Shotgun proteomics aids discovery of novel protein-coding genes, alternatice splicing, and "resurrected" pseudogenes in the mouse genome

M. Brosch, G. Saunders, A. Frankish, M. O. Collins, L. Yu, J. Wright, R. Verstraten, D. J. Adams, J. Harrow, J. S. Choudhary, and T. Hubbard

Genome Research **21** (2011) 756–767.

Speaker: Joseph Chuang-Chieh Lin

The Comparative & Evolutionary Genomics/Transcriptomics Lab.
Genomics Research Center, Academia Sinica
Taiwan

14 May 2014

Outline

Background

- Annotation efforts: automatic annotation systems (e.g., Ensembl) & manual annotation (e.g., VEGA, RefSeq).
- A high-throughput method providing orthogonal data for validation and confirmation of the protein-coding potential is also required.
- Efforts to combine genome annotation with protein MS: *proteomics* [Jaffe *et al.* 2004].
  - It serves as translational evidence.
- Peptide identification methods and significance measures are both required to be sensitive and accurate.

Background (contd.)

- Mascot Percolator [Brosch *et al.* 2009].
  - Mascot [Perkins *et al.* 1999]: a database search engine;
  - Percolator [Käll *et al.* 2007.]: a semi-supervised machine learning algorithm.
- Two significance measures:
  - q-value [Storey & Tibshirani 2003];
  - PEP (posterior error prob.) [Käll et al 2008]

Contribution of this paper

- A novel pipeline that integrates
  - highly sensitive & statistically robust peptide spectrum matching (PSM);
  - genome-wide protein-coding predictions

  to perform large-scale gene validation and discovery in the mouse genome for the first time.
- Validation of 32%, 17%, and 7% of all protein-coding genes, exons, and splice boundaries, resp.

Contribution of this paper (contd.)

- Strong evidence for identifying multiple AS translations from 53 genes & uncovered 10 entirely novel protein-coding genes.
  - 2 gene fusions (including a *Ins2-Igf2* fusion object).
  - 9 processed pseudogenes (unique peptide hits): not just transcribed but translated and resurrected into new coding loci.

FDR & PEP

PNAS

Statistical significance for genomewide studies

John D. Storey*† and Robert Tibshirani‡

*Department of Biostatistics, University of Washington, Seattle, WA 98195; and ‡Departments of Health Research and Policy and Statistics, Stanford University, Stanford, CA 94305

Journal of proteome research

Posterior Error Probabilities and False Discovery Rates: Two Sides of the Same Coin

Lukas Käll,† John D. Storey,†,‡ Michael J. MacCoss,† and William Stafford Noble*,†,§

Department of Genome Sciences, Biostatistics, Computer Science and Engineering, University of Washington, Seattle, Washington, 98195

Whereas the p value is a measure of significance in terms of the false positive rate, the q value is a measure in terms of the FDR . . .

A false positive rate of 5% means that on average 5% of the truly null features in the study will be called significant. A FDR of 5% means that among all features called significant, 5% of these are truly null on average.

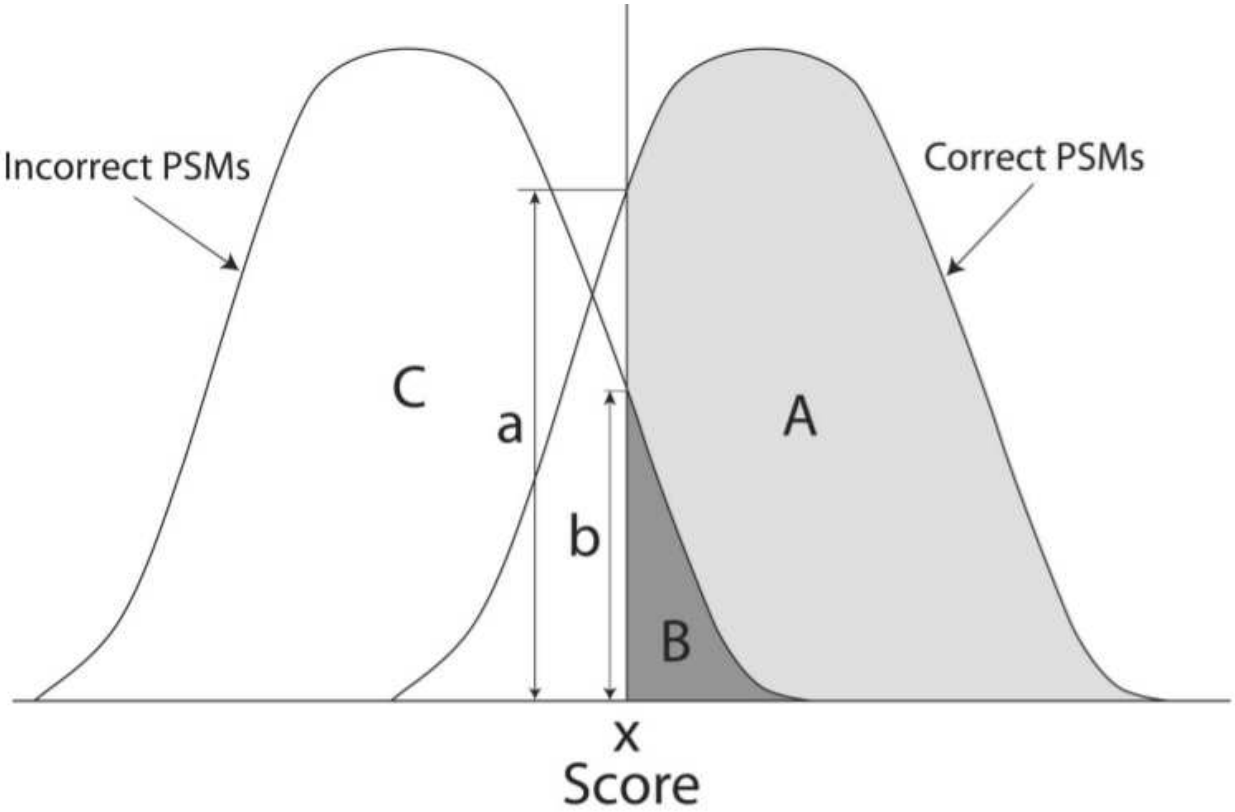
FDR & PEP

	Called significant	Called not significant	Total
Null true	F	$m_0 - F$	m_0
Alternative true	T	$m_1 - T$	m_1
Total	S	$m - S$	m

$$\frac{\text{no. false positive features}}{\text{no. significant features}} = \frac{F}{F+T} = \frac{F}{S},$$

$$\text{FDR} = \text{E}\left[\frac{F}{F+T}\right] = \text{E}\left[\frac{F}{S}\right].$$

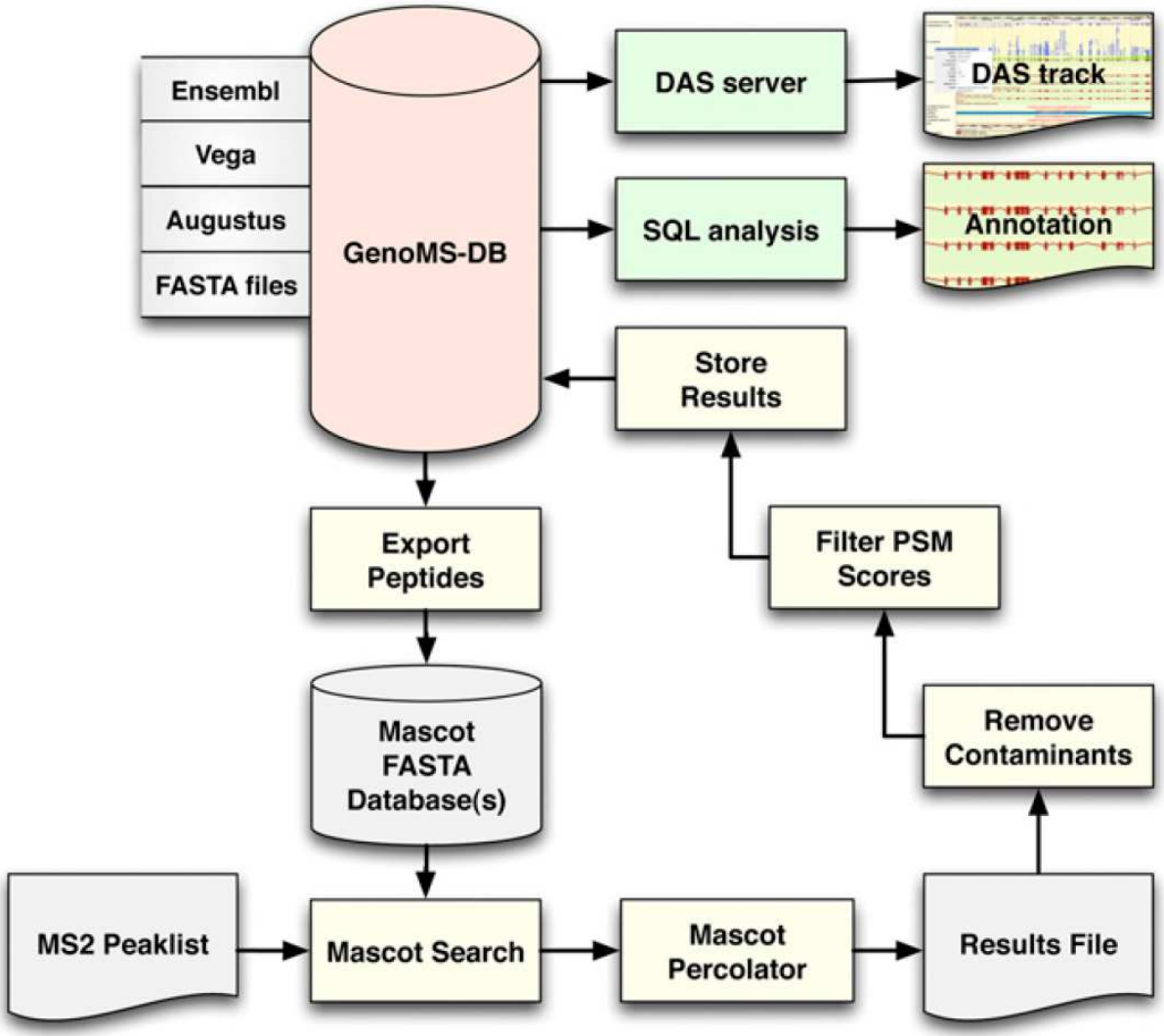
FDR & PEP



FDR = B/(A + B)

PEP = b/(a + b)

Overview of Genome Annotation Pipeline



MS/MS data

- 10,465,149 tandem MS spectra.
  - 729,583 spectra: in-house experiments
    - Nuclear protein extracts of murine ESCs & murine brain membrane fractions.
  - 9,735,566 spectra: PeptideAtlas project.
    - Sampling of mouse tissues including brain, liver, lung, heart, kidney, testes, and placenta.

GenoMS-DB database construction

- Gene products from
  - ⋆ Ensembl, VEGA, IPI digest in silico;
    - IPI: International Protein Index [Kersey *et al.* *Proteomics* 2004].
  - ⋆ predictions from Augustus.
- Ensembl Per API: to capture the peptide-genome mapping.

Automatic & manual annotation

- Perl-based Distributed Annotation System (DAS):
  - Visualize the identified peptides stored in GenoMS-DB as tracks in various genome browsers and curation tools.
- Manual annotation:
  - MS PSMs overlapping annotated loci $\rightarrow$ HAVANA.
  - Otherwise, follow the hierarchy:
    - RT-PCR $>$ species-specific transcriptional support $>$ rodent specific transcriptional support $>$ strong mammalian conservation $>$ paralogous gene transcriptional evidence.

Translated pseudogene analysis

- To select the parent of each identified translated pseudogene:
  - assign homology scoring of the putative translation of the processed pseudogene object against the SWISS-PROT data set;
  - (check) assign each of the PSMs aligning to the pseudogene loci to a parent protein by aligning to the compete UniProt database using HMMER.
- Gene orthologous to these parents: application of Ensembl website.
- Protein alignment: ClustalW2 (EBI).
- Identification of domains: InterProScan (EBI).

Generaton of high-confidence PSMs

- When considering q-value $< 1\% \rightarrow$ PEP $< 1\%$:
  - 1,124,724 peptides were identified (Ensembl, Vega).
  - 967,131 peptides were identified (Augustus).
- Only the best PEP and q-value score for each peptide sequence was considered ($\Rightarrow$ 95,606).
- Removing peptides matching common contaminants (3,260 removed).

Generaton of high-confidence PSMs

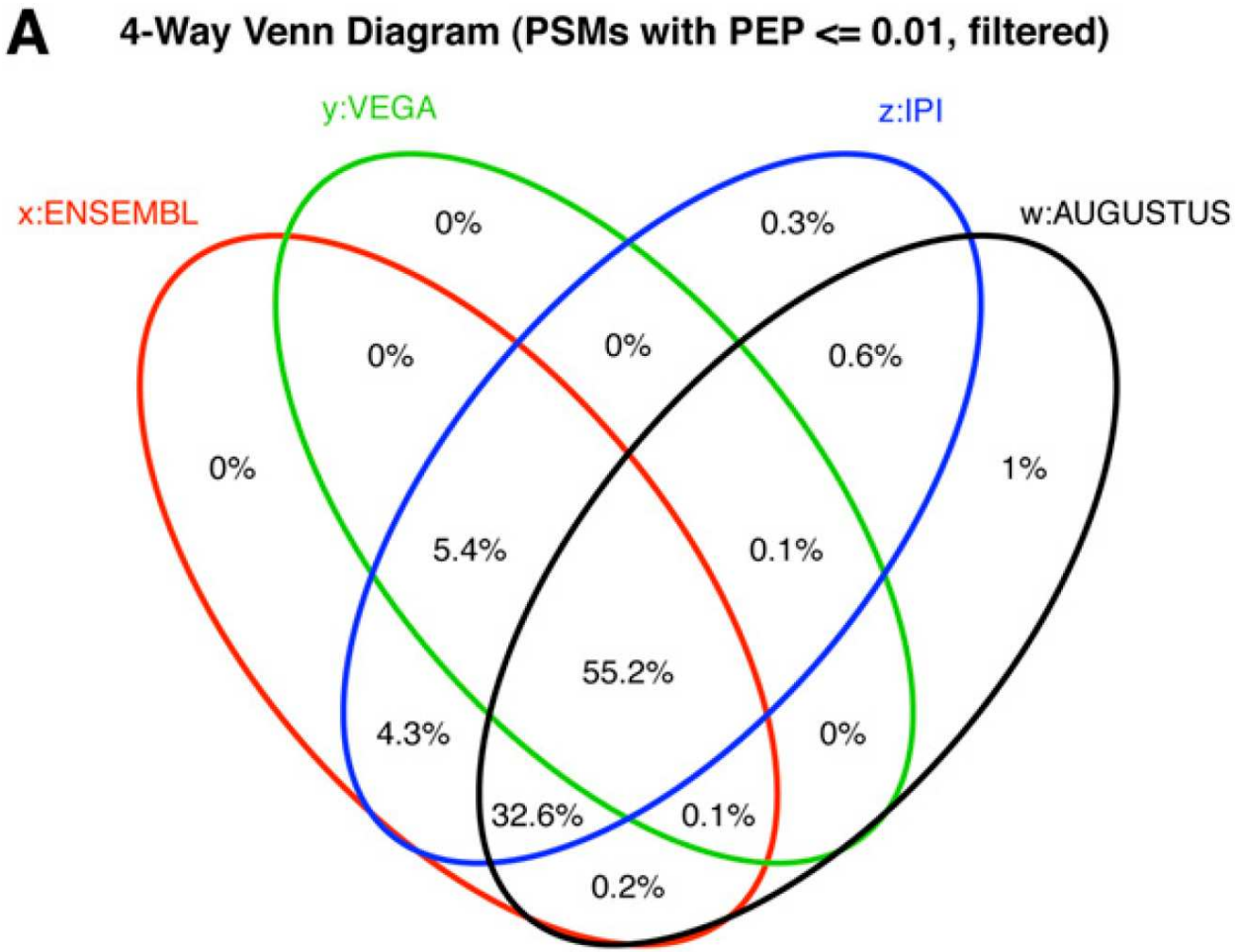
- When considering q-value $< 1\% \rightarrow$ PEP $< 1\%$:
  - 1,124,724 peptides were identified (Ensembl, Vega).
  - 967,131 peptides were identified (Augustus).
- Only the best PEP and q-value score for each peptide sequence was considered ($\Rightarrow$ 95,606).
- Removing peptides matching common contaminants (3,260 removed).

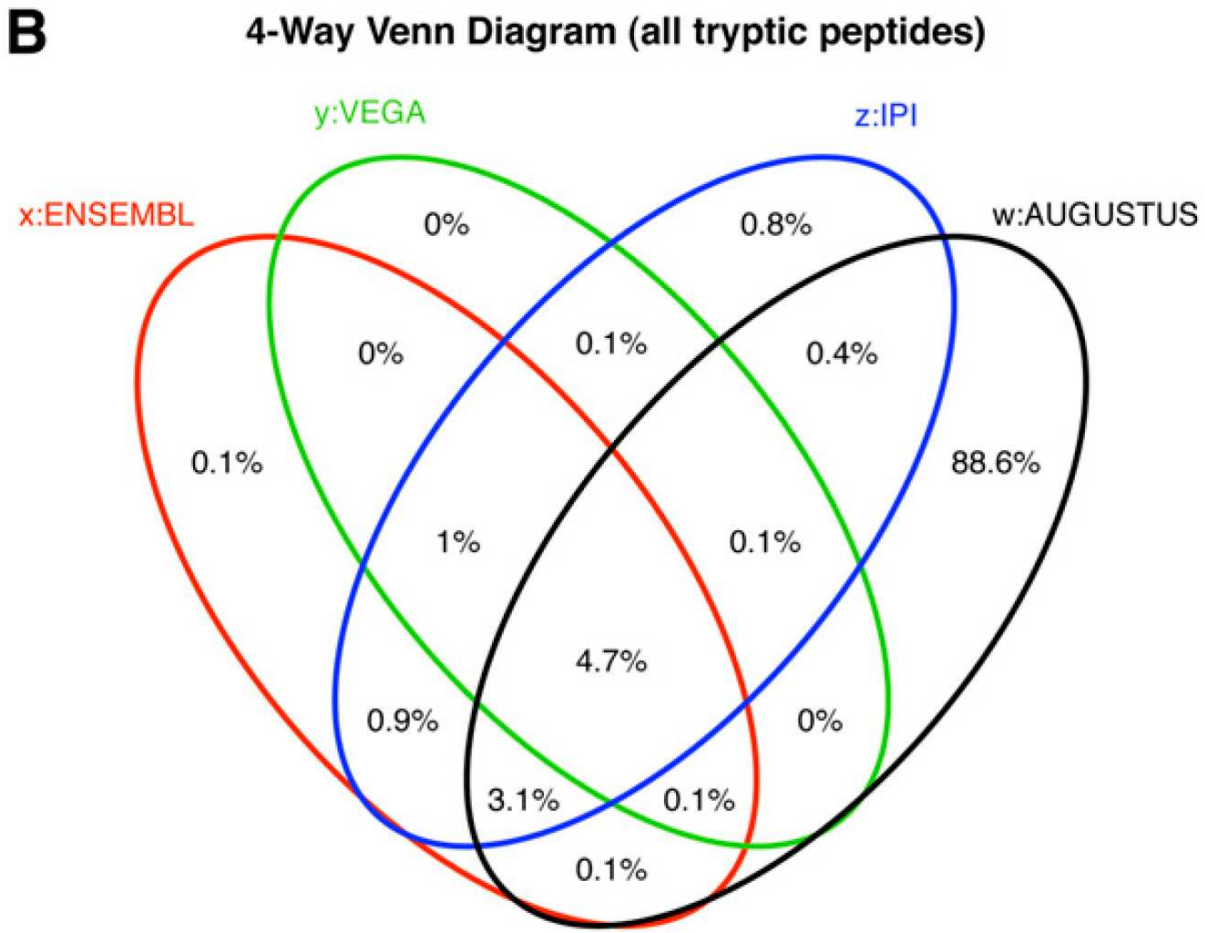
Generaton of high-confidence PSMs (contd.)

- Filtering peptides where isoforms attributed to amino acids that cannot be discriminated in low energy collision induced dissociation data (1,159 removed).
- Unambiguous mapping to one genomic locus ($\Rightarrow$ 76,029 remained).
- Testing whether semi-tryptic form of the peptide sequence mapped elsewhere ($\Rightarrow$ 758 cases removed).
- Testing whether one residue substitution/insertion/deletion could be identified elsewhere ($\Rightarrow$ 6,685 cases removed; 68,586 finally.)

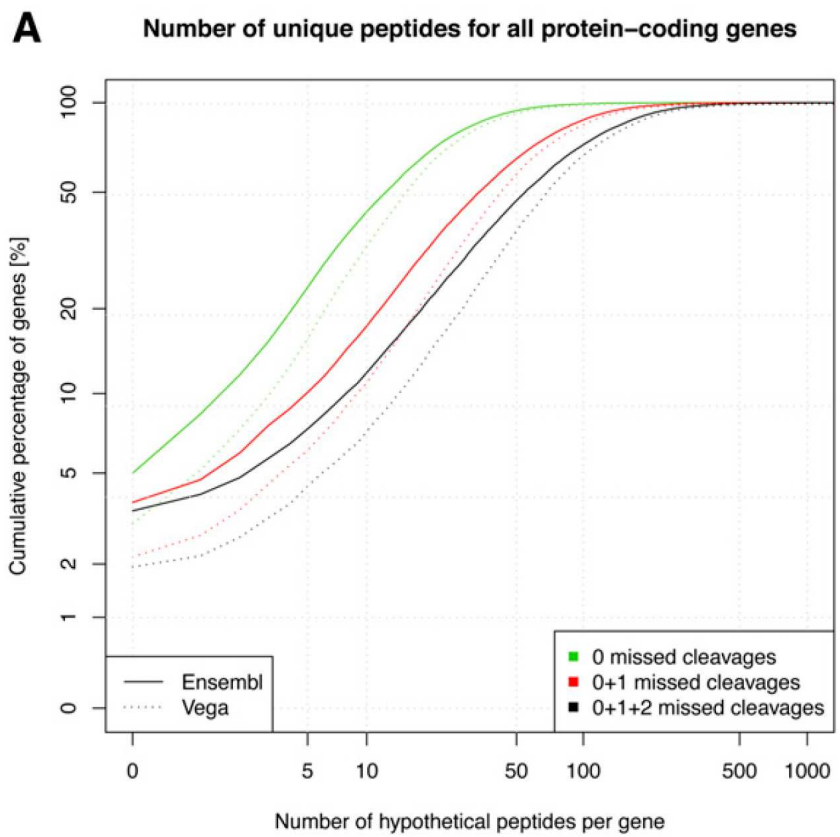
⋆ $1\% \leq \text{PEP} \leq 5\%$: exclusively used as supplement.

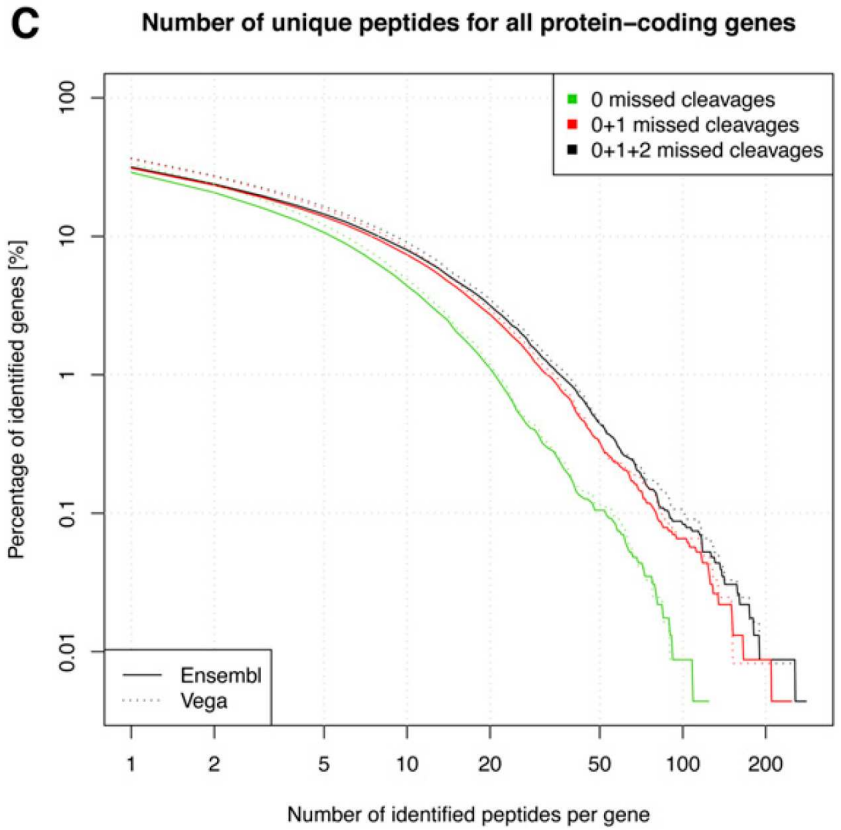
⋆ $\text{PEP} \leq 1\%$: primary annotation data source (58,574 cases).


A
4-Way Venn Diagram (PSMs with PEP <= 0.01, filtered)
y:VEGA
z:IPI
x:ENSEMBL
w:AUGUSTUS
0%
0.3%
0%
0%
0.6%
0%
1%
5.4%
0.1%
55.2%
4.3%
0%
32.6%
0.1%
0.2%


B
4-Way Venn Diagram (all tryptic peptides)
y:VEGA
z:IPI
x:ENSEMBL
w:AUGUSTUS
0%
0.8%
0%
0.1%
0.4%
0.1%
88.6%
1%
0.1%
4.7%
0.9%
0%
3.1%
0.1%
0.1%

Validation of Ensembl/VEGA gene annotation



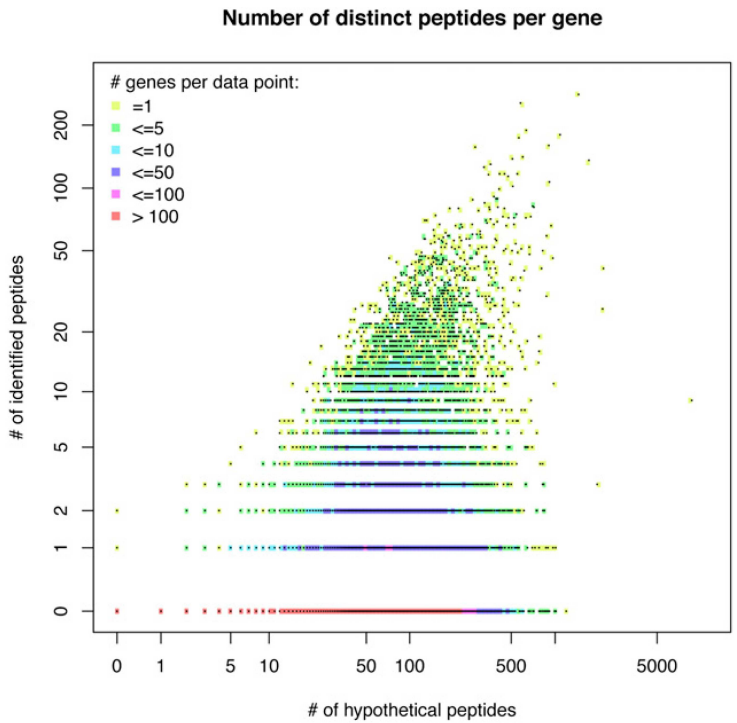


Validation of Ensembl/VEGA gene annotation

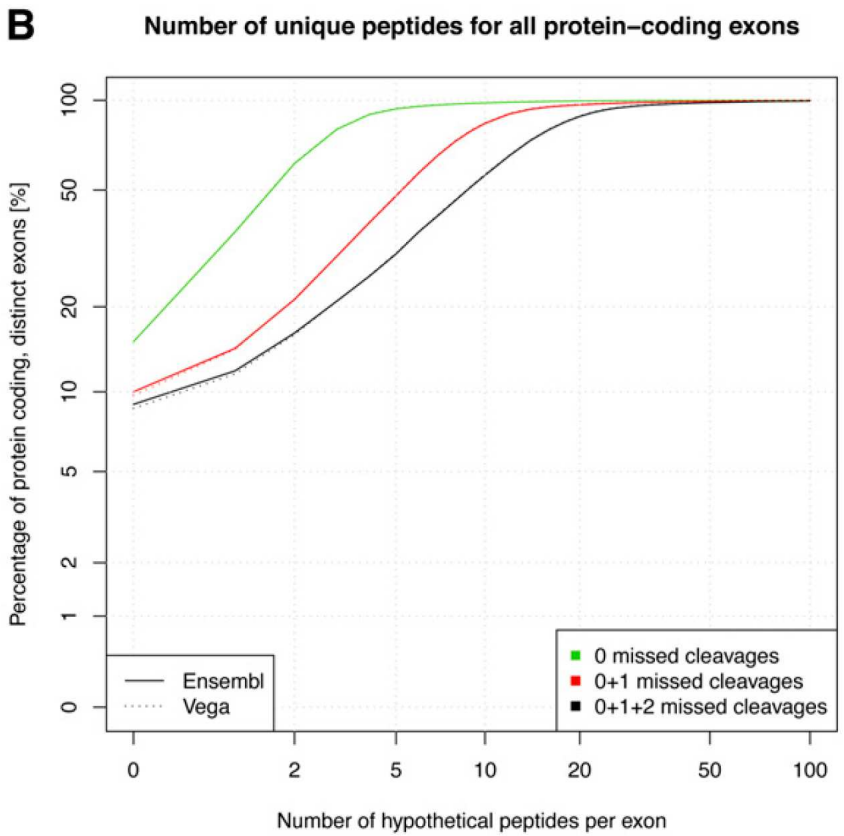
- Is there a linear model fitted?
  - gene products with more potential peptides
    $\Rightarrow$
    sampled peptides $\uparrow$

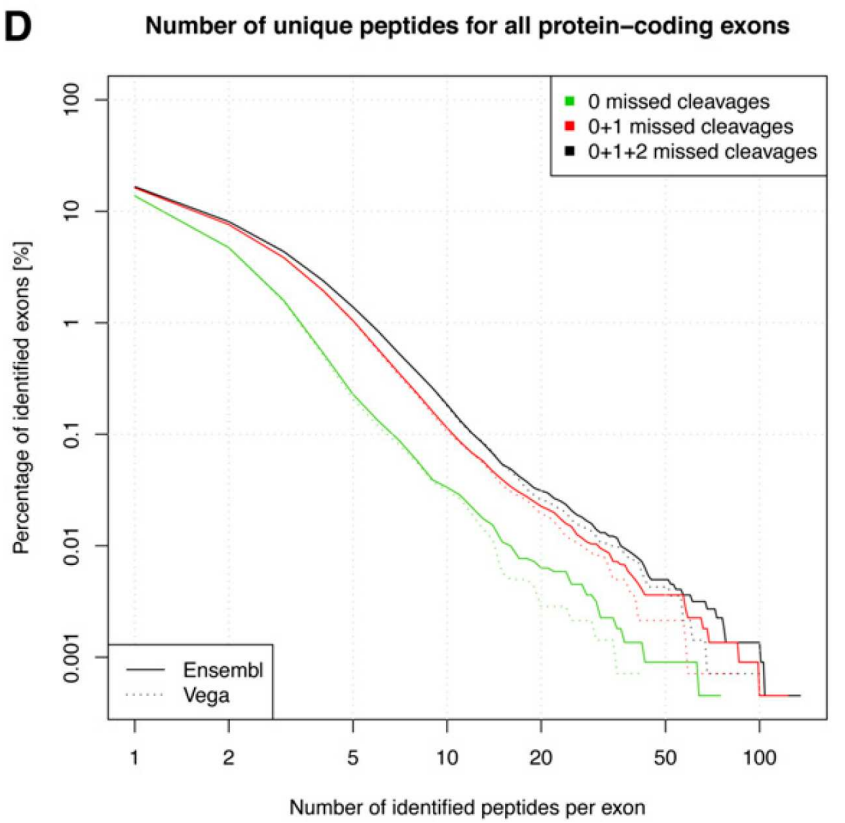
Validation of Ensembl/VEGA gene annotation

- Is there a linear model fitted?
  - gene products with more potential peptides
    $\Rightarrow$
    sampled peptides $\uparrow$



Validation of Ensembl/VEGA gene annotation (structure)





Validation of Ensembl/VEGA gene annotation (structure)

- Overall, 16.7% (7.1%) of the total Ensembl protein-coding exons (introns) could be validated by peptide identifications.

Validation of Ensembl/VEGA gene annotation (AS)

- Until recently, only limited evidence of expression of AS transcripts was available at the protein level.
- The majority of protein sequence is shared between the variant transcripts, differing only in small parts ($\Rightarrow$ **signatures**) of the translation products.
- Here, a total of 370 peptides enabled discrimination of 112 Ensembl transcripts in 53 genes.
  - 3.4% of all protein-coding genes with annotated multiple coding AS forms that can be discriminated by a peptide.

Validation of Ensembl/VEGA gene annotation (AS)

- Until recently, only limited evidence of expression of AS transcripts was available at the protein level.
- The majority of protein sequence is shared between the variant transcripts, differing only in small parts ($\Rightarrow$ **signatures**) of the translation products.
- Here, a total of 370 peptides enabled discrimination of 112 Ensembl transcripts in 53 genes.
  - 3.4% of all protein-coding genes with annotated multiple coding AS forms that can be discriminated by a peptide.

Validation of Ensembl/VEGA gene annotation (structure)

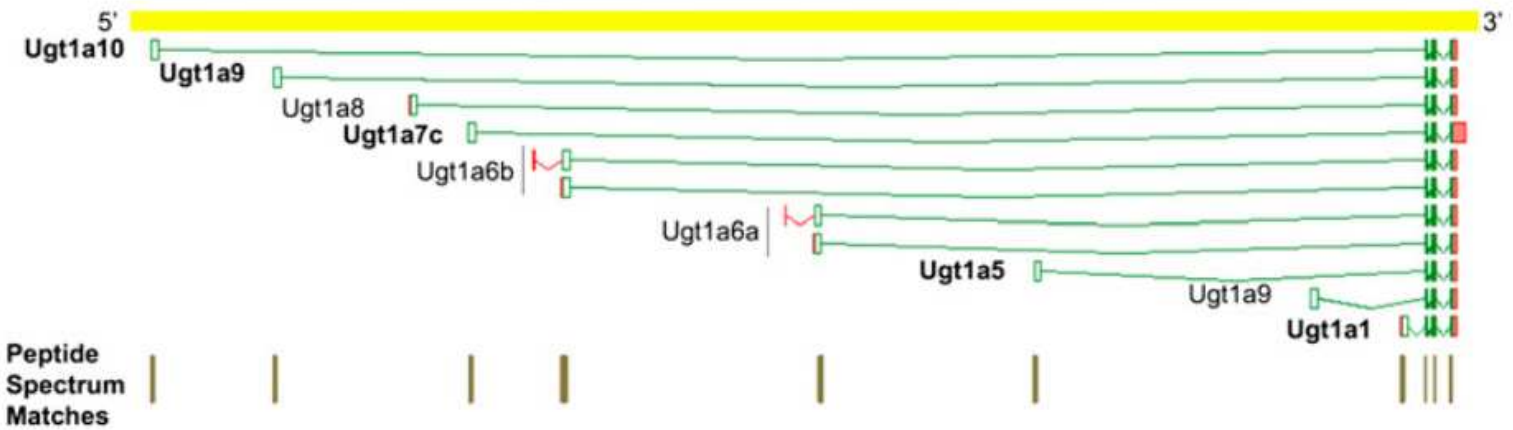


Figure 5. MS PSMs confirm the protein-coding potential of five alternatively translated products of the UDP-glucuronosyltransferase 1 family, polypeptide A6 (highlighted in bold). Ambiguous PSMs are shown for the two alternatively spliced transcripts of the Ugt1a6a and Ugt1a6b genes, respectively; and as clusters for each of the 3′ exons.

Manual identification of protein-coding novel loci and AS variants

- The more stringent criteria for the peptide identification.
  - PEP $\leq 1\%$ ($\Rightarrow$ q-value $< 0.14\%$).
  - For peptides not support by Ensembl & VEGA:
    - ≥ 2 peptides had to be identified (one having PEP < 0.01 and the second < 0.05).
- 36 MS PSMs were identified; 10 novel protein-coding loci were supported.

Table 1. Summary of novel protein-coding objects identified by PSMs

Transcript stable ID	Chromosome	Genomic clone	Mass spec tags aligning	Description	Additional Evidence
OTTMUST00000090068	6	AC165974.4	IVAAQQELLAQR RPDPGPSPLGAIPELGCRR RPDPGPSPLGAIPELGCR ENAGLLER IVAAQQELLAQRR LSRENAGLLER	Uni-exon novel orphan CDS	Strong mammalian conservation
OTTMUST00000090127	14	AC165148.2	AAEDEEVPAFFK DVAHLGPDPHR	Uni-exon novel orphan CDS	Mouse-specific transcriptional evidence
OTTMUST00000090128	7	AC113298.14	ASSAAAAAALSR AGAPGPASSPALLVLR	Uni-exon novel orphan CDS	Rodent-specific transcriptional evidence
OTTMUST00000090124	15	AC164597.11	FAKPPPPLLTSSESSTVEPPHMAR FGLHTEDLYER	CDS highly similar to de novo prediction EDL29334	Rodent-specific transcriptional evidence
OTTMUST00000090118	7	AC108827.10	SFVSHSHLQSHGR AFTHPSTVVLHK	CDS highly similar to de novo prediction EDL12440	Paralogous gene transcriptional evidence
OTTMUST00000090119	7	AC108827.10	AFAQSSSLQYHK NPPASAFQVVGLKACTTTAWPG	CDS highly similar to de novo prediction EDL12440	Paralogous gene transcriptional evidence
OTTMUST00000090503	13	AC154437.2	IITITGTQDQIQNAQYLLQNRR SLHELNPR	*Hnrnpk-2210016F16Rik* fusion object	Mouse-specific transcriptional evidence
OTTMUST00000090122	7	AC013548.13	ILGTSDSPVLFIHRPGTSGTTK APPALEGAANIDPASGSSSGQFRK LLVQPELQKPK	*Ins2-Igf2* fusion object	Mouse-specific transcriptional evidence
OTTMUST00000089966	5	AC162528.5	MDATPQDPDADFQELAK VATEQSTAEHQGPER AHSVENPAGQAPEAKPQPK FDQEAYAQTER EAPQSDSVGQQAGR ATQVSLLSARPEVATKPAVPAR GVASGHGSAVVSK HDLDAAPATK YDIVHASGER SGTEDMLEPSR	5′ Extension of novel protein (2900026A02Rik) CDS	Strong mammalian conservation
OTTMUST00000090346	X	AL450395.7	VKQEEQLQSVPAKEK YSLQPWQSTPFEQVSVTPDHDPA AAAASWSPPIDPPTSR SGLPVPSTSISSATAEDDVSPK SSEGQLPSTQPSQAFDVAK DIGQPTTTEAEVTTVQK	*Gm14569* locus	Strong mammalian conservation

Manual identification of protein-coding novel loci and AS variants

B

```
Mouse     MASAAEPPGTAAEYLQELTRIVAAQQELLAQRRRRIEELERQVAR
Rat       MASAAEPPGTAAEYLQELTRIVAAQQELLAQRRRRIEELERQVAR
Human     MASAAEPPGQAAEYLQELTRIVAAQQELLARRRRRIEELERQVAR
                 2                                3
Mouse     LSRENAGLLERHRRHLAACARRPDPGP----SPLGAIPELGCRRD
Rat       LSRENAGLLERHRRHLAACARRPDPGP----SPLGAIPELGSRRD
Human     LSRENAGLLERHRRHLAACARRPDPGPGPGPQPLGAIPELGGRRD

Mouse     K*
Rat       K*
Human     K*
```

Manual identification of protein-coding novel loci and AS variants

B

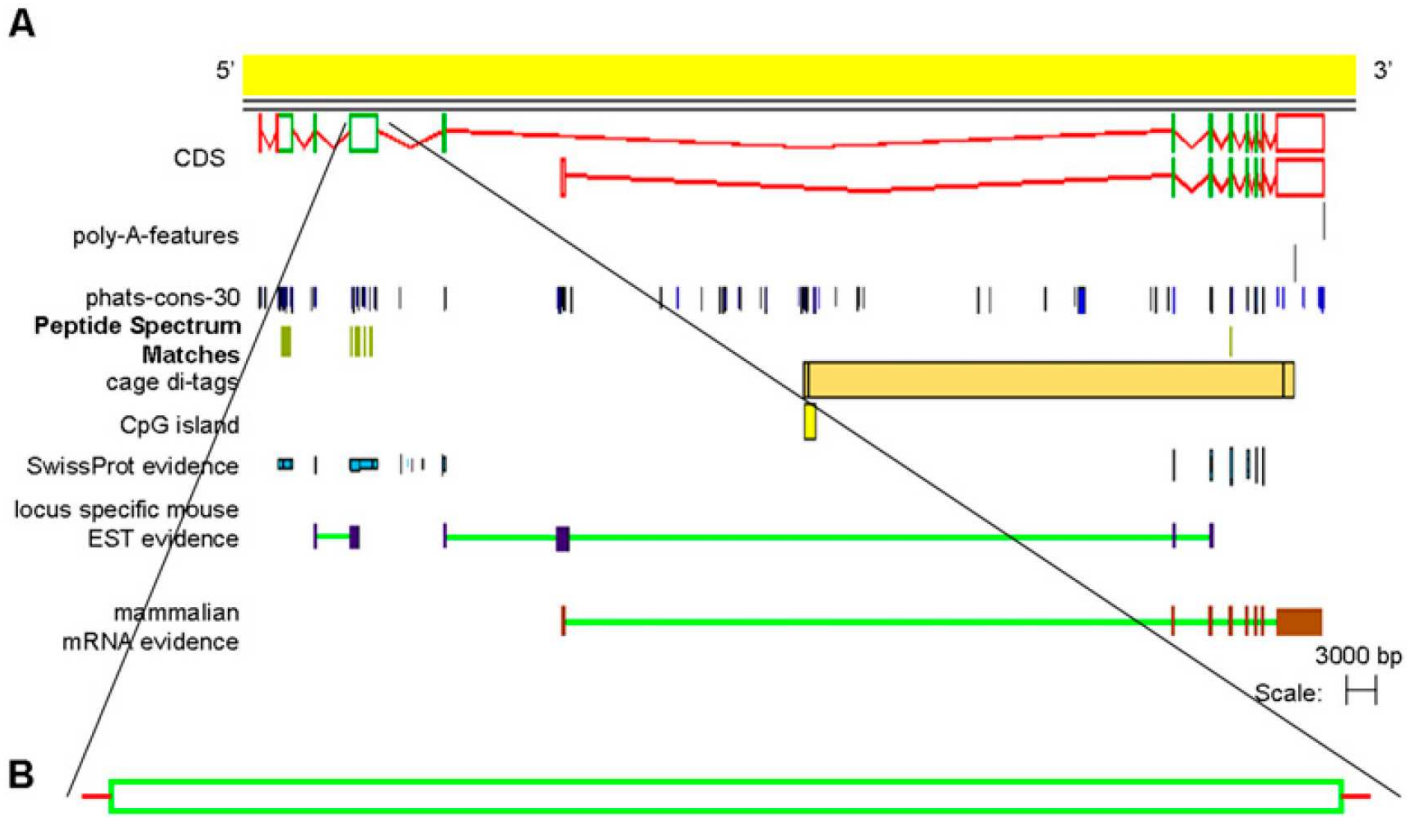
Insulin 2B peptide

```
Mouse   MALWMRFLPLLALLFLWESHPTQAFVKQHLCGSHLVEALYLVCGERGFFYTPMSRREVED
Human   MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAED

Mouse   PQDSAPSSPAATSSPRLEGPGTTTRAPPALEGAANIDPASGSSSGQFRKRILGTSDSPVL
Human   LQASALSLSSSTST-WPEGLDATARAPPALVVTANIGQAGGSSSRQFRQRALGTSDSPVL

Mouse   FIHRPGTSGTTKRLEYRGRVIKTELTVQEEEEEEENDYDDCRPLTQGSESPKAKLLVQPE
Human   FIHCPGAAGTAQGLEYRGRRVTTELVWEEVDSSPQPQGSESLPAQPPAQPAPQPEPQQAR

Mouse   LQKPKSRPVPMGIPVGKSMLV-LLISLAFALCCIAAYGPGETLCGGELVDTLQFVCSDRG
Human   EPSPEVSCCGLWPRRPQRSQN*GWQPAPASAPTARQRDTNGNPNGEVDAGASHLLGLRLV

Mouse   FYFSRPSS-RANRRSRGIVEECCFRSCDLALLETYCATPAKSERDV-STSQAVL-PDDFP
Human   LHCCLPPQ*DPVRRGAGGHPPVRLWGPRLLLQQARKPCEPSQPWHR*GVLFPQL*PGPPG
```

Preptin peptide

```
Mouse   RYPVGKFFQYDTWRQSAGRLRRGLPALLRARRGRM-LAKELKEFREAKRHRPLIVLPPKD
Human   DVLCYPRQVREGRVDPSDRASGQLPQIPRGQVLPI*HLEAVHPAPAQGPACPPACPPGSR

Mouse   PAHGG---------ASSEMSSNHQ*---------------
Human   ARQGARGVQGGQTSPSPDCSTHPRPRPRGRPPRDGQQSEV
```

Manual identification of protein-coding novel loci and AS variants



Resurrected pseudogenes

- Retrotransposed/processed pseudogenes have generally been considered as "dead on arrival".
- While the increasing number of transcribed retrotransposed genes creates additional candidate protein-coding loci [Bärtsch *et al.*, BMC Genomics 2008], there is no evidence that proteins originates from such loci.
- The MS data in this paper provides support for the translation of nine processed pseudogenes in the reference mouse genome.

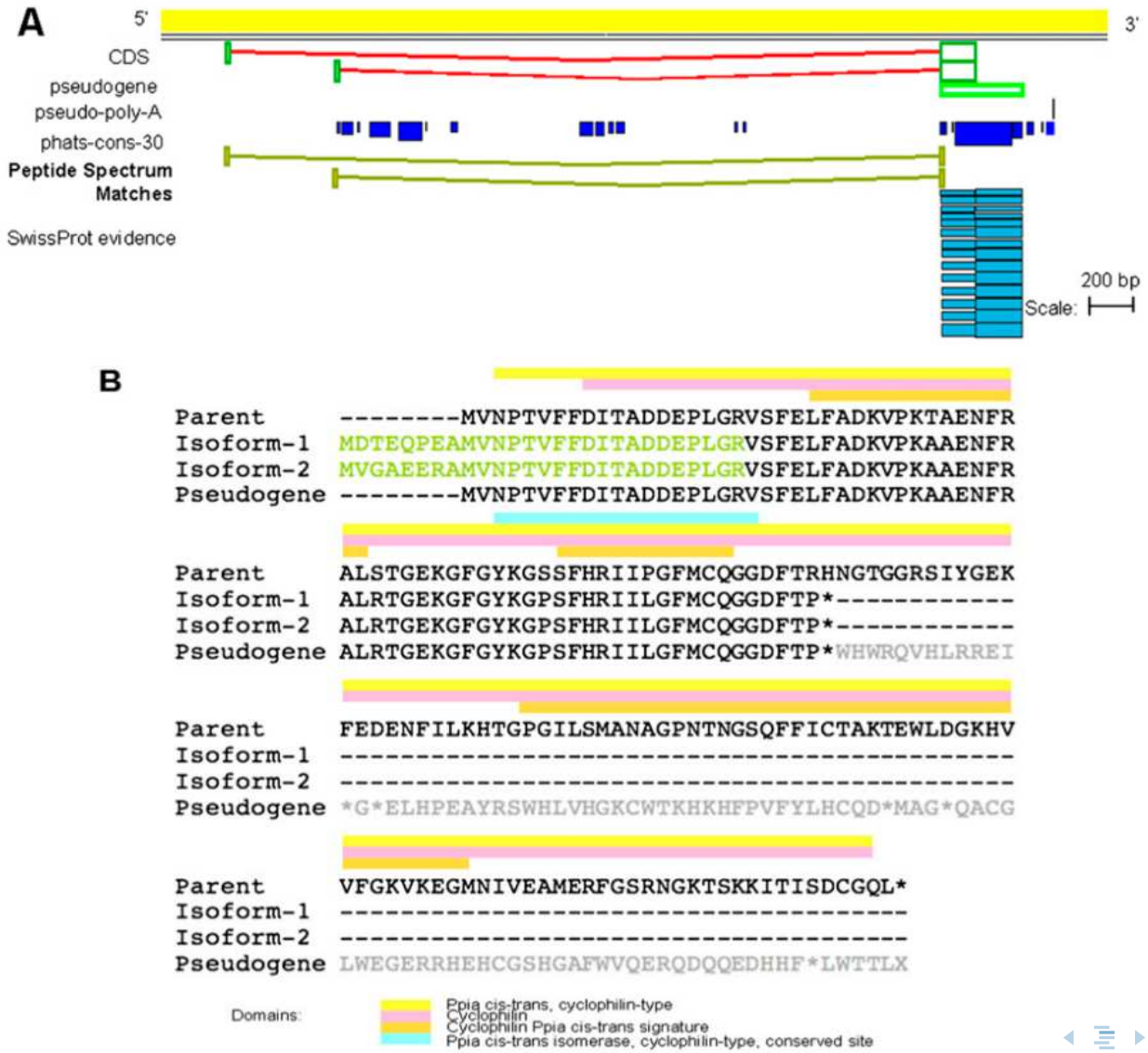
Resurrected pseudogenes (contd.)

- Each pseudogene is supported by ≥ 2 peptides.
- Unique mapping in the genome.
- Each PSM shows ≥ 2 amino acid substitutions compared with the translated parent protein sequence.
  - Each supporting PSM needed to be detected in ≥ 2 different tissues.

Resurrected pseudogenes (contd.)

- To ensure high confidence that these MS PSMs do indeed represent translations of these pseudogenic loci and NOT polymorphisms of the parent locus:
  - The residues substituted in our PSMs in comparison with the parent polypeptide are conserved in the amino acid sequences of the 1:1 rat and human orthologs;
  - No evidence of SNP/INDEL at these codon positions of the parent mouse locus.

Resurrected pseudogenes



Resurrected pseudogenes (contd.)

- Among the nine identified pseudogenes:
  - Only 2 shows syntentic ortholog in rat.
  - **None** possess human orthologs.
- However, the genes surrounding each translated mouse pseudogene show strong syntetic conservation with the equivalent rat and human loci (data not shown).
- Hypotheses to explain the detection:
  - Only relics of translation; generated until sufficient mutations are accrued $\Rightarrow$ NMD targets.
  - positive selection.
- Further investigation is required.

Discussion

- The mouse proteome is far from being saturated by MS-based peptide identifications.
  - However, MS data have become a richer and more valuable resource for genome annotation than 10 year ago.
- For the nine putative translated pseudogenic loci, whether they are able to produce functional protein is unclear.
- Among the 10 novel protein-coding loci, 8 of them can be found in the reference human genome.
  - **Note:** None of them was identified by either RefSeq or Ensembl annotation.

Thank you.