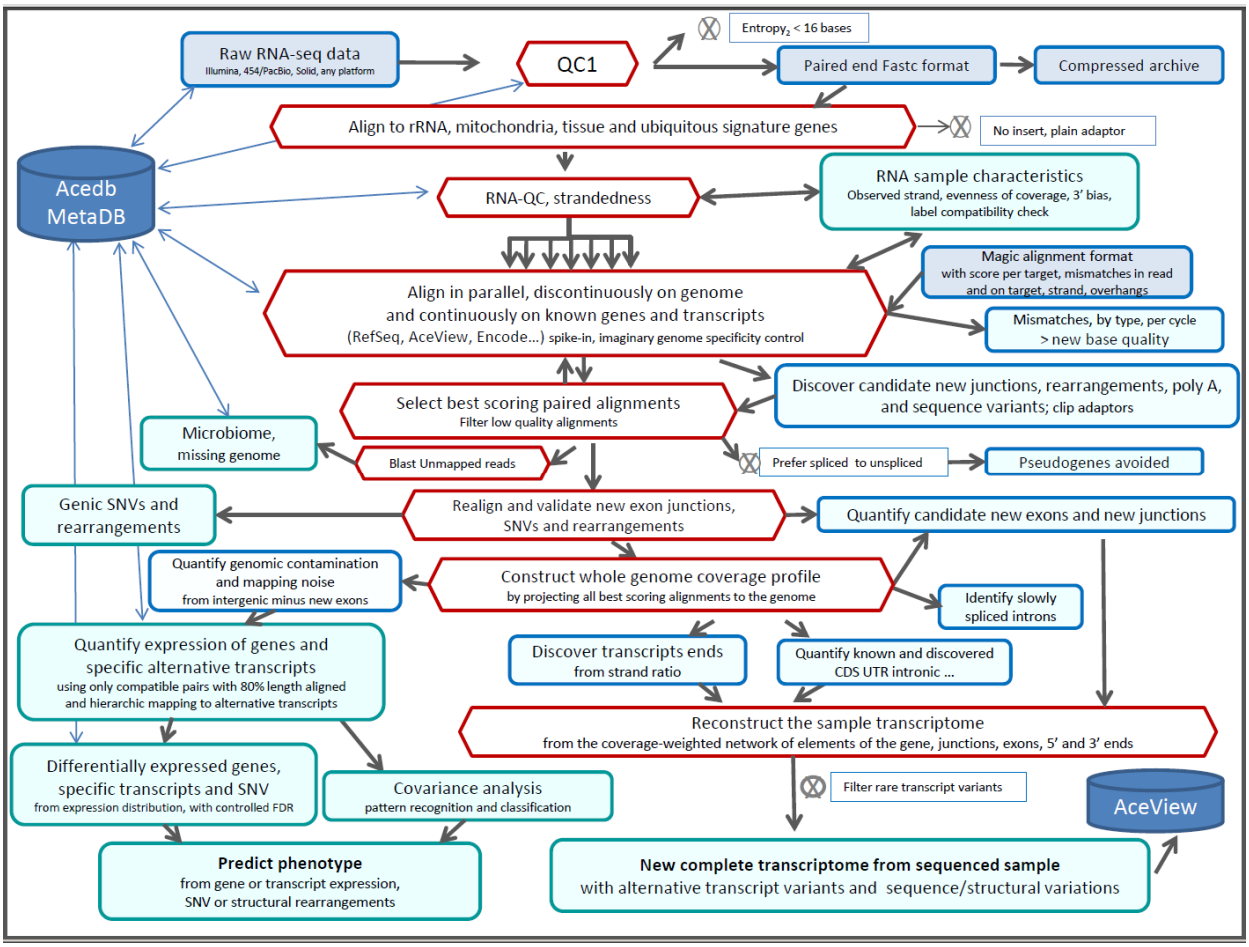
Supplementary Materials

for

Tissue-specific transcriptome sequencing analysis expands the non-human primate reference transcriptome resource (NHPRTR)

I. Supplementary Figures

Supplementary Figure 1. Overview of Magic pipeline for RNA-seq analysis.



The Magic code is available on the AceView NCBI website in the Downloads / Software tab / Magic, i.e. at ftp://ftp.ncbi.nlm.nih.gov/repository/acedb/Software/Magic. The package, written in C, is original, self-standing and open source, but is not modular. It includes its own aligner and software to perform the integrative data analysis presented here, and generates expression data tables as well as multiple quality control reports. Annotating new full-length genes and transcripts is implemented at NCBI but not

yet finalized for public access. Typically, one gigabase of RNA sequences are aligned on all targets using about 6 hours on a single CPU, 8 GB of RAM and 2 gigabytes of disk space. In practice there are no parameters to specify as Magic is auto-adaptive. Before use, simple experimental details have to be entered in the object oriented AceView/AceDB master database for each sequencing file (a 'run'), and relevant groups of runs manually created (for instance multiple runs for the same library, or groups of libraries per tissue or per species). Actions to be performed, such as expression index measures, transcript discovery, SNP detection, sample comparisons, DEG calls or covariance analysis, titration or serial profiling are manually specified for each group. The pipeline is then run with a simple command such as: "MAGIC ALIGN GENE_EXPRESSION INTRON_DISCOVERY SNP".

Supplementary Figure 2. Examples of the new display of gene expression across tissues and primates for each gene on the AceView website.

Expression profile of the top gene, SLC1A2, is conserved across all species and highest (top octile, dark green) in cerebellum, brain and liver, whereas the bottom gene, SLC5A1, shows an innovative strong usage in Human and Chimpanzee heart. The plot to the right shows the distribution of expression index for the gene in all tissues (green) compared to all measures across all genes and tissues (blue).

II. Supplementary Table

Supplementary Table 1. Summary of the data and quality control metrics **(MS excel file available at NAR Online)**. This table includes 7 sections and describes the 154 tissue libraries from non-human primates, made from total RNA, and ordered by species and tissue. 1) Data description (columns A to U). Columns B to K identify the samples, libraries and protocols, L to U give the statistics of the raw sequences. 2) Mapping results and unicity (columns V to AA). 3) Partition of all mapped reads per target, annotated gene or genome (columns AB to AK). 4) Properties of the libraries derived from mapping: average insert length, measured strand preservation, and number of mismatches per kilobase aligned

(columns AL to AN). 5) Mapping fate of read pairs (columns AO to BC), either compatible, non-compatible, or with a single read aligned (AP to AR) with further details on the types of pairs by location. 6) Factors and corrections for computing the gene index (columns BD to BJ): megabases aligned overall, in genes, in intergenic regions, lists of highly expressed genes, and insert length. 7) Number of RefSeq or AceView genes and transcripts significantly expressed (Columns BK to BM).

III. Supplementary Methods

Tissue samples and RNA-seq library preparation

All samples and RNAs were collected and prepared as described previously (1). The total RNA-Seq libraries were constructed using the Illumina TruSeq Stranded Total RNA Sample Prep Kit with Ribo-Zero Gold (catalog #RS-122-2201), which targets RNA for cDNA synthesis. This kit uses Ribo-Zero to remove rRNA prior to library preparation. Whole blood samples were additionally globin-depleted (prepared using the TruSeq Stranded Total RNA with Ribo-Zero Globin kit from Illumina. The libraries were quality-controlled for appropriate mass and insert size (also for evidence of DNA contamination) and quantitated using the BioAnalyzer 2100 system and qPCR (Kapa Biosystems, Woburn, MA). Sequencing was performed with the Illumina HiSeq system using HiSeq version 3 sequencing reagents. The libraries were clonally amplified on a cluster generation station using Illumina HiSeq version 3 cluster generation reagents to achieve a target density of approximately 700,000/mm2 in a single channel of a flow cell. Image analysis, base calling, and error estimation were performed using Illumina Analysis Pipeline (version 2.8). For rhesus macaque Indian-origin, some additional data were generated using the Illumina MiSeq system to supplement the number of reads up to a minimum level for all tissue libraries. HiSeq RTA analysis (version 1.13) and MiSeq Reporter (version 2.4.60) achieve comparable base-calling and quality thresholds because both machines use similar SBS chemistry.

RNA-seq data analysis using the MAGIC pipeline

1) <u>Overview</u>

In a complex RNA-seq project such as NHPRTR, with different species and multiple tissues, there is a critical need for an integrative tool allowing high level analysis, which should implement stringent quality control leading to rigorous extraction of signal from noise, and allow manipulating transparently very large datasets. The 'Magic' RNA-seq analysis pipeline tries to fulfill this need. It was originally developed as part of the large SEQC collaboration, and the version of 2012 was recently described in two articles (2,3). As the program continuously improves, we provide here a detailed description of the 2014 version employed for this study. The software is available online: ftp://ftp.ncbi.nlm.nih.gov/repository/acedb/Software/Magic.

The pipeline, detailed in Supplementary Figure 1, can perform RNA-seq to genome alignments for all sequencing platforms; it uses read-pair information and has no limits on the read length (20 bp to >100 kb). It meshes in a single integrative environment the quality control, alignment and coverage analysis, versatile grouping of samples by phenotype or function, quantification of expression at gene, transcript and exon junction levels, discovery and validation of new introns and exons. It also knows how to select

significant differential features, call SNV (substitutions and short indels) and rearrangements, and perform predictions using either stratified signature selection or covariance analyses.

Although the pipeline will run automatically and the data is very deep, an analysis of the 'Data_summary_and_QC' table (Supplementary Table 1) remains critical for proper interpretation of the results, because it can reveal the important experiment-specific biases such as batch effects, suboptimal quality of the library preparations, or problems occurring during the sequencing. Ignoring those may well lead to over-interpretation of the deep data and irreproducible conclusions.

Here we describe the application of the Magic pipeline to the NHPRTR project, which so far produced 212 libraries from isolated tissues or mixed tissues on 16 primates of medical interest (including the comparable Human BodyMap project: 19 libraries generated by Illumina in 2010). The main focus of this project is to better define the biology of primates and their relevance as models for human. We reasoned that one way to gain an integrated comparative view and to avoid intractable nomenclature problems would be to unify the gene view by centering the sequence analysis on the Human genome, more complete and better annotated than any of the other primates. This ambitious approach could be envisioned because AceView and its successor Magic can efficiently find complete discontinuous alignments of read pairs genome-wide even in the presence of a high density of sequence variations. In total 6.37 terabases of primates RNA sequence were generated and analyzed.

2) Method for Magic alignment of the RNA-seq data

Read pairs are sorted alphabetically by their sequence, split in blocks of up to 5 million reads or 300 Mb, and analyzed as single objects, using the fastc modification of the fasta format, which represents each pair on a single line as ATTGNCATGNC><CCTATTTGAT. The mapping strategy is to select from the reads a dense set of 16-base seeds with high information content (2-letter entropy greater than 8 bases, one seed every fifth base), to pre-scan the gene and genome targets using compact Bloom filters to select the seeds present 1 to 9 times (hence there is no requirement to mask the genome, which is treated as a single object), hash the selected seeds, scan the target again and extend the exact seed matches using a fast automaton developed for the AceView cDNA-to-genome aligner. Alignments may include substitutions, short insertions, or deletions, up to 3 bases long. Starting from a given position, the automaton zips all exact matches, and when encountering a mismatch, it incorporates substitutions or small indels based on the number of exact matches immediately downstream. This number is specified in a tiny cost table which depends on the sequencing platform, for instance favoring indels in Roche 454 (especially frequent in homo-polymers), or for SOLiD color code, double transitions followed by on the fly automatic error correction. Alignments stop if encountering 5 mismatches over 10 bases.

A special property of Magic is that it maps the read pairs in parallel i) discontinuously on the unmasked genome, and ii) continuously on a variety of gene annotations (RefSeq, AceView, Encode then recursively to the newly reconstructed transcriptome), slightly favoring mapping to the abundant mitochondrial and ribosomal RNAs. Once fragments have been aligned to all targets, a universal scoring system allows Magic to select the top scoring alignments and the best targets for each read pair. For RNA-Seq, each alignment is scored by gaining one point per base aligned and losing eight points per mismatch (base substitution, insertion or deletion); the length of the aligned region is then recursively

clipped to optimize this score. As a result, every aligned read in Magic is deeply rooted in its target, with sticky ends of 8 exact matches on both sides of the alignment. This high error cost may seem surprising, but is essential for accurate transcriptome reconstruction, allowing reliable recognition of substitutions or indels and precise cooperative discovery of exon-intron and rearrangements boundaries.

The target sequences for this project are available at: ftp://ftp.ncbi.nlm.nih.gov/repository/acedb/NHPRTR_Reference_Targets . In brief we used:

- The reference genome GRch37 (without patches)
- AceView genes (264,387 transcripts, from 55,836 genes, including 378,219 exon junctions from AceView 2010 v4, which is a subset of the 2010 version where more single exon genes were moved to the 'cloud' class and a few pseudogenes manually evaluated and filtered): AceView 2010 genes summarize the 9 million human cDNA sequences in GenBank, and RefSeq/Gene NM/NR transcript models from 2010 (37.1). The cDNAs define about twice as many genes and close to twice as many exon junctions than RefSeq (370,000 rather than 198,000 exon junctions in 2010). AceView 2010 annotates 205,691 spliced (multi-exon) transcript variants. Entrez Gene names are used when available; otherwise genes are named using an AceView identifier, stable across releases.
- The image of the RefSeq v104 on the genome (40,894 transcripts from 24,536 genes, January 2013),
- The mitochondrial genome (NC_012920.1),
- Ribosomal RNA 5S and 45S precursor (NR_046235 and NR_023363),
- Small non coding RNA genes (including 8,611 genes from Ensembl 37.70: 1,923 snRNA, 1,529 snoRNA, 3,110 miRNA and 2,049 misc-RNA, complemented by 625 tRNAs from Lowe, UCSC tRNA track, February 2012). Many of those sequences are already part of transcript models in AceView.
- Finally an imaginary genome, used as a mapping specificity control. The imaginary whole genome is constructed by complementing the bases (exchange A:T and G:C), but not reversing the order: it has exactly the same statistics as the genome but is a completely alien decoy. Hits to this target are false positives and their characteristics are used to tune the alignment quality filters and choose the minimum length aligned and maximum mismatch rate for the specific experiment.

3) Quantification of RNA-seq expression data

The number of reads and bases aligned in each gene are then collected and used to compute the gene expression index. To ensure high specificity of the mapping and reliable indexes, very stringent criteria are used: only fragments with both reads aligned contribute to the quantitative analyses; they represent 90.1% of all mapped fragments in NHPRTR. They further have to i) align at a unique site, in a unique gene on the proper strand (note that reads mapping to several transcripts of the same gene count as unique), ii) the alignments have to be quasi complete for both reads in the pair (at least 80% bases aligned per read, ending with 8 exact bases on each side of each read, i.e. greater than 160 bases aligned per pair); iii) the topology of the alignment has to be 'compatible', reads facing each other, and not too distant (short gene extensions are allowed). This set of criteria applies in the computation of the expression value, in the coverage plots (visible at UCSC), or in the determination of homozygous SNPs (minimum coverage 10, minimum allele frequency 95%, also posted at UCSC). The library preparation protocol used (total RNA UDG) has the advantage of being stranded, but enriches in intronic reads, and altogether after applying all constraints, only 41.1% of the aligned fragments fulfill our criteria. One

should remain aware that genes whose sequence evolved to become too distant from Human cannot be measured well.

The Magic normalized RNA-Seq index of expression is directly comparable to a normalized microarray logarithmic signal. The logarithm is introduced because gene expression is mostly log-normal. The index is actually similar to a $\log_2$ FPKM in that it is normalized to be independent of the size of the experiment and of the length of the gene. However, it dampens the sampling fluctuations and defines several corrections to increase robustness and comparability of the values, which are quantified in each library as part of the quality control, and applied to the F, K and M terms, defining a significant FPKM (sFPKM), with attempts to compensate for undesirable batch effects: i) the insert length of the library, which introduces a steric constraint on the coverage of short genes; ii) the eventual 3' bias, which limits the coverage of long genes; iii) the level of genomic contamination, sequencing and mapping noise, estimated from the density of intergenic reads outside new exons discovered in the experiment; this non-specific noise needs to be subtracted from each gene; iv) the presence of extremely highly expressed genes, removed from the effective library size.

Correcting for these measurable effects, which are library dependent and occasionally strong, stabilizes the expression values (index/sFPKM) across samples. All these corrections affect the normalization factor z, while maintaining the read counts n on the original scale, to correctly control the sampling fluctuations (see below).

The Magic index, resulting from optimization using the SEQC data, is computed as follows:

$$\text{Index} = \log_2 \left(n + \sqrt{(5+b)^2 + n^2}\right) - 1 + \log_2(z), \text{ with } z = 10^{12} r/N'l'$$

sFPKM= 2^{Index} / 1000 i.e. Index = $\log_2$(1000 sFPKM) (an sFPKM of 1 corresponds to an index ~10).

- n is the number of reads aligned in the gene. A read partially aligned counts as a fraction.
- b is the intergenic background noise density multiplied by the gene length.
- r is the average aligned length of the reads, in bases.
- N' /10^9 is the number of gigabases aligned to the genes with a RefSeq GeneID, after excluding the ribosomal and mitochondrial genes and the genes gathering more than 2% of the total number of bases aligned.
- l' is the length in bases of the gene effectively sequenceable in the particular experiment. It is influenced by
    - the 3' bias, measured over the cumulated profile of very long genes as the distance (in bp) between the region of maximal 3' expression (averaged over 500 bp) and the 5'-most region where expression dropped to 20% of the maximum. This length is used in each library as an upper bound for the sequenceable length of the gene l_{max}.
    - The insert length of the library: the insert has to be within the boundaries of the transcript, so the average length λ of the insert is subtracted from the length l of each

gene to compensate for the steric constraints. Indeed $l' = (l - \lambda)$ is the number of possible positions of the insert in the gene, and this correction matters for short genes, under-represented in libraries with long insert size.

- Finally, l' is restricted to the interval $[\lambda, l_{max}]$, since short genes cannot accommodate inserts of size longer than the gene and long genes where the library was polyA selected are limited by the 3' bias.
- Note that another effect occurs when inserts are much longer than their sequenced length and the gene is barely longer than the insert: then the central portion of the gene will not be sequenced (example: gene of 600 bp, library insert 500 bp, sequencing length paired end 50+50 bp, the 300 central bases will be missing from the sequence). We do not automatically correct for this yet. But if comparing libraries with different insert sizes, this may create the illusion of differential alternative splicing of the center of the gene.

Finally, sampling fluctuations affect the counts in low expressed genes, and all values (index or sFPKM) based on less than 10 reads are flagged NE/NA. The advantage of the normalized formula is that it uses an argsh function, equivalent to $\log_2$ for high counts, but creating a plateau for small counts, so that it squeezes the values associated to low expression, making a twofold change in the low range correspond to much more than a doubling of the read counts. For zero counts (0 reads in the gene), the value is flagged as Not-Expressed (NE/value), for 2 to 8 reads above the average intergenic coverage, it is flagged as Not-Accurate (NA/value), indicating high sampling error and a tradeoff between low precision due to high sampling error and low accuracy due to the proximity of the zero counts plateau. Yet the value behind the flag depends on the length of the gene and on the depth of the experiment: for an NE/value, the deeper the experiment, the smaller the value, indicating higher confidence that the gene is not expressed.

To use this information in the normalized tables, we recommend to memorize the NA/NE flags, then to perform the differential analysis using the values behind the flags, then to filter away all differential genes where the high value is flagged. Notice that the flagged values should not be rejected, because it is desirable to detect the eventual ON/OFF genes for which the OFF value is very low and probably flagged. This system may seem complex, but it allows more sensitivity and more specificity than a global threshold.

4) Quality control measures

The detailed quality control measures are available in the Supplementary Table 1. Briefly, in the 154 libraries, made from total RNA from non-human primates' tissues in the NHPRTR project, the library properties measured through the mapping results can be summarized as follows:

- The depth of each library is on average 140.49 million reads (SD 44.65); 13.37 gigabases of RNA-seq (SD 4.10) were generated. Baboon thymus is deeper (454 million reads, 45.8 Gb) (Table S1, columns N P).
- A low fragment multiplicity (1.18 fragments per unique sequence, SD 0.19, col L-N) indicates limited PCR amplification and proper representation of the transcriptome. A multiplicity of 1.9 to 2.3 is observed in colon

from mouse lemur, and 3 whole blood libraries, (Japanese macaque, ring-tailed lemur and squirrel monkey), with shorter read length.

- The average GC content of the libraries is 45.13% (SD 2.41%, col U), quite homogeneous, except for isolated libraries where the percentage of GC is above 50% (whole blood from 3 species, Japanese macaque, sooty mangabey or ring tailed lemur, and spleen and colon from mouse lemur) and all chimpanzee libraries, apart from the four from brain locations, which tend to be richer in GC (46 to 52%).
- Good quality mapping was obtained for 88.75% of the reads on average (SD 11.64%, depending on taxonomy distance to Human, Figure 2a). On average, each library has 123.9 million reads (SD 41.9 million, Col V), 65.18 million fragments (SD 21.7 million, Col AD) and 9.93 gigabases (SD 3.66 Gb, Col BD) well aligned, however, no library has less than 4 gigabases aligned. The largest is thymus from Baboon, with 35.8 gigabases aligned and the smallest are from lemurs. The 14 Rhesus macaque Indian libraries are in the low range, between 4.9 and 7.4 Gb.
- On average 86.66% of all fragments map uniquely (SD 11.8, Col X). 2.09% (SD 1.80%, col Y) map to 2 to 9 sites, with little impact of the species but, as expected for shorter sequencing length (51+51), strong enrichment of multiple mapping in 4 blood libraries (Japanese macaque 18.4%, sooty mangabey 12.3%, squirrel monkey 6.4% and owl monkey 5.2%)
- The ribosomal RNA was usually efficiently removed by the ribozero treatment (average 3.5% of all aligned reads map to rRNA, SD 6.0%, Col AB) although large variations are observed: for instance, all 14 Rhesus macaque, Indian, were remarkably depleted (average 0.13%) while 7 libraries had 15 to 42% rRNA (mouse lemur colon (42%) and spleen (36%), ring tailed lemur blood (22%); marmoset colon (21%), baboon colon (28%) and bone marrow (16%) and Chimpanzee lymph node (26%)). The mitochondrial sequence appears to have evolved fast, as there was not much mitochondrial RNA (0.82%, SD 1.29%, Col AC) in any tissue, with the exception of heart (10 to 2% depending on distance to human). While RefSeq models absorb on average 41.86% of the reads (SD 10.84%, with an increase to 54.6% in lemurs, Col AD), mapping on AceView transcripts allowed to allocate 49.16% of the reads (SD 10.06%, Col AE). Another 0.79% of the reads (SD 1.1%, Col AG) aligned to small non coding RNAs not previously annotated in AceView or RefSeq, remarkably much higher in all blood libraries than in any other tissue. But by far the best target was the genome (78.53% of the reads, SD 6.94%, down to 65.6% in lemurs, Col AF), since many of the reads map inside introns of pre-messengers.
- The average insert length of the libraries is homogeneous, a feature typical of the TruSeq protocol (147 bp SD 15 bp), except for the blood libraries, which have slightly longer inserts (average length 170 to 200 bp)(Col AL).
- The UDG protocol was used, and indeed the libraries are well stranded, although not perfectly: 9.29% of the reads escaped the selection and map on the forward strand of the transcripts (SD 3.24%, Col AM).
- Read pairs are in majority compatible (80.93%, SD 8.96%, Col AP), both reads facing each other in close proximity in a gene (for 58% of the compatible pairs) or in the genome (for the remaining 42%, mainly in introns), with a strong dependency on taxonomy. This comes in addition to a systematic decrease in blood libraries of the frequency of compatible pairs (to around 65% in all species) and a correlated increase in fragments where a single read is mapped. There are however a large number of incompatible pairs (9.00%, SD 2.27%, Col AQ), varying from around 5% in chimpanzee to ~9% in old world monkeys to ~11% in new world monkeys. A majority of those appear to reflect genomic rearrangements that occur locally (4.14% have incompatible topology, with reads not facing each other locally (column AY), and 1.59% link two genes (Col AZ)), or more broadly in the genome (3.19% link a gene to very distant genome and 0.05% link two distant genomic sites, Col BA). There are 10.07% (SD 7.75%) of the pairs where a single read could be aligned at the stringency required (Col BB, BC).
- As a result of the stringent mapping criteria applied (demanding in particular 160 bp aligned in compatible fragment) and the protocol used, on average 5.12 Gb of RNA per library (SD 1.89 Gb, Col BD-BE) mapped

inside genes. But in almost all libraries, some highly expressed genes absorbed a sizeable fraction of the reads (on average 7.5%, SD 7.6, Columns BH, BI include gene lists). The most frequent are the incompletely polyadenylated cytoplasmic RNAs RN7SL1, 2 and 3 belonging to the signal recognition particle that mediates co-translational insertion of secretory proteins into the lumen of the endoplasmic reticulum [supplied by OMIM], and RN7SK, a small nuclear RNA belonging to the 7SK nucleoprotein complex. These 7S RNA genes are preeminent in all NHP tissues, MYH7 and TTN in heart, TTN NEB and MYH2 in muscle, ALB, APOB, FGG/B/A in liver, GH1 and PRL in pituitary, and MALAT1 in many samples, especially in blood, kidney and pituitary. Overall, these few genes contribute on average 24.4% of the fragments in skeletal muscle, 13.0% in liver, 10.3% in blood, 9.7% in heart, 5.9% in bone marrow and pituitary gland.

5) Notes on the expression tables

When the same biological sample was sequenced in multiple runs (as is the case for 33 NHP tissues and 11 Human tissues in this project), reads mapped to the same gene were cumulated before computing the index. In groups per tissue or per sample, consisting of different individual samples, the index of the group is computed as the average index of all participants. Thus in additive (single sample) groups, we compute the log of the average of the number of reads, while in diverse tissue or species groups, we compute the average of the log.

For each annotation, 3 tables representing respectively the raw read counts, the Magic index and the significant FPKM sFPKM = 2^{Index} / 1000 are provided on the NHPRTR site. All expression tables are sorted by descending maximal index of expression across the 168 tissue samples from the 15 primates. The first columns contain metadata connecting AceView genes to RefSeq IDs, Entrez GeneIDs, and genomic coordinates based on the human reference build 37.

Data representation is provided gene by gene on the AceView human website (Supplementary Figure 2), which also provides a link to the coverage plots at UCSC, per tissue and per species and the support of each exon-junction in the primate data (mouse over the gene diagram).

References

1. Pipes, L., Li, S., Bozinoski, M., Palermo, R., Peng, X., Blood, P., Kelly, S., Weiss, J.M., Thierry-Mieg, J., Thierry-Mieg, D. *et al.* (2013) The non-human primate reference transcriptome resource (NHPRTR) for comparative functional genomics. *Nucleic Acids Res*, **41**, D906-914.
2. Wang, C., Gong, B., Bushel, P.R., Thierry-Mieg, J., Thierry-Mieg, D., Xu, J., Fang, H., Hong, H., Shen, J., Su, Z. *et al.* (2014) The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. *Nat Biotechnol*, **32**, 926-932.
3. SEQC/MAQC-III Consortium. (2014) A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat Biotechnol*, **32**, 903-914.