A database of orthologous exons in primates for comparative analysis of RNA-seq data

Ran Blekhman

Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853, USA

Abstract

RNA-seq technology facilitates the study of gene expression at the level of individual exons and transcripts. Moreover, RNA-seq enables unbiased comparative analysis of expression levels across species. Such analyses typically start by mapping sequenced reads to the appropriate reference genome before comparing expression levels across species. However, this comparison requires prior knowledge of orthology at the exon level. With this in mind, I constructed a database of orthologous exons across three primate species (human, chimpanzee, and rhesus macaque). The database facilitates cross-species comparative analysis of exon- and transcript-level regulation. A web application allowing for an easy database query: http://giladlab.uchicago.edu/orthoExon/

Introduction

Recently developed massively parallel sequencing technologies allow for investigation of gene expression profiles at unprecedented depth (Fu et al, 2009; Marioni et al, 2008; Mortazavi et al, 2008). Termed RNA-seq, this approach allows identification of exon-, allele- and isoform-specific expression, and provides many technical advantages over existing microarray technologies, such as higher accuracy of expression measurement and lower levels of background noise (Wang et al, 2009).

Like expression microarrays, RNA-seq can be applied to study variation in gene expression levels across species and individuals. However, a clear advantage of RNA-seq over traditional microarray technologies is the ability to compare expression at the levels of individual exons and transcripts (Gilad et al, 2009; Wang et al, 2009). Specifically, RNA-seq can be used to compare exon-level expression across different species (Blekhman et al, 2010). In practice, this can be done by sequencing mRNA samples from multiple species, mapping the short reads against the relevant reference, and excluding reads that do not map to a unique genomic location. Then, the numbers of reads that fall within exons can be compared across species to identify exon-level expression differences. However, this requires prior knowledge of exon-level orthology between species, and currently, no such database exists for primates. In addition, in many available annotations (such as Ensembl) a single base can be annotated as part of two or more different exons, which may complicate any comparison of exon-specific expression levels.

Here, I present an annotation of orthologous exons in three primate species: human, chimpanzee, and rhesus macaque. This database allows mapping of RNA-seq reads to a unique, single exon in each of the three primate species, and also provides information on orthology between exons in these species. Moreover, this dataset can be used to identify differences in alternative splicing between species by examining reads that span orthologous exon junctions.

Database description

To identify orthologous exons in human, chimpanzee, and rhesus macaque, I used a three-step strategy (Figure 1):

(1) For each annotated human exon, identify putative orthologous exons in chimpanzee and rhesus macaque;

(2) Exclude exons located in regions with repetitive sequence content in any of the species, to avoid ambiguity in RNA-seq read mapping;

(3) Merge exons from the same gene whose genomic locations overlap to create a final set of orthologous meta-exons.

A full description of the methodology is available in the Methods section. Briefly, as a starting set, I used all known human Ensembl exons. I then used Blat (Kent, 2002) to identify likely orthologous exons in the chimpanzee and rhesus macaque genomes. I included only exons with a high similarity between species, and did not allow for long gaps in the aligned exon sequences. Next, I excluded exons that might be positioned within repetitive regions in any of the three genomes, as such regions might be particularly susceptible to mapping biases, thus leading to detection of spurious differences in expression levels across species. To do so, I mapped the exons of each species against that species' genome using Blat, and excluded from further analysis exons whose sequence is highly similar to at least one additional region in the genome (see supplementary methods). I then excluded from the analysis any exons for which there are no good matches in both chimpanzee and rhesus macaque, resulting in a set of high-quality orthologous exon trios.

Finally, I merged regions of overlapping exons, to allow for a unique mapping of reads to a single, orthologous exon in each species. To do so, I identified all cases of overlapping exons (Ensembl annotations include a large number of overlapping exons), excluded exons that were overlapping in one or two, but not in all three

species, and combined the remaining set of overlapping exons as appropriate (Figure 1).

The full analysis outlined above was repeated twice, generating two versions of the database: (1) hg18-panTro2-rheMac2 and (2) hg19-panTro3-rheMac2. The final dataset defined 150,107 meta-exons (in 20,689 Ensembl genes) in version 1, and 187,889 meta-exons (in 30,030 Ensembl genes) in version 2.

I note that it is difficult to assess the quality of the definitions, since there is no comparable information on exon-level orthology. Nevertheless, it is possible to assess the quality of the gene-level orthology using available information. To do so, I compared the orthologous exon genomic positions in version 1 of the database with a set of orthologous gene trios in human, chimpanzee, and rhesus macaque, identified recently by (Kosiol et al, 2008). I found that 97.9%, 97.8%, and 97.8% of the genomic coordinates that I defined overlapped the genomic positions from (Kosiol et al, 2008) in the genomes of human, chimpanzee, and rhesus macaque, respectively (the locations of 97.7% of the genes overlapped in all three species), suggesting our method performs well.

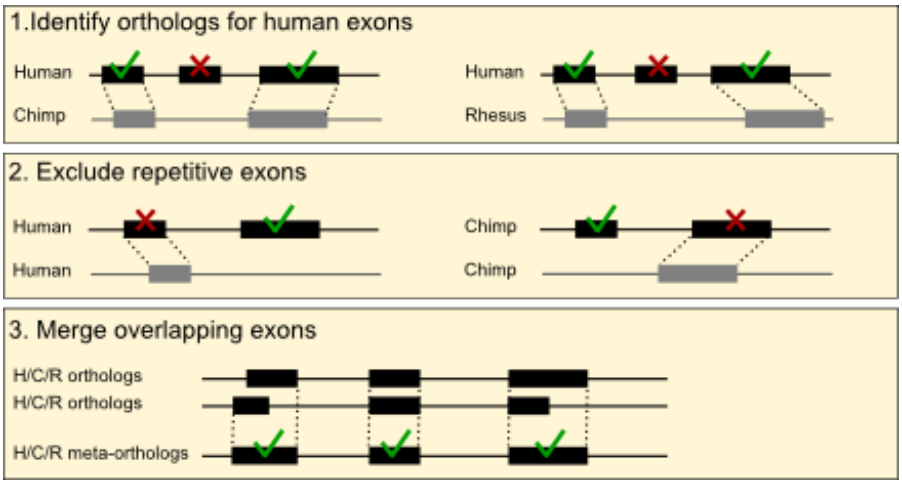


Fig. 1. An illustration of the methodology used to construct the database.

Database web interface

I have constructed a web application for database queries (http://giladlab.uchicago.edu/orthoExon/). The web application was designed using PHP, with a dedicated MySQL database serving as the back-end. The user interface was developed in AJAX, which allows for asynchronous data retrieval from the MySQL server without requiring a page reload.

For users interested in orthologous exon information for a specific gene, the interface is simple and intuitive, with a single text box where the user can input a gene name. AJAX was used to incorporate an auto-complete feature, which displays possible matching gene names from the database as the user is typing. After choosing a gene name, the database information available for that gene is displayed (Figure S1), including the genomic positions of all the orthologous exons in the three species. There are options to download the exon coordinates as a bed file, download a fasta file of the sequences, download the multi-species alignment of orthologous exons, and to view all exon regions in the UCSC genome browser with one click (Figure S2).

For users interested in the full dataset, the entire contents of both versions of the database can be downloaded from the website, and can be used in various genome-wide applications.

Discussion

I constructed a database of orthologous exons in human, chimpanzee, and rhesus macaque, and designed a web interface for easy data access and query. The database can be used in comparative analysis of RNA-seq data from these primate species to help identify differences in expression at the exon-level. Moreover, for each gene, a measure of its expression can be obtained by summing the number of reads mapped to its orthologous exons. This allows comparisons of expression at the gene-level to be performed. Another possible use for this database is in comparative analysis of alternative splicing, namely, identifying differences in splice variants across species.

This can be achieved by finding RNA-seq reads that map to orthologous exon-exon junctions.

A possible limitation of this database is the inclusion of orthology information from only three primate species, because our methodology requires a complete genome sequence for finding orthologous exons. However, as genome sequences from more species are completed, I will add information from additional primate species (such as orangutan and marmoset) to the database, thus extending its utility.

Methods

The approach used here is similar to the one used in (Blekhman et al, 2010), with several updates and additions. I chose to repeat here parts of the text from (Blekhman et al, 2010) to have a complete and accurate description of the pipeline used to construct the current version of the database.

Identifying human orthologs in chimpanzee and rhesus macaque. We first downloaded genomic coordinates for all human exons in the Ensembl database (http://www.ensembl.org, release 50 and release 64) (Hubbard et al, 2009). The following analyses was performed twice, with two sets of genome builds and starting exon data, generating two versions of the database:

(v1): hg18, panTro2, and rheMac2, using Ensembl version 50

(v2): hg19, panTro3, and rheMac2, using Ensembl version 64

I used Galaxy (http://main.g2.bx.psu.edu/) to extract the corresponding DNA sequences from the human genome, obtained sequence data for 520,023 exons in 36,397 Ensembl annotated genes in hg18, and 1,179,300 exons in 54,305 genes in hg19.

To find the orthologous sequences for the human exons in the non-human primates, I downloaded the full genome sequences of chimpanzee (*Pan troglodytes*, panTro2 and panTro3 drafts) and rhesus macaque (*Macaca mulatta*, rheMac2 draft) from the UCSC Genome Browser database (www.genome.ucsc.edu). Subsequently, for each human exon, I used Blat (Kent, 2002) to find the putative corresponding

positions in the chimpanzee genome (hg19-panTro3 and hg18-panTro2) and rhesus macaque genome (hg19-rheMac2 and hg18-rheMac2). Exons for which the best hit had less than 96% or 92% identity in chimpanzee or rhesus macaque respectively, were excluded (percent identity was defined as the proportion of bases in the query human exons for which a perfect match was found in the non-human primate target exon). Further, alignments with indels longer than 25bp were excluded. This resulted in the identification of 222,287 orthologous exons (in 28,299 genes) in hg18-panTro2, 193,632 orthologous exons (in 24,598 genes) in hg18-rheMac2, 944,458 orthologous exons (in 44,948 genes) in hg19-panTro3, and 827,950 orthologous exons (in 34,922 genes) in hg19-rheMac2,

Percent identify cutoff. In order to define a percent identity cutoff I mapped RNA-seq data to candidate orthologous exons, and identified a cutoff that maximized the number of orthologous exons while minimizing any possible biases that might result from using a human-specific set of exons as our input. The methodology for selecting a percent identity cutoff is identical to the one used in (Blekhman et al, 2010), therefore I will not repeat it in full here. Briefly, I used a range of possible cutoffs, and for each calculated the proportion of orthologous exons that were retained, and estimated the bias towards higher gene expression levels in humans. I selected cutoffs for which this bias is minimized and close to our a priori expectation that about 50% of genes will be more highly expressed in humans relative to chimpanzee or rhesus macaque.

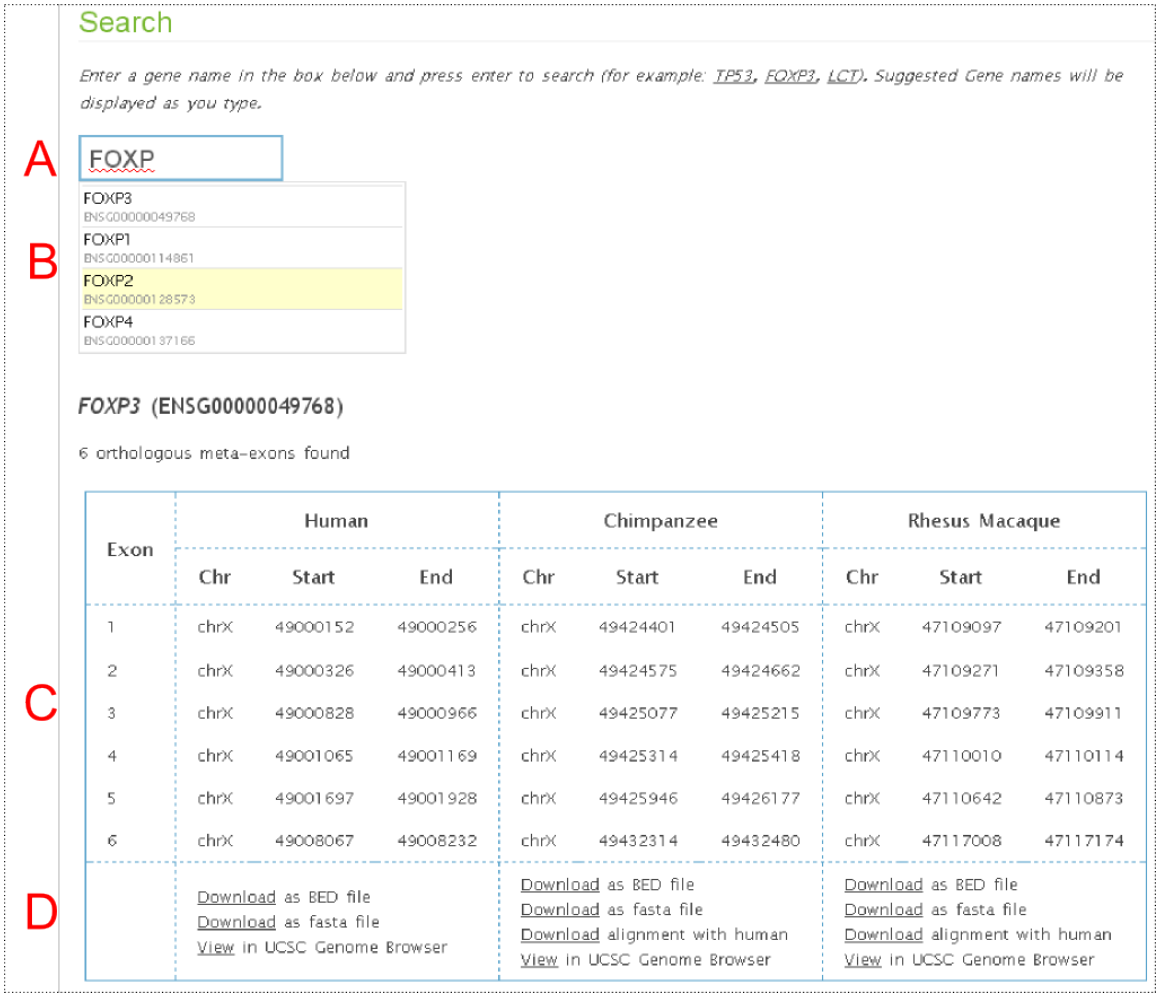
Excluding repetitive regions and joining overlapping exons. I excluded exons positioned within repetitive regions, in any of the three genomes, as such regions might be susceptible to mapping biases, thus potentially leading to the identification of spurious differences in expression levels between species. To do so, I used Blat to map all of a species' exons against its own genome, and excluded any exon for which (***i***) the best hit was not the original exon position, or (***ii***) the second-best hit had higher than 90% identity. Of the remaining set of orthologous exons, 163,487 and 291,868 were shared across the three species in versions 1 and 2, respectively. Finally, I identified cases where two (or more) orthologous exons overlapped one another in all three species, and combined such exons (I excluded

exons that overlapped in only a subset of the three species). This resulted in the definition of 150,107 meta-exons (in 20,689 Ensembl genes) in v1, and and 187,889 meta-exons (in 30,030 Ensembl genes) in v2.

Acknowledgements

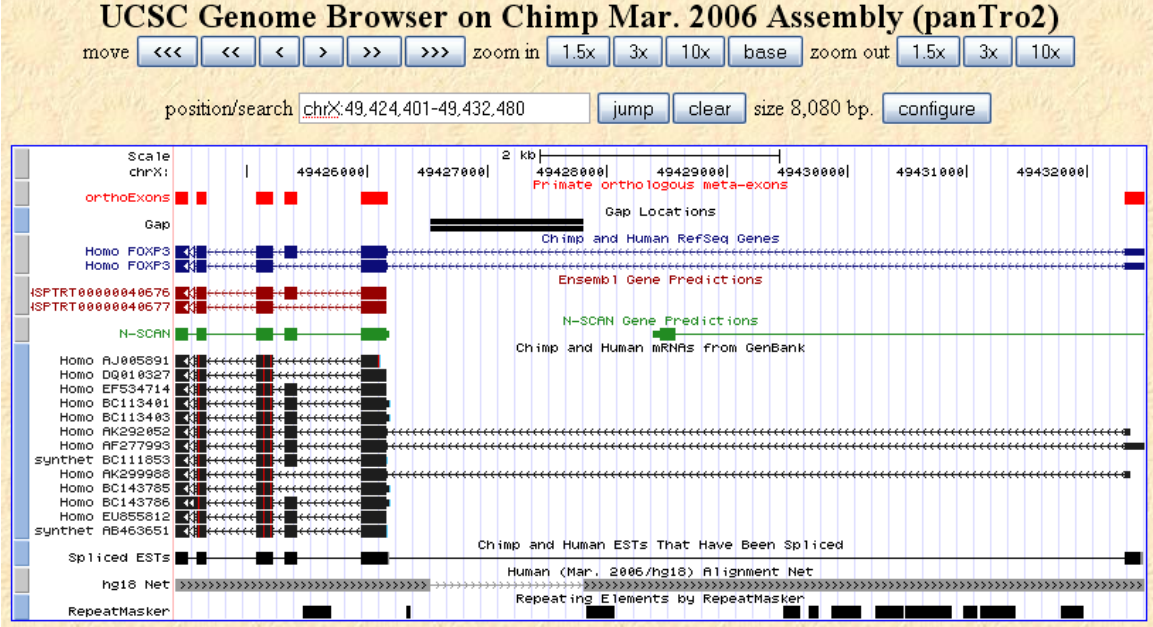
I would like to thank Matthew Stephens, Yoav Gilad, the members of the Gilad lab, and especially John Marioni for helpful discussions and comments.

Figure S1. Design of the database web application. (A) Input text box, where the user enters the name of a gene of interest. (B) auto-suggested gene names that match the text entered in A. (C) A table displaying the orthologous exon positions found in the database for *FOXP3*. (D) Links allowing the user to download the genomic coordinates as a BED file, download fasta files of the sequences, download the sequence alignment of orthologous exons and displaying the positions directly in the UCSC genome browser (see figure S2).



Exon	Human			Chimpanzee			Rhesus Macaque		
	Chr	Start	End	Chr	Start	End	Chr	Start	End
1	chrX	49000152	49000256	chrX	49424401	49424505	chrX	47109097	47109201
2	chrX	49000326	49000413	chrX	49424575	49424662	chrX	47109271	47109358
3	chrX	49000828	49000966	chrX	49425077	49425215	chrX	47109773	47109911
4	chrX	49001065	49001169	chrX	49425314	49425418	chrX	47110010	47110114
5	chrX	49001697	49001928	chrX	49425946	49426177	chrX	47110642	47110873
6	chrX	49008067	49008232	chrX	49432314	49432480	chrX	47117008	47117174
	Download as BED file Download as fasta file View in UCSC Genome Browser			Download as BED file Download as fasta file Download alignment with human View in UCSC Genome Browser			Download as BED file Download as fasta file Download alignment with human View in UCSC Genome Browser		

Figure S2. Displaying the chimpanzee (panTro2) orthologous exons for the human (hg18) *FOXP3* gene as a custom track (top, in red) in the UCSC genome browser.



References

Blekhman R, Marioni JC, Zumbo P, Stephens M, Gilad Y (2010) Sex-specific and lineage-specific alternative splicing in primates. *Genome research* **20:** 180-189

Fu X, Fu N, Guo S, Yan Z, Xu Y, Hu H, Menzel C, Chen W, Li Y, Zeng R, Khaitovich P (2009) Estimating accuracy of RNA-Seq and microarrays with proteomics. *BMC Genomics* **10:** 161

Gilad Y, Pritchard JK, Thornton K (2009) Characterizing natural variation using next-generation sequencing technologies. *Trends Genet* **25:** 463-471

Hubbard TJ, Aken BL, Ayling S, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Clarke L, Coates G, Fairley S, Fitzgerald S, Fernandez-Banet J, Gordon L, Graf S, Haider S, Hammond M, Holland R, Howe K, Jenkinson A, Johnson N, Kahari A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Kulesha E, Lawson D, Longden I, Megy K, Meidl P, Overduin B, Parker A, Pritchard B, Rios D, Schuster M, Slater G, Smedley D, Spooner W, Spudich G, Trevanion S, Vilella A, Vogel J, White S, Wilder S, Zadissa A, Birney E, Cunningham F, Curwen V, Durbin R, Fernandez-Suarez XM, Herrero J, Kasprzyk A, Proctor G, Smith J, Searle S, Flicek P (2009) Ensembl 2009. *Nucleic Acids Res* **37:** D690-697

Kent WJ (2002) BLAT--the BLAST-like alignment tool. *Genome Res* **12:** 656-664

Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A (2008) Patterns of positive selection in six Mammalian genomes. *PLoS Genet* **4:** e1000144

Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* **18:** 1509-1517

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10:** 57-63