

GENCODE EN-TE_x Proteogenomics

Functional Proteomics - Institute of Cancer Research



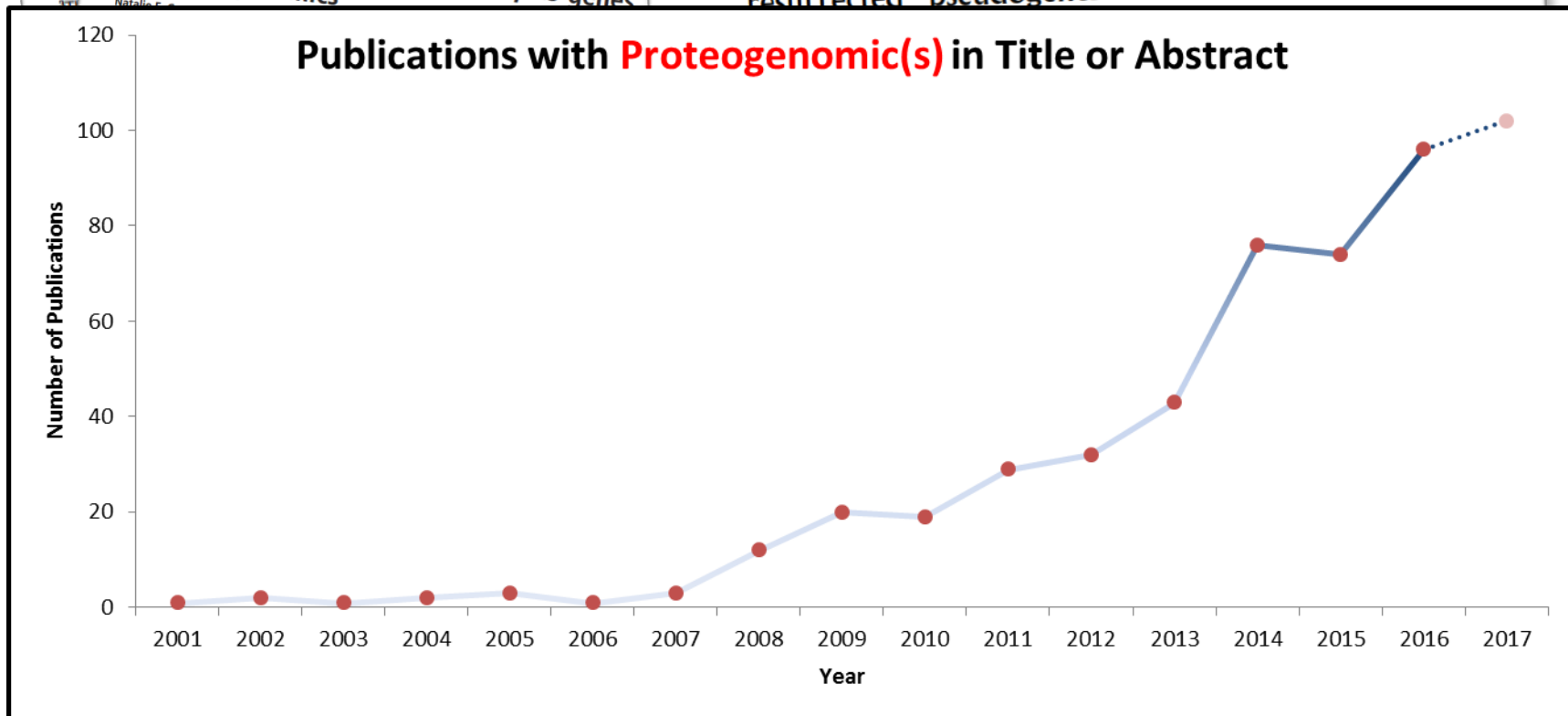
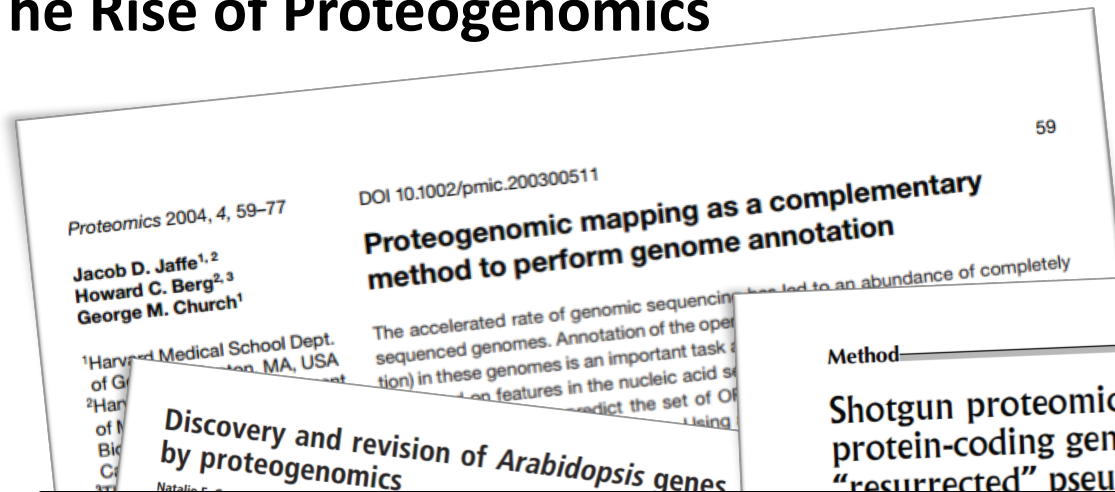
James Wright



Jyoti Choudhary

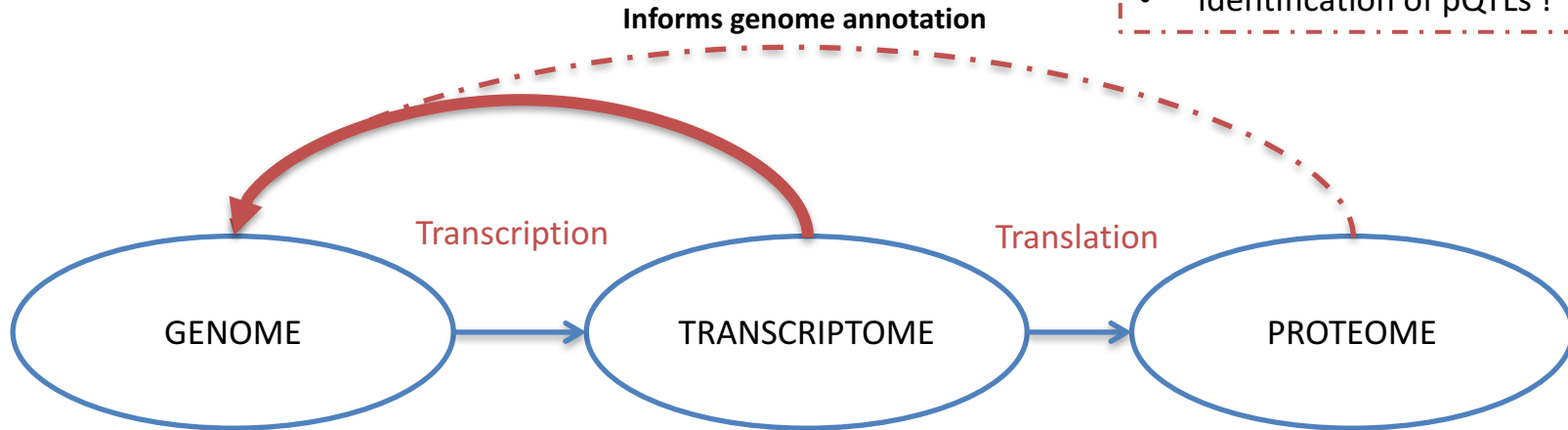


The Rise of Proteogenomics



What is Proteogenomics? ...good for?

- Identification of Novel Genes
- Confirms Protein Coding Potential
- Confirms Protein Coding Isoform
- Refinement of Gene Structure
- Identification of pQTLs !



- Static
- Genes
- Genotype
- Variation and Mutation
- QTLs (Quantitative Trait Loci)

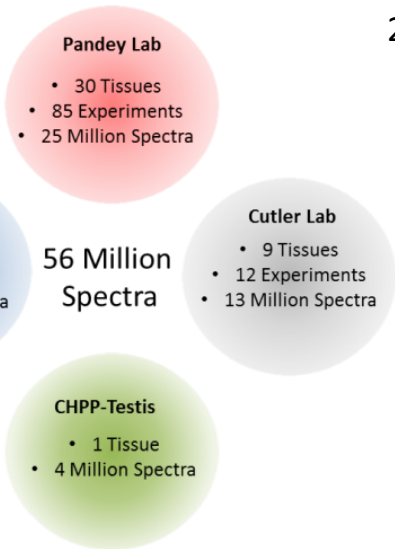
- Dynamic
- Transcripts
- Alternative Splicing
- Abundance
- Direct Regulation
- Noisy

- Super Dynamic
- Proteins
- Isoforms
- Abundance
- **Modification**
- **Interactions and Complexes**
- **Degradation**
- Direct and **Indirect Regulation**

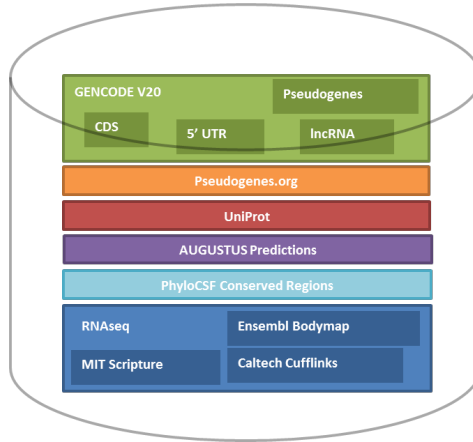
Only Moderate (0.5) Correlation Between Transcriptome and Proteome

Improving GENCODE reference gene annotation using a proteogenomics workflow

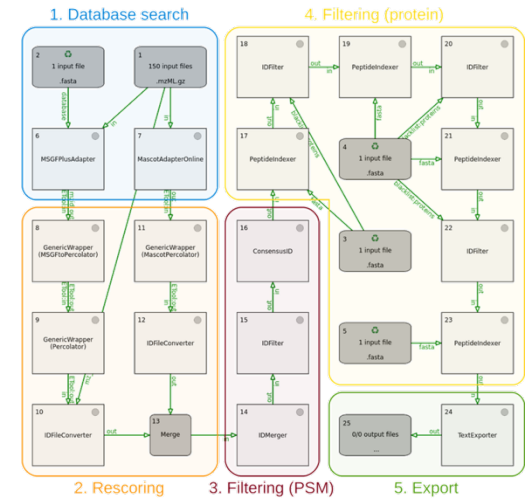
1. Datasets



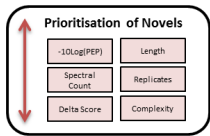
2. Search Space



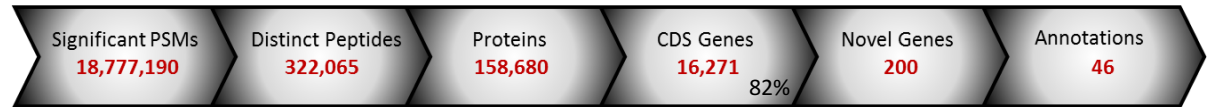
3. Analysis Pipeline



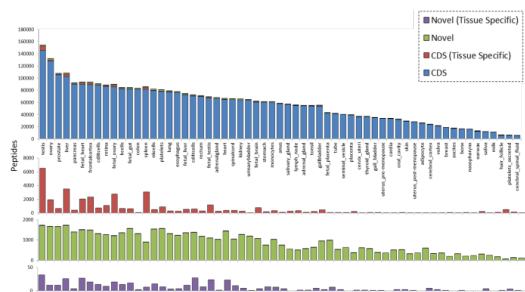
3. Prioritisation



4. Identifications



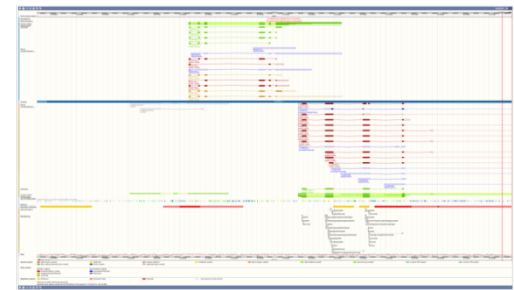
5. Tissue Analysis



5. Annotation



6. Genome Mapping



ftp://ngs.sanger.ac.uk/scratch/project/team17/msproteomicshub_significance/hub.txt

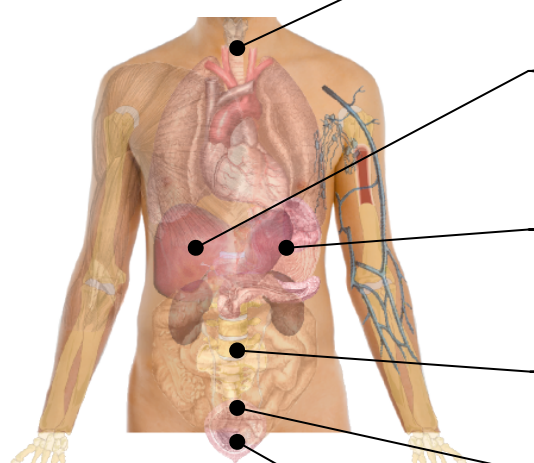
Nat Commun. 2016 Jun 2;7:11778.

Improving GENCODE reference gene annotation using a high-stringency proteogenomics workflow. Wright JC, Mudge J, Weisser H, Barzine MP, Gonzalez JM, Brazma A, Choudhary JS, Harrow J.

J Proteome Res. 2016 Dec 2;15(12):4686-4695. Epub 2016 Nov 10. Flexible Data Analysis Pipeline for High-Confidence Proteogenomics. Weisser H, Wright JC, Mudge JM, Gutenbrunner P1, Choudhary JS.

EN-TEt Tissue Samples

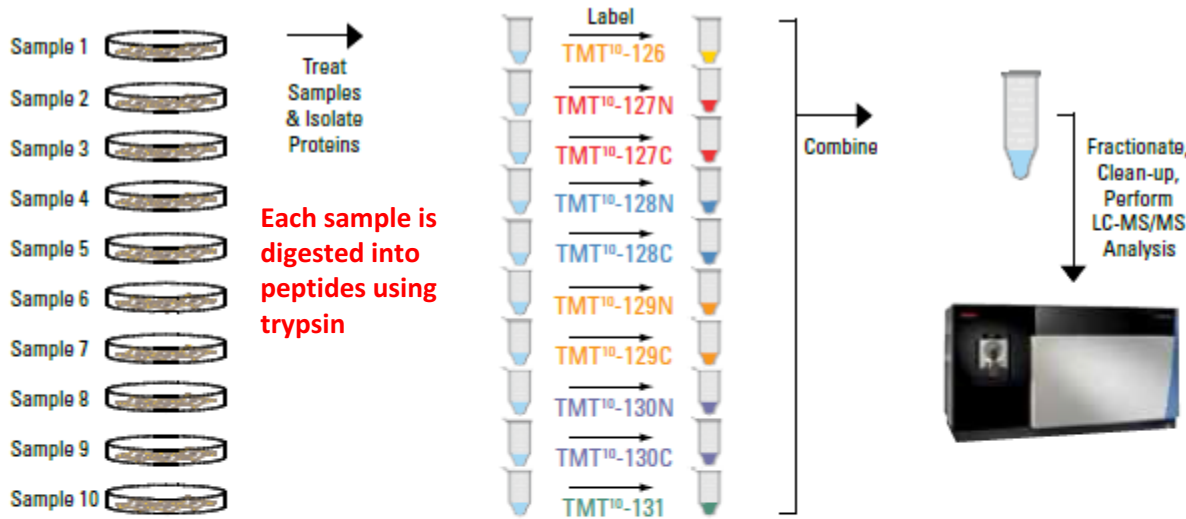
- 10 samples in 6 tissues across 4 donors



Sample	Donor	Sex	Age
Esophagus	ENC-003	female	53
Esophagus	ENC-004	female	51
Liver	ENC-003	female	53
Liver	ENC-003	female	53
Spleen	ENC-003	female	53
Spleen	ENC-004	female	51
Small Intestine	ENC-003	female	53
Prostate	ENC-001	male	37
Testis	ENC-001	male	37
Testis	ENC-002	male	54

- RNAseq and full genome data available for all samples

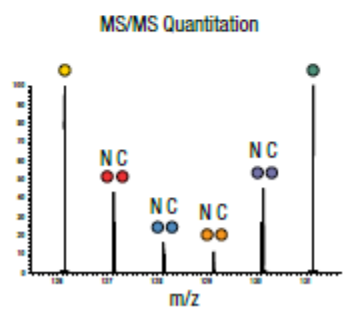
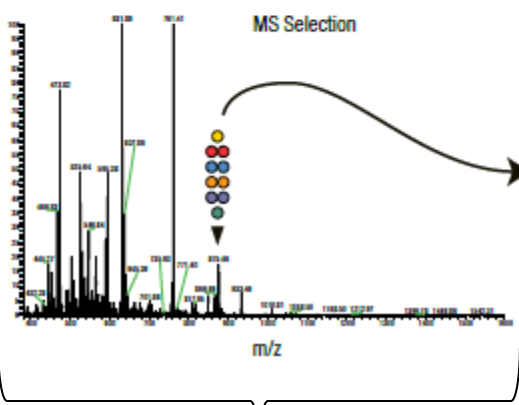
TMT Labelling Method



Each sample is digested into peptides using trypsin

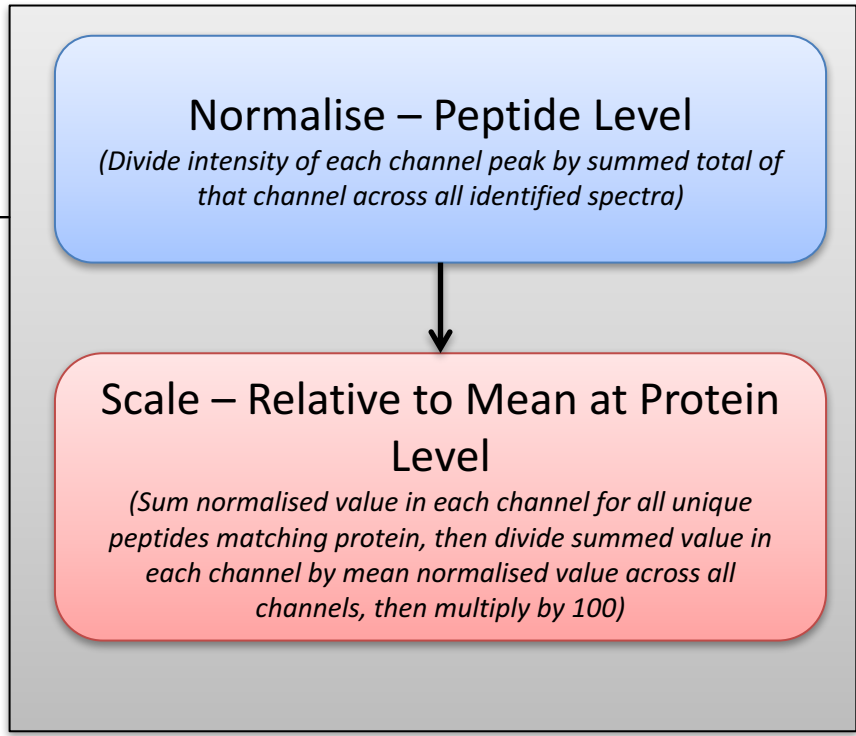


Lu Yu



Extract intensity of each quantitation peak

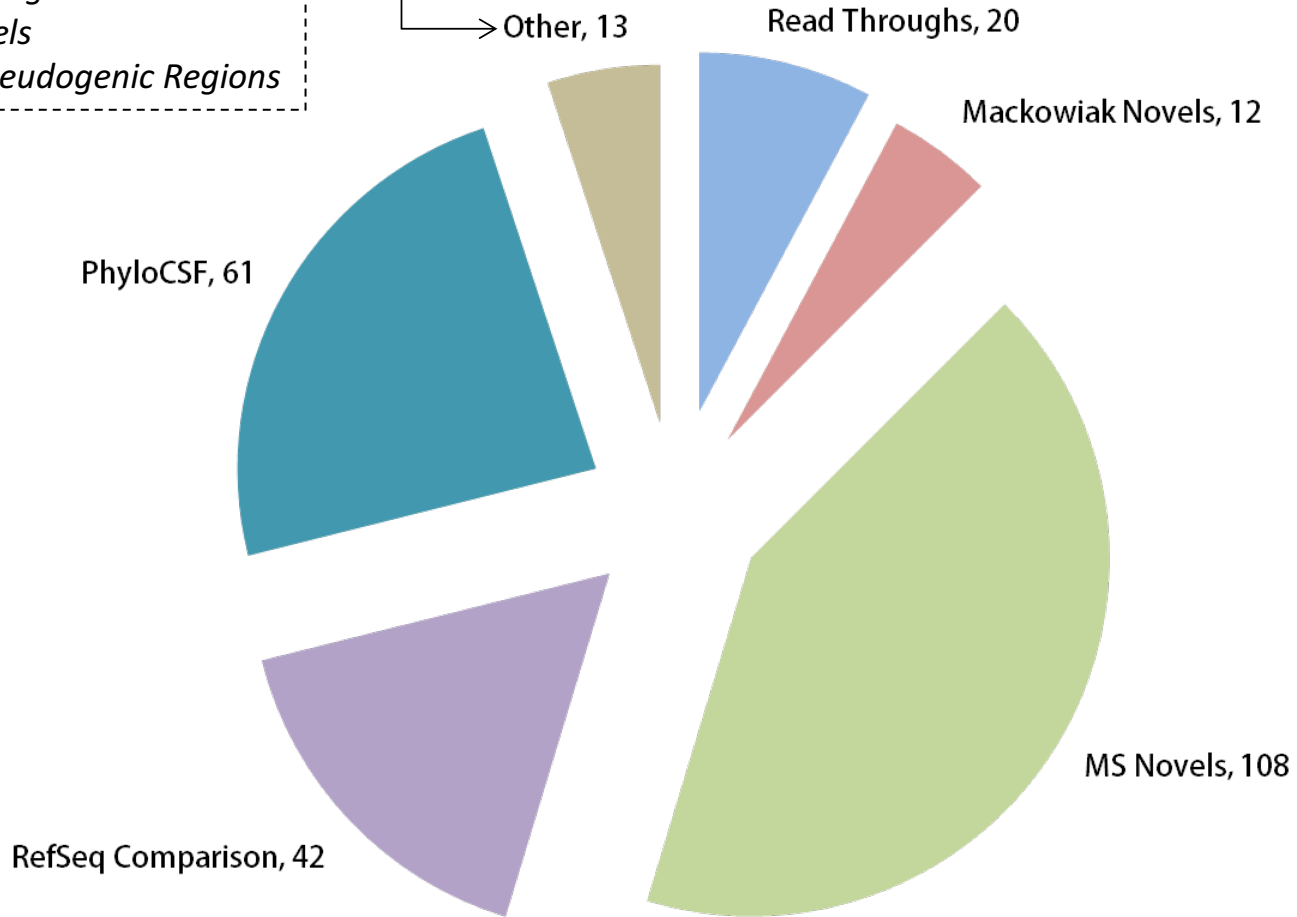
Identify Spectra
(Compare spectra to theoretical spectra generated from sequence database)



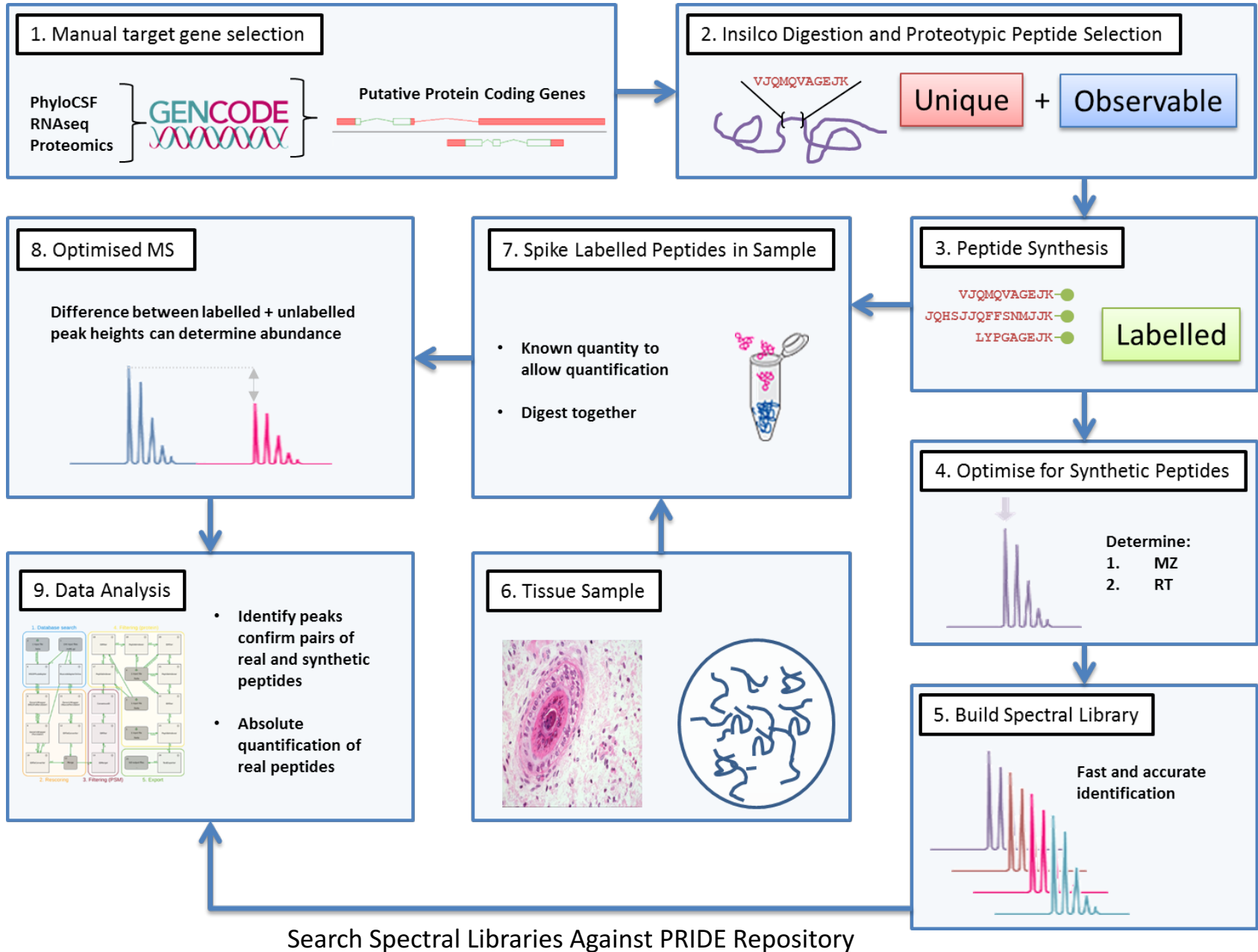
Synthetic Peptide Sources

256 Labelled Peptides Generated

- UniProt Comparison*
- Unitary Pseudo Genes*
- EIEE Novel Regions*
- Ulitsky Novels*
- CDS with Pseudogenic Regions*



Targeted Search Method



GENCODE Proteogenomics Sequence Sources

Type	Database
Known Proteins	GENCODE CDS UniProt SwissProt/Trembl NextProt* Known RefSeq (N)* Vega Update*
Pseudogenes	GENCODE Pseudogenes Yale Pseudogenes.org
InRNA	GENCODE InRNA Rory PacBio IncRNAs*
Speculative	GENCODE 5'UTR GENCODE 3'UTR* AUGUSTUS Halfwise* Model RefSeq (X)* PhyloCSF*
RNAseq	Jose SLRseq* Ulitsky (PLAR) RNAseq* Iyer RNAseq* Ensembl BodyMap
Contaminates	CRap HLA Sequences
Other	Synthetic Peptides* Sample Specific RNAseq and predicted variants* DecoyPYrat - Decoy Sequences

Pre-search Analysis

Collapse Redundant Sequences

Re-index sequences and generate non-redundant search database

In silico Digest Sequences

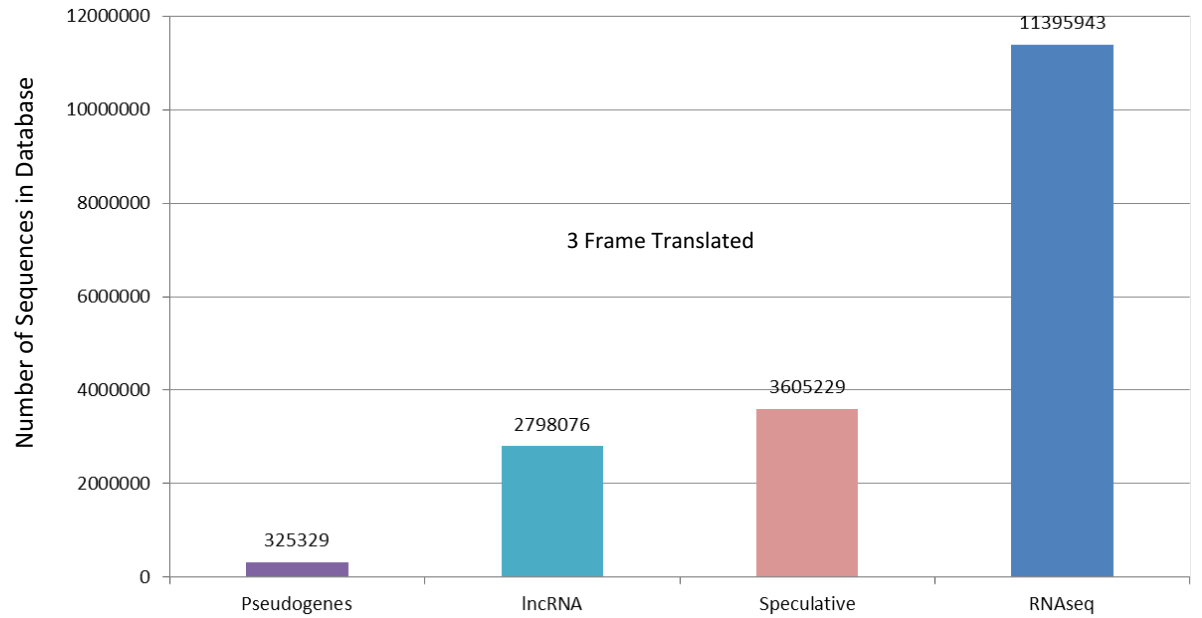
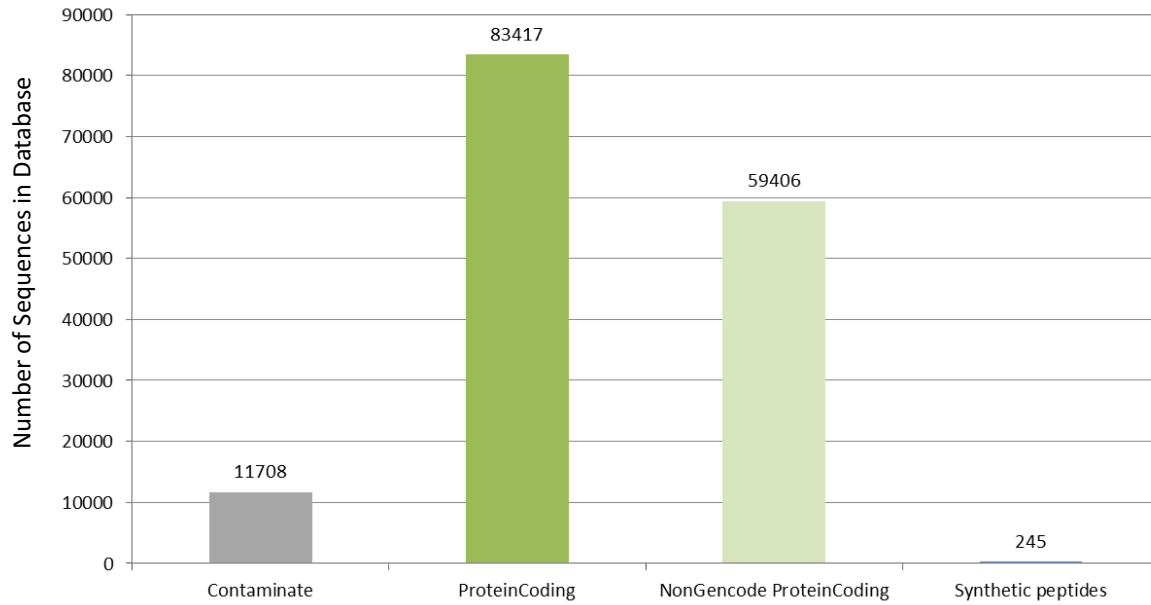
Generate table of gene / transcript specific peptides

3 Frame Translated De Novo Assembly of RNAseq from each sample

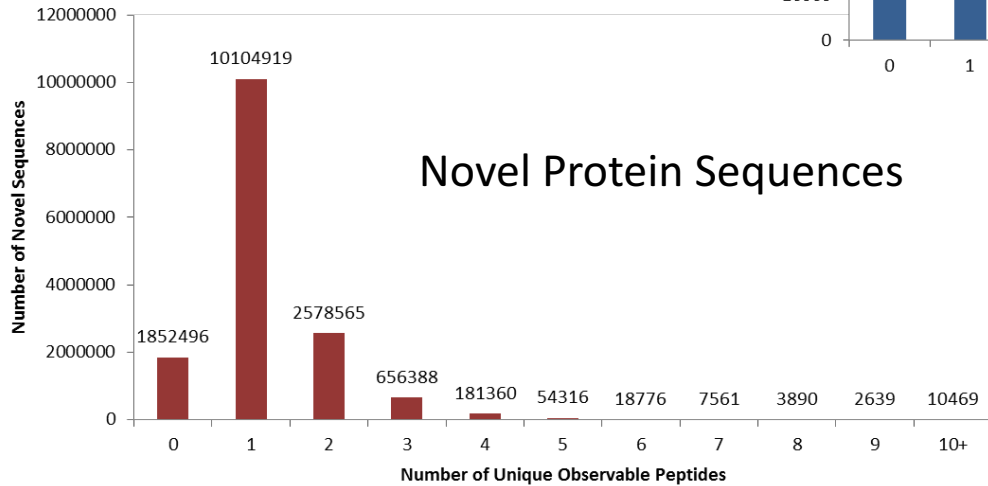
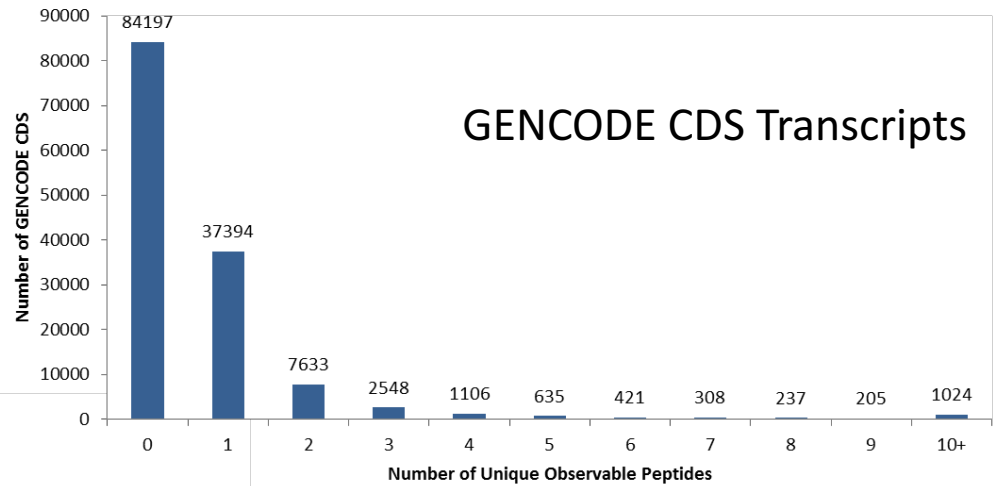
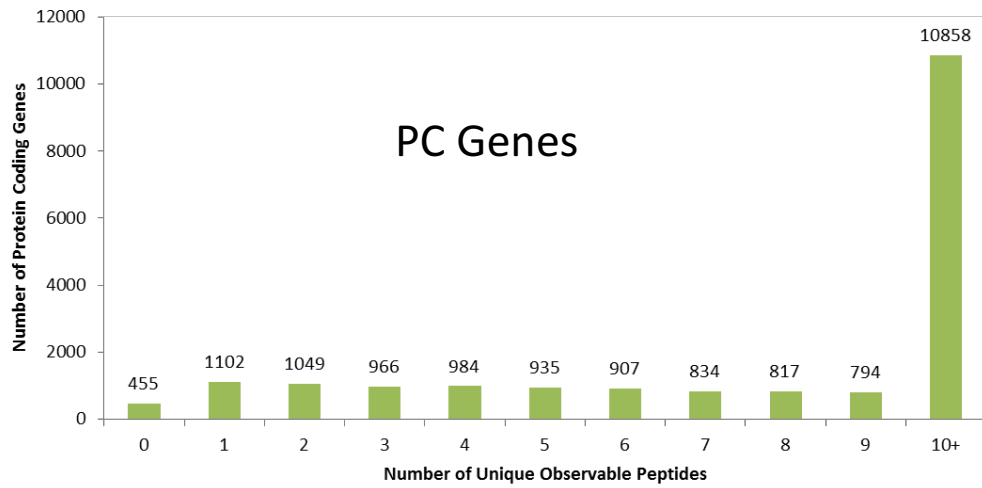
Sequence variants called and mapped using VEP (Variant Effect Predictor)

* New sequence sources not previously used

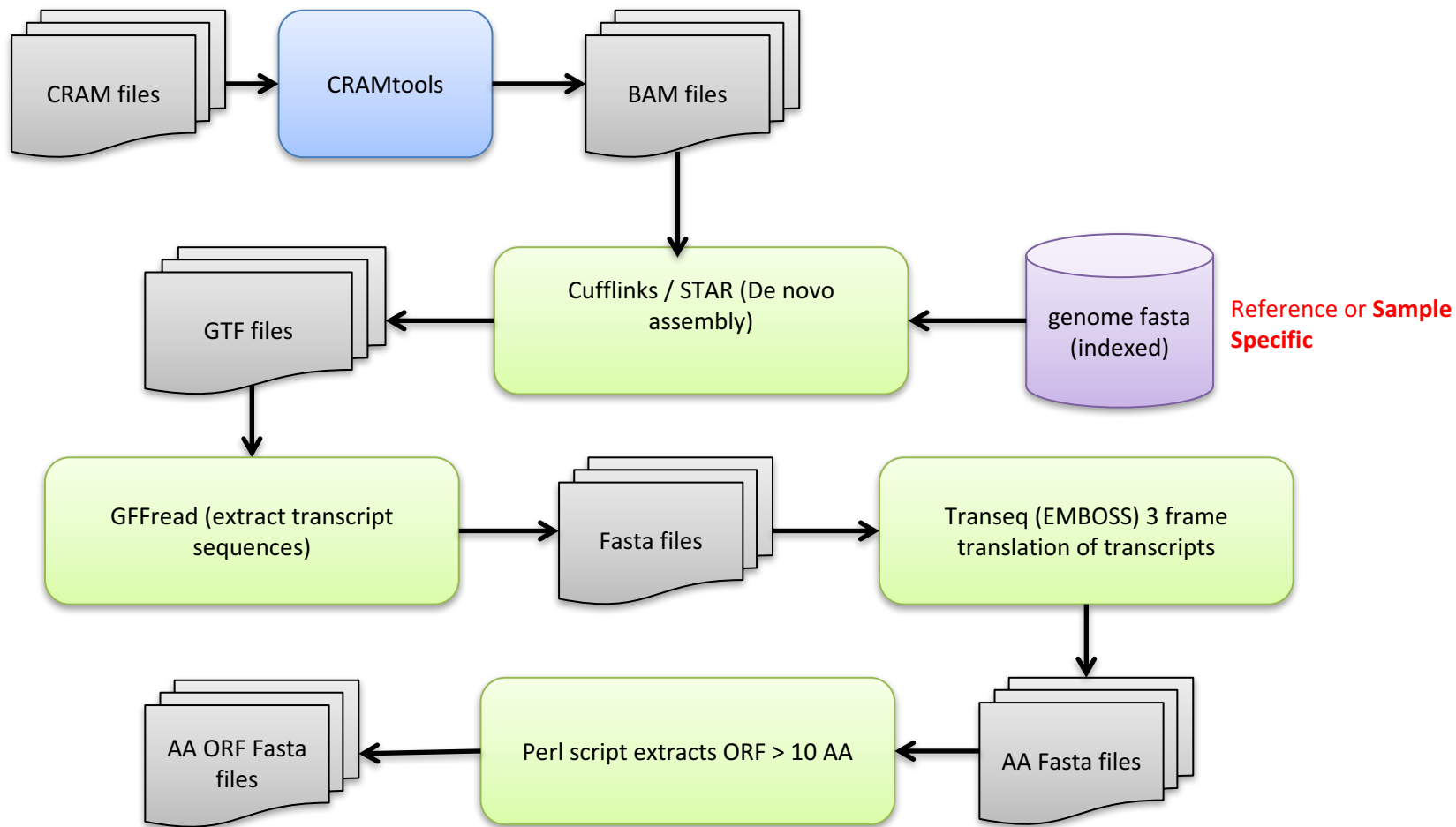
Number of Sequences in Each Category



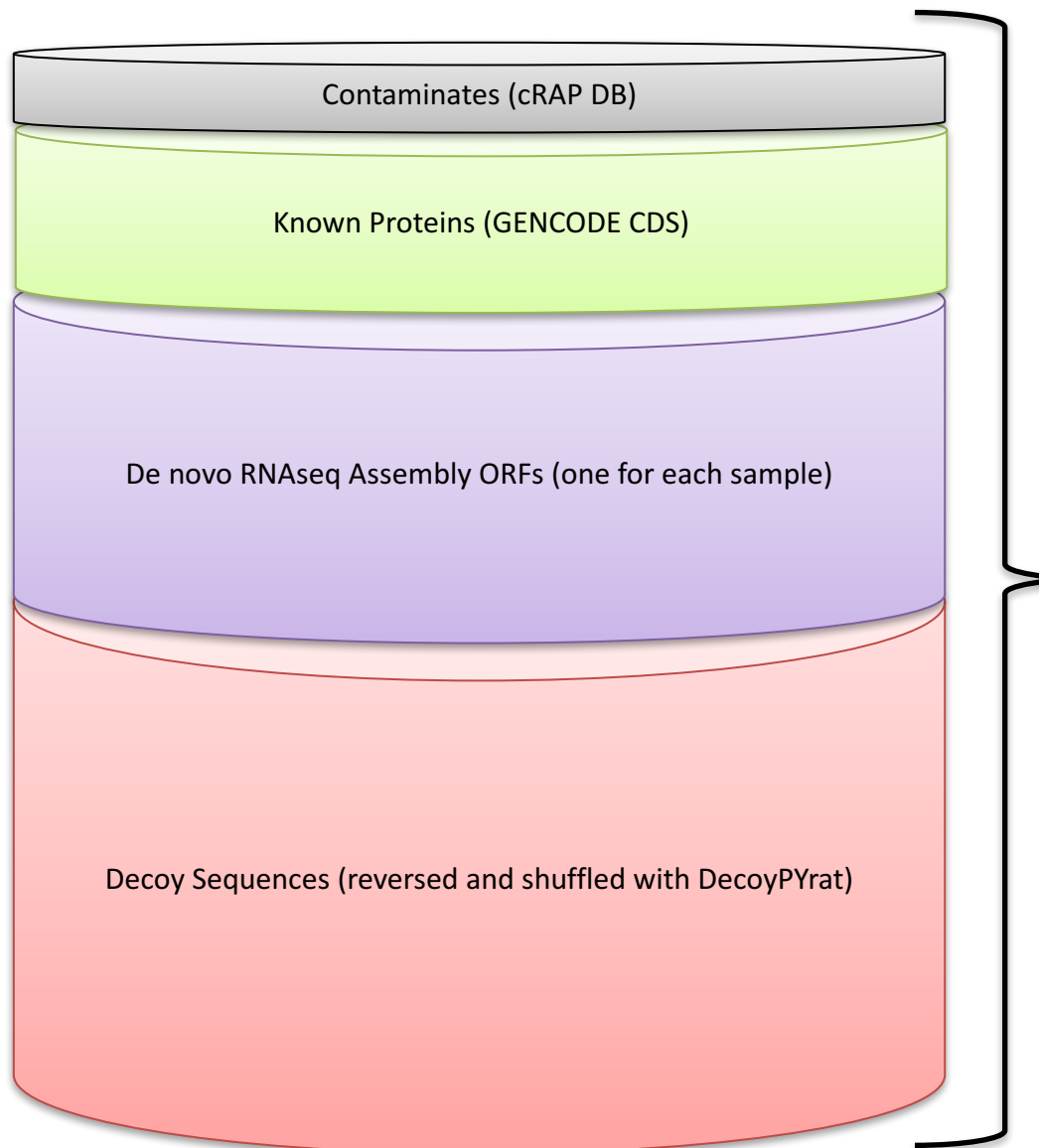
Observable Tryptic Peptides



RNAseq De novo Assembly



Sample Specific Sequence Database



Protein Clustering

(Redundant and subset sequences and collapsed to single protein sequence)

Protein Identifiers

(Each protein assigned unique identifier and protein type, separate table generated mapping each protein to original accessions and Ensembl gene / transcript ids)

Protein Types

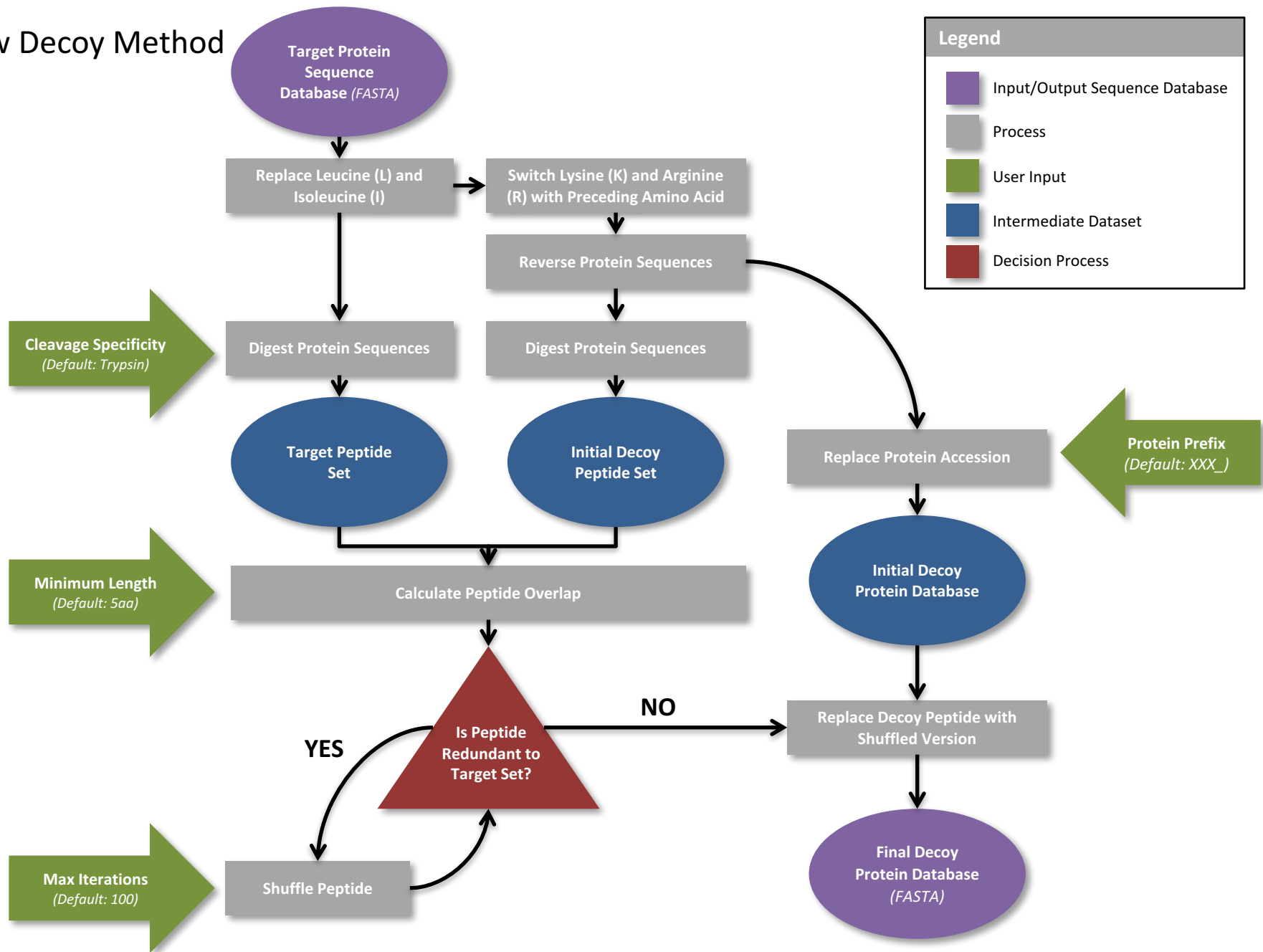
(Each protein assigned type)

CON = Contaminate, CDS = Reference Protein, NOV = Protein is unique to RNAseq assemblies, XXX = Decoy Sequence

Peptide Table

(Simulated tryptic digestion, to generate table linking short peptide sequence to original proteins)

New Decoy Method

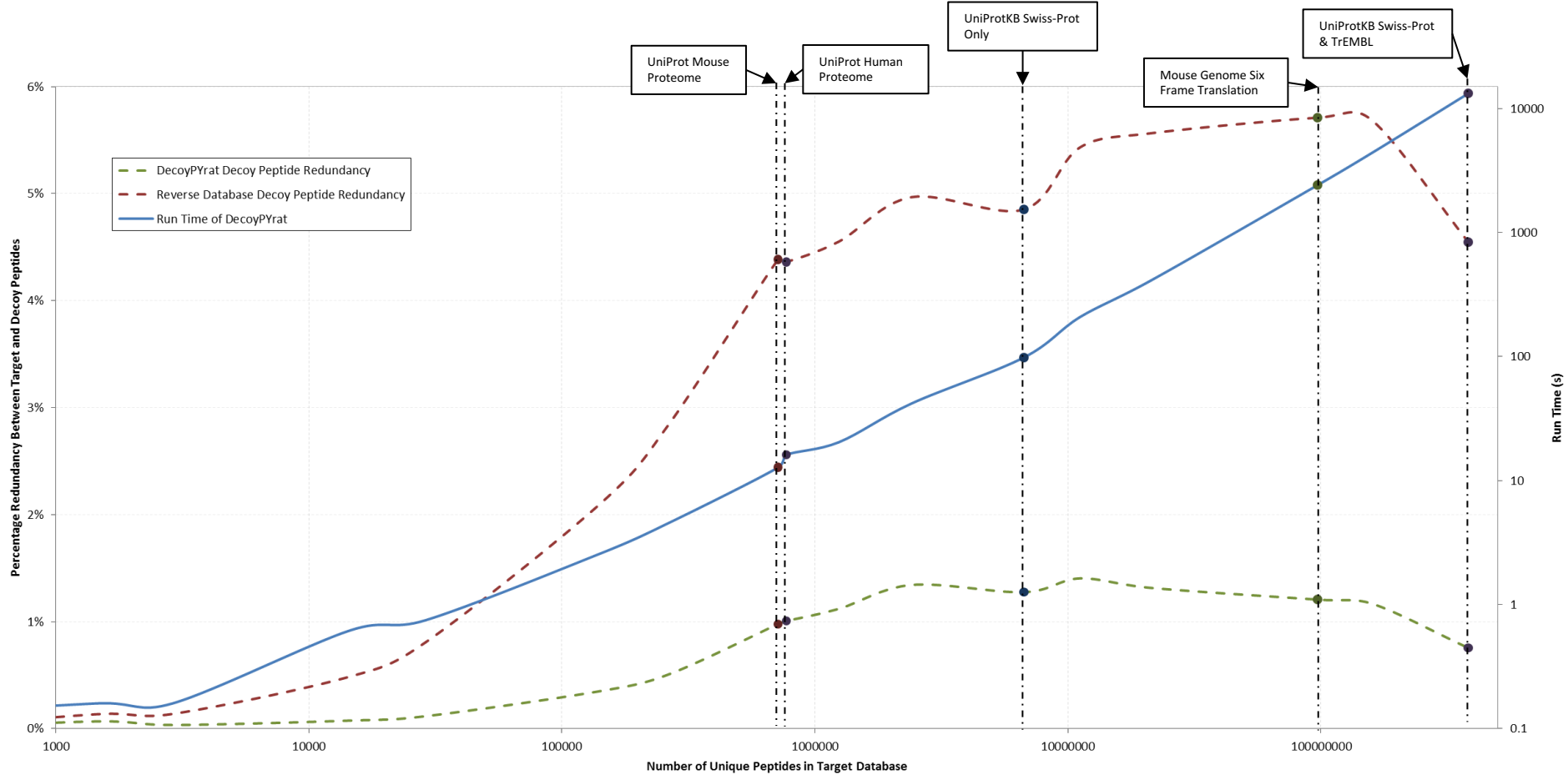


J Proteomics Bioinform. 2016 Jun 27;9(6):176-180.

DecoyPyrat: Fast Non-redundant Hybrid Decoy Sequence Generation for Large Scale Proteomics.

Wright JC, Choudhary JS.

DecoyPYrat vs Reversed Sequences

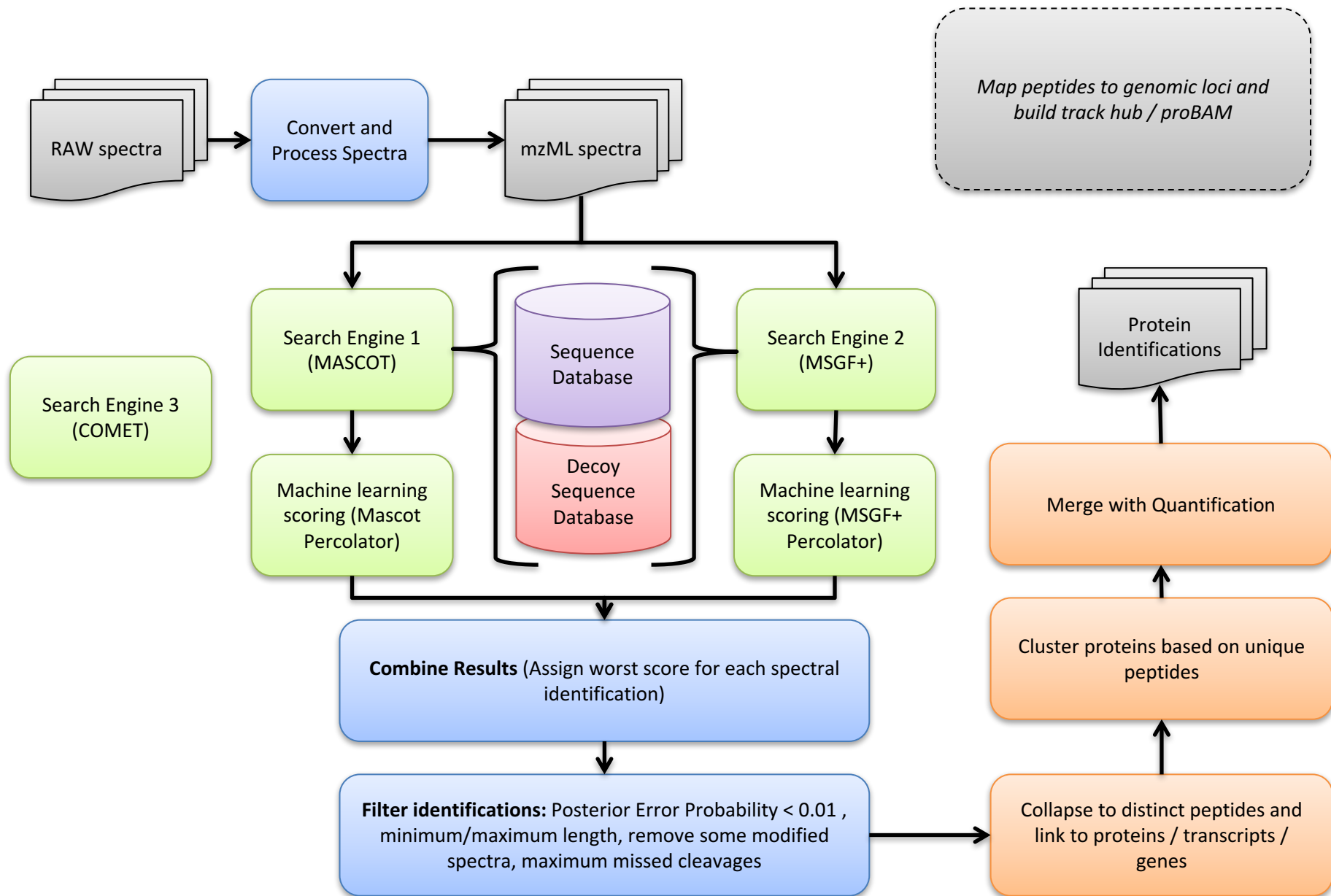


J Proteomics Bioinform. 2016 Jun 27;9(6):176-180.

DecoyPYrat: Fast Non-redundant Hybrid Decoy Sequence Generation for Large Scale Proteomics.

Wright JC, Choudhary JS.

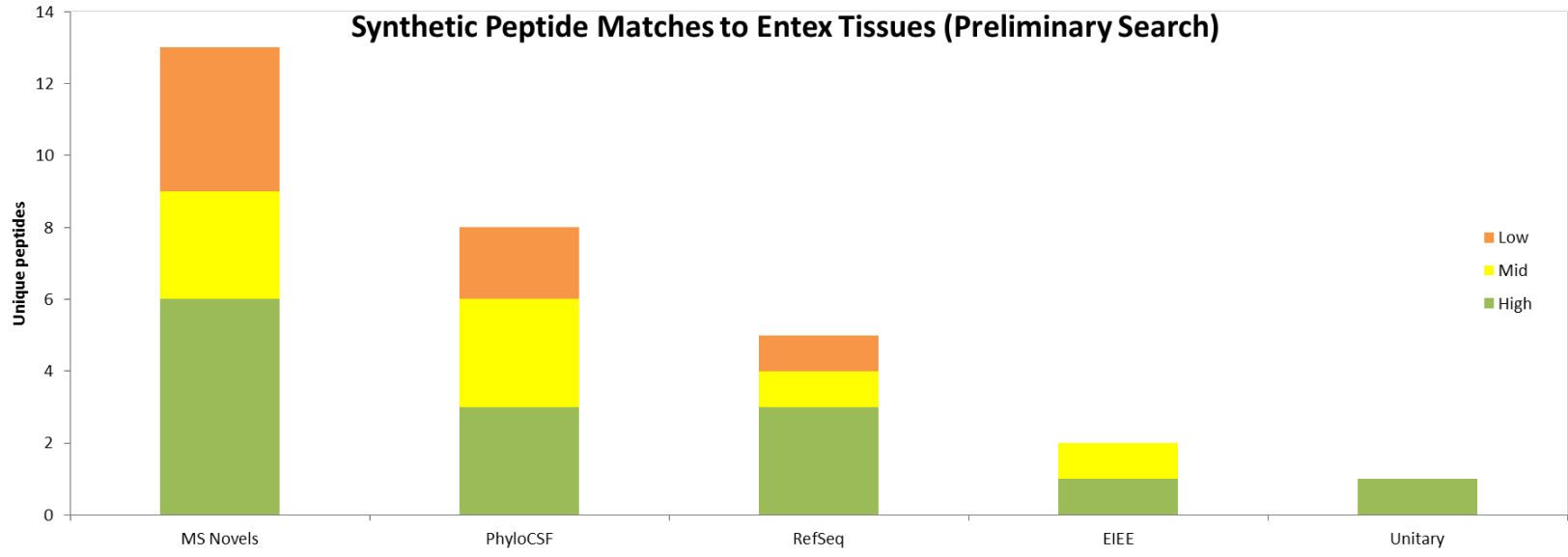
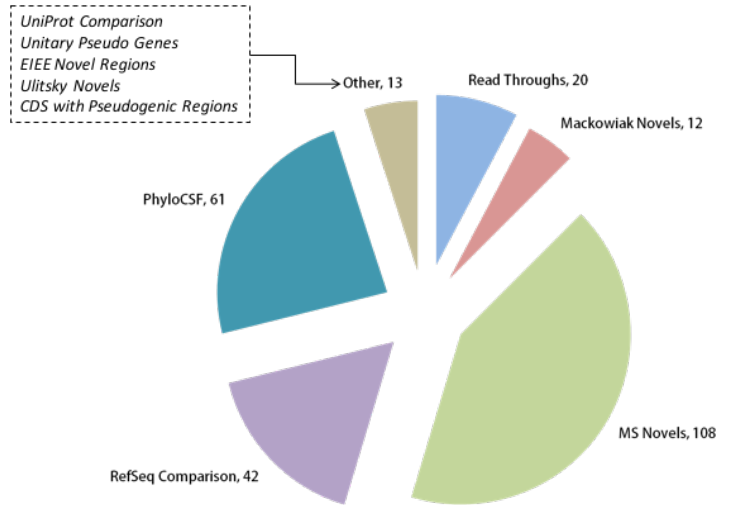
Proteogenomic Identification Workflow



Planned Analysis

- Identify novel protein coding genes and transcripts
 - Map all peptides to genome in a trackhub for use in annotation
- Examine correlation between RNAseq and Proteome
 - Improve on previous experiments by using the RNAseq to identify the Proteome
- Proteome analysis use RNAseq data to highlight missing proteome
 - Further targeted experiments to get deeper proteome
- Investigate differences in transcript and protein, abundances and isoforms between tissues and individuals
- If enough statistical power, link genetic features to proteome differences

Preliminary Results



MS analysis Peptides

Peptide	FDR	PEP	VEGA ID	Notes
X.FTLQGAESLVTYR.X	0.02%	2.52E-06	OTTHUMG00000021534	Novel protein similar to centromere protein V (CENPV). Possibly testis specific.
X.EAGSRDPLPSAPLPDPPAPAESPK.X	0.11%	2.92E-02	OTTHUMG00000021534	see above
X.DPLPSAPLPDPPAPAESPK.X	1.17%	1.71E-01	OTTHUMG00000021534	see above
X.SSGPYGGGGQYFAK.X	0.02%	3.71E-05	OTTHUMG00000010071	Heterogeneous nuclear ribonucleoprotein A1 pseudogene 6 (HNRNPA1P6). The CDS has multiple disablements. Nonetheless, the locus is transcribed and the peptide is found within a 24aa ORF. Potentially real, although the peptide does is actually not uniquely mapped; it also aligns to P09651.
X.MLNKPLELLVQDGK.X	0.02%	1.74E-04	OTTHUMG00000155842	GDP dissociation inhibitor 2 (GDI2) pseudogene. Has 6 mismatches to parent, so mapping seems ok. The CDS contains disablements, though the peptide is found in the second half which corresponds to an intact ORF. No conventional evidence for transcription, though intriguing. No changes made.
X.LSLEGHHSTPSGAYGSVK.X	0.02%	2.96E-03	OTTHUMG00000018326	Annexin A2 pseudogene 3 (ANXA2P3). The CDS has multiple disablements. However, the peptide sits within an ORF and seems uniquely mapped (3 differences to the parent). No conventional transcript evidence, though CAGE is very high and specific to certain epithelial cells. Intriguing though haven't changed anything.
X.TEQGPTGVTMTSNPLTWGQLK.X	0.04%	7.39E-03	OTTHUMG00000150367	ERVK3-1; doesn't add any new insights.
X.LALQDALNENSQLQESQK.X	0.04%	8.40E-03	OTTHUMG00000016925	meningioma expressed antigen 6 (coiled-coil proline-rich) (MGEA6) processed pseudogene. No conventional evidence for transcription, although HPA has strong support in testis, which matches a weak CAGE peak. The CDS is intact until a PTC 20aa from the end. Just 1aa different to the parent, with about a dozen paralogs. So could be misplaced, although potentially real. Probably we'd want to do some targeted transcriptomics before changing it.
X.HQGVMVGLGQKDSYVGEDAQSK.X	0.05%	1.44E-02	OTTHUMG00000177065	Actin, beta (ACTB) pseudogene. There are multiple disablements, although the peptide is found in a decent sized ORF based on the canonical ATG and has 3 mismatches to the parent. No conventional transcription evidence, although HPA apparently finds testis expression. Potentially real.
X.HQGVMVGLGQK.X	1.49%	2.04E-01	OTTHUMG00000177065	see above
X.KLLVSNVDQTLDDPYATFVK.X	0.11%	2.97E-02	OTTHUMG00000018684	Cofilin 1 (non-muscle) pseudogene 1. Clearly transcribed, especially in testis. The CDS has a PTC, although the peptide is within an ORF that covers the first half, using the canonical ATG. Has 5 difference to the parent. So potentially we could switch this already?
X.QTMQGLADGAGAVLLASEDAVK.X	2.63%	2.84E-01	OTTHUMG00000024126	Acetyl-Coenzyme A acyltransferase 2 (mitochondrial 3-oxoacyl-Coenzyme A thiolase) (ACAA2) pseudogene. One EST supports transcription, although the CDS has two PTCs. The peptide is spliced in a manner for which I see no evidence.
X.WGNAGAEYLMESTGVFTTMEK.X	3.80%	3.61E-01	OTTHUMG00000160001	Within GAPDHP44, with 4 mismatches to parent and in a fragment ORF. However, there are dozens of these loci, and we'd naturally worry the peptide is misplaced. No evidence for transcription.

PhyloCSF peptides

Peptide	FDR	PEP	VEGA ID	Notes
X.LQMDLDVTTTQLLPNAGGFLCR.X	0.02%	9.39E-06	OTTHUMG00000159843	MINDY family member, previously known as FAM188B2. It was a protein-coding gene in GENCODEv19, but switched to pseudogene as the only transcript evidence at that time supported a truncated CDS. Very weak expression in human CAGE and HPA, though Intropolis finds highest expression in early embryo. Mouse has CAGE limited to 3 ear cell experiments. PMID:20717163 previously demonstrated that certain of these exons are transcribed in retina as part of the CLRN1 locus immediately upstream (their coding potential was not examined). In context, these look like readthrough transcription events; both the canonical STOP of CLRN1 and the ATG used for MINDY4B are deeply conserved. We may not have expected to find peptide support, given the restricted expression profile. Nonetheless, no reason to be suspicious about the peptide data.
X.LSTGLSQNGGRSSAQPCPR.X	0.02%	1.33E-03	OTTHUMG0000019268	Uncharacterised protein, now C10orf143. The human cDNA evidence includes additional exons that induce frameshifts, hence the CDS was previously missed. Has general expression.
X.NNLSWLKEDTQLTNAK.X	0.03%	4.40E-03	OTTHUMG00000158801	MyoD family inhibitor domain containing 2 (MDFIC2). Human expression is generally weak, and mostly limited to cancer and cell line experiments. It was built originally on mouse evidence, where expression is most obvious in mesenchymal stem cells, certain brain experiments and limb.
X.EPGLETGTQAADCK.X	0.04%	1.06E-02	OTTHUMG00000173186	Uncharacterised protein, initially picked up in the C-HPP work. Previously described as a lncRNA p53 target by PMID:25524025. Low general expression in HPA.
X.LGVSGEPSPCTSTNR.X	0.07%	2.08E-02	OTTHUMG0000017712	Found by PhyloCSF; novel Myb/SANT-like DNA-binding domain containing protein, found within what were previously considered non-coding exons of HSPA14. All evidence indicates that transcription occurs from the shared HSAP14 promoter in human and mouse, with no evidence of differential expression between the two loci. It is thus not obvious how translation of the two CDS is distinguished. This CDS was initially suggested by a GENCODE reanalysis of the Kim et al and Wilhelm et al mass spectrometry datasets, although the supporting evidence was not considered strong enough to support annotation at that time. Clear general expression profile.
X.LQVQNGECPWQVSLQMSR.X	0.09%	2.58E-02	OTTHUMG00000163805	Novel protease, serine family member. Human has a premature termination codon in the final exon, giving a shorter protein compared to other mammals including apes. However, this PTC does not appear to affect the trypsin domain, so it was decided to represent the locus as coding as opposed to a unitary pseudogene. So the peptide support is nice. Human and mouse expression is mostly limited to testis.
X.YPDKLFGTNENL.X	3.02%	3.11E-01	OTTHUMG0000019742	SMIM27, previously known as TOPORS-AS1. General expression.
X.WLMAQQQELQQKEQELK.X	4.13%	3.79E-01	OTTHUMG00000132641	Uncharacterised protein. PMID:22196729 had previously reported this locus as a deeply conserved lncRNA, first identified in zebrafish. Expression is apparently testis specific.

Other peptides

Peptide	FDR	PEP	VEGA ID	Notes
EIEE Genes:				
X.DAAYYSYQNSSPK.X	0.03%	4.58E-03	OTTHUMT00000488129	CREBBP splice acceptor shift, originally built on mouse evidence but I can see human RNAseq support faint but discernible in read coverage graphs.
X.AMSLASLLTNTMEGK.X	0.15%	3.53E-02	OTTHUMT00000490140	SCN2A poison exon, with the final two aa only in the poison exon. This portion of CDS is conserved, although the larger region is deeply conserved like a regulatory exon.
RefSeq:				
X.QHASEGDGDQSPTQCAGMR.X	0.02%	4.26E-06	OTTHUMG00000119017	C2orf48 was switched to non-coding because the CDS has poor conservation beyond apes, although RefSeq have maintained it. We would not make this without experimental support. Has plenty of standard significance Pandey / Kuster support. The first two exons are well conserved in mammals, although not apparently as coding sequence. There is no CAGE support, thus readthrough from RRM2 upstream seems to be the route of transcription.
X.QLQLSVFQDLNQFSHCR.X	0.20%	4.46E-02	OTTHUMG00000119017	see above
X.LQEAGGPTGGCGVGGQPLGGR.X	0.03%	4.16E-03	OTTHUMG00000171567	This was CCDS45101 at one stage, though the ORF was removed due to lack of conservation and RefSeq agree it is LINC01599. It was a 5 exon CDS of 324aa, although 4 of these exons have no RNAseq support. It was based on a single anonymously submitted cDNA, which may well be dubious. However, there is also appreciable standard significance Pandey / Kuster support. I think if we were to make this we'd need to do some targeted RNA experimental work first, and right now I don't even think we should have the lincRNA.
X.AQLDLQHPQDRVVTCK.X	0.04%	9.22E-03	OTTHUMG00000176796	Actually on the PhyloCSF list, matched to an XP. Novel protein similar to TLE4. HPA supports testis expression; Intropolis strongly supports early embryo expression in multiple experiments.
X.LLPGFMCQGGDFTRPNGTDDK.X	4.34%	3.90E-01	OTTHUMG00000191983	Novel peptidylprolyl isomerase A family protein. Toby switched this to coding recently, presumably based on the RefSeq comparison. Nice to have peptide support, although it has just 1aa different to PPIAL4D.
Unitary Pseudogenes:				
X.EGVTNPSNSSQALLK.X	0.02%	2.62E-03	OTTHUMG00000177210	Tubulin tyrosine ligase-like family, member 13 (TTLL13). RefSeq have a pseudogene, and we did too until Toby changed it in 2016 (probably in the UniProt comparison, where it is A6NNM8). It looked pseudogenised in primates due to splice site loss, though RNAseq analysis supports a change in splicing. Nice example.

Summary

- Proteogenomic analysis of 6 different tissues across 4 individuals
- Build sample specific search space using genomic and RNAseq data
- Identify novel protein coding transcripts and alternative splicing
- Quantitative comparison of RNAseq and proteome across tissues and individuals
- Synthetic peptides for validation of potential protein coding transcripts
- Spectral library for use in PRIDE repository