Proteomics in non-model organisms: a new analytical frontier

Michelle Heck[1,2,3] and Benjamin A. Neely[4*]

[1] Emerging Pests and Pathogens Research Unit, USDA Agricultural Research Service, Ithaca, NY, USA
[2] Plant Pathology and Plant Microbe Biology Section, School of Integrative Plant Science, Cornell University, Ithaca, NY, USA
[3] Boyce Thompson Institute, Ithaca, NY, USA
[4] Chemical Sciences Division, National Institute of Standards and Technology, Charleston, SC, USA

*** Corresponding author:** BN; benjamin.neely@nist.gov

Abstract

For the last century we have relied on model organisms to help understand fundamental biological processes. Now, with advancements in genome sequencing, assembly and annotation, non-model organisms may be studied with the same advanced bioanalytical toolkit as model organisms. Proteomics is one such technique, which classically relies on predicted protein sequences to catalog and measure complex proteomes across tissues and biofluids. Applying proteomics to non-model organisms can advance and accelerate biomimicry studies, biomedical advancements, veterinary medicine, agricultural research, behavioral ecology, and food safety. In this post-model organism era we can study almost any species, meaning that many non-model organisms are in fact important emerging model organisms. Herein we focus specifically on eukaryotic organisms and discuss steps to generating sequence databases, analyzing proteomic data with or without a database, interpreting results, and future research opportunities. Proteomics is more accessible than ever before, and will continue to rapidly advance in the coming years, enabling critical research and discoveries in non-model organisms that were hitherto impossible.

Keywords: comparative biology, biomimicry, proteomics, genomics, non-model

Introduction

Beginning with fruit flies at the turn of the 20th century, model organisms have enabled biological discoveries by bringing testable yet complex biological systems to the lab. These organisms, including mice, rats, frogs, zebra fish, roundworms, yeast, maize, *Arabidopsis*, and *Escherichia coli*, are easily maintained in the lab, have well established methods and molecular tools, can be manipulated at the molecular level, and have a vast body of associated literature[1-2]. Over the last 100 years, these model organisms have been crucial to revealing fundamental biological truths and developing the theoretical framework for molecular biology. But as cutting-edge biomolecular analysis in non-model organisms is becoming more routine, we are on the cusp of a new frontier [3-4]. Over billions of years, natural selection has favored countless adaptations to allow organisms to inhabit and thrive in all corners of the earth. Through the study of non-model organisms, scientists are revealing insights into the molecular basis of a vast array of human conditions. Naturally occurring adaptations and susceptibilities inform the evolutionary and genetic underpinning of chronic and genetic disease in humans, from cancer to neurodegeneration (expertly reviewed by Stenvinkel *et al.*[5-6]). For instance, elephants and naked mole rats illuminate molecular mechanisms at play in cancer resistance[7-9], deer antlers hold clues to organ regeneration[10-11], diving mammals inform ischemia reperfusion injury resistance[12-16], hibernating arctic squirrels and grizzly bears teach lessons about neuroprotection and metabolism[17-19], and naked mole rats and bats possess secrets to longevity[20-23]. Plants are also treasure troves of information including how compartmentalization in plants avoids senescence[24-25]. This field of study, biomimetics, seeks to understand these adaptations and provide insight into chronic disease in humans, and is made possible by non-model organism research.

In addition to biomimetics, non-model organisms are important in other key areas, including agricultural research, veterinary medicine, behavioral ecology, and food safety. Emerging plant diseases in high value agricultural crops, such as citrus greening disease[26-28]

and cotton blue disease[29-33], are forcing rapid development of bioanalytical capabilities to study non-model plant species, plant pathogens and insect vectors of plant pathogens. Knowledge of these virus-interacting host proteins paves the way for genome editing approaches to develop durable resistance. Non-model organism research is also critical in veterinary medicine for industries such as zoos and aquariums or pets ($3 billion and $18 billion annual, respectively[34-35]), which rely on bioanalytical measurements for development and validation of new techniques translated from human medicine. In cases such as bats, genomic and proteomic analysis is being harnessed to better understand longevity, as well as their innate immunity and ability to serve as disease reservoirs[36-38]. Maybe less apparent is the applicability of proteomics to behavioral ecology (expertly reviewed by Valcu and Kempenaers[39]), such as eusociality in the naked mole-rat[40]. Lastly, with growing concerns of food authenticity, proteomic analysis in non-model organisms offers promising avenues of tracking and validation in aquaculture and agriculture[41-45]. These topics are economically critical and rely on our ability to rapidly develop cutting edge analytical capabilities in non-model organisms. Classical model organisms are vitally important, but there is an enormous promise of untold discoveries and applications by studying non-model organisms (Figure 1).

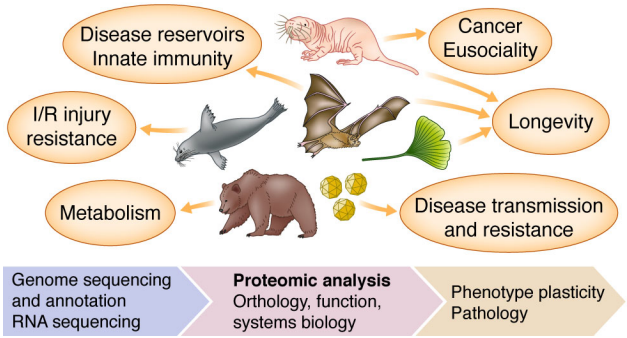


Figure 1. General overview of proteomics in non-model organisms. Some notable examples of non-model proteomics are shown (clockwise from top): naked mole-rat, bat, ginkgo, viruses, hibernating mammals, and diving mammals. I/R refers to ischemia reperfusion.

Given the current SARS-CoV-2 pandemic, it is important to specifically highlight the importance and potential of non-model organism research when studying zoonotic viruses. Although predicting zoonosis includes many non-molecular factors[46-47], it has been shown that sequence homology between host receptors (*e.g.*, ACE2 in SARS and SARS-CoV-2 infection) can predict zoonotic potential and species tropism[48-50]. For instance, SARS-CoV-2 entry potential correlates with ACE2 receptor homology between reservoirs and hosts[50]. Current studies of SARS-CoV-2 utilizing structural information such as receptor binding domain[51] and glycosylation of the viral spike protein[52] will continue to improve predictive power even further[53-54]. Undoubtedly proteomic analysis of these non-model reservoirs can provide empirical evidence of structural and glycosylation predictions on vital molecular targets, which is made possible only by directly studying the non-model organism. Furthermore, determining host-virus interactions by protein-protein interaction (PPI) proteomic analysis is critical to understanding infection, such as with Ebola virus[55], Zika virus[56] and SARS-CoV-2[57], but similar PPI studies in viral reservoirs themselves could be equally enlightening for innate and adaptive immunity as well as zoonotic spillover potential. Broad proteomic studies across hundreds of reservoir and host species will improve our understanding of viral infection, zoonotic prediction, and highlight possible treatments.

Proteomics in non-model systems is driven by genomics

Currently, the most common method of proteomic analysis is shotgun proteomics (also referred to as bottom-up proteomics). By digesting proteins to peptides, proteomics researchers use mass spectrometers and database search algorithms to identify and quantify near complete proteomes (such as recent studies in human[58-59] and yeast[60]). Traditional data-dependent acquisition (DDA) or newer techniques such as data-independent acquisition (DIA) or BoxCar[61] may be used if the search space can be defined using protein sequences and appropriate post-translational modifications. When using shotgun proteomics to study a non-model organism that

may not have an annotated genome available, the search space is the major hurdle[62-63]. If an annotated genome is available, any limitation that applies to proteomics in model organisms will also apply to a non-model (*e.g.*, studying glycosylation with DIA data is difficult regardless). Knowledge of current solutions and best practices in non-model organisms is limiting widespread adoption. For starters, in the past, genome sequencing and annotation could prove an insurmountable or very expensive problem to overcome, but this is no longer the case.

In recent years, post-2015, there has been an explosion of genome sequencing as the cost to generate accurate highly-contiguous genome assemblies has continued to drop[64]. There are ongoing large sequencing efforts (*e.g.*, Earth BioGenome Project, Vertebrate Genomes Project, DNA Zoo, Zoonomia Project, Bat1K Project, Bird 10 000 Genomes Project) that are beginning to hit their stride and soon will be releasing hundreds of genomes per year to achieve many thousands of genomes by completion. Of the roughly 5400 mammal species, currently 430 have genomes, while of the roughly 400 000 plant species, there are 630 plant genomes, and of the nearly 1 million named species of insects less than 500 genomes are complete[65] (Figure 2). When a genome is not available, especially in the case of mammals the most affordable path to *de novo* sequencing is to use short-read sequencing with proximity ligation techniques (such as Hi-C or Chicago). The resulting genome assemblies are accurate for transcriptomic and proteomic studies (*e.g.*, proteomic analysis using a Chicago-based mammalian assembly[66]). In cases of highly heterogeneous and repetitive genomes (*e.g.*, Mollusca), additional and more expensive techniques may be required such as long-read sequencing based on Oxford Nanopore Technologies or PacBio, as well as optical mapping. But in the case of many species, a *de novo* assembly can be generated in the matter of months for minimal cost and expertise. Once completed, quality measures including contiguity (*i.e.*, scaffold and contig N50) and completeness (*i.e.*, BUSCO[67]) can be evaluated, but since there are no universal quality thresholds, we recommend comparing quality metrics of phylogenetically similar published assemblies. Alternatively, proteogenomics represents

experimental validation of a genome assembly using proteomics[66, 68-69]. Peptide identification can help validate polymorphisms and predicted alleles while top-down proteomic methods can confirm predicted isoforms. Future development of genome annotation pipelines may include ways to integrate mass spectrometry-based proteomic data.

all	P
eukaryote	1308
bacteria	876
virus	106
archaea	54

eukaryotes	P	G	A
animals	601	1958	487
green plants	343	630	108
fungi	209	3210	0
apicomplexans	40	143	0
kinetoplastids	29	82	0
diatoms	11	13	0
ciliates	8	32	0
oomycetes	8	89	0
haptophytes	5	4	0
diplomonads	5	8	0
red algae	5	13	0
trichomonads	4	4	0
dinoflagellates	4	12	0
cellular slime molds	3	20	0
brown algae	3	5	0
forams	2	2	0
euglenoids	1	1	0
cryptomonads	1	2	0
other eukaryotes	25	490	0

animals	P	G	A
chordates	298	1060	323
arthropods	165	593	140
molluscs	55	44	8
nematodes	22	134	0
cnidarians	20	35	10
flatworms	17	39	0
echinoderms	11	17	2
segmented worms	7	4	0
sponges	2	2	1
placozoans	1	2	0
rotifers	1	10	0
brachiopods	1	1	1
priapulids	0	1	1
hemichordates	0	2	1

green plants	P	G	A
flowering plants	303	517	106
green algae	15	87	0
ferns	2	1	0
club-mosses	1	3	0
other green plants	18	16	0

chordates	P	G	A
mammals	151	430	138
bony fish	55	280	80
lizards	53	33	10
birds	19	240	76
amphibians	8	17	5
carilaginous fish	3	9	2
turtles	2	24	5
alligator	2	2	2
dinosaurs	2	0	0
lungfish	1	1	0

mammals	P	G	A
even-toed ungulates	32	120	26
rodents	28	93	25
carnivores	25	60	23
primates	20	63	28
odd-toed ungulates	11	12	4
rabbits & hares	5	3	2
anteaters and sloths	5	5	0
bats	4	37	12
elephants	4	1	1
herbivorous marsupials	3	3	2
elephant shrews	2	1	1
aardvarks	1	2	1
manatees and dugongs	1	1	1
carnivorous marsupials	1	2	1
possums	1	1	1
armadillos	1	3	1
egg-laying mammals	0	1	1
insectivores	0	8	3
golden moles	0	1	1
tenrecs	0	2	1
flying lemurs	0	1	1
tree shrews	0	3	1
pangolins	0	3	1

Figure 2. Distribution of species with proteomic data sets (P), genome assemblies (G), and genome annotations (A) across phylogenetic clades. This series of charts was created by cross-referencing taxon IDs of genome assemblies on Genbank and genome annotations on RefSeq (valid as of 12 March 2020) with published proteomic data sets on the Proteomics Identifications Database (PRIDE; as of 17 January 2020) in order to emphasize clades lacking proteomic analysis, or clades that should be focused on for genomic sequencing and annotation. Specific to Eukaryotes, there are approximately 1308 species with published proteomic results, 6718 with available genomes and 595 species with complete genomic annotations. Note that these charts do not show the total number of species per clade (*e.g.*, 950 000 named insect species[65]), or the number of proteomic data sets per species (*e.g.*, 68 % of the 12 660 PRIDE data sets are human and mouse). For this cataloging, we focused only on genome and annotation resources at NCBI and acknowledge that for non-animal non-flowering plants, there are other more appropriate repositories such as the Joint Genome Institute. Terms in tables are accepted common or scientific NCBI designations, except in a few cases under

‘mammals’ (the full mammal table with taxon IDs is available in Supplemental Table S1), and that these numbers may not be exact due to self-reporting on PRIDE and matching taxon identifiers between sources (note that the two dinosaur proteomic data sets are from *Brachylophosaurus canadensis* and *Tyrannosaurus rex*).

Genome sequencing and availability does not seem to be the main roadblock to proteomics in non-model species: the current bottleneck is genome annotation. Complete and accurate genome annotation is key to proteomic workflows. Broadly there are two paths to annotation: in-house annotation or making the data publicly available for annotation by NCBI or Ensembl. In the case of in-house annotation, publicly available and often free, high-performance computing (HPC) resources such as XSEDE (Extreme Science and Engineering Discovery Environment; including JetStream), ACI-REF (Advanced Cyberinfrastructure Research and Education Facilitators) Network, NCI (National Computational Infrastructure) Australia, PRACE (Partnership for Advanced Computing in Europe) and the ELIXIR (the European life-sciences Infrastructure for biological Information) network, have enabled novice bioinformaticians without a local HPC to annotate genome assemblies using pipelines like MAKER (see tutorials[70-71]). Although still an iterative task, this is made possible by having thousands of processing hours on a virtual cluster. The second path to genome annotation is using free public resources like NCBI RefSeq or Ensembl. Specifically, the RefSeq project maintains and curates new and updated genome annotations[72-73]. RefSeq has published guidelines on how genomes are selected for annotation[74], related to assembly contiguity, secondary RNA-seq data, and need by researchers. Once a high-quality genome assembly has been made public on DDBJ (DNA Data Bank of Japan), ENA (European Nucleotide Archive) or GenBank, it may be annotated and publicly released by RefSeq in a matter of weeks. But of the hundreds of genomes on GenBank, only a fraction have been annotated (Figures 2 and 3). What may not be evident is that a

researcher not associated with the original genome project but with access to RNA-seq data may upload new sequence read archives (SRA), which can help begin or improve annotations of already existing high-quality genome assemblies. This is also an excellent way to demonstrate interest and improve annotation as more secondary gene evidence and transcript diversity typically results in more complete and accurate annotations. For instance, multiple tissue single-stranded RNA sequencing data were used by RefSeq to annotate the California sea lion genome[75]. The current annotation system is ripe for democratization and team science approaches, especially for pre-existing publicly available genome assemblies.

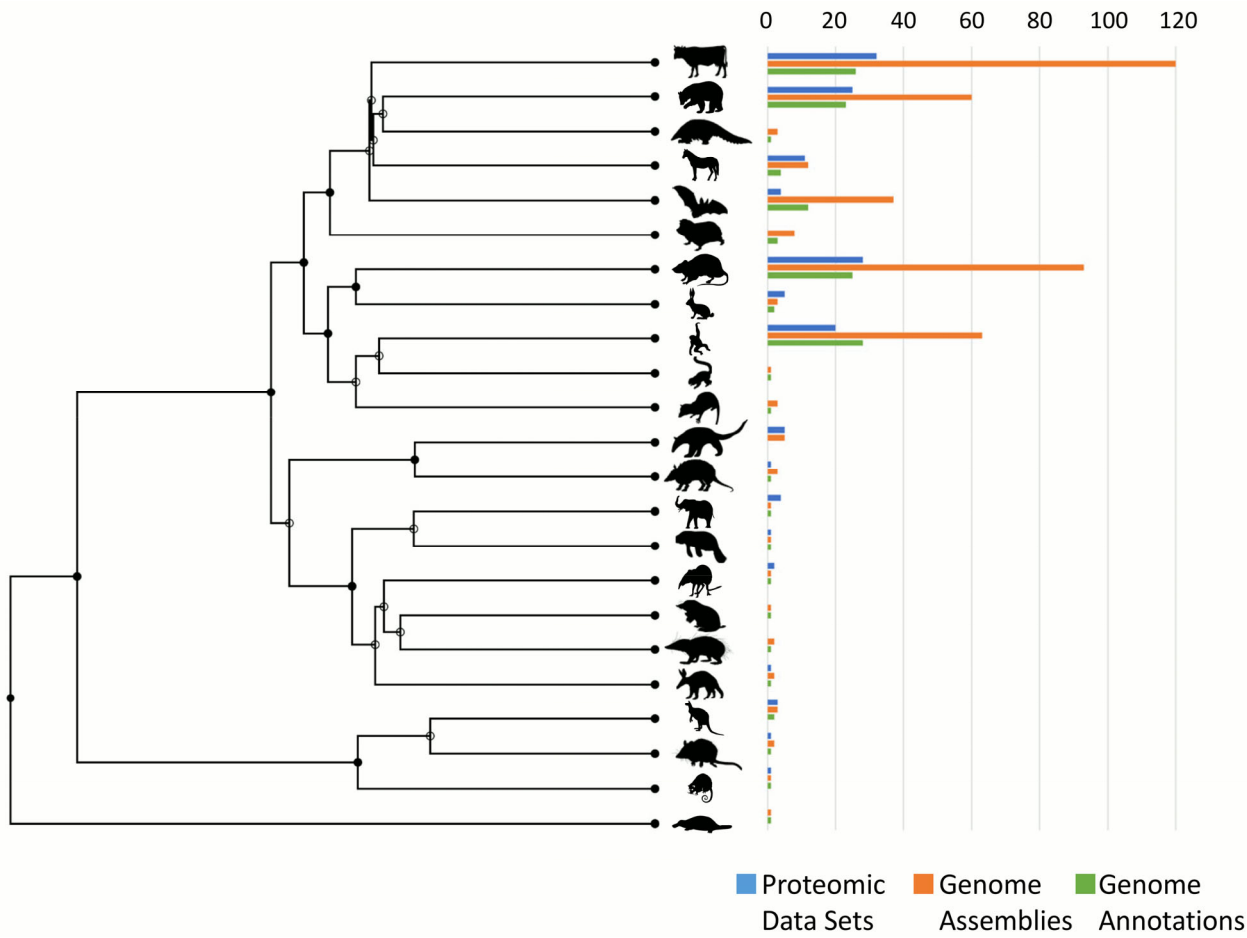


Figure 3. Distribution of mammals with proteomic data sets or genome annotations. Using a time tree (http://www.timetree.org), which incorporates fossil evidence[76], information

regarding the 23 groups of mammals with proteomic data sets on the Proteomics Identifications Database (PRIDE; as of 17 January 2020), genome assemblies on Genbank and genome annotations on RefSeq (valid as of 12 March 2020) was plotted. This is per taxa and does not include the number of proteomic data sets, genome assemblies or genome annotations within taxa. The order, from top to bottom is: even-toed ungulates, carnivores, pangolins, odd-toed ungulates, bats, insectivores, rodents, rabbits and hares, primates, flying lemurs, tree shrews, anteaters and sloths, armadillos, elephants, manatees and dugongs, elephant shrews, golden moles, tenrecs, aardvarks, herbivorous marsupials, carnivorous marsupials, possums, egg-laying mammals.

Following in-house or RefSeq genome annotation as described above, there will be a sequence database (a collection of protein sequences, typically as a FASTA, XML or GFF file) that can be used to search proteomic data. There are secondary resources such as UniProtKB that may import and evaluate annotations from RefSeq and Ensembl, but it is important to note that UniProtKB does not import all available high-quality annotations. Moreover, whereas RefSeq groups genome annotations by release or by user submission, the provenance of protein assignments from other resources may be difficult to determine. For these reasons, retrieving species-specific proteomes from UniProtKB, RefSeq and Ensembl will yield different protein databases, often based on different genome assemblies, annotation pipelines, and curation. Since interpretations of proteomic data in non-model organisms are only as useful as the functional annotation of the proteome, a careful manual inspection of automated annotations may be required before proceeding with biological and functional experiments, as even RefSeq annotations may not be perfect due to issues with homology-based assignments, likely because orthologs in well-annotated species may not always exist[77]. When proteins of interest are unknown or hypothetical, manual single protein curation is required using tools like UniRule[78]. In

insects, community and student-led annotation efforts have helped identify families of well-conserved genes and proteins[79-81], but species-specific genes and genes under positive selection are more difficult to annotate and require detailed, tissue-specific proteomic investigation[82-85]. As more annotations come online, and with continued curation by the community, pipeline generated genome annotations will continue to improve.

There is a wealth of public genome assemblies, even multiple genomes from the same species, challenging the concept of a reference genome. In this pan-genome era, researchers are realizing that one individual cannot capture the genomic (and therefore proteomic) diversity of a population or species. In other words, more than a single human, aphid or African elephant reference genome may be required to describe the genome or proteome. Moreover, even with reference genomes there will be numerous annotation versions as gene predictions improve, more evidence is acquired, assemblies are improved, or more genomes are completed. The reference human annotation is on its tenth release, and less obvious species like the cow are on their seventh release. For protein databases to evolve and effectively capitalize on pan-genomic data, a new framework will need to be developed.

Analyzing proteomic data from non-model organisms

Given the global resources devoted to genome annotation and how these data products are maintained and propagated throughout data hubs, when possible the greatest community benefits will come from applying resources to genome annotation (described above), as opposed to one-off transcriptomes. Yet there are still occasions where it is not possible due to resources, time, or simply availability of quality samples for genomic sequencing. If samples can be acquired with RNA stability in mind, it should be possible to generate predicted protein sequences using RNA sequencing (such as[86]; Figure 4A). In cases such as biofluids, choosing a relevant tissue to sequence may be difficult since protein provenance is unknown. For

instance, due to its proximity to all organs the complete blood proteome is translated in many tissues, not the blood itself[87]. Another consideration is that when using short-read RNA sequencing, *de novo* assembly is required, which can be computationally costly and may not provide sequence coverage of protein isoforms or very large proteins (*e.g.*, human titin isoform N2BA mRNA is over 100 000 residues in length). To overcome these issues, full-length RNA sequencing can be employed[88], negating the need for transcriptome assembly since the direct transcript evidence can define the search space. As the capability and accessibility of full-length RNA sequencing improves, this could become a viable parallel approach to support mass spectrometry-based proteomics.

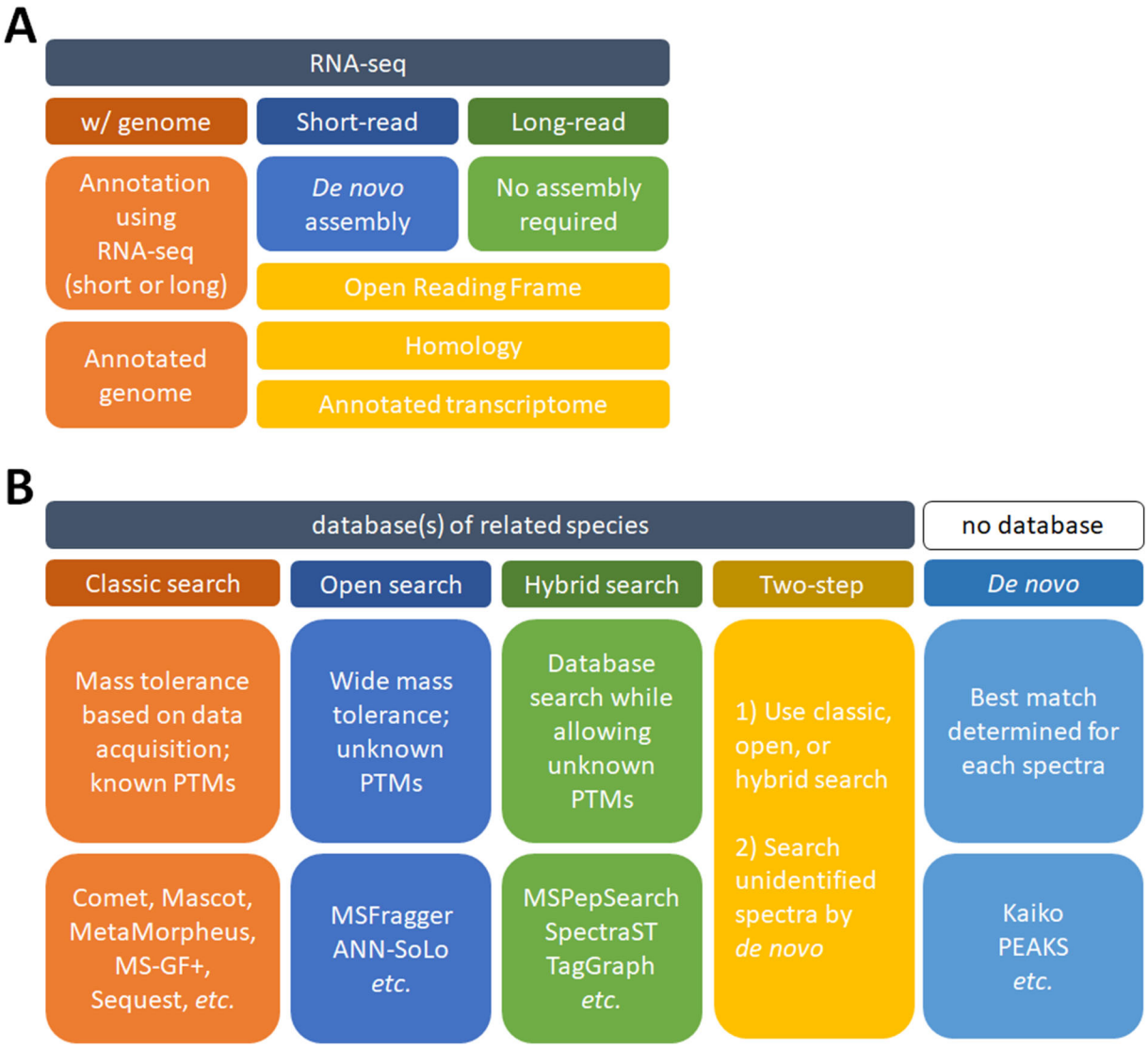


Figure 4. Different approaches to analyzing proteomic data without an annotated genome. **A.** Using nucleic acid sequencing to generate an annotated genome or transcriptome. These can be used to analyze shotgun proteomic data. **B.** Searching shotgun proteomic data without prior/relevant nucleic acid sequencing. The software listed is not exhaustive and there are many suitable alternatives.

Proteomic analysis of a non-model organism in the absence of any additional nucleic acid sequencing is also possible (Figure 4B). Homology searching is when the database of a closely related species is used for conventional database searching. In this case, only tryptic peptides with complete homology will be identified. Depending on the clade this may work better than others (*e.g.*, using the Weddell seal and Pacific walrus annotations to search California sea lion proteomic data provided valid conclusions[89]). Recently an approach was developed that can assist in selecting species by evaluating how well a homologous sequence database can explain an unknown proteome[63]. Other search approaches can involve using wide mass tolerances such that single amino acid variants (SAAVs) can be tolerated between the known and unknown species. This is accomplished by employing an open search with a wide mass tolerance, and searching using a closely related species with a tool like MSFragger[90] or ANN-SoLo[91]. Similarly, utilizing a hybrid spectral library approach may work with closely related species[92]. Also, TagGraph[93] should likewise be applicable with an unknown proteome by combining database and *de novo* searching to identify unknown post-translational modifications (PTMs). Another approach is to take the remaining unmatched spectra from any of these approaches and perform *de novo* searching with tools such as Kaiko[94], PEAKS[95], or those packaged within DeNovoGUI[96]. Using *de novo* proteomic techniques in conjunction with database searching, or as a stand-alone search, is a viable alternative when database searching of shotgun proteomic data isn't possible. Although the accuracy of *de novo* is still improving, recent and ongoing advancements continue to make *de novo* a valid approach. Finally, it is important to note that employing these search approaches may be useful even in well studied organisms where we have complete confidence in the validity of sequence databases. Until we appreciate and account for the genetic diversity within a species, there will always remain the possibility that the reference sequence database is incomplete or inaccurate.

In the given suggestions, it is assumed that the source of the proteome is known. In some cases there may be a primary or secondary unidentified species. In the case of unknown

disease agents, we can use proteomics where other methods fail due to antibody reactivity, RNA degradation or unknown PCR probes. This can present an enormous search space (such as the nearly 12 million proteins listed on UniProtKB for Fungi), but one that is not insurmountable given the numerous databases available, though high computational costs and significant false-discovery concerns exist. Conversely, we may be tasked with identifying non-model organisms within model systems. For instance, when studying the gut proteome there is the microbiome and host, but there are also remnants of ingested food. In an effort to correctly define search space and accurately identify mass spectra, we must account for numerous unknowns. In these cases, multiple search steps, greater reliance on known spectral libraries, or *de novo* is required.

Technical capabilities, hurdles, and future advancements

The discussion and solutions presented so far have focused on current approaches to define and evaluate the search space for shotgun proteomic analysis in non-model organisms, but there are other proteomic approaches worth mentioning. In contrast to digesting proteins prior to analysis, intact proteins can be analyzed using mass spectrometry by top-down proteomics[97-99]. Although top-down is not yet routinely used in non-model organisms, it would face similar hurdles as those in humans, but given accurate protein sequences top-down non-model proteomics is possible. Protein arrays, composed of antibodies, lectins, or aptamers, are another possible approach to proteomics in non-model organisms. Though these arrays are typically geared towards human proteins, such as the SOMAscan, PETAL[100], or Olink platforms, they may be used in other species with varying success when there is enough cross-reactivity between conserved epitopes[101-102]. It seems likely that arrays will be developed for animal-models if the market exists. There is a paucity of commercial antibodies against proteins from non-model organisms, therefore targeted mass spectrometry-based proteomics is a convenient alternative and a tremendously valuable tool to aid researchers interested in quantification of

specific proteins in non-model organisms[31, 103-104]. Overall, it seems only a matter of time before other proteomic techniques become routine in non-model organism studies.

Once protein identifications have been made, data interpretation in non-model organisms presents its own unique challenges. A typical workflow will result in a list of differentially abundant proteins, or proteins that are drivers of some difference, and this list is used for downstream analysis. Depending on the source of the non-model organism sequence database, the protein identification may be meaningful (*e.g.*, alpha-2 microglobulin) or it may simply be a generic locus or transcript ID which is unsuitable for biological interpretation. One possible first step is to convert the identifications to a known model-species by using BLAST to assign protein sequences to their model-species equivalent (a notable implementation is the PAW BLASTer[105]). When determining orthology, the conversion may still require manual inspection due to the high degree of homology between certain proteins (*e.g.*, human alpha-2 microglobulin and pregnancy zone protein are 72% identical). Once the protein identifications have been converted to a model-species, a typical downstream analysis is gene set enrichment or pathways analysis. The underpinning biology and literature inform these lists, which means results from a non-model organism may be irrelevant or at least filled with caveats. Currently, the Molecular Signatures Database (MSigDB v7.0) includes five species (the vast majority being human and mouse), the Reactome pathway database and WebGestalt have 15 and 12 species, respectively, while the Plant Reactome has 97 plant species[106] and the OmicsDB::Pathogens has 11 species[107]. But the further away, phylogenetically speaking, you move from the few well-studied model organisms, the less applicable the results may be. One alternative is to use co-expression or stabilized regression analysis[108], which can utilize abundance patterns to identify novel pathways or gene sets. Although these approaches are far from perfect, they do allow conditional conclusions to be reached and generation of testable hypotheses.

Proteomics provides unique insights into biology

Given the hurdles described, an apparent question is why use proteomics to study non-model organisms. Studying biology at the molecular level is not mutually exclusive to nucleic acid sequencing techniques, but there are cases where proteomics provides more relevant and actionable insights as compared to gene or transcript studies. Since the abundance of the majority of proteins does not correlate with transcript or gene abundance[109-115], proteomic analysis can provide a clearer understanding of the plasticity of phenotypes. Moreover, there are specific cases where proteomics can provide more relevant and actionable insights, such as biofluids (blood, urine, cerebrospinal fluid, insect hemolymph, *etc.*) where the mRNA in the fluid is not related to the proteome of the biofluid. Instead, the proteome of a biofluid is reflected by organs proximal to the fluid and any pathology causing flux between the fluid and organ(s). Also, it is important to note that genes do not change with age, but protein abundance, turnover, localization, modifications and interactions do change[116-117]. In other words, phenotypes may arise from unique combinations of protein interactions and protein abundance, which will not be captured from genomic information alone. In addition to examining protein abundance differences, proteomic analysis can capture measurements of functionally important endogenous peptides[118-119], PTMs[120] and proteoforms[121]. In the case of proteoforms especially, abundance is highly dynamic, such as histone H4 with over 40 proteoforms that may vary over two orders of magnitude[122]. It is not to say that other techniques cannot answer these questions, but it is important to note where proteomics is uniquely capable.

The complexity of the proteome is further magnified when one considers that different proteoforms can interact to form functional protein complexes with unknown roles in cellular homeostasis and pathogensis (*e.g.*, tumor suppressor protein PTEN[123]). Only proteomic methods can provide quantitative measurements of protein complex formation in cells at proteome-wide scales[29, 124-132], improving our understanding of the evolutionary origins of protein interactions, protein co-localization, and functional protein interaction networks[133]. Genetic and proteomic diversity also extends to an organism's microbiota[134]. This proteomic diversity is

critical to understanding biological phenotypes, including cancer[135] and transmission of viruses[134, 136]. Metaproteomic techniques can also provide insights into host-symbiont functional compartmentalization of proteomes, predictions on resident microbiota in eukaryotic host tissues and knowledge on co-evolutionary mechanisms regulating symbiosis in non-model organisms is crucial to our understanding of eukaryotic biology[137-139]. It is evident that proteomics can provide an additional modality to improve nucleic acid based analyses, while also providing unique insight into the molecular landscape responsible for a given phenotype or pathology.

Future Outlook

There is great potential in applying proteomics in non-model organisms and the technical hurdles are not formidable. This is best exemplified by a recent large scale proteomic study that cataloged 340 000 proteins from 100 species[140]. In the future, proteomic studies of non-model organisms will undoubtedly become more commonplace. As it becomes possible to define the molecular landscape in species across the tree of life, the next step is finding the non-model organism experts. But this doesn't need to be for an exotic species in a distant land. In the shallow waters along the eastern seaboard is a euryhaline stingray that can achieve and maintain an extraordinarily high renal transtubular osmotic gradient[141] and regenerate glomeruli[142]. Understanding the elasmobranch kidney could provide insight into chronic disease in humans. Exciting topics like this are all around us if we just look, explore and experiment. Proteomics researchers must work with biologists in these adjacent fields who can direct experimental design of naturally occurring phenotypes to empower comparative studies, and may be unaware of the power of proteomics or its accessibility via cores and collaborations. Empowering these scientists with advanced bioanalytical capabilities is a major goal of proteomics in non-model organisms. There is great potential to address pressing issues such as food supply security, help facilitate commerce, protect human and animal health, and use these

techniques in biomimetics to help accelerate biomedical breakthroughs. The post-model organism era opens the door to new applications of maturing technology, hopefully leading to new fundamental biological truths unattainable using classic model organisms.

Acknowledgements

The graphics in Figure 1 were created specifically for this manuscript by Debbie Maizels. Funding for MH by USDA ARS Project # 8062-22410-006-00-D. The authors wish to thank Juan Antonio Vizcaino (European Bioinformatics Institute) for providing current information on data sets at PRIDE, and Magnus Palmblad (Leiden University Medical Center) and Michael G. Janech (College of Charleston) for critical feedback. Identification of certain commercial equipment, instruments, software or materials does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products identified are necessarily the best available for the purpose.

Competing Interests: The authors declare that they have no competing financial interests.

Contributions: M.H. and B.N. conceived and wrote the manuscript.

Supporting Information:

The following supporting information is available free of charge at ACS website http://pubs.acs.org/

Supplementary Table S1. Specific NCBI taxon IDs and their designation for 23 mammal groups used in Figures 2 and 3.

References

1. Muller, B.; Grossniklaus, U., Model organisms--A historical perspective. *Journal of proteomics* **2010,** *73* (11), 2054-63.
2. Tang, B.; Wang, Y.; Zhu, J.; Zhao, W., Web resources for model organism studies. *Genomics, proteomics & bioinformatics* **2015,** *13* (1), 64-8.
3. Russell, J. J.; Theriot, J. A.; Sood, P.; Marshall, W. F.; Landweber, L. F.; Fritz-Laylin, L.; Polka, J. K.; Oliferenko, S.; Gerbich, T.; Gladfelter, A.; Umen, J.; Bezanilla, M.; Lancaster, M. A.; He, S.; Gibson, M. C.; Goldstein, B.; Tanaka, E. M.; Hu, C. K.; Brunet, A., Non-model model organisms. *BMC biology* **2017,** *15* (1), 55.
4. Goldstein, B.; King, N., The Future of Cell Biology: Emerging Model Organisms. *Trends in cell biology* **2016,** *26* (11), 818-824.
5. Stenvinkel, P.; Painer, J.; Johnson, R. J.; Natterson-Horowitz, B., Biomimetics - Nature's roadmap to insights and solutions for burden of lifestyle diseases. *Journal of internal medicine* **2020,** *287* (3), 238-251.
6. Stenvinkel, P.; Painer, J.; Kuro, O. M.; Lanaspa, M.; Arnold, W.; Ruf, T.; Shiels, P. G.; Johnson, R. J., Novel treatment strategies for chronic kidney disease: insights from the animal kingdom. *Nature reviews. Nephrology* **2018,** *14* (4), 265-284.
7. Callier, V., Core Concept: Solving Peto's Paradox to better understand cancer. *Proceedings of the National Academy of Sciences of the United States of America* **2019,** *116* (6), 1825-1828.
8. Vazquez, J. M.; Sulak, M.; Chigurupati, S.; Lynch, V. J., A Zombie LIF Gene in Elephants Is Upregulated by TP53 to Induce Apoptosis in Response to DNA Damage. *Cell reports* **2018,** *24* (7), 1765-1776.
9. Seluanov, A.; Gladyshev, V. N.; Vijg, J.; Gorbunova, V., Mechanisms of cancer resistance in long-lived mammals. *Nature reviews. Cancer* **2018,** *18* (7), 433-441.
10. Li, C.; Zhao, H.; Liu, Z.; McMahon, C., Deer antler--a novel model for studying organ regeneration in mammals. *The international journal of biochemistry & cell biology* **2014,** *56*, 111-22.
11. Dong, Z.; Haines, S.; Coates, D., Proteomic Profiling of Stem Cell Tissues during Regeneration of Deer Antler: A Model of Mammalian Organ Regeneration. *Journal of Proteome Research* **2020,** *19* (4), 1760–1775.
12. Zapol, W. M.; Liggins, G. C.; Schneider, R. C.; Qvist, J.; Snider, M. T.; Creasy, R. K.; Hochachka, P. W., Regional blood flow during simulated diving in the conscious Weddell seal. *Journal of applied physiology: respiratory, environmental and exercise physiology* **1979,** *47* (5), 968-73.
13. Halasz, N. A.; Elsner, R.; Garvie, R. S.; Grotke, G. T., Renal recovery from ischemia: a comparative study of harbor seal and dog kidneys. *The American journal of physiology* **1974,** *227* (6), 1331-5.
14. Vazquez-Medina, J. P.; Zenteno-Savin, T.; Elsner, R.; Ortiz, R. M., Coping with physiological oxidative stress: a review of antioxidant strategies in seals. *Journal of comparative physiology. B, Biochemical, systemic, and environmental physiology* **2012,** *182* (6), 741-50.
15. Elsner, R.; Oyasaeter, S.; Almaas, R.; Saugstad, O. D., Diving seals, ischemia-reperfusion and oxygen radicals. *Comparative biochemistry and physiology. Part A, Molecular & integrative physiology* **1998,** *119* (4), 975-80.
16. Allen, K. N.; Vazquez-Medina, J. P., Natural Tolerance to Ischemia and Hypoxemia in Diving Mammals: A Review. *Frontiers in physiology* **2019,** *10*, 1199.
17. Drew, K. L.; Rice, M. E.; Kuhn, T. B.; Smith, M. A., Neuroprotective adaptations in hibernation: therapeutic implications for ischemia-reperfusion, traumatic brain injury and neurodegenerative diseases. *Free radical biology & medicine* **2001,** *31* (5), 563-73.

18. Grabek, K. R.; Cooke, T. F.; Epperson, L. E.; Spees, K. K.; Cabral, G. F.; Sutton, S. C.; Merriman, D. K.; Martin, S. L.; Bustamante, C. D., Genetic variation drives seasonal onset of hibernation in the 13-lined ground squirrel. *Communications biology* **2019,** *2*, 478.
19. Jansen, H. T.; Trojahn, S.; Saxton, M. W.; Quackenbush, C. R.; Evans Hutzenbiler, B. D.; Nelson, O. L.; Cornejo, O. E.; Robbins, C. T.; Kelley, J. L., Hibernation induces widespread transcriptional remodeling in metabolic tissues of the grizzly bear. *Communications biology* **2019,** *2*, 336.
20. Austad, S. N., Methusaleh's Zoo: how nature provides us with clues for extending human health span. *Journal of comparative pathology* **2010,** *142 Suppl 1*, S10-21.
21. Wilkinson, G. S.; Adams, D. M., Recurrent evolution of extreme longevity in bats. *Biology letters* **2019,** *15* (4), 20180860.
22. Podlutsky, A. J.; Khritankov, A. M.; Ovodov, N. D.; Austad, S. N., A new field record for bat longevity. *The journals of gerontology. Series A, Biological sciences and medical sciences* **2005,** *60* (11), 1366-8.
23. Johnson, A. A.; Shokhirev, M. N.; Shoshitaishvili, B., Revamping the evolutionary theories of aging. *Ageing research reviews* **2019,** *55*, 100947.
24. Bernard, C.; Compagnoni, A.; Salguero-Gómez, R., Testing Finch's hypothesis: The role of organismal modularity on the escape from actuarial senescence. *Functional Ecology* **2020,** *34* (1), 88-106.
25. Wang, L.; Cui, J.; Jin, B.; Zhao, J.; Xu, H.; Lu, Z.; Li, W.; Li, X.; Li, L.; Liang, E.; Rao, X.; Wang, S.; Fu, C.; Cao, F.; Dixon, R. A.; Lin, J., Multifeature analyses of vascular cambial cells reveal longevity mechanisms in old Ginkgo biloba trees. *Proceedings of the National Academy of Sciences of the United States of America* **2020,** *117* (4), 2201-2210.
26. Chin, E. L.; Ramsey, J. S.; Mishchuk, D. O.; Saha, S.; Foster, E.; Chavez, J. D.; Howe, K.; Zhong, X.; Polek, M.; Godfrey, K. E.; Mueller, L. A.; Bruce, J. E.; Heck, M.; Slupsky, C. M., Longitudinal Transcriptomic, Proteomic, and Metabolomic Analyses of Citrus sinensis (L.) Osbeck Graft-Inoculated with "Candidatus Liberibacter asiaticus". *J Proteome Res* **2020,** *19* (2), 719-732.
27. Ramsey, J. S.; Chin, E. L.; Chavez, J. D.; Saha, S.; Mischuk, D.; Mahoney, J.; Mohr, J.; Robison, F. M.; Mitrovic, E.; Xu, Y.; Strickler, S. R.; Fernandez, N.; Zhong, X.; Polek, M.; Godfrey, K. E.; Giovannoni, J. J.; Mueller, L. A.; Slupsky, C. M.; Bruce, J. E.; Heck, M., Longitudinal Transcriptomic, Proteomic, and Metabolomic Analysis of Citrus limon Response to Graft Inoculation by Candidatus Liberibacter asiaticus. *J Proteome Res* **2020,** *19* (6), 2247-2263.
28. Kruse, A.; Fleites, L. A.; Heck, M., Lessons from One Fastidious Bacterium to Another: What Can We Learn about Liberibacter Species from Xylella fastidiosa. *Insects* **2019,** *10* (9), 300.
29. DeBlasio, S. L.; Chavez, J. D.; Alexander, M. M.; Ramsey, J.; Eng, J. K.; Mahoney, J.; Gray, S. M.; Bruce, J. E.; Cilia, M., Visualization of Host-Polerovirus Interaction Topologies Using Protein Interaction Reporter Technology. *J Virol* **2015,** *90* (4), 1973-87.
30. DeBlasio, S. L.; Johnson, R.; Mahoney, J.; Karasev, A.; Gray, S.; Cilia, M., Insights into the polerovirus-plant interactome revealed by co-immunoprecipitation and mass spectrometry. *Molecular Plant-Microbe Interactions* **2015,** *28* (4), 467-481.
31. DeBlasio, S. L.; Johnson, R.; Sweeney, M. M.; Karasev, A.; Gray, S. M.; MacCoss, M. J.; Cilia, M., Potato leafroll virus structural proteins manipulate overlapping, yet distinct protein interaction networks during infection. *Proteomics* **2015,** *15* (12), 2098-112.
32. DeBlasio, S. L.; Johnson, R. S.; MacCoss, M. J.; Gray, S. M.; Cilia, M., Model System-Guided Protein Interaction Mapping for Virus Isolated from Phloem Tissue. *J Proteome Res* **2016,** *15* (12), 4601-4611.
33. DeBlasio, S. L.; Xu, Y.; Johnson, R. S.; Rebelo, A. R.; MacCoss, M. J.; Gray, S. M.; Heck, M., The Interaction Dynamics of Two Potato Leafroll Virus Movement Proteins Affects

Their Localization to the Outer Membranes of Mitochondria and Plastids. *Viruses* **2018,** *10* (11), 585.
34. IBISWorld Zoos & Aquariums in the US industry statistics. https://www.ibisworld.com/united-states/market-research-reports/zoos-aquariums-industry/ (accessed 27 March).
35. APPA Pet Industry Market Size & Ownership Statistics. https://www.americanpetproducts.org/press_industrytrends.asp (accessed 27 March).
36. Jebb, D.; Huang, Z.; Pippel, M.; Hughes, G. M.; Lavrichenko, K.; Devanna, P.; Winkler, S.; Jermiin, L. S.; Skirmuntt, E. C.; Katzourakis, A.; Burkitt-Gray, L.; Ray, D. A.; Sullivan, K. A. M.; Roscito, J. G.; Kirilenko, B. M.; Dávalos, L. M.; Corthals, A. P.; Power, M. L.; Jones, G.; Ransome, R. D.; Dechmann, D.; Locatelli, A. G.; Puechmaille, S. J.; Fedrigo, O.; Jarvis, E. D.; Springer, M. S.; Hiller, M.; Vernes, S. C.; Myers, E. W.; Teeling, E. C., Six new reference-quality bat genomes illuminate the molecular basis and evolution of bat adaptations. *bioRxiv* **2019**, 836874.
37. Pavlovich, S. S.; Lovett, S. P.; Koroleva, G.; Guito, J. C.; Arnold, C. E.; Nagle, E. R.; Kulcsar, K.; Lee, A.; Thibaud-Nissen, F.; Hume, A. J.; Muhlberger, E.; Uebelhoer, L. S.; Towner, J. S.; Rabadan, R.; Sanchez-Lockhart, M.; Kepler, T. B.; Palacios, G., The Egyptian Rousette Genome Reveals Unexpected Features of Bat Antiviral Immunity. *Cell* **2018,** *173* (5), 1098-1110.e18.
38. Huang, Z.; Whelan, C. V.; Foley, N. M.; Jebb, D.; Touzalin, F.; Petit, E. J.; Puechmaille, S. J.; Teeling, E. C., Longitudinal comparative transcriptomics reveals unique mechanisms underlying extended healthspan in bats. *Nature ecology & evolution* **2019,** *3* (7), 1110-1120.
39. Valcu, C.-M.; Kempenaers, B., Proteomics in behavioral ecology. *Behavioral Ecology* **2014,** *26* (1), 1-15.
40. Mulugeta, E.; Marion-Poll, L.; Gentien, D.; Ganswindt, S. B.; Ganswindt, A.; Bennett, N. C.; Blackburn, E. H.; Faulkes, C. G.; Heard, E., Molecular insights into the pathways underlying naked mole-rat eusociality. *bioRxiv* **2017**, 209932.
41. Paredi, G.; Sentandreu, M. A.; Mozzarelli, A.; Fadda, S.; Hollung, K.; de Almeida, A. M., Muscle and meat: new horizons and applications for proteomics on a farm to fork perspective. *Journal of proteomics* **2013,** *88*, 58-82.
42. Ortea, I.; O'Connor, G.; Maquet, A., Review on proteomics for food authentication. *Journal of proteomics* **2016,** *147*, 212-225.
43. Piras, C.; Roncada, P.; Rodrigues, P. M.; Bonizzi, L.; Soggiu, A., Proteomics in food: Quality, safety, microbes, and allergens. *Proteomics* **2016,** *16* (5), 799-815.
44. Wulff, T.; Nielsen, M. E.; Deelder, A. M.; Jessen, F.; Palmblad, M., Authentication of fish products by large-scale comparison of tandem mass spectra. *J Proteome Res* **2013,** *12* (11), 5253-9.
45. Righetti, P. G.; Fasoli, E.; D'Amato, A.; Boschetti, E., Chapter 6 - Making Progress in Plant Proteomics for Improved Food Safety. In *Comprehensive Analytical Chemistry*, García-Cañas, V.; Cifuentes, A.; Simó, C., Eds. Elsevier: 2014; Vol. 64, pp 131-155.
46. Becker, D. J.; Albery, G. F.; Sjodin, A. R.; Poisot, T.; Dallas, T. A.; Eskew, E. A.; Farrell, M. J.; Guth, S.; Han, B. A.; Simmons, N. B.; Carlson, C. J., Predicting wildlife hosts of betacoronaviruses for SARS-CoV-2 sampling prioritization. *bioRxiv* **2020**, 2020.05.22.111344.
47. Letko, M.; Seifert, S. N.; Olival, K. J.; Plowright, R. K.; Munster, V. J., Bat-borne virus diversity, spillover and emergence. *Nature reviews. Microbiology* **2020,** *18*, 461–471.
48. Li, F., Receptor recognition and cross-species infections of SARS coronavirus. *Antiviral research* **2013,** *100* (1), 246-54.
49. Li, W.; Zhang, C.; Sui, J.; Kuhn, J. H.; Moore, M. J.; Luo, S.; Wong, S. K.; Huang, I. C.; Xu, K.; Vasilieva, N.; Murakami, A.; He, Y.; Marasco, W. A.; Guan, Y.; Choe, H.; Farzan, M., Receptor and viral determinants of SARS-coronavirus adaptation to human ACE2. *The EMBO journal* **2005,** *24* (8), 1634-43.

50. Zhao, X.; Chen, D.; Szabla, R.; Zheng, M.; Li, G.; Du, P.; Zheng, S.; Li, X.; Song, C.; Li, R.; Guo, J.-T.; Junop, M.; Zeng, H.; Lin, H., Broad and differential animal ACE2 receptor usage by SARS-CoV-2. *bioRxiv* **2020**, 2020.04.19.048710.
51. Lan, J.; Ge, J.; Yu, J.; Shan, S.; Zhou, H.; Fan, S.; Zhang, Q.; Shi, X.; Wang, Q.; Zhang, L.; Wang, X., Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature* **2020,** *581* (7807), 215-220.
52. Watanabe, Y.; Berndsen, Z. T.; Raghwani, J.; Seabright, G. E.; Allen, J. D.; Pybus, O. G.; McLellan, J. S.; Wilson, I. A.; Bowden, T. A.; Ward, A. B.; Crispin, M., Vulnerabilities in coronavirus glycan shields despite extensive glycosylation. *Nature communications* **2020,** *11* (1), 2688.
53. Pach, S.; Nguyen, T. N.; Trimpert, J.; Kunec, D.; Osterrieder, N.; Wolber, G., ACE2-Variants Indicate Potential SARS-CoV-2-Susceptibility in Animals: An Extensive Molecular Dynamics Study. *bioRxiv* **2020**, 2020.05.14.092767.
54. Damas, J.; Hughes, G. M.; Keough, K. C.; Painter, C. A.; Persky, N. S.; Corbo, M.; Hiller, M.; Koepfli, K.-P.; Pfenning, A. R.; Zhao, H.; Genereux, D. P.; Swofford, R.; Pollard, K. S.; Ryder, O. A.; Nweeia, M. T.; Lindblad-Toh, K.; Teeling, E. C.; Karlsson, E. K.; Lewin, H. A., Broad Host Range of SARS-CoV-2 Predicted by Comparative and Structural Analysis of ACE2 in Vertebrates. *bioRxiv* **2020**, 2020.04.16.045302.
55. Batra, J.; Hultquist, J. F.; Liu, D.; Shtanko, O.; Von Dollen, J.; Satkamp, L.; Jang, G. M.; Luthra, P.; Schwarz, T. M.; Small, G. I.; Arnett, E.; Anantpadma, M.; Reyes, A.; Leung, D. W.; Kaake, R.; Haas, P.; Schmidt, C. B.; Schlesinger, L. S.; LaCount, D. J.; Davey, R. A.; Amarasinghe, G. K.; Basler, C. F.; Krogan, N. J., Protein Interaction Mapping Identifies RBBP6 as a Negative Regulator of Ebola Virus Replication. *Cell* **2018,** *175* (7), 1917-1930.e13.
56. Shah, P. S.; Link, N.; Jang, G. M.; Sharp, P. P.; Zhu, T.; Swaney, D. L.; Johnson, J. R.; Von Dollen, J.; Ramage, H. R.; Satkamp, L.; Newton, B.; Hüttenhain, R.; Petit, M. J.; Baum, T.; Everitt, A.; Laufman, O.; Tassetto, M.; Shales, M.; Stevenson, E.; Iglesias, G. N.; Shokat, L.; Tripathi, S.; Balasubramaniam, V.; Webb, L. G.; Aguirre, S.; Willsey, A. J.; Garcia-Sastre, A.; Pollard, K. S.; Cherry, S.; Gamarnik, A. V.; Marazzi, I.; Taunton, J.; Fernandez-Sesma, A.; Bellen, H. J.; Andino, R.; Krogan, N. J., Comparative Flavivirus-Host Protein Interaction Mapping Reveals Mechanisms of Dengue and Zika Virus Pathogenesis. *Cell* **2018,** *175* (7), 1931-1945.e18.
57. Gordon, D. E.; Jang, G. M.; Bouhaddou, M.; Xu, J.; Obernier, K.; White, K. M.; O'Meara, M. J.; Rezelj, V. V.; Guo, J. Z.; Swaney, D. L.; Tummino, T. A.; Huettenhain, R.; Kaake, R. M.; Richards, A. L.; Tutuncuoglu, B.; Foussard, H.; Batra, J.; Haas, K.; Modak, M.; Kim, M.; Haas, P.; Polacco, B. J.; Braberg, H.; Fabius, J. M.; Eckhardt, M.; Soucheray, M.; Bennett, M. J.; Cakir, M.; McGregor, M. J.; Li, Q.; Meyer, B.; Roesch, F.; Vallet, T.; Mac Kain, A.; Miorin, L.; Moreno, E.; Naing, Z. Z. C.; Zhou, Y.; Peng, S.; Shi, Y.; Zhang, Z.; Shen, W.; Kirby, I. T.; Melnyk, J. E.; Chorba, J. S.; Lou, K.; Dai, S. A.; Barrio-Hernandez, I.; Memon, D.; Hernandez-Armenta, C.; Lyu, J.; Mathy, C. J. P.; Perica, T.; Pilla, K. B.; Ganesan, S. J.; Saltzberg, D. J.; Rakesh, R.; Liu, X.; Rosenthal, S. B.; Calviello, L.; Venkataramanan, S.; Liboy-Lugo, J.; Lin, Y.; Huang, X. P.; Liu, Y.; Wankowicz, S. A.; Bohn, M.; Safari, M.; Ugur, F. S.; Koh, C.; Savar, N. S.; Tran, Q. D.; Shengjuler, D.; Fletcher, S. J.; O'Neal, M. C.; Cai, Y.; Chang, J. C. J.; Broadhurst, D. J.; Klippsten, S.; Sharp, P. P.; Wenzell, N. A.; Kuzuoglu, D.; Wang, H. Y.; Trenker, R.; Young, J. M.; Cavero, D. A.; Hiatt, J.; Roth, T. L.; Rathore, U.; Subramanian, A.; Noack, J.; Hubert, M.; Stroud, R. M.; Frankel, A. D.; Rosenberg, O. S.; Verba, K. A.; Agard, D. A.; Ott, M.; Emerman, M.; Jura, N.; von Zastrow, M.; Verdin, E.; Ashworth, A.; Schwartz, O.; d'Enfert, C.; Mukherjee, S.; Jacobson, M.; Malik, H. S.; Fujimori, D. G.; Ideker, T.; Craik, C. S.; Floor, S. N.; Fraser, J. S.; Gross, J. D.; Sali, A.; Roth, B. L.; Ruggero, D.; Taunton, J.; Kortemme, T.; Beltrao, P.; Vignuzzi, M.; García-Sastre, A.; Shokat, K. M.; Shoichet, B. K.; Krogan, N. J., A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* **2020,** *583*, 459–468.

58. Muntel, J.; Gandhi, T.; Verbeke, L.; Bernhardt, O. M.; Treiber, T.; Bruderer, R.; Reiter, L., Surpassing 10 000 identified and quantified proteins in a single run by optimizing current LC-MS instrumentation and data analysis strategy. *Molecular omics* **2019,** *15* (5), 348-360.
59. Bekker-Jensen, D. B.; Kelstrup, C. D.; Batth, T. S.; Larsen, S. C.; Haldrup, C.; Bramsen, J. B.; Sorensen, K. D.; Hoyer, S.; Orntoft, T. F.; Andersen, C. L.; Nielsen, M. L.; Olsen, J. V., An Optimized Shotgun Strategy for the Rapid Generation of Comprehensive Human Proteomes. *Cell systems* **2017,** *4* (6), 587-599.e4.
60. Richards, A. L.; Hebert, A. S.; Ulbrich, A.; Bailey, D. J.; Coughlin, E. E.; Westphall, M. S.; Coon, J. J., One-hour proteome analysis in yeast. *Nature protocols* **2015,** *10* (5), 701-14.
61. Meier, F.; Geyer, P. E.; Virreira Winter, S.; Cox, J.; Mann, M., BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes. *Nature methods* **2018,** *15* (6), 440-448.
62. Cilia, M.; Tamborindeguy, C.; Rolland, M.; Howe, K.; Thannhauser, T. W.; Gray, S., Tangible benefits of the aphid Acyrthosiphon pisum genome sequencing for aphid proteomics: Enhancements in protein identification and data validation for homology-based proteomics. *J Insect Physiol* **2011,** *57* (1), 179-90.
63. Johnson, R. S.; Searle, B. C.; Nunn, B. L.; Gilmore, J. M.; Phillips, M.; Amemiya, C. T.; Heck, M.; MacCoss, M. J., Assessing Protein Sequence Database Suitability Using De Novo Sequencing. *Molecular & cellular proteomics : MCP* **2020,** *19* (1), 198-208.
64. Wetterstrand, K. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). www.genome.gov/sequencingcostsdata (accessed 10 March).
65. Vaughan, T. A.; Ryan, J. M.; Czaplewski, N. J., *Mammalogy*. 6th ed.; Jones & Bartlett Learning: U.S.A., 2013; p 756.
66. Neely, B. A.; Ellisor, D. L.; Davis, W. C., Proteomics as a metrological tool to evaluate genome annotation accuracy following *de novo* genome assembly: a case study using the Atlantic bottlenose dolphin (*Tursiops truncatus*). *bioRxiv* **2018**, 254250.
67. Simão, F. A.; Waterhouse, R. M.; Ioannidis, P.; Kriventseva, E. V.; Zdobnov, E. M., BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics (Oxford, England)* **2015,** *31* (19), 3210-2.
68. Volkening, J. D.; Bailey, D. J.; Rose, C. M.; Grimsrud, P. A.; Howes-Podoll, M.; Venkateshwaran, M.; Westphall, M. S.; Ane, J. M.; Coon, J. J.; Sussman, M. R., A proteogenomic survey of the Medicago truncatula genome. *Mol Cell Proteomics* **2012,** *11* (10), 933-44.
69. Carolan, J. C.; Caragea, D.; Reardon, K. T.; Mutti, N. S.; Dittmer, N.; Pappan, K.; Cui, F.; Castaneto, M.; Poulain, J.; Dossat, C.; Tagu, D.; Reese, J. C.; Reeck, G. R.; Wilkinson, T. L.; Edwards, O. R., Predicted effector molecules in the salivary secretome of the pea aphid (Acyrthosiphon pisum): a dual transcriptomic/proteomic approach. *J Proteome Res* **2011,** *10* (4), 1505-18.
70. Campbell, M. S.; Holt, C.; Moore, B.; Yandell, M., Genome Annotation and Curation Using MAKER and MAKER-P. *Current protocols in bioinformatics* **2014,** *48*, 4.11.1-39.
71. MAKER 2.31.9 with CCTOOLS Jetstream Tutorial. https://cyverse.atlassian.net/wiki/spaces/TUT/pages/258736333/MAKER+2.31.9+with+CCTOOLS+Jetstream+Tutorial (accessed 14 August).
72. Thibaud-Nissen, F.; Souvorov, A.; Murphy, T.; DiCuccio, M.; Kitts, P., Eukaryotic Genome Annotation Pipeline. 2nd ed.; National Center for Biotechnology Information (US): Bethesda (MD), 2013. https://www.ncbi.nlm.nih.gov/books/NBK169439/ (accessed 27 March 2020).
73. O'Leary, N. A.; Wright, M. W.; Brister, J. R.; Ciufo, S.; Haddad, D.; McVeigh, R.; Rajput, B.; Robbertse, B.; Smith-White, B.; Ako-Adjei, D.; Astashyn, A.; Badretdin, A.; Bao, Y.; Blinkova, O.; Brover, V.; Chetvernin, V.; Choi, J.; Cox, E.; Ermolaeva, O.; Farrell, C. M.; Goldfarb, T.; Gupta, T.; Haft, D.; Hatcher, E.; Hlavina, W.; Joardar, V. S.; Kodali, V. K.; Li, W.; Maglott, D.;

Masterson, P.; McGarvey, K. M.; Murphy, M. R.; O'Neill, K.; Pujar, S.; Rangwala, S. H.; Rausch, D.; Riddick, L. D.; Schoch, C.; Shkeda, A.; Storz, S. S.; Sun, H.; Thibaud-Nissen, F.; Tolstoy, I.; Tully, R. E.; Vatsan, A. R.; Wallin, C.; Webb, D.; Wu, W.; Landrum, M. J.; Kimchi, A.; Tatusova, T.; DiCuccio, M.; Kitts, P.; Murphy, T. D.; Pruitt, K. D., Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic acids research* **2016,** *44* (D1), D733-45.
74. NCBI Eukaryotic Genome Annotation Policy On Which Genomes Are Annotated. https://www.ncbi.nlm.nih.gov/genome/annotation_euk/policy/ (accessed 27 March).
75. NCBI Zalophus californianus Annotation Release 100. https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Zalophus_californianus/100/.
76. Hedges, S. B.; Marin, J.; Suleski, M.; Paymer, M.; Kumar, S., Tree of life reveals clock-like speciation and diversification. *Molecular biology and evolution* **2015,** *32* (4), 835-45.
77. Hosseinzadeh, S.; Ramsey, J.; Mann, M.; Bennett, L.; Hunter, W. B.; Shams-Bakhsh, M.; Hall, D. G.; Heck, M., Color morphology of Diaphorina citri influences interactions with its bacterial endosymbionts and 'Candidatus Liberibacter asiaticus'. *PLoS One* **2019,** *14* (5), e0216599.
78. MacDougall, A.; Volynkin, V.; Saidi, R.; Poggioli, D.; Zellner, H.; Hatton-Ellis, E.; Joshi, V.; O'Donovan, C.; Orchard, S.; Auchincloss, A. H.; Baratin, D.; Bolleman, J.; Coudert, E.; de Castro, E.; Hulo, C.; Masson, P.; Pedruzzi, I.; Rivoire, C.; Arighi, C.; Wang, Q.; Chen, C.; Huang, H.; Garavelli, J.; Vinayaka, C. R.; Yeh, L. S.; Natale, D. A.; Laiho, K.; Martin, M., UniRule: a unified rule resource for automatic annotation in the UniProt Knowledgebase. *Bioinformatics (Oxford, England)* **2020**.
79. Chen, W.; Hasegawa, D. K.; Kaur, N.; Kliot, A.; Pinheiro, P. V.; Luan, J.; Stensmyr, M. C.; Zheng, Y.; Liu, W.; Sun, H.; Xu, Y.; Luo, Y.; Kruse, A.; Yang, X.; Kontsedalov, S.; Lebedev, G.; Fisher, T. W.; Nelson, D. R.; Hunter, W. B.; Brown, J. K.; Jander, G.; Cilia, M.; Douglas, A. E.; Ghanim, M.; Simmons, A. M.; Wintermantel, W. M.; Ling, K. S.; Fei, Z., The draft genome of whitefly Bemisia tabaci MEAM1, a global crop pest, provides novel insights into virus transmission, host adaptation, and insecticide resistance. *BMC Biol* **2016,** *14* (1), 110.
80. International Aphid Genomics, C., Genome sequence of the pea aphid Acyrthosiphon pisum. *PLoS Biol* **2010,** *8* (2), e1000313.
81. Saha, S.; Hosmani, P. S.; Villalobos-Ayala, K.; Miller, S.; Shippy, T.; Flores, M.; Rosendale, A.; Cordola, C.; Bell, T.; Mann, H.; DeAvila, G.; DeAvila, D.; Moore, Z.; Buller, K.; Ciolkevich, K.; Nandyal, S.; Mahoney, R.; Van Voorhis, J.; Dunlevy, M.; Farrow, D.; Hunter, D.; Morgan, T.; Shore, K.; Guzman, V.; Izsak, A.; Dixon, D. E.; Cridge, A.; Cano, L.; Cao, X.; Jiang, H.; Leng, N.; Johnson, S.; Cantarel, B. L.; Richards, S.; English, A.; Shatters, R. G.; Childers, C.; Chen, M. J.; Hunter, W.; Cilia, M.; Mueller, L. A.; Munoz-Torres, M.; Nelson, D.; Poelchau, M. F.; Benoit, J. B.; Wiersma-Koch, H.; D'Elia, T.; Brown, S. J., Improved annotation of the insect vector of citrus greening disease: biocuration by a diverse genomics community. *Database (Oxford)* **2017,** *2017,* baz035.
82. Carolan, J. C.; Fitzroy, C. I.; Ashton, P. D.; Douglas, A. E.; Wilkinson, T. L., The secreted salivary proteome of the pea aphid Acyrthosiphon pisum characterised by mass spectrometry. *Proteomics* **2009,** *9* (9), 2457-67.
83. Bos, J. I.; Prince, D.; Pitino, M.; Maffei, M. E.; Win, J.; Hogenhout, S. A., A functional genomics approach identifies candidate effectors from the aphid species Myzus persicae (green peach aphid). *PLoS Genet* **2010,** *6* (11), e1001216.
84. Mathers, T. C.; Chen, Y.; Kaithakottil, G.; Legeai, F.; Mugford, S. T.; Baa-Puyoulet, P.; Bretaudeau, A.; Clavijo, B.; Colella, S.; Collin, O.; Dalmay, T.; Derrien, T.; Feng, H.; Gabaldon, T.; Jordan, A.; Julca, I.; Kettles, G. J.; Kowitwanich, K.; Lavenier, D.; Lenzi, P.; Lopez-Gomollon, S.; Loska, D.; Mapleson, D.; Maumus, F.; Moxon, S.; Price, D. R.; Sugio, A.; van Munster, M.; Uzest, M.; Waite, D.; Jander, G.; Tagu, D.; Wilson, A. C.; van Oosterhout, C.; Swarbreck, D.;

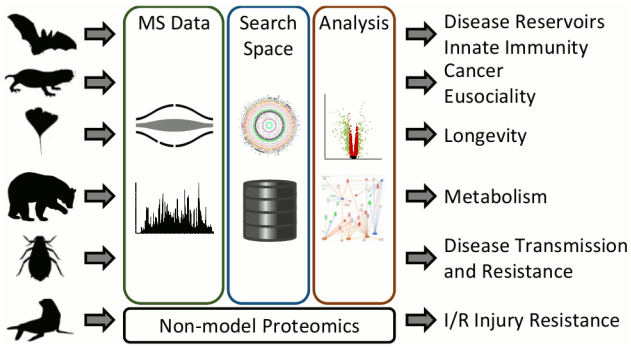
Hogenhout, S. A., Rapid transcriptional plasticity of duplicated gene clusters enables a clonally reproducing aphid to colonise diverse plant species. *Genome Biol* **2017,** *18* (1), 27.
85. Kruse, A.; Ramsey, J. S.; Johnson, R.; Hall, D. G.; MacCoss, M. J.; Heck, M., Candidatus Liberibacter asiaticus Minimally Alters Expression of Immunity and Metabolism Proteins in Hemolymph of Diaphorina citri, the Insect Vector of Huanglongbing. *J Proteome Res* **2018,** *17* (9), 2995-3011.
86. Guo, Q.; Li, D.; Zhai, Y.; Gu, Z., CCPRD: A Novel Analytical Framework for the Comprehensive Proteomic Reference Database Construction of NonModel Organisms. *ACS Omega* **2020,** *5* (25), 15370-15384.
87. Uhlén, M.; Karlsson, M. J.; Hober, A.; Svensson, A. S.; Scheffel, J.; Kotol, D.; Zhong, W.; Tebani, A.; Strandberg, L.; Edfors, F.; Sjöstedt, E.; Mulder, J.; Mardinoglu, A.; Berling, A.; Ekblad, S.; Dannemeyer, M.; Kanje, S.; Rockberg, J.; Lundqvist, M.; Malm, M.; Volk, A. L.; Nilsson, P.; Månberg, A.; Dodig-Crnkovic, T.; Pin, E.; Zwahlen, M.; Oksvold, P.; von Feilitzen, K.; Häussler, R. S.; Hong, M. G.; Lindskog, C.; Ponten, F.; Katona, B.; Vuu, J.; Lindström, E.; Nielsen, J.; Robinson, J.; Ayoglu, B.; Mahdessian, D.; Sullivan, D.; Thul, P.; Danielsson, F.; Stadler, C.; Lundberg, E.; Bergström, G.; Gummesson, A.; Voldborg, B. G.; Tegel, H.; Hober, S.; Forsström, B.; Schwenk, J. M.; Fagerberg, L.; Sivertsson, Å., The human secretome. *Science signaling* **2019,** *12* (609).
88. Amarasinghe, S. L.; Su, S.; Dong, X.; Zappia, L.; Ritchie, M. E.; Gouil, Q., Opportunities and challenges in long-read sequencing data analysis. *Genome biology* **2020,** *21* (1), 30.
89. Neely, B. A.; Prager, K. C.; Bland, A. M.; Fontaine, C.; Gulland, F. M.; Janech, M. G., Proteomic Analysis of Urine from California Sea Lions (Zalophus californianus): A Resource for Urinary Biomarker Discovery. *J Proteome Res* **2018,** *17* (9), 3281-3291.
90. Kong, A. T.; Leprevost, F. V.; Avtonomov, D. M.; Mellacheruvu, D.; Nesvizhskii, A. I., MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nature methods* **2017,** *14* (5), 513-520.
91. Bittremieux, W.; Laukens, K.; Noble, W. S., Extremely Fast and Accurate Open Modification Spectral Library Searching of High-Resolution Mass Spectra Using Feature Hashing and Graphics Processing Units. *J Proteome Res* **2019,** *18* (10), 3792-3799.
92. Burke, M. C.; Mirokhin, Y. A.; Tchekhovskoi, D. V.; Markey, S. P.; Heidbrink Thompson, J.; Larkin, C.; Stein, S. E., The Hybrid Search: A Mass Spectral Library Search Method for Discovery of Modifications in Proteomics. *J Proteome Res* **2017,** *16* (5), 1924-1935.
93. Devabhaktuni, A.; Lin, S.; Zhang, L.; Swaminathan, K.; Gonzalez, C. G.; Olsson, N.; Pearlman, S. M.; Rawson, K.; Elias, J. E., TagGraph reveals vast protein modification landscapes from large tandem mass spectrometry datasets. *Nature biotechnology* **2019,** *37* (4), 469-479.
94. Lee, J.-Y.; Mitchell, H. D.; Burnet, M. C.; Jenson, S. C.; Merkley, E. D.; Shukla, A. K.; Nakayasu, E. S.; Payne, S. H., Proteomics of natural bacterial isolates powered by deep learning-based de novo identification. *bioRxiv* **2018**, 428334.
95. Ma, B.; Zhang, K.; Hendrie, C.; Liang, C.; Li, M.; Doherty-Kirby, A.; Lajoie, G., PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid communications in mass spectrometry : RCM* **2003,** *17* (20), 2337-42.
96. Muth, T.; Weilnbock, L.; Rapp, E.; Huber, C. G.; Martens, L.; Vaudel, M.; Barsnes, H., DeNovoGUI: an open source graphical user interface for de novo sequencing of tandem mass spectra. *J Proteome Res* **2014,** *13* (2), 1143-6.
97. Kafader, J. O.; Melani, R. D.; Durbin, K. R.; Ikwuagwu, B.; Early, B. P.; Fellers, R. T.; Beu, S. C.; Zabrouskov, V.; Makarov, A. A.; Maze, J. T.; Shinholt, D. L.; Yip, P. F.; Tullman-Ercek, D.; Senko, M. W.; Compton, P. D.; Kelleher, N. L., Multiplexed mass spectrometry of individual ions improves measurement of proteoforms and their complexes. *Nature methods* **2020**, *17*, 391–394.

98. Rose, R. J.; Damoc, E.; Denisov, E.; Makarov, A.; Heck, A. J., High-sensitivity Orbitrap mass analysis of intact macromolecular assemblies. *Nature methods* **2012,** *9* (11), 1084-6.
99. Snijder, J.; Schuller, J. M.; Wiegard, A.; Lossl, P.; Schmelling, N.; Axmann, I. M.; Plitzko, J. M.; Forster, F.; Heck, A. J., Structures of the cyanobacterial circadian oscillator frozen in a fully assembled state. *Science (New York, N.Y.)* **2017,** *355* (6330), 1181-1184.
100. Wang, Z.; Li, Y.; Hou, B.; Pronobis, M. I.; Wang, M.; Wang, Y.; Cheng, G.; Weng, W.; Wang, Y.; Tang, Y.; Xu, X.; Pan, R.; Lin, F.; Wang, N.; Chen, Z.; Wang, S.; Ma, L. Z.; Li, Y.; Huang, D.; Jiang, L.; Wang, Z.; Zeng, W.; Zhang, Y.; Du, X.; Lin, Y.; Li, Z.; Xia, Q.; Geng, J.; Dai, H.; Yu, Y.; Zhao, X. D.; Yuan, Z.; Yan, J.; Nie, Q.; Zhang, X.; Wang, K.; Chen, F.; Zhang, Q.; Zhu, Y.; Zheng, S.; Poss, K. D.; Tao, S. C.; Meng, X., An array of 60,000 antibodies for proteome-scale antibody generation and target discovery. *Sci Adv* **2020,** *6* (11), eaax2271.
101. Loffredo, F. S.; Steinhauser, M. L.; Jay, S. M.; Gannon, J.; Pancoast, J. R.; Yalamanchi, P.; Sinha, M.; Dall'Osso, C.; Khong, D.; Shadrach, J. L.; Miller, C. M.; Singer, B. S.; Stewart, A.; Psychogios, N.; Gerszten, R. E.; Hartigan, A. J.; Kim, M. J.; Serwold, T.; Wagers, A. J.; Lee, R. T., Growth differentiation factor 11 is a circulating factor that reverses age-related cardiac hypertrophy. *Cell* **2013,** *153* (4), 828-39.
102. Löfgren, M.; Svala, E.; Lindahl, A.; Skiöldebrand, E.; Ekman, S., Time-dependent changes in gene expression induced in vitro by interleukin-1β in equine articular cartilage. *Research in veterinary science* **2018,** *118*, 466-476.
103. Neely, B.; Carlin, K.; Arthur, J.; McFee, W.; Janech, M., Ratiometric Measurements of Adiponectin by Mass Spectrometry in Bottlenose Dolphins (*Tursiops truncatus*) with Iron Overload Reveal an Association with Insulin Resistance and Glucagon. *Frontiers in endocrinology* **2013,** *4* (132), 1-13.
104. Pinheiro, P. V.; Ghanim, M.; Alexander, M.; Rebelo, A. R.; Santos, R. S.; Orsburn, B. C.; Gray, S.; Cilia, M., Host Plants Indirectly Influence Plant Virus Transmission by Altering Gut Cysteine Protease Activity of Aphid Vectors. *Molecular & cellular proteomics : MCP* **2017,** *16* (4 suppl 1), S230-s243.
105. Wilmarth, P. PAW_BLAST. https://github.com/pwilmart/PAW_BLAST.
106. Naithani, S.; Preece, J.; D'Eustachio, P.; Gupta, P.; Amarasinghe, V.; Dharmawardhana, P. D.; Wu, G.; Fabregat, A.; Elser, J. L.; Weiser, J.; Keays, M.; Fuentes, A. M.; Petryszak, R.; Stein, L. D.; Ware, D.; Jaiswal, P., Plant Reactome: a resource for plant pathways and comparative analysis. *Nucleic acids research* **2017,** *45* (D1), D1029-d1039.
107. Hansen, B. O.; Olsson, S., OmicsDB::Pathogens - A database for exploring functional networks of plant pathogens. *bioRxiv* **2020**, 2020.03.18.979971.
108. Pfister, N.; Williams, E. G.; Peters, J.; Aebersold, R.; Bühlmann, P., Stabilizing Variable Selection and Regression. *arXiv preprint arXiv:1911.01850* **2019**.
109. Gygi, S. P.; Rochon, Y.; Franza, B. R.; Aebersold, R., Correlation between protein and mRNA abundance in yeast. *Molecular and cellular biology* **1999,** *19* (3), 1720-30.
110. Foss, E. J.; Radulovic, D.; Shaffer, S. A.; Ruderfer, D. M.; Bedalov, A.; Goodlett, D. R.; Kruglyak, L., Genetic basis of proteome variation in yeast. *Nature genetics* **2007,** *39* (11), 1369-75.
111. Edfors, F.; Danielsson, F.; Hallstrom, B. M.; Kall, L.; Lundberg, E.; Ponten, F.; Forsstrom, B.; Uhlen, M., Gene-specific correlation of RNA and protein levels in human cells and tissues. *Molecular systems biology* **2016,** *12* (10), 883.
112. Liu, Y.; Beyer, A.; Aebersold, R., On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell* **2016,** *165* (3), 535-50.
113. Nusinow, D. P.; Szpyt, J.; Ghandi, M.; Rose, C. M.; McDonald, E. R., 3rd; Kalocsay, M.; Jane-Valbuena, J.; Gelfand, E.; Schweppe, D. K.; Jedrychowski, M.; Golji, J.; Porter, D. A.; Rejtar, T.; Wang, Y. K.; Kryukov, G. V.; Stegmeier, F.; Erickson, B. K.; Garraway, L. A.; Sellers, W. R.; Gygi, S. P., Quantitative Proteomics of the Cancer Cell Line Encyclopedia. *Cell* **2020,** *180* (2), 387-402.e16.

114. Salovska, B.; Zhu, H.; Gandhi, T.; Frank, M.; Li, W.; Rosenberger, G.; Wu, C.; Germain, P. L.; Zhou, H.; Hodny, Z.; Reiter, L.; Liu, Y., Isoform-resolved correlation analysis between mRNA abundance regulation and protein level degradation. *Molecular systems biology* **2020,** *16* (3), e9170.
115. Buccitelli, C.; Selbach, M., mRNAs, proteins and the emerging principles of gene expression control. *Nature reviews. Genetics* **2020**, *Online ahead of print*.
116. Lehallier, B.; Gate, D.; Schaum, N.; Nanasi, T.; Lee, S. E.; Yousef, H.; Moran Losada, P.; Berdnik, D.; Keller, A.; Verghese, J.; Sathyan, S.; Franceschi, C.; Milman, S.; Barzilai, N.; Wyss-Coray, T., Undulating changes in human plasma proteome profiles across the lifespan. *Nature medicine* **2019,** *25* (12), 1843-1850.
117. Kelmer Sacramento, E.; Kirkpatrick, J. M.; Mazzetto, M.; Baumgart, M.; Bartolome, A.; Di Sanzo, S.; Caterino, C.; Sanguanini, M.; Papaevgeniou, N.; Lefaki, M.; Childs, D.; Bagnoli, S.; Terzibasi Tozzini, E.; Di Fraia, D.; Romanov, N.; Sudmant, P. H.; Huber, W.; Chondrogianni, N.; Vendruscolo, M.; Cellerino, A.; Ori, A., Reduced proteasome activity in the aging brain results in ribosome stoichiometry loss and aggregation. *Molecular systems biology* **2020,** *16* (6), e9596.
118. Boschiero, C.; Dai, X.; Lundquist, P. K.; Roy, S.; de Bang, T.; Zhang, S.; Zhuang, Z.; Torres-Jerez, I.; Udvardi, M. K.; Scheible, W. R.; Zhao, P. X., MtSSPdb: the Medicago truncatula Small Secreted Peptide Database. *Plant Physiol* **2020**, *183 (1)*, 399-413.
119. Fleites, L. A.; Johnson, R.; Kruse, A. R.; Nachman, R. J.; Hall, D. G.; MacCoss, M.; Heck, M. L., Peptidomics Approaches for the Identification of Bioactive Molecules from Diaphorina citri. *J Proteome Res* **2020**, *19 (4)*, 1392-1408.
120. Zheng, Y.; Sweet, S. M.; Popovic, R.; Martinez-Garcia, E.; Tipton, J. D.; Thomas, P. M.; Licht, J. D.; Kelleher, N. L., Total kinetic analysis reveals how combinatorial methylation patterns are established on lysines 27 and 36 of histone H3. *Proc Natl Acad Sci U S A* **2012,** *109* (34), 13549-54.
121. Smith, L. M.; Kelleher, N. L.; Consortium for Top Down, P., Proteoform: a single term describing protein complexity. *Nat Methods* **2013,** *10* (3), 186-7.
122. Zheng, Y.; Huang, X.; Kelleher, N. L., Epiproteomics: quantitative analysis of histone marks and codes by mass spectrometry. *Curr Opin Chem Biol* **2016,** *33*, 142-50.
123. Malaney, P.; Uversky, V. N.; Dave, V., PTEN proteoforms in biology and disease. *Cell Mol Life Sci* **2017,** *74* (15), 2783-2794.
124. Cristea, I. M.; Williams, R.; Chait, B. T.; Rout, M. P., Fluorescent proteins as proteomic probes. *Mol Cell Proteomics* **2005,** *4* (12), 1933-41.
125. Cristea, I. M.; Carroll, J. W.; Rout, M. P.; Rice, C. M.; Chait, B. T.; MacDonald, M. R., Tracking and elucidating alphavirus-host protein interactions. *J Biol Chem* **2006,** *281* (40), 30269-78.
126. Moorman, N. J.; Sharon-Friling, R.; Shenk, T.; Cristea, I. M., A targeted spatial-temporal proteomics approach implicates multiple cellular trafficking pathways in human cytomegalovirus virion maturation. *Mol Cell Proteomics* **2010,** *9* (5), 851-60.
127. Cristea, I. M.; Chait, B. T. Affinity purification of protein complexes *Cold Spring Harb Protoc* [Online], 2011. http://www.ncbi.nlm.nih.gov/pubmed/21536765.
128. Miteva, Y. V.; Budayeva, H. G.; Cristea, I. M., Proteomics-based methods for discovery, quantification, and validation of protein-protein interactions. *Anal Chem* **2013,** *85* (2), 749-68.
129. Weisbrod, C. R.; Chavez, J. D.; Eng, J. K.; Yang, L.; Zheng, C.; Bruce, J. E., In Vivo Protein Interaction Network Identified with a Novel Real-Time Cross-Linked Peptide Identification Strategy. *J Proteome Res* **2013**, *12* (4), 1569-1579.
130. Zheng, C.; Weisbrod, C. R.; Chavez, J. D.; Eng, J. K.; Sharma, V.; Wu, X.; Bruce, J. E., XLink-DB: Database and Software Tools for Storing and Visualizing Protein Interaction Topology Data. *J Proteome Res* **2013**, *12* (4), 1989–1995.
131. Chavez, J. D.; Eng, J. K.; Schweppe, D. K.; Cilia, M.; Rivera, K.; Zhong, X.; Wu, X.; Allen, T.; Khurgel, M.; Kumar, A.; Lampropoulos, A.; Larsson, M.; Maity, S.; Morozov, Y.;

Pathmasiri, W.; Perez-Neut, M.; Pineyro-Ruiz, C.; Polina, E.; Post, S.; Rider, M.; Tokmina-Roszyk, D.; Tyson, K.; Vieira Parrine Sant'Ana, D.; Bruce, J. E., A General Method for Targeted Quantitative Cross-Linking Mass Spectrometry. *PLoS One* **2016,** *11* (12), e0167547.
132. Ramsey, J. S.; Chavez, J. D.; Johnson, R.; Hosseinzadeh, S.; Mahoney, J. E.; Mohr, J. P.; Robison, F.; Zhong, X.; Hall, D. G.; MacCoss, M.; Bruce, J.; Cilia, M., Protein interaction networks at the host-microbe interface in Diaphorina citri, the insect vector of the citrus greening pathogen. *R Soc Open Sci* **2017,** *4* (2), 160545.
133. Kuriyan, J.; Eisenberg, D., The origin of protein interactions and allostery in colocalization. *Nature* **2007,** *450* (7172), 983-90.
134. Cilia, M.; Tamborindeguy, C.; Fish, T.; Howe, K.; Thannhauser, T. W.; Gray, S., Genetics coupled to quantitative intact proteomics links heritable aphid and endosymbiont protein expression to circulative polerovirus transmission. *J Virol* **2011,** *85* (5), 2148-66.
135. Mertins, P.; Mani, D. R.; Ruggles, K. V.; Gillette, M. A.; Clauser, K. R.; Wang, P.; Wang, X.; Qiao, J. W.; Cao, S.; Petralia, F.; Kawaler, E.; Mundt, F.; Krug, K.; Tu, Z.; Lei, J. T.; Gatza, M. L.; Wilkerson, M.; Perou, C. M.; Yellapantula, V.; Huang, K. L.; Lin, C.; McLellan, M. D.; Yan, P.; Davies, S. R.; Townsend, R. R.; Skates, S. J.; Wang, J.; Zhang, B.; Kinsinger, C. R.; Mesri, M.; Rodriguez, H.; Ding, L.; Paulovich, A. G.; Fenyo, D.; Ellis, M. J.; Carr, S. A.; Nci, C., Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* **2016,** *534* (7605), 55-62.
136. Cilia, M.; Howe, K.; Fish, T.; Smith, D.; Mahoney, J.; Tamborindeguy, C.; Burd, J.; Thannhauser, T. W.; Gray, S., Biomarker discovery from the top down: Protein biomarkers for efficient virus transmission by insects (Homoptera: Aphididae) discovered by coupling genetics and 2-D DIGE. *Proteomics* **2011,** *11* (12), 2440-58.
137. Poliakov, A.; Russell, C. W.; Ponnala, L.; Hoops, H. J.; Sun, Q.; Douglas, A. E.; van Wijk, K. J., Large-scale label-free quantitative proteomics of the pea aphid-Buchnera symbiosis. *Mol Cell Proteomics* **2011,** *10* (6), M110 007039.
138. Ramsey, J. S.; Johnson, R. S.; Hoki, J. S.; Kruse, A.; Mahoney, J.; Hilf, M. E.; Hunter, W. B.; Hall, D. G.; Schroeder, F. C.; MacCoss, M. J.; Cilia, M., Metabolic Interplay between the Asian Citrus Psyllid and Its Profftella Symbiont: An Achilles' Heel of the Citrus Greening Insect Vector. *PLoS One* **2015,** *10* (11), e0140826.
139. Kruse, A.; Fattah-Hosseini, S.; Saha, S.; Johnson, R.; Warwick, E.; Sturgeon, K.; Mueller, L.; MacCoss, M. J.; Shatters, R. G., Jr.; Cilia Heck, M., Combining 'omics and microscopy to visualize interactions between the Asian citrus psyllid vector and the Huanglongbing pathogen Candidatus Liberibacter asiaticus in the insect gut. *PLoS One* **2017,** *12* (6), e0179531.
140. Müller, J. B.; Geyer, P. E.; Colaço, A. R.; Treit, P. V.; Strauss, M. T.; Oroshi, M.; Doll, S.; Virreira Winter, S.; Bader, J. M.; Köhler, N.; Theis, F.; Santos, A.; Mann, M., The proteome landscape of the kingdoms of life. *Nature* **2020,** *582*, 592–596.
141. Janech, M. G.; Fitzgibbon, W. R.; Chen, R.; Nowak, M. W.; Miller, D. H.; Paul, R. V.; Ploth, D. W., Molecular and functional characterization of a urea transporter from the kidney of the Atlantic stingray. *American journal of physiology. Renal physiology* **2003,** *284* (5), F996-f1005.
142. Elger, M.; Hentschel, H.; Litteral, J.; Wellner, M.; Kirsch, T.; Luft, F. C.; Haller, H., Nephrogenesis is induced by partial nephrectomy in the elasmobranch Leucoraja erinacea. *Journal of the American Society of Nephrology : JASN* **2003,** *14* (6), 1506-18.

For TOC Only


MS Data
Search Space
Analysis
Disease Reservoirs
Innate Immunity
Cancer
Eusociality
Longevity
Metabolism
Disease Transmission and Resistance
Non-model Proteomics
I/R Injury Resistance

Supporting Information for Proteomics in non-model organisms: a new analytical frontier

Michelle Heck[1,2,3] and Benjamin A. Neely[4*]

[1] Emerging Pests and Pathogens Research Unit, USDA Agricultural Research Service, Ithaca, NY, USA
[2] Plant Pathology and Plant Microbe Biology Section, School of Integrative Plant Science, Cornell University, Ithaca, NY, USA
[3] Boyce Thompson Institute, Ithaca, NY, USA
[4] Chemical Sciences Division, National Institute of Standards and Technology, Charleston, SC, USA

Table of Contents

Supplementary Table S1. Specific NCBI taxon IDs and their designation for 23 mammal groups used in Figures 2 and 3. The number of genome assemblies on Genbank and genome annotations on RefSeq (valid as of 12 March, 2020) with published proteomic data sets on the Proteomics Identifications Database (PRIDE; as of 17 January 2020) are listed as well.

taxon ID	mammals	proteomic data sets	genome assemblies	genome annotations
91561	even-toed ungulates	32	120	26
9989	rodents	28	93	25
33554	carnivores	25	60	23
9443	primates	20	63	28
9787	odd-toed ungulates	11	12	4
9979	rabbits & hares	5	3	2
948950	anteaters and sloths	5	5	0
9397	bats	4	37	12
9779	elephants	4	1	1
38609	Herbivorous marsupials	3	3	2
28734	elephant shrews	2	1	1
9815	aardvarks	1	2	1
9774	manateees and dugongs	1	1	1
38608	carnivorous marsupials	1	2	1
38605	possums	1	1	1
948951	armadillos	1	3	1
9255	egg-laying mammals	0	1	1
9362	insectivores	0	8	3
9389	golden moles	0	1	1
9369	tenrecs	0	2	1
30656	flying lemurs	0	1	1
9392	tree shrews	0	3	1
9971	pangolins	0	3	1