternate) and for those GVPs containing multiple SNPs (multi-ref = multiple reference alleles, multi-alt = multiple alternate alleles, both = containing both reference and alternate alleles in some combination). Notably, the reference alleles are most prevalent in detected GVPs, across all organisms.

When we examine the frequency of variant type by the types of genes associated with these GVPs, we find that distribution of variant types differs by gene type. In Figure 6 below, we delineate GVPs associated with keratins, keratin-associated proteins (KAPs), and other structural proteins. For keratins, the distribution of variant type is similar to the distribution observed across the different source organisms in Figure 5, where the vast majority are reference alleles, followed by alternate alleles, derived from single SNPs. Most notably however, we observe detection of alternate alleles much more frequently compared to the reference alleles in keratin-associated proteins (12 alternate alleles out of 20 GVP sequences attributed to KAPs). Also interesting is the prevalence of detecting GVPs belonging to other structural (non-keratin and non-KAP) proteins that contain multiple SNPs and both the reference and alternate alleles within a single sequence (25 out of 91 GVP sequences attributed to other structural proteins). The stark difference in distribution of variant types by gene type provides us with additional insight into which GVPs we can expect to detect in proteomics data. It is also quite obvious, and not at all surprising, that most of the detected human contaminant GVPs derive from keratins, as opposed to KAPs or other structural proteins, given how similar its variant type distribution profile is to the distribution profiles observed across all organisms in Figure 5.

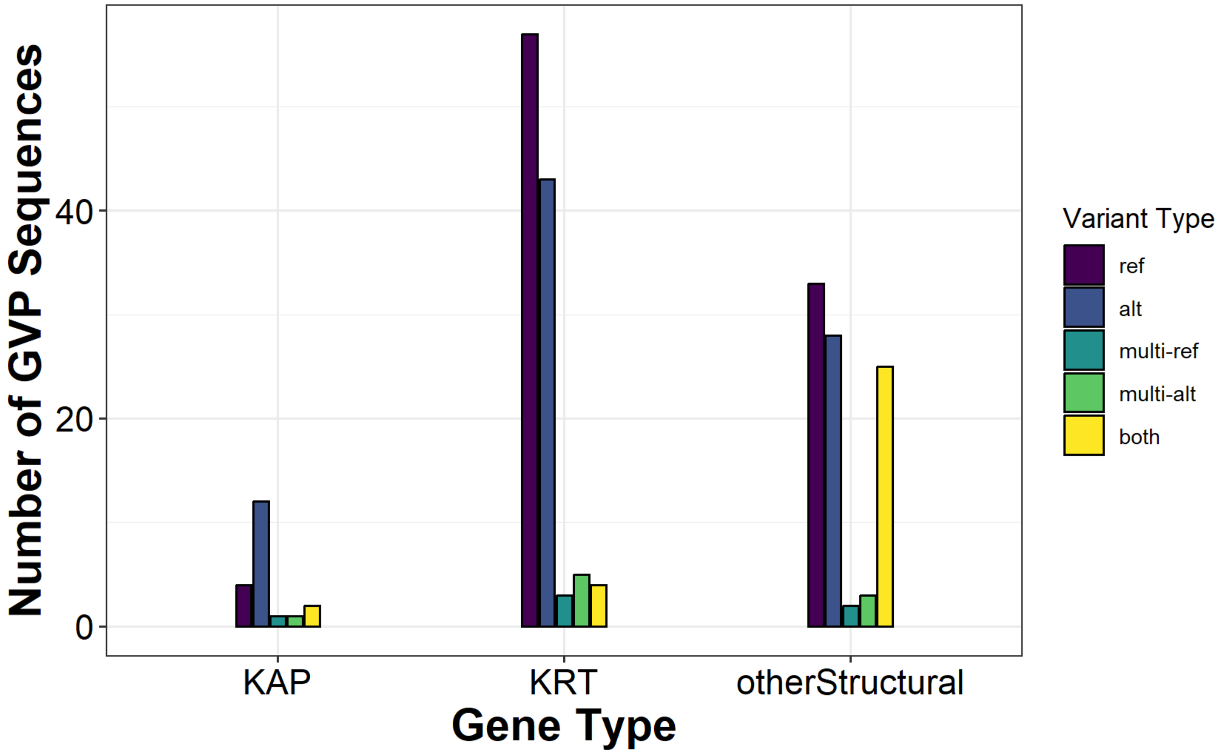

Figure 6. Barplot displaying distribution of GVP variant type, by type of gene associated with GVP sequence. Gene types include genes resulting in keratins (KRT), keratin-associated proteins (KAP), and other structural proteins. Variant types include those with single SNPs (ref = reference, alt = alternate) and for those GVPs containing multiple SNPs (multi-ref = multiple reference alleles, multi-alt = multiple alternate alleles, both = containing both reference and alternate alleles in some combination).

### 2.2.6 Common GVPs

Finally, we examined the extent to which we could detect a common subset of human contaminant GVPs within and across all the source organism datasets analyzed. Determining whether similar GVPs are detected within and across a diverse set of non-human organisms can provide us with insight into which GVPs are more likely to be detected, separate from the variant type analysis performed above, and a basis for identifying potential biomarkers of the “dark proteome” that could be robustly detected (as opposed to spurious detections), giving us more confidence in their detection.

We first considered GVP detection among PRIDE datasets from the same organism. How often are GVPs detected in more than one project? Are there any GVPs that are more often detected than others, or perhaps ubiquitous across projects? Figure 7 below addresses these questions.

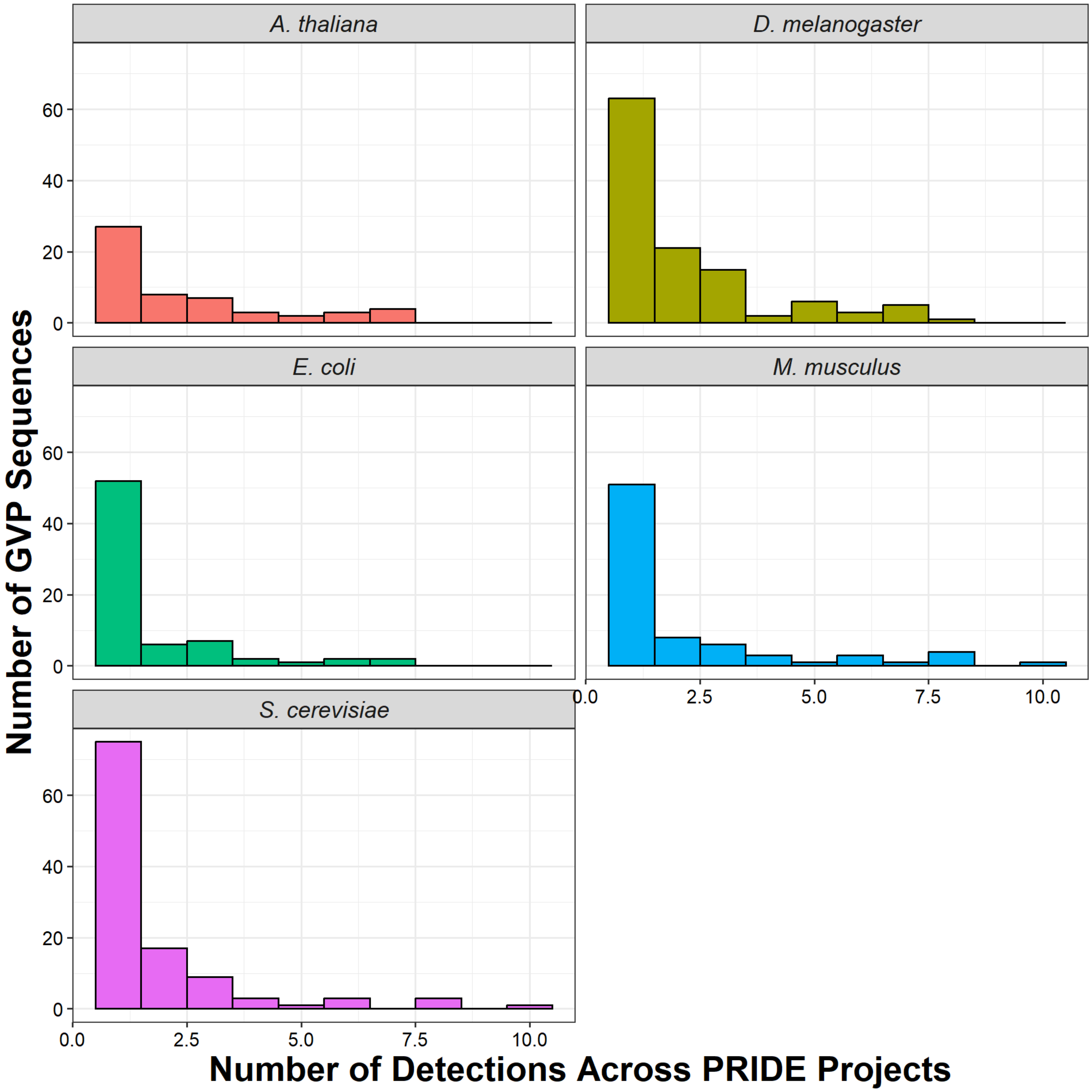

Figure 7. Histogram displaying distribution of number of human contaminant GVPs most often detected from PRIDE projects for each source organism.

It is obvious that the frequency of GVP detection within an organism varies by organism (Figure 7). In 4 out of the 5 organisms (the exception being *A. thaliana*), most GVPs are only detected in a single project. It is important to keep in mind that human contaminant GVP detection in non-human samples is expected to be challenging, owing to the fact that these are minor components and are very likely to be present at only trace levels. Given that challenges in complete peptide detection (for the source organism) using the DDA MS/MS paradigm already exist, it is not surprising that few human contaminant GVPs, that are minor traces, are detected in multiple projects.

Of note is that there are a handful of GVPs that are detected in almost all projects investigated per organism (e.g., 1 GVP detected in 10 *S. cerevisiae* projects out of 10 *S. cerevisiae* projects with any GVP detections). Table 4 below compiles the GVP sequences observed with the highest frequency across projects per source organism. Many of these sequences are also detected quite frequently in the other organisms (e.g., LAADDFR detected in most PRIDE projects for each organism) and are represented in the top 10 most frequently detected GVPs out of all GVP detections (Table 1). The high prevalence of these GVPs within projects from a single organism and generally observed frequently across projects from different organisms suggests that these robustly detected GVPs could potentially be utilized as biomarkers of the "dark" proteome, as evidence of trace components.

Table 4. Most frequently detected GVPs across PRIDE projects for each source organism.

| GVP | *D. melanogaster* | *S. cerevisiae* | *E. coli* | *A. thaliana* | *M. musculus* |
|---|---|---|---|---|---|
| VTMQNLNDR | 7 | 8 | 7 | 7 | 8 |
| LEQEIATYR | 7 | 6 | 4 | 7 | 6 |
| LAADDFR | 8 | 8 | 6 | 7 | 10 |
| AQYEEIAQR | 7 | 8 | 7 | 6 | 8 |
| FASFIDK | 6 | 10 | 5 | 7 | 8 |

When considering common GVPs across all source organisms, we observed 22 human contaminant GVPs to be detected in at least one project from each of the 5 source organisms of interest (Figure 8). Interestingly, this common set of GVPs is approximately equivalent to the number of GVPs only detected in *M. musculus* and *E. coli* (23 GVPs and 26 GVPs, respectively), and much greater than the number of GVPs detected solely in *A. thaliana* (7 GVPs). However, 41% of detected GVPs in *S. cerevisiae* and 44% of detected GVPs in *D. melanogaster* are solely detected in those respective organisms. Clearly, different sets of GVPs are detectable for each organism, but a not insignificant number are able to be detected among a diverse set of non-human organisms. This leads us to conclude that it is possible to assemble a core set of common human contaminant GVP biomarkers that are broadly detectable as part of the "dark" proteome.

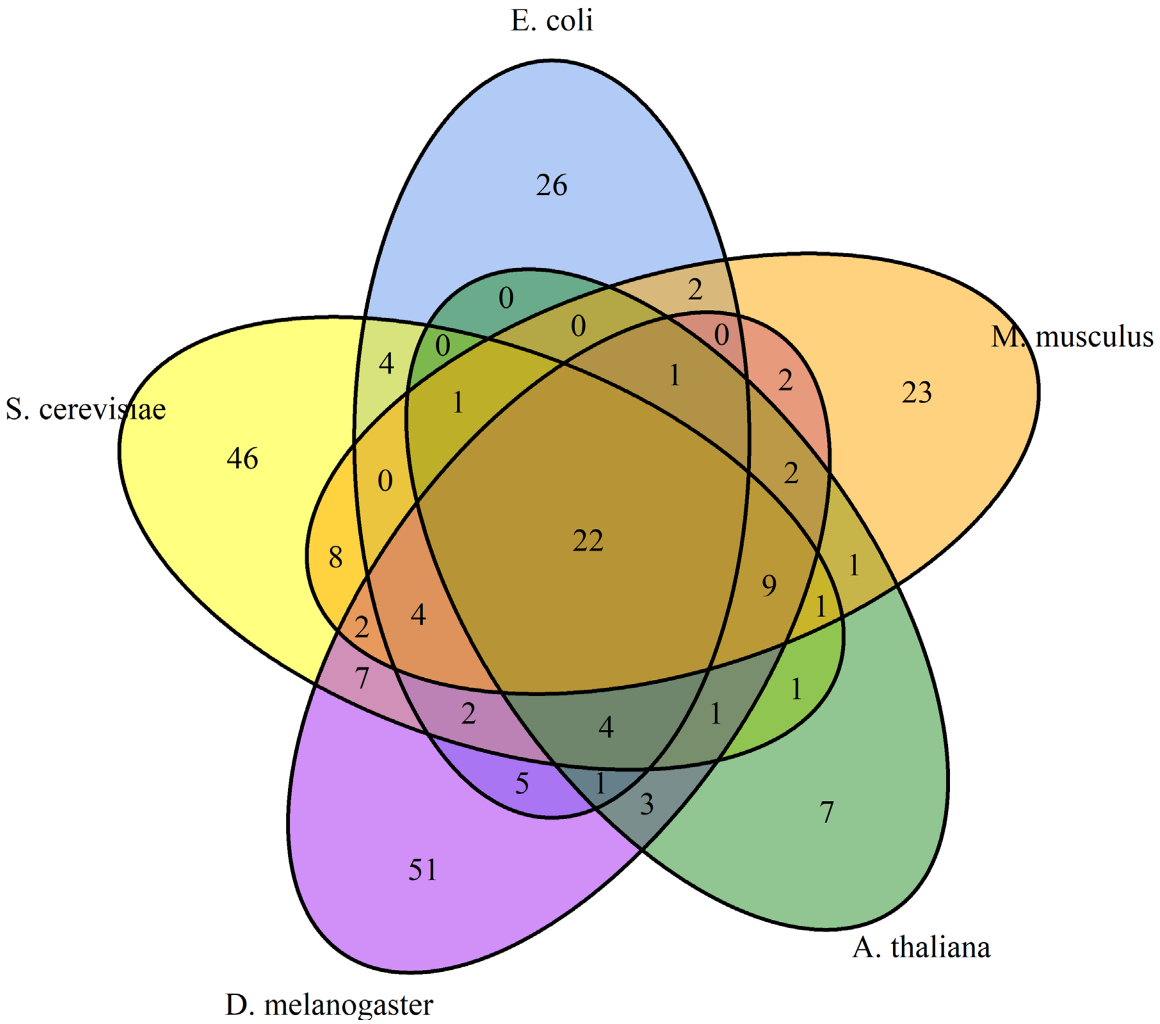

Figure 8. Venn diagram showing GVP sequence overlap among datasets from different organisms. 22 GVPs were found to be detected in at least one PRIDE project from each organism.

With the investigation of the PRIDE datasets, we have demonstrated the feasibility of human contaminant GVP detection across a diverse set of non-human organisms, characterized the biological origins of these GVPs, developed a better understanding of GVP detectability, as well as determined the feasibility of establishing a robust set of common GVPs as indicators of human contamination to uncover more of the “dark” proteome.

However, additional questions regarding human contaminant GVP detection remain. Chiefly, how reproducible or variable are human contaminant GVP detections in non-human samples prepared by the same individual? To address this question, we investigate GVP detection using another set of proteomics data, described below in Section 2.3.

## 2.3 Detection from In-House Repository Datasets

We apply a similar approach for GVP detection of datafiles with our in-house proteomics data repository as well as a similar analysis of detection results to that described in Section 2.2. The primary advantage of this in-house repository is the ability to link datafiles acquired to the individual who prepared these samples, so that we can investigate the variability in GVP profiles across samples prepared by the same individual. This will allow us to establish the extent of reproducibility or variability of detecting minor protein components across different samples handled by the same individual. In the previous analysis of PRIDE projects, we could only assume that all the datafiles belonging to a single project were prepared by the same individual. But because our in-house repository tracks the entirety of the sample preparation, data acquisition, and data analysis process, we can have confidence in knowing which samples were prepared by the same individual. However, to ensure proper data handling, as per our IRB exemption protocol, all sample preparer names were deidentified and instead assigned a numeric identifier following data download from the repository.

### 2.3.1 Dataset Selection

To select projects (termed campaigns henceforth), we surveyed the number of datafiles acquired between 2015 and 2020, grouped by organism. The top 20 are shown in Figure 9. We selected organisms based on the following criteria: (1) non-human, (2) phylogenetic diversity, and (3) a subset of organisms that are the same as those organisms selected from PRIDE projects. We ultimately decided on datafiles from samples originating from *M. musculus*, *A. thaliana*, *Rhodosporidium toruloides*, *E. coli* (BL21 strain), and *Bos taurus*.

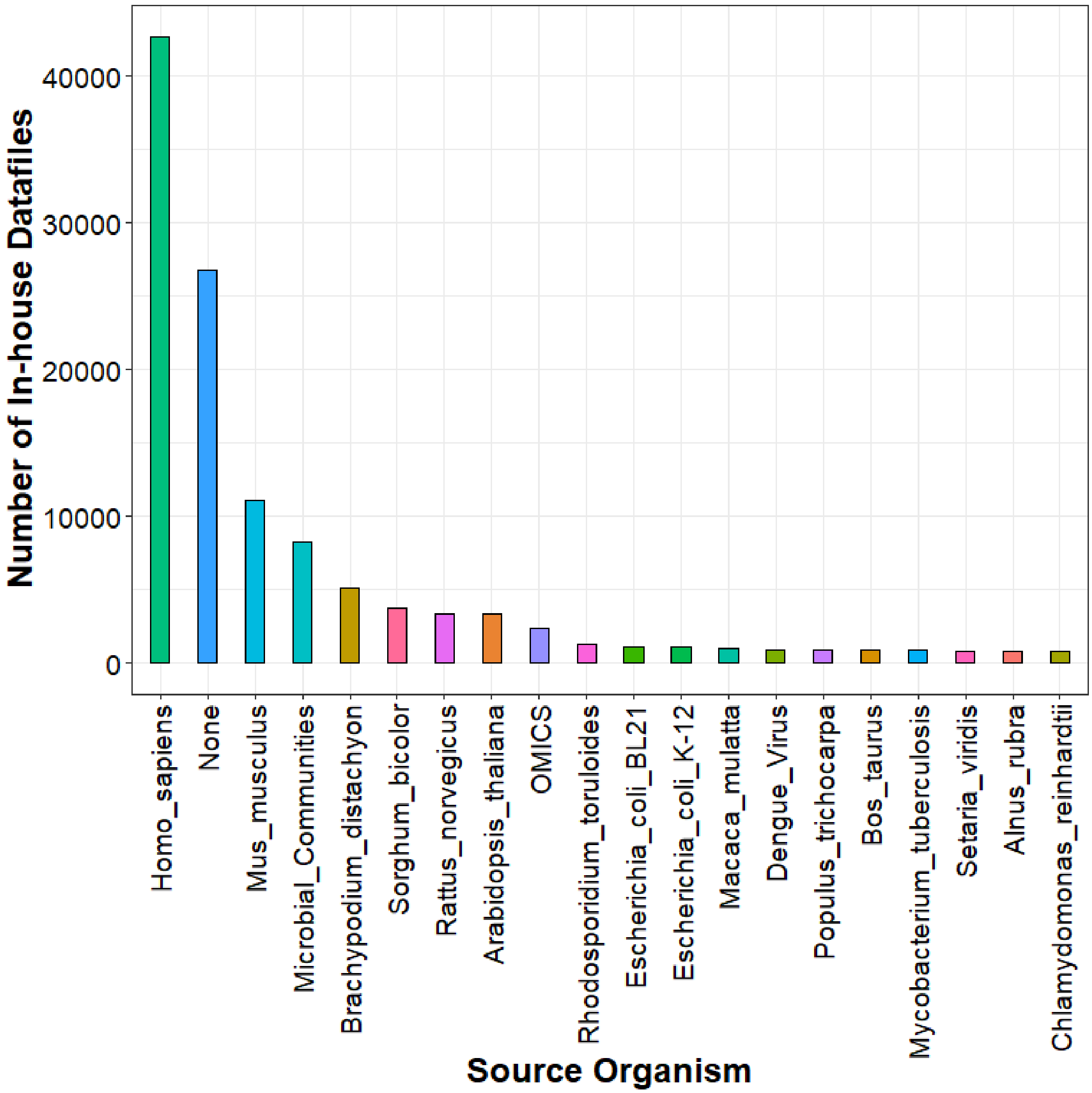

Figure 9. Barplot of the distribution of the number of data-dependent acquisition mass spectrometry datafiles deposited into the in-house repository by source organism, between 2015 and 2020. Datafiles are further associated with campaigns.

Of the selected organisms, we then surveyed a subset of campaigns (with the approval of the respective project managers for data reuse) that were prepared by different individuals. Figure 10 below displays the distribution of campaigns along with the number of datafiles per campaign for each sample preparer. We elected to use datafiles corresponding to samples prepared by individual 37, owing to the individual having prepared a great number of samples from most of the organisms in consideration (4 out of 5 organisms).

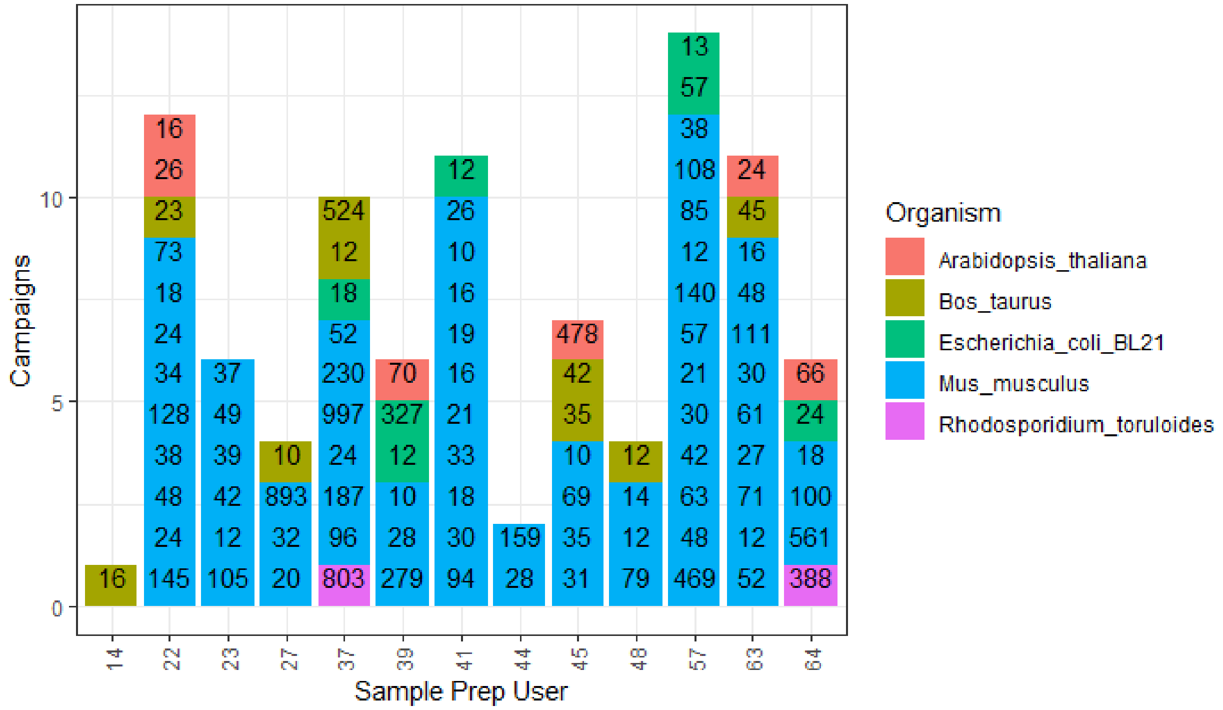

Figure 10. Barplot of the number of campaigns and datafiles, grouped by organism, of samples prepared by each individual, between 2015 and 2020. Sample preparer 37 prepared the largest number of samples that exhibit the greatest organism diversity.

In total, 10 campaigns, with datafiles ranging between 16 and 2059 (on average, 564 ± 697 (s.d.) datafiles), were considered for GVP detection analysis, to examine variability of GVP profiles from samples prepared by the same individual.

### 2.3.2 GVP Detection from a Single Sample Preparer

Using our in-house repository of proteomics data, we again demonstrate the feasibility of human contaminant GVP detection that is broadly applicable to non-human organisms. Of the 10 DDA projects selected for database search and GVP detection, we found that 9 campaigns contain detectable GVPs at both 1% and 5% peptide-level FDR control. Figure 11 below displays the number of detected unique GVP sequences per campaign, by organism, at 1% peptide-level FDR control. One striking observation with these campaigns is the low number of GVP detections across the four organisms (on average, 11 ± 7 (s.d.) GVPs per campaign), compared to the results observed for the PRIDE projects at the same 1% FDR level.

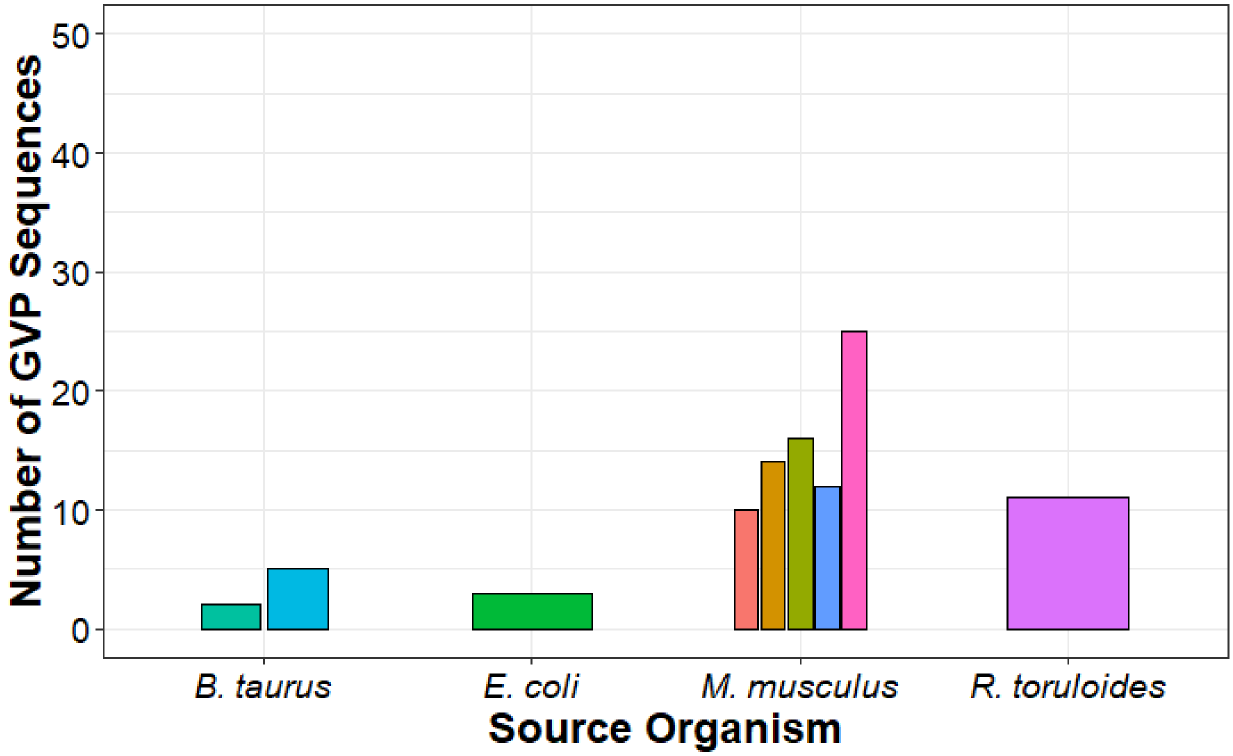

Figure 11. Barplot displaying the number of unique GVP sequences detected from in-house repository campaigns (each represented by a single bar), grouped by source organism, at 1% peptide-level FDR control, where each sample within the campaign was prepared by sample preparer 37.

We note that in contrast to the database search performed on the PRIDE datasets, the only post-translational modification considered for these datasets is cysteine carbamidomethylation. This difference likely contributes to fewer detected GVPs across the various campaigns. Interestingly enough, however, the *M. musculus* campaigns yielded higher number of detected GVPs (on average, 15 ± 6 (s.d.) GVPs) compared to campaigns from the other organisms, whereas the opposite trend was observed in PRIDE projects. This is likely owing to a greater number of datafiles within the *M. musculus* campaigns (between 24 and 2059 datafiles) than for the other organisms and also compared to *M. musculus* PRIDE projects.

At 5% peptide-level FDR control, we again observe a slight increase in number of detected human contaminant GVPs compared to 1% FDR, and this increase in detections is especially apparent for *E. coli* (15 GVPs), *M. musculus* (on average, 22 ± 10 (s.d) GVPs), and *R. toruloides* campaigns (26 GVPs).

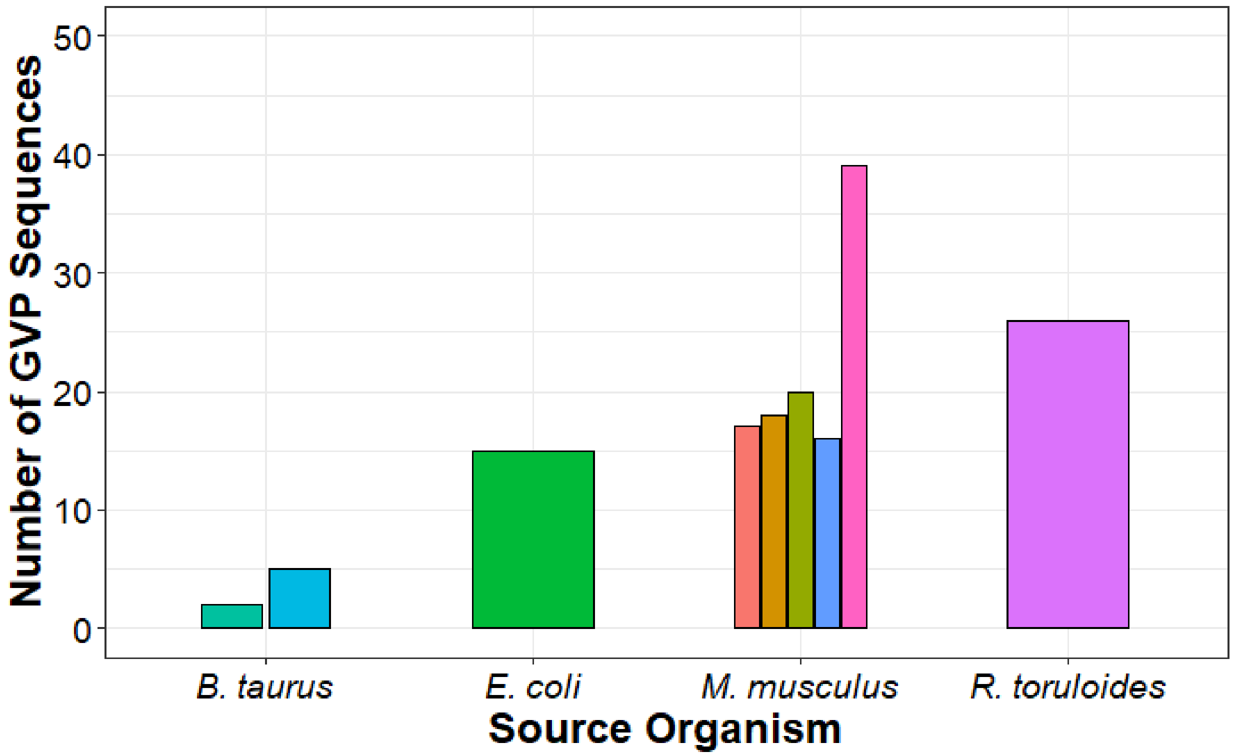

Figure 12. Barplot displaying the number of unique GVP sequences detected from in-house repository campaigns (each represented by a single bar), grouped by source organism, at 5% peptide-level FDR control, where each sample within the campaign was prepared by sample preparer 37.

Following further confirmation of feasibility of human contaminant GVP detection in proteomics data, we investigate, from the set of detected GVPs at 5% FDR, biological sources of these GVPs, conditions related to detectability, as well as feasibility of detecting a common set of these minor components derived from the same individual, across various non-human samples.

### 2.3.3 Proteins and Genes Associated with GVPs

Again, we observe that the most frequently detected human contaminant GVPs derive from cytoskeletal keratins, some of which are known to be abundant in skin (Table 5). Many of these GVPs (i.e., 8 GVPs) are also reported as most frequently detected GVPs in PRIDE projects (Table 1). Note that the organisms investigated in these campaigns are slightly different than the set examined in Section 2.2, indicating that there likely exists a core set of human contaminant GVPs (including a subset of the ones listed below) that are easily and broadly detectable across non-human organisms.

Table 5. List of top 11 most frequently detected human contaminant GVPs across in-house repository campaigns prepared by sample preparer 37.

| GVP sequence | Detection frequency | Gene name | Chromosome |
|---|---|---|---|
| AQYEEIAQR | 8 | KRT76 | 12 |

| | | | |
|---|---|---|---|
| VTMQNLNDR | 8 | KRT14 | 17 |
| LAADDFR | 7 | KRT13 | 17 |
| FLEQQNK | 6 | KRT6B | 12 |
| FLEQQNQVLETK | 6 | KRT74 | 12 |
| LEQEIATYR | 6 | KRT14 | 17 |
| IVLQIDNAR | 5 | KRT19 | 17 |
| ASLEAAIADAEQR | 4 | KRT8 | 12 |
| DYQELMNVK | 4 | KRT76 | 12 |
| FASFIDK | 4 | KRT75 | 12 |
| HSGIGHGQASSAVR | 4 | FLG | 1 |

Next, we examined which genes were more likely to be represented in detected human contaminant GVPs across campaigns of samples prepared by the same individual. We applied a minimum detection frequency of 2 GVPs. Six out of 9 genes in Table 6 were also reported in Table 2 as genes associated with the most GVP detections. Here, we see a mix of structural and cytoskeletal proteins, and 2 cuticular (hard) keratins, which represents a slightly different set of genes (more cytoskeletal keratins) than those associated with the most frequently detected GVPs in Table 5 above.

Table 6. List of top 9 predominant genes associated with detected human contaminant GVPs across in-house repository campaigns prepared by sample preparer 37.

| Gene name | GVP detection frequency | Chromosome |
|---|---|---|
| FLG | 17 | 1 |
| KRT13 | 4 | 17 |
| KRT74 | 4 | 12 |
| DST | 3 | 6 |
| IVL | 3 | 1 |
| KRT14 | 3 | 17 |
| KRT37 | 3 | 17 |

| | | |
|---|---|---|
| KRT77 | 3 | 12 |
| KRT83 | 3 | 12 |

### 2.3.4 GVP Protein Assignment Ambiguity

In our examination of protein assignment ambiguity for detected human contaminant GVPs, we similarly observed ambiguous assignment to mouse proteins (61 GVPs out of 158 unique GVP sequence detections across 9 campaigns) and sheep proteins (15 GVPs across 9 campaigns). Additionally, we found ambiguous protein assignment to cow proteins (25 GVPs across 9 campaigns). This was observed in both *B. taurus* campaigns as well as campaigns containing a mixture of *M. musculus* and *B. taurus* samples (though we labeled those mixed campaigns as dominated by either *M. musculus* or *B. taurus*, not both). With the exception of some GVPs mapping to mouse, sheep, or cow proteins, none of the detected contaminant GVPs mapped to the other source organisms (e.g., *E. coli*, *R. toruloides*).

Further investigation of the ambiguous mappings showed assignments to mouse, sheep, and cow keratins. As keratins are generally known to be highly homologous, it is not surprising to find non-human keratins as additional assignments of human contaminant GVPs (e.g., Keratin, type II cytoskeletal 73, from the KRT73 gene in cow, Keratin, type II cytoskeletal 2 oral, from the Krt76 gene in mouse).

Taking these observations and similarly those from Section 2.2.4, we can have more confidence that any detected human contaminant GVP from non-human proteomics data likely derives from human proteins rather than from the non-human source.

There are instances of a particular GVP sequence ambiguously assigned to multiple human proteins (e.g., DNAELENLIR mapped to KRT33B and KRT31); all instances are already reported in Table 3 above. All ambiguous human protein assignments observed here are also to keratins, which again, is not surprising owing to their high degree of homology to carry out similar structural functions.

### 2.3.5 GVP Alleles

Across the 9 campaigns, we observe human contaminant GVPs containing only a single SNP as the most prevalent condition (Figure 13). This aligns with observations in Section 2.2.5. All human contaminant GVPs with ambiguous assignments to human keratins and ambiguous assignments to different combinations of SNPs were removed for this analysis. Notably, the 7 GVPs detected in *B. taurus* campaigns only contain a single SNP. The maximum number of SNPs contained within a single GVP sequence for this set of proteomics data remains 5 SNPs, all carried within the gene FLG, which codes for filaggrin, a structural protein.

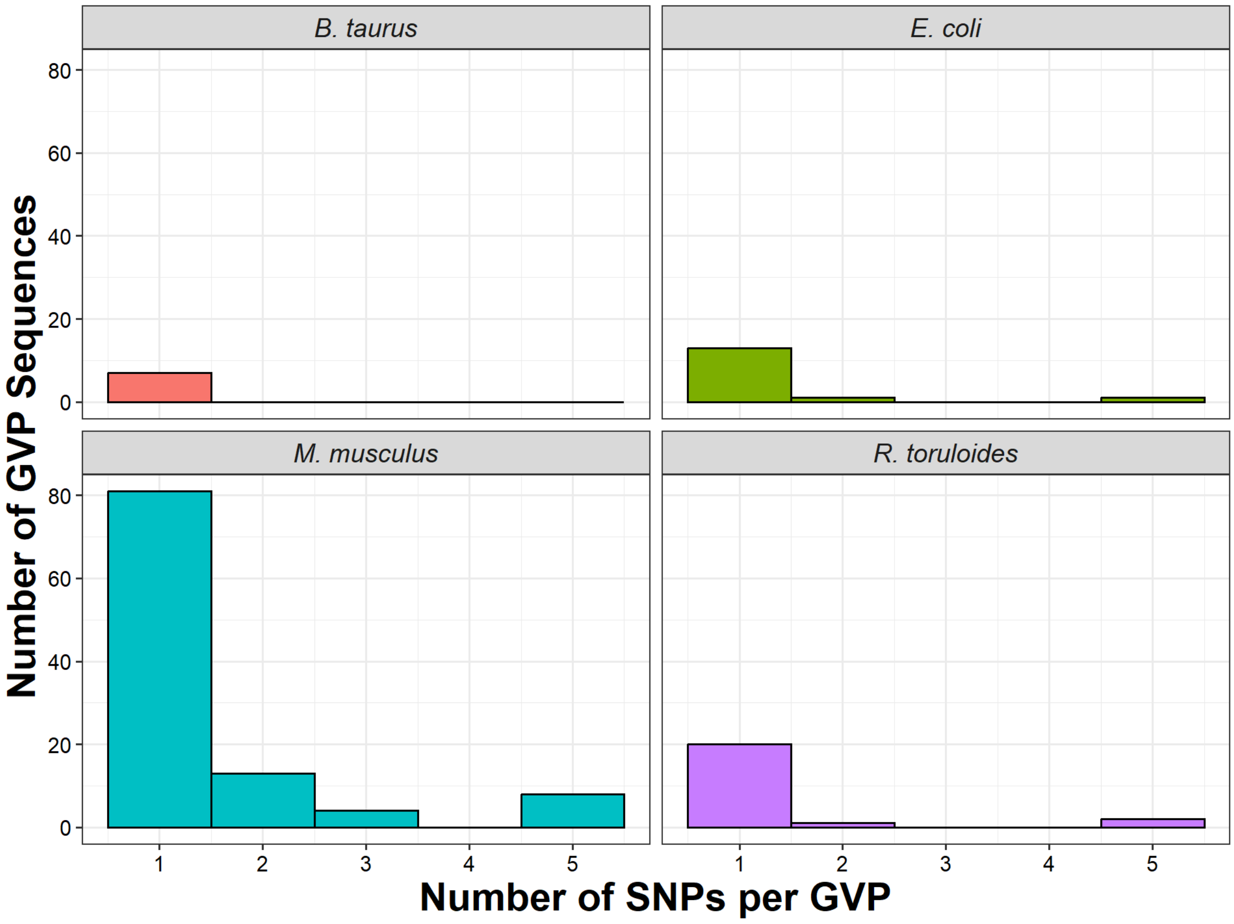

Figure 13. Histogram displaying the distribution of the number of SNPs contained in each detected human contaminant GVP, grouped by organism, across the in-house repository campaigns handled by sample preparer 37. Most GVPs contain only a single SNP.

We find that the distribution of variant types in this set of proteomics data exhibits little variability (Figure 14). For example, the variant type distribution is very similar across the *E. coli* and *R. toruloides* campaigns. All 7 detected GVPs from *B. taurus* campaigns are reference alleles. Clearly, the dominant form in detected GVPs is the reference allele. However, it is interesting that a large number of detected GVPs from *M. musculus* campaigns contain multiple SNPs that include both the reference and alternate alleles (17 out of 106 GVPs from *M. musculus* campaigns), even greater than the number of GVPs containing only the alternate allele for a single SNP (11 out of 106 GVPs). These alleles are the SNP products carried in FLG, KRT23, and KRTAP4-8, and on average, contain 4 ± 1 (s.d.) SNPs.

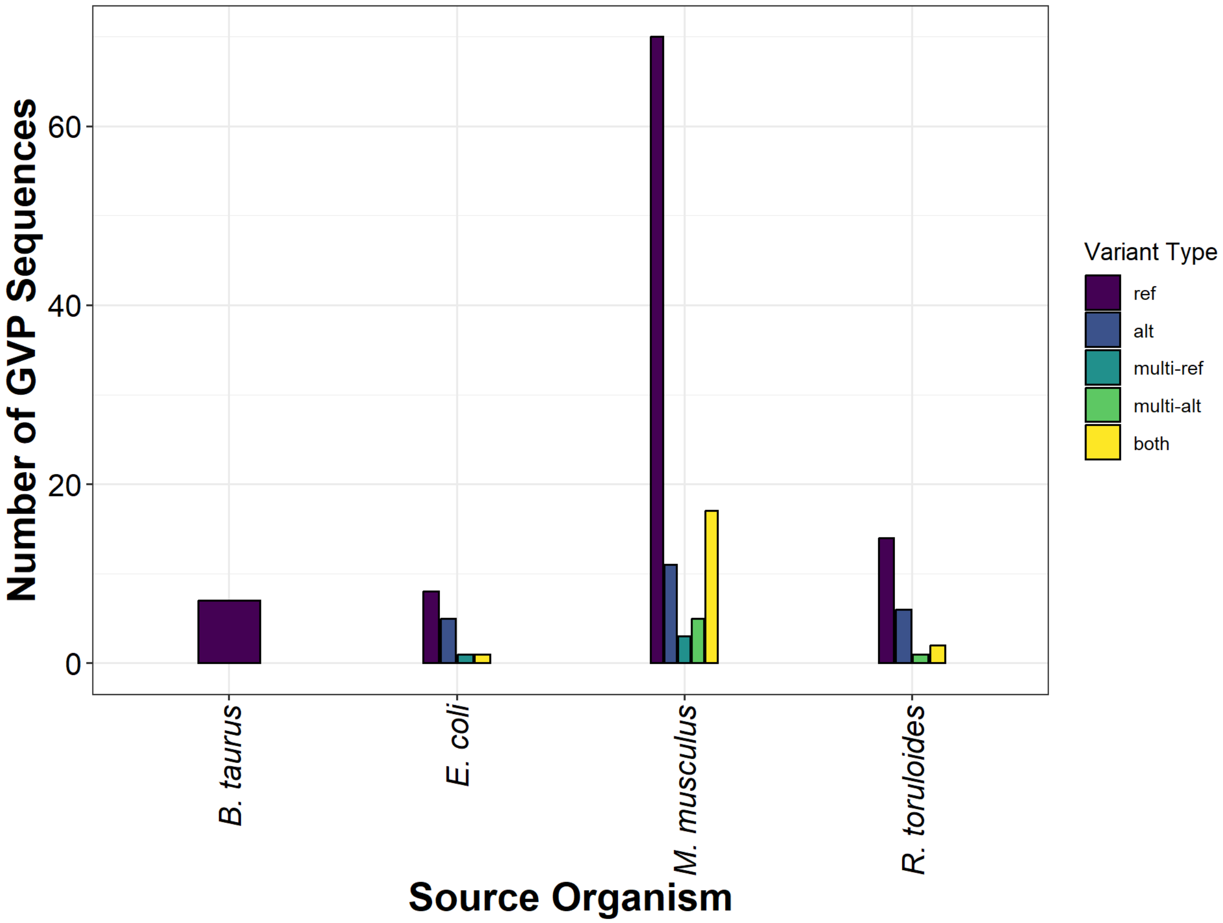

Figure 14. Barplot displaying the distribution of variant types among detected human contaminant GVPs, grouped by source organism, for campaigns from the in-house repository of samples handled by sample preparer 37. Variant types include those with single SNPs (ref = reference, alt = alternate) and for those GVPs containing multiple SNPs (multi-ref = multiple reference alleles, multi-alt = multiple alternate alleles, both = containing both reference and alternate alleles in some combination). Notably, the reference alleles are most prevalent in detected GVPs, across all organisms.

Slight differences exist between the distribution of variant type by gene type observed across these campaigns and that across PRIDE projects, likely owing to having a more modest total number of human contaminant GVPs detected across campaigns compared to PRIDE projects. In Figure 15 below, we find that alternate alleles occur much more frequently than reference alleles in GVPs from KAPs (though the gap is not very wide), reference alleles dominate GVPs from keratins, and multiple SNPs are more often contained in GVPs from structural proteins. Detection of the alternate alleles in this set of proteomics data is fairly low compared to PRIDE projects (Figure 6). Given that one of the main differences in this database search is the inclusion of fewer post-translational modifications, we hypothesize that this constraint and overall fewer detections of alternate alleles may be related.

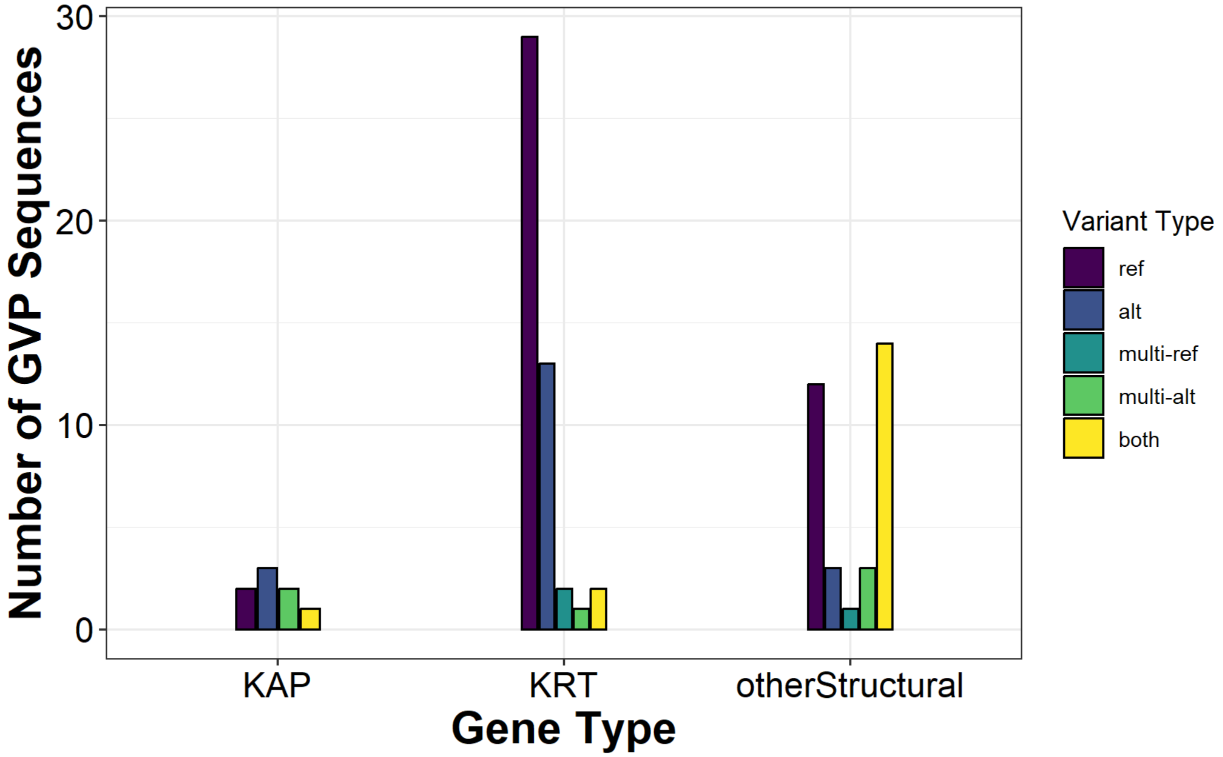

Figure 15. Barplot displaying distribution of GVP variant type, by type of gene associated with GVP sequence, among in-house repository campaigns prepared by sample preparer 37. Gene types include genes resulting in keratins (KRT), keratin-associated proteins (KAP), and other structural proteins. Variant types include those with single SNPs (ref = reference, alt = alternate) and for those GVPs containing multiple SNPs (multi-ref = multiple reference alleles, multi-alt = multiple alternate alleles, both = containing both reference and alternate alleles in some combination).

One commonality to the distribution of variant types observed here to the PRIDE project results is the similarity of the variant type profile for GVPs belonging to keratins to that of the profiles belonging to 2 out of the 4 organisms studied here: *E. coli* and *R. toruloides*. It is clear that the vast majority of detected human contaminant GVPs derive from keratins, and as such, the resemblance in variant type profiles is reasonable. It is more likely that detected human contaminant GVPs derive from keratins and contain only a single SNP exhibiting the reference allele, though detection of the alternate allele is also quite possible, depending on the SNP and proteotypicity of the alternate allele.

The observation that there exists some differences in the other organisms' variant type profiles likely implies that a different set of GVPs may be detectable among different non-human organisms, despite the samples having been handled by the same individual.

We examine the degree to which these GVPs are reproducible among different non-human samples handled by the same individual in the next section.

### 2.3.6 Common GVPs

Analysis of common GVPs within and across campaigns handled by the same individual can provide us with insight into sample-to-sample reproducibility or variability expected in detected human contaminant GVP profiles. In an ideal and complete peptide detection analysis, we would expect a highly reproducible set of detected GVPs, with little-to-no variation from proteomics sample-to-sample. However, given that DDA mass spectrometry yields an incomplete analysis, in that only the most abundant peptides are typically detected, detection of the trace components that are human contaminant GVPs in non-human samples are even more at a disadvantage. Thus, to demonstrate confidence in human contaminant GVP detection, with the alternative hypothesis being that they are spurious detections, it is important to characterize the extent to which we can detect the same GVPs in different samples when we expect the same GVPs to be found owing to the same human individual origin.

We first investigate how often GVP sequences are detected in multiple datafiles from the same campaign. Figure 16 below displays the distribution of GVP sequence detection within each campaign. Because only a single *R. toruloides* and *E. coli* campaign, respectively, were selected for analysis, no conclusions can be drawn regarding repeated GVP detections in multiple campaigns for these organisms. But for the other two organisms, it is obvious that most detected GVPs are only found in a single campaign, suggesting that human contaminant GVP detection can be quite variable in proteomics data from the same source organism, even when the samples are prepared by the same individual.

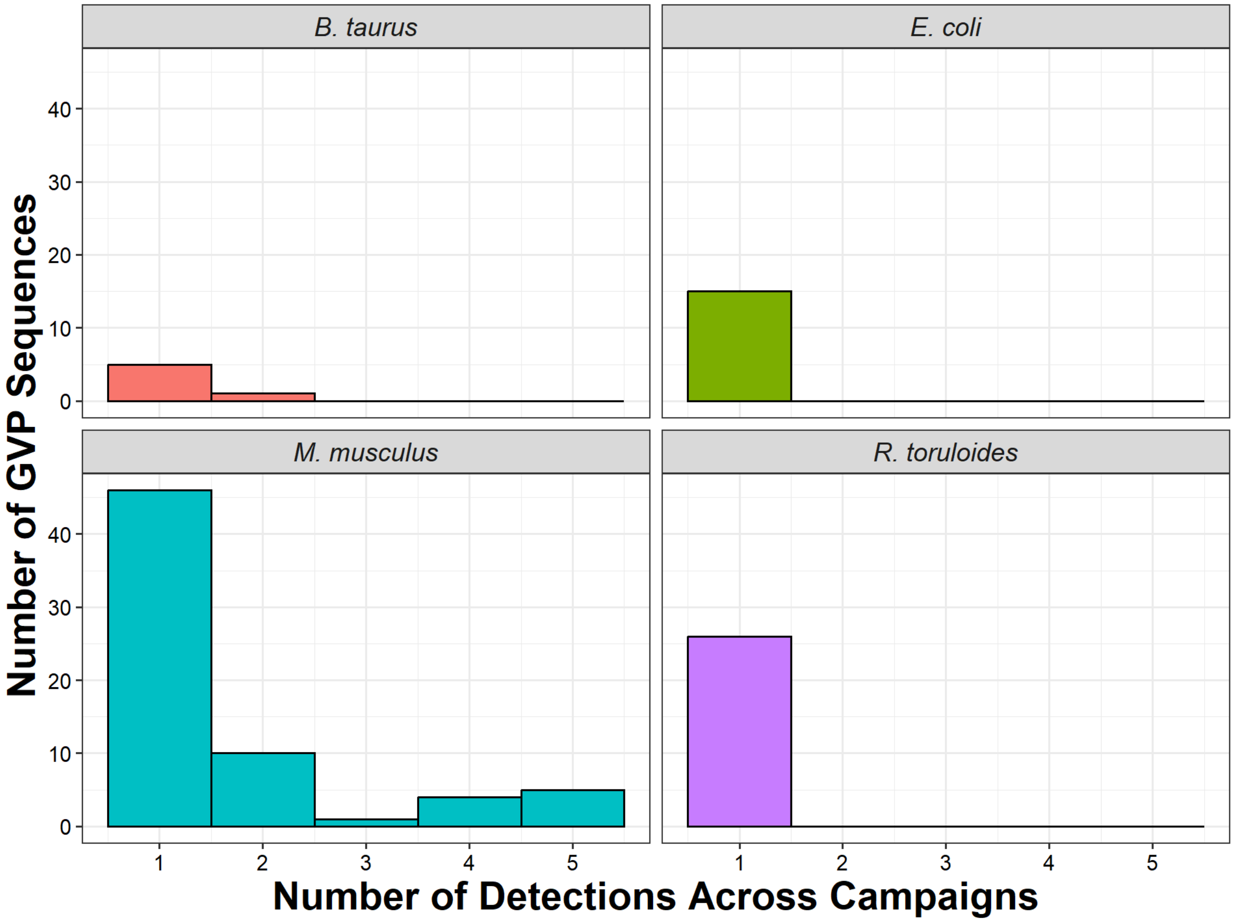

Figure 16. Histogram displaying distribution of number of human contaminant GVPs most often detected from select in-house repository campaigns for each source organism, wherein each sample from these campaigns was handled by sample preparer 37.

We then examined the feasibility in identifying a common set of detected human contaminant GVPs across all four non-human organisms of interest, wherein the GVP is detected in at least one campaign per organism. Figure 17 below displays the overlap in GVPs among campaigns from different organisms. Unfortunately, only a single human contaminant GVP (i.e., VTMQNLNDR) was found to be ubiquitous across the four organisms. With the exception of *B. taurus* campaigns, most of the detected human contaminant GVPs are only found in a single organism. Given the results from this analysis and the previous examination of GVP detection in multiple campaigns from the same organism, we observe a fair amount of variability in GVP detection even when the samples are handled by the same individual, and the extent of this variability may differ from source organism-to-organism.

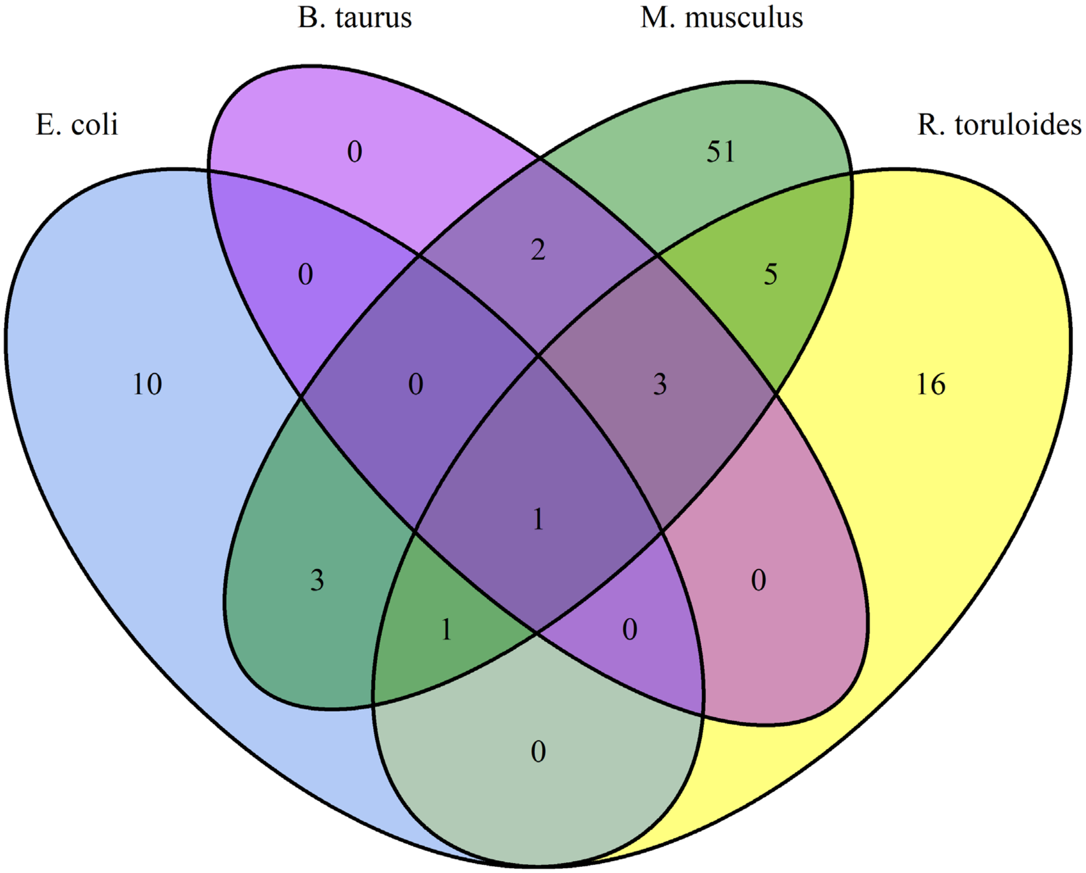

Figure 17. Venn diagram showing GVP sequence overlap among campaigns from different organisms. Only a single human contaminant GVP sequence overlap among all four organisms.

Human contaminant GVPs with higher detections and observed in at least two organisms are represented in Table 7 below. Note the scarcity of detections within each organism for each GVP owing to the limited number of campaigns investigated here. Not unexpectedly, there are more instances of GVP detection in two organisms as opposed to three organisms, again suggesting organism-related variability in GVP detection. However, despite the variability of these detections across organisms, many of these frequently detected GVPs are also reported as frequently detected in PRIDE projects (Table 4), thus providing confidence that these GVP detections are real, as opposed to spurious detections, despite the modest set of campaigns considered.

Table 7. Most frequently detected GVPs across campaigns handled by sample preparer 37 for each source organism, where each GVP is observed in more than one organism.

| GVP | *M. musculus* | *R. toruloides* | *E. coli* | *B. taurus* |
|---|---|---|---|---|
| DNAELENLIR | 2 | 1 | 0 | 0 |
| VTMQNLNDR | 5 | 1 | 1 | 1 |
| AQYEEIAQR | 5 | 1 | 0 | 2 |
| DSLENTLTESEAR | 1 | 1 | 0 | 0 |
| DYQELMNVK | 2 | 1 | 0 | 1 |
| FLEQQNK | 5 | 1 | 0 | 0 |
| LQFYQNR | 1 | 1 | 0 | 0 |
| LAADDFR | 5 | 1 | 0 | 1 |
| LEQEIATYR | 5 | 0 | 0 | 1 |
| SISVSVAGGALWGR | 1 | 0 | 1 | 0 |
| FLEQQNQVLETK | 4 | 1 | 1 | 0 |
| FASFIDK | 3 | 1 | 0 | 0 |
| LASELNHVQEVLEGYK | 2 | 0 | 1 | 0 |
| QVVSSSEQLQSYQAEIIELR | 1 | 0 | 1 | 0 |

We examined a limited number of campaigns here to investigate human contaminant GVP profile variability in non-human organisms with samples prepared by the same individual. But from these results, it is difficult to identify a set of reproducibly detected GVPs when we expect to find them. This difficulty in reproducible detection may be owing to the combination of challenges with incomplete peptide detection and abundance of minor human contaminant proteins. Additionally, human biology may play a role, as it is known that the extent of human contamination (e.g., via skin cell shedding) varies from individual-to-individual.

## 2.4 Comparison of GVP Detection Performance

The previous Section, Section 2.3, provided insight into human contaminant GVP detectability performance the two sets of proteomics data: PRIDE projects and in-house repository campaigns of samples handled by sample preparer 37. In this section, we aim to compare similarity of sets of detected human contaminant GVPs from these two efforts and draw general conclusions regarding GVP detection in DDA mass spectrometry-based proteomics data.

Similar human contaminant GVPs can be detected in proteomics datasets from different data sources, which provides further confidence that detected human contaminant GVPs from one source are real detections—not spurious—that can be reproduced, to some extent, in another data source, regardless of source organism. Figure 18 below displays the overlap in detected GVPs between all PRIDE projects and all campaigns prepared by sample preparer 37. Although many more GVP sequences are unique to PRIDE projects, a substantial number are shared with those detected in the campaigns.

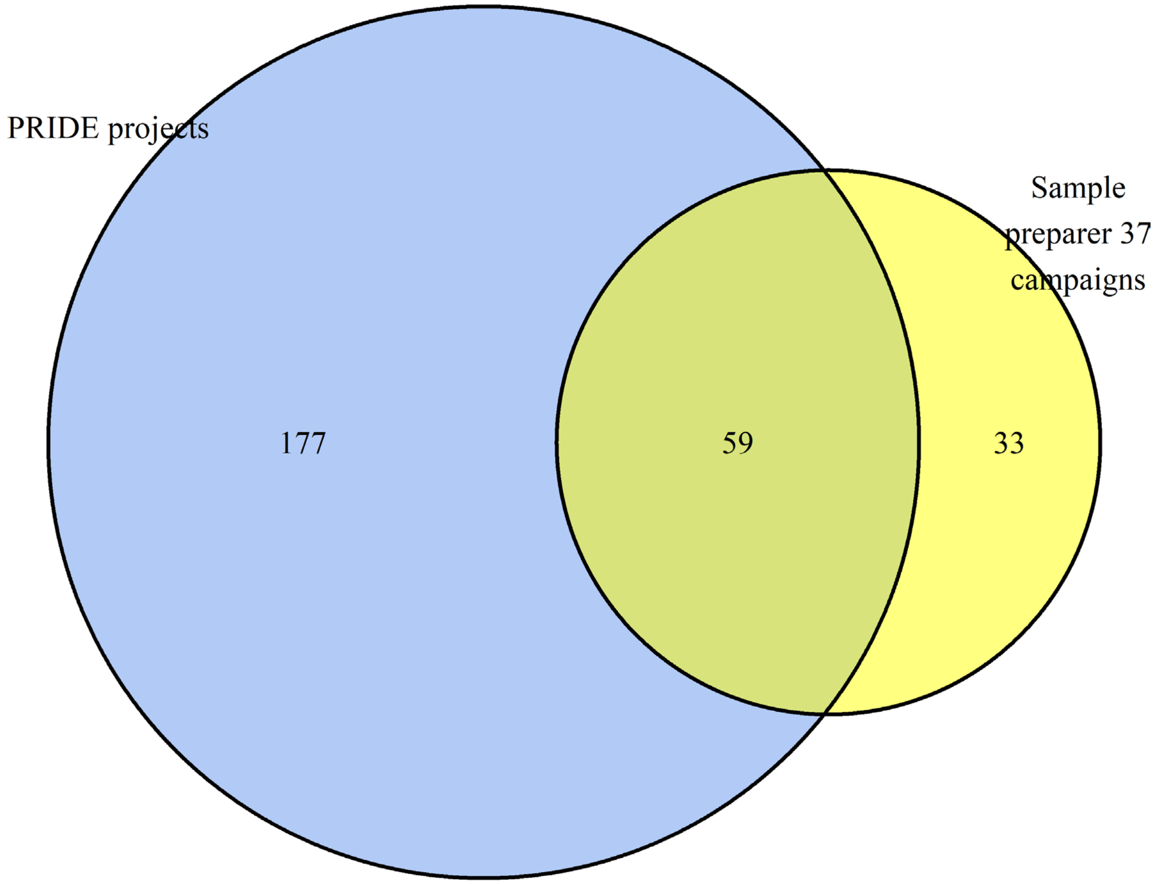

Figure 18. Venn diagram of human contaminant GVP overlap between GVPs detected in PRIDE projects and the in-house repository campaigns prepared by sample preparer 37.

In the previous sections, there were indications that human contaminant GVP detection may vary from organism-to-organism. To investigate the extent to which GVP detection in the two proteomics datasets containing the same source organism could yield similar GVPs, we further examined the overlap in detected human contaminant GVPs in *E. coli* (Figure 19) and *M. musculus* datasets (Figure 20), respectively. Datasets from these species were selected for analysis in both the PRIDE projects and in-house repository campaigns prepared by sample preparer 37. Note that the *E. coli* strain selected for analysis are different in these two sets of proteomics data: proteomics data from *E. coli* K-12 strain were examined in PRIDE projects while those belonging to the BL21 strain were examined in the campaigns. Given the difference in strain, we may observe some differences in detected GVPs.

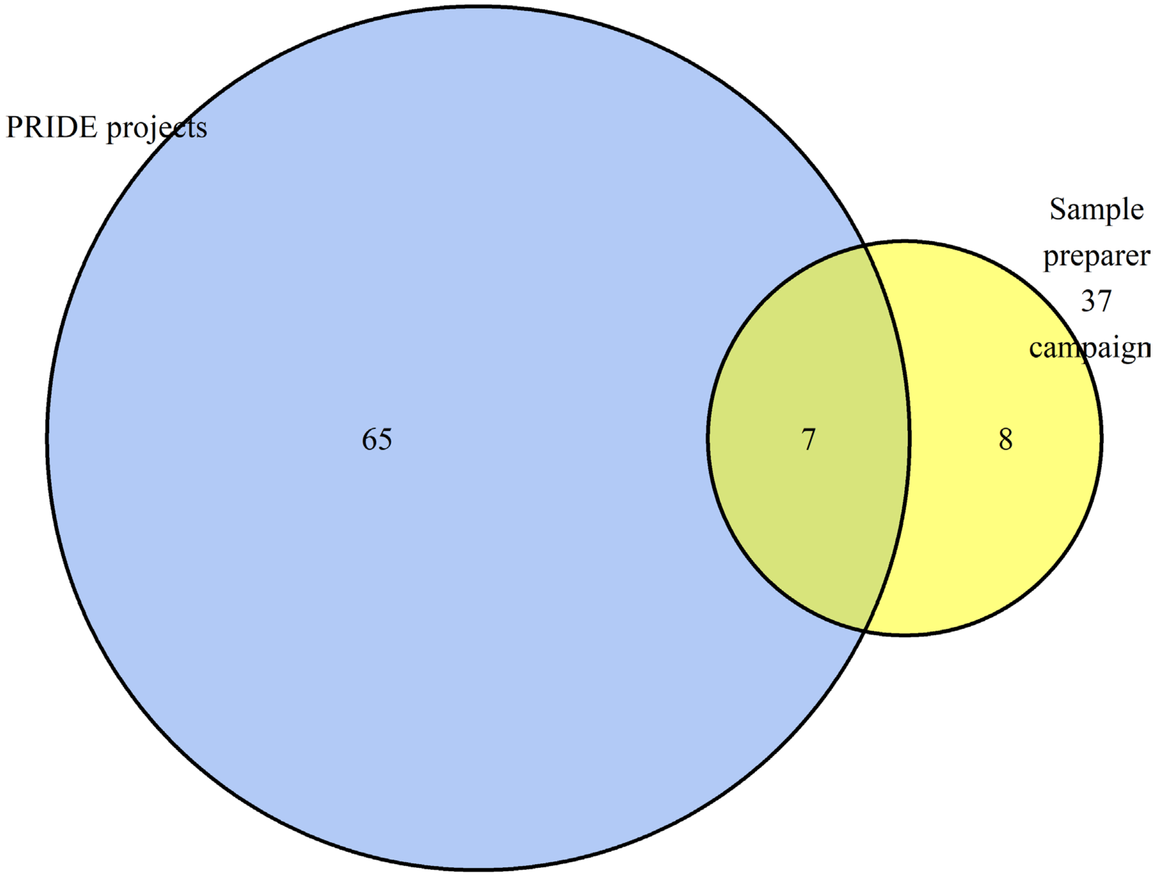

Figure 19. Venn diagram of human contaminant GVP overlap between GVPs detected in PRIDE *E. coli* (K-12 strain) projects and the in-house repository *E. coli* (BL21 strain) campaigns prepared by sample preparer 37.

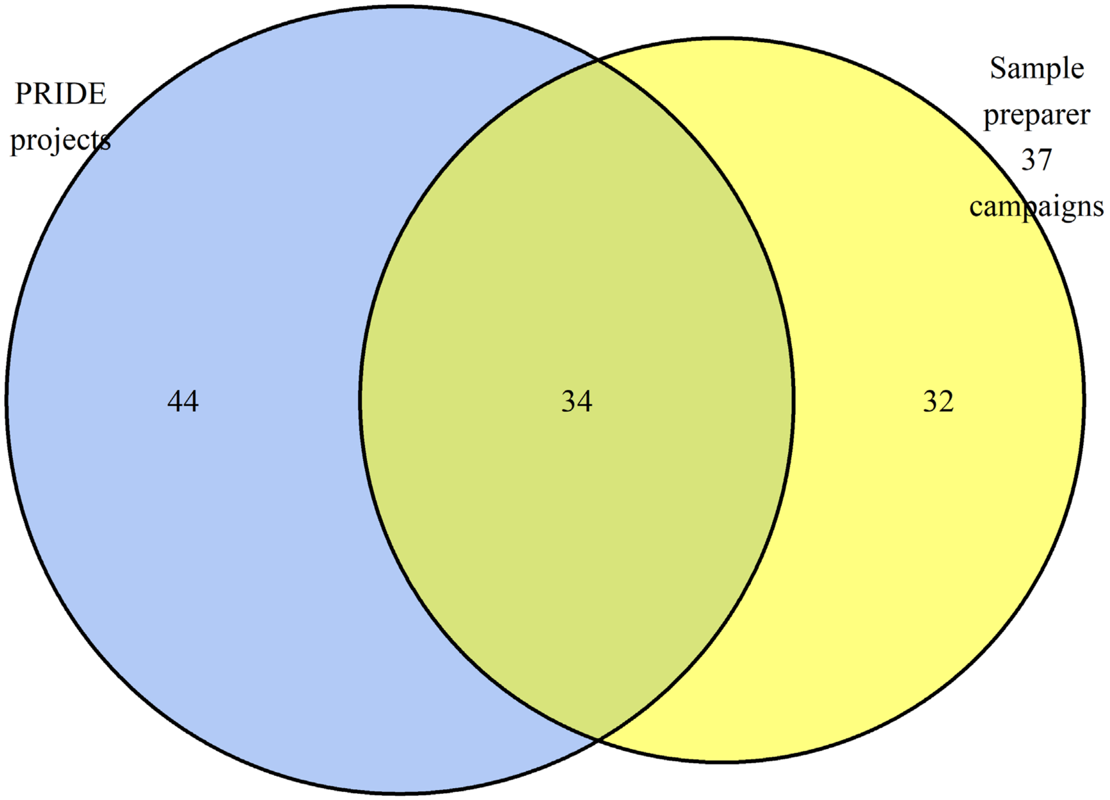

Figure 20. Venn diagram of human contaminant GVP overlap between GVPs detected in PRIDE *M. musculus* projects and the in-house repository *M. musculus* campaigns prepared by sample preparer 37.

We found that many more GVPs are shared between the *M. musculus* proteomics datasets from the two different data sources (Figure 20), compared to the GVP overlap in *E. coli* datasets (Figure 19). However, because of an imbalance of *E. coli* datasets between the two data sources—only a single *E. coli* campaign was analyzed, whereas 11 *E. coli* PRIDE projects were analyzed—few conclusions can be drawn regarding the GVP overlap and any effect of *E. coli* strain differences. We expect an increase in GVP overlap with analysis of more datasets, and with more balanced data. On the other hand, the number of analyzed *M. musculus* datasets between the two data sources are more balanced, and we see many more detected GVPs being shared (Figure 20). In fact, the number of shared GVPs in *M. musculus* datasets makes up the majority of shared GVPs across all proteomics datasets from both data sources. This observation again highlights the feasibility of detecting a common set of human contaminant GVPs in proteomics data from different data sources and among datasets from the same and different source organisms, respectively.

Taking these observations in GVP detection overlap between the two data sources, we can clearly see that while there exists some sample-to-sample variability in human contaminant GVP detection, there are a subset of GVPs that are more frequently detected within and broadly detected among proteomics datasets, regardless of source organism and data source. Implications of this effort include the potential of human contaminant GVP detection to advance understanding of the “dark” proteome, for a more complete characterization of a proteomic sample.

## 2.5 Conclusions

Overall, we not only demonstrate the feasibility of human contaminant GVP detection broadly in proteomics data, considering many more non-human organisms than previous efforts, but also develop a better understanding of GVP detectability, characterize the sample-to-sample variability in GVP detection, even when we expect to observe a similar profile, and identify a core set of GVPs that can potentially be used as markers indicative of the human contaminant traces portion of the “dark” proteome.

This effort focused on a portion of the “dark” proteome, and was primarily a qualitative assessment, but future efforts to continue to elucidate the “dark” proteome could provide a quantitative assessment of the extent to which elucidation of human contaminant GVPs provides a more complete protein analysis beyond the conventional consideration of the source organism. Investigation of other, unexpected traces and minor protein components belonging to the “dark” proteome could also provide additional insight.

That being said, alternative peptide detection strategies, such as data-independent acquisition (DIA) mass spectrometry, could be valuable for more extensive elucidation of the “dark” proteome, including human contaminant GVP detection, to enable a more complete proteomic sample analysis.

Additional capabilities compatible with human contaminant GVP detection can also be examined. One such capability is unknown source organism characterization, which enables an untargeted, unbiased approach to identifying the source organism in an unknown proteomic sample. This approach has demonstrated applicability in forensic and clinical samples. The next section, Section 3.0, examines the feasibility and accuracy of this combined capability, and assesses for any effects on human contaminant GVP detection.

# 3.0 MARLOWE-GVP Analysis

Standard mass spectrometry-based proteomics data analysis methods are predicated on knowing the organism composition of the sample being analyzed. Specifically, knowledge of the taxonomic composition of the sample is required to create the protein sequence database that is used during the database search step. However, there are samples, such as forensic and metaproteomics samples, where the organism composition is not known. In these situations, additional data analysis is required to first determine what species are present in the sample.

Several strategies have been developed for utilizing proteomics to understand the taxonomic composition of an unknown sample using as few assumptions about source organism as possible. One strategy is database searching of mass spectrometry-based proteomics data where the database contains proteins from all known organisms, such that one maximizes the likelihood that the unknown organisms in that sample are contained within that large database (e.g., MiCId[20]). A second method is to make no assumptions on the source organism by first performing *de novo* peptide identification and then applying species identification tools, such as UniPept and the lowest common ancestor approach[22], to assign *de novo*-sequenced peptides to organisms (e.g., MetaNovo[23]).

While these proteomics species identification tools have been successful, these methods are unable to fully characterize the protein content within a sample. This is because these methods do not account for contaminant proteins, such as media peptides and human keratins, that are artificially introduced into the sample during sample preparation. Therefore, new methods are needed for more complete characterization of samples of unknown origin.

In this work, we present a method that aims to provide more complete characterization of samples of unknown composition by combining an unknown organism characterization tool, MARLOWE (Chu et al. (2024), submitted), with genetically variant peptide (GVP) detection. GVPs are peptides that contain single amino acid polymorphisms that result from non-synonymous SNPs in protein coding regions of DNA. These peptides have been recently shown to be detectable by mass spectrometry in non-human samples[18]. We envision the ability to obtain two pieces of information through this combined, untargeted pipeline: unknown source organism identification and detection of human contaminant GVPs.

MARLOWE precedes GVP detection here, wherein MARLOWE returns a ranked list of potential organisms in a sample, which then informs the database search and subsequent GVP detection. MARLOWE utilizes *de novo* peptide sequencing of mass spectrometry data to detect highly-confident regions of peptides (called tags) that can be matched to a database of tryptic peptides. These tryptic peptides are mapped to organisms following a protein inference that relies on peptide strength[14]. From this process, a weighted score is then applied to each assigned organism, thus resulting in a list of potential source organisms by score. These potential source organisms can then be included in the database search as the source organism component, and combined with the GVP portion of the database so that GVP detection can also be performed.

We aim to examine the feasibility of this combined unknown source organism and human contaminant GVP detection capability. To our knowledge, this is the first method that aims to improve characterization of unknown samples by combining human contaminant GVP detection with organism characterization.

We demonstrate successful unknown source organism characterization and human contaminant GVP detection using this combined untargeted approach. We observe high accuracy (at least 90% correct characterization) in unknown organism characterization across a diverse set of non-human samples with varying degrees of sample complexity. Similar human contaminant GVP detection to the conventional GVP detection approach (using a database search containing the ground truth source organism) was achieved, with a high degree of similarity in GVP profiles. Through this effort that relies on an untargeted approach, we enable a more complete characterization of the "dark" proteome of unknown proteomics samples.

## 3.1 MARLOWE-GVP Pipeline Development

The bioinformatics pipeline combining MARLOWE and GVP detection is graphically depicted below (Figure 21). The subsections below describe the different steps of the pipeline in more detail. Briefly, *de novo* peptide sequencing is performed on unknown raw proteomics datasets, followed by MARLOWE, to determine a list of potential source organisms. These lists of potential organisms are then filtered, by applying a threshold determined by a trained support vector machine (SVM) machine learning model, and used in conjunction with the previously created GVP FASTA file in a subsequent database search to detect human contaminant GVPs.

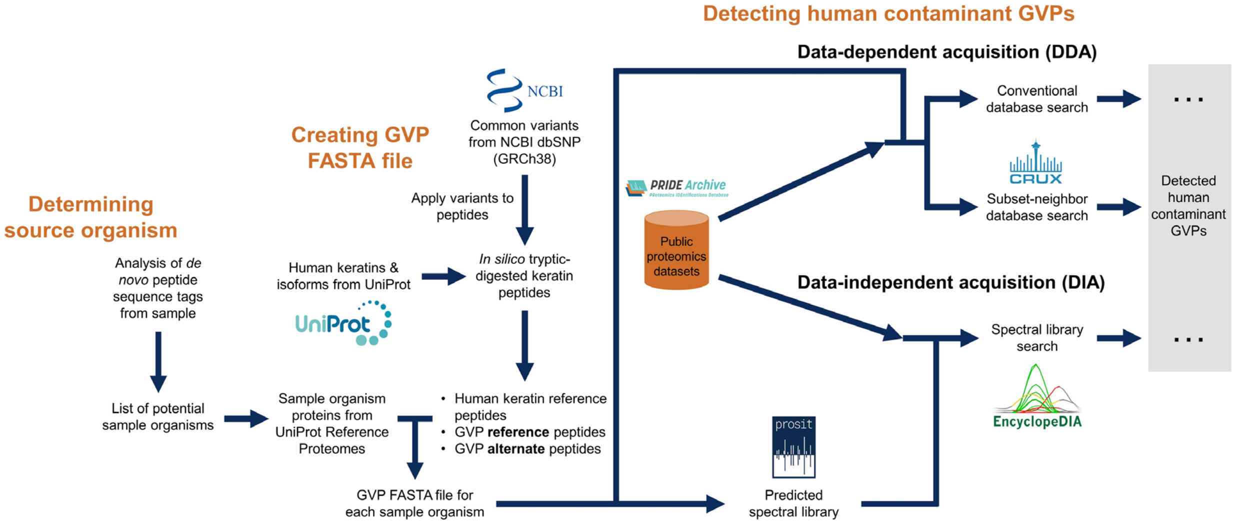
Detecting human contaminant GVPs
Data-dependent acquisition (DDA)
Creating GVP FASTA file
NCBI
Common variants from NCBI dbSNP (GRCh38)
Determining source organism
Apply variants to peptides
Analysis of *de novo* peptide sequence tags from sample
Human keratins & isoforms from UniProt
UniProt
*In silico* tryptic-digested keratin peptides
PRIDE Archive
Public proteomics datasets
Conventional database search
CRUX
Subset-neighbor database search
Detected human contaminant GVPs
Data-independent acquisition (DIA)
Spectral library search
EncyclopeDIA
List of potential sample organisms
Sample organism proteins from UniProt Reference Proteomes
Human keratin reference peptides
GVP **reference** peptides
GVP **alternate** peptides
GVP FASTA file for each sample organism
prosit
Predicted spectral library
. . .
. . .

Figure 21. Bioinformatics pipeline combining MARLOWE and GVP detection. SVM machine learning model-based filtering occurs following a list of potential source organisms produced by MARLOWE, and filtered results are then used to inform creation of the final FASTA file that combines GVP sequences of interest and source organism proteomes for database searching and GVP detection.

### 3.1.1 *De Novo* Peptide Sequencing

*De novo* peptide sequencing of mass spectrometry-based proteomics data was performed using PEAKS Online X (build 1.7.2022-08-03_160501, Bioinformatics Solutions)[37] and instrument-specific default parameters, which include preset precursor and fragment ion mass error tolerances, and fragmentation method. Post-translational modifications were included that align with those reported for each dataset.

Following *de novo* peptide sequencing, highly-confident regions of each peptide, called tags, were identified. Peptides with an average local confidence score below 50 and length below 7 amino acid residues were not considered. Each consecutive amino acid residue included in the tag (minimum length of 6 residues) must have a minimum local confidence score of 80. These lists of tags, obtained for each datafile within a project of interest, were then used as input to MARLOWE.

### 3.1.2 Source Organism Characterization with MARLOWE

Unknown source organism characterization was performed using MARLOWE, which contains the underlying KEGG (Kyoto Encyclopedia of Genes and Genomes) database of organisms' proteomes (downloaded July 2019) for tag-organism assignment. Contaminant filtering was not applied. Post-translational modifications included in the characterization aligned with those utilized in *de novo* peptide sequencing. Through application of peptide strength and protein inference, organism assignments were scored. Final potential source organism lists with associated scores were produced by filtering to organisms with at least 2 peptide assignments. These lists of potential source organisms, one list per datafile within each project, were then filtered using criteria determined by a support vector machine learning model, to identify the most promising source organism candidates for inclusion in the database search.

### 3.1.3 Support Vector Machine Model Development for Organism Filtering

We used a support vector machine (SVM) model, created using the sklearn Python package, to inform our downselection of MARLOWE's list of potential source organisms to be included in the FASTA file for the subsequent database search and GVP detection. The goal of the SVM classifier[38], which is a supervised machine learning method, is to determine the optimal support vector or hyperplane (for more than two variables) that maximizes the distance between known groups or classes in higher dimensional space. Thus, training the SVM model requires the feature input(s), i.e., the information that the model will use for class prediction, as well as the labeled classes, i.e., the groups that the model will later predict once trained.

In practice, in the training phase, our SVM model takes in different metrics associated with MARLOWE's source organism characterization as feature inputs along with a binary class label of whether the metric is associated with the correct or incorrect organism relative to the ground truth source organism. For example, a potential feature input is a list of taxonomic ranks for a single proteomics datafile, and the class labels of interest are correct and incorrect source organism. As such, within that list, there will be taxonomic ranks associated with both labels, though in this case, given that our proteomics datasets contain only a single source organism, only one taxonomic rank would be associated with the correct organism label, and the remaining ranks on the list would be labeled with incorrect organism.

Model training includes determining the support vector that maximizes the true positive and true negative determinations of correct and incorrect source organism given organism characterization metric(s) as feature input. Of the PRIDE datasets considered for this effort, we applied an 80/20 train/test split for model development. To avoid any bias in train/test splits towards any particular source organism, we applied this split at the PRIDE project level among datasets for each respective source organism.

In the test and deployment phase, the SVM model takes in the metrics resulting from MARLOWE's analysis of potential source organisms and predicts whether the metrics are associated with the correct or incorrect source organism.

We investigated different metrics from MARLOWE results that could be used as feature input by the SVM, to determine the optimal set that maximizes performance. After examining combinations of metrics as input, including taxonomic rank, normalized taxonomic score, and score ratio, we observed that many of the SVM models utilizing different combinations of these metrics as input performed similarly, with the exception of Model #3 that solely utilizes score ratio (Table 8).

Table 8. Performance of each SVM model.

| Feature Input Combinations Used | Accuracy | Precision | F1 | Classification Decision Equation |
|---|---|---|---|---|
| Taxonomic rank | 0.987 | 0.99 | 0.98 | $-2x_0 + 3.00 = 0$ |
| Normalized taxonomic score | 0.987 | 0.99 | 0.98 | $4.85x_0 - 2.11 = 0$ |
| Score ratio | 0.935 | 0.94 | 0.92 | $-2.97x_0 + 1.24 = 0$ |
| Normalized taxonomic score & rank | 0.987 | 0.99 | 0.98 | $-2x_1 + 3.00 = 0$ |
| Score ratio & rank | 0.987 | 0.99 | 0.98 | $-2x_1 + 3.00 = 0$ |
| Score ratio & normalized taxonomic score | 0.987 | 0.99 | 0.98 | $1.48x_0 + 6.47x_1 - 3.64 = 0$ |
| Rank & normalized taxonomic score & score ratio | 0.987 | 0.99 | 0.98 | $-2x_0 + 3.00 = 0$ |

SVM model results demonstrate that Model #1 using solely taxonomic rank performs as well as other models incorporating more metrics (e.g., Model #4 using both normalized taxonomic score and taxonomic rank). This indicates that the inclusion of additional feature inputs does not add value to the model in terms of improving performance. This observation is likely owing to the simplicity of datasets applied to SVM model for prediction. Since most datasets only include a single dominant organism, the SVM model is easily able to predict the correct organism, which is typically within the taxonomic group represented as the top-scoring or top-ranked hit returned by MARLOWE results.

As such, we elect to apply the simplest of SVM models, that is, Model #1 that utilizes the taxonomic group as feature input, and apply the optimal threshold determined by this model, that is, the first-ranked taxonomic group, as the threshold for filtering MARLOWE results of unknown datasets to the “true” source organism.

We expect that this approach could be easily adapted to and much more suitable for more complex datasets (i.e., datasets from samples containing more than one organism). It is likely that a different model, such as the model utilizing normalized taxonomic score and score ratio, may be more applicable and perhaps more performant on datasets exhibiting higher complexity.

### 3.1.4 MARLOWE-Informed GVP FASTA File Creation and Database Search

Based on results from the optimal SVM model in Section 3.1.3, we created FASTA files to include proteomes of organisms represented in the top-ranked taxonomic group returned by MARLOWE. Here source organism proteome FASTA file construction was customized for each datafile within a PRIDE project, as MARLOWE analysis is performed on each datafile.

To construct the source organism proteome FASTA file, as informed from MARLOWE-SVM model filtering, all organisms in the top-ranked taxonomic group with at least 2 peptide hits were retained, and their proteomes combined into a single source organism FASTA file. This file was then combined with the GVP FASTA file that contains human reference keratins, GVPs, and common contaminants as described in Frankenfield et al. (2022)[24]. Additional details regarding GVP database construction are described in Section 2.1.

We then used a database search to detect peptides and GVPs in proteomics datafiles. In this work, we used the Tide search engine[30] implemented within Crux[31]. For all of these searches, all other parameters were set to their default values, except --compute-sp=T and --pin-output=T. Only static cysteine carbamidomethylation was considered as a post-translational modification. Following the database search, all peptide detections for each datafile within the same PRIDE project were combined, and the false discovery rate was estimated using Percolator[32] within Crux. Results were filtered to not allow any matches with an FDR of more than 5% at the peptide-level[33]. The GVPs that were identified at the end of this pipeline were then reported. Prior to the database search, raw mass spectrometry files were converted to mzML format using either MSConvert within the Proteowizard[34] suite or ThermoFileRawParser[35].

## 3.2 Results of MARLOWE-GVP Analysis

### 3.2.1 MARLOWE Correct Organism Classification

Prior to applying SVM filtering, we examine results produced by conventional MARLOWE and compare correct classification rates, that is, whether the potential source organisms returned by MARLOWE match the ground truth organism.

In our manual curation of the PRIDE DDA projects, we found that 19 of the 55 projects were suitable for MARLOWE analysis. This represents 844 datafiles (across 19 projects) in total (Table 9).

Table 9. Number of datafiles, grouped by organism, from PRIDE projects on which conventional MARLOWE performance was evaluated

| Ground truth organism | Number of datafiles |
|---|---:|
| *Saccharomyces cerevisiae* | 573 |
| *Mus musculus* | 77 |
| *Escherichia coli* | 72 |
| *Arabidopsis thaliana* | 68 |

| | |
|---|---|
| *Drosophila melanogaster* | 54 |

Correct MARLOWE classification rates are reported below in Table 10. Correct classification is defined on two levels: (1) as the correct species detected within the top taxonomic group or (2) as the correct species detected within any taxonomic group from lists of potential organisms produced by MARLOWE.

We find that MARLOWE performs well across the organisms of interest, with high correct classification as the top ranked organism, and substantial improvement in correct classification for *E. coli* and *S. cerevisiae* samples when considering any rank (Table 10). This performance demonstrates that MARLOWE can easily characterize to the correct source organism when these organisms represent the primary source organism, which is expected, and is broadly applicable to a diverse set of organisms.

Table 10. Conventional MARLOWE correct classification rate for each organism, at two different levels (top rank, any rank)

| **Organism** | **Number of Datasets** | **% Correct Classification (top rank/any) *** |
|---|---|---|
| *A. thaliana* | 68 | 100/100 |
| *D. melanogaster* | 54 | 96.3/96.3 |
| *E. coli* | 72 | 88.9/**97.2** |
| *M. musculus* | 77 | 98.7/98.7 |
| *S. cerevisiae* | 573 | 91.1/**97.7** |

*Bold text indicates a sizeable increase in correct classification when the ground truth organism was detected at any rank (as opposed to the top rank)

In addition to characterizing performance by correct organism, we further wanted to examine MARLOWE's performance on samples of varying complexity. MARLOWE was originally developed on datasets derived from fairly complex, whole-cell lysates. These more complex samples ensured datasets had sufficient peptide detections belonging to multiple proteins that could inform organism source. However, we challenged MARLOWE with much simpler samples, that is, simpler peptide composition, including with datasets derived from fractionated and/or purified samples. With fractionation and/or purification, we expect detection of fewer peptides per treated sample and thus, lower protein diversity, that could be source-organism-informative.

To capture different levels of sample complexity, we categorized by experiment type (simple, moderate, or complex), which is meant to loosely classify a project based on the variety of peptides expected to be in each sample. It is expected that a greater variety of peptides (e.g., in complex samples) will make it easier for MARLOWE to correctly classify to the ground truth organism as there will be more peptide coverage and potentially more strong peptides and tags to support the characterization. The categories we used to label sample complexity are as follows:

- Complex: A whole cell lysate with no enrichment or fractionation was always labeled as complex.

- Moderate: If a sub-population of cells was isolated and used, this type of sample was labeled as moderate, as long as there were no further protein or peptide level selection steps. Projects that separated cytoplasm from membrane proteins were labeled as moderate level complexity. If a sample was a whole-cell lysate but “enriched” for certain proteins, it was labeled as moderate. Whole-cell lysates fractionated by a gel were labeled as moderate, as each portion would have a narrow range of proteins.

- Simple: Any project with immuno-purification or affinity-purification steps to isolate certain proteins or peptides was always labeled as simple. Any project that isolated complexes of proteins in combination with DNA was labeled as simple.

Interestingly, we note that MARLOWE’s performance across the varying sample complexities is comparable, and always achieving at least 90% correct classification (Table 11). This demonstrates MARLOWE’s broad applicability to analyze proteomics data for source organism characterization, as it is performant on samples spanning a wider range of sample preparation methods than intended during algorithm development.

Table 11. Conventional MARLOWE correct classification rate to the ground truth organism, organized by sample complexity.

| Sample Complexity | Number of Datasets | % Correct Classification (top rank/any) * |
|---|---|---|
| Simple | 585 | 89.9/**97.3** |
| Moderate | 176 | 98.9/99.4 |
| Complex | 83 | 98.8/98.8 |

*Bold text indicates a sizeable increase in correct classification when the ground truth organism was detected at any rank (as opposed to the top rank)

Now that we have established correct classification rates with conventional MARLOWE, given known ground truth, and demonstrated applicability not only to a diverse set of organisms but also to a wide range of sample complexities, we further examine performance of the entire MARLOWE-GVP pipeline where we may not know the source organism in unknown samples. Here, we are interested in utilizing MARLOWE to downselect to the most promising potential source organisms for a database search that includes GVP detection, and want to examine whether selection of organisms’ proteomes via an untargeted approach (i.e., MARLOWE) can affect GVP detection.

Our intuition is to avoid utilizing all possible source organisms returned by MARLOWE in the database search. Increasing the search space in a target-decoy-based database search algorithm will dilute the statistical power, thus resulting in more missed true targets compared to a database search with a smaller database. This will likely translate to detection of fewer human contaminant GVPs, as these GVPs represent a very minor component of the source organism sample. Therefore, we examine filtering criteria to restrict MARLOWE results of potential source organisms to create a reasonably-sized database for the database search that (1) maintains the flexibility of the untargeted nature of source organism characterization, because there is a possibility that the top-ranked hit from MARLOWE may not be the true source organism, but also (2) constrains the number of organisms that would be included in the database search to avoid low statistical power. As such, there needs to be criteria to downselect MARLOWE's lists to the most promising candidates.

Here, we apply a filtering threshold based on the results of an SVM model that was trained and optimized on PRIDE datasets for correct prediction of taxonomic groups.

### 3.2.2 SVM Organism Filtering

As demonstrated in Section 3.1.3, many of the SVM models we examined demonstrated similar performance. The one that we selected, Model #1 from Table 8 that utilizes solely taxonomic group rank as feature input, balanced performance with simplicity in model engineering, which we define as having the least number of necessary feature inputs that contribute to performance improvements.

This SVM model found that considering only the top-ranked taxonomic group was sufficient for correct source organism prediction. Thus, extrapolating this prediction to performing MARLOWE analysis of unknown samples, we expect that the top-ranked taxonomic group returned by MARLOWE will most likely contain the true source organism. Note that the top-ranked taxonomic group will likely contain more than one organism; MARLOWE implements unknown source organism characterization to taxonomic groups, which are organized by proteomic similarity of organisms. All proteomes from organisms in the top-ranked taxonomic group will be included in the source organism FASTA file. This file is then combined with the GVP FASTA file for the database search. The GVP FASTA file here will be the same GVP FASTA file used for the GVP detection-only workflow.

In general, when considering only the organisms within the top-ranked taxonomic group, we found that proteomes from 3 ± 5 (s.d.) organisms, on average, were included to represent the source organism portion of the FASTA file. We observed some variability in the number of included organisms' proteomes, especially when comparing PRIDE projects from different ground truth organisms. Figure 22 below displays the distribution of the number of organisms' proteomes included as the source organism portion of the FASTA file, grouped by the project's ground truth organism.

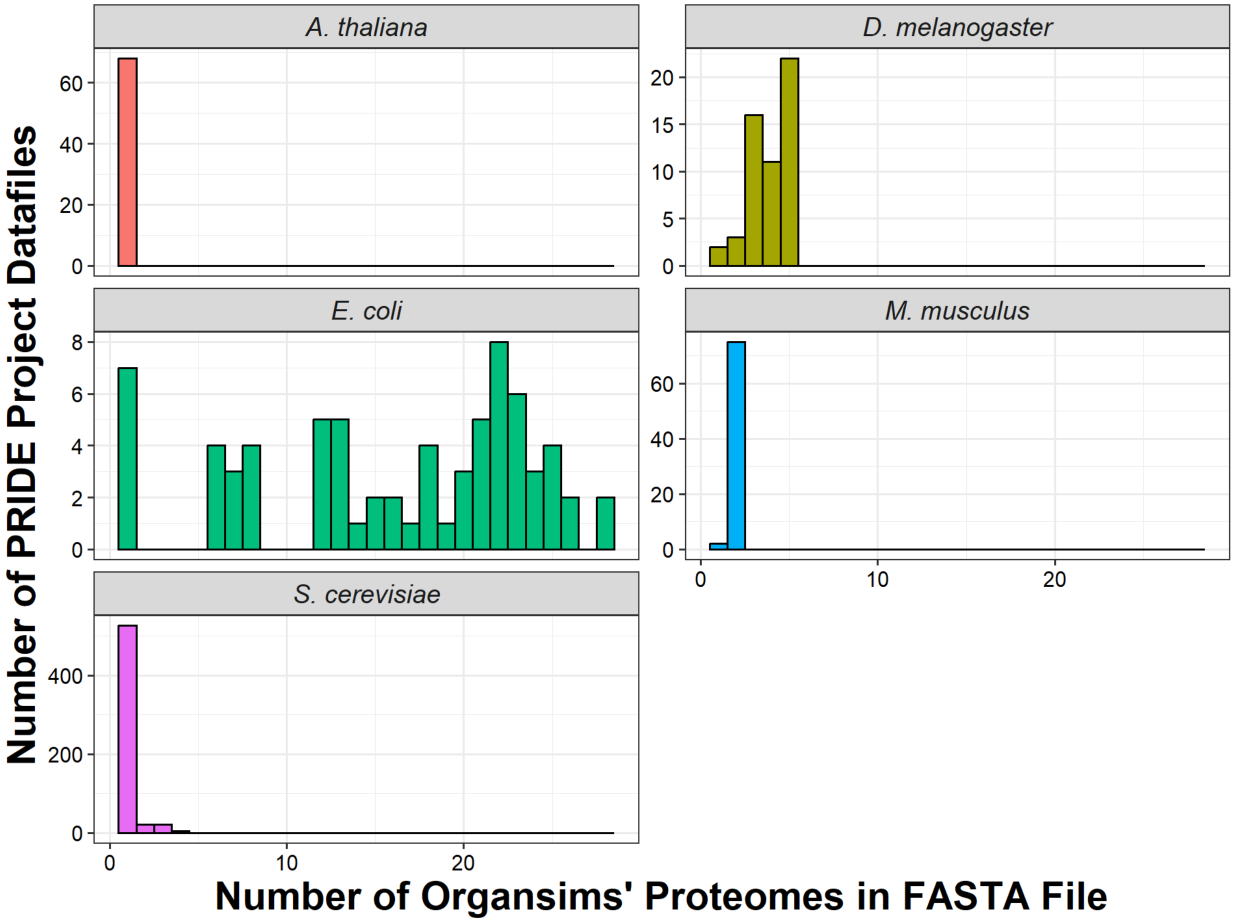

Figure 22. Histogram displaying the distribution of the number of organisms' proteomes included in the source organism FASTA file, as informed by SVM model filtering of MARLOWE's potential source organism lists, grouped by ground truth source organism, for downstream database search and GVP detection from select PRIDE projects. The vast majority of FASTA files created from this approach contain a single organism's proteome, though up to a maximum of 28 organisms' proteomes for a minority of FASTA files.

It is obvious that the vast majority of FASTA files informed by MARLOWE and SVM model filtering contain a single organism's proteomes (*A. thaliana* and *S. cerevisiae*), though up to 28 organisms' proteomes were included in FASTA files for *E. coli* projects (Figure 22). The large number of organisms' proteomes included in FASTA files for *E. coli* projects is likely owing to a high representation of *E. coli* strains in the KEGG database underlying MARLOWE and additional organisms (e.g., *Shigella flexneri*) that exhibits proteomic similarity to be contained within the same taxonomic group as *E. coli* (K-12 strain). The taxonomic group containing *E. coli* (K-12 strain) includes 79 organisms, but only those organisms with at least 2 peptide hits are included in the FASTA file, thus restricting to organisms that not only exhibit high proteomic similarity in general, but must demonstrate this in empirical measurements.

For projects belonging to other source organisms, the number of organisms' proteomes included in the FASTA file is much more modest (Figure 22). This distribution of organisms included in the database search aligns with our desire to create a reasonably-sized database that balances an untargeted approach to source organism characterization while minimizing the possibility of diluting statistical power in peptide detection.

These augmented MARLOWE-SVM informed FASTA files were then combined with the GVP FASTA file into a single complete FASTA file for database search and GVP detection. Note that of the 844 datafiles analyzed via MARLOWE, 62 returned MARLOWE results of potential source organisms where the top-ranked taxonomic group did not contain the ground truth source organism. In this case, the FASTA file for subsequent database search would not contain the true source organism. We expect that source organism peptide detections will be affected, however, human contaminant GVP detection may be less affected, as (1) the same GVP FASTA file was used, and (2) incorrect detections from having the incorrect organism proteome in the FASTA file for those datafiles may be diluted at the project level, as peptide detections are aggregated to the project level for FDR control. Effects of using such a FASTA file for database search and GVP detection will be examined in the next section.

### 3.2.3 GVP Detection Performance

GVP detection performance using the MARLOWE-GVP pipeline was examined. Of the 19 PRIDE DDA projects tested, 17 projects yielded detectable human contaminant GVPs at 5% peptide-level FDR control. When compared to the GVP detection-only pipeline (where only the ground truth organism's proteome was included in the database search) under the same FDR control conditions, we observed a systematic decrease in the number of detected GVPs with the MARLOWE-GVP pipeline (Figure 23).

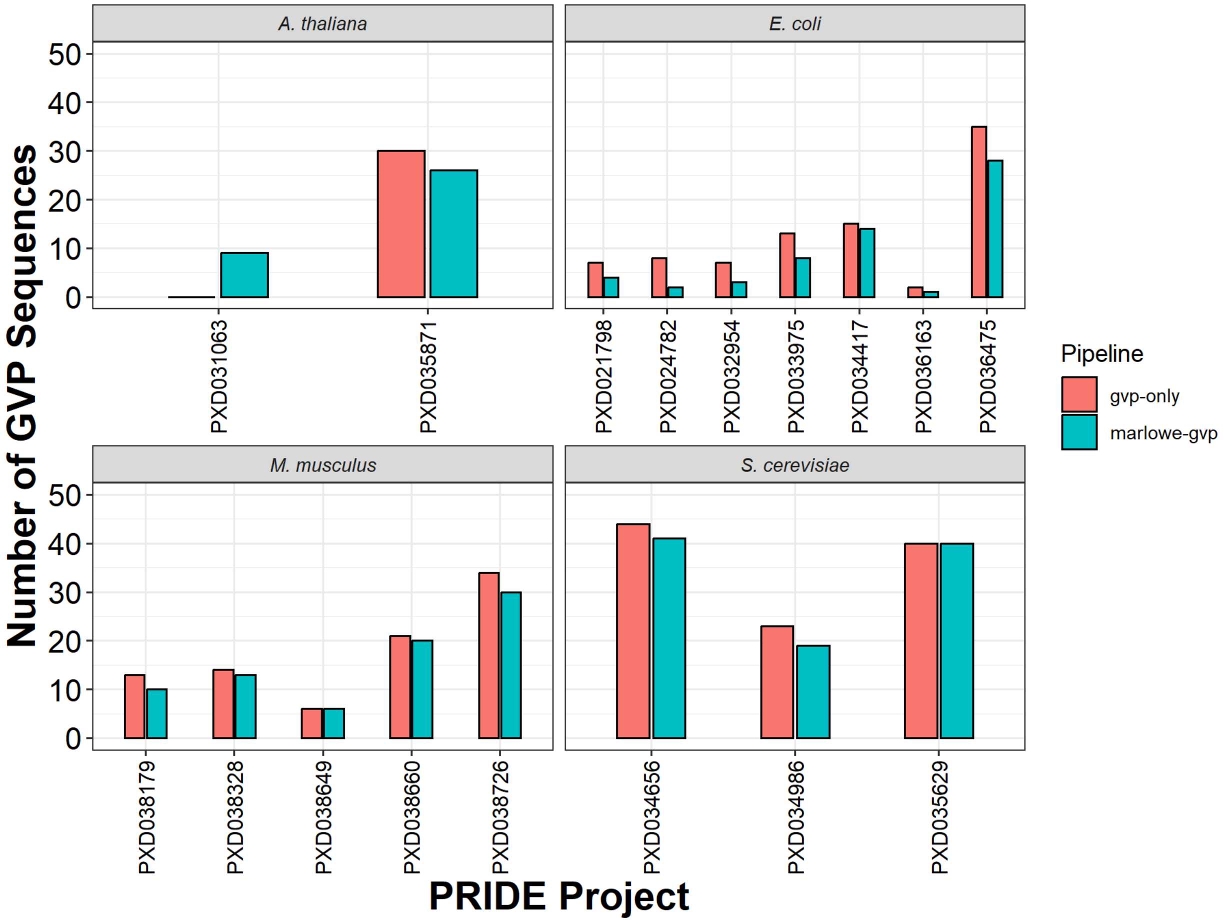

Figure 23. Comparison of the numbers of detected human contaminant GVPs per PRIDE project at 5% peptide-level FDR control, and grouped by ground truth organism, with the two

bioinformatics pipeline (gvp-only = GVP detection only; marlowe-gvp = via MARLOWE-GVP pipeline).

This systematic decrease in number of detected human contaminant GVPs in the MARLOWE-GVP workflow may be due to a slight difference in the database search step of the workflow between the two approaches. In the GVP detection-only workflow, modifications considered in the database search were matched to expected modifications as reported by each PRIDE project owner, either described in the project metadata or in the associated publications. However, in the MARLOWE-GVP workflow, the database search only considered cysteine carbamidomethylation as the sole post-translational modification (PTM).

To investigate the extent of and potential explanations for this systematic decrease in detected GVPs, we also examined and compared holistic peptide detection from both pipelines, which includes source organism peptide detection. On average, 20,903 ± 16,767 (s.d.) unique peptide sequences were detected using the GVP detection-only pipeline (N = 18 PRIDE projects), compared to 15,494 ± 11,546 (s.d.) unique peptide sequences detected with the MARLOWE-GVP pipeline (N = 18 projects). This decrease in peptide sequences was determined to be statistically significant via a paired t-test (p-value = 0.0136, df = 17). Given these consistent observations, we hypothesize that the difference in included PTMs for the database search between the GVP detection-only and MARLOWE-GVP pipelines may account for the systematic decrease in the number of GVPs as well as the total number of peptide sequences.

Future efforts will focus on performing the database search using a consistent set of PTMs, to ensure that downstream GVP detections can be fairly compared between the two pipelines, to examine the accuracy of the MARLOWE-GVP pipeline for GVP detection. We expect that use of a consistent set of database parameters will yield greater similarity in the total number of detected peptides and human contaminant GVPs between the two pipelines.

Despite the difference in database search parameters between the two pipelines, we examined the similarity in the set of detected human contaminant GVPs between the two pipelines: GVP detection-only and MARLOWE-GVP. To perform this comparison, we calculated the Jaccard index (i.e., taking the intersection of detected GVPs over the union of the two sets) between the two sets of GVPs per PRIDE project as a measure of GVP profile similarity. Figure 24 below displays the GVP profile similarity for each project, grouped by ground truth source organism.

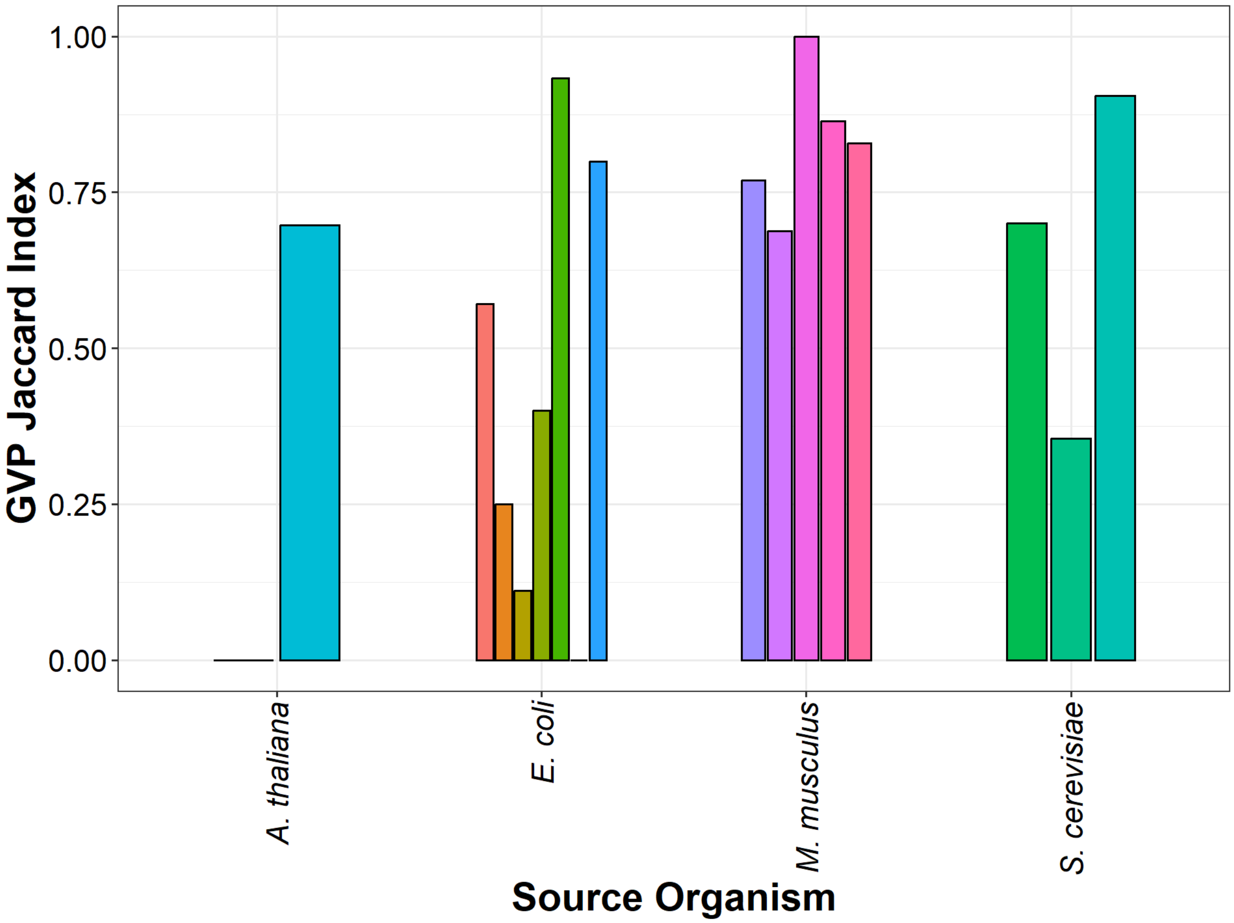

Figure 24. Barplot displaying similarity of human contaminant GVP profiles for each PRIDE project and grouped by ground truth source organism, between the GVP detection-only and MARLOWE-GVP pipelines.

Interestingly, despite the database search difference described above, we find that detected human contaminant GVP profiles using the two different pipelines share a high degree of similarity, which is fortuitous and also provides us with confidence in the GVP detection results from the MARLOWE-GVP pipeline. On average, we observe GVP profile similarities of 0.58 ± 0.33 (s.d.), on a 0 – 1 scale, where 1 is identity and 0 is completely dissimilar. Clearly, there is a core set of GVPs without post-translational modifications that are present in proteomics data and can be detected, and those are not affected by the difference in database search PTM parameters between the two pipelines. However, we expect to see a higher degree of GVP profile similarity between the two pipelines with a consistent set of PTMs applied to the database search parameters in future efforts.

While creating MARLOWE-SVM model informed FASTA files to capture the unknown source organism's proteome in an untargeted manner, we observed that of the 844 datafiles filtered via the SVM model, the top-ranked taxonomic group for 62 datafiles (from 7 PRIDE projects out of 19) did not contain the ground truth source organism. This represents, on average, 27 ± 36 % (s.d.) of each project. These projects also tended to be simple in sample complexity (e.g., *S. cerevisiae* project that utilized affinity purification during sample preparation). We investigate the effect of using the incorrect source organism's proteome during database search on human contaminant GVP detection and resulting GVP profiles, as compared to the profiles produced using the GVP detection-only pipeline.

Of these 7 projects, only 5 contained detectable human contaminant GVPs (Table 12). Jaccard indices for these projects are quite high, above the average Jaccard index across all projects, allowing us to infer that human contaminant GVP detection may not be affected by having a portion of database searches with the incorrect source organism.

Table 12. GVP profile similarity between GVP detection-only and MARLOWE-GVP pipelines for PRIDE projects containing datafiles searched with the incorrect source organism

| Ground truth source organism | PRIDE project | Number of GVPs (GVP detection-only) | Number of GVPs (MARLOWE-GVP) | GVP profile Jaccard index |
|---|---|---|---|---|
| *E. coli* | PXD021798 | 7 | 4 | 0.571 |
| *E. coli* | PXD034417 | 15 | 14 | 0.933 |
| *S. cerevisiae* | PXD034656 | 35 | 28 | 0.700 |
| *E. coli* | PXD036475 | 13 | 10 | 0.800 |
| *M. musculus* | PXD038179 | 44 | 41 | 0.769 |

Finally, to capture the core set of GVPs detected in both pipelines, we determined the top 10 most frequently detected human contaminant GVPs using the MARLOWE-GVP pipeline that are also detected using the GVP detection-only pipeline (Table 13). Most of these GVPs (9 GVPs) are reported as top detected GVPs in the proteomics data from PRIDE projects (Table 1) and 8 of these GVPs are reported as top detected GVPs in proteomics data from in-house repository campaigns, prepared by sample preparer 37 (Table 5).

Table 13. List of top 10 most frequently detected GVPs in PRIDE projects using the MARLOWE-GVP pipeline that are also detected from the GVP detection-only pipeline.

| GVP sequence | Detection frequency | Gene name | Chromosome |
|---|---|---|---|
| LAADDFR | 13 | KRT13 | 17 |
| AQYEEIAQR | 12 | KRT76 | 12 |
| FASFIDK | 11 | KRT75 | 12 |
| VTMQNLNDR | 11 | KRT14 | 17 |
| DYQELMNVK | 10 | KRT76 | 12 |

| | | | |
|---|---|---|---|
| FLEQQNQVLETK | 8 | KRT74 | 12 |
| LEQEIATYR | 8 | KRT14 | 17 |
| SLYGLGGSK | 8 | KRT6C | 12 |
| FLEQQNK | 7 | KRT6B | 12 |
| AEAEALYQTK | 6 | KRT78 | 12 |

We observe a high degree of similarity in GVP detection between both pipelines, despite the PTM parameter difference in database search. GVP detection also does not appear to be affected by using the incorrect source organism's proteome during database search. These results provide us with confidence into the accuracy of GVP detection using MARLOWE-GVP as a fully untargeted approach to determining the source organism for database search. We expect that addressing the PTM issue in future efforts for a true comparison of the pipelines will only produce an even greater level of similarity in GVP profiles, thus allowing us to be even more confident in the MARLOWE-GVP workflow as an alternative to the conventional database search approach.

## 3.3 Conclusions

We successfully demonstrate an end-to-end pipeline, MARLOWE-GVP, that combines two capabilities: untargeted unknown source organism characterization and human contaminant GVP detection. We further show the broad applicability of this approach to proteomics data from a diverse set of non-human organisms. This combined capability enables a more complete characterization of an unknown proteomics sample and advances our understanding of the "dark" proteome.

# 4.0 Concluding Remarks & Future Outlook

Through our two lines of effort, human contaminant GVP detection and application of MARLOWE-GVP pipeline, we advance characterization and understanding of the “dark” proteome, towards a more complete proteomic characterization of non-human samples of potentially unknown origin. The development and assessment of these two capabilities have provided us with a better understanding of untargeted and minor protein analysis, though limitations still exist—primarily detection variability likely owing to incomplete peptide detection. To continue to push the boundaries and further our elucidation of the “dark” proteome, future efforts should examine alternative and more complete peptide detection strategies, such as data-independent acquisition mass spectrometry, and investigate other potential trace components of unknown proteomics samples. Further characterization efforts of this “dark” proteome can find broad applications, including in forensic science, metaproteomics, and evolutionary biology.

# 5.0 References

(1) Chan, Q. W. T.; Rogalski, J.; Moon, K.-M.; Foster, L. J. The application of forensic proteomics to identify an unknown snake venom in a deceased toddler. *Forensic Science International* **2021**, *323*, 110820. DOI: https://doi.org/10.1016/j.forsciint.2021.110820.

(2) Kalb, S. R.; Goodnough, M. C.; Malizio, C. J.; Pirkle, J. L.; Barr, J. R. Detection of botulinum neurotoxin A in a spiked milk sample with subtype identification through toxin proteomics. *Analytical chemistry* **2005**, *77* (19), 6140-6146.

(3) Kalb, S. R.; Barr, J. R. Mass spectrometric identification and differentiation of botulinum neurotoxins through toxin proteomics. *Reviews in analytical chemistry* **2013**, *32* (3), 189-196.

(4) Gilquin, B.; Jaquinod, M.; Louwagie, M.; Kieffer-Jaquinod, S.; Kraut, A.; Ferro, M.; Becher, F.; Brun, V. A proteomics assay to detect eight CBRN-relevant toxins in food. *PROTEOMICS* **2017**, *17* (1-2), 1600357. DOI: https://doi.org/10.1002/pmic.201600357 (acccessed 2024/09/16).

(5) Dupré, M.; Gilquin, B.; Fenaille, F.; Feraudet-Tarisse, C.; Dano, J.; Ferro, M.; Simon, S.; Junot, C.; Brun, V.; Becher, F. Multiplex Quantification of Protein Toxins in Human Biofluids and Food Matrices Using Immunoextraction and High-Resolution Targeted Mass Spectrometry. *Analytical Chemistry* **2015**, *87* (16), 8473-8480. DOI: 10.1021/acs.analchem.5b01900.

(6) Merkley, E. D.; Jenson, S. C.; Arce, J. S.; Melville, A. M.; Leiser, O. P.; Wunschel, D. S.; Wahl, K. L. Ricin-like proteins from the castor plant do not influence liquid chromatography-mass spectrometry detection of ricin in forensically relevant samples. *Toxicon* **2017**, *140*, 18-31. DOI: https://doi.org/10.1016/j.toxicon.2017.10.004.

(7) Legg, K. M.; Powell, R.; Reisdorph, N.; Reisdorph, R.; Danielson, P. B. Verification of protein biomarker specificity for the identification of biological stains by quadrupole time-of-flight mass spectrometry. *ELECTROPHORESIS* **2017**, *38* (6), 833-845. DOI: https://doi.org/10.1002/elps.201600352 (acccessed 2024/09/16).

(8) Legg, K. M.; Powell, R.; Reisdorph, N.; Reisdorph, R.; Danielson, P. B. Discovery of highly specific protein markers for the identification of biological stains. *ELECTROPHORESIS* **2014**, *35* (21-22), 3069-3078. DOI: https://doi.org/10.1002/elps.201400125 (acccessed 2024/09/16).

(9) Yang, H.; Zhou, B.; Deng, H.; Prinz, M.; Siegel, D. Body fluid identification by mass spectrometry. *International Journal of Legal Medicine* **2013**, *127* (6), 1065-1077. DOI: 10.1007/s00414-013-0848-1.

(10) Chu, F.; Mason, K. E.; Anex, D. S.; Jones, A. D.; Hart, B. R. Hair Proteome Variation at Different Body Locations on Genetically Variant Peptide Detection for Protein-Based Human Identification. *Scientific Reports* **2019**, *9* (1), 7641. DOI: 10.1038/s41598-019-44007-7.

(11) Milan, J. A.; Wu, P.-W.; Salemi, M. R.; Durbin-Johnson, B. P.; Rocke, D. M.; Phinney, B. S.; Rice, R. H.; Parker, G. J. Comparison of protein expression levels and proteomically-inferred genotypes using human hair from different body sites. *Forensic Science International: Genetics* **2019**, *41*, 19-23. DOI: https://doi.org/10.1016/j.fsigen.2019.03.009.

(12) Parker, G. J.; Leppert, T.; Anex, D. S.; Hilmer, J. K.; Matsunami, N.; Baird, L.; Stevens, J.; Parsawar, K.; Durbin-Johnson, B. P.; Rocke, D. M.; et al. Demonstration of Protein-Based

Human Identification Using the Hair Shaft Proteome. *PLOS ONE* **2016**, *11* (9), e0160653. DOI: 10.1371/journal.pone.0160653.

(13) Mason, K. E.; Anex, D.; Grey, T.; Hart, B.; Parker, G. Protein-based forensic identification using genetically variant peptides in human bone. *Forensic Science International* **2018**, *288*, 89-96. DOI: https://doi.org/10.1016/j.forsciint.2018.04.016.

(14) Jarman, K. H.; Heller, N. C.; Jenson, S. C.; Hutchison, J. R.; Kaiser, B. L. D.; Payne, S. H.; Wunschel, D. S.; Merkley, E. D. Proteomics Goes to Court: A Statistical Foundation for Forensic Toxin/Organism Identification Using Bottom-Up Proteomics. *Journal of Proteome Research* **2018**, *17* (9), 3075-3085. DOI: 10.1021/acs.jproteome.8b00212.

(15) Boulund, F.; Karlsson, R.; Gonzales-Siles, L.; Johnning, A.; Karami, N.; Al-Bayati, O.; Åhrén, C.; Moore, E. R. B.; Kristiansson, E. Typing and Characterization of Bacteria Using Bottom-up Tandem Mass Spectrometry Proteomics*. *Molecular & Cellular Proteomics* **2017**, *16* (6), 1052-1063. DOI: https://doi.org/10.1074/mcp.M116.061721.

(16) Alves, G.; Wang, G.; Ogurtsov, A. Y.; Drake, S. K.; Gucek, M.; Sacks, D. B.; Yu, Y.-K. Rapid Classification and Identification of Multiple Microorganisms with Accurate Statistical Significance via High-Resolution Tandem Mass Spectrometry. *Journal of the American Society for Mass Spectrometry* **2018**, *29* (8), 1721-1737. DOI: 10.1007/s13361-018-1986-y.

(17) Mason, K. E.; Paul, P. H.; Chu, F.; Anex, D. S.; Hart, B. R. Development of a Protein-based Human Identification Capability from a Single Hair. *Journal of Forensic Sciences* **2019**, *64* (4), 1152-1159, https://doi.org/10.1111/1556-4029.13995. DOI: https://doi.org/10.1111/1556-4029.13995 (acccessed 2022/04/12).

(18) Chu, F.; Lin, A. Detecting genetically variant peptides in non-human samples. *bioRxiv* **2024**, 2024.2008.2022.609263. DOI: 10.1101/2024.08.22.609263.

(19) Pfrunder, S.; Grossmann, J.; Hunziker, P.; Brunisholz, R.; Gekenidis, M.-T.; Drissner, D. Bacillus cereus Group-Type Strain-Specific Diagnostic Peptides. *Journal of Proteome Research* **2016**, *15* (9), 3098-3107. DOI: 10.1021/acs.jproteome.6b00216.

(20) Alves, G.; Wang, G.; Ogurtsov, A. Y.; Drake, S. K.; Gucek, M.; Suffredini, A. F.; Sacks, D. B.; Yu, Y.-K. Identification of microorganisms by high resolution tandem mass spectrometry with accurate statistical significance. *Journal of The American Society for Mass Spectrometry* **2015**, *27* (2), 194-210.

(21) Sahl, J. W.; Vazquez, A. J.; Hall, C. M.; Busch, J. D.; Tuanyok, A.; Mayo, M.; Schupp, J. M.; Lummis, M.; Pearson, T.; Shippy, K. The effects of signal erosion and core genome reduction on the identification of diagnostic markers. *MBio* **2016**, *7* (5).

(22) Mesuere, B.; Debyser, G.; Aerts, M.; Devreese, B.; Vandamme, P.; Dawyndt, P. The Unipept metaproteomics analysis pipeline. *PROTEOMICS* **2015**, *15* (8), 1437-1442. DOI: https://doi.org/10.1002/pmic.201400361 (acccessed 2024/09/16).

(23) Potgieter, M. G.; Nel, A. J. M.; Fortuin, S.; Garnett, S.; Wendoh, J. M.; Tabb, D. L.; Mulder, N. J.; Blackburn, J. M. MetaNovo: An open-source pipeline for probabilistic peptide discovery in complex metaproteomic datasets. *PLOS Computational Biology* **2023**, *19* (6), e1011163. DOI: 10.1371/journal.pcbi.1011163.

(24) Frankenfield, A. M.; Ni, J.; Ahmed, M.; Hao, L. Protein Contaminants Matter: Building Universal Protein Contaminant Libraries for DDA and DIA Proteomics. *Journal of Proteome Research* **2022**, *21* (9), 2104-2113. DOI: 10.1021/acs.jproteome.2c00145.

(25) Dobbs, J. M.; Jenkins, M. L.; Burke, J. E. Escherichia coli and Sf9 Contaminant Databases to Increase Efficiency of Tandem Mass Spectrometry Peptide Identification in Structural Mass Spectrometry Experiments. *Journal of the American Society for Mass Spectrometry* **2020**, *31* (10), 2202-2209. DOI: 10.1021/jasms.0c00283.

(26) Mellacheruvu, D.; Wright, Z.; Couzens, A. L.; Lambert, J.-P.; St-Denis, N. A.; Li, T.; Miteva, Y. V.; Hauri, S.; Sardiu, M. E.; Low, T. Y.; et al. The CRAPome: a contaminant repository for affinity purification–mass spectrometry data. *Nature Methods* **2013**, *10* (8), 730-736. DOI: 10.1038/nmeth.2557.

(27) The UniProt, C. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research* **2023**, *51* (D1), D523-D531. DOI: 10.1093/nar/gkac1052 (acccessed 9/16/2024).

(28) Martin, F. J.; Amode, M. R.; Aneja, A.; Austine-Orimoloye, O.; Azov, Andrey G.; Barnes, I.; Becker, A.; Bennett, R.; Berry, A.; Bhai, J.; et al. Ensembl 2023. *Nucleic Acids Research* **2023**, *51* (D1), D933-D941. DOI: 10.1093/nar/gkac958 (acccessed 9/16/2024).

(29) Chen, S.; Francioli, L. C.; Goodrich, J. K.; Collins, R. L.; Kanai, M.; Wang, Q.; Alföldi, J.; Watts, N. A.; Vittal, C.; Gauthier, L. D.; et al. A genomic mutational constraint map using variation in 76,156 human genomes. *Nature* **2024**, *625* (7993), 92-100. DOI: 10.1038/s41586-023-06045-0.

(30) Park, C. Y.; Klammer, A. A.; Käll, L.; MacCoss, M. J.; Noble, W. S. Rapid and Accurate Peptide Identification from Tandem Mass Spectra. *Journal of Proteome Research* **2008**, *7* (7), 3022-3027. DOI: 10.1021/pr800127y.

(31) McIlwain, S.; Tamura, K.; Kertesz-Farkas, A.; Grant, C. E.; Diament, B.; Frewen, B.; Howbert, J. J.; Hoopmann, M. R.; Käll, L.; Eng, J. K.; et al. Crux: Rapid Open Source Protein Tandem Mass Spectrometry Analysis. *Journal of Proteome Research* **2014**, *13* (10), 4488-4491. DOI: 10.1021/pr500741y.

(32) Käll, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nature Methods* **2007**, *4* (11), 923-925. DOI: 10.1038/nmeth1113.

(33) Lin, A.; Short, T.; Noble, W. S.; Keich, U. Improving Peptide-Level Mass Spectrometry Analysis via Double Competition. *Journal of Proteome Research* **2022**, *21* (10), 2412-2420. DOI: 10.1021/acs.jproteome.2c00282.

(34) Chambers, M. C.; Maclean, B.; Burke, R.; Amodei, D.; Ruderman, D. L.; Neumann, S.; Gatto, L.; Fischer, B.; Pratt, B.; Egertson, J.; et al. A cross-platform toolkit for mass spectrometry and proteomics. *Nature Biotechnology* **2012**, *30* (10), 918-920. DOI: 10.1038/nbt.2377.

(35) Hulstaert, N.; Shofstahl, J.; Sachsenberg, T.; Walzer, M.; Barsnes, H.; Martens, L.; Perez-Riverol, Y. ThermoRawFileParser: Modular, Scalable, and Cross-Platform RAW File

Conversion. *Journal of Proteome Research* **2020**, *19* (1), 537-542. DOI: 10.1021/acs.jproteome.9b00328.

(36) Perez-Riverol, Y.; Bai, J.; Bandla, C.; García-Seisdedos, D.; Hewapathirana, S.; Kamatchinathan, S.; Kundu, Deepti J.; Prakash, A.; Frericks-Zipper, A.; Eisenacher, M.; et al. The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Research* **2022**, *50* (D1), D543-D552. DOI: 10.1093/nar/gkab1038 (acccessed 9/16/2024).

(37) Xin, L.; Qiao, R.; Chen, X.; Tran, H.; Pan, S.; Rabinoviz, S.; Bian, H.; He, X.; Morse, B.; Shan, B.; et al. A streamlined platform for analyzing tera-scale DDA and DIA mass spectrometry data enables highly sensitive immunopeptidomics. *Nature Communications* **2022**, *13* (1), 3108. DOI: 10.1038/s41467-022-30867-7.

(38) Bhavsar, H.; Panchal, M. H. A review on support vector machine for data classification. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)* **2012**, *1* (10), 185-189.

**Pacific Northwest National Laboratory**

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99354

1-888-375-PNNL (7665)

***www.pnnl.gov***