Proteomics and You
(and other tales of analytical biochemistry)

October 5, 2020
Cal Poly CS Guest Lecture

Ben Neely
Biochemical and Exposure Sciences Group
NIST Charleston

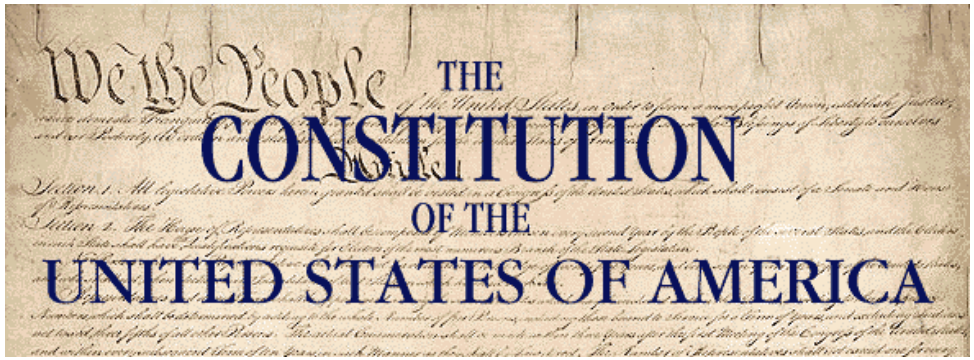
Identification of certain commercial equipment, instruments, software or materials does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products identified are necessarily the best available for the purpose.

National Institute of Standards & Technology

The History of Standards

"Uniformity in the currency, weights, and measures of the United States is an object of great importance, and will, I am persuaded, be duly attended to."

George Washington, State of the Union Address, 1790



Article I, Section 8: "The Congress shall have the power to…*fix the standard of weights and measures*"

The History of Standards

"When you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind."

Lord Kelvin (William Thomson), Lecture May 1883

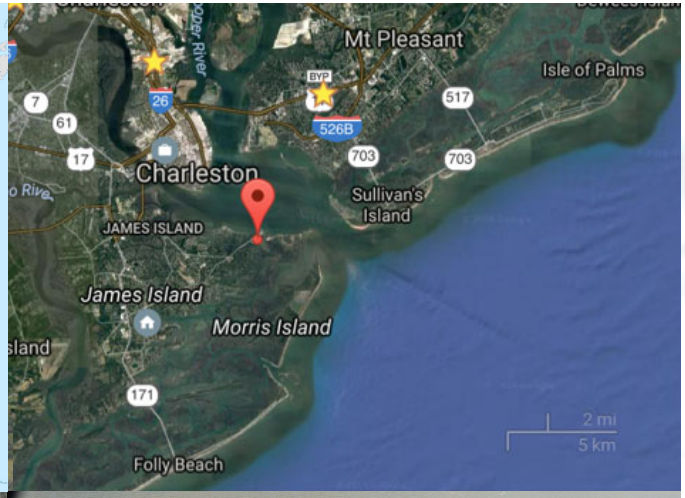
Basic Stats and Facts

Gaithersburg

Boulder

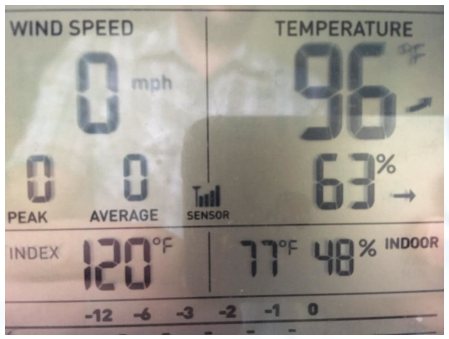
Collaborations:
JILA, JQI, HML,
IBBR, ChiMaD,
NCCoE, Stanford

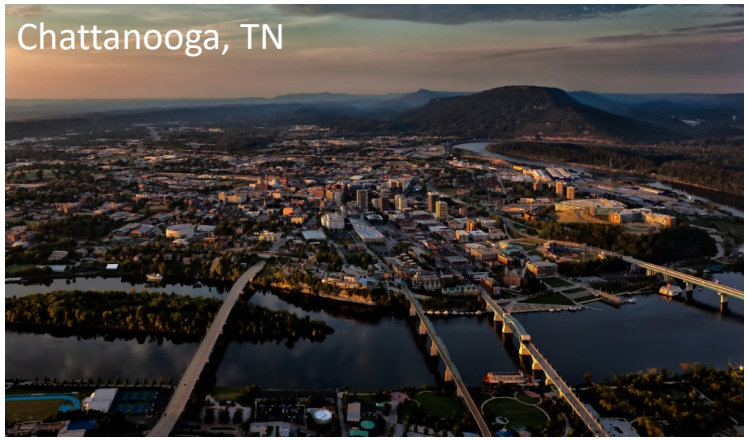

Charleston, SC


Mt Pleasant
Isle of Palms
Charleston
JAMES ISLAND
Sullivan's Island
James Island
Morris Island
Folly Beach
2 mi
5 km

Why do we need standards?

Horses can't be out if above 95°F or heat index ≥ 110°F (formerly was 125°F)





7 Bridges Marathon (2017)
Full was 0.63 miles long
Half was 0.63 miles short
Human error

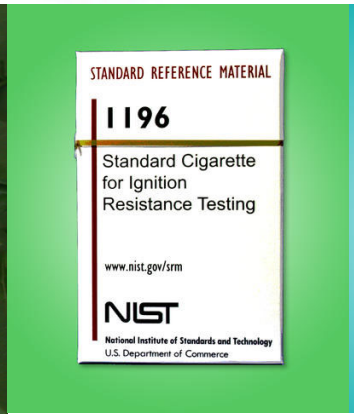
Why do we need standards?

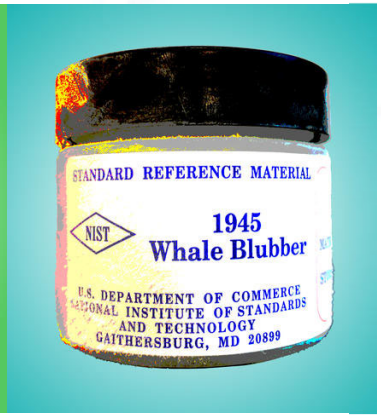
October 2017, admitted to falsifying data about the strength and durability of some copper and aluminum that was used in cars and trains and possibly planes and a space rocket, too. AND faked data about iron ore powder and materials used in DVDs and LCD screens.

Aaron Franklin calibrates weekly at minimum to keep meat cooking at 235.

National Institute of Standards & Technology

Report of Investigation





~~Proteomics?~~

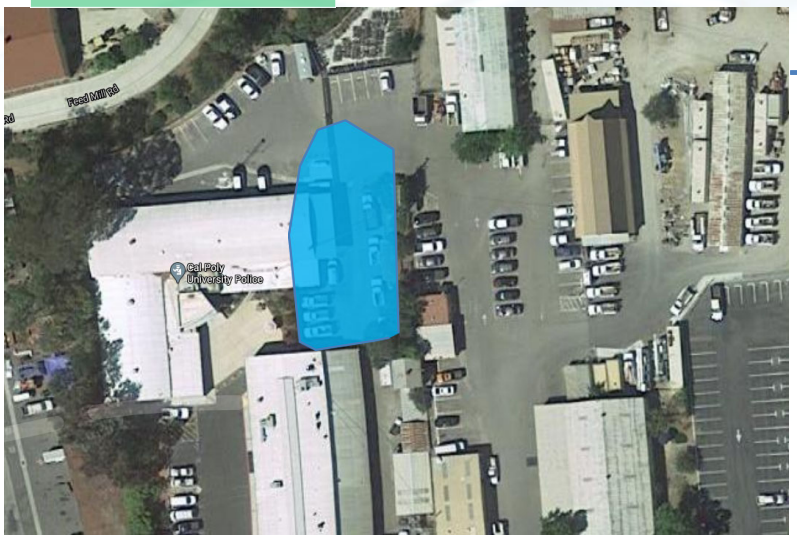
Mass Spectrometry?

Top-Down

car	weight
	1,852.1 lb
	1,807.8 lb
	2,602.5 lb

Mass Spectrometry and Proteomics

Imaging



- Determine different makes and models of cars (colors, type of tires, *etc.*).
- Quantify the number of each car type.

↓

- Can use the number of certain kinds of cars to predict the day of the week.
- Can use the composition of cars to infer parking lot type.
  - Ex. There are more police cars than we would expect, so it is probably a police station.

Bottom-up Shotgun Proteomics:


42.5 lb


7.4 lb


0.1 lb

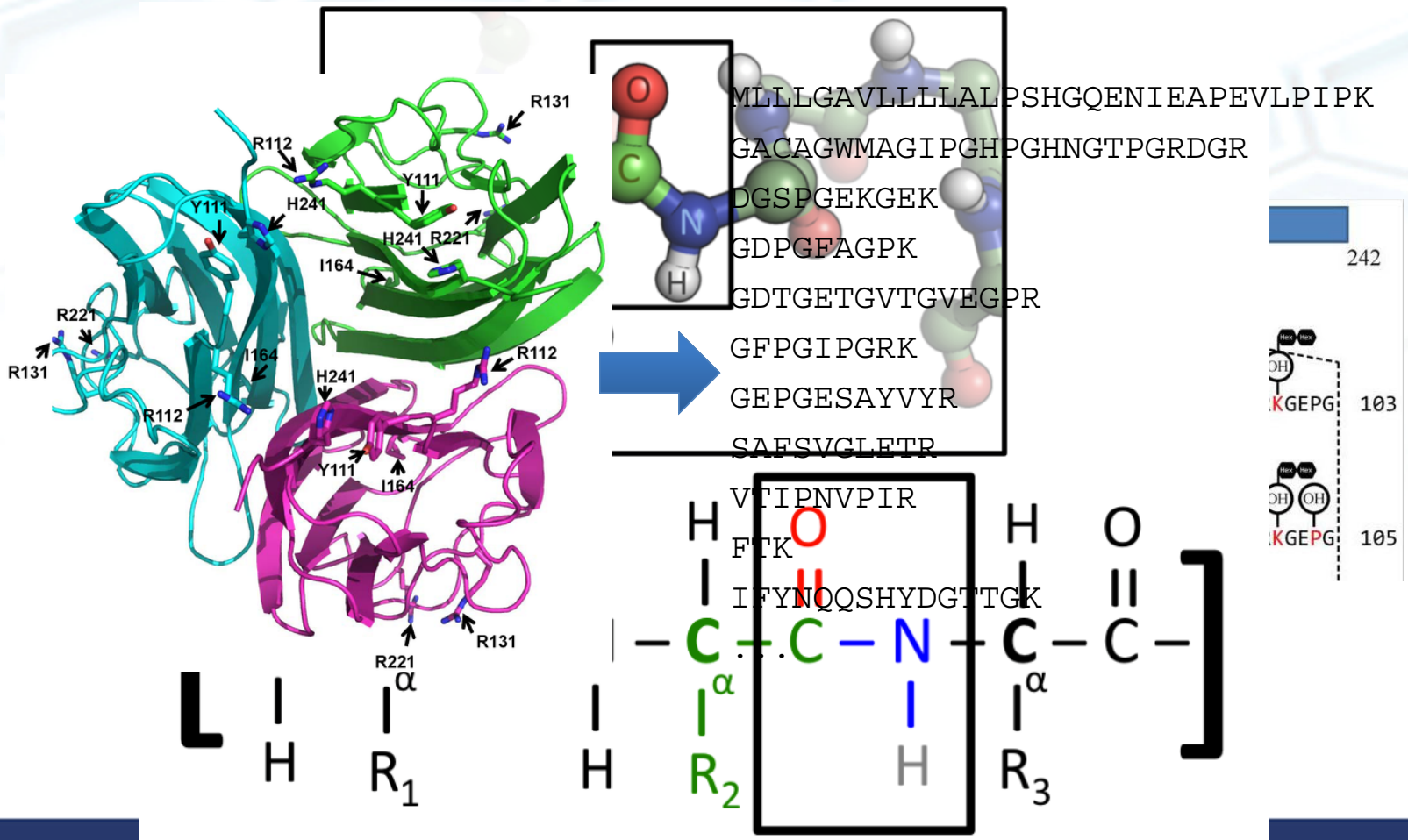
=

>95% confidence

* can identify anything that is known

Shotgun Proteomics 101

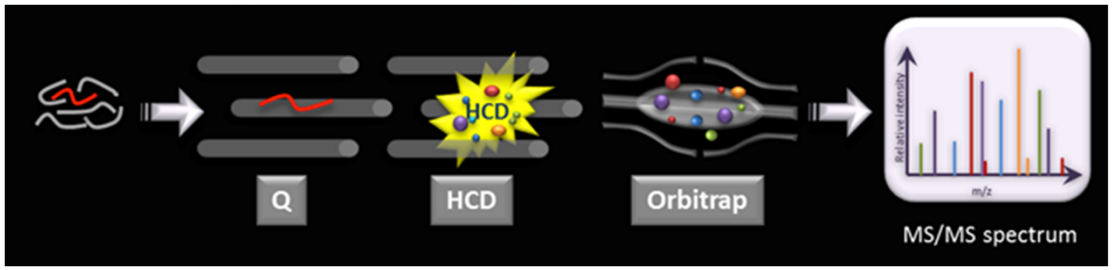
What is a protein? a peptide?



Shotgun Proteomics 101

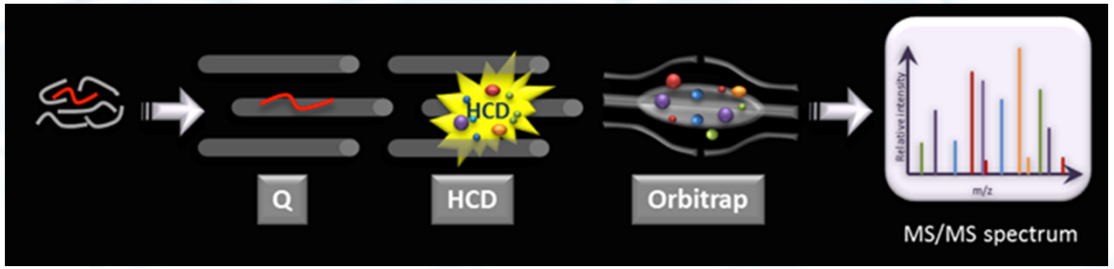
Digesting one protein and measuring one peptide (ex. IFYNQQSHYDGTTGK)

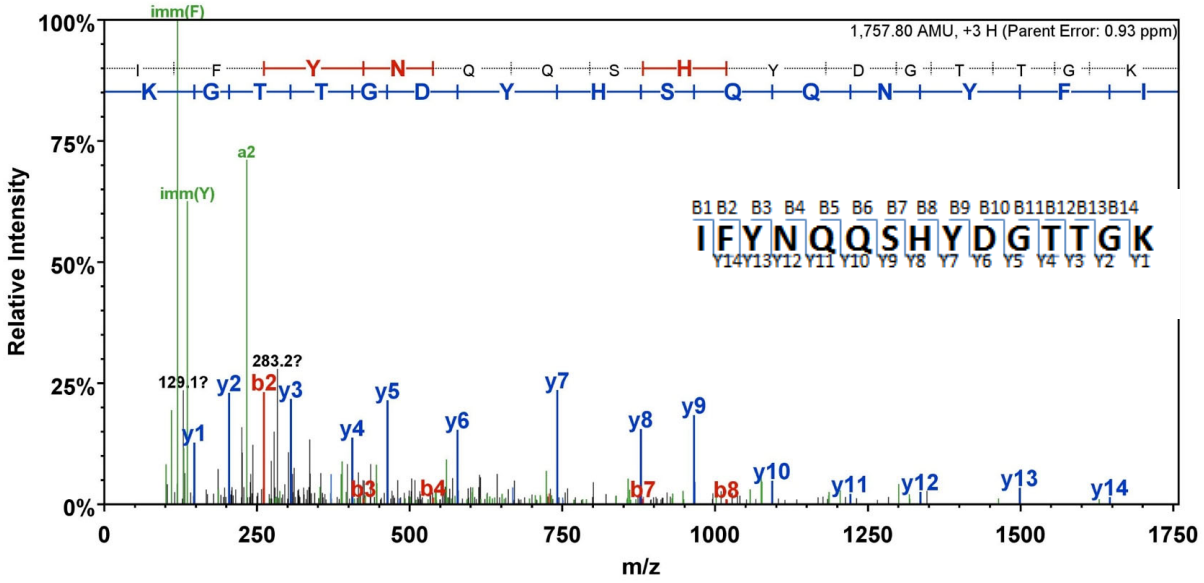
Digestion



Shotgun Proteomics 101

Digesting one protein and identifying one peptide (ex. IFYNQQSHYDGTTGK)

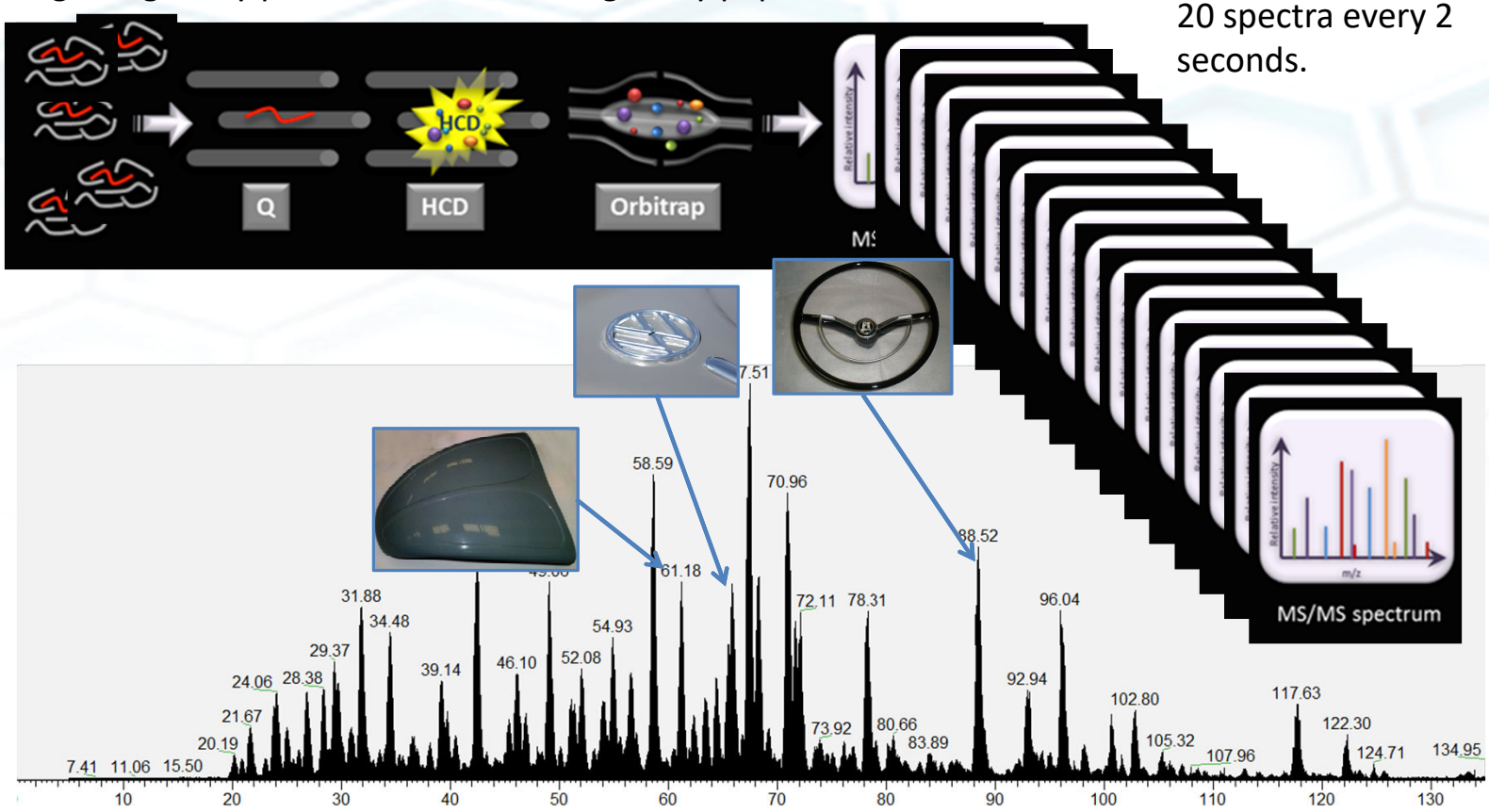




Shotgun Proteomics 101

Digesting many proteins and measuring many peptides.

Example: acquire 20 spectra every 2 seconds.



130 minute run: 70,000 spectra -> 40,000 peptides -> 7,500 proteins -> 4,000 protein groups

Shotgun Proteomics 101 - Quantification

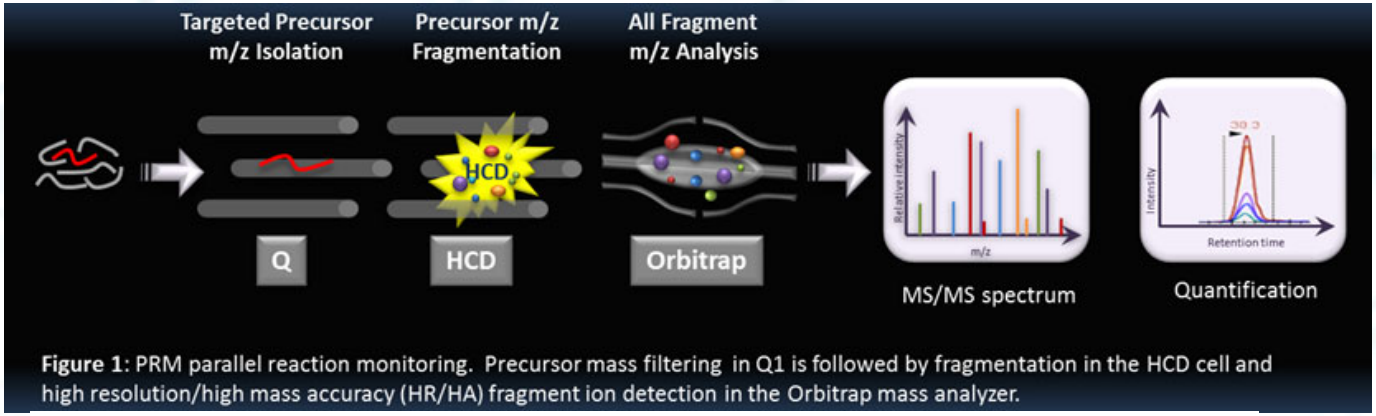
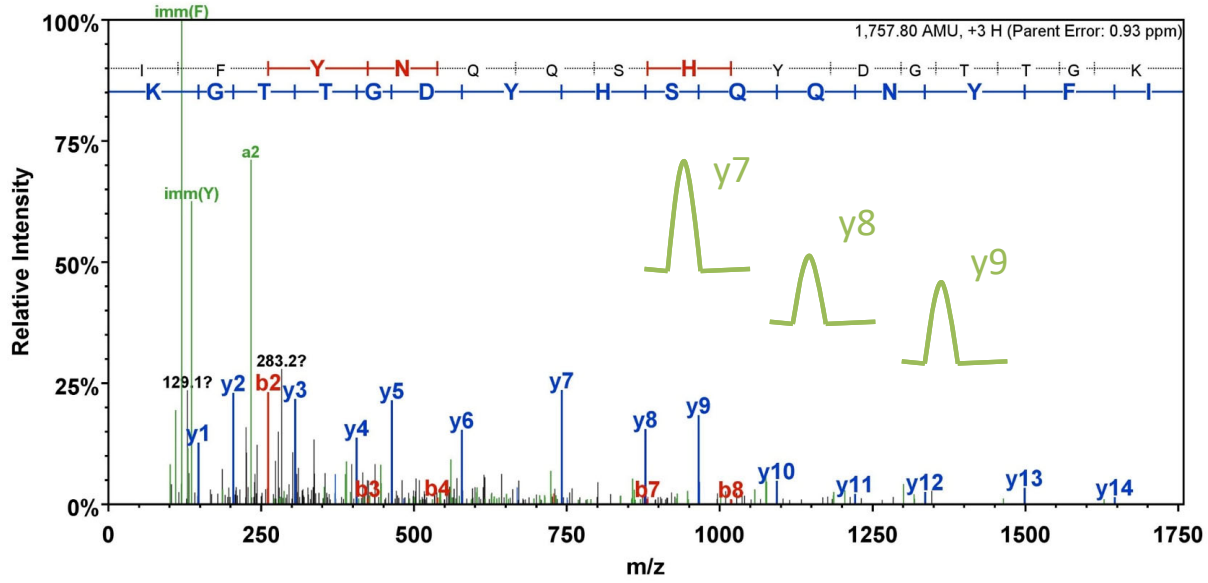
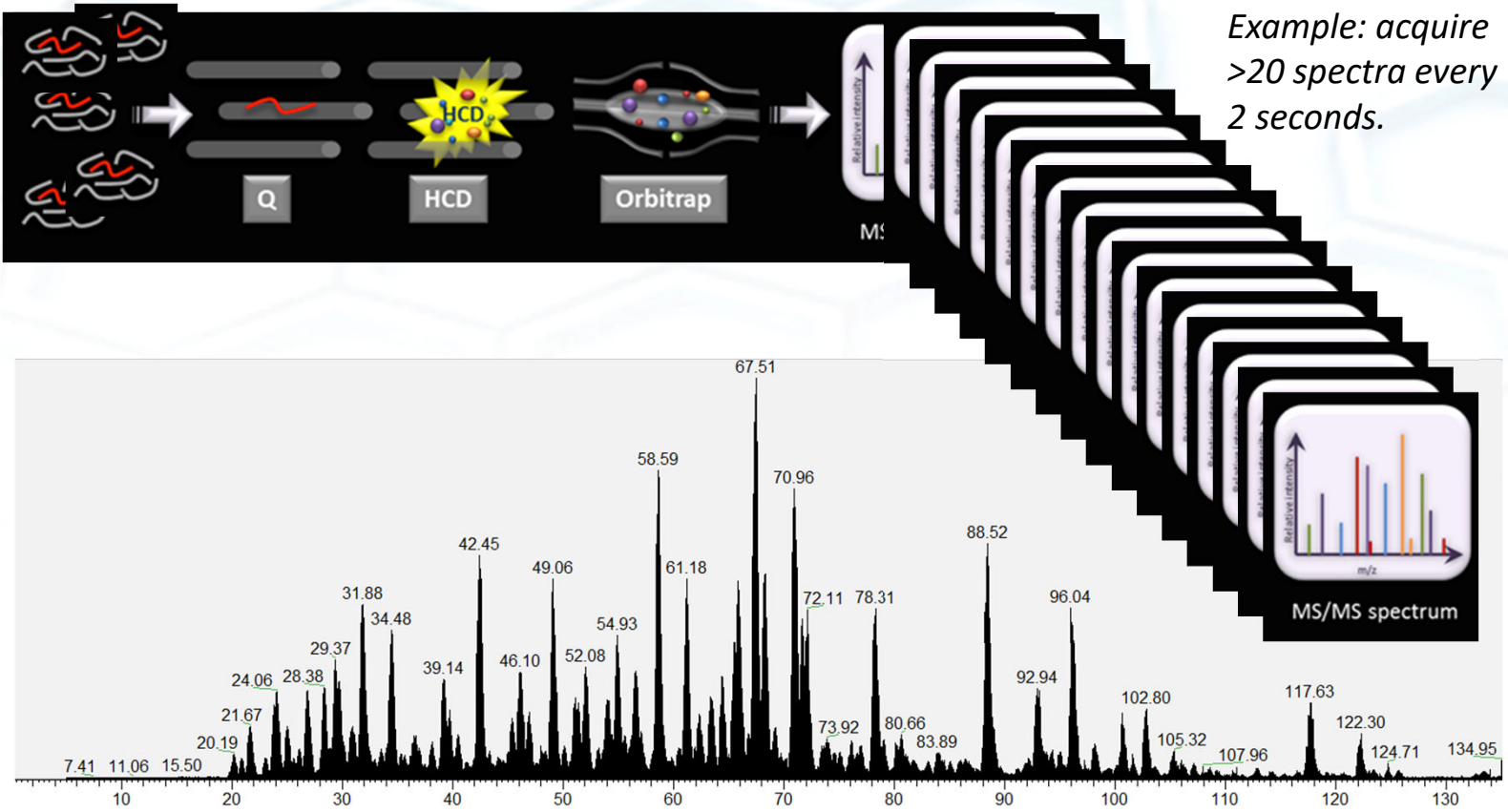


Figure 1: PRM parallel reaction monitoring. Precursor mass filtering in Q1 is followed by fragmentation in the HCD cell and high resolution/high mass accuracy (HR/HA) fragment ion detection in the Orbitrap mass analyzer.



Shotgun Proteomics 101 – no cars

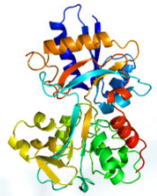
Digesting many proteins to identify many peptides and infer many proteins.



Example: acquire >20 spectra every 2 seconds.

130 minute run: 70,000 MS/MS spectra

Protein



~42 000 Protein isoforms

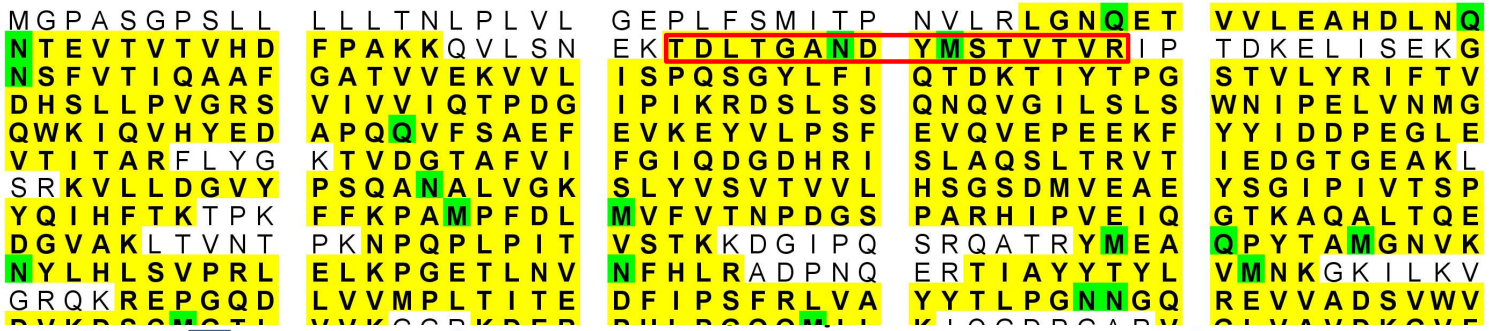
~ 2.5 million non-redundant tryptic peptides

Need a protein database (which is from an annotated genome) to make identifications*

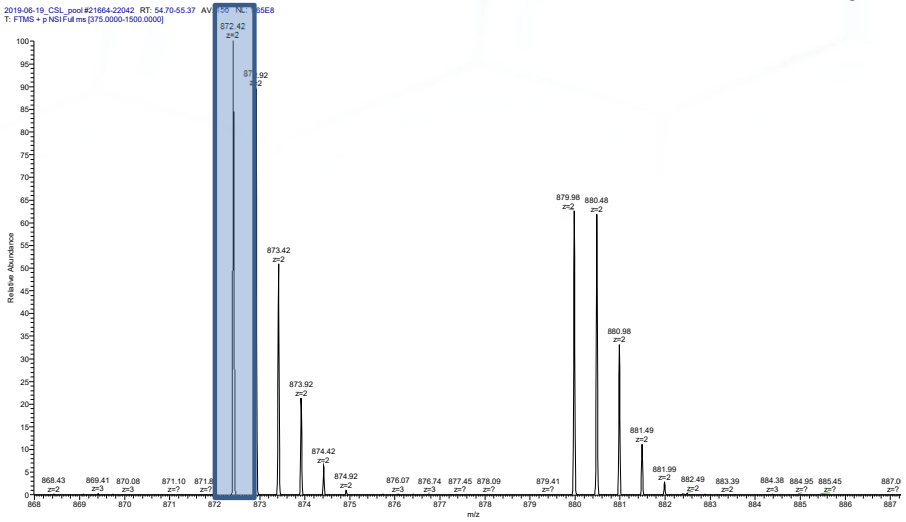
**de novo is beyond this discussion*

Shotgun Proteomics 101 – Probabilistic Pattern Matching

XP_027439774.1 (100%), 186,662.2 Da
complement C3 [Zalophus californianus]
107 exclusive unique peptides, 230 exclusive unique spectra, 371 total spectra, 1275/1661 amino acids (77% coverage)



130 minute run:
70 000 MS/MS spectra
40 000 peptides
4 000 protein



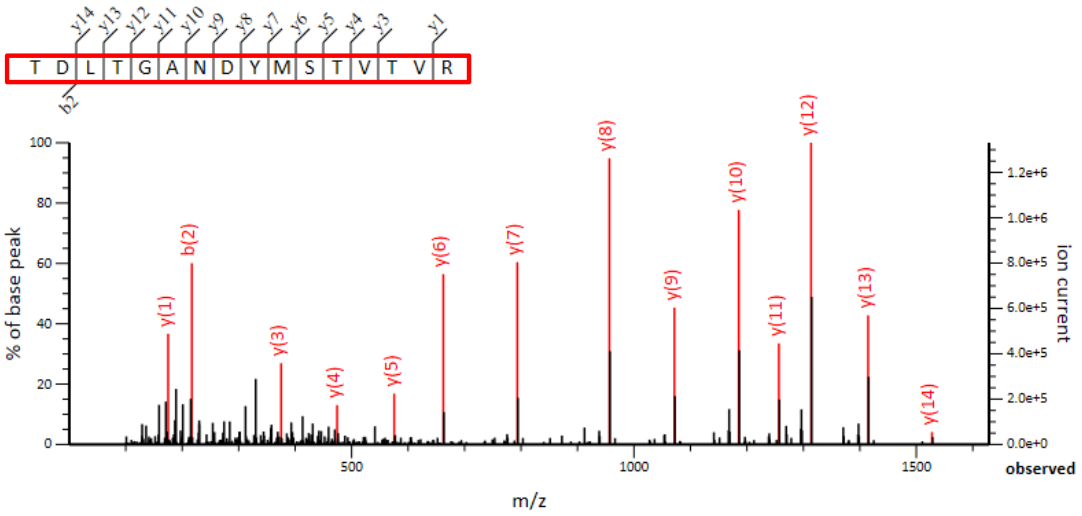
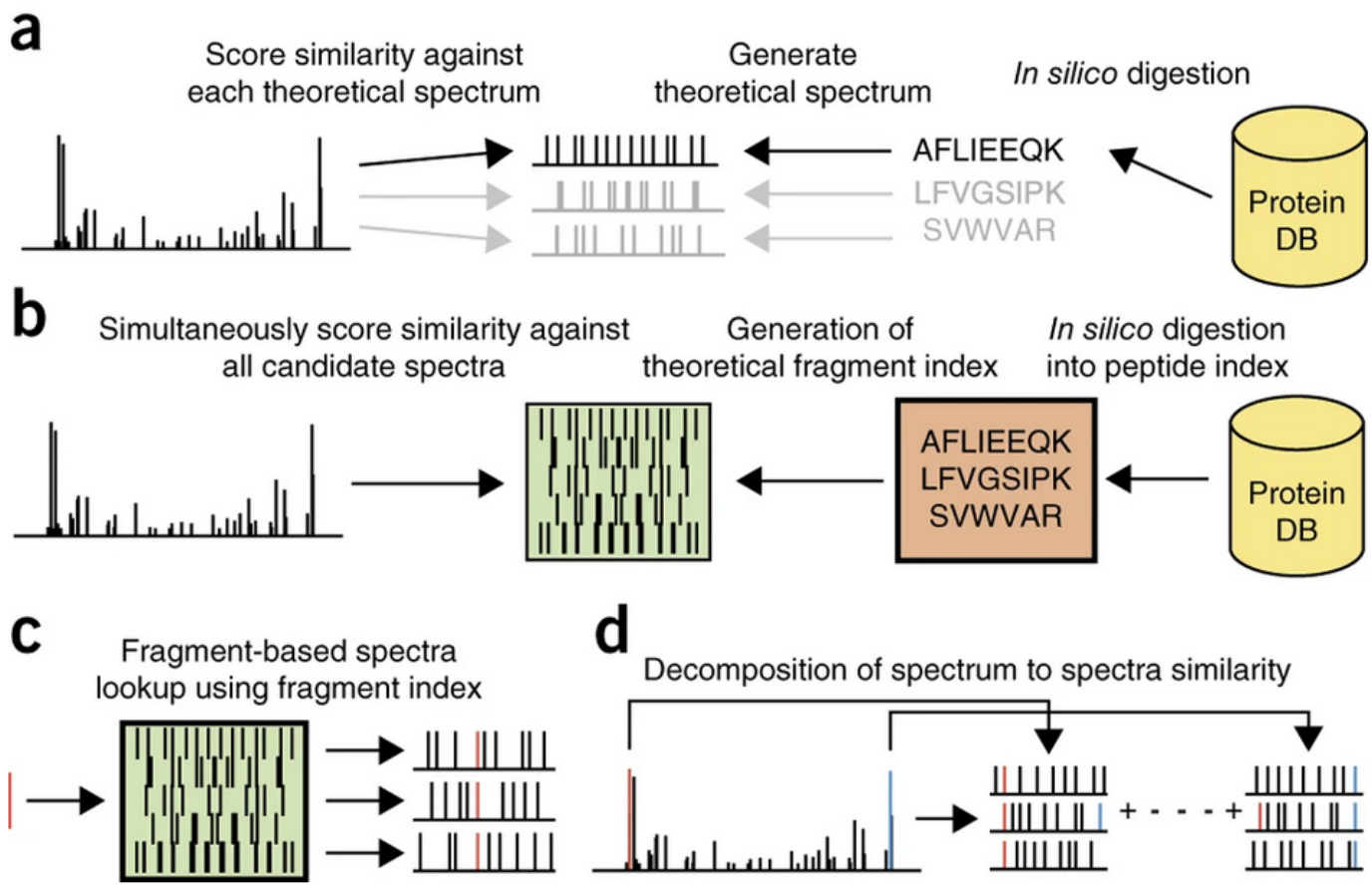
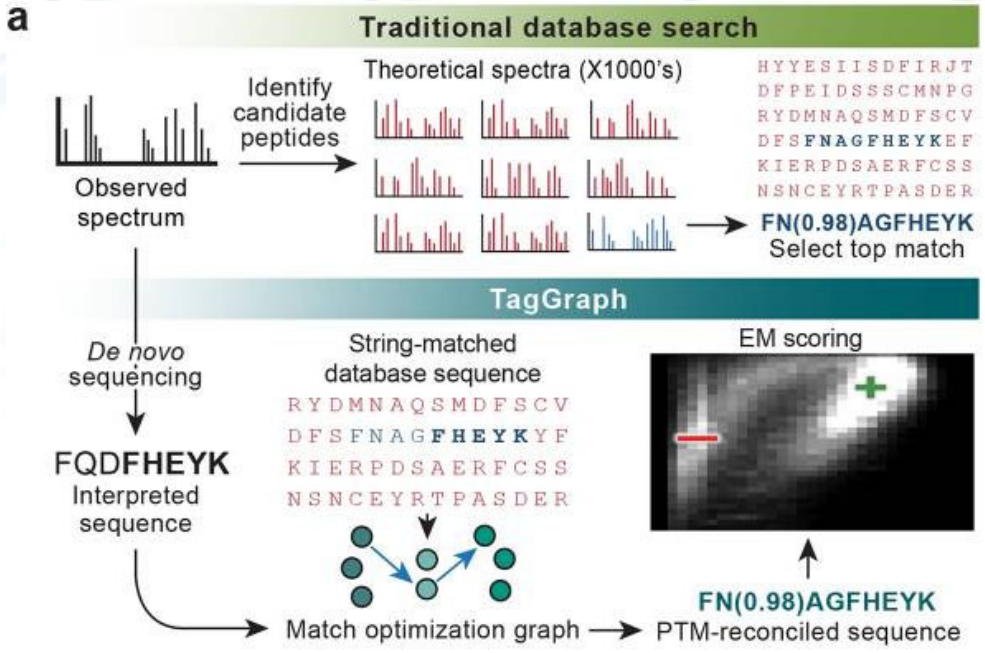


Figure 1: Database-search strategies and the MSFragger algorithm.

From: MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics



Some Super Cool Modern Techniques



nature biotechnology
ARTICLES
https://doi.org/10.1038/s41587-019-0067-5

TagGraph reveals vast protein modification landscapes from large tandem mass spectrometry datasets

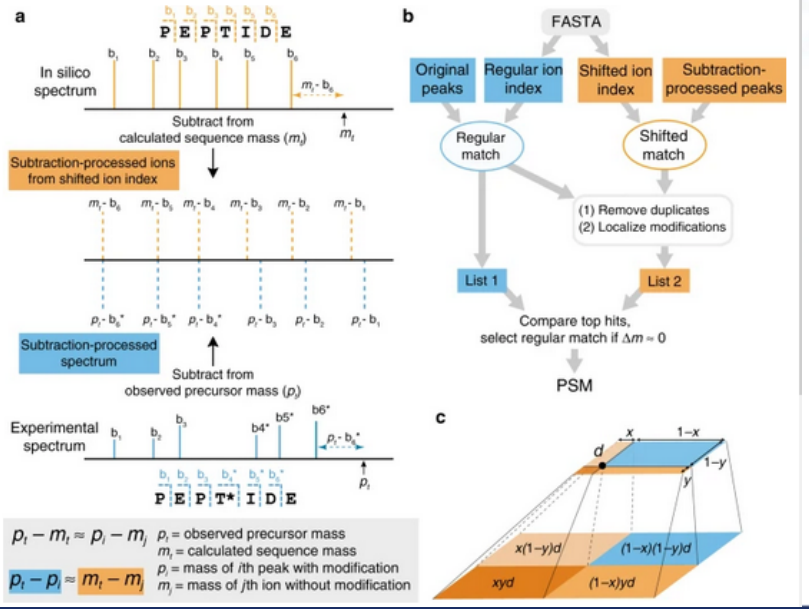
Arun Devabhaktuni, Sarah Lin, Lichao Zhang, Kavya Swaminathan, Carlos G. Gonzalez, Niclas Olsson, Samuel M. Pearlman, Keith Rawson and Joshua E. Elias*

nature COMMUNICATIONS
ARTICLE
https://doi.org/10.1038/s41467-020-17921-y OPEN

Identification of modified peptides using localization-aware open search

Fengchao Yu[1], Guo Ci Teo[1], Andy T. Kong[1], Sarah E. Haynes[1], Dmitry M. Avtonomov[1], Daniel J. Geiszler[1] & Alexey I. Nesvizhskii[1,2]✉

Fig. 1: Overview of the localization-aware open search strategy.

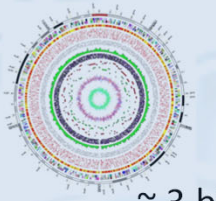


Proteomics?

~~Mass Spectrometry?~~

"without that light [of evolution] it becomes a pile of sundry facts some of them interesting or curious but making no meaningful picture as a whole.“
– Dobzhansky 1973

Human Proteomics



Human Genome

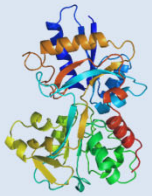
~ 3 billion ATCGs

Gene

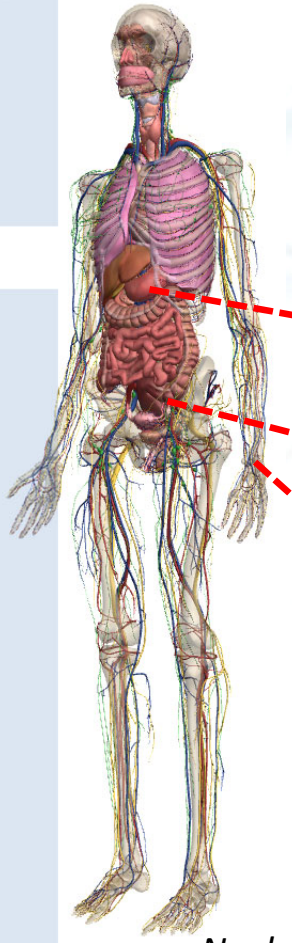
~20 000 protein coding sequences

mRNA

Protein



~42 000 Protein isoforms

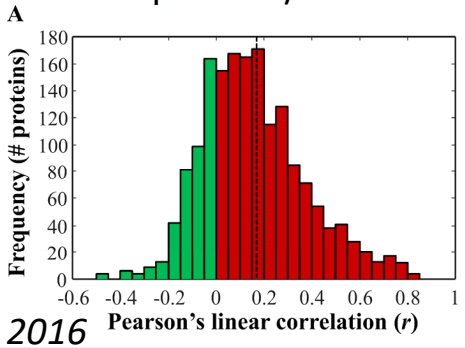


Proteomics

Liver 14 000 proteins
3000-5000 proteins/2hr MSMS

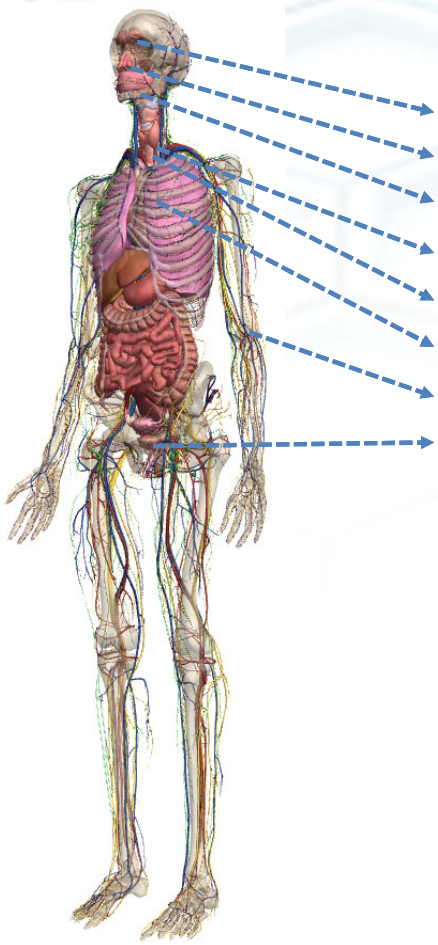
Kidney 14 800 proteins
3000-5000 proteins/2hr MSMS

Blood 3300 proteins
200-400 proteins/2hr MSMS



Neely et al., 2016

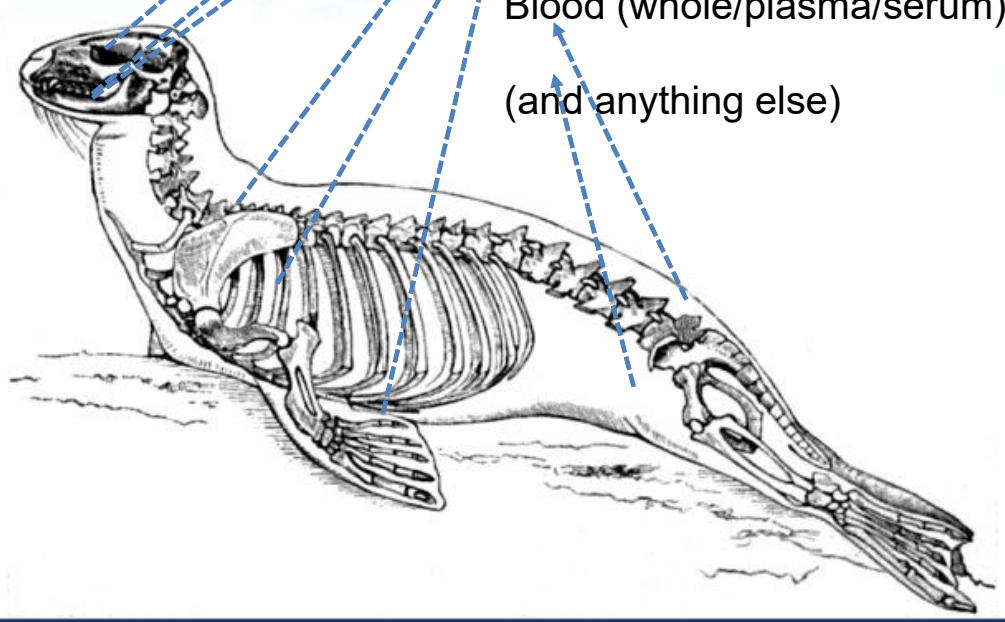
Proteomics



<u>Fluids:</u>
Tears
Nasal Mucus
Saliva
CSF
BAL
Breast Milk
Blood (whole/plasma/serum)
Urine
(and anything else)

<u>Tissues:</u>
Any

Proteomics



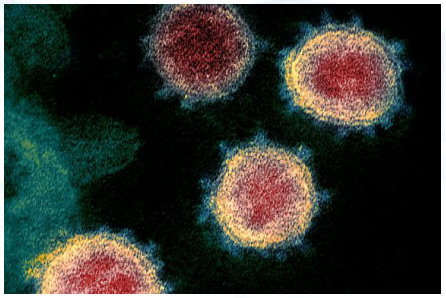
Fluids:
Tears
Nasal Mucus
Saliva
CSF
BAL
Breast Milk
Blood (whole/plasma/serum)

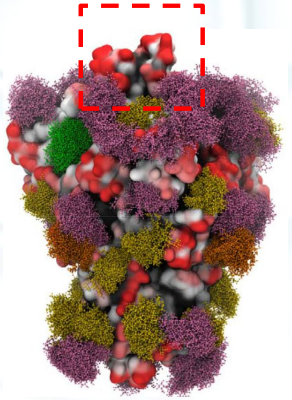
(and anything else)

Goals:

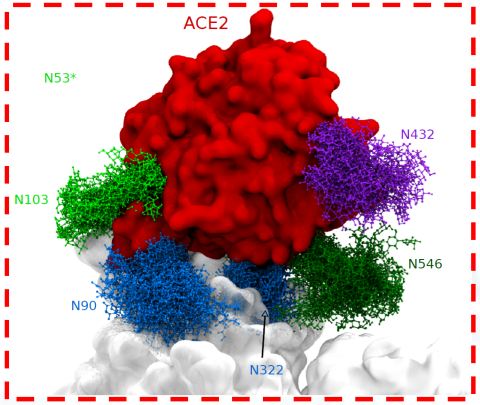
- Characterization
  - Compare across species
- Biomarker Discovery
  - Discrimination
  - Stratification
  - Monitoring treatment
- Systems Biology
  - Holistic understanding
  - Treatment targets
  - New hypotheses

SARS-CoV-2 and Mass Spec



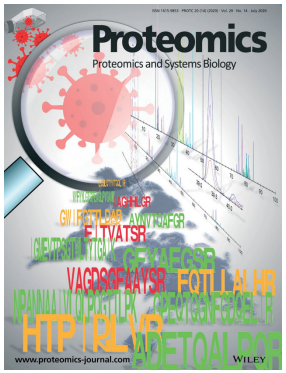


Grant et al. 2020

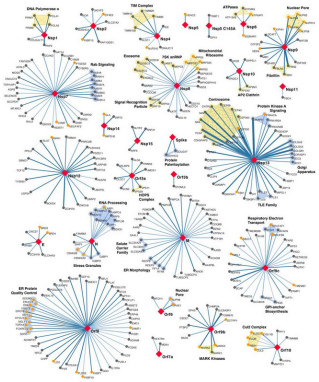


Zhao et al. 2020

Detection

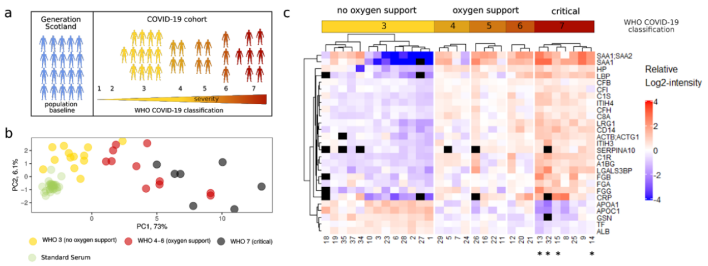


Interaction



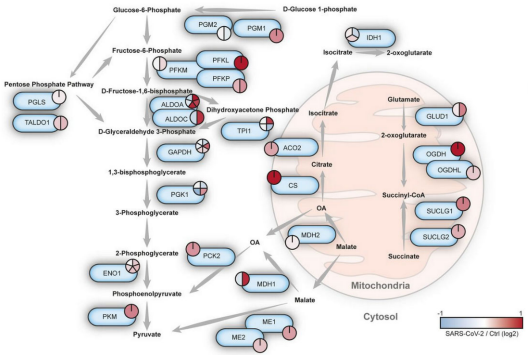
Gordon et al. 2020

Biomarkers



Messner et al. 2020

Host response

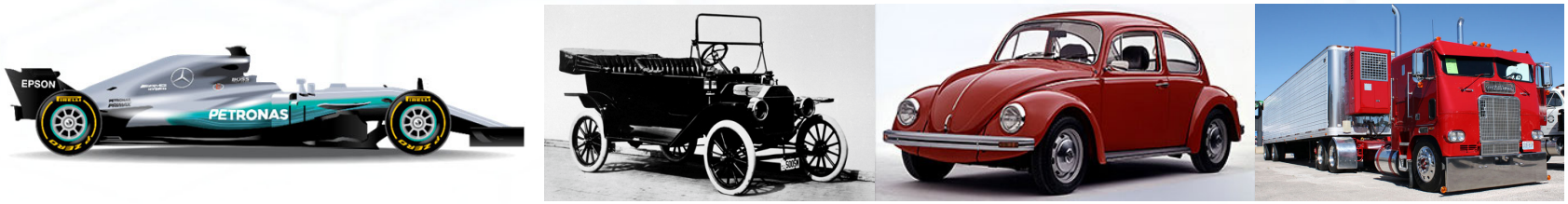


Klann et al. 2020

Comparative proteomics and biomimicry

Basic premises for today:

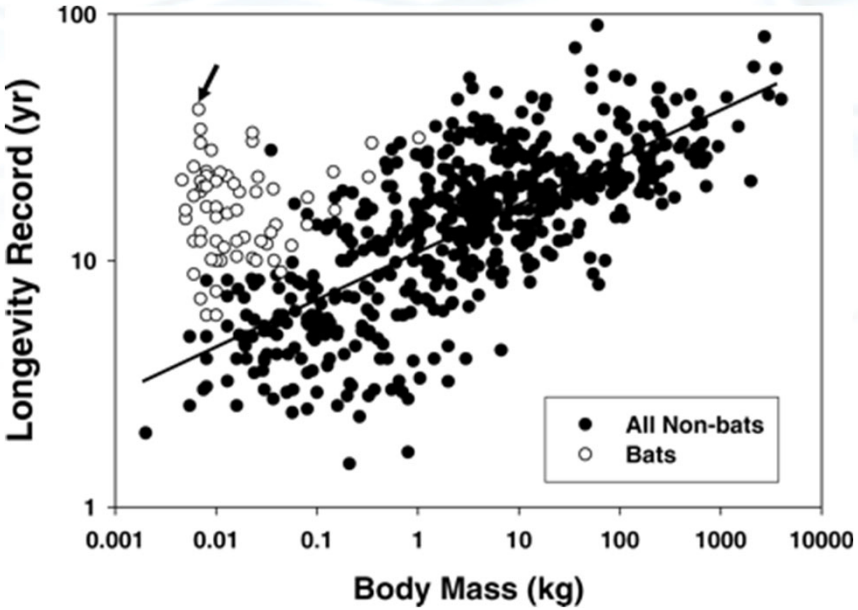
- Mammals are mammals are mammals
  - approximately same organs and same proteins (that are mostly orthologous)
  - There likely isn't some unique protein that confers superpower phenotypes

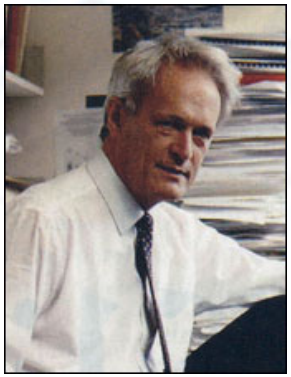


- For every biologically relevant phenotype, nature has a naturally occurring system

Biomimicry – longevity

Studying species that live longer than expected help us better understand aging in humans.

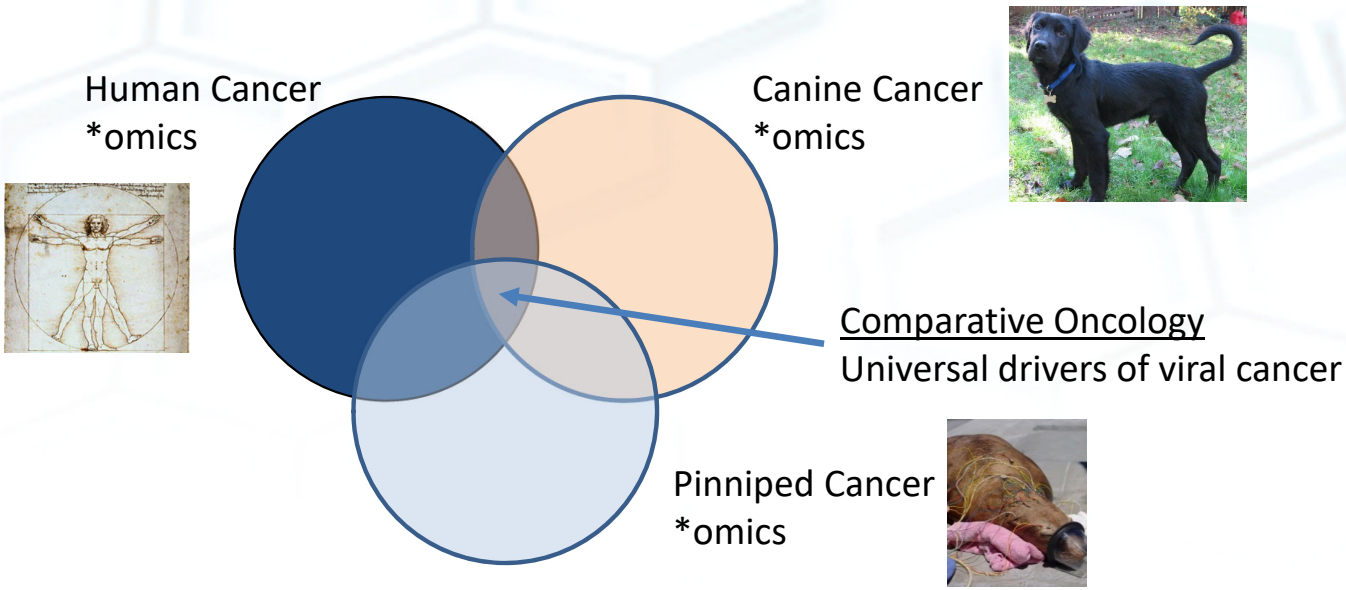




Sir Richard Peto

Podlutsky *et al.,* 2005, J Gerontol, doi: 10.1093/gerona/60.11.1366
Image credits: Manuel Ruedi; Shutterstock.com/belizar

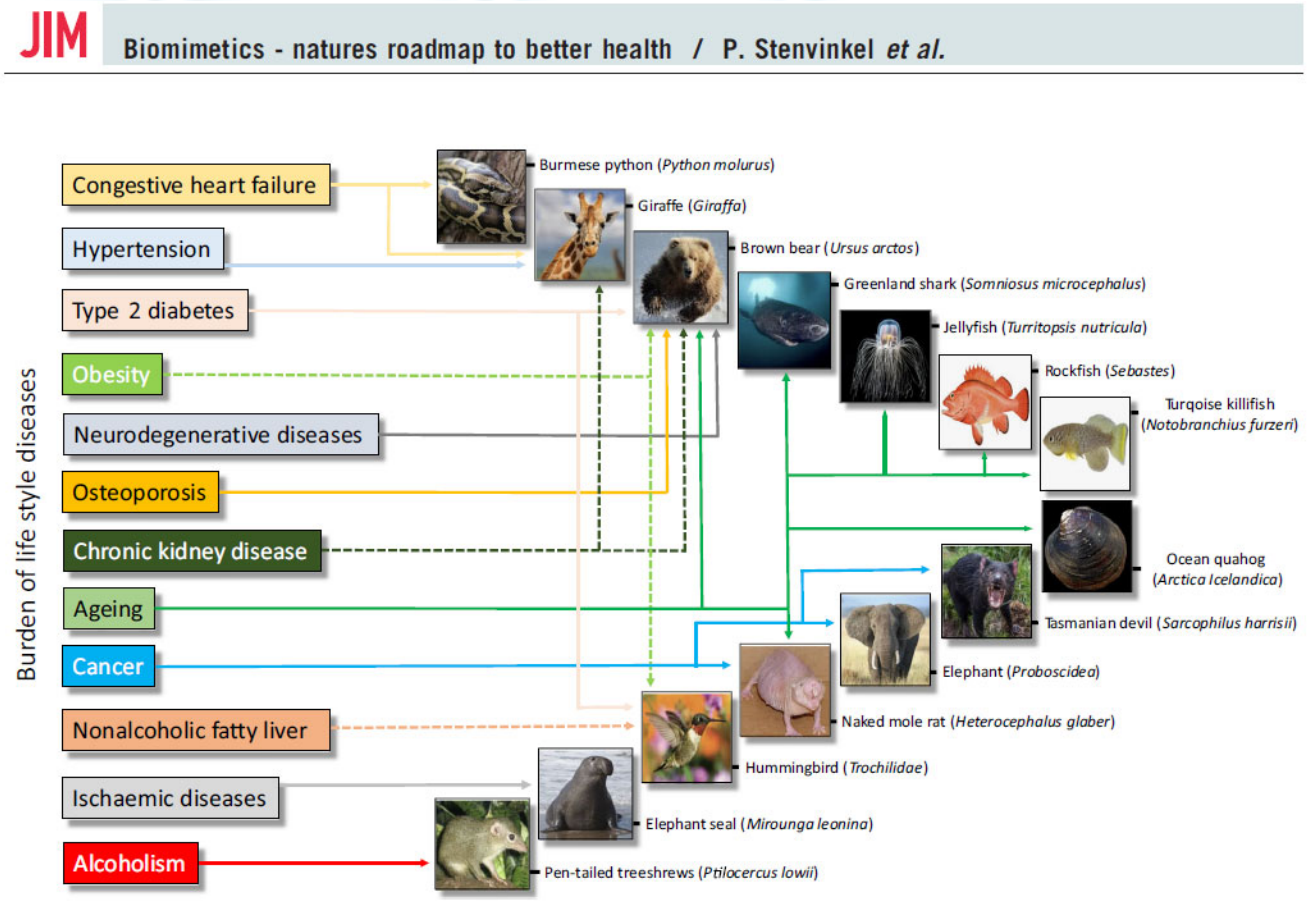
Comparative Oncology



- Need for non-rodent naturally occurring cancer models with human relevance
- 20% of human cancer is viral
- California sea lion has the highest cancer prevalence of wild mammals (26% have urogenital carcinoma)
- Identify aberrant cellular pathways of virally-induced cancers common to humans and animals.

Beyond Bats and Longevity

- For every biologically relevant phenotype, nature has a naturally occurring system



Mammals that dive and don't dive

Ability: Dive 1000's feet for over an hour

When diving:
- Slow heart rate
- Decrease flow to some organs

When re-surface:
- reperfusion with oxygen rich blood

Human equivalent:
- Heart attack, stroke, acute kidney injury

Non-diving relatives

Ability: cute best friend

**does not dive*

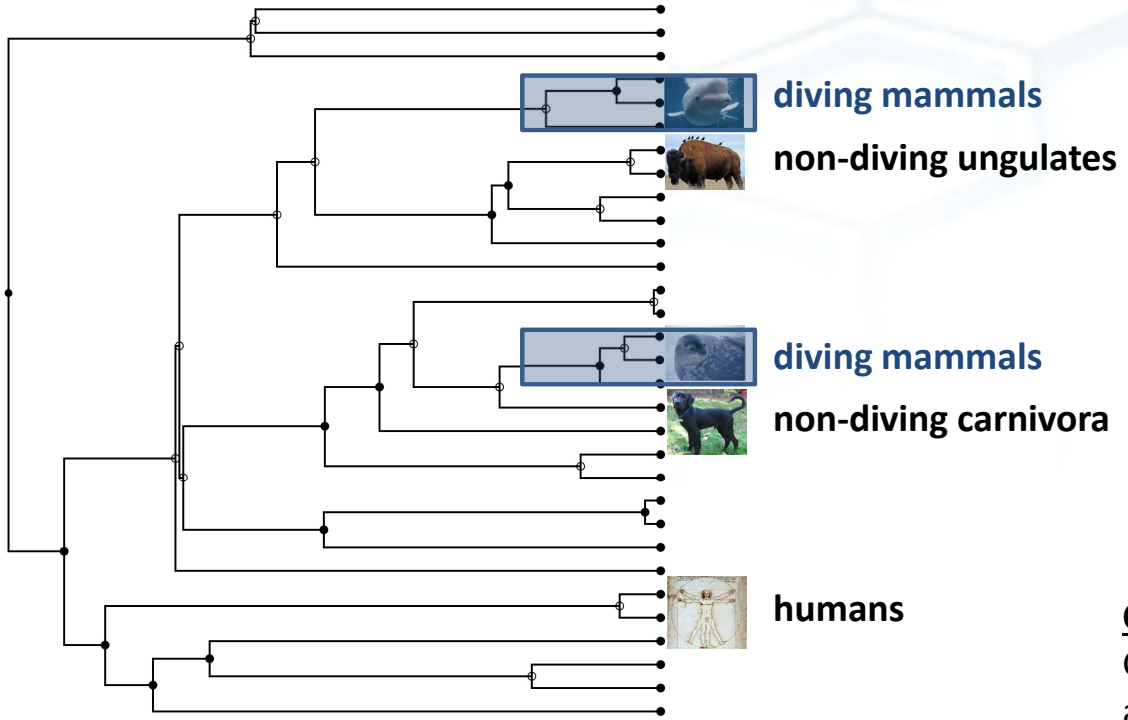
Ability: chase Kevin Costner

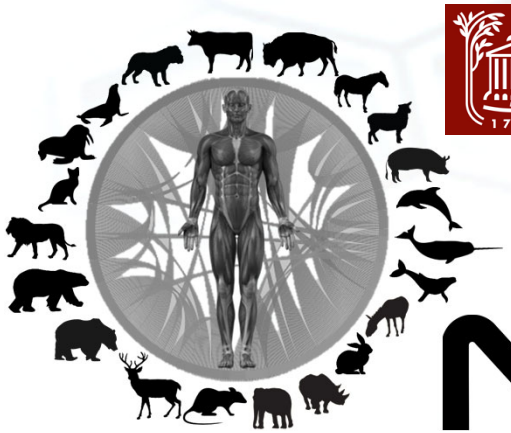
**does not dive*

Image credits: AFSC (NOAA); NPS; Mystic Aquarium

Biomimicry – Basic unknowns to answer larger questions

To ask these big questions, we have to begin exploring Mammals on a large scale.





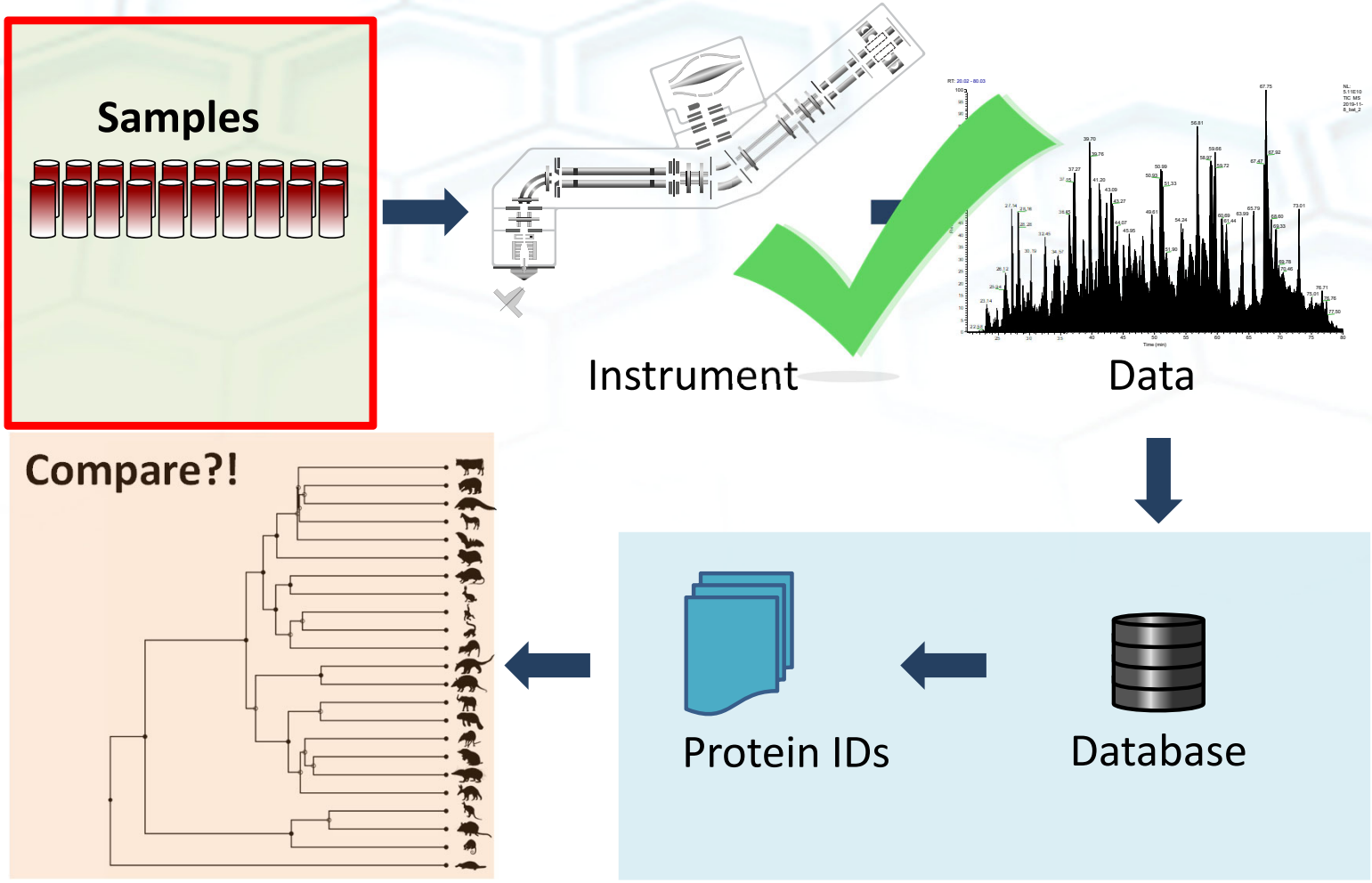
COLLEGE of CHARLESTON

NIST

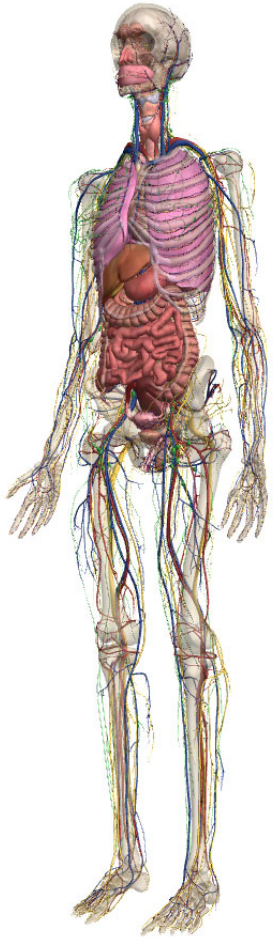
Comparative Mammalian Proteome Aggregator Resource

Generate data in a standard method from many mammals and make publicly available

Large Scale Comparative Proteomics – The three problems



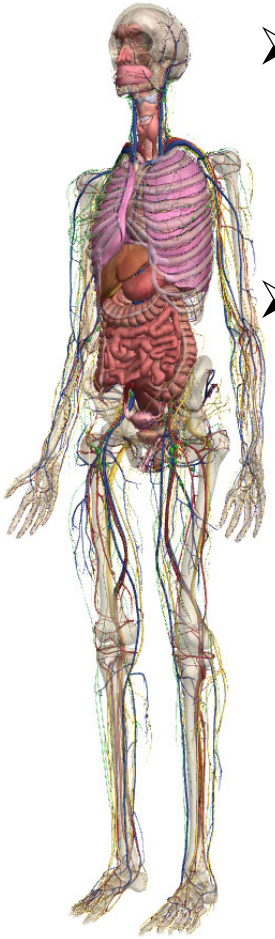
What is Blood?



- Blood is fairly easy to get
- Proximal to the whole system carrying nutrients/waste
  - Provides both direct and indirect information on organs (signaling/leaking)
- Contains known markers of disease/health status
  - Protein abundance and PTMs in blood inform chronic and acute phenotypes

1993 Haddaway's "What is Love" music video is basically a vampire party.

What is Blood?

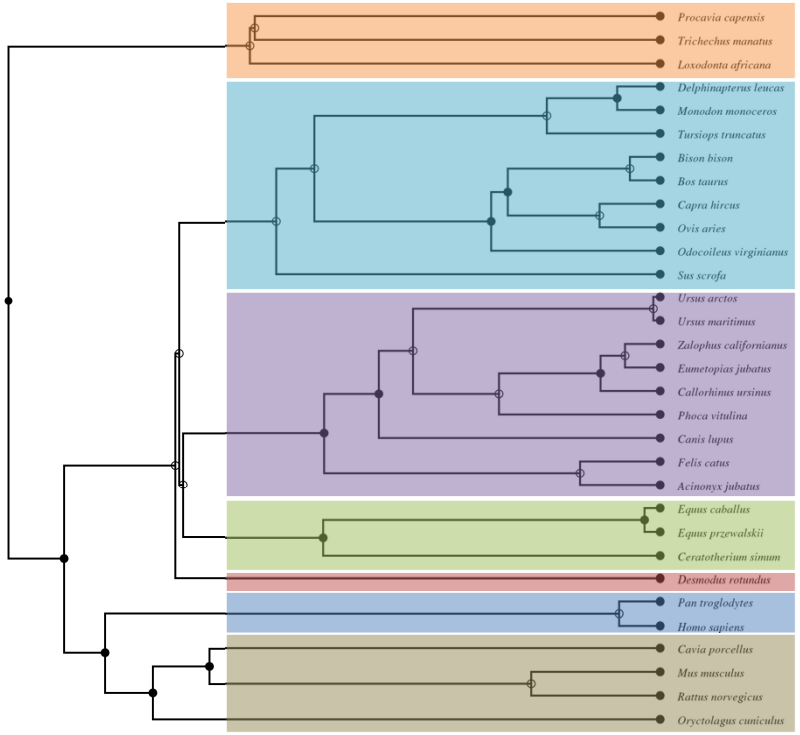


- Whole blood contains cells and plasma
  - Serum is fluid left after clotting
  - Plasma is fluid from unclotted
- Blood proteome is not from “blood”
  - Transcript levels ≠ protein levels

1993 Haddaway’s “What is Love” music video is basically a vampire party.

CoMPARe Phase I: serum from 31 species

CoMPARe Phase I: serum from 31 species



Afrotheria: African elephant, FL manatee, rock hyrax

Artiodactyla (even-toed ungulates): cow, bison, sheep, goats, pig, white-tailed deer

cetaceans: bottlenose dolphin, beluga, narwhal

Carnivora: brown bear, polar bear, CA sea lion, N fur seal, harbor seal, steller sea lion, dog, cat, cheetah

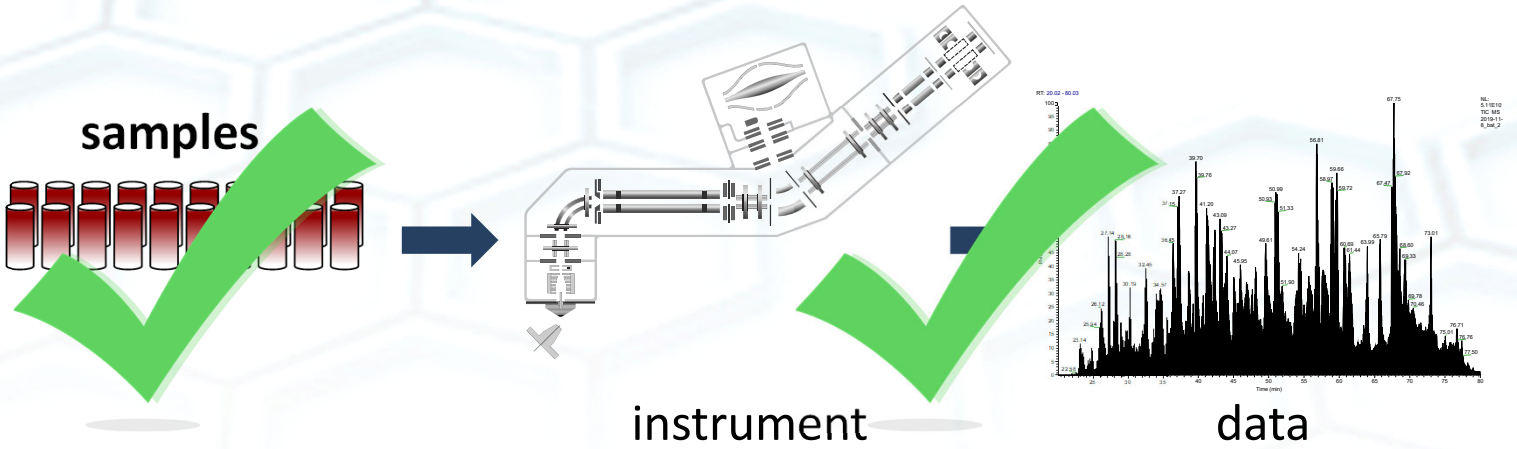
Perissodactyla (odd-toed ungulates): horse, Przewalski's horse, white rhino

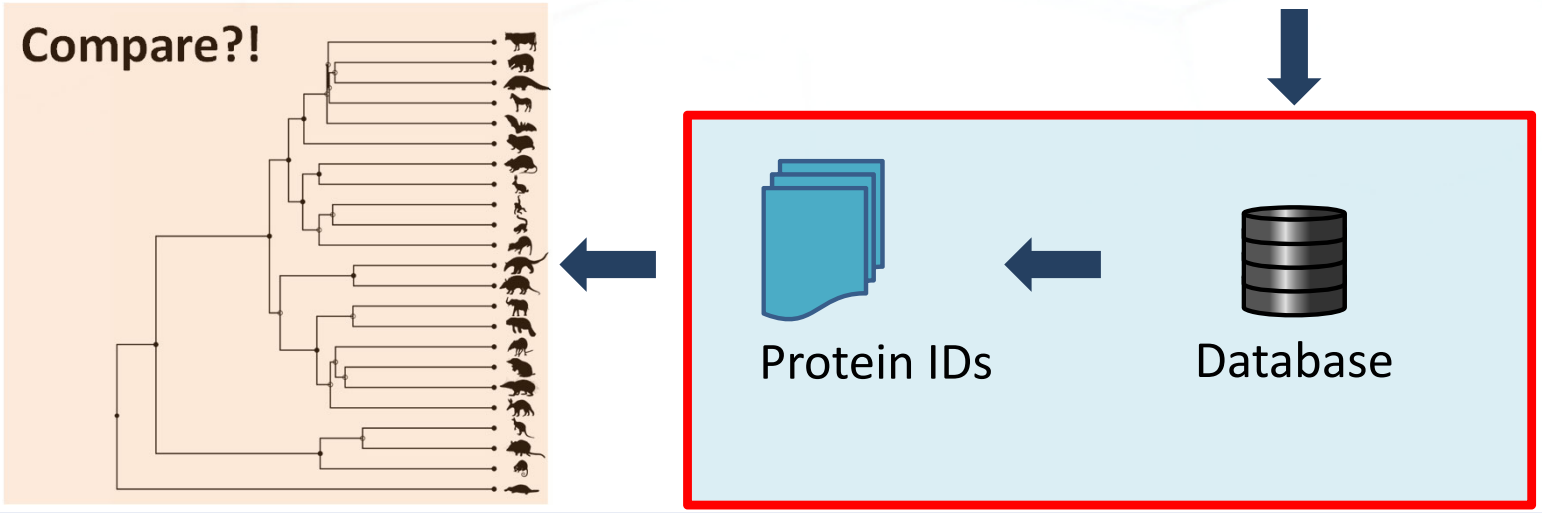
Chiroptera: vampire bat

Primates: human, chimp

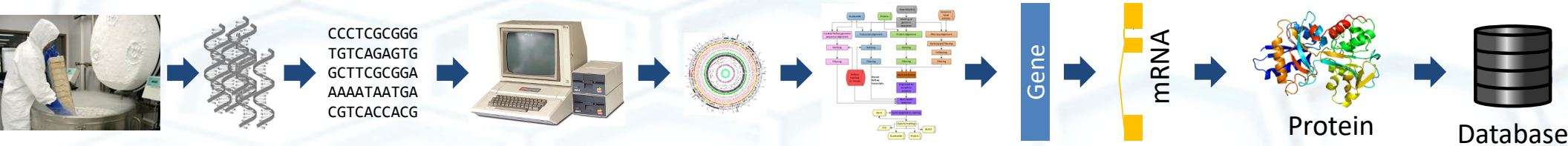
Glires: guinea pig, mouse, rat, rabbit

Large Scale Comparative Proteomics





You need a sequence database!



Genome Sequencing	Genome Assembly	Genome Annotation
	>10 % of mammals	<3 % of mammals

Fasta via…

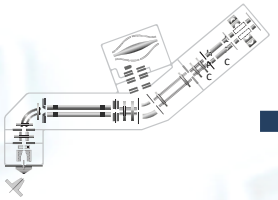
	Mammalian species with annotated genomes
NCBI RefSeq	144
DNA Zoo	109
e!Ensembl	42 (this number may be higher)
UniProt	*did not count

- 20 % of mammal species (~1400)
- Less than 50 genomes currently completed
- 10 species annotated on RefSeq

You need a sequence database! What if there isn't one?

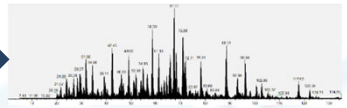
Biofluid (or tissue)

protein



Mass Spec

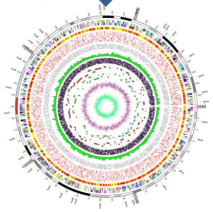
100,000s MS/MS spectra



Protein Identifications
(100s-10,000)

Single blood or tissue sample

DNA

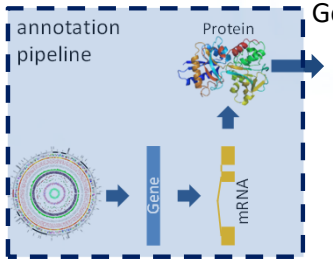


Genome

Tissue(s)

mRNA sequencing

NCBI RefSeq



Annotated Genome

DNA Zoo

VGP

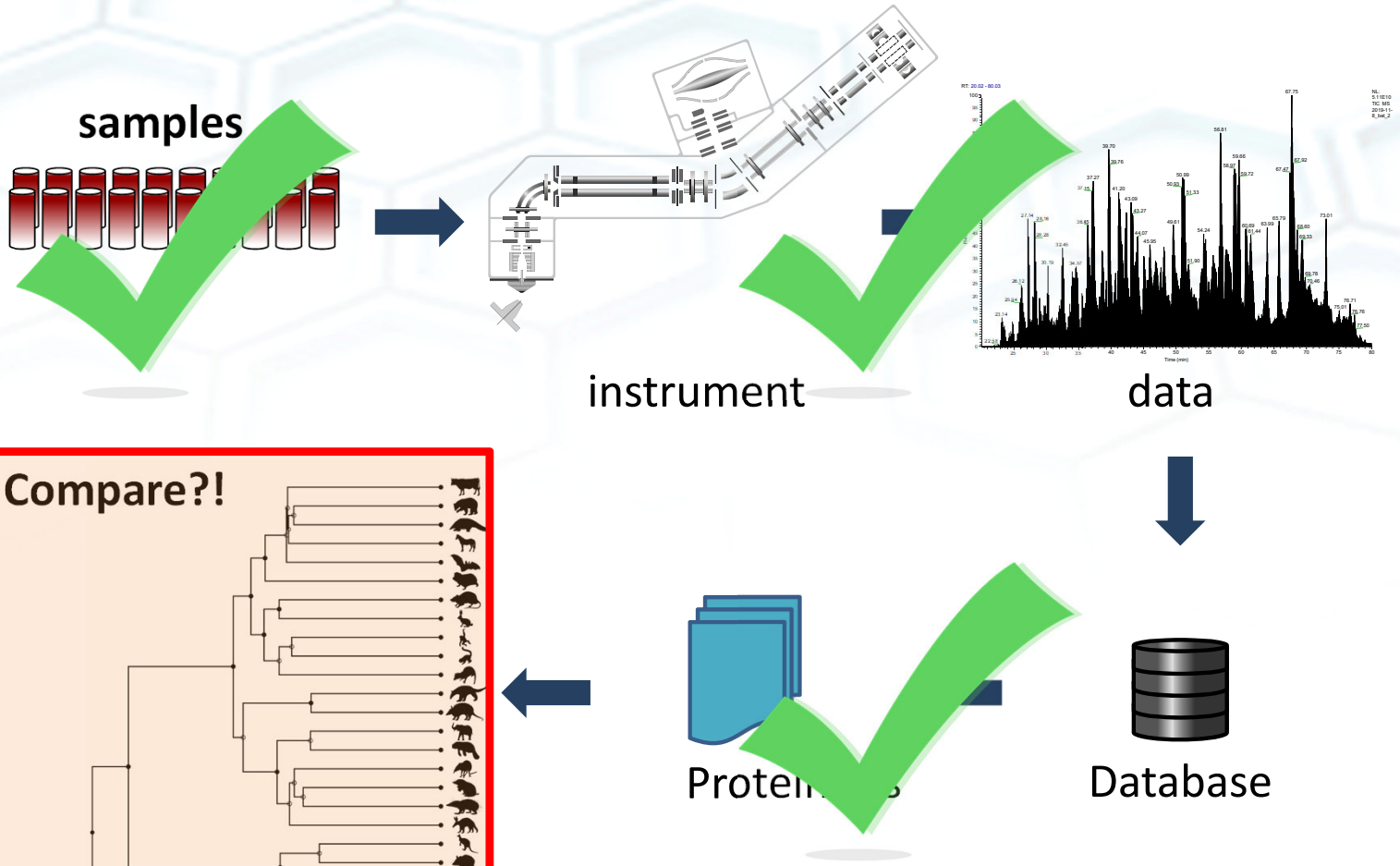
CA sea lion annotation completed December 2018

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

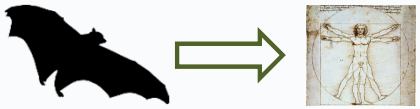
NIST

Ex. by next year, with major assistance from DNA Zoo, all marine mammals in NIST Biorepository will have genomes completed and publicly available.

Large Scale Comparative Proteomics



pwilmart / PAW_BLAST



Vampire Bat Clusterin:
>XP_024424850.1 clusterin isoform X1 [Desmodus rotundus]

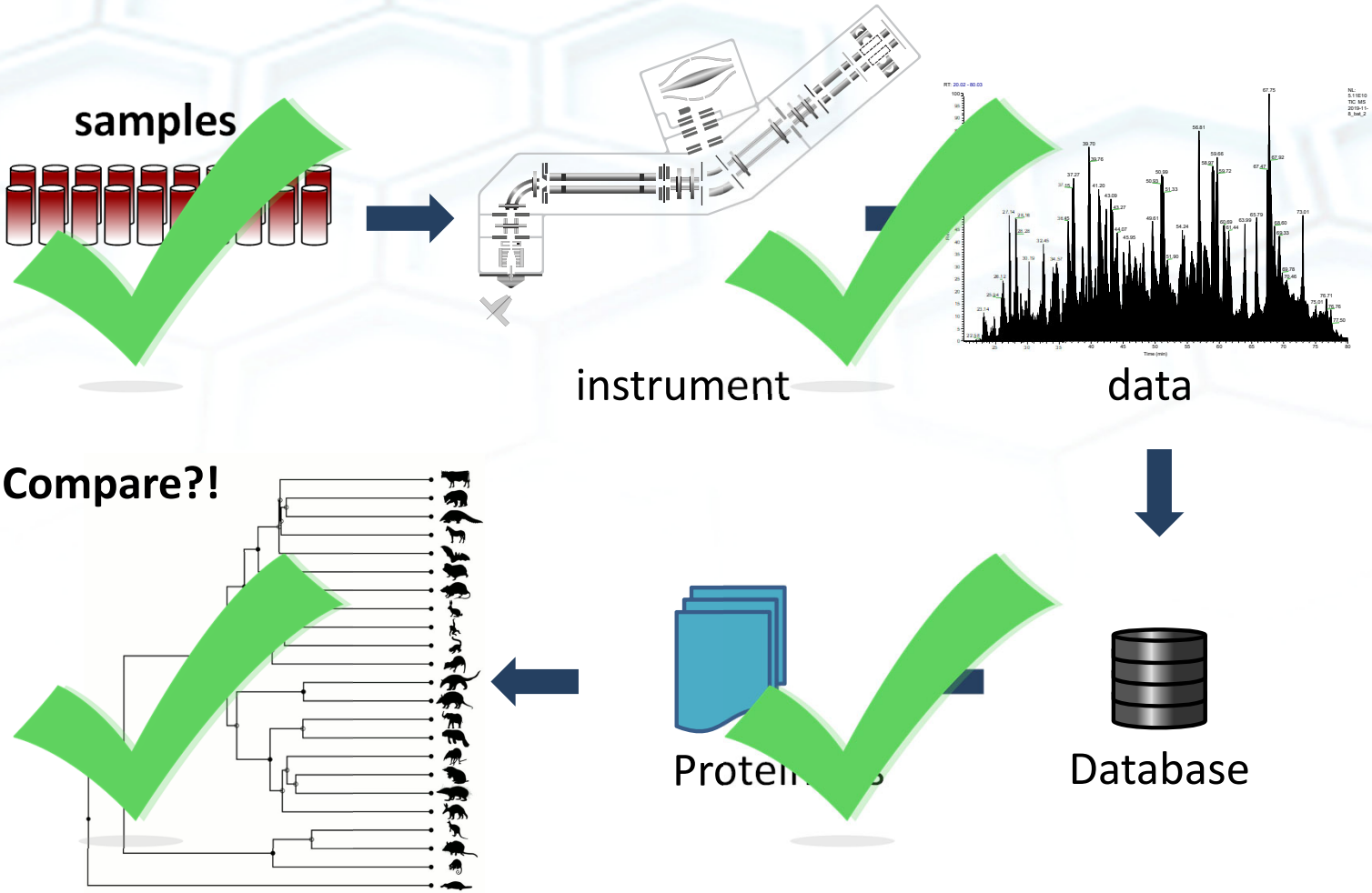
Human Ortholog
P10909 Clusterin CLU 62.3% ID

Rock Hyrax Clusterin:
>ENSPCAP00000007283.1 pep
scaffold:proCap1:scaffold_4568:239:20231:1
gene:ENSPCAG00000007764.1 transcript:ENSPCAT00000007781.1
gene_biotype:protein_coding transcript_biotype:protein_coding
gene_symbol:CLU description:clusterin [Source:HGNC
Symbol;Acc:HGNC:2095]

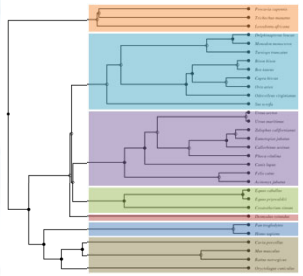
P10909 Clusterin CLU 53.6% ID

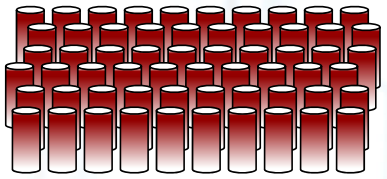
CAL POLY
SAN LUIS OBISPO

Large Scale Comparative Proteomics



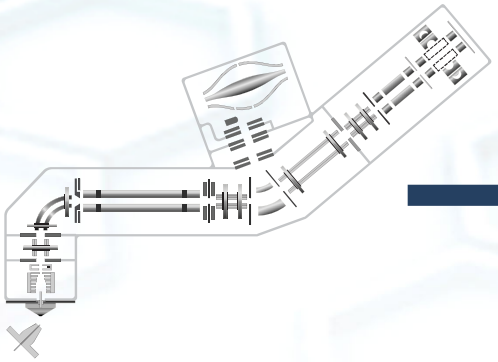
CoMPARe Phase I approach

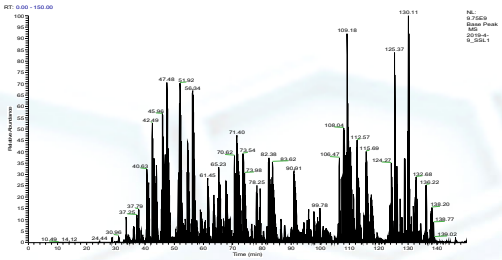




Serum from up to 4 individuals from each species
(+ NIST Serum)

Tryptic Digest





Raw files

Pd
v2.2

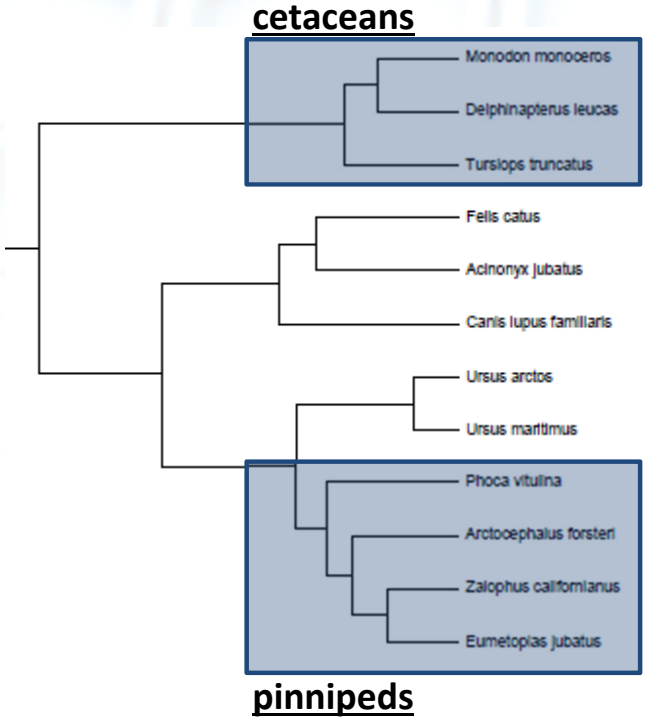
Minora Feature Detector
Sequest HT
Mascot

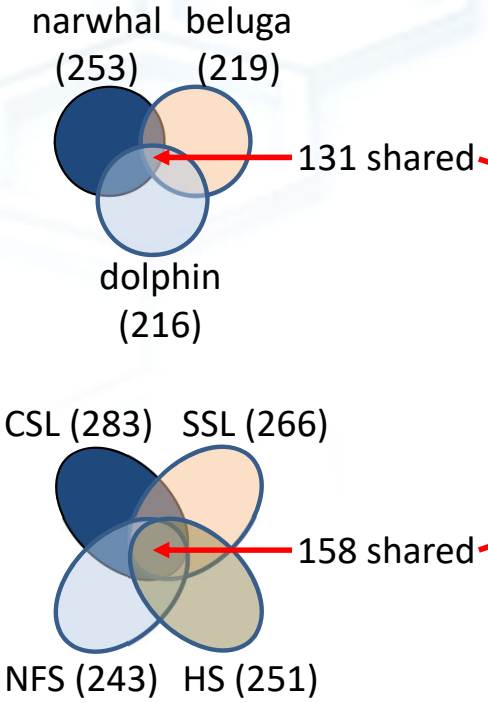
Fasta via…
NCBI RefSeq
DNA Zoo

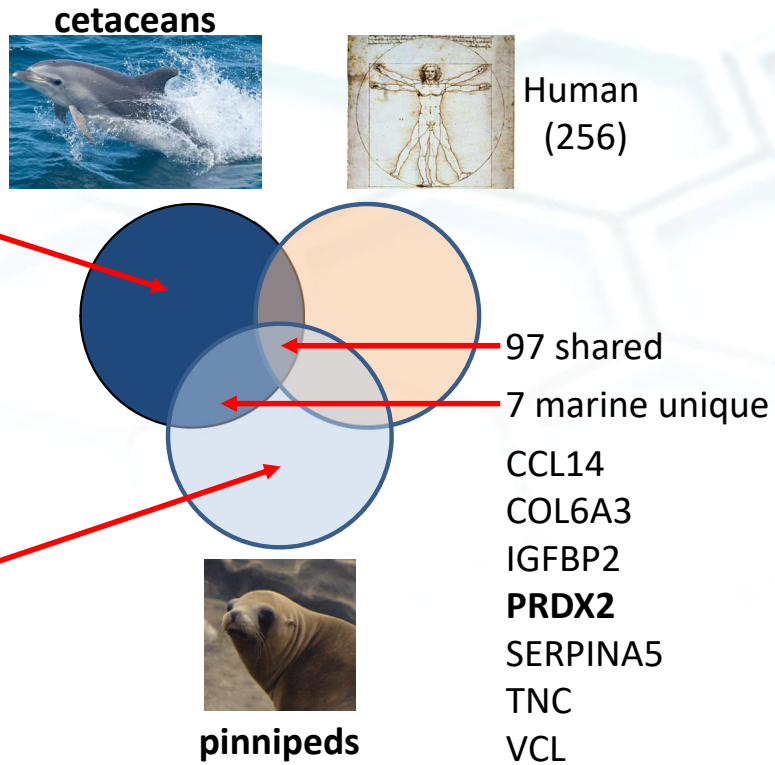
pwilmart / PAW_BLAST

Protein IDs
w/ MS1 quant
w/ human orthologs

CoMPARe Phase I Results: diving v human

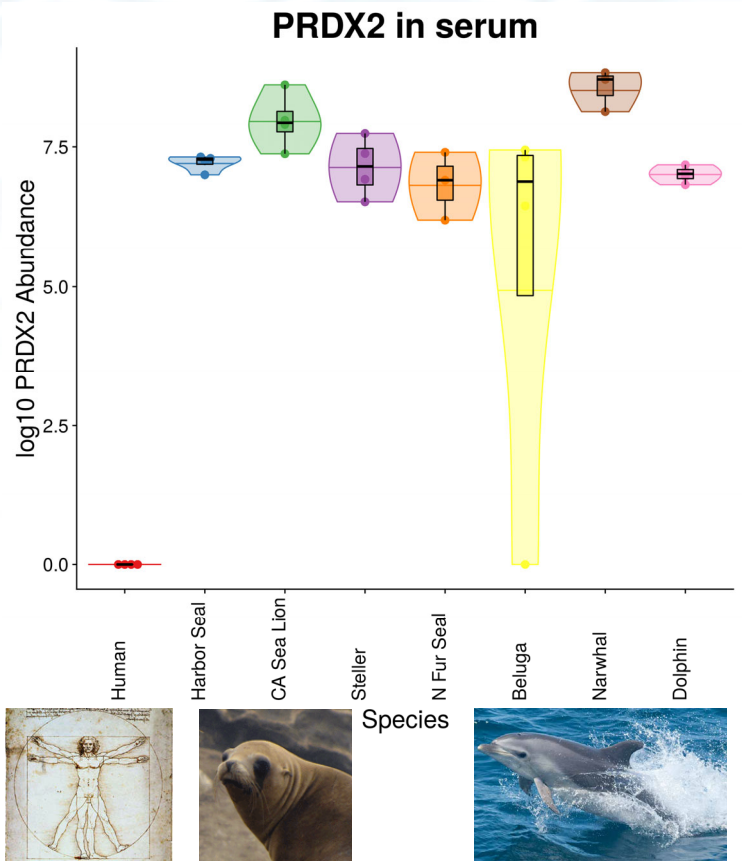




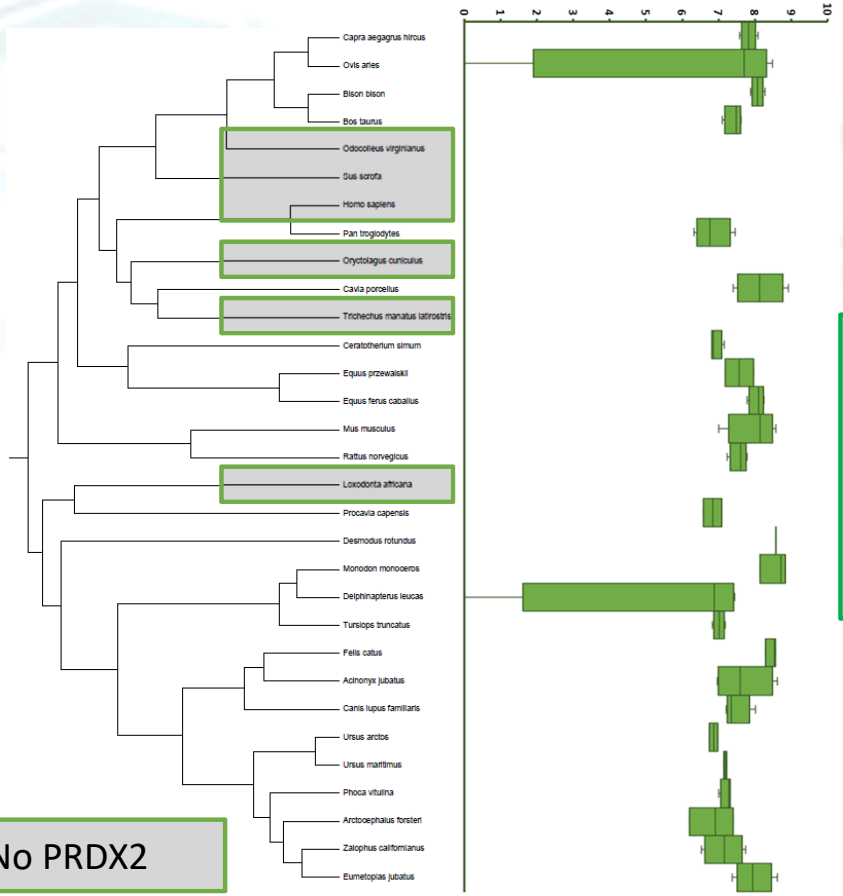


Note: these samples weren't collected while diving, and are more reflective of "normal" proteome.

PRDX2 - Peroxiredoxin-2



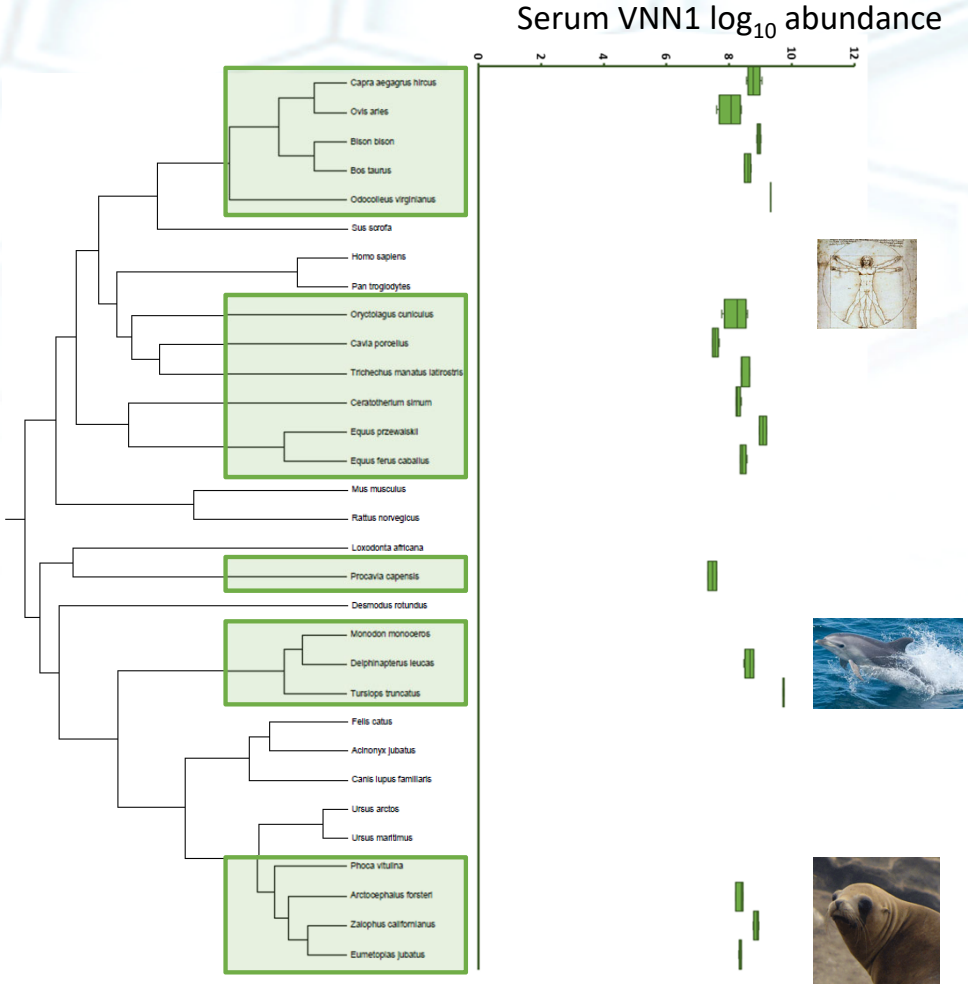
Serum PRDX2 log_{10} abundance

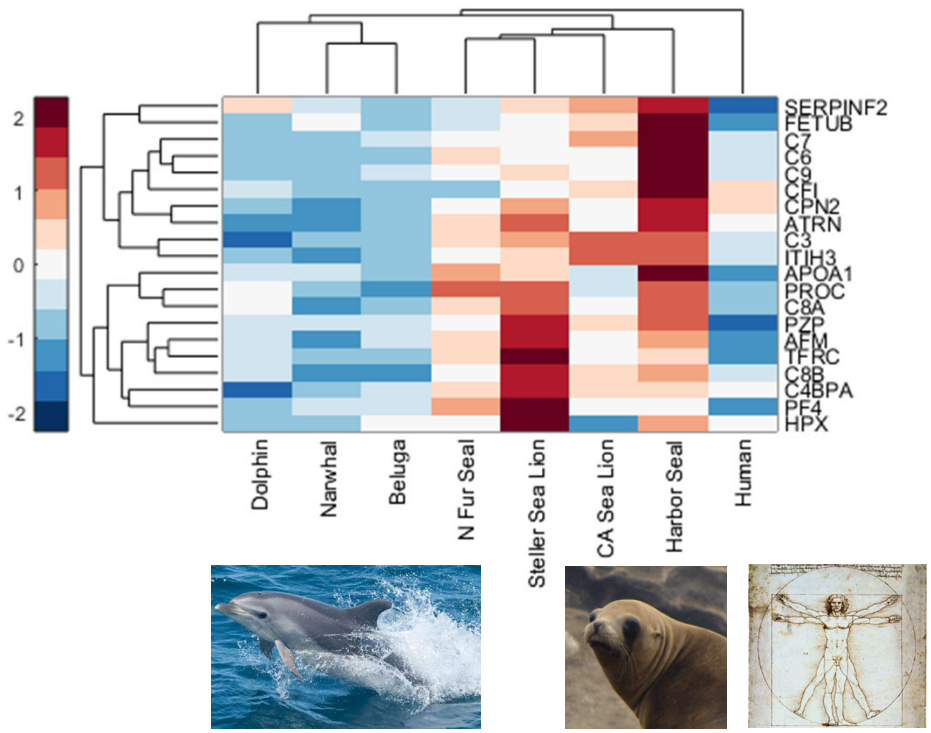


No PRDX2

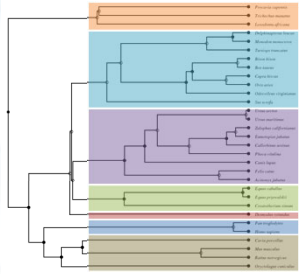
Missing in:
white-tailed deer
pig
human
FL manatee
African Elephant

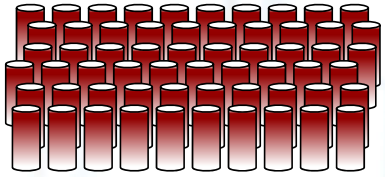
Comparisons by protein or between clades





Sample processing and data acquisition approach



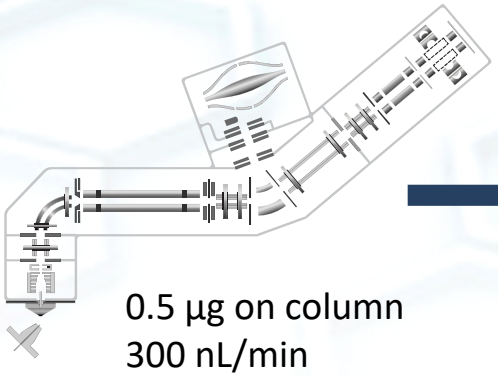


Serum from up to 4 individuals from each species

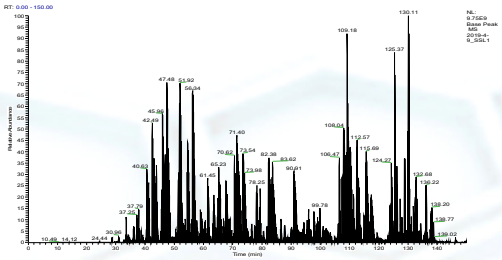
+ NIST SRM 909c Human Serum per digestion

Tryptic Digest

No depletion



0.5 µg on column
300 nL/min
2-hour, 2-step gradient



≈60,000 MS/MS spectra

Raw files

ID-free

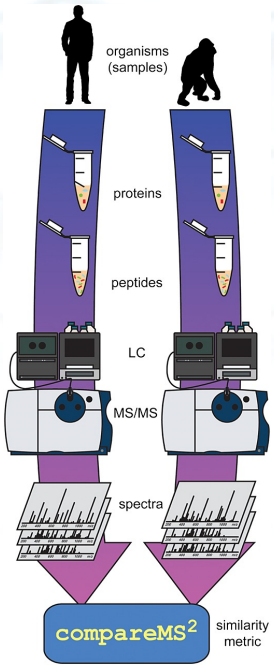
Thermo Raw File Parser

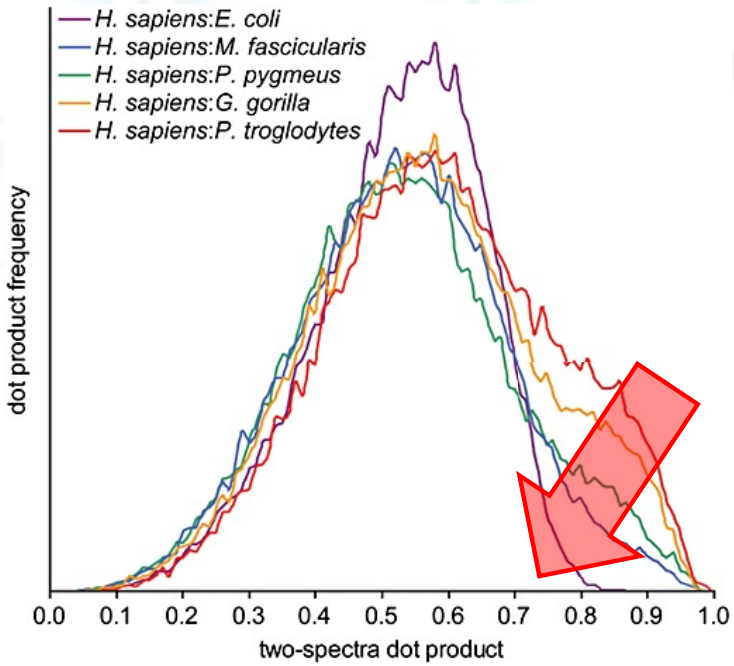
mgf

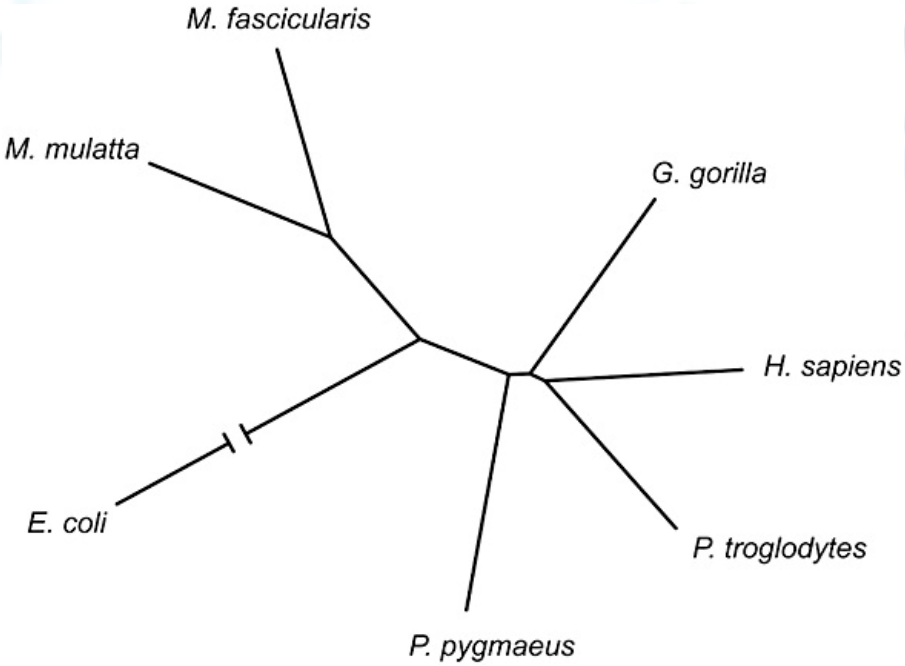
compareMS^2

similarity metric

ID-free spectrum-to-spectrum approach: compareMS2

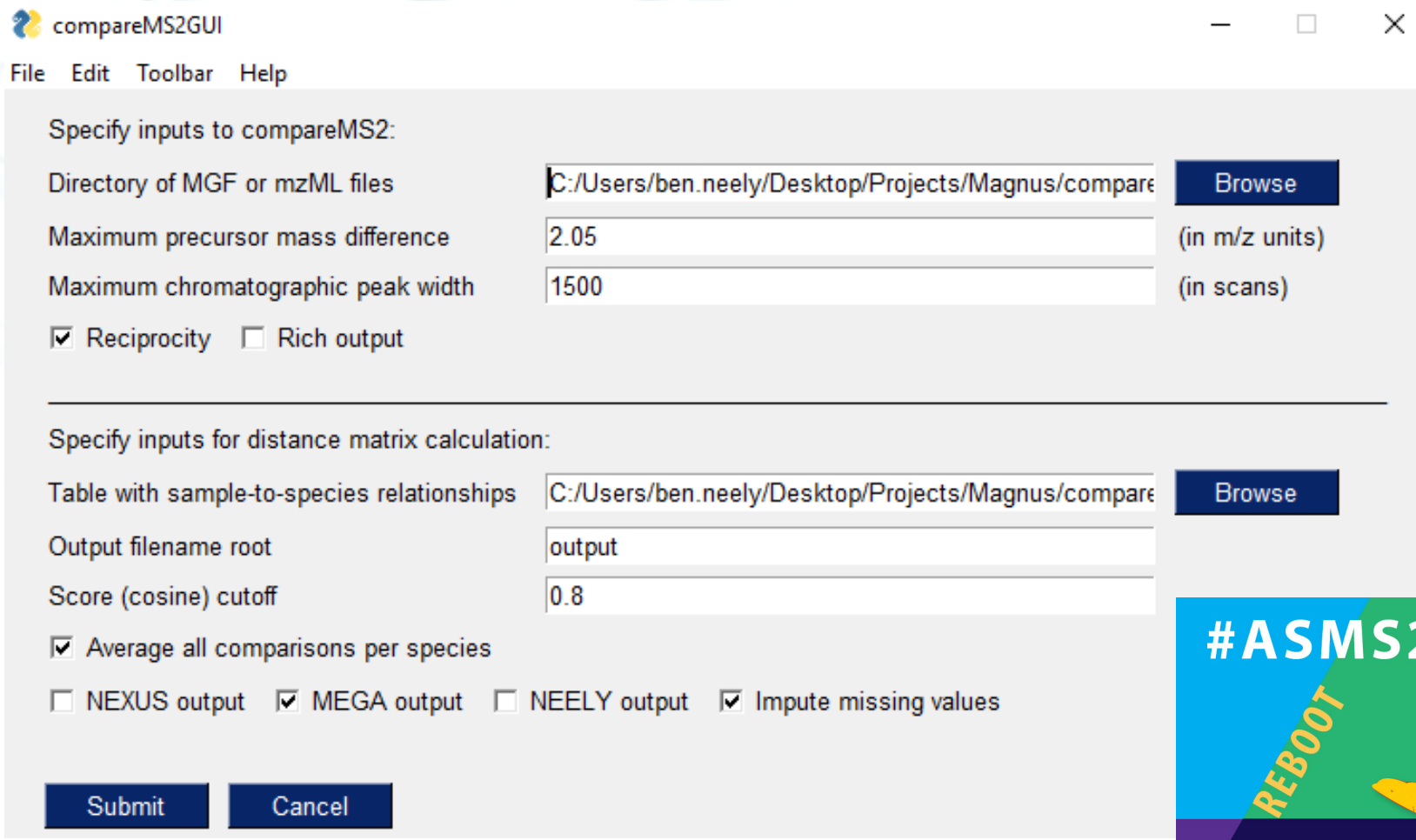






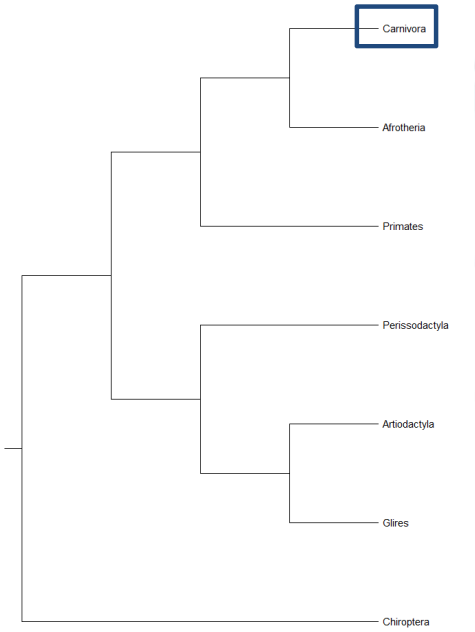
Palmblad and Deelder, 2012, Rap Com Mass Spec. https://doi.org/10.1002/rcm.6162

ID-free spectrum-to-spectrum approach: compareMS2



compareMS2 Results: 31 species undepleted serum DDA

Order Level Tree



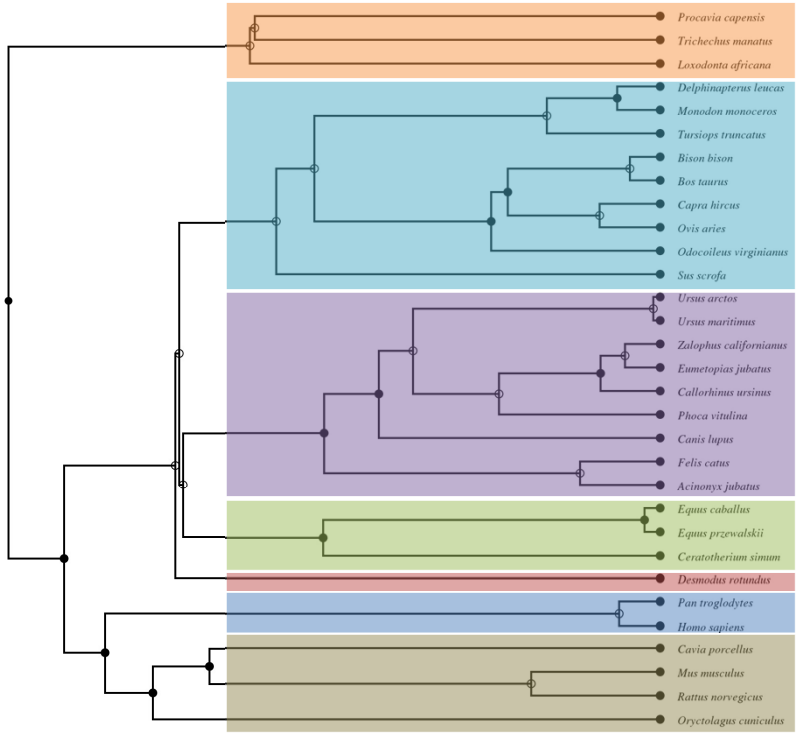
Carnivora

Felis catus
Acinonyx jubatus
Canis lupus familiaris
Ursus arctos
Ursus maritimus
Phoca vitulina
Arctocephalus forsteri
Zalophus californianus
Eumetopias jubatus

Arctocephalus forsteri 2
Arctocephalus forsteri 4
Arctocephalus forsteri 1
Zalophus californianus 2
Zalophus californianus 3
Zalophus californianus 1
Zalophus californianus 4
Eumetopias jubatus 1
Eumetopias jubatus 2
Eumetopias jubatus 3
Eumetopias jubatus 4

Unexpected benefit: independent overview of the quality of and similarity between all datasets

CoMPARe Phase I: serum from 31 species



Afrotheria: African elephant, FL manatee, rock hyrax

Artiodactyla (even-toed ungulates): cow, bison, sheep, goats, pig, white-tailed deer

cetaceans: bottlenose dolphin, beluga, narwhal

Carnivora: brown bear, polar bear, CA sea lion, N fur seal, harbor seal, steller sea lion, dog, cat, cheetah

Perissodactyla (odd-toed ungulates): horse, Przewalski's horse, white rhino

Chiroptera: vampire bat

Primates: human, chimp

Glires: guinea pig, mouse, rat, rabbit

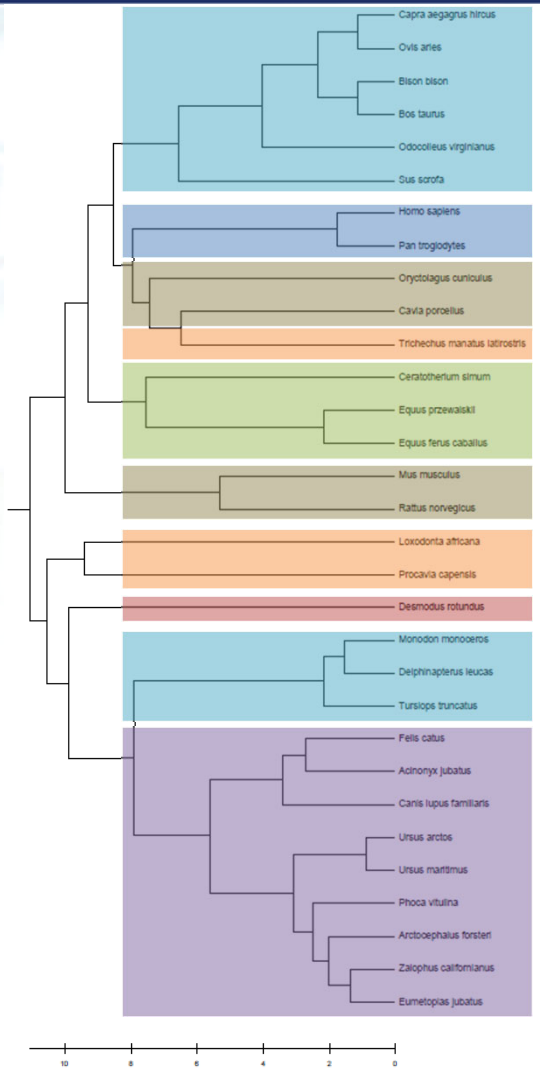
compareMS2 Results

Probability of 3 in 10^{40} to get the correct tree by chance

Relationships between samples are reflective of:

- sequence-dependent fragmentation spectra (i.e., genetic variability)
- Relative protein abundance
- post-translational modifications

- ID-free analysis provides useful applications beyond intended purpose



Artiodactyla (even-toed ungulates): goat, sheep, bison, cow, white-tailed deer, pig

Primates: human, chimp

Glires: rabbit, guinea pig

Afrotheria: FL manatee

Perissodactyla (odd-toed ungulates): white rhino, horse, Przewalski's horse

Glires: mouse, rat

Afrotheria: African elephant, rock hyrax

Chiroptera: vampire bat

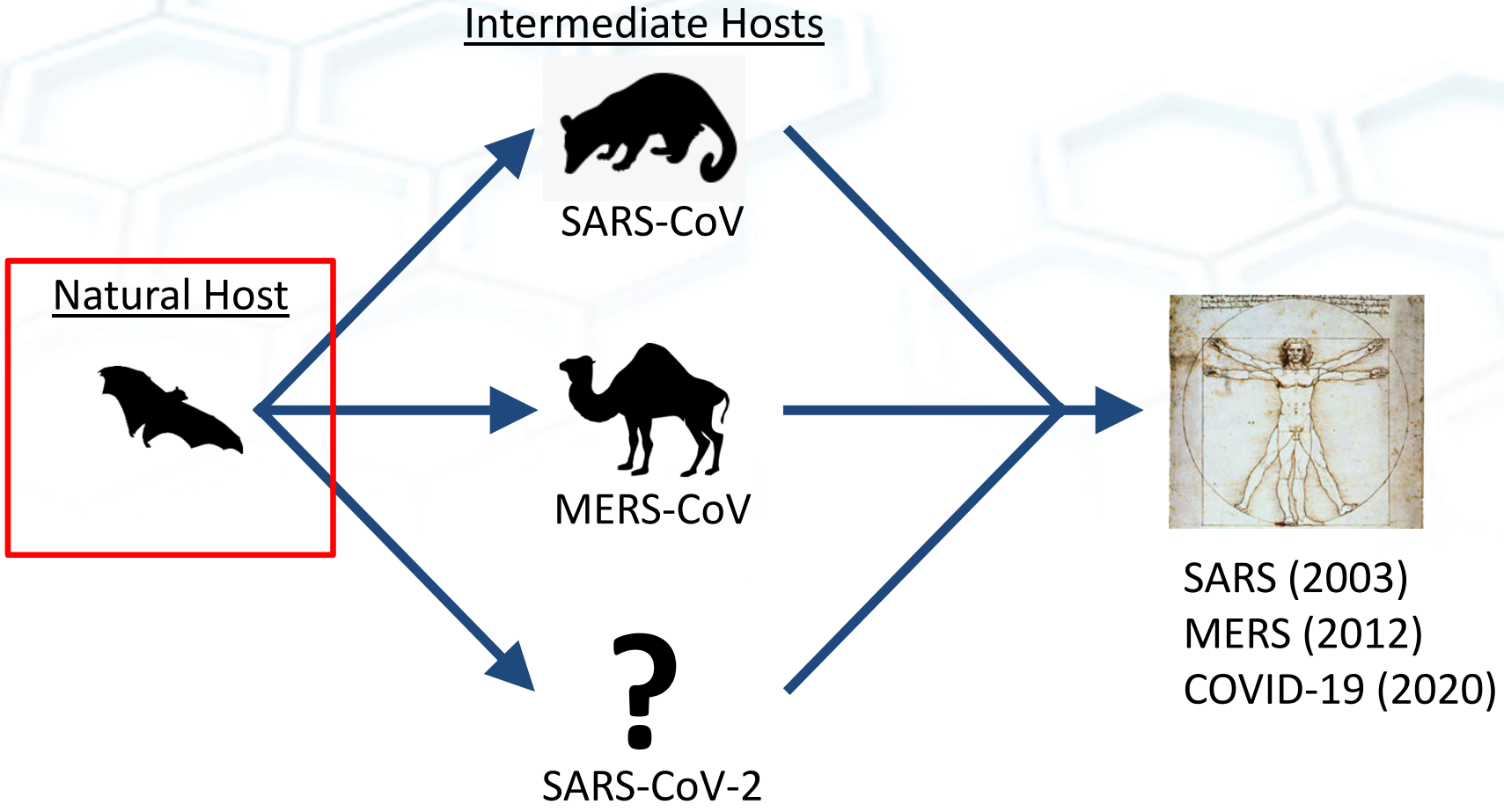
Artiodactyla (even-toed ungulates): narwhal, beluga, dolphin

Carnivora: cat, cheetah, dog, brown bear, polar bear, harbor seal, N fur seal, CA sea lion, steller sea lion

Individual species have broader impacts to stakeholders

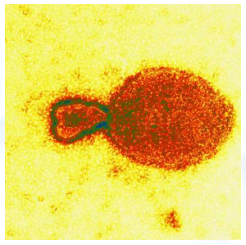
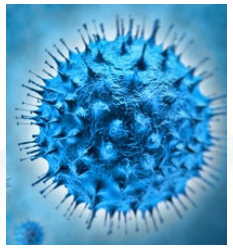
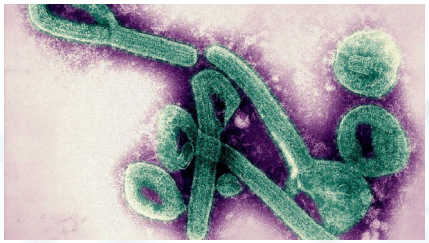
- Stakeholders have more bioanalytical tools at their disposal
- Mass Spec cores can offer more services using more species
- Funders (NIH/NSF) may be more likely to fund proteomic-based grants in more species

Novel to humans, not novel to other mammals

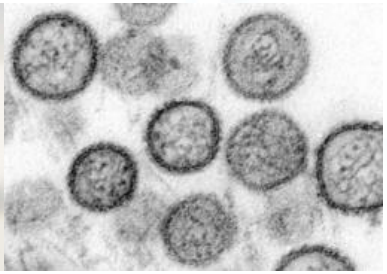


Other natural hosts and other viruses… new opportunities



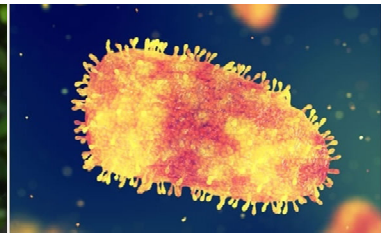





deer mouse

hantavirus

raccoon

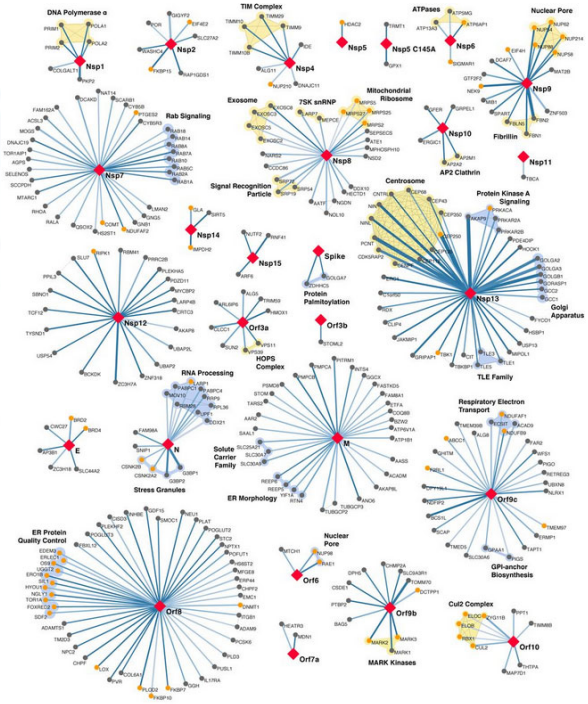



rabies

Verena
The Viral Emergence Research Initiative

Proteomics in host and reservoirs

- What could PPI studies in natural host tell us?



- **What do natural host even look like?**

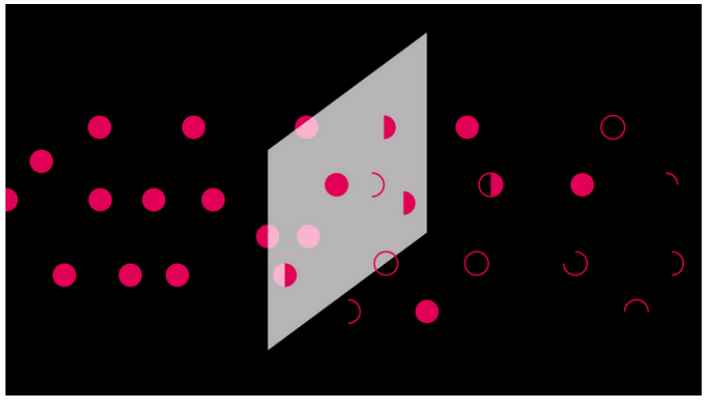
The Atlantic

HEALTH

Immunology Is Where Intuition Goes to Die

Which is too bad because we really need to understand how the immune system reacts to the coronavirus.

ED YONG



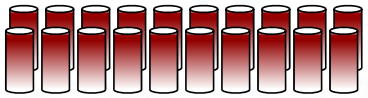
- Infection is not virulence
- Virulence is complicated and linked to the immune system
- Innate immune system is where it begins

Need to characterize the molecular landscape in non-human hosts and reservoirs

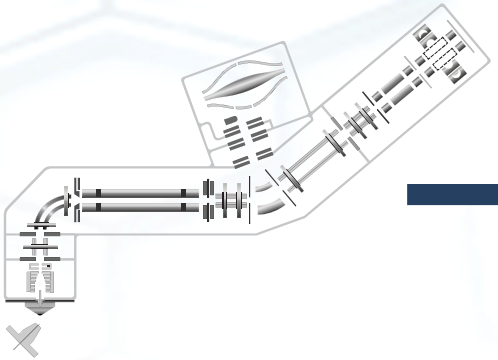
Image: Andrew Neel

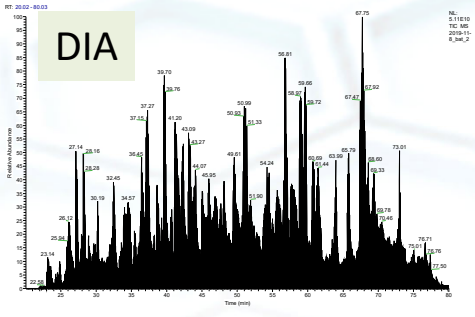
Ex. Vampire Bat DIA Study started because of COMPARE

The UNIVERSITY of OKLAHOMA



Serum from 17 bats
from two locations

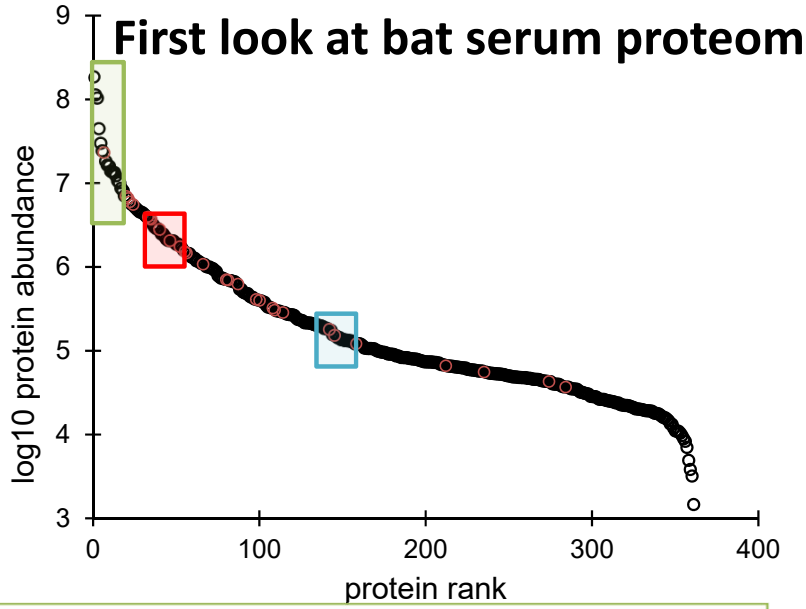


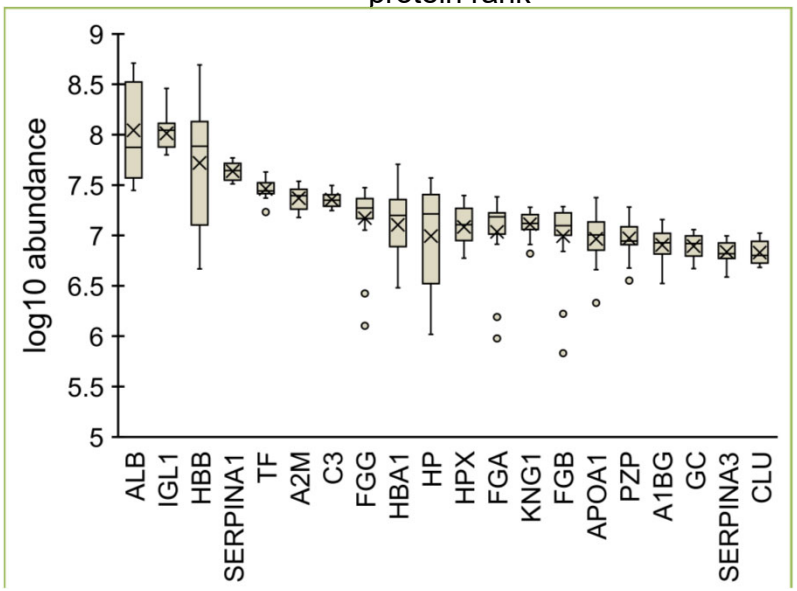


directDIA

pwilmart / PAW_BLAST

361 Protein IDs
w/ DIA quant
w/ human orthologs

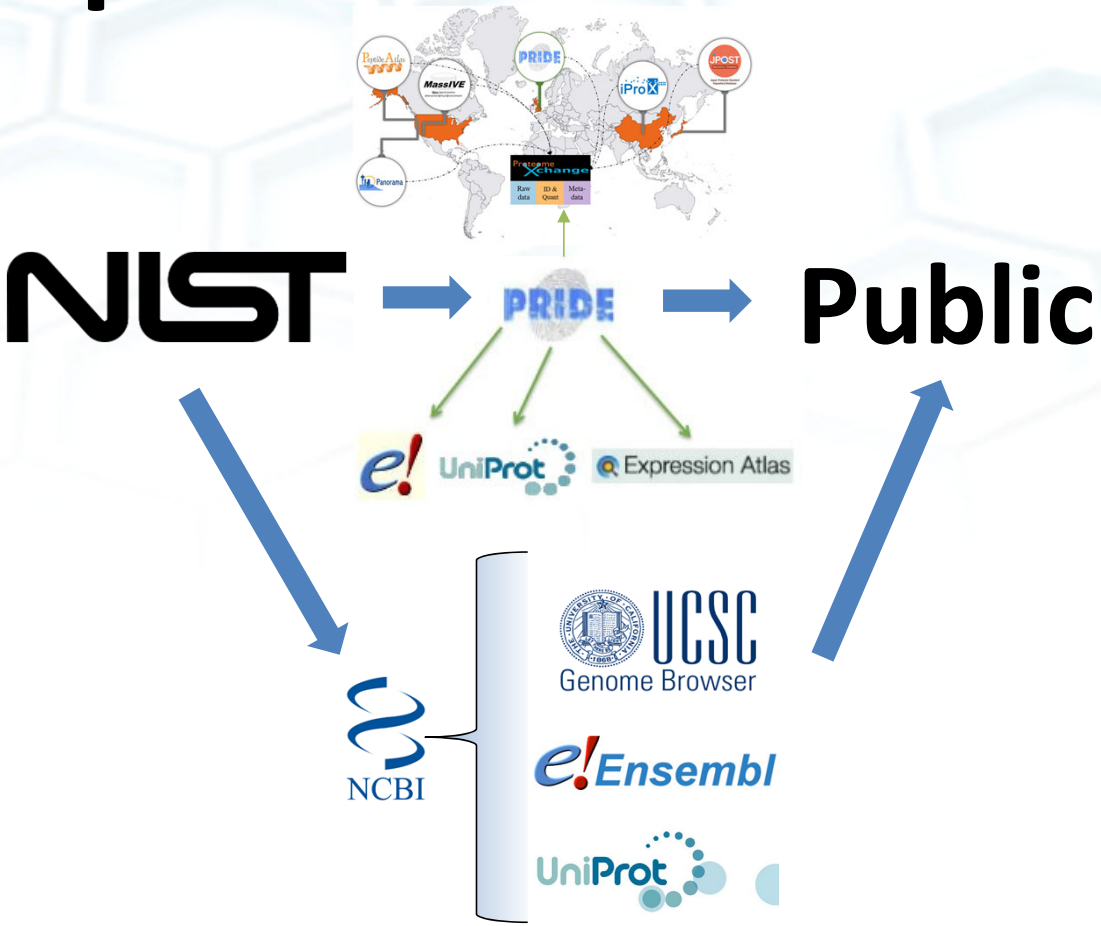




Complement Activation

Rank	Symbol	Name	log10 abundance
7	C3	Complement C3	7.4
20	CLU	Clusterin	6.8
21	SERPING1	Plasma protease C1 inhibitor	6.8
22	C9	Complement component C9	6.8
24	C4A	Complement C4-A	6.7
33	CFB	Complement factor B	6.6
34	CFH	Complement factor H	6.6
35	KRT1	Keratin, type II cytoskeletal 1	6.6
40	C6	Complement component C6	6.4
46	C8B	Complement component C8 beta chain	6.3
52	C5	Complement C5	6.2
55	C8A	Complement component C8 alpha chain	6.2
56	C7	Complement component C7	6.2
66	C4BPA	C4b-binding protein alpha chain	6.0
80	C8G	Complement component C8 gamma chain	5.8
81	C1QB	Complement C1q subcomponent subunit B	5.8
87	C1QC	Complement C1q subcomponent subunit C	5.8
98	C1R	Complement C1r subcomponent	5.6
100	CFI	Complement factor I	5.6
108	C1S	Complement C1s subcomponent	5.5
109	CFP	Properdin (Complement factor P)	5.5
114	C2	Complement C2	5.5
128	APOR	Apolipoprotein R*	5.3
131	MBL1	Mannose-binding protein A*	5.3
142	MBL2	Mannose-binding protein C	5.3
145	MASP1	Mannan-binding lectin serine protease 1	5.2
158	C1QA	Complement C1q subcomponent subunit A	5.1
212	MASP2	Mannan-binding lectin serine protease 2	4.8
235	COLEC11	Collectin-11	4.7
274	FCN1	Ficolin-1	4.6
284	FCN3	Ficolin-3	4.6

Data goes public

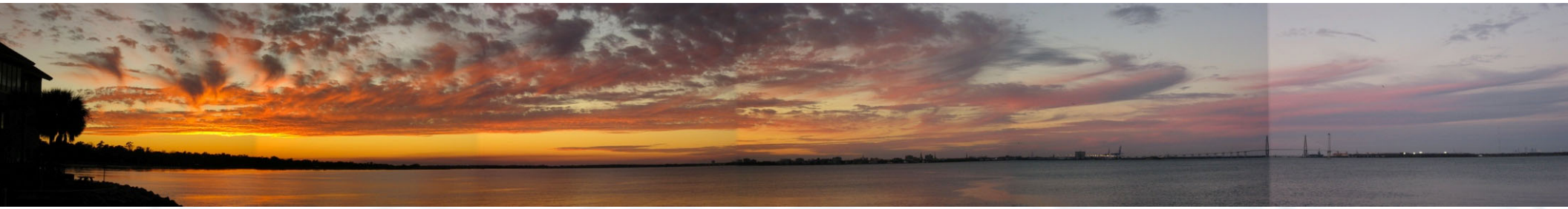


Summary

- Proteomics is the best
- The molecular landscape of most species is unknown and proteomics is uniquely positioned to survey, especially in biofluids
- We **can** generate and compare proteomic data from multiple mammal species
- More than “A v B” studies are required to limit false positives
- Clade-level comparisons are essential
  - maybe wild animals just look a certain way
  - OR maybe humans are the weird ones

Molecular cartography

Comparative Mammalian Proteome Aggregator Resource



Questions?

Alison Bland and Mike Janech

Magnus Palmblad

Phil Wilmarth

Daniel Becker