ORIGINAL ARTICLE

Assessment and improvement of Indian-origin rhesus macaque and Mauritian-origin cynomolgus macaque genome annotations using deep transcriptome sequencing data

Xinxia Peng[1,2], Lenore Pipes[3,4,5], Hao Xiong[1,2], Richard R. Green[1,2], Daniel C. Jones[6], Walter L. Ruzzo[6,7,8], Gary P. Schroth[9], Christopher E. Mason[3,4,5], Robert E. Palermo[1,2] & Michael G. Katze[1,2]

1 Department of Microbiology, University of Washington, Seattle, WA, USA
2 Washington National Primate Research Center, Seattle, WA, USA
3 Department of Physiology and Biophysics, Weill Cornell Medical College, New York, NY, USA
4 The HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, Weill Cornell Medical College, New York, NY, USA
5 Tri-Institutional Training Program in Computational Biology and Medicine, Weill Cornell Medical College, New York, NY, USA
6 Department of Computer Science & Engineering, University of Washington, Seattle, WA, USA
7 Department of Genome Sciences, University of Washington, Seattle, WA, USA
8 Fred Hutchinson Cancer Research Center, Seattle, WA, USA
9 Illumina, Inc., San Diego, CA, USA

Keywords
intergenic transcripts – non-coding RNA – non-polyadenylated RNA – RNAseq

Correspondence
Dr. Michael G. Katze, Department of Microbiology, University of Washington, Box 358070, RSN 960 Republican, Seattle, WA 98195-8070, USA.
Tel.: 206 732 6135;
fax: 206 732 6056;
e-mail: honey@uw.edu

Accepted April 03, 2014.

Abstract

Background The genome annotations of rhesus (*Macaca mulatta*) and cynomolgus (*Macaca fascicularis*) macaques, two of the most common non-human primate animal models, are limited.
Methods We analyzed large-scale macaque RNA-based next-generation sequencing (RNAseq) data to identify un-annotated macaque transcripts.
Results For both macaque species, we uncovered thousands of novel isoforms for annotated genes and thousands of un-annotated intergenic transcripts enriched with non-coding RNAs. We also identified thousands of transcript sequences which are partially or completely 'missing' from current macaque genome assemblies. We showed that many newly identified transcripts were differentially expressed during SIV infection of rhesus macaques or during Ebola virus infection of cynomolgus macaques.
Conclusions For two important macaque species, we uncovered thousands of novel isoforms and un-annotated intergenic transcripts including coding and non-coding RNAs, polyadenylated and non-polyadenylated transcripts. This resource will greatly improve future macaque studies, as demonstrated by their applications in infectious disease studies.

Introduction

Rhesus (*Macaca mulata*) and cynomolgus (*Macaca fascicularis*) macaques are the most common and best-studied non-human primate (NHP) animal models, including for AIDS research. As the rhesus macaque genome sequence first became available in 2007 [1], several draft macaque genomes have been published [2, 3]. However, due to the lack of extensive species-specific transcriptional data, the annotation of macaque genomes has heavily relied on sequence homology-based strategies, usually referenced against available human annotation [1–3]. These human-centric and largely automated approaches have been

successful in annotating the core set of well-conserved protein-coding transcripts, but much less efficient for annotating the complex alternatively spliced isoforms and less-conserved non-coding RNAs (ncRNAs).

One of the reasons for the lack of NHP-specific transcriptional data is that historically the gold standard method for transcript discovery has been cDNA sequencing (by Sanger technology), a high-cost and labor-insensitive effort. Fortunately, this barrier has been largely removed with the emergence of transcriptome deep sequencing (RNAseq) [4]. RNAseq has enabled the discovery of tens of thousands of unknown transcripts in mammalian transcriptomes, as exemplified by the huge number of novel isoforms and ncRNAs recently uncovered in the human transcriptome by the ENCODE project [5]. To assist NHP genome annotation, we produced large-scale RNAseq data for each of fifteen NHP species [6] and fully released the raw sequencing data through the Nonhuman Primate Reference Transcriptome Resource (NHPRTR) website (http://nhprtr.org).

Here, using this collection of deep RNAseq data, we report a comprehensive assessment and improvement of annotations of Indian-origin rhesus and Mauritian-origin cynomolgus macaque genomes. For both macaque species, we reconstructed thousands of novel isoforms for annotated genes and thousands of un-annotated intergenic transcripts which are enriched with non-coding RNAs from RNAseq analysis of mature mRNAs. We *de novo* assembled sequencing reads which were not aligned to the reference genomes and found that many of the assembled transcript contigs corresponded to macaque transcripts 'missing' from current macaque genome assemblies. In addition, we discovered un-annotated non-polyadenylated intergenic transcripts exclusively from RNAseq analysis of total RNAs. We used this set of new annotation to analyze RNAseq data collected from macaque virus infection studies and show that a large number of newly identified transcripts were differentially expressed during SIV infection of rhesus macaques or during Ebola virus infection of cynomolgus macaques. Together, these results demonstrate the biological relevance of newly identified transcripts and the importance of improving macaque annotations.

Materials and methods

mRNAseq data preprocessing and mapping

We trimmed adaptor sequences from mRNAseq raw sequencing reads using FAR (http://sourceforge.net/projects/theflexibleadap/), and we then used Bowtie [7] to align trimmed reads to a custom collection of mammalian ribosomal RNA sequences (rRNA) and remove reads from residual rRNAs. Next we mapped remaining reads to the corresponding macaque reference genome assemblies (rhesus: rheMac2 in UCSC, cynomolgus: FR874244-FR874264 in ENA). To speed up this read mapping step, we first aligned the cleaned reads to the reference genome using the fast mapper Bowtie. Then we aligned the remaining unmapped reads to the same reference genomes using the sensitive gapped aligner GSNAP [8], which detects reads spanning splicing junctions. GSNAP also allows short insertions and deletions, which covers both genetic variations and potential errors in the reference genome assembly.

Reference-based transcript assembly and filtering against reference annotation

Using the mRNAseq reads mapped to reference genomes, we used cufflinks [9] to *ab initio* assemble macaque transcripts independent of reference annotations. We compared the reconstructed macaque transcripts against reference annotations to identify novel splicing isoforms (class code 'j' assigned by cuffcompare) for reference-annotated genes (Ensembl version 70 for rhesus and [2] for cynomolgus macaque) and un-annotated intergenic transcripts (class code 'u' assigned by cuffcompare) using cuffcompare [9]. We further removed those macaque transcripts which: (i) had no introns to minimize potential un-processed transcripts or DNA contaminants, (ii) were too short (total exonic length <200 nt), (iii) likely partial (the length of first or last exon was <100 nt), (iv) spanned two or more reference-annotated genes to minimize potentially mis-assembled transcripts, or (v) were inside introns of another newly reconstructed transcript. The coding potential of all newly identified transcripts was calculated using CPAT [10].

De novo assembly of un-mapped mRNAseq reads and alignment of assembled transcript contigs

To identify macaque transcripts which are potentially missing from the available reference genome assemblies, we *de novo* assembled the remaining un-mapped mRNAseq reads using Trinity [11]. We then used BLAT [12] to align the assembled macaque transcript contigs (200 nt or longer) to both the human (hg19 in UCSC) and the corresponding macaque reference genome sequences to identify those macaque transcript contigs which were well aligned to the human genome but not to reference macaque genomes.

To determine if the identified macaque transcript contigs were indeed 'missing' from the macaque genome

J Med Primatol **43** (2014) 317–328

assemblies, we examined the alignment of rhesus genome (rheMac2) and human genome (hg19) assemblies provided by the UCSC genome browser (http://genome.ucsc.edu). Using UCSC nets and chains tools, we initially classified the hg19-aligned contigs into three distinct categories that explain their absence from rheMac2: completely missing (the contig does not align to rheMac2 but the hg19 alignment spans the entire contig), partially missing (the contig does not align to rheMac2 but the hg19 alignment partially spans the contig), and no human–rhesus genome alignment (the contig aligns to a region in hg19 that has no available genome alignment with rheMac2). The contigs that did not fall into these previously described categories were further analyzed to see whether they were within repetitive regions, segmental duplications, or low-complexity regions.

Total RNAseq *de novo* assembly and intergenic transcript identification

We preprocessed the total RNAseq reads using an approach similar to that described for mRNAseq data. Due to the relatively smaller size of total RNAseq data, we used Trinity to *de novo* assemble the full set of cleaned total RNAseq reads without first mapping the reference genomes. We first placed the assembled macaque transcript contigs (120 nt or longer) onto the corresponding macaque reference genome sequences using GMAP [13] and grouped those uniquely aligned transcript contigs as independent transcriptionally active regions (TARs) if their genomic coordinates overlapped. We then removed any TARs if their genomic coordinates overlapped with either reference-annotated transcripts or newly identified transcripts from mRNAseq data. Transcripts were further filtered out if: (i) the transcript had the total exonic length <200 nt (with two or more exons) or <120 nt (single exon, to cover putative snoRNAs or the like) or (ii) the length of the last or the first exon was <100 nt.

Next we selected the subset of TARs which had higher expression abundances in total RNAseq data than the corresponding mRNAseq data. Because the sequencing depths were too different between two datasets, we used Picard (http://picard.sourceforge.net) to randomly sample 3–4 sets of 50 million reads from mRNAseq data and 3–4 sets of 50 million reads from total RNAseq. Next we used HTSeq (http://www-huber.embl.de/users/anders/HTSeq/doc/overview.html) to obtain raw read counts for all TARs and reference-annotated genes. We normalized the raw read counts by the corresponding total read count, that is, the sum of raw read counts of all genes/TARs. For each gene/TAR, we calculated a metric Rtm, which was defined as the ratio between the minimum of normalized total RNAseq read counts and the maximum of normalized mRNAseq read counts. We calculated the distributions of the Rtms for genes/TARs from different annotation sources. We chose a threshold for Rtm which showed the best separation between different annotation sources. We selected the subset of TARs which had much higher Rtms as un-annotated intergenic transcripts derived from total RNAseq data, that is, they were assembled only from total RNAseq data and highly enriched in total RNAseq data.

Availability

All of the transcripts identified from this study can be downloaded from the NHPRTR website (http://nhprtr.org).

Results

Overview of macaque RNAseq data processing

In total, NHPRTR generated over 7.6 billion short sequence reads (3.5 billion 2 × 100 nt paired-end reads and 681 million single-end 100 nt reads) for the two macaque species studied here. This was accomplished by using complementary mRNAseq (mature polyadenylated mRNAs) and total RNAseq (rRNA depleted total RNAs) protocols [6]. Considering the large volume of the RNAseq dataset and the draft quality of available macaque genomes, we devised a streamlined analytic strategy to reconstruct macaque transcripts from the large-scale macaque RNAseq data (Fig. 1). Our strategy combined reference mapping with genetic variation aware steps and *de novo* transcript assembly. Also, mRNAseq data and total RNAseq data were processed independently and then contrasted to uncover non-polyadenylated transcripts.

Systematic *ab initio* transcript reconstructions with mRNAseq data reveal thousands of novel isoforms and intergenic transcripts in both macaque genomes

Based on mRNAseq reads aligned to reference macaque genome assemblies, we reconstructed macaque transcripts using the assembler Cufflinks separately for each macaque, independent of their reference genome annotations (Methods). To minimize potential contaminants of unspliced transcripts, we only kept spliced, that is, with at least one intron, transcripts. As shown in Table 1, for rhesus macaque, we identified 59,116 novel isoforms for 13,344 annotated genes and 4783 intergenic

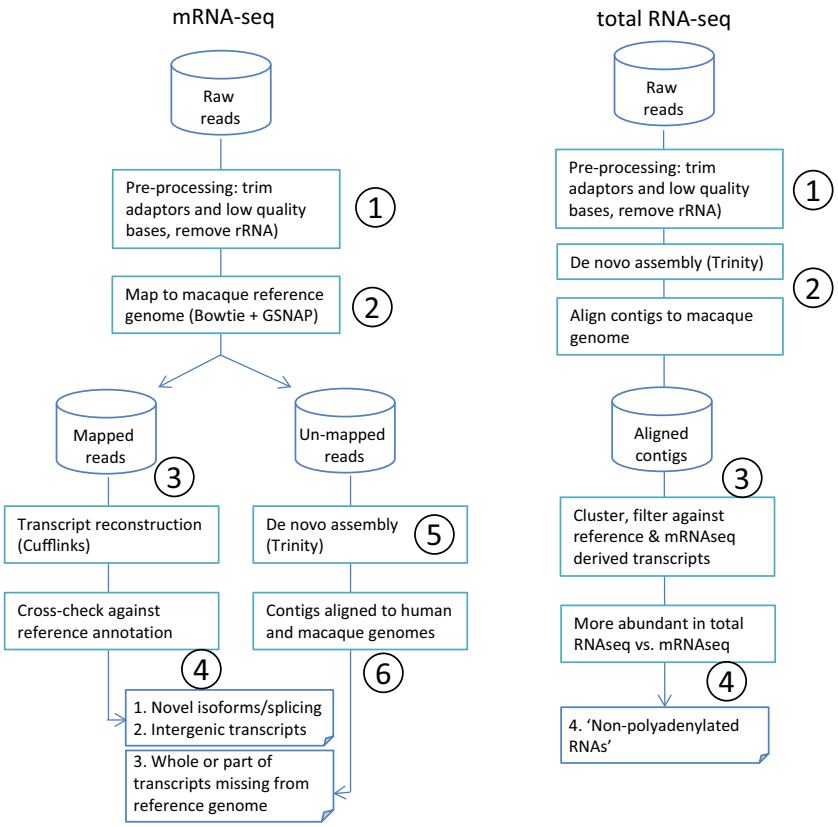


Fig. 1 Overview of the strategy for deriving macaque transcripts using species-specific RNAseq data. Left, the workflow for mRNAseq data. Step 1, the raw sequencing reads were cleaned through preprocessing procedures. Step 2, the cleaned reads were mapped to the corresponding macaque reference genome assembly. To speed up the process, the fast mapper Bowtie was first used to quickly place un-spliced reads onto the genome. Un-mapped reads were then aligned using the gapped aligner GSNAP, which also allowed indels. Step 3, macaque transcripts were reconstructed with mapped reads *ab initio* using Cufflinks, independent of reference annotation. Step 4, newly constructed macaque transcripts were compared against available reference annotation. Step 5, in parallel to Step 3, after GSNAP in Step 2, all un-mapped reads were *de novo* assembled into transcript contigs using Trinity. Step 6, contigs were aligned to both the human and macaque genomes to identify transcripts not covered by the corresponding macaque reference genome. Right, the workflow for total RNAseq data. Step 1, the raw sequencing reads were cleaned through preprocessing procedures. Step 2, the cleaned reads were directly *de novo* assembled into transcript contigs using Trinity. Step 3, contigs were aligned to the corresponding macaque reference genome and filtered against both reference-annotated and mRNAseq-derived transcripts to identify transcripts likely unique to total RNAseq data. Step 4, from these total RNAseq-derived transcripts, a subset of total RNAseq-enriched transcripts were selected by comparing the expression abundance measured by total RNAseq (cleaned total RNAseq reads were also aligned to reference genome separately) and mRNA seq. This subset of transcripts was considered as putative non-polyadenylated transcripts.

transcriptionally active regions (TAR, due to the inability to define the actual gene start and end boundaries). For cynomolgus macaque, we identified relatively fewer novel isoforms for reference-annotated genes but many more intergenic TARs. This was likely due to more genes annotated for rhesus plus the lower RNAseq coverage for the cynomolgus macaque genome (Table 1 in [6]). Figure 2 shows a few of examples of newly identified rhesus transcripts.

As shown in Fig. 3A, all of these newly constructed transcripts for rhesus macaque are longer than 200 nt, and most of them tend to be longer than 1 kb, similarly for both novel isoforms for reference-annotated genes and intergenic transcripts. But the coding potentials between novel isoforms and intergenic transcripts were very different (Fig. 3B). About 90% of novel isoforms were predicted to be protein coding, while about 80% of intergenic transcripts were predicted to be non-coding. Similar observations for cynomolgus macaque transcripts are shown in Fig. 3C, D. This shows that most of these intergenic transcripts are likely to be long non-coding RNAs. This also suggests that in general macaque protein-coding genes were much better covered by the

J Med Primatol **43** (2014) 317–328

Table 1 Summary of macaque annotations derived from RNA seq data

Species	Annotation source		Gene/TAR	Transcript
Indian rhesus macaque	Reference annotation		30,246	44,725
	mRNAseq:	Novel isoforms of	13,344	59,116
	Reference-based assembly (Cufflinks)	reference genes		
	Spliced (at least one intron)	Intergenic	4783	10,010
	total RNAseq:	Intergenic	2347	5561
	De novo assembly (Trinity)			
Mauritian cynomolgus macaque	Reference annotation		16,656	27,062
	mRNAseq:	Novel isoforms of	2234	9290
	Reference-based assembly (Cufflinks)	reference genes		
	Spliced (at least one intron)	Intergenic	7725	18,502
	total RNAseq:	Intergenic	2260	6143
	De novo assembly (Trinity)			

TAR, Transcriptionally Active Regions.

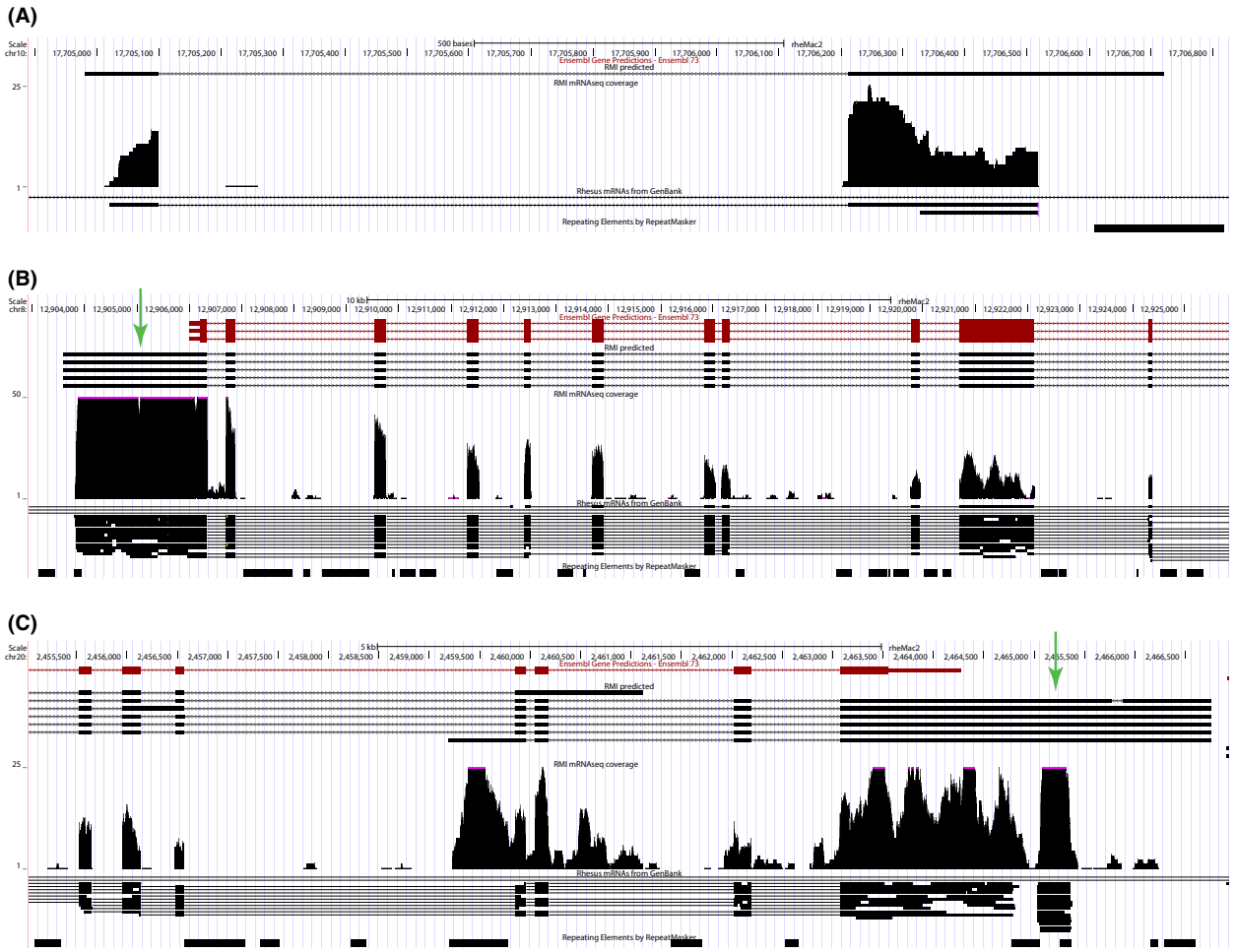


Fig. 2 UCSC genome browser view of example transcripts newly constructed from mRNAseq data for rhesus macaque. (**A**) An intergenic transcript on chromosome 10 (genomic position: chr10:17,704,891–17,706,826). From the top to the bottom, the tracks are as follows: (i) Ensembl annotation, (ii) rhesus transcripts constructed from mRNAseq data, (iii) read coverage of mRNAseq data (generated from 50 million randomly sampled reads), (iv) aligned rhesus mRNAs from Genbank, and (v) repeats by RepeatMasker. (**B**) An example of extended UTRs (indicated by green arrow) by novel isoforms, tracks are as follows: (A) The genomic location shown is chr8:12,902,941–12,925,847. (**C**) Another example of extended UTRs, tracks are as in (A) The genomic location shown is chr20:2454513–2466937.

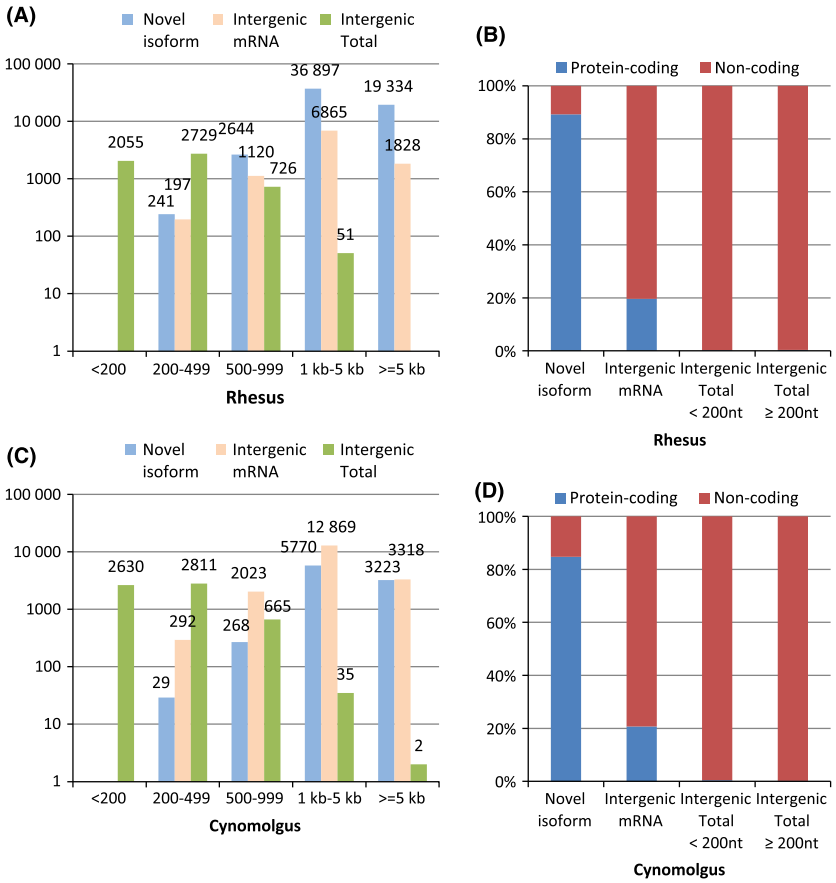


Fig. 3 Characterization of newly identified macaque transcripts. (**A**) The length distribution of newly identified transcripts for rhesus macaque, separately for each category. Novel: novel isoforms for reference-annotated genes. Intergenic mRNA, intergenic transcripts from mRNAseq data. Intergenic Total, intergenic transcripts from total RNAseq data. The x-axis shows the individual bins for transcript length. The y-axis is the number of transcripts within each bin ($\log_{10}$ scale), and the actual number of transcripts is labeled on top of each bar. (**B**) The percentages of coding vs. non-coding transcripts based on coding potential prediction, separately for each category of newly identified transcripts. Transcripts from Intergenic Total were divided into two subcategories by the transcript length, shorter than 200 and 200 nt or longer. (**C**) As in panel A, for cynomolgus macaque. (**D**) As in panel B, for cynomolgus macaque.

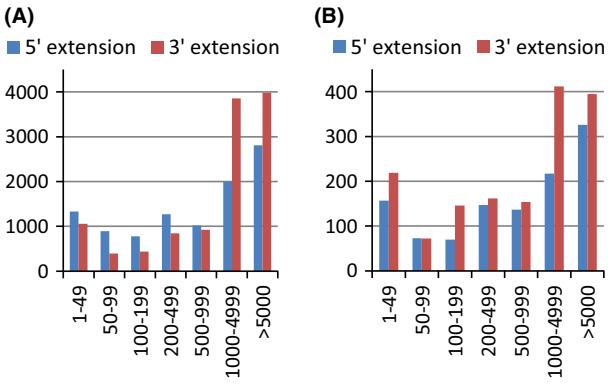
reference annotations, while the annotations of non-coding RNAs were very insufficient and can be improved by leveraging RNAseq data as collected here.

For reference-annotated genes, these novel isoforms also often extended beyond the 3′ or 5′ end of the corresponding gene, ranging from a few bases to over 5 kb (Fig. 4). In general, larger (>1 kb) extensions tended to be located at 3′ ends, suggesting the preference was likely due to oligo-dT priming of the 3′ end of mature mRNAs. As shown in Fig. 2B, C, some of those extensions were also supported by expressed sequence tag (EST) sequences, demonstrating that incorporation of previously collected ESTs and species–specific transcription, such as the RNAseq data here, will greatly facilitate the annotation of full-length macaque transcripts.

De novo assembly of un-mapped RNAseq reads recovers thousands of transcripts, partially or completely, missing from current macaque genome assemblies

On average, about 10% of cleaned mRNAseq reads were not aligned onto the macaque reference genome assemblies. We then investigated if these un-mapped reads represent bone fide macaque transcripts which were not fully covered by the available reference genome assemblies. We *de novo* assembled these un-mapped reads into transcript contigs and then aligned the assembled contigs to both the corresponding macaque genome assemblies and the human reference genome (Methods). In total, we obtained 51,408 contigs from rhesus and 21,931 contigs from cynomolgus mRNAseq data, all longer than 200 nt,

J Med Primatol **43** (2014) 317–328



Rhesus	
5'	3'
10 101	11 497
75.7%	86.2%

Cynomolgus	
5'	3'
1127	1560
50.4%	69.8%

Fig. 4 Summary of the extension of the ends of annotated genes by newly predicted isoforms. (**A**) Summary for rhesus macaque. The number of bases extended was based on genomic positions. Annotated genes were binned based on the number of bases extended beyond the annotated gene start and end positions (x-axis). The table below shows the total number of genes with 5′ or 3′ end extended and the percentage of the total number of genes with novel isoforms. (**B**) Summary for cynomolgus macaque, similarly as in (**A**).

which were well aligned to the human genome, but not to either macaque genome assembly, pointing to potential sequencing errors or gaps in current macaque assemblies.

We then examined the placement of these contigs on the alignment of the rhesus and human genome assemblies available from the UCSC genome browser. It turned out that 95% of these rhesus contigs were aligned to the alignment blocks where the rhesus genome assembly had gaps (Fig. 5A). In reference to the human annotation, these contigs overlapped 2272 annotated human genes, 52 of which were completely covered by these contigs. As an example (Fig. 6), the human gene CD1C was completely covered by multiple *de novo* assembled rhesus macaque contigs, representing potentially novel isoforms, while located within a gap in the rhesus genome assembly spanning the whole gene. Interestingly, these rhesus transcript contigs even extended beyond the annotation of the human CD1C gene, suggesting the UTR annotation of this human gene might not be complete either. Overall, for most of these human genes, these macaque contigs covered a relatively small portion (10–20%) of their exonic regions (Fig. 5B), suggesting the current rhesus macaque assembly contained many, but mostly relatively small, gaps or errors. In summary, our results show the RNAseq data contain many macaque transcript sequences which are missing from current macaque genome assemblies, and we were able to successfully recover these macaque transcripts using the analytical strategies developed here.

Thousands of macaque transcripts, mostly predicted to be non-coding RNAs, were exclusively uncovered by sequencing total RNAs but not mRNAs

As the mRNAseq data described above were designed to capture polyadenylated mature mRNAs through the use of oligo-dT priming during library creation, we analyzed total RNAseq data collected on the total

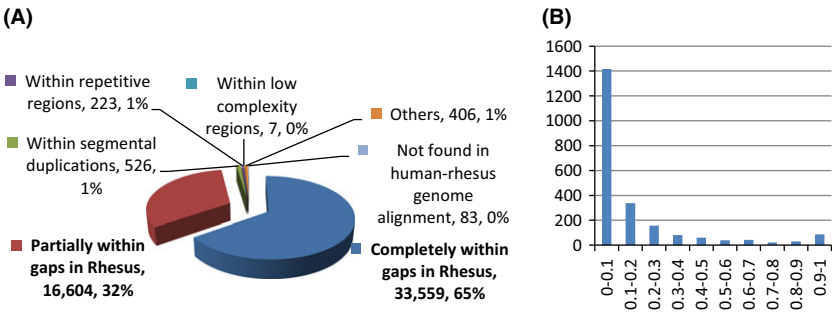


Fig. 5 Characterization of rhesus macaque transcript contigs derived from mRNAseq reads un-mapped to the rhesus genome assembly. (**A**) Classification of these macaque transcript contigs based on their placement on the rhesus–human genome alignments. The corresponding human genomic positions where the macaque contigs were aligned were classified as follows: (i) covered by a gap in rhesus genome assembly completely (completely within gaps in rhesus) or (ii) partially within gaps in rhesus. The remaining genomic locations were based on the overlap with human genome annotations: segmental duplication, repetitive regions, low-complexity regions, others. In total, 51,408 *de novo* assembled rhesus transcript contigs were aligned to the human genome, but not to the rhesus genome assembly. (**B**) Distribution of the portions of annotated human genes covered by at least one of these rhesus transcript contigs. In total, these transcript contigs overlapped 2272 annotated human genes.

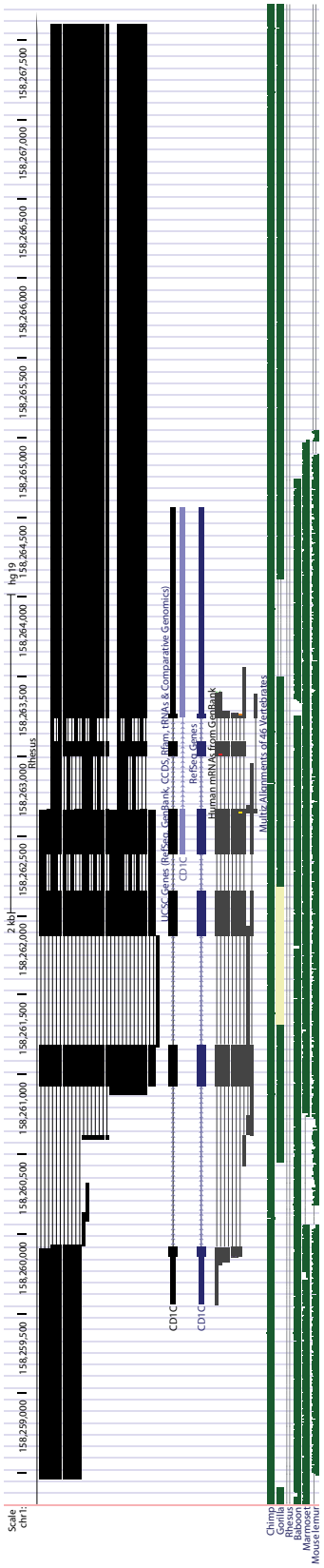


Fig. 6 An example of rhesus genes missing from the current rhesus genome assembly visualized using the UCSC genome browser. CD1C is a CD1 family member, structurally related to the major histocompatibility complex proteins, and mediates the presentation of primarily lipid and glycolipid antigens of self- or microbial origin to T cells. From top to bottom, the tracks shown are as follows: (i) the alignment of *de novo* assembled macaque transcript contigs to the human genome, (ii) UCSC human gene annotation, (iii) Refseq human gene annotation, (iv) the alignment of human mRNAs to the human genome, and (v) the alignment of multiple genomes against the human genome. Horizontal lines in the alignment show the gaps in the alignments for the corresponding genome which exist in the rhesus macaque assembly.

RNAs from the same samples to uncover non-polyadenylated macaque transcripts. To do so, we first assembled macaque transcripts from the total RNAseq data independently and then applied an *in silico* subtraction strategy. The subtraction strategy included the following: (i) the removal of those total RNAseq-derived transcripts overlapping any mRNAseq-derived transcripts or reference-annotated transcripts and (ii) the removal of those total RNAseq-derived transcripts that did not show higher expression abundance as quantified by total RNAseq compared to mRNAseq.

In brief, to focus on transcripts enriched by total RNAseq analysis, we quantified the expression abundance of the remaining TARs separately using mRNAseq and total RNAseq data. For each gene or newly identified TAR, we calculated a ratio of the expression abundance measured by total RNAseq to that by mRNAseq (Rtm). As shown in Fig. 7A, B, TARs derived from total RNAseq data alone were well separated from mRNAseq-derived TARs based on the metric Rtm, for both rhesus and cynomolgus macaques. Total RNAseq-derived TARs tended to have much higher Rtm, an indication that these TARs were transcribing non-polyadenylated transcripts which were not adequately captured by mRNAseq. Two examples are shown in Fig. 7C, D. As expected, mRNAseq-derived TARs had similar distributions of Rtms as reference-annotated genes, as current macaque reference annotations mostly cover protein-coding genes, which mostly transcribe polyadenylated transcripts. In addition, the coding potential analysis also showed that these total RNAseq-derived transcripts were predominantly non-coding (Fig. 3B, D). In summary, these results show that through RNAseq analysis of total RNAs, we were able to uncover thousands of TARs which were enriched with non-polyadenylated non-coding RNAs. This class of transcripts was not well covered by the conventional mRNAseq analysis, the most commonly used technique for transcriptome analysis.

J Med Primatol **43** (2014) 317–328

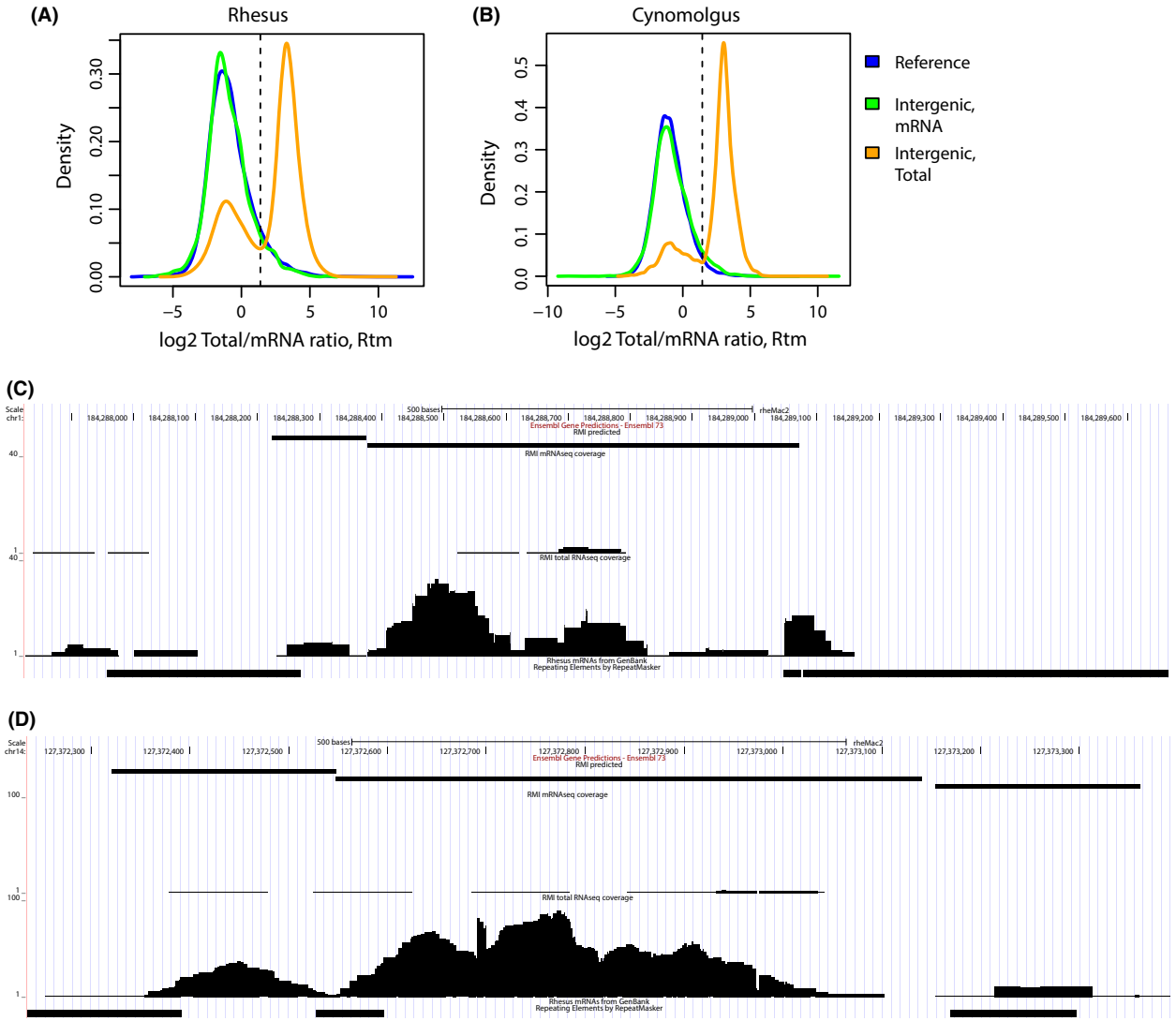


Fig. 7 Characterization of total RNAseq-derived intergenic transcripts. (**A**) The distribution of Rtm (log_2 scale) for rhesus genes or TARs from different annotation sources. Reference: reference-annotated genes (blue). Intergenic mRNA: Intergenic TARs derived from mRNAseq data (green). Intergenic Total: Intergenic TARs derived from total RNAseq data (orange). The dashed vertical line shows the local minimum between two peaks in the Rtm distribution of TARs derived from total RNAseq data. TARs with Rtms lying to the left of this line were removed from the final annotation. (**B**) Cynomolgus macaque, similar as in (**A**). (**C**) Browser view of an example of rhesus intergenic transcripts identified on chromosome 1 (genomic position: chr1:184,287,827–184,289,666) by total RNAseq. From top to bottom, the tracks shown are (i) total RNAseq-derived transcripts, (ii) read coverage by mRNAseq data, and (iii) read coverage by total RNAseq data, and) repeats. (**D**) Another example of rhesus intergenic transcripts identified on chromosome 14 (genomic position: chr14:127,372,237–127,373,392) by total RNAseq, similar as in (**C**).

A number of newly identified transcripts are related to the macaque host response to virus infection

To evaluate the biological relevance of newly identified transcripts, we investigated if their expression changed in the context of two separate virus infection studies. For a rhesus macaque study, we obtained an mRNAseq dataset derived from rectal samples obtained from rhesus macaques infected (by rectal inoculation) with simian immunodeficiency virus SIVmac (F. Barrenas, R.E. Palermo, B. Agricola, M.B. Agy, L. Aicher, V. Carter, L. Flanary, R.R. Green, R. McLain, Q. Li, W. Lu, R. Murnane, X. Peng, M.J. Thomas, J.M. Weiss, D.M. Anderson, M.G. Katze in preparation). We detected (one or more uniquely mapped reads in all 12 samples) the expressions of 2181 mRNAseq-derived TARs. Out of these detected TARs, 245 were differentially expressed (fold change >1.5 and raw *P*-value <0.05) between base line samples and samples from infected animals collected as early as day 3 and/or 6

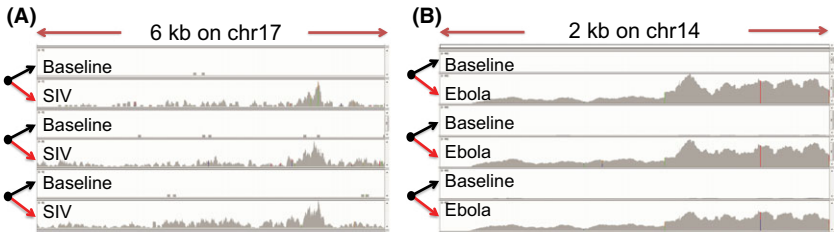


Fig. 8 Example of newly identified macaque intergenic transcripts differentially expressed during virus infection. (**A**) Read coverage (vertical peaks) of one rhesus macaque intergenic TAR which was upregulated in rectal samples 3 days after SIV infection. (**B**) Read coverage of one cynomolgus macaque intergenic TAR which was upregulated in PBMC samples 7 days after Ebola virus infection.

post-infection, and 135 of them had a false discovery rate <0.05. An example of a differentially expressed rhesus macaque TAR is shown in Fig. 8A. Similarly, from a cynomolgus macaque study, we obtained an mRNAseq dataset on peripheral blood mononuclear cell (PBMC) samples from animals infected with Ebola virus (F. Barrenas, R.E. Palermo, B. Agricola, M.B. Agy, L. Aicher, V. Carter, L. Flanary, R.R. Green, R. McLain, Q. Li, W. Lu, R. Murnane, X. Peng, M.J. Thomas, J.M. Weiss, D.M. Anderson, M.G. Katze, in preparation). Here we detected (one or more uniquely mapped reads in all 14 samples) the expression of 1175 TARs derived from the mRNAseq data. Out of the detected TARs, 824 were differentially expressed (fold change >1.5 and raw *P*-value <0.05) between base line samples and samples from infected animals collected at day 4 and/or 7 post-infection, and 714 of them had false discovery rate <0.05. An example of a differentially expressed cynomolgus TAR is shown in Fig. 8B. As illustrated by these two use cases, many of the newly identified macaque transcripts are likely to be biologically relevant as their expression changed significantly in response to virus infection.

Discussion

RNAseq is quickly becoming the dominant technology for transcript discovery [4, 5]. Here we used large-scale NHP RNAseq data [6] to systematically investigate transcriptional units present in the genomes of Indian-origin rhesus and Mauritian-origin cynomolgus macaques. We identified thousands of novel isoforms for reference-annotated genes and thousands of un-annotated intergenic transcripts for both macaque species. The finding of thousands of novel macaque transcripts is not surprising, as even for the well-studied human genome, thousands of novel isoforms and intergenic transcripts were uncovered recently using RNAseq [5]. However, our results highlight the huge gaps that exist in our understanding of macaque biology.

The majority of intergenic transcripts that we found were predicted to be non-coding RNAs, which is in agreement with similar studies [5]. This suggests that most macaque protein-coding genes have been reliably identified by sequence homology-based approaches, whereas non-coding RNAs are less likely to be identified by such approaches due to their weaker sequence conservation [14]. The biological relevance of ncRNAs is rapidly emerging in many research areas, including in HIV/AIDS research. For example, our recent RNAseq studies have found the expression of many ncRNAs is related to virus infection, including mice infected by severe acute respiratory syndrome coronavirus (SARS-CoV) [15] and human $CD4^+$ T cells infected by HIV-1 [16]. Zhang et al. [17] reported that the knockdown of the long ncRNA NEAT1 enhances HIV-1 production by increasing the export of Rev-dependent instability element-containing HIV-1 mRNAs from the nucleus to the cytoplasm. Here we also showed that many newly identified intergenic transcripts were differentially expressed during SIV infection of rhesus macaques or during Ebola virus infection of cynomolgus macaques. Not only do these results convincingly argue that many of novel transcripts identified in this study are biologically relevant but also that this collection of novel transcripts can serve as a resource for future macaque studies due to its more complete coverage.

By *de novo* assembly of un-mapped RNAseq reads, we also identified many macaque transcripts which are missing from existing macaque genome assembles. Our analyses both locate the specific genomic regions which need to be revised and provide the exact sequences for filling those gaps or correcting the errors. We found that the majority of these missing macaque transcripts only partially cover homologous human genes, based on their placement on the macaque–human genome alignments. This suggests that the current rhesus macaque genome assembly has already covered most macaque genes, but likely with relatively small 'holes' for many genes, mainly due to its draft quality. As currently most

J Med Primatol **43** (2014) 317–328

sequenced genomes are of draft quality, there is a great need for systematic and continuous improvements of these genome assemblies. Here we demonstrated that many of these 'missing' fragments could be readily identified by *de novo* assembly of un-mapped RNAseq reads. Because species-specific RNAseq data are becoming more readily available, new computational strategies are needed to harness this sequence information to improve reference genome assemblies in parallel to individual studies.

In summary, we used ultra-deep transcriptome sequencing data to systematically derive more complete genome annotations for the two macaque species most commonly used in research applications. The improved annotation added thousands of novel splicing isoforms to currently annotated genes and un-annotated intergenic transcripts. The improved annotation provides significantly better coverage of macaque non-coding RNAs, including both polyadenylated and non-polyadenylated transcripts. These results also demonstrate that the ultra-deep RNAseq data generated by the NHPRTR are extremely valuable for improving NHP genome annotation and potentially genome assembly.

Acknowledgments

Research reported in this publication was supported by the Office Of The Director, National Institutes of Health under Award numbers R24OD010445, P51OD010425 and R24OD011172 (to M.G.K.). This work was also funded by Public Health Service grant R01NS076465 (to C.E.M.). This work was funded by Public Health Service grants R24OD010445, P51OD010425 and R24OD011172 (to M.G.K.) and R01NS076465 (to C.E.M.).

Reference

1 Gibbs RA, Rogers J, Katze MG, Bumgarner R, Weinstock GM, Mardis ER, Remington KA, Strausberg RL, Venter JC, Wilson RK, Batzer MA, Bustamante CD, Eichler EE, Hahn MW, Hardison RC, Makova KD, Miller W, Milosavljevic A, Palermo RE, Siepel A, Sikela JM, Attaway T, Bell S, Bernard KE, Buhay CJ, Chandrabose MN, Dao M, Davis C, Delehaunty KD, Ding Y, Dinh HH, Dugan-Rocha S, Fulton LA, Gabisi RA, Garner TT, Godfrey J, Hawes AC, Hernandez J, Hines S, Holder M, Hume J, Jhangiani SN, Joshi V, Khan ZM, Kirkness EF, Cree A, Fowler RG, Lee S, Lewis LR, Li Z, Liu YS, Moore SM, Muzny D, Nazareth LV, Ngo DN, Okwuonu GO, Pai G, Parker D, Paul HA, Pfannkoch C, Pohl CS, Rogers YH, Ruiz SJ, Sabo A, Santibanez J, Schneider BW, Smith SM, Sodergren E, Svatek AF, Utterback TR, Vattathil S, Warren W, White CS, Chinwalla AT, Feng Y, Halpern AL, Hillier LW, Huang X, Minx P, Nelson JO, Pepin KH, Qin X, Sutton GG, Venter E, Walenz BP, Wallis JW, Worley KC, Yang SP, Jones SM, Marra MA, Rocchi M, Schein JE, Baertsch R, Clarke L, Csuros M, Glasscock J, Harris RA, Havlak P, Jackson AR, Jiang H, Liu Y, Messina DN, Shen Y, Song HX, Wylie T, Zhang L, Birney E, Han K, Konkel MK, Lee J, Smit AF, Ullmer B, Wang H, Xing J, Burhans R, Cheng Z, Karro JE, Ma J, Raney B, She X, Cox MJ, Demuth JP, Dumas LJ, Han SG, Hopkins J, Karimpour-Fard A, Kim YH, Pollack JR, Vinar T, Addo-Quaye C, Degenhardt J, Denby A, Hubisz MJ, Indap A, Kosiol C, Lahn BT, Lawson HA, Marklein A, Nielsen R, Vallender EJ, Clark AG, Ferguson B, Hernandez RD, Hirani K, Kehrer-Sawatzki H, Kolb J, Patil S, Pu LL, Ren Y, Smith DG, Wheeler DA, Schenck I, Ball EV, Chen R, Cooper DN, Giardine B, Hsu F, Kent WJ, Lesk A, Nelson DL, O'Brien WE, Prufer K, Stenson PD, Wallace JC, Ke H, Liu XM, Wang P, Xiang AP, Yang F, Barber GP, Haussler D, Karolchik D, Kern AD, Kuhn RM, Smith KE, Zwieg AS: Evolutionary and biomedical insights from the rhesus macaque genome. *Science* 2007; **316**:222–34.

2 Ebeling M, Kung E, See A, Broger C, Steiner G, Berrera M, Heckel T, Iniguez L, Albert T, Schmucki R, Biller H, Singer T, Certa U: Genome-based analysis of the nonhuman primate *Macaca fascicularis* as a model for drug safety assessment. *Genome Res* 2011; **21**:1746–56.

3 Yan G, Zhang G, Fang X, Zhang Y, Li C, Ling F, Cooper DN, Li Q, Li Y, van Gool AJ, Du H, Chen J, Chen R, Zhang P, Huang Z, Thompson JR, Meng Y, Bai Y, Wang J, Zhuo M, Wang T, Huang Y, Wei L, Li J, Wang Z, Hu H, Yang P, Le L, Stenson PD, Li B, Liu X, Ball EV, An N, Huang Q, Fan W, Zhang X, Wang W, Katze MG, Su B, Nielsen R, Yang H, Wang X: Genome sequencing and comparison of two nonhuman primate animal models, the cynomolgus and Chinese rhesus macaques. *Nat Biotechnol* 2011; **29**:1019–23.

4 Denoeud F, Aury JM, Da Silva C, Noel B, Rogier O, Delledonne M, Morgante M, Valle G, Wincker P, Scarpelli C, Jaillon O, Artiguenave F: Annotating genomes with massive-scale RNA sequencing. *Genome Biol* 2008; **9**:R175.

5 Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, Xue C, Marinov GK, Khatun J, Williams BA, Zaleski C, Rozowsky J, Roder M, Kokocinski F, Abdelhamid RF, Alioto T, Antoshechkin I, Baer MT, Bar NS, Batut P, Bell K, Bell I, Chakrabortty S, Chen X, Chrast J, Curado J, Derrien T, Drenkow J, Dumais E, Dumais J, Duttagupta R, Falconnet E, Fastuca M, Fejes-Toth K, Ferreira P, Foissac S, Fullwood MJ, Gao H, Gonzalez D, Gordon A, Gunawardena H, Howald C, Jha S, Johnson R, Kapranov P, King B, Kingswood C, Luo OJ, Park E, Persaud K, Preall JB, Ribeca P, Risk B, Robyr D, Sammeth M, Schaffer L, See LH, Shahab A, Skancke J, Suzuki AM, Takahashi H, Tilgner H, Trout D, Walters N, Wang H, Wrobel J, Yu Y, Ruan X, Hayashizaki Y, Harrow J, Gerstein M, Hubbard T, Reymond A, Antonarakis SE, Hannon G, Giddings MC, Ruan Y, Wold B, Carninci P, Guigo R, Gingeras TR: Landscape of transcription in human cells. *Nature* 2012; **489**:101–8.

6 Pipes L, Li S, Bozinoski M, Palermo R, Peng X, Blood P, Kelly S, Weiss JM, Thierry-Mieg J, Thierry-Mieg D, Zumbo P, Chen R, Schroth GP, Mason CE, Katze MG: The non-human primate reference transcriptome resource (NHPRTR) for comparative functional genomics. *Nucleic Acids Res* 2013; **41**:D906–14.

7 Langmead B, Trapnell C, Pop M, Salzberg SL: Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 2009; **10**:R25.

8 Wu TD, Nacu S: Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* 2010; **26**:873–81.

9 Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 2010; **28**:511–5.

10 Wang L, Park HJ, Dasari S, Wang S, Kocher JP, Li W: CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res* 2013; **41**:e74.

11 Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A: Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 2011; **29**:644–52.

12 Kent WJ: BLAT–the BLAST-like alignment tool. *Genome Res* 2002; **12**:656–64.

13 Wu TD, Watanabe CK: GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 2005; **21**:1859–75.

14 Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL: Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* 2011; **25**:1915–27.

15 Peng X, Gralinski L, Armour CD, Ferris MT, Thomas MJ, Proll S, Bradel-Tretheway BG, Korth MJ, Castle JC, Biery MC, Bouzek HK, Haynor DR, Frieman MB, Heise M, Raymond CK, Baric RS, Katze MG: Unique signatures of long noncoding RNA expression in response to virus infection and altered innate immune signaling. *MBio* 2010; **1**:e00206–10.

16 Chang ST, Sova P, Peng X, Weiss J, Law GL, Palermo RE, Katze MG: Next-generation sequencing reveals HIV-1-mediated suppression of T cell activation and RNA processing and regulation of noncoding RNA expression in a $CD4^+$ T cell line. *MBio* 2011; **2**: e00134–11.

17 Zhang Q, Chen CY, Yedavalli VS, Jeang KT: NEAT1 long noncoding RNA and paraspeckle bodies modulate HIV-1 posttranscriptional expression. *MBio* 2013; **4**:e00596–12.