Integrated Transcriptomic-Proteomic Analysis Using a Proteogenomic Workflow Refines Rat Genome Annotation*[S]

Dhirendra Kumar‡, Amit Kumar Yadav‡§, Xining Jia¶, Jason Mulvenna¶, and Debasis Dash‡¶

Proteogenomic re-annotation and mRNA splicing information can lead to the discovery of various protein forms for eukaryotic model organisms like rat. However, detection of novel proteoforms using mass spectrometry proteomics data remains a formidable challenge. We developed EuGenoSuite, an open source multiple algorithmic proteomic search tool and utilized it in our in-house integrated transcriptomic-proteomic pipeline to facilitate automated proteogenomic analysis. Using four proteogenomic pipelines (integrated transcriptomic-proteomic, Peppy, Enosi, and ProteoAnnotator) on publicly available RNA-sequence and MS proteomics data, we discovered 363 novel peptides in rat brain microglia representing novel proteoforms for 249 gene loci in the rat genome. These novel peptides aided in the discovery of novel exons, translation of annotated untranslated regions, pseudogenes, and splice variants for various loci; many of which have known disease associations, including neurological disorders like schizophrenia, amyotrophic lateral sclerosis, etc. Novel isoforms were also discovered for genes implicated in cardiovascular diseases and breast cancer for which rats are considered model organisms. Our integrative multi-omics data analysis not only enables the discovery of new proteoforms but also generates an improved reference for human disease studies in the rat model. *Molecular & Cellular Proteomics* 15: 10.1074/mcp.M114.047126, 329–339, 2016.

Characterization of bio-macromolecules in mammalian tissues is the most basic prerequisite for understanding normal physiology and disease pathophysiology. The complexity of regulatory and interaction networks created among these molecules, especially proteins, is among the highest in the brain (1, 2). Alternate splicing is a major contributor to this complexity, and it is tightly regulated at different stages of brain development (3). Although recent advances in nucleotide sequencing technologies have enabled comprehensive profiling of gene expression at the transcript level, it lacks the true depiction of protein isoforms (4). Although mass spectrometry (MS)-based proteomics provides a high-throughput method to probe proteins from biological samples, challenges remain in the analysis of MS data to harness its full power. A major limitation is the confident identification of various proteoforms expressed in a biological state or tissue. Proteoforms represent all protein forms expressed or observed from a given gene locus (5). Genome-wide detection of proteoforms in the brain would enable better genome annotation of protein coding regions. Despite microglia being an important cell type in the mammalian brain where they mediate neuro-inflammation processes and defense against pathogens, there are few comprehensive studies focused on protein discovery from microglia (6, 7). As microglia are proteomically underexplored and undergo high levels of splicing in brain, they are promising targets for the detection of novel proteins and isoforms. Accordingly, we re-analyzed publicly available proteomics data and integrated it with high-throughput transcriptomics data to refine rat genome annotations.

*Rattus norvegicus* (rat) is a widely used model organism to study various human diseases. Although the rat, mouse, and human genomes are comparable in size (8), there are large differences in the number of annotated proteins or transcripts (Table I). The relatively smaller number of annotations for the rat genome limits its utility as a model organism for studying human diseases. Although rats and mice are taxonomically close, alternative splicing of RNA leading to protein diversity is largely species-specific (9). Therefore, even though mouse proteome databases can assist in capturing some novel peptides from rat MS data, a rat-specific proteogenomic study would enable cataloguing an exhaustive list of novel peptides and proteins. Brosch et al. (10) utilized high-throughput proteomics data to annotate the mouse genome and discovered resurrected pseudogenes. A recent study by Low et al. (11)

From the ‡G. N. Ramachandran Knowledge Centre for Genome Informatics, Council ofScientific and Industrial Research-Institute of Genomics and Integrative Biology, South Campus, Sukhdev Vihar, New Delhi, India; ¶Infectious Diseases Program, QIMR Berghofer Medical Research Institute, Brisbane, Australia

Received December 23, 2014, and in revised form, November 2, 2015

Published, MCP Papers in Press, November 11, 2015, DOI 10.1074/mcp.M114.047126

Author contributions: D.K., J.M., and D.D. designed research; D.K., A.K.Y., X.J., J.M., and D.D. performed research; D.K., A.K.Y., and X.J. analyzed data; D.K. and D.D. wrote the paper.
carried out a comprehensive analysis on rat liver and highlighted the translation of many un-annotated but predicted genes. However, despite these studies, there has not been a significant effort to improve the genomic annotation of rodents by simultaneously analyzing transcriptomics and proteomics data.

Ensembl (12) is a preferred resource for annotations of eukaryotic genomes especially for model organisms like human, mouse, rat, and zebrafish. To annotate genomes, Ensembl relies on a computational pipeline that integrates ab initio prediction, sequence similarity, experimental transcripts, manual curation, and annotations from other sources to report a final set of proteins for an organism (13), and it is considered the gold standard for the annotation of many species. Besides ab initio prediction, experimental observation of peptides or proteins from a genomic locus directly confirms the existence of (one or more) alternative forms of a protein-coding gene or proteoform. Proteogenomic methods utilize these experimental observations to find protein coding features in genomic loci or, more appropriately, the transcripts. Similar proteogenomic studies have been particularly effective in improving genomewide annotations for various eukaryotic model organisms (10, 14–17). Proteogenomics is also gaining popularity as a preferred method to search for key new protein players in diseases like cancer (18, 19).

In this study, we first developed an in-house ITP pipeline for automated and integrative analysis of transcriptomics and proteomics data. ITP pipeline consists of two components. One component employs open-source algorithms for reference-based transcriptome assembly from RNA-seq reads. The other component is EuGenoSuite, which enables proteomic data analysis against this assembled transcriptome. We have shared the EuGenoSuite as open-source software, which builds and extends upon the foundations laid by our prokaryotic analysis pipeline, GenoSuite (20, 21). We demonstrate the effectiveness of this pipeline in discovering proteoforms and refining rat genome annotations in a high-throughput manner by analyzing publicly available RNA-seq and MS proteomics data. Additionally, we utilized alternative proteogenomic methods implemented in other pipelines, namely Enosi (22), Peppy (23), and ProteoAnnotator (24), for comparative analysis and comprehensive characterization of proteoforms.

### MATERIALS AND METHODS

RNA-seq Data Analysis—The RNA-seq dataset used in this work was part of the study by Merkin et al. (9) and was downloaded from the Sequence Read Archive repository at the NCBI (trace.ncbi.nlm.nih.gov). This paired end read dataset was obtained using the Illumina HiSeq 2000 and GAII platforms and represents triplicate data from nine different tissues of R. norvegicus. RNA-seq reads from the rat brain were processed by the transcriptome analysis part of our in-house pipeline (Fig. 1). First, quality check and trimming of reads were performed using the tools FastQC and Trimmomatic, respectively (25). Filtered paired end reads were then mapped on to the reference rat genome Rnor_5 from Ensembl by STAR aligner (26). Transcripts were assembled using Cufflinks (27). Reference annotation was provided as input at the mapping and transcript assembly steps. Cuffmerge from the Cufflinks suite was used to merge transcriptome from individual samples. The gffread utility from the Cufflinks was used for FASTA conversion from transcript assembly GTF. A Perl script to automate various components of transcript assembly and database creation from RNA-seq data is provided as supplemental File 1. For quantitative estimation of transcripts in individual samples, Cuffdiff from the Cufflinks suite was used to calculate "fragments per kilobase of transcript per million mapped reads" (FPKM) values for each assembled transcript.

Proteomics Data Analysis—Raw rat microglia proteomics data generated by Bell-Temin et al. (6) was downloaded from the PRIDE repository (28). The PRIDE XML files (accession numbers 18380–18393) were downloaded and converted to MGF using PRIDE-Inspector (29). MGF files were searched against a concatenated target-decoy database comprising a three-frame (+) translation of transcript assembly generated from RNA-seq reads. 115 contaminant protein sequences from the cRAP database were included in the database. The database contained 4,736,666 protein sequence entries. MS/MS spectral data were searched against this database using the OMSSA (30) and XTandem (31) peptide identification algorithms, employing the following search parameters: trypsin as the protease with one allowed missed cleavage; 20 ppm precursor ion tolerance; 0.5-Da product ion tolerance; carbamidomethylation of cysteine as fixed modification; and oxidation of methionine and peptide N-terminal acetylation as variable modifications. Peptide spectral match (PSM) results from both algorithms were integrated in a manner similar to that described for GenoSuite (20). False discovery rate (FDR)¹ was calculated using Equation 1

\[
FDR (\%) = \frac{\text{Qualified DECOY PSM count}}{\text{Qualified TARGET PSM count}} \times 100 \quad (\text{Eq. 1})
\]

Protein grouping was carried out using the parsimony method (32) as implemented in our previous work (33). In cases where multiple protein isoforms shared the same number of peptides within a group, the protein identification representing the transcript with the highest FPKM was considered representative of the group. Protein level FDR was calculated with respect to the target and decoy protein groups. Best PSM q value for each protein was considered as a protein group metric to sort target and decoy protein identifications; qualified proteins were counted with respect to best PSM q value, and FDR was estimated using Equation 2,

¹ The abbreviations used are: FDR, false discovery rate; FPKM, fragments per kilobase of transcript per million mapped reads; MGF, Mascot Generic Format; PSM, peptide spectral match; ITP, integrated transcriptomic-proteomic; PA, ProteoAnnotator.
Development of EuGenoSuite—EuGenoSuite is a multi-algorithmic proteomic analysis component of our in-house eukaryotic proteogenomic analysis pipeline ITP (Fig. 1). The combination of multiple search engines and result integration increases both sensitivity and specificity in peptide discovery (33, 34) and should be a beneficial search engines and result integration increases both sensitivity and specificity in peptide discovery (33, 34) and should be a beneficial tool for proteogenomic analysis. In EuGenoSuite, only OMSA and X!Tandem searches are configured to search MS data against extra-large eukaryotic databases in a timely fashion. It has also been successfully used in our recent work to identify missing human proteins (37).

Protein inference is a major challenge for all eukaryotic shotgun proteomics studies where multiple peptides are shared between isoforms or alternative transcripts. For this purpose, we utilized our own protein inference tool, ProteinAssembler, soon to be integrated with the ProteoStats (38) library. EuGenoSuite can be used as a generic multi-algorithmic proteomic search tool. EuGenoSuite is distributed as free software at sourceforge. The source code is available on request.

Proteogenomic Analysis—Enosi, Peppy, and ProteoAnnotator tools were also employed for proteogenomic analysis. MS data searches with these pipelines were carried out using the same search parameters as used in EuGenoSuite. These pipelines require different search databases other than the assembled transcriptome searched in EuGenoSuite. Enosi’s splice graph database (16) was built from BAM files generated by aligning raw RNA-seq reads against the reference rat genome (rn5) using STAR aligner. MS data were searched for each sample separately using MS-GF+ (40) that is integrated with Enosi. Peppy searches MS data against a six-frame translated genome and thus the reference genome was passed as input. Peppy does not allow a variable modification as a parameter for translated genome and three-frame translated transcriptome to distinguish splice junction peptides, only detectable in the translated genomic regions. Rat brain-derived RNA-seq FPKM values were also considered as a support for the remaining 220 high confident novel peptides. We calculated the average FPKM values from Cuffdiff estimates from three biological replicates and used it as an additional metric for novel loci. From the 220 novel peptides, 177 were supported by one or more transcripts with a minimum FPKM of 0.5. These novel peptides not only highlight missing annotations in the Ensembl rat genome annotations but also represent protein isoforms expressed in microglia. A genome-wide view of transcriptomic and proteomic results is presented in Fig. 2A.

RESULTS

Proteogenomic Findings Expand the Microrganal Proteomic Landscape—RNA-seq and MS data were analyzed using our in-house eukaryotic proteogenomic analysis pipeline ITP (Fig. 1). Three replicates of rat brain RNA-seq data were analyzed to catalog transcripts expressed or annotated. By using Cufflinks without applying an expression filter, a total of 101,640 unique transcripts could be assembled from ~400 million input paired end reads. These transcripts were then translated to amino acid sequences in three reading frames. This translated transcriptome database was used for protein identification from the ~2 million publicly available mass spectra generated in an independent study (6). The tandem mass spectra were generated from the rat microglia cell line HAPI (43). From an automated analysis of MS data from three biological replicates with two technical replicates each, a total of 4,431 proteins were identified at a protein FDR of <1%, of which 21 were contaminant proteins. Protein summaries from individual sample searches are provided in supplemental File 2. To minimize the suspected high false positives in eukaryotic proteogenomic analyses, we considered peptide identifications at a PSM level FDR of ≤1% and detected in both replicates of each sample. In total 11,503 peptides were identified of which 10,963 mapped to annotated Ensembl proteins, 235 to contaminant proteins, and 305 to un-annotated regions (supplemental File 3).

Two hundred and sixty five (87%) of the novel peptides mapped exclusively to a unique locus in the genome. To eliminate the possibility of novel peptides originating from known genes as a result of non-synonymous mutations, we mapped novel peptides to the annotated proteins with one residue mutated each time throughout the peptide. It revealed that 45 (~17%) of the novel peptides might arise from annotated genes, and these were not considered while annotating novel translated genomic regions. Rat brain-derived RNA-seq FPKM values were also considered as a support for the remaining 220 high confident novel peptides. We calculated the average FPKM values from Cuffdiff estimates from three biological replicates and used it as an additional metric for novel loci. From the 220 novel peptides, 177 were supported by one or more transcripts with a minimum FPKM of 0.5. These novel peptides not only highlight missing annotations in the Ensembl rat genome annotations but also represent protein isoforms expressed in microglia. A genome-wide view of transcriptomic and proteomic results is presented in Fig. 2A.
Rat Genome Re-annotation Using Proteogenomics

Fig. 1. Schematic flowchart of ITP pipeline for eukaryotic proteogenomic analysis. A reference-based transcriptome assembly from RNA-seq reads was utilized as base for proteomic search by EuGenoSuite and proteogenomic analysis.

Fig. 2. Proteogenomic findings. A, genome-wide view of transcriptomic and proteomic findings in this study. Transcripts, their FPKM values, and peptides identified by EuGenoSuite integrated in ITP were plotted against rat reference genome (rn5). B, comparison of novel peptide discovery by different eukaryotic proteogenomic pipelines. Numbers in parentheses denote novel peptides identified by the pipeline. The novel peptides identified at <1% FDR (in both technical replicates) that mapped uniquely to the genome and proteome (with one amino acid mutation allowed) from four pipelines were compared. The pipelines compared were ITP, ProteoAnnotator (PA), Peppy, and Enosi. Each pipeline identified a set of novel peptides not covered by other pipelines.
Analysis by Enosi, Peppy, and ProteoAnnotator with similar search and filter criteria identified an additional 272 novel peptides in both the technical replicates at a PSM FDR of <1%. A comparative analysis of novel peptides by individual pipelines is presented in Fig. 2B. Besides detecting a large common set of peptides, every pipeline discovers an exclusive set of novel peptides from MS data, primarily attributable to differences in the search databases used by the pipelines. Related analysis of overall peptide discovery by the pipelines is represented in the supplemental Fig. 1. Barring Peppy, the other three pipelines had comparable performances in terms of the number of peptide identifications. It should be noted that both ITP and PA use the same set of peptide identification algorithms, and thus, most of the peptide identifications are shared between these two pipelines (supplemental Fig. 1). It is reported that peptides identified by two or more pipelines have fewer errors than the ones identified by only one pipeline (34, 36). In fact, when we consider only the high confidence novel peptides (identified by two or more pipelines), performances of ITP, PA, and Enosi are similar. Although IPT identified 159 such high confidence peptides, PA, Peppy, and Enosi identified 156, 74, and 142 peptides, respectively (Fig. 2B).

Searching MS/MS data against a transcriptome database, as in ITP, or an ab initio prediction database, as in ProteoAnnotator, allows for the prediction of gene or transcript structure covering novel peptides. Similar inferences for loci covering peptides identified from genome-translated databases and splice graph databases were difficult to derive. For this, we mapped all 492 novel peptides cumulatively identified by four pipelines onto the assembled transcripts and then onto the genes predicted by Augustus and GENSCAN to infer the gene structure and their probable transcripts. 371 of these could be mapped to at least one of the transcripts, predicted genes, or annotated proteins from other sources of rat proteome. The best scoring PSMs for these were manually inspected to assess the quality of the match. Eight novel peptides were not considered for further analysis due to their poor spectral matches. The remaining 363 high confidence peptides led to the refinement of annotations for 249 gene loci in total. From the 363 novel peptides, 35 mapped to intronic regions, 44 to UTRs of annotated genes, 141 to intergenic regions, 20 to loci annotated as non-coding, 37 to different translation frames, 4 to opposite strands, and 82 were novel splice junction peptides with reference to the Ensembl genome annotation. The 129 (492–363) novel peptides that were not considered for gene loci annotation refinements were largely identified exclusively by Enosi or Peppy. It should be noted that ITP and PA consider transcripts and gene predictions, respectively, to build the search database; thus, all peptide identifications from these two pipelines also have additional supporting information. Numbers of novel peptides identified exclusively by ITP, PA, Peppy, and Enosi are 61, 71, 41, and 120, respectively (Fig. 2B). Information such as the number of selected peptides, distribution to different proteogenomic categories, and so on are provided in supplemental Table 1 and supplemental File 4 for the novel peptides identified exclusively by one pipeline. A summary of all novel peptides and the associated information about their identification and genomic loci is presented in supplemental File 4. Representative annotated PSMs of novel peptides are presented in supplemental File 5.

**Discovery of Novel Exons by Intronic Peptides**—Peptides mapping to intronic regions identified new exons for 28 annotated genes. Fig. 3 depicts an example of novel exon discovery in the gene Xpot. Two peptides mapped to a region in the genome marked as intron by both Ensembl and NCBI-RefSeq annotations. The observed transcript, peptides, and conserved orthologs suggest seven new exons for the rat Xpot gene that codes for exportin-T protein, a nuclear export receptor for tRNAs. Although it might be a variant transcript from the Xpot locus, it is the prominent proteoform expressed in microglia.

**Translation of Annotated UTRs**—Forty-four peptides mapped to exons that are annotated as non-coding and represent UTR annotations of 27 genes. Detection of peptides from such regions highlights a major limitation of the annotation efforts carried out using transcripts as the only experimental information. As depicted in Fig. 4A, the translation initiation site annotation needs to be adjusted for the Ganab gene (neutral a-glucosidase AB) where three peptides identified by ITP mapped to UTR exons. The amended exon structure is also supported by similar exon patterns in transcripts and conserved orthologs from related mammals. Additionally, these peptide identifications were verified by manual inspection of PSMs (supplemental File 5).

**Novel Peptides Rat Out Mis-annotated Pseudogenes**—Interestingly, from peptides mapping to non-coding loci, three pseudogenes were observed to be translated, as detected by two or more unique novel peptides. Eleven more pseudogene products were detected by single unique peptide hits. We excluded peptides similar to known genes, and thus these peptides were uniquely mapped to pseudogenes. Fig. 4B highlights a pseudogene locus on chromosome 4 where five peptides mapped uniquely. Transcript evidence and the identification of multiple peptides from a pseudogene, distinct from its paralogs, authenticate pseudogene translation. This locus was also supported by a GENESCAN prediction. In related organisms, orthologs of this locus are annotated for protein poly(C)-binding protein 1 (PCBP1). This protein family is known to have four similar but distinct gene members. High similarity between this locus and its paralogs might be an explanation for its annotation as a pseudogene in the absence of a translated product. Recently, using GENESCAN predictions for protein identification from rat liver, Low et al. (11) identified this protein by multiple peptides indicating widespread expression of its protein product.
Novel Genes and Gene Extensions—As many as 141 peptides fall in genomic regions that lack any gene annotation. Such peptides indicate the translation of novel genes. As shown in Fig. 5A, four transcript isoforms and nine novel peptides map to a locus where Ensembl annotation lacks any gene. However, annotation from other sources like NCBI-RefSeq and orthologs in related organisms indicates the presence of the gene Gmps that codes for guanine monophosphate synthase. Considering the complexity of eukaryotic gene models, we also verified these peptide detections with transcript structure and annotated gene overlap. A substantial part (49 of 141) of these intergenic peptides is actually part of known genes with wrongly annotated boundaries, for example the Dock2 gene locus as presented in Fig. 5B. The Ensembl annotation suggests a gene for the locus that is short in comparison with mouse or human Dock2. However, observed transcripts and five intergenic peptides mapping downstream to the annotated gene suggest similarity in exon pattern to its orthologs in mouse and human. Notably, the Dock2 gene has been associated with the esophageal adenocarcinoma (42, 44), and a better annotation for this locus will assist in clinical studies.

Novel Splice Isoforms Expressed in Rat Microglia—Eighty-two novel peptides could not be mapped to the genome or annotated protein set indicating that these are novel splice junction peptides. These novel splice variants are supported by transcripts, ab initio predictions and peptides. 49 peptides indicate novel splice variants for 43 annotated genes. An additional 27 of these exon junctions are also predicted by GENESCAN or Augustus but are not included in Ensembl annotations probably because of the lack of experimental observation. As shown in Fig. 6, two novel splice junction peptides suggest the existence of a new exon in the highly conserved tars gene product that is involved in the protein translation process. We further checked the status of the Tars gene in related genomes to find that this new exon is present in most of them. The combination of conservation and transcriptomic and proteomic observations strongly suggests the addition of this exon for tars gene annotation in the rat genome.

Novel Proteoforms Are from Genes Implicated in Diseases—For the 249 loci on which novel peptides were mapped, we queried for their reported disease associations using DisGeNET. Forty two genes were found to have significant disease associations (supplemental File 6). These human diseases include neurological disorders like amyotrophic lateral sclerosis (TARDBP), epilepsy (PCDH19), and schizophrenia (SBN01) where dis-regulation in the novel proteoforms might have some role and thus should be further investigated. Association with other diseases such as myocardial ischemia (EIF2A), osteoarthritis (IMMT, GLS, and DDX3X), and various carcinomas of non-brain tissues (DOCK2, RBM3, CNTN6, and AHNAK) might indicate new functional attributes of the novel proteoforms identified from these genes.
Annotation Resources Need Better Synchronization—Our proteogenomic analysis is based on the Ensembl genome annotation, which is a preferred source for eukaryotic annotations. However, multiple other sources also provide ab initio predictions and annotated proteomes. We compared our findings with Uniprot rat proteome, Ensembl ab initio predictions, NCBI predictions, and results from the proteogenomic analysis of rat liver by Low et al. (11). A significant fraction of novel peptides from our study were supported by the ab initio or automated prediction from Ensembl, GENESCAN, and NCBI adding confidence to our findings. Surprisingly, many of the 363 novel peptides also mapped to annotated genes or proteins from other sources. Sixty four of the novel peptides mapped to SwissProt rat proteins, an additional 30 to NCBI-RefSeq proteins, and a further 39 peptides to the genes identified in a study by Low et al. (11), where GENESCAN predictions as well as reference annotations were used to identify proteins.

FIG. 4. Proteomic detection of translation products of annotated non-coding regions. A, translation of annotated UTR exons for Ganab gene locus. Four peptides mapped to exons annotated as non-coding in both Ensembl and RefSeq reference annotation. Bar thickness in reference annotation tracks represent coding status. Thin bar, non-coding. Thick bar, coding. Track descriptions: same as Fig. 3. B, translation of pseudogene. Five peptides uniquely mapping to this locus suggest an active gene for the locus. Track descriptions: same as Fig. 3. Ensembl predicts a pseudogene for the locus, although RefSeq lacks any gene annotation. The locus is similar to PCBP1 in related organisms.

DISCUSSION

The discovery of protein isoforms from mammalian tissues is a complex, multi-step, and challenging process. Various approaches have been proposed to create a compact yet comprehensive search database for peptide and protein detection from MS data (22, 45, 46). The most promising of these utilizes RNA-seq to create a sample or tissue-specific search database. In this work, we first harnessed this approach to create a rat brain-specific search database that potentially includes all possible isoforms in the search space. The database not only represents un-annotated transcribed regions but also enables the detection of splice variants and the differentiation between coding and non-coding tran-
scripts. However, because the transcriptomics and proteomic data were not from the same sample, we utilized parallel approaches to search against splice graph (Enosi), translated genome (Peppy), and \textit{ab initio} predictions (ProteoAnnotator) to maximize proteogenomic peptide discovery. This work is potentially the first ever comparison of the performance of alternative proteogenomic pipelines for novel discoveries. Although the ProteoAnnotator performed better in the number

![Fig. 5. Novel peptides mapping to intergenic regions. Track descriptions: same as Fig. 3. A, novel gene. Nine peptides mapped to a genomic region where Ensembl lacks any gene annotation. Transcripts suggest two non-overlapping loci. However, RefSeq annotation and conservation suggest Gmpos gene for the locus. The gene and observed variants should be added to the Ensembl rat genome annotation. B, extension of annotated gene. Five peptides map to a genomic region that lacks any gene annotation in both Ensembl and RefSeq. Observed transcripts and genomic conservation for the locus suggest these peptides to be part of the Dock2 gene product.](image1)

![Fig. 6. Detection of splice variants. Track descriptions: same as Fig. 3. Two splice junction peptides map to an intronic regions for the Tars gene. Transcript suggests a splice variant with a new exon for the locus that is conserved in mouse and human genomes.](image2)
of qualified PSMs because of its smaller search database, Peppy reported the least number of peptides due to its extremely large genome-translated search database. Notably, ITP detects the highest number of novel peptides from the subset of peptides identified by more than one pipeline. It indicates that integration of the assembled transcriptome leads to high confidence peptide detection because peptides jointly identified by multiple pipelines include fewer errors.

Although our approach allows both the discovery of novel peptides and the characterization of novel transcript structure with relative FPKM expression, using multiple pipelines may be a more comprehensive strategy. Each pipeline identifies an exclusive set of peptides that bring novel discoveries as well as incorrect hits. Although we enhanced the sensitivity by combining results from multiple pipelines, we controlled the erroneous hits by applying multiple stringent filters. These quality checks are important to ensure low false positives in peptide detection because several of the novel translation events, such as splice variants, are identified using a single exon junction spanning peptides. In addition to FDR filtering, peptides were selected if detected in both replicates and mapped to a unique region in the genome. However, it should be noted that doing such comprehensive analysis by multiple tools is technically challenging and resource-intensive.

Eukaryotic proteogenomic analyses are complicated by novel peptides arising from non-synonymous mutation events. We eliminated such peptides from further consideration by iteratively mapping novel peptides with a single point mutation at any position in the peptide. Nearly 20% of the novel peptides had such similarities to known proteins and would have contributed to the false positives. This method is especially helpful in the cases where peptides map to pseudogenes similar to active paralog genes. By utilizing only the high confidence peptides, we could discover a range of novel exons, splice variants, translation of annotated UTRs, pseudogenes, new genes, and the extension of gene boundaries.

We also utilized FPKM values from RNA-seq data for confidence assessment of novel peptides identified through ITP. Because microglia are a major subset of the brain, microglial transcripts are expected to be present in RNA from the brain. However, FPKM values could vary widely, and the absence or low FPKM value of a transcript in a brain sample does not necessarily reflect protein expression in the microglia. This is further complicated by the moderate correlation between current RNA and protein quantitation methods (4). A notable example of this is in our data was the Dock2 gene product, which failed the FPKM filter of 0.5.

The majority of eukaryotic genome annotation methods depends on the detection of transcripts and/or expressed sequence tags. However, these methods inherently lack the power to annotate protein translation-specific features like translation initiation site, coding DNA sequence, UTR, or pseudogenes. This limitation can be overcome by integrating large scale MS proteomics data in the annotation process. Numerous novel peptides refine the annotation of exons and loci from non-coding to coding regions and thus improve the identification of overall coding potential in the rat genome. For example, a number of pseudogenes are observed to be transcribed and translated in our study. Our results are supported by the resurrection of pseudogenes in rodents based on a similar study in mice (10). The fact that many annotations missing in Ensembl reference annotation were present in other sources, like UniProt and NCBI, highlights a need to synchronize major genome annotation resources and to include proteogenomic studies in the process of annotation. That better genome annotation can be achieved by the integration of high-throughput multi-omics data as achieved in our work is well supported by the recent study by Low et al. (11), in which numerous novel protein forms were discovered in the rat liver by integrating genomics, transcriptomics, and proteomics data. Nearly 1/5th of the novel proteins detected in our work were also detected by Low et al. (11). Capturing genomic information allowed Low et al. (11) to detect variant peptides resulting from non-synonymous mutations in addition to detection of splice junction peptides by integrating RNA-seq data. However, analysis of brain microglia (versus liver) and comprehensive characterization by multiple proteogenomic pipelines resulted in a higher number of novel peptides and splice variants as compared with Low et al. (11).

A number of the genes for which we have now provided refined annotations were also significantly associated with diseases. These disease associations included neurological disorders like schizophrenia. Novel protein forms could add value in understanding such complex neurological diseases. Interestingly, genome annotations could also be refined for genes implicated in diseases affecting tissues other than the brain such as cardiovascular disease and breast cancer, for both of which rats are considered to be an excellent model system (39). Novel proteoforms for such genes could add microglia-specific functional roles to these genes. However, these studies would benefit the most when the rat genome has a comprehensive gene annotation of gene boundaries and isoforms.

CONCLUSION

Integration of data from multiple omic sources revealed various new proteoforms not previously reported in the rat genome. These novel discoveries highlighted the shortcomings in the present annotation of the rat genome. An improved annotation as achieved in our work creates a better reference for human disease studies in this model species. Recent accumulation of large scale rat transcriptomics and genomics data in public repositories presents an opportunity to improve annotation. However, the development of methods for the integration of high-throughput proteomics data with these public repositories will provide the most accurate and complete genome annotations. EuGenoSuite and ITP are such endeavors that have been used to comprehensively annotate
the rat genome with proteomics and transcriptomics data with the highest levels of stringency and automation.

Acknowledgments—We thank Dr. Trambayab Basak, Dr. Shantanu Sengupta, Dr. Beena Pillai and Dr. Shivshankar H. Nagaraj for helpful discussions and valuable suggestions. We thank Dr. Arjit Mukhopadhyay for critically reviewing the manuscript. We thank Devi Prasanna Das for help in collating the supplemental material. D.D. acknowledges Council of Scientific and Industrial Research-in silico Biology project for computational infrastructure.

This work was supported by Department of Science and Technology, India INSPIRE Senior Research Fellowship (to D.K.). D.D. acknowledges Genomics (BSC-0121) Project for publication charges.

REFERENCES

1. Yeo, G., Holste, D., Kreiman, G., and Burge, C. B. (2004) Variation in alternative splicing across human tissues. Genome Biol. 5, R74.

2. Xu, Q., Modrek, B., and Lee, C. (2002) Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. Nucleic Acids Res. 30, 3754–3766.

3. Mazin, P., Xiong, J., Liu, X., Yan, Z., Zhang, X., Li, M., He, L., Somel, M., Yuan, Y., Phoeben, Ch., P., Li, N., Hu, Y., Fu, N., Ning, Z., Zeng, R., Yang, H., Chen, W., Gelfand, M., and Khaitovich, P. (2013) Widespread splicing changes in human brain development and aging. Mol. Syst. Biol. 9, 633.

4. Nagaraj, N., Wisniewski, J. R., Geiger, T., Cox, J., Kircher, M., S., Wellington, A. M., Abrahamson, N., Nerieman, W. C., Haviak, P. H., Pasko, D., Tonelatto, P. J., Twigg, S., Ponting, C. P., Duarte, J. M., Rice, S., Goodstadt, L., Beatson, S. A., Ames, R. D., Winter, E. E., Webber, C., Brandt, P., Nakayuka, G., Adetobi, M., Chiaromonte, F., Elshiti, L., Eswara, P., Hardison, R. C., Hou, M., Kolbe, D., Makova, K., Miller, W., Nekrutkenko, A., Rieme, C., Schwartz, S., Taylor, J., Yang, S., Zhang, Y., Lindpanthor, K., Andrews, T. D., Caccamo, M., Clamp, M., Cuppen, E., Curwen, V., Durbin, R., Eyras, E., Searle, S. M., Cooper, G. M., Batzoglou, S., Brudno, M., Sidow, A., Stone, E. A., Venter, J. C., Payseur, B. A., Bourque, G., Lopez-Otin, C., Fuente, A. S., Chakraborti, K., Chatjeri, S., Dewey, C., Pachter, L., Bray, N., Yap, V. B., Caspi, A., Tesler, G., Pevez, P. A., Huassiler, D., Roskin, K. M., Baertch, R., Clasow, H., Furey, T. S., Hinnichs, A., Karolichik, D., Kent, W. J., Rosenbloom, K. R., Trumbower, H., Weirach, M., Cooper, D. N., Stenson, P. S., P., Ma, B. A., Nguyen, H. T., A., Arumugam, M., Sharpnay, P., Sharpnay, T., Taylor, M. S., Rihtriman, H., Mudunuri, U., Peterson, J., Guyer, M., Felsenfeld, A., Old, S., Mockrin, S., and Collins, F. (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature 428, 493–521.

9. Merkin, J., Russell, C., Chen, P., and Burge, C. B. (2012) Evolutionary dynamics of gene and isoform regulation in mammalian tissues. Science 338, 1593–1599.

10. Brosch, M., Saunders, G. I., Frankish, A., Collins, M. O., Yu, L., Wright, J., Verstraten, R., Adams, D. J., Harrow, J., Choudhary, J. S., and Hubbard, T. (2011) Shotgun proteomics aids discovery of novel protein-coding genes, alternative splicing, and “resurrected” pseudogenes in the mouse genome. Genome Res. 21, 756–767.

11. Low, T. Y., van Haeisch, S., van den Toorn, H., Giansanti, P., Cristabal, A., Toonen, P., Schafer, S., Hübner, N., van Breukelen, B., Mohamed, S., Cuppen, E., Heck, A. J., and Guérard, V. (2013) Quantitative and qualitative proteome characteristics extracted from in-depth integrated genomics and proteomics analysis. Cell Rep. 5, 1469–1478.

12. Flicek, P., Amode, M. R., Barrett, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gill, L., Giron, C. G., Gordon, L., Hourlier, T., Hunt, S., Johnson, N., Juettamann, T., Kähäri, A. K., Keenan, S., Kulesha, E., Martin, F. J., Malreul, T., McLaren, W. M., Miletich, D. N., Rag, N., Pig, M., Pidigallati, M., Prithchard, B., Pritchard, E., Riat, H. S., Ruffier, M., Sheppard, D., Taylor, K., Thomann, A., Trevisanion, S. J., Ulluo, A., Wilder, S. P., Wilson, M., Zadussa, A., Aken, B. L., Birney, E., Cunningham, F., Harrow, J., Herrero, R., Hubbard, T. J., Kinsella, R., Muffato, M., Parker, A., Spudich, G., Yates, A., Zerbino, D. R., and Searle, S. M. (2014) Ensemble 2014. Nucleic Acids Res. 42, D749–D755.

13. Cawley, C., Eyras, E., Andrews, T. D., Clarke, L., Mungin, E., Searle, S. M., and Clamp, M. (2004) The Ensembl automatic gene annotation system. Genome Res. 14, 942–950.

14. Kim, M., Sint, P. M., Getnet, D., Nirujogi, R. S., Manda, S., S., Chae-krady, R., Madugundu, A. K., Kelkar, D. S., Isserlin, R., Jain, S., Thomas, J. K., Muthusamy, B., Leal-ROJAS, P., Kumar, P., Sahasrabuddhe, N. A., Balakrishnan, L., Advani, J., George, B., Renuse, S., Selvan, L. D., Patil, A. H., Nair, V., Nag, D. B., Pignatelli, M., Prithchard, B., Pritchard, E., Riat, H. S., Ruffier, M., Sheppard, D., Taylor, K., Thomann, A., Trevisanion, S. J., Vullo, A., Wilder, S. P., Wilson, M., Zadussa, A., Aken, B. L., Birney, E., Cunningham, F., Harrow, J., Herrero, R., Hubbard, T. J., Kinsella, R., Muffato, M., Parker, A., Spudich, G., Yates, A., Zerbino, D. R., and Searle, S. M. (2014) Ensemble 2014. Nucleic Acids Res. 42, D749–D755.
thieson, T., Lemeer, S., Schnatbaum, K., Reimer, U., Wenschuh, H., Mollenhauer, M., Slotta-Huspenina, J., Boese, J. H., Bantscheff, M., Gerstmaier, A., Faerber, F., and Kuster, B. (2014) Mass-spectrometry-based draft of the human proteome. Nature 509, 582–587

16. Woo, S., Cha, S. W., Merrihew, G., He, Y., Castellana, N., Guest, C., MacCoss, M., and Bafna, V. (2014) Proteogenomic database construction driven from large scale RNA-seq data. J. Proteome Res. 13, 21–28

17. Castellana, N. E., Payne, S. H., Shen, Z., Stanke, M., Bafna, V., and Briggs, S. P. (2008) Discovery and revision of Arabidopsis genes by proteogenomics. Proc. Natl. Acad. Sci. U.S.A. 105, 21034–21038

18. Zhang, B., Wang, J., Wang, X., Zhu, J., Liu, Q., Shi, Z., Chambers, M. C., Zimmerman, L. J., Shaddock, K. F., Kim, S., Davies, S. R., Wang, S., Wang, P., Kingsinger, C. R., Rivers, R. C., Rodriguez, H., Townsend, R. R., Ellis, M. J., Carr, S. A., Tabb, D. L., Coffey, R. J., Siebos, R. J., and Liebler, D. C. (2014) Proteogenomic characterization of human colon and rectal cancer. Nature 513, 382–387

19. Halvey, P. J., Wang, X., Wang, J., Bhat, A. A., Dhwaw, P., Li, M., Zhang, B., Liebler, D. C., and Siebos, R. J. (2014) Proteogenomic analysis reveals unanticipated adaptations of colorectal tumor cells to deficiencies in DNA mismatch repair. Cancer Res. 74, 387–397

20. Kumar, D., Yadav, A. K., Kadimi, P. K., Nagaraj, S. H., Grimmond, S. M., and Dash, D. (2013) Proteomic reconstruction of Bradyrhizobium japonicum USDA110 using GenoSuite, an automated multi-algorithmic pipeline. Mol. Cell. Proteomics 12, 3388–3397

21. Kumar, D., Mondal, A. K., Yadav, A. K., and Dash, D. (2014) Discovery of rare protein coding genes in model methylotroph Methylobacterium extorquens AM1. Proteomics 14, 2780–2794

22. Castellana, N. E., Shen, Z., He, Y., Walley, J. W., Cassidy, C. J., Briggs, S. P., and Bafna, V. (2014) An automated proteogenomic method uses mass spectrometry to reveal novel genes in Zea mays. Mol. Cell. Proteomics 13, 157–167

23. Risk, B. A., Spitzer, W. J., and Giddings, M. C. (2013) Peppy: proteogenomic search software. J. Proteome Res. 12, 3019–3025

24. Ghai, F., Krishna, R., Perkins, S., Collins, A., Xia, D., Westling, J., and Jones, A. R. (2014) ProteoAnnotator–open source proteogenomics annotation software supporting PSI standards. Proteomics 14, 2731–2741

25. Bolger, A. M., Lohse, M., and Usadel, B. (2014) Trimmmomatic: a flexible trimmer for illumina sequence data. Bioinformatics 30, 2114–2120

26. Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21

27. Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. 28, 511–515

28. Martens, L., Hermjakob, H., Jones, P., Adamski, M., Taylor, C., States, D., Gevaert, K., Vandekerckhove, J., and Apweiler, R. (2005) PRIDE: the proteomics identifications database. Proteomics 5, 3537–3545

29. Wang, R., Fabregat, A., Rios, D., Oveille, D., Foster, J. M., Côte, R. G., Griss, J., Csordas, A., Perez-Riverol, Y., Reisinger, F., Hermjakob, H., Martens, L., and Vizcaino, J. A. (2012) PRIDE Inspector: a tool to visualize and validate MS proteomics data. Nat. Biotechnol. 30, 135–137

30. Geer, L. Y., Markey, S. P., Kowalak, J. A., Wagner, L., Xu, M., Maynard, D. M., Yang, X., Shi, W., and Bryant, S. H. (2004) Open mass spectrometry search software. Mol. Cell. Proteomics 15.1

31. Craig, R., and Beavis, R. C. (2004) TANDEM: matching proteins with tandem mass spectra. Bioinformatics 20, 1466–1467

32. Nesvizhskii, A. I., and Aebersold, R. (2005) Interpretation of shotgun proteomic data: the protein inference problem. Mol. Cell. Proteomics 4, 1419–1440

33. Yadav, A. K., Bhawardj, G., Basak, T., Kumar, D., Ahmad, S., Priyadarshini, R., Strub, A. K., Dain, S., and Sengupta, S. (2011) A systematic analysis of eluted fraction of plasma post immunoadfinity depletion: implications in biomarker discovery. PLoS One 6, e24442

34. Jones, A. R., Siepen, J. A., Hubbard, S. J., and Paton, N. W. (2009) Improving sensitivity in proteome studies by analysis of false discovery rates for multiple search engines. Proteomics 9, 1220–1229

35. Tanner, S., Shu, H., Frank, A., Wang, L. C., Zandi, E., Mumber, M., Pevzner, P. A., and Bafna, V. (2005) InsPeCT: identification of posttranslationally modified peptides from tandem mass spectra. Anal. Chem. 77, 4626–4639

36. Yadav, A. K., Kumar, D., and Dash, D. (2011) MassWiz: a novel scoring algorithm with target-decoy based analysis pipeline for tandem mass spectrometry. J. Proteome Res. 10, 2154–2160

37. Kumar, D., Jain, A., and Dash, D. (2015) Probing the missing human proteome: a computational perspective. J. Proteome Res.

38. Yadav, A. K., Kadimi, P. K., Kumar, D., and Dash, D. (2013) ProteoStats—a library for estimating false discovery rates in proteomics pipelines. Bioinformatics 29, 2799–2800

39. Iannaccone, P. M., and Jacob, H. J. (2009) Rats! Dis. Model. Mech. 2, 206–210

40. Kim, S., and Pevzner, P. A. (2014) MS-GF+ makes progress towards a universal database search tool for proteomics. Nat. Commun. 5, 5277

41. Bajic, V. B., and Karlin, S. (1998) Finding the genes in genomic DNA. Curr. Opin. Struct. Biol. 8, 348–354

42. Bauer-Mehren, A., Rautachka, M., Sanz, F., and Furlong, L. I. (2010) DisGeNET: a Cytoscape plugin to visualize, integrate, search and analyze gene-disease networks. Bioinformatics 26, 2924–2926

43. Cheepesthorn, P., Radov, L., Menzies, S., Reid, J., and Connor, J. R. (2001) Characterization of a novel brain-derived microglial cell line isolated from neonatal rat brain. Glia 35, 63–62

44. Dulak, A. M., Stojanov, P., Peng, S., Lawrence, M. S., Fox, C., Stewart, C., Bandla, S., Imamura, Y., Schumacher, S. E., Shefler, E., McKenna, A., Carter, S. L., Cibulskis, K., Sivachenko, A., Saksena, G., Voet, D., Ramos, A. H., Auclair, D., Thompson, K., Sougnez, C., Onofrio, R. C., Guiducci, C., Beroukhim, R., Zhou, Z., Lin, L., Lin, J., Reddy, R., Chang, A., Landrenau, R., Pennathur, A., Ogino, S., Luketich, J. D., Golub, T. R., Gabriel, S. B., Lander, E. S., Beer, D. G., Godfrey, T. E., Getz, G., and Bass, A. J. (2013) Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. Nat. Genet. 45, 478–486

45. Wang, X., and Zhang, B. (2013) customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search. Bioinformatics 29, 3235–3237

46. Nagaraj, S. H., Waddell, N., Madugundu, A. K., Wood, S., Jones, A., Mandiyam, R. A., Nones, K., Pearson, J. V., and Grimmond, S. M. (2015) PGTools: a software suite For proteogenomic data analysis and visualization. J. Proteome Res. 14, 2255–2266