Shotgun proteomics aids discovery of novel protein-coding genes, alternative 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


**Speaker:** Joseph Chuang-Chieh Lin

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

14 May 2014
Outline

1. Introduction

2. Materials & methods

3. Results

4. Discussion
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]
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 \textit{Ins2-Lgf2} fusion object).
  - 9 processed pseudogenes (unique peptide hits): not just transcribed but translated and resurrected into new coding loci.
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

<table>
<thead>
<tr>
<th></th>
<th>Called significant</th>
<th>Called not significant</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null true</td>
<td>$F$</td>
<td>$m_0 - F$</td>
<td>$m_0$</td>
</tr>
<tr>
<td>Alternative true</td>
<td>$T$</td>
<td>$m_1 - T$</td>
<td>$m_1$</td>
</tr>
<tr>
<td>Total</td>
<td>$S$</td>
<td>$m - S$</td>
<td>$m$</td>
</tr>
</tbody>
</table>

\[
\frac{\text{no. false positive features}}{\text{no. significant features}} = \frac{F}{F + T} = \frac{F}{S'}
\]

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

\[
\text{FDR} = \frac{B}{A + B}
\]

\[
\text{PEP} = \frac{b}{a + b}
\]
Overview of Genome Annotation Pipeline

- GenoMS-DB
  - Ensembl
  - Vega
  - Augustus
  - FASTA files
  - DAS server
  - SQL analysis
  - Store Results
  - Export Peptides
  - Mascot FASTA Database(s)
  - MS2 Peaklist
  - Mascot Search
  - Mascot Percolator
  - Filter PSM Scores
  - Remove Contaminants
  - DAS track
  - Annotation
  - Results File
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;
    - predictions from Augustus.

- Ensembl Per API: to capture the peptide-genome mapping.
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 → 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).
Generator 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).
Generation 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).
Generation 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-trypptic 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)
4-Way Venn Diagram (all tryptic peptides)
Validation of Ensembl/VEGA gene annotation
Validation of Ensembl/VEGA gene annotation

- Is there a linear model fitted?
  - gene products with more potential peptides
    ⇒ sampled peptides ↑
Validation of Ensembl/VEGA gene annotation

- Is there a linear model fitted?
  - gene products with more potential peptides
  - sampled peptides ↑
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 (⇒ *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 (⇒ 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.
The more stringent criteria for the peptide identification.

- PEP \leq 1\% \Rightarrow q\text{-value} < 0.14\%.
- For peptides not support by Ensembl & VEGA:
  - \geq 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

<table>
<thead>
<tr>
<th>Transcript stable ID</th>
<th>Chromosome</th>
<th>Genomic clone</th>
<th>Mass spec tags aligning</th>
<th>Description</th>
<th>Additional Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>OTTMUST0000090068</td>
<td>6</td>
<td>AC165974.4</td>
<td>IVAQQQELLQAQR RPDPGSPSLGAPELGCR RPDPGSPSLGAPELGCR ENAGLIER IVAQQQELLQAQR LSRENAGLIER</td>
<td>Uni-exon novel orphan CDS</td>
<td>Strong mammalian conservation</td>
</tr>
<tr>
<td>OTTMUST0000090127</td>
<td>14</td>
<td>AC165148.2</td>
<td>AAEDEEVPAFFK DVAHLGDPHR</td>
<td>Uni-exon novel orphan CDS</td>
<td>Mouse-specific transcriptional evidence</td>
</tr>
<tr>
<td>OTTMUST0000090128</td>
<td>7</td>
<td>AC113298.14</td>
<td>ASSAAAAAALS AGACPASAPALLVLR</td>
<td>Uni-exon novel orphan CDS</td>
<td>Rodent-specific transcriptional evidence</td>
</tr>
<tr>
<td>OTTMUST0000090124</td>
<td>15</td>
<td>AC164597.11</td>
<td>FAKPPPPPLTSSSTEVEPPHRMAR FGLHTEODYER</td>
<td>CDS highly similar to de novo prediction EDL29334</td>
<td>Rodent-specific transcriptional evidence</td>
</tr>
<tr>
<td>OTTMUST0000090118</td>
<td>7</td>
<td>AC108827.10</td>
<td>SFVSHSLQSHGR AFTHPSTVVLHK</td>
<td>CDS highly similar to de novo prediction EDL12440</td>
<td>Paralogous gene transcriptional evidence</td>
</tr>
<tr>
<td>OTTMUST0000090119</td>
<td>7</td>
<td>AC108827.10</td>
<td>AFAQSSSLQYHK NPPASAFAQVGLKACTTTAWPG</td>
<td>CDS highly similar to de novo prediction EDL12440</td>
<td>Paralogous gene transcriptional evidence</td>
</tr>
<tr>
<td>OTTMUST0000090503</td>
<td>13</td>
<td>AC154437.2</td>
<td>IITITQDQDQNAQYLLQNRR SLHELNP</td>
<td>Hnnpk-2210016F16Rik fusion object</td>
<td>Mouse-specific transcriptional evidence</td>
</tr>
<tr>
<td>OTTMUST0000090122</td>
<td>7</td>
<td>AC013548.13</td>
<td>ILGTSDSPVLFHHRPGTSGTK APPALEGANIDPASGSSGQFRK LLVQPELQPK</td>
<td>In52-lgf2 fusion object</td>
<td>Mouse-specific transcriptional evidence</td>
</tr>
<tr>
<td>OTTMUST0000089966</td>
<td>5</td>
<td>AC162528.5</td>
<td>MDATPOQDPDADFQELAK VATEQSTAEOGPER AHSVENPAGQAEAPKQPK FDQEAYAQTER EAPOSDSCQ QAGR ATQYLLSRAPVATKPAVPAR GVASHGHSAVYSK HDLDAAPATK YDINHAMPMR SGTEDMLEP5R</td>
<td>5' Extension of novel protein (2900026A0Rik) CDS</td>
<td>Strong mammalian conservation</td>
</tr>
<tr>
<td>OTTMUST0000090346</td>
<td>X</td>
<td>AL4S0395.7</td>
<td>VKQEEQLQSVPAKEK YSLQPKWQSTFEOQVSVPDPDHPA AAAASWSPPIDPPTSR SGLVPVSTSISSATAEDVQPK SSEGQLPSTQPSOAFDVAK DIGQPTTTEAEVTVQK</td>
<td>Gm14569 locus</td>
<td>Strong mammalian conservation</td>
</tr>
</tbody>
</table>
Manual identification of protein-coding novel loci and AS variants

A

uni-exon CDS

Peptide Spectrum Matches

locus specific mouse EST evidence

Scale: 100 bp

B

Mouse

MASAAEPPGTAAYLQELTRIVAAQQELLAQRRRIEELERQVAR

Rat

MASAAEPPGTAAYLQELTRIVAAQQELLAQRRRIEELERQVAR

Human

MASAAEPPGQAAYLQELTRIVAAQQELLARRRIEELERQVAR

Mouse

LSRENAGLLEHRHRHLACARRPDPGP----SPLGAIPELGCRD

Rat

LSRENAGLLEHRHRHLACARRPDPGP----SPLGAIPELGRRD

Human

LSRENAGLLEHRHRHLACARRPDPGPQPLGAIPELGRRD

Mouse

K*

Rat

K*

Human

K*
Manual identification of protein-coding novel loci and AS variants
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 originate 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 $\geq 2$ peptides.
- Unique mapping in the genome.
- Each PSM shows $\geq 2$ amino acid substitutions compared with the translated parent protein sequence.
  - Each supporting PSM needed to be detected in $\geq 2$ 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.