The power of 3 for shotgun proteomics: proteases, databases, and search engines.

Dr Delphine Vincent

01 May 2020

Introduction

Proteases

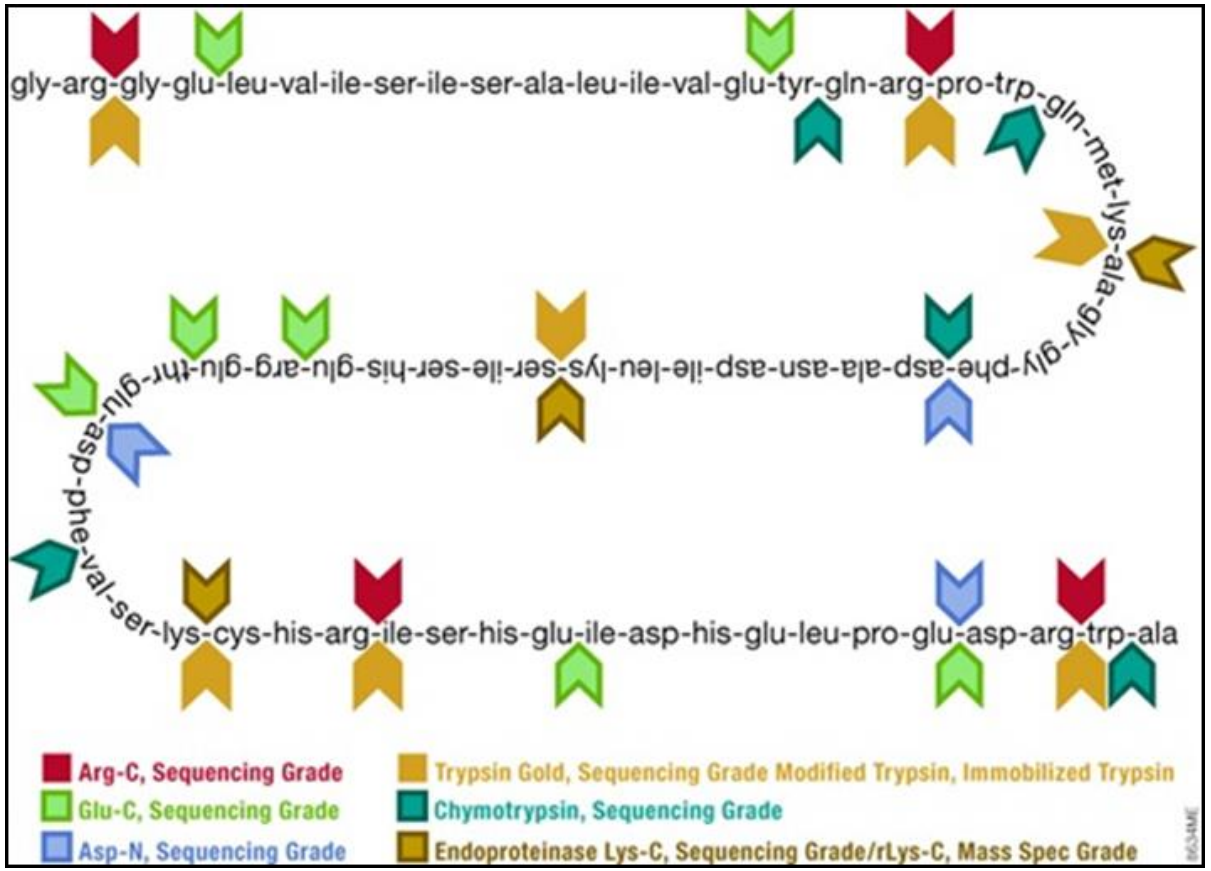
Multiple protease strategy to increase the proteome depth and sequence coverage.
Lots of proteases to choose from.
Orthogonal proteases yield complementary results

Proteases most widely used in proteomic analysis:

1.Trypsin (R, K) by far the most widely used protease in proteomic analysis.

2.Other proteases and cleavage reagents
- Glu-C (E)
- Lys-C (K)
- Chymotrypsin (F, W, Y)
- Asp-N (N)

3.Non specific proteases
- Subtilysin
- Pepsin (F, L)
- Proteinase K
- Pronase (mixture of endo- and exo-proteinases)
- Elastase (A, G, S, V)
- Thermolysine (I, M, F, W, Y, V)



Protease selectivity: the less residues targeted, the more selective the enzyme.
Protease orthogonality: different proteases cleave different AA residues.

Databases

A protein sequence database is required to match an acquired spectrum to its theoretical counterpart. The database comprises the AA sequences of all proteins that are expected in the sample. This is why specific protein databases arising from genome sequencing projects of the species of interest are ideal. However, if that is not readily available, sequences from a closely related species must be explored. Issues related to database search include variant protein, sequencing errors, or homologous protein from another species [Creasy et al 2002].

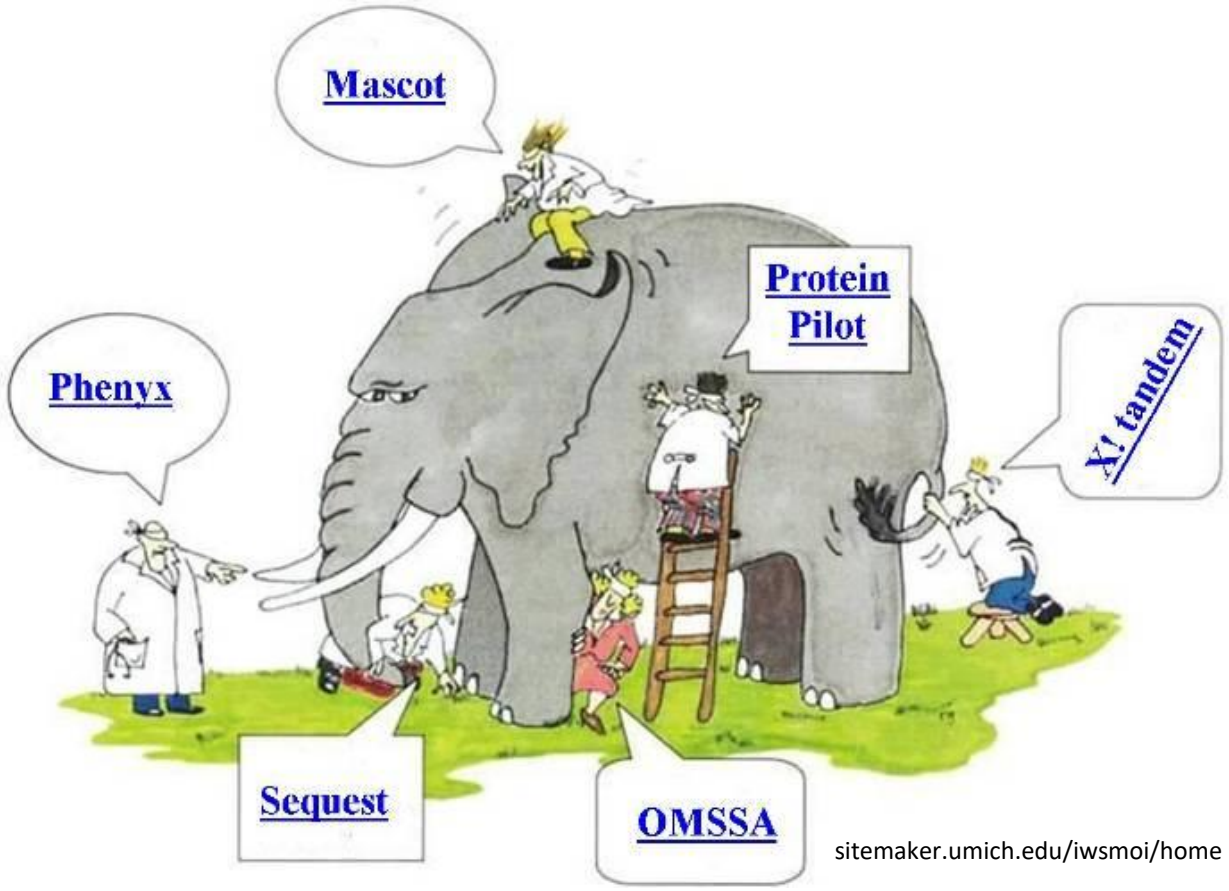
Several *Cannabis sativa* genomes have been sequenced [van Bakel et al 2011] [Grassa et al 2018] [Laverty et al 2018].

Number of genes varies from 27819 to 34589 [Kovalchuk et al 2020].

Database specificity: a specific database only contains sequences from the species of interest

Search engines

A plethora of search engines are available to the proteomics community to turn tandem mass spectra of peptides into AA sequences (Mascot, Masslynx, MS-Tag/MS-Seq, PeptideSearch, PepFrag, ProbID, SEQUEST, SpectrumMill, X!Tandem, Peaks, MaxQuant, InSpect, MSGFDB, OMSSA, MyriMatch, Paragon, ProteinPilot, Phenyx, PeptideProphet, Sonar, Andromeda, Protein Prospector, InsPecT, Morpheus, MS Amanda, SimSpectraST, ProLuCID, etc…).

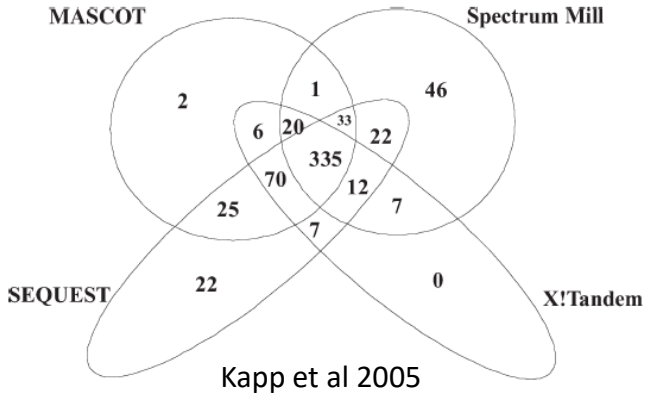


sitemaker.umich.edu/iwsmoi/home

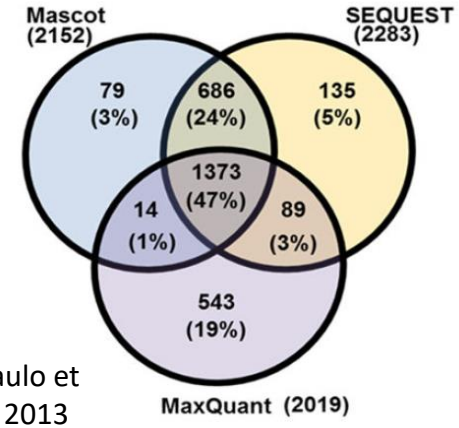
Same fundamental elements:

- read protein sequence databases
- emulate enzymatic cleavage to peptides
- extrapolate post-translational modifications (PTMs)
- apply a tolerance of observed precursor and fragment masses
- predict fragment ions for each peptide sequence
- compare observed and expected fragments

Shared results across search engines but specificity always reported.



Kapp et al 2005



Paulo et al 2013

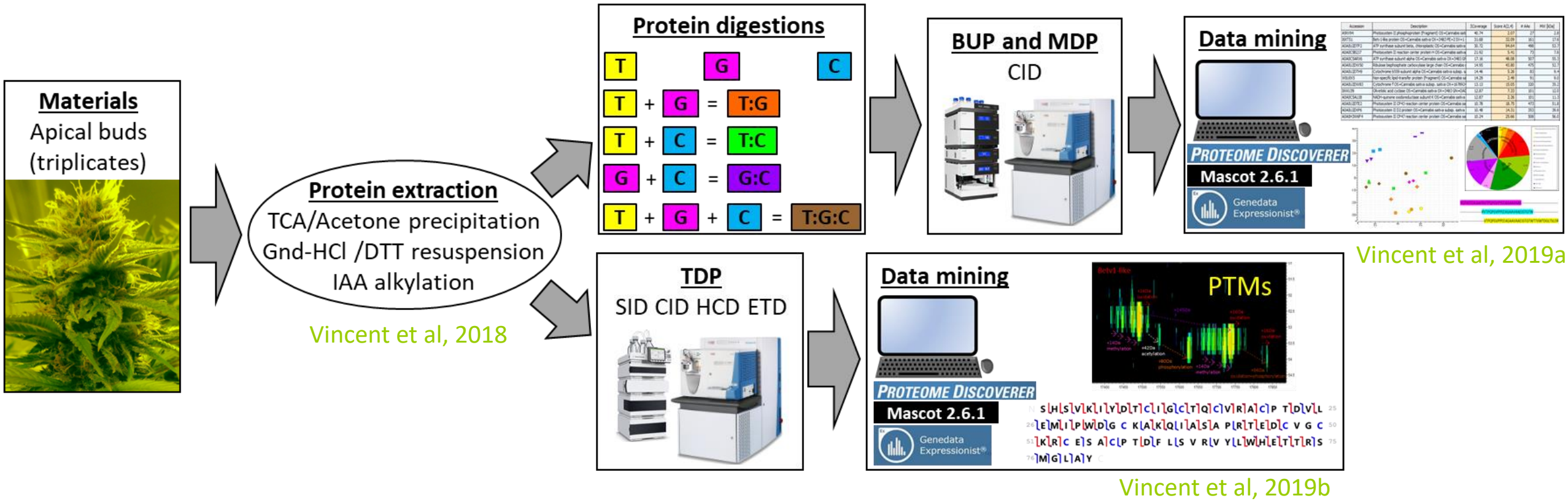
What has been achieved so far in cannabis proteomics at Agribio?

1st experiment: optimisation of protein extraction from mature buds for BUP (trypsin).

2nd experiment: optimisation of protein digestion for protein identification using BUP and MDP.

3rd experiment: optimisation of intact protein analysis and sequencing using TDP.

4th experiment: optimisation of database search.



Materials - Methods

Sample preparation and analysis

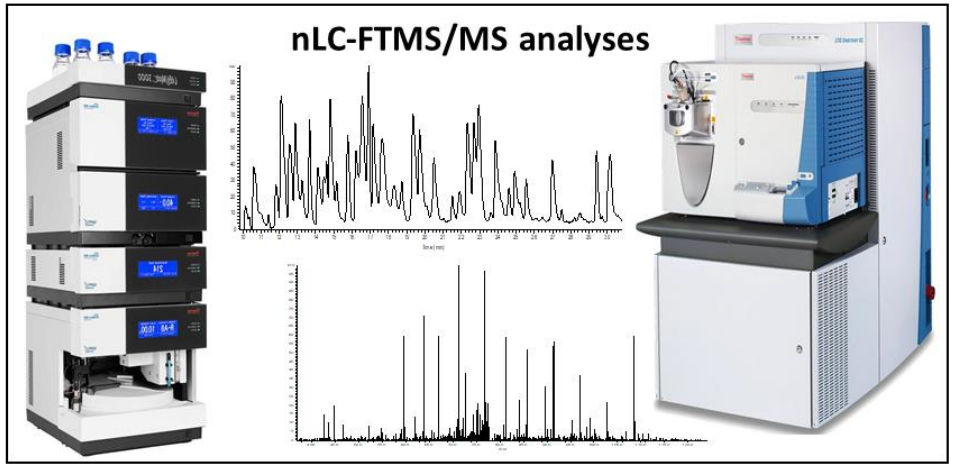
Apical buds (triplicates)

Protein extraction
TCA/Acetone precipitation
Gnd-HCl /DTTresuspension
IAA alkylation

Single-step protein digestion using 4 proteases

Protease name	*Protease code*	*AA targeted*	*Terminus targeted*	*Selectivity*	*Efficiency*
rAsp-N	A	D	N-term	High	85%
Chymotrypsin	C	F, W, Y	C-term	Low	<80%
Trypsin/Lys-C	TL	R, K	C-term	Intermediate	>90%

nLC-FTMS/MS analyses



Data analysis

PROTEOME DISCOVERER 1.4

Thermo SCIENTIFIC SEQUEST

MATRIX SCIENCE MASCOT

Genedata Expressionist®

DB searched	Size	Specificity
SP21	Small	High
Uniprot515	Small	High
JO29k	Intermediate	High
Homemade95k	Large	High
SPGP40k	Intermediate	Low

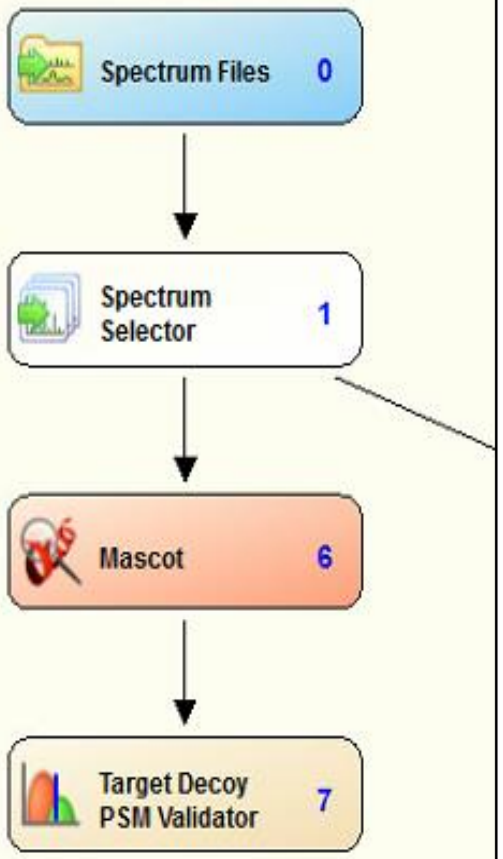
Five databases

Order	DB name	Sources	Nb entries	Date	Description	Algorithm	Taxonomy
1	SP21	https://www.uniprot.org/uniprot/?query=taxonomy:%22Rosales%20[3744]%22%20cannabis%20organism:sativa&fil=reviewed%3Ayes https://patentscope.wipo.int/search/en/detail.jsf?docId=WO2015196275 https://patentscope.wipo.int/search/en/detail.jsf?docId=WO2011017798&_cid=P11-K8DWCD-64087-1	19 from SwissProt + CBCAS (patent WO2015/196275) + GOT (patent WO2011/017798) = **21**	Feb 2020	Yes	SEQUEST MASCOT	C. sativa
2	Uniprot515	https://www.uniprot.org/uniprot/?query=taxonomy%3A%22Rosales+%5B3744%5D%22+cannabis+organism%3Asativa https://patentscope.wipo.int/search/en/detail.jsf?docId=WO2015196275 https://patentscope.wipo.int/search/en/detail.jsf?docId=WO2011017798&_cid=P11-K8DWCD-64087-1	513 from uniprot + 2 patents = **515**	Feb 2020	Yes	SEQUEST MASCOT	C. sativa
3	JO29k	https://www.cannabisdraftmap.org/	**29,057**	Dec 2019	Yes	SEQUEST *	C. sativa
4	Homemade95k	https://www.uniprot.org/uniprot/?query=taxonomy%3A%22Rosales+%5B3744%5D%22+cannabis+organism%3Asativa https://www.ncbi.nlm.nih.gov/protein (cannabis sativa) AND "Cannabis sativa"[porgn: txid3483] http://medicinalplantgenomics.msu.edu/pub/data/MPGR/Cannabis_sativa/	513 + 2 patents + 37,143 + 57,411 = **95,069**	Feb 2020	Yes Yes No	SEQUEST MASCOT	C. Sativa
5	SPGP40k	https://www.uniprot.org/uniprot/?query=reviewed:yes%20taxonomy:33090	**39,800**	Feb 2020	Yes	SEQUEST MASCOT	Green plants

* JO29k cannot be parsed in Mascot due to duplicate rows.

Data file search

Proteome Discoverer 1.4



MASCOT

1. Input Data	
Protein Database	
Enzyme Name	
Maximum Missed Cleavage Sites	9
Instrument	ESI-TRAP
Taxonomy	
1.1 Peptide Scoring Options	
Peptide Cut Off Score	10
Peptide Without Protein Cut Off Score	5
1.2 Protein Scoring Options	
Use MudPIT Scoring	Automatic
Protein Relevance Threshold	20
Protein Relevance Factor	1
2. Tolerances	
Precursor Mass Tolerance	10 ppm
Fragment Mass Tolerance	0.8 Da
Use Average Precursor Mass	False
3. Modification Groups	
From Quan Method	
4. Dynamic Modifications	
1. Dynamic Modification	Acetyl (N-term)
2. Dynamic Modification	Oxidation (M)
3. Dynamic Modification	Acetyl (K)
4. Dynamic Modification	Methyl (K)
5. Dynamic Modification	Phospho (ST)
6. Dynamic Modification	Phospho (Y)
7. Dynamic Modification	NAG (N)
8. Dynamic Modification	
9. Dynamic Modification	
5. Static Modifications	
1. Static Modification	Carbamidomethyl (C)
2. Static Modification	
3. Static Modification	
4. Static Modification	
5. Static Modification	
6. Static Modification	

Enzyme Name: AspN, Chymotrypsin, Trypsin

Protein Database: SP21, Uniprot515, JO29k, Homemade95k, SPGP40k

SEQUEST

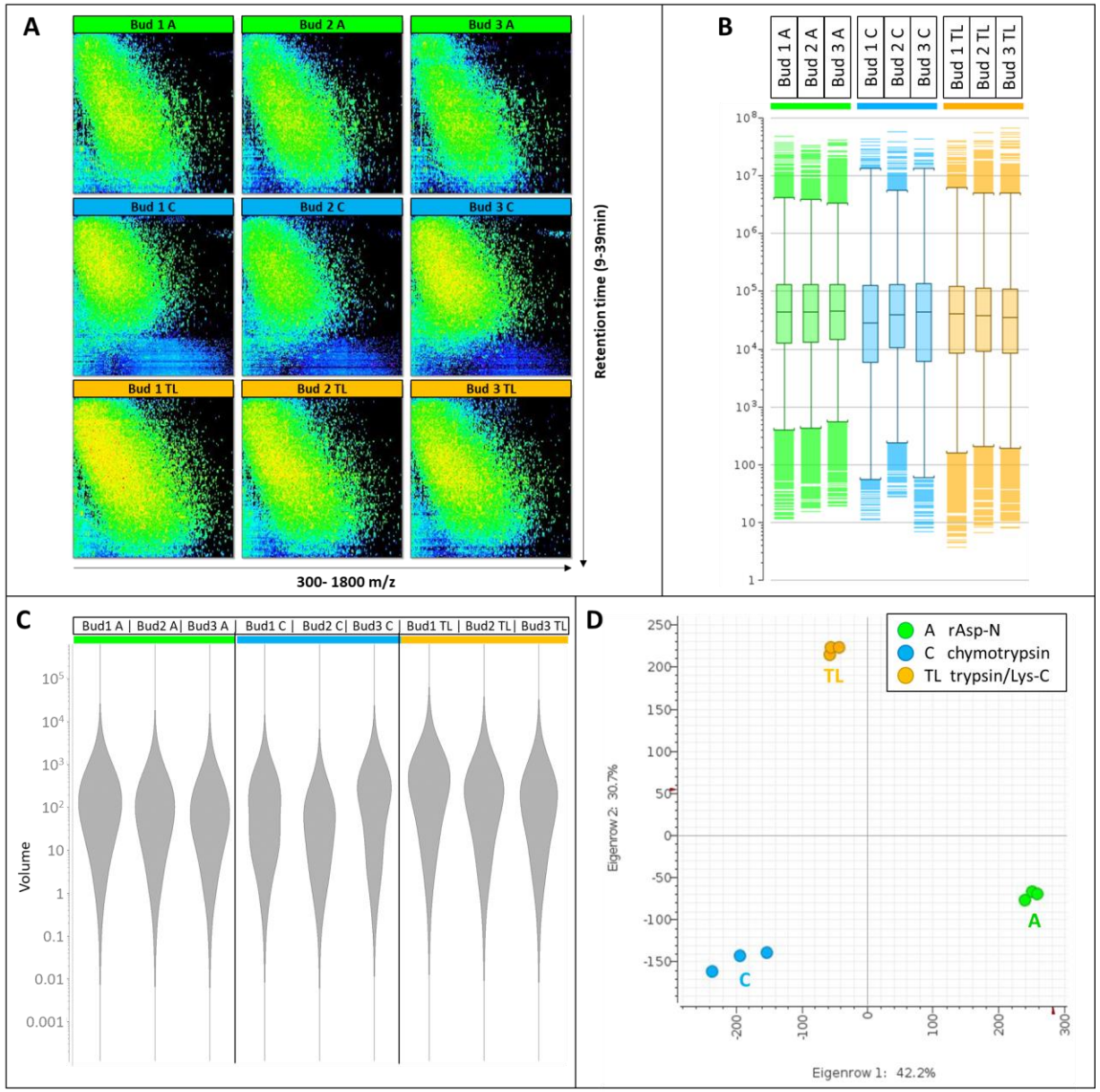
1. Input Data	
Protein Database	
Enzyme Name	
Maximum Missed Cleavage Sites	12
1.1 Peptide Scoring Options	
Maximum Peptides Considered	500
Maximum Peptides Output	10
Calculate Probability Scores	False
Absolute XCorr Threshold	0.4
Fragment Ion Cutoff Percentage	0.1
Peptide Without Protein XCorr Threshold	1.5
1.2 Protein Scoring Options	
Maximum Protein References Per Peptide	100
Protein Relevance Threshold	1.5
Peptide Relevance Factor	0.4
2. Tolerances	
Precursor Mass Tolerance	10 ppm
Fragment Mass Tolerance	0.8 Da
Use Average Precursor Mass	False
Use Average Fragment Masses	False
3. Ion Series	
Use Neutral Loss a Ions	True
Use Neutral Loss b Ions	True
Use Neutral Loss y Ions	True
Weight of a Ions	0
Weight of b Ions	1
Weight of c Ions	0
Weight of x Ions	0
Weight of y Ions	1
Weight of z Ions	0
4. Dynamic Modifications	
Max. Modifications Per Peptide	4
N-Terminal Modification	Acetyl / +42.011 Da (Any N-Terminus)
C-Terminal Modification	None
1. Dynamic Modification	Oxidation / +15.995 Da (M)
2. Dynamic Modification	N-acetyl-D-glucosamine / +221.090 Da (N)
3. Dynamic Modification	Methyl / +14.016 Da (K)
4. Dynamic Modification	Acetyl / +42.011 Da (K)
5. Dynamic Modification	Phospho / +79.966 Da (S, T, Y)
6. Dynamic Modification	None
5. Static Modifications	
Peptide N-Terminus	None
Peptide C-Terminus	None
1. Static Modification	Carbamidomethyl / +57.021 Da (C)
2. Static Modification	None
3. Static Modification	None
4. Static Modification	None
5. Static Modification	None

Enzyme Name: AspN, Chymotrypsin, Trypsin

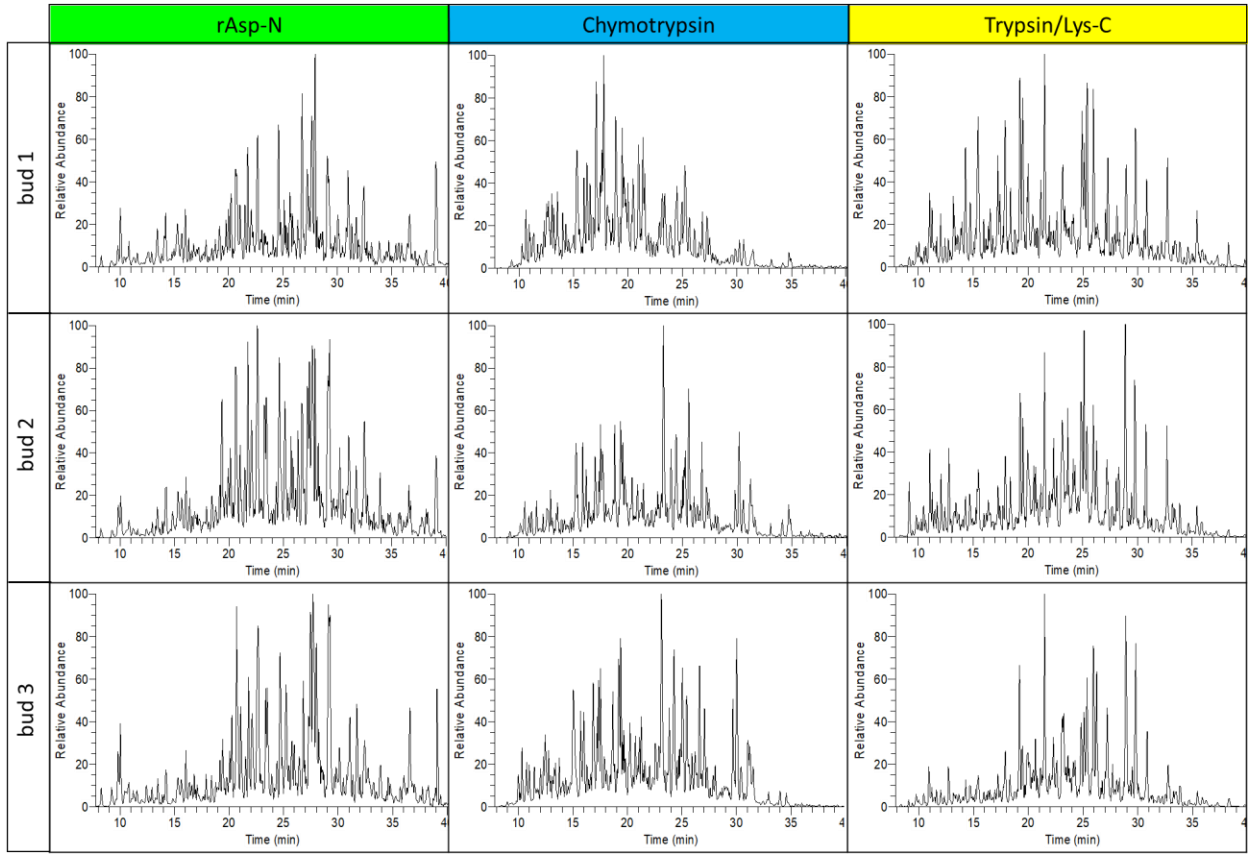
Protein Database: SP21, Uniprot515, JO29k, Homemade95k, SPGP40k

Results

Statistical analyses

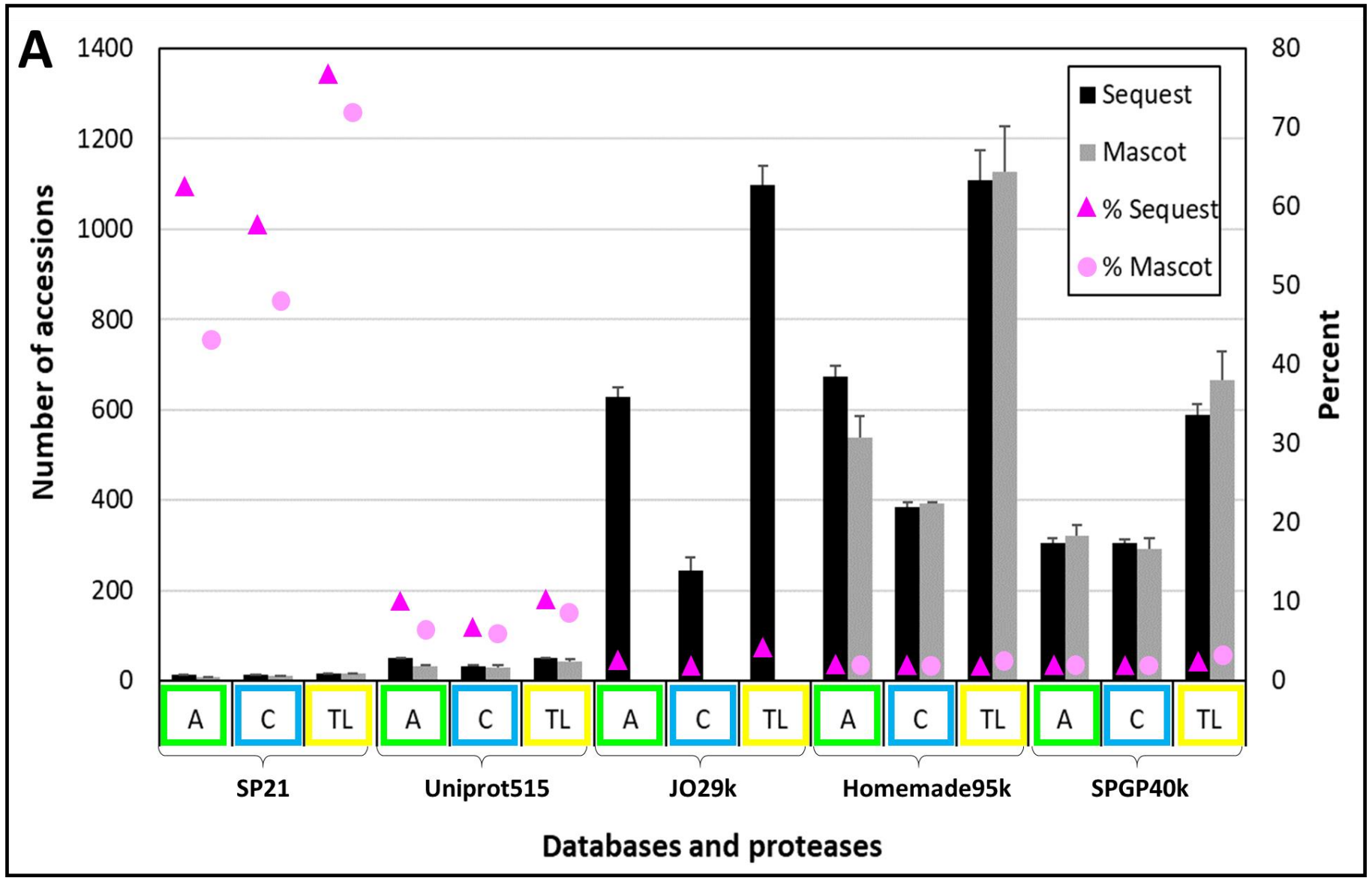


TICs



Good reproducibility
Protease patterns very different (triangle on PCA plot).
TL pattern display more peaks.

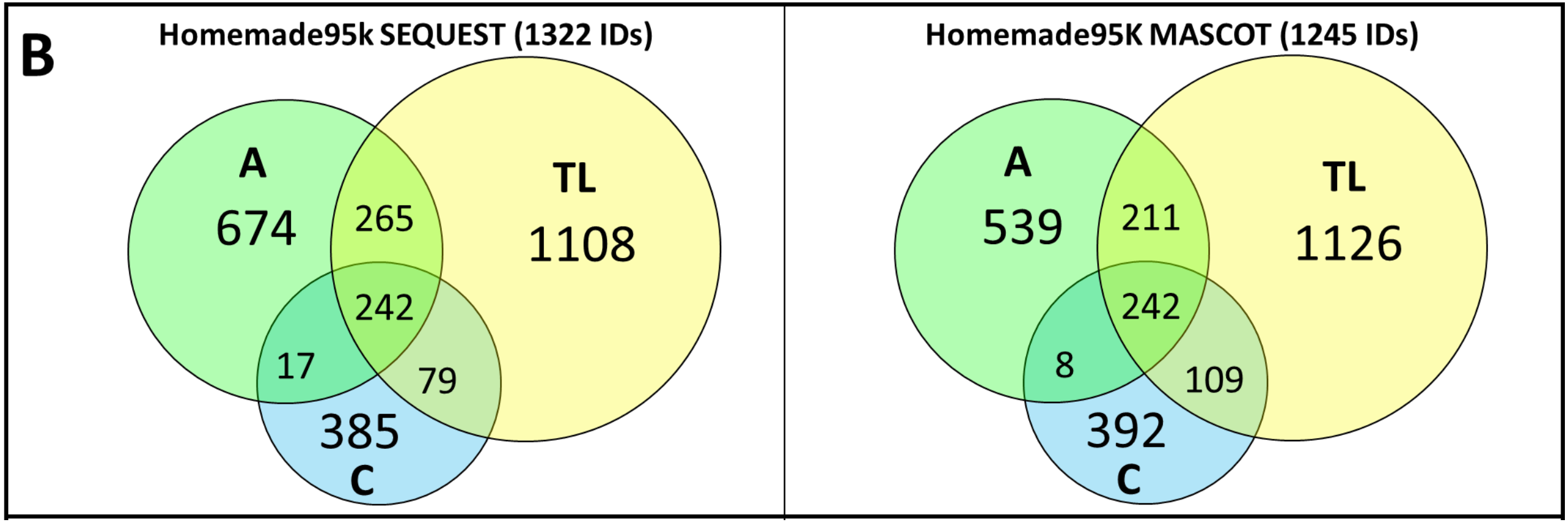
Number of IDs



The larger the database the longer the list of proteins identified.

Opposite trend with % relative to database size (pink dots).

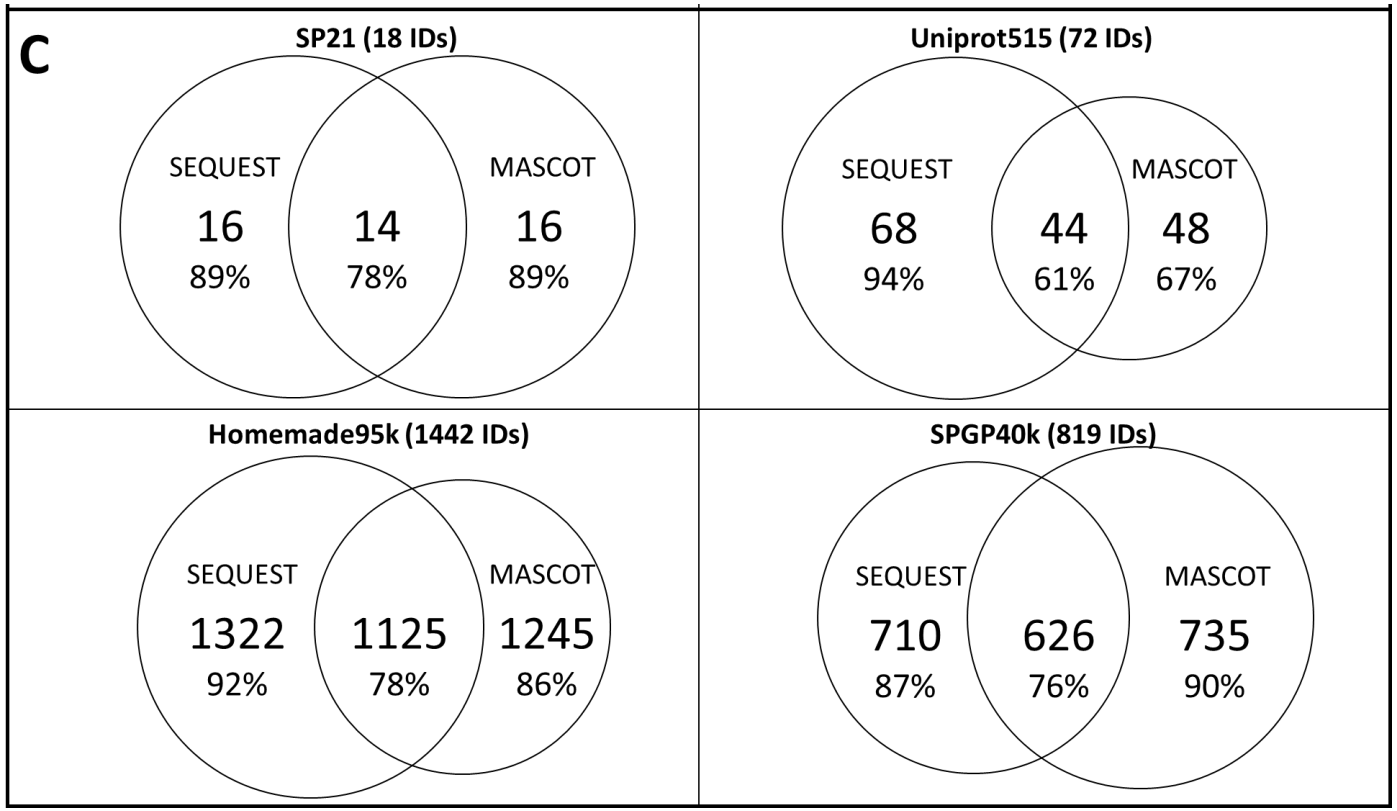
Proteases

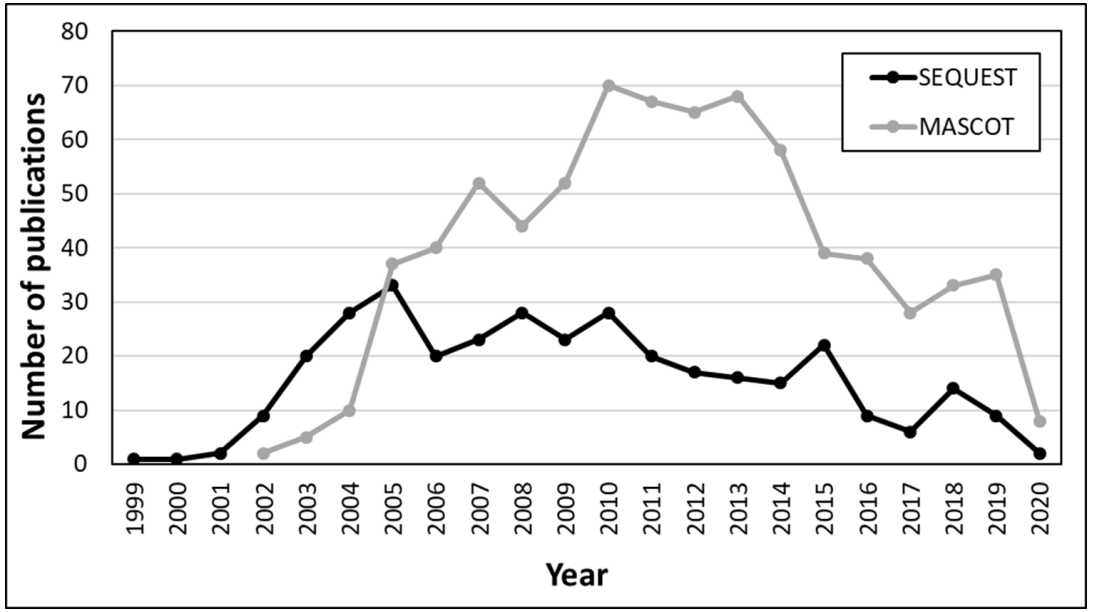


Number of IDs: TL > A > C

Some shared accessions across proteases but also mostly complementary.

Search algorithms

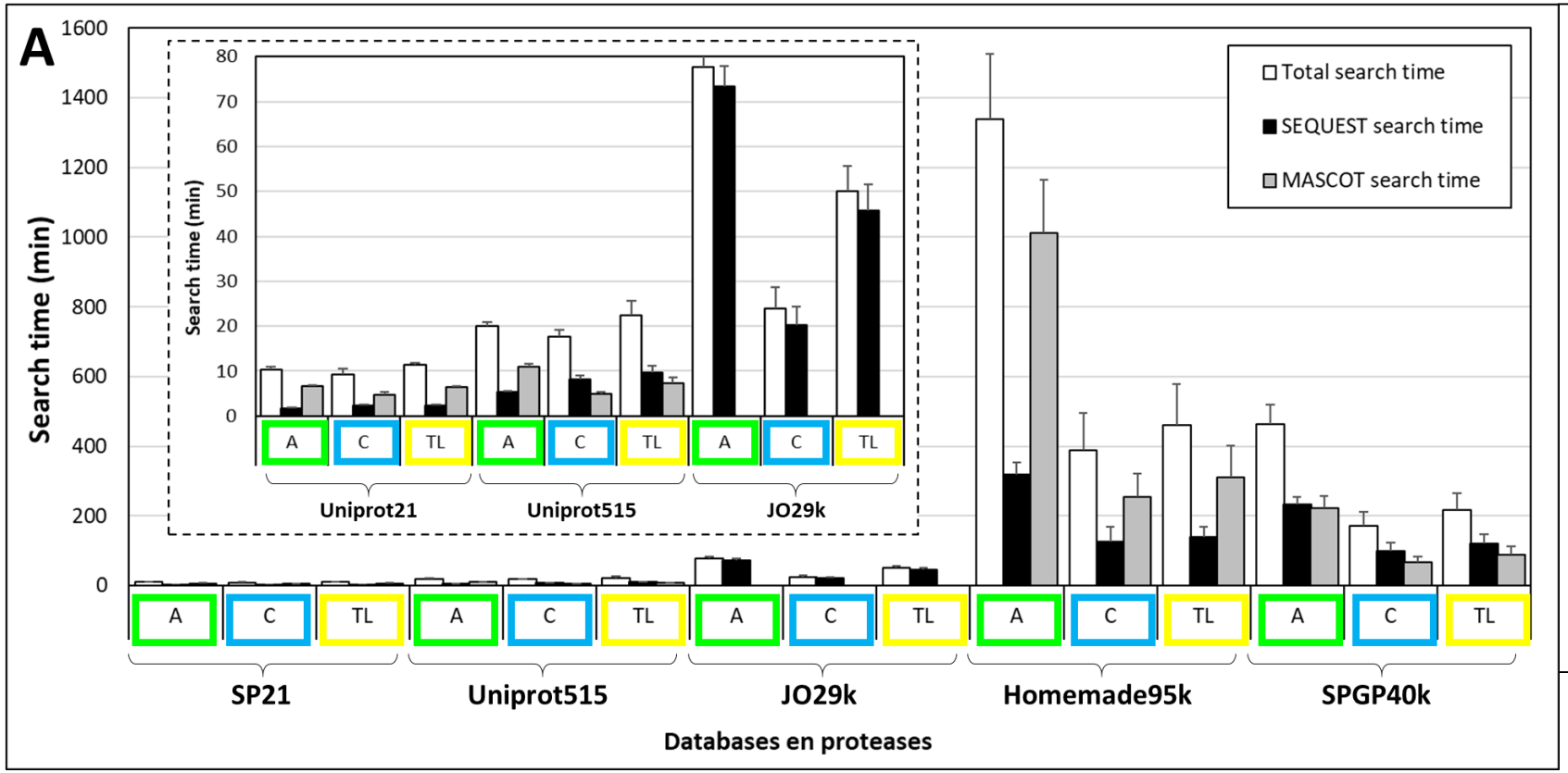


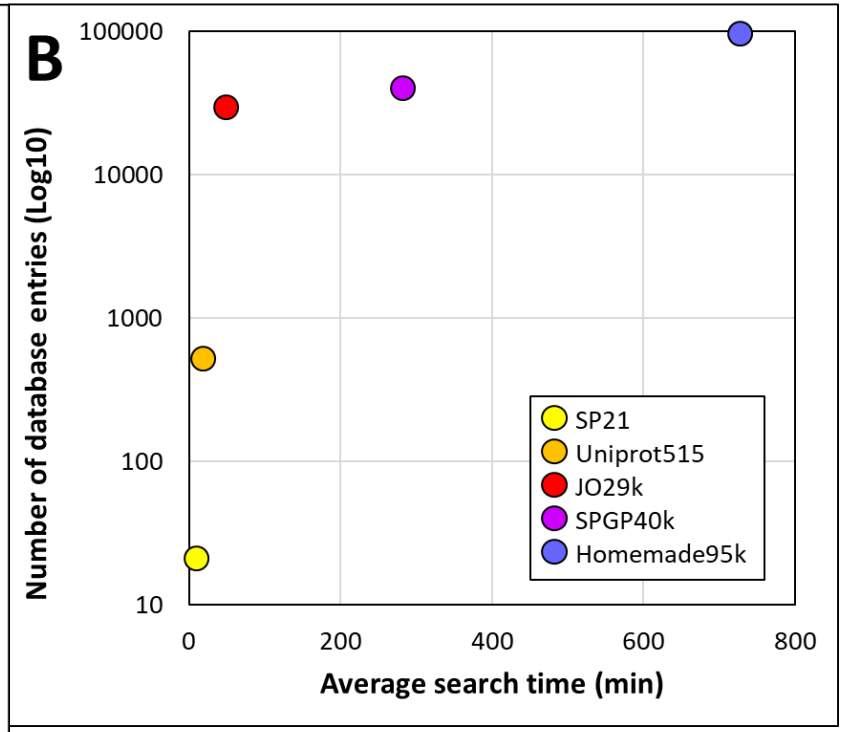


Even though SEQUEST (1994) predates Mascot (1999), the latter is more popular.

SEQUEST yields more IDs than Mascot.
Most accessions are common across both algorithms but many are unique to each search engine.

Search times





The larger the database, the longer the search.
The type of protease also affects the duration of the search (Asp-N takes longer).
SEQUEST is faster than Mascot.

Miscleavages and peptide size

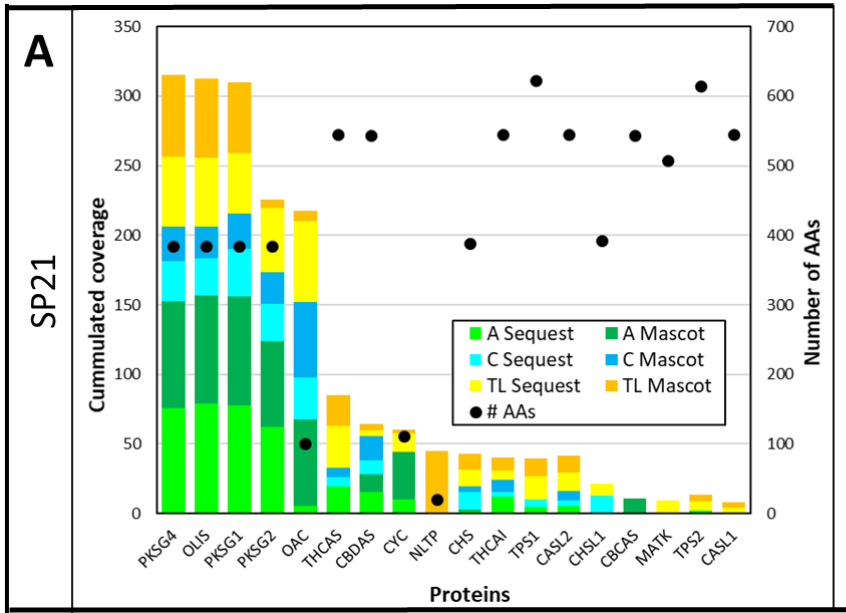
# miscleavage	SP21	Uniprot515	JO29k	Homemade95k	SPGP40k
0	116	433	2822	5818	2060
1	33	95	282	1091	403
2	20	51	32	339	140
3	7	16	13	158	60
4	8	9	5	54	28
5	1	1	6	22	7
6	4	3	4	8	5
7	2	3	1	8	4
8	1	0	3	5	1
10	1	0	1	1	1
TOTAL	193	611	3169	7504	2709
TOTAL miscleavage=0	116	433	2822	5818	2060
TOTAL miscleavage>0	77	178	347	1686	649
% miscleavage>0	39.9	29.1	10.9	22.5	24.0
ELPD	39	255	2475	4132	1411

A. Peptide length	Uniprot21	Uniprot515	JO29k	Homemade95k	SPGP40k
min	626.4	626.4	969.5	604.3	604.3
max	7600.9	6385.2	6724.5	6993.1	6448.6
average	2123.2	2023.2	2173.6	1975.8	1866.0
SD	1099.7	1048.9	791.1	830.3	776.8

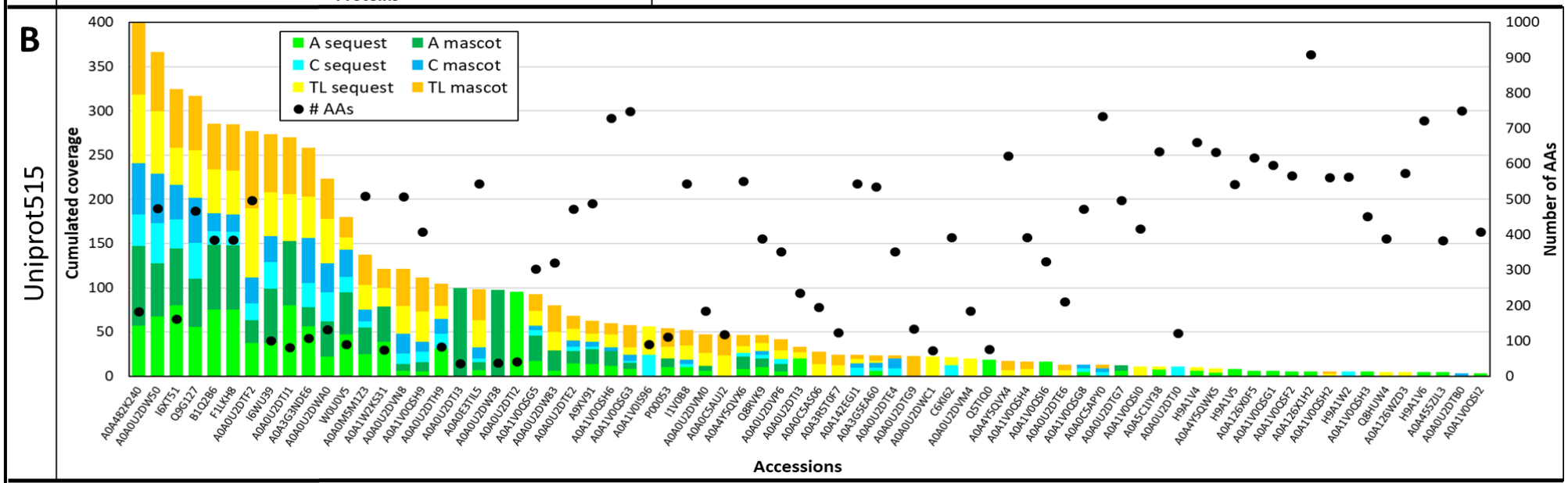
B. Protease	Database	min mass	max mass	average mass	SD mass
A	Uniprot21	1006.6	7600.9	2475.2	1166.7
A	Uniprot515	631.3	5994.1	2363.4	1192.1
A	JO29k	969.5	6724.5	2280.9	905.8
A	Homemade95k	653.4	6375.2	2147.2	939.1
A	SPGP40k	653.4	6448.6	2028.9	929.2
C	Uniprot21	774.4	5520.9	1807.1	927.0
C	Uniprot515	704.4	5520.9	1779.1	793.0
C	JO29k	1034.6	6061.9	2108.9	776.2
C	Homemade95k	789.5	6954.3	1901.9	724.2
C	SPGP40k	789.5	5121.4	1832.0	581.4
TL	Uniprot21	626.4	5303.5	2007.0	1058.9
TL	Uniprot515	626.4	6385.2	1926.4	1015.7
TL	JO29k	1055.5	6369.2	2112.1	705.8
TL	Homemade95k	604.3	6369.2	1922.4	789.4
TL	SPGP40k	604.3	6369.2	1795.0	706.0

Up to 10 miscleavages but mostly 0-3.
Peptide length: 0.6-7.6 kD, average 2 kD.
Asp-N produces longer peptides (average 2.2 kD) than TL and C (average 1.9 kD).
Long peptides carry more miscleavages.

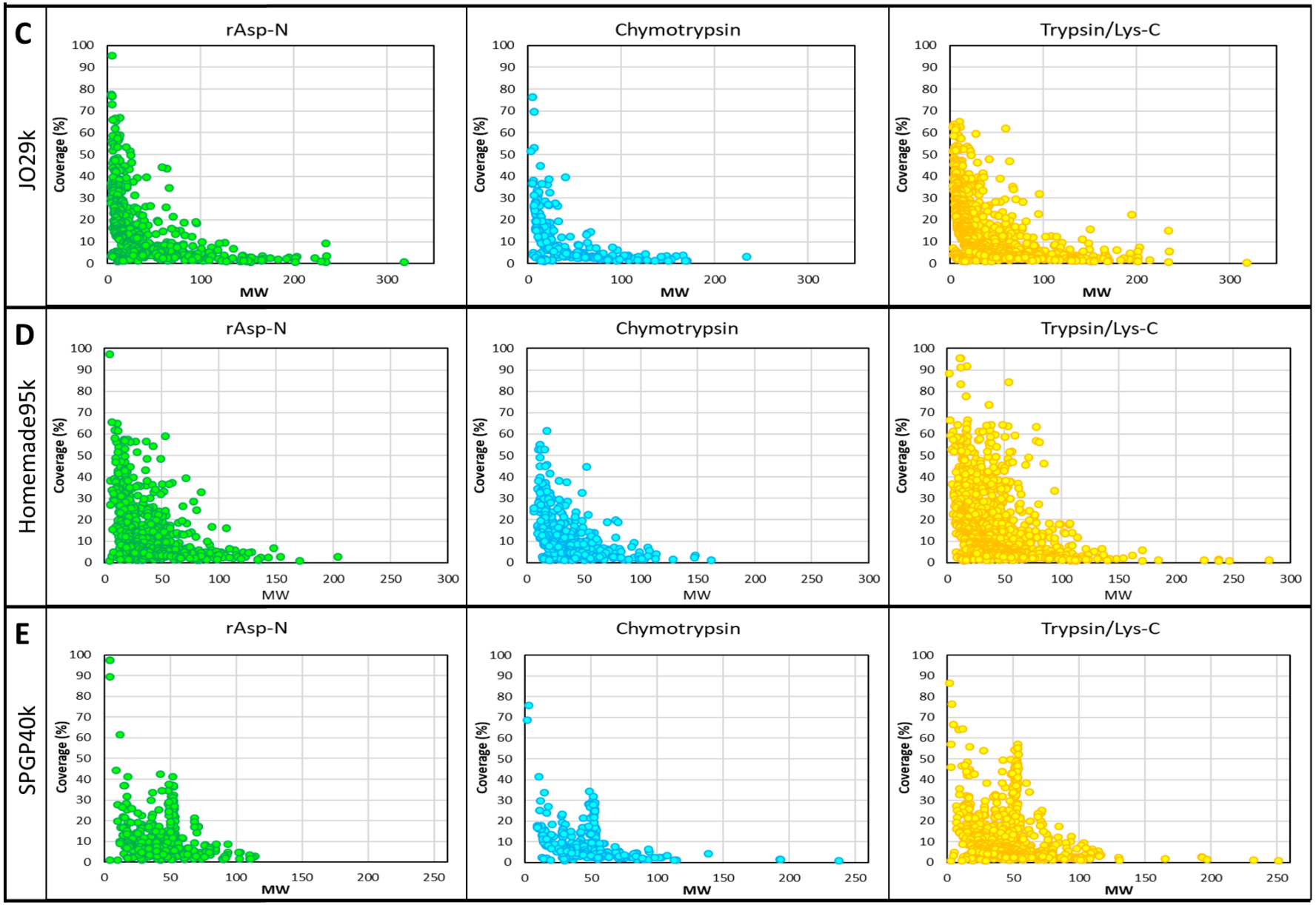
Sequence coverage



Short proteins achieve greater sequence coverage (e.g. OAC 101AAs).
Coverage ranking per protease: TL >= Asp-N > C



Sequence coverage



Short proteins achieve greater sequence coverage is confirmed with the largest databases.

Similar trend across all proteases.

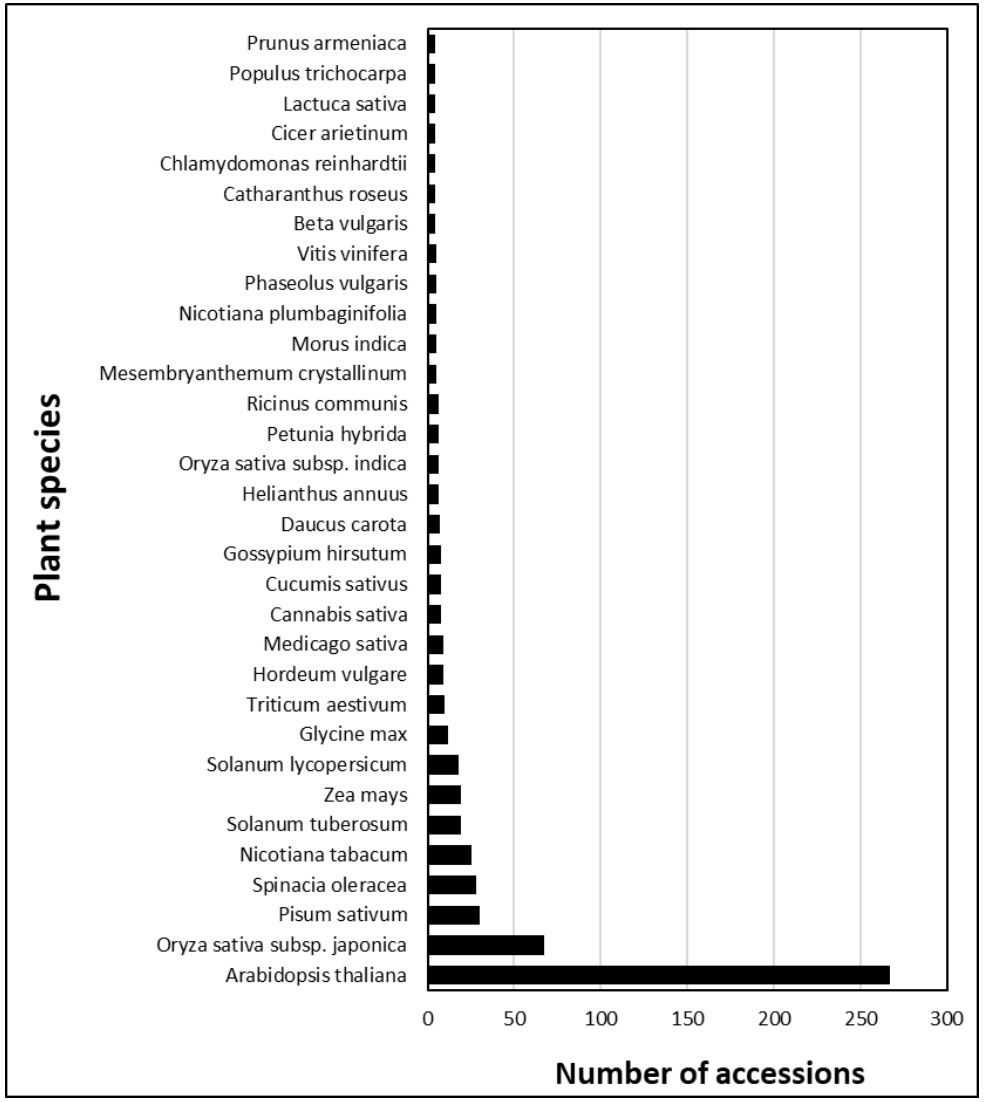
Negative relationship between protein MW and sequence coverage.

Post-translational modifications

PTM	Uniprot21	Uniprot515	JO29k	Homemade95k	SPGP40k
Carbamidomethyl (C)	34	94	493	602	226
N-term acetyl (K)	21	16	27	91	44
Acetyl (K)	47	32	47	132	71
Methyl (K)	61	49	114	163	158
NAG (N)	10	5	9	17	7
Oxidation (M)	18	24	43	66	90
Phospho (STY)	86	57	100	201	71
TOTAL PTMs	277	277	833	1272	667
# identified peptides	344	611	3169	7504	2709
# unmodified peptides	192	450	2255	5593	1834
# modified peptides	152	161	914	1911	875
% modified peptides	44.2	26.4	28.8	25.5	32.3

Cannabis proteins are heavily modified.

Database specificity and Gene Ontology (GO)



A. 72 IDs from Uniprot515

- biological_process (50 results)
  - transport (6 results)
  - metabolic process (48 results)
    - nitrogen compound metabolic process (21 results)
    - biosynthetic process (33 results)
      - ATP synthesis coupled proton transport (5 results)
    - secondary metabolic process (2 results)
    - electron transport chain (7 results)
    - cellular metabolic process (45 results)
    - primary metabolic process (34 results)
    - small molecule metabolic process (10 results)
    - organic substance metabolic process (36 results)
  - cellular process (46 results)
  - response to stimulus (2 results)
  - biological regulation (2 results)

B. 819 IDs from SPGP40k

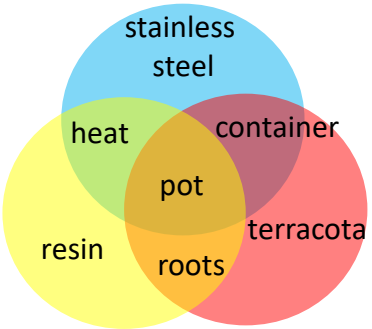
- biological_process (713 results)
  - immune system process (5 results)
  - cell adhesion (1 results)
  - circadian rhythm (5 results)
  - metabolic process (581 results)
    - protein glycosylation (3 results)
    - NADH metabolic process (4 results)
    - NADP metabolic process (7 results)
    - nitrogen compound metabolic process (340 results)
    - catabolic process (128 results)
    - biosynthetic process (330 results)
    - secondary metabolic process (7 results)
      - phenylpropanoid metabolic process (3 results)
      - phenylacetate catabolic process (1 results)
      - glycosinolate metabolic process (2 results)
      - secondary metabolite biosynthetic process (5 results)
      - olivetolic acid biosynthetic process (1 results)
    - methylation (6 results)
    - pigment metabolic process (16 results)
    - hormone metabolic process (3 results)
    - cellular metabolic process (537 results)
    - primary metabolic process (452 results)
    - small molecule metabolic process (224 results)
    - ATP metabolic process (76 results)
    - oxidation-reduction process (41 results)
    - organic substance metabolic process (508 results)
  - cellular process (628 results)
  - pollen tube guidance (1 results)
  - reproductive process (19 results)
  - killing of cells of other organism (1 results)
  - multicellular organismal process (32 results)
  - developmental process (40 results)
  - multicellular organism reproduction (1 results)
  - growth (5 results)
  - response to stimulus (174 results)
  - localization (77 results)
  - multi-organism process (34 results)
  - biological regulation (116 results)
  - cellular component organization or biogenesis (85 results)
  - detoxification (8 results)

The non-specific database (SPGP40k) helps identify homologous proteins which can then be categorise using GO. More annotations mean deeper insight into the biology of the system of interest.

Conclusions

- **Proteases**: the more the better but if you had to choose one, pick trypsin. It delivers every time!
- **Databases**: size does matter! The larger the better! If possible, chose a specific database.
- **Search engines**: I think the more the better, but the jury is still out on this!
- **Results published:** https://doi.org/10.3390/proteomes8020013

What do this 3 items have in common?



CULTURE VICTORIA


Thank you!