BIOINFORMATICS ORIGINAL PAPER

Vol. 31 no. 1 2015, pages 25–32
doi:10.1093/bioinformatics/btu608

Genome analysis

Advance Access publication September 12, 2014

Improved rat genome gene prediction by integration of ESTs with RNA-Seq information

Liping Li[1,2], Enguo Chen[3], Chun Yang[4], Jun Zhu[2], Pushkala Jayaraman[5], Jeffrey De Pons[5], Catherine C. Kaczorowski[4], Howard J. Jacob[4,5], Andrew S. Greene[4], Matthew R. Hodges[4], Allen W. Cowley, Jr[4], Mingyu Liang[4], Haiming Xu[2,*], Pengyuan Liu[3,4,*] and Yan Lu[1,4,*]

[1]Department of Gynecologic Oncology, The Affiliated Women's Hospital, School of Medicine, Zhejiang University, Hangzhou, Zhejiang 310006, [2]Institute of Bioinformatics, Zhejiang University, Hangzhou, Zhejiang 310058, China, [3]Division of Respiratory Medicine, Sir Run Run Shaw Hospital, School of Medicine, Zhejiang University, Hangzhou, Zhejiang 310058, China, [4]Department of Physiology and the Cancer Center, Medical College of Wisconsin, Milwaukee, WI 53226, USA and [5]Human Molecular and Genetics Center, Medical College of Wisconsin, Milwaukee, WI 53226, USA

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: RNA-Seq (also called whole-transcriptome sequencing) is an emerging technology that uses the capabilities of next-generation sequencing to detect and quantify entire transcripts. One of its important applications is the improvement of existing genome annotations. RNA-Seq provides rapid, comprehensive and cost-effective tools for the discovery of novel genes and transcripts compared with expressed sequence tag (EST), which is instrumental in gene discovery and gene sequence determination. The rat is widely used as a laboratory disease model, but has a less well-annotated genome as compared with humans and mice. In this study, we incorporated deep RNA-Seq data from three rat tissues—bone marrow, brain and kidney—with EST data to improve the annotation of the rat genome.

Results: Our analysis identified 32 197 transcripts, including 13 461 known transcripts, 13 934 novel isoforms and 4802 new genes, which almost doubled the numbers of transcripts in the current public rat genome database (rn5). Comparisons of our predicted protein-coding gene sets with those in public datasets suggest that RNA-Seq significantly improves genome annotation and identifies novel genes and isoforms in the rat. Importantly, the large majority of novel genes and isoforms are supported by direct evidence of RNA-Seq experiments. These predicted genes were integrated into the Rat Genome Database (RGD) and can serve as an important resource for functional studies in the research community.

Availability and implementation: The predicted genes are available at http://rgd.mcw.edu.

Contact: hmxu@zju.edu.cn or pliu@mcw.edu or yanlu76@zju.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on December 2, 2013; revised on September 3, 2014; accepted on September 8, 2014

1 INTRODUCTION

The rat, *Rattus norvegicus*, is the first mammalian species to be domesticated for scientific research, with work dating back >150 years (Lindsey, 1979). The rat has become a widely used disease model in the fields of physiology, pharmacology, toxicology, nutrition, behavior and immunology (Aitman *et al.*, 2008). The genome sequence of the Brown Norway (BN) rat strain is the third completed mammalian genome after the human and mouse genomes (Atanur *et al.*, 2013; Gibbs *et al.*, 2004); however, its genome annotation has progressed slowly. Genome annotation is the process of attaching biological information to sequences (Stein, 2001). One of the most important steps of genome annotation is to find regions of genomic sequence that encode genes. This includes annotation of protein-coding genes and RNA genes, such as microRNAs, small nucleolar RNA and long non-coding RNA. In this article, we focused on improving annotation of protein-coding genes in the rat. Compared with the human and mouse genomes, the current build of rat genome rn5 is considerably less comprehensive. As of November 2013, the human build hg19 annotated 37 556 mRNAs and 18 667 consensus protein-coding genes, and the mouse build mm10 annotated 28 646 mRNAs and 19 962 genes. These are compared with 17 938 mRNAs and 16 744 genes in the rat build rn5.

Existing annotation methods fall into two groups according to the type of data used for gene discovery (Stanke *et al.*, 2006a). The first group consists of *ab initio* programs, which are intrinsic methods based on gene content and signal detection (Mathé *et al.*, 2002). All the programs in this group use only queried genomic sequences as input for gene discovery. The second group is composed of extrinsic methods, which comprise all programs that use data such as cDNA or protein other than the query genomic sequence to improve the accuracy of gene finding (Atanur *et al.*, 2013). A few programs, such as AUGUSTUS (Stanke *et al.*, 2006b), combine extrinsic and *ab initio* approaches by mapping protein and expressed sequence tag (EST) data to the genome to validate *ab initio* predictions (Stanke *et al.*, 2006b). AUGUSTUS is one of the most accurate programs for gene discovery due, in large part, to its ability to incorporate multiple 'hints' into gene prediction models, making it a flexible program capable of using information from various sources. These hints are pieces of extrinsic evidence, which includes the location of exons, introns and biological signals of a given input DNA

*To whom correspondence should be addressed.

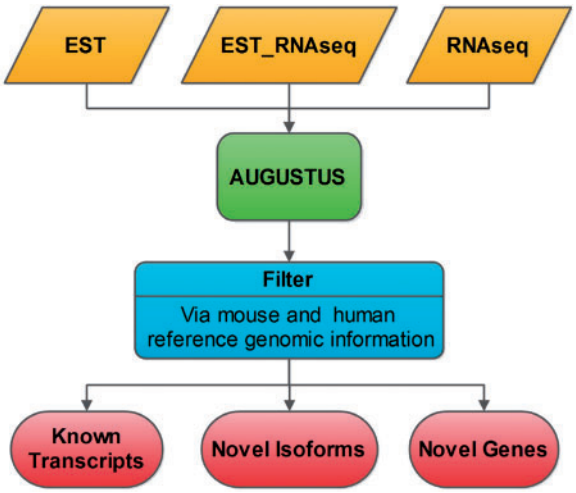

Fig. 1. A flowchart for predicting protein-coding genes with incorporation of RNA-Seq and EST data. The analysis pipeline of gene prediction includes (**1**) extracting extrinsic evidence from EST and/or RNA-Seq data, (**2**) predicting genes with AUGUSTUS and (**3**) filtering protein-coding genes with phylogenetic conservation across species

sequence and can be retrieved from EST, RNA-Seq and other data sources.

ESTs are short subsequences of a cDNA library, and are one source of extrinsic data. They are an enormous resource for determining the exon–intron structures of genes. However, ESTs are often incomplete and only reflect small pieces of full mRNAs (Mathé *et al.*, 2002). RNA-Seq (also called whole-transcriptome sequencing) is a recently emerging technology that uses the capabilities of next-generation sequencing to detect and quantify entire transcripts. This technology is a more rapid, comprehensive and cost-effective approach in the discovery of novel genes and transcripts compared with EST data from conventional Sanger sequencing (Roberts *et al.*, 2011). RNA-Seq mapping of short reads in exon junctions can reveal the precise location of exon/intron boundaries, to a single-nucleotide resolution (Wang *et al.*, 2009), and thus has the potential to significantly improve existing genome annotations (Denoeud *et al.*, 2008). Moreover, because a larger number of reads can be obtained with deep sequencing, RNA-Seq has the sensitivity to detect transcripts for genes with low expression levels, which are often missed by EST analysis (Denoeud *et al.*, 2008).

In this study, we developed an analysis pipeline in which we incorporate high-throughput RNA-Seq data from three rat tissues (kidney, brain and bone marrow) into a gene prediction model based on the AUGUSTUS algorithm (Fig. 1). We compared our predicted protein-coding gene sets with public datasets and found that RNA-Seq significantly improved genome annotation and identified novel genes and gene isoforms in the rat. These predicted genes were deposited into the RGD (http://rgd.mcw.edu) (Supplementary Fig. S1) and may serve as an important resource for the research community.

2 METHODS AND MATERIALS

2.1 RNA-Seq experiments

Tissues from bone marrow, brain and kidney were harvested from BN or Sprague Dawley (SS) rats or rats with a mixed genetic background of BN and SS (Supplementary Table S1). This allowed the identification of new loci, alternatively spliced isoforms and tissue-specific exons (in known loci). mRNA from these tissues was sequenced using the Illumina HiSeq 2000 Sequencing system with 100 bp paired-end reads. RNA-Seq library preparation and sequencing were previously described (Kaczorowski *et al.*, 2013).

Before data analysis, raw sequences with low quality (base quality <13) at both read ends were trimmed, and any reads <25 bp were removed. Reads were aligned to the rat reference genome (rn5) with TopHat v2.0.0 (Trapnell *et al.*, 2009). Each alignment was assigned a mapping quality score by TopHat, which is the Phred-scaled probability that the alignment is incorrect. Only reads with mapping quality >1 are retained for the analysis. Aligned sequence files in BAM format were subsequently used for generating extrinsic evidence for the location and the structure of genes.

2.2 Predicting protein-coding genes

To predict protein-coding genes, we generated two types of extrinsic transcribed evidence (i.e. hints) that are extracted from EST and RNA-Seq. Rat EST sequences were downloaded from the UCSC genome database (http://hgdownload.soe.ucsc.edu/goldenPath/rn5/bigZips/). The EST sequences were then aligned to the rat genome using the Blat program (Kent, 2002) to generate intron and exon hints. The BAM files from RNA-Seq contain information on the junction structure of exons and introns. The intron and exonpart hints were extracted from the BAM files using the bam2hints module in AUGUSTUS (Stanke *et al.*, 2006a). These RNA-Seq hints were generated for bone marrow, brain and kidney tissues separately. Then, each hint was used for protein-coding gene prediction, implemented in AUGUSTUS. To evaluate the impact of integration of EST with RNA-Seq on genome annotation, the hints from EST and RNA-Seq were further combined (i.e. EST_RNASeq) and used for subsequent gene prediction (Fig. 1).

After gene prediction, we evaluated whether the identified untranslated regions (UTRs) are transcripted and further extended UTRs to adjacent regions that are transcribed using RNA-Seq data. UTRs whose values of reads per kilobase per million mapped reads (RPKM) are at least 90% of their corresponding protein-coding sequence (CDS) are considered as transcribed UTRs. To extend the UTRs to adjacent transcribed regions, we first used the Cufflinks program (Trapnell *et al.*, 2010) to construct transcripts/exons from RNA-Seq reads. Then, we checked whether the upstream (i.e. 5′UTR, immediately before start codon) and the downstream (i.e. 3′UTR, immediately after stop codon) of AUGUSTUS-predicted genes are transcribed based on their RPKM values. The UTR could then be further extended to adjacent exons if there were a minimum three supporting reads spanning between the new exon and the exon containing the start or stop codon. Using the above strategy, the UTRs could be extended to further adjacent exons. The

same RPKM thresholds were applied to these newly included exons.

2.3 Filtering protein-coding genes

To more accurately identify transcripts and genes, we filtered results from AUGUSTUS by considering their phylogenetic conservation across species and homology with known genes and transcripts. Briefly, we first downloaded human (hg19) and mouse (mm10) gene annotation data from the UCSC genome database. To obtain orthologous homology, we used the UCSC liftOver utility (http://hgdownload.soe.ucsc.edu) to convert genome coordinates and genome annotations from hg19 to rn5 and from mm10 to rn5. Then, we filtered AUGUSTUS-predicted genes with <80% of orthologous homology to either mouse or human. After filtering, the transcripts predicted from RNA-Seq and EST_RNASeq were combined in each tissue (Supplementary Fig. S2).

2.4 Comparing the newly predicted protein-coding genes with rat RefSeq genes

After filtering, we evaluated the coverage and quality of predicted transcripts and genes in each of the three rat tissues based on rat RefSeq genes (rn5) as a reference set. As the current RefSeq gene annotation is not complete, we defined transcripts that show at least 95% similarity in protein-coding regions to the rat RefSeq transcripts as known transcripts, genes/transcripts that do not show any overlap with the rat RefSeq database as novel transcripts and the remaining as novel isoforms. Finally, transcripts predicted from various resources and tissues were integrated and combined with genomic annotation to generate a list of consensus gene prediction models (Supplementary Fig. S2).

2.5 Evaluating retained and/or excluded exons in transcript isoforms

In general, if multiple transcript isoforms exist in the same tissue, these retained or excluded exons lead to unbalanced reads mapped in neighboring exons. To examine these unbalanced reads, we used the module 'pileup' in the samtools tool (Li *et al.*, 2009) to extract reads at each positions within retained or excluded exons and their neighboring exons. We then used two-sample *t*-test statistics to evaluate the difference in read counts between retained or excluded exons and their neighboring exons. The Benjamini–Hochberg method was used to control false discovery rate (FDR) in the multiple testing procedures (Benjamini and Hochberg, 1995).

3 RESULTS

3.1 Sequence depth of RNA-Seq

To detect transcripts for genes with low expression levels, we combined RNA-Seq reads from different replicates and different experimental conditions from the same tissues (Supplementary Table S1). As a result, we obtained 369 million Illumina HiSeq 2000 RNA-Seq reads from six libraries of bone marrow, 600 million reads from 12 libraries of brain and 1372 million reads from 12 libraries of kidney. More than 88% of these reads could be mapped unambiguously with TopHat (Trapnell *et al.*, 2009) to the rat genome (rn5) and intron and exon junctions. Such a high depth of sequence coverage in our study allowed for the detection of extremely low-expressed transcripts with RPKM as low as 0.01 in each rat tissue (Trapnell *et al.*, 2010).

3.2 New rat genome annotation with RNA-Seq data

By incorporating EST information with RNA-Seq data from bone marrow into gene prediction models, we annotated 16 834 transcripts, including 8857 known transcripts, 5937 novel isoforms (within known genes) and 2040 new genes, that show no overlaps with any genes in the public rat genome database (rn5). Similarly, we annotated 17 883 transcripts in brain, including 9571 known transcripts, 6221 novel isoforms and 2091 new genes, and 17 076 transcripts in kidney, including 8623 known transcripts, 6245 novel isoforms and 2208 new genes (Fig. 2, Supplementary Fig. S3 and Supplementary Table S2). The analysis using the combined hints from EST and RNA-Seq (i.e. EST_RNASeq) yielded 10%, 8% and 4% more transcripts than that using RNA-Seq hints alone in bone marrow, brain and kidney, respectively. Interestingly, >25% novel isoforms and gene transcripts were uniquely identified by the analysis using hints either from RNA-Seq or from EST_RNASeq. Among novel isoforms, ~50% tend to have new start or stop codons (Supplementary Table S3). About one-third of these transcripts are expressed in a tissue-specific manner

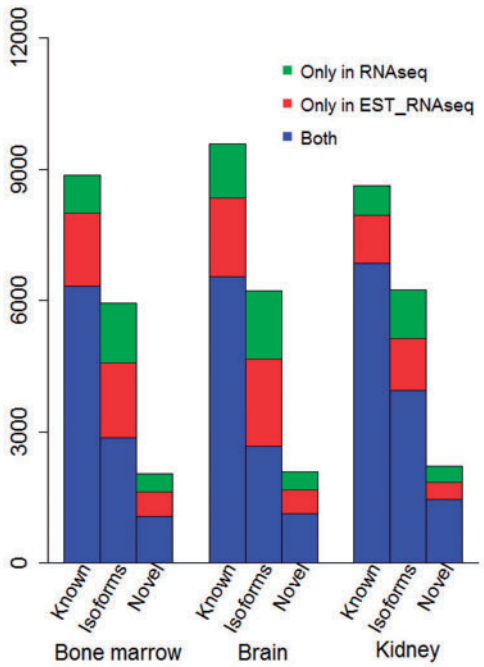

Fig. 2. Results of gene prediction from various resources and tissues. Known transcripts are those that show at least 95% similarity in protein-coding regions to the rat RefSeq transcripts, novel genes/transcripts are those that do not show any overlap with the rat RefSeq database (rn5) and the others are novel isoforms. Known: known transcripts within known genes; isoforms: novel isoforms within known genes; novel: transcripts within novel genes

(Supplementary Fig. S4). When combining transcripts from the three tissues, we identified 32 197 transcripts, including 13 461 known transcripts, 13 934 novel isoforms and 4802 new genes, which had 13 349 loci overlapping with the rat genome. In comparison, the public rat genome database (rn5) contains 17 938 transcripts that span 16 744 loci on the rat genome.

Approximately 80% of AUGUSTUS-predicted genes have transcribed UTRs that are supported by RNA-Seq reads. The predicted UTR sizes are similar between the analyses using hints from RNA-Seq and EST_RNASeq, whereas the UTR sizes vary among the three different tissues. Bone marrow and brain tend to have smaller UTRs than kidney. The predicted 5′UTRs from our analyses are, on average, two times larger than those of the rat RefSeq database (rn5) (271 versus 139 bp). The predicted 3′UTRs in the bone marrow are smaller than those of RefSeq, whereas 3′UTRs in the other two tissues are larger than those of RefSeq (Supplementary Table S4).

3.3 Improving protein-coding gene prediction

There are several improvements in the rat genome annotation resulting from incorporation of EST and RNA-Seq data into gene prediction models. First, our analysis predicted longer and more complete transcripts compared with the public database. In bone marrow, each transcript has a mean of 14 protein-coding exons and an average of 2108 bp of coding sequence. Similar statistics on gene structures were observed in brain and kidney. This is in comparison with a mean of nine protein-coding exons and an average of 1520 bp of coding sequences in the rat RefSeq transcript (Table 1). For example, Rfwd2 is a RING finger gene on rat chromosome 13, which is a major candidate gene responsible for the development of salt-sensitive hypertension in SS rats (Moreno *et al.*, 2011). The full-length transcript of Rfwd2 in the rat RefSeq (rn5) has 13 coding exons, which contain only the WD2 repeat domain and lack the RING finger domain. However, the newly predicted Rfwd2 has 20 coding exons and contains both RING finger and WD2 repeat domains. The gene structure of our predicted Rfwd2 in rat is more homologous to the same gene in mouse, human and other species (Fig. 3A).

In addition, the analysis identified more isoforms at each gene locus compared with the public database. Our prediction yielded a median of two transcripts per gene, whereas the rat RefSeq database has a median of one transcript per gene. For example, Ccnt2 (cyclin T2) encodes for an immunoglobulin-like domain-containing receptor. Our analysis identified a novel isoform of the full-length Ccnt2 gene, which is expressed in all of the three tissues, and is similar to the homologous gene in human, mouse and other species (Fig. 3B).

3.4 Identifying novel genes and isoforms

Another major advantage of our gene prediction analyses is the discovery of novel genes and isoforms. In our analysis of bone marrow, we identified 2040 novel loci that show no overlap with protein-coding genes in the public rat RefSeq database, and 5937 transcripts that are novel splice isoforms (Fig. 2). Similarly, we identified 2091 novel genes and 6221 novel isoforms in brain, and 2208 novel genes and 6245 novel isoforms in kidney.

We compared the characteristics of these novel genes with known genes that are presented in the public rat RefSeq database. We observed that our predicted novel genes had larger CDS size and more isoforms per gene than the known genes in RefSeq detected by the same analysis pipeline (Supplementary Table S5). In addition, novel genes predicted by our analysis tend to be expressed at lower levels than the known genes that overlap the public rat RefSeq database (Fig. 4). This may in part explain the reason why some of these new loci are not present in the public database.

We searched for alternative sources of evidence for these novel genes in addition to our sequencing and analyses. Among the novel genes we discovered, 719 are supported by RNA-Seq reads only, 69 are supported by EST only, 3965 are supported by both RNA-Seq reads and EST sequences and 115 are predicted *ab initio* from genomic sequences (Fig. 5). For example, Rbm27 (RNA-binding motif protein 27) encodes for a gene with a C3H1-type zinc finger and an RNA recognition motif domain. Our analysis identified Rbm27 as a novel gene on rat chromosome 18 that was not previously annotated in the public database (rn5). This is supported not only by our RNA-Seq data but also by homologous proteins in mouse and human (Fig. 6A).

We further investigated transcript evidence for these novel isoforms. If there is only one isoform for a full-length gene locus presented in a specific tissue, we directly checked whether the retained or excluded exons identified in the novel isoforms were supported by RNA-Seq or EST reads. As a result, 1017 of the 1088 exons that were either retained or excluded in novel isoforms show transcript evidence in RNA-Seq and/or EST data (Fig. 7A). If multiple transcript isoforms exist in the

Table 1. Basic characteristics of predicted protein-coding genes

Parameters	Bone marrow	Brain	Kidney	EST	RefSeq
Number of genes	11 789 (268)	12 033 (537)	11 792 (528)	10 595	16 744
Number of genes on plus strand	5957 (131)	6066 (261)	5902 (256)	5385	8401
Number of genes on minus strand	5832 (137)	5967 (276)	5890 (272)	5210	8343
Number of transcripts	16 834 (4083)	17 883 (6029)	17 076 (5326)	12 511	17 938
Mean CDS length per transcript (bp)	2108	2172	2169	1767	1519
Mean numbers of exon per transcript	13.8	14.2	14.4	11.6	8.9
Mean numbers of isoform per gene	1.4	1.5	1.4	1.2	1.1

Note: Numbers in the parentheses are numbers of genes unique to the tissue.

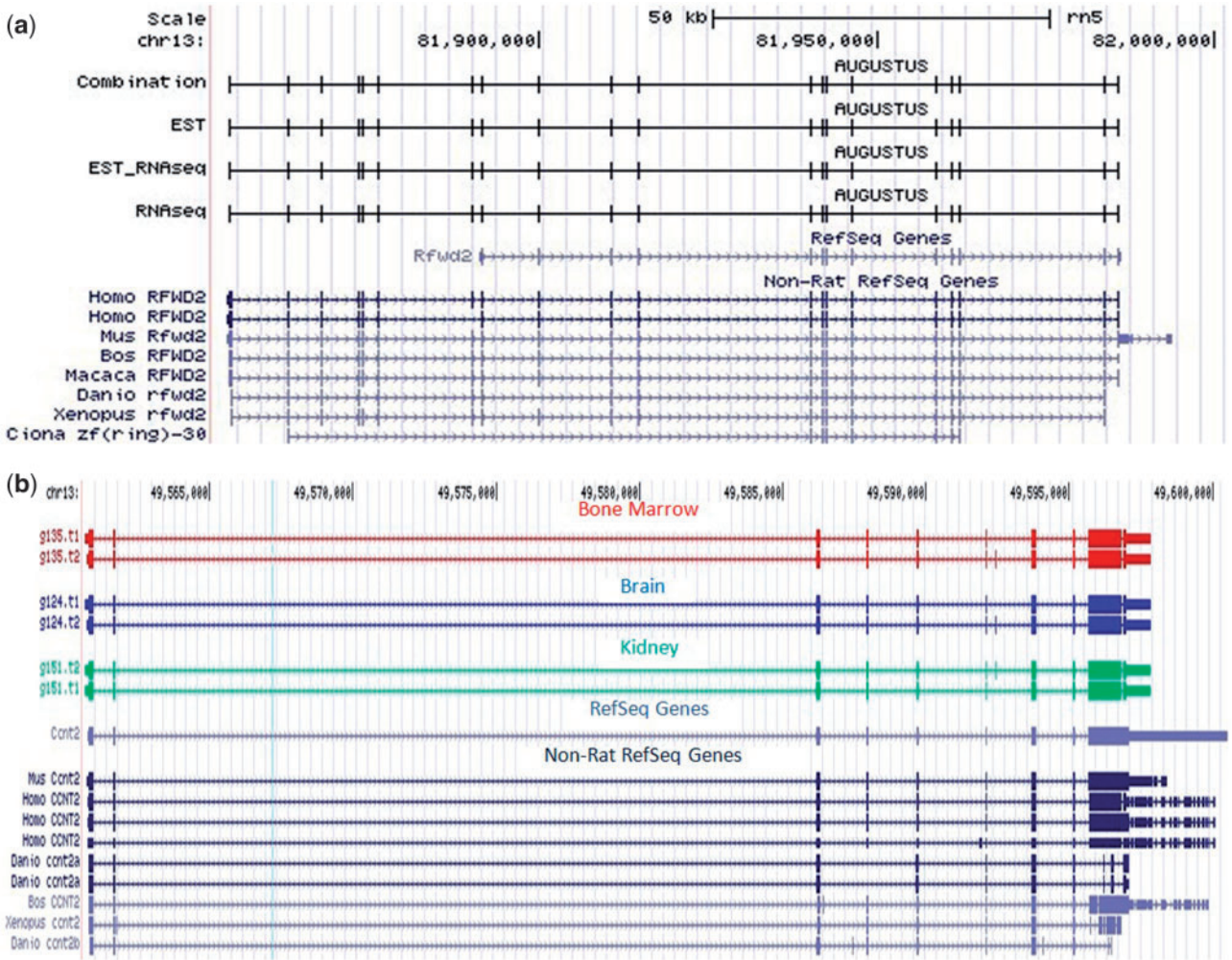

Fig. 3. Improved protein-coding prediction. (**A**) Rfwd2 is an example showing longer, complete transcripts predicted by our analysis; (**B**) Ccnt2 is an example showing more isoforms per gene predicted by our analysis

same tissue, these retained or excluded exons generally lead to unbalanced reads mapped in their neighboring exons. We thus examined whether RNA-Seq reads were disproportionally mapped in retained or excluded exons and their neighboring exons. Interestingly, 12 733 of the 12 846 exons reveal evidence for alternative splicing (FDR <0.05) (Fig. 7B). For example, myocyte enhancer factor 2a (Mef2a) encodes for a DNA-binding transcription factor that activates many muscle-specific, growth factor-induced and stress-induced genes. Our analysis identified three different transcript isoforms in kidney. One isoform contains both exons 7 and 8, and the other two contain either exon 7 or exon 8. All of the three isoforms contain adjacent exons 6 and 9. As expected, we observed lower peaks of RNA-Seq reads at exons 7 and 8 than their adjacent peaks at exons 6 and 9 (Fig. 6B).

4 DISCUSSION

Compared with the human and mouse genomes, the rat genome is not as highly annotated. In this study, we improved the prediction of protein-coding genes after incorporation of RNA-Seq evidence of transcription from bone marrow, brain and kidney tissues. Our analysis yielded nearly twice the number of transcripts in the public rat genome database (rn5) that currently contains 17 938 transcripts that span 16 744 loci. Specifically, we identified 32 197 transcripts, including 13 461 known transcripts, 13 934 novel isoforms and 4802 new genes. In general, the new analysis predicted longer and more complete transcripts and discovered more isoforms per gene locus when compared with the public database.

A large majority of novel genes and isoforms identified in our analyses resulted from the use of RNA-Seq data. This was because the RNA-Seq data contain abundant evidence from transcribed sequences that may be unavailable in ESTs. In addition, the high depth of sequence coverage of the RNA-Seq in our experiments allows the detection of extremely low-expressed transcripts in each rat tissue, which are usually missed by EST methodology. We used RPKM to measure mRNA abundance in RNA-Seq experiments, and found that novel genes tend to have smaller RPKM than known genes

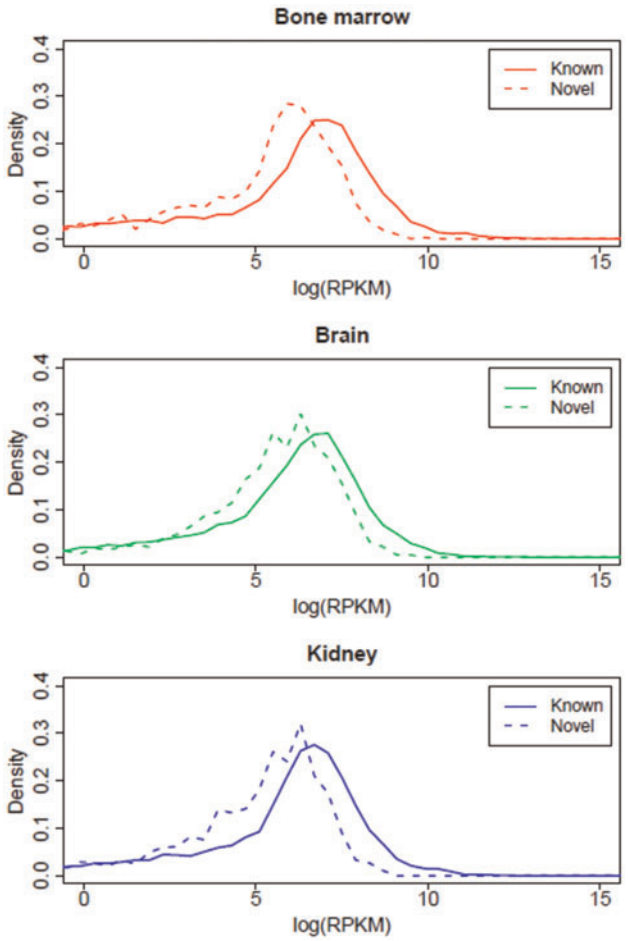

Fig. 4. Read coverage depth of known and novel transcripts. (**A**) Bone marrow, (**B**) Brain and (**C**) Kidney. RPKMs (Reads Per Kilobase of exon per Million mapped reads) are used to measure mRNA abundance in RNA-Seq experiments. The predicted novel genes tend to be expressed at lower levels than known genes in each of the three rat tissues

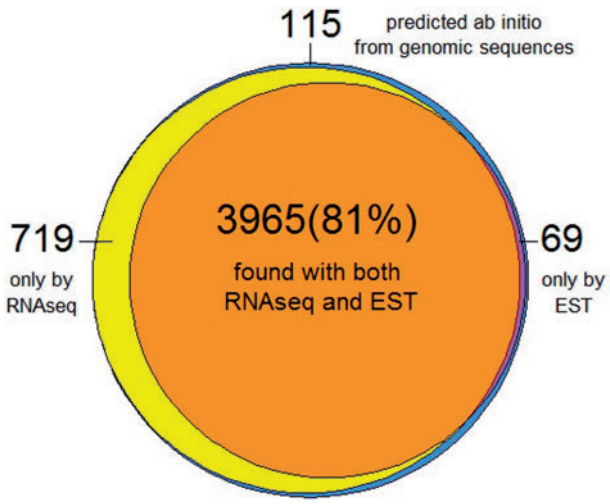

Fig. 5. Venn diagram for novel genes supported by evidence of various sources. Venn diagram shows the intersection of novel genes supported by RNAseq (yellow) or EST only (purple), or predicted *ab initio* from genomic sequences (blue), or both by RNAseq and EST (orange)

that are present in the public database. In addition to their lower expression, novel transcript isoforms within known gene loci are usually expressed in a tissue-specific manner. Because of these features, deep RNA sequencing of multiple tissues (and from different experimental conditions) may be ideal for predicting protein-coding genes in less-well-annotated eukaryotic genomes. This method can provide a more rapid, comprehensive and cost-effective discovery of novel genes and transcripts compared with ESTs from conventional Sanger sequencing (Roberts *et al*. 2011).

We noticed that although the large majority of predicted transcripts are common between RNASeq and EST_RNASeq, they also had unique sets of transcripts. Integration of different sets of extrinsic hints may lead to different predicted protein-coding genes/transcripts. AUGUSTUS is based on a generalized hidden Markov model and searches the most likely gene structure given queried genomic sequences and extrinsic hints (Stanke *et al*., 2006a). As a consequence of this, the introduction of hints changes the relative likelihood of the gene structures. The introduction of hints has two effects on the prediction of gene structure: the 'bonus effect' and 'malus effect'. The former increases the likelihood of a particular gene structure that is compatible with the hint, whereas the latter decreases that likelihood (Stanke *et al*., 2006a). Combined hints from RNA-Seq and EST (i.e. the EST_RNASeq method) do not necessarily increase the relative likelihood of a particular gene structure that was initially predicted by the RNASeq.

Several caveats from our analysis should be acknowledged. First, the rat shows high genetic similarity to both human and mouse (Aitman *et al*., 2008; Atanur *et al*., 2013; Su *et al*., 2004). In our approach, to more accurately identify transcripts and genes, we filtered results from AUGUSTUS by considering their phylogenetic conservation across species and homology with known gene and transcripts. Although such a filtering strategy is extremely useful to identify a list of transcripts/genes with high confidence (Supplementary Table S2), it could potentially overlook a small proportion of species-specific genes in the rat genome. It should also be noted that the number of the identified transcripts with high confidence is almost doubled when compared with the current public database. Second, we have demonstrated that many of novel transcript isoforms are expressed in a tissue-specific manner (Supplementary Fig. S4). In our study, this included only three tissues, and thus genome annotation could be further improved through analysis of more tissues under additional experimental conditions. Third, to detect transcripts with low expression levels, we combined RNA-Seq reads from multiple libraries of the same tissues from different genetic backgrounds. This might overlook some strain-specific transcripts. Nevertheless, such strain-specific transcripts account for <1% of total number of the identified transcripts (Supplementary Table S6).

In summary, we have demonstrated that incorporation of RNA-Seq with EST data into a gene prediction procedure significantly improves annotation of the rat genome. Our study has identified 40% more high-confidence novel genes and/or isoforms in the rat. The large majority of novel genes and isoforms are supported by direct evidence from RNA-Seq

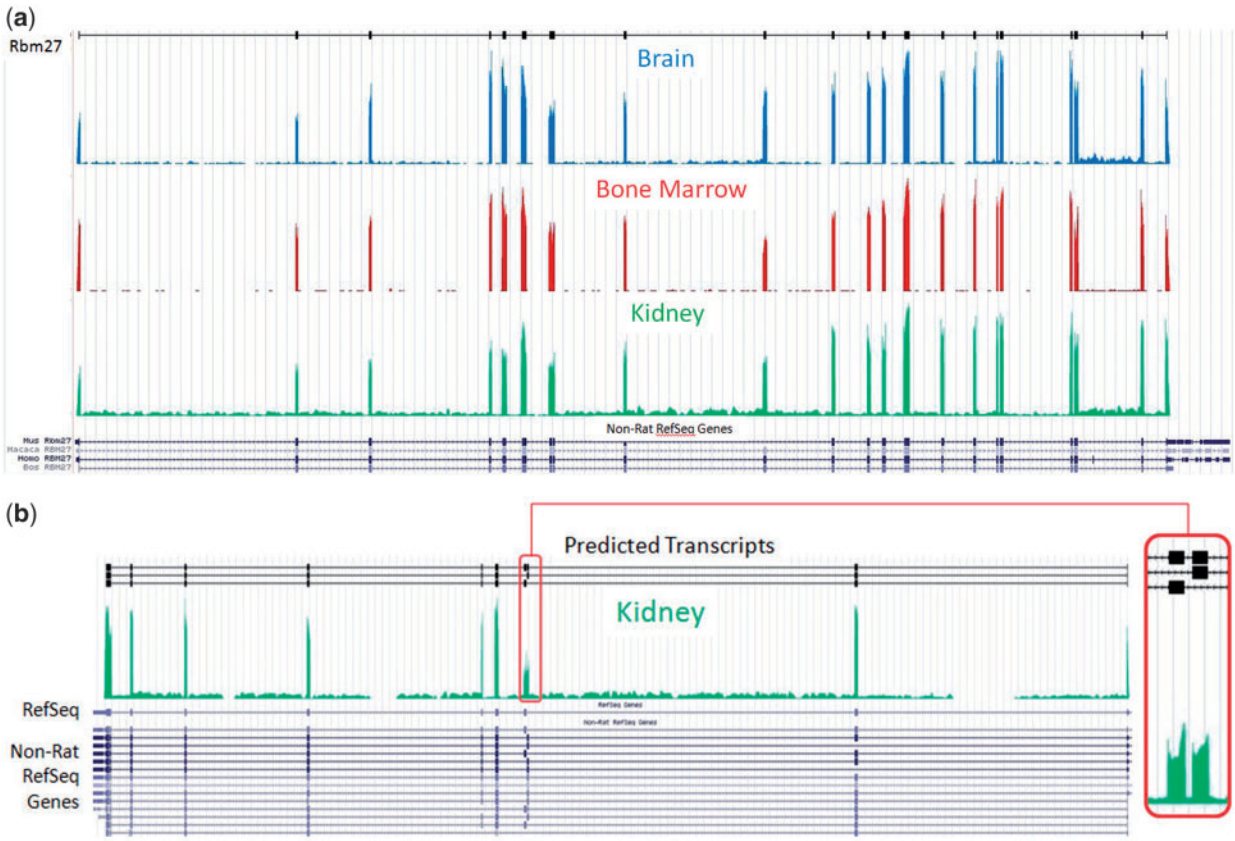

Fig. 6. Novel genes or isoforms supported by RNA-Seq data. (**A**) Rbm27 is an example showing our predicted novel genes supported by RNA-Seq data; (**B**) Mef2a is an example showing our predicted novel isoforms supported by RNA-Seq data

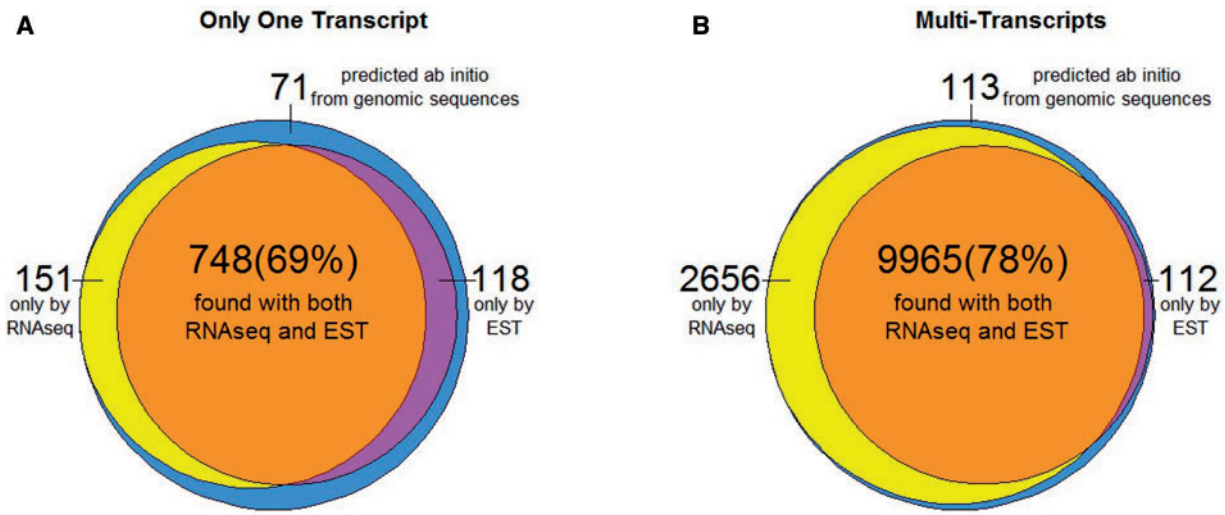

Fig. 7. Venn diagram for novel isoforms supported by evidence of various sources. (**A**) Single transcript per locus. (**B**) Multiple transcripts per locus

experiments. These predicted genes were integrated into the RGD (http://rgd.mcw.edu) and can serve as an important resource for further studies by the research community. Accessing the predicted genes in the RGD is detailed in Supplementary Fig. S1.

ACKNOWLEDGEMENTS

The authors thank Haris G. Vikis and three anonymous reviewers for reading and commenting on the manuscript. They also thank Gregory McQuestion for providing system support for various programs used in the study.

Funding: This work has been support in part by start-up from Advancing a Healthier Wisconsin Fund (FP00001701 and FP00001703), the Louisiana Hope Research Grant provided by Free to Breathe, Women Health Research Program at the Medical College of Wisconsin, National Natural Science Foundation of China (No. 31401125, 31271608, 81472420 and 81372514), and the Fundamental Research Funds for the Central Universities of China.

Conflict of interest: none declared.

REFERENCES

Aitman,T.J. *et al.* (2008) Progress and prospects in rat genetics: a community view. *Nat. Genet.*, **40**, 516–522.

Atanur,S.S. *et al.* (2013) Genome sequencing reveals loci under artificial selection that underlie disease phenotypes in the laboratory rat. *Cell*, **154**, 691–703.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.*, **57**, 289–300.

Denoeud,F. *et al.* (2008) Annotating genomes with massive-scale RNA sequencing. *Genome Biol.*, **9**, R175.

Gibbs,R.A. *et al.* (2004) Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*, **428**, 493–521.

Kaczorowski,C.C. *et al.* (2013) Targeting the endothelial progenitor cell surface proteome to identify novel mechanisms that mediate angiogenic efficacy in a rodent model of vascular disease. *Physiol. Genomics*, **45**, 999–1011.

Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.

Li,H. *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Lindsey,J.R. (1979) Historical foundations in the laboratory rat. Baker,H.J. *et al.* (eds.) Academic press, New York, NY, pp. 1–36.

Mathé,C. *et al.* (2002) Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res.*, **30**, 4103–4117.

Moreno,C. *et al.* (2011) Narrowing a region on rat chromosome 13 that protects against hypertension in Dahl SS-13BN congenic strains. *Am. J. Physiol. Heart Circ. Physiol.*, **300**, H1530–H1535.

Roberts,A. *et al.* (2011) Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics*, **27**, 2325–2329.

Stanke,M. *et al.* (2006a) Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics*, **7**, 62.

Stanke,M. *et al.* (2006b) AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. *Genome Biol.*, **7** (**Suppl. 1**), S11 11–18.

Stein,L. (2001) Genome annotation: from sequence to biology. *Nat. Rev. Genet.*, **2**, 493–503.

Su,A.I. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA*, **101**, 6062–6067.

Trapnell,C. *et al.* (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.

Trapnell,C. *et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotech.*, **28**, 511–515.

Wang.,Z. *et al.* (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**, 57–63.