Bioinformatics, 38(19), 2022, 4474–4480
https://doi.org/10.1093/bioinformatics/btac548
Advance Access Publication Date: 10 August 2022
Original Paper

OXFORD

Genome analysis

Detection of orthologous exons and isoforms using EGIO

Jinfa Ma [1,2], Jane Y. Wu [3,*] and Li Zhu [1,2,*]

[1]State Key Laboratory of Brain and Cognitive Science, Institute of Biophysics, Chinese Academy of Sciences, Beijing, China, [2]College of Life Sciences, University of Chinese Academy of Sciences, Beijing, China and [3]Department of Neurology, Center for Genetic Medicine, Lurie Cancer Center, Northwestern University Feinberg School of Medicine, Chicago, IL 60611, USA

*To whom correspondence should be addressed.
Associate Editor: Can Alkan

Received on March 30, 2022; revised on June 15, 2022; editorial decision on August 3, 2022; accepted on August 5, 2022

Abstract

Motivation: Alternative splicing is an important mechanism to generate transcriptomic and phenotypic diversity. Existing methods have limited power to detect orthologous isoforms.

Results: We develop a new method, EGIO, to detect orthologous exons and orthologous isoforms from two species. EGIO uses unique exonic regions to construct exon groups, in which process dynamic programming strategy is used to do exon alignment. EGIO could cover all the coding exons within orthologous genes. A comparison between EGIO and ExTraMapper shows that EGIO could detect more orthologous isoforms with conserved sequence and exon structures. We apply EGIO to compare human and chimpanzee protein-coding isoforms expressed in the frontal cortex and identify 6912 genes that express human unique isoforms. Unexpectedly, more human unique isoforms are detected than those conserved between humans and chimpanzees.

Availability and implementation: Source code and test data of EGIO are available at https://github.com/wu-lab-egio/EGIO.

Contact: zhuli@ibp.ac.cn or jane-wu@northwestern.edu

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

1 Introduction

Alternative pre-mRNA splicing, the process in which multiple transcripts are generated from a single genetic locus, is a robust mechanism for the expansion of transcriptomes, thereby contributing to functional and phenotypic diversity (Keren *et al.*, 2010; Ule and Blencowe, 2019). It is estimated that ∼95% of human protein-coding genes undergo alternative splicing (Baralle and Giudice, 2017; Pan *et al.*, 2008). Alternative splicing may result in the formation of protein products with distinct structures, a major mechanism for evolutionarily adaptive changes (Long *et al.*, 2013; O'Bleness *et al.*, 2012; Reyes *et al.*, 2013). As a vigorous mechanism proposed for phenotypic novelty (Bush *et al.*, 2017; Kahles *et al.*, 2018), alternative splicing plays an important role in regulating the expression of genes essential for cell differentiation, lineage determination, and organ development (Baralle and Giudice, 2017; Feng *et al.*, 2021).

A number of studies have been published on genetic and genomic features unique to the human lineage (Khrameeva *et al.*, 2020; O'Bleness *et al.*, 2012; Xu *et al.*, 2018). Although certain human unique features associated with alternative splicing have been reported, such as predominant splicing pattern and differential exon usage (Reyes *et al.*, 2013; Xiong *et al.*, 2018), human unique transcript isoforms have not been examined at a transcriptomic level, because existing methods have limited powers to detect all the orthologous isoforms, as reviewed in (Chakraborty *et al.*, 2021). ExTraMapper, a recently published method, could detect much more orthologous exons and isoforms than previous methods, such as OrthoExon and Exalign (Chakraborty *et al.*, 2021). However, ExTraMapper uses similarity to determine the isoform orthology, which might detect more isoforms that are not in perfect match for exon structures. Several terms have been proposed when studying orthologous isoforms or alternative splicing isoforms, such as equally spliced variant (Takeda *et al.*, 2008), splicing orthology (Zambelli *et al.*, 2010), splicing orthologous CDS (Jammali *et al.*, 2019), all emphasizing the conservation of both exon sequence and exon structure in defining orthologous isoforms.

Here, we describe exon group ideogram-based detection of orthologous exons and orthologous isoforms (EGIO), which uses a dynamic programming strategy to construct exon groups (EGs, a set of corresponding exon mappings) and exon group ideogram (EGI, a set of 5′–3′ arranged EGs). Orthologous isoforms can be detected under the direction of EGI. Comparative studies showed that EGIO was a robust method for detecting orthologous isoforms and could cover all the coding exons of orthologous gene pairs. Compared to

ExTraMapper, EGIO could detect more orthologous isoforms that were perfect matches in both exon sequence and exon structures. We applied EGIO to detect orthologous isoforms with newly assembled transcriptomes of the frontal cortex between humans and chimpanzees. Our results showed that human unique isoforms frequently arose from novel combinations of existing exons and were prevalent in the human frontal cortex.

2 Materials and methods

2.1 Overview of EGIO

Orthologous exons and isoforms were detected within orthologous gene pairs. Orthologous exons were first detected by reciprocal BlastN and then confirmed by the collinearity test: exons arranged in corresponding orders. However, the BlastN-based method was suitable for detecting large 1-to-1 exons but not for small exons or non-1-to-1 exons. To overcome these problems, a dynamic programming strategy was applied to do exon alignment, in which a set of unique coding exon regions (defined as united exons in Fu and Lin, 2012; to simplify the description, all the unique exon regions were called united exons, though a unique exon region might only map to only one exon) were first detected following local pairwise alignment and application of identity-based score frame and identity guided backtrace. The dynamic programming was operated with the guidance of reciprocal BlastN because of the stringent characteristics of reciprocal BlastN. After dynamic programming, EG was detected, then followed by EGI construction. Orthologous isoforms were transcripts containing exons in the same EGs, with the identity score of each corresponding exon passing the identity threshold. We named this pipeline EGIO (Exon Group Ideogram based detection of Orthologous Exons and Orthologous Isoforms). An overview of the EGIO pipeline is shown in Figure 1.

2.2 Detection of orthologous exons using reciprocal BlastN

The percentage of exons with orthologs was significantly lower in the UTRs than in the coding regions, and the percentage of orthologous exons with equal lengths was also lower in the UTRs than in the coding regions (Fu and Lin, 2012). Thus, we only focused on exons in the coding regions. If an exon contained both UTR and the coding region, the coding region was extracted as an individual exon. To detect orthologous exons, exon pairs should pass both the homology test and the collinearity test (to avoid duplicated exons, see Supplementary Fig. S1). We adopted published criteria (Fu and Lin, 2012; Yu *et al.*, 2004) to perform the homology test, and homologous exons should pass the following criteria: (i) Candidates with a significant BlastN E-value ($\leq$1e-5); (ii) Having $\geq$80% residues in both sequences included in the BlastN alignment; (iii) Candidates were the best hits, might be more than one hits; (iv) Conditions (i), (ii), and (iii) must be true reciprocally. BLAST+ (Altschul *et al.*, 1997) was used to do reciprocal BlastN. To be defined as orthologous exons, the homologous exons were further tested for exon collinearity. For isoforms not containing duplicated exons, if homologous exons were all located in corresponding positions, these homologous exons were considered orthologous exons. For isoforms containing duplicated exons, the first best-matched homologous exon(s) were defined as orthologous exon(s).

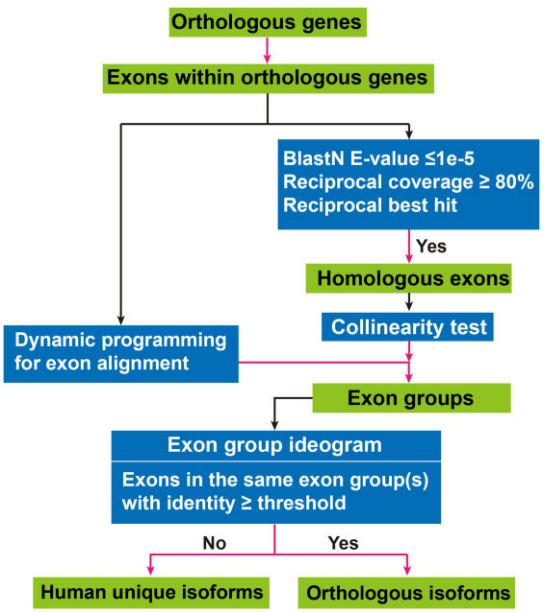

Fig. 1. A flow chart for the EGIO pipeline. Black and claret-red lines represent input to and output from the programs or filters, respectively. Green boxes mark the outputs from the databases or criteria for operation

2.3 Building EGI using a dynamic programming approach

Orthologous isoform detection is based on orthologous exonic structures. However, orthologous exons are not exclusively in the 1–1 relationship between species, and 1–N and N–1 relationships have been reported in OrthoExon and Exalign (Fu and Lin, 2012; Pavesi *et al.*, 2008). The BlastN-based method could detect large 1–1 orthologous exons but has limitations in detecting 1–N, N–1, and small orthologous exons, as well as lacking the ability to detect new exons. In addition, some exons overlap with others due to alternative splice site usage, contributing to redundant comparisons. Therefore, we simplified the exon alignment into a sequence alignment question and used a dynamic programming strategy to do exon alignment. First, a set of 5′–3′ arranged united exons were detected. Then, the identity of any of the two united exons between two species was calculated to construct a $(j+1)\times(i+1)$ identity matrix, where j and i represented total united exon numbers in two species, respectively. To detect 1–N and N–1 relationships, the identity was kept if coverage of either of the two united exons was no less than the coverage threshold (which was set as 80% in this study); otherwise, the identity was set as 0 to avoid noisy results:

$$\text{identity}(i,j) = \begin{cases} \text{identity}(i,j), & \text{if coverage}_{i \text{ or } j} \geq 80\% \\ 0, & \text{if coverage}_{i \text{ and } j} < 80\% \end{cases}$$

Next, a scored frame was operated. Given an united exon pair i and j, the score $F(i, j)$ was determined by $F(i-1, j)$, $F(i, j-1)$, $F(i-1, j-1)$, identity(i, j) and gap penalty parameter d:

$$F(i,j) = \max \begin{cases} F(i-1,j) + \text{score}\big(\text{identity}(i,j)\big) + d \\ F(i,j-1) + \text{score}\big(\text{identity}(i,j)\big) + d \\ F(i-1,j-1) + \text{score}\big(\text{identity}(i,j)\big) \end{cases}$$

A match score was obtained if the identity score exceeded the threshold; otherwise, a mismatch penalty would be given. With the principle of diagonal first, the position mismatch was given a gap penalty d. Finally, relationships of exon matches, including orthologous united exons, as well as new exon and exon loss, were detected by finding the best pathway to get the best matches from the bottom right to the top left of the scoring matrix (Eddy, 2004). However, exon alignment was more complex if one united exon could occur more than once in 1–N or N–1 relationships, so a matched exon region might be used more than once. There were two types of relationships in the non-diagonal pathway during the backtrace, true matches and gaps. The relationship was determined by the identity:

$$\text{type}(i,j) = \begin{cases} \text{map}, & \text{if identity}(i,j) \geq 80\% \\ \text{gap}, & \text{if identity}(i,j) < 80\% \end{cases}$$

This dynamic programming strategy had linear constraints. Therefore, this method was suitable for analyzing genes with or without duplicated exons, as well as those with 1–N or N–1 orthologous

exons. To integrate the collinearity test into the dynamic programming process, the united exons were arranged in the 3′–5′ direction before the analysis so that the first matches were confirmed first. After dynamic programming-based exon alignment, EGs were constructed with all the exons within the same orthologous united exons classified into the same EG. Exons with no orthologous exons were classified as a single EG. Using a local rather than global strategy during alignment may lead to a reduced detection stringency and an increase in the false-positive rate. To minimize the false-positive rate, we used orthologous exons detected by BlastN as anchors. If there was a conflict in exon mapping between results of dynamic programming and BlastN, the BlastN result would be taken as the final match. EGI was built following EG construction (See the pipeline diagram in Fig. 2).

2.4 Confirming non-1-to-1 exon mappings

To confirm non-1-to-1 mappings, an extra mapping score (MapS) frame was applied. A non-1-to-1 mapping was confirmed only if the MapS of non-1-to-1 exon mapping was more than any single exon mapping:

$$\mathrm{MapS}_{\mathrm{non-1-1}} > \mathrm{MapS}_{\mathrm{exon}\ j}\ (j = 1,\ 2, \ldots)$$

where j belongs to non-1-1 mappings. The MapS was calculated based on the sequence alignment of each nucleotide (nt):

$$\mathrm{MapS} = \sum_{i=1}^{n} \mathrm{score}_{\mathrm{nt}\ i}$$

The score of each nucleotide was calculated as follows:

$$\mathrm{score}_{\mathrm{nt}} = \begin{cases} 10, & \text{if match} \\ -10, & \text{if mismatch} \\ -8 \times \alpha^{m-1}, & \text{if gap} \end{cases}$$

For continuous gaps, the first gap would be given a −8 penalty, whereas the continuous gaps would be less punished by multiplying the former gap penalty by a constant α ($0<\alpha<1$, which was set as 0.8 in EGIO); m was the position of a gap in the continuous gaps.

2.5 Detection of orthologous isoforms using an EGI-based strategy

Orthologous isoforms were detected following EGI construction. Orthologous isoforms were defined as transcripts containing exons belonging to the same EGs, with each corresponding exon pair no less than the identity threshold (Supplementary Fig. S1). Briefly, isoforms of different species were arranged based on EGI (Supplementary Fig. S2, *FCRL5* was shown as an example), then a comparison was operated between any two isoforms to confirm their orthology. With EGI, the collinearity test was easy to operate. Therefore, the orthology of isoforms would be guaranteed once the corresponding exon sequence exceeded the identity threshold.

3 Results

3.1 EGIO is a robust method to detect orthologous isoforms

EGIO involves several essential parameters, including sequence/exon identity, coverage of local pairwise alignment, score frame of dynamic programming, and size of microexon, defined as exon with 3–15 nt (Irimia *et al.*, 2014), or 3–27 nt (Gonatopoulos-Pournatzis and Blencowe, 2020). To optimize the parameters, we detected orthologous exons and isoforms of humans and monkeys. The thresholds for sequence identity and coverage were set to 80% as previously described (Fu and Lin, 2012; Yu *et al.*, 2004).

We generated a set of reference orthologous isoforms by sequence alignment and exonic arrangement (Fig. 3A). EGIO used local alignment so that a microexon might be aligned to more than one exon; thus microexons should be isolated for additional analysis to confirm the orthology. However, we found EGIO was insensitive to microexon size, even larger microexon thresholds would decrease its sensitivity (Fig. 3B). We also evaluated the effects of different score frames of dynamic programming on the results. As shown in Figure 3C, the sensitivity was relatively stable if the mismatch penalty was no less than the gap penalty. This was because the exon structures of orthologous genes were not always in a perfect 1–1 match, and a lower gap penalty allowed easily opening a gap than generating a mismatch. To be noted, the precision of EGIO was indeed stable under different microexon size thresholds and dynamic programming score frames.

3.2 EGIO detects orthologous isoforms with perfect exon mappings

To evaluate the performance of EGIO, we compared EGIO results with that from ExTraMapper (Chakraborty *et al.*, 2021). In comparison with previous methods, i.e. InParanoid and OrthoExon for exons and Exalign for isoforms, ExTraMapper detected more orthologous exons and isoforms (Chakraborty *et al.*, 2021). Compared with ExTraMapper, EGIO used united exons, rather than individual exons, to detect orthologous exons, thereby reducing redundant mappings (Fig. 4A). EGIO detected nearly the same orthologous exon mappings as ExTraMapper (Fig. 4B, 0.6% versus 1.6% pipeline-unique results). We then chose orthologous isoforms with a coding similarity score of 0.8, 0.9, and 1.0 detected by ExTraMapper as the reference to calculate the sensitivity and precision of EGIO. As shown in Figure 4C, EGIO was sensitive to coding sequence conserved isoforms. However, with the increasing coding similarity score of the reference, the precision of EGIO decreased due to the detection of more orthologous isoform mappings. On the

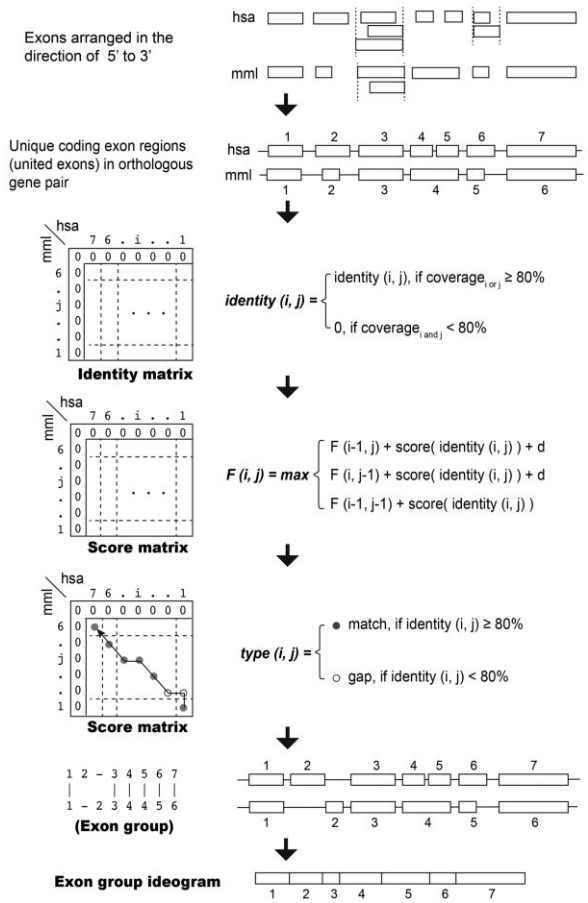

Fig. 2. A flow chart for building an EGI using a dynamic programming strategy

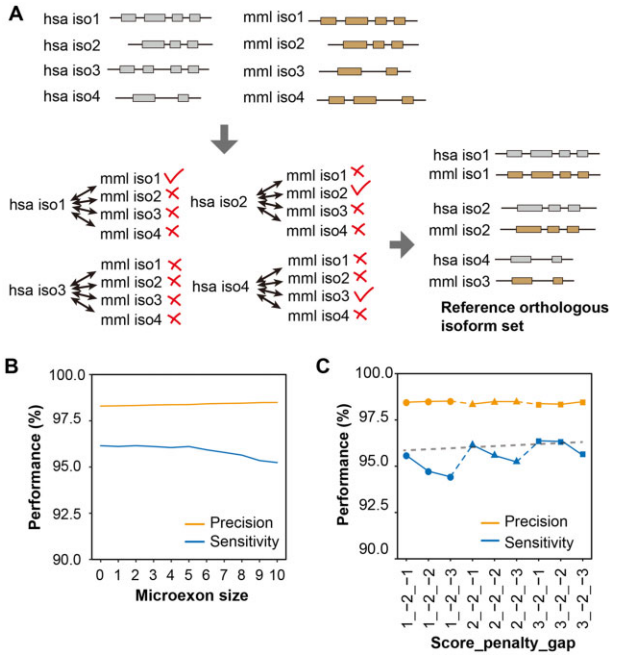

Fig. 3. Sensitivity and precision of the EGIO pipeline using different parameters. The EGIO pipeline involves four important parameters, including identity and coverage during local pairwise alignment, microexon size, and the score frame of dynamic programming. The identity and coverage are all set to 80%. (**A**) A diagram of the exon-sequence-based method to generate reference data of orthologous isoforms. Any two isoform combinations in an orthologous gene pair are compared, and they are orthologous isoform candidates only if the two isoforms are composed of the same number of exons and each corresponding exon passes the homologous test (identity ≥80%). Although isoforms with a split exon or a fusion exon are excluded and isoforms containing duplicated exons may be included, these isoforms represent only a small fraction of the total isoforms. In addition, this reference dataset may also contain false orthologous isoforms that fail the collinearity test, as shown in Supplementary Figure S1 Mechanism 4. hsa, homo sapiens; mml, macaca mulatta. (**B**) Sensitivity and precision with different threshold settings of microexon size (score frame: match score: 2, mismatch penalty: −2, gap penalty: −1). (**C**) Sensitivity and precision with different score frames (microexon size: 2). To achieve higher sensitivity, the mismatch score should be no less than the gap penalty, indicated by the dashed line. Sensitivity = TP/(TP+FN); Precision = TP/(TP+FP). TP, true positives, defined as isoforms detected by both EGIO and exon-sequence based methods; FN, false negatives, defined as isoforms are detected only by exon-sequence based methods; FP, false positives, defined as isoforms that are detected only by EGIO

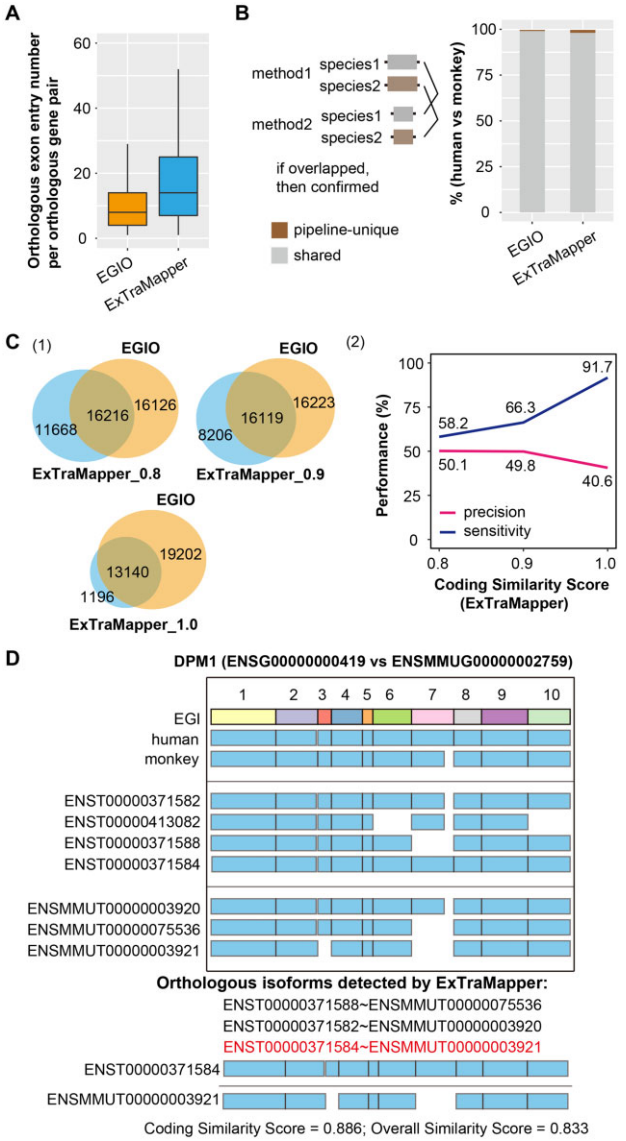

Fig. 4. Comparison between EGIO and ExtraMapper. ExTraMapper provides processed results of humans versus monkeys using Ensembl reference transcriptomes (release version 102). The same transcriptomes were used to detect orthologous exons and isoforms with EGIO. (**A**) Box plots of orthologous exon entry number per orthologous gene pair. EGIO uses united exons to detect orthologous exons, which eliminates redundant results. (**B**) Comparison of orthologous exons detected by EGIO and ExTraMapper. Because of the reason described in (A), orthologous exons might not be the perfect match. If an orthologous exon pair of one method overlaps with that of another, this orthologous exon pair is confirmed. (**C**) Performance of EGIO. Orthologous isoforms with coding similarity scores over 0.8, 0.9, and 1.0 detected by ExTraMapper are used as the reference. The left panel shows the Venn diagram, and the right panel shows the performance of EGIO. (**D**) An example to illustrate that ExTraMapper uses a less stringent definition of orthologous isoforms. Sensitivity = TP/(TP+FN); Precision = TP/(TP+FP). TP, true positives; FN, false negatives; FP, false positives. TP, isoforms detected by both EGIO and ExTraMapper; FN, isoforms detected only by ExTraMapper; FP, isoforms detected only by EGIO. Parameters for human versus monkey using EGIO: identity: 80%, coverage: 80%, microexon size: 2, match score: 2, mismatch penalty: −2, gap penalty: −1

other hand, ExTraMapper detected orthologous isoforms with less stringent exon structure constraints. For example, human ENST00000371584 and monkey ENSMMUT00000003921 of gene DPM1 were counted as orthologous isoforms by ExTraMapper, whereas these transcripts were not defined as an orthologous isoform pair in EGIO analysis (Fig. 4D). Taken together, our comparison showed that EGIO had higher accuracy and stringency, and it could detect more orthologous exon mappings and identify orthologous isoform that was conserved in both sequences and exon structures.

Comparison of human versus mouse showed similar results with that of human versus monkey (Supplementary Fig. S3), in which EGIO detected nearly the same exonic region mappings as ExTraMapper but more orthologous isoforms than ExTraMapper. In addition, EGIO detected more orthologous exon regions than OrthoExon and InParanoid (Supplementary Fig. S3A). Exalign was designed to align exonic structures based on exon length (Pavesi *et al.*, 2008). Compared with Exalign, EGIO also detected more orthologous isoforms (Supplementary Fig. S3B, see orthologous isoform detection by Exalign in Supplementary Methods). When using different EGIO orthologous isoform sets with various coding similarities as queries, EGIO showed similar performance (Supplementary Fig. S3C and Fig. 4C), because orthologous isoforms detected by EGIO had more considerable coding similarities, and changes in coding similarity thresholds would not significantly affect query sample volume.

3.3 Construction of human and chimpanzee transcriptomes using frontal cortex samples

Since chimpanzees are the closest relatives to humans, we compare orthologous exons and isoforms between chimpanzees and humans. We focus on datasets from the frontal cortex because the following reasons: alternative splicing is more prevalent in the mammalian nervous system (Raj and Blencowe, 2015); the frontal cortex plays a critical role in cognitive function. Nine datasets (GSE100796, GSE124439, GSE135036, GSE47966, GSE49379, GSE58604, GSE59288, GSE68719, and GSE80655) in GEO were included, containing 111 human and 57 chimpanzee frontal cortex samples (Supplementary Table S1). Over 90% of reads were retained after read-trimming for GSE124439, GSE135036, GSE47966, GSE49379, GSE58604, GSE59288, GSE68719, and over 80% of reads were retained for GSE100796 and GSE80655 (Supplementary Fig. S4A). RNA-seq read alignment was performed using STAR (Dobin *et al.*, 2013) and HISAT2 (Kim *et al.*, 2015) on adapter-trimmed reads to the reference genome (Ensembl release version 102). The unique alignment rate was over 75% for all the samples using HISAT2 and STAR (Supplementary Fig. S4B).

StringTie (Pertea *et al.*, 2015) and TACO (Niknafs *et al.*, 2017) were used to detect transcripts and merge individual transcriptomes after genome-wide alignment. To reduce potential false-positive results, we used two pipelines, STAR/StringTie/TACO (SST) and HISAT2/StringTie/TACO (HST), to assemble transcriptomes from RNA-seq datasets. Next, Gffcompare (Pertea and Pertea, 2020) was used to compare the newly assembled transcriptomes, which generated 59 331 and 38 418 overlapping transcripts (class code: =) in humans and chimpanzees (Supplementary Fig. S5A). Possible noisy transcripts were then filtered as described in the Supplementary Methods section based on a previous study (Pertea *et al.*, 2018). Finally, a total of 26 188 and 17 101 filtered transcripts for human and chimpanzee samples were obtained, respectively (Supplementary Fig. S5A). We compared the filtered transcripts with Ensembl reference transcriptomes using Gffcompare. As shown in Supplementary Figure S5B, 78.9% and 42.9% of transcripts showed exact matches to Ensembl annotations (class code: =) in humans and chimpanzees; whereas the remaining transcripts in classes c, i, j, k, o, u, x, y were newly identified transcripts. Finally, 5523 novel transcripts in humans and 9759 novel transcripts in chimpanzees were detected. The protein-coding potential of these novel transcripts was then detected using CPC2 (Kang *et al.*, 2017) and CPAT (Wang *et al.*, 2013).

To build more complete transcriptomes, we merged our newly identified transcripts and Ensembl reference transcriptomes as the final new reference transcriptomes (The newly assembled transcriptomes can be downloaded at https://github.com/wu-lab-egio/EGIO_example_source, see the whole workflow for transcriptome construction in Supplementary Fig. S6). We focused on protein-coding genes and protein-coding transcripts (which were referred to as 'isoforms' in this study), including those annotated to be nonsense-mediated decay (NMD) in Ensembl. In the new reference transcriptomes, 20 094 and 24 140 protein-coding genes were identified for humans and chimpanzees, respectively, of which 121 in humans and 838 in chimpanzees were newly assembled genes that have not been annotated in Ensembl. As shown in Supplementary Figure S5C, human protein-coding genes have more annotated protein-coding isoforms than chimpanzees, with the average transcript numbers per gene being 5.3 (106 764/20 094) and 2.4 (57 982/24 140) in human and chimpanzee, respectively.

3.4 Human unique splicing isoforms are prevalent in the frontal cortex

EGIO requires pre-defined orthologous gene pairs as guidance. Using InParanoid and BlastP analyses (Supplementary Methods), 18 762 orthologous gene pairs between humans and chimpanzees were detected, of which 16 582 orthologous gene pairs were recorded by InParanoid and the rest 2180 pairs were detected by reciprocal BlastP (Supplementary Data 1). All of these orthologous gene pairs correspond to 89.3% (17 939/20 094) of human protein-coding genes in the above reference transcriptomes. Inparanoid used non-redundant UniProt ID to detect orthologous genes; there were $1-N$ or $N-1$ orthologous genes after transforming UniProt ID into Ensembl Gene ID; thus numbers of genes with orthologous relatives were less than 18 762.

Following EGIO analysis, we compared the isoforms of orthologous genes in the human frontal cortex samples with those in chimpanzees (see Supplementary Data 2 for EGs and Supplementary Data 3 for a complete list of protein-coding isoforms). A total of 100 467 human isoforms were included in our analysis, and 96 655 of them were annotated in Ensembl (Fig. 5A, left panel), whereas the remaining 3812 (~3.8%) were previously unannotated transcripts in our newly assembled human transcriptome (Fig. 5A, right panel). An analysis using the newly assembled chimpanzee transcriptome led to the identification of a total of 1900 additional orthologous isoforms in humans, including 1349 from annotated isoforms and 551 from unannotated isoforms, respectively (Fig. 5A). Finally, a total of 42 011 orthologous isoforms and 58 456 human unique isoforms were identified. Surprisingly, more human unique isoforms were detected than orthologous isoforms, although there are only <2% sequence differences between human and chimpanzee over the entire genomes (Hacia, 2001). Detection of these human unique isoforms suggests that alternative splicing may play an even more

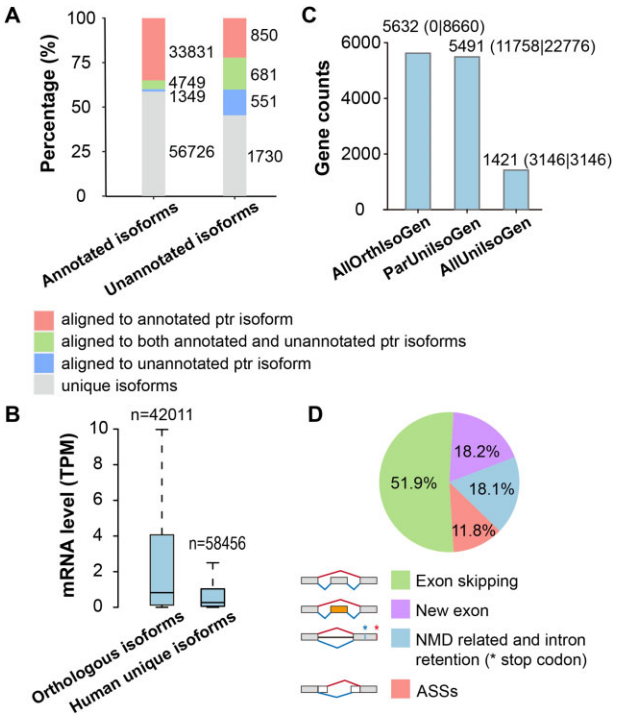

Fig. 5. A summary of human unique isoforms expressed in the frontal cortex. (**A**) Bar graphs of different types of human transcript isoforms. Annotated isoforms are those in the Ensembl reference transcriptomes, and unannotated isoforms are those newly assembled transcripts. (**B**) Box plots of mRNA levels of human protein-coding isoforms. Outliers are not included. (**C**) Bar plots of gene numbers of different gene groups. Protein coding isoforms with expression levels at TPM ≥1 are included in the analysis. Their corresponding genes are classified into three different groups: genes that express isoforms which are all orthologous (AllOrthIsoGen), genes that express both orthologous isoforms and human unique isoforms (ParUniIsoGen); genes that express all isoforms as human unique isoforms (AllUniIsoGen). In brackets are the numbers of human unique isoform numbers | the total number of expressed isoforms in different gene groups. (**D**) Different mechanisms to generate human unique isoforms. Exon skipping is the most frequent type. When an isoform contains exon(s) not in the orthologous exon group, it is classified into the new exon group. The NMD type is defined by the same criteria as in Ensembl, including intron events that lead to premature stop codons. If a human unique isoform passes the collinearity test (all the corresponding exons in the same EGs) but fails the homologous test (i.e. with sequence identity <80%), it will be classified into the group of alternative splice site selection (ASSs)

important role than previously anticipated in the origin of human unique higher nervous activity.

Using the newly assembled reference transcriptome, we quantified mRNA expression levels using Salmon (Patro *et al.*, 2017). As shown in Figure 5B, a significant fraction of human unique isoforms had lower expression levels than orthologous isoforms. Some isoforms with very low expression levels may be noises. Therefore, we used mean transcripts per million (TPM) $\geq$1 as the cut-off to define expressed human unique isoforms. After this expression filtering, 34 582 expressed isoforms encoded by 12 544 genes were obtained. Among these 12 544 genes, 5632 genes contained all expressed isoforms as orthologous isoforms, 5491 genes showed a part of expressed isoforms as orthologous isoforms, whereas 1421 genes expressed all splicing isoforms as human unique isoforms (Fig. 5C; and see *RTN4R* in Supplementary Fig. S2C as an example for genes expressing all isoforms as human unique isoforms; see Supplementary Data 4 for gene list of different groups).

Expressed isoforms were then analyzed and classified according to different splicing mechanisms, including exon skipping, the inclusion of new exon, NMD-related, and alternative splice site selection. Exon skipping generates 51.9% of human unique isoforms and the inclusion of new exons contributes to 18.2% of human unique isoforms, respectively (Fig. 5D). NMD contributes to 18.1% of human unique isoforms. At last, about 11.8% of human unique isoforms are generated by usage of alternative splice sites, as in 1−*N* and *N*−1 exons (Supplementary Fig. S2).

To illustrate the potential role of human-unique isoforms, we took *MBP* as an example (Supplementary Fig. S7). *MBP* encodes myelin basic protein and is mainly expressed in the nervous system (Supplementary Fig. S7A). Compared with the chimpanzee, orthologous isoforms ENST00000578193 and ENST00000359645 are downregulated in human, whereas human unique isoform ENST00000397865 is upregulated, which makes it the second highest expressed isoform (Supplementary Fig. S7B). Compared with the most abundant isoform ENST00000397866 in human and ENSPTRT00000092864 in chimpanzee, ENST00000397865 lacks the exon in EG 11 (Supplementary Fig. S7B). Based on protein domain annotation, the skipped exon 11 encodes a domain associated with experimental autoimmune encephalomyelitis (EAE). Two EAE domains exist in ENST00000397866. Skipping of the second EAE domain encoded by exon 11 might contribute to the adaption in human, leading to high expression of ENST00000397865. However, the hypothesis needs further experimental validation.

4 Discussion

The concept of clusters of orthologous groups has been used at the gene level (Tatusov *et al.*, 1997). Extending from orthologous groups, we propose to use the concept of EGs to study alternative splicing isoforms. EG refers to a group of exons of the same origin based on their positions and sequences. The first step of EGIO is to transform isoform comparison to sequence alignment (Fig. 2). To our knowledge, it is the first study in which dynamic programming is applied to systematic analyses in exon alignment and detection of orthologous exon mappings based on exon sequence. This dynamic programming strategy enables us to detect all kinds of orthologous exons (even microexon as small as only one nucleotide). In addition, the linear constraint of dynamic programming enhances the detection of orthologous isoforms. With a BlastN-guided model, EGIO takes advantage of the dynamic programming strategy and the stringent criteria for the reciprocal BlastN method. In this study, we only include protein-coding genes. With proper adjustment of parameters, the EGIO pipeline can also be used for studying non-protein-coding genes.

It has been suggested that alternative splicing should be considered when defining orthologous genes at the transcript level (Jia *et al.*, 2010; Zambelli *et al.*, 2010). In addition to gene sequences, conservation of exon/intron structures should be included when defining orthologous genes. Compared with ExTraMapper, EGIO utilizes more stringent constraints of exon structure conservation, which enables the detection of more orthologous isoforms with perfect alignment, both in sequence and in exon structure. Using more stringent criteria, EGIO still detects more orthologous isoforms than ExTraMapper (Fig. 4C). The current pipeline of EGIO also detects more *N*–*N* orthologous isoforms, whereas ExTraMapper could detect more orthologous isoforms with a 1–1 relationship. One limitation of the current EGIO pipeline is that those transcript isoforms that are distinct only in the UTR regions are missing in our data output because the UTR regions are excluded from our analyses.

Although orthologous genes have been studied in primates, much less is known about global gene expression at the isoform level. This is critical because functionally distinct or even antagonistic isoforms can be generated from the same genetic locus (Ule and Blencowe, 2019; Wu *et al.*, 2003). Therefore, orthologous gene pairs may not be directly translated into functional equivalency when species-distinct or unique alternative splicing exists. We systematically apply EGIO to compare human and chimpanzee transcriptomes in the frontal cortex. Our study has revealed that 1421 orthologous genes express all isoforms as human unique isoforms. Recently, a nanopore-based single-molecule peptide detection has been reported for detecting individual proteins at the single amino acid resolution (Brinkerhoff *et al.*, 2021). Such tools, when combined with transcript isoform analyses, will enable us to detect different protein isoforms from individual genes.

A variety of methods have been developed to detect differential alternative splicing, local splicing variation, and quantify isoform expression, including rMATs, LeafCutter, LSVs, and sleuth (Li *et al.*, 2018; Pimentel *et al.*, 2017; Shen *et al.*, 2014; Vaquero-Garcia *et al.*, 2016). EGIO can provide information on orthologous exons and isoforms, and it can be used in combination with these quantitative methods to compare differentially alternative splicing events and differentially expressed isoforms among different species. Such comparative studies at global transcript isoform levels between humans and other species will advance our understanding of human unique gene expression and regulation.

Acknowledgements

The authors thank Dr Wen Wang (Northwest University, China) and Dr Yong Zhang (Institute of Zoology, CAS, China) for their valuable suggestions on orthologous isoform detection. The authors thank Dr Warren A. McGee and Dr David Kuo for their valuable suggestions on RNA-seq pipelines. The authors thank Dr Yu-Heng Lu (Harvard Medical School) for his suggestions to optimize EGIO parameters.

Funding

This work was supported by the National Key R&D Program of China [2019YFA0508603 to L.Z. and J.M.]; the National Natural Science Foundation of China [31971075 to L.Z.]; and the National Institutes of Health [R01CA175360 to J.Y.W.].

Conflict of Interest: none declared.

Data availability

RNA-seq datasets used in this study are available under accession numbers GSE100796, GSE124439, GSE135036, GSE47966, GSE49379, GSE58604, GSE59288, GSE68719 and GSE80655. Control samples of the frontal cortex in RNA-seq datasets are included in the study. The source data of EGIO and new transcriptomes of human/chimpanzee are deposited on github online; the link is listed in the article. All the other data are incorporated into the article and its online supplementary material.

References

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Baralle,F.E. and Giudice,J. (2017) Alternative splicing as a regulator of development and tissue identity. *Nat. Rev. Mol. Cell Biol.*, **18**, 437–451.

Brinkerhoff,H. *et al.* (2021) Multiple rereads of single proteins at single-amino acid resolution using nanopores. *Science*, **374**, 1509–1513.
Bush,S.J. *et al.* (2017) Alternative splicing and the evolution of phenotypic novelty. *Philos. Trans. R Soc. Lond. B Biol. Sci.*, **372**, 20150474.
Chakraborty,A. *et al.* (2021) ExTraMapper: exon- and transcript-level mappings for orthologous gene pairs. *Bioinformatics*, **37**, 3412–3420.
Dobin,A. *et al.* (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
Eddy,S.R. (2004) What is dynamic programming? *Nat. Biotechnol.*, **22**, 909–910.
Feng,H. *et al.* (2021) Complexity and graded regulation of neuronal cell-type-specific alternative splicing revealed by single-cell RNA sequencing. *Proc. Natl. Acad. Sci. USA*, **118**, e2013056118.
Fu,G.C. and Lin,W.C. (2012) Identification of gene-oriented exon orthology between human and mouse. *BMC Genomics*, **13** (Suppl. 1), S10.
Gonatopoulos-Pournatzis,T. and Blencowe,B.J. (2020) Microexons: at the nexus of nervous system development, behaviour and autism spectrum disorder. *Curr. Opin. Genet. Dev.*, **65**, 22–33.
Hacia,J.G. (2001) Genome of the apes. *Trends Genet.*, **17**, 637–645.
Irimia,M. *et al.* (2014) A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell*, **159**, 1511–1523.
Jammali,S. *et al.* (2019) SplicedFamAlign: CDS-to-gene spliced alignment and identification of transcript orthology groups. *BMC Bioinformatics*, **20**, 133.
Jia,Y. *et al.* (2010) Refining orthologue groups at the transcript level. *BMC Genomics*, **11** (**Suppl. 4**), S11.
Kahles,A. *et al.*; Cancer Genome Atlas Research Network. (2018) Comprehensive analysis of alternative splicing across tumors from 8,705 patients. *Cancer Cell*, **34**, 211–224.e6.
Kang,Y.J. *et al.* (2017) CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res.*, **45**, W12–W16.
Keren,H. *et al.* (2010) Alternative splicing and evolution: diversification, exon definition and function. *Nat. Rev. Genet.*, **11**, 345–355.
Khrameeva,E. *et al.* (2020) Single-cell-resolution transcriptome map of human, chimpanzee, bonobo, and macaque brains. *Genome Res.*, **30**, 776–789.
Kim,D. *et al.* (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods*, **12**, 357–360.
Li,Y.I. *et al.* (2018) Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.*, **50**, 151–158.
Long,M. *et al.* (2013) New gene evolution: little did we know. *Annu. Rev. Genet.*, **47**, 307–333.
Niknafs,Y.S. *et al.* (2017) TACO produces robust multisample transcriptome assemblies from RNA-seq. *Nat. Methods*, **14**, 68–70.
O'Bleness,M. *et al.* (2012) Evolution of genetic and genomic features unique to the human lineage. *Nat. Rev. Genet.*, **13**, 853–866.
Pan,Q. *et al.* (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.*, **40**, 1413–1415.
Patro,R. *et al.* (2017) Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods*, **14**, 417–419.
Pavesi,G. *et al.* (2008) Exalign: a new method for comparative analysis of exon-intron gene structures. *Nucleic Acids Res.*, **36**, e47.
Pertea,G. and Pertea,M. (2020) GFF utilities: gffRead and GffCompare. *F1000Res.*, **9**, 304.
Pertea,M. *et al.* (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.*, **33**, 290–295.
Pertea,M. *et al.* (2018) CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome Biol.*, **19**, 208.
Pimentel,H. *et al.* (2017) Differential analysis of RNA-seq incorporating quantification uncertainty. *Nat. Methods*, **14**, 687–690.
Raj,B. and Blencowe,B.J. (2015) Alternative splicing in the mammalian nervous system: recent insights into mechanisms and functional roles. *Neuron*, **87**, 14–27.
Reyes,A. *et al.* (2013) Drift and conservation of differential exon usage across tissues in primate species. *Proc. Natl. Acad. Sci. USA*, **110**, 15377–15382.
Shen,S. *et al.* (2014) rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc. Natl. Acad. Sci. USA*, **111**, E5593–E5601.
Takeda,J-I. *et al.* (2008) Low conservation and species-specific evolution of alternative splicing in humans and mice: comparative genomics analysis using well-annotated full-length cDNAs. *Nucleic Acids Res.*, **36**, 6386–6395.
Tatusov,R.L. *et al.* (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.
Ule,J. and Blencowe,B.J. (2019) Alternative splicing regulatory networks: functions, mechanisms, and evolution. *Mol. Cell*, **76**, 329–345.
Vaquero-Garcia,J. *et al.* (2016) A new view of transcriptome complexity and regulation through the lens of local splicing variations. *Elife*, **5**, e11752.
Wang,L. *et al.* (2013) CPAT: coding-potential assessment tool using an alignment-free logistic regression model. *Nucleic Acids Res.*, **41**, e74.
Wu,J.Y. *et al.* (2003) Alternative pre-mRNA splicing and regulation of programmed cell death. *Prog. Mol. Subcell. Biol.*, **31**, 153–185.
Xiong,J. *et al.* (2018) Predominant patterns of splicing evolution on human, chimpanzee and macaque evolutionary lineages. *Hum. Mol. Genet.*, **27**, 1474–1485.
Xu,C. *et al.* (2018) Human-specific features of spatial gene expression and regulation in eight brain regions. *Genome Res.*, **28**, 1097–1110.
Yu,H. *et al.* (2004) Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Res.*, **14**, 1107–1118.
Zambelli,F. *et al.* (2010) Assessment of orthologous splicing isoforms in human and mouse orthologous genes. *BMC Genomics*, **11**, 534.